

**ANALISIS POLA PENJUALAN MENGGUNAKAN MODEL
CRISP-DM DENGAN ALGORITMA K-MEANS (STUDI
KASUS: WAHYU JAYA GROUP CABANG SEMARANG
TAHUN 2023)**

SKRIPSI

Diajukan untuk Memenuhi Sebagian Syarat Guna Memperoleh
Gelar Sarjana Matematika
dalam Ilmu Matematika



Oleh : **RAHMAT BAGUS PANGESTU**
NIM : 2008046040

FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI WALISONGO
SEMARANG
2024

PERNYATAAN KEASLIAN

Yang bertanda tangan di bawah ini :

Nama : Rahmat Bagus Pangestu
NIM : 2008046040
Jurusan/Program Studi : Matematika/ Matematika

menyatakan bahwa skripsi yang berjudul :

**ANALISIS POLA PENJUALAN MENGGUNAKAN MODEL
CRISP-DM DENGAN ALGORITMA K-MEANS (STUDI KASUS:
WAHYU JAYA GROUP CABANG SEMARANG TAHUN 2023)**

secara keseluruhan adalah hasil penelitian/karya saya sendiri,
kecuali bagian tertentu yang dirujuk sumbernya.

Semarang, 16 Oktober 2024
Pembuat pernyataan,



Rahmat Bagus Pangestu
NIM : 2008046040



KEMENTERIAN AGAMA R.I.
UNIVERSITAS ISLAM NEGERI WALISONGO
FAKULTAS SAINS DAN TEKNOLOGI
Jl. Prof. Dr. Hamka (Kampus II) Ngaliyan Semarang
Telp. 024-7601295 Fax. 7615387

PENGESAHAN

Naskah skripsi berikut ini :

Judul : **Analisis Pola Penjualan Menggunakan Model CRISP-DM dengan Algoritma K-Means (Studi Kasus: Wahyu Jaya Group Cabang Semarang Tahun 2023)**

Penulis : Rahmat Bagus Pangestu

NIM : 2008046040

Jurusan : Matematika

Telah diujikan dalam sidang *tugas akhir* oleh Dewan Penguji Fakultas Sains dan Teknologi UIN Walisongo dan dapat diterima sebagai salah satu syarat memperoleh gelar sarjana dalam Ilmu Matematika.

Semarang, 16 Oktober 2024

DEWAN PENGUJI

Penguji I,

Zulaikha, M.Si

NIP : 199204092019032027

Penguji III,

Nur Khasanah, M.Si

NIP : 199111212019032625

Penguji II,

Ariska Kurnia Rachmawati,

M.Sc

NIP : 198908112019032019

Penguji IV,

Emy Siswanah, M.Sc

NIP : 198702022011012014

Pembimbing,

Ariska Kurnia Rachmawati, M.Sc

NIP : 198908112019032019

NOTA DINAS

Semarang, 19 September 2024

Yth. Ketua Program Studi Matematika
Fakultas Sains dan Teknologi
UIN Walisongo Semarang

Assalamu'alaikum warahmatullahi wabarakatuh

Dengan ini diberitahukan bahwa saya telah melakukan bimbingan, arahan dan koreksi naskah skripsi dengan:

Judul : Analisis Pola Penjualan Menggunakan Model
CRISP-DM dan Algoritma K-Means: Studi Kasus
Wahyu Jaya Group Cabang Semarang Tahun 2023
Nama : Rahmat Bagus Pangestu
NIM : 2008046040
Jurusan : Matematika

Saya memandang bahwa naskah skripsi tersebut sudah dapat diajukan kepada Fakultas Sains dan Teknologi UIN Walisongo untuk diujikan dalam Sidang Munaqasyah.

Wassalamu'alaikum warahmatullahi wabarakatuh

Pembimbing,



Ariska Kurnia Rachmawati, M.Sc
NIP : 19890811 201903 2 019

ABSTRAK

Penelitian ini menganalisis pola penjualan di Wahyu Jaya Group Cabang Semarang pada tahun 2023 menggunakan model *CRISP-DM* (*Cross-Industry Standard Process for Data Mining*) dengan algoritma *K-Means*. Tujuan penelitian ini adalah untuk menemukan pola penjualan yang relevan guna meningkatkan kinerja bisnis dan strategi penjualan produk di Kota Semarang. Model *CRISP-DM* dipilih karena strukturnya yang sistematis dalam mengelola proyek data mining, mulai dari pemahaman bisnis hingga penyebaran model. Data penjualan yang digunakan berasal dari bulan November 2023 dan dianalisis dengan algoritma *K-Means* untuk mengelompokkan produk berdasarkan wilayah di Kota Semarang. Algoritma *K-Means* mengelompokkan data penjualan berdasarkan karakteristik tertentu, sementara metode *Silhouette Coefficient* digunakan untuk menilai kualitas kluster dan menentukan jumlah kluster optimal, yaitu 2 kluster dengan nilai struktur kluster sebesar 0,6081 yang menunjukkan kualitas kluster yang cukup kuat dan potensi tumpang tindih rendah.

Data yang dianalisis mencakup 16 kecamatan di Kota Semarang dan diolah menggunakan perangkat lunak seperti *MATLAB* 2013, *RStudio* 2024.09.0+375, dan *Microsoft Excel* 2013. Hasil penelitian menunjukkan bahwa penerapan model *CRISP-DM* dengan algoritma *K-Means* efektif dalam mengidentifikasi segmen pasar potensial dan memberikan wawasan bagi pengambilan keputusan strategis perusahaan. Dua kluster dalam pola penjualan produk AMDK jenis botol dan gelas menunjukkan bahwa kluster pertama memiliki tingkat penjualan tinggi dengan pangsa pasar 62,01%, sedangkan kluster kedua menunjukkan tingkat penjualan rendah dengan pangsa pasar 37,99%. Implikasi dari hasil ini memungkinkan perusahaan merancang pendekatan yang lebih efektif di setiap wilayah, seperti meningkatkan promosi di wilayah dengan penjualan rendah atau mempertahankan strategi yang berhasil di wilayah dengan penjualan tinggi. Hal ini diharapkan tidak hanya membantu mengenali pola penjualan,

tetapi juga menyediakan dasar yang kuat bagi perusahaan untuk memperkuat posisi kompetitif dan meningkatkan kinerja bisnis di masa mendatang.

Kata kunci : *CRISP-DM, K-Means, Silhouette Coefficient*, Penjualan, Segmentasi Pasar

KATA PENGANTAR

Dengan membaca *Alhamdulillah* sebagai bentuk rasa syukur ke hadirat Allah SWT atas segala rahmat, hidayah, serta karunia-Nya, penulis dapat menyelesaikan penulisan skripsi yang berjudul "*Analisis Pola Penjualan Menggunakan Model CRISP-DM dengan Algoritma K-Means (Studi Kasus: Wahyu Jaya Group Cabang Semarang Tahun 2023)*" dengan baik.

Skripsi ini ditulis untuk memenuhi salah satu syarat memperoleh gelar Sarjana Strata 1 (S1) pada Program Studi Matematika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Walisongo Semarang. Dalam penulisan skripsi ini, penulis menghadapi beberapa kesulitan sehingga penulisan ini tidak lepas dari bantuan berbagai pihak. Oleh karena itu, penulis mengucapkan terima kasih kepada:

1. Bapak Prof. Dr. Nizar, M.Ag., selaku Rektor Universitas Islam Negeri Walisongo Semarang.
2. Bapak Prof. Dr. H. Musahadi, M.Ag., selaku Dekan Fakultas Sains dan Teknologi.
3. Ibu Any Muanalifah, M.Si., selaku Ketua Prodi Matematika.
4. Ibu Ariska Kurnia Rachmawati, M.Sc., selaku Dosen Pembimbing yang telah meluangkan waktu dan pikiran serta memberikan bimbingan, pengarahan, dan saran yang bermanfaat dalam proses penulisan skripsi ini.
5. Ibu Zulaikha, M.Si., selaku Dosen Wali yang telah meluangkan waktu dan pikiran serta memberikan bimbingan,

pengarahan, dan saran yang bermanfaat selama kuliah dan proses penulisan skripsi ini.

6. Bapak/Ibu Dosen dan Staf Program Studi Matematika Fakultas Sains dan Teknologi UIN Walisongo Semarang yang telah mendidik, mengarahkan dan membantu selama kuliah dan proses penulisan skripsi ini.
7. Bapak, Ibu, dan Kakak yang selalu memberikan doa, dukungan, dan motivasi kepada penulis selama proses penulisan skripsi ini.
8. Teman-teman angkatan 2020 Program Studi Matematika UIN Walisongo Semarang yang telah membantu dalam kegiatan selama kuliah dan telah memberikan keceriaan, semangat serta kebersamaannya selama ini.
9. Diri sendiri yang telah bertahan hingga penyelesaian dalam proses penulisan skripsi ini.
10. Serta, kepada semua pihak yang telah membantu penulis dalam proses penulisan skripsi ini yang tidak dapat disebutkan namanya satu per satu.

Atas segala kekurangan dan kelemahan dalam penulisan skripsi ini, penulis mengharapkan saran dan kritik yang membangun. Semoga skripsi ini dapat menjadi bacaan yang bermanfaat serta dapat dikembangkan lebih lanjut oleh peneliti-peneliti selanjutnya.

DAFTAR ISI

HALAMAN JUDUL	i
PERNYATAAN KEASLIAN	ii
PENGESAHAN	iii
NOTA PEMBIMBING	iv
KATA PENGANTAR	viii
DAFTAR ISI	ix
DAFTAR TABEL	xi
DAFTAR GAMBAR	xii
DAFTAR LAMPIRAN	xiv
BAB I PENDAHULUAN	1
A. Latar Belakang Penelitian	1
B. Rumusan Masalah	5
C. Tujuan Penelitian	5
D. Manfaat Penelitian	6
E. Batasan Masalah	6
BAB II LANDASAN PUSTAKA	8
A. Landasan Teori	8
1. Data	8
2. <i>Data mining</i>	9
3. Cross-Industry Standard Process for Data mining	13
4. Klastering	16
5. Algoritma <i>K-Means</i>	22
6. Metode <i>Silhouette Coefficient</i>	25
7. Tendensi Pusat	29
8. <i>Interquartile Range</i>	30
9. Bisnis Penjualan dan Pemasaran	31
B. Kajian Pustaka	34
BAB III Metode Penelitian	38
A. Sumber Data	38
B. Populasi dan Sampel	39

C.	Atribut Penelitian	39
D.	Teknik Pengumpulan Data	42
E.	Studi Literatur	42
F.	Teknik Analisis Data	42
BAB IV Hasil Penelitian dan Pembahasan		50
A.	Pemahaman Bisnis (<i>Business Understanding</i>)	50
B.	Pemahaman Data (<i>Data Understanding</i>)	51
C.	Persiapan Data (<i>Data Preparation</i>)	68
D.	<i>Metode Silhoutte Coefficient</i>	78
E.	Pemodelan (<i>Modeling</i>)	87
F.	Evaluasi (<i>Evaluation</i>)	108
G.	Penyebaran (<i>Deployment</i>)	110
BAB V PENUTUP		113
A.	Kesimpulan	113
B.	Saran	114
DAFTAR PUSTAKA		115
Lampiran-lampiran		120

DAFTAR TABEL

Tabel	Judul	Halaman
Tabel 2.1	Interpretasi Nilai $s(o)$	27
Tabel 2.2	Interpretasi Nilai S	28
Tabel 3.1	Atribut Sumber data	40
Tabel 3.2	Atribut Perhitungan Penelitian	41
Tabel 4.1	Hasil Proses Identifikasi Data	53
Tabel 4.2	Hasil Proses Identifikasi Tipe Data	54
Tabel 4.3	Hasil Distribusi Jumlah Penjualan Produk dan Pendapatan berdasarkan tanggal	56
Tabel 4.4	Hasil Proses Distribusi Jumlah Penjualan Produk dan Pendapatan berdasarkan Status Dokumen	58
Tabel 4.5	Hasil Proses Distribusi Jumlah Penjualan Produk dan Pendapatan berdasarkan lokasi wilayah	59
Tabel 4.6	Hasil Proses Distribusi Jumlah Penjualan Produk dan Pendapatan berdasarkan lokasi wilayah	61
Tabel 4.7	Hasil Proses Deskripsi Data	68
Tabel 4.8	Hasil Proses Deskripsi Atribut Kolom	69

DAFTAR GAMBAR

Gambar	Judul	Halaman
Gambar 1.1	Tingkat Konsumsi AMDK di Kota Semarang (Sumber: BPS Provinsi Jawa Tengah)	2
Gambar 2.1	Siklus Fase <i>CRISP-DM</i> (Sumber: Taylor (2017))	13
Gambar 2.2	Proses Klastering	17
Gambar 3.1	<i>Flowchart</i> Model <i>CRISP-DM</i> (Sumber: Shearer (2000))	43
Gambar 3.2	Tahapan Modeling Algoritma <i>K-Means</i> (Sumber: Hans dkk (2012))	47
Gambar 4.1	Hasil Proses Identifikasi <i>Missing Value</i>	63
Gambar 4.2	Hasil Proses Identifikasi Data Duplikat	64
Gambar 4.3	Hasil Proses Identifikasi <i>Outliers</i>	65
Gambar 4.4	Data Proses	67
Gambar 4.5	Hasil Proses Reduksi Atribut Kolom	70
Gambar 4.6	Hasil Proses Reduksi Atribut Kolom Nama Produk	71
Gambar 4.7	Hasil Proses pemeriksaan dan penghapusan nilai yang hilang (<i>missing values</i>)	73
Gambar 4.8	Hasil Proses pemeriksaan dan penghapusan data duplikat	74

Gambar 4.9	Hasil Proses pemeriksaan dan penghapusan <i>Outliers</i>	75
Gambar 4.10	Identifikasi Data Setelah Pemeriksaan Kualitas Data	76
Gambar 4.11	Hasil Proses <i>Intergrasi Data</i>	77
Gambar 4.12	Hasil Proses Evaluasi Data Jumlah Klaster 2	79
Gambar 4.13	Hasil Program Nilai <i>Silhouette Coefficient</i> untuk Jumlah Klaster 2	85
Gambar 4.14	Hasil Program Nilai <i>Silhouette Coefficient</i> Untuk Beberapa Jumlah Klaster	86
Gambar 4.15	Visual <i>Silhouette Coefficient</i> Untuk Beberapa Jumlah Klaster	87
Gambar 4.16	Data Penelitian <i>Pemodelan</i>	88
Gambar 4.17	<i>centroid</i> Awal dalam Iterasi 1	89
Gambar 4.18	Hasil Proses Manual Iterasi Pertama	97
Gambar 4.19	Hasil Proses Komputasi Iterasi Pertama	97
Gambar 4.20	Hasil Proses Manual Perhitungan <i>centroid</i> Baru	101
Gambar 4.21	Hasil Proses Komputasi Perhitungan <i>centroid</i> Baru	101
Gambar 4.22	Hasil Proses Komputasi Iterasi Kedua	103
Gambar 4.23	Hasil Proses Komputasi Iterasi Ketiga	104
Gambar 4.24	Hasil Proses Komputasi Iterasi Keempat	105
Gambar 4.25	Visual Algoritma <i>K-Means</i>	106
Gambar 4.26	Hasil Proses Terminasi	107
Gambar 4.27	Hasil Proses Evaluasi <i>Threshold Mean dan Standar Deviasi</i>	109
Gambar 4.28	Peta Hasil klaster Penjualan Produk AMDK Kota Semarang	111

DAFTAR LAMPIRAN

	Halaman
Lampiran 1	Sumber Data 120
Lampiran 2	Surat Perizinan Penelitian 121
Lampiran 3	<i>Script RStudio 2024.09.0+375 Data Understanding</i> 122
Lampiran 4	<i>Script RStudio 2024.09.0+375 Data Preparation</i> 129
Lampiran 5	<i>Script MATLAB 2013 Silhouette Coefficient</i> 133
Lampiran 6	<i>Script MATLAB 2013 Algoritma K-Means</i> 139
Lampiran 7	Daftar Riwayat Hidup 145

BAB I

PENDAHULUAN

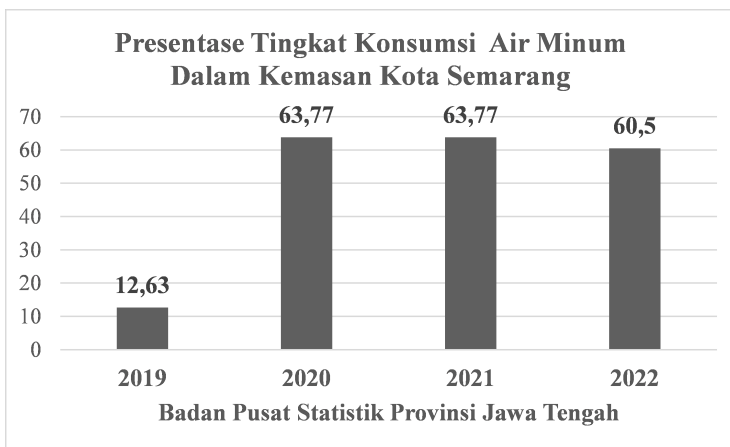
A. Latar Belakang Penelitian

Sumber daya air adalah sumber daya alam yang penting bagi kehidupan flora, fauna, dan manusia. Ketersediannya terbatas dan penyebarannya tidak merata, baik secara tempat maupun waktu. Air juga memiliki nilai ekonomis, terutama dalam industri seperti makanan dan minuman. Oleh karena itu, reformasi pengelolaan sumber daya air diperlukan untuk mengatasi kemiskinan, meningkatkan ketahanan pangan dan energi, serta melestarikan alam. Beberapa kebijakan, seperti Undang-Undang Nomor 7 Tahun 2004 tentang Sumber Daya Air (UU SDA) dan Peraturan Presiden tentang Dewan Sumber Daya Air telah disusun untuk mendukung prinsip Pengelolaan Sumber Daya Air Terpadu (IWRM). Pengelolaan ini mencakup perencanaan, pelaksanaan, pemantauan, dan evaluasi dengan melibatkan masyarakat dan dunia usaha (Purwanto dan Susanto, 2015).

Menurut Supriadi & Dewi (2018), pengelolaan sumber daya air dapat diberikan kepada pihak non-pemerintah berdasarkan Peraturan Pemerintah Nomor 121 Tahun 2015. Delegasi ini memerlukan persetujuan dari pemerintah pusat dan daerah serta harus sesuai dengan ketentuan konstitusi. Pengelolaan tersebut tidak boleh mengganggu kebutuhan air masyarakat, khususnya masyarakat yang tinggal di dekat sumber air, sebagaimana diatur dalam Pasal 2 Peraturan Pemerintah Nomor 121 Tahun 2015. Pengelolaan ini strategis karena berkaitan dengan aktivitas bisnis

yang memerlukan air sebagai bahan pokok, seperti produksi air minum dalam kemasan (AMDK).

Badan Pusat Statistik Provinsi Jawa Tengah merilis data tingkat konsumsi AMDK Kota Semarang dari tahun 2019 hingga 2022 (BPS, 2022). Di bawah ini adalah diagram batang yang menunjukkan perbandingan tingkat konsumsi AMDK selama periode tersebut.



Gambar 1.1. Tingkat Konsumsi AMDK di Kota Semarang
(Sumber: BPS Provinsi Jawa Tengah)

Perbandingan tingkat konsumsi Air Minum Dalam Kemasan (AMDK) di Kota Semarang pada Gambar 1.1 menunjukkan peningkatan signifikan sebesar 63,77% dari tahun 2019 hingga 2021 akibat pandemi. Namun, tingkat konsumsi menurun menjadi 60,50% pada tahun 2022. Penurunan ini menimbulkan kekhawatiran di kalangan perusahaan produksi dan penjualan AMDK terkait dampaknya terhadap pendapatan. Perusahaan Wahyu Jaya cabang Semarang berdiri pada tahun 1997 bergerak di bidang distribusi AMDK bekerja sama dengan mitra seperti

Danone AQUA, Danone VIT, dan perusahaan lainnya. Perusahaan ini menghadapi persaingan pasar dan tingkat penurunan konsumsi AMDK masyarakat sehingga memerlukan strategi untuk menganalisis pola penjualan guna mengidentifikasi produk-produk berpotensi penjualan tertinggi dan terendah di Kecamatan Kota Semarang. Oleh karena itu, salah satu pendekatan untuk mengatasi masalah adalah *data mining*.

Data mining adalah proses analisis mendalam yang bertujuan mengidentifikasi informasi tersembunyi dalam data kompleks. Proses ini mencakup ekstraksi, evaluasi, pencarian model, dan penilaian data yang dikumpulkan untuk mengungkap pola-pola penting. Sasaran utama dari *data mining* adalah mengubah data yang belum diproses menjadi pemahaman praktis untuk mendukung pengambilan keputusan. Istilah proses menunjukkan pentingnya pendekatan terencana dalam penggunaan alat komputer.

Dalam konteks ini, pentingnya pendekatan terencana ditunjukkan melalui model standar yang digunakan dalam industri adalah *Cross-Industry Standard Process for Data Mining (CRISP-DM)*. Model ini menyediakan kerangka kerja terstruktur untuk mengelola siklus hidup proyek *data mining*, mulai dari pemahaman bisnis hingga penerapan dan evaluasi model. Dengan demikian, *CRISP-DM* menjadi panduan yang efektif untuk mengimplementasikan *data mining* secara efisien dan terencana karena fleksibilitas dan struktur sistematis. Penerapan pemodelan dalam *CRISP-DM* menggunakan klastering *data mining*.

Klastering adalah teknik dalam *data mining* yang mengelompokkan data berdasarkan kesamaan atau perbedaan dalam karakteristiknya. Metode ini termasuk dalam pembelajaran

tak terawasi (*unsupervised learning*) karena data dikelompokkan tanpa memerlukan target atau label tertentu. Tujuannya adalah menemukan pola dalam data yang belum diberi nama yang sangat berguna dalam menyelesaikan masalah kompleks yang sulit dipecahkan dengan algoritma lain (Muflikhah et al, 2018). Klastering diterapkan dalam *CRISP-DM* sebagai tahap pemodelan menggunakan algoritma *K-Means*. Algoritma ini dipilih karena mudah dipahami, relatif cepat proses klasteringnya, dan dapat mudah dikombinasikan dengan metode lain seperti *Silhouette Coefficient*. Metode *Silhouette Coefficient* digunakan untuk menilai jumlah klaster optimal sebelum pemodelan serta kualitas klaster setelah tahap pemodelan (Iriadi dan Priatno, 2020).

Model dan metode yang digunakan dalam penelitian ini merujuk kepada studi yang dilakukan oleh Núñez Arias et al (2021) berjudul *CRISP-DM and K-Means Neutrosophic in the Analysis of Risk Factors for Hearing Loss in Children*. Penelitian tersebut menunjukkan bahwa algoritma *K-Means* dan model *CRISP-DM* dapat diterapkan dalam bidang kesehatan. Dengan dasar tersebut, penelitian ini akan difokuskan pada penerapan di bidang bisnis distribusi penjualan dengan penambahan metode *Silhouette Coefficient* untuk meningkatkan akurasi pengelompokan. Selanjutnya, penelitian di bidang lingkungan yang dilakukan oleh Dhewayani et al (2022) berjudul *Implementasi K-Means Clustering untuk Pengelompokan Daerah Rawan Bencana Kebakaran Menggunakan Model CRISP-DM*. Penelitian ini juga menunjukkan bahwa kombinasi algoritma *K-Means*, metode *Silhouette Coefficient*, dan model *CRISP-DM* dapat efektif dalam mengidentifikasi daerah rawan kebakaran sehingga memungkinkan tindakan pencegahan yang tepat. Mengikuti jejak

penelitian tersebut, fokus penelitian ini akan diarahkan ke sektor bisnis distribusi penjualan produk AMDK di wilayah Semarang. Penelitian selanjutnya oleh Marsono (2019) yang berjudul *Analisis Data Mining Pada Strategi Penjualan Produk PT Aquasolve Sanaria Dengan Menggunakan Metode K-Means Clustering*. Penelitian ini juga membuktikan efektivitas algoritma *K-Means* dalam konteks bisnis.

Berdasarkan temuan dari penelitian-penelitian sebelumnya, penulis berencana melaksanakan studi penelitian dengan menggunakan algoritma *K-Means* yang dilengkapi dengan metode *Silhouette Coefficient* serta model *CRISP-DM* dengan fokus pada sektor bisnis distribusi penjualan menggunakan perangkat lunak seperti *Microsoft Excel 2013*, *RStudio 2024.09.0+375*, dan *MATLAB 2013*. Oleh karena itu, penulis tertarik untuk melakukan penelitian berjudul Analisis Pola Penjualan menggunakan Model *CRISP-DM* dengan Algoritma *K-Means* (Studi Kasus: Wahyu Jaya Group Cabang Semarang Tahun 2023).

B. Rumusan Masalah

Berdasarkan latar belakang yang telah dibahas, maka rumusan masalah yang diperoleh yaitu Bagaimana penerapan model *CRISP-DM* dengan algoritma *K-Means* dapat menganalisis pola penjualan di Wahyu Jaya Group Cabang Semarang pada November 2023 serta implikasinya terhadap strategi bisnis?

C. Tujuan Penelitian

Berdasarkan rumusan masalah penelitian ini dapat diperoleh tujuan masalah dalam penelitian ini yaitu Untuk menerapkan

model *CRISP-DM* sehingga dapat mengimplementasikan setiap tahap dalam model *CRISP-DM* untuk memahami proses analisis data penjualan dengan menggunakan algoritma *K-Means* dengan metode *Silhouette Coefficient* sehingga dapat mengelompokkan data penjualan dan mengidentifikasi pola yang relevan.

D. Manfaat Penelitian

Berdasarkan rumusan masalah dan tujuan penelitian ini dapat diperoleh manfaat masalah dalam penelitian ini, yaitu:

1. Penelitian ini diharapkan dapat menambah pengetahuan tentang penerapan *data mining*, *CRISP-DM*, dan algoritma *K-Means*.
2. Penelitian ini dapat dijadikan sebagai referensi untuk penelitian selanjutnya.
3. Penelitian ini akan membantu perusahaan Wahyu Jaya cabang Semarang untuk tetap relevan dan berkelanjutan dalam bisnis AMDK.

E. Batasan Masalah

Adapun batasan masalah yang dalam penelitian ini sebagai berikut, yaitu:

1. Penelitian ini terbatas pada penggunaan model *CRISP-DM* untuk analisis data penjualan dengan fokus akan diberikan pada tahapan-tahapan *CRISP-DM* dari pemahaman bisnis hingga penerapan hasil.

2. Teknik metode *data mining* yang digunakan adalah klastering pada algoritma *K-Means* dengan metode *Silhouette Coefficient* .
3. Data yang digunakan adalah data transaksi penjualan produk AMDK jenis botol dan gelas yang hanya pada Seluruh Kecamatan Kota Semarang bulan November tahun 2023.
4. Analisis akan terbatas pada produk tertentu dari perusahaan Wahyu Jaya cabang Semarang, yaitu: AQUA 1500ML, AQUA 220ML, AQUA 330ML, AQUA 600ML, MIZONE 500ML, VIT 200ML, VIT 330ML, VIT 550ML, VIT 1500ML.

BAB II

LANDASAN PUSTAKA

A. Landasan Teori

1. Data

Menurut Muflikhah et al (2018), data merupakan sekumpulan informasi yang mencakup berbagai objek beserta atribut atau karakteristik yang melekat pada masing-masing objek. Atribut ini dikenal juga sebagai variabel yang merujuk pada sifat yang dimiliki oleh setiap objek. Persiapan pemrosesan data sangat memerlukan pengecekan terhadap kualitas data yang mencakup beberapa aspek penting seperti *outliers*, *missing value*, dan data duplikat. Berikut penjelasan masing-masing aspek, yaitu:

1.1. *Outliers*

Outliers adalah data yang memiliki nilai yang sangat berbeda dari data lainnya dan dapat memengaruhi hasil analisis serta model prediktif. Identifikasi *outlier* penting untuk memastikan bahwa analisis data tetap representatif dan bebas bias.

1.2. *Missing value*

Missing value merujuk pada data yang tidak tersedia dalam *dataset* yang biasanya disebabkan oleh kesalahan pengukuran atau masalah teknis. Penanganan *missing value* sangat penting untuk mencegah analisis yang bias serta memastikan hasil yang akurat.

1.3. Data duplikat

Data duplikat yang muncul dalam *dataset* dapat merusak hasil analisis sehingga penting untuk menghapusnya guna menjaga integritas analisis data.

Pengecekan dan penanganan ketiga aspek tersebut merupakan langkah awal yang penting dalam pemrosesan data untuk memastikan bahwa data yang digunakan berkualitas dan siap dianalisis lebih lanjut (Muflikhah et al, 2018).

2. Data mining

Data mining adalah proses analisis mendalam yang melibatkan pemecahan data kompleks untuk mengidentifikasi informasi tersembunyi dan berharga yang sebelumnya tidak terdeteksi. Proses ini melampaui sekadar ekstraksi dan evaluasi data melalui sistem otomatis atau semi otomatis dengan tujuan mengungkap pola-pola penting. Selain itu, *data mining* juga mencakup pencarian model, abstraksi, dan penilaian terhadap data yang telah dikumpulkan. Sasaran utama dari *data mining* adalah mentransformasikan data yang belum diproses menjadi pemahaman praktis, mendukung pengambilan keputusan yang mendalam, serta memberikan kemampuan kepada perusahaan atau organisasi untuk beradaptasi dengan cepat terhadap dinamika data dan perubahan kondisi ekonomi (Muflikhah et al, 2018).

Istilah proses dalam konteks *data mining* menekankan pentingnya pendekatan terencana, bukan sekadar pemanfaatan alat komputer untuk otomatisasi. *Cross-Industry Standard Process for Data Mining (CRISP-DM)* adalah model standar *data mining*

yang menyediakan kemudahan dalam implementasi secara efektif, efisien, dan terstruktur, baik untuk skala besar maupun kecil. Model ini menawarkan kerangka kerja yang jelas untuk mengatur siklus hidup proyek *data mining*, mulai dari pemahaman bisnis hingga penerapan dan evaluasi model (Iriadi dan Priatno, 2020).

Peranan *data mining* sangat penting dalam analisis data. Menurut Larose (2014), peranan *data mining* dibagi menjadi beberapa aspek, yaitu:

2.1. Deskripsi

Identifikasi pola dan perkembangan dalam data sangat penting. Sebagai contoh, analisis yang dilakukan oleh seorang pengamat pemilu menunjukkan bahwa individu yang baru kehilangan pekerjaan cenderung tidak mendukung calon presiden yang sedang menjabat. Fenomena ini dapat dijelaskan melalui analisis data eksploratif yang memanfaatkan grafik untuk mengeksplorasi dan mengenali pola serta tren dalam data.

2.2. Estimasi

Dalam konteks *data mining*, estimasi merujuk pada proses memprediksi nilai numerik dari atribut target berdasarkan atribut prediktor. Atribut target adalah hasil yang diprediksi seperti tekanan darah, sementara atribut prediktor adalah faktor-faktor yang mendasari proses prediksi tersebut. Model yang dibangun berdasarkan data yang lengkap digunakan untuk memperkirakan nilai target pada dataset baru. Metode estimasi meliputi jaringan saraf tiruan dan teknik statistik, termasuk regresi linier. Proses estimasi berbeda dengan klasifikasi, karena menghasilkan nilai

numerik alih-alih kategori.

2.3. Prediksi

Prediksi dalam *data mining* berfokus pada hasil masa depan, mirip dengan klasifikasi dan estimasi. Contoh-contohnya meliputi meramalkan nilai saham, menghitung peningkatan kecelakaan akibat kenaikan batas kecepatan, analisis statistik untuk menentukan pemenang *World Series*, dan proyeksi keberhasilan molekul dalam riset farmasi. Selain metode statistik seperti *point estimation* dan *simple linear regression*, metode *k-nearest neighbor*, *decision trees*, dan *neural networks* juga digunakan dalam *data mining* dan *knowledge discovery*.

2.4. Klasifikasi

Klasifikasi dalam *data mining* terdiri dari empat komponen utama yaitu atribut target, atribut prediktor, *dataset* pelatihan, dan *dataset* pengujian. Atribut target berfungsi sebagai label untuk objek yang diklasifikasikan, sedangkan atribut prediktor adalah atribut independen yang digunakan dalam proses ini. *dataset* pelatihan mengandung contoh berlabel yang digunakan untuk melatih model dengan tujuan utama dalam pembelajaran terawasi adalah memprediksi kelas objek baru. Klasifikasi menetapkan atribut target berdasarkan atribut yang ada dan mengukur akurasi model menggunakan *dataset* pengujian. Faktor penting dalam pemilihan model klasifikasi meliputi akurasi, kecepatan, dan interpretabilitas dengan contoh model seperti *Naive Bayes*, *Decision Tree*, *Neural Network*, *Genetic Algorithm*, dan *K-nearest Neighbor*.

2.5. Klastering

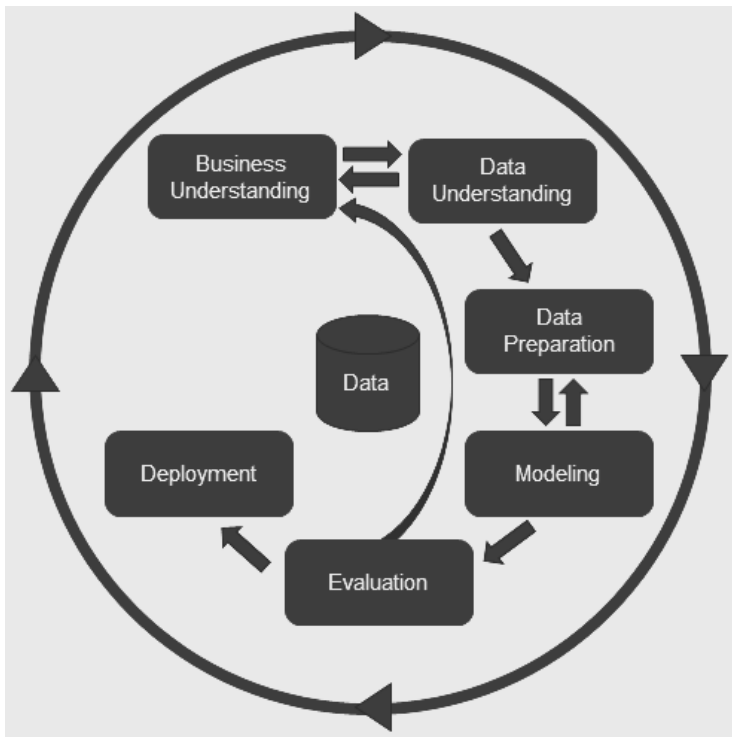
Klastering dalam *data mining* merupakan metode pembelajaran tanpa pengawasan (*unsupervised learning*) yang mengelompokkan data menjadi kelompok homogen berdasarkan kesamaan karakteristik. Berbeda dengan klasifikasi, klastering tidak memerlukan atribut target yang diketahui, melainkan mengidentifikasi kelompok berdasarkan struktur dalam data. Proses ini mengelompokkan data yang serupa dari *dataset* yang lebih besar berdasarkan atribut yang relevan. Klastering bermanfaat untuk mengidentifikasi pola tersembunyi dan dapat diterapkan dalam berbagai bidang seperti segmentasi pasar, deteksi anomali, dan analisis data multidimensi. Terdapat empat pendekatan utama dalam klastering: *partition clustering*, *grid-based clustering*, *hierarchical clustering*, dan *density-based clustering* (Jollyta et al, 2020).

2.6. Asosiasi

Asosiasi dalam *data mining* adalah proses analitis yang bertujuan untuk mengidentifikasi hubungan signifikan antara atribut dalam kumpulan data besar. Metode ini sering digunakan dalam analisis afinitas atau *market basket analysis* untuk mendeteksi pola pembelian item. Aturan asosiasi yang dihasilkan mengikuti format "Jika ada *antecedent* (sebab), maka ada *consequent* (akibat)" dengan nilai dukungan dan kepercayaan yang menunjukkan frekuensi kombinasi item serta probabilitas kemunculan item *consequent* saat item *antecedent* dibeli.

3. Cross-Industry Standard Process for Data mining

Cross-Industry Standard Process for Data Mining (CRISP-DM) adalah sebuah model standar yang digunakan dalam *data mining* di bidang industri atau bisnis untuk mengelola proyek *data mining*. Model ini dikembangkan oleh tiga perusahaan pionir dalam bidang *data mining*, yaitu *Daimler Chrysler (Daimler-Benz)*, *SPSS (ISL)*, dan *NCR*. *CRISP-DM* menggambarkan siklus hidup proyek *data mining* yang terdiri dari enam tahapan (Iriadi dan Priatno, 2020). Menurut Larose (2014) terdapat enam tahapan yang dijelaskan di bawah ini, yaitu:



Gambar 2.1. Siklus Fase *CRISP-DM* (Sumber: Taylor (2017))

Siklus fase *CRISP-DM* yang ditampilkan pada Gambar 2.1 terdiri dari enam fase dalam proses *data mining* yang akan dijelaskan secara mendetail di bawah ini, yaitu:

3.1. **Pemahaman Bisnis (Fase *Business Understanding*)**

Business Understanding adalah tahap awal yang berfokus pada pemahaman tujuan serta kebutuhan bisnis dari penelitian atau proyek. Pada tahap ini, masalah bisnis yang ingin dipecahkan diidentifikasi, dan tujuan analisis ditetapkan secara jelas. Selain itu, batasan yang mungkin dihadapi selama penelitian juga dipertimbangkan. Hasil akhir dari tahap ini adalah pernyataan yang jelas mengenai tujuan penelitian yang ingin dicapai, serta indikator yang akan digunakan untuk mengukur keberhasilannya.

3.2. **Pemahaman Data (Fase *Data Understanding*)**

Data Understanding dimulai dengan proses pengumpulan data yang relevan untuk penelitian atau proyek ini. Pada tahap ini, data dieksplorasi secara mendalam untuk memahami karakteristik awal dan mengevaluasi kualitasnya, termasuk mengidentifikasi adanya *missing value*, *Outliers*, dan data duplikat. Pemahaman yang diperoleh dari tahap ini sangat penting untuk memastikan bahwa data yang digunakan benar-benar representatif dalam menyelesaikan masalah bisnis. Hasil dari tahap ini adalah gambaran awal tentang data, termasuk masalah yang perlu ditangani sebelum melanjutkan ke tahap berikutnya.

3.3. Persiapan Data (*Data Preparation*)

Data Preparation adalah fase krusial dalam proses *CRISP-DM* yang berfokus pada menyiapkan data agar siap diolah pada tahap pemodelan. Tahap ini mencakup langkah-langkah seperti membersihkan data dari kesalahan atau nilai *error* melakukan transformasi pada variabel atau atribut dan mengurangi dimensi data jika diperlukan, dan menggabungkan data dari berbagai sumber jika ada dua sumber data atau lebih. Tujuan dari tahap ini adalah untuk menghasilkan *dataset* final yang bersih dan siap digunakan dalam proses pemodelan. Kualitas data yang dihasilkan pada tahap ini sangat menentukan kualitas model yang akan dibangun.

3.4. Pemodelan (*Modeling*)

Modeling adalah tahap di mana teknik pemodelan diterapkan untuk menemukan pola atau hubungan dalam data. Pada tahap ini, berbagai metode *data mining* seperti *decision tree*, *regresion*, dan *clustering* dapat digunakan. Pengaturan parameter model dan pengujian beberapa pendekatan sering dilakukan untuk mendapatkan model yang optimal. Hasil dari tahap ini adalah model yang telah dilatih menggunakan data dan siap untuk dievaluasi lebih lanjut.

3.5. Evaluasi (*Evaluation*)

Evaluation bertujuan untuk memastikan bahwa model yang telah dibangun relevan dengan tujuan bisnis yang telah ditetapkan. Pada tahap ini, kinerja model diukur menggunakan metrik yang sesuai, seperti akurasi, *precision*, *error rate* tergantung pada jenis model yang digunakan.

Selain itu, hasil dari model juga harus diperiksa untuk memastikan kesesuaiannya dengan kebutuhan dan tujuan bisnis. Jika model dinilai memadai, tahap ini diakhiri dengan rekomendasi penerapan model dalam sistem bisnis.

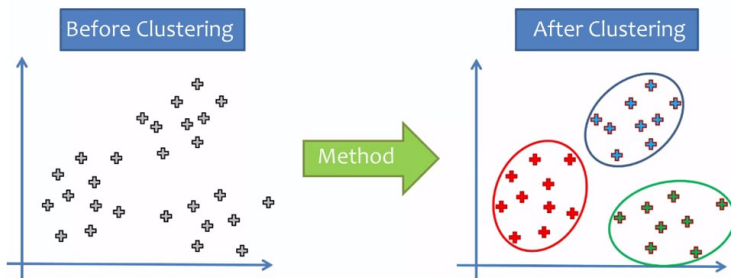
3.6. **Penyebaran (*Deployment*)**

Deployment adalah tahap akhir di mana model atau wawasan yang telah dihasilkan diimplementasikan dalam sistem bisnis. Pada tahap ini, hasil analisis diintegrasikan ke dalam proses bisnis yang ada, dan pengguna dilatih untuk memanfaatkan model secara efektif. Penerapan dapat berupa laporan, dashboard, atau sistem otomatis yang menerapkan model secara langsung dalam pengambilan keputusan bisnis. Hasil akhirnya adalah sistem yang dapat dioperasikan oleh bisnis untuk mencapai hasil yang diinginkan.

4. **Klustering**

Klustering dalam *data mining* adalah metode pembelajaran tanpa pengawasan (*unsupervised learning*) yang mengelompokkan data ke dalam kelompok homogen berdasarkan karakteristik serupa. Berbeda dengan klasifikasi, klustering tidak memerlukan atribut target yang sudah diketahui. Metode ini mengidentifikasi kelompok berdasarkan struktur data yang ada dengan mengumpulkan data yang memiliki kesamaan dari *dataset* yang lebih besar. Klustering berguna untuk menemukan pola tersembunyi dan dapat diterapkan dalam segmentasi pasar, deteksi anomali, dan analisis data multidimensi (Jollyta et al, 2020).

Dalam pembelajaran tanpa pengawasan, data dikelompokkan menjadi kluster berdasarkan kesamaan antar objek. Objek yang jaraknya dekat dikelompokkan dalam kluster yang sama, dengan tujuan membuat objek dalam satu kluster sehomogen mungkin, sementara kluster yang berbeda sejauh mungkin satu sama lain. Pendekatan ini memungkinkan identifikasi pola dan struktur yang tersembunyi dalam data tanpa memerlukan atribut target yang ditentukan sebelumnya, seperti yang ditunjukkan dalam gambar di bawah ini.



Gambar 2.2. Proses Klastering

Gambar 2.2 menunjukkan bahwa sebelum proses klastering, data tersebar ke segala arah tanpa memperlihatkan pola karakteristik yang serupa. Setelah melalui proses klastering, setiap kluster dibentuk berdasarkan kemiripan karakteristik data. Menurut Muflikhah et al (2018), ukuran yang digunakan dalam klastering mencakup kemiripan dan ketidakmiripan. Ukuran-ukuran ini sangat penting dalam mengelompokkan objek ke dalam kluster, karena menentukan sejauh mana hubungan antar objek berdasarkan kedekatan atau perbedaan karakteristiknya. Berikut adalah beberapa ukuran yang digunakan, yaitu:

4.1. Ukuran Kemiripan

Ukuran kesamaan digunakan untuk mengukur sejauh mana dua atau lebih objek memiliki kesamaan dalam hal atribut atau fitur yang digunakan dalam klastering. Tujuan dari ukuran ini adalah untuk menemukan objek yang mirip satu sama lain sehingga dapat dikelompokkan dalam klaster yang sama. Beberapa metode umum yang digunakan untuk menghitung kesamaan antara objek adalah sebagai berikut:

- 4.1.1. **Cosinus Similarity** adalah Ukuran ini mengukur sudut antara dua vektor fitur dalam ruang dimensi tinggi. Nilai cosinus dari sudut tersebut menunjukkan tingkat kemiripan antara dua objek. Nilai 1 menandakan bahwa kedua vektor identik (kemiripan maksimum), sementara nilai 0 menunjukkan bahwa kedua vektor tidak memiliki kemiripan sama sekali.
- 4.1.2. **Kovarian** adalah Ukuran ini mengukur sejauh mana dua variabel berubah secara bersamaan. Jika kedua variabel meningkat bersamaan, maka kovarian bernilai positif, menunjukkan adanya kemiripan. Sebaliknya, kovarian negatif menunjukkan pola perubahan yang berlawanan, sedangkan nilai nol menunjukkan tidak adanya hubungan antara kedua variabel.
- 4.1.3. **Korelasi** adalah Ukuran yang lebih distandarisasi dibandingkan kovarian, korelasi mengukur hubungan linear antara dua variabel. Nilai korelasi berkisar antara -1 hingga 1. Nilai mendekati 1 menunjukkan kemiripan yang tinggi, sementara nilai mendekati -1 menunjukkan ketidakmiripan yang signifikan.

4.2. Ukuran Ketidakmiripan

Ukuran ketidakmiripan atau jarak digunakan untuk mengukur sejauh mana dua objek berbeda. Semakin besar nilai jarak antara dua objek, semakin tidak mirip kedua objek tersebut. Beberapa metode yang umum digunakan untuk mengukur jarak dalam klastering antara lain, yaitu:

4.2.1. **Jarak Euclidean** (*Euclidean Distance*) adalah salah satu ukuran jarak yang paling umum digunakan. Ukuran ini menghitung panjang garis lurus antara dua obyek dalam ruang n dimensi, sehingga sangat efektif untuk data numerik.

$$dist(p, q) = \sqrt{\sum_{i=1}^n (p_i - q_i)^2}$$

dimana p dan q adalah dua titik dalam ruang n dimensi, dan p_i serta q_i adalah koordinat dari titik tersebut.

4.2.2. **Jarak Manhattan** (*Manhattan Distance*), yang juga dikenal sebagai jarak *city block* adalah Ukuran menghitung jarak antara dua obyek dengan mengikuti sumbu vertikal dan horizontal. Ukuran ini cocok digunakan pada data diskrit atau kategorikal.

$$dist(p, q) = \sum_{i=1}^n |p_i - q_i|$$

dimana p dan q adalah dua titik dalam ruang n dimensi, dan p_i serta q_i adalah koordinat dari titik tersebut.

4.2.3. **Jarak Minkowski** adalah generalisasi dari jarak

Euclidean dan *Manhattan* yang memungkinkan pengukuran jarak dengan berbagai derajat kebebasan.

$$dist(p, q) = (\sum_{i=1}^n (p_i - q_i)^{1/m})$$

dimana p dan q adalah dua titik dalam ruang n dimensi, dan p_i serta q_i adalah koordinat dari titik tersebut dengan m adalah derajat kebebasan.

4.3. Metode Klastering

Metode klastering merupakan alat penting dalam analisis data yang digunakan untuk mengelompokkan objek berdasarkan kesamaan atau karakteristik tertentu. Berikut ini adalah beberapa metode klastering dasar yang umum digunakan, menurut Hans dkk (2012), yaitu:

4.3.1. *Partitioning Methods*

Metode partisi adalah teknik klastering yang membagi data ke dalam beberapa kelompok secara terpisah. Proses ini menghasilkan jumlah klaster, di mana nilai jumlah klaster tidak melebihi jumlah total data dan setiap klaster harus berisi minimal satu data. Metode ini menggunakan perhitungan jarak untuk menentukan pembagian klaster dan terus diperbaiki secara bertahap. Contoh algoritma yang sering digunakan dalam metode partisi adalah *K-Means* dan *K-Medoids*, yang bekerja secara efektif pada data berukuran kecil hingga menengah.

4.3.2. *Hierarchical Methods*

Metode *Hierarchical* adalah teknik klastering yang menyusun data dalam struktur pohon. Terdapat dua pendekatan utama yaitu **Agglomeratif** yang menggabungkan objek secara bertahap menjadi klaster lebih besar dan **Divisif** yang memecah satu klaster besar menjadi klaster lebih kecil. Kedua pendekatan ini menggunakan ukuran jarak atau kepadatan untuk membentuk klaster, membantu dalam memahami struktur hierarkis data.

4.3.3. ***Density-based Methods***

Metode berbasis kepadatan adalah pendekatan klastering yang fokus pada kepadatan data. Metode ini memungkinkan pembentukan klaster yang kompleks dan penanganan kebisingan dengan mempertimbangkan kepadatan di sekitar objek. Dengan membagi data menjadi klaster eksklusif, metode ini dapat mengungkap struktur hierarkis dalam subruang data, serta efektif dalam mengidentifikasi pola kompleks dan mengatasi ketidakpastian. Oleh karena itu, metode ini merupakan alat penting dalam analisis data dan pengelompokan berdasarkan distribusi kepadatan.

4.3.4. ***Grid-based Methods***

Metode berbasis *grid* membagi ruang data menjadi sel-sel kecil yang membentuk struktur *grid*, di mana proses klastering dilakukan. Pendekatan ini sangat efisien dalam menangani data spasial karena waktu pemrosesan bergantung pada jumlah sel, bukan jumlah objek. Selain itu, metode ini dapat dikombinasikan

dengan teknik klastering lain, seperti metode berbasis kepadatan atau hierarkis, untuk memperdalam analisis dan menemukan pola yang lebih kompleks dalam *dataset*.

5. Algoritma *K-Means*

Algoritma *K-Means* adalah algoritma klastering dalam metode *Partitioning Methods* yang membagi objek menjadi sejumlah klaster berdasarkan karakteristik tertentu. Algoritma ini mempartisi data sedemikian rupa sehingga objek yang memiliki kemiripan ditempatkan dalam klaster yang sama, sementara objek yang berbeda karakteristiknya dikelompokkan ke klaster lain. Secara historis, bentuk inti dari *K-Means* ditemukan oleh berbagai peneliti dari disiplin ilmu yang berbeda. Beberapa penelitian paling berpengaruh dilakukan oleh Lloyd (1982), Forgey (1965), Friedman dan Rubin (1967), serta McQueen (1967) (Jollyta et al, 2020). Berikut ini adalah langkah-langkah dasar untuk algoritma *K-Means* menurut Hans dkk (2012), yaitu:

5.1. Inisialisasi

Langkah pertama dalam algoritma *K-Means* adalah menentukan jumlah klaster k yang diinginkan. Kemudian, pilih k objek o dari *dataset* D sebagai *centroid* awal. Pemilihan ini dapat dilakukan dengan menentukan jumlah klaster k secara acak atau menggunakan metode lain, seperti *Silhouette Coefficient*, untuk memastikan pemilihan awal yang lebih optimal. Metode ini membantu meningkatkan kualitas klastering dengan memilih *centroid* yang lebih representatif sejak awal. Dengan simbol k adalah jumlah

klaster yang akan dibentuk. o adalah objek data. D adalah *dataset* atau kumpulan objek data yang akan dikelompokkan.

5.2. Penugasan

Untuk setiap objek o dalam *dataset* D , hitung jarak antara objek o dan setiap *centroid* dalam klaster k menggunakan jarak *Euclidean*. Jarak ini diperoleh dengan menghitung jarak *Euclidean* antara objek o dan *centroid* C_j dari setiap klaster ke- j yang menggunakan rumus, yaitu:

$$\text{dist}(o, C_j) = \sqrt{\sum_{i=1}^d (o_i - C_{ji})^2} \quad (2.1)$$

Keterangan:

- C_j adalah *centroid* dalam setiap klaster j dari k .
- $\text{dist}(o, C_j)$ adalah jarak *Euclidean* antara objek o dan *centroid* C_j .
- d adalah jumlah atribut dalam ruang data.
- o_i adalah nilai dari atribut ke- i dari objek o .
- C_{ji} adalah nilai dari atribut ke- i dari *centroid* C_j .

Setelah jarak *Euclidean* dihitung, Kemudian objek (o) dikelompokkan ke klaster C_j dengan *centroid* terdekat dengan rumus, yaitu:

$$C_i = \text{argmin}_j(\text{dist}(o, C_j)) \quad (2.2)$$

Keterangan:

- $dist(o, C_j)$ adalah jarak *Euclidean* antara objek o dan *centroid* C_j .
- $argmin_j$ adalah mencari nilai j yang menghasilkan nilai minimum $dist(o, C_j)$.
- C_i adalah kluster yang dipilih untuk objek o berdasarkan *centroid* terdekat.

5.3. Pembaruan *centroid*

Setelah setiap objek o memiliki kluster C_i , langkah berikutnya adalah menghitung ulang *centroid*. *centroid* baru c_j dihitung sebagai rata-rata dari seluruh objek o yang termasuk dalam kluster C_i . Rumus untuk menghitung baru adalah sebagai berikut:

$$c_j = \frac{1}{|C_i|} \sum_{o \in C_i} o \quad (2.3)$$

Keterangan:

- c_j adalah *centroid* baru untuk kluster C_i dihitung sebagai rata-rata dari semua objek o yang termasuk dalam kluster tersebut.
- $|C_i|$ adalah jumlah objek dalam kluster C_i .
- o adalah objek data yang berada dalam kluster C_i .

5.4. Iterasi

Langkah iterasi dalam algoritma *K-Means* melibatkan pengulangan langkah penugasan dan pembaruan *centroid* baru sehingga tidak ada perubahan dalam penugasan kluster.

Hal ini terjadi ketika objek-objek tidak lagi berpindah antar kluster. Kondisi ini dikenal sebagai konvergensi yang menjadi titik akhir dari algoritma *K-Means*.

5.5. Terminasi

Langkah terminasi dalam algoritma *K-Means* terjadi ketika *centroid* dari setiap kluster tidak lagi berubah atau ketika penugasan objek ke kluster telah stabil. Kondisi ini menghasilkan pembagian objek ke dalam kluster yang meminimalkan variasi di dalam setiap kluster.

6. Metode *Silhouette Coefficient*

Metode *Silhouette Coefficient* digunakan untuk menilai kualitas pengelompokan data dalam kluster. Metode ini mengukur seberapa dekat data dengan klasternya sendiri serta seberapa jauh data tersebut dari kluster lain (Davis, 2007). Langkah-langkah perhitungan *Silhouette Coefficient* dapat merujuk pada referensi Dewi dan Pramita (2019) sebagai berikut, yaitu:

6.1. Menghitung Nilai Rata-rata Jarak Intrakluster (a)

Untuk setiap objek o dalam kluster C_i , hitung rata-rata jarak antara objek tersebut dengan semua objek lain dalam kluster yang sama. Rata-rata jarak intrakluster ini dinotasikan sebagai $a(o)$. Rumus untuk menghitung a adalah sebagai berikut:

$$a(o) = \frac{1}{|C_i| - 1} \sum_{o' \in C_i, o' \neq o} dist(o, o') \quad (2.4)$$

Keterangan:

1. $a(o)$ adalah rata-rata jarak antara objek o dan objek lain o' di dalam klaster sama C_i .
2. $|C_i|$ adalah jumlah objek o dalam klaster C_i .
3. $\sum_{o' \in C_i, o' \neq o} \text{dist}(o, o')$ adalah jarak *Euclidean* antara objek o dan objek lainnya o' dalam klaster yang sama.

6.2. Menghitung Nilai Rata-rata Jarak Antar klaster Terdekat (b)

Hitung rata-rata jarak antara objek o dan semua objek lain o' dalam klaster terdekat lainnya C_j , dengan ketentuan $C_j \neq C_i$. Rata-rata jarak antar-klaster ini dinotasikan sebagai $b(o)$ dengan $j \neq i$. Rumus untuk menghitung b adalah sebagai berikut:

$$b(o) = \frac{1}{|C_j|} \sum_{o' \in C_j} \text{dist}(o, o') \quad (2.5)$$

Keterangan:

- $b(o)$ adalah rata-rata jarak antara objek o dan objek lain o' di dalam klaster lainnya C_j .
- $|C_j|$ adalah jumlah objek dalam klaster terdekat lainnya C_j .
- $\min_{j \neq i}$ adalah nilai minimum dari jarak rata-rata objek o ke klaster lain $b(o)$.
- $\sum_{o' \in C_j} \text{dist}(o, o')$ adalah jarak *Euclidean* antara objek o dan objek lainnya o' dalam klaster yang berbeda C_j .

6.3. Menghitung Nilai *Silhouette* untuk Setiap Objek ($s(o)$)

Hitung koefisien *Silhouette* untuk setiap objek o dengan rumus berikut:

$$s(o) = \frac{b(o) - a(o)}{\max(a(o), b(o))} \quad (2.6)$$

Keterangan:

- $s(o)$ adalah koefisien *silhouette* untuk objek o yang menunjukkan seberapa baik objek o dikelompokkan dalam klasternya. Nilai $s(o)$ berkisar antara -1 dan 1 dengan diinterpretasikan sebagai berikut, yaitu:

Tabel 2.1. Interpretasi Nilai $s(o)$

Interpretasi Nilai $s(o)$	Keterangan
Jika $s(o) \approx 1$	objek o sangat cocok dengan klasternya
Jika $s(o) \approx 0$	objek o berada dibatas antara dua klaster dengan nilai mendekati 1 berarti cocok, dan mendekati -1 berarti tidak cocok.
Jika $s(o) \approx -1$	objek o lebih dekat ke klaster lain dibanding klasternya

- $a(o)$ adalah rata-rata jarak antara objek o dan objek lain o' di dalam klaster sama C_i .
- $b(o)$ adalah rata-rata jarak antara objek o dan objek lain o' di dalam klaster lainnya C_j .
- $\max(a(o), b(o))$ adalah fungsi *max* memastikan nilai $s(o)$ akan berada dalam rentang -1 hingga 1 dengan membandingkan kedekatan objek dalam klaster dan klaster lain.

6.4. Menghitung Nilai Rata-rata *Silhouette Coefficient* (S)

Langkah terakhir adalah menilai kualitas keseluruhan klastering dengan menghitung rata-rata nilai $s(o)$ untuk semua objek o dalam *dataset* D . Rata-rata ini merupakan *Silhouette Coefficient* dari klastering. Rumus untuk menghitung nilai S adalah:

$$S = \frac{1}{n} \sum_{o \in D} s(o) \quad (2.7)$$

Keterangan:

- S adalah rata-rata koefisien *silhouette* untuk seluruh *dataset* D . Interpretasikan Nilai S menurut Dhewayani et al (2022) sebagai berikut yaitu:

Tabel 2.2. Interpretasi Nilai S

Interpretasi Nilai S	Keterangan
$0.7 < S \leq 1$	Struktur klaster sangat kuat
$0.5 < S \leq 0.7$	Struktur klaster Cukup Kuat atau standar
$0.25 < S \leq 0.5$	Struktur klaster lemah
$S \leq 0.25$	Struktur klaster sangat lemah Struktur ini dianggap kualitas gagal

- n adalah Jumlah objek dalam *dataset* D .
- $s(o)$ adalah Koefisien *silhouette* untuk setiap objek o .

Metode *Silhouette Coefficient* memiliki beberapa kelebihan dibandingkan metode lain, seperti *elbow*, *doubin*, dan metode lainnya. Menurut Davis (2007), kelebihan metode ini antara lain:

1. Skala Nilai Rentang Jelas

Nilai *Silhouette Coefficient* berkisar dari -1 hingga 1. Nilai positif menunjukkan bahwa objek berada di klaster yang tepat, dengan nilai mendekati 1 sebagai indikator terbaik. Nilai mendekati 0 menunjukkan bahwa objek berada di perbatasan antara dua klaster, sedangkan nilai negatif menunjukkan bahwa objek mungkin lebih cocok di klaster lain (Dhewayani et al, 2022).

2. Evaluasi Dalam Berbagai Tingkatan

Silhouette Coefficient mengukur kualitas klastering dengan mempertimbangkan kohesi (seberapa dekat titik data dalam klaster yang sama) dan separasi (seberapa jauh titik data dari klaster yang berbeda). Evaluasi dilakukan pada tingkat data individu serta pada keseluruhan klaster.

3. Aplikasi Luas

Silhouette Coefficient dapat digunakan dengan berbagai algoritma klastering, seperti *K-Means*, *Agglomerative Clustering*, dan *DBSCAN*. Hal ini menjadikan *Silhouette Coefficient* sebagai alat yang serbaguna untuk memvalidasi klaster di berbagai konteks dan jenis data.

7. Tendensi Pusat

Ukuran tendensi pusat menurut Kurniawan et al (2019) adalah statistik yang mencerminkan kecenderungan data untuk berkumpul di sekitar nilai tengah suatu distribusi. Nilai ini dianggap representatif terhadap hasil pengukuran data yang dihimpun. Ukuran rata-rata berfungsi sebagai indikator sentral

atau lokasi distribusi frekuensi. Tiga ukuran tendensi pusat yang umumnya digunakan adalah *mean*, median, dan modus.

7.1 **Mean (rata-rata)**

Mean atau rata-rata hitung adalah total semua nilai dibagi jumlah observasi. Ukuran ini memberikan ringkasan awal yang menggambarkan data numerik dengan mempertimbangkan semua nilai pengamatan sehingga sangat berguna sebagai langkah awal dalam penelitian.

7.2 **Median**

Median adalah nilai sentral dalam suatu kumpulan data yang telah diurutkan. Apabila jumlah data ganjil, median merupakan nilai yang membagi kumpulan data menjadi dua bagian yang sama besar. Sebaliknya, jika jumlah data genap, median dihitung dengan cara mengambil rata-rata dari dua nilai yang terletak di tengah.

7.3 **Modus**

Modus adalah nilai yang paling sering muncul dalam suatu kumpulan data. Meskipun sering dianggap kurang signifikan dibandingkan dengan nilai rata-rata (*mean*) dan nilai tengah (median), keberadaan modus tetap memiliki peranan yang penting. Dalam sebuah *dataset*, mungkin terdapat dua atau lebih modus, atau bahkan tidak ada modus sama sekali.

8. **Interquartile Range**

IQR (*Interquartile Range*) menurut Kurniawan et al (2019) adalah ukuran yang digunakan untuk mengukur sebaran data dengan mengabaikan *outliers*. IQR dihitung dari selisih antara

kuartil atas (Q3) dan kuartil bawah (Q1), yang membantu dalam memahami kecenderungan data, terutama saat data tidak simetris. IQR mencakup 50% data tengah, sehingga kurang memberikan gambaran yang utuh tentang keseluruhan sebaran data. Selain itu, IQR juga dapat digunakan untuk mendeteksi *outliers*. Data yang berada di luar batas atas ($Q3 + 1,5 \times IQR$) atau batas bawah ($Q1 - 1,5 \times IQR$) dianggap sebagai *outliers*. Namun, IQR dapat menyesatkan jika data yang dianalisis terlalu sempit.

Ketika memeriksa sekumpulan data, sering kali ditemukan nilai atau posisi yang jauh berbeda dari data lainnya, yang juga disebut sebagai *outliers*. Untuk mendeteksi *outliers*, IQR dapat digunakan. Pertama, tentukan nilai Q1 (kuartil bawah) dan Q3 (kuartil atas). Selanjutnya, hitung nilai IQR dengan mengurangkan Q1 dari Q3. Setelah itu, cari batas bawah dan batas atas menggunakan rumus $Q1 - 1,5 \times IQR$ dan $Q3 + 1,5 \times IQR$. Data yang berada di luar batas ini dianggap sebagai *outliers*.

9. Bisnis Penjualan dan Pemasaran

Menurut Sutarno (2012), bisnis adalah usaha terorganisir untuk memproduksi dan menjual produk atau layanan demi memperoleh laba serta kesejahteraan masyarakat. Dalam konteks ini, penjualan memiliki peran vital dan pemasaran menjadi kunci untuk mencapai tujuan tersebut. Menurut Kotler and Keller (2016), pemasaran didefinisikan sebagai kegiatan yang bertujuan untuk menciptakan, mengkomunikasikan, dan menukar penawaran bernilai guna memenuhi kebutuhan pelanggan. Manajemen pemasaran mencakup pemilihan pasar sasaran dan membangun hubungan dengan pelanggan melalui nilai yang unggul. Pemasar berusaha mengundang respons dari prospek

untuk mempengaruhi permintaan produk dan mencapai tujuan organisasi. Pengusaha mengelola barang, jasa, dan pengalaman sambil merespons kondisi permintaan yang beragam. Tujuan utama pemasaran adalah memahami kebutuhan pelanggan sehingga produk dan layanan dapat terjual dengan sendirinya, serta menciptakan nilai yang memenuhi kebutuhan masyarakat secara efektif.

Pangsa pasar dan segmen pasar merupakan dua konsep krusial dalam dunia pemasaran. Pangsa pasar merujuk pada persentase penjualan produk suatu perusahaan dibandingkan dengan total penjualan di industri yang bersangkutan. Sementara itu, segmen pasar adalah proses pengelompokan konsumen berdasarkan karakteristik atau perilaku mereka dengan tujuan mengembangkan strategi pemasaran yang lebih efektif. Untuk meningkatkan pangsa pasar, perusahaan perlu menargetkan segmen yang sesuai (Davis, 2007).

Menurut Sutarno (2012), proses pemasaran melibatkan analisis situasi yang mencakup analisis lingkungan internal dan eksternal untuk mengidentifikasi peluang, memahami kemampuan perusahaan, serta lingkungan operasionalnya. Hal ini penting untuk memenuhi kebutuhan pelanggan dan menjaga keseimbangan antara permintaan dan penawaran. Analisis ini dapat dibagi menjadi dua kategori utama sebagai berikut:

1. Analisis Lingkungan Internal

Fokus pada faktor mikro perusahaan seperti produksi, pemasaran, modal, dan sumber daya manusia.

2. Analisis Lingkungan Eksternal

Menghadapi faktor makro yang mempengaruhi perusahaan

termasuk pelanggan, kompetitor, serta aspek politik dan sosial.

Segmentasi pasar adalah proses pembagian pasar menjadi kelompok pelanggan yang memiliki kebutuhan serupa. Pemasar melakukan segmentasi pasar untuk mengidentifikasi kelompok yang dapat dilayani secara efektif. Menurut Sutarno (2012), kategori segmentasi pasar meliputi:

1. **Segmentasi Geografis**

Pembagian pasar berdasarkan lokasi geografis, seperti negara, wilayah, kota, atau lingkungan.

2. **Segmentasi Demografis**

Pembagian pasar berdasarkan karakteristik demografis seperti usia, jenis kelamin, pendapatan, pendidikan, dan status sosial.

3. **Segmentasi Psikografis**

Pembagian pasar berdasarkan gaya hidup, kepribadian, nilai, dan minat konsumen.

4. **Segmentasi Perilaku**

Pembagian pasar berdasarkan perilaku konsumen termasuk pola pembelian, frekuensi penggunaan, dan respon terhadap produk.

Menurut Davis (2007), untuk menentukan apakah pangsa pasar tinggi atau rendah. Rumus berikut dapat digunakan untuk menghitung pangsa pasar yaitu:

$$M_{it} = \frac{S_{it}}{\sum S_t} \quad (2.8)$$

Keterangan:

- M_{it} adalah Pangsa pasar perusahaan i di waktu t yang ditunjukkan dengan hasil presentase.
- S_{it} adalah Penjualan produk perusahaan i dalam waktu t .
- S_t adalah Jumlah semua Penjualan produk perusahaan dalam waktu t .

Penentuan pangsa pasar sesuai dengan hasil persamaan 2.8, memberikan informasi yang berguna untuk menganalisis penjualan produk, khususnya di lokasi tertentu. Perbandingan total penjualan produk sejenis seperti gelas atau botol, baik dari perusahaan sendiri maupun dari pesaing (jika data tersedia) sehingga dapat memperoleh wawasan yang berharga. Informasi ini membantu mengidentifikasi produk dengan penjualan tertinggi atau terendah sehingga menyesuaikan strategi penjualannya.

B. Kajian Pustaka

Berikut adalah beberapa penelitian terdahulu yang menjadi referensi dalam penelitian ini, yaitu:

1. Penelitian Núñez Arias et al (2021)

Penelitian yang dilakukan oleh Núñez Arias et al (2021) dengan judul *CRISP-DM and K-means neutrosophic in the analysis of risk factors for hearing loss in children* pada jurnal *Neutrosophic Computing & Machine Learning* bertujuan untuk mengidentifikasi faktor-faktor risiko yang menyebabkan gangguan pendengaran pada anak-anak usia 0 hingga 36 bulan di Rumah Sakit Universitas Provinsi Holguin,

Kuba. Pendekatan yang digunakan adalah model *CRISP-DM* dan algoritma *K-Means neutrosophic* untuk menganalisis data dan menghasilkan segmentasi informasi yang akurat. Temuan utama penelitian ini adalah identifikasi risiko dini terhadap gangguan pendengaran *sensorineural* yang dapat membantu praktisi kesehatan dalam merumuskan strategi pencegahan di masa depan.

Kesamaan penelitian tersebut dengan penelitian penulis terletak pada metode yang digunakan, yaitu *CRISP-DM* dan algoritma *K-Means*. Namun, penelitian penulis berfokus pada bidang bisnis dengan tambahan metode *Silhouette Coefficient* untuk mendukung algoritma *K-Means* dalam menghasilkan klaster yang lebih optimal.

2. Penelitian Dhewayani et al (2022)

Penelitian yang dilakukan oleh Dhewayani et al (2022) berjudul *Implementasi K-Means Clustering untuk Pengelompokan Daerah Rawan Bencana Kebakaran Menggunakan Model CRISP-DM* dalam *Jurnal Teknologi dan Informasi (JATI) Vol.12 No.1 Edisi Maret 2022* bertujuan untuk mengelompokkan daerah yang memiliki potensi kebakaran tinggi, guna meningkatkan kesadaran masyarakat terhadap risiko kebakaran. Hasil penelitian menunjukkan bahwa daerah dalam klaster 3 memiliki indeks potensi kebakaran tertinggi, dengan rata-rata 87 kejadian per tahun. Penelitian ini menggunakan model *CRISP-DM* dan algoritma *K-Means*, serta didukung oleh metode *Silhouette Coefficient*.

Perbedaan utama dengan penelitian penulis adalah fokus yang diarahkan pada bidang bisnis, serta penggunaan

beberapa perangkat lunak seperti *MATLAB 2013* , *RStudio 2024.09.0+375*, dan *Ms. Excel 2013*. Selain itu, penelitian ini akan dilakukan di wilayah kecamatan Kota Semarang.

3. Penelitian Khumaidi (2020)

Penelitian oleh Khumaidi (2020) berjudul *Data Mining for Predicting the Amount of Coffee Production Using CRISP-DM Method* dalam jurnal *TECHNO Nusa Mandiri* bertujuan untuk memprediksi hasil produksi kopi menggunakan metode *CRISP-DM* dan algoritma *multiple linear regression*. Hasil penelitian menunjukkan bahwa luas areal perkebunan, curah hujan, tekanan udara, dan intensitas penyinaran matahari memiliki pengaruh signifikan terhadap variasi hasil produksi kopi dengan koefisien determinasi sebesar 91,96%. Evaluasi prediksi menunjukkan akurasi yang baik, dengan nilai *RMSE* sebesar 0,3477.

Kesamaan dengan penelitian penulis terletak pada penggunaan metode *CRISP-DM*. Namun, penelitian ini berbeda karena fokusnya adalah wilayah tertentu dengan penerapan algoritma *K-Means*.

4. Penelitian Fauzan et al (2023)

Penelitian oleh Fauzan et al (2023) berjudul *Penerapan Data Mining untuk Menganalisis Penjualan Air Minum dalam Kemasan Selama Masa Pandemi COVID-19* dalam *Jurnal Mahasiswa Teknik Informatika (JATI)* bertujuan untuk menganalisis data penjualan air minum dalam kemasan selama pandemi COVID-19 menggunakan metode *K-Means* dan *KNN*. Hasil penelitian menunjukkan bahwa algoritma *K-Means* berhasil mengidentifikasi tingkat penjualan

tertinggi dan terendah, sementara *KNN* mencapai tingkat akurasi sebesar 85%. Penelitian ini juga mengungkapkan bahwa penjualan air minum dalam kemasan terdampak signifikan oleh pandemi, dan analisis data mining dapat membantu perusahaan menjaga stabilitas penjualan.

Kesamaan dengan penelitian ini adalah penggunaan algoritma *K-Means* dan data berupa penjualan produk dari perusahaan distribusi AQUA. Namun, penelitian penulis menggunakan model *CRISP-DM* dan berfokus pada produk AQUA dalam kemasan jenis botol dan gelas.

5. Penelitian Marsono (2019)

Penelitian oleh Marsono (2019) yang berjudul *Analisis Data Mining pada Strategi Penjualan Produk PT Aquasolve Sanaria dengan Menggunakan Metode K-Means Clustering dalam Jurnal Teknologi Sistem Informasi dan Sistem Komputer TGD (J-SISKO TECH)* bertujuan untuk mengelola data PT Aquasolve Sanaria melalui metode *K-Means* klastering. Hasil penelitian menunjukkan bahwa pengumpulan data pasokan dan produksi sangat penting untuk menetapkan strategi penjualan yang efektif di perusahaan. Penerapan *data mining* dengan algoritma *K-Means* memungkinkan perusahaan merumuskan strategi penjualan berdasarkan hasil klasterisasi.

Persamaan dengan penelitian ini terletak pada penggunaan algoritma *K-Means*. Namun, penelitian penulis menggunakan model *CRISP-DM* yang dilengkapi dengan algoritma *K-Means*, serta berfokus pada data wilayah penjualan.

BAB III

Metode Penelitian

A. Sumber Data

Menurut Abubakar (2021), sumber data dalam penelitian dapat diperoleh melalui cara-cara berikut:

1 *Person*

Data yang berasal dari orang melalui wawancara atau angket.

2 *Place*

Data yang berasal dari tempat, diperoleh melalui observasi kondisi statis maupun dinamis, seperti ruangan dan kelengkapan alat.

3 *Paper*

Data berupa simbol, seperti huruf, angka, gambar, atau kertas, misalnya laporan dan buku.

Sumber data dalam penelitian ini adalah data *Paper* berupa laporan transaksi dari Perusahaan distribusi Wahyu Jaya Cabang Kota Semarang yang mencakup wilayah Kota Semarang dan sekitarnya pada periode November 2023. Data ini dapat dilihat pada Lampiran 1 yang menunjukkan bahwa seluruh data transaksi dari tanggal 1 hingga 30 November 2023 mencakup 25.210 data transaksi dengan 15 atribut kolom.

B. Populasi dan Sampel

Populasi adalah kumpulan seluruh objek atau subjek yang memiliki karakteristik tertentu (Kesumawati dkk, 2017). Dalam penelitian ini, populasi terdiri atas data transaksi penjualan produk AMDK di perusahaan distribusi Wahyu Jaya Cabang Kota Semarang selama bulan November 2023 yang mencakup 25.210 transaksi pada 16 kecamatan di Kota Semarang, sebagaimana dirinci pada Lampiran 1. Penelitian ini menggunakan teknik pengambilan sampel sensus atau sampel jenuh, yaitu teknik yang melibatkan seluruh anggota populasi sebagai sampel karena jumlah populasi dianggap terbatas (Widodo, 2017). Dengan demikian, sampel penelitian meliputi data transaksi penjualan produk AMDK di perusahaan distribusi Wahyu Jaya Cabang Kota Semarang selama bulan November 2023 yang mencakup 25.210 transaksi pada 16 kecamatan di Kota Semarang.

C. Atribut Penelitian

Menurut Muflikhah et al (2018), data adalah kumpulan objek yang memiliki berbagai atribut, di mana atribut atau variabel tersebut merepresentasikan karakteristik dari masing-masing objek. Atribut-atribut ini diklasifikasikan ke dalam beberapa jenis berdasarkan karakteristiknya, yaitu:

1. **Atribut Nominal** adalah atribut ini menunjukkan kategori tanpa urutan atau nilai numerik. Contohnya adalah jenis kelamin atau warna produk.
2. **Atribut Ordinal** adalah atribut yang memiliki urutan, namun jaraknya tidak harus sama. Contohnya adalah peringkat

kepuasan pelanggan.

3. **Atribut Interval** adalah atribut ini memiliki skala dengan jarak yang konsisten antar nilai, tetapi tidak memiliki titik nol absolut. Contohnya adalah suhu dalam *Celsius*.
4. **Atribut Rasio** adalah atribut yang memiliki nilai nol absolut dan interval yang seragam, memungkinkan perbandingan proporsional. Contohnya adalah berat atau jumlah produk.

Atribut dalam penelitian ini terdiri dari 15 atribut dengan tipe data yang bervariasi yang dirancang untuk mencakup karakteristik objek yang relevan dalam sumber data penelitian. Berikut ini adalah daftar atribut yang terdapat dalam sumber data penelitian beserta tipe data masing-masing:

Tabel 3.1. Atribut Sumber data

No	Atribut	Tipe Data	Keterangan
1.	Nomor	Nominal	Nomor data
2.	Nomor Dokumen	Nominal	Nomor Kode Transaksi
3.	Tanggal Dokumen	Interval	Tanggal Transaksi
4.	Status Dokumen	Nominal	Applied: Transaksi Berhasil Void/Draft: Transaksi Gagal
5.	Nomor Sales Order	Nominal	Nomor Kode Sales Pengirim
6.	ID Pelanggan	Nominal	ID Pelanggan
7.	Nama Pelanggan	Nominal	Nama Toko Pelanggan
8.	Alamat	Nominal	Alamat Toko Pelanggan
9.	ID Produk	Nominal	ID Nama Produk
10.	Nama Produk	Nominal	Nama Produk
11.	Jumlah Produk	Ratio	Jumlah Transaksi Produk
12.	Satuan Produk	Nominal	Satuan Produk

No	Atribut	Tipe Data	Keterangan
13.	Harga Produk	Ratio	Harga Produk
14.	Total Diskon Produk	Ratio	Diskon Transaksi Produk
15.	DPP (Pendapatan Bersih)	Ratio	Pendapatan Bersih

Tabel 3.1 menjelaskan 15 atribut pada sumber data, termasuk tipe data dan keterangan untuk setiap kolom. Tabel ini berfungsi sebagai menentukan atribut yang akan digunakan dalam perhitungan penelitian ini. Berikut ini adalah daftar atribut perhitungan dalam penelitian beserta tipe data masing-masing:

Tabel 3.2. Atribut Perhitungan Penelitian

No	Atribut	Tipe Data
1.	Alamat	Nominal
2.	AQUA 1500ML	Ratio
3.	AQUA 220ML	Ratio
4.	AQUA 330ML	Ratio
5.	AQUA 600ML	Ratio
6.	Mizone Activ Lychee Lemon 550ML	Ratio
7.	VIT 1500ML	Ratio
8.	VIT 200ML	Ratio
9.	VIT 330ML	Ratio
10.	VIT 550ML	Ratio

Tabel 3.2 di atas menunjukkan atribut-atribut yang digunakan dalam perhitungan *K-Means*. Atribut Alamat bertipe nominal yang berupa kategori tanpa urutan atau nilai numerik. Sementara itu, atribut seperti AQUA 1500ML, AQUA 220ML, hingga VIT 550ML bertipe rasio yang bernilai numerik dan memiliki nol yang menunjukkan tidak adanya jumlah produk. Dalam analisis *K-Means*, atribut rasio ini digunakan untuk mengelompokkan data

ke dalam klaster berdasarkan kedekatan nilai antar atribut.

D. Teknik Pengumpulan Data

Teknik pengumpulan data sangat penting dalam penelitian untuk memperoleh data yang valid mengenai atribut yang diteliti (Sugiyono, 2014). Data penelitian dapat diperoleh dari sumber primer atau sekunder. Data primer adalah data yang dikumpulkan secara langsung melalui kuesioner, survei, atau metode lain, sedangkan data sekunder adalah data yang diperoleh dari sumber lain seperti dokumentasi, arsip, atau data perusahaan (Sundari dkk, 2024). Penelitian ini menggunakan data sekunder berupa basis data perusahaan distribusi Wahyu Jaya Cabang Kota Semarang untuk periode November 2023.

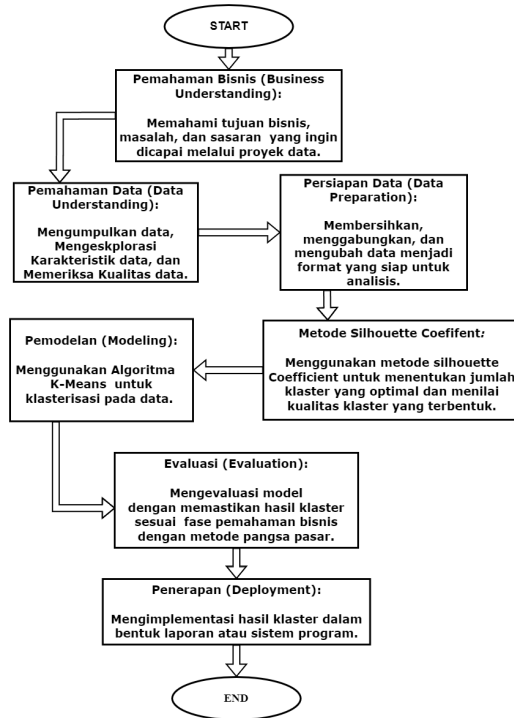
E. Studi Literatur

Studi literatur adalah metode penting dalam penulisan karya ilmiah yang mencakup pengumpulan, pembacaan, pencatatan, dan pengelolaan data pustaka. Metode ini krusial untuk penelitian yang bertujuan mengembangkan aspek teori dan manfaat praktis (Sundari dkk, 2024). Penelitian ini menggunakan studi literatur dengan merujuk pada jurnal dan buku dalam Daftar Pustaka.

F. Teknik Analisis Data

Metode analisis data yang digunakan dalam penelitian ini adalah model CRISP-DM dengan algoritma *K-Means* dan metode *Silhouette Coefficient* untuk mengelompokkan produk gelas dan botol dari Perusahaan Distribusi Wahyu Jaya Cabang Kota

Semarang berdasarkan wilayah 16 Kecamatan di Kota Semarang.. Software yang digunakan adalah *Matlab versi R2013a*, *RStudio versi 2024.09.0+375*, dan *Microsoft Excel 2013*. Model fase CRISP-DM yang diterapkan dalam teknik Analisis Data adalah sebagai berikut:



Gambar 3.1. *Flowchart Model CRISP-DM* (Sumber:Shearer (2000))

Berikut ini adalah penjelasan mengenai Gambar 3.1 yang memaparkan setiap komponen alur model CRISP-DM secara lengkap, yaitu:

1. **START**

Langkah untuk memulai aliran program.

2. **Pemahaman Bisnis (*Business Understanding*)**

Tahap Pemahaman Bisnis (*Business Understanding*) bertujuan untuk menentukan masalah serta tujuan penelitian atau proyek dengan mempertimbangkan syarat dan batasan yang relevan terhadap produk-produk perusahaan distribusi Wahyu Jaya Cabang Kota Semarang, khususnya jenis gelas dan botol berbagai ukuran berdasarkan wilayah kecamatan di Kota Semarang.

Selanjutnya, masalah dan tujuan penelitian ini akan diinterpretasikan ke dalam proses *Data Mining* dengan model *CRISP-DM* dengan metode Algoritma *K-Means* dan metode *Silhouette Coefficient* yang telah dipilih dalam penelitian ini. Sasaran yang ingin dicapai adalah memastikan adanya pemahaman bisnis yang mendalam agar analisis data dapat dilakukan secara efektif dan menghasilkan solusi yang tepat serta relevan bagi permasalahan dan tujuan penelitian atau proyek.

3. **Pemahaman Data (*Data Understanding*)**

Tahap Pemahaman Data (*Data Understanding*) dimulai dengan pengumpulan data yang relevan untuk penelitian atau proyek. Pada tahap ini, data dieksplorasi secara mendalam untuk memahami karakteristik awal dan

mengevaluasi kualitasnya, termasuk mengidentifikasi *missing value*, *outliers*, dan data duplikat. Setelah pengumpulan data, dilakukan analisis sederhana untuk memahami karakteristik dasar dari data yang diperoleh. Pemeriksaan kualitas data juga dilakukan untuk memastikan bahwa data yang digunakan benar-benar berkualitas baik dan representatif dalam menyelesaikan masalah bisnis. Hasil dari tahap ini adalah gambaran awal mengenai data, termasuk masalah yang perlu ditangani sebelum melanjutkan ke tahap berikutnya. Jika data telah melewati proses pemeriksaan dan pembersihan ini, data tersebut disebut sebagai Data Proses.

4. Persiapan Data (*Data Preparation*)

Tahap Persiapan Data (*Data Preparation*) adalah fase penting dalam proses *CRISP-DM* yang bertujuan menyiapkan data agar siap digunakan dalam tahap pemodelan. Tahap ini melibatkan beberapa langkah, termasuk pembersihan data dari *missing value*, *outliers*, dan data duplikat. Selain itu, dilakukan transformasi atribut untuk memastikan data berada dalam format yang sesuai, serta reduksi dan integrasi data jika terdapat fitur yang tidak relevan atau sumber data yang berbeda. Melalui tahap ini, data dari berbagai sumber dapat digabungkan menjadi satu *dataset* yang koheren dan konsisten. Kemudian, pembersihan data dari *missing value*, *outliers*. Hasil dari Persiapan Data (*Data Preparation*) adalah *dataset* akhir yang bersih, terstruktur, dan siap digunakan dalam tahap pemodelan. Kualitas data yang dihasilkan pada tahap ini sangat penting karena akan memengaruhi kualitas model yang dibangun dan pada akhirnya, menentukan

keberhasilan penelitian.

5. **Metode *Silhouette Coefficient***

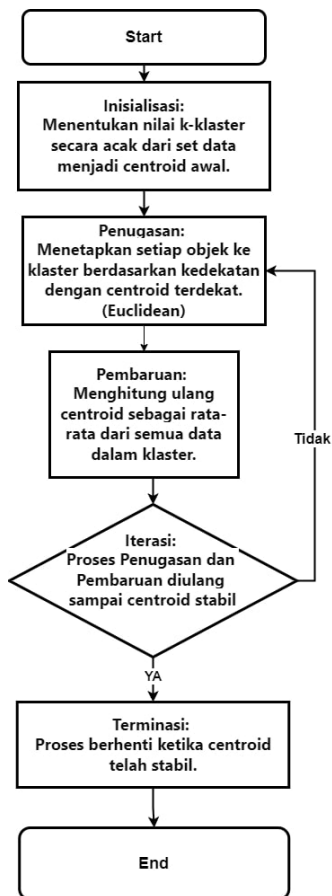
Sebelum memasuki tahap pemodelan (*Modeling*), metode *Silhouette Coefficient* akan digunakan untuk menentukan jumlah kluster yang optimal dan mengevaluasi kualitas pengelompokan data. Proses ini dilakukan dengan mengukur kekuatan kluster berdasarkan persamaan 2.4, 2.5, 2.6, dan 2.7.

6. **Pemodelan (*Modeling*)**

Tahap Pemodelan (*Modeling*) adalah tahap penting dalam proses analisis data, di mana teknik-teknik pemodelan diterapkan untuk mengidentifikasi pola atau hubungan tertentu dalam data yang telah disiapkan. Pada tahap ini, berbagai metode *data mining* seperti klusterisasi dapat digunakan sesuai dengan tujuan analisis yang ingin dicapai. Salah satu metode klusterisasi yang digunakan adalah algoritma *K-Means* yang berfungsi untuk mengelompokkan data penelitian ke dalam beberapa kluster berdasarkan kemiripan karakteristiknya.

Setelah menentukan jumlah kluster optimal, algoritma *K-Means* diterapkan dengan menggunakan rumus persamaan yang tercantum dalam persamaan 2.1, 2.2, dan 2.3. Algoritma ini secara bertahap mengelompokkan data ke dalam kluster-kluster hingga menemukan hasil akhir yang paling optimal sehingga setiap data dalam satu kluster memiliki karakteristik sama. Hasil dari tahap pemodelan ini adalah model yang telah dilatih menggunakan data yang siap untuk dievaluasi lebih lanjut.

Tahap Pemodelan yang diperlihatkan pada Gambar 3.1 memaparkan algoritma *K-Means* yang digunakan pada tahapan pemodelan. Berikut alur penelitian untuk langkah-langkah dalam algoritma *K-Means*, yaitu:



Gambar 3.2. Tahapan Modeling Algoritma *K-Means* (Sumber: Hans dkk (2012))

Berikut ini adalah penjelasan dalam Gambar 3.2 dalam alur penelitian tahapan pemodelan menggunakan algoritma *K-Means*, yaitu:

- 6.1. **Start:** Langkah untuk memulai aliran program.
- 6.2. **Inisialisasi:** Langkah untuk menentukan nilai kluster secara acak dari *dataset* menjadi *centroid* awal yang telah diambil dari Metode *Shiloutte Coefficient*.
- 6.3. **Penugasan:** Langkah untuk menetapkan setiap objek ke kluster berdasarkan kedekatan dengan *centroid* terdekat menggunakan jarak *Euclidean* dalam persamaan 2.1 dan persamaan 2.2 sebagai penentuan kedekatan data dengan kluster.
- 6.4. **Pembaruan:** Jalur jika *centroid* belum stabil. Langkah untuk menghitung ulang *centroid* baru sebagai rata-rata dari semua data dalam kluster dalam persamaan 2.3.
- 6.5. **Iterasi:** Setelah *centroid* baru dihitung, maka langkah ini menghitung ulang *centroid* baru sebagai rata-rata dari semua data dalam kluster dengan mengulang-ulang langkah penugasan dan langkah pembaruan hingga *centroid* stabil. Apakah *centroid* stabil?.
- 6.6. **Terminasi:** Proses Iterasi akan berhenti ketika *centroid* stabil dan Hasil akan ditampilkan.
- 6.7. **End:** Langkah untuk memberhentikan aliran program.

7. Evaluasi (*Evaluation*)

Tahap Evaluasi (*Evaluation*) bertujuan untuk memastikan bahwa model yang telah dibangun sesuai dengan masalah dan tujuan bisnis yang telah ditetapkan pada tahap *Business Understanding*. Pada tahap ini, efektivitas model dievaluasi menggunakan metode pangsa pasar untuk menganalisis penjualan dengan dimulai dari kategori terendah hingga tertinggi sesuai dengan persamaan 2.8. Setelah proses evaluasi model selesai dan hasilnya memenuhi kriteria yang ditetapkan, tahap ini ditutup dengan memberikan rekomendasi untuk penerapan (*Deployment*) model ke dalam sistem bisnis agar hasil analisis dapat diimplementasikan secara efektif.

8. Penerapan (*Deployment*)

Tahap Penerapan (*Deployment*) merupakan tahap akhir dalam model *CRISP-DM*, di mana evaluasi hasil model yang telah dihasilkan diimplementasikan ke dalam sistem bisnis untuk memberikan solusi terhadap masalah yang diidentifikasi serta mencapai tujuan penelitian. Pada tahap ini, hasil analisis diintegrasikan ke dalam proses bisnis yang ada, misalnya dalam bentuk laporan atau format lainnya yang dapat langsung mendukung pengambilan keputusan secara efektif. Implementasi ini bertujuan untuk memastikan bahwa wawasan yang diperoleh dari analisis dapat diterapkan dalam praktik dan memberikan nilai tambah bagi organisasi.

9. *END*

Langkah untuk menghentikan aliran program.

BAB IV

Hasil Penelitian dan Pembahasan

Bab ini membahas hasil analisis pola penjualan menggunakan model *CRISP-DM* dan algoritma *K-Means* di Wahyu Jaya Group Cabang Semarang. Pembahasan dilakukan untuk mengidentifikasi pola penjualan yang ditemukan serta memberikan interpretasi terhadap hasil pengelompokan yang dihasilkan. Selain itu, bab ini juga mengevaluasi efektivitas penerapan metode yang digunakan dalam membantu perusahaan memahami distribusi dan tren penjualan produk. Berikut ini proses hasil dan pembahasan analisis pola penjualan menggunakan model *CRISP-DM* dan algoritma *K-Means* di Wahyu Jaya Group Semarang, yaitu:

A. Pemahaman Bisnis (*Business Understanding*)

Tahap Pemahaman Bisnis dalam penelitian ini bertujuan untuk mengidentifikasi masalah dan merumuskan tujuan analisis untuk Perusahaan Distribusi Wahyu Jaya Cabang Semarang. Perusahaan ini mendistribusikan produk AMDK, seperti *AQUA* dan *VIT* dalam berbagai ukuran. Namun, perusahaan kesulitan dalam menemukan pola penjualan dan optimalisasi segmentasi pasar di Kota Semarang.

Permasalahan utama yang dihadapi adalah penurunan tingkat konsumsi AMDK dalam beberapa tahun terakhir yang menimbulkan kekhawatiran akan dampaknya terhadap pendapatan dan pertumbuhan bisnis. Untuk tetap kompetitif di pasar, Perusahaan Wahyu Jaya perlu memahami pola penjualan dan preferensi konsumen di setiap kecamatan yang

menjadi cakupan distribusinya. Pemahaman ini akan membantu perusahaan menentukan strategi pemasaran yang lebih efektif dan spesifik untuk setiap wilayah.

Dalam konteks penelitian ini, metode *data mining* dengan pendekatan model *CRISP-DM* dengan algoritma *K-Means* serta metode *Silhouette Coefficient* dipilih sebagai alat analisis. Pendekatan ini diharapkan mampu memberikan wawasan mengenai distribusi dan karakteristik penjualan AMDK berbagai wilayah di kecamatan Kota Semarang, serta membantu mengelompokkan area dengan tingkat penjualan rendah dan tinggi. Pemahaman bisnis yang mendalam penting untuk memastikan proses *data mining* sesuai kebutuhan perusahaan. Dengan menetapkan tujuan dan batasan, analisis data bisa dirancang lebih tepat dan menghasilkan solusi yang relevan untuk permasalahan di perusahaan distribusi Wahyu Jaya Cabang Kota Semarang.

B. Pemahaman Data (*Data Understanding*)

Tahap Pemahaman Data (*Data Understanding*) dalam penelitian ini dimulai dengan pengumpulan data yang relevan dari perusahaan distribusi Wahyu Jaya Cabang Semarang. Selanjutnya, data tersebut dieksplorasi secara mendalam untuk memahami karakteristik awalnya, termasuk jenis dan distribusi nilai dalam setiap kolom. Selain itu, dilakukan pemeriksaan terhadap kualitas data dengan tujuan memastikan bahwa data yang digunakan memiliki kualitas yang baik. Setelah melalui tahap eksplorasi dan pemeriksaan kualitas data, dilakukan analisis sederhana untuk memperoleh gambaran dasar mengenai karakteristik data, seperti

distribusi penjualan per wilayah dan variasi jumlah produk. Data yang telah melalui tahap pemeriksaan dan pembersihan ini kemudian disebut sebagai Data Proses. Berikut adalah hasil dan pembahasan mengenai langkah-langkah dalam tahapan ini:

1. Pengumpulan Data

Langkah pengumpulan data membahas sumber data yang dijelaskan pada BAB 3, di mana data yang diperoleh mencakup informasi transaksi penjualan produk Air Minum Dalam Kemasan (AMDK) dari berbagai merek dan ukuran pada bulan November tahun 2023. Untuk melihat keseluruhan data, informasi ini dilampirkan pada Lampiran 1 yang mencakup 25.210 data transaksi dengan total 15 atribut. Data ini telah mendapatkan izin akses penelitian dari perusahaan, dengan mematuhi peraturan yang ditetapkan oleh perusahaan, sebagaimana dijelaskan dalam Lampiran 2.

2. Analisis Karakteristik Awal

Proses analisis karakteristik ini termasuk jenis dan distribusi nilai dalam setiap atribut kolom dilakukan menggunakan *software RStudio 2024.09.0+375* dan *Microsoft Excel 2013* dengan *Script Program* pada Lampiran 3. Untuk memahami karakteristik awal dari data, langkah-langkah penting yang dapat dilakukan adalah sebagai berikut:

2.1 Identifikasi Data dan Tipe Data

Identifikasi data dan tipe data adalah langkah awal yang penting dalam analisis data, karena membantu peneliti memahami jenis informasi yang terkandung dalam

setiap kolom *dataset*. Proses identifikasi data dan tipe data menggunakan *software RStudio 2024.09.0+375* akan menampilkan hasil sebagai berikut:

Tabel 4.1. Hasil Proses Identifikasi Data

Informasi Data	Jumlah Data	
Jumlah Kolom	15	
Jumlah Baris	25210	
Atribut Kolom	Jumlah Nilai Unik	Contoh Nilai Unik
Nomor	25210	1,2,3,4,5
Nomor Dokumen	13644	852-1662303, dll
Tanggal Dokumen	26	2023-11-01, dll
Status Dokumen	3	Applied, Draft, Void
Nomor Sales Order	13092	852-1601167, dll
ID Pelanggan	3195	852-0023925, dll
Nama Pelanggan	2965	SN STORE - WARUNG MAGY - TOKO, dll
Alamat	17	Semarang Utara, NA, dll
ID Produk	22	145141, dll
Nama Produk	19	VIT 1500ML, AQUA 220ML MIZONE ACTIV LYCHEE LEMON 500ML, dll
Jumlah Produk	198	2,100,200,74, dll
Satuan Produk	6	BOX, BOTOL,NA, dll
Harga Produk	64	, 14000, 0, -30000, dll
Total Diskon Produk	223	0, 45000, NA, dll
DPP (Pendapatan Bersih)	804	75675.6757, 0, dll

Tabel 4.2. Hasil Proses Identifikasi Tipe Data

No	Atribut	Tipe Data
1.	Nomor	Character (Nominal)
2.	Nomor Dokumen	Character (Nominal)
3.	Tanggal Dokumen	POSIXct (Interval)
4.	Status Dokumen	Character (Nominal)
5.	Nomor Sales Order	Character (Nominal)
6.	ID Pelanggan	Character (Nominal)
7.	Nama Pelanggan	Character (Nominal)
8.	Alamat	Character (Nominal)
9.	ID Produk	Character (Nominal)
10.	Nama Produk	Character (Nominal)
11.	Jumlah Produk	Numeric (Ratio)
12.	Satuan Produk	Character (Nominal)
13.	Harga Produk	Numeric (Ratio)
14.	Total Diskon Produk	Numeric (Ratio)
15.	DPP (Pendapatan Bersih)	Numeric (Ratio)

Hasil identifikasi data dan tipe data dalam Tabel 4.1 menunjukkan bahwa terdapat total 25.210 data transaksi penjualan yang tercatat dengan 15 atribut kolom beserta jumlah nilai pada setiap atribut dan contoh sampel di dalamnya. Berdasarkan sampel contoh atribut, ditemukan beberapa masalah seperti nilai *NA* (*Missing value*) dan angka negatif pada beberapa atribut yang memerlukan penelitian lebih lanjut. Sementara itu, Tabel 4.2 menampilkan tipe data untuk setiap atribut kolom yang meskipun berbeda nama, memiliki makna yang sama.

2.2 Distribusi Data

Analisis distribusi data ini harus sesuai dengan

tahap *Business Understanding* dengan menampilkan penjualan produk AMDK di setiap kecamatan di Kota Semarang. Dalam analisis pola penjualan menggunakan algoritma *K-Means*, beberapa atribut kolom yang relevan adalah atribut kolom-kolom yang dapat membantu mengidentifikasi segmen atau kelompok dalam data penjualan. Biasanya, atribut kolom yang relevan meliputi informasi kuantitatif terkait volume penjualan dan pendapatan, serta faktor tambahan yang mungkin memengaruhi pola penjualan. Berikut adalah langkah-langkah dalam menganalisis distribusi data menggunakan *RStudio 2024.09.0+375* dengan *Script* program pada Lampiran 3:

2.2.1 Distribusi Jumlah Penjualan Produk dan Pendapatan berdasarkan tanggal

Distribusi jumlah produk terjual adalah atribut yang mencatat kuantitas unit produk yang terjual dalam setiap transaksi sehingga mencerminkan volume dan performa penjualan. Dalam konteks penelitian ini yang bertujuan untuk memahami pola penjualan dan segmentasi pasar, Jumlah produk terjual memainkan peran penting dalam memberikan gambaran tentang permintaan produk, popularitas produk tertentu, serta frekuensi penjualan dalam rentang waktu tertentu. Berikut ini adalah hasil analisis distribusi jumlah penjualan produk:

Tabel 4.3. Hasil Distribusi Jumlah Penjualan Produk dan Pendapatan berdasarkan tanggal

Tanggal	Jumlah Produk	Pendapatan Bersih
01-11-2023	19693	260.648.883
02-11-2023	21345	268.715.394
03-11-2023	19538	273.167.027
04-11-2023	19931	300.105.856
06-11-2023	19441	247.132.695
07-11-2023	18787	241.643.387
08-11-2023	22238	325.519.302
09-11-2023	19997	283.641.305
10-11-2023	19490	258.226.306
11-11-2023	17936	246.724.505
13-11-2023	20349	277.988.973
14-11-2023	20126	292.486.640
15-11-2023	22180	317.475.621
16-11-2023	20848	283.245.405
17-11-2023	17112	233.053.263
18-11-2023	16415	223.285.838
20-11-2023	18154	244.493.964
21-11-2023	21681	317.054.658
22-11-2023	18887	272.034.333
23-11-2023	21229	287.625.685
24-11-2023	17355	231.565.901
25-11-2023	15233	229.975.360
27-11-2023	18670	240.017.991
28-11-2023	17294	251.315.658
29-11-2023	16772	230.574.257
30-11-2023	19780	288.236.545

Hasil data distribusi penjualan harian pada Tabel 4.3 untuk bulan November 2023 menunjukkan variasi jumlah produk terjual dan pendapatan

bersih yang signifikan. Jumlah produk terjual berkisar antara 17.770 hingga 22.711 unit per hari, sedangkan pendapatan bersih harian bervariasi antara 200 hingga 310 juta. Hubungan positif antara jumlah produk terjual dan pendapatan bersih menunjukkan bahwa peningkatan penjualan menghasilkan pendapatan yang lebih tinggi, meskipun perbedaan harga antar produk juga memengaruhi total pendapatan harian. Fluktuasi ini mungkin disebabkan oleh faktor-faktor seperti promosi khusus, hari tertentu dengan tingkat penjualan lebih tinggi, atau tren mingguan yang menunjukkan peningkatan penjualan pada akhir pekan. Data ini akan dianalisis lebih lanjut untuk mengidentifikasi pola mingguan atau musiman yang dapat membantu dalam perencanaan stok dan strategi pemasaran.

2.2.2 Distribusi Jumlah Penjualan Produk dan Pendapatan berdasarkan Status Dokumen

Distribusi penjualan produk berdasarkan status dokumen dalam data mentah ini bertujuan untuk menganalisis jumlah transaksi penjualan produk sesuai dengan status transaksi yang tercatat. Penjelasan mengenai setiap kategori status dokumen dapat ditemukan dalam Gambar 3.1, yang mencakup tiga nilai unik beserta keterangan masing-masing, yaitu *Applied*, *Draft*, dan *Void*. Proses analisis ini dilakukan dengan menyajikan hasil distribusi penjualan produk berdasarkan

status dokumen. Berikut ini ditampilkan hasil analisis distribusi penjualan produk per status dokumen:

Tabel 4.4. Hasil Proses Distribusi Jumlah Penjualan Produk dan Pendapatan berdasarkan Status Dokumen

Status Dokumen	Jumlah Produk	Pendapatan Bersih
Applied	496.537	6.862.927.769
Draft	0	0
Void	3944	63.026.982

Hasil distribusi penjualan berdasarkan Status Dokumen menunjukkan bahwa status *Applied* atau transaksi berhasil memiliki jumlah produk tertinggi, yaitu sebanyak 496.537 unit, dengan pendapatan bersih sebesar 6.862.927.769. Sementara itu, status *Void* atau transaksi gagal hanya mencatat 3.944 unit produk dengan pendapatan bersih sebesar 63.026.982. Status *Draft* yang menunjukkan transaksi dalam tahap perencanaan, tidak mencatat jumlah produk maupun pendapatan. Meskipun status *Void* mencatat penjualan sebesar 3.944 unit produk, transaksi dengan status ini akan dihapus karena tidak relevan dengan penelitian ini. Proses penghapusan transaksi *Void* akan dilaksanakan pada Fase *Data Preparation*.

2.2.3 Distribusi Jumlah Penjualan Produk dan Pendapatan berdasarkan Lokasi Wilayah

Distribusi jumlah penjualan produk dan

pendapatan berdasarkan lokasi wilayah memberikan gambaran mengenai sebaran penjualan dan pendapatan di berbagai wilayah. Analisis ini bertujuan untuk memahami pola penjualan serta mengidentifikasi wilayah dengan performa penjualan yang tinggi dan rendah. Hasil distribusi jumlah penjualan produk dan pendapatan berdasarkan lokasi wilayah akan ditampilkan sebagai berikut:

Tabel 4.5. Hasil Proses Distribusi Jumlah Penjualan Produk dan Pendapatan berdasarkan lokasi wilayah

Alamat	Jumlah Produk	Pendapatan Bersih
Banyumanik	1649	18.664.955
Candisari	19946	210.877.398
Gajahmungkur	35219	468.982.432
Gayamsari	462	7.674.719
Genuk	7294	140.415.991
Gunungpati	26792	280.400.753
Mijen	59946	644.472.870
Ngaliyan	99382	1.281.793.919
Pedurungan	5206	38.064.113
Semarang Barat	100948	1.499.893.820
Semarang Selatan	28117	374.481.790
Semarang Tengah	49818	910.413.568
Semarang Timur	1247	28.937.973
Semarang Utara	24119	431.298.243
Tembalang	16412	187.538.784
Tugu	10163	140.989.144
NA	13761	261.054.279

Hasil distribusi penjualan dan pendapatan per

alamat pelanggan yang ditampilkan dalam Gambar 4.5 mencakup 17 kategori menunjukkan jumlah produk terjual dan pendapatan bersih di setiap kecamatan di Kota Semarang. Kota Semarang terdiri dari 16 kecamatan, sedangkan Gambar 4.5 mencakup 17 kategori, termasuk satu kategori *NA* (*missing value*) untuk alamat di luar Kota Semarang dengan jumlah penjualan produk sebesar 13.761 unit. Kategori *NA* (*missing value*) ini akan diproses lebih lanjut pada tahap berikutnya.

2.2.4 Distribusi Jumlah Penjualan Produk dan Pendapatan berdasarkan Jenis Produk (Nama Produk)

Distribusi jumlah penjualan produk dan pendapatan berdasarkan jenis produk (nama produk) memberikan wawasan mengenai performa masing-masing produk di pasar. Analisis ini penting untuk memahami produk-produk dari perusahaan Wahyu Jaya Semarang yang bekerja sama dengan berbagai perusahaan, seperti *Danone* dan lainnya. Hasil analisis ini diharapkan dapat membantu dalam merumuskan strategi pengelolaan produk, pengaturan inventaris, dan pemasaran yang lebih efektif, serta dalam menentukan prioritas pengembangan produk di masa mendatang. Berikut adalah hasil distribusi jumlah penjualan produk dan pendapatan berdasarkan jenis produk (nama produk):

Tabel 4.6. Hasil Proses Distribusi Jumlah Penjualan Produk dan Pendapatan berdasarkan lokasi wilayah

Nama Produk	Jumlah Produk	Pendapatan Bersih
AQUA 1500ML	31457	1.376.450.929
AQUA 220ML	9862	273.195.158
AQUA 330ML	9761	313.170.595
AQUA 5GLN BTL	3145	0
AQUA 5GLN BTL - NO VALUE	25	0
AQUA 600ML	38701	1.508.811.291
AQUA 750ML	2	153.144
AQUA GLN ISI	177986	2.552.848.739
AQUA TISSUE	178016	0
DISPENSER BAJA PUTIH 2 KERAN	34	1.066.930
MIZONE ACTIV LYCHEE LEMON 500ML	4043	144.256.534
VIT 1500ML	2566	73.894.689
VIT 200ML	17701	312.098.329
VIT 330ML	1472	39.391.239
VIT 550ML	8119	223.186.617
VIT GLN BTL	-332	0
VIT GLN ISI	8968	107.430.559
VIT TISSUE	8955	0

Tabel 4.6 ini menampilkan distribusi penjualan berdasarkan nama produk, dengan tiga atribut kolom yaitu Nama Produk, Jumlah Produk, dan Pendapatan Bersih. Produk jenis botol dan gelas seperti AQUA 1500ML dan AQUA 600ML menunjukkan volume penjualan tinggi dan kontribusi pendapatan yang signifikan,

masing-masing dengan penjualan sebanyak 31.457 unit dan 3.8862 unit, serta menghasilkan lebih dari 1,37 miliar dan 1,5 miliar juta dalam pendapatan. Sebaliknya, beberapa entri seperti *AQUA 5GLN BTL - NO VALUE* dan *NA* menunjukkan pendapatan nol atau data yang hilang (*missing value*), yang mungkin disebabkan oleh transaksi penjualan yang tidak lengkap atau produk yang sudah tidak tersedia. Pemilihan nama produk akan disesuaikan dengan kajian penelitian ini sehingga proses penyaringan nama produk yang tidak relevan akan dilakukan pada tahap *Data Preparation*.

3. Pemeriksaan dan Pembersihan Kualitas Data

Pemeriksaan dan pembersihan kualitas data dilakukan setelah mengidentifikasi permasalahan seperti data yang tidak lengkap, *missing value*, dan kendala lainnya yang terdapat pada karakteristik beberapa atribut kolom. Berikut ini merupakan proses pemeriksaan dan pembersihan kualitas data mentah menggunakan *RStudio 2024.09.0+375*:

3.1. Pemeriksaan dan Pembersihan nilai yang hilang (*missing values*)

Proses pemeriksaan dan penghapusan nilai yang hilang (*missing values*) merupakan langkah penting dalam pengolahan data. Pada tahap ini, data akan diperiksa terlebih dahulu untuk mengidentifikasi nilai yang hilang, kemudian dilakukan tindakan berupa penghapusan baris atau kolom yang terpengaruh

atau menggantinya dengan nilai estimasi yang sesuai, seperti *mean*, median, atau modus. Tujuan dari langkah ini adalah untuk memastikan bahwa analisis yang dilakukan menghasilkan hasil yang akurat dan dapat diandalkan. Berikut ini adalah hasil proses pemeriksaan dan pembersihan nilai yang hilang (*missing values*):

```
> # Deteksi dan penanganan missing values
> cat("Deteksi Missing Value:\n")
    Deteksi Missing Value:

> rows_before <- nrow(data)
> cat(sprintf("Jumlah baris sebelum menangani missing value: %d\n", rows_before))
    Jumlah baris sebelum menangani missing value: 25210

> # Menangani missing values
> data_cleaned <- na.omit(data)
> rows_after <- nrow(data_cleaned)
> cat(sprintf("Jumlah baris yang dihapus: %d\n", rows_before - rows_after))
    Jumlah baris yang dihapus: 3954

> cat(sprintf("Jumlah baris setelah menangani missing value: %d\n", rows_after))
    Jumlah baris setelah menangani missing value: 21256

> # Menampilkan beberapa baris data untuk verifikasi
> cat("Beberapa baris data setelah pembersihan:\n")
Beberapa baris data setelah pembersihan:
> print(head(data_cleaned))
# A tibble: 6 × 15
  Nomor Nomor Dokumen Tanggal Dokumen Status Dokumen Nomor Sales Order ID
Pelanggan
  <chr> <chr>          <dtm>          <chr>          <chr>          <chr>
1 1 852-1662303 2023-11-01 00:00:00 Applied 852-1601167 852-0023925
2 2 852-1662732 2023-11-01 00:00:00 Applied 852-1601336GL 852-0017206
3 3 852-1662732 2023-11-01 00:00:00 Applied 852-1601336GL 852-0017206
4 4 852-1662732 2023-11-01 00:00:00 Applied 852-1601336GL 852-0017206
5 5 852-1662732 2023-11-01 00:00:00 Applied 852-1601336GL 852-0017206
6 6 852-1662904 2023-11-01 00:00:00 Applied 852-1601563 852-0021978
# 19 more variables: `Nama Pelanggan` <chr>, `Alamat` <chr>, `ID Produk` <chr>, `Nama Pr
oduk` <chr>, `Jumlah Produk` <dbl>, `Satuan Produk` <chr>, `Harga` <dbl>, `Total Diskon`
<dbl>, `Pendapatan Bersih` <dbl>
```

Gambar 4.1. Hasil Proses Identifikasi *Missing Value*

Hasil proses dimulai dengan mendeteksi dan

menangani *missing values* pada Gambar 4.1. Jumlah transaksi sebelum penanganan adalah 25.210, dan setelah menghapus data kosong dengan fungsi *na.omit*, jumlahnya menjadi 21.256. Sebanyak 3.954 transaksi dihapus akibat *missing values*, dan beberapa data transaksi pertama dari data yang telah dibersihkan ditampilkan untuk verifikasi.

3.2. Pemeriksaan dan Pembersihan Data Duplikat

Proses pemeriksaan dan penghapusan data duplikat merupakan langkah penting dalam pengolahan data untuk memastikan keakuratan dan keandalan hasil analisis. Pada tahap ini, data diperiksa secara saksama untuk mengidentifikasi adanya entri yang sama atau duplikat, baik secara keseluruhan maupun sebagian. Selanjutnya, data duplikat dihapus agar tidak menimbulkan bias dalam analisis, mengganggu interpretasi hasil, atau mengurangi efisiensi proses pengolahan data. Berikut ini adalah proses pemeriksaan dan pembersihan data duplikat:

```
> # Deteksi duplikat
> cat("Deteksi Duplikat:\n")
      Deteksi Duplikat:
> duplicate_rows <- data_cleaned[duplicated(data_cleaned), ]
> num_duplicate_rows <- nrow(duplicate_rows)
> cat(sprintf("Jumlah baris duplikat: %d\n", num_duplicate_rows))
      Jumlah baris duplikat: 0

> if (num_duplicate_rows > 0) {
+   cat("Contoh baris duplikat:\n")
+   print(head(duplicate_rows))
+ } else {
+   cat("Tidak ada baris duplikat ditemukan.\n")
+ }
```

Tidak ada baris duplikat ditemukan.

Gambar 4.2. Hasil Proses Identifikasi Data Duplikat

Hasil yang ditampilkan pada Gambar 4.2 menunjukkan bahwa data mentah tidak mengandung transaksi duplikat. Dengan demikian, jumlah data yang akan digunakan untuk proses selanjutnya tetap sebanyak 21.256 transaksi.

3.3. Pemeriksaan dan Pembersihan Outliers

Proses pemeriksaan dan penghapusan data *outliers* adalah langkah penting dalam pengolahan data untuk memastikan keandalan dan akurasi analisis. *Outliers* adalah nilai yang secara signifikan berbeda dari data lainnya, yang dapat disebabkan oleh kesalahan pengukuran, entri data yang salah, atau variasi ekstrem yang tidak sesuai dengan pola umum. Proses ini mengidentifikasi data penjualan yang menyimpang, seperti nilai negatif dengan menggunakan metode IQR (*Interquartile Range*), sebelum dilakukan pembersihan data. Berikut adalah langkah-langkah pemeriksaan dan pembersihan data *outliers*:

Gambar 4.3. Hasil Proses Identifikasi *Outliers*

> # Deteksi outliers menggunakan metode IQR

Kolom	Jumlah Produk	Harga	Total Diskon	Pendapatan Bersih
Minimum	-424	52500	-75000	-2207207,207
Q1	2	15750	0	18018,018
Median	7	20000	0	86486,4865
Q3	20	47000	6000	219594,5946
Maximum	500	84995	3500800	22745945,95
IQR	18	31250	6000	201576,5766
Batas_Bawah	-25	31125	-9000	-284346,8469
Batas_Atas	47	93875	15000	521959,4595
Jumlah_Outliers	3672	15	2461	2906

```

> # Menggabungkan hasil deteksi outliers dari semua kolom
    menjadi satu data frame
> outliers_summary_df <- do.call(rbind, outliers_summary)
> # Menyimpan hasil deteksi outliers ke file Excel

    Hasil deteksi outliers telah disimpan ke file Excel.
-----
> # Menghapus baris yang terkena outliers

    Jumlah baris terkena outliers: 5935

    Jumlah baris setelah menghapus outliers: 15321

> # Menyimpan data setelah pembersihan
> cleaned_file_path <- "C:/Users/62857/OneDrive/Documents
    /Skripsi/data_Proses.xlsx"
Data setelah pembersihan telah disimpan ke file Excel.
> cat("-----\n")
-----

```

IQR (*Interquartile Range*) menurut Kurniawan (2019), adalah ukuran yang digunakan untuk mengukur sebaran data dengan mengabaikan *outliers*. Hasil pada Gambar 4.3 menunjukkan proses deteksi dan penghapusan *outliers* menggunakan metode IQR pada beberapa atribut kolom numerik seperti Jumlah Produk, Harga, Total Diskon, dan Pendapatan Bersih. Setelah menghitung statistik dasar untuk setiap atribut, batas bawah dan atas untuk mendeteksi *outliers* ditentukan. Berdasarkan hasil analisis, ditemukan 3.672 data *outliers* pada atribut kolom Jumlah Produk dan 2.906 data *outliers* pada atribut kolom Pendapatan Bersih. Dengan demikian, sebanyak 5.935 data transaksi yang mengandung *outliers* dihapus, sehingga total data transaksi berkurang dari 21.191 menjadi 15.321. Data yang telah dibersihkan disimpan dalam file *dataProses.xlsx* untuk analisis lebih lanjut sehingga memastikan hasil yang akurat dan bebas.

Nomor	Nomor Dokumen	Tanggal Dokumen	Status Dokumen	Nomor Sales Order	ID Pelanggan	Nama Pelanggan	Alamat	ID Produk	Nama Produk	Jumlah Produk	Satuan Produk	Harga	Total Diskon	Pendapatan Bersih
1	2-1662	#####	Applied	2-160115	52-00239	TORE - W	Arang Sel	145141	LYCHEE	2	BOX	42000	0	75675.7
2	2-1662	#####	Applied	2-160133	52-00172	AGY - TOK	arang Ut	745601	GLN	0	BOTOL	14000	0	0
3	2-1662	#####	Applied	2-160133	52-00172	AGY - TOK	arang Ut	293101	TISS	0	EMBAF	0	0	0
4	2-1662	#####	Applied	2-160155	52-00015	S - BUDI J	arang Ba	016993	5GLN	-6	BOTOL	-30000	0	0
5	2-1662	#####	Applied	2-160155	52-00231	AHS FENI	Mijen	745591	JA GLN	30	BOTOL	15750	0	425676
6	2-1662	#####	Applied	2-160155	52-00231	AHS FENI	Mijen	193101	JA TIS	30	EMBAF	0	0	0
7	2-1662	#####	Applied	2-160155	52-00203	VIDAH - T	embalang	745591	JA GLN	20	BOTOL	17000	0	306306
8	2-1662	#####	Applied	2-160155	52-00203	VIDAH - T	embalang	193101	JA TIS	20	EMBAF	0	0	0
9	2-1662	#####	Applied	2-160155	52-00203	VIDAH - T	embalang	016993	5GLN	1	BOTOL	30000	0	0
10	2-1662	#####	Applied	2-160149	52-00015	S SANTOS	Gunungpa	293101	TISS	30	EMBAF	0	0	0
11	2-1662	#####	Applied	2-160149	52-00015	S SANTOS	Gunungpa	016993	5GLN	28	BOTOL	30000	0	0
12	2-1663	#####	Applied	2-160177	52-00239	KUTUT - T	Candisari	745591	JA GLN	20	BOTOL	17000	0	306306
13	2-1663	#####	Applied	2-160177	52-00239	KUTUT - T	Candisari	193101	JA TIS	20	EMBAF	0	0	0
14	2-1663	#####	Applied	2-160177	52-00106	SRI - TOK	Candisari	745591	JA GLN	30	BOTOL	17000	0	459459
15	2-1663	#####	Applied	2-160177	52-00106	SRI - TOK	Candisari	193101	JA TIS	30	EMBAF	0	0	0
16	2-1663	#####	Applied	2-160176	52-00217	KO - RAN	Gayamsar	745591	JA GLN	10	BOTOL	17000	0	153153
17	2-1663	#####	Applied	2-160176	52-00217	KO - RAN	Gayamsar	193101	JA TIS	10	EMBAF	0	0	0
18	2-1663	#####	Applied	2-160176	52-00209	O KO BARO	Candisari	745591	JA GLN	20	BOTOL	17000	0	306306
19	2-1663	#####	Applied	2-160176	52-00209	O KO BARO	Candisari	193101	JA TIS	20	EMBAF	0	0	0
20	2-1663	#####	Applied	2-160176	52-00209	TOKO DEP	Candisari	193101	JA TIS	45	EMBAF	0	0	0
21	2-1663	#####	Applied	2-160175	52-00173	TUNG - T	Candisari	745591	JA GLN	20	BOTOL	17000	0	306306
22	2-1663	#####	Applied	2-160175	52-00173	TUNG - T	Candisari	193101	JA TIS	20	EMBAF	0	0	0
23	2-1663	#####	Applied	2-160175	52-00167	GGENG - T	Candisari	745591	JA GLN	15	BOTOL	17000	0	229730
15314	2-1677	#####	Applied	2-161785	52-00236	RI - WARU	Tugu	745611	JA 600	1	BOX	47000	0	42342.3
15315	2-1677	#####	Applied	2-161785	52-00236	RI - WARU	Tugu	745531	JA 150	1	BOX	52500	0	47297.3
15316	2-1677	#####	Applied	2-161785	52-00236	RI - WARU	Tugu	157095	T 5501	1	BOX	32500	0	29279.3
15317	2-1677	#####	Applied	2-161785	52-00114	EH-2D - T	arang Ba	745611	JA 600	2	BOX	47000	6000	79279.3
15318	2-1677	#####	Applied	2-161785	52-00135	KARTI - A	arang Ba	745611	JA 600	2	BOX	47000	6000	79279.3
15319	2-1677	#####	Applied	2-161785	52-00135	KARTI - A	arang Ba	157095	T 5501	1	BOX	32500	2500	27027
15320	2-1677	#####	Applied	2-161785	52-00135	KARTI - A	arang Ba	145141	LYCHEE	1	BOX	42000	3000	35135.1
15321	2-1677	#####	Applied	2-161785	52-00039	GUNG - W	arang Ba	157095	T 5501	2	BOX	32500	5000	54054.1

Gambar 4.4. Data Proses

C. Persiapan Data (*Data Preparation*)

Pada tahap Persiapan Data, langkah-langkah krusial diambil untuk memastikan kualitas data sebelum analisis dilakukan. Langkah pertama adalah deskripsi data yang berasal dari fase Pemahaman Data untuk mengetahui karakteristik dan struktur data. Selanjutnya, proses *reduksi dan integrasi* data bertujuan untuk menyederhanakan *dataset* dengan mengurangi fitur yang tidak relevan serta menggabungkan data dari berbagai sumber atau kolom. Terakhir, pembersihan data dilakukan untuk memperbaiki data tidak relevan termasuk penanganan nilai hilang, duplikasi, dan *outliers* sehingga data siap untuk analisis. Berikut ini adalah penjelasan langkah-langkah tersebut secara detail dengan menggunakan *RStudio 2024.09.0+375* dengan *script* Lampiran 4, yaitu:

1. Deskripsi Data

Deskripsi data yang diperoleh dari fase Pemahaman Data dalam Gambar 4.4 memberikan proses pemahaman mendalam yang jelas mengenai karakteristik dan struktur data yang akan dianalisis. Hasil proses deskripsi data yang dilakukan menggunakan *RStudio 2024.09.0+375* ditampilkan sebagai berikut:

Tabel 4.7. Hasil Proses Deskripsi Data

Informasi Data	Jumlah Data
Jumlah Kolom	15
Jumlah Baris	15321

Tabel 4.7 menampilkan hasil dari fase Pemahaman Data dalam Gambar 4.4. Hasil proses ini menunjukkan bahwa

data memiliki 15 atribut kolom dengan total 15.321 data transaksi. Setiap atribut dilengkapi dengan informasi mengenai tipe data, jumlah nilai unik, dan rincian lainnya sebagai berikut:

Tabel 4.8. Hasil Proses Deskripsi Atribut Kolom

Atribut Kolom	Jumlah Nilai Unik	Contoh Nilai Unik
Nomor	15321	1,2,3,4,5
Nomor Dokumen	8334	852-1662303, dll
Tanggal Dokumen	26	2023-11-01, dll
Status Dokumen	2	Applied, Void
Nomor Sales Order	8322	852-1601167, dll
ID Pelanggan	2797	852-0023925, dll
Nama Pelanggan	2601	SN STORE - WARUNG MAGY - TOKO, dll
Alamat	16	Semarang Utara, dll
ID Produk	20	145141, dll
Nama Produk	18	VIT 1500ML, AQUA 220ML MIZONE ACTIV LYCHEE LEMON 500ML, dll
Jumlah Produk	67	2,100,200,74, dll
Satuan Produk	4	BOX, BOTOL, Lembar, Unit
Harga Produk	44	, 14000, 0, dll
Total Diskon Produk	44	0, 45000, dll
DPP (Pendapatan Bersih)	302	75675.6757, 0, dll

Tabel 4.7 menunjukkan hasil proses deskripsi yang mengonfirmasi bahwa 15 atribut kolom telah bersih dari nilai NA (*Missing Value*) dan nilai negatif yang terlihat, namun tetap memerlukan analisis lebih lanjut. Pada atribut

Status Dokumen, nilai *Void* masih tercantum sehingga perlu dihapus pada langkah reduksi data berikutnya.

2. Reduksi Data

Reduksi data adalah proses menghapus fitur yang tidak relevan atau berkontribusi rendah pada analisis, sehingga mengurangi kompleksitas, risiko *overfitting*, dan waktu pemrosesan. Hasilnya adalah *dataset* yang lebih bersih, ringkas, dan fokus, mendukung pengambilan keputusan yang lebih baik. Dengan demikian, analisis hanya mencakup informasi penting, memungkinkan evaluasi yang lebih mendalam terhadap karakteristik dan performa produk, serta meningkatkan efisiensi dan akurasi analisis. Hasil proses *reduksi data* menggunakan *RStudio 2024.09.0+375* adalah sebagai berikut:

```
> # PROSES REDUKSI DATA
> # Menghapus kolom selain yang diperlukan
> data <- data %>% select("Nomor", "Status Dokumen", "Alamat",
                           "Nama Produk", "Jumlah Produk")

Kolom telah dihapus sehingga hanya tersisa
kolom yang diperlukan. Proses telah berhasil.
-----

> # Menghapus baris dengan 'Void' di kolom 'Status Dokumen'
> data <- data %>% filter(`Status Dokumen` != "Void")
> removed_void_rows <- initial_rows - nrow(data)
> cat(sprintf("\nJumlah baris dengan 'Void' di 'Status Dokumen'
yang dihapus: %d\n", removed_void_rows))

Jumlah baris dengan 'Void' di 'Status Dokumen' yang dihapus: 52
> cat("Proses penghapusan baris dengan nilai 'Void' telah berhasil.
\n")

Proses penghapusan baris dengan nilai 'Void'
telah berhasil.
-----
```

Gambar 4.5. Hasil Proses Reduksi Atribut Kolom

Gambar 4.5 menunjukkan bahwa atribut kolom selain Nomor, Status Dokumen, Alamat, Nama Produk, dan Jumlah

Produk telah dihapus dari data sehingga hanya atribut kolom yang diperlukan yang tersisa yang sangat penting bagi penelitian ini agar sesuai dengan tujuan. Proses penghapusan atribut kolom berhasil dilakukan dengan memastikan data menjadi lebih fokus dan sesuai dengan kebutuhan analisis. *Reduksi data* pada atribut kolom Status Dokumen dilakukan dengan menghapus nilai-nilai yang tidak relevan, seperti *Void* untuk memastikan hanya data yang valid yang tersisa untuk dianalisis. Proses ini mencakup penghapusan baris dengan nilai *Void* menghitung jumlah baris data transaksi yang dihapus, dan menampilkan pesan konfirmasi bahwa penghapusan telah berhasil dengan total 52 baris data transaksi dihapus dari data berdasarkan nilai *Void*.

```
>
> # Menghapus jenis nilai tertentu di kolom 'Nama Produk'
> unwanted_values <- c("VIT TISSUE", "AQUA 5GLN BTL", "VIT GLN ISI",
  "AQUA GLN ISI", "AQUA TISSUE", "DISPENSER BAJA PUTIH 2 KERAN",
  "VIT GLN BTL", "AQUA 750ML", "AQUA 5GLN BTL - NO VALUE")

  Jumlah baris dengan jenis nilai tertentu yang dihapus: 5740
> cat("Proses penghapusan jenis nilai tertentu pada
  'Nama Produk' telah berhasil.\n")

  Proses penghapusan jenis nilai tertentu pada
  'Nama Produk' telah berhasil.
```

Gambar 4.6. Hasil Proses Reduksi Atribut Kolom Nama Produk

Reduksi data pada kolom Nama Produk difokuskan untuk mempertahankan hanya produk yang relevan, seperti Produk AMDK Jenis Gelas dan Botol dalam berbagai ukuran sesuai dengan tahapan pemahaman bisnis. Gambar 4.6 menunjukkan bahwa beberapa nilai tertentu, seperti VIT TISSUE, AQUA 5GLN BTL, dan lainnya telah dihapus dari

atribut kolom Nama Produk dengan total 5.740 baris data transaksi yang dihapus. Proses ini bertujuan untuk membersihkan data dari produk yang tidak relevan sehingga analisis menjadi lebih fokus pada produk yang signifikan dan sesuai dengan tujuan penelitian.

3. Pemeriksaan Kualitas Data

Pembersihan data dalam penelitian merupakan langkah penting sebelum tahapan pemodelan, karena dapat mengurangi bias dan kesalahan serta mendukung pengambilan keputusan yang lebih tepat dan valid. Proses ini mencakup identifikasi terhadap *missing values*, data duplikat, dan *outliers*. Pembersihan data dilakukan menggunakan *RStudio 2024.09.0+375* dengan langkah-langkah sebagai berikut:

3.1. Pemeriksaan dan Pembersihan nilai yang hilang (*missing values*)

Proses pemeriksaan dan penghapusan nilai yang hilang (*missing values*) merupakan langkah penting dalam tahapan persiapan data. Proses ini serupa dengan langkah pada tahap pemahaman data. Dalam tahap ini, data akan diperiksa untuk mengidentifikasi nilai yang hilang, kemudian dilakukan tindakan berupa penghapusan baris atau kolom terpengaruh, atau menggantinya dengan nilai estimasi yang sesuai, seperti *mean*, *median*, atau *modus*. Tujuan dari proses ini adalah untuk menghasilkan analisis yang akurat dan dapat diandalkan. Berikut ini ditampilkan hasil proses pemeriksaan dan pembersihan nilai yang hilang:

```

> # -----
> # 3. Pemeriksaan Kualitas
> # Memeriksa missing values
> cat("\nPemeriksaan Kualitas Data:\n")

Pemeriksaan Kualitas Data:

> missing_values <- sapply(data, function(x) sum(is.na(x)))
# Menghitung jumlah missing values per kolom
>
> if (any(missing_values > 0)) {
+   cat("Terdapat missing values dalam data:\n")
+   for (i in 1:length(missing_values)) {
+     if (missing_values[i] > 0) {
+       cat(sprintf(" Kolom: %s, Jumlah missing values: %d\n",
+         colnames(data)[i], missing_values[i]))
+     }
+   }
+ } else {
+   cat("Tidak ada missing values dalam data.\n")
+ }

Tidak ada missing values dalam data.

> cat("-----\n")
-----

```

Gambar 4.7. Hasil Proses pemeriksaan dan penghapusan nilai yang hilang (*missing values*)

Gambar 4.7 menunjukkan proses pendeteksian dan penghapusan *missing value* yang menghasilkan data bebas dari *missing value* karena tidak ditemukan adanya *missing values*.

3.2. Pemeriksaan dan Pembersihan Data Duplikat

Deteksi dan pengecekan data duplikat merupakan langkah penting untuk memastikan keunikan dan akurasi setiap entri dalam data. Dengan mengidentifikasi dan menghapus duplikasi, integritas data dapat terjaga sehingga analisis dapat dilakukan berdasarkan data yang valid dan tidak terdistorsi.

Berikut ini adalah hasil dari proses pemeriksaan dan pembersihan data duplikat.

```
> # Memeriksa data duplikat
> cat("\nMemeriksa Data Duplikat:\n")

Memeriksa Data Duplikat:

> duplicate_rows <- duplicated(data) | duplicated(data,
  fromLast = TRUE) # Memeriksa duplikat
> num_duplicates <- sum(duplicate_rows)
>
> if (num_duplicates > 0) {
+   cat(sprintf("Terdapat %d baris duplikat dalam data.\n",
  num_duplicates))
+ } else {
+   cat("Tidak ada baris duplikat dalam data.\n")
+ }

Tidak ada baris duplikat dalam data.

> cat("-----\n")
-----
```

Gambar 4.8. Hasil Proses pemeriksaan dan penghapusan data duplikat

Gambar 4.8 menunjukkan hasil proses deteksi dan penghapusan data duplikat. Dari hasil proses ini, tidak ditemukan data duplikat sehingga tidak ada data transaksi yang perlu dihapus.

3.3. Pemeriksaan dan Pembersihan Outliers

Proses deteksi dan penghapusan data *outliers* penting untuk memastikan keandalan analisis yang dilakukan menggunakan metode IQR (*Interquartile Range*). *Outliers* adalah nilai yang sangat berbeda dari data lainnya yang bisa terjadi akibat kesalahan pengukuran, kesalahan entri, atau variasi ekstrem. Menurut Kurniawan (2019), IQR mengukur sebaran

data tanpa memperhitungkan *outliers*. Proses ini bertujuan untuk mengidentifikasi nilai penjualan yang menyimpang, termasuk nilai negatif dengan menggunakan metode IQR. Berikut ini adalah hasil pemeriksaan dan pembersihan data *outliers*:

```
> # Memeriksa dan Menghapus Data Outliers serta
  Menghapus Angka Negatif:
```

**Memeriksa dan Menghapus Data Outliers serta
Menghapus Angka Negatif:**

```
> numeric_columns <- data %>% select_if(is.numeric)
  # Memilih kolom numerik
>
> for (col_name in colnames(numeric_columns)) {
+   # Menghitung batas bawah dan atas untuk deteksi outlier
+   q1 <- quantile(numeric_columns[[col_name]], 0.25,
+     na.rm = TRUE)
+   q3 <- quantile(numeric_columns[[col_name]], 0.75,
+     na.rm = TRUE)
+   iqr <- IQR(numeric_columns[[col_name]], na.rm = TRUE)
+   lower_bound <- q1 - 1.5 * iqr
+   upper_bound <- q3 + 1.5 * iqr
+   # Menemukan posisi outlier dan nilai negatif
+   outliers_or_negatives <-
+     (numeric_columns[[col_name]] < lower_bound |
+      numeric_columns[[col_name]] > upper_bound |
+      numeric_columns[[col_name]] < 0)+
+   # Menghapus baris dengan outliers dan nilai negatif
+     jika ditemukan
+   if (num_outliers_or_negatives > 0) {
+     data <- data[!outliers_or_negatives, ]}
```

**Kolom: Jumlah Produk,
835 outliers/angka negatif telah dihapus.**

Gambar 4.9. Hasil Proses pemeriksaan dan penghapusan *Outliers*

Gambar 4.9 menunjukkan hasil proses deteksi dan penghapusan *outliers*. Proses ini mengidentifikasi dan menghapus sebanyak 835 *outliers*. Hal ini membuktikan pentingnya pemeriksaan ulang

kualitas data untuk menjaga integritas data sehingga analisis dapat dilakukan berdasarkan data yang valid. Setelah pemeriksaan kualitas data selesai, data akan ditampilkan sebagai berikut:

```
> # Menampilkan ringkasan data akhir
> cat("\nRingkasan Data Akhir:\n")

Ringkasan Data Akhir:
> final_num_rows <- nrow(data)
> final_num_cols <- ncol(data)
> cat(sprintf("Jumlah kolom: %d\n", final_num_cols))
Jumlah kolom: 5
> cat(sprintf("Jumlah baris: %d\n", final_num_rows))
Jumlah baris: 8694
> cat("-----\n")
-----
> View(data)
> cat("-----\n")
-----
```

Gambar 4.10. Identifikasi Data Setelah Pemeriksaan Kualitas Data

Gambar 4.10 menunjukkan hasil identifikasi data setelah pemeriksaan kualitas yang mengungkapkan bahwa data memiliki 8.694 transaksi dengan 5 kolom atribut yang sesuai dengan permasalahan dan tujuan penelitian, yaitu Nomor, Alamat, Status Dokumen, Nama Produk, dan Jumlah Produk. Selanjutnya, data akan disesuaikan oleh penulis melalui tahap Integrasi Data.

4. Intergrasi Data

Integrasi data adalah proses penggabungan informasi dari berbagai sumber untuk membentuk satu kumpulan data yang konsisten dan siap dianalisis. Proses ini menggabungkan data yang terpisah menjadi satu kesatuan sehingga mengurangi ketidakonsistenan dan pengulangan.

Dengan demikian, *integrasi data* memungkinkan analisis yang lebih efisien dan akurat karena semua elemen penting disatukan dalam satu struktur. Dalam penelitian ini, atribut kolom yang digunakan untuk *integrasi data* mencakup atribut kolom Alamat, kolom Nama Produk, dan kolom Jumlah Produk sebagai data numerik. Berikut adalah hasil *integrasi data* menggunakan *pivot data* dengan *RStudio 2024.09.0+375* dan *Microsoft Excel 2013*, yaitu:

```
> # 5. Integrasi Data
> # Membuat data frame baru dengan baris sebagai 'Alamat'
  dan kolom sebagai 'Nama Produk'
> pivot_data <- data %>%
+   group_by('Alamat', 'Nama Produk') %>%
+   summarise(Jumlah = sum('Jumlah Produk')) %>%
+   spread(key = 'Nama Produk', value = Jumlah, fill = 0)
'summarise()' has grouped output by 'Alamat'. You can
override using the '.groups' argument.
>
> # Menampilkan hasil data dalam jendela terpisah
> cat("\nHasil data setelah pivot:\n")
```

Hasil data setelah pivot:

```
> print(head(pivot_data, 20))
```

Alamat	AQUA 1500ML	AQUA 220ML	AQUA 330ML	AQUA 600ML	MIZONE ACTIV LYCHEE LEMON 500ML	VIT 1500ML	VIT 200ML	VIT 330ML	VIT 550ML
Banyumanik	45	2	7	30	9	8	27	4	8
Candisari	277	94	112	311	56	68	274	31	107
Gajahmungkur	403	212	142	498	150	112	439	65	176
Gayamsari	8	10	2	19	0	3	15	0	4
Genuk	31	18	23	43	19	8	63	1	5
Gunungpati	174	82	23	180	95	48	236	19	41
Mijen	501	216	229	507	182	75	489	74	142
Ngaliyan	962	376	227	1092	313	182	892	129	340
Pedurungan	9	1	0	20	9	2	10	0	1
Semarang Barat	839	308	247	1140	212	235	661	111	374
Semarang Selatan	256	115	196	393	95	108	249	62	133
Semarang Tengah	114	42	50	187	52	53	89	30	110
Semarang Timur	49	14	5	45	8	5	33	5	5
Semarang Utara	294	111	92	361	100	104	243	31	164
Tembalang	106	48	35	140	45	35	100	23	29
Tugu	190	62	46	254	49	31	179	29	25

```
> # Menyimpan data setelah pivot ke file Excel
> cleaned_file_path <- "C:/Users/62857/OneDrive/Documents
/FILE program RSTUDIO dan MATLAB/Persiapan Data (Data Preparation)
> write_xlsx(pivot_data, path = cleaned_file_path)
```

Gambar 4.11. Hasil Proses *Intergrasi Data*

Gambar 4.11 menunjukkan pembuatan data baru melalui *pivot* dengan baris data yang berasal dari atribut kolom Alamat dan atribut kolom yang berasal dari Nama Produk. Kolom Jumlah Produk yang berisi data numerik digunakan untuk perhitungan dalam penelitian ini. Hasil *pivot data* ini terdiri dari 16 baris data dan 10 atribut kolom, yaitu 1 atribut kolom bertipe *Character* dan 9 atribut kolom bertipe *Numeric*. *Pivot data* ini akan digunakan pada fase selanjutnya sebagai data penelitian.

D. Metode Silhoutte Coefficient

Setelah persiapan data, metode *Silhouette Coefficient* diterapkan pada tabel data penelitian dari Gambar 4.10 untuk menilai kualitas kluster dengan memilih jumlah kluster yang optimal. Proses ini melibatkan perhitungan jarak antar objek dalam kluster (**a**) dan antar kluster (**b**), diikuti dengan perhitungan nilai *Silhouette Coefficient* (**(s(o))**) untuk setiap objek yang berkisar antara -1 hingga 1. Perhitungan nilai **S** digunakan untuk menilai kualitas seluruh kluster. Nilai di atas 0,7 menandakan struktur yang kuat, sementara di bawah 0,25 menunjukkan struktur yang tidak jelas. Metode ini memungkinkan evaluasi yang sederhana dan penerapan dengan berbagai algoritma klustering untuk menentukan jumlah kluster optimal sebelum menggunakan *K-Means*. Berikut ini contoh penjelasan proses perhitungan *Silhouette Coefficient* menggunakan *MATLAB 2013* dengan jumlah kluster yang diuji berkisar antara 2 hingga 5 dan hasilnya akan ditampilkan sebagai contoh jumlah kluster adalah 2 sebagai berikut:

1. Menghitung Nilai Rata-rata Jarak Intraklaster (a)

Proses metode *Silhouette Coefficient* menggunakan data penelitian dari Gambar 4.10 disajikan dengan bantuan *software MATLAB 2013*, sedangkan *script*-nya ditampilkan pada Lampiran 5. Langkah pertama dalam metode ini adalah menghitung jarak antara objek data dan data lainnya dalam klaster yang sama (a). Data penelitian ini telah mengetahui klaster setiap objek menggunakan *K-Means* pada *software MATLAB 2013*. Sebagai contoh untuk jumlah klaster sebanyak 2, hasil klasterisasi yang diperoleh menunjukkan setiap objek masuk ke dalam klaster tertentu sebagai berikut:

Evaluasi untuk jumlah klaster: 2
Hasil akhir clustering:

Data		Klaster	
Banyumanik	Data 1	Klaster 2	
Candisari	Data 2	Klaster 2	
Gajahmungkur	Data 3	Klaster 1	
Gayamsari	Data 4	Klaster 2	
Genuk	Data 5	Klaster 2	
Gunungpati	Data 6	Klaster 2	
Mijen	Data 7	Klaster 1	
Ngaliyan	Data 8	Klaster 1	
Pedurungan	Data 9	Klaster 2	
Semarang Barat	Data 10	Klaster 1	
Semarang Selatan	Data 11	Klaster 2	
Semarang Tengah	Data 12	Klaster 2	
Semarang Timur	Data 13	Klaster 2	
Semarang Utara	Data 14	Klaster 2	
Tembalang	Data 15	Klaster 2	
Tugu	Data 16	Klaster 2	

Gambar 4.12. Hasil Proses Evaluasi Data Jumlah Klaster 2

Setelah data dikelompokkan ke dalam klaster, dilakukan perhitungan nilai jarak antara objek data dengan data lain dalam klaster yang sama (a) yang hasilnya ditampilkan secara manual. Berikut ini adalah salah satu contoh perhitungan manual untuk menunjukkan proses perhitungan jarak *Euclidean* antara Banyumanik dan setiap objek lain dalam Klaster 1:

dist(Banyumanik, Candisari) =

$$= \sqrt{(45 - 277)^2 + (2 - 94)^2 + (7 - 112)^2 + (30 - 311)^2 + (9 - 56)^2} \\ \sqrt{+(8 - 68)^2 + (27 - 274)^2 + (4 - 31)^2 + (8 - 107)^2} \\ = \sqrt{229622} = 479, 189$$

dist(Banyumanik, Gayamsari) =

$$= \sqrt{(45 - 8)^2 + (2 - 10)^2 + (7 - 2)^2 + (30 - 19)^2 + (9 - 0)^2} \\ \sqrt{+(8 - 3)^2 + (27 - 15)^2 + (4 - 0)^2 + (8 - 4)^2} \\ = \sqrt{1861} = 43, 139$$

dist(Banyumanik, Genuk) =

$$= \sqrt{(45 - 31)^2 + (2 - 18)^2 + (7 - 23)^2 + (30 - 43)^2 + (9 - 19)^2} \\ \sqrt{+(8 - 8)^2 + (27 - 63)^2 + (4 - 1)^2 + (8 - 5)^2} \\ = \sqrt{2291} = 47, 864$$

dist(Banyumanik, Gunungpati) =

$$= \sqrt{(45 - 174)^2 + (2 - 82)^2 + (7 - 23)^2 + (30 - 180)^2 + (9 - 95)^2} \\ \sqrt{+(8 - 48)^2 + (27 - 236)^2 + (4 - 19)^2 + (8 - 41)^2} \\ = \sqrt{99788} = 315, 892$$

dist(Banyumanik, Pedurungan) =

$$= \sqrt{(45 - 9)^2 + (2 - 1)^2 + (7 - 0)^2 + (30 - 20)^2 + (9 - 9)^2} \\ \sqrt{+(8 - 2)^2 + (27 - 10)^2 + (4 - 0)^2 + (8 - 1)^2} \\ = \sqrt{1836} = 42, 849$$

dist(Banyumanik, Semarang Selatan) =

$$= \sqrt{(45 - 256)^2 + (2 - 115)^2 + (7 - 196)^2 + (30 - 393)^2 + (9 - 95)^2} \\ \sqrt{+(8 - 108)^2 + (27 - 249)^2 + (4 - 62)^2 + (8 - 133)^2} \\ = \sqrt{310449} = 557,180$$

dist(Banyumanik, Semarang Tengah) =

$$= \sqrt{(45 - 114)^2 + (2 - 42)^2 + (7 - 50)^2 + (30 - 187)^2 + (9 - 52)^2} \\ \sqrt{+(8 - 53)^2 + (27 - 89)^2 + (4 - 30)^2 + (8 - 110)^2} \\ = \sqrt{51657} = 227,282$$

dist(Banyumanik, Semarang Timur) =

$$= \sqrt{((45 - 49)^2 + (2 - 14)^2 + (7 - 5)^2 + (30 - 45)^2 + (9 - 8)^2} \\ \sqrt{+(8 - 5)^2 + (27 - 33)^2 + (4 - 5)^2 + (8 - 5)^2} \\ = \sqrt{445} = 21,095$$

dist(Banyumanik, Semarang Utara) =

$$= \sqrt{(45 - 294)^2 + (2 - 111)^2 + (7 - 92)^2 + (30 - 361)^2 + (9 - 100)^2} \\ \sqrt{+(8 - 104)^2 + (27 - 243)^2 + (4 - 31)^2 + (8 - 164)^2} \\ = \sqrt{279886} = 529,043$$

dist(Banyumanik, Tembalang) =

$$= \sqrt{(45 - 106)^2 + (2 - 48)^2 + (7 - 35)^2 + (30 - 140)^2 + (9 - 45)^2} \\ \sqrt{+(8 - 35)^2 + (27 - 100)^2 + (4 - 23)^2 + (8 - 29)^2} \\ = \sqrt{26877} = 163,942$$

dist(Banyumanik, Tugu) =

$$= \sqrt{(45 - 190)^2 + (2 - 62)^2 + (7 - 46)^2 + (30 - 254)^2 + (9 - 49)^2} \\ \sqrt{+(8 - 31)^2 + (27 - 179)^2 + (4 - 29)^2 + (8 - 25)^2} \\ = \sqrt{102469} = 320,108$$

Setelah perhitungan manual jarak *Euclidean* antara Banyumanik dan setiap objek lain dalam Klaster 1 selesai, langkah berikutnya adalah menghitung nilai **a** menggunakan rumus pada persamaan 2.4 dengan hasil sebagai berikut:

$$\mathbf{a}(\text{Banyumanik}) = (479,189 + 43,139 + 47,864 + 315,892 + 42,849 + 557,180 + 227,282 + 21,095 + 529,043 + 163,942 + 320,108) / (12 - 1)$$

$$\mathbf{a}(\text{Banyumanik}) = 249,780$$

Proses manual ini akan diulang hingga semua objek selesai dihitung. Hasil perhitungan manual ini akan dijadikan referensi untuk hasil program yang menggunakan *MATLAB 2013*. Detail *script* lebih lanjut dapat dilihat pada Lampiran 5.

2. Menghitung Nilai Rata-rata Jarak Antar kluster Terdekat (b)

Setelah nilai **a** diperoleh, dilanjutkan dengan perhitungan jarak *Euclidean* ke semua titik data lain dalam kluster berbeda, kemudian rata-rata nilai tersebut diambil sebagai nilai **b**. Langkah ini memberikan informasi tentang kedekatan objek dengan kluster terdekat di luar kluster asalnya. Proses perhitungan **b** dilakukan menggunakan persamaan 2.5 dengan bantuan *MATLAB 2013*. Detail *script* dapat dilihat pada Lampiran 5. Berikut adalah salah satu contoh perhitungan manual jarak *Euclidean* antara *Banyumanik* dan setiap objek lain dalam kluster lain, dengan hasil sebagai berikut:

$$\begin{aligned} \text{dist}(\text{Banyumanik}, \text{Gajahmungkur}) &= \\ &= \sqrt{(45 - 403)^2 + (2 - 212)^2 + (7 - 142)^2 + (30 - 498)^2 +} \\ &\quad \sqrt{(9 - 150)^2 + (8 - 112)^2 + (27 - 439)^2 + (4 - 65)^2 + (8 - 176)^2} \\ &= \sqrt{641899} = 801,186 \end{aligned}$$

dist(Banyumanik, Mijen) =

$$= \sqrt{(45 - 501)^2 + (2 - 216)^2 + (7 - 229)^2 + (30 - 507)^2 + \\ \sqrt{(9 - 182)^2 + (8 - 75)^2 + (27 - 489)^2 + (4 - 74)^2 + (8 - 142)^2} \\ = \sqrt{801263} = 895,133$$

dist(Banyumanik, Ngaliyan) =

$$= \sqrt{(45 - 962)^2 + (2 - 376)^2 + (7 - 227)^2 + (30 - 1092)^2 + \\ \sqrt{(9 - 313)^2 + (8 - 182)^2 + (27 - 892)^2 + (4 - 130)^2 + (8 - 340)^2} \\ = \sqrt{3154026} = 1775,958$$

dist(Banyumanik, Semarang Barat) =

$$= \sqrt{(45 - 839)^2 + (2 - 308)^2 + (7 - 247)^2 + (30 - 1140)^2 + \\ \sqrt{(9 - 212)^2 + (8 - 235)^2 + (27 - 661)^2 + (4 - 111)^2 + (8 - 374)^2} \\ = \sqrt{2653871} = 1629,071$$

Setelah perhitungan manual jarak *Euclidean* antara Banyumanik dan setiap objek lain dalam Klaster 2 selesai, langkah berikutnya adalah menghitung nilai **b** menggunakan rumus pada persamaan 2.5 dengan hasil sebagai berikut:

• **Nilai b Banyumanik**

$$\mathbf{b} \text{ (Banyumanik)} = \frac{(801,186 + 895,133 + 1775,958 + 1629,071)}{4} \\ = 1275,337$$

Proses manual ini akan diulang hingga semua objek selesai dihitung. Hasil perhitungan manual ini akan dijadikan referensi untuk hasil program yang menggunakan *MATLAB 2013*. Detail *script* lebih lanjut terdapat pada Lampiran 5.

3. Menghitung nilai *Silhouette* untuk setiap objek (**s(o)**)

Setelah nilai **a** dan **b** diperoleh, perhitungan *Silhouette Coefficient* untuk setiap titik data (**s(o)**) dihitung

menggunakan rumus pada Persamaan 2.6, di mana nilai *Silhouette Coefficient* berkisar antara -1 hingga 1. Nilai positif menunjukkan bahwa objek berada di klaster yang tepat dan semakin mendekati 1, semakin baik objek tersebut dikelompokkan. Sebaliknya, jika nilainya mendekati -1 ini menandakan bahwa objek mungkin lebih cocok berada di klaster lain. Proses perhitungan $s(o)$ dilakukan menggunakan persamaan 2.5 dengan bantuan *MATLAB 2013*. Detail *script* lebih lanjut dapat dilihat pada Lampiran 5. Berikut adalah salah satu contoh perhitungan manual untuk menghitung nilai *Silhouette* setiap objek:

$$s(o) = \frac{1275.337 - 249.780}{\max(249.780, 1275.337)} = \frac{1025.557}{1275.337} = 0,8041$$

Proses manual ini akan diulang untuk semua objek hingga selesai. Hasil perhitungan manual ini akan digunakan sebagai referensi untuk hasil program yang menggunakan *MATLAB 2013*. Detail *script* lebih lanjut dapat dilihat pada Lampiran 5.

4. Menghitung Nilai Rata-rata *Silhouette Coefficient* (S)

Langkah terakhir adalah menilai kualitas keseluruhan klaster dengan menghitung rata-rata nilai $s(o)$ untuk semua objek. Nilai S yang lebih dari 0,7 mengindikasikan struktur klaster yang kuat, sementara nilai di bawah 0,25 menunjukkan struktur yang kurang jelas. Proses perhitungan ini dilakukan menggunakan persamaan 2.7 dengan bantuan *MATLAB 2013*. Detail *script* lebih lanjut dapat dilihat pada Lampiran 5. Berikut adalah salah satu contoh perhitungan manual untuk menghitung nilai S :

$$\mathbf{S} = (0,8041 + 0,5850 + (-0,0983) + 0,7922 + 0,8070 + 0,7354 + 0,1136 + 0,5362 + 0,7911 + 0,5428 + 0,4708 + 0,7824 + 0,8080 + 0,5126 + 0,8068 + 0,7401)/16$$

$$\mathbf{S} = 9,7298/16$$

$$\mathbf{S} = 0,6081$$

Hasil perhitungan manual ini digunakan sebagai referensi untuk hasil program yang menggunakan *MATLAB 2013*. Detail *script* lebih lanjut dapat dilihat pada Lampiran 5. Berikut ini ditampilkan hasil program:

Evaluasi untuk jumlah kluster: 2
Detail perhitungan Silhouette Coefficient:

Data	a(i)	b(i)	s(i)
Banyumanik	249.7802	1275.3368	0.8041
Candisari	334.1704	805.3258	0.5850
Gajahmungkur	664.0227	598.7797	-0.0983
Gayamsari	271.8024	1308.2941	0.7922
Genuk	241.3674	1250.7643	0.8070
Gunungpati	259.7316	981.4365	0.7354
Mijen	617.2988	696.4332	0.1136
Ngaliyan	728.7049	1571.0054	0.5362
Pedurungan	273.8185	1311.0508	0.7911
Semarang Barat	651.0192	1423.8996	0.5428
Semarang Selatan	400.7310	757.2394	0.4708
Semarang Tengah	233.4119	1072.8321	0.7824
Semarang Timur	241.9627	1260.2760	0.8080
Semarang Utara	373.0347	765.3229	0.5126
Tembalang	215.5477	1115.8263	0.8068
Tugu	251.2764	966.6489	0.7401

Nilai Silhouette Coefficient global: 0.6081

Gambar 4.13. Hasil Program Nilai *Silhouette Coefficient* untuk Jumlah Kluster 2

Hasil program nilai *silhouette* dengan jumlah kluster 2, yang ditampilkan dalam Gambar 4.13 menunjukkan nilai **a**, **b**, dan **s(o)** untuk setiap objek. Perhitungan manual sebelumnya untuk objek Banyumanik sesuai dengan hasil program yang

ditampilkan dalam Gambar 4.13. Hal ini membuktikan bahwa nilai **a**, **b**, dan **s(o)** untuk setiap objek dapat digunakan dalam penelitian ini. Nilai **S** yang ditampilkan dalam Gambar 4.13 menghasilkan interpretasi sebesar 0,6081, yang menunjukkan struktur klaster yang cukup kuat.

Setelah perhitungan untuk jumlah klaster 2 selesai, proses perhitungan akan dilanjutkan untuk jumlah klaster lainnya hingga mencapai 5 sesuai dengan langkah-langkah yang telah dijelaskan sebelumnya. Berikut ini adalah perhitungan nilai rata-rata *Silhouette Coefficient* (**S**) untuk berbagai jumlah klaster:

Nilai *Silhouette Coefficient* global untuk semua jumlah klaster:

Jumlah Klaster | Nilai SC Global

2		0.6081
3		0.5977
4		0.5408
5		0.5991

Jumlah klaster optimal: 2

Nilai *Silhouette Coefficient* global untuk jumlah klaster optimal: 0.6081

Interpretasi nilai *Silhouette Coefficient* global:

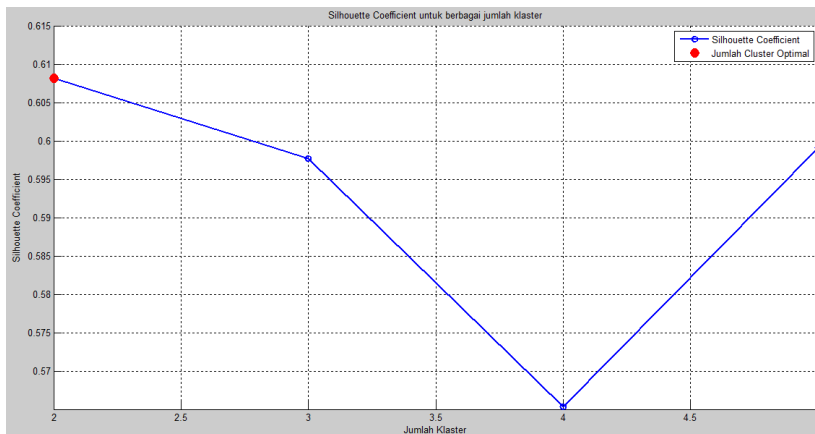
Struktur klaster yang standar.

Visualisasi *Silhouette Coefficient*:

Gambar 4.14. Hasil Program Nilai *Silhouette Coefficient* Untuk Beberapa Jumlah Klaster

Analisis terhadap nilai rata-rata *Silhouette Coefficient* (**S**) pada berbagai jumlah klaster, seperti yang ditunjukkan dalam Gambar 4.14, menunjukkan bahwa nilai tertinggi yaitu 0,6081 diperoleh dengan 2 klaster. Hal ini mengindikasikan

bahwa struktur kluster yang terbentuk lebih baik dibandingkan dengan jumlah kluster lainnya. Berikut ini akan ditampilkan tahap visualisasi sesuai dengan penjelasan dalam Gambar 4.14:



Gambar 4.15. Visual *Silhouette Coefficient* Untuk Beberapa Jumlah Kluster

Gambar 4.15 menghasilkan jumlah kluster yang paling optimal, yaitu sebanyak 2 kluster. Jumlah kluster ini telah dibuktikan melalui perhitungan manual maupun program komputasi. Dengan demikian, penulis memutuskan untuk menggunakan 2 kluster dalam penelitian ini dan akan diterapkan pada tahapan pemodelan selanjutnya.

E. Pemodelan (*Modeling*)

Setelah menetapkan jumlah kluster sebanyak 2 menggunakan metode *Silhouette Coefficient*, fase pemodelan dilanjutkan dengan teknik *data mining*, yaitu algoritma *K-Means*. Algoritma *K-Means*

membagi data ke dalam kluster, di mana objek dengan karakteristik serupa dikelompokkan dalam kluster yang sama, sedangkan objek dengan karakteristik berbeda dimasukkan ke kluster yang lain. Perhitungan ini dilakukan menggunakan *MATLAB 2013* dengan rincian *script* tersedia pada Lampiran 6. Berikut ini adalah contoh perhitungan manual untuk menghitung nilai *S*:

1. Inisialisasi

Langkah pertama dalam algoritma *K-Means* adalah inisialisasi. Proses ini dimulai dengan menentukan jumlah kluster yang dapat dilakukan secara acak atau menggunakan metode tertentu seperti *Silhouette Coefficient*. Selanjutnya, objek dipilih sebagai *centroid* awal. Berikut adalah data penelitian yang digunakan dalam hasil proses inisialisasi:

Alamat	AQUA 1500ML	AQUA 220ML	AQUA 330ML	AQUA 600ML	MIZONE ACTIV LYCHEE LEMON 500ML	VIT 1500ML	VIT 200ML	VIT 330ML	VIT 550ML
Banyumanik	45	2	7	30	9	8	27	4	8
Candisari	277	94	112	311	56	68	274	31	107
Gajahmungkur	403	212	142	498	150	112	439	65	176
Gayamsari	8	10	2	19	0	3	15	0	4
Genuk	31	18	23	43	19	8	63	1	5
Gunungpati	174	82	23	180	95	48	236	19	41
Mijen	501	216	229	507	182	75	489	74	142
Ngaliyan	962	376	227	1092	313	182	892	130	340
Pedurungan	9	1	0	20	9	2	10	0	1
Semarang Barat	839	308	247	1140	212	235	661	111	374
Semarang Selatan	256	115	196	393	95	108	249	62	133
Semarang Tengah	114	42	50	187	52	53	89	30	110
Semarang Timur	49	14	5	45	8	5	33	5	5
Semarang Utara	294	111	92	361	100	104	243	31	164
Tembalang	106	48	35	140	45	35	100	23	29
Tugu	190	62	46	254	49	31	179	29	25

Gambar 4.16. Data Penelitian *Pemodelan*

Setelah data penelitian dalam Gambar 4.16 diperoleh dari tahapan persiapan data, kemudian dilakukan proses

inisialisasi dengan memilih objek oleh penulis sebagai *centroid* awal. Berikut adalah objek yang terpilih, yaitu:

ITERASI 1										
Alamat	AQUA 1500ML	AQUA 220ML	AQUA 330ML	AQUA 600ML	MIZONE ACTIV LYCHEE LEMON 500ML	VIT 1500ML	VIT 200ML	VIT 330ML	VIT 550ML	Klaster
Semarang Utara	294	111	92	361	100	104	243	31	164	C1
Tugu	190	62	46	254	49	31	179	29	25	C2

Gambar 4.17. *centroid* Awal dalam Iterasi 1

centroid awal pada Gambar 4.17 dipilih dengan menetapkan objek Semarang Utara sebagai C1 dan objek Tugu sebagai C2. Dalam konteks inisialisasi *centroid* untuk algoritma *K-Means*, pemilihan objek Semarang Utara dan Tugu dimaksudkan untuk mewakili dua pola penjualan yang berbeda berdasarkan distribusi nilai produk. Dengan memilih Semarang Utara dan Tugu sebagai *centroid* awal, algoritma *K-Means* diharapkan dapat mengidentifikasi dua klaster berbeda yaitu satu untuk wilayah dengan penjualan tinggi dan satu lagi untuk wilayah dengan penjualan rendah.

2. Penugasan

Langkah ini dilakukan setelah pemilihan *centroid* awal, yaitu dengan menghitung jarak *Euclidean* antara setiap objek dan *centroid*. Pada iterasi pertama, proses perhitungan secara manual dan komputasi jarak pada C1 dan C2 ini dilakukan menggunakan rumus pada persamaan 2.1 dengan hasil sebagai berikut:

- **Proses Perhitungan Manual Jarak *Euclidean* Antara C1 dan Objek**

dist(Banyumanik, C1) =

$$\begin{aligned}
 &= \sqrt{(45 - 294)^2 + (2 - 111)^2 + (7 - 92)^2 + (30 - 361)^2 + (9 - 100)^2} \\
 &\quad \sqrt{+(8 - 104)^2 + (27 - 243)^2 + (4 - 31)^2 + (8 - 164)^2} \\
 &= \sqrt{279886} \\
 &= 529,04
 \end{aligned}$$

dist(Candisari, C1) =

$$\begin{aligned}
 &= \sqrt{(277 - 294)^2 + (94 - 111)^2 + (112 - 92)^2 + (311 - 361)^2 + (56} \\
 &\quad \sqrt{-100)^2 + (68 - 104)^2 + (274 - 243)^2 + (31 - 31)^2 + (107 - 164)^2} \\
 &= \sqrt{10920} \\
 &= 104,50
 \end{aligned}$$

dist(Gajahmungkur, C1) =

$$\begin{aligned}
 &= \sqrt{(403 - 294)^2 + (212 - 111)^2 + (142 - 92)^2 + (498 - 361)^2 + (150} \\
 &\quad \sqrt{-100)^2 + (112 - 104)^2 + (439 - 243)^2 + (65 - 31)^2 + (176 - 164)^2} \\
 &= \sqrt{85631} \\
 &= 292,63
 \end{aligned}$$

dist(Gayamsari, C1) =

$$\begin{aligned}
 &= \sqrt{(8 - 294)^2 + (10 - 111)^2 + (2 - 92)^2 + (19 - 361)^2 + (0 - 100)^2} \\
 &\quad \sqrt{+(3 - 104)^2 + (15 - 243)^2 + (0 - 31)^2 + (4 - 164)^2} \\
 &= \sqrt{315807} \\
 &= 561,97
 \end{aligned}$$

dist(Genuk, C1) =

$$\begin{aligned}
 &= \sqrt{(31 - 294)^2 + (18 - 111)^2 + (23 - 92)^2 + (43 - 361)^2 + (19} \\
 &\quad \sqrt{-100)^2 + (8 - 104)^2 + (63 - 243)^2 + (1 - 31)^2 + (5 - 164)^2} \\
 &= \sqrt{258061} \\
 &= 508
 \end{aligned}$$

dist(Gunungpati, C1) =

$$\begin{aligned}
 &= \sqrt{(174 - 294)^2 + (82 - 111)^2 + (23 - 92)^2 + (180 - 361)^2 + (95} \\
 &\quad \sqrt{-100)^2 + (48 - 104)^2 + (236 - 243)^2 + (19 - 31)^2 + (41 - 164)^2} \\
 &= \sqrt{71246} = 266,92
 \end{aligned}$$

dist(Mijen, C1) =

$$\begin{aligned}
 &= \sqrt{(501 - 294)^2 + (216 - 111)^2 + (229 - 92)^2 + (507 - 361)^2 + (182 \\
 &\sqrt{-100)^2 + (75 - 104)^2 + (489 - 243)^2 + (74 - 31)^2 + (142 - 164)^2} \\
 &= \sqrt{164373} \\
 &= 405,43
 \end{aligned}$$

dist(Ngaliyan, C1) =

$$\begin{aligned}
 &= \sqrt{(962 - 294)^2 + (376 - 111)^2 + (227 - 92)^2 + (1092 - 361)^2 + (313 \\
 &\sqrt{-100)^2 + (182 - 104)^2 + (892 - 243)^2 + (130 - 31)^2 + (340 - 164)^2} \\
 &= \sqrt{1582466} \\
 &= 1258
 \end{aligned}$$

dist(Pedurungan, C1) =

$$\begin{aligned}
 &= \sqrt{(9 - 294)^2 + (1 - 111)^2 + (0 - 92)^2 + (20 - 361)^2 + \\
 &\sqrt{(9 - 100)^2 + (2 - 104)^2 + (10 - 243)^2 + (0 - 31)^2 + (1 - 164)^2} \\
 &= \sqrt{318574} \\
 &= 562,42
 \end{aligned}$$

dist(Semarang Barat, C1) =

$$\begin{aligned}
 &= \sqrt{(839 - 294)^2 + (308 - 111)^2 + (247 - 92)^2 + (1140 - 361)^2 + (212 \\
 &\sqrt{-100)^2 + (235 - 104)^2 + (661 - 243)^2 + (111 - 31)^2 + (374 - 164)^2} \\
 &= \sqrt{1221629} \\
 &= 1105,27
 \end{aligned}$$

dist(Semarang Selatan, C1) =

$$\begin{aligned}
 &= \sqrt{(256 - 294)^2 + (115 - 111)^2 + (196 - 92)^2 + (393 - 361)^2 + (95 \\
 &\sqrt{-100)^2 + (108 - 104)^2 + (249 - 243)^2 + (62 - 31)^2 + (133 - 164)^2} \\
 &= \sqrt{15299} \\
 &= 123,69
 \end{aligned}$$

dist(Semarang Tengah, C1) =

$$\begin{aligned}
 &= \sqrt{(114 - 294)^2 + (42 - 111)^2 + (50 - 92)^2 + (187 - 361)^2 + (52 \\
 &\sqrt{-100)^2 + (53 - 104)^2 + (89 - 243)^2 + (30 - 31)^2 + (110 - 164)^2} \\
 &= \sqrt{100739} = 317,39
 \end{aligned}$$

dist(Semarang Timur, C1) =

$$\begin{aligned}
 &= \sqrt{(49 - 294)^2 + (14 - 111)^2 + (5 - 92)^2 + (45 - 361)^2 +} \\
 &\sqrt{(8 - 100)^2 + (5 - 104)^2 + (33 - 243)^2 + (5 - 31)^2 + (5 - 164)^2} \\
 &= \sqrt{265181} \\
 &= 514,96
 \end{aligned}$$

dist(Semarang Utara, C1) =

$$\begin{aligned}
 &= \sqrt{(294 - 294)^2 + (111 - 111)^2 + (92 - 92)^2 + (361 - 361)^2 + (100} \\
 &\sqrt{-100)^2 + (104 - 104)^2 + (243 - 243)^2 + (31 - 31)^2 + (164 - 164)^2} \\
 &= \sqrt{0} \\
 &= 0
 \end{aligned}$$

dist(Tembalang, C1) =

$$\begin{aligned}
 &= \sqrt{(106 - 294)^2 + (48 - 111)^2 + (35 - 92)^2 + (140 - 361)^2 + (45} \\
 &\sqrt{-100)^2 + (35 - 104)^2 + (100 - 243)^2 + (23 - 31)^2 + (29 - 164)^2} \\
 &= \sqrt{137927} \\
 &= 371,39
 \end{aligned}$$

dist(Tugu, C1) =

$$\begin{aligned}
 &= \sqrt{(190 - 294)^2 + (62 - 111)^2 + (46 - 92)^2 + (254 - 361)^2 + (49} \\
 &\sqrt{-100)^2 + (31 - 104)^2 + (179 - 243)^2 + (29 - 31)^2 + (25 - 164)^2} \\
 &= \sqrt{58133} \\
 &= 241,11
 \end{aligned}$$

Proses perhitungan manual jarak *Euclidean* antara C1 dan setiap objek telah diperoleh. Selanjutnya, proses perhitungan manual jarak *Euclidean* antara C2 dan setiap objek akan dilakukan dengan hasil sebagai berikut:

- **Proses Perhitungan Manual Jarak *Euclidean* Antara C2 dan Objek**

dist(Banyumanik, C2) =

$$\begin{aligned}
 &= \sqrt{(45 - 190)^2 + (2 - 62)^2 + (7 - 46)^2 + (30 - 254)^2} \\
 &\sqrt{+(9 - 49)^2 + (8 - 31)^2 + (27 - 179)^2 + (4 - 29)^2 + (8 - 25)^2} \\
 &= \sqrt{102469} \\
 &= 320,11
 \end{aligned}$$

dist(Candisari, C2) =

$$\begin{aligned}
 &= \sqrt{(277 - 190)^2 + (94 - 62)^2 + (112 - 46)^2 + (311 - 254)^2 + (56} \\
 &\sqrt{-49)^2 + (68 - 31)^2 + (274 - 179)^2 + (31 - 29)^2 + (107 - 25)^2} \\
 &= \sqrt{33369} \\
 &= 182,67
 \end{aligned}$$

dist(Gajahmungkur, C2) =

$$\begin{aligned}
 &= \sqrt{(403 - 190)^2 + (212 - 62)^2 + (142 - 46)^2 + (498 - 254)^2 + (150} \\
 &\sqrt{-49)^2 + (112 - 31)^2 + (439 - 179)^2 + (65 - 29)^2 + (176 - 25)^2} \\
 &= \sqrt{245080} \\
 &= 495,06
 \end{aligned}$$

dist(Gayamsari, C2) =

$$\begin{aligned}
 &= \sqrt{(8 - 190)^2 + (10 - 62)^2 + (2 - 46)^2 + (19 - 254)^2} \\
 &\sqrt{+(0 - 49)^2 + (3 - 31)^2 + (15 - 179)^2 + (0 - 29)^2 + (4 - 25)^2} \\
 &= \sqrt{124352} \\
 &= 352,64
 \end{aligned}$$

dist(Genuk, C2) =

$$\begin{aligned}
 &= \sqrt{(31 - 190)^2 + (18 - 62)^2 + (23 - 46)^2 + (43 - 254)^2} \\
 &\sqrt{+(19 - 49)^2 + (8 - 31)^2 + (63 - 179)^2 + (1 - 29)^2 + (5 - 25)^2} \\
 &= \sqrt{88336} \\
 &= 297,21
 \end{aligned}$$

dist(Gunungpati, C2) =

$$\begin{aligned}
 &= \sqrt{(174 - 190)^2 + (82 - 62)^2 + (23 - 46)^2 + (180 - 254)^2} \\
 &\sqrt{+(95 - 49)^2 + (48 - 31)^2 + (236 - 179)^2 + (19 - 29)^2 + (41 - 25)^2} \\
 &= \sqrt{12671} = 112,57
 \end{aligned}$$

dist(Mijen, C2) =

$$\begin{aligned}
 &= \sqrt{(501 - 190)^2 + (216 - 62)^2 + (229 - 46)^2 + (507 - 254)^2 + (182 \\
 &\sqrt{-49)^2 + (75 - 31)^2 + (489 - 179)^2 + (74 - 29)^2 + (142 - 25)^2} \\
 &= \sqrt{349374} \\
 &= 591,08
 \end{aligned}$$

dist(Ngaliyan, C2) =

$$\begin{aligned}
 &= \sqrt{(962 - 190)^2 + (376 - 62)^2 + (227 - 46)^2 + (1092 - 254)^2 + (313 \\
 &\sqrt{-49)^2 + (182 - 31)^2 + (892 - 179)^2 + (130 - 29)^2 + (340 - 25)^2} \\
 &= \sqrt{2139877} \\
 &= 1462,83
 \end{aligned}$$

dist(Pedurungan, C2) =

$$\begin{aligned}
 &= \sqrt{(9 - 190)^2 + (1 - 62)^2 + (0 - 46)^2 + (20 - 254)^2} \\
 &\sqrt{+(9 - 49)^2 + (2 - 31)^2 + (10 - 179)^2 + (0 - 29)^2 + (1 - 25)^2} \\
 &= \sqrt{125773} \\
 &= 354,64
 \end{aligned}$$

dist(Semarang Barat, C2) =

$$\begin{aligned}
 &= \sqrt{(839 - 190)^2 + (308 - 62)^2 + (247 - 46)^2 + (1140 - 254)^2 + (21 \\
 &\sqrt{2 - 49)^2 + (235 - 31)^2 + (661 - 179)^2 + (111 - 29)^2 + (374 - 25)^2} \\
 &= \sqrt{1736148} \\
 &= 1317,63
 \end{aligned}$$

dist(Semarang Selatan, C2) =

$$\begin{aligned}
 &= \sqrt{(256 - 190)^2 + (115 - 62)^2 + (196 - 46)^2 + (393 - 254)^2 + (95 \\
 &\sqrt{-49)^2 + (108 - 31)^2 + (249 - 179)^2 + (62 - 29)^2 + (133 - 25)^2} \\
 &= \sqrt{74684} \\
 &= 273,28
 \end{aligned}$$

dist(Semarang Tengah, C2) =

$$\begin{aligned}
 &= \sqrt{(114 - 190)^2 + (42 - 62)^2 + (50 - 46)^2 + (187 - 254)^2 + (52 \\
 &\sqrt{-49)^2 + (53 - 31)^2 + (89 - 179)^2 + (30 - 29)^2 + (110 - 25)^2} \\
 &= \sqrt{26500} = 162,79
 \end{aligned}$$

dist(Semarang Timur, C2) =

$$\begin{aligned}
 &= \sqrt{(49 - 190)^2 + (14 - 62)^2 + (5 - 46)^2 + (45 - 254)^2} \\
 &\quad \sqrt{+(8 - 49)^2 + (5 - 31)^2 + (33 - 179)^2 + (5 - 29)^2 + (5 - 25)^2} \\
 &= \sqrt{92196} \\
 &= 303,64
 \end{aligned}$$

dist(Semarang Utara, C2) =

$$\begin{aligned}
 &= \sqrt{(294 - 190)^2 + (111 - 62)^2 + (92 - 46)^2 + (361 - 254)^2 + (100} \\
 &\quad \sqrt{-49)^2 + (104 - 31)^2 + (243 - 179)^2 + (31 - 29)^2 + (164 - 25)^2} \\
 &= \sqrt{58133} \\
 &= 241,11
 \end{aligned}$$

dist(Tembalang, C2) =

$$\begin{aligned}
 &= \sqrt{(106 - 190)^2 + (48 - 62)^2 + (35 - 46)^2 + (140 - 254)^2 + (45} \\
 &\quad \sqrt{-49)^2 + (35 - 31)^2 + (100 - 179)^2 + (23 - 29)^2 + (29 - 25)^2} \\
 &= \sqrt{26694} \\
 &= 163,38
 \end{aligned}$$

dist(Tugu, C2) =

$$\begin{aligned}
 &= \sqrt{(190 - 190)^2 + (62 - 62)^2 + (46 - 46)^2 + (254 - 254)^2 + (49} \\
 &\quad \sqrt{-49)^2 + (31 - 31)^2 + (179 - 179)^2 + (29 - 29)^2 + (25 - 25)^2} \\
 &= \sqrt{0} \\
 &= 0
 \end{aligned}$$

Setelah hasil jarak *Euclidean* antara C1 dan C2 diperoleh, setiap objek akan dikelompokkan ke dalam *centroid* terdekat menggunakan rumus pada Persamaan 2.2. Hasilnya adalah sebagai berikut:

$$\mathbf{Objek} = MIN(C1 : C2)$$

$$\mathbf{Banyumanik} = MIN(529,04; 320,11) = 320,11$$

$$\text{Candisari} = \text{MIN}(104, 5; 182, 67) = 104, 5$$

$$\text{Gajahmungkur} = \text{MIN}(104, 5; 182, 67) = 104, 5$$

$$\text{Gayamsari} = \text{MIN}(561, 97; 352, 64) = 352, 64$$

$$\text{Genuk} = \text{MIN}(508; 297, 21) = 297, 21$$

$$\text{Gunungpati} = \text{MIN}(266, 92; 112, 57) = 112, 57$$

$$\text{Mijen} = \text{MIN}(405, 43; 591, 08) = 405, 43$$

$$\text{Ngaliyan} = \text{MIN}(1258; 1462, 8) = 1258$$

$$\text{Pedurungan} = \text{MIN}(564, 42; 354, 64) = 354, 64$$

$$\text{Semarang Barat} = \text{MIN}(1105, 3; 1317, 6) = 1105, 3$$

$$\text{Semarang Selatan} = \text{MIN}(123, 69; 273, 28) = 123, 69$$

$$\text{Semarang Tengah} = \text{MIN}(317, 39; 162, 79) = 162, 79$$

$$\text{Semarang Timur} = \text{MIN}(514, 96; 303, 64) = 303, 64$$

$$\text{Semarang Utara} = \text{MIN}(0; 241, 11) = 0$$

$$\text{Tembalang} = \text{MIN}(371, 39; 163, 38) = 163, 3$$

$$\text{Tugu} = \text{MIN}(241, 11; 0) = 0$$

Setelah perhitungan manual pada iterasi pertama selesai, perhitungan komputasi akan dilakukan menggunakan *MATLAB 2013* dengan rincian *script* tersedia pada Lampiran 6. Hasil perhitungan manual ini digunakan sebagai referensi agar perhitungan komputasi bebas dari bias dan sesuai dengan tahap pemahaman bisnis. Berikut adalah hasil perhitungan secara komputasi:

Hasil Kalster Iterasi 1

Alamat	AQUA 1500ML	AQUA 220ML	AQUA 330ML	AQUA 600ML	MIZONE ACTIV LYCHEE	VIT 1500ML	VIT 200ML	VIT 330ML	VIT 550ML	Klaster
Banyumanik	45	2	7	30	9	8	27	4	8	2
Candisari	277	94	112	311	56	68	274	31	107	1
Gajahmungkur	403	212	142	498	150	112	439	65	176	1
Gayamsari	8	10	2	19	0	3	15	0	4	2
Genuk	31	18	23	43	19	8	63	1	5	2
Gunungpati	174	82	23	180	95	48	236	19	41	2
Mijen	501	216	229	507	182	75	489	74	142	1
Ngaliyan	962	376	227	1092	313	182	892	130	340	1
Pedurungan	9	1	0	20	9	2	10	0	1	2
Semarang Barat	839	308	247	1140	212	235	661	111	374	1
Semarang Selatan	256	115	196	393	95	108	249	62	133	1
Semarang Tengah	114	42	50	187	52	53	89	30	110	2
Semarang Timur	49	14	5	45	8	5	33	5	5	2
Semarang Utara	294	111	92	361	100	104	243	31	164	1
Tembalang	106	48	35	140	45	35	100	23	29	2
Tugu	190	62	46	254	49	31	179	29	25	2

Gambar 4.18. Hasil Proses Manual Iterasi Pertama

Centroid awal:									
Centroid 1:	294.00	111.00	92.00	361.00	100.00	104.00	243.00	31.00	164.00
Centroid 2:	190.00	62.00	46.00	254.00	49.00	31.00	179.00	29.00	25.00
Iterasi 1:									
Hasil Klaster: [2 1 1 2 2 2 1 1 2 1 1 2 2 1 2 2]									

Gambar 4.19. Hasil Proses Komputasi Iterasi Pertama

Hasil perhitungan komputasi dan manual pada iterasi pertama yang ditampilkan dalam Gambar 4.18 dan 4.19 menunjukkan kesamaan hasil klaster. Hal ini membuktikan bahwa proses klasterisasi telah sesuai dengan tahapan pemahaman bisnis, serta *script* program mendukung penelitian ini.

3. Pembaruan *centroid Baru*

Setelah menghitung jarak *Euclidean* antara setiap objek dan *centroid* pada iterasi pertama, tahap berikutnya adalah menghitung *centroid* baru untuk iterasi kedua. Perhitungan ini akan dilakukan secara manual dan komputasi menggunakan rumus pada Persamaan 2.3 dengan hasil sebagai berikut:

• Perhitungan Manual *centroid Baru C1*

$$\mathbf{C1} = (Candisari + Gajahmungkur + Mijen + Ngaliyan + SemarangBarat + SemarangSelatan + SemarangUtara)/7$$

C1 (AQUA1500ML)

$$\begin{aligned} &= (277 + 403 + 501 + 962 + 839 + 256 + 294)/7 \\ &= 3532/7 \\ &= 504,57 \end{aligned}$$

C1 (AQUA220ML)

$$\begin{aligned} &= (94 + 212 + 216 + 376 + 308 + 115 + 111)/7 \\ &= 1432/7 \\ &= 204,57 \end{aligned}$$

C1 (AQUA330ML)

$$\begin{aligned} &= (112 + 142 + 229 + 227 + 247 + 196 + 92)/7 \\ &= 1.245/7 \\ &= 177,86 \end{aligned}$$

C1 (AQUA600ML)

$$\begin{aligned} &= (311 + 498 + 507 + 1092 + 1140 + 393 + 361)/7 \\ &= 4.302/7 \\ &= 614,57 \end{aligned}$$

C1 (Mizone Activ Lychee Lemon 500ML)

$$\begin{aligned}
 &= (56 + 150 + 182 + 313 + 212 + 95 + 100)/7 \\
 &= 1108/7 \\
 &= 158,29
 \end{aligned}$$

C1 (VIT1500ML)

$$\begin{aligned}
 &= (68 + 112 + 75 + 182 + 235 + 108 + 104)/7 \\
 &= 884/7 \\
 &= 126,29
 \end{aligned}$$

C1 (VIT200ML)

$$\begin{aligned}
 &= (274 + 439 + 489 + 892 + 661 + 249 + 243)/7 \\
 &= 3247/7 \\
 &= 463,86
 \end{aligned}$$

C1 (VIT330ML)

$$\begin{aligned}
 &= (31 + 65 + 74 + 130 + 111 + 62 + 31)/7 \\
 &= 504/7 \\
 &= 72
 \end{aligned}$$

C1 (VIT550ML)

$$\begin{aligned}
 &= (107 + 176 + 142 + 340 + 374 + 133 + 164)/7 \\
 &= 1.436/7 \\
 &= 205,14
 \end{aligned}$$

• Perhitungan Manual centroid Baru C2

$$\begin{aligned}
 \mathbf{C2} = & (Banyumanik + Gayamsari + Genuk + Gunungpati + \\
 & Pedurungan + SemarangTengah + SemarangTimur + \\
 & Tembalang + Tugu)/9
 \end{aligned}$$

C2 (AQUA1500ML)

$$= (45 + 8 + 31 + 174 + 9 + 114 + 49 + 106 + 190)/9$$

$$= 726/9$$

$$= 80,67$$

C2 (AQUA220ML)

$$= (2 + 10 + 18 + 82 + 1 + 42 + 14 + 48 + 62)/9$$

$$= 279/9$$

$$= 31$$

C2 (AQUA330ML)

$$= (7 + 2 + 23 + 23 + 0 + 50 + 5 + 35 + 46)/9$$

$$= 191/9$$

$$= 21,22$$

C2 (AQUA600ML)

$$= (30 + 19 + 43 + 180 + 20 + 187 + 45 + 140 + 254)/9$$

$$= 918/9$$

$$= 102$$

C2 (Mizone Activ Lychee Lemon 500ML)

$$= (9 + 0 + 19 + 95 + 9 + 52 + 8 + 45 + 49)/9$$

$$= 286/9$$

$$= 31,78$$

C2 (VIT1500ML)

$$= (8 + 3 + 8 + 48 + 2 + 53 + 5 + 35 + 31)/9$$

$$= 193/9$$

$$= 21,44$$

C2 (VIT200ML)

$$= (27 + 15 + 63 + 236 + 10 + 89 + 33 + 100 + 179)/9$$

$$= 752/9$$

$$= 83,56$$

C2 (VIT330ML)

$$= (4 + 0 + 1 + 19 + 0 + 30 + 5 + 23 + 29)/9$$

$$= 111/9$$

$$= 12,33$$

C1 (VIT550ML)

$$= (8 + 4 + 5 + 41 + 1 + 110 + 5 + 29 + 25)/9$$

$$= 228/9$$

$$= 25,33$$

Setelah manual perhitungan selesai dengan hasil yang diharapkan sebagai referensi, proses perhitungan secara komputasi dilakukan untuk memastikan kesesuaian dengan tujuan penelitian pada tahap pemahaman bisnis. Berikut ini ditampilkan perbandingan hasil perhitungan manual dan komputasi:

Alamat	AQUA 1500ML	AQUA 220ML	AQUA 330ML	AQUA 600ML	MIZONE ACTIV LYCHEE LE MON 500ML	VIT 1500ML	VIT 200ML	VIT 330ML	VIT 550ML
C1	504.57	204.57	177.86	614.57	158.29	126.29	463.86	72.00	205.14
C2	80.67	31.00	21.22	102.00	31.78	21.44	83.56	12.33	25.33

Gambar 4.20. Hasil Proses Manual Perhitungan *centroid* Baru

Centroid:									

Centroid	AQUA 1500ML	AQUA 220ML	AQUA 330ML	AQUA 600ML	MIZONE ACTIV 500ML	VIT 1500ML	VIT 200ML	VIT 330ML	VIT 550ML
Centroid 1	504.57	204.57	177.86	614.57	158.29	126.29	463.86	72.00	205.14
Centroid 2	80.67	31.00	21.22	102.00	31.78	21.44	83.56	12.33	25.33

Gambar 4.21. Hasil Proses Komputasi Perhitungan *centroid* Baru

Hasil pada Gambar 4.20 dan 4.21 menunjukkan bahwa perhitungan *centroid* baru yang dilakukan secara manual dan komputasi menghasilkan hasil yang sama. Hal ini

membuktikan bahwa proses manual dan komputasi tersebut sesuai dengan tahap pemahaman bisnis dan dapat digunakan dalam penelitian ini. Perhitungan *centroid* baru bertujuan untuk menentukan *centroid* baru yang akan digunakan dalam perhitungan jarak *euclidean* pada iterasi kedua.

4. Iterasi

Pada tahap ini, proses iterasi dilaksanakan untuk menjalankan langkah-langkah penugasan dan pembaruan *centroid* baru. Setiap langkah menyebabkan perubahan posisi *centroid* sedikit demi sedikit hingga mencapai kestabilan atau konvergen. Proses iterasi ini dilakukan secara manual dan melalui komputasi dengan mengikuti prosedur tahapan penugasan dan pembaruan *centroid* yang telah diterapkan pada tahap sebelumnya.

Pada iterasi pertama, langkah-langkah tersebut dijalankan secara manual untuk memberikan pemahaman mendasar mengenai perubahan posisi *centroid*. Kemudian mulai iterasi kedua, proses dilanjutkan dengan metode komputasi menggunakan *MATLAB 2013*. Rincian kode *script* untuk implementasi komputasi dapat dilihat pada Lampiran 6. Proses komputasi ini dipilih karena terbukti memberikan hasil yang konsisten dan sama dengan perhitungan manual sehingga mempercepat proses iterasi.

Berikut ini disajikan tahapan iterasi lanjutan setelah iterasi pertama, yang mencakup posisi *centroid* baru setelah setiap iterasi, serta langkah-langkah rinci dari iterasi kedua hingga iterasi berakhir pada hasil konvergen. Hasil akhir ini menunjukkan kesesuaian antara hasil perhitungan

komputasi dan manual, serta memberikan gambaran yang jelas mengenai perubahan posisi *centroid* pada setiap langkah hingga mencapai kestabilan, yaitu :

4.1. Proses Perhitungan Iterasi Kedua

Berikut ini hasil perhitungan proses iterasi kedua secara lengkap yaitu:

Iterasi 2:					
Hasil Klaster: [2 2 1 2 2 2 1 1 2 1 1 2 2 2 2 2]					
Centroid:					

Centroid	AQUA 1500ML	AQUA 220ML	AQUA 330ML	AQUA 600ML	
MIZONE ACTIV	500ML	VIT 1500ML	VIT 200ML	VIT 330ML	VIT 550ML
Centroid 1	592.20	245.40	208.20	726.00	
190.40	142.40	546.00	88.40	233.00	
Centroid 1	117.91	44.00	35.91	144.55	
40.18	33.18	115.36	15.73	45.36	

Gambar 4.22. Hasil Proses Komputasi Iterasi Kedua

Gambar 4.22 menunjukkan hasil iterasi kedua algoritma *K-Means*, di mana objek data terbagi dalam dua klaster dengan distribusi yang lebih merata dibandingkan iterasi sebelumnya. Pada iterasi ini, *centroid* 1 mempertahankan nilai rata-rata tinggi pada beberapa produk seperti AQUA 1500ML dan AQUA 600ML mencerminkan dominasi kategori tersebut dalam klaster ini dengan beberapa penyesuaian untuk produk baru yang bergabung. Di sisi lain, *centroid* 2 menunjukkan peningkatan signifikan pada rata-rata produk seperti VIT 200ML dan VIT 550ML

mengindikasikan kecenderungan berbeda antar kluster dengan karakteristik produk yang menonjol sesuai masing-masing *centroid*. Posisi *centroid* baru ini juga akan digunakan sebagai acuan pada iterasi ketiga.

4.2. Proses Perhitungan Iterasi Ketiga

Berikut ini hasil perhitungan proses iterasi ketiga secara lengkap yaitu:

Iterasi 3:							
Hasil Klaster: [2 2 1 2 2 2 1 1 2 1 2 2 2 2 2 2]							
Centroid:							

Centroid	AQUA 1500ML	AQUA 220ML	AQUA 330ML	AQUA 600ML	MIZONE ACTIV 500ML	VIT 1500ML	VIT 200ML VIT 330ML VIT 550ML

Centroid 1	676.25	278.00	211.25	809.25	214.25	151.00	
620.25	95.00	258.00					
Centroid 2	129.42	49.92	49.25	165.25	44.75	39.42	
126.50	19.58	52.67					

Gambar 4.23. Hasil Proses Koputasi Iterasi Ketiga

Gambar 4.23 menunjukkan hasil pada iterasi ketiga algoritma *K-Means*, di mana objek data tetap terbagi dalam dua kluster dengan distribusi yang konsisten, mirip dengan iterasi sebelumnya. *centroid* 1 masih menunjukkan nilai yang lebih tinggi pada produk seperti AQUA 1500ML dan AQUA 600ML yang mengindikasikan dominasi produk tersebut dalam kluster ini. Sementara itu, *centroid* 2 mempertahankan nilai rata-rata yang lebih rendah dengan sedikit peningkatan pada produk seperti AQUA 220ML dan

VIT 200ML. Hal ini menunjukkan bahwa setiap kluster mulai mencapai stabilitas dengan karakteristik produk yang relatif konsisten pada masing-masing *centroid*. Posisi *centroid* baru ini akan digunakan sebagai acuan pada iterasi keempat.

4.3. Proses Perhitungan Iterasi Keempat

Berikut ini hasil perhitungan proses iterasi keempat secara lengkap yaitu:

Iterasi 4:					
Hasil Klaster: [2 2 1 2 2 2 1 1 2 1 2 2 2 2 2 2]					
Centroid:					

Centroid	AQUA 1500ML	AQUA 220ML	AQUA 330ML	AQUA 600ML	
MIZONE ACTIV	500ML	VIT 1500ML	VIT 200ML	VIT 330ML	VIT 550ML

Centroid 1	676.25	278.00	211.25	809.25	
214.25	151.00	620.25	95.00	258.00	
Centroid 2	129.42	49.92	49.25	165.25	
44.75	39.42	126.50	19.58	52.67	

Konvergen setelah 4 iterasi.					

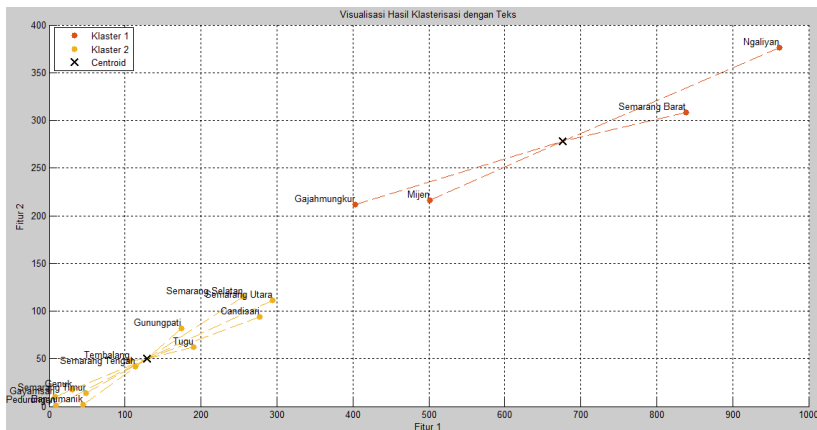
Gambar 4.24. Hasil Proses Koputasi Iterasi Keempat

Pada iterasi keempat, seperti ditunjukkan pada Gambar 4.24 bahwa algoritma *K-Means* menghasilkan klasterisasi yang telah mencapai konvergensi. Proses klasterisasi membentuk dua klaster dengan distribusi objek data yang konsisten, serupa dengan iterasi ketiga. *centroid* 1 mempertahankan nilai rata-rata tinggi untuk

produk seperti AQUA 1500ML dan AQUA 600ML yang menandakan dominasi produk tersebut dalam kluster ini. Sementara itu, *centroid* 2 juga mempertahankan nilai rata-rata lebih rendah untuk sebagian besar produk menunjukkan bahwa karakteristik kluster yang stabil. Karena tidak ada perubahan posisi *centroid* pada iterasi ini, klusterisasi telah mencapai konvergensi setelah empat iterasi.

5. Terminasi

Langkah terminasi dalam algoritma *K-Means* tercapai ketika *centroid* setiap kluster stabil atau penugasan objek ke kluster tidak berubah. Hasil kluster pada iterasi keempat akan ditampilkan pada wilayah Kecamatan Kota Semarang sesuai dengan *centroid*. Proses ini menggunakan *MATLAB 2013* dengan rincian kode pada Lampiran 6. Berikut adalah hasil proses terminasi secara visual:



Gambar 4.25. Visual Algoritma *K-Means*

Visual hasil klasterisasi dalam setiap wilayah Kecamatan Kota Semarang akan ditampilkan sesuai Penjelasan Gambar 4.25 sebagai berikut :

Hasil Klaster dengan Teks:
Klaster 1:

Gajahmungkur, , , , , , , ,
Mijen, , , , , , , ,
Ngaliyan, , , , , , , ,
Semarang Barat, , , , , , , ,

Jumlah data: 4

Klaster 2:

Banyumanik, , , , , , , ,
Candisari, , , , , , , ,
Gayamsari, , , , , , , ,
Genuk, , , , , , , ,
Gunungpati, , , , , , , ,
Pedurungan, , , , , , , ,
Semarang Selatan, , , , , , , ,
Semarang Tengah, , , , , , , ,
Semarang Timur, , , , , , , ,
Semarang Utara, , , , , , , ,
Tembalang, , , , , , , ,
Tugu, , , , , , , ,
Jumlah data: 12

Gambar 4.26. Hasil Proses Terminasi

Analisis klaster dalam Gambar 4.25 dan 4.26 ini membagi wilayah Kecamatan Kota Semarang menjadi dua kelompok berdasarkan karakteristik tertentu. Klaster 1 mencakup empat wilayah Gajahmungkur, Mijen, Ngaliyan, dan Semarang Barat yang memiliki pola data serupa namun berbeda dari wilayah lain. Sementara itu, Klaster 2 terdiri dari 12 wilayah seperti Banyumanik, Candisari, dan Genuk yang memiliki pola serupa tetapi berbeda dari Klaster 1. Jumlah wilayah yang lebih besar di Klaster 2 menunjukkan perbedaan karakteristik yang lebih signifikan. Penentuan karakteristik pada klaster 1 dan 2 akan dijelaskan secara lengkap dalam tahapan selanjutnya.

F. Evaluasi (*Evaluation*)

Tahap evaluasi dilakukan setelah pemodelan selesai untuk menilai kualitas dan efektivitas algoritma *K-Means* dalam mengelompokkan data. Kualitas struktur klaster dievaluasi menggunakan *Silhouette Coefficient* seperti ditunjukkan pada Gambar 4.14. Hasil evaluasi menunjukkan nilai 0,6081 yang menunjukkan kualitas struktur klaster cukup kuat atau standar. Nilai ini mengindikasikan bahwa struktur klaster terbentuk dengan baik, dengan potensi tumpang tindih yang kecil.

Evaluasi hasil klaster menunjukkan struktur klaster yang cukup baik sehingga dapat digunakan dalam penelitian ini. Selanjutnya, hasil klaster akan dievaluasi dalam konteks pemahaman bisnis untuk memberikan solusi atas permasalahan yang dihadapi. Pada tahap pemahaman bisnis, disebutkan bahwa perusahaan Wahyu Jaya Group Cabang Semarang

menghadapi tantangan berupa penurunan konsumsi masyarakat Kota Semarang dan persaingan penjualan dengan perusahaan lain. Untuk mengatasi permasalahan ini, penelitian ini akan menggunakan *Threshold Mean dan Standar Deviasi* sebagai ambang batas untuk mengidentifikasi segmentasi pasar Kota Semarang. Proses ini dilakukan menggunakan *MATLAB 2013* dengan rincian *script* pada Lampiran 6. Hasil proses ditampilkan sebagai berikut:

Analisis Karakteristik Data Berdasarkan Kluster:
Kategori Threshold Mean dan Standar Deviasi:

Sangat Rendah: < 16.04%
Rendah: 16.04% - 41.51%
Sedang: 41.51% - 58.49%
Tinggi: 58.49% - 83.96%
Sangat Tinggi: > 83.96%

Hasil Threshold Mean dan Standar Deviasi:
Klaster 1:
Pangsa Pasar Klaster 1: 62.01%
Kategori Penjualan: Tinggi

Klaster 2:
Pangsa Pasar Klaster 2: 37.99%
Kategori Penjualan: Rendah

Gambar 4.27. Hasil Proses Evaluasi *Threshold Mean dan Standar Deviasi*

Analisis karakteristik data membagi klaster berdasarkan ambang batas *mean dan standar deviasi* ke dalam lima kategori: Sangat Rendah, Rendah, Sedang, Tinggi, dan Sangat Tinggi dengan perhitungan persamaan dalam *MATLAB 2013* dengan rincian *script* pada Lampiran 6. Klaster 1 memiliki pangsa pasar sebesar 62,01% dan dikategorikan sebagai Tinggi menunjukkan bahwa wilayah-wilayah dalam klaster ini memiliki performa penjualan yang lebih baik. Sebaliknya, Klaster 2 dengan pangsa pasar sebesar 37,99% masuk dalam kategori Rendah yang mencerminkan performa penjualan yang relatif lebih rendah dibandingkan Klaster 1.

G. Penyebaran (*Deployment*)

Tahap penyebaran atau *deployment* merupakan langkah akhir, di mana model atau wawasan yang diperoleh diimplementasikan dalam sistem bisnis. Pada tahap ini, hasil analisis diintegrasikan ke dalam proses bisnis yang ada, serta dilakukan pelatihan bagi pengguna agar dapat memanfaatkan model dengan efektif. Penerapan hasil dapat berbentuk laporan, dashboard, atau sistem otomatis yang mendukung pengambilan keputusan bisnis secara langsung. Hasil akhir dari tahap ini adalah sistem yang dapat dijalankan oleh bisnis untuk mencapai tujuan yang diinginkan. Berikut adalah hasil tahapan penyebaran, yaitu:

1. Laporan Analisis Penjualan Produk AMDK Hasil Klasterisasi Untuk Bisnis

Dalam penelitian ini, tahap penyebaran menghasilkan laporan penjualan produk AMDK tertinggi dan terendah pada perusahaan Wahyu Jaya Group Cabang Semarang.

Berikut ini adalah hasil penjualan produk AMDK tertinggi dan terendah dengan *Microsoft Excel 2013* membentuk peta penjualan tertinggi dan terendah, yaitu:



Gambar 4.28. Peta Hasil klaster Penjualan Produk AMDK Kota Semarang

Gambar 4.28 menunjukkan peta klaster penjualan produk AMDK di Kota Semarang untuk Wahyu Jaya Semarang. Peta ini membagi wilayah administratif kota ke dalam kecamatan yang diberi warna berbeda untuk memudahkan identifikasi menjadi panduan visual dalam perencanaan dan pengelolaan kota serta menunjukkan distribusi penjualan. Warna kuning mewakili klaster 1 sebagai wilayah dengan penjualan tertinggi, sedangkan warna coklat menunjukkan klaster 2 sebagai wilayah dengan penjualan terendah, seperti terlihat pada Gambar 4.25 dan 4.26. Kecamatan Ngaliyan, Semarang Barat, Mijen, dan Gajahmungkur termasuk dalam klaster 1 dengan rata-rata penjualan tertinggi, sementara kecamatan lainnya dalam klaster 2 memiliki penjualan lebih

rendah. Setelah pembagian kecamatan, total penjualan produk AMDK dianalisis untuk mengidentifikasi produk populer sebagai dasar perencanaan distribusi dan strategi pemasaran. Gambar 4.24 menunjukkan AQUA 1500ML dan AQUA 600ML sebagai produk dengan penjualan terbaik di centroid 1 dan 2, sehingga dapat dijadikan acuan oleh Wahyu Jaya Semarang dalam strategi distribusi dan pemasaran.

2. Rekomendasi Strategi Penjualan Hasil Klasterisasi Untuk Bisnis

Rekomendasi bagi Wahyu Jaya Group Semarang setelah penelitian ini adalah memaksimalkan keuntungan, distribusi, dan stok di kecamatan dengan penjualan tinggi yang termasuk dalam klaster 1. Sementara itu, strategi penetrasi pasar perlu diterapkan di kecamatan dengan penjualan rendah yang berada di klaster 2, misalnya melalui promosi, peluncuran produk baru, atau perencanaan distribusi untuk produk populer seperti AQUA 1500ML dan AQUA 600ML. Dengan strategi ini, hasil klasterisasi dapat digunakan untuk mengoptimalkan distribusi dan meningkatkan penjualan di seluruh wilayah Kota Semarang, sehingga analisis ini relevan dan aplikatif dalam strategi bisnis nyata.

BAB V

PENUTUP

A. Kesimpulan

Penerapan model *CRISP-DM* yang dikombinasikan dengan algoritma *K-Means* terbukti efektif dalam menganalisis pola penjualan di Wahyu Jaya Group Cabang Semarang pada November 2023. Pendekatan ini berhasil mengelompokkan data penjualan ke dalam beberapa klaster berdasarkan karakteristik wilayah dengan memberikan wawasan mendalam tentang preferensi konsumen dan tren penjualan. Analisis ini juga menentukan jumlah klaster optimal menggunakan *Silhouette Coefficient* dengan hasil optimal sebanyak 2 klaster dan nilai struktur klaster sebesar 0,6081 yang menunjukkan kualitas klaster cukup kuat dengan potensi tumpang tindih rendah.

Algoritma *K-Means* menghasilkan klasterisasi dalam 4 iterasi dengan membagi wilayah kecamatan Kota Semarang menjadi dua kelompok berdasarkan karakteristik tertentu. Klasterisasi ini memudahkan identifikasi wilayah dengan tingkat penjualan tinggi dan rendah menggunakan *Threshold Mean* dan *Standar Deviasi*. Hasilnya adalah klaster 1 mencakup empat kecamatan, yaitu Gajahmungkur, Mijen, Ngaliyan, dan Semarang Barat dengan pangsa pasar sebesar 62,01% dan dikategorikan sebagai tinggi sehingga menunjukkan performa penjualan yang lebih baik di wilayah ini. Sementara itu, klaster 2 mencakup dua belas kecamatan dengan pangsa pasar sebesar 37,99% dan dikategorikan sebagai rendah sehingga mencerminkan performa penjualan yang relatif lebih rendah. Wilayah dikategorikan rendah

ini memerlukan perhatian khusus dalam strategi distribusi dan pemasaran produk AMDK di Kota Semarang.

Implikasi dari hasil ini memungkinkan perusahaan merancang pendekatan yang lebih efektif di setiap wilayah, misalnya meningkatkan promosi di wilayah berpenjualan rendah atau mempertahankan strategi yang berhasil di wilayah berpenjualan tinggi. Implikasi lainnya Dengan demikian, penerapan model *CRISP-DM* dan *algoritma K-Means* tidak hanya membantu mengenali pola penjualan, tetapi juga menyediakan dasar yang kuat bagi perusahaan untuk memperkuat posisi kompetitif dan meningkatkan kinerja bisnis di masa mendatang.

B. Saran

Berdasarkan kesimpulan yang diperoleh diharapkan bagi peneliti selanjutnya dapat mengembangkan penelitian ini dengan metode analisis klaster lainnya atau *K-Means* pada data yang lebih besar agar memberikan pengaruh signifikan terhadap metode klaster. Bagi perusahaan Wahyu Jaya Group Cabang Semarang diharapkan penelitian dijadikan referensi untuk meningkatkan penjualan produk AMDK sehingga dapat memberikan kontribusi bagi masyarakat melalui penciptaan lapangan kerja dan peningkatan akses air minum bersih.

DAFTAR PUSTAKA

- Abubakar, H. R. (2021). *Pengantar metodologi penelitian* (Edition Pertama). SUKA-Press UIN Sunan Kalijaga. ISBN : 978-632-7816-25-6.
- Badan Pusat Statistik Provinsi Jawa Tengah. (2020). *Statistik Air Bersih Provinsi Jawa Tengah 2019*. diakses dari <https://jateng.bps.go.id/publication/2020/12/30/1312ebe13368ada2a3832833/statistik-air-bersih-provinsi-jawa-tengah-2019.html> pada tanggal 1 oktober 2024, 21:01 pm.
- Badan Pusat Statistik Provinsi Jawa Tengah. (2021). *Statistik Air Bersih Provinsi Jawa Tengah 2020*. diakses dari <https://jateng.bps.go.id/publication/2021/12/30/35d62d308e18472d5b41a7c0/statistik-air-bersih-provinsi-jawa-tengah-2020.html> pada tanggal 1 oktober 2024, 21:01 pm.
- Badan Pusat Statistik Provinsi Jawa Tengah. (2022). *Statistik Air Bersih Provinsi Jawa Tengah 2021*. diakses dari <https://jateng.bps.go.id/publication/2022/12/30/7aec10230f1c7555fb949515/statistik-air-bersih-provinsi-jawa-tengah-2021.html> pada tanggal 1 oktober 2024, 21:01 pm.
- Badan Pusat Statistik Provinsi Jawa Tengah. (2023). *Statistik Air Bersih Provinsi Jawa Tengah 2022*. diakses dari <https://jateng.bps.go.id/publication/2023/12/21/b8573f603e6c51d646d47a73/statistik-air->

- bersih-provinsi-jawa-tengah-2022.html pada tanggal 1 oktober 2024, 21:01 pm.
- Davis, J. (2007). *MAGIC NUMBERS for sales management: Tolak Ukur Mengevaluasi sukses penjualan* (Edition Pertama). Jakarta: Elex Media Komputindo. ISBN : 978-979-27-1479-1.
- Dhewayani, F. N., Amelia, D., Alifah, D. N., Sari, B. N., & Jajuli, M. (2022). Implementasi K-Means Clustering untuk Pengelompokkan Daerah Rawan Bencana Kebakaran Menggunakan Model CRISP-DM. *Jurnal Teknologi Dan Informasi*, 12, 64–77.<https://doi.org/10.34010/jati.v12i1>.
- Dewi, D. A. I. C., & Pramita, D. A. K. (2019). Analisis Perbandingan Metode Elbow dan Silhouette pada Algoritma Clustering K-Medoids dalam Pengelompokan Produksi Kerajinan Bali. *Matrix: Jurnal Manajemen Teknologi dan Informatika*, 9(3), 102-109.
- Fauzan, M.F., Purnamasari, A.I. and Dwilestari, G.. (2023). Penerapan Data Mining Untuk Menganalisis Penjualan Air Minum Dalam Kemasan Selama Masa Pandemi COVID-19. *JATI (Jurnal Mahasiswa Teknik Informatika)*., 7(1), pp.700-706.
- Gorunescu, F. (2011). *Data Mining* (Vol. 12). Springer Berlin Heidelberg. <https://doi.org/10.1007/978-3-642-19721-5>.
- Han, J., Kamber, M., & Pei, J. (2012). *Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems)* (Edition Third). Morgan Kaufmann is an imprint of Elsevier.

- Iriadi, N., & Priatno, A. I. (2020). *Penerapan Data Mining dengan Rapid Miner* (Edition pertama). Graha Ilmu.
- JacopSoft Consultoria y Capacitacion. (2024). *Clustering*. diakses dari <https://www.jacobsoft.com.mx/esmx/clustering-analysis/clustering/>
- Jollyta, D., Ramdhan, W., & Zarlis, M. (2020). *Konsep data mining dan penerapan*. Deepublish.
- Jollyta, D., Siddik, M., Mawengkang, H., & Efendi, S. (2020). *TEKNIK EVALUASI CLUSTER: Solusi Menggunakan Python dan Rapidminer*. Deepublish.
- Kesumawati, N., Allen, M.R., dan Novita S. (2017). *Pengantar Statistika Penelitian* (Edition pertama). Depok: Rajawali Pers. ISBN: 978-602-425-132-1.
- Kurniawan, R., Sohibien, G. P. D., & Rahani, R. (2019). *Cara Mudah Belajar Statistik Analisis Data & Eksplorasi*. Prenada Media.
- Khumaidi, A. (2020). DATA MINING FOR PREDICTING THE AMOUNT OF COFFEE PRODUCTION USING CRISP-DM METHOD. *Techno Nusa Mandiri: Journal of Computing and Information Technology As an Accredited Journal Rank.*, 17(1). www.unkris.ac.id
- Kotler, P., & Keller, K. L. (2016). *Marketing Management (15e ed)*. Pearson India Education Services Pvt. Ltd. ISBN: 978-93-528-6880-3
- Larose, D. T., & Larose, C. D. (2014). *Discovering Knowledge in Data*. Wiley. <https://doi.org/10.1002/9781118874059>

- Lubis, M. S. (2018). *Metodologi penelitian* (Edition pertama). Deepublish.
- Marsono. (2019). Analisis Data Mining Pada Strategi Penjualan Produk PT Aquasolve Sanaria Dengan Menggunakan Metode K-Means Clustering. *J-SISKO TECH: Jurnal Teknologi Sistem Informasi Dan Sistem Komputer TGD*, 32(1), 32–41.
- Muflikhah, L., Ratnawati, D. E., & Mardi Putri, R. R. (2018). *Data mining* (Edition pertama). UB Press.
- Mulaab. (2017). *DATA MINING: Konsep dan Aplikasi*. Malang: Media Nusa Creative. 978-602-6743-98-5.
- Núñez Arias, E. B., González Nuñez, B. M., Nonell Fernández, L., & Rodríguez Pupo, J. M. (2021). CRISP-DM and K-means neutrosophic in the analysis of risk factors for hearing loss in children. *Neutrosophic Computing & Machine Learning*, 16.
- Purwanto, M. Y. J., & Susanto, A. (2015). *Pengantar Pengelolaan Sumber Daya Air*. Universitas Terbuka, Kementerian Riset, Teknologi, dan Pendidikan Tinggi, Tangerang.
- Rachmawati, A. K., Saleh, M., & Ramdani, M. N. (2021). Comparison and Prediction of Data Mining Models to Determine the Classification of Family Planning Program User Status. *Indonesian Journal of Mathematics Education*, 4(2), 66-73.
- Rachmawati, A. K., & Miasary, S. D (2021). Peramalan Penyebaran Jumlah Kasus Covid19 Provinsi Jawa Tengah dengan Metode ARIMA., *Zeta-Math Journal*, 6(1), 11-16.
- Shearer, C., (2000). The CRISP-DM model: the new blueprint for data mining. *Journal of Data Warehousing*, 5: 12-22.

- Sundari, U. Y., Panudju, A. A. T., Nugraha, A. W., Purba, F., Erlina, Y., Nurbaiti, N., . & Pereiz, Z. (2024). *Metodologi Penelitian*. Padang : CV Gita Lentera. ISBN : 978-623-88909-9-6.
- Supriadi, & Dewi Kemala Sari. (2018). *Hukum pengelolaan sumber daya air di Indonesia* (Pertama). Suluh Media.
- Sutarno (2012). *Serba-Serbi Manajemen Bisnis* (Edition Pertama). Yogyakarta: Graha Ilmu. ISBN : 978-979-756-964-8.
- Sugiyono. 2014. *Statistika untuk Penelitian*. Bandung: Alfabeta.
- Taylor, J. (2017). *Four problems with CRISP-DM and how to fix them*. kdnuggets. Retrieved from <https://www.kdnuggets.com/2017/01/four-problems-crisp-dm-fix.html>
- Widodo. (2017). *Metodologi Penelitian Populer & Praktis*. Jakarta: Rajawali. ISBN : 978-602-475-088-1.

Lampiran 1. Sumber Data

Nomor	Nomor Dokumen	Tanggal Dokumen	Status Dokumen	Nomor Sales Order	ID Pelanggan	Nama Pelanggan	Alamat	ID Produk	Nama Produk	Jumlah Produk	Satuan Produk	Harga	Total Diskon	DPP/Pen. dapat. Bersih)
1	852-	#####	Applied	852-	852-	SN	Semara	145141	MIZONE	2	BOX	42000	0	75676
2	852-	#####	Applied	852-	852-	MAGY -	Semara	74560	VIT	0	BOTOL	14000	0	0
3	852-	#####	Applied	852-	852-	MAGY -	Semara	29310	WIT	0	LEMBA	0	0	0
4	852-	#####	Applied	852-	852-	MAGY -	Semara	74559	AQUA	100	BOTOL	15750	0	1418919
5	852-	#####	Applied	852-	852-	MAGY -	Semara	19310	AQUA	100	LEMBA	0	0	0
6	852-	#####	Applied	852-	852-	AHS -	Ncaliva	19310	AQUA	200	LEMBA	0	0	0
7	852-	#####	Applied	852-	852-	AHS -	Semara	74559	AQUA	200	BOTOL	15750	0	2837838
8	852-	#####	Applied	852-	852-	AHS -	Semara	19310	AQUA	75	LEMBA	0	0	0
9	852-	#####	Applied	852-	852-	AHS -	Semara	74559	AQUA	75	BOTOL	15750	0	1064189
10	852-	#####	Applied	852-	852-	AHS -	Semara	1016993	AQUA	6	BOTOL	30000	0	0
11	852-	#####	Applied	852-	852-	AIR	Candisa	74559	AQUA	50	BOTOL	15750	0	709459
12	852-	#####	Applied	852-	852-	AIR	Candisa	19310	AQUA	50	LEMBA	0	0	0
13	852-	#####	Applied	852-	852-	AHS FENI	Mijen	74559	AQUA	30	BOTOL	15750	0	425676
14	852-	#####	Applied	852-	852-	AHS FENI	Mijen	19310	AQUA	30	LEMBA	0	0	0
15	852-	#####	Applied	852-	852-	ANAN	Mijen	74559	AQUA	50	BOTOL	15750	0	709459
16	852-	#####	Applied	852-	852-	ANAN	Mijen	19310	AQUA	50	LEMBA	0	0	0
17	852-	#####	Applied	852-	852-	AL WIDAH	Tembal	74559	AQUA	20	BOTOL	17000	0	306306
18	852-	#####	Applied	852-	852-	AL WIDAH	Tembal	19310	AQUA	20	LEMBA	0	0	0
19	852-	#####	Applied	852-	852-	AL WIDAH	Tembal	1016993	AQUA	1	BOTOL	30000	0	0
20	852-	#####	Applied	852-	852-	CENTRAL	Mijen	74559	AQUA	50	BOTOL	15750	0	709459
21	852-	#####	Applied	852-	852-	CENTRAL	Mijen	19310	AQUA	50	LEMBA	0	0	0
22	852-	#####	Applied	852-	852-	TIRTO	Gunung	74560	VIT	30	BOTOL	14000	45000	337838
23	852-	#####	Applied	852-	852-	TIRTO	Gunung	29310	VIT	30	LEMBA	0	0	0
24	852-	#####	Applied	852-	852-	TIRTO	Gunung	74559	AQUA	50	BOTOL	15750	0	709459
25	852-	#####	Applied	852-	852-	TIRTO	Gunung	19310	AQUA	50	LEMBA	0	0	0
.....														
25196	852-	#####	Applied	852-	852-	MIE	Semara	134578	AQUA	3	BOX	32000	0	86486
25197	852-	#####	Applied	852-	852-	MIE	Semara	74561	AQUA	13	BOX	47000	0	550450
25198	852-	#####	Applied	852-	852-	BU	Tugu	74561	AQUA	1	BOX	47000	0	42342
25199	852-	#####	Applied	852-	852-	BU	Tugu	74553	AQUA	1	BOX	52500	0	47297
25200	852-	#####	Applied	852-	852-	BU	Tugu	157095	VIT	1	BOX	32500	0	29279
25201	852-	#####	Applied	852-	852-	POC/TIGA		74561	AQUA	2	BOX	47000	6000	79279
25202	852-	#####	Applied	852-	852-	POC/TIGA		157095	VIT	1	BOX	32500	2500	27027
25203	852-	#####	Applied	852-	852-	POC/TIGA		145141	MIZONE	1	BOX	42000	3000	35135
25204	852-	#####	Applied	852-	852-	SOLEH-	Semara	74561	AQUA	2	BOX	47000	6000	79279
25205	852-	#####	Applied	852-	852-	WR KARTI	Semara	74561	AQUA	2	BOX	47000	6000	79279
25206	852-	#####	Applied	852-	852-	WR KARTI	Semara	157095	VIT	1	BOX	32500	2500	27027
25207	852-	#####	Applied	852-	852-	WR KARTI	Semara	145141	MIZONE	1	BOX	42000	3000	35135
25208	852-	#####	Applied	852-	852-	HO TEL	Semara	157095	VIT	80	BOX	32500	10000	225252
25209	852-	#####	Applied	852-	852-	AGUNG -	Semara	157095	VIT	2	BOX	32500	5000	54054
25210	852-	#####	Applied	852-	852-	CV/NARWA	Semara	74553	AQUA	200	BOX	49000	240000	8612613

Lampiran 2. Surat Perizinan Penelitian

**CV. WAHYU JAYA**
Kawasan Industri Candi Blok 3 No 2

Semarang , 02 Mei 2024

Perihal : Balasan Permohonan Izin Penelitian

Yth. Bapak/Ibu
Fakultas SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI WALISONGO SEMARANG
di Tempat

Dengan hormat,

Yang bertanda tangan di bawah ini:

Nama : Iwan Sugiyanto
Jabatan : Kepala Depo Cv Wahyu Jaya Semarang

Menerangkan bahwa:

Nama : Rahmat Bagus Pangestu
NIM : 2008046040

Telah kami setuju untuk mengadakan penelitian di kantor CV.Wahyu Jaya Semarang dengan judul penelitian: "Analisis Pola Penjualan Menggunakan Pendekatan CRISP-DM dan Algoritma K-means"(Studi Kasus : CV.Wahyu Jaya Semarang Kawasan Industri Candi Semarang)

Demikian surat balasan ini kami sampaikan. Atas Perhatian & Kerjasamanya saya ucapkan Terima kasih.

KA. Depo Wahyu Jaya


Iwan Sugiyanto

CV. WAHYU JAYA
SEMARANG

Lampiran 3. Script RStudio 2024.09.0+375 Data Understanding

```

library(readxl)
library(dplyr)
library(tidyr)
library(writexl)
clc <- function()
cat("&O14")
clc()
starttime <- Sys.time()
data <- readexcel("Skripsi/November data awal - Copy.xlsx", sheet
= "Sheet1", coltypes = c("text", "text", "date", "text", "text", "text",
"text", "text", "text", "text", "numeric", "text", "numeric",
"numeric", "numeric")
)
cat("Data telah dibaca dari file Excel.&n")
cat("&nRingkasan Data:&n")
numrows <- nrow(data)
numcols <- ncol(data)
cat(sprintf("Jumlah kolom: cat(sprintf("Jumlah baris:
cat("Peringatan: Data memiliki", numrows, "baris, yang mungkin
mempengaruhi performa.&n")
cat("-----&n")
cat("Nama-nama Kolom:&n")
colnameslist <- colnames(data)
for (i in seq(1, numcols, by = 3)
)

```

```

endindex <- min(i + 2, numcols)
cat(sprintf("colnameslist[i], if (i + 1 <= numcols)
colnameslist[i + 1] else "", if (i + 2 <= numcols)
colnameslist[i + 2] else ""))
)
cat("—————&n")
cat("Jumlah Nilai Unik, Tipe Data, dan Contoh Nilai Unik per
Kolom:&n")
summarylist <- list()
for (columnname in names(data))
)
column <- data[[columnname]] numunique <-
length(unique(column))
)
type <- class(column)
uniquevalues <- unique(column)
examplevalues <- head(uniquevalues, 5)
examplevaluesstr <- paste(examplevalues, collapse = ", ")
summarylist[[columnname]] <- list("Kolom" = columnname,
"Jumlah Nilai Unik" = numunique, "Tipe Data" = type, "Contoh
Nilai Unik" = examplevaluesstr)
cat(sprintf("&nKolom: cat(sprintf("Jumlah Nilai Unik:
cat(sprintf("Tipe Data: cat(sprintf("Contoh Nilai Unik:
summarydf <- bindrows(lapply(summarylist, as.data.frame))
)
summaryfilepath <-
"C:/Users/62857/OneDrive/Documents/Skripsi/summarydata.xlsx"
writexlsx(summarydf, path = summaryfilepath)
cat("Ringkasan jumlah nilai unik, tipe data, dan contoh nilai unik

```

```

per kolom telah disimpan ke file Excel.&n")
cat("-----&n")
cat("-----&n")
distributionbydate <- data groupby('Tanggal Dokumen')
summarize(JumlahProduk = sum('Jumlah Produk', na.rm = TRUE)
, PendapatanBersih = sum('Pendapatan Bersih', na.rm = TRUE)
)
arrange('Tanggal Dokumen')
cat("Distribusi Penjualan berdasarkan Tanggal Dokumen:&n")
print(distributionbydate, n = 30)
writexlsx(distributionbydate, path =
"C:/Users/62857/OneDrive/Documents/Skripsi/
distribusipenjualantanggal.xlsx")
cat("Hasil distribusi penjualan berdasarkan Tanggal Dokumen
telah disimpan ke file Excel.&n")
cat("-----&n")
distributionbystatus <- data groupby('Status Dokumen')
summarize(JumlahProduk = sum('Jumlah Produk', na.rm = TRUE)
, PendapatanBersih = sum('Pendapatan Bersih', na.rm = TRUE)
)
arrange('Status Dokumen')
cat("Distribusi Penjualan berdasarkan Status Dokumen:&n")
print(distributionbystatus)
writexlsx(distributionbystatus, path =
"C:/Users/62857/OneDrive/Documents/Skripsi/
distribusipenjualanstatus.xlsx")
cat("Hasil distribusi penjualan berdasarkan Status Dokumen telah
disimpan ke file Excel.&n")
cat("-----&n")

```

```

distributionbyaddress <- data groupby('Alamat')
summarize(JumlahProduk = sum('Jumlah Produk', na.rm = TRUE)
, PendapatanBersih = sum('Pendapatan Bersih', na.rm = TRUE)
)
arrange('Alamat')
cat("Distribusi Penjualan berdasarkan Alamat:&n")
print(distributionbyaddress)
writexlsx(distributionbyaddress, path =
"C:/Users/62857/OneDrive/Documents/Skripsi/
distribusipenjualanalamat.xlsx")
cat("Hasil distribusi penjualan berdasarkan Alamat telah
disimpan ke file Excel.&n")
cat("-----&n")
distributionbyproduct <- data groupby('Nama Produk')
summarize(JumlahProduk = sum('Jumlah Produk', na.rm = TRUE)
, PendapatanBersih = sum('Pendapatan Bersih', na.rm = TRUE)
)
arrange('Nama Produk')
cat("Distribusi Penjualan berdasarkan Nama Produk:&n")
print(distributionbyproduct)
writexlsx(distributionbyproduct, path =
"C:/Users/62857/OneDrive/Documents/Skripsi/
distribusipenjualanproduk.xlsx") cat("Hasil distribusi penjualan
berdasarkan Nama Produk telah disimpan ke file Excel.&n")
cat("-----&n")
cat("Deteksi Missing Value:&n")
missingvaluespercolumn <- sapply(data, function(x)
sum(is.na(x)
)

```

```

)
totalmissingvalues <- sum(missingvaluespercolumn)
rowsbefore <- nrow(data)
cat("Jumlah missing value per kolom:&n")
print(missingvaluespercolumn)
cat(sprintf("Jumlah total missing value: cat(sprintf("Jumlah baris
sebelum menangani missing value:
missingvaluesdf <- data.frame( Kolom =
names(missingvaluespercolumn)
, JumlahMissingValue = missingvaluespercolumn )
writexlsx(missingvaluesdf, path =
"C:/Users/62857/OneDrive/Documents/Skripsi/
Jumlahmissingvalueperkolom.xlsx")
cat("Hasil Jumlah missing value per kolom telah disimpan ke file
Excel.&n")
cat("—————&n")
datacleaned <- na.omit(data)
rowsafter <- nrow(datacleaned)
cat(sprintf("Jumlah baris setelah menangani missing value:
cat(sprintf("Jumlah baris yang dihapus: cat("Beberapa baris data
setelah pembersihan:&n")
print(head(datacleaned)
)
cat("—————&n")
cat("Deteksi Duplikat:&n")
duplicaterows <- datacleaned[duplicated(datacleaned)
, ] numduplicaterows <- nrow(duplicaterows)
cat(sprintf("Jumlah baris duplikat:
if (numduplicaterows > 0)

```

```

cat("Contoh baris duplikat:&n")
print(head(duplicaterows)
)
else cat("Tidak ada baris duplikat ditemukan.&n")
cat("-----&n")
datanoduplicates <- datacleaned
numericcols <- sapply(datanoduplicates, is.numeric)
datanumeric <- datanoduplicates[, numericcols]
outlierssummary <- list()
rowswithoutliers <- vector()
for (colname in colnames(datanumeric)
)
coldata <- datanumeric[[colname]] Q1 <- quantile(coldata, 0.25,
na.rm = TRUE)
Q3 <- quantile(coldata, 0.75, na.rm = TRUE)
IQR <- Q3 - Q1 lowerbound <- Q1 - 1.5 * IQR upperbound <- Q3 +
1.5 * IQR outlierindices <- which(coldata < lowerbound | coldata >
upperbound)
cat(sprintf("&nKolom: cat(sprintf(" Minimum: cat(sprintf(" Q1:
cat(sprintf(" Median: cat(sprintf(" Q3: cat(sprintf(" Maximum:
cat(sprintf(" IQR: cat(sprintf(" Batas Bawah: cat(sprintf(" Batas
Atas: cat(sprintf(" Jumlah Outliers:
outlierssummary[[colname]] <- data.frame(
Kolom = colname,
Minimum = min(coldata, na.rm = TRUE)
, Q1 = Q1,
Median = median(coldata, na.rm = TRUE)
, Q3 = Q3,
Maximum = max(coldata, na.rm = TRUE)

```

```
, IQR = IQR,
BatasBawah = lowerbound,
BatasAtas = upperbound,
JumlahOutliers = length(outlierindices))
rowswithoutliers <- unique(c(rowswithoutliers,
outlierindices)
)
outlierssummarydf <- do.call(rbind, outlierssummary)
writexlsx(outlierssummarydf, path =
"C:/Users/62857/OneDrive/Documents/Skripsi/outlierssummary.xlsx")
cat("Hasil deteksi outliers telah disimpan ke file Excel.&n")
cat("—————&n")
cat(sprintf("Jumlah baris terkena outliers: datafinal <-
datanoduplicates[-rowswithoutliers, ]
cat(sprintf("Jumlah baris setelah menghapus outliers:
cleanedfilepath <-
"C:/Users/62857/OneDrive/Documents/Skripsi/dataProses.xlsx"
writexlsx(datafinal, path = cleanedfilepath)
cat("Data setelah pembersihan telah disimpan ke file Excel.&n")
cat("—————&n")
endtime <- Sys.time()
totaltime <- endtime - starttime
cat(sprintf("Waktu Eksekusi:
cat("—————&n")
cat("Penyelesaian Script&n")
```

Lampiran 4. Script RStudio 2024.09.0+375 Data Preparation

```

library(readxl)
library(dplyr)
library(writexl)
library(tidyr)
clc <- function() cat("014")
clc()
starttime <- Sys.time()
data <- readexcel("Skripsi/dataProses.xlsx", sheet = "Sheet1",
coltypes = c("text", "text", "date", "text", "text", "text", "text", "text",
"text", "text", "numeric", "text", "numeric", "numeric", "numeric"))
cat("Data telah dibaca dari file Excel dataproses.n")
endtime <- Sys.time() timetaken <- endtime - starttime
cat(sprintf("Waktu membaca data:
cat("—————n")
cat("nRingkasan Data:n") numrows <- nrow(data) numcols <-
ncol(data) cat(sprintf("Jumlah kolom: cat(sprintf("Jumlah baris:
cat("Peringatan: Data memiliki", numrows, "baris, yang mungkin
mempengaruhi performa.n")
cat("—————n")
cat("nRingkasan Kolom:n") summarylist <- list() for (i in
1:numcols) colname <- colnames(data)[i] coldata <- data[[i]]
coltype <- class(coldata)[1] numdata <- length(coldata)
numuniquevalues <- length(unique(na.omit(coldata)))
cat(sprintf("nKolom: cat(sprintf(" Jumlah data: cat(sprintf("
Jumlah jenis nilai:
summarylist[[i]] <- list( Kolom = colname, 'Tipe Data' = coltype,

```

```

'Jumlah Data' = numdata, 'Jumlah Jenis Nilai' = numuniquevalues )
if (is.character(coldata)) uniquevalues <-
unique(na.omit(coldata)) numdisplay <- min(numuniquevalues,
20) maxvaluewidth <- 30 for (j in 1:numdisplay) cat(sprintf(" if (j
if (numuniquevalues > 20) cat(sprintf(" ... ( cat("n", rep("-",
maxvaluewidth + 10), "n", sep = "")
summarydf <- bindrows(lapply(summarylist, as.data.frame))
summaryfilepath <-
"C:/Users/62857/OneDrive/Documents/Skripsi/summarydataProses.xlsx"
writexlsx(summarydf, path = summaryfilepath) cat("Ringkasan
jumlah nilai unik, tipe data, dan contoh nilai unik per kolom telah
disimpan ke file Excel.n")
cat("-----n")
data <- data cat("nKolom telah dihapus sehingga hanya tersisa
kolom yang diperlukan.nProses telah berhasil.n")
cat("-----n")
initialrows <- nrow(data) data <- data removedvoidrows <-
initialrows - nrow(data) cat(sprintf("nJumlah baris dengan 'Void'
di 'Status Dokumen' yang dihapus: cat("Proses penghapusan baris
dengan nilai 'Void' telah berhasil.n")
cat("-----n")
unwantedvalues <- c("VIT TISSUE", "AQUA 5GLN BTL", "VIT GLN
ISI", "AQUA GLN ISI", "AQUA TISSUE", "DISPENSER BAJA PUTIH 2
KERAN", "VIT GLN BTL", "AQUA 750ML", "AQUA 5GLN BTL - NO
VALUE") initialrows <- nrow(data) data <- data
removedproductrows <- initialrows - nrow(data)
cat(sprintf("nJumlah baris dengan jenis nilai tertentu yang
dihapus: cat("Proses penghapusan jenis nilai tertentu pada 'Nama
Produk' telah berhasil.n")

```

```

cat("-----n")
cat("nProses Deteksi dan Pengecekan Missing Value:n")
missingvalues <- sum(is.na(data)) if (missingvalues > 0)
cat(sprintf("Data memiliki data <- na.omit(data) cat("Missing
values telah dihapus dari data.n") else cat("Data tidak memiliki
missing values.n") cat("-----n")
cat("nProses Deteksi dan Pengecekan Data Duplikat:n")
duplicaterows <- data[duplicated(data), ] if (nrow(duplicaterows)
> 0) cat(sprintf("Data memiliki data <- data[!duplicated(data), ]
cat("Data duplikat telah dihapus.n") else cat("Data tidak memiliki
baris duplikat.n") cat("-----n")
cat("nProses Deteksi Outliers Menggunakan Metode IQR:n")
numericcolumns <- sapply(data, is.numeric) outlierlist <- list()
outliersdetected <- FALSE
for (colname in names(data)[numericcolumns]) coldata <-
data[[colname]] Q1 <- quantile(coldata, 0.25, na.rm = TRUE) Q3
<- quantile(coldata, 0.75, na.rm = TRUE) IQR <- Q3 - Q1
lowerbound <- Q1 - 1.5 * IQR upperbound <- Q3 + 1.5 * IQR
outlierindices <- which(coldata < lowerbound | coldata >
upperbound) numoutliers <- length(outlierindices)
if (numoutliers > 0) outliersdetected <- TRUE
cat(sprintf("nKolom: cat(sprintf(" Minimum: cat(sprintf(" Q1:
cat(sprintf(" Median: cat(sprintf(" Q3: cat(sprintf(" Maximum:
cat(sprintf(" IQR: cat(sprintf(" Batas Bawah: cat(sprintf(" Batas
Atas: cat(sprintf(" Jumlah Outliers:
outlierlist[[colname]] <- list( Kolom = colname, JumlahOutliers =
numoutliers, NilaiBatasBawah = lowerbound, NilaiBatasAtas =
upperbound )
data <- data[-outlierindices, ]

```

```

if(!outliersdetected) cat("Tidak ada outliers yang terdeteksi.n")
else cat("nOutliers telah dihapus dari data.n")
cat("-----n")
outlierdf <- bindrows(lapply(outlierlist, as.data.frame))
outlierfilepath <-
"C:/Users/62857/OneDrive/Documents/Skripsi/outliersdataProses.xlsx"
writexlsx(outlierdf, path = outlierfilepath) cat("Informasi outliers
telah disimpan ke file Excel.n")
cat("-----n") cat("nData yang
sudah dibersihkan akan disimpan ke file Excel baru.n")
pivotdata <- data groupby(Alamat, 'Nama Produk')
summarise(Jumlah = sum('Jumlah Produk')) spread(key = 'Nama
Produk', value = Jumlah, fill = 0)
cat("nHasil data setelah pivot:n") print(head(pivotdata, 20))
View(pivotdata) cat("-----n")
dirpath <- "C:/Users/62857/OneDrive/Documents/Skripsi/" if
(!dir.exists(dirpath)) dir.create(dirpath, recursive = TRUE)
cat("Direktori telah dibuat.n")
cleanedfilepath <- file.path(dirpath, "dataPenelitian(3).xlsx")
writexlsx(pivotdata, path = cleanedfilepath) cat("Data pivot telah
disimpan ke file Excel dataPenelitian.xlsx.n")
cat("-----n")

```

Lampiran 5. Script MATLAB 2013 Silhouette Coefficient

```
% Menghapus command window
clc;
clear;
% 1. Membaca Data dari File Excel filename =
'dataPenelitian(3).xlsx';
sheet = 'Sheet1';
range = 'A1:J17';
[num, txt, raw] = xlsread(filename, sheet, range);
% Membulatkan data numerik ke bilangan bulat untuk analisis
numRounded = round(num);
% Menampilkan data mentah dengan format kolom yang rapi
disp('Data mentah:');
[nRowsRaw, nColsRaw] = size(raw);
% Menentukan lebar kolom maksimum untuk setiap kolom
colWidths = zeros(1, nColsRaw);
for j = 1:nColsRaw
colWidths(j) = max(cellfun(@(x) length(num2str(x)),
raw(:, j)));
end
% Mengurangi lebar kolom maksimum dengan satu karakter agar
tidak terlalu lebar
colWidths = colWidths - 1;
% Menampilkan data mentah dengan garis pemisah vertikal
for i = 1:nRowsRaw
fprintf('|');
for j = 1:nColsRaw
```

```

fprintf(' %-*s |', colWidths(j), num2str(rawi, j));
end
fprintf('n');
end
% Menampilkan keterangan untuk data numerik tanpa
menampilkan datanya
disp('Data numerik telah dibulatkan ke format bilangan bulat.');
```

% 2. Membuat Ringkasan Data

```

disp('Ringkasan Data:');
% Menampilkan jumlah kolom
fprintf('Jumlah kolom: %dn', nColsRaw);
% Menampilkan nama-nama kolom
fprintf('Nama-nama kolom: ');
for j = 1:nColsRaw
fprintf('%s', raw1, j);
if j < nColsRaw
fprintf(', ');
else
fprintf('n');
end
end
% Menampilkan jumlah baris
fprintf('Jumlah baris: %dn', nRowsRaw - 1); % Mengurangi 1
untuk baris header
```

% 3. Menentukan Jumlah Kluster Optimal dan Menghitung Silhouette Coefficient

```

% Rentang jumlah kluster yang akan diuji
kRange = 2:5;
bestK = 0;
```

```

bestSilhouette = -inf;
% Menyimpan nilai Silhouette Coefficient global untuk semua k
globalSilhouettes = zeros(length(kRange), 1);
for idx = 1:length(kRange)
k = kRange(idx);
fprintf('nEvaluasi untuk jumlah kluster: %dn', k);
% Melakukan clustering dengan jumlah kluster k labels =
kmeans(numRounded, k);
% Menampilkan hasil akhir clustering
fprintf('Hasil akhir clustering:n');
fprintf('-----n');
for i = 1:size(numRounded, 1)
fprintf('Data %3d: Kluster %2dn', i, labels(i));
end
% Matriks jarak Euclidean
distances = pdist2(numRounded, numRounded);
% Inisialisasi variabel untuk menyimpan nilai a(i) dan b(i)
a = zeros(size(numRounded, 1), 1);
b = zeros(size(numRounded, 1), 1);
% Menghitung nilai a(i) dan b(i)
for i = 1:size(numRounded, 1)
clusteri = labels(i);
% Jarak rata-rata dalam kluster yang sama (a(i))
clusteriindices = (labels == clusteri);
a(i) = mean(distances(i, clusteriindices));
% Jarak rata-rata ke kluster lain yang terdekat (b(i))
b(i) = inf;
for c = 1:k
if c = clusteri

```

```

clusterindices = (labels == c);
b(i) = min(b(i), mean(distances(i, clusterindices)));
end end end
% Menghitung nilai Silhouette Coefficient (s(i))
s = (b - a) ./ max(a, b);
% Menghitung nilai Silhouette Coefficient global
globals = mean(s);
globalSilhouettes(idx) = globals;
% Memperbarui jumlah kluster optimal jika nilai globals lebih
baik
if globals > bestSilhouette
bestSilhouette = globals;
bestK = k;
end
% Menampilkan detail nilai a(i), b(i), s(i), dan nilai global
fprintf('Detail perhitungan Silhouette Coefficient:n');
fprintf('-----n');
fprintf('Data | a(i) | b(i) | s(i) n');
fprintf('-----n');
for i = 1:size(numRounded, 1)
fprintf('%4d | %8.4f | %8.4f | %8.4fn', i, a(i), b(i), s(i));
end fprintf('-----n');
fprintf('Nilai Silhouette Coefficient global: %.4fn', globals);
end
% Menampilkan nilai Silhouette Coefficient global untuk semua
jumlah kluster
fprintf('nNilai Silhouette Coefficient global untuk semua jumlah
kluster:n');
fprintf('-----n');

```

```

fprintf('Jumlah Kluster | Nilai SC Global\n');
fprintf('—————n');
for idx = 1:length(kRange)
fprintf(' %2d | %.4fn', kRange(idx),
globalSilhouettes(idx));
end fprintf('—————n');
% Menampilkan jumlah kluster optimal
fprintf('nJumlah kluster optimal: %dn', bestK);
fprintf('Nilai Silhouette Coefficient global untuk jumlah kluster
optimal: %.4fn', bestSilhouette);
% Interpretasi
disp('Interpretasi nilai Silhouette Coefficient global:');
if bestSilhouette > 0.7
disp('Struktur kluster yang kuat.');
```

```
elseif bestSilhouette > 0.5
disp('Struktur kluster yang standar.');
```

```
elseif bestSilhouette > 0.25
disp('Struktur kluster yang lemah.');
```

```
else
disp('Tidak memiliki struktur yang jelas.');
```

```
end
% 5. Visualisasi Silhouette Coefficient
fprintf('Visualisasi Silhouette Coefficient:n');
```

```
figure;
hold on;
plot(kRange, globalSilhouettes, '-o', 'LineWidth', 2,
'MarkerSize', 6, 'DisplayName', 'Silhouette Coefficient');
```

```
plot(bestK, bestSilhouette, 'ro', 'MarkerSize', 10,
'MarkerFaceColor', 'r', 'DisplayName', 'Jumlah Cluster Optimal');
```

```
% Highlight optimal cluster
title('Silhouette Coefficient untuk berbagai jumlah klaster');
xlabel('Jumlah Klaster');
ylabel('Silhouette Coefficient');
grid on;
legend('show'); % Menampilkan legend
hold off;
fprintf('—————n');
```

Lampiran 6. Script MATLAB 2013 Algoritma K-Means

```

clc;
clear;
% 1. Membaca Data dari File Excel filename =
'dataPenelitian(3).xlsx';
sheet = 'Sheet1';
range = 'A1:J17';
[num, txt, raw] = xlsread(filename, sheet, range);
numRounded = round(num);
disp('Data mentah:'); [nRowsRaw, nColsRaw] = size(raw);
colWidths = zeros(1, nColsRaw); for j = 1:nColsRaw
colWidths(j) = max(cellfun(@(x) length(num2str(x)), raw(:, j)));
end
colWidths = colWidths - 1;
for i = 1:nRowsRaw
fprintf('|');
for j = 1:nColsRaw
fprintf('
end
fprintf('n');
end
disp('Data numerik telah dibulatkan ke format bilangan bulat.');
```

% 2. Membuat Ringkasan Data

```

disp('Ringkasan Data:');
fprintf('Jumlah kolom:
fprintf('Nama-nama kolom: ');
for j = 1:nColsRaw
```

```

fprintf('if j < nColsRaw
fprintf(', ');
else fprintf('n');
end end
fprintf('Jumlah baris:
% 3. Melakukan Clustering dengan Jumlah Klaster yang Dapat
Ditentukan
rng('default');
optimalClusters = 2;
maxIter = 1000;
tolerance = 1e-6;
clusterIdx = kmeans(numRounded, optimalClusters, 'Replicates',
50);
initialCentroidsIdx = randperm(size(numRounded, 1),
optimalClusters);
initialCentroids = numRounded(initialCentroidsIdx, :);
disp('Centroid awal:'); for i = 1:optimalClusters
fprintf('Centroid
fprintf('fprintf('n');
end
for iter = 1:maxIter
dists = pdist2(numRounded, initialCentroids, 'euclidean');
[ , idx] = min(dists, [], 2);
newCentroids = zeros(optimalClusters, size(numRounded, 2));
for i = 1:optimalClusters
newCentroids(i, :) = mean(numRounded(idx == i, :), 1);
end
fprintf('Iterasi fprintf('Hasil Klaster:
fprintf('Centroid:n');

```

```

fprintf('—————n');
fprintf('| for j = 1:size(numRounded, 2)
fprintf('Var
end fprintf('n—————n');
for i = 1:optimalClusters
fprintf('| fprintf('fprintf('n');
end
fprintf('—————n');
if norm(newCentroids - initialCentroids, 'fro') < tolerance
fprintf('Konvergen setelah
break;
end initialCentroids = newCentroids;
end
% 4. Menampilkan Hasil Klaster dengan Teks
fprintf('Hasil Klaster dengan Teks:n');
if size(txt, 1) > 1
meanCluster1 = mean(numRounded(clusterIdx == 1, :), 1);
meanCluster2 = mean(numRounded(clusterIdx == 2, :), 1);
if mean(meanCluster1) > mean(meanCluster2)
clusterIdx = 3 - clusterIdx;
end
for k = 1:optimalClusters
fprintf('nKlaster clusterindices = find(clusterIdx == k);
fprintf('Indeks data: ');
fprintf('fprintf('n');
fprintf('—————n');
for i = 1:numel(clusterindices)
rowIdx = clusterindices(i) + 1;
fprintf('end

```

```

fprintf('—————n');
fprintf('Jumlah data:
fprintf('—————n');
end
else disp('Data teks kosong atau hanya terdiri dari satu baris.');
```

e

% 5. Visualisasi Hasil Klasterisasi dengan Teks

```

colors = 0.8500 0.3250 0.0980
; 0.9290 0.6940 0.1250;
0.4940 0.1840 0.5560;
0.4660 0.6740 0.1880;
0.3010 0.7450 0.9330;
0.6350 0.0780 0.1840;
0.75 0.75 0.75;
;
figure;
hold on;
for k = 1:optimalClusters
clusterData = numRounded(clusterIdx == k, :);
scatter(clusterData(:, 1), clusterData(:, 2), 50, colors(k, :), 'filled',
'DisplayName', sprintf('Klaster
for i = 1:size(clusterData, 1)
plot(clusterData(i, 1), newCentroids(k, 1),
clusterData(i, 2), newCentroids(k, 2), '-', 'Color',
colors(k, :), 'HandleVisibility', 'off');
end end
plot(newCentroids(:, 1), newCentroids(:, 2), 'kx', 'MarkerSize',
12, 'LineWidth', 2, 'DisplayName', 'Centroid');
if size(txt, 1) > 1
```

```

for i = 1:size(numRounded, 1)
text(numRounded(i, 1), numRounded(i, 2), txti + 1,
'VerticalAlignment', 'bottom', 'HorizontalAlignment', 'right');
end end
title('Visualisasi Hasil Klasterisasi dengan Teks');
xlabel('Fitur 1');
ylabel('Fitur 2');
grid on;
legend('show'); hold off;
% 6. Analisis Karakteristik Data Berdasarkan Klaster
totalsales = sum(numRounded(:));
marketshares = zeros(optimalClusters, 1);
for i = 1:optimalClusters
clusterdata = numRounded(clusterIdx == i, :);
clustertotalsales = sum(clusterdata(:));
marketshares(i) = clustertotalsales / totalsales * 100;
end
meanmarketshare = mean(marketshares);
stdmarketshare = std(marketshares);
thresholds = [meanmarketshare - 1.5 * stdmarketshare, ...
meanmarketshare - 0.5 * stdmarketshare, ...
meanmarketshare + 0.5 * stdmarketshare, ...
meanmarketshare + 1.5 * stdmarketshare];
fprintf('Threshold Mean dan Standar Deviasi:n');
fprintf('—————n');
fprintf('Sangat Rendah: <
fprintf('Rendah:
fprintf('Sedang:
fprintf('Tinggi:

```

```

fprintf('Sangat Tinggi:
> fprintf('—————n');
% Analisis karakteristik data berdasarkan kluster
fprintf('Analisis Karakteristik Data Berdasarkan Klaster:n');
fprintf('—————n');
for i = 1:optimalClusters
fprintf('Kluster clusterdata = numRounded(clusterIdx == i, :);
clustertotalsales = sum(clusterdata(:));
marketshare = clustertotalsales / totalsales * 100
;
fprintf('Pangsa Pasar Kluster
if marketshare > thresholds(4)
category = 'Sangat Tinggi';
elseif marketshare > thresholds(3)
category = 'Tinggi';
elseif marketshare > thresholds(2)
category = 'Sedang';
elseif marketshare > thresholds(1)
category = 'Rendah';
else category = 'Sangat Rendah';
end
fprintf('Kategori Penjualan: fprintf('—————n');
end
fprintf('—————n');

```

Lampiran 7. Daftar Riwayat Hidup

DAFTAR RIWAYAT HIDUP

1. Identitas Diri

Nama Lengkap : Rahmat Bagus Pangestu
Tempat, Tanggal Lahir : Kota Semarang, 10 Januari 2001
Alamat : Wahyu Asri XII No. C/64 RT 07/RW 06,
Kel. Tambakaji,
Kec. Ngaliyan, Kota Semarang
Nomor HP : 085701142666
Email : rahmatpangestu128@gmail.com

2. Riwayat Pendidikan

- (a) SD Negeri Puwoyoso 03 lulus tahun 2013
- (b) SMP Negeri 18 Semarang lulus tahun 2016
- (c) SMA Negeri 08 Semarang lulus tahun 2019
- (d) UIN Walisongo Semarang

Semarang, 16 Oktober 2024

Rahmat Bagus Pangestu
NIM. 2008046040