LAPORAN UTS

PENERAPAN MODEL LRFMV DAN DEMOGRAFI BEBRBASIS ALGORITMA K-MEANS DALAM SEGMENTASI PERILAKU PELANGGAN RETAIL ONLINE

ANALISIS DATA BISNIS

Kelompok 7

Prithalia Ibra Cardine	187221037
Atria Nur Farradina	187221030
Mochamad Taufiqul Hafizh	082111633002

PROGRAM STUDI S-1 SISTEM INFORMASI

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS AIRLANGGA

2025

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR LAMPIRAN

BAB I
PENDAHULUAN

1.1 Latar Belakang

Pertumbuhan pesat *e-commerce* dalam beberapa tahun terakhir telah mengubah cara pelanggan berbelanja dan mendorong perusahaan untuk menyesuaikan diri dengan kebutuhan pelanggan yang terus berubah karena pelanggan merupakan aset berharga perusahaan (Shankar Awasthi & Professor, 2022). Kondisi ini juga menciptakan persaingan antarperusahaan sehingga perusahaan harus memahami karakteristik, kebiasaan, dan perilaku masing-masing segmen pelanggan untuk menemukan pelanggan baru, menetapkan strategi penting, mengelola hubungan pelanggan, dan meningkatkan keuntungan (Siagian et al., 2021). Tidak hanya itu, faktor demografis seperti usia, jenis kelamin, pendapatan, dan lokasi juga memainkan peran penting dalam mempengaruhi pola belanja dan preferensi pelanggan di platform *online* (Karaniya Wigayha et al., 2025).

Salah satu pendekatan populer dalam analisis perilaku pelanggan adalah model *Length, Recency, Frequency, Monetary, Volume* (LRMFV), yang merupakan pengembangan dari model RFM dan LRFM dengan menambahkan dimensi *Volume* (V) untuk menganalisis hubungan antara kuantitas pembelian dan profitabilitas pelanggan, sehingga memperkaya pemetaan perilaku dan meningkatkan akurasi segmentasi (Mahfuza et al., 2022). Model ini terbukti mampu memberikan gambaran yang lebih akurat mengenai nilai dan aktivitas pelanggan dalam konteks transaksi ritel. Di sisi lain, penggabungan faktor demografis seperti usia, jenis kelamin, dan wilayah ke dalam analisis perilaku pelanggan juga dapat memperkaya pemahaman terhadap keragaman karakter pelanggan serta meningkatkan relevansi hasil segmentasi (Ho et al., 2023). Namun, kedua pendekatan ini masih dikembangkan secara terpisah dan belum diintegrasikan ke dalam kerangka kerja analisis perilaku pelanggan yang menyeluruh.

Selain itu, algoritma K-Means banyak digunakan dalam analisis perilaku dan penelitian segmentasi pelanggan karena kesederhanaan dan efisiensinya dalam mengelompokkan data berbasis perilaku. Namun, algoritma ini hanya optimal untuk data numerik, sementara variabel kategorikal harus terlebih dahulu dikonversi menggunakan *one-hot encoding* agar dapat diproses (Perišić & Pahor, 2023). Untuk mengatasi keterbatasan ini, algoritma K-Prototypes diciptakan, yang secara langsung mengelompokkan data numerik dan kategorikal tanpa modifikasi tambahan dengan menggabungkan konsep K-Means dan K-Modes (Huang, 1997).

Namun, algoritma K-Prototypes belum banyak diterapkan dalam analisis perilaku pelanggan ritel online.

Berdasarkan permasalahan yang telah diidentifikasi, penelitian ini berfokus pada analisis perbandingan efektivitas antara algoritma K-Means dan K-Prototypes dalam mengolah model LRFMV Demografi untuk menghasilkan segmentasi pelanggan yang lebih spesifik pada dataset ritel online Inggris. Tujuan utama dari penelitian ini adalah mengusulkan model LRFMV Demografi yang mampu membantu bisnis memperoleh pemahaman yang lebih mendalam mengenai perilaku sekaligus karakteristik demografis pelanggan. Untuk mencapai tujuan tersebut, kedua algoritma clustering diterapkan secara paralel terhadap dataset model LRFMV, kemudian hasil segmentasi yang dihasilkan dievaluasi dengan metrik *Adjusted Rand Index* (ARI) dan *Adjusted Mutual Information* (AMI) guna mengukur *similarity* dan efektivitas klasterisasi dari masing-masing metode. Selanjutnya, dilakukan *Cohort Analysis* terhadap segmen pelanggan yang terbentuk untuk memberikan wawasan tambahan mengenai dinamika perilaku pelanggan dari waktu ke waktu.

1.2 Rumusan Masalah

Berdasarkan permasalahan yang telah diuraikan, penelitian akan permasalahan terkait bagaimana perbandingan efektivitas antara K-Means dan K-Prototypes dalam mengolah data LRFMV+Demografi untuk menghasilkan segmentasi pelanggan yang spesifik pada dataset ritel online UK?

1.3 Tujuan Penelitian

Tujuan penelitian ini adalah mengusulkan dan menguji model LRFMV-Demografi (LRFMDV) untuk memperoleh pemahaman yang holistik mengenai karakteristik dan perilaku segmen pelanggan.

1.4 Manfaat Penelitian

Manfaat penelitian dibagi menjadi dua, yakni:

1. Secara teoritis, penelitian ini menyajikan perbandingan metodologis untuk segmentasi perilaku pelanggan dalam bisnis retail. Kontribusi utamanya adalah evaluasi kuantitatif menggunakan metrik ARI dan AMI untuk mengukur efektivitas antara algoritma K-Prototypes dan K-Means dalam menangani model LRFM-V + Demografi.
2. Secara praktis, penelitian ini menghasilkan model LRFMV Demografi yang holistik bagi bisnis untuk memahami segmen pelanggan secara mendalam. Selain itu, penerapan Cohort Analysis memberikan wawasan dinamis tentang perubahan perilaku pelanggan.

1.5 Batasan Penelitian

Batasan penelitian akan menegaskan area penelitian yang dilakukan dan mempertajam hasil penelitian. Batasan tersebut, yakni:

1. Penelitian ini terbatas pada dataset "Online Retail" dari Kaggle, yang hanya mencakup transaksi dari satu perusahaan ritel online yang berbasis di Inggris. Data ini juga terbatas secara temporal, yaitu hanya mencakup periode 2009-2010 dan 2010-2011.
2. Model LRFM-Demografi-Volume yang diusulkan sangat bergantung pada ketersediaan data demografis. Dataset yang digunakan tidak memiliki data demografis pelanggan yang sebenarnya sehingga terpaksa menggunakan variabel 'Country' sebagai pendekatan untuk data demografis.
3. Perusahaan dalam dataset ini memiliki model bisnis yang sangat spesifik, yaitu "menjual hadiah unik untuk segala kesempatan" dan melayani banyak "pelanggan grosir". Hasil segmentasi dari penelitian ini tidak dapat digeneralisasi secara langsung ke industri e-commerce lain yang memiliki model bisnis berbeda.
4. Cohort Analysis yang diterapkan bersifat deskriptif. Analisis ini dapat menunjukkan apa yang terjadi pada perilaku segmen dari waktu ke waktu (misalnya, penurunan atau peningkatan retensi), tetapi tidak secara otomatis menjelaskan mengapa perubahan perilaku tersebut terjadi.

BAB II
REPRESENTASI DATA

2.1 Dataset

Dataset yang digunakan dalam penelitian ini adalah "Online Retail II Dataset" yang dapat diakses pada , yaitu data transaksi *non-store online retail* yang berbasis dan terdaftar di Inggris. Dataset ini memuat seluruh transaksi yang terjadi selama periode 1 Desember 2009 hingga 9 Desember 2011, dengan cakupan dua tahun kegiatan penjualan. Perusahaan yang menjadi sumber data ini bergerak dalam bidang penjualan produk *unique all-occasion giftware*, dengan sebagian besar pelanggannya merupakan pembeli grosir.

Dataset ini bersumber dari platform Kaggle, diunggah oleh Bojan Tunguz, berdasarkan data yang dikumpulkan oleh Dr. Daqing Chen dari *London South Bank University*. Secara keseluruhan, dataset ini berisi 1.067.371 baris informasi transaksi individual yang dicatat pada level *invoice*, dengan delapan atribut utama. Deskripsi mendetail untuk setiap atribut disajikan pada Tabel 2. 1.

Tabel 2. 1 Cuplikan Dataset

Nama Atribut	Deskripsi	Tipe Data
InvoiceNo	Nomor faktur transaksi berupa 6 digit unik. Jika diawali huruf "C", berarti transaksi dibatalkan.	Nominal
StockCode	Kode produk berupa 5 digit unik untuk setiap item.	Nominal
Description	Nama produk yang dijual.	Nominal
Quantity	Jumlah unit produk yang dibeli per transaksi.	Numerik
InvoiceDate	Tanggal dan waktu terjadinya transaksi.	Numerik
UnitPrice	Harga per unit barang dalam mata uang pound sterling (£).	Numerik
CustomerID	Kode pelanggan berupa 5 digit unik untuk setiap pelanggan.	Nominal
Country	Negara tempat pelanggan berada.	Nominal

2.2 Pengolahan Data

Pengolahan data adalah serangkaian operasi terstruktur yang mentransformasikan data mentah menjadi informasi yang bermakna dan terorganisir. Operasi ini mencakup pembersihan, transformasi, agregasi, dan penataan data untuk memastikan kualitas, konsistensi, dan relevansi. Tujuan utamanya adalah merepresentasikan data dalam format yang siap dianalisis atau digunakan untuk pengambilan keputusan.

2.2.1 Dimensi Data

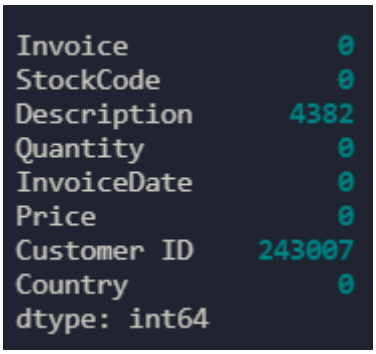
Dimensi data untuk dataset Online Retail tahun 2009-2010 adalah (**525461**, **8**) sementara pada tahun 2010-2011 adalah (**541910**, **8**). Artinya, data penjualan pada tahun 2010-2011 lebih banyak 16.449 dari pada data penjualan tahun 2009-2010. Distribusi data pembeli dicek melalui perhitungan *unique values* pada kolom "Customer ID" yang dapat dilihat pada

Tahun	Jumlah Pembeli
2009-2010	4383
2010-2011	4372

Data kemudian di-merge yang menghasilkan dimensi (**1067371, 8**) dengan total "Customer ID" yang melakukan pembelian berulang sebanyak 5881 pelanggan sejak tahun 2009 hingga 2011.

2.2.2 *Null Handling*

Pengecekan *null* dilakukan pada dataset retail tahun 2009-2011 yang sudah digabungkan. Hasil pengecekan *null* dapat dilihat pada **Gambar 2. 1.**



Gambar 2. 1 Hasil Pengecekan *Null*

"Customer ID" memiliki 243.007 baris *null.* Hal tersebut mengindikasikan bahwa dari 1.067.371 data penjualan retail, terdapat 22.8% transaksi yang tidak memiliki data pelanggan secara valid. *Null handling* dilakukan untuk memastikan bahwa tidak ada "Customer ID" yang kosong. Nilai "Customer ID" kosong mengindikasikan bahwa pembelian tidak terdokumen secara valid sehingga perlu untuk dieliminasi dari penelitian. Hasil dari *null handling* menghasilkan 824.364 data transaksi yang memiliki "Customer ID".

Null pada “Description” dapat diabaikan karena tidak berpengaruh pada penelitian secara langsung. Fungsi kolom “Description” dapat dilihat pada **Tabel 2. 1**.

2.2.3 *Minus Value Handling*

Nilai negatif pada kolom “Quantity” akan dihapus karena mencerminkan transaksi yang tidak sah atau tidak valid. Jumlah transaksi sebelum penghapusan adalah 824.364. Penghapusan dilakukan menggunakan **Algoritma 2. 1**.

Algoritma 2. 1 Validasi Nilai “Quantity”

$$retail = retail[retail["Quantity"] > 0]$$
$$retail["Quantity"].value\_counts().head()$$

Setelah dilakukan seleksi, data transaksi yang tersisa adalah sebanyak 805.620. Artinya, terdapat sebanyak 18.744 transkasi dengan “Quantity” pembelian bernilai negatif.

2.2.4 Seleksi Transaksi Berhasil

Pada “Invoice”, nilai yang diawali dengan huruf kode “C” mengindikasikan bahwa transaksi dibatalkan. Dengan pengetahuan tersebut, maka setelah menghapus transkasi tanpa “Customer ID” dan “Quantity” negatif, dari 805.620 akan dilakukan seleksi transaski berhasil menggunakan **Algoritma 2. 2.**

Algoritma 2. 2 Seleksi Transaksi Berhasil

$$retail = retail[\sim retail["Invoice"].astype(str).$$
$$str.contains("C", na = False)]$$

Seleksi menghasilkan dimensi 805620, artinya tidak ada data transkasi gagal tersisa. Hal tersebut menunjukkan bahwa transaksi gagal telah tereliminasi bersama dengan transaksi tanpa “Customer ID” dan “Quantity” negatif.

2.3 Data *Coding*

Data coding adalah proses mengubah data mentah—baik berupa teks, angka, atau kategori menjadi bentuk yang lebih terstruktur dan terstandarisasi sehingga mudah dipahami dan dianalisis. Proses ini berupa pemberian label, pengelompokan, pengkodean numerik, ataupun transformasi nilai berdasarkan aturan tertentu untuk menyederhanakan data dan memungkinkan perbandingan antarvariabel secara lebih akurat dan sistematis.

2.3.1 Kuantitas

Pelabelan kelompok “Quantity” pembelian dilakukan karena total kuantitas pembelian yang beragam, yakni **438** *unique values*. “Quantity” terbanyak adalah 80.995 barang. Pelabelan kelompok kuantitas dapat dilihat pada **Tabel 2. 2.**

Tabel 2. 2 Pelabelan Kelompok "Quantity"

Kuantitas Pembelian	Label	Jumlah Kelompok
1-5	*Very Small*	406.930
6-50	*Small*	375.303
51-100	*Medium*	13.428
101+	*Bulk*	9.959

2.3.2 Negara

Pelabelan "Country" dilakukan karena kolom tersebut terdiri dari 41 negara, sehingga menyulitkan proses analisis demografi LRMFV tanpa melakukan pengelompokan. Pelabelan kemudian dilakukan berdasarkan benua, dengan pengecualian negara dominan seperti "United Kingdom" yang dipisahkan ke kategori tersendiri karena kontribusi jumlah transaksinya sangat besar dan secara statistik dapat menciptakan bias jika digabungkan dengan kelompok benua lain. Pengelompokkan didasarkan pada **Tabel 2. 3.** Pengelomopokkan menghasilkan 7 label.

Tabel 2. 3 Pelabelan Kelompok "Country"

Negara	Label	Jumlah Transaksi Per Kelompok Negara
United Kingdom, Channel Islands	UK (United Kingdom)	726.865
Europe Community, France, Belgium, EIRE, Germany, Portugal, Denmark, Netherlands, Poland, Spain, Italy, Cyprus, Greece, Norway, Austria, Sweden, Finland, Switzerland, Malta, Lithuania, Czech Republic	Europe	73457
Japan, Korea, Hong Kong, Singapore, Thailand, Lebanon, Saudi Arabia, United Arab Emirates, Bahrain, RSA	Asia	1572
USA, Canada, Brazil, West Indies	America	785
Nigeria	Afrika	30
Iceland	Iceland	253
Unspecified	Other	843

2.3.3 Bulan

Pelabelan bulan transaksi dilakukan sebab data "InvoiceDate" masih dalam format DD/MM/YY. Hal tersebut dapat menyulitkan analisis data sehingga perlu dilakukan perubahan format angka MM ke karakter nama bulan. Hasil dari pelabelan dapat dilihat pada **Tabel 2. 4.**

Tabel 2. 4 Pelabelan "InvoiceDate" berdasarkan Bulan Transaksi

Nomor Bulan	Label	Jumlah Transaksi
1	January	43010
2	February	43297
3	March	59479
4	April	49882
5	May	56966
6	June	58376
7	July	53860
8	August	53406
9	September	74632
10	Oktober	99120
11	November	124861
12	December	88731

2.3.4 Total Pembelian

Total pembelian dihitung dengan mengalikan "Price" terhadap "Quantity". Sebelum dilakukan pelabelan, total pembelian dikelompokkan berdasarkan "Customer ID" untuk memudahkan analisis. Hasil pelabelan terdiri dari 5881 baris dengan detail label dilihat pada **Tabel 2. 5.**

Tabel 2. 5 Pelabelan "Total Spending"

Total Pembelian	Label	Jumlah Kelompok Pembelian
0 - 100£	*Low*	218
100£ - 500£	*Medium*	1822
500£ - 2000£	*High*	2186
> 2000£	*Very High*	1655

2.4 Tabulasi

Tabulasi merupakan proses esensial dalam pengolahan data yang mengubah data mentah menjadi format tabel yang ringkas dan terstruktur. Proses ini menyederhanakan penyajian data agar mudah dibaca dan menunjukkan distribusi atau hubungan antar variabel untuk mempermudah penarikan kesimpulan.

2.4.1 Distribusi Kelompok Kuantitas pada Setiap Bulan Transaksi

Tabulasi "Quantity_Group" terhadap "InvoiceMonth" dapat dilihat pada **Gambar 2. 2**. Hasil tabulasi menunjukkan pola distribusi jumlah unit per transaksi sepanjang tahun. Secara umum, dua kelompok dengan volume terkecil Very Small (1–5) dan Small (6–50)

mendominasi seluruh bulan. Hal tersebut mencerminkan bahwa mayoritas transaksi melibatkan pembelian dalam jumlah kecil.

Quantity_Group	Very Small (1-5)	Small (6-50)	Medium (51-100)	Bulk (101+)
InvoiceMonth				
January	20953	20564	773	720
February	21516	20385	804	592
March	29828	27894	967	790
April	24181	24250	858	593
May	27672	27498	1062	734
June	29563	26961	1044	808
July	25601	26607	993	659
August	24241	27209	1086	870
September	33181	39089	1378	984
October	51564	45062	1414	1080
November	70868	51235	1689	1069
December	47762	38549	1360	1060

Gambar 2. 2 Distribusi Kelompok Kuantitas pada Bulan Transaksi

Peningkatan signifikan terlihat pada bulan September hingga November, terutama pada kategori Very Small dan Small, yang dapat mengindikasikan musim belanja atau peningkatan permintaan menjelang akhir tahun. Sementara itu, kelompok Medium (51–100) dan Bulk (101+) tetap muncul tetapi dalam jumlah jauh lebih rendah, mencerminkan bahwa transaksi dalam jumlah besar relatif jarang. Pola ini menegaskan bahwa perilaku pembelian pelanggan mayoritas bersifat low-volume dan cenderung meningkat pada kuartal akhir tahun.

2.4.2 Distribusi Kelompok Negara pada Setiap Bulan Transaksi

Tabulasi Country_Group terhadap InvoiceMonth dapat dilihat pada **Gambar 2. 3**. Tabulasi ini menunjukkan bahwa United Kingdom (UK) mendominasi transaksi secara signifikan di setiap bulan, dengan volume yang jauh lebih tinggi dibandingkan kelompok negara lainnya. Hal ini konsisten dengan fakta bahwa mayoritas pelanggan dalam dataset berasal dari UK.

Country_Group	Afrika	America	Asia	Europe	Iceland	Other	UK
InvoiceMonth							
January	1	105	102	4793	29	64	37789
February	0	0	272	3773	0	16	39115
March	0	10	87	5276	0	0	53979
April	0	54	139	3824	24	30	45776
May	0	16	51	5418	0	47	51258
June	0	93	53	5684	18	9	52240
July	0	103	212	5000	0	287	48081
August	29	59	75	5533	22	194	47380
September	0	62	149	7691	0	110	66497
October	0	184	165	9670	87	34	88725
November	0	66	129	10179	0	52	114222
December	0	33	138	6616	73	0	81803

Gambar 2. 3 Distribusi Kelompok Negara pada Setiap Bulan Transaksi

Kelompok Europe berada di posisi kedua dengan jumlah transaksi yang stabil dan cukup besar, meskipun tetap jauh di bawah UK. Sementara itu, kelompok America, Asia, Iceland, Africa, dan Other mencatat jumlah transaksi yang relatif kecil dan fluktuatif sepanjang tahun. Peningkatan transaksi yang paling menonjol terjadi pada bulan September hingga November, terutama pada UK dan Europe, yang mengindikasikan adanya peningkatan aktivitas belanja menjelang akhir tahun. Secara keseluruhan, pola ini menegaskan bahwa pasar utama penjualan retail berada di UK, diikuti oleh beberapa negara Eropa, sedangkan kontribusi dari wilayah lain bersifat minor.

2.4.3 Distribusi Kelompok Negara terhadap Kelompok Kuantitas

Hasil tabulasi Quantity_Group terhadap Country_Group bisa dilihat pada **Gambar 2. 4**. Tabulasi menunjukkan bahwa sebagian besar transaksi berasal dari United Kingdom (UK), dengan dominasi yang sangat jelas pada kelompok Very Small (1–5) dan Small (6–50), masing-masing mencapai ratusan ribu transaksi. Hal ini menegaskan bahwa pola pembelian pelanggan dari UK cenderung berorientasi pada pembelian dalam jumlah kecil.

Quantity_Group	Very Small (1-5)	Small (6-50)	Medium (51-100)	Bulk (101+)
Country_Group				
Afrika	27	3	0	0
America	252	527	2	4
Asia	292	1111	113	56
Europe	17612	51071	2481	2293
Iceland	47	205	0	1
Other	310	524	8	1
UK	388098	320885	10576	7306

Gambar 2. 4 Distribusi Kelompok Negara terhadap Kelompok Kuantitas

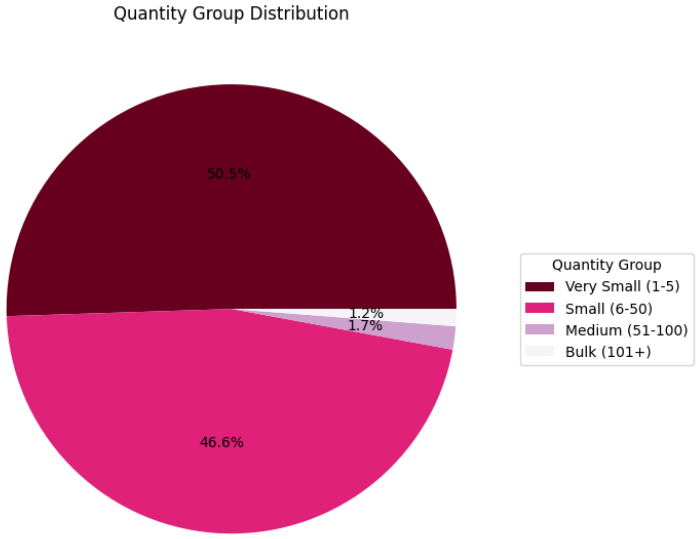
Negara-negara Eropa lain berada di urutan kedua, mencatat volume signifikan terutama pada kategori Small dan Very Small, meskipun tetap jauh lebih rendah dibandingkan UK. Sementara itu, kelompok negara lain seperti America, Asia, Africa, Iceland, dan Other menunjukkan volume transaksi yang jauh lebih kecil dan sebagian besar juga terkonsentrasi pada pembelian dalam jumlah rendah. Transaksi pada kategori Medium (51–100) dan Bulk (101+) muncul tetapi dalam jumlah terbatas, umumnya berasal dari UK dan beberapa negara Eropa.

2.5 Penyajian Data

Penyajian Data menampilkan hasil pengolahan data dalam bentuk yang terstruktur dan mudah dipahami, baik melalui tabel, grafik, maupun visualisasi lainnya. Hasilnya dapat digunakan untuk mengkomunikasikan temuan kunci penelitian secara efektif.

2.5.1 Kuantitas

Grafik distribusi Quantity Group pada **Gambar 2. 5** menunjukkan bahwa sebagian besar transaksi didominasi oleh kelompok pembelian dalam jumlah kecil. Kategori Very Small (1–5) merupakan yang terbesar dengan porsi sekitar 50,5%, diikuti oleh kategori Small (6–50) sebesar 46,6%.

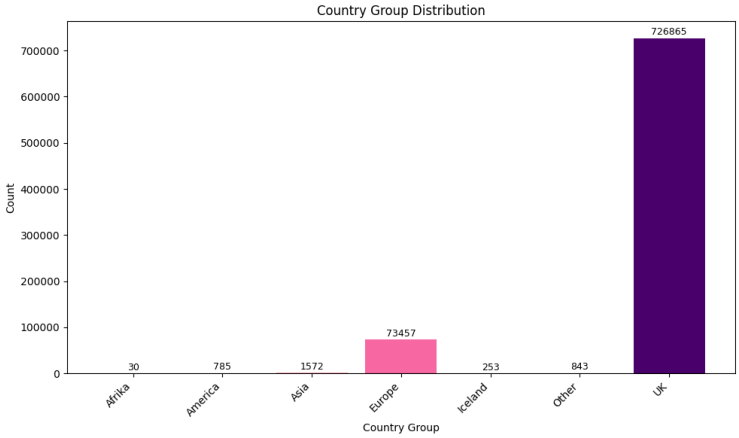


Gambar 2. 5 Distribusi Kelompok Kuantitas

Sementara itu, kategori Medium (51–100) dan Bulk (101+) hanya menyumbang 1,7% dan 1,2%, yang mencerminkan bahwa transaksi dengan jumlah unit lebih besar relatif jarang terjadi. Pola ini menegaskan bahwa mayoritas pelanggan cenderung melakukan pembelian berukuran kecil, sehingga karakteristik penjualan retail pada dataset ini lebih mengarah pada transaksi low-volume.

2.5.2 Negara

Visualisasi “Country Group Distribution” pada **Gambar 2. 6** menunjukkan bahwa distribusi data sangat didominasi oleh kelompok UK, yang memiliki jumlah hitungan atau Count tertinggi 726.865. Kelompok kedua terbesar adalah Europe dengan 73.457 hitungan, namun selisihnya sangat jauh dibandingkan UK.



Gambar 2. 6 Distribusi Kelompok Negara

Sementara itu, kelompok-kelompok lainnya seperti Asia (1.572), Other (843), America (785), Iceland (253), dan Afrika (30) memiliki Count yang sangat kecil, menunjukkan bahwa data yang dianalisis hampir seluruhnya berasal dari atau diklasifikasikan sebagai UK.

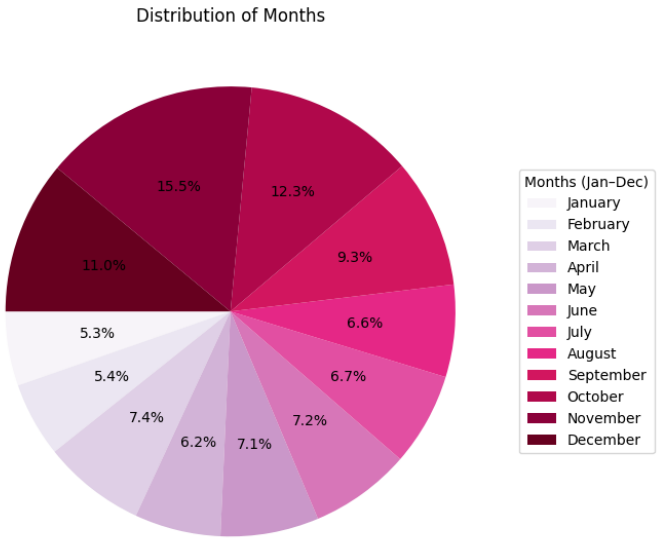
Peta dunia juga dibuat untuk menampilkan distribusi pelanggan di seluruh dunia dapat dilihat pada **Gambar 2. 9**, di mana intensitas warna ungu menunjukkan konsentrasi Jumlah Pelanggan yang lebih tinggi. Berdasarkan visualisasi ini, terlihat jelas bahwa basis pelanggan sangat terkonsentrasi di wilayah Eropa Barat terutama Britania Raya serta di negara-negara maju seperti Amerika Utara, AS dan Kanada. Sebaliknya, sebagian besar wilayah dunia lainnya termasuk sebagian besar Afrika, Asia, dan Amerika Selatan, diwarnai dengan warna putih atau merah muda sangat muda menandakan bahwa jumlah pelanggan di wilayah-wilayah tersebut sangat rendah dibandingkan dengan area yang dominan

Gambar 2. 7 Peta Persebaran Negara

2.5.3 Bulan

Diagram "Distribution of Months" pada **Gambar 2. 8** menunjukkan adanya variasi musiman yang sangat signifikan, dengan data yang sangat terkonsentrasi di akhir tahun. Bulan Desember memiliki proporsi tertinggi sebesar 15.5%, diikuti oleh November (12.3%) dan Oktober (11.0%), yang secara kolektif menyumbang lebih dari sepertiga total distribusi data. Sebaliknya, data menunjukkan titik terendah di sekitar tengah tahun dan awal tahun, di mana

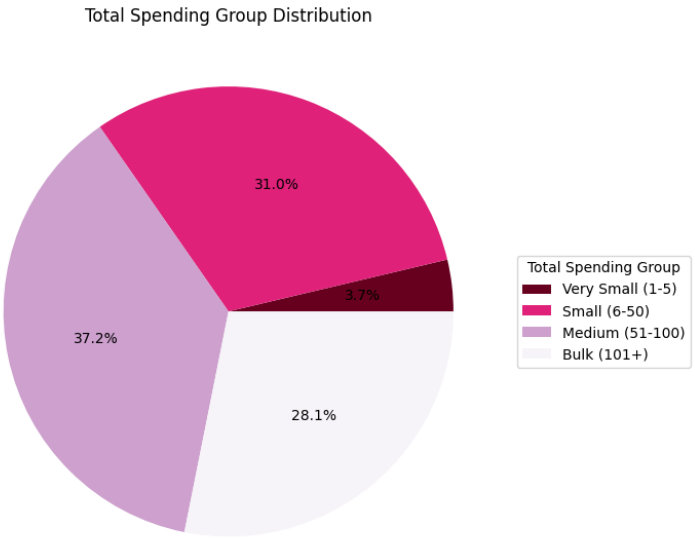
bulan Juni merupakan yang terendah dengan 5.3%, sementara bulan-bulan seperti Mei dan Januari juga berada di bawah 6.5%. Pola ini dengan jelas mengindikasikan bahwa aktivitas atau kejadian yang diukur oleh data ini sangat didominasi oleh bulan-bulan di kuartal terakhir.



Gambar 2. 8 Distribusi Transaksi per Bulan

2.5.4 Total Pembelian

Diagram "Total Spending Group Distribution" pada **Gambar 2. 9** menunjukkan kelompok pembelanjaan terbesar adalah Medium (51-100), yang menyumbang bagian dominan sebesar 37.2% dari total, menunjukkan bahwa mayoritas pembelanjaan berada dalam kisaran menengah ini.



Gambar 2. 9 Distribusi Kelompok Total Pengeluaran

Kelompok terbesar kedua adalah Small (6-50) dengan kontribusi 31.0%, diikuti oleh kelompok Bulk (101+) yang mencakup 28.1% dari total pembelanjaan. Kontrasnya, kelompok Very Small (1-5) hanya menyumbang bagian terkecil, yaitu 3.7%

2.6 Statistika Deskriptif

Proses ini berfokus pada pengumpulan, pengolahan, penyajian, dan analisis ringkasan data tanpa membuat kesimpulan atau generalisasi ke populasi yang lebih luas. Tujuannya adalah untuk mendeskripsikan atau memberikan gambaran karakteristik utama suatu kumpulan data, biasanya melalui ukuran-ukuran seperti nilai rata-rata (*mean*), median, modus, serta dispersi data seperti standar deviasi dan rentang.

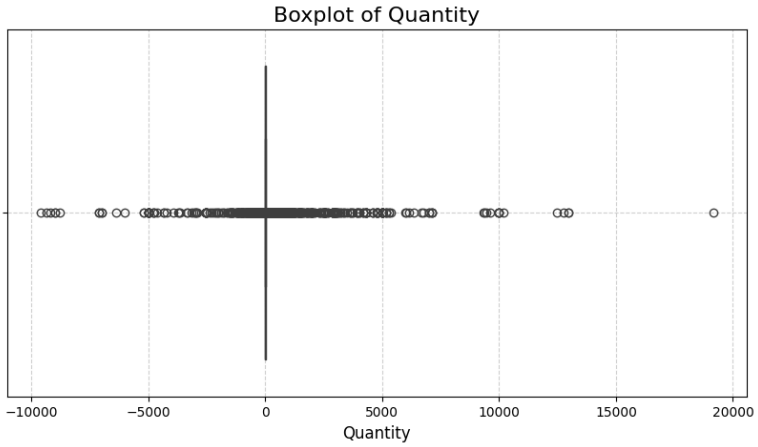
2.6.1 Kuantitas

Data kuantitas mencakup total 805.620 entri transaksi dengan hasil statistika deskriptif dapat dilihat pada **Gambar 2. 10**. Rata-rata (*mean*) kuantitas yang dipesan per baris transaksi adalah 13.31 unit. Namun, nilai median (*50%*) hanya 5 unit dan 75% dari transaksi memiliki kuantitas 12, menunjukkan bahwa distribusi data sangat miring ke kanan (*positively skewed*). Kecenderungan ini dikonfirmasi oleh deviasi standar (*std*) yang sangat tinggi (144.31), jauh melampaui rata-rata yang mengindikasikan adanya *outliers* ekstrem. Kondisi *outlier* tersebut dapat dilihat pada **Gambar 2. 11.**

```
count    805620.000000
mean         13.307665
std         144.306739
min           1.000000
25%           2.000000
50%           5.000000
75%          12.000000
max       80995.000000
Name: Quantity, dtype: float64
```

Gambar 2. 10 Statistika Deskriptif Kuantitas

Outlier ini terlihat dari nilai maksimum (*max*) yang mencapai 80.995 unit. Nilai minimum (*min*) adalah 1, yang mengonfirmasi tidak ada nilai negatif (retur) dalam rangkuman ini.



Gambar 2. 11 Diagram *Boxplot* Kuantitas

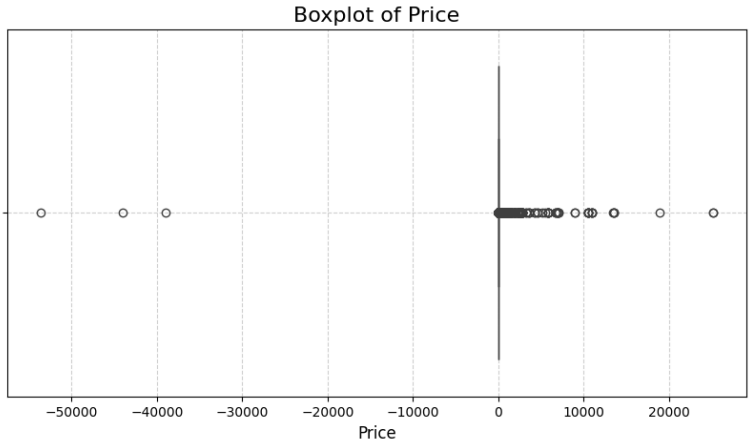
Boxplot Kuantitas menunjukkan bahwa mayoritas data sangat terkonsentrasi di sekitar nilai nol (0) dan memiliki kuantitas kecil. Namun, data didominasi oleh *outliers* ekstrem. Terdapat *outliers* positif yang mengindikasikan pesanan sangat besar dan *outliers* negatif yang masif. Pencilan negatif ini adalah bukti kuat adanya retur atau pembatalan bervolume tinggi yang harus dibersihkan dari *dataset* sebelum analisis metrik bisnis.

2.6.2 Harga Barang

Rata-rata (*mean*) harga adalah **3.21**, namun nilai mediannya (*50%*) hanya **1.95**, dan 75% dari harga berada di bawah **3.75**. Sama seperti kuantitas, adanya perbedaan besar antara rata-rata dan median, serta deviasi standar (**29.19**) yang relatif tinggi, menunjukkan distribusi yang sangat **miring ke kanan** karena adanya pencilan harga yang tinggi. Hal ini terlihat jelas dari nilai maksimum (*max*) yang mencapai **10.953.50**, menunjukkan keberadaan satu atau lebih barang dengan harga yang sangat ekstrem. Nilai minimum (*min*) adalah **0**, yang mengindikasikan adanya item yang mungkin diberikan secara gratis atau merupakan sampel, dan perlu diverifikasi sebelum perhitungan pendapatan dilakukan.

```
count    805620.000000
mean          3.206279
std          29.197901
min           0.000000
25%           1.250000
50%           1.950000
75%           3.750000
max       10953.500000
Name: Price, dtype: float64
```

Gambar 2. 12 Statistika Deskriptif Harga Barang



Gambar 2. 13 Diagram *Boxplot* Harga Barang

Boxplot Harga menunjukkan bahwa mayoritas harga sangat terkonsentrasi di dekat nol dengan rentang yang sangat sempit. Adanya harga negatif yang besar dan beberapa harga positif yang sangat tinggi mengindikasikan nilai negatif harus dihapus sebagai kesalahan *entry* sebelum analisis harga yang valid.

2.6.3 Total Pembelian

```
count      5881.000000
mean       3017.076888
std       14734.128619
min           0.000000
25%         347.800000
50%         897.620000
75%        2304.180000
max      608821.650000
Name: TotalSpend, dtype: float64
```

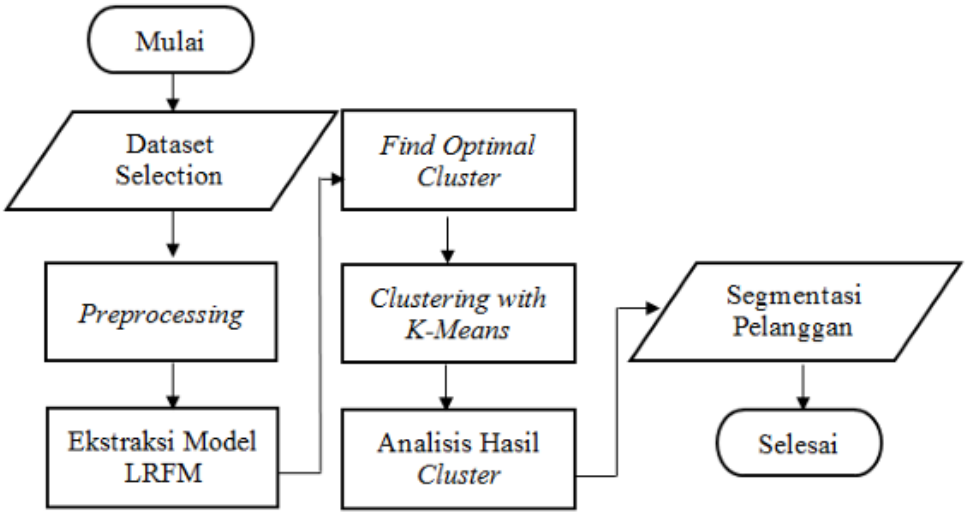
Data Total Pembelian (*TotalSpend*) mencakup 5.881 entri transaksi pelanggan. Data menunjukkan distribusi yang sangat miring ke kanan karena nilai rata-rata (*mean*) sebesar 3.017,08 jauh lebih tinggi daripada nilai median (*50%*) yang hanya 897,62. Mayoritas pembelanjaan pelanggan berada di sisi yang lebih rendah, dengan 75% dari total pembelanjaan berada di bawah 2.304,18. Deviasi standar (*std*) yang sangat besar bersama dengan nilai maksimum (*max*) yang ekstrem mengindikasikan adanya *outliers* di mana pelanggan dengan pengeluaran sangat besar mendistorsi rata-rata. Nilai minimum (*min*) adalah 0, menunjukkan adanya kemungkinan transaksi yang batal atau data yang tidak sempurna.

BAB III
METODOLOGI

3.1 Metodologi Paper

Penelitian berjudul *"E-Commerce Customer Segmentation Using K-Means Algorithm and Length, Recency, Frequency, Monetary (LRFM) Model"* yang dilakukan oleh Siagian, A., Ginting, F., dan Fadlisyah, A. (2021) bertujuan untuk mengelompokkan pelanggan *e-commerce* berdasarkan perilaku pembelian dan mengusulkan penerapan model LRFM (*Length, Recency, Frequency, Monetary*) sebagai dasar dalam segmentasi pelanggan *e-commerce* menggunakan algoritma K-Means Clustering.

Penelitian ini menggunakan Dataset dari UCI Machine Learning Repository yang memuat transaksi penjualan daring perusahaan berbasis di Inggris pada periode 1 Desember 2010 hingga 9 Desember 2011. Dataset dapat diakses pada **Lampiran 1**. Dataset ini terdiri dari 541.909 data dengan 8 atribut data. Detail dataset telah dijelaskan pada **Tabel 2. 1.** Metode penelitian yang digunakan pada paper untuk menghasilkan segmentasi pelanggan *e-commerce* digambarkan dalam bentuk *flowchart* yang terdapat pada **Gambar 3. 1**.



Gambar 3. 1 Kerangka kerja yang digunakan dalam paper

3.1.1 Dataset Selection

Pada tahap ini dilakukan pemilihan data transaksi pelanggan dari *Online Retail Dataset* yang relevan dengan kebutuhan analisis, yaitu atribut yang dapat dikonversi menjadi variabel perilaku pelanggan. Pemilihan difokuskan pada kolom *CustomerID*, *InvoiceDate*, *Quantity*, dan *UnitPrice* untuk membentuk model perilaku berdasarkan aktivitas pembelian pelanggan.

3.1.2 Preprocessing

Data yang telah dipilih kemudian melalui proses pembersihan untuk menjamin validitas analisis. Langkah-langkah yang dilakukan meliputi penghapusan data duplikat, nilai nol atau negatif pada Quantity dan UnitPrice, serta baris tanpa CustomerID. Deteksi outlier dilakukan menggunakan metode Interquartile Range (IQR), dan seluruh variabel numerik dinormalisasi menggunakan teknik Min-Max Scaling agar berada dalam rentang 0–1 untuk menghindari dominasi satu variabel terhadap variabel lainnya.

3.1.3 Ekstraksi Model LRFM

Tahap ini mengubah data transaksi mentah menjadi representasi perilaku pelanggan menggunakan empat dimensi utama:

a. *Length* (L): jangka waktu antara transaksi pertama dan terakhir pelanggan,
b. *Recency* (R): waktu sejak transaksi terakhir hingga tanggal acuan analisis,
c. *Frequency* (F): jumlah transaksi yang dilakukan pelanggan, dan
d. *Monetary* (M): total nilai pembelian yang dihitung dari hasil perkalian *Quantity* × *UnitPrice*.

Proses ini menghasilkan satu entri data per pelanggan yang mencerminkan pola interaksi dan kontribusi finansial mereka terhadap perusahaan.

3.1.4 Penentuan Jumlah *Cluster* Optimal

Untuk menentukan jumlah klaster yang paling sesuai, digunakan tiga metode evaluasi yaitu *Elbow Method*, *Silhouette Coefficient*, dan *Davies-Bouldin Index (DBI)*. Nilai *k* optimal dipilih berdasarkan titik keseimbangan antara variasi dalam klaster dan pemisahan antar klaster.

3.1.5 Clustering with K-Means

Dengan nilai *k* yang telah ditentukan, data pelanggan diklasterisasi menggunakan algoritma *K-Means* dengan jarak *Euclidean* sebagai ukuran kesamaan. Proses iterasi dilakukan hingga posisi pusat *centroid* stabil dan tidak terjadi perubahan signifikan antar iterasi.

3.1.6 Analisis Hasil *Cluster*

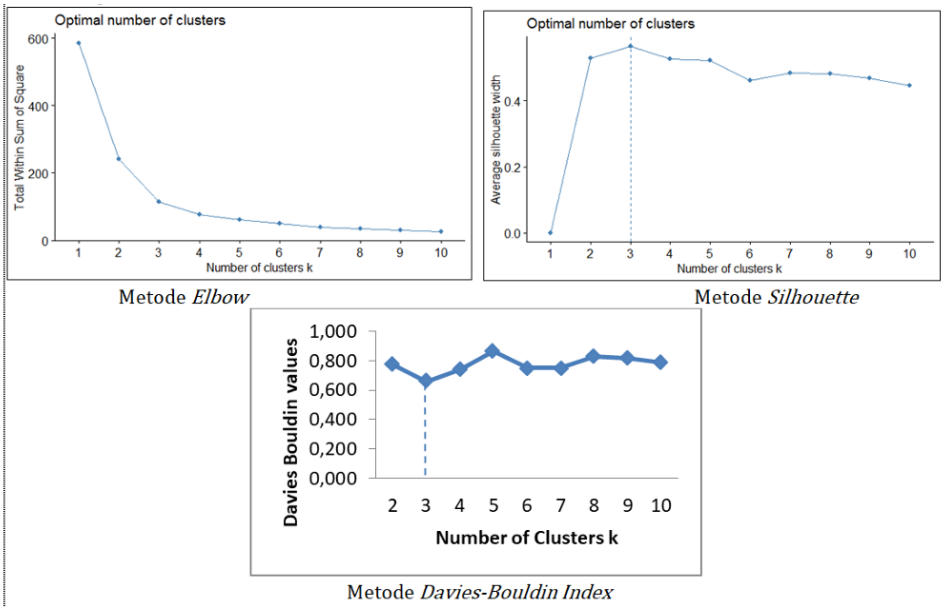
Setiap klaster dianalisis dengan menghitung rata-rata dan deviasi standar untuk masing-masing variabel LRFM. Hasil tersebut dibandingkan dengan rata-rata keseluruhan dataset untuk mengidentifikasi karakteristik dominan tiap klaster. Tanda peningkatan (↑) atau penurunan (↓) digunakan untuk menggambarkan kecenderungan perilaku pelanggan pada masing-masing segmen.

3.1.7 Segmentasi Pelanggan

Tahap akhir adalah interpretasi hasil klasterisasi ke dalam Customer Value Matrix dan Customer Loyalty Matrix. Melalui pemetaan ini, setiap klaster diberi label sesuai perilaku dominannya, yaitu *New Customers* (pelanggan baru), *Lost Customers* (pelanggan tidak aktif), dan *Core Customers* (pelanggan loyal). Segmentasi ini digunakan untuk memberikan pemahaman yang lebih jelas tentang distribusi perilaku pelanggan serta menjadi dasar bagi strategi pemasaran yang lebih terarah.

3.1.8 Hasil

Hasil dari penerapan algoritma K-Means terhadap model LRFM (*Length, Recency, Frequency, Monetary*) menghasilkan pembentukan tiga klaster (k = 3) sebagai konfigurasi optimal berdasarkan hasil evaluasi menggunakan *Elbow Method, Silhouette Coefficient, dan Davies-Bouldin Index* pada gambar 3. Nilai k = 3 dipilih karena memberikan keseimbangan terbaik antara homogenitas dalam klaster dan perbedaan antar klaster. Visualisasi hasil pencarian K-Means Optimal dapat dilihat pada **Gambar 3. 2.**



Gambar 3. 2 Visualisasi Grafik Metode Penentuan Cluster

Hasil proses klasterisasi pada **Gambar 3. 3** yang menghasilkan data pelanggan terbagi menjadi tiga kelompok dengan jumlah anggota masing-masing:

- Cluster 1: 1.404 pelanggan
- Cluster 2: 878 pelanggan
- Cluster 3: 1.324 pelanggan

```
K-means clustering with 3 clusters of sizes 1404, 878, 1324

Cluster means:
          LN        RN          FN          MN
1 0.08010816 0.1341800 0.004905234 0.002178432
2 0.06782414 0.6186037 0.003306823 0.001624878
3 0.68118586 0.1151438 0.010083553 0.004866305
```

Gambar 3. 3 Hasil *Clustering* K-Means

Setiap klaster memiliki nilai rata-rata dan deviasi standar yang berbeda untuk masing-masing variabel LRFM, yang menunjukkan adanya variasi pola transaksi pelanggan terlihat pada **Gambar 3. 4.**

Cluster	Jumlah Pelanggan	L	R	F	M	Simbol LRFM
1	1404	0,115882585	0,100009328	0,004872944	0,002067239	L↓ R↓ F↓ M↓
2	878	0,12101734	0,13877689	0,003497927	0,001679752	L↓ R↑ F↓ M↓
3	1324	0,17110053	0,102759397	0,006653794	0,003110657	L↑ R↓ F↑ M↑
Rata-rata	3606	0,136000152	0,113848539	0,005008221	0,002285883	

Gambar 3. 4 Metriks Nilai Model LRFM

Secara umum, nilai *Frequency* dan *Monetary* menjadi pembeda utama antar klaster, sementara *Recency* dan *Length* membantu mengidentifikasi tingkat aktivitas pelanggan dalam jangka waktu tertentu. Visualisasi hasil klasterisasi ditampilkan dalam bentuk grafik dua dimensi berdasarkan kombinasi variabel LRFM, yang menunjukkan pemisahan kelompok pelanggan secara jelas.

Berdasarkan hasil klasterisasi yang dihasilkan oleh algoritma *K-Means*, pelanggan terbagi ke dalam tiga segmen utama yang merepresentasikan variasi perilaku dan nilai bisnis berbeda. Interpretasi hasil analisis variabel LRFM menunjukkan bahwa setiap klaster memiliki pola interaksi dan tingkat kontribusi finansial yang khas terhadap perusahaan. Setiap klaster merepresentasikan segmen pelanggan dengan perilaku dan nilai bisnis yang berbeda, yaitu:

1. Cluster 1 diidentifikasi sebagai kelompok pelanggan baru (*New Customers*). Klaster ini memiliki nilai *Length*, *Frequency*, dan *Monetary* yang relatif rendah, menunjukkan bahwa pelanggan dalam kelompok ini baru melakukan sedikit transaksi dengan nilai pembelian kecil. Aktivitas mereka masih terbatas, sehingga berpotensi menjadi target untuk strategi promosi dan peningkatan loyalitas.
2. Cluster 2 menggambarkan pelanggan tidak aktif (*Lost Customers*). Nilai *Recency* pada klaster ini tinggi, menandakan bahwa pelanggan sudah lama tidak bertransaksi. Sementara itu, *Frequency* dan *Monetary* rendah, menunjukkan adanya penurunan minat

atau keterlibatan dalam pembelian. Segmen ini penting untuk diidentifikasi agar perusahaan dapat merancang strategi retensi, seperti program reaktivasi pelanggan.

3. Cluster 3 dikategorikan sebagai pelanggan inti (*Core Customers*). Klaster ini memiliki nilai *Frequency* dan *Monetary* tertinggi dengan *Recency* rendah, yang berarti pelanggan dalam kelompok ini melakukan pembelian berulang dan baru-baru ini bertransaksi. Mereka merupakan pelanggan paling loyal dan bernilai tinggi yang memberikan kontribusi signifikan terhadap pendapatan perusahaan.

Melalui pemetaan hasil ke dalam *Customer Value Matrix* dan *Customer Loyalty Matrix*, penelitian ini menegaskan bahwa model LRFM efektif dalam menggambarkan tingkat loyalitas dan nilai pelanggan. Hasil ini memberikan dasar strategis bagi perusahaan untuk membedakan perlakuan terhadap tiap segmen, fokus pada retensi pelanggan inti, reaktivasi pelanggan hilang, dan peningkatan keterlibatan pelanggan baru. Dengan demikian, penelitian ini menunjukkan bahwa pendekatan berbasis data dapat digunakan untuk memahami perilaku pelanggan secara lebih komprehensif dan mendukung pengambilan keputusan pemasaran yang lebih tepat sasaran.

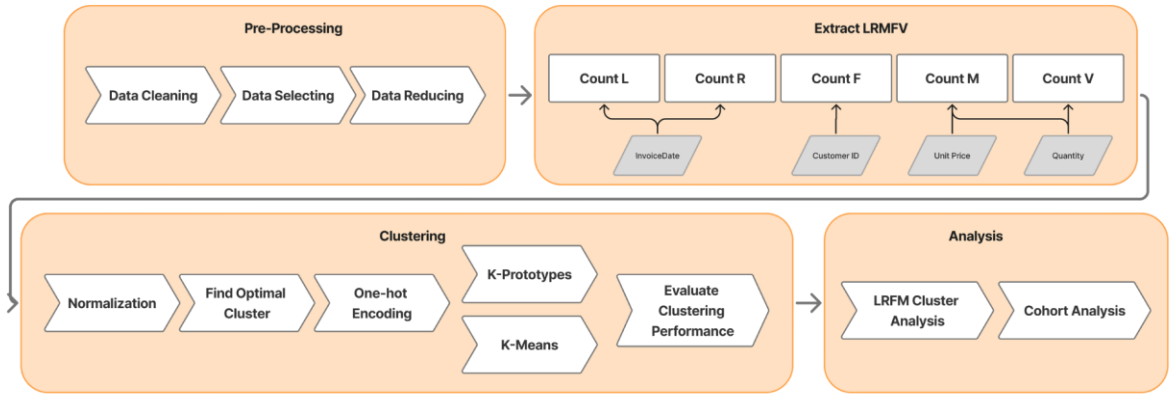
3.2 Metodologi Keterbaruan

Model transaksional tradisional, seperti RFM (*Recency, Frequency, Monetary*), telah lama menjadi standar namun seringkali dianggap kurang memadai untuk mendapatkan pemahaman yang holistik dan mendalam. Segmentasi yang lebih kaya memerlukan integrasi data yang lebih luas, menggabungkan data perilaku transaksional yang diperluas menjadi LRFMV (*Length, Recency, Frequency, Monetary, Value)* dengan data Demografis pelanggan. Penggunaan model LRFMV terbukti sebagai pendekatan yang efisien untuk segmentasi pelanggan (Mahfuza R. et al., 2022), sementara perluasan model RFM dengan analisis demografi juga ditekankan untuk pemahaman perilaku pelanggan yang lebih komprehensif (Thanh Ho et al., 2023).

Perbandingan metodologis ini akan dievaluasi secara kuantitatif, sebagaimana dicantumkan menggunakan metrik ARI (*Adjusted Rand Index*) dan AMI (*Adjusted Mutual Information*) untuk mengukur Performa Clustering antara K-Means dengan K-Prototypes. Selain itu, untuk melengkapi analisis statis hasil segmentasi berdasarkan *LRFM Index* dan *Customer Loyalty Matrix*, penelitian ini akan menerapkan *Cohort Analysis* pada segmen yang terbentuk. Hal ini bertujuan untuk memberikan wawasan dinamis yang berharga tentang bagaimana perilaku segmen pelanggan berkembang dari waktu ke waktu. Penelitian ini

merumuskan masalah utama utama, yakni bagaimana perbandingan efektivitas antara K-Means dan K-Prototypes dalam mengolah data LRFMV+Demografi untuk menghasilkan segmentasi pelanggan yang spesifik pada dataset ritel online UK?

Metodologi yang diajukan dapat dilihat pada **Gambar 3. 5.** Penelitian ini diawali dengan fase Pre-Processing, mencakup *Data Cleaning* untuk menghilangkan inkonsistensi, *Data Selecting* untuk memilih variabel yang relevan dan *Data Reducing* untuk memproses *dataset* mentah. Selanjutnya, dilakukan fase *Extract LRFMV* untuk mendapatkan fitur segmentasi pelanggan, di mana *Length* (L) dan *Recency* (R) dihitung dari "InvoiceDate", *Frequency* (F) dari "Customer ID", *Monetary* (M) dari "UnitPrice" dikali "Quantity", dan *Value* (V) dari "Quantity".



Gambar 3. 5 Metodologi Keterbaruan

Data hasil ekstraksi ini kemudian memasuki fase *Clustering* yang didahului oleh *Normalization* dan penentuan jumlah kluster optimal (*Find Optimal Cluster*) menggunakan *Elbow, Sillhoute Score,* dan *Davis-Bouldien Index*. Untuk perbandingan, data "Country" disiapkan melalui *One-hot Encoding* sebelum diterapkan pada K-Means, sementara K-Prototypes dapat memproses data campuran secara langsung. Kinerja kedua algoritma kemudian dievaluasi menggunakan ARI dan AMI. Terakhir, segmen yang terbentuk dianalisis secara mendalam melalui *LRFM Cluster Analysis* dan diperkaya dengan analisis dinamis menggunakan *Cohort Analysis* untuk memahami perilaku pelanggan dari waktu ke waktu.

3.2.1 Pre-Processing

Berdasarkan analisis deskriptif data dan visualisasi terutama *boxplot* Kuantitas dan Harga, terlihat jelas adanya masalah kualitas data seperti nilai negatif dan *outliers* ekstrem pada variabel kunci. Oleh karena itu, sebelum dapat melakukan analisis lanjutan atau pemodelan

yang akurat Pra-Pemrosesan berfokus pada penghapusan data negatif dan nol yang tidak valid serta penanganan *outliers* untuk memastikan data yang digunakan benar-benar merepresentasikan transaksi pembelian yang valid.

A. Data Cleaning

Proses *Data Cleaning* merupakan langkah awal krusial yang bertujuan untuk memastikan kualitas dan konsistensi data. Kegiatan utama meliputi penanganan nilai yang hilang (*missing values*), penghapusan duplikasi data, dan identifikasi serta penanganan *outliers* yang dapat memengaruhi hasil klastering. Fokus utama adalah pada pembersihan data transaksi, seperti memastikan nilai "Quantity" dan "UnitPrice" adalah positif, serta memvalidasi kelengkapan data kunci seperti "InvoiceNo", "CustomerID", dan "InvoiceDate" agar data siap digunakan untuk ekstraksi fitur LRFM. Proses ini dijabarkan lebih lengkap pada **Bab 2.**

B. Data Selecting

Setelah data dibersihkan, tahap *Data Selecting* dilakukan untuk memfokuskan analisis hanya pada data yang relevan dengan tujuan penelitian, yaitu segmentasi pelanggan ritel *online* di Inggris (*UK*). Pemilihan data ini mencakup penyaringan transaksi yang dilakukan di wilayah UK serta memastikan ketersediaan variabel-variabel yang dibutuhkan untuk pembentukan model LRFMV dan analisis Demografi, yaitu InvoiceDate (untuk L dan R), CustomerID (untuk F), UnitPrice (untuk M), Quantity (untuk V), dan informasi geografis/demografis (Country).

C. Data Reducing

Langkah Data Reducing dalam konteks ini berfungsi ganda sebagai penyelesaian pembersihan data dan fokus pada data yang dapat diatribusikan ke pelanggan. Proses ini melibatkan penghapusan semua transaksi yang tidak memiliki *Customer ID* (nilai kosong/hilang), karena transaksi tersebut tidak dapat digunakan untuk membangun profil pelanggan individu (LRFMV). Selain itu, dilakukan eliminasi data transaksi di mana Quantity bernilai negatif atau nol, karena ini mengindikasikan transaksi pengembalian barang (*returns*) atau data tidak valid yang dapat mendistorsi perhitungan *Monetary* (M) dan *Value* (V) dari pelanggan.

3.2.2 *Extract LRMFV*

Setelah menyelesaikan seluruh proses *Data Cleaning* dan *Data Reducing* pada data transaksi mentah, langkah selanjutnya adalah ekstraksi fitur LRFMV yang telah dijelaskan, yakni menghitung *Length* (L), *Recency* (R), *Frequency* (F), *Monetary* (M), dan *Value* (V) untuk setiap pelanggan. Hasil dari fase ini adalah sebuah *dataset* matriks fitur di level pelanggan, di mana setiap baris mewakili satu pelanggan dan setiap kolom berisi nilai LRFMV serta variabel Demografi (misalnya, Country). *Dataset* akhir ini, yang memiliki tipe campuran (numerik

LRFMV dan kategorikal Demografi), kemudian menjadi input utama untuk fase Clustering, diawali dengan Normalization agar semua fitur memiliki skala yang seragam, memastikan tidak ada fitur yang mendominasi perhitungan jarak, serta mempersiapkan data untuk proses segmentasi menggunakan algoritma K-Means dan K-Prototypes.

A. *Count L (Length)*

Length (L) atau Durasi Pelanggan mengukur lama waktu sejak pelanggan pertama kali melakukan transaksi hingga tanggal terakhir analisis (atau tanggal *snapshot*). Penghitungan L dilakukan dengan mencari tanggal transaksi pertama (InvoiceDate minimum) untuk setiap Customer ID dan menghitung selisih waktu antara tanggal tersebut dengan tanggal *snapshot* yang ditentukan. Nilai L yang lebih tinggi menunjukkan pelanggan yang telah lama terikat dengan bisnis.

B. *Count R (Recency)*

Recency (R) mengukur seberapa baru pelanggan melakukan transaksi terakhir. Penghitungan R dilakukan dengan mengambil tanggal transaksi terbaru (InvoiceDate maksimum) untuk setiap Customer ID dan menghitung selisih waktu (biasanya dalam hari) antara tanggal tersebut dengan tanggal *snapshot* analisis. Nilai R yang rendah (dekat dengan nol) mengindikasikan pelanggan yang baru-baru ini aktif dan lebih mungkin merespons promosi.

C. *Count F (Frequency)*

Frequency (F) mengukur seberapa sering pelanggan melakukan pembelian. Penghitungan F didasarkan pada jumlah total transaksi unik (InvoiceNo yang berbeda) yang dilakukan oleh setiap Customer ID. Nilai F yang tinggi mencerminkan pelanggan yang sangat loyal dan sering berinteraksi dengan platform ritel.

D. *Count M (Money)*

Monetary (M) mengukur jumlah uang total yang telah dihabiskan oleh pelanggan. Penghitungan M dilakukan dengan mengakumulasikan total nilai moneter dari semua transaksi (UnitPrice dikalikan Quantity lalu dijumlahkan) untuk setiap Customer ID. M menunjukkan daya beli dan kontribusi finansial pelanggan terhadap pendapatan bisnis.

E. *Count V (Value)*

Value (V) dalam model LRFMV yang diperluas ini mengukur total kuantitas (volume) barang yang dibeli oleh pelanggan. Penghitungan V dilakukan dengan mengakumulasikan total Quantity dari semua item yang dibeli oleh setiap Customer ID. Fitur ini memberikan dimensi tambahan mengenai volume konsumsi, membedakannya dari nilai moneter murni (M), yang

penting untuk memahami preferensi pembelian dalam jumlah besar (volume transaksi) oleh segmen tertentu.

3.2.3 *Clustering*

Setelah menyelesaikan proses ekstraksi, didapatkan *dataset* di level pelanggan yang terdiri dari fitur numerik LRFMV dan variabel kategorikal Demografi. *Dataset mixed-type* ini kemudian menjadi input untuk fase *Clustering* untuk menjalani normalisasi fitur numerik dan *One-Hot Encoding* untuk kategorikal.

A. *Normalization*

Proses normalisasi adalah langkah pra-pemrosesan data numerik (LRFMV) yang vital sebelum algoritma *clustering* dijalankan. Tujuannya adalah untuk menyeragamkan skala dari semua variabel numerik tersebut, sehingga tidak ada satu fitur pun (misalnya, *Monetary* yang nilainya besar) yang mendominasi perhitungan jarak. Dengan menormalisasi data, semua variabel memiliki bobot yang setara dalam penentuan klaster, yang pada gilirannya meningkatkan akurasi dan keadilan hasil segmentasi.

B. *Find Optimal Cluster*

Langkah pencarian kluster optimal bertujuan untuk menentukan jumlah kelompok (k) yang paling representatif dalam *dataset*. Jumlah k yang tepat akan menghasilkan klaster yang kohesif secara internal dan terpisah secara eksternal. Penentuan ini dilakukan melalui perbandingan hasil dari tiga metrik evaluasi yang berbeda: *Elbow Method*, *Silhouette Score*, dan *Davies-Bouldin Index* (DBI). Penggunaan berbagai metrik memastikan bahwa pemilihan k bersifat kuat dan tidak bias.

C. *One-Hot Encoding* Demografi

One-Hot Encoding (OHE) adalah teknik yang diterapkan secara spesifik pada variabel kategorikal Demografi, seperti "Country", sebagai persiapan untuk algoritma K-Means. Karena K-Means hanya beroperasi pada data numerik, OHE mengubah setiap kategori menjadi kolom biner. Langkah ini diperlukan untuk memungkinkan perbandingan kinerja antara K-Means (yang memerlukan data yang sepenuhnya dinumerisasi) dan K-Prototypes (yang mampu menangani data kategorikal secara langsung).

D. K-Means & K-Prototypes

Pada tahap ini, *dataset* akan diuji coba menggunakan dua algoritma *clustering* utama. K-Means diterapkan pada *dataset* LRFMV yang telah dinormalisasi dan variabel Demografi yang telah di-OHE. Sementara itu, K-Prototypes diterapkan pada *dataset* LRFMV yang dinormalisasi dan variabel Demografi kategorikal asli. Tujuannya adalah untuk

membandingkan efektivitas dan hasil segmentasi yang dihasilkan oleh kedua algoritma dalam menangani model LRFMDV.

E. *Evaluate Clustering Performance*

Langkah terakhir dari fase *clustering* adalah mengevaluasi hasil klustering. Evaluasi ini dilakukan untuk mengukur kesamaan antara hasil klastering yang dihasilkan oleh K-Means dan K-Prototypes. Metrik yang digunakan adalah *Adjusted Rand Index* (ARI) dan *Adjusted Mutual Information (AMI)*. Kedua metrik ini memberikan skor kuantitatif antara 0 dan 1, di mana skor yang lebih tinggi menunjukkan tingkat persetujuan yang lebih besar antar hasil *clustering*, menjadi kontribusi utama dalam perbandingan metodologis penelitian ini.

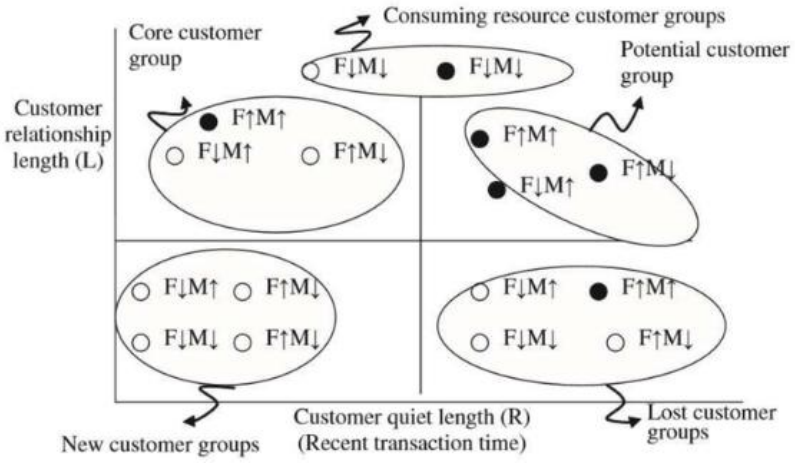
3.2.4 *Analysis*

Pada fase ini, *dataset* yang telah dikelompokkan akan diinterpretasikan secara mendalam melalui LRFM *Cluster Analysis* untuk memberi label pada segmen berdasarkan matriks loyalitas. Analisis ini kemudian dilengkapi dengan *Cohort Analysis* guna memberikan wawasan dinamis mengenai evolusi perilaku dan retensi setiap segmen pelanggan seiring berjalannya waktu.

A. LRFM *Cluster Analysis*

Setelah proses *clustering* selesai dan segmen pelanggan telah terbentuk, LRFM Cluster Analysis dilakukan untuk menginterpretasikan dan memberikan label deskriptif pada setiap klaster. Analisis ini menggunakan metrik LRFMV yang dinormalisasi dan dihitung dengan membandingkan nilai rata-rata (mean) dari setiap aspek L, R, F, M, dan V dalam setiap klaster terhadap nilai rata-rata keseluruhan *dataset*. Berdasarkan perbandingan ini (apakah nilai klaster berada *di atas* atau *di bawah* rata-rata *dataset*), setiap klaster diklasifikasikan ke dalam kategori segmen pelanggan yang bermakna. Proses ini memungkinkan identifikasi karakteristik unik setiap segmen berdasarkan pola perilaku transaksi mereka.

Untuk memperkaya interpretasi klaster, analisis ini akan mengadopsi kerangka *Customer Loyalty Matrix* (Chang., et al. 2004) pada **Gambar 3. 6.** Matriks ini membagi pelanggan berdasarkan dimensi *Customer Relationship Length* (L) pada sumbu vertikal dan *Customer Quiet Length* (R) atau waktu transaksi terbaru pada sumbu horizontal.



Gambar 3. 6 *Customer Loyalty Matrix* (berdasarkan Cheng., et al. 2004)

Setiap kuadran matriks ini merepresentasikan kelompok pelanggan yang berbeda, seperti "*Core customer group*", "*New customer groups*", "*Potential customer group*", "*Consuming resource customer groups*", dan "*Lost customer groups*", berdasarkan kombinasi L dan R, serta indikator Frequency (F) dan Monetary (M). Dengan memplot posisi rata-rata setiap klaster pada matriks ini, kita dapat secara visual mengklasifikasikan dan memahami tingkat loyalitas serta potensi setiap segmen pelanggan secara lebih mendalam.

B. *Cohort Analysis*

Sebagai pelengkap dari analisis klaster statis, Cohort Analysis diterapkan pada segmen yang telah diidentifikasi. *Cohort Analysis* merupakan metode untuk melacak dan memvisualisasikan bagaimana perilaku sekelompok pelanggan (kohort) berubah dari waktu ke waktu. Kohort biasanya didefinisikan berdasarkan bulan akuisisi atau karakteristik segmen awal. Penerapan ini memungkinkan peneliti untuk memperoleh wawasan dinamis, misalnya, bagaimana tingkat retensi (*retention rate*) atau frekuensi pembelian suatu segmen tertentu berkembang setelah mereka diklasifikasikan, memberikan pemahaman yang lebih kaya mengenai nilai seumur hidup (*Lifetime Value*) pelanggan.

DAFTAR PUSTAKA

[1] Ho, T., Nguyen, S., Nguyen, H., Nguyen, N., Man, D. S., & Le, T. G. (2023). An Extended RFM Model for Customer Behaviour and Demographic Analysis in Retail Industry. *Business Systems Research*, *14*(1), 26–53. https://doi.org/10.2478/bsrj-2023-0002

[2] Huang, Z. (1997). CLUSTERING LARGE DATA SETS WITH MIXED NUMERIC AND CATEGORICAL VALUES. *Open Journal of Applied Sciences, Vol.5 No.6, June 16, 2015*. https://api.semanticscholar.org/CorpusID:3007488

[3] Karaniya Wigayha, C., Rolando, B., & Wijaya, A. J. (2025). *A DEMOGRAPHIC ANALYSIS OF CONSUMER BEHAVIORAL PATTERNS ON DIGITAL E-COMMERCE PLATFORMS* (Vol. 1, Issue 2). https://journal.dinamikapublika.id/index.php/Jumder

[4] Mahfuza, R., Islam, N., Toyeb, M., Emon, M. A. F., Chowdhury, S. A., & Alam, M. G. R. (2022). LRFMV: An efficient customer segmentation model for superstores. *PLoS ONE*, *17*(12 December). https://doi.org/10.1371/journal.pone.0279262

[5] Perišić, A., & Pahor, M. (2023). Clustering mixed-type player behavior data for churn prediction in mobile games. *Central European Journal of Operations Research*, *31*(1), 165–190. https://doi.org/10.1007/s10100-022-00802-8

[6] Shankar Awasthi, K., & Professor, A. (2022). PROCEEDINGS OF INTERNATIONAL CONFERENCE ON "INTERNATIONAL CONFERENCE ON FOSTERING INNOVATIONS USING CONSUMER BEHAVIOR ANALYSIS IN THE E-COMMERCE PROCEEDINGS OF INTERNATIONAL CONFERENCE ON "INTERNATIONAL CONFERENCE ON FOSTERING INNOVATIONS USING.

[7] Siagian, R., Pahala Sirait, P. S., & Halima, A. (2021). E-Commerce Customer Segmentation Using K-Means Algorithm and Length, Recency, Frequency, Monetary Model. *JOURNAL OF INFORMATICS AND TELECOMMUNICATION ENGINEERING*, *5*(1), 21–30. https://doi.org/10.31289/jite.v5i1.5182

LAMPIRAN

Lampiran 1 Dataset *Online Retail* Tahun 2009-2011

Dataset Online Retail	https://www.kaggle.com/datasets/lakshmi25npathi/online-retail-dataset