LAPORAN PRAKTIK KERJA LAPANGAN KONVERSI STUDI INDEPENDEN MBKM

ANALISIS SEGMENTASI PELANGGAN MODEL RFM DENGAN MENGGUNAKAN METODE PENGKLASTERAN K-MEANS

STUDI INDEPENDEN DI YAYASAN BAKTI ACHMAD ZAKY

RFM MODEL CUSTOMER SEGMENTATION ANALYSIS USING K-MEANS CLUSTERING METHOD

INDEPENDENT STUDY AT YAYASAN BAKTI ACHMAD ZAKY

MUHAMMAD FATHI FARHAT

24010120130115

DEPARTEMEN MATEMATIKA

FAKULTAS SAINS DAN MATEMATIKA

UNIVERSITAS DIPONEGORO

SEMARANG

2022

LAPORAN PRAKTIK KERJA LAPANGAN KONVERSI STUDI INDEPENDEN MBKM

ANALISIS SEGMENTASI PELANGGAN MODEL RFM DENGAN MENGGUNAKAN METODE PENGKLASTERAN K-MEANS

STUDI INDEPENDEN DI YAYASAN BAKTI ACHMAD ZAKY

RFM MODEL CUSTOMER SEGMENTATION ANALYSIS USING K-MEANS CLUSTERING METHOD

INDEPENDENT STUDY AT YAYASAN BAKTI ACHMAD ZAKY

Diajukan untuk memenuhi ujian mata kuliah Praktik Kerja Lapangan

MUHAMMAD FATHI FARHAT

24010120130115

DEPARTEMEN MATEMATIKA

FAKULTAS SAINS DAN MATEMATIKA

UNIVERSITAS DIPONEGORO

SEMARANG

2022

HALAMAN PENGESAHAN

LAPORAN PRAKTIK KERJA LAPANGAN

ANALISIS SEGMENTASI PELANGGAN MODEL RFM DENGAN MENGGUNAKAN METODE PENGKLASTERAN K-MEANS STUDI INDEPENDEN DI YAYASAN BAKTI ACHMAD ZAKY

Telah dipersiapkan dan disusun oleh:

MUHAMMAD FATHI FARHAT

24010120130115

An. Ketua Departemen Matematika Sekretaris Prodi S-1 Matematika	Pembimbing
Dr. Dra. Titi Udjiani, M.Si. NIP. 196402231991022001	Dr. Susilo Hariyanto, S.Si., M.Si. NIP. 19741014200012001

KATA PENGANTAR

Puji syukur kehadirat Tuhan Yang Maha Esa atas segala rahmat dan karunia-Nya sehingga penulis dapat menyelesaikan Laporan Praktik Kerja Lapangan Konversi Studi Independen MBKM berjudul "A**nalisis Segmentasi Pelanggan Model RFM dengan Menggunakan Metode Pengklasteran K-Means".**

Dalam penyusunan laporan ini, penulis menyampaikan terima kasih yang sebesar-besarnya kepada:

1. Ibu Prof. Dr. Widowati, S.Si., M.Si. selaku Dekan Fakultas Sains dan Matematika Universitas Diponegoro.
2. Bapak Dr. Susilo Hariyanto, S.Si., M.Si. selaku Ketua Departemen Matematika Fakultas Sains dan Matematika Universitas Diponegoro sekaligus Dosen Pembimbing yang telah memberikan bimbingan dalam proses pembuatan laporan ini.
3. Bapak Robertus Heri S.U., S.Si., M.Si. selaku Koordinator Mata Kuliah Praktik Kerja Lapangan
4. Bapak Muhammad Fahmi selaku Mentor dalam kegiatan Studi Independen Startup Campus – Data Science Track di Yayasan Bakti Achmad Zaky, yang membimbing saya selama kegiatan berlangsung
5. Kedua Orang Tua saya yang senantiasa mendukung keputusan saya untuk terus belajar dan mengembangkan diri.
6. Semua pihak yang telah memberikan dukungan dan bantuannya yang tidak dapat penulis tulis satu persatu.

Penulis menyadari bahwa Laporan Praktik Kerja Lapangan Konversi Studi Independen MBKM ini masih kesalahan dan kekurangan. Oleh karena itu, penulis mengharapkan kritik dan saran dari pembaca. Semoga tulisan ini dapat bermanfaat bagi setiap pihak.

Semarang, 12 Desember 2022

Penulis

DAFTAR ISI

DAFTAR GAMBAR

BAB I
PENDAHULUAN

1.1. Latar Belakang

Dewasa ini, *data discovery* dianggap sebagai bagian dasar dari kekuatan suatu bisnis karena dapat memudahkan para pegawai dan petinggi perusahaan untuk dapat memahami sampai menangani permasalahan yang ada melalui data. *Data discovery* adalah suatu proses pengumpulan data yang relevan dan pengolahan secara visual hingga menjadi wawasan yang berguna bagi keputusan bisnis suatu perusahaan. Beberapa perusahaan *startup unicorn* di bidang *e-commerce* seperti Shopee dan Tokopedia sangat memperhatikan proses *data discovery* untuk mengolah data pelanggan yang begitu banyak untuk memberikan wawasan berupa rekomendasi terbaik yang relevan dengan profil pelanggan melalui pengolahan data historis yang rapi dan terstruktur.

Suatu bisnis yang sedang berkembang membutuhkan strategi segmentasi pasar untuk memahami kebutuhan pelanggan. Pengolahan segmentasi pasar dapat mengelompokan konsumen ke dalam segmen konsumen secara alami maupun diciptakan secara artifisial yang memiliki preferensi atau karakteristik produk yang serupa. Segmentasi pasar memiliki keuntungan diantaranya perusahaan memiliki fokus dalam penjualannya, meningkatkan keunggulan persaingan bisnis, dapat memperluas proses maupun jangkaun penjualan, kepercayaan pelanggan terhadap pebisnis baik jasa maupun produk, memiliki komunikasi terhadap promosi produk yang dinilai oleh pelanggan, dan segmentasi dapat meningkatkan keunggulan dan meningkatkan kekuatan merk [1].

Startup campus merupakan program *bootcamp* intensif yang diinisiasi oleh Yayasan Bakti Achmad Zaky untuk mengembangkan talenta digital yang banyak dibutuhkan di era modern ini. Startup Campus Data Science Track adalah Studi Independen Bersertifikat Data Science yang memberi pencapaian belajar bagi peserta agar mampu menyelesaikan permasalahan/ tantangan bisnis menggunakan pemodelan kuantitatif dan teknik analisis data, serta mampu mengkomunikasikan hasil analisis serta menyajikannya menggunakan teknik visualisasi data yang efektif dan mampu mendemonstrasikan keahlian analisis data statistik atau teknik machine learning untuk pengambilan keputusan bisnis. Dengan demikian, program ini memiliki relevansi yang kuat dengan

topik yang diangkat penulis dalam laporan ini, yaitu "Analisis Segmentasi Pelanggan Model RFM dengan Menggunakan Metode Pengklasteran K-Means".

1.2. Rumusan Masalah

1. Bagaimana sistematika analisis segmentasi pelanggan dengan model RFM?
2. Bagaimana sistematika metode pengklasteran K-Means untuk menunjang analisis segmentasi pelanggan?
3. Bagaimana hubungan model RFM dan metode pengklasteran K-Means untuk analisis segmentasi pelanggan?

1.3 Tujuan

1. Mengetahui sistematika analisis segmentasi pelanggan dengan model RFM.
2. Mengetahui sistematika metode pengklasteran K-Means untuk menunjang analisis segmentasi pelanggan.
3. Mengetahui hubungan model RFM dan metode pengklasteran K-Means untuk analisis segmentasi pelanggan.

1.4 Manfaat

1. Bagi Mahasiswa

   Menambah wawasan serta pengalaman dalam mengaplikasikan ilmu tentang data sains yang sudah dipelajari.
2. Bagi Instansi Pendidikan

   Menambah arsip kepustakaan dalam pengembangan serta pengaplikasian ilmu tentang data sains.
3. Bagi Yayasan Bakti Achmad Zaky

   Laporan ini diharapkan dapat menjadi bukti bahwa penulis telah mengikuti program Studi Independen dengan baik.

BAB II
LANDASAN TEORI

2.1 Segmentasi Pelanggan

Keadaan Pasar adalah suatu hal yang akan memberikan keuntungan untuk meningkatkan produktivitas penjualan suatu produk baik, secara konvensional maupun secara online. Untuk mengetahui keadaan pasar, dibutuhkan strategi segmentasi pelanggan dengan cara mengelompokan pelanggan ke dalam segmen pelanggan secara alami maupun diciptakan secara artifisial yang memiliki preferensi atau karakteristik produk yang serupa.

Segmentasi pelanggan adalah proses membagi pelanggan biasa maupun potensial menjadi kelompok-kelompok yang berbeda atau segmen. Tujuan utama segmentasi pelanggan adalah untuk menemukan kelompok pelanggan yang memiliki pola konsumsi yang serupa. Melalui segmentasi pelanggan, bisnis dapat lebih fokus pada segmen yang secara signifikan meningkatkan penjualan dan keuntungan melalui pelanggan potensial[2].

2.2 Model RFM

Model RFM (*Recency, Frequency dan Monetary*) merupakan model segmentasi yang membedakan pelanggan melalui tiga variabel yaitu *recency*, *frequency* dan *monetary*. Model RFM telah banyak diterapkan dalam beberapa bidang, terutama dalam dunia pemasaran. Dengan mengadopsi model RFM, perusahaan dapat secara efektif mengidentifikasi pelanggan yang berharga dan akan digunakan sebagai pengembangan strategi pemasaran yang efektif. Model RFM sering digunakan untuk segmentasi pasar[3]. Model RFM terdiri dari *Recency, Frequency*, dan *Monetary* yang memiliki pengertian sebagai berikut:

2.2.1. Recency

Recency merupakan variabel untuk mengukur nilai pelanggan berdasarkan rentang waktu (tanggal, bulan, tahun) transaksi terakhir pelanggan sampai saat ini. Semakin kecil rentang waktu maka nilai *recency* semakin besar.

2.2.2. Frequency

Frequency merupakan variabel untuk mengukur nilai pelanggan berdasarkan jumlah transaksi yang dilakukan pelanggan dalam satu periode. Semakin banyak jumlah transaksi yang dilakukan maka nilai f semakin besar.

2.2.3. Monetary

Monetary merupakan variabel untuk mengukur nilai pelanggan berdasarkan jumlah besaran uang yang dikeluarkan pelanggan dalam satu periode. Semakin banyak jumlah besaran uang yang dikeluarkan pelanggan maka nilai *M* semakin besar [4].

2.3. Metode Pengklasteran K-Means

2.3.1. Pengertian K-Means

Metode K-Means merupakan metode non hirarki yang berasal dari metode data klasterisasi. Pendekatan K-Means mengambil strategi *greedy* (serakah) yang menghasilkan partisi baru dengan menugaskan setiap pola ke pusat klaster terdekat dan menghitung pusat klaster baru. K-Means mengklasifikasikan data tertentu yang ditetapkan melalui sejumlah klaster (k cluster). Idenya adalah mendefinisikan nilai pusat k (k centroid), satu untuk setiap klaster.

Metode K-Means dimulai dengan pembentukan partisi klaster di awal kemudian secara interaktif partisi klaster ini diperbaiki hingga tidak terjadi perubahan yang signifikan pada partisi klaster. Metode K-Means mempartisi data ke dalam klaster sehingga data berkarakteristik sama dimasukan kedalam satu klaster yang sama dan data yang berkarakteristik berbeda dikelompokkan ke dalam klaster yang lain. Tujuan dari pengelompokan data ini adalah untuk meminimalkan fungsi objektif yang diatur dalam proses pengelompokan, yang pada umumnya berusaha meminimalkan variasi dalam suatu kelompok dan memaksimalkan variasi antar kelompok [5].

Ketika dihadapkan suatu permasalahan yang membutuhkan teknik *data mining*, maka metode pengklasteran K-Means merupakan metode yang cocok untuk memudahkan pengelompokkan sejumlah data besar secara teliti. Pada penelitian yang dilakukan oleh Kusuma et al (2019) untuk menentukan kelas unggulan, pengelompokan siswa-siswi berprestasi menggunakan metode K-Means mempunyai ketelitian yang cukup tinggi terhadap ukuran objek, sehingga algoritma relatif lebih terukur dan efisien untuk pengolahan objek dalam jumlah besar[6].

2.3.2. *Distance Space* Untuk Menghitung Jarak Antara Data dan *Centroid*

Beberapa *distance space* telah diimplementasikan dalam menghitung jarak (*distance*) antara data dan centroid termasuk di antaranya L_1 (*Manhattan/City Block*) *distance space*, L_2 (*Euclidean*) *distance space*, dan *Lp* (*Minkowski*) *distance space*. Jarak antara dua titik

x_1 dan x_2 pada *Manhattan/City Block distance space* dihitung dengan menggunakan rumus sebagai berikut[7]:

$$D_{L_1}(x_2, x_1) = \|x_2 - x_1\| = \sum_{j=1}^{p} |x_{2j} - x_{1j}|$$

Dimana

P: Dimensi Data

$|...|$: Nilai absolut

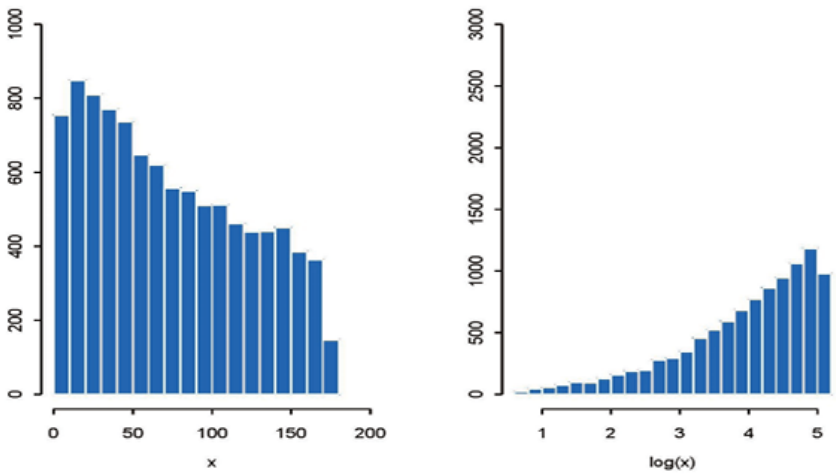
Sedangkan untuk L_2 (*Euclidean*) *distance space,* jarak antara dua titik dihitung menggunakan rumus sebagai berikut:

$$D_{L_1}(x_2, x_1) = \|x_2 - x_1\| = \sqrt{\sum_{j=1}^{p} (x_{2j} - x_{1j})^2}$$

2.3.3. Transformasi Log

Transformasi log merupakan metode transformasi yang paling banyak digunakan untuk mengatasi data menceng atau *skewed data. Skewness* atau kemencengan adalah tingkat ketidaksimetrisan atau kejauhan simetri dari sebuah distribusi normal data. Ketika distribusi suatu data tidak normal, pentransformasian data dapat dilakukan untuk membuat data menjadi senormal mungkin sehingga meningkatkan validitas analisis statistiknya. Kendati demikian, transformasi log tidak bisa sepenuhnya menjamin untuk menghilangkan kemencengan data menjadi berdistribusi normal.

Gambar 2.1. Grafik perubahan distribusi data menceng sebelum dan sesudah



ditransformasi log [8].

2.3.3. Silhoutte Score

Dalam melakukan pengklasteran, klastering dengan Silhoutte Score yang tinggi akan menunjukkan kualitas pengklasteran yang lebih maksimal[9]. Indeks Silhoutte pertama

kali ditemukan oleh Rousseeuw pada tahun 1987 untuk menentukan kesesuaian setiap titik untuk himpunan yang disebut sebagai klaster. Indeks Silhoutte didefinisikan sebagai:

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

$a(i)$: Jarak rata-rata titik ke-i ke seluruh titik pada klaster

$b(i)$: Jarak rata-rata minimum titik ke-i ke seluruh titik pada klaster lainnya

Ketika nilai $S(i)$ mendekati 1, maka titik ke-i cocok dengan klaster yang ditetapkan. Sebaliknya, ketika nilai $S(i)$ mendekati 0 atau negatif, maka titik ke-i tidak cocok untuk klaster yang ditetapkan. Silhoutte Score rata-rata yang menunjukkan kualitas setiap pengklasteran didefinisikan sebagai:

$$S(i) = \frac{1}{n} \sum_{S_i \in S} S(i)$$

BAB III
METODOLOGI PENELITIAN

3.1.Waktu dan Tempat

Kegiatan Praktik Kerja Lapangan ini dilaksanakan pada tanggal 18 Agustus sampai 24 Desember 2022 secara daring di PT. Achmad Zaky Foundation dalam bentuk Studi Independen dengan nama program yaitu Startup Campus Data Science Track. Pada program ini, penulis diajarkan berbagai cara untuk membuat suatu keputusan bisnis melalui data. Salah satunya adalah dengan melakukan analisis segmentasi metode RFM menggunakan metode pengklasteran K-Means.

3.2. Alat dan Prosedur Kerja

Alat yang digunakan dalam pengolahan data adalah Google Colab dengan menggunakan bahasa pemrograman Python. Adapun prosedur kerja yang dilakukan adalah sebagai berikut:

1. Identifikasi masalah.

   Masalah yang akan dibahas adalah menentukan segmentasi pelanggan untuk memudahkan dalam pengambilan keputusan bisnis perusahaan.

2. Pengumpulan Data.

   Data yang digunakan dalam kegiatan Praktik Kerja Lapangan ini adalah data sekunder yang diberikan oleh mitra yaitu dataset publik yang bisa diunduh melalui website Kaggle.com

3. Persiapan Data.

   Melakukan proses penguploadan data, penggabungan *(merge)* data, penanganan *missing value* (data kosong), hingga perubahan tipe data.

4. Analisis model RFM.

   Menentukan *Recency, Frequency*, dan *Monetary* dan finalisasi tabel RFM dari data.

5. Pengklasteran K-Means.

   Memastikan tidak ada fitur yang berdistribusi miring (*skewed distribution*), dan menghitung silhouette score untuk menentukan jumlah klaster ideal.

6. Eksplorasi klaster.

   Melakukan eksplorasi atau analisis mendalam terhadap masing-masing klaster yang mewakili segmen-segmen pelanggan.

BAB IV

HASIL DAN PEMBAHASAN

4.1. Persiapan Awal Data

4.1.1. Mengimpor *Library* Awal untuk Pengolahan Dataset

```
#mengimpor semua library yang dibutuhkan
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Gambar 4.1. Pengimporan Library yang Digunakan

4.1.2. Mengupload Dataset ke Colab

Seluruh dataset yang akan digunakan akan diupload menggunakan fungsi pd.read_csv(). Terdapat 9 dataset berbeda yang berisi data pelanggan suatu *e-commerce* berdasarkan kriteria tertentu. Adapun dataset yang digunakan dalam laporan ini adalah sebagai berikut:

1. Customers dataset
2. Geolocation dataset
3. Items dataset
4. Order payments dataset
5. Order reviews dataset
6. Products dataset
7. Orders dataset
8. Sellers dataset
9. Product category name translation dataset

4.1.3. Melakukan *Merge* Atau Penggabungan Data

Pada langkah sebelumnya, telah diketahui bahwa dataset-dataset disediakan secara terpisah sehingga perlu dilakukan penggabungan seluruh dataset ke dalam satu dataset baru dengan fungsi *merge*.

```
data = customers.merge(orders, on = 'customer_id') \
                .merge(items, on = 'order_id') \
                .merge(payments, on = 'order_id') \
                .merge(reviews, on = 'order_id') \
                .merge(products, on = 'product_id') \
                .merge(name_translations, on = 'product_category_name') \
                .merge(sellers, on = 'seller_id')
```

Gambar 4.2. Penggabungan Data

Selanjutnya, dengan fungsi pd.shape() diketahui bahwa dataset yang telah di-*merge* berukuran 115609x40. Artinya terdapat sebanyak 115.609 data pelanggan dan 40 fitur pada dataset baru.

4.1.4. Menyesuaikan Tipe Data

Pada Langkah ini, tipe data pada beberapa fitur perlu disesuaikan dengan keadaan sesungguhnya. Dalam hal ini, fitur yang menunjukkan waktu atau tanggal perlu diubah menjadi bertipe data Datetime agar dapat diidentifikasi sebagai waktu sehingga mendapat perlakuan khusus saat pengolahan data.

```
#mengubah tipe data pada timestamps menjadi date_time
for feature in ['order_purchase_timestamp', 'order_approved_at', 'order_delivered_carrier_date',
                'order_delivered_customer_date', 'order_estimated_delivery_date', 'shipping_limit_date',
                'review_creation_date', 'review_answer_timestamp']:
    data[feature] = pd.to_datetime(data[feature], errors = 'raise', utc = False)
```

Gambar 4.3. Penyesuaian Tipe Data Menjadi Datetime

4.1.5. Menghilangkan Fitur Berisi *Missing Value* atau Data

Pada Langkah ini, akan dihapus fitur-fitur yang berisi data kosong dalam jumlah besar. Sementara pada fitur yang berisi data kosong dalam jumlah kecil hanya akan dihilangkan baris yang berisi data kosong saja. Melalui pengecekan data kosong dengan menggunakan fungsi data.isnull().sum(), tampak bahwa tidak ada data kosong pada sebagian besar fitur. Data kosong kebanyakan disebabkan oleh pelanggan yang tidak memberikan ulasan untuk produk yang dibeli dan beberapa pengiriman yang gagal, yaitu pada fitur review_comment_title dan review_comment_message.

```
#menghilangkan fitur yang memiliki data kosong dengan jumlah besar
data.drop(['review_comment_title', 'review_comment_message'], axis=1, inplace = True)
#menghapus baris yang berisi data kosong dari fitur yang jumlah data kosongnya tidak besar
data.dropna(axis = 0, inplace = True)
```

Gambar 4.4 Menghapus Fitur Atau Baris Berisi Data Kosong

```
#mengecek kembali data yang kosong
missing_data = data.isnull().sum().sort_values(ascending = False)
missing_data
```

```
customer_id                    0
product_description_lenght     0
payment_value                  0
review_id                      0
review_score                   0
review_creation_date           0
review_answer_timestamp        0
product_category_name          0
product_name_lenght            0
product_photos_qty             0
customer_unique_id             0
product_weight_g               0
product_length_cm              0
```

Gambar 4.5 Pengecekan Kembali Missing Value

4.2. Analisis RFM

4.2.1. Menghitung *Recency*

Recency adalah indikator untuk mengukur seberapa baru (*recent*) pelanggan melakukan pembelian. Perhitungan *recency* pada data ini dihitung berdasarkan waktu pembelian terakhir pelanggan.

```
#Menggabungkan data berdasarkan ID pelanggan untuk mengetahui pembelian terakhir tiap pelanggan
df_recency = data.groupby(by = 'customer_unique_id', as_index = False)['order_purchase_timestamp'].max()
#Menghitung recency berdasarkan pembelian terakhir
df_recency["LastPurchaseDate"] = df_recency["LastPurchaseDate"].dt.date
recent_date = data['order_purchase_timestamp'].dt.date.max()
df_recency['Recency'] = df_recency['LastPurchaseDate'].apply(lambda x: (recent_date - x).days)
df_recency.head()
```

	customer_unique_id	LastPurchaseDate	Recency
0	0000366f3b9a7992bf8c76cfdf3221e2	2018-05-10	111
1	0000b849f77a49e4a4ce2b2a4ca5be3f	2018-05-07	114
2	0000f46a3911fa3c0805444483337064	2017-03-10	537
3	0000f6ccb0745a6a4b88665a16c9f078	2017-10-12	321
4	0004aac84e0df4da2b147fca70cf8255	2017-11-14	288

Gambar 4.6 Perhitungan *Recency*

4.2.2. Menghitung *Frequency*

Pada perhitungan ini, akan dihitung seberapa sering (*frequent*) tiap pelanggan membeli produk. Langkah yang dilakukan adalah dengan menggabungkan ID pelanggan untuk mengetahui jumlah ID unik pembelian tiap pelanggan.

```
#menghitung jumlah ID pembelian pelanggan yang sama
frequency_df = data.groupby(["customer_unique_id"]).agg({"order_id": "nunique"}).reset_index()
frequency_df.rename(columns = {"order_id": "Frequency"}, inplace = True)
frequency_df.head()
```

	customer_unique_id	Frequency
0	0000366f3b9a7992bf8c76cfdf3221e2	1
1	0000b849f77a49e4a4ce2b2a4ca5be3f	1
2	0000f46a3911fa3c0805444483337064	1
3	0000f6ccb0745a6a4b88665a16c9f078	1
4	0004aac84e0df4da2b147fca70cf8255	1

Gambar 4.7 Perhitungan *Frequency*

4.2.3. Menghitung *Monetary*

Pada perhitungan ini, akan diketahui total pengeluaran masing-masing pelanggan dalam membeli produk.

```
#mengitung total pengeluaran tiap pelanggan
monetary_df = data.groupby('customer_unique_id', as_index = False)['payment_value'].sum()
monetary_df.rename(columns = {"payment_value": "Monetary"}, inplace = True)
monetary_df.head()
```

	customer_unique_id	Monetary
0	0000366f3b9a7992bf8c76cfdf3221e2	141.90
1	0000b849f77a49e4a4ce2b2a4ca5be3f	27.19
2	0000f46a3911fa3c0805444483337064	86.22
3	0000f6ccb0745a6a4b88665a16c9f078	43.62
4	0004aac84e0df4da2b147fca70cf8255	196.89

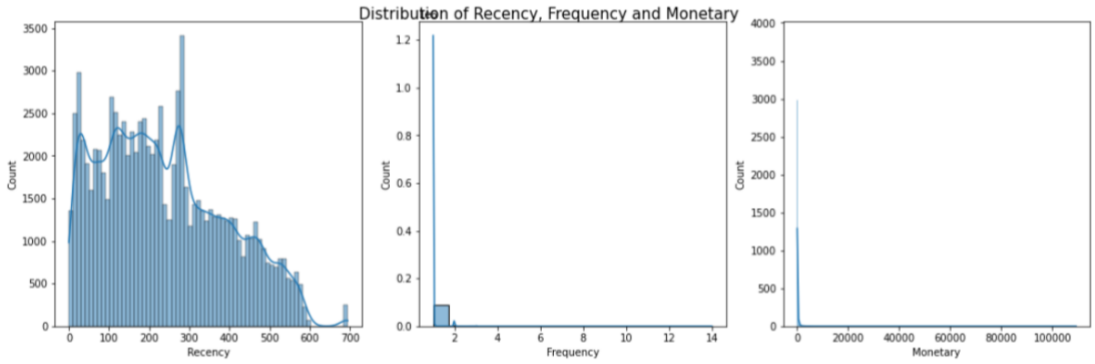
Gambar 4.8. Perhitungan *Monetary*

4.3. Pengklastern K-Means

Pada bagian ini, akan dilakukan pengklasteran dengan metode K-Means untuk menentukan banyaknya kluster yang ideal dari fitur RFM yang telah dibentuk. Namun, sebelum melakkan pengklasteran akan diuji beberapa indikator yaitu pengujian data yang menceng (*skewed data*) dan Silhoutte Score.

4.3.1. Transformasi Log Pada Data yang Menceng (*Skewed Data)*

Sebelum melakukan pengklasteran K-Means kita perlu memastikan apakah terdapat kemencengan (*skewness*) pada distribusi seluruh fitur RFM.

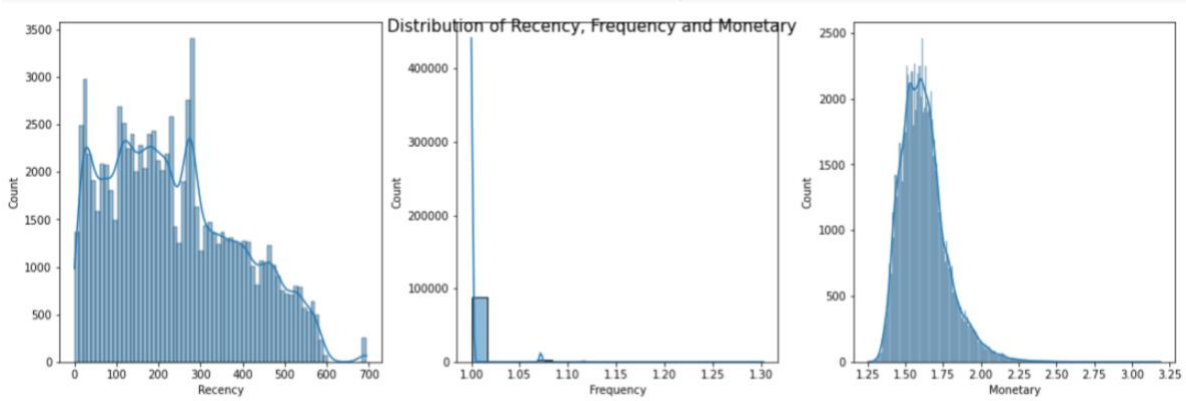


Gambar 4.9 Grafik Distribusi RFM

Dari grafik di atas terlihat bahwa fitur *frequency* dan *monetary* menceng ke kanan. Untuk mengatasinya maka dilakukan transformasi log pada data.

```
rfm_transformed = rfm.copy()
for feature in rfm.columns[2:]:
    rfm_transformed[feature] = rfm_transformed[feature].apply(lambda x: np.power(x, (1/10)))
```

Gambar 4.10. Fungsi Transformasi Log



Gambar 4.11 Grafik DIstribusi RFM Setelah Ditransformasi Log

Hasil transformasi log menunjukkan distribusi fitur *monetary* menjadi lebih berdistribusi normal. Sementara itu, pada distribusi fitur *frequency* tidak terdapat perubahan signifikan dan masih berdistribusi miring. Terlihat bahwa distribusi fitur *frequency* tidak bisa ditranformasi lebih jauh sehingga data akan digunakan seadanya.

4.3.2. Perhitungan Silhoutte Score

Langkah selanjutnya adalah menerapkan pengklasteran K-Means pada data terskala, lalu menentukan Silhoutte Score pada setiap jumlah kluster untuk menentukan berapakah jumlah kluster yang ideal untuk segmentasi pelanggan.

```
#penskalaan data
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
scaled_rfm = scaler.fit_transform(rfm.drop('customer_unique_id', axis = 1))

scaled_rfm_df = pd.DataFrame(scaled_rfm, columns = rfm.columns[1:])
```

Gambar 4.12. Penskalaan Data

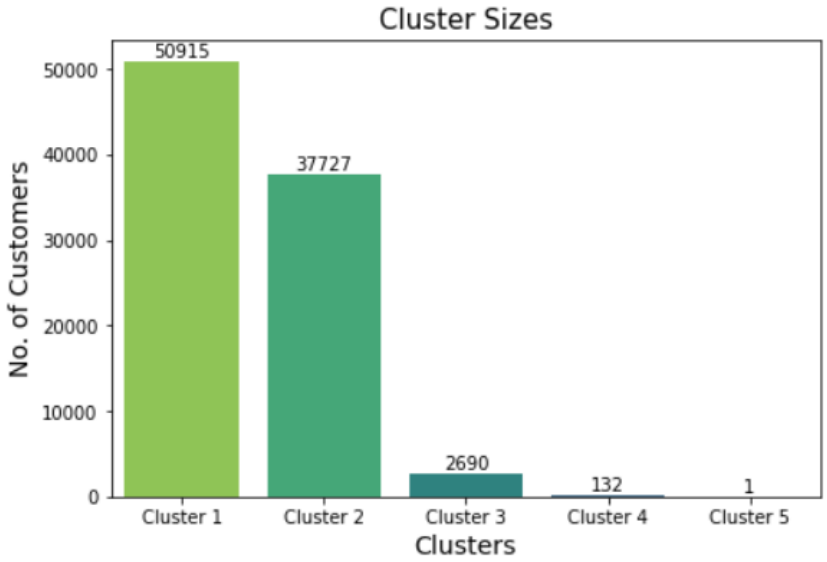
Gambar 4.13. Grafik Silhoutte Score

Dari grafik Silhoutte Score di atas, tampak bahwa Silhoutte Score mencapai titik tertinggi ketika jumlah kluster adalah 5. Dengan demikian, diperoleh bahwa jumlah klaster ideal dari metode pengklasteran K-Means di atas adalah sebanyak 5 klaster. Analisis segmentasi pelanggan akan dilakukan dengan membagi pelanggan menjadi 5 segmen yang berbeda.

4.4. Eksplorasi Klaster

Pada bagian ini, akan dilakukan eksplorasi klaster pada fitur RFM yang telah terbentuk sebelumnya, serta menganalisis segmentasi pelanggan model RFM dengan 5 klaster.

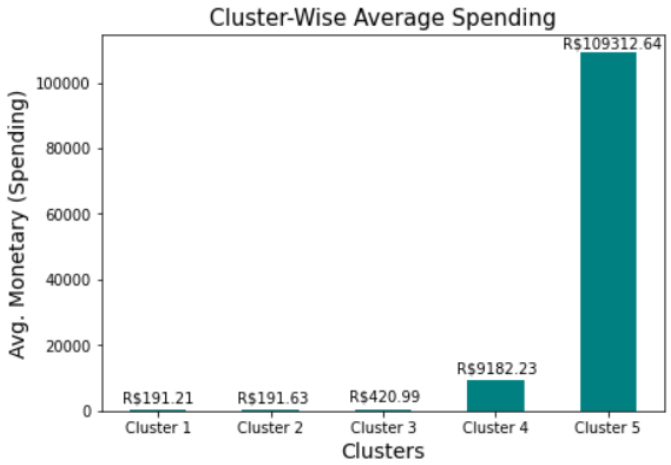
4.4.1. Distribusi Jumlah Pelanggan



Gambar 4.14. Grafik Jumlah Pelanggan Tiap Klaster

Dari grafik, terlihat bahwa pelanggan paling banyak tergolong pada klaster 1 yaitu sebanyak 50.915 pelanggan. Sementara itu, klaster 5 memiliki jumlah pelanggan paling sedikit yaitu 1 pelanggan.

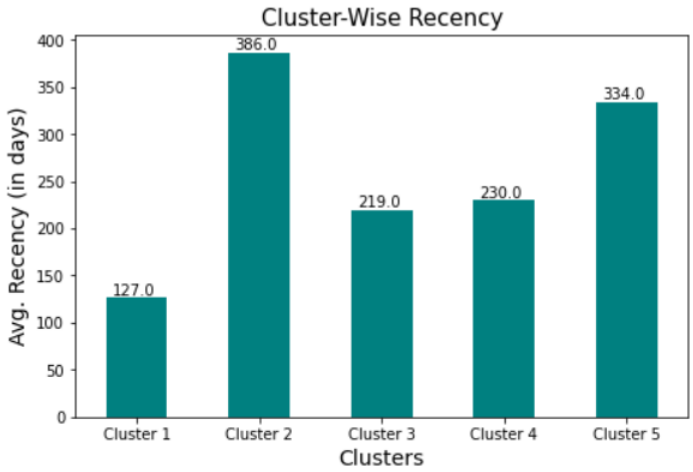
4.4.2. Rata-Rata Pengeluaran Pelanggan



Gambar 4.15. Grafik Rata-Rata Pengeluara Pelanggan

Dari grafik, terlihat jelas alasan mengapa klaster 5 hanya berisi 1 pelanggan. Jumlah pengeluaran pelanggan pada klaster 5 adalah yang paling besar diabnding klaster lainnya. Jumlah pengeluaran pelanggan pada klaster 1 paling sedikit, sementara pengeluaran pelanggan pada klaster 4 meningkat secara signifikan dibanding klaster sebelumnya.

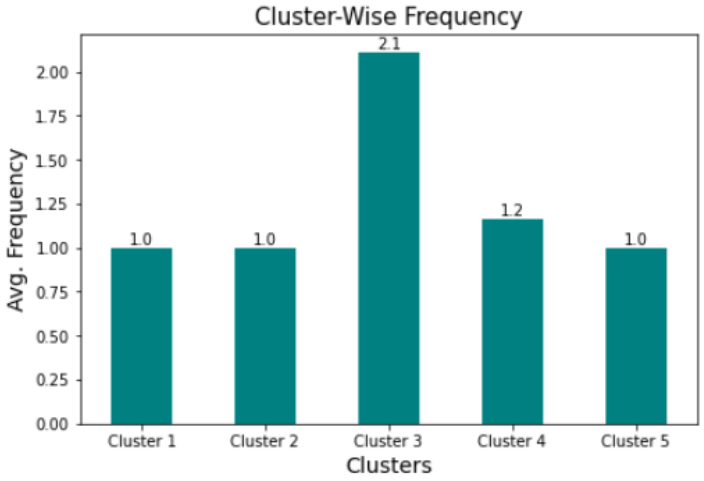
4.4.3. Rata-Rata Jumlah Hari Setelah Pembelian Terakhir



Gambar 4.16. Grafik Rata-Rata Jumlah Hari Setelah Pembelian Terakhir

Dari grafik, terlihat bahwa pelanggan di klaster 1 adalah pelanggan yang paling baru melakukan pembelian, sedangkan pelanggan dari klaster 2 sudah lebih dari satu tahun (rata-rata) sejak pembelian terakhir. Untuk pelanggan dari cluster 3 dan 4, masing-masing memiliki rata-rata sekitar 219 dan 230 hari sejak pembelian terakhir mereka.

4.4.4. Rata-Rata Jumlah Pembelian Produk



Gambar 4.17. Rata-Rata Jumlah Pembelian Produk

Pelanggan dari klaster 3 adalah pelanggan yang paling sering membeli dengan rata-rata 2,1 kali pembelian per pelanggan, sedangkan klaster lainnya memiliki rata-rata sekitar 1 kali pembelian per pelanggan.

4.4.5. Analisis Segmentasi Pelanggan

Berdasarkan eksplorasi klaster yang telah dilakukan pada Langkah-langkah sebelumnya, dapat diambil kesimpulan pola pelanggan masing-masing klaster atau segmen:

1. Klaster 0: Diartikan sebagai pelanggan baru/pembelanja rendah karena pengeluaran mereka adalah yang terkecil di antara semua klaster. Jumlah pelanggan pada klaster ini adalah yang terbanyak dibanding klaster lain sehingga menghadirkan peluang terbesar bagi perusahaan untuk meningkatkan basis pelanggan mereka. Pemasaran khusus terhadap segmen pelanggan ini dapat membantu meningkatkan relevansi perusahaan di antara mereka dan mengubah mereka menjadi pelanggan yang lebih loyal.
2. Klaster 1: Pelanggan dalam klaster ini merupakan konversi yang gagal dari pelanggan Klaster 0 (pelanggan baru/pembelanja rendah) menjadi Klaster 2 (Loyal/Pengeluaran Sedikit Lebih Tinggi). klaster ini mewakili pelanggan perusahaan yang dikategorikan sebagai *churn* (kehilangan pelanggan).
3. Klaster 2: Klaster ini berisi pelanggan yang loyal/pengeluaran sedikit lebih linggi. Pelanggan dalam klaster ini memiliki frekuensi pembelian tertinggi dan pengeluaran yang sedikit lebih tinggi dibandingkan dengan pelanggan baru. Pemasaran terhadap pelanggan ini harus fokus pada program loyalitas.

4. Klaster 3: Klaster ini berisi pelanggan dengan pengeluaran besar karena mereka banyak berbelanja (tidak termasuk Klaster 4 (anomali)) dan merupakan pembeli terbanyak kedua setelah Klaster 2 (loyal/pengeluaran sedikit lebih tinggi). Pemasaran ke pelanggan ini harus fokus pada rekomendasi produk karena mereka tampaknya tertarik untuk menghabiskan uang dalam jumlah besar sehingga menghasilkan keuntungan yang lebih tinggi.
5. Klaster 4: Hanya ada satu pelanggan di Klaster ini dan pelanggan ini telah menghabiskan uang dengan jumlah yang sangat besar untuk pembeliannya. Klaster ini dapat diartikan sebagai anomali.

BAB V
PENUTUP

5.1 Kesimpulan

Berdasarkan analisis yang telah dilakukan, didapatkan kesimpulan sebagai berikut:

1. Prosedur kerja dimulai dari identifikasi masalah, pengumpulan data, persiapan data, analisis model RFM, pengklasteran K-Means, dan eksplorasi klaster
2. Hasil dari analisis segmentasi pelanggan pada data adalah terbentuknya 5 klaster yang mewakili segmen pelanggan dengan pola konsumsi yang berbeda-beda.

5.2 Saran

Berdasarkan analisis data di atas, penulis memiliki kendala pada perhitungan Silhoutte Score untuk menentukan jumlah klaster ideal, yaitu akibat sangat lambatnya pemprosesan rumus Silhoutte Score di Colab. Dengan demikian, penulis menyarankan mencari perhitungan lain untuk menentukan jumlah klaster ideal metode K-Means.

DAFTAR PUSTAKA

[1] A. Surahman, A. F. Octaviansyah, and D. Darwis, "Ekstraksi Data Produk E-Marketplace Sebagai Strategi Pengolahan Segmentasi Pasar Menggunakan Web Crawler," *Sistemasi*, vol. 9, no. 1, p. 73, 2020, doi: 10.32520/stmsi.v9i1.580.

[2] C. Wang, "Efficient customer segmentation in digital marketing using deep learning with swarm intelligence approach," *Inf Process Manag*, vol. 59, no. 6, p. 103085, 2022, doi: 10.1016/j.ipm.2022.103085.

[3] B. E. Adiana, I. Soesanti, and A. E. Permanasari, "ANALISIS SEGMENTASI PELANGGAN MENGGUNAKAN KOMBINASI RFM MODEL DAN TEKNIK CLUSTERING," *Jurnal Terapan Teknologi Informasi*, vol. 2, no. 1, pp. 23–32, Apr. 2018, doi: 10.21460/JUTEI.2018.21.76.

[4] W. Taqwim, … N. S.-J., and undefined 2019, "Analisis Segmentasi Pelanggan Dengan Rfm Model Pada Pt. Arthamas Citra Mandiri Menggunakan Metode Fuzzy C-Means Clustering," *download.garuda.kemdikbud.go.id*.

[5] A. Sulistiyawati, E. S.-J. T. Kompak, and undefined 2021, "Implementasi Algoritma K-means Clustring dalam Penetuan Siswa Kelas Unggulan," *ejurnal.teknokrat.ac.id*, vol. 15, no. 2.

[6] A. S. Kusuma and K. S. Aryati, "Sistem Informasi Akademik Serta Penentuan Kelas Unggulan Dengan Metode Clusttering Dengan Algoritama K-Means Di Smp Negeri 3 Ubud," *Jurnal Sistem Informasi dan Komputer Terapan Indonesia (JSIKTI)*, vol. 1, no. 3, 2019, doi: 10.33173/jsikti.29.

[7] Yudi Agusta, "K-Means – Penerapan, Permasalahan dan Metode Terkait," *Jurnal Sistem dan Informatika*, vol. 3, no. Februari, 2007.

[8] C. Feng *et al.*, "Log-transformation and its implications for data analysis," *Shanghai Arch Psychiatry*, vol. 26, no. 2, 2014, doi: 10.3969/j.issn.1002-0829.2014.02.

[9] M. Ay, L. Özbakır, S. Kulluk, B. Gülmez, G. Öztürk, and S. Özer, "FC-Kmeans: Fixed-centered K-means algorithm," *Expert Syst Appl*, vol. 211, p. 118656, Jan. 2023, doi: 10.1016/j.eswa.2022.118656.

LAMPIRAN

Lampiran 1. Poster Laporan PKL

Kampus Merdeka INDONESIA JAYA | Startup Campus

ANALISIS SEGMENTASI PELANGGAN MODEL RFM DENGAN MENGGUNAKAN PENGKLASTERAN K-MEANS

Muhammad Fathi Farhat - 24010120130115

Latar Belakang

Segmentasi pelanggan adalah hal yang menguntungkan dalam segi bisnis sehingga diperlukan strategi segmentasi pelanggan yang akurat dan berbasis data.

Tujuan

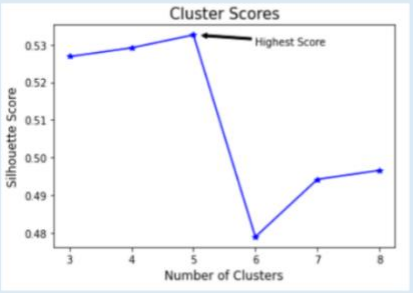
Menguji keakuratan hasil analisis segmentasi pelanggan model RFM dengan menggunakan pengklasteran K-Means

Prosedur Kerja

- Identifikasi masalah
- Pengumpulan data
- Persiapan data
- Analisis model RFM
- Pengklasteran K-Means
- Eksplorasi klaster

Hasil dan Pembahasan

PKL ini menggunakan data sekunder dari mitra yaitu dataset publik yang diakses melalui Kaggle.com



Gambar 1. Analisis Jumlah Klaster Ideal dengan Silhoutte Score

Hasil proses pengklasteran diperoleh 5 klaster ideal:

- Klaster 0: Pelanggan Baru
- Klaster 1: Pelanggan *Churn*
- Klaster 2: Pelanggan Loyal
- Klaster 3: Pelanggan dengan Pengeluaran Besar
- Klaster 4: Anomali

Kesimpulan

Pengklasteran K-Means dapat mengelompokkan pola konsumsi pelanggan ke dalam segmen pelanggan dengan sesuai.

Saran

Penulis menyarankan mencari perhitungan lain dalam menentukan jumlah klaster ideal karena lambatnya pemrosesan rumus Silhoutte Score di Python.