Rakamin Academy

Homework Unsupervised Learning

Anggota Kelompok:

Muhammad Hanif Fajari
Ja'far Shadiq Alatas
Fauzan Kharim
Aditya Ninda Putri
Debi Anggita Sasti

Exploratory Data Analysis

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 62988 entries, 0 to 62987
Data columns (total 23 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   MEMBER_NO          62988 non-null  int64
 1   FFP_DATE           62988 non-null  object
 2   FIRST_FLIGHT_DATE  62988 non-null  object
 3   GENDER             62985 non-null  object
 4   FFP_TIER           62988 non-null  int64
 5   WORK_CITY          60719 non-null  object
 6   WORK_PROVINCE      59740 non-null  object
 7   WORK_COUNTRY       62962 non-null  object
 8   AGE                62568 non-null  float64
 9   LOAD_TIME          62988 non-null  object
 10  FLIGHT_COUNT       62988 non-null  int64
 11  BP_SUM             62988 non-null  int64
 12  SUM_YR_1           62437 non-null  float64
 13  SUM_YR_2           62850 non-null  float64
 14  SEG_KM_SUM         62988 non-null  int64
 15  LAST_FLIGHT_DATE   62988 non-null  object
 16  LAST_TO_END        62988 non-null  int64
 17  AVG_INTERVAL       62988 non-null  float64
 18  MAX_INTERVAL       62988 non-null  int64
 19  EXCHANGE_COUNT     62988 non-null  int64
 20  avg_discount       62988 non-null  float64
 21  Points_Sum         62988 non-null  int64
 22  Point_NotFlight    62988 non-null  int64
dtypes: float64(5), int64(10), object(8)
```

Setelah dilakukan pengecekan, dataset terdiri dari 62988 baris dan 23 kolom.

Kemudian kami melakukan pengecekan ternyata tidak ada duplicated data.

Setelah itu kami melakukan pengecekan dan ada beberapa kolom yang terdapat missing value yaitu GENDER, WORK_CITY, WORK_PROVINCE, WORK_COUNTRY, AGE, SUM_YR_1,SUM_YR_2.

Data yang mengandung tanggal dapat diubah menjadi datetime.

Exploratory Data Analysis

Rakamin Academy

Berikut merupakan ringkasan statistik dari kolom numerik

```
# ringkasan statistik dari kolom numerik
df[nums].describe().T
```

	count	mean	std	min	25%	50%	75%	max
MEMBER_NO	62988.0	31494.500000	18183.213715	1.0	15747.750000	31494.500000	47241.250000	62988.0
FFP_TIER	62988.0	4.102162	0.373856	4.0	4.000000	4.000000	4.000000	6.0
AGE	62568.0	42.476346	9.885915	6.0	35.000000	41.000000	48.000000	110.0
FLIGHT_COUNT	62988.0	11.839414	14.049471	2.0	3.000000	7.000000	15.000000	213.0
BP_SUM	62988.0	10925.081254	16339.486151	0.0	2518.000000	5700.000000	12831.000000	505308.0
SUM_YR_1	62437.0	5355.376064	8109.450147	0.0	1003.000000	2800.000000	6574.000000	239560.0
SUM_YR_2	62850.0	5604.026014	8703.364247	0.0	780.000000	2773.000000	6845.750000	234188.0
SEG_KM_SUM	62988.0	17123.878691	20960.844623	368.0	4747.000000	9994.000000	21271.250000	580717.0
LAST_TO_END	62988.0	176.120102	183.822223	1.0	29.000000	108.000000	268.000000	731.0
AVG_INTERVAL	62988.0	67.749788	77.517866	0.0	23.370370	44.666667	82.000000	728.0
MAX_INTERVAL	62988.0	166.033895	123.397180	0.0	79.000000	143.000000	228.000000	728.0
EXCHANGE_COUNT	62988.0	0.319775	1.136004	0.0	0.000000	0.000000	0.000000	46.0
avg_discount	62988.0	0.721558	0.185427	0.0	0.611997	0.711856	0.809476	1.5
Points_Sum	62988.0	12545.777100	20507.816700	0.0	2775.000000	6328.500000	14302.500000	985572.0
Point_NotFlight	62988.0	2.728155	7.364164	0.0	0.000000	0.000000	1.000000	140.0

Exploratory Data Analysis

Rakamin Academy

Berikut merupakan ringkasan statistik dari kolom kategorik

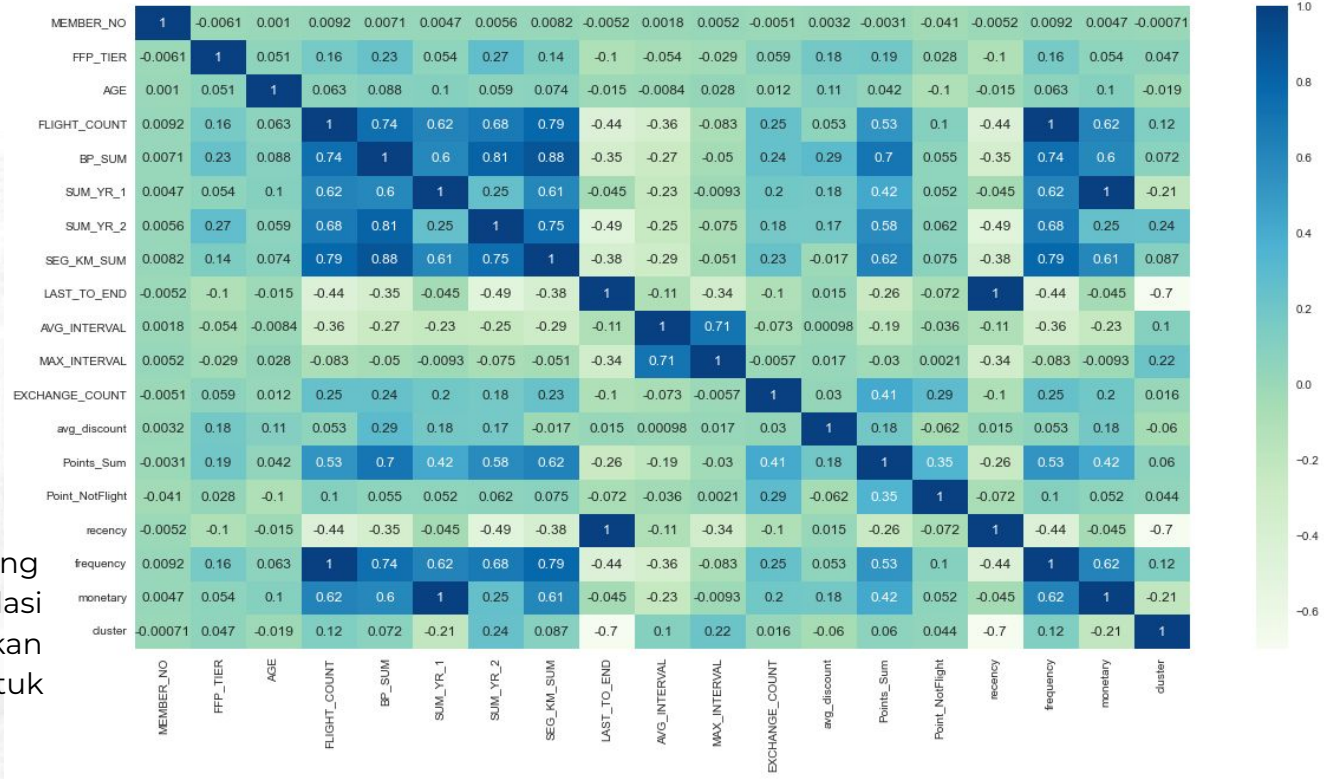
```
# categorical columns
df[cats].describe().T
```

	count	unique	top	freq
FFP_DATE	62988	3068	1/13/2011	184
FIRST_FLIGHT_DATE	62988	3406	2/16/2013	96
GENDER	62985	2	Male	48134
WORK_PROVINCE	59740	1165	guangdong	17509
WORK_COUNTRY	62962	118	CN	57748
WORK_CITY	60719	3234	guangzhou	9386
LOAD_TIME	62988	1	3/31/2014	62988
LAST_FLIGHT_DATE	62988	731	3/31/2014	959

Exploratory Data Analysis

Fitur yang berkorelasi adalah:

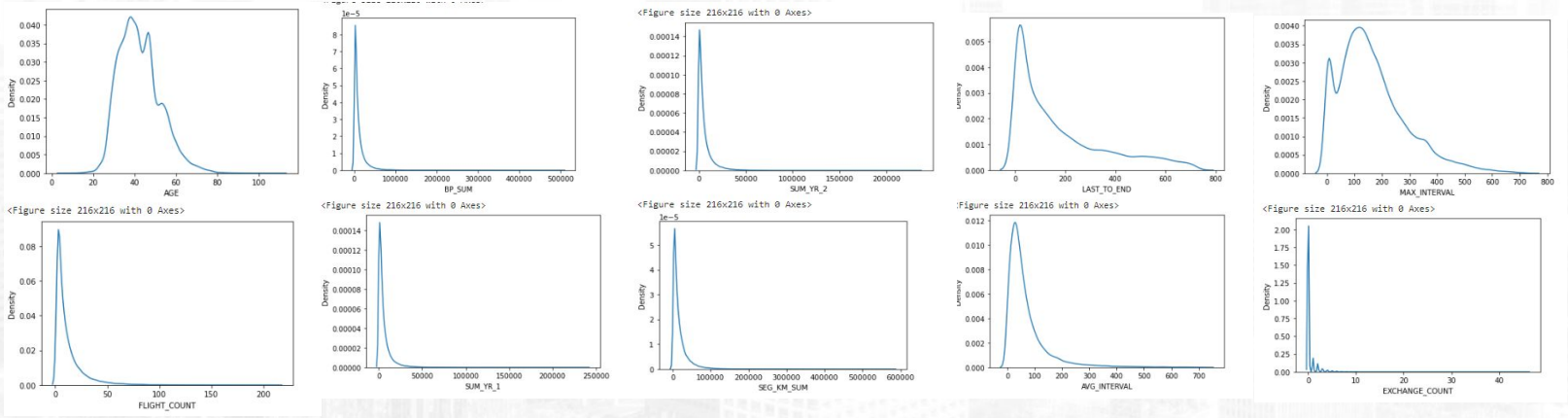
1. BP_SUM,
2. SUM_YR_2
3. SEG_KM_SUM
4. MAX_INTERVAL
5. Points_Sum
6. recency
7. frequency
8. monetary



Terdapat beberapa kolom yang hampir tidak memiliki korelasi dengan kolom manapun, akan dijadikan pertimbangan untuk digunakan

1. Exploratory Data Analysis

b. Keluarkan statistik kolom baik numerik maupun kategorikal, cari bentuk distribusi setiap kolom (numerik), dan jumlah baris untuk setiap unique value (kategorikal)



Setelah melakukan pengecekan distribusi dengan KDE plot dapat dilihat bahwa distribusi kolom rata-rata memiliki skew positif yang ekstrim

2. Feature Selection & Feature Engineering

Kami memilih fitur dengan menggunakan model RFM.

Recency :
Fitur yang digunakan Recency yaitu LAST_FLIGHT_DATE (Tanggal penerbangan terakhir) dikurangi dengan LOAD_TIME (Tanggal data diambil)

Frequency :
Fitur yang digunakan yaitu FLIGHT_COUNT (: Jumlah penerbangan Customer)

Monetary :
SUM_YR_1 (Fare Revenue)

Setelah memilih fitur yang akan digunakan kami melakukan Data Processing untuk fitur fitur tersebut.

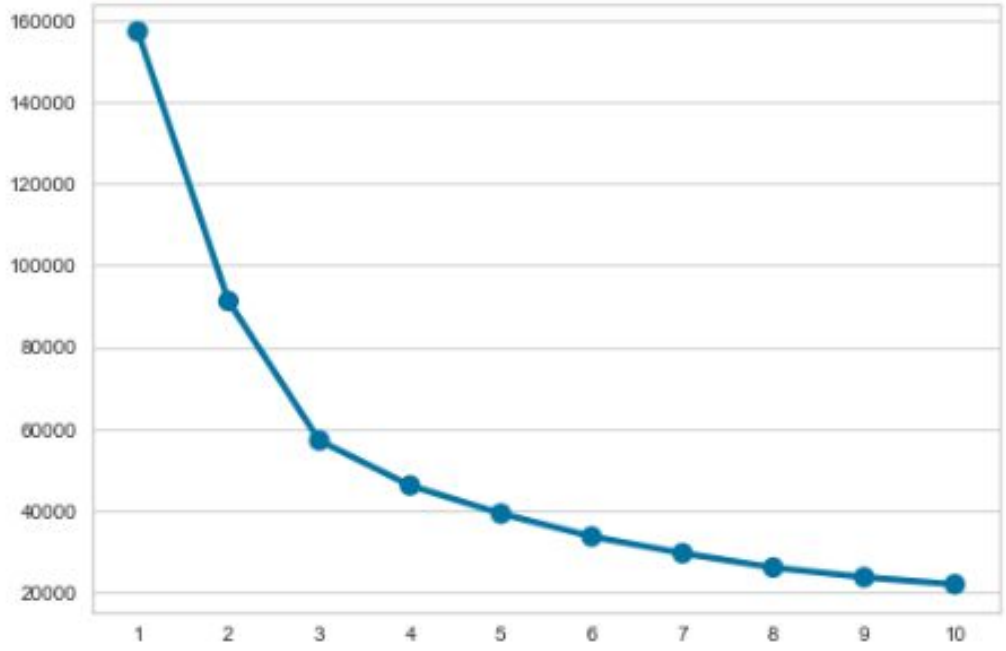
Terdapat missing value pada kolom Monetary (SUM_YR_1). Kemudian kami lakukan penghapusan (drop) pada data yang mengandung missing value.

Kemudian kami melakukan handling outlier dengan menggunakan IQR.

Setelah itu kami melakukan Scalling dengan StandardScaler.

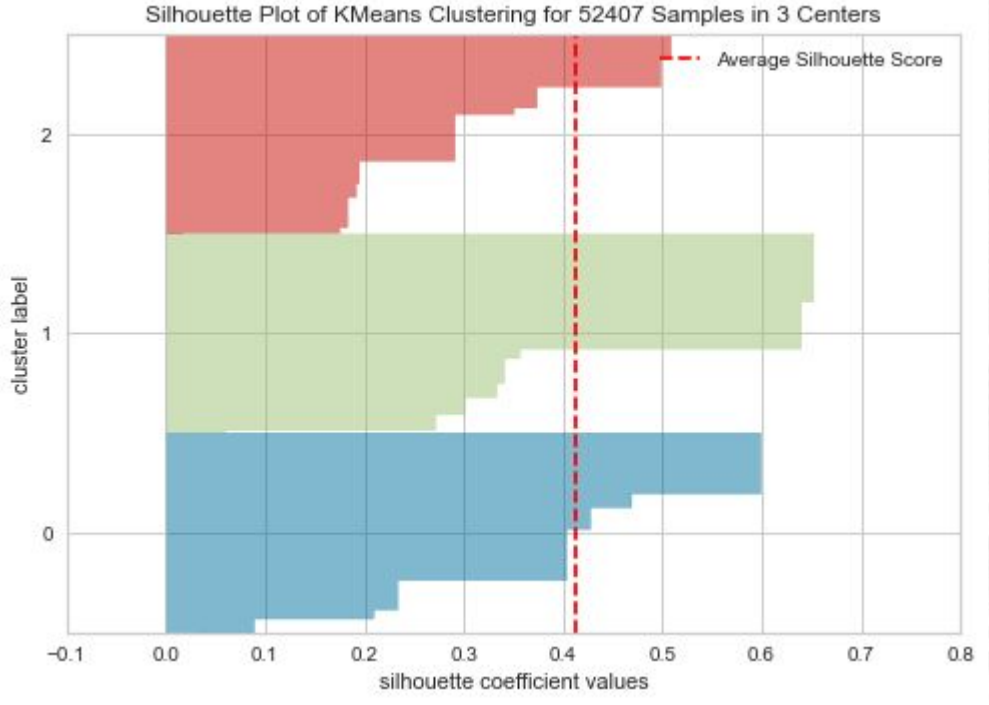
3. Modelling - KMeans

Berdasarkan elbow plot dibawah ini, jumlah cluster yang optimal adalah sebanyak 3 cluster. Dikarenakan mulai dari cluster 3 ke cluster 4 dan seterusnya, penurunan nilai inersia tidak terlalu besar. Sehingga apabila melakukan clustering sebanyak 4 atau lebih cluster tidak akan terlalu berarti.



3. Modelling

Nilai evaluasi silhouette score adalah diantara -1 hingga 1. Semakin mendekati 1, cluster yang dihasilkan terpisah dengan baik dengan cluster lainnya. Sedangkan nilai evaluasi silhouette score clustering K-means kami adalah sebesar 0.5.



3. Interpretasi Cluster

Dari hasil clustering dengan k-means diperoleh hasil statistik mean dan median dari setiap fitur di setiap cluster adalah sebagai berikut:

	cluster	recency			frequency			monetary			
		mean	min	max	mean	min	max	mean	min	max	count
0	0	347.472776	58	605	3.897426	2	27	2511.326363	0.0	12584.0	17558
1	1	80.852277	0	604	15.177704	5	31	6122.085636	1211.0	12590.0	17411
2	2	73.265168	0	192	5.796078	2	31	902.297282	0.0	12400.0	17438

cluster 0 (Risk to Churn Customer) :
Customer yang sudah lama tidak bertransaksi dengan jumlah penerbangan paling rendah. Customer ini mengeluarkan uang dengan jumlah rata-rata dari masing-masing cluster.

cluster 1 (Potential to be loyalist customer):
Customer yang biasa terbang dengan maskapai kita dengan jumlah penerbangan yang banyak dan mengeluarkan uang paling banyak

cluster 2 (Average Customer):
Customer yang paling sering menggunakan maskapai kita akan tetapi jarang melakukan penerbangan dan mengeluarkan uang paling sedikit.

4. Rekomendasi Strategi Bisnis

Dari hasil clustering sebelumnya, kami memberikan rekomendasi sebagai berikut:

cluster 0 (Risk to Churn Customer) :
Meminta tim marketing untuk riset mengapa customer ini sudah lama tidak menggunakan maskapai kita. Promo untuk terbang kembali dengan tujuan dan harga yang sesuai dengan karakteristik.

cluster 1 (Potential to be loyalist customer):
Memberikan promo referal kepada pelanggan untuk mengajak kerabat atau teman untuk menggunakan maskapai kita.

cluster 2 (Average Customer):
Memberikan promo untuk penerbangan selanjutnya dengan waktu yang terbatas pada saat pembelian tiket (Loyalty Program).