Airline Customer Value Analysis

Airline customer value analysis with K-Means clustering approach in python.

Anggota:

1. Aisyah Khairunnisa A
2. Budi Dwi Ananto
3. David Prayogo
4. Gayatri A
5. Hafidh Rizky
6. Muharlan
7. Shafa Amelia
8. Sony Monthona

Exploratory data analysis

```
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)
memory usage: 11.1+ MB
```

Intrepetasi dataset :

- Terdapat 23 kolom
- Memiliki range index 62.988 baris
- Terdapat missing values pada fitur GENDER, WORK_CITY, WORK_PROVINCE, WORK_COUNTRY, AGE, SUM_YR_1, SUM_YR_2
- Tipe data pada fitur FFP_DATE, FIRST_FLIGHT_DATE, LOAD_TIME, LAST_FLIGHT_DATE sebaiknya diubah menjadi datetime/timestamp
- Tipe data pada fitur AGE sebaiknya diubah menjadi integer

Exploratory data analysis

[5]:

	MEMBER_NO	FFP_TIER	AGE	FLIGHT_COUNT	BP_SUM	SUM_YR_1	SUM_YR_2	SEG_KM_SUM	LAST_TO_END	AVG_INTERVAL	MAX_INTERVAL	EXCHANGE_COUNT	avg_discount	Points_Sum	Point_NotFlight
count	62988.000000	62988.000000	62568.000000	62988.000000	62988.000000	62437.000000	62850.000000	62988.000000	62988.000000	62988.000000	62988.000000	62988.000000	62988.000000	62988.0000	62988.000000
mean	31494.500000	4.102162	42.476346	11.839414	10925.081254	5355.376064	5604.026014	17123.878691	176.120102	67.749788	166.033895	0.319775	0.721558	12545.7771	2.728155
std	18183.213715	0.373856	9.885915	14.049471	16339.486151	8109.450147	8703.364247	20960.844623	183.822223	77.517866	123.397180	1.136004	0.185427	20507.8167	7.364164
min	1.000000	4.000000	6.000000	2.000000	0.000000	0.000000	0.000000	368.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.0000	0.000000
25%	15747.750000	4.000000	35.000000	3.000000	2518.000000	1003.000000	780.000000	4747.000000	29.000000	23.370370	79.000000	0.000000	0.611997	2775.0000	0.000000
50%	31494.500000	4.000000	41.000000	7.000000	5700.000000	2800.000000	2773.000000	9994.000000	108.000000	44.666667	143.000000	0.000000	0.711856	6328.5000	0.000000
75%	47241.250000	4.000000	48.000000	15.000000	12831.000000	6574.000000	6845.750000	21271.250000	268.000000	82.000000	228.000000	0.000000	0.809476	14302.5000	1.000000
max	62988.000000	6.000000	110.000000	213.000000	505308.000000	239560.000000	234188.000000	580717.000000	731.000000	728.000000	728.000000	46.000000	1.500000	985572.0000	140.000000

Interpretasi data pada kolom nums :

- Kolom - kolom sekilas ada beberapa fitur yangg asimetrik distribusinya (mean dan median berbeda signifikan)
- Kolom FLIGHT_COUNT, BP_SUM, SUM_YR_1 ,SEG_KM_SUM, LAST_TO_END, AVG_INTERVAL, MAX_INTERVAL, Points_Sum, Point_NotFlight tampaknya skewed ke kanan
- Kolom FFP_TIER merupakan kolom kategori
- Kolom SUM_YR_1 dan SUM_YR_2 seharusnya integer karena merepresentasi jumlah/ sum
- Kolom avg_discount, MEMBER_NO, dan FFP_TIER memiliki distribusi normal
- Terdapat value 0 pada kolom SUM_YR_1 dan SUM_YR_2 yang tampak kurang tepat sehingga perlu dilakukan investigasi lebih lanjut
- Terdapat customer dengan usia 110 tahun yang tampak tidak normal, lebih baik dihapus

Exploratory data analysis

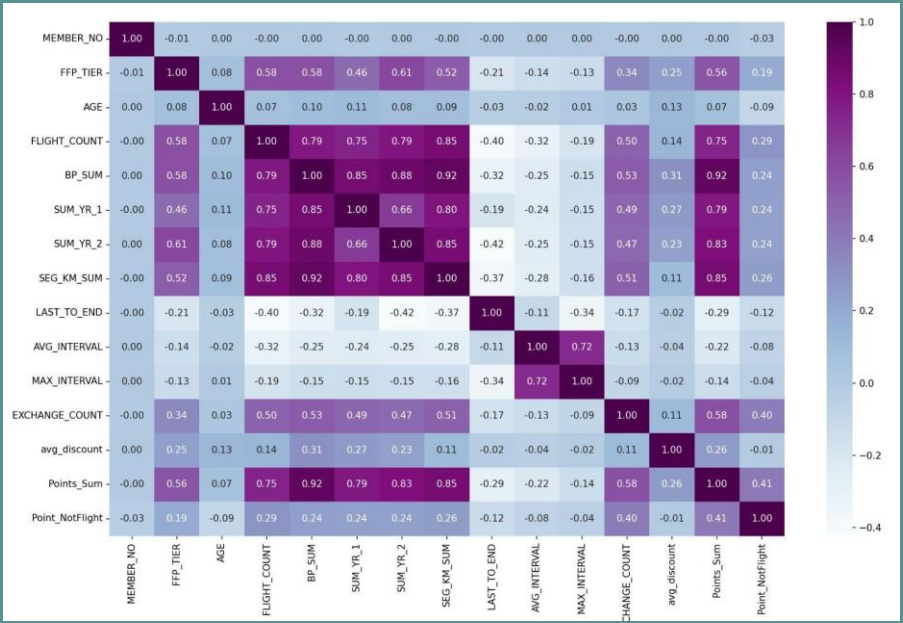
[6]:

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

Interpretasi data pada kolom cats:

- Pada kolom GENDER menunjukan bahwa user didominasi Male
- Pada kolom WORK_CITY didominasi penumpang dari guangzhou
- Pada kolom WORK_PROVINCE didominasi penumpang dari guangdong
- Pada kolom WORK_COUNTRY didominasi penumpang dari negara bagian CN
- Pada kolom LOAD_TIME terlihat bahwa data diambil pada tanggal 3/31/2014, sehingga dapat digunakan sebagai cutoff date
- Pada kolom LAST_FLIGHT_DATE didominasi datetime 3/31/2014
- Pada kolom GENDER terdapat 2 nilai yaitu Male, Female
- Pada kolom WORK_COUNTRY diduga terdapat singkatan nama negara CN, JP (japan), HK(Hongkong), SG (Singapore)

Exploratory data analysis



Berdasarkan plot di samping, berikut adalah fitur-fitur yang memiliki korelasi satu sama lain >0.5:

- FIRST_FLIGHT_DATE dengan FFP_DATE = 0.86
- WORK_PROVINCE dengan WORK_CITY = 0.81
- WORK_PROVINCE dengan WORK_COUNTRY = 0.70
- WORK_PROVINCE dengan WORK_CITY = 0.85
- BP_SUM dengan FLIGHT_COUNT = 0.79
- BP_SUM dengan FFP_TIER = 0.58
- BP_SUM dengan EXCHANGE_COUNT = 0.53
- FLIGHT_COUNT dengan FFP_TIER = 0.58
- SUM_YR_1 dengan FLIGHT_COUNT = 0.75
- SUM_YR_1 dengan BP_SUM = 0.85
- SUM_YR_1 dengan SUM_YR_2 = 0.66
- SUM_YR_1 dengan SEG_KM_SUM = 0.80
- SUM_YR_1 dengan Points_Sum = 0.79
- SUM_YR_2 dengan FFP_TIER = 0.61

Exploratory data analysis

- SUM_YR_2 dengan FLIGHT_COUNT = 0.80
- SUM_YR_2 dengan BP_SUM = 0.88
- SUM_YR_2 dengan SEG_KM_SUM = 0.85
- SUM_YR_2 dengan Points_Sum = 0.83
- SEG_KM_SUM dengan FFP_TIER = 0.52
- SEG_KM_SUM dengan FLIGHT_COUNT = 0.85
- SEG_KM_SUM dengan BP_SUM = 0.92
- SEG_KM_SUM dengan EXCHANGE_COUNT = 0.51
- Points_Sum dengan FFP_TIER = 0.56
- Points_Sum dengan FLIGHT_COUNT = 0.75
- Points_Sum dengan BP_SUM = 0.92
- Points_Sum dengan SEG_KM_SUM = 0.85
- Points_Sum dengan EXCHANGE_COUNT = 0.58
- MAX_INTERVAL dengan AVG_INTERVAL = 0.72

Data pre-processing & feature engineering

- Handling missing values dengan cara menghapus baris apabila nilainya < 5% dari keseluruhan data (5%=3.149,4), dan mengisi sisanya dengan modus.
- Handling outlier pada data yang mempunyai nilai anomali, seperti:
  a. Tanggal '2014/2/29' pada kolom LAST_FLIGHT_DATE.
  b. Usia 110 pada kolom AGE.
  c. Terdapat baris dimana harga tiketnya adalah 0, total diskon 0 tetapi memiliki jarak terbang, yang artinya customer melakukan penerbangan. Data dengan karakteristik tersebut tergolong tidak normal sehingga sebaiknya dihapus.
  d. Menghapus data duplikat (nol duplikat).
  e. Mengubah tipe data pada kolom AGE dari float menjadi integer.
  f. Mengubah tipe data pada kolom FFP_DATE, FIRST_FLIGHT_DATE, LOAD_TIME, LAST_FLIGHT_DATE dari integer menjadi datetime karena menunjukkan waktu.
  g. Melakukan feature engineering kolom MEMBERSHIP_DURATION yang diperoleh dari kolom LOAD_TIME dan FFP_DATE yang berguna untuk clustering.

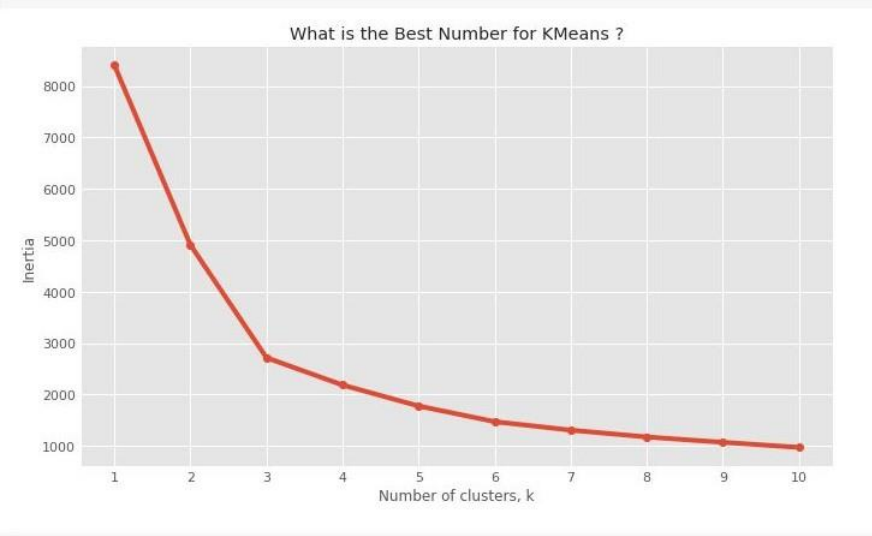
Feature selection

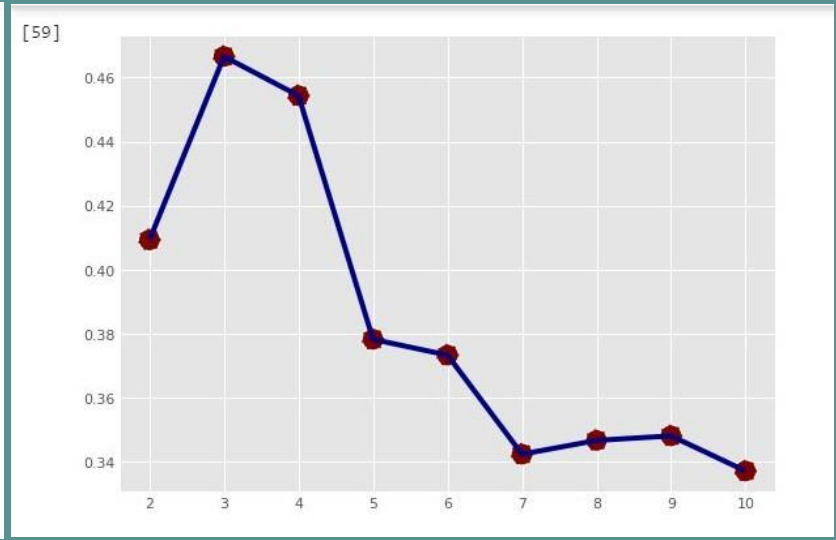
Tujuan dari clustering: ingin membuat clustering customer menjadi beberapa kategori seperti low value, middle value & high value dengan memperhitungkan prinsip RFM.

- Recency -> kolom LAST_TO_END: Jarak penerbangan terakhir ke pesanan penerbangan paling akhir.
- Frequency -> kolom FLIGHT_COUNT: Jumlah penerbangan customer.
- Monetary -> kolom SEG_KM_SUM: Total jarak (km) penerbangan yang sudah dilakukan.
- Fitur yang tidak kalah pentingnya adalah MEMBERSHIP_DURATION yang menunjukkan periode membership dalam bulan. Fitur ini merupakan hasil feature engineering dari FFP_DATE dan LOAD_TIME.

Karena distribusi pada fitur-fitur pilihan tersebut merupakan skew positif maka perlu dilakukan transformasi sehingga dilakukan normalisasi.

Clustering (K-means)



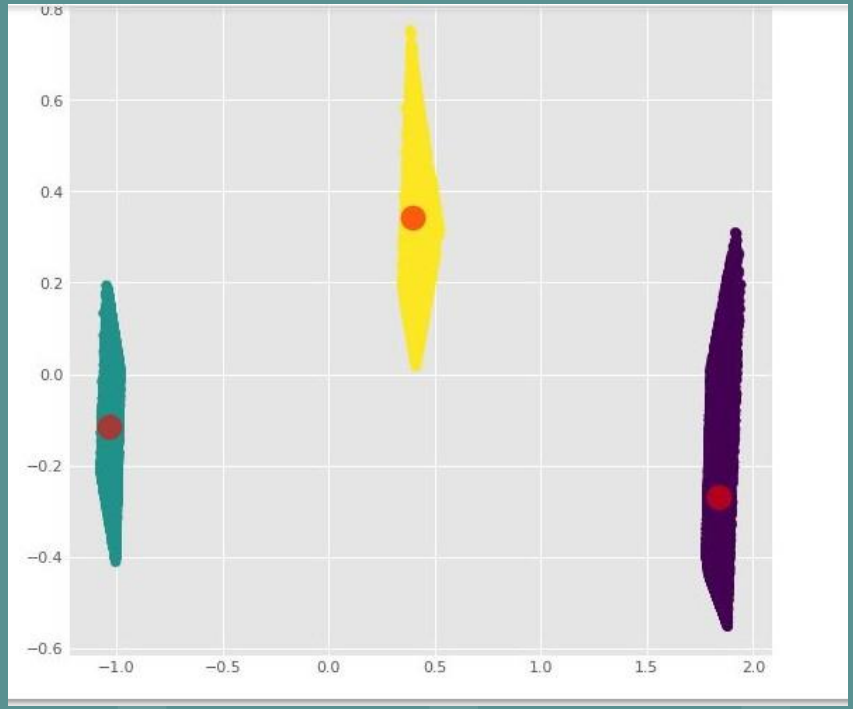


Dari hasil visualisasi dan silhouette score terlihat bahwa jumlah cluster yang optimal adalah 3.

Clustering (K-means)

	PC 1	PC 2
0	0.362591	0.580623
1	0.368782	0.497111
2	0.379502	0.241637
3	0.346911	0.278847
4	0.376613	0.333881

Berikut adalah hasil evaluasi menggunakan PCA. Terlihat bahwa setiap cluster sudah terpisahkan dengan baik.



Clustering (K-means)

	clusters	total_members
0	1	28405
1	2	18849
2	0	12033

clusters	MEMBERSHIP_DURATION	LAST_TO_END	FLIGHT_COUNT	SEG_KM_SUM
0	31.0	467.0	3.0	4575.0
1	28.0	71.0	8.0	11768.0
2	80.0	66.0	10.0	14608.0

Setelah melabeli setiap customer sesuai dengan clusternya, berikut adalah jumlah dari masing-masing cluster.

Business insight

Berdasarkan hasil dari clustering, diketahui terdapat 3 kategori customer berdasarkan periode membership, jumlah penerbangan, jarak penerbangan terakhir ke pesanan penerbangan terakhir, dan total jarak dari penerbangan yang sudah dilakukan, yaitu:

- Cluster 0: Low value customer.

  a. Jarak penerbangan terakhir ke pesanan penerbangan terakhir paling lama

  b. Durasi membershp paling baru

  c. Jumlah penerbangan sedikit

  d. Total jarak yang ditempuh sedikit

- Cluster 1: Middle value customer, cukup loyal.

  a. Jarak penerbangan terakhir ke pesanan penerbangan terakhir menegah/sedang

  b. Durasi membershp menengah

  c. Jumlah penerbangan sedang

  d. Total jarak yang ditempuh menegah/sedang

Business insight

- Cluster 2: High value customer, sangat loyal.

  a. Jarak penerbangan terakhir ke pesanan penerbangan terakhir paling dekat

  b. Durasi membershp paling lama

  c. Jumlah penerbangan paling banyak

  d. Total jarak yang ditempuh banyak

Berdasarkan cluster tersebut, pihak maskapai sebaiknya meningkatkan interaksi kepada customer cluster 0 dan 1 agar mereka lebih sering menggunakan maskapai tersebut, bisa dengan cara memberikan promo ataupun diskon lainnya. Sedangkan untuk customer pada cluster 2 sebaiknya diberikan fasilitas khusus seperti ealry boarding time, vip lounge, dan sebagainya agar customer tersebut tetap loyal kepada maskapai.