Rakamin Academy

Homework Unsupervised Learning

Dokumen Airlines.csv
Project Raka-Max

Dataset Information

Code	Description
MEMBER_NO-b	ID Member
FFP_DATE	Frequent Flyer Program Join Date
FIRST_FLIGHT_DATE	Tanggal Penerbangan pertama
GENDER	Jenis Kelamin
FFP_TIER	Tier dari Frequent Flyer Program
WORK_CITY	Kota Asal
WORK_PROVINCE	Provinsi Asal
WORK_COUNTRY	Negara Asal
AGE	Umur Customer
LOAD_TIME	Tanggal data diambil
FLIGHT_COUNT	Jumlah penerbangan Customer
BP_SUM	Rencana Perjalanan
SUM_YR_1	Fare Revenue
SUM_YR_2	Votes Prices
SEG_KM_SUM	Total jarak(km) penerbangan yg sudah dilakukan
LAST_FLIGHT_DATE	Tanggal penerbangan terakhir
LAST_TO_END	Jarak waktu penerbangan terakhir ke pesanan penerbangan paling akhir
AVG_INTERVAL	Rata-rata jarak waktu
MAX_INTERVAL	Maksimal jarak waktu
EXCHANGE_COUNT	Jumlah penukaran
avg_discount	Rata rata discount yang didapat customer
Points_Sum	Jumlah poin yang didapat customer
Point_NotFlight	point yang tidak digunakan oleh members

Dataset di samping berisikan informasi member sebuah program frequent flyer dari suatu maskapai penerbangan.

Setiap row mewakili informasi customer yang mengikuti program dari maskapai tersebut.

Goals

membuat segmentasi member program penerbangan dengan data yang sudah disediakan di atas.

Pre-Processing I

Data Cleaning and Handling

Rakamin Academy

```
MEMBER_NO              0
FFP_DATE               0
FIRST_FLIGHT_DATE      0
GENDER                 3
FFP_TIER               0
WORK_CITY           2269
WORK_PROVINCE       3248
WORK_COUNTRY          26
AGE                  420
LOAD_TIME              0
FLIGHT_COUNT           0
BP_SUM                 0
SUM_YR_1             551
SUM_YR_2             138
SEG_KM_SUM             0
LAST_FLIGHT_DATE       0
LAST_TO_END            0
AVG_INTERVAL           0
MAX_INTERVAL           0
EXCHANGE_COUNT         0
avg_discount           0
Points_Sum             0
Point_NotFlight        0
dtype: int64
```

Fitur yang memiliki missing values.

- Fitur WORK_CITY dan WORK_PROVINCE akan kita impute menggunakan modus dari fitur tersebut.
- Sedangkan fitur seperti GENDER, WORK_COUNTRY AGE, SUM_YR_1 dan SUM_YR_2 akan kita drop karena memiliki porsi dibawah 1% dari keseluruhan data.

Datatype setiap fitur.

```
#   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   MEMBER_NO          61856 non-null  int64
 1   FFP_DATE           61856 non-null  object
 2   FIRST_FLIGHT_DATE  61856 non-null  object
 3   GENDER             61856 non-null  object
 4   FFP_TIER           61856 non-null  int64
 5   WORK_CITY          59701 non-null  object
 6   WORK_PROVINCE      58747 non-null  object
 7   WORK_COUNTRY       61856 non-null  object
 8   AGE                61856 non-null  float64
 9   LOAD_TIME          61856 non-null  object
 10  FLIGHT_COUNT       61856 non-null  int64
 11  BP_SUM             61856 non-null  int64
 12  SUM_YR_1           61856 non-null  float64
 13  SUM_YR_2           61856 non-null  float64
 14  SEG_KM_SUM         61856 non-null  int64
 15  LAST_FLIGHT_DATE   61856 non-null  object
 16  LAST_TO_END        61856 non-null  int64
 17  AVG_INTERVAL       61856 non-null  float64
 18  MAX_INTERVAL       61856 non-null  int64
 19  EXCHANGE_COUNT     61856 non-null  int64
...
 21  Points_Sum         61856 non-null  int64
 22  Point_NotFlight    61856 non-null  int64
```

- Beberapa fitur yang menunjukan tanggal akan dirubah agar memiliki tipe data yang sesuai guna prosesi EDA yang lebih tepat.

Data Cleaning and Handling

Rakamin Academy

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
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
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
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
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
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
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
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
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

- Fitur Age memiliki max value 110 yang mana merupakan sebuah anomali.

Feature Transformation

Rakamin Academy

Setelah dilakukan transformasi fitur kita mendapati bahwa:

	FFP_DATE	FIRST_FLIGHT_DATE	GENDER	WORK_CITY	WORK_PROVINCE	WORK_COUNTRY	LOAD_TIME	LAST_FLIGHT_DATE
count	61436	61436	61436	61436	61436	61436	61436	61436
unique	3067	3404	2	3189	1149	117	1	730
top	2011-01-13 00:00:00	2013-02-16 00:00:00	Male	guangzhou	guangdong	CN	2014-03-31 00:00:00	2014-03-31 00:00:00
freq	182	92	46972	11406	20330	56332	61436	956
first	2004-11-01 00:00:00	1905-12-31 00:00:00	NaN	NaN	NaN	NaN	2014-03-31 00:00:00	2012-04-01 00:00:00
last	2013-03-31 00:00:00	2015-05-30 00:00:00	NaN	NaN	NaN	NaN	2014-03-31 00:00:00	2014-03-31 00:00:00

Ada data yang menunjukan penerbangan pertama pada tahun 1905. (Tidak Mungkin)

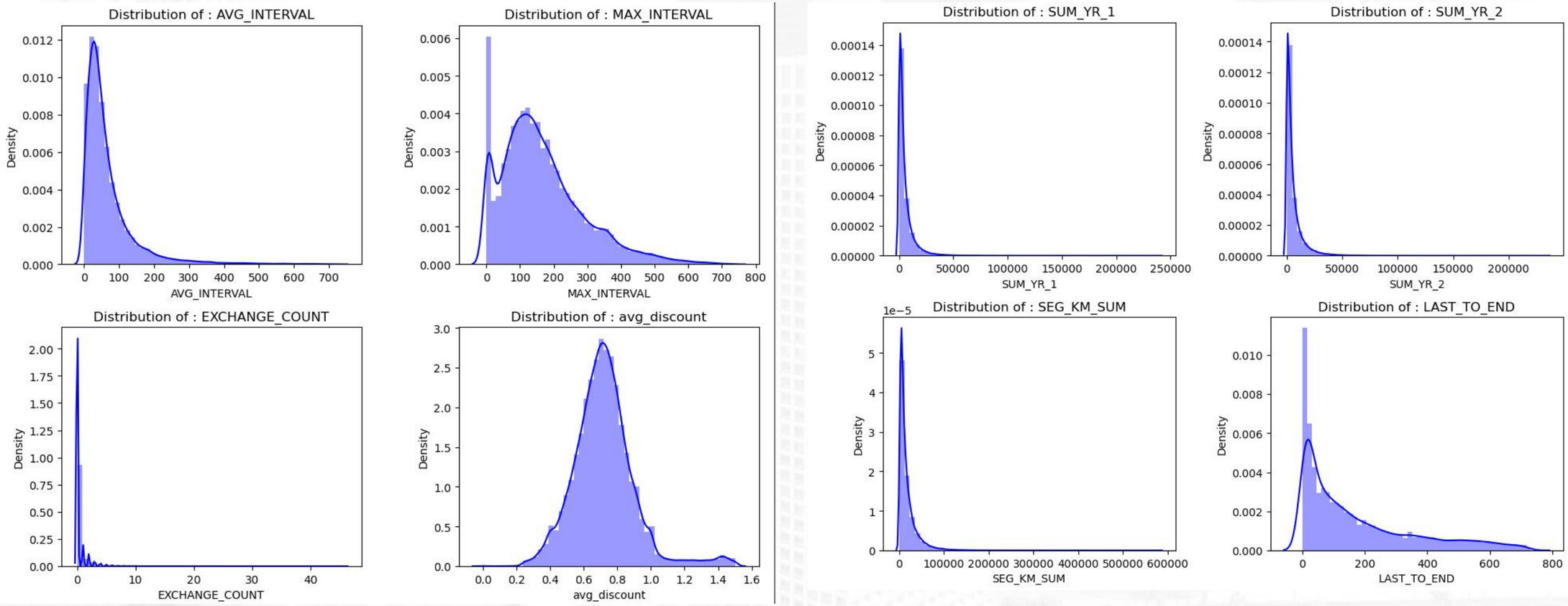
LAST_FLIGHT_DATE
2014/2/29 0:00:00
2014/2/29 0:00:00
2014/2/29 0:00:00
2014/2/29 0:00:00
2014/2/29 0:00:00
...
2014/2/29 0:00:00
2014/2/29 0:00:00
2014/2/29 0:00:00
2014/2/29 0:00:00
2014/2/29 0:00:00

Data menunjukan ada sejumlah row penerbangan terakhir pada tanggal 29-Februari-2014 yang mana tahun 2014 tidak memiliki tanggal 29 Februari.

Rakamin Academy

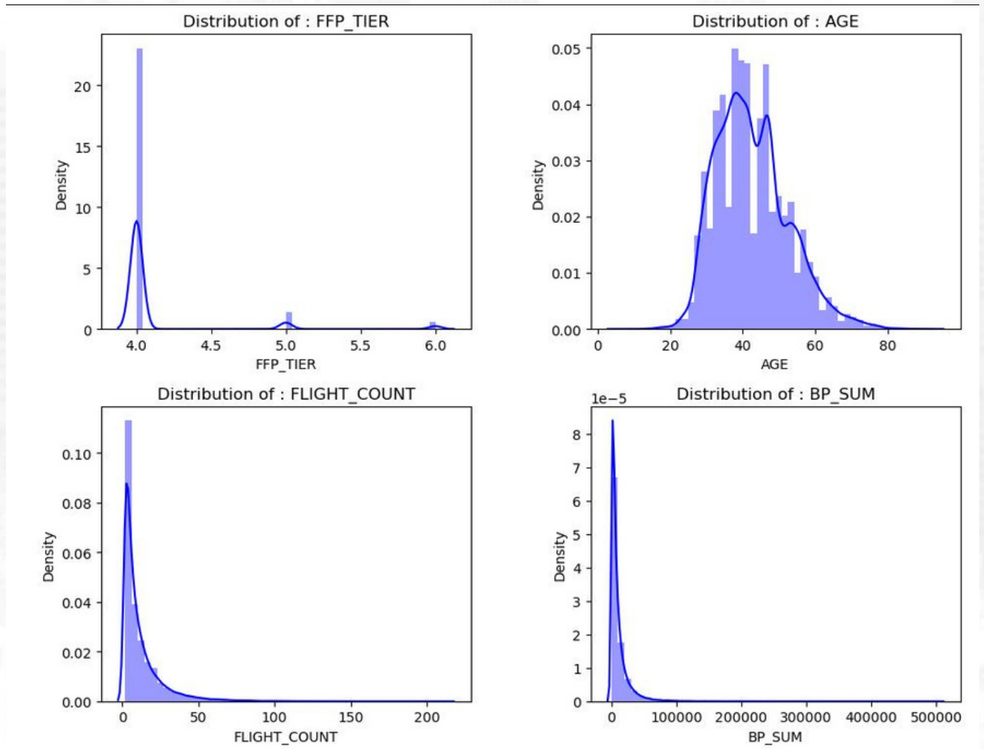
EDA

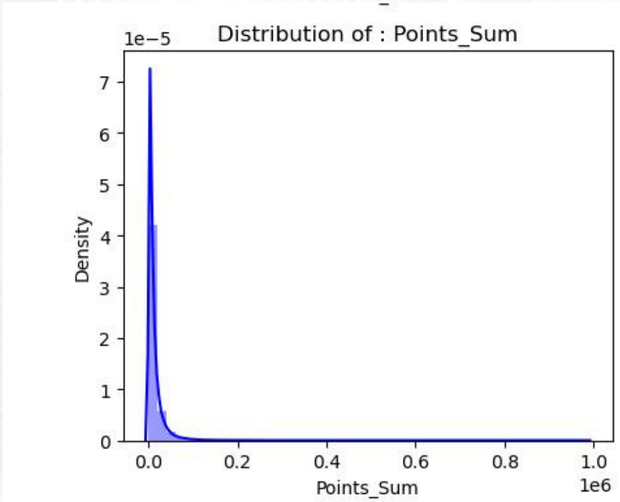
Univariate Data Analysis

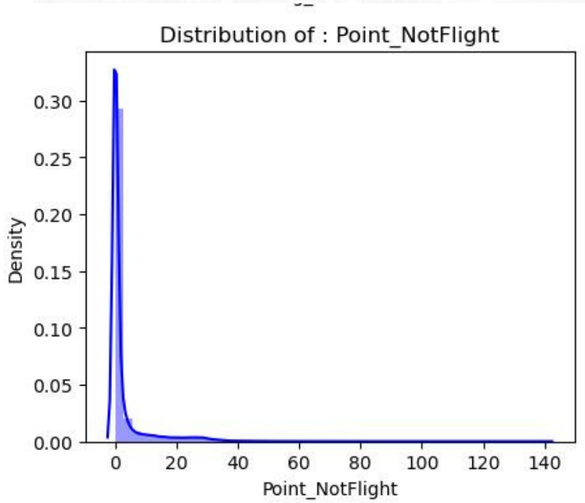


Mayoritas fitur memiliki distribusi yang skew kecuali avg_discount memiliki distribusi yang normal.

Univariate Data Analysis

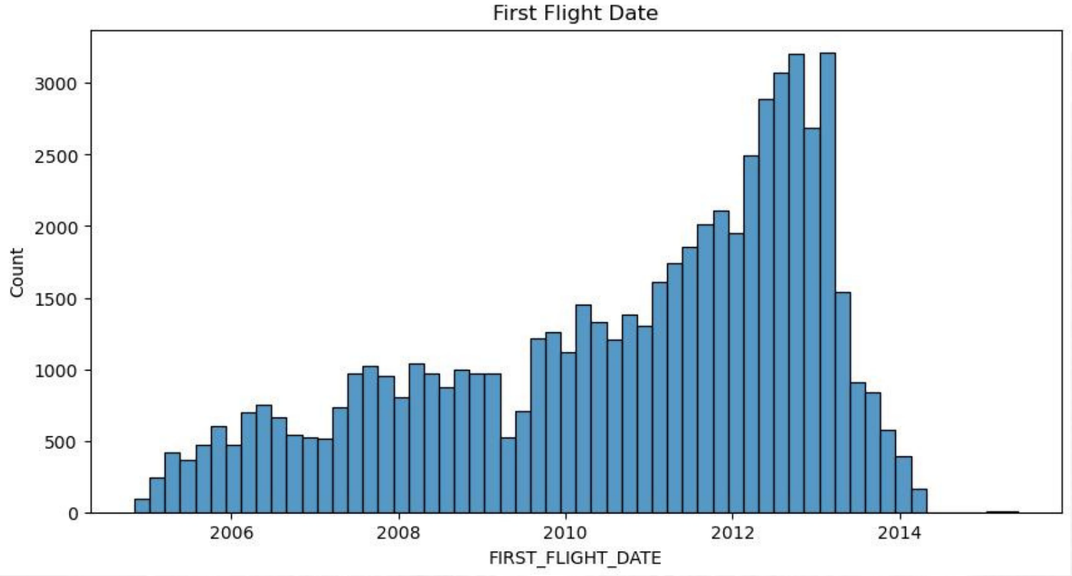


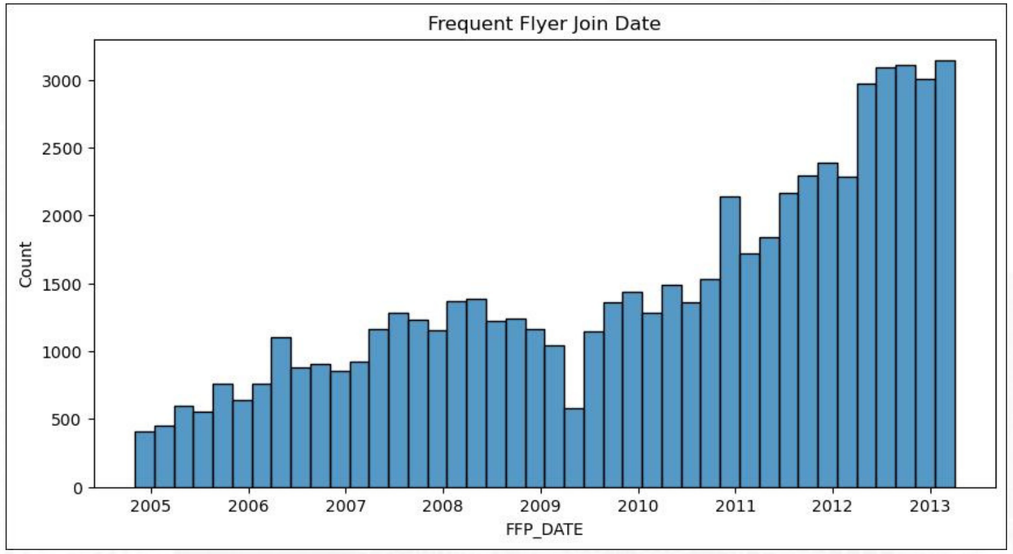


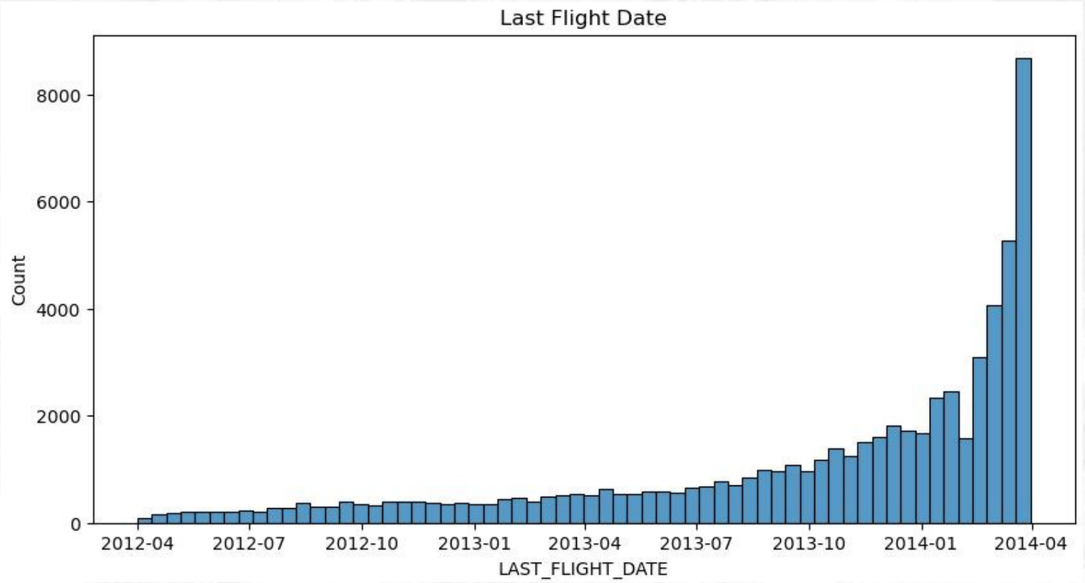


Mayoritas fitur memiliki distribusi yang skew kecuali AGE memiliki distribusi yang normal.

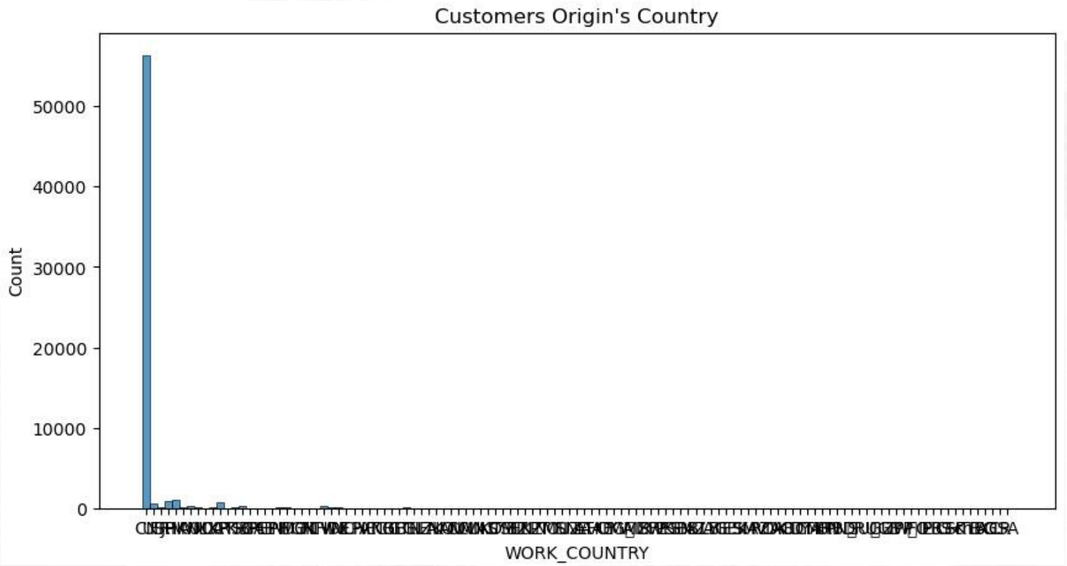
Univariate Data Analysis

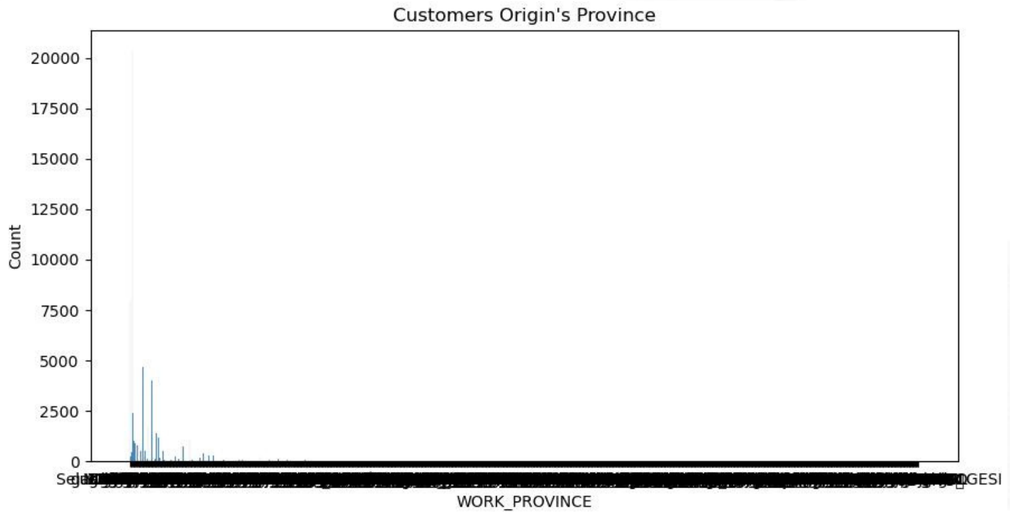


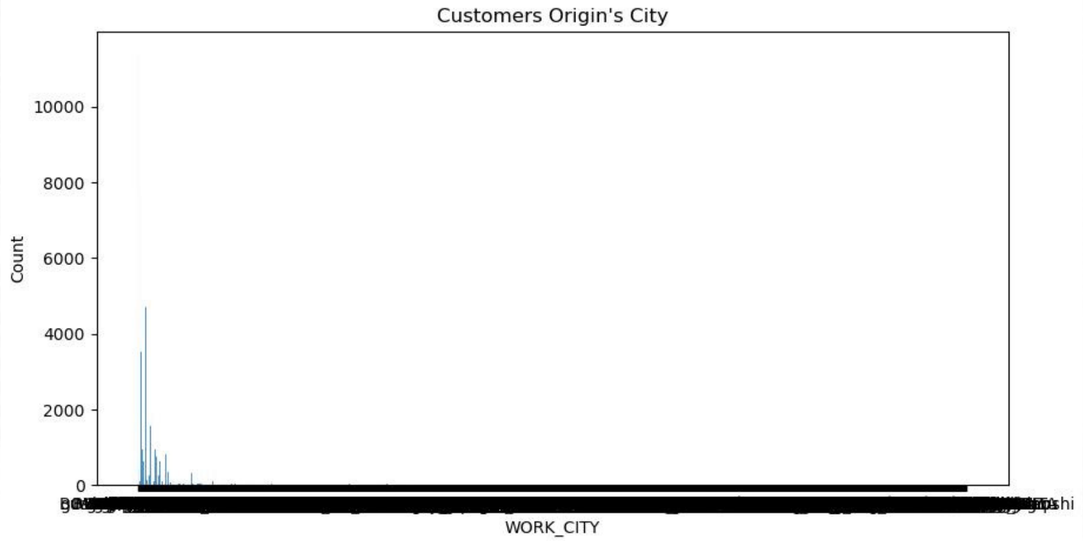




Univariate Data Analysis

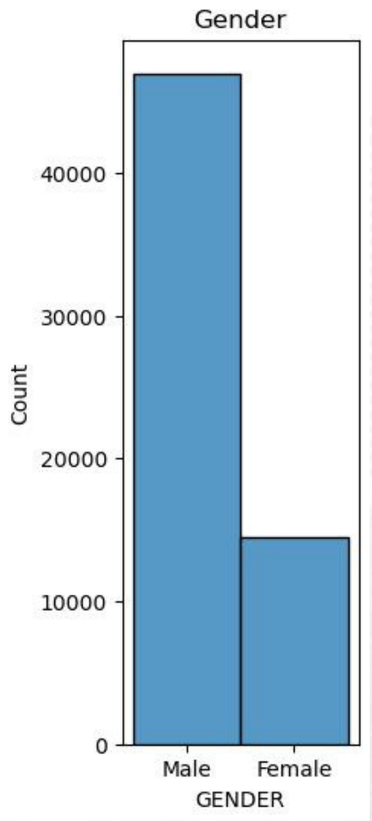


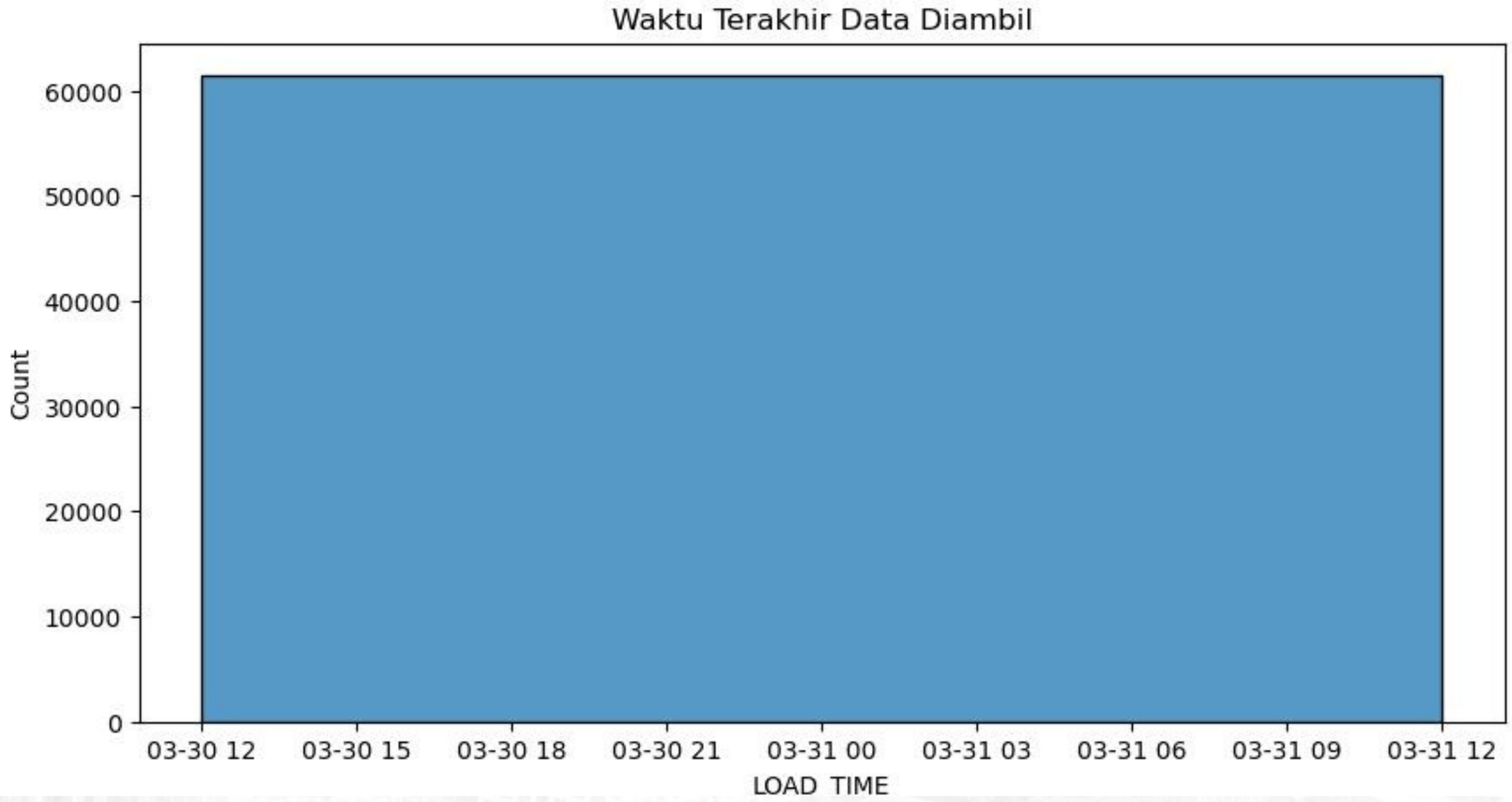




Ketiga fitur ini memiliki unique value yang sangat banyak.

Univariate Data Analysis





Kebanyakan member adalah laki-laki

Fitur Load_Time hanya memiliki satu value yaitu 03-31 00

Pre-Processing II

Feature Selection

Menggunakan **RFM** kita akan menentukan fitur apa saja yang akan kita ikut sertakan dalam proses clustering.

- Recency -> Kapan terakhir kali member melakukan penerbangan, bersama dengan order terakhirnya. 'LAST_TO_END'
- Frequency -> Berapa kali member melakukan penerbangan. 'FLIGHT_COUNT'
- Monetary -> Seberapa besar **Purchase Power** member? 'SEG_KM_SUM'

Kita akan menambahkan fitur 'avg_discount' dan 'Points_Sum' yang berisikan informasi penting terhadap kontribusi member kepada maskapai penerbangan.

Data Scaling

Rakamin Academy

Before scaling

	LAST_TO_END	FLIGHT_COUNT	SEG_KM_SUM	avg_discount	Points_Sum
count	61434.000000	61434.000000	61434.000000	61434.000000	61434.000000
mean	173.566006	11.939463	17275.642608	0.721644	12660.379480
std	181.479396	14.118850	21054.586002	0.184850	20615.000973
min	1.000000	2.000000	368.000000	0.000000	0.000000
25%	28.000000	3.000000	4848.250000	0.612500	2838.000000
50%	107.000000	7.000000	10149.000000	0.711779	6420.000000
75%	262.000000	15.000000	21430.750000	0.808988	14422.750000
max	731.000000	213.000000	580717.000000	1.500000	985572.000000

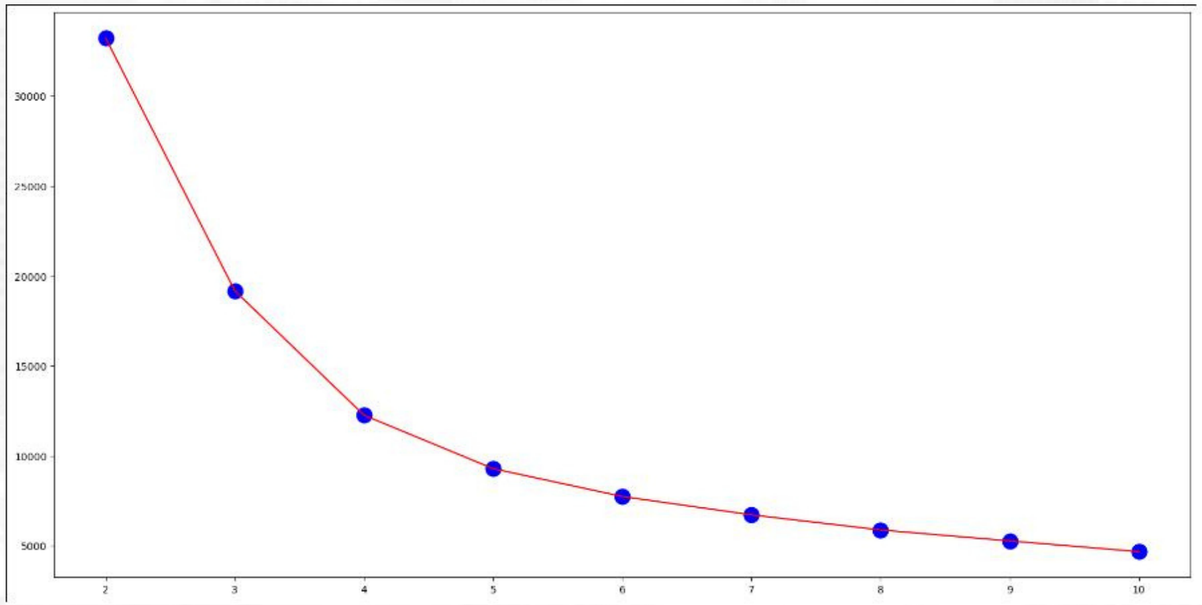
After scaling

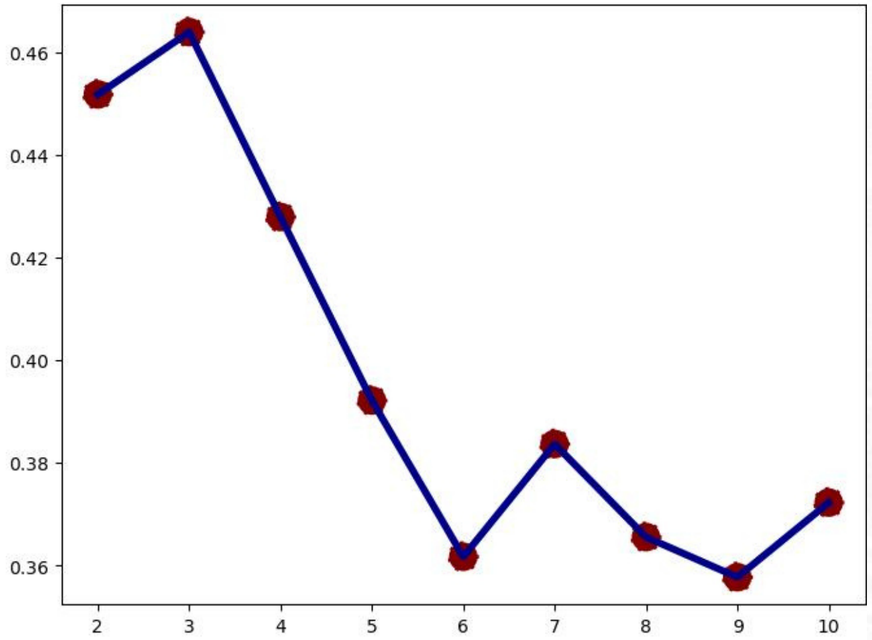
	LAST_TO_END	FLIGHT_COUNT	SEG_KM_SUM	avg_discount	Points_Sum
count	61434.000000	61434.000000	61434.000000	6.143400e+04	61434.000000
mean	0.236392	0.047106	0.029134	1.080101e-15	0.012846
std	0.248602	0.066914	0.036279	1.000008e+00	0.020917
min	0.000000	0.000000	0.000000	-3.903983e+00	0.000000
25%	0.036986	0.004739	0.007720	-5.904505e-01	0.002880
50%	0.145205	0.023697	0.016854	-5.336874e-02	0.006514
75%	0.357534	0.061611	0.036293	4.725202e-01	0.014634
max	1.000000	1.000000	1.000000	4.210790e+00	1.000000

Clustering

Clustering

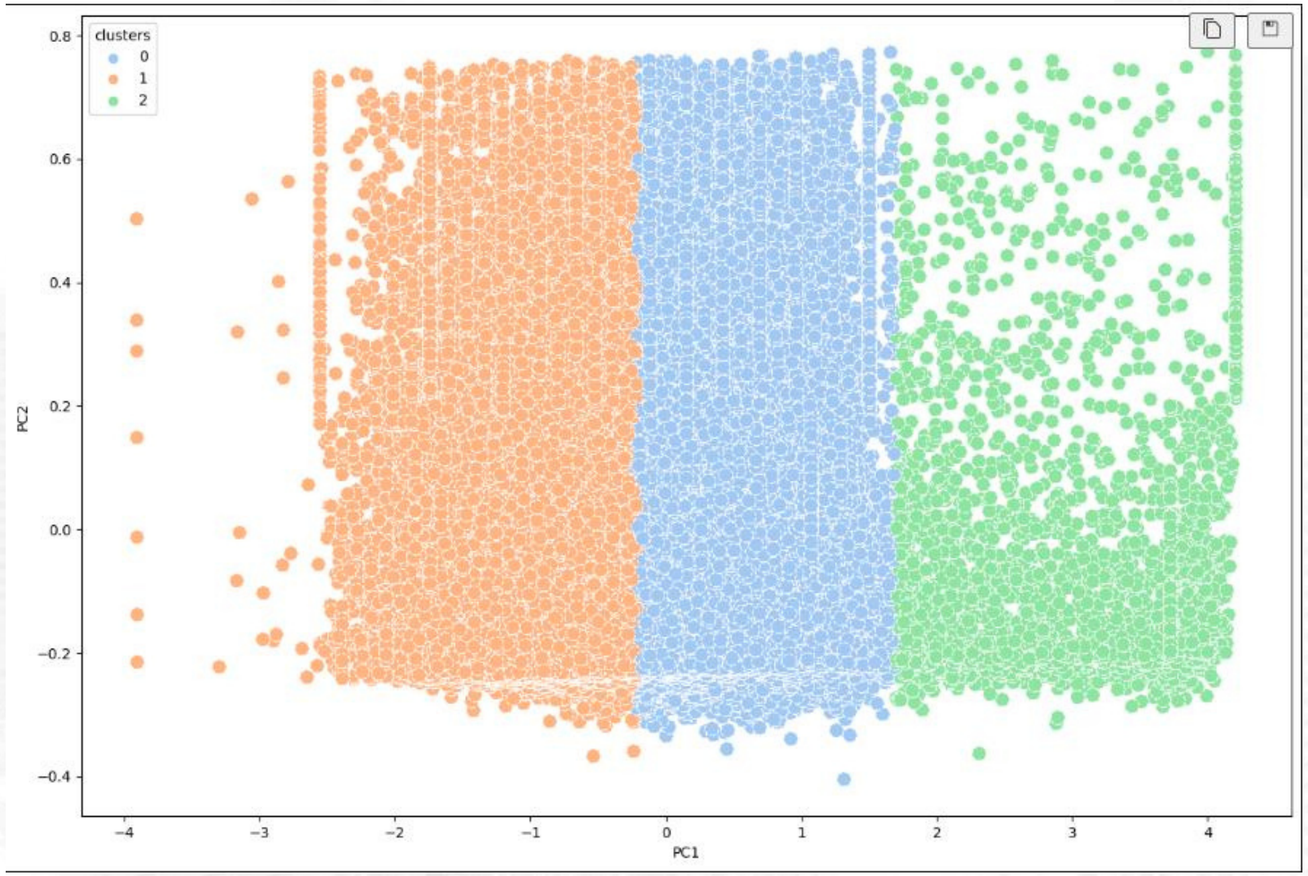
Kita akan menentukan berapa cluster yang akan kita *assign* dalam dataset ini menggunakan Elbow Method dan Silhouette Score.





Dari grafik di atas kita menentukan untuk mengambil cluster sejumlah 3.

Cluster Scatter Plot with PCA.



clusters	LAST_TO_END	FLIGHT_COUNT	SEG_KM_SUM	avg_discount	Points_Sum
0	102.0	8.0	10864.0	0.781283	8168.0
1	114.0	6.0	9140.0	0.590000	4243.5
2	78.0	9.0	12885.0	1.274964	16220.0

```
new_df['clusters'].value_counts()
```

```
0    33363
1    25496
2     2575
Name: clusters, dtype: int64
```

1. Cluster 0 :
- Merupakan member Tier tengah dengan akumulasi point ~8000.
- Merupakan member Tier tengah dengan rata-rata diskon di ~ 0.78
- Merupakan cluster dengan member paling banyak.
- Member dengan jarak penerbangan middle.
- Member dengan frekuensi penerbangan middle.

2. Cluster 1 :
- Merupakan member Tier rendah dengan akumulasi point ~4000
- Merupakan member penerbangan kelas rendah dengan rata-rata diskon di ~ 0.59 ()
- Merupakan cluster dengan member sedang.
- Member dengan jarak penerbangan paling rendah.
- Member dengan frekuensi penerbangan paling sedikit.

3. Cluster 2 :
- Merupakan member Tier atas dengan akumulasi point ~16000
- Merupakan member penerbangan kelas atas dengan rata-rata diskon di ~1.27
- Merupakan Cluster dengan member paling sedikit.
- Member dengan jarak penerbangan paling jauh.
- Member dengan frekuensi penerbangan paling banyak.