Passenger Segmentation in Airline : A Clustering Analysis

Suci Rahma Nura

Suci Rahma Nura

- Data Science Project Based Intern IDX Partners X Rakamin
- Data Science Intern 360DigiTMG
- Credit Analyst PT. BFI Finance Indonesia, Tbk
- Bachelor's degree in Mathematics Andalas University

Table of contents

0. Flowchart Modelling

Step 1

Data Understanding

- Missing and duplicate values check

Step 2

Exploratory Data Analysis

- Statistical Summary
- Univariate, Bivariate and Multivariate Analysis

Step 3

Data Preprocessing

- Missing Value and Outliers Handling
- Selection columns for RFM Modelling
- Feature Scalling

Step 4

Modeling and Evaluation

- Elbow Method
- Clustering Visualization

Step 5

Analysis and Recommendations

- Insight
- Recommendations

01

Background and Objective

Passengers

Passengers are crucial to airlines because they are the **primary source of revenue**, and their satisfaction drives loyalty and repeat business.

Business Background

Airlines must **prioritize** passenger satisfaction, service quality, and reputation in a competitive industry. Clustering passengers based on their characteristics helps **improve** personalized services, pricing strategies, and operational efficiency.

Objective

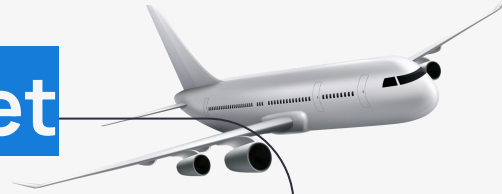
For grouping segment customers based on their criteria so that the company can take appropriate actions for each passenger group, **ultimately benefiting** the company.

02 Data Understanding

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

Dataset



- The dataset was sourced from ww.Kaggle.com.
- Have 62.988 **rows** and 23 **column**

Missing Value Checking

In [ ]:
```
data_null = df.isnull().sum().reset_index()
data_null.columns = ['feature','missing_value']
data_null['percentage'] = round((data_null['missing_value']/len(df))*100,2)
data_null = data_null.sort_values('percentage', ascending=False).reset_index(drop=True)
data_null = data_null[data_null['percentage']>0]
data_null
```

Out[48]:

	feature	missing_value	percentage
0	WORK_PROVINCE	3248	5.16
1	WORK_CITY	2269	3.60
2	SUM_YR_1	551	0.87
3	AGE	420	0.67
4	SUM_YR_2	138	0.22
5	WORK_COUNTRY	26	0.04

We will identify the chategory of missing value

Duplicate Data

In [ ]:
```
df.duplicated().sum()
```

Out[50]:
```
0
```

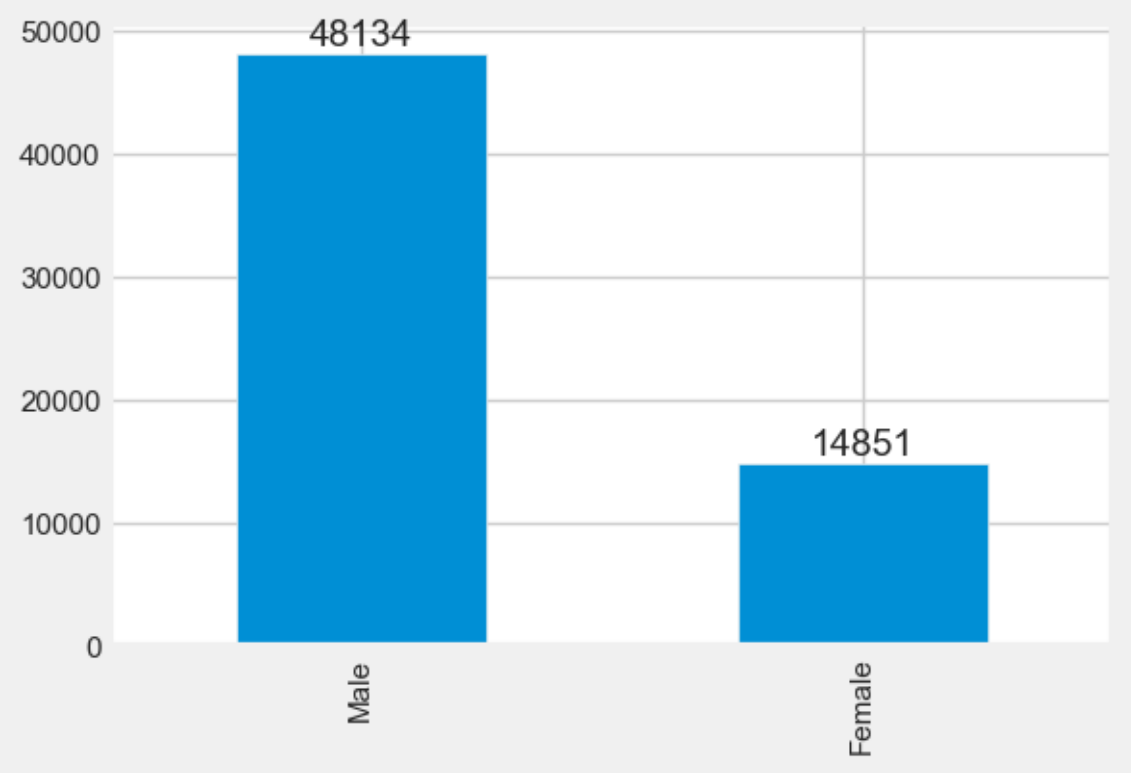
Tidak terdapat duplicate data

Dataset

- There are **some missing values**. And they will be handled in the next step.
- There is **no duplicate** values

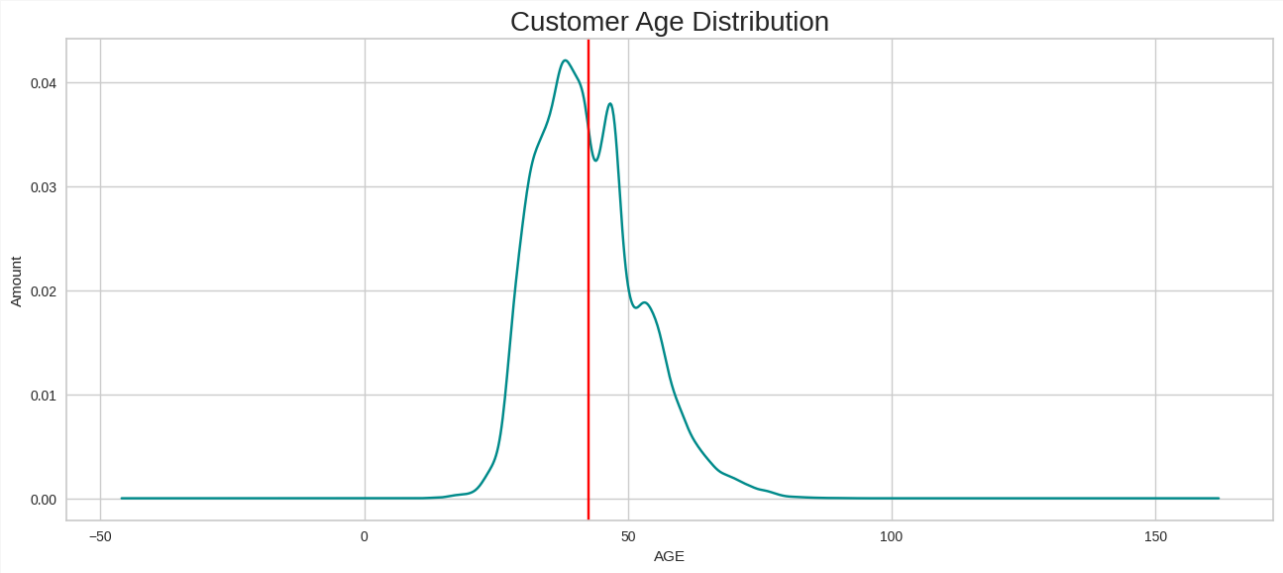
03 Exploratory Data Analysis

Distribution by Gender



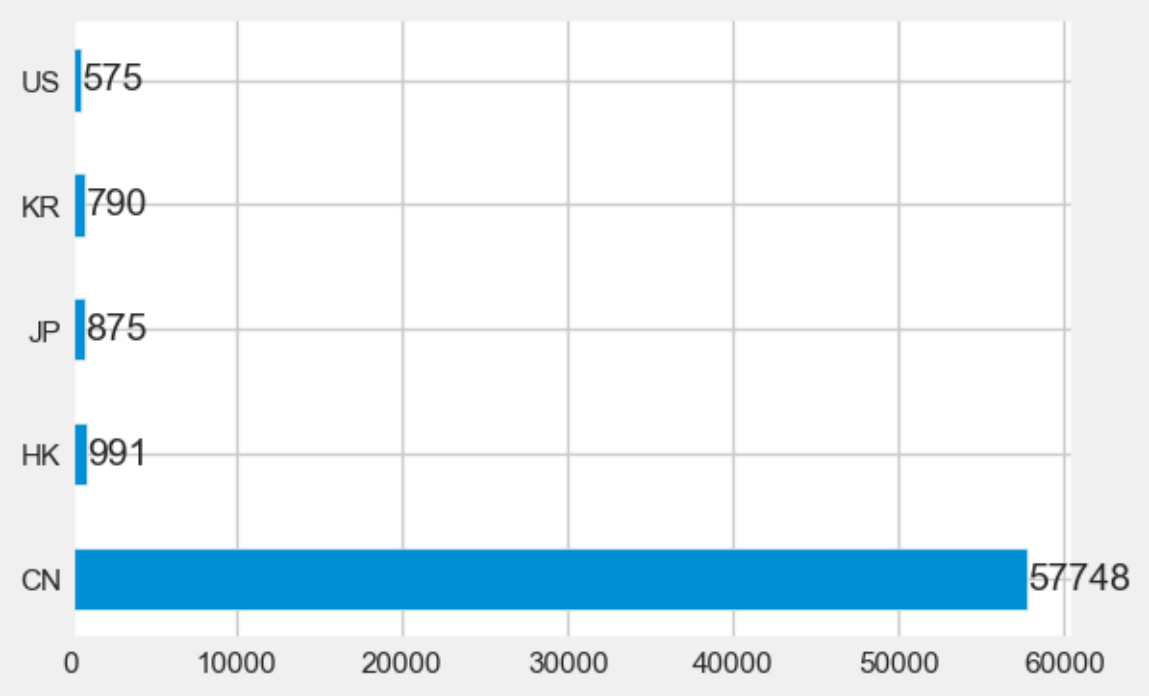
Based on gender, **most** customers are male.

Distribution by Age



The **average age** of customers is around 41 years old

Distribution by Countrty



In the top 5 countries that use airlines services, **CN** is the number 1 airlines user followed by HK and JP

04

Data Pre Processing

Missing Value and Outliers Handling

Missing Values

"The dataset contains **more than 5%** missing values. We will address the missing data based on the type of missing data, either by using the **mode** or by filling in with '**other**'."

Outliers

Handling Outlier using **IQR (Inter-Quartile Range) :**

$$IQR = Q_3 - Q_1$$

Q_3 : Quantile 75th
Q_1 : Quantile 25th

Feature Scalling : Standar Scalling

Standard Scaling

```
# because it is unsupervised so there is no need to split the data.

from sklearn.preprocessing import StandardScaler
df_std = df_handling.copy()
scale = StandardScaler()
kolom_all = [x for x in ['LAST_TO_END','FLIGHT_COUNT','SEG_KM_SUM']]
for kolom in kolom_all:
  df_std[kolom] = scale.fit_transform(np.array(df_std[kolom]).reshape(-1,1))
df_std.describe()
```

	MEMBER_NO	LAST_TO_END	FLIGHT_COUNT	SEG_KM_SUM
count	54366.000000	5.436600e+04	5.436600e+04	5.436600e+04
mean	31502.348821	-2.509366e-17	3.345822e-17	-5.018732e-17
std	18176.328744	1.000009e+00	1.000009e+00	1.000009e+00
min	1.000000	-1.042134e+00	-9.658866e-01	-1.213653e+00
25%	15827.250000	-8.244570e-01	-8.241034e-01	-7.766996e-01
50%	31527.000000	-3.331287e-01	-3.987538e-01	-3.201364e-01
75%	47247.750000	5.500183e-01	4.519454e-01	5.165826e-01
max	62988.000000	2.739228e+00	3.429393e+00	3.397366e+00

Variables with different scales (*check in range minimum values to maximum values each columns)* can **heavily** impact certain machine learning algorithms. Rescaling variables to have similar scales helps avoid **bias** in the model.

What is RFM ?

Recency

How **recent** or current a passenger's interaction or transaction with the airline is.

Frequency

How **often** passengers travel with the airline within a specific time frame.

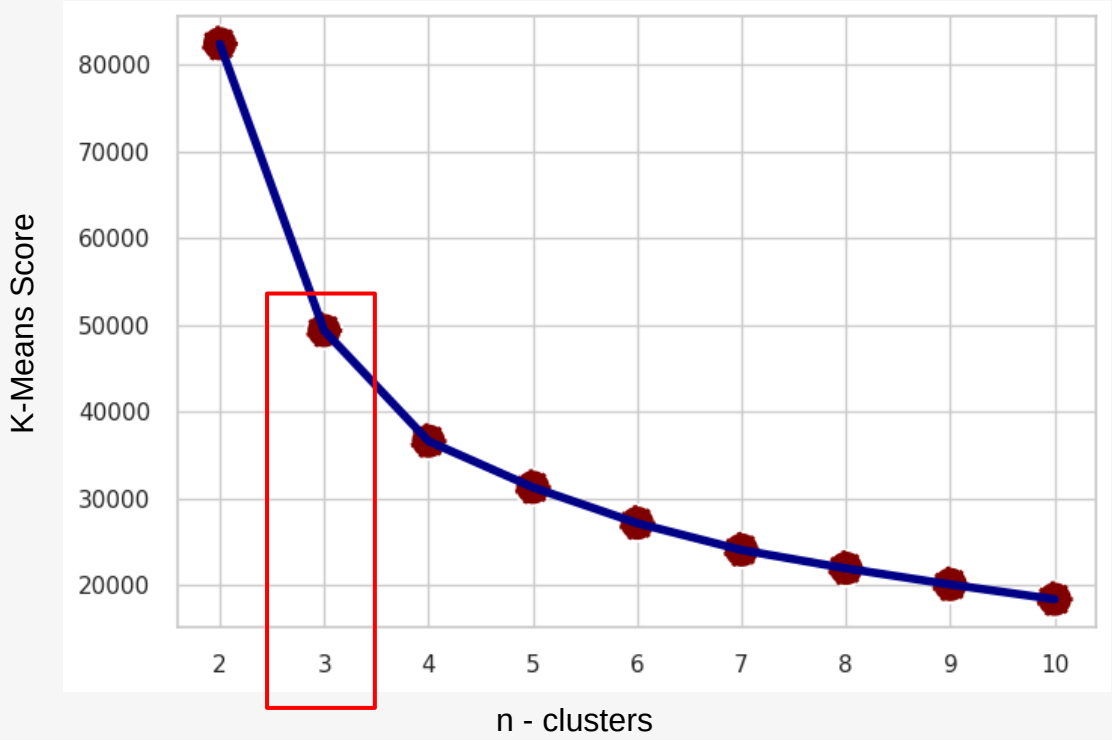
Monetary

The flight **distance** traveled by passengers.

05

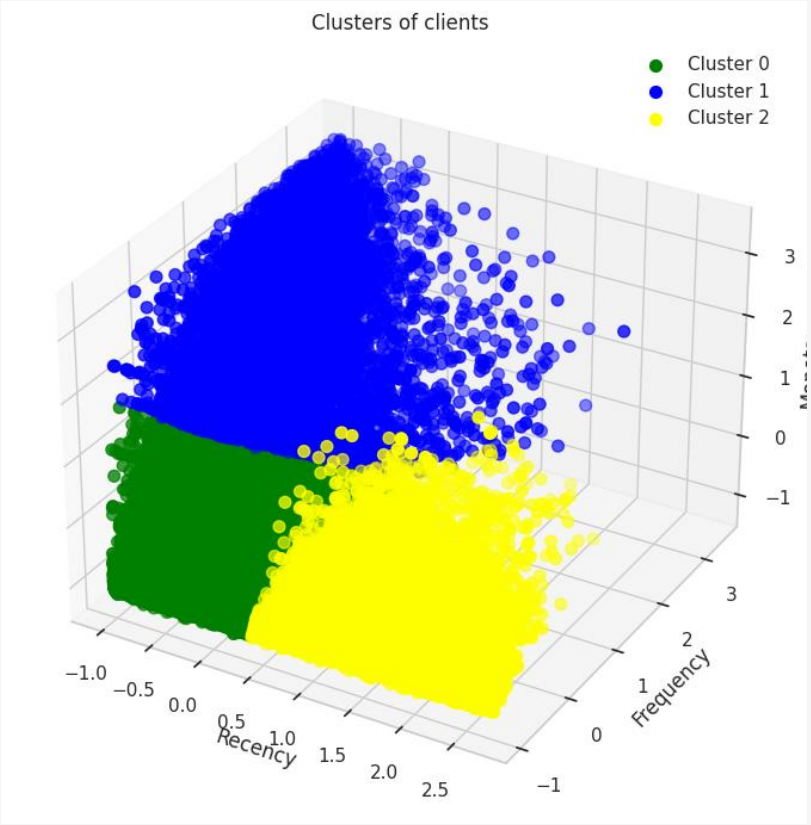
Modeling and Analysis

Elbow Method : to find the cluster



According to the graph, the best K value for **K-Means Clustering** is K = 3.

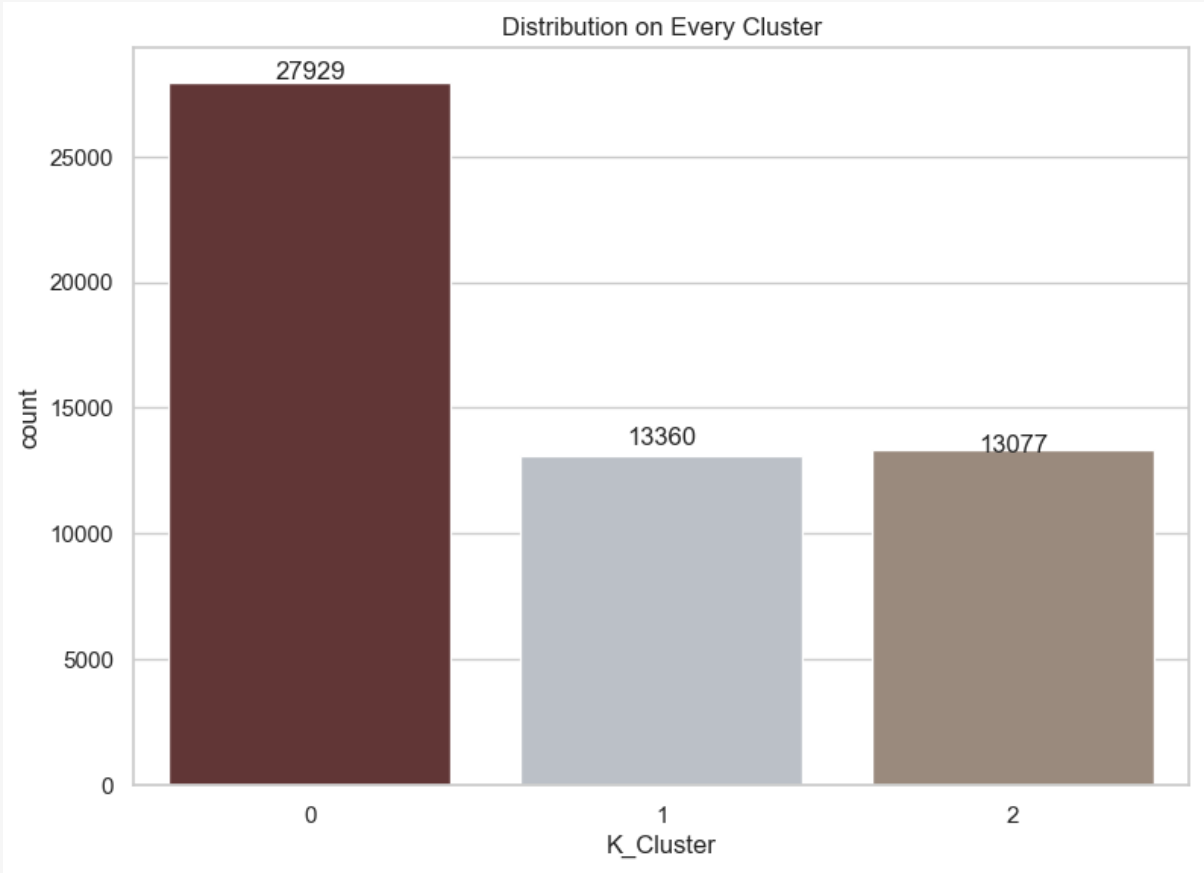
Visualization of the three clusters.



The results is best because of the **boundaries** can be **distinguished**

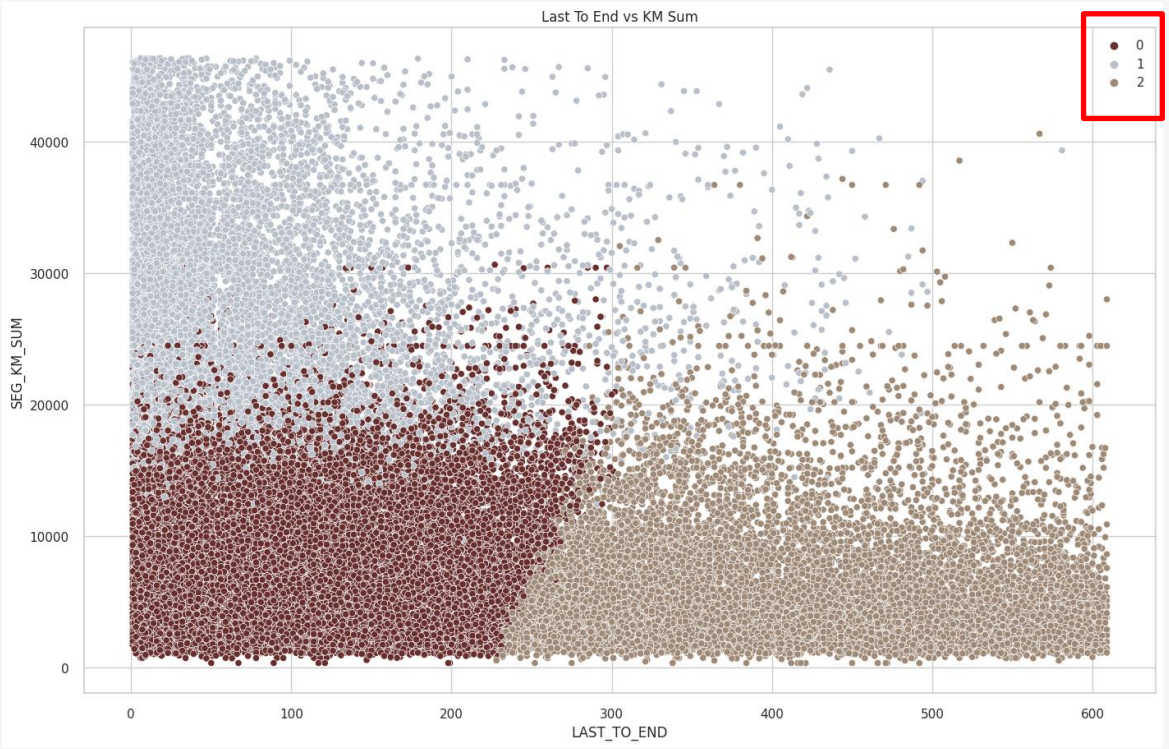
```
#plot 3D dengan hasil cluster : RFM
fig = plt.figure(figsize=(12,8))
dx = fig.add_subplot(111, projection='3d')
colors = ['green', 'blue', 'yellow', 'red','black']
for i in range(0,3):
    dx.scatter(data_cluster[data_cluster.K_Cluster == i].LAST_TO_END,
    data_cluster[data_cluster.K_Cluster == i].FLIGHT_COUNT,
    data_cluster[data_cluster.K_Cluster == i].SEG_KM_SUM,
            c = colors[i], label = 'Cluster ' + str(i), s=50)
dx.set_title('Clusters of clients')
dx.set_xlabel('Recency')
dx.set_ylabel('Frequency')
dx.set_zlabel('Monetary')
dx.legend()
plt.show()
```

Visualization of the three clusters.



From the data displayed the **most** clusters are at cluster **0** followed by cluster 2, and then 1

Visualization Clustering Recency & Monetary Value



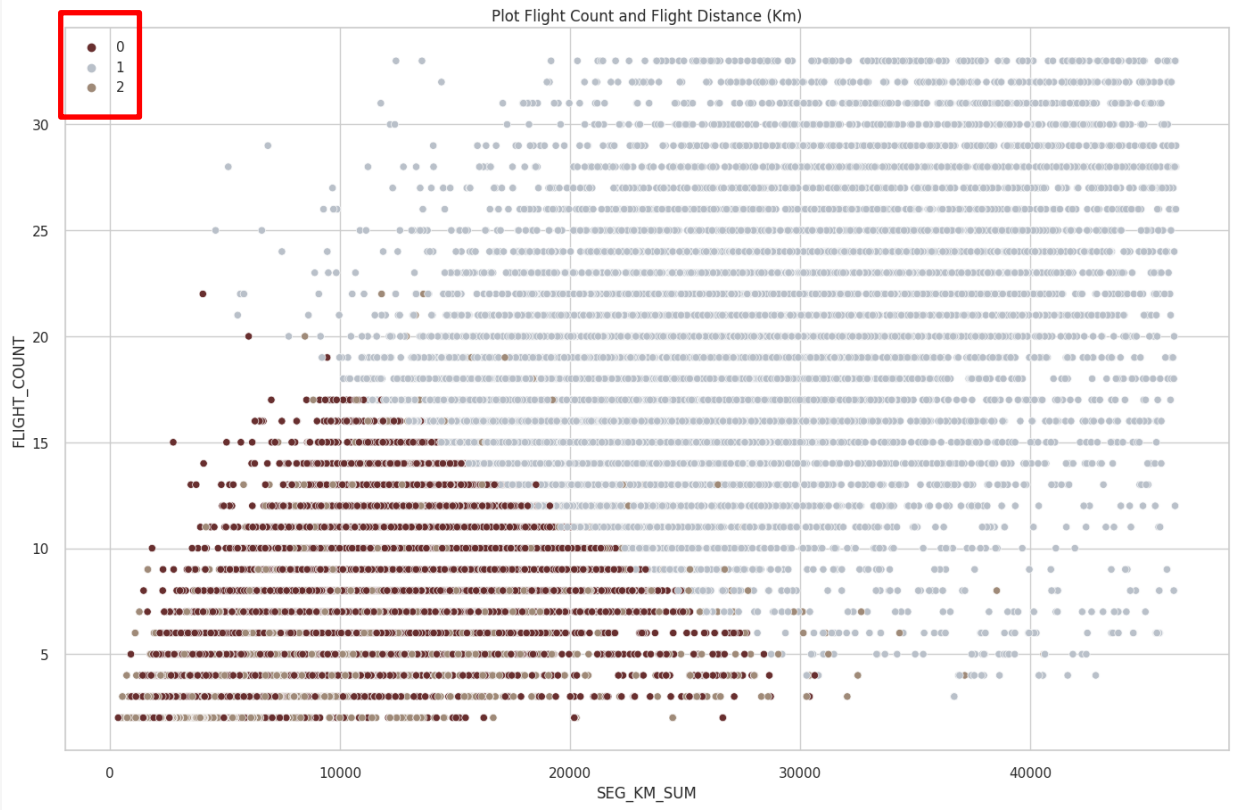
3 clusters

0 : New customer with **low** flight distance

1 : New customer with **high** flight distance

2 : Old customer with **low** flight distance

Visualization Clustering Recency & Frequency



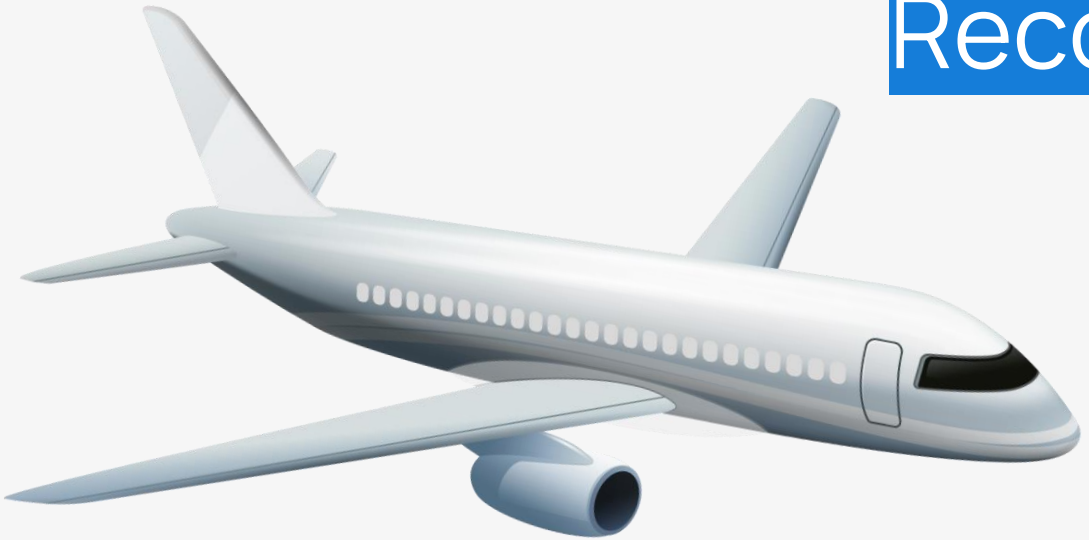
3 clusters

0 and **2** : New customer with **low** flight count tends to **high** flight distance

1 : New customer with **high** flight count and flight distance

06

Analysis and Recommandations



Analysis

Cluster 1

high value customers

Profiling

- Lowest level flight recency (most recent)
- Highest level of flight frequency
- Highest level of airline mileage
- Longest membership duration
- Lowest number of cluster members

Cluster 0

middle value customer

Profiling

- Mid-level flight recency
- Lowest to middle airline mileage
- Membership duration between cluster 1 and cluster 2
- Highest number of cluster members

Cluster 2

low value customer

Profiling

- Highest level flight recency (longest)
- Lowest rate flight frequency
- Lowest level of airline mileage
- In the average discount level, it shows the lower level flight class
- Most recent membership duration
- Middle number of cluster members

Recommendations

Cluster 1

high value customers

They are the most ideal type of customer, they contribute the most to the airline. They are also loyal. Airline needs to provide special management for these customers and improve their satisfaction, such as give **a free trials services** in term of some condition.

Cluster 0

middle value customer

As a potential customer, airline can encourage these customers to increase transactions. We have to give more offers such as **cashback or point** that make the customer have a feeling or a rush to use the offers again in the airlines.

Cluster 2

low value customer

Airline needs to **increase interaction** with these customers and needs to take certain marketing strategies to extend this customer cycle. Give a promotion that such as **discount or holiday voucher** to get the customers back to use Airlines.

Thank You!

Feel Free to Contact me!

sucirahma.srn@gmail.com

https://github.com/eseren

https://www.linkedin.com/in/sucisrn/