

Airline Customer Clustering using K-Means Algorithm

Unsupervised Machine Learning
Project Homework



Content Direction

- 
- A blurred background image showing a person's hands typing on a laptop keyboard, with a white coffee mug to the left.
- 1 EDA
 - 2 Pre-Process & Feature Engineering
 - 3 Modeling & Evaluasi
 - 4 Interpretasi Hasil

Exploratory Data Analysis

Descriptive Analysis

1

1 Feature & Labels

- 22 *usable features*, mengecualikan '**MEMBER_NO**'
Categorical : 8 Features, **Numerical** : 14 Features

2 Missing Values

- Terdapat Null Values pada '**SUM_YR_1, AGE, SUM_YR_2, WORK_PROVINCE, WORK_CITY, WORK_COUNTRY, GENDER**' dengan presentase tidak mencapai 10%

3 Weird Column Summaries

- **FFP_TIER** memiliki *unique_values* yang sedikit untuk dikategori sebagai *numerical*
- **EXCHANGE_COUNT** memiliki nilai yang sama dari min – Q3 yaitu 0
- **WORK_PROVINCE, WORK_CITY, WORK_COUNTRY** menunjukan *customer CN* mendominasi data

4

Duplicated Data

- Data tidak memiliki nilai duplikat

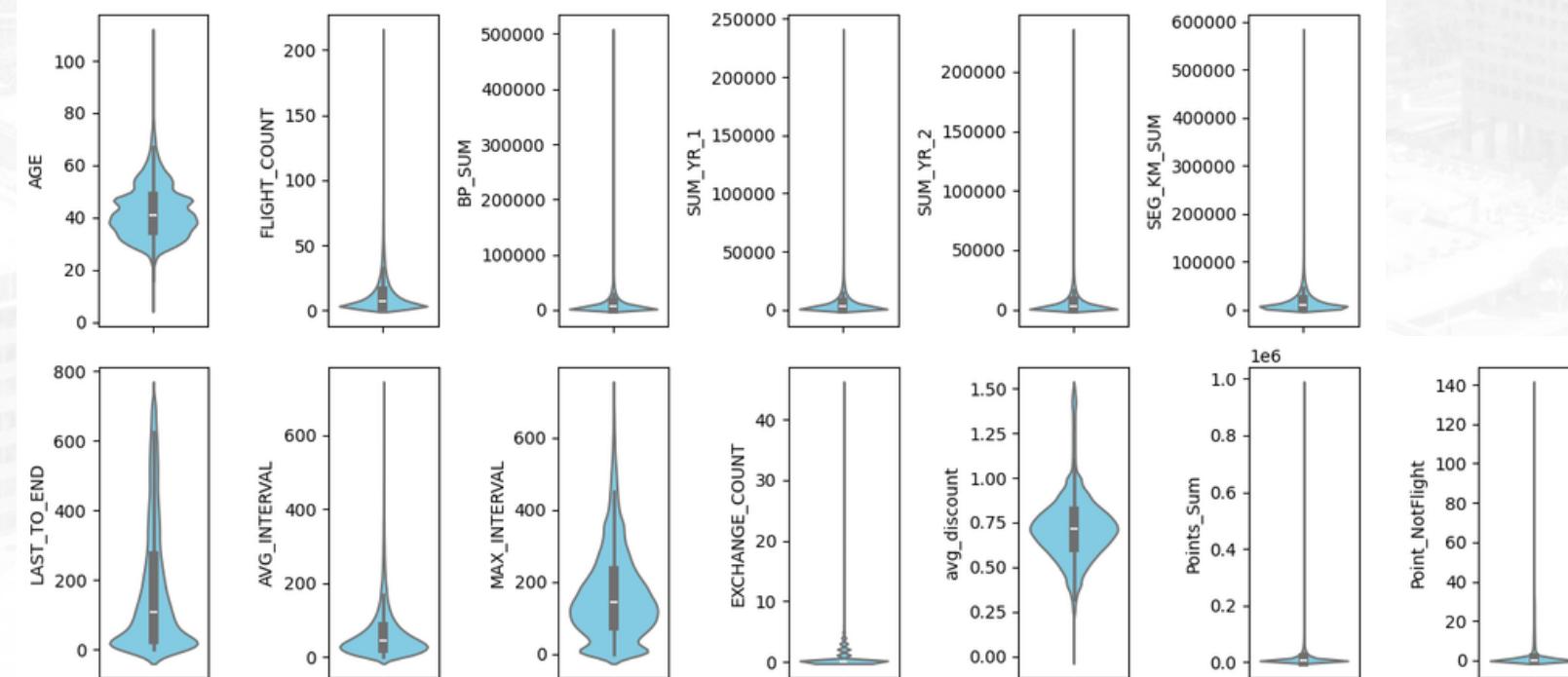
Exploratory Data Analysis

Univariate Analysis

1

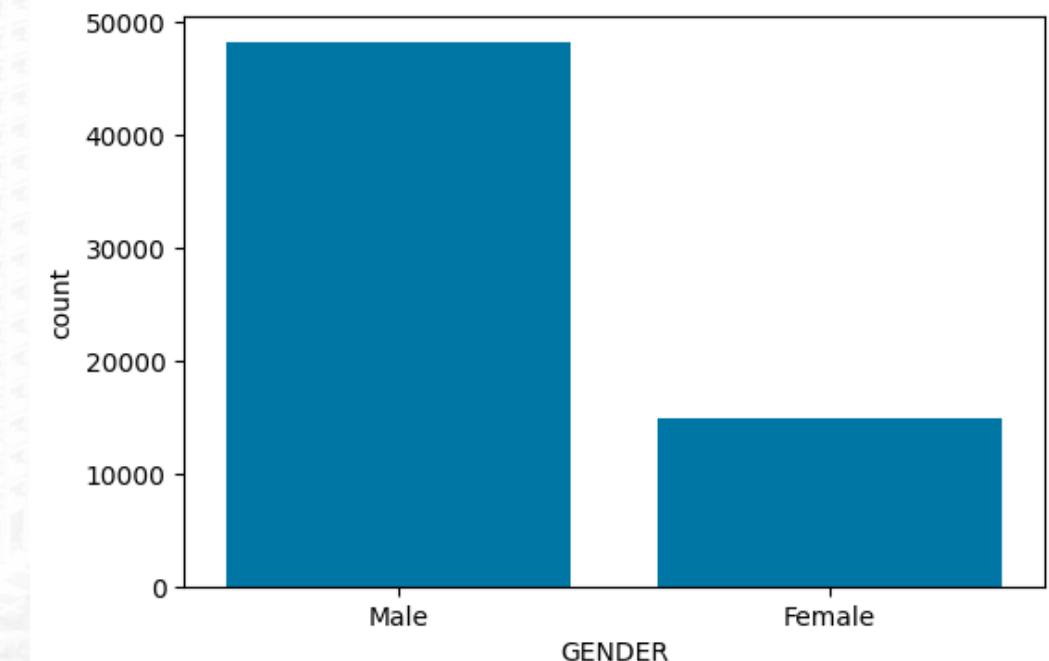
1 Numerical Feature Plot

- Semua numerical Feature memiliki outlier secara statistika
- Kebanyakan distribusi *right skewed* kecuali AGE & AVG_DISCOUNT yang normal



2 Categorical Feature Plot

- Semua **Categorical Feature** kecuali **Gender** memiliki terlalu banyak *Unique Values* untuk **WORK area** & bertipe **DateTime** sehingga untuk dapat di plot perlu dirubah bentuknya



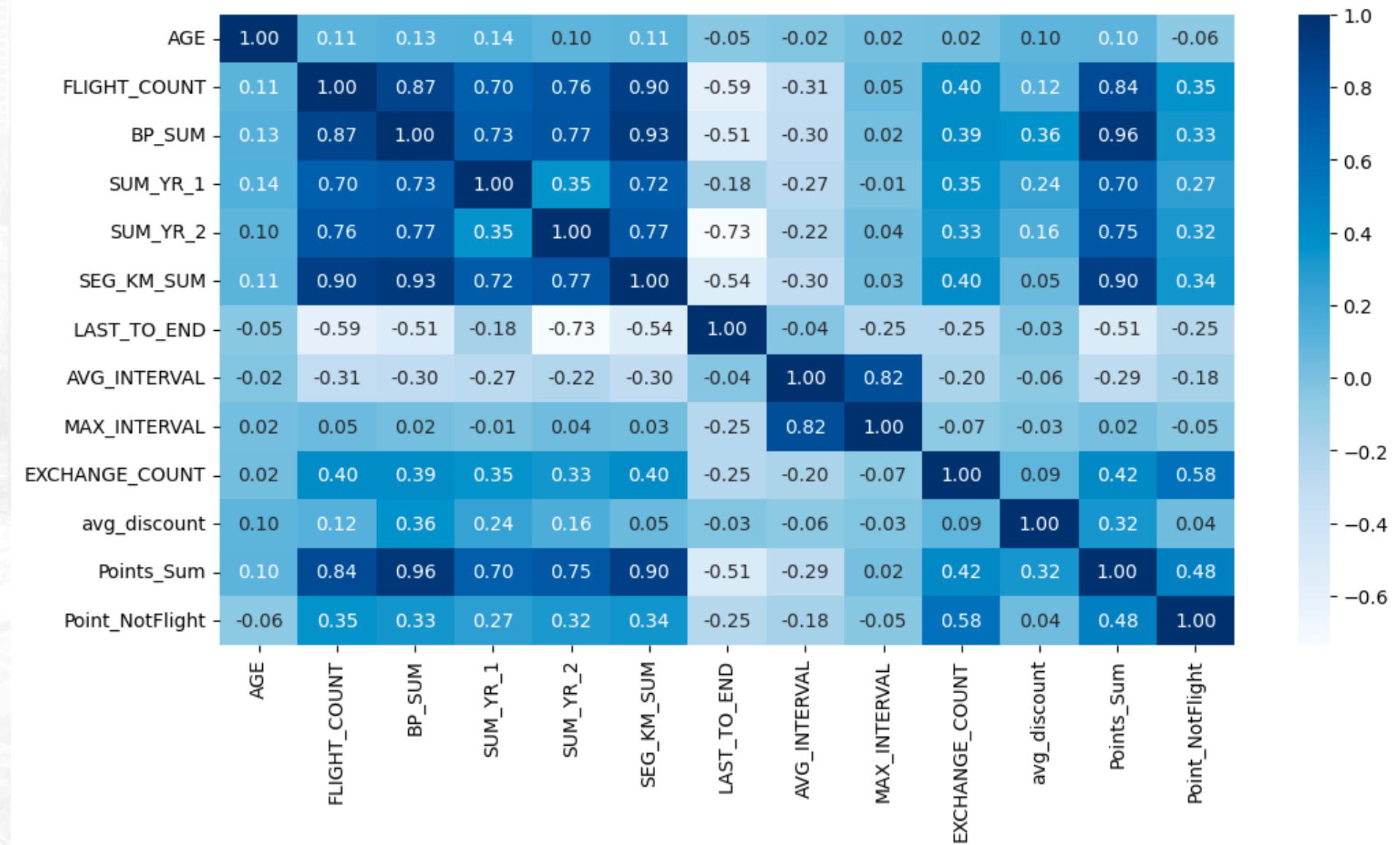
Exploratory Data Analysis

Multivariate Analysis

1

1 Correlation Heatmap Numerical Feature

- Banyak Feature yang memiliki korelasi lebih dari 0.7 dan dicurigai sebagai feature redundan, oleh karena itu akan dilakukan *feature selection* yang lebih baik di *pre-processing*.



Pre-Processing

Data Cleansing

2

1 Missing Values Handling

Drop the Feature
menggunakan dropna

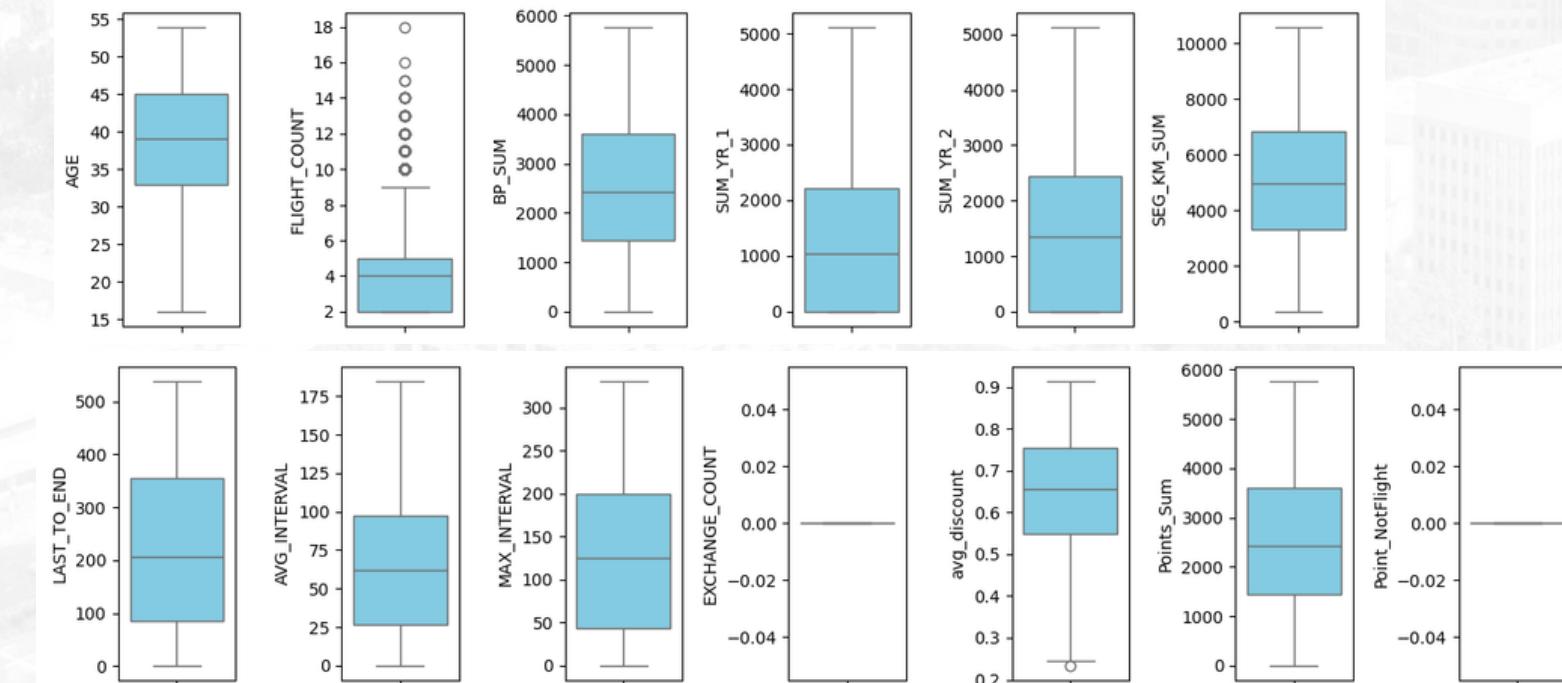
2 Duplicated Data Handling

Tidak Perform kerena data
tidak memiliki duplication

3 Outlier Handling

After Outlier handling

Perform menggunakan IQR satu iterasi,
EXCHANGE_COUNT & Point_NotFlight,
menjadi *feature* yang akan di *drop* karena
hanya memiliki satu nilai saja setelah *handling*



Pre-Processing

Data Cleansing

2

4 Feature Extraction

*FFP_DATE, FIRST_FLIGHT_DATE,
LOAD_TIME, LAST_FLIGHT_DATE*

*WORK_CITY, WORK_PROVINCE,
WORK_COUNTRY*

Awalnya berupa object value kemudian diubah menjadi DATETIME & **hanya diambil tahunnya saja** agar dapat lebih mudah dipahami

Feature tersedia dalam bentuk object dengan unique value yang sangat banyak hingga lebih dari 100, sehingga agar lebih mudah dipahami **diambil 4 unique value teratas kemudian sisanya dijumlahkan dan di-extract menjadi nilai baru dengan nama 'other'**

Dari yang telah diubah LOADTIME kemudian menjadi **feature yang di drop** karena hanya menjelaskan tanggal data diubah menjadi csv, sehingga semua baris memiliki nilai yang sama.

Selain itu berdasarkan **3 feature WORK area, hanya dipilih WORK_COUNTRY sebagai representatif** dan sisanya di drop. Alasannya karena country saja telah representatif.

Pre-Processing

Feature Engineering

2

1 Feature Encoding

6 categorical feature
19 categorical feature

| | | |
|---------------------|---------|------------------|
| Categorical Feature | Ordinal | label encoding |
| | Nominal | One Hot Encoding |

2 Feature Transformation StandardScaler()

Diterapkan pada semua kolom kecuali untuk *categorical feature* yang menggunakan OHE. Dengan tujuan menyamakan standar deviasi menjadi 1 dan mean menjadi 0.

3 Feature Selection

Before Pre-Processing 22 feature

After Pre-Processing **29 feature**

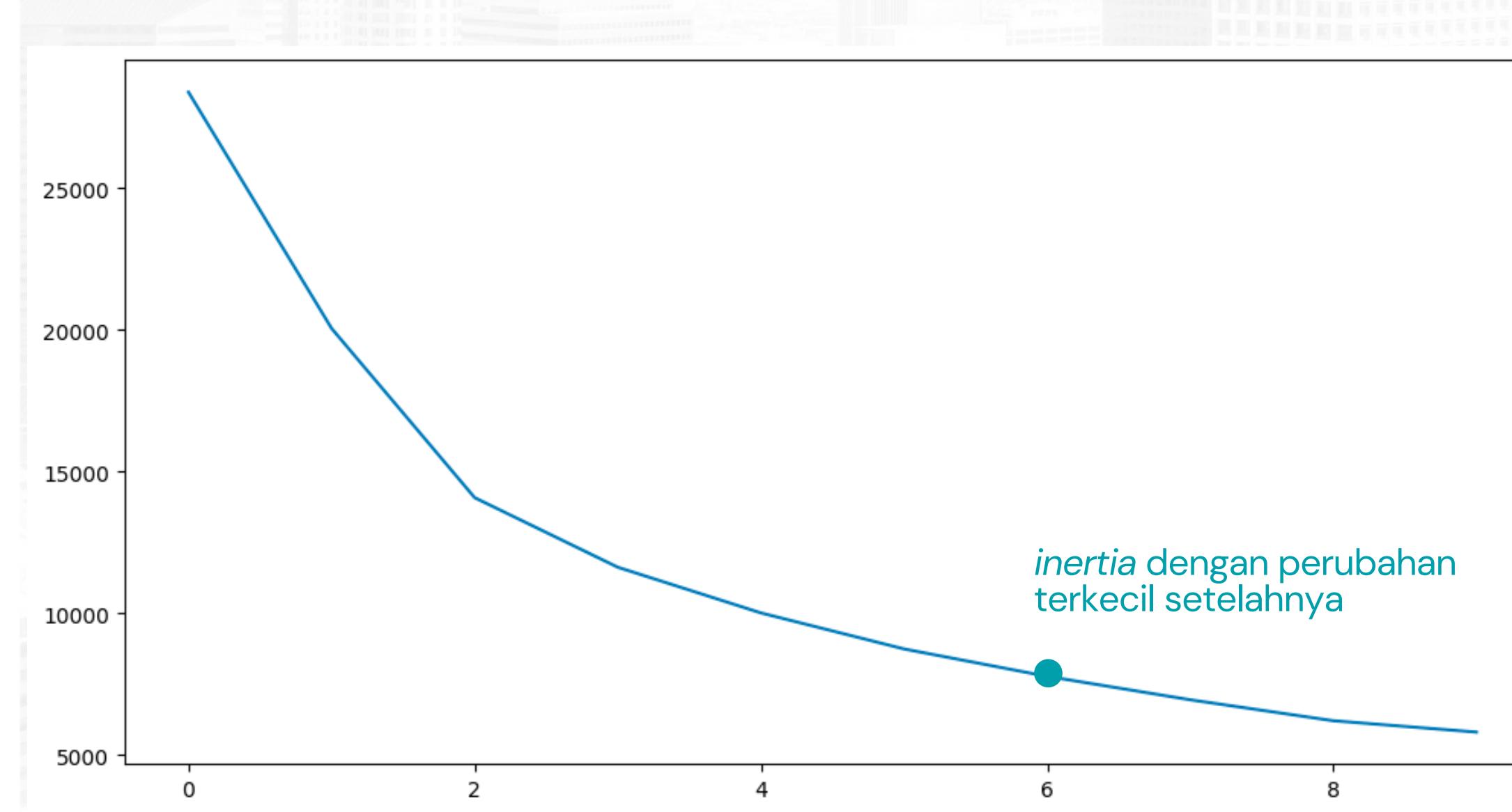
Used to produce Cluster 5 Feature

AGE, gender_Male, gender_Female,
FLIGHT_COUNT, work_country_CN

Dipilih *feature* diatas adalah untuk dapat membagi kelompok *customer* dari *China* atas gender, usia, dan jumlah penerbangan

Modeling Elbow Method

3



Elbow method digunakan untuk menemukan jumlah cluster optimal.

Dengan *feature* yang dimiliki dan tujuan untuk klasifikasi *customer profile* dipilih rentang dari 2 hingga 10 agar tidak terlalu banyak.

Berdasarkan visualisasi *elbow method*, ditunjukkan bahwa nilai k dengan pengurangan inertia paling kecil berada pada **k=6**. Namun hasil ini belum terlalu jelas terlihat sehingga dilakukan pencarian sillhoute score lebih lanjut.

Modeling Silhouette score

3

```
For n_clusters=2, the silhouette score is 0.2833031
For n_clusters=3, the silhouette score is 0.3110320
For n_clusters=4, the silhouette score is 0.3244150
For n_clusters=5, the silhouette score is 0.3229782
For n_clusters=6, the silhouette score is 0.3388407
For n_clusters=7, the silhouette score is 0.3307019
For n_clusters=8, the silhouette score is 0.3181769
For n_clusters=9, the silhouette score is 0.3387377
For n_clusters=10, the silhouette score is 0.336220
```

Berikut merupakan hasil silhouette score dari pencarian jumlah *cluster* optimal.

Hasil terbaik nampak muncul pada jumlah *cluster* sebanyak 6 dengan score sejumlah 0.3388.

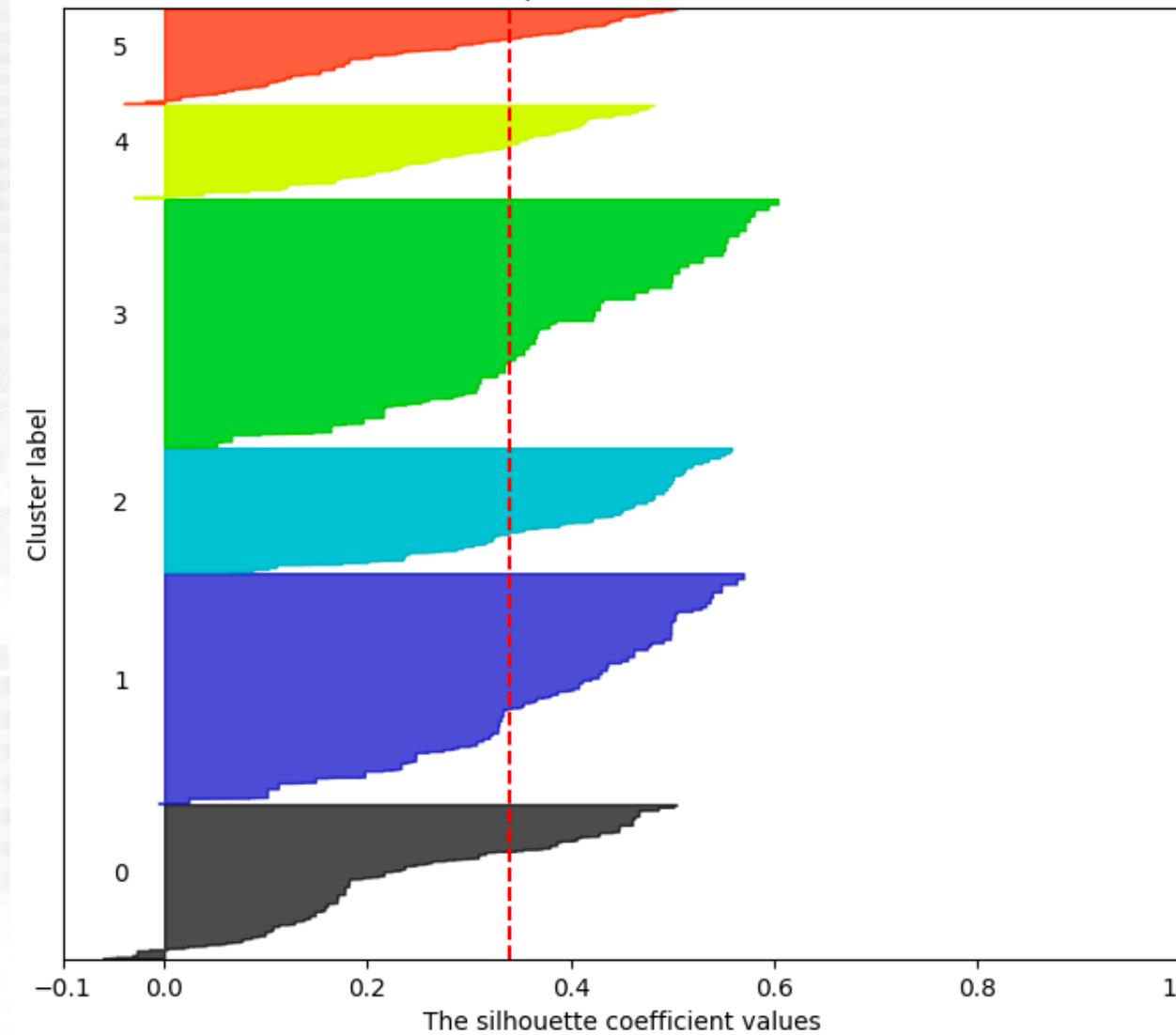
Maka dengan ini diputuskan the *optimal cluster number* adalah 6.

Modeling K-Means with the best cluster number

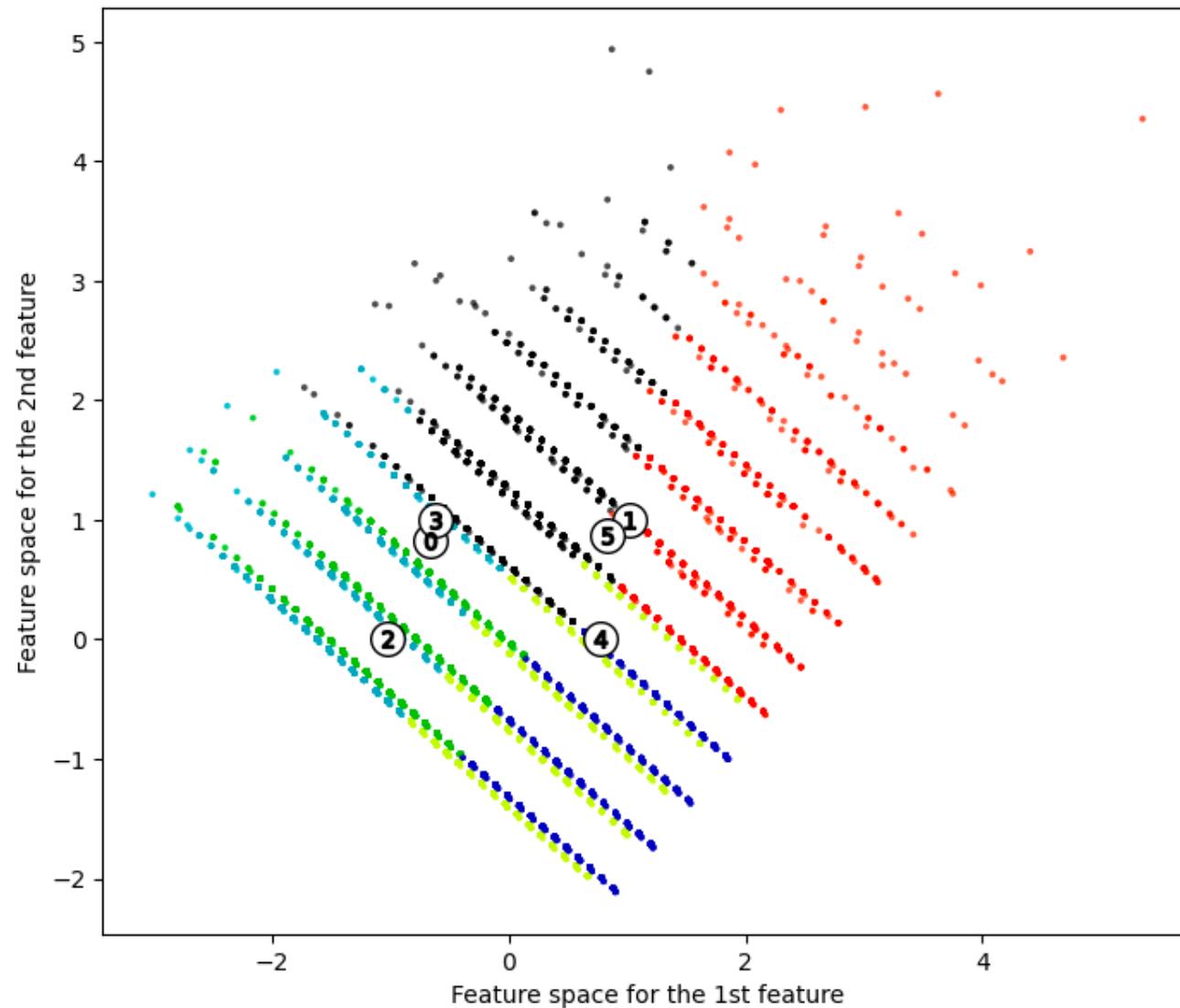
3

Silhouette analysis for KMeans clustering on sample data with n_clusters = 6

The silhouette plot for the various clusters.



The visualization of the clustered data.



Berikut merupakan penampakan *cluster* yang dihasilkan dari 5 *feature* yang telah dipilih.

Didapatkan *cluster* yang cukup terurai walaupun terdapat beberapa *cluster* dengan *centroid* yang berdekatan.

Hasil visualisasi didapatkan melalui penggunaan PCA terhadap hasil *fitting*.

Interpretasi Deskripsi Cluster & Rekomendasi

4

*Hasil interpretasi didapatkan dengan mengambil *describe* dari masing-masing cluster dan *inversed scaled feature AGE & FLIGHT_COUNT* sehingga dapat langsung diambil *insight*.

Cluster 0

| | AGE | gender_Male | gender_Female | FLIGHT_COUNT | work_country_CN | Cluster_Id |
|------|-----------|-------------|---------------|--------------|-----------------|------------|
| mean | 34.098884 | 0.826688 | 0.173312 | 6.223285 | 0.965444 | 0.0 |
| 50% | 34.000000 | 1.000000 | 0.000000 | 6.000000 | 1.000000 | 0.0 |
| std | 4.146622 | 0.378617 | 0.378617 | 1.248024 | 0.182701 | 0.0 |

Cluster 1

| | AGE | gender_Male | gender_Female | FLIGHT_COUNT | work_country_CN | Cluster_Id |
|------|-----------|-------------|---------------|--------------|-----------------|------------|
| mean | 46.660981 | 1.0 | 0.0 | 3.278962 | 0.891613 | 1.0 |
| 50% | 46.000000 | 1.0 | 0.0 | 3.000000 | 1.000000 | 1.0 |
| std | 3.666538 | 0.0 | 0.0 | 1.092588 | 0.310923 | 0.0 |

Cluster 2

| | AGE | gender_Male | gender_Female | FLIGHT_COUNT | work_country_CN | Cluster_Id |
|------|-----------|-------------|---------------|--------------|-----------------|------------|
| mean | 31.381328 | 0.0 | 1.0 | 3.111111 | 0.948060 | 2.0 |
| 50% | 32.000000 | 0.0 | 1.0 | 3.000000 | 1.000000 | 2.0 |
| std | 4.062896 | 0.0 | 0.0 | 1.110424 | 0.221978 | 0.0 |

Customer memiliki presentase asal 96% dari China dengan usia rerata 34, didominasi 82% pria & memiliki rata-rata jumlah penerbangan sebanyak 6

Customer memiliki presentase asal 90% dari China dengan usia rerata 47, didominasi 100% pria & memiliki rata-rata jumlah penerbangan sebanyak 3

Customer memiliki presentase asal 94% dari China dengan usia rerata 31, didominasi 100% wanita & memiliki rata-rata jumlah penerbangan sebanyak 3

Interpretasi Deskripsi Cluster & Rekomendasi

4

*Hasil interpretasi didapatkan dengan mengambil *describe* dari masing-masing cluster dan *inversed scaled feature AGE & FLIGHT_COUNT* sehingga dapat langsung diambil *insight*.

Cluster 3

| | AGE | gender_Male | gender_Female | FLIGHT_COUNT | work_country_CN | Cluster_Id |
|------|-----------|-------------|---------------|--------------|-----------------|------------|
| mean | 34.389512 | 1.0 | 0.0 | 2.801451 | 0.955475 | 3.0 |
| 50% | 35.000000 | 1.0 | 0.0 | 3.000000 | 1.000000 | 3.0 |
| std | 4.278301 | 0.0 | 0.0 | 0.811594 | 0.206293 | 0.0 |

Cluster 4

| | AGE | gender_Male | gender_Female | FLIGHT_COUNT | work_country_CN | Cluster_Id |
|------|-----------|-------------|---------------|--------------|-----------------|------------|
| mean | 44.875766 | 0.0 | 1.0 | 3.411199 | 0.926509 | 4.0 |
| 50% | 44.000000 | 0.0 | 1.0 | 3.000000 | 1.000000 | 4.0 |
| std | 4.602820 | 0.0 | 0.0 | 1.302946 | 0.261055 | 0.0 |

Cluster 5

| | AGE | gender_Male | gender_Female | FLIGHT_COUNT | work_country_CN | Cluster_Id |
|------|-----------|-------------|---------------|--------------|-----------------|------------|
| mean | 45.352487 | 0.874786 | 0.125214 | 7.749571 | 0.918525 | 5.0 |
| 50% | 46.000000 | 1.000000 | 0.000000 | 7.000000 | 1.000000 | 5.0 |
| std | 4.553556 | 0.331104 | 0.331104 | 1.732020 | 0.273681 | 0.0 |

Customer memiliki presentase asal 95% dari China dengan usia rerata 34, didominasi 100% pria & memiliki rata-rata jumlah penerbangan sebanyak 3

Customer memiliki presentase asal 93% dari China dengan usia rerata 45, didominasi 100% wanita & memiliki rata-rata jumlah penerbangan sebanyak 3

Customer memiliki presentase asal 92% dari China dengan usia rerata 45, didominasi 87% pria & memiliki rata-rata jumlah penerbangan sebanyak 8

Interpretasi Deskripsi Cluster & Rekomendasi

4

Berdasarkan hasil interpretasi *customer* beberapa insight menariknya dengan rekomendasi yang dapat diberikan adalah sebagai berikut.

1 Insights menarik

- 4 cluster yaitu **cluster 1 hingga 4** dari 6 yang tersedia merupakan *cluster* yang memiliki *customer* dengan **rata-rata jumlah penerbangan terendah yaitu sebanyak 3** selama periode penerbangan 2004 hingga 2014.
- **Cluster 5** merupakan *cluster* yang memiliki *customer* dengan **rata-rata jumlah penerbangan tertinggi sebanyak 8**.

2 Recomendation

- Berdasarkan profilenya, *customer* pria maupun wanita dari usia 31 hingga 47 tahun di China perlu diberikan **promosi khusus seperti benefit nilai penukaran poin yang lebih banyak** sehingga dapat meningkat dari segi rata-rata jumlah penerbangannya.
- Memastikan **memberikan pelayanan sesuai standar saat ini dengan kontrol yang lebih baik** untuk memastikan *customer retain* akan penggunaan layanan penerbangan.

Team Collaborator



Project Mentor



Project Leader/QA



Business Division



Technical ML Division



Thank You

[bagusatya08/K-Means-Clustering-...](#)

Implementation of Clustering Using K-Means Algorithm into Airline Customer Dataset

1 Contributor 0 Issues 0 Stars 0 Forks

bagusatya08/K-Means-Clustering-Airline_Customer:
Implementation of Clustering Using K-Means Algorithm...

Implementation of Clustering Using K-Means Algorithm into Airline Customer Dataset - GitHub - bagusatya08/K-Means-Clustering-Airline_Customer: Implementation of Clustering Using K-Means Algorith...



