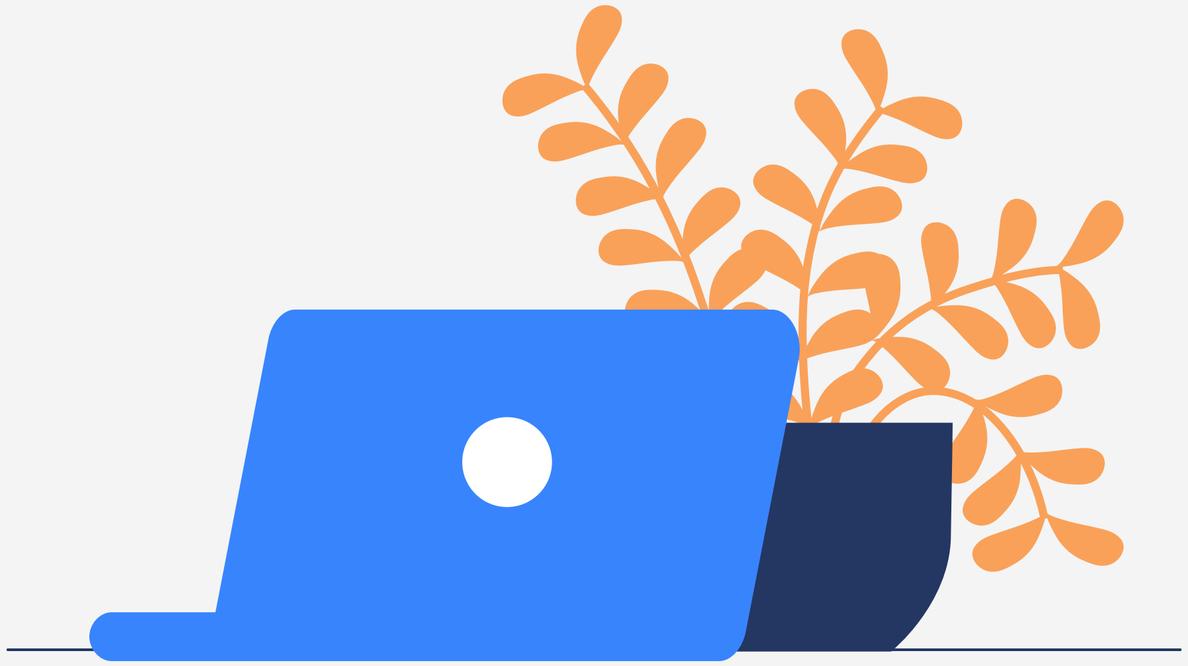
Kmeans Clustering for flight data

Intellegend

EDA



Data kosong dan duplikat

Tidak ada data yang duplikat, namun terdapat beberapa kolom yang memiliki missing value. Berikut cara penanganannya:

- GENDER, WORK_CITY, WORK_PROVINCE, WORK_COUNTRY : isi missing value dengan modus
- AGE, SUM_YR_1, SUM_YR_2 : isi missing value dengan median

Tipe data

Terdapat beberapa kolom dengan tipe data salah, berikut nama kolom dan pembenarannya:

- FFP_DATE : date
- FIRST_FLIGHT_DATE : date
- LAST_FLIGHT_DATE : date
- LOAD_TIME : date, namun LOAD_TIME akan di drop karena hanya memiliki 1 unique value yang berarti tidak memiliki informasi untuk clustering

Berada di range yang tepat

Beberapa data memiliki max value yang sangat jauh dengan data yang lainnya, akan dilakukan outlier removal untuk case ini

Kesalahan pada penginputan data

ada beberapa value LAST_FLIGHT_DATE yang memiliki tanggal yang aneh, seperti 2014/2/29 karena tanggal tersebut tidak ada. maka akan kita buat menjadi missing value dan akan diisi dengan median

Sebaran data dan preprocess

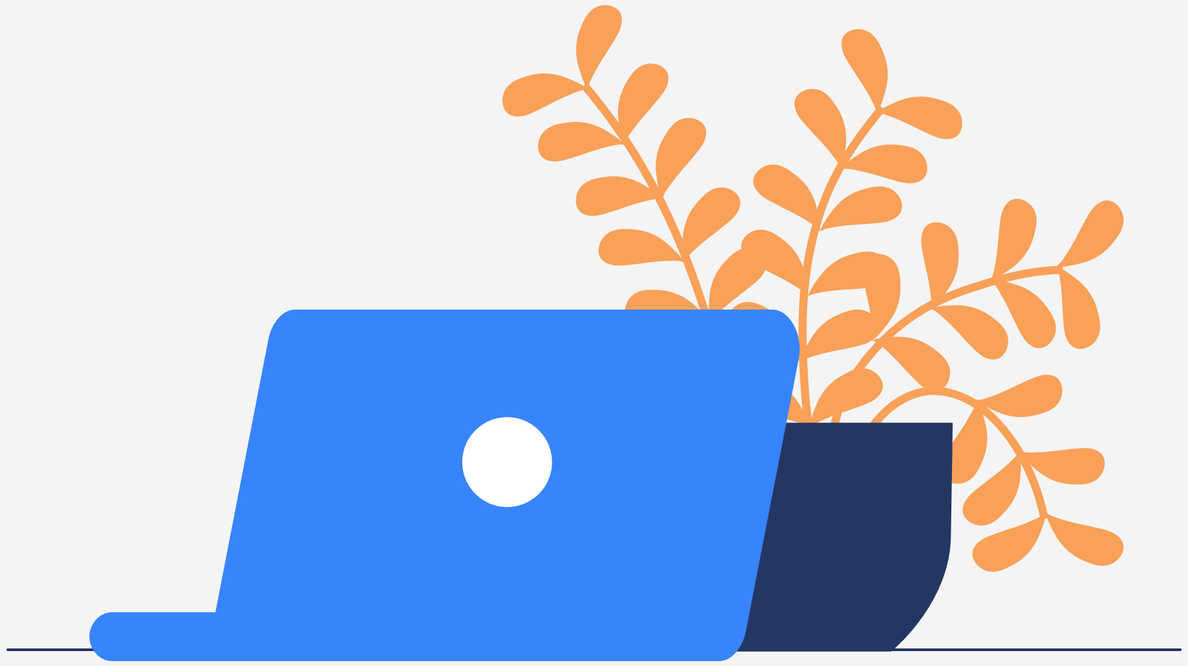
- MEMBER_NO : akan di drop karena memiliki unique value yang sangat banyak (uniform)
- AGE, FLIGHT_COUNT, BP_SUM, SUM_YR_1, SUM_YR_2, SEG_KM_SUM, LAST_TO_END, AVG_INTERVAL, MAX_INTERVAL, EXCHANGE_COUNT, Points_Sum, Point_NotFlight : memiliki right skewed distribution, akan dilakukan boxcox transformation agar data mendekati distribusi normal

Kolom-kolom yang berkorelasi kuat (correlation > 0.7):

- 'FLIGHT_COUNT' dengan 'BP_SUM', 'SUM_YR_1', 'SUM_YR_2', 'SEG_KM_SUM', 'Points_Sum'
- 'BP_SUM' dengan 'FLIGHT_COUNT', 'SUM_YR_1', 'SUM_YR_2', 'SEG_KM_SUM', 'Points_Sum'
- 'SUM_YR_1' dengan 'FLIGHT_COUNT', 'BP_SUM', 'SUM_YR_2', 'SEG_KM_SUM', 'Points_Sum'
- 'SUM_YR_2' dengan 'FLIGHT_COUNT', 'BP_SUM', 'SUM_YR_1', 'SEG_KM_SUM', 'Points_Sum'
- 'SEG_KM_SUM' dengan 'FLIGHT_COUNT', 'BP_SUM', 'SUM_YR_1', 'SUM_YR_2', 'Points_Sum'
- 'Points_Sum' dengan 'FLIGHT_COUNT', 'BP_SUM', 'SUM_YR_1', 'SUM_YR_2', 'SEG_KM_SUM'
- 'AVG_INTERVAL' dengan 'MAX_INTERVAL'

Kolom-kolom yang berkorelasi kuat ini akan di drop jika kolom yang berkorelasi kuat dengannya akan dipakai, agar tidak menimbulkan multicorelation

Feature



Feature Selection

fitur yang akan digunakan untuk clustering adalah fitur yang termasuk kedalam RFM, yaitu Recency, Frequency, dan Monetary. Fitur-fitur tersebut adalah:

- 'LAST_TO_END' : fitur ini termasuk kedalam Recency karena berisi data tentang kapan jarak waktu penerbangan terakhir ke pesanan penerbangan paling akhir
- 'LAST_FLIGHT_DATE' : fitur ini termasuk kedalam Recency karena berisi data tentang kapan terakhir kali customer terbang, nantinya akan dilakukan feature engineering pada fitur ini untuk mengambil tahun, bulan dan hari.
- 'FLIGHT_COUNT' : fitur ini termasuk kedalam Frequency karena berisi data tentang jumlah penerbangan yang telah dilakukan oleh customer
- 'SUM_YR_1' : fitur ini termasuk kedalam Monetary karena berisi data tentang revenue yang telah diberikan customer tersebut kepada perusahaan.

Preprocessing & Feature engineering

Drop missing value

Karena feature yang dipakai hanya sedikit, maka missing value akan di drop agar data tetap original. selain itu karena missing value hanya sebagian kecil (<1%) dari data yang ada maka drop missing value dapat dilakukan karena data yang tersisa masih banyak.

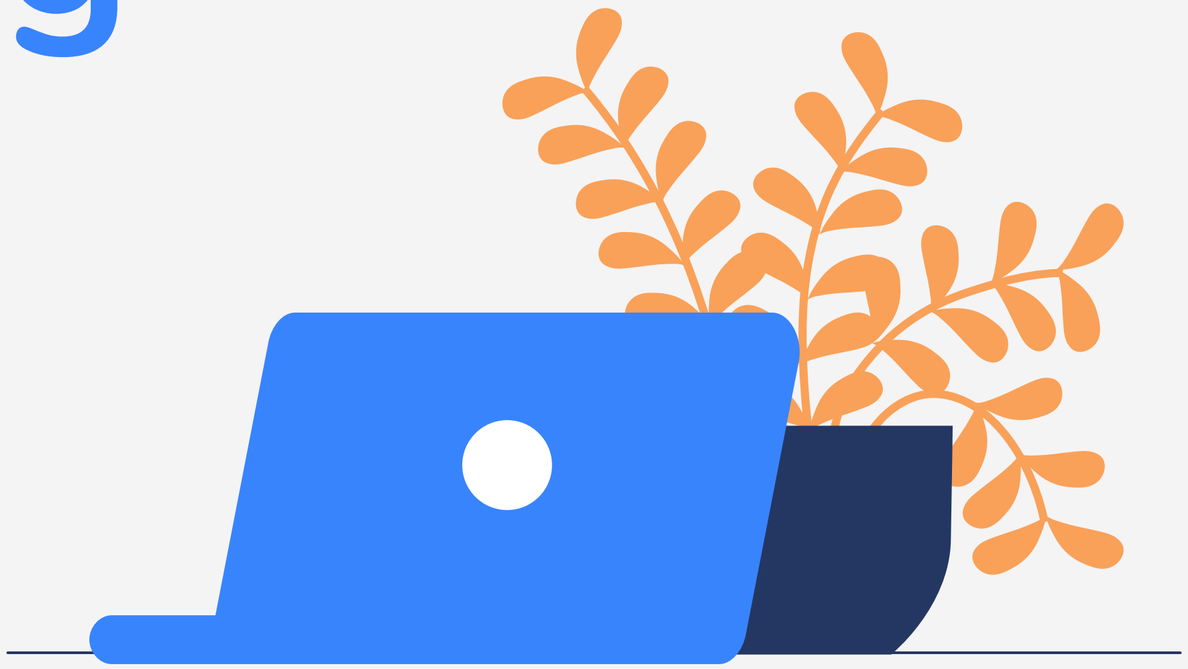
Feature engineering

Memisahkan Tahun, bulan dan tanggal dari kolom LAST_FLIGHT_DATE menjadi kolom-kolom baru

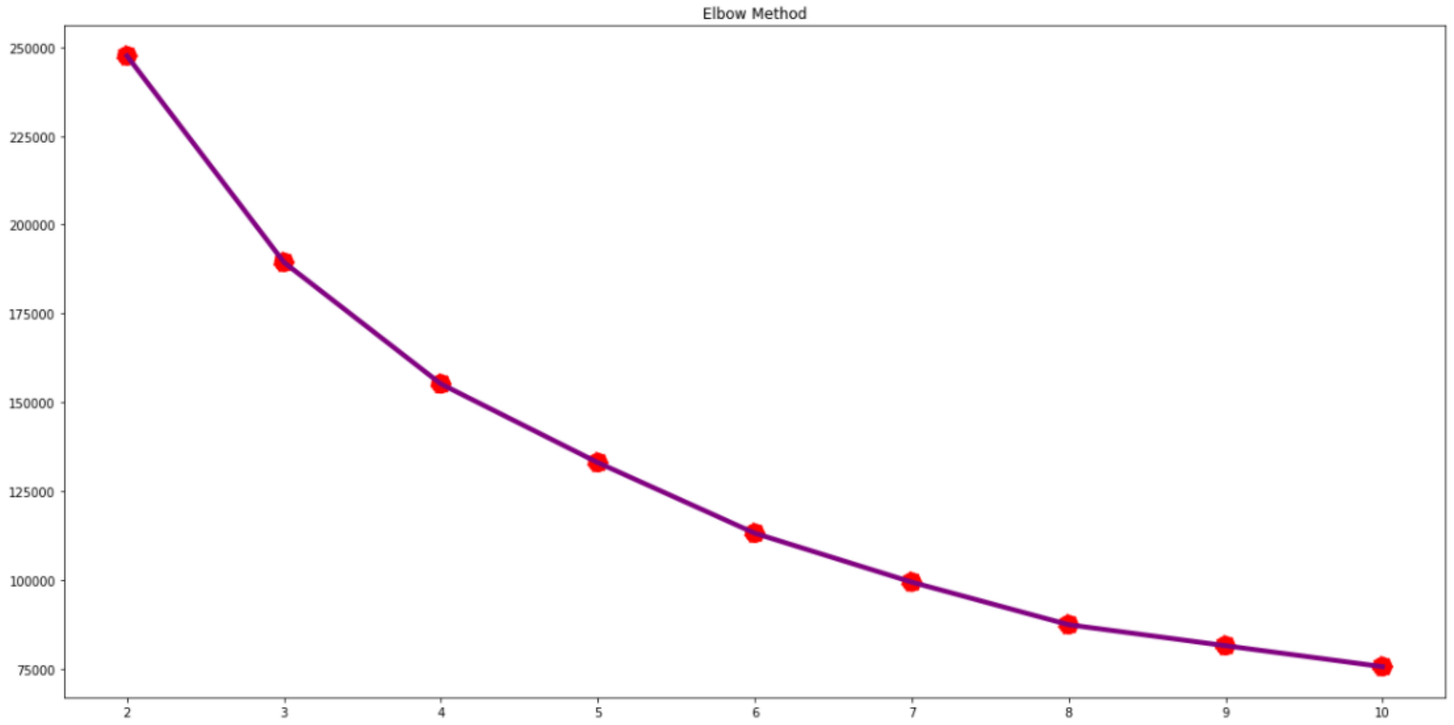
Standardization (Feature scaling)

Apply standardScaler pada data agar scale data sama, sehingga dapat dipakai pada Kmeans clustering

Clustering



Menemukan Jumlah Cluster yang optimal



Penurunan Inertia = 22.288261211597106% Jumlah cluster = 3

Penurunan Inertia = 18.475147905197637% Jumlah cluster = 4

Penurunan Inertia = 13.174234784645936% Jumlah cluster = 5

Penurunan Inertia = 13.785089821103208% Jumlah cluster = 6

Penurunan Inertia = 12.198562561771613% Jumlah cluster = 7

Penurunan Inertia = 13.036733320044162% Jumlah cluster = 8

Penurunan Inertia = 6.855185715545173% Jumlah cluster = 9

Penurunan Inertia = 7.00301645448487% Jumlah cluster = 10

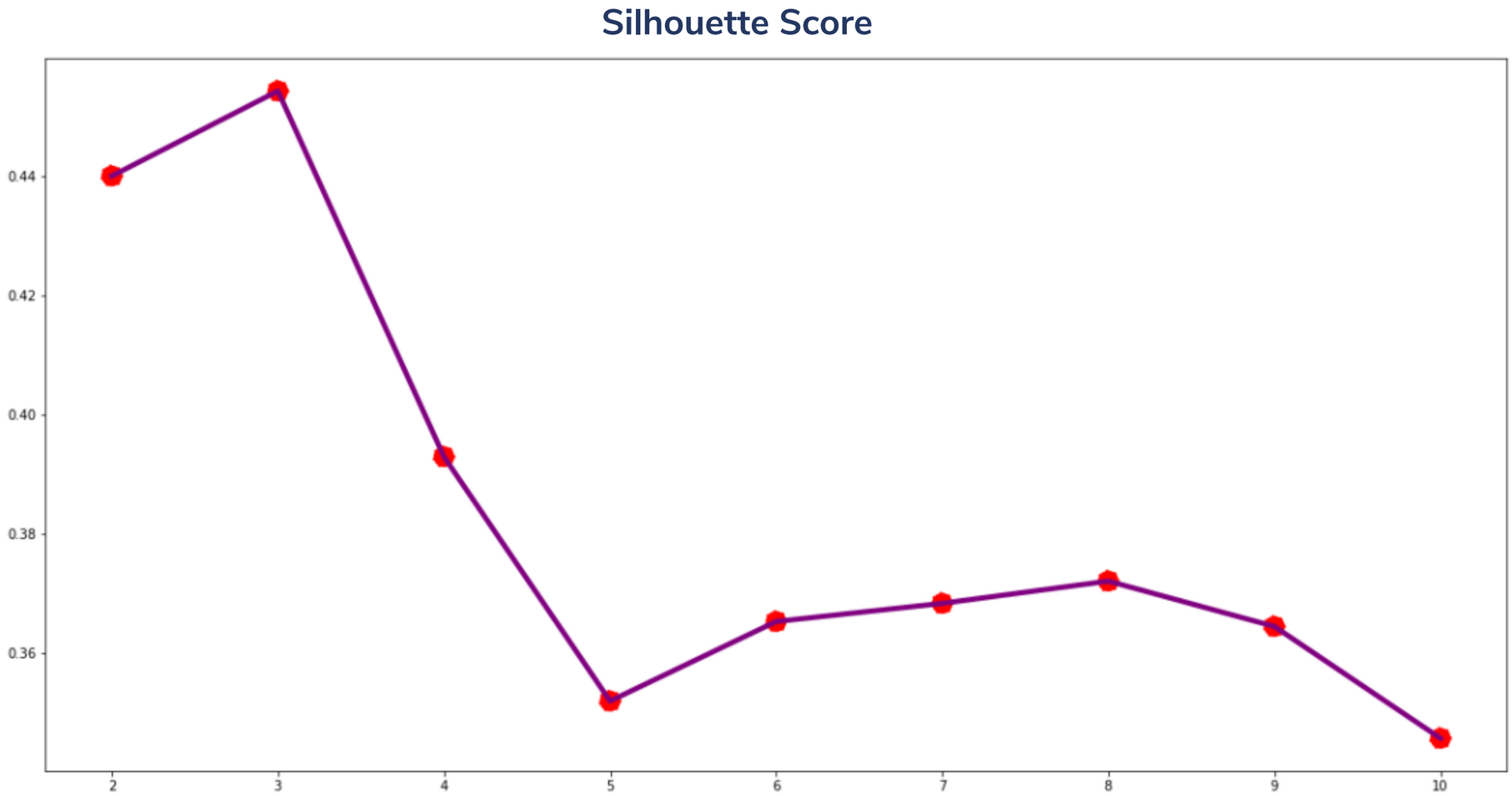
Penurunan Inertia = 6.726869833652965% Jumlah cluster = 11

Penurunan Inertia = 5.953040268029228% Jumlah cluster = 12

Penurunan Inertia = 6.131814052145788% Jumlah cluster = 13

Penurunan Inertia = 4.586248113529456% Jumlah cluster = 14

Menemukan Jumlah Cluster yang optimal



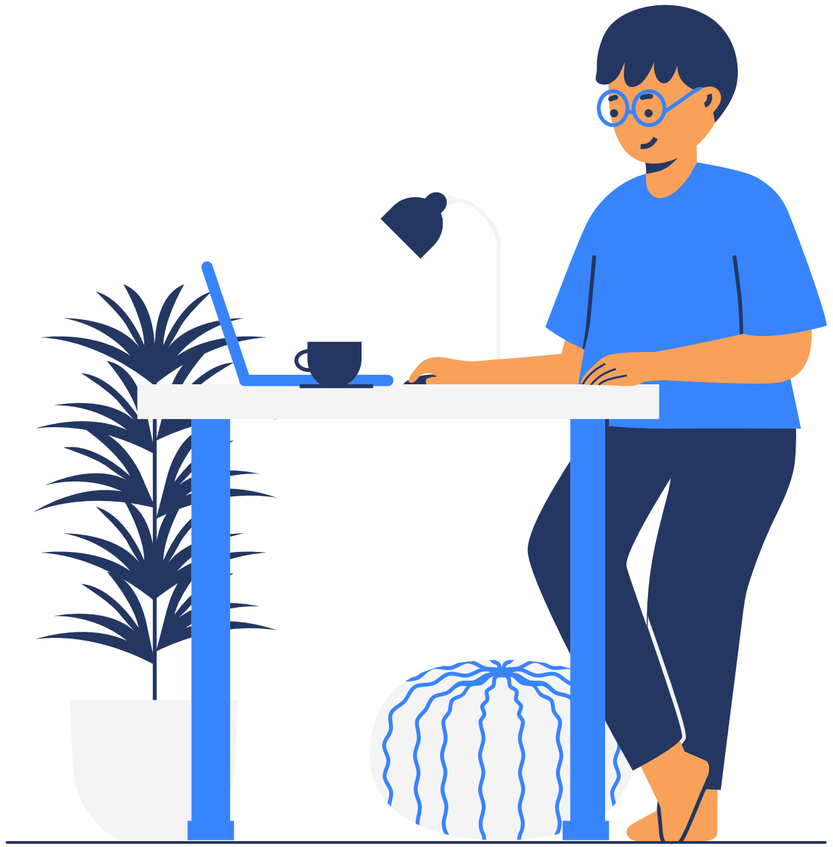
Dari penurunan diatas dapat kitalihat bahwa jumlah cluster 3 adalah yang paling optimal, karena setelah jumlah cluster 3, penurunan inertia sudah tidak terlalu signifikan, serta silhouette score sudah mulai turun.

Clustering menggunakan Kmeans

```
kmeans = KMeans(n_clusters=3, random_state = 42)
kmeans.fit(df_cluster.values)
```

```
KMeans(n_clusters=3, random_state=42)
```

```
df_cluster['label'] = kmeans.labels_
```



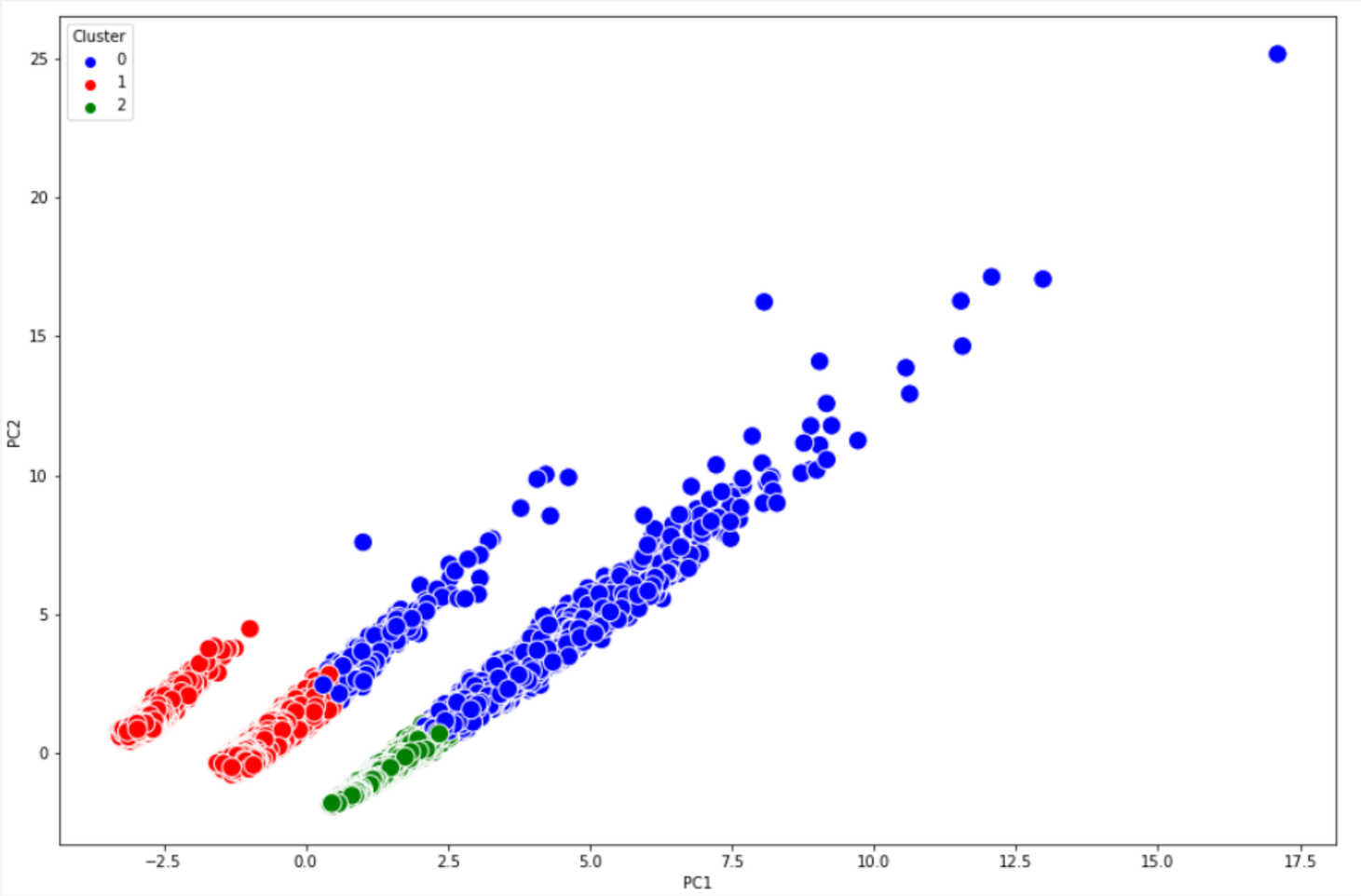
Evaluasi Cluster

PCA

Karena jumlah feature terlalu banyak, maka akan sulit di visualisasikan. Karena itu kita akan menggunakan PCA dan mengurangi jumlah feature menjadi 2

kita hanya mendapatkan 68.94% dari total variance data saat menggunakan 2 components

Visualisasi dengan 2 Components



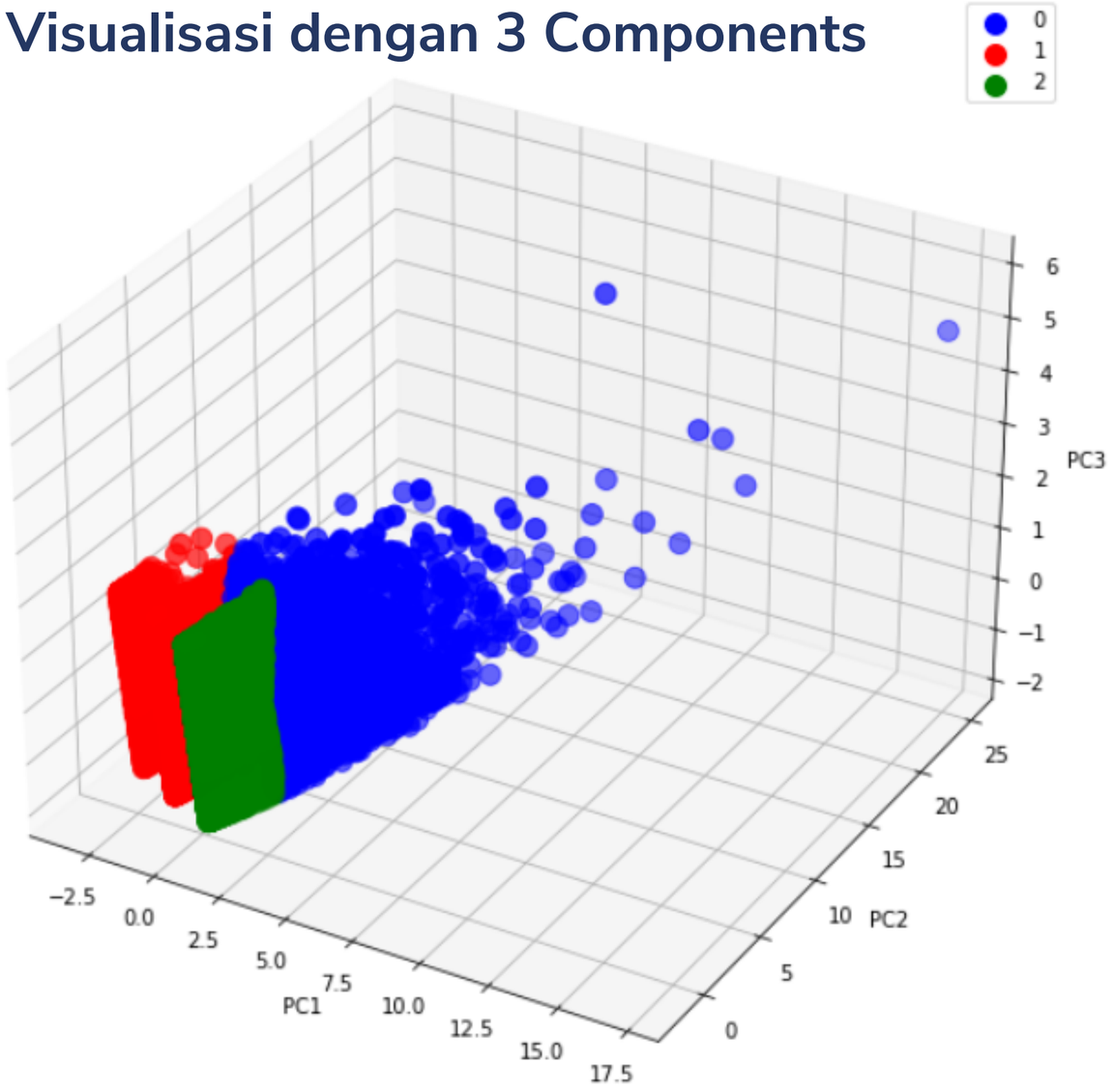
Evaluasi Cluster

PCA

karena terlalu banyak variance yang terbuang pada PCA maka visualisasi menjadi tidak terlalu jelas, karena itu kita akan mencoba untuk memperbanyak components pada PCA

Variance yang dapat kita pertahankan dari data naik menjadi 85.09%, dengan ini informasi yang terbuang pada PCA menjadi berkurang

Visualisasi dengan 3 Components



Interpretation

Customer pada tiap cluster

Dari ciri-ciri dari setiap cluster diatas dapat disimpulkan bahwa:

- Cluster 0: First class customer, mempunyai tier yang tinggi, serta sering memakai jasa penerbangan walaupun jaraknya hanya dekat sekalipun. jenis penerbangan yang rata-rata dipakai oleh customer ini adalah penerbangan yang mahal (fare revenue yang besar dibandingkan jumlah dan jarak penerbangan)
- Cluster 1: Business class customer, mempunyai tier yang rendah, serta jarang memakai jasa penerbangan, rata-rata penerbangan yang dilakukan adalah penerbangan jauh dengan waktu yang cukup lama. Jenis penerbangan yang rata-rata dipakai oleh customer ini adalah penerbangan yang sedang harganya (fare revenue yang lumayan besar dibandingkan jumlah dan jarak penerbangan)
- Cluster 2: Economy class customer, mempunyai tier yang rendah, namun lumayan sering memakai jasa penerbangan, namun rata-rata penerbangan yang dilakukan adalah penerbangan yang murah (fare revenue yang sedikit dibandingkan jumlah dan jarak penerbangan),

Rekomendasi bisnis

Customer yang berada di cluster 1 (business class customer) jarang melakukan penerbangan, jika dilihat avg discount yang dimiliki lebih rendah daripada customer di cluster 0 yang sering melakukan penerbangan, dan terlihat jika customer ini enggan untuk naik pesawat economy yang murah. Selain itu customer di cluster 1 memiliki jumlah paling banyak.

Karena itulah rekomendasi bisnis yang tepat untuk dilakukan adalah membuat seasonal discount untuk penerbangan jarak jauh di kelas bisnis agar customer di cluster 1 dapat lebih sering terbang.

THANKYOU