Homework Solution - Clustering

Team:

1. Erick Rafly Keliat
2. Kukuh Utama Putra
3. Lukman Nulhakim
4. Rahmat Arif Ramadhan
5. Zildjian Rachmat

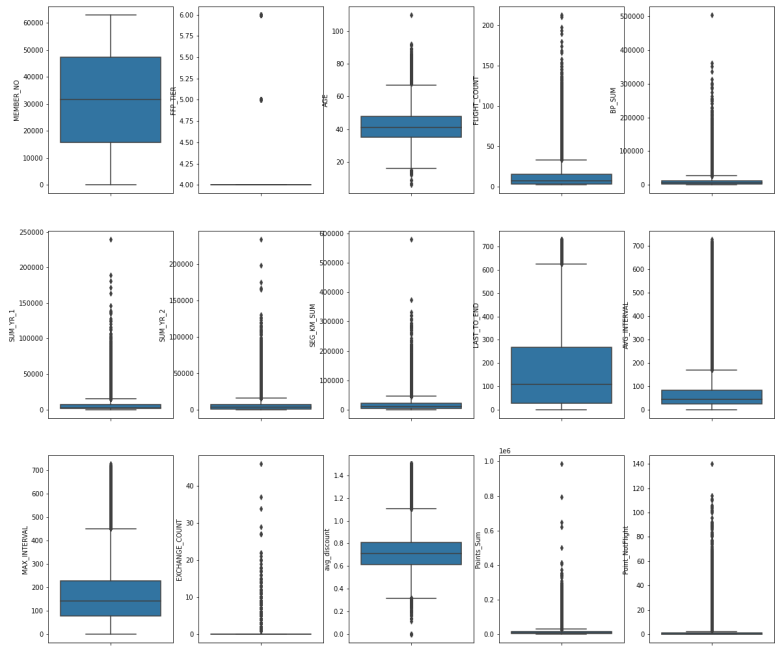
Exploratory Data Analysis

Deskripsi menggunakan fungsi `info()` untuk mengetahui kondisi dari dataset (tipe data, adanya nilai null, hingga jumlah baris)

- Terdapat 62,988 rows

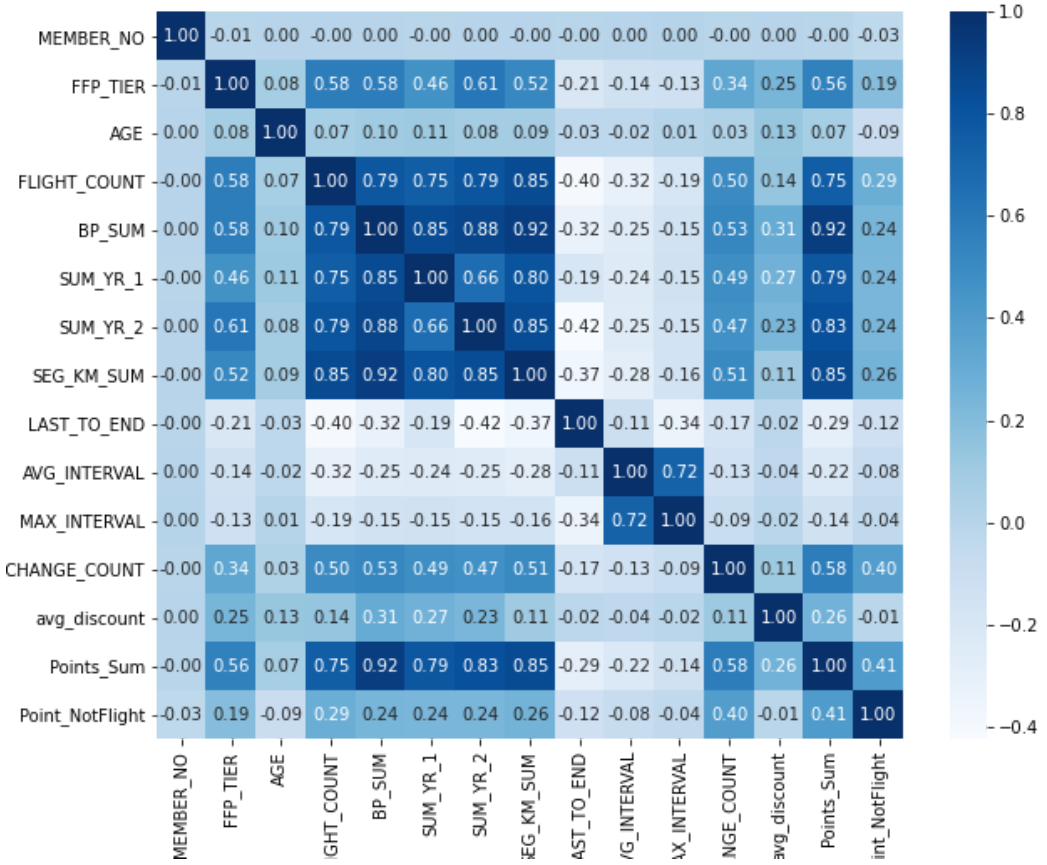
- terdapat 6 kolom yang mempunyai nilai NULL, sehingga tidak diperlukan drop null values

Outlier:



- Pada setiap feature terdapat outlier yang harus dibersihkan

Multivariate Analysis



disini kita dapat melihat data yang memiliki korelasi diatas 0.8 antar feature

- Point_Sum dan BP_Sum

- Point_Sum dan Sum_Yr_2

- Point_Sum dan Sum_Yr_1

- Seg_Km_sum dan Flight_Count

- Seg_Km_sum dan Sum_Yr_1

- Seg_Km_sum dan Sum_Yr_2

- Seg_Km_sum dan Point_sum

- Seg_Km_sum dan BP_sum

- Bp_sum dan Sum_yr2

- Flight_count dan Sum_yr2

- Flight_count dan Sum_yr1

- Flight_count dan BP_Sum

- Flight_count dan Point_SUm

dan juga terdapat feature yang tidak memiliki korelasi dengan feature lain

- Member no

- Age

- last_to_end

- avg_interval

- max interval

- avg_discount

Menghapus Missing Value

Terdapat 7000 rows missing value pada description, jumlahnya masih tidak terlalu besar dengan total data yang ada. jadi dapat langsung dihapus saja

Menghapus Outlier

Agar persebaran data pada cluster tidak banyak yang memencil dan tidak terlalu banyak cluster yang terbentuk. Setelah reduksi, diperoleh 28470 rows.

Feature Engineering

Menambah Feature Baru:

1. Membuat data date
2. Membuat feature lama menjadi member dari data pertama terbang dengan data terakhir terbang (lama_menjadi_member)
3. Melakukan One Hot Encoding pada Gender
4. Menambah feature biaya_per_km dari feature SUM YR dibagi SEG KM SUM

Berdasarkan proses EDA serta fitur engineering, kita akan memilih feature berikut untuk dimasukkan ke dalam modelling :

- FFP_TIER

- FLIGHT_COUNT

- AVG_INTERVAL

- EXCHANGE_COUNT

- avg_discount

- lama_menjadi_member

- biaya_per_km

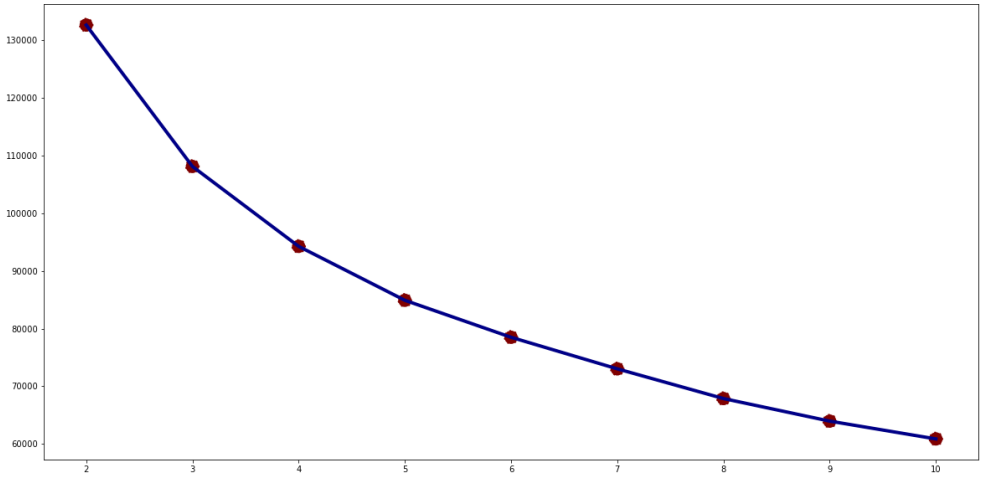
- Points_Sum

- Point_NotFlight

Standarisasi Feature

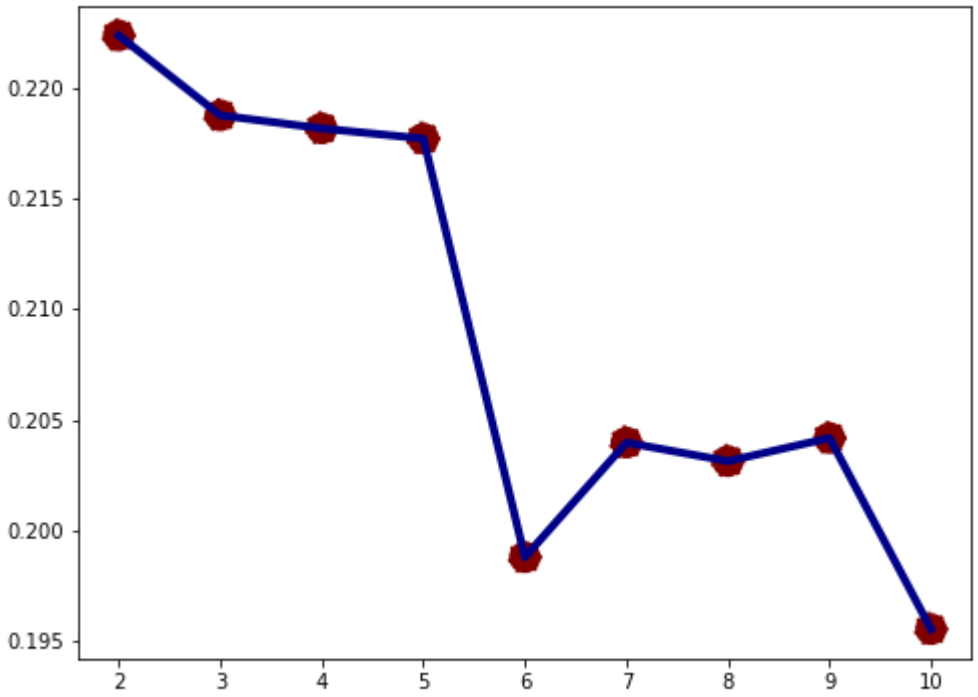
Diperlukan standarisasi data agar proses modelling lebih mudah dilakukan dikarenakan jarak data tidak jauh

Modeling



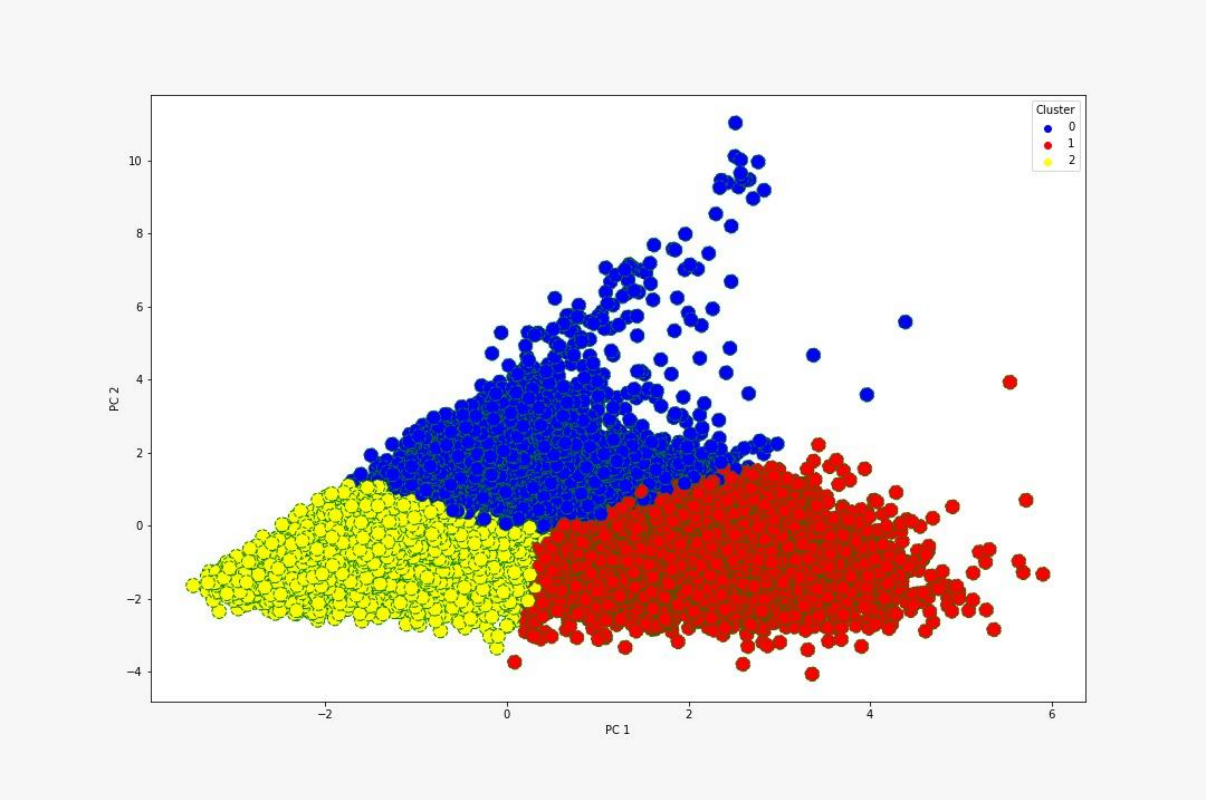
Dari hasil, proses inertia untuk melihat total cluster yang sebaiknya diambil adalah 3 cluster

Untuk memastikan apakah total cluster yang kita ambil sudah baik, kita dapat melihat dengan menggunakan silhouette proses



Berdasarkan hasil silhoutte , 3 cluster juga memiliki nilai yang tinggi. maka bisa memperkuat kesimpulan kita untuk mengambil jumlah cluster sebanyak 3 cluster

Summary (Analysis Hasil Clustering)



Terdapat 3 clustering dari Machine Learning dengan Model KMeans. Insight yang didapat adalah:

cluster 0	customer dengan value tertinggi yang dapat dilihat dari biaya perjalanan per km yaitu dengan rata-rata sebesar 0,83 Jauh perjalanannya adalah 64,48
cluster 1	customer dengan value sedang yang dapat dilihat dari biaya perjalanan per km yaitu dengan rata-rata sebesar 0,61 Jauh perjalanannya adalah 48,98
cluster 2	customer dengan value kecil yang dapat dilihat dari biaya perjalanan per km yaitu dengan rata-rata sebesar 0,45 Jauh perjalanannya adalah 66,77

Appendix

Pada prosesnya semua anggota melakukan proses dari awal sampai akhir, kemudian setiap orang melaporkan pekerjaannya untuk ditarik kesimpulan sebagai bahan buat kerjaan kelompok

Erick Rafly Keliat : EDA, Feture Enginering, Modeling, Summary

Kukuh Utama Putra : EDA, Feture Enginering, Modeling, Summary

Lukman Nulhakim : EDA, Feture Enginering, Modeling, Summary

Rahmat Arif Ramadhan : EDA, Feture Enginering, Modeling, Summary

Zildjian Rachmat : EDA, Feture Enginering, Modeling, Summary