Homework Solution - Clustering

Team:

1. Talenta Nanda
2. Dwirany Dian Audry Sinaga
3. Ichwan
4. Mega Dwi Astuti
5. Talenta Nanda

Submission:

1. Report:

<https://docs.google.com/document/d/1E8_kgR3sTpNN-vzitcAeMuOKACMQaAKKPciNB_plI5s/edit>

1. Notebook:

<https://colab.research.google.com/drive/1t75ktwLfUdSYDVU6fzd_2xf-ThLHcOmR>

Exploratory Data Analysis dan Preprocessing

* Berdasarkan missing values, dilakukan pengisian data pada feature AGE, SUM\_YR\_1, dan SUM\_YR\_2. Selain itu, dilakukan drop pada feature WORK\_CITY, WORK\_PROVINCE, WORK\_COUNTRY dan GENDER
* Dari hasil heatmap dan referensi ebook diketahui bahwa fitur penting berdasarkan sumber yang dibaca, dipilih menggunakan model LRFMC dimana fitur yang digunakan untuk model ini adalah: load\_time, ffp\_date, last\_to\_end, flight\_count, seg\_km\_sum, avg\_discount
* Fitur yang nilai korelasinya rendah dan dianggap tidak berhubungan dalam penyelesaian masalah akan di drop dari dataset: MEMBER\_NO, AGE, GENDER, EXCHANGE\_COUNT, SUM\_YR\_1, SUM\_YR\_2, POINT\_NOTFLIGHT, AVG\_INTERVAL, MAX\_INTERVAL
* Distribusi pelanggan banyak terpusat di umur produktif (27 – 55 tahun). Hal ini bisa menjadi insight bahwa pelanggan melakukan perjalanan bisnis
* Pelanggan Loyal bisa dilihat dari pembagian flight count per tahun menjadi member
* Pelanggan lama bisa dilihat dari tanggal data diambil dikurang FFP join date

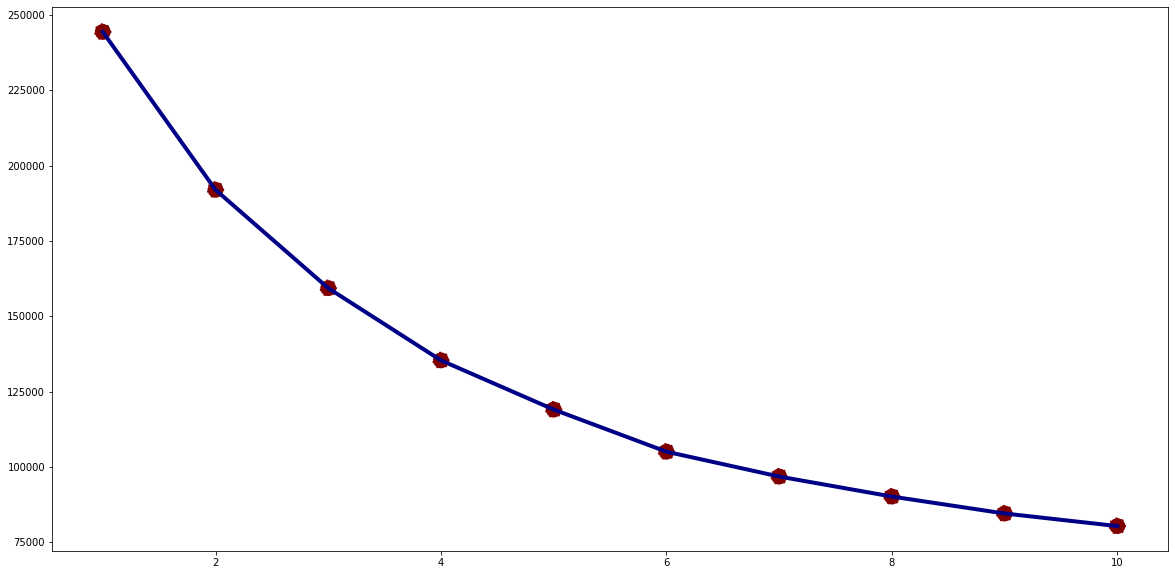
Feature yang tidak ada akan ditambahkan dalam feature engineering

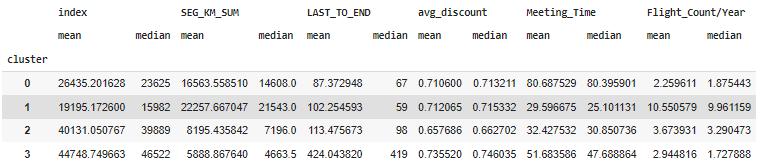
Feature Engineering

* Dari feature LAST\_FLIGHT\_DATE, FIRST\_FLIGHT\_DATE, dan FLIGHT COUNT, dilakukan perhitungan untuk menentukan rata-rata penerbangan per tahun dengan output pada feature Flight\_Count/Year
* Dibuat feature Meeting\_Time untuk menghitung jumlah bulan antara LOAD\_TIME dan FPP\_DATE
* Setelah feature engineering, drop feature yang tidak digunakan kembali, dan handling outlier, dilakukan standarisasi pada dataset sehingga data dapat siap untuk modelling

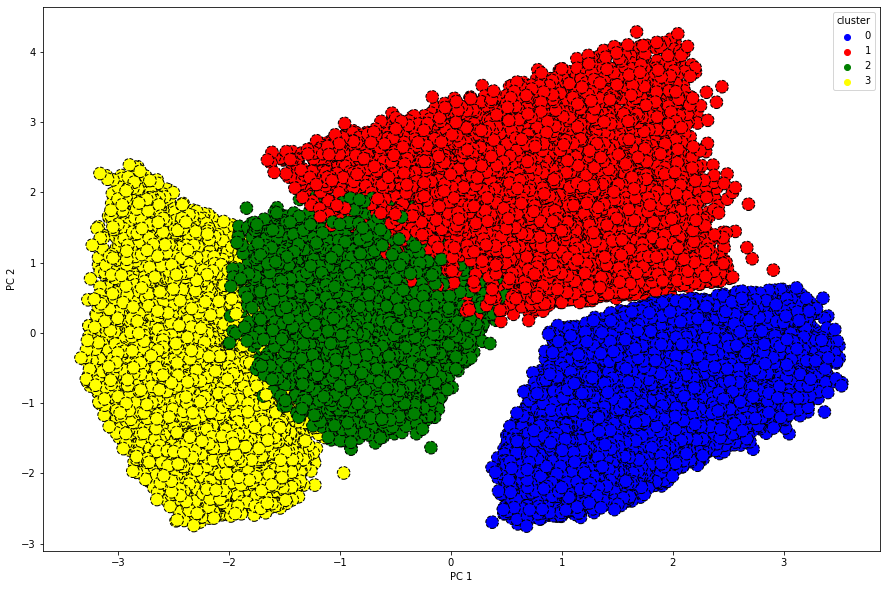
Modeling

Melakukan clustering menggunakan k-means dimana berdasarkan elbow method digunakan jumlah cluster yaitu sebanyak 4 cluster.



Dari hasil clustering dengan k-means diperoleh hasil mean dan median dari setiap fitur di setiap cluster adalah sebagai berikut:  


Lalu dilakukan PCA untuk melihat visualisasi hasil clustering dalam gambar 2D, dengan hasil sbb:

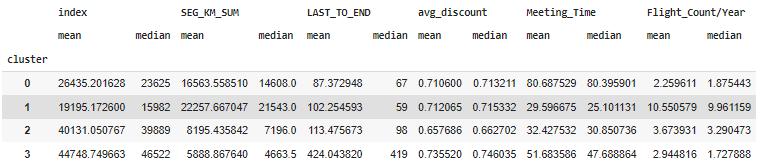


Summary (Analysis Hasil Clustering)

Simpulkan secara singkat hasil pengerjaan teman-teman. Analysis karakteristik setiap cluster dan berikan insight

Berdasarkan hasil pengerjaan, didapatkan beberapa kesimpulan sebagai berikut:

1. Feature yang mempengaruhi dalam segmentasi pelanggan diantaranya yakni total jarak penerbangan, jarak waktu penerbangan terakhir ke pesanan penerbangan paling akhir, rata-rata diskon yang didapatkan pelanggan, durasi pelanggan menjadi member, dan rata-rata jumlah penerbangan per tahun.
2. Berdasarkan cluster yang telah diidentifikasi, dapat diketahui karakteristik masing-masing cluster secara umum adalah sebagai berikut:



Terlihat dalam tabel bahwa rata-rata diskon yang didapatkan tiap pelanggan tidak terlalu berpengaruh pada segmentasi pelanggan, dilihat dari nilai feature yang tidak terlalu jauh antar cluster yang terbentuk.

Adapun penjelasan terhadap ke-4 cluster yang terbentuk adalah sebagai berikut:

1. Cluster 0 merupakan pelanggan yang telah mendaftar sebagai member cukup lama (rata-rata 81 bulan) dengan frekuensi terbang sedang (rata-rata 16.564 km) atau sering melakukan penerbangan jarak jauh mengingat jumlah penerbangan per tahunnya cenderung rendah (rata-rata 2 kali per tahun).
2. Cluster 1 merupakan pelanggan baru (rata-rata 30 bulan) dengan frekuensi terbang tinggi (rata-rata 11 penerbangan per tahun), dengan penerbangan jarak jauh (rata-rata 22.258 km).
3. Cluster 2 merupakan pelanggan dengan durasi menjadi member, frekuensi terbang, dan jarak penerbangan sedang namun memiliki jarak waktu penerbangan terakhir ke pesanan penerbangan paling akhir rendah.
4. Cluster 3 merupakan pelanggan dengan durasi menjadi member, frekuensi terbang, dan jarak penerbangan sedang namun memiliki jarak waktu penerbangan terakhir ke pesanan penerbangan paling akhir tinggi.

Appendix

Tuliskan kontribusi pengerjaan masing-masing anggota tim di bagian ini

Kontribusi pengerjaan homework:

1. Data Pre-processing & EDA: Avinska Valeria Gunawan & Ichwan
2. Feature Engineering: Mega Dwi Astuti
3. Modelling & Evaluation: Talenta Nanda & Dwirany Dian Audry Sinaga
4. Laporan: Semua

Kesulitan dalam pengerjaan tugas ini diantaranya adalah:

Berhubung tugas ini ialah Unsupervised Learning dan target tidak ada dan fitur yang tersedia banyak, maka dalam pengerjaannya sedikit sulit untuk menentukan target dari fitur yang ada dan menentukan fitur penting yang akan menjadi fitur dalam analisis K-Means.