Laporan Tugas Clustering Kartu Kredit (K-Means Clustering)

- Dataset :
  Dataset yang digunakan yaitu dataset kartu kredit yang menyimpan data penggunaan kartu kredit.

	CUST_ID	BALANCE	BALANCE_FREQUENCY	PURCHASES	ONEOFF_PURCHASES	INSTALLMENTS_PURCHASES	CASH_ADVANCE	PURCHASES_FR
0	C10001	40.900749	0.818182	95.40	0.00	95.40	0.000000	
1	C10002	3202.467416	0.909091	0.00	0.00	0.00	6442.945483	
2	C10003	2495.148862	1.000000	773.17	773.17	0.00	0.000000	
3	C10004	1666.670542	0.636364	1499.00	1499.00	0.00	205.788017	
4	C10005	817.714335	1.000000	16.00	16.00	0.00	0.000000	
...	...	...	...	...	...	...	...	...
8945	C19186	28.493517	1.000000	291.12	0.00	291.12	0.000000	
8946	C19187	19.183215	1.000000	300.00	0.00	300.00	0.000000	
8947	C19188	23.398673	0.833333	144.40	0.00	144.40	0.000000	
8948	C19189	13.457564	0.833333	0.00	0.00	0.00	36.558778	
8949	C19190	372.708075	0.666667	1093.25	1093.25	0.00	127.040008	

  Jumlah row data yang terdapat pada dataset yaitu 8950 row. Semua attribut yang terdapat pada dataset akan digunakan selain “CUST_ID” karena tidak memengaruhi data pada penggunaan kartu kredit.

  Selain itu semua data yang digunakan merupakan data numeric sehingga tidak perlu mengubah data categorical menjadi numerical.

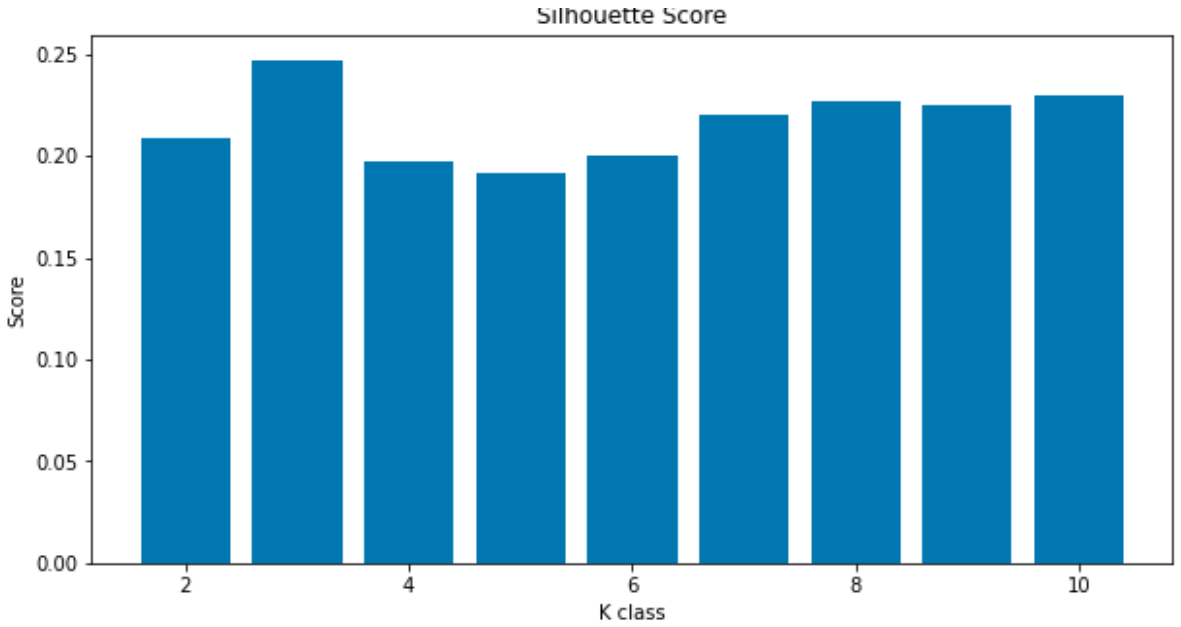
- Preproccess :
  - Missing Value
    Pada dataset yang digunakan terdapat beberapa missing value pada kolom MINIMUM_PAYMENTS sejumlah 313 row dan pada kolom CREDIT_LIMIT sejumlah 1 row.

```
CUST_ID                             0
BALANCE                             0
BALANCE_FREQUENCY                   0
PURCHASES                           0
ONEOFF_PURCHASES                    0
INSTALLMENTS_PURCHASES              0
CASH_ADVANCE                        0
PURCHASES_FREQUENCY                 0
ONEOFF_PURCHASES_FREQUENCY          0
PURCHASES_INSTALLMENTS_FREQUENCY    0
CASH_ADVANCE_FREQUENCY              0
CASH_ADVANCE_TRX                    0
PURCHASES_TRX                       0
CREDIT_LIMIT                        1
PAYMENTS                            0
MINIMUM_PAYMENTS                  313
PRC_FULL_PAYMENT                    0
TENURE                              0
dtype: int64
```

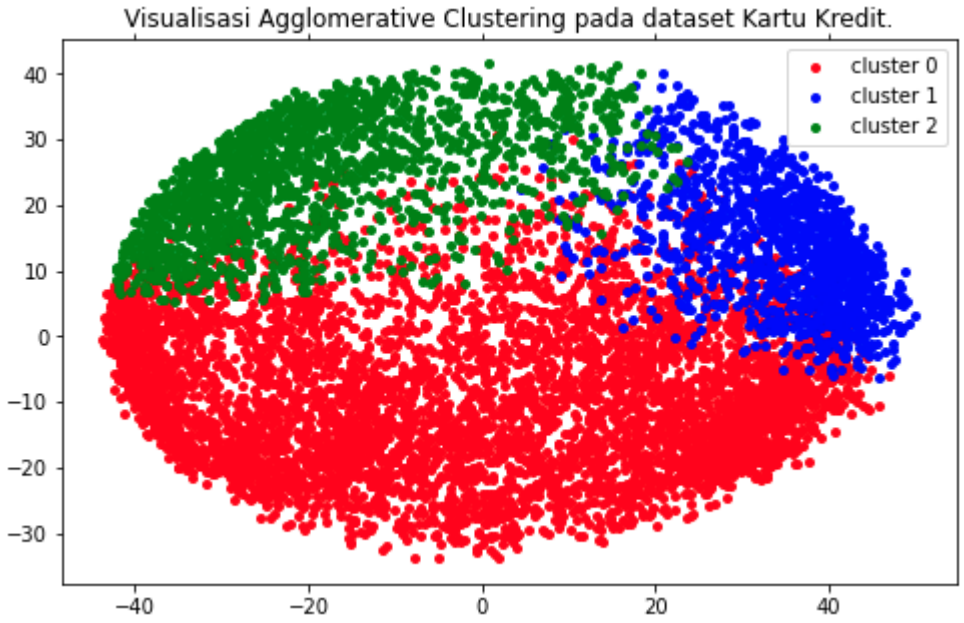
Dikarenakan jumlah data yang ada dan jumlah row missing value sangat berbeda jauh, maka dalam case ini row yang memiliki missing value akan di drop. Sehingga data yang diolah sejumlah 8636 row.

- Scaler
  Karena setiap kolom value-nya berupa numeric tidak ada yang kategorikal, maka langsung membuat scaler supaya range value disesuaikan supaya dapat dicluster. Pada kasus ini digunakan StandardScaler()

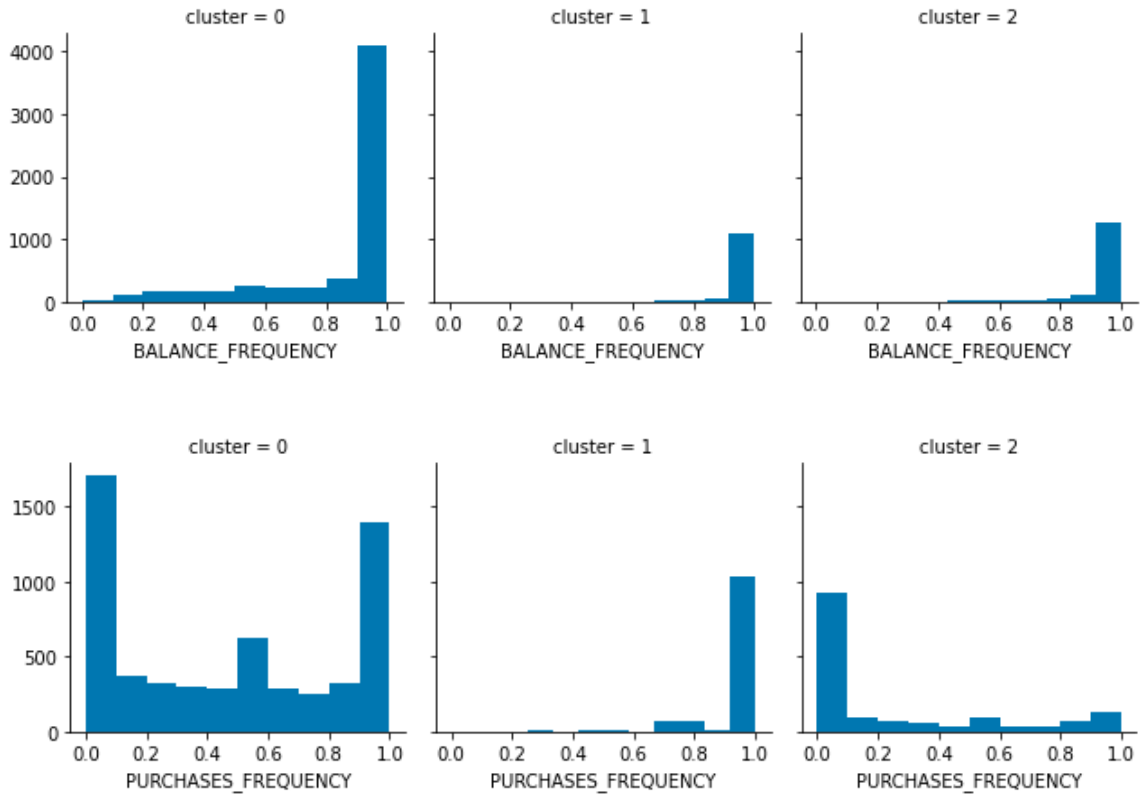
- Silhouette Score
  Silhouette Score dihitung dengan mencoba meng-cluster menggunakan algoritma KMeans, dengan mengiterasikan jumlah "k" mulai dari 2 hingga 10 cluster. Setelah itu menggunakan fungsi dari library sklearn yaitu silhouette_score(). Setelah dihitung, diplot menggunakan bar plot, dan menghasilkan nilai dengan 3 cluster memiliki nilai silhouette paling tinggi dari yang lain.

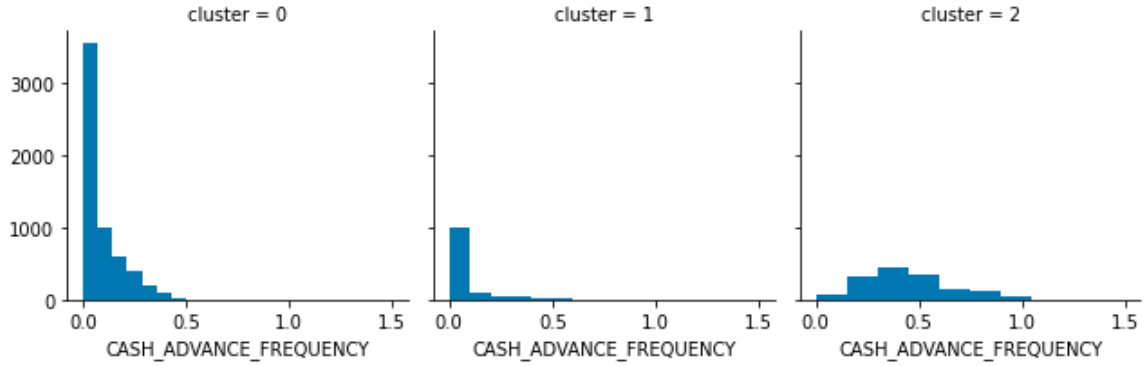


- Clustering :
  Menggunakan algoritma KMeans Clustering, dengan nilai K yang telah dicari yang optimal menggunakan silhouette score, maka menggunakan K = 3.



- Analisis Hasil :
  - Diasumsikan bahwa kelompok yang memerlukan kartu kredit adalah kelompok yang sering memakai kartu kredit untuk bertransaksi. Maka dari kolom kolom yang ada akan dianalisis melalui kolom **BALANCEFREQUENCY, PURCHASESFREQUENCY, CASHADVANCEFREQUENCY**
  - Persebaran data pada setiap cluster dari kolom yang dianalisis





- Ditinjau dari kolom frekuensi, dapat dilihat bahwa cluster 0 lebih sering menggunakan kartu kredit, dari penarikan tunai, top up, dan pembayaran.
- Maka dapat ditarik kesimpulan bahwa **cluster 0** lebih tepat untuk diberikan kartu kredit dan membutuhkan lebih sering dari cluster yang lain