# ZF data scientist intern test report

Chiu Pit Ho, Patrick

In the project, I use python on jupyterLab for the whole process. Also, I use some python libraries, such as:

- *Pandas* - performing data processing
- *Numpy* - performing data processing
- *Datetime* - performing date calculation
- Os - finding system path
- *Matplotlib* - plotting graph
- *Sklearn* - performing K-means clustering

In the part of performing k-means algorithm, the way of determining the number of clusters is by using the **Elbow method on sum of squared estimates of errors (SSE)**.

I perform 4 different visions of it, which are :

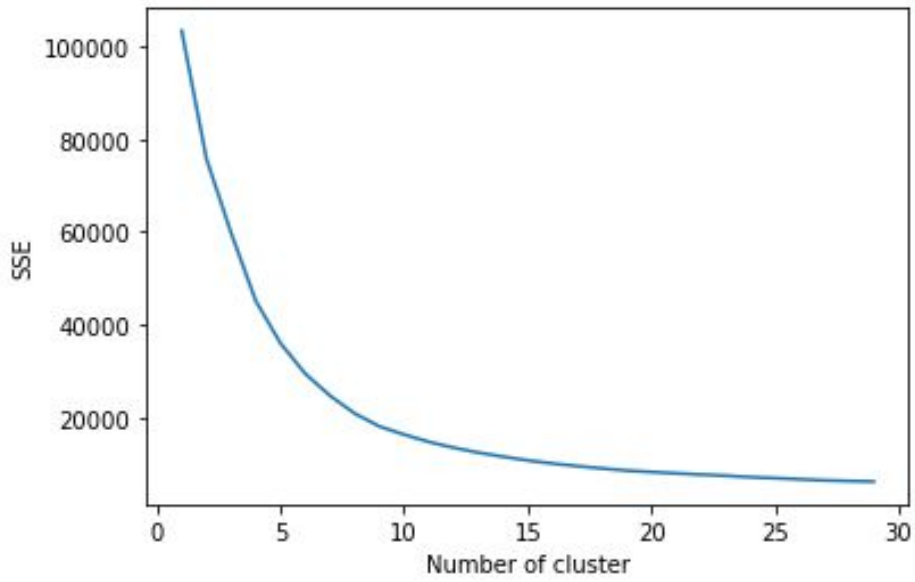
1. K-means
2. **K-means with normalized data**
3. K-means with removed Outliers
4. K-means with removed Outliers and normalized

**I think the second option - K-means with normalized data is most suitable.**
**Because its SSE is the least, also consider the whole data set.**

~~~~~~~~~~~~~~~~~~
~~~~~~~~~~~~~~~~~~

## Result of K-means with normalized data

- # of Cluster vs SSE




Cluster numbers: 8

- Cluster centers:

| | R | F | M |
|---|---|---|---|
| 1 | 1082.0938716 | 4.09484436 | -0.13496922 |
| 2 | 339.28852293 | 8.38319791 | -0.2172272 |
| 3 | 591.71428571 | 10.32432432 | 2.86695697 |
| 4 | 218. | 1540 | -0.31267877 |
| 5 | 341.51001466 | 28.08573522 | 0.02095136 |
| 6 | 657.546875 | 6.49424913 | -0.2517597 |
| 4 | 689. | 5 | 97.79743996 |
| 8 | 745.87845304 | 6.96685083 | 9.40538398 |

- The mean of the RFM data:

| R | F | M |
|---|---|---|
| 567.214669 | 9.545750 | -0.037446 |

Frankly speaking, I didn't know what an RFM model is at the beginning. I thought some customers with weird behavior can be defined as outriders and could be removed. But after I studied some information on the Internet. I think it is not good to remove a single customer, no matter how weird his behavior is.

~~~~~~~~~~~~~~~~~~

## File list:

| | |
|---|---|
| RFM_df.csv | The only RFM data for each customer |
| RFM_Kmeans.csv | Customer code, R, F, M and the segmentation label via using K-Means |
| RFM_Kmeans_nor.csv | Customer code, R, F, M and the segmentation label via using K-means with normalized data |
| RFM_Kmeans_ro.csv | Customer code, R, F, M and the segmentation label via using K-means with removed Outliers |
| RFM_Kmeans_ro_nor.csv | Customer code, R, F, M and the segmentation label via using K-means with removed Outliers and normalized |
| test.ipynb | The ipynb file with my code |
| test.html | The html vision of test.ipynb |
| ZF data scientist intern test by patrick.pdf | The report of this test |
~~~~~~~~~~~~~~~~~~