DAT Client segmentation

https://towardsdatascience.com/customer-segmentation-in-python-9c15acf6f945

La segmentation des clients est utile pour comprendre les sous-populations démographiques et Psychographie de nos clients dans une analyse de rentabilisation.

Segment customers based on their buying behavior on the market.

Les étapes à suivre

- **Gather the data**
- **Create Recency Frequency Monetary (RFM) table**
- **Manage skewness and scale each variable**
- **Explore the data**
- **Cluster the data**
- **Interpret the result**

1 - Data Gathering

The dataset given by the company is a transactional data that contains transactions,

Each row represents the transaction that occurs. It includes the product name, price, the client ID and the month

Here is the size of the dataset.

```
(7245522, 8)
```

2 - Create The RFM Table

Once we have the data, we will make the data easier to conduct an analysis.

To segmenting customer, there are some metrics that we can use, such as **when the customer buys the product** for last time, **how frequent the customer buy the product**, and h**ow much the customer pays for the product**. We will call this segmentation as RFM segmentation.

To make the RFM table, we can create these columns, such as Recency, Frequency, and MonetaryValue column.

We already have for each client the month he perched the item Recency columns.

To create the frequency column, we can count how much transactions by each customer for each product

Lastly, to create the monetary value column, we can sum all transactions for each customer.

Out[7]:

CLI_ID	LIBELLE	Recency	Frequency	Monetary
1490281	CR JR PARF BIO.SPE AC.SENT.50ML	10	1	7.45
	EAU MICELLAIRE 3 THES FL200ML	10	1	5.95
	GD JDM4 PAMPLEMOUSSE FL 200ML	10	2	6.66
	GD JDM4 TIARE FL 200ML	10	1	1.67
13290776	EDT UN MATIN AU JARDIN 100ML MUGUET	9	1	13.00
...	...	...	...	...
997385337	VAO BRIL ROSE SOMPTUEUX 14 CN3 5.5ML	6	1	4.45
	VAO BRIL ROUG/SIENN 33 AX/SO CN3 5,5ML	9	1	4.45
	VAO HIBISCUS ROUGE ETE13 ANI LU4 3ML	5	1	3.90
	VAO PASTEL PARME 03 MANUC CN3 5.5ML	9	1	8.90

Out[7]:

CLI_ID	LIBELLE	Recency	Frequency	Monetary
1490281	CR JR PARF BIO.SPE AC.SENT.50ML	10	1	7.45
	EAU MICELLAIRE 3 THES FL200ML	10	1	5.95
	GD JDM4 PAMPLEMOUSSE FL 200ML	10	2	6.66
	GD JDM4 TIARE FL 200ML	10	1	1.67
13290776	EDT UN MATIN AU JARDIN 100ML MUGUET	9	1	13.00
...	...	...	...	...
997385337	VAO BRIL ROSE SOMPTUEUX 14 CN3 5.5ML	6	1	4.45
	VAO BRIL ROUG/SIENN 33 AX/SO CN3 5,5ML	9	1	4.45
	VAO HIBISCUS ROUGE ETE13 ANI LU4 3ML	5	1	3.90
	VAO PASTEL PARME 03 MANUC CN3 5.5ML	9	1	8.90

Out[175]:

CLI_ID	LIBELLE	Recency	Frequency	MonetaryValue
1490281	CR JR PARF BIO.SPE AC.SENT.50ML	10	1	7.45
	EAU MICELLAIRE 3 THES FL200ML	10	1	5.95
	GD JDM4 PAMPLEMOUSSE FL 200ML	10	2	6.66
	GD JDM4 TIARE FL 200ML	10	1	1.67
13290776	EDT UN MATIN AU JARDIN 100ML MUGUET	9	1	13.00
	EDT UN MATIN AU JARDIN 100ML LILAS	12	2	43.18
	GD LILAS FP FL200ML	12	3	17.28
	LAIT LILAS FP FL200ML	12	2	19.30
	LAIT VELOUTE COCO PN2 400ML	9	1	5.50

Right now, the dataset consists of **recency**, **frequency**, and **monetary** value column. But we cannot use the dataset yet because we have to preprocess the data more.

Manage Skewness and Scaling

Objectif: we have to make sure that the data meet these assumptions, we have to manage the skewness of the variables.

There are some methods that we can use to manage the skewness, they are,

- **log transformation**
- **square root transformation**
- **box-cox transformation**

We choose to usr box-cox transformation method for our case.

when calculate the skewness we get:

Frequency	Recency	Monetary
0,3449801005	0,6604840139	-0,8367210143

By using the box-cow transformation, we will have data that less skewed, we can transform the RFM table,

	Recency	Frequency	Monetary
5911144	4.387815	0.0	1.644826
5911145	6.769914	0.0	1.644826
5911146	3.565543	0.0	1.574061
5911147	6.769914	0.0	2.072351
5911148	6.769914	0.0	2.150838

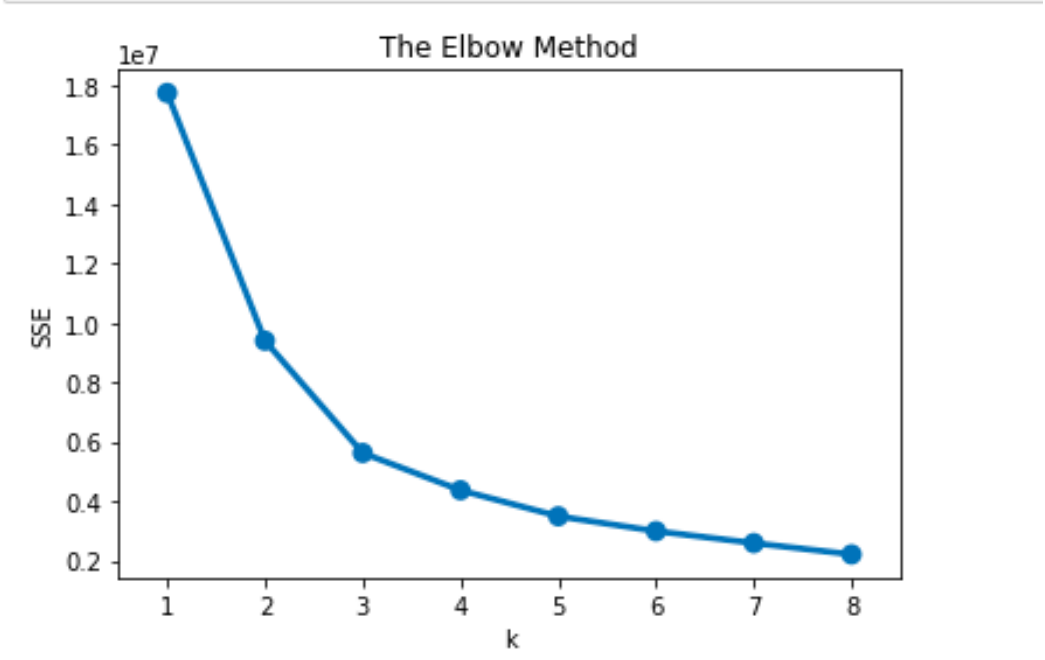
Still we can't use data. If we look at the plot once more, each variable don’t have the same **mean** and **variance**. We have to normalize it. To normalize, we can use **StandardScaler** object from scikit-learn library to do it. The code will look like this,

	0	1	3
0	-0.2602543	-0.88305407	-0.37502666
1	-0.2602543	-0.88305407	-0.59662906
2	-0.2602543	0.9661512	-0.48759124
3	-0.2602543	-0.88305407	-1.57980502
4	-1.38982604	-0.88305407	0.25083063
5	1.27355643	0.9661512	2.06894371
6	1.27355643	1.51681677	0.61853584
7	1.27355643	0.9661512	0.77104925
8	-1.38982604	-0.88305407	-0.67030744

Modelling

Right after we preprocess the data, now we can focus on modelling. To make segmentation from the data, we can use the K-Means algorithm to do this.

First determine which hyperparameter fits to the data:



How to interpret the plot? The x-axis is the value of the k, and the y-axis is the SSE value of the data. We will take the best parameter by looking at where the k-value will have a linear trend on the next consecutive k.

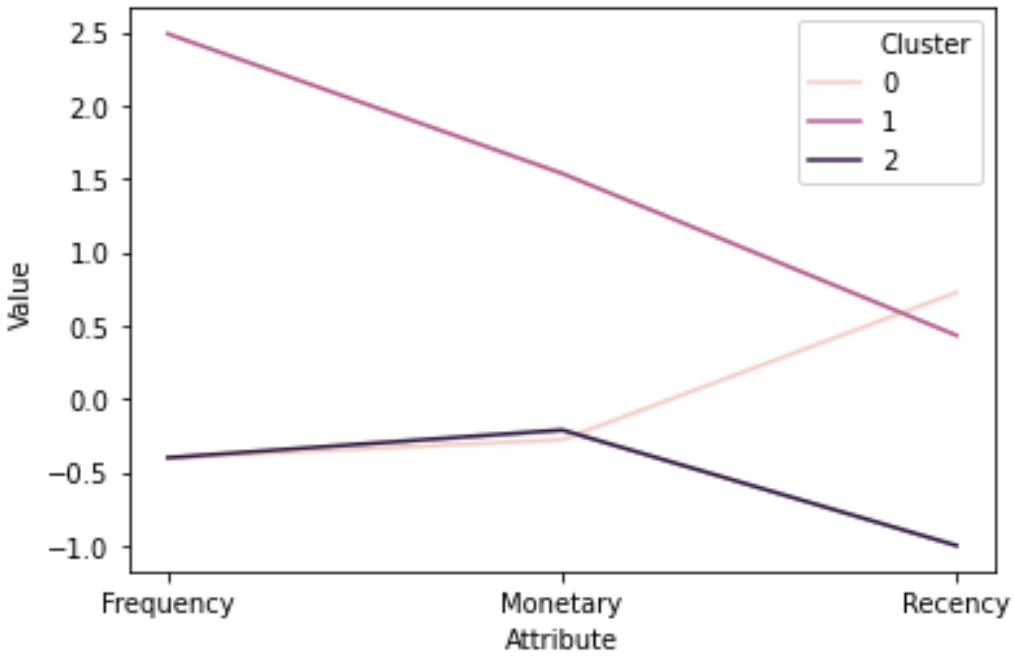
Based on our observation, the k-value of 3 is the best hyperparameter for our model because the next k-value tend to have a linear trend. Therefore, our best model for the data is **K-Means with the number of clusters is 3**.

Interpret The Segment

We can summarize the RFM table based on clusters and calculate the mean of each variable.

	Recency	Frequency	Monetary	
	mean	mean	mean	count
Cluster				
0	9.71	1.00	5.66	2742343
1	8.66	2.62	54.01	821345
2	3.54	1.00	6.19	2347461

Besides that, we can analyze the segments using snake plot. It requires the normalized dataset and also the cluster labels. By using this plot, we can have a good visualization from the data on how the cluster differs from each other. We can make the plot by using this code,



By using this plot, we know how each segment differs. It describes more than we use the summarized table.

We infer that cluster 0 is frequent, spend more, and they buy the product recently. Therefore, it could be the cluster of a **loyal customer**.

Then, the cluster 1 is less frequent, less to spend, but they buy the product recently. Therefore, it could be the cluster of **new customer**.

Finally, the cluster 2 is less frequent, less to spend, and they buy the product at the old time. Therefore, it could be the cluster of **churned customers**.