ECODATION INTERNSHIP DOCUMENTATION

PROJECT: TO CLASSIFY CREDIT CARD CUSTOMER USING WITH **K-MEANS ALGORITHM**

STUDENT NAME / SURNAME: Emirhan ASLANKARAYİĞİT

Output 1: Analyse the dataset using with the Jupyter Notebook

This dataset whose format is csv file is consisted of **8950 rows** and **18 columns**:

That means it has 8950 Customer ID and 18 features belong to customer:

- What are our dataset's features namely?
  - **CUST_ID**
  - **BALANCE**
  - **BALANCE_FREQUENCY**
  - **PURCHASES**
  - **ONEOFF_PURCHASES**
  - **INSTALLMENTS_PURCHASES**
  - **CASH_ADVANCE**
  - **PURCHASES_FREQUENCY**
  - **ONEOFF_PURCHASES_FREQUENCY**
  - **PURCHASES_INSTALLMENTS_FREQUENCY**
  - **CASH_ADVANCE_FREQUENCY**
  - **CASH_ADVANCE_TRX**
  - **PURCHASES_TRX**
  - **CREDIT_LIMIT**
  - **PAYMENTS**
  - **MINIMUM_PAYMENTS**
  - **PRC_FULL_PAYMENT**
  - **TENURE**

Output1:

Out[2]:

	CUST_ID	BALANCE	BALANCE_FREQUENCY	PURCHASES	ONEOFF_PURCHASES	INSTALLMENTS_PURCHASES	CASH_ADVANCE	PURCHASES_FRE
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	

Output 2: Correlation is done to understand the relationship between the features that located on the frame. (Using ***corr()***)

Output2:

```
In [20]: data.corr()
```

Out[20]:

	BALANCE	BALANCE_FREQUENCY	PURCHASES	ONEOFF_PURCHASES	INSTALLMENTS_PURCHASES	CASH_ADVAI
BALANCE	1.000000	0.322412	0.181261	0.164350	0.126469	0.49
BALANCE_FREQUENCY	0.322412	1.000000	0.133674	0.104323	0.124292	0.09
PURCHASES	0.181261	0.133674	1.000000	0.916845	0.679896	-0.05
ONEOFF_PURCHASES	0.164350	0.104323	0.916845	1.000000	0.330622	-0.03
INSTALLMENTS_PURCHASES	0.126469	0.124292	0.679896	0.330622	1.000000	-0.06
CASH_ADVANCE	0.496692	0.099388	-0.051474	-0.031326	-0.064244	1.00
PURCHASES_FREQUENCY	-0.077944	0.229715	0.393017	0.264937	0.442418	-0.21
ONEOFF_PURCHASES_FREQUENCY	0.073166	0.202415	0.498430	0.524891	0.214042	-0.08
PURCHASES_INSTALLMENTS_FREQUENCY	-0.063186	0.176079	0.315567	0.127729	0.511351	-0.17
CASH_ADVANCE_FREQUENCY	0.449218	0.191873	-0.120143	-0.082628	-0.132318	0.62
CASH_ADVANCE_TRX	0.385152	0.141555	-0.067175	-0.046212	-0.073999	0.65
PURCHASES_TRX	0.154338	0.189626	0.689561	0.545523	0.628108	-0.07
CREDIT_LIMIT	0.531283	0.095843	0.356963	0.319724	0.256499	0.30
PAYMENTS	0.322802	0.065008	0.603264	0.567292	0.384084	0.45
MINIMUM_PAYMENTS	0.398684	0.132569	0.093860	0.048755	0.132172	0.14
PRC_FULL_PAYMENT	-0.318959	-0.095082	0.180379	0.132763	0.182569	-0.15
TENURE	0.072692	0.119776	0.086288	0.064150	0.086143	-0.06

We have detected **by taking values greater than 0.7** as a result of **corr() method**, because the most suitable data to process are found by this way. It is called **"very high correlation". By this way, we obtained Boolean data like True, False to realise.**

Out[21]:

	BALANCE	BALANCE_FREQUENCY	PURCHASES	ONEOFF_PURCHASES	INSTALLMENTS_PURCHASES	CASH_ADVAI
BALANCE	True	False	False	False	False	F
BALANCE_FREQUENCY	False	True	False	False	False	F
PURCHASES	False	False	True	True	False	F
ONEOFF_PURCHASES	False	False	True	True	False	F
INSTALLMENTS_PURCHASES	False	False	False	False	True	F
CASH_ADVANCE	False	False	False	False	False	
PURCHASES_FREQUENCY	False	False	False	False	False	F
ONEOFF_PURCHASES_FREQUENCY	False	False	False	False	False	F
PURCHASES_INSTALLMENTS_FREQUENCY	False	False	False	False	False	F
CASH_ADVANCE_FREQUENCY	False	False	False	False	False	F
CASH_ADVANCE_TRX	False	False	False	False	False	F
PURCHASES_TRX	False	False	False	False	False	F
CREDIT_LIMIT	False	False	False	False	False	F
PAYMENTS	False	False	False	False	False	F
MINIMUM_PAYMENTS	False	False	False	False	False	F
PRC_FULL_PAYMENT	False	False	False	False	False	F
TENURE	False	False	False	False	False	F

As a result of this operation(correlation):

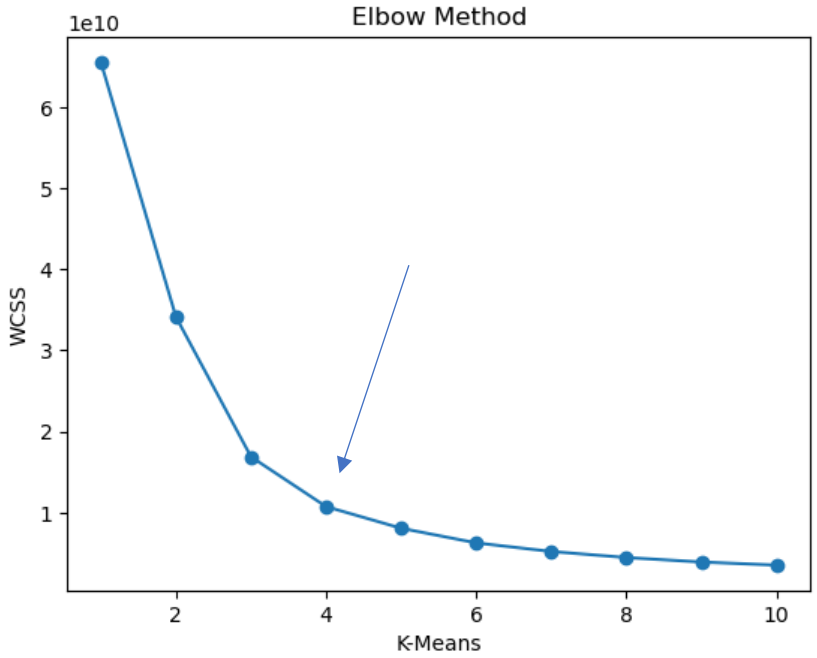
- **PURCHASES / ONEOFF_PURCHASES**
- **CASH_ADVANCE_TRX / CASH_ADVANCE_FREQUENCY**
- **PURCHASES_INSTALLMENTS_FREQUENCY / PURCHASES_FREQUENCY**

Data given above have determined that there is a relation between these features, because their correlation is categorized on this group: "Very High Correlation". Corr() method is easier way to classify the customers.

Output 3: Determination of the number of clusters with the elbow method

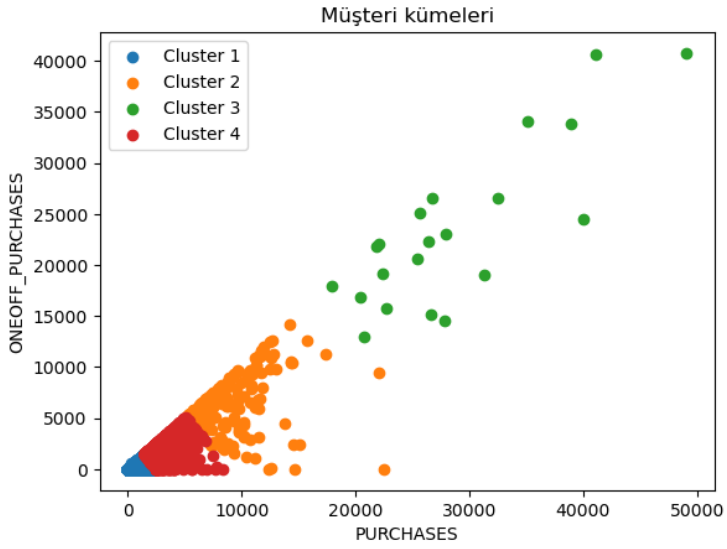
If we take the two features belong to dataset,

1) **PURCHASES**
   ONEOFF_PURCHASES
   The elbow method's graph is like that:



The most suitable number of cluster(s) is "4" according to Elbow method.

OUTPUT 4 for this part: To classify the customer using with the K-Means algorithm



According to this scatter graph, we can observe that there are 4 groups of customer segment.

Customers who belong to "Cluster 2" and "Cluster 3" are less than other clusters in terms of numerical.

For “Cluster 2”, minimum purchase 5176.62 and maximum purchase 22500.0.

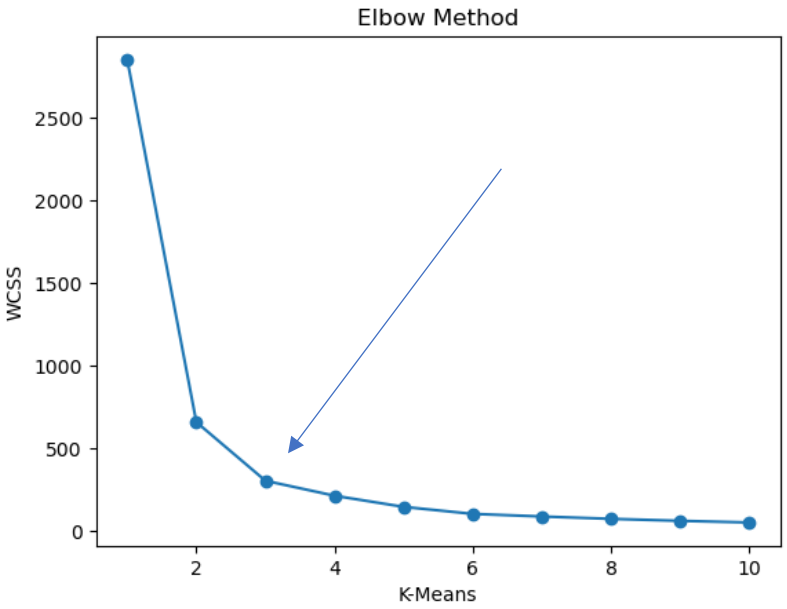
For “Cluster 3”, minimum purchase 17945.0 and maximum purchase 49039.57.

For “Cluster 2”, minimum one off-purchases 0.0 and maximum one off-purchases 14215.0.

For “Cluster 3”, minimum one off-purchases 13007.07 and maximum one off-purchases 40761.25.

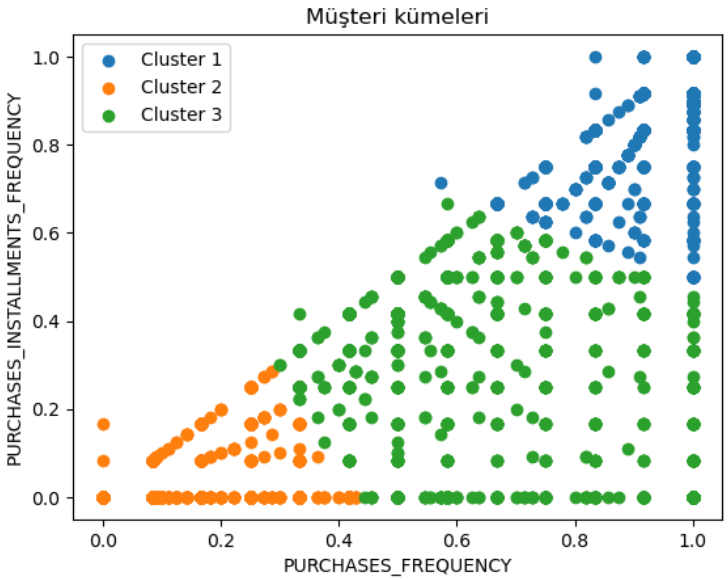
2) **PURCHASES_INSTALLMENTS_FREQUENCY / PURCHASES_FREQUENCY**

The elbow method’s graph is like that:



The most suitable number of cluster(s) is “3” according to Elbow method.

OUTPUT 4 for this part: To classify the customer using with the K-Means algorithm



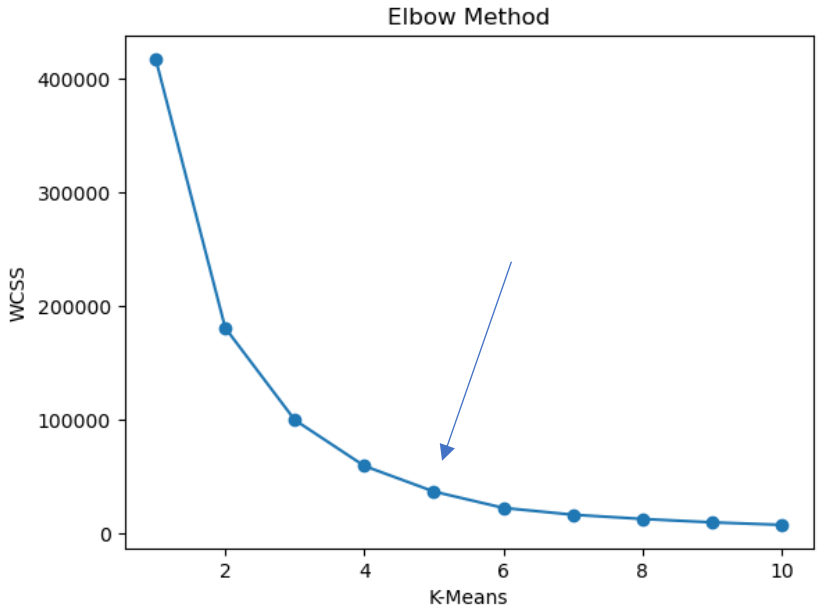
According to this scatter graph, we can observe that there are 3 groups of customer segment.

Customers who belong to "Cluster 2" are more than other clusters in terms of numerical.

For "Cluster 2", "**PURCHASES_FREQUENCY**" is changed by this range: (0.0, 0.428571)

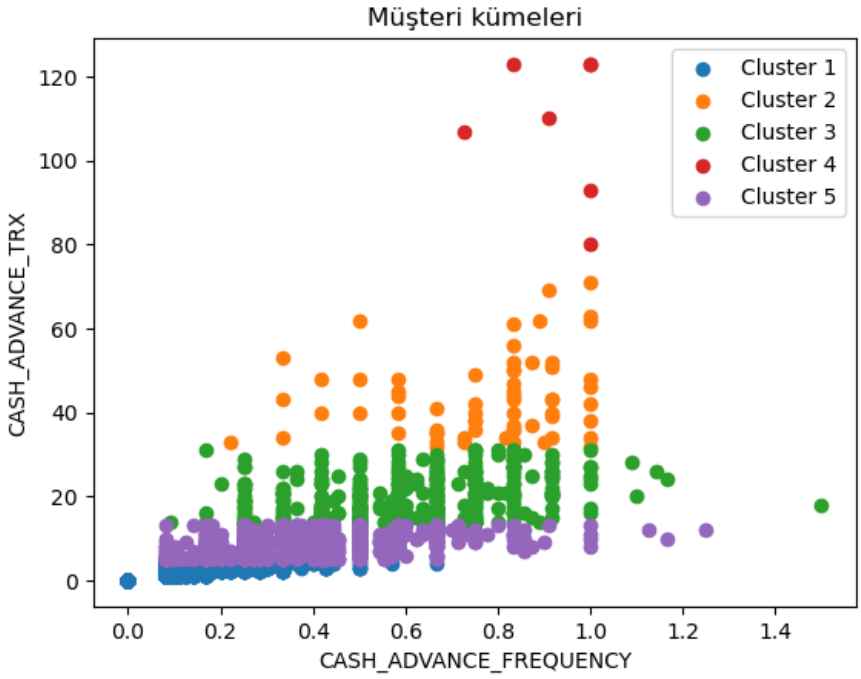
3) **CASH_ADVANCE_TRX / CASH_ADVANCE_FREQUENCY**

The elbow method's graph in this categorization is like that:



The most suitable number of clusters is "5" according to Elbow Method.

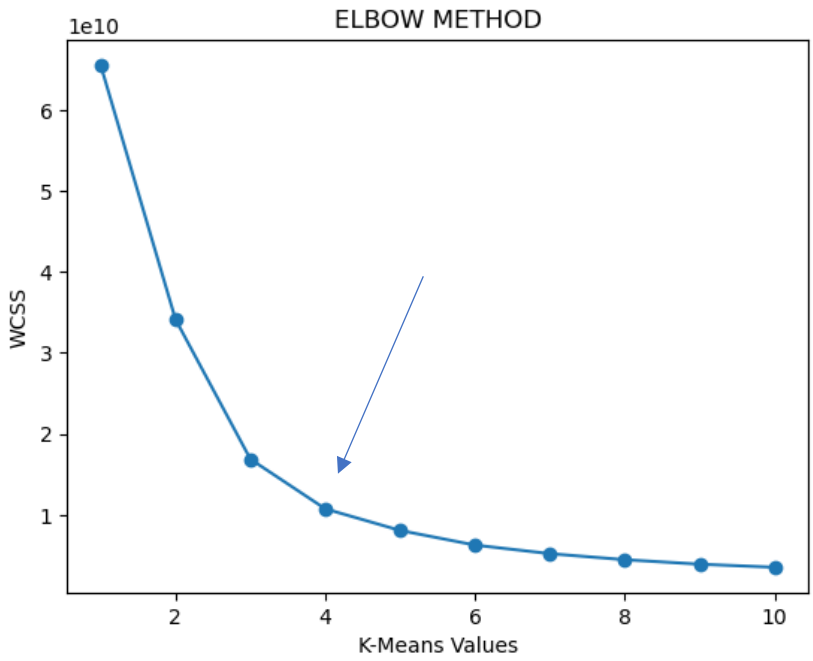
OUTPUT 4 for this part: To classify the customer using with the K-Means algorithm



7 customers are categorized in Cluster 4 that represents by this color: Red

Let's observe the **6 features** we found together and do **PCA**.

If we handle 6 features using with Elbow Method:



The optimum value of number of clusters is “4”.

That means **"4" is namely the optimum cluster number for:**

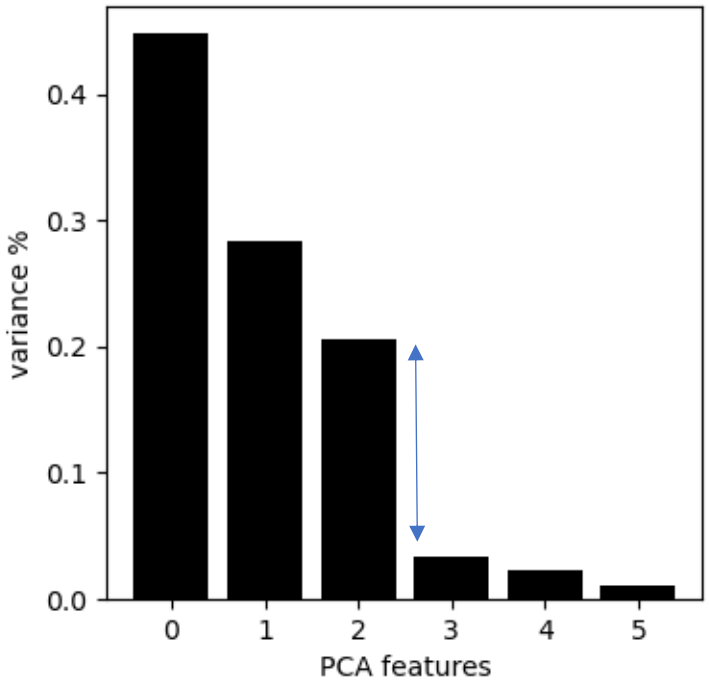
1) PURCHASES / ONEOFF_PURCHASES (Positive)

2) PURCHASES_FREQUENCY / PURCHASES_INSTALLMENTS_FREQUENCY (Positive)

3) CASH_ADVANCE_FREQUENCY / CASH_ADVANCE_TRX (Positive)

(all of these categories)

PCA (Principal Components Analysis)



We have observed that “3” is the official number for these features.

If we visualize it,