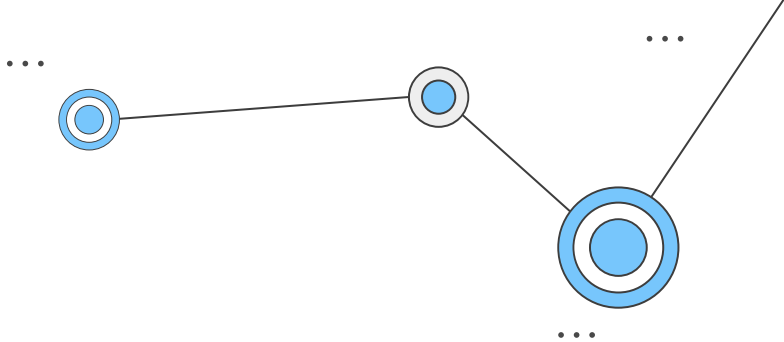
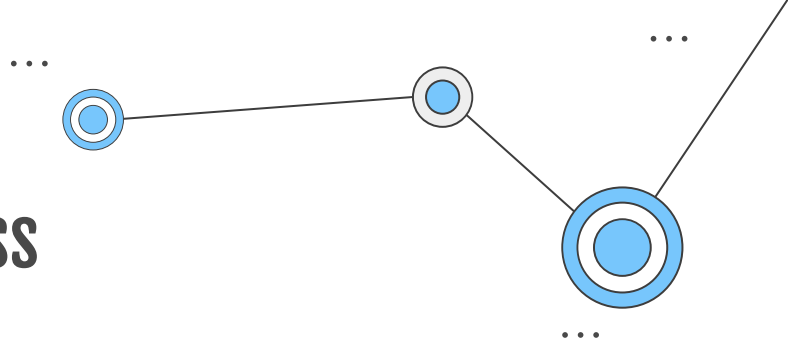
Customer Segmentation for Promotional Campaign Strategy

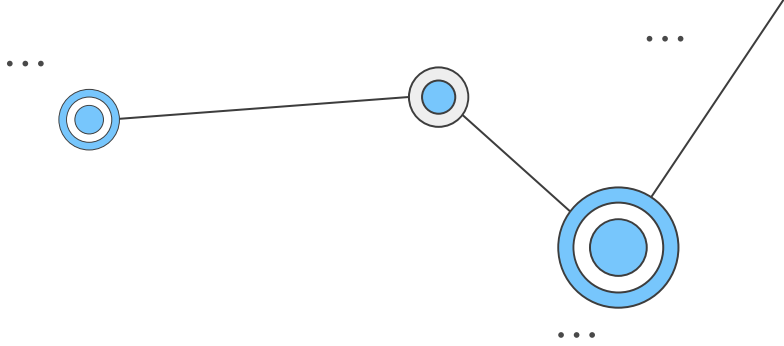


- Cleaning
- EDA
- MODEL - CLUSTERING
- RESULTS



Cleaning Process

- Dropped columns that had the same information across the board
- Dropped Nan values
- Got rid of irregular registrations in the Marital Status column
- Dropped outliers found in the Income column
- Processed categorical values in order to make them usable in the model
- Processed registrations that contained dates to make them usable in the model

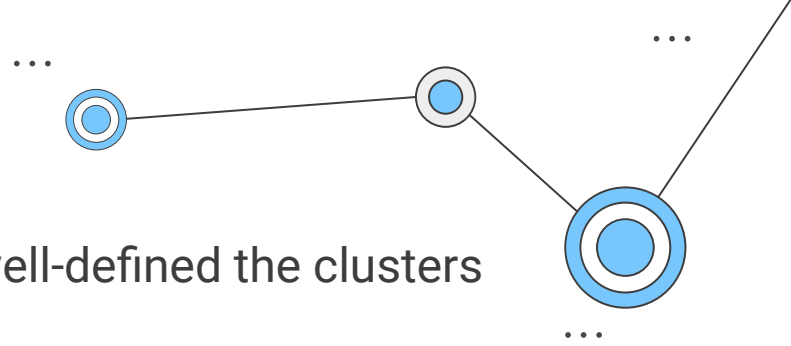


Exploratory Data Analysis Insights

The Amount of money spent on different product categories is highly correlated to the annual Income of the customer

Correlation Heatmap

Clustering using K-means



Silhouette score: a metric used to measure how well-defined the clusters are.

High silhouette score = Good

Low silhouette score = Bad

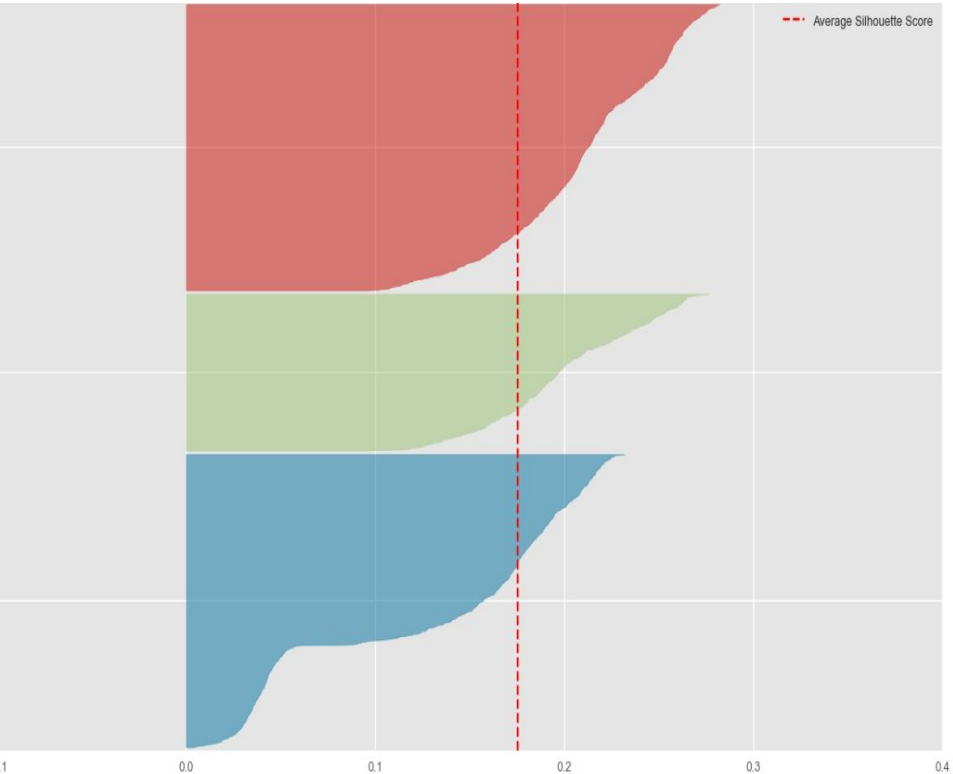
Two Clustering Paths

1. After Scaling our data
2. Without Scaling our data

1. Clustering on Scaled data

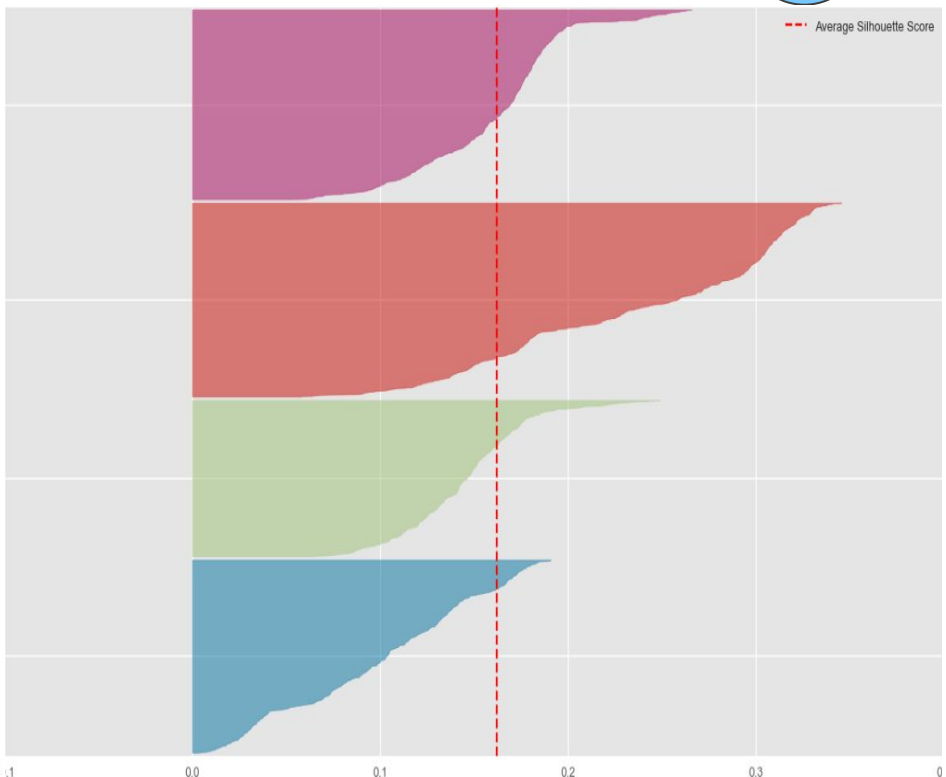
Clusters = 3

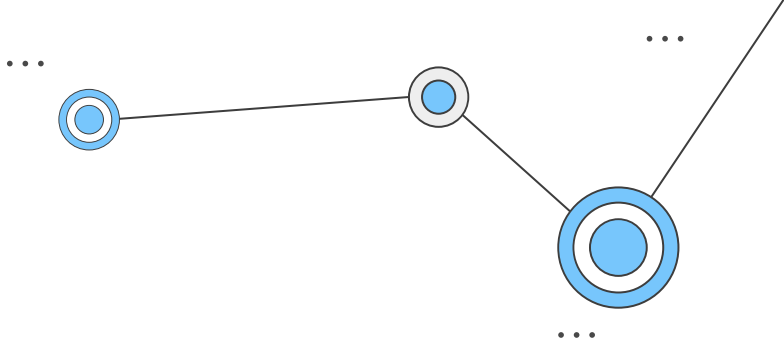
Silhouette score : 0.014 (low)



Clusters = 4

Silhouette score : 0.02 (low)

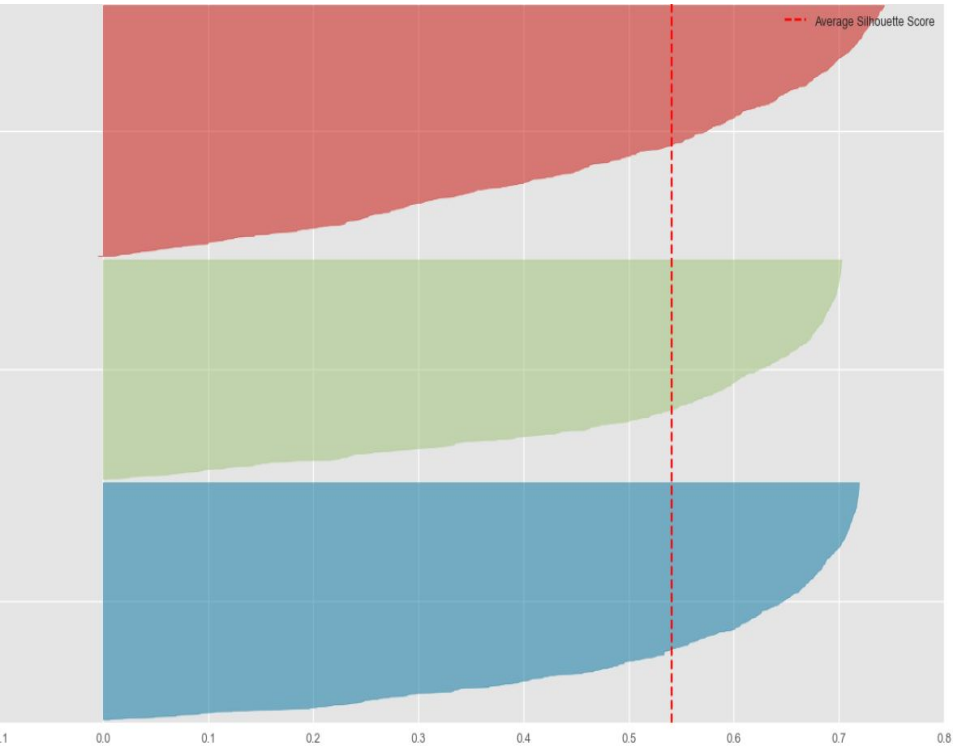




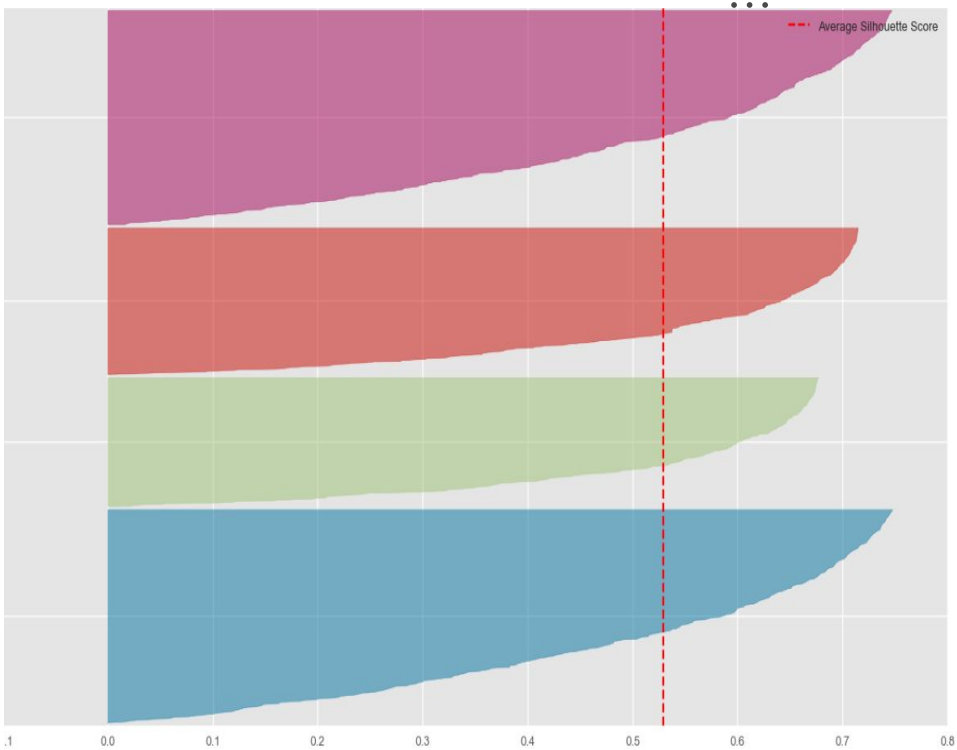
Clustering on Scaled data does not create meaningful results

2. Clustering Without Scaling

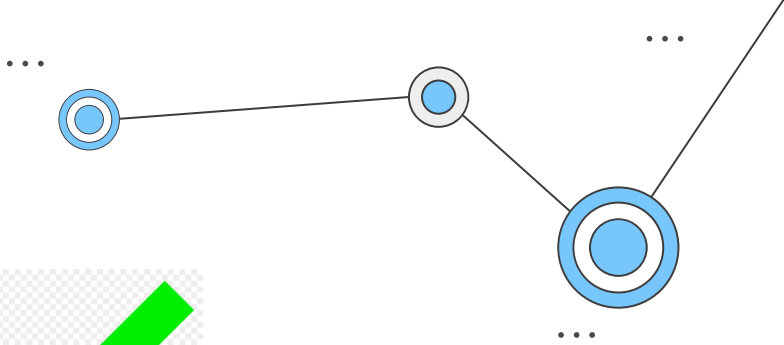
Clusters = 3
Silhouette score : 0.54 (high)



Clusters = 4
Silhouette score: 0.52 (high)

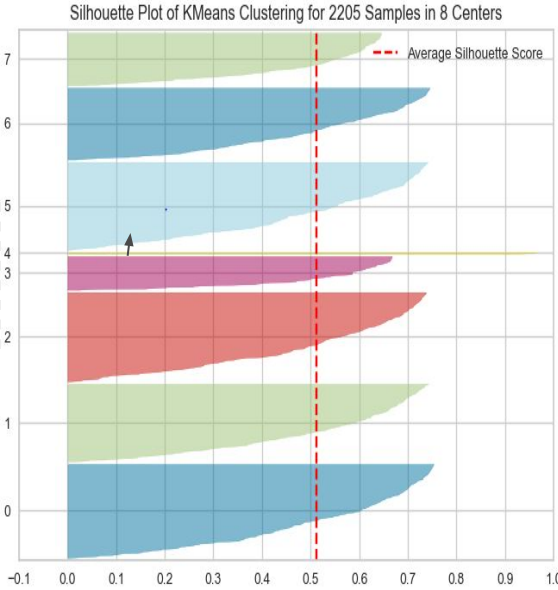


Choosing Number of Clusters

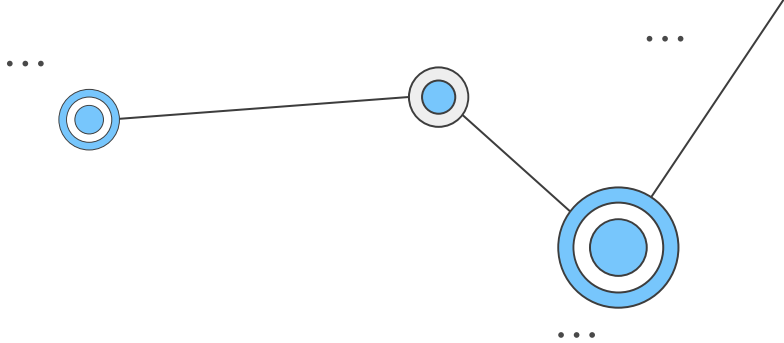


Any choice between 2-6 clusters is acceptable.

More than 7 clusters creates imbalances that we do not want



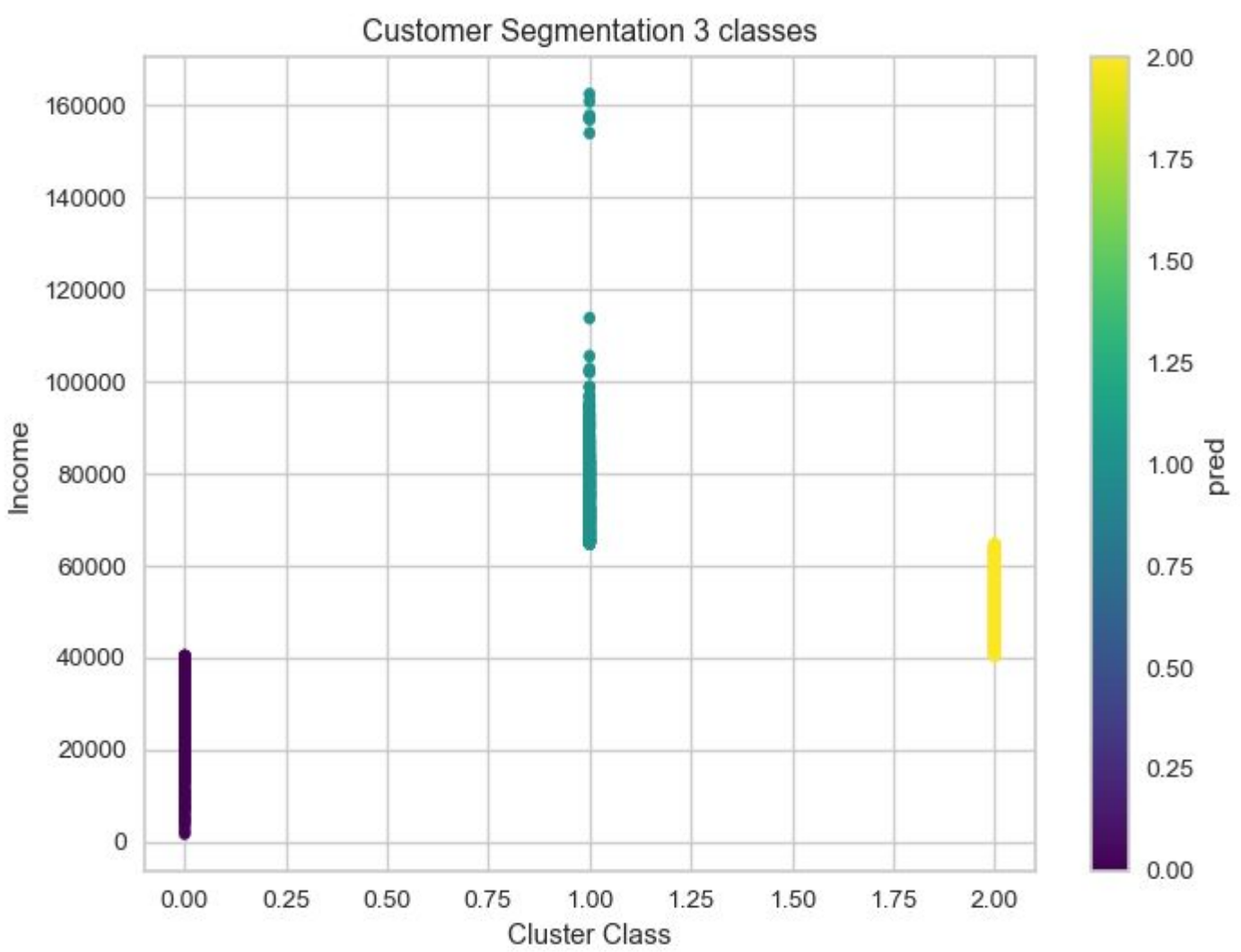
Explaining Customer Clusters

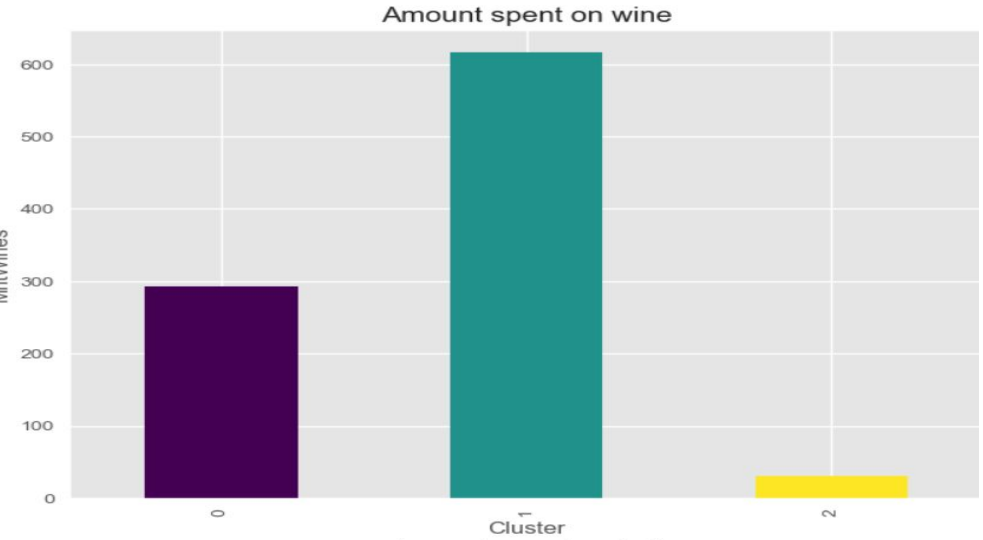


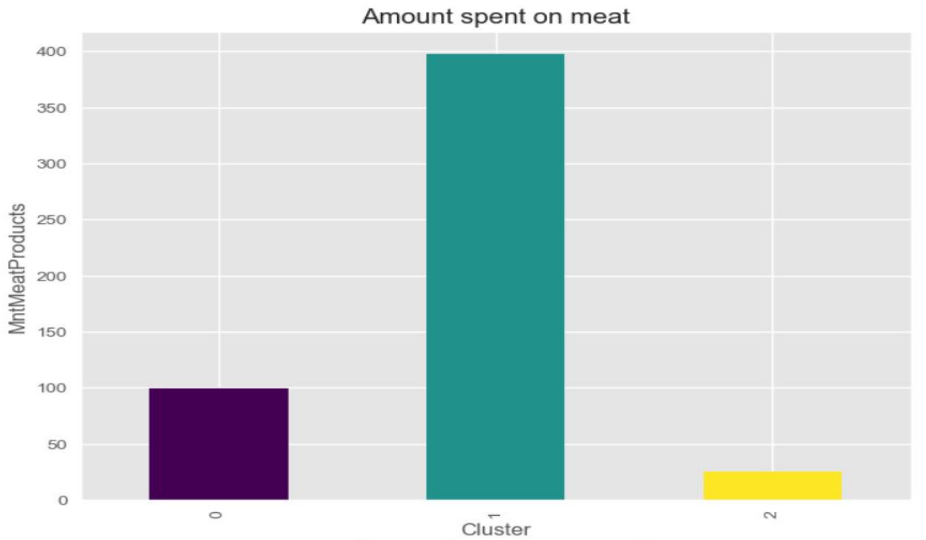
- The algorithm splits customers into different clusters based on their income

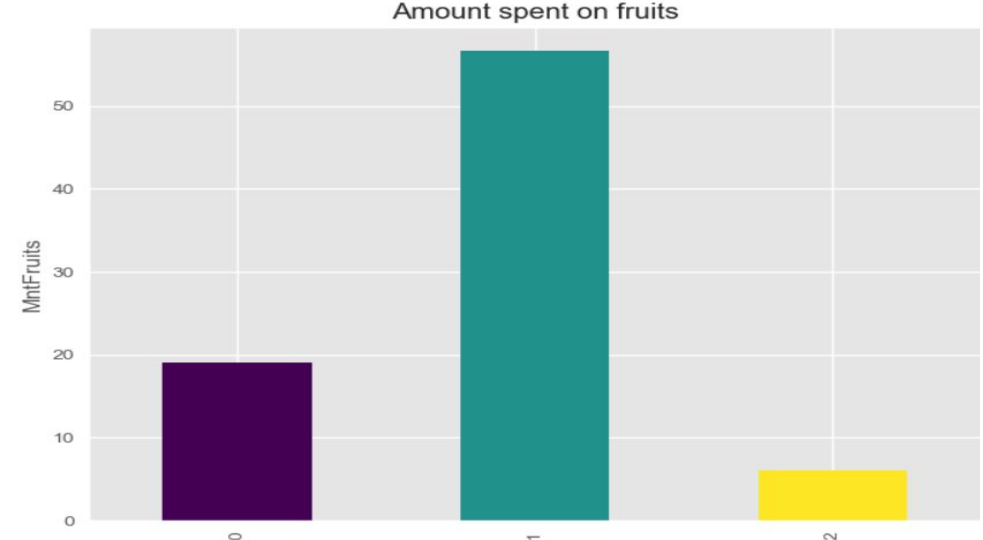
- Customers with Higher income spent more but buy similar amounts of products.
  - This means that they tend to make more expensive purchases

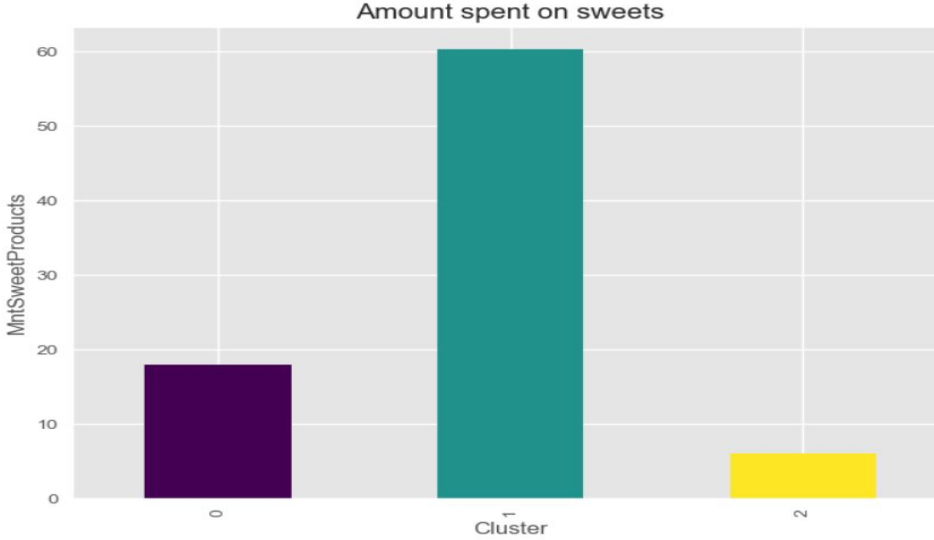
- Customers with Higher income buy less products on discounts


Customer Segmentation 3 classes
Income
160000
140000
120000
100000
80000
60000
40000
20000
0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
Cluster Class
pred
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00

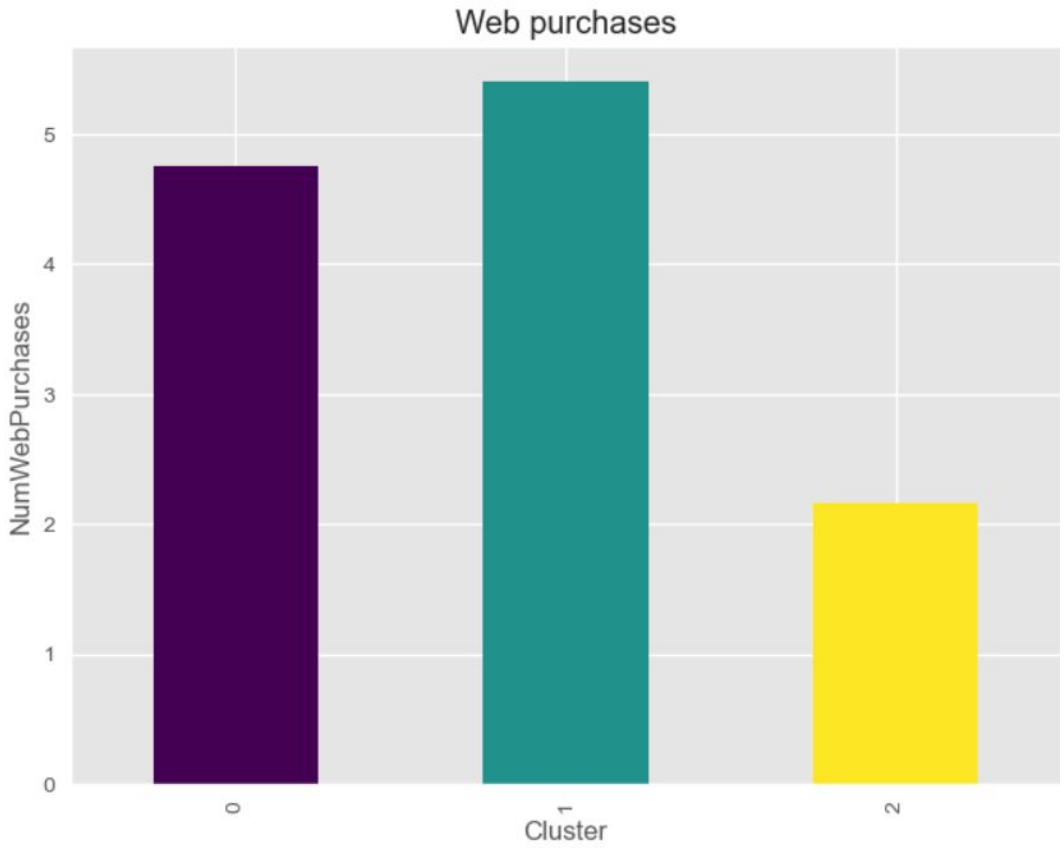

Amount spent on wine
MntWines
600
500
400
300
200
100
0
0
1
2
Cluster


Amount spent on meat
MntMeatProducts
400
350
300
250
200
150
100
50
0
0
1
2
Cluster


Amount spent on fruits
MntFruits
50
40
30
20
10
0
0
1
2


Amount spent on sweets
MntSweetProducts
60
50
40
30
20
10
0
0
1
2
Cluster

Customers with Higher income spent more but make similar amount of purchases



Customers with higher income do not take into consideration discounts

Closing Thoughts

- The number of clusters should be decided according to the number of different promotional campaigns the store is willing to run, and they should not exceed 6

- To lessen the impact of customer income in the clustering process we will need more data on the purchasing behavior of each customer (e.g. type of products bought, purchasing frequency)

Thank You

Nikolaos Tsakalas
Athanasios Alexandrakis