Customer Segmentation

Springboard Capstone Project
Ryan Paik

Problem

A global superstore has transaction orders from 2011 to 2015. The company wants to identify their customer behavior and actions. The focus is to use RFM segmentation and to classify future customers.

Data obtained from DataWorld

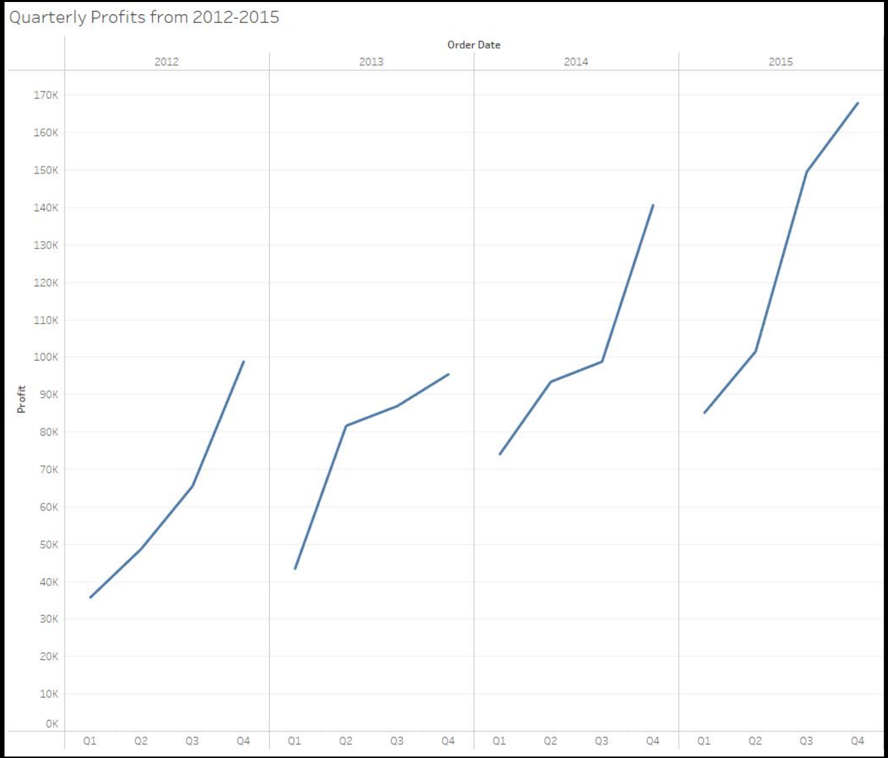
Who does this benefit?

The Global Superstore

- Investors
- Decision Makers
- Marketing Team

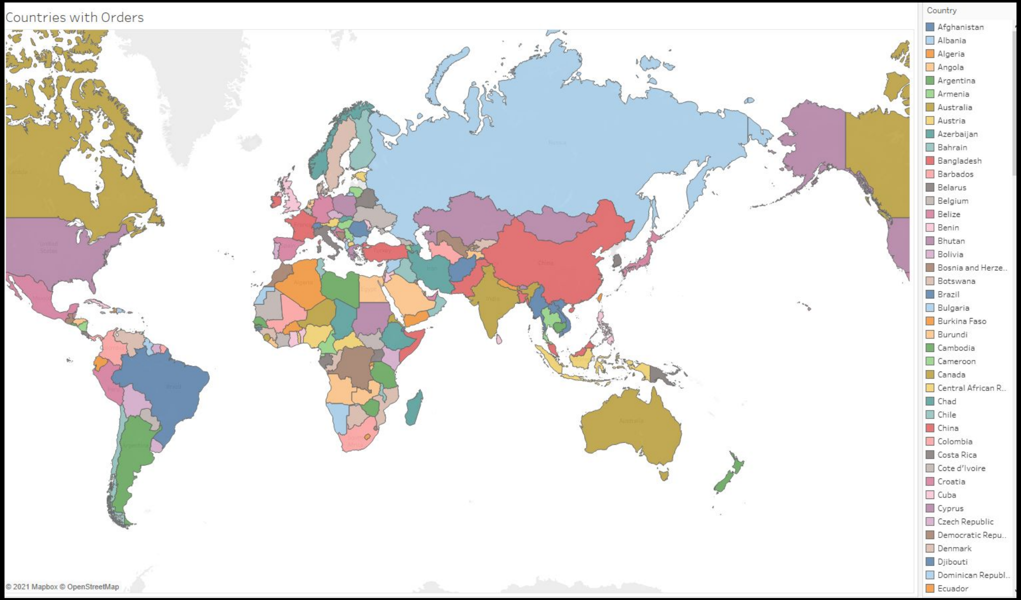
EDA

The Global Superstore is in a positive uptrend with their profits. Q4 has the highest peak each year.



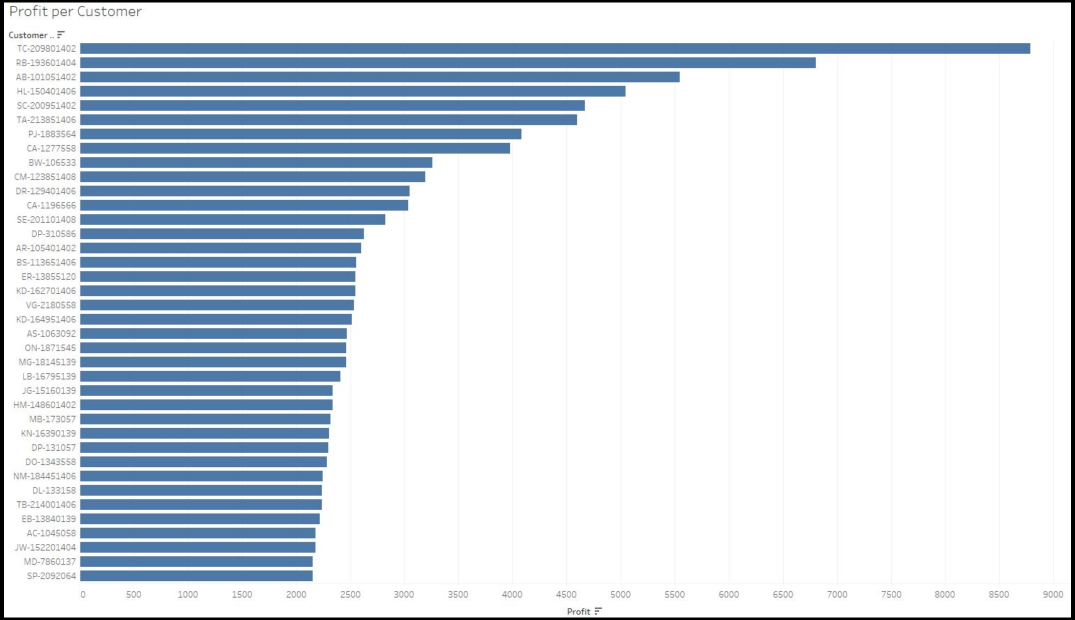
Location

- Well established internationally
- Data contains Region, Country, State, and City



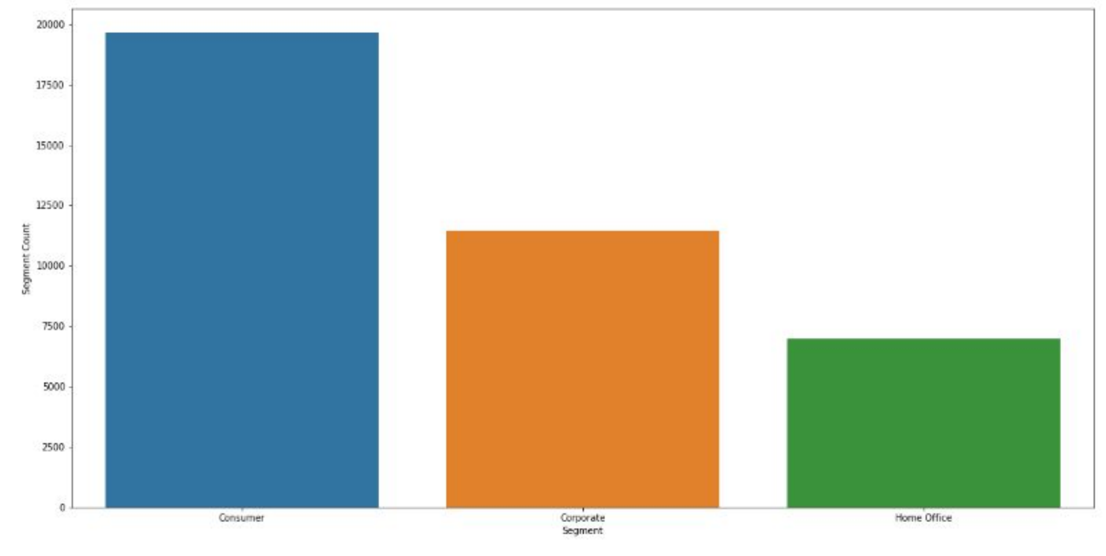
Customers

The highest profiting customer is almost $9,000.



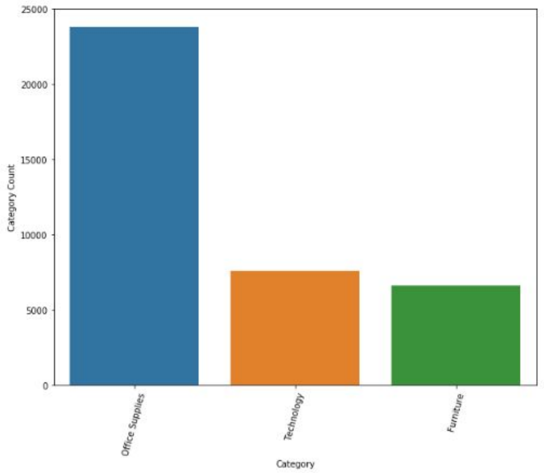
Customers

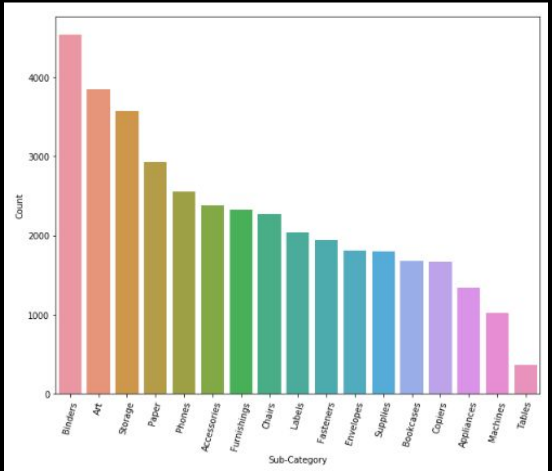
Consumers are the largest type of customers.



Products

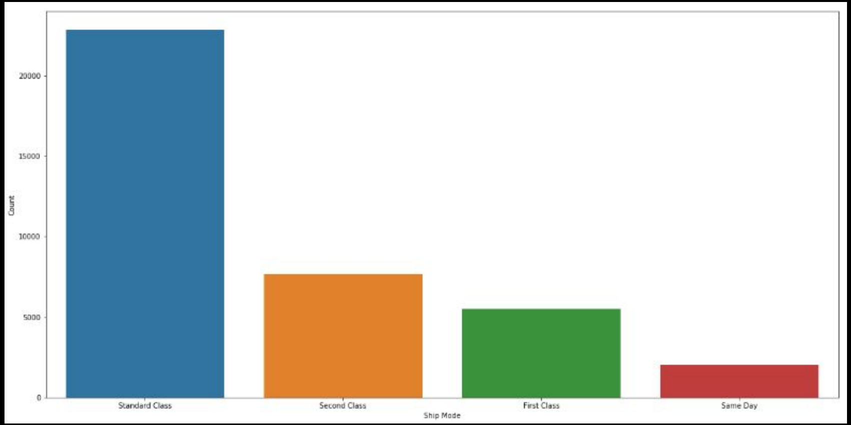
- Office Supplies is largest category based on orders
- Blinders and Art sub-categories have the largest number of orders

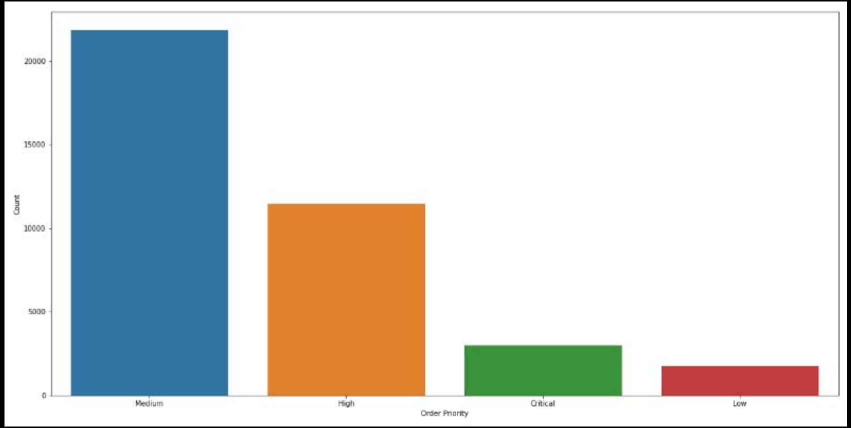




Products

- Majority of customers use standard class for shipping method
- Medium is the highest used order priortiy



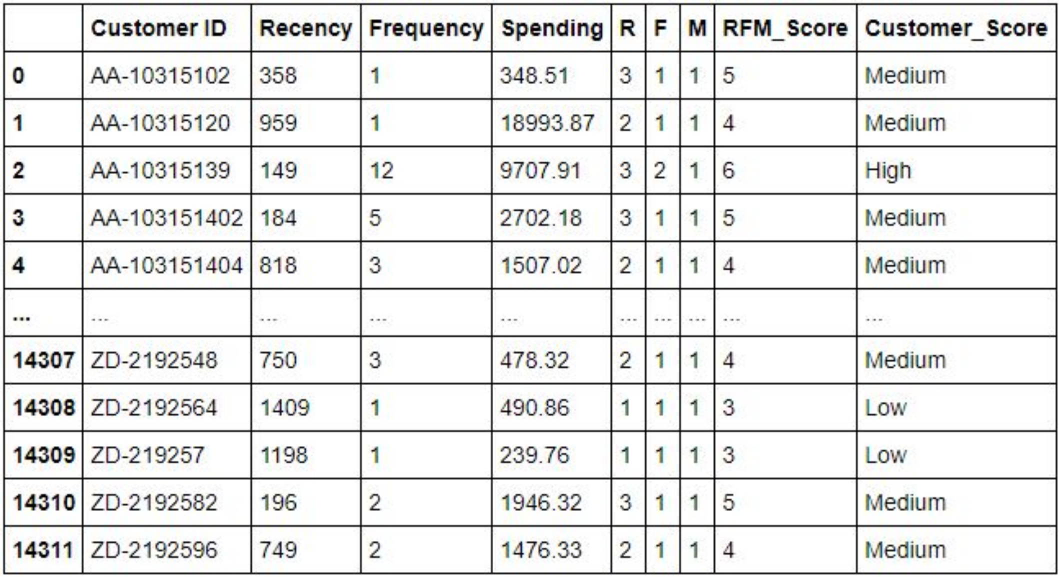


RFM Segmentation

- Recency: most recent activity from customer
- Frequency: how often the customer makes a transaction
- Monetary: the amount of spent by customer

Segmentation

Each customer score was created based on RFM. The highest score possible is 9 and lowest is 3. Based on their score, the customer is divided into Elite, High, Medium, and Low.

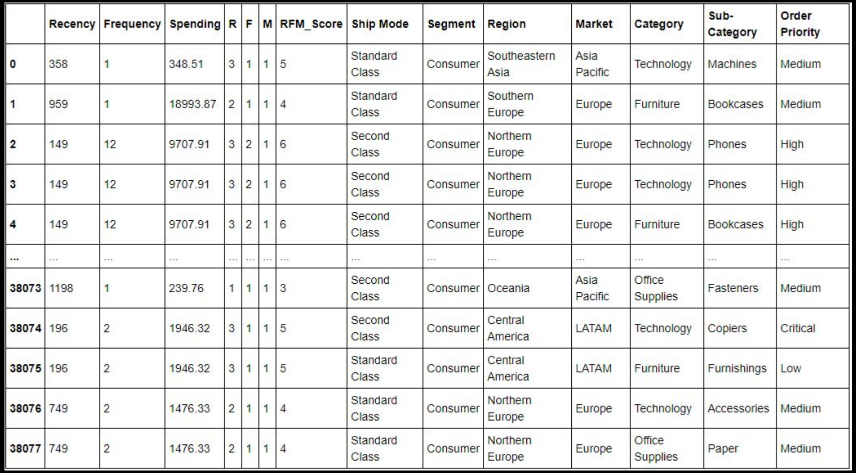


	Customer ID	Recency	Frequency	Spending	R	F	M	RFM_Score	Customer_Score
0	AA-10315102	358	1	348.51	3	1	1	5	Medium
1	AA-10315120	959	1	18993.87	2	1	1	4	Medium
2	AA-10315139	149	12	9707.91	3	2	1	6	High
3	AA-103151402	184	5	2702.18	3	1	1	5	Medium
4	AA-103151404	818	3	1507.02	2	1	1	4	Medium
...	...	...	...	...	...	...	...	...	...
14307	ZD-2192548	750	3	478.32	2	1	1	4	Medium
14308	ZD-2192564	1409	1	490.86	1	1	1	3	Low
14309	ZD-219257	1198	1	239.76	1	1	1	3	Low
14310	ZD-2192582	196	2	1946.32	3	1	1	5	Medium
14311	ZD-2192596	749	2	1476.33	2	1	1	4	Medium

Pre-Processing

Test data was made and added more categorical data to have a higher specificity in classification.

- Log was used to unskew the data
- Get_dummies was performed on categorical data
- StandardScaler and Robust Scaler was used to scale the data



	Recency	Frequency	Spending	R	F	M	RFM_Score	Ship Mode	Segment	Region	Market	Category	Sub-Category	Order Priority
0	358	1	348.51	3	1	1	5	Standard Class	Consumer	Southeastern Asia	Asia Pacific	Technology	Machines	Medium
1	959	1	18993.87	2	1	1	4	Standard Class	Consumer	Southern Europe	Europe	Furniture	Bookcases	Medium
2	149	12	9707.91	3	2	1	6	Second Class	Consumer	Northern Europe	Europe	Technology	Phones	High
3	149	12	9707.91	3	2	1	6	Second Class	Consumer	Northern Europe	Europe	Technology	Phones	High
4	149	12	9707.91	3	2	1	6	Second Class	Consumer	Northern Europe	Europe	Furniture	Bookcases	High
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
38073	1198	1	239.76	1	1	1	3	Second Class	Consumer	Oceania	Asia Pacific	Office Supplies	Fasteners	Medium
38074	196	2	1946.32	3	1	1	5	Second Class	Consumer	Central America	LATAM	Technology	Copiers	Critical
38075	196	2	1946.32	3	1	1	5	Standard Class	Consumer	Central America	LATAM	Furniture	Furnishings	Low
38076	749	2	1476.33	2	1	1	4	Standard Class	Consumer	Northern Europe	Europe	Technology	Accessories	Medium
38077	749	2	1476.33	2	1	1	4	Standard Class	Consumer	Northern Europe	Europe	Office Supplies	Paper	Medium

Machine Learning Models

Type: Unsupervised Learning - Clustering

Datasets: Standard scaled and Robust scaled

Models: KMeans, Agglomerative Clustering, DBSCAN

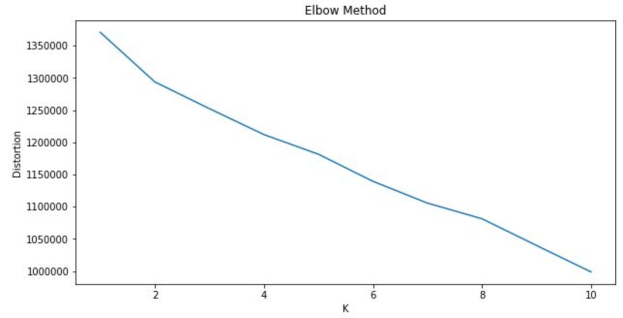
Model Evaluation: Silhouette Coefficient and Davies Bouldin

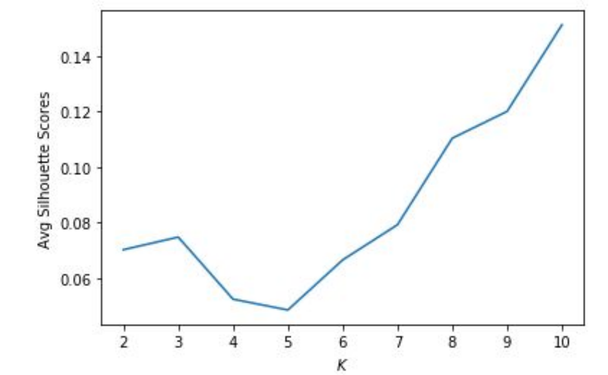
Standard Scaling - KMeans

Elbow method is ambiguous. Silhouette Method showed optimal number of clusters is 10.

Silhouette Coefficient: 0.144

Davies Bouldin: 2.475



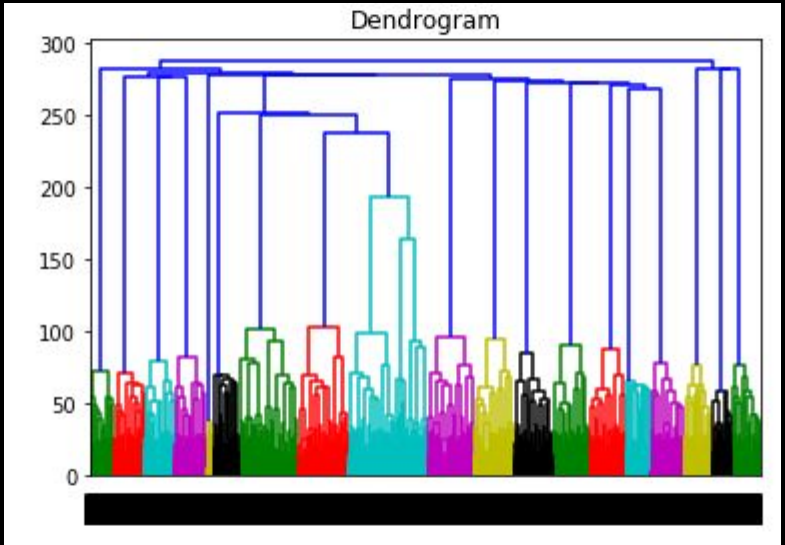


Standard Scaling - Agglomerative Clustering

Optimal number of clusters is 19.

Silhouette Coefficient: 0.220

Davies Bouldin: 1.730



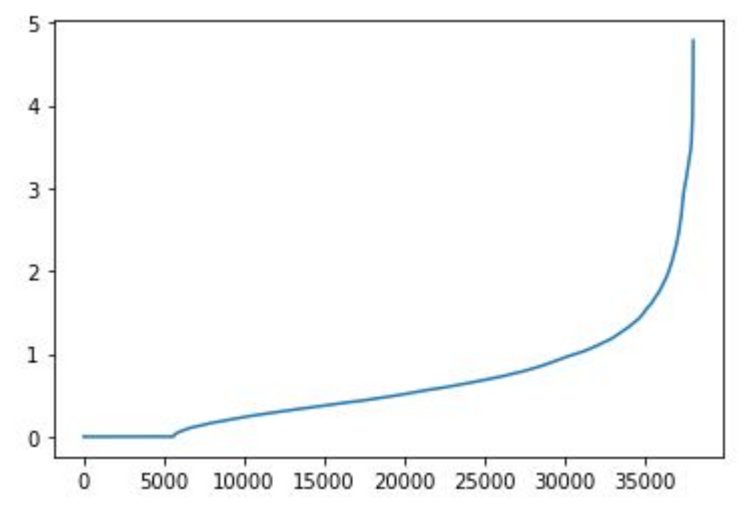
Standard Scaling - DBSCAN

Eps: 2.5

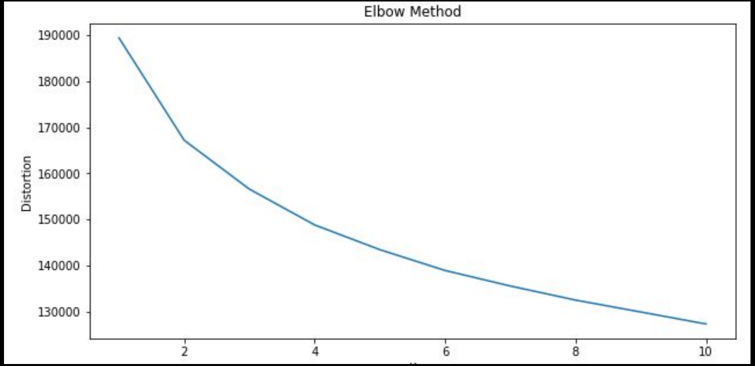
Min_samples: 200

Silhouette Coefficient: -0.184

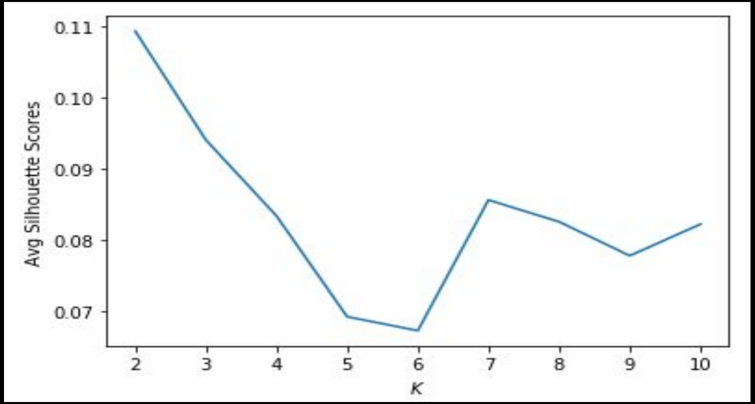
Davies Bouldin: 1.744



Robust Scaling - KMeans



Elbow method is ambiguous. Silhouette Method showed optimal number of clusters is 7.



Silhouette Coefficient: 0.080

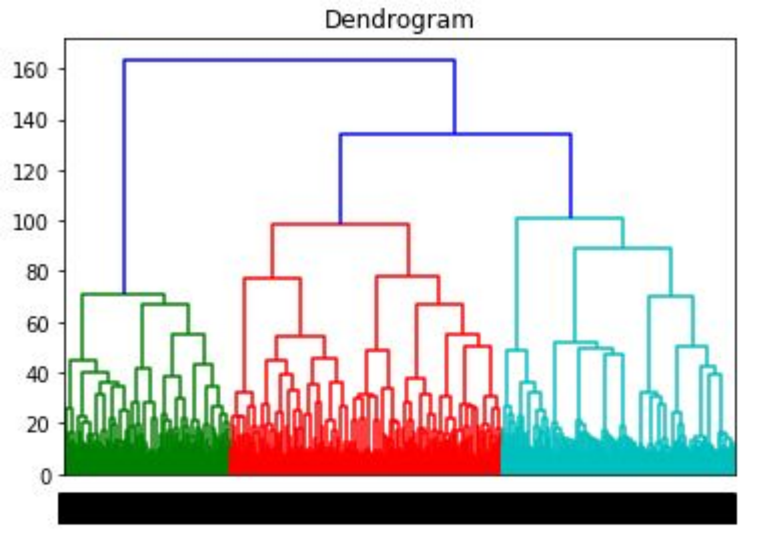
Davies Bouldin: 2.571

Robust Scaling - Agglomerative Clustering

Optimal number of clusters is 3.

Silhouette Coefficient: 0.051

Davies Bouldin: 3.415



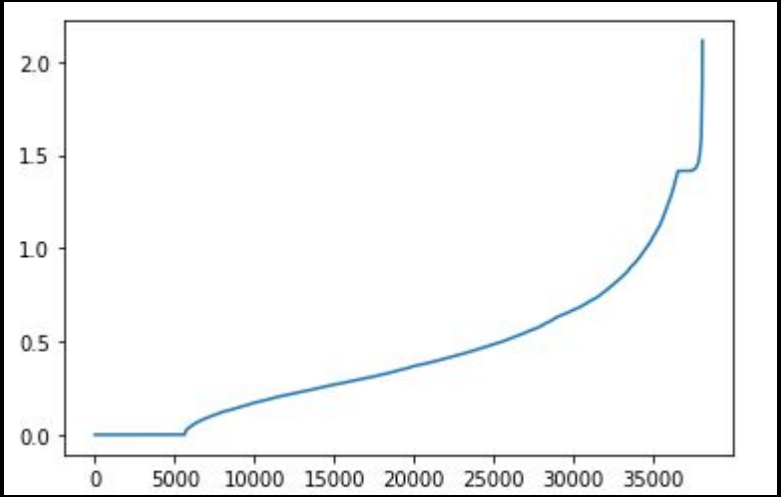
Robust Scaling - DBSCAN

Eps: 1.55

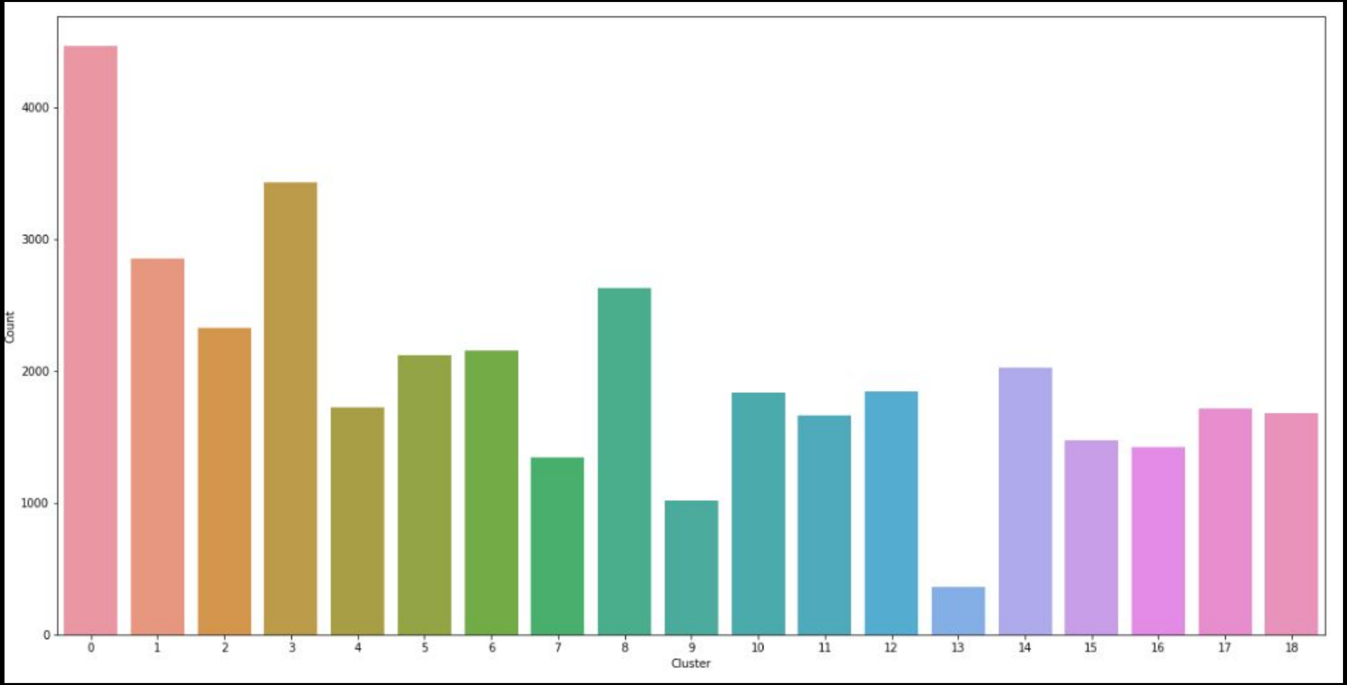
Min_samples: 150

Silhouette Coefficient: 0.090

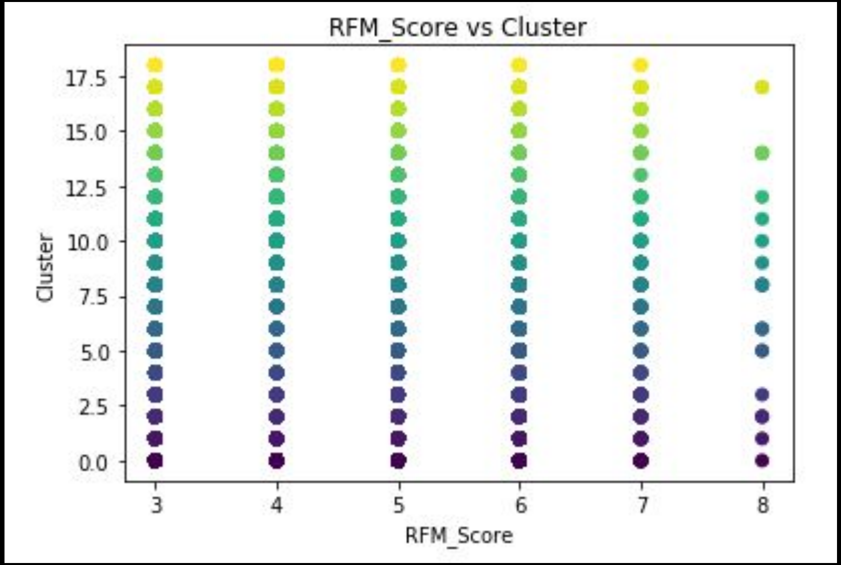
Davies Bouldin: 4.201

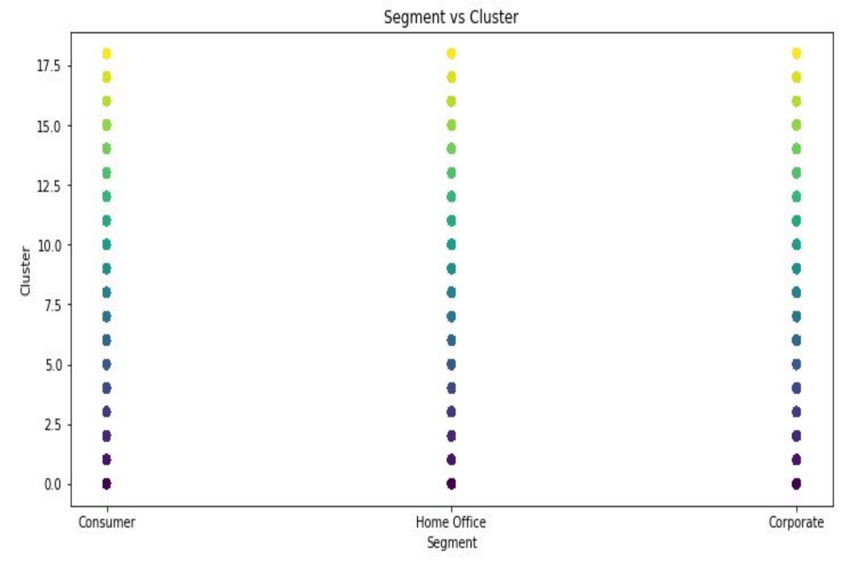


Standard Scaling - Agglomerative Clustering



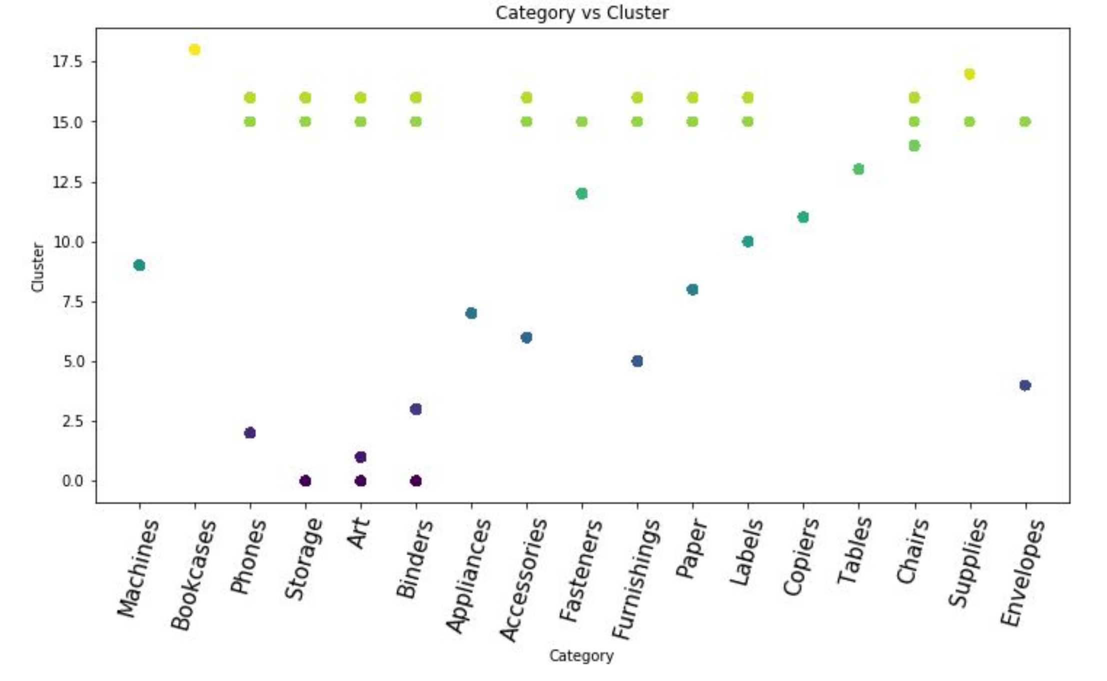
Standard Scaling - Agglomerative Clustering





Many of the other features were similar in difficulty in identifying specific customers among the clusters based on features.

Standard Scaling - Agglomerative Clustering



Conclusions

Model: Agglomerative Clustering with standard scaled method

Future Improvements:

The dataset requires further analysis since based on individual feature, there are indistinguishable clusters among the features.

- Each categorical features can be analysed based on its distribution.
- Other categorical features can be used in the analysis if able to increase computational power.