Ryan Paik

Final Report:
Customer Segmentation

Problem Statement

A global superstore has obtained data transaction orders from 2011 to 2015. The business wants to learn about their customer's behavior for the business to scale efficiently and effectively. To identify customer's behavior, the RFM (recency, frequency, and monetary) method will be used to distinguish and identify customers. This method will use their most recent purchase was, how frequently the customer makes an order, and how much the customer has spent to differentiate customer's behavior. Once customers are segmented, other features that are provided with the data will be used to classify future customers based on previous customer's behaviors.

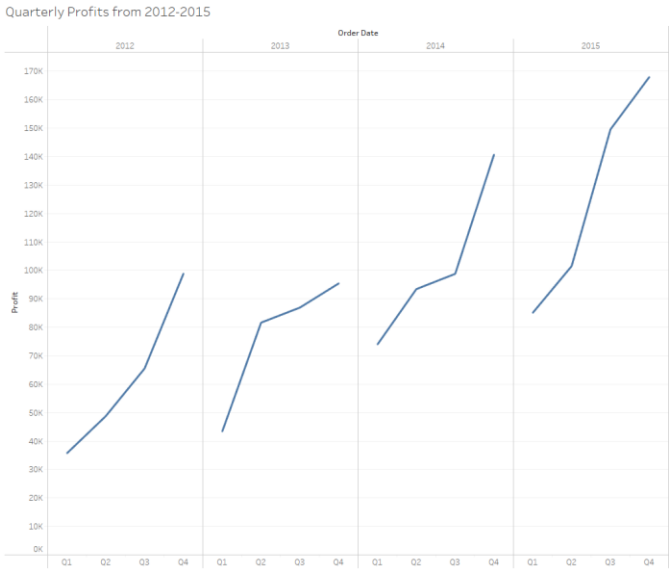
Data Wrangling

The data contains transaction orders and has 24 columns and 51,290 rows. The data has multiple information about customers, location, products, and profit. The only missing data is in the postal code column and 80.5% of the column is missing. This column cannot be filled and will be dropped. Also, the profit column contains 672 of 0 values and these transactions will be dropped. Furthermore, the discount column is removed to focus on customers that have increased business profit. The data is ready to begin customer segmentation.

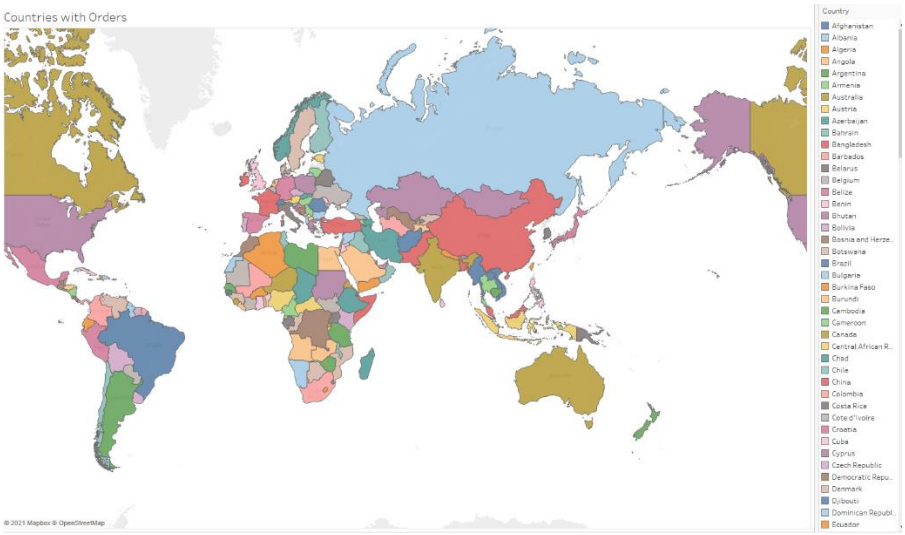
Exploratory Data Analysis

The dataset columns can be separated into 4 different categories to do a more efficient data analysis: sales, location, customer, and products.

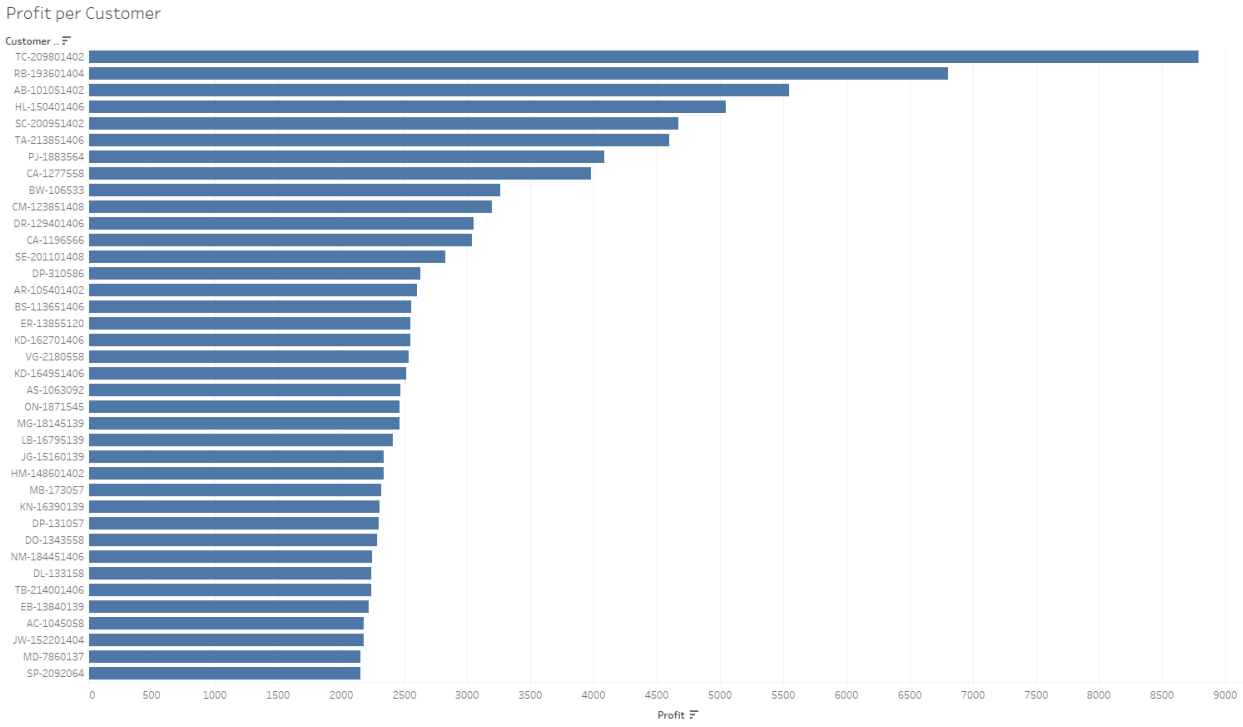
Sales: The sales of the business have been increasing overall throughout the past 4 years. A pattern can be seen that the profits is peaked during Q4, which can be assumed given the holidays. The positive increase is a good indication that the customers that will be segmented contains reliable data.



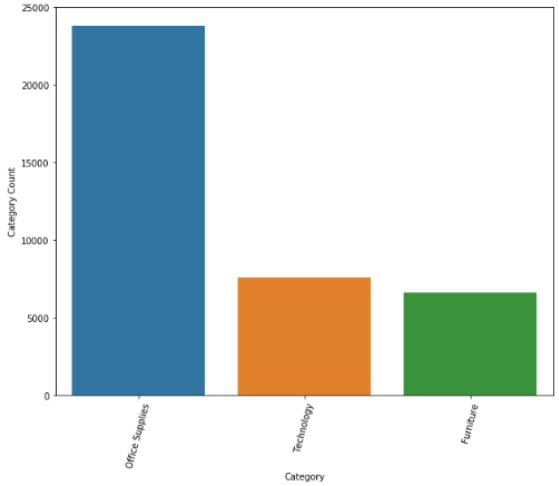
Location: The business is well established internationally as there are 151 unique countries, 929 unique states, and 3257 unique cities with orders. Due to the vast amount of categorical data, the region will be used as a feature for the models. All other location columns will be dropped.

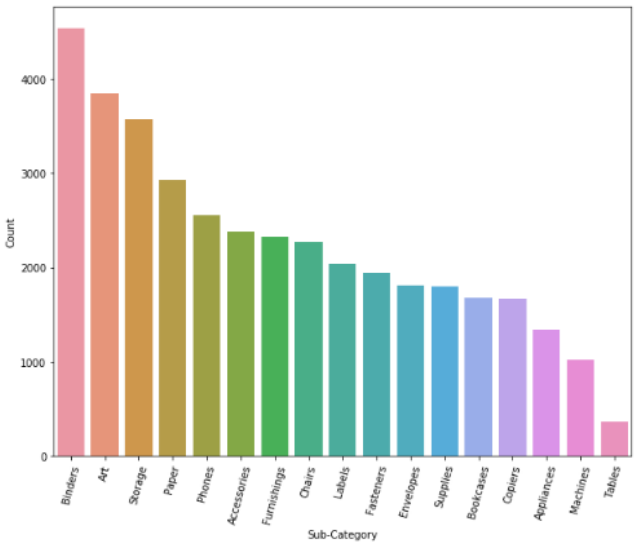


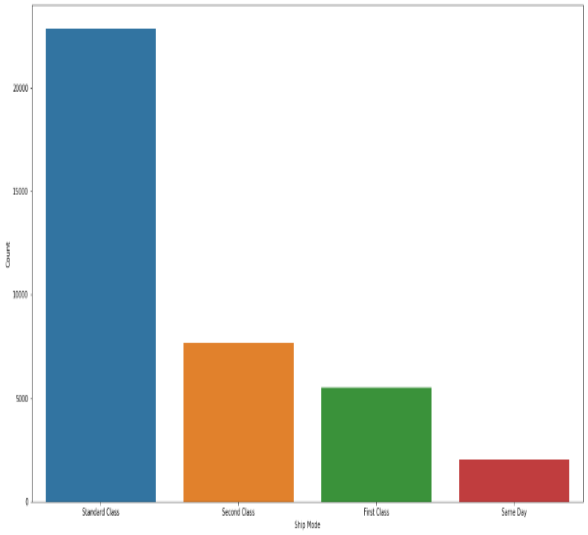
Customer: The profit per customer can be seen here. TC-209801402 is the highest profiting customer for the business.

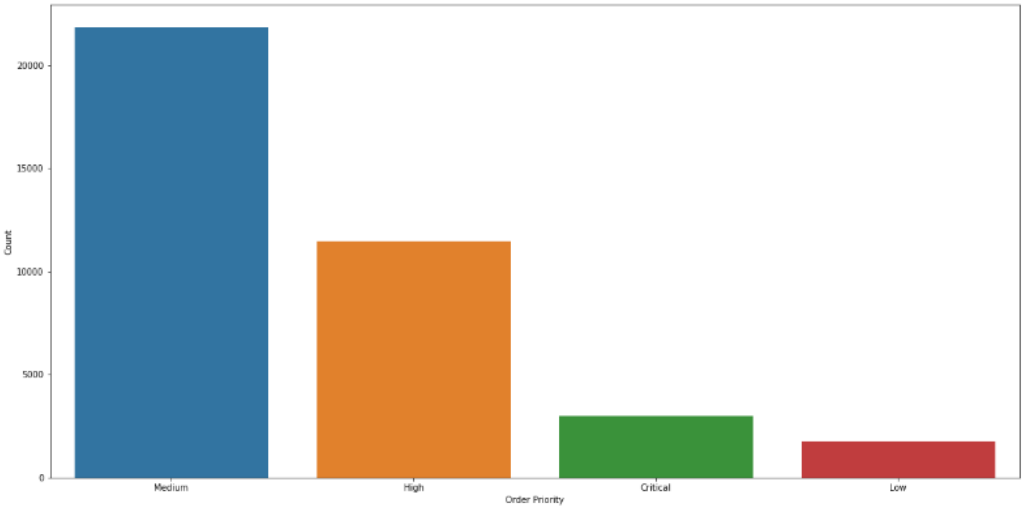


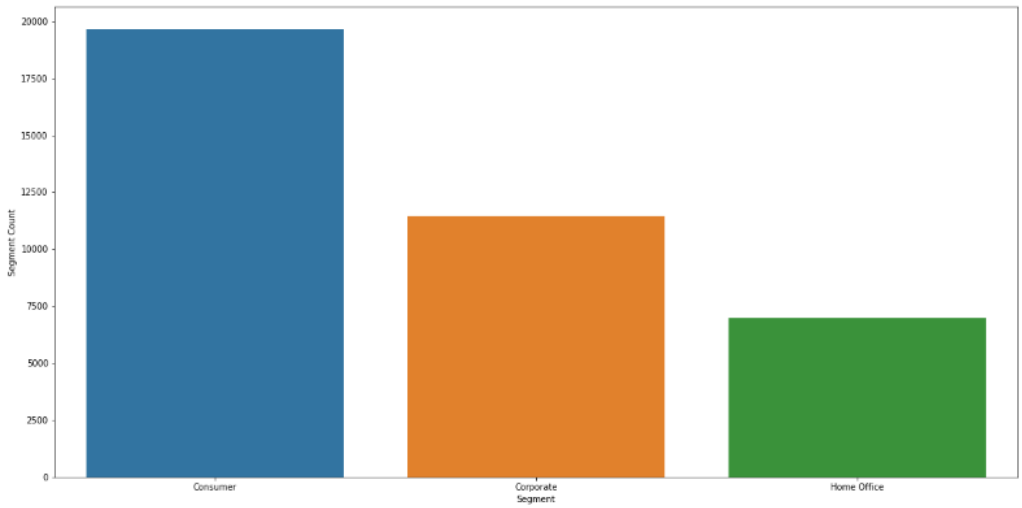
Product: The products have many categorical data that can be used with feature selection in classifying future customers.




Count
4000
3000
2000
1000
0
Binders
Art
Storage
Paper
Phones
Accessories
Furnishings
Chairs
Labels
Fasteners
Envelopes
Supplies
Bookcases
Copiers
Appliances
Machines
Tables
Sub-Category


20000
15000
10000
5000
0
Standard Class
Second Class
First Class
Same Day
Ship Mode


Count
20000
15000
10000
5000
0
Medium
High
Critical
Low
Order Priority


Segment Count
20000
17500
15000
12500
10000
7500
5000
2500
0
Consumer
Corporate
Home Office
Segment

Pre-Processing

The customer's behavior will be segmented by the RFM analysis. RFM stands for Recency, Frequency, and Monetary value and each category is a key representation of customer traits.

- R: most recent activity from customer
- F: how often the customers make a transaction
- M: the amount of spending by customer

Recency:

The recency column is made by obtaining the last date of orders. This date is used and subtracted from the other dates. This will identify how many days the order was made in comparison to the most recent date. This order date is grouped with customer ID.

Frequency:

The frequency column identifies how many orders each customer was made. A groupby was used on the customer and a sum of count was aggregated for each customer.

Monetary:

The monetary column was named as spending and was found similarly to the frequency method. The Revenue was column was used and summed for each customer.

Once the RFM column was created for each customer, the columns were divided by quantiles and 3 groups were created for each group. The group they are identified as is the score they are given for each column. The RFM score is made by adding each column. A higher-level customer will have a higher RFM score. The maximum score a customer can obtain is 9 points and the lowest is 3 points. Each customer is segmented by the RFM score: 3 points is low, 4-5 points is medium, 6-7 points is high, and 8-9 is elite. With these customer segmentations, the data is merged back with the data to use categorical data to classify future customers. The categorical data that will be used are Ship Mode, Segment, Region, Market, Sub-Category, and Order Priority.

The RFM columns were heavily right skewed. To un-skew the date, the RFM columns were logged. The categorical data were changed to numerical values by using the get_dummies feature. This in turned changed the data set from 14 columns to 36 columns. The data is unskewed and logged to have a better normal distribution and the categorical data was changed into numerical values by scaling. 2 different datasets were created based on different scaler methods: Standard and Robust. Standard scaling uses the mean and standard deviation whereas the Robust method uses outliers. The data is ready to be tested by models. PCA was not done since there was enough computational power with the number of dimensions.

Model Selection

In this unsupervised machine learning dataset, 3 machine learning models were used: KMeans, Agglomerative Clustering, and DBSCAN. Both standard scaling and robust scaled datasets were used. To identify the number of clusters for KMeans, the elbow and silhouette method were used. For the hierarchical clustering, the hierarchical graph was evaluated to identify the number of clusters. For the DBSCAN, epsilon was identified with Nearest Neighbors and trial-and-error was done for min_samples. To compare each model, the silhouette coefficient and Davis Bouldin score were used. Overall, the agglomerative clustering with the standard scaling dataset had the best silhouette and Davies Bouldin score. The agglomerative clustering had 19 number of clusters. Although the agglomerative clustering was the best model, many of the features were indistinguishable among each other. The best feature that was distinguishable were Sub-Categories.

Future Improvements

The dataset can require further analysis based on the categorical features. The distribution can be evaluated to understand why the machine learning models are having difficulty separating each cluster. Furthermore, there were many other features that were included in the dataset; however, were not used due to computational power and dimensionality. The other features can be used to do further analysis.