SAINT MARY'S UNIVERSITY SINCE 1802

One University. One World. Yours.

Data and Text Mining

MCDA5580

Team Members:

NAME	A#	Email
Manoj Bandaru	A00433174	Manoj.bandaru@smu.ca
Aditya Tandon	A00432835	Aditya.Tandon@smu.ca
Khagesh Pandya	A00431429	Khagesh.pandya@smu.ca

Contents

List of Figures

List of Tables

1. Executive Summary:

We have online retail transactions data to make business insights and to understand customer and products in a way such that organization can execute different strategies on various departments like marketing, inventory storage, offers etc. Based on the data and by taking in account of appropriate features, different clusters are created, where each cluster contains same attributes. Based on clusters we can decide on which segment needs attention, also the preventive measures required.

2. Objective:

Objective is to do market basket analysis or customer segmentation by using appropriate datamining algorithms and methodology. The main objective of this report is to use K-means clustering algorithm and cluster the data into different segments. The end goal is to make business insights based on the clusters and segments formed.

3. Data Summary:

Data used in the assignment is about Online Retail store transactions. Transactional data contains multiple entry of different products for single transaction. So, it is mandatory to perform some operations to get appropriate features for customers and products.

Columns:

- **Invoice No:** It is unique transaction ID to identify transaction.
- **StockCode:** StockCode is unique ID for product.
- **Quantity:** Number of quantities sold for each product in a given transaction.
- **Description:** Description about product.
- **Invoice Date:** Transaction date and time in string.
- **Unit Price:** Unit price for each product.
- **CustomerID:** Unique ID for the given customer.
- **Country:** From where transaction took place.
- **Invoice Date Time:** Timestamp for the transaction.

Data is in range of 1/12/2010 to 31/12/2011. The entire year will give a better insight, so we've taken into consideration of only 2011-year data for clustering.

The Data we have is about Online Retail transaction data, using SQL we changed the OLTP to OLAP data for both Product and Customer analysis. The useful columns in this transactional data has been identified and made into OLAP data used for clustering and for giving useful Business insights.

As per our observation, we segregated and aggregated required columns for forming clusters. The limitations of this data are, we cannot perform clustering based on Demographics, Climate, Store Typed, Store capacity, Competition because the data doesn't give required information. Further, based on the INVOICE time and date we divided into 4 quarters. The product data was divided into 4 different tables, based on which quarter of the year data lies in, but the customer data was not divided like that.

Open Refine was used to detect the outliers and to remove them. The data that doesn't lie near the trend line made using the data points and hence were removed as outliers to form better clusters.

4. Design/Approach/Method:

4.1 Selecting the appropriate data:

Wrote an SQL command to filter out the customer data and the product data from the online Retail Database. After forming different tables for both the customer data and product data, we had to find additional parameters for both the cases.

4.2 Feature selection to support the analysis:

Features are selected for Customer Data which were Distinct Products bought, Number of Products Bought, Revenue, Visits, Average Spend and Standard Deviation, excluding the CustomerID which was used to map the cluster data with the original data.

Four features were selected for Product Data which were Baskets, Distinct Customers, Average Price and Total Revenue, excluding the StockCode which was needed for mapping of the data.

4.3 Clean the data and remove outliers:

Using Open refine we cleaned the data along with removing the outliers that were out looking out of the way, seeing the trendline of by plotting the data.

4.4 Scale Data:

Scaling the data is required to put the data points in a similar range, as they are skewed in the starting and we want them to be in a normal distribution, to make better clusters.

4.5 Determine the appropriate number of clusters and perform clustering:

It's important to determine the number of clusters. We used two methods for determining the clusters, which were Elbow method and the Silhouette Score. Elbow method was counted by calculating the 'tot.withinss' of the clusters, which is total value of the Sum of Squares between the data points inside the cluster. After getting all the 'tot.withinss' values, we plot the elbow point. Whereas, Silhouette Score is calculated using the difference between distance of one random data point from all the data points within the cluster and with the points of the neighboring cluster divided by the difference which is greater than the two mentioned above. After that, we performed K-means clustering to divide the data points in the clusters.

4.6 De-normalize data:

After performing Clustering, we must map the data to perform Cluster Analysis. To do that, we have to un-scale the data or De-Normalize the data.

5. Feature Selection:

Feature selection was taken based on the attributes which are useful for forming better clustering and to get better business insights.

The Customer Cluster Features are as below:

Feature Name	Measurement	Description
CustomerID	N/A	Unique field for identifying items
Distinct Products Bought	COUNT(DISTINCT)	Total number of products that have purchased by the customer at some point
Number of Products bought	COUNT(DISTINCT)	Total number of products bought
Revenue	SUM	Total Revenue generated by the customer
Visits	COUNT(DISTINCT)	Total number of visits of customer
Avg_Spent	Average	Average spent by Customers
Standard Deviation	Std.dev	Standard Deviation of time between purchases for each customer

Table 1: Customer Cluster Features

The Product Cluster has following attributes:

Feature Name	Measurement	Description
StockCode	N/A	Unique field for identifying items
Distinct_Customers	COUNT(DIST)	Total number of distinct Customers who have purchased the item at some point
Average_Price	Avg_Price	The average price of the item. This compensates for changes in price across the dataset and does not include instances where the price is $0.00, which may indicate a special price calculated for BOGO or coupons.
Total_Revenue	SUM	Total Revenue generated by the customer
Baskets	COUNT(DIST)	Total number of visits of customer

Table 2: Product Features Descrtiption

6. Data Cleansing/ Outlier Removal:

In any analysis of clustering Data cleansing and Outlier removal is very important as it might mislead taking business decisions and forming clusters, in our analysis we used an Open Refine tool for outlier removal. Also, some outliers are removed by using R by identifying the data points which make a significant difference in Cluster formation.

In Open Refine, the data has been loaded and transformed every into numeric. For each column we created numeric facet then changing the facet minimum and maximum values. For product cluster we have four attributes and customer clustering has six, as the process of removing outliers we created facets for all attributes, and we neglected values which is not in data distribution in facet based on all attributes for all quarter-product clusters and customer clusters.

Outlier Removed for Customer Data:

- No. of Product Bought – Greater than [1000]
- Distinct Products Bought - Greater than [300]
- Revenue – Greater than [30,000]
- Visits – Greater than [20]
- Average Spent – Greater than [8000]
- Standard Deviation – Greater than [260]

Outlier Removed for Product Data:

- Average Price - Greater than [50]
- Total Revenue – Greater than [13,000]
- Distinct Customers - Greater than [150]
- Baskets - Greater than [200]

7. Cluster Analysis:

After preprocessing data is passed to k means clustering algorithm. Which returned clustered object. To identify best number of required clusters elbow method is used and to make more accurate silhouette method is used support the results.

7.1 Product Clusters:

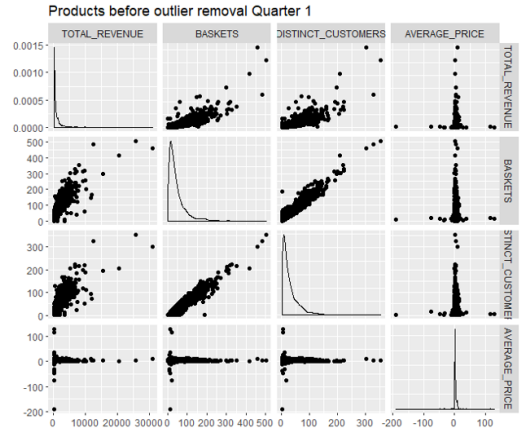
To perform clustering on product the data of given year is divided into different quarters. Quarter based clustering will give more information to form strategy. As buying pattern of the customers are will be same for the quarters it will allow company to take in reference for future as well.

Quarter1: 1/2011 to 3/2011
Quarter2: 4/2011 to 6/2011
Quarter3: 7/2011 to 9/2011
Quarter4: 10/2011 to 12/2011

To understand relationship between clusters scatterplot is used. Scatterplot also indicate potential number of outliers which we need to remove from data. It also indicates trend of given data with relationship with each other's.

7.1.1 Scatterplot Matrix Before Outlier Removal:

Quarter 1:



Quarter 2:

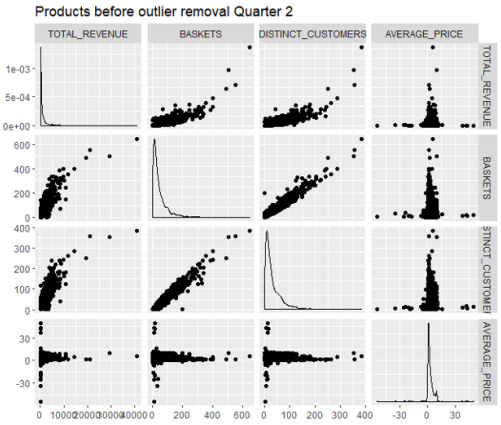
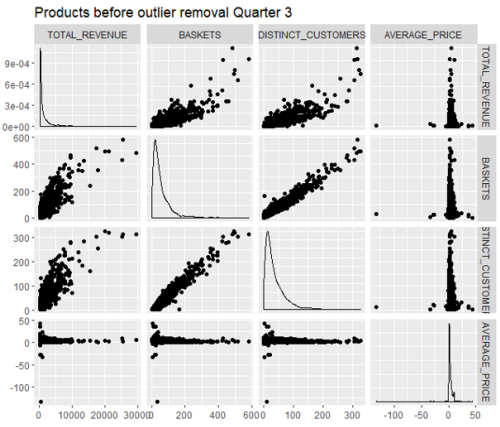


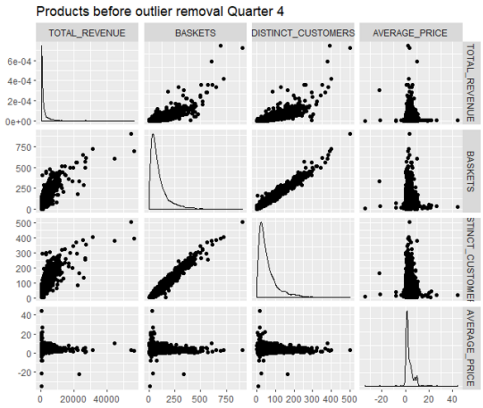
Figure 1 Scatter plot before outlier removal for Quarter 1 & 2

Quarter 3:



Quarter 3 Scatterplot Before Outlier Removal

Quarter 4:



Quarter 4 Scatterplot Before Outlier Removal

Figure 2 Scatter plot before outlier removal for Quarter 3 & 4

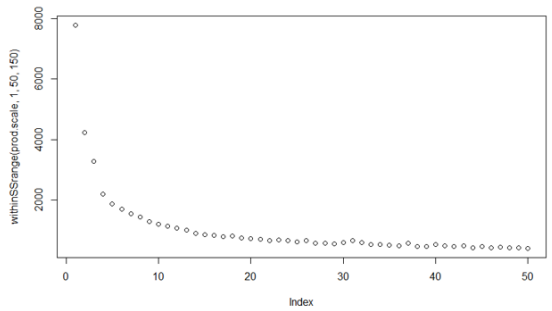
7.1.2 Number of Clusters:

It is important to identify optimal number of clusteres which can give us valuable information about the each segement as well as it will have less total error. Total distance of cluster center and data points should be less and distance between clusters sould be greater.Large number of cluster do not give idea about population and small number of clusters contains scattered data point. So, to identify optimal number of clusters in given data there are certain methods can be used from which number of clusters can be identify.

1. **Elbow Function:**
   Below elbow function is plot of **Cluster No Vs within sum of squared errors.** Function is converging to minimum value of within SSE. Optimal point lies between curve where function get diverge. As shown in below optimal number of clusters falls into the given circle.
2. **Silhouette Score:**
   Silhouette score is another measurement to decide optimal number of clusters in data. It measures how much similar a given data point is in its own cluster and compare to another cluster. It ranges from -1 to +1. High silhouette score indicates that data are matched in given clusters.

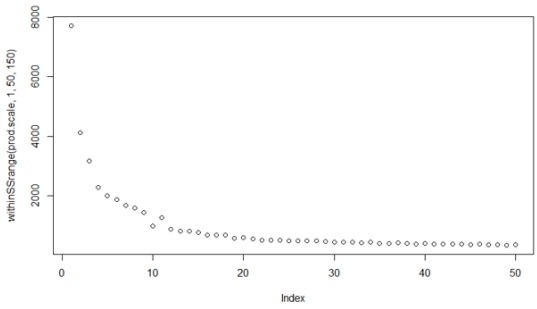
Elbow Function Plot:

Quarter 1:



Elbow function for quarter 1

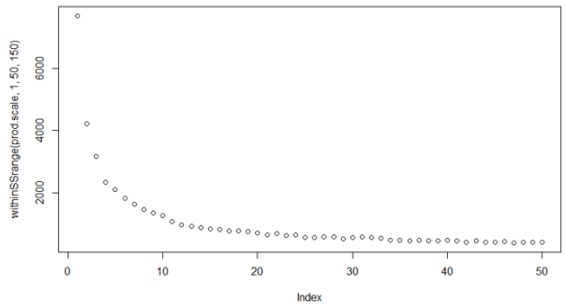
Quarter 2:



Elbow function for quarter 2

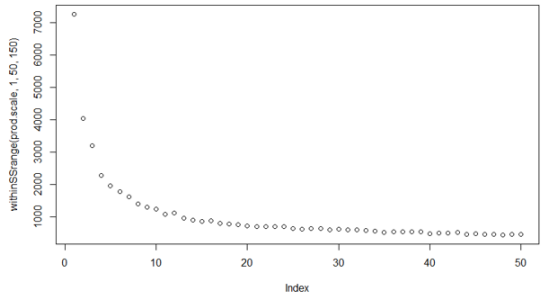
Figure 3 Elbow Point for Quarter 1 & 2

Quarter 3:



Elbow function for quarter 3

Quarter 4:



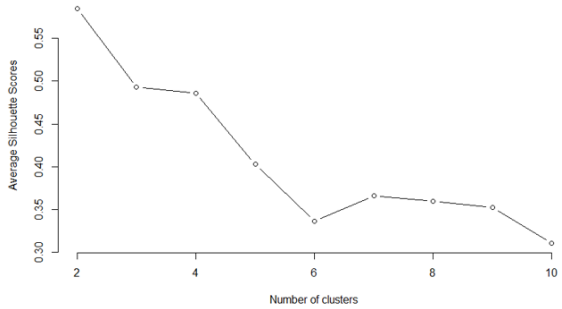
Elbow function for quarter 4

Figure 4 Elbow Point for Quarter 1 & 2

Silhouette Score:

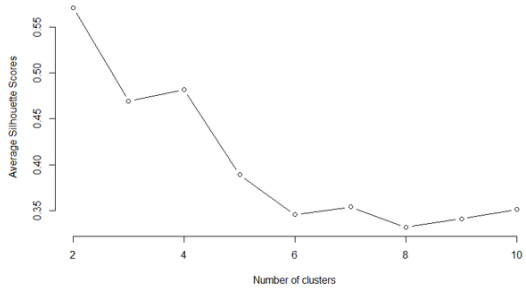
Here silhouette score is used to support the decision for the elbow function.

Quarter 1:



Silhouette Score for quarter 1

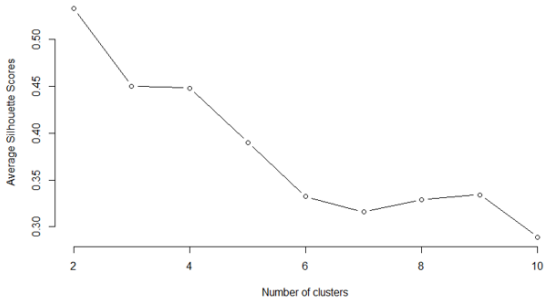
Quarter 2:



Silhouette Score for quarter 2

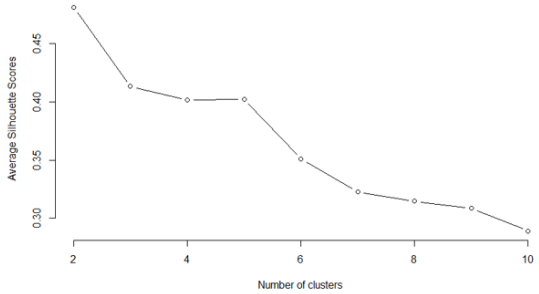
Figure 5 Silhouette Point for Quarter 3 &4

Quarter 3:



Silhouette Score for quarter 3

Quarter 4:



Silhouette Score for quarter 1

Figure 6 Silhouette Score for Quarte 3 & 4

7.1.3 Clusters for Quarters:

Steps are performed for clustering:

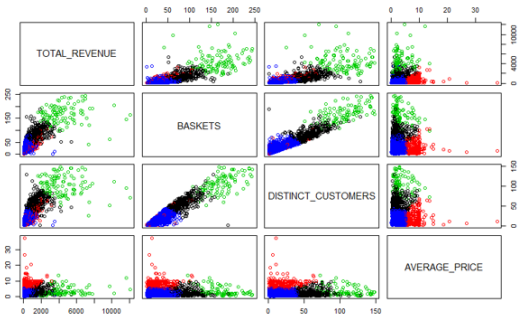
1. **Scaling**: As we have multiple attributes and all attributes are in different range of values, scaling the data is necessary. We can scale the data using multiple functions rather we used scale function to get all values into one scale.

2. **K-means Clustering**: Based on elbow function and silhouette score number of clusters are decided and scaled data are passed to the k-mean algorithm to form clusters.

3. **Un-scaling:** After clustering data is un-scale (DE normalize the data). Then data is bind to the clusters with cleaned data and plotting the clusters give the below results.

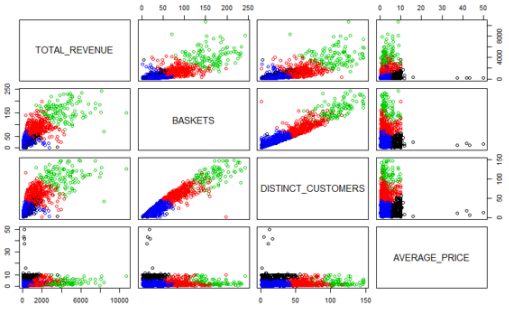
Scatterplot After Clustering:

Quarter 1:



Scatterplot after clustering for quarter 1

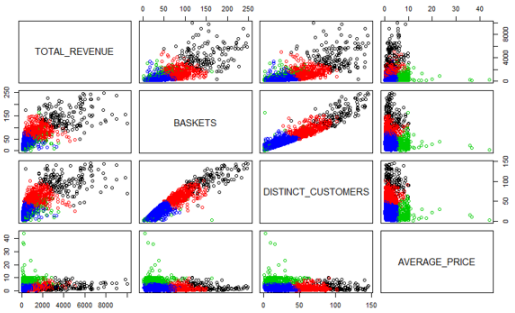
Quarter 2:



Scatterplot after clustering for quarter 2

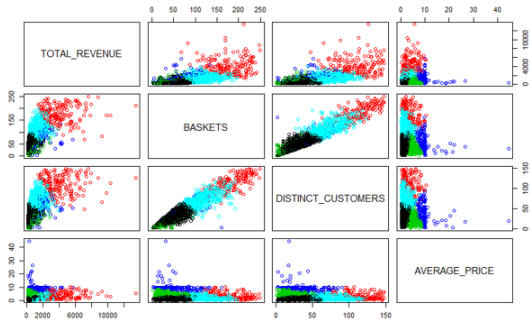
Figure 7 Scatter plot after clustering for quarter 1 & 2

Quarter 3:



Scatterplot after clustering for quarter 3

Quarter 4:



Scatterplot after clustering for quarter 4

Figure 8 Scatter plot after clustering for quarter 3 & 4

Number Of cluster Quarters:

Quarter 1: **4**

Quarter 2: **4**

Quarter 3: **4**

Quarter 4: **5**

7.2 Customer Clusters:

Based on the Feature selection, customer clustering has been done, Initially in Linear regression statistical modelling we try to analyze and visualize the correlation between 2 numeric variables(Bivariate Variables).This relation is visualized using scatterplot.

7.2.1 Customer cluster before Outlier Removal:

As given in the below image to make cluster more accurate and centered, outlier needs to be removed from the data. Outlier removal is crucial as otherwise clusters will be biased to the one side.

Figure 9 Scatterplot matrix for customers before outlier removal

7.2.2 Customers After Outlier Removal

As we have a picture of outliers in above scatter plot, we moved the file to open refine to remove outliers which might make cluster biased and inaccurate. In Open refine we changed all attributes to numeric before creating numeric facets for every attribute. Based on the distribution of data overall in the facet we got 1839 values which we thought will be better data for clustering and analysis. The below scatter plot will depict you the better picture after outlier removal.

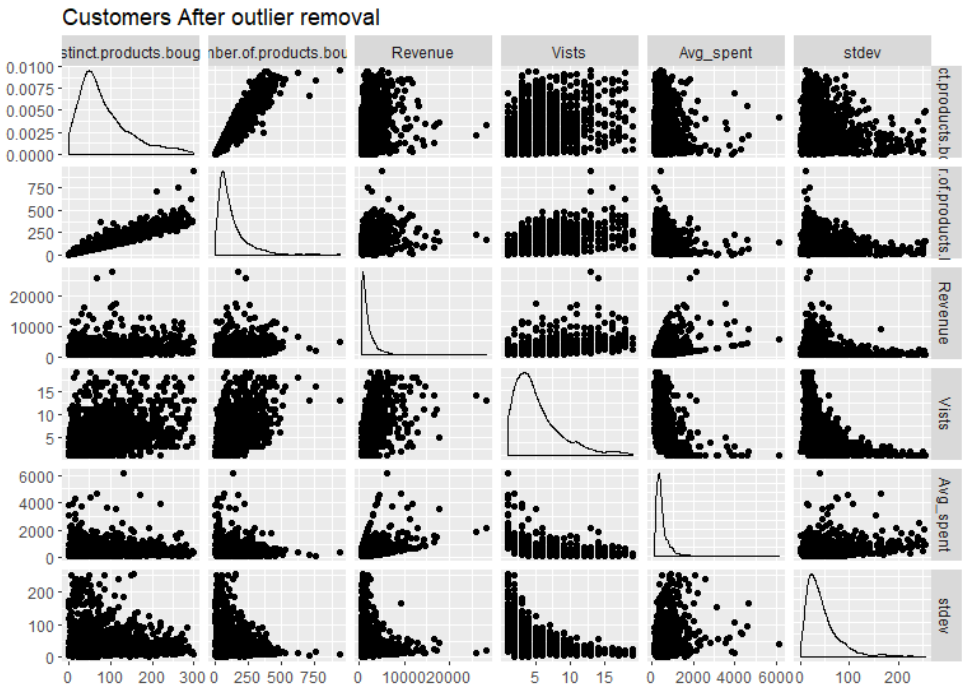


Figure 10 Scatterplot matrix for customers after outlier removal

It is important to identify optimal number of clusteres which can give us a idea about the population as well as it will have less total error. So, to identify number of clusters in given data there are certain methods can be used from which number of clusters can be identify.

Elbow Function:

Silhouette Score:

1. Elbow Function:

Below elbow function is plot of **Cluster No Vs within sum of squared errors.** Function is converging to minimum value of within SSE. Optimal point lies between curve where function get diverge. As shown in below optimal number of clusters falls into the given circle.

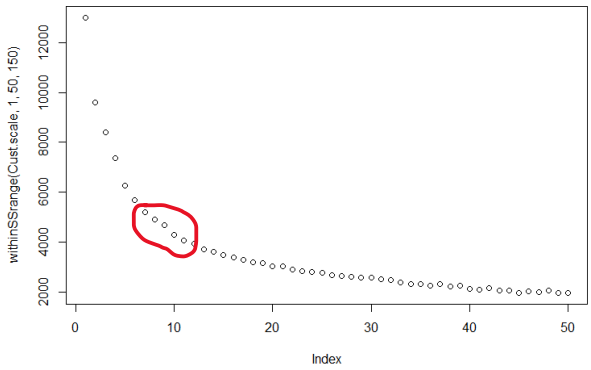


Figure 11 Elbow function for customer features

2. Silhouette Score:

Silhouette score is also another method to identify number of clusters in the data. Here silhouette score to support elbow function method. silhouette score is clearly indicating 5 number of clusters for the customers.

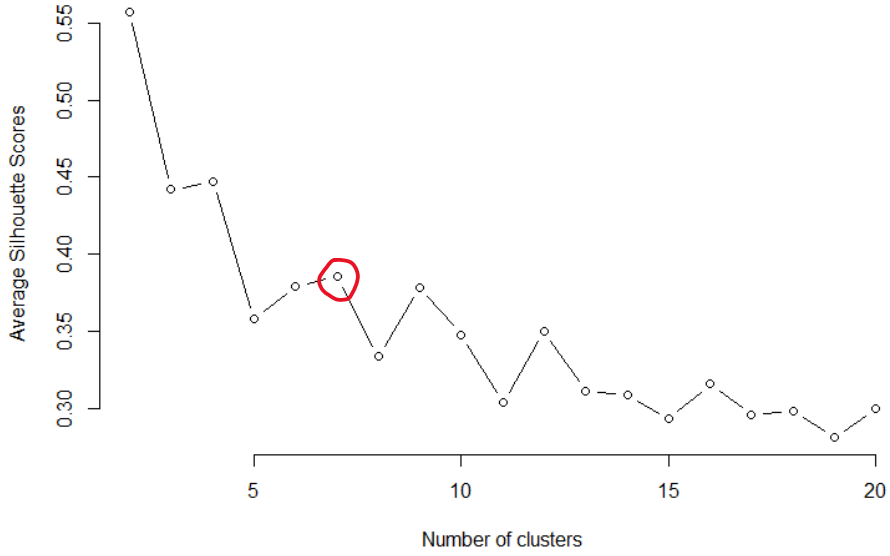


Figure 12 Scatterplot matrix for customers before outlier removal

7.2.3 Clusters:

As we have multiple attributes and all attributes are in different range of values, scaling the data is necessary. We can scale the data using multiple functions rather we used scale function to get all values into one scale. After performing K-means clustering algorithm on the data based on the elbow point and silhouette co-efficient. Then, unscaled the data (de-normalize the data). Then binding the clusters with cleaned data and plotting the clusters give the below results.

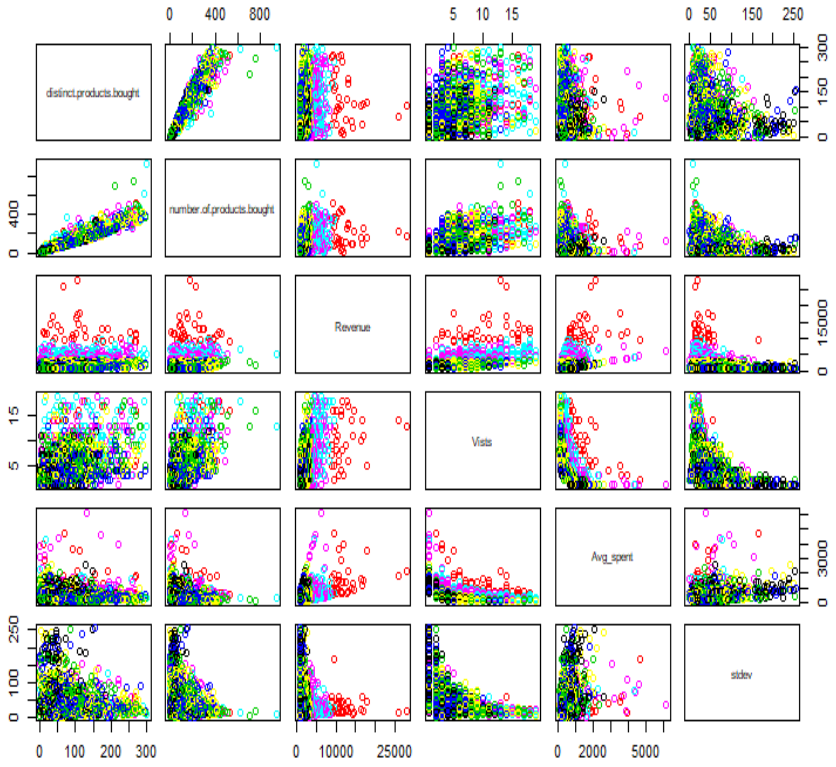


Figure 13 Scatterplot matrix for customers before outlier removal

7.2.4 Principal Component Analysis:

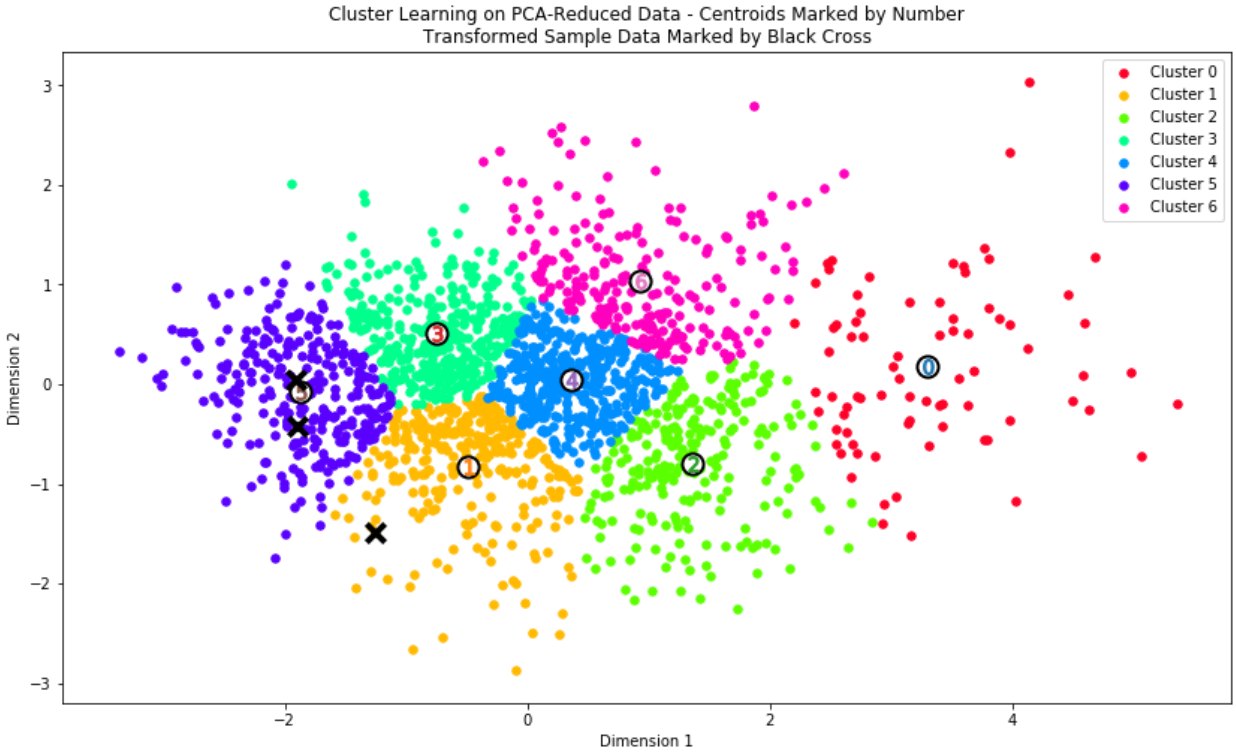


Figure 14 Scatterplot matrix for customers before outlier removal

Principal Component Analysis is used to reduce the dimension of the data to visualize the data in 2D. In PCA, we reduce the dimension by taking the meaningful features from the different features and merging them in order to get the maximum variance of that data. Thus, we reduced the 5-dimension data to two dimensions, and plotted it using Python Library Sci-Kit Learn. We ran K-means on the data set and formed seven cluster by finding the optimal clusters using Silhouette Score. Here we get the plotted graph and can clearly see the data having seven different clusters formed in different colors and hence can derive a lot of information from this graph.

8. Cluster Profiling:

Based on RFM model there are some measurements which needs to be considered to decide clusters characteristics. Creating profile demand curves for items based on comparable products and product categories is a powerful way to set the level of expected future sales and minimize the risk involved in launching a new product.

8.1 Customer Clusters:

Customer Cluster profiling is done based on the cluster centers and observing the sizes of cluster, ranking, segmentation and interpretation has been done. Below are the results for each cluster. All these ranking of clusters is done based on the RFM model. Recency, Frequency and Monetary.

Purchase frequency is calculated by the number of visits, also re verified it by calculating Purchase frequency with formula:

1. Purchase Frequency: Number of Orders/ Number of Unique Customers
2. Time Between Purchase: 365 days/ Purchase Frequency

Cluster Segment	Description	Recommendation
Cluster1(Big Spenders)	• Spend the most in less visits • Frequency of visits is less • Standard Deviation for visits is comparatively more	Market your most expensive customers.
Cluster2(Lost Cheap Customers)	• Last Purchased Long ago • Purchased few • Spent little	Don't spend too much trying to re-acquire
Cluster3(Loyal Customers)	• Bought most frequently (Visits) • Giving good revenue • Most recently bought	Provide discounts on daily consumed products
Cluster4(Big Customers)	• Recently Bought customers • More frequently visiting customers • High revenue generators	Continue Monitoring
Cluster5(Big Spenders)	• Spend the most in less visits • Frequency of visits is less • Standard Deviation for visits is comparatively more	Market your most expensive customers.
Cluster6(Loyal Customers)	• Bought most frequently (Visits) • Giving good revenue • Most recently bought	Provide discounts on daily consumed products
Cluster7(Almost Lost)	• Haven't purchased for some time • But purchased frequently and spend the most	Aggressive price discount

Table 3 Customer Cluster Profiling

8.1.1 Visualization of attributes per cluster:

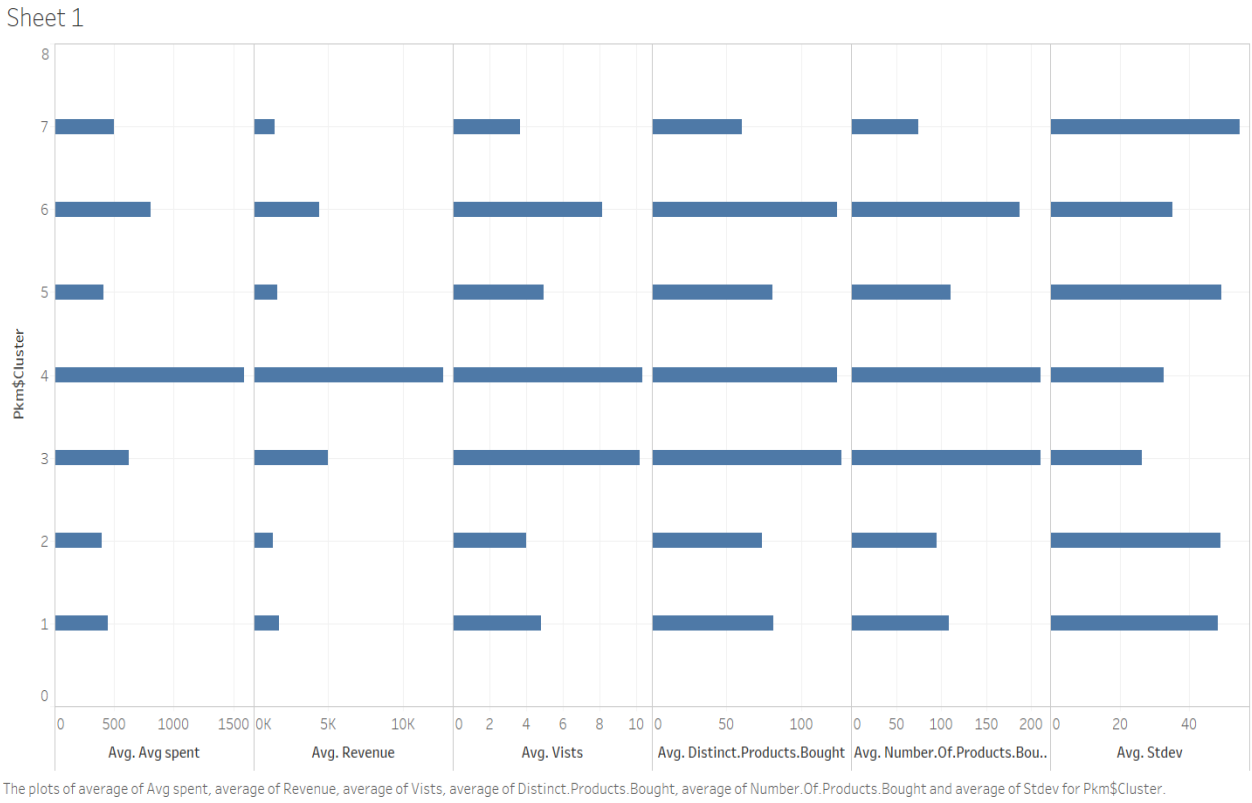


Figure 15 Bar Chart of Customer Cluster

8.2 Product Cluster:

Product Cluster profiling is done based on the cluster centers and observing the sizes of each cluster. Ranking, segmentation and interpretation has been done. Below are the results for each cluster. All these ranking of clusters is done based on the Monetary and Frequency labels.

As mentioned earlier products are segregated into quarters and cluster profiling and interpretation has been done based on quarters. The analysis as per quarter can give you how online store products are getting sold quarterly, there by inventory can be made available and strategies can be planed.

Quarter -1:

Cluster Segment	Description	Cluster Type
Cluster -1	• Second Highest Baskets size • Highest Revenue • Low average price	Highest potential
Cluster -2	• Highest average price • Lowest revenue • Less Baskets size	Discount should be given

Cluster -3	• Second Highest Revenue • Maximum basket size • Medium average price	Main attraction
Cluster- 4	• Medium revenue • Least average price • Least distinct customer	Daily usage

Table 4 Quarter 1 Product Profiling

Quarter -2:

Cluster Segment	Description	Cluster Type
Cluster -1	• Highest average price • Lowest revenue • Least Baskets size	Discount should be given
Cluster -2	• Second Highest Baskets size • Highest Revenue • Low average price	Highest potential
Cluster -3	• Second Highest Revenue • Maximum basket size • Medium average price	Main attraction
Cluster- 4	• Medium revenue • Least average price • Least distinct customer	Daily usage

Table 5 Quarter 2 P roduct Profiling

Quarter 3:

Cluster Segment	Description	Cluster Type
Cluster -1	• Second Highest Revenue • Maximum basket size • Medium average price	Main attraction
Cluster -2	• Second Highest Baskets size • Second Highest Revenue • Low average price	Highest potential
Cluster -3	• Highest average price • Lowest revenue • Less baskets size	Discount should be given
Cluster- 4	• Medium revenue • Least average price • Least distinct customer	Daily usage

Table 6 Quarter 3 Product Profiling

Quarter 4:

Cluster Segment	Description	Cluster Type
Cluster -1	• Medium revenue • Least average price • Least distinct customer	Daily usage
Cluster -2	• Second Highest Revenue • Maximum basket size • Medium average price	Main attraction
Cluster -3	• Second Highest Baskets size • Second Highest Revenue • Low average price	High potential
Cluster- 4	• Highest average price • Lowest revenue • Less baskets size	Discount Should be given
Cluster -5	• Second highest revenue • Second least average price • Average basket size	Best price item model

Table 7 Quarter 4Product Profiling

Recommendation Table for Products:

Cluster Category	Recommendation
High Sale Products	Especially for this category Stocking and inventory management will be helpful
Main attraction	More of these category products should be there and it should be more advertise.
High potential	As it has less basket size some business decision is required to sell more products.
Bargain Products	Discount Should be given to increase sell
Best price items model	Monitoring is required to increase more revenue

Table 8 Recommendation Table for Product Cluster

8.2.1 Visualization of attributes per cluster:

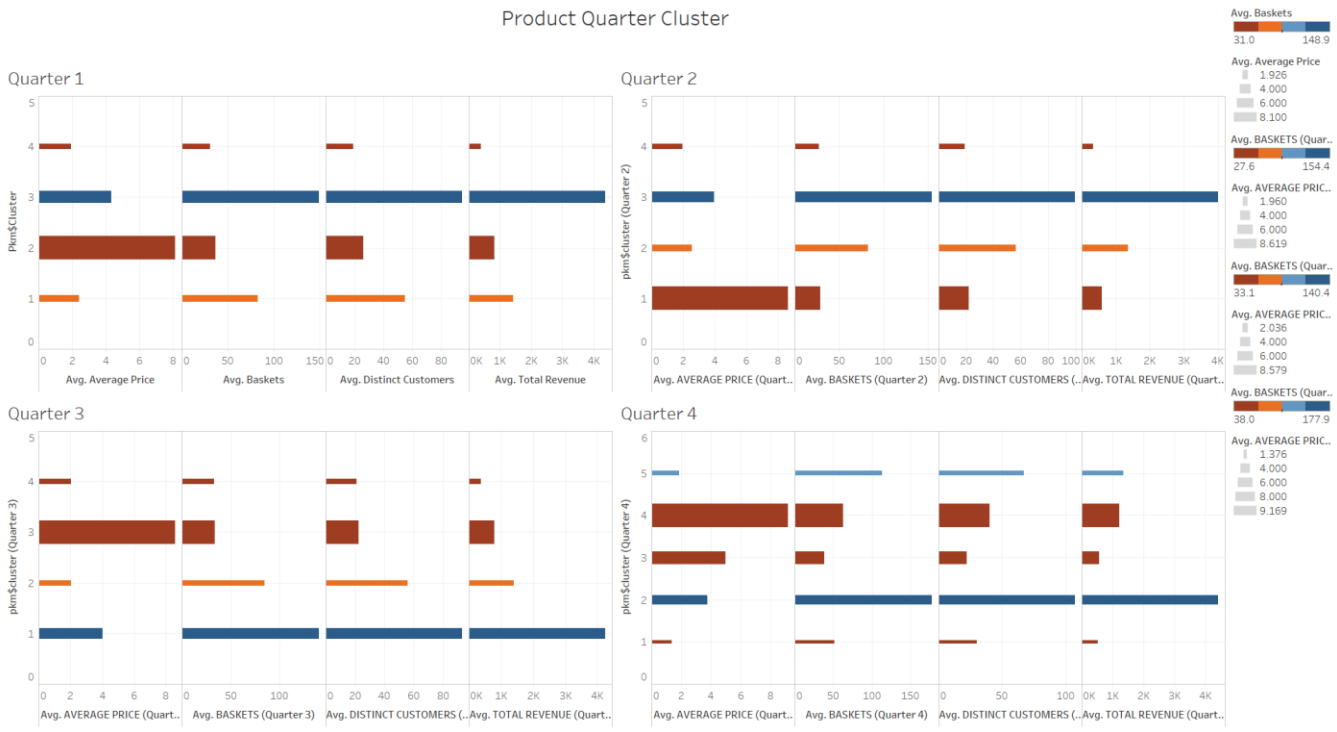


Figure 16 Bar Chart of Product Cluster

9. Conclusion and Next steps:

To conclude, above analysis can be help for driving business decisions for online selling. Further analysis can be performed to find association between products and much more. Which can eventually increase revenue and will lead to generate more profit.

Further, the customer purchase behavior has been interpreted by the purchase frequency and time between purchases.

Purchase Frequency to analyze customer behavior:

```
Purchase Freq. for Cluster1 4.837037037037037
Purchase Freq. for Cluster2 3.997167138810198
Purchase Freq. for Cluster3 10.221374045801527
Purchase Freq. for Cluster4 10.352941176470589
Purchase Freq. for Cluster5 4.9634464751958225
Purchase Freq. for Cluster6 8.148809523809524
Purchase Freq. for Cluster7 3.6666666666666665
```

Figure 17 Purchase Frequency of Clusters

Time between Purchases:

```
Time Between Purchases for Cluster1 75.4594180704441
Time Between Purchases for Cluster2 91.31467044649186
Time Between Purchases for Cluster3 35.709484690067214
Time Between Purchases for Cluster4 35.25568181818181
Time Between Purchases for Cluster5 73.53761178327196
Time Between Purchases for Cluster6 44.7918188458729
Time Between Purchases for Cluster7 99.54545454545455
```

Figure 18 Time Between Purchases for different Clusters

9.1 Decision Tree Classification on Cluster Profiling

For Next steps, we performed Decision tree Classification on the results of the Products Quarter Data. Trained the decision tree model on the quarter data, labels as our classification of clusters like High Potential and other 4 categories. We then used the trained model to test on our raw Product data and get the results of exactly which category the products belong to.

```
n= 17

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 17 13 1 (0.24 0.24 0.24 0.24 0.059)
   2) BASKETS=149.3,160.6,165.8,177.1 4  0 2 (0 1 0 0 0) *
   3) BASKETS=113.9,22.5,22.6,25.7,26.7,28.7,29.0,33.2,47.0,53.9,81.7,81.9,84.0 13  9 1 (0.31 0 0.31 0.31 0.077)
     6) AVERAGE_PRICE=1.4,1.8,1.9,8.6,8.7,8.8,9.9 9  5 1 (0.44 0 0 0.44 0.11)
      12) TOTAL_REVENUE=278.1,291.8,354.4,487.7 4  0 1 (1 0 0 0 0) *
      13) TOTAL_REVENUE=1252.4,1388.6,463.1,544.6,686.5 5  1 4 (0 0 0 0.8 0.2)
        26) TOTAL_REVENUE=1252.4,463.1,544.6,686.5 4  0 4 (0 0 0 1 0) *
        27) TOTAL_REVENUE=1388.6 1  0 5 (0 0 0 0 1) *
     7) AVERAGE_PRICE=2.2,2.4,2.6,4.9 4  0 3 (0 0 1 0 0) *
```

Figure 19 Decision Tree Nodes

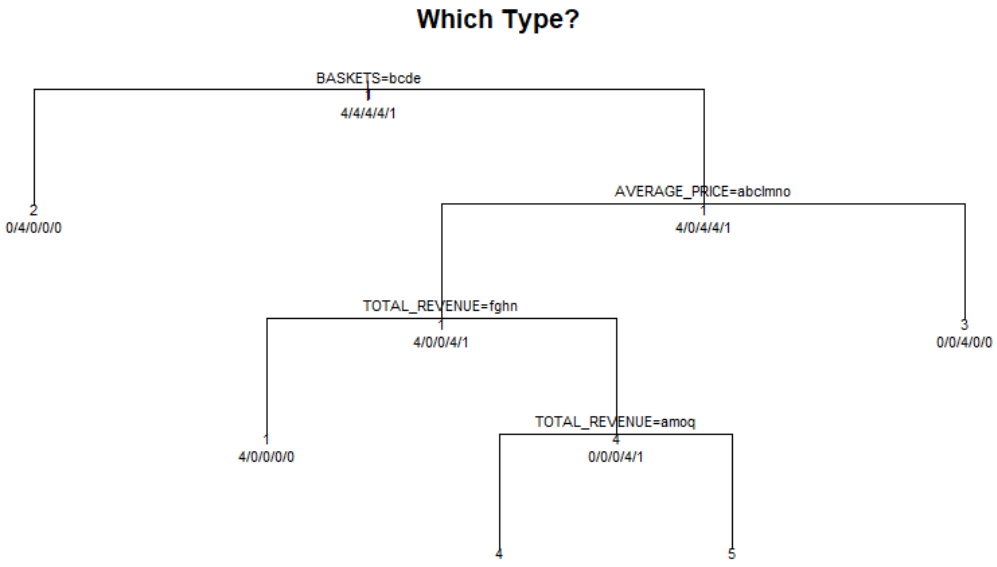


Figure 20 Decision Tree Structure of Training for Product – Quarter wise Data

10. References:

Donnelly, K. (2017). *How to Calculate Customer Lifetime Value*. [online] Shopify. Available at: https://www.shopify.ca/blog/customer-lifetime-value [Accessed 24 May 2019].

Burkard, K. (2018). *11 Key Retention Metrics You Need to Know*. [online] Blog.smile.io. Available at: https://blog.smile.io/retention-metrics-you-need-to-know [Accessed 24 May 2019].

McEachern, A. (2018). *How to Calculate Purchase Frequency, and 3 Tips to Improve It!*. [online] Blog.smile.io. Available at: https://blog.smile.io/how-to-calculate-purchase-frequency [Accessed 24 May 2019].

Correia, J. (2016). *How RFM Analysis Boosts Sales | Blast Analytics & Marketing*. [online] Blast Analytics & Marketing. Available at: https://www.blastam.com/blog/rfm-analysis-boosts-sales [Accessed 24 May 2019].

Scikit-learn.org. (n.d.). *sklearn.decomposition.PCA — scikit-learn 0.21.2 documentation*. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html [Accessed 24 May 2019].

En.wikipedia.org. (n.d). *Silhouette (clustering)*. [online] Available at: https://en.wikipedia.org/wiki/Silhouette_(clustering) [Accessed 24 May 2019].