“A study on online retail data set to understand the characteristics of customer segments that are associated with the business”

Author: Shashank Badre (M12383328)
First Reader: Professor Peng Wang
Second Reader: Professor Yichen Qin

Business Analytics Program
University of Cincinnati, College of Business

Table of Contents

List of Figures

List of Tables

Summary

Online retailers in the world who happen to have a small business and are new entrants in the market are keen on using data mining and business intelligence techniques to better understand existing and potential customer base. However, such small businesses often lack expertise and technical know how to perform requisite analysis. This study will help such online retailers to understand the approach and different ways the data can be utilized to gain insights into its customer base. This study is done on an online retail data set to understand characteristics of different segments of customers. Based on these characteristics the study will explain which customers segments contribute high monetary value and which customer segments contribute low monetary value to the business.

Characteristics of different customer segments have been analyzed using a combination of Recency, Frequency and Monetary (RFM) value analysis, Principal Component Analysis, K means, Kernel K means and Hierarchical clustering techniques. Combination of RFM analysis, Principal Component Analysis and K means clustering technique has produced the best results for this study.

It has been found that there are 6 distinct customer segments in the data set. Characteristics of these segments are discussed in detail in the study. However, some important findings with respect to these customer segments are given below.

- 17 customers in **cluster 2** are high value customers
- 2 customers in **cluster 4** have potential to be high value customers in future
- 79 customers in **cluster 3** and 1416 customers in **cluster 1** are moderate value customers
- Customers in **cluster 5** are low value active customers and customers in **cluster 6** are low value inactive customers

1. Introduction

Retailers who sell their products online are often concerned with the following questions.

- Who are the most profitable customers?
- Who are the least profitable customers?
- What different types of customers do the online retailer has?
- What are the characteristics of customers?
- Which customers buy frequently on retailer's website?
- Understanding most loyal and least loyal customers

This study aims at solving above questions faced by small online retailers. The study shows how data mining and business intelligence techniques can be used to generate insights on customer segments, revenue contribution etc.

For this study, we have an Online retail business in United Kingdom which sells its products on Amazon. It has collected data for almost over a year and stored all the transactions that occurred in the later part of 2010 and for a complete duration of 2011. Majority of its customers are in United Kingdom and hence for this study I have considered only UK customers for customer segmentation and further analysis.

Dataset source: https://archive.ics.uci.edu/ml/datasets/Online+Retail#

2. Methodology Adopted

To address the questions faced by small online retailers, following Methodology has been adopted.

- Calculate recency, frequency and monetary value for each customer ID
- RFM analysis, Principal Component Analysis, K-means clustering, Kernel K-means clustering, and Hierarchical clustering technique has been used to find out clusters in the data set
- Using results from the clustering approach, customers will be segmented, and revenue contributed by each segment of customers will be measured

Using clustering techniques, we will be able to perform customer segmentation. By measuring revenue contributed by each segment of customers, we will be able to understand the characteristics of the most and least profitable customers. This study will eventually allow online retailers to understand attributes for different segment of customers so that it can target these customers by using personalized marketing campaigns.

Data Structure

- Online retailer data set is a transactional data that records all transactions occurred in between 01/12/2010 and 09/12/2011
- There are 541909 observations and 8 variables in the data set

The detailed structure of the data set is given below.

InvoiceNo – A character variable of 6-digit nominal number uniquely assigned to each transaction

StockCode - A character variable of 5-digit nominal number uniquely assigned to each product

Description - This is a character variable and it shows description of a product

Quantity - This is a character variable and it shows quantity of a product per transaction

InvoiceDate - This is a datetime variable. It shows the date of the transaction

UnitPrice - This is numeric variable. It shows the unit price of a product in a transaction

CustomerID - A character variable of 5-digit nominal number uniquely assigned to each customer

Country - This is a character variable. It shows the name of the country where each customer resides

3. Data Pre-Processing

Online retail raw data set consists of 541909 observations and 8 variables. I have already mentioned variables in the data structure.

- Delete all canceled transaction records. Invoice no. that starts with "C" denote canceled transaction records
- Some observations in Description contains "?" and does not provide accurate information and such observations are not considered for this study
- 1454 entries in Description and 135080 entries in Customer ID variable are null
- For this study, observations with complete information are taken for further analysis
- Many entries in variable Quantity contain negative values and such observations are omitted for the analysis
- Format for Invoice date is set to YYYY-MM-DD
- Derived variable "Year of Purchase" is added to the data set and it contains year in which transaction was carried out by the customer
- Derived variable "Days Since" is added to the data set and the entries are the difference in days between transaction date and fixed date of "2012-01-01"
- 0.1% values in variables Quantity and Unit Price are treated as outliers and hence observations with quantity less than 504 and unit price less than 42 are considered for the analysis
- Derived variable "Total Price" is added to the data set and the values in this variable are obtained by multiplying unit price and quantity, thus it shows the total value for a transaction

	Quantity	UnitPrice	CustomerID	Days Since	Total Price
mean	12	3	15294	175	20
sd	26	3	1713	113	53
median	6	2	15159	154	12
minimum	1	0	12347	23	0
maximum	500	42	18287	396	3285

Table 1: Summary table for cleaned data set

4. Exploratory Data Analysis

Let us look at distribution of quantity and unit price variables in terms of median and quartiles.

Figure 1: Boxplots for Quantity and Unit Price variables

Interpretation of Box plots

- Boxplot for quantity shows that median number of quantity bought in a transaction is 6 and 75% quartile is 12
- Boxplot for Unit Price shows that median number of unit price for a transaction is 1.95 dollars and 75% quartile is 3.75 dollars

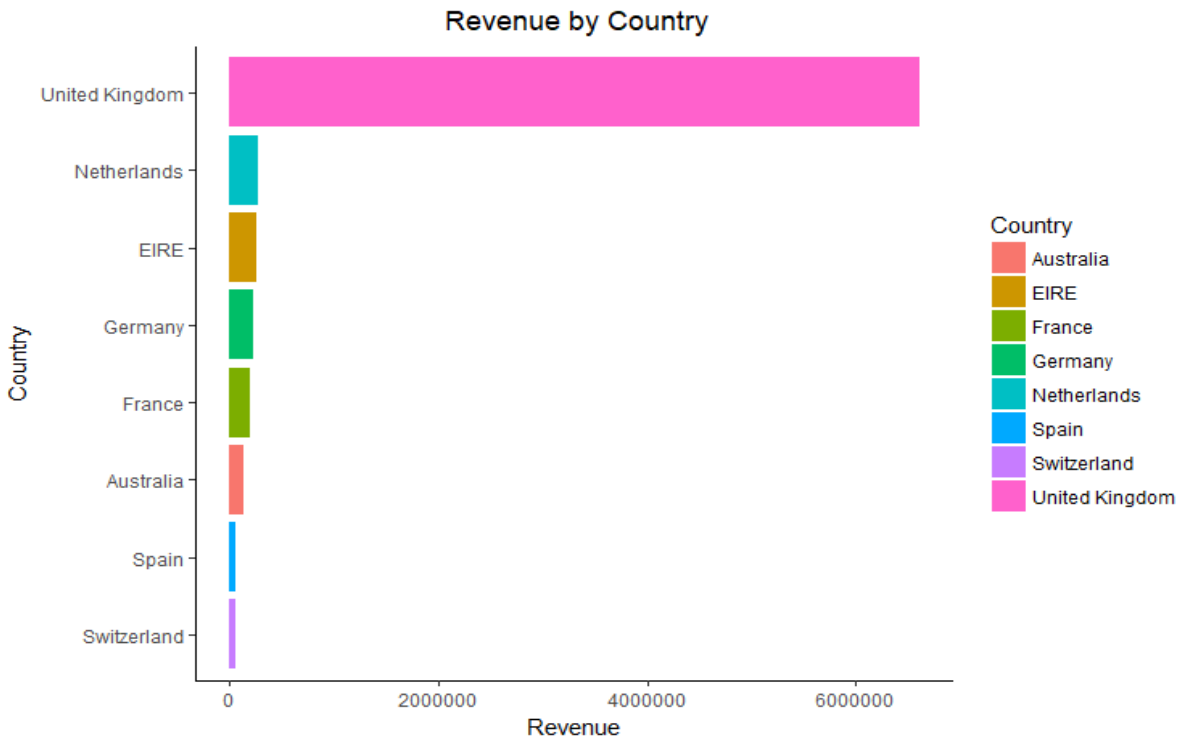

Figure 2: Revenue generated Country wise

It can be seen from the above bar chart that most of the revenue is generated by an online retailer from United Kingdom. Netherlands is at second position when it comes to revenue generation.

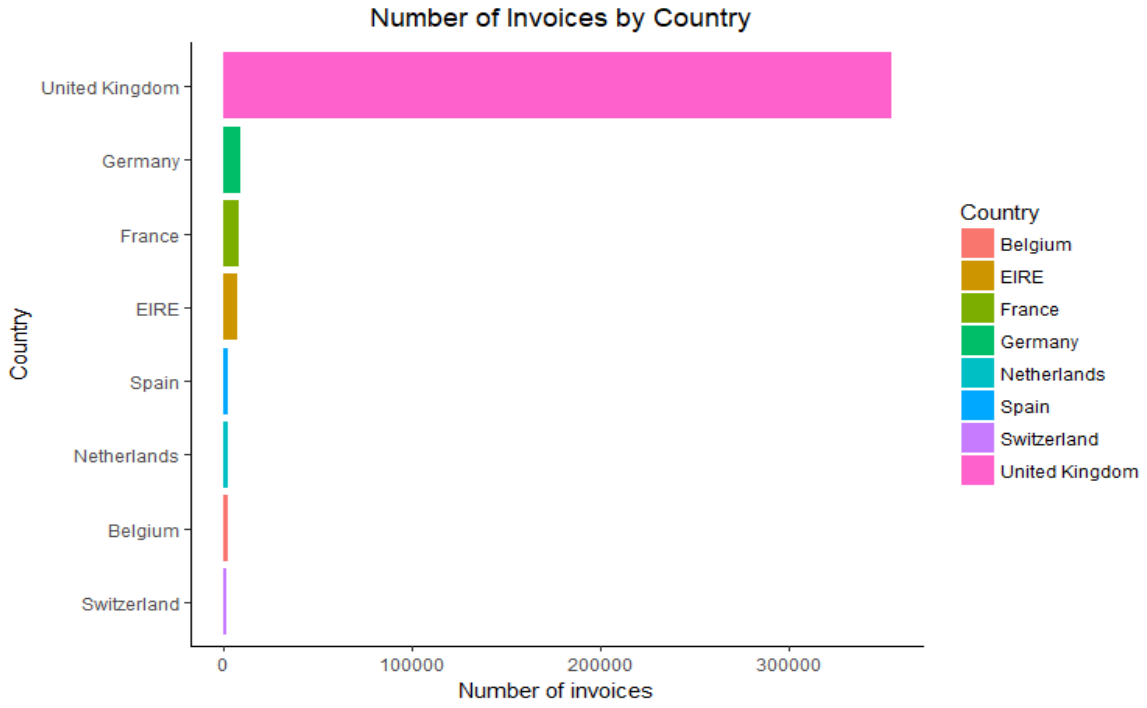

Figure 3: Number of Invoices Country wise

It can be seen from the above bar chart that United Kingdom is at top when it comes to number of invoices i.e. number of transactions. Germany is at second position when it comes to generating number of invoices.

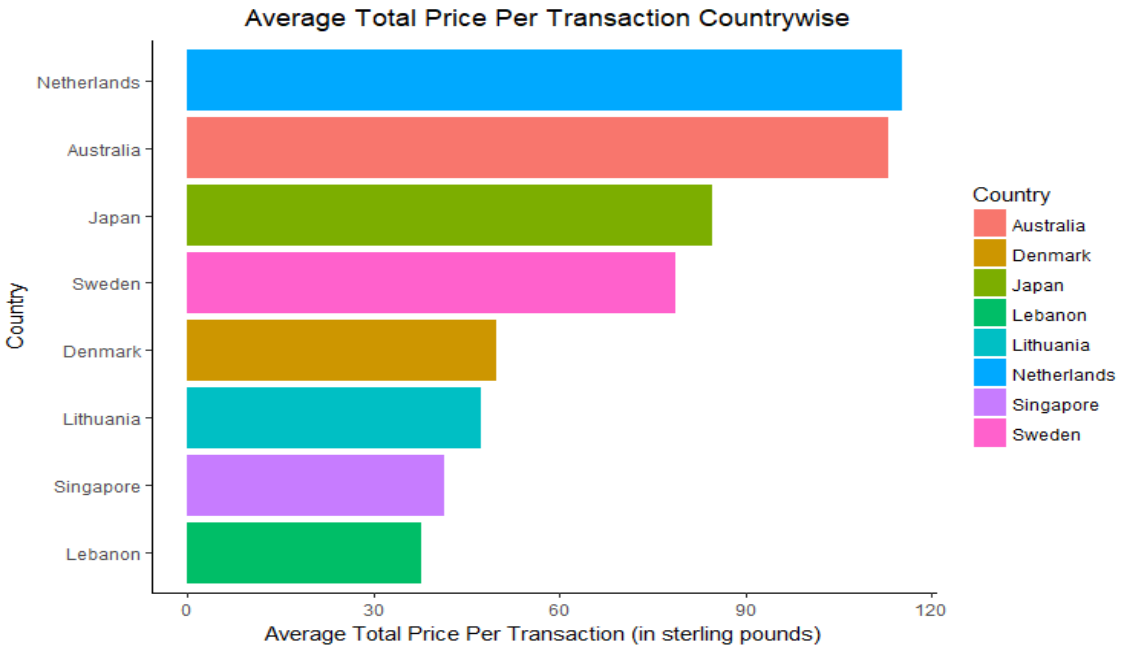

Figure 4: Average total price per transaction Country wise

It can be seen from the above bar chart that Netherland contributes the highest revenue when it comes to a average total price per transaction.

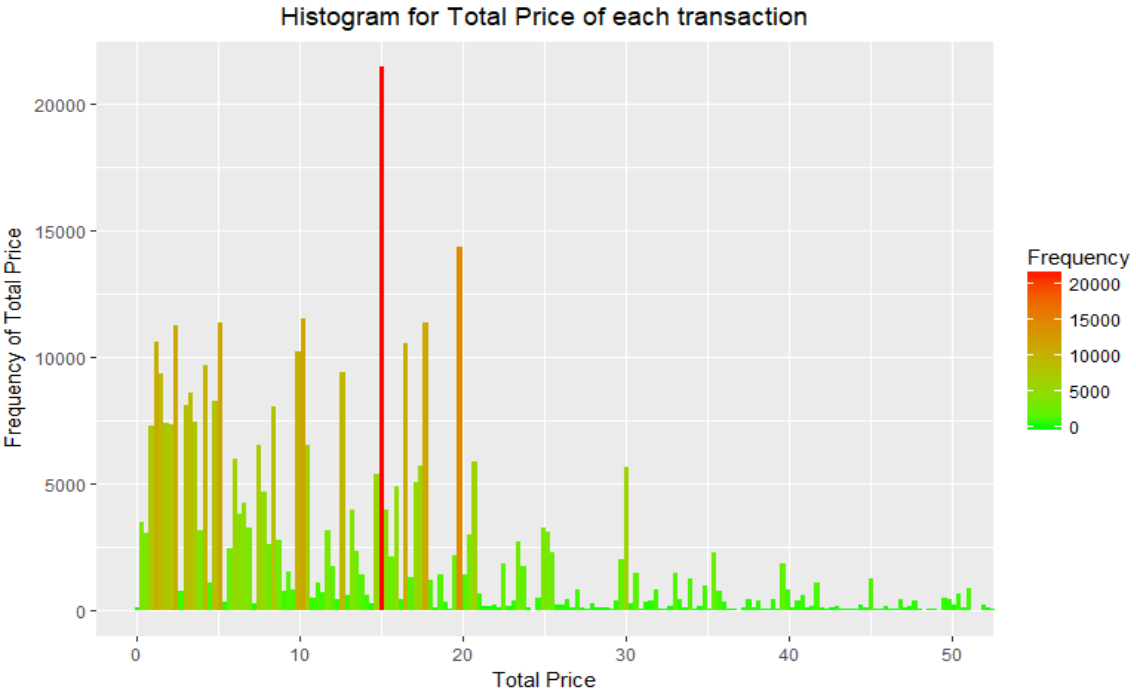

Figure 5: Distribution of total price for each transaction

Above distribution of total price for each transaction shows that frequency is highest for transaction that amounts to 15 sterling pounds of purchases.

4. Data Preparation for RFM analysis and Clustering

To understand different segments of customers, I have performed clustering. To perform clustering I have performed Recency, Frequency and Monetary analysis on cleaned data set to derive a new data set. RFM analysis is a technique used in marketing to determine quantitatively to understand the best customer segments by examining how recently a customer has transacted with the business i.e. Recency, how frequently the customer has been transacting with the business i.e. Frequency and how much revenue the customer has contributed to the business i.e. Monetary.

As stated above, we have the cleaned data set. Now, for RFM analysis, new derived data set is prepared in a following way.

- ***Customer ID*** is selected from the cleaned data set
- Difference in days in between recent transaction of a customer and fixed date '2012-01-01' is taken as ***Recency*** for the customer
- Difference in days in between first transaction of a customer and fixed date '2012-01-01' is taken as ***First purchase*** for the customer
- Number of a transactions done by a customer with an online retailer is taken as ***Frequency*** of the customer transactions with the business
- Sum of Total Price of all transactions of a customer is taken as ***Revenue Contribution*** by a customer to the online retail business
- Transaction with the minimum amount of all transactions of a customer with the business is taken as ***Minimum Purchase value***
- Transaction with the maximum amount of all transactions of a customer with the business is taken as ***Maximum Purchase value***
- Average amount of all transactions of a customer with the business is taken as ***Average Purchase value***

	Recency *(in days)*	First Purchase *(in days)*	Frequency *(in days)*	Revenue Contribution (£)	Minimum Purchase Value (£)	Maximum Purchase Value (£)	Average Purchase Value (£)
mean	115	246	91	1693	11	94	28
sd	99	118	218	5798	53	173	66
median	73	272	41	639	5	50	17
minimum	23	24	1	3	0	2	1
maximum	396	396	7847	228882	2003	3285	2125

Table 2: Summary of RFM analysis

4.1 K means clustering

K means clustering is a data mining technique used to perform cluster analysis. It aims to partition observations in a data set into individual clusters. Each observation belongs to a one single cluster with the nearest mean. All observations in a cluster exhibit almost similar characteristic.

To perform K means clustering scaling is very important as it ensures scales of each variable is same.

One important question while performing K-means clustering is to decide on how many clusters would be appropriate for to analyze for the business. Elbow method and Silhouette analysis on data set allows to understand the optimal number of clusters that are present in the data set.

For K means clustering, I have chosen Recency, Frequency, Revenue Contribution and Average Purchase value variables from the derived data set. Using elbow method and silhouette analysis it was found that the data set can be segmented in 4 different clusters.

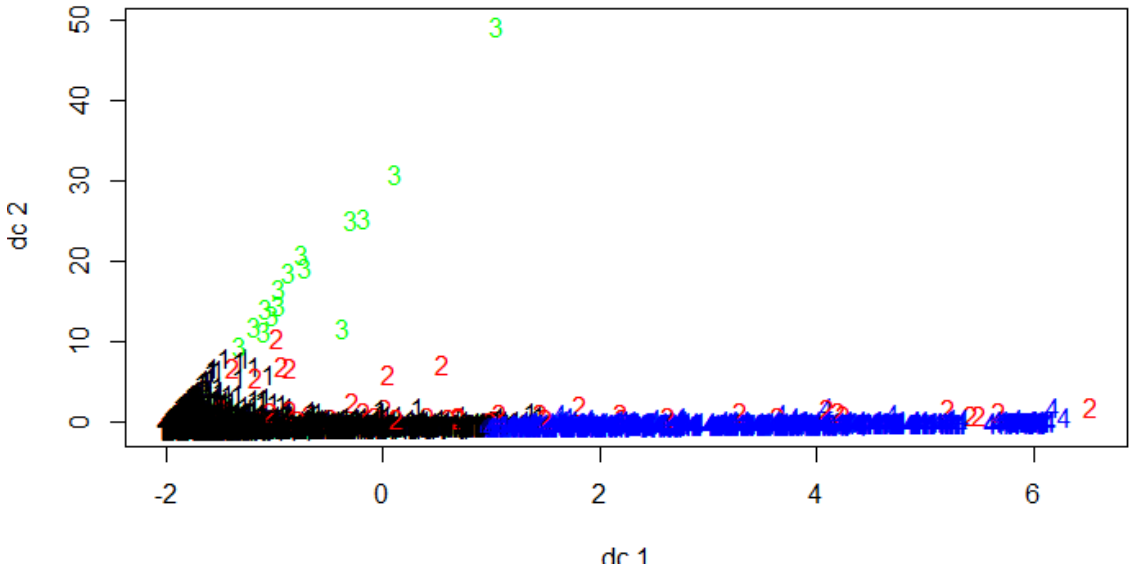

Figure 6: Clusters obtained in the data set using K means clustering approach

Interpretation of the K means clustering plot is given below.

- Above plot shows 4 different clusters plotted on x and y dimensions
- Cluster 1 and Cluster 4 are densely populated
- Cluster 1 is black in color, cluster 2 is red in color, cluster 3 is green and cluster 4 is blue in color
- Cluster 2 which is red in color seems to be fragmented in between clusters 1 and 4

Cluster	**Number of Observations**	**Recency *(in days)***	**Frequency *(in days)***	**Revenue Contribution (£)**	**Average Purchase Value (£)**
1	2873	63	103	1710	23
2	44	149	16	5715	447
3	15	29	1986	68653	130
4	969	267	28	425	23

Table 3: Characteristics of clusters using K means clustering

Interpretation of the above table is given below.

- Number of observations column shows number of records belonging to a cluster
- Recency column shows on an average *Recency* statistic for each customer belonging to a cluster (For example: On an average a customer in cluster 1 had transacted 63 days before the date of '2012-01-01')
- Frequency column shows on an average *Frequency* statistic for each customer belonging to a cluster (For example: On an average a customer in cluster 3 has 1986 transactions with the business)
- Revenue Contribution column shows on an average how much *revenue contributed* by a customer to the business belonging to a cluster (For example: On an average a customer in cluster 4 has contributed 68653 sterling pounds to the business)
- Average purchase column shows on an average the transaction amount for each transaction by a customer with the business (For example: On an average a customer in cluster 2 has transaction amount of 447 pounds for a single transaction with the online retailer)

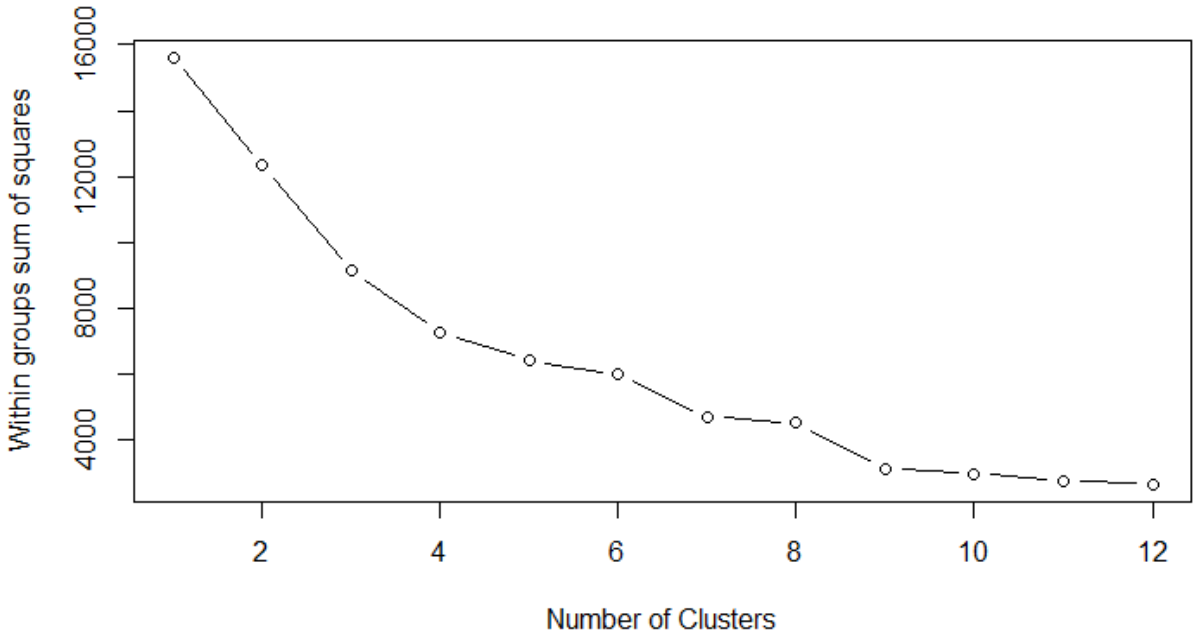

Figure 7: Elbow Method to select optimal number of clusters

Elbow plot shows optimal number of clusters as 4 to be selected. We can see the significant change in slope after cluster number 4.

Let us look at silhouette width to decide on choosing optimal number of clusters.

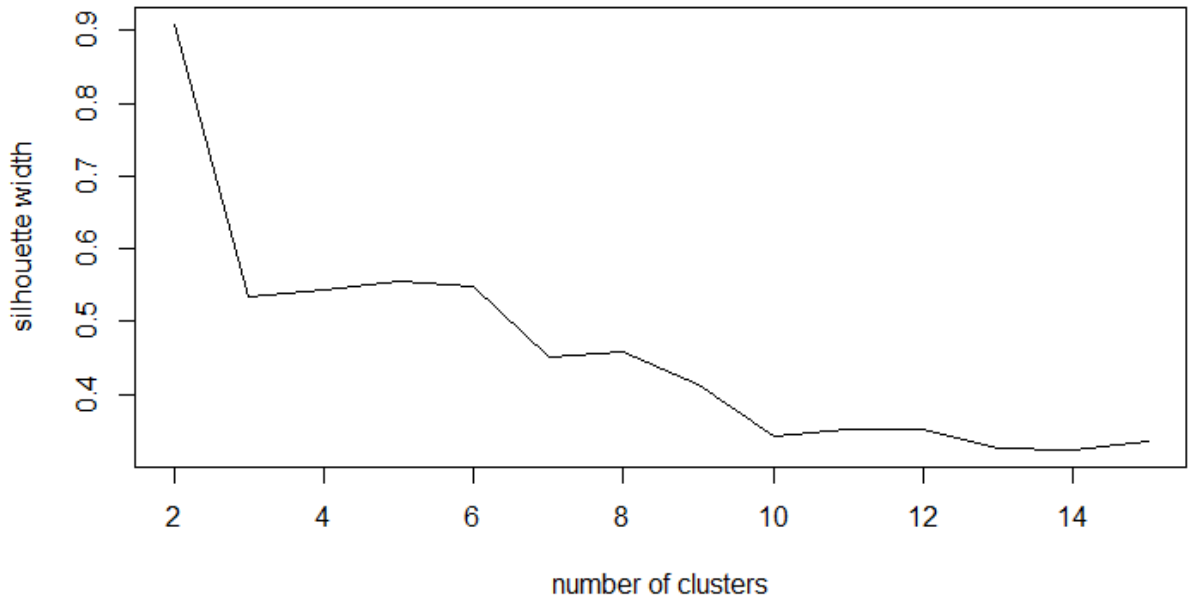

Figure 8: Silhouette width for K means clustering

Silhouette width plot shows maximum silhouette width at 2. However, silhouette width at 4, 5 and 6 are almost similar and one of these cluster number can be chosen for clustering the data set.

By analyzing elbow plot and silhouette width plots, it can be concluded that K means clustering on the data set performs best when the cut off cluster value is chosen at 4.

4.2 Hierarchical Clustering

Hierarchical clustering is a data mining technique used to perform cluster analysis. It aims to build a hierarchy of clusters. For this study I have used Agglomerative approach for performing hierarchical clustering. Agglomerative approach is a “bottom up” approach where each observation is treated as an individual cluster. These clusters are merged into a hierarchy depending on the distance in between these clusters. Thus, to perform hierarchical clustering we need to build distance matrix between observations first and then perform hierarchical clustering analysis on this distance matrix to create a hierarchy of clusters.

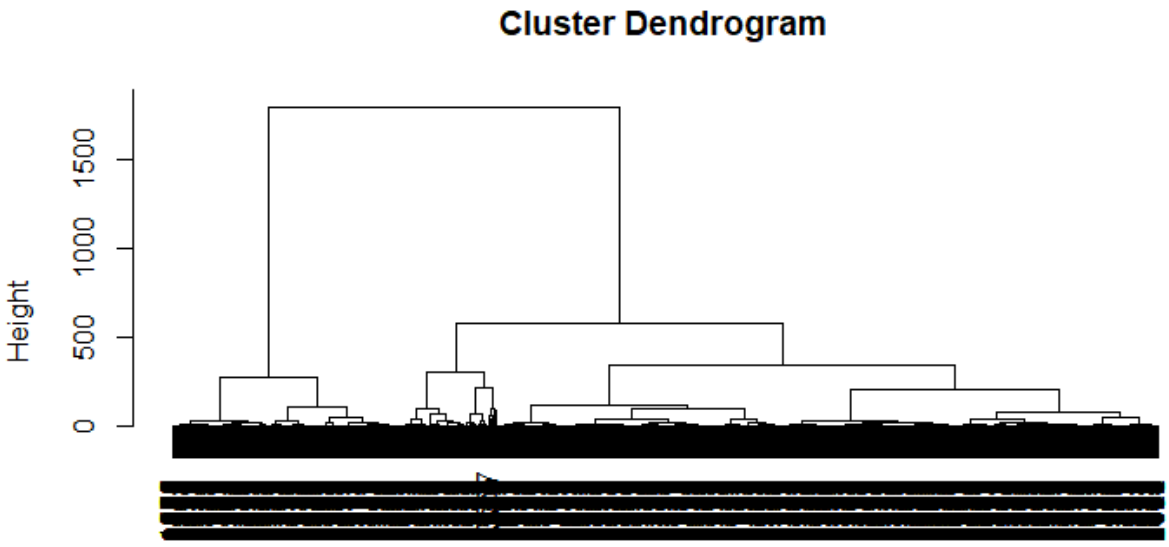

Figure 9: Hierarchy of clusters

Above plot is a cluster dendrogram. Dendrogram shows hierarchy of different clusters. Height shows the distance between the observations. To calculate distance between observations Ward method has been used.

Cut dendrogram in such a way that we get 6 clusters. Let us analyze characteristics of each of these individual clusters.

Cluster	Number of Observations	Recency *(in days)*	Frequency *(in days)*	Revenue Contribution (£)	Average Purchase Value (£)
1	1491	44	103	1616	17
2	120	123	248	12631	261
3	380	206	41	510	17
4	1130	96	35	619	25
5	543	315	24	362	23
6	237	42	434	6707	28

Table 4: Hierarchical clusters aggregate statistics

Interpretation of the above table is same as the interpretation of the table for K means clustering statistics which is mentioned above.

- It can be concluded that customers belonging to cluster 2 are high value customers as revenue contribution and average purchase value is very high for this segment of customers
- Customers belonging to cluster 6 are moderate to high value clusters as revenue contribution from each of these customers is on an average 6707 sterling pounds
- Customers belonging to cluster 1 are moderate value customers as revenue contribution from each of these customers is on an average 1616 sterling pounds
- Customers belonging to cluster 3,4 and 5 are low value customers as revenue contribution from each of these customers is on an average very low and the recency and frequency statistics are not that great for these set of customers

Thus 120 customers in cluster 2 are high value customers, 237 customers in cluster 6 are moderate to high value customers, 1491 customers in cluster 1 are moderate value customers and customers in cluster 3, 4 and 5 are low value clusters.

Marketing efforts should be focused excessively on customers in cluster 2 and cluster 6. Customers in clusters 3,4 and 5 can be given least priority when it comes to targeting and personalized marketing campaigns.

4.3 Kernel K means clustering

Kernel K means calculates Kernel function to compute similarity of objects. In kernel k means, for each data point we simply loop through centroid and select the kernel function value which is maximum.
Number of centers for performing Kernel K means clustering is taken as 5 and hence we can see 5 clusters in the data set.

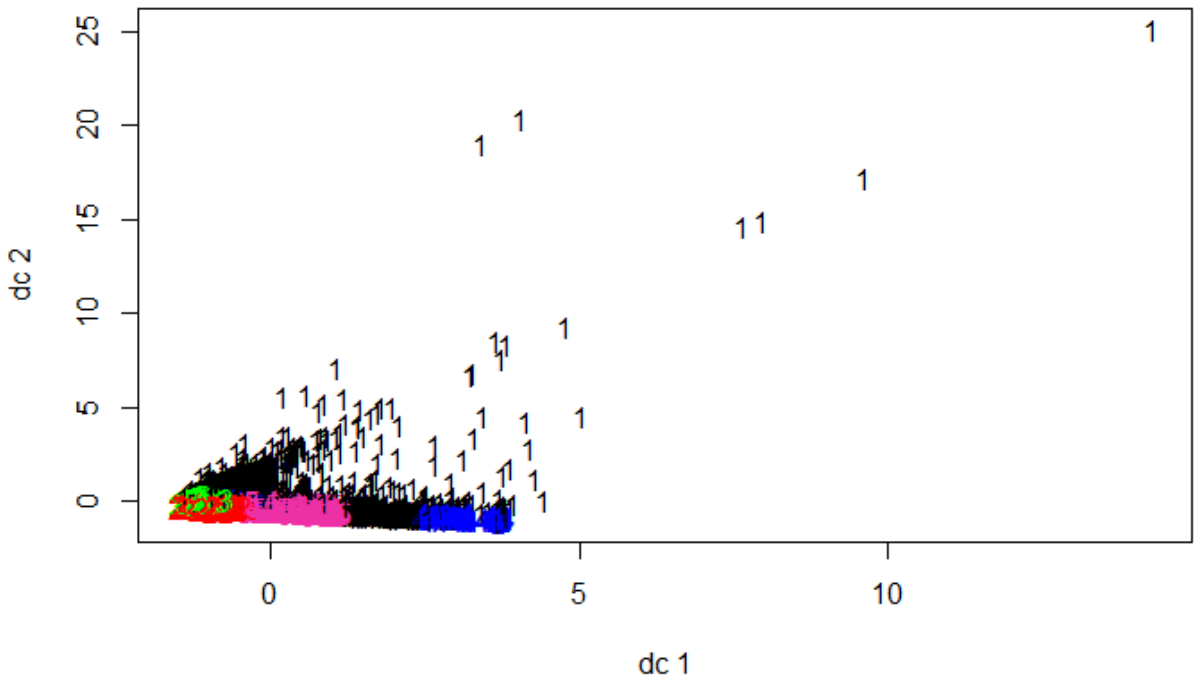

Figure 10: Kernel K means clusters

It is seen that kernel K means clustering has not produced more distinguishing clusters as compared to K means and Hierarchical clustering.

4.4 Principal Component Analysis with K means clustering

Principal Component Analysis is unsupervised learning approach. PCA involves computing principal components and its role in explaining the data set. PCA is performed on a set of variables with no dependent variable in the data set. PCA is a dimension reduction technique which allows most of the variability in the data set to be explained using fewer variables. PCA is used to reduce the number of variables and avoid multicollinearity.

Principal Component Analysis finds a low-dimensional representation of a data set that contains as much of the variation as possible. Each dimension is a linear combination of p variables in the data set and thus it reduces the number of plots necessary for visual analysis.

The way these principal components are calculated is given below.

The first principal component has the largest variance and it is a linear combination of p features.

$$Z1=\phi_{11}X_1+\phi_{21}X_2+\ldots+\phi_{p1}X_p$$

The second principal component is the linear combination of p features that has maximal variance out of all linear combinations that are uncorrelated with Z1.

$$Z2=\phi_{12}X_1+\phi_{22}X_2+\ldots+\phi_{p2}X_p$$

ϕ_1 and ϕ_2 are first principal and second principal component loading vectors.

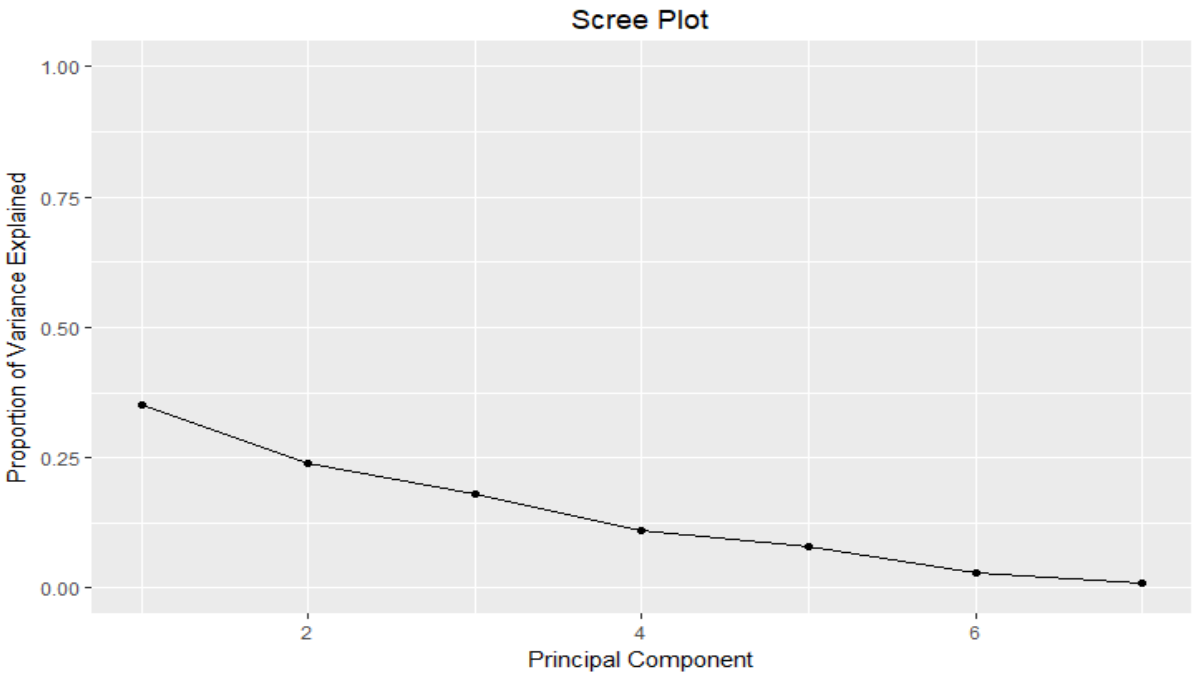

Figure 11: Scree Plot for Principal Component Analysis

Scree plot of Principal Component Analysis shows Proportion of variance explained by different principal components.

Variance explained by PC1 is 35%, PC2 is 24%, PC3 is 18%, PC4 is 11%, PC5 is 8%, PC6 and PC 7 is 3% and 1% respectively.

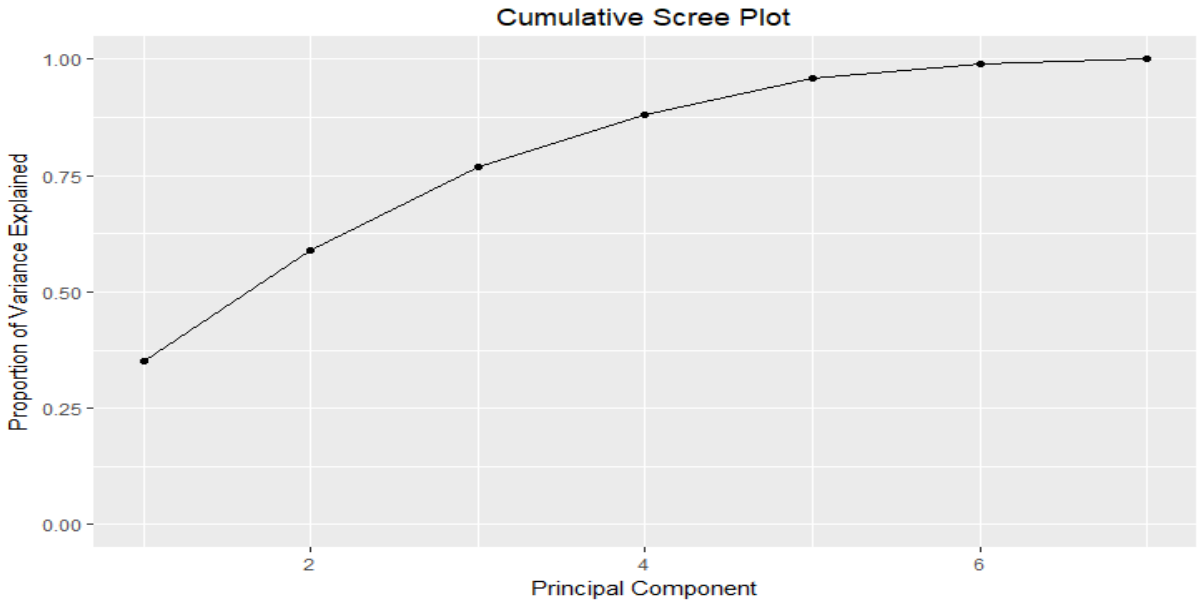

Figure 12: Cumulative Scree Plot for Principal Component Analysis

Cumulative scree plot shows cumulative percent of variance explained by principal components. Combination of PC1, PC2, PC3 and PC4 explains 88% percentage of variance in the data set and hence for our further analysis on clustering I have considered Principal components 1,2,3 and 4.

We prepare a new data frame consisting of customer IDs and its respective PC1, PC2, PC3 and PC4 scores. Principal component scores for sample 6 customer IDs is shown below.

	Cust_ID	PC1	PC2	PC3	PC4
1	12747	0.9402322	-0.8498426	0.22657371	-0.1628614
2	12748	5.8203285	-15.2572192	0.16513771	11.3836529
3	12749	0.2137822	-0.7919091	-0.68059392	0.1932551
4	12820	-0.2273298	-0.3475213	0.00201779	0.2791192
5	12821	-0.5778183	0.7094222	0.82633378	-0.1668786
6	12822	-0.2730007	0.3148239	-0.95935938	-0.3081841

Table 5: Principle Component Scores for sample customers

Now I have performed K means clustering on Principal Component data set. Using elbow plot and average silhouette width analysis, it was found that cluster value of 6 performs best cluster analysis on principal component data set.

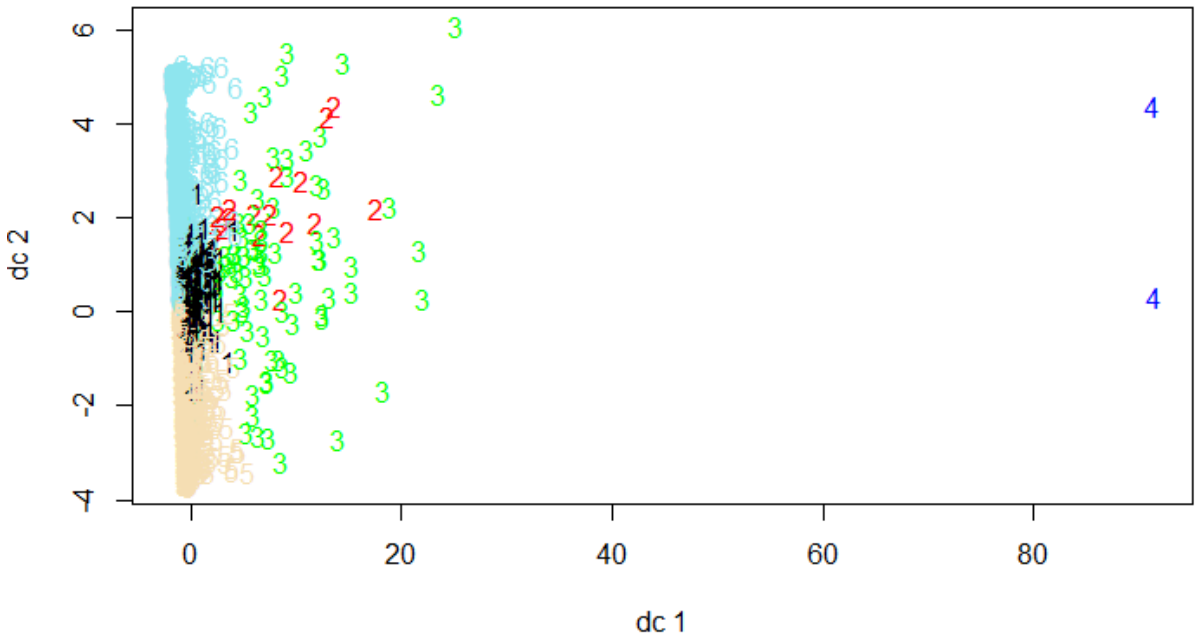

Figure 13: K means clustering on Principal Component data set

Interpretation of the K means clustering plot is given below.

- Above plot shows 6 different clusters plotted on x and y dimensions
- Cluster 6 and Cluster 5 are densely populated
- Cluster 1 is black in color, cluster 2 is red in color, cluster 3 is green, cluster 4 is dark blue in color, cluster 5 is light brown in color and cluster 6 is aqua blue in color

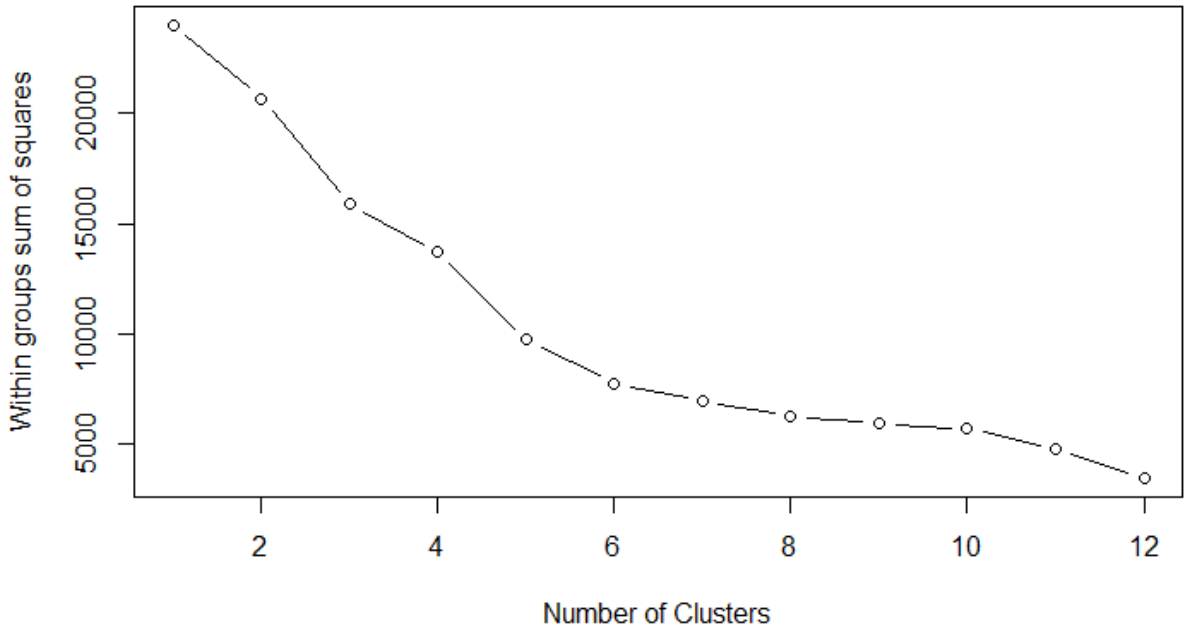

Figure 14: Elbow plot for K means clustering on Principal Component data set

Elbow plot shows optimal number of clusters as 6 to be selected. We can see the significant change in slope after cluster number 6.

Let us look at silhouette width to decide on choosing optimal number of clusters.

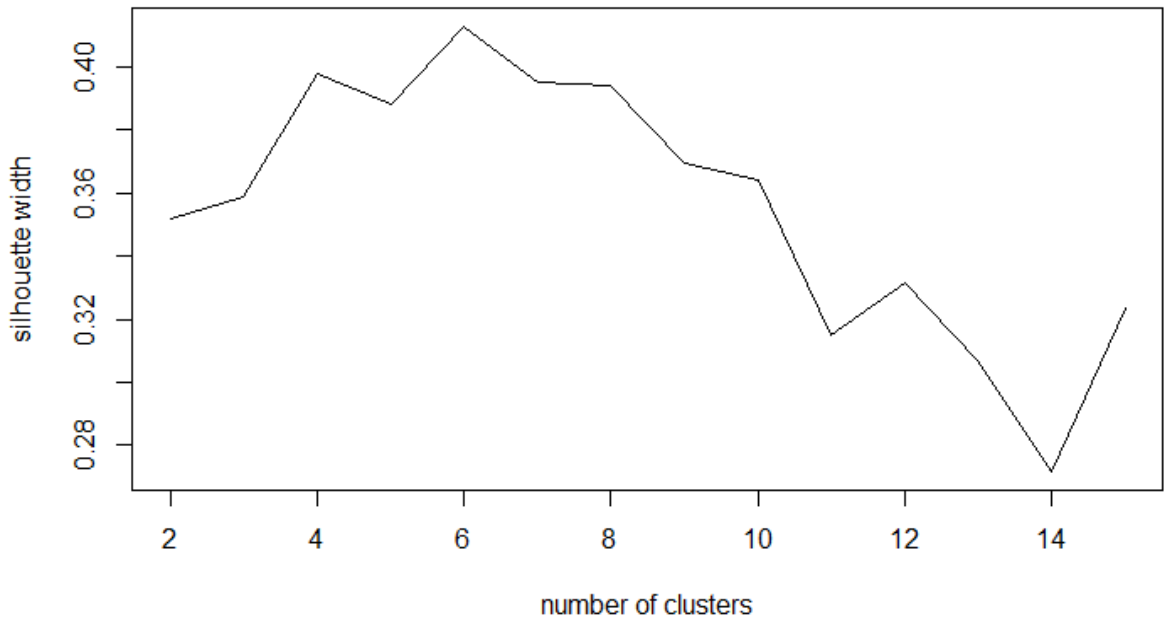

Figure 15: Silhouette width for K means clustering on Principal Component Data

Silhouette width plot shows maximum silhouette width at 6. By analyzing elbow plot and silhouette width plots, it can be concluded that K means clustering on the data set performs best when the cut off cluster value is chosen at 6.

Customer Segment	Number of Customers	Recency *(in days)*	First Purchase *(in days)*	Frequency	Revenue Contribution (£)	Minimum Purchase value (£)	Maximum Purchase value (£)	Average Purchase value (£)
1	1416	59	336	165	2693	4	108	22
2	17	30	373	1607	66079	9	1389	201
3	79	138	267	29	4582	154	590	263
4	2	33	193	2	4188	1956	2117	2064
5	1449	70	116	43	612	7	61	21
6	938	268	308	29	438	9	56	22

Table 6: Characteristics of customer segments i.e. clusters for K means clustering and PCA

Interpretation of the above table is given below.

- It can be concluded that customers belonging to cluster 2 are high value customers as revenue contribution is 66079 pounds per customer, frequency of transactions is very high at 1607 transactions per customer and average purchase value is 263 for this segment of customers
- Customers belonging to cluster 4 seems to be relatively new set of customers who have the potential to be high value customers in future as average purchase value for these customers is 2064
- Customers belonging to cluster 3 are moderate value customers as revenue contribution from each of these customers is on an average 4582 sterling pounds, average purchase value for these customers is 263 and these are active customers whose recency statistics is pretty good
- Customers belonging to cluster 1 are moderate value customers as revenue contribution from each of these customers is on an average 2693 sterling pounds and these are active customers whose recency statistics is pretty good
- Customers belonging to cluster 5 are active, low value customers as revenue contribution from each of these customers is on an average 612 sterling pounds, average purchase value for these customers is 21, recency statistics is pretty good suggesting they are still doing transactions with the business
- Customers belonging to cluster 6 are low value, inactive customers as revenue contribution from each of these customers is on an average 438 sterling pounds and recency statistics is bad for these set of customers suggesting they are not doing business with the online retailer

4.5 Principal Component Analysis with Hierarchical clustering

I have applied Hierarchical clustering on Principal Component Analysis. The dendrogram is such a way that 5 clusters can be seen in the data set.

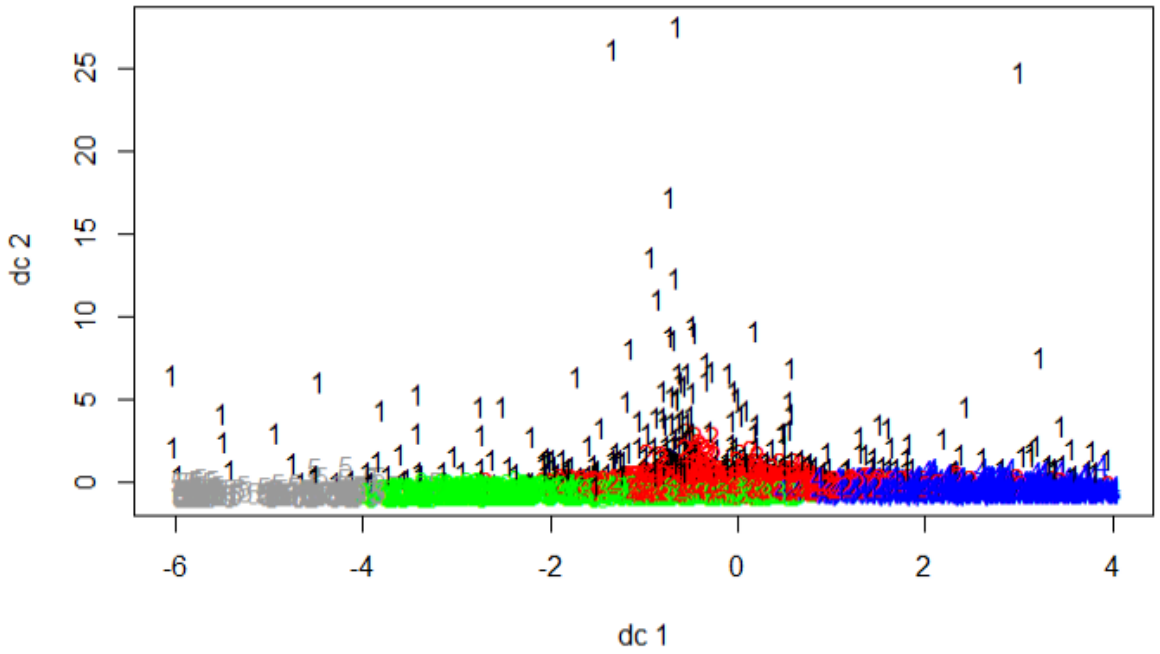

Figure 16: Hierarchical clustering on Principal Component data set

Interpretation of the Hierarchical means clustering plot is given below.

- Above plot shows 5 different clusters plotted on x and y dimensions
- Cluster 3 and Cluster 4 are densely populated
- Cluster 1 is black in color, cluster 2 is red in color, cluster 3 is green, cluster 4 is dark blue in color, cluster 5 is grey in color

K means clustering on Principal Component data set has produced more interpretable and distinguishing different clusters as compared to Hierarchical clustering on Principal Component data.

4.6 Principal Component Analysis with Kernel K means clustering

I have applied Kernel K means clustering on Principal Component Analysis. Number of centers for performing Kernel K means clustering is taken as 5 and hence we can see 5 clusters in the data set.

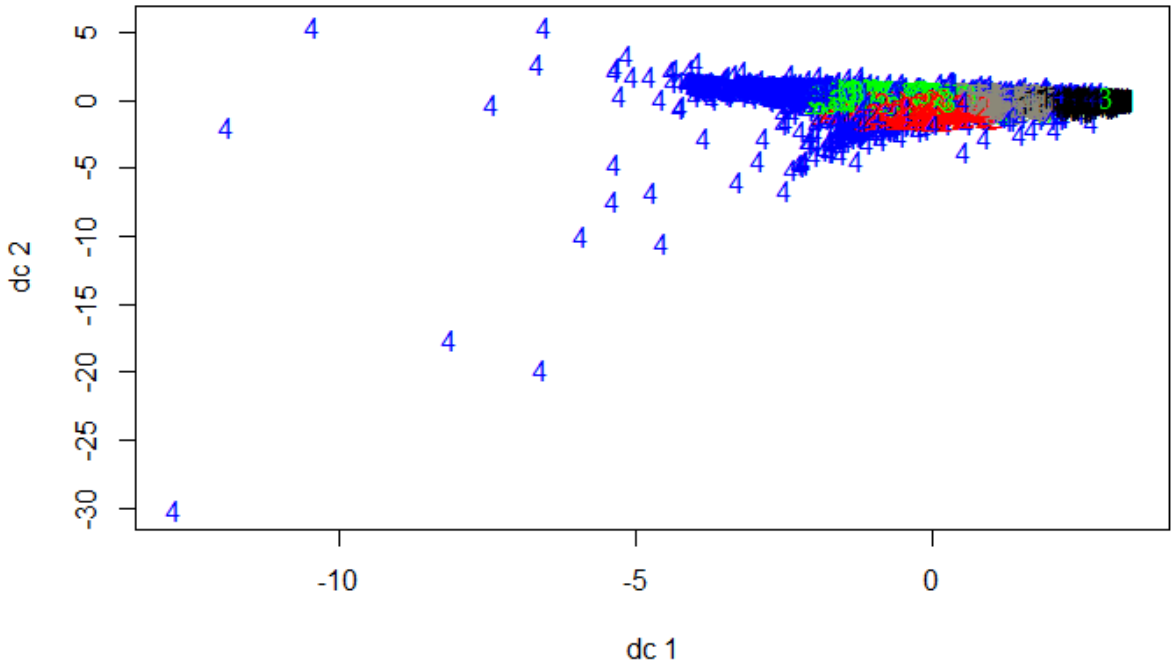

Figure 17: Kernel K means clustering on Principal Component data set

Interpretation of the Hierarchical means clustering plot is given below.

- Above plot shows 5 different clusters plotted on x and y dimensions
- Clusters are difficult to distinguish for Kernel K means clustering technique

K means clustering on Principal Component data set has produced more interpretable and distinguishing clusters as compared to Hierarchical clustering on Principal Component data and Kernel K means on Principal Component data.

5. Conclusion

I have used two different approaches to clustering. Clustering with Principal component Analysis and clustering without Principal component Analysis. It has been found that both approaches have worked fine when it comes to understanding clusters in the data set. However, clustering with principal component analysis seems to be more superior as it captures detail information and hence conclusion is proposed based on Principle Component Analysis approach.

Using K-means clustering on Principle Component data set, it has been found that 17 customers in **cluster 2** are high value customers, 2 customers in **cluster 4** have potential to be high value customers in future, 79 customers in **cluster 3** and 1416 customers in **cluster 1** are moderate value customers, customers in **cluster 5** are low value active customers and customers in **cluster 6** are low value inactive customers.

Marketing efforts should be highly focused on customers in cluster 2 and cluster 4. Significant marketing efforts should be focused on cluster 3 and cluster 1 as they have been contributing steadily to online retailer's business.

Customers in clusters 5 and 6 can be given least priority when it comes to targeting and personalized marketing campaigns.

By identifying these customer segments and its worth to the business, Customer lifetime value analysis can be performed. As this data is captured only for a year, additional data is needed to perform cohort analysis. Additional data and approach to calculate the Customer lifetime value is given below.

$$CLV = [\ \$M - \$R] \times [(1 + d) / (1 + d - r)]$$

- Contribution Margin ($M) for a customer belonging to a customer segment
- Additional data recorded for several years to perform cohort analysis and understand customer retention rate (r)
- Discount rate (d) to calculate net present value of future cash flows
- Marketing spend ($R) to retain a customer
- Transition of customers from one segment to another segment over the years

6. Acknowledgement

Thanks to Dr Daqing Chen for compiling Online Retail data set which helped in understanding the variables in the data set.

Thanks to Professor Ed Winkofsky and Professor Michael Fry for giving me an opportunity to complete the capstone project.

Special thanks to Professor Peng Wang and Professor Yichen Qin for clarifying doubts and helping me in successfully completing this project.

7. References

[1] "k-means clustering," [Online]. Available: https://en.wikipedia.org/wiki/K-means_clustering.

[2] D. Chen, S. L. Sain and K. Guo, "Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining," *Journal of Database Marketing & Customer Strategy Management,* 2012.

[3] B. Boehmke, "Principal Components Analysis," [Online]. Available: http://uc-r.github.io/pca.