**Data Analysis and forecast of Online Retail**

*A project report submitted in fulfilment of the requirement for the*

**BAN 693 – CAPSTONE PROJECT**

Submitted by

**Moumita Bhattacharya (dh7328)**

**Vine Baskaran (cx2528)**

Master of Science in Business Analytics California State University, East Bay

Under the guidance of

**Prof. Jimmy Wu**

****

Department of Master of Business Analytics

December 2021

**CONTENTS**

**1. Goal of Analysis 3**

**2. Tools Used 3**

**3. Data Collection 3**

**4. Data Description 3**

**5. Problem Statement 4**

**6. Data Analysis Approach** **4**

6.1. Data Cleaning and preparation  **4**

6.2. Data Visualization  **4**

6.3. Clustering Algorithms  **6**

6.4. Classification Algorithms **10**

6.5. Apriori Algorithm  **15**

6.6. Time Series Algorithm  **19**

**7. Results: 21**

**8. Limitation: 23**

**9. Conclusion: 24**

**10. Reference:** **25**

## **1. Goal of Analysis:** The goal of the project is to predict total sales for Online Retail with different timeline among different countries. Clustering models such as K Means hierarchical and non-hierarchical are used to produce RMF clusters. Classification models such as KNN, Logistic, SVM, PCA, Naive Bayes and Decision trees are used for total sales prediction and Apriori Algorithm for Market Basket Analysis. The historical data of 2010 and 2011, was considered to develop the time series model

**2. Tools Used:** Models were built using the R programming language in RStudio and Python programming language in Jupyter Notebook to predict accurate values.

**3. Data Collection:** The dataset was collected from Kaggle. Historical data of past 13 months, from Q4 2011 to Q4 2012, was considered to develop the model. (https://www.kaggle.com/lakshmi25npathi/online-retail-dataset).

**4. Data Description**

* InvoiceNo: Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
* StockCode: Nominal, a 5-digit integral number uniquely assigned to each distinct product.
* Description: Product (item) name. Nominal.
* Quantity: The quantities of each product (item) per transaction. Numeric.
* InvoiceDate: Numeric, the day and time when each transaction was generated.
* UnitPrice: Numeric, Product price per unit in sterling.
* CustomerID: Nominal, a 5-digit integral number uniquely assigned to each customer.
* Country: Nominal, the name of the country where each customer resides.

**5. Problem Statement and Research Questions**

The goal of this project is to predict the department wide monthly sales for a store. This should then help to optimize the manufacturing process in different locations and therefore to increase income while lowering costs. It should be possible to feed in past sales data from a department and to get the predicted monthly sales. It also calculates what product sold the most? What time should the company display advertisements to maximize likelihood of customers buying products and to target which customer? Which city and which country has the highest sales? Identify customer churn rate and customer lifetime value of customers?

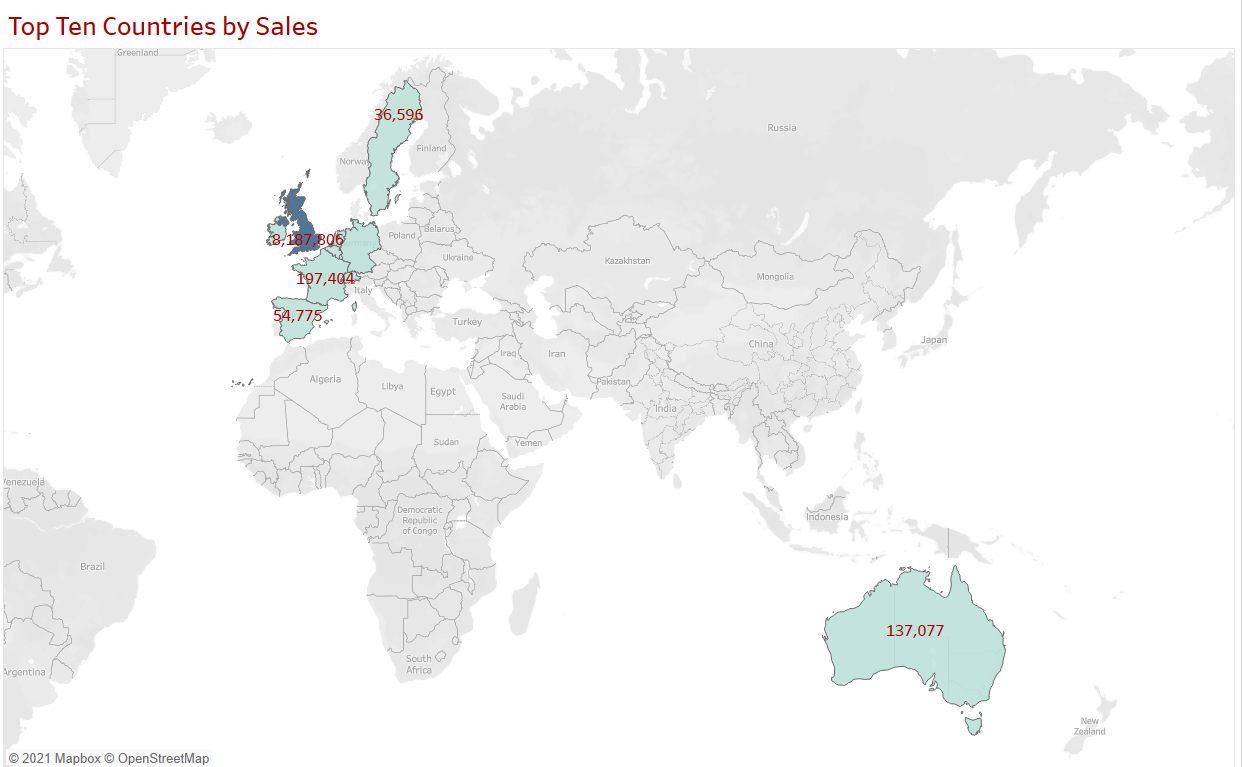
**6. Data Analysis Approach**

**6.1. Data Cleaning**: By importing necessary libraries, incorrect spellings, extra whitespaces, irregular values, null values, duplicates and transactions with price zero, post, cancelled are removed. Date with incorrect format is corrected. Also highly skewed data is removed for further analysis.

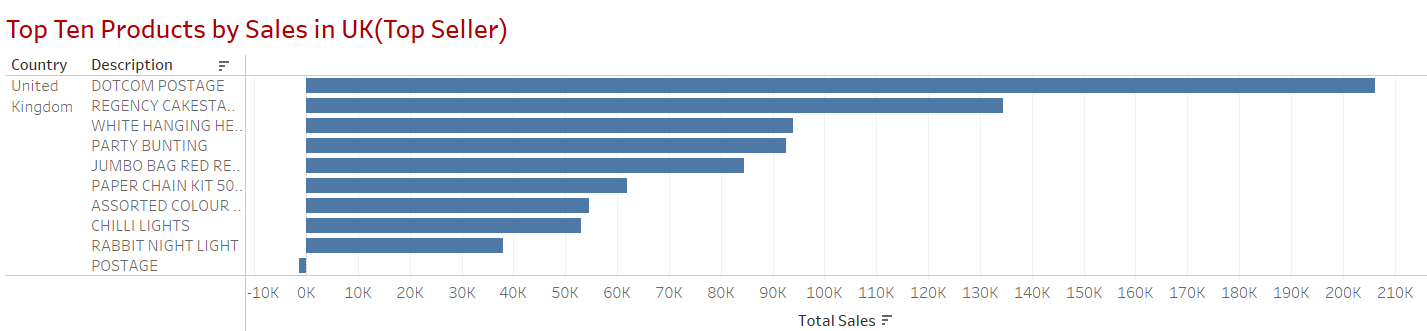
**Data Preparation:** Calculated and created new data columns for recency, frequency, monetary, total unit, total transaction, total price, total sales, average order value, purchase rate, customer value and customer lifetime value to identify the churn rate and repeat rate of customers. Variables were normalized using the center scale method. Euclidean distance was used to calculate the normalized distance for RMF analysis to target specific customers

**6.2 Data Visualization:**

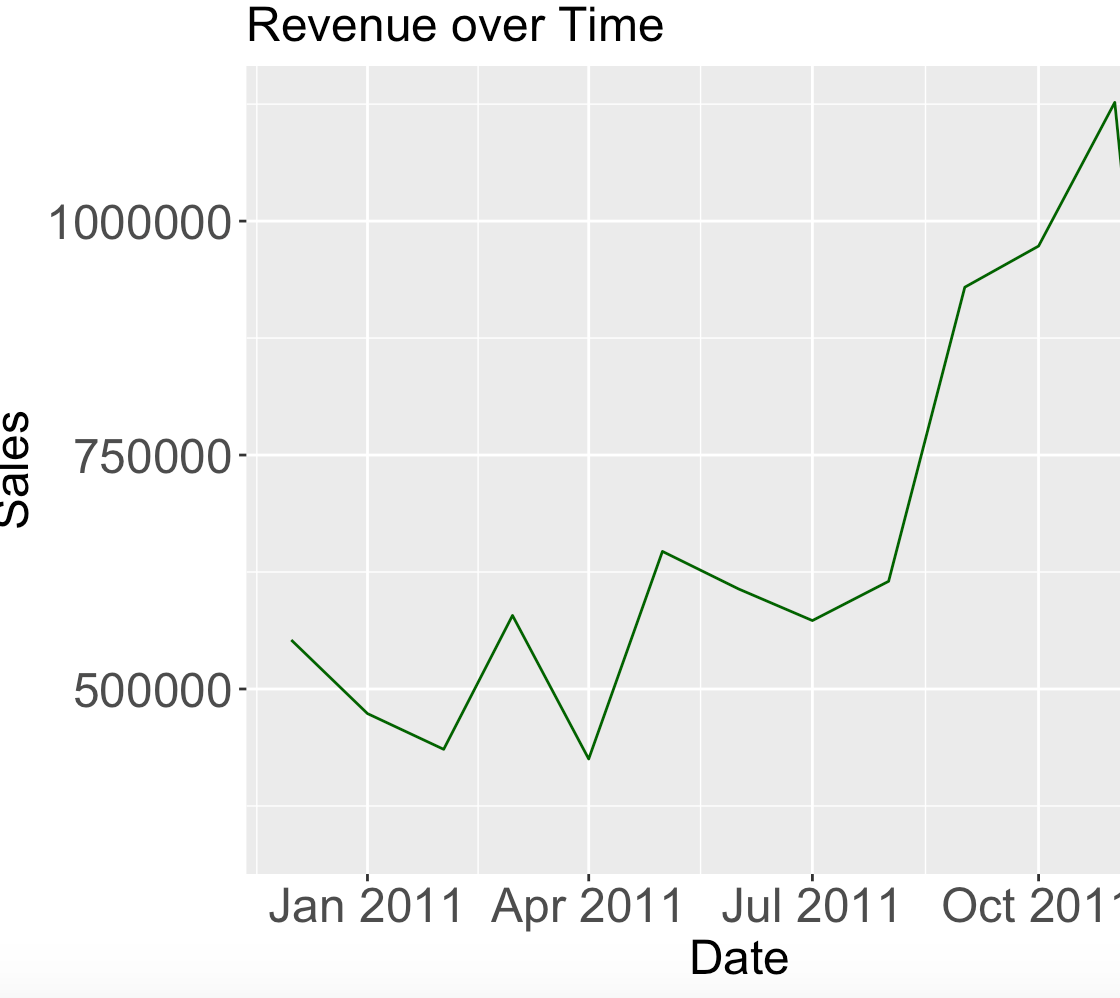
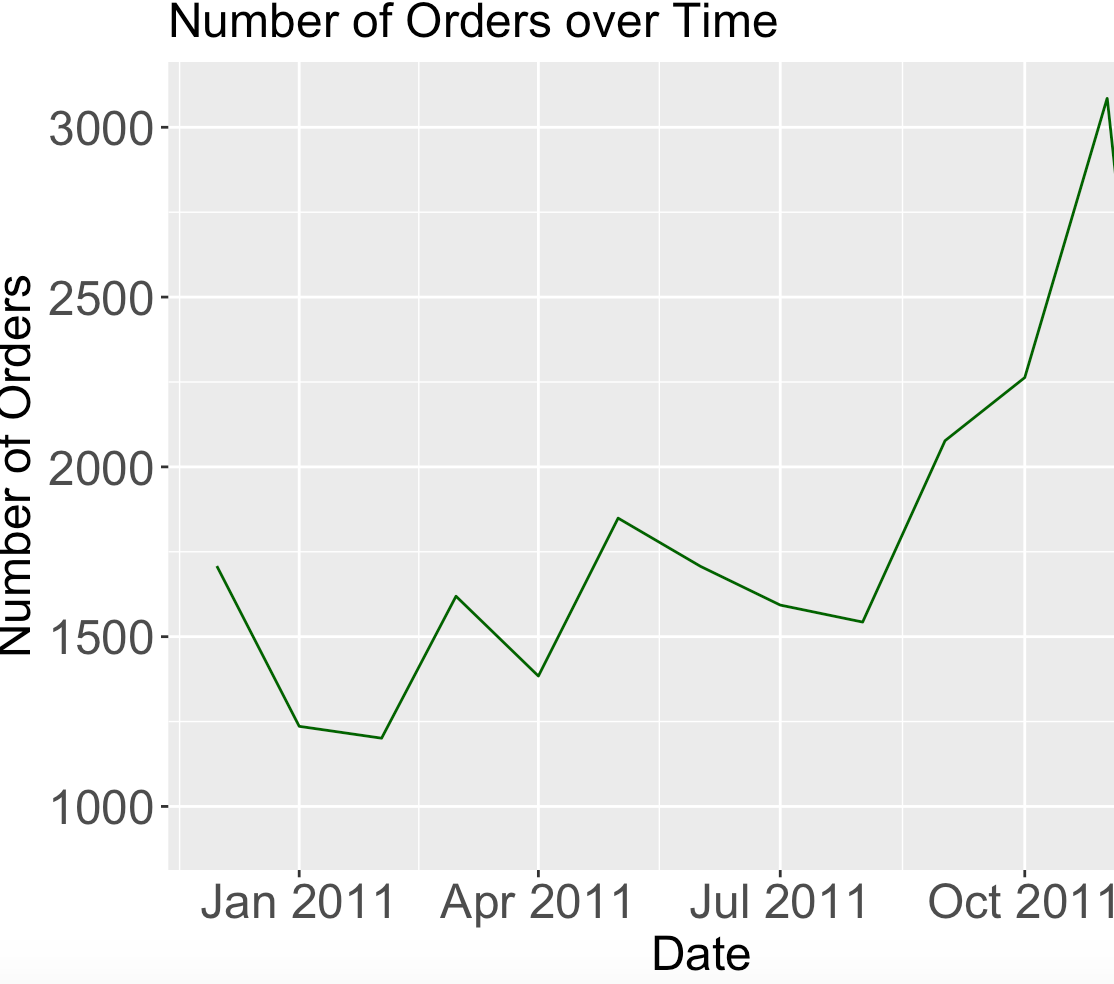
Through different angles of data visualizations, we have identified the most selling country, city as well as popular products throughout the year that determines the total profits levels. Amongst the European countries, the top seller country is the United Kingdom followed by Netherland, EIRE, Germany, France dominating the market.

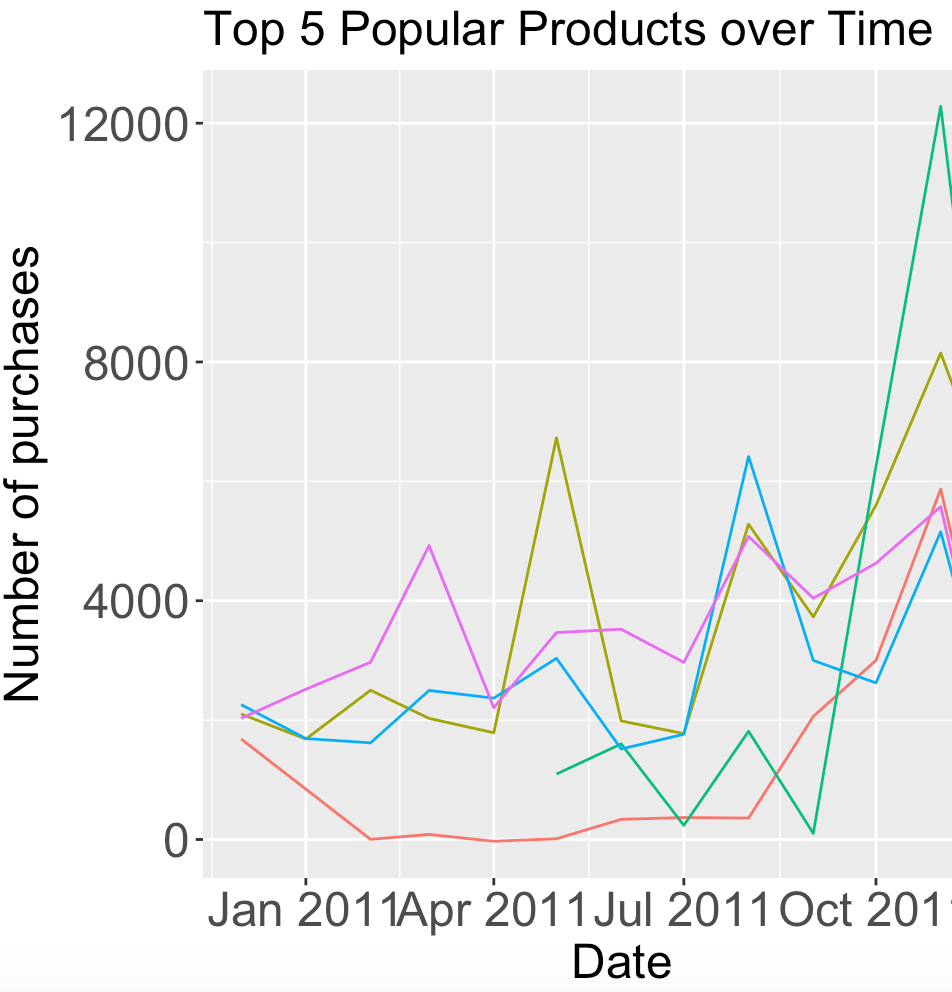
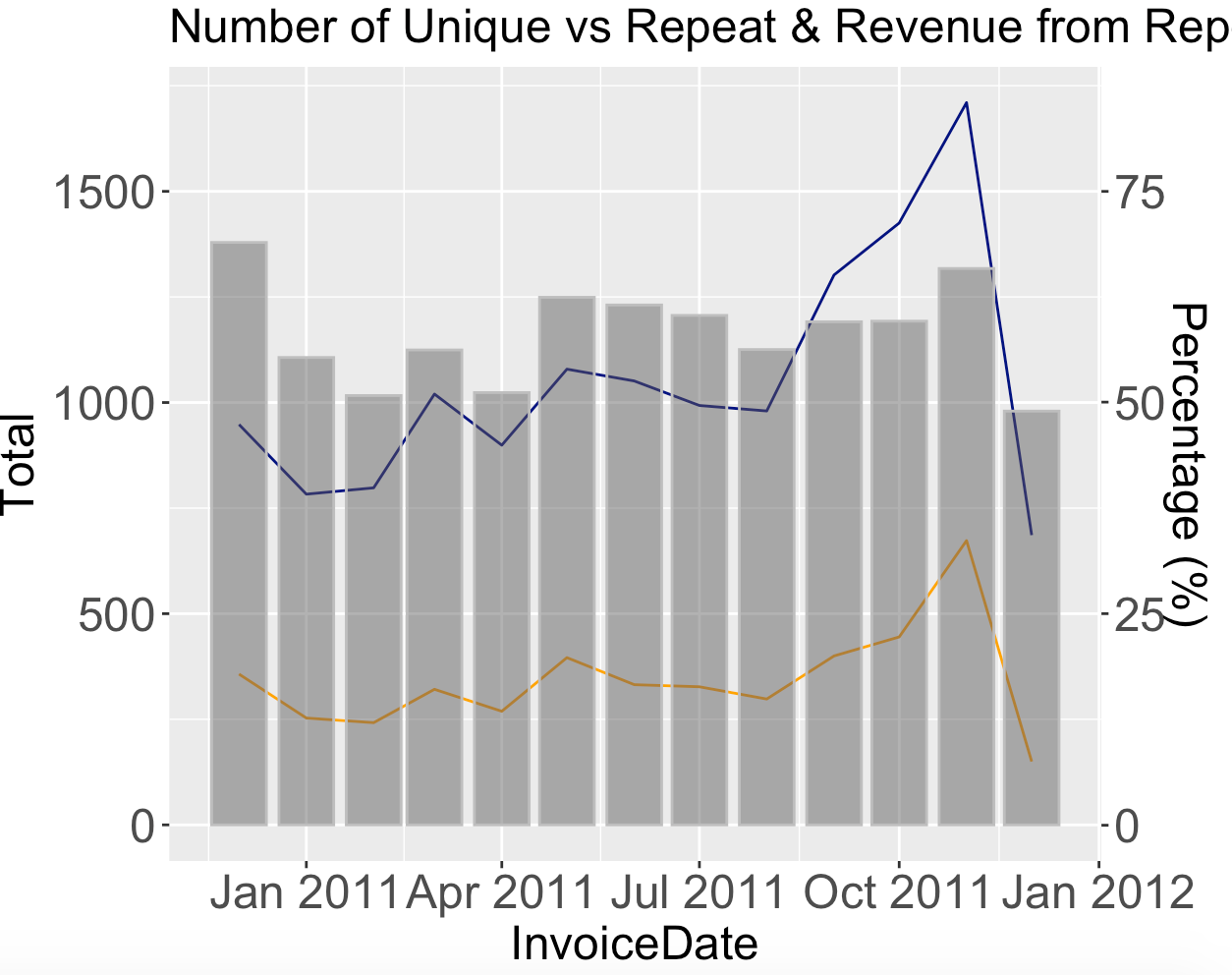
****

From the below picture, we can see that the top ten seller products are Dotcom Postage, Regency Cake, White Hanging, Party Bunting etc.

****

The dataset which we will be analyzing is for a UK-based online retail store which majorly sells gift products. The customer base of the company are mainly wholesale retailers. So, the obvious selling trend of the market is high in the holiday session compared to other months.

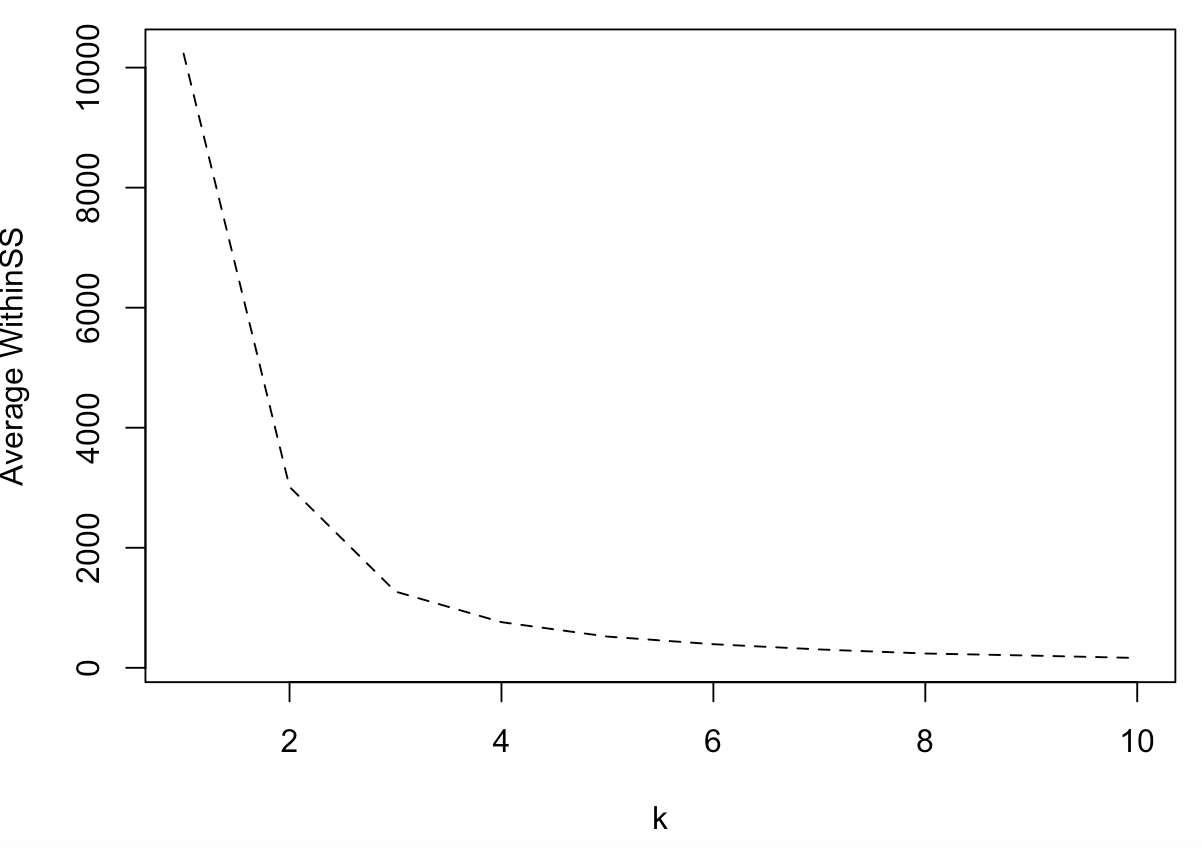
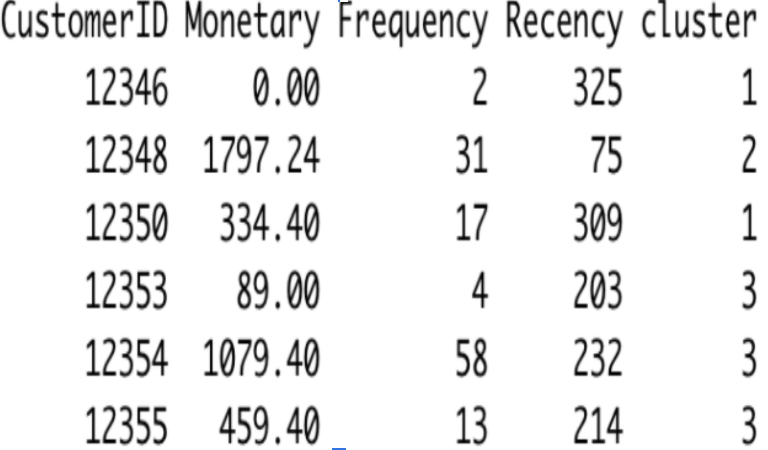
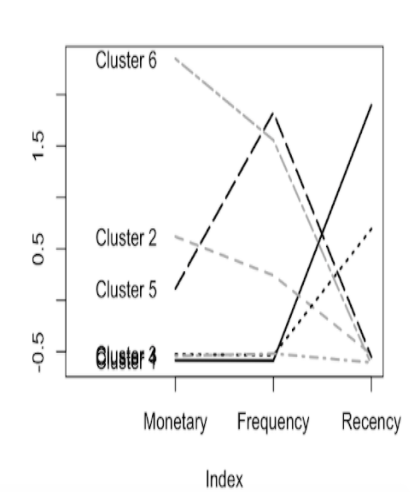




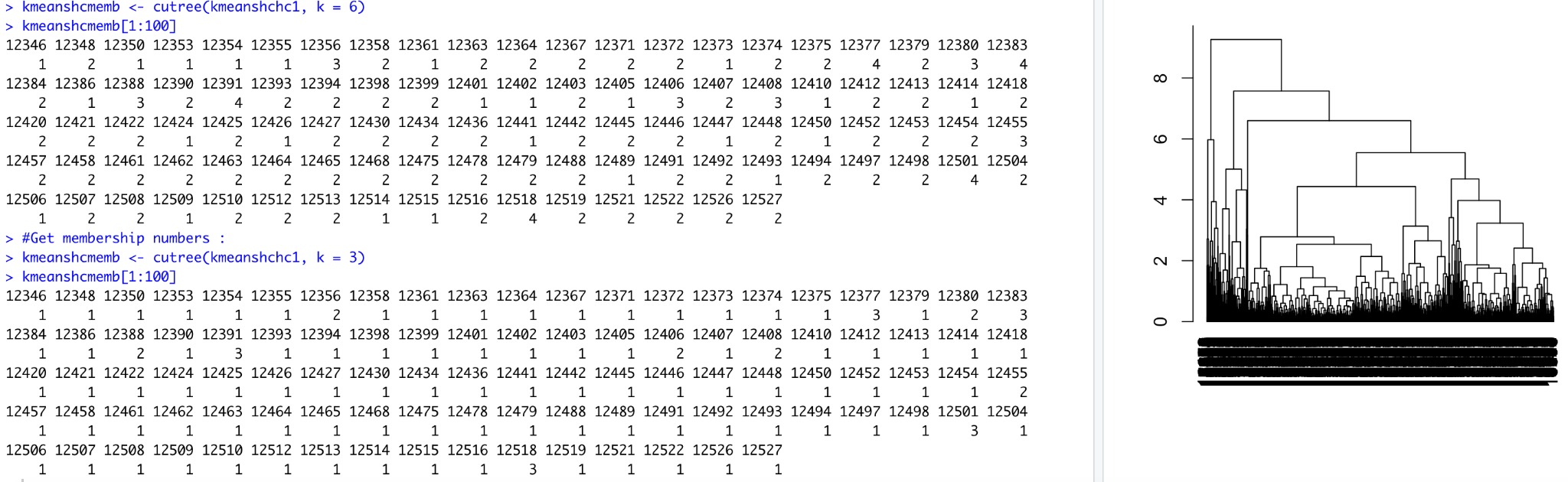
The above plots represent the number of orders over time, revenue over time, number of unique, repeated items and number of purchases of top 5 products over time. It is observed from the trend that the sale during the end of year is much higher compared to any of the months of the year.

**6.3 Clustering Algorithms:**

**6.3.1 K Means Non-Hierarchical clustering:** K creates random centroids and allocates some of these records to each of these clusters. The plot shows the centers of low and high values of each of all 6 clusters in the table based on variables (RMF). The table below shows clusters formed with the label name customer id where Customer ID 12346 & 12350 are in cluster 1 and Customer ID 12353,12354 & 12355 are in cluster 3. In order to pick the best value of k the elbow curve is used. **k= 3** has the best accuracy

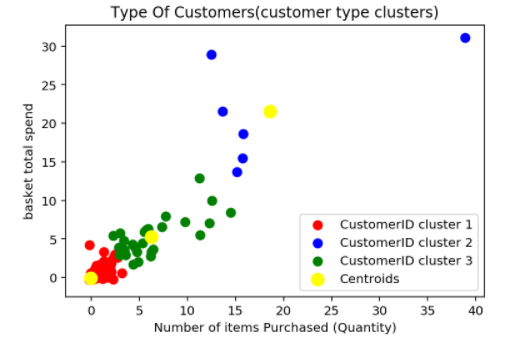
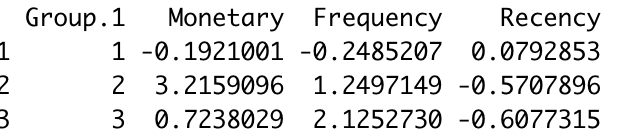


**6.3.2 K Means Hierarchical clustering:** The cluster dendrogram plot below shows how clusters are merged step by step with the label name Customer ID using **Complete method**

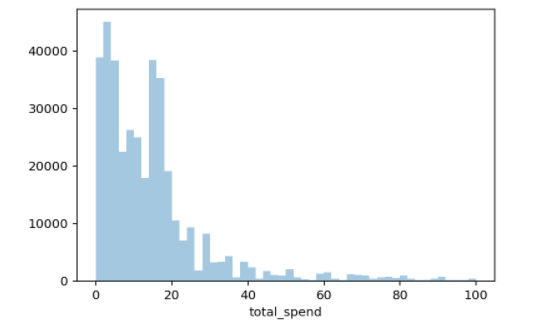
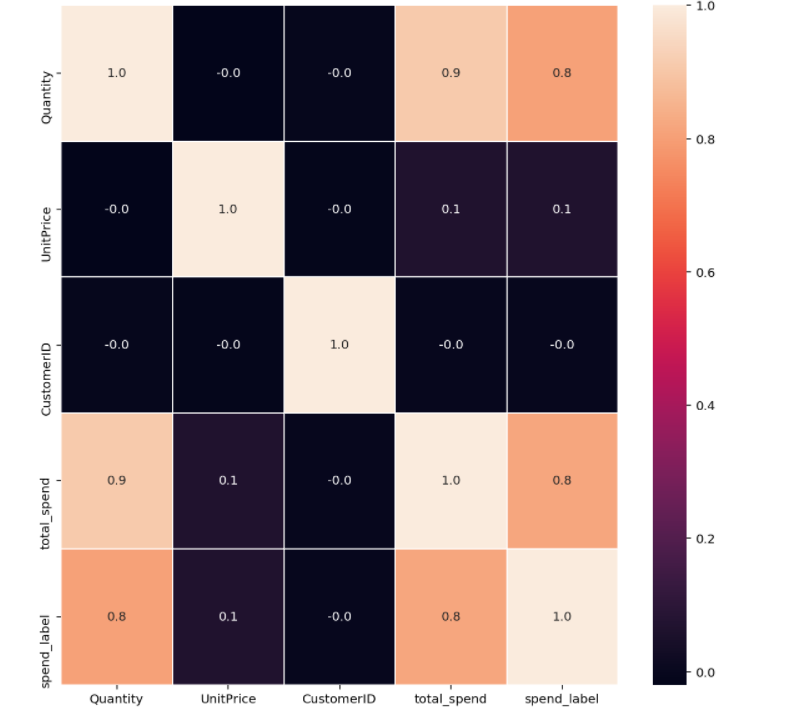


The output result of cluster (k=6) means the first observation is in cluster 1, the second observation is in cluster 2 and so on. Similarly, the output result of cluster (k=3) means the first observation is in cluster 1, the second observation is in cluster 1 and so on. Each cluster aggregates to split the data into subsets that compute summary statistics for each. The subset split is based on the cluster (k=3) and cluster (k=6) and summarizes each subset by their mean. Picking clusters with k =3 as it provides better accuracy to target specific customers.

**With 3 clusters observations** - average monetary is high in cluster 2, average frequency is high in cluster 3 and the average recency is high in cluster 1. The below K-Means Clustering separates the customers based on numbers of items and total spend. Cluster 1 has the least number of items sold whereas cluster 2 has the highest number of items sold.



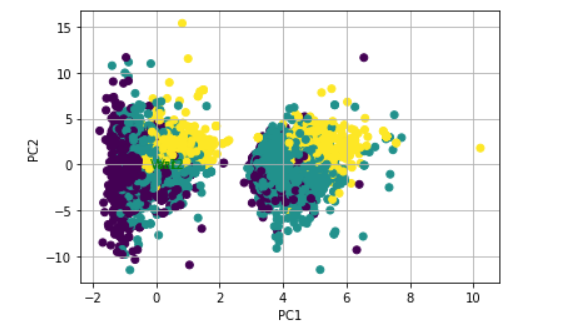
Before Modeling a cursory examination of correlation to identify potentially problematic variables from the model training dataset. Multicollinearity is detected by correlation heat map(Figure below) and variance inflation factor (VIF). Removal of some independent variables that are highly correlated with each other amended this issue. Unsurprisingly, total spend correlates with quantity and the spend\_label. As the spend\_label is very imbalanced, for the purposes of our research we used the lower value baskets for prediction.Below is what we have identified a range suitable for binning: data[(data['total\_spend'] >0) & (data['total\_spend'] < 100)].Here we are repeating the labeling of total spend from low (0), medium (1), and high (2), but with the lower range of values. After taking three labels of baskets - labels represent 0 - low value baskets, 1 - medium value baskets, 3 - higher value baskets.



Here Heatmap (left fig) represents the correlation between each of the attributes we considered for our model and the total spending (right fig) which is mostly restricted below 100.

**PCA model:** From the below visualization analysis of the dataset, the majority of variance in the model is in the 1st principle component. Total spend is sufficient to predict the basket size label.

From visualization of the principal components, the labels are well separated/clustered, facilitating machine learning.



**6. 4 Classification Models to Predict Total Sales**

**Data Partition:** Dataset were divided into three partitions 50%, 30% and 20% respectively: training, validation and test data sets to develop models and check performance using the validation data and test data.

**6.4.1 K-NN Algorithm:**

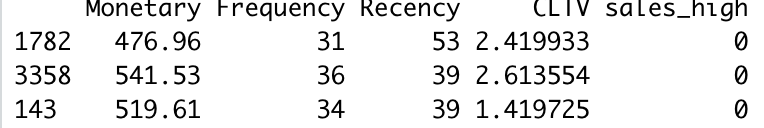
Assume New Data X predictor: "Monetary"=500,"Frequency"=35, "Recency"=50, "CLTV"=3

Y Variable Sales: Is the total sales higher or lower than mean total sales? Considering Class 0 as No(Sales less than mean Sales) and Class 1 as Yes(Sales greater than mean Sales). Predicting if total sales was higher than total mean sales using new data. Predictor variables are standardized to put them on comparable scales. Typically, the value of k is the lowest error rate in validation data. K=3 with 0.957 has the best accuracy that is used to find high accuracy in validation data. Below are the confusion matrix accuracy measures.

* In test data using train data, among 489, 472 are predicted correctly, the misclassified are 17. Among 194, 178 are predicted correctly; the misclassified are 16.
* In test data using full data, among 487, 479 are predicted correctly the misclassified are 8. Among 196, 187 are predicted correctly; the misclassified are 9.

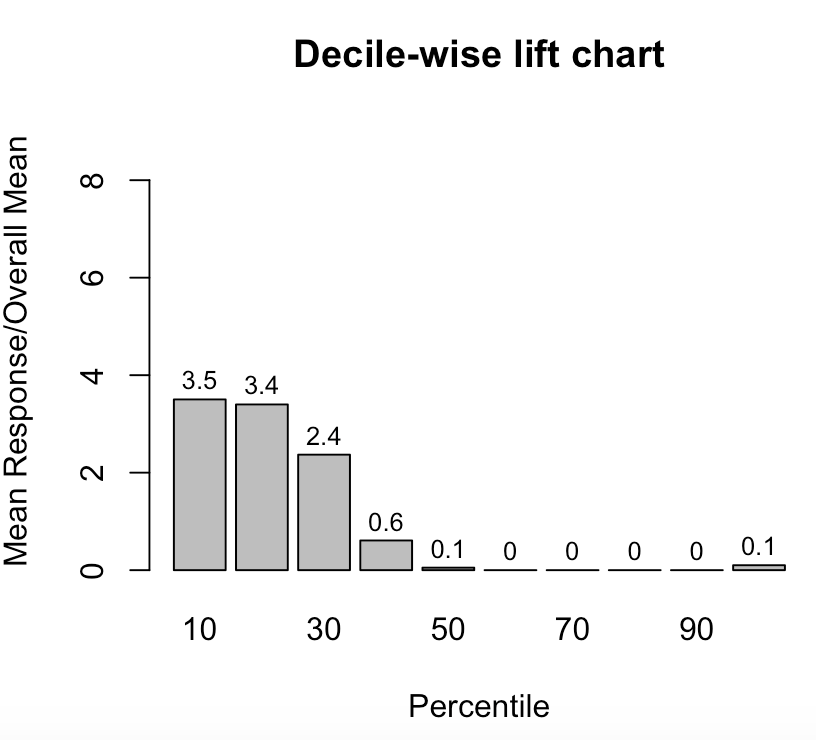
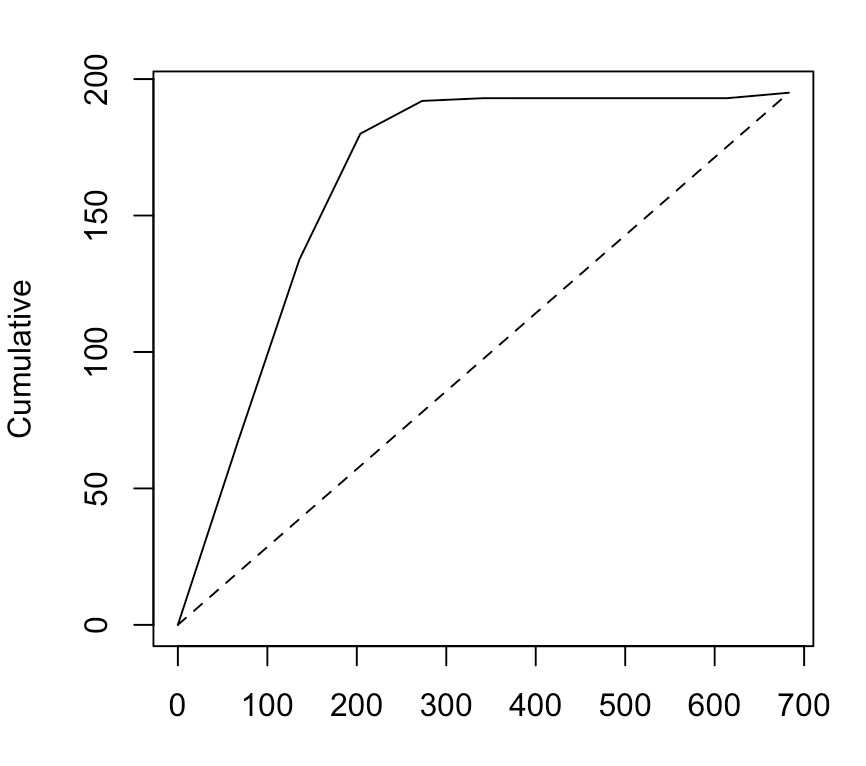
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **KNN Confusion Matrix and Statistics** | **Accuracy** | **Sensitivity** | **Specificity** | **Predicated** **Class** |
| **Training data to train and test data as test** | 0.9517 | 0.9672 | 0.9128 | Class ‘0’ – No Sales less than mean Sales) |
| **Full data to train and test data to test** | 0.9751 | 0.9816 | 0.9590 | Class ‘0’ – No (Sales less than mean Sales) |

K-NN compares each record from the new dataset to k=3 nearest records in the dataset. Records 1782, 3358 and 143 are the nearest k neighbors picked by Euclidean distance measure. There are ‘3’ zeros for sales, therefore the final predicted class is 0, by default it chooses the majority. It means sales is less than mean sales



**6.4.2 Logistic:**

The first 5 actual and predicted records from validation and test data have the prediction probabilities based on variables (RMF) and CLTV. However, using the gains table the performance of validation/test data for other records can be identified more accurately. Validation and test data are grouped into 10 different groups where each group has 102-103 records and 68-69 records respectively. Cumu N is the cumulative total of the records of each row. Ex test data: second row 68 + 68 records, the third row will have the first 3 records 68+68+68.Mean resp for test data: from 68 records, mean score is 100% which means 100% of 68 is predicted to be class 0 and Mean cum resp: out of cumulative records 68 + 68 = 136, 99% of 136 is predicted to be class 0 - where the Lift index is the ratio between mean resp and highest cumulative mean resp(last row) out of all the records. From Gain%, cume pct. of total resp: out of 68 records, 34.9 % predicted to be class 0 in test data.

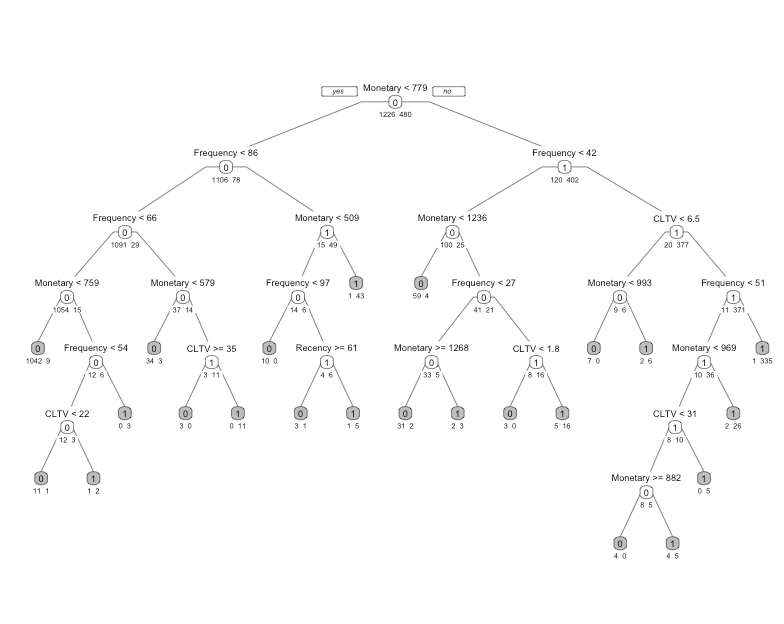


The Lift curve is pulled away from the base prediction line and from the first few tall bars of the decile-wise lift chart depicts the prediction model is good where it gives the ratio between mean response and overall mean. In test data using train data, among 490, 471 are predicted correctly, the misclassified are 19. Among 193, 176 are predicted correctly; the misclassified are 17.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Logit - Confusion Matrix** | **Accuracy** | **Sensitivity** | **Specificity** | **Predicated**  **Class** |
| **Testing data** | 0.9473 | ​​ 0.9652 | 0.9026 | Class ‘0’ – No (Sales less than mean Sales) |

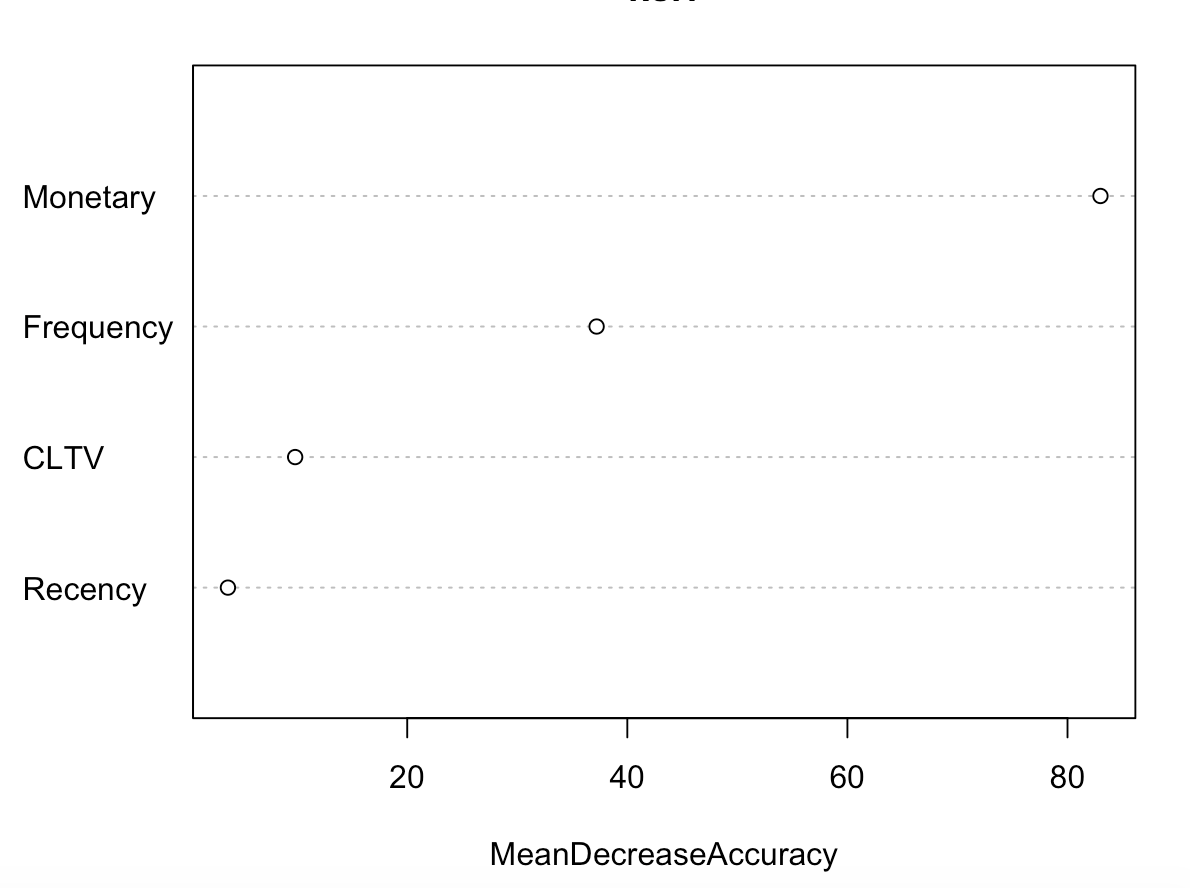
**6.4.3 Decision Trees:** Classification tree with default parameters based on a variables (RMF) and CLTV - rules based on the tree if

* Monetary > 779 and Frequency > 42 then predicted class = 1.
* 509 > Monetary < 779 and Frequency > 86 then predicted class = 1

.

To involve more predictors, a deeper tree was developed( cp=0), then pruned by lower cp to find a better parsimonious pruned tree with cp=0.00208333 at split 10 to avoid overfitting of the training data where cross validation pruning was added as a cushion to minimum error as it helps to pick the right cp level to stop tree growth

**Random Forest:** Variable importance plot provides a list of the most significant variables in descending order by a mean decrease in Gini. The top variables Monetary followed by Frequency contribute more to the model than the bottom ones i.e. CLTV and Recency.

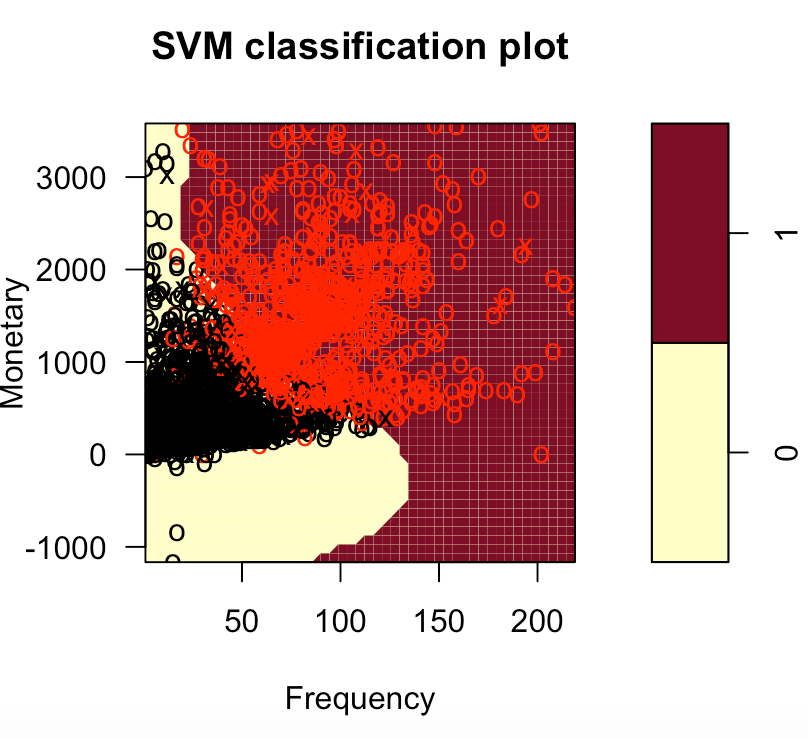
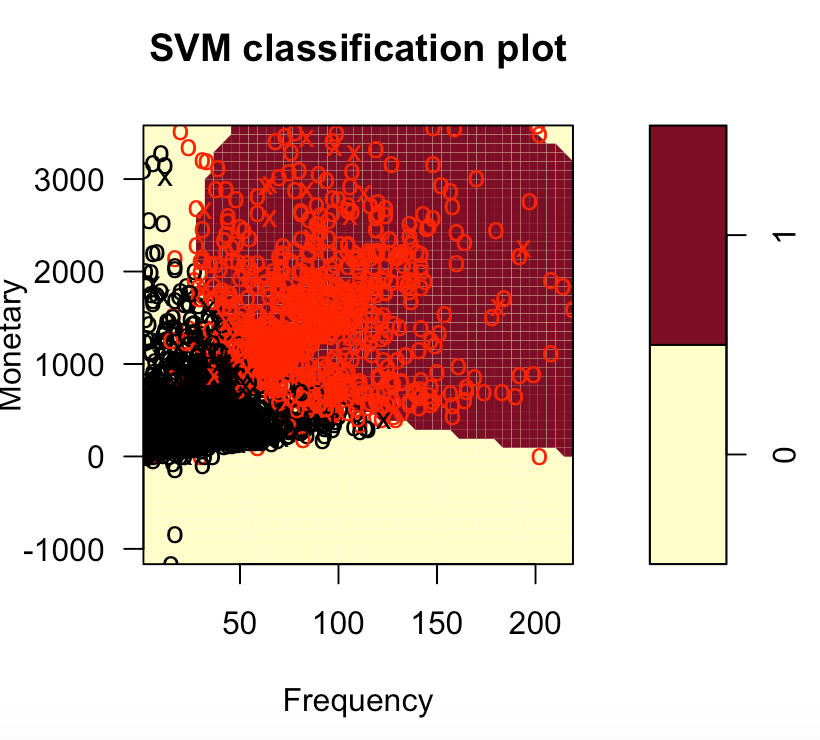


|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Confusion Matrix** | **Accuracy** | **Sensitivity** | **Specificity** | **Predicated** **Class** |
| **Classification Trees Test data** | 0.9488 | ​​0.9775 | 0.8769 | Class ‘0’ - No (Sales less than mean Sales) |
| **Random Trees Test data** | 00.9561 | 0.9693 | 0.9231 | Class ‘0’ - No(Sales less than mean Sales) |
| **Boosted Trees**  **Test data** | 0.9561 | 0.9693 | 0.9231 | Class ‘0’ – No (Sales less than mean Sales) |

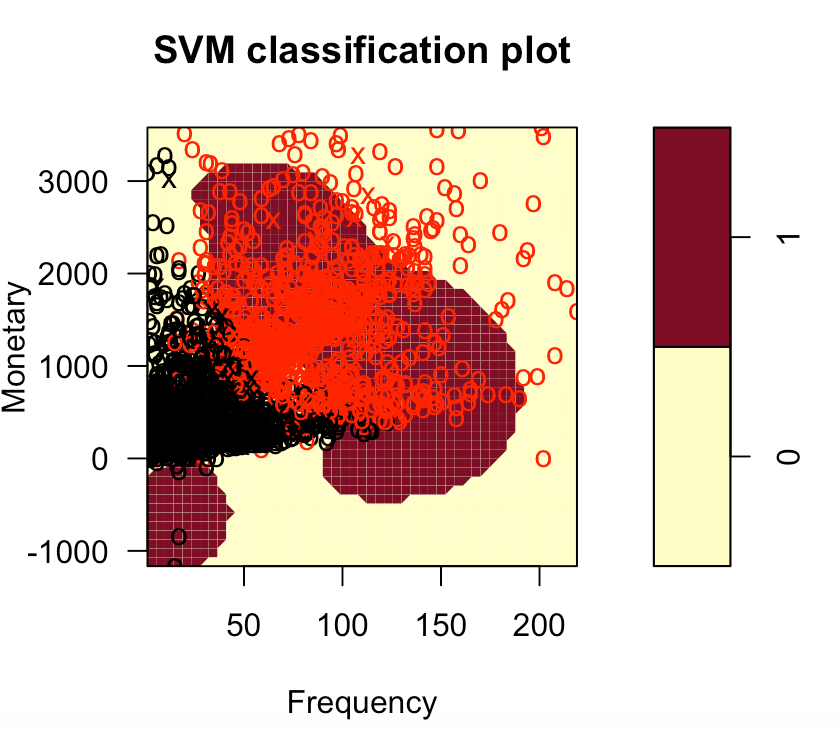
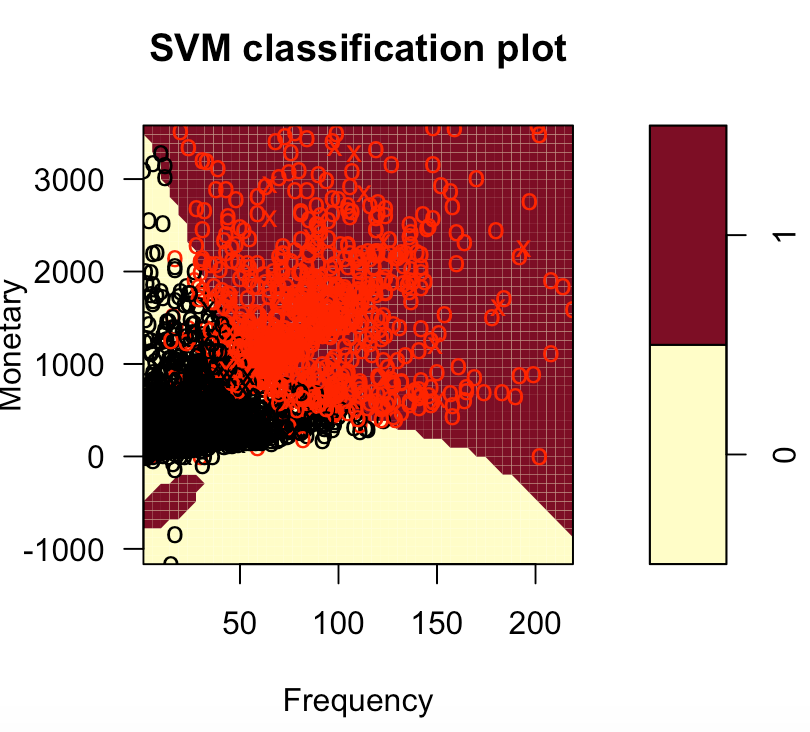
* CT: In test data using train data, 477 are predicted correctly, the misclassified are 24. 171 are predicted correctly; the misclassified are 11.
* Random Trees: In test data using train data, 473 are predicted correctly, the misclassified are 15. 180 are predicted correctly; the misclassified are 15.
* Boosted Trees: In test data using train data, 473 are predicted correctly, the misclassified are 15. 180 are predicted correctly; the misclassified are 15.

**6.4.4 SVM:** Based on a variables (RMF) and CLTV for predicting total mean sales, compare radial kernel with large γ, small γ (0.1 vs 0.5), large and small misclassification cost (1 vs 10000)

* Low gamma tends to underfit while high gamma tends to overfit.From the lhs plot: Lower the cost, the lower the penalty to misclassification and low gamma corresponds to large dispersion in the clusters of points.From the rhs plot: Lower the cost, the lower the penalty to misclassification and higher gamma corresponds to low dispersion in the clusters of points and depicts changes in the boundary compared to lhs plot

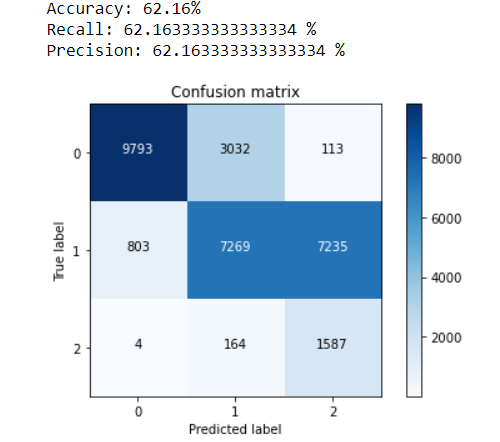


* From the fig cost with 100,000 has larger misclassifications compared to cost with 1 that has small misclassification.



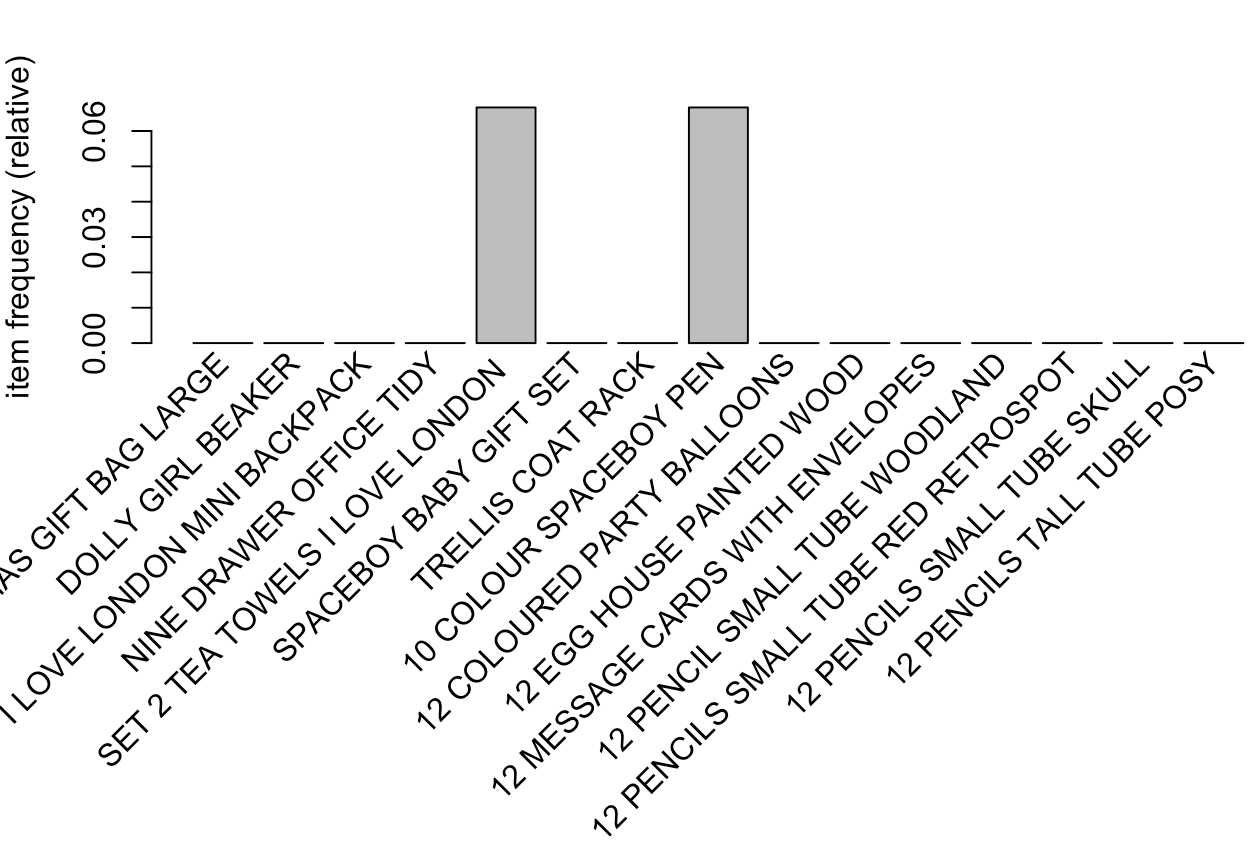
* From the lhs plot: Higher the cost, the higher the penalty to misclassification and low gamma corresponds to large dispersion in the clusters of points. From the rhs plot: Higher the cost, the higher the penalty to misclassification and higher gamma corresponds to low dispersion in the clusters of points and depicts changes in the boundary compared to lhs plot. Accuracy for this model is approximately 66% where it predicted the sales is less than mean sales

**6.4.5 Naive Bayes:** This method will provide an indication of whether or not a model can be built using this dataset without using more computationally expensive methods. Trained GaussianNB model is used for prediction on the test data, the following accuracy is approximately 63%. The model accuracy predicts how likely a customer is to be purchasing a low, medium, or high value basket. The model is little biased toward the small basket size, could be due to bias in the training set.

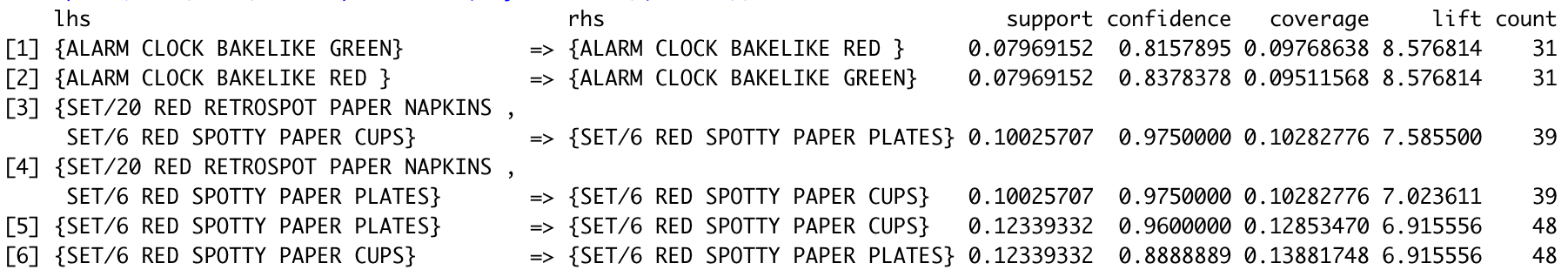


**6.5 Aprior:** Considering a few European countries for market basket analysis to identify what items were bought together. Binary index matrix is created by converting the columns binary (1 if purchase- irrespective of the number and 0 if no purchase) and converted into a matrix form.

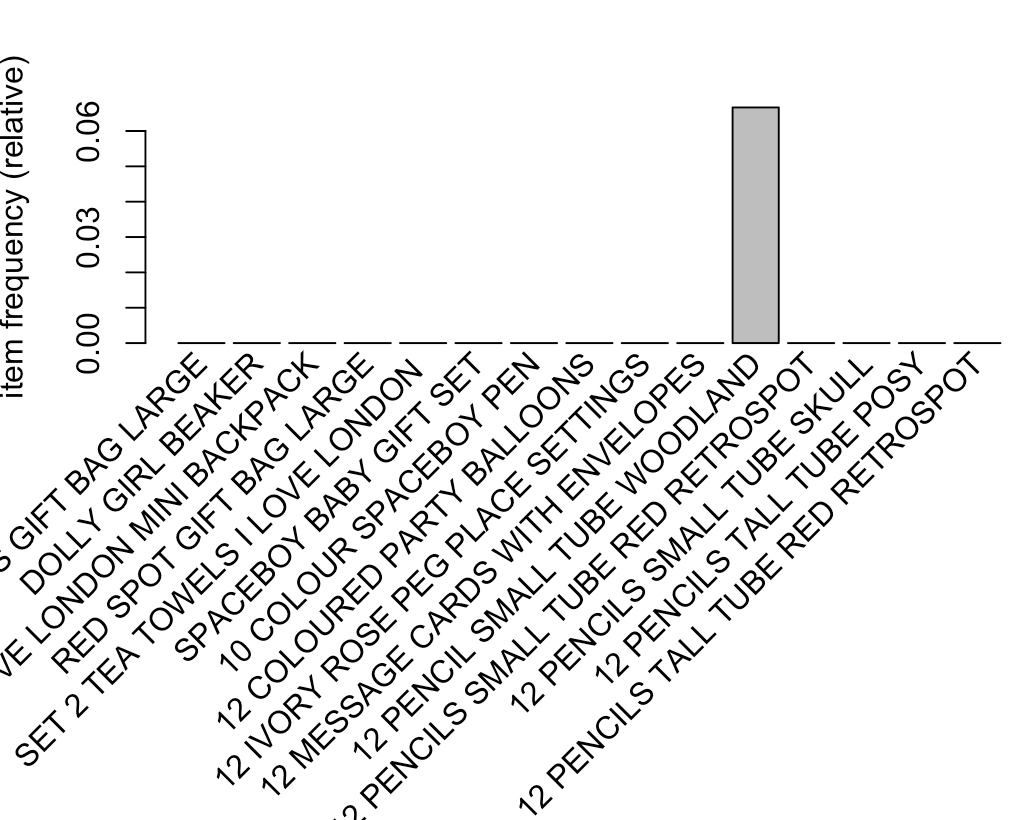
**France:** Item Frequency for first 15 transactions is shown below

****

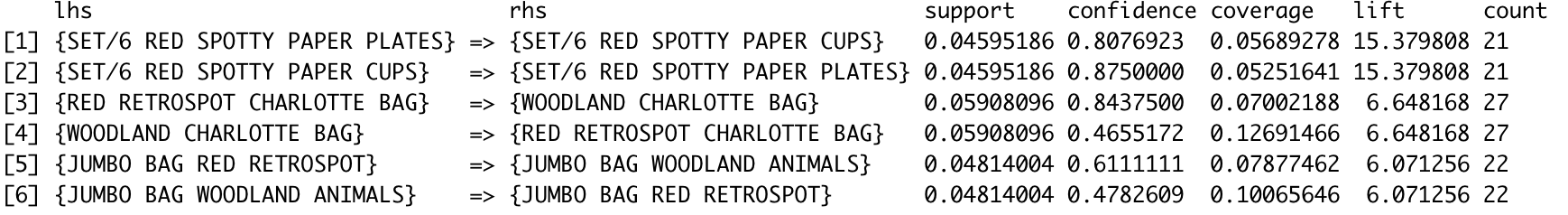
By Inspecting the first six rules, sorted by their lift. Rule 1 & 2 has the highest lift (8.576814) with confidence 0.8157895 and 0.8378378. From Rule 1, P(Alarm clock bakelite red|Alarm clock bakelike green) - The probability of purchasing Alarm clock bakelike red is 81.57% if the product purchased is Alarm clock bakelike green. With lift we can check the probability of Alarm clock bakelike red purchased. P (Alarm clock bakelike red) = Confidence/Lift = 0.8157895/8.576814 = 0.09511 = 9%

****

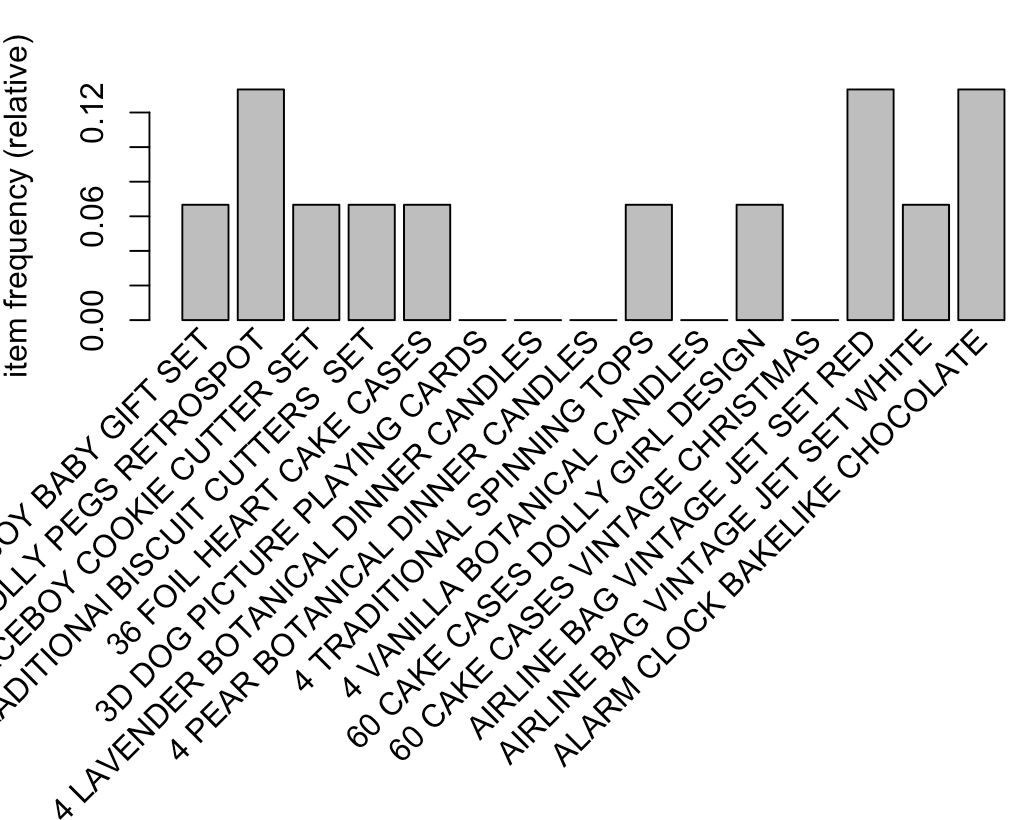
**Germany:** Item Frequency for first 15 transactions is shown below

****

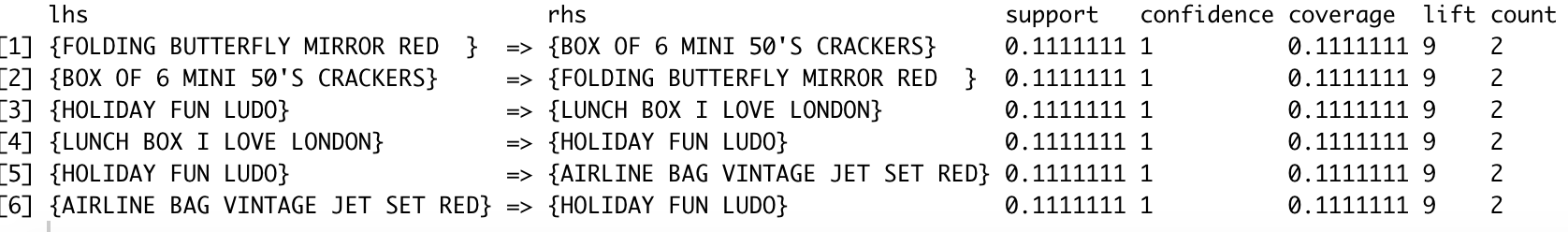
By Inspecting the first six rules, sorted by their lift. Rule 1 & 2 has the highest lift(15.379808) with confidence 0.8076923 and 0.87500.From Rule 1, P(Set/6 Red Spotty Paper cups|Set/6 Red Spotty Paper Plates) - The probability of purchasing Set/6 Red Spotty Paper cups is 80.76% if the product purchased is Set/6 Red Spotty Paper Plates. With lift we can check the probability of Set/6 Red Spotty Paper cups purchased. P (Set/6 Red Spotty Paper cups) = Confidence/Lift = 0.8076923/15.379808 = 0.05251641 = 5%

****

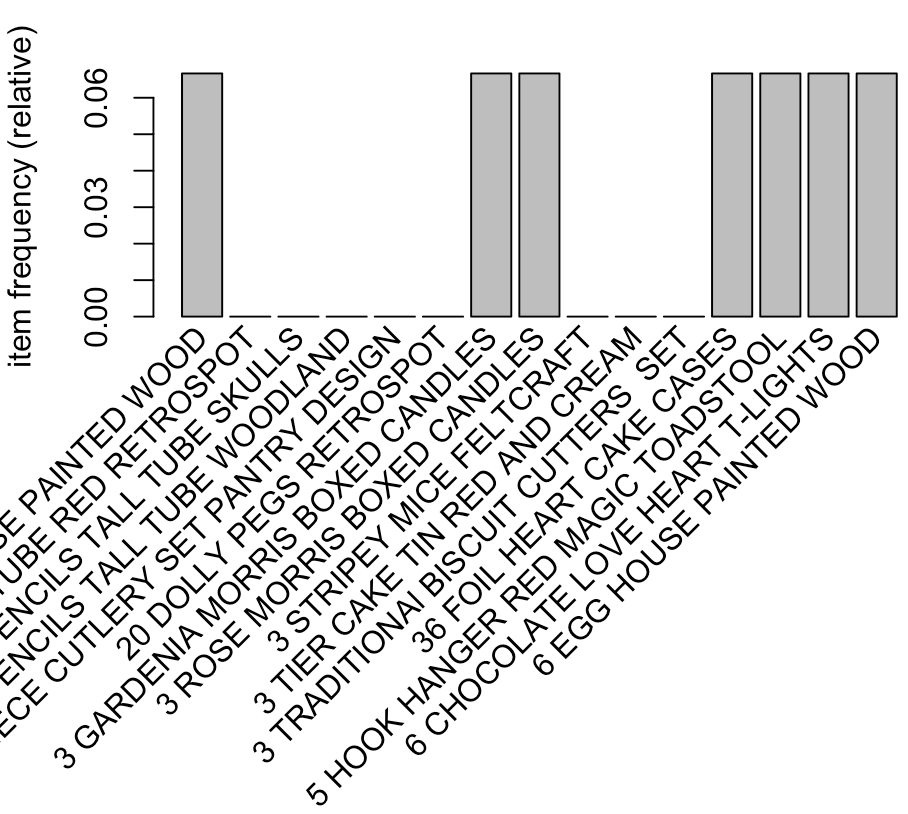
**Denmark:** Item Frequency for first 15 transactions is shown below

****

By Inspecting the first six rules, sorted by their lift. All 6 Rules have the highest lift(9) with confidence 1.From Rule 1, P(Box of 6 Mini 50’s crackers|Folding butterfly mirror red) - The probability of purchasing a Box of 6 Mini 50’s crackers is 90% if the product purchased is Folding butterfly mirror red. With lift we can check the probability of Box of 6 Mini 50’s crackers purchased. P(Box of 6 Mini 50’s crackers) = Confidence/Lift = 1/9 = 0.1111 = 11%.

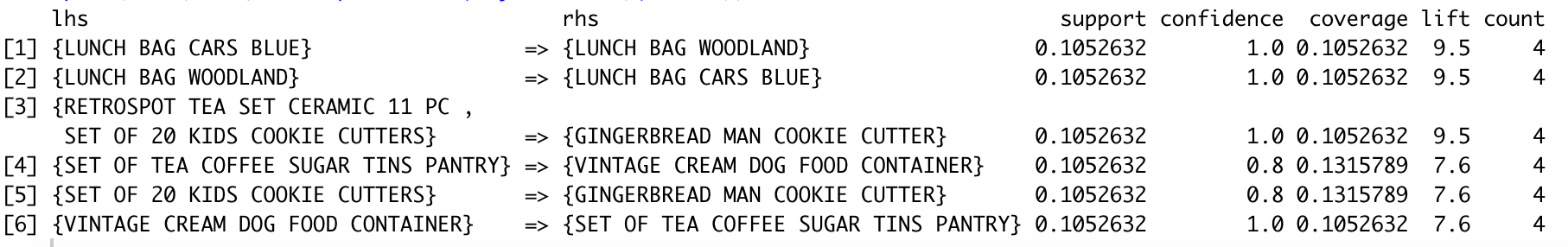
****

**Italy:** Item Frequency for first 15 transactions is shown below

****

By Inspecting the first six rules, sorted by their lift. Rule 1,2,3 has the highest lift (9.5) with confidence 1. From Rule 1, P (Lunch Bag woodland|Lunch bag cars blue) - The probability of purchasing Lunch Bag woodland is 95% if the product purchased is Lunch bag cars blue

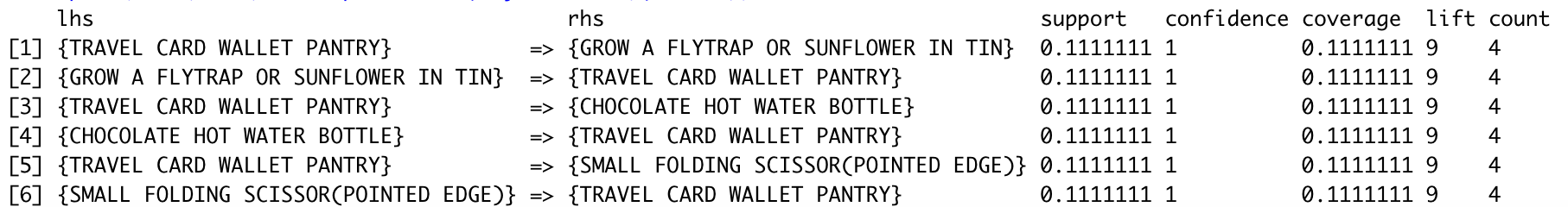
With lift we can check the probability of Lunch Bag woodland purchased. P (Lunch Bag woodland) = Confidence/Lift = 1/9.5 = 0.1052 = 10.5%

****

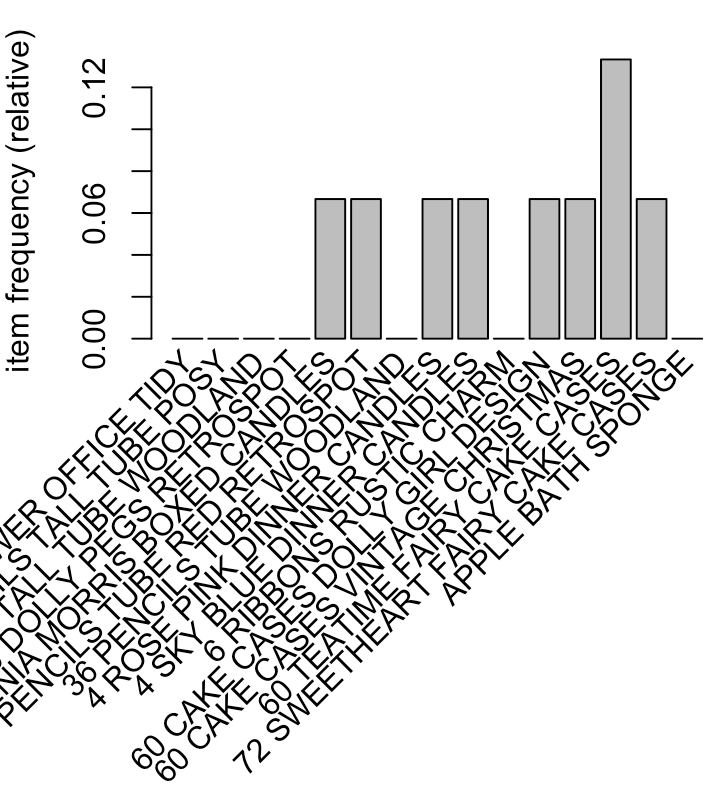
**Norway:** Item Frequency for first 15 transactions is shown below

****

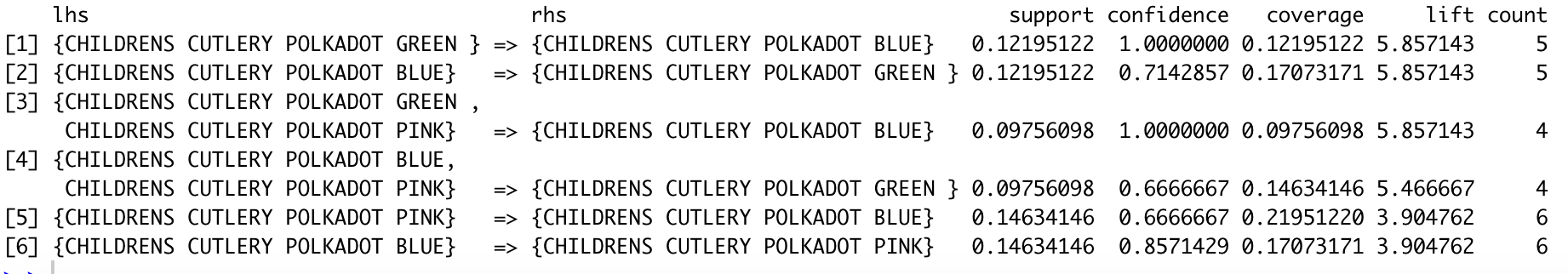
By Inspecting the first six rules, sorted by their lift. All 6 rules have the highest lift(9) with confidence 1.From Rule 1, P(Grow a flytrap or sunflower in tin|Travel card wallet pantry) - The probability of purchasing Grow a flytrap or sunflower in tin is 90% if the product purchased is Travel card wallet pantry. With lift we can check the probability of a Grow a flytrap or sunflower in tin is purchased. P (Grow a flytrap or sunflower in tin) = Confidence/Lift = 1/9 = 0.11= 11%

****

**Finland:** Item Frequency for first 15 transactions is shown below

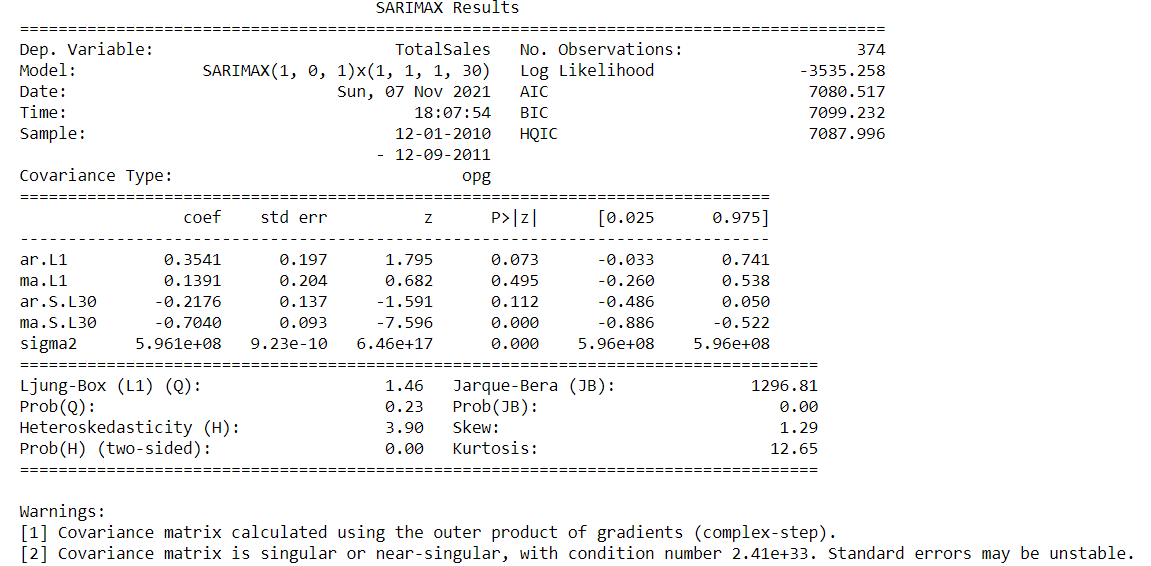
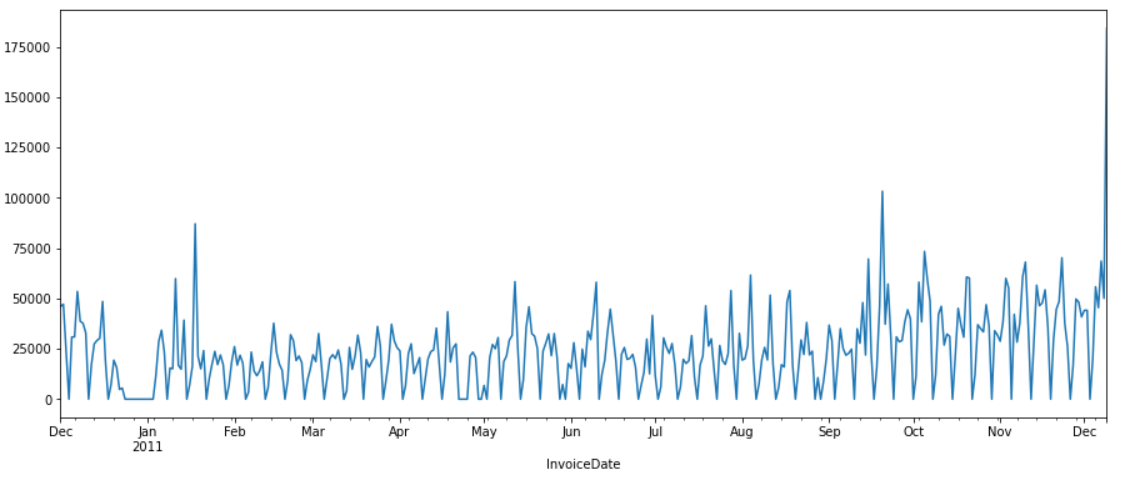
****

By Inspecting the first six rules, sorted by their lift. Rule 1,2 & 3 has the highest lift (5.857143) with confidence 1, 0.714 and 1 respectively. From Rule 1, P (Childrens cutlery polkabot blue|Childrens cutlery polkabot green) - The probability of purchasing Childrens cutlery polkabot blue is 58.57% if the product purchased is Childrens cutlery polkabot green. With lift we can check the probability of Childrens cutlery polkabot blue purchased. P (Childrens cutlery polkabot blue) = Confidence/Lift = 1/5.857143 = 0.17 = 17%

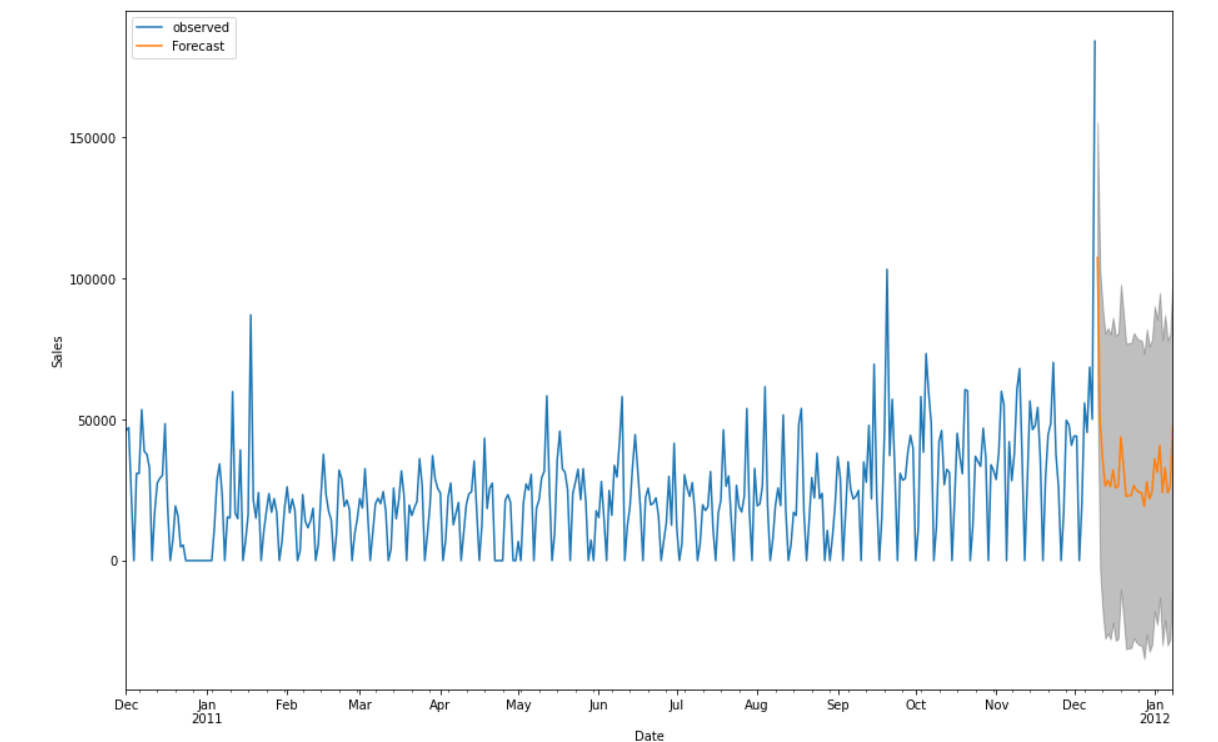
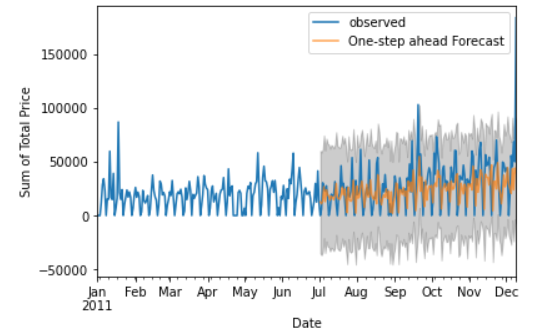
****

**6.6 Time Series:**

Stats models were used to calculate different observations (AIC, BIC). For future prediction, we have considered last 6 months data as training data to predict 2012 sale value. By 2011, the total no of sales was more than our predicted sales, after running the training data for the last 6 months of 2011



Different labels of Total Sales between December 2010 to December 2011 to predict the January 2012 Total Sales.



The predicted Sales showed a compatible growth over the month of January. The time series plot for the returns is shown in fig. (above left). This plot exhibits no periodicity, but we do have a trend effect due to the random walk nature. After transformation using log returns, the trend effect is eliminated as seen in fig. (above right). Thus, we do not have to apply any differencing technique to make the series stationary; therefore, the integration order is zero. Generally, the volatility of the series is fairly uniform for the years 2010 to 2011. For the years of 2010 to 2011, volatility of the series was highly non-uniform and more pronounced around 2011.

## **Results:**

* Majority of the purchases are made from the UK when compared to other countries.
* The top seller products are Dotcom Postage, Regency Cake, White Hanging, Party Bunting and top 5 popular products over time with StockCode are 22086, 22197, 23084, 84879 and 85099B.
* The Trend of sale during the end of year is much higher compared to any of the months of the year. It is recommended to have high inventory during the end of the year to plan for purchases during the end of the year and probably limit inventory for the rest of the year.
* Revenue over time during the beginning of the year is around 50,000$ and revenue towards the end of the year is doubled and number of orders being placed is around 1500 and double during the end of the year
* Customer ID 14647 has the highest Monetary (​​279489.02), Customer ID 17841 has the highest Frequency (7812) and Customer ID 12791, 13747, 14729, 16583, 17908, 17968 and 18074 have the highest recency (373)
* Customer ​​14911 has the highest Average order value (**AOV**) that tracks the average dollar amount spent each time a customer places an order and Customer 14096 has the highest **Customer LifeTime Value**
* Target cluster 2 customers for high **Monetary** as customers who spend a lot of money are more likely to spend money in the future and provide high value to business
* Target cluster 3 customers for high **Frequency** as first-time customers may be good targets for follow-up advertising to convert them into more frequent customers.
* Target cluster 1 customers for high **Recency** as customers who recently made a purchase will still have the product on their mind and are more likely to purchase the product again.
* The top variables Monetary and Frequency have **high predictive power** while CLTV and Recency has low predictive power in classifying the total sales
* Target cluster 2 for highest number of **items sold** and target less on Cluster 1 customer as it has least number of items sold
* Sales is less than **average Sales** predicted from all the classification modelswhichgives a granular view at the results of every sale. Measuring average sales by customer can deliver useful insights such as how many dollars customers are spending at the point of sale, and how that data compares to historical data.
* Among the few European countries: The probability of purchasing **Lunch Bag woodland** is 95% if the product purchased is **Lunch bag cars blue** has the highest probability while the probability of purchasing **Childrens cutlery polkabot blue** is 58.57% if the product purchased isChildren's cutlery polkabot green is the least probability
* The model accuracy predicts the customer is to be purchasing a low value basket - that determines the amount spent by shopper at one transaction
* Number of purchases tends to increase during the holiday season. It is not unusual that gift products demand is higher in the final quarter of the year. Additionally, there might be a possibility that the retail customers tend to stockpile gift supplies during these times as the price of products might be relatively down.
* As per our dataset, suppose the store started its operations from December 2009, the repeat purchase rate seems to be trending upwards. In a way, enhancing the customer acquisition strategy and retaining those customers for a longer period can prove to be more beneficial.
* Stores having **a wide variety of products** **catalogue** tend to have **a higher repeat purchase rate**. Customers have varied options to choose from the product offerings, increasing the repeat purchases.

## **Limitation:**

An imbalance exists in the location data, with the majority of purchases from the UK and also the dataset is available for only 13 months from 2010 to 2011. The UK could be removed from the dataset in order to model other countries, however an alternative approach could be to build country specific models. Similarly, the dataset provided for this assignment contains an imbalance in that there is a bimodal distribution with the majority of baskets clustering at a lower value and a small number of higher value baskets. For model training, the highest value baskets were removed as "outliers", however this has removed the most valued customers from the dataset. An alternative approach to removing the high value baskets would be to use oversampling methods to create synthetic data for higher spend baskets. This approach is computationally expensive however, and beyond the scope of this assignment.

## 

## 

## 

## 

## **Conclusions:**

With the given constraints, the data analysis section of this work identified a number of useful metrics in the data, including:

The frequency of invoices throughout the year. Customer segmentation based on RFM analysis to target existing and new customers. Total mean sales prediction based on variables monetary, frequency and customer lifetime value. Item Frequency and items bought together. The number of orders placed per customer, and per country. The value of baskets per customer and per country.The contents of the highest spending customer baskets.

The observation of 'per customer baskets' and their value enabled the classification of customers based on their level of spend. This information enabled the construction of a proof of concept machine learning model to predict new customers' level of spend. This model could be used to predict income levels for the business and combined with apriori modeling.

A further model was constructed to predict the country of origin of customers. This model would enable prediction of trends in geographic regions based on items purchased.

In addition, a Time Series model was also developed to enable the prediction of future Sale value, and the heat map correlation showed between the purchase of different items in the inventory.

## 

## 

## 

## 

## **Reference:**

**Online Resource:**

Online Retail Dataset| Kaggle : https://www.kaggle.com/lakshmi25npathi/online-retail-dataset

Data analytics in E-commerce Retail:<https://towardsdatascience.com/data-analytics-in-e-commerce-retail-7ea42b561c2f>

Online Retail Data analysis using R: https://rstudio-pubs-static.s3.amazonaws.com/430563\_d38c12b53d724fa6852949b1f3e4ffbf.html

**Book:**

Python Data Science Handbook: Essential Tools for Working with Data by Jake VanderPlas

ISBN-13: 978-1491912058

# Python Data Science Handbook: Essential Tools for Working with Data 1st Edition

ISBN-13: 978-1491912058

# R for Data Science: Import, Tidy, Transform, Visualize, and Model Data 1st Edition

ISBN-13: 978-1491910399

**Development References:**

https://www.udemy.com/

https://code4startup.com/

https://www.youtube.com/

<https://github.com/>