MERCER UNIVERSITY

SCHOOL OF BUSINESS

Data Mining

Report On

Predictive Sales Patterns of Online Retail

Prepared by

Monish K Pinapala

Goutham Yallapu

Table of Contents

1. Introduction

In the ever-evolving landscape of modern commerce, businesses face a myriad of challenges and opportunities in managing sales, inventory, and operational efficiency. A thorough grasp of sales trends, the capacity to predict future sales, and continual process improvement are all necessary components of a successful and competitive business. Furthermore, reducing product returns is a critical concern for organizations looking to improve customer happiness and optimize operations. Thus, we aim to determine any relationship that would explain the above through our analysis of the dataset.

1.1 Research Question:

- What are the sales patterns, and how do they vary across different dimensions, including products/items, regions, and time intervals as well as other relevant factors?

In this project, we are trying to analyse the sales patterns which helps in understanding how consumers interact with different products. It also reveals preferences and buying habits, which are crucial for effective marketing strategies. Sales variability across various regions can highlight geographical preferences and market potential. This knowledge is crucial for targeted marketing campaigns and regional product customization. Predictive analytics can be used for budget allocations, pricing strategies and marketing strategy development.

2. Literature Review:

A research paper published Daqing Chen, Sri Laing Shin & Sun Go (2012), the dataset they used in their research is of a UK based registered online business which consists of transactions occurring from 1 January 2011 to 31 December 2011. It has 406,830 observations with 9 variables and the main purpose of their research was to help businesses understand their customers better. Via RFM model-based clustering analysis they have analyzed the data based on profitability of specific categories of customers, purchase behaviour patterns, and sales patterns across products, regions, seasons etc.. and the information and methods they have used in the paper will help us understand the topic better and carry our research effectively.

Bhupathiraju and Raghavendra (2022) used prior transactional data to show the applicability of client segmentation strategies for an online retail shop. To categorize clients based on their purchasing histories and behaviours, they used RFM (Recency, Frequency, Monetary Value) analysis and K-means clustering algorithms. Predicting client lifetime value, finding product connections, and conducting more analysis on the items each group purchases are some of the major aspects that were covered. The authors come to the conclusion that careful model interpretation and meticulous data preparation are essential to guaranteeing the extraction of useful insights. In our study, we Build predictive models to forecast future customer value/spending, likelihood of churn, response to promotions, etc. Could use techniques like regression, random forests, or neural networks.

Furthermore, in this study, we aim to construct predictive models for forecasting future sales patterns and analyses. To achieve this, we will employ various techniques, including regression and random forests.

3. Data and its sources:

3.1 Data Gathering

For this study, we have taken the dataset "Online Retail" (Refer to appendix 8.1) containing up to 230,000 observations and 8 variables. These multiple data points provide a rich and diversified source of information, allowing for extensive research and valuable insights into various elements of the subject matter. In addition to the above, it can be used to identify trends and patterns, and explore relationships between variables.

3.2 Data Cleaning and Pre-processing

In the dataset, there were a total of 57,626 missing values in the dataset, where 57,007 were part of CustomerID and 619 were part of Description. After omitting the null observations, dataset is left with 172,993 observations with 8 variables. Furthermore, we have conducted a thorough examination of the dataset to ensure its integrity and reliability. During this process, we discovered the presence of extreme outliers or data points that deviated greatly from the remainder of the data distribution. These outliers were evident across multiple critical variables and were likely to skew our research, potentially biasing our conclusions. To address this issue, we removed these extreme outliers. This step was crucial to improve the accuracy and relevance of our subsequent research.

Unit Price: After a logarithmic translation, the histogram of the Log-Transformed Unit Price displays a normal-like distribution of unit prices. The most common prices cluster toward the center of the distribution, with fewer instances of extremely low or high costs. This shows that following the transformation, unit price changes within a modest range with no extreme outliers, allowing for more trustworthy statistical analysis. (Refer Figure 1)

Country: The bar graph shows that the United Kingdom leads sales compared to the next four countries, which have significantly smaller sales volumes: Germany, France, EIRE, and Belgium. The importance of sales in the United Kingdom corresponds to the company's headquarters location, implying a strong home market influence on sales success. (Refer Figure 3).

Analysis with Quantity and Price (High Quantity Low Price and Low Price High Quantity):

According to the plot, items with lower unit prices tend to sell better, and the data point trend may indicate that there is a negative correlation between the quantity ordered and the unit price. The dense cluster of points towards the bottom of the dependent variable and towards the higher end of the independent variable indicates that there is a concentration of data points at lower unit prices and higher quantities. The unit price usually decreases as the quantity increases. (Refer Figure 4)

3.3 Reducing Data Dimensions

In addition to above, we have removed all observations under UnitPrice variable who's value is less than or equal to 0.01 as the data is missing the prices for these records and needs to be omitted for a clean dataset. We have created 8 new dummy variables, one of the dummy variable we have created is InvoiceDummy, which is a binary variable, containing values of 0 and 1, where 1 represents the InvoiceIds that start with C, which means cancelled orders and 0 represents successful orders. We created this variable to separate the sales with returns/cancellations from the whole population of orders. Another dummy variable is HighSale. This variable represents 80,297 as low sales when compared with 80,188 as high sales is almost 1:1, which is a fairly balanced distribution. In this case, we can conclude that the dataset is not under or over sampled. Additionally, we have created Log_Quantity and Log_UnitPrice for Furthermore, this balanced dataset could be utilized directly to train and predict the classification models.

To conclude with, we have created Seasons which are divided into 4 groups, namely Winter, Spring, Summer, Fall to narrow down trends and understand customer's seasonal spending behaviour. Grouped country to group UK as one and the rest of the countries as Others as UK is where the company is based out of and the data would be skewed. To conclude with, Cust.Frequency and Item.Frequency which represents frequency of occurrence in the Customer ID's and Description respectively. After creating the new variables, we are left with 160,485 observations with 14 variables which can be partitioned to predict the dataset.

3.4 Data Mining Task and its importance in Marketing:

The predictive data mining task in this project is focused on predicting sales patterns based on several characteristics such as product categories, geographies, time intervals, and other relevant factors. This is critical in marketing analytics since it aids in analyzing consumer behaviors, optimizing product distribution, and anticipating future sales patterns. Predictions will be defined through analytical models, which may apply techniques such as regression analysis to determine correlations between sales and influencing variables. These forecasts are critical for strategic marketing planning and decision-making.

3.5 Data Partition:

We have parted data into 60% Training with 96,291 observations and 40% Validation with 64,194 observations with 16 Variables each (8 original and 8 dummy Variables). By using most of the data for training, we ensure that our models have a comprehensive learning phase, capturing the underlying trends and patterns across various variables while the validation data acts as new, unseen data for the models, enabling us to assess how well our predictions generalize beyond the data on which the model was trained.

4. Methodology

Dependent Variable: Log_Quantity

Independent Variable: Log_UnitPrice + Season + GroupedCountry + Item.Frequency + Cust.Frequency

4.1 Multiple Linear Regression:

Multiple Linear Regression (MLR) is a statistical technique used to model the relationship between one dependent variable (outcome) and two or more independent variables (regressors). This technique allows for the evaluation of the impact of individual variables on sales while controlling for other factors, providing insights into how different factors influence sales patterns in the dataset. Because of its ability to examine numerous influences on a target variable at the same time, this method is critical in marketing analytics.

The regression analysis (Refer Output 1) results indicate several significant factors affecting Log_Quantity (the log-transformed quantity of sales). A negative coefficient for Log_UnitPrice (-0.579) suggests that higher prices lead to lower sales quantities. Seasonal variations show a positive impact on sales, with Spring, Summer, and Winter associated with increases in Log_Quantity, as indicated by their positive coefficients (0.121, 0.102, and 0.041, respectively). Sales in the UK (GroupedCountryUK) have a significant negative impact on quantities sold, with a coefficient of -0.667. The frequency of item and customer transactions also plays a role, with moderate and frequent transactions positively influencing sales quantity. This model's R^2 value of 0.272 implies that about 27.2% of the variance in Log_Quantity is explained by these independent variables.

4.2 Random Forest Model:

A Random Forest model is a type of machine learning algorithm that is especially good at handling complex data sets for tasks like classification (identifying what group something belongs to) and regression (predicting a number). As our dependent variable is a continuous numerical variable, we have performed regression analysis. It helps in determining the significance of each variable in prediction, providing significant insights into complex linkages within the data. In our case, it provided a reliable method for forecasting sales patterns while considering a wide range of contributing elements.

The model explains 19.56% of the variation in Log_Quantity. The mean of squared residuals, which measures the average squared difference between observed and anticipated values, is 0.9622. This shows that, while the model has some predictive power, a considerable percentage of the variance in Log_Quantity remains unexplained, emphasizing potential areas for model development or research of other explanatory factors. (Refer Output 2)

4.3 Regression Tree:

Regression Trees are a sort of decision tree that is used to predict continuous outcomes. In our model, we used Regression Trees to forecast sales amounts which enabled us to view and comprehend how various aspects such as product kind, geography, and time intervals influence sales quantities.

In tree construction, the model mostly used GroupedCountry and Log_UnitPrice. The root node error was 1.1962, which is a measure of the overall variance in the dependent variable that the model is attempting to explain. The complexity parameter (CP) values and accompanying splits indicate how the model accuracy improves with each split. The n=96291 in the tree represents the number of observations used. The importance of nation grouping and unit price in calculating sales quantities is highlighted in this model. (Refer Output 3)

4.4 Recency, Frequency and Monetary Model

Recency, Frequency, and Monetary (RFM) model is a marketing analysis method that examines how recently a consumer has purchased (Recency), how frequently they purchase (Frequency), and how much the customer spends (Monetary). It allowed us to target marketing initiatives and prioritize high-value customers, which was critical for optimizing marketing efforts and improving customer relationship management more effectively.

RFM Segment Characteristics demonstrates a clear trend among RFM scores ranging from 3 to 15. Lower scores indicate that these customers are less engaged, as seen by broader recency ranges and minimal frequency and monetary values. As RFM scores rise, there is a significant decrease in recency and an increase in frequency and monetary values, indicating increased consumer engagement and value. Customers with the highest RFM scores have recent contacts, frequent purchases, and high spending, indicating that they are the most valuable group. This distinction assists in adapting marketing techniques to different target categories. (Refer Output 4)

5. Empirical Results:

Application of Techniques:

We have applied three predictive analytics techniques: Multiple Linear Regression, Random Forest Model, and Regression Tree which were chosen for their robustness and applicability for the dataset.

Challenges:

In our initial model iterations, we tested total sales as the dependent variable. However, we faced issues with multicollinearity among the independent variables, which could distort the statistical significance and reliability of the model's coefficients. Then, we tried testing for the Quantity. Also, a notable challenge in our research was highly dispersed nature of the sales quantity data. This skewness could have impacted the accuracy of our models. To mitigate this issue, we applied a logarithmic transformation to the quantity variable, resulting in a Log_Quantity dependent variable. In the similar manner, also we have applied a log transformation to the Unit price variable, resulting in Log_UnitPrice predictor.

Model Iterations:

In our research, the Multiple Linear Regression model was iterated numerous times to account for multicollinearity and ensure a normal distribution of residuals. We changed the independent variables and experimented with various transformations. Iterations in the Random Forest Model focused on determining the ideal number of trees to balance accuracy and overfitting, modifying parameters such as maximum tree depth. In the similar way, the Regression Tree methodology required numerous attempts to identify the optimal tree complexity, cutting overly specialized branches to improve generalizability. Each iteration intended to improve model accuracy and interpretability in response to the particular problems of our research.

6. Conclusion and Recommendations:

6.1 Results Conclusion

Models	Variance Explained (R^2)	Root Mean Square Error (RMSE)
Multiple-Linear Regression	**27.2**	**0.93**
Random Forest	19.5	0.98
Regression Tree	23.7	0.95

Multiple Linear Regression model has the highest R^2 of 27.2%, indicates a better fit in explaining the variance in sales, paired with a lower RMSE of 0.93. Random Forest model, although less effective in explanation of the variance at 19.5%, still holds reasonable RMSE of 0.98. To conclude with, Regression Tree model explaining 23.7% of the variance and having an RMSE of 0.95, making it moderately effective in prediction accuracy. In our model comparison, we chose Multiple Linear Regression (MLR) as our primary predictive model because it has a reduced Root Mean Square Error (RMSE) of 0.93 when compared to the other models. A lower RMSE suggests a more precise model-to-data fit, making MLR the most dependable choice for our predictive analytics.

Marketing Conclusion:

From a marketing standpoint, the most important factor influencing the quantity of products sold is the product's unit price (Log_UnitPrice), according to the Random Forest model's Variable Importance Plot. This implies that driving sales volume requires a pricing strategy. The noteworthy contribution of GroupedCountry suggests that sales quantity is influenced by geographical market segmentation as well, suggesting that pricing and marketing strategies may have differing effects on different regions.

The lesser importance of Item.Frequency and Cust.Frequency highlights that while the frequency of item purchases and customer buying habits do contribute to sales volume, they are not as influential as price and location. Seasonal variations also play a role, indicating that marketing campaigns and stock levels could be adjusted seasonally to optimize sales.

To put it simply, marketers should concentrate on competitive pricing tactics and target their advertising at particular geographic areas. Furthermore, marketing campaigns can be improved for increased customer engagement and sales performance by having a better understanding of seasonal trends and customer purchase patterns.

6.2 Recommendations:

Based on our observations from the model's businesses should take into account the following strategies in order to comprehend sales patterns and their variations across various dimensions:

- **Dynamic Pricing:** Modify prices based on how sensitively demand is met, possibly with the aid of tools like dynamic pricing algorithms.
- **Seasonal Marketing:** Create stock plans and focused marketing campaigns for various times of the year.
- **Regional Targeting:** Adapt goods and advertising strategies to local markets while taking into account the distinctive characteristics of various locales.
- **Retention and Loyalty:** Promote recurring purchases with frequent shopper discounts or loyalty programs.

- **Data-Driven Decision Making:** To identify more precise patterns and trends, keep fine-tuning the sales models using more detailed data from various time periods, geographical areas, and product categories. Businesses can better align their sales strategies with the demand patterns of their clientele by utilizing these insights.

7. References

Chen, Daqing, Sai Laing Sain, and Kun 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 19 (2012): 197-208.

Bhupathiraju, Gowtham Varma, and V. R. T. S. Raghavendra. "Data mining for the online retail industry: Customer segmentation and assessment of customers using RFM and k-means." (2022).

8. Appendix:

8.1 Data Dictionary

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

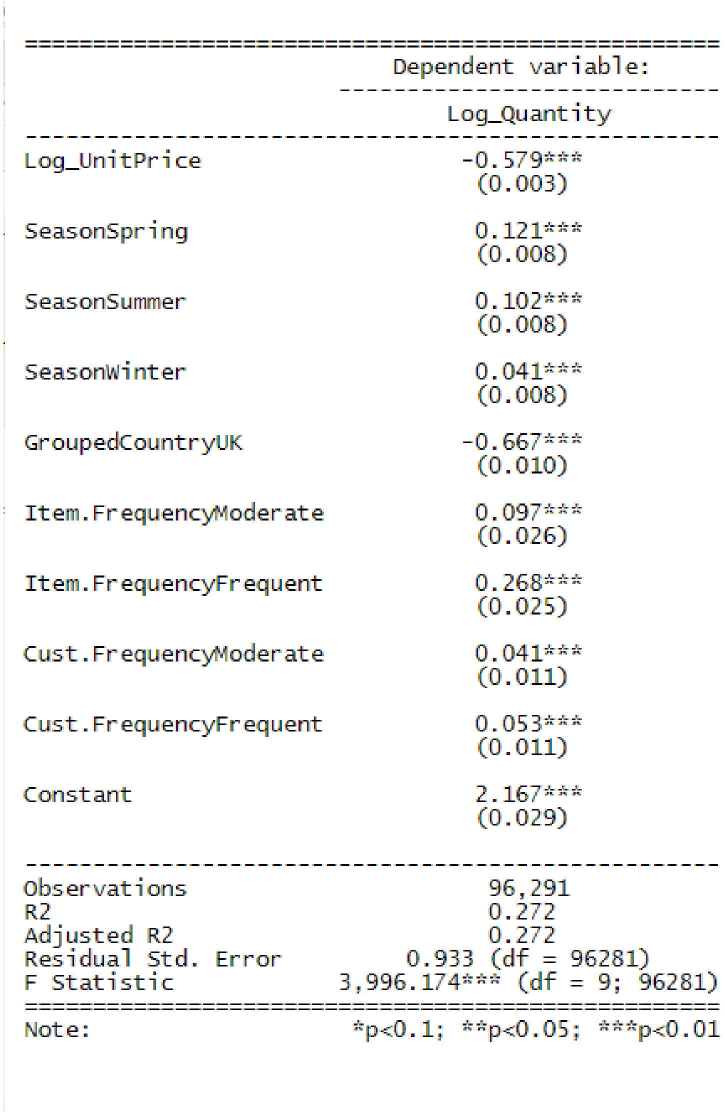
Dummy variables

Variable	**Type**	**Description**
Grouped Country	Categorical	Countries are grouped UK as one and the rest of the countries as Others
Season	Categorical	Months were divided into 4 groups such as Winter, Spring, Summer, Fall
Invoice Dummy	Categorical	Invoice are categorized into 2 groups where 1 represents the InvoiceIds that start with C, which means cancelled orders and 0 represents successful orders.

Cust.Frequency	Categorical	Customers are grouped into 3 categories based on their frequency of occurrence such as Rare, Moderate and Frequent
Item.Frequency	Categorical	Descriptions are grouped into 3 categories based on their frequency of occurrence such as Rare, Moderate and Frequent
Log_Quantity	Numerical	Log transformed values of Quantity
Log_Unitprice	Numerical	Log transformed values of Unit Price

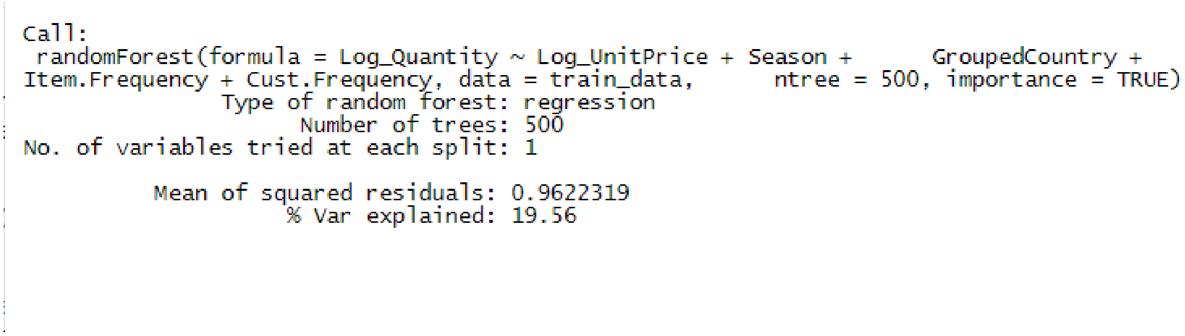
8.2 Result outputs

Output 1: Multiple Linear Regression



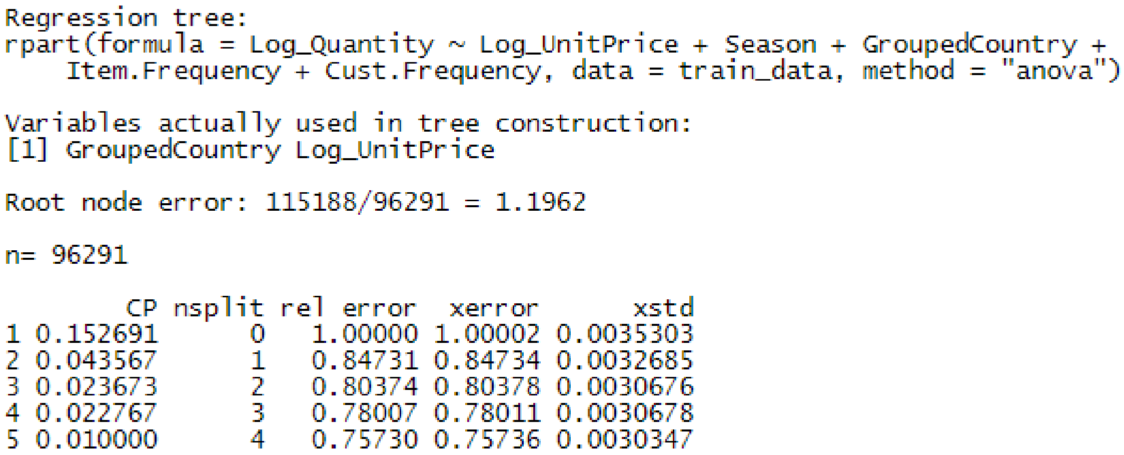
```
===================================================
                        Dependent variable:
                    ---------------------------
                            Log_Quantity
---------------------------------------------------
Log_UnitPrice                 -0.579***
                               (0.003)

SeasonSpring                   0.121***
                               (0.008)

SeasonSummer                   0.102***
                               (0.008)

SeasonWinter                   0.041***
                               (0.008)

GroupedCountryUK              -0.667***
                               (0.010)

Item.FrequencyModerate         0.097***
                               (0.026)

Item.FrequencyFrequent         0.268***
                               (0.025)

Cust.FrequencyModerate         0.041***
                               (0.011)

Cust.FrequencyFrequent         0.053***
                               (0.011)

Constant                       2.167***
                               (0.029)

---------------------------------------------------
Observations                    96,291
R2                              0.272
Adjusted R2                     0.272
Residual Std. Error       0.933 (df = 96281)
F Statistic           3,996.174*** (df = 9; 96281)
===================================================
Note:                   *p<0.1; **p<0.05; ***p<0.01
```

Output 2: Random Forest Model



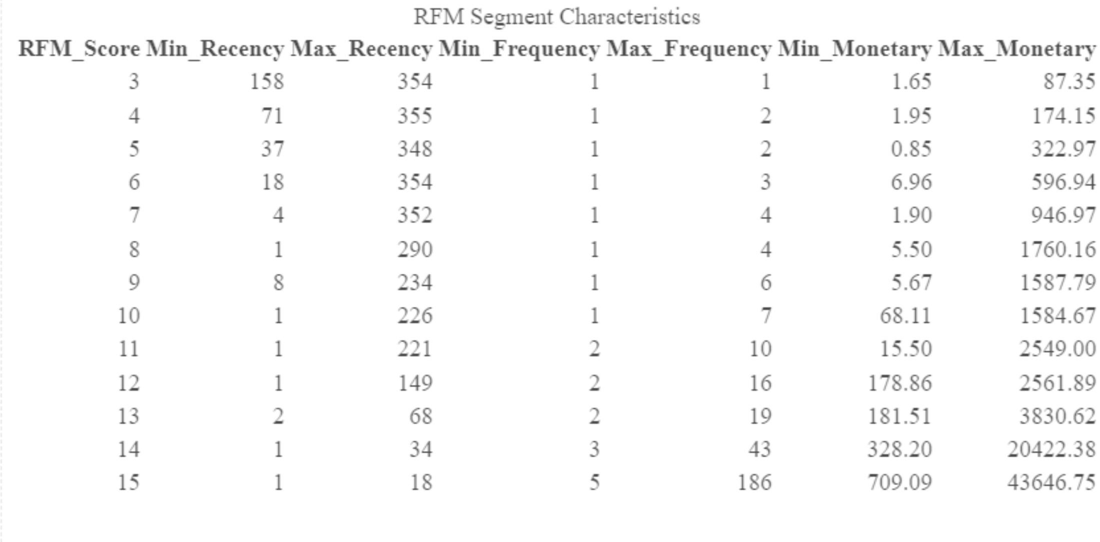
```
Call:
 randomForest(formula = Log_Quantity ~ Log_UnitPrice + Season +      GroupedCountry +
Item.Frequency + Cust.Frequency, data = train_data,      ntree = 500, importance = TRUE)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 1

          Mean of squared residuals: 0.9622319
                    % Var explained: 19.56
```

Output 3: Regression Tree



```
Regression tree:
rpart(formula = Log_Quantity ~ Log_UnitPrice + Season + GroupedCountry +
    Item.Frequency + Cust.Frequency, data = train_data, method = "anova")

Variables actually used in tree construction:
[1] GroupedCountry Log_UnitPrice

Root node error: 115188/96291 = 1.1962

n= 96291

        CP nsplit rel error  xerror      xstd
1 0.152691      0   1.00000 1.00002 0.0035303
2 0.043567      1   0.84731 0.84734 0.0032685
3 0.023673      2   0.80374 0.80378 0.0030676
4 0.022767      3   0.78007 0.78011 0.0030678
5 0.010000      4   0.75730 0.75736 0.0030347
```

Output 4: Recency Frequency Monetary



RFM Segment Characteristics

RFM_Score	Min_Recency	Max_Recency	Min_Frequency	Max_Frequency	Min_Monetary	Max_Monetary
3	158	354	1	1	1.65	87.35
4	71	355	1	2	1.95	174.15
5	37	348	1	2	0.85	322.97
6	18	354	1	3	6.96	596.94
7	4	352	1	4	1.90	946.97
8	1	290	1	4	5.50	1760.16
9	8	234	1	6	5.67	1587.79
10	1	226	1	7	68.11	1584.67
11	1	221	2	10	15.50	2549.00
12	1	149	2	16	178.86	2561.89
13	2	68	2	19	181.51	3830.62
14	1	34	3	43	328.20	20422.38
15	1	18	5	186	709.09	43646.75

8.3 Data Visualizations:

Figure 1: Unit Price

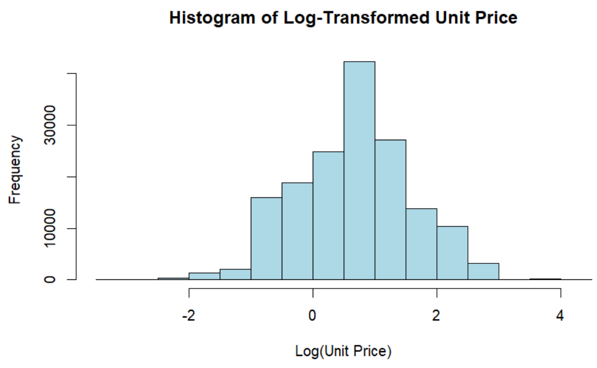


Figure 2: Sale Patters per Day of the week

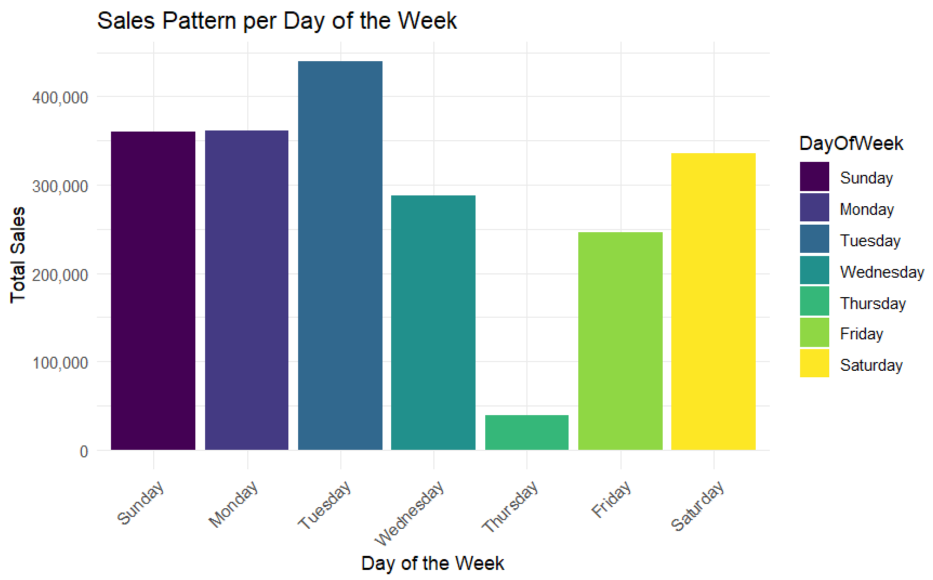


Figure 3:

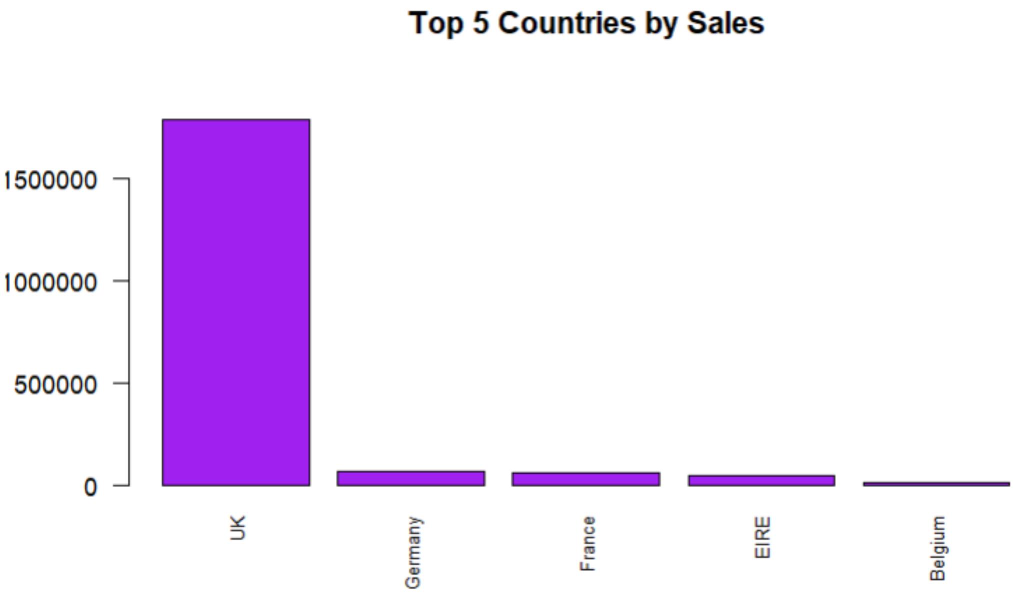


Figure 4: Sale Patterns by region

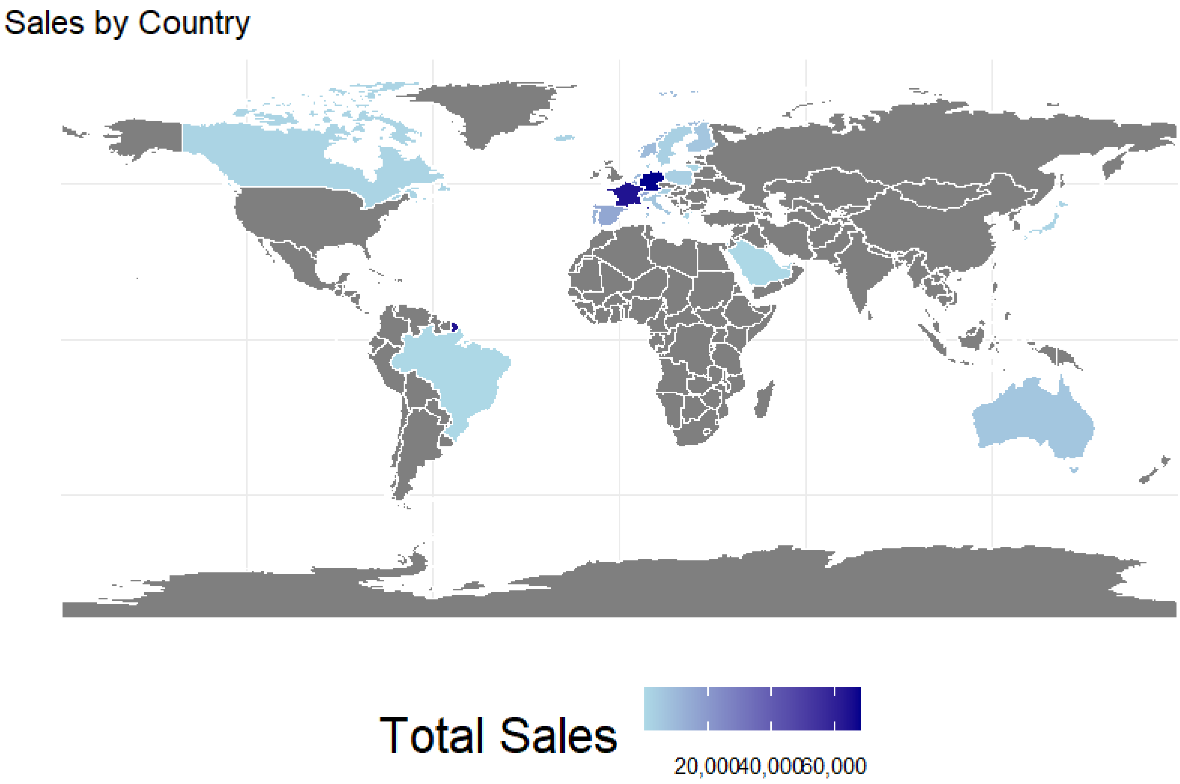


Figure 5: Sale patterns by Quantity and Price

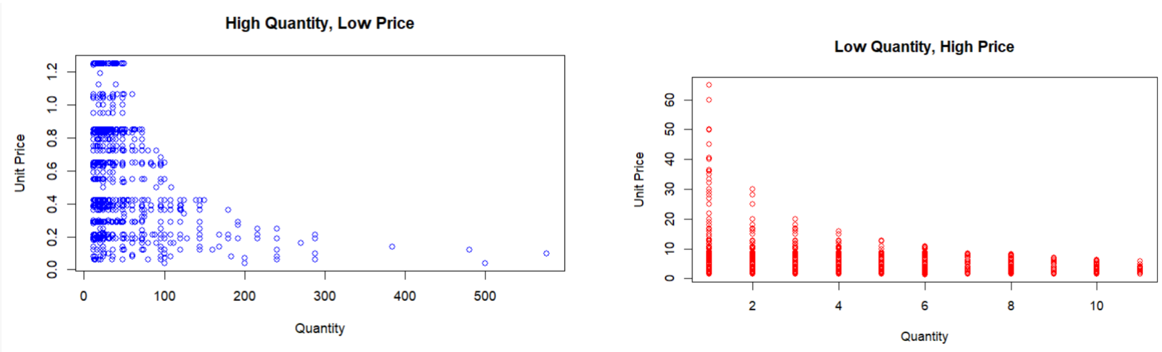
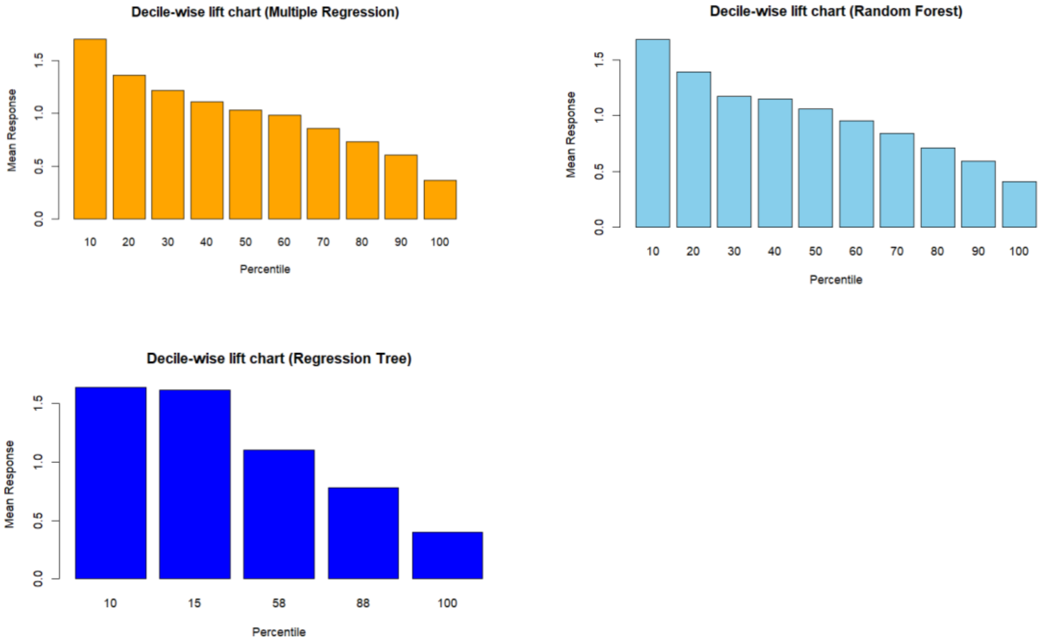


Figure 6: Decile Charts


Decile-wise lift chart (Multiple Regression)
Mean Response
Percentile
Decile-wise lift chart (Random Forest)
Mean Response
Percentile
Decile-wise lift chart (Regression Tree)
Mean Response
Percentile