Online Retail and Customer Spending Analysis Using R

1. **Introduction**

In today’s data-driven landscape, data has emerged as the new gold, shaping the way we perceive and interact with the world. As we navigate the age of information, understanding data patterns becomes paramount. Not only does it provide a glimpse into what the future might hold, but it also unveils the intricate trends and behaviors that drive markets.

One such area profoundly impacted by data is online shopping. The surge in e-commerce during and after the pandemic revolutionized the way people purchase products. With the click of a button, consumers gained unprecedented ease and convenience. As a result, e-commerce has seamlessly integrated itself into nearly every company’s operations, transcending industry boundaries.

However, not all companies are equally adept at deciphering the wealth of data at their disposal. While some organizations have mastered the art of analyzing customer behaviors and sales trends, others find themselves at a crossroads. They grapple with questions like: Where do we begin? What factors should we examine to truly understand our market?

For our project, the task is to unravel the unknown for online shopping behaviors. Through different analyses such as linear regression, random forest, logistic regression, and more, we aim to offer valuable insights. We will address a few different key objectives such as customer behavior exploration. By tapping into the potential information that is not seen we will uncover patterns, preferences, and pain points. Another objective will be examining if there are trends relating to the popularity of certain products during specific seasons or months. Lastly we will be identifying key factors that are crucial for driving success. Pinpointing these factors will assist businesses in making informed decisions.

1. **Dataset**

The [dataset](https://archive.ics.uci.edu/dataset/352/online+retail), Online Retail[[1]](#footnote-1), contains all of the transactions from 12/1/2010 to 12/9/2011 for a UK based non-store online retail. The company’s products are mostly unique all-occasion gifts, and the customers are mainly wholesalers. The data contained sales and returns and originally consisted of 541,909 rows.

The dataset consists of 8 variables:

* *InvoiceNo*: Categorical, 6-digit unique number assigned to each transaction
* *StockCode*: Categorical, 5 to 9-digit unique string assigned to each product
* *Description*: Categorical, description of the product or product name
* *Quantity*: Integer, quantity of each product per transaction
* *InvoiceDate*: DateTime, date and time the transaction was generated
* *UnitPrice*: Continuous, product price per unit
* *CustomerID*: Categorical, 5 digit unique number assigned to each customer
* *Country*: Categorical, name of country where the customer resides

1. **Data Cleaning**

The main part of cleaning the dataset was to put each product into categories, based on the description. This was done through Excel. First, product descriptions that did not refer to actual products were removed. This includes items such as “bank charges” or other company expenses. The *StockCode* for the removed items mostly consisted of non-numerical codes. Second, rows were removed where the *UnitPrice* was 0.00 or the *StockCode* was N/A.

Once the rows were cleaned up, a pivot table was created to get the 3,813 unique *Descriptions* and *StockCodes*, which helped to form a new table for categories. In this new table, the *Description* was used to make new categorical variables for each product. These new variables were:

* *Category*: 25 total categories such as Food, Decoration, Gifts, Furniture
* *Item*: 1 to 2 word description of the item such as “candleholder” or “playing cards”
* *Occasion*: If the product related to occasions such as Birthday or Christmas
* *Design*: If the product had one of the common designs or brands such as Spaceboy, Dolly Girl, or polkadots

*Category* and *Item* were included for each product, but *Occasion* and *Design* were not included for every one. *Occasion* existed for 260 products and *Design* existed for 534 products.

The Category table and the cleaned dataset were put into RStudio and joined on *StockCode*. *totalSale* was added to each row, which is *Quantity \* UnitPrice*. Multiple tables were created from this table, which include 7 sub-tables of more general categories, one for sales and one for returns, and one that removed NA from *CustomerID* and *InvoiceNo*.

**Rows:**

* All Retail: 536,357
* Sales: 527,659
* Returns: 8,698

1. **General Analysis**

Total Sales: $10,024.418

Total Returns: $232,760

Total Sales - Returns: $9,791,658

Most Common *Category*: Decorations, 17.7%% of products

Most Common *Occasion*: Christmas, 69% of 260 products

Most Common *Design*: Retrospot, 17.2% of 534 products

*Country* with Most Purchased Products: United Kingdom, 91.7% of 5,421,769 products

Average *UnitPrice*: $3.26

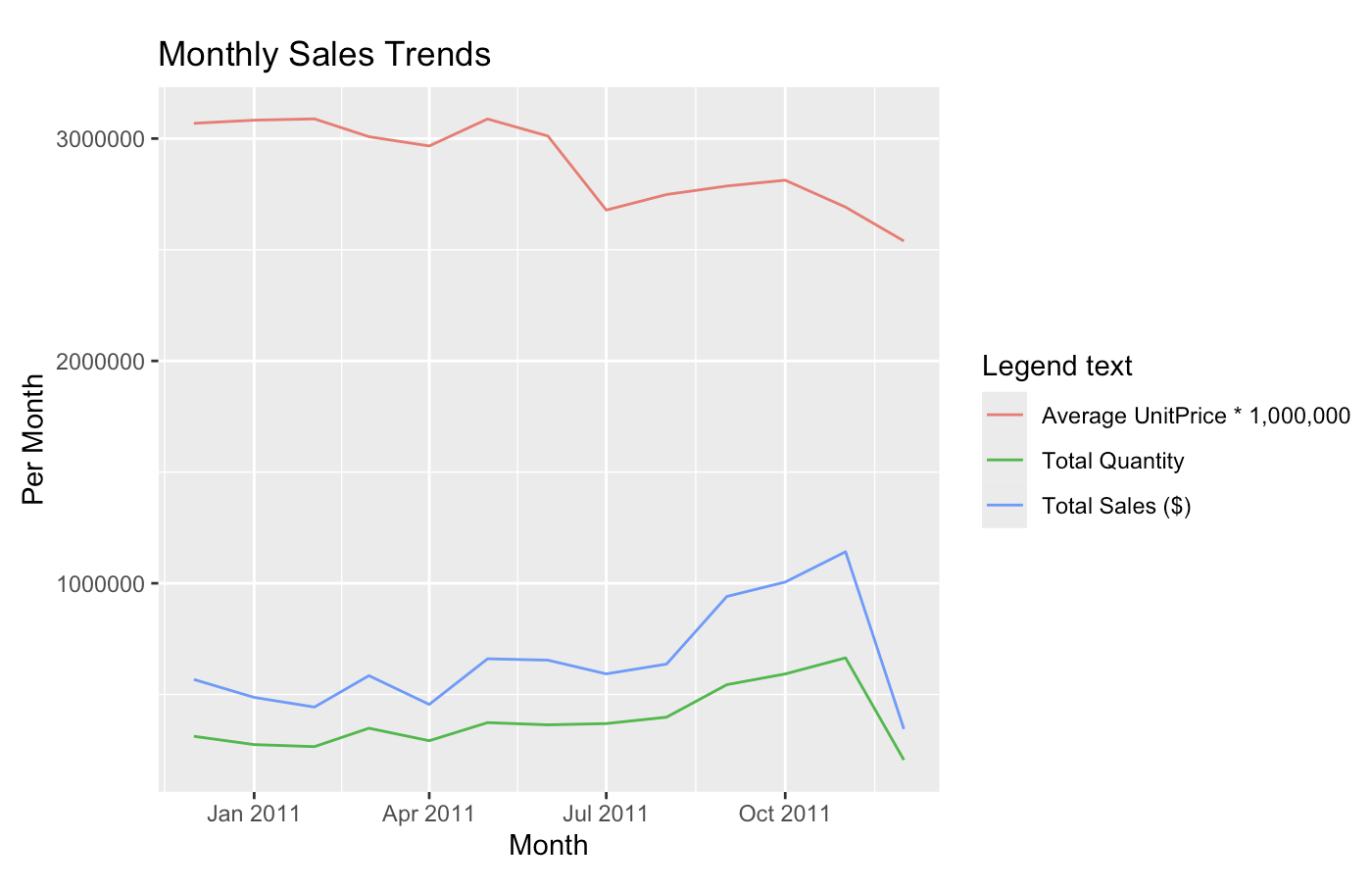
Largest Average *UnitPrice*: Lights, $5.13

Smallest Average *UnitPrice*: Incense, $0.51

The largest category, Decorations had an average *UnitPrice* of $2.82 and made up the largest % of available products, % of products sold and % of total sales, at 17.75%, 16.4%, and 15.24% respectively.

The Travel category had an average *UnitPrice* of $2.47 and made up 5.5% of available products, 11.48% of products sold, and 12.66% of total sales. On the other hand, the Jewelry category had an average *UnitPrice* of $4.93 and made up 7.83% of available products, 0.78% of products sold, and 0.48% of total sales. It may be better for the company to decrease the number of Jewelry products that are available, as these make up a large percent of the products, but make up a very low amount of sales, and increase the available travel products as these make up a lower percent of products but a larger percent of sales. Additionally, increasing the *UnitCost* of the travel products and decreasing the *UnitCost* of the jewelry products could have potential.

1. **Monthly Analysis**

****

***Figure 1***

When the number of sales in terms of *Quantity* and $, the average *UnitPrice*

decreased. There could be an opportunity for the company to increase these prices instead of decreasing them. This happened for the *Occasion* of Christmas and Easter, where the average *UnitPrice* decreased starting around 4-5 months before the holiday. Additionally, this analysis can show customers when it is the best time to buy specific products.

1. **Linear Regression**

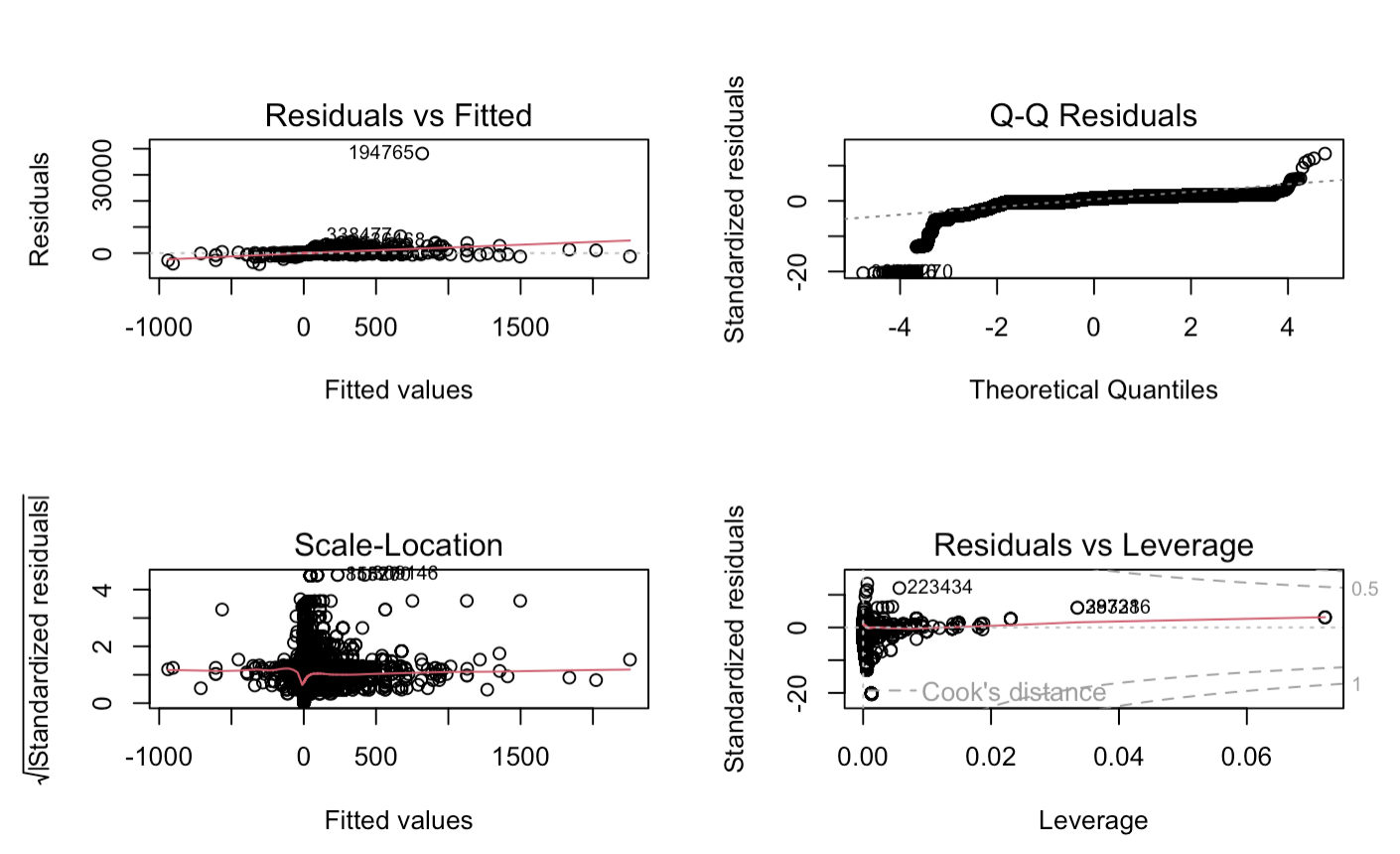
A linear regression model is developed on the dataframe, where Y is “*totalSale*” and the predictor variables are “*Quantity”, “UnitPrice*”, and “*Category*”. The equation of the regression:

lm(formula = totalSale ~ Quantity + UnitPrice + Category, data = retail2)  
  
**Data Preparation and Model Description**

In response to observed heteroscedasticity and potential influential outliers in the initial linear regression model, a refined approach was employed using the Weighted Least Squares (WLS) method. Prior to modeling, the data was filtered to adjust for influential observations, and weights were calculated to stabilize variance across the data spectrum.

The model, formulated with *totalSale* as the dependent variable, explores the relationship between total sales and several predictors: *Quantity, UnitPrice*, and various *Category* indicators. The adjusted dataset, retail2\_filtered, contains cleaned and possibly transformed values ensuring a more robust dataset for analysis.

lm(formula = totalSale ~ Quantity + UnitPrice + Category, data = retail2\_filtered, weights = weights).



***Figure 2***

**Coefficients**

The model estimates suggest significant effects from most categories on total sales, with particularly strong impacts noted for *Quantity* and *UnitPrice*. Notable category effects include negative impacts from *Category: Jewelry* and positive impacts from *Category: Music* and *Category: Travel,* among others. The Intercept and several categories demonstrated less influence or negative associations with the total sales.

**Model Diagnostics**

Weighted Residuals: The distribution of residuals has significantly improved, showcasing a narrower spread which indicates better conformity to the homoscedasticity assumption.

Residual Standard Error: Reduced to 0.5259, indicating a tighter clustering of residuals around zero, a desirable outcome reflecting less error variance in the model predictions.

Multiple R-squared: Increased to 0.6163, suggesting that the model explains approximately 61.63% of the variability in total sales, which is an improvement from the initial model.

F-statistic: The overall fit of the model is statistically significant with a p-value substantially less than 0.05, confirming that the model predictors collectively have a significant effect on the response variable.

**Conclusion and Business Insights**

This refined linear regression model, utilizing WLS, provides a more reliable analysis of factors influencing total sales. By addressing the issues of heteroscedasticity and influential points, the model offers robust insights that can aid in making informed business decisions, particularly in pricing and stock management across different product categories.

1. **Market Basket Analysis**

**Introduction and Methodology**

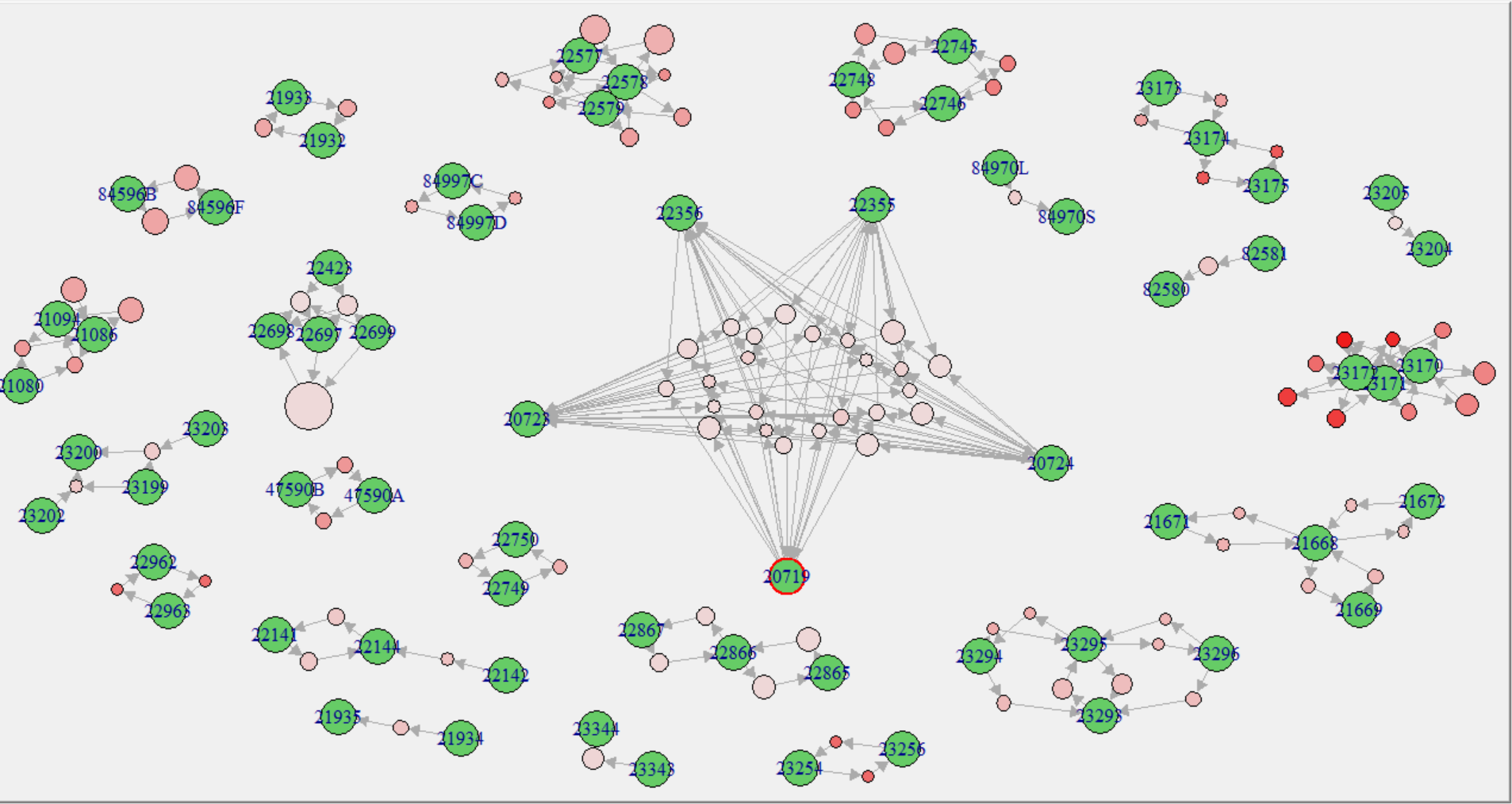
Market Basket Analysis is a data mining technique used to discover relationships between items within large datasets of transactional records. It is often used in retail to identify patterns of items frequently purchased together. This analysis is conducted using the association rule learning method, particularly through the Apriori algorithm, which is adept at identifying frequent itemsets and generating rules from them.

**Data Preparation and Analysis Description**

The dataset for this analysis comprises transaction data, where each transaction represents items bought together by a customer. The analysis results in the extraction of association rules indicating which items (or groups of items) are commonly bought in conjunction with others. The rules are evaluated and ranked using various metrics such as support, confidence, lift, and others to determine the strength and usefulness of the identified associations.

**Network Diagram Explanation**

The network diagram visually represents the relationships between different items (or itemsets) identified in the market basket analysis. Each node (circle) represents an item, labeled by its *StockCode*, and the links (edges) between them show that these items are frequently bought together.



***Figure 3***

Node Size and Color: Larger and uniquely colored nodes (like 0711 in red) might indicate key items that appear frequently across multiple transactions or rules.

Clustered Nodes: Items that are closely grouped together, connected by multiple edges, represent itemsets that are commonly bought together. For example, the large cluster in the middle of the diagram likely represents a significant set of items that are strongly associated.

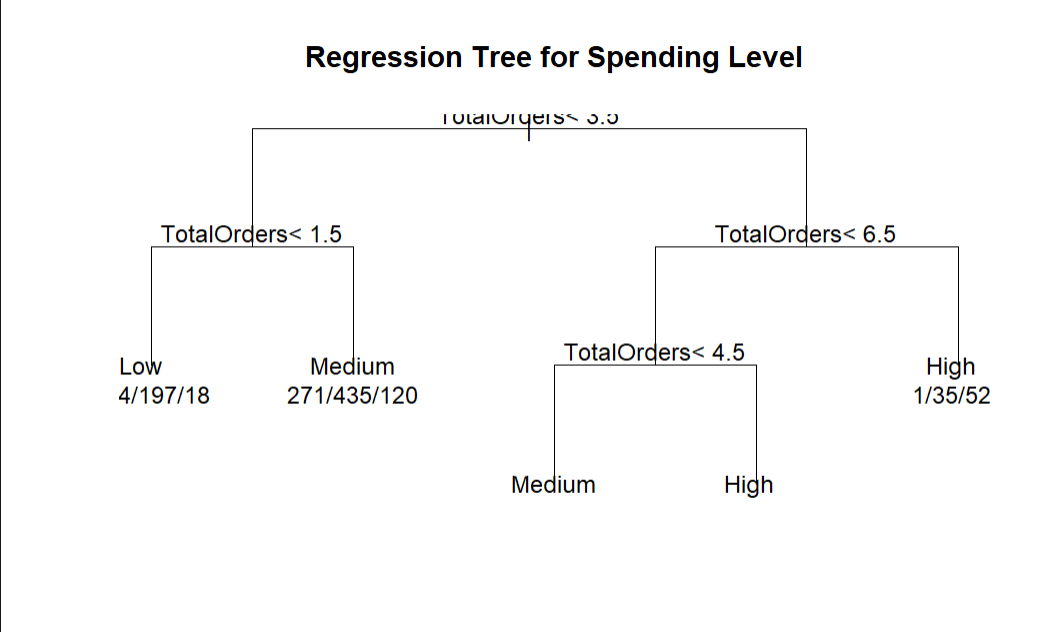
1. **Customer Spending Category Classification**

In this analysis, a multi-class classification task was performed on an online retail dataset to categorize customers into three spending categories: High Spenders, Medium Spenders, and Low Spenders. The classification was based on the total sales made by each customer. The process began with data preparation, where the total sales for each customer were calculated by aggregating the sales data grouped by *customerID.* This information was then merged back into the main dataset to create a new column representing each customer's total spending. Quantiles were used to determine the spending category thresholds, and each customer was assigned to one of the three categories based on their total spending. Feature engineering was also performed to create additional relevant features, such as the frequency of visits, which represents the number of transactions made by each customer. Subsequently, the dataset was split into training and testing sets, and several classification models were trained and evaluated. The models used *Spending\_Category* as the target variable and *Frequency\_of\_Visit*, *UnitPrice*, and *Quantity* as predictor variables. These steps enriched the dataset and tailored it for more effective predictive analysis.

The models created for this analysis includes Regression Tree, Logistic Regression, Naive Bayes Classification, and Random Forest Classifier.

**8.1 Regression Tree Analysis**

A decision tree was developed to predict customer spending levels using the “rpart” package in R, focusing on the total number of orders. The tree was pruned based on complexity parameters to optimize its performance and prevent overfitting. The pruned tree effectively segmented customers into distinct groups based on their order frequencies, with critical thresholds identified at 1.5 and 4.5 total orders for differentiating spending levels. The model demonstrated an accuracy of approximately 69.23%, indicating a reasonable level of predictiveness. The ROC curve analysis further confirmed the model's efficacy, with the curve approaching the top left corner, suggesting a good balance between sensitivity and specificity.



***Figure 4***

The regression tree generated in **Figure 4** provides clear insights into customer spending behavior:

* Low Spending Customers: Typically have fewer than 1.5 orders.
* Medium Spending Customers: Fall between 1.5 and 4.5 orders.
* High Spending Customers: Exceed 4.5 orders, with a significant likelihood of being classified as high spenders above 6.5 orders.

These findings suggest that increasing customer engagement and order frequency could potentially elevate their spending levels. Specifically, the model highlighted a substantial increase in the likelihood of high spending as customers' total orders surpass 4.5, underscoring an actionable threshold for targeted marketing strategies.

**8.2 Logistic Regression:**

In the logistic regression model, the multinom() function was used to train a multinomial logistic regression model. The confusionMatrix() function was used to evaluate the model's performance by comparing the predicted spending categories with the actual spending categories in the test data.

**Model Output:**

The model achieved an accuracy of 0.6664, indicating that it correctly predicted the spending category for approximately 66.64% of the customers in the test data.

However, it has a lower sensitivity for Medium Spenders (51.58%) compared to Low Spenders (83.33%) and High Spenders (65.04%). This suggests that the model may have some difficulty correctly identifying Medium Spenders. The specificity values are high for Low Spenders (82.86%) and High Spenders (91.61%), indicating that the model is effective in identifying customers who do not belong to these spending categories. However, the specificity for Medium Spenders is lower (75.60%).

The precision values are reasonably high for High Spenders (79.98%) and Low Spenders (70.56%), suggesting that when the model predicts a customer as belonging to these spending categories, it is often correct. The precision for Medium Spenders is lower (51.00%).

**8.3 Naive Bayes Classifier**

In the Naive Bayes classifier model, the naiveBayes() function was used to train the model. The confusionMatrix() function was used to evaluate the model's performance.

**Model Output:**

The Naive Bayes classifier achieved an accuracy of 0.5412, indicating that it correctly predicted the spending category for approximately 54.12% of the customers in the test data. This accuracy is lower compared to the logistic regression model.

The model has a high sensitivity for Low Spenders (90.55%) but lower sensitivities for Medium Spenders (46.21%) and High Spenders (26.43%). This suggests that the model is effective in identifying Low Spenders but struggles to correctly identify Medium and High Spenders. The specificity values are relatively high for High Spenders (98.43%) and moderate for Medium Spenders (70.45%) and Low Spenders (62.62%), indicating that the model is effective in identifying customers who do not belong to the High Spenders category but has some difficulty identifying customers who are not Medium or Low Spenders.

The precision is highest for High Spenders (89.67%), followed by Low Spenders (54.42%) and Medium Spenders (43.50%). This suggests that when the model predicts a customer as a High Spender, it is often correct, but it has lower precision when predicting Low and Medium Spenders.

**8.4 Random Forest Classifier**

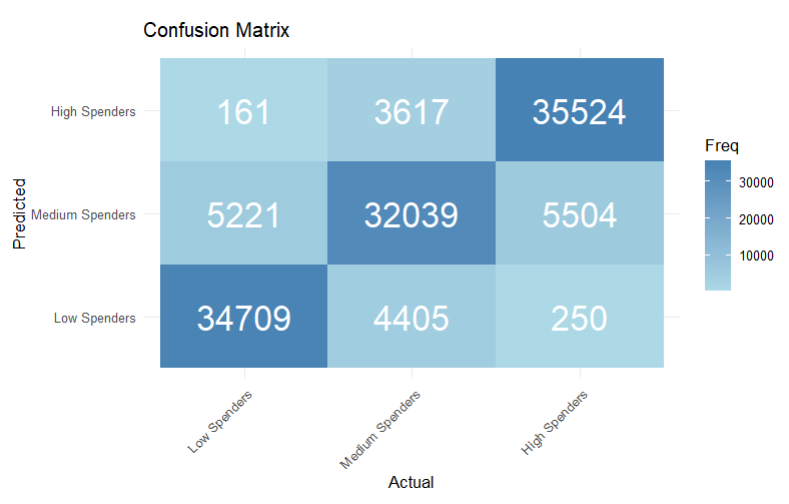
In the random forest classifier model, the randomForest() function was used to train the model, and the performance is evaluated using the confusion matrix.

**Model Output:**

The confusion matrix for this model is shown in **Figure 5**. The random forest classifier achieved an accuracy of 0.8422, indicating that it correctly predicted the spending category for approximately 84.22% of the customers in the test data. This accuracy is significantly higher compared to both the logistic regression and Naive Bayes models.

The model has high sensitivities for all three spending categories: Low Spenders (86.58%), Medium Spenders (79.98%), and High Spenders (86.06%). This suggests that the model is effective in correctly identifying customers in each spending category. The specificity values are also high for all three categories, indicating that the model is effective in identifying customers who do not belong to a particular spending category. The specificity is particularly high for Low Spenders (94.28%) and High Spenders (95.29%).

The precision (positive predictive value) is high for all three categories: Low Spenders (88.17%), Medium Spenders (74.92%), and High Spenders (90.39%). This indicates that when the model predicts a customer as belonging to a specific spending category, it is often correct.



***Figure 5***

From the confusion matrix, we can see that the model performs well in identifying low spenders (34,709 correctly classified), medium spenders (32,039 correctly classified), and high spenders (35,524 correctly classified). However, it tends to misclassify some medium spenders as low spenders (4,405 cases) and some medium spenders as high spenders (5,504 cases). Additionally, it misclassifies a small number of high spenders as medium spenders (3,617 cases) and a few low spenders as medium spenders (5,221 cases).

**8.5 Conclusion** Based on the analysis, the Random Forest Classifier emerges as the most effective model for this task, providing high accuracy, sensitivity, specificity, and precision across all categories. It is recommended for operational use in classifying customers by spending categories in similar datasets.

**9. Discussion**

The analysis conducted to understand online customer spending behavior offers invaluable insights into the dynamics of e-commerce and its intersection with consumer preferences. Leveraging various analytical techniques which include regression analysis, decision trees, and market basket analysis businesses can gain a nuanced understanding of customer segments, product performance, and key drivers of customer spending behavior. The insights gathered enable strategic decision-making, empowering businesses to optimize product offerings, refine market strategies, and allocate resources effectively. The development of predictive models opens avenues for future applications in forecasting customer behavior, facilitating agile decision-making and maintaining a competitive advantage in the rapidly evolving digital landscape. However it is crucial to acknowledge the limitations related to data quality, model assumptions, and ethical considerations as well. Addressing challenges like these ensures responsible use of data and fosters trust among consumers and stakeholders. Ultimately driving sustainable growth and long term success in the online retail industry.

In conclusion the analysis not only highlights the current trends in online shopping, which can be helpful information for both customers and companies, but also paves the way for future advancements in predictive analytics and data driven strategies. Utilizing big data, businesses can anticipate customer needs to adapt strategies to changing market dynamics and deliver personalized experiences for consumers. Algorithms can be developed to tailor towards e-commerce allowing companies to continue to integrate into the digital world. Overall this project offers what is already happening and the future of e-commerce and personalized consumer markets.

1. <https://archive.ics.uci.edu/dataset/352/online+retail> [↑](#footnote-ref-1)