Contents

[**1.0** **Introduction** 3](#_Toc197533534)

[1.1 Data Description 3](#_Toc197533535)

[1.2 Assumptions 4](#_Toc197533536)

[1.3 Hypothesis & Objectives 5](#_Toc197533537)

[**2.0** **Data Preparation** 7](#_Toc197533538)

[2.1 Data Import 7](#_Toc197533539)

[2.2 Data Cleaning & Pre-processing & Validation 8](#_Toc197533540)

[**3.0** **Data Analysis** 13](#_Toc197533541)

[3.1 Analytical Techniques Utilized 13](#_Toc197533542)

[3.1.1 Chi-Square Test 13](#_Toc197533543)

[3.1.2 Descriptive Statistics 15](#_Toc197533544)

[3.1.3 Graphical Representations 15](#_Toc197533545)

[3.2 Exploratory Data Analysis (EDA) 16](#_Toc197533546)

[3.3 Objectives 27](#_Toc197533547)

[*3.3.1* *To analyse how income levels affect spending patterns (Amount Spent) and purchase frequency (Total Purchases)* 27](#_Toc197533548)

[*3.3.2* *To identify what impacts gender spending patterns* 33](#_Toc197533549)

[*3.3.3* *To rank product categories by profitability (Total Revenue) and customer engagement* 38](#_Toc197533550)

[*3.3.4* *To assess how payment method correlates with spending habits* 41](#_Toc197533551)

[*3.3.5* *To map purchasing power and preferences by geographic region* 44](#_Toc197533552)

[*3.3.6* *To uncover the age-based trends in product ratings* 49](#_Toc197533553)

[3.4 Extra Features 51](#_Toc197533554)

[**4.0** **Conclusion** 54](#_Toc197533555)

[**5.0** **Workload Matrix** 56](#_Toc197533556)

[**6.0** **References** 57](#_Toc197533557)

# **Introduction**

In the retail industry, many companies face underlying issues pertaining to capturing customer preferences, inventory control and improvement of shopping experience. The retail industry continuously makes drastic changes powered by the implementation of data analytics for survival and success (retailcloud, 2024).

The use case of data analytics in the retail industry can provide targeted communication customers, whereby retailers can easily segment their customer base based on the insights generated. Information extracted from the data can pertain to purchasing patter and preferences allows retailers to create personalized marketing campaigns that best matches. According to (retailcloud, 2024), studies showed that customized, data-oriented marketing campaigns outperforms traditional marketing methods by 20%, therefore resulting in higher engagement and better conversion rates.

Furthermore, the application of data analytics on the retail industry can also generate a more accurate demand forecasting used for efficient inventory control measurements. Online retail analysis mainly focuses on projecting customer needs by analysing historical sales data, seasonal trends and market dynamics (retailcloud, 2024). This process reduces overstocking and understocking issues, therefore ensuring that the right products are available at the right time and therefore improving customer satisfaction.

In this project, the developers will dive into exploring and visualizing data and relationships of variables amongst each other to understand more on consumer backgrounds and their purchasing behaviour. The developers will also identify factors that measure customer satisfaction, while also generating insights for stakeholders on a set of recommended courses of action.

## Data Description

The dataset provided describes customer demographics and purchasing history. The data attributes are detailed as follows:

* **Transaction ID**: A unique identifier for each transaction
* **Customer ID**: A unique identifier for each customer
* **Name**: Customer’s full name
* **Email**: Customer’s email address
* **Phone**: Contact number of the customer
* **Address**: Physical address of the customer
* **Age**: Age of the customer
* **Gender**: Gender of the customer
* **Income**: Income bracket or level of the customer
* **Customer Segment**: Classification of customers based on behaviour or demographics
* **Total Purchases**: Total number of purchases made by the customer
* **Amount Spent**: Total monetary amount spent by the customer
* **Product Category**: Category to which the purchased product belongs
* **Product Brand**: Brand name of the product
* **Product Type**: Type or model of the product purchased
* **Feedback**: Customer’s feedback or rating related to the product or service received
* **Shipping Method**: The method used to deliver the purchased products
* **Payment Method**: The method of payment chosen by the customer
* **Order Status**: Status of the order (e.g., shipped, delivered)
* **Products**: Product Purchased
* **Ratings**: Ratings given by customers on different products (target)
* **(City, State, Zipcode, Country):** Geographic details of the customer.
* **(Year, Month, Time, Date):** Date components extracted from the last purchase date.

## Assumptions

* The dataset represents the population being studied
* There are no duplicate entries unless explicitly checked and handled
* All record values are within plausible ranges (e.g. no negative ages)
* External factors (such as economic conditions) do not significantly impact the data
* All recorded transactions are legitimate
* Aggregating data (e.g., daily into monthly) does not lose critical insight
* Certain outliers can be excluded if they are data errors
* It is assumed that the purchase amounts units are all presented in USD despite the regional differences.

## Hypothesis & Objectives

**Hypothesis 1**:

*“Customers in higher income brackets spend more per transaction that those in lower income brackets, but their purchase frequency (Total Purchase) is similar”*

**Objective 1:**

*“To analyse how income levels affect spending patterns (Amount Spent) and purchase frequency (Total Purchases)”*

* Analysis 1-1: Are “Amount Spent” and “Total Purchases” similar across different income categories?
* Analysis 1-2: Do high-income customers favour specific “Product Categories” or “Brands”
* Analysis 1-3: Does “Payment Method” mediate the income spending relationship?

**Hypothesis 2:**

*“Females are more likely to be ‘Premium’ customers than their Male counterpart”*

**Objective 2:**

*“To identify what impacts gender spending patterns”*

* Analysis 2-1: Is there a relationship between gender and customer segmentation?
* Analysis 2-2: Is there “Income” equality amongst the genders?
* Analysis 2-3: Do females have a higher purchasing power than males?

**Hypothesis 3:**

*“Electronics (Product Category) have the highest average order value (Amount Spent) but the lowest purchase frequency (Total Purchase)”*

**Objective 3:**

*“To rank product categories by profitability (Total Revenue) and customer engagement”*

* Analysis 3-1: Does a high number of orders correlate with high purchase amount (Amount Spent)?
* Analysis 3-2: Is there a correlation between “Product Brand” and “Ratings” within categories?

**Hypothesis 4:**

*“Credit card users (Payment Method) spend 50% more per transaction than PayPal users”*

**Objective 4:**

*“To assess how payment method correlates with spending habits”*

* Analysis 4-1: Are certain “Product Categories” related to specific payment methods?
* Analysis 4-2: Does the customer use of a specific payment method change over the year?

**Hypothesis 5:**

*“Customers from the USA (Country) have on average a higher purchasing power (Amount Spent) than customers from other nations by at least 20%.”*

**Objective 5:**

*“To map purchasing power and preferences by geographic region”*

* Analysis 5-1: What are the income levels of customers in each country?
* Analysis 5-2: Is there a difference in purchasing power (Amount Spent) over the year in each country?
* Analysis 5-3: Are there regional trends in Product Category preferences?

**Hypothesis 6:**

*“Customers aged 18 – 25 have much higher standards when purchasing products compared to customers who are older”*

**Objective 6:**

*“To uncover the age-based trends in product ratings”*

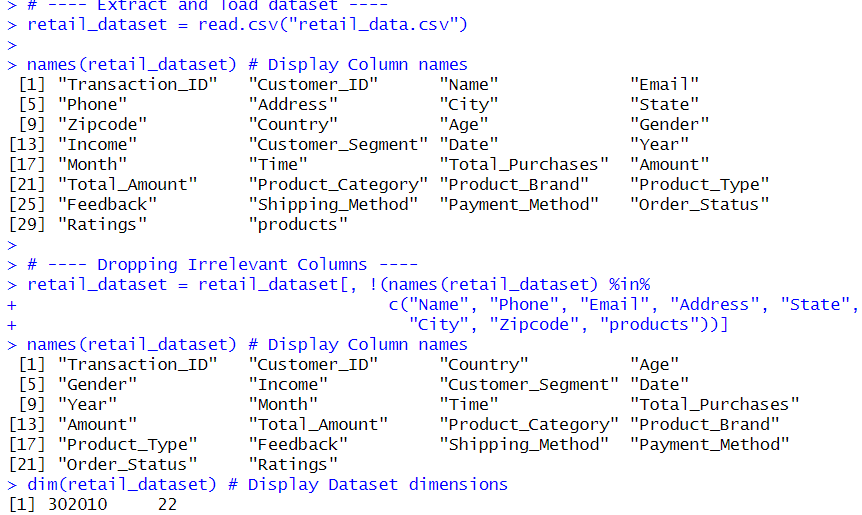
* Analysis 6-1: Is there an underlying relationship between age and product ratings?
* Analysis 6-2: Are product “Ratings” influenced by “Income Levels”?

# **Data Preparation**

## Data Import



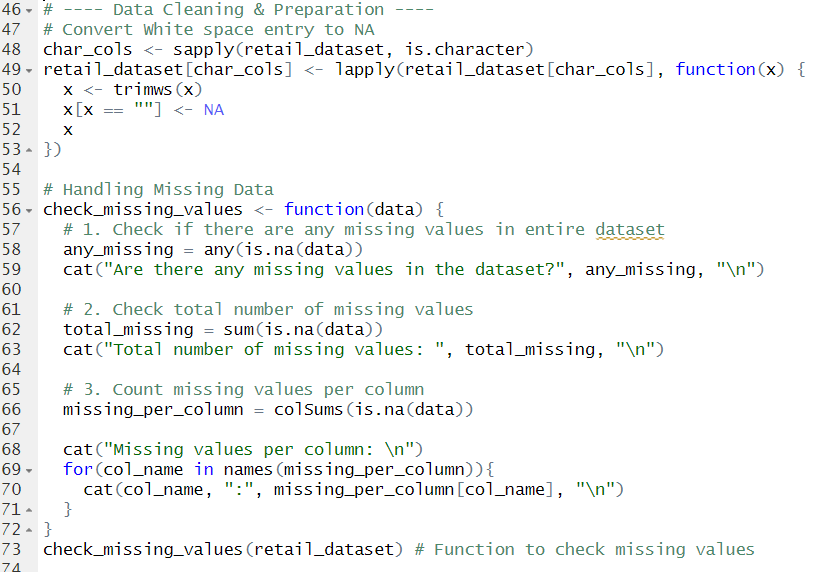
Code Snippet 1: Loading Dataset & Removing Irrelevant Columns



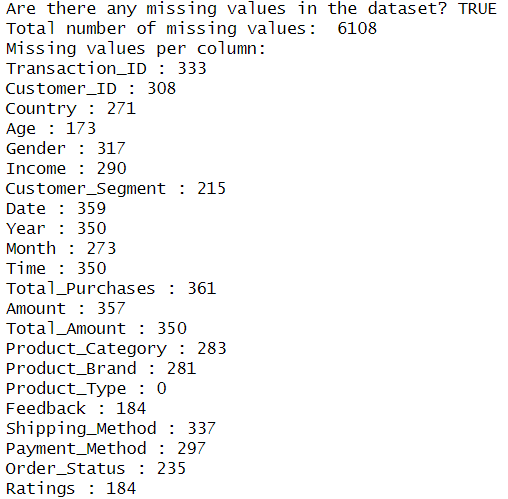
Code Snippet Result 1

The code above shows the process of extracting the dataset from the csv file and loading it onto a data frame called ‘retail\_dataset’. Essentially, this data frame will be the base where all the cleaning, preprocessing and analysis will be made. The code also shows the deletion of the columns "Name", "Phone", "Email", "Address", "State", "City", "Zipcode", "products" due to their irrelevancy to the preprocessing and analysis process. The ‘names’ function is called to display all the column names. The ‘dim’ function shows the dimensions of the dataset, containing 30210 rows and 22 columns.

## Data Cleaning & Pre-processing & Validation

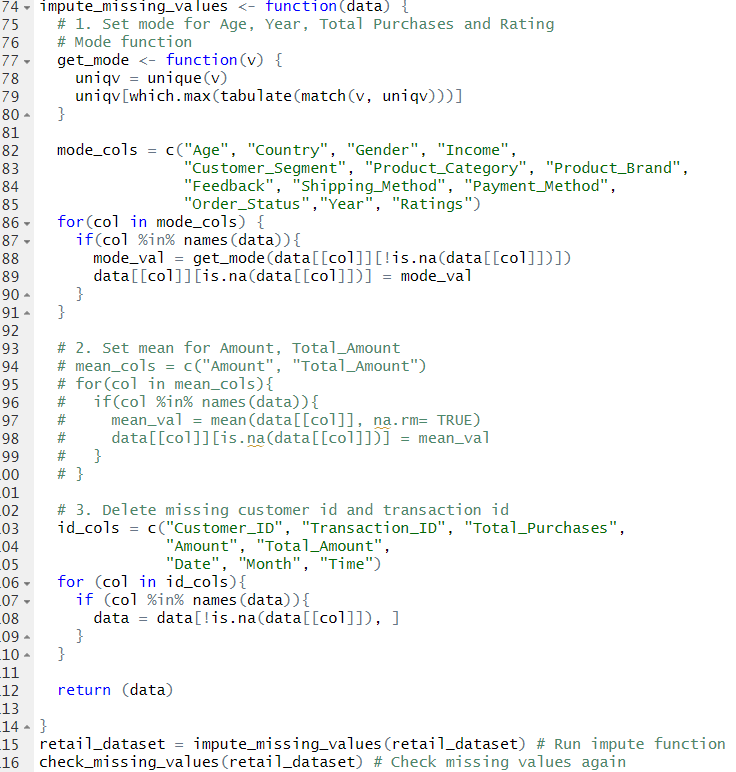


Code Snippet 2: Converting Whitespace to NA & Checking for NA values

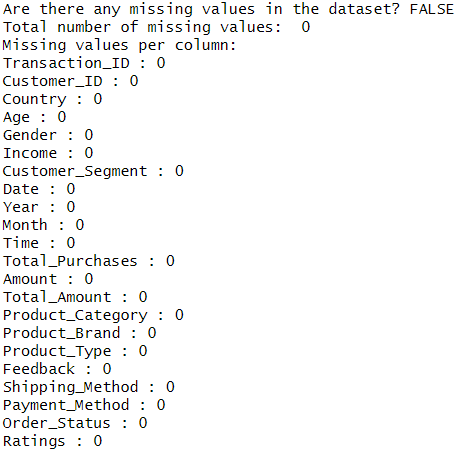


Code Snippet Result 2

In the dataset, there exists missing data, this can come in the form of either missing whitespace or value of NA, therefore, it is necessary to clean the data beforehand. The code from lines 48-53 is used to extract any missing value in the records within the dataset and replacing it with value ‘NA’ (see line 51). Onwards from line 56-72, is the function ‘check\_missing\_values’ used to extract if there are any missing values (= NA), their total amount and their distribution within each column in the dataset. The result shown above shows that only a minor fraction of the dataset being Null, which does not significantly impact the dataset.

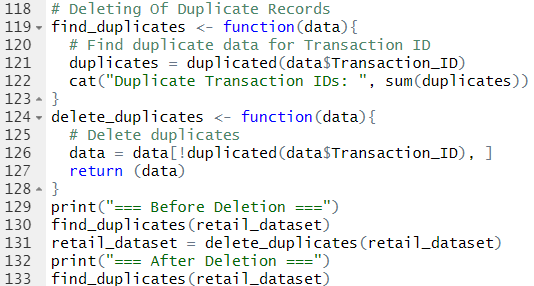


Code Snippet 3: Imputation Of Missing Values

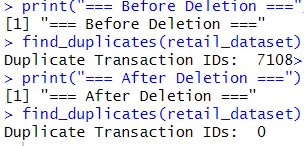


Code Snippet Result 3

The code snippet above shows the function for imputing the missing values viewed earlier in the ‘check\_missing\_values’ function. The function ‘impute\_missing\_values' contains different methods of imputation of the columns accordingly. For example, columns ‘Age’, ‘Country’, ‘Gender’, etc are imputed using the mode since they are categorical values while ‘Customer\_ID’, ‘Transaction\_ID’, ‘Amount’, ‘Total Amount’, etc are imputed by having their records being deleted due to the significance of data being present and its validation. The ‘check\_missing\_values’ function is called again, showing no result of NULL values being present in the dataset.

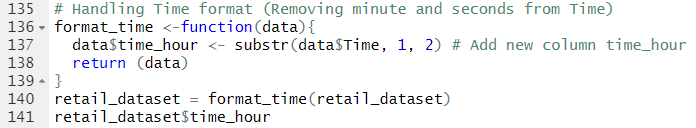


Code Snippet 4: Finding & Deleting Duplicates



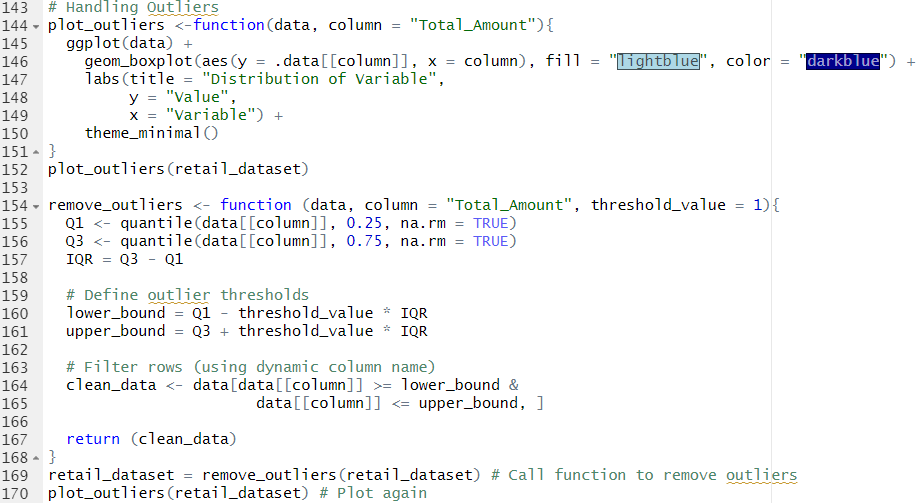
Code Snippet Result 4

The code snippet above shows the functions ‘find\_duplicates’ and ‘delete\_duplicates’ used to extract and delete duplicates of data present in the ‘Transaction\_ID’ column. The ‘Transaction\_ID' column is the primary key for the data and therefore, should not contain any duplicates. The ‘delete\_duplicates’ function extracts the duplicated rows and deletes them.

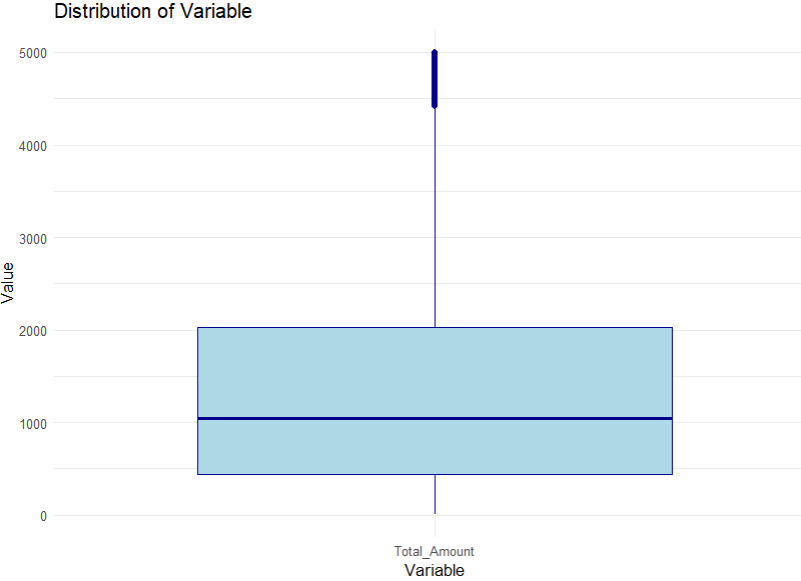


Code Snippet 5: Transforming Time Column

The code snippet above shows the ‘format\_time’ function used to transform or change the ‘Time’ column into a new column which displays only the hour of the time presented (since minute and seconds will not be used in the analysis).



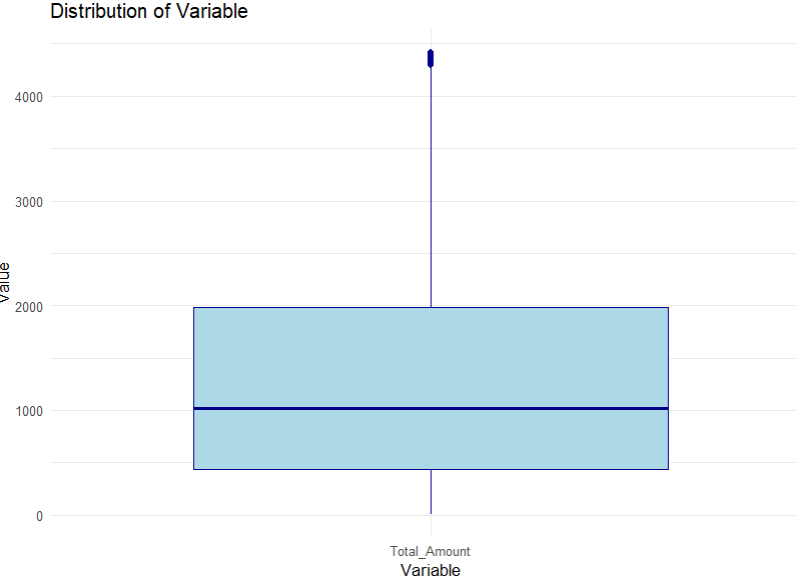
Code Snippet 6: Plotting & Removing Outliers



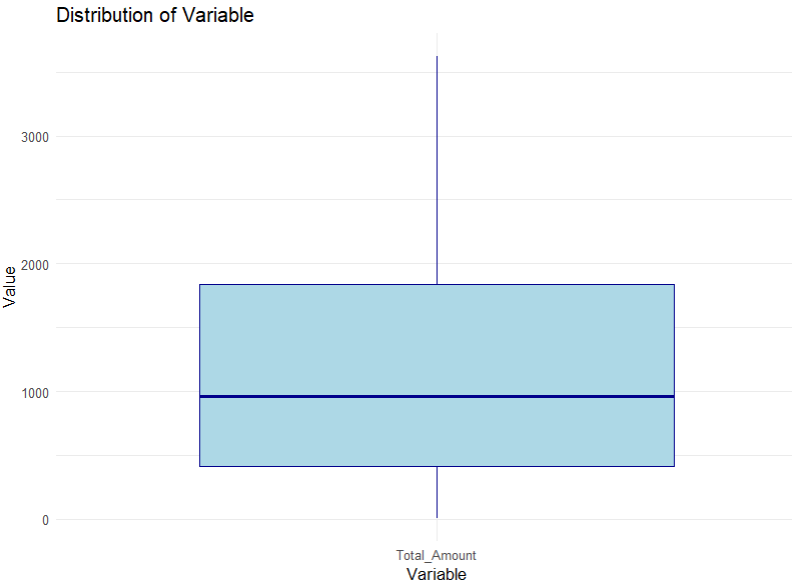
Code Snippet Result 5: Plot Function Before Deletion Of Outliers

The code above shows the functions ‘plot\_outliers’ and ‘remove\_outliers’ used to view and delete records containing outliers pertaining to the column specified by the user respectively. Both functions contain the dataset and column parameters which consider the dataset provided and the column to plot or delete outliers for. The default value is set to the column ‘Total\_Amount’ though the developer can change that by providing the argument as needed. The ‘plot\_outliers’ function uses ggplot2 package to generate a box and whiskers plot showing the column data distribution.

In the ‘remove\_outliers’ function, the IQR method of outlier removal is used. This method decides that any data point in a boxplot that is more than 1 IQR points below the first quartile data or more than 1 IQR points above the third quartile data is considered an outlier (Chaudhary, 2024). Typically, the default threshold would be 1.5 IQR points, however, as seen by the figure below, there still exists outliers in the dataset (denoted by blue circles). As for when using 1 IQR points, the threshold becomes much stricter.



Code Snippet Result 6: When Using Default Threshold



Code Snippet Result 7: When Using 1 IQR Points For Threshold



Code Snippet 7: Deleting Time & Transaction ID Columns

The code above shows the subset function being used to remove ‘Transaction\_ID’ and ‘Time’ columns from the dataset. These columns are no longer needed for the analysis process.

# **Data Analysis**

## Analytical Techniques Utilized

### Chi-Square Test

A Chi-Square test or particularly the Chi-Square Test of Independence is a hypothesis testing methods of identifying whether 2 variables in each dataset have dependency on one another. The purpose of this test is used to decide if the variables are related or not (jmp Statistical Recovery, n.d.). The following formula is used to determine the Chi-Squared value or .

where:

* **O:** the observed frequency
* **E:** the expected frequency

**Applying the Chi-Square Test:**

1. Define the null and alternative hypothesis before collecting data
2. Decide on the alpha value α which is relies on the rate of risk willing to take drawing the wrong conclusion. For example, given that α = 0.1 when testing for independence, this means that the developer has a 10% risk of drawing the conclusion that the 2 variables are independent when they are not.
3. Checking the data for any errors
4. Checking the assumptions for the test
5. Performing the test and drawing conclusion

**Degrees of freedom (df):** used to determine if a null hypothesis can be rejected based on the total number of variables and samples within the experiment (Ganti, Rasure, & Kvilhaug, 2024). Reference is made using the Chi-Squared table to get the p-value.

**The degree of freedom is calculated as:**

where:

* **r:** number of rows
* **d:** number of columns

**Comparison to critical value or p-value:**

If > critical value (or p-value < α), reject null hypothesis and accept alternate hypothesis, else accept null hypothesis.

The assumption for the dataset to perform Chi-Squared test include:

* Data must be in frequency count (not percentage)
* Expected frequencies should be more than or equal to 5 for most cells
* Observations must be independent

### Descriptive Statistics

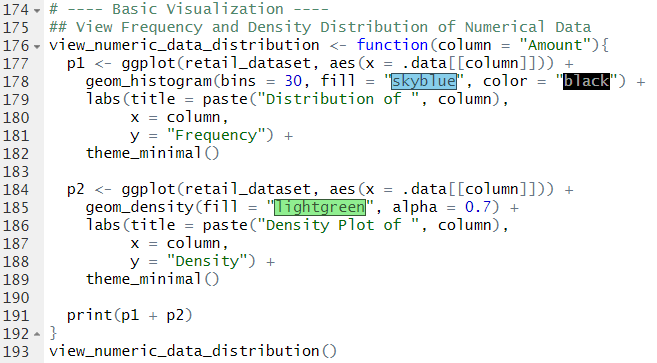
Descriptive Statistics is the concept of summarizing numerical and categorical data in a precise and informative way such as measuring centre, variability, shape and location of dataset to describe key features about the data (Green, Manski, Hansen, & Broatch, 2023). Descriptive statistics generates the foundation for quantitative analysis and insights into the distribution and characteristics of the dataset. Some of the key elements in descriptive statistics include:

* **Measure of central tendency:** Measures the average of the values of distribution, which includes the mean, median and mode.
* **Dispersion:** Measures the degree of spread or distribution of the data. Though only used for ordinal and interval scale data, it can be used to identify elements such as Range, Interquartile Range, Variance and Standard Deviation.
* **Shape:** Is highly important in understanding the type of analysis to perform on the data, especially when detecting skewness (the extent to which the data is distributed, either positive or negative), Kurtosis (the extend of which the distribution is flat or peaked) or Modality (the number of peaks in the distribution).

### Graphical Representations

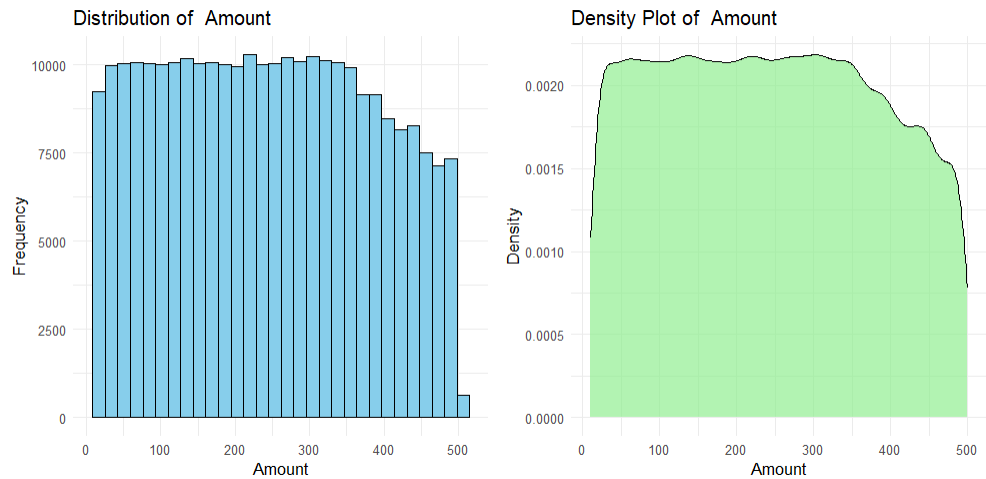
Datasets can be represented in a variety of different ways and are highly applicable in identifying relationships between variables in the dataset. Graphs measure the shapes of distribution, indicating any relationships or outliers, therefore, enabling comparison to be made between the various distributions. Graphical representations are useful for data cleaning and exploratory analysis and can also help developers familiarize themselves with the dataset before analysis (Green, Manski, Hansen, & Broatch, 2023).

## Exploratory Data Analysis (EDA)



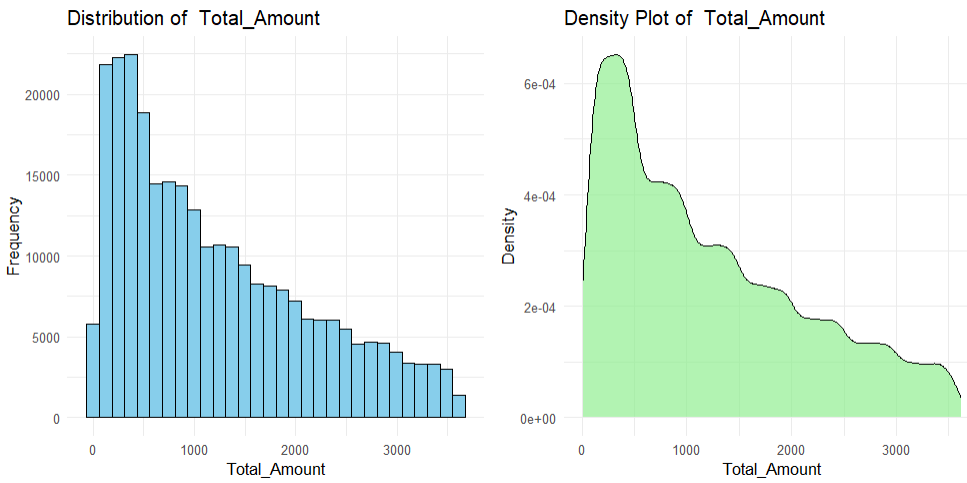
Code Snippet 8: view\_numeric\_data\_distribution Function

The code above shows the ‘view\_numeric\_data\_distribution’ designed to understand numerical data distribution through its frequency and Density. The 2 graphs will be plotted next to each other.



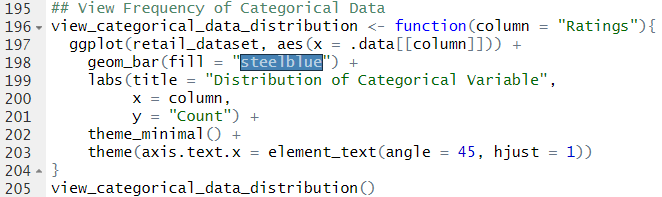
Code Snippet Result 8: Amount Distribution

The plot above shows the results of the frequency and density distributions of the Amount purchased. From the 2 plots shown, there exists an even distribution of frequency amongst the amount paid per product and the data points demonstrate a flatter, wider spread in both histograms. As for density distribution, a low-density data, visualized by a flatter, wider spread region, shows that data is spread out over a wide range.



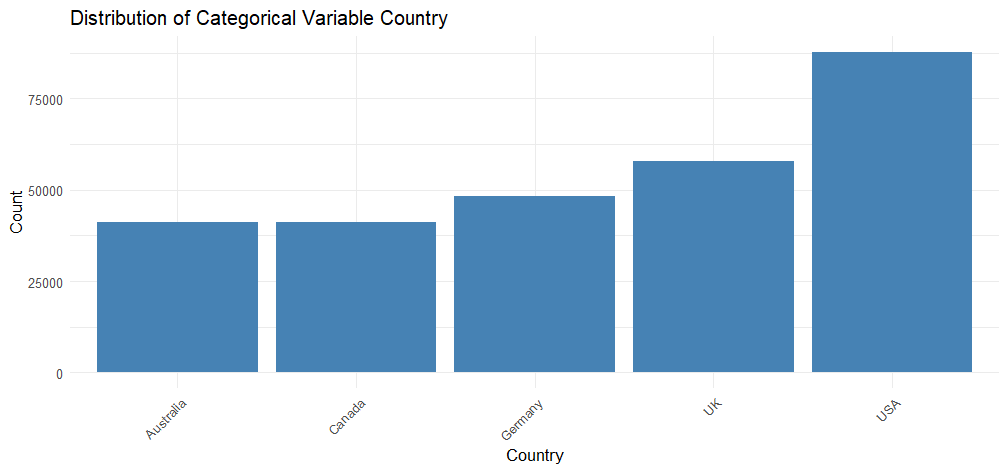
Code Snippet Result 9: Total Amount Distribution

As seen by the plot above, it can be observed that the data on both frequency and density is skewed to the left, showing most total purchases being between 0-1000 USD. These graphs, in comparison with the visualization presented by the previous graphs, can tell us that **most customers purchase either low volume on highly expensive products or high volume on less expensive products, though most of these purchases are budgeted at the cost of no more than 1000 USD**. This is evaluated by the peak presented on the left side of the density graph where data is closely packed in a small range.



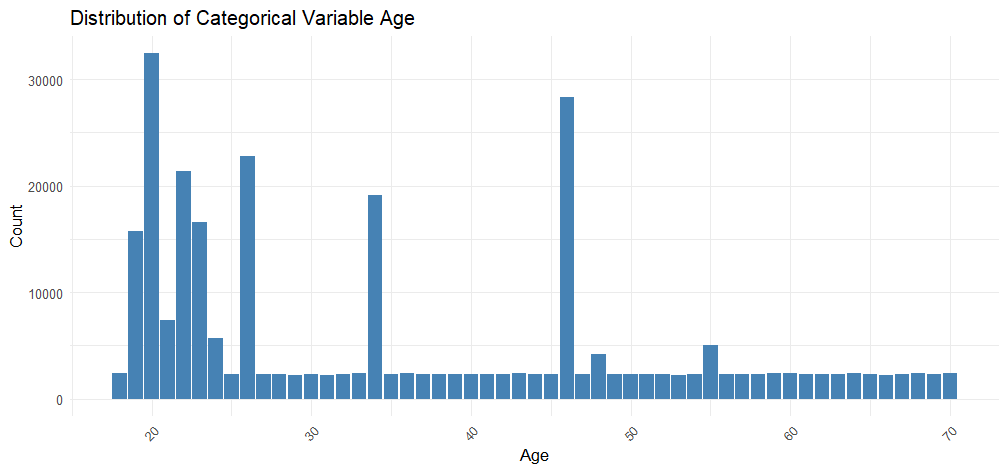
Code Snippet 9: view\_categorical\_data\_distribution Function

The code snippet above shows the view\_categorical\_data\_distribution function which plots a chart of data distribution for categorical values. Unlike numerical values, which are plotted by density graphs and frequency histograms, categorical values are plotted by only frequency of categories using bar charts. The function contains a column parameter which allows the developer to specify the column name to plot the data for, making the function reusable while making the code DRY (Do not Repeat Yourself). The default column name is set to ‘Rating’.



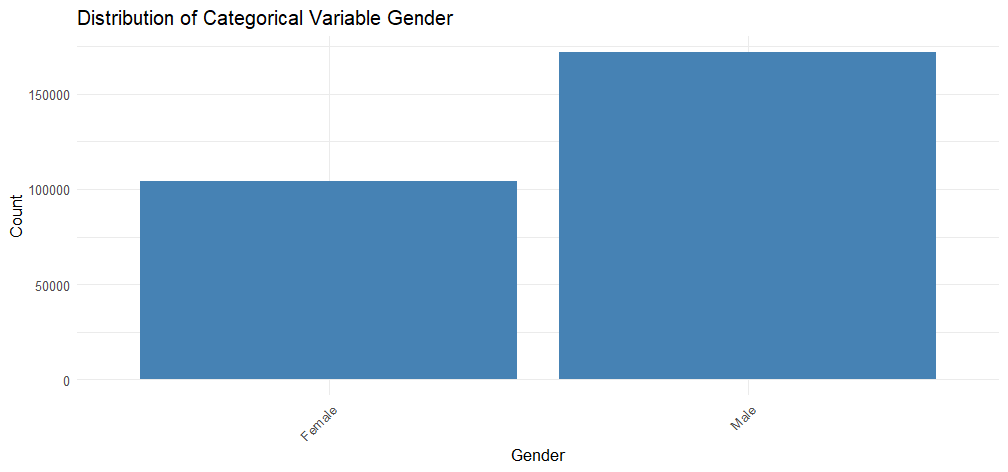
Code Snippet Result 10: Country Distribution

The figure above shows the ‘Country’ distribution of the dataset presented using a boxplot chart. From the figure, we can deduce that the most common purchases of customers in the dataset were in the order of ‘USA’ then ‘UK’, ‘Germany’, ‘Canada’ and then ‘Australia’. The ‘USA’ shows significant margin in count over the other countries.



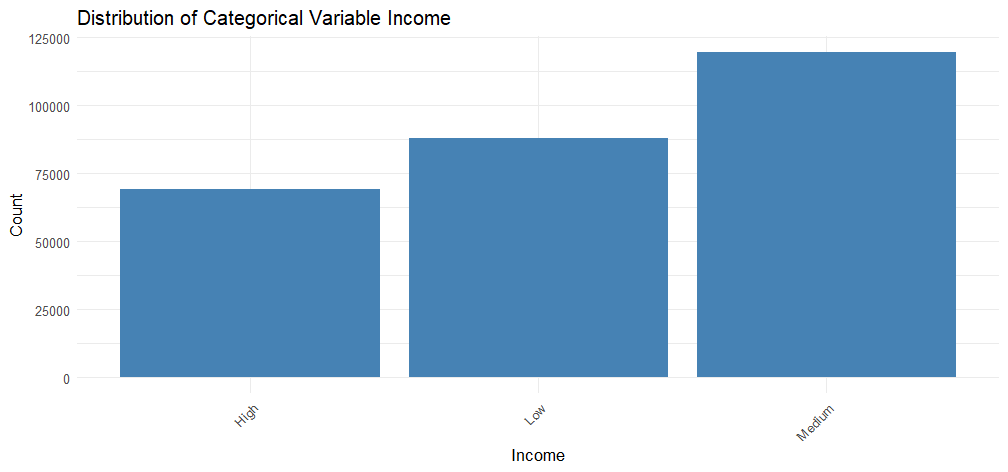
Code Snippet Result 11: Age Distribution

The figure above shows the ‘Age’ distribution amongst the customer dataset, whereby most customers fall between the ages 18 – 25 inclusive. However, from the plot, we can also see certain spikes between the ages of customers, especially in customers aged 34 and 46.



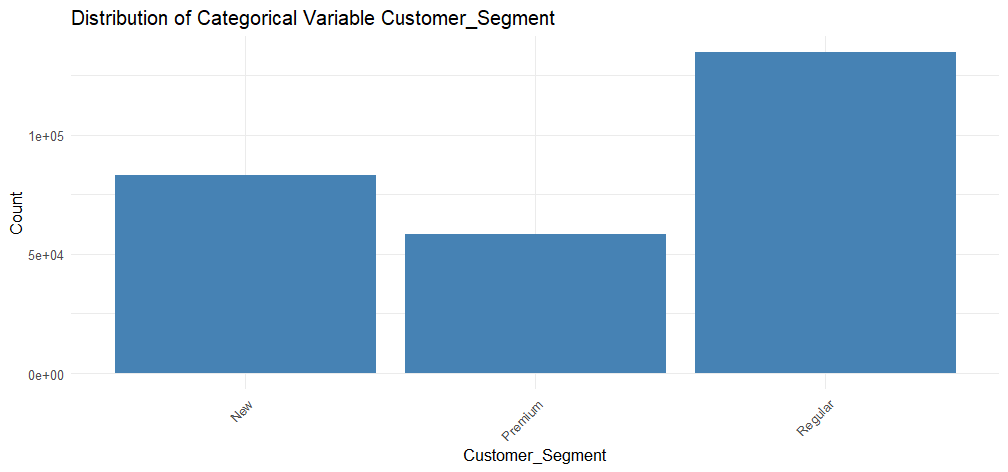
Code Snippet Result 12: Gender Distribution

The figure above shows the Gender distribution of the customers in the retail dataset, whereby most customers are in the ‘Male’ and showing a nearly 1.5x the amount of ‘Female’ consumers in the dataset.



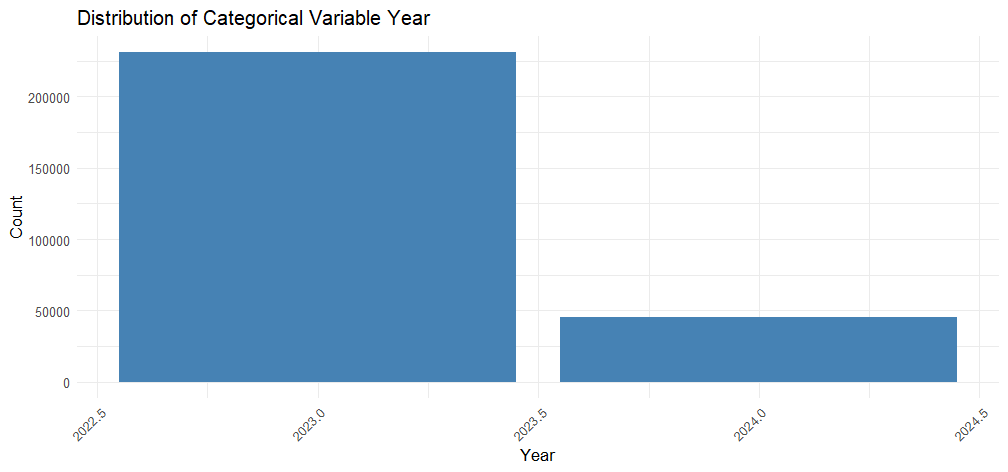
Code Snippet Result 13: Income Distribution

The figure above shows the Income distribution amongst the consumers in the dataset, whereby most consumers are in the ‘Medium’ category, followed up by the ‘Low’ category and then by the ‘High’ category.



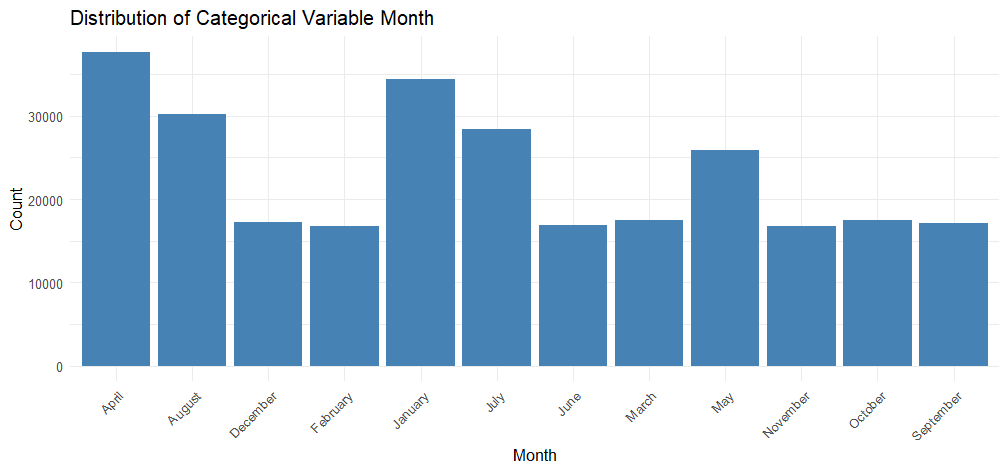
Code Snippet Result 14: Customer Segment Distribution

The figure above shows the Customer Segment distribution amongst the consumers of the dataset, whereby most customers fall in the ‘Regular’ category, followed by the ‘New’ and then the ‘Premium’ category. The dataset shows a drastic difference in customer segment, whereby ‘Regular’ customers are nearly as large as ‘Premium’ and ‘New’ customers combined.



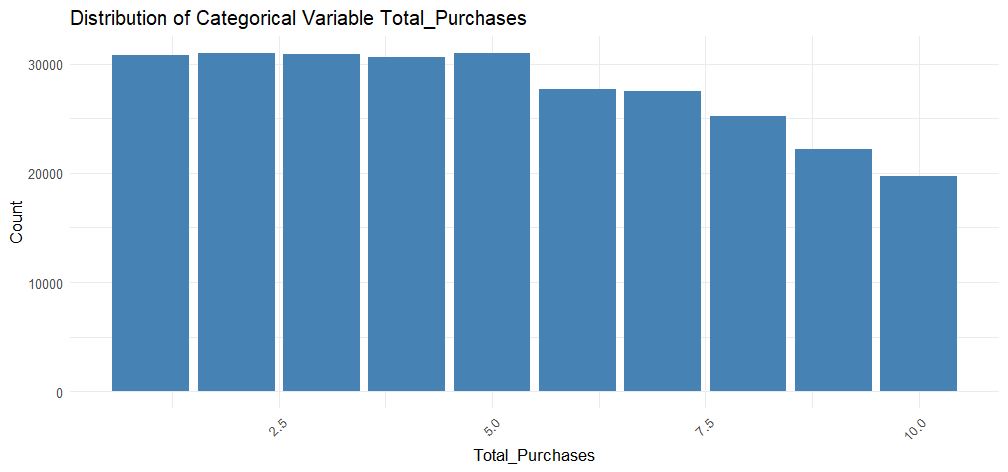
Code Snippet Result 15: Year Distribution

The figure above shows the distribution of the years of the transactions made in the dataset. From the dataset, it shows that most customers purchased items in the year 2023 with a significant difference compared to the year 2024.



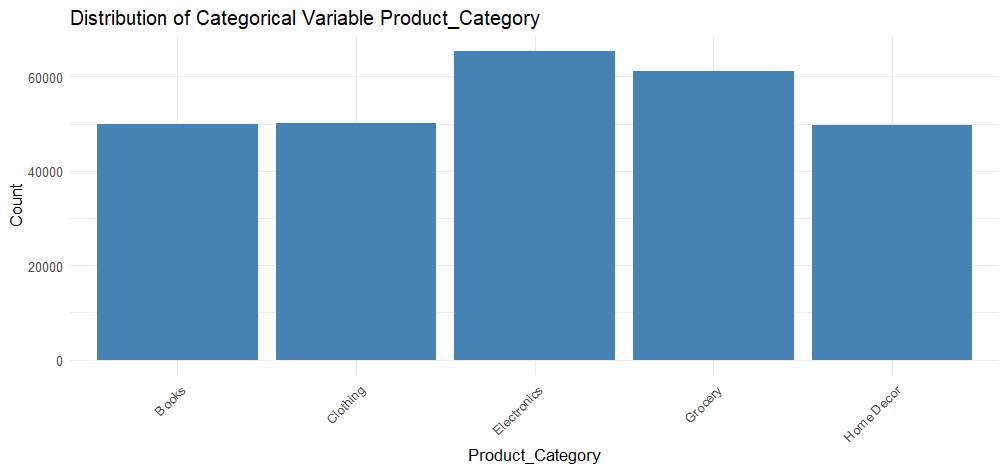
Code Snippet Result 16: Month Distribution

The figure above shows a boxplot of the transaction months of items purchased from the dataset, whereby most purchases were made in April, August, January, July and May. The rest of the month show a highly similar purchase count with very little variance.



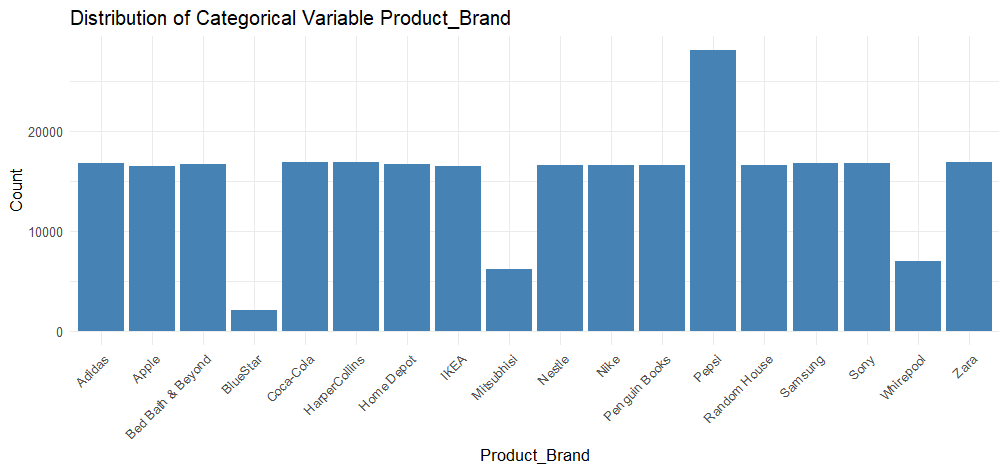
Code Snippet Result 17: Purchase Count Distribution

The figure above shows the purchase count distribution, meaning the distribution of item quantity per transaction in the dataset. The dataset distribution of ‘Total Purchases’ shows less variance in purchase count per transaction, however there is a slight decrease in purchase count after 5 amounts, meaning that customers are more likely to purchase 5 or less quantities per item.



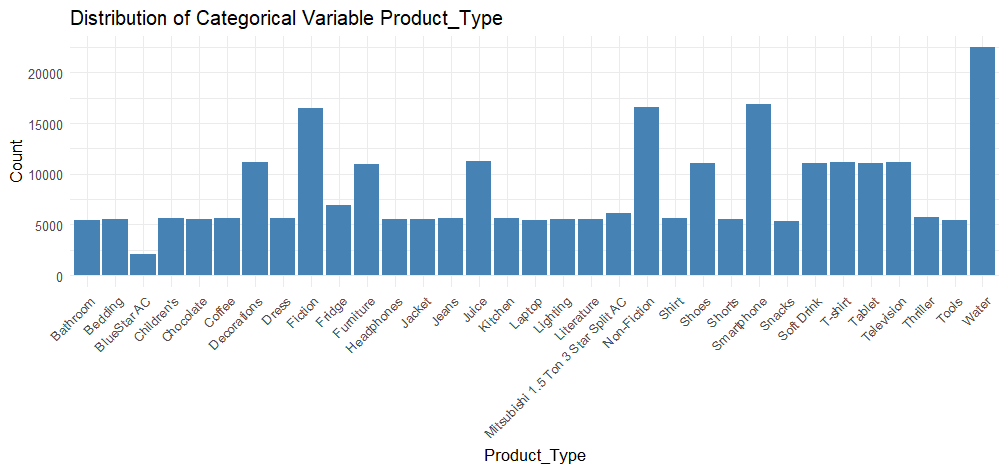
Code Snippet Result 18: Product Category Distribution

The figure above shows the product category distribution of the retail dataset, whereby most items purchased fall into the ‘Electronics’ category and ‘Grocery’ category, while ‘Home Décor’, ‘Books’ and ‘Clothing’ have slightly similar purchase counts.



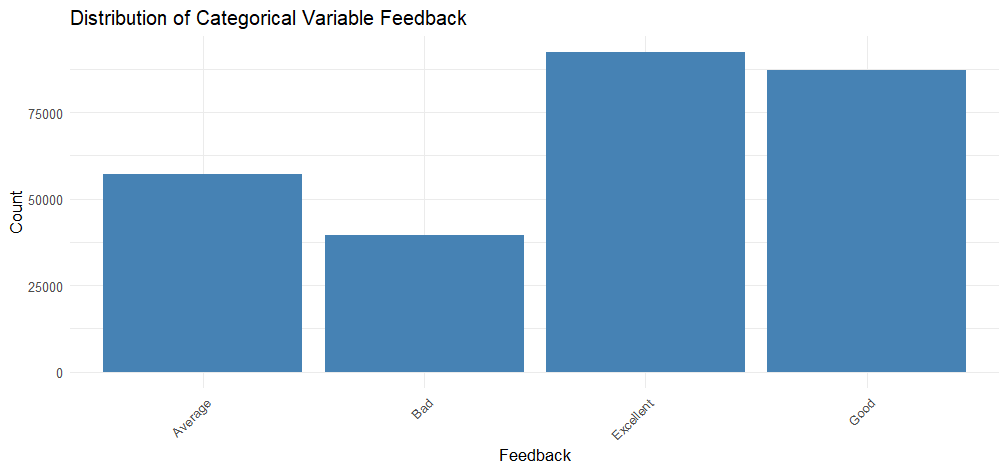
Code Snippet Result 19: Product Brand Distribution

The figure above shows the distribution of product brands of the products purchased, whereby most items purchased are from the ‘Pepsi’ brand, significantly above most other brands. Other brands on the other hand have very similar counts, excluding ‘BlueStar’, ‘Mitsubishi’ and ‘Whirepool’ which have very low purchase counts.



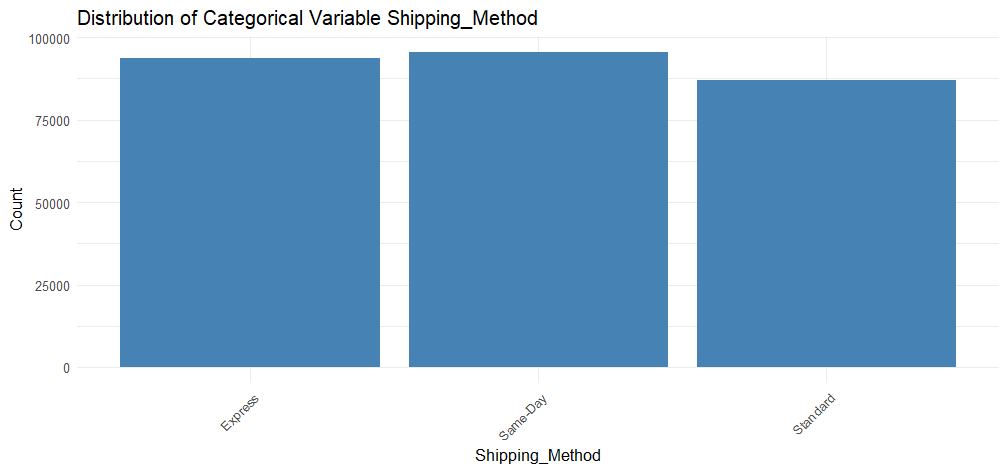
Code Snippet Result 20: Product Type Distribution

The figure above shows the distribution of product types of items purchased from the retail dataset. ‘Water’ has the most purchase counts compared to any other product type. This is explained due to the high amount of demand on ‘Pepsi’ and ‘Coca-Cola’ brands, which are both categorized as ‘Water’. Other significant product types are ‘Fiction’ books, ‘Non-fiction’ books and ‘Smart Phones’.



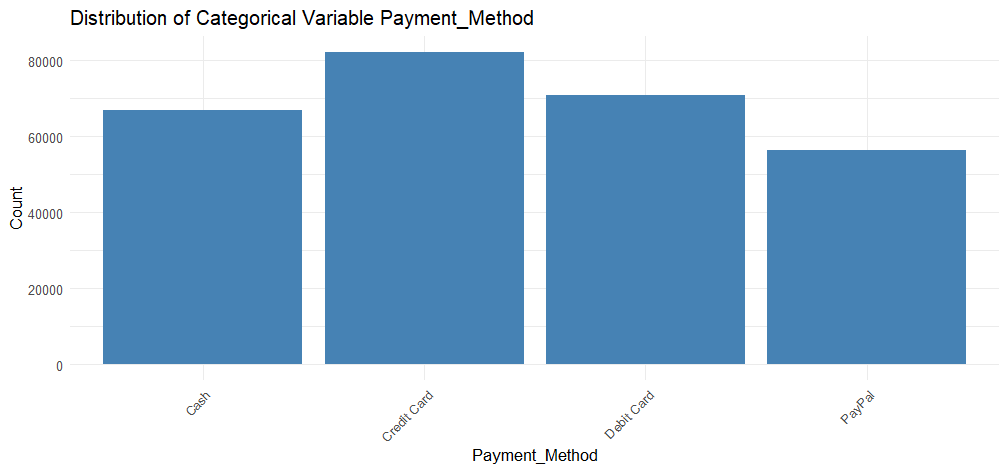
Code Snippet Result 21: Feedback Distribution

The figure above shows the ‘Feedback’ distribution of items purchased whereby most customers provided ‘Excellent’ and ‘Good’ feedback, followed by ‘Average’ and then ‘Bad’. The dataset shows a larger portion of the consumers have positive feedback towards their purchase of items.



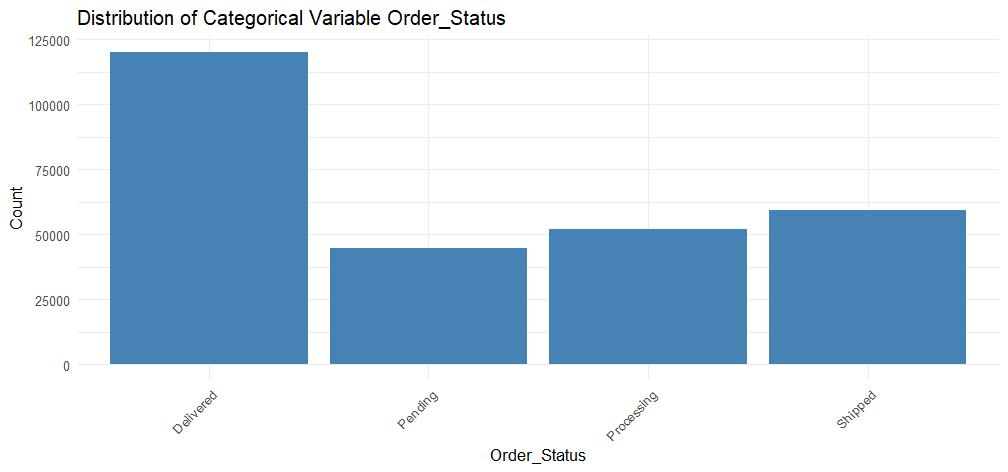
Code Snippet Result 22: Shipping Method Distribution

The figure above shows the Shipping Method distribution of items purchased from the dataset, detailing that purchase methods are highly similar in count and have less variance amongst one another.



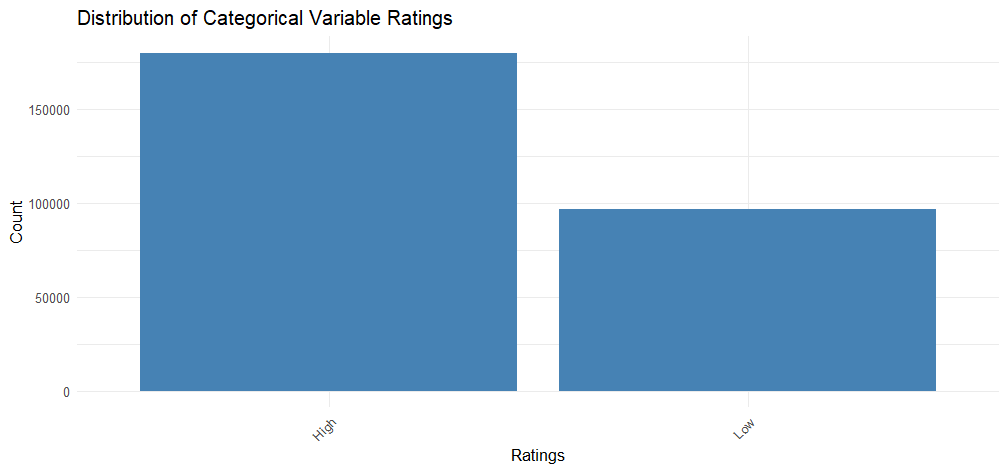
Code Snippet Result 23: Payment Method Distribution

The figure above shows the payment method distribution of items purchased, whereby most customers purchased their items through ‘Credit Card’ and ‘Debit Card’ as compared to customers who purchased through ‘Cash’ and ‘PayPal’.



Code Snippet Result 24: Order Status Distribution

The figure above shows the order status of items purchased from the retail dataset, whereby most items purchased have been ‘Delivered’ as compared to other items that are either ‘Pending’, ‘Processing’ or ‘Shipped’.



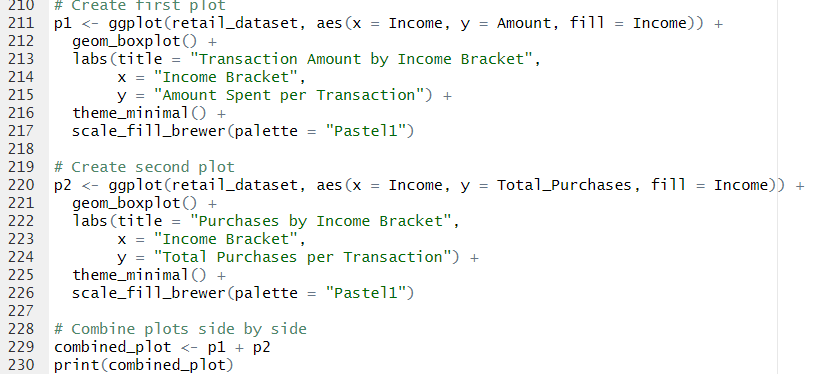
Code Snippet Result 25: Rating Distribution

The figure above shows the Ratings of items purchased by the consumers, whereby there is a significance in ‘High’ ratings as compared to ‘Low’ ratings from the transactions made. This means that consumers have had a more positive rating of the item purchased and received than negative.

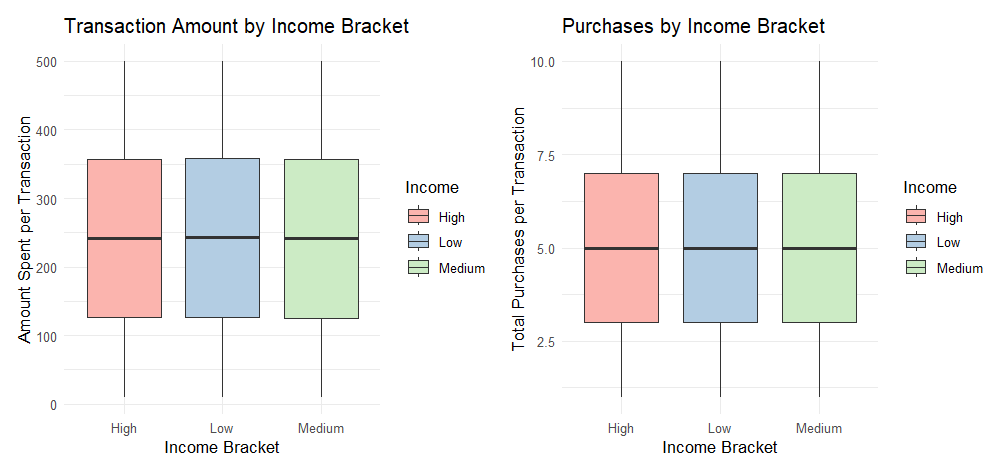
## Objectives

### *To analyse how income levels affect spending patterns (Amount Spent) and purchase frequency (Total Purchases)*

#### Analysis 1-1: Are “Amount Spent” and “Total Purchases” similar across different income categories?



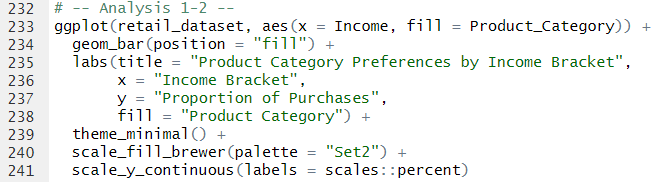
Code Snippet 10: Analysis 1-1



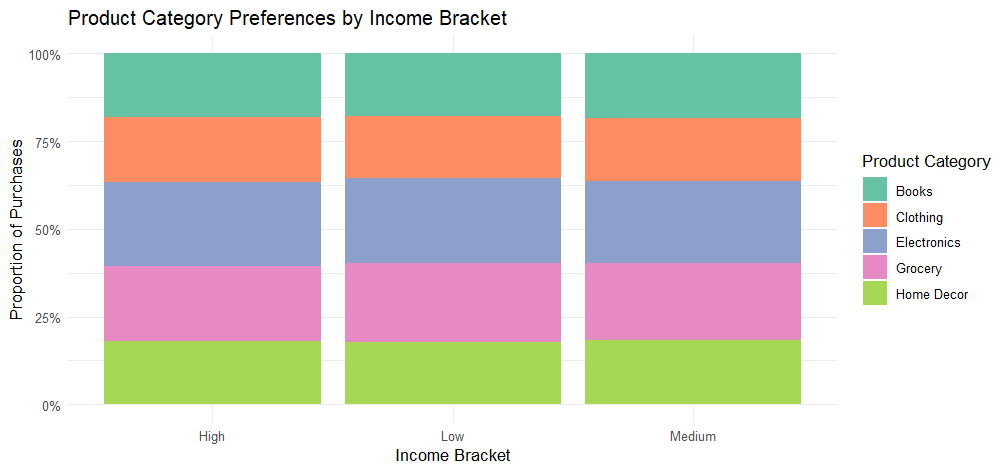
Code Snippet Result 26: Analysis 1-1

The code snippet alongside the snippet result shown above represents a combined box plot diagram showing the relationship between Income brackets and amount spent comparatively with total purchases. From the plot diagram, we can deduce that amount spent per transaction and total purchases do not have a relationship with the income bracket as seen as the mean, upper and lower quartiles are nearly identical across all income categories. Therefore, we can deduce that Income does not affect the spending amount per transaction.

#### Analysis 1-2: Do high-income customers favour specific “Product Categories” or “Brands”?

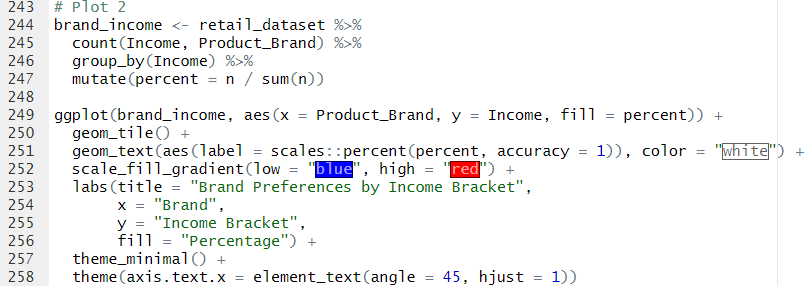


Code Snippet 11: Analysis 1-2 (Product Categories)



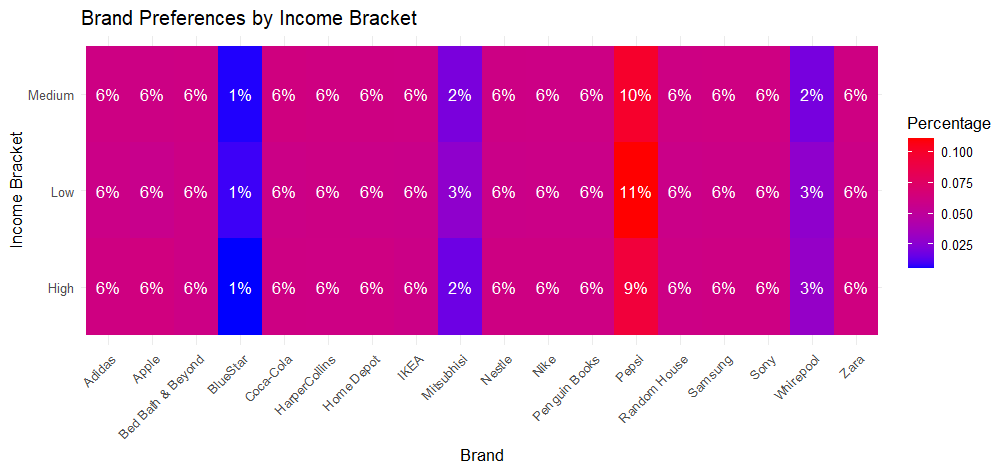
Code Snippet Result 27: Analysis 1-2 (Product Categories)

The code above shows a stacked bar chart to reflect the relationship between product categories and income bracket. The bar chart shows that the proportion of purchases is nearly equivalent across all income categories, therefore there is no evidence to prove that different income categories can affect the spending of different product categories.



Code Snippet 12: Analysis 1-2 (Brands)

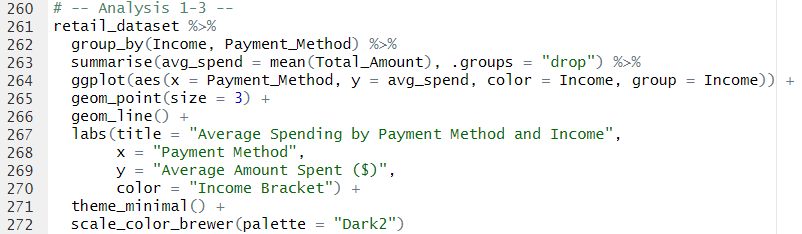
The code above draws out a heatmap matrix detailing the relationship between the purchasing power of different brands across different income categories. The code from lines 244 – 247 groups different brands within different income groups and is grouped by income bracket. Subsequently, a new data variable is created named ‘percent’ which calculates the percentage of each brand within each income group. All this data is entered onto a new Data Frame, named ‘brand\_income’ which is then inserted onto the ggplot function for plotting.



Code Snippet 13: Analysis 1-2 (Brands)

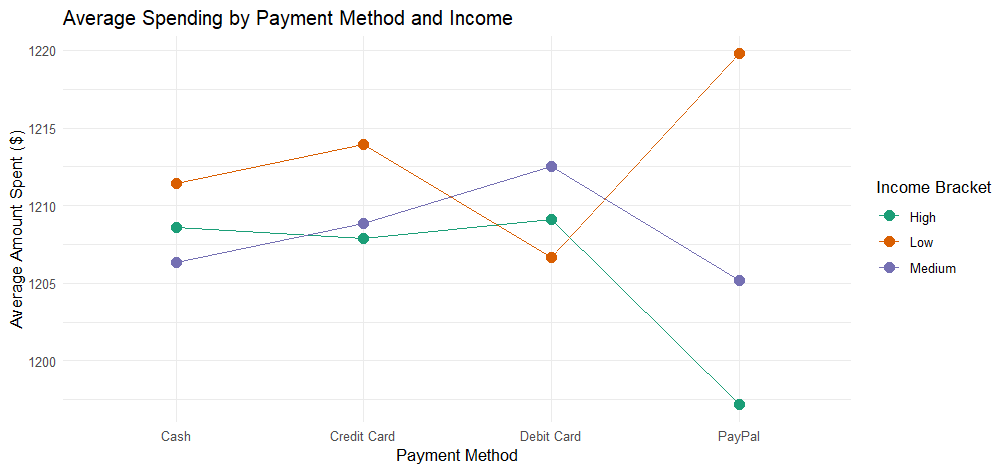
From the heatmap seen above, we can deduce that the proportion of spending per each brand across different income groups is seen as highly correlated, with spending patterns being nearly identical across various brands for the various income groups. The only exceptions can be made is the purchasing of ‘Misubishi’ where ‘Low’ income customers are more likely to purchase the brand than other income groups, ‘Whirepool’ where ‘Medium’ income customers are less likely to purchase the brand than non ‘Medium’ income customers as well as ‘Pespsi’ where ‘Low’ income customers are more likely to purchase the brand followed up by ‘Medium’ and ‘High’ income customers. Nevertheless, the graphs represent no significant relationships amongst the spending patterns of different customer incomes for different brands.

#### Analysis 1-3: Does “Payment Method” mediate the income spending relationship?



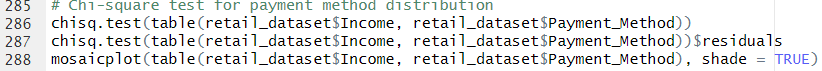
Code Snippet 14: Analysis 1-3 (Line Plot)

The code above draws out a line plot of the average amount spent for each different income groups for the various payment methods. First, the data is grouped by Income and Payment Method and the average spending is determined by the mean () function on line 263. Subsequently, the data is plotted using the ggplot function where the x-axis represent the payment method and y-axis represent the average spending patterns. The colors and grouping are based on the income brackets where each line connects points for the same income group across payment methods.

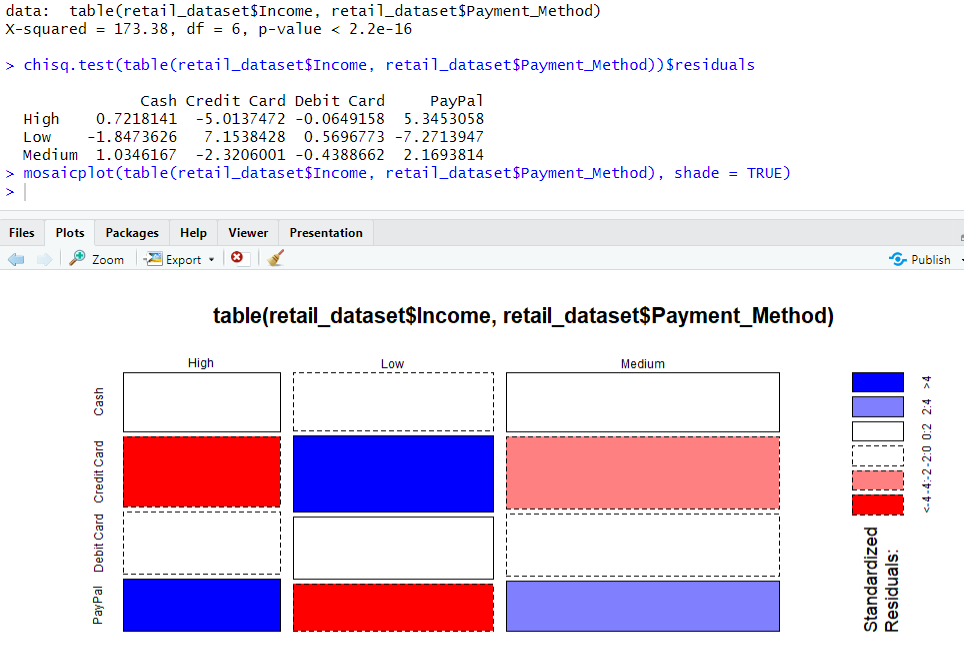


Code Snippet Result 28 (Line Plot)

The line plot above shows very interesting results. The graph shows that ‘Low’ income customers generally have a slightly higher average spending amount compared to different categories across different payment methods (except for Debit Card). Though the close margin of average spending amount is more differed in Paypal, the information is not enough to prove the significance of payment method on spending amount.



Code Snippet 15: Analysis 1-3 (Chi-Square Test)

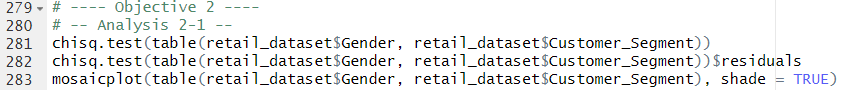


Code Snippet Result 29: Analysis 1-3 (Chi-Square Test)

The code above alongside the resulting information and mosaic plot shows the simple implementation of the Chi-Squared test on the dataset to identify the dependency between the income groups and payment method. The resulting p-value is shown as 2.2e-16 which is significantly small and thus, there exists a very dependable relationship between the 2 variables. The residual function on the other hand shows the relationship between the categories across the variables, whereby ‘Credit Cards’ are much more common among low-income customers than expected but less common among high/medium income. As for ‘PayPal’, the information shows the opposite trend where it is preferred by high/medium income but avoided by low-income households, while ‘Cash’ is slightly more favored by medium income households and less by low-income households. Lastly, there is no information to prove any significance in relations between ‘Debit Card’ and different incomes.

### *To identify what impacts gender spending patterns*

#### Analysis 2-1: Is there a relationship between gender and customer segmentation?



Code Snippet 16: Analysis 2-1

The code above utilizes the Chi-Squared test function to compare the relationship between gender and customer segment and then extracting the residuals as well as plotting those residuals onto a mosaic plot.



Code Snippet Result 30: Analysis 2-1

The mosaic plot figure displays the relationship amongst the genders and customer segment categories. The figure shows that Females are more likely to be ‘New’ members as well as ‘Premium’ members but are less likely to be ‘Regular’ members. On the other hand, Male genders are more likely to be only ‘Regular’ members only. There is no relationship of males being either ‘Premium’ or ‘New’ members.

#### Analysis 2-2: Is there “Income” equality amongst the genders?

A group of black text

AI-generated content may be incorrect.

Code Snippet 17: Analysis 2-2

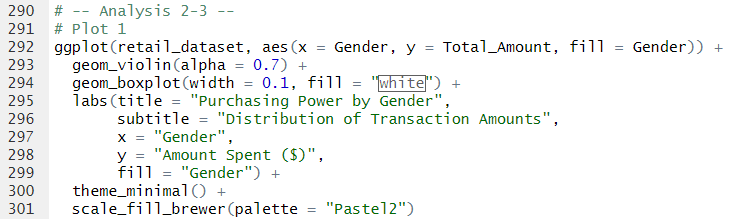
The code snippet above shows the implementation of the Chi-Squared function to evaluate the dependency of the Gender and Income variable. The residuals are extracted and plotted using a mosaic plot.



Code Snippet Result 31: Analysis 2-2

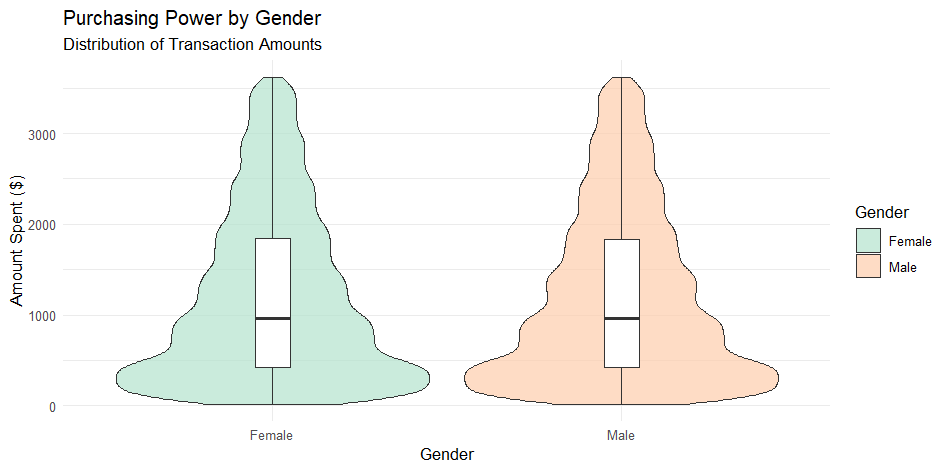
The resulting shows that there exists a relationship between gender and income, specifically where in both ‘High’ and ‘Medium’ income classes. Males are more likely to be in the ‘High’ income category and Females are not likely to be in the ‘High’ income category as they have a negative residual value. However, for ‘Medium’ income classes, it is vice versa as Females are more likely to be in that category than males. As for ‘Low’ income classes, females are more likely to be in that category than males by a close margin. Therefore, we can say that there exists an income inequality amongst the genders.

#### Analysis 2-3: Do females have a higher purchasing power than males?



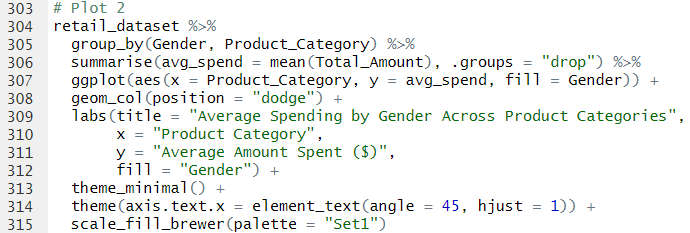
Code Snippet 18: Analysis 2-3 (Violin Plot)

The code above shows the implementation of a violin plot which is the combination of a box plot and a kernel density plot to show the distribution of total amount spent across the genders. The violin plot generates a much better understanding of the data’s shape, spread and central tendency compared to a standard boxplot. The width of the violin at different points represents the probability density of the data (meaning how likely values are at different points). A wider section means more data points are concentrated there. Many violin plots contain a mini boxplot inside, showing the median, quartiles and sometimes outliers. The ggplot library helps with the creation of the violin plot.



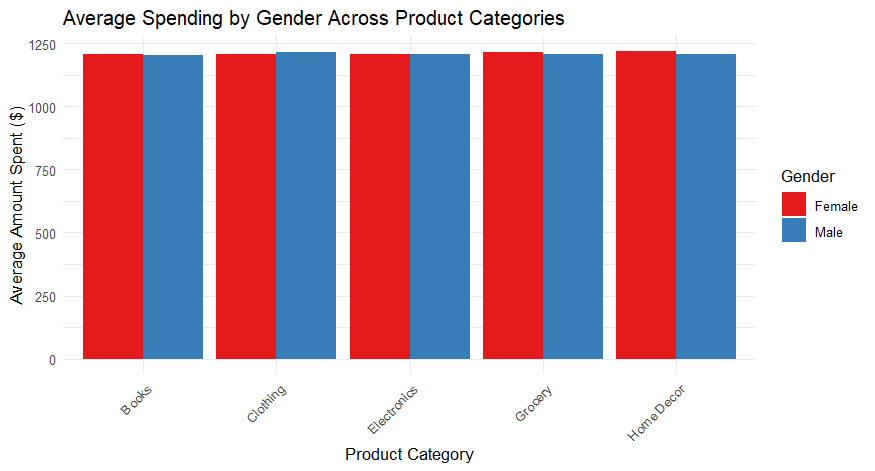
Code Snippet Result 32: Analysis 2-3 (Violin Plot)

The resulting code shows the distribution of amount spent per genders in the dataset. It can be evaluated that the distribution of the amount spent per gender is relatively similar, therefore both male and female genders both have similar spending amount per transaction. The distribution on both gender category shows identical peaks and identical boxplots as well. Therefore, we can conclude that both males and females have similar purchasing power.



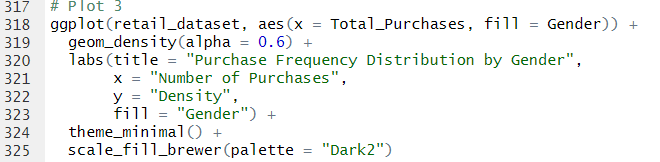
Code Snippet 19: Analysis 2-3 (Bar Chart)

The code above denotes the implementation of a bar chart to understand the relationship between the total amount spent amongst genders per product category. The implementation of the code first groups gender and product category together and the average mean spend value is calculated as a result for each group. The .groups function ensures the grouping is removed after the summarization.



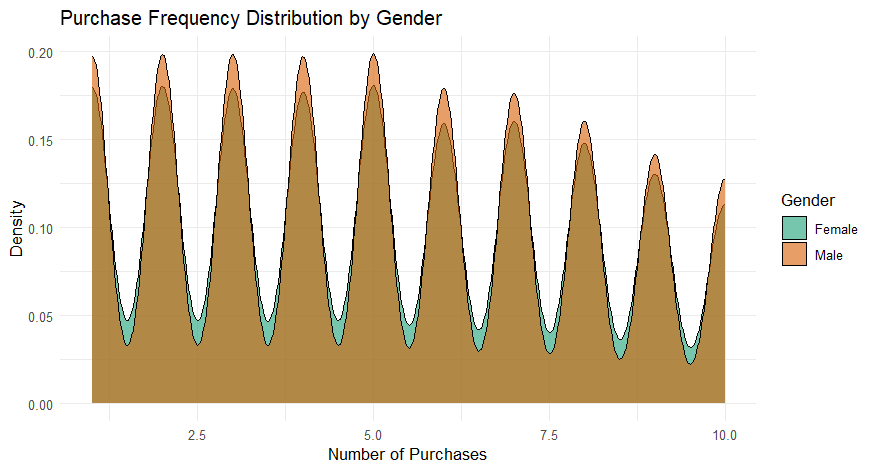
Code Snippet Result 33: Analysis 2-3 (Bar Chart)

The resulting information shows the purchasing power of genders distributed across different product categories. From the graph, we can see that both male and females have nearly identical spending amount for ‘Electronics’ and ‘Books’. However, the data shows that males spend more on average than females on ‘Clothing’ and females spend slightly more on ‘Home Décor’ and ‘Grocery’.



Code Snippet 20: Analysis 2-3 (Density Plot)

The code above shows the implementation of a density plot to show the quantity per transaction for customers for each gender. The code takes in the ‘Total\_Purchases’ variable as its x-axis and is plotted for both genders (using the fill function).

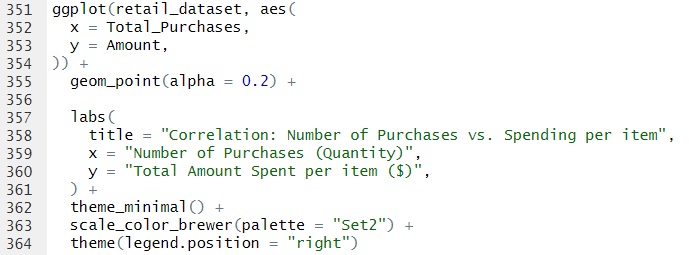


Code Snippet Result 34: Analysis 2-3 (Density Plot)

The density plot shown above shows no relationship between purchasing quantity and gender as the peaks of both gender distributions are nearly identical and there is no data to show any difference in skewness of density plots for both genders. Therefore, we can conclude that while both genders have a disparity in income, nevertheless, both genders do have identical purchasing power.

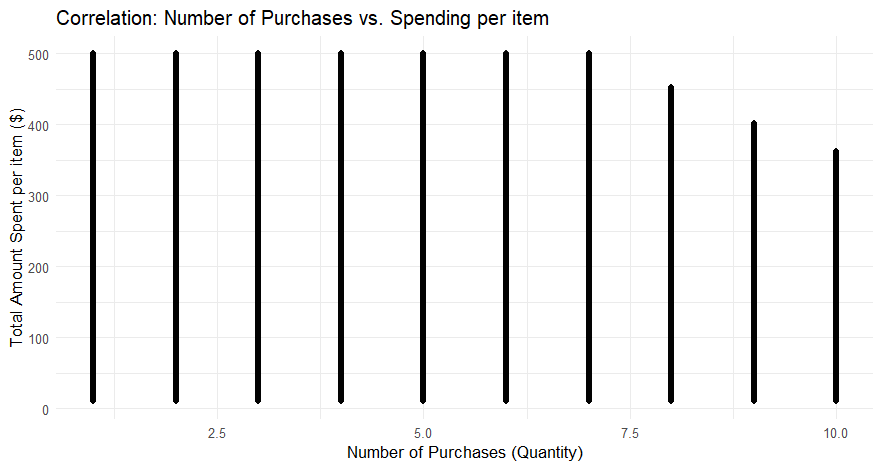
### *To rank product categories by profitability (Total Revenue) and customer engagement*

#### Analysis 3-1: Does a high number of transactions correlate with high purchase amount (Amount Spent)?



Code Snippet 21: Analysis 3-1

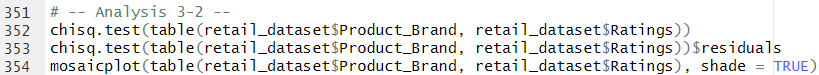
The code above is used to create a scatter plot to identify the relationship between purchase amount per product and quantity purchased. The developer used the ggplot2 library to extract the geom\_point function to draw the scatter plot and assign ‘Total\_Purchases’ and ‘Amount’ for the x and y axis respectively.



Code Snippet Result 35: Analysis 3-1

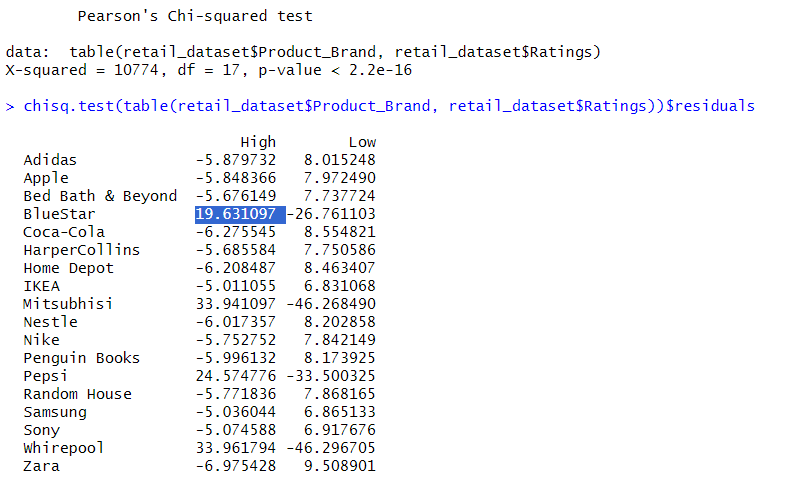
The result of running the code shows that there exists no correlation between purchase amount and quantity per item purchased. However, the scatter plot shows that when the quantity purchased goes higher than 7, the amount spent per item drops, therefore showing a negative correlation. In other words, the customer is likely to increase the purchase quantity of an item given that it is cheaper. This is only applied when the price of the item falls below USD 500.

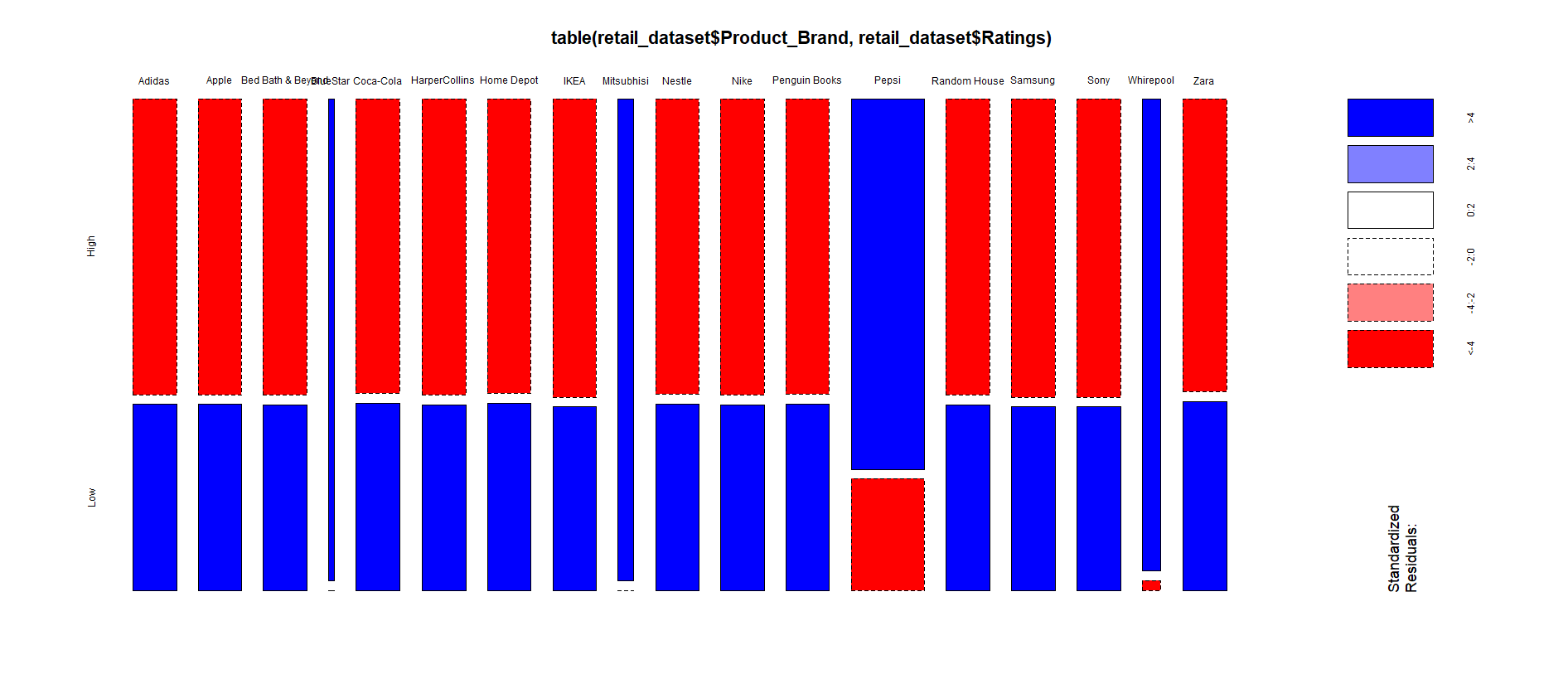
#### Analysis 3-2: Is there a relationship between “Product Brand” and “Ratings” within categories?



Code Snippet 22: Analysis 3-2

The code above once again shows the implementation of the Chi-Squared test on the dataset, particularly with regards to ‘Product\_Brand’ and ‘Rating’ variables to identify the relationship between them. The residuals are extracted and plotted on a mosaic plot.



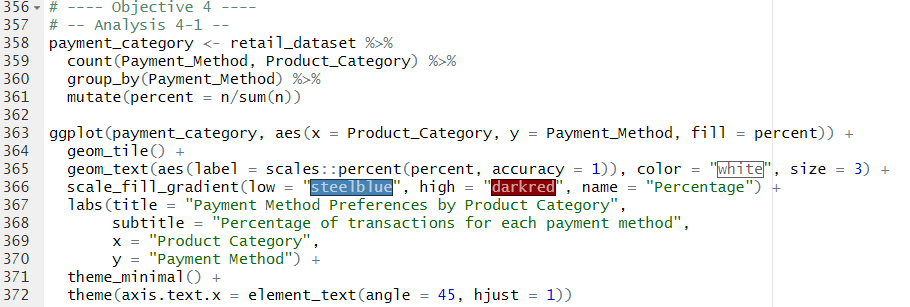


Code Snippet Result 36: Analysis 3-2

The resulting mosaic plot shows that there exists a relationship between the brand of the product and its rating, particularly for certain brands such as “Mitsubishi”, “Pepsi”, “Whirepool” and “Bluestar”. While the data shows most brands receiving predominantly ‘Low’ ratings, the 4 brands that stood out (mentioned earlier) have shown extremely significant ‘High’ ratings comparatively. Therefore, the relationship between product brand and rating is certain, such that customers are likely to give these brands a ‘High’ rating while providing a ‘Low’ rating to other brands.

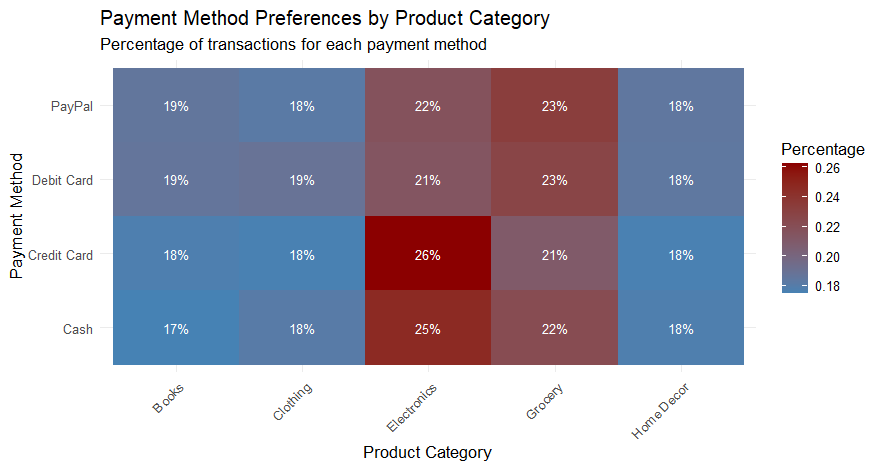
### *To assess how payment method correlates with spending habits*

#### Analysis 4-1: Are certain “Product Categories” related to specific payment methods?



Code Snippet 23: Analysis 4-1

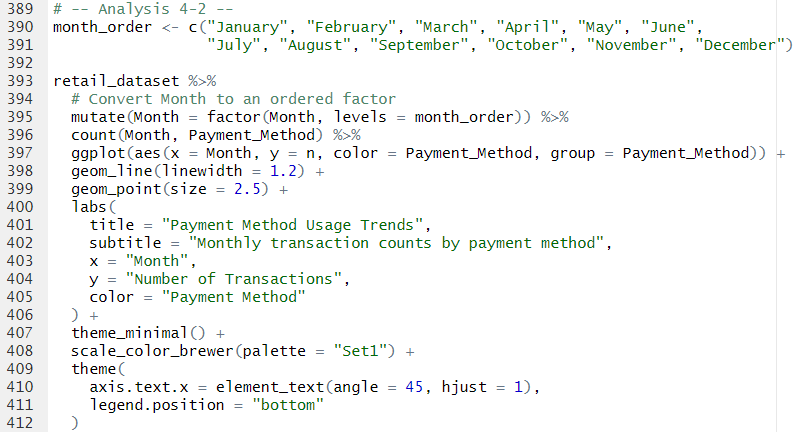
The code above shows the implementation of a heatmap matrix to denote the relationship between payment method and product category, particularly with how the product category impacts the payment method for the transactions made. The code first creates a ‘payment\_category’ table by setting up the pipeline to extract the number of transactions for each unique combination of payment method and product category. The data is then grouped by payment method and the percentage of transactions for each product category with each payment method is made. The data is then plotted.



Code Snippet Result 37: Analysis 4-1

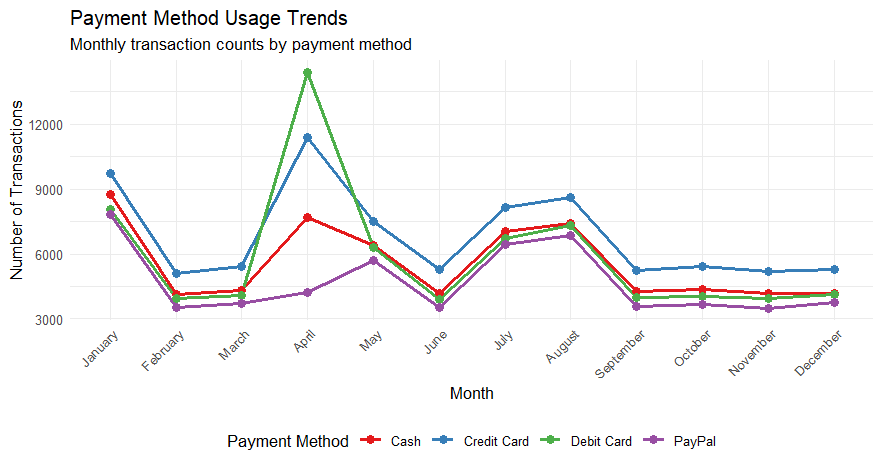
The result of the code snippet shows the promising results, particularly with regards to the purchase of ‘Electronic’ products. Customers who purchase ‘Electronics’ are likely to use credit cards or cash compared to other payment methods and debit cards as well as PayPal are likely to be used to pay for groceries or books. Apart from that, there exists no relationship between the product category and payment method. This is mostly regarding ‘Clothing’ and ‘Home Décor’.

#### Analysis 4-2: Does the customer use of a specific payment method change over the year?



Code Snippet 24: Analysis 4-2

The code shows the implementation of a multiple line plot of the number of transactions made each month alongside their payment method. The code first orders the data according to the months, defined by the month\_order vector created on line 376. A pipeline is created which first creates a new column named ‘Month’ to get the ordered month of the Data Frame. Subsequently, the count of ‘Month’ and ‘Payment Method’ is plotted using the geom\_line function.

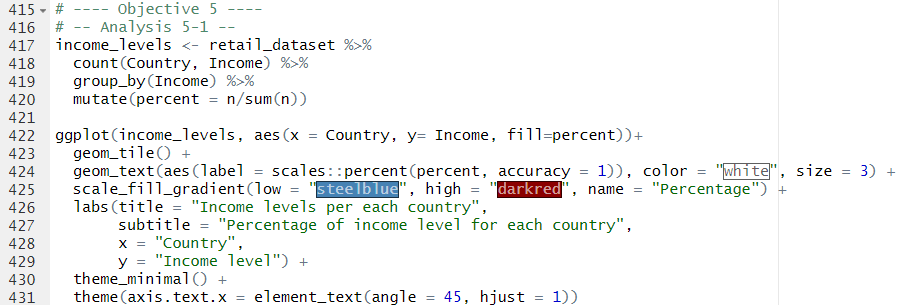


Code Snippet Result 38: Analysis 4-2

The resulting code shows the number of transactions per month, each separated by their payment method. From the graph, we can see that April was the month where the purchasing power was at its peak, followed by January, August and July in order. It can also be seen that a spike in Debit Card payment is significant in April despite Credit Cards being in the lead throughout the year.

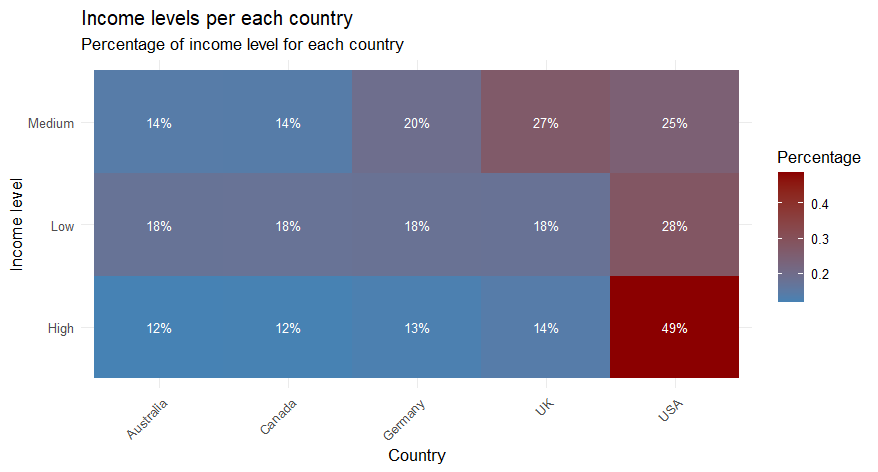
### *To map purchasing power and preferences by geographic region*

#### Analysis 5-1: What are the income levels of customers in each country?



Code Snippet 25: Analysis 5-1

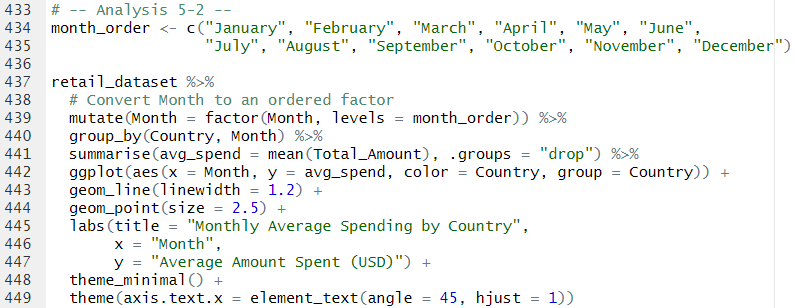
The code above implements a heatmap matrix to show the distribution of income levels per country. First, the retail dataset is extracted, and the count function is used to count the income levels per each country. The data is then grouped based on Income levels and a new column “percent” is created to extract the percentage of customers from a country in that income level



Code Snippet Result 39: Analysis 5-1

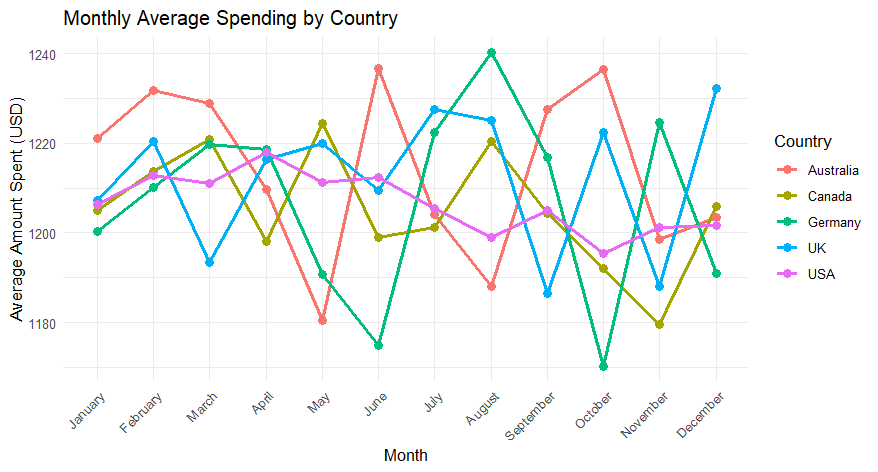
The heatmap matrix shows that the majority of ‘High’ and ‘Low’ income customers are from the USA with a 49% majority of the population. However, customers from the UK fall into most ‘Medium’ income categories with a 27% majority. The statistics shows high disparity between the income categories in the USA as it has a high number of ‘Low’ and ‘High’ income customers. However, in countries such as Germany, Canada and Australia, there is a much lower income disparity among the population. The majority of the UK fall into the ‘Medium’ income category while there is a slighter edge in more people being in the ‘Low’ income category than ‘High’ income category.

#### Analysis 5-2: Is there a difference in purchasing power (Amount Spent) over the year in each country?



Code Snippet 26: Analysis 5-2

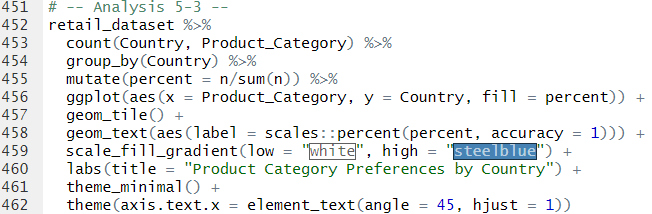
The code above shows the implementation of a line plot to show the total purchasing amount per country over the year. The code requires the correct order of the months; therefore the developers create a vector to set the order of the months and a mutate function is called to create a new column within the order of the months. The code then groups the data by the ‘Country’ and ‘Month’ columns and a summary of average spending per each grouped country and month is created using the summarise function. The data is then plotted using the ggplot function, where the x-axis represent the months, y-axis represent the average spending per country.



Code Snippet Result 40: Analysis 5-2

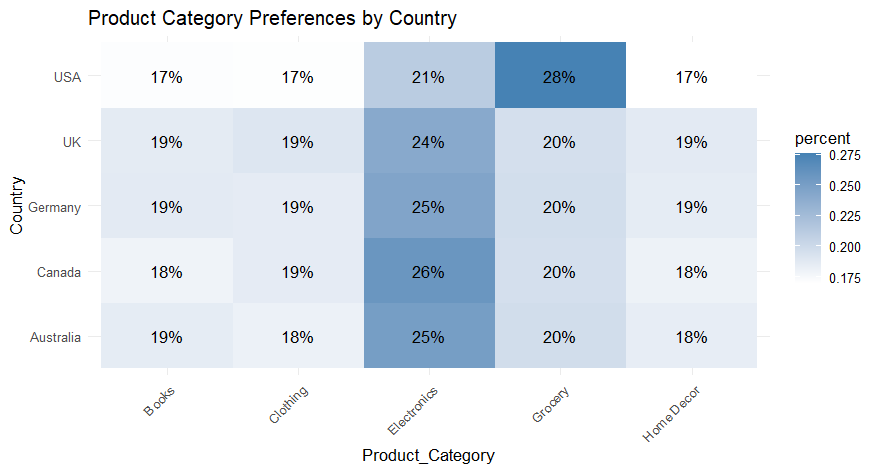
The resulting graph shows the seasonal spending patterns of customers per country. In Australia, customer spending is at its highest in January, February, March, June, September and October compared to other countries, whereas Canada in May, Germany in April, August and November and the UK in July and December. As for the US, it holds no regional trends compared to other countries, but it does see a decline in average spending over the year. This gives us the intuition for future stakeholders on when and where regional trends are at its peak to stock up its production and boost its advertising.

#### Analysis 5-3: Are there regional trends in Product Category preferences?



Code Snippet 27: Analysis 5-3

The code above implements a heatmap matrix to show the relationship between product category preferences per country. The code counts the category preferences per country using the count function and subsequently groups the data per country. The mutate function creates the percentage column using the mutate function to show the ratio of customers who purchased a specific product category from the total purchases in that country.

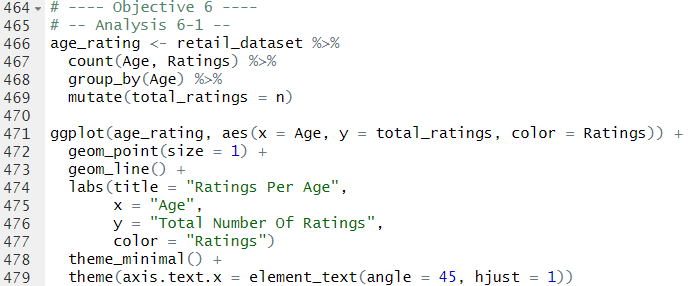


Code Snippet Result 41: Analysis 5-3

The plot above shows the distribution of purchases of product categories per country. The heatmap shows customers in the USA spend mostly on Groceries with a majority of 28%, followed by Electronics with 21% of purchases. As for the UK, Germany, Canada and Australia, most customers purchase Electronics while other product categories share an equal distribution of number of purchases.

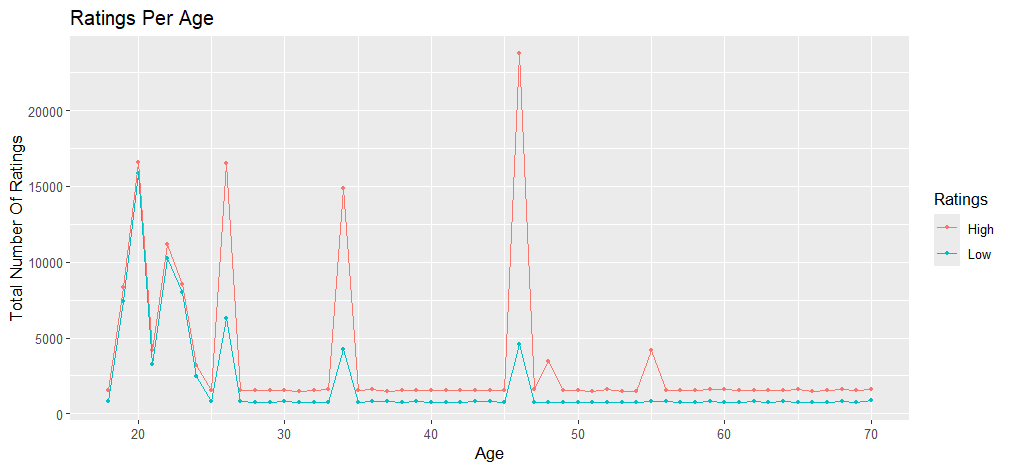
### *To uncover the age-based trends in product ratings*

#### Analysis 6-1: Is there an underlying relationship between age and product ratings?



Code Snippet 28: Analysis 6-1

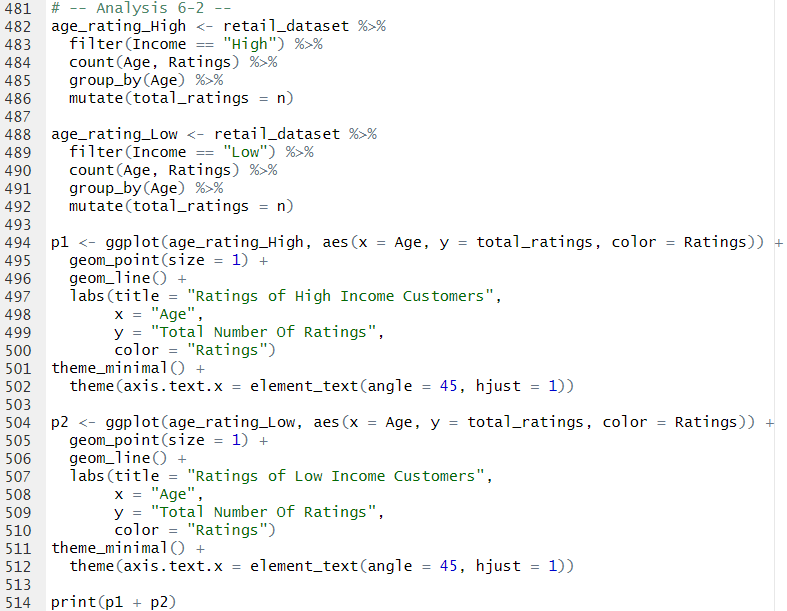
The code above shows the implementation of a line plot to show the relationship between age and rating per customer. The code counts the ratings per age category using the count function and then groups the data based on age while a total\_ratings column is created to calculate the total number of ratings. The data is plotted using the ggplot function.



Code Snippet Result 42: Analysis 6-1

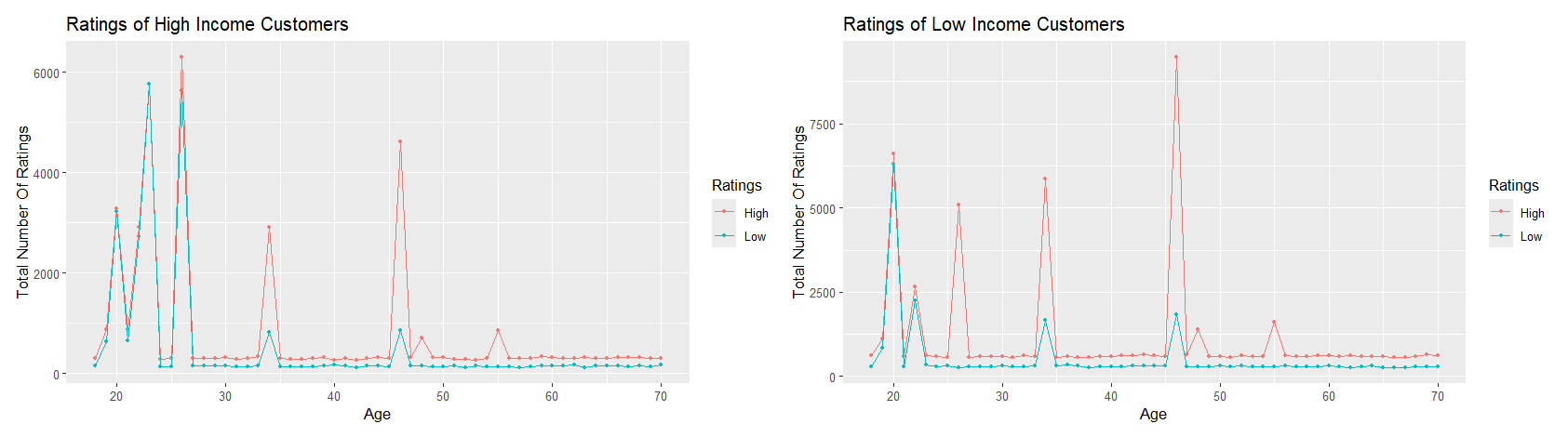
The line graph shows that while most of the population provides a ‘High’ rating, the graph displays a very close margin in ratings for people aged between 10 – 25 years old whereby, an even distribution of ‘High’ and ‘Low’ ratings is made. However, for customers aged 27, 35 and 46, there is a much greater margin between the ratings where most customers provided ‘High’ ratings compared to ‘Low’ ratings. The rest of the data shows a constant difference in the number of ratings.

#### Analysis 6-2: Are product “Ratings” influenced by “Income Levels”?



Code Snippet 29: Analysis 6-2

Much similar to the code in Analysis 6-1, the code above utilizes two plots (p1 and p2) to plot the graphs side by side to show the difference in ratings per age in both ‘High’ and ‘Low’ income categories. There exists a filter function on line 483 and 489 to separate the graphs and highlight the respective income category.



Code Snippet Result 43: Analysis 6-2

The plots show a much better view in the ratings per age for each income category, whereby ‘High’ income customers aged between 20 – 25 gave more ‘Low’ ratings than ‘Low’ income customers. However, the graphs do also show that customers aged 20 gave more ‘Low’ ratings from ‘Low’ income categories. In the end, we can summarise that there exists a relationship between age and customer ratings, whereby customers aged between 20 – 25 years old tend to have much higher expectations of products than other age groups, especially if they fall in the ‘High’ income category.

## Extra Features

1. ***impute\_missing\_values() –* Robust Data Cleaning**

**Purpose:** Handles missing values by imputing mode for categorical columns and removing NA entries for critical IDs.

**Importance:**

* Ensure data is complete by replacing missing categorical values such as Income and Product Category with the most frequent value (mode).
* Drops rows with missing transactional data such as Customer\_ID, Amount, which preserves analysis integrity.
* Avoids bias in visualization such as box plots and bar charts by ensuring no gaps in key variables like Income or Product Category.

1. ***chisq.test() with Mosaic Plots* – Categorical Relationships**

**Purpose:** Tests dependencies between categorical variables such as Income, Payment Method, Gender and Customer Segment.

**Importance:**

* Revealed statistically important relationships such as:
  + Low-income customers preferring credit cards
  + Females being more likely “Premium” members (via residuals in mosaic plots).

**Impact:**

* Quantified gender-income disparities and payment method preferences, therefore driving targeted marketing recommendations.

1. ***geom\_tile() for Heatmaps* – Brand/Income & Country Analysis**

**Purpose:** Visualizes proportional relationship such as Product Brand popularity by Income or Country

**Importance:**

* Highlighted brand-specific trends:
  + Mitsubishi and Pepsi were favoured by low-income customers.
  + The USA showed extreme income disparity (49% low/high income).

**Impact:**

* Guided regional inventory strategies such as stocking Mitsubishi in low-income areas.

1. ***geom\_violin() for Density/Violin Plots* – Gender Spending**

**Purpose:** Compares the distribution of numerical variables such as Total Amount across categories such as Gender.

**Importance:**

* Confirmed no gender-based differences despite income disparities.
* Showed males spend more on clothing, females on groceries (using geom\_col())

**Impact:**

* Debunked assumptions about gender driven purchasing power, emphasising product-category targeting instead.

1. ***geom\_line + Faceting* – Age/Rating Graphs**

**Purpose:** Analyses how ratings vary by Age and Income such as high-income young customers give more low ratings.

**Importance:**

* Identified critical age groups (20 – 25 years) as harsh reviewers, especially in high-income brackets.
* Suggested quality improvements for electronics/books targeting younger demographics.

**Impact:**

* Directly linked customer satisfaction to age and income, informing product development

# **Conclusion**

The study examined multiple elements of customer purchasing activity and income-based variations along with gender proportions and product assessment procedures to reveal important discoveries. Studied data **shows income levels create no substantial change in customer purchasing behaviours other than isolated brand preferences between Mitsubishi and Pepsi**. Gender dynamics within income distribution show male domination of high-income areas together with female dominance of middle-income zones **but both groups have similar purchasing capacities.**

Research findings reveal that age alongside income levels have a major impact on customer rating conduct. **High-income youth between 20 – 25 years old deliver more negative ratings to products because they possess elevated quality standards**. The ratings given to **Mitsubishi and Whirlpool turned out higher than other brands which demonstrates how consumers value brand perception when determining satisfaction**. Payment methods demonstrated direct income-related trends because customers from lower-income backgrounds mostly chose credit cards yet customers from higher-income groups chose PayPal.

The USA exhibits the maximum economic inequality between its citizenry because its customer base comprises equal proportions of both low and high earners, yet most UK customers earn medium-income levels. The timing of peak spending shifted between different countries which could benefit companies through strategic marketing plans.

**Recommendations:**

* Businesses should focus on upgrading product quality alongside customer support for younger high-income customers since they tend to award lower ratings.
* Mitsubishi and Pepsi companies should market their products by highlighting the positive perceptions about their products that exist among consumers. Other types of brands need to study their rating issues as their customer feedback primarily shows negative markings.
* Businesses need to synchronize their inventory and marketing activities with the busiest buying months of each nation (Germany experiences peak buying during April while the United Kingdom peaks in December).
* Payment options can be improved by encouraging high-income customers to use PayPal and developing financial incentives using credit cards for low-income groups to boost customer satisfaction.

**Limitations and Future Directions:**

* The data only contained transactional data which means it is missing key subjective elements including emotional attachment to brands along with loyalty indicators. Further studies should adopt sentiment analysis of customer reviews to evaluate how ratings differ from each other. The evolution of customer spending behaviours over time with respect to economic developments needs further investigation through longitudinal research methods. Researchers should implement experimental testing of pricing strategies to establish if bulk discounts beyond seven products consistently increase sales volumes.
* Income levels alongside gender differences shape some aspects of human conduct but purchasing patterns show consistent patterns across different populations. The wide disparity in customer ratings and brand identities requires businesses to design specific improvements in product standards combined with customer experience initiatives. Operational optimization and customer satisfaction together with revenue growth becomes attainable for businesses that use these insights.

# **References**

Chaudhary, S. (24 January, 2024). *Why 1.5 Is Used in the IQR Rule for Outlier Detection*. Retrieved from builtin: https://builtin.com/articles/1-5-iqr-rule#:~:text=The%20IQR%20method%20of%20outlier,data%20is%20considered%20an%20outlier.

Ganti, A., Rasure, E., & Kvilhaug, S. (28 February, 2024). *Degrees of Freedom in Statistics Explained: Formula and Example*. Retrieved from Investopedia: https://www.investopedia.com/terms/d/degrees-of-freedom.asp

Green, J., Manski, S., Hansen, T., & Broatch, J. (2023). Descriptive statistics. *International Encyclopedia of Education (Fourth Edition)*, 723-733. doi:https://doi.org/10.1016/B978-0-12-818630-5.10083-1

jmp Statistical Recovery. (n.d.). *The Chi-Square Test*. Retrieved from Statistics Knowledge Portal: https://www.jmp.com/en/statistics-knowledge-portal/chi-square-test

retailcloud. (4 October, 2024). *Key Benefits of Using Data Analytics in the Retail Industry*. Retrieved from retailcloud: https://retailcloud.com/benefits-data-analytics-in-retail-industry/