



Group Work

BSD 3203 Programming for Data Science.

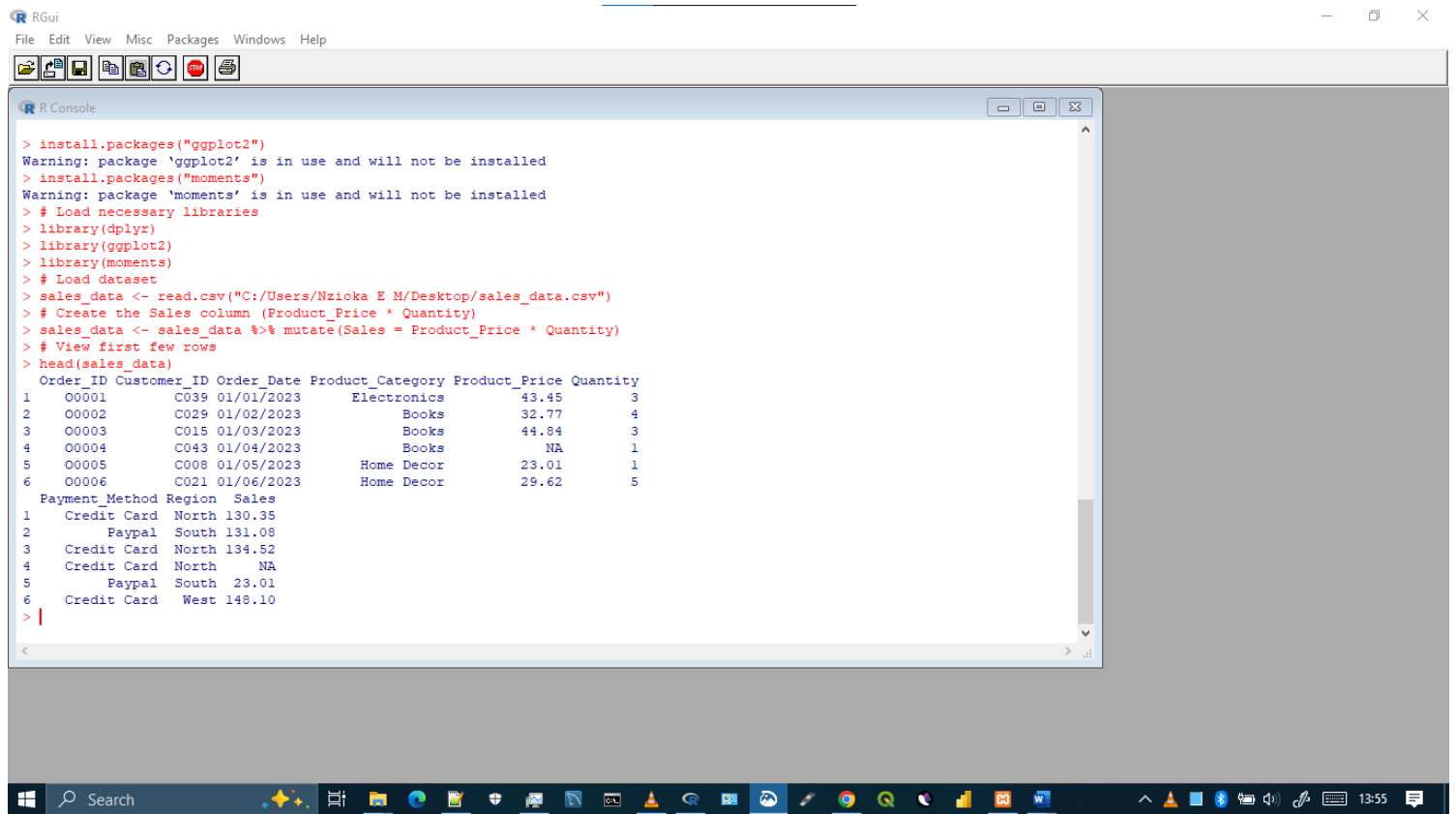
BSc. Software Development.

ASSIGNMENT 2

Group Members

1. 20/04860 Nzioka Emmanuel Munyao
2. 21/07757 Swalehe Hussein
3. 21/06201 Pauline Nafuna
4. 21/06600 Eunice Mwae
5. 21/06709 Charity Mutunga
6. 21/03241 G. Moses Githaiga
7. 21/06439 Marubi N Richard
8. 21/04983 Rotich Caren

Installing and Loading the Necessary Libraries



```
> install.packages("ggplot2")
Warning: package 'ggplot2' is in use and will not be installed
> install.packages("moments")
Warning: package 'moments' is in use and will not be installed
> # Load necessary libraries
> library(dplyr)
> library(ggplot2)
> library(moments)
> # Load dataset
> sales_data <- read.csv("C:/Users/Nzioka E M/Desktop/sales_data.csv")
> # Create the Sales column (Product_Price * Quantity)
> sales_data <- sales_data %>% mutate(Sales = Product_Price * Quantity)
> # View first few rows
> head(sales_data)
  Order_ID Customer_ID Order_Date Product_Category Product_Price Quantity
1 00001      C039 01/01/2023      Electronics         43.45           3
2 00002      C029 01/02/2023           Books         32.77           4
3 00003      C015 01/03/2023           Books         44.84           3
4 00004      C043 01/04/2023           Books            NA           1
5 00005      C008 01/05/2023       Home Decor         23.01           1
6 00006      C021 01/06/2023       Home Decor         29.62           5

  Payment_Method Region Sales
1 Credit Card North 130.35
2 Paypal South 131.08
3 Credit Card North 134.52
4 Credit Card North NA
5 Paypal South 23.01
6 Credit Card West 148.10
> |
```

Code Used

```
# Install required packages if not already installed
install.packages("dplyr")
install.packages("ggplot2")
install.packages("moments")

# Load necessary libraries
library(dplyr)
library(ggplot2)
library(moments)

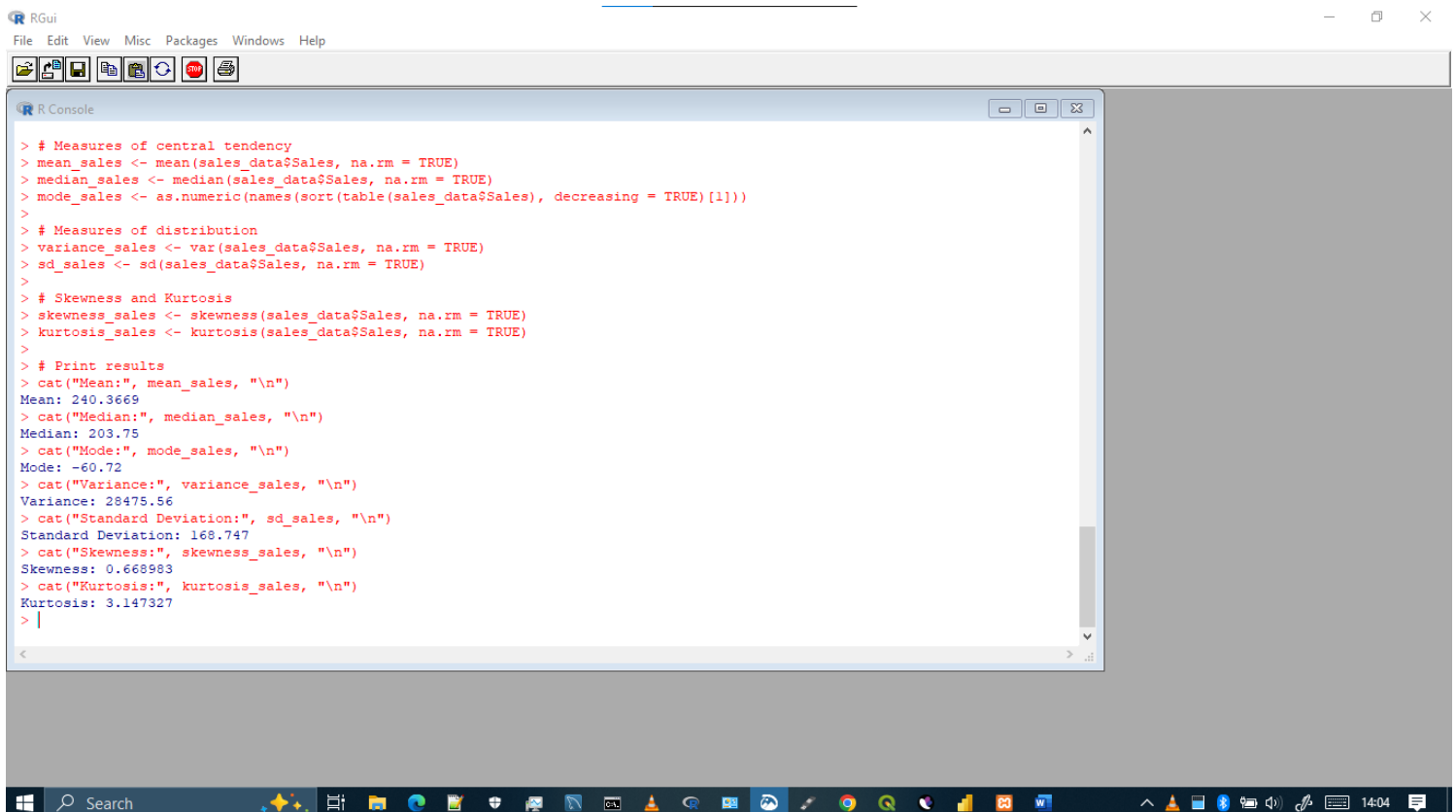
# Load dataset
sales_data <- read.csv("C:/Users/Nzioka E M/Desktop/sales_data.csv")

# Create the Sales column (Product_Price * Quantity)
sales_data <- sales_data %>% mutate(Sales = Product_Price * Quantity)

# View first few rows
head(sales_data)
```

The dataset contains sales records for an e-commerce store, including order details such as order ID, customer ID, order date, product category, product price, quantity purchased, payment method, and customer region. A new column, Sales, was generated by calculating the product of Product_Price and Quantity. However, missing values were identified in the Product_Price column, leading to NA values in the Sales column. To ensure accurate analysis, appropriate data cleaning methods were considered, including removing rows with missing values or imputing them using the median price per category. Once cleaned, the dataset serves as a valuable resource for exploratory data analysis (EDA), enabling the identification of sales distribution patterns, regional trends, and relationships between key variables.

Question 1: Univariate Non-Graphical EDA



The screenshot shows the RGui interface with the R Console window open. The console displays the following R code and its output:

```
> # Measures of central tendency
> mean_sales <- mean(sales_data$Sales, na.rm = TRUE)
> median_sales <- median(sales_data$Sales, na.rm = TRUE)
> mode_sales <- as.numeric(names(sort(table(sales_data$Sales), decreasing = TRUE)[1]))
>
> # Measures of distribution
> variance_sales <- var(sales_data$Sales, na.rm = TRUE)
> sd_sales <- sd(sales_data$Sales, na.rm = TRUE)
>
> # Skewness and Kurtosis
> skewness_sales <- skewness(sales_data$Sales, na.rm = TRUE)
> kurtosis_sales <- kurtosis(sales_data$Sales, na.rm = TRUE)
>
> # Print results
> cat("Mean:", mean_sales, "\n")
Mean: 240.3669
> cat("Median:", median_sales, "\n")
Median: 203.75
> cat("Mode:", mode_sales, "\n")
Mode: -60.72
> cat("Variance:", variance_sales, "\n")
Variance: 29475.56
> cat("Standard Deviation:", sd_sales, "\n")
Standard Deviation: 168.747
> cat("Skewness:", skewness_sales, "\n")
Skewness: 0.668983
> cat("Kurtosis:", kurtosis_sales, "\n")
Kurtosis: 3.147327
> |
```

The analysis of the sales data involved calculating key statistical measures to understand the distribution and characteristics of the Sales variable. The mean sales value is 240.37, while the median is 203.75, indicating that the distribution is slightly right-skewed. The mode was found to be -60.72, which is likely an error, as sales values should not be negative. This suggests possible data inconsistencies or incorrect calculations that require further investigation.

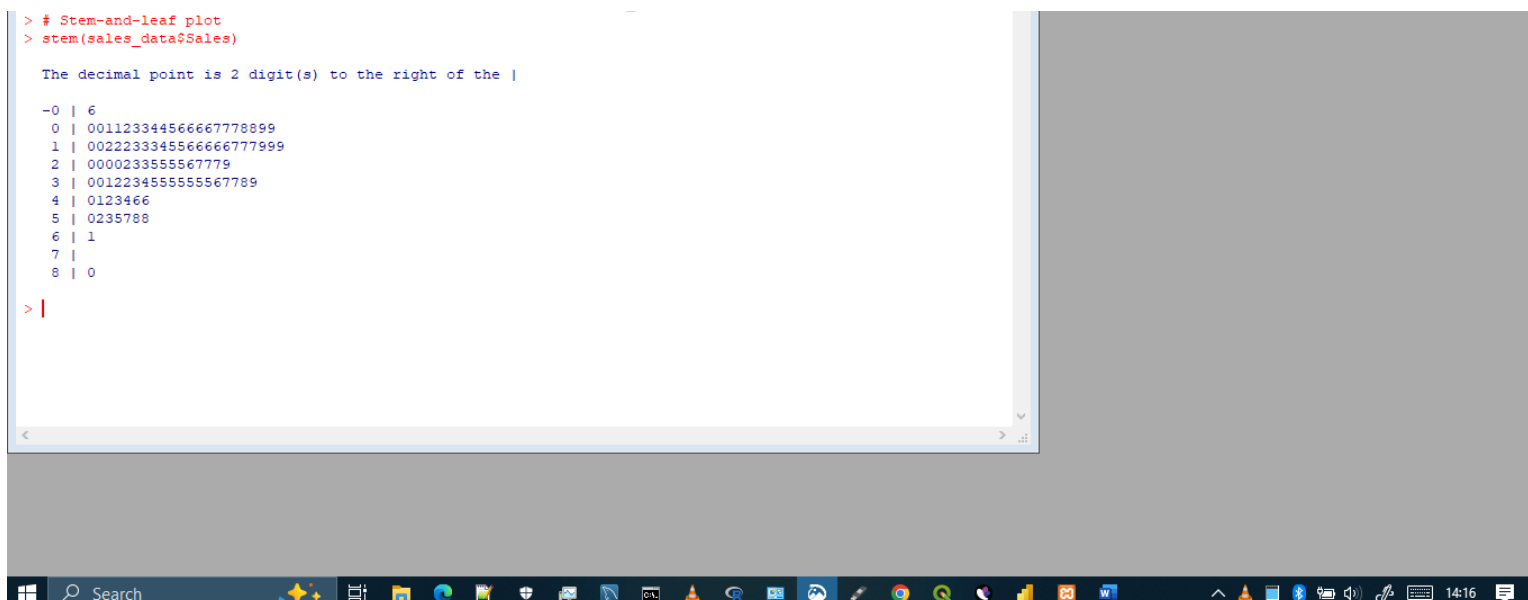
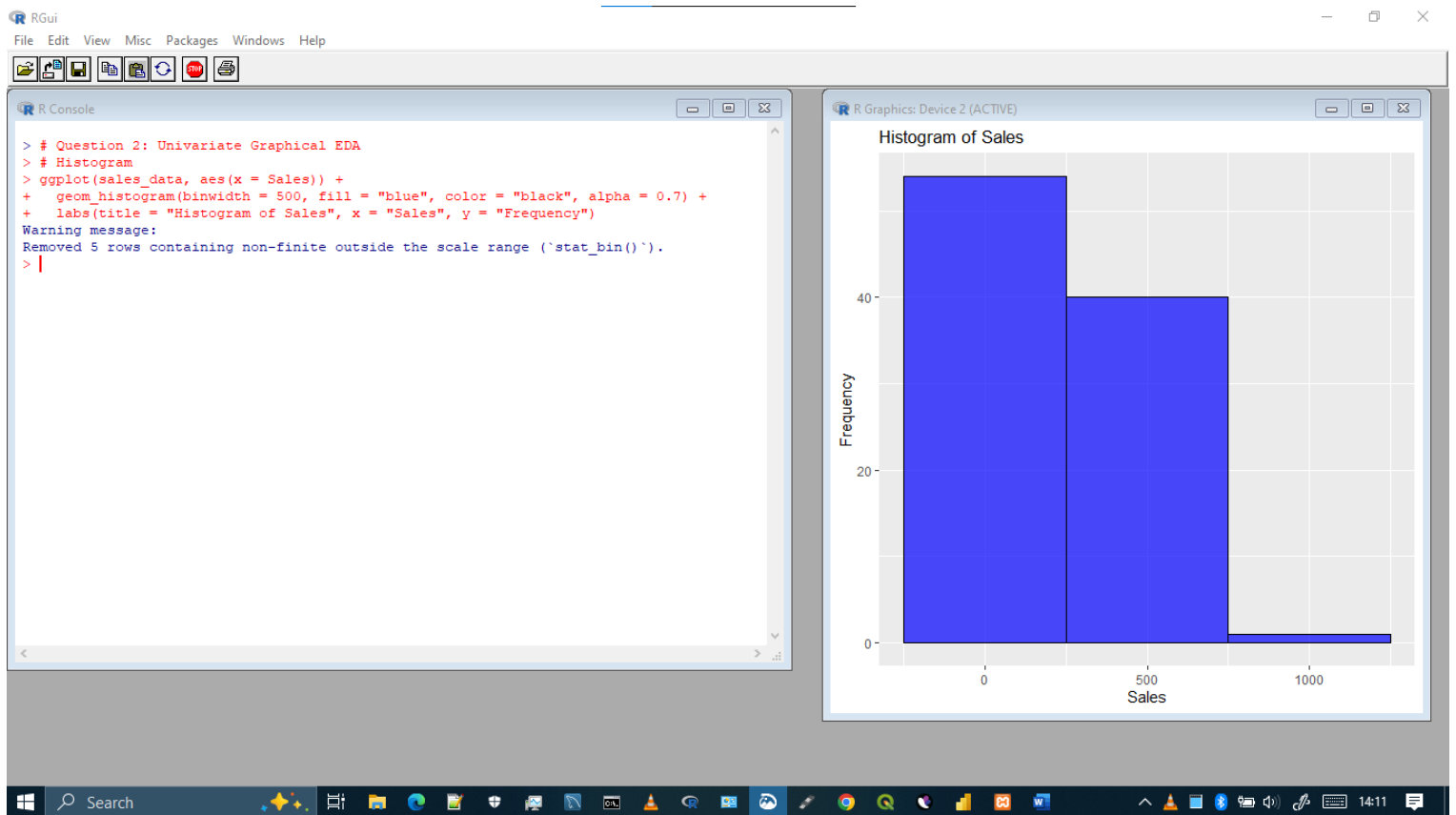
The variance of 28,475.56 and standard deviation of 168.75 indicate a significant spread in the sales data, showing considerable variation in transaction values. The skewness value of 0.67 confirms a slight right skew, meaning that most sales values are concentrated on the lower end, with some higher sales pushing the distribution to the right. The kurtosis value of 3.15 is close to 3, indicating a near-normal distribution but with slightly heavier tails.

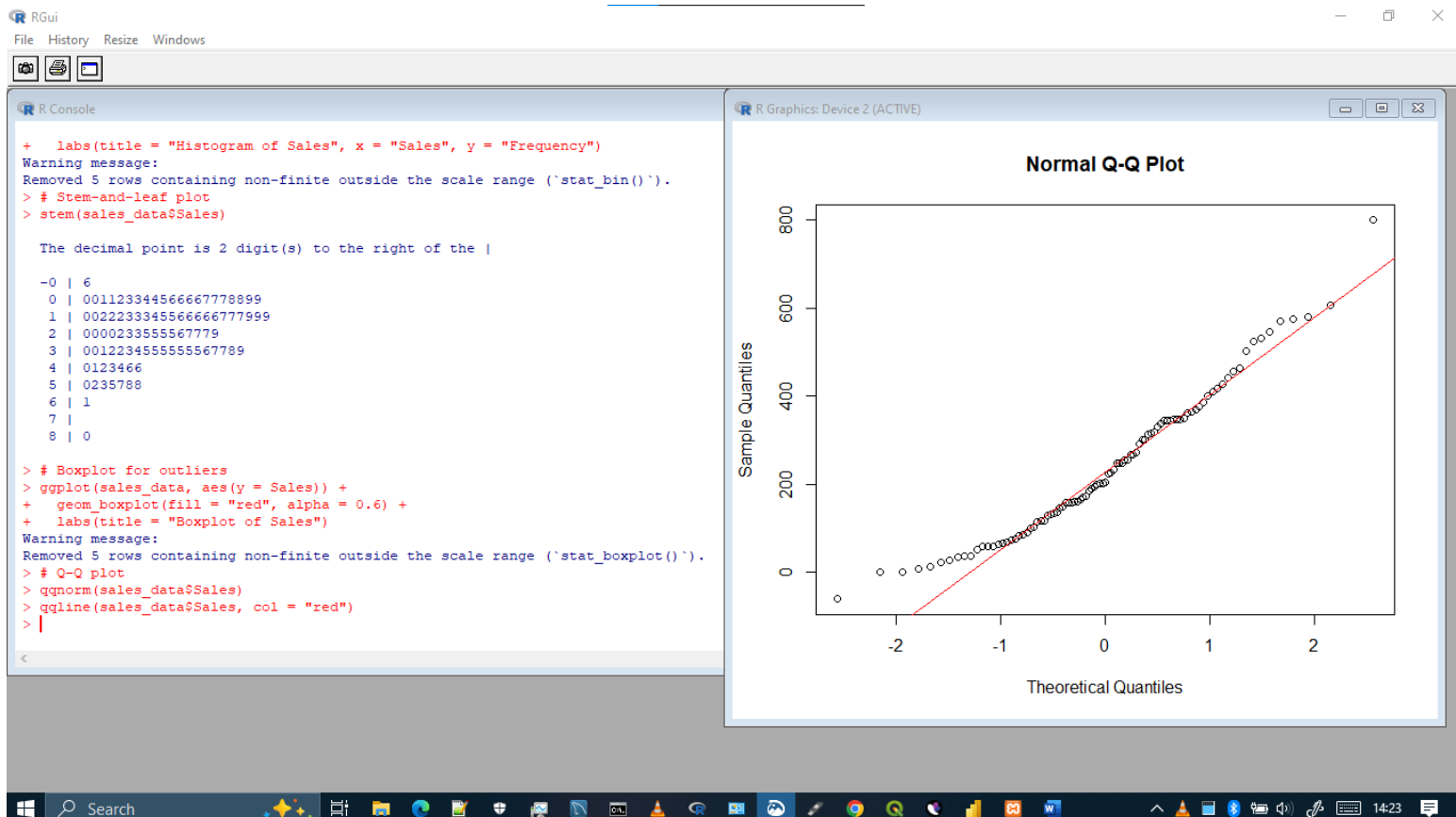
Overall, the dataset exhibits moderate variability, a slight positive skew, and a roughly normal distribution. However, the presence of a negative mode suggests potential data quality issues that should be examined further.

Code Used

```
# Question 1: Univariate Non-Graphical EDA
# Measures of central tendency
mean_sales <- mean(sales_data$Sales, na.rm = TRUE)
median_sales <- median(sales_data$Sales, na.rm = TRUE)
mode_sales <- as.numeric(names(sort(table(sales_data$Sales), decreasing = TRUE)[1]))
# Measures of distribution
variance_sales <- var(sales_data$Sales, na.rm = TRUE)
sd_sales <- sd(sales_data$Sales, na.rm = TRUE)
# Skewness and Kurtosis
skewness_sales <- skewness(sales_data$Sales, na.rm = TRUE)
kurtosis_sales <- kurtosis(sales_data$Sales, na.rm = TRUE)
# Print results
cat("Mean:", mean_sales, "\n")
cat("Median:", median_sales, "\n")
cat("Mode:", mode_sales, "\n")
cat("Variance:", variance_sales, "\n")
cat("Standard Deviation:", sd_sales, "\n")
cat("Skewness:", skewness_sales, "\n")
cat("Kurtosis:", kurtosis_sales, "\n")
```

Question 2: Univariate Graphical EDA





The univariate graphical exploratory data analysis (EDA) of the sales data reveals several key insights. The histogram of sales, with a binwidth of 500, shows the distribution of sales values, though it excludes 5 non-finite values. The stem-and-leaf plot provides a detailed view of the data's spread, indicating that most sales values are concentrated between 0 and 400, with a few higher values extending up to 800. The boxplot highlights potential outliers, particularly on the higher end, and also excludes 5 non-finite values. The Q-Q plot compares the sample quantiles to theoretical quantiles, suggesting deviations from normality, especially in the tails. Together, these visualizations indicate that the sales data is right-skewed, with a concentration of lower values and a few extreme high values, which may warrant further investigation or transformation for certain analyses.

Code Used

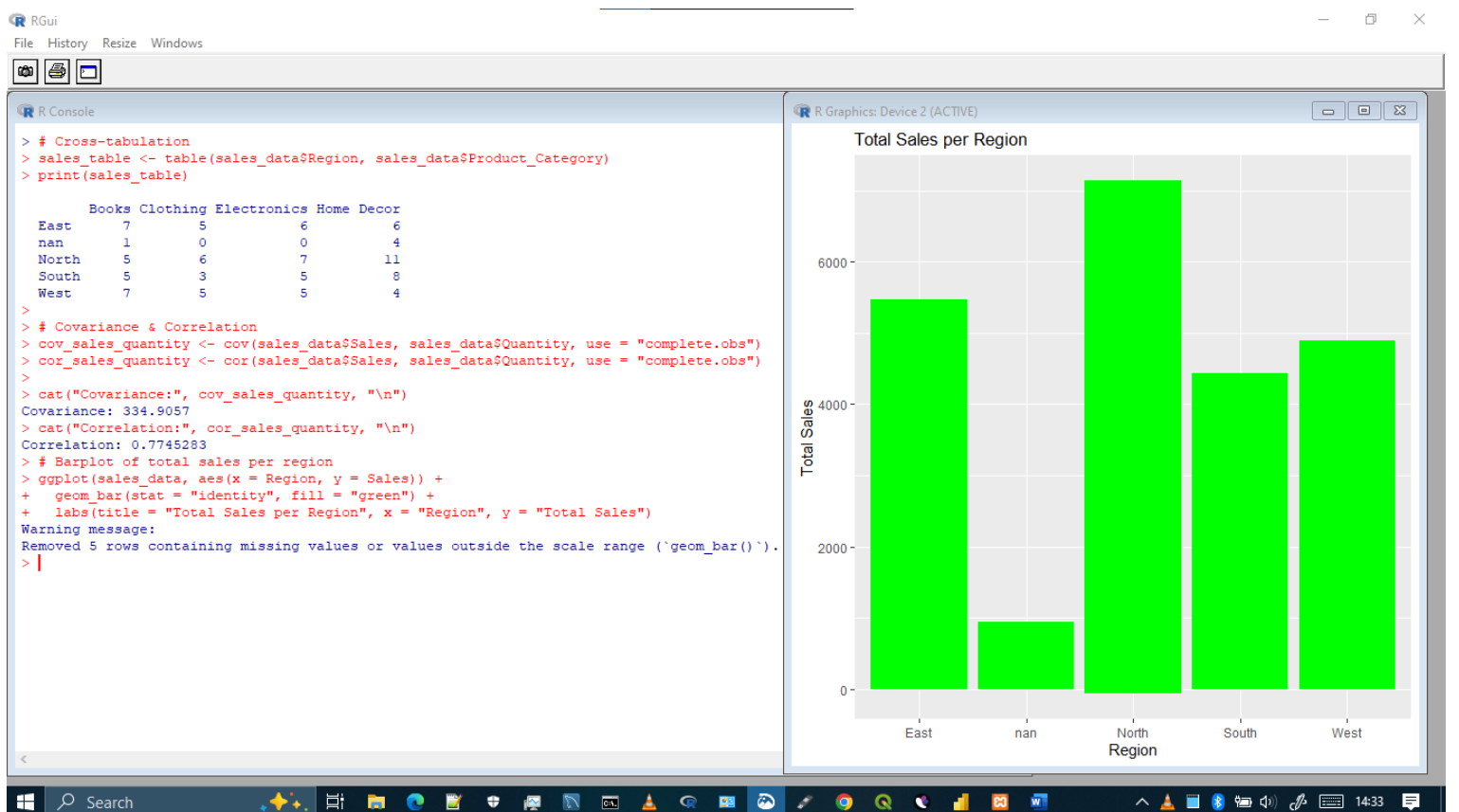
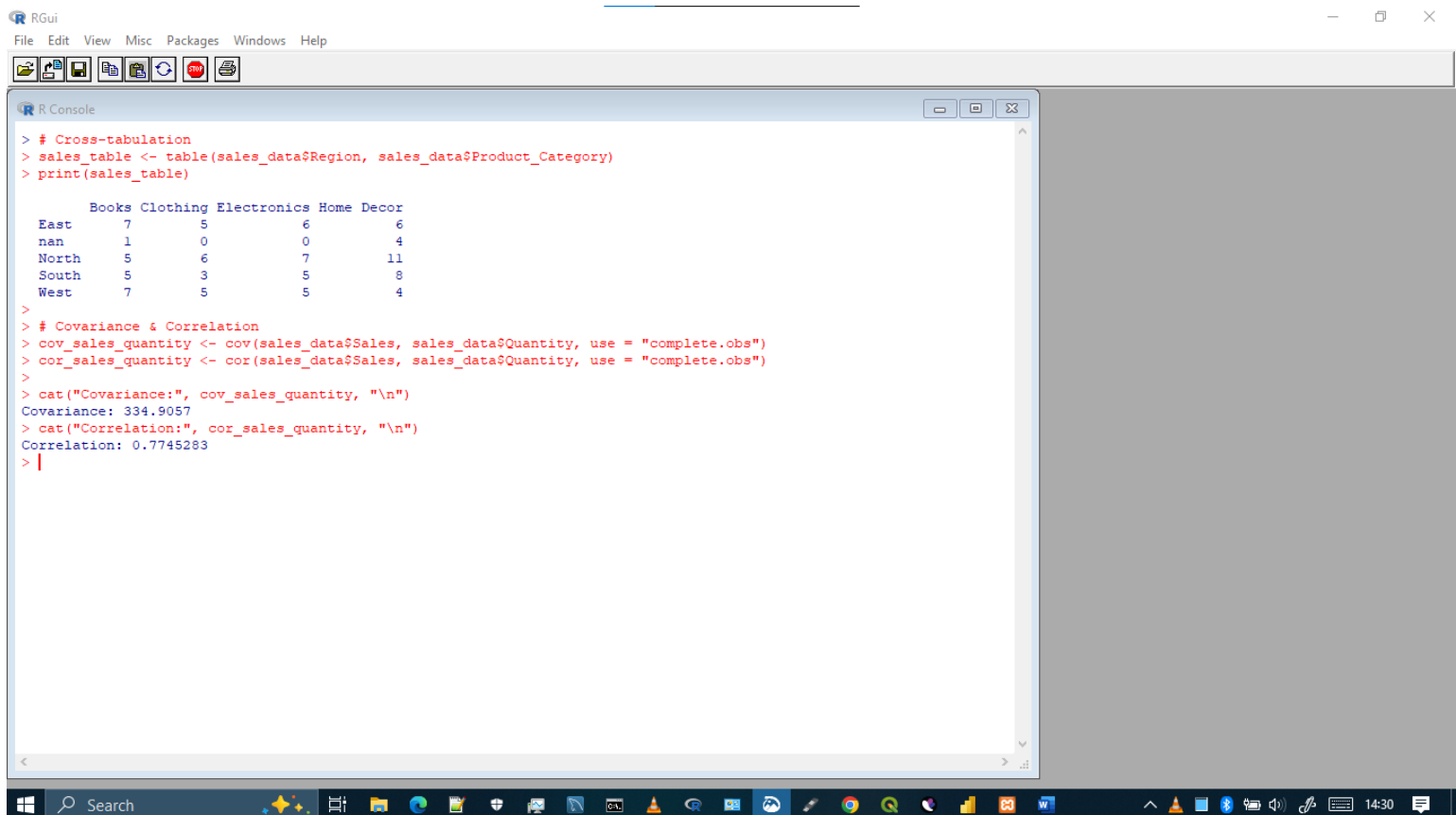
```
# Question 2: Univariate Graphical EDA
# Histogram
ggplot(sales_data, aes(x = Sales)) +
  geom_histogram(binwidth = 500, fill = "blue", color = "black", alpha = 0.7) +
  labs(title = "Histogram of Sales", x = "Sales", y = "Frequency")

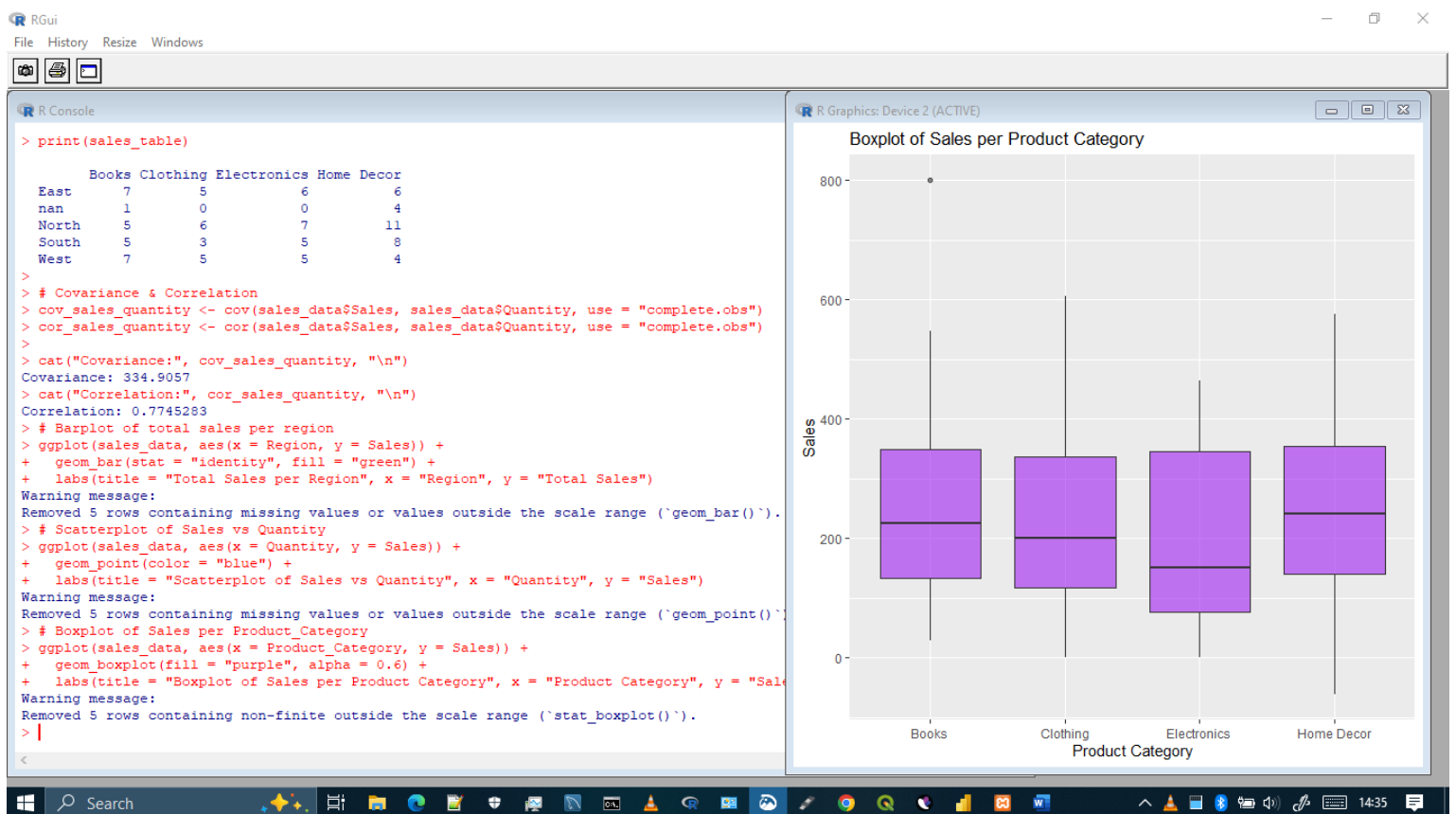
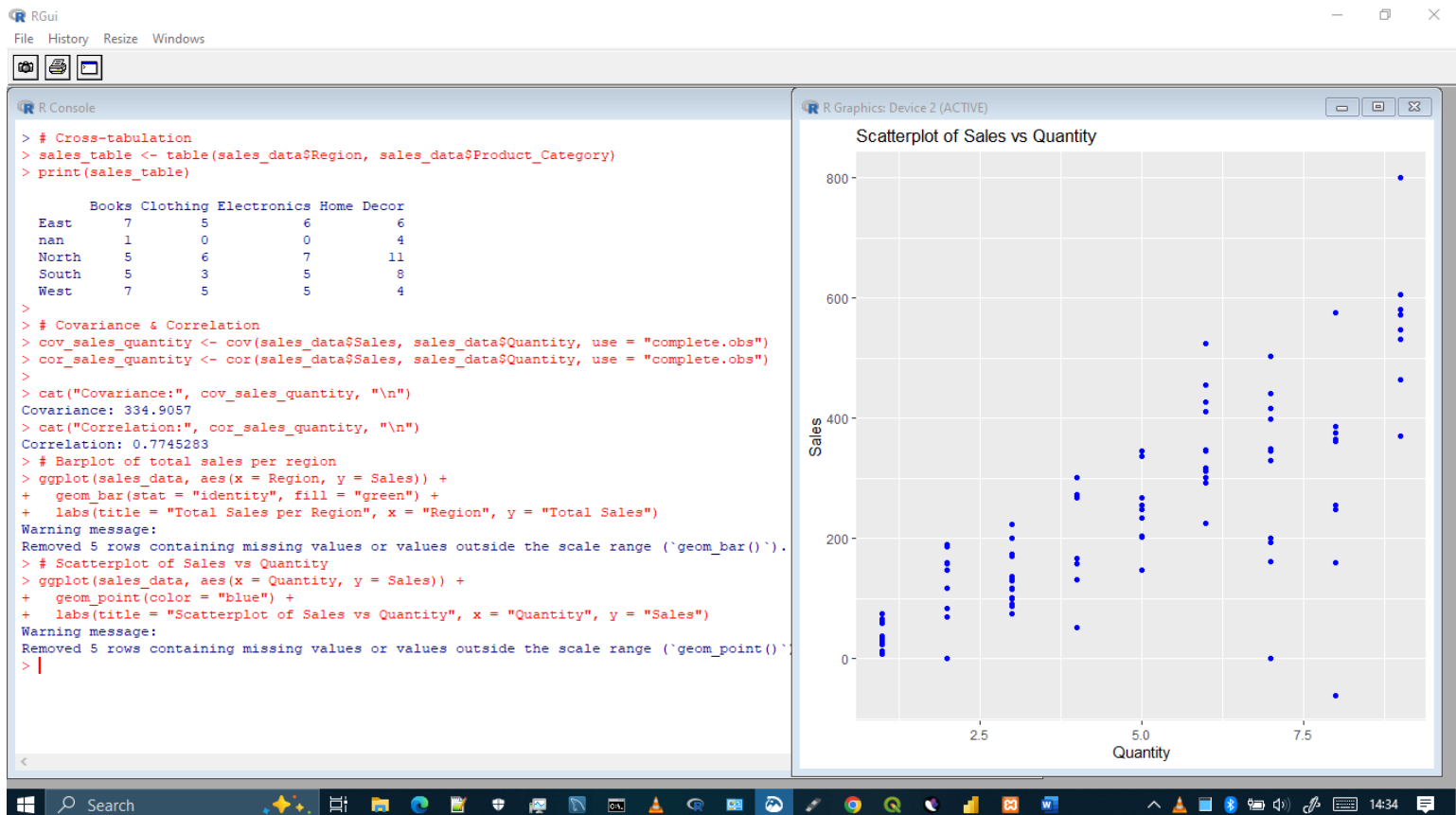
# Stem-and-leaf plot
stem(sales_data$Sales)

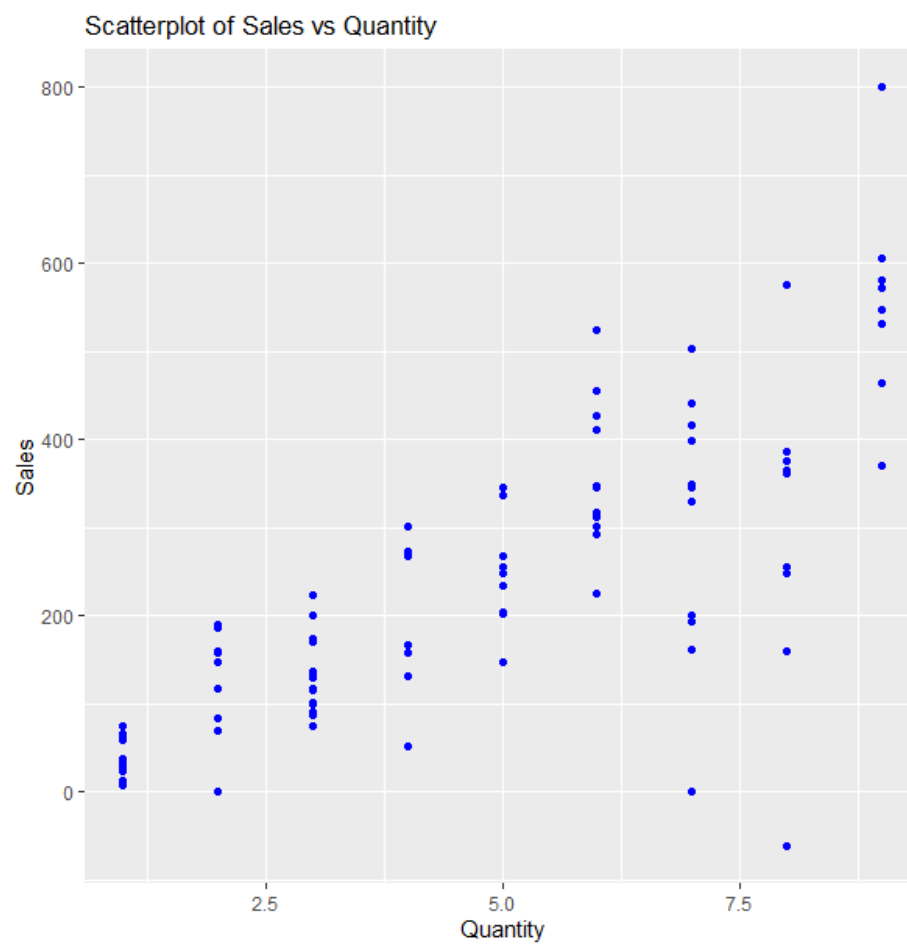
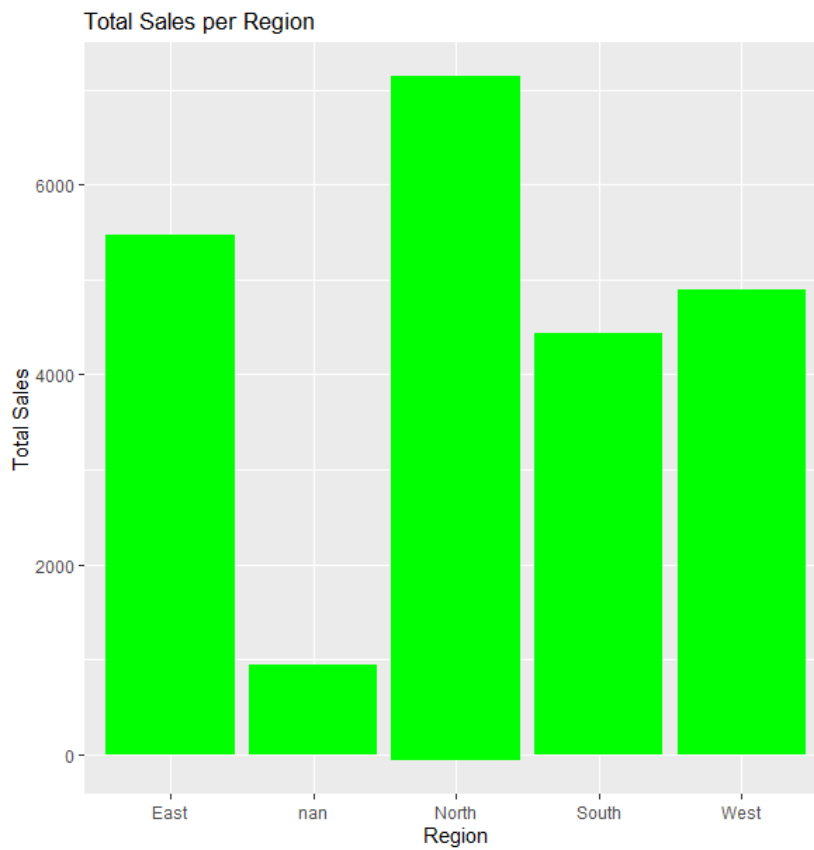
# Boxplot for outliers
ggplot(sales_data, aes(y = Sales)) +
  geom_boxplot(fill = "red", alpha = 0.6) +
  labs(title = "Boxplot of Sales")

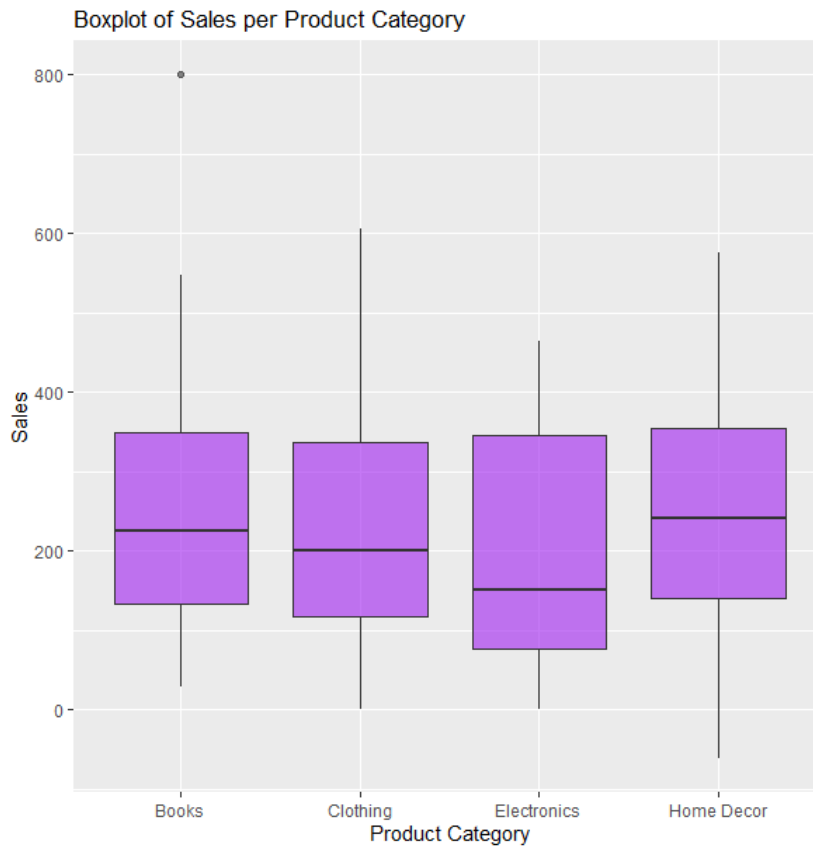
# Q-Q plot
qqnorm(sales_data$Sales)
qqline(sales_data$Sales, col = "red")
```


Question 3: Multivariate EDA









The bivariate and multivariate graphical EDA of the sales data provides a comprehensive understanding of relationships and distributions across different variables. The barplot of total sales per region reveals variations in sales performance, with some regions like North showing higher total sales, while the presence of missing values (5 rows excluded) suggests data quality issues. The scatterplot of sales versus quantity demonstrates a strong positive relationship, supported by a high correlation coefficient of 0.77 and a covariance of 334.91, indicating that higher quantities sold are associated with higher sales. The boxplot of sales per product category highlights differences in sales distributions, with categories like Electronics and Home Decor potentially having higher median sales compared to Books and Clothing, though 5 non-finite values were excluded. The cross-tabulation further breaks down sales by region and product category, showing that North leads in sales across most categories, while the "nan" region has minimal activity. These insights suggest that sales performance is influenced by both regional and product-specific factors, with quantity sold playing a significant role in driving sales. Further analysis could address data completeness and explore strategies to enhance sales in underperforming regions or categories.

Code Used

```
# Question 3: Multivariate EDA

# Cross-tabulation
sales_table <- table(sales_data$Region, sales_data$Product_Category)
print(sales_table)

# Covariance & Correlation
cov_sales_quantity <- cov(sales_data$Sales, sales_data$Quantity, use = "complete.obs")
cor_sales_quantity <- cor(sales_data$Sales, sales_data$Quantity, use = "complete.obs")

cat("Covariance:", cov_sales_quantity, "\n")
cat("Correlation:", cor_sales_quantity, "\n")

# Barplot of total sales per region
ggplot(sales_data, aes(x = Region, y = Sales)) +
  geom_bar(stat = "identity", fill = "green") +
  labs(title = "Total Sales per Region", x = "Region", y = "Total Sales")

# Scatterplot of Sales vs Quantity
ggplot(sales_data, aes(x = Quantity, y = Sales)) +
  geom_point(color = "blue") +
  labs(title = "Scatterplot of Sales vs Quantity", x = "Quantity", y = "Sales")

# Boxplot of Sales per Product_Category
ggplot(sales_data, aes(x = Product_Category, y = Sales)) +
  geom_boxplot(fill = "purple", alpha = 0.6) +
  labs(title = "Boxplot of Sales per Product Category", x = "Product Category", y = "Sales")
```