Semester 2, 2023/2024

TTTR6124 STATISTICAL METHODS FOR DATA ANALYTICS

Assignment 3

Prepared by:

Name	Matric Number
MOHAMED MOUBARAK MOHAMED MISBAHOU MKOUBOI	P139575

Lecturer:

ASSOC. PROF. DR. SHAHNORBANUN SAHRAN

Title of the research

Analysis of Supermarket Sales Data

Introduction

The number of supermarkets in most populated cities is expanding, resulting in intense market competition. Data analysis is critical in the service industry, particularly in supermarkets, to better understand sales patterns and customer behaviour.

Problem Statements

1. How do sales vary by branch and product line?
2. What are the sales patterns over time?
3. Do sales vary depending on consumer demographics?

Research Objectives

1. Analyze sales distribution across branches and product lines.
2. Identify sales trends during a specified period.
3. To study the effect of consumer demographics on sales.

Methodology

To transform and visualize the dataset in a way that people will understand, we need data analysts who perform data analytics to give such results. Data analytics collects, transforms, and organises data to conclude, predict results, and make informed decisions. Figure 1 demonstrates how we analysed data from a supermarket sales dataset.

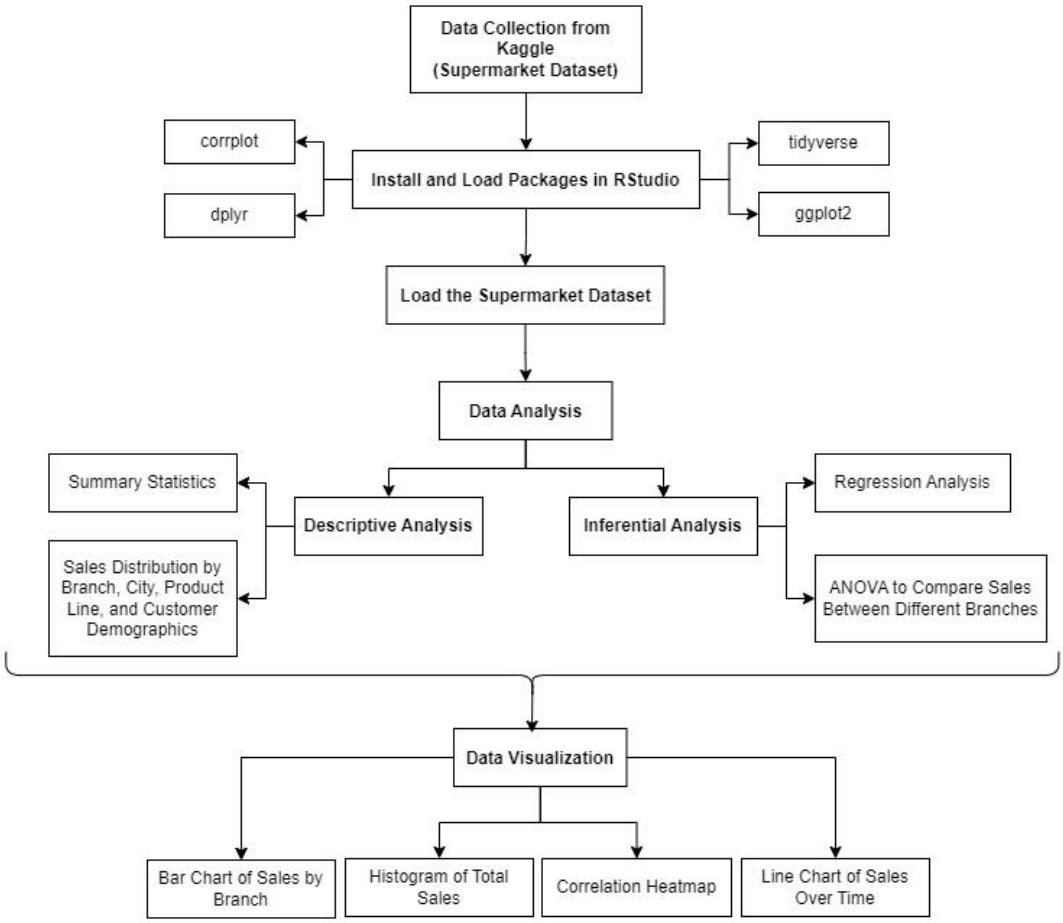

Figure 2. Data analysis on supermarket sales dataset

1. Dataset:
   Supermarket Sales Dataset by Aung Pyae from Kaggle.

2. Tools:
   RStudio for data analysis and visualizations.

3. Data Description:
   The number of supermarkets in most major cities is growing, and market competition is strong. The dataset represents one of the grocery company's historical sales, which were recorded in three distinct branches during three months.

   It consists of 1000 instances and 17 attributes. Except for invoice ID, branch, city, customer type, product line, and payment, which are numerical data types, date and time are interval data types, and gender is binary, the rest of the attributes are numerical data types. Table 1 has descriptions for each attribute.

Table 1: Description of the dataset

Attributes	Description	Data Type
Invoice id	Computer-generated sales slip invoice identification number	Nominal
Branch	Branch of supercenter (3 branches are available identified by A, B and C)	Nominal
City	Location of supercenters	Nominal

Customer type	Type of customers, recorded by Members for customers using member cards and Normal for those without member card	Nominal
Gender	Gender type of customer	Binary
Product line	General item categorization groups - Electronic accessories, Fashion accessories, Food and beverages, Health and beauty, Home and lifestyle, Sports and travel	Nominal
Unit price	The price of each product is $	Numerical
Quantity	Number of products purchased by customer	Numerical
Tax	5% tax fee for customer buying	Numerical
Total	Total price including tax	Numerical
Date	Date of purchase (Record available from January 2019 to March 2019)	Interval
Time	Purchase time (10 am to 9 pm)	Interval
Payment	Payment used by the customer for purchase (3 methods are available – Cash, Credit card and Ewallet)	Nominal
COGS	Cost of goods sold	Numerical
Gross margin percentage	Gross margin percentage	Numerical
Gross income	Gross income	Numerical
Rating	Customer stratification rating on their overall shopping experience (On a scale of 1 to 10)	Numerical

4. Data Analysis:
   - Descriptive analysis includes summary data and box plots for sales distribution.
   - Inferential analysis includes ANOVA and regression.
   - Visualisations include histograms, bar charts, line graphs, and correlation heatmap.

Data analysis

1. Descriptive Analysis:

✧ Summary statistics reveal the central tendency and distribution of total sales and quantity.

```
# Summary statistics for sales and quantities
summary(supermarket_sales$Total)
summary(supermarket_sales$Quantity)
```

Code Snippet 1

Table 2: Summary statistics for sales

Min	1st Qu	Median	Mean	3rd Qu	Max
10.68	124.42	253.85	322.97	471.35	1042.65

Table 3: Summary statistics for quantities

Min	1st Qu	Median	Mean	3rd Qu	Max
1.00	3.00	5.00	5.51	8.00	10.00

✧ Box plots show the sales distribution by branch, city, product line, and client demographic.

```
# Sales distribution by branch
ggplot(supermarket_sales, aes(x = Branch, y = Total)) +
  geom_boxplot() +
  theme_minimal() +
  ggtitle("Sales Distribution by Branch")

# Sales distribution by city
ggplot(supermarket_sales, aes(x = City, y = Total)) +
  geom_boxplot() +
  theme_minimal() +
  ggtitle("Sales Distribution by City")

# Sales distribution by product line
ggplot(supermarket_sales, aes(x = Product.line, y = Total)) +
  geom_boxplot() +
  theme_minimal() +
  ggtitle("Sales Distribution by Product Line")

# Sales distribution by customer demographics
ggplot(supermarket_sales, aes(x = Gender, y = Total)) +
  geom_boxplot() +
  theme_minimal() +
  ggtitle("Sales Distribution by Gender")

# Sales distribution by customer type
ggplot(supermarket_sales, aes(x = Customer.type, y = Total)) +
  geom_boxplot() +
  theme_minimal() +
  ggtitle("Sales Distribution by Customer Type")
```

Code Snippet 2

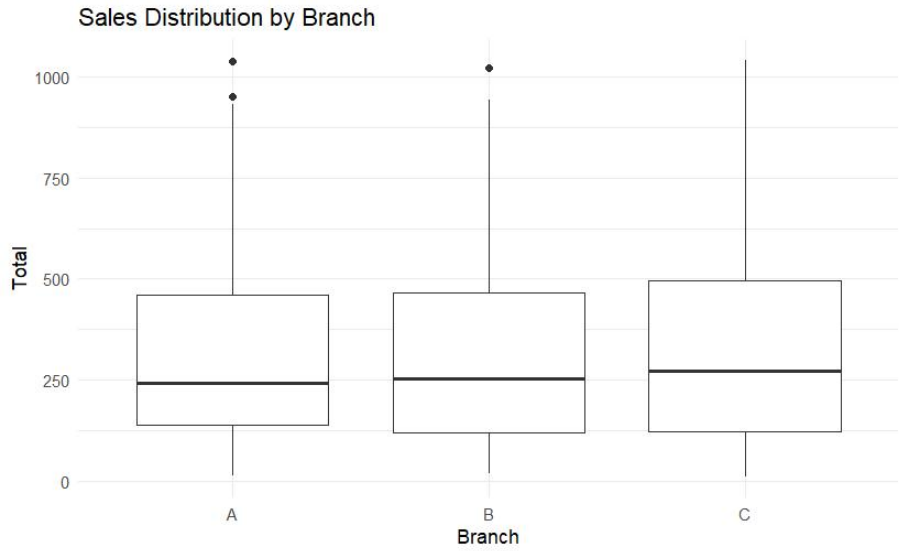

Figure 2. Sales Distribution by Branch

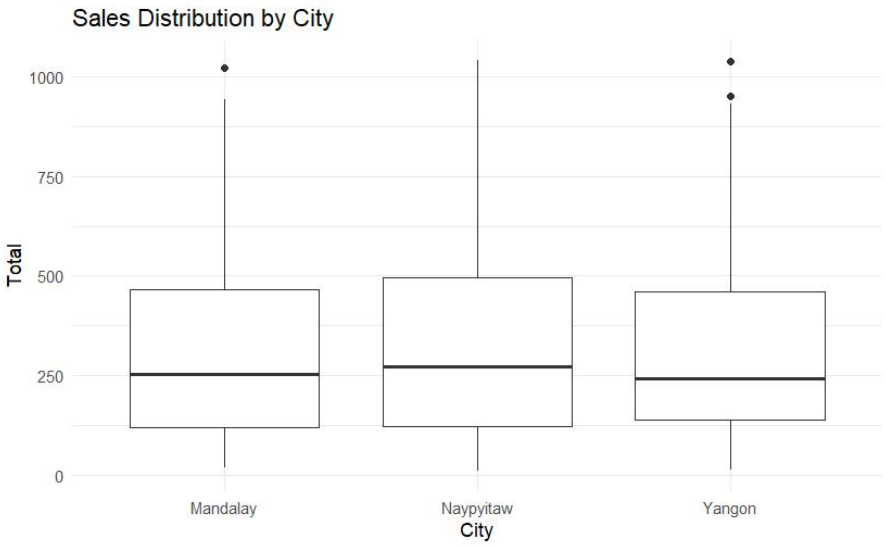

Figure 3. Sales Distribution by City

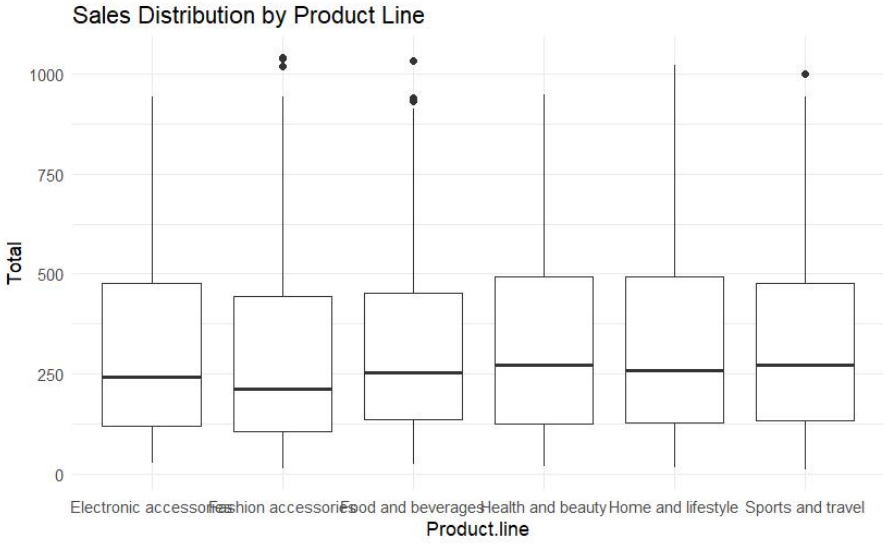

Figure 4. Sales Distribution by Product Line

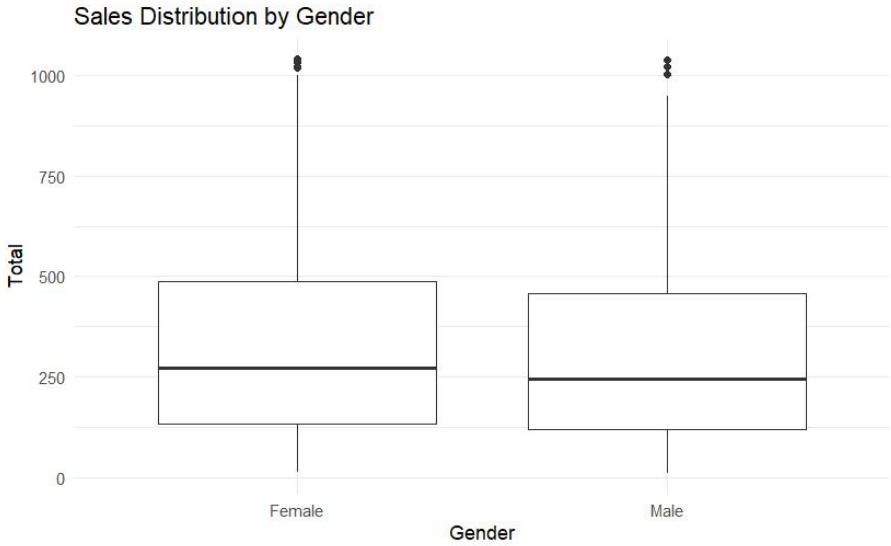

Figure 5. Sales Distribution by Gender

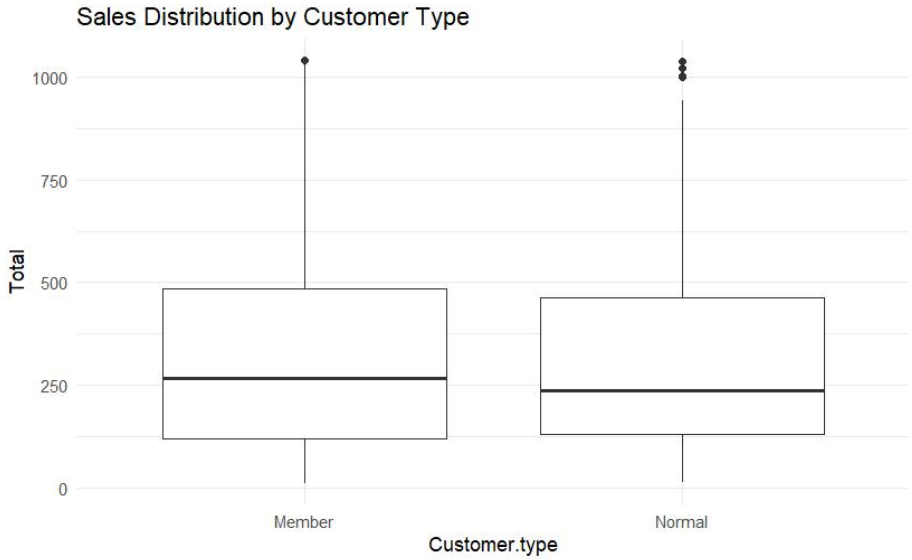

Figure 6. Sales Distribution by Customer Type

These visualisations (Figures 2, 3, 4, 5, and 6) provide an excellent overview of the sales distribution across categories. They demonstrate that there is no significant difference in sales by customer type, gender, product line, city, or branch. The sales distributions are quite constant, with a median sales of 250 and a few high-value outliers in all categories.

2. Inferential Analysis:

- ANOVA results show if there are statistically significant changes in sales among branches (Table 4) and product lines (Table 5).

```
# ANOVA to compare sales between branches
anova_result <- aov(Total ~ Branch, data = supermarket_sales)
summary(anova_result)

# ANOVA to compare sales among product lines
anova_result_product <- aov(Total ~ Product.line, data = supermarket_sales)
summary(anova_result_product)
```

Code Snippet 3

Table 4: Results of ANOVA to compare sales between branches

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Branch	2	106988	53494	0.885	0.413
Residuals	997	60292151	60474		

Table 5: Results of ANOVA to compare sales among product lines

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Branch	5	102506	20501	0.338	0.89
Residuals	994	60296633	60661		

- Regression analysis examines the link between sales and a variety of predictors such as quantity, unit price, tax, COGS, and gross income(Table 6, Figures 7 and 8).

```r
# Simple Linear Regression to explore the relationship between sales and other
variables
regression_result <- lm(Total ~ Quantity + Unit.price + Tax.5. + cogs + gross
.income, data = supermarket_sales)

# Summary of the regression model
summary(regression_result)

# Residuals Plot
ggplot(data = supermarket_sales, aes(x = predict(regression_result), y = residuals
(regression_result))) +
  geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red") +
  theme_minimal() +
  ggtitle("Residuals vs Fitted Plot") +
  xlab("Fitted Values") +
  ylab("Residuals")

# Predicted vs Actual Plot
ggplot(data = supermarket_sales, aes(x = predict(regression_result), y = Total)) +
  geom_point(alpha = 0.5) +
  geom_abline(slope = 1, intercept = 0, color = "red") +
  theme_minimal() +
  ggtitle("Predicted vs Actual Plot") +
  xlab("Predicted Values") +
  ylab("Actual Values")
```

Code Snippet 4

- ✓ **Call:**

$$\text{lm(formula} = \text{Total} \sim \text{Quantity} + \text{Unit.price} + \text{Tax.5.} + \text{cogs} + \text{gross.income, data} = \text{supermarket\_sales)}$$

- ✓ **Residuals:**

Min = -2.610e-13
1Q = -5.880e-14
Median = -2.860e-14
3Q = -7.500e-15
Max = 2.904e-11

- ✓ **Coefficients:** (2 not defined because of singularities)

Table 6: Summary of Regression Results

Variable	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.499e-12	1.478e-13	1.014e+01	< 2e-16
Quantity	-3.863e-14	2.372e-14	-1.629e+00	0.10373
Unit price	-7.603e-15	2.398e-15	-3.170e+00	0.00157
Tax 5%	2.100e+01	7.658e-15	2.742e+15	< 2e-16
COGS	NA	NA	NA	NA
Gross income	NA	NA	NA	NA

- ✓ **Residual Standard Error:** 9.394e-13 on 996 degrees of freedom
- ✓ **Multiple R-squared:** 1.0, **Adjusted R-squared:** 1.0
- ✓ **F-statistic:** 2.281e+31 on 3 and 996 DF, **p-value:** < 2.2e-16

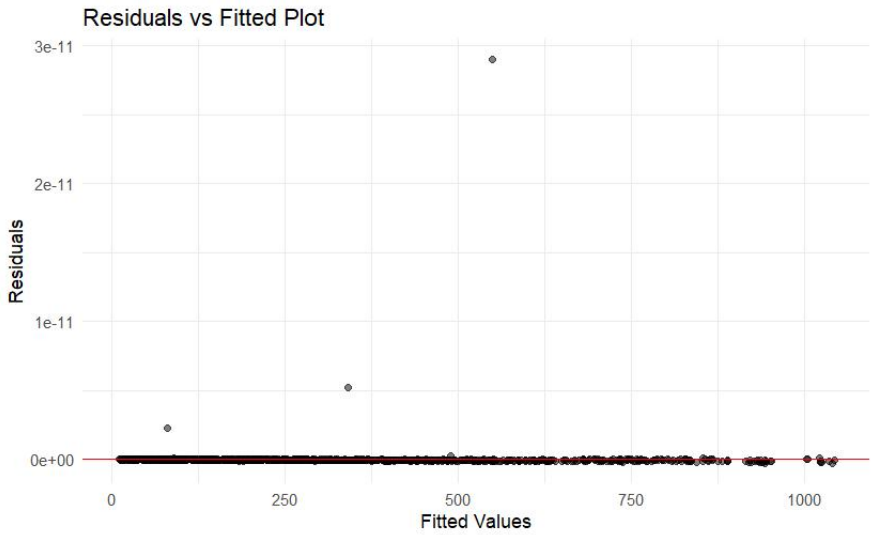

Figure 7. Residuals vs Fitted Plot

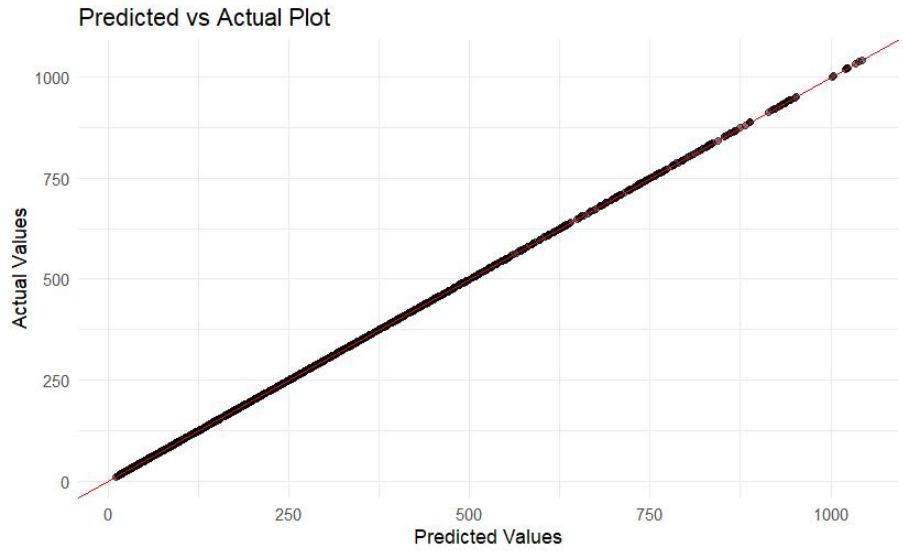

Figure 8. Predicted vs Actual Plot

The results show that unit price and tax 5% are significant predictors of total sales (p-values < 0.05). However, Quantity, COGS, and Gross revenue make no significant improvements to the model, as evidenced by their high p-values or lack of definition due to singularities. The model has an extremely high R-squared value, indicating that it explains nearly all of the variance in the Total sales data.

3. Visualizations:

✓ Histograms show the distribution of total sales.

```
# Histogram of Total Sales
ggplot(supermarket_sales, aes(x = Total)) +
  geom_histogram(binwidth = 10, fill = "blue", color = "black", alpha = 0.7) +
  theme_minimal() +
  ggtitle("Histogram of Total Sales") +
  xlab("Total Sales") +
  ylab("Frequency")
```

Code Snippet 5

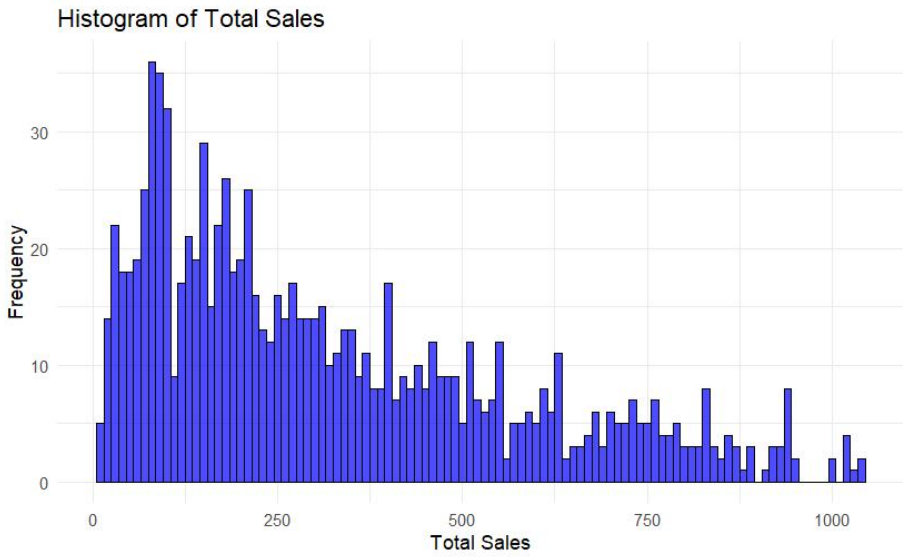

Figure 9. Histogram of Total Sales

✓ Bar charts highlight the total sales by branch.

```
# Bar chart of Sales by Branch
ggplot(supermarket_sales, aes(x = Branch, y = Total, fill = Branch)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  ggtitle("Total Sales by Branch") +
  xlab("Branch") +
  ylab("Total Sales")
```

Code Snippet 6

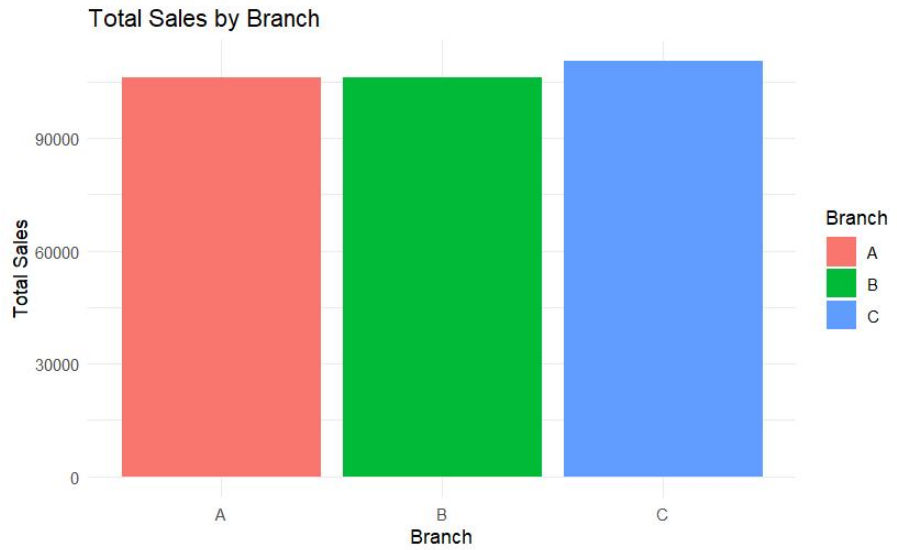

Figure 10. Total Sales by Branch

✓ Line charts track the sales trends over time.

```r
# Convert Date column to Date type if not already in that format
supermarket_sales$Date <- as.Date(supermarket_sales$Date, format = "%m/%d/%Y")

# Line chart of Sales Over Time
ggplot(supermarket_sales, aes(x = Date, y = Total)) +
  geom_line(color = "blue") +
  theme_minimal() +
  ggtitle("Sales Over Time") +
  xlab("Date") +
  ylab("Total Sales")
```

Code Snippet 7

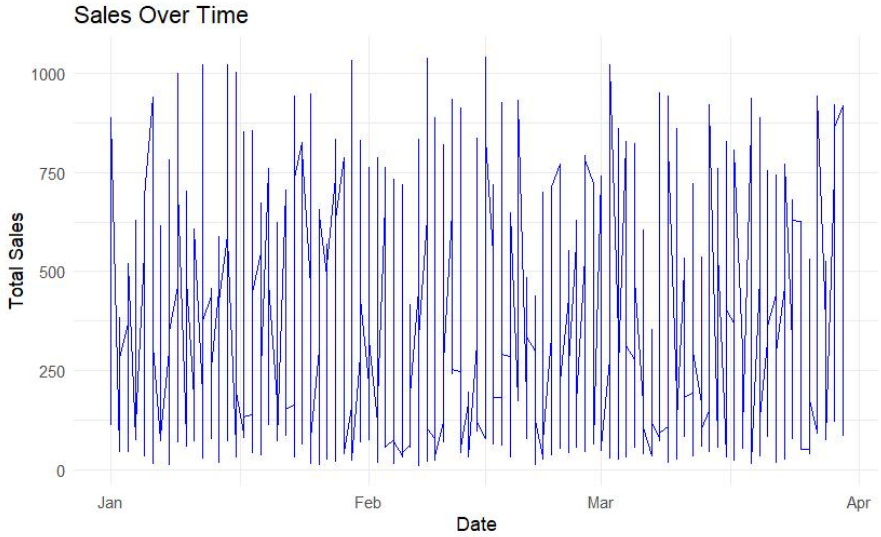

Figure 11. Sales Over Time

✓ Correlation heatmaps reveal the relationships between numeric variables.

```r
# Select numeric columns for correlation matrix
numeric_columns <- supermarket_sales %>% select(where(is.numeric))

# Remove columns with zero standard deviation
numeric_columns <- numeric_columns %>% select_if(~ sd(.) != 0)

# Calculate the correlation matrix
correlation_matrix <- cor(numeric_columns, use = "complete.obs")

# Plot the correlation heatmap
corrplot(correlation_matrix, method = "color", type = "upper",
         tl.col = "black", tl.srt = 45,
         addCoef.col = "black", number.cex = 0.7,
         col = colorRampPalette(c("red", "white", "blue"))(200))
```

Code Snippet 8

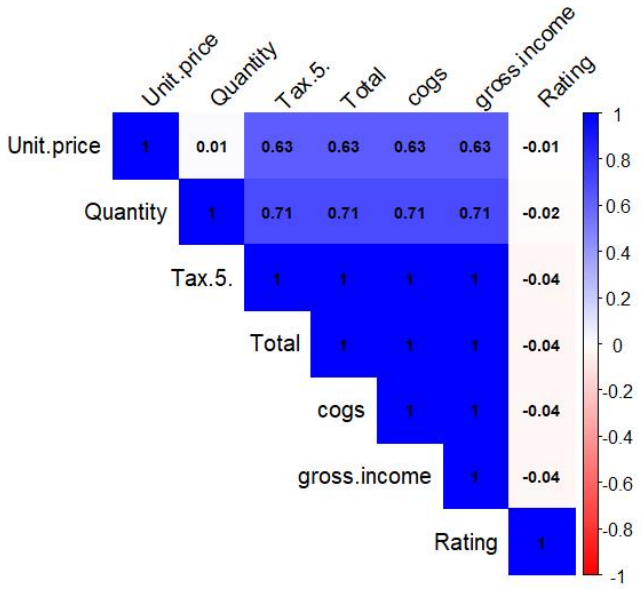

Figure 12. Correlation Heatmap

Figures 9, 10, 11, and 12 provide a complete overview of the data, showing sales distribution, branch performance, trends over time, and important variable correlations.

Discussion and Conclusion

The analysis provides information about supermarket sales patterns and customer behaviour. Key findings include sales differences among branches and product lines, monthly sales trends, and the impact of client demographics on sales. These insights can help supermarkets plan their operations and marketing strategies to increase sales and satisfy consumers.

Reference

Aung Pyae. (2019). *Supermarket sales*. Kaggle.com. https://www.kaggle.com/datasets/aungpyaeap/supermarket-sales

Seda. (2021, September 27). *Chapter 3 Exploring Data | Statistics with R - Practical Sessions*. Bookdown.org. https://bookdown.org/mrenna/statbook/exploring-data.html#descriptive-statistics

Seda. (2021, September 27). *Chapter 4 Statistical Inference | Statistics with R - Practical Sessions*. Bookdown.org. https://bookdown.org/mrenna/statbook/statistical-inference.html

GeeksforGeeks. (2020, April 12). *Data Visualization in R*. GeeksforGeeks; GeeksforGeeks. https://www.geeksforgeeks.org/data-visualization-in-r/