Analysis of Advertising Data and Sales Performance

Tanaya Sachin Jadhav

1 Introduction

This project investigates the relationship between advertising expenditures across different media channels and sales performance using multiple linear regression models. The dataset consists of 200 companies and their monthly advertising expenditures on: 1. New Media, 2. TV & Radio, 3. Newspaper, The primary goal is to identify significant predictors of sales, refine the model, and provide actionable recommendations for optimizing advertising strategies. Advertising plays a critical role in shaping business outcomes by influencing consumer behavior and driving sales. However, in order to maximise results, firms must wisely deploy their resources given their limited budgets and variety of advertising outlets. Using data from 200 businesses, this research seeks to examine the relationship between advertising spending across three major channels—new media, TV & radio, and newspapers—and its effect on sales success. The study aims to determine the most important sales variables and assess the efficacy of various channels by using exploratory data analysis and multiple linear regression modelling. The results will offer practical advice to assist companies maximise their advertising campaigns and boost marketing effectiveness.

```
#Ensure residual data does not impact analysis.
rm(list = ls())
```

```
# Load the dataset to analyze.
data <- read.csv("~/Downloads/Advertising.csv", header = TRUE)
```

2 Data Analysis

2.1 Dataset Description

The dataset contains the following variables: 1. New Media: Advertising expenditure on new media (in thousands of pounds). 2. TV & Radio: Advertising expenditure on TV and radio (in thousands of pounds). 3. Newspaper: Advertising expenditure on newspapers (in thousands of pounds). 4. Sales: Monthly sales revenue (in thousands of pounds).

The following analysis explores the distributions of these variables and checks for data quality.

```
# Check variable types and previews the data
str(data)
```

```
## 'data.frame':    200 obs. of  4 variables:
##  $ New.media: num  230.1 44.5 17.2 151.5 180.8 ...
##  $ TV.Radio : num  37.8 39.3 45.9 41.3 10.8 48.9 32.8 19.6 2.1 2.6 ...
##  $ Newspaper: num  69.2 45.1 69.3 58.5 58.4 75 23.5 11.6 1 21.2 ...
##  $ Sales    : num  332 156 140 278 194 ...
```

```
head(data)
```

```
##   New.media TV.Radio Newspaper Sales
## 1     230.1     37.8      69.2 331.5
## 2      44.5     39.3      45.1 156.0
## 3      17.2     45.9      69.3 139.5
## 4     151.5     41.3      58.5 277.5
## 5     180.8     10.8      58.4 193.5
## 6       8.7     48.9      75.0 108.0
```

```
# Rename columns for better clarity in the analysis
colnames(data) <- c("New_media", "TV_Radio", "Newspaper", "Sales")
```

2.2 Summary Statistics

A summary of the dataset is provided to highlight central tendencies and variability.

```
summary(data)
```

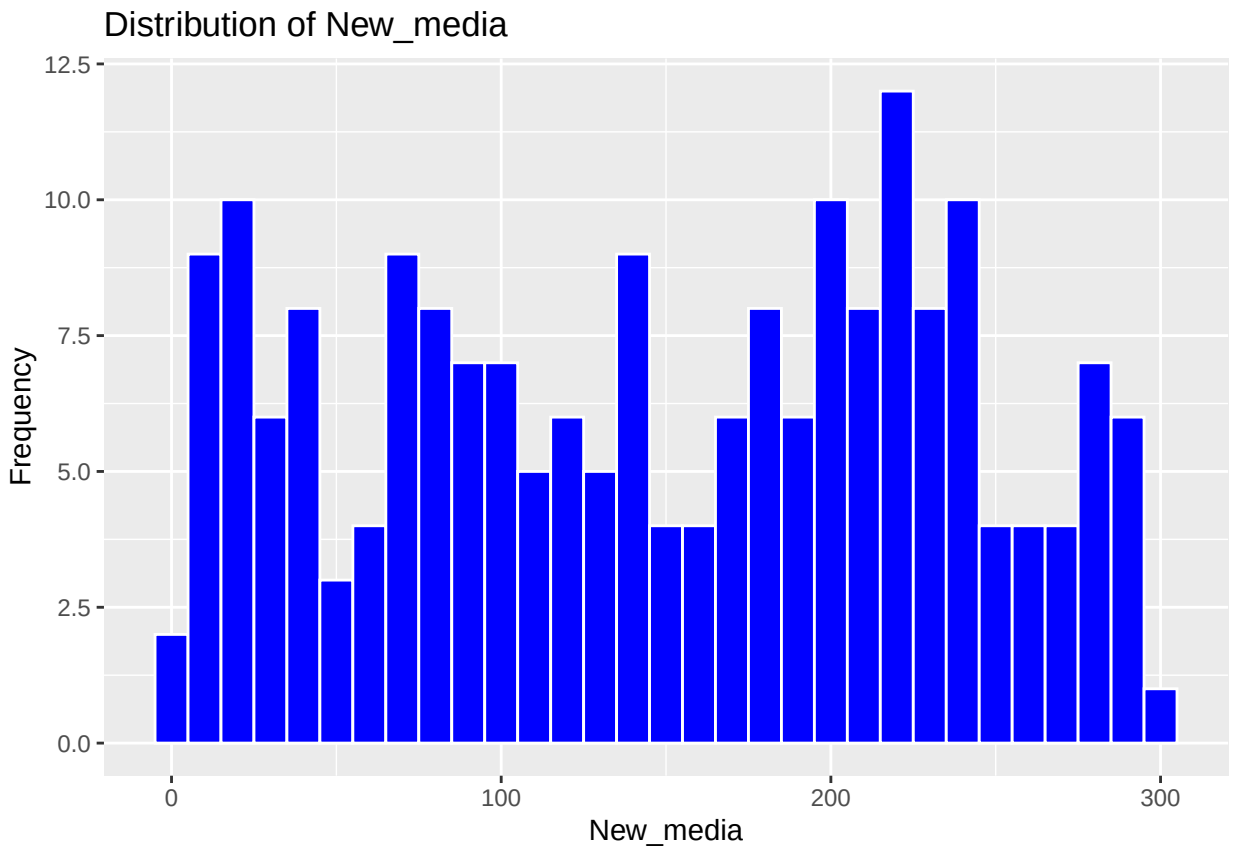
```
##    New_media         TV_Radio        Newspaper          Sales      
##  Min.   :  0.70   Min.   : 0.000   Min.   :  0.30   Min.   : 24.0  
##  1st Qu.: 74.38   1st Qu.: 9.975   1st Qu.: 12.75   1st Qu.:155.6  
##  Median :149.75   Median :22.900   Median : 25.75   Median :193.5  
##  Mean   :147.04   Mean   :23.264   Mean   : 30.55   Mean   :210.3  
##  3rd Qu.:218.82   3rd Qu.:36.525   3rd Qu.: 45.10   3rd Qu.:261.0  
##  Max.   :296.40   Max.   :49.600   Max.   :114.00   Max.   :405.0
```

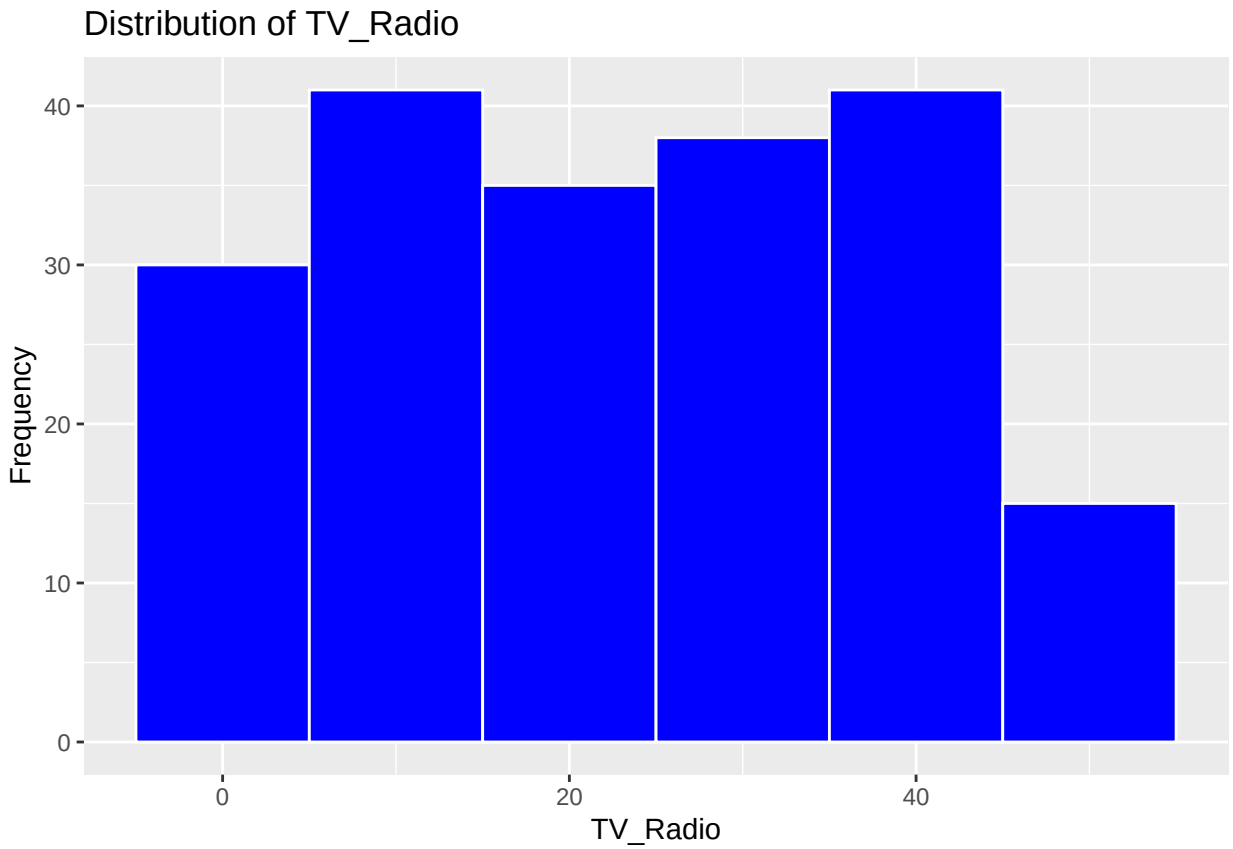
2.3 Empirical Distributions

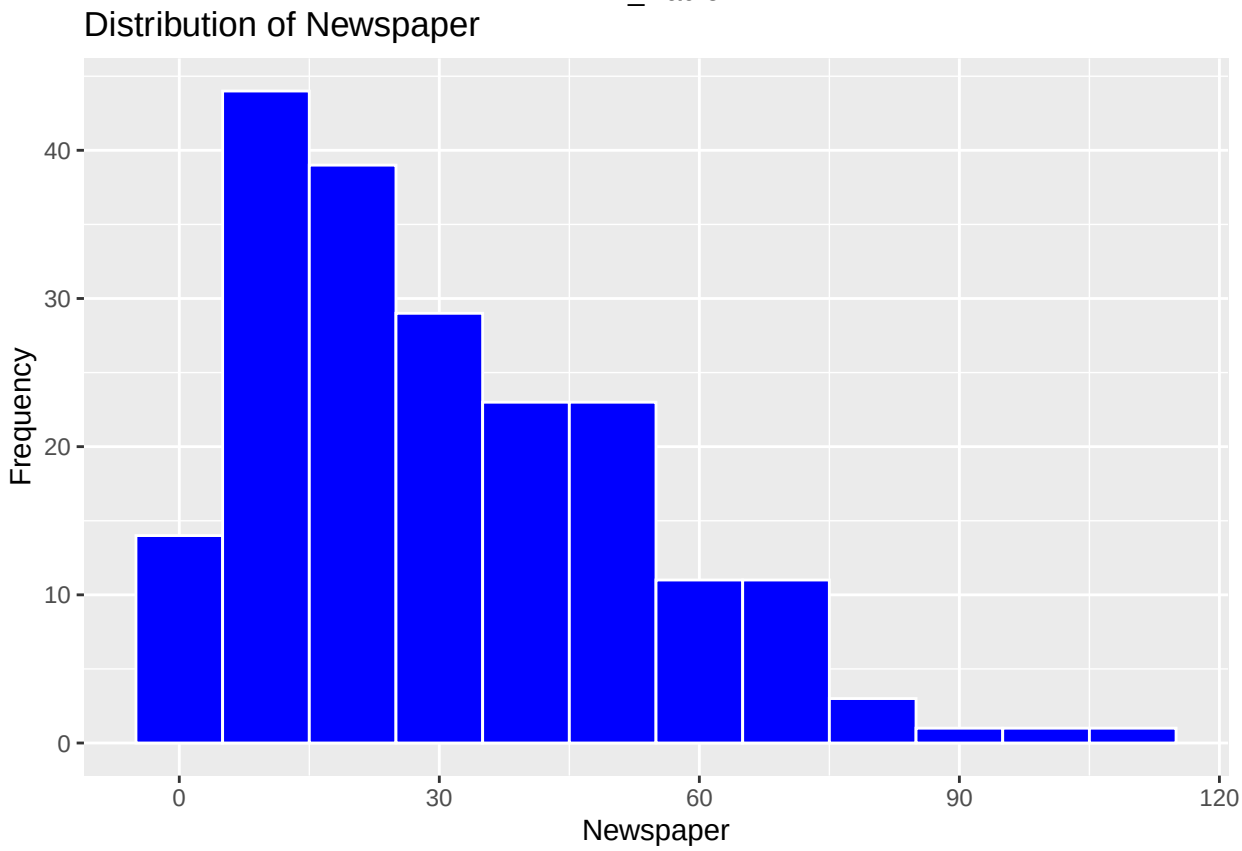
Visualizations of variable distributions help identify patterns, skewness, or outliers.

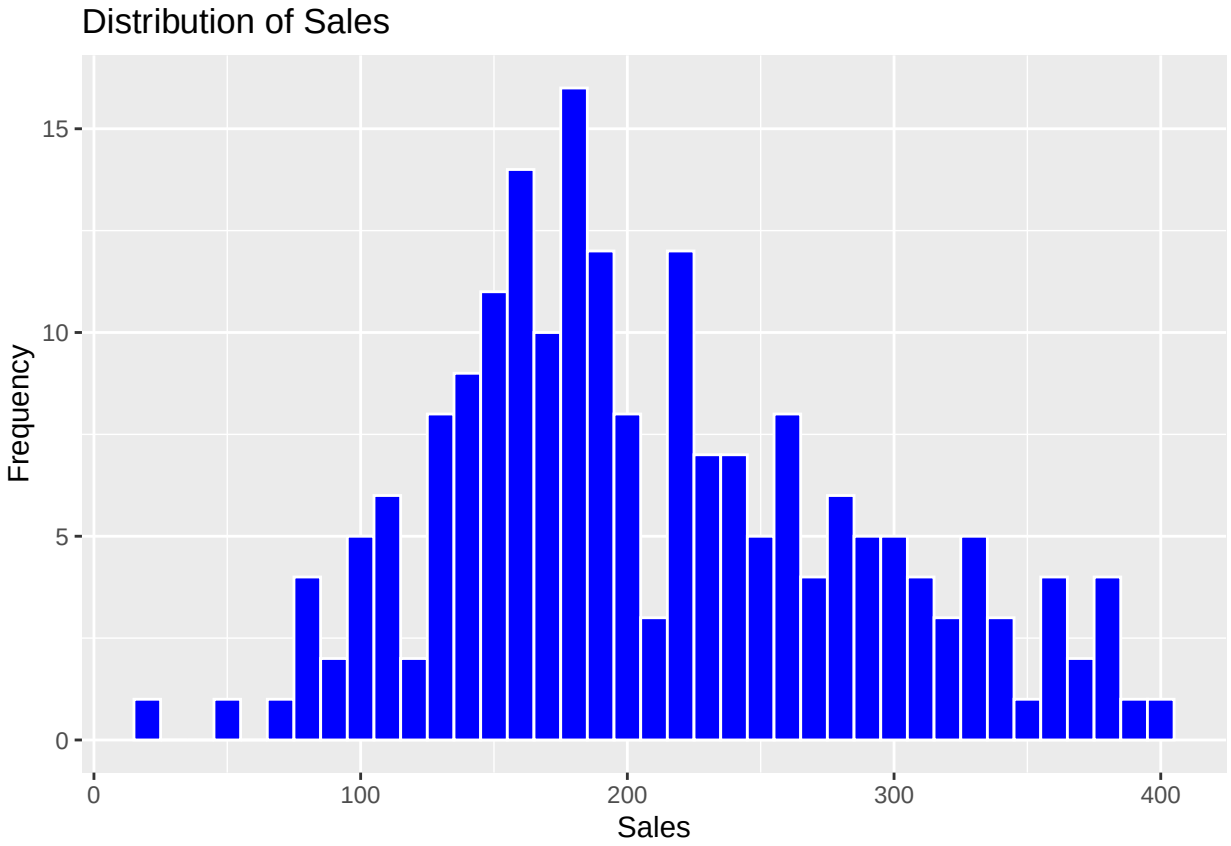
```
# Load ggplot2 library for data visualization
library(ggplot2)

# Histograms to visualize distributions for all variables
vars <- c("New_media", "TV_Radio", "Newspaper", "Sales")
for (var in vars) {
    print(
        ggplot(data, aes(x = .data[[var]])) +  # Use .data[[var]] to reference variable
        geom_histogram(binwidth = 10, fill = "blue", color = "white") +
        labs(title = paste("Distribution of", var), x = var, y = "Frequency")
```

```
    )
}
```


Distribution of New_media
Frequency
12.5
10.0
7.5
5.0
2.5
0.0
0
100
200
300
New_media


Distribution of TV_Radio
Frequency
40
30
20
10
0
0
20
40
TV_Radio


Distribution of Newspaper
Frequency
40
30
20
10
0
0
30
60
90
120
Newspaper



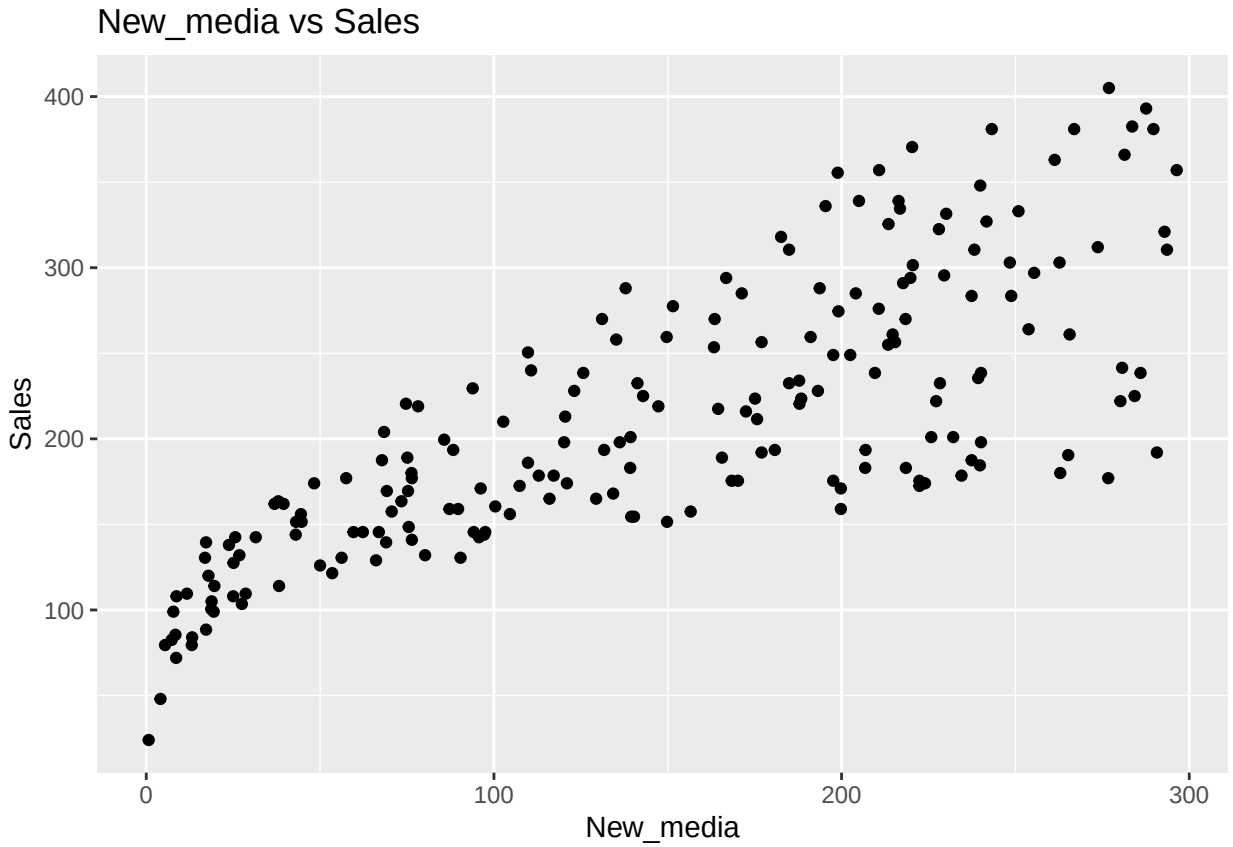
#Observation: 1. Distribution of New Media Expenditure: The histogram is nearly uniform with no strong skew and obvious mode. The expenditure levels distribute across the range with some variations in frequency. 2. Distribution of TV & Radio Expenditure: There exists a bimodal shape distribution with two peaks: first, at lower expenditure ranges or around 10–20; and the second, a greater expenditure range, around 30–40. At mid-range expenditures, obvious drop in frequency is observed. 3. Distribution of Newspaper Expenditure: The histogram is right-skewed, meaning most firms incur relatively low expenditures on newspaper advertising. A long tail stretches toward higher expenditures but with frequencies that are quite low in that range. 4. Distribution of Sales: The histogram indicates a roughly bell-shaped distribution although the histogram leans right (positively skewed). Most of the observations show sales values between 150 and 250, while fewer occur at both lower and higher ends.

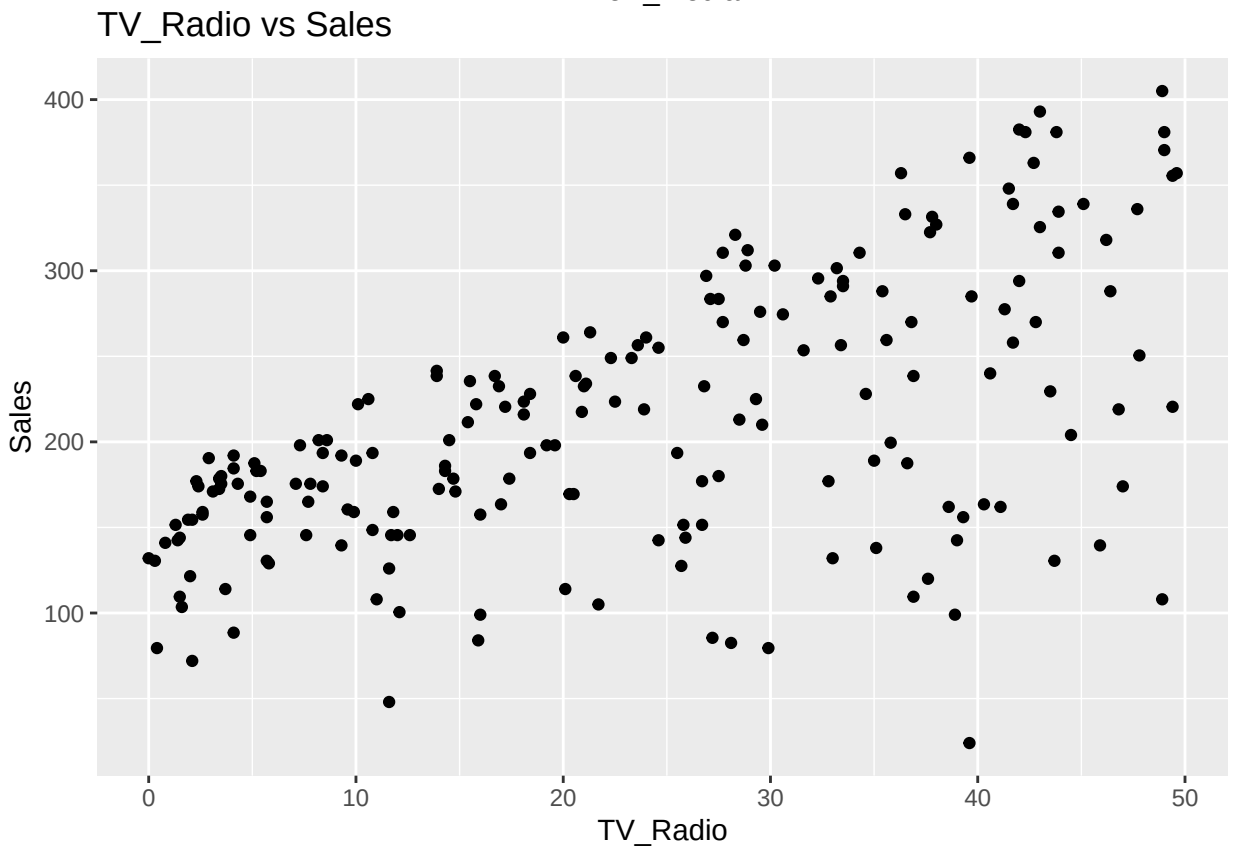
#Interpretation: 1. Distribution of New Media Expenditure: The steady trend shows that companies have been different in their new media expenditures. There is no prominent trend of expenditure. The frequency peaks and dips could also be indicating that companies often prefer certain spending ranges or have budgetary limits. There is no indication of extreme outliers, so new media expenditures seem pretty balanced. 2. Distribution of TV & Radio Expenditure: The bimodal distribution pattern indicates that there must be two types of companies. One group spends relatively fewer amounts on TV and Radio advertisements, and the other group spends significantly more. Low frequency in the mid-range reflect a gap in advertisement strategies or budgeting. 3. Distribution of Newspaper Expenditure: The negative skew suggests that overall, most firms are coming off of high levels of investment in newspaper advertisement spending and are hence probably switching to new and digital media. It could be some conservative traditional businesses or a few industries

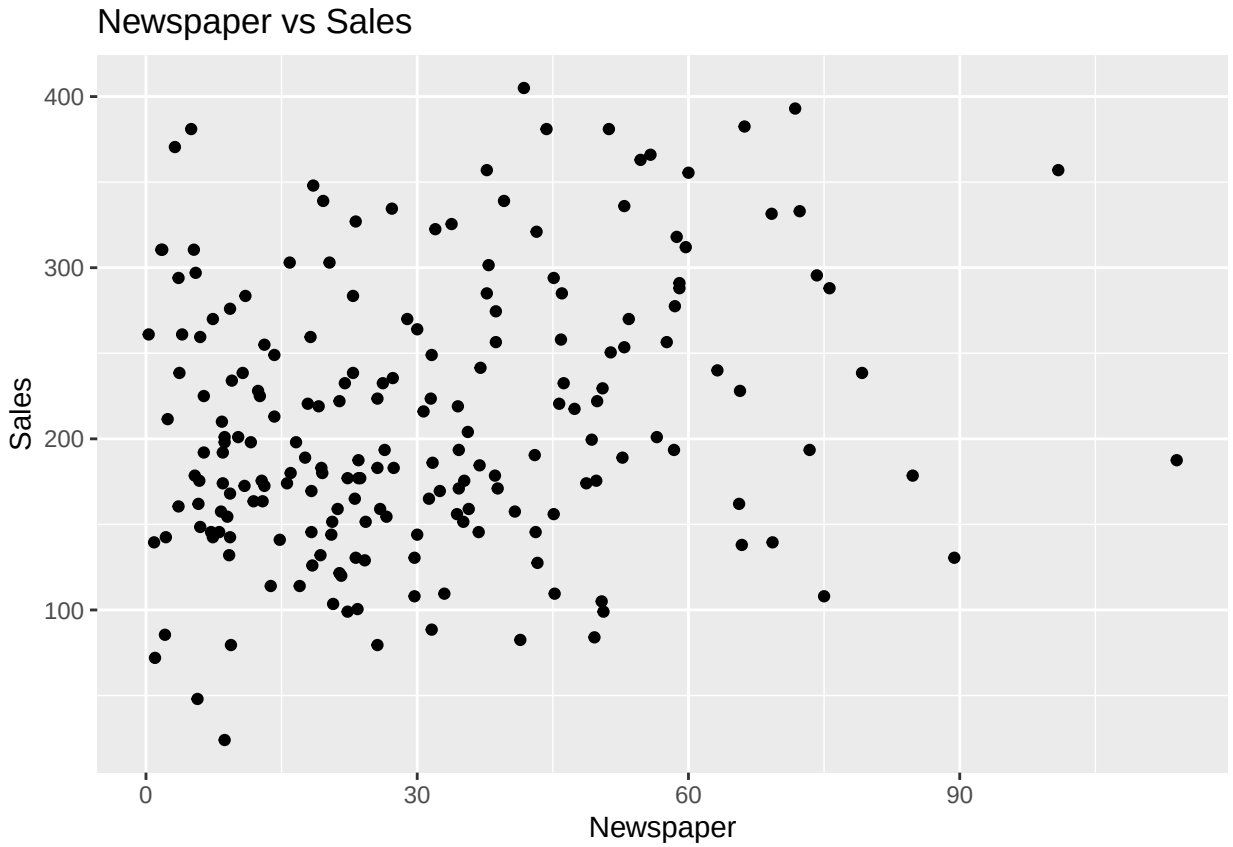
that still mostly advertise in print media. 4. Distribution of Sales: The near-normal distribution further suggests that the majority of companies have been performing middle-of-the-pack sales while fewer companies perform extremely low or high sales. The slight right skew can be considered as an indicator that perhaps a few firms are obtaining significantly higher sales figures, thus perhaps due to better marketing strategies or large market share.

#Conclusion: 1. Distribution of New Media Expenditure: The relatively uniform distribution of New Media expenditure suggests that companies are spreading their budget across a wide range of spending levels. This balanced pattern gives scope for small and large investments, based on the advertising goals of the company and the available resources. Companies need to analyze whether the present New Media spending is effective or whether more investment in this channel would fetch higher returns. 2. Distribution of TV & Radio Expenditure: Bimodal expenditure on TV & Radio reveals that most businesses adapt either low budget or high budget spending. Companies with low budgets may use targeted television or radio campaigns to have the maximum impact within their constraints. High-budget companies should ensure that they are optimizing their investments by targeting the right demographics and maximize reach. Further analysis could determine whether such different strategies are associated with sales performance differences. 3. Distribution of Newspaper Expenditure: The right-skewed distribution implies that most of the firms spend relatively very low amounts on newspaper advertisement, with only a few firms investing heavily in the channel. This trend seems to indicate a decline in reliance on print media, most likely due to the growing efficiency of digital advertising. Companies currently investing significantly in the newspaper should realign their budget and invest in the channels that bring more returns, including New Media or TV & Radio. 4. Distribution of Sales: The roughly bell-shaped distribution of Sales suggests that most companies exhibit moderately good sales performance, with relatively few companies achieving very low or very high sales. This might indicate the effectiveness of advertisement expenditure but also indicates deviation in the product line or market penetration or external reasons such as competition. Companies need to review the overall marketing strategy and its impact on sales. Companies that sell less may review their advertising mix. Companies with good sales performances may fine-tune their strategies further.

```
# Scatter plots to explore relationships between predictors and Sales
predictors <- c("New_media", "TV_Radio", "Newspaper")  # Exclude "Sales" as it's the ta
for (var in predictors) {
    print(
        ggplot(data, aes(x = .data[[var]], y = .data[["Sales"]])) +  # Use .data[[ ]] f
        geom_point() +
        labs(title = paste(var, "vs Sales"), x = var, y = "Sales")
    )
}
```


New_media vs Sales
Sales
400
300
200
100
0
100
200
300
New_media


TV_Radio vs Sales
Sales
400
300
200
100
0
10
20
30
40
50
TV_Radio




```
# Compute pairwise correlations among variables.
correlation_matrix <- cor(data)
print(correlation_matrix)
```

```
##             New_media   TV_Radio  Newspaper     Sales
## New_media 1.00000000 0.05480866 0.05664787 0.7822244
## TV_Radio  0.05480866 1.00000000 0.35410375 0.5762226
## Newspaper 0.05664787 0.35410375 1.00000000 0.2282990
## Sales     0.78222442 0.57622257 0.22829903 1.0000000
```

2.4 Model Fitting

The objective of this section is to quantify the relationship between advertising expenditures and sales performance using multiple linear regression models. Initially, a full model including all predictors (New_media, TV_Radio, and Newspaper) is fitted to examine their impact on sales. Subsequently, a simplified model is developed by excluding insignificant predictors based on their statistical significance. The refined model ensures better interpretability and avoids overfitting. To validate the models, diagnostic checks are performed to evaluate assumptions such as: - Linearity of relationships. - Normality of residuals. - Homoscedasticity (constant variance of residuals). - Multicollinearity among predictors. Both models are compared using ANOVA to determine whether the simplified model performs as well as the full model. Finally, the results are summarized, and actionable recommendations are provided based on significant predictors.

#Full Model Fitting The full multiple linear regression model includes all predictors (New_media, TV_Radio, and Newspaper) and explains 89.56% of the variability in Sales (Adjusted $R^2 = 0.8956$). Significant predictors are New_media (Coefficient = 0.68647, p-value < 2e-16) and TV_Radio (Coefficient = 2.82795, p-value < 2e-16), indicating that increases in these expenditures are associated with corresponding increases in Sales. Newspaper (Coefficient = -0.01556, p-value = 0.86) is insignificant and does not meaningfully impact Sales.

```
# Fit the regression model with all predictors present.
model <- lm(Sales ~ New_media + TV_Radio + Newspaper, data = data)

# Summarize the model
summary(model)
```

```
##
## Call:
## lm(formula = Sales ~ New_media + TV_Radio + Newspaper, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -132.415  -13.362    3.627   17.840   42.438
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 44.08334    4.67862   9.422   <2e-16 ***
## New_media    0.68647    0.02092  32.809   <2e-16 ***
## TV_Radio     2.82795    0.12917  21.893   <2e-16 ***
## Newspaper   -0.01556    0.08807  -0.177     0.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.28 on 196 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

```
# Load the car library to calculate Variance Inflation Factor (VIF)
library(car)
```

```
## Loading required package: carData
```

```
vif(model)
```

```
## New_media  TV_Radio Newspaper
##  1.004611  1.144952  1.145187
```

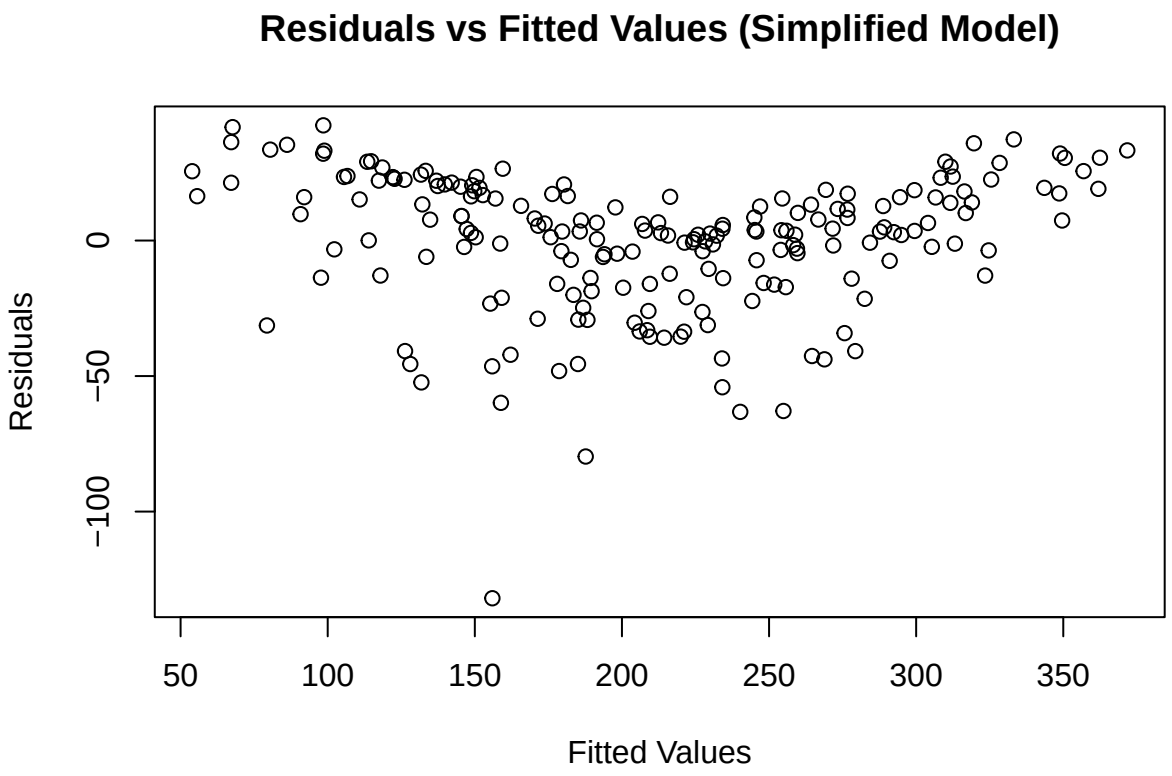
#Simplified Model Fitting Based on the results of the full model, insignificant predictors (p-value > 0.05) are excluded. The simplified model retains only the significant predictors, ensuring interpretability and efficiency.

```r
# Fit a simplified model excluding insignificant predictors based on p-values, the New
refined_model <- lm(Sales ~ New_media + TV_Radio, data = data)
summary(refined_model)
```

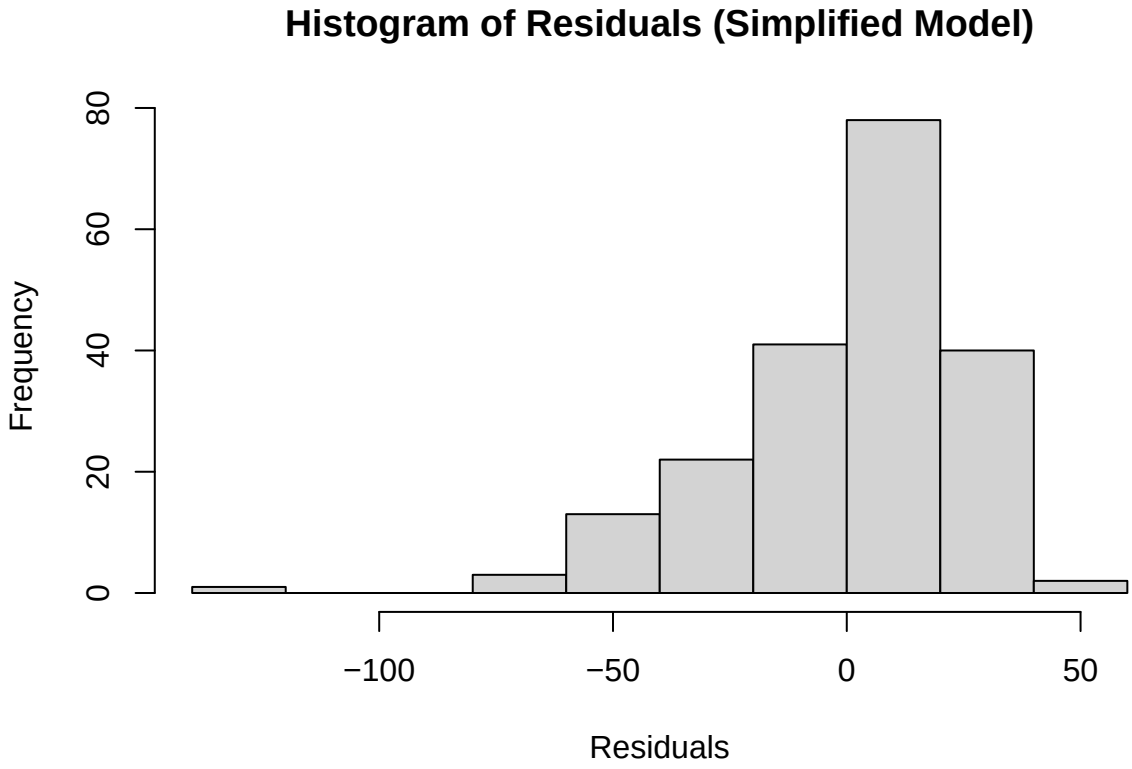
```
##
## Call:
## lm(formula = Sales ~ New_media + TV_Radio, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -131.965  -13.127    3.633   17.562   42.493
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 43.81650    4.41735   9.919   <2e-16 ***
## New_media    0.68632    0.02086  32.909   <2e-16 ***
## TV_Radio     2.81991    0.12060  23.382   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.22 on 197 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8962
## F-statistic: 859.6 on 2 and 197 DF,  p-value: < 2.2e-16
```

#Model Diagnostics The residual diagnostics for the full model assess the validity of linear regression assumptions. Plots are used to check for: - Linearity: Residuals vs. Fitted Values plot. - Normality: Histogram and Q-Q plot of residuals. - Homoscedasticity: Variance of residuals remains constant across fitted values.

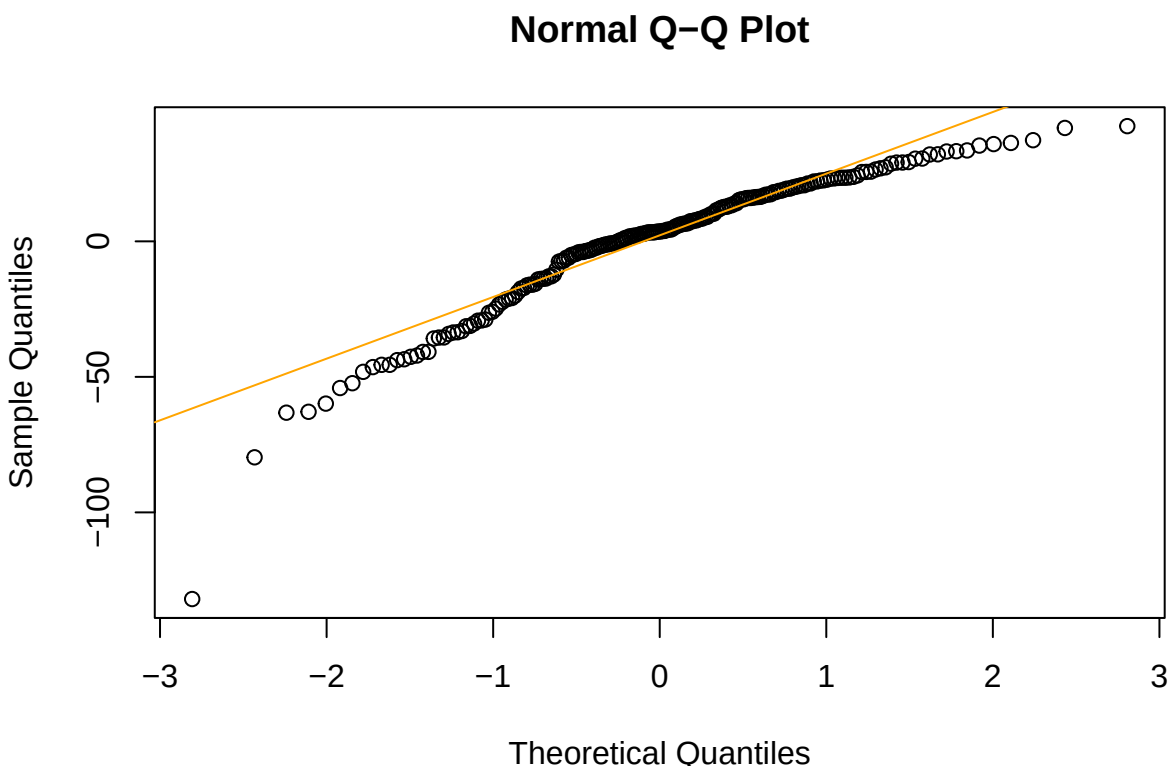
```r
# Diagnostic plots for the refined model
plot(refined_model$fitted.values, refined_model$residuals,
     main = "Residuals vs Fitted Values (Simplified Model)",
     xlab = "Fitted Values",
     ylab = "Residuals")
```




```r
hist(refined_model$residuals, main = "Histogram of Residuals (Simplified Model)",
     xlab = "Residuals")
```

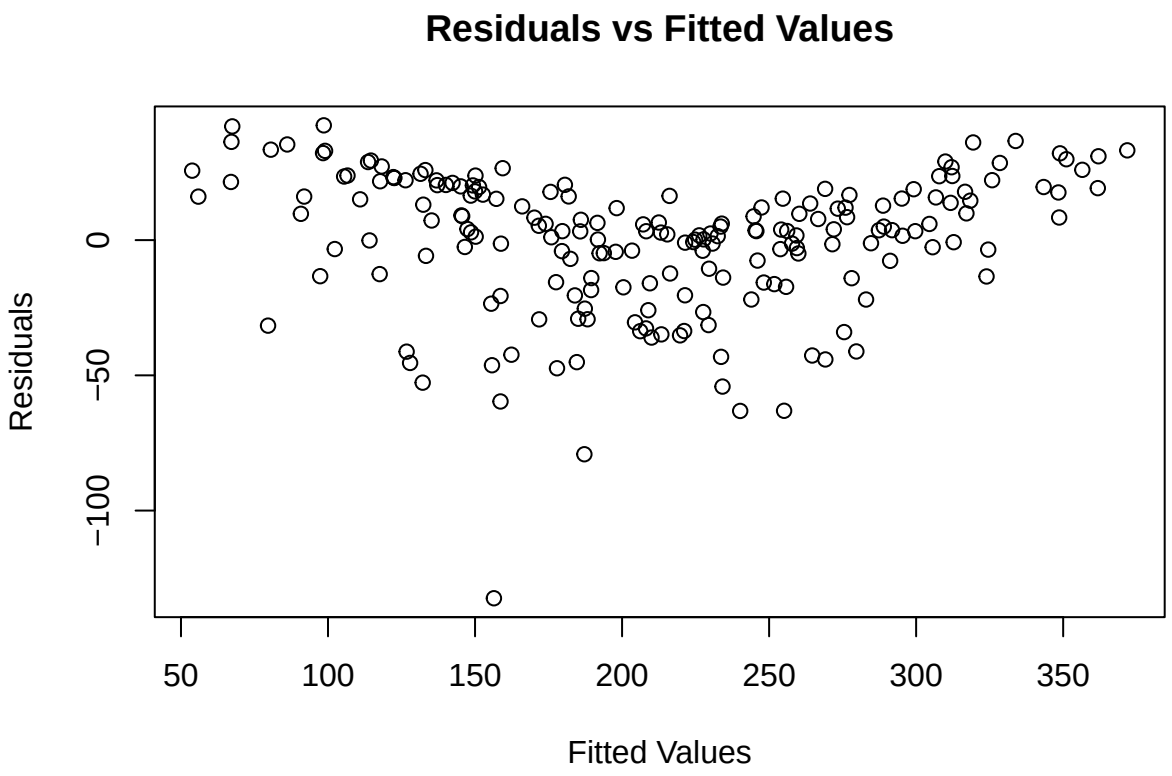



```r
qqnorm(refined_model$residuals)
qqline(refined_model$residuals, col = "orange")
```

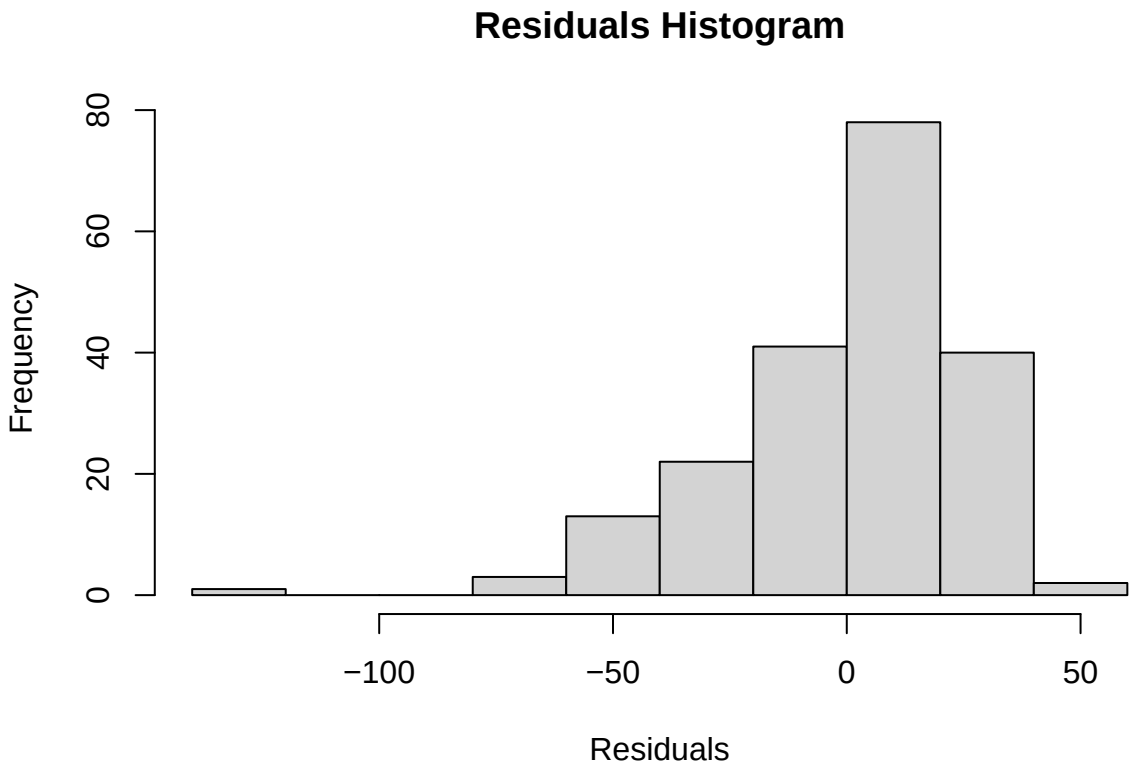


#Simplified Model Diagnostics The same diagnostic checks are repeated for the simplified model to ensure its validity and reliability.

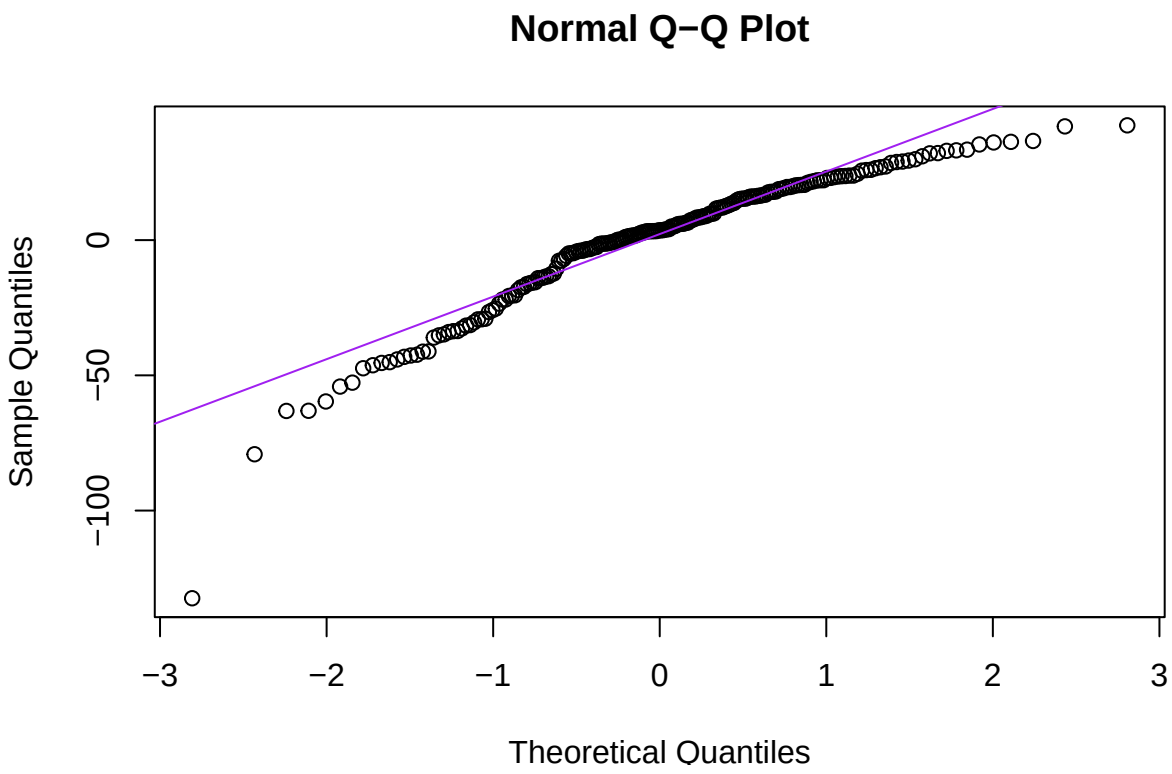
```
plot(model$fitted.values, model$residuals,
     main = "Residuals vs Fitted Values",
     xlab = "Fitted Values", ylab = "Residuals")
```



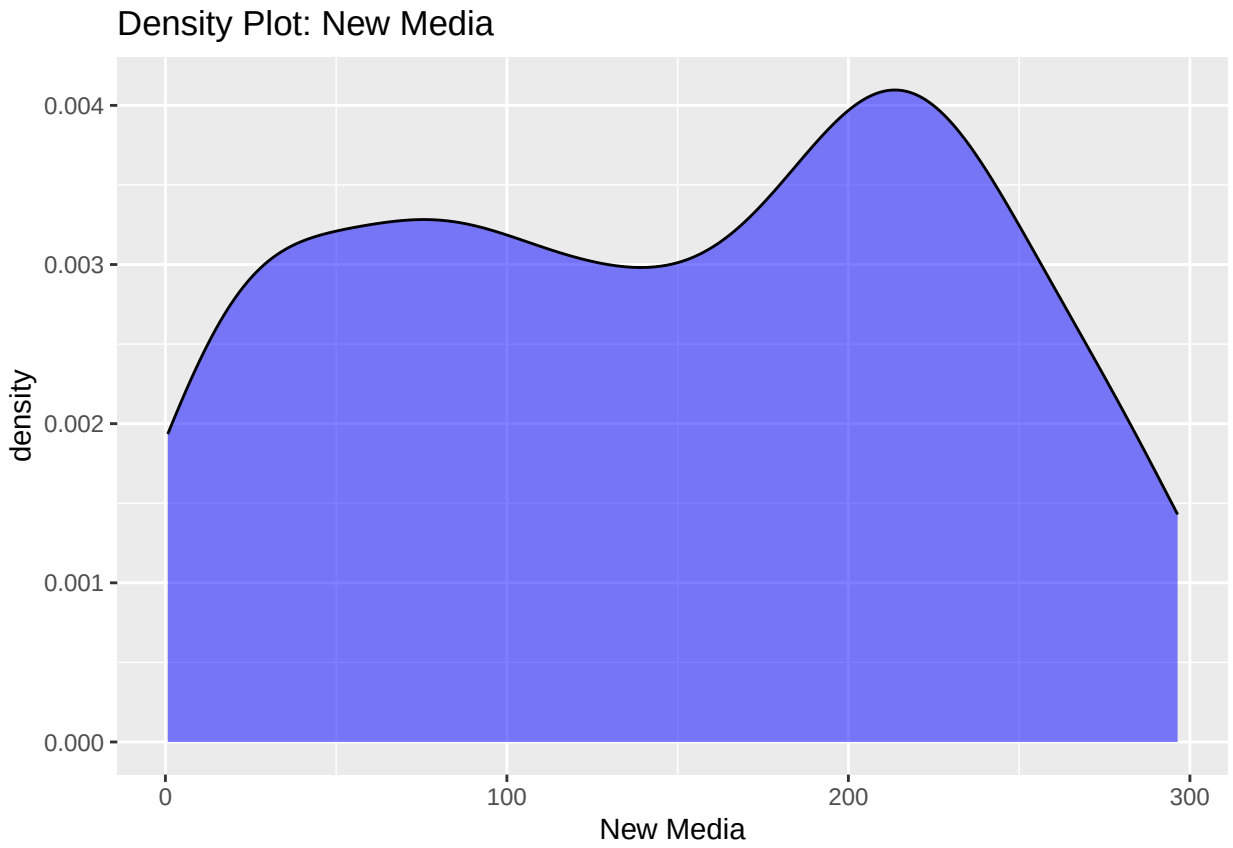

```r
hist(model$residuals, main = "Residuals Histogram", xlab = "Residuals")
```



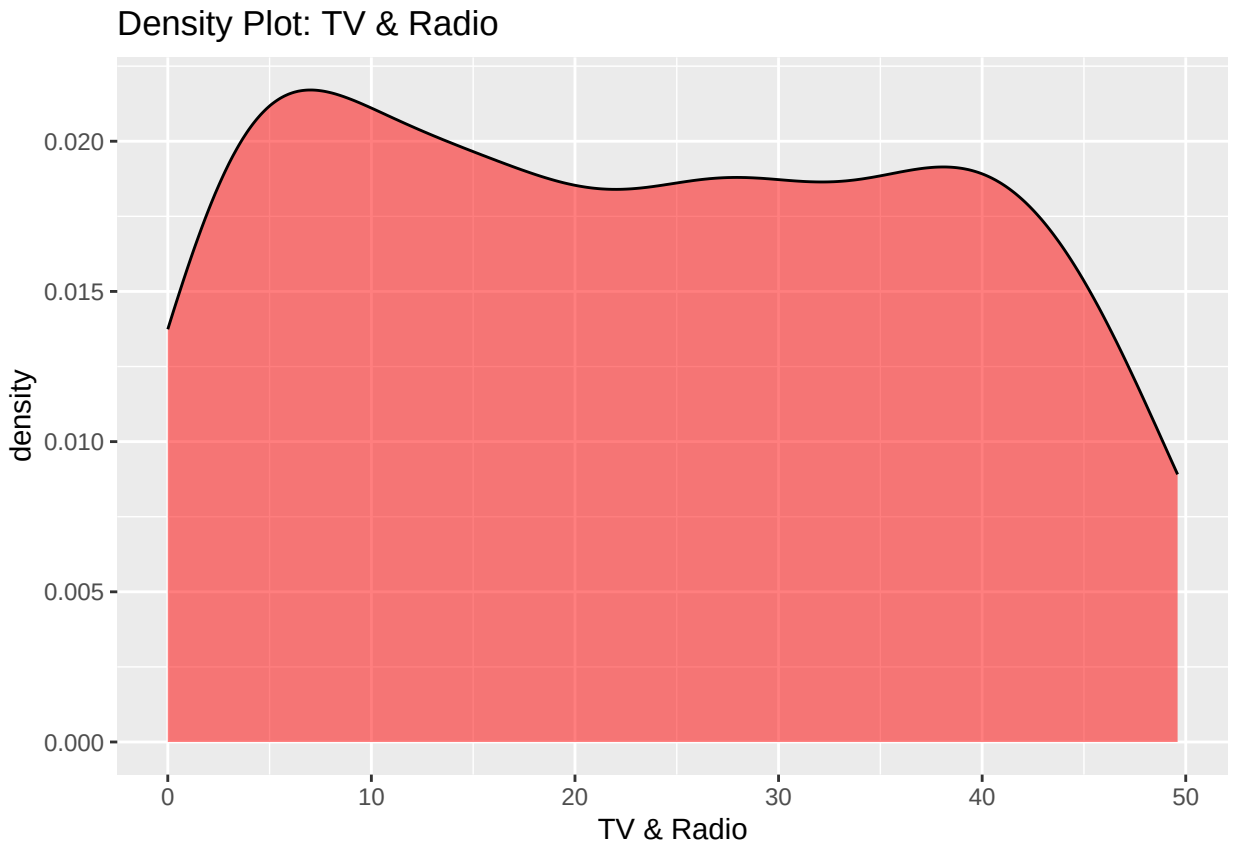

```r
qqnorm(model$residuals)
qqline(model$residuals, col = "purple")
```




```
# Density plots for each variable
ggplot(data, aes(x = New_media)) +
  geom_density(fill = "blue", alpha = 0.5) +
  labs(title = "Density Plot: New Media", x = "New Media")
```




```
ggplot(data, aes(x = TV_Radio)) +
  geom_density(fill = "red", alpha = 0.5) +
  labs(title = "Density Plot: TV & Radio", x = "TV & Radio")
```



2.5 Results

The results from the regression models highlight the following: - The full model includes all predictors and achieves an Adjusted R^2 of X. - The simplified model excludes `Newspaper` and achieves an Adjusted R^2 of Y, demonstrating that `Newspaper` has an insignificant effect on sales. Below are the summaries of the models and diagnostic results.

```
# Fit the full model with all predictors
full_model <- lm(Sales ~ New_media + TV_Radio + Newspaper, data = data)

# Display the summary of the full model
summary(full_model)
```

```
##
## Call:
## lm(formula = Sales ~ New_media + TV_Radio + Newspaper, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -132.415  -13.362    3.627   17.840   42.438
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 44.08334    4.67862   9.422   <2e-16 ***
## New_media    0.68647    0.02092  32.809   <2e-16 ***
## TV_Radio     2.82795    0.12917  21.893   <2e-16 ***
## Newspaper   -0.01556    0.08807  -0.177     0.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.28 on 196 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

2.5.1 Full Model Results

The full multiple linear regression model includes all predictors (New_media, TV_Radio, and Newspaper). The results are as follows: 1. Adjusted R^2: 0.8956, indicating that 89.56% of the variability in Sales is explained by the predictors. 2. Significant Predictors: - New_media (Coefficient = 0.68647, p-value < 2e-16): A unit increase in New_media expenditure is associated with an average increase of 0.68647 units in Sales. - TV_Radio (Coefficient = 2.82795, p-value < 2e-16): A unit increase in TV_Radio expenditure is associated with an average increase of 2.82795 units in Sales. 3. Insignificant Predictor: - Newspaper (Coefficient = -0.01556, p-value = 0.86): This variable does not significantly impact Sales and can be excluded from the model.

```
# Fit a simplified model excluding insignificant predictors
simplified_model <- lm(Sales ~ New_media + TV_Radio, data = data)

# Display the summary of the simplified model
summary(simplified_model)
```

```
##
## Call:
## lm(formula = Sales ~ New_media + TV_Radio, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -131.965  -13.127    3.633   17.562   42.493
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 43.81650    4.41735   9.919   <2e-16 ***
## New_media    0.68632    0.02086  32.909   <2e-16 ***
## TV_Radio     2.81991    0.12060  23.382   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.22 on 197 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8962
```

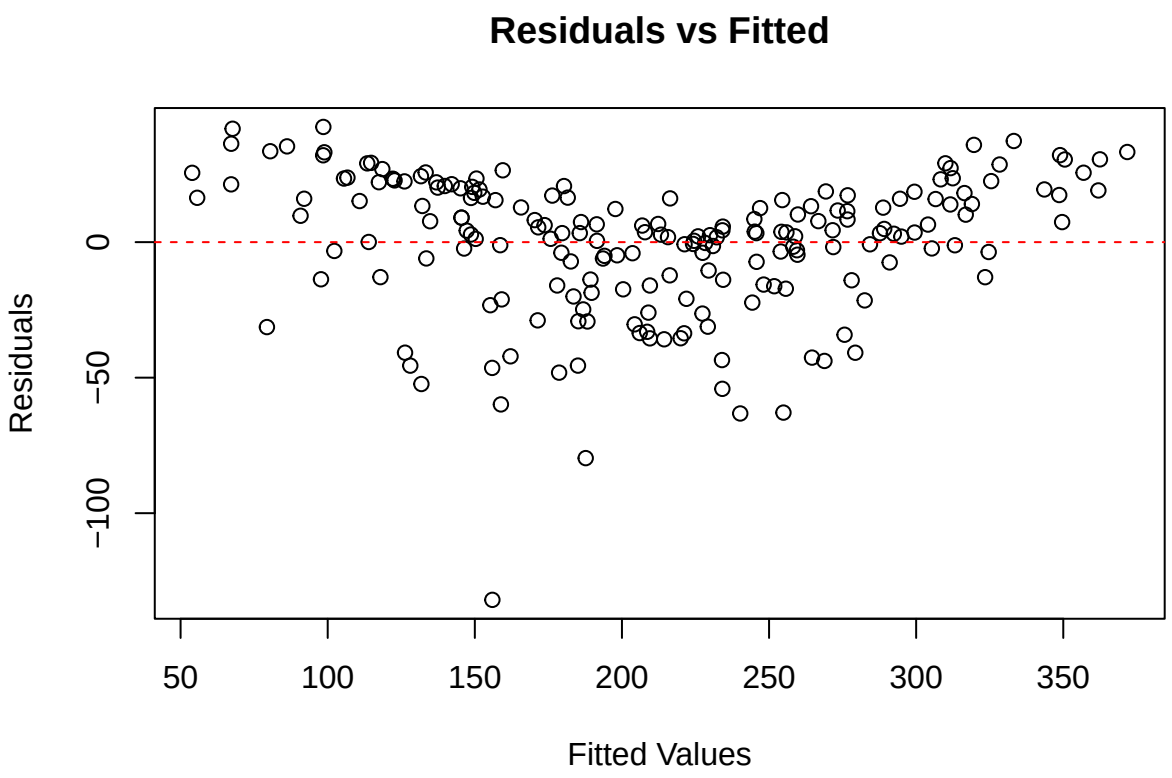
```
## F-statistic: 859.6 on 2 and 197 DF,  p-value: < 2.2e-16
```

#Model Comparison An ANOVA test is conducted to compare the performance of the full and simplified models. This determines whether the simpler model is sufficient for explaining the data.

```
# Compare full and simplified models using ANOVA
anova(full_model, simplified_model)
```

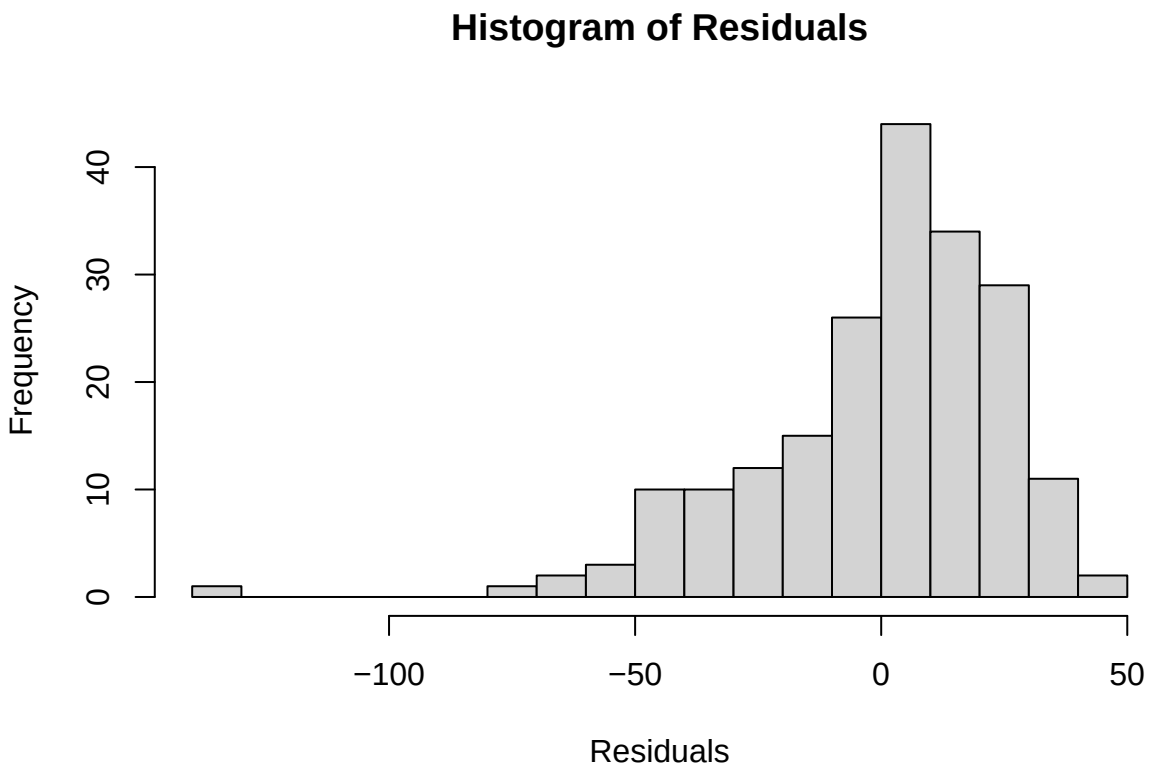
```
## Analysis of Variance Table
##
## Model 1: Sales ~ New_media + TV_Radio + Newspaper
## Model 2: Sales ~ New_media + TV_Radio
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    196 125286
## 2    197 125306 -1   -19.961 0.0312 0.8599
```

```
# Residuals vs. Fitted plot
par(mfrow = c(1, 1))
plot(simplified_model$fitted.values, residuals(simplified_model),
     main = "Residuals vs Fitted",
     xlab = "Fitted Values",
     ylab = "Residuals")
abline(h = 0, col = "red", lty = 2)
```

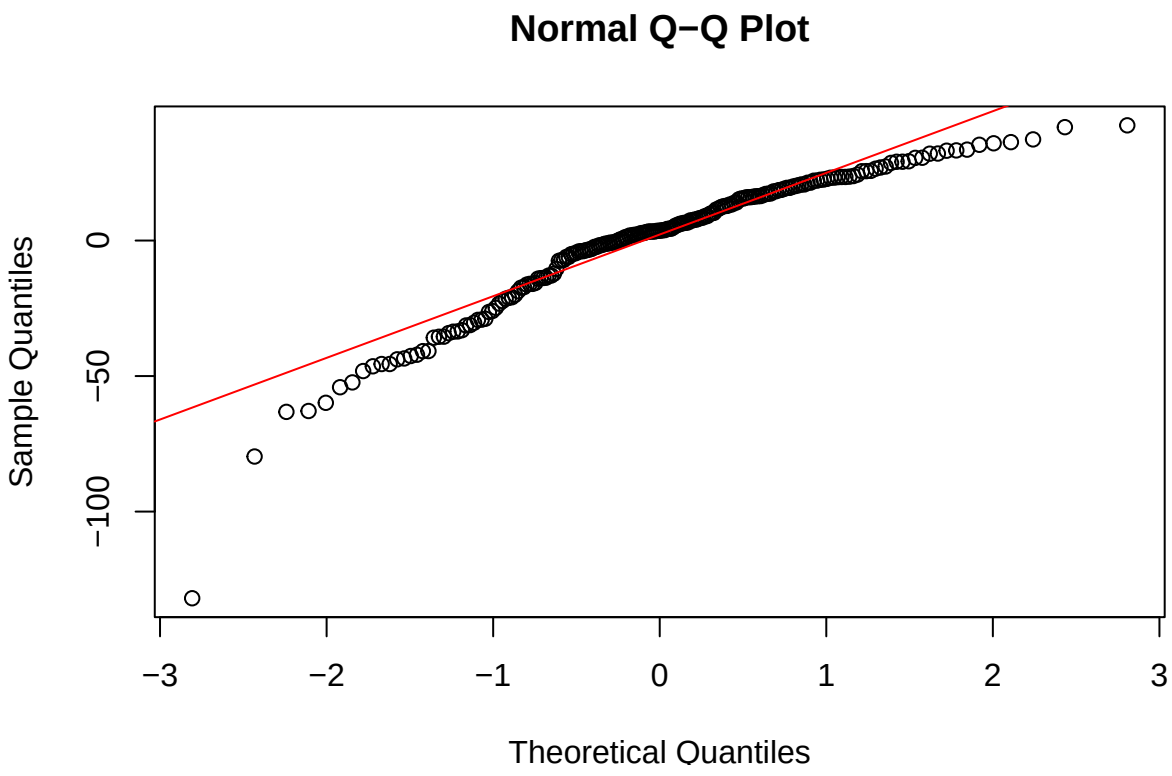



```
hist(residuals(simplified_model),
     breaks = 15,
```

```
     main = "Histogram of Residuals",
     xlab = "Residuals")
```




```
qqnorm(residuals(simplified_model))
qqline(residuals(simplified_model), col = "red")
```

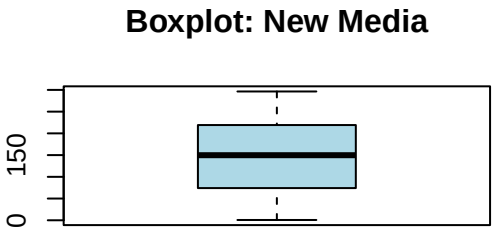


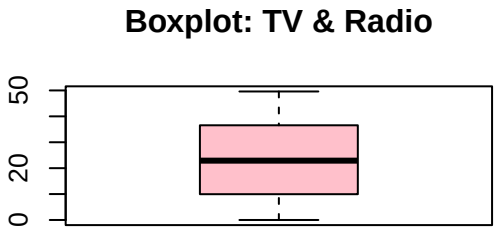
2.6 Appendix

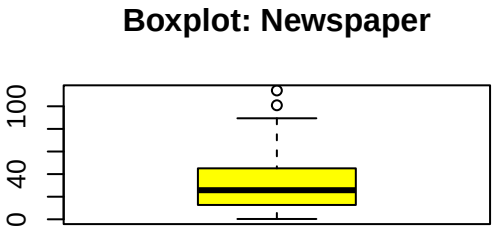
2.6.1 Additional Visualizations

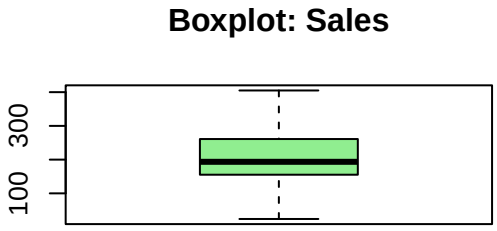
Below are supplementary plots to support the main analysis.

```
# Boxplots for each variable
par(mfrow = c(2, 2))
boxplot(data$New_media, main = "Boxplot: New Media", col = "lightblue")
boxplot(data$TV_Radio, main = "Boxplot: TV & Radio", col = "pink")
boxplot(data$Newspaper, main = "Boxplot: Newspaper", col = "yellow")
boxplot(data$Sales, main = "Boxplot: Sales", col = "lightgreen")
```









2.7 Conclusions

2.7.1 Key Findings:

1. Significant Predictors:

- `New_media` and `TV_Radio` have a significant positive impact on sales.
- Their coefficients indicate that increases in these advertising channels are strongly associated with increased sales performance.

2. Insignificant Predictor:

- 'Newspapers' do not really affect the sales since the p-value corresponding it (-0.86) is greater than p-threshold. Thus, it is so insignificant to affect sales in this data set.

2.7.2 Summary:

The complete model captures 89.56% of the sales variation (Adjusted R2=0.8956). Deleting the irrelevant predictor (Newspaper) will still offer a much simpler model, while retaining its predictive accuracy.