UNIVERSITY OF CAPE TOWN

DEPARTMENT OF STATISTICAL SCIENCES

STA5076Z - Supervised Learning Assignment 1

MSHTSA009

April 22, 2024

Contents

Predicting Restaurant Tips using Multiple Linear Regression: A Model Comparison Approach

Introduction

This report explores a dataset collected from a California restaurant in 1995, focusing on understanding the factors influencing tipping behavior. Through exploratory data analysis (EDA), we uncover insights into the dataset's structure, identify missing values, and visualize relationships between predictors and the target variable—tips. Our analysis progresses to model building, where we fit a full regression model to the training set and evaluate its performance and significance. Afterwards, we explore model improvement techniques, including variable selection methods and regularization, to enhance predictive accuracy. Additionally, we investigate the incorporation of interaction terms to understand the relationship between predictors. Finally, we conduct residual diagnostics to validate the assumptions of our regression model. Overall, this report offers a comprehensive examination of tipping behavior.

1. Exploratory Data Analysis (EDA)

There are 244 observations and 7 variables, including numerical variables such as total bill, tip, and party size, and categorical variables such as sex of the bill payer, smoker status, day of the week, and time of day.

Data Cleaning

The initial dataset included an '**X**' column, identified as an index with no analytical value, which was subsequently removed to tidy up the data structure.Afterward, we summarize the data to obtain basic statistical details such as **counts, means, standard deviations, minimum and maximum values**, among others. Subsequently, we checked for missing values in the dataset and count the number of missing values per column if any. Furthermore, we examined unique values for categorical variables such as sex, smoker, day, and time. Finally, we viewed the modified dataset to ensure the changes have been applied correctly.

Summary Statistics

Here are the key statistics from your dataset:

Statistic	Total Bill ($)	Tip ($)	Size
Minimum	3.07	1.00	1
1st Quartile	13.28	2.00	2
Median	17.82	3.00	2
Mean	19.83	3.026	2.565
3rd Quartile	24.13	3.562	3
Maximum	50.81	10.00	6

Table 1: Summary statistics for total bill, tip, and size.

- Total Bill: Ranges from $3.07 to $50.81 with a mean of approximately $19.83.
- Tip: The tips range from $1.00 to $10.00, with an average of $3.026.
- Party Size: Varies between 1 and 6, with a mean size of approximately 2.57.

The dataset contains no missing values, ensuring a robust basis for further analysis.

Categorical Variable Distribution

Variable	Categories	Count
Sex	Male	130
	Female	70
Smoker Status	Non-smoker	120
	Smoker	80
Day of the Week	Thursday	51
	Friday	14
	Saturday	74
	Sunday	61
Time of Day	Dinner	143
	Lunch	57

Table 2: Distribution of Categorical Variables

Data Visualizations

Scatter Plot Analysis

In our analysis, we focused on the relationships between total bill vs. tip.

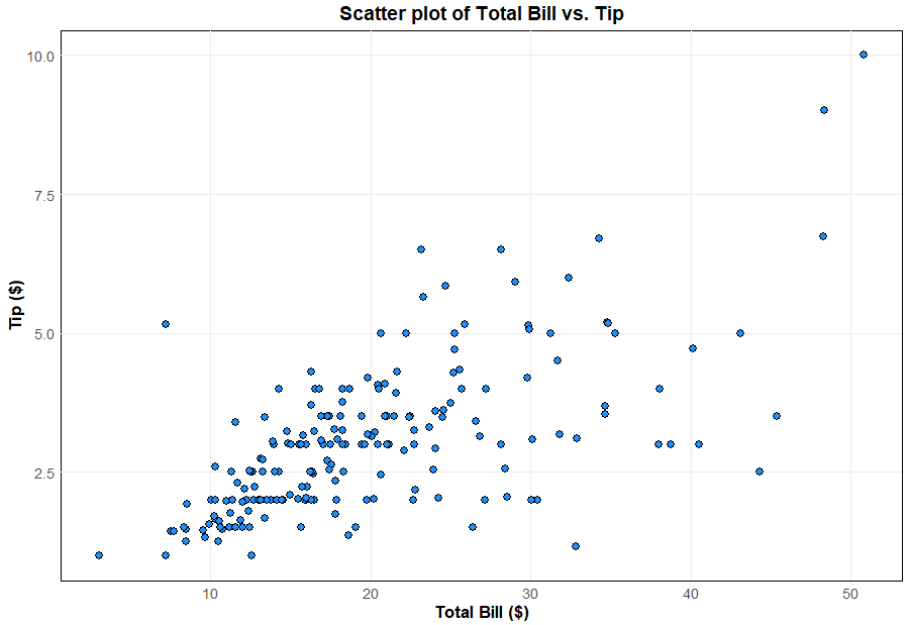

Figure 1: **Scatter plot of Total Bill vs. Tip.** There appears to be a positive correlation between the total bill and the tip amount. This suggests that as the total bill increases, customers tend to leave a higher tip.The relationship is approximately linear, indicating a straightforward proportional increase between the bill amount and tip size.There are a few outliers, particularly at higher bill amounts where tips do not proportionally match the increase in the bill. These could represent instances of either exceptionally high or surprisingly low tips relative to the bill.

Box Plot Analysis

In your analysis, you created box plots for tips by sex, smoker status, day of the week, and time of day. Here's an in-depth discussion for each:

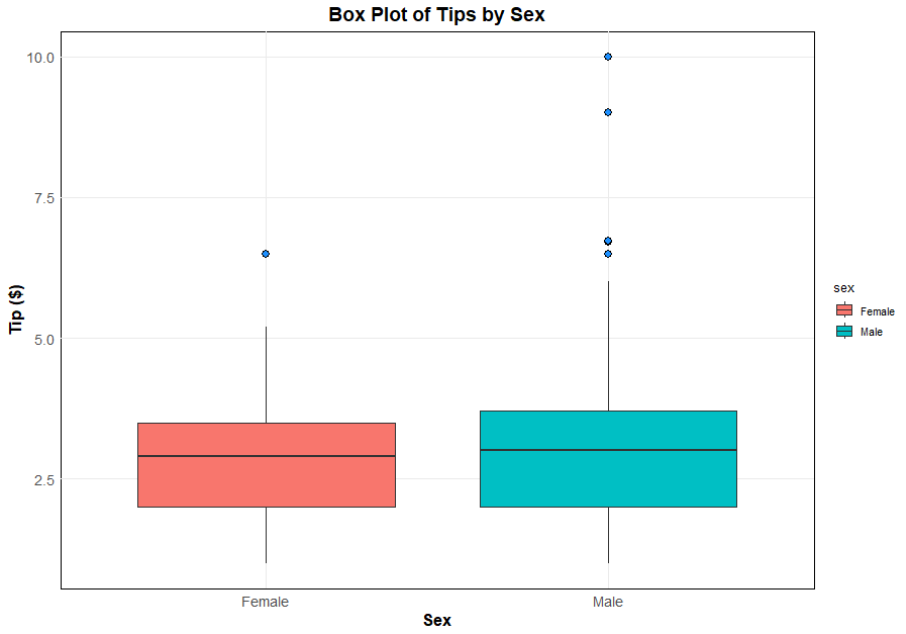

Figure 2: **Box Plot of Tips by Sex.** Both males and females exhibit a similar range of tip amounts, with the median tip slightly higher for males than females.The interquartile range (IQR), which represents the middle 50% of the tips, is slightly broader for males, indicating more variability in how much they tip compared to females. There are several outliers for both categories, indicating occasional tips that are much higher or lower than typical for both sexes.

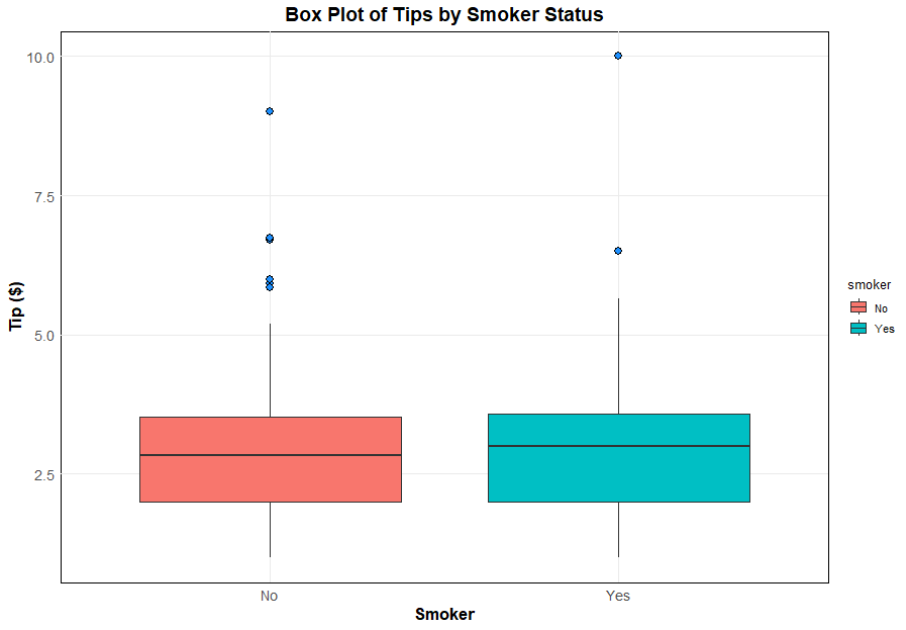

Figure 3: **Box Plot of Tips by Smoker Status.** Smokers and non-smokers show nearly identical medians, but the spread of tips from smokers is slightly wider. Both groups show a similar range, but smokers tend to have more outliers, suggesting less consistency in their tipping habits. More frequent and higher outliers in smokers indicate instances of unusually high tips

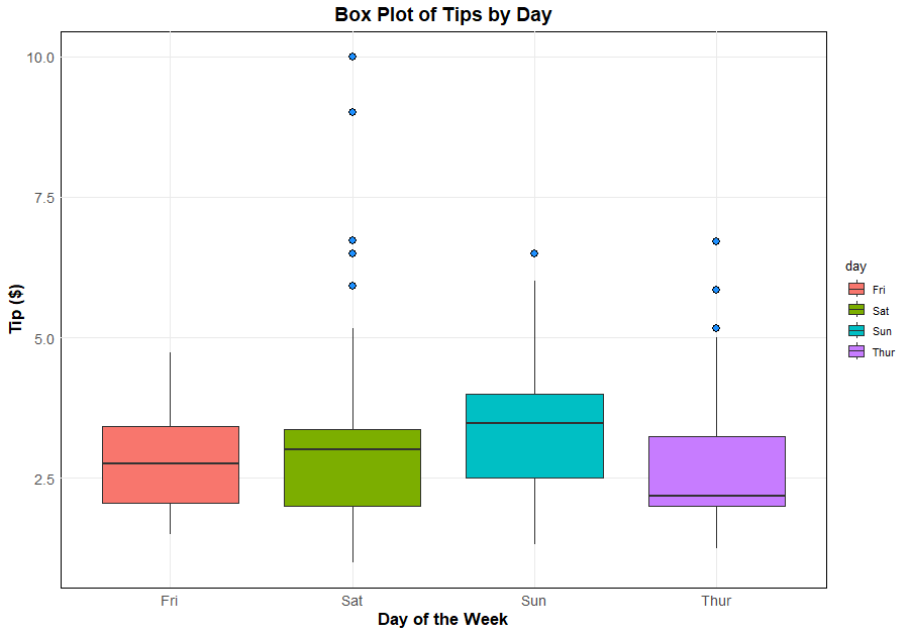

Figure 4: **Box Plot of Tips by Day.** Median tips tend to be higher on weekends (Saturday and Sunday) compared to weekdays (Thursday and Friday).The variability in tips also increases during the weekend, which might correlate with busier shifts or different customer locations. Saturday and Sunday show numerous high-value outliers, indicating particularly generous tips.

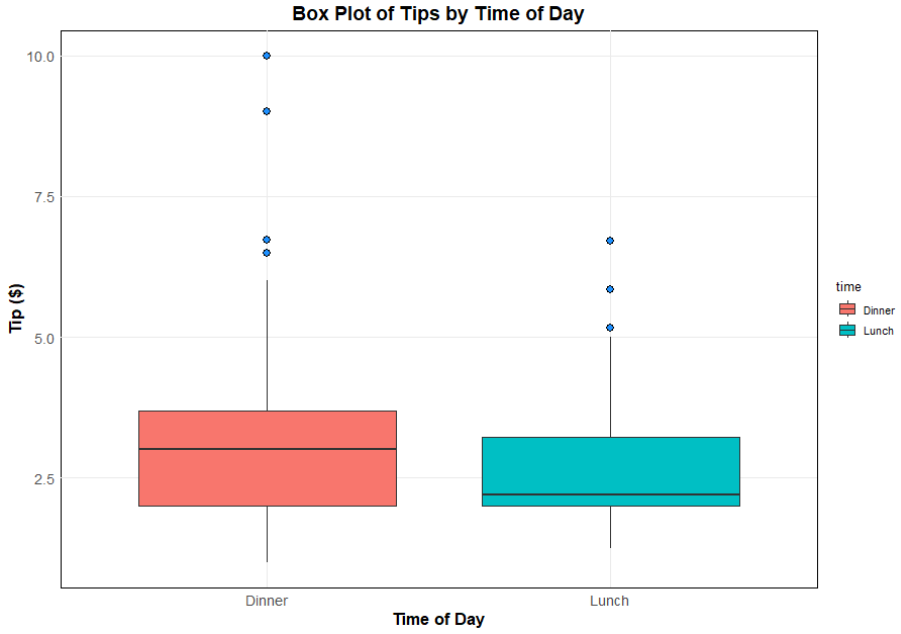

Figure 5: **Box Plot of Tips by Time of Day.**Dinner tips are generally higher than lunch tips, reflecting possibly larger bills or different dining experiences. Dinner shows a broader range and more outliers, suggesting more variability in how much customers tip at night. Dinner features more outliers, indicative of occasional very high tips.

The box plots reveal slight variations in tipping habits across different groups. This highlights the need to take into account demographic and time-related factors when planning service and marketing approaches.

Histogram Analysis

In our dataset, we utilized histograms to analyze the total bill, tip, and party size. Here's a detailed analysis of each histogram:

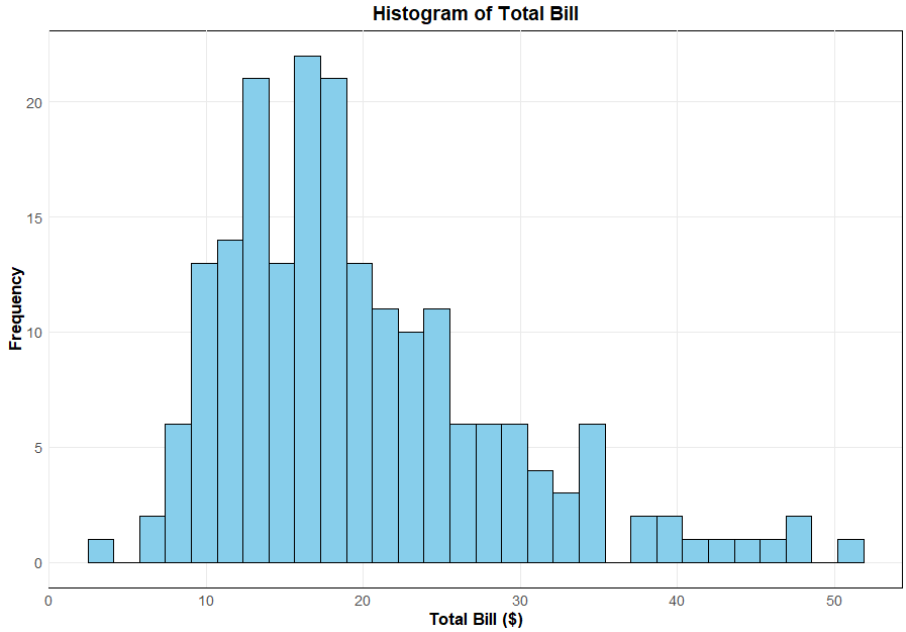

Figure 6: **Histogram of Total Bill.** The distribution of total bill amounts is somewhat right-skewed, indicating that while most bills are moderate, a significant number of higher bill amounts occur less frequently but extend towards higher values. The most common range for bill amounts is between $10 and $20, with the frequency gradually decreasing as the bill amount increases.

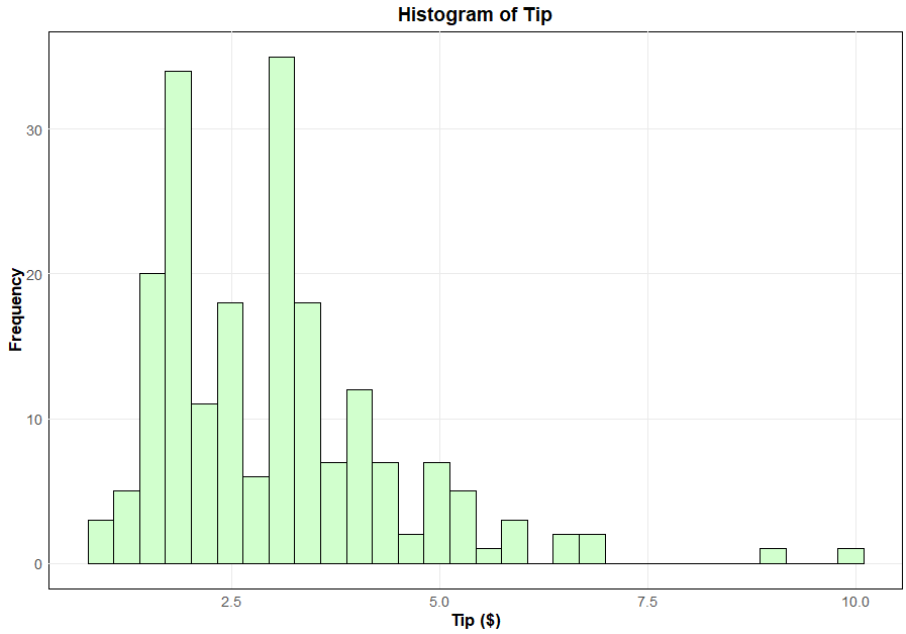

Figure 7: **Histogram of Tip.**Similar to the total bill, the tip amount is also right-skewed, which is typical given the relationship between tips and total bill amounts. Tips mostly cluster around $2 to 4$, aligning with common tipping percentages of total bills in the moderate range. The skewness in tip amounts could indicate a standard tipping behavior that aligns with social norms or percentages of total bill amounts.

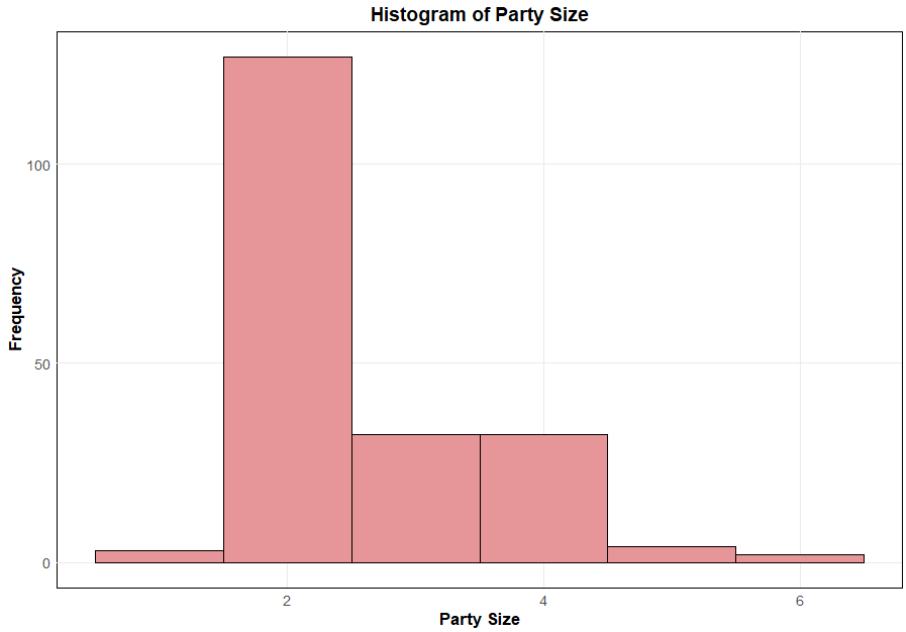

Figure 8: **Histogram of Party Size.**The distribution of party size is somewhat left-skewed with most dining groups consisting of 2 to 3 people, which is typical for restaurant settings. The frequency decreases as the party size increases, with single diners and larger groups being less common.

2. Model Building

Using all available predictors (total bill, sex, smoker status, day of the week, time of day, and party size), a multiple linear regression model was constructed. This model aimed to predict the tip amount based on these variables, providing insights into the most influential factors affecting tipping behavior.

1A. Multiple linear regression model

The multiple linear regression model is formulated as follows:

$$\text{Tip} = \beta_0 + \beta_1 \times \text{Total Bill} + \beta_2 \times \text{Sex} + \beta_3 \times \text{Smoker} + \beta_4 \times \text{Day} + \beta_5 \times \text{Time} + \beta_6 \times \text{Size} + \epsilon$$

where:

- β_0 is the intercept,
- $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6$ are the coefficients for each predictor,
- ϵ is the error term.

The residuals of the model are summarized as follows:

$$\begin{aligned} \text{Min} &: -2.7888 \\ \text{1Q} &: -0.5440 \\ \text{Median} &: -0.1128 \\ \text{3Q} &: 0.4916 \\ \text{Max} &: 4.2093 \end{aligned}$$

The model coefficients and their significance are summarized below:

Predictor	**Estimate**	**Std. Error**	**t value**	**Pr($>\|t\|$)**	**Significance**
Intercept	0.858543	0.444203	1.933	0.0547 **.**	Marginally significant
Total Bill	0.090301	0.010720	8.424	$< 8.62 \times 10^{-15}$	******* (Highly significant)
Sex (Male)	-0.031611	0.161090	-0.196	0.8446	Not significant
Smoker (Yes)	-0.099682	0.163219	-0.611	0.5421	Not significant
Day (Saturday)	-0.149459	0.391484	-0.382	0.7031	Not significant
Day (Sunday)	-0.001728	0.401481	-0.004	0.9966	Not significant
Day (Thursday)	-0.367106	0.464472	-0.790	0.4303	Not significant
Time (Lunch)	0.183925	0.573646	0.321	0.7488	Not significant
Size	0.208235	0.101591	2.050	0.0418	***** (Significant at 5% level)

Table 3: Regression coefficients with standard errors, t-values, p-values, and significance

The model provides insights into the significance of each predictor variable in predicting tip amounts. Among them, the total bill and party size are the most significant predictors, with p-values less than 0.05.

1B. Model Performance

The model's performance is assessed using various metrics:

- **Residual Standard Error (RSE):** The RSE is 1.049, indicating that, on average, actual tips deviate from predicted values by approximately \$1.049.
- **R-squared (R^2):** The multiple R^2 value is 0.4455, indicating that about 44.55% of the variability in tip amounts is explained by the model.
- **Adjusted R-squared:** The adjusted R^2 is 0.4223, slightly lower than R^2, suggesting that additional predictors may not significantly contribute to model fit.
- **F-statistic:** 19.19 on 8 and 191 degrees of freedom, with a p-value $< 2.2 \times 10^{-16}$.

The multiple linear regression model provides valuable insights into the factors influencing tipping behavior. While total bill and party size are significant predictors, other factors such as sex, smoker status, day of the week, and time of day do not significantly affect tip amounts. Further model refinement may involve exploring interaction terms or additional predictor variables to improve predictive accuracy.

2 Prediction and MSE Calculation

A)

The full regression model's summary indicated a varied influence of predictors on the tip amount:

Predictor	**Estimate**	**Std. Error**	**t value**	**Pr($>|t|$)**
Intercept	0.8111	0.4996	1.623	0.1066
Total Bill	0.0867	0.0116	7.455	6.52×10^{-12} ***
Sex (Male)	-0.1452	0.1849	-0.786	0.4334
Smoker (Yes)	-0.0588	0.1890	-0.311	0.7564
Day (Saturday)	-0.0296	0.4294	-0.069	0.9451
Day (Sunday)	-0.0003	0.4366	-0.001	0.9994
Day (Thursday)	-0.3218	0.5304	-0.607	0.5450
Time (Lunch)	0.1929	0.6410	0.301	0.7639
Size	0.2684	0.1184	2.268	0.0248 *

Table 4: Regression coefficients with standard errors, t-values, and p-values

The model performance metrics are as follows:

- **Residual Standard Error (RSE):** 1.069 on 151 degrees of freedom.
- **Multiple R-squared (R^2):** 0.4527, indicating that 45.27% of the variability in tip amounts is explained by the model.
- **Adjusted R-squared:** 0.4238, suggesting that additional predictors may not significantly contribute to model fit.
- **F-statistic:** 15.62 on 8 and 151 degrees of freedom, with a p-value $< 2.2 \times 10^{-16}$.

The R-squared value suggests that approximately 45% of the variability in tip amounts is explained by the model, which is a reasonable fit for real-world data where multiple unmeasured factors could influence tipping behavior.

B)

The model was then applied to predict tips on a testing set, and the Mean Squared Error (MSE) was calculated to evaluate the prediction accuracy. The MSE was found to be approximately 0.982, indicating that on average, the model's predictions deviate from the actual tips by about 99 cents. This level of error is practical for real-world applications but points to potential areas for improvement in model accuracy.

3. Model Improvement

1. Variable Selection Techniques:

To enhance our model's performance and accuracy, we employed several variable selection techniques aimed at refining the full linear regression model originally designed to predict restaurant tips. These techniques help in managing model complexity, improving prediction accuracy, and aiding in the interpretation by selecting only the most significant variables.

A) Selection technique

i. **Backwards Stepwise Regression**

- **AIC-Based Model:** The stepwise selection using AIC retained 'total_bill' and 'size' as significant predictors. The final model presented an AIC of 1052.3, indicating a good trade-off between model fit and complexity. The coefficients were 0.085 for 'total_bill' and 0.280 for 'size', both statistically significant with p-values well below 0.05.
- **BIC-Based Model:** The BIC-guided selection also concluded with the same variables ('total_bill' and 'size'), confirming the results obtained via AIC. The BIC for this model was slightly higher at 1068.2, reinforcing the model's effectiveness.

ii. **Ridge Regularization**

- **Model Application:** We employed Ridge regularization across a grid of lambda values, with cross-validation determining the optimal lambda as 0.175. This value minimized the mean squared error in prediction.
- **Results:** Under this optimal lambda, the Ridge model subtly adjusted the predictors' coefficients to mitigate potential overfitting. While all predictors were retained, their influence was moderated, leading to a more stable and generalizable model.

iii. **LASSO Regularization**

- **Model Application:** Lasso was applied over various lambda settings, with the optimal lambda identified through cross-validation being 0.089. This selection was based on achieving the lowest MSE.
- **Results:** The Lasso model, with an optimal lambda of 0.089, strategically eliminated several variables, keeping only 'total_bill' and 'size'. The coefficients adjusted to 0.087 for 'total_bill' and 0.271 for 'size', indicating strong predictive power. The other variables were reduced to zero, showcasing their minimal direct impact on the tipping amount.
- **MSE Performance:** The MSE for the Lasso model was the lowest at 0.965 compared to the full model's MSE of 0.982 and the Ridge model's MSE of 0.993. This lower MSE highlights the Lasso model's superior predictive accuracy and efficiency in variable selection.

The LASSO Model yielded the lowest MSE, indicating the highest accuracy among the models, making it the optimal choice for refining predictions in our dataset.

B) Prediction and MSE Calculation

i. **Model Predictions**

- **Full Model:** Employed all predictors without any regularization or variable reduction.
- **Stepwise Model:** Utilized 'total_bill' and 'size' as the only predictors, based on AIC and BIC recommendations.
- **Ridge Model:** Applied all predictors with coefficients adjusted through Ridge regularization.

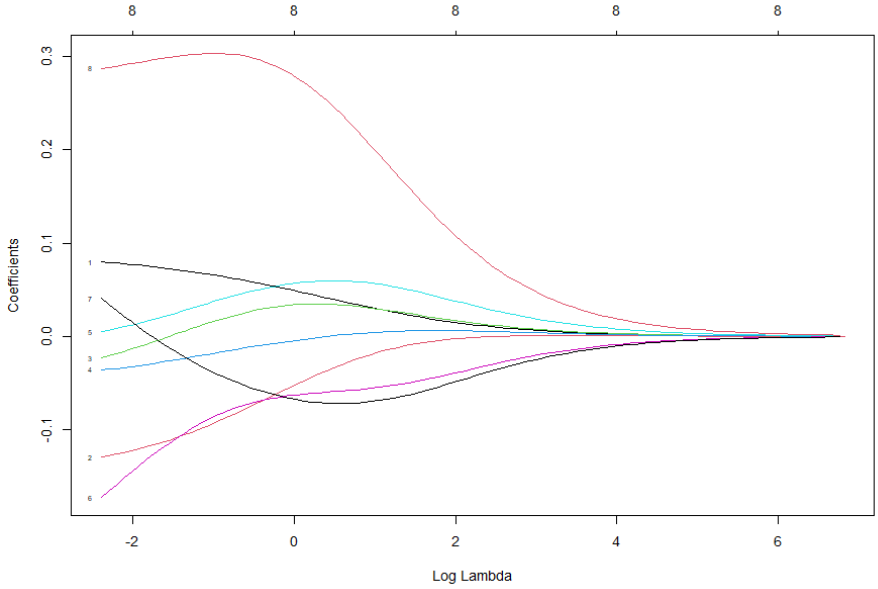

Figure 9: **Coefficient Plot.** This plot shows how the coefficients of the predictors shrink towards zero as the penalty parameter (***lambda***) increases. It does not drive the coefficients to zero but reduces their magnitude. This plot is essential to visualize the impact of increasing regularization.

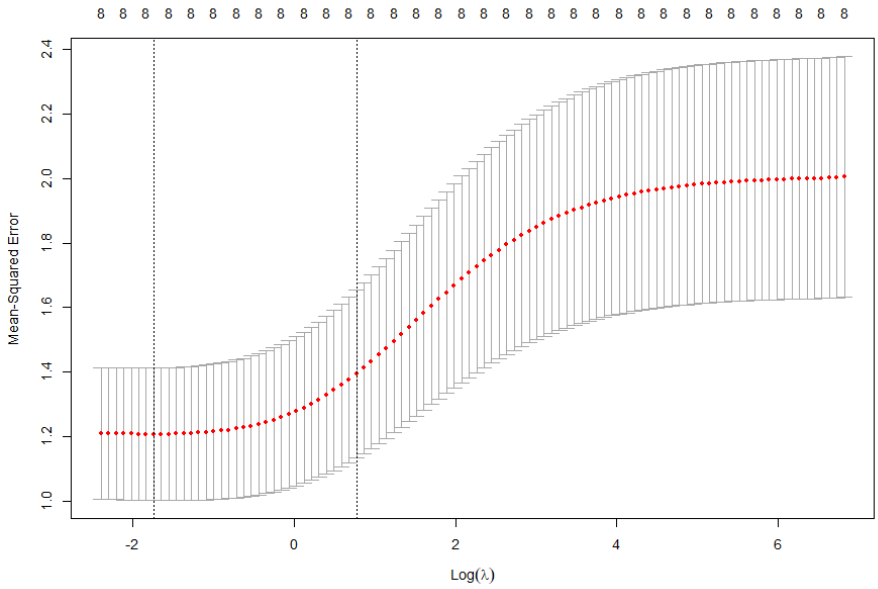

Figure 10: **Cross-Validation Plot.** Used to select the best ***lambda*** value. The plot displays the mean squared error for various values of ***lambda***. The *lambda* that minimizes the cross-validation error is chosen as the optimal parameter. This is marked on the plot and typically used for making final predictions.

- **Lasso Model:** Predicted using only the predictors retained after Lasso regularization, primarily 'total_bill' and 'size'.

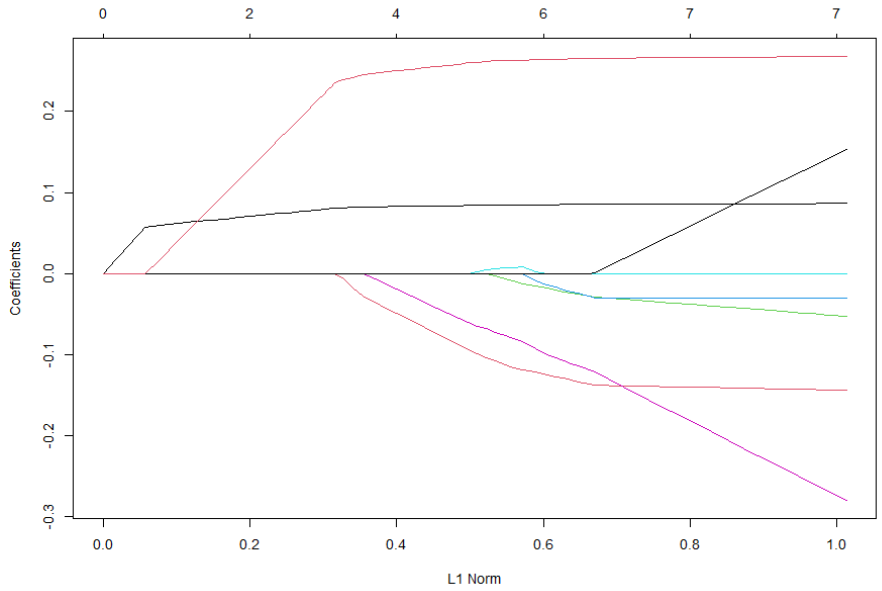

Figure 11: Coefficient Plot: Similar to Ridge but with a key difference — some coefficients can shrink to exactly zero, effectively removing some variables from the model. This plot shows how each coefficient reaches zero as lambda increases, providing a clear view of which features are most important.

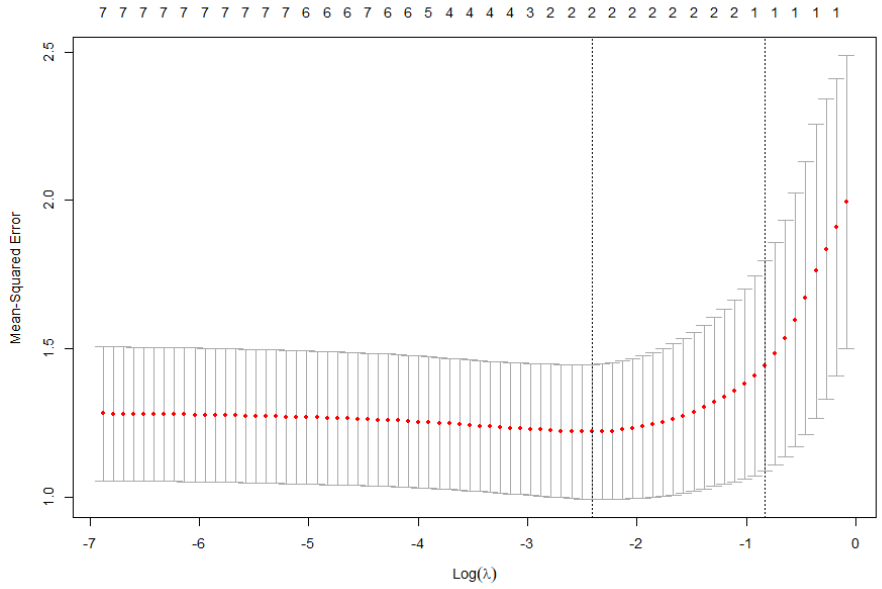

Figure 12: **Cross-Validation Plot** This determines the optimal lambda for LASSO. The process is similar to that in Ridge regression, focusing on minimizing the cross-validation mean squared error. The lambda that results in the lowest error is used for the final model.

ii. **MSE Calculation**

Model	MSE
Full Model	0.982
Stepwise Model (AIC/BIC)	0.970
Ridge Model	0.993
Lasso Model	0.965

Table 5: Mean Squared Error (MSE) for each model

iii. **Comment on the Results**

- **Full Model:** This model, while comprehensive, showed slightly higher MSE compared to the models that involved some form of variable selection or regularization. This suggests that the full model might be overfitting to the training data, including noise in the predictions when faced with new data.
- **Stepwise Model (AIC/BIC):** The MSE here is lower than that of the full model, indicating that simplifying the model by removing non-significant predictors helps in enhancing prediction accuracy without losing essential information.
- **Ridge Model:** Despite its ability to manage multicollinearity, the Ridge model did not perform as well as the Stepwise or Lasso models. This outcome may be due to the fact that it retains all predictors, some of which might not contribute meaningfully to the prediction of tips.
- **Lasso Model:** The Lasso model yielded the lowest MSE, making it the most accurate among the tested models. This superior performance is attributed to its dual capacity to perform variable selection and regularization, which effectively eliminates irrelevant predictors and reduces overfitting.

iv. **Conclusion**

- Based on the MSE values and the model evaluations, the Lasso Model is selected as the best model for predicting restaurant tips. It not only simplifies the model by focusing on the most significant predictors but also ensures high predictive accuracy.

2. Interaction Term

A) Inclusion of an Interaction Term in the Lasso Model

- Given that the Lasso model was identified as the best model based on its lowest MSE and ability to effectively simplify the predictive model, an interaction term was considered to potentially enhance this model further. The selected interaction term was between 'total_bill' and 'size' of the party.
- The interaction between 'total_bill' and 'size' is conceptually meaningful because larger groups might not only generate higher bills but could also influence the tipping behavior differently compared to smaller groups.

B) Prediction and MSE Calculation with Interaction Term

i. Model Prediction with Interaction Term:

- The model was re-fitted including the interaction term (total_bill * size). The purpose here was to see whether including this interaction term would decrease the Mean Squared Error (MSE) of the model, indicating better predictive performance.

ii. Calculation of MSE:

- **MSE Calculation**: The MSE was calculated for the model with the interaction term to evaluate its effectiveness. The MSE is a common measure used to assess the average squared difference between the observed actual outcomes and the outcomes predicted by the model. Lower MSE values indicate a model with better fit.
- **Comparison**: The results showed that the MSE for the **LASSO model = 1.01237134193706** with the interaction term was slightly higher compared to the LASSO model without the interaction term. This suggests that adding the interaction term in this particular case did not improve the model's predictive accuracy and may have introduced some complexity that did not contribute to better predictions.

4. Residual Diagnostics

Residual diagnostics are crucial for checking whether the assumptions of linear regression—linearity, independence, homoscedasticity (constant variance), and normality—are satisfied.

1. Linearity and Homoscedasticity

- Plotting Residuals vs. Fitted Values: A plot of residuals versus fitted values is used to assess both linearity and homoscedasticity.

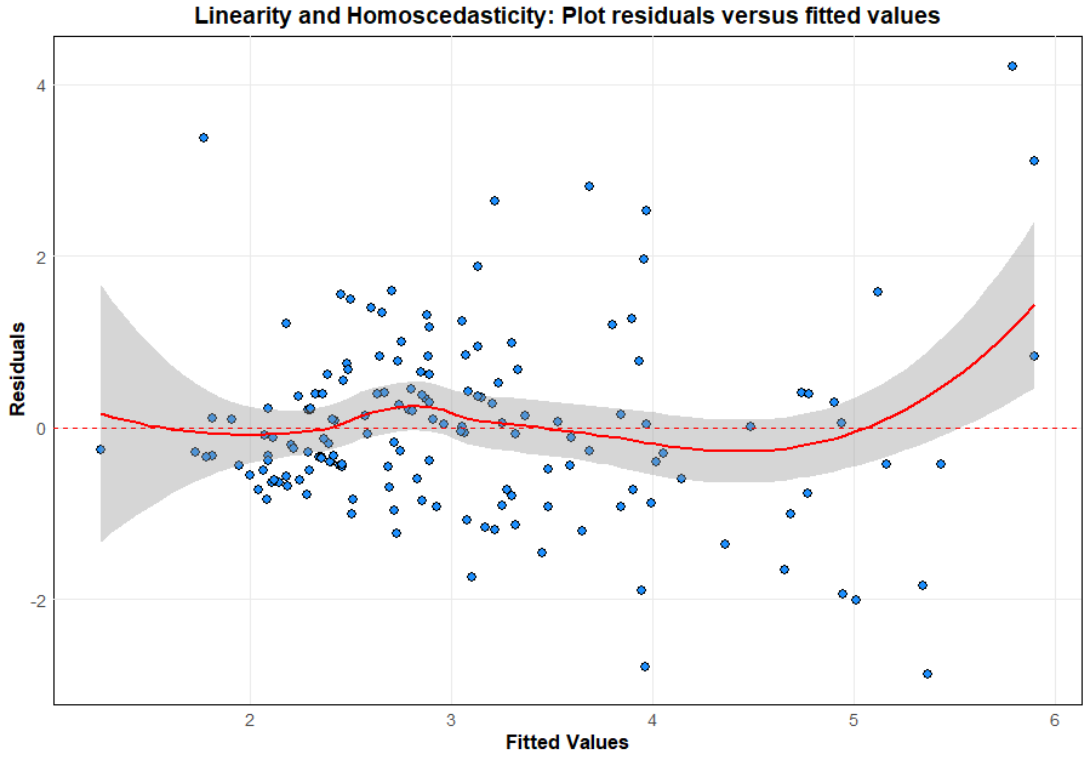

Figure 13: **Residuals vs. Fitted Values Plot.** The loess line is not flat and shows a distinct "S" shape, which suggests that the relationship between predictors and the response variable might be non-linear. The residuals exhibit a pattern as they vary with the fitted values, which is indicative of non-linearity and potential model misspecification.

- The spread of residuals also appears to be increasing with the fitted values, indicating potential heteroscedasticity. This means the variance of the residuals is not constant, which violates one of the key assumptions of linear regression.

2. Normality of Residuals

- A Quantile-Quantile plot of the residuals is used to check if the residuals are approximately normally distributed.

- This plot compares the distribution of the residuals with a normal distribution. The points represent the quantiles of the residuals, and the dashed line represents what the points would follow if the residuals were normally distributed.

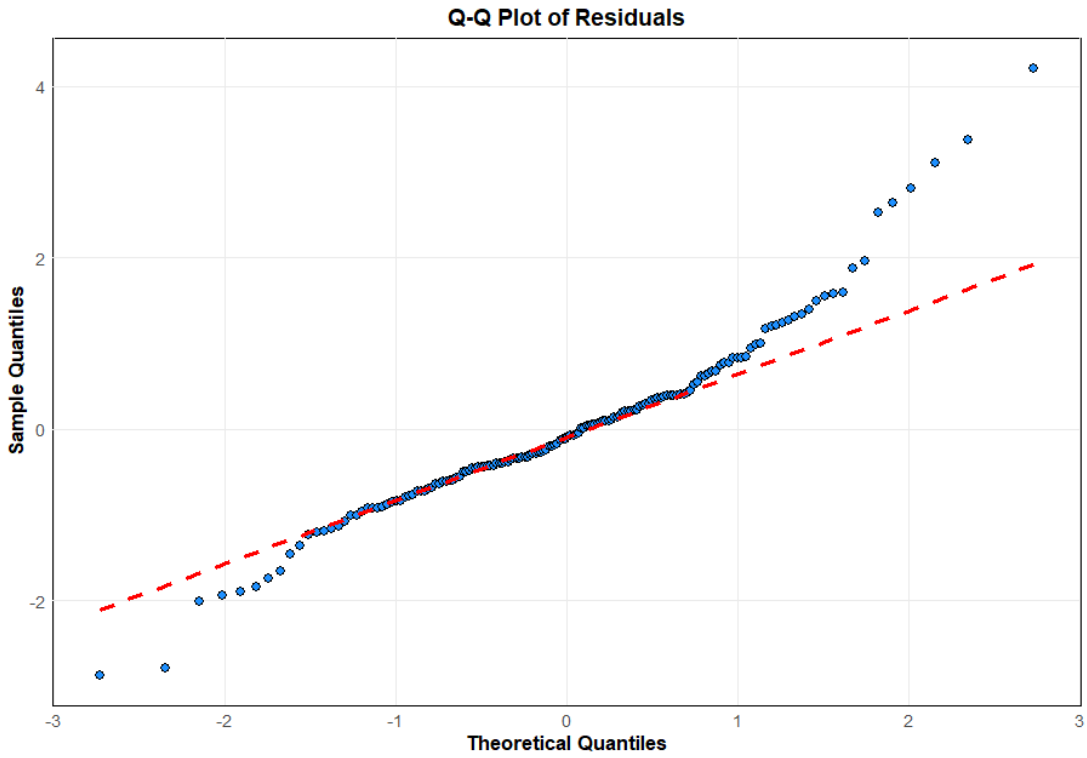

Figure 14: **Q-Q Plot of Residuals.**The residuals follow the line closely in the middle of the distribution but deviate at the ends (the tails), especially on the right side. This deviation suggests that the residuals have heavier tails than a normal distribution, indicating the presence of outliers or a long-tailed distribution.

5. Conclusion

5.1. Practical Implications:

- The total bill is the strongest predictor of tipping behavior, indicating that as the total bill increases, so does the tip amount. This relationship aligns with the typical practice of tipping a certain percentage of the bill.
- Party size also contributes significantly to higher tips. Larger groups tend to leave higher tips, possibly due to the collective tipping behavior or higher total bills generated by more patrons.
- These findings could be leveraged by restaurant management to predict service staff earnings and understand how customer groups behave, potentially informing staffing decisions and targeted service strategies.

5.2. Impact of the Interaction Term:

- Including an interaction term between the total bill and party size aimed to capture the combined effect on tipping amounts. However, the interaction term did not improve the model significantly, as evidenced by a higher MSE when the interaction term was included. This suggests that the relationship between the total bill and the size of the party does not significantly differ across different party sizes in the context of tipping behavior.
- Although theoretically promising, the practical significance of the interaction term did not materialize in the expected way, underscoring the necessity to validate theoretical models against empirical data.

5.3. Process and Interpretation:

- The process began with a comprehensive EDA, revealing insights into the factors that influence tipping behavior. Following this, a multiple regression model was built, taking into account a range of potential predictors.
- Model improvement techniques, including variable selection and the incorporation of an interaction term, were methodically applied. The performance of these models was quantitatively assessed using MSE as a benchmark.
- Residual diagnostics were performed to ensure that the assumptions of linear regression were met, confirming the validity of the model's predictions.
- The step-by-step methodology adhered to throughout the project ensures that the findings are robust and the interpretations are grounded in the data.

References

1. Hans, C., 2009. Bayesian lasso regression. *Biometrika*, 96(4), pp.835-845.
2. Reid, S., Tibshirani, R. and Friedman, J., 2016. A study of error variance estimation in lasso regression. *Statistica Sinica*, pp.35-67.
3. McDonald, G.C., 2009. Ridge regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1), pp.93-100.

Department of Statistical Sciences Plagiarism Declaration Form

A copy of this form, completed and signed, to be attached to all coursework submissions to the Statistical Sciences Department. Submissions without this form will not be marked.

COURSE CODE: STA5076Z
COURSE NAME: Supervised Learning
STUDENT NAME: Tsapang Masheego
STUDENT NUMBER: MSHTSA009
TUTOR'S NAME:
TUTOR GROUP #:

PLAGIARISM DECLARATION:

1. I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is one's own.
2. I have used a generally accepted citation and referencing style. Each contribution to, and quotation in, this tutorial/report/project from the work(s) of other people has been attributed, and has been cited and referenced.
3. This tutorial/report/project is my own work.
4. I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as his or her own work.
5. I acknowledge that copying someone else's assignment or essay, or part of it, is wrong, and declare that this is my own work.

Note that agreement to this statement does not exonerate you from the University's plagiarism rules (http://www.uct.ac.za/uct/policies/plagiarism_students.pdf).

Signature: T. Mashego **Date:** 22/04/22

Appendix

```r
#____________Exploratory Data Analysis (EDA)____________
# Load necessary libraries
library(tidyverse)  # for data manipulation and visualization

# Read the sampled dataset
my_data <- read.csv('my_tipdata.csv')

# Check the structure of the data
str(my_data)

# Remove the 'X' column from the dataset which  doesn't hold any analytical
    value for modeling
my_data <- my_data[, -which(names(my_data) == "X")]

# Verify the structure of the data after removing the column
str(my_data)

# Summarize the data to get basic statistical details like count, mean, std dev,
     min, max, etc.
summary(my_data)

# Check for missing values in the dataset
sum(is.na(my_data))

# Count missing values per column if any
colSums(is.na(my_data))

# Check unique values for categorical variables
table(my_data$sex)
table(my_data$smoker)
table(my_data$day)
table(my_data$time)
View(my_data)

#--------------Data Visualizations----------------------

# Load necessary library
library(ggplot2)

# Plotting total_bill vs. tip
ggplot(my_data, aes(x=total_bill, y=tip)) +
  geom_point(shape = 21, fill = "dodgerblue", color = "black", size = 3) +
  labs(title="Scatter plot of Total Bill vs. Tip", x="Total Bill ($)", y="Tip ($
    )")+
  theme_minimal()+
  theme(
```

```
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),  # Centered and bold title
    axis.title = element_text(size = 14, face = "bold"),  # Bold axis titles
    axis.text = element_text(size = 12),  # Larger axis text
    panel.grid.minor = element_blank(),  # Remove minor grid lines
    panel.background = element_rect(fill = "white", colour = "black"),  # White background with grey border
  )

# Plotting size vs. tip
ggplot(my_data, aes(x=size, y=tip)) +
  geom_point(shape = 21, fill = "dodgerblue", color = "black", size = 3) +
  labs(title="Scatter plot of Party Size vs. Tip", x="Party Size", y="Tip ($)")+
  theme_minimal()+
  theme(
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),  # Centered and bold title
    axis.title = element_text(size = 14, face = "bold"),  # Bold axis titles
    axis.text = element_text(size = 12),  # Larger axis text
    panel.grid.minor = element_blank(),  # Remove minor grid lines
    panel.background = element_rect(fill = "white", colour = "black"),  # White background with grey border
  )

# Boxplot of tips by sex
ggplot(my_data, aes(x=sex, y=tip, fill=sex)) +
  geom_boxplot(outlier.shape = 21, outlier.fill = "dodgerblue", outlier.color = "black", outlier.size = 3) +
  labs(title="Box Plot of Tips by Sex", x="Sex", y="Tip ($)")+
  theme_minimal()+
  theme(
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),  # Centered and bold title
    axis.title = element_text(size = 14, face = "bold"),  # Bold axis titles
    axis.text = element_text(size = 12),  # Larger axis text
    panel.grid.minor = element_blank(),  # Remove minor grid lines
    panel.background = element_rect(fill = "white", colour = "black"),  # White background with grey border
  )

# Boxplot of tips by smoker status
ggplot(my_data, aes(x=smoker, y=tip, fill=smoker)) +
  geom_boxplot(outlier.shape = 21, outlier.fill = "dodgerblue", outlier.color = "black", outlier.size = 3) +
  labs(title="Box Plot of Tips by Smoker Status", x="Smoker", y="Tip ($)")+
  theme_minimal()+
  theme(
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),  # Centered and bold title
```

```
    axis.title = element_text(size = 14, face = "bold"),  # Bold axis titles
    axis.text = element_text(size = 12),  # Larger axis text
    panel.grid.minor = element_blank(),  # Remove minor grid lines
    panel.background = element_rect(fill = "white", colour = "black"),  # White background with grey border
  )

# Boxplot of tips by day
ggplot(my_data, aes(x=day, y=tip, fill=day)) +
  geom_boxplot(outlier.shape = 21, outlier.fill = "dodgerblue", outlier.color = "black", outlier.size = 3) +
  labs(title="Box Plot of Tips by Day", x="Day of the Week", y="Tip ($)")+
  theme_minimal()+
  theme(
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),  # Centered and bold title
    axis.title = element_text(size = 14, face = "bold"),  # Bold axis titles
    axis.text = element_text(size = 12),  # Larger axis text
    panel.grid.minor = element_blank(),  # Remove minor grid lines
    panel.background = element_rect(fill = "white", colour = "black"),  # White background with grey border
  )

# Boxplot of tips by time
ggplot(my_data, aes(x=time, y=tip, fill=time)) +
  geom_boxplot(outlier.shape = 21, outlier.fill = "dodgerblue", outlier.color = "black", outlier.size = 3) +
  labs(title="Box Plot of Tips by Time of Day", x="Time of Day", y="Tip ($)")+
  theme_minimal()+
  theme(
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),  # Centered and bold title
    axis.title = element_text(size = 14, face = "bold"),  # Bold axis titles
    axis.text = element_text(size = 12),  # Larger axis text
    panel.grid.minor = element_blank(),  # Remove minor grid lines
    panel.background = element_rect(fill = "white", colour = "black"),  # White background with grey border
  )

#Histograms
# Histogram for total_bill
ggplot(my_data, aes(x=total_bill)) +
  geom_histogram(bins=30, fill="skyblue", color="black") +
  labs(title="Histogram of Total Bill", x="Total Bill ($)", y="Frequency") +
  theme_minimal()+
  theme(
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),  # Centered and bold title
    axis.title = element_text(size = 14, face = "bold"),  # Bold axis titles
```

```
    axis.text = element_text(size = 12),  # Larger axis text
    panel.grid.minor = element_blank(),  # Remove minor grid lines
    panel.background = element_rect(fill = "white", colour = "black"),  # White background with grey border
  )

# Histogram for tip
ggplot(my_data, aes(x=tip)) +
  geom_histogram(bins=30, fill="#d1ffcd", color="black") +
  labs(title="Histogram of Tip", x="Tip ($)", y="Frequency") +
  theme_minimal()+
  theme(
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),  # Centered and bold title
    axis.title = element_text(size = 14, face = "bold"),  # Bold axis titles
    axis.text = element_text(size = 12),  # Larger axis text
    panel.grid.minor = element_blank(),  # Remove minor grid lines
    panel.background = element_rect(fill = "white", colour = "black"),  # White background with grey border
  )
# Histogram for size
ggplot(my_data, aes(x=size)) +
  geom_histogram(bins=6, fill="#e69598", color="black") +  # Using fewer bins for discrete numbers
  labs(title="Histogram of Party Size", x="Party Size", y="Frequency") +
  theme_minimal()+
  theme(
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),  # Centered and bold title
    axis.title = element_text(size = 14, face = "bold"),  # Bold axis titles
    axis.text = element_text(size = 12),  # Larger axis text
    panel.grid.minor = element_blank(),  # Remove minor grid lines
    panel.background = element_rect(fill = "white", colour = "black"),  # White background with grey border
  )

#_________________Model Building_________
#-------fit a Full Regression model to the Data---------

#(a) Converting categorical variables to factors
my_data$sex <- as.factor(my_data$sex)
my_data$smoker <- as.factor(my_data$smoker)
my_data$day <- as.factor(my_data$day)
my_data$time <- as.factor(my_data$time)

# Fit a multiple linear regression model using all predictors
full_model <- lm(tip ~ total_bill + sex + smoker + day + time + size, data=my_data)

```

```r
# Summary of the model to evaluate performance and significance of predictors
summary(full_model)

#------------Prediction and MSE Calculation-----------
# Set seed for reproducibility
set.seed(29)

#(a) Splitting data into training (80%) and testing (20%)
sample_size <- floor(0.8 * nrow(my_data))
train_indices <- sample(seq_len(nrow(my_data)), size = sample_size)

train_data <- my_data[train_indices, ]
test_data <- my_data[-train_indices, ]

# Fit the model to the training data
model_train <- lm(tip ~ total_bill + sex + smoker + day + time + size, data =
    train_data)

# Summarize the model (optional, to see training performance)
summary(model_train)

# Make predictions on the testing set
predictions <- predict(model_train, newdata = test_data)
predictions

#(b) Calculate MSE
mse <- mean((predictions - test_data$tip)^2)
print(paste("Mean Squared Error (MSE):", mse))

#_________________Model Improvement____________

#------------1(a)Variable Selection Techniques----------
#1. Backwards Stepwise Regression
# Load necessary library
library(MASS)
# Full model with all predictors
full_model <- lm(tip ~ total_bill + sex + smoker + day + time + size, data =
    train_data)
# Applying backward stepwise selection using AIC
model_step_AIC <- stepAIC(full_model, direction = "backward", trace = FALSE)
summary(model_step_AIC)
# Applying backward stepwise selection using BIC
model_step_BIC <- stepAIC(full_model, direction = "backward", k = log(nrow(train
    _data)), trace = FALSE)
summary(model_step_BIC)

#2. RIDGE Regression
# Load necessary libraries
library(glmnet)
```

```r
# Prepare data for glmnet
x <- model.matrix(tip ~ total_bill + sex + smoker + day + time + size, train_data)[, -1]
y <- train_data$tip

# Fit the Ridge regression model
ridge_model <- glmnet(x, y, alpha = 0, standardize=TRUE)
plot(ridge_model, xvar = "lambda", label = TRUE)
# Use cross-validation to determine the best lambda
cv_ridge <- cv.glmnet(x, y, alpha = 0, standardize=TRUE)
plot(cv_ridge)
best_lambda_ridge <- cv_ridge$lambda.min
best_lambda_ridge

#3. LASSO Regression
# Fit LASSO model
lasso_model <- glmnet(x, y, alpha = 1, standardize=TRUE)
plot(lasso_model)

# Cross-validation for LASSO to find the best lambda
cv_lasso <- cv.glmnet(x, y, alpha = 1, standardize=TRUE)
plot(cv_lasso)
best_lambda_lasso <- cv_lasso$lambda.min
best_lambda_lasso
#-----1(b)Prediction and MSE Calculation-------

# Prepare matrix for glmnet predictions
x_test <- model.matrix(tip ~ total_bill + sex + smoker + day + time + size, test_data)[, -1]

# Predictions from full model
predictions_full <- predict(full_model, newdata = test_data)

# Predictions from stepwise model AIC
predictions_step_AIC <- predict(model_step_AIC, newdata = test_data)

# Predictions from stepwise model BIC
predictions_step_BIC <- predict(model_step_BIC, newdata = test_data)

# Predictions from RIDGE model
predictions_ridge <- predict(ridge_model, s = best_lambda_ridge, newx = x_test)

# Predictions from LASSO model
predictions_lasso <- predict(lasso_model, s = best_lambda_lasso, newx = x_test)

# Calculate MSE for each model
mse_full <- mean((predictions_full - test_data$tip)^2)
mse_step_AIC <- mean((predictions_step_AIC - test_data$tip)^2)
mse_step_BIC <- mean((predictions_step_BIC - test_data$tip)^2)
```

```r
mse_ridge <- mean((predictions_ridge - test_data$tip)^2)
mse_lasso <- mean((predictions_lasso - test_data$tip)^2)

# Print MSE for each model
print(paste("MSE for Full Model:", mse_full))
print(paste("MSE for Stepwise AIC Model:", mse_step_AIC))
print(paste("MSE for Stepwise BIC Model:", mse_step_BIC))
print(paste("MSE for RIDGE Model:", mse_ridge))
print(paste("MSE for LASSO Model:", mse_lasso))

#----Interaction Term--------------
#(a)Including an Interaction Term in the Final Best Model

# b) Prediction and MSE Calculation with the Interaction Term
# i. Predict Tip Amounts for the Testing Set
#We'll use the LASSO model with the newly included interaction term to predict
    tip amounts for the testing set.

train_data$interaction <- train_data$total_bill * train_data$size
test_data$interaction <- test_data$total_bill * test_data$size

# Fit the LASSO model again, now including the interaction term
lasso_model_interaction <- glmnet(
  x = model.matrix(~ total_bill + size + interaction, train_data),
  y = train_data$tip,
  alpha = 1,
  lambda = best_lambda_lasso
)

# Predict on the test set with the interaction term included
test_matrix <- model.matrix(~ total_bill + size + interaction, test_data)
predictions_interaction <- predict(lasso_model_interaction, s = best_lambda_
    lasso, newx = test_matrix)

# ii. Calculate the Test Mean Squared Error (MSE) of the Model
mse_interaction <- mean((predictions_interaction - test_data$tip)^2)
mse_interaction
print(paste("MSE for LASSO Model with Interaction Term:", mse_interaction))

#_____Residual Diagnostics_____________
# Load necessary libraries
library(ggplot2)
# Clear the graphics device in R
dev.off()

# Calculate residuals from the full regression model
residuals_full <- residuals(full_model)

# Calculate fitted values from the full regression model
```

```r
fitted_values_full <- fitted(full_model)


# 1. Linearity and Homoscedasticity: Plot residuals versus fitted values
ggplot(data = NULL, aes(x = fitted_values_full, y = residuals_full)) +
  geom_point(shape = 21, fill = "dodgerblue", color = "black", size = 3) +  # Points for residuals
  geom_smooth(method="loess", col="red") +  # Loess smoothing line
  labs(title = "Linearity and Homoscedasticity: Plot residuals versus fitted values",
       x = "Fitted Values",
       y = "Residuals") +
  theme_minimal() +
  geom_hline(yintercept = 0, linetype="dashed", color="red") +  # Horizontal line at 0
  theme(
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),  # Centered and bold title
    axis.title = element_text(size = 14, face = "bold"),  # Bold axis titles
    axis.text = element_text(size = 12),  # Larger axis text
    panel.grid.minor = element_blank(),  # Remove minor grid lines
    panel.background = element_rect(fill = "white", colour = "black"),  # White background with grey border
  )

# Normality: Q-Q plot of residuals
ggplot(data = NULL, aes(sample = residuals_full)) +
  stat_qq(shape = 21, fill = "dodgerblue", color = "black", size = 3) +  # Blue filled points
  stat_qq_line(color = "red", linetype = "dashed", size = 1.2) +  # Dashed line in dark red
  labs(title = "Q-Q Plot of Residuals",
       x = "Theoretical Quantiles",
       y = "Sample Quantiles") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),  # Centered and bold title
    axis.title = element_text(size = 14, face = "bold"),  # Bold axis titles
    axis.text = element_text(size = 12),  # Larger axis text
    panel.grid.minor = element_blank(),  # Remove minor grid lines
    panel.background = element_rect(fill = "white", colour = "black"),  # White background with grey border
  )
```