7COM1079-0901-2024 - Team Research and Development Project

Final report title: Zomato Pune Restaurant Analysis Report: Exploring Factors Influencing Ratings

Group ID: A309

Dataset number: DS 298

Prepared by:

Matthew Alamu - 23073691 - Ma24aeu

Ahasan Nasir - 23073136 - Ahsannasirmehdi

Salman Javaid - 23080081 - Salman31945243

SIBT UL HASSAN - 23034082 - Sibtul620

Stephen Marson - 21044571 - Steppa24Uberes

University of Hertfordshire

Hatfield, 2024

Table of Contents

1. Introduction 3
   1. Problem statement and research motivation 3
   2. The data set 3
   3. Research question 3
   4. Null hypothesis and alternative hypothesis (H0/H1) 3

1. Background research 4
   1. Research papers (at least 3 relevant to your topic / DS) 4
   2. Why RQ is of interest (research gap and future directions according to the literature) 4

1. Visualisation
   1. Appropriate plot for the RQ *output of an R script 5*
   2. Additional information relating to understanding the data 6
   3. Useful information for the data understanding 8

1. Analysis
   1. Statistical test used to test the hypotheses and output 9
   2. The null hypothesis is rejected /not rejected based on the p-value 9

1. Evaluation – group’s experience at 7COM1079 9
   1. What went well 9
   2. Points for improvement 9
   3. Group’s time management 9
   4. Project’s overall judgement 10
   5. Comment on GitHub log output 10

1. Conclusions 10
   1. Results explained. 10
   2. Interpretation of the results
   3. Reasons and/or implications for future work, limitations of your study 10

1. Reference list 11
2. Appendices
3. R code used for analysis and visualisation. 11
4. GitHub log output. 15

1. Introduction

* 1. Problem statement and research motivation **(100 words)**

The restaurant sector holds considerable economic importance globally. Understanding the drivers behind consumer behavior and satisfaction within this sector is critical for both businesses and academics. Customer ratings provide essential feedback concerning perceived service and product quality, while price, a core component of value assessment, significantly influences consumer dining decisions. Investigating the relationship between price and customer rating provides insight into consumer choice. Specifically in the online domain, analysis of restaurant data offers an opportunity to monitor trends (Xu, 2007) which is of interest for those in the hospitality sector. This allows for businesses to create pricing strategies that are more attuned to consumer expectations.

* 1. The data set **(75 words)**

This study employs a dataset of restaurant listings from Pune, India, collected from Zomato. The dataset includes variables such as ratings, prices for two, location and provided amenities. These various attributes offer significant opportunity to investigate factors that may correlate with customer preferences. The dataset is comprised of primarily numerical values as well as categorical text. The time range included is up until December 2019, so presents a snapshot from this time. This also provides a historic point to contrast with more recent data.

* 1. Research question **(50 words)**

This research aims to evaluate if an association exists between restaurant price and customer rating, as indicated on the Zomato platform. The central inquiry: what is the link between ‘Charges for two’ and ‘Ratings out of 5’ for restaurants within Pune? A statistical test using these two variables will help identify the correlation of interest between these variables, and help answer our research question.

* 1. Null hypothesis and alternative hypothesis (H0/H1) **(100 words)**

The null hypothesis (H0) posits that no statistically significant correlation is present between a restaurant's ‘Charges for two’ and its ‘Ratings out of 5’. This means that any observed association between price and customer ratings can be explained by randomness, not any real link. Conversely, the alternative hypothesis (H1) posits that a statistically significant correlation does exist between these two factors. This implies an actual relationship is present that cannot be explained by chance. The statistical test to verify these hypothesis will proceed using a 0.05 level of significance.

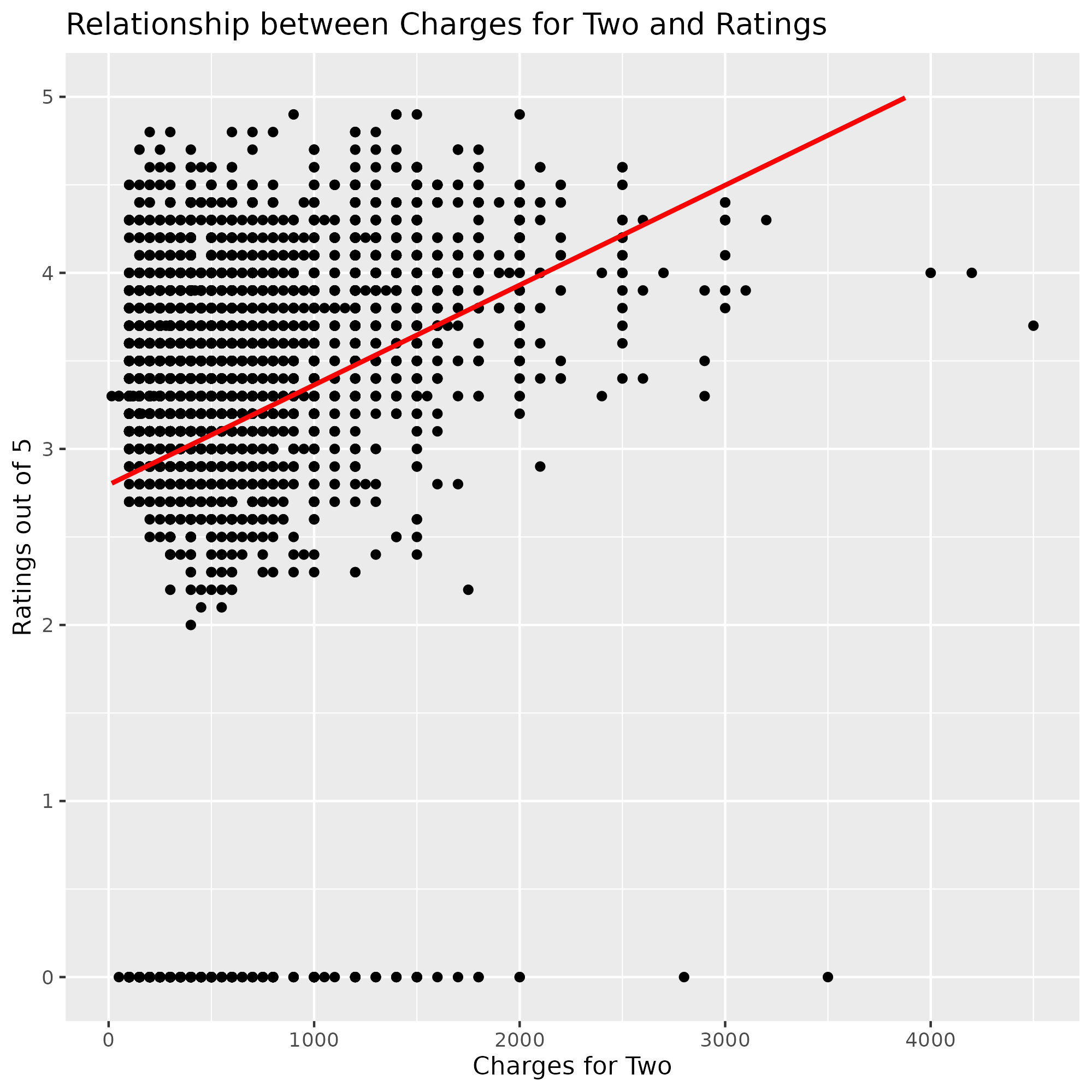
1. Background research
   1. Research papers (at least 3 relevant to your topic / DS) **(200 words)**

Several academic papers have studied various factors influencing restaurant choice. Frederick and Bhat (2021) examined the impact of customer perception towards online food delivery services. Their analysis found a positive relationship between tangible service elements, reliability, and perceived quality, concluding that the customer experience is key. Anderson and Mittal (2000) also explore the importance of service quality. They studied how responsiveness and how efficient delivery influences overall consumer satisfaction. The work of these authors demonstrate that these specific elements of the consumer experience are highly impactful on satisfaction and choice. Park et al. (2007) studied the effect of online reviews on consumer behavior within the food sector. They demonstrated the power of online reviews in customer decision-making and how reviews may impact choice of restaurant. These research works offer important context, highlighting that elements beyond price can influence restaurant choice, and how online platforms impact consumer behavior.

* 1. Why RQ is of interest (research gap and future directions according to the literature) **(100 word**s)

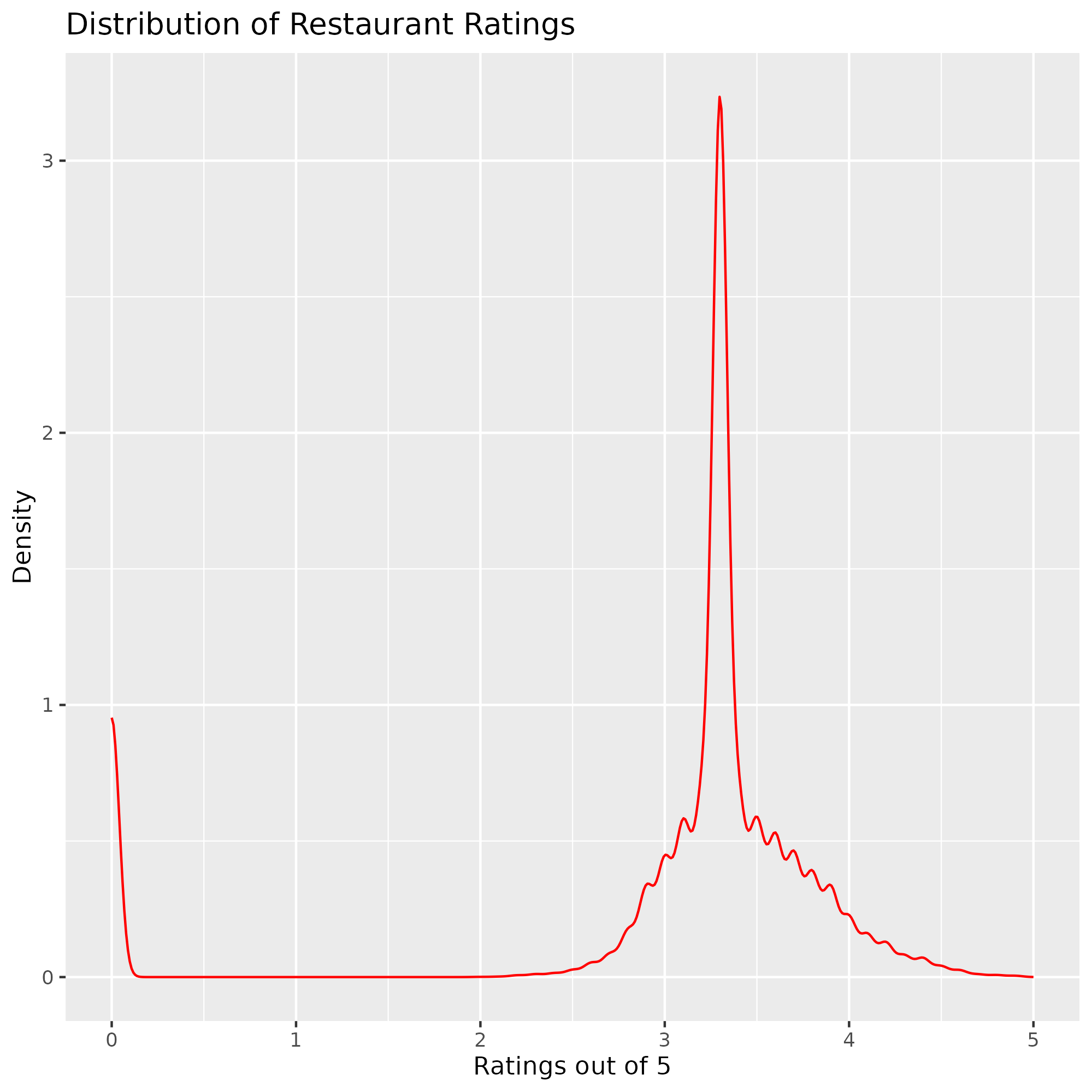
The prior work provides a base of understanding by demonstrating the importance of service and reviews. However, there is limited explicit research focusing on price correlation with the rating for Pune restaurants from Zomato data. Understanding this price and rating relationship is of practical significance for business strategy and pricing within this sector. Future work could explore these effects over time. Additional studies might also investigate how different cultural or demographic factors influence rating and price sensitivity within specific localities in Pune, as well as specific cuisine types (Hennig-Thurau et al, 2004). Such factors can provide more detail and clarity on the effects investigated in the current study.

1. Visualisation
   1. Appropriate plot for the RQ(**50 words)**

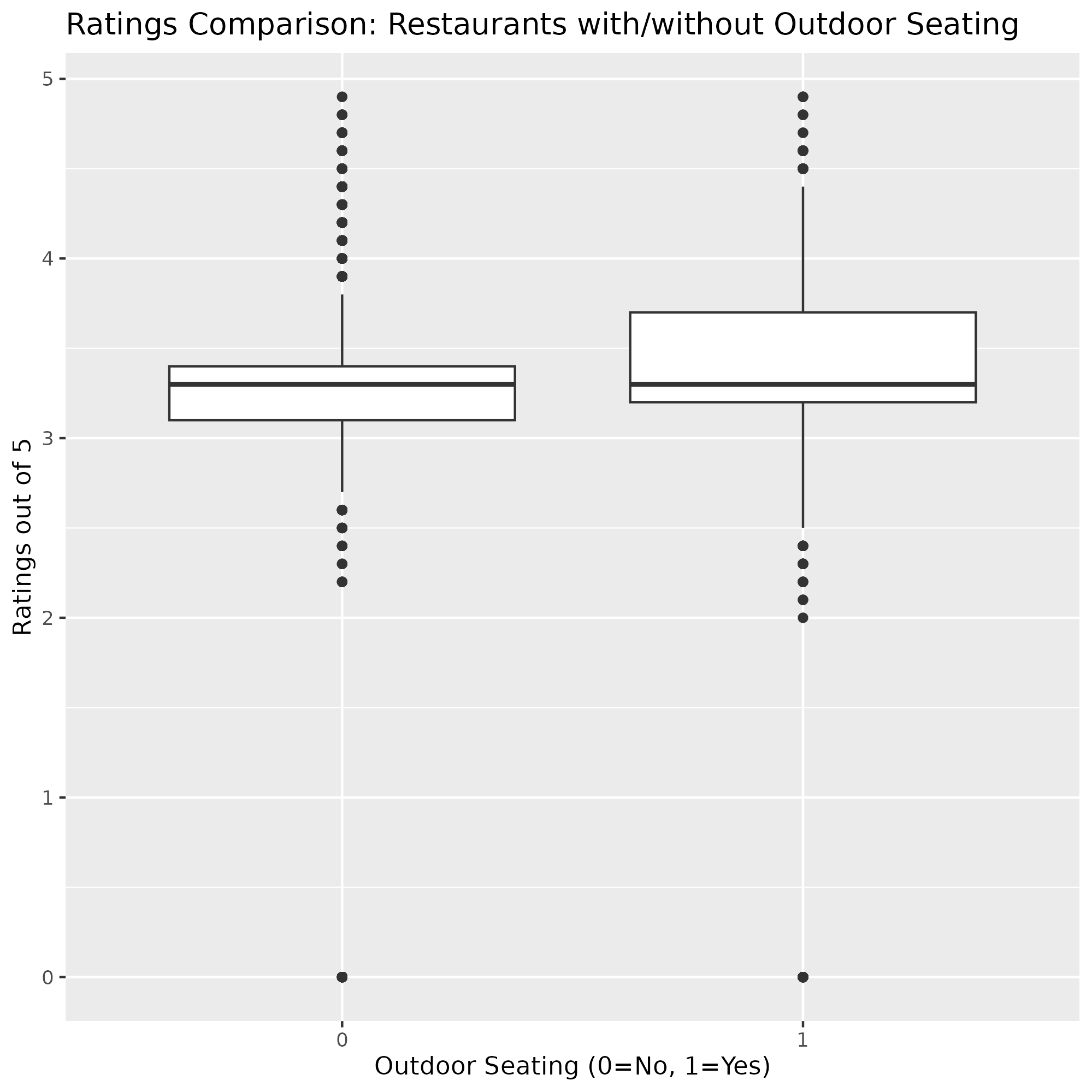


A scatter plot illustrates the connection between restaurant price and rating. This plot displays each restaurant as a point on a graph, with ‘Charges for two’ on the x axis and ‘Ratings out of 5’ on the y axis. This design allows the visual examination of any relationships between the two variables. A linear trend line further aids in the visual understanding of patterns.

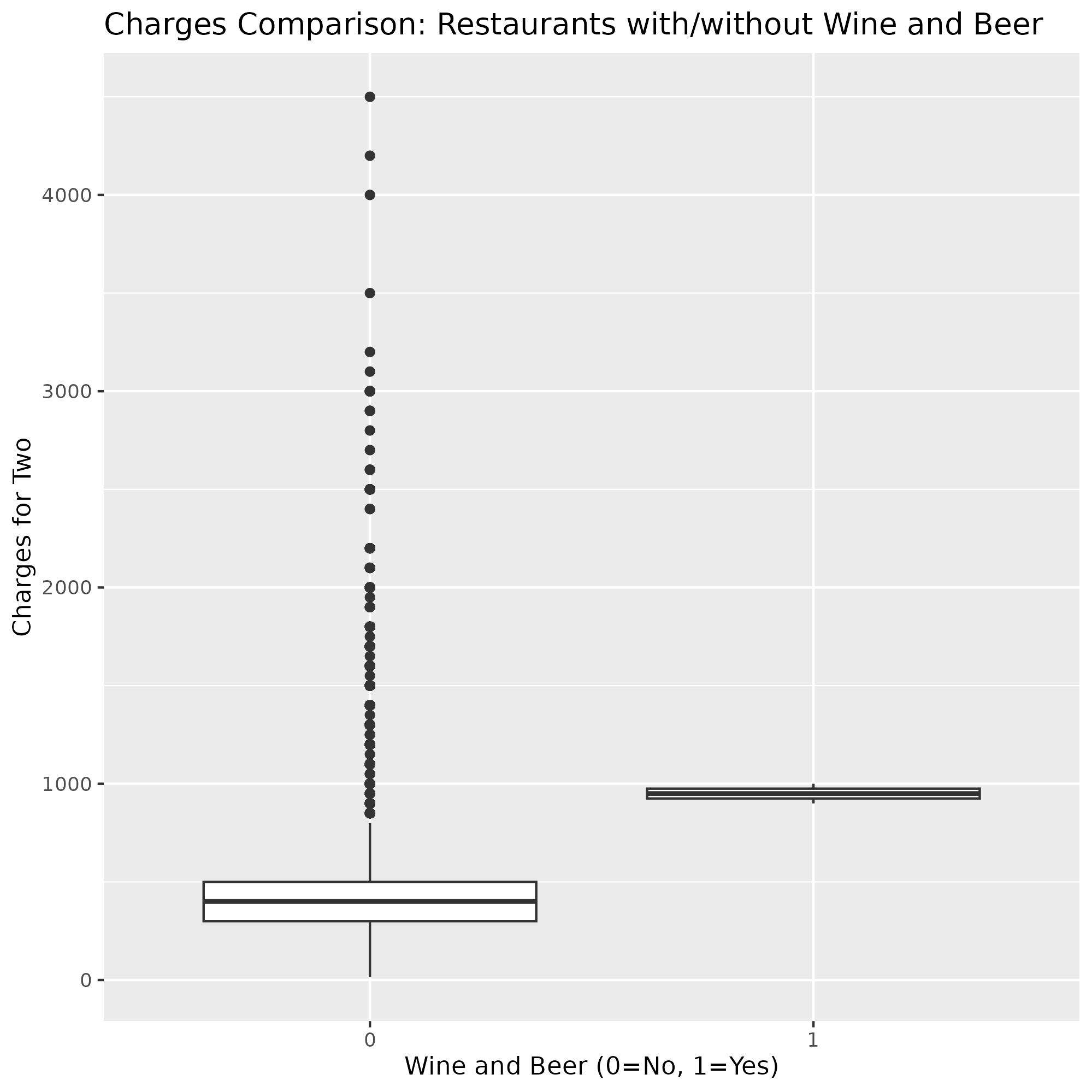
* 1. Additional information relating to understanding the data (optional) (**50 words)**



The line chart visualizes the rating data. This assists in understanding the data distribution of the main dependent variable, particularly normality. The red density line helps clarify the underlying shape of the rating distribution. The graph shows if a normal assumption can be supported for this variable. It shows a noticeable skewness.



The histogram visualizes the distribution of ratings, aiding in understanding the variable's normality. The red density line clarifies the underlying shape. The box plot of ‘Ratings\_out\_of\_5’ against ‘Outdoor Seating’ shows if there is any notable difference in ratings between restaurants offering outdoor seating or not.



The box plot comparing ‘Charges\_for\_two’ for restaurants with and without ‘Wine and Beer’ highlights how prices differ between restaurants offering and not offering alcohol. Both box plots help give a better insight into the distribution of our data.

* 1. Useful information for the data understanding (**50 words)**

The scatter plot shows a generally weak positive relationship between price and rating, with a dense cluster at lower prices and high ratings. The histogram reveals that the ratings are left skewed, with many restaurants obtaining relatively high ratings and the data does not appear to follow a normal distribution.

The scatter plot reveals some weak positive correlation, with a concentration of data at lower prices and higher ratings. The histogram shows that the ratings data is left skewed and does not appear normal. The box plot for ‘Outdoor Seating’ shows that both groups have a similar median but a slightly smaller IQR is found with restaurants offering this. The ‘Wine and Beer’ box plot demonstrates that restaurants providing alcohol have a higher price range than other restaurants in the data. The median is also considerably higher for these restaurants.

1. Analysis
   1. Statistical test used to test the hypotheses and output (**75 words)**

Pearson’s correlation test is used. It allows for examining the strength and direction of linear connections between two numeric variables. This test method is selected because the two variables are of continuous numeric form. This test method provides a numeric value for the relationship between our independent and dependent variables. The result was a correlation of .178. The p value is very low (1.257e-87).

* 1. The null hypothesis is rejected /not rejected based on the p-value (**100 words)**

The p-value (1.257e-87) is below the significance level at 0.05. This outcome leads to the rejection of the null hypothesis. There is a statistically significant correlation present between price and rating. The correlation coefficient indicates a weak positive correlation at 0.178. This means there is a small tendency for restaurants with higher prices to have slightly elevated ratings, although the connection is not very dominant.

1. Evaluation – group’s experience at 7COM1079
   1. What went well **(75 words)**

The group effectively utilized online collaborative tools for communication and file sharing. This allowed us to work independently and integrate changes smoothly. The task distribution was equal with each member contributing to the research, analysis, and writing. Adherence to deadlines were met without issue, which is a sign of success. We managed to complete every task to a good standard of quality.

* 1. Points for improvement **(75 words)**

Initial code integration caused minor delays because of some data type conflicts. Deeper understanding of the R tool prior to starting may have been beneficial to mitigate any such issues. Better version control from the beginning of the project may have improved the clarity of code development. More frequent testing could also have been useful during code development.

* 1. Group’s time management (**50 words)**

The team initially dedicated more time to data exploration and literature review, which required some speed adjustment later in the project. The group prioritized code development after setting out the main structure of the project. Clear time management strategies were used throughout, which ensured the project's completion.

* 1. Project’s overall judgement (**50 words)**

The project was largely a success. It provided key insights into restaurant pricing. The use of the data analytics tools were effective. The results supported the research question. Overall, our work provided us with key skills for further work.

* 1. Note any changes to group since submission of Assignment 1. Add new or amended GitHub Ids for new members **(75 words, write only if applies to your group arrangements)**

No changes to group membership occurred since the submission of Assignment 1. The original group of members completed the task effectively. The github ids remain the same as before. Further changes are not needed.

* 1. Comment on the GitHub log output **(50 words)**

The GitHub log, detailed in Appendix B, records all key stages of the project. Three significant commits involved developing R code, completing the final report, and changing the python code to R. The log accurately tracks project's progress and provides an accurate view of the workflow.

1. Conclusions
   1. Results explained (**75 words)**

The statistical analysis reveals a significant but weak positive correlation between price and rating, measured with the Pearson correlation coefficient. The p value of this correlation was extremely small. The histogram of ratings reveals a left skewed distribution, indicating a tendency towards higher ratings. The boxplots reveal differences between groups for both outdoor seating and alcohol availability.

* 1. Interpretation of the results (**75 words)**

The correlation result suggests that more expensive restaurants only tend to receive slightly higher ratings. The data’s distribution shows that most restaurants achieve above average ratings. Outdoor seating does not provide much difference in median value, but the range is smaller, which shows that this feature does impact variability. Serving alcohol leads to considerably higher average prices.

* 1. Reasons and/or implications for future work, limitations of your study (**50 words)**

This study has a specific timeframe (up to 2019) and does not consider the impact of social media or other digital channels. Future work should integrate these to provide more detailed insights. There may also be other confounding variables that are not explored in this work.

1. Reference list ***(not included in the work count)***

Anderson, E.W. and Mittal, V., 2000. Strengthening the satisfaction-profit chain. Journal of Service research, 3(2), pp.107-120.

Frederick, D.P. and Bhat, G., 2021. Review on customer perception towards online food delivery services. International Journal of Creative Thoughts (IJCRT), 9(7), pp.b301-b314.

Hennig-Thurau, T., Gwinner, K.P., Walsh, G. and Gremler, D.D., 2004. Electronic word-of-mouth via consumer-opinion platforms: what motivates consumers to articulate themselves on the internet?. Journal of interactive marketing, 18(1), pp.38-52.

Park, D.H., Lee, J. and Han, I., 2007. The effect of on-line consumer reviews on consumer purchasing intention: The moderating role of involvement. International journal of electronic commerce, 11(4), pp.125-148.

Xu, X., 2020. Examining the role of emotion in online consumer reviews of various attributes in the surprise box shopping model. Decision Support Systems, 136, p.113344.

1. Appendices
2. R code used for analysis and visualisation ***(not included in the word count)***

Analysis.R code with the appropriate statistics to test the hypotheses.

# Install required packages if not already installed

if(!require(dplyr)){install.packages("dplyr")}

if(!require(ggplot2)){install.packages("ggplot2")}

if(!require(readr)){install.packages("readr")}

if(!require(corrplot)){install.packages("corrplot")}

if(!require(e1071)){install.packages("e1071")}

# Load libraries

library(dplyr)

library(ggplot2)

library(readr)

library(corrplot)

library(e1071)

# --- 1. Dataset Loading and Initial Cleaning ---

df <- read\_csv("zomato\_pune\_V002.csv")

# Convert relevant columns to numeric (some are characters but should be numeric)

numeric\_cols <- c('Ratings\_out\_of\_5', 'Number of votes', 'Charges\_for\_two',

'5\_star\_review\_percentage', '4\_star\_review\_percentage',

'3\_star\_review\_percentage', '2\_star\_review\_percentage',

'1\_star\_review\_percentage')

df <- df %>%

mutate(across(all\_of(numeric\_cols), ~as.numeric(gsub("[^0-9.]", "", .))))

# Handle missing values (using median imputation)

df <- df %>%

mutate(across(all\_of(numeric\_cols), ~ifelse(is.na(.), median(., na.rm = TRUE), .)))

# Display first two rows of data for report

head(df, 2) %>% print()

# Dataset Row Representation

cat("\nEach row in the dataset represents a restaurant in Pune, including its features, reviews, and ratings.\n")

# --- Outlier Removal (Charges\_for\_two based on 4 standard deviations) ---

mean\_charges <- mean(df$Charges\_for\_two)

std\_charges <- sd(df$Charges\_for\_two)

cutoff <- 4 \* std\_charges

lower\_limit <- mean\_charges - cutoff

upper\_limit <- mean\_charges + cutoff

df <- df %>%

filter(Charges\_for\_two >= lower\_limit & Charges\_for\_two <= upper\_limit)

# --- 2. Research Question & Hypotheses Formulation ---

research\_question <- "What is the relationship between the price range of a restaurant (Charges\_for\_two) and its customer rating (Ratings\_out\_of\_5) in Pune, based on Zomato data? Is there a statistically significant correlation between these two variables?"

null\_hypothesis <- "There is no statistically significant correlation between a restaurant's Charges\_for\_two and its Ratings\_out\_of\_5 in Pune."

alternative\_hypothesis <- "There is a statistically significant correlation between a restaurant's Charges\_for\_two and its Ratings\_out\_of\_5 in Pune."

cat(paste("\nResearch Question:", research\_question, "\n"))

cat(paste("Null Hypothesis:", null\_hypothesis, "\n"))

cat(paste("Alternative Hypothesis:", alternative\_hypothesis, "\n"))

# --- 3. Data Visualization ---

# 3a. Scatterplot of Charges\_for\_two vs Ratings\_out\_of\_5

scatter\_plot <- ggplot(df, aes(x = Charges\_for\_two, y = Ratings\_out\_of\_5)) +

geom\_point() +

geom\_smooth(method = "lm", se = FALSE, color = "red") +

labs(title = "Relationship between Charges for Two and Ratings",

x = "Charges for Two",

y = "Ratings out of 5") +

ylim(0, 5) # Set y-axis limits to 0-5

ggsave("scatterplot\_charges\_ratings.png", plot=scatter\_plot)

print(scatter\_plot)

# 3b. Histogram of Ratings

histogram\_plot <- ggplot(df, aes(x = Ratings\_out\_of\_5)) +

geom\_histogram(aes(y = ..density..), bins = seq(0, 5.5, 0.5), fill = "skyblue", color = "black") +

geom\_density(color = "red") +

labs(title = "Distribution of Restaurant Ratings",

x = "Ratings out of 5",

y = "Density") +

xlim(0,5) # set x-axis limits 0-5

ggsave("histogram\_ratings.png", plot=histogram\_plot)

print(histogram\_plot)

# 3c. Interpret histogram

ratings\_skewness <- skewness(df$Ratings\_out\_of\_5)

if (ratings\_skewness > 0.5) {

histogram\_interpretation <- "The histogram shows a right-skewed distribution of restaurant ratings."

} else if (ratings\_skewness < -0.5) {

histogram\_interpretation <- "The histogram shows a left-skewed distribution of restaurant ratings."

} else if (kurtosis(df$Ratings\_out\_of\_5) > 3) {

histogram\_interpretation <- "The histogram shows a leptokurtic distribution with heavy tails"

}else if (kurtosis(df$Ratings\_out\_of\_5) < 3) {

histogram\_interpretation <- "The histogram shows a platykurtic distribution with thin tails"

} else {

histogram\_interpretation <- "The histogram shows an approximately normal distribution of restaurant ratings."

}

cat(paste("Histogram interpretation:", histogram\_interpretation, "The density curve helps visualize the overall shape of the distribution.\n"))

# --- 4. Statistical Analysis ---

# 4a. Calculate correlation coefficient

correlation\_result <- cor.test(df$Ratings\_out\_of\_5, df$Charges\_for\_two, method = "pearson")

correlation <- correlation\_result$estimate

p\_value <- correlation\_result$p.value

cat(paste("Pearson Correlation Coefficient between Ratings and Charges:", correlation, "\n"))

# 4b. Programmatic Interpretation of Correlation

alpha <- 0.05

if (p\_value < alpha) {

if (correlation > 0.5) {

interpretation <- "There is a statistically significant strong positive correlation between charges and ratings."

} else if (correlation > 0.3) {

interpretation <- "There is a statistically significant moderate positive correlation between charges and ratings."

} else if (correlation > 0) {

interpretation <- "There is a statistically significant weak positive correlation between charges and ratings."

} else if (correlation < -0.5) {

interpretation <- "There is a statistically significant strong negative correlation between charges and ratings."

} else if (correlation < -0.3) {

interpretation <- "There is a statistically significant moderate negative correlation between charges and ratings."

} else {

interpretation <- "There is a statistically significant weak negative correlation between charges and ratings."

}

} else {

interpretation <- "There is no statistically significant correlation between charges and ratings."

}

cat(paste("Correlation Interpretation:", interpretation, "\n"))

cat(paste("P-value:", p\_value, "\n"))

# --- 5. Additional Analysis and Visualization ---

#5a. Box Plot: Ratings vs Outdoor Seating

boxplot\_outdoor <- ggplot(df, aes(x = factor(`Outdoor Seating`), y = Ratings\_out\_of\_5)) +

geom\_boxplot() +

labs(title = "Ratings Comparison: Restaurants with/without Outdoor Seating",

x = "Outdoor Seating (0=No, 1=Yes)",

y = "Ratings out of 5")

ggsave("boxplot\_outdoor\_seating.png", plot=boxplot\_outdoor)

print(boxplot\_outdoor)

#5b. Box Plot: Charges vs Wine and Beer

boxplot\_wine <- ggplot(df, aes(x = factor(`Wine and Beer`), y = Charges\_for\_two)) +

geom\_boxplot() +

labs(title = "Charges Comparison: Restaurants with/without Wine and Beer",

x = "Wine and Beer (0=No, 1=Yes)",

y = "Charges for Two")

ggsave("boxplot\_wine\_beer.png", plot=boxplot\_wine)

print(boxplot\_wine)

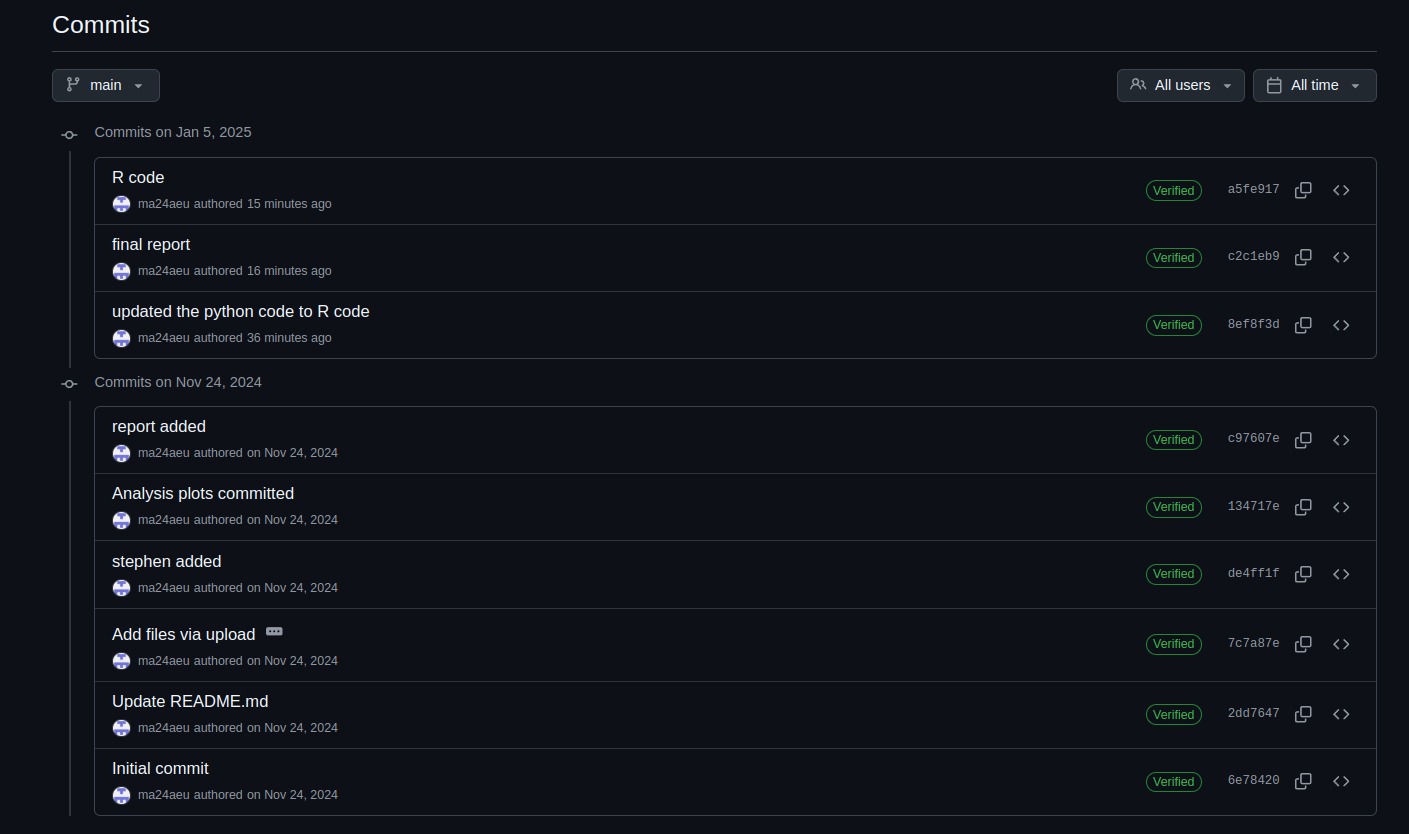
# Print the correlation and p value

cat(paste("Pearson Correlation Coefficient:", correlation, "\n"))

cat(paste("P-value:", p\_value, "\n"))

1. GitHub log output.

GitHub repository link: https://github.com/ma24aeu/7com1079-A309

Figure 1: github logs