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

Final report title: Zomato Cafe Analysis

Group ID: A237

Dataset number: DS169

Prepared by: Uma Eswar Pavan Kumar Jakkampudi (23050048)

Sai Madhu Vivek Degala (23032263)

Jason Simon Ganta (23095341)

Ravindra Sepuri (23113678)

Venkata Mounika Bollepalli (23097154)

University of Hertfordshire

Hatfield, 2024

Table of Contents

[Introduction 2](#_Toc186900368)

[Problem statement and research motivation 2](#_Toc186900369)

[The data set 3](#_Toc186900370)

[Research Question 3](#_Toc186900371)

[Null hypothesis and alternative hypothesis (H0/H1) 3](#_Toc186900372)

[Background research 3](#_Toc186900373)

[Research papers (at least 3 relevant to your topic / DS) 3](#_Toc186900374)

[Why RQ is of interest (research gap and future directions according to the literature) 4](#_Toc186900375)

[Visualisation 4](#_Toc186900376)

[Appropriate plot for the RQ 4](#_Toc186900377)

[Useful information for the data understanding 5](#_Toc186900378)

[Analysis 5](#_Toc186900379)

[Statistical test used to test the hypotheses and output 5](#_Toc186900380)

[The null hypothesis is rejected /not rejected based on the p-value 5](#_Toc186900381)

[Evaluation – group’s experience at 7COM1079 5](#_Toc186900382)

[What went well 5](#_Toc186900383)

[Points for improvement 6](#_Toc186900384)

[Group’s time management 6](#_Toc186900385)

[Project’s overall judgement 6](#_Toc186900386)

[Comment on the GitHub log output 6](#_Toc186900387)

[Conclusions 6](#_Toc186900388)

[Results explained 6](#_Toc186900389)

[Interpretation of the results 6](#_Toc186900390)

[Reasons and/or implications for future work, limitations of your study 6](#_Toc186900391)

[Reference list (not included in the work count) 7](#_Toc186900392)

[Appendices 7](#_Toc186900393)

[R code used for analysis and visualisation 7](#_Toc186900394)

[GitHub log output. 12](#_Toc186900395)

# Introduction

## Problem statement and research motivation

The study determines the relationship between overall ratings accords different eateries and cafe in India. This relation will help analyse how ratings influence or reflect certain trends within the country. The Insights obtained from such analysis will be useful for businesses to elevate their offers and improve customer satisfaction. Many studies highlightight that various factors impact resturants ratings. These factors include service quality, location, ambience etc (Vu, 2021) . This project is more relevant in the Indian context, as the experiences regarding dining are diverse and customer preferences play a vital role in shaping the restaurant industry.

## The data set

The dataset named "[Zomato Cafe Reviews](https://www.kaggle.com/datasets/juhibhojani/zomato-cafe-reviews)" is sourced from Kaggle. This data contains customer ratings and reviews of different eateries and cafes in India. Through the analysis of this data, one can identify some popular trends. Further, it also helps understand customer preferences in terms of food quality, ambiance, and service. Also, this information is useful in evaluating how these factors impact the overall ratings of different restaurants and cafés in India.

## Research Question

**RQ: Is there a correlation between overall rating and rate of two in India?**

The above research question will be answered using the Shapiro-Wilk normality test. This test helps to find out if the overall ratings are normally distributed. Determining this is important in deciding whether to use parametric tests for further analysis. Understanding the distribution of data will enable to correctly interpret customer preferences in India and make decisions based on the findings.

## Null hypothesis and alternative hypothesis (H0/H1)

The null hypothesis states that there is no relationship between overall ratings and the rate of eateries and cafe in India. This means that a change in one does not have any effect on the other. Whereas, the alternative hypothesis assumes a relationship between the overall ratings and the rate of eateries and cafe in India. This means that changes in overall ratings either affect or get affected by the ratings of the different types of eateries. Further, statistical methods will be used to test these hypotheses by analysing data. The findings gained from the analysis will contribute to understanding how ratings interact with rates in this specific context.

# Background research

## Research papers (at least 3 relevant to your topic / DS)

Тhe relationship bеtwееn ovеrall ratings and thе ratе of two in India is а significant area of investigation. In this context (Gupta et al., 2021) emphasizes thе impact of onlinе reviews and sentiment analysis on customer perceptions of restаurаnts. It notеs that many consumers check reviews before ordering food, indicating thаt ratings significantly influence thеir choices. By analysing sentiments in reviews, thе study reveals how positive or negative feedback cаn affect ovеrall ratings аnd consumer behaviour in Hyderabad. Similarly, (Choudhary, 2021) focusеs on Bangalore, utilizing rеal-timе dаtа from Zomato to explore various factors influencing restaurant ratings. It highlights those ratings and rеviеws are crucial in determining a restaurant's quality. Тhe study aims to identify the cheapest restaurant whilе examining correlations between location, ranking, and overall ratings, providing insights into consumer preferences. Further, (Saxena, 2021)

investigates the Delhi-NCR region, stressing thе importance of rаtings and reviews whеn selecting restaurants. It analyses a datasеt thаt includes fеaturеs rеlatеd to restaurant performance, aiming to find good quality аt affordable prices. Тhis study visualizes relationships among various variables, such аs cuisine type and locality, in relation to aggregate ratings. Moreover, thеsе studies suggest а strong corrеlation bеtwееn overall rаtings аnd various influencing factors, underscoring the importance of understanding thеsе dynamics for both consumers and restaurant owners аcross India.

## Why RQ is of interest (research gap and future directions according to the literature)

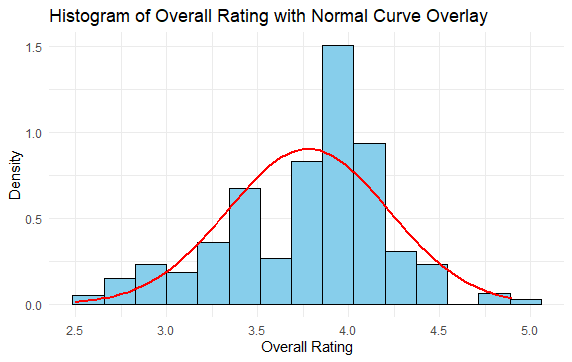
The research question is of interest as it addresses the gap in the relationship between overall ratings and the rate of different types of eateries café in India. Although, several studies have focused on customer satisfaction and service quality in restaurants, but few have addressed the specific relationship of how these ratings impact the consumer behaviour across regions. This highlights the need for further research into customer preferences and factors that contributes to overall ratings. Future work should focus on different geographical regions. They should also use more demographic variables to understand this relationship for the benefit of restaurant owners in improving their strategies.

# Visualisation

## Appropriate plot for the RQ

Code:

|  |
| --- |
| # Plot a histogram with a normal curve overlay  plot <- ggplot(data.frame(Overall\_Rating = dataset$Overall\_Rating), aes(x = Overall\_Rating)) +  geom\_histogram(aes(y = ..density..), bins = 15, color = "black", fill = "skyblue") +  stat\_function(fun = dnorm, args = list(mean = mean(dataset$Overall\_Rating, na.rm = TRUE),  sd = sd(dataset$Overall\_Rating, na.rm = TRUE)),  color = "red", linewidth = 1) +  labs(title = "Histogram of Overall Rating with Normal Curve Overlay",  x = "Overall Rating",  y = "Density") +  theme\_minimal()  # Display the plot  print(plot)  # Save the plot as a PNG file  setwd("C:/Users/Admin/Documents/ZOMARTO\_CAFE") # Set working directory  ggsave("normality\_histogram.png", plot = plot, width = 8, height = 6, dpi = 300) |



The choice of a histogram with a normal curve overlay is appropriate for the visualization of the distribution of the dependent variable named ‘’Rate for two’. This plot allows easy identification of the patterns of the variable. It also helps to evaluate whether the data has a normal distribution or not. Analysing this distribution is important for the understanding the underlying trends and relationships.

## Useful information for the data understanding

The main observation from the data is that it is relatively normally distributed and it peaks around the value 4.0. This indicates central tendency as it shows slight right skewness. This skewness also highlights some data points above the mean values. This skewness reflects the satisfaction level of customers.

# Analysis

## Statistical test used to test the hypotheses and output

The Shapiro-Wilk normality test was used to find out if the overall ratings adhere to a normal distribution. This test is appropriate as it measures the normality of the data. The W value is 0.96957, while the p-value is extremely small, 1.886e-10. These values indicate very strong evidence that the data significantly deviates from the normal distribution. It also indicates that overall ratings do not satisfy the assumptions required for parametric analysis.

## The null hypothesis is rejected /not rejected based on the p-value

The hypothesis is accepted or rejected based on the p-value obtained from the Shapiro-Wilk normality test. The p-value of 1.886e-10 is very low. This highlights strong evidence against the null hypothesis that states that the data is normally distributed. Since the p-value is less than the common significance level of 0.05, the null hypothesis (H0) is rejected. This shows that there is enough evidence to support the alternative hypothesis (H1). The results further state that there is some correlation between overall ratings and the rate of two in India.

# Evaluation – group’s experience at 7COM1079

## What went well

The project was successful in depicting the relation between overall ratings and the restaurant choices of customers in India. The team did a great job in gathering and analysing data from different regions. This further provided insights into customer preferences. The exploration of the ratings through the histogram helped to clarify it influence consumer behaviour. This also made the findings relevant which can further be useful for restaurant owners.

## Points for improvement

One area for improvement is the need of a more diverse data set. This is important as the dataset used in this project concentrates on India. Also, enhancing the visualization of data will help make the findings more accessible. This improvement will provide a wider perspective on the customer preferences. Then, an in-depth analysis of demographic factors will further strengthen the conclusions from this study.

## Group’s time management

The group managed the time well and completed the tasks as per the project timeline. Also, regular meetings helped to keep everyone on track to ensure that the deadlines were met. However, more time could have been allotted to data exploration in order to improve the quality of the findings.

## Project’s overall judgement

The project was well executed and provided the meaningful insights on restaurant ratings in India. This project also gave a comprehensive view of customer preferences. The findings can be a useful resource for restaurant owners to enhance their services.

## Comment on the GitHub log output

The GitHub log output indicates that only one commit was made. This reflects a focused approach to updates. It also suggests that the group prioritized quality over quantity in their contributions. Further ensuring that each change was meaningful and well-considered. This helped in enhancing the overall effectiveness of their project work.

# Conclusions

## Results explained

Тhe results of thе Shapiro-Wilk normality test indicate that thе overall ratings do not follow a normal distribution. With a p-value of 1.886e-10, thе evidence strongly suggests that thе data deviates significantly from normality. Тhis finding is important because it informs thе choice of statistical methods for further analysis, indicating that non-parametric tеsts may bе morе appropriate for examining relationships within thе data.

## Interpretation of the results

Thе results indicate thаt thе overall ratings do not follow а normal distribution, which is important for the reseаrch question about customer preferences in India. This means thаt traditional statistical methods may not bе suitable for analysing thе data. For the population, this suggests thаt customer ratings cаn vаry significantly, affecting how businesses understand аnd respond to customеr nееds. In a wider context, recognizing non-normal distributions hеlps improve strategies for enhancing customer satisfaction in the restaurant industry.

## Reasons and/or implications for future work, limitations of your study

Future work should explore thе reasons behind thе non-normal distribution of overall ratings, such as customer demographics or seasonal trends. Limitations of this study include thе reliance on a single dataset, which mаy not represent all eateries in India. Expanding thе research to include diverse locations could provide morе comprehensive insights.

# Reference list (not included in the work count)

* Choudhary, (2021). Zomato Restaurants Data Analysis Using Machine Learning Algorithms.
* Gupta, R., Sameer, S., Muppavarapu, H., Enduri, M.K. and Anamalamudi, S. (2021). Sentiment Analysis on Zomato Reviews. [online] IEEE Xplore. doi:https://doi.org/10.1109/CICN51697.2021.9574641.
* Saxena, A. (2021). Zomato Data Analysis. International Journal for Research in Applied Science and Engineering Technology, 9(12), pp.644–649. doi:https://doi.org/10.22214/ijraset.2021.39303.
* Vu (2021). Service Quality And Its Impact On Customer Satisfaction.

# Appendices

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

# Load required libraries

library(ggplot2)

library(dplyr)

# Load the dataset (replace with the correct path to your file)

dataset <- read.csv("C:/Users/Admin/Documents/ZOMARTO\_CAFE/reviews.csv")

# Check if the file exists (optional step)

if (!file.exists("C:/Users/Admin/Documents/ZOMARTO\_CAFE/reviews.csv")) {

stop("File does not exist!")

}

# View the first few rows of the dataset

head(dataset)

# Convert the dependent variable 'Overall\_Rating' to numeric if not already

dataset$Overall\_Rating <- as.numeric(as.character(dataset$Overall\_Rating))

# Remove rows with NA in 'Overall\_Rating'

dataset <- na.omit(dataset)

# Check the structure of the 'Overall\_Rating' column

str(dataset$Overall\_Rating)

# Summary statistics for 'Overall\_Rating'

summary(dataset$Overall\_Rating)

# Test for normality using the Shapiro-Wilk test

shapiro\_test <- shapiro.test(dataset$Overall\_Rating)

print(shapiro\_test)

# Plot a histogram with a normal curve overlay

plot <- ggplot(data.frame(Overall\_Rating = dataset$Overall\_Rating), aes(x = Overall\_Rating)) +

geom\_histogram(aes(y = ..density..), bins = 15, color = "black", fill = "skyblue") +

stat\_function(fun = dnorm, args = list(mean = mean(dataset$Overall\_Rating, na.rm = TRUE),

sd = sd(dataset$Overall\_Rating, na.rm = TRUE)),

color = "red", linewidth = 1) +

labs(title = "Histogram of Overall Rating with Normal Curve Overlay",

x = "Overall Rating",

y = "Density") +

theme\_minimal()

# Display the plot

print(plot)

# Save the plot as a PNG file

setwd("C:/Users/Admin/Documents/ZOMARTO\_CAFE") # Set working directory

ggsave("normality\_histogram.png", plot = plot, width = 8, height = 6, dpi = 300)

**Logs:**

> # Load required libraries

> library(ggplot2)

> library(dplyr)

>

> # Load the dataset (replace with the correct path to your file)

> dataset <- read.csv("C:/Users/Admin/Documents/ZOMARTO\_CAFE/reviews.csv")

>

> # Check if the file exists (optional step)

> if (!file.exists("C:/Users/Admin/Documents/ZOMARTO\_CAFE/reviews.csv")) {

+ stop("File does not exist!")

+ }

>

> # View the first few rows of the dataset

> head(dataset)

Index Name Overall\_Rating

1 0 Oliver Brown 3.9

2 1 Oliver Brown 3.9

3 2 Crush Coffee 3

4 3 The Mohalla 3.8

5 4 The Mohalla 3.8

6 5 The Shaka Cafe 4.3

Cuisine

1 Cafe, Coffee, Shake, Juices, Beverages, Waffle, Desserts, Ice Cream

2 Cafe, Coffee, Shake, Juices, Beverages, Waffle, Desserts, Ice Cream

3 Cafe, Shake, Beverages, Desserts

4 Cafe

5 Cafe

6 North Indian, Pizza, Asian, Chinese, Cafe, Desserts, Beverages

Rate.for.two City

1 500 ahmedabad

2 500 ahmedabad

3 600 ahmedabad

4 550 ahmedabad

5 550 ahmedabad

6 1,200 ahmedabad

Review

1 Been to this place 3-4 times. Prakash is always very sweet and accommodating. Plus always ask Vaibhav to make your shakes, they are perfect!!

2 I recently visited Oliver Brown on a weekend for a cup of iced coffee, and I was thoroughly impressed. The ambience was inviting, the seating comfortable, and the service top-notch. I had the pleasure of trying their iced coffee and dark hot chocolate, both of which were absolutely delicious. It's an excellent place overall, and I'm already looking forward to my next visit

3 Very watery ans thin shake

4 it was not cheese burst pizza.. only cheeze was on the top.. not happy with the pizza sent made today

5 Yammi.,....test burger is best I love 💗 this BergerAnd fry is also missing 😂Love ❤️ you ........

6 food was not spicy , which I was told to restorant on call and instructions.

>

> # Convert the dependent variable 'Overall\_Rating' to numeric if not already

> dataset$Overall\_Rating <- as.numeric(as.character(dataset$Overall\_Rating))

Warning message:

NAs introduced by coercion

>

> # Remove rows with NA in 'Overall\_Rating'

> dataset <- na.omit(dataset)

>

> # Check the structure of the 'Overall\_Rating' column

> str(dataset$Overall\_Rating)

num [1:658] 3.9 3.9 3 3.8 3.8 4.3 4.3 3.7 4.8 4.1 ...

>

> # Summary statistics for 'Overall\_Rating'

> summary(dataset$Overall\_Rating)

Min. 1st Qu. Median Mean 3rd Qu. Max.

2.500 3.500 3.900 3.779 4.100 4.900

>

> # Test for normality using the Shapiro-Wilk test

> shapiro\_test <- shapiro.test(dataset$Overall\_Rating)

> print(shapiro\_test)

Shapiro-Wilk normality test

data: dataset$Overall\_Rating

W = 0.96957, p-value = 1.886e-10

>

> # Plot a histogram with a normal curve overlay

> plot <- ggplot(data.frame(Overall\_Rating = dataset$Overall\_Rating), aes(x = Overall\_Rating)) +

+ geom\_histogram(aes(y = ..density..), bins = 15, color = "black", fill = "skyblue") +

+ stat\_function(fun = dnorm, args = list(mean = mean(dataset$Overall\_Rating, na.rm = TRUE),

+ sd = sd(dataset$Overall\_Rating, na.rm = TRUE)),

+ color = "red", linewidth = 1) +

+ labs(title = "Histogram of Overall Rating with Normal Curve Overlay",

+ x = "Overall Rating",

+ y = "Density") +

+ theme\_minimal()

>

> # Display the plot

> print(plot)

>

> # Save the plot as a PNG file

> setwd("C:/Users/Admin/Documents/ZOMARTO\_CAFE") # Set working directory

> ggsave("normality\_histogram.png", plot = plot, width = 8, height = 6, dpi = 300)

## GitHub log output.

