PIMPRI CHINCHWAD EDUCATION TRUST'S

PIMPRI CHINCHWAD COLLEGE OF ENGINEERING

Department of Computer Engineering

Report of Mini Project

- Subject: Data Exploration and Visualization (DEVL)
- Academic Year: 2024-25
- Semester III
- Submitted by :-

1. Rohan Adkine(123B1B076)
2. Piyush Ahirao (123B1B077)
3. Disha Andre (123B1B078)
4. Anish Pote (123B1B079)

- Submitted to :- Mrs. Harsh Talele.
- Date:- 23-10-2024
- Sign:-

Project Title: Data Analysis of Restaurants

- **DataSet used:**
  Data of Zomato RestaurantAnalysis.
  Link :- https://www.kaggle.com/code/payamamanat/zomato-restaurant-analysis

- **About this dataset:** - This dataset contains information about various restaurants listed on the Zomato platform, including details such as restaurant names, ratings, types of cuisines served, and services offered (like online ordering and table booking). Researchers and analysts can use this dataset to track restaurant performance, analyze customer preferences, and study trends in the restaurant industry, especially in terms of online ordering and table bookings.

- **Details of the data:**
  The dataset contains **7,105 observations** and **12 variables** as follows:

Sr. No.	Column Name	Description
1	`restaurant name`	The name of the restaurant. (String)
2	`restaurant type`	The type of restaurant (e.g., Quick Bites, Casual Dining). (String)
3	`rate (out of 5)`	The average rating of the restaurant, on a scale of 1 to 5. (Float)
4	`num of ratings`	The number of ratings given to the restaurant. (Integer)
5	`avg cost (two people)`	The average cost for two people dining at the restaurant. (Float)
6	`online_order`	Whether the restaurant accepts online orders. (`Yes` / `No`) (String)
7	`table booking`	Whether the restaurant allows table bookings. (`Yes` / `No`) (String)
8	`cuisines type`	The types of cuisines offered at the restaurant (e.g., Italian, Chinese). (String)
9	`area`	The area or neighborhood where the restaurant is located. (String)
10	`local address`	The detailed address of the restaurant. (String)
11	`Unnamed: 0`	An unnecessary column, possibly an index. (Integer)
12	`Unnamed: 0.1`	Another unnecessary column, likely another index. (Integer)

- **Introduction :-** This report provides an in-depth analysis of a the above dataset, which contains various attributes such as restaurant types, ratings, average costs for two people, and geographical information like areas and addresses. The analysis includes data cleaning, visualization, and insights into the distribution of restaurants, their costs, and customer ratings. The goal is to extract actionable insights to understand the restaurant landscape better.

- **Data Loading and Preparation :-**

1. **Loading Necessary Libraries :-** To manipulate, clean, and visualize the data, we start by loading the necessary Python libraries. These include:
   Syntax :-

```
# Import necessary libraries
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error
```

2.Loading the Data :- The dataset is loaded using the read.csv() function. It is important to specify the correct file path where the dataset is stored, ensuring the data is accessible.

Syntax :- df = pd.read_csv("C:/Users/hp/Downloads/zomato.csv")

3.Checking Column Names :- We examine the column names to understand the structure of the dataset and identify which variables are available for analysis. This step is essential for knowing the type of data we have and determining whether any column needs further cleaning or renaming.

Syntax :-

```
# Check column names
print("Column Names:")
print(df.columns)
```

4.Dropping Unnecessary Columns :- Certain columns might not add value to the analysis. For example, columns that are repeated, unnamed, or hold redundant information can be removed to streamline the dataset.

Syntax :-

```
# Drop unnecessary columns if they exist
columns_to_drop = ["Unnamed: 0.1", "Unnamed: 0"]
df.drop(columns=[col for col in columns_to_drop if col in df.columns], inplace=True)
```

5.Quick Data Scan :- The info() function is used to get a concise view of the data types and contents of each column, while describe() provides a quick statistical overview of the numerical columns. This helps identify data ranges, missing values, and potential outliers.

Syntax :-

```
# Quick scan of dataset
print("Dataset Overview:")
print(df.info())
print(df.describe())
```

6.Handling Missing Values :- Missing values are a common issue in datasets. For this dataset, specific columns like ratings and average costs may have missing data. We fill the missing values with the column mean, which prevents skewing the analysis.

Syntax :-

```
# Number of missing values per column
missing_values = df.isna().sum()
print("Missing Values per Column:")
print(missing_values)
```

```
# Fill missing values in 'rate (out of 5)' and 'avg cost (two people)' columns with mean
for col in ["rate (out.of.5)", "avg cost (two.people)"]:
    if col in df.columns:
        df[col].fillna(df[col].mean(), inplace=True)
```

7.Checking for Duplicates :- Duplicate rows can affect the integrity of the analysis. We check for duplicates and, if any are found, they can be removed. This ensures each restaurant is represented only once.

Syntax :-

```
# Number of duplicates
duplicate_count = df.duplicated().sum()
print(f"Number of duplicate rows: {duplicate_count}")
```

8.Summary of Counts :- We summarize key metrics of the dataset, including the total number of restaurants, the unique number of restaurant types, and the distribution across local addresses and areas. This provides a snapshot of the dataset's scope and variety.

```
# Summary of counts
summary_count = {
    'total_rows': len(df),
    'total_restaurant_types': df['restaurant type'].nunique() if 'restaurant type' in df.columns else None,
    'total_unique_addresses': df['local address'].nunique() if 'local address' in df.columns else None,
    'total_unique_areas': df['area'].nunique() if 'area' in df.columns else None,
    'total_unique_cuisines': df['cuisines type'].nunique() if 'cuisines type' in df.columns else None,
}
print("Summary of Counts:")
print(summary_count)
```

```
Summary of Counts:
{'total_rows': 7105, 'total_restaurant_types': 81, 'total_unique_addresses': 90, 'total_unique_areas': 30, 'total_unique_cuisines': 2175}
```

➢ Exploratory Data Analysis (EDA)

1. Table Booking Relation with Online Order

```
Unique values in 'table booking': ['No' 'Yes']

Counts with Percentages for Table Booking:
  table_booking  count  percentage
0            No   6361   89.528501
1           Yes    744   10.471499
```

- **Insights :-**
  - Partnering with delivery apps allows hotels to focus on enhancing kitchen capacity rather than expanding physical space.
  - This approach enables efficient food service while minimizing overhead costs. By prioritizing delivery convenience, hotels can boost customer satisfaction and loyalty.
  - Adapting to the growing demand for food delivery positions hotels competitively in the market.

2. Recommendation for new Restaurant Location

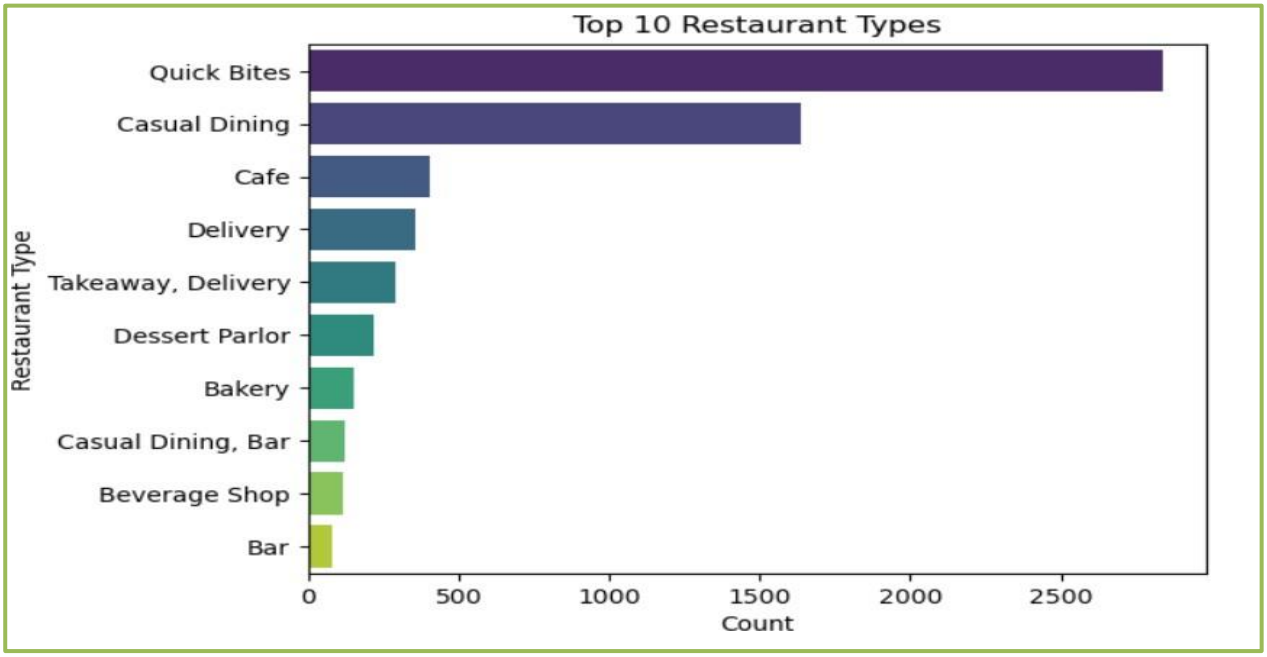
```
Cuisines with Maximum Frequency:
Cuisine Type: North Indian, Frequency: 3237
```

```
Area with the Least Number of Restaurants:
Area: Residency Road, Restaurant Count: 46
```

- **Insights :-**
  - **Demand**: North Indian cuisine has the highest ratings, indicating strong customer interest and loyalty.
  - **Location Advantage**: Residency Road likely has fewer North Indian restaurants, reducing competition and increasing your potential customer base.
  - **Growth Opportunity**: Opening a restaurant here could attract diners seeking North Indian options, leading to quicker brand recognition.

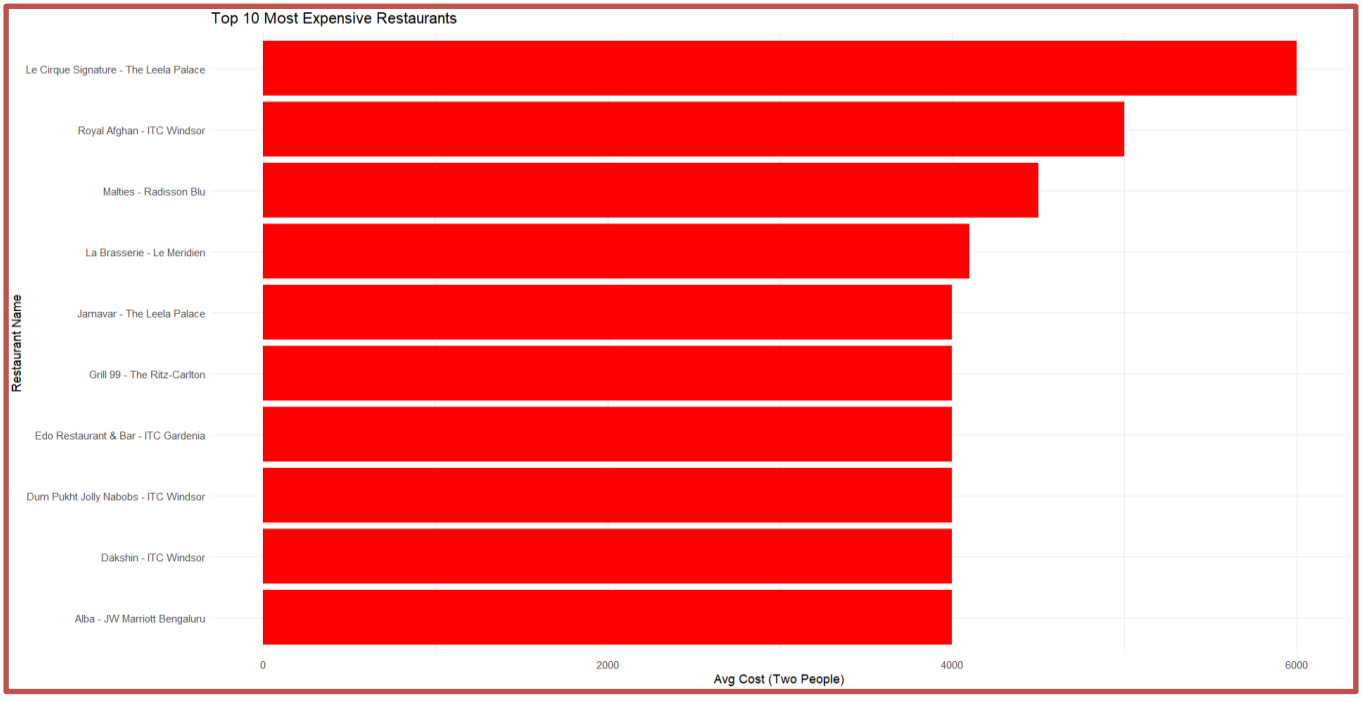
➢ Visualizations and Insights

1. Bar Plot: Top 10 most common restaurant types



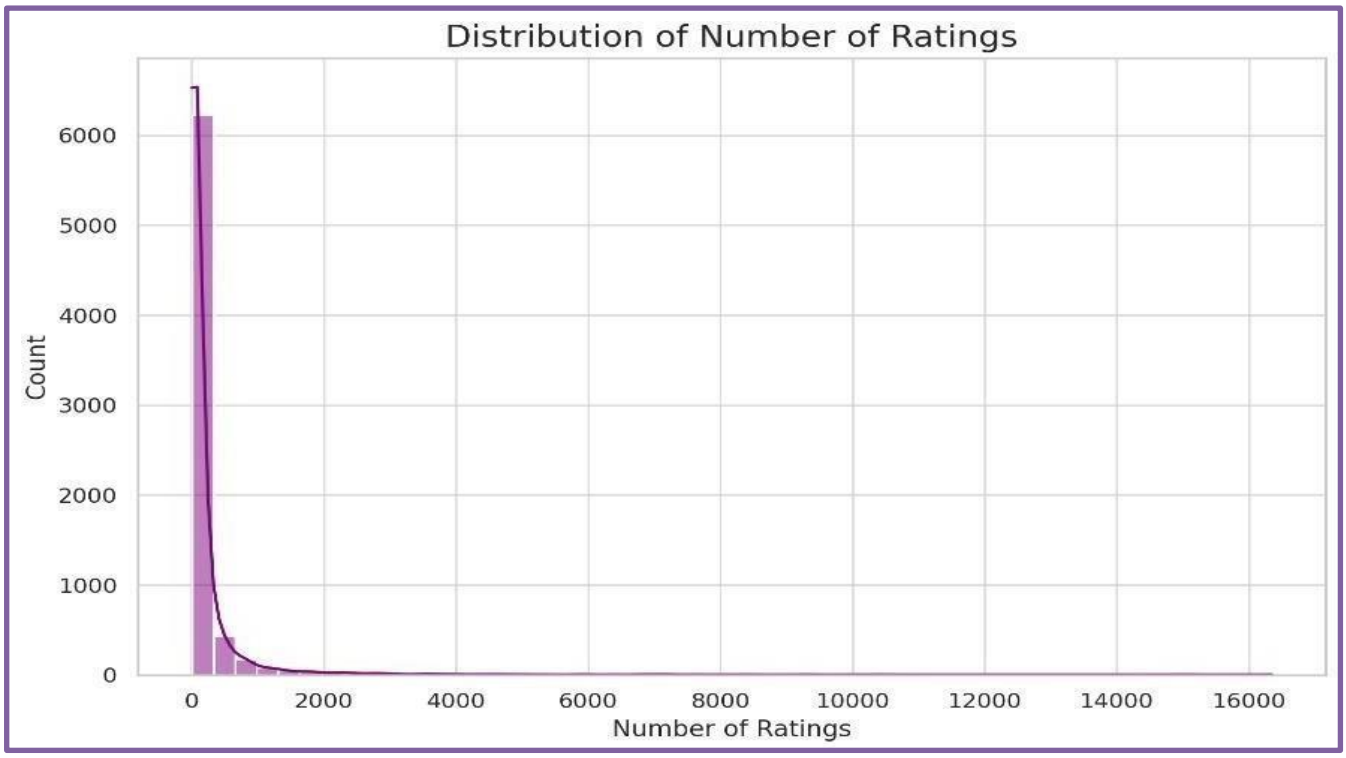
- **Insight:** The bar plot reveals the ten most popular restaurant types. For example, cafes, casual dining, and quick bites are likely to appear at the top, showing that these dining formats are favored in the analyzed area.

2. Bar Plot: Top 10 most expensive restaurants



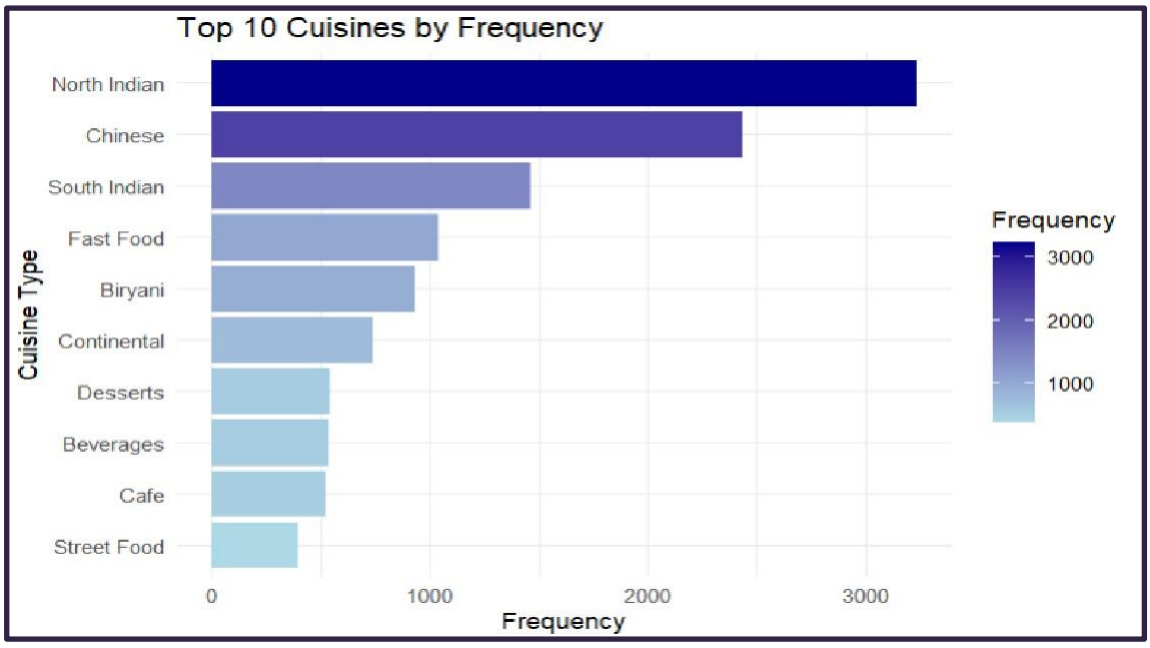
- **Insight:** This visualization identifies the most expensive restaurants, helping consumers and businesses understand which establishments cater to higher-end dining preferences.

3. Histogram Plot : Distribution of Number of Ratings



- **Insights :** The majority of restaurants have a smaller number of ratings, with only a few having many reviews. The distribution is right-skewed.

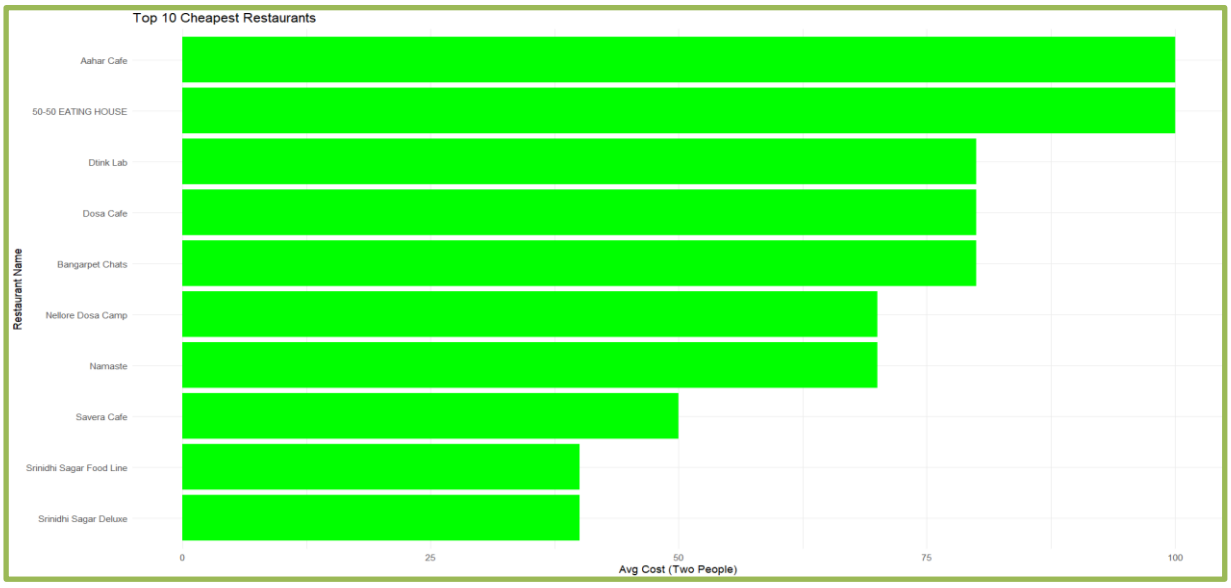
4. Box Plot : Top Cuisines by Frequency



- **Insights :** The most popular cuisine types are displayed. This can provide insights into customer preferences.

5 . Bar Plot: Top 10 most cheapest restaurants

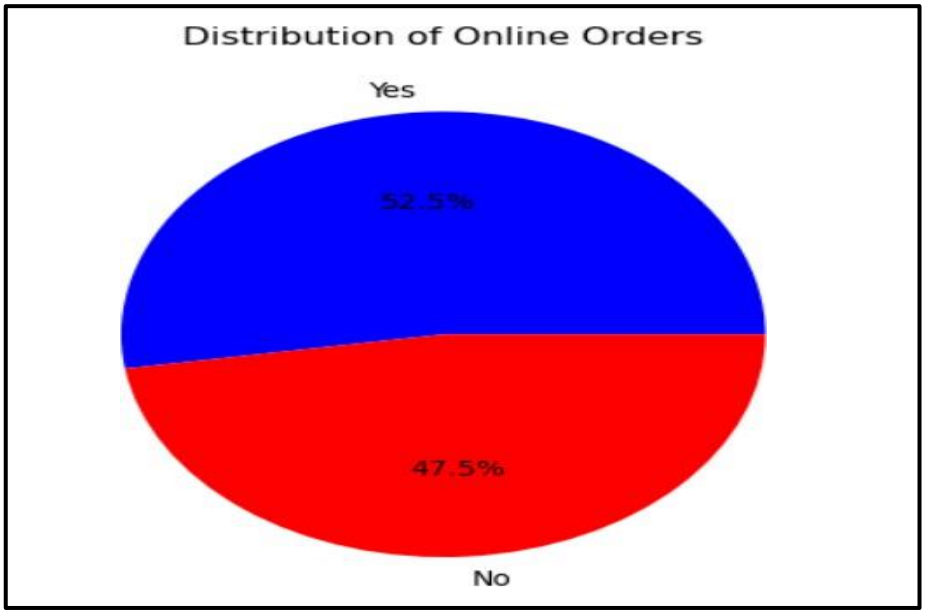
- This plot showcases the ten cheapest restaurants. It is especially helpful for budget-conscious customers or for comparing different pricing strategies.



- **Insight:** The cheapest restaurants are displayed here, indicating where customers can dine for a lower cost. This plot also helps business owners understand the pricing landscape and competition.

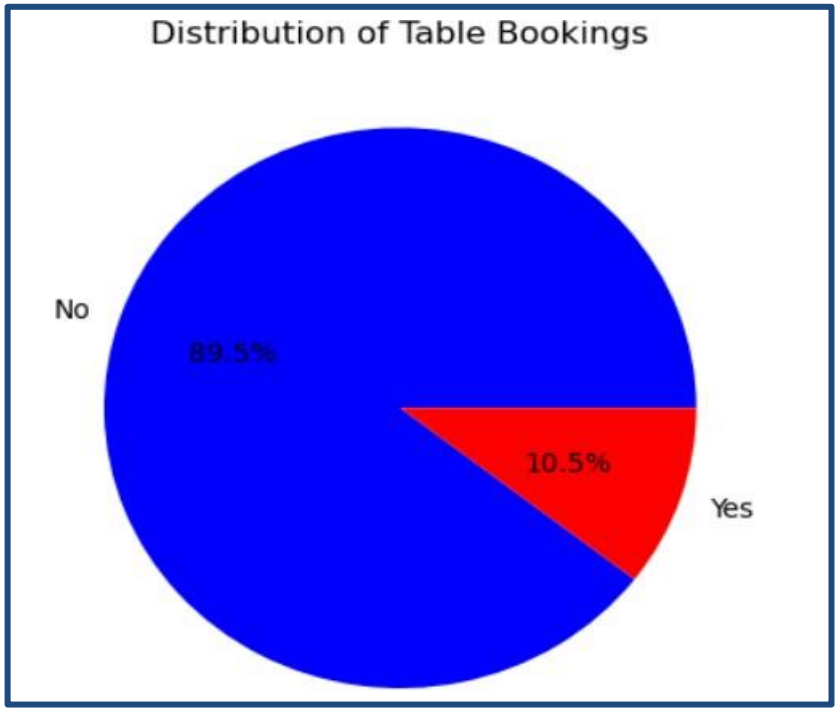
6. Pie Plot: Online Order Distribution

- This pie plot shows how many restaurants offer online ordering services. In today's digital world, online ordering has become an important feature for customer convenience.



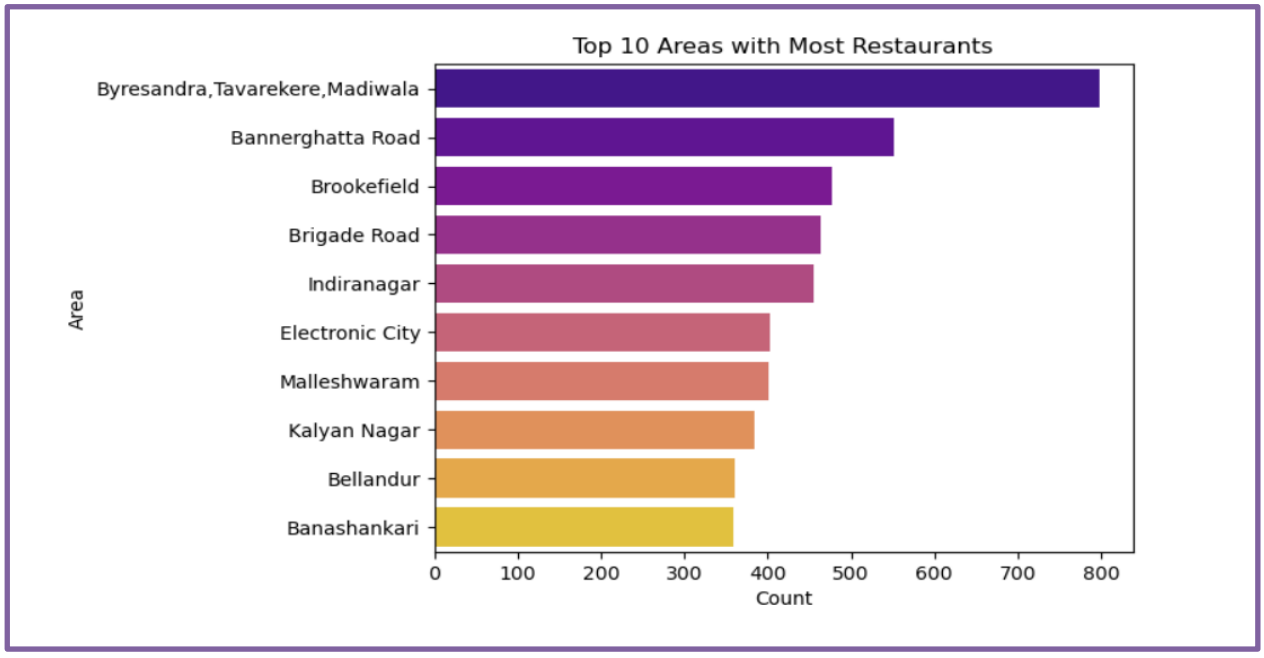
- **Insight:** A high proportion of restaurants offering online ordering highlights the growing trend of digital convenience. If the proportion is low, this could indicate an area of opportunity for restaurants to expand their service options.

7. Pie Chart : Table Booking



- **Insight :** Around 89.5% of Zomato users placed online orders without booking a table, showing a strong preference for delivery or pickup. Only 10.5% of users opted for table reservations, indicating limited usage of this feature.

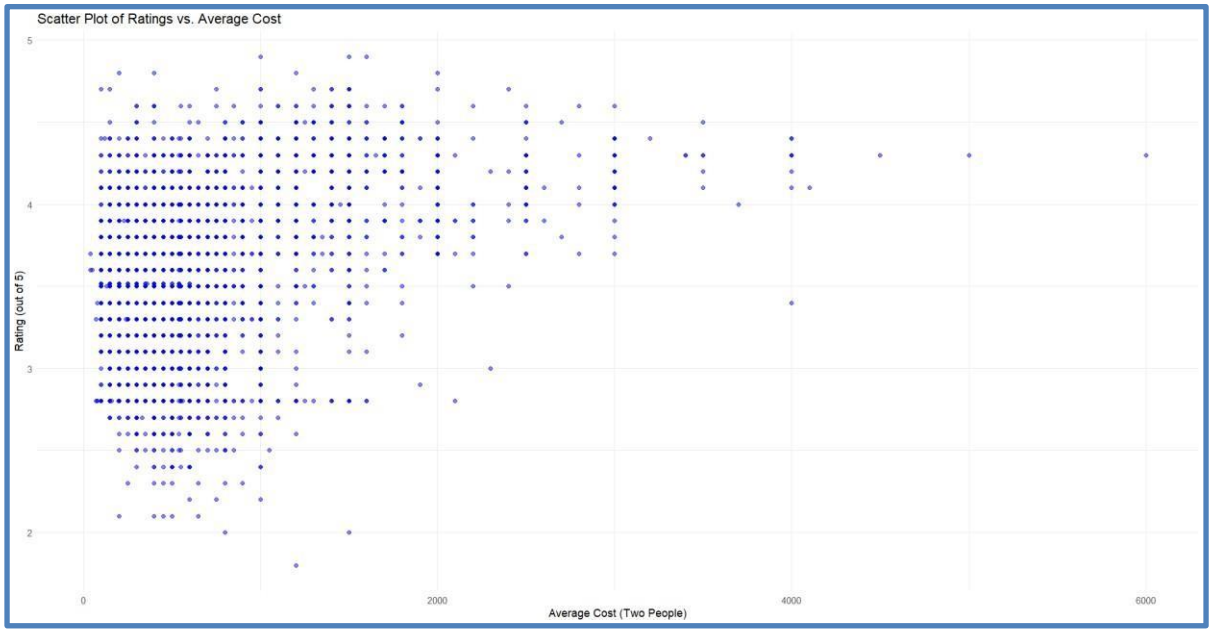
8. Bar Plot : Top 10 Areas with Most Restaurants



- **Insight:** This bar plot reveals which areas have the largest number of restaurants, suggesting potential zones of high demand. Business owners may use this information to consider expansion into less saturated areas, while customers may use it to find areas with more dining options.

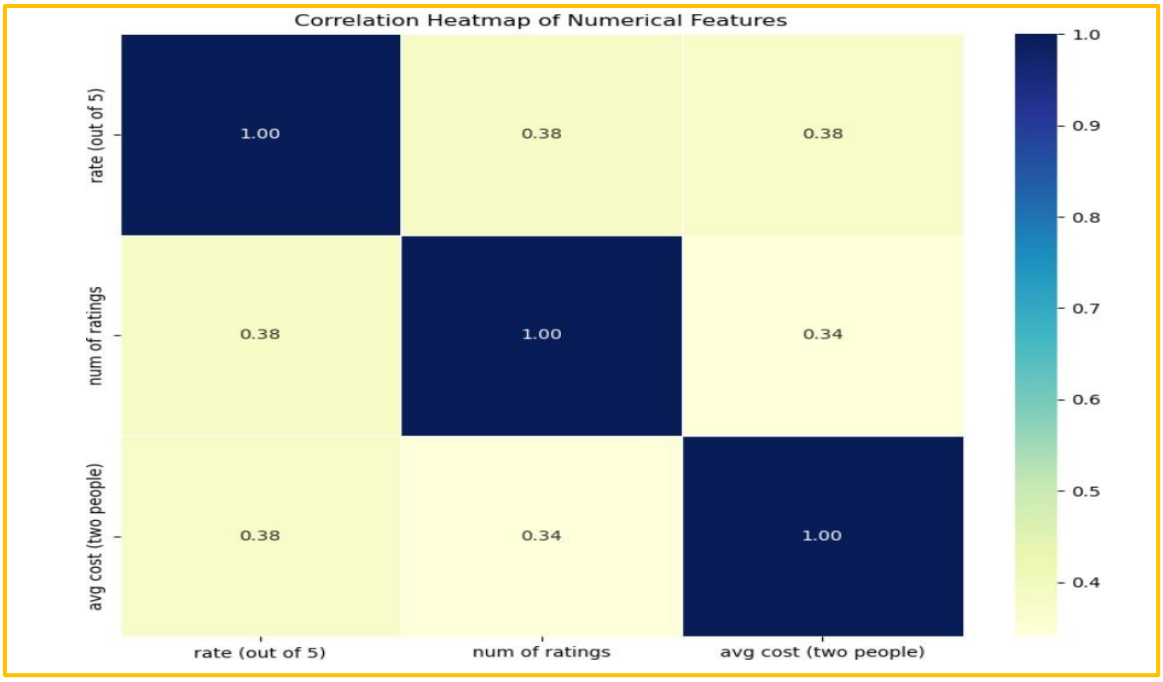
9 . Scatter Plot: Ratings vs. Average Cost

- This scatter plot explores the relationship between restaurant ratings and the average cost for two people. The aim is to see if there is a correlation between the cost of a meal and customer satisfaction.

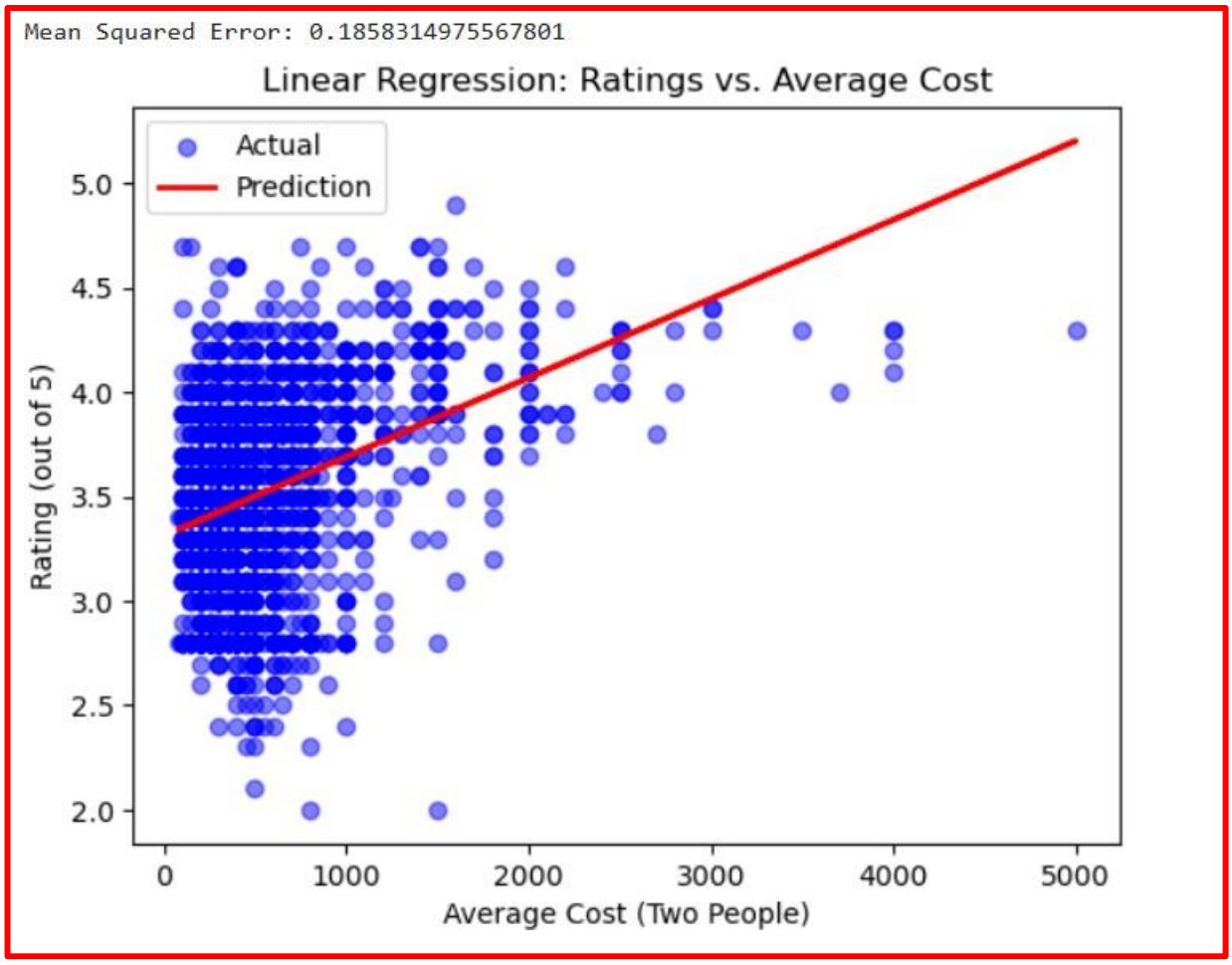


- **Insight:** The scatter plot provides an overview of whether high-cost restaurants tend to have higher ratings, suggesting a potential relationship between price and quality, or whether lower-cost restaurants can also achieve high ratings.

10. Heatmap of Numeric Columns in Zomato Dataset (Correlation)



➢ Linear Regression: Predict ‘rating based on average cost



▪ Insights :-

1. **Weak Linear Relationship**: The regression line shows a slight upward trend, but the broad spread in actual ratings suggests cost alone doesn’t predict ratings well.
2. **Clustered Data at Lower Costs**: Most data points are in the lower cost range (under $1,000), with ratings varying widely, implying that other factors are more influential at this level.
3. **Outliers at High Costs**: A few high-cost outliers show higher ratings, but the model overestimates in this range, likely due to limited data points.
4. **Model Performance (MSE)**: The low Mean Squared Error indicates small prediction errors, but the high variance in actual ratings limits the model's practical predictive power.

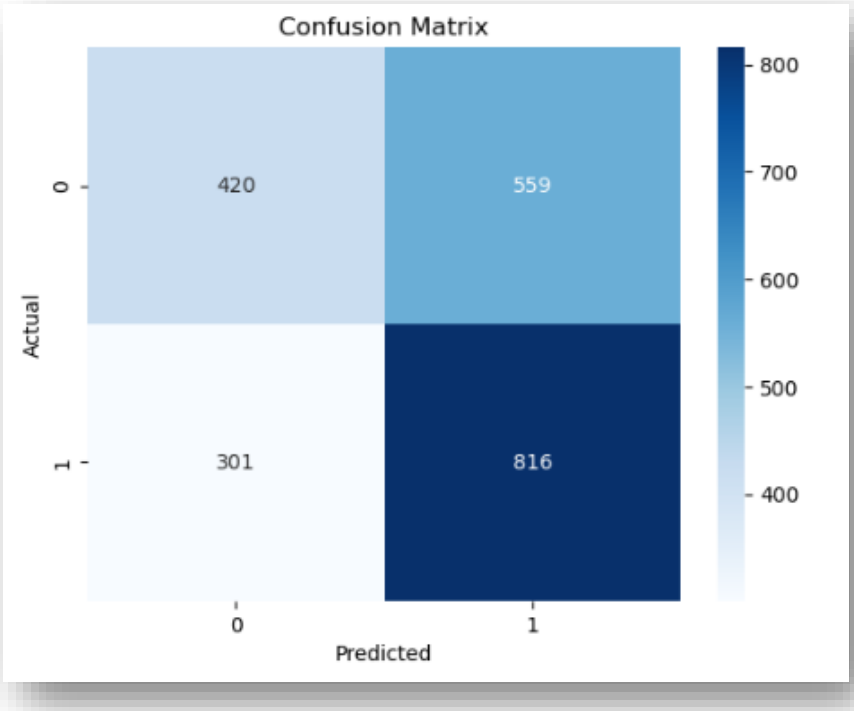
➢ Logistic Regression

1. Model Building
A **Logistic Regression** model was chosen for this binary classification task, as it is efficient for predicting categorical outcomes.

2. Model Evaluation
The model was evaluated on the test set using the following metrics:

•**Accuracy**: 58.97%

- **Confusion Matrix**:



1. **True Negatives (420)**: Correctly predicted restaurants without online ordering.
2. **True Positives (816)**: Correctly predicted restaurants with online ordering.
3. **False Positives (559)**: Predicted as having online ordering but don't.
4. **False Negatives (301)**: Predicted as not having online ordering but do.

- **Classification Report**:

Metric	Precision	Recall	F1-Score
Class 0	0.58	0.43	0.49
Class 1	0.59	0.73	0.65
Macro Avg	0.59	0.58	0.57
Weighted Avg	0.59	0.59	0.58

➢ Conclusion

The analysis of this restaurant dataset reveals valuable insights into several aspects of the dining landscape. Key findings include:

1.Common Restaurant Types: Cafes, casual dining, and quick bites dominate the dining scene, reflecting popular customer preferences.

2.Restaurant Ratings: The majority of restaurants are well-rated, with a noticeable cluster around higher rating values, suggesting good customer experiences.

3.Cost Insights: There is a wide range of restaurant prices, from affordable options to more expensive, premium establishments, catering to different customer segments.

4.Geographical Distribution: Certain areas and addresses have a much higher concentration of restaurants, indicating zones of high competition. However, this also presents opportunities for expansion into less saturated areas.

5.Online Ordering: Many restaurants offer online ordering, reflecting the growing digital convenience trend in the restaurant industry.

- Partnering with delivery apps allows hotels to focus on enhancing kitchen capacity rather than expanding physical space.
- This approach enables efficient food service while minimizing overhead costs. By prioritizing delivery convenience, hotels can boost customer satisfaction and loyalty.
- Adapting to the growing demand for food delivery positions hotels competitively in the market.

6. Logistic Regression

•**Accuracy (58.97%)**: The model correctly predicts the online ordering status for around 59% of restaurants.

•**Precision (Class 1: 59%)**: When the model predicts a restaurant offers online ordering, it is correct 59% of the time.

•**Recall (Class 1: 73%)**: The model identifies 73% of restaurants that actually offer online ordering.

•**F1-Score (Class 1: 65%)**: Balances precision and recall, indicating moderate performance in predicting online ordering availability.

➢ Model Performance Analysis: Online Ordering Prediction

The model's performance in predicting whether restaurants offer online ordering can be summarized as follows:

1.High Recall:

1. The model demonstrates **high recall**, successfully identifying a large proportion of restaurants that offer online ordering. This means that most of the restaurants with online ordering capabilities are accurately captured by the model.

2.Moderate Precision:

Despite the high recall, the model's **precision** is moderate. This indicates that it occasionally incorrectly classifies restaurants that do not offer online ordering as those that do, leading to some false positives. Therefore, while the model is effective at identifying online-ordering restaurants, there is room for improvement in reducing misclassifications.

3.Influential Features: The model appears to utilize several key features in its predictions, including:

1. **Restaurant Type**: Different types of restaurants (e.g., fast food, casual dining) may have different propensities to offer online ordering.
2. **Ratings**: Restaurants with higher customer ratings are more likely to offer online ordering, which the model may use as a signal.
3. **Geographic Area**: The model may consider the location or region of the restaurant, as certain areas might have more restaurants with online ordering capabilities.

4.Baseline Performance and Areas for Improvement:

Overall, the model serves as a **solid baseline** for predicting the availability of online ordering. However, **precision improvement** is necessary to reduce false positives and enhance overall reliability. Further model tuning, feature engineering, or the inclusion of additional data sources could contribute to more accurate and consistent predictions.