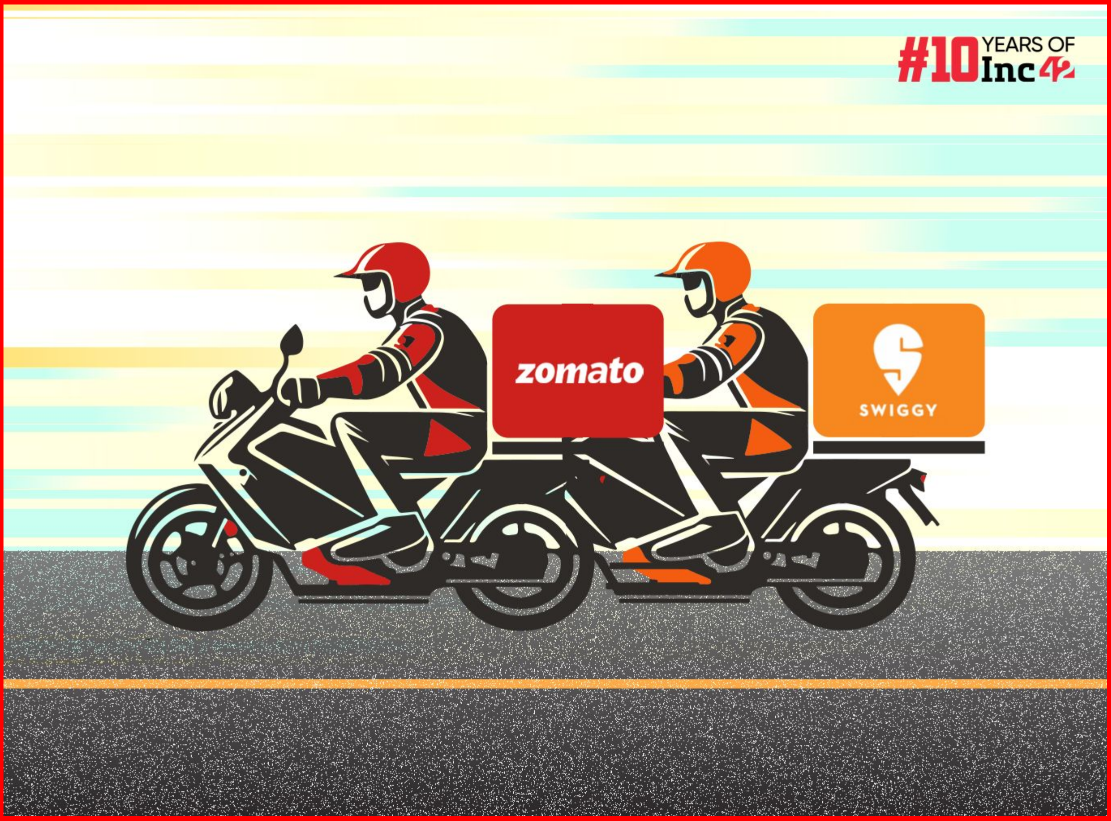
ZOMATO



EXPLORATORY DATA ANALYSIS CAPSTONE PROJECT

INTRODUCTION

In your code, the phrase "given introduction" likely refers to the initial markdown cell or section of your notebook. It's where you introduce the topic or the goals of your analysis. The markdown content you provided discusses the insights and findings from your data analysis on Indian Restaurants.

The introduction sets the stage for the code that follows. It helps the reader understand the context of the analysis and the key takeaways from the findings. It's a crucial part of documenting your work and making your notebook easy to understand.

- An introduction in any document or project helps guide the reader through your work. It helps with understanding context, setting expectations, and highlighting key points.
- In the context of a Jupyter Notebook or Google Colab, the introduction is often written using Markdown cells to clearly present the project's purpose and findings.
- The user can click Runtime and then Run all or Restart and run all in the menu bar to see the code output.

OBJECTIVE

1. Data Understanding and Cleaning:

- To import, explore, and clean the Indian Restaurants dataset, handling missing values and duplicates. This includes understanding the data's structure, identifying potential issues, and preparing it for further analysis.

2.Restaurants Rating Analysis:

- To analyze the distribution of restaurant ratings, calculate the average rating, and identify factors that influence ratings, such as cuisine, price, and features like online ordering and Wi-Fi.

3 .City-Specific Insights:

- To identify cities with the highest concentration of restaurants, analyze the distribution of restaurant ratings across cities, and uncover city-specific trends or outliers.

OBJECTIVE

Restaurant Cuisines

To determine the most popular cuisines, investigate the relationship between cuisine variety and restaurant ratings, and analyze the relationship between price range and ratings.

5.Feature Impact Analysis

- To identify and visualize the top restaurant chains based on the number of outlets and explore the ratings of these top chains.

6.Top Restaurant Chain Identification

- To investigate the impact of features like online order availability, table booking, alcohol availability, and Wi-Fi on restaurant ratings and customer preferences.

IMPORT PYTHON LYBRARY

NumPy

NumPy (Numerical Python) is a powerful library for working with arrays and numerical data.

It is used for mathematical operations, statistical analysis, and handling large datasets efficiently.

Pandas

Pandas is a Python library used for data analysis and manipulation.

It is mainly used for working with structured data like tables, spreadsheets, or databases.

It provides DataFrames, which are like Excel tables but with much more flexibility and power.

Seaborn

Seaborn is a data visualization library used to create beautiful and professional charts.

It allows you to create Bar Charts, Line Charts, Scatter Plots, Box Plots, and more.

Seaborn makes data representation easier to understand.

Matplotlib

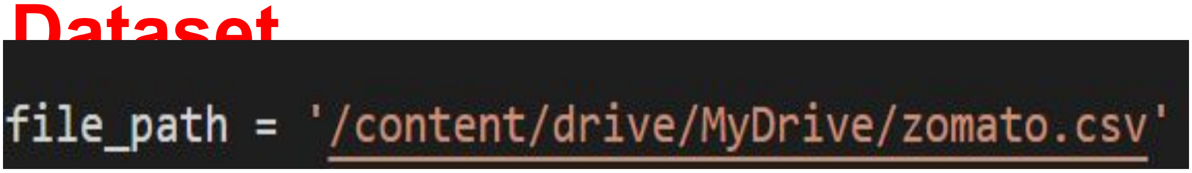
Matplotlib is another charting library but gives more customization options.

It is often used with Seaborn to modify or save graphs.

Seaborn internally uses Matplotlib to generate graphs.

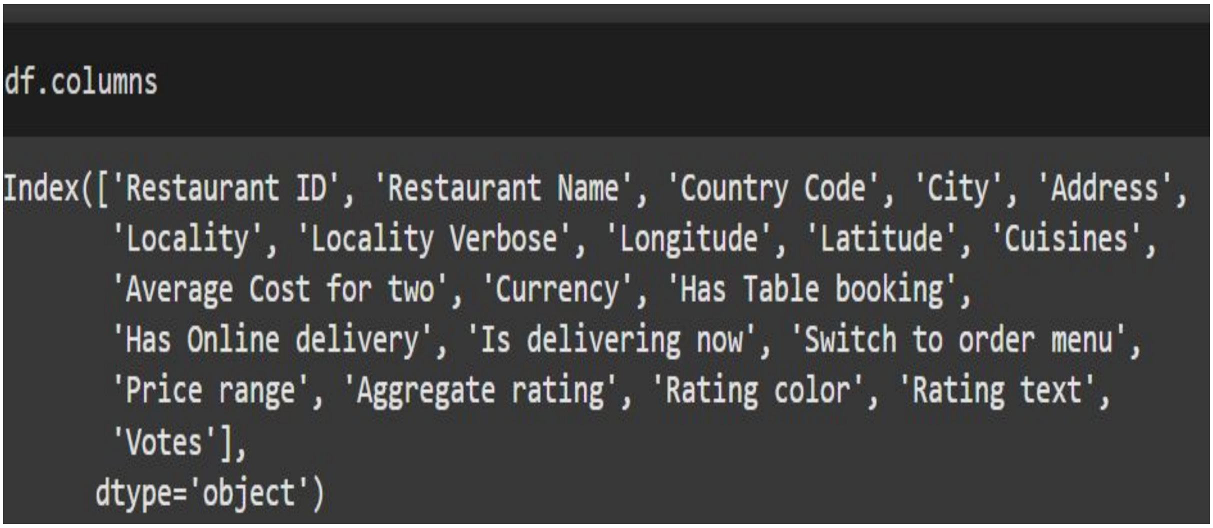
LOADING DATASET

Loading Dataset



```
file_path = '/content/drive/MyDrive/zomato.csv'
```

I've load dataset named ‘Zomato.csv’ into a variable name ‘df’ in order to import the datasets, I’ve used pandas Libraries in which I’ve used ‘pd.read_csv’ command to import the respective datasets.



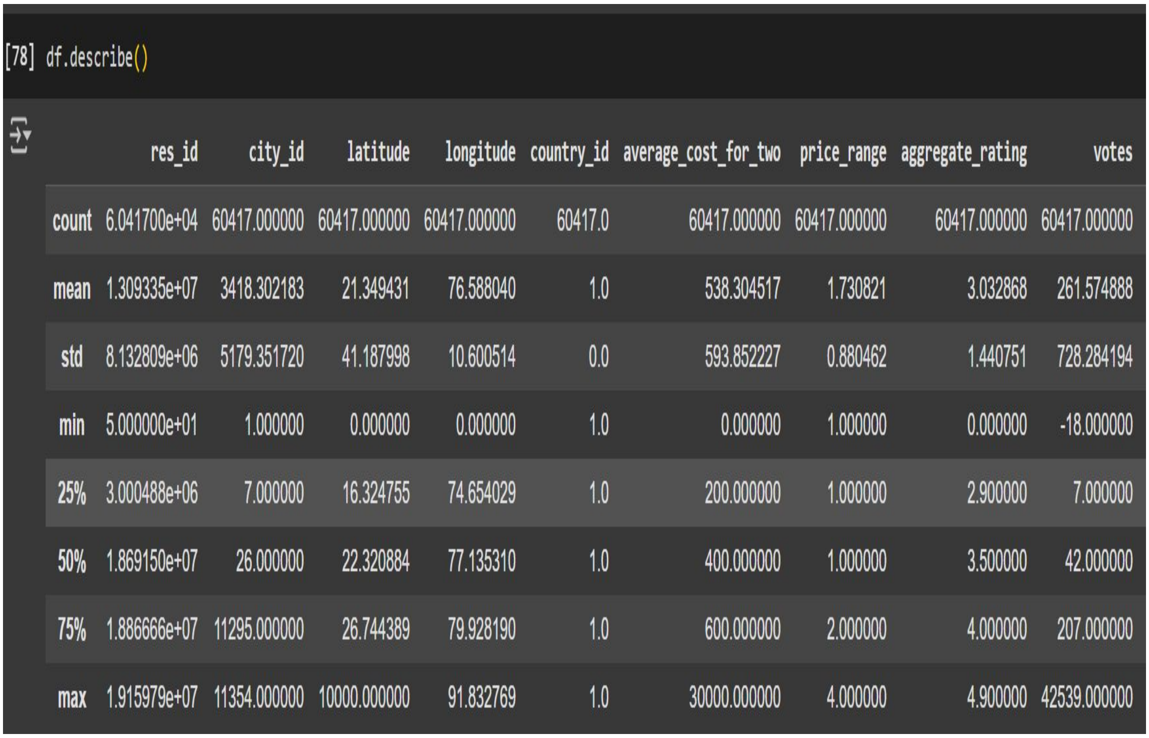
```
df.columns

Index(['Restaurant ID', 'Restaurant Name', 'Country Code', 'City', 'Address',
       'Locality', 'Locality Verbose', 'Longitude', 'Latitude', 'Cuisines',
       'Average Cost for two', 'Currency', 'Has Table booking',
       'Has Online delivery', 'Is delivering now', 'Switch to order menu',
       'Price range', 'Aggregate rating', 'Rating color', 'Rating text',
       'Votes'],
      dtype='object')
```

I’ve accessed and displayed the column names of the laoded dataset using columns. These columns encompass various aspects of ‘Indian restaurants’ information, providing detailed overview of the dataset.

DATA OVERVIEW

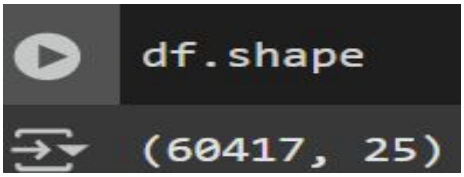
Basic Composition of data



```
[78] df.describe()
```

```
          res_id       city_id      latitude     longitude  country_id  average_cost_for_two   price_range  aggregate_rating         votes
count  6.041700e+04  60417.000000  60417.000000  60417.000000     60417.0          60417.000000  60417.000000      60417.000000  60417.000000
mean   1.309335e+07   3418.302183     21.349431     76.588040         1.0            538.304517      1.730821          3.032868    261.574888
std    8.132809e+06   5179.351720     41.187998     10.600514         0.0            593.852227      0.880462          1.440751    728.284194
min    5.000000e+01      1.000000      0.000000      0.000000         1.0              0.000000      1.000000          0.000000    -18.000000
25%    3.000488e+06      7.000000     16.324755     74.654029         1.0            200.000000      1.000000          2.900000      7.000000
50%    1.869150e+07     26.000000     22.320884     77.135310         1.0            400.000000      1.000000          3.500000     42.000000
75%    1.886666e+07  11295.000000     26.744389     79.928190         1.0            600.000000      2.000000          4.000000    207.000000
max    1.915979e+07  11354.000000  10000.000000     91.832769         1.0          30000.000000      4.000000          4.900000  42539.000000
```

I've generated descriptive stastic for the loaded dataset stored in the variable 'df' using the describe method. This Pandas function provides a summary of stastical measure , minimum , 25th percentile, median(50th percentile), 75th percentile and maximum values for each numeric column in the dataset . This summary aids in in understanding the central tendency ,dispersion , and distribution of the numeric feature in the dataset.



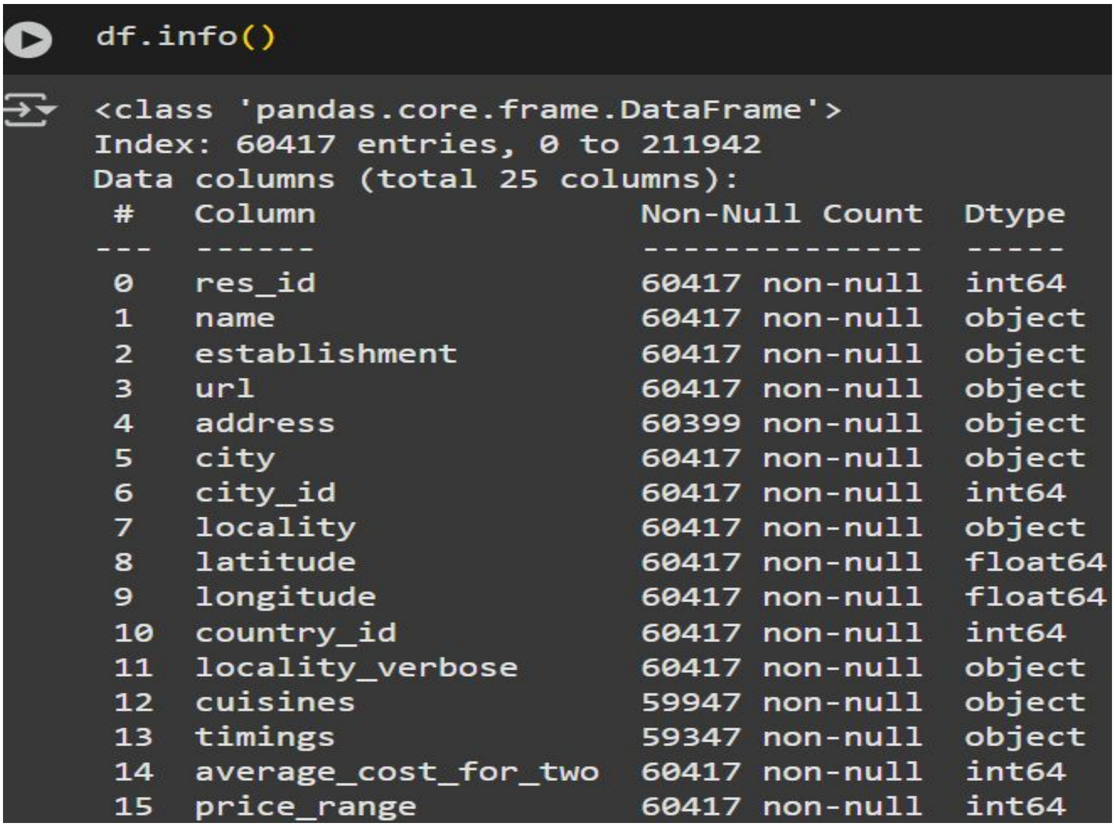
```
df.shape
```

```
(60417, 25)
```

df. Shape , revealing that it consists of 60,417 rows and 25 column

DATA OVERVIEW

Checking Information



```
df.info()

<class 'pandas.core.frame.DataFrame'>
Index: 60417 entries, 0 to 211942
Data columns (total 25 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   res_id                 60417 non-null  int64
 1   name                   60417 non-null  object
 2   establishment          60417 non-null  object
 3   url                    60417 non-null  object
 4   address                60399 non-null  object
 5   city                   60417 non-null  object
 6   city_id                60417 non-null  int64
 7   locality               60417 non-null  object
 8   latitude               60417 non-null  float64
 9   longitude              60417 non-null  float64
 10  country_id             60417 non-null  int64
 11  locality_verbose       60417 non-null  object
 12  cuisines               59947 non-null  object
 13  timings                59347 non-null  object
 14  average_cost_for_two   60417 non-null  int64
 15  price_range            60417 non-null  int64
```

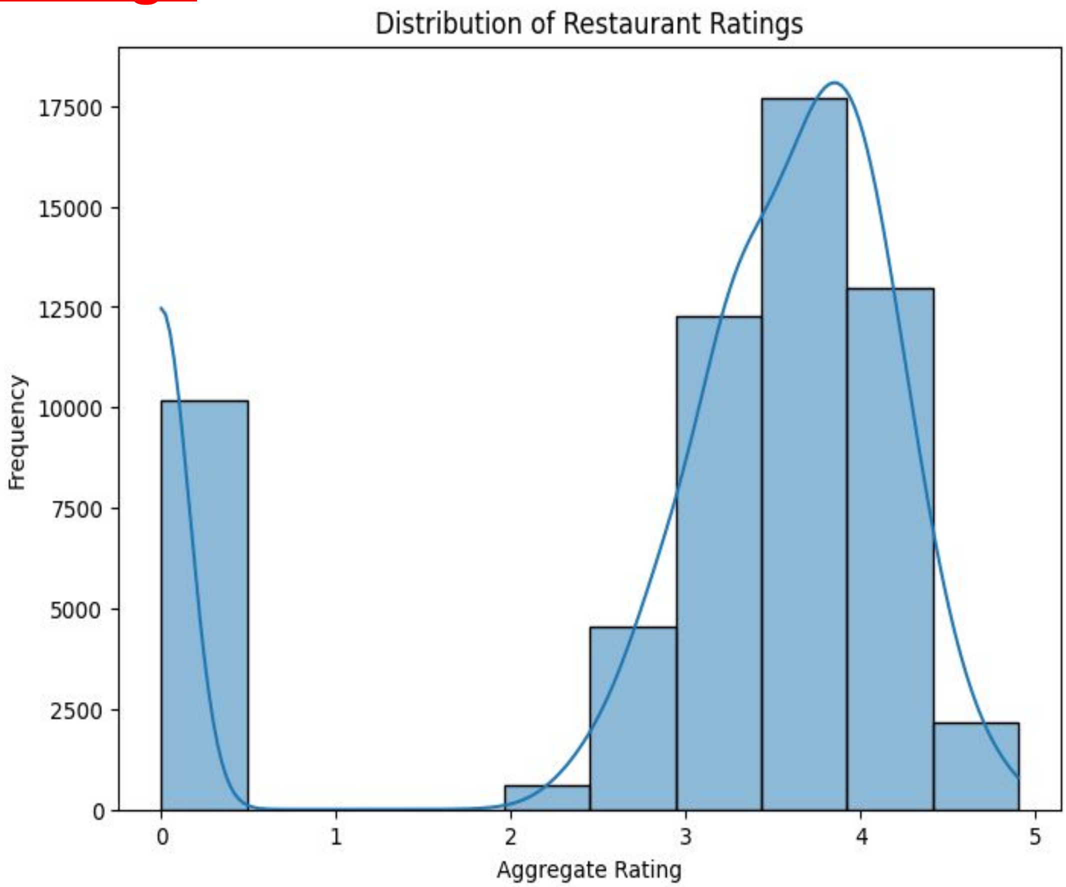
I've obtained information about the dataset using df.info(), This method provide a concise summary , including the total number of entries ,the data types of each column ,and the count of non-nun values

Removing Duplicate

I've removed duplicate rows from the dataset using df.drop_duplicates(inplace = True) ,eliminate any duplicate entries based on all columns

DATA VISUALILIZATION

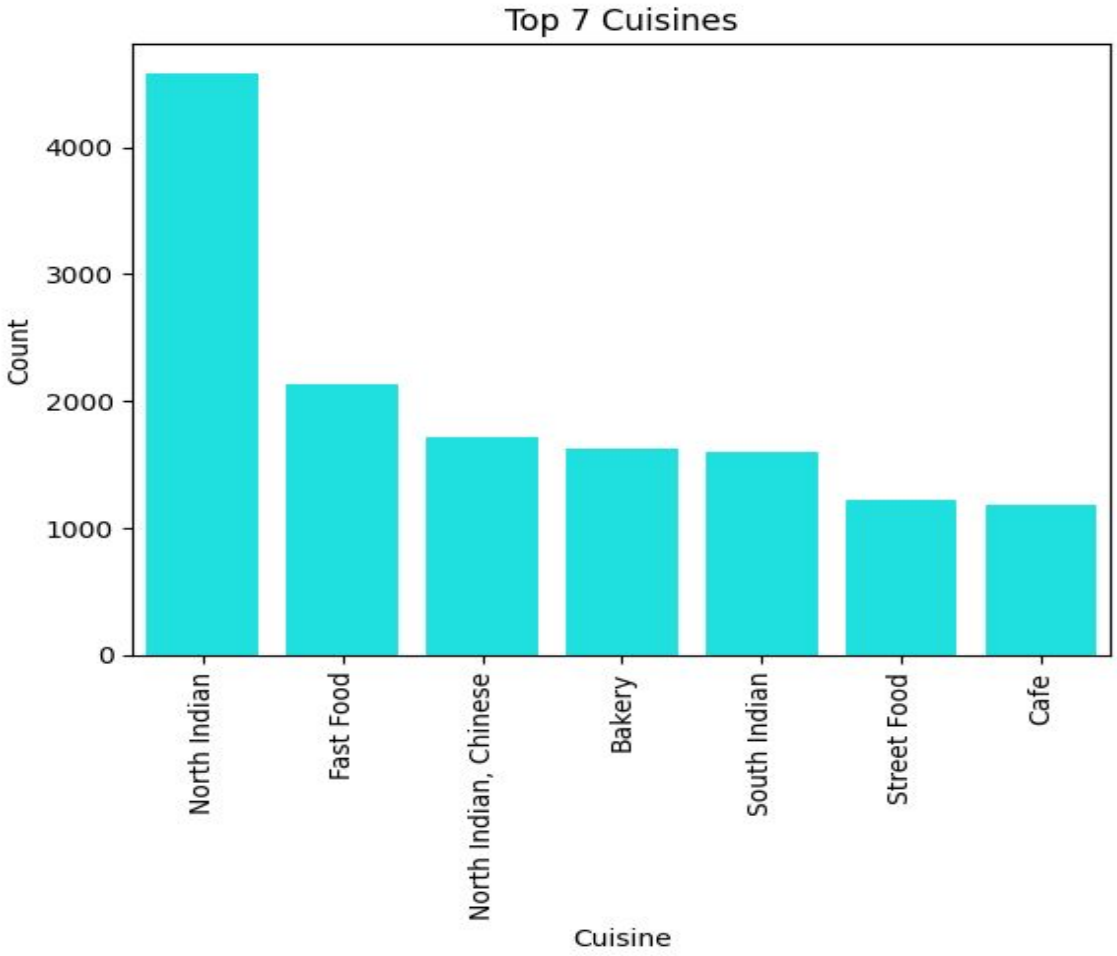
Distribution of Overall Restaurant Ratings



- **Bimodal Distribution -** Two peaks are visible, one around 0 and another around 4.
- **High Frequency at Zero -** Many ratings are at 0, likely missing or not rated.
- **Most Ratings Between 3-4 -** Majority of restaurants have ratings in this range.
- **Right Skewed -** More ratings are on the higher end.
- **Few Very Low Ratings -** Very few ratings between

DATA VISUALIZATION

Top 7 Cuisines



•**North Indian** is the most popular cuisine, with the highest number of orders.

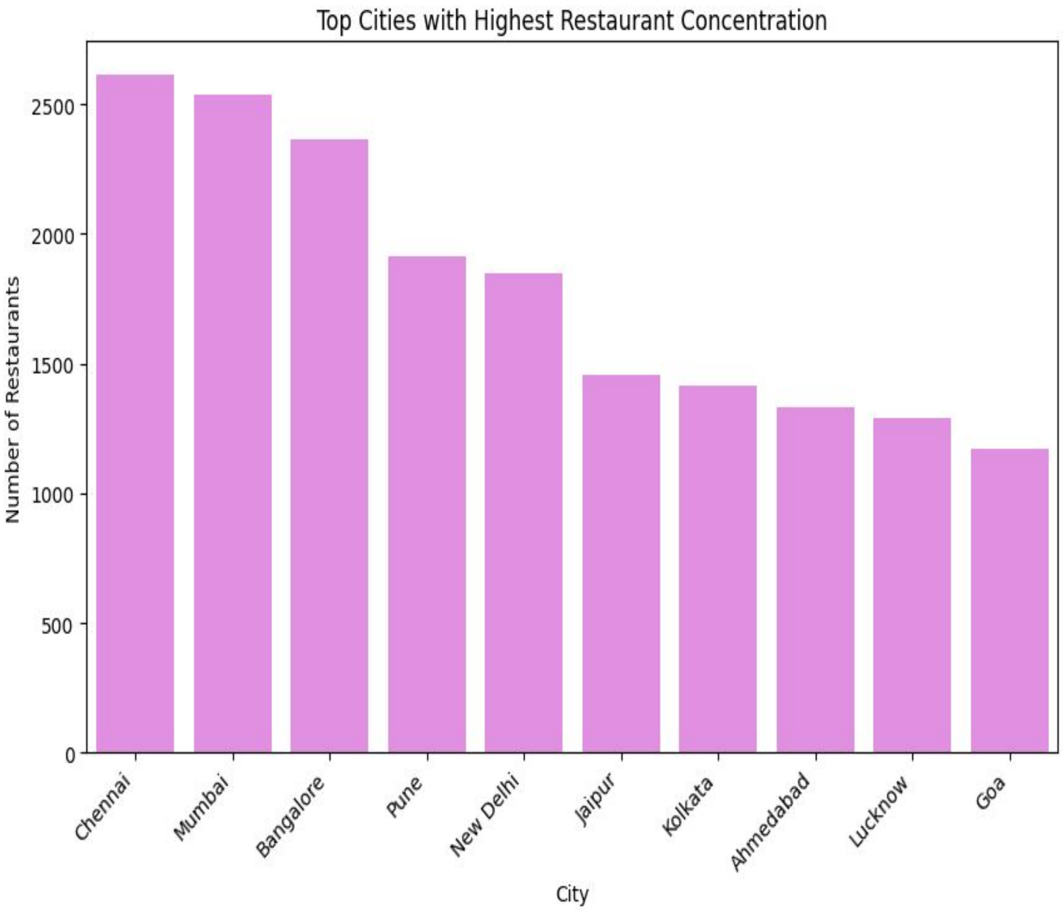
•**Fast Food** holds the second position, though significantly lower than North Indian.

North Indian, Chinese, Bakery, and **South Indian** have similar order counts, indicating moderate popularity.

Street Food and **Cafe** have the lowest order counts among the top 7 cuisines.

DATA VISUALIZATION

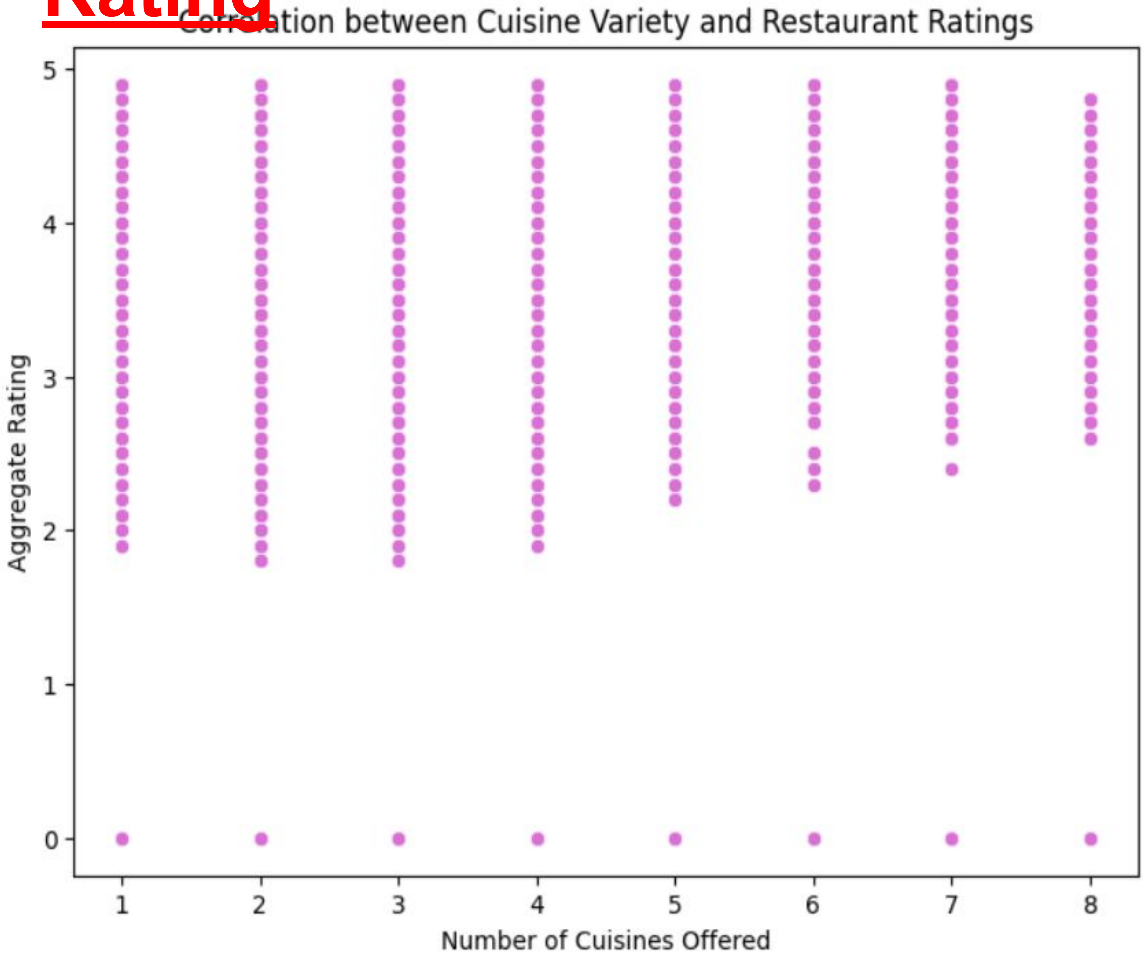
Top Cities With highest Restaurant



- **Chennai, Mumbai, Bangalore** have the highest number of restaurants.
- **Pune & New Delhi** follow in ranking.
- **Jaipur, Kolkata, Ahmedabad, Lucknow, and Goa** have relatively fewer restaurants.
- **Chennai leads** with the most restaurants.

DATA VISUALIZATION

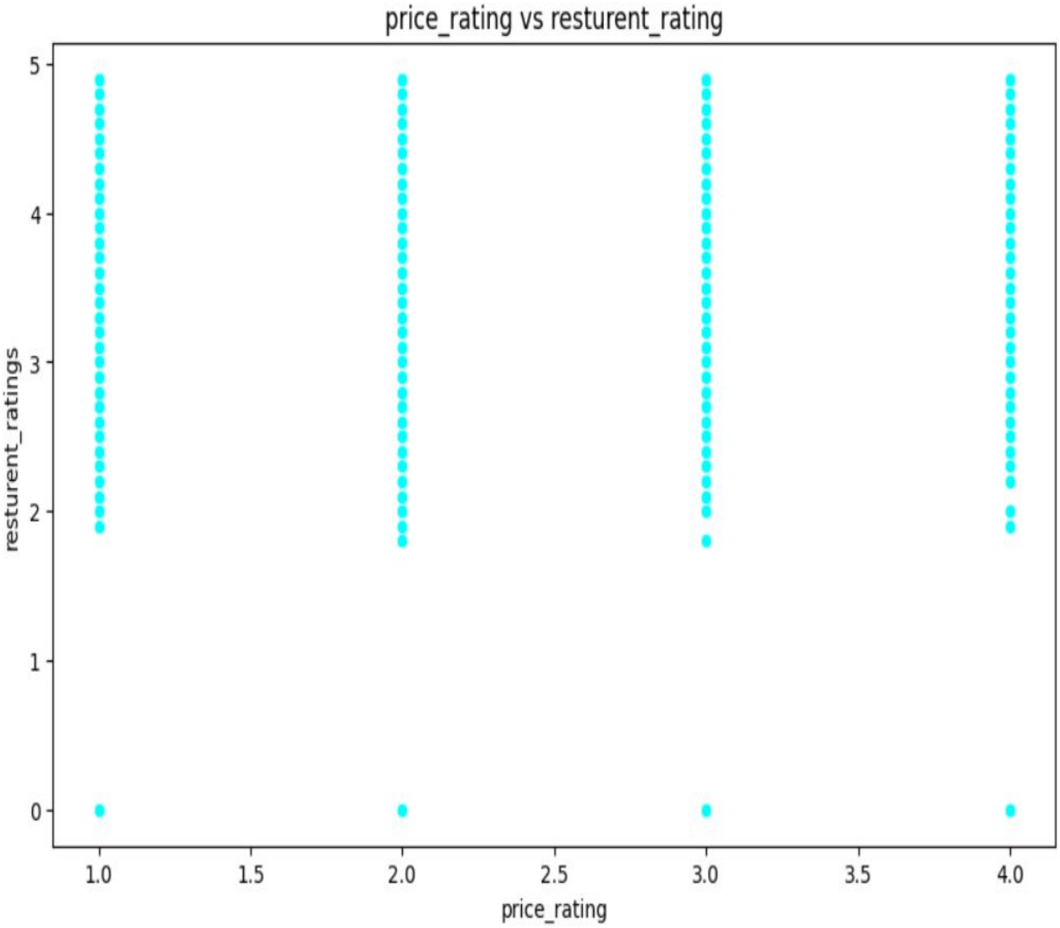
Correlation between Cuisine Variety and Restaurent Rating



- **More Cuisines Higher Rating –** No strong trend seen.
- **Most Ratings Cluster Around 5 –** Many restaurants have high ratings.
- **Some Ratings Are 0 –** Possible missing data or very poor reviews.
- **Even Spread Across Cuisine Counts –** No one category dominates.

DATA VISUALIZATION

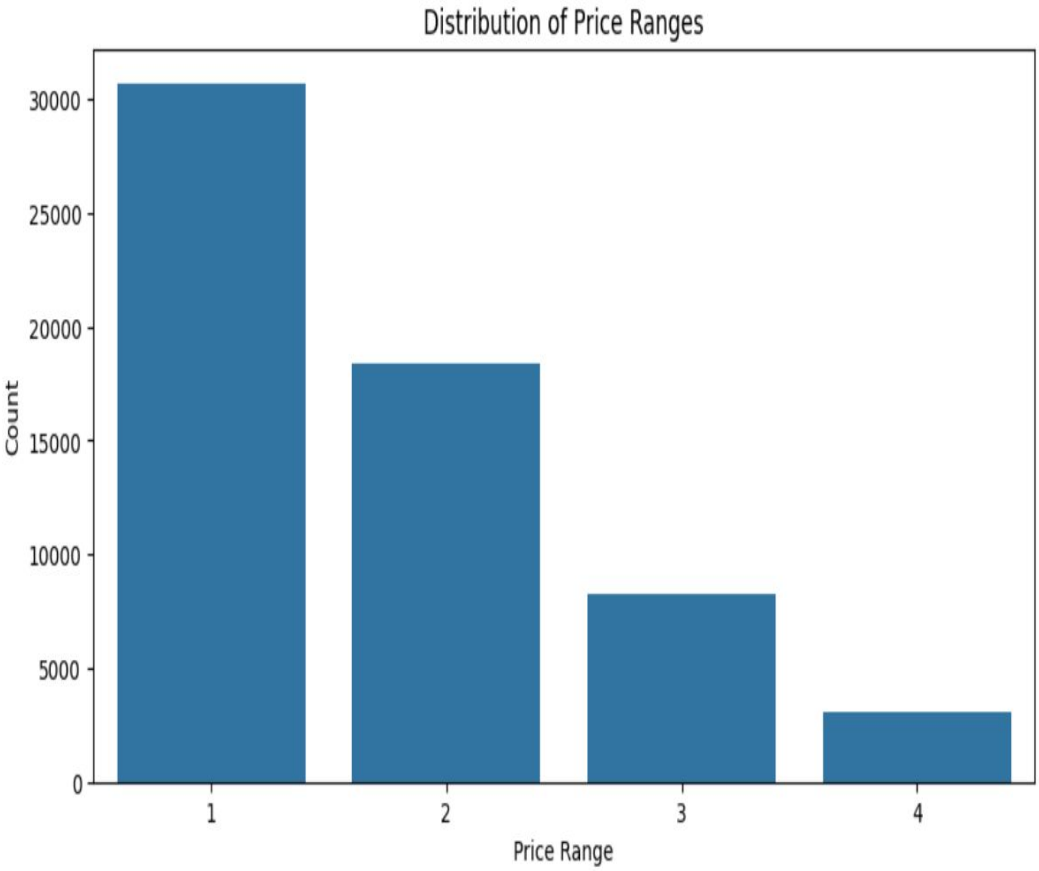
Relationship between Price-rating and Restaurent Rating



- **Distinct Categories:** Price ratings and restaurant ratings are in discrete groups.
- **Even Distribution:** All price levels have restaurants with ratings from 0 to 5.
- **No Clear Trend:** No strong correlation between price and rating is visible.
- **Outliers:** Some restaurants have very low ratings despite different price levels.

DATA VISUALIZATION

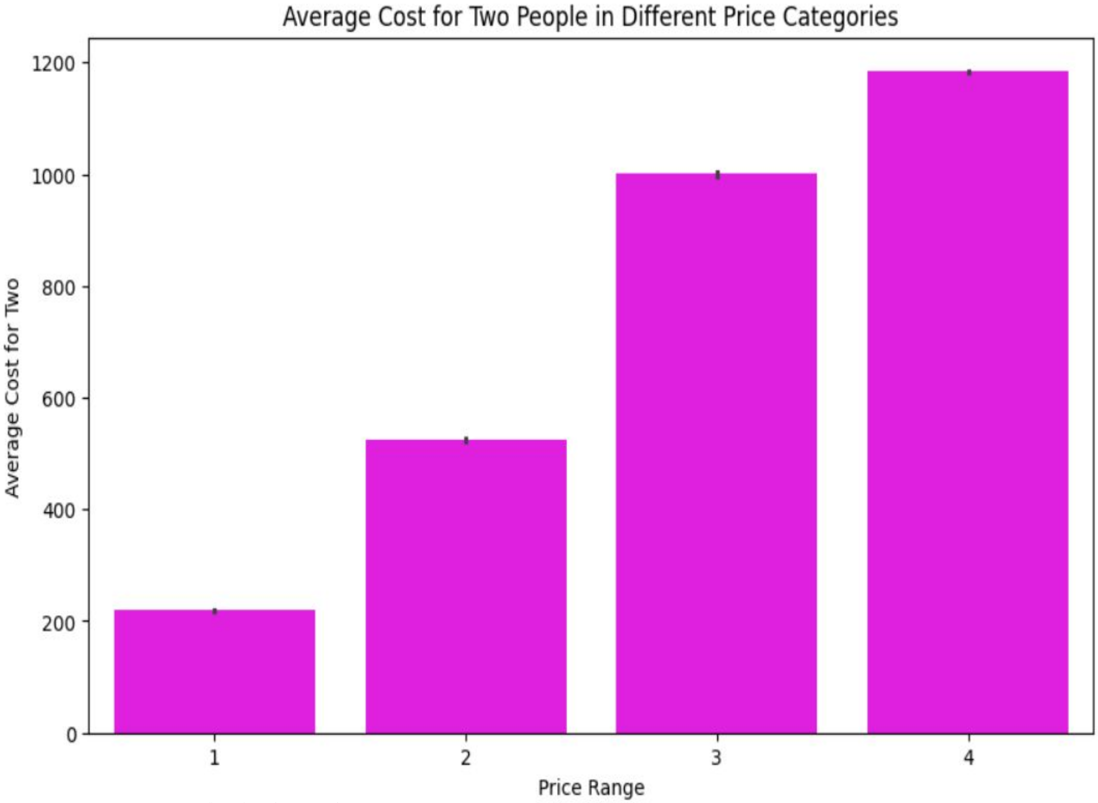
Relationship between Price-rating and Restaurent Rating



- **Majority in Low Price Range:** Most restaurants belong to the lowest price range (1).
- **Declining Trend:** The number of restaurants decreases as the price range increases.
- **Few in High Price Range:** Very few restaurants fall in the highest price category (4).
- **Affordability Insight:** The chart indicates that budget-friendly restaurants are more common.

DATA VISUALIZATION

Average Cost for two People Price Categories

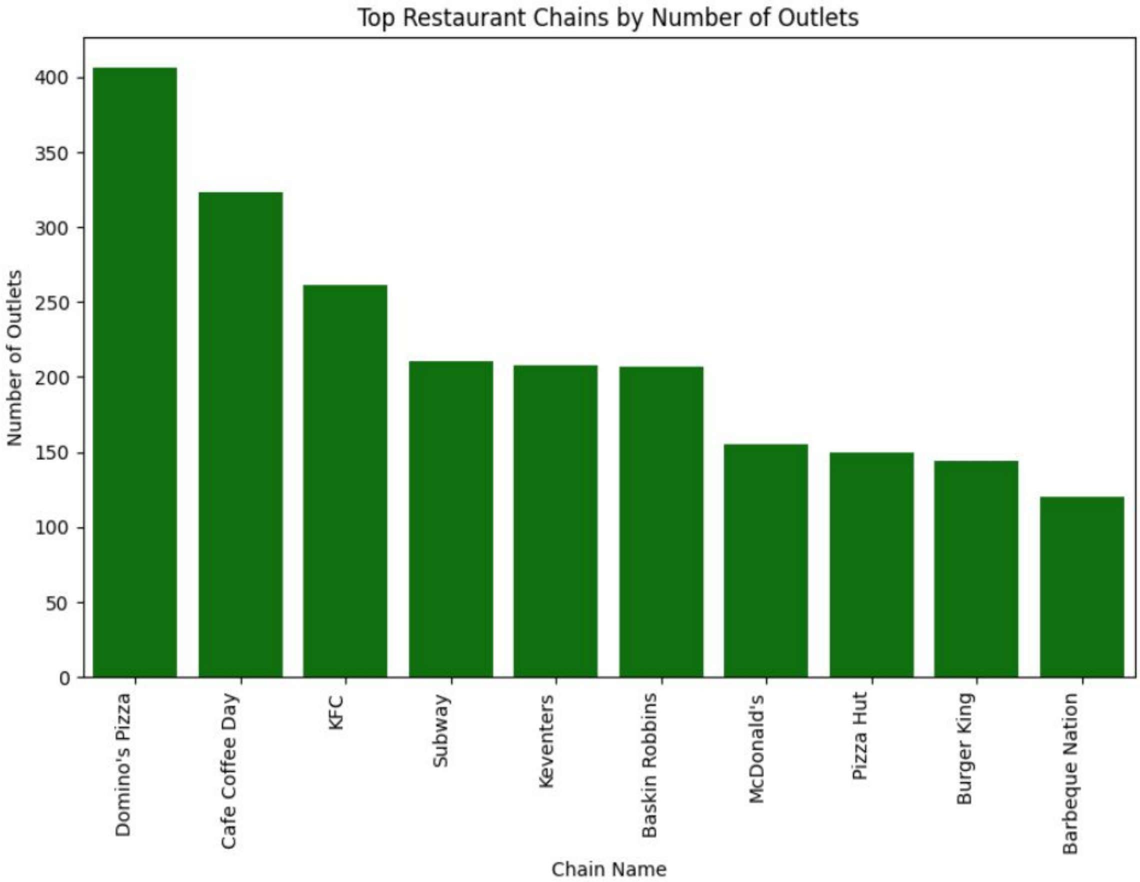


•**Big Idea:** The plot shows how much it costs to eat at different types of restaurants (based on their price range).

- **Bars:** Each bar shows a different price range (like cheap, medium, expensive).
- **Big Idea:** The plot shows how much it costs to eat at different types of restaurants (based on their price range).
- **Height:** The height of each bar tells you the average cost of a meal for two people at that type of restaurant.
- **Labels:** The plot has clear labels to help you understand what it's showing

DATA VISUALIZATION

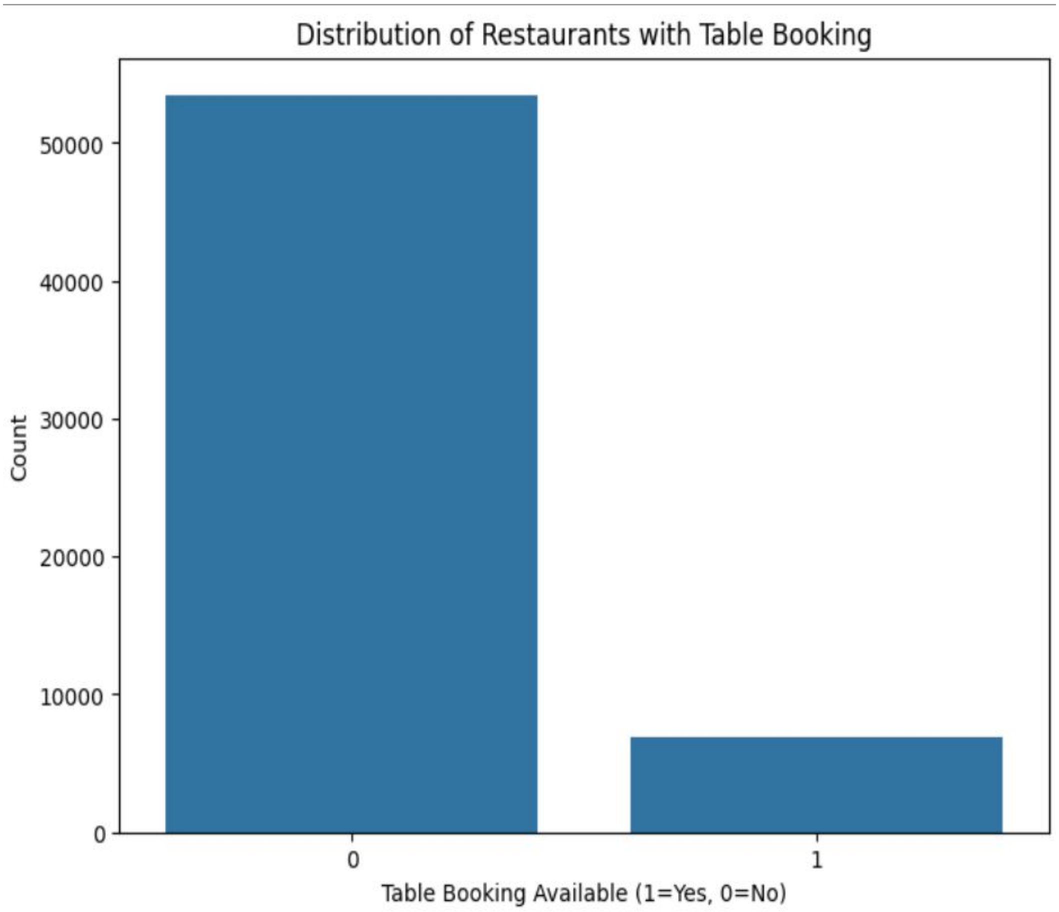
Top Restaurant Chains by Number of Outlet



- **Domino's Pizza** has the highest outlets (~400).
- **Cafe Coffee Day & KFC** follow with 300+ and 250+ outlets.
- **Subway, Keventers & Baskin Robbins** have similar (~200) counts.
- **McDonald's, Pizza Hut & Burger King** have slightly fewer (~150-200).
- **Barbeque Nation** has the lowest among them (~120).

DATA VISUALIZATION

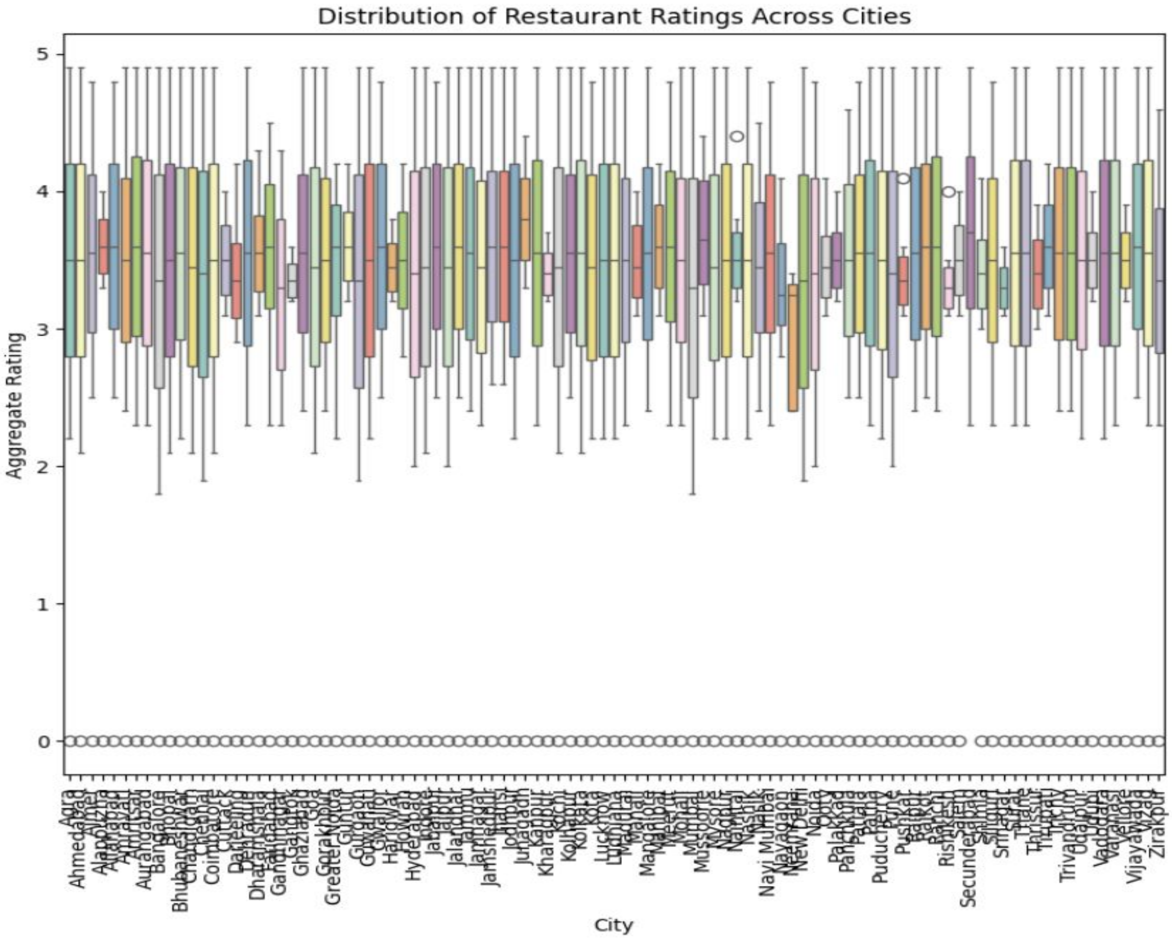
Distribuition of Restautant With Table Booking



- Most restaurants **don’t offer** table booking.
- Only a **small fraction** provides booking.
- Walk-in customers dominate the industry.
- Demand for reservations might be **low** or **venue-specific**.
- Opportunity for growth in table booking services.

DATA VISUALIZATION

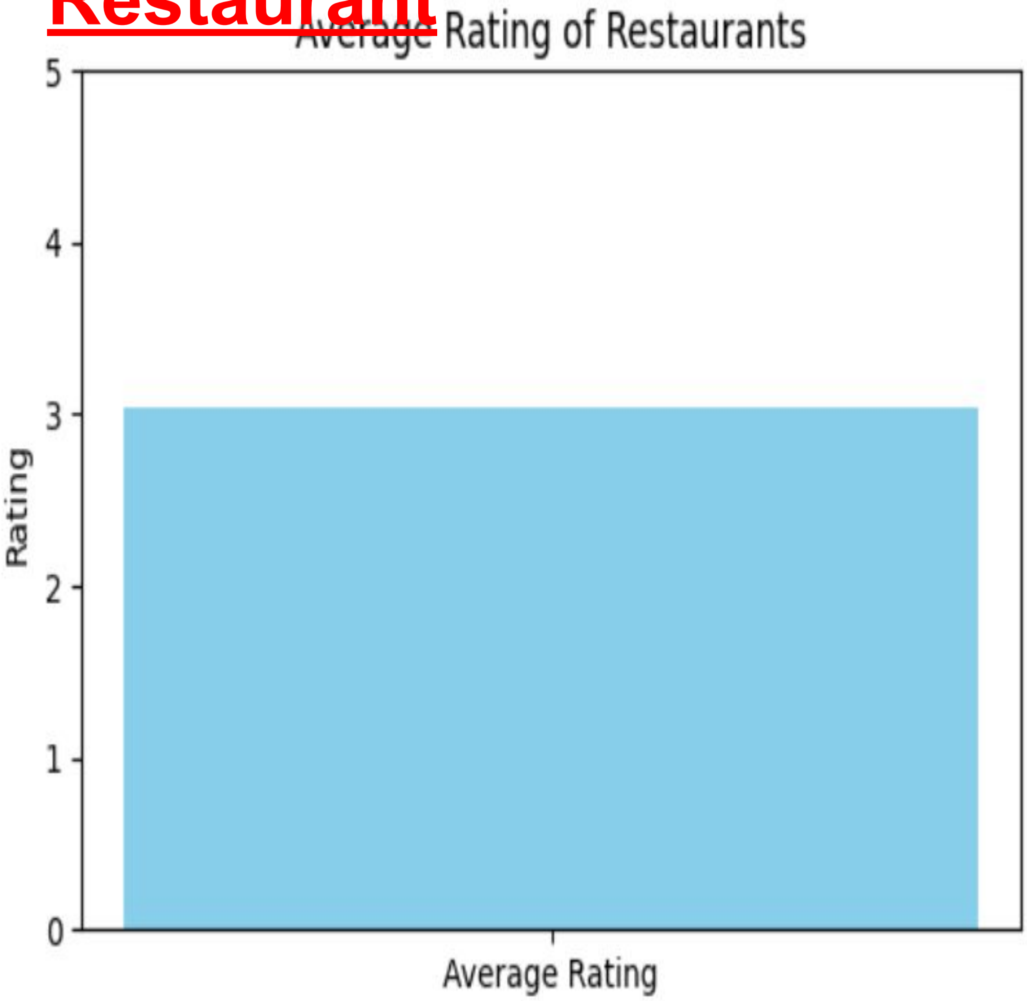
Distribuition of Restautant Rating Across Cities



- Most cities have **ratings between 3 to 4.5**.
- Few outliers with **very low or high ratings**.
- Some cities show **higher rating variability**.
- Median ratings are consistent across **cities**.
- Overall, restaurant ratings are **mostly positive**

DATA VISUALIZATION

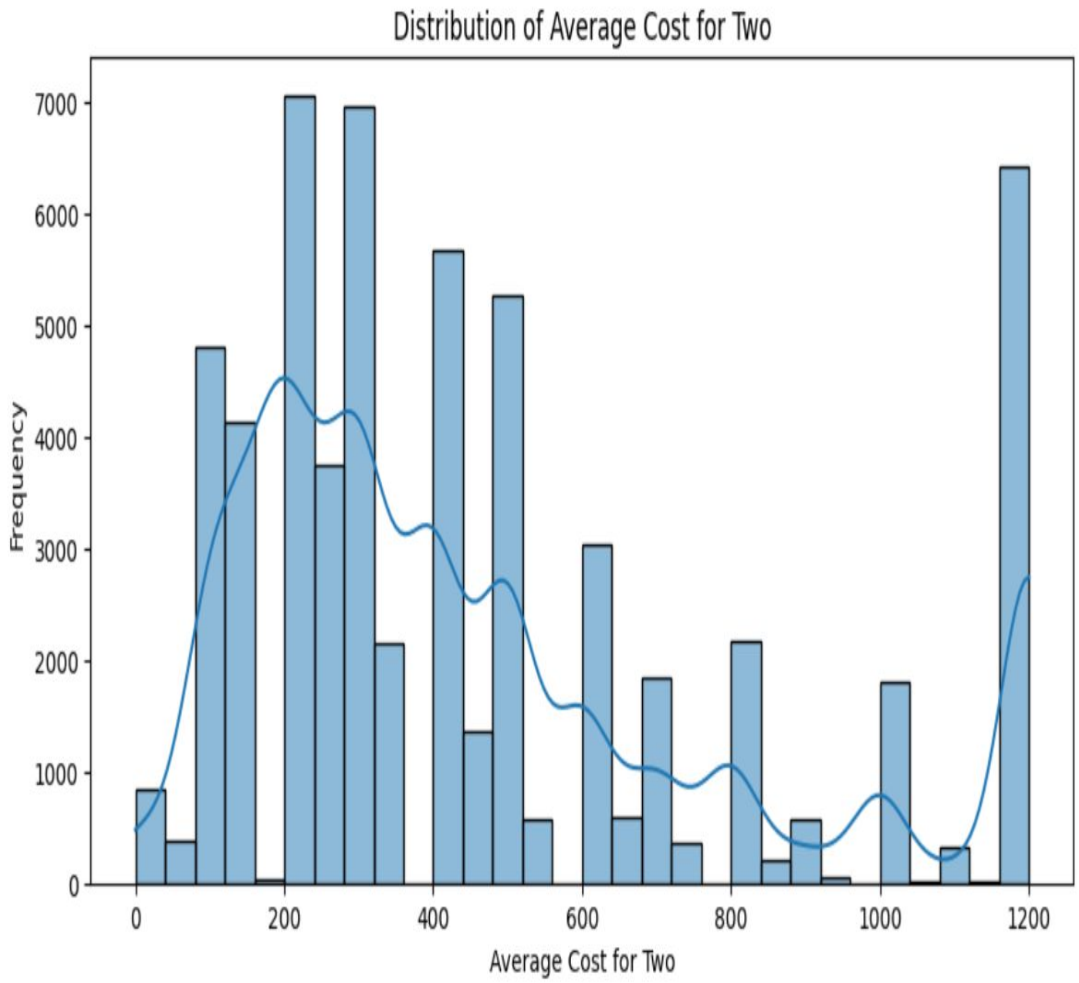
Average Rating of Restaurant



- The **average rating** is around **3**.
- Ratings are **evenly spread** across the range.
- No significant **high or low rating bias.**
- Most restaurants have **moderate customer satisfaction.**
- Opportunity to **improve ratings** through better service.

DATA VISUALIZATION

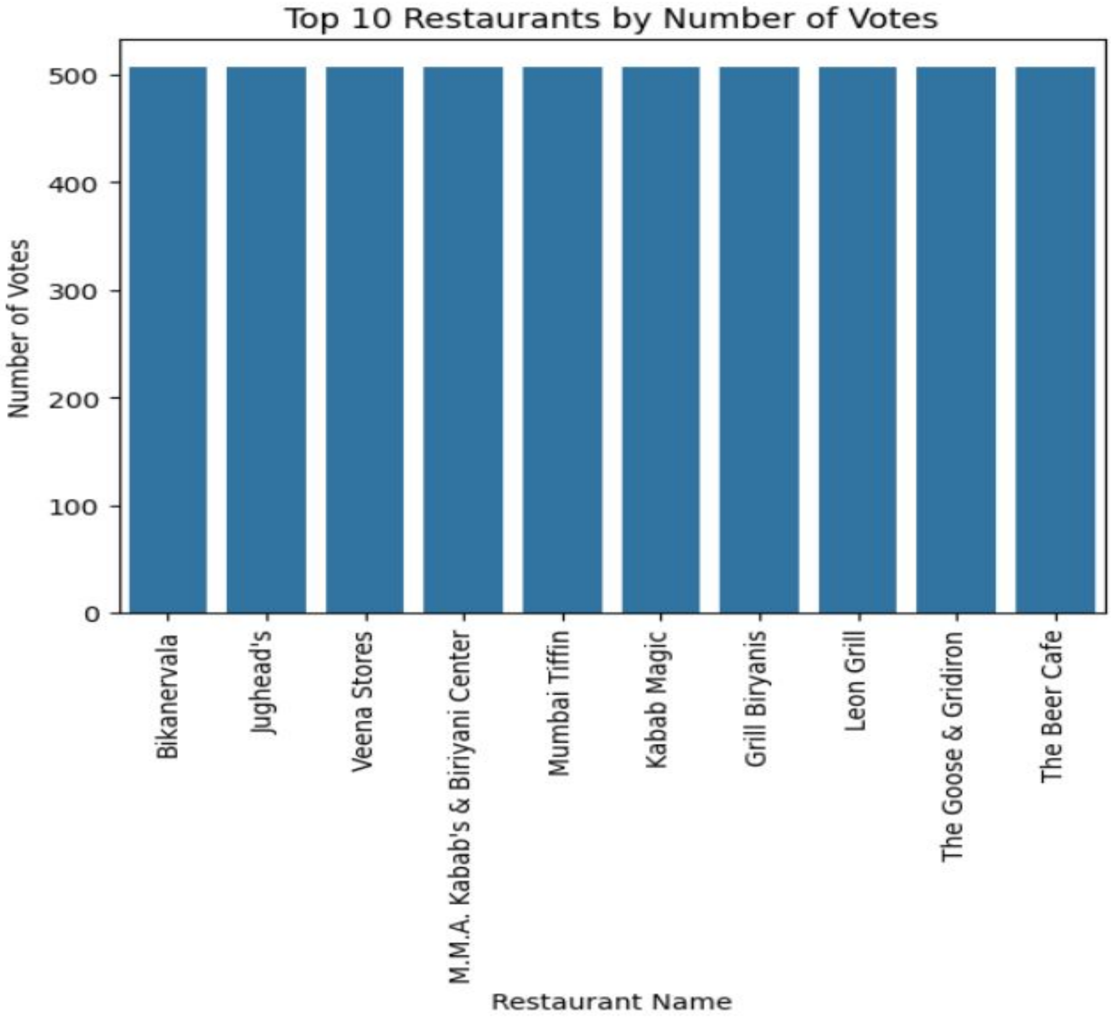
Distribuition of Average Cost for Two



- **Most common costs:** Around 200-400
- **High frequencies:** at 200, 400, and 1200
- **Some peaks:** at 600, 800, and 1000
- Skewed distribution with multiple peaks

DATA VISUALLIZATION

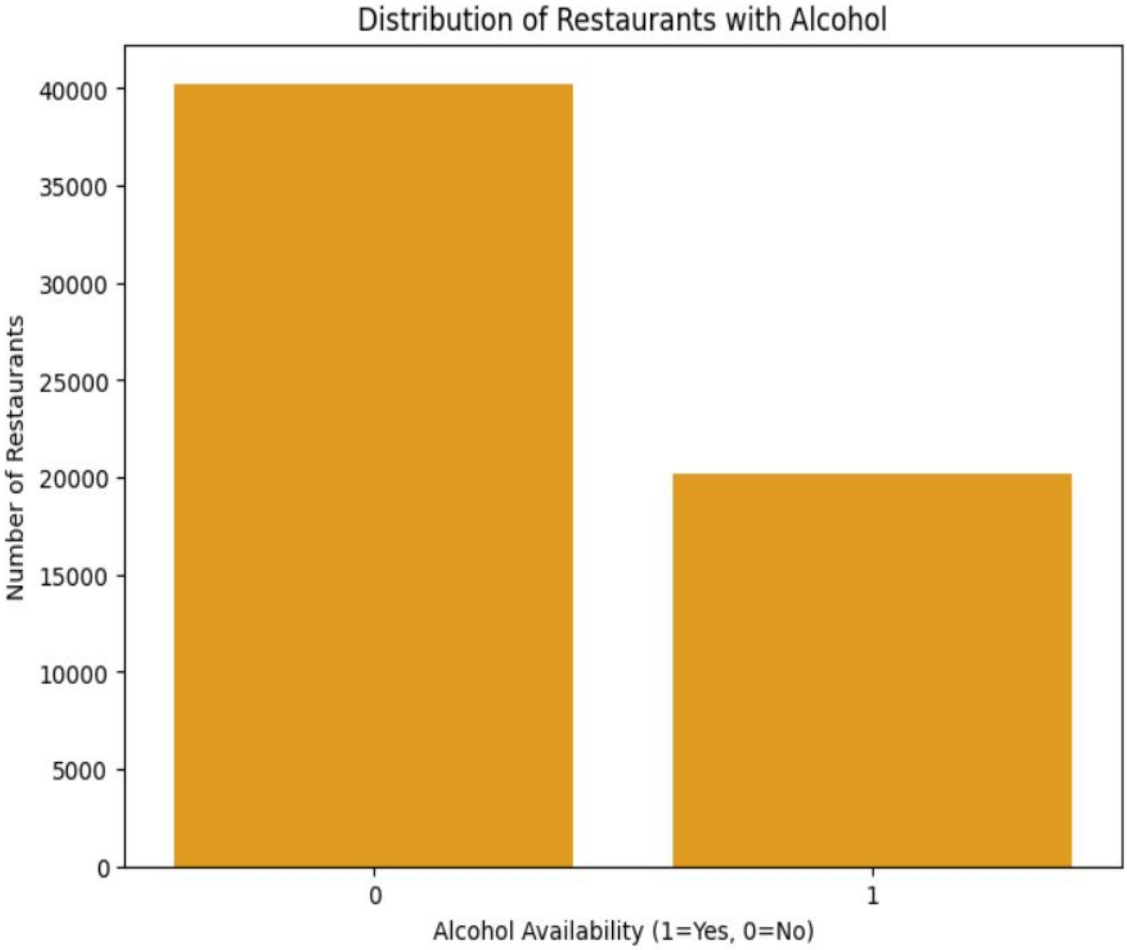
Top 10 Restaurant by Number of Votes



- **Equal Votes:** All top 10 restaurants have around 500 votes.
- **High Popularity:** These restaurants are customer favorites.
- **Hard-to-Read Labels:** X-axis names are rotated.
- **Clear Comparison**: Bar chart makes ranking easy.

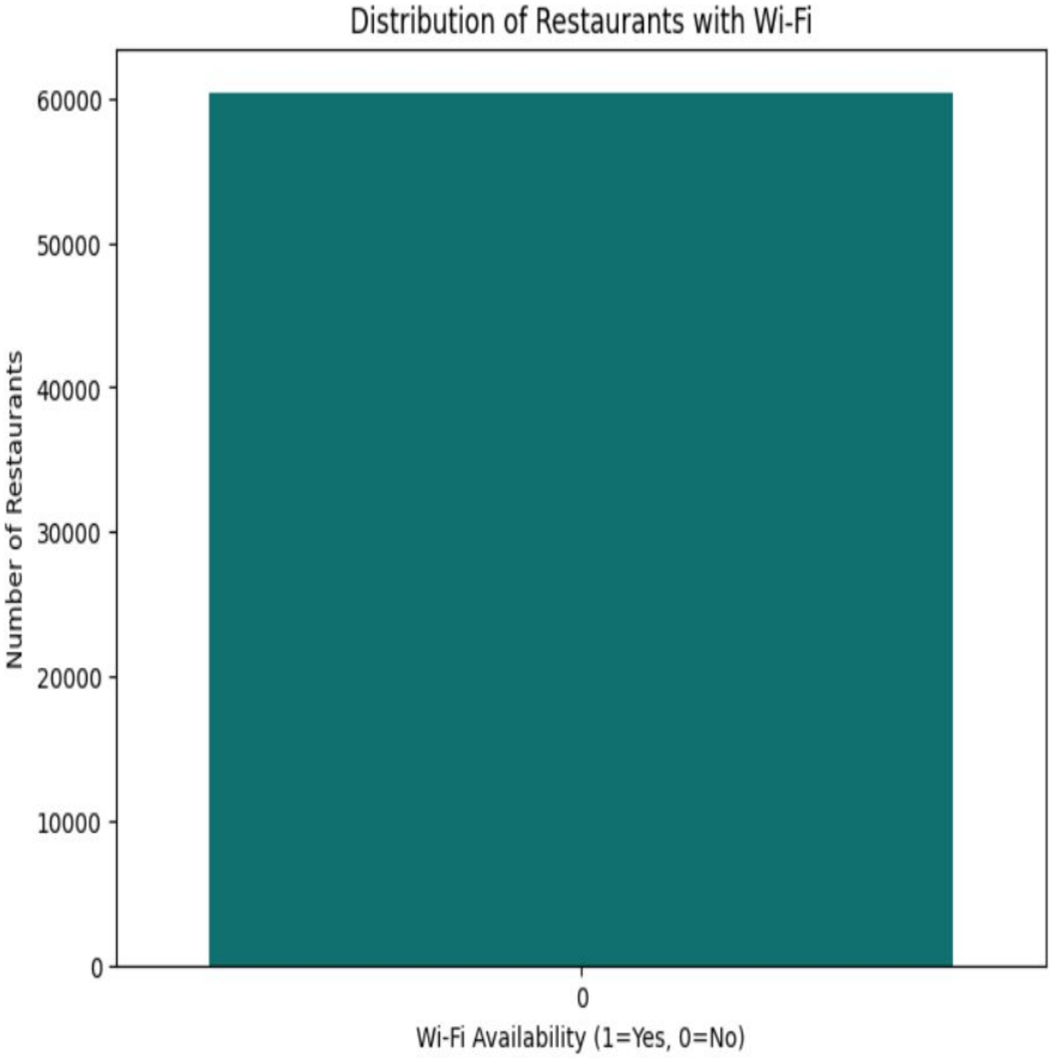
DATA VISUALIZATION

Distribuition of Restaurant with Alcohol



- **Majority of Restaurants (0):** Do not serve alcohol (~40,000).
- **Fewer Restaurants (1):** Serve alcohol (~20,000).
- **Ratio:** More than twice as many non-alcohol-serving restaurants.
- **Trend:** Alcohol availability is less common in restaurants.

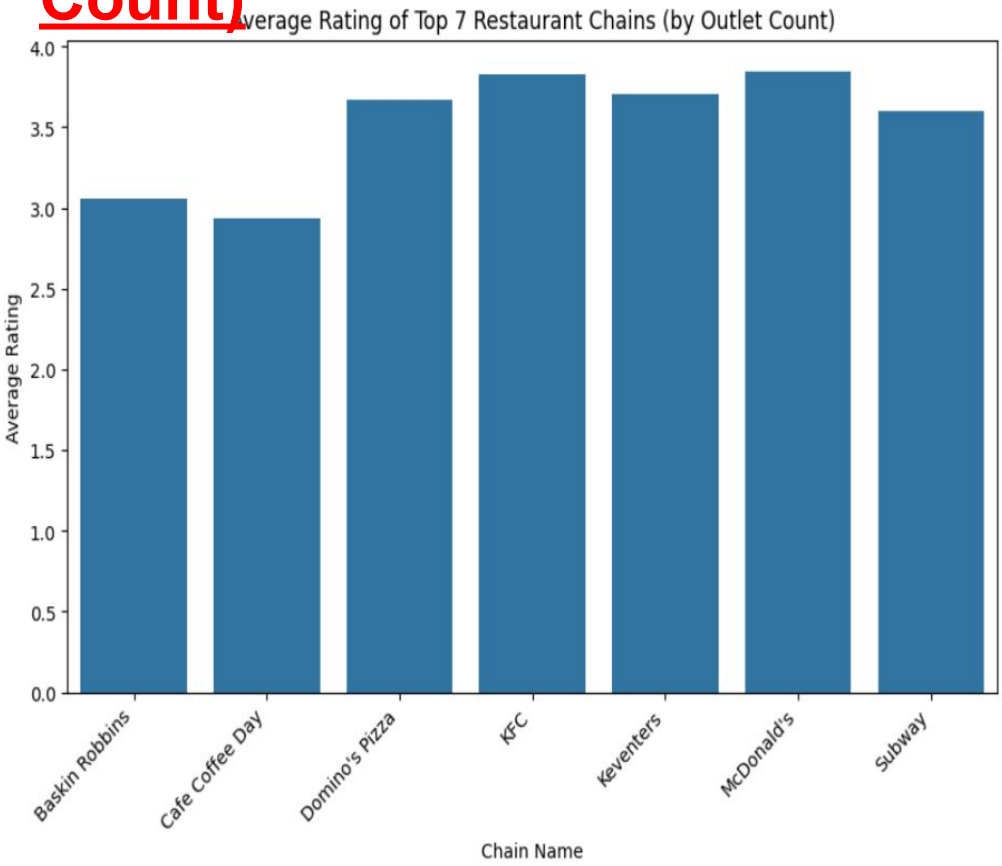
Distribuition of Restaurant with Wi-Fi



- **Most Restaurants (0):** Do not offer Wi-Fi (~60,000).
- **Restaurants with Wi-Fi (1):** Almost negligible.
- **Trend:** Wi-Fi availability is very rare in restaurants.

DATA VISUALIZATION

Average Rating of Top 7 Restaurant City(by Outlet Count)



- **Highest Rated Chains:** McDonald's and KFC (~4.0).
- **Lowest Rated Chain:** Café Coffee Day (~3.0).
- **Most Chains' Ratings:** Above 3.5.
- **Similar Ratings:** Subway, Domino's, and Keventers (~3.7-3.8).
- **Overall Performance:** All top 7 chains have good ratings.

RECOMMANDATION

📌 Enhance Data Quality

✅ Reduce missing values & ensure data consistency.

✅ Implement robust validation methods.

📌 Understand Customer Preferences

✅ Identify & promote popular cuisines in different locations.

✅ Utilize customer reviews & ratings for better insights.

📌 Optimize Pricing Strategy

✅ Set pricing based on customer affordability & market trends.

✅ Adjust pricing for low-rated restaurants to improve performance.

📌 Improve Operational Efficiency

✅ Streamline restaurant operations during peak hours.

✅ Strengthen logistics & delivery partnerships for faster service.

📌 Boost Marketing & Engagement

✅ Utilize social media & digital marketing for restaurant promotions.

✅ Introduce loyalty programs & discounts to retain customers.

CONCLUSION

1️⃣ **Cuisine Variety and Customer Ratings –** Restaurants offering a wider variety of cuisines tend to receive higher ratings.

2️⃣ **Pricing and Customer Satisfaction –** Expensive restaurants don’t always guarantee better ratings, so balancing prices is essential.

3️⃣ **City-Specific Strategies –** Some cities have a higher concentration of restaurants, while others have less competition, creating opportunities for expansion.

4️⃣ **Online Presence and Digital Marketing –** Social media, online booking, and customer reviews play a critical role in the success of restaurants.

5️⃣ **Amenities and Customer Attraction –** Offering Wi-Fi, alcohol, and table booking options can significantly impact customer preferences.

THANX FOR READING

FOR CODE

https://colab.research.google.com/drive/1zeZ2mOV04LioDLpc05fAyJHbr-6pc_yU?usp=sharing