Milestone 1 Report: Data Collection, Preprocessing, and Exploratory Data Analysis (EDA)

Course: CAP5771 – Spring 2025
Name: Rushang Sunil Chiplunkar
Date: February 23, 2025

---

1. Objective of the Project

The primary goal of this project is to develop a recommendation tool that aggregates and analyzes diverse datasets from Zomato, Swiggy, and Indian Restaurants to recommend restaurants in Bangalore. Currently, there is no solution that consolidates reviews and review counts from these platforms, which is a real-world problem for users seeking comprehensive insights. For this milestone, I am focusing on collecting and preprocessing data as well as performing exploratory data analysis. The long-term aim is to build an interactive tool that collects user input on preferred cuisines, addresses, costs, and localities, and recommends restaurants based on their interests.

2. Type of Tool

Planned Tool Options

- **Interactive Dashboard**: An interface to visualize trends, ratings, costs, cuisines, and other key metrics of restaurants in Bangalore using Streamlit.
- **Recommendation Engine**: A system that suggests personalized restaurant options based on user preferences and historical data, allowing users to input their preferences.

At this stage, my work lays the foundation for whichever final tool is chosen by integrating and analyzing data from multiple sources.

3. Data to Be Used

I am using three primary datasets containing restaurant information for Bangalore:

- **Swiggy Dataset**: Extracted from a JSON file, this dataset provides details such as restaurant names, areas, ratings, rating counts, cost for two, addresses, cuisines, and menu items.
- **Zomato Dataset**: Provided in CSV format, this dataset includes restaurant name, address, location, cuisines, ratings, vote counts, and approximate cost for two.

- **Indian Restaurants Dataset**: This dataset offers similar details specific to Indian restaurants in Bangalore. It has been filtered to include relevant information such as restaurant name, location, locality, cuisines, ratings, votes, and cost.

Each dataset has been cleansed and standardized so that key features (e.g., name, address, rating, and cost for two) align for further analysis and integration.

4. Tech Stack

Programming Language

- Python

Libraries & Frameworks

- **Data Processing**:
  - ijson: For streaming JSON parsing as the file is almost 1 GB, making it harder to process.
  - Pandas: For data manipulation and cleaning.
  - NumPy: For numerical operations.
- **Data Visualization**:
  - Matplotlib and Seaborn: To generate charts, histograms, box plots, and other visualizations for the EDA.
- **Additional Tools**:
  - Jupyter Notebook / Python scripts: For developing and testing the data pipeline.
  - Git & GitHub: For version control and repository management.

Future milestones will also incorporate machine learning libraries such as Scikit-Learn for modeling.

5. Project Timeline & Future Tasks

Milestone 1: Data Collection, Preprocessing, and Exploratory Data Analysis (EDA)

Timeline: February 5, 2025 – February 21, 2025 (2 weeks)
Tasks:

- **Data Collection**:
  - Acquire datasets from approved sources (Swiggy, Zomato, and Indian Restaurants).
  - Verify dataset accessibility and document their properties (dimensions, attributes, and source URLs).

- **Data Preprocessing**:
    - Clean and preprocess data (e.g., handling missing values, converting data types, and addressing inconsistencies).
    - Standardize column names and formats across datasets.
- **Exploratory Data Analysis**:
    - Compute descriptive statistics (mean, median, standard deviation, etc.).
    - Identify patterns and anomalies using visualizations (histograms, scatter plots, box plots).
    - Merge datasets and remove duplicates to form a unified dataset.

Future Milestones (Overview)

Milestone 2: Feature Engineering, Feature Selection, and Data Modeling

Timeline: February 22, 2025 – March 28, 2025 (5 weeks)
Tasks:

- **Feature Engineering** (February 22 – March 1): Create new features from existing data to enhance model performance.
- **Feature Selection** (March 2 – March 8): Identify and select key variables that significantly impact the model.
- **Data Modeling** (March 9 – March 28): Build and train initial predictive models using selected features.

Milestone 3: Evaluation, Interpretation, Tool Development, and Presentation

Timeline: March 29, 2025 – April 30, 2025 (5 weeks)
Tasks:

- **Model Evaluation** (March 29 – April 5): Evaluate model performance using appropriate metrics and validation techniques.
- **Interpretation** (April 6 – April 12): Interpret the results and derive actionable insights from the models.
- **Tool Development** (April 13 – April 23): Develop the final tool (dashboard/recommendation engine) based on the evaluated models.
- **Presentation** (April 24 – April 30): Prepare and deliver the final presentation showcasing the project outcomes.

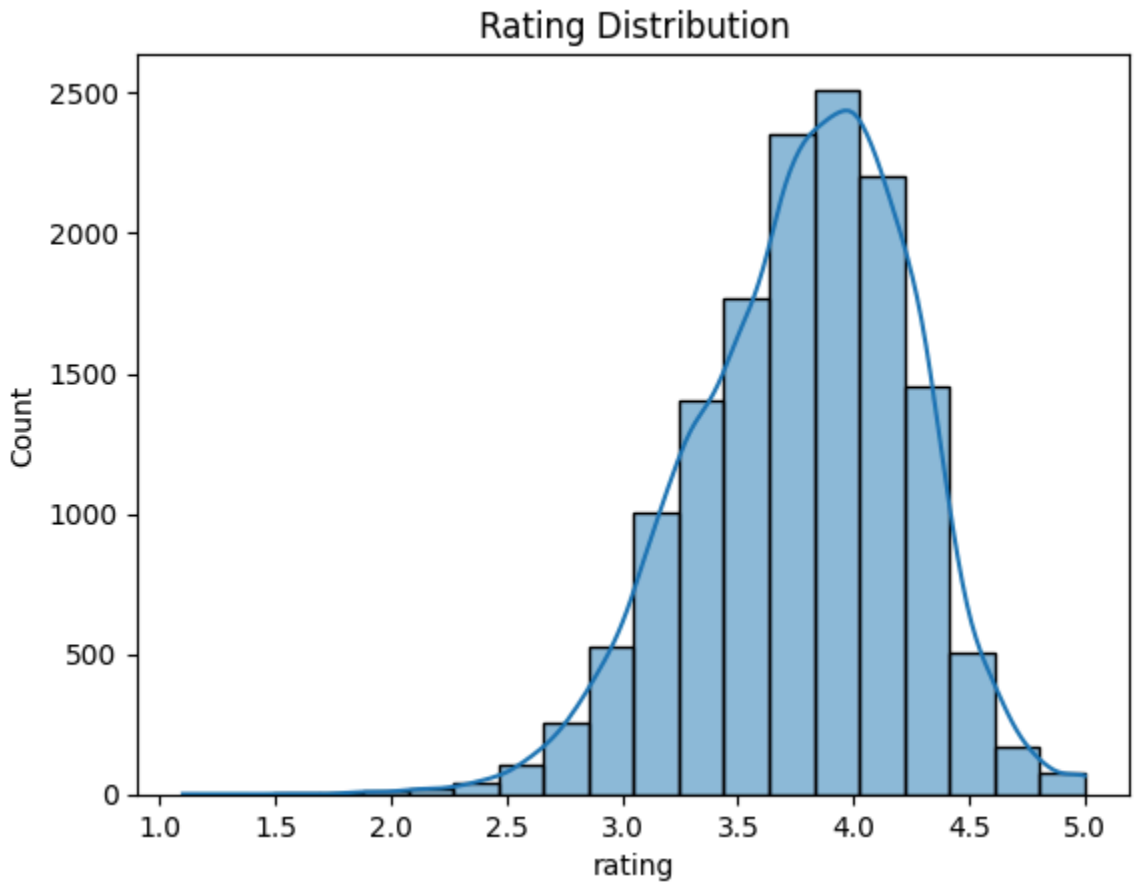
6. Exploratory Data Analysis (EDA) Report and Key Insights

Summary of EDA

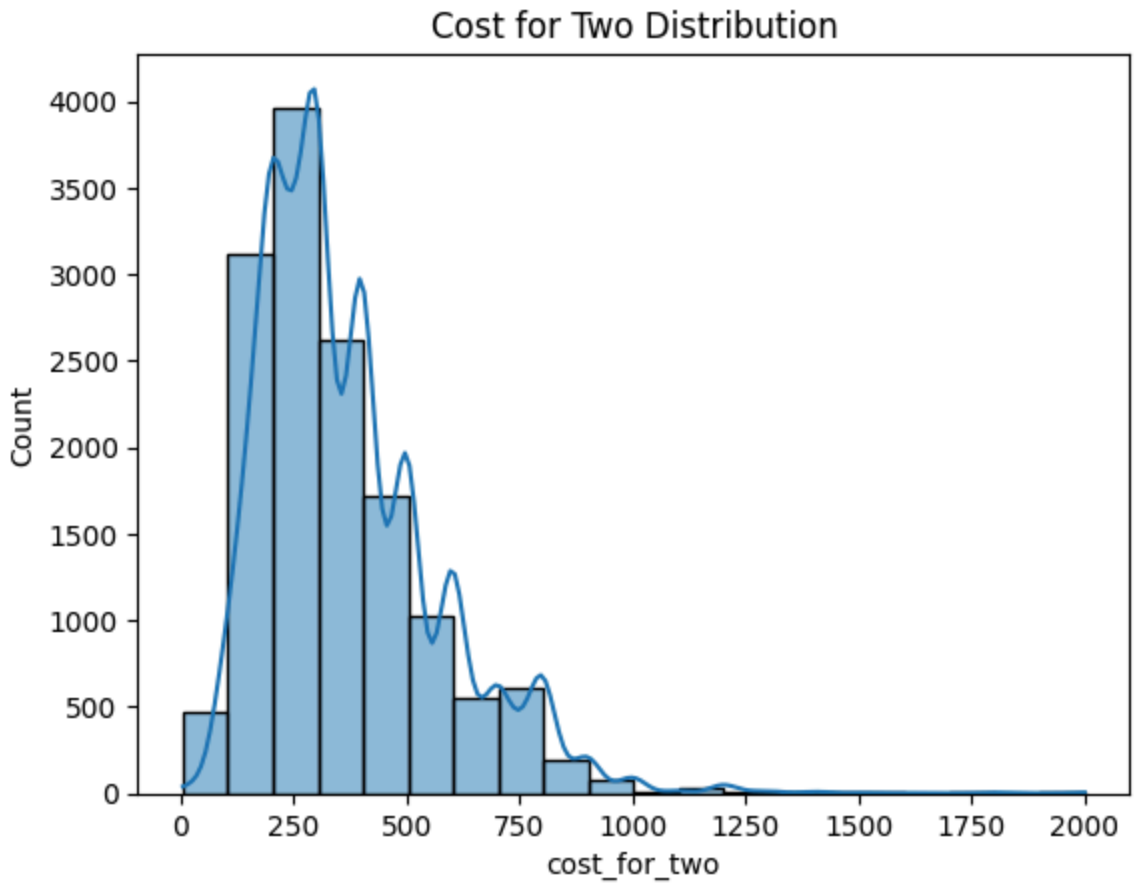
- **Data Integration**: After preprocessing and merging the datasets from Swiggy, Zomato, and Indian Restaurants, duplicate entries were removed (based on restaurant name and address), resulting in a consolidated dataset with unique restaurant records.
- **Data Quality**: Initial analyses revealed missing values in some fields (e.g., rating counts and costs), which were handled by removing incomplete records. Standardization of rating and cost columns was performed to ensure consistency.

Key Insights and Visualizations

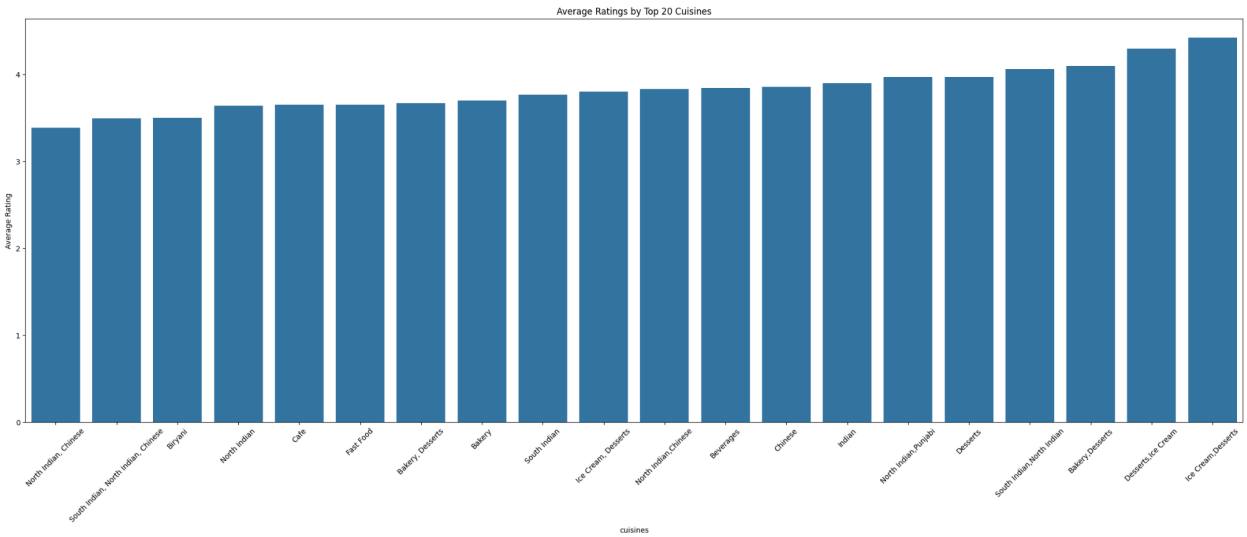
- **Ratings Distribution**: The majority of restaurants have ratings clustered between 3.5 and 4.5, indicating generally positive customer experiences.



- **Cost Analysis**: Most restaurants in the dataset have a cost for two between 125 INR and 500 INR.



- **Cuisine Popularity**: Dessert establishments are among the highest rated in Bangalore.

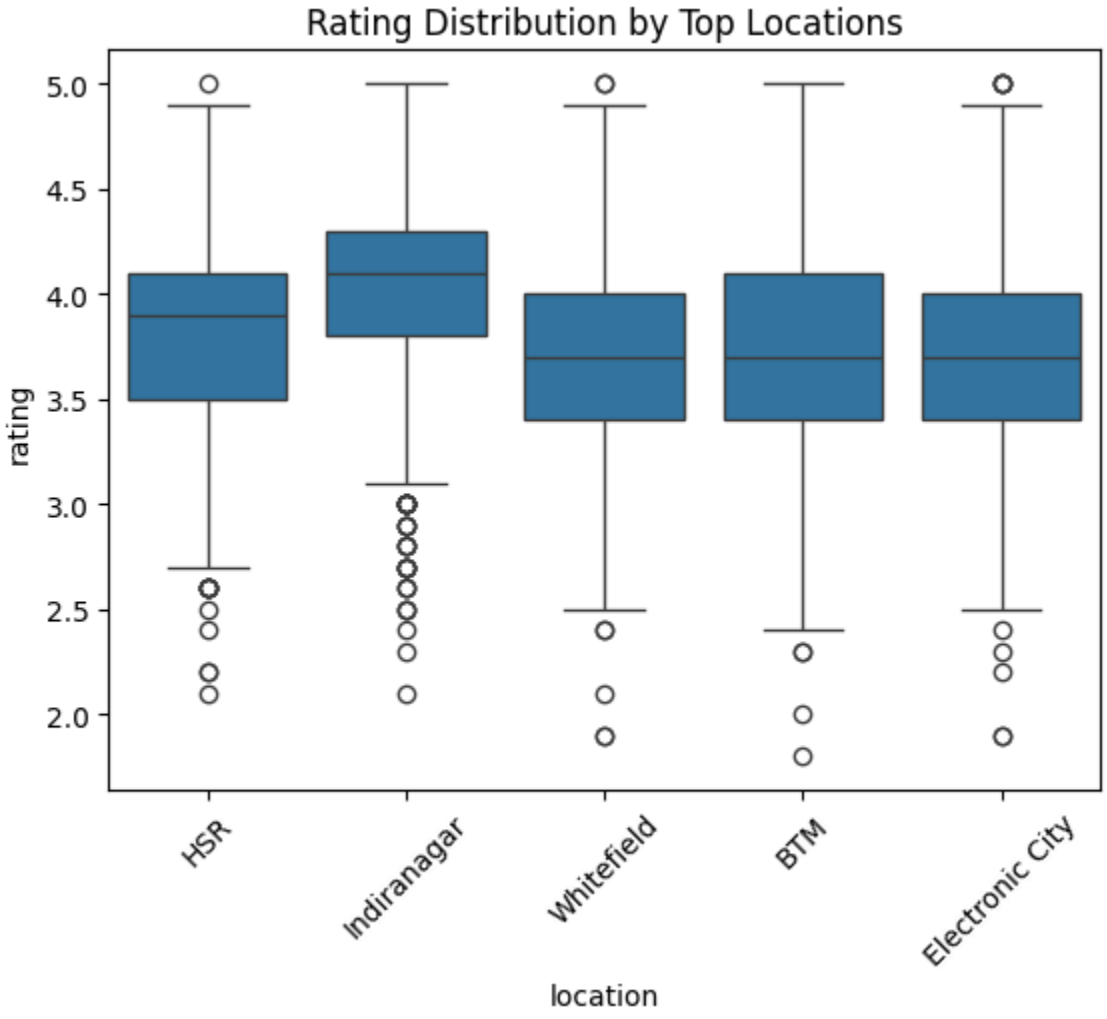


- **Descriptive Statistics Table**: Key metrics such as mean rating, median cost, and standard deviation are computed for numerical columns.

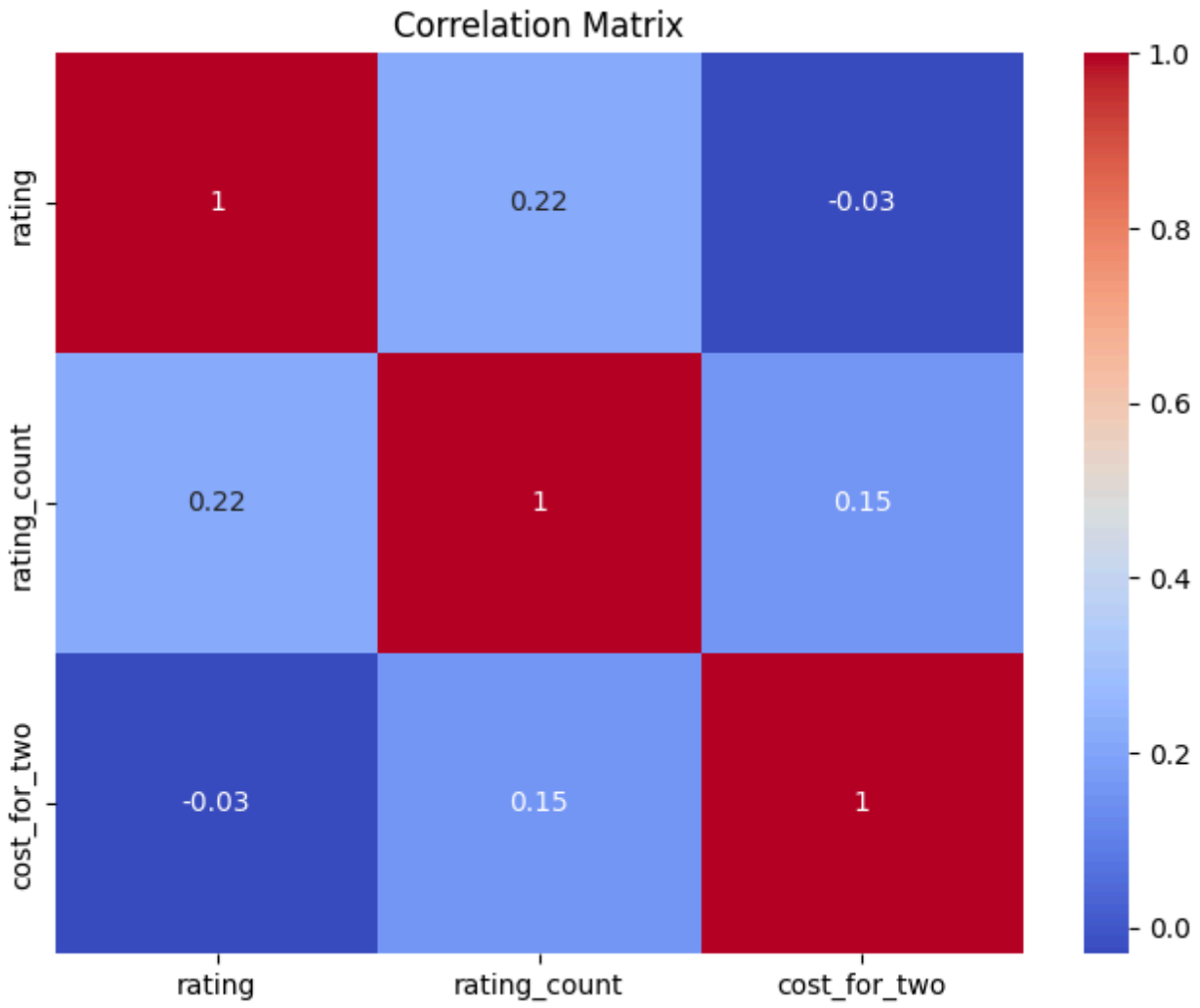
```
...   Numerical Summary:
                rating  rating_count  cost_for_two
      count  14393.000000  14393.000000  14393.000000
      mean       3.783812    149.154728    372.446606
      std        0.465240    356.140090    197.608786
      min        1.100000      0.000000      5.000000
      25%        3.500000     20.000000    230.000000
      50%        3.800000     50.000000    300.000000
      75%        4.100000    100.000000    500.000000
      max        5.000000  14654.000000   2000.000000
```

- **Cost vs Rating Relationship**: A scatter plot of 'cost for two' versus 'rating' shows a weak positive correlation, suggesting that higher-rated restaurants tend to have slightly higher costs.

- **Rating Distribution by Top Locations**: The data indicates that certain areas have a higher concentration of highly-rated restaurants.



- **Correlation Matrix**: There is a weak positive correlation between rating count and cost for two, and a moderate positive correlation between rating count and rating.



- **Normality Test**: The Shapiro-Wilk test indicates that both 'cost for two' and 'rating' distributions deviate significantly from normality, suggesting that non-parametric methods may be more appropriate for further analysis.

```
Normality Tests:
rating: p-value = 0.0000
rating_count: p-value = 0.0000
cost_for_two: p-value = 0.0000
```