Restaurant Data Analysis: Level 1 Report

This report outlines the comprehensive findings from the Level 1 phase of the restaurant data analysis project. The focus of this stage is on data exploration, preprocessing, and descriptive analysis, aiming to gain insights into the dataset, identify potential data issues, and highlight key features for further investigation in subsequent stages.

Task 1: Data Exploration and Preprocessing

1.1 Dataset Overview

The dataset used in this analysis consists of **9551 rows** and **21 columns**, representing a wide range of information about various restaurants. Each record contains valuable details about the restaurants' attributes, which include:

- **Basic Information**: Restaurant ID, name, and location (city, address, latitude, longitude, locality).
- **Operational Details**: Availability of table bookings, online delivery options, and delivery status.
- **Customer Experience**: Average cost for two, cuisine types, and aggregate ratings (rating color, rating text, and number of votes).
- **Financial Indicators**: Price range and currency used.

This dataset offers a rich source of information, providing the foundation for both descriptive and predictive analytics.

1.2 Missing Values Analysis

During data exploration, missing values were identified, specifically in the **"Cuisines"** column, where 9 missing values were detected. In order to ensure data integrity and completeness, these missing values were replaced with the label **"Unknown."** This strategy allows for better handling of incomplete data and avoids potential biases in the analysis.

1.3 Data Type Review and Conversion

The dataset contains two primary types of data:

- **Numerical**: Columns like average cost for two and aggregate ratings, which use integer and float data types.
- **Categorical**: Columns such as city, cuisines, and delivery options, which are represented as strings (object data type).

No data type conversion was necessary, as the data types for all variables were found to be appropriate for their respective analyses.

1.4 Target Variable Analysis

The **target variable**, **"Aggregate Rating,"** displays a notable class imbalance. A detailed examination of the rating distribution revealed:

- A significant concentration of ratings at **0.0** and within the range of **3.0 to 4.0**.
- Many restaurants have either very low or no ratings, while a minority of restaurants received moderate to high ratings.

This imbalance poses a challenge for modeling and suggests the need for specialized techniques, such as **oversampling**, **undersampling**, or adjusting **class weights** during model training, to ensure that the model does not favor the majority class.

Task 2: Descriptive Analysis

2.1 Statistical Summary of Numerical Columns

A summary of the numerical columns revealed important insights:

- **Average Cost for Two**: The data displays a wide range, with costs varying significantly across restaurants, indicating a diverse restaurant market.
- **Votes**: There is a high variance in the number of votes, with some restaurants receiving significantly more votes than others. This suggests that a small number of restaurants dominate customer engagement.

Key statistics such as **mean**, **median**, and **standard deviation** were calculated for these numerical columns, helping identify the central tendencies and spread of data.

2.2 Distribution of Categorical Variables

Analysis of categorical variables provided further insights:

- **Country Code**: The dataset is predominantly focused on a single country, leading to **skewness** in country representation.
- **City**: The top three cities with the highest concentration of restaurants are **New Delhi, Gurgaon, and Noida**, which likely reflects the urbanization trends and concentration of restaurants in major metropolitan areas.
- **Cuisines**: The most popular cuisine types are **North Indian, Chinese, and Italian**, indicating strong customer preferences towards these food categories.

This distribution of variables plays a key role in understanding the geographical and culinary trends in the restaurant industry, helping stakeholders make data-driven decisions.

Task 3: Geospatial Analysis

3.1 Visualizing the Distribution of Aggregate Ratings

A key component of the analysis was visualizing restaurant data geospatially. A **heatmap** was used to visualize the distribution of **aggregate ratings** across various locations. The heatmap revealed clusters of high-density restaurant locations, particularly in urban areas.

The spatial distribution of ratings revealed:

- **Highly Rated Restaurants**: Limited geographic areas where restaurants received exceptional ratings (above 4.5).
- **Low to No Ratings**: A significant number of restaurants received 0.0 ratings, indicating either **unrated** restaurants or those with **low customer satisfaction**.

This geospatial analysis confirms the **urban-centric** nature of the dataset, with a clear concentration of restaurants and varying rating patterns across different cities.

3.2 Insights from Geospatial Analysis

Key insights derived from the geospatial data include:

- Most restaurants receive **average ratings** between 3.0 and 4.0, with very few achieving exceptional ratings above 4.5.

- A considerable number of restaurants have a rating of **0.0**, which could indicate a lack of customer feedback or dissatisfaction, necessitating further investigation into why these establishments struggle to attract ratings.

These findings suggest a potential market opportunity for businesses to focus on improving customer satisfaction in order to elevate their ratings.

Conclusion

The **Level 1** analysis of the restaurant dataset provided valuable insights into the structure and characteristics of the data. The key findings from this phase include:

- **Handling Missing Data**: Missing values were addressed to ensure data completeness.
- **Class Imbalance**: The imbalance in aggregate ratings presents a challenge for future modeling tasks.
- **Descriptive Statistics**: Numerical and categorical variables revealed diverse pricing strategies and cuisine preferences, along with a concentration of restaurants in urban areas.
- **Geospatial Insights**: Geospatial analysis highlighted the uneven distribution of restaurant ratings and the prevalence of unrated or low-rated establishments.