Restaurant Dataset Analysis

Level 1

Task 1. Data Exploration and Preprocessing

1.1. Dataset Overview

- **Number of rows and Columns:**
    - The dataset comprises a total number of rows and columns. This dimension provides an understanding of the dataset's size and helps gauge its complexity. For instance, a large number of rows might indicate a comprehensive dataset, whereas many columns could suggest a rich dataset with various features.
- **Initial Data Preview:**
    - An initial review of the dataset's first few rows was conducted. This preview helps to get a sense of the data's structure, including the column names and the types of data stored. It provides a snapshot of how data is organized and highlights any immediate issues, such as inconsistent formatting.

1.2. Handling Missing Values

- **Missing Values Check:**
    - A thorough check for missing values in each column was performed. Missing values can affect the quality of analysis, so identifying their presence is crucial. The count of missing values for each column was determined to understand where data is lacking.
- **Handling Missing Values:**
    - Missing values in the 'Cuisines' column were addressed by filling them with the placeholder 'Unknown'. This approach allows the dataset to remain complete while indicating that some data was not provided.

1.3. Data Type Conversion

- **Data Types Review:**
    - The data types of columns were reviewed before and after conversion to ensure compatibility with analytical methods. Proper data types are essential for accurate calculations and visualizations. For example, numerical data should be in numeric formats, and categorical data should be in object formats.

1.4. Target Variable Analysis

- **Distribution of Target Variable ('Aggregate rating'):**
    - The distribution of the 'Aggregate rating' variable was analyzed to understand how ratings are spread across different values. A histogram with a kernel density estimate (KDE) curve was used to visualize this distribution, providing insights into the overall rating trends.
- **Class Distribution:**
    - The frequency of each class within the 'Aggregate rating' variable was examined to detect any class imbalances. Identifying such imbalances is crucial for ensuring that predictive models are not biased towards any particular class.

Task 2. Descriptive Analysis

2.1. Statistical Measures

- **Basic Statistics:**
    - Basic statistical measures, including mean, median, and standard deviation, were computed for numerical columns. These statistics provide insights into the central tendency (mean and median) and variability (standard deviation) of the data. This helps in understanding the overall distribution and spread of numerical values.
- **Median Values:**
    - The median values for numerical columns were calculated to provide a measure of the central value of each distribution. The median is less affected by outliers compared to the mean and thus offers a robust central value.

2.2. Distribution of Categorical Variables

- **Country Code Distribution:**
    - The distribution of the 'Country Code' variable was analyzed to show how many restaurants are present in each country. This helps in understanding the dataset's geographical coverage and identifying countries with the highest number of entries.
- **City Distribution:**
    - The distribution of restaurants across different cities was explored to highlight the top cities with the highest number of restaurants. This analysis identifies key urban areas with significant restaurant presence.
- **Cuisines Distribution:**
    - The distribution of different cuisines was analyzed to identify the most popular cuisines in terms of restaurant count. This information can be useful for understanding culinary trends and preferences.

Task 3. Geospatial Analysis

3.1. Restaurant Locations Map

- **Map Visualization:**
    - A map visualization of restaurant locations was created using latitude and longitude data. This map provides a visual representation of restaurant distribution and highlights areas with higher concentrations.
- **Heatmap and Markers:**
    - A heatmap was overlaid on the map to visualize areas with high restaurant density. Markers were added to the map for each restaurant, displaying details such as the restaurant's name, address, and aggregate rating. This visualization helps in understanding geographical patterns and clusters of restaurant activity.

3.2. Distribution Analysis

- **Distribution Across Cities:**
    - The distribution of restaurants across different cities was analyzed, focusing on the top 10 cities with the most restaurants. This helps in identifying urban areas with the highest restaurant density.
- **Distribution Across Countries:**
    - The distribution of restaurants across countries was visualized to show how restaurant counts vary globally. This analysis reveals which countries have the most significant number of restaurants.

3.3. Correlation Analysis

- **Correlation with Ratings:**
    - Scatter plots were used to explore the relationship between restaurant ratings and their latitude/longitude coordinates. These plots help in understanding whether there is any geographical pattern or trend related to ratings.
- **Correlation Matrix:**
    - A correlation matrix was computed to analyze the relationships between latitude, longitude, and aggregate ratings. This matrix provides insights into how location coordinates correlate with ratings.
- **Ratings Distribution by City and Country:**
    - Boxplots were created to show the distribution of ratings across the top 10 cities and countries. These plots highlight variations in ratings by location and can indicate whether certain locations have consistently higher or lower ratings.

Tools and Techniques Used

- **Data Exploration:**
    - Pandas was used for data manipulation and exploration, including handling missing values and checking data types.
- **Data Visualization:**

    - `Matplotlib` and `Seaborn` were employed for creating various plots, such as histograms, bar charts, and scatter plots, to analyze distributions and relationships within the dataset.
    - `Folium` was used to create interactive maps for geospatial analysis, including heatmaps and marker clusters.
- **Statistical Analysis:**
    - Basic statistical measures like mean, median, and standard deviation were calculated using `Pandas` methods to summarize numerical data.
- **Geospatial Visualization:**
    - `Folium`'s HeatMap and MarkerCluster plugins were utilized to visualize restaurant locations and density on maps.

Conclusion

The analysis of the restaurant dataset provided valuable insights into the distribution and characteristics of restaurants across different locations and cuisines.

Key findings include:

- **Geographical Distribution:** The data reveals significant restaurant densities in specific cities and countries, with visualizations highlighting geographical patterns and clusters.
- **Culinary Trends:** The most popular cuisines and the distribution of restaurants by cuisine were identified, offering a glimpse into current culinary trends and preferences.
- **Ratings Analysis:** The relationship between restaurant ratings and their geographical locations was explored, revealing insights into how ratings vary across different regions.

Overall, the analysis demonstrates the dataset's richness and offers actionable insights for stakeholders interested in understanding restaurant distribution, culinary trends, and geographic patterns.