Exploratory Data Analysis on Cuisine Rating Dataset

Author: Aditya Gupte

Institute: IIT Gandhinagar

Date: May 16, 2024

Introduction

The goal of this project is to perform an Exploratory Data Analysis (EDA) on the "Cuisine Rating" dataset, which contains information about customers' preferences for different cuisines, their demographic details, and ratings for food, service, and overall experience. EDA is a crucial step in any data analysis project as it helps to understand the data, identify patterns, and gain insights that can guide further analysis or modelling.

Dataset

The dataset used for this project is a CSV file named "Cuisine_rating.csv". It contains the following columns:

- User ID
- Area code
- Location
- Gender
- YOB (Year of Birth)
- Marital Status
- Activity
- Budget
- Cuisines
- Alcohol (consumption habits)
- Smoker (smoking habits)
- Food Rating
- Service Rating
- Overall Rating
- Often A S (Unknown column)

The dataset has 200 rows, representing 200 customers.

Methodology

The EDA process followed these steps: data loading and exploration, univariate analysis, bivariate analysis, multivariate analysis, and statistical summaries. The EDA process followed a structured methodology to comprehensively explore and analyze the dataset.

Firstly, the data loading and exploration phase involved importing the necessary libraries, reading the dataset into a pandas DataFrame, and performing initial checks on the data. This included examining the first few rows using the head() function and checking for missing values using the isnull().sum() function. These steps provided an initial understanding of the data structure and quality.

Secondly, univariate analysis was conducted to gain insights into the distribution and characteristics of individual variables in the dataset. For categorical variables such as location, gender, marital status, activity, and cuisine preferences, bar plots and count plots were created using matplotlib and seaborn to visualize the distribution of each category. This helped identify the most prevalent categories and any potential imbalances or skewness in the data. For numerical variables like budget and ratings, histograms and kernel density estimates (KDE) were plotted to visualize the distribution and identify any outliers or unusual patterns.

Thirdly, bivariate analysis was performed to explore the relationship between different variables in the dataset. Bar plots were created to visualize the mean overall rating for different categories of cuisine, activity, marital status, gender, and location. This allowed for the identification of any potential associations or trends between these variables and the overall rating. Additionally, a scatter plot matrix was created to visualize the pairwise relationships between different numerical features in the dataset, such as overall rating, food rating, service rating, and budget. Violin plots were used to compare the distribution of overall rating across different categories, and a heatmap was created to visualize the distribution of overall rating across different combinations of location and cuisine.

Results and Insights

The EDA process revealed several interesting insights about the dataset:

1. The dataset covers a diverse range of locations, with varying representations across different areas.
2. The gender distribution shows a slight skew towards females.
3. Most customers are either single or married, with a smaller portion being divorced.
4. The majority of customers are either students or professionals.
5. The most popular cuisines in the dataset are Japanese, Indian, and Chinese, followed by French, Italian, and Filipino.
6. Alcohol consumption habits vary widely, with a significant portion of customers being either non-drinkers or social drinkers.
7. Smoking habits are more polarized, with a larger portion of customers being either non-smokers or regular smokers.
8. Overall, the ratings for food, service, and overall experience are generally positive, with a slight skew towards higher ratings.
9. There are some noticeable patterns between overall rating and factors like cuisine, activity, marital status, gender, and location.
10. Certain combinations of location and cuisine seem to have higher or lower overall ratings.

All the graphs and plots are present in the python notebook. The report contains expalantion of steps undertaken while coding the project.

Conclusion

The Exploratory Data Analysis on the Cuisine Rating dataset provided valuable insights into customer preferences, behaviors, and satisfaction levels related to different cuisines and dining experiences. The visualizations and statistical summaries revealed patterns and relationships between various features, which can be useful for businesses in the food and hospitality industry to better understand and cater to their customers' needs and preferences.