PREDECTIVE ANALYTICS ASSIGMENT

MUHAMMAD ABDULLAH
28-FMS-BSBIN-F21
ASSIGMENT

ZOMATO DATA SET(A FOOD CHAIN)

The dataset is titled "Zomato", and it contains information about various restaurants, focusing on key aspects such as their names, locations, average ratings, price range, customer votes, and whether or not they offer online delivery and table booking services. This type of dataset is commonly used for analysis related to restaurant performance, customer preferences, and service availability.

Key Features of the Dataset:

1. Restaurant Name: The name of the restaurant.

2. City: The city where the restaurant is located.

3. Cuisines: The types of cuisines offered at the restaurant (e.g., Indian, Chinese, etc.).

4. Average Cost for Two: The average cost of a meal for two people at the restaurant.

5. Has Table booking: Indicates whether the restaurant offers table booking (Yes/No).

6. Has Online delivery:Specifies whether the restaurant provides online delivery (Yes/No).

7. Votes: The number of votes or ratings the restaurant has received from customers.

8. Aggregate rating: The average rating of the restaurant given by customers (ranging from 1 to 5).

9. Price range:The price category of the restaurant (on a scale from 1 to 4, with 1 being the lowest).

10. Is delivering now Indicates whether the restaurant is currently delivering food.

11. Switch to order menu: Indicates whether the restaurant's menu is available for online ordering.

Purpose of the Dataset:

The main goal of this dataset is to analyze various aspects of restaurants, such as services, pricing, and customer ratings, to determine which restaurants are more favorable or preferred by customers. By using this dataset, various analytical models can be built, such as linear regression, clustering, or other machine learning techniques to evaluate restaurant performance.

Potential Analyses:

1. Analyzing the Relationship Between Customer Ratings and Votes:You could examine whether restaurants that receive more customer votes tend to have higher average ratings.

2. Impact of Table Booking or Online Delivery on Ratings: Analyze whether restaurants offering online delivery or table booking services tend to have higher customer ratings.

3. Relationship Between Price and Rating: Compare the price range and customer ratings to see if cheaper restaurants have higher ratings or if higher-priced restaurants are more highly rated.

Usefulness of the Dataset:

- Performance Analysis: This dataset can help analyze restaurant performance based on customer feedback.

- Preference Analysis: It can be used to identify what services customers value more (e.g., table booking or online delivery).

- Price Sensitivity: You can analyze how price affects customer satisfaction and the overall rating of the restaurant.

Overall, this dataset provides valuable insights into restaurant services, pricing strategies, and customer preferences, which can be used to guide business decisions or predict customer behavior.

CORRELATION TELLING

The correlation matrix shows the relationship between the different numerical features in the dataset. Each cell represents the correlation between two features. The darker the red, the more strongly the features are positively correlated. The darker the blue, the more strongly the features are negatively correlated.

Here's a breakdown of the notable correlations:

Restaurant ID and Country Code: There's a weak negative correlation. This suggests that restaurants with higher IDs are less likely to be in specific countries.

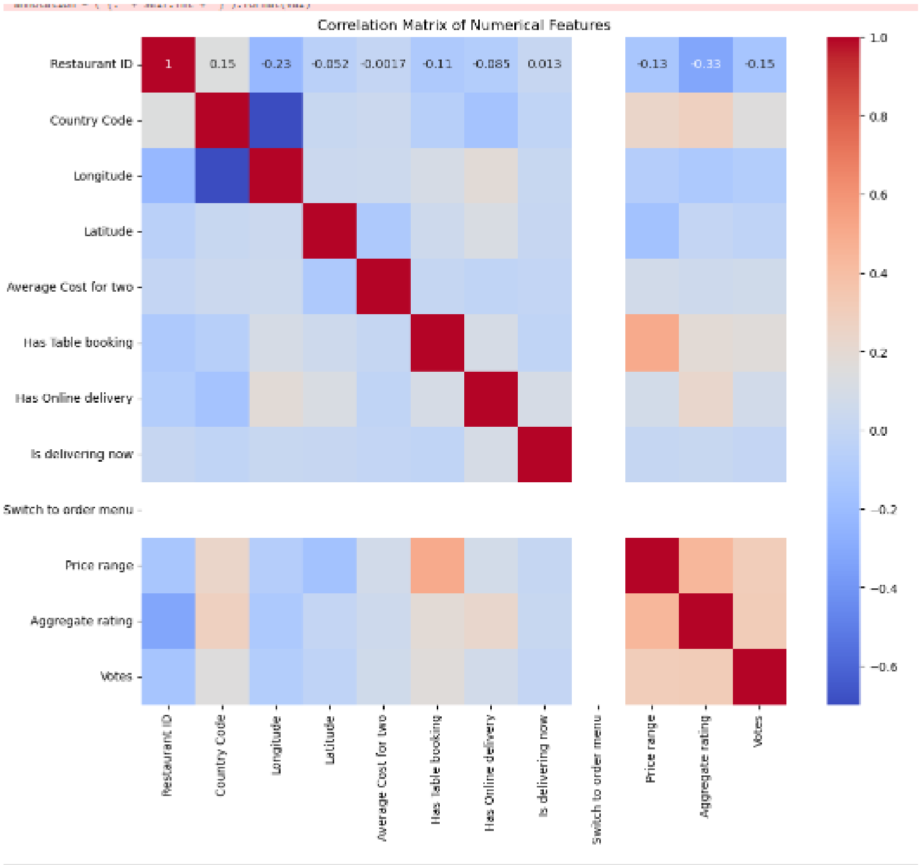
Average Cost for Two and Aggregate Rating: There's a moderate positive correlation. This suggests that more expensive restaurants tend to have higher aggregate ratings. This could be due to higher quality food or service associated with higher prices.

Aggregate Rating and Votes: There's a strong positive correlation. This is expected since restaurants with higher ratings tend to receive more votes. This suggests popularity and good service are closely linked.

Price Range and Votes: There's a moderate positive correlation. This suggests restaurants with higher price ranges receive more votes. This is likely because people are more likely to vote for restaurants they consider a good value for their money, or are more likely to vote for a restaurant they spend a significant amount at.

Price Range and Aggregate Rating: There's a moderate positive correlation. This suggests that higher-priced restaurants have higher aggregate ratings. This could be related to higher quality food or service associated with higher prices.

The remaining correlations are weaker, indicating less strong relationships.



BOX PLOT REPRESENTATION

The box plots represent the distribution of numerical features in a dataset related to restaurants. Here's a breakdown of each attribute:

Restaurant ID: A unique identifier for each restaurant. The box plot shows that the ID values are evenly distributed within a certain range, with no significant outliers.

Country Code: Represents the country where the restaurant is located. The box plot indicates that there's one outlier value with a very high country code, while the majority of restaurants are clustered around a lower code range.

Longitude: The geographical longitude of the restaurant. The box plot displays a wide range of longitude values with a few outliers on the lower end.

Latitude: The geographical latitude of the restaurant. The box plot showcases a cluster of latitude values with a few outliers on both the lower and upper ends.

Average Cost for Two: The average cost of a meal for two people at the restaurant. The box plot reveals a significant number of outliers with very high average costs, suggesting that the data likely contains restaurants with various price points.

Has Table Booking: A binary feature (0 or 1) indicating whether the restaurant offers table bookings. The box plot shows a very tight distribution of values around 0 and 1, indicating that most restaurants either do or don't offer table bookings.

Has Online Delivery: Another binary feature (0 or 1) signifying if the restaurant provides online delivery services. The box plot suggests that the majority of restaurants offer online delivery, with very few outliers.

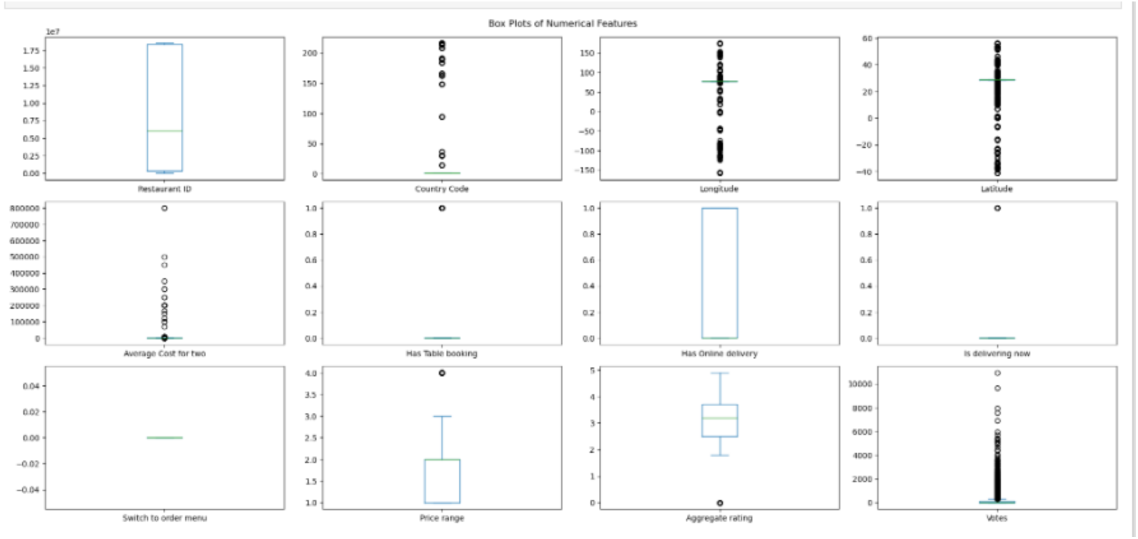
Is Delivering Now: A binary feature (0 or 1) showing if the restaurant is currently offering delivery. The box plot indicates a cluster of values around 0 and 1, implying that delivery availability is common but also experiences fluctuations.

Switch to Order Menu: This attribute is unclear without additional context. Its box plot shows a very tight distribution around 0, potentially suggesting a binary feature with most values being 0.

Price Range: A categorical feature indicating the price range of the restaurant. The box plot reveals that most restaurants fall into a specific price range, with a few outliers on the higher end.

Aggregate Rating: The average rating of the restaurant. The box plot displays a cluster of ratings around a specific value, with a few outliers at the higher end.

Votes: The number of votes the restaurant has received. The box plot shows a significant number of outliers with a high number of votes, indicating that the dataset contains popular



HISTROGRAM REPRESENTATION

The histograms show the distribution of the numerical features in the dataset. Here's a breakdown of each attribute:

Restaurant ID: This feature appears to have a unique identifier for each restaurant. The histogram shows a large number of restaurants with IDs clustered around a particular range, with fewer restaurants at higher ID values.

Country Code: The histogram suggests that the majority of restaurants are located in a single country (represented by the peak around 0). There are a few restaurants in other countries, but their representation is much smaller.

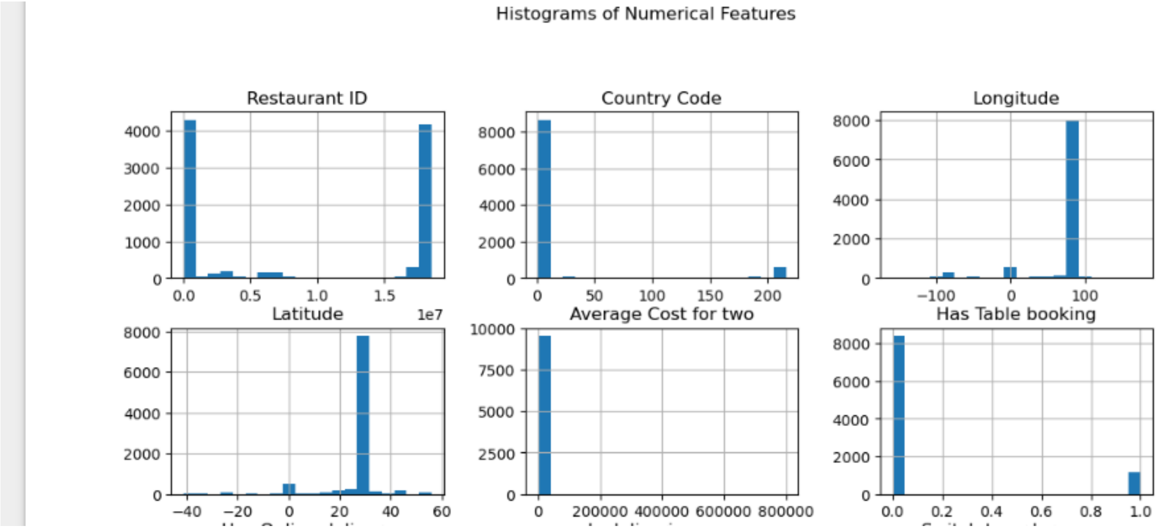
Longitude: The histogram displays a strong peak around a particular longitude value, indicating that most of the restaurants are located in a specific geographical area.

Latitude: The histogram shows a similar pattern to Longitude, with a peak around a specific latitude value. This further confirms that the data likely focuses on a particular region.

Average Cost for two: The histogram shows that most restaurants have an average cost for two that falls within a particular price range. The data is skewed towards lower prices.

Has Table Booking: The histogram indicates that a majority of restaurants offer table booking, but there is a small minority of restaurants that do not.

Overall, the histograms provide valuable insights into the distribution of these numerical features and suggest that the data likely represents a specific geographic region with a particular price range. The data set may be biased toward restaurants that are more commonly found in that region.



Summary of Visuals:

1. **Histograms** – Show the distribution of numerical data.
2. **Correlation Matrix (Heatmap)** – Show relationships between numerical variables.
3. **Box Plots** – Highlight the presence of outliers and visualize data spread.
4. **Actual vs Predicted Scatter Plot** – Show how well the Linear Regression model performed.

5. **Residual Plot –** Highlight errors in the model's predictions.