REPORTING AND VISUALIZATION OF FIRST MONTH TASK

PRESENTED BY,

Name: SINU S MARIAM

Designation: Machine Learning Intern

Organization: COGNIFYZ TECHNOLOGIES

Batch Date: 23/06/2024 to 23/09/2024

Third Month Task: Report of First- and Second-Month Task

TASK 1 - PREDICT RESTAURANT RATINGS

1.1 SUMMARY OF EXPLORATORY DATA ANALYSIS:

The following are the key findings obtained during Data Analysis:

- Total Number of Restaurants Analysed: 9551
- Distinct Number of Restaurant Brands Analysed: 7437
- Number of Countries where Restaurants are spread: 15
- Number of Cities where Restaurants are spread: 140
- Number of Cuisines/ Combination of Cuisines analysed: 1825

The below figure shows the screenshot of Tableau visualization of analysis of data regarding Restaurant Information:

The link of complete Viz of Tableau Public is :
https://public.tableau.com/app/profile/sinu.s.mariam/viz/Restaurant_analysis_Ratings_Dashboard/Dashboard1?publish=yes

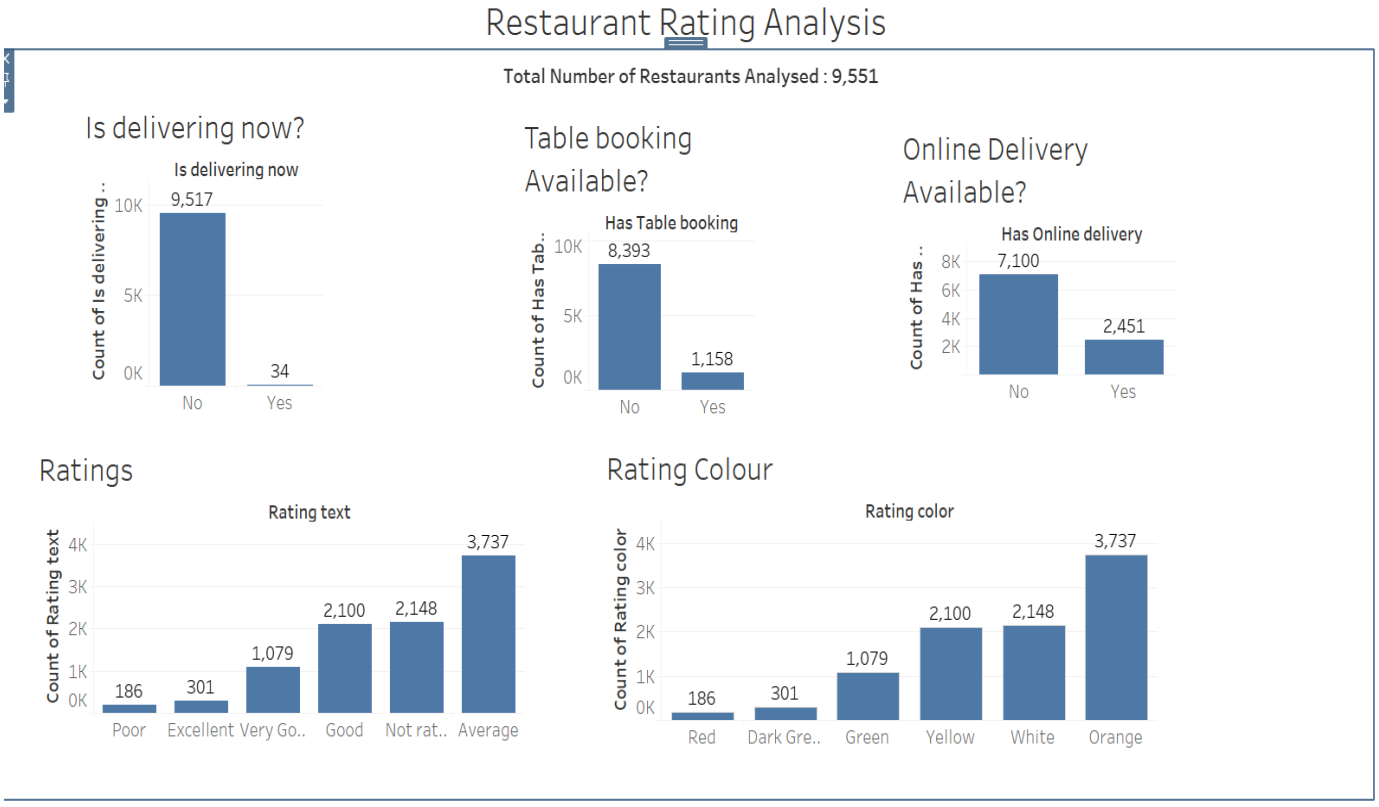

Fig 1.1: Screenshot of TABLEAU Public Visualization Repor

1.2 DATA INSIGHTS

1.2.1 HEAT MAP

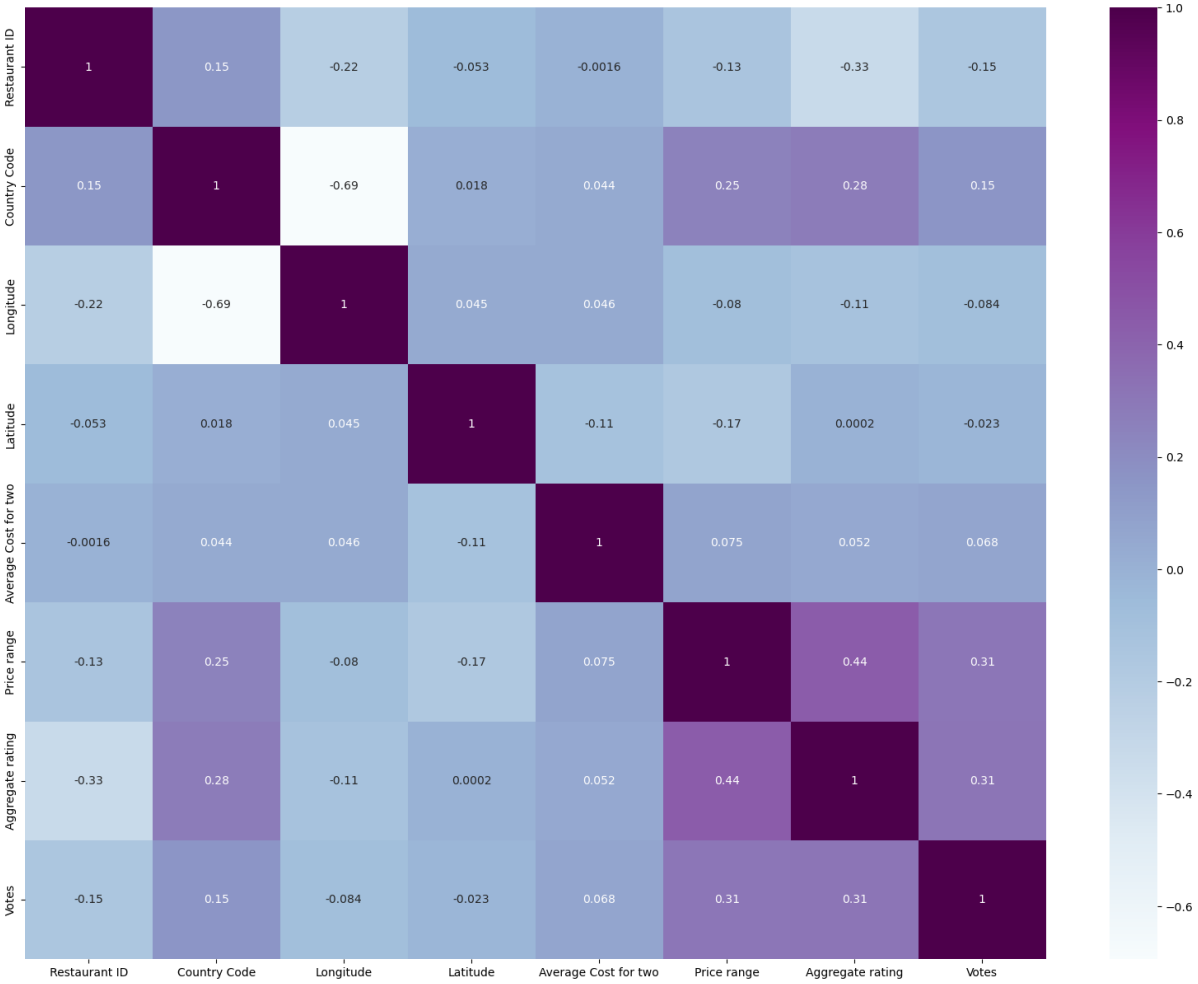

Fig 1.1 : Heat Map of Feature Variables Affecting the Prediction of Target Variable 'Aggregate rating'

- Variables Price range, Votes and Country Code have better correlation values (above 0.2) with target 'Aggregate rating' variable
- Factor 'Average Cost for two' is having low value (0.052) with target 'Aggregate rating' variable
- Latitude and Longitude have very low correlation or negative correlation with target 'Aggregate rating' variable
- Columns 'Longitude' and 'Latitude' can be dropped because of their very low correlation values with target variable ‘Aggregate rating'.
- Also, there are details from 15 countries (15 different country codes are available), and geographical data can be taken into consideration from columns 'City' and 'Locality.

- The column 'Address' can be dropped because its content is available in columns such as 'City' and 'Locality'
- The column 'Locality Verbose' can be dropped since its entries are a combination of columns 'City' and 'Locality'
- The column 'Restaurant ID' can be dropped because it does not have relevant Info
- The column 'Switch to order menu' can be dropped because it has only 1 unique value
- The column 'Rating text' and 'Rating color' have same no: of Unique values. So, either one can be dropped.

1.2.2 PLOTS OF DISTRIBUTION OF VARIABLES

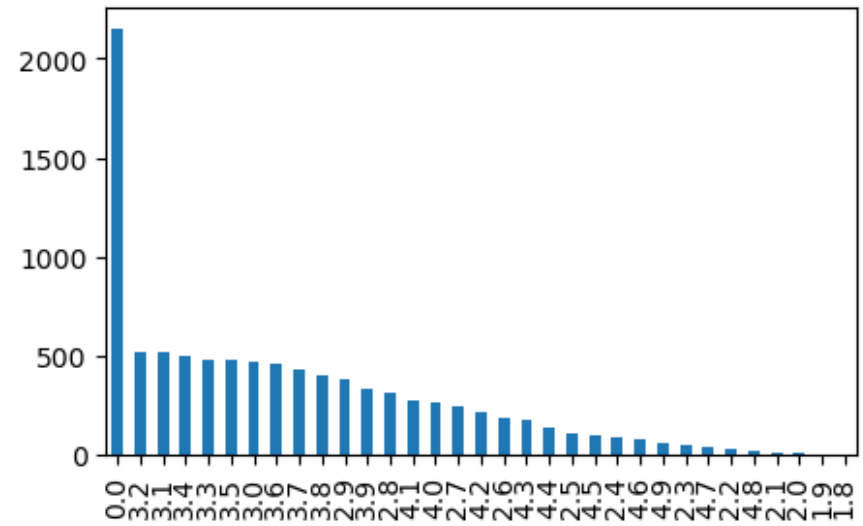

Fig 1.2: Histogram of Target variable 'Aggregate rating'

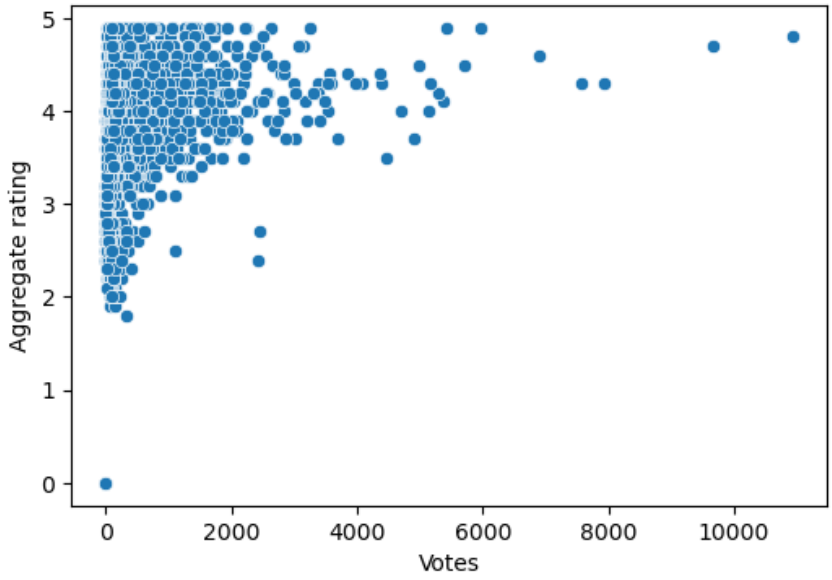

Fig 1.3: Scatter Plot of Target variable 'Aggregate rating' and Feature variable ‘Votes’

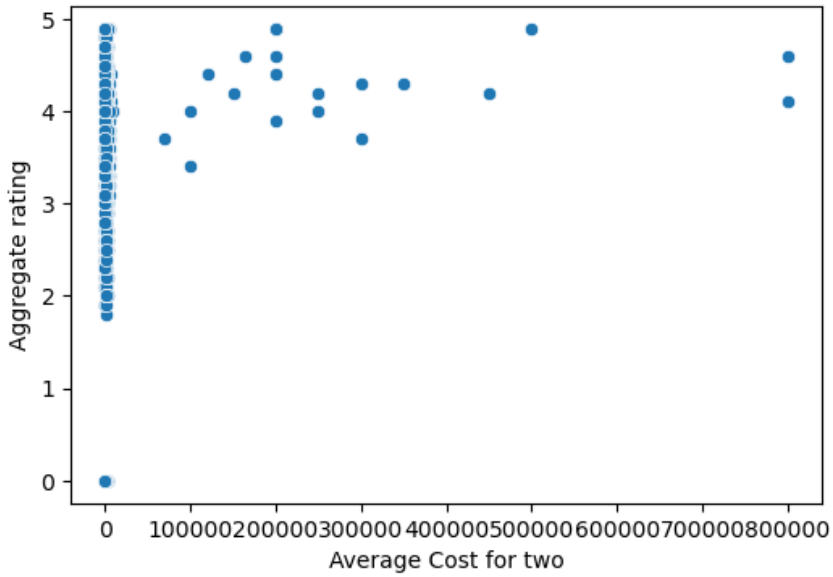

Fig 1.4: Scatter Plot of Target variable 'Aggregate rating' and variable ‘Average Cost for two’

- From Histogram of Target variable 'Aggregate rating' (Fig 1.2), it can be seen that a large number of restaurants obtained ZERO Ratings.
- From Scatter Plot of Target variable 'Aggregate rating' and Feature variable ‘Votes’ (Fig 1.3), it is noted that there is evidently good linear relation between them except for the cluster seen for entries where Votes are given as ZERO value.
- From Scatter Plot of Target variable 'Aggregate rating' and variable ‘Average Cost for two’ (Fig 1.4), it can be seen that there is somewhat a linear relation between them for higher values.

1.3 POTENCIAL ASPECTS IDENTIFIED DURING DATA ANALYSIS

- Linear Regression Model and Decision Tree Model were tested for prediction of Target variable 'Aggregate rating'.
- When analysed with predictive models such as Linear regression, Decision Tree for prediction of Target variable 'Aggregate rating', Mean Squared Error and Coefficient of Determination or R Squared Value (r2) were determined for model evaluation.
- Decision Tree Regression has high Coefficient of Determination or R Squared Value (0.985) compared to that of Linear Regression Model
- Decision Tree Regression also has very low Mean Squared Error (0.032) compared to that of Linear Regression Model.

TASK 2 - RESTAURANT RECOMMENTATION

- Create a restaurant recommendation system based on user preferences.
- Test the recommendation system by providing sample user preferences and evaluating the quality of recommendations.

2.1: POTENTIAL ASPECTS CONSIDERED DURING DATA ANALYSIS

- Here Recommendation system is developed using Content-Based Filtering System.
- Dataset does not contain user information, So Restaurant recommendation system is based on features of restaurants themselves.
- We are trying to Recommend restaurants that have similar 'Cuisine Types' with highest ratings in a preferred City.
- All columns except -'Cuisines', 'City','Aggregate rating', 'Rating text' can be dropped for analysis.
- Apply TF-ID (Term Frequency-Inverse Document Frequency) as a part of Natural Language Processing and extract features from text data using TF-IDF.
- To compute similarity between restaurants based on features 'Cosine Similarity metrics' is used here.
- For Sample Results We are giving a restaurant name as input and the model Recommends restaurants that have similar 'Cuisine Types' with highest ratings in the city where the Input restaurant resides.

- Sample Result:

```
# Print Recommendation for Restauarant similar to 'Rasoi Ghar'
print(get_recommendations("Rasoi Ghar", cos_dis, data_frame))
```

	Restaurant Name	Cuisines	City	Aggregate rating
598	Grand Barbeque Buffet Restaurant	Indian, Asian	Dubai	4.4
586	Rasoi Ghar	Indian	Dubai	4.3
590	Carnival By Tresind	Indian	Dubai	4.9
597	Tresind - Nassima Royal Hotel	Indian	Dubai	4.9

```
# Print Recommendation for Restauarant similar to 'Y Cafe & Restaurant'
print(get_recommendations("Y Cafe & Restaurant", cos_dis, data_frame))
```

	Restaurant Name	Cuisines	City	Aggregate rating
864	Kalsang AMA Cafe	Cafe	Dehradun	4.2
854	Y Cafe & Restaurant	Cafe	Dehradun	4.0
859	Razzmatazz	Cafe	Dehradun	3.9
856	First Gear Cafe	Cafe, Chinese	Dehradun	3.9

Fig 2.1: Sample Result for Restaurant Recommendation System

TASK 3 - RESTAURANT RECOMMENTATION

- Develop a machine learning model to classify restaurants based on their cuisines.

3.1: POTENTIAL ASPECTS IDENTIFIED DURING DATA ANALYSIS

- To develop a machine learning model to classify restaurants based on their cuisines, we need only 2 columns for analysis - 'Restaurant Name' and 'Cuisines'
- The variable 'Cuisines' has 1825 unique values in column.

```
Name of the variable : Cuisines
North Indian                                              936
North Indian, Chinese                                     511
Chinese                                                   354
Fast Food                                                 354
North Indian, Mughlai                                     334
                                                         ...
Bengali, Fast Food                                          1
North Indian, Rajasthani, Asian                             1
Chinese, Thai, Malaysian, Indonesian                        1
Bakery, Desserts, North Indian, Bengali, South Indian       1
Italian, World Cuisine                                      1
Name: Cuisines, Length: 1825, dtype: int64
```

Fig 3.1: Unique values and its count in variable 'Cuisines'

- 'Cuisines' are Label Encoded and Extraction of features are done from text data using TF-IDF as part of Natural Language Processing (NLP).
- Linear Regression Model and Decision Tree Model were tested for prediction of Target variable 'Aggregate rating'.
- Random Forest Model obtained better Accuracy, Precision, Recall and f1 score compared to Logistic Regression Model.