COGNIFYZ DATA SCIENCE INTERNSHIP LEVEL 3 REPORT

About the Level

Level 3 of the Cognifyz Data Science Internship focuses on the following tasks:

1. Predictive Modelling
2. Customer Preference Analysis, and
3. Data Visualization.

Task 1: Predictive Modelling

- Build a regression model to predict the aggregate rating of a restaurant based on available features.
- Split the dataset into training and testing sets and evaluate the model's performance using appropriate metrics.
- Experiment with different algorithms (e.g., linear regression, decision trees, random forest) and compare their performance.

Task 2: Customer Preference Analysis

- Analyse the relationship between the type of cuisine and the restaurant's rating.
- Identify the most popular cuisines among customers based on the number of votes.
- Determine if there are any specific cuisines that tend to receive higher ratings.

Task 3: Data Visualization

- Create visualizations to represent the distribution of ratings using different charts (histogram, bar plot, etc.).
- Compare the average ratings of different cuisines or cities using appropriate visualizations.
- Visualize the relationship between various features and the target variable to gain insights.

RESULTS

Task 1: Predictive Modelling

Four different regression models were built namely Linear Regression, Decision Tree, Random Forest and Support Vector Machine to predict the aggregate rating of a restaurant

Matthew Adenyo | adenyomatthew@gmail.com | linkedin.com/in/matth-adenyo

based on the available features. The features are: Average Cost for two, Votes, Price range, Has Table Booking and Has Online Delivery. The performance of the models was evaluated and their rmse and r-squared are recorded in the table below for comparison.

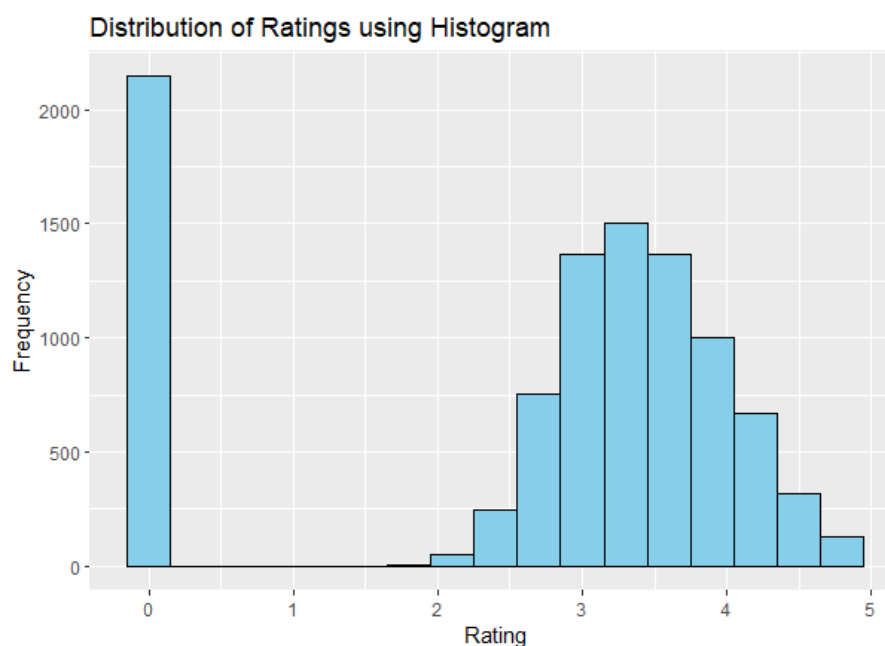| Models	rmse	r-squared
Linear Regression	1.294572	0.267201
Decision Tree	0.3984173	0.9308097
Random Forest	0.3363971	0.950567
Support Vector Machine	1.458372	0.2205025

From the table above, Decision Tree and Random Forest perform well as they have a lower rmse and r-squared closer to 1. However, Random Forest Algorithm outperform all the models making it the preferred choice for this prediction.
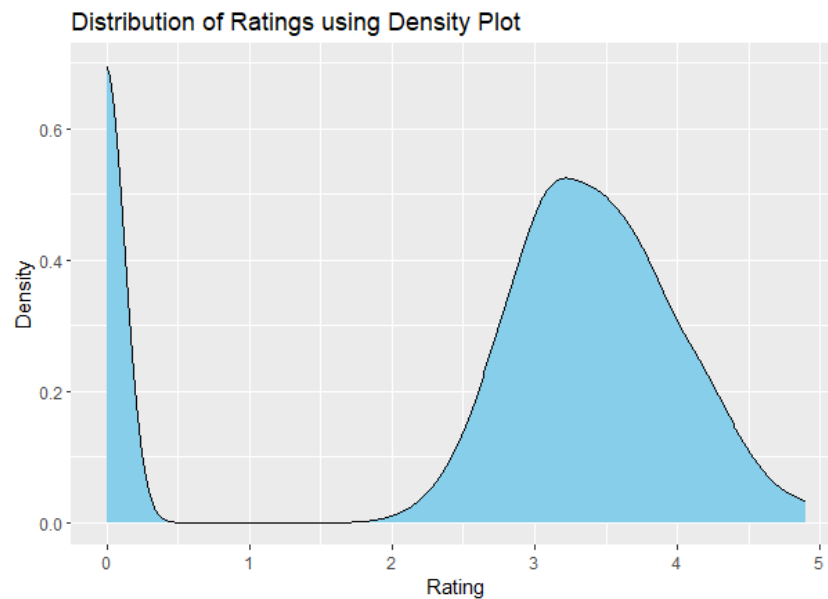
Task 2: Customer Preference Analysis

Based on the number of votes, North Indian, Mughlai and Chinese cuisines are the most popular cuisines. Also, most cuisines such as American, BBQ, Sandwich, Burger, Grill Caribbean, Seafood, Coffee and Tea among others have the same average rating of 4.9.

Task 3: Data Visualization

The majority of the rating given is between 3-4.

Distribution of Ratings using Histogram

Matthew Adenyo | adenyomatthew@gmail.com | linkedin.com/in/matth-adenyo

Distribution of Ratings using Density Plot

Based on the average rating, Inner City is the most rated city followed by Quezon City, Makati City, Pasig City, Mandaluyong City and Beechworth.

Average Ratings by City (Top 10)
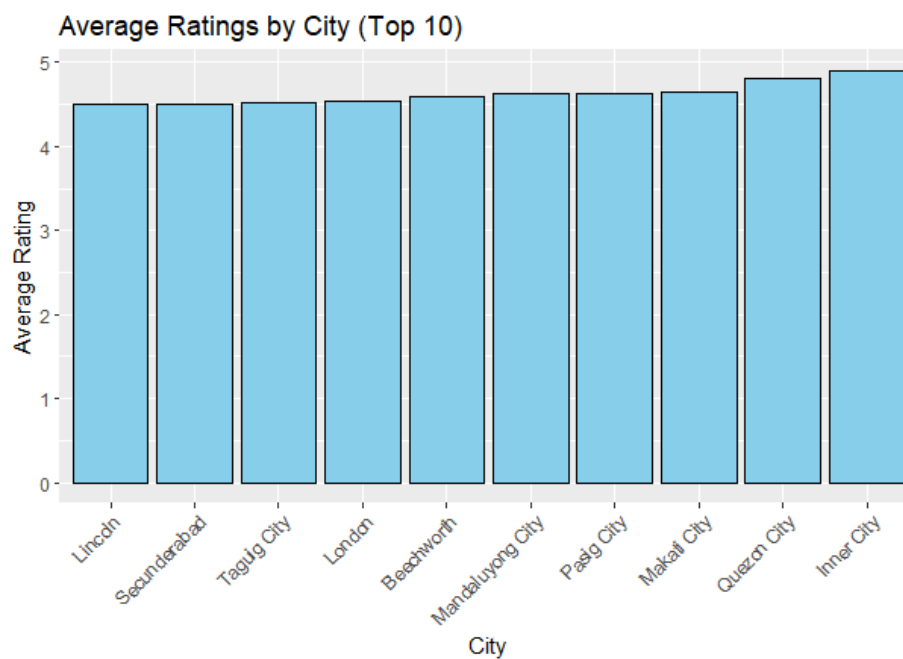
Also, Aggregate rating has a positive correlation with Votes, Price range, Has Table Booking and Has Online Delivery. Among them all, Price range has a strong positive correlation.

Correlation Matrix
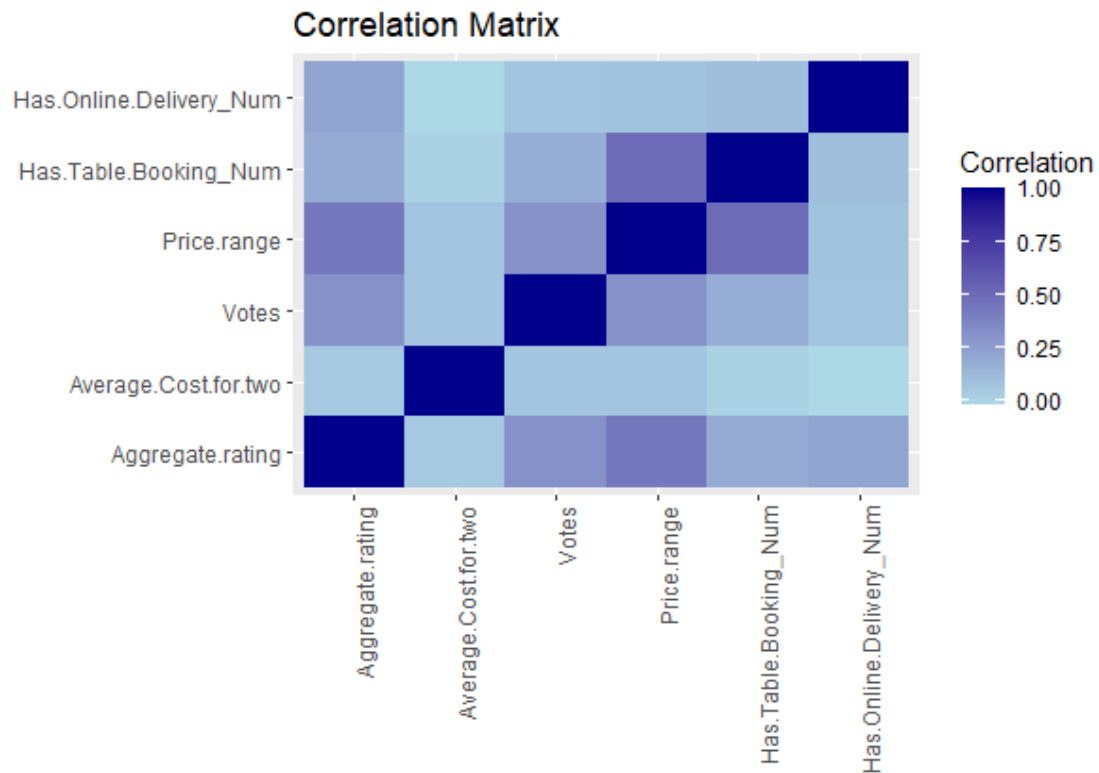
Conclusion

This segment of the project has highlighted the critical role of predictive modelling, customer preference analysis, and data visualization in uncovering actionable insights and driving strategic decision-making.

The customer preference analysis has served as a cornerstone in understanding the needs and preferences of the target audience.

Also, the utilization of data visualization techniques has helped to communicate complex insights in a clear, concise, and impactful manner.