DATA SCIENCE IN RESTAURANT ANALYTICS CAPSTONE PROJECT

Gabriel Amao
April 2025

Project Objective

- Analyze restaurant data to identify customer preferences
- Understand the impact of cuisine, cost, and services on ratings.
- Build Machine Learning models to predict restaurant ratings

Dataset Overview

Source:
(https://raw.githubusercontent.com/Oyeniran20/axia_class_cohort_7/refs/heads/main/Dataset%20.csv)
CSV dataset of restaurant reviews (9,551 records)

Key columns: Cuisines, City, Average Cost, Rating, Votes, Services

Target variable: Aggregate rating

9 Missing values in the Cuisines column
2,148 restaurant entries with a rating of 0.0 (treated as 'Not Rated')

Dataset Summary

There was a total of 9,551 restaurant entries across multiple countries

The summary of the features of the dataset includes: Cuisines, Cost, Ratings, Location, Services

I found 2,148 restaurants to have a rating of (0.0 rating). I treated them as NaN (Unrated)

There were only 9 missing values in the 'Cuisines' column

Methodology

I began by cleaning the data. Here ,missing values were removed for the exploratory analysis.

The dataset was explored using graphs to observe the patterns.

For better analysis, I changed some columns into more useful formats ("feature engineering").

Finally, I built the models to predict a restaurant’s rating based on the data provided.

EXPLORATORY DATA ANALYSIS

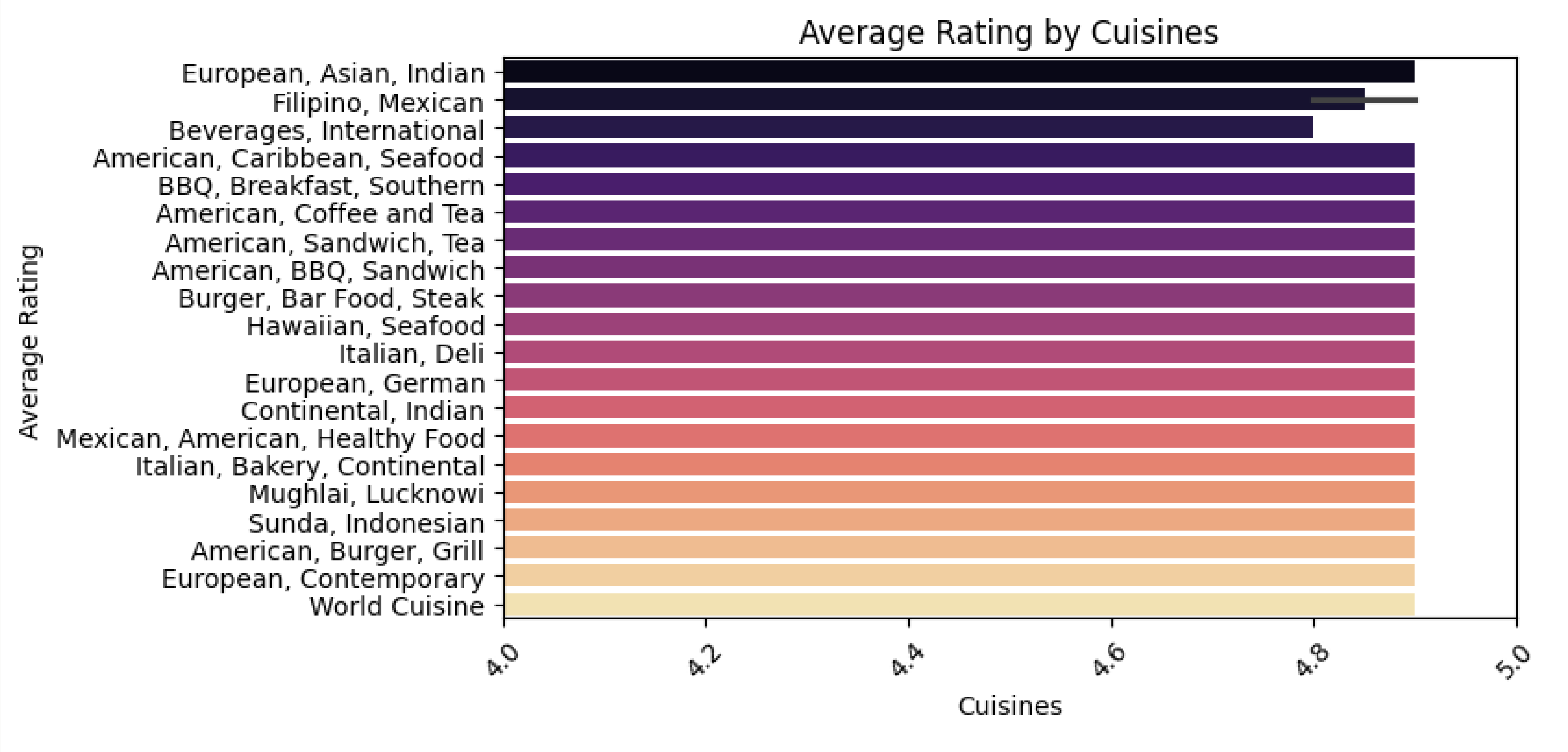
Visual Insights From the Ratings Distribution

- The aggregate ratings ranged from 0.0 to 5.0.

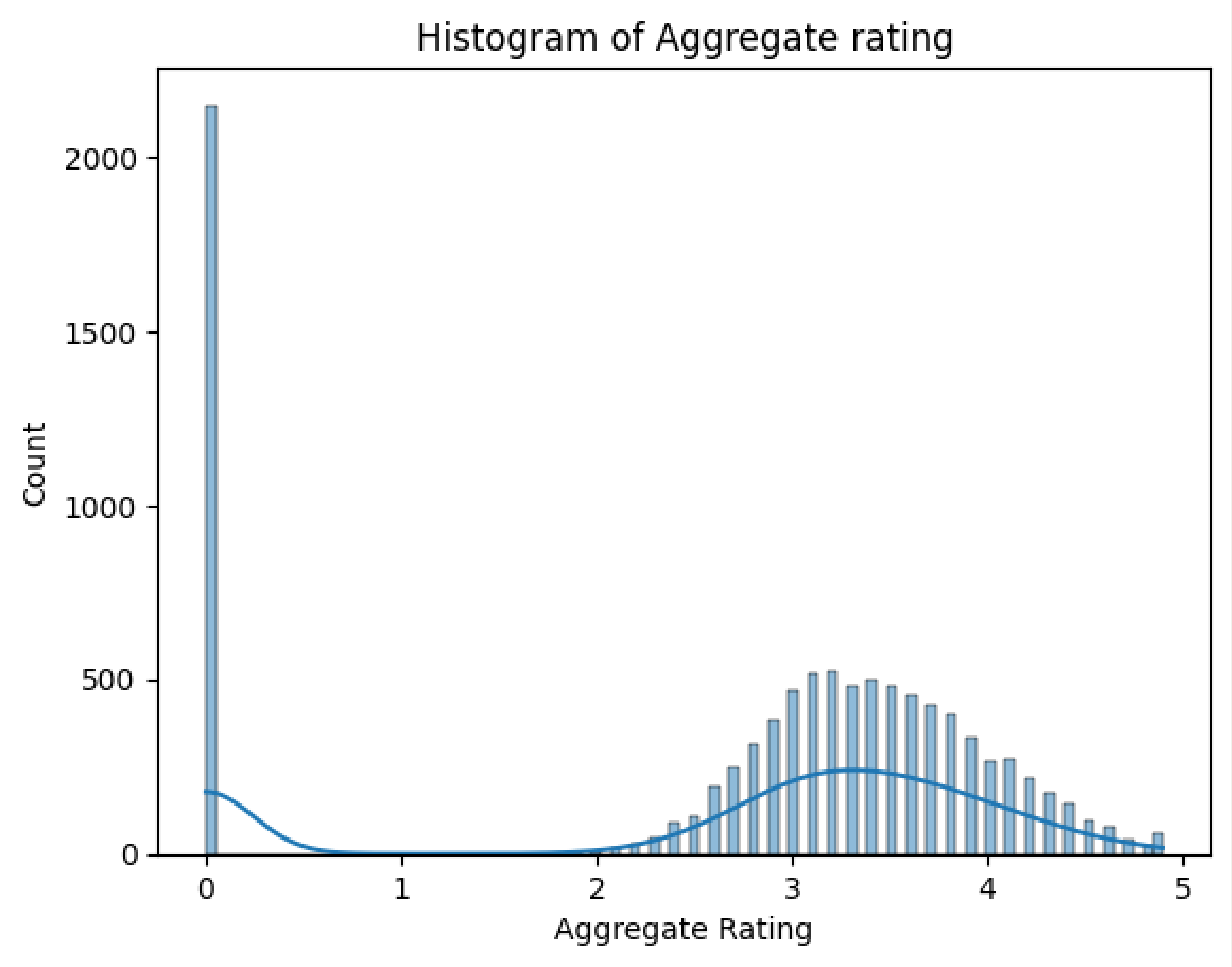
- However, the bar chart showed most ratings clustered between 3.5 and 4.5 values.

- A good number of restaurants were rated 0.0. Iclassified them as unrated entries.

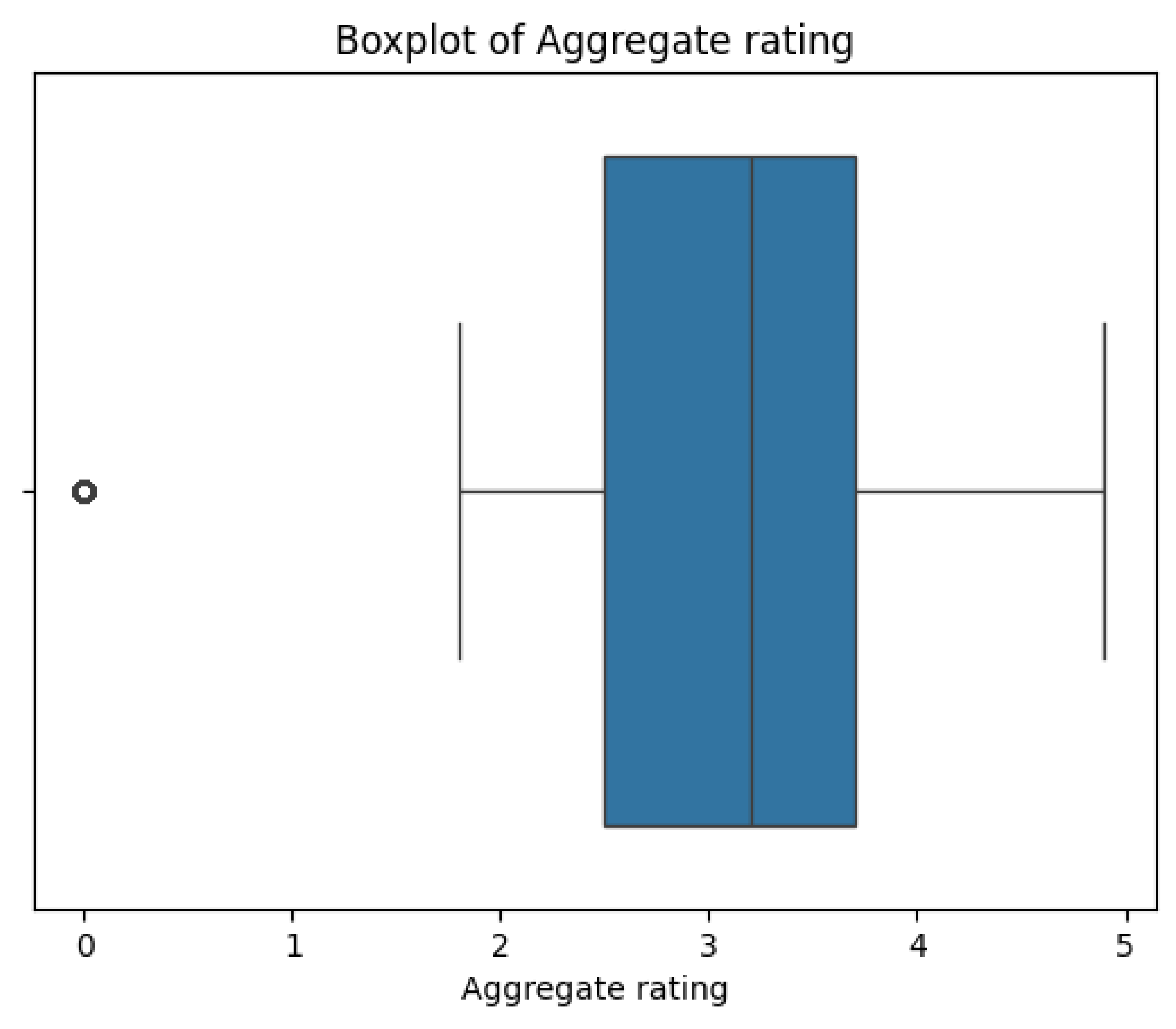
Average Rating By Cuisines



Histogram of Aggregate rating



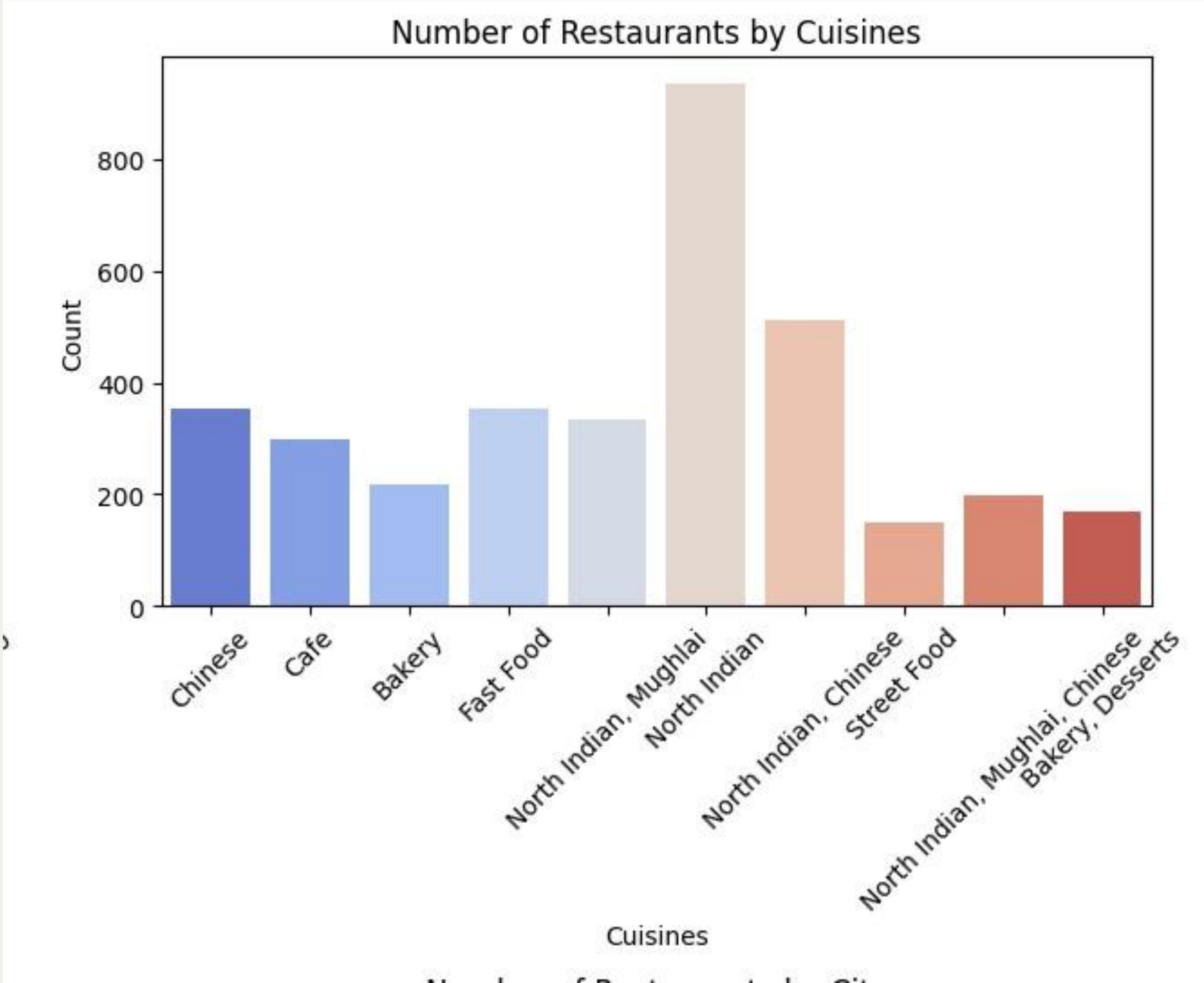
Boxplot of Aggregate rating



Cuisine & Cost Analysis

- After filtering the cuisines by votes (votes ≥50), The Top-rated were: Italian, Chinese, South Indian
- North Indian was the most popular cuisines but had an average rating
- Grouped Average Cost for two into Low, Medium, High, Luxury

Number of Restaurants by Cuisines



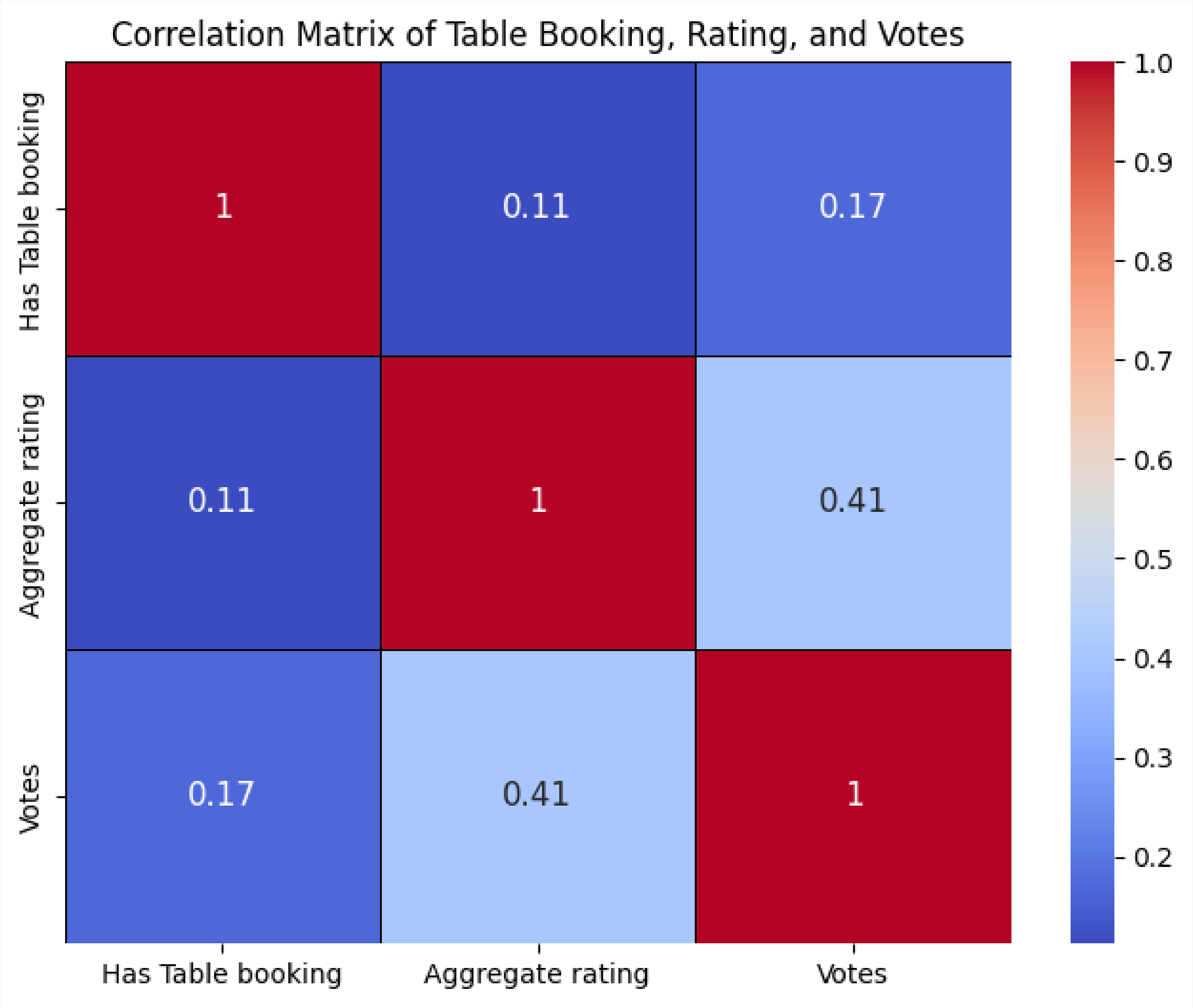
Insights on Customer Preferences

Services like table booking and delivery slightly improve ratings

Restaurants with more votes often had better ratings

High-cost restaurants had higher ratings

if they can be improved, there is an opportunity for popular but lower rated cuisines


Correlation Matrix of Table Booking, Rating, and Votes
Has Table booking
Aggregate rating
Votes
1
0.11
0.17
0.11
1
0.41
0.17
0.41
1
Has Table booking
Aggregate rating
Votes
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2

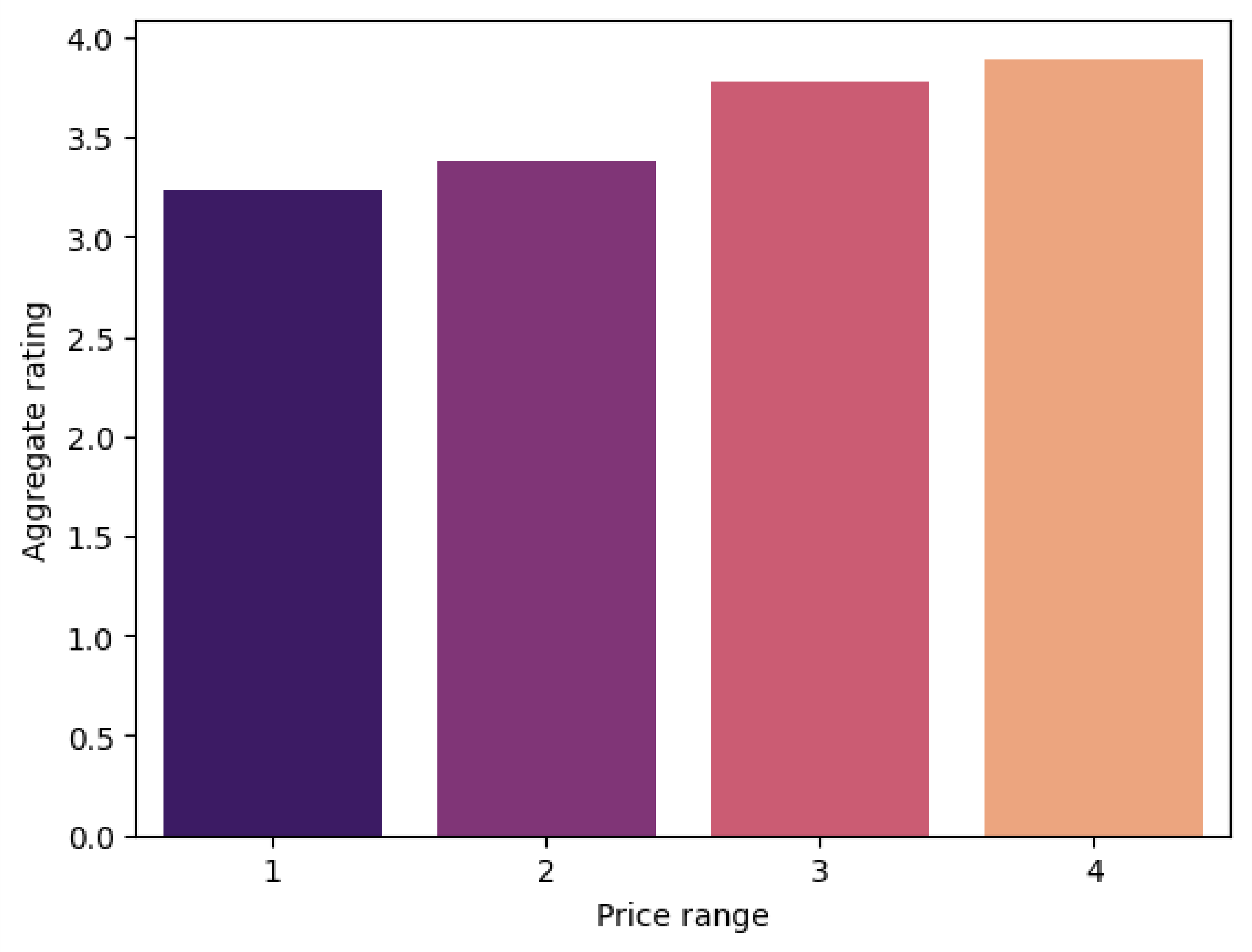
Cost Effect on Ratings

- Restaurants with a higher price range had better ratings.

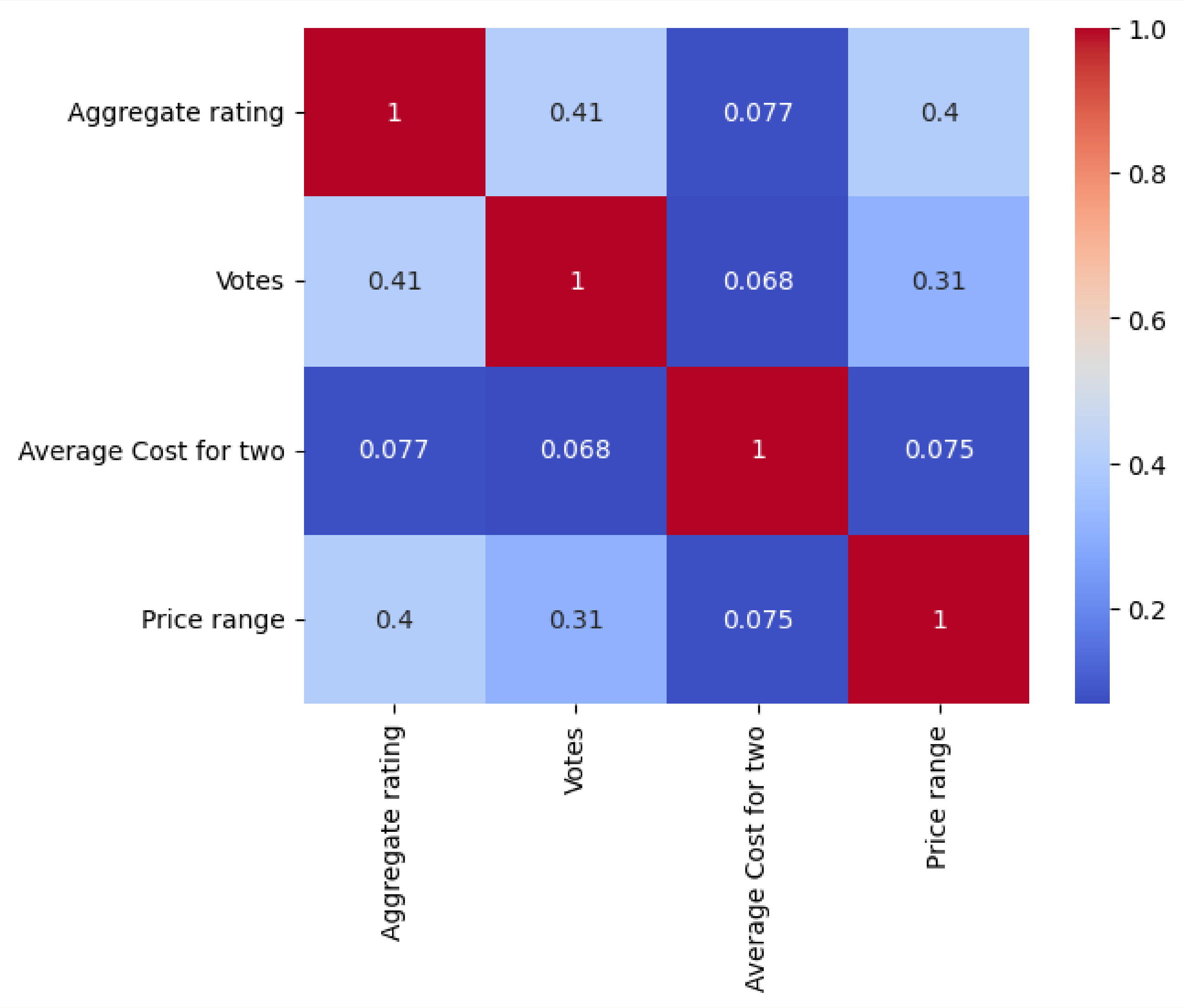
- Luxury restaurants are the best rated restaurants. Thus, we can see that customers prefer restaurants with better services even if the prices are slightly higher.

- There was a weak correlation between Votes and prices.

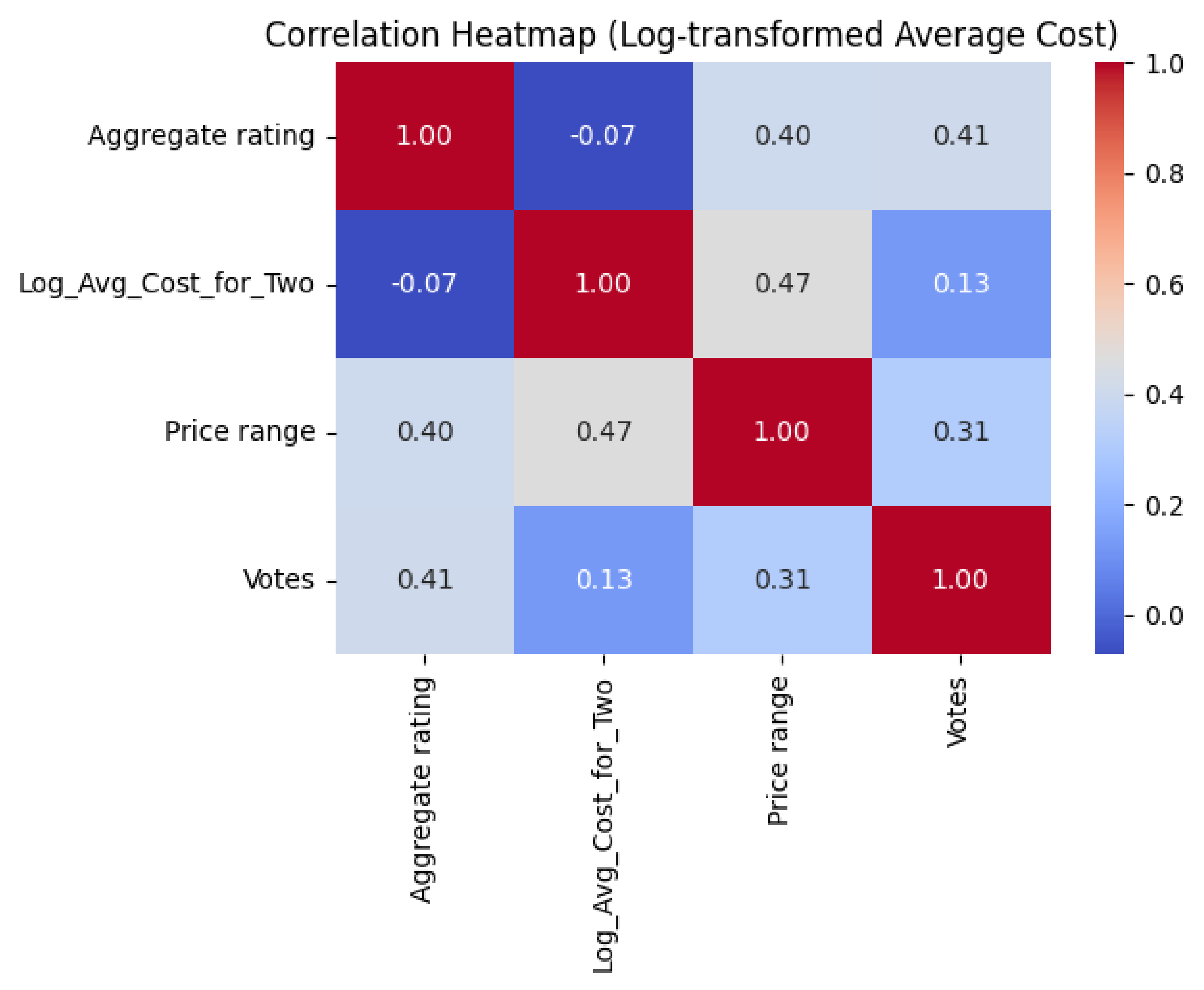
Histogram Showing Price Range vs Aggregate Rating



Correlation Heatmap showing Average Cost



Correlation Heatmap showing the Log tansformed Average Cost



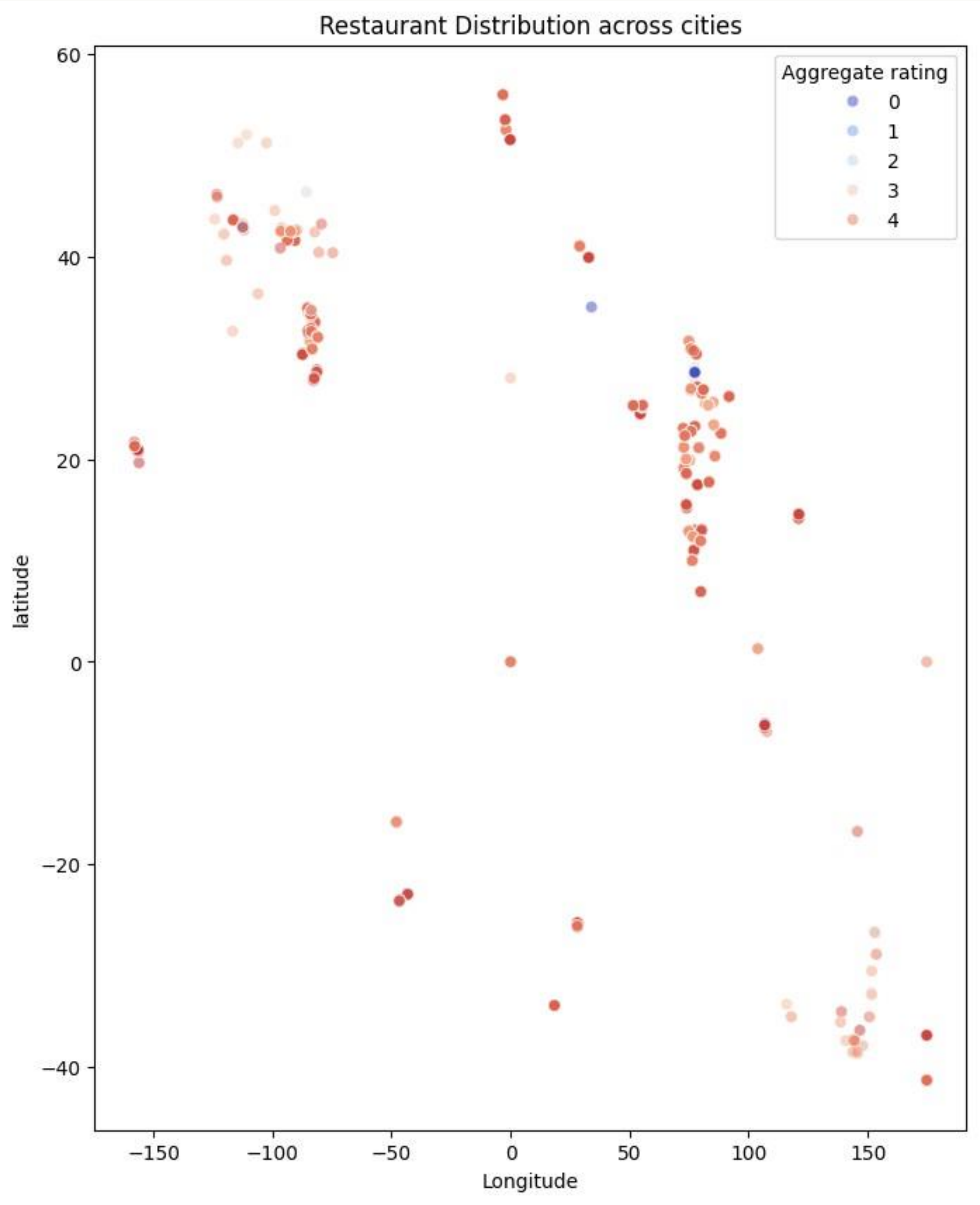
Geospatial Analysis

The Latitude and Longitude of each restaurant was used to map their clusters

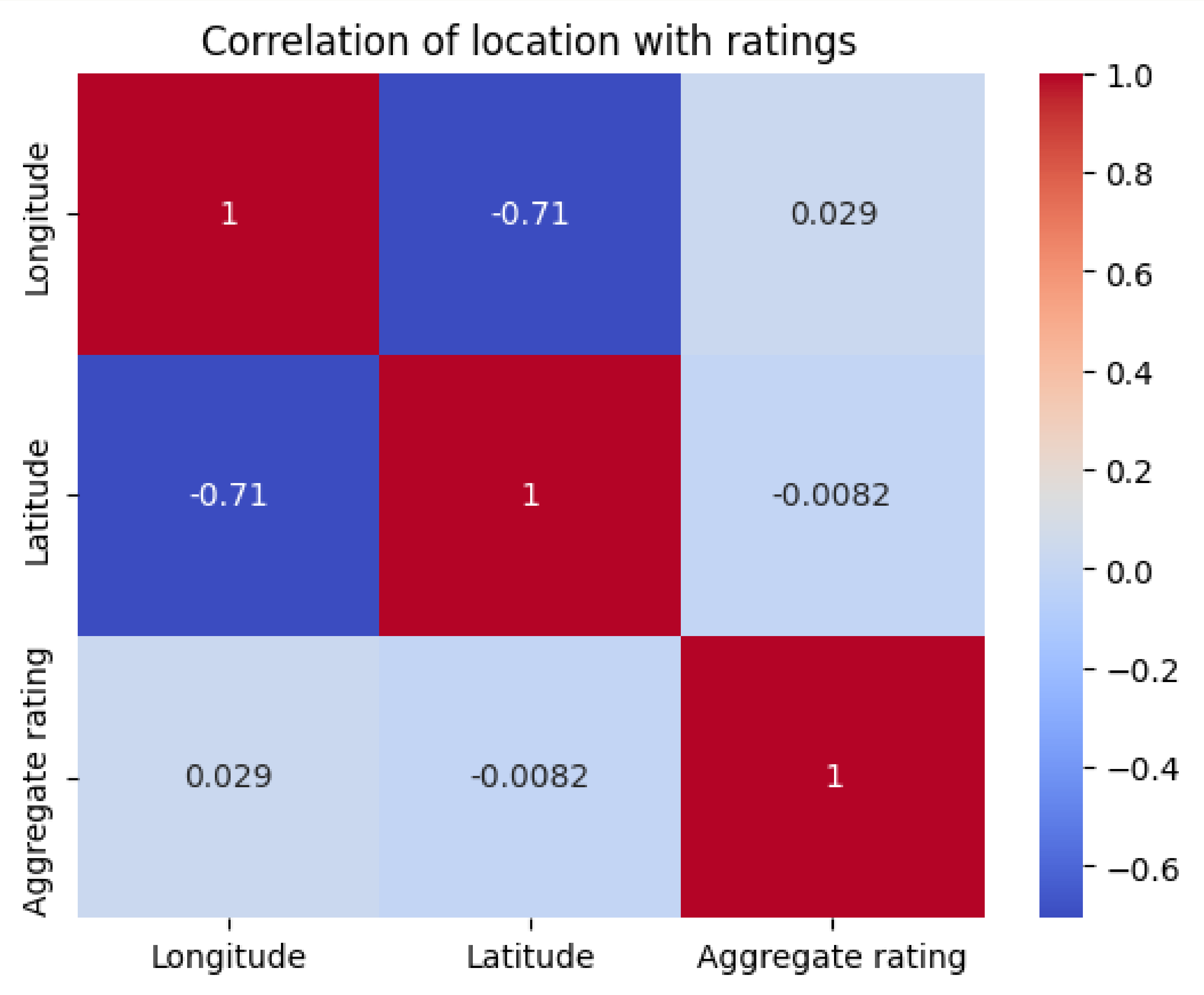
Dense clusters were observed in New Delhi and its surroundings.

It was observed that the Location vs Aggregate Rating had a weak correlation showing that the rating did not depend on the location of the restaurant.

Restaurant Distribution Across Cities



Correlation of location with ratings



New Features for Modeling

For modelling, I changed some features in the dataset. A summary is below:

- I dropped the ID-based fields and other columns that gave no information about the variations in Aggregate rating.

- I encoded the columns below so they could be used for the linear regression model:
  - Cuisines (also grouped missing values as 'Other')
  - City, Locality, and service flags

- I created new features such as:
  - 'High_cost' (used average cost for two to group restaurants as high-end and low-end) (avg cost >500)
  - Total_cuisines How many cuisines it served

- I used the Log-transformed average cost due to help with the skewness.

Machine Learning Models

Model Building

- Target: Aggregate rating

- Train-Test split: 80/20

- Models used:
  - Linear Regression
  - Random Forest Regressor
  - XGBoost Regressor

- Imputed missing values in target using mean strategy

Model Evaluation Metrics

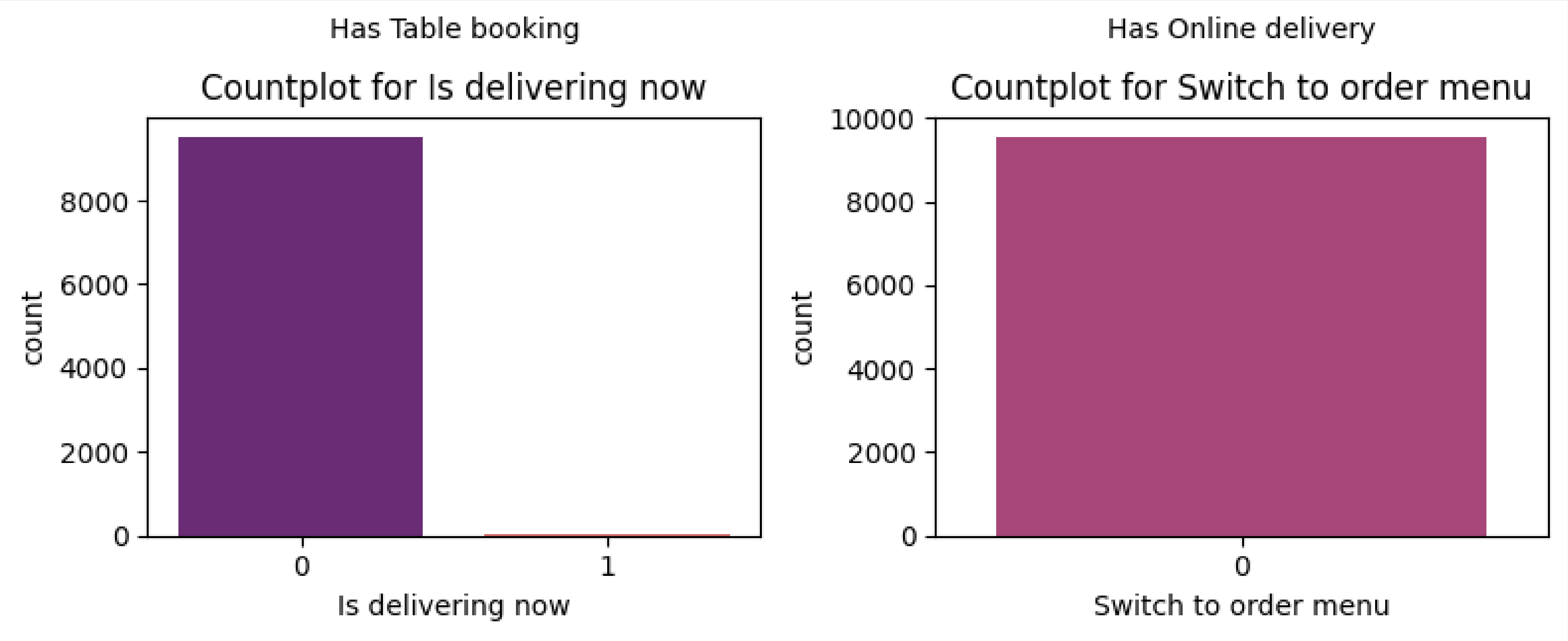
The Best Model is the Random Forest because of its high R-squared value and lower RMSE.

It also captures the non-linear relationships. This makes it best suited for all metrics

Limitations

In my analysis, I was had the following challenges with the dataset:

- The 0.0 ratings were quite substantial. This was bad for the model. Therefore, mean ratings were imputed.
- The Cuisine represntation was not balanced : most of it was from India
- There was really no correlation of the ratings with the location of the restaurant. I could not derive insights from it.
- Some features were constant for all restaurants(e.g., 'Switch to order menu')


Has Table booking
Has Online delivery
Countplot for Is delivering now
8000
6000
4000
2000
0
count
0
1
Is delivering now
Countplot for Switch to order menu
10000
8000
6000
4000
2000
0
count
0
Switch to order menu

Insights

Votes had a moderate correlation with ratings. This means that the more pouplar cuisines had better rating.

Using the Voted filter, the Italian cuisine rated highest among the popular cuisines.(votes ≥50)

The higher the cost of the restaurant, the higher the votes and ratings

Restaurants that had the option of bookings and delivery had an improvement in their ratings compared to others.

My Random Forest model had a prediction with an estimation of 82% accuracy

Recommendations

The dataset can be improved by:

- Including customer reviews to better understand how people feel.

- Adding time-based trends to analyse peak hours and non-peak hours.

- Classifying the ratings into tiers for a better analysis such as " High", "Medium", or "Low" instead of giving individual ratings.

Conclusion

- In this project, we studied what makes some restaurants get higher ratings than others.

- To do this, the dataset was first explored. Here, columns such as -cuisine, price range, location, and services like delivery and table booking- were studied.

- The mean ratings were checked, then filtered by votes to remove bias and Italian Cuisines were found to be the highest rated.

- Also, more expensive restaurants was seen to be rated better.

- After the dataset was explored, a machine learning model was built, to predict a restaurant's rating based on its features.

- The best model was found to be Random Forest.

- The insights given can help restaurant owners improve their services and better understand what their customer wants.

Thank You!

This report was compiled by

Gabriel Amao

ojoayoamao9418@gmail.com

www.linkedin.com/in/ojoayo