yelp

Predict Restaurant Success Using Yelp Data

Problem Statement

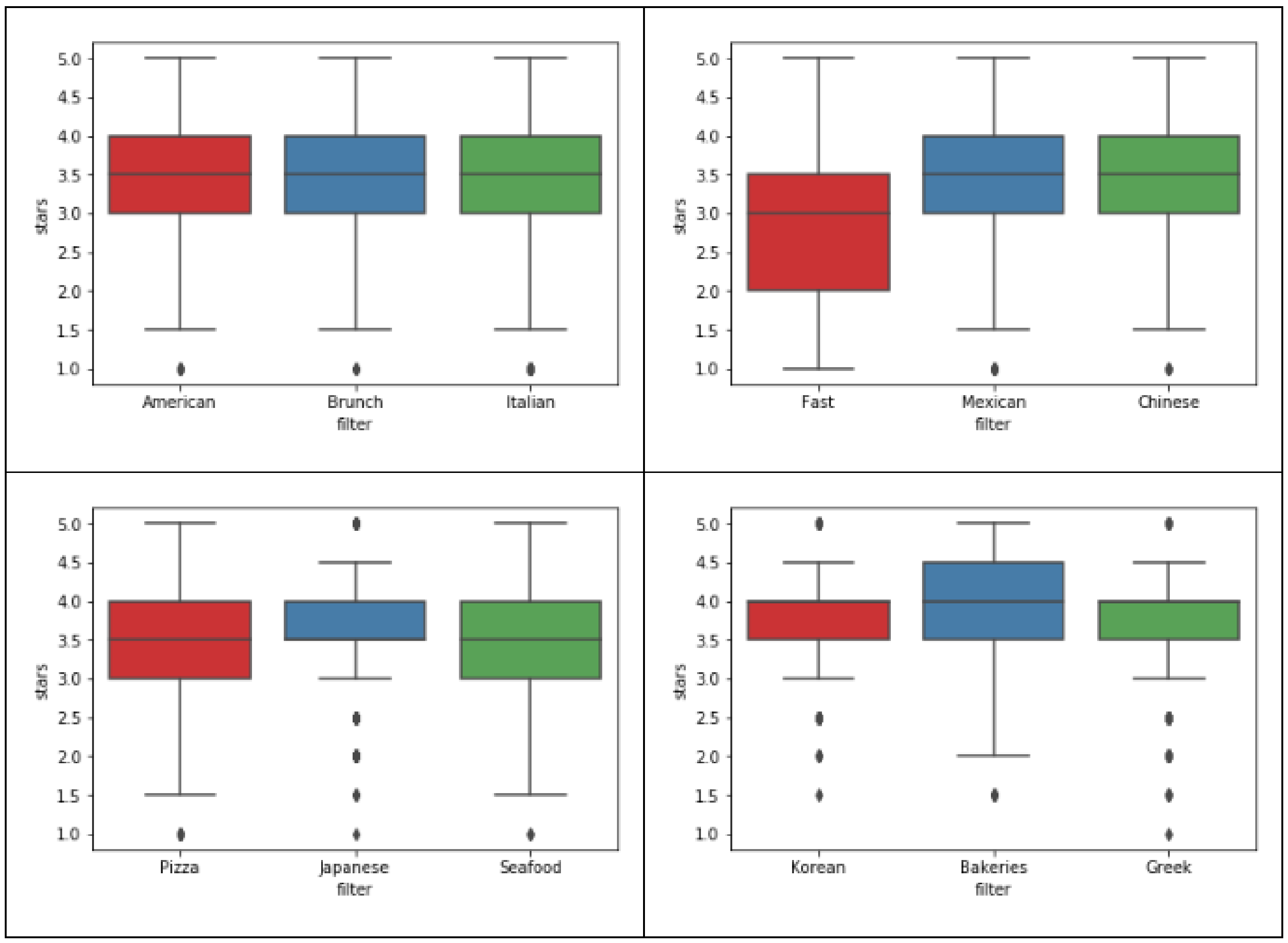
Today, many people look at Yelp reviews to decide which restaurant to go. Customers are likely to go to the restaurants where they have high ratings and positive feedback. Therefore, Yelp star rating could be a good indicator to predict the success of a new restaurant.

This project will build a model that predicts the Yelp star ratings of a new restaurant to help new restaurant owners make business decisions more effectively.

Exploratory Data Analysis

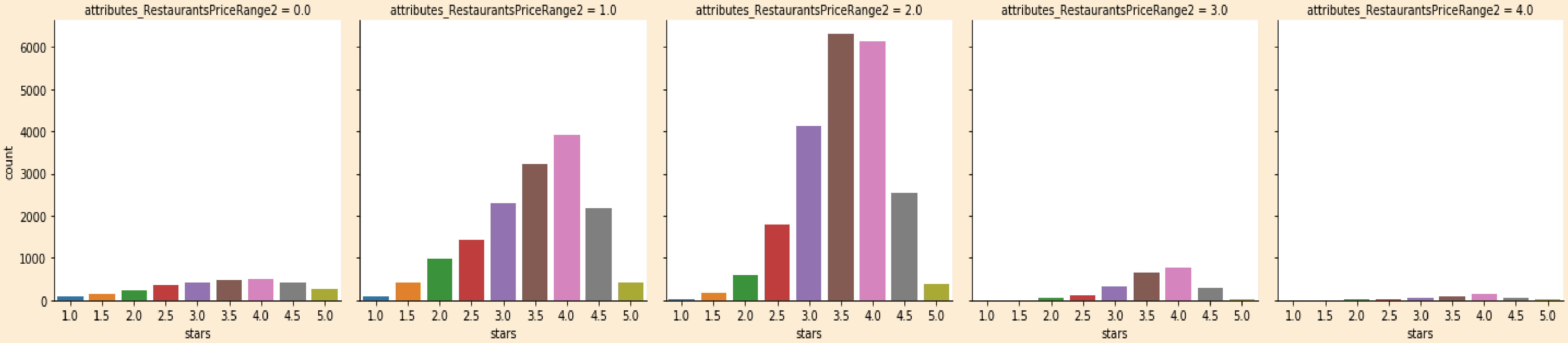
yelp

Restaurant Categories and Star Rating



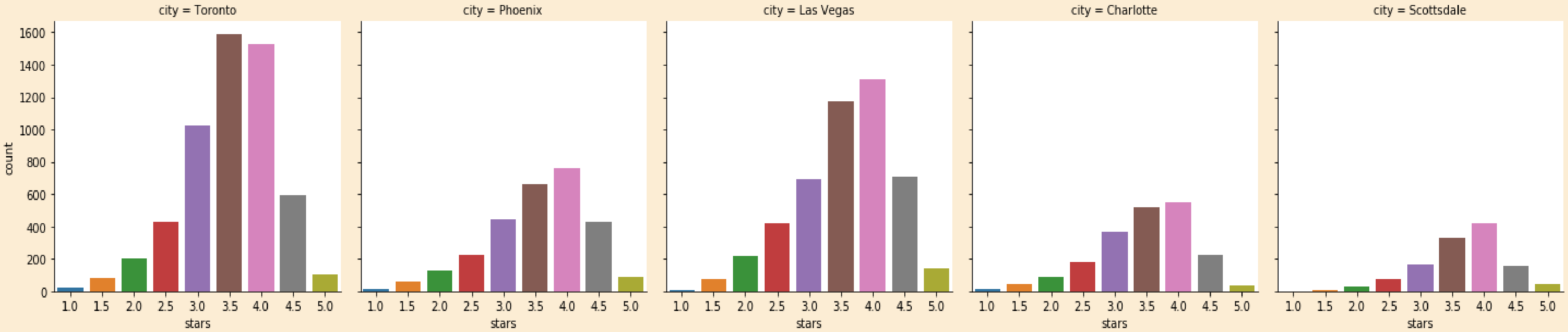
- Fast food has lower median than other categories
- Bakeries have the highest median
- Japanese, Korean, and Greek have higher median than 3.5

Price Range and Star Rating



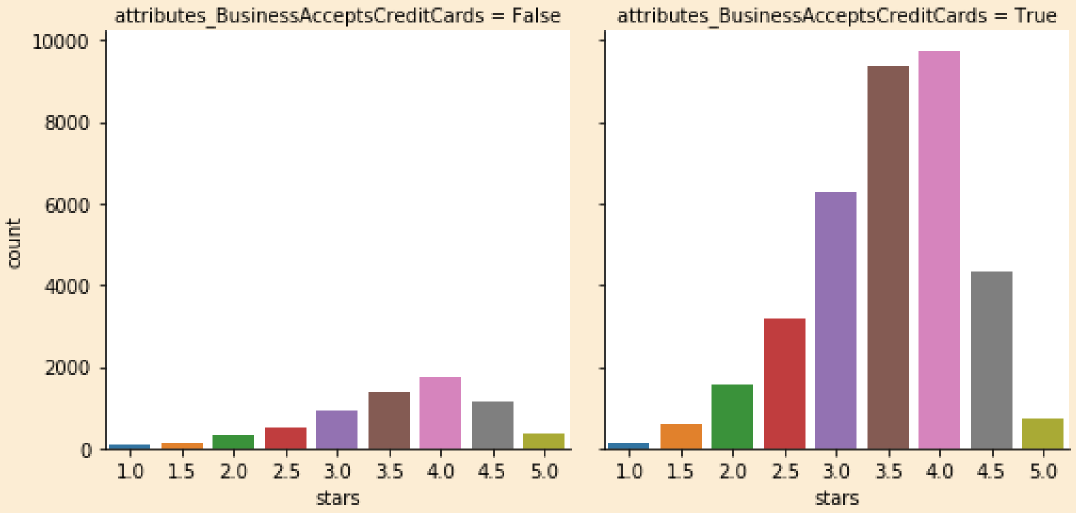
Affordable restaurants are available to more customers and it leads to more ratings

Price Range and Star Rating



The data has 652 cities. The top 5 cities are Las Vegas, Toronto, Phoenix, Charlotte, and Scottsdale. All of them have high number of restaurants that have 3.5 or 4.0 star rating.

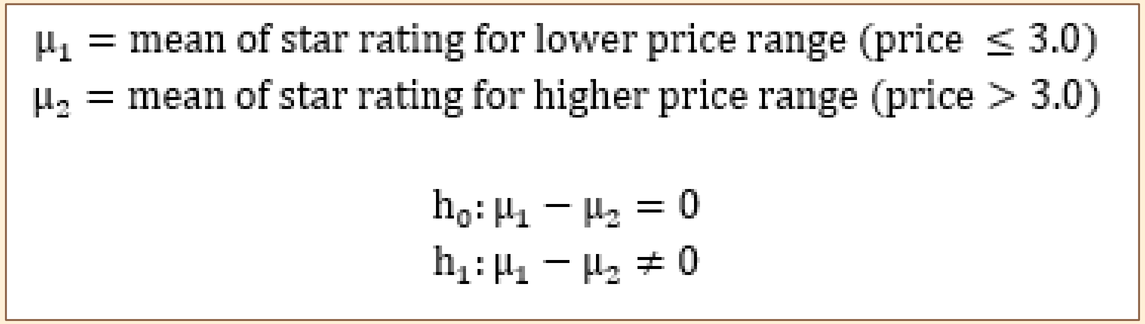
Credit Card Acceptance and Star Rating



The graph tells us that more restaurants accepts credit card but there is no relationship between credit card acceptance and star rating.

Statistical Analysis

Hypothesis Testing between different prices



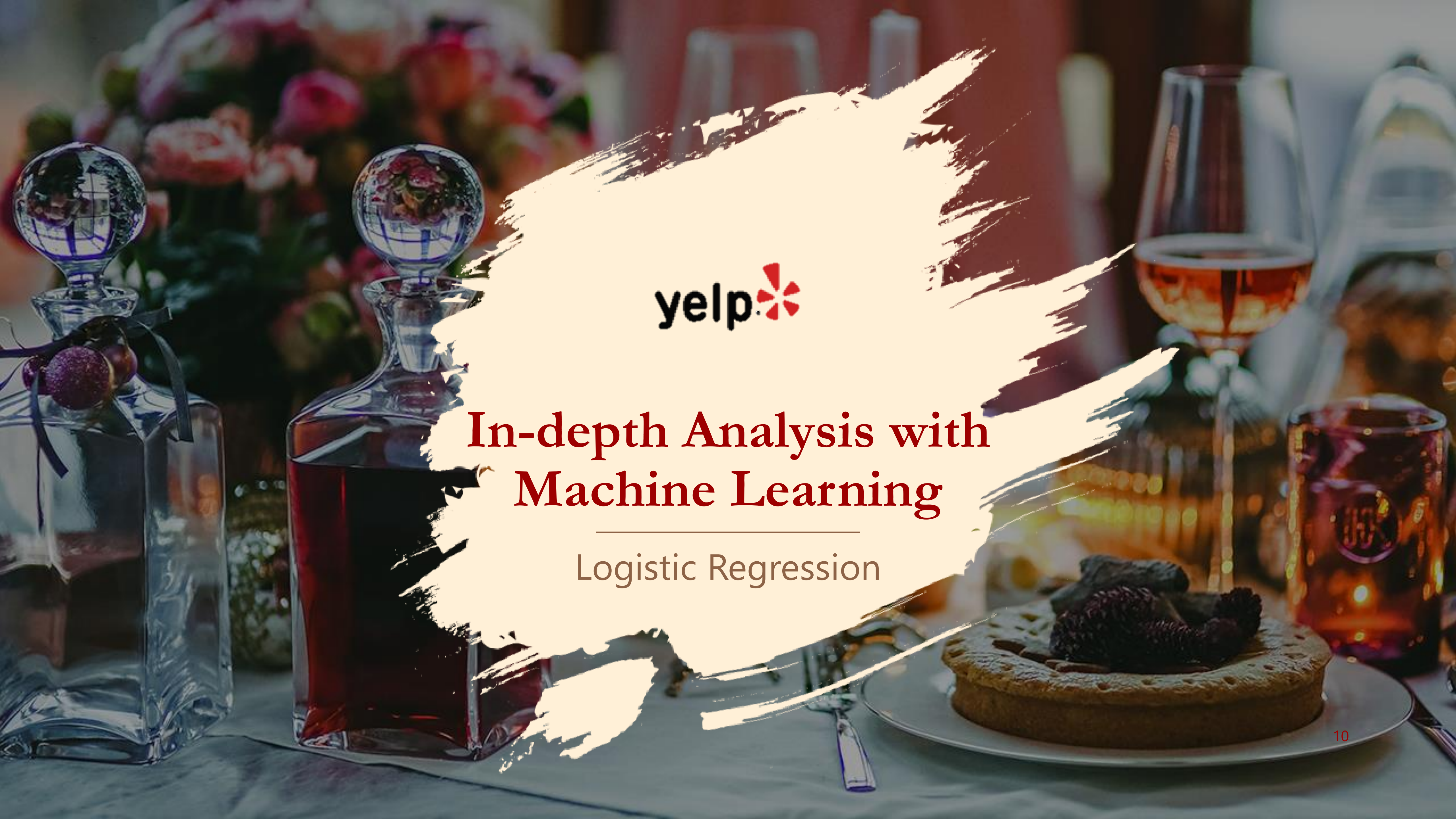
μ_1 = mean of star rating for lower price range (price ≤ 3.0)

μ_2 = mean of star rating for higher price range (price > 3.0)

$$h_0: \mu_1 - \mu_2 = 0$$

$$h_1: \mu_1 - \mu_2 \neq 0$$

The statistical significance test with α = 0.05 showed p-value of 0. Therefore, the null hypothesis was rejected. It was concluded that there is a significant difference in star rating between lower and higher price range.



In-depth Analysis with Machine Learning

Logistic Regression

Logistic Regression

Rescaled Data

Due to data imbalance, star rating above 4.0 is categorized as 1 and the rest as 0

F-1 score

68% of the variability of the response variable was explained with the model

Train Test Data Split

The model was trained using 80% of data and remaining 20% is used to evaluate the performance of the model

Feature Importance

The biggest negative contributor is operating hours on Monday and the positive one is operating hours on Thursday

Restaurant

LOGO

THANK YOU!

Jaesuk Kim

jaesukkim4402@gmail.com

Please see Jupyter Notebook for further analysis