Title: Restaurants and Consumer Analysis

Team members

Overview

In this project, we will utilize restaurant and consumer data from the Yelp dataset to visualize interesting patterns and predict ratings and performance of a restaurant based on consumer sentiments and reviews. We begin with the exploratory analysis of the datasets including the study of the distribution of ratings and cuisine types with respect to the business categories and restaurant locations. We will present the relevant data visualization features with interactive widgets and the demo statistical models and machine learning models in R shiny app. The goal is to help both customers and merchants understand the restaurant better by providing deep insights through visualizations and sentiments.

Dataset Characteristics

Reference: https://www.yelp.com/dataset/documentation/main

This dataset was originally from the Yelp Dataset Challenge website. It contains business information, user information, user`s review, check-in information, and tips information. All original datasets are in JSON format, and we need to convert them into CSV format. Because of the large sizes of the primary dataset, we will mainly focus on the business information and review information for visualizations and models and extract relevant information from other charts.

Methods

We will perform explanatory analysis and will use various interactive techniques to better understand the dataset and give insights to users of the big picture. The ratings/reviews for brands, various categories, cuisines and numerous trends (seasonal, festive and demographic) are understood in this section.

1) *Data Wrangling*

Tidy and clean the dataset with NA/Null values, filter the dataset to US restaurants, understand the correlation between variables, perform feature reduction, and remove bias.

2) *Data Visualization*

- Plot word clouds showcasing the most common cuisine types and the most probable words from the ratings based on reviews, filter by cuisine and city.
- Plot frequency distributions of states vs cuisine and frequency distributions of top cities vs reviews counts.
- Plot graphical maps to filter the restaurants by ratings, city names, and states and also to differentiate them by cuisine.

Models

1) *Sentiment Analysis*

Understand the trends in user behaviours and classify words that contribute to positive and negative sentiments for key attributes in a rating category. Will focus on the top three cities based on review counts and conduct the sentiment analysis on the relevant users` reviews.

2) *Classification*

Results from sentimental analysis will be used to predict the user rating based on text reviews. This will help determine if it's worth trying a new restaurant or not. We will use various machine learning algorithms like Naive Bayes, Support Vector Machines and Random Forest to understand which words affect the classification the most and then further predict the 5 classes of rating.