Improving Success Rates of Restaurants

Milestone Report, Capstone Project 1

What is the problem I want to solve?

Restaurants are a big part of modern life and they make up a significant portion of small businesses everywhere. They are known for high turnover and failure rates, especially within the first year of business.

Failed restaurants are costly to the restaurateurs, to the industry and to the diners. When restaurants fail as businesses, the restaurateurs themselves bear economic burden, the food services industry lose potential growth from the contribution of new ideas and economy, and the diners lose out on opportunities for new enjoyable experiences. This study in data science aims to help restaurateurs to incorporate more success factors into their strategies when opening new restaurants.

What dataset is used for analysis?

In order to make inferences about restaurants in general, a decent sample of restaurant data is needed for analysis. Yelp provides just the dataset needed at https://www.yelp.com/dataset. The dataset is downloaded and unpacked into 3 files in json lines format. Each file is gigabytes in size.The first file contains 192,609 businesses on Yelp with over 1.2 million business attributes like hours, parking, availability, and ambience. The second file includes the aggregated check-ins over time for each of the 192,609 businesses. The third file contains the reviews of the 192609 businesses. The file containing the reviews is over 5 gigabytes in size.

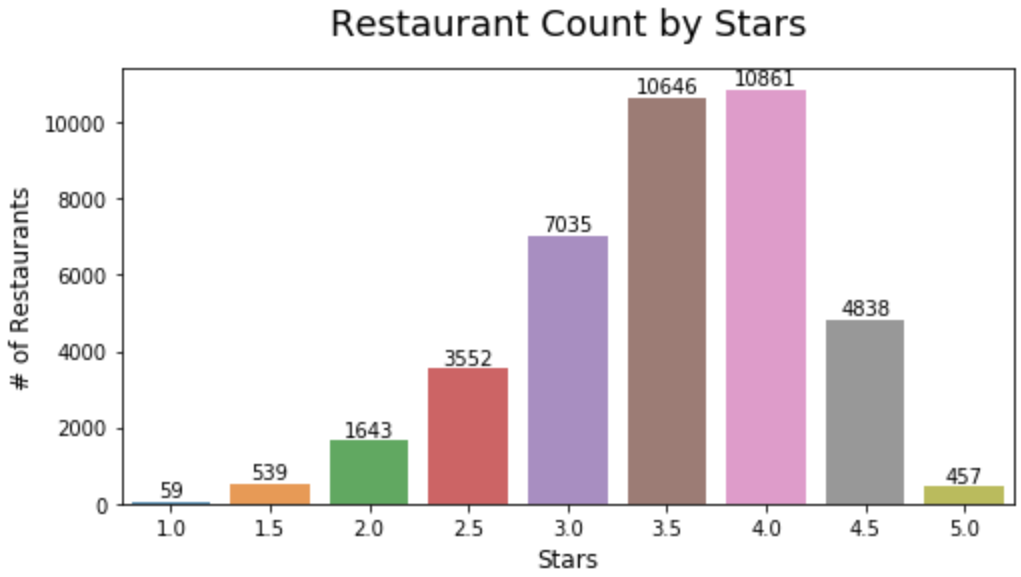
Data wrangling activities:

- Read files that are in json lines format.
- Format json strings.
- Create pandas dataframes from json strings.
- Filter businesses into restaurants dataframe.
- Filter reviews for restaurant reviews by restaurant standing.
- Filter restaurants outliers - those with less than 10 reviews.
- Triage restaurants into those of good (4-5 stars), moderate (2.5-3.5 stars) and poor (1-2 stars) standing into dataframes.
- Triage checkins by restaurants standing (good, moderate, poor) into 3 restaurant checkin dataframes.
- Parse and reshape date columns of restaurant checkin dataframes.
    - Split original date string by comma
    - Strip each split date string of leading whitespace
    - Convert each date string to datetime value
    - Insert each date value with its business id into a dictionary
    - Create list of the dictionaries containing business ids and dates
    - Create dataframe from list
- Reset index for all resulting dataframes before using for analysis.

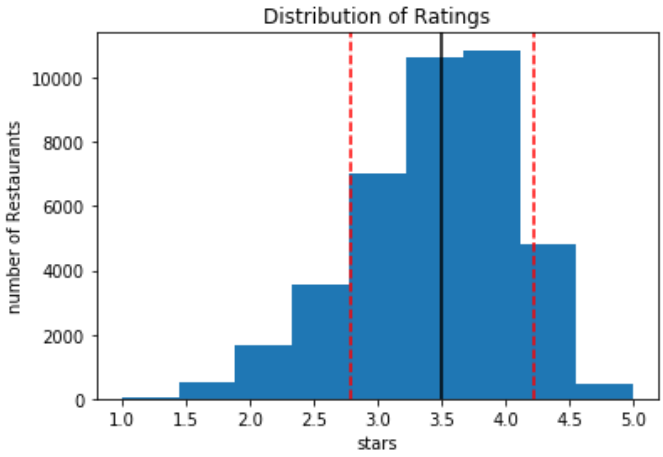
Exploratory Data Analysis

Ratings Distribution

Yelp restaurant dataset contains significantly more restaurants in moderate and good standing than those in poor standing. Most restaurants have 3-4 stars in rating, with 1 star ratings being the rarest.

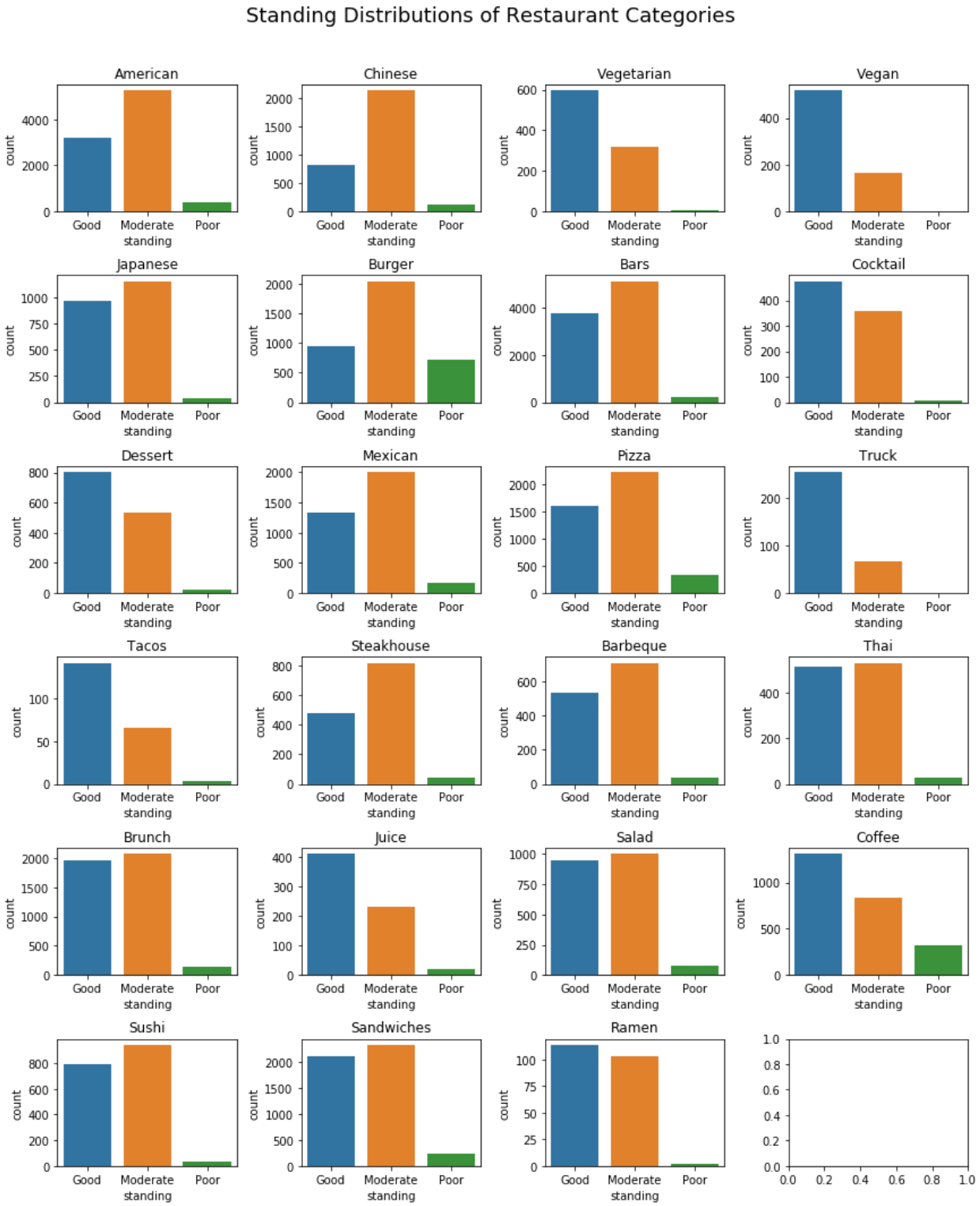


The restaurants average 3.5 stars in rating and 0.71 in standard deviation.



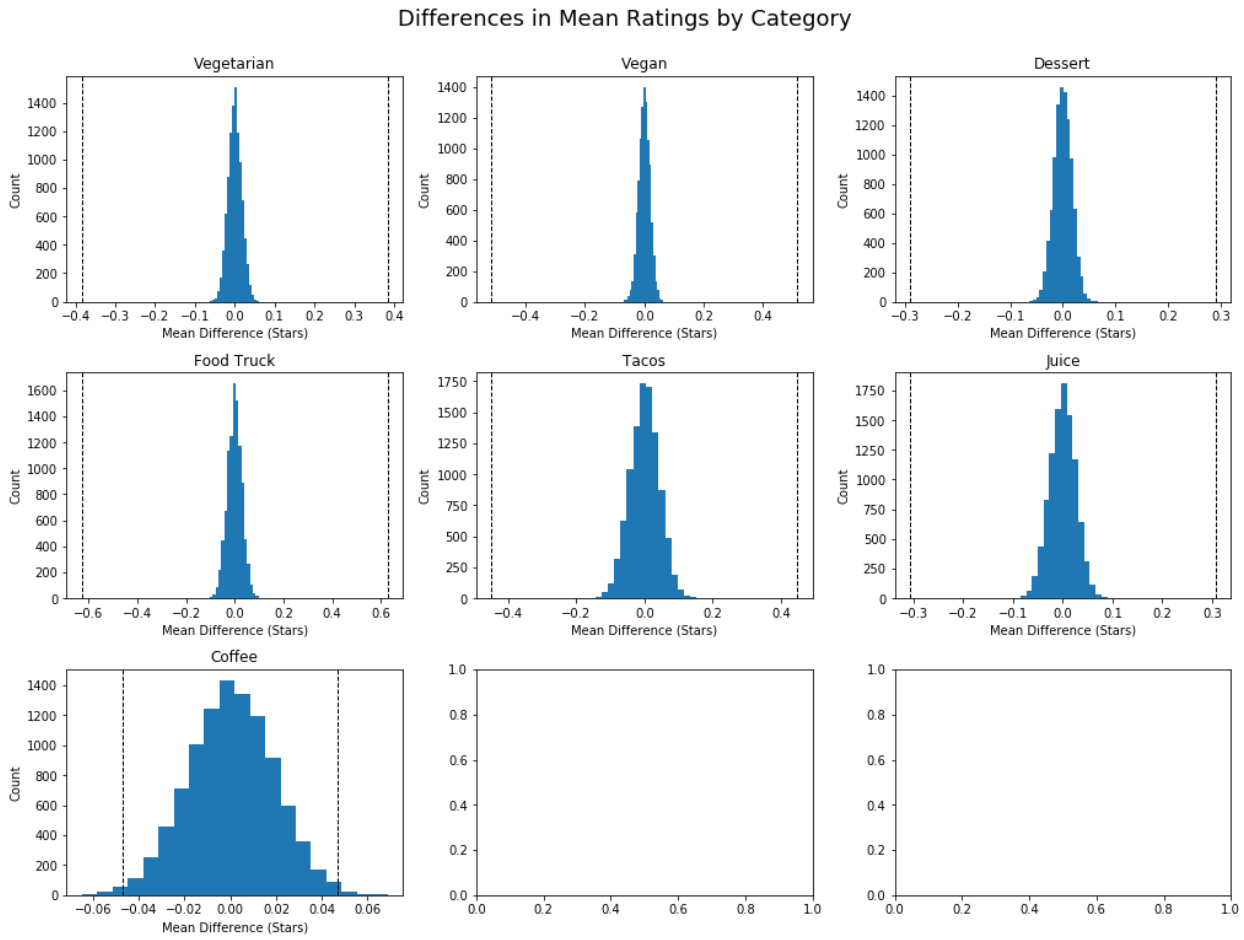
Restaurant Categories

Yelp has over 7000 categories for restaurants. Restaurants of type Vegetarian, Vegan, Food Truck, Cocktail, Dessert, Tacos, Juice, Coffee, and Ramen have standing distributions that differ from that of the larger population. These categories contain more good restaurants than moderate and poor restaurants.



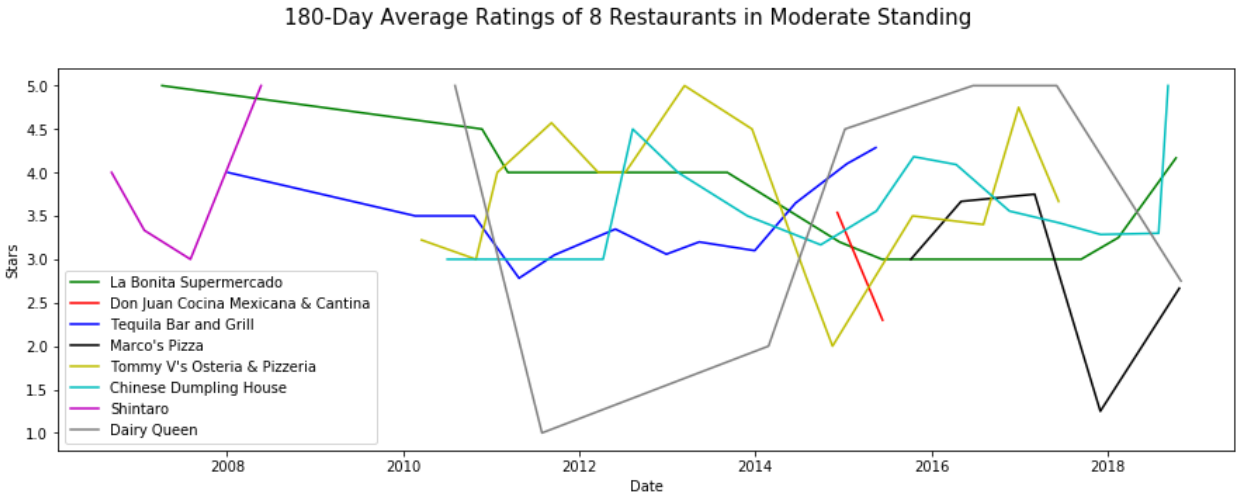
Using bootstrap inference to perform a t-test on the hypothesis of whether mean ratings differ by category, significant relationship is found between restaurant category and restaurant standing. New restaurateurs might benefit from knowing restaurants in what categories are easier to start

up and manage. Perhaps keeping it simple with a food truck or focusing on beverages is the way to go for a new restaurateur in the business.

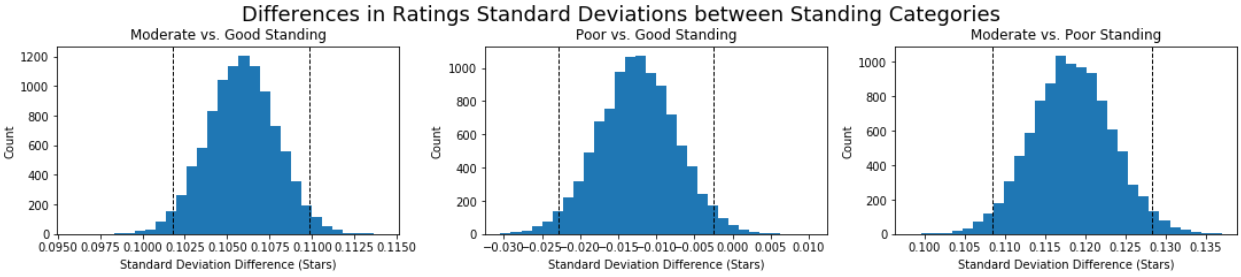


Standard Deviations in Ratings by Standing

Moderate restaurants fluctuate more widely in performance over time than good and poor restaurants. Poor restaurants fluctuate more widely in performance over time than good restaurants. Good restaurants perform with higher consistency over time.



Using bootstrap inference to perform t-tests, the resulting p-values of 0 suggests there are significant differences in the standard deviations of ratings depending on restaurant standing.



The larger variances in ratings of moderate restaurants suggests to future restaurateurs the importance and advantage of starting on the right foot, and the difficulty of improving consistency in performance once the restaurant is up and running.