TRUE REVIEW

A Personalized Yelp Recommendation Engine

Dmitriy Kats
June 2019

Table of Contents

Dmitriy Kats
Capstone 2 – Personalized Yelp Restaurant Recommendation Engine

1. Introduction and Executive Summary

Many times, we find ourselves on Yelp, looking for a new restaurant in the area with each one having 4000+ reviews. There is no way we can go through that many reviews to understand how people rate and what they value. Our recommendation system will take a given user's reviews and compare to others'. Based on content and dining style, we'll present more relevant restaurants with a better fit. A content / collaborative based recommendation engine that will also average the star ratings and predict what a given user might rate the restaurant.

2. Acquiring and Aggregating the Data

The data were obtained from the Yelp Challenge Dataset, containing:

business - Contains business data including location data, attributes, and categories.

review - Contains full review text data including the user_id that wrote the review and the business_id the review is written for.

checkins - Checkins on a business.

tip - Tips written by a user on a business. Tips are shorter than reviews and tend to convey quick suggestions.

user - User data including the user's friend mapping and all the metadata associated with the user.

All data was placed into a Google storage bucket and then loaded into BigQuery for easy access.

In order to reduce the size of the Pandas data we're working with we:

- Queried the tables only with restaurants that are not fast-food
- Joined the restaurant data with review data
- Removed any users with no friends and less than 200 reviews
- Limited the analysis to one city

3. Cleaning the data

The data are relatively clean, given this is Yelp's official public dataset. However, in order to analyze and model, we'll need to significantly reduce the focus and scope of the data.

3.1. Data Query

Since the data was loaded into Google BigQuery, we'll first query the entire business data set, which contains all business, in addition to our desired restaurant data. Merging the entire business and review datasets would take too long on a local machine, so we'll let Google handle it. We'll then save the resulting dataframe as a CSV for easy access locally.

As stated in section 2, we'll focus on only the described attributes of the dataset.
Resulting csv file will contain the following:
City Name: Scottsdale

- Number of Users: 76,011 3,119 (4%)
- Number of Restaurants: 1,322 → after filtering → 1,210 (92%)
- Number of Reviews: 173,062 24,738 (14%)

GitHub code

4. Exploratory Data Analysis (EDA)

Yelp has been around since 2004, helping people find great local businesses by presenting an easy to read star rating along with more detailed users' reviews. By the beginning of 2019, the number of users and reviews has exploded to over 100M and 184M, respectively. We are going to look at a small subset of the data in order to draw insights in review trends and restaurant trends and use these insights to guide users to a better experience by predicting their personal rating of a specific restaurant. We'll first look at the businesses that are tagged as a non-fast-food restaurant. We have data from 2004 through the middle of 2017.

4.1. Ratings Over Time

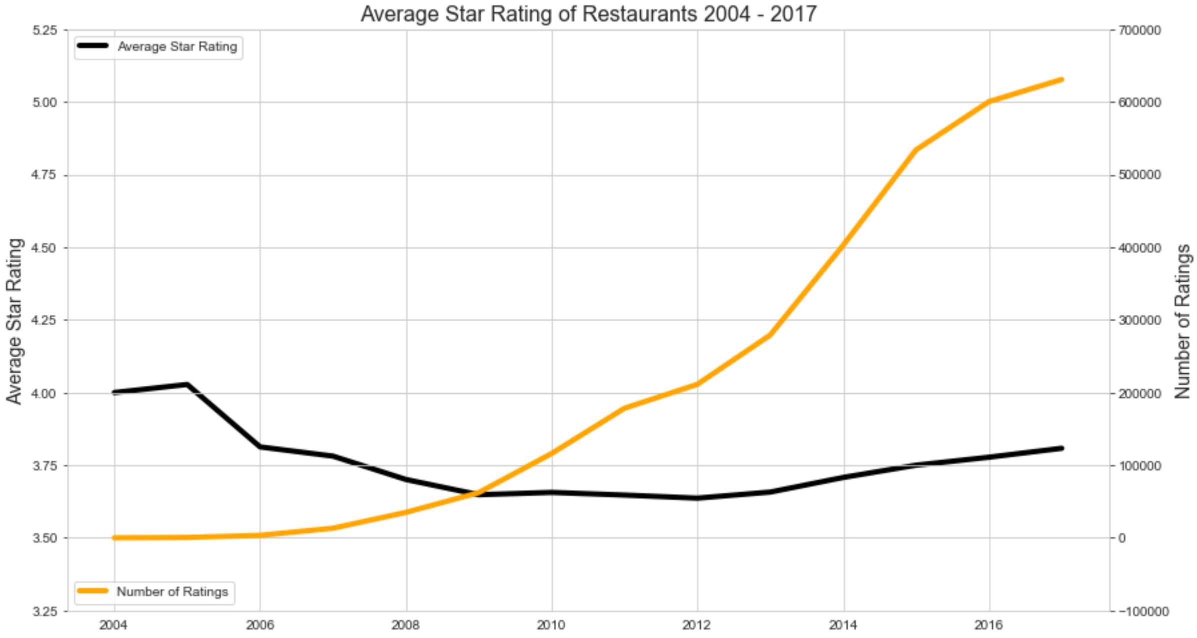


Figure 1- Average Restaurant Rating Over Time

****NOTE THE Y-AXIS SCALE FOR STAR RATINGS DOES NOT START AT A MINIMUM****

We can see from above that the average restaurant rating has fluctuated over time, starting out at around 4 stars, dipping down to 3.6, and currently on an uptrend at around 3.8 stars. We also note that the number of reviews has sky rocketed starting at about the same time as the downward trend on the average star rating. We'll take a look at the relationship between number of reviews and average star rating later in the EDA, but for now, let's look at what the distribution of ratings looks like over this same time-frame.

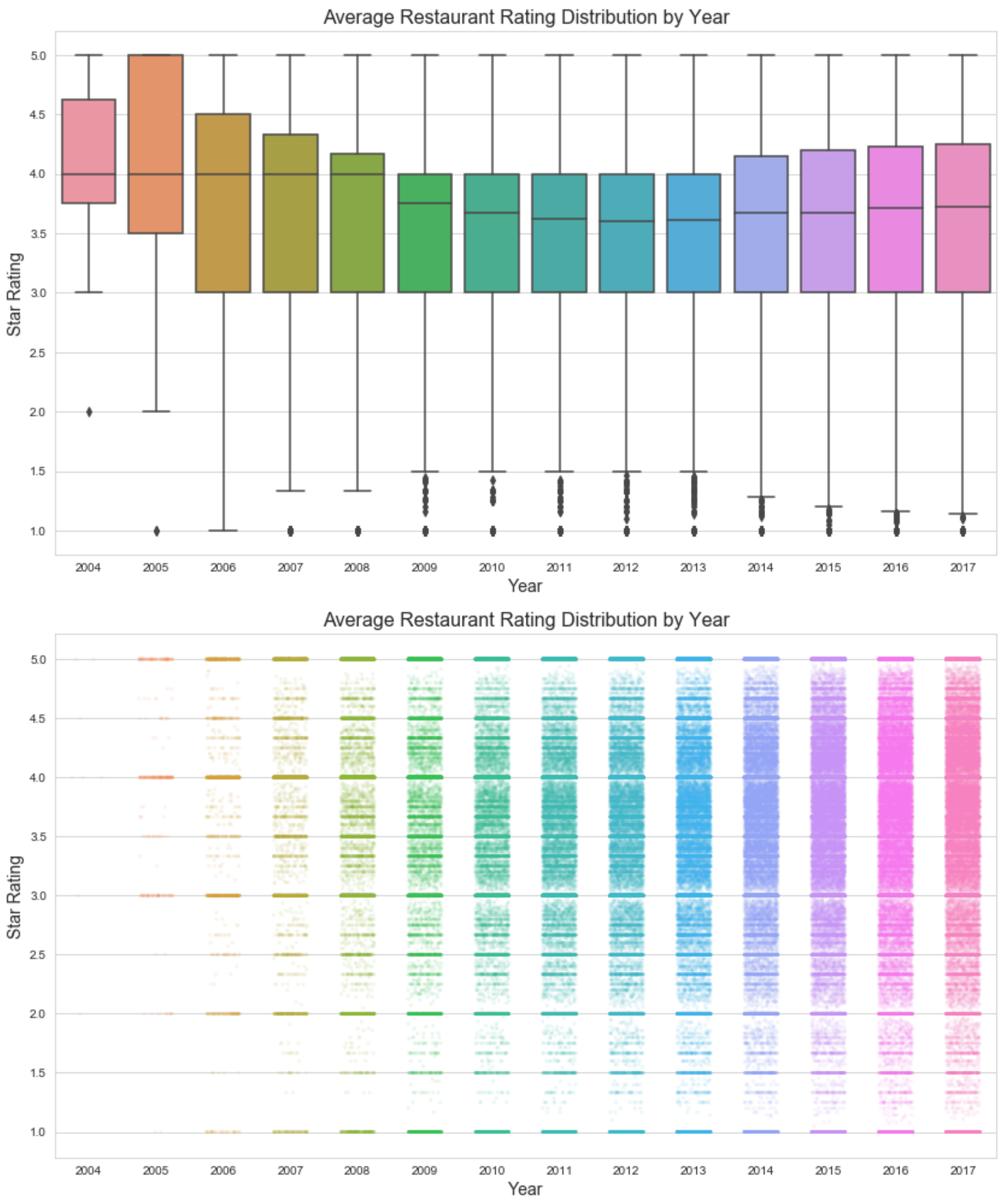


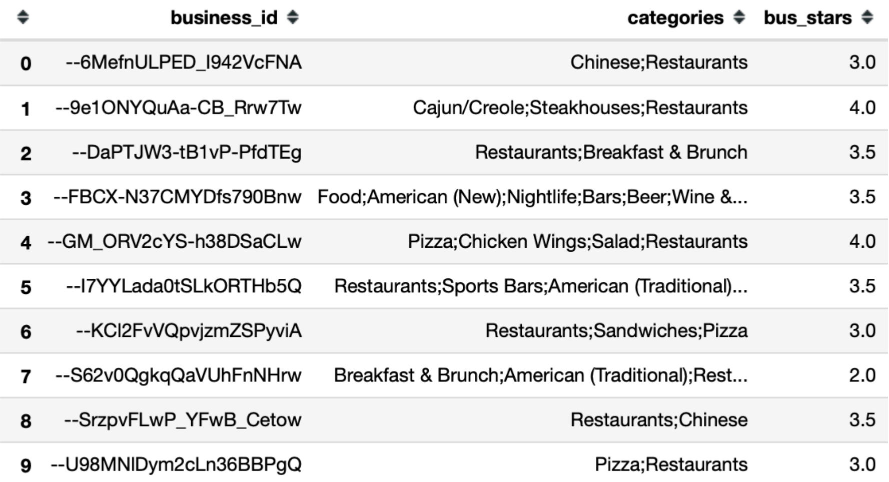
Figure 2 - Distribution of Star Ratings Over Time

Above figures show a distribution of average reviews over the years. The two different figures help us visualize how the spread has increased and how the median / mean shifted over the years.

4.2. Restaurant Categories

In this section we'll take a look at the category breakdown from all the restaurants in the dataset. It should be noted that these tags are manually entered by Yelp users and there are some discrepancies. Additionally, each restaurant is tagged with multiple tags as such the number of occurrences does not add up to number of restaurants. In general, we should get a good idea of the types of restaurants we have.

We took out the words: "Restaurant"," Food", and "Nightlife" since these don't typically contribute to actual food category. We kept "Bar" since some restaurants offer bar food, which can be considered a food category.



⬍	business_id ⬍	categories ⬍	bus_stars ⬍
0	--6MefnULPED_I942VcFNA	Chinese;Restaurants	3.0
1	--9e1ONYQuAa-CB_Rrw7Tw	Cajun/Creole;Steakhouses;Restaurants	4.0
2	--DaPTJW3-tB1vP-PfdTEg	Restaurants;Breakfast & Brunch	3.5
3	--FBCX-N37CMYDfs790Bnw	Food;American (New);Nightlife;Bars;Beer;Wine &...	3.5
4	--GM_ORV2cYS-h38DSaCLw	Pizza;Chicken Wings;Salad;Restaurants	4.0
5	--I7YYLada0tSLkORTHb5Q	Restaurants;Sports Bars;American (Traditional)...	3.5
6	--KCl2FvVQpvjzmZSPyviA	Restaurants;Sandwiches;Pizza	3.0
7	--S62v0QgkqQaVUhFnNHrw	Breakfast & Brunch;American (Traditional);Rest...	2.0
8	--SrzpvFLwP_YFwB_Cetow	Restaurants;Chinese	3.5
9	--U98MNlDym2cLn36BBPgQ	Pizza;Restaurants	3.0

Table 1 - Example of Restaurant Category Labels from the DataFrame

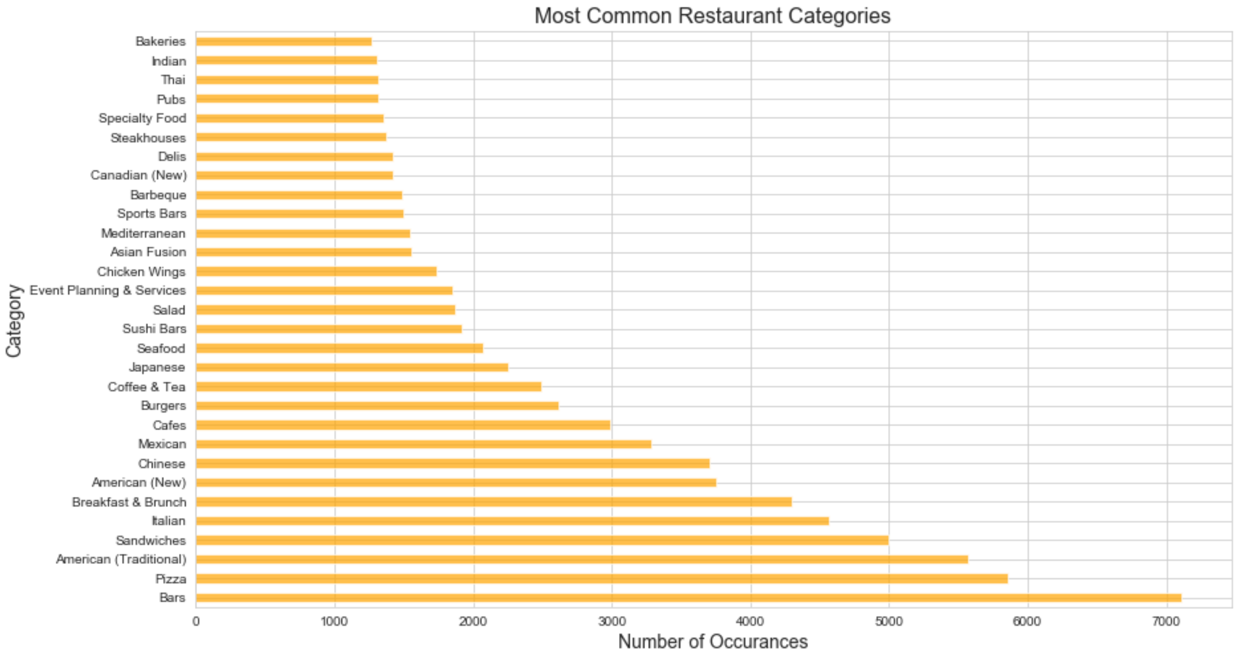


Figure 3 - Number of Category Labels

From the figure above we can see how the categories breakdown among all the restaurants. It's interesting that top category is "bar", followed by mostly junk food / lunch places. We'll use these categories in our feature engineering section, so it's important to understand the breakdown and the fact that same label may be shared among a group of different restaurants.

4.3. Ratings in Different Cities

Yelp operates all over the world; however, their public dataset is for a limited number of cities. We want to select one city for our analysis. Let's now look at what cities we have in our dataset and the distribution of reviews.

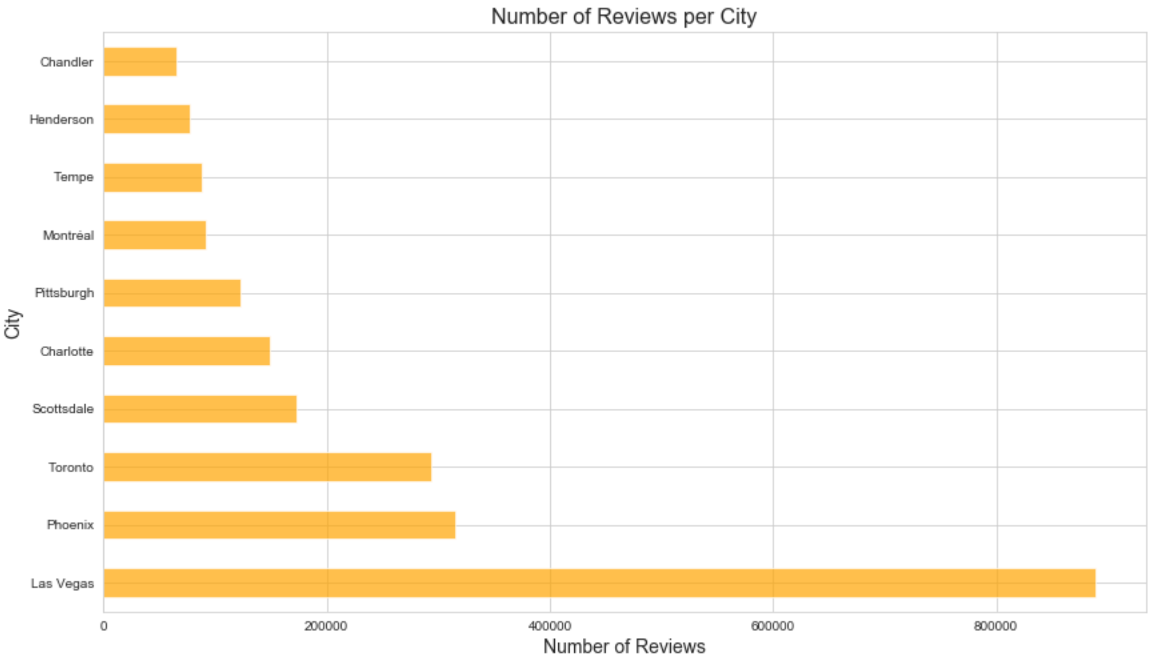


Figure 4 - Number of Reviews for Top Cities

Las Vegas has the most reviews, by far. However, we should look at other city statistics to decide which city to choose. We ultimately want to create a user-restaurant matrix that is not sparse. For this we'll need to look at a ratio between the number of reviews for a given city and the total possible combination of reviews (users x restaurants). Minimizing the sparsity will produce a better collaborative recommendation engine.

```
('Las Vegas', 0.02),
('Scottsdale', 0.017),
('Henderson', 0.017),
('Tempe', 0.016),
('Chandler', 0.013),
('Phoenix', 0.012),
('Charlotte', 0.007),
('Pittsburgh', 0.006),
('Toronto', 0.003),
('Montréal', 0.003)
```

Although Vegas has the highest sparsity, we'll pick the next city on the list, Scottsdale. This should be a much smaller dataset to deal with.

4.4. Highest Rated and Reviewed Restaurant

Before we look into an individual city, let’s take a deeper dive into one of the highest rated and reviewed restaurants in our dataset: Gordon Ramsay BurGR in Las Vegas, with over 2000 5-star ratings. We’ll take a look at user checkins and reviews for this restaurant and compare to the other highly rated restaurants in Vegas located on The Strip.

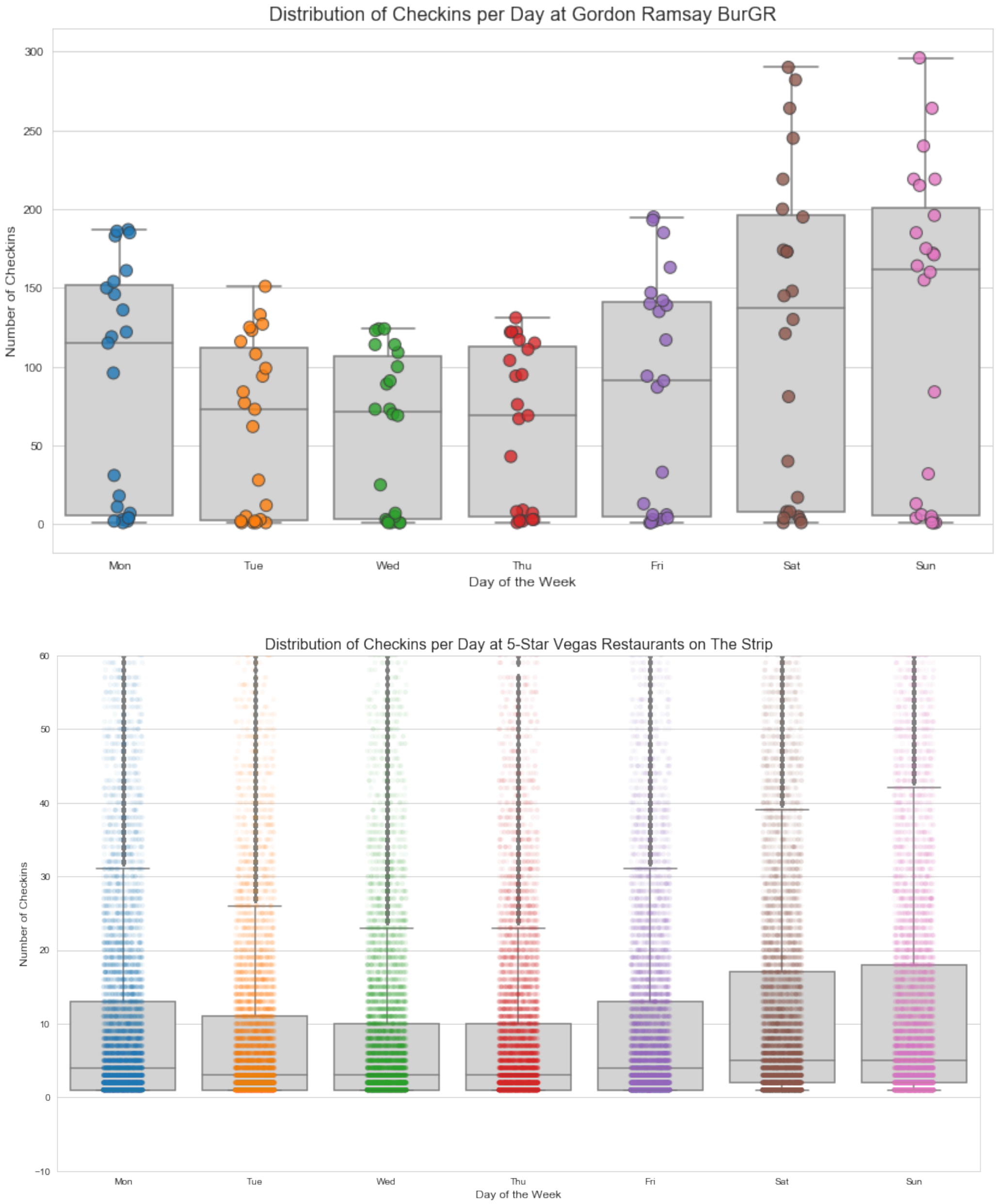


Figure 5 - Number of Checkins Each Day for Gordon Ramsay BurGR (TOP) and Other High Rated Restaurants on The Strip (BOTTOM)

We see a similar pattern for the rest of The Strip restaurants as we do for Gordon Ramsay BurGR, where the middle of the week tends to be slower and the checkins pick up during the weekend and into Monday.

Since Gordon Ramsay BurGR is a highly rated restaurant, we can see how the average rating looks over time and if there any spikes in distribution of ratings to check for any potential influx of "fabricated" ratings.

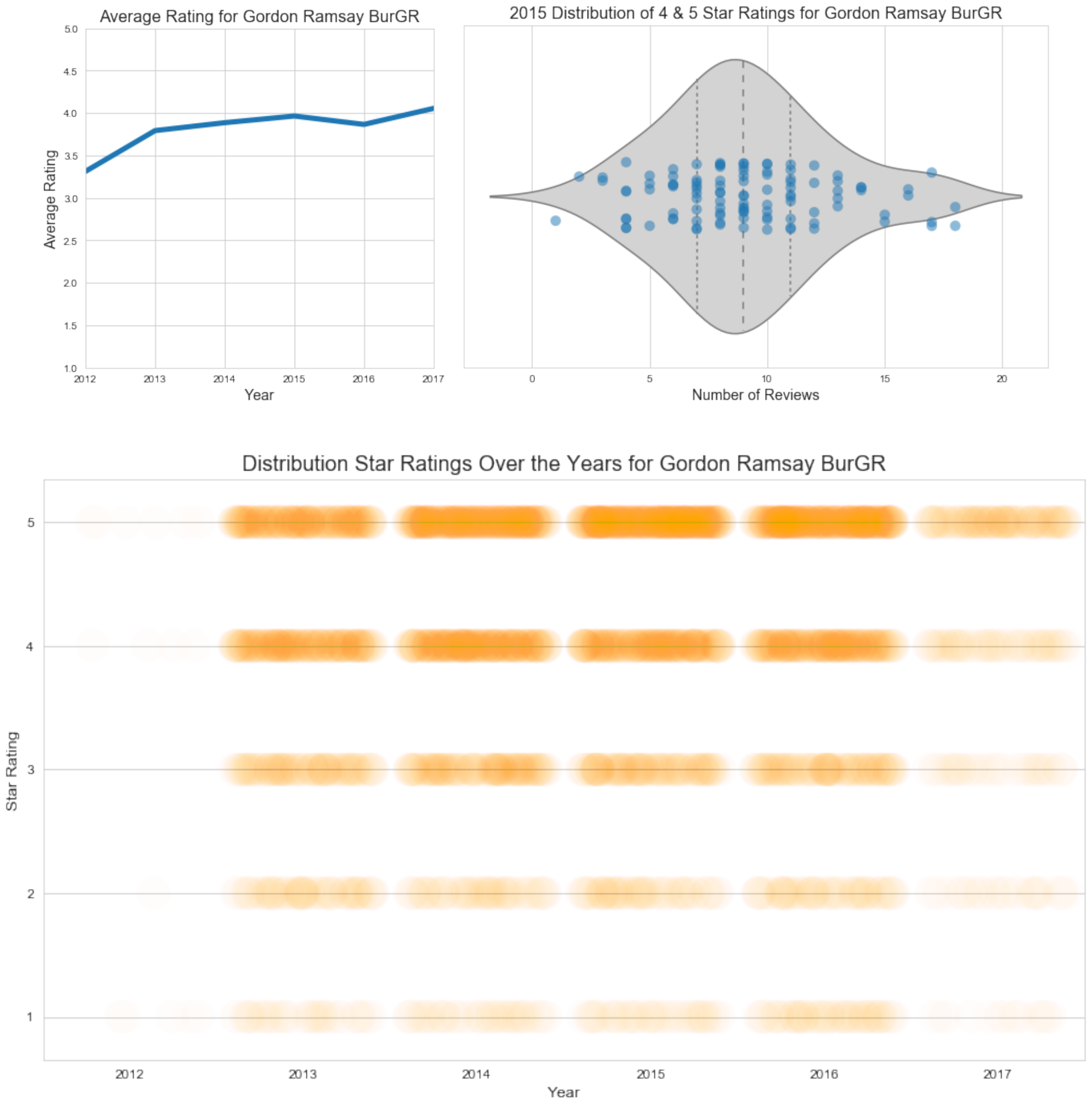


Figure 6 - Detailed Figures for Gordon Ramsay BurGR

In general, we don't see any spikes in number of reviews and the average remains relatively flat over time.

4.5. Users and Their Friend

Yelp also provided us with user data, along with who they are connected to on the platform. Let's take a look at how different users review. Let's first see how review based on how many friends they have.

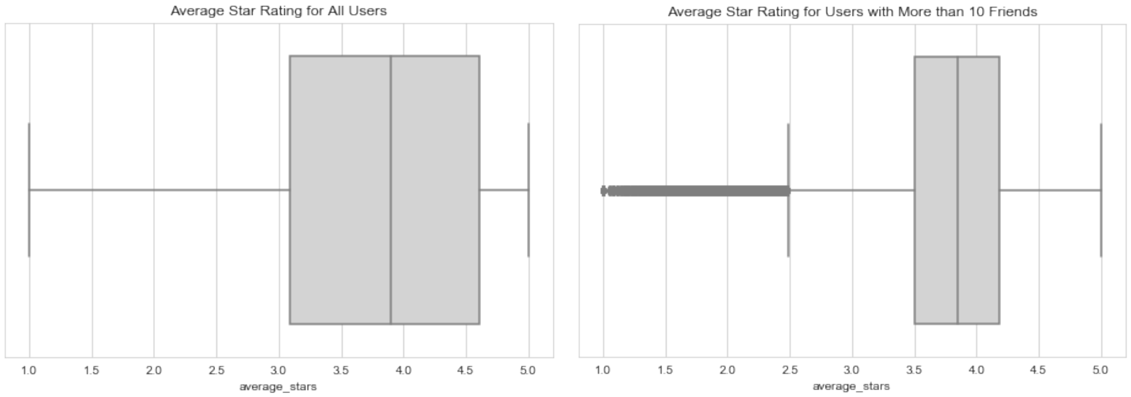
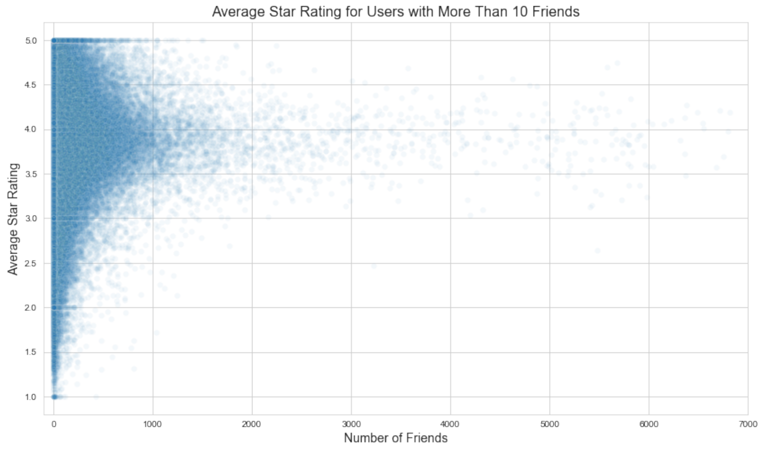
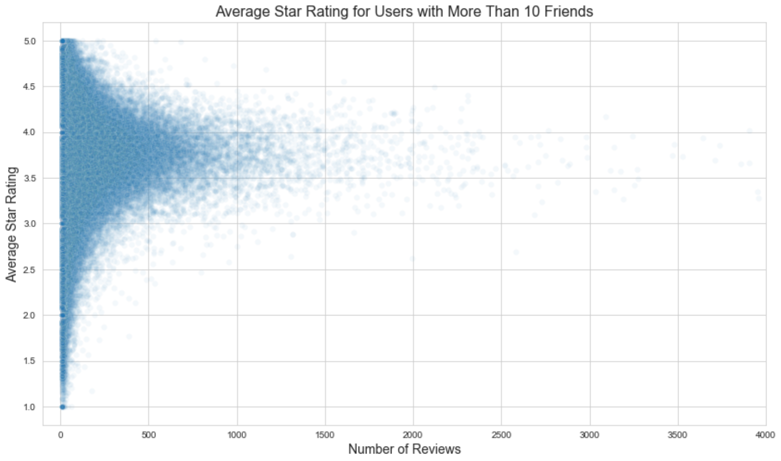


Figure 7 - Rating Distribution for All Users (Left) and Users with more than 10 Friends (Right)

Next let’s see what the actual scatter of ratings looks like.



We notice an interesting phenomenon, the more friends a user has the closer the average rating are to the mean. This will presumably hold for a scatter plot looking at number of reviews vs average star rating.



5. EDA Conclusions

6. Modeling

There are many algorithms available to for recommendation systems. Below is a table with ones we'll focus in this analysis.

	Advantages	Disadvantages

Table 2 Modeling Techniques

7. Feature Engineering and Data Balancing

8. Conclusions and Next Steps