Digging In: Yelp Data Wrangling

BY SAM GUTENTAG
MARCH 2, 2018

Overview

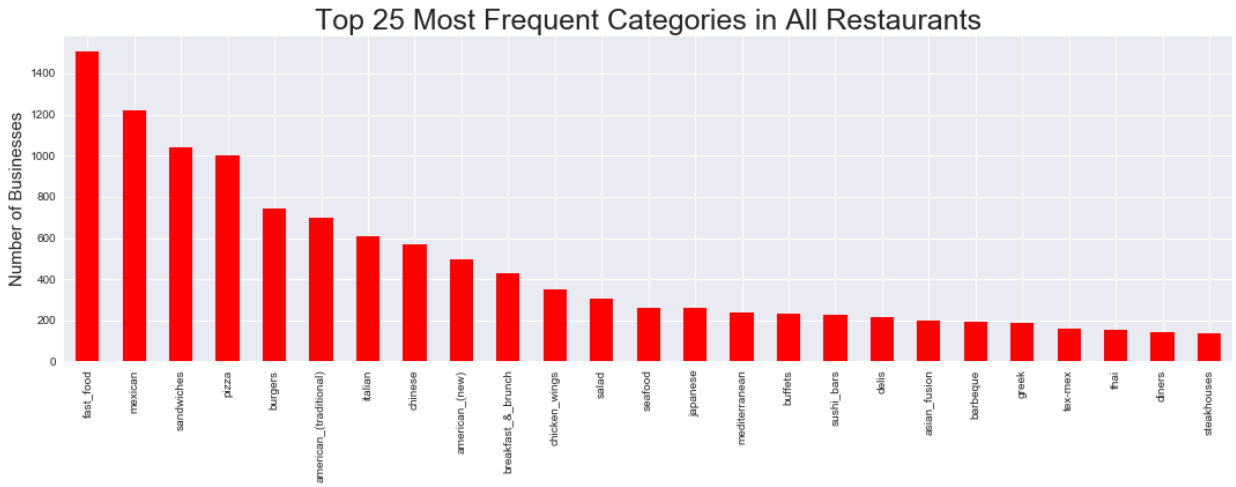
The data provided by the Yelp Academic Dataset site is very well structured JSON. Broken into 5 files, the data set consists of 'business.json', 'review.json', 'user.json', 'checkin.json', and 'tips.json' files.

Business Data

The 'business.json' file consists of data for each business included in the datasets. Each business is assigned a unique 'business_id' attribute that can be used to cross reference reviews, tips, and checkins reported in the other data files. Business data contains information on the name, location, neighborhood, the total review count, the business' star rating, several attributes such as parking, validation, byob, and several other factors. These are largely ignored in the wrangling as they are all user reported and by default are set to nan, only updating to True or False when enough users report an update.

Business data also includes an array of cuisine categories for each business. This categories array and location are the largest filters used to limit the data set to only locations that are considered a 'restaurant' and are in the state of Arizona. A list of restaurant cuisines was manually collected from all possible reporting categories and a list of 221 possible options was retained. This list is used to cross reference each business, removing any location that included a category value not in this list of restaurant categories. Finally, a secondary data frame of only locations not classifying as 'Fast Food' were created with a similar process.

This reduced the initial set of 174,566 businesses to just 7066 total restaurants, 5561 non fast food restaurants and 1505 Fast Food restaurants.

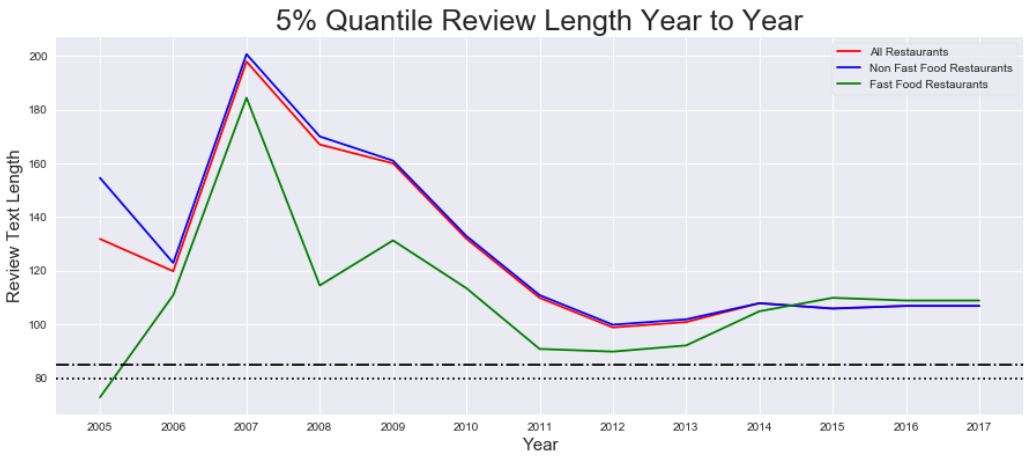


Review Data

The original 'review.json' contains over 5.26 million restaurant reviews, after we have pruned down our selection of businesses to include only Arizona Restaurants, we can filter these reviews to include only those reviews given to restaurants in this list. We end up with a set of Arizona Restaurant Reviews consisting of 501,250 reviews split into 40,349 Fast Food Reviews and 460,901 Non Fast Food Reviews.

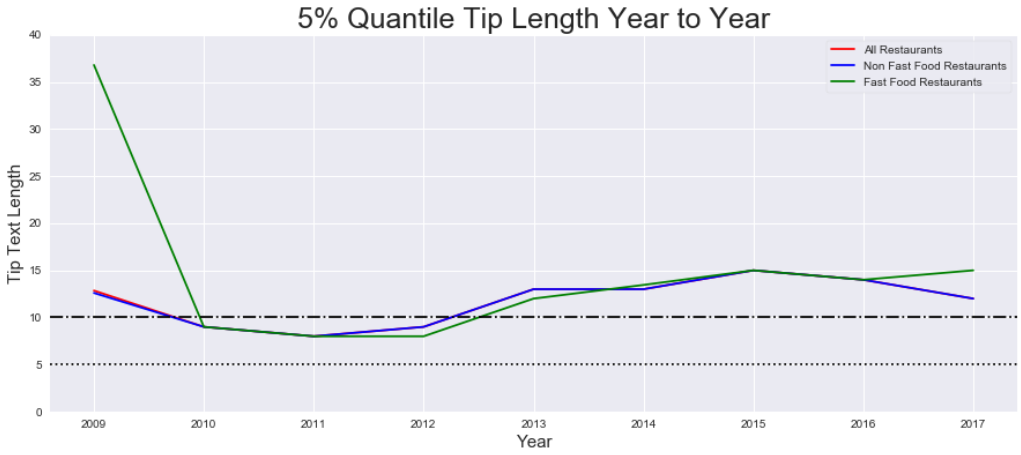
Next up, we clean up reviews by stripping out reviews with a text length less than 80 characters. The 80 character limit was reached by looking at the 5% Quantile of the review lengths year to year. Review lengths at this metric only dip below 80 characters in early 2005 so it is a safe line to draw in the sand.

Our analysis will look at these 495,894 Restaurant reviews, split 455,987 for Non Fast Food and 39907 Fast Food Restaurants.



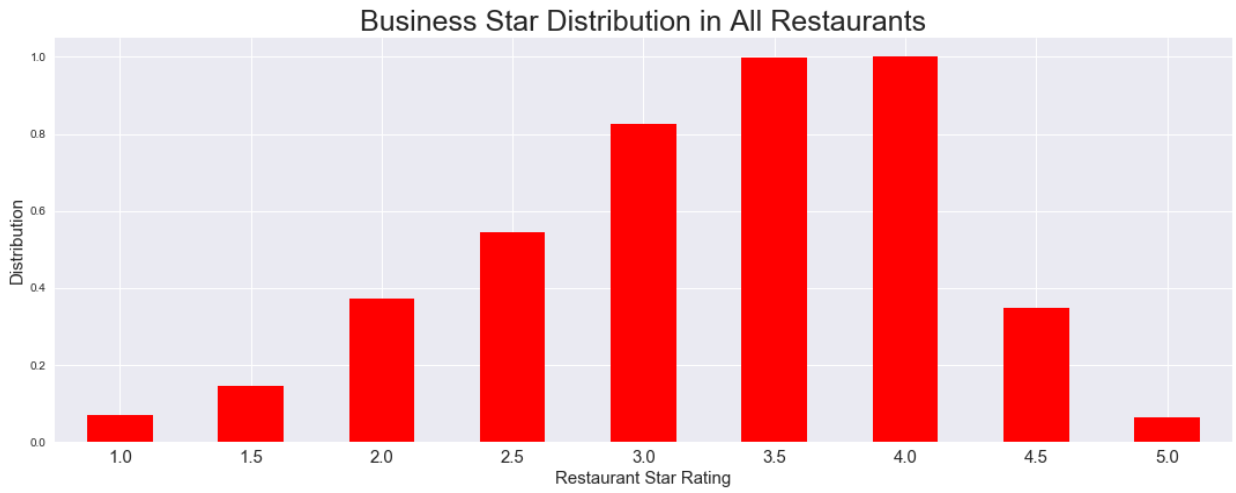
Tip Data

The original 'tip.json' file consists of 1,098,324 tips which we quickly reduce in the same method as the review data described above. We prune down to 1119,274 Arizona restaurant tips, with 12,052 Fast Food and 107,222 Non Fast Food Tips. Next, we prune out tips less than 10 characters with the same method as reviews. Finally, tips are scrubbed of commas, for the benefit of writing out to csv files.



Next Steps

A quick visual exploratory data analysis of the reviews we will be using show some interesting trends. Namely, reviews are skewed towards the higher 4 or 5 star range overall, but when subsetting only Fast Food Locations, we see a distinct difference, where it looks like reviews are much more likely to doll out criticism and low ratings.




Business Star Distribution in Fast Food Restaurants
Distribution
0.0
0.2
0.4
0.6
0.8
1.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Restaurant Star Rating


Review Star Distribution in Fast Food Restaurants
Distribution
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
Review Star Rating


Business Star Distribution in Non Fast Food Restaurants
Distribution
0.0
0.2
0.4
0.6
0.8
1.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Restaurant Star Rating


Review Star Distribution in Non Fast Food Restaurants
Distribution
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
Review Star Rating