Final Paper

Introduction

Whenever a person hears about a business or think about starting their own, they are always told that a business's success is all about location, location, location. It's hard to quantify how much of a business's success is determined by its location, but we were curious to see if location was linked to higher star ratings on Yelp. Yelp is a service that publishes crowd-sourced reviews about local business online. Additionally, we wanted to see if there was any relationship between certain attributes of restaurants and their ratings on Yelp. Addiontally, we wanted to explore the the behavior of reviewers when they travel to places outside their "home" location.

Aside from analyzing the success and behavior of reviewers and businesses, we also decided to explore the relationship between the text in reviews and the number of stars it received. In our project, we used a deep dataset of Round 7 of the Yelp Academic Dataset challenge, available at: [https://www.yelp.com/dataset_challenge/dataset]. The dataset includes five json files of business, check-in, user, review, and tip data, but our research only involved looking at the business, user, and review datasets.

Exploring the Data

1. Data Properties

The dataset includes 2.2 million Yelp reviews for 77,455 businesses across the United States. To gain access to the dataset, a name and email had to be provided and the terms of use had to be accepted. The schematic for the business, user, and review datasets were:

{ 'type': 'business', 'business_id': (encrypted business id), 'name': (business name), 'neighborhoods': [(hood names)], 'full_address': (localized address), 'city': (city), 'state': (state), 'latitude': latitude, 'longitude': longitude, 'stars': (star rating, rounded to half-stars), 'review_count': review count, 'categories': [(localized category names)] 'open': True / False (corresponds to closed, not business hours), 'hours': { (day_of_week): { 'open': (HH:MM), 'close': (HH:MM) }, ... }, 'attributes': { (attribute_name): (attribute_value), ... }, }

{ 'type': 'review', 'business_id': (encrypted business id), 'user_id': (encrypted user id), 'stars': (star rating, rounded to half-stars), 'text': (review text), 'date': (date, formatted like '2012-03-14'), 'votes': {(vote type): (count)}, }

{ 'type': 'user', 'user_id': (encrypted user id), 'name': (first name), 'review_count': (review count), 'average_stars': (floating point average, like 4.31), 'votes': {(vote type): (count)}, 'friends': [(friend user_ids)], 'elite': [(years_elite)], 'yelping_since': (date, formatted like '2012-03'), 'compliments': { (compliment_type): (num_compliments_of_this_type), ... }, 'fans': (num_fans), }

2. Data Pre-Processing

Data pre-processing were the steps taken to collect and prepare the input data for data mining and visualization.

Stage One: JSON to CSV

Converting the raw data from JSON was quite easy using R libraries. We began by reading in the data and writing the tidied data-frame to CSV objects to be read in our analyses. Once this was completed, other R libraries were used to wrangle and tidy the data for their final transformations.

Stage Two: Data Wrangling

We began by filtering for businesses with more than 300 reviews in the United States. Looking more closely at the business dataset, we noticed a nested data-frame of attributes for each business, so we cleaned and gathered the data so that each row contained one business attribute. Business data set was used to see what makes a business popular and successful; especially by looking at where most-reviewed are clustered and the influence of major business attributes.

Review data was cleaned for sentiment analysis using a library called "tidytext". After reading in the review dataset, we randomly sampled 500,000 observations. With this, we tokenized each word in a text review into a single row in the data-frame and removed stop words like "a", "and", "the", etc. since they had no indication of the reviewers' attitudes. Using built-in lexicons, we assigned each word a valence score and computed the mean score for each review. Additionally, we created per-word summaries, computing the number of business and reviews the word appeared in and the average star rating based on each review. We filtered for words that appeared in more than 10 businesses and 500 reviews to exclude rare words with strange valence scores and ratings. Using these cleaned datasets, we were able to create visualizations and make inferences discussed later.

Lastly, business data, user data and review data were all used to analyze the users' behaviors when they are in unfamiliar locations User data was cleaned to filter for active users with a significant amount of reviews. Cleaned review data and business data were then joined together with user data. We created an additional boolean column called 'most_revieSwed', after computing a user's "home" (state where they reviewed the most). Using this, we cleaned up and created a new data frame with average star ratings of familiar places and unfamiliar places per user. In the next section, we discuss this further along with results.

Stage Three: Pre-Processing using 'tidytext' for Sentiment Analysis

Sentiment analysis, or opinion mining, makes use of natural language processing to analyze pieces of text and determine the attitude of the speaker or writer. To perform sentiment analysis, we researched various text mining libraries. First, "**tm**" was used to remove stop words, common English words that would give no information, but the process of splitting text reviews into one word-per-row and removing stop words was too inefficient. After more research, we came across "**tidytext**", a library for text mining. Using this library, cleaning and aggregating the review data was extremely easy and efficient. We then performed tidy analysis to strip reviews and tokenize each word into a single row in the data-frame.

To explore this data, we researched different lexicons and found ANEW (Affective Norms for English Words), a "sentiment lexicon" that scores words for valence. However, upon further research, we came across an article written by Finn Arup Nielsen, where he compares a sentiment lexicon he created to ANEW. His lexicon, AFINN, contains a list of English words that are assigned an integer value from -5 (negative) to +5 (positive) based on the word's sentiment. He concluded that his word list better classifies sentiment rating than ANEW on Twitter sentiment analysis. Since Yelp text reviews are similar to tweets, we used AFINN for this sentiment analysis.

Results and Discussion

A. Location Mining and Business Attributes

Our first hypothesis was that successful businesses would be clustered at certain states. We examined the data set and wanted to know about the relationship between the number of reviews and the star ratings. It turned out that when number of reviews increases, the star ratings gets clustered around 3.5 out of 5. Hence, we hypothesized that high number of reviews, especially those locations with most reviewed businesses, will have positive influence to the star ratings.

We first tried to plot this information with points on a map using longitude and latitude of the businesses. Points with different sizes stacked up on each other, so the size and color of the points did not tell much

information. Even though 900+ points were plotted, there were really only six primary clusters of points so there was no point in having so many small individual points.

Instead, we ended up with a bar graph, which clearly shows which state has the most number of businesses with over 300 reviews. There were 6 distinct regions in total. Interestingly, the graph shows that the majority of the businesses is located in two states: about 62% of the filtered businesses were located in Nevada and another 31% was in Arizona.

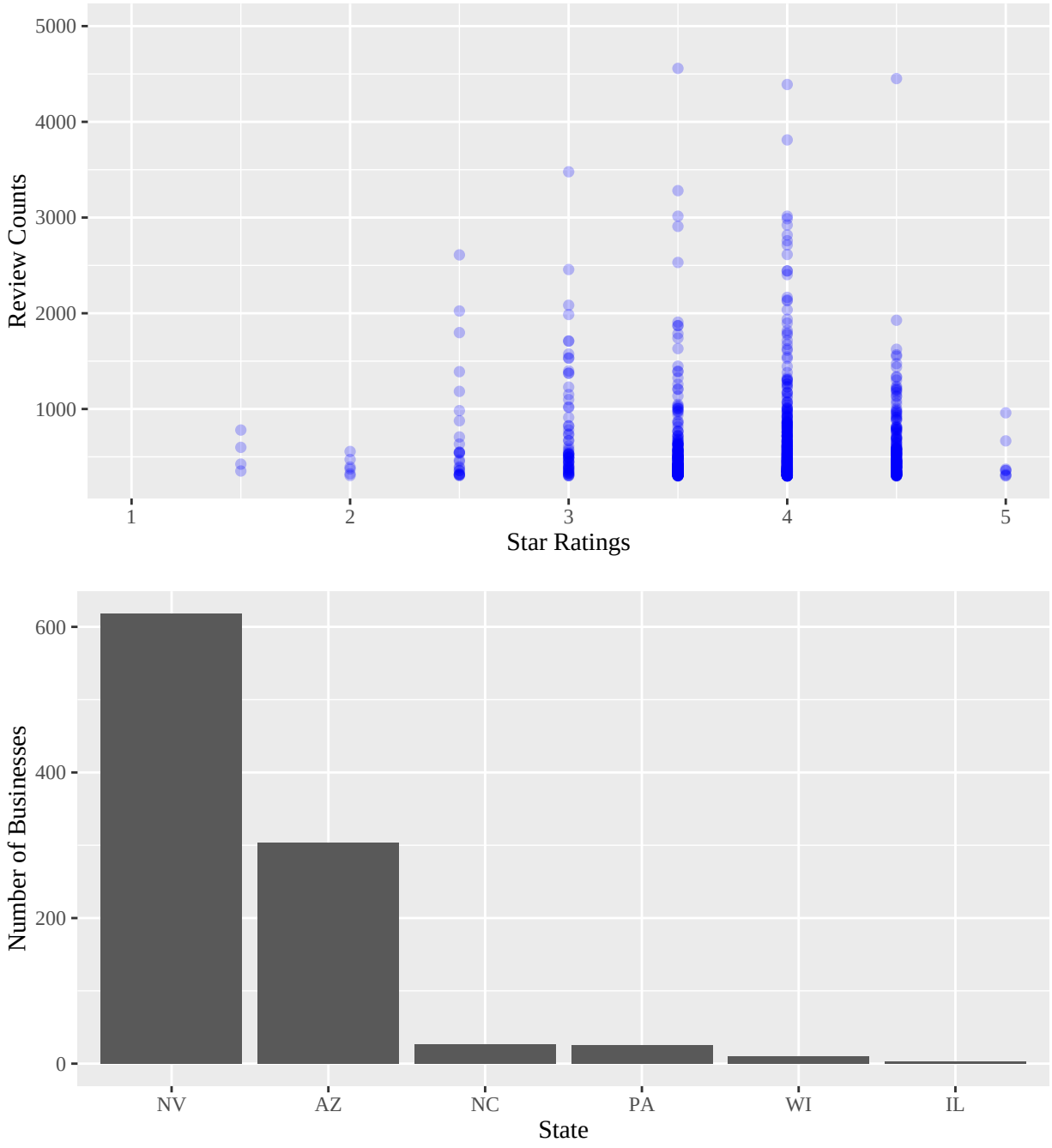

Figure 1: Number of businesses with more than 300 reviews

We also hypothesized that there were certain business attributes that would help businesses succeed in terms of number of reviews and ratings. We first tried to include two representative plots, 'Noise Level' and 'Attire

Type'. However, subsetting too much of the data set made us lose valuable information about the dataset. Choosing only two random attributes and analyzing was too narrow an exploration of the data.

As such, six major attributes were chosen. There were slight differences in star ratings based on some categories. For example, businesses that accept credit cards seemed to have higher ratings compared to those that don't. Interestingly, restaurants that do not offer delivery showed slightly higher star ratings than those who do. There were no major findings from the plots, however. Except for slight differences in average ratings of certain attributes, there was no significant difference evident. The user review ratings were evenly spread out even if the specific attribute existed or not. Our findings pointed out that business attributes do not have huge influences over user ratings.

Figure 2: User ratings on business attributes

B. Sentiment Analysis

We hypothesized that words with higher valence scores would produce higher Yelp star ratings for businesses. First, using the R library "wordcloud", we formed two word clouds for the 100 most frequent words in text reviews with one-star and five-star ratings to determine if there was a distinct difference between the words used.

Figure 3: 100 most common words for one-star and five-star rated businesses

While words like "food" and "service" were the most common, there is a noticeable difference between the two word clouds generated. To further explore this difference, we generated boxplots of mean sentiment score versus business star rating to crudely determine if there was a positive trend for valence score and business rating.

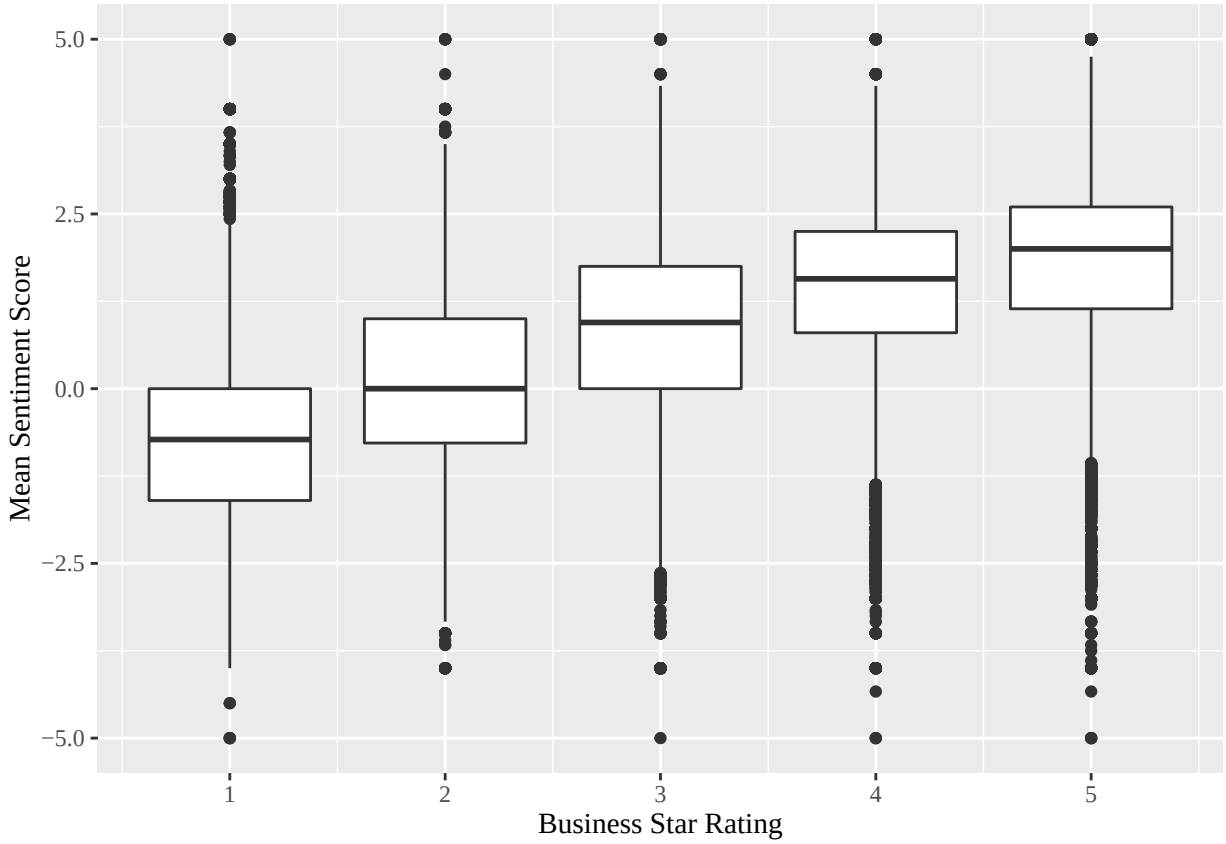

Figure 4: Mean Sentiment Score vs. Business Star Rating

There was indeed a positive correlation, despite some words with high valence scores appearing in reviews with low ratings and vice versa. Now to determine which words were indicative of a positive review and a negative review and the accuracy of our valence score classifier, using the word summaries data we generated a scatterplot labeled by word, colored by AFINN score, and plotted by average star rating versus word frequency. Additionally, we overlaid a horizontal line of the mean restaurant rating as a metric to determine the accuracy of our classifier. We would expect all words above to be blue and all words below to be red.


```
## [1] "Positivity Accuracy: 0.643410852713178"
```

```
## [1] "Negativity Accuracy: 0.869918699186992"
```

Figure 5: Average Yelp Rating vs. Frequency of Words

For the most part, our classifier was accurate. 65.02% of positive words sampled lied above the mean restaurant rating, and even better, 86.59% of negative words lied below the mean star rating. Therefore, it

seems reasonable to infer that there is indeed a positive correlation between valence score and Yelp business rating.

However, we also wanted to see these words plotted by their valence score versus average star rating to visualize the actual trend between sentiment and rating, rather than just viewing what percentage of words lied above the mean rating.

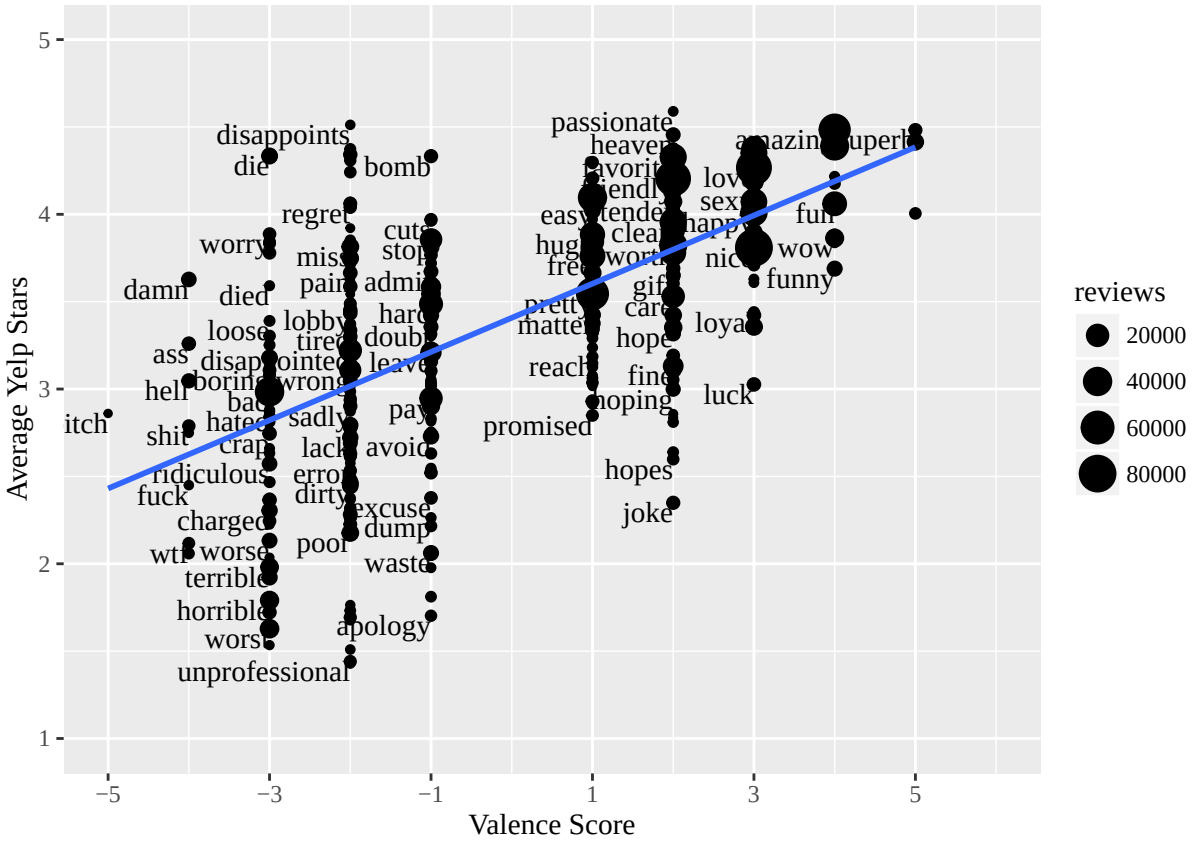

Figure 6: Average Yelp Star Rating vs. Valence Score

Based on a simple linear regression line, words with higher scores are clustered towards 4 and 5-star ratings, while most negative words have pretty low average ratings. Some words like "damn" had higher star ratings despite their low AFINN score, but that could be attributed to writers' different expressions like "damn good" for really good food.

Lastly, we wanted to see what the mean rating was for each AFINN score since that would clearly indicate a correlation. We expected that lower scores would have lower mean business ratings, and higher scores would have higher mean ratings.

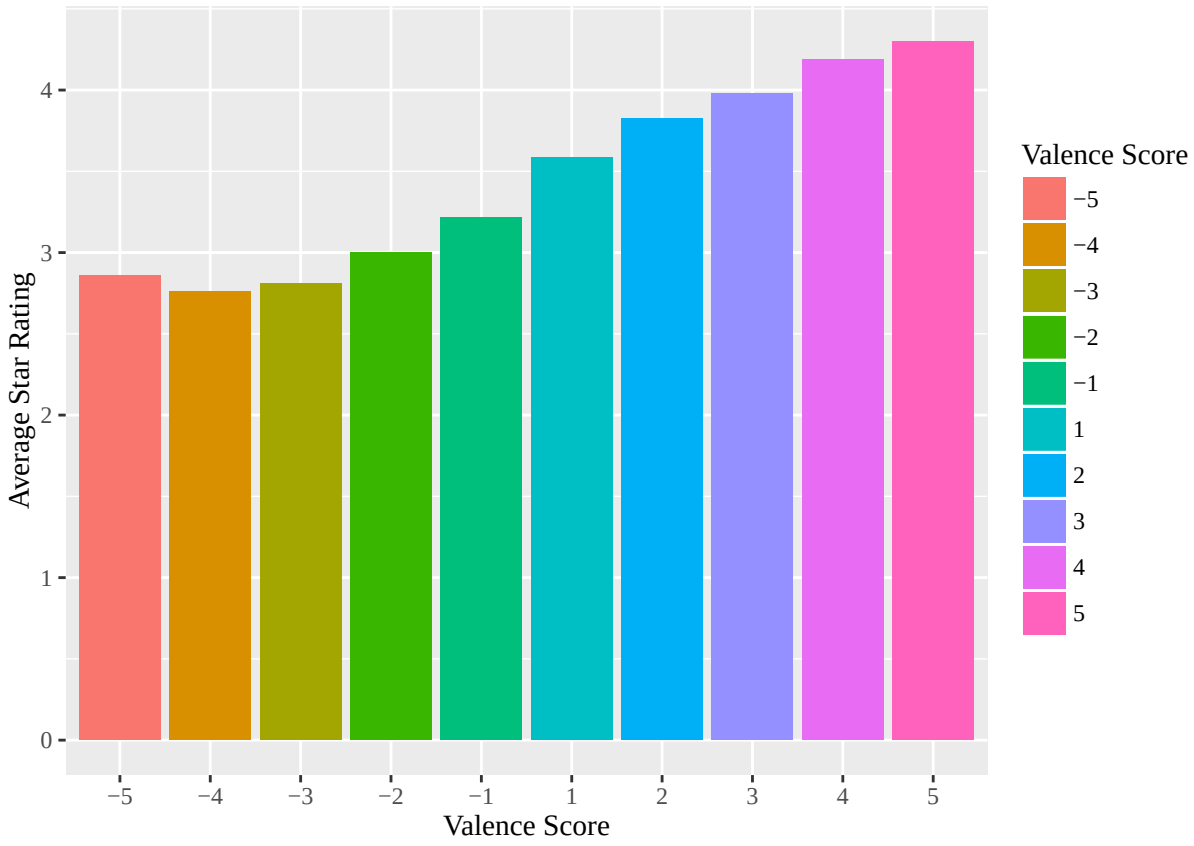

Figure 7: Bar Graph of Average Yelp Rating per AFINN sentiment score

As expected, higher AFINN scores did in fact result in higher mean business ratings. Interestingly, words with a -5 AFINN rating had a higher rating than words with a -4 and -3 AFINN rating. We're not exactly sure what would've caused this, but it seems like an interesting thing to explore for further research.

C. User Behavior when Travelling

Looking at the user and review datasets combined, we wanted to find out what would affect people's behavior of rating businesses. Since the dataset given by Yelp was academic, this did not include every single review by every user in data. We filtered to focus on active users that have written more than 50 reviews in total and have 10 reviews. This seemed like a reasonable constraint on the data that would remove any outlier behavior and would still capture the overall trend in user behavior when travelling.

We hypothesized that users would give relatively higher star ratings when travelling than when visiting local restaurants that they are familiar with.

First, we faceted the graph after selecting the top 15 users with most reviews. However, since facet focuses rather on individual data than overall trends, it was hard to conceptualize the findings. Individual patterns varied, so to capture the overall trend, we plotted their behaviors into one graph.

Another approach we took was to plot users' ratings in a histogram. From the filtered user dataset, one familiar region was chosen per user. This was based on where the user wrote the most reviews. Two average star ratings were calculated for a user: one for reviews in familiar state and another from reviews when travelling to unfamiliar places. The best visualization for such a graph was using a density plot, shown below.

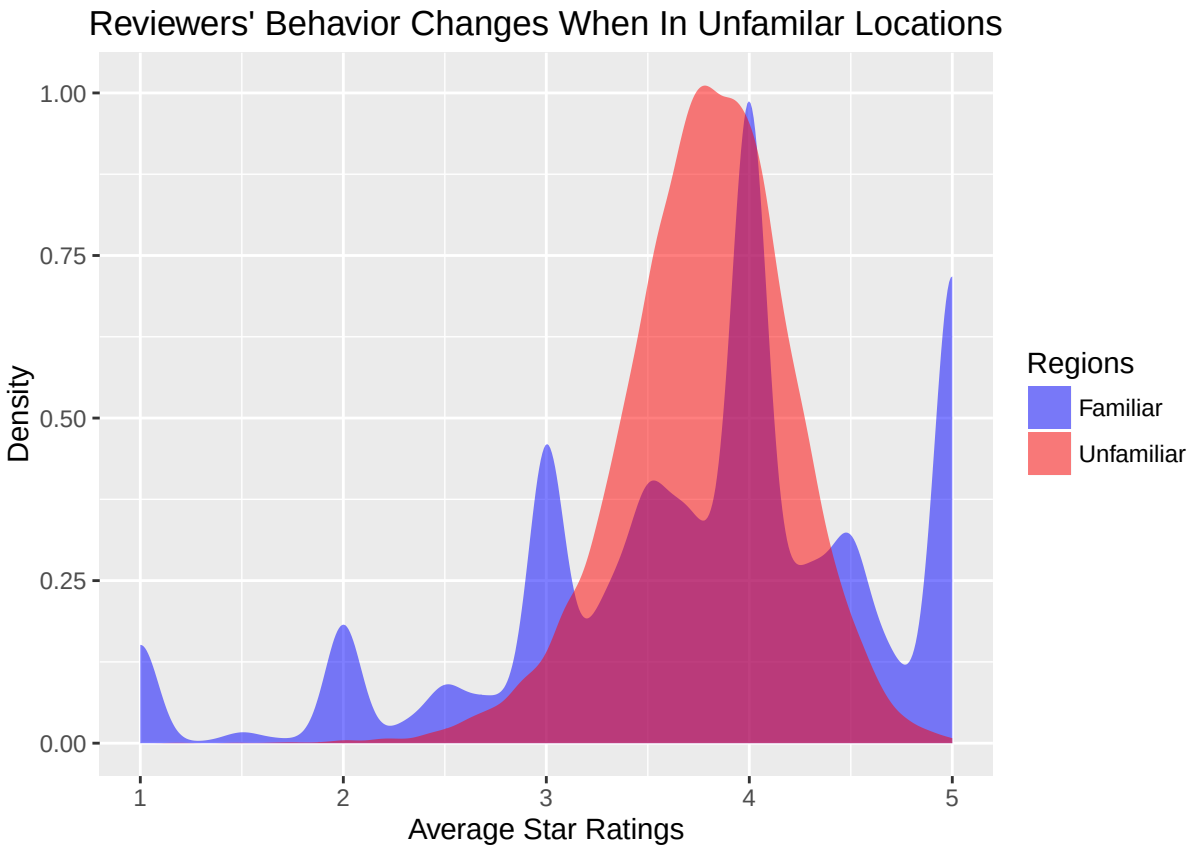

Figure 8: "Reviewers' Behavior Changes When In Unfamilar Locations"

The density curve for unfamiliar region is smooth and the majority of users tend to give star ratings in between 3 to 4.5 out of 5. Number of people giving 1 or 2 stars while traveling is barely seen. On the other hand, the ratings in familiar places were rather widely distributed. There was a significant portion of users giving star ratings from 1 through 5, even though the peak of the curve is around 4.

The mean rating was 3.718 for unfamiliar and 3.762 for familiar places. However, the standard deviation for the familiar locations plot was 0.41, and unfamiliar one was 1.37 so star rating, explained by the variablity between users who travelled.

Conclusion

In conclusion, we investigated several questions related to the Yelp datasets, and our results were promising.

There is a close relationship between star ratings and number of reviews–the popularity of the business. There are certain locations where most reviewed businesses are clustered; hence, location also has some influence over the star ratings. However, no significant correlations between business attributes and the review ratings were found.

Using natural language processing and text mining, we confirmed our original hypothesis that a positive correlation exists between AFINN sentiment score and star rating. 86.59% of words with negative scores appeared in reviews that were below the mean rating of all reviews from our sample.

Lastly, by examining the behavior of users while travelling showed that most users tend to give mediocre ratings when in unfamiliar places. On the other hand, the ratings were evenly distributed while in familiar locations. Surprisingly the mean of each situation was about the same.

Future Work

We would use cross-validation to create a training set and test set for our sample of reviews from the dataset. We can create a classifier using sentiment analysis and test our trained classifier against the test set to predict business ratings based on the text of the reviews. Additionally, we could look at possible correlations between restaurant reviews and ratings and tips, looking at the tips data provided with the dataset.

At last, we could use full data set instead of an academic, truncated data provided for data challenge. This way we could utilize more information to make even accurate plots and analysis. Using this, a comparison of user trends in different countries will be interesting to further investigate.