CAPSTONE PROJECT #1

Yelp Recommendation System

Author: Jerome Gonzaga

Problem Statement

Current Issue: It can often be very difficult to find a good restaurant, good mechanic, good barber shop, good cafe, etc. People can find a business on yelp and see their associated star rating but it's hard to trust the star rating. Many times I find that people look at reviews and photos to determine if the business fits what they are looking for and even this method can result in mixed conclusions about a business.

Problem to solve: Is there a way to optimize searching for a business based on other characteristics? If so, what sort of patterns in the dataset can we find in order to base the searches on?

Data Used:

1. https://www.yelp.com/dataset

Documentation for the available data: https://www.yelp.com/dataset/documentation/main

Data Wrangling

The following datasets, which all come from yelp.com, were used for this project:

1. business.json
2. review.json
3. user.json
4. checkin.json
5. tip.json
6. photo.json

Each of the JSON files was extracted and converted to CSV files. The CSV files were then extracted to Pandas Dataframes using the read_csv() Pandas function. Some of the columns in the datasets needed to be cleaned up further. The following actions were performed on some of the datasets:

- Unpacking nested dictionaries within some of the columns and add additional columns for the nested attributes.
- Reformat 'date' columns to proper 'datetime' formats
- Fill in empty data points with blank entries
- Remove entries that included 'null' hours

Here is a tabular description of each of the cleaning up DataFrame datasets:

Dataset	**Attributes Included**
df_business	location data, attributes, and categories
df_photo	Photo id value, associated business id value, caption, category label of photo

df_review	Full review text, user_id of reviewer, review rating, review date
df_tip	Tips and quick suggestions written by users for a business
df_checkin	Business id values, comma-separated list of check-ins
df_user	User's name, number of reviews, votes sent and received, average rating of all reviews, date user joined Yelp, unique user id

To clean up the large dataset further, I filtered out the df_review DataFrame to only include one yelp user. Given the information from the Yelp datasets, the users with the most amount of business reviews available would provide the most information to work with for further analysis.

Initially, I was going to filter out the datasets by city but after critically thinking about this project, it made more sense to base a Yelp search off of a user's review history. A user's review history provides insight into what the user likes and dislikes about certain businesses.

The business DataFrame that I created from the business csv file needed some clean-up as well. To do so I observed and did the following:

1. The "attributes" column had a lot of information in the column. Each entry in the "attributes" column was a dictionary that was represented by a string.
2. Each entry in the "attributes" column was converted from a string type to a dict type.
3. There were nested dictionaries within each dict
4. I split the dict in the "attributes" column into separate columns for each attribute listed in the dict.
5. The 'hours" column also had dicts represented by strings. I split up the dict in the "hours" column by each day.
6. The "hours" column was a datetime format that is not conventional. The column was adjusted to a proper python datetime format.

To deal with any missing values I performed the following:

1. All values that were 'NaN' were changed to blank cells
2. Any rows that were blank or had "NaN" in the 'hours' column were deleted. The rows that did not have any data in the 'hours' column were more likely to be businesses that closed down.

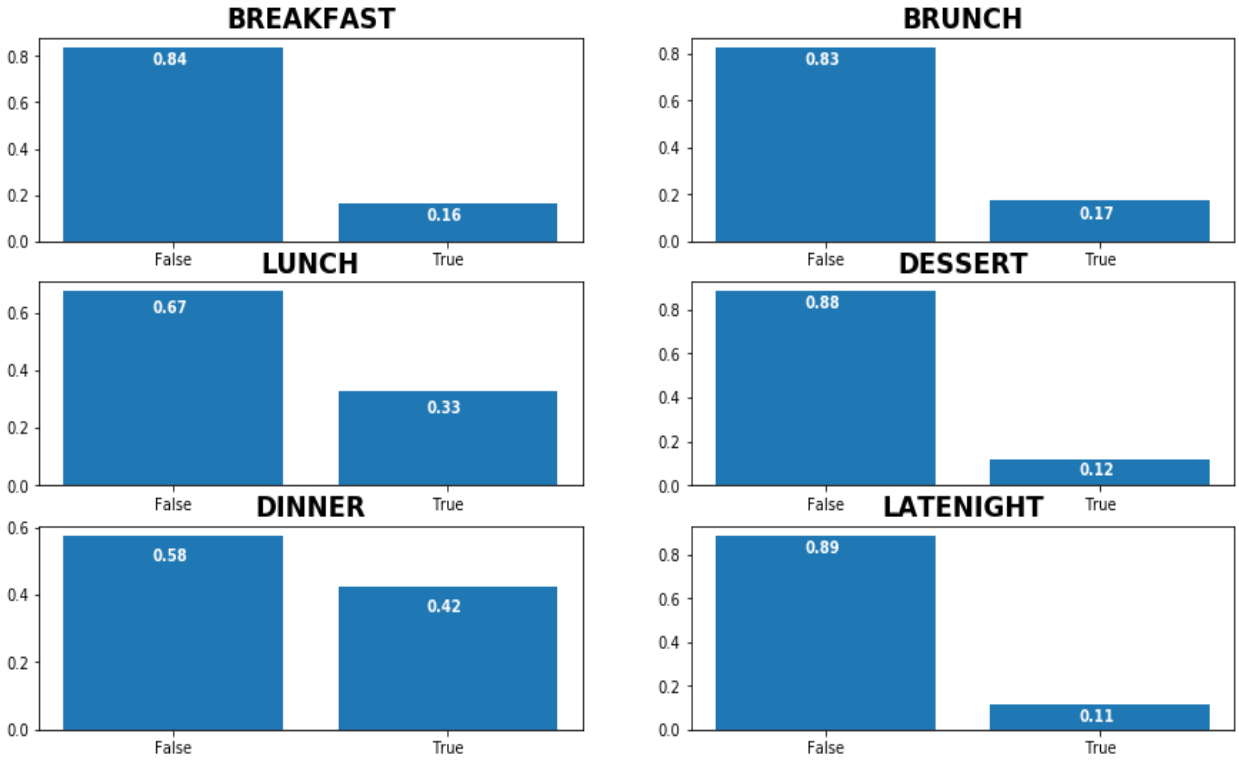
I used the describe method on each of the DataFrames and after reviewing the results of the describe method, it did not seem like there were any outliers. One observation to note about the data though is that some of the top 25-percentile of the data was much larger than the mean and the mode. These data points were kept and assumed as valid entries. Sometimes there are real users or businesses that have thousands more reviews compared to others.

Data Story:

There is a lot of data from the Yelp datasets and it was a challenge to figure out where to start analyzing the data. As a way to begin investigating the data, I started to look at some of the

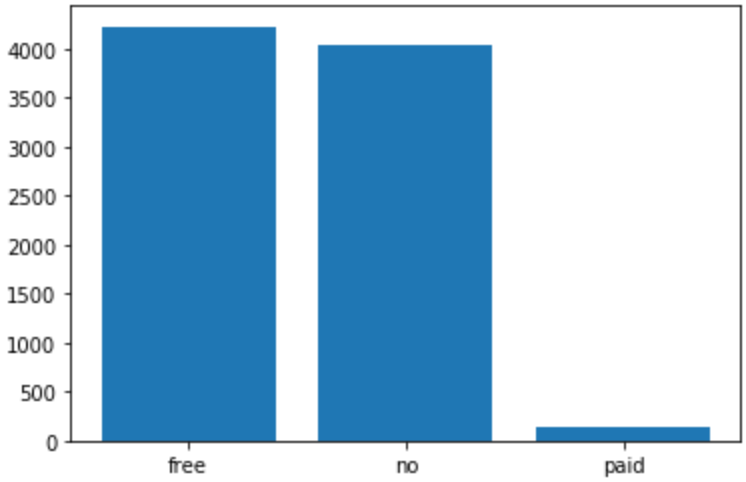
attributes listed for businesses. As a personal interest, I started by plotting some numbers regarding restaurants to see how many restaurants consider themselves good for certain meals.

The following plot shows the percentage of restaurants that self-reported whether they were good for different meals throughout the day:

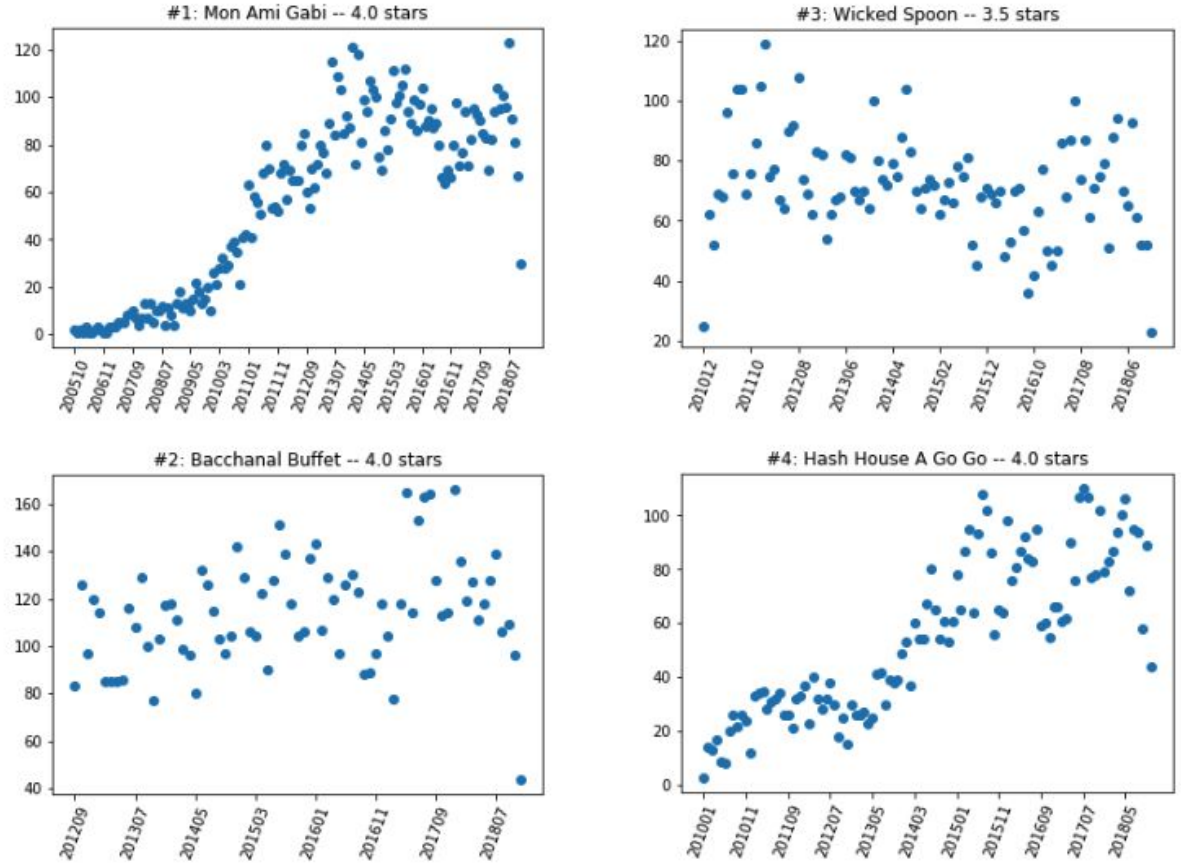


I thought that this was an interesting plot to investigate because it is often the case that people look for restaurants to go to through Yelp. So, what this plot tells me is that a lot of businesses state that they are not good for certain meals, but there are a good number of businesses that state they are a good place for dinner.

Another interesting investigation was taking a look at whether a business offered free WiFi, paid WiFi or no WiFi at all. This attribute is a common desire for many people because of the need to surf the web or check email. The following plot shows the number of businesses for free, paid, or no WiFi:



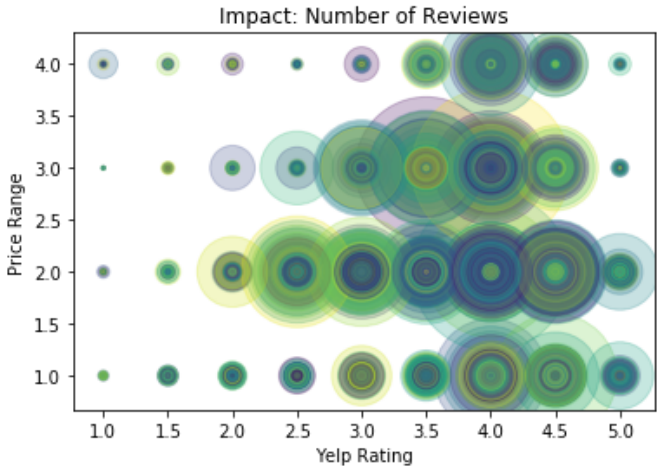
Then as a way to start investigating data associated with certain businesses, I then plotted the number of reviews received by a business over time. As an example, here are plots of the top 4 businesses that had the most number of reviews in the Yelp dataset - time versus number of reviews received in a month:



It is interesting to note that some businesses had a positive slow for the number of reviews received in a month over time while some top businesses had a negative or even close to 0

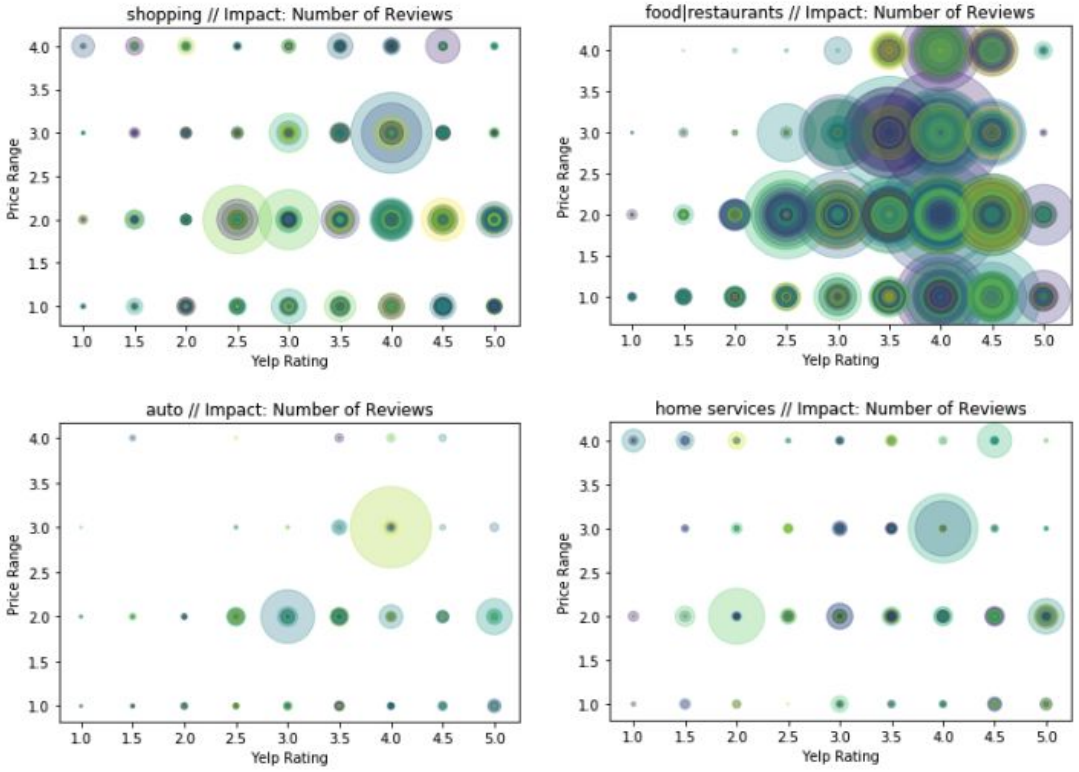
slope. What this shows me is that there are certain businesses that were well liked and visited/reviewed compared to other businesses.

One other interested plot that I created to investigate the Yelp dataset is the price range of the business versus the posted Yelp rating versus the number of reviews that a business had:



The larger the circle around a certain data point tells me how many reviews a particular business had. So, what this plot above tells me is that there are a lot of businesses that have a price range of about 2 to 3 dollar signs and a posted Yelp rating of 3 to 4.5 stars that especially have a lot of reviews. This was intriguing to find because it tells me that most people visit and review businesses that have a medium price range and are slightly above average star rating.

The following graphs tell me that the most amount of reviews can be found for restaurants and food businesses, compared to auto businesses or even shopping businesses, which means that it will be worthwhile to put more investigation into how people choose to visit or review businesses under the food and restaurant category.



As a way to begin diving into what Yelp users say about businesses under the food and restaurant categories, I created this word cloud to show the most popular words used in tips for these businesses. From an initial look at the most popular words used in tips for restaurants, looks like people made suggestions about a restaurant's happy hour, whether or not something was amazing, tips about burgers, an also tips about food prices at a restaurant.

Based on all the plots, I can tell that a lot of reviews and tips were made for businesses specifically under the food and restaurant categories and there seems to be a convergence on the most popular price ranges and posted Yelp ratings that people will visit and review. So, as I dive further into this project, I can focus on the posted Yelp ratings, the price range, and the text used on the reviews and tips in order to create some sort of recommendation system.

Thus one hypothesis that I can test is the following:

- A user looking for a restaurant or food business to visit will likely desire to go to (and positively rate) a business that has a medium price range, more reviews, and a slightly higher than average posted Yelp rating.

Statistical Data Analysis:

To test this hypothesis further, I am going to select the Yelp user with the most amount of posted reviews so that I have the most amount of data to work with. One thought that came to mind regarding this hypothesis is that a Yelper will typically decide on whether to go to a specific business based on the information of the business posted on Yelp - whether it is the business' Yelp rating, reviews from other Yelpers, the number of attributes, or even based on the photos. However, how much does this information really affect whether or not a Yelper likes the business? How much does a Yelper's rating of a business get affected by whether a business provides information on attributes, pictures, or even the number of tips or reviews?

Some questions I need to ask through the data: *Is there any relationship between the number of reviews, attributes, check-ins, tips, Yelp rating or quality of business pictures and whether a user up-rates or down-rates a business? If a business has similar characteristics to that of the business that the Yelper up-rated, is it more likely that a Yelper will go to that similar business?*

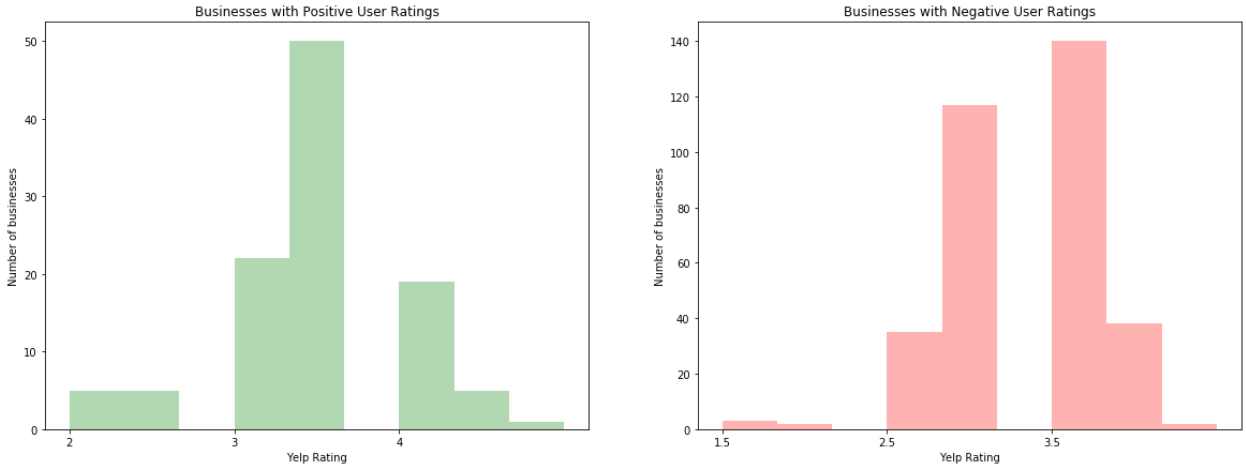
(Note: An up-rating is when the user rates the business higher than the posted Yelp rating. A down-rating is when the user rates the business lower than the posted Yelp rating.)

The Yelp user that I investigated further (user `CxDOIDnH8gp9KXzpBHJYXw`) has a total of 2056 businesses that the Yelper visited and reviewed under the food and restaurant category. It is interesting to note that the top types of businesses that this user went to within the restaurant/food category:

1. Chinese
2. Nightlife
3. Bars
4. Japanese
5. Coffee & Tea
6. Canadian (New)
7. Cafes
8. Breakfast & Brunch
9. Italian
10. Sushi bars
11. Desserts
12. Asian Fusion
13. Korean

This list of top types of food/restaurant businesses is very particular to this Yelp user and this kind of filtering should be considered within a Yelp recommendation system.

The following plot shows me the Yelp ratings and the number of businesses that the user positively rated (meaning rated higher than the posted Yelp rating) or negatively rated (rated lower than the posted Yelp rating).

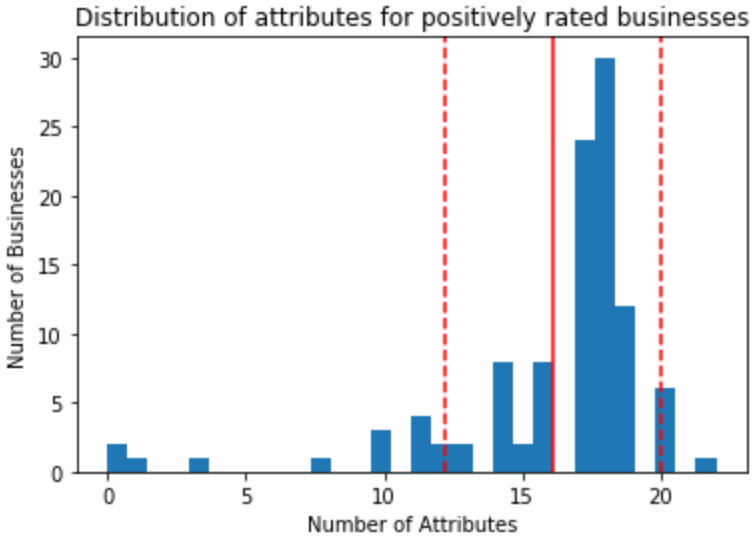


What this plot above tells me is that this user negatively rated many businesses with a 3 or 3.5 while positively rating most businesses with a 3.5 or higher.

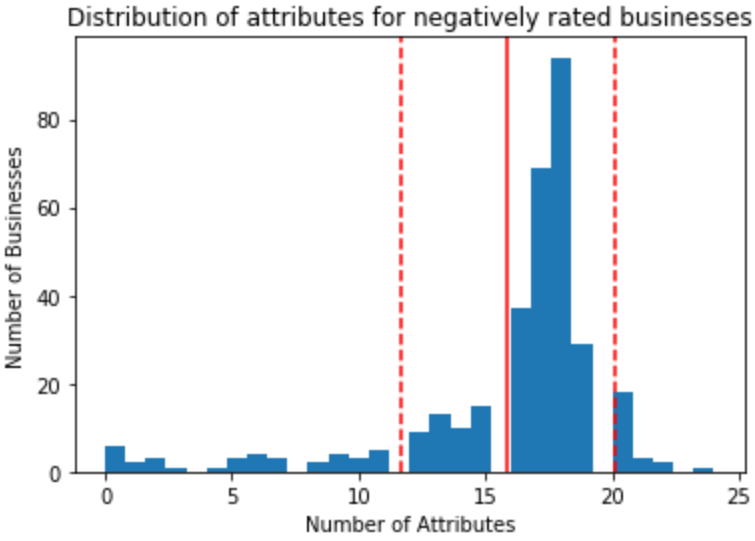
Now it will be interesting to investigate whether certain attributes or features of a business affect whether or not a user positively or negatively rates a business. The following parts describe an attribute that is investigated, a null hypothesis, an alternative hypothesis, a plot of the data associated with the attribute, and some statistical calculations and conclusions made regarding the attributes/features.

Part 1: Number of Attributes Posted

Null Hypothesis: The number of attributes posted on Yelp affects the Yelper's rating of the business
Alternative Hypothesis: The number of attributes posted on Yelp does not affect the Yelper's rating of the business



The mean and standard deviation of the number of attributes are 16.11 and 3.92, respectively.



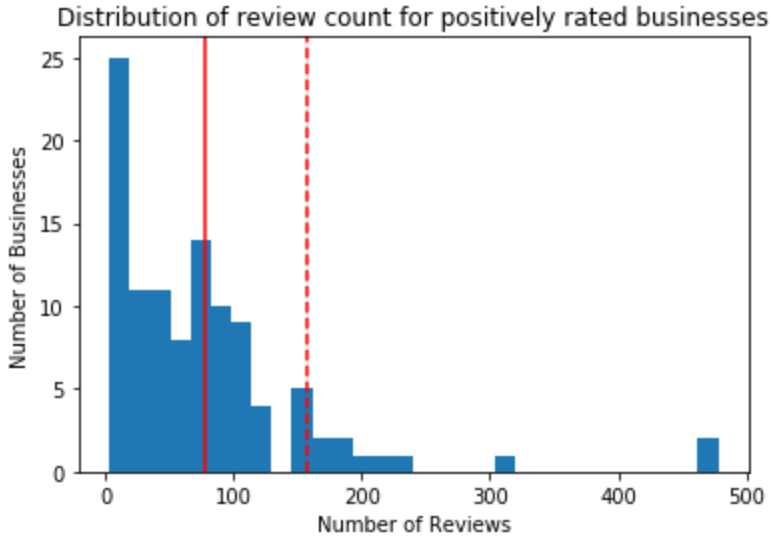
The mean and standard deviation of the number of attributes are 15.88 and 4.21, respectively.

The p-value is very high at more than 60%. The null hypothesis - that the number of attributes posted affects a Yelper's rating - does not have to be rejected.

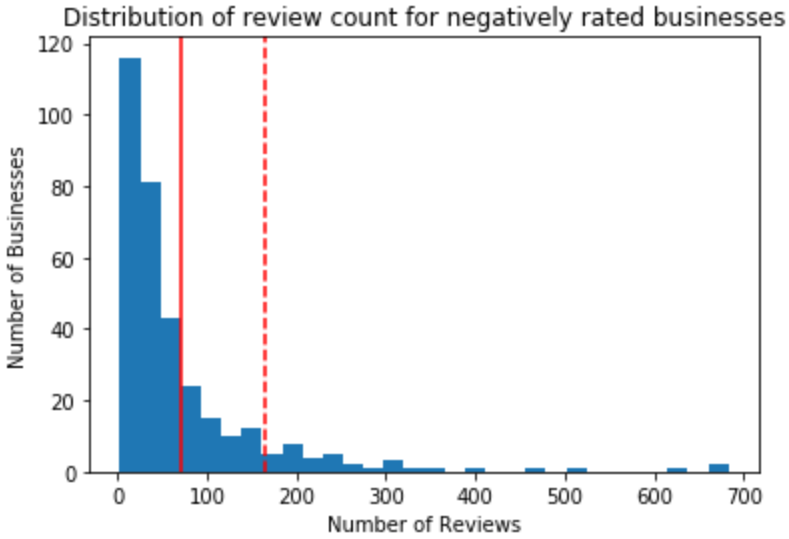
Part 2: Number of Yelp Reviews

Null Hypothesis: The number of reviews posted on Yelp affects the Yelper's rating of the business

Alternative Hypothesis: The number of reviews posted on Yelp does not affect the Yelper's rating of the business



The mean and standard deviation of the number of reviews are 77.32 and 80.03, respectively.

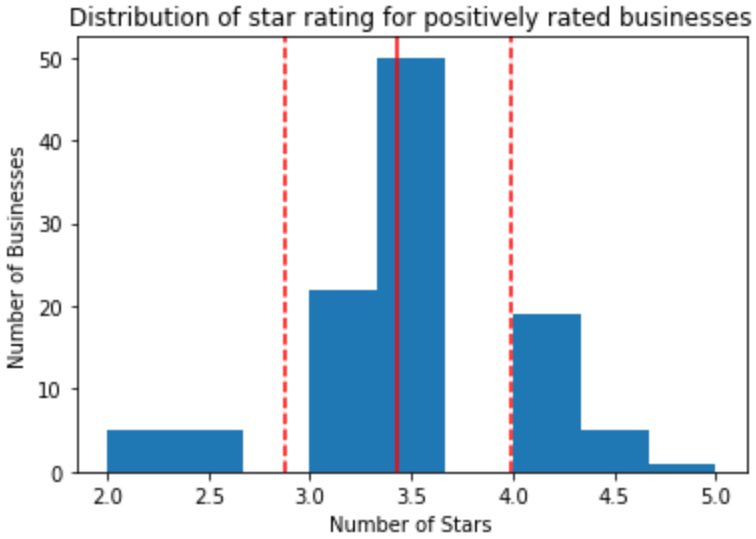


The mean and standard deviation of the number of reviews are 71.45 and 94.18, respectively.

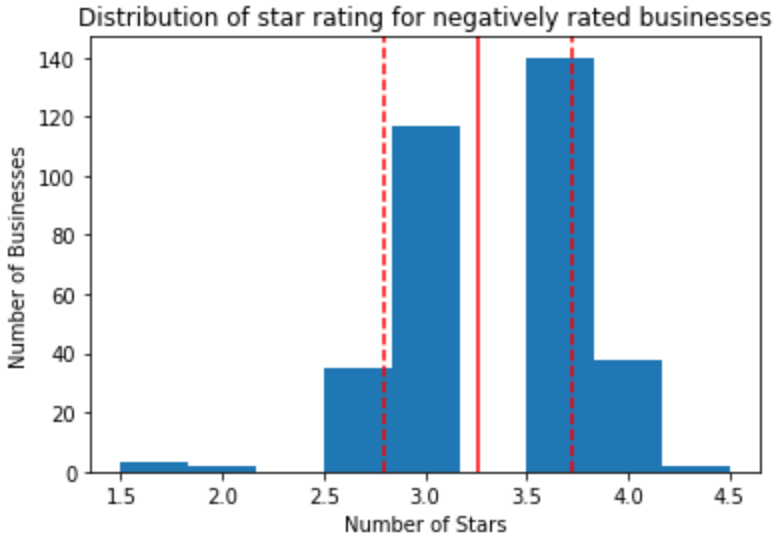
The p-value is very high at more than 50%. The null hypothesis - that the number of Yelp reviews online affects a Yelper's rating - does not have to be rejected.

Part 3: Yelp Business Rating

Null Hypothesis: The star rating posted on Yelp affects the Yelper's rating of the business
Alternative Hypothesis: The star rating posted on Yelp does not affect the Yelper's rating of the business



The mean and standard deviation of the number of reviews are 3.43 and 0.56, respectively.



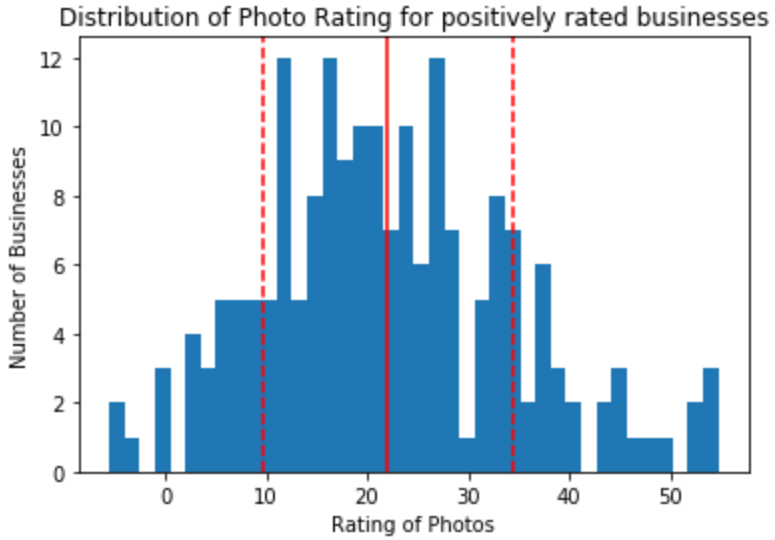
The mean and standard deviation of the number of reviews are 3.26 and 0.46, respectively.

The p-value is less than 0.5%. The null hypothesis - that the business' posted Yelp rating affects a Yelper's rating - can be rejected.

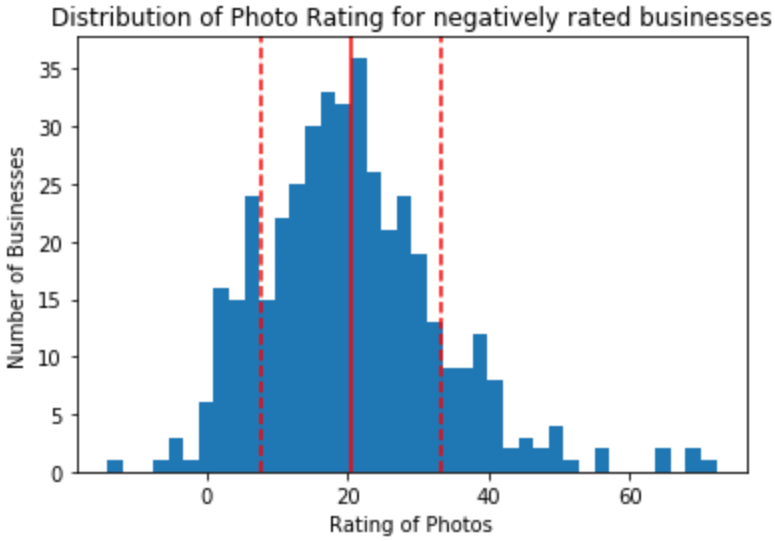
Part 4: Quality of Posted Photos

Null Hypothesis: The quality of the photos posted on the business' Yelp page affects the Yelper's rating of the business

Alternative Hypothesis: The quality of the photos posted on the business' Yelp page does not affect the Yelper's rating of the business



The mean and standard deviation of the photo ratings are 22.04 and 12.39, respectively.



The mean and standard deviation of the photo ratings are 20.53 and 12.74, respectively.

The p-value is somewhat high at about 17%. The null hypothesis - that the quality of the business photos affects a Yelper's rating - does not have to be rejected; however, it should be noted that based on the quality rating, the businesses that were negatively rated by the Yelper has a mean score that is better than that of businesses that were positively rated. This result needs to be further analyzed to see if the quality rating test is valid and trustworthy.

More data analysis and investigation will be required to determine how the number of reviews, the posted reviews, the business photos and the attributes can affect whether or not a user will like visiting a particular business or not.

In-Depth Analysis

As a way to start the in-depth analysis, the features considered for machine learning were the review texts written by users. The target for the machine learning model was to correctly predict the user's Yelp rating score given the review text that's fed to the model. The idea of the machine learning model is that the model gets all of the user's review text from a particular business category of interest and the associated Yelp score, analyzes the review text to determine the top words used by the user per Yelp rating score, then after being trained with this data, predict the Yelp rating score of the review text of that same particular business category from other user's. Based on the predicted Yelp rating scores, a list of recommended businesses is provided.

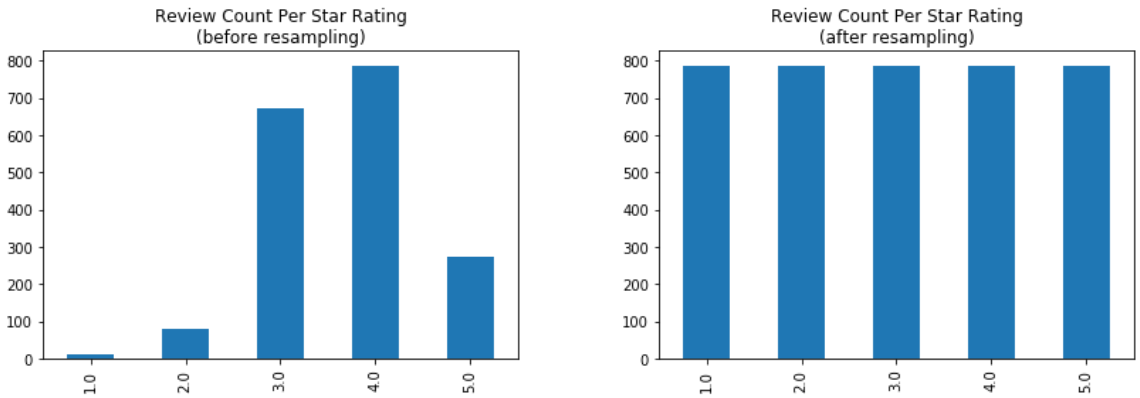
Building the Model

Since the Yelp rating scores associated with the review texts are provided, this is a supervised machine learning. There were a large number of features that could be considered for the model to predict the Yelp rating score. Some of the potential features that can be explored include:

- The number of check-ins a business received over a period of time
- The number of reviews received by businesses
- The text from the tips provided by users for businesses
- The businesses that a user has in common with friends and/or fans
- The attributes offered by a business

There are a large number of features that can be explored but for the purposes of this project, only a user's review texts were considered. In the future, one or even more of these other features can be considered in order to better predict a user's Yelp rating score.

Before the model analyzes the review texts, one significant step that was implemented was to ensure that each classification has the same amount of samples. Before performing this step, I had originally provided the model the unaltered review text dataset. Often, there would be more reviews associated with one or two of the Yelp rating scores (ie. more reviews for a rating of 3 versus a rating of 5). As a result, the model predictions were heavily skewed by the review texts for one or two particular rating scores. So, a resampling step was performed before the review text dataset was fed to the model so that for rating scores that had less review texts a random sampling was performed to duplicate review texts for a specific classification so that the number of reviews per classification would be equal.



Example of resampling review texts to equalize the number of reviews per star rating

This model takes each of the review texts, by row, and passes each review text through a pipeline. The pipeline first vectorizes the review using CountVectorizer(). During the CountVectorizer step much of the text preprocessing is done here including removing stopwords that are very common in the english language, stemming the words to get rid of variations in word tense, and also tokenizing review text phrases using 1-gram and bigrams, After each of the review texts are vectorized, the output matrix is scaled to now be heavily skewed by some of the outlier tokens. Then the matrix is then passed onto the TruncatedSVD step to find the features have a more significant impact on the prediction of the Yelp rating score. Lastly in the pipeline, the OneVsRestClassifier and LogisticRegression. The review texts are split into a training set and a test set in order to verify the performance of the model.

Evaluating the Model

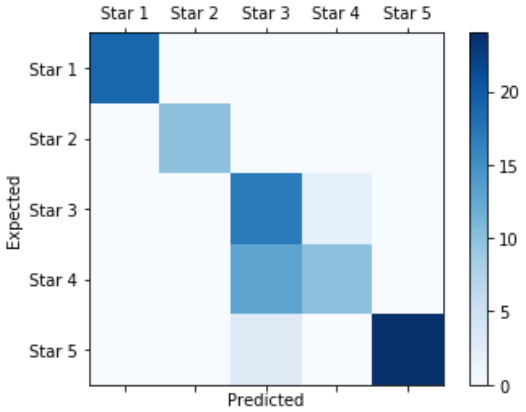
As an example, here are the output metrics and recommendations for one user for the business category of “breakfast, food, restaurant”:

Training Accuracy: 1.0
Testing Accuracy: 0.816

Classification Report:

	precision	recall	f1-score	support
1.0	1.00	1.00	1.00	19
2.0	1.00	1.00	1.00	10
3.0	0.52	0.89	0.65	19
4.0	0.83	0.43	0.57	23
5.0	1.00	0.89	0.94	27
accuracy			0.82	98
macro avg	0.87	0.84	0.83	98
weighted avg	0.87	0.82	0.82	98

Confusion Matrix:



The model determined that the top 20 tokens used as features for predicting Yelp rating scores were the following:

1. churro tot
2. 2 shrimp
3. 2 slice
4. 18 brunch
5. beauti
6. 10 awhil
7. 7 pm
8. abl walk
9. 10 clam
10. 6 ashley
11. allow use
12. 50 great
13. addendum
14. 3 dip
15. onlin went
16. 10th floor
17. almond orang
18. adjust
19. 3 salsa
20. 3 salsa

The reason for duplicates for the tokens is possibly due to the resampling preprocessing step performed on the review texts. The truncated words are a result of the stemming preprocessing.

Utilizing the Model - Providing Recommendations

After training and verifying the model, the model is used to provide recommendations to the user for businesses that most probably will fit the user's likings (based on the user's review text from other similar businesses). One consideration to note, however, for this recommendation system at this time is that the recommendations are based on both the quality and the quantity of the user's reviews. If a user does not have many reviews for a specific category of businesses, then the model will not have any past data to base the recommendations off of.

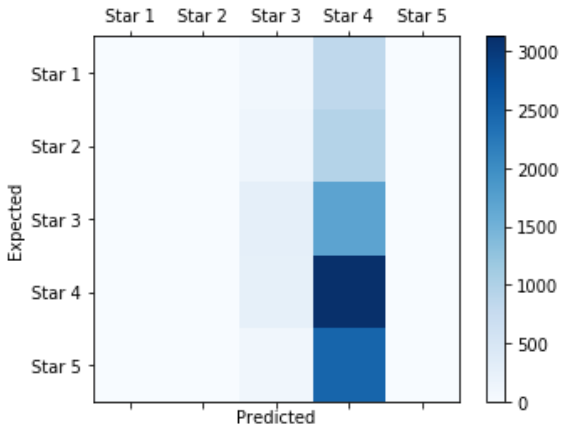
As an example of putting the predictive power of the model into practice, a city was chosen to find recommended businesses for a user to visit. The city Toronto was chosen for the large amount of reviews available for businesses in that area. The review texts for businesses of the same category as above from other users within Toronto was extracted from the df_review dataset and provided to the recommendation model.

It is interesting to see the classification report and the confusion matrix for the review text from other users.

Classification Report:

	precision	recall	f1-score	support
1.0	0.00	0.00	0.00	937
2.0	0.00	0.00	0.00	1110
3.0	0.32	0.14	0.19	1985
4.0	0.34	0.92	0.50	3386
5.0	0.00	0.00	0.00	2582
accuracy			0.34	10000
macro avg	0.13	0.21	0.14	10000
weighted avg	0.18	0.34	0.21	10000

Confusion Matrix:



The classification report and confusion matrix above shows that this user in the examples above would likely rate most other breakfast restaurants as a 3 or a 4 based on the user's review text. The fact that, based on other user's review texts, the predicted Yelp rating score is 3 or 4 shows that the model sees the commonality between the user's words for 3 and 4 Yelp rating scores but maybe the user's words to describe 5-Yelp rating scores are not used by other users. The model is relatively good at predicting businesses that the user would rate as 3 or 4, but not that good at predicting scores of 1, 2 or 5. The reason for this is possibly due to the primary features selected in the model.

Based on the model predictions, the businesses that this model would recommend are the following as a 5-star rating and a 4-star rating.

Predicted 5-star rated business:

1. Bonjour Brioche

Predicted 4-star rated business:

1. Egg Sunrise Grill
2. Nord Bistro

3. Trius + Aim
4. Bailey's Cafe
5. Eggsmart
6. Café Polonez
7. Sunny Morning
8. Huevos Gourmet
9. Rashers
10. Takht-e Tavoos

Future Improvements to Consider

Currently this recommendation model utilizes review texts from Yelp users in order to provide recommended businesses to visit. The model takes review texts from a user, analyzes the text, provides a Yelp rating score based on the text and then provides recommendations based on what the reviewer writes. However, there may be many other factors that contribute to the reason a user rates a business a certain score. Some of these factors can include, but are not limited to:

- Hours of operation
- Available business attributes (ambiance, wifi availability, parking, delivery, etc.)
- How new (or old) a business is
- Number of reviews that a business has
- Location
- Number of tables
- Busyness of business
- Quality, look or feel of the business, food, drinks, service, atmosphere, etc.

As an example, given that the testing accuracy of the model is about 81.6%, there is room for improvement and the additional factors above can be added to the model in order to provide better predictions for the Yelp rating score. In the future, the model can be modified and updated to include more features that are hopefully statistically independent.