Juliette Wong
jnwong
Due: 10/1/2020

Assignment #2: Exploratory Data Analysis for Yelp data

Introduction

Yelp.com is one of the most popular restaurant and merchant information websites in the United States. It allows users to rate different businesses, place reviews about their experiences, leave tips for other potential customers, and check-in to restaurants and stores. Many individuals use a combination of these to determine whether or not to go to a restaurant or store.

Since the restaurants' and stores' reviews and star ratings are crowd-sourced, there is often a lack of consistency in ratings. Some customers may be harsher raters than others, leading to lower ratings despite having the same experience. Also, other factors, such as location and when the customer visited the business, may affect how a customer gives ratings.

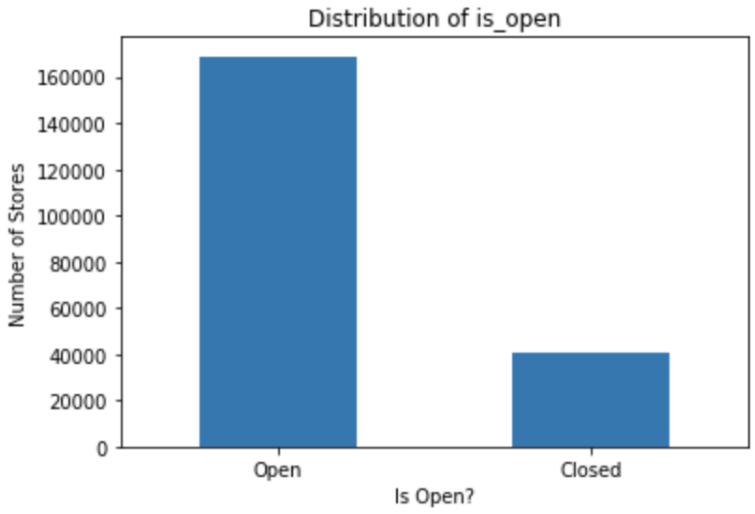
This report explores the following questions:

- What factors contribute to the number of stars received on Yelp?
- Are the categories of stores the same across different locations?
- How has the user's interaction of different Yelp features changed over time?
- What is the difference between reviews and tips?

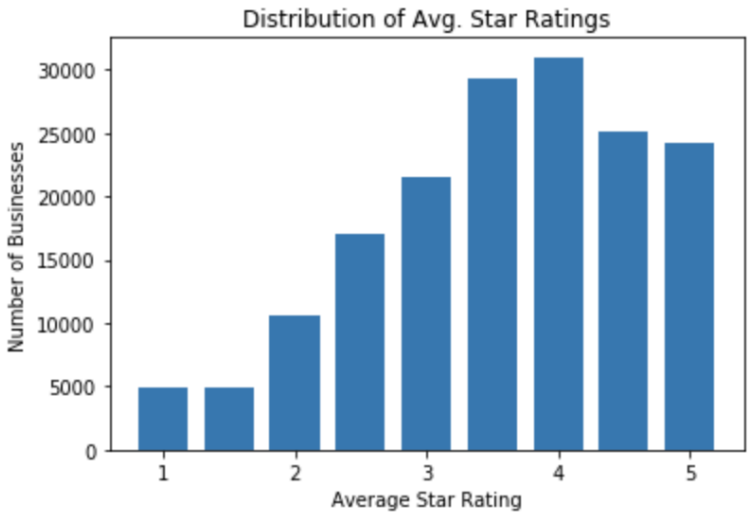
Methods

This report uses multiple related JSON files published by Yelp. This dataset does not contain information from all businesses available on Yelp but instead focuses on the metropolitan areas Montreal, Calgary, Toronto, Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas, Madison, and Cleveland. In total, there are 209,393 businesses in the dataset. The business file includes a unique business ID, the name, geographical information, average star rating, number of reviews, whether the company is open, and the categories and other attributes of the store. The reviews file contains the date of the review, the associated business id and rating, and the text from the review itself. The tips file contains the business id, a timestamp, and the user's suggestion. Finally, the check-ins JSON file includes the business id and a list of dates when a customer has checked in. This dataset and its documentation can be found online.

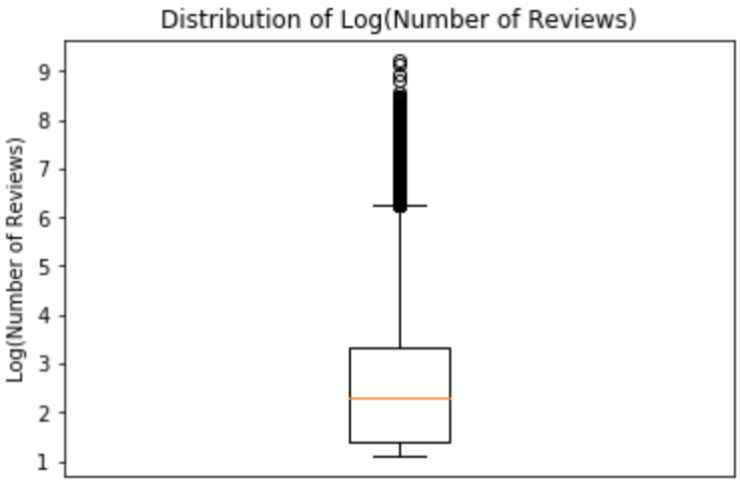
We begin by plotting individual variables to assess the quality of the datasets and the distributions of the variables.



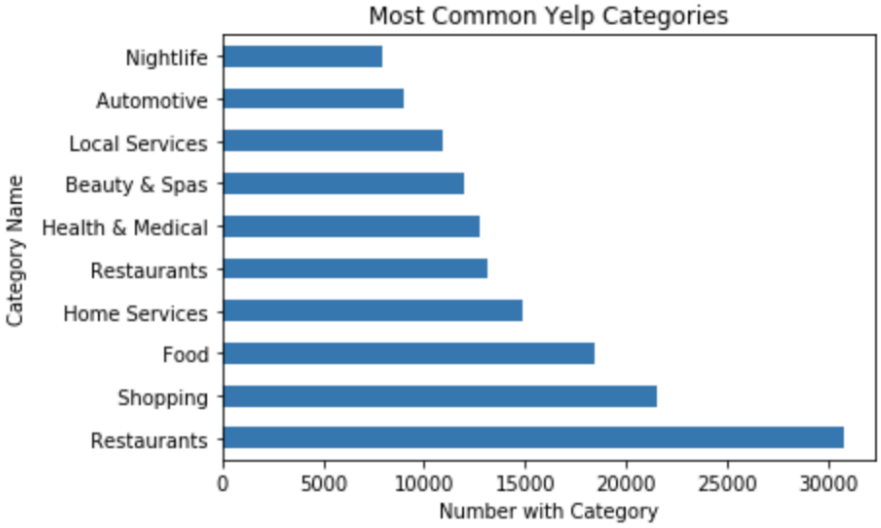
This chart shows the distribution of open and closed businesses in the Yelp business dataset. We found that of the 209,393 businesses, 40,090 businesses are now labeled as closed. For the purposes of this report, we choose to filter out these businesses, as we do not have information about when these businesses have closed, and so these data points may skew our results when conducting time series analysis on reviews.



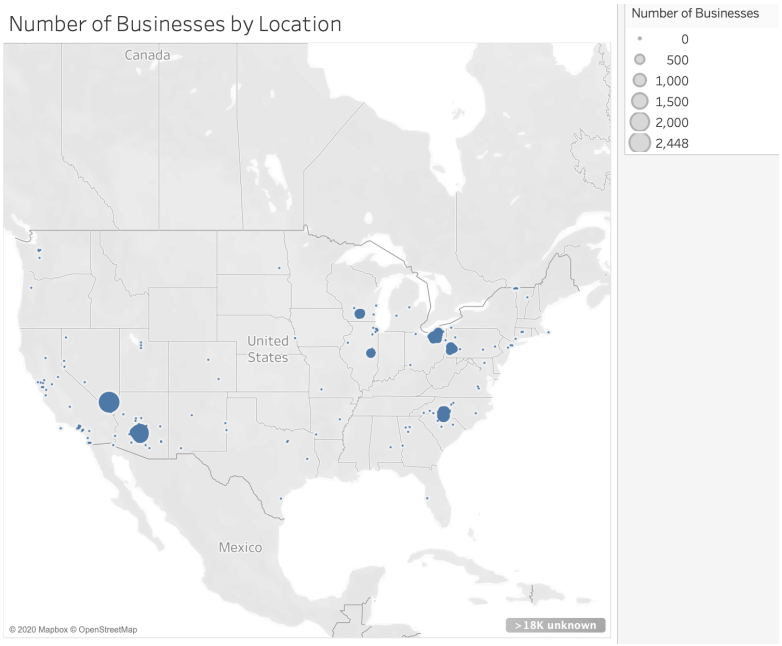
The chart above shows the distribution of average star ratings, rounded to the nearest half star, for the businesses that are labeled as open on Yelp. We see that the data is skewed to the left, showing that there are more businesses with higher average ratings than there are with lower average ratings. In addition, we see that the most common average rating for a business is 4 stars, followed by 3.5 stars.



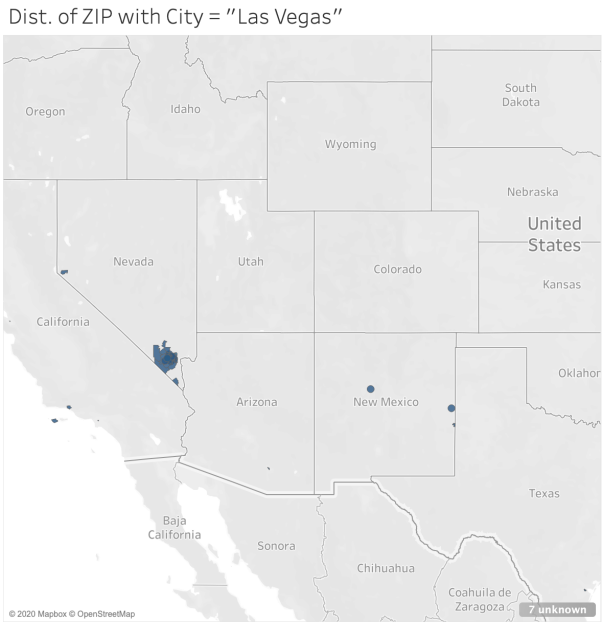
This is a boxplot showing the distribution of the natural logarithm of the number of reviews at different businesses. A logarithmic transformation was applied to this variable to make the data easier to see, as the data was extremely right-skewed. Even with the transformation, the distribution of the number of reviews is still right-skewed, with a median of 9 reviews, but a maximum of 10,129 reviews.



Next, we look at the top 10 most common categories among the businesses. There are various categories for the businesses, so the top 10 were chosen to see what the most common businesses in Yelp are. Note that a business can have multiple categories; for example, Yelp can label a business with the categories “Restaurants” and “Food.” When creating this plot, we noticed that there were some businesses that were not labeled with any categories. Since we were planning on analyzing the different categories in different cities, for the purposes of this report, these businesses were omitted.



Finally, we look at the geographical locations of the businesses in the dataset by creating a map with the number of businesses by ZIP code. As mentioned earlier, not all businesses were included in the dataset, as we would expect more populated places like San Francisco or New York City to have more businesses than what we see on the map. We see that there are over 18k unknown values, suggesting that the data is not clean.

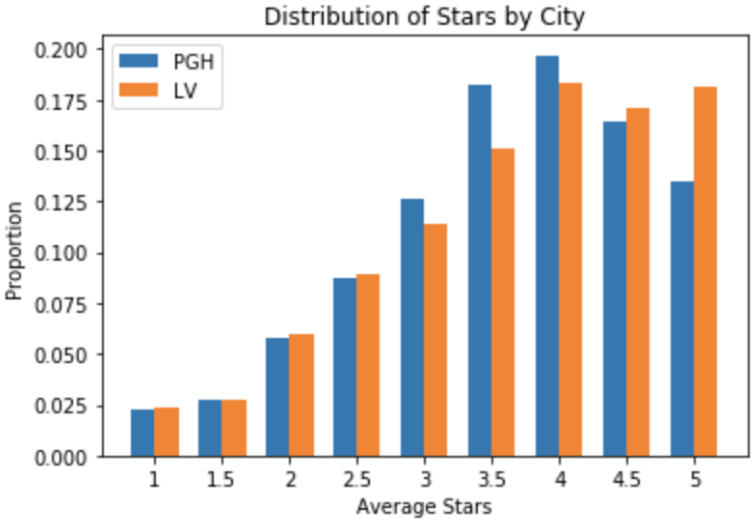


Looking further to see how messy our dataset is, we look at the number of businesses that have the city labeled as "Las Vegas." We see that there are other data points outside of the Las Vegas, NV area that we did not expect, such as the points in California and New Mexico. Therefore, when subsetting by

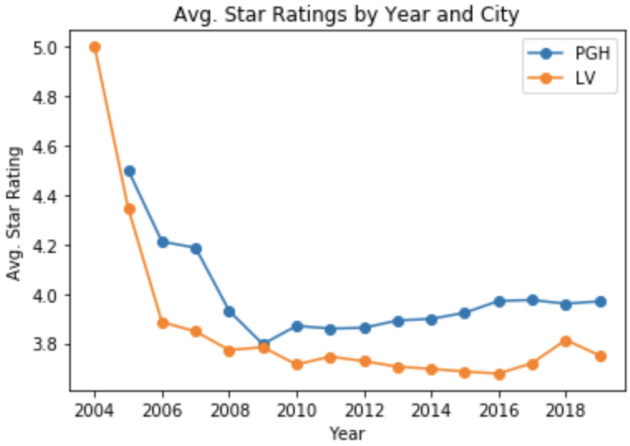
location, we will be filtering by both the city name and the state name to make sure that we are correctly choosing what data to look at.

Results and Insights

Part 1: What factors contribute to the number of stars received on Yelp?

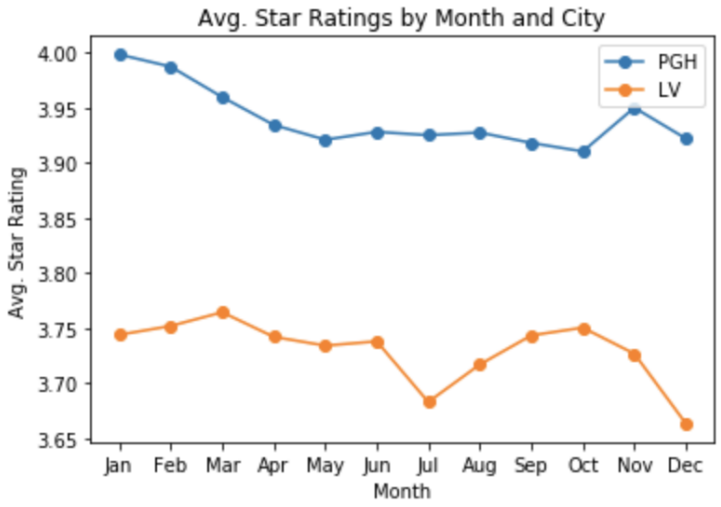


We filtered the data to consist of two cities: Las Vegas, NV, and Pittsburgh, PA, chosen because of the large number of businesses in both cities. The relative proportions of the stars were plotted because the two cities have different numbers of businesses. While Pittsburgh seems to have more reviews with lower average stars (such as 3 or 3.5 stars) than Las Vegas does, we see that there is not a significant difference in the distribution of stars between the two cities. Thus, the location of a business may not be a significant factor in the ratings.

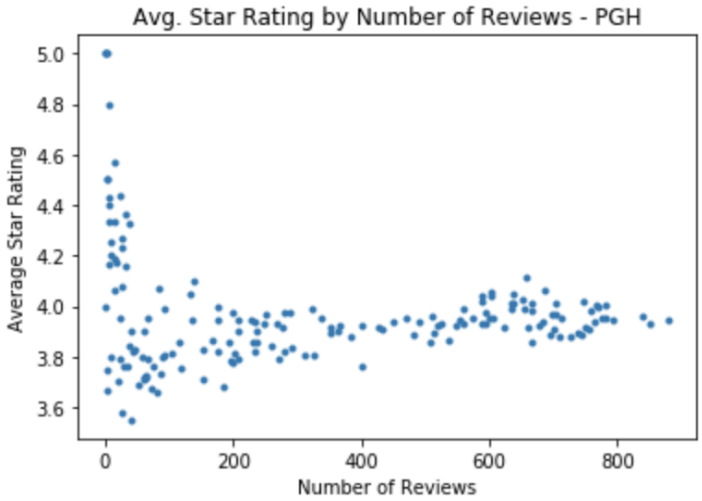


Next, we wanted to see if the year a review was made plays a factor in the average star rating. Using the same two cities, we see that the average star rating was a lot higher before 2009 than after 2009. This may be due to the shift in how customers use Yelp. Initially, customers may have only posted reviews when they had extremely positive experiences, leading to higher ratings; however, after seeing

negative reviews by others, customers may have been more comfortable with posting negative reviews themselves, which might drop the average star ratings.



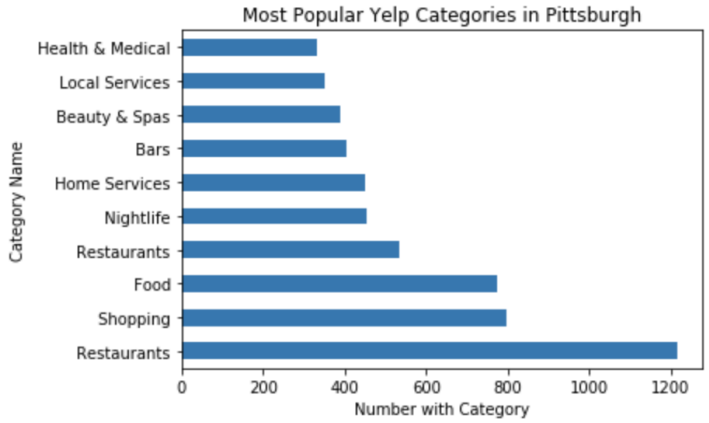
In addition to seeing how ratings changed by year, we wanted to see if the month a review was made affected the number of stars the review received. In both Pittsburgh and Las Vegas, we see a slight negative trend. However, the drop in average star ratings between the months is less than 0.5, suggesting that there is no association between the month of the review and its rating.



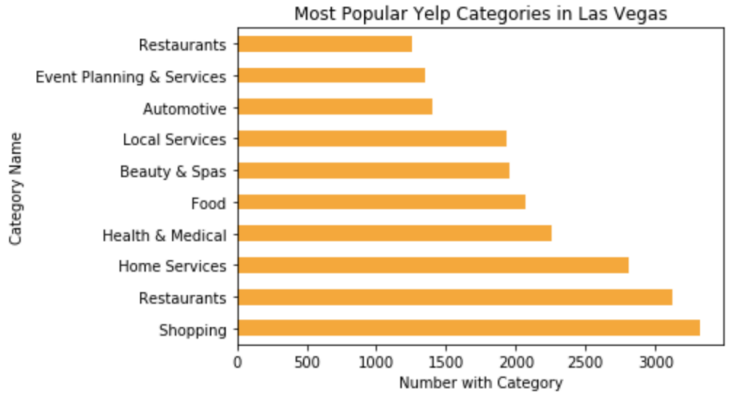
Next, we wanted to see whether there was a relationship between the number of reviews and the average star ratings. For this, we used monthly data from 100 Pittsburgh businesses that had the most overall reviews. We see that with a small number of reviews, the average star ratings are the most volatile and has the largest range of possible ratings. However, not including the high outliers, there is a small positive correlation between the number of reviews and average ratings, suggesting that with more reviews, the average rating for a business may increase.

Part 2: Are the categories of stores the same across different locations?

Similar to part 1, we chose to compare the cities of Las Vegas and Pittsburgh, as we believe that they should have different distributions of business categories based on our knowledge of the two cities. As a reminder, a business can have more than one category associated with it.

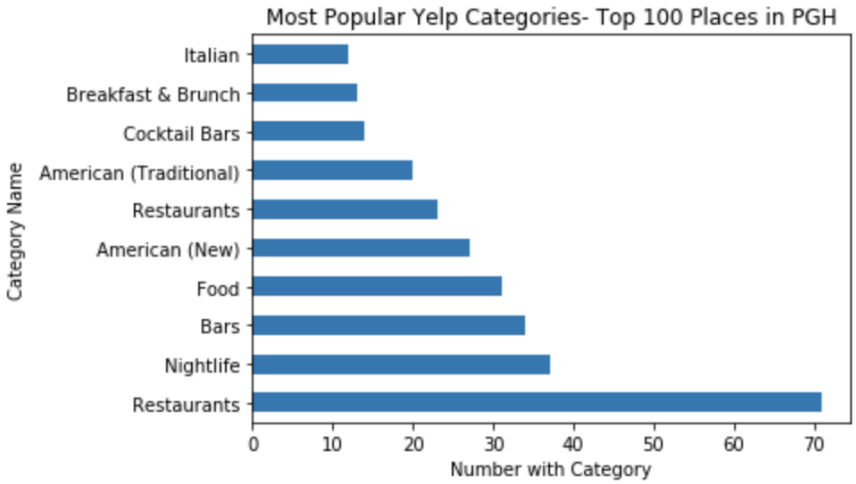


The two bar plots show the top 10 most popular categories in both Las Vegas and Pittsburgh businesses. Pittsburgh's most common category is "Restaurants" with a relatively large jump between the first and second most common category ("Shopping"), and most of the top 10 categories are related to food or drink. However, the category "Health & Medical" is also one of the top categories in Pittsburgh, probably because healthcare is one of the top five industries in Pittsburgh.

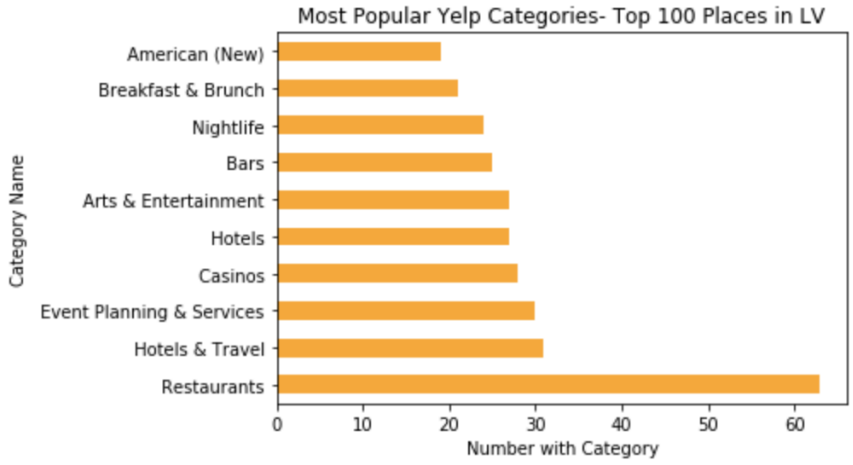


The top 10 most popular categories in Las Vegas are similar, but they are in a different order than Pittsburgh's. We see that there is more emphasis on businesses that are not related to food or drink than there was in Pittsburgh's bar graph. This makes sense, as one of Las Vegas' top industries is tourism. We also see that "Event Planning & Services," a category that was not seen in Pittsburgh's top 10, is in the top 10 categories for Las Vegas, also emphasizing the tourism industry in the city.

Looking at the above categories led to the following question: **Do the most popular businesses in the city have a similar distribution of categories?** To accomplish this, we assumed that there is an association between the number of reviews and popularity of business: the more reviews a business has, the more popular we assumed the business was. We, therefore, filtered the data to the 100 businesses in each city that had the most reviews, and looked at the categories of these businesses.

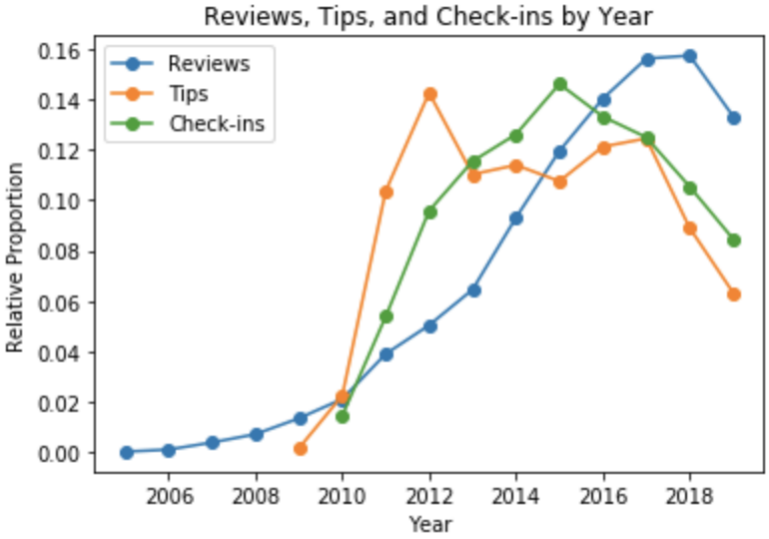


This bar chart shows the categories of the top 100 places in Pittsburgh. Like the top 10 categories for all of Pittsburgh, we notice that the categories are related to food and drink. However, unlike the previous distribution, all 10 of the most common categories are related to food and drink, and even describe different cuisines (such as “Italian”) or types of meals (“Breakfast and Brunch”). Therefore, we can conclude that while non-dining businesses are listed in Pittsburgh, they probably aren’t as reviewed as much as places related to food and drink.

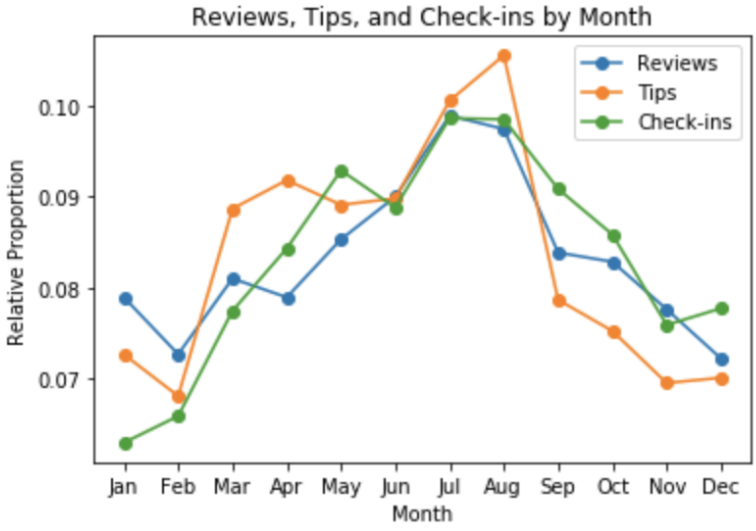


Similar to the bar chart for the top 100 reviewed places in Pittsburgh, this bar chart for the 100 most reviewed businesses does show some similarities to the bar chart for the top 10 categories in Las Vegas. However, this bar chart does emphasize the tourism industry more than the original bar chart did, as it has the categories “Hotel,” “Arts & Entertainment,” and “Casinos”. From this chart alone, one can infer how important the tourism and entertainment industries are for the economy of the city.

Part 3: How has the user's interaction of different Yelp features changed over time?



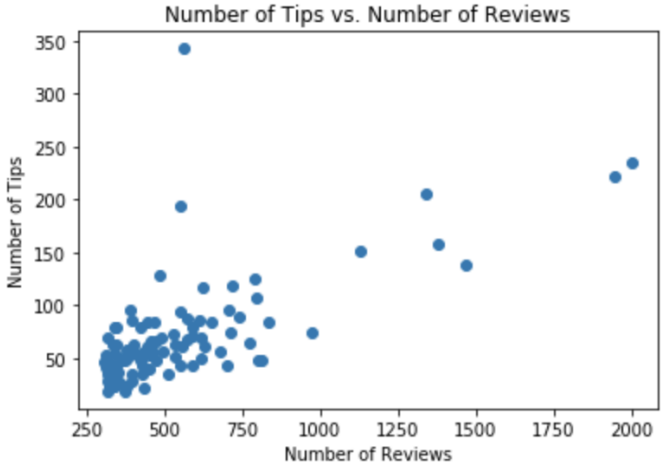
To see how users have interacted with different features in Yelp over time, we aggregated the number of reviews, tips, and check-ins by year, and plotted the time series. The above chart shows the relative proportion of these items compared to the total number of these items for each. To help with the runtime of the data, only reviews, tips, and check-ins from businesses in Pittsburgh were used. With the exception of 2019, as the years increase, the number of reviews conducted by users increased. The tips and check-in data had a more complex pattern: both seemed to have sharp increases in their usage between the years 2010-2012, but then the usage of these features gradually slowed down or dropped. This may be because customers realized that the tips feature was similar to the reviews feature, or because there was no reward for checking into a business an additional time. The lower proportion of tips, reviews, and check-ins in 2019 are perhaps because most Yelp pages were already established with lots of reviews and tips by then, so users felt less inclined to add additional information to the business' pages.



Next, we wanted to see if there was seasonality in the data. We aggregated the data by month (regardless of year) and compared the relative levels of reviews, tips, and check-ins. We found that all three line plots had a similar shape: on average, the level of interaction increased from January to August, and then decreased from August to December. We hypothesize that this is due to both the weather and the season. When the weather is warmer, outdoor seating is available, and many places have summer specials, which may contribute to a user's overall experience at the business. In addition, the summertime is when more people are free, as people in the education industry have their summer breaks, so they may spend more time visiting different businesses and restaurants.

After plotting the above two, an additional question was raised: **Is there a correlation between the number of reviews, tips, and check-ins?** The number of reviews, tips, and check-ins were aggregated over combinations of years and months (ex: number of reviews, tips, and check-ins in January 2015), and a correlation plot was created. We saw that all three variables were positively correlated against each other. Check-ins had relatively strong correlations with reviews and tips, signifying that the more check-ins there were, the more reviews and tips there were. Reviews and tips had a weaker correlation of 0.31, which may be due to the fact that tips were basically shorter versions of reviews, so users were more likely to only write either a tip or a review, but not both.

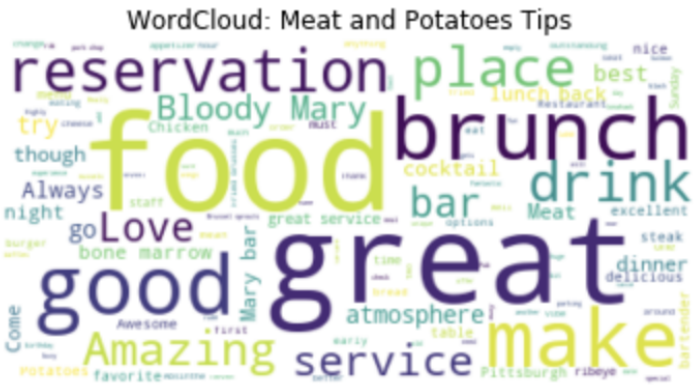
Part 4: What is the difference between reviews and tips?



This is a scatterplot of the number of reviews versus the number of tips for the top 100 most reviewed businesses in Pittsburgh. We can see that there is a positive correlation between the number of tips and the number of reviews: the more reviews there are, on average, the more tips that business would have on its Yelp page. This shows that even though tips and reviews may be similar in concept, users utilize both when writing suggestions for future customers to read.

We also wanted to use the text from the reviews and tips to see if we can see a difference in the content of these features. For this part of the report, we chose to use the Pittsburgh restaurant Meat and Potatoes' reviews and tips. We chose this restaurant because it is the Yelp business with the highest number of reviews in all of Pittsburgh, so we believed that it would be a good example to see the difference between reviews and tips.

This is a word cloud for the reviews written for Meat and Potatoes. There are many words in the word cloud, but we see that words with a positive connotation, such as "great" and "delicious" are large, suggesting these words appear commonly in their reviews. It is hard to see menu items within this word cloud. We can infer that reviews consist of what happened and how they felt about the dining or overall experience. For example, they could say that the food that they ordered was delicious.



In contrast, this is the word cloud for the tips written for Meat and Potatoes. We can see that there are still descriptive words such as “Love,” “good,” and “Amazing.” However, it is easier to see specific menu items in this word cloud, such as “Bloody Mary” and “bone marrow.” In addition, there are fewer words in this word cloud, suggesting that tips are shorter and are more straight to the point in suggesting future customers what to order or look out for.

Summary and Discussion

In summary, we found that what year the review is placed affects the star ratings. However, factors such as the number of reviews, the month the review was posted, and the location of the business does not play as big of a factor in ratings as expected. We found that the distribution of stores in Pittsburgh and Las Vegas is somewhat similar, but they also highlight the specialized industries of each city (such as entertainment for Las Vegas). In addition, the categories of the top 100 most reviewed businesses in either location have a different distribution than all of the businesses in that city. Yelp reviews, tips, and check-ins as a whole have increased throughout the years but unfortunately have dropped in 2019. They are also more frequent during the summer months than the rest of the year. Finally, there is a positive correlation between the number of tips and the number of reviews, and the tips contain more straightforward suggestions on what future customers should look out for.

In the future, additional analysis in other attributes about the business, such as whether the company provides parking, has takeout or delivery, and what the business’ hours are, should be conducted. These factors could play an important role in a customer’s experience and therefore, might influence a customer’s review. We would also extend our analysis to incorporate more cities, to see if there’s a pattern between the size of a metropolitan area and the categories of the businesses. Finally, we would like to incorporate the social aspect of Yelp, and explore variables such as the number of friends a user has and how “cool”, “funny”, or “useful” other Yelp users found the review to be.