Capstone 2 - Milestone Report 1

The Problem:

The Black Pearl, a new hotel and casino plans to open up a location right on the Las Vegas Strip. As they continue to plan out the design and amenities for their hotel, they want to take a closer look at what the public is saying about their soon to be rivals in the area. They want to understand what topics are correlated with good and bad reviews, and to be able to classify sentiment of their own reviews on any platform to track their performance. By analyzing the yelp reviews of the current resorts and casinos on the strip they'll be able to gain insight into their competition, and the star ratings will provide a good benchmark to analyze sentiment moving forward.

The Data:

The Yelp public dataset is available online through their dataset challenge webpage, and is also available for download on Kaggle. The complete dataset contains 5 different tables, but the two that are being used in the scope of this project are the business table and the reviews table. The business table contains company specific information for the different businesses in the Yelp public dataset such as name, location, company star rating, and review count. Each business is given a unique business ID that allows you to join information among the different tables in the dataset.

The other table being used for analysis is the reviews tabel. This table contains a unique review ID for each review, the business ID the review was regarding, and the text of the review. The review rating, as well as information on if that particular review was voted as cool, helpful, or funny is included as well.

Both of these tables were downloaded in a JSON format. The lines were read in one by one, and then converted to a JSON string. This JSON string was then used to create a dataframe using the .read_json() method available in Pandas.These steps were completed separately for both the business and review files. The next step was to filter out the business dataframe to only include businesses in the hotel category. This way when the business data is merged with the review data, only the reviews for hotels would be included. This made the review file, originally containing 4.39 GB of data, more manageable for analysis.

The filtered hotel data and the reviews data were then merged together on the unique business_id variable using an inner join. Because the The Black Pearl is specifically interested in resort casinos on the Las Vegas Strip, another step of filtering was done to make sure we were only working with joint hotel/casinos and only those located on the Las Vegas Strip.

As both the business and review dataframe had a star variable, those column names had to be changed to review rating and company rating to differentiate the two metrics, and avoid confusion moving forward. Afterwards, the unneeded columns were dropped to make sure we were only focusing our analysis on useful information. The attributes column, containing valuable information for restaurant type businesses, was dropped along with state, city, and postal code information that was the same for every business in this view. Latitude and longitude were also dropped as the focus of this project is in one specific area. Null values were then filled in for the address field with the string 'Not Available' to make it clear this information was not provided. Lastly, the index was set to review_id, and a column called text length was added so the length of each review could be compared between different review ratings as well.

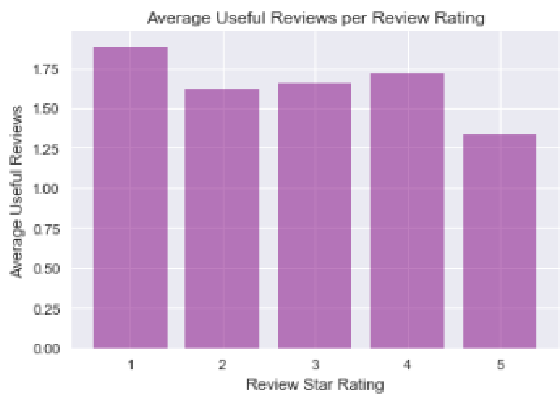
The final step before diving into the exploratory analysis was to check for any outliers in the data. I started by looking for any companies with 0 reviews, and although I didn't find any, I found a big range in the number of reviews per company with 3 as a minimum and 4,041 as the max. Although these are outliers, they do not indicate any errors, and were not moved from the dataset. Next I checked to make sure the review ratings and company ratings all fell within the 1-5 star possibilities. All of the ratings fell within the expected threshold, and I moved on to checking the useful, funny, or cool designations for the yelp reviews. Each of these 3 variables should only have be marked with a 1 or a 0 for each review. A 1 in the useful column means that the review was considered useful, for example, and 0 means it was not considered useful. Each of the 3 variables, useful, funny, and cool only contained either 1's or 0's as expected. Lastly, I took a look at the text length variable to check if there were any one or two word reviews. The minimum text length for any review was 15 which is still long enough to convey sentiment. At this point the data was considered ready to explore further.

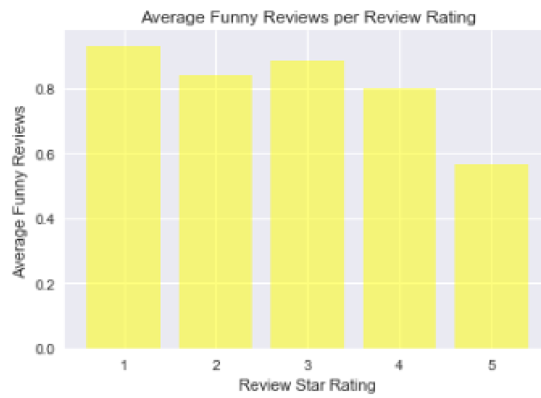
Analysis and Testing:

In order to get a better understanding of what contributes to high and low rated reviews, we'll first have to take a look at the makeup of our of our sample.The analysis started off by breaking down the total number of reviews by review rating, and I found that 1 and 4 star reviews are the most common in this dataset with 15,639 and 15,556 reviews respectively. 2 star reviews had the smallest number at 9,139.

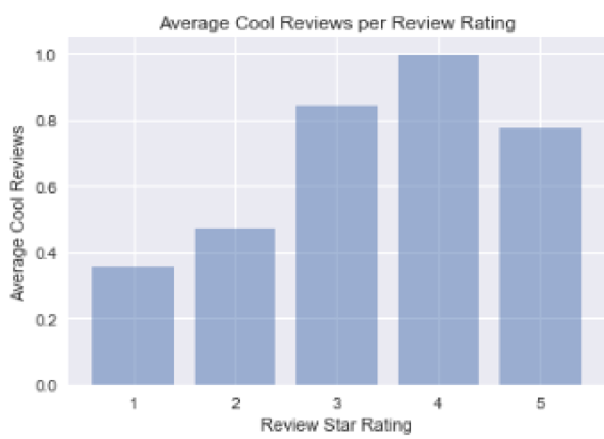
When I looked at the number of useful, funny, and cool labeled reviews by review rating I found that 1 and 4 star reviews contained the most amount of these designations. As the 1 and 4 star reviews had a bigger pool to draw from, I compared the averages of the useful, funny, and cool reviews across the star ratings as well.

When looking at the average usefulness across the different star ratings, 1 star reviews ranked the highest. Overall the lower rated reviews contained higher useful scores than higher rated reviews did. This was the case with the average funny reviews as well. 1 star reviews were again ranked the highest, and like the useful designation, funny reviews tended to have higher average scores in the low review ratings compared to the high ones.



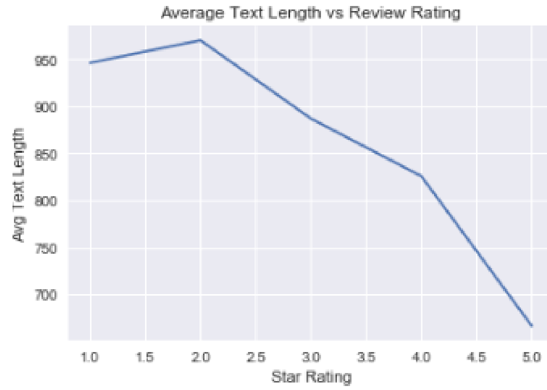


Cool reviews, however, had the highest average score in 4 star reviews. Unlike reviews marked as useful or funny, the cool reviews were found in greater numbers on the higher rated reviews.

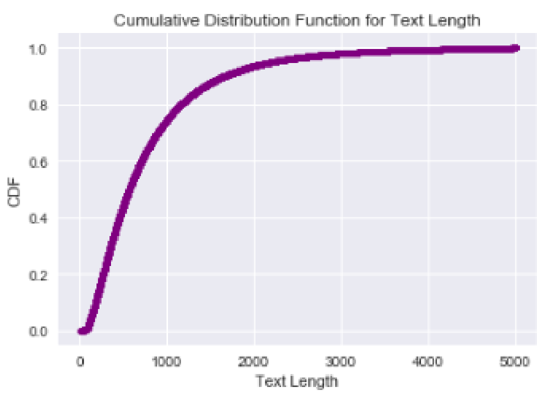


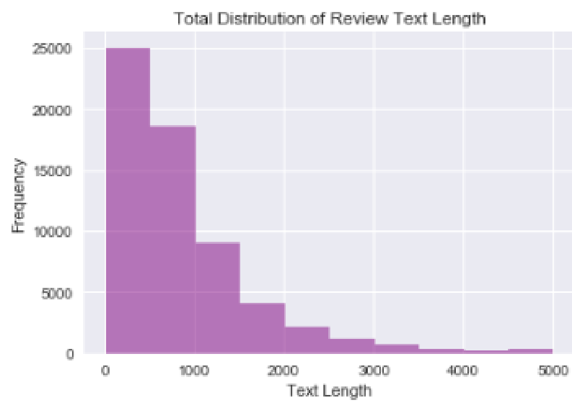
As the data seems to represent lower reviews being more frequently marked as useful or funny, and higher reviews more frequently marked as cool, t-tests were run to confirm whether or not their was a significant difference in these designations between high (above 3 stars) and low (below 3 star) review ratings. The t-tests were run for each of the 3 designations, and the results all pointed to a significant difference between the high and low rated review sets. After generating bootstrap samples for the mean, the probability of useful and funny designations being higher in lower rated reviews and cool designations being higher in higher rated results was tested, and the results confirmed our initial analysis.

The next variable of interest was the text length of each review. When we look at average text length by review rating we find an overall trend of shorter text lengths being associated with higher review ratings. It seems that when something has gone wrong for a customer, it requires more explanation than a positive experience does. There are a number of outliers for text length for each of the review ratings though , so a hypothesis test was conducted to determine if there truly is a significant difference in text length for high and low rated reviews.



The results of the t-test assuming the average text length was equal in high and low review ratings showed a p-value far below the .05 significance level. We can confidently say there is a significant difference in the text length between review ratings. The text lengths range from a minimum value of 15 to a maximum value of 5,000, but the average was found to be a length of 864.





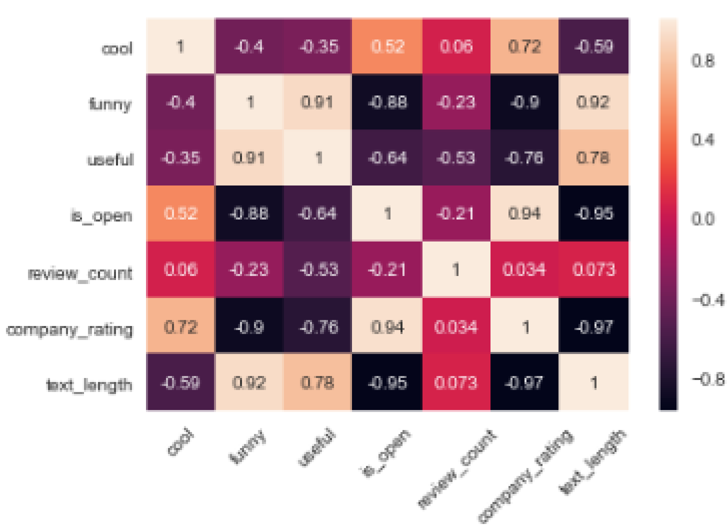
A cumulative distribution function was generated for the text length variable and we see that 70% of the reviews fall below a length of 1000, and 40% of the reviews fall below 500. The distributions of review lengths holds the same shape with low, high, and the middle 3 star ratings.

Looking at the average text length by company rating it became apparent that the higher text lengths were more frequently associated with lower rated companies. This builds off of our findings that text length is longer for lower star reviews. In regard to the number of reviews by company rating, you'll see that companies with ratings between 2.5 and 3.5 have the highest number of ratings. Companies rated on either extreme of the spectrum tend to have a fewer number of total reviews. Generally speaking, this pattern makes sense as more reviews give the opportunity for both positive and negative experiences to be represented, and can cause the average to fall towards the middle of the rankings. To check if there was an important correlation between the number of reviews and the company rating however, a t-test was performed to see if there was a significant difference between the number of reviews in high and low rated companies.

The resulting p-value computed from the t-test was above the .05 significance level at .25. These results did not allow us to reject the null hypothesis, and did not allow us to find that the number of reviews had a significant impact on the overall company rating.

I found that the review counts ranged from a minimum value of 3 to a maximum value of 4,041 reviews across all companies in this dataset. The average company review count is 760, and we find the majority of review counts falling below 500. Although we previously looked at the useful, funny, and cool designations as they related to review ratings, it made sense to see if the patterns we found held true for company ratings as well. T-tests were again run for the three variables to determine if there was a significant difference between high (above a 3 star rating) and low (below a 3 star rating) companies. The results of our testing found that there actually isn't a significant difference in average amount of useful or funny ratings between high and low rated companies. For cool reviews however, the results proved that the average cool rating was more likely to be associated with high rated companies.

The last step before examining the text itself was to look closer at correlations between all the numeric variables.



The data was grouped by review rating, and a heat map was generated to easily see how the variables related to one another. A number 1 represents a 100% positive correlation and - 1 would represent a 100% negative correlation. From this heatmap we can see that text length is positively correlated with funny and useful reviews, and is very negatively correlated with company ratings.

We can also see that the company rating is positively correlated with cool reviews, but negatively correlated with useful ones.The review count showing almost no correlation with company ratings, backs up our findings regarding this question posed earlier. One other thing to note is that funny reviews are often found useful, but they have a strong negative correlation with company ratings. Customers seem to appreciate honest reviews about negative experiences so they can avoid making the same mistakes themselves. Positive experiences don't relate to the useful designation in the same way.

After looking through the other variables in the dataset, the text of the reviews themselves were analyzed further. The text was split up into two groups, one star and five star reviews. These were chosen in hopes of look at the most polarizing language for customer experiences.

Each of the two groups were tokenized separately, converted to lowercase, and filtered to only contain alpha characters. I then used the english stopwords from the nltk.corpus package to make sure redundant and unimportant words were removed. Words like hotel, stay, stayed, and 'u' were added to this set of stopwords after an initial run through. I found that they were not helpful in finding common themes, and were frequently appearing in both 1 and 5 star reviews. Once the stopwords were removed, each group was lemmatized and converted to a bag of words.

Looking at the most common words in 5 star reviews I found that words like; nice, pool, good, service, clean, really, love, staff, also, and would were most prominent. In the 1 star reviews words like even, never, back, could , told, service, front, desk, said, and check were appearing most often. There were some common themes in both the 1 star and 5 star review texts. Room was mentioned most frequently in both groups of reviews, and time, like, and service were both found in the top words as well.

Word clouds were generated for each group to better visualize the words associated with both the positive and negative reviews. You'll find the word cloud for the 1 star reviews directly to the left, and the word cloud representing the the 5 star reviews below. Looking at these world clouds we see some additional words of importance stand out for positive and negative reviews. Words like called, nothing, and fee popping up in in the 1 star word cloud are easy to associate with negative hotel experiences. On the positive side, words like restaurant, buffet, food, and

drink appearing in the 5 star word cloud bring our attention to other important factors that go into a positive experience.

In addition to looking at the 1 and 5 star review text, I also wanted to gain some insight on the reviews for top rated companies currently on the Las Vegas Strip. While looking closer at who the top competition will be once The Black Pearl opens its doors, all companies having less than 1000 reviews were taken out of the equation. This left 21 companies, and only 4 of them had a company rating above 3.0. These four were M Resort Spa and Casino, ARIA Resort and Casino, South Point Hotel, Casino & Spa, and Red Rock Casino Resort & Spa. The M Resort Spa and Casino had a company rating of 4.0 and the remaining three had company ratings of 3.5.

Looking at all of the reviews for these companies with over a 3 star rating, I was able to gain some insight into what the positive reviews for these companies consisted of. The common words in the general 5 star review group were found in these reviews as well. Words like great, nice, like, pool, good, service, would, really, also, and clean were most common in the top companies positive reviews. Another few words that appeared more frequently specifically for these companies however, were restaurant, buffet, and food, letting us know what these companies are doing really well.

A word cloud was generated for the top companies positive reviews as well. Looking closer at the world cloud below, bed, spa, and bathroom were also mentioned frequently in these company's reviews.

Summary of Findings:

- 1 and 4 star reviews are most common
- Reviews marked as useful or funny are more frequently associated with low review ratings while reviews marked as cool are more frequently associated with high ratings
- Longer review lengths are associated with low rated reviews, and this feeds into longer review lengths being associated with lower rated companies as well
- The number of reviews alone does not have a significant impact on the overall company rating

- There isn’t a significant difference in the average useful or funny reviews between high and low company ratings, but reviews marked as cool are more likely to be associated with high rated companies
- Text length is positively correlated with funny and useful reviews, and is very negatively correlated with company ratings.
- Funny reviews are positively correlated with useful ones
- Common themes in five star reviews: room, time, service, nice, good, love, everything, staff, clean, pool, restaurant, buffet, food, and drink
- Common themes in one star reviews: room, time, even, never, back, could , told, service, front, desk, said, checked, called, nothing, and fee
- Common themes in top companies positive reviews closely resembled the total 5 star reviews, but had restaurant, buffet, food appear more often and they also included bed, spa, and bathroom frequently