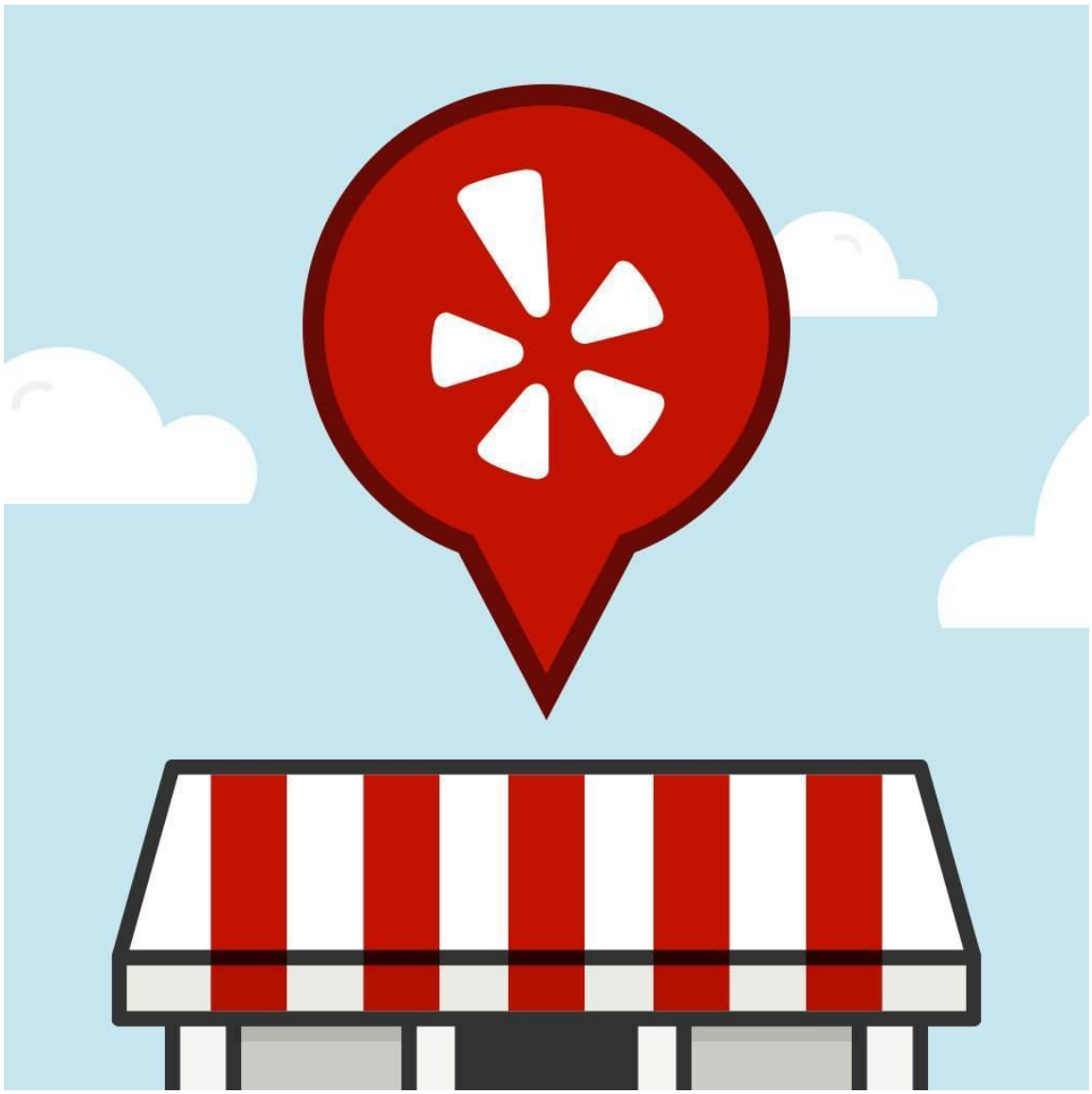

Yelp Analysis

A Practical Analysis of Ratings

INTRODUCTION

The Yelp dataset contains various dimensions of data that includes information on businesses, users, and more. Looking at this dataset, our group was inundated by many questions that raised our interest because there are many potential ways to extract meaningful information from this data and implement the results in improving business strategies.

Yelp is a database which is very popular for storing information about service businesses, especially restaurants. Aside from the name, address, hours of operation or photos provided by the businesses, Yelp gathers plenty of information from the users of their services. In fact, Yelp can be described as a platform in which users share their experience from different businesses and use other people's information when considering using a new business's service. Users can even rate each other on usefulness of the comments they share about a business. For such a database, which is heavily dependent on the data that users provide, it is very important to figure out what portion of this data actually represents practical value. One way of gauging the consistency of the users' comments with practical value is the number of users who find a comment useful.

Another important aspect in Yelp dataset is users' ratings. Prospective customers potentially judge businesses by these ratings. Therefore, it is critical to determine what portion of these ratings are ranked as useful by other users. To make our analysis more practical we decided to determine if there is any relation between the ratings and the usefulness of the data.

Furthermore, we are going to analyze what impact these ratings will have to the overall popularity of restaurants. Based on the information that we gathered in the data, we conducted an analysis to determine how we can improve the performance of these restaurants.

STORYLINE

Since Yelp is a very popular source of information and it is heavily dependent on user-provided data, it is critical to obtain an understanding of users' behavior. For instance, how likely is it that a 5-star review be useful to users? Or is there any relationship between the star-ratings and the level of usefulness? For our analysis, we define usefulness as the ability of the rating to provide practical value. To obtain a visualization on the potential relationship between star-ratings and usefulness of the data, we used a pie chart. Since the number of incidents in different ratings might vary and we did not want to let the difference between our distinct sample sizes (count of different star-rating) affect our results, we found the pie chart as an efficient tool to illustrate which portion of the useful ratings belong to each star rating. Furthermore, we needed to see the proportions of the total reviews for each star-rating to obtain a logical intuition regarding the usefulness of each star-rating. Comparing the two pie charts we came to the realization that the 1-star ratings are the most likely to be useful and the 5-star ratings are the least likely to be useful.

Getting a grasp of a relationship between the star ratings and the applicable use of the ratings, we would like to see what is the impact of a rating on the overall popularity of a restaurant. Considering the dataset limitations, analyzing the number of restaurant check-ins can be useful in estimating the overall popularity of a restaurant compared with other restaurants in the dataset. Based on the data, we found that users are most likely to check-in to 4-star restaurants, which are the most popular, and are least likely to check-in to 1-star restaurants, which are the least popular, as expected.

Since we have obtained how user ratings can impact other users behaviors and consequently the popularity of the restaurants, we would like to see what would users with the

highest number of followers, known as social media influencers, consider as problems or weak points when evaluating the quality of a restaurant. Obtaining this information will be very useful in determining what factors about a restaurant's weaknesses are most likely to be communicated by the most popular users on Yelp.

Based on the information we gathered in the first stage of our analysis, we determined that 1-star ratings are the most practical ratings. As a result, we based our analysis in this stage on the 1-star rated restaurant. Excluding stop words, we searched for the most frequent words that were brought up by the users with most followers regarding 1-star rated restaurants. An interesting finding is that service was found to be the most frequent word and food was mentioned less frequently than service. Also, rude and staff were among the most frequent words. Based on this data, it can be concluded that if a restaurant is providing poor service it is more likely to be communicated by the most popular Yelp users than if a restaurant provides poor quality food. In addition, the attitude of the staff can be a very influential trigger for users to write a bad review for a restaurant.

In conclusion, our understanding of the impact of restaurants' star-ratings on the overall popularity of a restaurant and verifying that low star ratings is most practical information,, we believe that this analysis can be a good guide to restaurants with poor star-ratings. Since bad service and attitude within the staff are potentially the biggest triggers for influential Yelp users, focusing on improving service quality and staff attitudes will be a very effective way for restaurants to improve their star-rating and consequently reputation in an affordable manner. If restaurant owners implement this strategy, it will help them move from a 1-star rated restaurant to the coveted 4-star ratings.

CHALLENGES AND SUCCESSES

The initial downloading of the dataset caused issues due to its significant size. We originally had a few group members attempt the JSON download and a few group members attempt the SQL downloads. JSON files gave us a basic view of our data and helped us on preparing questions. We excluded some columns because of the quality of data.

After inserting data into MySQL database, we originally tried to find the relationship between Stars and Useful reviews, however, we got extremely small slopes when we attempted to do logistic regression on Rating and Useful. Also, we found that the Rating does not provide a true normal distribution because people tend to rate higher than 3. So, the rating distribution shows a left skewness. Then, we realized that it was a good idea to compare Usefulness in proportion. We figured out that pie chart would be a representative way to show the results and found that 1-star review has a small proportion in the total review, but it covers a bigger area in the useful review.

Nevertheless, it took us several hours to debate on the questions and the potential routes to meaningful solutions. It appears that all the columns have some type of relationship with stars. Therefore, we chose Stars as our main topic and came up with other questions based on it. Some of us thought that we should focus on user side. For example, what is good information for a user and how can Yelp improve their user experience. Others thought that it would be better to understand the business side to find out the best way to enhance the businesses to attract more customers. Ultimately, we decided to focus on the business side.

We considered the total number of check-ins as a measure of popularity since there is no data about the number of customers who walked into each restaurant and tried to find some useful information on the relationship between Popularity and Stars. Nevertheless, we couldn't use aggregate function twice while we were working within a single SQL query. So, we created a new data frame to store columns and used the function again in that temporary table to finally get a bar plot.

Next, we thought that the restaurants with more photos might have a better rating. Unfortunately for us, those photos are uploaded by consumers and not the restaurant themselves

so we didn't follow through on this idea. However, we still considered photos posted as a good measure of popularity because even within one check-in, a consumer can post multiple pictures, which typically represents a good experience at the restaurant.

Then, we were faced with a big challenge of how we wanted to analyze the text data within the Tips table. We analyzed the restaurants with 1-star because we want to know that what words are common in the consumers' suggestions for lowly-rated restaurants. At first, we tried to eliminate some stopwords but we couldn't get rid of the line breaks. After attempting several command lines and python filters that didn't work we finally discovered a new package called nltk (Natural Language Toolkit) to get the most common words for tips.

In conclusion, we obtained meaningful information from the analysis by finding that the Stars didn't really match the Usefulness for highly-rated restaurants. However, other Yelp users can relate to the frustration and anger of a 1-star rating so those reviews are very useful. In addition, the 5-stars group doesn't have the highest volume of customers. This can be attributed to the price of foods, the geographical location or some other reasons and business owners can focus on improving their restaurant to the more coveted 4-star region. Moreover, words like service comes before food in reviews for 1-star rated restaurants, which indicates that these 1-star restaurants should improve their services in order to attract more customers and improve their Yelp rating.

Ultimately, the yelp dataset is large but also provides a few limitations.. For example, if we can have more data on the price range of each restaurant and the volume of customer each day, we would be able to add to our analysis by providing further analysis on prices and revenue. For further analysis, we may also want to understand the 4-star restaurants better to understand why these restaurants are so popular, even when compared to 5-star restaurants.

TASK ASSIGNMENT

After Xin successfully imported the dataset into his notebook, we mainly programmed on his computer. We worked together during the meeting to brainstorm ideas and discuss the storylines together. Xin had the main responsibility of writing and modifying code and other teammates

help him when he got stuck. Parisa was mainly responsible for the introduction and telling a story based on our notebook. Hongbo documented all challenging times and better moments. Xinran wrote down the workflow part. Justin recorded the Youtube video for the project and proofread the final report.

WORKFLOW

Part 1 - Loading data into mysql database

- Imported python modules: pymysql, matplotlib.pyplot, pandas, pandas.io.sql.
- Connected to the mysql local host and create a database named "yelp_db".
- Showed all the databases in the MySQL local-host and used the newly created database "yelp_db".

Part 2 - Data presentation

- Selected all the first three rows in the *review*, *business*, *checkin*, *photo*, *tip*, *user* and *attribute* datasets and print the results.

Part 3 - Usefulness for average rating

- We grouped the total number of *useful reviews* by different *stars* in the *review* dataset and plot a bar chart to show the quantity by *star* and a pie chart to show the proportion by *star*.
- We grouped the total number of *reviews* by different *stars* in the *review* dataset and plot a bar chart to show the quantity by *star* and a pie chart to show the proportion by *star*.

Part 4 - Check-ins for average rating

- We created a temporary table named *temp* which contains the total number of check-in *counts* grouped by different *business* and the *business*' corresponding *star*.
- We grouped the total number of check-ins of all the business by different *star* in the *temp* table and plot a bar chart to show the quantity by *star* and a pie chart to show the proportion by *star*.

Part 5 - Check-ins for average rating

- We selected the *name* of business, *caption* of the photo, *label* of the photo from the joined table of *business* and *photo* where the *label* is “food” and print the first five rows.
- We grouped the total number of *label* by the *name* of business and show the first ten rows.
- We created another temporary table named *temp1* which contains *business ID* and total number of *photos* with label “food” grouped by different *business* in the descending order.
- We printed the top three *business ID* with the most number of *photos.*
- We grouped the total number of *photos* with label “food” of all the business by the different *star* of the business from the joined table of *temp1, business* and *photo* and plot a bar chart to show the quantity by *star* and a pie chart to show the proportion by *star*.

Part 6 - Text analysis of *tips* for one-star rating restaurant

- We selected the *name* of business, *business ID, name* of business with one-star rating and *text* of the tip for the business from the joined table of *business* and *tip* and show the top ten rows.
- We extracted each *text* from the *tip* written for one-star rating restaurant and show the first ten texts.
- We saved the content of the *text* to "tip_for_1star.txt" and reload the file in Jupyter notebook.
- We split the text word for word and ignore the cases and print out the frequency of the words in descending order.
- We imported python packages: nltk, io, stopwords, word_tokenize
- We got a list of stop words from the stopwords module.
- We selected the words in "tip_for_1star.txt" that are not in stop words and save them in “filteredtext.txt”.
- We sorted the top 30 common words in “filteredtext.txt”.

Part 7 - Text analysis of *tips* for five-star rating restaurant

- We selected the *attribute name* and *value* of business from the joined table of *attribute* and *business* where restaurants got a five-star rating and show the top ten rows.

- We saved the content of the *attribute name* and *value* of business to " attributes.txt " and reload the file in Jupyter notebook.
- We selected the words in " attributes.txt " that are not in stop words and save them in “filteredattr.txt”.
- We sorted the top 30 common words in “filteredattr.txt”.