Yelp Reviews

A Comprehensive Analysis of Review Trends

Dominic Deckys
University of Colorado
Boulder, Colorado
dominic.deckys@colorado.edu

Michael Gigiolio
University of Colorado
Boulder, Colorado
michael.gigiolio@colorado.edu

Dante Pasionek
University of Colorado
Boulder, Colorado
dante.pasionek@colorado.edu

1 PROBLEM STATEMENT & MOTIVATION

This paper and the research conducted seek to investigate factors that contribute to the performance and overall perception of a business. Through analysis of reviews on the popular restaurant critique website Yelp®, and the application of advanced data mining techniques, this work intends to reveal trends that define a restaurant in the public eye. This research draws motivation from businesses serving the public and their need to appeal to their customer base as effectively as possible. By determining what features are correlated with an establishment being viewed positively, this paper aims to provide a useful service to companies seeking to improve their operation.

2 LITERATURE SURVEY

While there currently has not been any work conducted that is similar to the research detailed here, many other studies have been performed using Yelp data. Notably, the company Yelp itself regularly hosts what is known as *Yelp Dataset Challenge* [1]. In this challenge, students are provided data from Yelp in a competition to help practice their data mining techniques. Other research includes:

(1) *What Yelp Fake Review Filter Might Be Doing?*[2]
An analysis on how Yelp determines fake reviews placed to artificially alter the standing of a business. An investigative report on the algorithmic analysis and determination of what qualities Yelp uses to determine the potential fraudulent nature of the review.

(2) *Integrating web-based data mining tools with business models for knowledge management*[3]
An analysis of how conducting data mining of a business' internal resources and information can allow companies to review and enhance the current business model and performance.

(3) *Yelp's Top 100 Places to Eat in Canada for 2018*[4]
Yelp conducted their own research to create a list of the 100 best rated restaurants in Canada based on both rating and volume. This study focused specifically on Canadian residents and was conducted only with reviews made by local people to better tailor the results to permanent Canadian residents.

(4) *What Was The Hottest New Restaurant in Your State in 2017?*[5]
Much like the study outlined in the previous article, Yelp's data team performed data analytics upon their American reviews in order to provide a list of the most popular restaurants started in 2017 by state.

(5) *Local Economic Outlook*[6]
Yelp conducted analysis of it's review data and produced a list of 50 cities, 50 neighborhoods, and 10 business categories that were found to be the most ideal places and business models for the success of small businesses. In addition, they determined the economic opportunity in these areas so as to aid current or prospective restaurant startup owners in making large scale choices that will most likely lead to their success. This analysis was performed using repetitive determinations of which small businesses were most likely to remain open in the subsequent three months over the course of two years.

This previous research will ideally aid in the analysis of the effects of location, hours of business and other factors on the overall impression of an establishment. Additionally, these studies will provide insight concerning specific methods of analysis on Yelp data in particular.

3 DATASET

The dataset used in this analysis was obtained through www.kaggle.com and is aptly named *Yelp Dataset*[7]. This dataset contains approximately 5,200,000 user reviews, with information on more than 174,000 businesses over a total of 11 metropolitan areas. This research will focus primarily on attributes such as `stars` for review ratings as well as `latitude`, `longitude` and `postal_code` to determine location of a business. Besides the attributes mentioned above, it's important to know what other attributes this research will be using.

In addition to the other tables, the data set contains a `business_hours` table, consisting of the hours of operation stored as a string of 24-hour time. Each day of the week is represented as an attribute, with each unique business id as a primary key. Establishments that are do not operate on certain days or are no longer in business have their hours represented as `None`.

The largest table in this dataset is `yelp_reviews` which contains information about the individual reviews. Specifically, the table

[1] https://www.yelp.com/dataset/challenge
[2] https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/viewFile/6006/6380
[3] https://www.sciencedirect.com/science/article/pii/S0167923602000982
[4] https://www.yelpblog.com/2018/02/109791
[5] https://www.yelpblog.com/2017/12/best-new-restaurant-every-state-2017
[6] https://www.yelpblog.com/2017/10/local-economic-outlook
[7] https://www.kaggle.com/yelp-dataset/yelp-dataset

contains the business reviewed, the number of stars given, the date the business was reviewed and comments made. This research will necessarily combine numerous attributes drawn from many of the tables provided in order to best identify trends.

4 PROPOSED WORK

Initially, the dataset will need to be cleaned and preprocessed. In order to accomplish this, numerous modifications are required to be made so that the dataset can be most effectively and efficiently worked with. These methods include:

- Several of the attributes, such as the name and address of the businesses, contain information represented as a string surrounded by quotation marks. In order to properly process and represent this data, the quotation marks will need to be removed so that only the relevant data is left.

- The `postal code` data column will need to be removed. As the data set includes reviews from several different countries, the postal codes often take different formats. For example, a postal code such as `L3P 1X3` represents a location in Canada, whereas `85286` is a postal code in the United States. This will require any algorithm processing the postal code to be able to distinguish between multiple different styles to properly determine location, and as the dataset already contains the latitude and longitude coordinates of the business, there is no need to expend so much effort to obtain less exact information.

- The `neighborhood` attribute will likely need to be removed. This results from the fact that this column has a large enough amount of missing values that render it functionally useless for processing. In addition, this information, even if it were to be completely present, does not provide the same utility as the other location services due to its inherent ambiguity.

- It would be useful to add a column to the `business_hours` table to represent the total number of hours open per week as well as average per day. This information, now easily accessible, would facilitate the comparison of other attributes to open hours.

- Creating a normalized score of a business by weighing the average star rating, the number of reviews left and number of check-ins. This is beneficial as a business with a high rating and very few or a single review may not have qualities allowing it to compare to other highly rated businesses. Conversely, a business with a low score and only a few reviews may be improperly represented as worse than its actual merit.

5 EVALUATION METHODS

This research is conducted primarily using Python and its peripherals. This includes, though is not currently limited to, Pandas, NumPy, and SciPy. Pandas will allow the CSV files to be converted into a dataframe, permitting easy operation of NumPy and SciPy upon the dataset. By using mathematical tools such as Ordinary Least-Squares Regression (OLS), confidence intervals and p-values, this analysis intends to find meaningful patterns regarding the interaction between business hours, ratings, and location. Specifically, the use of confidence intervals will aid in the determination of possible correlations between the aforementioned attributes. In turn, multiple regression techniques will allow for more accurate predictions of a business' performance in regards to certain criteria.

Methods similar to those executed in the prior research described in the literature survey will be implemented. For example, weight of the overall rating of a restaurant will be determined not only by the number of average stars, but also by the number of reviews, so as to not skew the data due to a restaurant having a insufficient sample of reviews. This results from the proposition that the comparison of average ratings for specific types of businesses, compounded with the various attributes that contribute to a higher or lower overall rating, will assist establishments in improving their consumer appeal.

6 TOOLS

A variety of tools will be used in this analysis, the programming tools to be used are detailed below.

- Python 3: A high level programming language

- Pandas: An open source Python library used for data manipulation and analysis. This library serves to format the data so that the other libraries used can operate upon it.

- NumPy: An open source Python library used for numerical computation on datasets and other forms of information. This will allow the performance of operations on the millions of values contained within the review data.

- SciPy: An open source Python library used for technical computing. This will allow the performance of complex computations such as regression and the calculation of confidence intervals.

In addition to these programming tools and peripherals, mathematical tools will also be required for this analysis to determine correlations and trends.

- Regression: This research anticipates to use a variety of regression techniques to aid in the modeling of trends found in the data, as well as to assist in the accurate prediction of future results.

- Confidence Intervals: Confidence intervals will aid in the determination of the statistical significance of any patterns that are located.

7 MILESTONES

The current main milestones of this project are detailed below, and will be updated with each project sprint.

(1) Remove the neighborhood attribute

(2) Distinguish between United States and Canadian postal codes.

(3) Clear out NaN, Null and Missing values.

(4) Parse open business hours to a single unit to help later computations.

(5) Create normalization scores for each review to use in overall analysis.

8 MILESTONES COMPLETED

Thus far, the data for `yelp_reviews.csv, yelp_business.csv` and `yelp_hours.csv` have been cleaned. As detailed above, all of the files have been loaded into Pandas DataFrames and removed all missing or null attributes. The following sections of the analysis will explain, at length, the cleaning process for each file.

8.1 Data Reduction

8.1.1 Reviews. The reviews are the primary target of what this analysis focuses on, and, accordingly, thorough processing is necessary to promote ease of access. The cleaning of reviews primarily centered around removing attributes not contained within the scope of this research. This included attributes such as `text`, containing the physical content of the review, and various review reactions (`funny`, `cool`, etc). The `text` attribute was removed because it made up the bulk of the information, and therefore memory requirement, of the review category, and there are presently no plans to make use of it. The reactions attributes have been removed due to their lack of relevant information. In an attempt to save on memory, any attribute that was not specifically intended to be used for a predefined purpose was removed. Fortunately, the Python Pandas library allows for very easy handling of this data reduction. The reviews file must be read into the program using python's `TextFileReader` object while still maintaining an awareness of the space requirements. Parsing the data in chunks (currently 500,000 entries at a time) allows the Jupyter Notebook to more readily apply certain techniques such as KMeans without requiring too much time or processing power.
The cleaned reviews DataFrame contains the following attributes: `review_id, user_id, business_id, stars` and `date`. The specific uses of all these attributes is listed in Section **PUT FUTURE SECTION HERE**.

8.1.2 Businesses. The cleaning conducted on the `yelp_business.csv` file closely mirrors that of the previous section with `yelp_reviews.csv`. Notably, the cleaning focused on the removal of attributes irrelevant to the main analysis. This included removing the `neighborhood` attribute which was not only missing values, but represented inferior method of determining location, especially when compared to using the `latitude` and `longitude` attributes. While this research may later involve using attributes such as `categories`, it is currently unclear how doing so will provide any interesting knowledge and how it can be applied. This attribute is a delimited list of business types and features, and parsing it will require the creation of a separate function specifically tailored to this data in order to put it to effective research use.

8.1.3 Users. This data set is used less significantly than some of the others. In particular, it is rarely used in standalone form, but instead in conjunction with other data. By synthesizing the other data sets with information on the users that wrote a specific review or visited a particular business, much more room for analysis is opened up. By comparisons involving a user's average stars or review total with features concerning their reviews or the businesses they reviewed, trends may become apparent regarding user preference and perception. With concern to data reduction, networking traits such as a user's friends or user reactions were removed. Once sufficiently reduced, this dataset consists of four attributes: `user_id, name, review_count` and `average_stars`.

8.2 Data Integration

Data integration for this research was achieved using the Python Pandas library, particularly the ability to concatenate multiple DataFrames into a single, unified DataFrame. This tool allows us to perform operations on data originating from different sets at the same time.

8.2.1 Business and Review. Combining the Business DataFrame (containing information about the businesses reviewed by Yelp users) with the Review DataFrame (information regarding each individual review) allows for the representation of the location of a review as well as the stars given. This combination requires plotting on three dimensions, but shows the distribution of stars, not only of a particular business, but also of locations. This process of clustering utilized in `Figure 1` will be discussed at length later on.

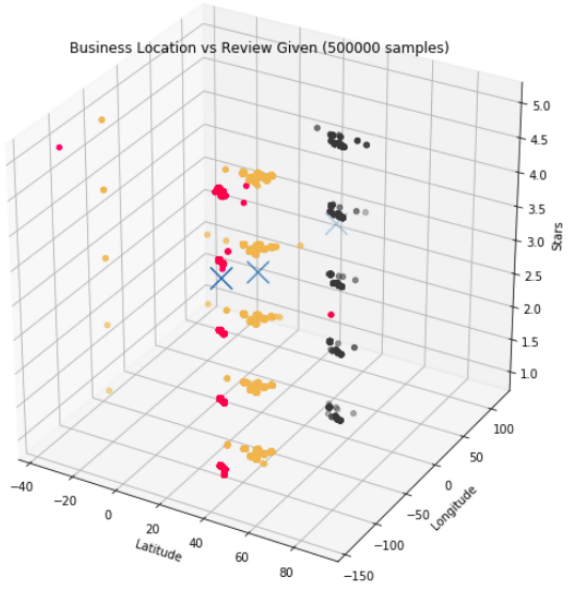

Figure 1: Using KMeans on combined dataset

8.2.2 Review and Users. Both the Review and Users DataFrames contain a `user_id` attribute on which the tables can be joined. Further analysis is required to determine if this aggregation will yield interesting information.

8.3 Data Cleaning

Currently, the vast majority of data cleaning revolves around removing NaN, Null, or missing values. The `postal_code` attribute requires the most attention to this end. The principal issue regarding this atttribute is malformed zip codes. Similarly, in order to effectively integrate data as listed in Sections 8.2.1 and 8.2.2, any Null values must be dropped. Failing to correctly perform this will lead to the failure of numerous tests, such as K-means.

9 RESULTS SO FAR

This research has uncovered a fair number of small, interesting patterns in the data, primarily identified within reviews and business location. Presently, the primary data mining technique used has been clustering, mostly through use of the K-Means clustering method.

9.1 Analysis of Business Location

Initially, it's important to understand where the businesses being analyzed lie. Using K-Means clustering with $n = 3$ clusters, it becomes clear that the businesses lie in three distinct clusters, along with several outliers.

The three clusters are clearly displayed above, each colored differently. Importantly, the locations of the three centroids of the data were pinpointed as (34.51432538, -113.22797375), (41.5357435, -80.26255264), and (53.03809423, 1.81518388). These centroids represent the location where the majority of the businesses computed in this set were centered around. In this particular example, only about 20% of the total data set was used. Though the above image may not appear to have 1 million data points, the businesses are largely clustered in small areas spread out over a large map, resulting a large amount of overlap. This does not actually tell us of anything we did not previously know, but grouping the businesses in this manner makes future, more specific analysis based on location much easier.

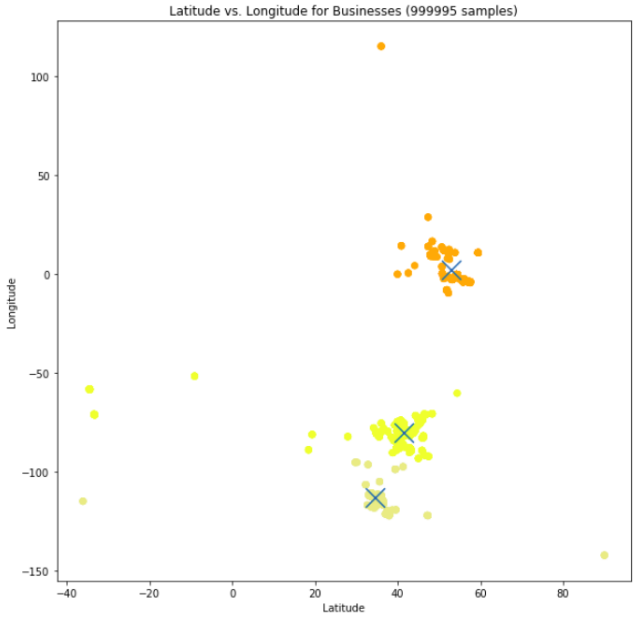

Figure 2: Using KMeans to cluster businesses

Similarly, clustering was performed on on two clusters closest to each other to find patterns of significance in a smaller subset of the data. This can be seen in the figure below.

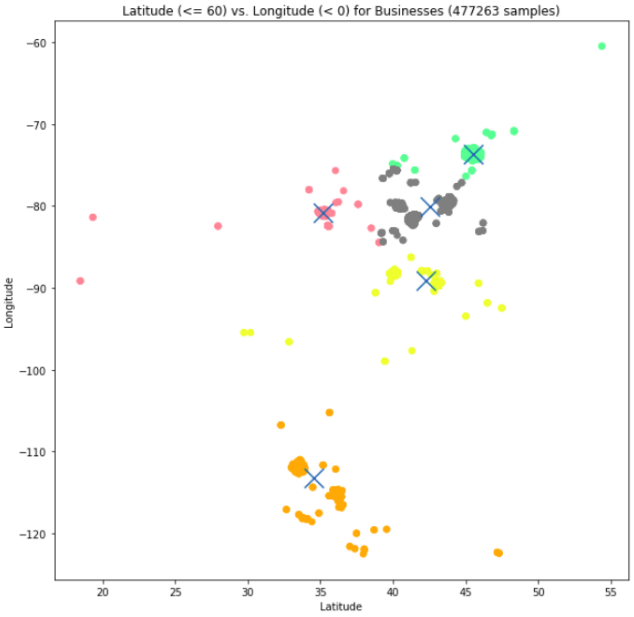

Figure 3: Using KMeans to cluster businesses

9.2 Review and Star Count

When analyzing the correlation between a high number of reviews and a person's willingness to give out stars, an interesting trend appears. Though some people have a tenancy to give abnormally high or low stars, the amount of stars dedicated users give on average tends to be more or less uniform. Interestingly, a similar trend is true in the case of the review count of a business. An establishment with a high number of reviews is likely to have a score falling in a very small range, with the overwhelming majority of businesses with more than 5000 reviews having an average of four stars. This is interesting as it would appear as though those companies with the most patrons, though having a high star rating, are not, in fact, the best available options based on customer feedback. This trend is demonstrated in Figure 4.

9.3 User Bias and Normalization

As shown in the figures above, different users tend to provides stars in different ways, resulting in different averages per person. Because of this, it is difficult to tell what a user actually thinks of

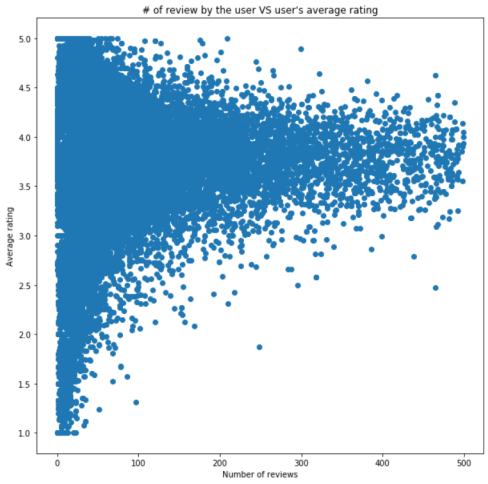

Figure 4: Comparing the quantity of user reviews to their average rating

an establishment comparatively, as a 3 star rating could be poor for someone who generally rates 5 stars, but is quite good for a user who commonly rates 1 star. Interesting information could be uncovered through a normalization of a user's reviews based on their average. This allows comparison of user reviews to each other without them being skewed by each user's unique tenancies.

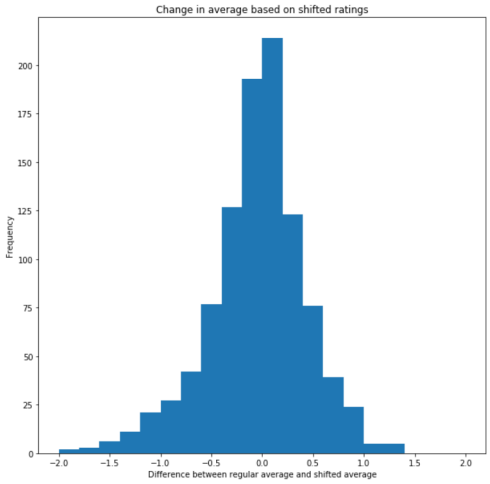

Figure 5: Change in business stars based on shifted user scores

As shown in Figure 5, represented by a box plot for a random sample of 1000 businesses, the change in the business' average rating following user normalization is somewhat significant. In order to find these new averages, each business has its corresponding reviews taken and have the individual review scores adjusted in comparison to the star average of the user ID that made the review. The score of each review is increased or decreased based on its deviation from the posting user's average stars.

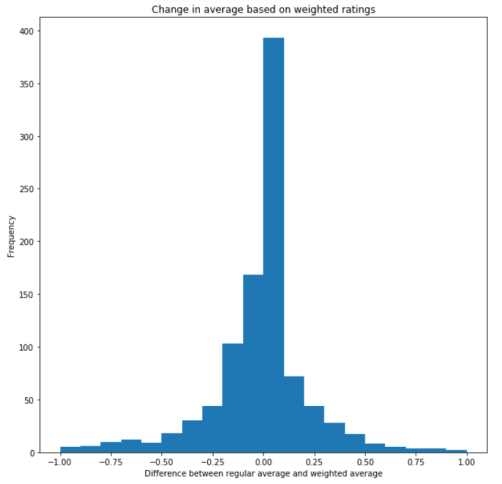

Figure 6: Comparing the quantity of user reviews to their average rating

Figure 6 shows how the business scores change if all the reviews are weighted by the formula: $weight = 1.0 + 0.01 * (review\_count - 50)$ for users with greater than 50 reviews. This allows those more dedicated reviewers, who likely have more experience, to have a greater effect on the rating of a business.
Both bias adjustments demonstrate a change in the overall rating of a business. Some businesses had their ratings change by an entire star, which is significant on a scale with a range of 4. This shows that the quality of a business is not necessarily entirely determinable by their sheer number of stars alone, resulting from the fact that some businesses may be more likely to be reviewed by those with a propensity towards favorable or unfavorable reviews, or simply by those who rarely review and have not demonstrated their abilities yet.

10 FUTURE MILESTONES

(1) *Stars of a Business and Average Patron Stars*
Initial analysis of these attributes shows that, when visualized, many of them are concentrated around certain, unexpected values. Presuming a normal distribution, most of the values would average around 3. However, preliminary results indicate that the mean value of business stars and patron stars is approximately 4. Additional analysis is required to determine the presence and significance of any underlying connection.

(2) *Continuing Clustering*
Clustering has demonstrated its usefulness to this research. However, it is implausible to think that all significant patterns have already been revealed. Focusing on the three centroids lain out in Figure 1 may aid the discovery of additional patterns between business locations, stars, and reviews.

(3) *Regression Analysis*
By application of certain regression techniques, this research

intends to create models in order to better predict the locations in which a business is more likely to find success, along with where patrons tend to give the highest reviews. Success will be determined by the present state of a business, being open for operation or permanently shut down.

(4) *Further Research*
The information presented in this analysis provides hopeful preliminary results. Most of the knowledge gained has come from a select few attributes. Specifically, most of our patterns and information has been obtained from business locations, business rating, user reviews, the average user review. Despite what has been uncovered, it is extremely unlikely that all trends, or even the majority, have been found and, as a result, there is still much work to be done.