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 ABSTRACT

It's important to provide consumers with easy access information in regards to their favorite local businesses. Yelp is one of the primary mediums in which people gather information about businesses closest to them. 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 various data mining techniques, this work intends to reveal trends that define a business in the public eye.

Specifically analyzing locations of businesses, average star rating, and total number of reviews, this research finds several interesting preliminary results. Most notably, we find interesting correlations between average business rating and average stars given by users. Similarly, and unsurprisingly this research found that the majority of data is found by users who have made less than 100 reviews, and businesses with less than 500 reviews. We also find that the highest rated businesses with the most reviews are usually centered around a 3.6±0.2 rating which is what we will consider the average review throughout this analysis. Finally, when clustering on business locations this analysis finds that there are nine main clusters in which these businesses are located. Overall, it the data mined in this project suggests that there is no major influence involving reviews on Yelp and the overall success of the business that is being reviewed.

2 INTRODUCTION

This paper and the research conducted seek to investigate factors that contribute and influence the overall perception of a business. This analysis uses a variety of techniques, however because the data focuses heavily on location, KMeans clustering is one of the main techniques used. This research is interested in finding the location of businesses with the highest ratings are located. It will also analyze credible reviews by normalizing the ratings of businesses based on the average review stars from a user. Clustering in 3-dimensions of latitude, longitude and business rating also helped determine what an average business looked like, and ultimately allowed the analysis of businesses both above and below the average. Using methods such as normalization, clustering and regression, these applied techniques help locate successful businesses to provide information to consumers of where the best businesses are. As mentioned above, this research is aimed at trying to provide this kind of information to people who intend on visiting these cities. While this research was unable to provide a comprehensive analysis of all domestic businesses, through the analysis of the nine main clusters this research is able to provide the best locations to travel in which the businesses successful and popular.

These questions are important on a personal level, to inform consumers of businesses with which other consumers had a positive experience.

While personal experience may vary on a case by case basis and includes many other factors. Statistically speaking a business with many high reviews is likely going to be more enjoyable than a business with many low reviews. What's important is sifting through all of these businesses and picking the ones who have the highest ratings with the most amount of reviews. In order to make suggestions of where to visit, consumers want places with better reviews, which is what this research intends to do.

3 RELATED WORK

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.

[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

(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.

4 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 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.

This dataset contains a variety of data types, mostly quantitative. Attributes such as 'state', and 'category' are both nominal attributes. The 'is_open' attribute is binary, mostly all other attributes are numeric, an ratio scaled (meaning that there is an inherent zero point) with the exception of longitude and latitude. The latitude and longitude attributes are interval scaled since they do not have an inherent zero point.

5 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] 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

6 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.

7 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. Python allows for importation of modules and libraries to further enhance its utility.
- 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.
- Seaborn: An open source Python library for creating and interpreting high level statistical information. This allows the generation and application of logistic regression lines.
- Folium: An open source Python library for visualizing map data in an effective and intuitive way. Folium facilitates the creation of complex and interactive maps.
- Matplotlib: An open source Python library for creating simple but effective graphical representations of data. Matplotlib allows for extensive visual customization.

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.
- Clustering: Clustering allows for the determination of natural groupings within the data.

8 TECHNIQUES USED

Many techniques were used in this project, spanning from data cleaning to analyzing. Each technique is detailed below.

8.1 Data Cleaning

(1) *Data Reduction*
Attributes that were missing too many values were dropped from this analysis, especially location attributes as longitude and latitude offer more precision when determining locations.

(2) *Modifying inconsistent data*
Attributes (mostly strings) were modified to be consistent with the other data points in the same category. Zip codes also modified to distinguish between US and foreign locations.

(3) *Data integration*
Data integration was an integral component in the data analysis. For example, when focusing on how users and businesses affected one another, it was important to include and match data from both tables so that all the necessary values needed to analyze any patterns in the dataset were easily available.

8.2 Clustering

(1) *Clustering using K-Means*
K-Means clustering was one of the primary techniques used in this analysis in order to determine inherent groupings within the data. Similarly, clustering in 3 dimensions by including multiple attributes was also used to help analyze patterns.

(2) *Clustering using density*
Density based scans were also used on several attributes, including location and review count, in order to analyze patterns in the data.

8.3 Outliers

(1) *Outlier Analysis*
During clustering, specifically when using K-Means, the data set was framed to focus on different clusters to help prevent the centroid location of the overall group being skewed as a result of several significant outlying data points.

8.4 Regression

(1) *Logistic & Linear Regression*
Regression was used help determine how the average star count of reviews affected a business' performance, as well as to help describe trends in user's average star ratings versus their total review count.

9 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.

9.1 Data Reduction

9.1.1 Reviews. The reviews are the primary target of what this analysis focuses on, and, accordingly, thorough processing is necessary to promote ease of analysis. 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 were no plans to put it to any use. 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 of 500,000 entries at a time allows the Jupyter Notebook to more readily apply certain techniques such as K-Means 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 are listed in Section 10.

9.1.2 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.

9.1.3 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`.

9.2 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.

9.2.1 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.

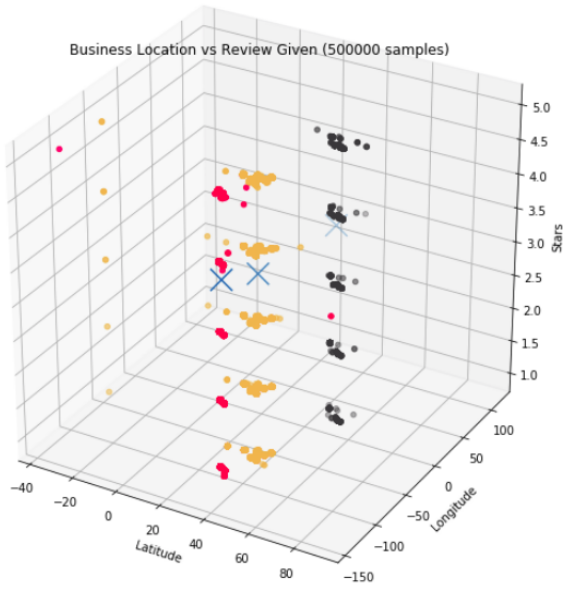

Figure 1: Using KMeans on combined dataset

9.2.2 Review and Users. Both the Review and Users DataFrames contain a `user_id` attribute on which the tables can be joined. This has led to some interesting information about the distribution of review scores in respect to user averages.

9.3 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 attribute is malformed zip codes. Similarly, in order to effectively integrate data as listed in Sections 8.1.2, any Null values must be dropped. Failing to correctly perform this will lead to the failure of numerous tests, such as K-means.

9.3.1 Business Hours. In order to explore any possible relationship between a business' success and their hours, the data from the file `yelp_business_hours.csv` was used. This file included the business id, and seven additional columns representing the days of the week containing strings such as "09:30-22:45". Primarily, all null values and rows containing "00:00-00:00", which indicates that a business is not open, needed to be removed. In order for this information to be useful it cannot be in a string format. As a result, the necessary data cleaning consisted of a script that iterated through the entire data set and parsed the strings into integer values. After all the stringers were workable numbers, the script calculated the difference between the open and close times and created a new column called `hours_open` to store this information.

9.3.2 Business Categories. Additionally, the data from the file `yelp_business.csv` contains the column `categories`. This data is stored as a large string of categories separated by ";". In order to gain information from this data it must be parsed. The categories that this research sought to examine were "Coffee & Tea", "Restaurants", and "Open Late". So again the script iterates through all of the data looking for these categories and adding a new column for each respectively.

10 RESULTS

This research has uncovered a fair number of small, interesting patterns in the data, primarily identified within reviews and business location. The primary data mining technique used has been clustering, mostly through use of the K-Means clustering method. This analysis specifically was interested in clustering based on location and by star count in order to determine if there exist any areas with a noteworthy difference in the amount of stars the businesses there received. In addition, logistic regression was performed in order to determine correlation and predict how the star count of a business relates to its chance of success.

10.1 K-Means and its use

As the subsequent sections greatly detail, K-Means clustering is one of the primary data mining techniques used throughout this research. It is particularly useful when trying to find patterns in the location of the dataset. One of the questions this analysis sought to answer was to see how the location of a business impacted that business' success. By applying clustering techniques to different variables, the process provides interesting and unique results. Additionally, by clustering in three, three different variables can be plotted against one to determine if any of them are related when there is uncertainty. Figure 1 is provides an example of how this works by clustering based on longitudinal and latitudinal location as well as the star rating of a business. 2D clustering demonstrates that there are 3 large clusters based on location. 3D clustering analysis shows the same information for location as the 2D clustering, but also that there is a relatively uniform distribution of star reviews in those locations. The clusters further center around the mean star value of the data. While this may not seem useful, the knowledge and application section of this research discusses that location does not have an impact on the success of a business, and ultimately should play less of a major role in the typical real estate discourse involving the creation of a new business.

10.2 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 with a different color. 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 locations where the majority of the businesses handled in this set were centered around. In this particular example, only about 20% of the total data set was used. The above image actually plots one million data points. However, it does not appear to have this much data as the businesses are overwhelmingly clustered in small areas spread out over a large map. As such there is a massive

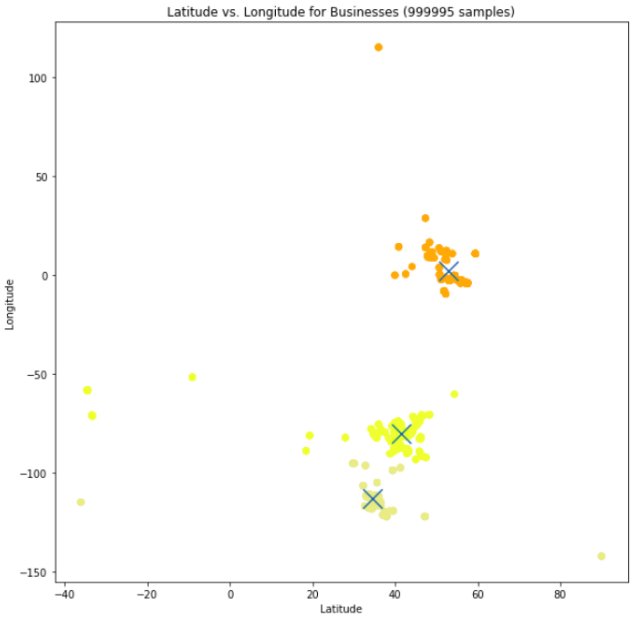

Figure 2: Using KMeans to cluster businesses

amount of overlap. Doing this process did not actually reveal anything not previously known, but grouping the businesses in this manner facilitated future, more specific analysis based on location. Similarly, clustering was performed on the clusters closest to each other to find patterns of significance in a smaller subset of the data. This is evident in the figure below.

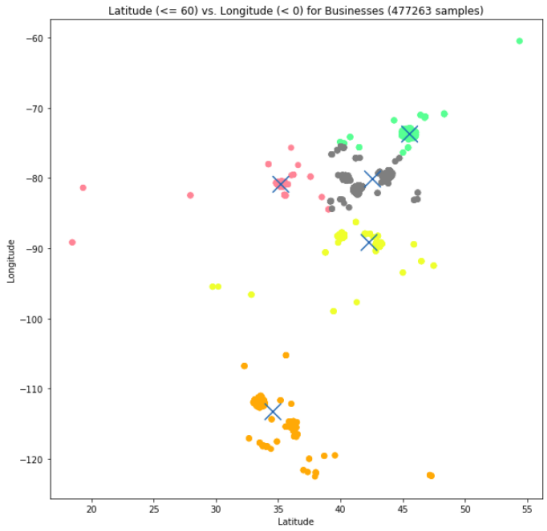

Figure 3: Using K-Means to cluster businesses

10.3 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.

10.4 User Bias and Normalization

As shown in Figure 4, different users tend to provide stars in different ways, resulting in different averages per person. Because of this, it is difficult to tell what a user actually thinks of 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 can 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.

As shown in Figure 5, represented by a box plot for a random sample

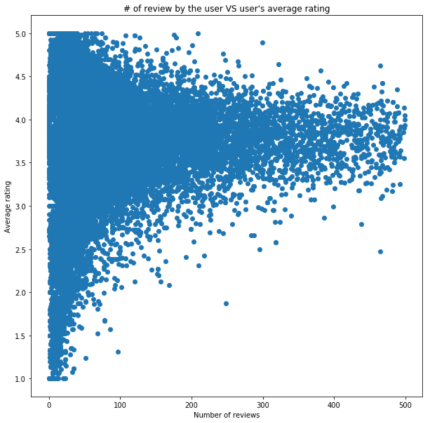

Figure 4: Comparing the quantity of user reviews to their average rating

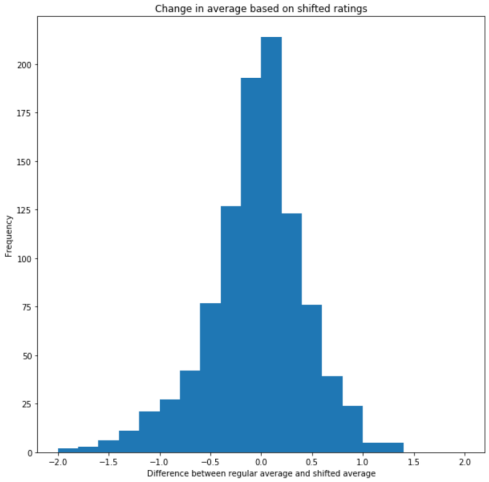

Figure 5: Change in business stars based on shifted user scores

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 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

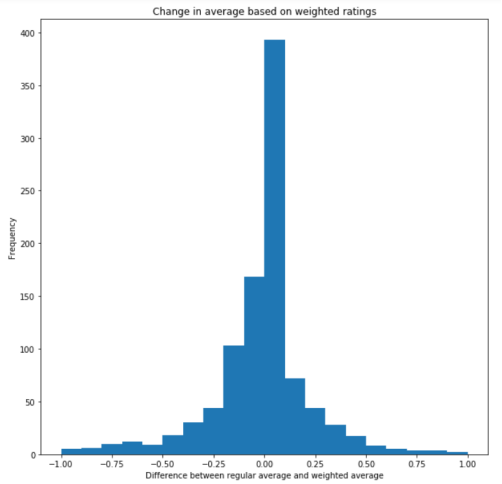

Figure 6: Comparing the quantity of user reviews to their average rating

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 yet refined their abilities.

10.5 Different Types of Businesses

In trying to determine the factors that make a business successful, this study looked to examine the distribution of reviews on different types of businesses.

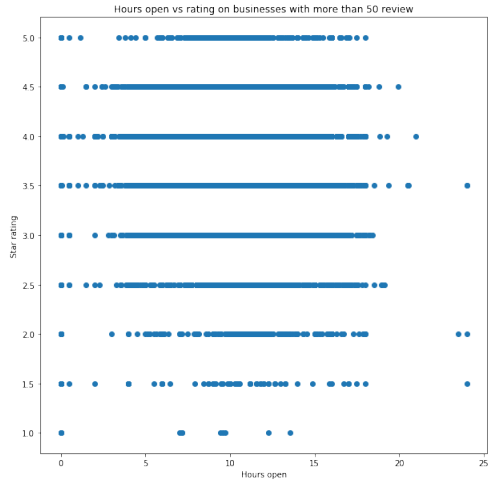

Figure 7: The correlation between hours open and star rating of a business

10.5.1 Hours Open. As can be seen immediately from cleaning the data, a large portion of businesses chose not to list their hours on yelp at all. Not taking those businesses into account, there is an approximately regular distribution of open hours with an average of about 10 hours per day. In figure 7, the hours a business is open are compared with the star rating. No obvious correlation can be seen, and the correlation coefficient as calculated by pandas for the data is ≈ -0.0267. Since something with a correlation coefficient of 0 is entirely uncorrelated, this data largely lacks any correlation. This demonstrates that the star rating of a business has almost nothing to do with how many hours a day it is open. This, in combination with the Logistic Regression in section 10.6, leads to the conclusion that the general success of a business and it's success on Yelp are more or less unrelated and that Yelp is not an effective way to gauge business performance.

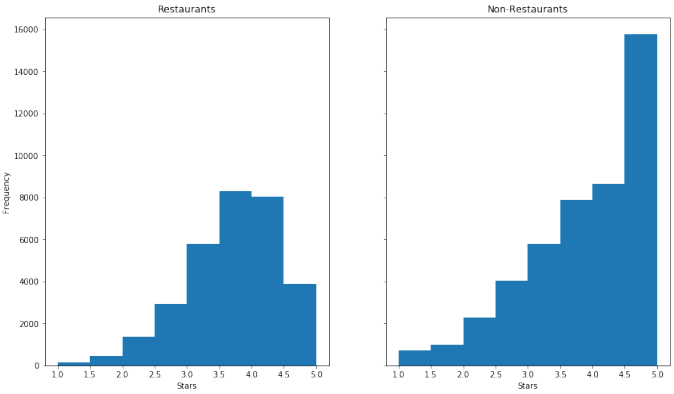

Figure 8: The distribution of yelp review for restaurants vs non restaurants

10.5.2 Restaurants vs Others. Although Yelp is mainly known as a service to use for reviewing restaurants, about half of the businesses on the service are not actually restaurants but instead other forms of establishment. As can be seen in Figure 8, there is a massive difference in the review distribution between restaurants and non-restaurants. There is not such a dramatic different in other types of establishments. For example this study also examined coffee shops in comparison to the entire dataset and found little to no difference in the distribution of the reviews. This demonstrates that Yelp is generally far more critical of restaurants than it is of other types of businesses. Additionally, it shows that it is there is a significantly higher likelihood that a non-restaurant business will receive a five star review.

10.5.3 Late Night vs Daytime. Figure 9 looks surprisingly similar to Figure 8 in that there is a dramatic difference in the distribution of reviews between establishments open later than 10 vs establishments that close before 10pm. Businesses open later than 10pm are significantly less likely to receive a five star review. It is highly unlikely that, in Figures 8 and 9, the businesses that are non-restaurants and close early are simply better businesses. As a result, this implies that there exists a significant bias among reviewers and that there are different expectations for these types of establishments.

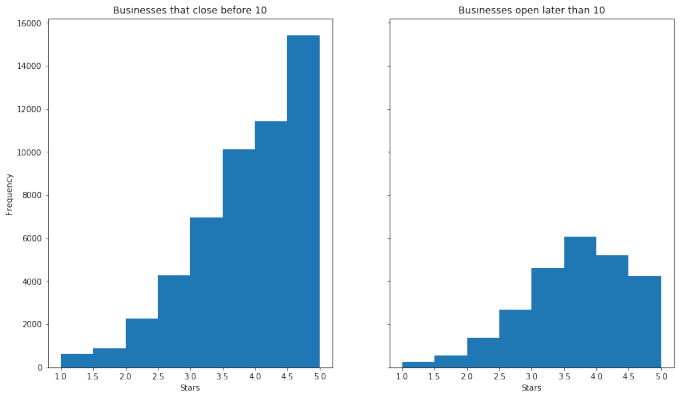

Figure 9: The distribution of yelp review for establishments open later than 10pm vs establishments that close before 10pm

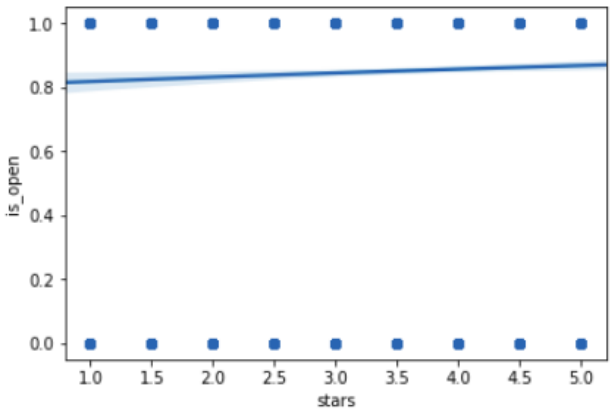

Figure 10: Logistic Regression on `is_open` and `stars`

10.6 Logistic Regression

In most situations throughout this analysis, linear regression would not be a logical means of prediction for this data as nothing has any sort of strictly linear relationship. This is mostly attributed to the type of data this is; latitude and longitude are interval scaled data types[8] Other attributes such as star ratings which was ratio scaled[9]. As seen in Figures 2 and 3, the location data, while clustered, has no obvious further patterns. Other attributes such as `neighborhood` as mentioned in section 9.3 had been dropped entirely due to inconsistent and missing values. At this point, it should be obvious why longitude and latitude would be the likely candidate for location determination. Despite all of this seemingly non-linear data, logistic regression became useful for reinforcing knowledge. Similar to the common applications of linear regression, this analysis instead attempted to use logistic regression to predict whether certain variables could indicate whether or not the business had shut down.

10.6.1 The Application of Logistic Regression. It's important to note that logistic regression is only useful when using binary data. Thankfully the `is_open` attribute is a boolean value with 1 meaning the business is currently open and 0 conversely meaning the business is currently shut down. What one would logically expect in formulating a hypothesis that some variable did effect the overall performance of a business is a regression line that has a distinct curve to predict the open status of a business. At a certain threshold (and for the purposes of the experiments performed in this analysis) such as 0.5, any point that is above that threshold would be considered open, and any point at or below would suggest the business has closed. If one were able to predict if a certain star rating affected if a business is open, then this analysis would push for businesses to achieve this lower limit in order to stay open. As seen in the following sections, this simply isn't the case.

10.6.2 Logistic Regression: `is_open` *vs* `stars`. As seen in figure 10, when using logistic regression on the desired variables, the results did not indicate any sort of immediate pattern between the stars of a business and the continued operation of said business. As has been mentioned before, a typical regression line that implies a relationship would have a clear curve indicating that change in prediction. This graph however does not contain that and ultimately fails to provide an effective prediction about how a star rating might influence a business. This graph does however inform us that star ratings, in fact, DO NOT have a remarkable impact on the success of the business. Thus, an overall poor star rating does not necessarily indicate the business would be affected at all. This knowledge indicates that businesses ought not to put significant focus into their Yelp rating. This could theoretically be because people are more likely to leave a review if they are upset with the service they received, but the average customer may not want to take the time leave a review whenever they patron to a business. At this point in the analysis it's important to determine how Yelp reviews should be weighed in important business decisions concerning performance and public perception. Should they be specifically interested in knowing the state of mind of the reviewers? Are they simply on one extreme or the other, only leaving reviews when lasting impressions are made on the person? These questions are far outside the scope of this analysis but are interesting topics to consider nonetheless.

10.6.3 Logistic Regression `is_open` *and* `review_count`. Similarly to the process listed in the section above, it was thought that the `review_count` attribute might influence business operations. However, again, it can be seen in Figure 11 that this simply is not the case. There is a similar pattern in both of these suggesting that `review_count` is not a factor at play because the regression line doesn't indicate any remarkable impact. Both of the graphs referenced strongly imply that neither `review_count` or `stars` influence the success of a business. Even though some restaurants may have received many more than 400 reviews, in order to illustrate the point, the data had been trimmed down to only visualize businesses with up to 400.

[8]There is no inherent zero point in this data
[9]Scaled at equal intervals, with an inherent zero point

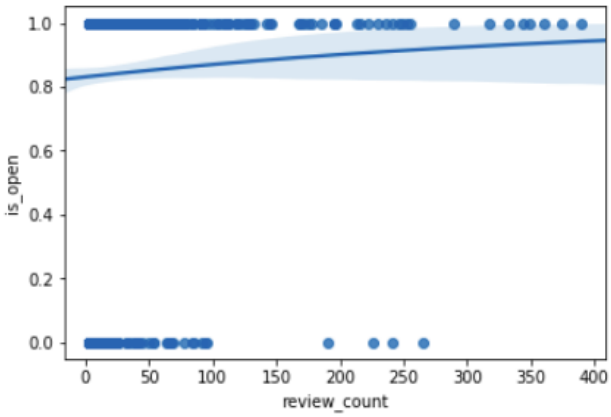

Figure 11: Logistic Regression on `is_open` and `review_count`

10.7 Density Based Analysis

Density based clusters obviously exist in the location data. However, this analysis was interested more in discovering density patterns in reviews and stars. At first, figure 4 shows a general trend in where most of the data lies. As this analysis has established, the mean for stars usually is between 3.5 and 3.9. While this might be interesting, perhaps what's more interesting in seeing just exactly *how many* of these reviews lie within this range (and others).

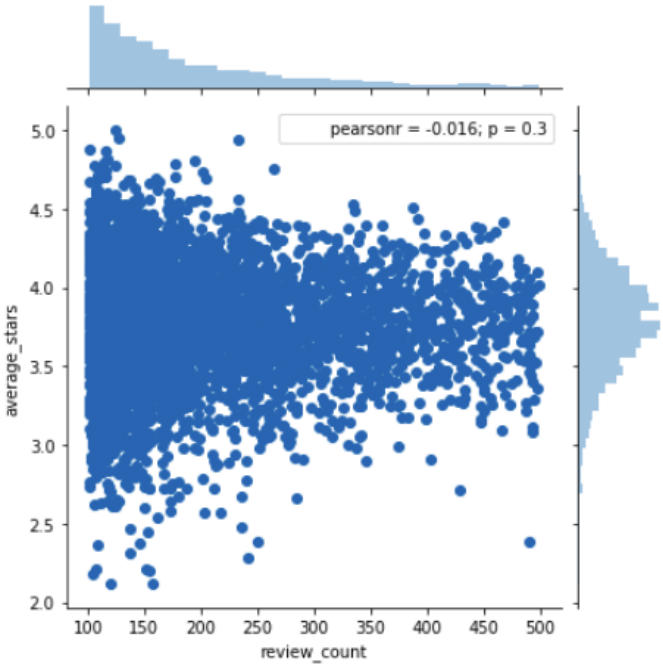

Figure 12: Dual plot of average stars and review counts

10.7.1 Density analysis on Stars vs Number of Reviews. Seaborn allowed for the creation of Figure 13. This specific plot is calculating the Pearson Linear Correlation Coefficient as well as plotting the densities of stars given average rating. As we can see, there is fairly normally distributed data amongst the stars attribute, centering around 3.75. The data is centered around 3.75, which is a logical prediction as it is approximately the mean. Similarly, this histogram of review counts shows a very high density of reviews towards the lower end of the spectrum. Meaning that most businesses likely don't receive over 100 or 200 yelp reviews, at least not presently. Prior sections have hinted at how most of the data in this data set is non-linear, and this can be seen in this graph as well. The correlation coefficient is nearly 0, which is what we'd expect for this kind of ratio-scaled data. One interesting note is that businesses with more than 100 reviews don't have less than a two star rating based on the samples taken.

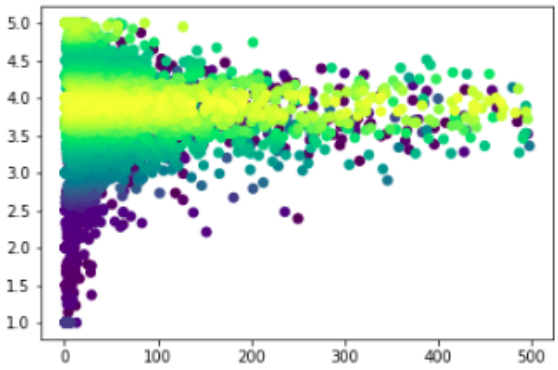

Figure 13: Dual plot of average stars and review counts

To help better visualize the ideas presented thus far, Figure 13 provides color to indicate the density of various regions of the plot. The yellow regions represent areas of high density reviews. We can see that the most of these reviews lie between 3.5 and 4, while the less dense purple regions surround the outside. Another interesting observation is the small clustering of 5 star ratings at the top, implying that the two highest density regions for reviews at about 3.75 and 5.

10.8 Density Visualizations

Up until now, this analysis hasn't shown where most of these businesses are when actually placed on a map of the continental United States. The dataset states that this data is spread across 11 metropolitan areas, and those remained unknown until placed on the map seen in the figure below.

Figure 14 gives a great visualization of where this data was collected. It also brings up an interesting problem with this analysis overall; the clusters of cities do not span the continental United States. We see there is a large portion of the data near the upper East cost, and only two larger clusters near the west coast but nothing in between. It is important to know how this could influence this analysis, specifically asking the question of these cities are an accurate representation of the entire country. A more thorough dataset would be needed in order to determine if this had an impact

Figure 14: Visualization of the main cities in the Yelp dataset

on the findings of this research. In order to better help understand if this data *does* accurately represent most of the country, this analysis looks closer at the cities to see any intra-city patterns. Figure 15 shows a zoomed in view of Scottsdale and Phoenix, Arizona to better determine if:

(1) This data would be able to accurately represent the greater area of the country
(2) There might be any patterns, or specific locations where businesses are doing particularity poorly.

This information will allow businesses to understand exactly how big of a factor location is when making important real estate decisions. Typically, businesses (especially restaurants) tend to view location as extremely important. The idea is that the right location could drastically improve a business'. This makes sense since customers are likely to go to a business that is easily accessible, in a *good part* of town or that has amenities such as on-site parking.

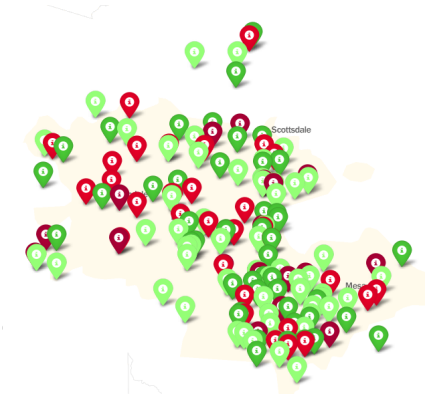

Figure 15: Visualization of Phoenix, Arizona

In Phoenix there are obviously many businesses. Perhaps what is most interesting about Figure 15 is that there's no clear areas where poorly rated businesses are and where successful businesses are. In this figure, the highest rating businesses are dark green and the lowest rated are dark red. Poorly rated businesses are intermingled with successful businesses and there are not any boundaries distinguishing any noticeable pattern. This leads this analysis to believe that bad reviews don't necessarily have an impact on success, an idea which was mentioned earlier in Section 10. This knowledge is helpful (and will be reiterated in the next section) because it removed the importance of location from business' decision to open.

Another finding that could be put to use is the normalization data from Section 10.4. Because it is clear that some users have a propensity to rate certain ways, it may be useful to normalize user ratings so that businesses that users find exceptional are properly recognized as such, even if their audience is particularly critical and only give a three star rating to even the best businesses. Additionally, people who are too nice to give poor ratings to businesses shouldn't unfairly tip an establishment's rating simply as a result of that. By doing this, Yelp could alter its functionality to provide a more fair and, arguably, more accurate rating.

11 KNOWLEDGE GAINED

As many prior sections have mentioned, there is actually a surprising amount of knowledge that was gathered from the data. Typically, one might think that the data we had found would be relevant in any sort of situation. This research was unable to find a pattern that might change the Yelp industry as we know it. That is not a problem, and in fact being able to draw conclusions about information that wasn't found can be just as powerful as data that shows significant patterns. Figures 14, 15, and 2 all reinforce the finding that location doesn't have an impact on how successful a business might be. Similarly, the logistic regression shown in figures 10 and 11 suggest that the number of reviews and stars of a business do not have any impact on the success either. This research also finds that people generally tend to give more 3.5 star reviews than anything else and also finds that even successful businesses get bad reviews. It is thought that business hours affect the success of a business and ones that close earlier tend to do better overall. It

also is thought that restaurants generally perform worse than other businesses which is likely caused not by one's experience but the quality of the food. This could be a more sensitive measurement, since people aren't buying a service necessarily (as you would to get your taxes done) but buying a good which not everyone might like. There is not set standard for food and everyone's taste differs. Overall, this analysis shows there is no major influence on the success of a business.

12 APPLICATIONS

We can use this information in order to help businesses understand what factors do and don't influence success. Specifically, location isn't important, and it should be removed from real estate decisions because this analysis finds no solid information linking location to the success of a businesses. Similarly, we find that Yelp reviews as a whole do not have an influence on the success of a businesses. If one wanted to measure that it would be important to make a more comprehensive review system that took other factors about a reviewer in order to better understand what exactly makes a successful businesses. This information can be used to prevent people from opening restaurants since they are more poorly reviewed, however it should be made clear that everyone will receive bad reviews and that doesn't necessarily have an impact on how bad a businesses is.