Restaurant Analysis Report

0 - Introduction

We all love eating, but what lies behind all of the restaurant data displayed to us? In this document, we discuss how restaurant data can be used to determine restaurant popularity. Using data of business and services on Yelp, we are able to predict a restaurant's success and discover any trends in user rating behaviour. Throughout the document, we explain how we collected our data and the process we took to extract-transform-load. With our modified data, we perform several analysis to explore the following questions:

1. Are restaurants more popular during certain days of the year?
2. Does the nationality of a restaurant have an effect on the ratings it gets?
3. Does a user's rating change overtime as he/she reviews many businesses? Does the user raise standards? Lower standards?
4. Is a restaurant successful or not?

1 - Data Usage

I: Data Collection

The first task was to collect the data. A subset of Yelp's businesses, reviews, and user data of information about businesses across 11 metropolitan areas in four countries was made available from Kaggle. The full set of information is accessible through the following link: https://www.kaggle.com/yelp-dataset/yelp-dataset/version/6?select=yelp_business.csv. The datasets that are used and cleaned are described below.

II: Data Description

A) yelp_academic_dataset_business.json

This file is a dataset of businesses from the Yelp database. Each row contains basic information for a business. Much of this information is contained within the attributes column. The attributes column is a dictionary for each row with up 17 more columns of information to dissect. Some examples from the attributes dictionary include Alcohol, Wi-fi, Ambience, and restaurant attire to name a few. Information on the businesses location is also given, along with the rating of the business.

B) yelp_academic_dataset_review.json

This file is a dataset of reviews from the Yelp database. Each row contains information about a user's review. The basic information includes the review message, the star rating. Some more notable and interesting attributes of the data included statistics of the review such as how many users found the review funny, useful, or cool.

C) yelp_academic_dataset_user.json

This file contained information of all users who have done a review from the Yelp data set. The basic information that comes from this data set is how many reviews the user has done, their average rating, and their creation date. Some interesting statistics that came from this data set were how many "fans" (i.e. followers) they had, how many people rated their reviews as useful, funny or cool, and how many of the reviews include photos.

III: Data Cleaning

A) clean_business.py

This file takes the *yelp_academic_dataset_business.json* file. As we are doing analyses on restaurants, we filter out the non-restaurant businesses. Since there is a very large amount of data and many computationally-heavy operations will be performed on such data, we take the data from the city with the largest number of reviews, which in this case, turns out to be Toronto.

B) business_reviews_users_merged.py

This file takes the results from *clean_business.py* and merge those results with the users data as well as the reviews data. We have a one to many relationship between the users data frame and the reviews data frame so we joined those together on *user_id*. Then, we have a one to many relationship on the business data and the user+reviews data so we join these together on *business_id.*

We ran into some difficulty initially trying to merge these using just pandas. With the size of the reviews data being ~ 6.5 GB and users data being ~ 3.5 GB, the data would read into pandas but would take forever to do operations. We soon switched over to spark where spark was able to handle reading, filtering and joining, then writing the data to pandas much faster. This is essentially a small ETL pipeline before the data is really ready to be analyzed.

IV: Limitations, Problems, Potential Changes

The data was last updated in 2019, so, some information may be slightly outdated. For example, some restaurants in the data may have closed down since then, especially due to the COVID-19 pandemic. In retrospect, it would have been a good idea to download all of the datasets from the kaggle link mentioned above as merging all datasets would have provided more information for each business, review, and user. In addition, the ideal option would have been to download the single file from https://www.yelp.com/dataset. We were not able to do this, however, because the file was extremely large and difficult to open and work with.

2 - Analysis

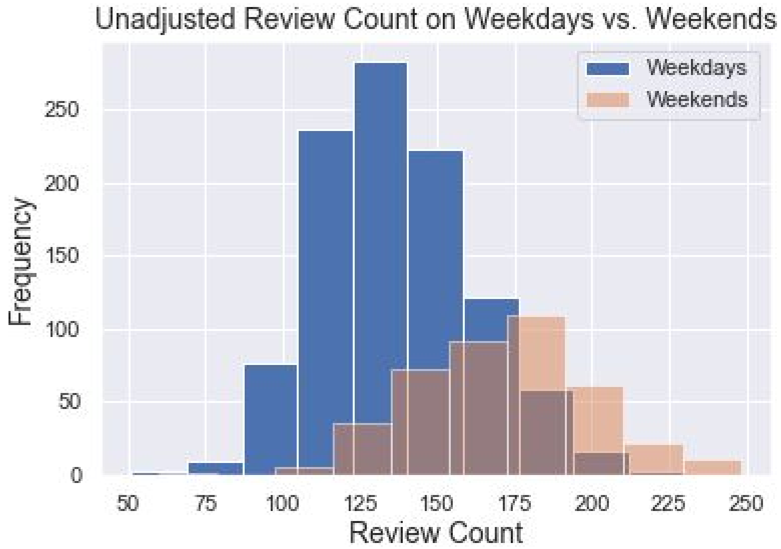
I: Restaurant Popularity

A) Introduction

In this section, we address the question: are restaurants more popular during certain days of the year? We do this using the business_reviews_users_merged.csv file. We answer this question by looking at the number of reviews submitted on the days we are interested in. We only take data from 2016 and onwards. The reason for this is that daily reviews before 2015 tend to be lower in quantity and can easily cause a long left tail which would not properly represent the review count distribution. In addition, it is more relevant to take more recent reviews as old reviews are not as relevant. Restaurants could have undergone many changes over the years, resulting in a popularity shift, thus, a shift in review counts. Restaurant popularity is determined by the number of reviews a restaurant gets. In addition to this, delayed reviews are accounted for. Some users may leave a rating a day after their visit. We conduct separate analyses for this case and denote these analyses as the adjusted analyses, and the other analyses as the unadjusted analyses.

B) Unadjusted Weekends vs. Weekdays

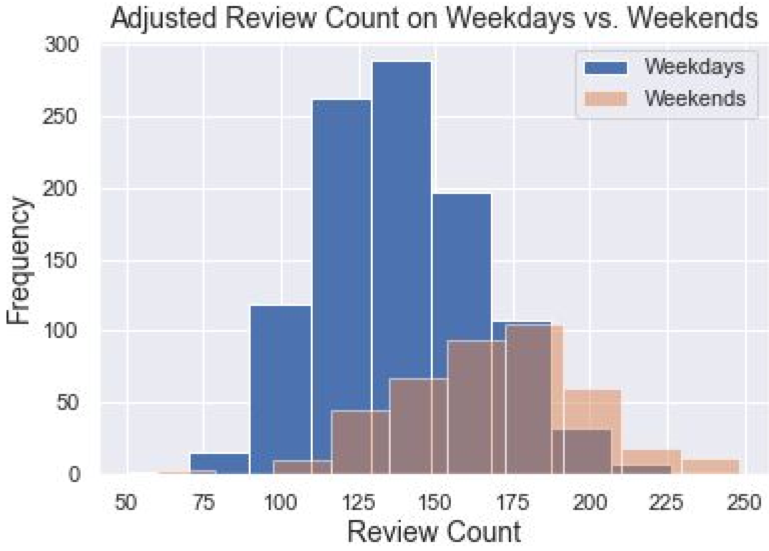
First, we answer the sub-question: are restaurants more popular on weekends or weekdays? We grouped the data using the dayofweek method in Pandas. From this, we can get a number ranging from zero to six which corresponds to which day of the week the review was left. Weekdays and weekends are separated into the variables x and y respectively. We obtain a count for the number of reviews left on each day of the week. Several statistical tests are conducted on this data. First, we conduct normal tests on both datasets. Weekdays had a normality p-value of 3.4e-05, failing the test, and weekends had a normality p-value of .75, passing the t-test. Though when we take a look at the histogram of the distribution of these two datasets, we can see that they appear to be normal enough, so it is still viable to conduct a t-test on this data. The equal-variance test fails with a p-value of .04, so we conduct a t-test assuming unequal variance. The t-test is used to compare means of the weekend and weekday means. The p-value resulting from this is 1.8e-78, failing the t-test. We can conclude that there is a difference between the means of the data. Some extra analyses were also conducted. We outputted the total reviews on weekdays vs. weekends, which yielded counts 141,116 and 70,501 respectively, indicating more reviews are submitted during weekdays. However, this is largely due to the fact that there are five weekdays and only two weekends, so next we outputted the average reviews per day for weekdays vs. weekends, yielding 136.9 and



171.1 respectively, indicating that weekends are much busier on a daily average. Observing the histogram, it is clear that the mean of weekends is larger than weekdays. From all of this data, we can conclude that restaurants are much more popular on weekends than weekdays.

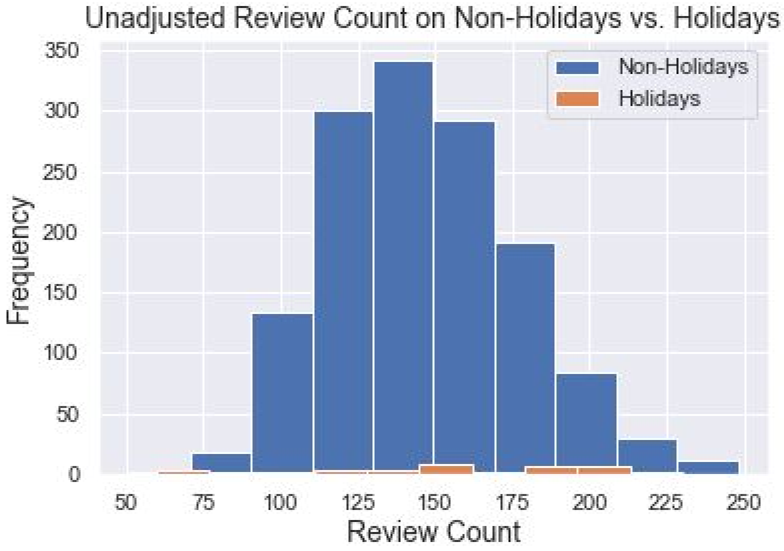
C) Adjusted Weekends vs. Weekdays

Second, we answer the sub-question: are restaurants more popular on adjusted weekends or weekdays? We must shift the data to account for one day late reviews. We do this by taking weekdays = days [1,5], and weekends = days [6, 0]. From here, we conduct the same analyses on the variables x2 and y2 representing weekdays and weekends respectively. First, we conduct normal tests on both datasets. Weekdays had a normality p-value of 2.3e-07, failing the test, and weekends had a normality p-value of .16, passing the t-test. Once again, when we take a look at the histogram of the distribution of these two datasets, we can see that they appear to be normal enough, so it is still viable to conduct a t-test on this data. The equal-variance test fails with a p-value of .005, so we conduct a t-test assuming unequal variance. The p-value resulting from this is 4.5e-59, failing the t-test. We can conclude that there is a difference between the means of the data. The total reviews on weekdays vs. weekends yielded counts 142,297 and 69,320 respectively, indicating more reviews are submitted during weekdays. The average reviews per day for weekdays vs. weekends, yielded 138.0 and 168.2 respectively, indicating that weekends are much busier on a daily average. Observing the histogram, it is clear that the mean of weekends is larger than weekdays. From all of this data, we can conclude that restaurants are much more popular on weekends than weekdays when taking into account delayed reviews.



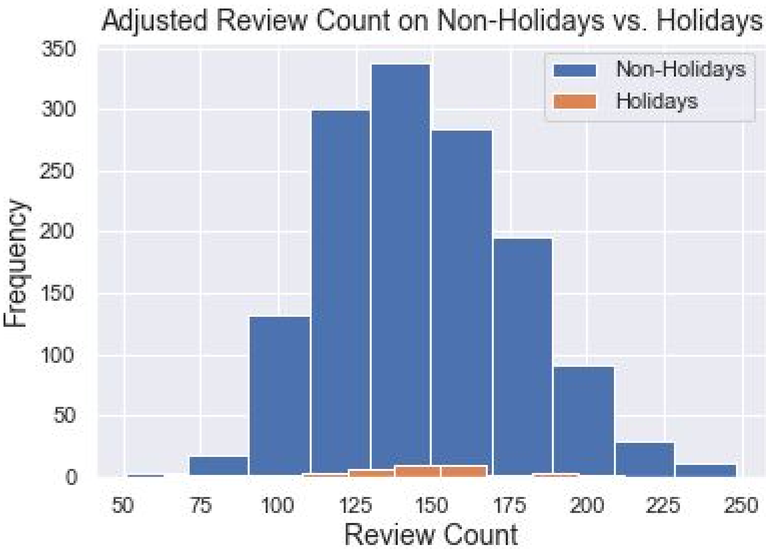
D) Unadjusted Non-Holidays vs. Holidays

Third, we answer the sub-question: are restaurants more popular on non-holidays or holidays? To define a non-holiday and holiday, we manually gathered data from https://www.statutoryholidays.com/. Data from 2016, 2017, 2018, and 2019 was taken. Since we are only conducting these analyses in Ontario, we only take days which are classified as holidays in Ontario. Holidays are not the same day every year, so a list of statutory holidays was made to denote the holidays. To define non-holidays, we simply take the complement of this list. From here, we conduct the same analyses on the variables x3 and y3 representing non-holidays and holidays respectively. First, we conduct normal tests on both datasets. Non-holidays had a normality p-value of 4.2e-09, failing the test, and holidays had a normality p-value of .25, passing the t-test. Once again, when we take a look at the histogram of the distribution of these two datasets, we can see that they appear to be normal enough, so it is still viable to conduct a t-test on this data. The equal-variance test fails with a p-value of 5.2e-05, so we conduct a t-test assuming unequal variance. The p-value resulting from this is 0.40, passing the t-test. We can conclude that there is not any significant difference between the means of the data. The total reviews on non-holidays vs. holidays yielded counts 205,811 and 5,806



respectively, indicating more reviews are submitted during non-holidays. The average reviews per day for non-holidays vs. holidays, yielded 146.5 and 152.8 respectively, indicating no large difference between the two datasets. Observing the histogram, it appears that there is indeed no large difference between the means of the non-holidays and holidays. From all of this data, we can conclude that the popularity of restaurants on non-holidays vs. holidays are not different.

E) Adjusted Non-Holidays vs. Holidays



Fourth, we answer the sub-question: are restaurants more popular on adjusted non-holidays or holidays? We must shift the data by one day to take into account for one day late reviews. We cannot simply shift the data the way we did on the weekends vs. weekday questions as holidays occur on different days each year. To tackle this problem, we imported timedelta from datetime. This method allows us to shift each date in the dataframe by a specified number of days, which is one in this case. For example, it will take a date like '2016-01-01' and convert it to '2016-01-02'. From here, we conduct the same analyses on the variables x4 and y4 representing non-holidays and holidays respectively. First, we conduct normal tests on both datasets. Non-holidays had a normality p-value of 4.4e-08, failing the test, and holidays had a normality p-value of .14, passing the t-test. Once again, when we take a look at the histogram of the distribution of these two datasets, we can see that they appear to be normal enough, so it is still viable to conduct a t-test on this data. The equal-variance test passes with a p-value of .34, so we conduct a t-test assuming equal variance. The p-value resulting from this is 0.65, passing the t-test. We can conclude that there is not any significant difference between the means of the data. The total reviews on non-holidays vs. holidays yielded counts 206,130 and 5,487 respectively, indicating more reviews are submitted during non-holidays. The average reviews per day for non-holidays vs. holidays, yielded 146.7 and 144.4 respectively, indicating no large difference between the two datasets. Observing the histogram, it appears that there is indeed no large difference between the means of the non-holidays and holidays. From all of this data, we can conclude that the popularity of restaurants on non-holidays vs. holidays are not different when taking into account delayed reviews.

F) Limitations, Problems, Potential Changes

It could have been a good idea to include a non-parametric test such as the U-test, as it does not assume anything about the distribution, though we felt the current number of tests performed were adequate. By cutting off reviews earlier than 2016, it can, in a way, be seen as p-hacking since we are changing the shape of the distribution in a way that makes it more normal, though our reasons for doing this are justifiable. When looking at the number of holidays, it is quite a small number. The histogram shows this. As a result, the answer of the total number of reviews on non-holidays vs. holidays is rather obvious. Whether we assumed unequal variance on the t-test for adjusted non-holidays for holidays or not, this did not change the p-value of the test, so we question this parameter's usefulness. It was also difficult to tell if adjusting the values produced any more accurate results as accuracy was difficult to define. Another oversight was that Yelp actually had a checkin.json file which indicates the exact time people would check in at a given restaurant. Using this would have yielded more accurate results for the popularity of restaurants rather than using user reviews. Though using user reviews helps us determine the popularity of submitting Yelp reviews at specific times of the year.

II: Nationality Comparison

A) Introduction

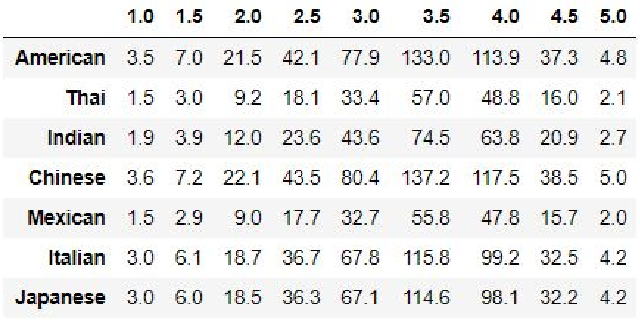
In this section, we address the question: does the nationality of a restaurant have an effect on the ratings it gets? We do this using the business_cleaned.csv file. To answer this question, we chose some of the most popular nationalities and ignored the less popular ones as their data could be easily misrepresented due to a small number of restaurant locations. We chose the following nationalities: American, Thai, Indian, Chinese, Mexican, Italian, and Japanese.

B) Chi-Square Test

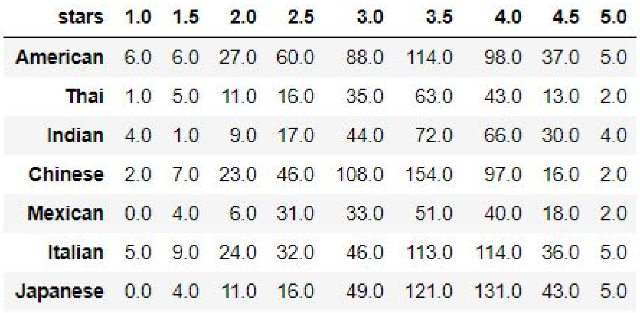
We group each of the stated nationalities by their rating out of five. From here, we made a table where each category was a nationality and each column represented the number of times this rating has been observed in the dataset. By transposing this table, we arrived at the proper form of the contingency table for the chi-squared test. Each category had well over five observations, so we were able to proceed with the test using the contingency table on the right.

The expected table given from calling chi2_contingency is on the right. By inspection, it is clear that the values are quite different from the previous table. The p-value yielded was 1.0e-09, failing the test and allowing us to conclude that the nationality of the restaurant does in fact have an effect on the ratings.

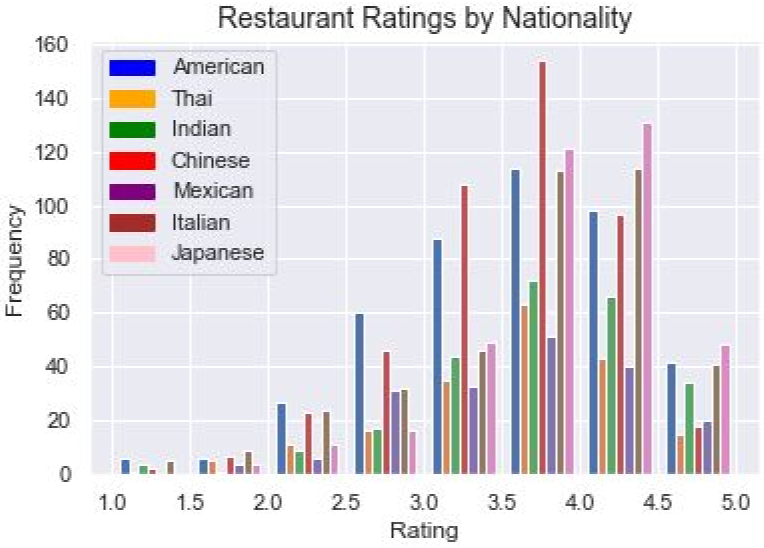
The below histogram displays the frequency of ratings of each restaurant nationality.



	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
American	3.5	7.0	21.5	42.1	77.9	133.0	113.9	37.3	4.8
Thai	1.5	3.0	9.2	18.1	33.4	57.0	48.8	16.0	2.1
Indian	1.9	3.9	12.0	23.6	43.6	74.5	63.8	20.9	2.7
Chinese	3.6	7.2	22.1	43.5	80.4	137.2	117.5	38.5	5.0
Mexican	1.5	2.9	9.0	17.7	32.7	55.8	47.8	15.7	2.0
Italian	3.0	6.1	18.7	36.7	67.8	115.8	99.2	32.5	4.2
Japanese	3.0	6.0	18.5	36.3	67.1	114.6	98.1	32.2	4.2

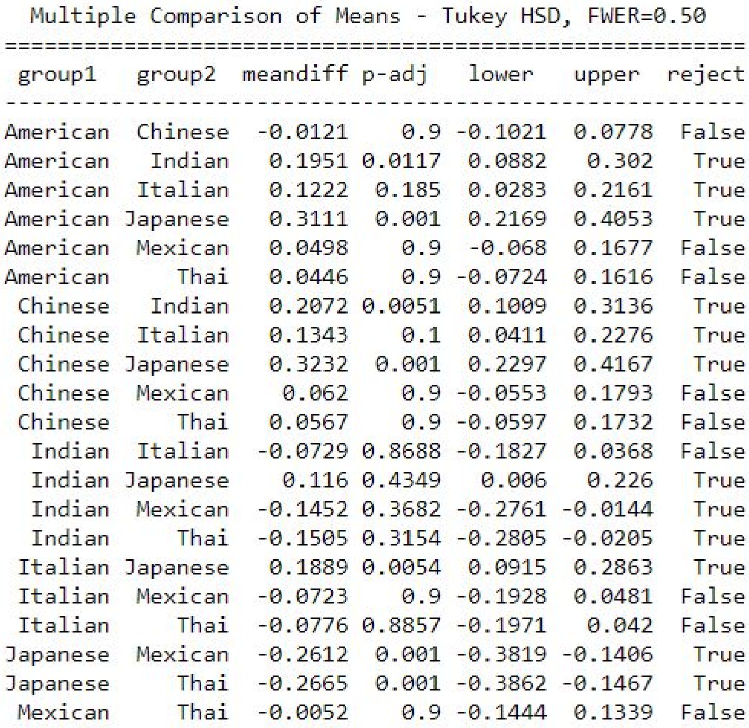


stars	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
American	6.0	6.0	27.0	60.0	88.0	114.0	98.0	37.0	5.0
Thai	1.0	5.0	11.0	16.0	35.0	63.0	43.0	13.0	2.0
Indian	4.0	1.0	9.0	17.0	44.0	72.0	66.0	30.0	4.0
Chinese	2.0	7.0	23.0	46.0	108.0	154.0	97.0	16.0	2.0
Mexican	0.0	4.0	6.0	31.0	33.0	51.0	40.0	18.0	2.0
Italian	5.0	9.0	24.0	32.0	46.0	113.0	114.0	36.0	5.0
Japanese	0.0	4.0	11.0	16.0	49.0	121.0	131.0	43.0	5.0

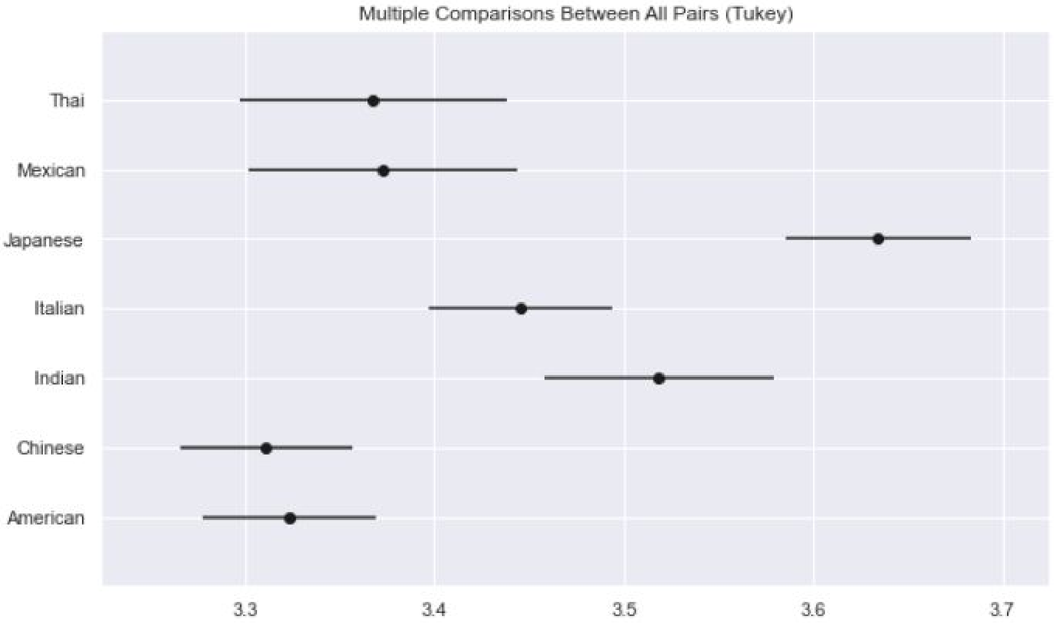


C) ANOVA & Post-Hoc Analysis

By performing ANOVA on the star ratings of each nationality of restaurant, we get a p-value of 1.0e-10, meaning that we can conduct a post-hoc analysis. Concatenating the restaurant nationalities, calling pd.melt, and dropping NaN values allowed us to arrive at the correct format of data for a Pairwise Tukey HSD test. To the right are the results of the test. We can see that there are many cases where we must reject the null hypothesis of the two nationalities having a similar average rating. Below is the graph of the average ratings. We can see that Japanese restaurants have a significantly higher average than other restaurants whereas Chinese and American ones have lower averages.



```
 Multiple Comparison of Means - Tukey HSD, FWER=0.50
=========================================================
 group1   group2  meandiff p-adj   lower    upper  reject
---------------------------------------------------------
American  Chinese  -0.0121    0.9 -0.1021  0.0778  False
American   Indian   0.1951 0.0117  0.0882   0.302   True
American  Italian   0.1222  0.185  0.0283  0.2161   True
American Japanese   0.3111  0.001  0.2169  0.4053   True
American  Mexican   0.0498    0.9  -0.068  0.1677  False
American     Thai   0.0446    0.9 -0.0724  0.1616  False
 Chinese   Indian   0.2072 0.0051  0.1009  0.3136   True
 Chinese  Italian   0.1343    0.1  0.0411  0.2276   True
 Chinese Japanese   0.3232  0.001  0.2297  0.4167   True
 Chinese  Mexican    0.062    0.9 -0.0553  0.1793  False
 Chinese     Thai   0.0567    0.9 -0.0597  0.1732  False
  Indian  Italian  -0.0729 0.8688 -0.1827  0.0368  False
  Indian Japanese    0.116 0.4349   0.006   0.226   True
  Indian  Mexican  -0.1452 0.3682 -0.2761 -0.0144   True
  Indian     Thai  -0.1505 0.3154 -0.2805 -0.0205   True
 Italian Japanese   0.1889 0.0054  0.0915  0.2863   True
 Italian  Mexican  -0.0723    0.9 -0.1928  0.0481  False
 Italian     Thai  -0.0776 0.8857 -0.1971   0.042  False
Japanese  Mexican  -0.2612  0.001 -0.3819 -0.1406   True
Japanese     Thai  -0.2665  0.001 -0.3862 -0.1467   True
 Mexican     Thai  -0.0052    0.9 -0.1444  0.1339  False
```



D) Limitations, Problems, Potential Changes

Performing Tukey's HSD test would have been better on datasets of equal length. Some of the nationalities contain as many as twice the number of rows as other nationalities. The nationalities with a smaller number of locations have a less accurate mean score compared to nationalities with more locations. If we had more time, a good idea would have been to randomly truncate the number of locations to a certain number so that each nationality has an equal number of samples

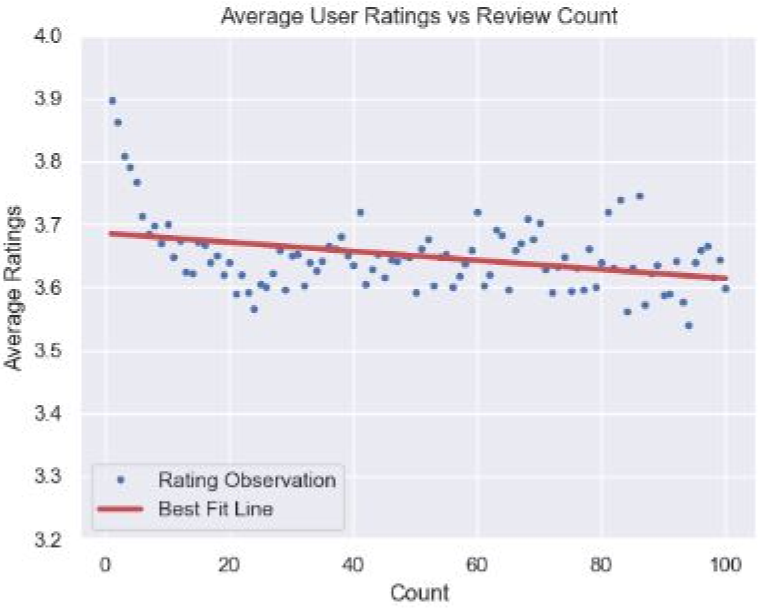
III: User Review Trend

A) Introduction

In this section, we address the question: Is there a relation between the amount of reviews a user does and the score of the review? We do this using the business_reviews_users_merged.csv file. To answer this question, we restrict our analyses to a user's first 100 reviews. Users must have made 100 reviews for this to work. The reason for this is that not many users have a high review count and would not have an impact when the review count becomes high. This causes data to become heavily reliant on the few users who have more reviews. We then chose to work with three groups: elite, non-elite and a combination of the two groups. Note that an elite user is a user verified by Yelp who has made a significant contribution to the community. From here, we gave each user's review a number in increasing order by the date the review was made. Using this, we can take the average of each user's rating by their number and analyse the change in rating as user's make more reviews.

B) Linear Regression

Looking at the linear regression graphs for each group, we noticed two trends that appear whether the user is an elite or non-elite. Firstly, we can see from our regression line that as users make more reviews, their ratings become lower. This can signify many things like users' standards of a restaurant. As a user makes more reviews, they can make more comparisons with previous restaurants they have previously rated and be more critical with the rating they give. Secondly, user ratings start off high followed by a sharp decline in rating

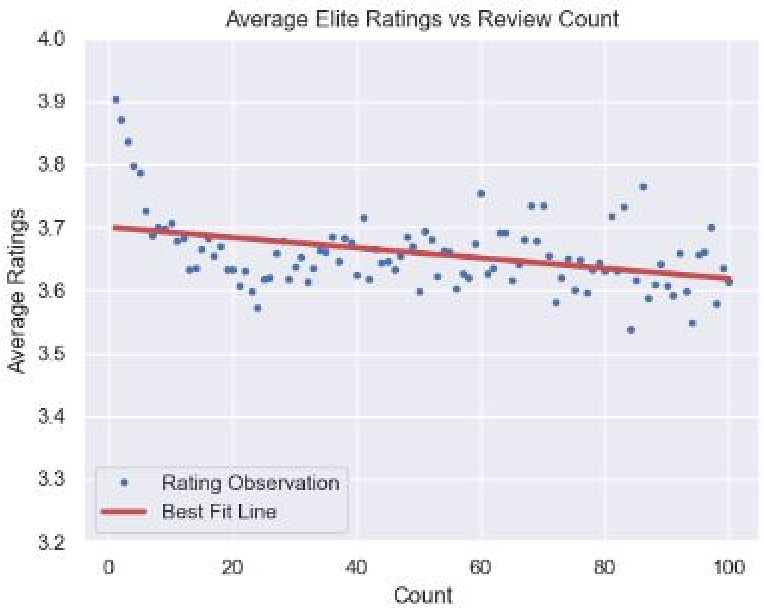


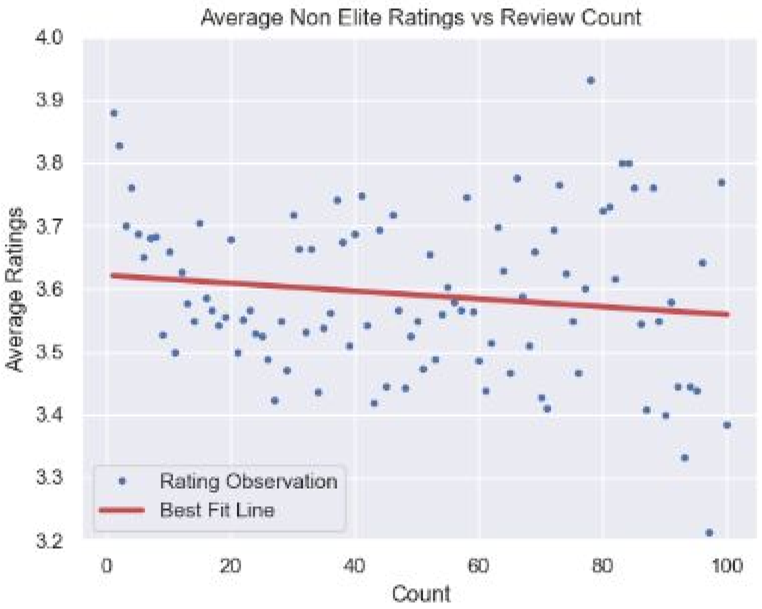
in the first few reviews. Users' standards are quite low when they first give out a rating but with the addition of more reviews made, they can make better, less careless ratings.

Comparing the elite vs non-elite graph we can see a noticeable difference in how scattered the observations are between the two graphs. From the elite graph, we can see that elite users are more consistent with their ratings by the lowly scattered data and close range of rating ranging between 3.6 and 3.7. While the non-elite users give out more bias ratings as can be observed by the highly scattered data observations on the non-elite graph. Non-elite users are more subjective and would give lower ratings as can be observed by the decrease in rating between 3.4 and 3.7 compared to the elite range of 3.6 to 3.7.

C) Limitations, Problems, Potential Changes

It could have been better if we considered users who did not make 100 reviews to see how much impact they would have on our findings. Another problem we came across is the amount of reviews made by elite and non-elite users. There were significantly more reviews made by elite users than non-elite users, giving us a more accurate best fit line for the elite users. If we had more time, we would have done an analysis for potential causes that lead to a users rating like service and pricing.



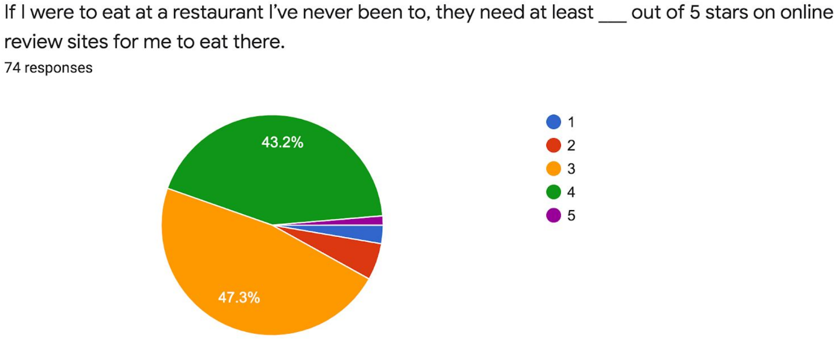


IV: Restaurant Score Prediction

A) Introduction

In this analysis, we wanted to predict whether or not your restaurant is deemed "successful". Now, this is a very open-ended question that can be interpreted many different ways. The way we went about exploring this was on the assumption of whether or not a Yelp user would eat at a restaurant they've never been to or not based on their online reviews. We surveyed 74 people using google forms and posed the question: "If I were to eat at a restaurant I've never been to, they need at least ___ out of 5 stars on online review sites for me to eat there." The results we found were quite interesting.

We found that on average, if your restaurant had a rating of about 3.4, you would likely receive new customers. With this information, we wanted to predict whether your restaurant would receive newer customers easily (i.e. are “successful”) based on information from the dataset. This was a perfect set-up for Naive-Bayes.


If I were to eat at a restaurant I've never been to, they need at least ___ out of 5 stars on online review sites for me to eat there.

74 responses

B) Naive Bayes Classifier

In a similar project predicting restaurant closure (Alifierakis, 2018), this project did quite a bit of feature engineering to get a linear regression model to work. We wanted to take some features of the data as inputs and predict whether or not a restaurant would garner new customers or not. Based on our survey, you could obtain new customers if you had a score of about 3.4 or higher. One last thing to note is that Yelp only scores restaurants on multiples of 0.5 using their own algorithm. While it would be ideal to consider the average rating of the restaurant, what a user would see is the score Yelp gave the restaurant so we will use this to determine whether a restaurant can acquire new customers.. The problem using Naive-Bayes can now be posed as: “Given these features about a business, can we predict whether a restaurant will have 3.5 or more stars?”

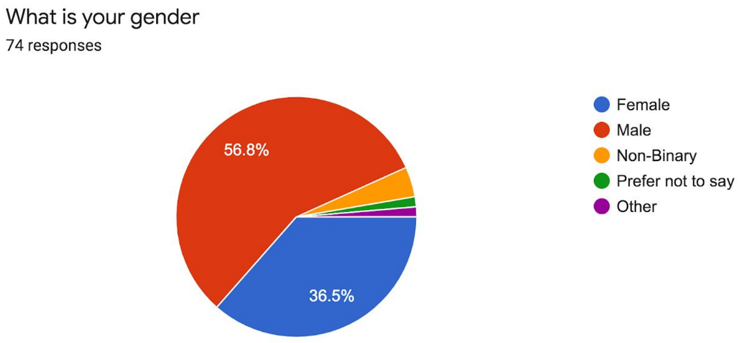
Feature engineering was definitely the most intensive part of building the Naive Bayes model. A lot of the data would have to be one-hot encoded as well as normalize similar strings to mean the same thing. After cleaning the data and extracting even more information out of the attributes column, we end up with up to 45 different possible input observations. Below is a table that showed which combination of input observations would give the best p-score of predicting whether a restaurant could acquire new customers. Note that we ran each one multiple times and took the max.

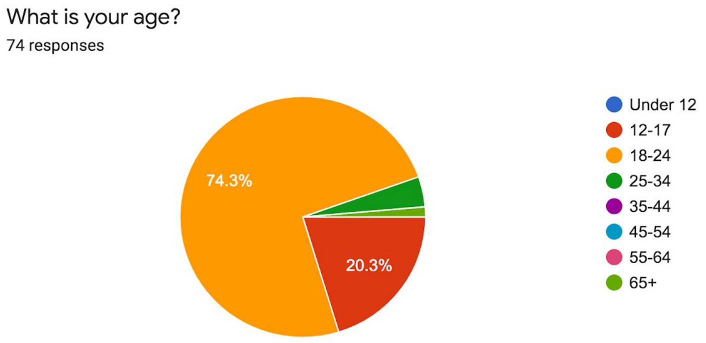
Input Features	**classifier-score**
'Review_count', 'RestaurantsPriceRange'	**0.63**
'Review_count', 'RestaurantsPriceRange', 'RestaurantsAttire'	0.55
'Review_count', 'RestaurantsPriceRange', 'RestaurantsAttire', ‘romantic’, ‘upscale’	0.49
'RestaurantsPriceRange', 'RestaurantsAttire'	0.46
'Review_count', 'RestaurantsAttire'	0.46

As the table shows, 'Review_count' and 'RestaurantsPriceRange' were the minimum number of input features needed to get 65% accuracy on the validation data. Any other features did not add to the model accuracy and leaving any one of those three out would reduce the accuracy. We believe these features make the most sense; a lot of reviews are usually associated with restaurants that are really popular and the price range is usually a big factor in whether or not you liked your food or not. The last input, Restaurant Attire, had two options of casual or dressy, and we believe this likely helped drive the score up because dressy restaurants are usually higher-end and have better food, hence the higher score.

C) Limitations, Problems, Potential Changes

Acknowledging the bias of our survey results is the first thing that comes to mind. Taking a look at who did the surveys, we have that our results heavily favor those in the 18-24 age group who are male.





18-24 year olds as well as males do not represent whether the entire population would deem 3.5 to be the appropriate star rating to consider a new restaurant. A future consideration would be to sample a wider group of people in different age ranges and genders.

Although exploring whether a restaurant would score 3.5 or higher is useful, it certainly is not the only factor in determining whether or not you will get new customers. There were a couple options we could have considered to try to make a more accurate prediction in customer gain.

Photos are typically associated with a review and with that comes the number of people who found that review helpful. We could have considered giving more weight to reviews that included pictures as well as reviews that lots of people rated useful.

Talking to friends who did the survey, they also informed the team that reading recent reviews was another factor in determining whether or not they would like to try this restaurant. Again, we could have considered some other factors like appropriately weighing reviews with more recency higher.

In the future, we would like to include these points in our analysis to see if a Naive Bayes model can give an even better prediction. Another cool piece of information to look into would be whether surrounding businesses affect the model's score or not. We originally considered this given that we have the latitude and longitude of each business (and could calculate the haversine distance) but given the intensity of the feature engineering, have decided to leave out. Finally, what may be even more useful to business owners is what features of their yelp page would give them a certain rating rather than just successful (> 3.5 stars) or not successful.

3 - Project Experience Summary

Rajan

- Cleaned a Yelp businesses dataset so that only restaurants which are currently open in Toronto are included
- Performed statistical analyses on the popularity of restaurants on certain days of the year, in particular, weekends vs. weekdays, and non-holidays vs. holidays
- Conducted categorical & pairwise statistical tests on various restaurant nationalities and observed a large difference amongst nationalities

Brandon

- Performed statistical analyses on the user review trends of elite, non-elite and both groups together
- Cleaned and processed restaurant data to be used in problem analysis
- Frequently communicated with team members and assisted in finding research questions

Rashid

- Created an ETL pipeline that took the raw data and outputted relevant information ready to be analyzed
- Prepared and conducted a survey to use for accuracy in building classification model
- Prepared business data to be fitted into a Naive Bayes classifier to predict if a restaurant would receive a certain star rating

References

- https://towardsdatascience.com/using-yelp-data-to-predict-restaurant-closure-8aafa4f72ad6