Factors of a Yelp Review and Their Influence

By: Dean Papadopoulos, Kripanjali Dhungana, Norman Morris, Zheding Zhao

Introduction:

Food is the way to every human's heart and trying different cuisines is one of our biggest adventures. People prefer to eat outside at least 2 to 3 times a week, however, most of us do not know what kind of cuisine we want to try, or which restaurant do we want to try. Should we stick to our go to restaurant again or try something new? Once we decide on a specific cuisine and a specific restaurant, we want to check out the reviews to see what other people think about the place. These are the people who have already tried the place and the food. Everyone is different and thus seeks different things when trying out a place. Tasty, high-quality food is one thing that is in most cases non-negotiable for everyone, but besides that some people want ambience whereas others want fast service. Some people want to go to a place where they have live music whereas some people want to sit and have a conversation while they dine. It has become a trend these days for people to sit in cafes and get their work done while sipping on their beverage or grabbing a bite.

It is all about preferences and for some it is quality service. From a user's perspective these are the obvious questions that come to mind when choosing a restaurant. However, from a business perspective what are the concerns that may arise? Or just as a fun project what would we want to know about who is writing these reviews or see if there is a certain region where people tend to eat out more. How about where the demographic of a more health-conscious population resides over not so health conscious, or how can restaurants improve their service based on the reviews, or even Yelp given tough competitors such as TripAdvisor or Foursquare. Hence, to explore questions as such we are conducting a marketing analysis specifically on the restaurant reviews on Yelp.

The main goal of this marketing analysis project we conducted is to attempt to answer some questions based on the datasets we chose. We have Yelp's datasets that we combined with the income and education data from the U.S. Census, as well as the unemployment rate. We did an initial exploratory analysis to learn about the datasets and how we can formulate questions we could get answers to. We brainstormed below questions that could be interesting to explore and conduct analysis on.

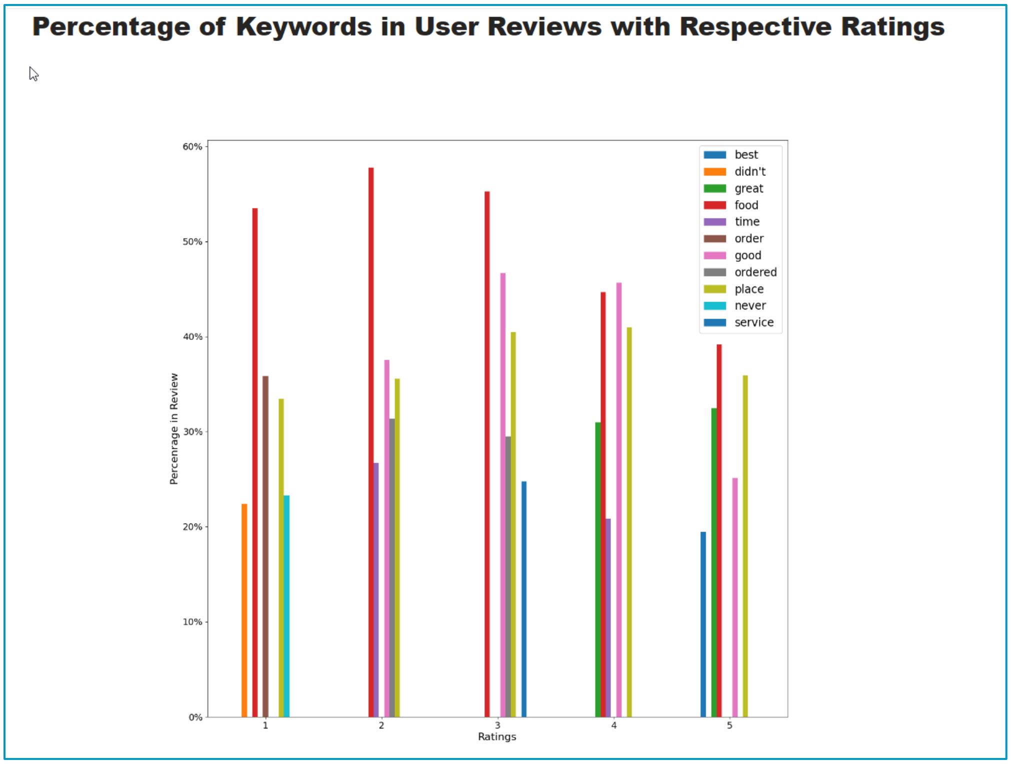
1. What are some keywords for good reviews and bad reviews?
2. Is there a correlation between review score and length of review text?
3. What is the relationship between median income and ratings on expensive restaurants?
4. What is the relationship between median income and number of restaurants?

5. Are people who leave more reviews likely to be more critical than people who leave less reviews?
6. Does lower unemployment result in more favorable reviews on Yelp?
7. Can classification machine learning models be used to predict a business' rating based on its available attributes?

Research:

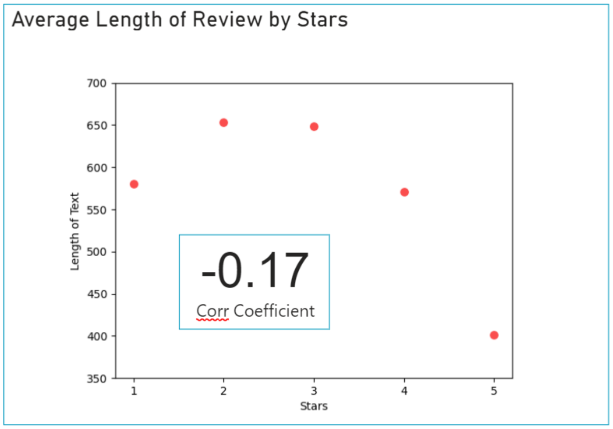
Our next step after formulating the hypothetical questions was to prepare to analyze our data. We started off by using Kafka in Azure Databricks where we created a topic, a producer that produces messages to a pre-made topic, and a consumer that consumes messages from the pre-made topic. We then performed the ETL process on the datasets in Databricks and loaded the data to MS SQL server and Data Lake. After loading the data in SQL server, we started exploring answers to our initial hypothesis. We then exported the data to Power BI and created visualizations that answered our questions.

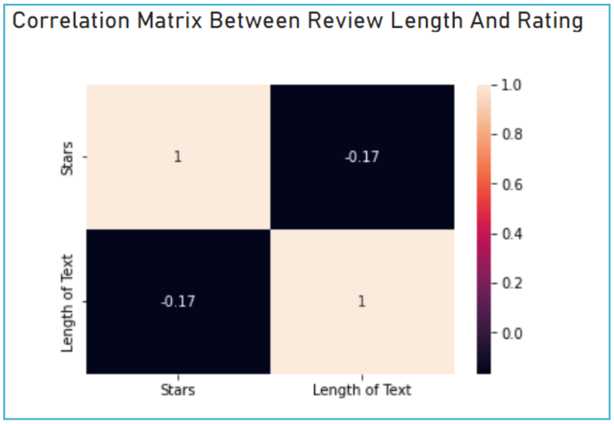
Keywords for reviews



As shown in the above clustered bar chart, there are some common keywords across all levels of rating such as food, place, (order and ordered if grouped as one). In addition, there are keywords and belongs to certain ratings, such as never and didn't for one star and best for five stars, which is reasonable considering the meaning of these words. The word time is also mentioned from two stars and five stars.

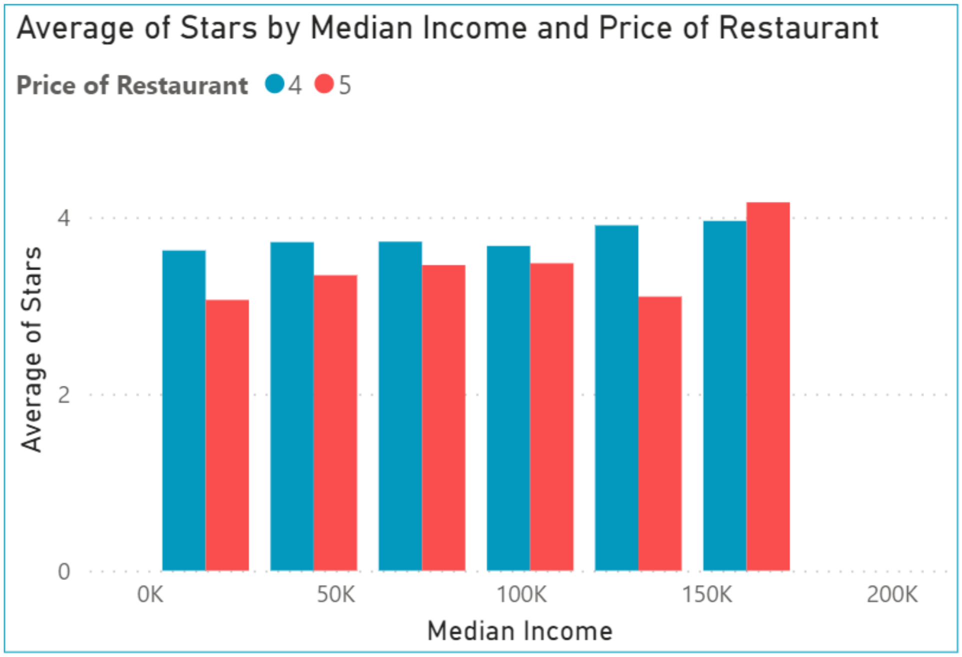
Review score compared to review length


Average Length of Review by Stars


Correlation Matrix Between Review Length And Rating

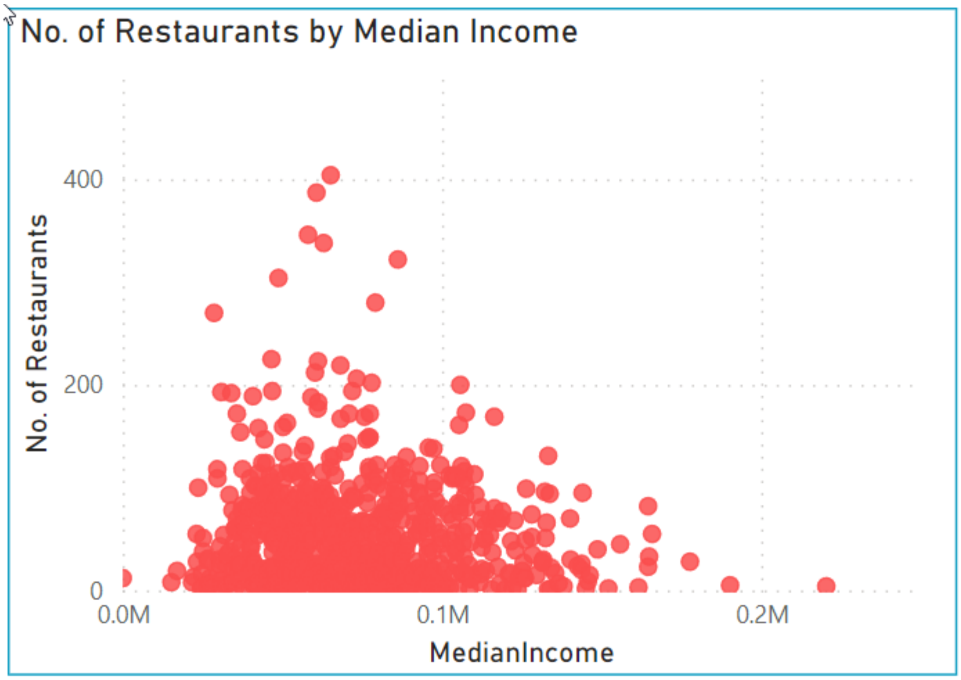
To see if these two fields were correlated a correlation matrix was run which discovered a slight negative correlation of -0.17. This means that longer reviews can sometimes lead to a worse rating. After this a scatter plot was created displaying the average length of characters for each star.

Median income and ratings for expensive restaurants


Average of Stars by Median Income and Price of Restaurant

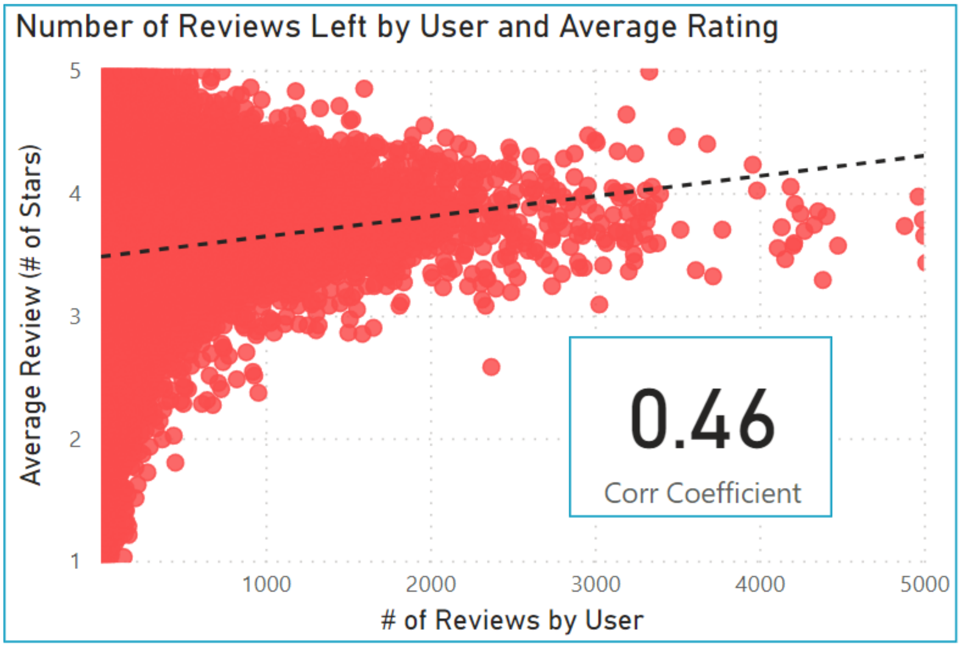
Interested in seeing how the median income of a zip code can impact the average rating of expensive restaurants, the bar chart above was created. The legend displays the price of the restaurants with 5 being the most expensive. This graph isn't extremely informative; however, it does showcase that highest earning zip codes have the highest rated expensive restaurants.

Median income vs number of restaurants



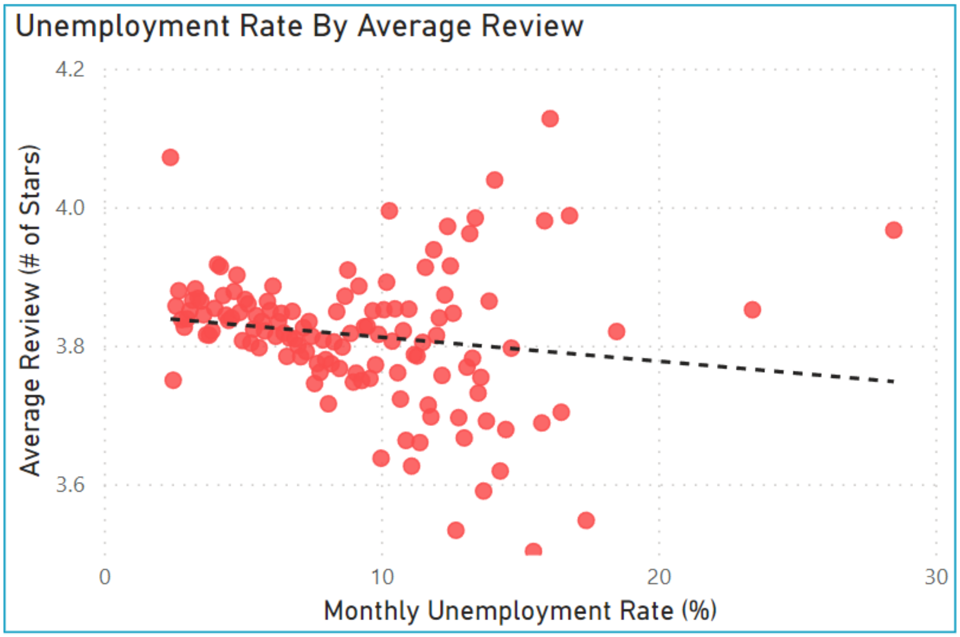
In the above graph, we see the number of restaurants plotted by the median income. Generally, there seem to be more restaurants in business around the area where the income is between 50k-150k as seen in the graph. For the median income of 65k we can see the number of restaurants at over 400 while for a median household income of 220k the number of restaurants is just 4. With this we can tell that where the higher median income is, the number of restaurants isn't as much as lower median income.

User Review Rating vs Frequency



Above we see a moderate positive correlation between the number of reviews and the average review ratings for those users. While not correlated enough to be of predictive value, it was far more predictive than any other criteria we tested, such as the number of compliments given or received for reviews. This is somewhat expected, as many people join Yelp to review a restaurant for a particularly great or poor experience, but the upward trend continues into the hundreds and thousands of reviews per user. We also can see that reviews tend to moderate much more as the number of reviews increases.

Unemployment Rate and Review Values



In the graph above, you see the monthly unemployment rate plotted by the average review value for the states of their businesses. There is little correlation between the two, with a coefficient of about –0.01. However, we do see that the reviews are much more moderate in areas and times that the unemployment rate is lower.

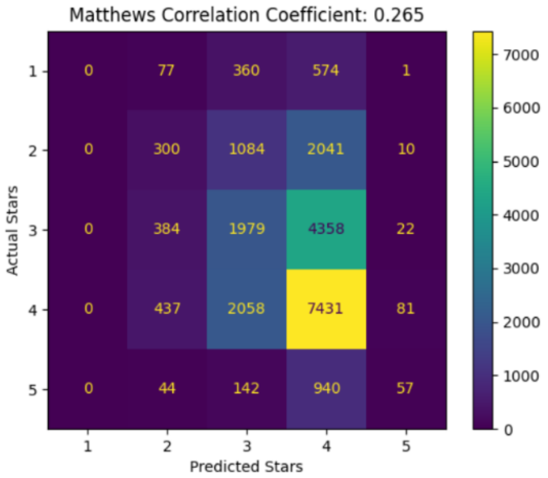
Business Attribute Prediction Model

In order to find out the relationship between a business' rating and its attributes, a machine learning model is used to predict business ratings based on their attributes. After ETL process, the attributes left and selected for machine learning purposes are Alcohol Level, Bike Parking Availability, Credit Card Acceptance, Caters Availability, Kids-Friendliness, TV Availability, Noise Level, Outdoor Seating Availability, Attire Requirements, Delivery Availability, Group-Friendliness, Price Range, Reservation Options, Take Out Options, and Wi-Fi Availability.

During the model selection stage, several classification methods, such as logistic regression and random forest have been considered, but the Extreme Gradient Boost (XGBoost) has been used to make the prediction. XGBoost was selected for two reasons. First, it has an algorithm that handles missing values. During the ETL stage of the project, any attributes that have more than half of value missing are removed. However, there are still missing values in the attributes that are left over, which cannot be simply imputed by the averages or zeroes. Secondly, the XGBoost is good for multi-class classification by its "multi:softmax" objective. Therefore, XGBoost is a good model for the datasets and objectives of this project.

After the model is selected, the next step is to optimize the model. Grid search is used to optimize the model. The parameters tuned through grid search are learning_rate (step size shrinkage used in update to prevents overfitting) max_depth (the maximum depth of a tree), n_estimators (number of trees used), min_child_weight (minimum sum of instance weight (hessian) needed in a child). After the grid search, the optimal model is selected and exported for visualizations.

Confusion Matrix on Prediction of Ratings Based on Business Attributes



The Matthew Correlation Coefficient is 0.265, which is relatively low. It means that for most of the predictions, the predicted stars if not the same as the actual stars, which means that a business' rating cannot be predicted purely based on the business attributes. A possible explanation could be that the data source converted some of the numeric variables (average price, loudness by average dB) to categorical variables (price range, and loudness in categories), which loses some accuracy. Another explanation is that businesses, especially restaurants, have many factors that cannot be quantified easily such as the freshness of food sources and dining environments, but these factors affect ratings of restaurants. In conclusion, the Optimized XGBoost model shows that the A business’ rating on Yelp.com cannot be determined solely by its available attributes.

Conclusion:

Looking back at our initial questions and combining them with our research process, there are a few conclusions we reached. We found some common keywords in the reviews across all levels of ratings, some words were only significant to one star and some to five stars and these words aligned to the levels of stars they received. When we checked the correlation between the ratings and length of the review, we found out that longer reviews could lead to a negative rating. While looking at the median income of a zip code impacting the average rating of expensive restaurants, we found out that the zip codes that have higher median salary had more expensive restaurants. Similarly, we looked at the number of restaurants by median income and concluded that there are more restaurants where the median income is somewhere between 50k-150k. The number of reviews and average ratings for users had a moderate positive correlation as expected since people join Yelp to write a review. We also found that as the number of reviews increased, they also tend to moderate more. We also compared the monthly unemployment rate with the average reviews, and we found a very little correlation between these two, however, we see that the reviews are much more moderate in areas and times where unemployment rate is lower. We also performed a machine learning model where we considered methods like logistics regression and random forest, but we used XGBoost to make the prediction since it could handle missing values and this method well fits for multi-class classification. We used grid search to optimize the model and tune the parameters. Then we used the optimal model to answer our questions through visualizations.

After our initial hypothesis and research, some of our recommendations to the restaurants are that they should focus on their individual advantages instead of adding features that are in business attributes. Besides quality food, service is also another important part of a restaurant’s ratings. User reviews could also be promoted, encouraging reviews from active Yelp users.

Sources:

Yelp. (n.d.). Yelp dataset. Retrieved August 8, 2022, from https://www.yelp.com/dataset/documentation/main

Department of Labor, U. S. (2022, March 14). *Civilian noninstitutional population and associated rate and ratio measures for model-based areas*. U.S. Bureau of Labor Statistics. Retrieved August 8, 2022, from https://www.bls.gov/lau/rdscnp16.htm#data

Bureau, U. S. C. (n.d.). *EDUCATIONAL ATTAINMENT*. Explore census data. Retrieved August 8, 2022, from https://data.census.gov/cedsci/table?q=S1501%3A+EDUCATIONAL+ATTAINMENT&tid=ACSST5Y2020.S1501

Bureau, U. S. C. (n.d.). *INCOME IN THE PAST 12 MONTHS (IN 2020 INFLATION-ADJUSTED DOLLARS)*. Explore census data. Retrieved August 8, 2022, from https://data.census.gov/cedsci/table?q=S1901%3A+INCOME+IN+THE+PAST+12+MONTHS+%28IN+2020+INFLATION-ADJUSTED+DOLLARS%29&g=0100000US%248600000&y=2020&tid=ACSST5Y2020.S1901