PROJECT REPORT

Aryan Jain and Kyle Hwang

COMP_SCI 396: Introduction to the Data Science Pipeline

Prof. Huiling Hu

Introduction and Motivation

Essentially, the project concerns itself with inspecting the quality of food by state and by cuisine across the yelp datasets. This will allow us to analyse which states have the highest rated and quality of food in each respective cuisine as well as which foods rank the highest in each state. Rankings are made based on both, qualitative and quantitative statistics and data analysis. Quantitatively, comparisons between states' respective cuisines will involve the analysis of different measures such as the average, median, quartiles, and the variance of the star ratings. Qualitative comparisons include sentiment analysis of reviews and tips. Through this project, our hope is to aid consumers and businesses in their food-related decisions. For example, consumers would be able to consult with our analysis to determine the best places to eat a meal or the best suited restaurant for them. On the other hand, businesses would be able to identify gaps or shortcomings in the market allowing them to potentially expand their businesses.

Dataset

The project will use the datasets "yelp_academic_dataset_business.json" and "yelp_academic_dataset_tips.json". The business dataset will allow us to come to conclusions comparing statistics such as the mean and the variance between states and cuisines. The tips dataset is used for sentiment analysis, which will allow for a more continuous comparison of the quality of the restaurants and the cuisines in each state, rather than the discrete values of the star ratings given in the business dataset.

Data Cleaning

In order to clean the dataset, we first needed to drop all blank or n/a values from the columns of ‘categories’ and ‘state’. We then further investigated the dataset and noticed that the Yelp Business dataset only contained a sufficient amount of businesses in 9 states: Colorado, Oregon, Florida, Georgia, British Columbia (Canada), Ohio, Texas, Massachusetts, and Washington. The other states included in the dataset only contained 1-20 total businesses which we concluded was an insignificant and insufficient amount of data to analyze and compare with other states. Thus, we filtered the dataset to only include businesses from these 9 states. Next, we needed to create a filter such that only restaurants and places that served food were considered in our project and pipeline. Through some manual investigation we noticed that all places that tend to serve food have certain keywords under their categories column, which has the datatype, String. Thus, in order to filter out all the non-restaurant businesses we used the pandas library and searched through each record’s ‘categories’ string to identify specific substrings that would classify them as a food-serving restaurant. The key substrings are: ‘Restaurants’, ‘Pizza’, ‘American’, ‘Chinese’, ‘Indian’, ‘Italian’, ‘Korean’, and ‘Thai’. Through the find function we were able to return indexes of -1, if the substring was not found, and then were able to create a column in the DataFrame for each substring before then adding in values of a record’s find function in its respective columns. This allowed us to condition particular columns to not equal -1 which then left us with a DataFrame of a specific category. These were the steps we took to clean and manage our dataset to leave us with a complete dataset of solely food-related businesses in states with a sufficient number of restaurants.

Exploratory Data Analysis

After we cleaned and filtered our dataset, we were able to perform several means of Exploratory Data Analysis. We noticed that the category of 'Pizza' was quite omnipresent in the Yelp dataset and thus we first decided to note which state had the best Pizza places.

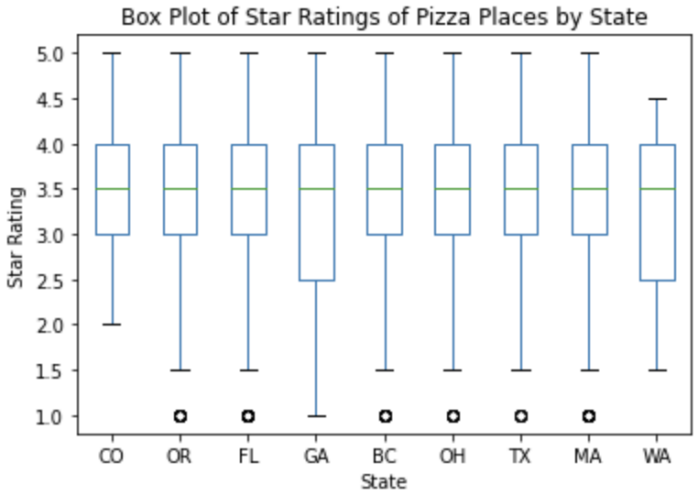

Figure 1: Side-by-Side Boxplot of Pizza Ratings by State

Figure 1 above shows the Side-by-Side Boxplot of Pizza Ratings by State. We were quite surprised with the results as the median and the upper quartiles were consistent throughout each of the states with each state's median pizza place rating at 3.5 stars in the Yelp dataset. Further, each state's 75th percentile pizza place is rated at 4.0 stars. The 25th percentile in each state varied from rating as high as 3.0 stars down to 2.5 stars in Georgia. Lastly, the lower end of each state's range concluded around 2.0 stars with only one state, Georgia, having enough pizza chains rated as poorly as 1.0 to consider the rating a non-outlier.

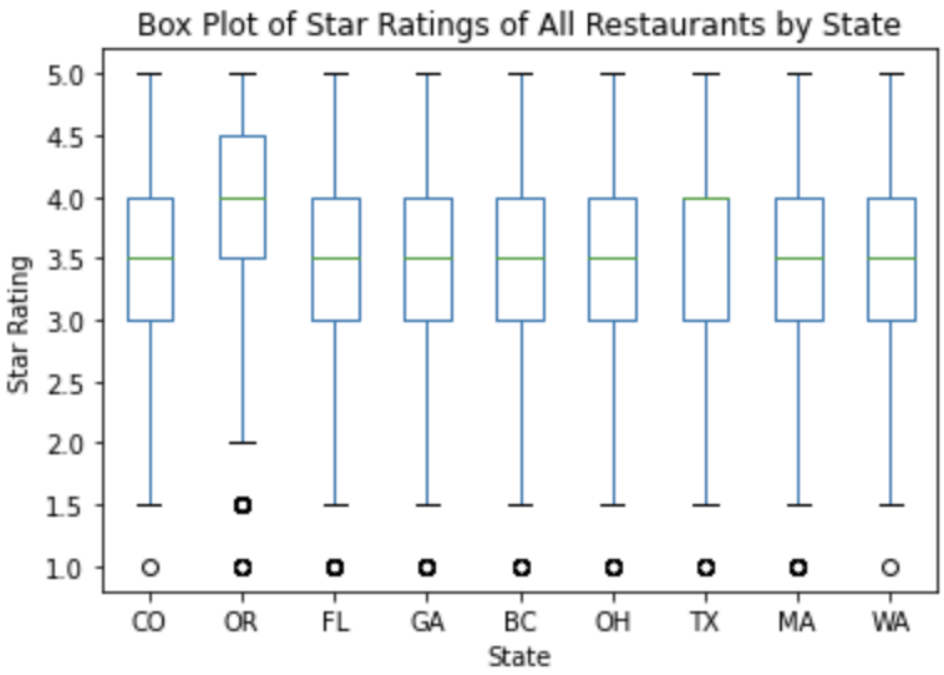

Figure 2: Side-by-Side Boxplot of Restaurant Ratings by State

We then aimed to perform the same analysis on all restaurant ratings in each of the 9 states and compare them. The results of this analysis are shown in Figure 2 in a side-by-side boxplot. Immediately, we noticed that Oregon and Texas have the highest median ratings of restaurants out of all the states with ratings of 4.0. However, with a higher 25th and 75th percentile ratings, Oregon seems to have the highest rated restaurants of all 9 states. This allows consumers to eat freely and worry less about choosing the right place to eat in Oregon in order to achieve the satisfaction of a good meal. We further noticed that other than Oregon and Texas, similar to the pizza places, the states are extremely similar with average ratings of 3.5 stars for restaurants. This analysis also allowed us to see that pizza places are considered average in each of the states as they have the same median ratings as other cuisines in the state as well.

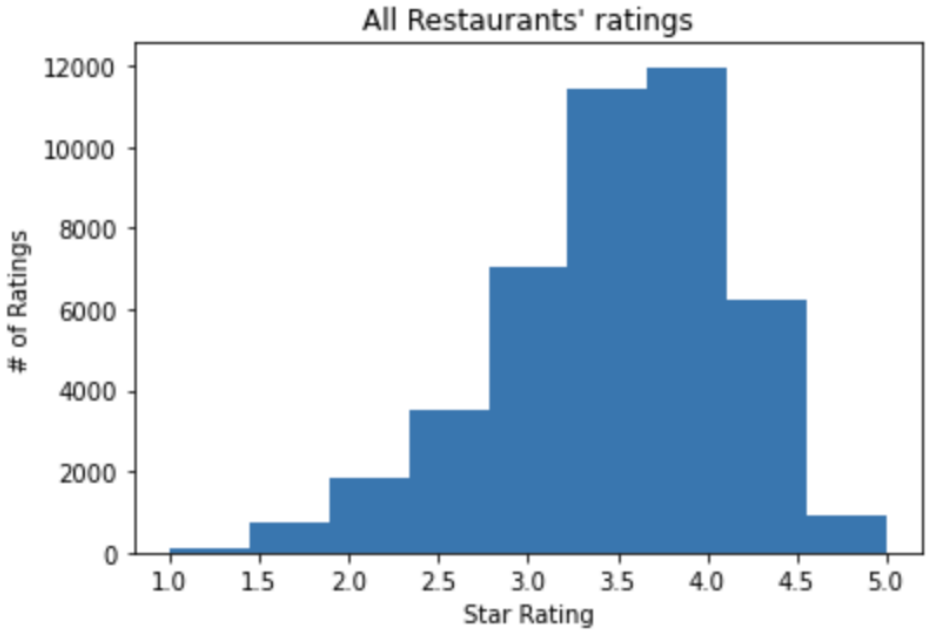

Figure 3: Bar Histogram of Restaurant Ratings in all states

Figure 3 shows the distribution of restaurant ratings in all states. The figure shows a slightly left skewed distribution of ratings of all restaurants with a mode of a rating of 4.0 stars. This follows closely to the boxplots depicted in figures 1 and 2 where the upper tail of the interquartile range was also at 4.0. Further, the distribution illustrates a mean and a median around 3.5 stars which also mirrors the boxplots above.

We then wanted to explore how each cuisine ranks and rates individually in each of the states and thus we created 9 box plots depicting each cuisine's ratings in their respective state.

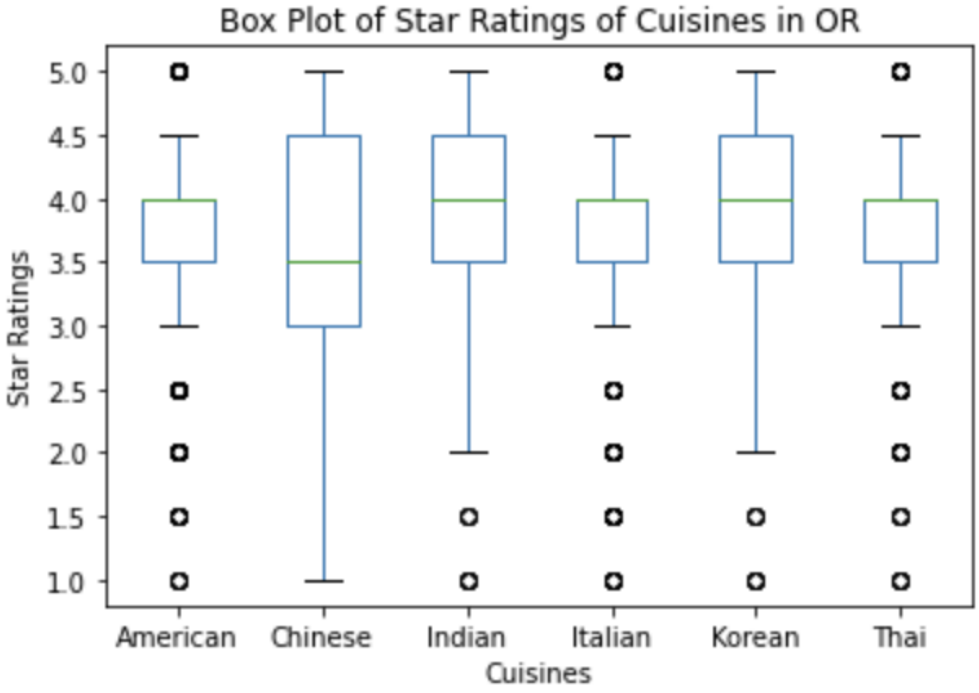

Figure 4: Side-by-Side boxplot of each cuisine's ratings in the state of Oregon

Figure 4 shows one of the boxplots and depicts the breakdown of each cuisine's ratings in the state of Oregon. From the analysis, we can identify that Chinese Food has the lowest median of all other cuisines in the states with a median rating of 3.5, lower than that of all other cuisines (4.0). Further, we see that Indian and Korean Food have the highest interquartile rating ranging from 4.5 down to 3.5 stars showing that Consumers new to Oregon should choose to eat Indian or Korean if choosing a restaurant at random in order to obtain the highest chance of satisfaction. By performing the same analysis for all states, which can be seen in our code and HTML, we were able to come to similar conclusions as above for the other states as well.

So far, these values that we analyzed involve discrete data points, specifically, the stars, since the Yelp dataset provides them at 0.5 intervals. So, we decided to analyze the tips dataset to get a more continuous value on how the customers feel about the restaurants.

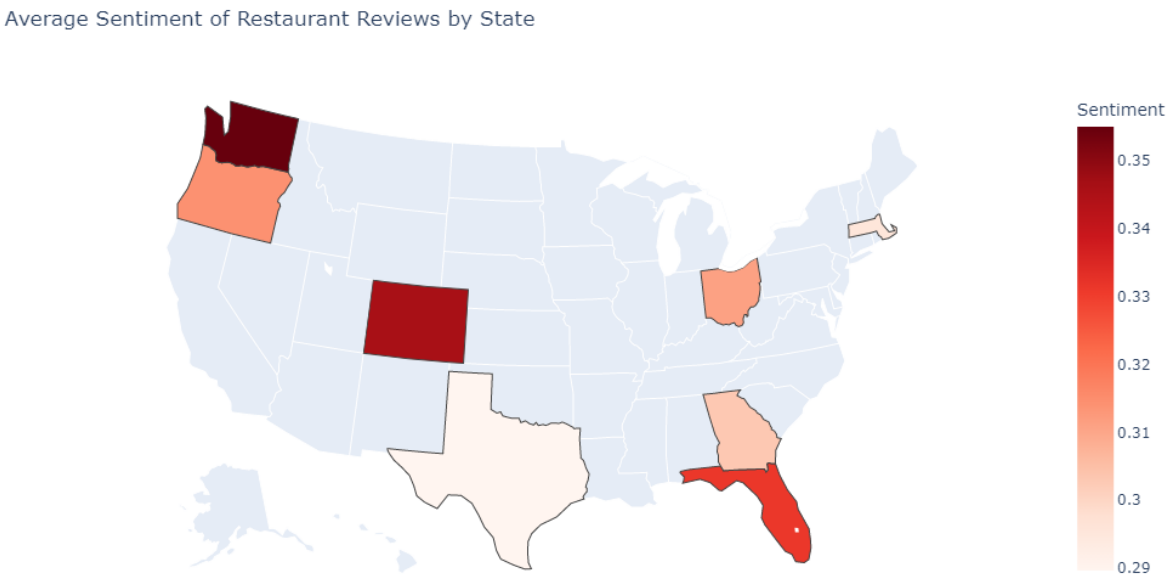

Figure 5: U.S. Map Plot of Sentiment Values in All Restaurants

Figure 5 shows the sentiment analysis of the tips from the tip dataset of all restaurants in the specified states. Upon observation, it can be seen that Texas and Massachusetts have the lowest sentiment whereas Washington has the highest. But overall, we saw that the sentiment across these states were relatively consistent, seeing as how the range of the sentiment goes from 0.29 to 0.35 when the sentiment ranges from -1.0 to +1.0.

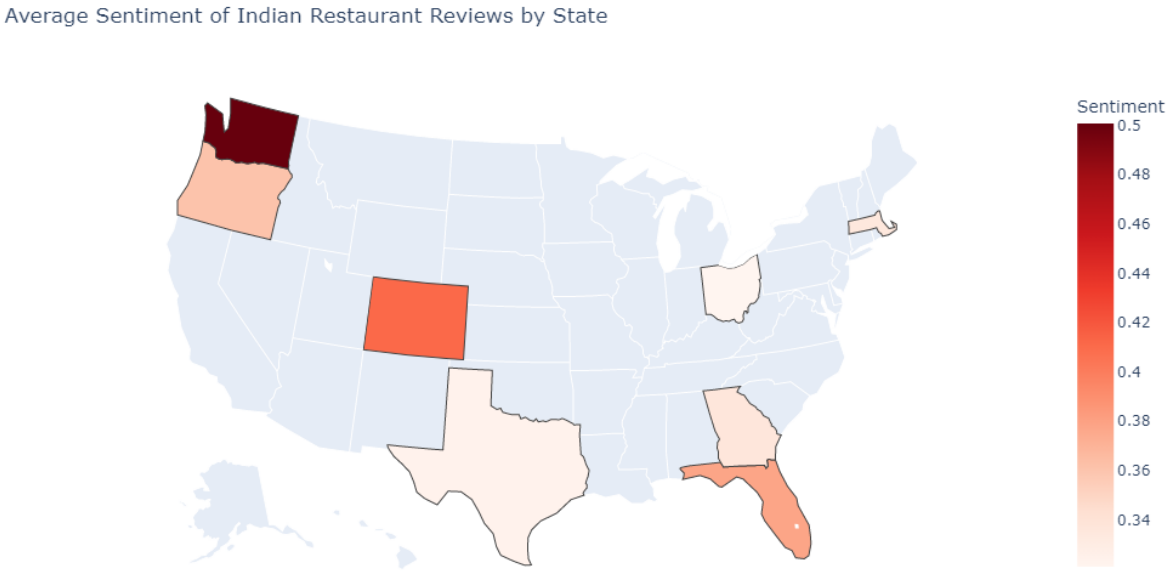

Figure 6: U.S. Map Plot of Sentiment Values in Indian Restaurants

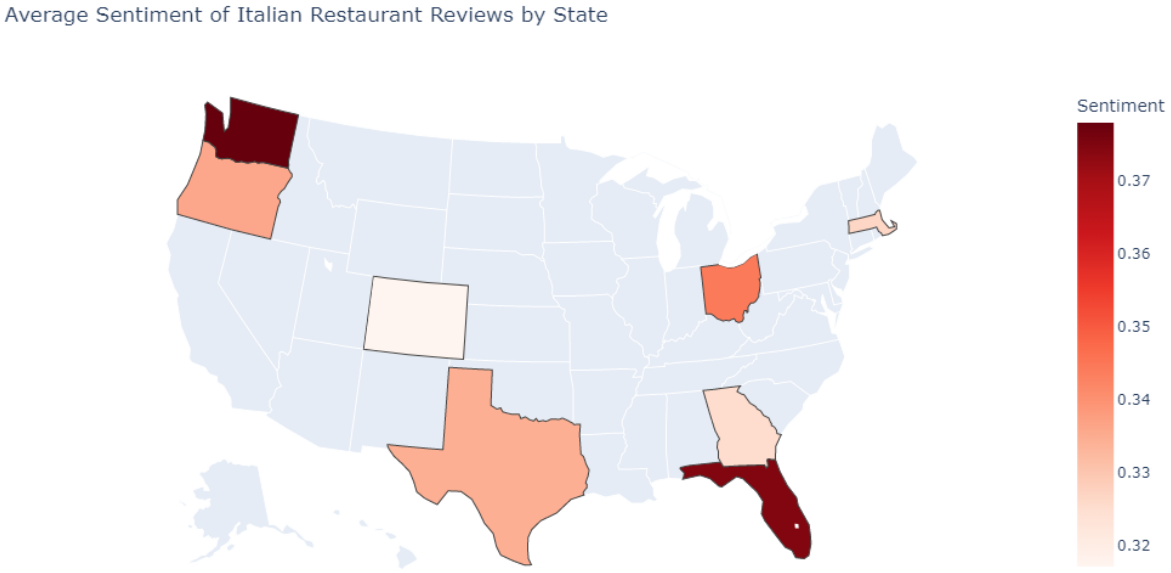

Figure 7: U.S. Map Plot of Sentiment Values in Italian Restaurants

We then wanted to look at individual cuisines in these particular regions. Figures 6 and 7 show two particular cuisines, namely Indian and Italian. Out of all of the cuisines analyzed,

Indian restaurants had the highest variance in their sentiment values, whereas Italian restaurants had the lowest. Something that can also be observed is that the patterns of the sentiment in all restaurants generally follow in the cuisines, like how Washington has the highest sentiment values.

Summary of Findings

In our box plots, we found that overall, the quartiles and the interquartile range were pretty similar considering all restaurants. Our histogram shows that the ratings were left-skewed and the mode of the ratings was 4.0 stars, which is consistent with our findings in the box plot of all restaurants. When considering each region we analyzed with each cuisine, the ratings had more outlier points. In our example of Oregon, Chinese food had the lowest median of 3.5 whereas the rest had 4.0, and Korean had the highest upper quartile, leading us to believe that Korean food would have the highest chance of being good. In our sentiment analysis, we gained an insight on how different states feel about different cuisines. Overall, we observed that Texas and Massachusetts seem to have the lowest sentiment value whereas Washington has the highest. This mostly carries over when observing each of the 6 cuisines we analyzed. But considering all restaurants, the variance of the sentiment values is relatively small. Also, we noticed that the average sentiment values are all slightly positive. The cuisine with the largest range in sentiment values was Indian, ranging from around 0.34 to 0.5, whereas the smallest range came from Italian, ranging from around 0.32 to 0.37.

Potential Implications and Improvements

Upon completion of the project, the primary stakeholder of the study would be consumers with the anticipated outcome mostly helping restaurant goers decide where to eat when choosing a restaurant at random. However, secondarily, the study could also help restaurant owners and stakeholders. By being able to compare states by cuisine but also cuisines within states, consumers, especially those new to their state or location, would be able to refer to the study to determine which cuisine is considered “a safe choice” in their state. For example, according to our sentiment analysis, among all the cuisines, it seems that Italian food has the least variance, which would make for a safer option, as opposed to Indian, which has the highest variance. Further, consumers who enjoy food, such as foodies, would be able to plan vacations or trips around the collated ratings as found by the study. For example, Washington had the highest sentiment value for Indian food, and Indian food had the highest variance, so an Indian food enthusiast could plan a trip to Washington for their Indian food.

On the other hand, this study could also help businesses recognize underperforming cuisines in areas and therefore, address the demand by expanding into the area. For example, a successful Thai chain could realize that Texas underperforms in the quality of their food and that opening a branch in Texas could be a successful endeavor as the quality of food in the surrounding area is subpar. Similarly, an Indian franchise in New Jersey looking to expand their menu out of the niche of Indian food could realize that Chinese food in the state is the lowest ranked and may choose to undergo menu changes to accommodate Chinese dishes.

As for limitations we faced, and improvements we would like to make should we continue working on this project, we have identified three main ideas. The first limitation we faced is the Yelp Dataset incompleteness. The Yelp Datasets only contain sufficient information

on business in 8 states with other states only listing 1-20 total businesses on Yelp. This incompleteness led to many gaps in our analysis and research such as not being able to identify key states' ratings.

Secondly, the datasets also contain many businesses with incomplete information such as null values in the categories column. This led to our data pipeline filtering out many businesses that, if we had the time we would've manually inspected and consulted an expert on to determine the nature of the businesses, were filtered out. As said in our data cleaning section, many of the restaurants had null values in areas that we needed. However, in most cases, we would be able to look up the restaurant with null values and fill in the columns manually. This way, we would have more data to work with and get a more accurate result in our analysis.

Lastly, our data pipeline only included the analysis of 6 cuisines. The 6 cuisines we handpicked seemed to be the most occurring in the yelp datasets, however, there are many other cuisines of food that were not represented in our analysis. Further, through association analysis, we noticed that the 6 cuisines we picked overlapped with many other cuisines that we did not include. However, had we been able to, we would have analyzed each cuisine possible and present in the datasets.