Favourable Cuisines by Selected Cities

# Introduction

When an entrepreneur is considering opening a new restaurant, one of the first decisions to be made is to select the cuisine. The success of the restaurant relies on ensuring there is a market given the cuisine. **My research question is given a metropolitan location what are the most favouable cuisines.**

To answer this question, I will use data from the Yelp Open Dataset which provides over 6 million reviews across 10 metropolitan areas of various businesses. The Yelp review dataset has many attributes that permit the identification of restaurants as well as their cuisine of focus. By using text sentiment analysis of the reviews this project will determine the cuisine preferences of selected metropolitan areas. This analysis will focus on text classification and sentiment analysis methodologies. R will be used as the primary analytical tool.

As well, I will look to see if there is a positively correlated relationship between positive sentiment scores to higher star ratings provided in the reviews. Since a higher number of stars indicates the reviewer is more positive about the restaurant, if will be a good indicator of whether or not the sentiment scores are of value.

# Literature Review

Although a literature review did not produce any published analytics of determining cuisine preferences by location, there were a few articles that provide guidance in the analytical approach of this analysis.

Although the Yelp review dataset contains a “number of stars rating”, it cannot reliably be used to determine the sentiment of the reviewer. Yu-Han Chen and John Merrick in their 2017 article titled “**Real time Yelp reviews analysis and response solutions for restaurant owners”[[1]](#endnote-1)[[2]](#footnote-1)** found that reviewers have different standards and a 3-star reviewer’s sentiment maybe negative or positive. They tried to normalize the reviewers’ ratings, but this failed, and they concluded the star ratings may not be a reliable indicator of sentiment.

# Dataset

This analysis will focus on two datasets: “business.json” and “review.json”. The “business.json” will provide the attributes to determine if the business is a restaurant, the location of the restaurant and the cuisine. Specific attributes that will be used are the “Business ID”, “City”, “Business Attributes” (to determine if the business is a restaurant) and “Categories” (to determine the cuisine). The “review.json” will provide the data required to classify the sentiment. Specific attributes that will be used are “Business ID” and “Text” (which contains the data for the sentiment of the review). The datasets were obtained from the following URL:

<https://www.yelp.com/dataset>

# Approach

**Step 1: Download and extract both datasets from the “tar” file.**

**Step 3: Pick the top 10 cuisines in the dataset and create a subset by filtering for those cuisines in the business.json dataset.**

**Step 8: By city determine the most favourable cuisines based on average sentiment score.**

**Step 4: Pick the top 10 cities in the dataset and create a subset by filtering for those cities in the business.json dataset.**

**Step 7: Compare the means of the sentiment scores of each star rating to determine if the means are different and positively correlate with the star ratings.**

**Step 2: Load the dataset into two dataframes and select restaurant businesses.**

**Step 5: Merge the reviews.json dataset with the filtered business.json dataset based on the business\_id of the filtered business.json.**

**Step 6: Apply NLP techniques to the “text” attribute to extract keywords and create a field to capture the sentiment score.**

**Step 1: Download and extract both datasets from the “tar” file.**

Go to the <https://www.yelp.com/dataset> website and click on “Download Dataset”. In the next webpage, fill in the required information and the dataset will be downloaded into your “Download” directory as a “tar” file. Right click on the “tar” file and extract into the directory of where the data will be stored. The extracted files will be of type “JSON”.

**Step 2: Load the dataset into two dataframes and select restaurant businesses.**

In R load “yelp\_academic\_dataset\_business.json” and “yelp\_academic\_dataset\_review.json” using the “readr” library and the function “read\_lines”. For each file the lines are then converted into a single JSON string and then parsed and converted to a tribble dataframes. The business category field is cleaned (removing spaces) and a word frequency is executed to find the best word to filter just restaurants. It is determined the word “Restaurant” is the best to filter. Using the “dplyr” library, he business dataset is filter to include only “Restaurant” in the category field.

**Step 3: Pick the top 10 cuisines in the dataset and create a subset by filtering for those cuisines in the business.json dataset.**

Using the same methodology, the word frequency list is scanned to determine the top ten cuisines. Using the “dplyr” library, execute the “filter” function to create a subset dataframe of just records that have the top ten cuisines. A new field called “cuisine” is created to label each record’s cuisine.

**Step 4: Pick the top 10 cities in the dataset and create a subset by filtering for those cities in the business.json dataset.**

Once again a word frequency count is used, this time on the field “city”, to determine the top 10 cities. The dataframe is then filtered to include only cities in the top 10. Some clean up is required to consolidate the city names.

**Step 5: Merge the reviews.json dataset with the filtered business.json dataset based on the business\_id of the filtered business.json.**

Merge the dataframes together by using the “match” function in r. “business.json” and “review.json” will be merged using a left join with the attribute “Business ID”. Columns “useful”, “funny”, “cool” and “date” are removed as they are not needed for the analysis. Any analysis of number of records by city and cuisine is produced to ensure sample sizes are sufficient for the analysis.

**Step 6: Apply NLP techniques to the “text” attribute to extract keywords and create a field to capture the sentiment score.**

The sentiment analysis will use the tidytext framework; we need to use the “unnest\_tokens” function. This will create a record for each word in the review comments field. We will use the AFINN lexicon for our sentiment analysis. It provides a list of negative and # positive works and assigns a value ranging from -5 to +5. The lower the score the more negative # the sentiment of the word. Using an inner\_join by using the word field in AFINN, we then group\_by review\_id to build a sentiment score of the review based on the mean AFINN value of each review

**Step 7: Compare the means of the sentiment scores of each star rating to determine if the means are different and positively correlate with the star ratings.**

A box plot is created to check to visually inspect if there is a relationship between mean sentiment value and the numbers of stars.

To determine if the means of each star category is statistically significant, we perform an ANOVA test since we have more than two categories to compare to each other. Before we perform the test we must confirm all ANOVA assumptions have been met. If they are not all met, we will need to use a non parametric test such as Kruskal Wallis. An post hoc analysis (Dunn Test) is undertaken to determine which means are different and to conclude if the sentiment values are positively correlated with the star ratings.

**Step 8: By city determine the most favourable cuisines based on average sentiment score.**

Similar to step 7 we determine if the ANOVA test can undertaken between sentiment and cuisine. If any of the assumptions are not met we will need to use a non parametric test such as Kruskal Wallis. An post hoc analysis (Dunn Test) is undertaken to determine which means are different and to conclude what are the cuisine preferences by city. Lastly we will answer our research question.

# Results

Based on the dataset I have determined the cities and cuisines to explore. They are shown in the following table as well as the number of reviews for each to illustrate their sample sizes.

Figure : Count of Reviews



From above it is seen that Canadian(new) is not a category in American cities. The lowest count of reviews is Sushibars in Mesa, at 79 there are still enough and all others our over 100.

We then use r to undertake a sentiment analysis of reviews which assigns a range of -5 to 5 (5 is most favourable) to reach review. To determine if the sentiment analysis is valid, we create a boxplot of the sentiment scores by each star category and find that there is a positive correlation between the two variables.

Figure : Boxplot of Average Sentiment Score by Star Rating

A screenshot of a cell phone

Description automatically generated

Based on my analysis it can be concluded that the average sentiment scores are significantly different and are positively correlated with the number of stars ratings. Therefore it is concluded that the sentiment analysis is valid measure of preference.

The next task is to use sentiment score to determine which cuisines are favoured by city. The figure below presents the average sentiment score by city.

Figure : Average Sentiment Score by City and Cuisine



Note, all means are statistically different from each other. Some observations of note:

1. Both Canadian cities had Canadian(new) as their favourite cuisine.
2. The most overall favourite cuisine is Mediterranean with only one city (Scottsdale) that didn’t rank in the top 3.
3. There are two cuisines that do not rank as the top 3 of any city, there are Mexican and Chinese. For these cuisines it is important that an entrepreneur that wants to open these types of cuisine undertake additional micro analysis of specific areas of the cities to find localized interest in these cuisines.

# Conclusions

I have shown that we can create a model that will identify favourite cuisines by city using Yelp data. This is very useful in that anyone wanting to open an restaurant will know which cuisines review favourable reviews by city. Favourable reviews will most likely result in additional business thus creating a good environment for a successful business.

1. [↑](#endnote-ref-1)
2. https://nycdatascience.com/blog/student-works/real-time-yelp-reviews-analysis-response-solutions-restaurant-owners/ [↑](#footnote-ref-1)