**1. Overall Project Objectives**

Yelp is an application to provide the platform for customers to write reviews and provide a star-rating. A research indicates that a one-star increase led to 59% increase in revenue of independent restaurants. Therefore, we see great potential of Yelp dataset as a valuable insights repository.

The main purpose of our project is to conduct thorough analysis on 7 different cuisine types of restaurants which are Korean, Japanese, Chinese, Vietnamese,Thai, French and Italian, figure out what makes a good restaurant and what concerns customers, and then make recommendations of the future improvement and profit growth. Specifically, we will mainly analyze customers' reviews and figure out reasons why customers love or dislike the restaurant. For example, there may be great reviews primarily due to the friendly service, or negative reviews about high price. Meanwhile, we will also compare among those 7 different cuisine types and figure out differences from reviews and gain valuable insights to make customized recommendations to different types of restaurants.

**2. Description of Data**

The Yelp dataset is downloaded from Kaggle website. In total, there are 5,200,000 user reviews, information on 174,000 business. we will focus on two tables which are business table and review table. Attributes of business table are as following:

* business\_id: ID of the business
* name: name of the business
* neighborhood
* address: address of the business
* city: city of the business
* state: state of the business
* postal\_code: postal code of the business
* latitude: latitude of the business
* longitude: longitude of the business
* stars: average rating of the business
* review\_count: number of reviews received
* is\_open: 1 if the business is open, 0 therwise
* categories: multiple categories of the business

Attribues of review table are as following:

* review\_id: ID of the review
* user\_id: ID of the user
* business\_id: ID of the business
* stars: ratings of the business
* date: review date
* text: review from the user
* useful: number of users who vote a review as usefull
* funny: number of users who vote a review as funny
* cool: number of users who vote a review as cool

**3. Direction of Analysis**

**Exploratory data analysis**

* Count the number of each cuisine type of restaurants
* Count the number of reviews in each cuisine type of restaurants
* Visualize the distribution of restaurants according to the ratings and cuisine types of restaurants.

**Review Analysis**

* Clean the category column in business table into different cuisine types of restaurants and find out how the reviews can help this type of restaurants improve in the future.
* Refer the business id in business table to review table, collect all the reviews of this type of restaurants and perform sentiment analysis to analyze frequent words in positive and negative reviews.
* Implement SVM model to get relatively positive and negative words and get score of each word.
* Get top 10 positive words and negative words in each cuisine type of restaurants in order to figure out the reason of high score and low score.
* Compare among different types of restaurants to figure out the advantages and disadvantages, then we can generate a series of recommendations to this type of restaurants for the future development.
* Overall, recommendations may have different topics which is but not limited to service, food, or decoration, etc. Our analysis is generally based on review words which we can tell from such as rude, overpriced, and slow to find out which aspect in this type of restaurants that they could improve.

**4. Summary of Progress**

**Selection and Filtering**

* Filter out 50 states of US.
* Filter out all restaurants of US.

**Cleaning**

* Categorize all restaurants by cuisine type using the matching keywords.
* Delete all records with null category.
* Remove quotation marks of name and address columns.
* Label restaurants above rating of 4 as positive, below rating 3 label as negative, label rating 3 as neural.
* Drop rows with neural label.
* Apply 'bag of words': the frequencies of various words appeared in each review as features and conduct SVM model to get score of each word.

**Reshaping and Reindexing**

* Reindex the data frame.
* Build a new column to input the new category name and delete the previous column.
* Convert array to dataframe.

**Visualization**

* Visualize the distribution of restaurants and reviews by category using seaborn.
* Visualize the distribution of restaurants and reviews by rating.
* Visualize top 10 negative words and positive words in each cuisine type.

**Manipulation**

* Get a 'polarity score' (a value that reflects the polarity of sentiment) towards each restaurant category, the sentiment score of each word was first multiplied by its frequency, and then normalized by the total number of reviews for the specific category of restaurants.

**Merging multiple data sets**

* Merge business table and review.