Task List

1. **Parse JSON data into loadable format (e.g. csv)**  [YZ]

- subtask1: research methods to parse JSON data

- subtask2: ...

1. **Load the formatted data into Postgres SQL**  [YZ]
2. **Identify a list of key questions which can be answered by our project** [XY]

Our project has three goals: 1) helping customers find the popular restaurants in a particular neighborhood; 2) helping researchers identify the fraud users; 3) understanding how the genders of the customers that affect their reviews and tips. To achieve our goals, we need to answer their related questions:

1. **To find out the popular restaurants in a particular location**
   1. **What are the top 10 cities in the country that have the most business (businees\_id) of restaurants?**
      1. We will first focus on the cities which have enough restaurants and thus have the potential to generate more subjective results
      2. We will later on employ our analysis method to the less popular cities
   2. **What are the restaurant categories in each city?**
      1. We could know what type of restaurants is more popular in a city
      2. We could identify the category of interest for 3-1-c)
   3. **What are the top 10 restaurants in a category in 3-1-b) that have the most review\_counts?**
      1. The restaurants with more review\_counts means they have been visited by more customers, and thus indicate some part of the popularity of a restaurant
   4. **What is the star of the restaurants that are obtained in c?** 
      1. We believe that if the restaurants have more review\_counts, the star of a restaurants have higher potential to subjective
      2. The restaurant with highest star could be the most popular restaurants in the particular location

We will apply these analysis methods from b to c to any other cities of interest.

1. **To identify the fraud users**

Before we start analysing the data, we need to decide the restaurant of interest. We will first focus on the most popular restaurant obtained from section 3-1-d) because the popularity enable it has more user data points that can provide more subjective results.

* 1. **How many review that the restaurant have each day?**
  2. **What is the averaged review times that restaurant have each day?**
  3. **How many checkin that the restaurant have each day?**
  4. **What is the averaged checkin times that restaurant have each day?**
     1. We need to know the averaged checkin time and review time to identify the dates of interest
     2. We may not consider the dates if they have few review/checkin times
  5. **What is the averaged stars of a restaurant in each date (in the dates of interest)?** 
     1. The averaged stars are calculated based on the given star in each date from ‘review.json’
  6. **How does the averaged dately stars of a particular business change with time?**
     1. By joining ‘review.json’ with ‘business,json’, we can average starts of a restaurant in each date since the opening date
     2. The stars-dating plot shows how consistent of the product quality in the restaurant is
  7. **Which customer (user\_id) gives the star that is 1 point higher or lower than the average star at the date of interest?**
     1. We can compare star given by each user with the average with the averaged star in that particular date
     2. If the difference between given star and the average star out of 1, we believe the user is suspicious - giving abnormal ratings

We will later on employ our analysis method to the restaurants of interest.

1. **Understanding how the genders of the customers that affect their reviews and tips**
   1. **What is the gender distribution of customers?**
      1. We will start with the restaurants in 3-1-d) which may have the most users
   2. **How many male customers give tips?**
   3. **How many female customers give reviews?**
      1. We are interested about whether male customers prefer give tips or reviews
   4. **How many compliments (including hot, cute, plain, cool, and funny) that given by each male/female customer to the restaurants they have visited?**
   5. **What is the averaged compliments of a male/female customer?**
2. **Understand relation/attribute and identify fields of interest**  [XY]

We need the relation *(business.json)* to address the problems in section 3 - 1) and the relations (review.json + checkin.json + review.json + user.json) for section 3 -2) and 3).

We will choose the related attributes based on needs. For example for section 3-1), we will select the attributes ‘categories’, ‘city’, ‘state’, ‘business\_id’ from relation ‘business.json’.

1. **Perform data cleaning** [RW]

Before we get into profiling the data and extracting derived relations, we need to perform some clean up. This process is mainly based on the understanding we have developed on the dataset during step 4. For example, suppose we identify the “text” field in the REVIEW file as relevant, and also want to limit its character length to between 10 and 5000, the Yelp max character length, then we would clean up the REVIEW file by adding a column named *text\_len* that contains the length of the review body text. After that based on this column we can easily filter out reviews that doesn’t fit our standard and we are left with reviews with reasonable lengths. Similar procedures will be performed for other fields of interest.

1. **profile the data set (missing values, mean…)** [RW]

We will mainly perform three types data profiling: single column profiling, cross-column analysis, and inter-table analysis. Single column profiling will focus on improving understanding of frequency distribution of different values, types, and use of each column (e.g. average length of Yelp reviews). Embedded value dependencies can be exposed in a cross-columns analysis (e.g. relationship between a business’s average rating and its location). Finally, overlapping value sets possibly representing foreign key relationships between entities can be explored in an inter-table analysis (e.g. the set of businesses that a user has reviewed). [1]

1. **Extract derived relations from the original dataset (e.g. abnormal)** [RW]

This step consists of rather straightforward SQL operations, but requires more high-level interpretations and guidances. The goal is to output relations that include attributes that can help answer certain questions. During this step, information gathered in step 6 the profiling stage could help immensely. For example, if we are interested in how a user’s propensity to generally rate highly or poorly is, we can extract data from all the businesses that the user has rated in the business.csv file. Then we can include important columns such as the business’ average stars, review counts, the user’s star for that business and his/her average rating.

1. **Design the ERD** [All]

Yelp Raw Dataset

* yelp\_academic\_dataset\_business.json
  + business\_id, name, neighborhood, address (#, st.), city, state, postal\_code, latitude, longitude, stars, review\_count, is\_open, attributes, categories, hours
* yelp\_academic\_dataset\_checkin.json
  + time, business\_id
* yelp\_academic\_dataset\_photo.json
  + photo\_id, business\_id, caption, label
* yelp\_academic\_dataset\_review.json
  + review\_id, user\_id, business\_id, stars, date, text, useful, funny, cool
* yelp\_academic\_dataset\_tip.json
  + text, date, likes, business\_id, user\_id
* yelp\_academic\_dataset\_user.json
  + user\_id, name, review\_count, yelping\_since, friends, useful, funny, cool, fans, elite, average\_stars, compliment\_hot, compliment\_more, compliment\_profile, compliment\_cute, compliment\_plain, compliment\_cool, compliment\_funny, compliment\_writer, compliment\_photos

**Reference**

[1] WikiPedia. “Data Profiling.” URL: <https://en.wikipedia.org/wiki/Data_profiling>