Group Assignment 2:

Describing our data:

1) List of the dataset(s) acquired, together with their locations, the methods used to acquire them, and any problems encountered. Record problems encountered and any resolutions achieved.

We extracted our datasets from https://www.yelp.com/dataset/challenge . The only dilemma we had with this data set is that when we downloaded it via yelp all files were in json format. So, we have to convert the essential dataset(s) into a .csv format by creating a python script.

2) Describe the data that has been acquired, including the format of the data, the quantity of data (for example, the number of records and fields in each table), the identities of the fields, and any other surface features which have been discovered. Evaluate whether the data acquired satisfies the relevant requirements.

The yelp data set downloaded are records of metadata about businesses' , the users who reviewed or gave tips about the business, photos taken of the business, tips given, and reviews give. After doing a visual inspection of the data we found that overall our data sets are structured or semi structured. We are considering some of the datasets to be semi structured, as some of the datasets have an attributes have some inconsistencies of a json being stored in the column, while other entries do not. We will be doing our machine learning tasks with relational data. The yelp dataset is quite huge. We will have more than enough for model training and testing.

1. **Business Table:**

The business table consists of approximately ~ 84,349 entries.

The business table consists of 60 features.
The fields that the business table contain are:

`address` longtext,

`attributes` longtext,

`attributes_AcceptsInsurance` longtext,

```
`attributes_AgesAllowed` longtext,
`attributes_Alcohol` longtext,
`attributes_Ambience` longtext,
`attributes_BYOB` longtext,
`attributes_BYOBCorkage` longtext,
`attributes_BestNights` longtext,
`attributes_BikeParking` longtext,
`attributes_BusinessAcceptsBitcoin` longtext,
`attributes_BusinessAcceptsCreditCards` longtext,
`attributes_BusinessParking` longtext,
`attributes_ByAppointmentOnly` longtext,
`attributes_Caters` longtext,
`attributes_CoatCheck` longtext,
`attributes_Corkage` longtext,
`attributes_DietaryRestrictions` longtext,
`attributes_DogsAllowed` longtext,
`attributes_DriveThru` longtext,
`attributes_GoodForDancing` longtext,
`attributes_GoodForKids` longtext,
`attributes_GoodForMeal` longtext,
`attributes_HairSpecializesIn` longtext,
`attributes_HappyHour` longtext,
`attributes_HasTV` longtext,
`attributes_Music` longtext,
`attributes_NoiseLevel` longtext,
`attributes_Open24Hours` longtext,
`attributes_OutdoorSeating` longtext,
`attributes_RestaurantsAttire` longtext,
`attributes_RestaurantsCounterService` longtext,
`attributes_RestaurantsDelivery` longtext,
`attributes_RestaurantsGoodForGroups` longtext,
`attributes_RestaurantsPriceRange2` longtext,
`attributes_RestaurantsReservations` longtext,
`attributes_RestaurantsTableService` longtext,
`attributes_RestaurantsTakeOut` longtext,
`attributes_Smoking` longtext,
`attributes_WheelchairAccessible` longtext,
`attributes_WiFi` longtext,
`business_id` longtext,
`categories` longtext,
`city` longtext,
`hours` longtext,
`hours_Friday` longtext,
`hours_Monday` longtext,
`hours_Saturday` longtext,
`hours_Sunday` longtext,
`hours_Thursday` longtext,
`hours_Tuesday` longtext,
`hours_Wednesday` longtext,
`is_open` int(11) DEFAULT NULL,
`latitude` longtext,

`longitude` longtext,

`name` longtext,

`postal_code` longtext,

`review_count` int(11) DEFAULT NULL,

`stars` int(11) DEFAULT NULL,

`state` char(63) DEFAULT NULL
```

Discoveries: We have attributes of restaurants and also categories of a restaurant.

2. **Checkin Table:**

The checkin table consists of approximately ~ 167,416 entries

The checkin table has 2 columns.

```
`date` longtext,

`business_id` longtext
```

Discoveries: Before looking more thoroughly at the dataset, we though date was individual dates not more than one date.

3. **Photo Table:**

The photos table consists of approximately ~ 199,899 entries.

The photo table has four columns.

```
`photo_id` varchar(255) DEFAULT NULL,

`business_id` varchar(255) DEFAULT NULL,

`label` varchar(255) DEFAULT NULL,

`caption` varchar(255) DEFAULT NULL
```

Discoveries: N/A

4. **Review Table:**

The review table consists of approximately ~ 5,809,913 entries.

The review table has 9 columns.

```
 `business_id` varchar(255) DEFAULT NULL,
  `cool` int(11) DEFAULT NULL,
  `date` varchar(255) DEFAULT NULL,
  `funny` int(11) DEFAULT NULL,
  `review_id` varchar(255) DEFAULT NULL,
  `stars` varchar(255) DEFAULT NULL,
  `text` longtext,
  `useful` int(11) DEFAULT NULL,
  `user_id` varchar(255) DEFAULT NULL
```

Discoveries: N/A

5. **Tips Table**

The tip table consists of approximately ~ 1,132,544 entries.

The tips table has five columns.

```
 `business_id` varchar(255) DEFAULT NULL,
  `compliment_count` int(11) DEFAULT NULL,
  `date` varchar(255) DEFAULT NULL,
  `text` varchar(255) DEFAULT NULL,
  `user_id` varchar(255) DEFAULT NULL
```

Discoveries: N/A

6. **User Table:**

The user table consists of ~ 1,699,849 entries.

The user table has 22 columns.

```
          `yelping_since` varchar(255) DEFAULT NULL,
    `useful` int(11) DEFAULT NULL,
    `compliment_photos` int(11) DEFAULT NULL,
    `compliment_list` int(11) DEFAULT NULL,
    `compliment_funny` int(11) DEFAULT NULL,
    `compliment_plain` int(11) DEFAULT NULL,
    `review_count` int(11) DEFAULT NULL,
    `friends` text,
    `fans` int(11) DEFAULT NULL,
    `compliment_note` int(11) DEFAULT NULL,
    `funny` int(11) DEFAULT NULL,
    `compliment_writer` int(11) DEFAULT NULL,
    `compliment_cute` int(11) DEFAULT NULL,
    `average_stars` varchar(255) DEFAULT NULL,
    `user_id` varchar(255) DEFAULT NULL,
`compliment_more` int(11) DEFAULT NULL,
    `elite` varchar(255) DEFAULT NULL,
    `compliment_hot` int(11) DEFAULT NULL,
    `cool` int(11) DEFAULT NULL,
    `name` varchar(255) DEFAULT NULL,
    `compliment_profile` int(11) DEFAULT NULL,
    `compliment_cool` int(11) DEFAULT NULL
```

Discoveries: N/A

The data that we are utilizing to compute our ML tasks, satisfies the requirements of what we want to accomplish. The biggest concern we have is that the business table needs to be restructured and we need to analyze what is the best way to preprocess data in this table and how we can transform features in a clean manner.