**Capstone Project 1: Data Wrangling**

Yelp’s dataset provides json files for business, checkin, photo, review, tip, and user. For my first capstone project, I will be using the json file related to businesses. The files contains 188,593 rows and 15 columns of data.

When I first tried to read in the file into Jupiter Notebooks, I got an error. Upon looking at the structure of the data in a text editor, I saw that a curly bracket was missing. Using the find and replace feature, I was able to add the curly bracket where needed for successful reading of the file into Jupiter Notebooks.

I imported the pandas and json packages to read in the file and put into a data frame to get better visibility into how the data was structured in the columns.

Using descriptive python methods as describe, shape, info, I got a better sense of the different data types and null values that exist. 5 out of the 15 columns don’t have complete information. Since the dataset contained a lot of data, I decided to drop rows where there were null values. Since I’m mainly interested in restaurants and food places, I further filtered out any rows that didn’t have ‘Food’ or ‘Restaurants’ as one of their categories. This brought the dataset down to 53,585 rows.

The attributes and hours columns had nested data and needed to be pulled out into individual columns so that the info could be accessed at the most granular level. I pulled out all the nested data from the attributes columns and merged them back into the main data frame.

The data when plotted in a histogram resembles a normal distribution.

After looking through the data, I noticed that there are outliers related to location. The dataset is described as being from 10 metropolitan areas but the location related columns shows that there are more than 10 areas in the dataset. I plotted the latitude and longitude points and saw that most of the restaurants and food places had latitude measurements greater than 20 degrees and longitude measurements less than 50 degrees. Since there were only 23 businesses that were showing as outside, I filtered these out too.

After all the data wrangling, I now have a dataset that has 53,562 businesses with 95 columns.