**Yelp Dataset**

Exploration and analysis on Yelp Datasets are both fun and challenging. The Yelp datasets available to be downloaded have a larger file size than the data we worked on in our weekly exercises, thus requiring more time for clean-up, slicing/dicing, and processing. I had to export multiple JSON files to dataframes and use panda dataframe join to merge to have a single combined dataset. Each JSON file presents in a different data dimension where I had to code to derive new variables for this project such as:

* Business File provides a unique row for each Business ID. The business attribute column is establishment features in a dictionary-type column where I had to loop through and identify pricing information and other attributes that value=True (i.e. Take Out, dine-in, Parking, etc.).
* Check-in File provides a unique row for each Business ID, but the check-in dates are stored in a single column where dates are comma delimited. I created a new variable of check-in count for each business\_id by parsing on each date and also exploding and converting dates to determine the count per year.
* Tip File provides Tip data by User ID and Business ID. I had to do some aggregation to get a count for each business and then join the main dataset.

## What do you feel was missed during the analysis?

Yelp Reviews and Users data are available, but unfortunately, I was faced with technical and environment limitation where my computing resource is limited, and it was not able to handle loading the JSON data to dataframe without jupyter notebook timing out due to memory issue. With that, I was only able to analyze the three datasets above, where I had to filter to only consider records with categories with Food and Restaurants which narrowed the data to 64,616 records. I feel that the User Reviews and User Data will provide more opportunities to perform EDA on Yelp dataset and get maximum insight and discover patterns such as user sentiments, spot anomalies, clean and filter out records impacting distribution and check assumptions.

## Were there any variables you felt could have helped in the analysis and any assumptions made you felt was incorrect?

Besides the two datasets that I was not able to load as mentioned above, I could have explored more the business attributes/features. I did include it in my EDA but for only features with a value of TRUE, assuming these are features available in the establishment. There are more features that present other values besides True, which could have counted but was filtered out because of this assumption.

The business dataset also presents attributes such as establishment location and hours which I could have also used to gain insight but will require a transformation in order to create new variables that will be useful input to my EDA, such as: calculating the total time establishment is open by using the hours data element, determine if the location and close proximity similar businesses that may impact review and ratings using address information.

What challenges did you face, what did you not fully understand?

I was initially faced with a challenge in data preparation and creating new variables where I had to determine how to read a dictionary column in a python dataframe. This took me a while to code a function to read through this column and extract the data I needed for analysis. Once I have the variables I need, I start looking at the data distribution. I see that most of my variables are not normally distributed, and I don’t quite fully understand what could have caused why data is not normal. I hoped for more time to identify whether there are correctable factors I can implement to see if my data is really normally-distributed. I believe that if I was able to present and use a normalized distributed variable, the result of this EDA could have provided more meaningful insight into the data and resulted in an accurate hypothesis test and regression analysis result.

Outcome of your EDA

I have seen several main points on business in Yelp with my exploratory and data analysis:

1. How ratings are impacted by multiple factors such as attribute features, number of check-ins, tips and user reviews
2. Compared ratings, attribute, and reviews between open and closed establishments and what could have potentially impacted their outcome.

This project has allowed me to use the knowledge and practice the skills I gained from working on different weekly exercises from this class:

* + Read, Clean-up, and Transform data
  + Helps identify errors in datasets
  + Determine the distribution and understand the relationship among them
  + Conduct hypothesis testing and regression analysis.

This has been a great semester, and I am looking forward to continuing to learn and apply these skills to future EDA projects.