**OSEMN Process for Working on Extracted Fields from the Yelp Reviews Data Set**

**Rahul Kumar Nalubandhu and Sandra Estrada**

**Introduction**

Conducting a review analysis gives businesses the opportunity to improve customer experience, identify service gaps and gain real time insights amongst many other benefits. This analysis dives into customer reviews entered in Yelp for businesses in different states throughout the United States.

**Problem**

The issue with having little to no customer reviews is that this can negatively impact sales. Customer reviews are an important channel to attract customers and increase sales. A benefit in analyzing reviews at the business, city and state level will provide insight into which states and/or businesses have the least customer engagement and allow for proper intervention.

**Obtain Data**

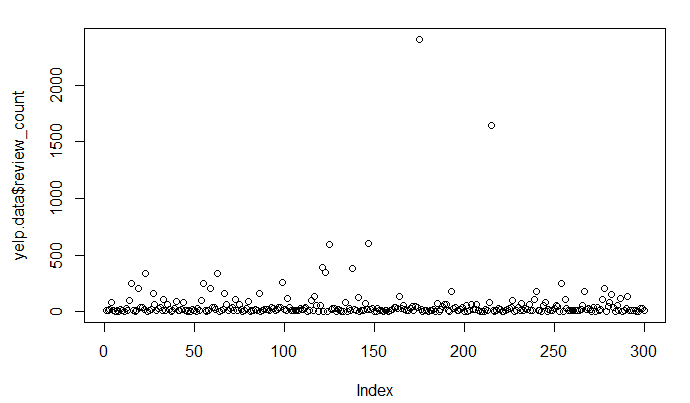
Yelp is a one-stop platform which enables customers to connect with businesses. More than 80 million people visit this platform in a month to find businesses and service providers. Customers are given the ability to leave reviews and request quotes from local businesses amongst many other things. In return, local business owners are given the ability to communicate with their customers and respond to reviews to build trust with their customers. The customer review data set is acquired directly through Yelp. The data set is 4.04GB (1 point) and split into multiple Json files (2 points) which contain businesses, reviews, and user data. In addition, the data has punctuation (1 point) and has more than one type of related data (2 points). Based on the point system requirements provided, the yelp data is a 6-point data set.

**Scrub Data**

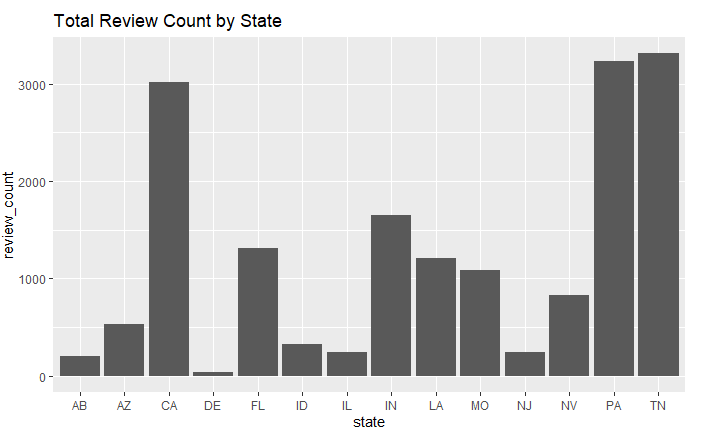
Prior to conducting the analysis for the yelp data set, the data is extracted and consolidated into a csv and Json file for the following fields: business name, city, state, and review count. Visual studio is the primary application utilized to read and extract data in the python programming language. In addition, all null values are deleted from the data set. Null values are deleted to not compromise the integrity of the analysis.

**Explore Data**

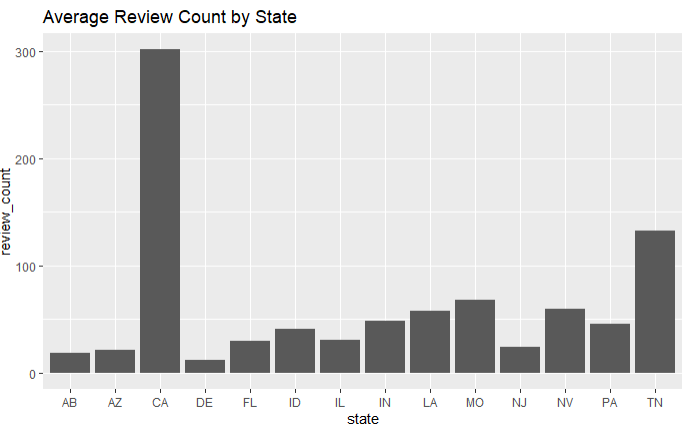
For the exploration stage, the csv data is loaded into R-studio. To begin the exploration stage the review count field is plotted. There are two evident outliers for businesses having more than 1,000 reviews on yelp. Most of the businesses have a review count of less than 500 reviews.

  
*Graph 1*

The review count variable is also examined at the state level for the total and average number of reviews. These observations indicate the states with the most reviews are Pennsylvania and Tennessee followed by California. On the opposite end of the spectrum, Delaware is the state with the least total and average number of reviews. On average, the state whose businesses have more reviews compared to other states is California.



*Graph 2*



*Graph 3*

**Model Data**

After reviewing the dataset, a decision tree is the most suitable modeling technique for this analysis. A decision tree will contain the different factors that determine if businesses are likely to face less customer reviews due to the state they are in or even the demographics of their target customers. While this is deemed the best approach for this analysis, more data needs to be gathered and analyzed to properly complete this.

**Interpret Data**

The business with the most reviews in this data set is Santa Barbara Shellfish Company located in CA with 2,404 reviews. For the companies with the lowest count of reviews, there are a few with only 5 reviews. An interesting observation is that most businesses are in FL. This observation is aligned with the results from the Average Review Count by State plot.





**Tasks Completed per Team Member**

Rahul Kumar Nalubandhu – Read, extracted data from the original data set, removed null values, converted the Json data into a csv format file, and assisted in portion of the r analysis.

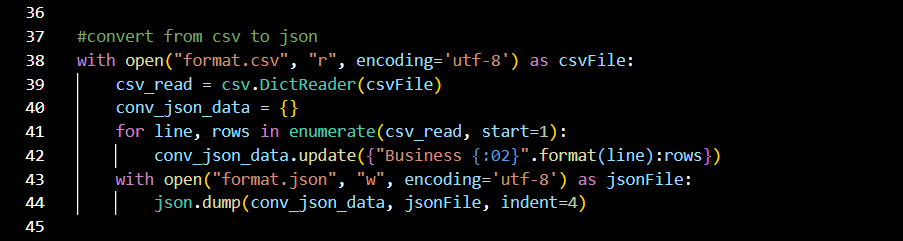
Sandra Estrada – Converted the csv format into a Json format file, completed the r analysis, and wrote the OSEMN report.

**Appendix**

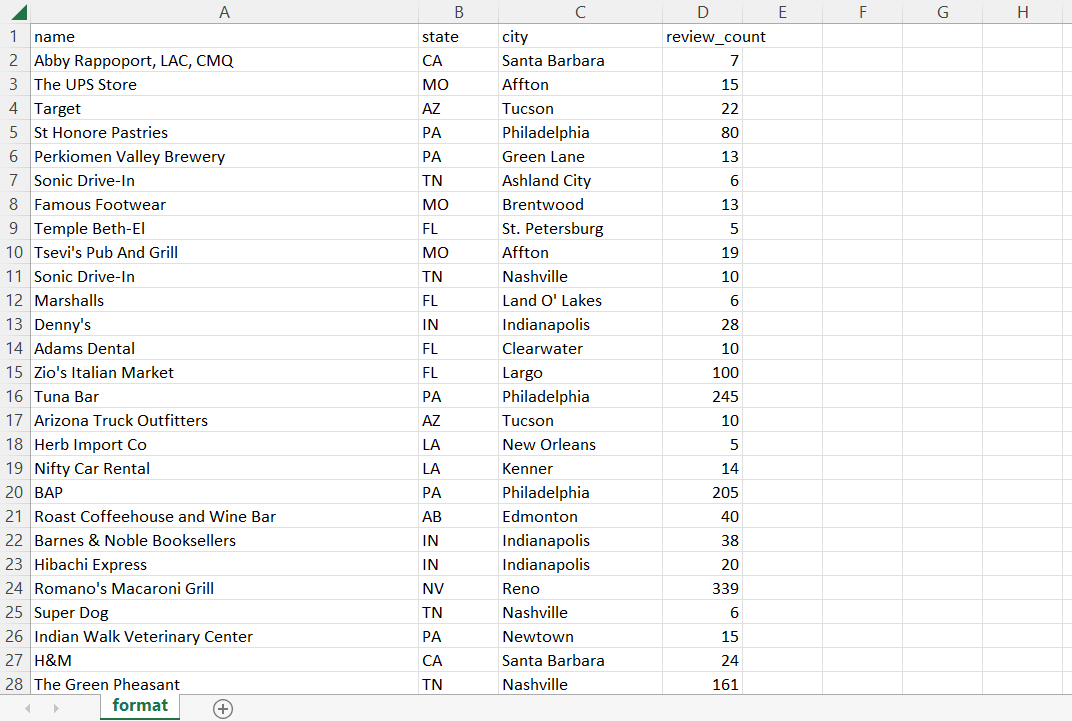
Initial Data:



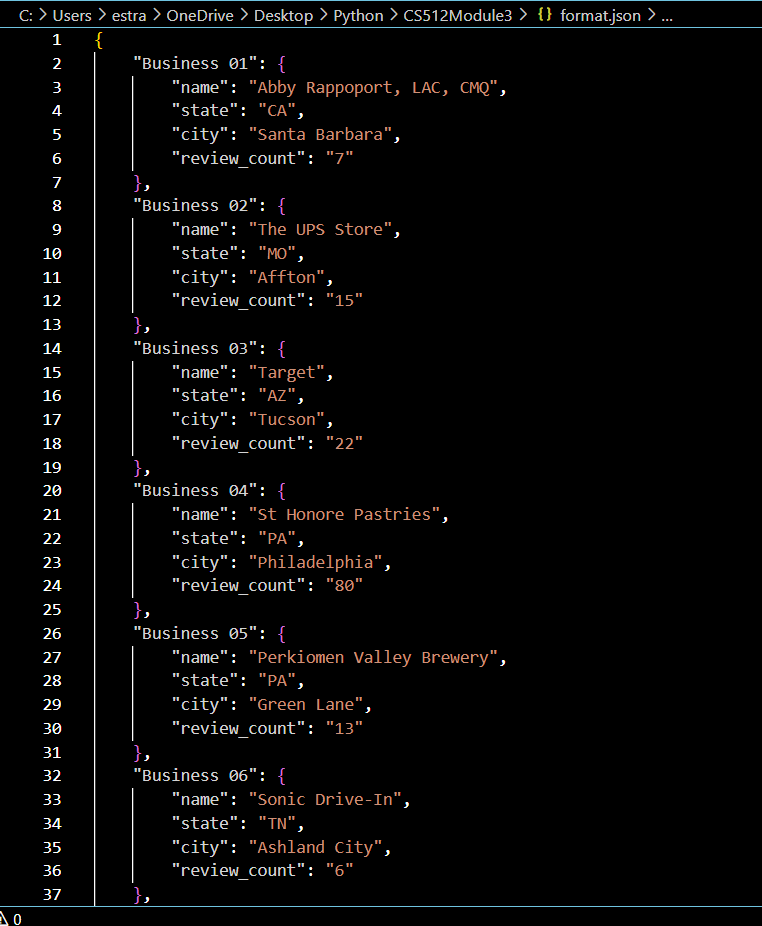
Python code for conversions:



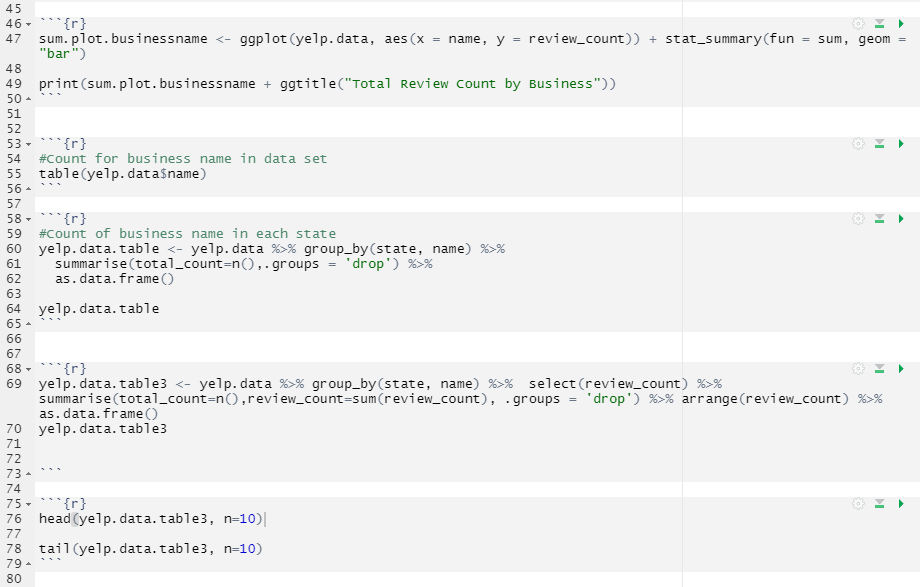
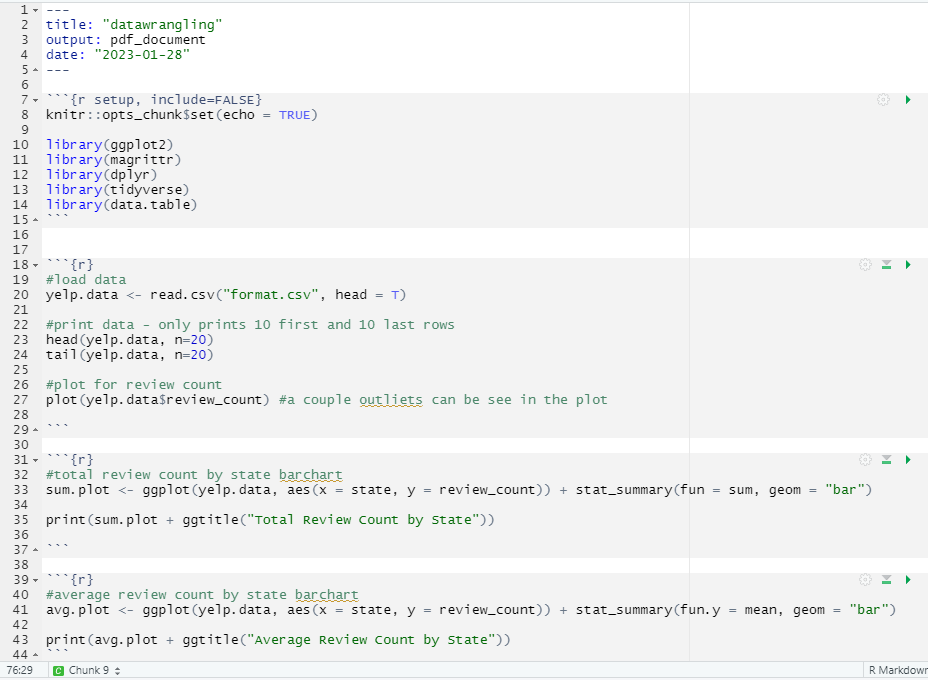
CSV Sample:



Json Sample :



R-Studio code:



**References**

e-satisfaction. (n.d.). *7 reasons why customer reviews are important*. Retrieved January 27, 2023, from [https://www.e-satisfaction.com/7-reasons-why-customer-reviews-are-important/#:~:text=Analyzing%20reviews%20left%20by%20your,what%20your%20customers%20truly%20want](https://oregonstateuniversity.sharepoint.com/sites/CS512TeamProjects/Shared%20Documents/General/e-satisfaction.%20(n.d.).%207%20reasons%20why%20customer%20reviews%20are%20important.%20Retrieved%20January%2027,%202023,%20from%20https:/www.e-satisfaction.com/7-reasons-why-customer-reviews-are-important#:~:text=Analyzing%20reviews%20left%20by%20your,what%20your%20customers%20truly%20want)

*Convert CSV to JSON using python with new headers*. Stack Overflow. (1969, July 1). Retrieved January 27, 2023, from <https://stackoverflow.com/questions/73089933/convert-csv-to-json-using-python-with-new-headers>

Rana, D. S. (2022, September 12). *Learn Review Analysis: Why, how, data sources with Free Trial*. Learn Review Analysis | Why, How, Data Sources with Free Trial. Retrieved January 29, 2023, from <https://www.repustate.com/blog/review-analysis/#:~:text=Review%20analysis%20allows%20you%20to,new%20sales%20opportunities%2C%20and%20more.>

*Indexing and selecting data - pandas 1.5.3 documentation*. (n.d.). Retrieved January 29, 2023, from <https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy.>

Yelp for Business. (2023, January 23). *Plan, start, grow, and advertise your small business*. Yelp for Business. Retrieved January 29, 2023, from https://business.yelp.com/