# Project Yelp Datasets

Team Members:

* LaKeisha Ringo
* John Giglio
* Pride Hawkins
* Stephen Wray

Scope:

For this project, we will be reviewing Yelp Datasets to answers the questions and hypothesis below:

Sources of data:

From Kaggle there was a competition that pulled data from Yelp and

the dataset was converted from JSON to CSV format.

This dataset is a subset of Yelp's businesses, reviews, and user data.

It was originally put together for the Yelp Dataset Challenge which is a chance for students to conduct research

or analysis on Yelp's data and share their discoveries.

In the dataset you'll find information about businesses across 11 metropolitan areas in four countries.

**In total, there are**:

5,200,000 user reviews

Information on 174,000 businesses

The data spans 11 metropolitan areas

https://www.kaggle.com/yelp-dataset/yelp-dataset/home

Data Items will include:

* Yelp Business & attributes
* Yelp Users
* Yelp reviews
* Yelp tips

Key Questions:

1) The sentiment analysis of the review and the stars are related. (Or not for our null hypothesis).

2) Can we say certain categories of businesses are preferred in certain states or geographical areas.

3)The top influencers are representative of the overall population's reviews.

(for example, can we say a certain group of top influencers match the overall populations opinions)

4) There will be no differences in reviews for a given company that is a chain business or restaurant.

Rough Breakdown of Tasks:

Review use of API.

Clean and review any data munging necessary for better analysis.

Create subsets of each data form to help each question above.

Review and test

Develop visualizations based upon data subsets.

Summarize analysis.