STOR 565 Project Proposal

Group 2 Members:

- Jessica Ho
- Katherine Wang
- Alex Kan
- Ishan Shah
- Svetak Sundhar

Problem Description

- **Problem**: Restaurant owners currently have plenty of great locations that they could place their restaurant. This analysis seeks to inform restaurants what attributes they need to have to be successful in the location they choose.
- We want to determine how restaurants can best improve themselves to improve ratings. What features make a good business? Determine the overall rating of a restaurant depending on how reviews change over time.
- **Predictions**: Predict rating (demographics, average review sentiment, geographic location, restaurant features and categories, user information)
    - How well a shop will be rated based upon demographics
    - What makes a helpful review

Data Set Description

- Data Size: 3GB in total from Yelp
- Source: https://www.kaggle.com/yelp-dataset/yelp-dataset/data
    - Business attributes (location, stars, review count, hours, type of business, etc.)
    - Individual user reviews (star ratings, actual reviews word for word, user information, usefulness of review, etc.)
- Demographic Data for US Zip Codes from simplyanalytics.com
    - Variables related to population, income, education level, racial/gender/age distribution

ML Techniques

- LASSO to do variable selection
- Principal Component Analysis
- Support Vector Machines
- Convolutional Neural Networks
- With clustering, we can determine which features are most important for a business' overall review (unsupervised learning) (might use this technique)

Challenges and Solutions

- Joining together all of the different tables
- Reducing the very high dimensionality of the data