Project 1: Exploratory Data Analysis
Proposal

Ameeshi Goel and Pavan Gollapalli

https://www.yelp.com/dataset

Motivation:

As hungry college students, the question of where to get food crosses our minds far too often. When deciding where to get food from, however, we often face the problem of where to get food. To resolve this issue, we have turned to apps like Yelp or Google Maps to understand the opinions of people like ourselves who have also dined at these places and left reviews and ratings of the locations. To that end, attempting to understand the dynamics at play behind these tools (How does a user's reviewing habits factor into their opinions of places? Should we trust these ratings?)

Related Work:

We first figured out common interests and on realizing that food was a big one, we thought of looking for food related datasets. We wanted to find a dataset that would give us some insight into how people pick the places they eat at, how often they frequent the places they like and how they rate these places, among other questions. Moreover, Karan's and George's presentations in class motivated us to find a dataset that could be used to model several interesting and meaningful questions. Looking through the public datasets on github (https://github.com/awesomedata/awesome-public-datasets), we stumbled across the 'Yelp Dataset Challenege' and thought it fit our criteria well.

Data:

This data is downloadable from Yelp's website, which offers the dataset as part of their semiannual data challenge. The data comes in both JSON and SQL format, but we have chosen to use the JSON format due to the fact that we will be conducting our analysis in Pandas and Jupyter. There are 6 JSON files, each representing a different table:

- Business.json (174,000): business_id, name, neighborhood, address, city, state, postal code, latitude, longitude, stars, review_count, is_open, attributes, categories, hours
- Review.json (5,200,000): review_id, user_id, business_id, stars, date, text, useful, funny, cool

- User.json (5,200,000): user_id, name, review_count, yelping_since, friends, useful, funny, cool, fans, elite, average_stars, compliment_hot, etc (other compliment columns)
- Checkin.json (146,350): time, business_id
- Tips.json (1,100,000): text, date, likes, business_id, user_id

Questions:

1. What time of day do most businesses get checkins? What about each business category?
    - This question could be addressed in pandas as the checkins table is organized by storing an object of checkins per time slot for each business. By combining all the objects stored in the "time" attribute of each business's row, it can be seen which time slots on which days (ex: Tuesday at 6pm). In order to break this down by category, we can join the Checkin table with the Business table on each checkin's business_id attribute, and use the corresponding business's category data to sort the number of checkins in each time slot.
2. What business categories do men tend to prefer, and what categories do women tend to frequent?
    - As the Yelp dataset does not contain gender information for its users, we will likely have to preprocess the user data slightly to add gender information; as we are provided the first name of each user, we can likely infer the gender from this information using an API. Once we have this information, we can join the review table to the user table and business table, aggregate reviews by gender and category, and compare the results.
3. What zip codes have the highest average ratings for business? Do these businesses have more reviews, and are those reviews read more?
    - Each business has location data, so it is possible to group businesses by location and then average their ratings. By then joining the reviews table, it becomes possible to count the number of reviews for a business and use the overall "useful" count for those reviews to see if people read and interact with the reviews more than in areas with poorly reviewed businesses.

Possible Findings and Implications:
We want to be able to have an impactful bunch of answers at the end of our analysis. We think we'd probably discover the following:

1. The zip code is a good indicator of the kind of the reviews business get. In other words, people from different states tend to be a more positive or negative group of people, which is reflected directly in the general sentiment of the reviews. This would make it easier for businesses to know where to branch out and perhaps, where to shut down from.
2. The category of a business determines the usual mass of check-in hours at that place. Therefore, a category like "Gastropub" would probably have more check-ins at night than in the day. And conversely, a category like "Bakery" would naturally have more check-ins in the day than in the evening.
3. Men and women have different preferences in categories. This would make it easier for businesses to realize if they're crowd has a tilt towards one gender and, if they need, how to make it more conducive for the other gender.