I am using the Yelp academic dataset for my project, which is a large dataset offered by Yelp to the academic community (https://www.yelp.com/academic_dataset) . There is a data mining competition associated with the dataset, but according to the terms and conditions the dataset is free to be used for any academic project. The dataset is comprised of 250 closest businesses to 28 US universities (Harvard and MIT are the only Boston schools included). This is a very large dataset comprised of five different json files:

yelp_academic_dataset_business.json	Contains information about the individual businesses (business id, business name, address, hours, stars, categories (food, bars, healthcare, etc), longitude, latitude, number of reviews, whether it is closed
yelp_academic_dataset_checkin.json	Check-ins for individual businesses (not planning on using this data)
yelp_academic_dataset_review.json	Contains reviews submitted by users for businesses, each review has a business id associated with it so the review can be linked to the business json file. The actual review is text, but there is also a stars field (1-5), a date field, and a rating field (useful, cool, funny)
yelp_academic_dataset_tip.json	Tips from users for businesses. Tips are in text. (not planning on using this data)
yelp_academic_dataset_user.json	Data about individual users – average stars (1-5), number of fans, friends (id numbers), review count, compliments, first name, date that they created their account

The data is a mix of text, dates, longitude and latitudes, id numbers, and integers representing stars, ratings etc. Luckily the dataset is very dense, there doesn't appear to be any missing data, except in cases where, for example, a user does not have any friends, and therefore their friend list is empty.

Data Analysis:
1. When I was looking through the data I noticed that many of the users (at least 1/3 , maybe ½) had only written one review. I hypothesize that the reviews of users who only posted one review will be more polarized towards 5 stars and 1 star, whereas the overall distribution of reviews will have more of a normal distribution. I believe this will be the case because someone who only wrote one review likely had a very positive or very negative experience, it is also possible that some companies created dummy accounts to positively rate themselves.
2. I hypothesize that similar businesses have similar check-in data. For example, coffee shops probably have the most check-ins in the morning, whereas bars probably have the most check-ins at night. I am going to try clustering to see if

similar businesses are clustered together based on the check-in data
3. I would like to use logistic regression to predict whether a business is closed (0) or open(1) – I believe there will be indicators in the data (eg. low rating), that suggest whether a business has failed.
4. I would also like to try to predict whether a review is positive or negative based on the text using sentiment analysis.
5. Finally, I would like to construct a graph using the friend data from the users file (like a mini social network), in order to investigate things like – do people with more friends write more reviews?