This project uses data from the Yelp Dataset Challenge, which has rules against providing the data to a third party.

I imported data from 3 main files, which are formatted in json.

User.json is 1.5 GB, a list of users and metadata about their reviews and the compliments they have given and received from other users, as well as their friend connections. There were too many users for me to use all at once, so I assigned each user a unique number at random and chose a sample of 20,000 users. Using the readline function to scan through the file 100,000 rows at a time, I converted only the lines of the users in the sample using the jsonlite library to convert the json data to a dataframe. I modified some of the fields, like simplifying a list of friends to a number of total friends, and a list of years the user was awarded elite status to the number of years that they got the award. The date that they registered with yelp was converted to a numeric value, the number of days since 1970. Once I had the dataframe with all the users in my sample, I used the R save function to store the dataframe for later use.

Review.json is 3.7 GB, a list of reviews and unique IDs of the user and business that they reviewed, and the number of compliments that review received. The users in my sample were added to a hashmap that made it simple to check if a user ID was present, and the review file was scanned using readline. If a review’s user ID was present in my sample, I added it to the reviews dataframe. The review text itself was not included, as I wasn’t doing text analysis for the project and the text took up the vast majority of space in the file. Since the lines of the json file are all formatted the same way, I was able to parse the user ID from the beginning and check if it was in my sample before converting the whole row to json, which greatly improved the performance. Even with this, scanning through the file took almost a half hour. Once the result is saved as a dataframe, it can be saved and loaded in less than a second.

Business.json is 129 MB and contains geographic and rating information about businesses, among other metadata. This file was mainly used for getting rating information from the business ID, so I only imported the fields related to the rating and geographic location including the business name. To save space, only the businesses that were rated by at least one user from my sample were saved.