Data Manipulation Exercise
Programming Camp, Stanford Economics
Brian Higgins and Ciaran Rogers

In this problem set we will complete a simple research project. The objective is to see your creative side when having data in your hands, the same way that you may end up having to use it when doing real research. We also want you to familiarize with data manipulation techniques. You will have to import data, clean it, and generate results based on a series of questions. Some of the questions are ill-defined and somewhat open ended —— that is a feature and not a bug.

For this problem set, you can use any software that generates reproducible code (e.g., Stata, Python, MATLAB, or R). The profession as a whole tends to use Stata —— but we seem to be evolving so if you are familiar with other languages, be our guest. We recommend you invest in learning at least basic Stata, as you will probably end up needing it at some point at least to replicate someone else's findings.

Please, attempt to work through this problem set on your own. You will have problem sets and exams that include programming, so it is important that you have the confidence to code on your own. That said, we are happy to help if you end up getting stuck setting something up or while coding. Feel free to email us, and we could chat through Zoom. When submitting your problem set, please submit a writeup in PDF and your code (.do, .m, .py, .r file). If you are using something other than Stata, a Jupyter notebook or an R notebook are just fine as well. Work that is not reproducible will not count as delivered.

Data

To make it interesting, we will be using two datasets that give a glimpse onto the set of restaurants that you will find around Stanford. Data comes from TripAdvisor. The first dataset, restaurants, contains a list of 1,108 restaurants that are less than 10 miles away from the Econ department. The second dataset, ratings, contains a subset of the reviews (3,342 of them) that you can find for these restaurants on TripAdvisor. Variables names in each of these datasets are self-explanatory.

Exercises

1. Load the restaurants dataset onto Stata / R / Python. What is the average rating in the dataset? What about the median one?

2. The restaurants data contains three different price levels, represented by dollar signs ($). What is the average rating for each of them?

3. Now, let's assume you are up to a snack while working in the Econ department. What are the five closest restaurants to Landau? Consider that the department is located at lat, lng = (37.428769, -122.165958). Bonus points if you take into consideration the curvature of the earth (hint: google geodesic distance).

4. Now, let's start working with restaurant chains. We define a chain to be a set of two or more restaurants that share the same name. How many chains are there in the data? What's the number of restaurants that are part of chains in the data? What are the 5 biggest chains? How many restaurants do each of these 5 chains have? Are these top chains well rated?

5. One particular chain that you will get to know well is Coupa Café. You will find yourself going there for a coffee quite often —— and you'll find their locations quite convenient. They have both locations in campus, and a few outside of it. We want to find out how many locations does Coupa have on campus. Using the address variable, a restaurant is on campus if their city is "Stanford". How many of the Coupas are located on campus? By the way, are there any other restaurants on campus?

6. Now, after a few visits to Coupa, you may want to go somewhere else to get your coffee. Note that cuisine0, cuisine1, and so forth contain information on the type of food that these restaurants serve. We define a cafe to be a restaurant that lists "Cafe" in any of their cuisine types. What are the five coffee shops that are closest to the Econ department? Repeat this for your favorite type of food.

7. We will now use the ratings data to get a glimpse of what other people dining in the Bay Area think of the restaurants near Stanford. We want to find specific key words in the reviews. We will categorize reviews based on four types of answers: (a) great, (b) friendly, (c) fresh, (d) ok. To find reviews that mention each of these words, you need to get all of the reviews to lowercase, then look for the specific word you want. Next, you can define a categorical variable (0-1) for isGreat = 1 if the review mentions "great", or "Great", or "GREAT", or "GrEaT". Construct one of each of these categorical variables for (a) through (d).

8. How do these variables correlate with the user ratings? Run a regression of the user rating on all of the variables you created in Q7, and a constant. Next, merge the price level in dollar signs from the restaurants data and construct a categorical variable going from 1 to 3 based on this price level. Repeat your regression controlling in addition for the price level.

9. Finally, you hear a friend hypothesize that tourists are stricter than locals with the food they consume abroad that comes from their country of origin. For example, this implies that French people are stricter with the French cuisine they eat in California than they are with Mexican cuisine in California. In your data, for instance, this may reflect in:
    a. French people leaving worse reviews for French restaurants than for other food types.
    b. French restaurants having worse reviews if they come from French nationals than if they come from other nationalities.

    Focus on 7 types of cuisines: Mexican, Indian, Japanese, Chinese, American, Italian, French. Use the location given by userLocation to determine where did a person visit from. Do you find evidence to support your friend's hypothesis?