### Python: Data, Science & Design Project 1

For our midterm project in my Python: Data, Science & Design class, we had to choose a data set we were interested in and work with it. In the process, we were to come up with a question about the data and explore it.

### My Data Set

From the very beginning, I knew that I wanted to work with voter data. The New York Times published “[An Extremely Detailed Map of the 2016 Election](https://www.nytimes.com/interactive/2018/upshot/election-2016-voting-precinct-maps.html)” last summer, which I became slightly obsessed with. I started my search for a data set from there, where I came across [Ryne Rohla](http://rynerohla.com/index.html/), the PhD candidate whose data the New York Times used for their interactive map. On Rohla’s website, I found a data set of votes for the 2016 presidential election within a mile of major universities. The data set includes 301 universities from around the country, with information on which state and athletic conferenc eeach university is in, as well as the voting data with six presidential candidates.

### My Question

What is the correlation between votes in universities in the North and South?

I aimed to answer my first question by mapping my data, à la New York Times. This way, I could see if there was any correlation between presidential voting in different parts of the country. I was particularly curious to see if voting within universities corresponded to expectations of votes for the city/state they are in. For example, do students at Boston University vote the way we expect the majority of Boston, Massachusetts to vote?

### Using the Data

Figuring out how to map the data was a lot harder than I originally thought it would be. I knew that in order to map my data, I would need the geolocation (latitude and longitude) of each university. The data set I chose did not have this information, so I had to download GeoPy. GeoPy gave me the exact location under a column named “location” and the geolocation under a column named “point.”

The “point” column was the important one for me, but the problem was that it gave me the latitude and longitude together in one column, as a tuple. In order to map my data, I would need the latitude and longitude to be in separate columns. To fix this, I had to loop through the data to append the latitude (df.point[0]) to an array called “lat” and the longitude (df.point[1]) to an array called “lon”. Then, I used these arrays to create two columns by the same name.

Now, I could finally map the data. For mapping, I used Folium, which is what we learned to use in class. The way I imagined the map was that at each university, there would be a circle with a radius of one mile, colored either blue for a majority of Clinton votes or red for a majority of Trump votes. This was easier said than done.

With Folium, you can create circles at a certain geolocation. What I had to figure out what how to 1) create a circle at EVERY geolocation, and 2) create circles with a radius of one mile. Figuring this out has been quite a challenge for me. I tried two different ways of doing the first step, both to no avail.

The first way was to create two arrays, one called latP and one called longP, to which I appended all the latitudes and longitudes by looping through the data set. Then, when creating the circles, I put latP and longP as the latitude and longitude. This resulted in only the first university on the list being marked. The second was to loop through the universities and create the circles within that for loop. For latitude and longitude, I just put lat and long, which are the names of the columns in my data set. The result was the same as that of the previous method. Also, both of the maps I created are very glitchy and don’t zoom in, which was not the case when I used a single instance for my latitude and longitude.

### In the Future

If I have more time with this data set, there is one more question I would like to try to answer: what is the correlation between votes in state and private universities? I imagine that I could do so by parsing through the data and comparing universities with “State” in their name to those without “State” in their name. I plan on grouping the universities into these two categories, and having each be the x-variable on a bar graph. Along the y-axis, I would show votes for Donald Trump versus votes for Hillary Clinton (in a stacked bar, with blue being the mean percentage of votes for Clinton, and red being the mean percentage of votes for Trump).