Webscraping

I decided to scrape another website, CareerBuilder.com.

In the file, GET1_Success - you can see how I was able to scrape one page and put it into a data frame that I could merge with the data set from indeed.

Scraping CareerBuilder was not easy and I took several different appraoches.

To start, I tried iterating over a list of keywords and passing it through the urls but found that the website could potentially give me different url results.

I had to iterate through 2 different lists of cities (because the first instance of a 2 word anything would be a - first and a + second.

I also had to iterate through the original list of keywords, one of them being data-science.

It proved to be too much to tackle at once so I decided to generate a list of urls based on similar principals.In MAKE LIST OF URLS, I do just this, however, just like the last time I tried downloading a lot of information from a website, my computer was unable to handle the traffic.

While I was able to capture all the data into soup, the speed of my computer slowed down tremendously and I was unable to work at a reasonable pace to finish replicating what I had down in GET1_Success.

Clean/EDA/Extract

Project4_Clean/eda/extract was how I took the dataset provided from indeed and cleaned it. I decided to consider monthly salaries as yearly salaries by multiplying their values by 12.

Visuals

Project4_Visuals represents the JD (job descriptions) and JT (job titles) I observed in addition to overall observations about the data set as a hole.

Regressions

Project4.ipynb is where I tried to put it all together. I believe I am at a place where I understand the core concepts and will be working hard to improve my models.