NFL Capstone - Data Wrangling

For this project, I would be scraping data from Pro-Football-Reference.com using BeautifulSoup. PFR is one of the better places to collect NFL and NCAA statistics as it is revered for its seemingly unlimited potential for analysis.

I first imported the appropriate packages for wrangling the data. I then initialized the pandas dataframe I need for the NFL draft data that I would be scraping. After that, I used BeautiulSoup to scrape Pro Football Reference to get a list of players who were drafted in the NFL from 1970-2012. I then isolated only the quarterbacks in the dataframe.

It turned out that the statistics included in the NFL draft dataframe I scraped included only the career NFL statistics for each player, not their college data (Which is what I would need for analysis later on). What I then needed to do was to get a list of names for each player in the dataframe along with a list of their respective urls for their college statistics on PFR. I had to get rid of periods in names (E.g., B.J., P.J.), apostrophes in names (E.g. O'Sullivan) and get rid of the indicator of whether or not they were eventually inducted into the NFL Hall of Fame (Denoted by ' HOF').

After getting the list of names and using that to make a list of urls for each player's college statistics page on PFR, I needed to initialize two dataframes: One for each player's passing statistics and one for their rushing statistics (Both of these categories are located in separate tables in the HTML). I then performed another scrape with BeautifulSoup to get the above statistics, but this time, while running the for loop, I had to account for bad web pages (And there were quite a few of them). Most of them were players that never really had much playing time, thus not a lot of statistics to report for them.

The next step in the process was merging the existing dataframes together. I performed a merge on the rushing and passing tables on the respective url links for each player. I also reordered the columns for ease of use. I noticed that some wide receivers were inadvertently included in the merged table (Which were clear due to all of the NaN values assigned to different passing categories. I rectified this issue by dropping all NaNs from the table. I had everything I needed except for the name of each player. Therefore, I had to create a function that used regular expressions to take the name of each player from their respective url links. After I used that function, I assigned a new column, 'name', to the merged dataframe.

The final step was cleaning the data and taking the career Approximate Value index for each player in the NFL draft dataframe and merging it with my newly merged passing/rushing college statistics dataframe. The reason for this is because I will be using the career AV as a response variable later on. After that, I dropped various columns that were deemed irrelevant for my project's purposes (A lot of columns that were empty and/or not useful). After that, I converted all data types that needed to be to numeric instead of object for easier analysis down the road. I also filled all missing value with 0. To wrap up, I converted the dataframe to csv for quick reference later.