NFL Capstone - Data Storytelling

For this section of my capstone project, I knew that I needed to use visualization tools in order to gain key insights into the data that doesn't stand out in the data frames alone. To start, what I decided to do was to use Seaborn's distplot tool to plot a histogram of each player's career Approximate Value score. Interestingly, the distribution showed that the vast majority of the players' career AV scores center around 0. One logical explanation is simply this: It is extremely hard for a player to become successful in the NFL. In each draft there are a lot of quarterbacks taken and frankly not very many of them go on to have great careers. So the distribution is telling us that being a quarterback in the NFL good enough to have a career AV score more than 25 (The highest is around 190 in our data frame) is a very rare event.

I then decided to look at box plot distributions for each quantitative college statistic available in the data frame. The most interesting plots were for the following stats:

- **Passing completion percentage**
    - I noticed in this box plot, that while the median looked normal (Around 55%), there were some outliers with a passing completion percentage of 100%! Upon further exploration, there were a handful of players that only had 2 or 3 attempts (And completed all of them), which caused the outliers to be seen in the plot.
- **Adjusted passing yards per attempt**
    - For this box plot, those same players that were giving us trouble for passing completion percentage were prevalent here as well. While most players saw around 7 yds, there were some that were much higher (Greater than 10 yds!). After exploring the data, not only did I see the same culprits as mentioned above, I also saw some big names like Sam Bradford, Tim Tebow, Cam Newton and Russell Wilson.
- **Passing touchdowns**
    - For this box plot, the median hovered around 35, but there were some extreme outliers. Some of which were over 120 passing touchdowns in their college careers! I filtered the data frame to include players with at least 85 or more passing touchdowns and found that (As expected) some all time greats were present here. Peyton Manning, Drew Brees, Phillip Rivers, and Sam Bradford to name a few. At this point, I'm thinking that there has to be a positive relationship between passing touchdowns and NFL career AV!
- **Interceptions**
    - I took a look at the distribution for career college interceptions and of course there were some fairly insane outliers (The median was around 25, but there were some players with greater than 60!), but I'm mostly interested in seeing what kind of relationship interceptions has with career AV. One would think that the higher the interceptions, the worse the career AV score will be.
- **Rushing touchdowns**

- For this box plot, again there were some huge outliers. After further exploration, I noticed something interesting: Most of the players included here are notable for not having statistically great NFL careers (Tim Tebow and Vince Young stand out here). It made me wonder how valuable it is to have a quarterback that mostly rushes instead of passes? Does that have a negative impact on their career AV?

The final step was to take a closer look at the various quantitative college stats and how they relate to career AV using scatter plots. For ease, what I decided to do was use Seaborn's pairplot tool, which allows a user to see all of the scatter plots conveniently at one time. It was here where things started to get quite interesting:

Notably, there was absolutely no clear relationship between career AV and pass completion percentage, pass yards/attempt and adjusted pass yards/attempt. How could this possibly be? One would think that these college stats would give you at least SOME insight into what their career AV score in the NFL would eventually be, but that does not seem to be the case here. The rest of the passing statistics seemed to have only a very slightly positive relationship with career AV. The conclusion is - Career AV does not seem to be hinged much on any player's college passing stats.

What about rushing stats? As you might be able to guess, the relationships between career AV and the various rushing stats seems to be even less positive than the passing stats.

The natural question now is: What is going on here? That is what I hope to figure out with some statistical analysis in the next section.