**DSC 530 Final Summary**

My question coming into this was “Is the modern-day NBA more selfish than ever before?”. A selfish basketball player can be defined as someone who shoots, doesn’t pass, and doesn’t focus on playing defense. I figured the best way to go about answering my question was to look and see if points scored are up and assists are down, and if steals, blocks, and turnovers are down, meaning that less defense is being played.

My first step was to do some initial exploratory data analysis. I noticed some cool observations, like there were years that were without data for steals, blocks, turnovers, and 3-pointers. That was because at the time they were not official stats. Steals and blocks didn’t become an official stat until 1974, turnovers not until 1978, and 3-pointers until 1980. I also noticed that someone had played 88 games in a season, which seems like an error since the NBA has 82 game seasons. Looking into that, the player could have been traded and therefore played more games than the allotted 82. Once I realized that, I noticed that in the initial dataset, that player would be represented 3 times, one time for their first team, one time for their second, and a third time that totals it together. I realized that I must clean this data as to not duplicate stats. The data treats this as 3 separate players, so in those situations, I had to use the totals of those players and remove the lines where they were broken out by each team. Also, I was easily able to identify seasons that were shortened due to lockouts where the teams played less games. 1999 and 2012 had lockouts. In addition, I found out the shot clock was introduced in the 1955 season, which will not have much of an impact on the data since the data started in 1950. Finally, as I loaded the data in, any blank value came across as a ‘NaN’, so I made those a ‘0’ in the dataset. Overall, I learned quite a bit from performing EDA.

There are no basketball statistics that measure selfishness of a player, so assumptions had to be made. If someone scores a lot of points and does not have a lot of assists, they could be labeled a selfish player, but they might not have any assists because their teammates missed all their potential opportunities. Over the course of seasons and seasons of data, that should average out and not be a factor. The main challenge I faced was trying to manipulate the data to get it into the desired format. I tried to group each variable by decade and see if there were any trends on a decade to decade basis. If I was able to create a data frame summed by year, as opposed to decade it might have been different. However, the decades are sums of the years, so I don’t anticipate that would have changed results too drastically overall. Another challenge that I started to realize towards the end of the project was that maybe my dataset wasn’t the best choice, such as when conducting the hypothesis test. I don’t feel like I missed much during the analysis. I felt I had the variables I needed to try and answer the question to the best of my ability, it was just a tough question to ask and a tough dataset to work with for what I was trying to answer.