**DSC 530**

**Term Project - Using 2020 NBA Data to Predict Positions**

**David Brehm**

This EDA project on NBA player stats was very informative. The correlation matrix in particular I thought was very insightful. A lot of the correlations seemed clear having a background in the sport, but some were more surprising. For example, frontcourt players tend to get more blocks because they are taller and have the role of rim defense. It would make sense that this role also gets more rebounds. As a player’s assists increase, their turnovers will usually increase as well from the assumed additional playmaking responsibilities. One that was a bit surprising though was points and turnovers being the highest correlation. That is probably a function of a larger offensive role on the team similar to the relationship between assists and turnovers, but I did not expect it to be that strong.

The model only achieving 50% accuracy was a bit of a disappointment. In retrospect I believe there are a couple of things that could be done to improve this.

* My assumption to just filter by games started might have not been the right direction. Even though that gets the dataset closer to players having similar roles, there is still a range of minutes played which affects their stats per game. If the predictors were scaled relative to minutes played, this might improve accuracy.
* Feature selection. I’ve heard of something called Recursive Feature Elimination which might have given a different seven predictors than the ones I chose. There could be variables missing in this analysis that would have been better predictors.

One aspect of this EDA project that I did not really understand was the result of the hypothesis test I chose. The process of shuffling one variable so that they are unrelated makes sense, however when the correlation iterations came back with a wide range, I was not exactly sure how to interpret that. The only textbook example had a result that was consistent with the initial correlation.