DSC 423 - Winter 2021

Sports/Video Game Group: El Dorado

Ken Thomas Wei Tong Su (James Su) Zhong Xie
Chunnan Liu Devin Carroll

We started our data analysis process by splitting the variables into logically consistent groups: shooting metrics, defensive metrics, non-shooting offensive metrics, and draft-related variables. We then explored the relationship the variables in each group had with our initial lines of inquiry: field goal percentage, PER, and usage rate. Unfortunately, we did not find many variables that looked like predictors for those variables. We did find that several of our variables were strongly correlated with a player's total points in a season and decided to pursue that as our dependent variable instead. A few interesting trends we found in our variables is that many of them are clearly right skewed and have many outliers. It is our belief that because the distribution for Minutes Played is heavily right-skewed, many players in our dataset do not have enough playing time to accumulate counting statistics, such as Points, Assists, Blocks, etc., at a level comparable with the smaller number of players with consistent playing time. These journeymen bench players are shrinking the Interquartile Range making the better players in the league appear as outliers and skewing the distribution to the right. Based on our analysis of the relationship between variables we have narrowed down our focus on variables to: Games, Games Started, Minutes Played, Assists, Steals, Turnovers, Free Throws Made, Free Throw Attempts, 2-Point Attempts, 3-Point Attempts, and Field Goal Attempts.

We believe that players with high values of Usage Rate (USG) correlate strongly with their total points in a season because it helps represent that player's shooting volume over the course of the season. To be a top scorer in the NBA, a player needs to shoot quite often and therefore take a high percentage of a team's shot. The correlation between usage rate and our response variable points is 0.49.

The easiest way to score in basketball comes from free throws. Free throws (FT) are granted when fouls are committed by the other team. Although free throws are only granted one, two or three at a time, we wish to understand the relationship between free throws on the overall points scored. The correlation between free throws and points scored is 0.93.

Rebounds, both offensive and defensive (ORB and DRB), put the ball in that player's hands to initiate the offense. Offensive rebounds lead to easier chances near the opponent's basket while defensive rebounds often lead to fast break opportunities which are also considered to be efficient chances. It is our opinion that a player's total rebounds impact a player's total points by giving that player easy scoring opportunities in many cases.

The correlation between assists (AST) and our response variable points (PTS) is 0.74. This correlation is likely driven by opportunity. Individuals in a position to score points are also able to assist in scoring points. Assists is correlated with explanatory variables such as games (G) and minutes played (MP), which support the case for opportunity being a contributing factor.

Minutes played is highly correlated to our response variable points (PTS), with a correlation of 0.99. It is also correlated to other variables that are highly correlated to points. The amount of time spent in a game dictates points made and other variables.

Like the minutes played variable, the correlation between Games (G) and points is 0.93. Players with higher games played tended to have a higher chance of scoring more points, and other variables correlated to points. Games started are the same as Games (G). However, the

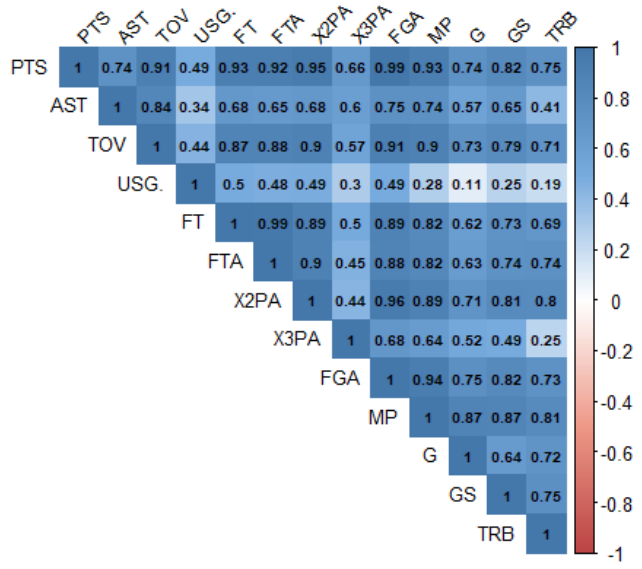
difference is the number of games played as a starter. We wanted to observe if games started will contribute an increase on variables that are highly correlated to points.

The correlation between turnovers (TOV) and our response variable points (PTS) is 0.91. Players who score more points are going to be targeted by the defense more often and as a result turnover the ball more often. Turnovers are also highly correlated with other shooting variables such as two-point attempts (X2PA) and field goal attempts (FGA).

The correlation between 2-point field goal attempts (X2PA) and our response variable points (PTS) is 0.95 which is highly correlated to our response variable. Shooting for 2 points is the main type of lead scoring method, so players will make more points when 2-point attempts get higher.

The correlation between 3-point field goal attempts (X3PA) and our response variable points (PTS) is 0.66. Shooting for 3 points is harder than other scoring methods with a low goal rate because the player must shoot behind the 3 points line which is much further. Also, normally only PG,SF and SG have the chance or ability to make 3-point attempts, so X3PA had lower correlation between PTS than other variables.

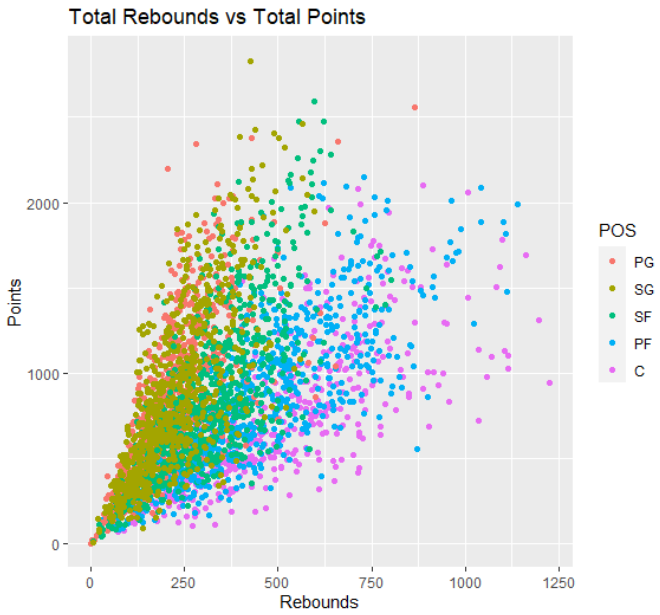
The correlation between field goal attempts (FGA) and our response variable points (PTS) is 0.99. Field goal attempts which include both 2-point field goal attempts and 3-point field goal attempts. FGA has a high correlation with PTS because it includes all main types of lead scoring methods.



A correlation matrix was created using our response and predictor variables. The thirteen variables chosen for the matrix were selected from an earlier matrix screening of approximately 56 variables. They were selected because they all demonstrated a high correlation with our response variable, points (above 0.6). Among the explanatory variables there exist several high correlations. Minutes played for instance tends to correlate positively with variables that result from more opportunity – field goal attempts, assists and turnovers would be examples. An

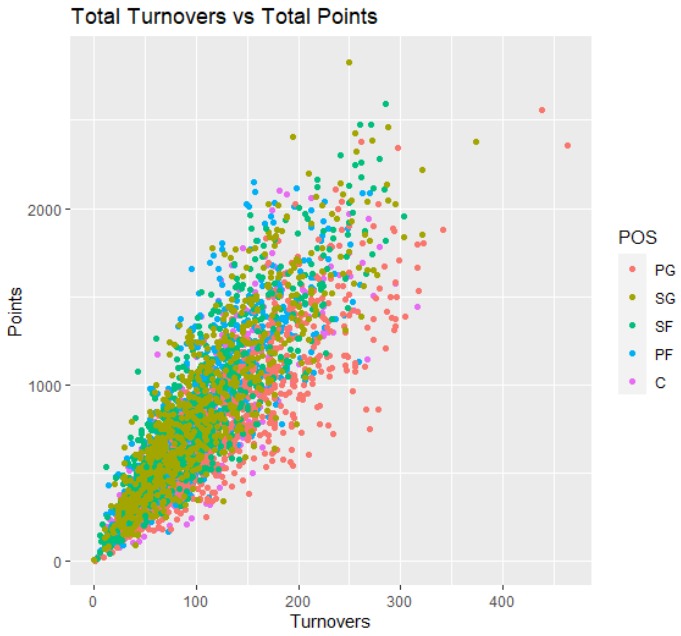
interesting note is that there were very few negative correlations amongst the 56 variables examined in this study.

Total Rebounds vs. Total Points



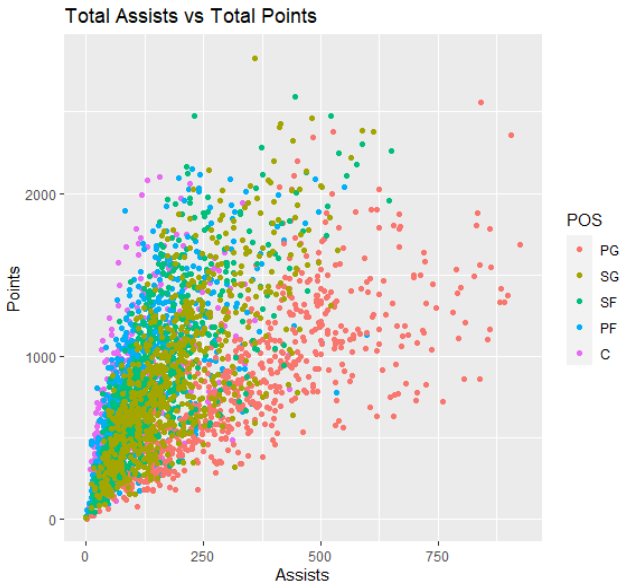
There appears to be a positive but hard to see a linear relationship, so we might consider a 2nd order term. PG, SG and SF had larger slopes than PF and C, which means that PG,SG and SF can make more points but less rebounds than PF and C because of the difference of height.

Total Turnovers vs. Total Points



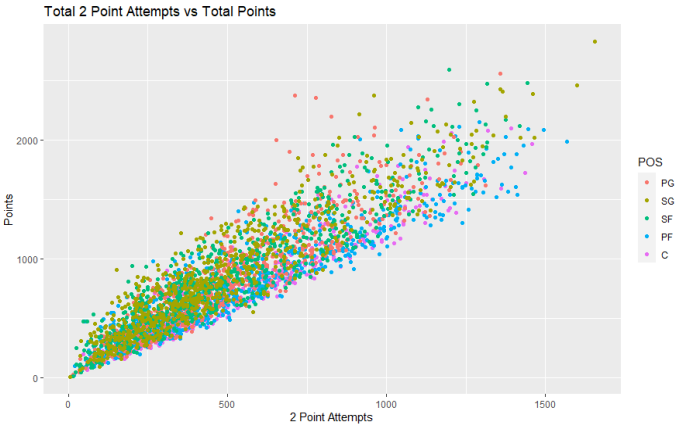
There appears to be a clear positive linear relationship. From this plot, PG had a smaller slope than other positions because PG usually has more time to keep the ball and has a larger possibility to lose the ball to opponents. TOV is a good variable that we might consider in the future.

Total Assists vs. Total Points



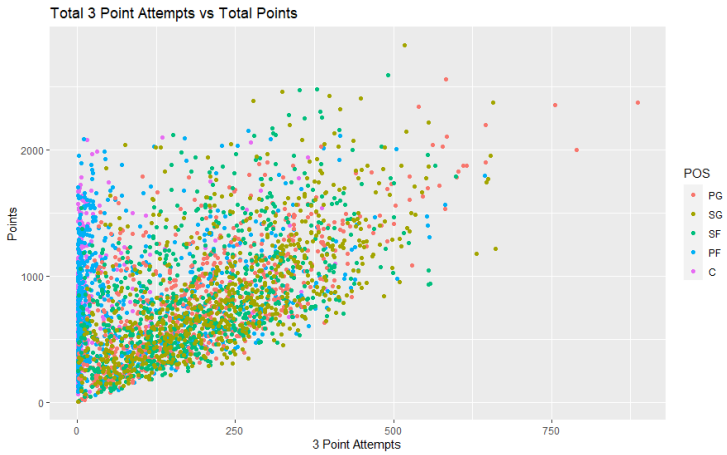
This plot also appears positive, but it is hard to see a linear relationship, which might benefit with 2nd order terms. We can see from this plot that PG obviously had more assists than other positions because their job is helping teammates to make points.

Total 2 Point Attempts vs. Total Points



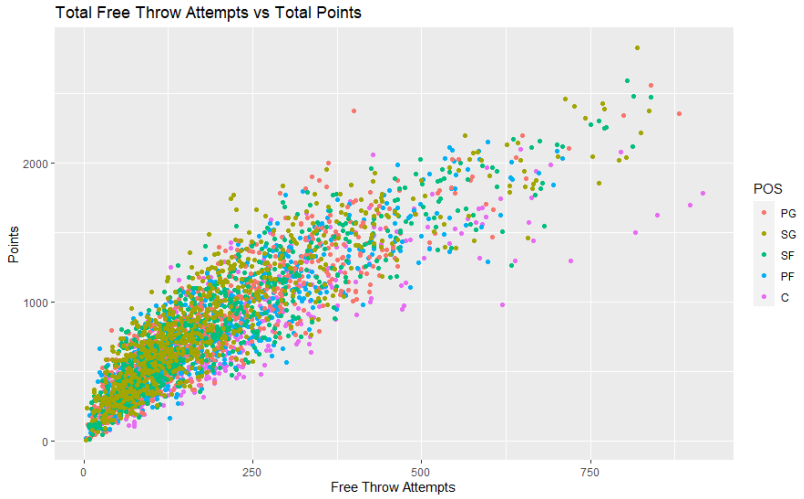
This plot shows that the relationship between 2-point attempts and total points is positive and strong, likely because two points is among the easiest ways to score points. Even Centers (C) of the past knew how to get two points in the paint. As a result, there is little difference amongst positions in this plot.

Total 3 Point Attempts vs. Total Points



This plot shows the relationship between 3-point attempts and total point is positive with moderate strength. This is likely because Centers (C) and Power Forwards (PF) in earlier times did not shoot the three-point shot. If we ignore the Center(C) and Power forward(PF), the strength would be stronger than the picture. Moreover, since the 2010s the three points became more emphasized, resulting in outliers.

Total Free Throw Attempts vs. Total Points



The relationship between total free throw attempts and total points is also positive and strong. The free throw attempts come from fouls committed by the other team, so with more free throw attempts, the total points are likely to increase. Some centers (C) have high free throw attempts because they have strong influence in the paint, but are not good at shooting the ball, such as Shaquille O'Neal. Their points from free throws were not in the linear range. However, this plot demonstrates that free throws can amount to many points.

Our team plans on using multiple regression to predict a player's total points in a season, the dependent variable, with a set of roughly thirteen independent variables. The variables we believe to be most impactful in predicting total points are Games, Games Started, Minutes Played, Assist, Steals, Turnovers, Free Throws Made, Free Throw Attempts, 2-Point Attempts, 3-Point Attempts, and Field Goal Attempts. We anticipate some of these variables will overlap, however, we have chosen to include all three so that we can explore those interactions and redundant predictive power before making a final decision. The specific areas we anticipate redundancy and possibly an interaction are: Games, Games Started, and Minutes Plays as they all roughly suggest a player's playing time then 2-Point Attempts, 3-Point Attempts, Field Goals Attempts, and Usage Rate because all four variables speak roughly to a player's shooting volume. Usage Rate in particular we believe is worth exploring as part of an interaction term with the other variables.