Section 1 - Introduction

The NBA is comprised of 30 teams each playing 82 games for a total of 1,230 games played per season. We will be investigating the connection between teams' performance statistics in the following categories: possessions, effective field goal rates, free throw rates, offensive/defensive rebounds, and turnovers. Our group has conducted an exhaustive analysis of many variables that can be measured in an NBA game to determine the total points that might be scored in a matchup of any two teams in the NBA. We hope that you find our results interesting and will be able to follow along with our methods, analysis, and discussion to come to the same conclusions. In order to measure these variables, we have broken them down into simplified categories to display their impact: possessions, field goal attempts, offensive/defensive rebounds, and turnovers. Within these categories are variables that show different data types which can help us predict the total points scored in a given game. If we don't account for how these variables impact the total points scored in the game, then our prediction will not be valid. This report will use regression analysis to identify trends to propose a model for estimating a team's offensive and defensive measurements to normalize a team's statistics with respect to an opponent in each game. We believe that these attributes will be the building blocks to help predict the number of points scored in a game. It is possible that some combinations of other categorical predictors would yield even better results and that can be worth further consideration in the future.

Section 2 - Data Preparation

The dataset contains all 30 NBA teams that were in the league from 2016-2018 (two full seasons). Half of the data points are related to how team A performs on offense and defense. The other half is represented how team B performs against each of the other 29 possible opponents. We gathered data from the 2016-2017 and the 2017-2018 seasons and combined them into one dataset using weighted averages based on the number of possessions in each season.

In order to have the best results we choose total possessions (T Poss off/def), Avg Total seconds (T Seconds off/def), Avg Total points per possession (T Points off/def), and % of offensive/defensive possessions after made the shot (S % off/def), Avg seconds after made shot on offense/defense (S seconds off/def), Avg Points per possession offense/defense after made shot (S points off/def). Knowing the total amount of possessions on offense and defense and breaking them down to the number of seconds after the shot, rebound, and turnovers will help us in figuring out the total number of points that each team will score in a given game.

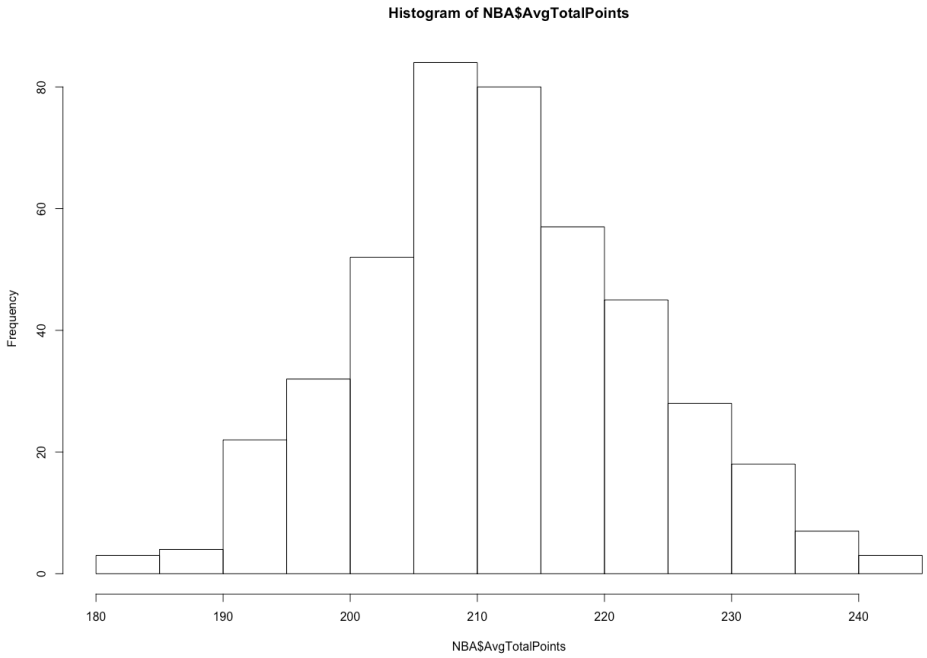
For each team, we will be looking at the offensive and defensive sides to determine how many points a team could score and allow an opponent to score to determine the total number of points scored in a future basketball game. The data that we will consider is the offensive or defensive: points per 100 possessions, effective field goal percentage, turnover percentage, rebounding percentage, and free throw rate per 100 field goal attempts.

After processing our dataset, we have 30 teams as data objects and 130 independent variables to predict how many points will be scored in a future game (dependent variable). We believe all the dependent variables in our data set can be used to predict points scored by any given team.

Section 3 - Data Analysis

Due to a large number of variables determining the total average points scored per game, we were able to narrow down the measurements by breaking them down into the following categories: possessions, effective field goal rates, free throw rates, offensive/defensive rebounds, and turnovers. Within these categories, we decided that some attributes were more worth pointing out than others. First, we start by showing that the AvgTotalPoints across the whole sample are just about normally distributed.

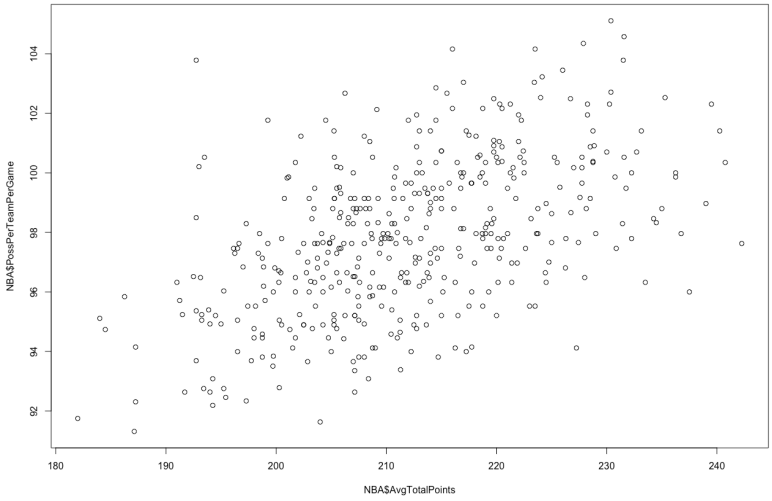
Figure 1



In Figure 1's histogram, we can see that there is a normal distribution of Total Average Points Scored. There is a slight skew to the right in the distribution, meaning there are a higher number of teams that score above mean than below. The histogram in figure 1 shows 68% of the total average points scored fall between 200 and 220, about 95% fall between 190 and 230, and about 99.7% fall between 180 and 240.

The next thing that we wanted to do was show the impact of the number of possessions in a game on the observed AvgTotalPoints (dependent variable). To do this we created a scatter plot and as you can see from figure 2 there is a general upward trend in this relationship.

Figure 2



There seems to be a pretty strong positive relationship between the projected number of possessions in a game and the average total points scored per game. This is because the more opportunities to score there are the more potential points there are to be scored. We believe that this data combined with each team's percent of possessions ending in a made shot, rebound or turnover, as well as the team's effective field goal percentage, can provide an even clearer picture in how many points two teams would score combined in any given matchup.

	AvgTotalPoints
AvgTotalPoints	1.00000000
PerAfterTurnoverOffTeamOne	0.19910065
SecondsAfterTurnoverOffTeamOne	-0.18999152
PointsAfterTurnoverOffTeamOne	0.37551602
PerAfterTurnoverDefTeamOne	0.16540761
SecondsAfterTurnoverDefTeamOne	-0.05984457
PointsAfterTurnoverDefTeamOne	0.19573145
TurnoverPerDefTeamOne	0.20557568
PerAfterTurnoverOffTeamTwo	0.11589156
SecondsAfterTurnoverOffTeamTwo	-0.31904994
PointsAfterTurnoverOffTeamTwo	0.13720359
PerAfterTurnoverDefTeamTwo	0.02436625
SecondsAfterTurnoverDefTeamTwo	-0.18704579
PointsAfterTurnoverDefTeamTwo	0.20391136
TurnoverPerOffTeamTwo	0.24967598
TurnoverPerDefTeamTwo	0.16817024

In the correlation matrix above, we ran our turnover variables to see what their correlation was in regards to the total average points scored. A correlation of 1 means the variables compare perfectly to each other while a correlation of -1 means the opposite. We can use this information to determine what specific variables will be useful or not useful in the future.

Section 4 - Model Building

Introduction

On a cold, Wednesday, November 22nd, 1950 in Minneapolis, Minnesota, the dominant Minneapolis Lakers played against the Fort Wayne Pistons. (Blitz) The Lakers were led by a superstar by the name of George Mikan (DePaul graduate). (Blitz) The Fort Wayne Pistons knowing that they were no match for the Lakers came up with a brilliant, yet simple plan to hold onto the ball as long as possible so that if the Lakers could not get possession of the ball, they would not be able to score. Everyone who was not part of the Pistons, including the referees, fans, and Lakers, all became frustrated with the actions of the Pistons and subsequently booed them. At half time, the Lakers led with a score of 13 to 11.(Blitz) The second half was even worse where the Lakers themselves decided they had enough of the Pistons antics and they themselves wanted to hold onto the ball believing the mantra "if you can't beat'em, join 'em." The two teams ended up scoring a paltry 13 combined points in the second half. (Blitz) The Fort Wayne Pistons ended up winning the game by a score of 19 to 18, leading to the lowest scoring basketball game in NBA history but also ushered in an era of low scoring basketball games. This was a huge nightmare for the NBA and threatened its very existence. Low possession basketball with minimal actions in shooting, field goals, and free throws, rebounds, and turnovers were a recipe for disaster for generating fan interest. This game was the reason why a 24-second shot clock was added to the game of basketball and why modern basketball games are offensively driven with the more points scored generally believed to be a more "exciting" game by the fans. Our group wanted to look into various variables such as possessions, field goals, and free throw attempts, rebounds, and turnovers to see if we could determine the total points that would be scored in a matchup of any two NBA teams.

Throughout the NBA season, there are a wide array of opponents. This leads to a variation of points being scored in a given game throughout the season. The problem we are facing is trying to predict the total points scored between two teams given their difference in variables that lead to predicting their points scored. In order to measure these variables, we have broken them down into simplified categories to display their impact: possessions, effective field goal rates, free throw rates, offensive/defensive rebounds, and turnovers. Within these categories are variables that show different data types which can help us predict the total points scored in a given game. If we do not account for how these variables impact the total points scored in the game, then our prediction will not be valid.

When collecting the data, we wanted to focus on how the pace of play for each team affects the number of points that two teams would score combined in a game. This is a different study than most standard datasets can provide, therefore we went to a few different sources to create our dataset. The three websites that we used to compile our data were impredicable.com, cleaningtheglass.com, and oddshark.com. Each of these sites provided some unique pieces of information for our study. Impredicable provided the per possession results for each team depending on if the opposing team made a shot the previous possession, missed a shot or turned the ball over. Cleaningtheglass provided us some statistics on how effective a team is in the categories of shooting percentage, offensive rebounding percentage, and free throw rate. And finally oddshark provided us with previous game results between each of the teams in the league. Once we found these datasets we scraped them from their respective websites and combined

them into a single dataset where each row is one matchup between two teams and the columns are their respective stats in each of the independent variables.

Our group will try to determine the total number of points scored by any two teams when they meet for an NBA game using the data from the 2016-2017 and 2017-2018 seasons. We will be focused primarily on looking at independent variables of each team's two seasonal average points scored, possessions, effective field goal rates, free throw rates, offensive and defensive rebounds, and the total number of turnovers per game. We have divided the project into parallel but distinct lines of work by each taking some independent variables to determine the total points that will be scored in an NBA game. No one will be relying on anyone else to complete their work and we all look forward to reaching our individual milestones, as stated in the following paragraphs, to give each of us an opportunity for you to demonstrate the skills we have learned in the class.

The main goal of this project is to predict the average total points scored between two teams in a game. We will be trying to calculate the total average points scored in future NBA games based on statistics from two NBA seasons. Shimu Wan will work with total points scored after a shot, offensive/defensive rebound, turnover.

Field goal and free throw rates of each team in a basketball matchup gives us a good overall sense of how many total points might be scored by each team. The higher the field goal rates, the more likely you are to increase your points total. These are important variables that will help our group make a multiple regression model to determine the total points scored between two teams in a basketball game. Kimoon Ryu will work with the effective field goal and free throw rate variables and try to incorporate them into our group's model to determine the total points scored between two teams in a basketball matchup.

Rebounds are among the most important aspects of winning in basketball games. Defensive rebounds limit an opponent's offense to one shot and offensive rebounds will add extra possessions. Although rebounds do not contribute directly to a team's ability to score, they play a vital role in providing opportunities to do so. For example, let's imagine a team having a field goal percentage is 50% and also averaging 8 offensive rebounds per game. Simply based on these statistics a team can potentially an additional 8-12 points hustling for rebounds. Thomas Le will look into rebounds to build a model to identify correlations in predicting the overall points scored.

In an NBA game, turnovers have shown to have an impact on the number of points that can be scored. Christian Craig will work with models in this category to show turnovers' impact on the total points scored in a given game. Despite the percentage of turnovers going down over the years. They still have a significant impact on the total points scored in a game. This is noted in an article by Jared Dubin of FiveThirtyEight when he writes about Dean Oliver's findings for the impact on offensive efficiency when stating "Oliver pegged turnover rate as being responsible for 25 percent of the variance in offensive efficiency" (Dubin). This shows the relevancy of this category, and we will be worth further exploring how this impacts our dependent variable.

Josh Hall will be responsible for trying to create a model with all of our data and working as a comparison for the rest of our individual models. He will be working with the possessions,

effective field goals and free throws rates, offensive and defensive rebounds, and the total number of turnovers per game to come up with a model that determines the average total points that will be scored between any two NBA teams. We look forward to comparing Shimu's, Kimoon's, Thomas', and Christian's models to Josh's model to determine how similar they are to one another.

Shimu Wan

I am focused on looking at all the points scored after an action such as a rebound, turnover, shot for the offensive and defensive teams.

Points Scored Model 1

```
Residuals:
     Min       1Q   Median       3Q      Max
-28.2093  -6.1632  -0.0376   6.2758  25.7124

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                        71.1099    49.3996   1.439 0.150748
Book1$PointsAfterMadeShotOffTeamOne 257.2194    73.8119   3.485 0.000544 ***
Book1$PointsAfterReboundOffTeamOne   68.2301    54.1993   1.259 0.208765
Book1$PointsAfterTurnoverOffTeamOne  72.8990    17.1909   4.241 2.74e-05 ***
Book1$PointsAfterMadeShotDefTeamOne  -5.0626    65.7670  -0.077 0.938678
Book1$PointsAfterTurnoverDefTeamOne   9.4761    16.3454   0.580 0.562397
Book1$PointsAfterReboundDefTeamOne  -25.1073    52.1545  -0.481 0.630478
Book1$PointsPossOffTeamOne           -2.2044     1.1736  -1.878 0.061017 .
Book1$PointsPossDefTeamOne           -0.2768     1.1421  -0.242 0.808593
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.878 on 426 degrees of freedom
Multiple R-squared:  0.2321,    Adjusted R-squared:  0.2177
F-statistic:  16.1 on 8 and 426 DF,  p-value: < 2.2e-16
```

Next, I checked all the variables and started to remove some variables with much non-significant t-tests values (> 0.05), which resulted in me keeping PointsAfterMadeShotoffTeamOne, PointsAfterTurnoveroffTeamOne, and PointsPossoffTeamOne. (See below)

Points Scored Model 2

```
Residuals:
     Min       1Q   Median       3Q      Max
-27.4535  -6.5667  -0.2287   6.4384  26.2568

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      26.3046    16.7019   1.575 0.116002
Book1$PointsAfterMadeShotOffTeamOne 173.5556    46.6017   3.724 0.000222 ***
Book1$PointsAfterTurnoverOffTeamOne  60.1031    13.5922   4.422 1.24e-05 ***
Book1$PointsPossOffTeamOne           -0.6363     0.4855  -1.311 0.190678
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.858 on 431 degrees of freedom
Multiple R-squared:  0.2261,    Adjusted R-squared:  0.2207
F-statistic: 41.98 on 3 and 431 DF,  p-value: < 2.2e-16
```

When running the VIF, PointPossoffTeamOne had a result of 11.35. After checking that the removal of PointPossoffTeamOne wouldn't impact other independent variables significantly, it was removed for multicollinearity reasons.

Points Scored Model 3

```
> vif(m2)
Book1$PointsAfterMadeShotOffTeamOne Book1$PointsAfterTurnoverOffTeamOne
                           9.698265                                1.664830
         Book1$PointsPossOffTeamOne
                          11.358060
```

Points Scored Model 4

```
Residuals:
    Min      1Q  Median      3Q     Max
-26.769  -6.419  -0.319   6.385  26.736

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                        26.00      16.71   1.556    0.121
Book1$PointsAfterMadeShotOffTeamOne   116.84      17.30   6.753 4.70e-11 ***
Book1$PointsAfterTurnoverOffTeamOne    52.17      12.18   4.283 2.27e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.867 on 432 degrees of freedom
Multiple R-squared:  0.223,     Adjusted R-squared:  0.2194
F-statistic:    62 on 2 and 432 DF,  p-value: < 2.2e-16
```

In an effort to try to improve our model, I started to create second-order terms to see if there was an improvement in adjusted R-Squared. This involved creating a new variable by squaring the original variable. When adding second-order terms, the F-test and T-test became insignificant and the adjusted R-squared did not increase. (See Figure below) The final model will now go back to the Figure above. The final model only has the variables PointAfterMadeShotoffTeamOne and PointAfterTurnoveroffTeamOne. In my final model, the F-Test is significant, a p-value less than 0.05, which means at least one of my betas is not equal to

zero and the T-Test says all the individual variables are significant, a p-value less than 0.05, and we can reject the null hypothesis that the individual beta is equal to zero. The adjusted R-squared of 0.2194 means that 21.94% of all the variability of the average points scored in a game is due to points scored after an action on the defensive and offensive sides.

Points Scored Model 5

```
Residuals:
    Min      1Q  Median      3Q     Max
-26.966  -6.377  -0.219   6.431  26.954

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                       260.72     678.95   0.384    0.701
Book1$PointsAfterMadeShotOffTeamOne  -246.63    1282.80  -0.192    0.848
Book1$PointsAfterTurnoverOffTeamOne   -20.55     351.13  -0.059    0.953
Book1$PAMSOT2                         173.29     613.52   0.282    0.778
Book1$PATOTO2                          29.79     145.81   0.204    0.838

Residual standard error: 9.888 on 430 degrees of freedom
Multiple R-squared:  0.2233,    Adjusted R-squared:  0.2161
F-statistic:  30.9 on 4 and 430 DF,  p-value: < 2.2e-16
```

Kimoon Ryu

I am focused on looking at the free throw and effective field goal rates as my explanatory variables. I have a feeling that the more attempts you have at the basket that there will be a high correlation with the number of points scored. I will run regression in R to verify my claims.

I started the analysis of our data for field goals and free throws by adding all of the data and running a summary of the data. This is model 1. This model has an adjusted R-squared of 0.2195 which means that 21.95% of the variation in the total points scored between two NBA teams is explained by the variables. The F-test being < 2.2e-16 means that at least one of the betas is not zero and significant. I see many individual T-test results that are significant to not significant (> 0.05).

Field Goals Model 1

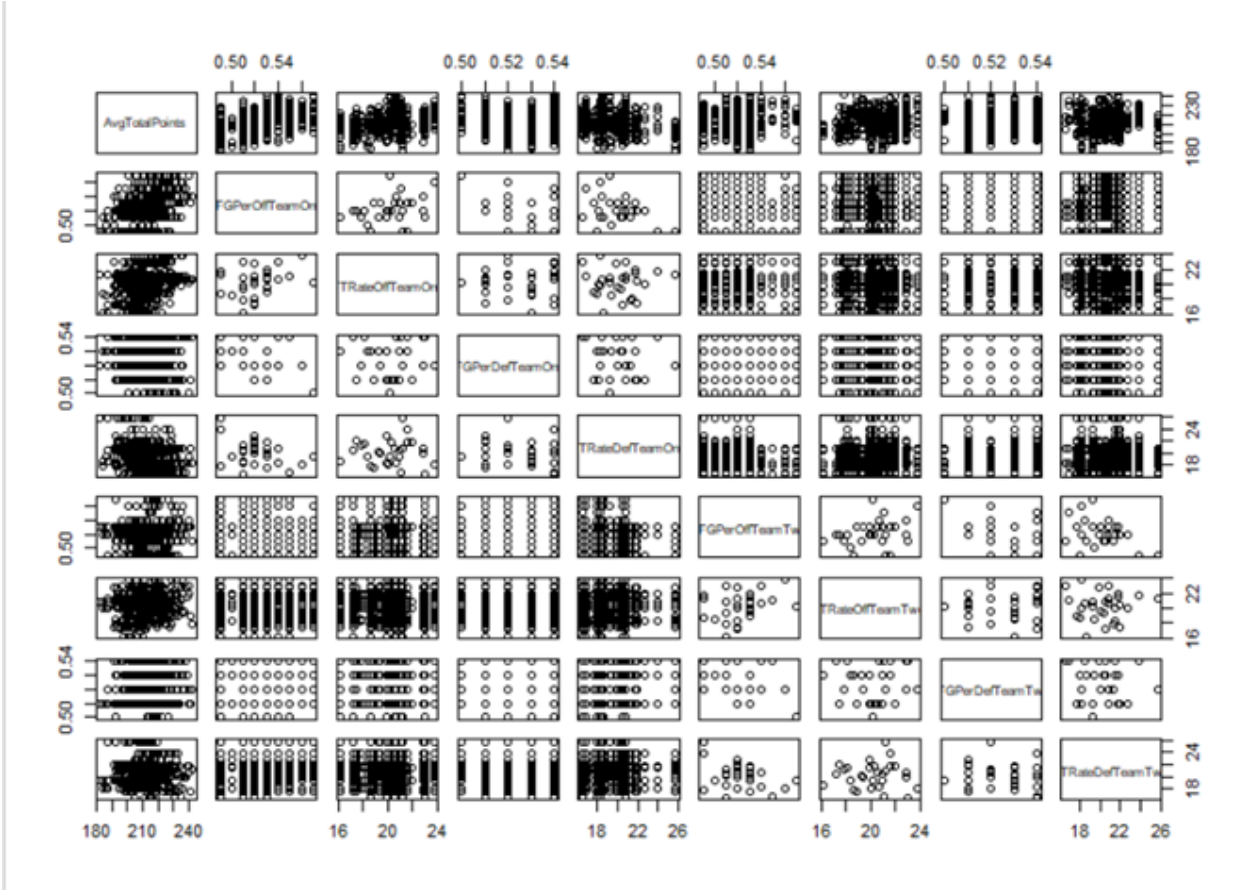
```
> m1 <- lm(AvgTotalPoints ~., data = NBA)
> summary(m1)

Call:
lm(formula = AvgTotalPoints ~ ., data = NBA)

Residuals:
     Min       1Q   Median       3Q      Max
-23.4626  -6.1079  -0.7437   6.8980  31.1914

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -15.8915    52.9490  -0.300  0.76423
eFGPerOffTeamOne 148.7950    29.2342   5.090 5.39e-07 ***
FTRateOffTeamOne   1.0145     0.2652   3.826  0.00015 ***
eFGPerDefTeamOne -32.9511    49.1116  -0.671  0.50262
FTRateDefTeamOne  -0.7456     0.2747  -2.714  0.00692 **
eFGPerOffTeamTwo  87.2210    37.4632   2.328  0.02037 *
FTRateOffTeamTwo   0.7399     0.3071   2.409  0.01641 *
eFGPerDefTeamTwo 179.3606    42.8181   4.189 3.41e-05 ***
FTRateDefTeamTwo   0.3609     0.2890   1.249  0.21244
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.866 on 426 degrees of freedom
Multiple R-squared:  0.2339,    Adjusted R-squared:  0.2195
F-statistic: 16.26 on 8 and 426 DF,  p-value: < 2.2e-16
```

I will next check for correlations between the variables by running a correlation matrix of the model. And then I will plot the various independent variables in the data to see if I could see any correlations. We are looking for correlations close to 0.9 and in this set of data we see that none exist. Another way to visualize correlations is by making a plot of all the variables present.

Field Goals Model 2

```
> cor(NBA)
                 AvgTotalPoints eFGPerOffTeamOne FTRateOffTeamOne eFGPerDefTeamOne FTRateDefTeamOne
AvgTotalPoints       1.00000000     0.3711488434       0.27201404      -0.06548315     -0.211554560
eFGPerOffTeamOne     0.37114884     1.0000000000       0.33834413      -0.31552115     -0.243022995
FTRateOffTeamOne     0.27201404     0.3383441320       1.00000000       0.04055682     -0.079049006
eFGPerDefTeamOne    -0.06548315    -0.3155211490       0.04055682       1.00000000     -0.359940493
FTRateDefTeamOne    -0.21155456    -0.2430229949      -0.07904901      -0.35994049      1.000000000
eFGPerOffTeamTwo     0.04740112    -0.0319797543      -0.03009414       0.04441135     -0.047254726
FTRateOffTeamTwo     0.14355852     0.0090207819      -0.03475282      -0.02467504      0.007255198
eFGPerDefTeamTwo     0.16072372    -0.0004671354       0.02237882      -0.01628795     -0.028489228
FTRateDefTeamTwo     0.03991152     0.0356735970      -0.01237178      -0.02921829      0.027123475
                 eFGPerOffTeamTwo FTRateOffTeamTwo eFGPerDefTeamTwo FTRateDefTeamTwo
AvgTotalPoints       0.04740112      0.143558519      0.1607237244       0.03991152
eFGPerOffTeamOne    -0.03197975      0.009020782     -0.0004671354       0.03567360
FTRateOffTeamOne    -0.03009414     -0.034752821      0.0223788160      -0.01237178
eFGPerDefTeamOne     0.04441135     -0.024675043     -0.0162879473      -0.02921829
FTRateDefTeamOne    -0.04725473      0.007255198     -0.0284892282       0.02712348
eFGPerOffTeamTwo     1.00000000      0.252026443     -0.3655743320      -0.44994395
FTRateOffTeamTwo     0.25202644      1.000000000     -0.0049515391       0.10286663
eFGPerDefTeamTwo    -0.36557433     -0.004951539      1.0000000000       0.10044190
FTRateDefTeamTwo    -0.44994395      0.102866634      0.1004418959       1.00000000
```

Field Goals Model 3



A check for multicollinearity was done by testing for the VIF (Variance Inflation Factor) of the independent variables. We see that none of the explanatory variables are close to the 10 value that causes us concern.

Field Goals Model 4

```
> vif(m1)
eFGPerOffTeamOne FTRateOffTeamOne eFGPerDefTeamOne FTRateDefTeamOne eFGPerOffTeamTwo FTRateOffTeamTwo
        1.549090         1.175135         1.489239         1.392743         1.656896         1.158386
eFGPerDefTeamTwo FTRateDefTeamTwo
        1.183997         1.354378
```

I used the backward selection process to eliminate some independent variables to make the model simpler and have a better adjusted R-squared value is possible. The explanatory variable with the highest T-test value (0.50262) was DeFGPerDefTeamOne. I eliminated that one and found that the R-squared goes up to 0.2205 from 0.2195. Next, FTRateDefTeamTwo (T-test = 0.215156) will be eliminated to see what happens. I came to the best model possible with only first-order terms using free throw and effective field goal rates for determining the average points scored. (Shown below)

Field Goals Model 5

```
> m3 <- lm(AvgTotalPoints ~. -eFGPerDefTeamOne - FTRateDefTeamTwo, data = NBA)
> summary(m3)

Call:
lm(formula = AvgTotalPoints ~ . - eFGPerDefTeamOne - FTRateDefTeamTwo,
    data = NBA)

Residuals:
     Min       1Q   Median       3Q      Max
-23.0066  -6.0570  -0.6762   6.6534  30.7345

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      -19.5889    35.9998  -0.544 0.586629
eFGPerOffTeamOne 158.9938    25.6970   6.187 1.43e-09 ***
FTRateOffTeamOne   0.9713     0.2602   3.732 0.000215 ***
FTRateDefTeamOne  -0.6566     0.2408  -2.726 0.006667 **
eFGPerOffTeamTwo  64.0693    32.5867   1.966 0.049931 *
FTRateOffTeamTwo   0.8417     0.2966   2.837 0.004765 **
eFGPerDefTeamTwo 174.8478    42.5591   4.108 4.78e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.866 on 428 degrees of freedom
Multiple R-squared:  0.2303,    Adjusted R-squared:  0.2195
F-statistic: 21.35 on 6 and 428 DF,  p-value: < 2.2e-16
```

In the model above, Offensive eFG Team One, Offensive FT Rate Team One, Defensive FT Rate Team One, Offensive eFG Rate Team Two, Offensive FT Rate Team Two, and Defensive eFG Team Two are responsible for 21.95% of the variation of the points scored by two teams in the NBA. Since the F-Test is significant (< 0.05) and the T-test says all the individual variables are significant (< 0.05), we can say that this model is a legitimate possibility.

No variables needed to be transformed because all of the variables are numerical variables and not categorical.

I want to try interaction terms between the Team Two defensive eFG/FT with the Team One offensive eFG/FT as well as Team One defensive eFG/FT with TeamTwo offensive eFG/FT. I have a feeling that the defense of one team could influence the offensive production of the eFG/FT of the other team. Unfortunately, none of the interaction terms (shown below) were significant so I slowly removed one at a time to see if the significance would increase. This did not happen and those terms were eliminated from the model.

Field Goals Model 6

```
NBA$eFGoTeam1dTeam2 <-  NBA$eFGPerOffTeamOne * NBA$eFGPerDefTeamTwo
NBA$eFGdTeam1oTeam2 <-  NBA$eFGPerDefTeamOne * NBA$eFGPerOffTeamTwo
NBA$FToTeam1dTeam2 <-  NBA$FTRateOffTeamOne * NBA$FTRateDefTeamTwo
NBA$FTdTeam1oTeam2 <-  NBA$FTRateDefTeamOne * NBA$FTRateOffTeamTwo
```

Field Goals Model 7

```
> m4 <- lm(AvgTotalPoints ~.-eFGPerDefTeamOne - FTRateDefTeamTwo, data = NBA)
> summary(m4)

Call:
lm(formula = AvgTotalPoints ~ . - eFGPerDefTeamOne - FTRateDefTeamTwo,
    data = NBA)

Residuals:
     Min       1Q   Median       3Q      Max
-23.4560  -6.1025  -0.7684   6.8636  31.1301

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       26.24250  554.62835   0.047   0.9623
eFGPerOffTeamOne  34.61565 1055.53093   0.033   0.9739
FTRateOffTeamOne   0.66814    0.39073   1.710   0.0880 .
FTRateDefTeamOne  -0.30104    2.92190  -0.103   0.9180
eFGPerOffTeamTwo 118.69531   62.62307   1.895   0.0587 .
FTRateOffTeamTwo   1.17796    2.83942   0.415   0.6785
eFGPerDefTeamTwo  64.96547 1054.32410   0.062   0.9509
eFGoTeam1dTeam2  218.48891 2019.35141   0.108   0.9139
eFGdTeam1oTeam2  -63.28929   94.35226  -0.671   0.5027
FToTeam1dTeam2     0.01664    0.01420   1.171   0.2421
FTdTeam1oTeam2    -0.02220    0.14461  -0.153   0.8781
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.891 on 424 degrees of freedom
Multiple R-squared:  0.2336,    Adjusted R-squared:  0.2156
F-statistic: 12.93 on 10 and 424 DF,  p-value: < 2.2e-16
```

I cannot think of a reason why squaring one of the variables would result in being helpful for the modeling, but because I am not an expert in linear regression or basketball, I want to try those variables too. Surprisingly, many of the variables when squared and turned into a second-order variable became significant. (Figure below). I slowly started eliminating second-order variables one by one where the T-test value was not significant.

Field Goals Model 8

```
> m5 <- lm(AvgTotalPoints ~.-eFGPerDefTeamOne - FTRateDefTeamTwo, data = NBA)
> summary(m5)

Call:
lm(formula = AvgTotalPoints ~ . - eFGPerDefTeamOne - FTRateDefTeamTwo,
    data = NBA)

Residuals:
     Min       1Q   Median       3Q      Max
-24.6991  -6.1135  -0.5599   6.6921  31.1799

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.282e+01  1.242e+03  -0.010  0.99177
eFGPerOffTeamOne  -2.346e+03  1.081e+03  -2.169  0.03061 *
FTRateOffTeamOne   1.083e+01  4.320e+00   2.507  0.01254 *
FTRateDefTeamOne   7.179e+00  3.907e+00   1.837  0.06685 .
eFGPerOffTeamTwo  -3.114e+03  1.146e+03  -2.717  0.00686 **
FTRateOffTeamTwo   1.166e+01  5.921e+00   1.969  0.04958 *
eFGPerDefTeamTwo   4.685e+03  4.689e+03   0.999  0.31837
eFGPerOffTeamOneSQ 2.355e+03  1.023e+03   2.303  0.02176 *
FTRateOffTeamOneSQ -2.408e-01  1.078e-01  -2.234  0.02603 *
eFGPerDefTeamOneSQ 2.466e+01  4.890e+01   0.504  0.61429
FTRateDefTeamOneSQ -1.921e-01  9.515e-02  -2.018  0.04419 *
eFGPerOffTeamTwoSQ 3.072e+03  1.096e+03   2.802  0.00531 **
FTRateOffTeamTwoSQ -2.769e-01  1.475e-01  -1.877  0.06122 .
eFGPerDefTeamTwoSQ -4.286e+03  4.475e+03  -0.958  0.33880
FTRateDefTeamTwoSQ  5.510e-03  6.958e-03   0.792  0.42882
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.692 on 420 degrees of freedom
Multiple R-squared:  0.2711,    Adjusted R-squared:  0.2468
F-statistic: 11.16 on 14 and 420 DF,  p-value: < 2.2e-16
```

I ended up with my final model (Figure below). Offensive eFG Team One, Offensive FT Rate Team One, Defensive FT Rate Team One, Offensive eFG Rate Team Two, Offensive FT Rate Team Two, Defensive eFG Team Two, Offensive eFG Team One Squared, Offensive FT Rate Team One Squared, Defensive FT Rate Team One Squared, and Offensive eFG Rate Team Two Squared are responsible for 24.48% of the variation of the points scored by two teams in the NBA. One variable, Defensive FT Rate Team One no longer has a significant T-test, but due to the second-order Defensive FT Rate Team One Squared having a significant T-test, we are keeping the first-order term in the model. For all other variables, since the F-test is significant (< 0.05) and the T-test says all the individual variables are significant (< 0.05), we can say that this model is a legit possibility.

Field Goals Model 9

```
> m6 <- lm(AvgTotalPoints ~.-eFGPerDefTeamOne - FTRateDefTeamTwo, data = NBA)
> summary(m6)

Call:
lm(formula = AvgTotalPoints ~ . - eFGPerDefTeamOne - FTRateDefTeamTwo,
    data = NBA)

Residuals:
     Min       1Q   Median       3Q      Max
-24.4824  -5.8785  -0.7505   7.0761  30.4680

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        1.212e+03  3.985e+02   3.043  0.00249 **
eFGPerOffTeamOne  -2.239e+03  1.042e+03  -2.147  0.03233 *
FTRateOffTeamOne   1.009e+01  4.263e+00   2.367  0.01839 *
FTRateDefTeamOne   6.903e+00  3.806e+00   1.814  0.07042 .
eFGPerOffTeamTwo  -2.826e+03  1.099e+03  -2.571  0.01047 *
FTRateOffTeamTwo   6.463e-01  3.007e-01   2.149  0.03216 *
eFGPerDefTeamTwo   1.671e+02  4.195e+01   3.984 7.99e-05 ***
eFGPerOffTeamOneSQ 2.249e+03  9.827e+02   2.288  0.02260 *
FTRateOffTeamOneSQ -2.219e-01  1.062e-01  -2.089  0.03732 *
FTRateDefTeamOneSQ -1.864e-01  9.334e-02  -1.997  0.04649 *
eFGPerOffTeamTwoSQ 2.765e+03  1.052e+03   2.629  0.00887 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.705 on 424 degrees of freedom
Multiple R-squared:  0.2622,    Adjusted R-squared:  0.2448
F-statistic: 15.07 on 10 and 424 DF,  p-value: < 2.2e-16
```

Thomas Le

The goal of any competitive sport is to outscore your opponent. In the NBA, the range of scores two teams has vary through a number of variables. In our study, we will attempt to predict the total number of points in any given match. We have identified categories such as possessions, field goal attempts, rebounds, and turnovers. We will examine the impact each variable has on the total overall score.

NBA Championship caliber teams rebound successfully. Rebounds earn possessions and control pace. When you can limit your opponent to a single shot every time down the floor, you increase your chances of winning. Hard fought defense is squandered if the possession does not culminate with a defensive rebound. This study will look at the impact rebounds have on total points scored in a game.

Rebounding Model 1

	AvgTotalPoints
AvgTotalPoints	1.00000000
PossessionInMatchupperTeam	0.52068719
TotalPossessionsinamatchup	0.52068719
OeFG%TeamOne	0.37114884
AvailReboundTeamOne	-0.04457296
OORB%TeamOne	0.09032989
OffORBTeamOne	0.05408780
DefORBTeamOne	0.16729326
TeamOneRebTotal	0.15869279
DeFG%TeamTwo	0.16072372
Avail Rebound TeamTwo	0.27619566
OORB%TeamTwo	0.08359169
OffORBTeamTwo	0.19433295
DefORBTeamTwo	-0.09889869
TeamTwoRebTotal	0.01968006
TotalRebounds	0.13006833

Model 1 is a correlation matrix including all of the important variables regarding rebounding with respect to AvgTotalPoints. The data includes information on possessions per matchup, field goal percentages, offensive and defensive rebounds for both teams and total rebounds. Although TotalRebounds does not have a high correlation with AvgTotalPoints, we will need to maintain that variable as it is the basis of this study. We will begin to eliminate variables below .10 to trim the data set that includes:

1. AvailReboundTeamOne
2. OORB%TeamOne
3. OffORBTeamOne
4. OORB%TeamTwo
5. DefORBTeamTwo
6. TeamTwoRebTotal

Rebounding Model 2

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -1670.1640   778.3924  -2.146   0.0325 *
PossessionInMatchupperTeam   -12.6926     7.2816  -1.743   0.0820 .
`OeFG%TeamOne`               965.9873   745.8512   1.295   0.1960
DefORBTeamOne                -2.6241     0.6556  -4.003 7.39e-05 ***
TeamOneRebTotal               1.8138     0.4387   4.135 4.27e-05 ***
`DeFG%TeamTwo`             2299.1816  1337.1202   1.720   0.0862 .
`Avail Rebound TeamTwo`      15.0407    16.1138   0.933   0.3511
TotalRebounds                 7.7226     7.5378   1.025   0.3062
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.913 on 427 degrees of freedom
Multiple R-squared:  0.3733,    Adjusted R-squared:  0.363
F-statistic: 36.33 on 7 and 427 DF,  p-value: < 2.2e-16
```

Before finalizing Model 2, variables such as TotalPossessionsinamatchup and OffORBTeamTwo were not defined because of singularities, therefore removed. Based on the F test of p-value: < 2.2e-16 we can reject the null hypothesis accept the alternative that at least one beta is not zero. Adj R-squared shows that 36% of the variability in y is explained by the model. TotalRebounds, the measured variable, is high compared to the default alpha of .05. All of the T-tests are high

with the exception of DefORBTeamOne and TeamOneRebTotal. We will look to improve the model as Model 3 but removing extraneous variables.

Rebounding Model 3

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -314.9502    37.8418  -8.323 1.16e-15 ***
`OeFG%TeamOne`   388.1163    30.5803  12.692  < 2e-16 ***
DefORBTeamOne     -2.3569     0.6306  -3.737 0.000211 ***
TeamOneRebTotal    1.7794     0.4387   4.056 5.93e-05 ***
`DeFG%TeamTwo`   276.8351    44.5542   6.213 1.23e-09 ***
TotalRebounds      1.9060     0.2587   7.367 9.02e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.924 on 429 degrees of freedom
Multiple R-squared:  0.3688,    Adjusted R-squared:  0.3614
F-statistic: 50.13 on 5 and 429 DF,  p-value: < 2.2e-16
```

Although the Adjusted R-squared remained the same, we were able to reduce the number of variables. The mean square of the errors, displayed below, is 78.53965 since we are using rebounds to predict average total points.

Rebounding Model 4

```
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)               26.927     17.035    1.58     0.11
PossessionInMatchupperTeam  3.149      0.217   14.54  < 2e-16 ***
TotalRebounds              -1.318      0.197   -6.71  6.2e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.09 on 432 degrees of freedom
Multiple R-squared:  0.34,      Adjusted R-squared:  0.337
F-statistic:  111 on 2 and 432 DF,  p-value: <2e-16
```

Through some reevaluation, I had included unnecessary data while developing my model to use rebounds as a predictor of total points scored. I was trying to force the data to show that rebounds is a predictor of total points. However, in correlation with PossessionInMatchupperTeam, TotalRebounds demonstrated a fair relationship with total points scored. In the final model, the Adjusted R - squared is a correct reflection of the data at 33.7% of the variability in AvgTotalPoints is explained with this model. The Durbin-Watson test shows that the variables are independent.

The findings on rebounds thus far have proven that there is a modest correlation with total points scored. Rebounds, with their reliable uncertainty, are an important part of basketball. The randomness of rebounding has its own place in the game. To elaborate, rebounding is not often the key, but good teams will find ways to limit their opponent's amount of possessions. When comparing the statistics of rebounds and total points scored, the randomness is more prevalent compared to looking at the impact of rebounds in the scope of total possessions.

Christian Craig

In every NBA game in a season, one of the statistics that is influential and plays a big role in the outcome of total points scored is turnovers. Whether it's on the offensive or defensive end, the influence of this statistic can be captured in numbers and can determine the total points scored in a given game. As a group we have narrowed what results to look for in the following categories: possessions, field goal attempts, offensive/defensive rebounds, and turnovers. These categories all play a major role in what determines the total points scored in a game and will be accounted for.

Since turnovers have shown to have an impact on the number of points that can be scored. I will work with models in this category to show its impact on the total points scored in a given game. When working with all of the turnover variables at once, some will help determine the ability to predict total points more than others. It is my responsibility to identify, add, or remove turnover variables that influence (and don't influence) the ability to predict our dependent variable. Below is our model that contains all of the turnover data:

Turnovers Model 1

```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                -40.6152    58.7841  -0.691 0.489997
PerAfterTurnoverOffTeamOne  -2.6515     0.9454  -2.805 0.005273 **
SecondsAfterTurnoverOffTeamOne  1.0067     0.9925   1.014 0.311018
PointsAfterTurnoverOffTeamOne  93.5675    13.7559   6.802 3.57e-11 ***
PerAfterTurnoverDefTeamOne   1.7881     0.6686   2.674 0.007777 **
SecondsAfterTurnoverDefTeamOne -1.5116     1.8249  -0.828 0.407984
PointsAfterTurnoverDefTeamOne  38.6083    10.6909   3.611 0.000342 ***
TurnoverPerDefTeamOne        3.2252     0.7523   4.287 2.25e-05 ***
PerAfterTurnoverOffTeamTwo  -1.9996     1.1905  -1.680 0.093780 .
SecondsAfterTurnoverOffTeamTwo -3.2544     1.2486  -2.606 0.009478 **
PointsAfterTurnoverOffTeamTwo  20.2616    16.0001   1.266 0.206094
PerAfterTurnoverDefTeamTwo  -1.4183     1.0425  -1.360 0.174405
SecondsAfterTurnoverDefTeamTwo -1.2952     2.1434  -0.604 0.545995
PointsAfterTurnoverDefTeamTwo  46.6873    12.0217   3.884 0.000120 ***
TurnoverPerOffTeamTwo        1.3284     0.9108   1.458 0.145454
TurnoverPerDefTeamTwo        1.6668     0.9328   1.787 0.074672 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.19 on 419 degrees of freedom
Multiple R-squared:  0.3462,    Adjusted R-squared:  0.3228
F-statistic: 14.79 on 15 and 419 DF,  p-value: < 2.2e-16
```

As shown above, the f-test (p-value:<2.2e16) tells us that at least one of the betas is not 0. The adjusted R-Squared tells us that 32.28% of the variability in the total points scored is predicted by our model. With turnovers being one of the five model categories, this not too bad. Notice that in the individual t-tests SecondsAfterTurnoverOffTeamOne, SecondsAfterTurnoverDefTeamOne, PerAfterTurnoverOffTeamTwo, PointsAfterTurnoverOffTeamTwo, PerAfterTurnoverDefTeamTwo, SecondsAfterTurnoverDefTeamTwo, TurnoverPerOffTeamTwo, and TurnoverPerDefTeamTwo all have a p-value higher than our default alpha of .05. Below in model 2, you will see those taken out to show what the updated model looks like without the variables that didn't pass the T-test. We can notice now that with taking those variables out, that the impact was minimal to the adjusted R-squared. Now 31.79% of the variability in total average points scored is predicted by our model. Once we combine our models as a team, we can expect the adjusted R-squared to

increase because, with more categories, we will increase our ability to predict the variability of the total average points scored.

We then can run our model to check the f-test, p-test, and adjusted R-squared. In Model 2, shown below. We ran our model with the variables that past the correlation threshold.

Turnovers Model 2

```
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                   13.5711    24.9020   0.545  0.58605
PerAfterTurnoverOffTeamOne    -2.6288     0.9231  -2.848  0.00461 **
PointsAfterTurnoverOffTeamOne 87.7921    11.8964   7.380 8.35e-13 ***
PerAfterTurnoverDefTeamOne     1.7069     0.6172   2.765  0.00593 **
PointsAfterTurnoverDefTeamOne 42.8849     9.6331   4.452 1.09e-05 ***
TurnoverPerDefTeamOne          3.3205     0.7277   4.563 6.61e-06 ***
SecondsAfterTurnoverOffTeamTwo -5.4885    0.8932  -6.145 1.84e-09 ***
PointsAfterTurnoverDefTeamTwo  39.8759    8.9014   4.480 9.61e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.223 on 427 degrees of freedom
Multiple R-squared:  0.3289,    Adjusted R-squared:  0.3179
F-statistic:  29.9 on 7 and 427 DF,  p-value: < 2.2e-16
```

With our current selected variables, we want to check to see how close or far away the errors are from the regression line (the actual values). By using the formula “mean(modl2$residuals^2)” we are able to see that our mean square error is 8350368. This makes sense since we are only using one category to predict the average total points scored between two teams. When we combine our models, we can expect this value to go down.

Next, we will use backward selection to see if we get difference recommended variables to use in our model. For this, we will include the variables that were kept in after the cutoff of correlation. The results are below:

Turnovers Model 3

```
                               Df Sum of Sq   RSS    AIC
<none>                                      35541 1941.3
- PointsAfterTurnoverOffTeamTwo  1     172.3 35713 1941.5
- PerAfterTurnoverDefTeamTwo     1     227.6 35769 1942.1
- TurnoverPerOffTeamTwo          1     237.0 35778 1942.2
- TurnoverPerDefTeamTwo          1     298.0 35839 1943.0
- PerAfterTurnoverOffTeamTwo     1     368.6 35910 1943.8
- PerAfterTurnoverDefTeamOne     1     535.8 36077 1945.9
- SecondsAfterTurnoverOffTeamTwo 1     591.2 36132 1946.5
- PerAfterTurnoverOffTeamOne     1     778.7 36320 1948.8
- PointsAfterTurnoverDefTeamOne  1    1627.2 37168 1958.8
- TurnoverPerDefTeamOne          1    1835.5 37376 1961.2
- PointsAfterTurnoverDefTeamTwo  1    2192.1 37733 1965.4
- PointsAfterTurnoverOffTeamOne  1    4425.3 39966 1990.4
```

The final backward regression model shows us the lowest possible AIC given the variables put in. This means that the least amount of information is lost by having these independent variables in the model. The backward regression model in Model 3 includes five more variables than our

regression model in Model 2, but since we know those variables don't pass the t-test, we can make the decision to leave them out of the model.

To further test the significance of our independent variables (see what gives the highest possible adjusted R-squared), we will use forward stepwise selection. The independent variables included in this model will also be the variables that were kept after the correlation test. The results are below:

Turnovers Model 4

```
Start:  AIC=1945.48
AvgTotalPoints ~ PerAfterTurnoverOffTeamOne + SecondsAfterTurnoverOffTeamOne +
    PointsAfterTurnoverOffTeamOne + PerAfterTurnoverDefTeamOne +
    SecondsAfterTurnoverDefTeamOne + PointsAfterTurnoverDefTeamOne +
    TurnoverPerDefTeamOne + PerAfterTurnoverOffTeamTwo + SecondsAfterTurnoverOffTeam
Two +
    PointsAfterTurnoverOffTeamTwo + PerAfterTurnoverDefTeamTwo +
    SecondsAfterTurnoverDefTeamTwo + PointsAfterTurnoverDefTeamTwo +
    TurnoverPerOffTeamTwo + TurnoverPerDefTeamTwo
```

Model 4 shows the final forward stepwise model that was provided. This showed all of the original variables that were in our model. Both Model 3 and Model 4 had additional variables, because of the increased adjusted R-squared, below we will run model showing backward regression recommended variables since we already know what the model looks like with all the variables included:

Turnovers Model 5

```
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -58.5956    37.9548  -1.544  0.12338
PointsAfterTurnoverOffTeamTwo  22.4040    15.6646   1.430  0.15339
PerAfterTurnoverDefTeamTwo     -1.6220     0.9867  -1.644  0.10094
TurnoverPerOffTeamTwo           1.4418     0.8596   1.677  0.09421 .
TurnoverPerDefTeamTwo           1.7132     0.9108   1.881  0.06066 .
PerAfterTurnoverOffTeamTwo     -2.2711     1.0856  -2.092  0.03703 *
PerAfterTurnoverDefTeamOne      1.5863     0.6289   2.522  0.01203 *
SecondsAfterTurnoverOffTeamTwo -3.2388     1.2224  -2.650  0.00836 **
PerAfterTurnoverOffTeamOne     -2.8114     0.9246  -3.041  0.00251 **
PointsAfterTurnoverDefTeamOne  42.3435     9.6333   4.396 1.40e-05 ***
TurnoverPerDefTeamOne           3.3923     0.7267   4.668 4.08e-06 ***
PointsAfterTurnoverDefTeamTwo  50.1495     9.8299   5.102 5.10e-07 ***
PointsAfterTurnoverOffTeamOne  86.5586    11.9412   7.249 2.02e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.177 on 422 degrees of freedom
Multiple R-squared:  0.3434,    Adjusted R-squared:  0.3247
F-statistic: 18.39 on 12 and 422 DF,  p-value: < 2.2e-16
```

The adjusted R-squared in model 5 (including T2DefTO) is .3247, while the adjusted R-squared in model 2 is .3179. Since the impact of keeping the five additional variables is so minimal, and since those variables do not pass the t-test with their high p-values, we will continue to leave the recommended backward regression variables out.

To further look into the accuracy of our model, we will check for multicollinearity in our independent variables. This will ensure we do not have rounding errors, incorrect beta estimates,

wrong positive or negative values, or t-tests giving back incorrect information. For this, we will first run a test check-in for a variable inflation factor greater than 10, using the independent variables in Model 3. In model 6 (below) the VIF for the independent variables seems to be fairly low.

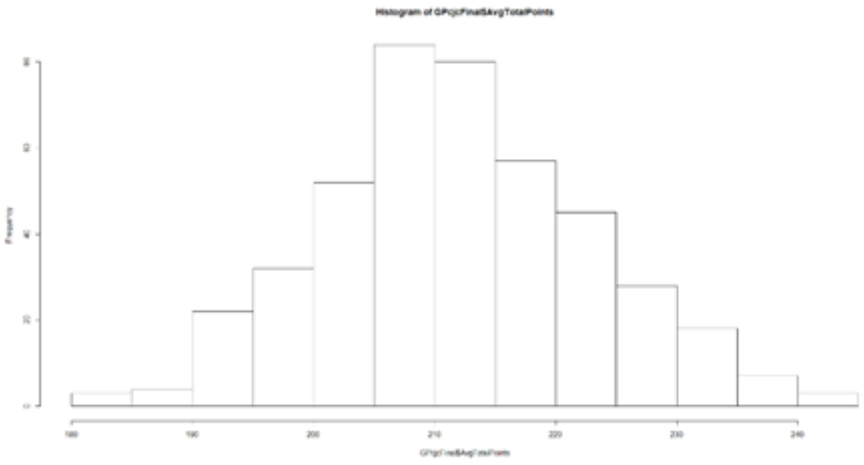
Turnovers Model 6

```
> vif(modl2)
    PerAfterTurnoverOffTeamOne  PointsAfterTurnoverOffTeamOne
                      2.488836                       1.456995
    PerAfterTurnoverDefTeamOne  PointsAfterTurnoverDefTeamOne
                      1.092079                       1.311608
         TurnoverPerDefTeamOne SecondsAfterTurnoverOffTeamTwo
                      2.244290                       1.085166
 PointsAfterTurnoverDefTeamTwo
                      1.019487
```

In summary we will use the variables that are provided in model 2. The variables associated in the model are appropriately correlated with the dependent variable, pass the t-test with given an alpha of .05, and give us a moderately high adjusted r-squared seeing as how this is just one category of variables in our data set. We took into account whether we should add or omit the independent variables using backward and forward stepwise regression. Also, we also checked for multicollinearity by checking the VIF. All of these steps were essential in choosing our final variables.

In our model, we will now check to see if there is a need to transform our dependent variable. This involves checking the distribution, to make sure it is not skewed.

Turnovers Model 7



The model above (model 7) shows that the distribution is normal. If the distribution was not normal, we would have to take the log of the dependent variable. All of the independent variables didn't have any heavily skewed distribution. As a result, we do not have to transform or log any of the variables for turnovers.

The below Model shows the addition of second-order terms:

Turnovers Model 8

```
Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                             -497.83958  291.25336  -1.709  0.08813 .
PerAfterTurnoverOffTeamOne               -20.52826   13.89818  -1.477  0.14041
PointsAfterTurnoverOffTeamOne            869.84720  421.65239   2.063  0.03973 *
PerAfterTurnoverDefTeamOne                 0.51229    0.77254   0.663  0.50762
PointsAfterTurnoverDefTeamOne             50.30326   10.83080   4.644 4.56e-06 ***
TurnoverPerDefTeamOne                    -26.49197   13.98147  -1.895  0.05880 .
SecondsAfterTurnoverOffTeamTwo            59.10373   24.31070   2.431  0.01547 *
PointsAfterTurnoverDefTeamTwo             40.17737    9.33911   4.302 2.10e-05 ***
GPcjcFinal$PerAfterTurnoverOffTeamOneSQ    1.08705    0.87903   1.237  0.21690
GPcjcFinal$PointsAfterTurnoverOffTeamOneSQ -324.76062 174.21283 -1.864  0.06299 .
GPcjcFinal$PerAfterTurnoverDefTeamOneSQ     1.11693    0.51093   2.186  0.02936 *
GPcjcFinal$SecondsAfterTurnoverOffTeamTwoSQ -3.39499   1.28276  -2.647  0.00843 **
GPcjcFinal$PointsAfterTurnoverDefTeamTwoSQ   0.02622    0.02332   1.124  0.26162
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.09 on 422 degrees of freedom
Multiple R-squared:  0.3557,    Adjusted R-squared:  0.3374
F-statistic: 19.42 on 12 and 422 DF,  p-value: < 2.2e-16
```

After analyzing Model 8 we can see that we will not add second-order terms to the model. This is because the addition of them has not only muddied our t-test but also has had little impact on the adjusted R-squared. There was also the attempt to add in the second-order term one by one. All of them had minimal impact in regards to the adjusted R-squared as well as negatively impacting the t-tests.

Within our model we checked for interaction terms, and the most significant increase was with the interaction was Team 1 points per possession after turnover and Team 1 points allowed per possession after turnover. This is shown in the model below:

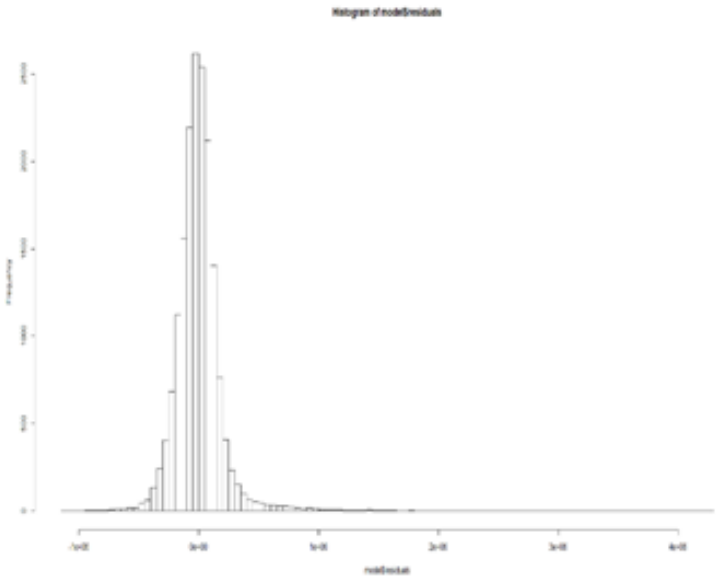
Turnovers Model 9

```
Coefficients:
                                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                                           1100.8555   503.4163   2.187  0.02930 *
PerAfterTurnoverOffTeamOne                              -1.6552     1.0235  -1.617  0.10657
PointsAfterTurnoverOffTeamOne                         -791.5724   406.8271  -1.946  0.05235 .
PerAfterTurnoverDefTeamOne                               1.7263     0.6147   2.809  0.00521 **
PointsAfterTurnoverDefTeamOne                         -824.9465   401.4358  -2.055  0.04049 *
TurnoverPerDefTeamOne                                    2.9919     0.7404   4.041 6.32e-05 ***
SecondsAfterTurnoverOffTeamTwo                          -5.3915     0.8905  -6.054 3.10e-09 ***
PointsAfterTurnoverDefTeamTwo                           39.6788     8.8638   4.476 9.75e-06 ***
PointsAfterTurnoverOffTeamOne:PointsAfterTurnoverDefTeamOne  699.1150   323.2997   2.162  0.03114 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.184 on 426 degrees of freedom
Multiple R-squared:  0.3362,    Adjusted R-squared:  0.3237
F-statistic: 26.97 on 8 and 426 DF,  p-value: < 2.2e-16
```

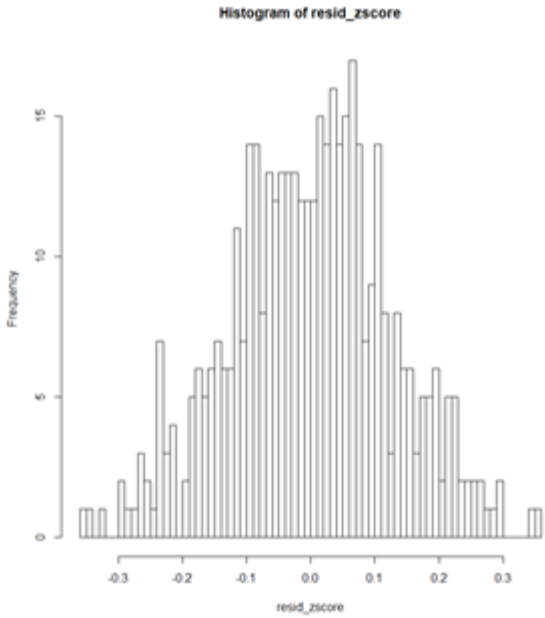
The interaction term has increased the adjusted R to .2923, but has also muddied the t-test of Team 1 points after turnover. Model 9 will be the new model moving forward, but could be subject to change after discussion with the group due to the increased t-test.

Next, we will perform a residual analysis of our variables. Below is a histogram that will check the distribution of the residuals, checking for normality:

Turnovers Model 10



Turnovers Model 11



As shown above, there appears to be a normal distribution, but there are some outliers. In Model 11 we can see that approximately 95% of our residuals fall within 2 standard deviations. Before this, we also calculated that our sum was equal to 0:

Turnovers Model 12

```
[1] -1.550843e-13
```

After checking for normality, we can check for independence using the Durbin Watson test. This is shown in the model below:

Turnovers Model 13

```
 lag Autocorrelation D-W Statistic p-value
   1      0.05873022       1.871206   0.098
 Alternative hypothesis: rho != 0
```

As shown above, the data for turnovers is above .05, we cannot reject the null hypothesis and need to look further into the data. In other words, this means that our data is dependent on each other. After taking out our interaction variable, we were able to lower our Durbin Watson Test, but we still cannot accept the alternative hypothesis. Reference the adjusted score below:

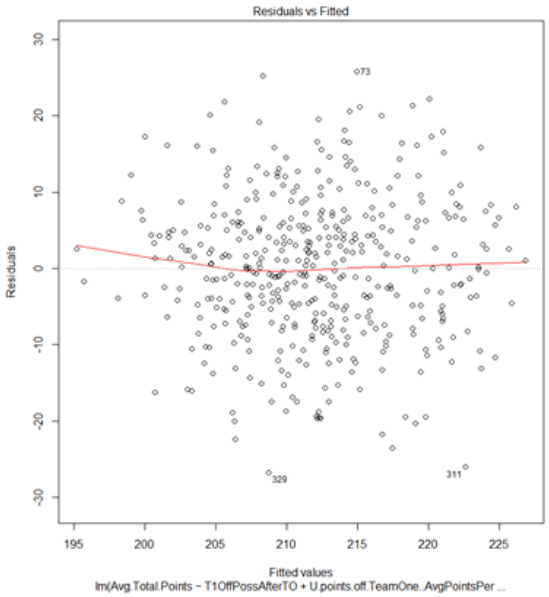
Turnovers Model 14

```
lag Autocorrelation D-W Statistic p-value
  1      0.06893457       1.850052   0.084
Alternative hypothesis: rho != 0
```

There is still a little dependence in our data set. At this time, we cannot take out any further variables because it will decrease our adjusted R-squared, so we will keep the interaction term in our model

We are now checking for constant variance in our model. In order to do this, we will check a residual vs. fitted plot. This can be seen in the model below:

Turnovers Model 15



Per Model 14, it can be interpreted as homoscedastic since the variances on the left appear to match the variances on the right. Despite there being some outliers, we can determine that there is some consistency amongst the residuals, as it appears 95% are within two standard deviations.

Since our interaction variable significantly increased our variable dependence, we are back to using model 3 as our final model. The model is also shown below:

Turnovers Model 16

```
Coefficients:
                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                     1100.8555   503.4163   2.187  0.02930 *
PerAfterTurnoverOffTeamOne                        -1.6552     1.0235  -1.617  0.10657
PointsAfterTurnoverOffTeamOne                   -791.5724   406.8271  -1.946  0.05235 .
PerAfterTurnoverDefTeamOne                        1.7263     0.6147   2.809  0.00521 **
PointsAfterTurnoverDefTeamOne                 -824.9465   401.4358  -2.055  0.04049 *
TurnoverPerDefTeamOne                            2.9919     0.7404   4.041 6.32e-05 ***
SecondsAfterTurnoverOffTeamTwo                  -5.3915     0.8905  -6.054 3.10e-09 ***
PointsAfterTurnoverDefTeamTwo                   39.6788     8.8638   4.476 9.75e-06 ***
PointsAfterTurnoverOffTeamOne:PointsAfterTurnoverDefTeamOne 699.1150   323.2997   2.162  0.03114 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.184 on 426 degrees of freedom
Multiple R-squared:  0.3362,    Adjusted R-squared:  0.3237
F-statistic: 26.97 on 8 and 426 DF,  p-value: < 2.2e-16
```

For our final model for turnovers, we can look at the f-test that states the p-value is <2.2e-16. This tells us that we can reject the null hypothesis, that at least one of the betas is not equal to zero. If we then look at each of the individual t-tests we can see that all of the variables pass (less than .05) except for "PerAfterTurnoverOffTeamOne". This is because the interaction term, that includes this variable, passes the t-test. With our p-values passing the t-tests we can then assume we can use our estimates that are showing in the model. The adjusted R-squared is .3237. This means that 32.37% of the variability in our dependent variable is explained by our model. For being a subsection of our data, this can be interpreted that turnovers can predict some part of the total points scored in a given game. We also checked the residuals in our analysis which included, adding the residuals to 0, showing distribution plots to display normality, performing the Durbin Watson Test to show the independence amongst variables, and showing homoscedastic variance by displaying the Residual vs Fitted model.

Josh Hall

In the model-building stage, I was tasked with the entire dataset. My goal was to build the best model I could within reason of having over 60 dependent variables. The goal of the project is to predict the total score of a game given the two teams in the matchup for each of the possible matchups between the 30 NBA teams.

We started with a huge dataset, 89 independent variables, and one of my main goals in this section was to attempt to limit it as much as possible without losing any of our predictive power. The first thing I did was to remove some variables from the dataset that were giving us the same information. In this dataset that was Wins, Losses, Exp Wins and Win Diff. These binary style stats do not actually impact the score of a game just tell you who won it. For the purposes of this study, we are not interested in who wins games or what their records are just how their pace of play and efficiency on offense and defense affect the total points score in a matchup.

The next thing that I decided to do was to look at the plots of our dependent variable versus all of our independent variables. Looking to find any obvious patterns that might lead you to believe you need a second-order model with any of the variables. There was nothing that jumped out at me but that has been the case in the past and second-order variables have been very helpful so I went ahead and tried a few different combinations and nothing improved the model.

At this point, I ran a model on the whole dataset with every first-order variable included. Since a few of the variables came up as NA there is a good possibility of multicollinearity across a few of the variables. Therefore I decided to check out the correlation matrix for the entire dataset. It was at this point that something became very obvious, we have the total points and seconds per possession as well as all three types that make up those numbers. They obviously had a lot of correlation and since all of the data in the total columns are shown in the three categories that make it up we should be able to remove that without any loss of information. This will shrink the data by 8 columns because we can remove the total points and seconds for TeamOne offense and defense as well as TeamTwo's totals for offense and defense.

I then ran the model again with those variables removed. This eliminated all but 2 NAs in the model estimate and it has the exact same Adjusted R-squared as the previous model showing that they were just excess variables and no information was lost. I then looked at the correlation matrix again to see if the last few NAs had a high correlation with anything. It does not look like

they do, therefore I did not want to remove those variables yet because I think they are going to be valuable variables in the future when we start to build some interaction terms. There will be future cleaning of the data that hopefully resolves this issue. These are the results of that model.

Entire Data Set Model 1

```
Call:
lm(formula = AvgTotalPoints ~ . - TeamOne - TeamTwo, data = NBA)

Residuals:
     Min       1Q   Median       3Q      Max
-20.0454  -5.0745  -0.0642   4.4702  20.1565

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                    82.71036 1470.11747   0.056    0.955
TotalPossOffTeamOne             0.09652    0.10195   0.947    0.344
PerAfterMadeShotOffTeamOne     59.59089  401.42946   0.148    0.882
SecondsAfterMadeShotOffTeamOne  0.31270   17.89413   0.017    0.986
PointsAfterMadeShotOffTeamOne 431.07976  303.44554   1.421    0.156
PerAfterReboundOffTeamOne     255.86379  727.35545   0.352    0.725
SecondsAfterReboundOffTeamOne  -7.12657   13.66961  -0.521    0.602
PointsAfterReboundOffTeamOne   -1.63166  234.06447  -0.007    0.994
PerAfterTurnoverOffTeamOne    599.25167  670.71693   0.893    0.372
SecondsAfterTurnoverOffTeamOne -3.39950    8.49185  -0.400    0.689
PointsAfterTurnoverOffTeamOne -85.74358   80.85851  -1.060    0.290
TotalPossDefTeamOne            -0.10387    0.06512  -1.595    0.112
PerAfterMadeShotDefTeamOne   -269.54596  735.59319  -0.366    0.714
SecondsAfterMadeShotDefTeamOne -13.61940   15.84991  -0.859    0.391
PointsAfterMadeShotDefTeamOne  72.06703  422.57360   0.171    0.865
PerAfterReboundDefTeamOne    -221.86210  988.46599  -0.224    0.823
SecondsAfterReboundDefTeamOne  -2.81302   19.33081  -0.146    0.884
PointsAfterReboundDefTeamOne   -0.88483  319.72989  -0.003    0.998
PerAfterTurnoverDefTeamOne   -161.53301  602.88016  -0.268    0.789
SecondsAfterTurnoverDefTeamOne -6.37478    9.14102  -0.697    0.486
PointsAfterTurnoverDefTeamOne   6.01588   45.91174   0.131    0.896
PointsPossOffTeamOne           -1.84697    3.56671  -0.518    0.605
eFGPerOffTeamOne              -13.23520  140.96409  -0.094    0.925
FTRateOffTeamOne               -2.28556    2.44807  -0.934    0.351
PointsPossDefTeamOne           -1.01379   13.77526  -0.074    0.941
eFGPerDefTeamOne              403.25196  574.67072   0.702    0.483
FTRateDefTeamOne                1.95211    1.95256   1.000    0.318
TotalPossOffTeamTwo             0.03867    0.05875   0.658    0.511
PerAfterMadeShotOffTeamTwo    -29.98966  251.68293  -0.119    0.905
SecondsAfterMadeShotOffTeamTwo  7.42134   14.79494   0.502    0.616
PointsAfterMadeShotOffTeamTwo 337.37926  309.41478   1.090    0.276
PerAfterReboundOffTeamTwo    -318.78346  435.19318  -0.733    0.464
SecondsAfterReboundOffTeamTwo   9.76871   10.68139   0.915    0.361
PointsAfterReboundOffTeamTwo   89.82358  123.27826   0.729    0.467
PerAfterTurnoverOffTeamTwo   -251.05190  697.92667  -0.360    0.719
SecondsAfterTurnoverOffTeamTwo -7.93921    5.49489  -1.445    0.149
PointsAfterTurnoverOffTeamTwo  51.51790   89.71686   0.574    0.566
TotalPossDefTeamTwo            -0.01343    0.06332  -0.212    0.832
PerAfterMadeShotDefTeamTwo   -559.04710  416.41064  -1.343    0.180
SecondsAfterMadeShotDefTeamTwo -11.04228   14.65444  -0.754    0.452
PointsAfterMadeShotDefTeamTwo  38.79867  481.04716   0.081    0.936
PerAfterReboundDefTeamTwo    -366.22106  533.29012  -0.687    0.493
SecondsAfterReboundDefTeamTwo  15.08207   11.03026   1.367    0.172
PointsAfterReboundDefTeamTwo  162.77259  284.53847   0.572    0.568
PerAfterTurnoverDefTeamTwo   -382.80869  454.26412  -0.843    0.400
SecondsAfterTurnoverDefTeamTwo -1.21745    6.56285  -0.186    0.853
PointsAfterTurnoverDefTeamTwo  52.46854   31.89147   1.645    0.101
PointsPossOffTeamTwo           -2.44880    2.86480  -0.855    0.393
eFGPerOffTeamTwo                9.62390  124.98273   0.077    0.939
FTRateOffTeamTwo               -1.67516    1.35651  -1.235    0.218
PointsPossDefTeamTwo           -3.83163    8.46157  -0.453    0.651
eFGPerDefTeamTwo              190.82470  313.38340   0.609    0.543
FTRateDefTeamTwo                0.93412    1.83127   0.510    0.610
PossPerTeamPerGame              3.09417    5.29191   0.585    0.559

Residual standard error: 7.885 on 381 degrees of freedom
Multiple R-squared:  0.5623,    Adjusted R-squared:  0.5014
F-statistic: 9.236 on 53 and 381 DF,  p-value: < 2.2e-16
```

Here we can see that our adjusted R-squared is over 50% but none of our T-tests are significant. This is a telltale sign that there is a fair amount of overlap in our independent variables. This means that there is some more weeding out of variables to be done. To be sure, I checked the VIF of the model and there were a lot of variables over the 10 threshold, meaning we have a lot of multicollinearity in our model.

The next thing I did was to add interaction variables that I thought were going to be important for improving the model beyond the first-order model. Possessions being the main focus of our research led me to believe that using that as an interaction variable with the Percent of time certain possession occur in a matchup that was an obvious place to start. Another thing I noticed when I was diving a little deeper into the individual variables is that PointsAfterMadeShotOff and SecondsAfterMadeShotOff looked close to log curves. Therefore I transformed those variables and figured it was time to try and limit the model so I ran a backward elimination to see what a potential final model would look like at this point.

Entire Data Set Model 2

```
Call:
lm(formula = AvgTotalPoints ~ TotalPossOffTeamOne + SecondsAfterMadeShotOffTeamOne +
    PointsAfterMadeShotOffTeamOne + PerAfterReboundOffTeamOne +
    SecondsAfterReboundOffTeamOne + PointsAfterReboundOffTeamOne +
    PerAfterTurnoverOffTeamOne + SecondsAfterTurnoverOffTeamOne +
    PointsAfterTurnoverOffTeamOne + TotalPossDefTeamOne + PerAfterMadeShotDefTeamOne +
    SecondsAfterMadeShotDefTeamOne + PerAfterReboundDefTeamOne +
    SecondsAfterReboundDefTeamOne + PointsAfterReboundDefTeamOne +
    PerAfterTurnoverDefTeamOne + SecondsAfterTurnoverDefTeamOne +
    PointsAfterTurnoverDefTeamOne + PointsPossOffTeamOne + FTRateOffTeamOne +
    PointsPossDefTeamOne + eFGPerDefTeamOne + FTRateDefTeamOne +
    TotalPossOffTeamTwo + PointsAfterMadeShotOffTeamTwo + PerAfterReboundOffTeamTwo +
    PerAfterTurnoverOffTeamTwo + SecondsAfterTurnoverOffTeamTwo +
    PerAfterMadeShotDefTeamTwo + PointsAfterMadeShotDefTeamTwo +
    PerAfterReboundDefTeamTwo + SecondsAfterReboundDefTeamTwo +
    PerAfterTurnoverDefTeamTwo + PointsAfterTurnoverDefTeamTwo +
    FTRateOffTeamTwo + PossPerTeamPerGame + log(SecondsAfterMadeShotOffTeamOne) +
    log(PointsAfterMadeShotOffTeamOne) + PerAfterReboundOffTeamOne:PossPerTeamPerGame +
    PerAfterMadeShotDefTeamOne:PossPerTeamPerGame + PerAfterTurnoverDefTeamOne:PossPerTeamPerGame +
    PerAfterTurnoverOffTeamTwo:PossPerTeamPerGame + PerAfterMadeShotDefTeamTwo:PossPerTeamPerGame +
    PerAfterReboundDefTeamTwo:PossPerTeamPerGame + PerAfterTurnoverDefTeamTwo:PossPerTeamPerGame +
    FTRateDefTeamOne:PossPerTeamPerGame + FTRateOffTeamTwo:PossPerTeamPerGame +
    eFGPerDefTeamOne:PossPerTeamPerGame, data = NBA)

Residuals:
     Min       1Q   Median       3Q      Max
-17.6745  -4.5858  -0.4805   4.3200  18.9301

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                               3.817e+04  1.501e+04   2.542 0.011398 *
TotalPossOffTeamOne                       5.991e-01  1.827e-01   3.279 0.001135 **
SecondsAfterMadeShotOffTeamOne            1.033e+03  3.593e+02   2.875 0.004261 **
PointsAfterMadeShotOffTeamOne             1.148e+04  4.183e+03   2.745 0.006328 **
PerAfterReboundOffTeamOne                -6.977e+03  3.429e+03  -2.035 0.042564 *
SecondsAfterReboundOffTeamOne            -3.119e+01  9.052e+00  -3.445 0.000633 ***
PointsAfterReboundOffTeamOne              1.483e+03  5.133e+02   2.889 0.004085 **
PerAfterTurnoverOffTeamOne                2.231e+03  5.810e+02   3.840 0.000144 ***
SecondsAfterTurnoverOffTeamOne           -8.069e+01  2.653e+01  -3.042 0.002513 **
PointsAfterTurnoverOffTeamOne            -3.016e+02  9.567e+01  -3.152 0.001749 **
TotalPossDefTeamOne                      -7.068e-01  2.134e-01  -3.313 0.001011 **
PerAfterMadeShotDefTeamOne               -6.733e+03  1.770e+03  -3.805 0.000165 ***
SecondsAfterMadeShotDefTeamOne           -1.750e+02  5.875e+01  -2.979 0.003073 **
PerAfterReboundDefTeamOne                -3.452e+03  1.265e+03  -2.729 0.006642 **
SecondsAfterReboundDefTeamOne             6.355e+01  2.467e+01   2.576 0.010373 *
PointsAfterReboundDefTeamOne              7.820e+02  2.847e+02   2.747 0.006303 **
PerAfterTurnoverDefTeamOne               -1.078e+04  3.236e+03  -3.332 0.000947 ***
SecondsAfterTurnoverDefTeamOne           -5.536e+01  1.869e+01  -2.961 0.003251 **
PointsAfterTurnoverDefTeamOne             1.327e+02  5.493e+01   2.416 0.016168 *
PointsPossOffTeamOne                     -3.984e+01  1.277e+01  -3.119 0.001951 **
FTRateOffTeamOne                         -1.704e+01  5.490e+00  -3.104 0.002048 **
PointsPossDefTeamOne                     -3.018e+01  1.135e+01  -2.659 0.008165 **
eFGPerDefTeamOne                         -4.230e+03  2.531e+03  -1.671 0.095439 .
FTRateDefTeamOne                         -1.829e+01  1.358e+01  -1.347 0.178734
TotalPossOffTeamTwo                       4.017e-02  1.248e-02   3.219 0.001397 **
PointsAfterMadeShotOffTeamTwo             1.290e+02  3.573e+01   3.611 0.000345 ***
PerAfterReboundOffTeamTwo                -3.120e+02  1.125e+02  -2.773 0.005827 **
PerAfterTurnoverOffTeamTwo                7.117e+03  2.945e+03   2.417 0.016116 *
SecondsAfterTurnoverOffTeamTwo           -4.842e+00  2.363e+00  -2.049 0.041105 *
PerAfterMadeShotDefTeamTwo                6.423e+03  3.866e+03   1.661 0.097441 .
PointsAfterMadeShotDefTeamTwo            -1.093e+02  6.675e+01  -1.638 0.102280
PerAfterReboundDefTeamTwo                 7.827e+03  4.349e+03   1.800 0.072719 .
SecondsAfterReboundDefTeamTwo             8.592e+00  4.989e+00   1.722 0.085823 .
PerAfterTurnoverDefTeamTwo                1.206e+04  4.398e+03   2.742 0.006393 **
PointsAfterTurnoverDefTeamTwo             2.417e+01  1.185e+01   2.040 0.042023 *
FTRateOffTeamTwo                         -2.243e+01  1.448e+01  -1.550 0.122032
PossPerTeamPerGame                       -2.454e+01  5.302e+01  -0.463 0.643680
log(SecondsAfterMadeShotOffTeamOne)      -1.859e+04  6.438e+03  -2.888 0.004093 **
log(PointsAfterMadeShotOffTeamOne)       -8.363e+03  3.617e+03  -2.312 0.021298 *
PerAfterReboundOffTeamOne:PossPerTeamPerGame   7.069e+01  3.492e+01   2.024 0.043633 *
PerAfterMadeShotDefTeamOne:PossPerTeamPerGame  3.370e+01  1.384e+01   2.435 0.015364 *
PerAfterTurnoverDefTeamOne:PossPerTeamPerGame  9.039e+01  3.218e+01   2.809 0.005224 **
PerAfterTurnoverOffTeamTwo:PossPerTeamPerGame -7.458e+01  3.031e+01  -2.461 0.014292 *
PerAfterMadeShotDefTeamTwo:PossPerTeamPerGame -6.644e+01  3.987e+01  -1.666 0.096443 .
PerAfterReboundDefTeamTwo:PossPerTeamPerGame  -8.076e+01  4.485e+01  -1.801 0.072529 .
PerAfterTurnoverDefTeamTwo:PossPerTeamPerGame -1.242e+02  4.494e+01  -2.763 0.006005 **
FTRateDefTeamOne:PossPerTeamPerGame           2.484e-01  1.393e-01   1.784 0.075216 .
FTRateOffTeamTwo:PossPerTeamPerGame           2.237e-01  1.483e-01   1.508 0.132289
eFGPerDefTeamOne:PossPerTeamPerGame           7.714e+01  2.622e+01   2.942 0.003460 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.5 on 386 degrees of freedom
Multiple R-squared:  0.5988,    Adjusted R-squared:  0.549
F-statistic:    12 on 48 and 386 DF,  p-value: < 2.2e-16
```

Obviously, this was a big improvement over the previous model and we were able to remove a lot of the unnecessary variables. I tried for a long amount of time to beat this model and it never happened. I am not saying it is the best model but it is by far the best one that I found.

Section 5 - Discussion

For our tests, we made the assumption that teams performed consistently not considering circumstances such as injuries, trades, or opponent schedule difficulty. It is possible that some combination of other categorical predictors would yield better results, but we will reserve that for future consideration. We explored areas to enhance our model with interaction terms, second-order terms, and transformations, but ultimately concluded the given first-order terms were sufficient.

Our results suggest several key findings in that no interaction terms, second-order terms or transformations worked on our individual models using the parallel but separate variables of our data. One of our group members had access to the whole data to make a model and he was able to find interaction terms and second-order terms to help give a better adjusted R- squared. Each of our own models got an adjusted R-squared around 0.3 each, but the one with the whole data was able to come up with an adjusted R-squared of 0.549. Currently, our 0.549 adjusted R-squared means that 54.9% variability in total points scored is explained by the independent variables that we have chosen for our model. We will work on making sure all parts of our individual models are added to the final model that includes all of the data to come up with the best adjusted R-squared for the final report.

When looking at just the effective field goal (eFG) rates and free throw (FT) rates in an NBA game, we noticed that the offensive eFG for team one, offensive eFG for team two and defensive eFG for team two are significant in our model. We thought there might be a potential issue of multicollinearity between the offensive eFG for team one and defensive eFG for team two, but that was not the case when using the variance inflation factor (VIR) to check. This could be due to the team two defensive eFG stat accounting for the average of all 29 teams in the NBA and not just the opponent (team one) in the matchup. What is also surprising is that all FT T-test values were significant except for the defensive FT variable of team one. We would think that any variable that is significant in team one should also be significant in team two and vice versa. This warrants further investigation and was also seen with another group member working with the whole dataset.

Having the whole dataset to work with allowed me to have a couple of major findings while attempting to create the best model possible. The first thing that I did to improve past the first-order model was to create interaction terms. The one that stood out the most in my research was multiplying the percent of possessions that end in a given manner. Either a made shot, rebound or turnover, by the average number of total possessions we predict for that matchup. In the best model that I found we kept seven of the twelve interaction terms created in this way. All of those had p-values lower than 0.05. I still think that there are potential improvements to the model in this way and I will explore the ones presented by my groupmates further.

After adding those interaction terms, I attempted to add some second-order terms and transform some of the variables. This had some mixed results that I think might be due to the high multicollinearity of our dataset. Before we settle on the final model, I am going to run a ridge regression on the data to try and mitigate that affect a bit. The issues here were that data from TeamOne and TeamTwo should, for the most part, be equally relevant. It does not seem like it is right now. I think something we could do is randomize which team in the matchup is TeamOne and TeamTwo since it is alphabetical right now. The major finding is that taking the Log value of points after a made shot for TeamOne had a good effect on the model but if I did it for both TeamOne and TeamTwo the model was not quite as good. Only about a 1% difference in Adjusted R-squared so it is up to the eye of the beholder on how important that is but, in this situation, I believe that to be a significant amount for what we are trying to predict. Therefore, the current best model only transforms TeamOne's points after making a shot. A second-order term had a similar issue with TeamOne working and TeamTwo making that model a little worse. This term was average seconds of possession after a made shot. The negative effects of the TeamTwo portion were about 0.5% in this instance which still seems significant to me but in the end, I think this variable will end up in the final model.

In regards to turnovers, one of the major findings was that implementing interaction terms minimally improved the adjusted R-squared, but also came at the cost of decreasing our independence of errors amongst variables. As a group, it is important to know that there will be a tradeoff between choosing an increase of adjusted R-squared and a lack of independence amongst the errors of the independent variables (increased correlation of errors). With an increased Durbin Watson Test, we can be left with a concern that the errors are dependent on one another. When errors are dependent on one another, we then lose the ability to explain why the errors have occurred. In this particular case, since the removal of our interaction term only brought the Durbin Watson test down by .01, it seemed worth the increased adjusted R-squared to leave it in.

Good teams rebound successfully. Rebounds earn possessions and control the pace. When you can limit your opponent to a single shot every time down the floor, you increase your chances of winning. Hard fought defense is squandered if the possession does not culminate with a defensive rebound. I will look at the impact rebounds have on total points scored in a game. My model began by including all of the pertinent variables regarding rebounding from the data. I trimmed the model down to exclude columns such as offensive rebounding and defensive rebounding percentages because it was hard to draw a conclusion to predict points while including percentages. Based on my final model, I was able to infer rebounds have a modest impact predicting total points when coupled with possessions. Rebounds do not have a direct relation to teams scoring points, but this attribute provides extra opportunities to do so much like the other facets we are examining within this study.

When using the game data such as points scored off of turnovers, rebounds, and the average points per offensive/defensive possession, I found that secondary and interaction terms did not help with my model. I tried creating second-order and interaction terms to help make an improvement on adjusted R-Squared, by making various combinations of my variables, but I found out that the more second-order and interaction terms that I had led to a lower Adjusted R-

Squared. I ended up making a model off of only first-order terms and this led to the highest adjusted R-Squared using the variables that I had at my disposal.

In regards to the findings in each category, even when using all of our variables to predict the total number of points scored between two teams, we had difficulty achieving this goal with high precision. Our effective field goals for team one and team two were significant in our model while showing no signs of multicollinearity. In our dataset as a whole, we were able to add interaction terms, but due to high multicollinearity present, we hope to mitigate this by running ridge regression in some of our models. The turnover model was able to show a minimal impact adding an interaction term. When trying to add interaction or second-order terms to turnovers, the independence amongst variables decreased significantly. In addition, our points variables were unable to show any benefits in adjusted R-squared from the creation of interaction or second-order terms. When trying to add second-order terms to predict the total number of points, in some cases, the adjusted R-squared lowered. Rebounds were looked into and showed that they had a minimal impact on the total points scored between two teams.

Section 6 - Conclusion

Multiple regression models were built by individual members of our group to determine the total number of points scored by any two teams when they meet for an NBA game using the data from the 2016-2017 and 2017-2018 NBA seasons. We focused primarily on looking at independent variables of each NBA team's two seasonal averages in the following categories: possessions, field goal attempts, free throw rates, offensive/defensive rebounds, and turnovers. We have divided the project into parallel but distinct lines of work by most group members taking some independent variables to determine the total points that will be scored in an NBA game and with one group member working with the whole data set to make an all-encompassing model. We saw first hand how our individual models, that worked with a subset of the data, were only able to come up with a partial amount of the adjusted R-squared that the person working with the whole data set was able to come up with. We are happy to report that even though we ended up with a final model adjusted R-squared of 0.549, we feel that we have succeeded in using this project as an opportunity for us to demonstrate the skills we have learned in the class and we believe that we have all reached our individual and team milestones for the project.

Section 7 - Reference

1. Blitz, Matt. “The Lowest Scoring Game in NBA History and the Fix That Saved Professional Basketball.” Today I Found Out, 23 Dec. 2013, http://www.todayifoundout.com/index.php/2013/12/lowest-scoring-game-nba-historyinvention-saved-professional-basketball/.
2. Dubin, Jared. “The NBA's Other Offensive Revolution: Never Turning The Ball Over.” FiveThirtyEight, FiveThirtyEight, 14 Mar. 2019, https://fivethirtyeight.com/features/the-nbas-other-offensiverevolution-never-turning-the-ball-over/.
3. “Team Efficiency and Four Factors.” Cleaning the Glass, https://cleaningtheglass.com/stats/team/1/team#tab-four_factors.
4. “NBA Schedule.” *Odds Shark*, https://www.oddsshark.com/stats/gamelog/basketball/nba/20722.
5. “NBA Win Probability Game Box Scores.” *Inpredictable*, http://stats.inpredictable.com/nba/ssnTeamPoss.php.