Jack Boydell

Professor Smalley

MATH 239

Final Report: Statistical Modeling and Analysis Results of College/International Statistics in Addition to NBA Statistics and Salaries

Table of Contents

1.0 Introduction

This report uses statistical methods and data analysis to investigate the relationship between National Basketball Association (NBA) player performance statistics and their salaries for the 2017-2018 season. Judged and paid according to their play on the court, general managers across the NBA consider these same metrics when signing and trading for certain players. Looking at the data as a whole, each player makes up a row with his corresponding 2017-2018 salary and 30+ statistics that measure certain performance aspects on the court. A separate self-created data frame holds 40 randomly stratified sampled players' college or international statistics, opening up a second path of analysis into the relationship between performance in the NBA and prior to entering the league.

The main question of whether certain basketball performance statistics show regression correlation to individual player salaries drives the overall investigation towards the ultimate goal of prediction. These beginning relationships form the basis for more complex methods of tree plots and multiple linear regression: Are certain statistics helpful in explaining and predicting a player's salary in a given season? Does a player's statistics in college or overseas prior to entering the NBA help explain and predict their minutes per game (MPG) once they are in the league?

This analysis will progress from applying simple linear regression and looking for any correlation between salary and certain performance based player statistics to investigating the predicting power of tree diagrams to eventual final methods of multiple linear regression. Through these statistical methods of looking at data, it is possible to analyze certain decisions made by general managers across the NBA to pay certain players a specific salary and discuss if some players should demand higher or lower pay based on their performance in the 2017-2018 regular season. The application of college or international metric data opens up a new course of investigation into whether these stats prior to entrance in the NBA can help predict performance and statistics once a player is in the league.

2.0 Data Wrangling/Application

In order to begin the process of analyzing the data on NBA player salaries for the 2017-2018 and various individual player statistics for the same year, it was first necessary to combine these two separate datasets into one cohesive data frame. Due to certain formatting differences in the two columns titled "Player" (one in each dataset), Excel was used as an outside source to make a slight change in the construction of player names in order to then use RStudio to create the similarities that allowed for the eventual joining of the two sets of data. While the dataset with NBA player stats provides great metrics for totals for the regular season, it was important to manipulate the data to create variables of certain metrics on a per game basis. This manipulation was achieved through simple division and the "mutate" function in R as a column for points per game (PPG) was created by dividing total points for the season (PTS) by the number of games played (G). Similar per game metrics were created for minutes, assists, rebounds, and turnovers on a per game basis. Table 2.1 shows an example entry (Alex Abrines) of the final data frame that will be used for analysis after joining and mutating are completed.

Table 2.1

```
##   Rk          Player Pos Age  Tm  G  GS   MP  FG  FGA   FG. X3P X3PA
## 1  1    Alex Abrines  SG  24 OKC 75  8 1134 115  291  0.395  84   221

##    X3P. X2P X2PA  X2P.  eFG.  FT FTA   FT. ORB DRB TRB AST STL BLK TOV  PF
## 1 0.380  31   70  0.443 0.540  39  46 0.848  26  88 114  28  38   8  25 124

##    PTS   X  season17_18      MPG       PPG       APG       RPG       TOG
## 1  353 185     5725000   15.12000  4.706667  0.3733333  1.520000  0.3333333
```

In hopes of incorporating analysis of NBA player's college and/or international stats prior to their entrance into the NBA, a stratified sample was taken based on the five number summary of the salary dataset. Separated into four quarters by salary minimum, first quartile, median, third quartile, and the maximum, a random stratified sample was performed to sample ten player names from each of the four stratas. This total of 40 NBA players from the 2017-2018 season were then separately researched to compile their basketball statistics for the final year of college or the last international season immediately prior to first entering the NBA, whether that be through the National Collegiate Athletic Association (NCAA) in the United States or professional leagues around the world. The college/team and country (if outside the NCAA system) are specified for each player in the dataset. This self-created data frame was created via Google Sheets with all data coming from either Basketball Reference or Sports Reference (citation included in Citations section). Using similar join methods in RStudio, these 40 sampled players were compiled in a data frame with their stats from 2017-2018 in the NBA and their final year of college or international play before their individual entrance into the league. Table 2.2 shows an example entry (Willy Hernangomez) from the data frame that will be used for analysis after wrangling is completed.

Table 2.2

```
##          Player        Pos  Age  Tm  G GS   MP  FG  FGA   FG.  X3P X3PA  X3P.
## 1 Willy Hernangomez C    23   TOT 48  1   495  91   164   0.555  5     12    0.417

##  X2P X2PA  X2P.  eFG.  FT FTA  FT. ORB DRB TRB AST STL BLK TOV  PF PTS

## 1  86  152   0.566 0.570  59    90  0.656  61   122  183    33  18    16    35   64  246

##           Team   cMPG cPPG cFGM cFGA  cFGp cAPG cRPG   MPG      PPG

## 1      Real Madrid 12.6   18.8     7.5     11.2   0.667   1.2    11.0  10.31250  5.125000

##       APG      RPG        TOG

## 1 0.6875000   3.812500   0.7291667
```

[Key difference]ex. cPPG = college/international points per game, PPG = NBA points per game

To begin the process of applying regression, a pairs correlation plot was created to look for any initial associations between salaries and certain statistics. The pairs plot is effective in providing an overview of any correlations in the data that will be further tested and examined. After choosing a preliminary seven or so statistic categories to test for salary correlation, point per game(PPG) and games started rate (GSR), a self created statistic, were shown to have the highest correlation with salary for the 2017-2018 season. The same pairs correlation plot was created for the college/international and NBA comparison with the response chosen to be minutes per game in the NBA (MPG). College/international assists per game (cAPG) and field goal percentage (cFGp) were shown to have the highest initial correlations with NBA minutes per game. Table 2.3 defines certain abbreviations and labels for basketball statistics that are important to understand and refer back as the analysis continues.

Table 2.3

Pos = Position	FG = Field Goals Made	PF = Personal Fouls
Tm = Team	FGA = Field Goals Attempted	PTS = Points
G = Games Played	FG. = Field Goal Percentage	TOV = Turnovers
GS = Games Started	X3P = 3 Point Makes	AST = Assists
MP = Minutes Played	X3PA = 3 Point Attempts	MPG = Minutes per game
FT = Free Throws Made	X3P. = 3 Point Percentage	APG = Assists per game
FTA = Free Throws Attempted	X2P = 2 Point Makes	RPG = Rebounds per game
FT. = Free Throw Percentage	X2PA = 2 Point Attempts	TOG = Turnovers per game
ORB = Offensive Rebounds	X2P. = 2 Point Percentage	PPG = Points per game
DRB = Defensive Rebounds	STL = Steals	eFG. = Effective FG Percentage
TRB = Total Rebounds	BLK = Blocks	GSR = Games Started Rate

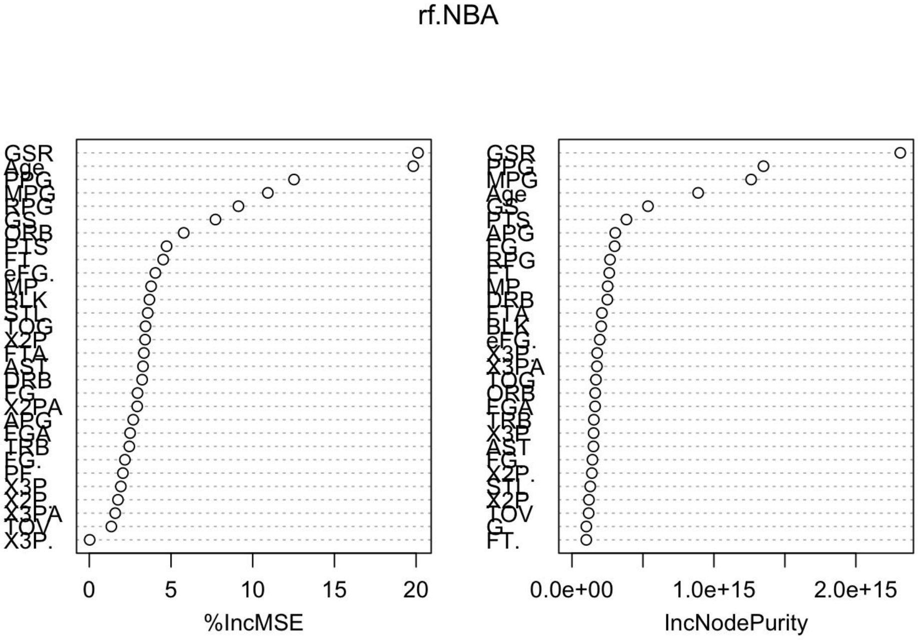
Response: season17_18 = Season Salary * GSR = GS/G

*Note: the “c” before some of these statistics denotes it to college or international season

3.0a Analysis of NBA Statistics and Salaries

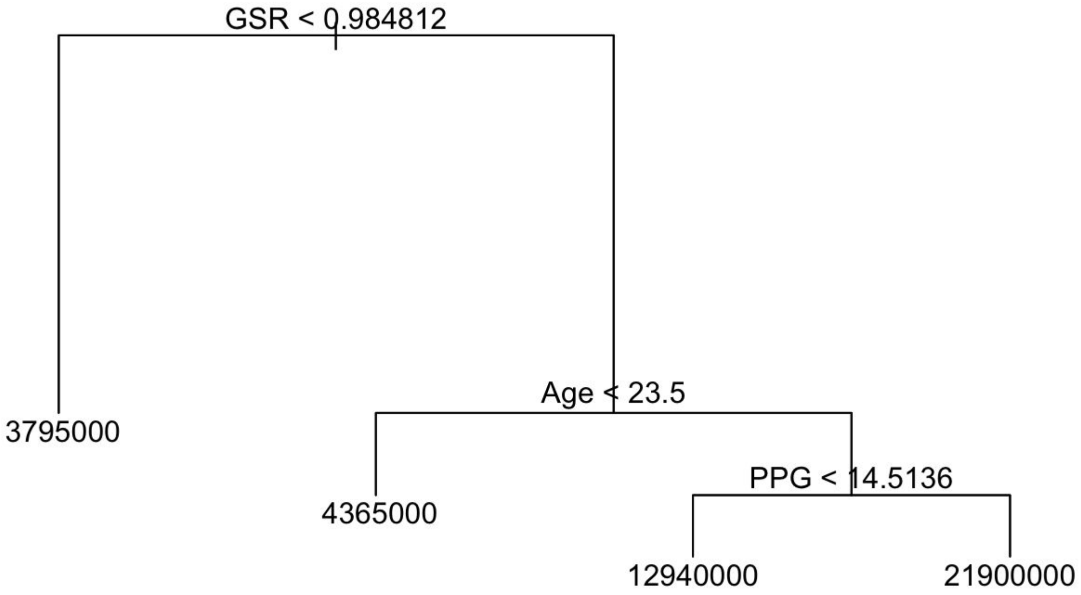
While the pairs correlation plots are helpful in highlighting potential correlations between explanatory variables and the response, it is not a strong enough method to form convincing variable selection. In order to get an idea of the variables that have significant influence, the randomForest package can be used to create an importance plot that examines the percent increase of the mean squared error (MSE) that each variable is responsible for. This comprehensive visual is helpful in accounting for what basketball statistics are more important in explaining a player's salary in the 2017-2018 season and should be tested as a part of the final regression model. Using this algorithm, a random 50-50 split was performed to separate the data into testing and training sets. The training set was used to fit the randomForest model that provides the input for the creation of the variable importance plot. The training and testing technique allows for a portion of the data to be used to create the model and the other to test its appropriateness. Figure 3.1 is the output of the variable importance plot and highlights certain explanatory statistics (including PPG, that have more importance in explaining a player's salary, season17_18).

Figure 3.1



The next step taken in explaining season salary with certain basketball statistics was to apply tree-based methods and visuals to the data set. Tree methods work to segment the predictor space into separate regions, providing simple and useful interpretations with some predicting power. The branches (internal nodes) and leaves (terminal nodes) illustrate splits of certain variables based on being greater than or less than a certain value of a statistic. As it is computationally impossible to look at all possible partitions, this tree method uses recursive binary splitting until the stopping criteria is met. Figure 3.2 illustrates the pruned regression tree for NBA salaries in the 2017-2018 season after cross-validation was performed to find the appropriate number of terminal nodes, set at 4.

Figure 3.2



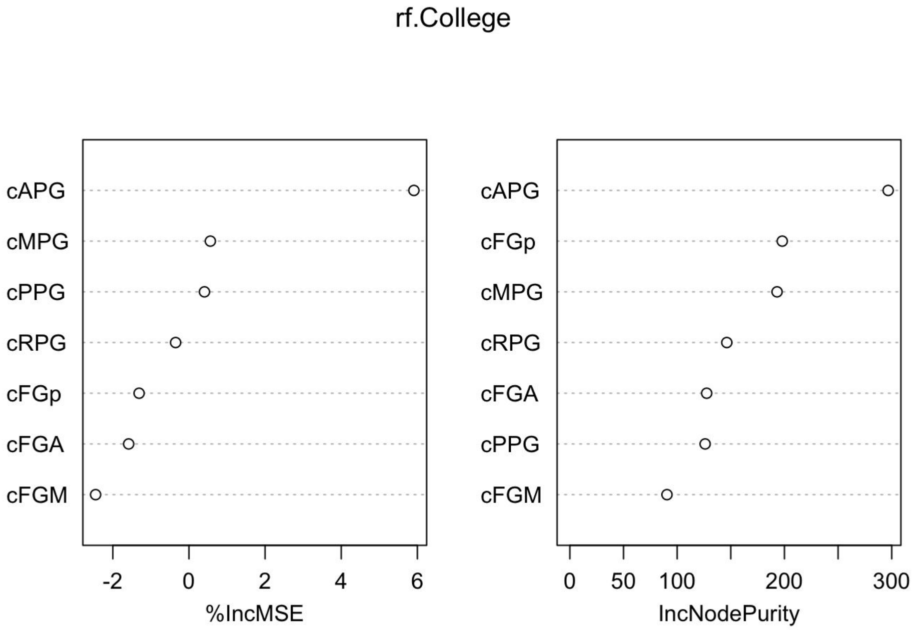
Following the diagram, a player with a GSR greater than 0.984812 and younger than 23.5 years would expect a season salary of around $4,365,000 in the 2017-2018 season. While this sequence of prediction is easy to follow, this regression tree is limited in its predicting power as the terminal nodes are represented by a single value that is applied to a large number of observations, some closer to this estimate than others.

To create the final model based on the statistical data, forward selection was performed on five chosen statistics: PPG, Age, GSR, GS, MPG, and TOG. These explanatory variables were chosen to be tested in the final multiple regression model based on their high initial correlations, large influence on the percent increase in the mean squared error, and presence in the tree diagram. After testing each variable in a simple linear regression format and recording the associated residual sum of squares of each, PPG was determined to be the first statistic in model with the smallest RSS. The process of forward selection was continued in a similar fashion until Age and GSR were added into the model. At the point of potentially adding a fourth statistic, GS, MPG, and TOG were all found to hold no significance. With the final model created, a random sample of one player was taken to compare the actual and predicted salary based on the three basketball stats: PPG + Age + GSR. Using the predict function, Allen Crabbe's stats (13.2 PPG, 25 years old, GSR = 0.9066667) with the regression model predicted a season salary of $11,381,341 with his actual salary for the 2017-2018 season being $19,332,500.

3.0b Analysis of College/International and NBA Statistics (Sample)

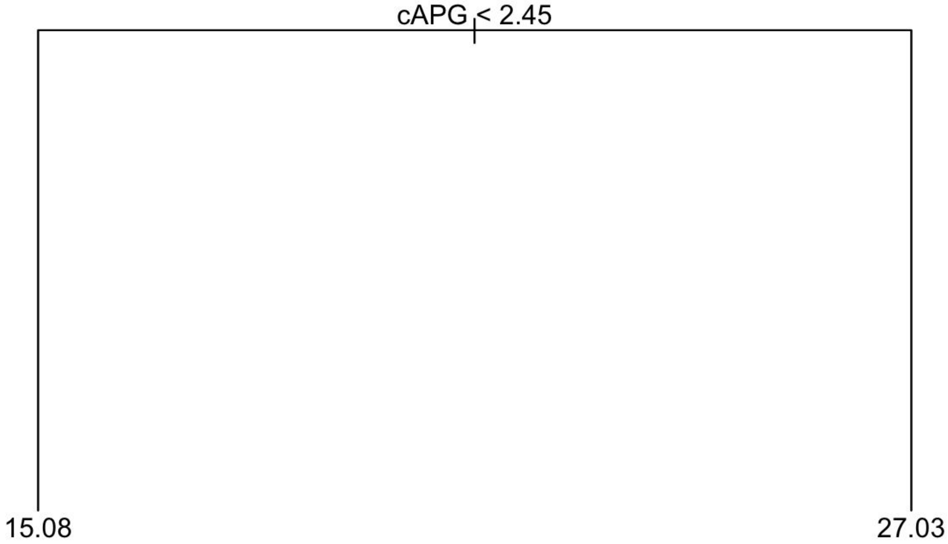
Similar to the NBA statistics and salaries analysis, it was also helpful to create a variable importance plot to begin to visualize certain college/international statistics after initial pairs correlation plots were created. While the previous importance plot contained all the variables from the NBA dataset, this analysis required limiting the explanatory variables to only statistics from college/international (identified with a "c"). This subset stems from the original question of interest that is looking into whether certain a player's basketball performance stats in the season prior to entering the NBA can help explain their MPG (chosen as response earlier) once they are in the league. Figure 3.3 shows the percent increase in mean squared error (MSE) of each of these explanatory variables.

Figure 3.3



Similarly, a tree plot was also constructed as a simple and useful method to observe what college/international statistics might explain minutes on a per game basis (MPG) for a player once he enters the league. Figure 3.4 depicts the tree plot with three terminal nodes split at cAPG.

Figure 3.4



In comparison to the previous tree plot withe the NBA salaries and statistics, this diagram is much more simplistic and required no additional pruning. The difference in complexity between the two could potentially be attributed to the number of variables and the limited total sample size of 40 in the college/international dataset. Following the diagram, a player in his college or international season prior to entering the NBA who averaged more than 2.45 APG would expected around 27.03 MPG in the 2017-2018 NBA season.

In constructing the final regression model, forward selection was again utilized but in this case only cAPG was shown to be significant. After simple regression techniques identified cAPG as having the lowest residual sum of squares, the adding of any other variable (cFGM, cFGA, cFGp, cRPG, cPPG, cMPG) was deemed insignificant in explaining MPG in the NBA. From this conclusion, it is important to consider the limited sample size of 40 players in the college/international dataset and the difficulty in explaining MPG in an NBA season with past stats from college or overseas.

4.0 Conclusions and Discussion

While the multiple regression model results identified PPG, Age, and GSR as the three explanatory variables that were significant in explaining a player's salary in the 2017-2018 season, it is important to consider the entire analysis when engaging in discussion. These three statistics hold convincing evidence of having a relationship with salary as PPG, Age, and GSR were also found as predictors in the pruned tree plot. Before analysis began, Age was not thought to be a key explanatory statistic because there are many examples in the NBA where younger players are paid more than older players and vice versa. However, when bringing in further context of the NBA's rules and regulations, rookie (typically younger players) contracts are restricted to a certain dollar by the position in which they were drafted. As these rookies grow older and finish out these initial contracts, they are eligible to sign more lucrative deals for a higher annual salary. When looking at the final prediction function and the testing of Allen Crabbe and his stats, it is interesting to consider the application of these statistical methods. With a residual of almost $8,000,000, these results ask the question if Crabbe is being overpaid for his performance or if there is another metric that justifies his salary of upwards of $19,000,000.

In comparing the two separate pathways of analysis, it is evident that the NBA statistics and salaries investigation yielded more convincing results of a relationship than the College/International data. The smaller sample size of 40 players must be taken into account as well as the time removed between the final season of college or international basketball for a player and his 2017-2018 NBA season. However, the results portrayed in the tree diagram (Figure 3.4) make contextual sense in the landscape of the modern NBA. Being a guard dominated league (guards typically average the most assists), guards might expect more playing time than other positions like centers or power forwards. Nonetheless, relationships between college and NBA basketball should be further explored to inform athletes of the statistics that best translate to the professional level.

5.0a Appendix: NBA Statistics and Salaries

~~~~
```{r}
library(tidyverse)
library(tidyr)
library(leaps)
library(tree)
library(randomForest)
library(GGally)
```
~~~~

Packages required to run R code and analysis.

~~~~
```{r}
salary<-read.csv("NBA_season1718_salary (1).csv", header=TRUE,
      stringsAsFactors = FALSE)
head(salary)
player_stats<-read.csv("nba_extra2.csv", header=TRUE,
      stringsAsFactors = FALSE)
head(player_stats)
```
~~~~

Loading and labeling of the two datasets that will be manipulated and joined to create the final dataset that will be used in the NBA Statistics and Salaries analysis.

~~~~
```{r}
player_stats2<-player_stats%>%
  filter(is.na(Rk)==FALSE)
player_stats3 <- player_stats2[-c(24, 25, 28, 29, 45, 46, 59, 60, 68, 69, 70,
73, 74, 77, 78, 97, 98, 102, 103, 118, 119, 122, 123, 142, 143, 145, 146, 167,
168, 174, 175, 180, 181, 190, 191, 202, 203, 229, 230, 237, 238, 242, 243, 250,
251, 252, 259, 260, 266, 267, 276, 277, 288, 289, 299, 300, 306, 307, 314, 315,
319, 320, 328, 329, 331, 332, 342, 342, 343, 344, 368, 369, 386, 387, 401, 402,
419, 420, 425, 426, 427, 439, 440, 442, 443, 450, 451, 454, 455, 471, 472, 481,
482, 493, 494, 496, 497, 523, 524, 535, 536, 550, 551, 568, 569, 585, 586, 602,
603, 604, 606, 607, 610, 611, 623, 624, 642, 643, 650, 651, 660, 661),]
head (player_stats3)
```
~~~~

Due to the format of the Kaggle dataset "nba_extra2" accounting for certain players playing for multiple teams in one year (trades, free-agent signings, etc.), some players originally had three rows or sets of observations: one for each team they played for and a total row (the two separate team stats together). These instances were manually identified and removed using this section of code. After this procedure, each NBA player only had one row/set of observations that accounts for his total statistics from that season.

~~~~
```{r}
playerS<-separate(data=player_stats3,col=Player,into=c("Player", "nickname"),
sep=":")
head(playerS)
~~~~

~~~
```
~~~

In this section of code the separate function is used to separate the "Player" and "nickname" by ':' so the two datasets could be joined by the similar column of "Player" each player's full name.

~~~
```{r}
data<-left_join(playerS, salary)
data<-data%>%
  mutate(MPG = MP/G, PPG = PTS/G, APG = AST/G, RPG = TRB/G, TOG = TOV/G, GSR =
GS/G)
data<-na.omit(data)
attach(data)
```
~~~

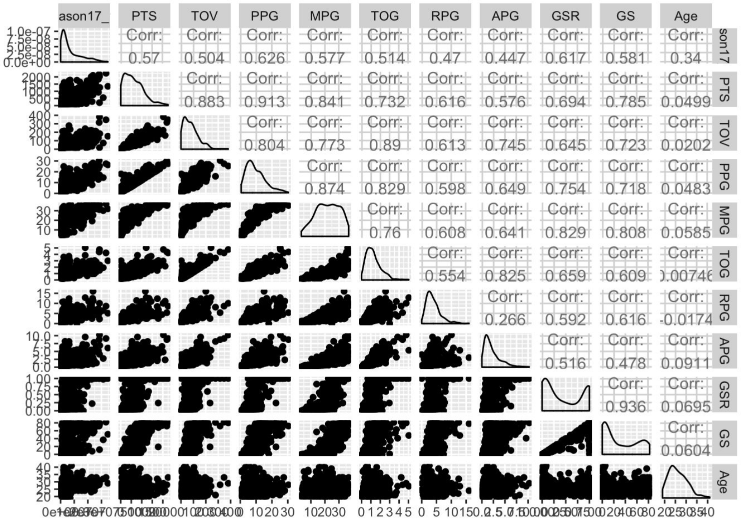
This section of code joins "playerS" and "salary" using the left_join function and creates the per game basis metrics using the mutate function. Na.omit works to remove any empty values (NAs) from the dataset.

~~~
```{r}
data.pure <- data
data.pure$Player <- data.pure$nickname <- data.pure$Tm <- data.pure$Pos <-
data.pure$Rk <- data.pure$X <- NULL
```
~~~

The final dataset titled “data” is copied and any non-numerical or unimportant variable is removed. This eliminates any confusion in later outputs of pairs correlation plots and puts a focus on the statistics that will be used in analysis.

~~~
```{r}
stats_salary_cor <- data %>%
  select(season17_18, PTS, TOV, PPG, MPG, TOG, RPG, APG, GSR, GS, Age)
ggpairs(stats_salary_cor)
cor(stats_salary_cor)[,"season17_18"]
```
~~~

A pairs correlation plot is created to highlight any initial correlation between 2017-2018 season salaries and certain basketball performance statistics.



~~~
```{r}
set.seed(1)
train<-sample(1:nrow(data.pure), nrow(data.pure)/2)
NBA.test <- data.pure[-train, "season17_18"]
rf.NBA<-randomForest(season17_18~., data=data.pure, subset=train,
importance=TRUE)
yhat.rf<-predict(rf.NBA, newdata=data.pure[-train,])
mean((yhat.rf-NBA.test)^2)

importance(rf.NBA)
varImpPlot(rf.NBA)
```
~~~

Using the randomForest package, a training subset is taken from the data.pure dataset and applied to produce a variable importance plot (discussed in the body of the report, Figure 3.1).

~~~
```{r}
set.seed(1)
train<-sample(1:nrow(data.pure), nrow(data.pure)/2)
NBA.test <- data.pure[-train, "season17_18"]
tree.NBA<-tree(season17_18~., data.pure, subset=train)
summary(tree.NBA)

plot(tree.NBA)
text(tree.NBA, pretty = 0)
```
~~~

The same train subset is created and applied to the tree algorithm to create the recursive binary splitting diagram (discussed in the body of the report).

~~~
```{r}
cv.NBA<-cv.tree(tree.NBA)
plot(cv.NBA$size, cv.NBA$dev, type='b')
~~~

~~~
prune.NBA<-prune.tree(tree.NBA, best=4)
plot(prune.NBA)
text(prune.NBA, pretty=0)
```
~~~

Cross-validation for the number of terminal nodes (4) is performed and used in coordination with the prune.tree function is modify the tree plot (Figure 3.2).

~~~
```{r}
mod1a <- lm(season17_18~MPG, data = data.pure)
anova(mod1a)      # RSS = 1.4034e+16

mod1b <- lm(season17_18~PPG, data = data.pure)
anova(mod1b)     # RSS = 1.2803e+16

mod1c <- lm(season17_18~Age, data = data.pure)
anova(mod1c)     # RSS = 1.8605e+16

mod1d <- lm(season17_18~GS, data = data.pure)
anova(mod1d)     # RSS = 1.3925e+16

mod1e <- lm(season17_18~GSR, data = data.pure)
anova(mod1e)     #RSS = 1.3041e+16

mod1f <- lm(season17_18~TOG, data = data.pure)
anova(mod1f)     # RSS = 1.5474e+16
```
~~~

First step of forward selection: testing each explanatory variable in a simple regression format and comparing residual sum of squares (RSS). PPG determined to be the first variable in the model.

~~~
```{r}
mod2a <- lm(season17_18~PPG + Age, data = data.pure)
anova(mod2a)     # RSS = 1.078e+16

mod2b <- lm(season17_18~PPG + GSR, data = data.pure)
anova(mod2b)     # RSS = 1.1779e+16

mod2c <- lm(season17_18~PPG + MPG, data = data.pure)
anova(mod2c)     # RSS = 1.2722e+16

mod2d <- lm(season17_18~PPG + GS, data = data.pure)
anova(mod2d)     # RSS = 1.2045e+16

mod2e <- lm(season17_18~PPG + TOG, data = data.pure)
anova(mod2e)     # TOG not significant
```
~~~

Second step of forward selection: Age determined to be the second variable in the model.

~~~
```{r}
mod3a <- lm(season17_18~PPG + Age + GSR, data = data.pure)
anova(mod3a)     #RSS = 9.8925e+15

mod3b <- lm(season17_18~PPG + Age + MPG, data = data.pure)
anova(mod3b)     # MPG not significant

mod3c <- lm(season17_18~PPG + Age + GS, data = data.pure)
anova(mod3c)     # RSS = 1.0108e+16

mod3d <- lm(season17_18~PPG + Age + TOG, data = data.pure)
anova(mod3d)     # TOG not significant
```
~~~

Third step of forward selection: GSR is determined to be the third variable in the model.

~~~
```{r}
mod4a <- lm(season17_18~PPG + Age + GSR + MPG, data = data.pure)
anova(mod4a)     # MPG not significant

mod4b <- lm(season17_18~PPG + Age + GSR + GS, data = data.pure)
anova(mod4b)     # GS not significant

mod4c <- lm(season17_18~PPG + Age + GSR + TOG, data = data.pure)
anova(mod4c)     # TOG not significant
```
~~~

Fourth step of forward selection: No other variables determined to be significant.

~~~
```{r}
set.seed(52)
data[sample(nrow(data),1),]
final.model <- lm(season17_18~PPG + Age +GSR, data = data.pure)
anova(final.model)
new.data<- data.frame(PPG = 13.2, Age = 25, GSR = 1)
final.test <- predict(final.model, new.data)
final.test
```
~~~

This section of code takes a random sample for one row/player from the data set and uses the predict function to input Allen Crabbe's stats to create a predicted salary output.

5.0b Appendix: College/International and NBA Statistics

~~~~
```{r}
library(tidyverse)
library(tidyr)
library(leaps)
library(tree)
library(randomForest)
library(GGally)
```
~~~~

Packages required to run R code and analysis.

~~~~
```{r}
firstQT<-data%>%
  filter(season17_18<1465920)
dim(firstQT)
set.seed(1)
samp1<-sample(121, 10)
sampFirst<-firstQT[samp1,]

secondQT<-data%>%
  filter(season17_18>=1465920 & season17_18<3028410)
dim(secondQT)
set.seed(2)
samp2<-sample(122, 10)
sampSecond<-secondQT[samp2,]

thirdQT<-data%>%
  filter(season17_18>=3028410 & season17_18<9539057)
dim(thirdQT)
set.seed(3)
samp3<-sample(122, 10)
sampThird<-thirdQT[samp3,]

fourthQT<-data%>%
  filter(season17_18>=9539097 & season17_18<=34682550)
dim(fourthQT)
set.seed(4)
samp4<-sample(122, 10)
sampFourth<-fourthQT[samp4,]
```
~~~~

This set of code performs a stratified sample based on the five number summary of the overall dataset (all the players' salaries in the NBA during the 2017-2018 season) to create four stratas based on salary in dollar amounts. The results of each strata were used to separately research and create a dataframe with these 40 players (10 from each strata). This new dataset will be used for statistical analysis.

~~~
```{r}
college.data <- read.csv("Stratified Sample_NBA stats_Sheet1.csv",
      header = TRUE, stringsAsFactors = FALSE)
head(college.data)
college<-read.csv("Stat Project_College_International Data_Sheet1.csv",
      header=TRUE, stringsAsFactors = FALSE)
head(college)
```
~~~

Loading and labeling of the two data sets that will be used in the College and NBA statistics analysis.

~~~
```{r}
pro.college <- left_join(college.data, college)
pro.college<-pro.college%>%
  mutate(MPG = MP/G, PPG = PTS/G, APG = AST/G, RPG = TRB/G, TOG = TOV/G)
pro.college <- na.omit(pro.college)
head(pro.college)
```
~~~

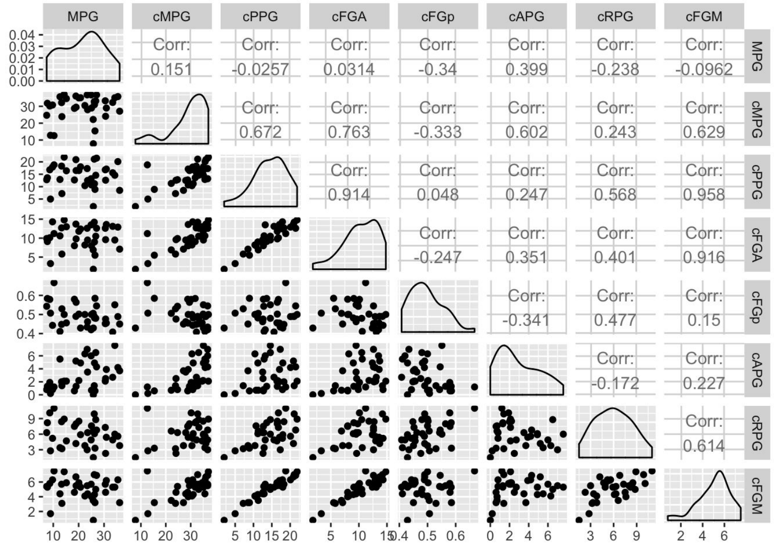
This section of code joins "college.data" and "college" by the Player column using the left_join function and creates the per game basis metrics using the mutata function. Na.omit works to remove any empty values (NAs) from the dataset.

~~~
```{r}
pro.college <- na.omit(pro.college)
pro.college.pure <- pro.college
pro.college.pure$Player <- pro.college.pure$Tm <- pro.college.pure$Team <-
pro.college.pure$Pos <- NULL
```
~~~

The final dataset titled “pro.college” is copied and any non-numerical or unimportant variable is removed. This eliminates any confusion in later outputs of pairs correlation plots and puts a focus on the statistics that will be used in analysis.

~~~
```{r}
stats_college_cor <- pro.college %>%
  select(MPG, cMPG, cPPG, cFGA, cFGp, cAPG, cRPG, cFGM)
ggpairs(stats_college_cor)
cor(stats_college_cor)[,"MPG"]
```
~~~

A pairs correlation plot is created to highlight any initial correlation between MGP in the 2017-2018 NBA season and certain basketball performance statistics from the college or international season prior to entering the league.



~~~
```{r}
set.seed(2)
train<-sample(1:nrow(pro.college.pure), nrow(pro.college.pure)/2)
set.seed(2)
rf.College<-randomForest(MPG~cMPG+cRPG+cPPG+cAPG+cFGM+cFGA+cFGp,
data<-pro.college.pure, subset=train, importance=TRUE)
yhat.rf<-predict(rf.College, newdata=pro.college.pure[-train,])
mean((yhat.rf-NBA.test)^2)
importance(rf.College)
varImpPlot(rf.College)
```
~~~

Using the randomForest package, a training subset is taken from the pro.college.pure dataset and applied to produce a variable importance plot (discussed in the body of the report, Figure 3.3).

~~~
```{r}
set.seed(2)
trainC<-sample(1:nrow(pro.college.pure), nrow(pro.college.pure)/2)
College.test <- data.pure[-train, "MPG"]
tree.College<-tree(MPG~cMPG+cRPG+cPPG+cAPG+cFGM+cFGA+cFGp, pro.college.pure,
subset=trainC)
summary(tree.College)
plot(tree.College)
text(tree.College, pretty = 0)
cv.College<-cv.tree(tree.College)
plot(cv.College$size, cv.College$dev, type='b')
```
~~~

This code sets a seed for the train subset of the pro.college.pure dataset and uses that subset to apply the tree algorithm and plot (Figure 3.4). No pruning by cross-validation needed in this instance.

~~~
```{r}
MOD1a <- lm(MPG~cRPG, data = pro.college.pure)
anova(MOD1a)     # RSS = 2324.44

MOD1b <- lm(MPG~cAPG, data = pro.college.pure)
anova(MOD1b)     # RSS = 2070.7

MOD1c <- lm(MPG~cFGA, data = pro.college.pure)
anova(MOD1c)     # RSS = 2440.96

MOD1d <- lm(MPG~cFGp, data = pro.college.pure)
anova(MOD1d)     # RSS = 2178.41

MOD1e <- lm(MPG~cPPG, data = pro.college.pure)
anova(MOD1e)     # RSS = 2462.14

MOD1f <- lm(MPG~cMPG, data = pro.college.pure)
anova(MOD1f)     # RSS = 2407.75
```
~~~

First step of forward selection: testing each explanatory variable in a simple regression format and comparing residual sum of squares (RSS). cAPG determined to be the first variable in the model.

~~~
```{r}
MOD2a <- lm(MPG~cAPG + cFGA, data = pro.college.pure)
anova(MOD2a)

MOD2b <- lm(MPG~cAPG + cPPG, data = pro.college.pure)
anova(MOD2b)

MOD2c <- lm(MPG~cAPG + cFGM, data = pro.college.pure)
anova(MOD2c)

MOD2d <- lm(MPG~cAPG + cMPG, data = pro.college.pure)
anova(MOD2d)

MOD2e <- lm(MPG~cAPG + cRPG, data = pro.college.pure)
anova(MOD2e)

MOD2f <- lm(MPG~cAPG + cFGp, data = pro.college.pure)
anova(MOD2f)
```
~~~

Second step of forward selection: No other variables determined to be significant, cAPG is the only variable to hold significance with the response of MPG in the NBA.

6.0 Citations

"Basketball Statistics and History." *Basketball Reference*, www.basketball-reference.com/.

Camli, Baris. "NBA Player Stats 2017-2018." *Kaggle*, 4 Nov. 2018, www.kaggle.com/mcamli/nba17-18.

"Sports Reference: Sports Stats, Fast, Easy, and up-to-Date." *Sports Reference*, www.sports-reference.com/.