Jeffrey Hou

Data 110

3/2/19

Exam 1 -NBA Statistics

Data: <https://www.kaggle.com/drgilermo/nba-players-stats-20142015>

For this assignment, I chose a dataset which I have a big interest in. Since a small child, I’ve followed NBA players and watched numerous NBA games. This dataset analyzes the top 490 players in the NBA in 2014-2015. This data set is very important to me since even before I was involved with data science, I have personally tried to analyze this type of data when looking at my favorite players and their statistics. As an avid Stephen Curry fan, I used to look up his shooting statistics just to figure out what made him so successful in scoring the ball. In the NBA, they do provide data for analysts to study to come up with either a prediction model or just a general trend on how the players can improve their shots, either through shot charts or hot zones. Using R, I am learning how to analyze and organize data, and one day in the future maybe I can work as a statistician in the NBA.

In this data set, there are 34 variables. This data set was pulled directly from the website Kaggle, which has a huge database of datasets. This dataset stood out to me since it was directly related to one of my hobbies. The data set variables are:

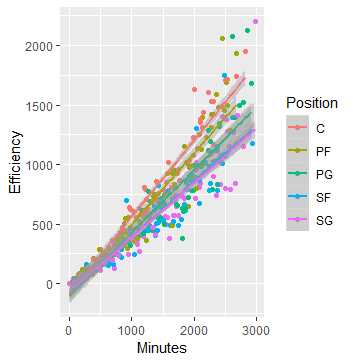
* Name
* Games Played
* MIN
* PTS
* FGM
* FGA
* FG%
* Age
* Birth Place
* College
* Experience
* Height
* Team
* Position
* Weight
* BMI
* 3PM
* 3PA
* 3P%
* FTM
* FTA
* FT%
* OREB
* DREB
* REB
* AST
* STL
* BLK
* TOV
* PF
* EFF
* AST/TOV
* STL/TOV

Types of variables:

* Factors- Experience
* Integer-Games Played, Minutes, Points, FGM, FGA, X3PM, X3PA, FTM, FTA, OREB, DREB, REB, AST, STL, BLK, TO, PF, EFF, Age,
* Numerals- FG., FT., X3P., AST.TOV, STL.TOV, Height, Weight, BMI
* Character: Name, Birth Place, Birth Date, College, Position, Team

The variable types vary from Factor, to Integer, to Numerals. For the factor type, the only variable that was needed to be a factor was experience, since you could rank someone from the least experienced to the most experienced. The variables that I changed from factor to character were College, Name, Position, Birth Date, Birth Place, and Team. Those variables do not have a general level system, as we can’t say a Position has a higher ranking or a college is higher ranked than another one.

The second part of cleaning up, I decided to rename all the variables into something that the public would understand, so all abbreviations are their full names. I also found a typo in the “collage” variable. I used the names function in R to rename it to college. The last problem I ran into when trying to create a graph was trying to get rid of all the NA values so they would not show up as missing points on my graph. Using na.strings, I was able to tell the program to change any variables that were “ “ ie empty into a NA value. The only problem with that was that it would still include NA as a variable under all the positions. To get rid of that problem, I used the scale colour discrete to hide the NA values in the graph. Using the GGplot2 package, I was able to use the geom\_jitter which allows for clearer visualization as it helps to avoid overplotting.



For the graph, I decided to analyze two factors, Minutes and Efficiency based on the position they play. I was expecting a positive correlation, since it would only make sense that the more minutes someone plays, the higher their efficiency would be. Efficiency is the only variable in basketball that can be used to accurately judge if a player is good or bad at the game. The efficiency variable is made from a formula: ((Points + Rebounds + Assists + Steals + Blocks) - ((Field Goals Att. - Field Goals Made) + (Free Throws Att. - Free Throws Made) + Turnovers)). With all these variables, to have a high rating, you must be good at all assets of the game, or else that one variable will significantly pull down your efficiency rating.

The one thing that was interesting to me was how all positions start very similar within the first 1000 minutes, but after that, some positions have a stronger positive trend compared to others. Centers surprisingly start to have a stronger efficiency rating the longer they play, compared to SGs and SFs which have a much lower efficiency rating. This would make sense since centers are usually shooting closer to the basketball, which means they have a better chance of making the shot compared to SGs and SFs who usually shoot the ball from a farther distance. A higher percentage of shots going in means less turnovers, which is one of the key variables in the Efficiency formula. One thing the graph does a good job of is using colors to designate the positions, which allows you to see very quickly which variables represent which dots. The regression lines also help since it shows the approximate average trend of all the data within each variable. One other very interesting thing this graph shows is how SGs usually struggle with efficiency, but there is one outlier at 3000 who defies all odds and has a sky-high efficiency rating. This player most likely is a strong all-around player, which means he knows his strengths and weaknesses and plays to his strengths. This allows him to keep his efficiency rating high. Looking through my data, I was able to find that the player that represents this dot is James Harden, who currently is a superstar and MVP in the NBA. Crazy how data 15 years ago can accurately predict future NBA all-stars.

RCODE:

#dataset is https://www.kaggle.com/drgilermo/nba-players-stats-20142015

library(ggplot2)

library(plyr)

nba<-read.csv("C:/Users/Jeffo\_000/Desktop/Tennis/nba.csv",header=T, na.strings="") # reads in the data from nba.csv and gives NA values for all empty values

nba$Collage<-as.character(nba$Collage) #changes variables from factor to character

nba$Name<-as.character(nba$Name)

nba$Pos<-as.character(nba$Pos)

nba$Birthdate<-as.character(nba$Birthdate)

nba$Birth\_Place<-as.character(nba$Birth\_Place)

nba$Team<-as.character(nba$Team)

str(nba) #shows the types of variable that are in the datatset

nba<-rename(nba, c("Collage" = "College")) #rename variables into simple terms

nba<-rename(nba, c("Birth\_Place" = "Birth Place"))

nba<-rename(nba, c("X3PM" = "3PM"))

nba<-rename(nba, c("X3PA" = "3PA"))

nba<-rename(nba, c("X3P."="3P"))

nba<-rename(nba, c("Games.Played"="Games Played"))

nba<-rename(nba, c("TOV"="TO"))

nba<-rename(nba, c("AST.TOV"="Assists/TO"))

nba<-rename(nba, c("STL.TOV"="Steals/TO"))

nba<-rename(nba, c("PTS"="Points"))

nba<-rename(nba, c("MIN"="Minutes"))

nba<-rename(nba, c("Pos"="Position"))

nba<-rename(nba, c("EFF"="Efficiency"))

ggplot(data=nba, aes(x=Minutes, y=Efficiency, color=Position))+ #plots data based on x,y variables

geom\_jitter(data=nba,aes(x=Minutes, y=Efficiency, color=Position))+ #avoids overplotting

geom\_smooth(method="lm")+ #creates regression lines for each variable

scale\_colour\_discrete(na.translate = F) #omits the NA values in the Position variable when plotting