ggplot for Sports Analytics

Bryan Stapleton

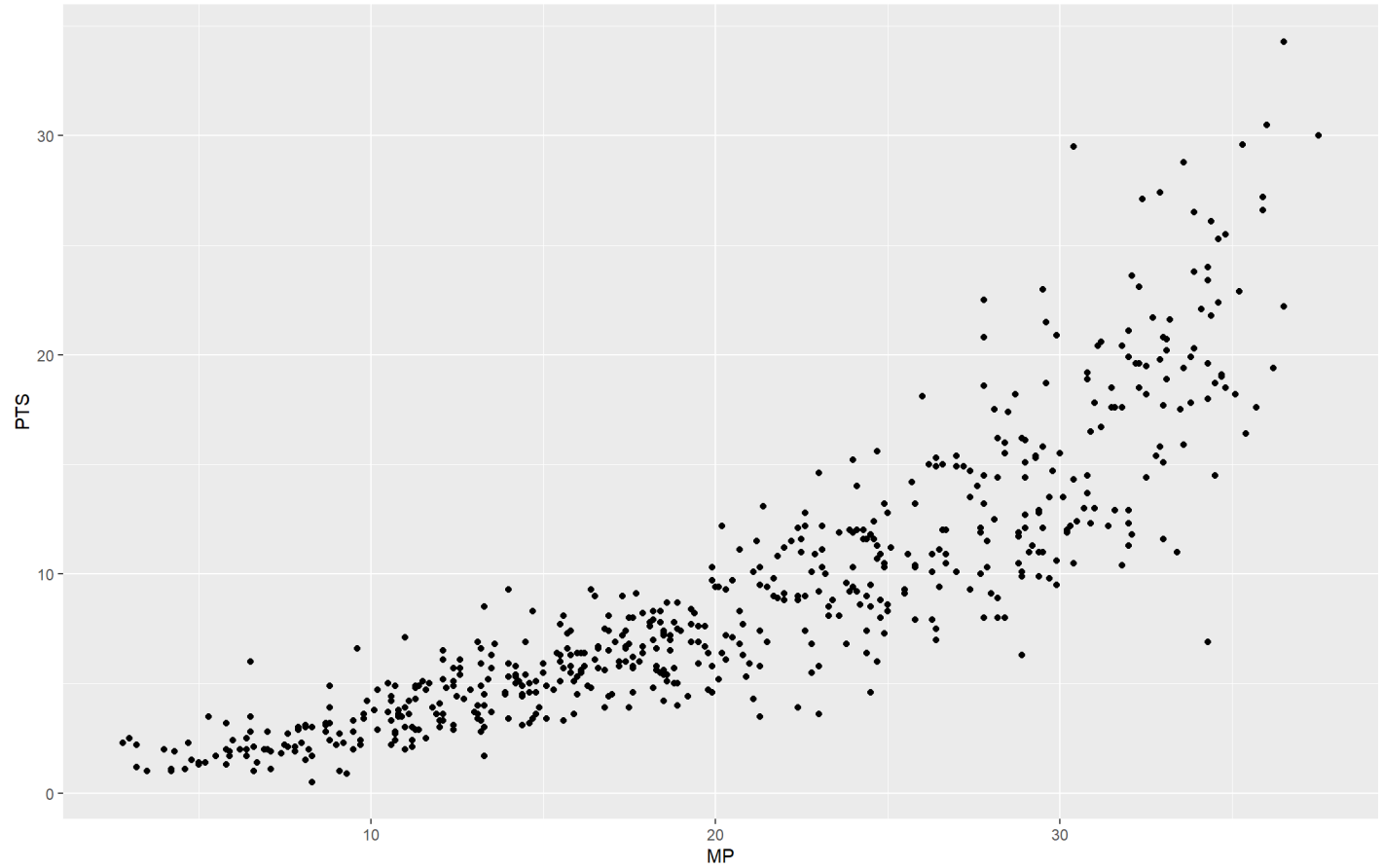
September 30, 2020

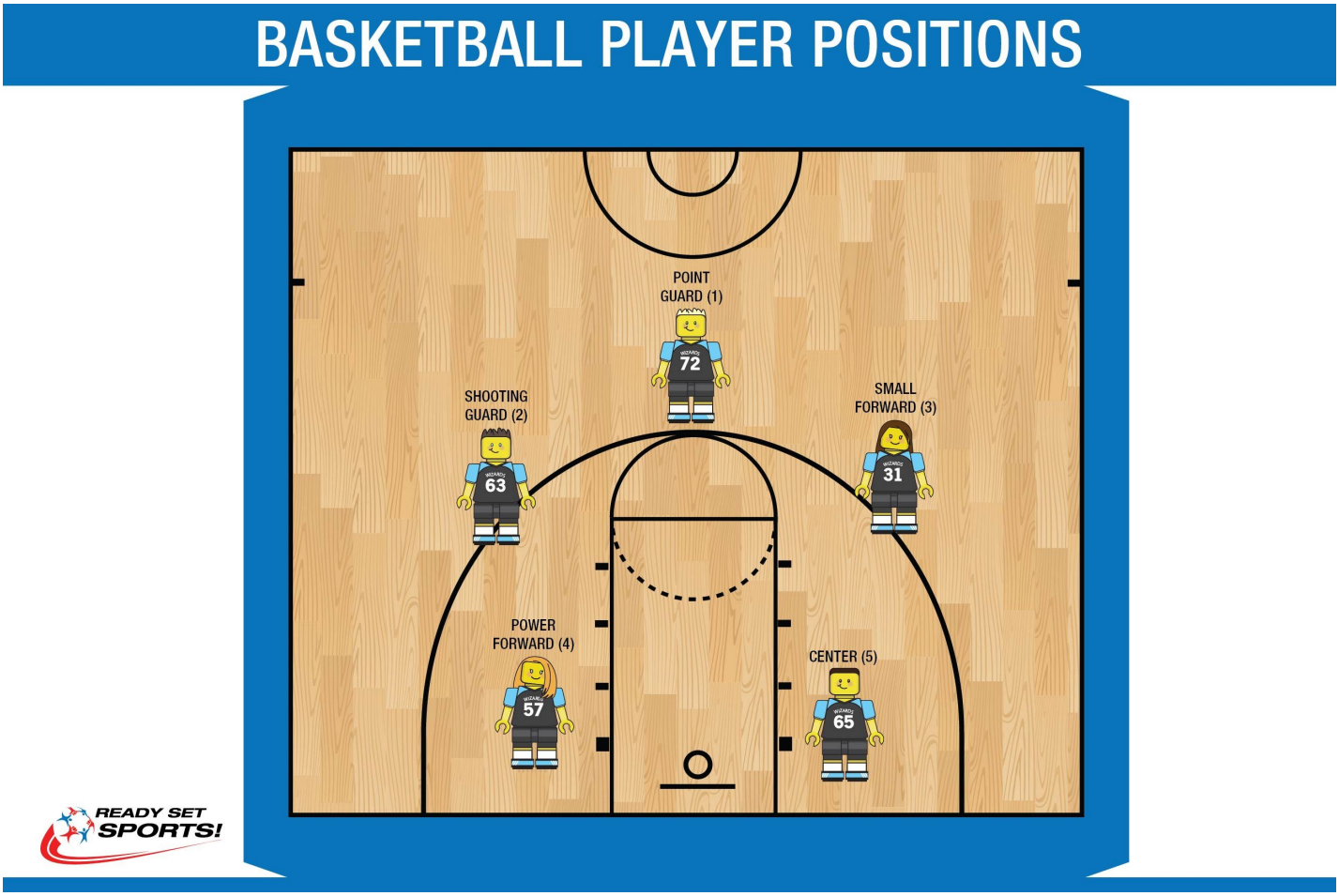
What is it ggplot2 is a data visualization package for R that breaks graphs into scales and layers. It is part of the tidyverse which is a collection of R packages (dplyr, readr, etc.) designed for data science.
How does it Work Let me show you! I got my data From baksetball-reference.com. I am look at the 2019-2020 NBA season.

```
NBA2020 <- read.csv("C:/Users/bbsta/Downloads/Finalformatnba.csv")
library(ggplot2)
#install.packages("plyr")
NBA2020 <- na.omit(NBA2020)
```

Relationship between minutes played and points.

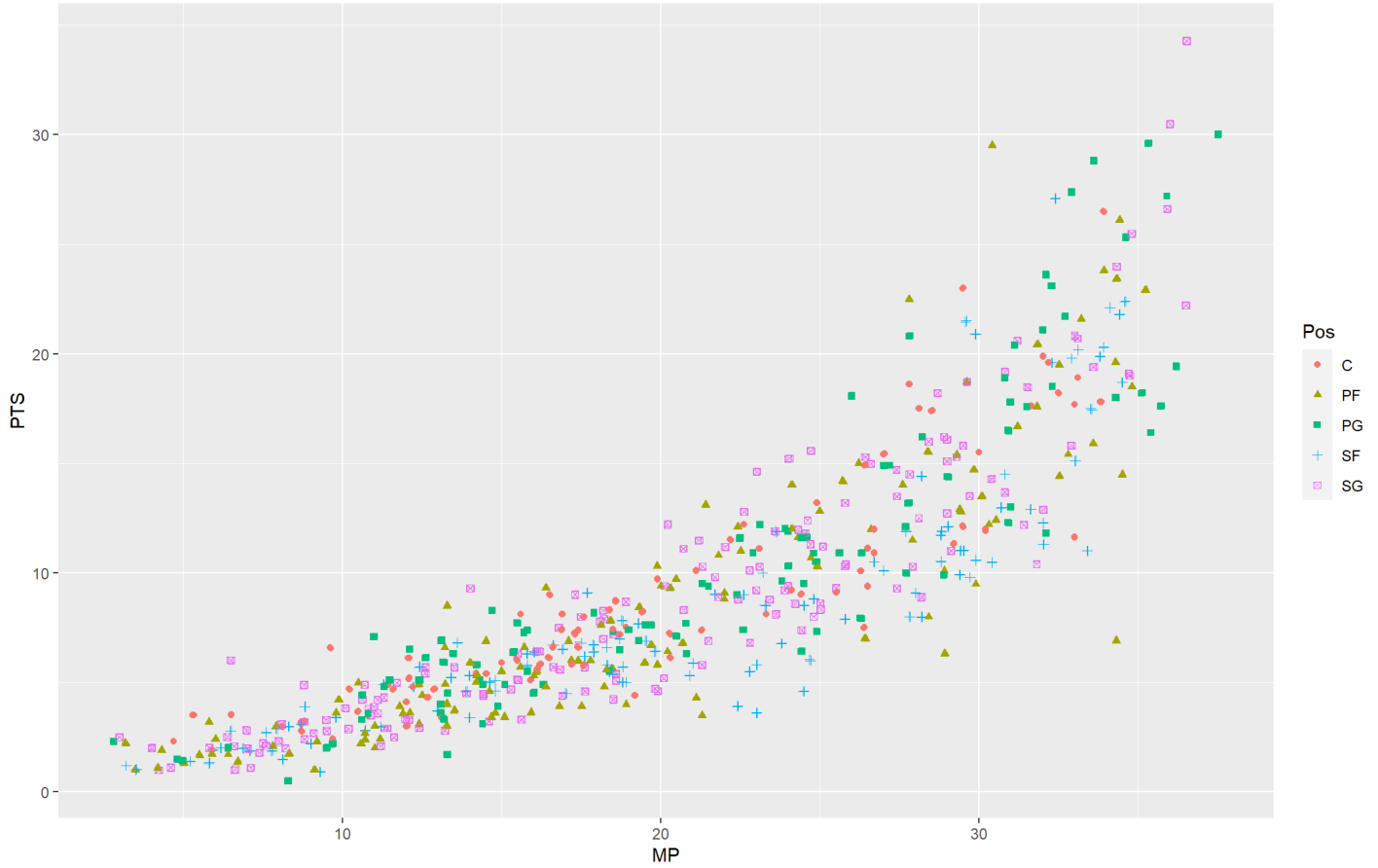
```
#attach(NBA2020)
ggplot(data = NBA2020, mapping = aes(x = MP, y=PTS)) +
  geom_point()
```





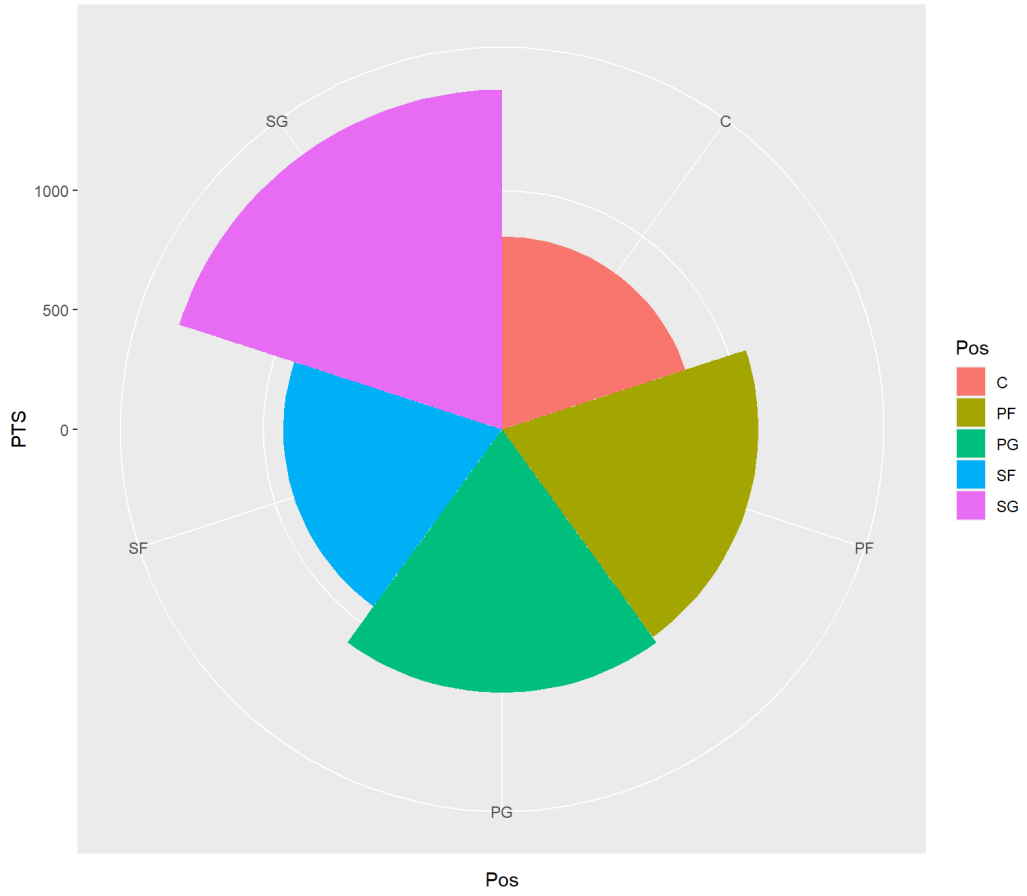
Set the color and shape parameter player position.Geom_Jitter adds a bit of randomness to the points to prevent overplotting.

```
library(plyr)
NBA2020$Pos <- revalue(NBA2020$Pos, c("C-PF" = "C", "SF-PF" ="PF", "SF-SG" = "SF", "SG-PF"= "SG"
, "PF-C" ="PF"))
ggplot(data = NBA2020, mapping = aes(x = MP, y=PTS, color = Pos, shape = Pos)) +
  geom_point() + geom_jitter()
```



coor_polar plots it on a polor coordinate system. We can see that Shooting Gaurds contribute most to scoring.However this may be due to the largest number of players at this position.

```
ggplot(data = NBA2020,mapping = aes(x = Pos, y = PTS, fill = Pos)) +
  geom_bar(stat = "identity",width = 1) +
  coord_polar()
```

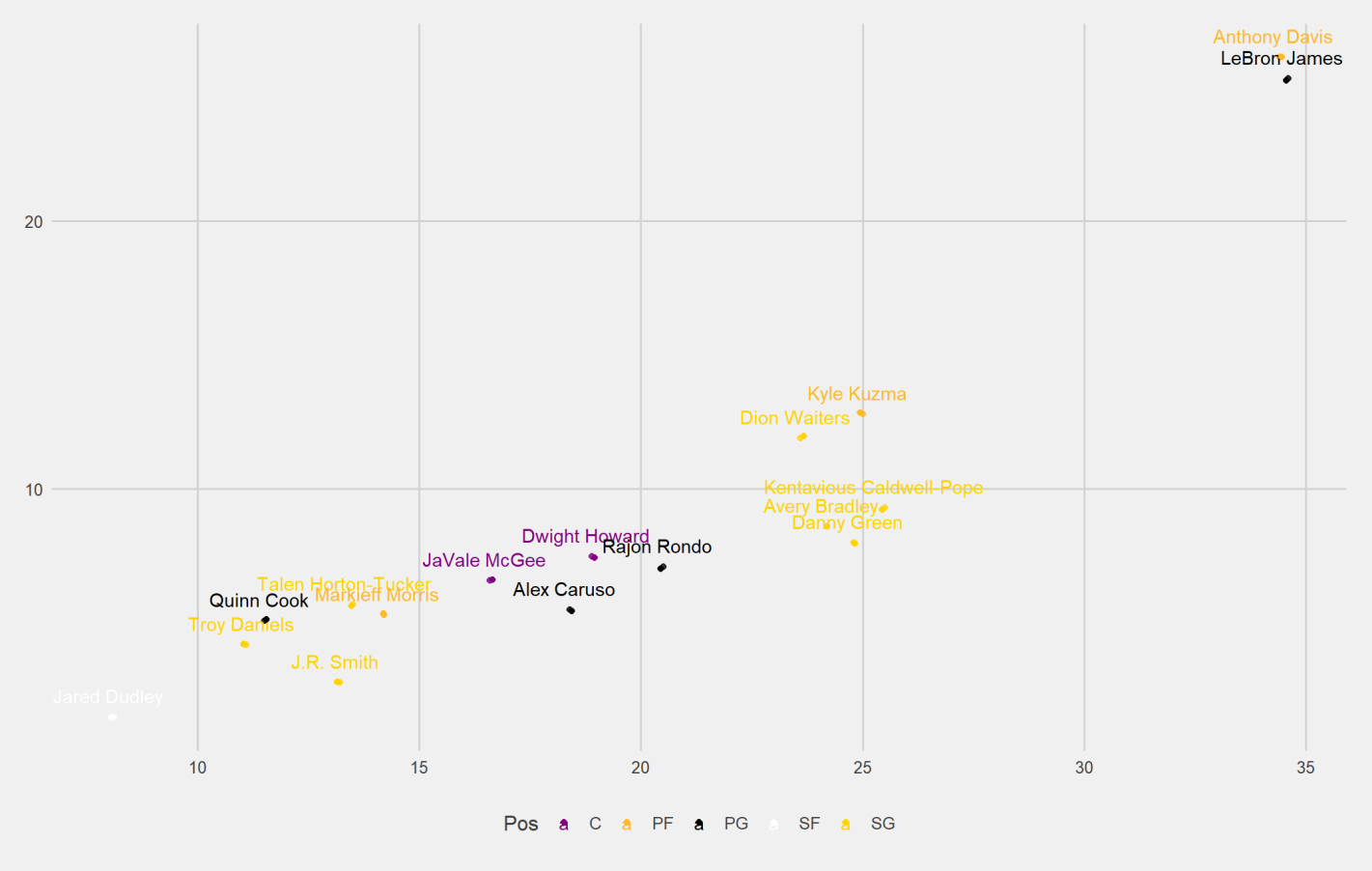



```
count(NBA2020$Pos)
```

```
##    x freq
## 1  C   91
## 2 PF  127
## 3 PG  104
## 4 SF  110
## 5 SG  159
```

Lakers 2020 Season Lets looks at just the Los Angeles Lakers. I can subset my data to just rows where team is equal to “LAL”. I set the the color of my data points to be the Lakers colors. I used the package ggthemes to have my chart look like the ones created at 538. You can use Geom_text to add labels to the data points.

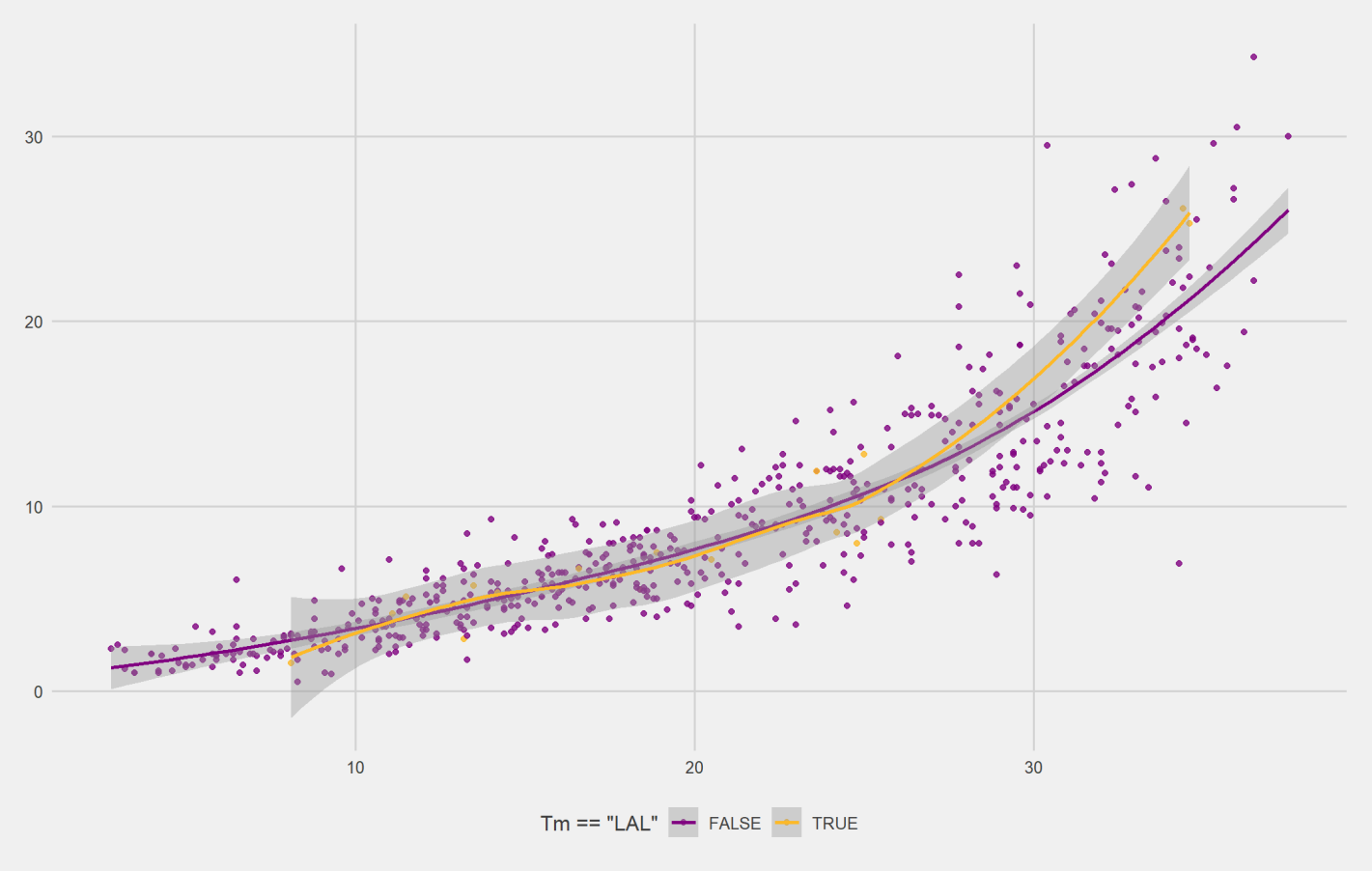
```
#install.packages("ggthemes")
lakers <- subset(NBA2020, Tm =="LAL")
lakers$Player <- sub('\\\\.*','',lakers$Player)
library(ggthemes)
ggplot(data = lakers, mapping = aes(x = MP, y=PTS, color = Pos)) +
  geom_point(alpha = .8) + scale_color_manual(values = c("#800080", "#FDB927", "#000000","#FFFFF
F","#FFD300"))  +
geom_text(aes(label=ifelse(Tm=="LAL",as.character(Player),'')),hjust=.55,vjust=-1) + geom_jitter
() + theme_fivethirtyeight()
```



geom_smooth fits a regression line through our data. Notice the larger confidence interval around the lakers line as there are way fewer data points. Lebron and AD are high leverage points.

```
ggplot(data = NBA2020, mapping = aes(x = MP, y=PTS, color = Tm == "LAL")) +
  geom_point(alpha = .8) + scale_color_manual(values = c("#800080", "#FDB927")) + theme_fivethir
tyeight() + geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Effective Field Goal Percentage

$$eFG\% = \frac{(FGM + (0.5 X 3PTM))}{FGA}$$

FGM = Field Goals Made (2PT and 3PT)

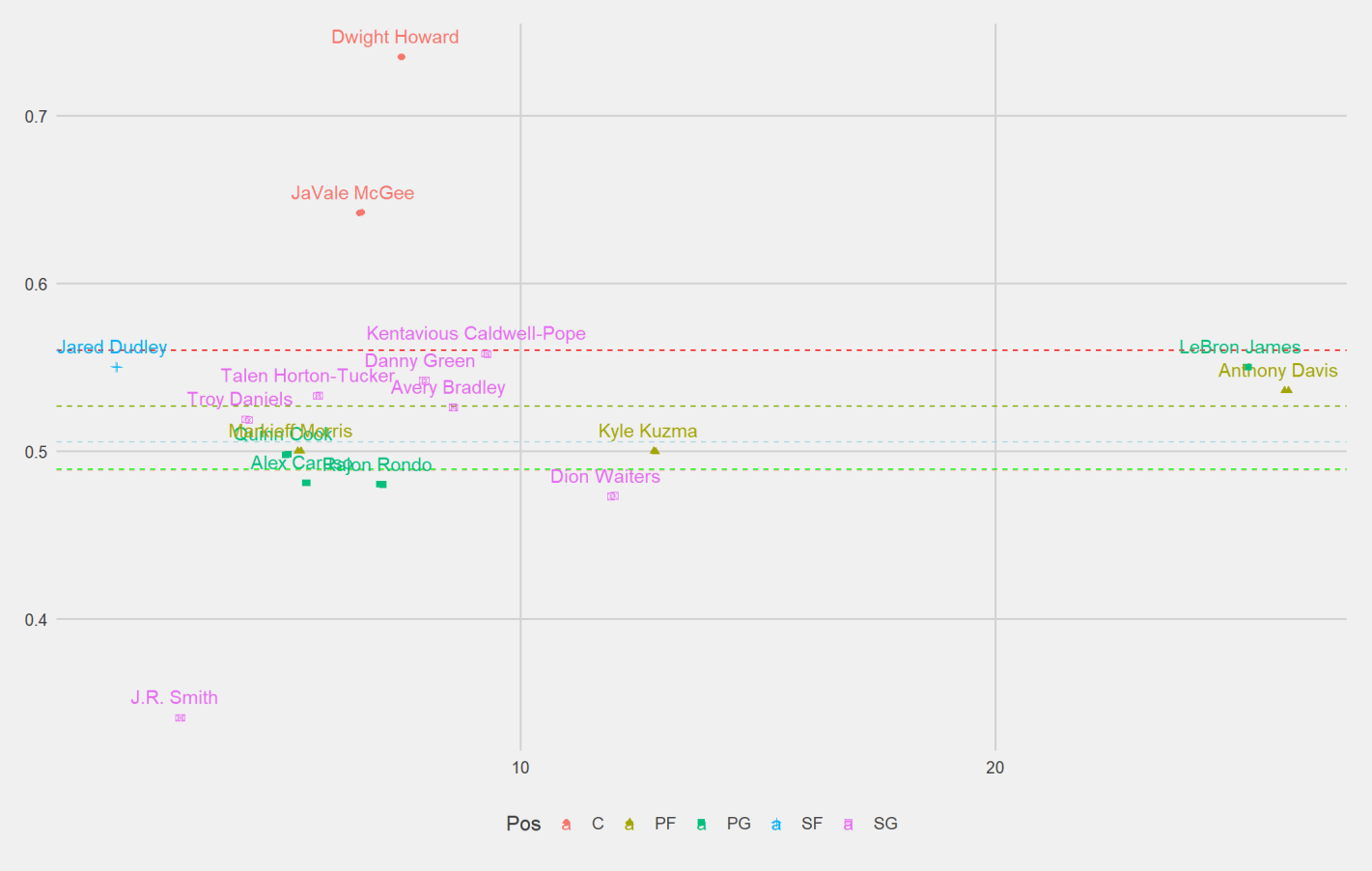
3PTM = Three Point Goals Made

FGA = Field Goals Attempted (2PT and 3PT)

How does the Lakers scoring compare to the rest of the league?

How the lakers eFG compares to the average NBA player at their position? I Use geom_hine to add a horizontal line at the mean eFG for each position.

```
ggplot(data = lakers, mapping = aes(x = PTS, y=eFG., color = Pos, shape = Pos)) +
  geom_point()+ theme_fivethirtyeight() + geom_hline(yintercept = centermean, col = "red", linet
ype = "dashed") + geom_hline(yintercept = PFmean, col = "#7CAE00", linetype = "dashed") + geom_h
line(yintercept = SFmean, col = "lightblue", linetype = "dashed") + geom_hline(yintercept = SGme
an, col = "pink", linetype = "dashed") + geom_hline(yintercept = PGmean, col = "green", linetype
= "dashed")+ geom_text(aes(label=ifelse(Tm=="LAL",as.character(Player),'')),hjust=.55,vjust=-1)
 + geom_jitter()
```



Advanced Statistics

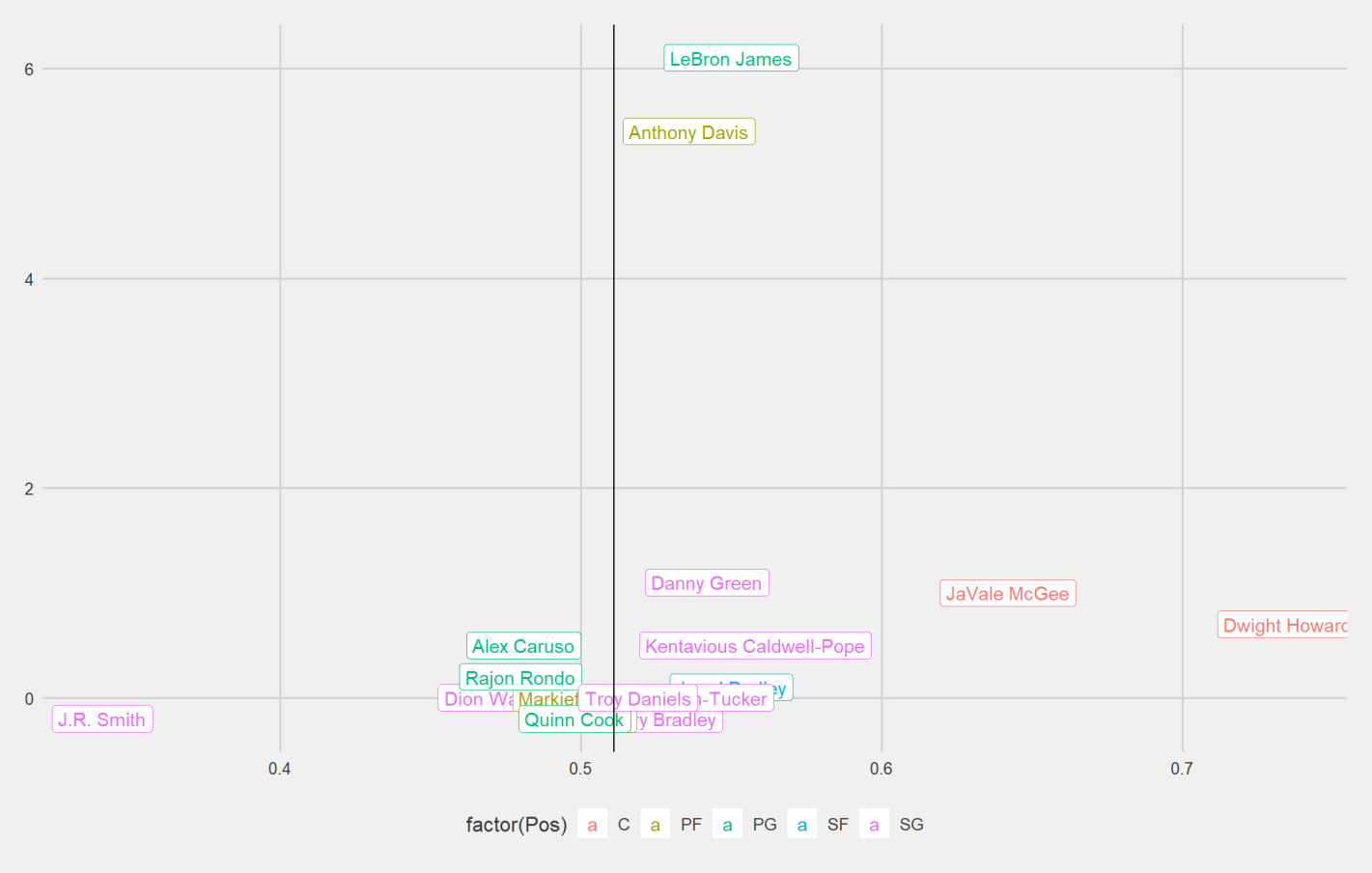
```
adv_stats <- read.csv("C:/Users/bbsta/Downloads/adv_stats.csv")
NBAefG <- subset(NBA2020, select = c(Player, eFG.))
adv_stats <- subset(adv_stats, select = -c(X, X.1))
adv_stats <- na.omit(adv_stats)
lakers_adv_stats <- subset(adv_stats, Tm == "LAL")
lakers_adv_stats$Player <- sub('\\\\.*','',lakers_adv_stats$Player)
lakersefG <- subset(lakers, select = c(Player, eFG.))
lakerstotal <- merge(lakersefG,lakers_adv_stats, by ="Player")
NBAvorp <- subset(adv_stats, select = c(Player, VORP))
total <- merge(NBA2020,NBAvorp, by ="Player")
totalvars <- subset(total, select = -c(Player, Pos, Age, Tm, G, MP, Rk, GS, FG))
```

VORP is an estimate of a players overall contribution to a team, measured vs what a theoretical “replacement player” would provide. To calculate VORP, the formula is simply: [BPM – (-2.0)] * (% of minutes played)*(team games/82), Where BPM is box plus/minus(how many points above league average per 100 possessions played.) The advantage over BPM is it accounts for minutes played!Here are the highest VORPs for the 2020 season. Lebron had a VORP of 6.1 and AD has a VORP of 5.4.

```
head(sort(adv_stats$VORP, decreasing = TRUE),n=10)
```

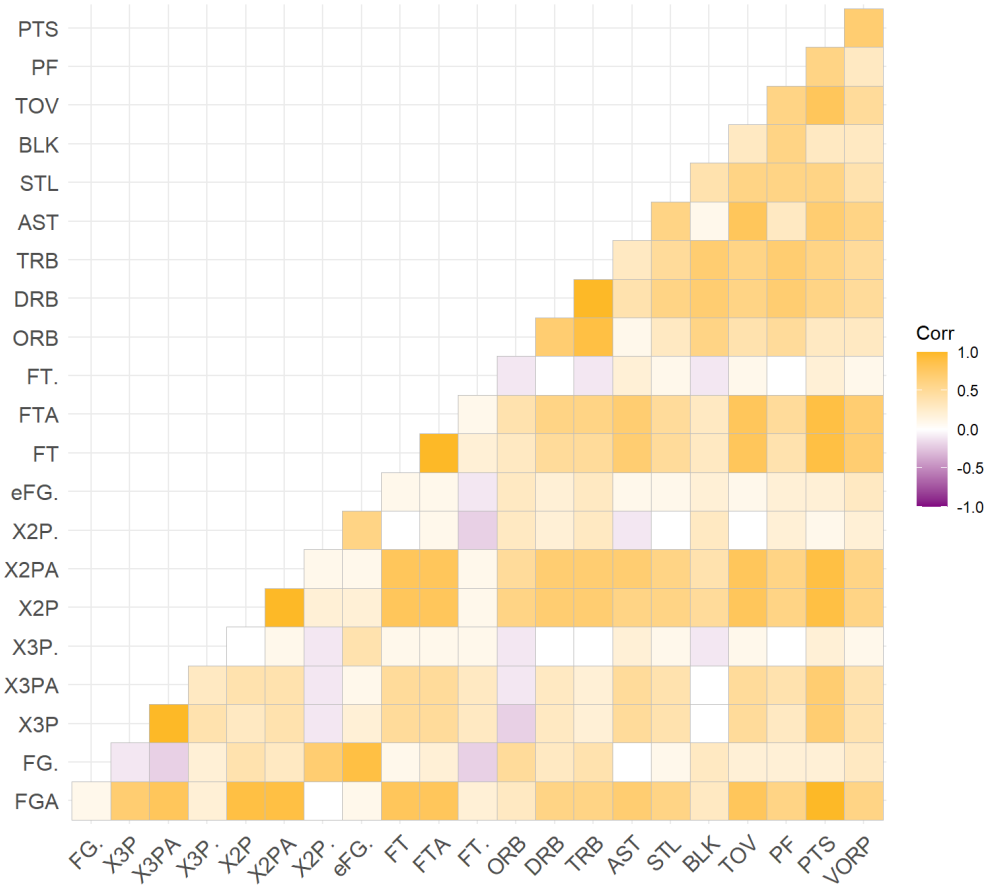
```
##  [1] 7.3 6.6 6.1 5.9 5.5 5.4 5.4 5.1 3.7 3.5
```

```
ggplot(data = lakerstotal, aes(eFG., VORP )) +
  geom_label(aes(label = Player, colour = factor(Pos))) + geom_vline(xintercept =mean(NBA2020$eF
G.)) + theme_fivethirtyeight()
```



Is EFG% the most strongly correlated variable with VORP?

```
#install.packages("ggcorrplot")
library(ggcorrplot)
cormatrix <- round(cor(totalvars),1)
ggcorrplot(cormatrix, method = "square", type = "lower", colors = c("#800080", "white", "#FDB92
7"), insig = "blank")
```



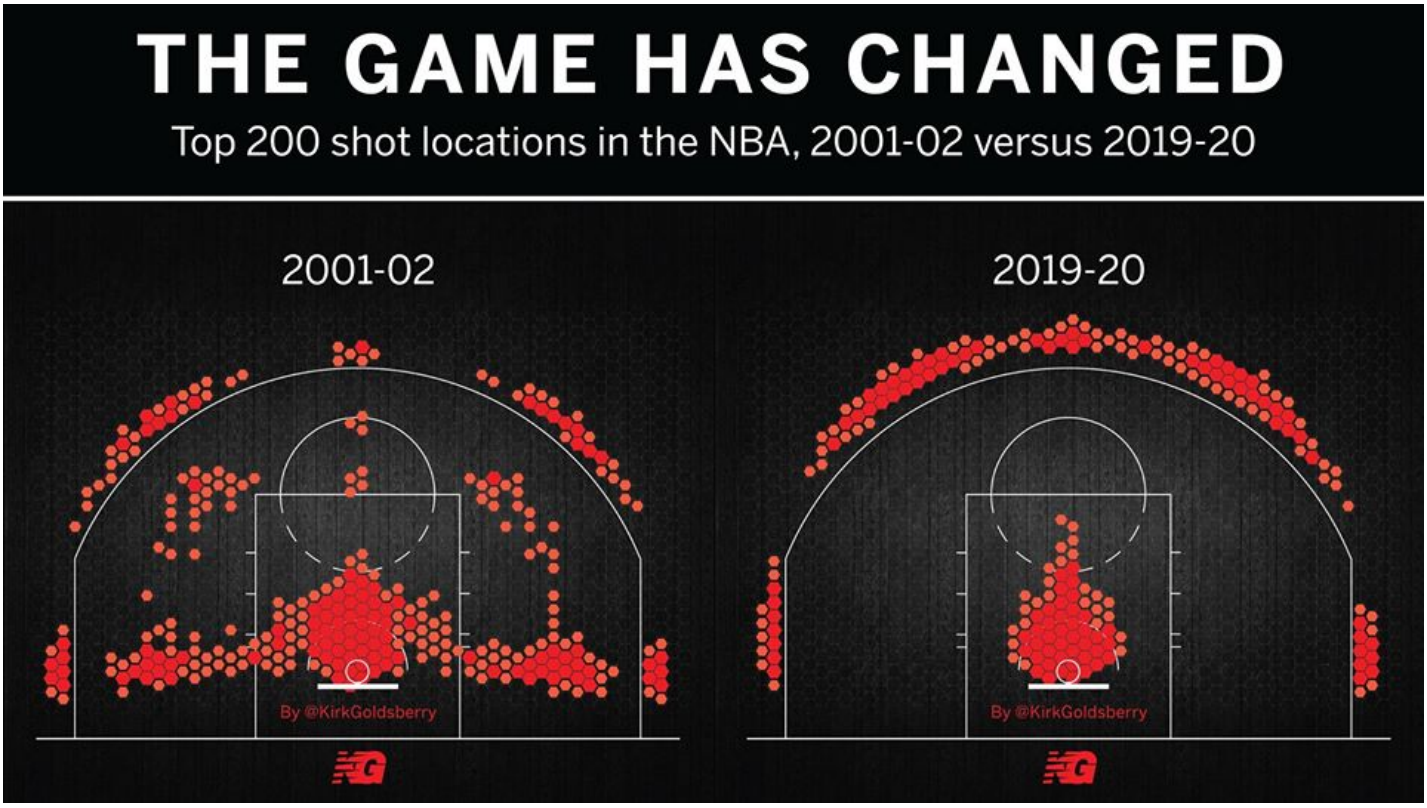
Free Throw attempts and Free throw makes are most strongly correlated with VORP.

```
cor(totalvars$VORP,totalvars)
```

```
##            FGA       FG.       X3P      X3PA      X3P.       X2P      X2PA
## [1,] 0.5790805 0.2656792 0.3813684 0.3687593 0.1131504 0.6155526 0.5745877
##           X2P.      eFG.        FT       FTA       FT.       ORB       DRB
## [1,] 0.1764015 0.2598229 0.6678809 0.6614907 0.1050329 0.3213332 0.536232
##            TRB       AST       STL       BLK       TOV        PF       PTS VORP
## [1,] 0.5048543 0.5599281 0.4456159 0.3237726 0.5459768 0.2855095 0.6562415    1
```

Why is cool ggplot is easy to use and fairly intuitive. ggplot allows you to a better understanding of the game of basketball and makes it easy for non fans to understand it at a deeper level. You can easily compare players, teams, organizations with a data based approach.

Drawbacks/Limitations There are base r packages that can create many of the plots that we created today.If you need to create a quick exploratory graphic that just you are looking at, its easier to just use the base package. If you do not want to download more packages or learn the syntax of ggplot it may not be worth your time.
One popular graphic is the shot chart(seen below). This can be easier to create in Tableu(maybe Python).Although it is possible to do in R. Many heatmap based plots are easier to create in base R with heatmap().


THE GAME HAS CHANGED
Top 200 shot locations in the NBA, 2001-02 versus 2019-20
2001-02
2019-20
By @KirkGoldsberry
By @KirkGoldsberry

Shot Chart