Final Report

Kecheng Liang

12/12/2018

Introduction

Basketball is one of the most popular sports in the world. National Basketball Association(NBA) is the largest league for this sport. There are lots of interesting data in the game. In basketball, an assist is attributed to a player who passes the ball to a teammate in a way that leads to a score by field goal, meaning that he or she was "assisting" in the basket. Because an assist can be scored for the passer even if the player who receives the pass makes a basket after dribbling the ball. In some situtation it becomes hard to define whether it is a assist. We may think that player who play in the home game are more easily get the tenth assist when the player already have nine assists. Same thing may happen in rebound. I want to do the analysis whether it really happens. Also I will do other interesting graph to show the miracle NBA data.

Data

The data is downloaded from website and the data from 2012 to 2018. It is well organized with 51 variables and I removed some useless variables.

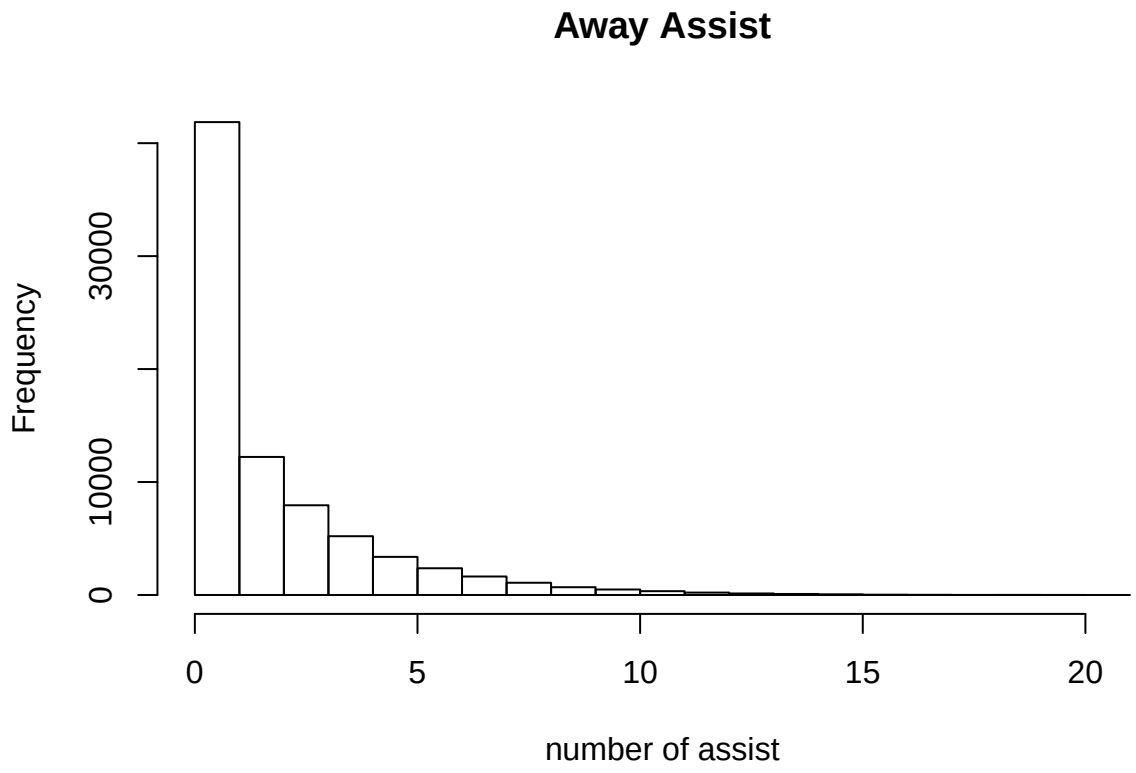
Table 1: Variables explanation

Variables	Explanation
teamAbbr	Abbreviation of team
teamConf	Identifies conference of team
teamLoc	Identifies whether team is home or visitor
teamRslt	Identifies whether team has won or lost
playDispNm	Player display name
playStat	Identifies staring status of player
playMin	Player minutes on floor
playPos	Player position during game
playHeight	Player height
playWeight	Player weight
playPTS	Points scored by player
playAST	Assists made by player
playTO	Turnovers made by player
playSTL	Steals made by player
playBLK	Blocks made by player
playPF	Personal fouls made by player
playFGA	Field goal attempts made by player
playFGM	Field goal shots made by player
playFG.	Field goal percentage made by player
play2PA	Two point attempts made by player
play2PM	Two point shots made by player
play2P.	Two point percentage made by player
play3PA	Three point attempts made by player
play3PM	Three point shots made by player
play3P.	Three point percentage made by player

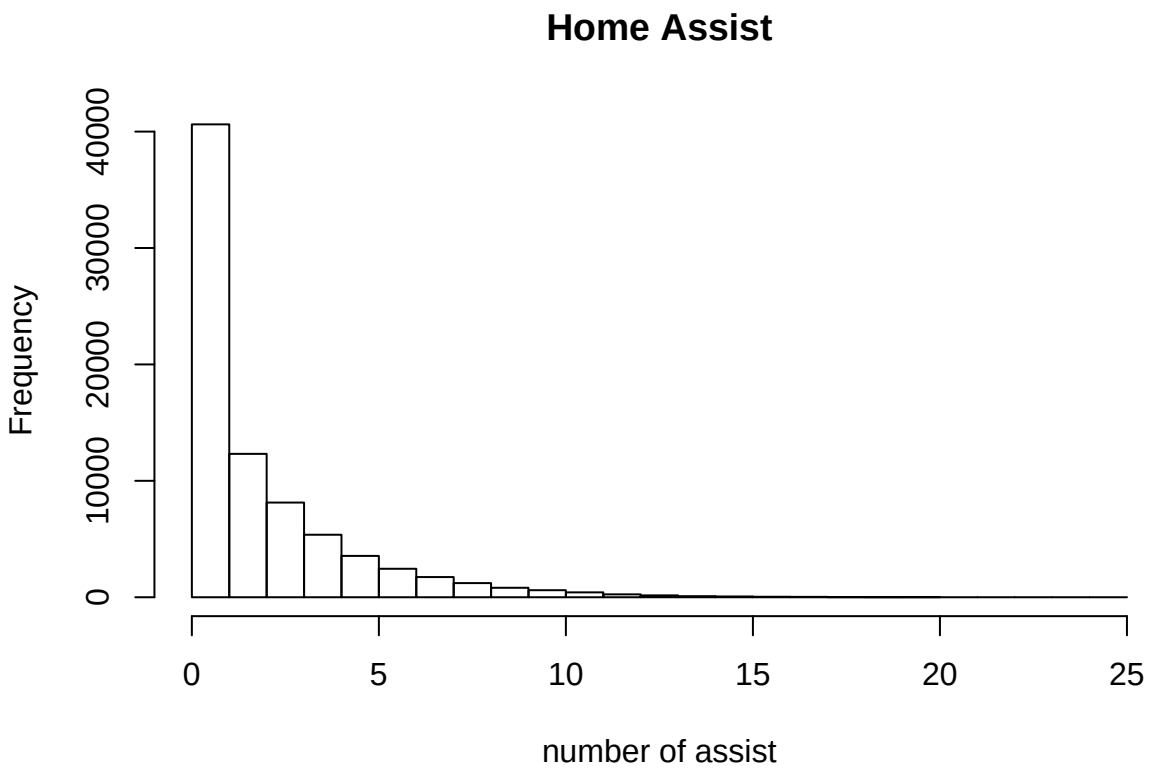
Variables	Explanation
playFTA	Free throw attempts made by player
playFTM	Free throw shots made by player
playFT.	Free throw percentage made by player
playORB	Offensive rebounds made by player
playDRB	Defensive rebounds made by player
playTRB	Total rebounds made by player
opptAbbr	AAbbreviAtion of opponent
opptConf	Identifies conference of opponent

EDA for Assists and Rebounds

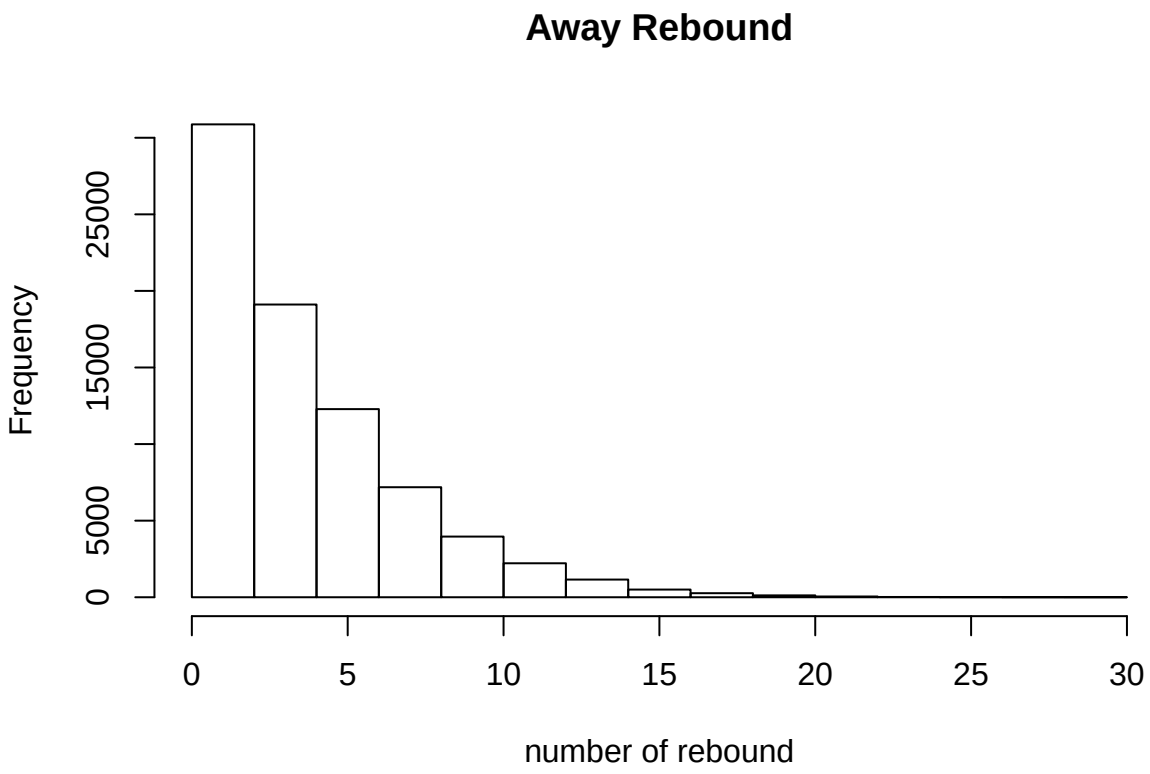
```
#Assist and team location
hist(x=away_data$playAST,main = "Away Assist",xlab = "number of assist")
```



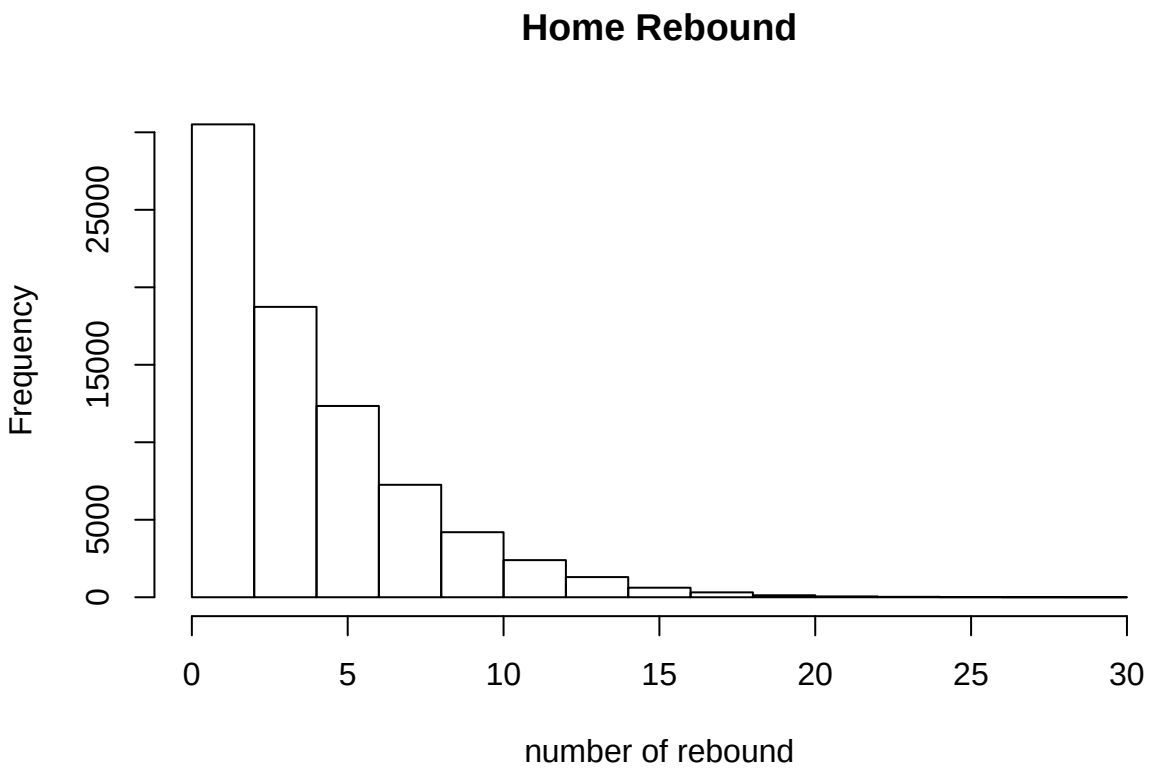

```
hist(x=home_data$playAST,main = "Home Assist",xlab = "number of assist")
```




```
#Rebound and team location
hist(x=away_data$playTRB,main = "Away Rebound",xlab = "number of rebound")
```




```
hist(x=home_data$playTRB,main = "Home Rebound",xlab = "number of rebound")
```



Chi square test

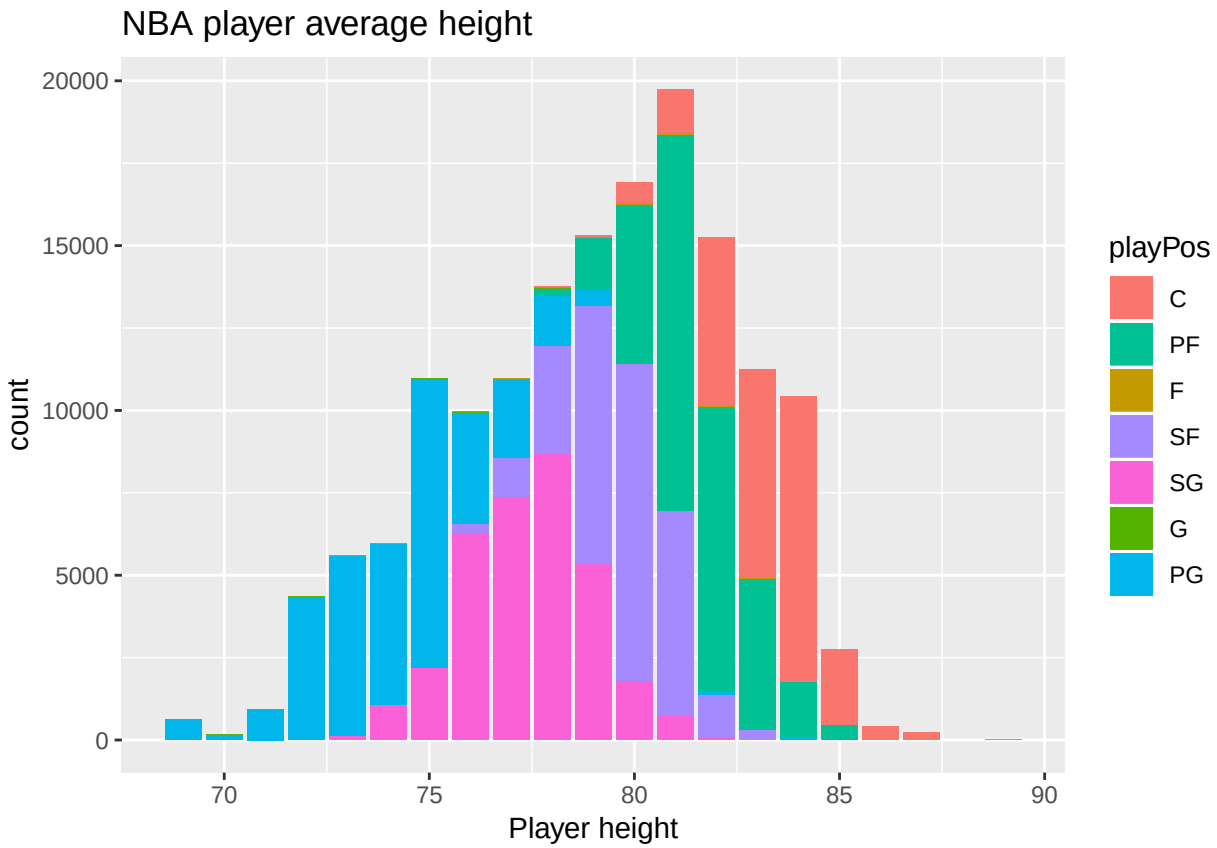
```
##
##  Pearson's Chi-squared test
##
## data:  total_ast
## X-squared = 95.549, df = 23, p-value = 8.177e-11
```

```
##
##  Pearson's Chi-squared test
##
## data:  total_trb
## X-squared = 64.517, df = 29, p-value = 0.000164
```

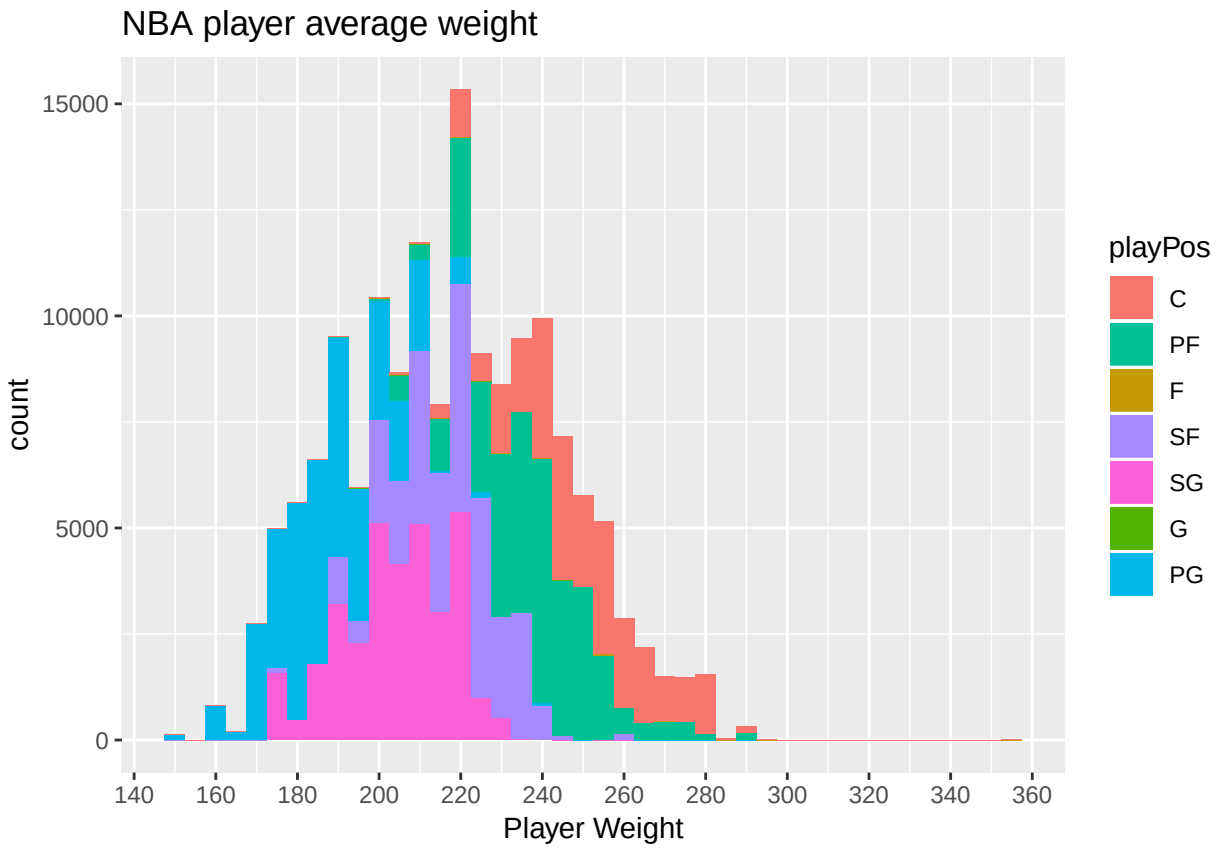
By doing the chi square test, we found that the both assist and rebound do not have significance difference between home game and away game.

EDA for Interesting NBA Stats

```
#NBA player height and weight related to position
ggplot(data = nba_total,aes(x=playHeight))+geom_bar(aes(fill=playPos))+
  ggtitle("NBA player average height")+xlab("Player height")+
  scale_fill_discrete(breaks=c("C","PF","F","SF","SG","G","PG"))
```

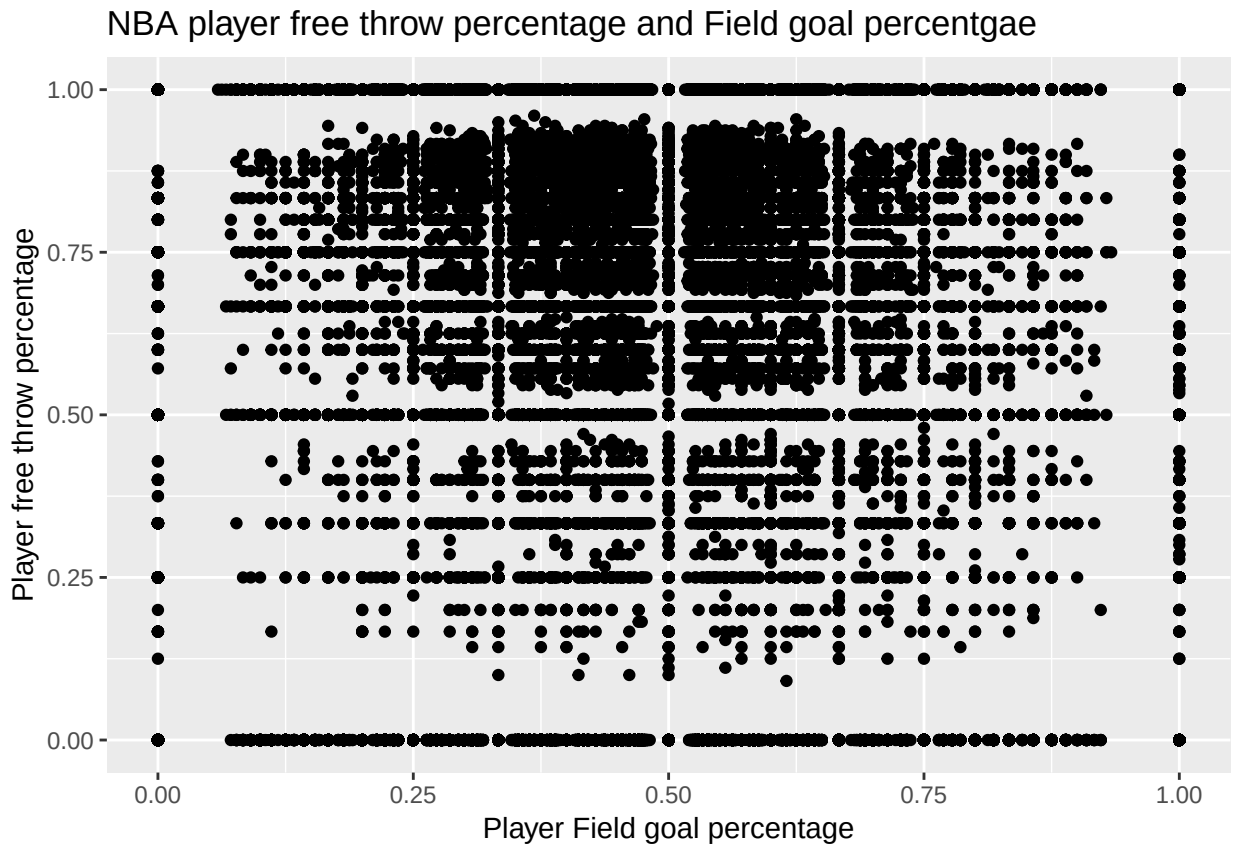



```
ggplot(data = nba_total,aes(x=playWeight,fill=playPos))+geom_histogram(binwidth = 5)+
  ggtitle("NBA player average weight")+
  scale_x_continuous(name='Player Weight',breaks=seq(140,360,20))+
  scale_fill_discrete(breaks=c("C","PF","F","SF","SG","G","PG"))
```

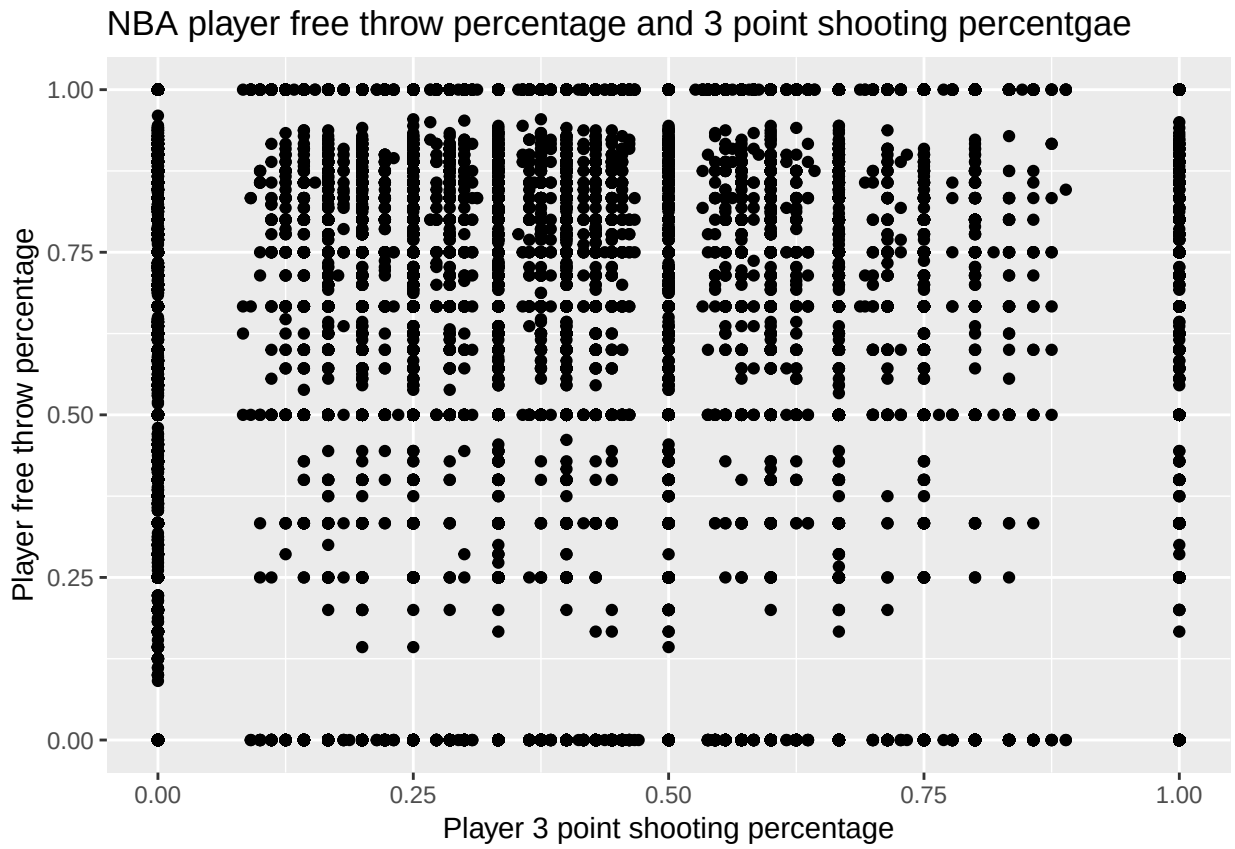


Those above graphs show the distribution of NBA players height and weight.

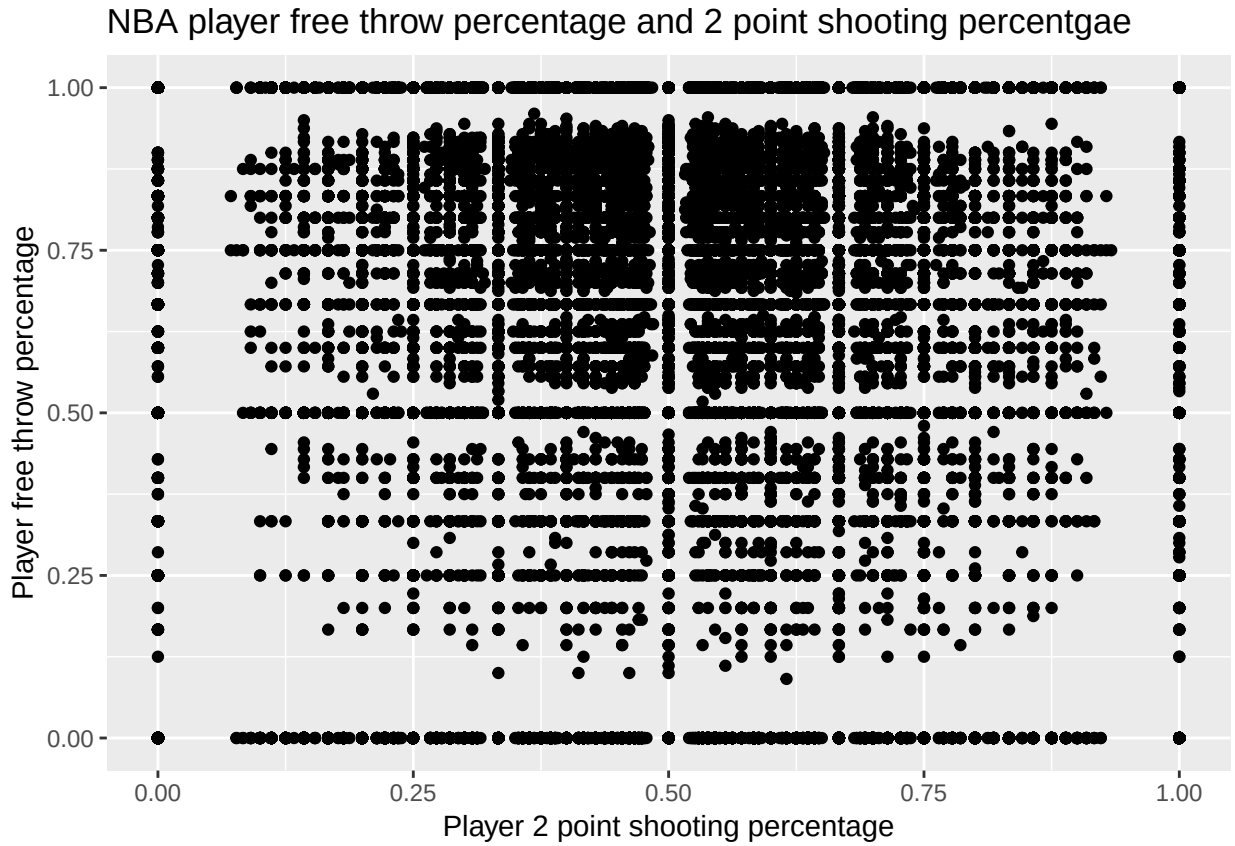
```
#NBA player relationship between free throw and shooting percentage
ggplot(data = nba_total,aes(x=playFG.,y=playFT.))+geom_point()+
  ggtitle("NBA player free throw percentage and Field goal percentgae")+
  xlab("Player Field goal percentage")+ylab("Player free throw percentage")
```




```
ggplot(data = nba_total,aes(x=play3P.,y=playFT.))+geom_point()+
  ggtitle("NBA player free throw percentage and 3 point shooting percentgae")+
  xlab("Player 3 point shooting percentage")+ylab("Player free throw percentage")
```

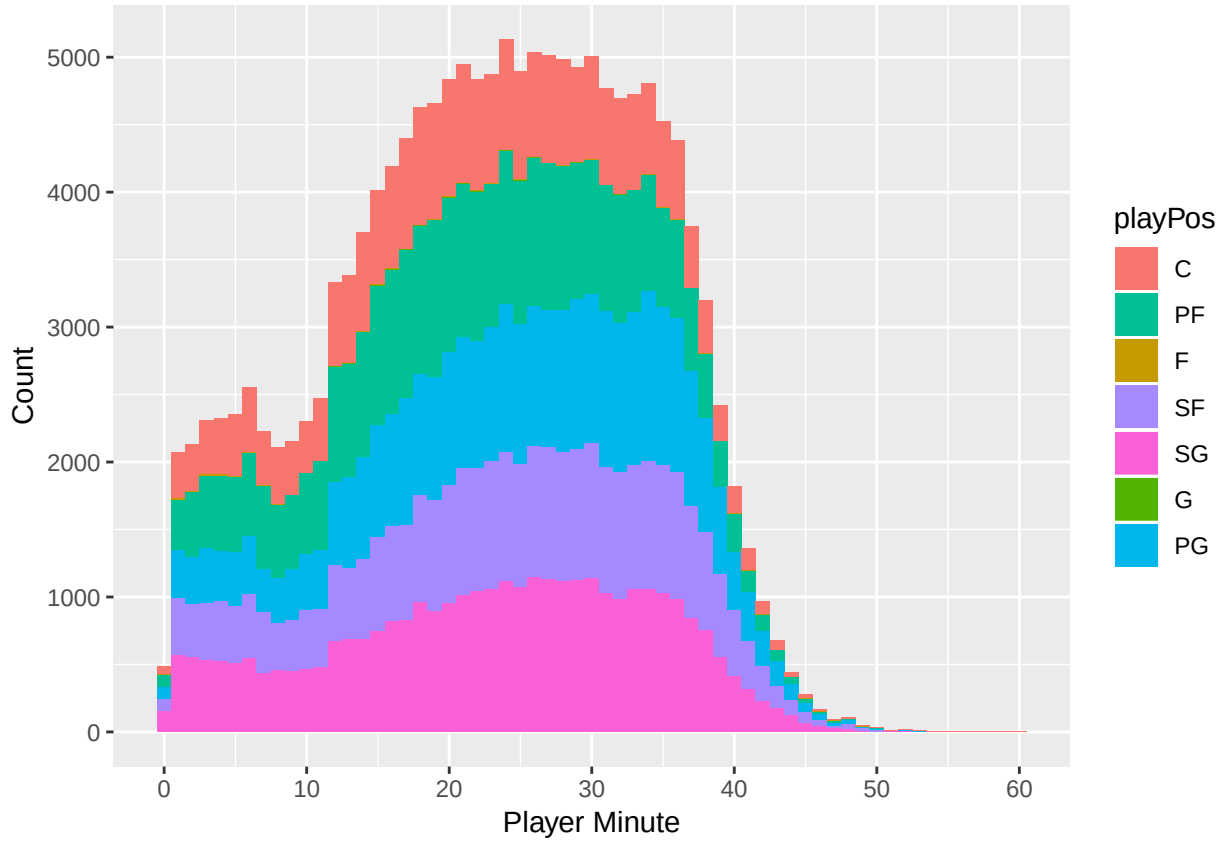



```
ggplot(data = nba_total,aes(x=play2P.,y=playFT.))+geom_point()+
  ggtitle("NBA player free throw percentage and 2 point shooting percentgae")+
  xlab("Player 2 point shooting percentage")+ylab("Player free throw percentage")
```

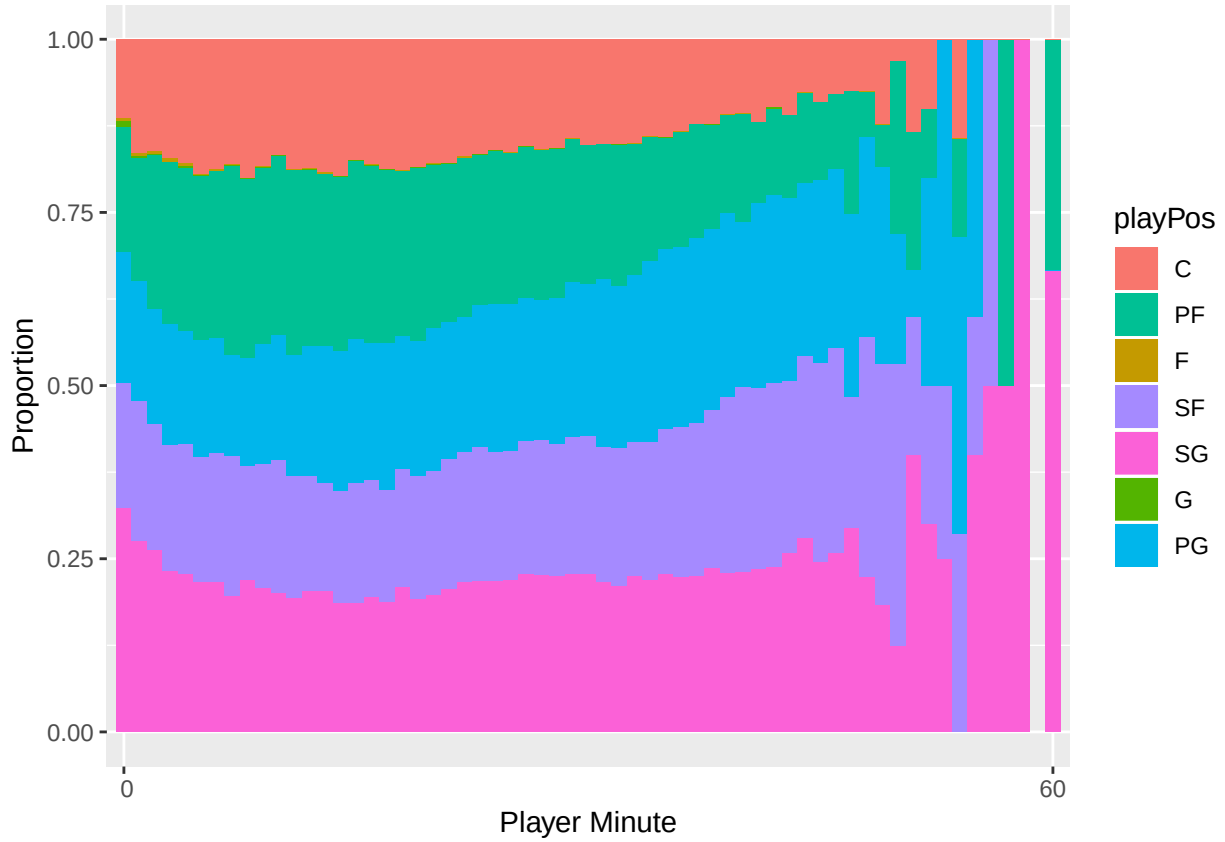


Those above graphs show the relationship between free throw and shooting percentage.

```
#Relationship between position and play minutes
ggplot(data = nba_total,aes(x=playMin,fill=playPos))+geom_histogram(binwidth = 1)+
  scale_x_continuous(name="Player Minute",breaks=seq(0,60,10))+
  scale_y_continuous(name="Count",breaks=seq(0,6000,1000))+
  scale_fill_discrete(breaks=c("C","PF","F","SF","SG","G","PG"))
```

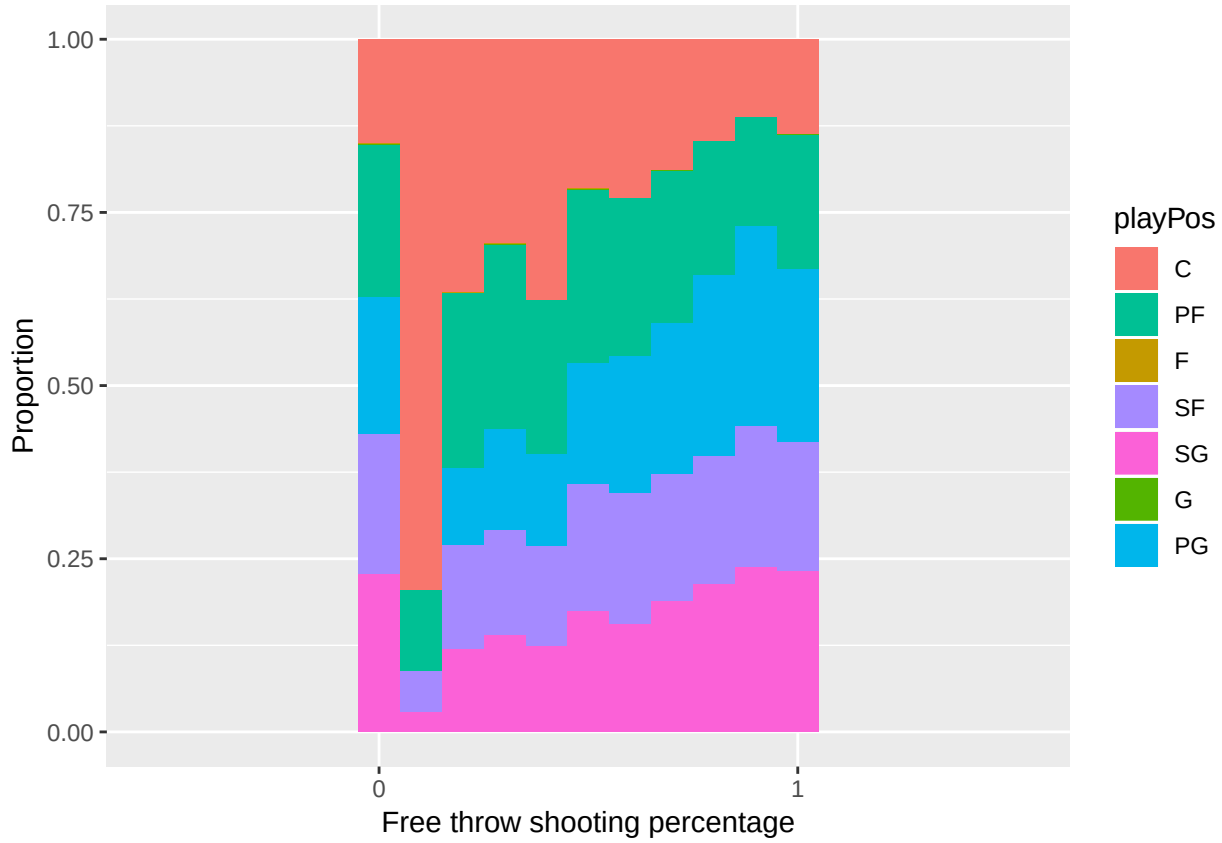



```
ggplot(data = nba_total,aes(x=playMin,fill=playPos))+geom_histogram(position = "fill",binwidth = 1)+
  scale_x_discrete(name='Player Minute',breaks=seq(0,60,10),limits=c(0,60))+
  ylab("Proportion")+scale_fill_discrete(breaks=c("C","PF","F","SF","SG","G","PG"))
```

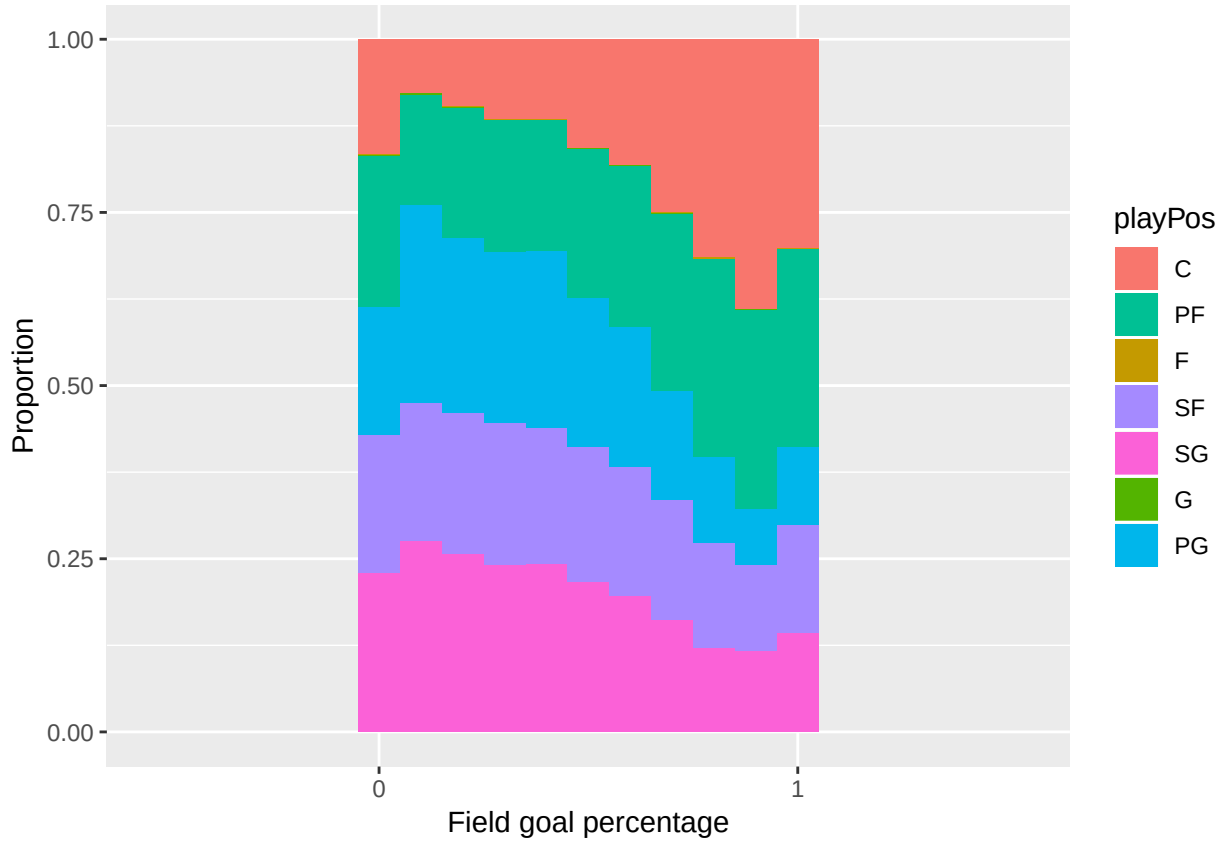


Those graphs show the relationship between player minutes and the position of player.

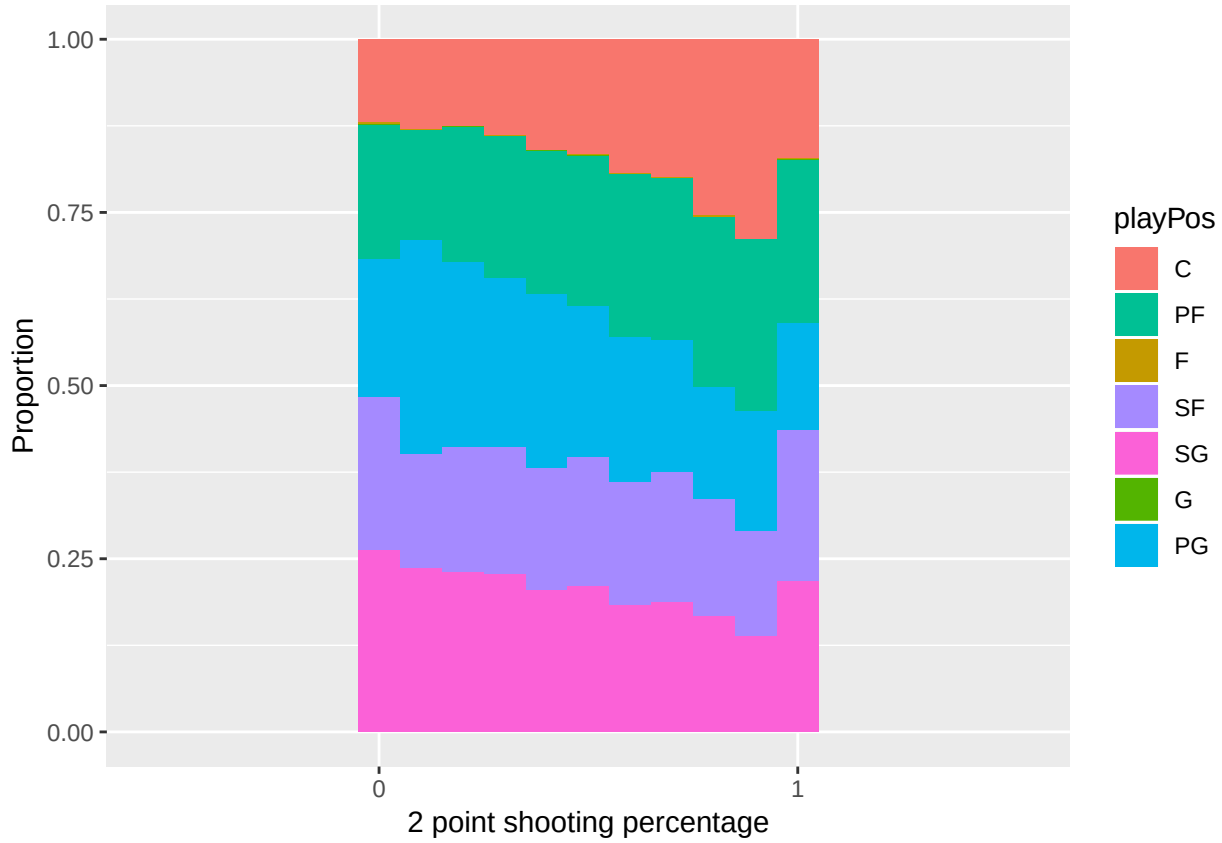
```
#Relationship between position and shooting percentage
ggplot(data = nba_total,aes(x=playFT.,fill=playPos))+geom_histogram(position = "fill",binwidth = 0.1)+
  scale_x_discrete(name="Free throw shooting percentage",breaks=seq(0,1,0.1),limits=c(0,1))+
  ylab("Proportion")+scale_fill_discrete(breaks=c("C","PF","F","SF","SG","G","PG"))
```



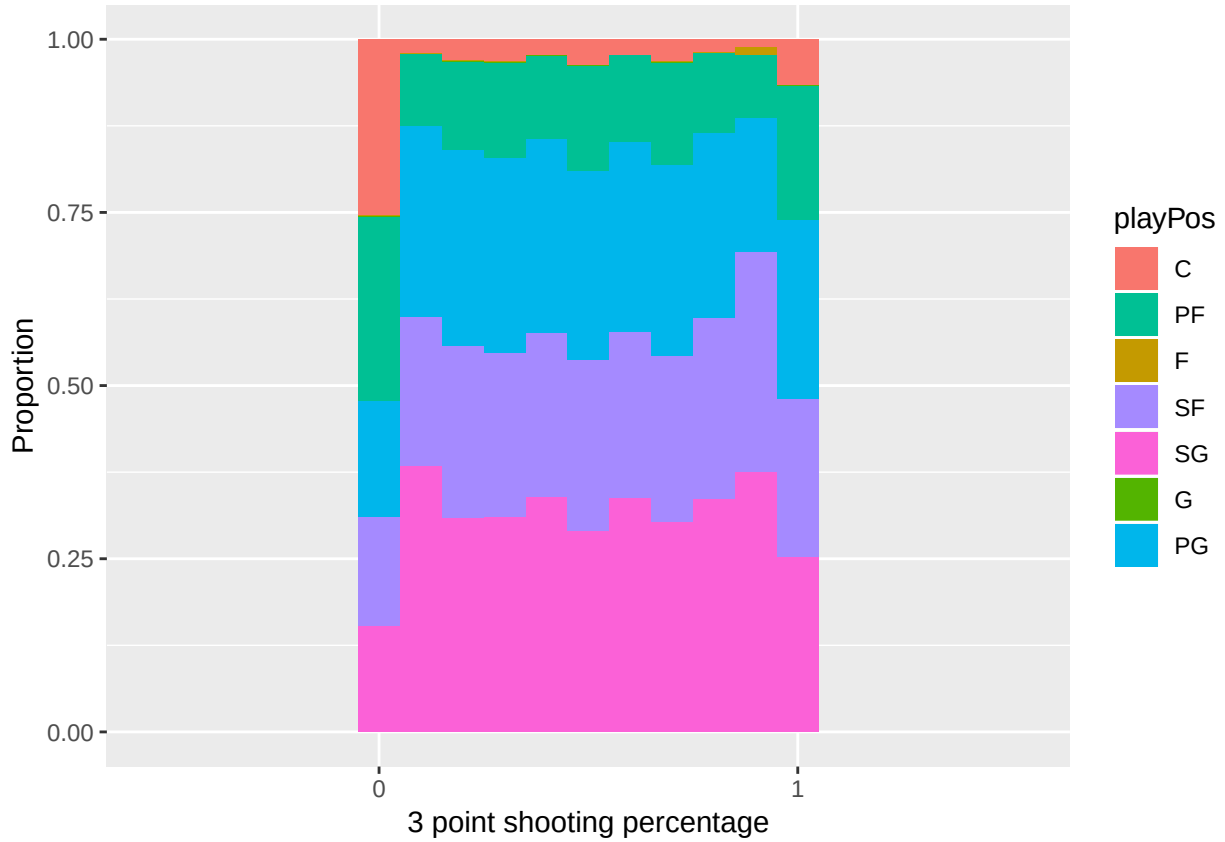

```
ggplot(data = nba_total,aes(x=playFG.,fill=playPos))+geom_histogram(position = "fill",binwidth = 0.1)+
  scale_x_discrete(name="Field goal percentage",breaks=seq(0,1,0.1),limits=c(0,1))+
  ylab("Proportion")+scale_fill_discrete(breaks=c("C","PF","F","SF","SG","G","PG"))
```




```
ggplot(data = nba_total,aes(x=play2P.,fill=playPos))+geom_histogram(position = "fill",binwidth = 0.1)+
  scale_x_discrete(name="2 point shooting percentage",breaks=seq(0,1,0.1),limits=c(0,1))+
  ylab("Proportion")+scale_fill_discrete(breaks=c("C","PF","F","SF","SG","G","PG"))
```

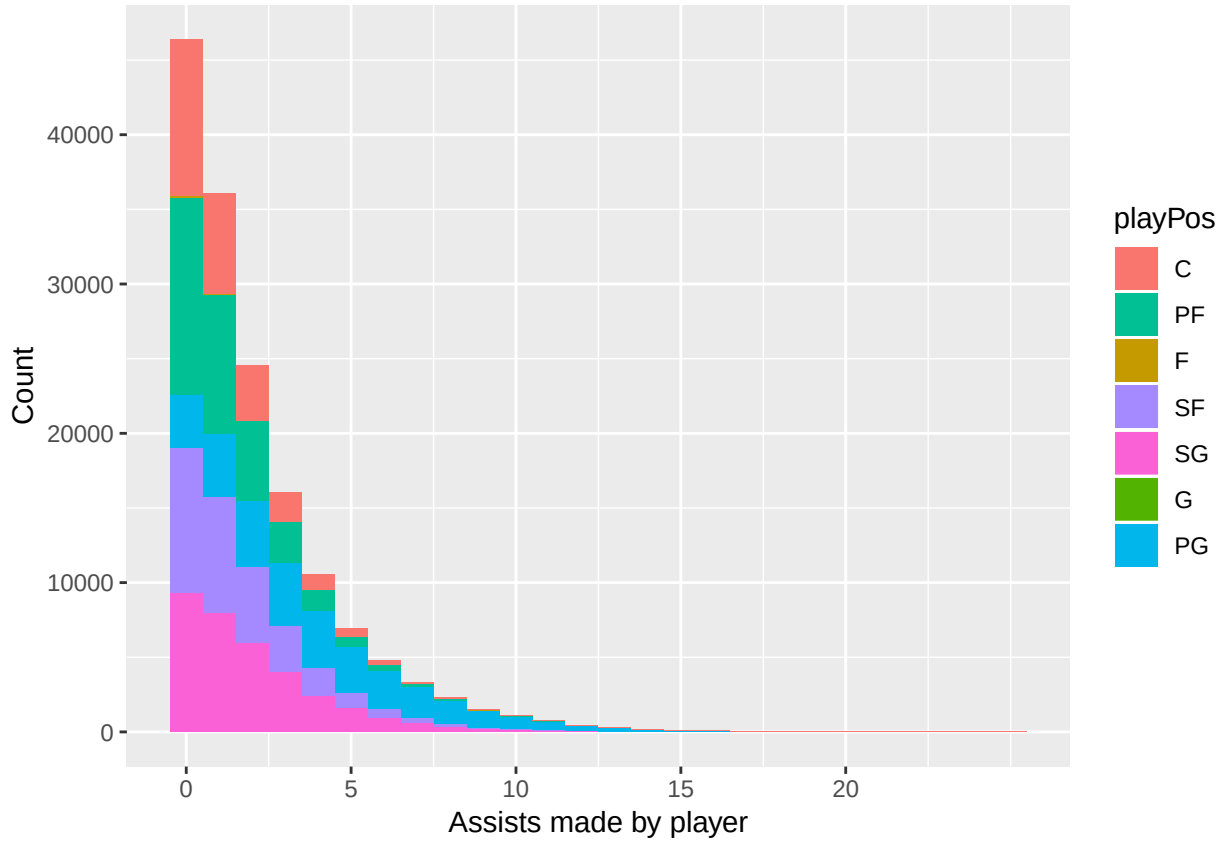



```
ggplot(data = nba_total,aes(x=play3P.,fill=playPos))+geom_histogram(position = "fill",binwidth = 0.1)+
  scale_x_discrete(name="3 point shooting percentage",breaks=seq(0,1,0.1),limits=c(0,1))+
  ylab("Proportion")+scale_fill_discrete(breaks=c("C","PF","F","SF","SG","G","PG"))
```

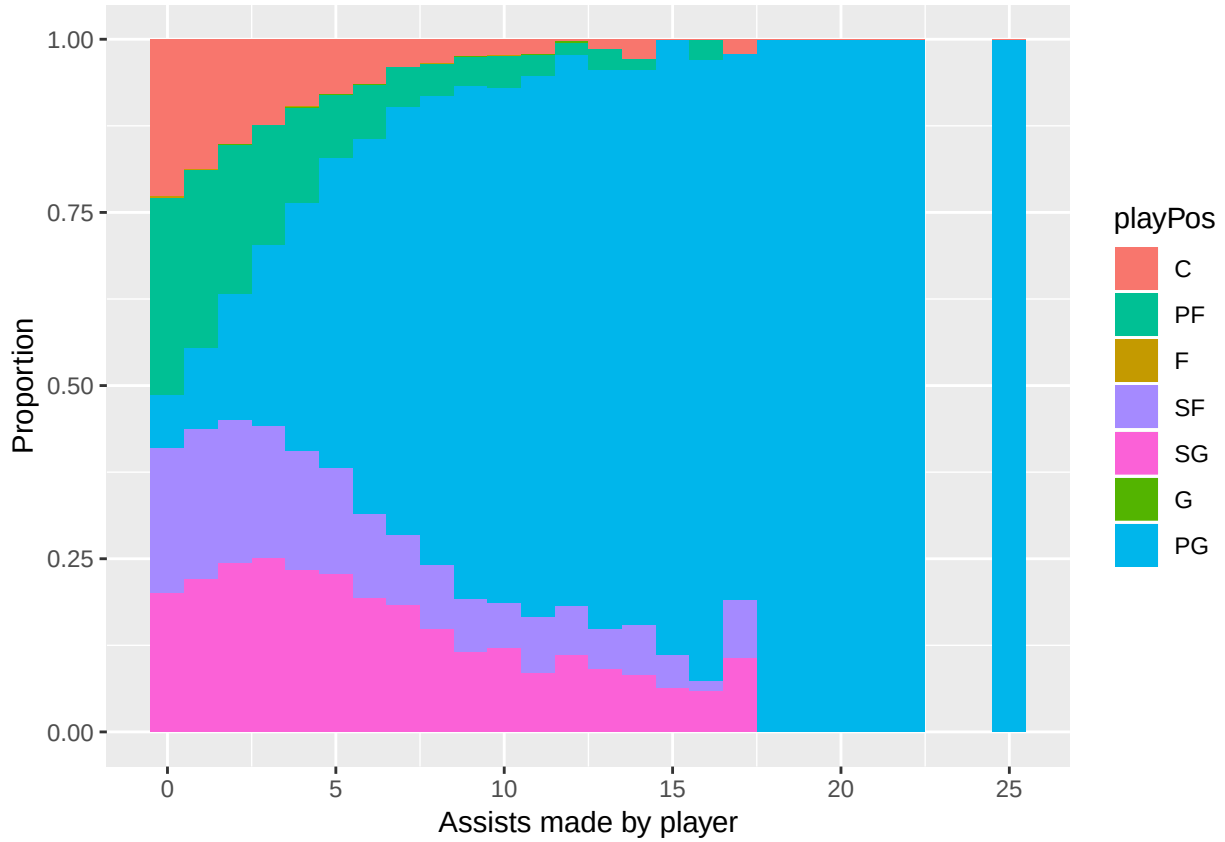


Those above graphs show the relationship between position and shooting percentage.

```
#Relationship between position and basic stats
ggplot(data = nba_total,aes(x=playAST,fill=playPos))+geom_histogram(binwidth = 1)+
  scale_x_continuous(name="Assists made by player",breaks=seq(0,20,5))+
  ylab("Count")+scale_fill_discrete(breaks=c("C","PF","F","SF","SG","G","PG"))
```

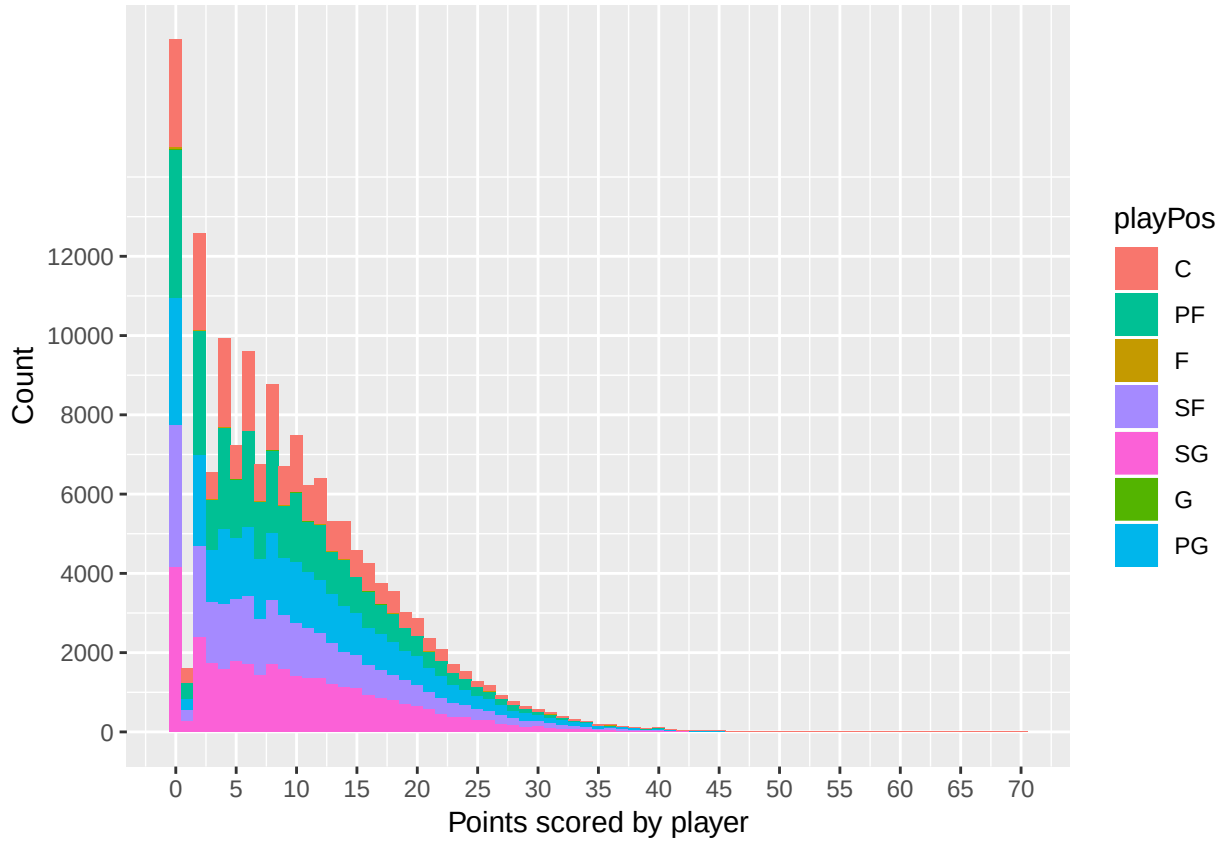



```
ggplot(data = nba_total,aes(x=playAST,fill=playPos))+geom_histogram(position="fill",binwidth = 1)+
  scale_x_continuous(name="Assists made by player",breaks=seq(0,25,5))+
  ylab("Proportion")+scale_fill_discrete(breaks=c("C","PF","F","SF","SG","G","PG"))
```

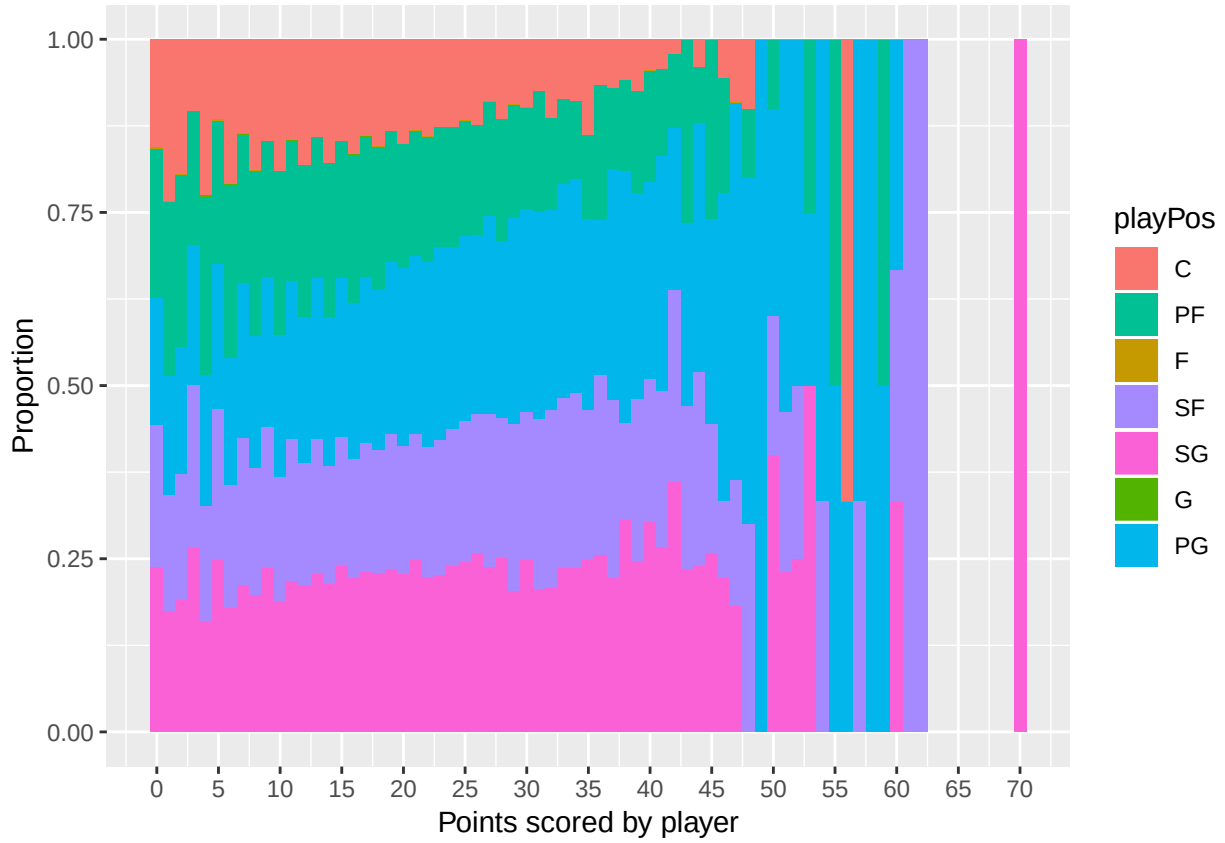


Those two graphs show that PG are much easier to get high number of assists.

```
ggplot(data = nba_total,aes(x=playPTS,fill=playPos))+geom_histogram(binwidth = 1)+
  scale_x_continuous(name="Points scored by player",breaks=seq(0,70,5))+
  scale_y_continuous(name="Count",breaks=seq(0,13000,2000))+
  scale_fill_discrete(breaks=c("C","PF","F","SF","SG","G","PG"))
```

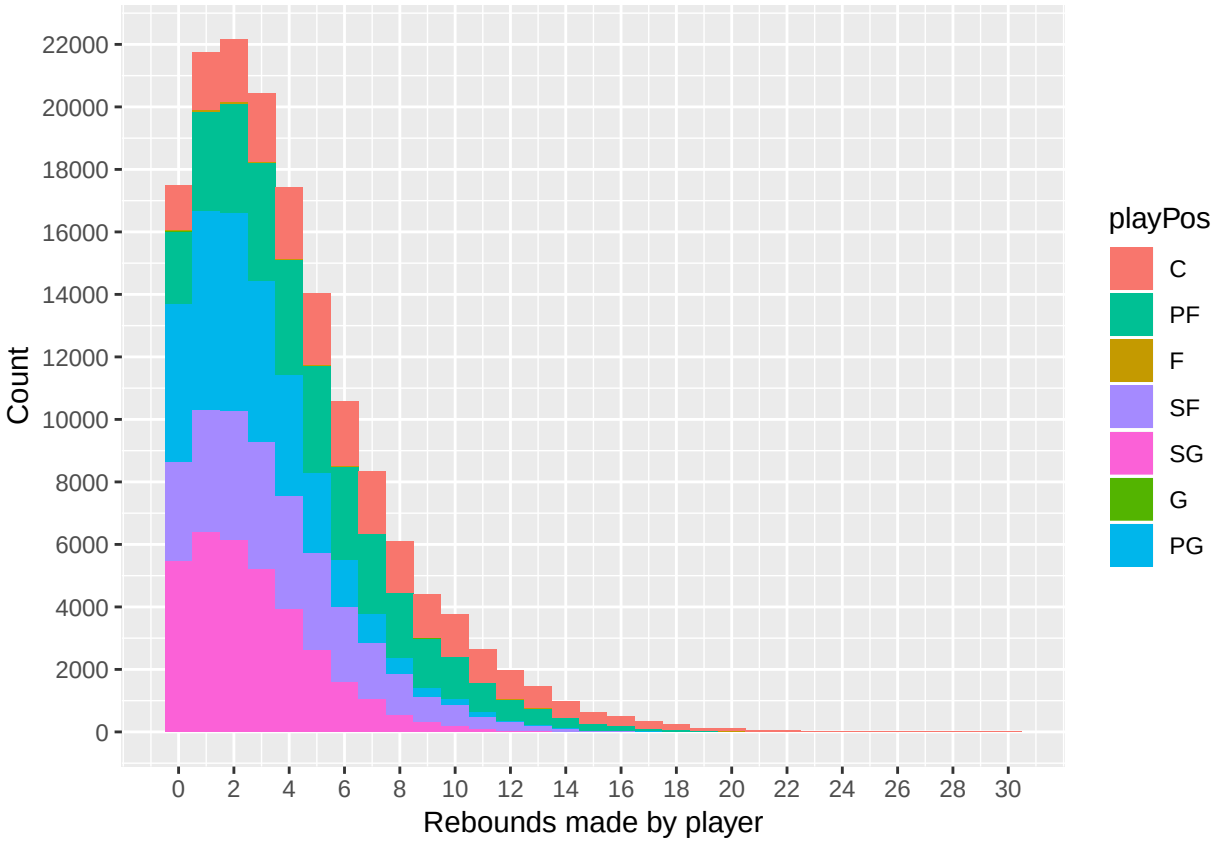



```
ggplot(data = nba_total,aes(x=playPTS,fill=playPos))+geom_histogram(position="fill",binwidth = 1)+
  scale_x_continuous(name="Points scored by player",breaks=seq(0,70,5))+
  ylab("Proportion")+scale_fill_discrete(breaks=c("C","PF","F","SF","SG","G","PG"))
```

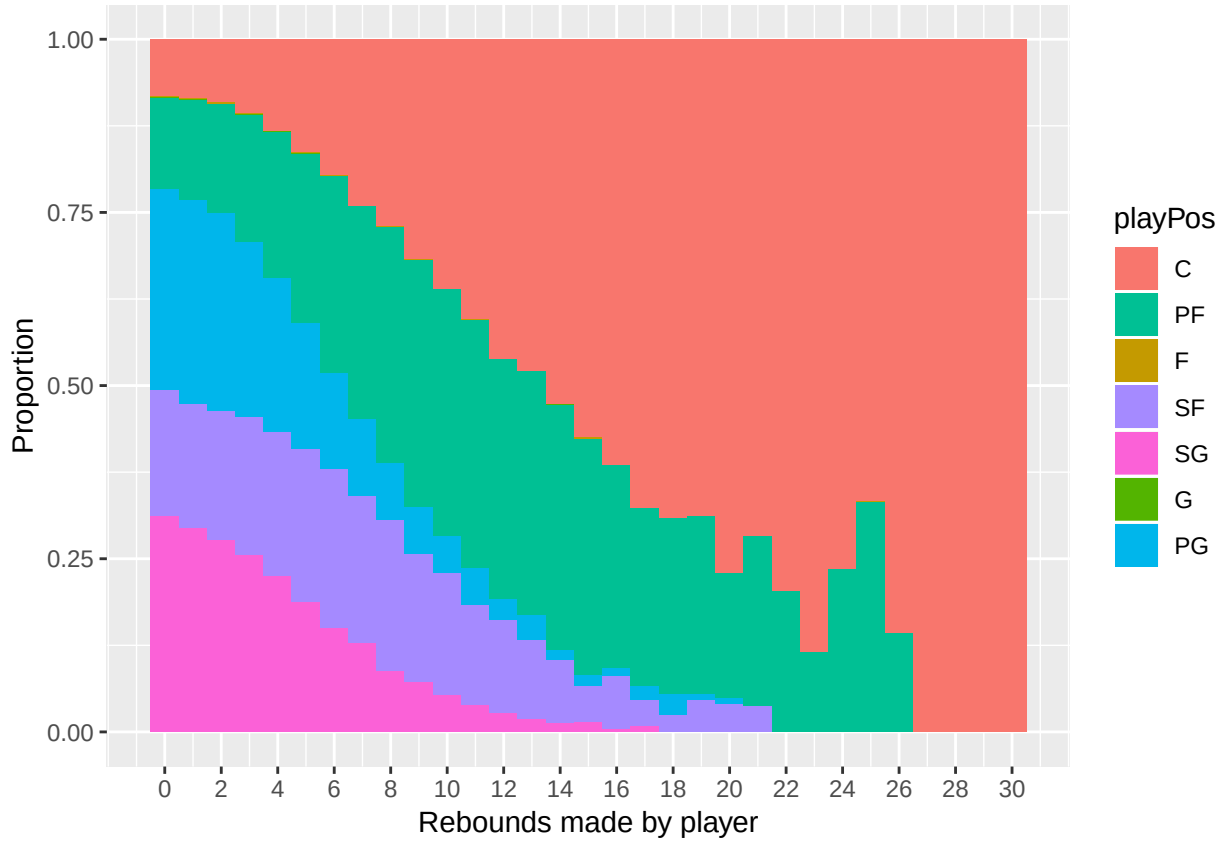


Those two above graphs show that SG, PG, and SF are much easier to score more points.

```
ggplot(data = nba_total,aes(x=playTRB,fill=playPos))+geom_histogram(binwidth = 1)+
  scale_x_continuous(name="Rebounds made by player",breaks=seq(0,30,2))+
  scale_y_continuous(name="Count",breaks=seq(0,22000,2000))+
  scale_fill_discrete(breaks=c("C","PF","F","SF","SG","G","PG"))
```

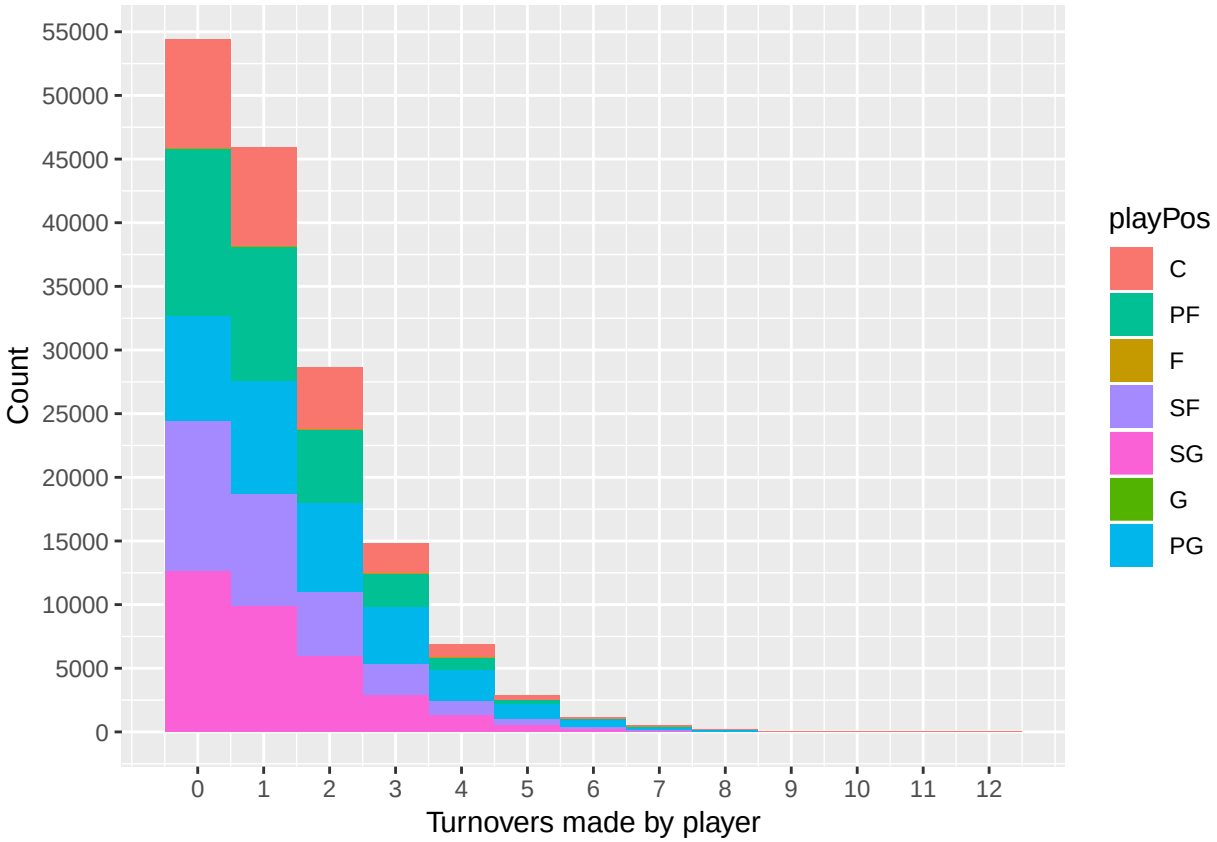



```
ggplot(data = nba_total,aes(x=playTRB,fill=playPos))+geom_histogram(position="fill",binwidth = 1)+
  scale_x_continuous(name="Rebounds made by player",breaks=seq(0,30,2))+
  ylab("Proportion")+scale_fill_discrete(breaks=c("C","PF","F","SF","SG","G","PG"))
```

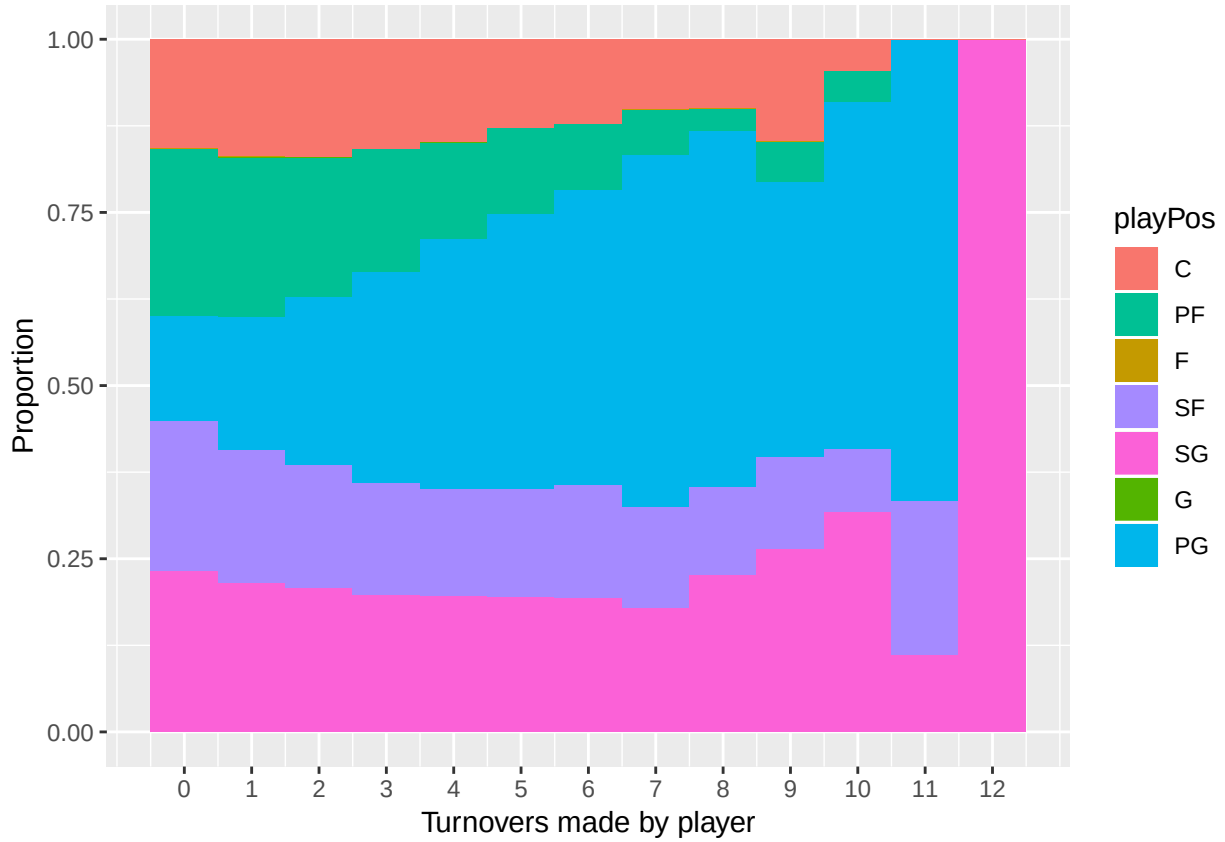


Those two above graphs show that the Center and Power forward are much easier to get more rebounds.

```
ggplot(data = nba_total,aes(x=playTO,fill=playPos))+geom_histogram(binwidth = 1)+
  scale_x_continuous(name="Turnovers made by player",breaks=seq(0,12,1))+
  scale_y_continuous(name="Count",breaks=seq(0,55000,5000))+
  scale_fill_discrete(breaks=c("C","PF","F","SF","SG","G","PG"))
```

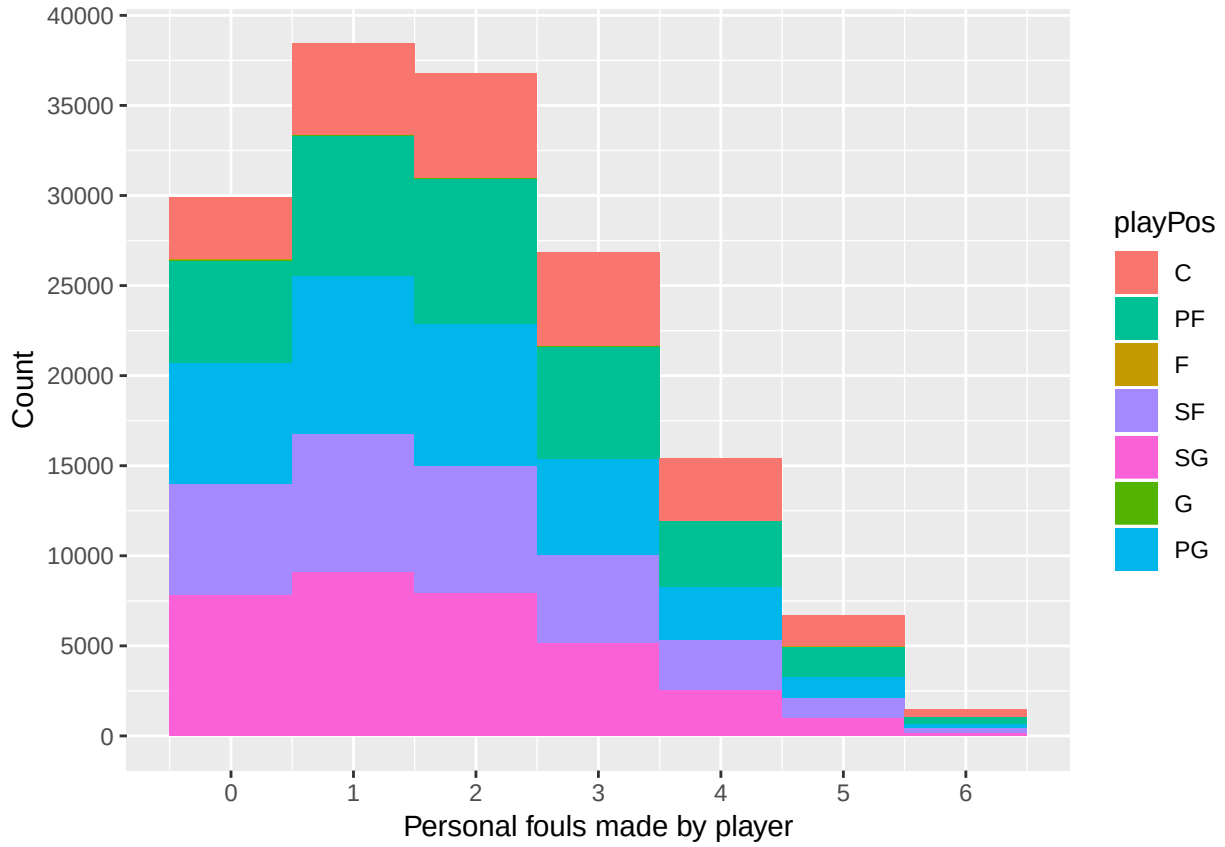



```
ggplot(data = nba_total,aes(x=playTO,fill=playPos))+geom_histogram(position="fill",binwidth = 1)+
  scale_x_continuous(name="Turnovers made by player",breaks=seq(0,12,1))+
  ylab("Proportion")+scale_fill_discrete(breaks=c("C","PF","F","SF","SG","G","PG"))
```

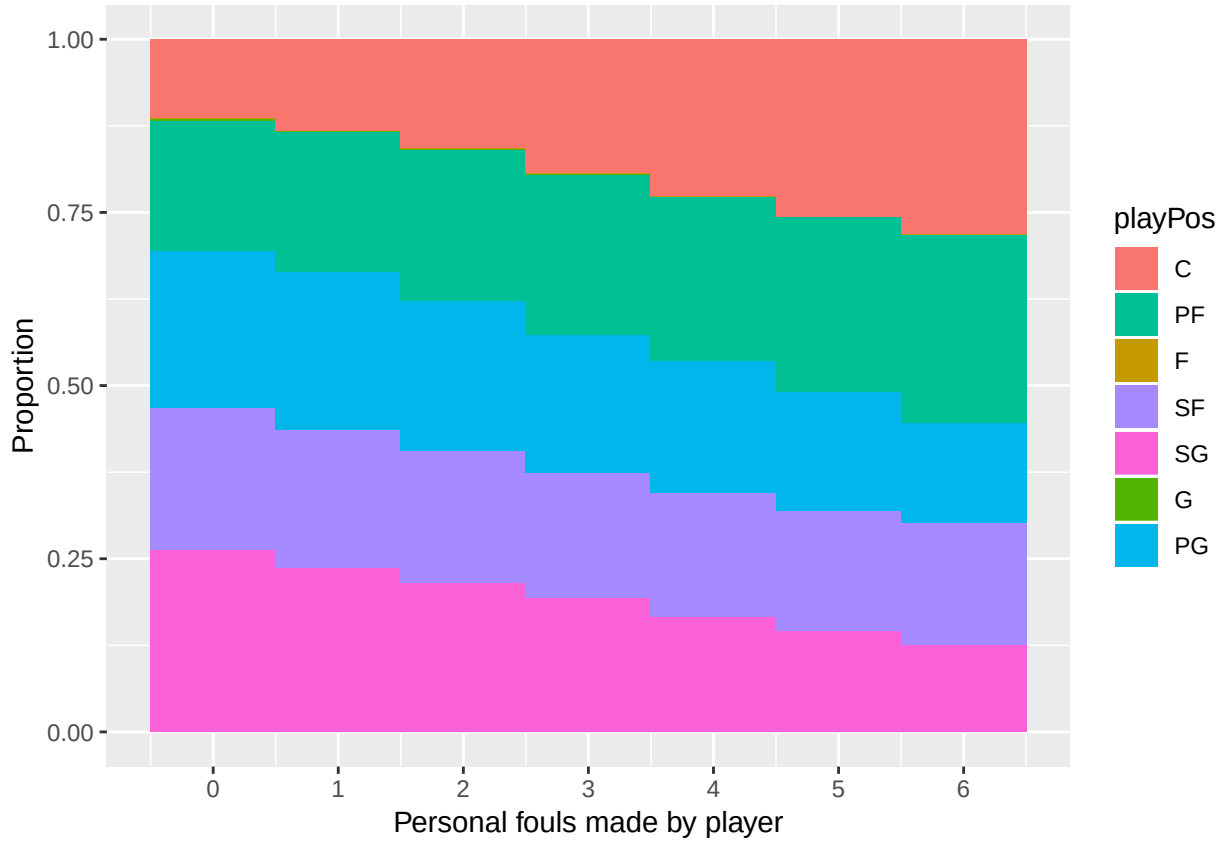


Those two above graphs show that PG and SG are much easier to have more turnovers.

```
ggplot(data = nba_total,aes(x=playPF,fill=playPos))+geom_histogram(binwidth = 1)+
  scale_x_continuous(name="Personal fouls made by player",breaks=seq(0,6,1))+
  scale_y_continuous(name="Count",breaks=seq(0,40000,5000))+
  scale_fill_discrete(breaks=c("C","PF","F","SF","SG","G","PG"))
```

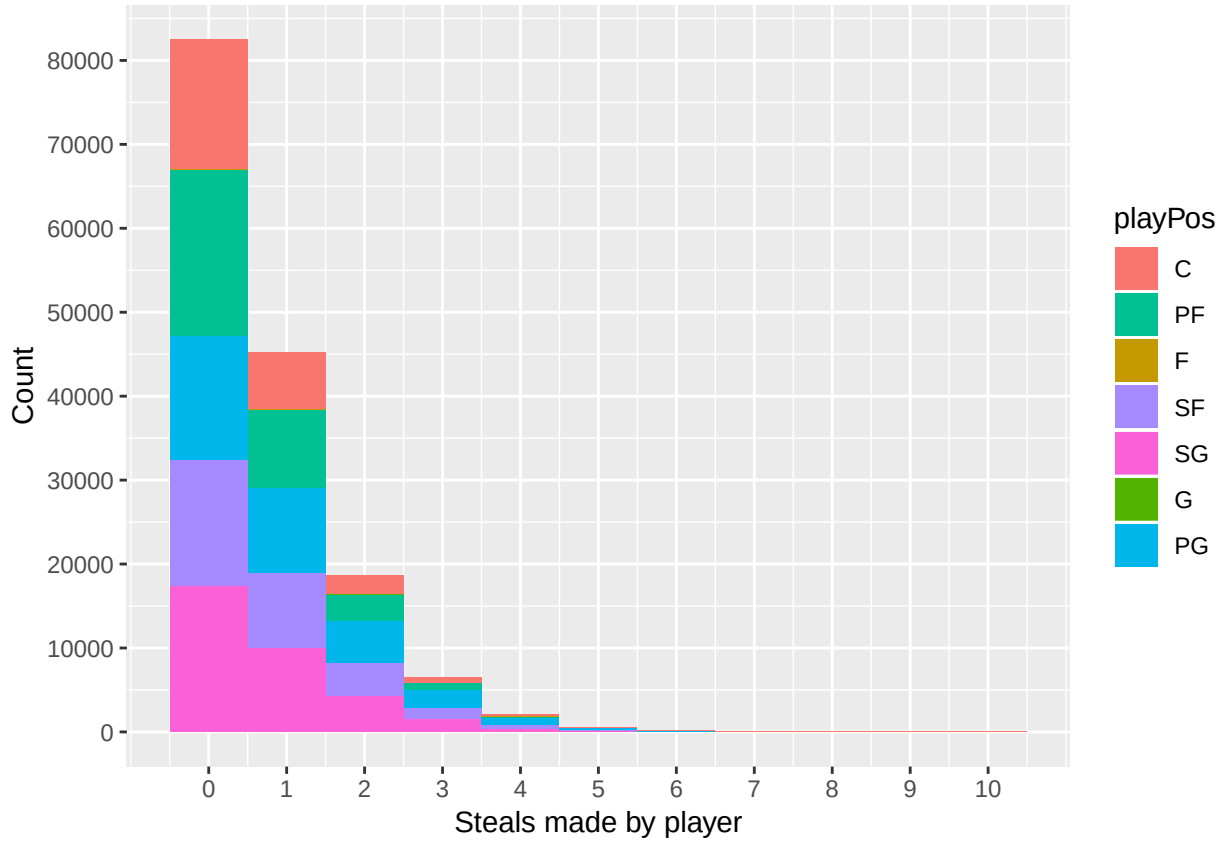



```
ggplot(data = nba_total,aes(x=playPF,fill=playPos))+geom_histogram(position="fill",binwidth = 1)+
  scale_x_continuous(name="Personal fouls made by player",breaks=seq(0,6,1))+
  ylab("Proportion")+scale_fill_discrete(breaks=c("C","PF","F","SF","SG","G","PG"))
```

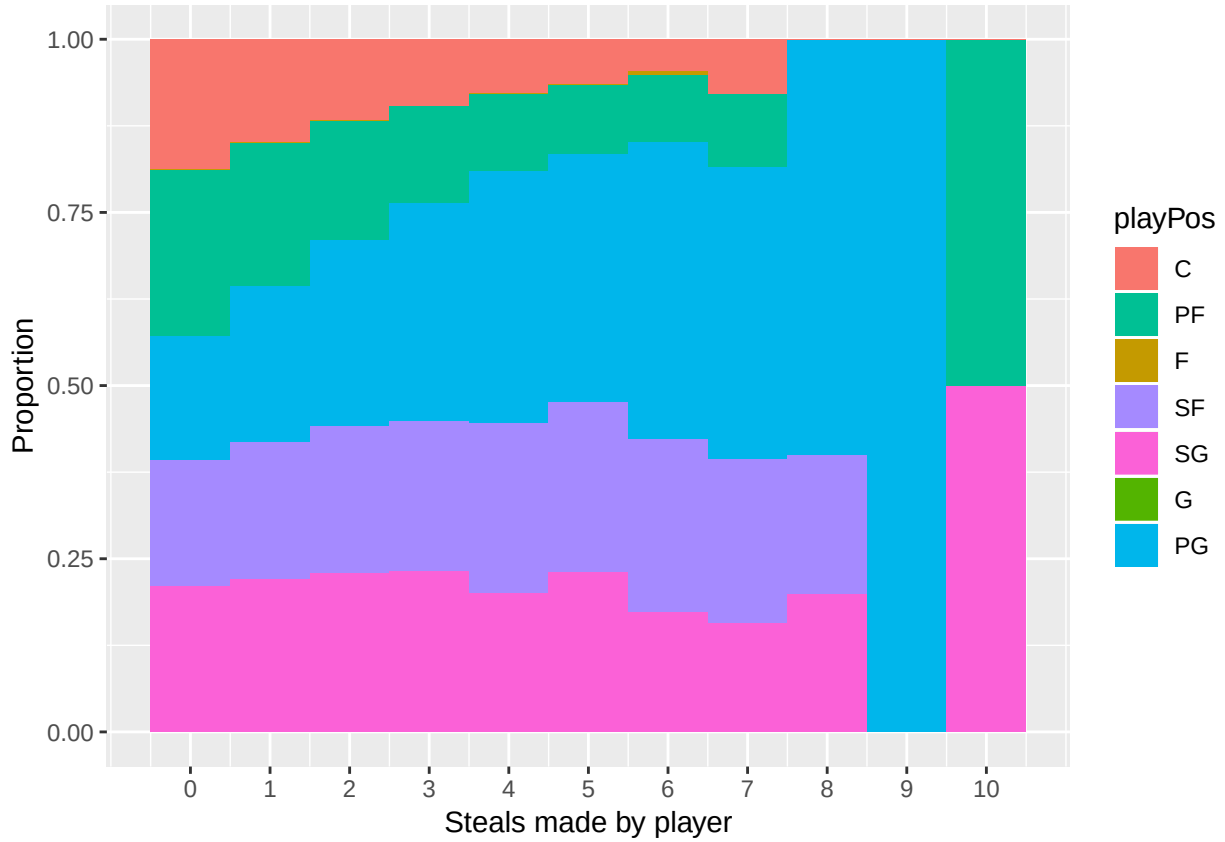


Those two above graphs show that center and power forward are much easier to have more fouls.

```
ggplot(data = nba_total,aes(x=playSTL,fill=playPos))+geom_histogram(binwidth = 1)+
  scale_x_continuous(name="Steals made by player",breaks=seq(0,10,1))+
  scale_y_continuous(name="Count",breaks=seq(0,90000,10000))+
  scale_fill_discrete(breaks=c("C","PF","F","SF","SG","G","PG"))
```

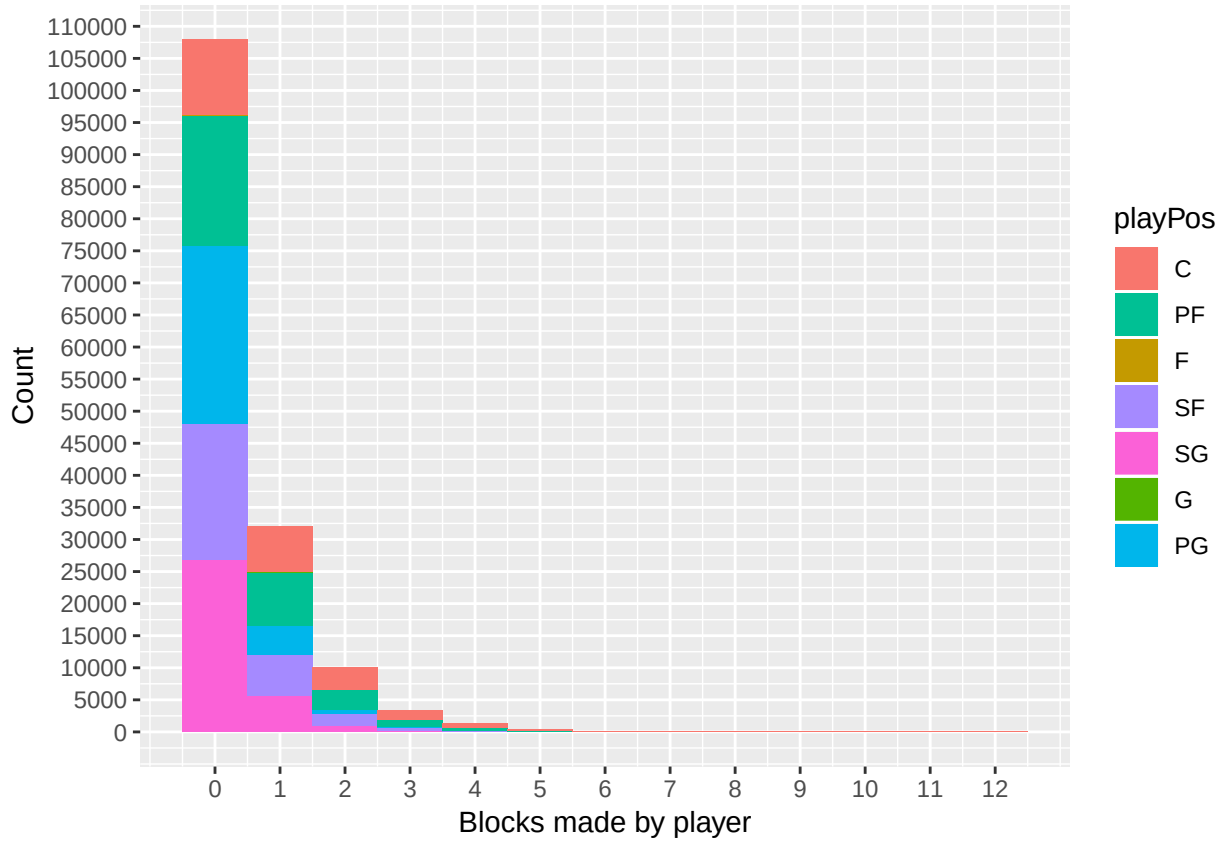



```
ggplot(data = nba_total,aes(x=playSTL,fill=playPos))+geom_histogram(position="fill",binwidth = 1)+
  scale_x_continuous(name="Steals made by player",breaks=seq(0,10,1))+
  ylab("Proportion")+scale_fill_discrete(breaks=c("C","PF","F","SF","SG","G","PG"))
```

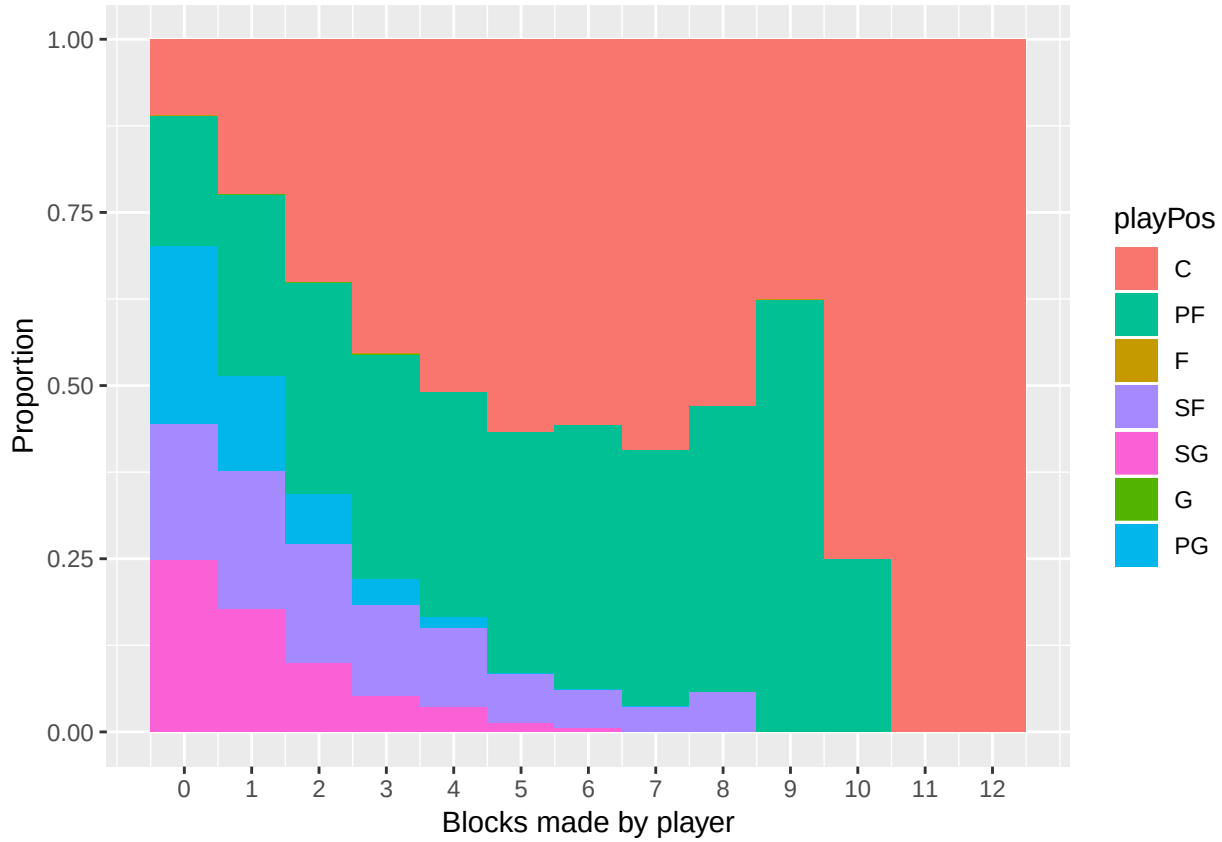


Those two above graphs show that pg are much easier to get steals.

```
ggplot(data = nba_total,aes(x=playBLK,fill=playPos))+geom_histogram(binwidth = 1)+
  scale_x_continuous(name="Blocks made by player",breaks=seq(0,12,1))+
  scale_y_continuous(name="Count",breaks=seq(0,110000,5000))+
  scale_fill_discrete(breaks=c("C","PF","F","SF","SG","G","PG"))
```




```
ggplot(data = nba_total,aes(x=playBLK,fill=playPos))+geom_histogram(position="fill",binwidth = 1)+
  scale_x_continuous(name="Blocks made by player",breaks=seq(0,12,1))+
  ylab("Proportion")+scale_fill_discrete(breaks=c("C","PF","F","SF","SG","G","PG"))
```



Those two above graphs show that C and PF are much easier to get blocks.