Problem Set 6

Subeom Lee

2019-03-12

```
## Warning: package 'knitr' was built under R version 3.5.3
```

Questions

```r
#http://asbcllc.com/nbastatR/index.html

library(nbastatR)
library(future)
library(stringi)
library(tidyverse)
library(lubridate)
library(texreg)
library(broom)
library(knitr)
library(ggpubr)
library(ggrepel)
library(janitor)
library(plotly)
library(reticulate)

plan(multiprocess)

# Run only when needed
# game_logs(seasons = 1947:2019, result_types = c("team", "player"))
# dataGameLogsTeam$Team = substring(dataGameLogsTeam$slugMatchup, 1, 3)

# Run when you updated data
# save(df_nba_player_dict, file='df_nba_player_dict.Rdata')
# save(dataGameLogsTeam, file='dataGameLogsTeam.Rdata')
# save(dataGameLogsPlayer, file='dataGameLogsPlayer.Rdata')

# load('df_nba_player_dict.Rdata')
# load('dataGameLogsTeam.Rdata')
# load('dataGameLogsPlayer.Rdata')

load('BaseEnvironment.Rdata')



# avg <- aggregate(dataGameLogsTeam[, 24:46], list(dataGameLogsTeam$yearSeason, dataGameLogsTeam$Team), mean)
# colnames(avg)[1] <- "Year"
# colnames(avg)[2] <- "Team"
#
#
#
# avgplot <- avg %>%
#             filter(Team %in% c('GSW', 'CHI', 'HOU', 'LAL')) %>%
#             ggplot(aes(x=Year, y=pctFG3Team, colour=Team)) +
#             geom_line()
# avgplot
#
# avg2 <- aggregate(dataGameLogsTeam[, 24:46], list(dataGameLogsTeam$yearSeason), mean)
# colnames(avg2)[1] <- "Year"
# min2 <- aggregate(dataGameLogsTeam[, 24:46], list(dataGameLogsTeam$yearSeason), min)
# colnames(min2)[1] <- "Year"
```

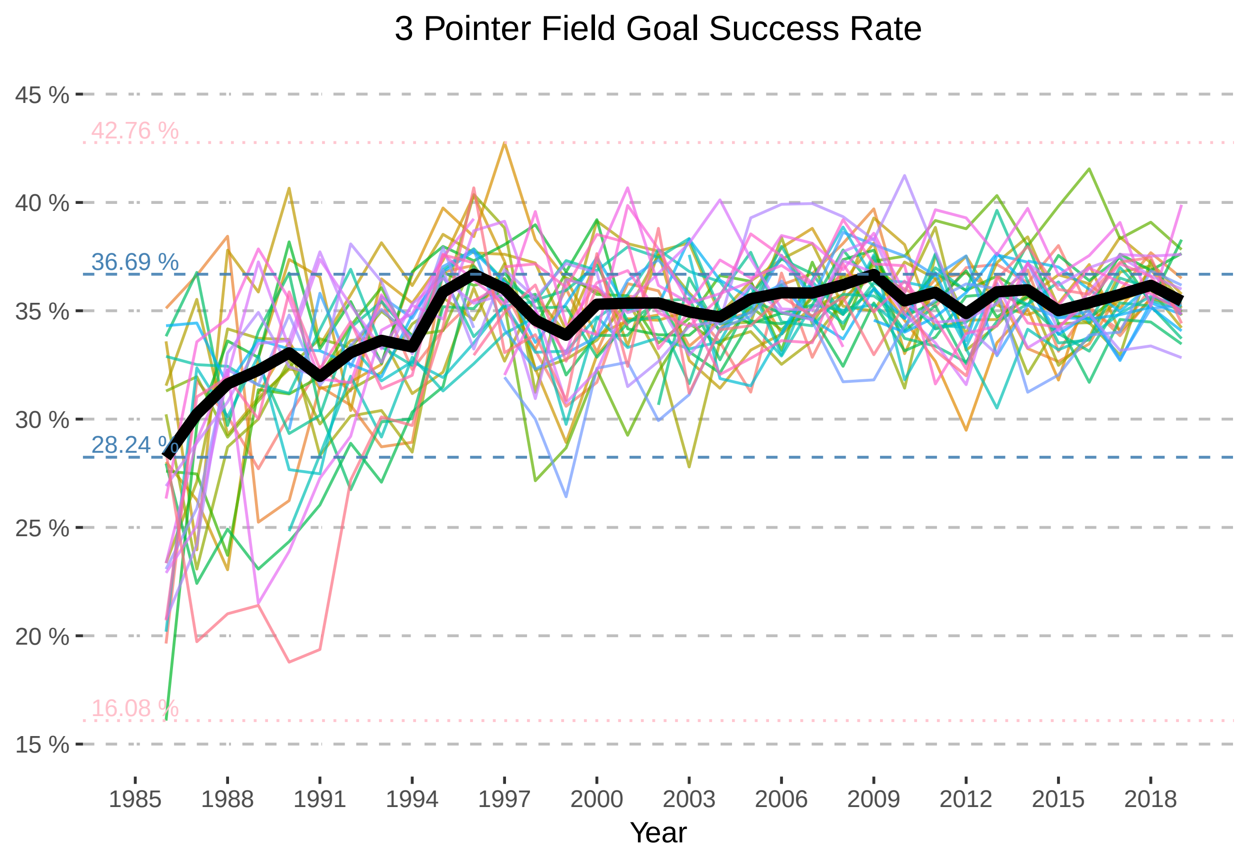
```
# max2 <- aggregate(dataGameLogsTeam[, 24:46], list(dataGameLogsTeam$yearSeason), max)
# colnames(max2)[1] <- "Year"
#
# avgplot2 <- avg2 %>%
#               filter(Year >= 1986) %>%
#               ggplot(aes(x=Year, y=pctFG3Team*100)) +
#               geom_path(colour='violet', size=2)
# avgplot2
#
# avgplot3 <- avg2 %>%
#               ggplot(aes(x=Year, y=pctFG2Team*100)) +
#               geom_path(colour='red', size=2)
# avgplot3
#
# avgminmax <- ggplot() +
#                 geom_line(data=min2, aes(x=Year, y=pctFG3Team*100), size=1, colour='green') +
#                 geom_line(data=max2, aes(x=Year, y=pctFG2Team*100), size=1, colour='red') +
#                 geom_smooth(data=avg2, aes(x=Year, y=pctFG3Team*100), size=2, colour='black')
# avgminmax
#
# avg1986 <- avg %>% filter(Year>=1986)
# avg21986 <- avg2 %>% filter(Year>=1986)
#
# avgcombinedall <- ggplot() +
#                 geom_line(data=avg1986, aes(x=Year, y=pctFG3Team*100, colour=Team), size=0.5, show.legend=FALSE) +
#                 geom_line(data=avg1986, aes(x=Year, y=pctFG2Team*100, colour=Team), size=0.5, show.legend=FALSE) +
#                 geom_line(data=avg21986, aes(x=Year, y=pctFG3Team*100), size=2, colour='black')+
#                 geom_line(data=avg21986, aes(x=Year, y=pctFG2Team*100), size=2, colour='black')
# avgcombinedall
#
# avgfiltered <- avg %>% filter(Team %in% c('GSW', 'CHI', 'HOU', 'LAL'))
#
# avgcombined <- ggplot() +
#                 geom_line(data=avgfiltered, aes(x=Year, y=pctFG3Team*100, colour=Team), size=1, show.legend=FALSE) +
#                 geom_line(data=avgfiltered, aes(x=Year, y=pctFG2Team*100, colour=Team), size=1, show.legend=FALSE) +
#                 geom_smooth(data=avg2, aes(x=Year, y=pctFG3Team*100), size=2, colour='black')
# avgcombined
#
# avgcombined2 <- avgcombined +
#                 geom_smooth(data=avg2, aes(x=Year, y=pctFG2Team*100), size=2, colour='red')
# avgcombined2
#
# avgcombined3 <- avgcombined2 +
#                 geom_smooth(data=avg2, aes(x=Year, y=pctFTTeam*100), size=2, colour='green')
# avgcombined3
#
# avgcombined4 <- avgcombined3 +
#                 geom_smooth(data=avg2, aes(x=Year, y=ptsTeam), size=2, colour='purple')
# avgcombined4

# ggplotly(p=ggplot2::last_plot())

# library(ggplot2)
# library(ggpubr)
# theme_set(theme_pubr())
#
# figure <- ggarrange(avgplot, avgplot2,
#                     labels = c("Each Team", "All Teams"),
#                     ncol = 1, nrow = 2)
# figure

# climate <- read.csv('ps5_data.csv')
# a <- ggplot(climate) +
#       xlab('Year') +
#       ylab('Temperature(°C)') +
#       theme(panel.border=element_rect(colour="black", fill=NA), panel.background=element_rect(fill=NA),
#             panel.grid=element_line(color="grey")) +
```

```
#       geom_smooth(aes(Year, Lowess.5.), colour="blue", size=1) +
#       geom_line(aes(Year, No_Smoothing), colour="grey", size=1) +
#       geom_point(aes(Year, No_Smoothing), shape=1, size=3)
#
```

Team level questions

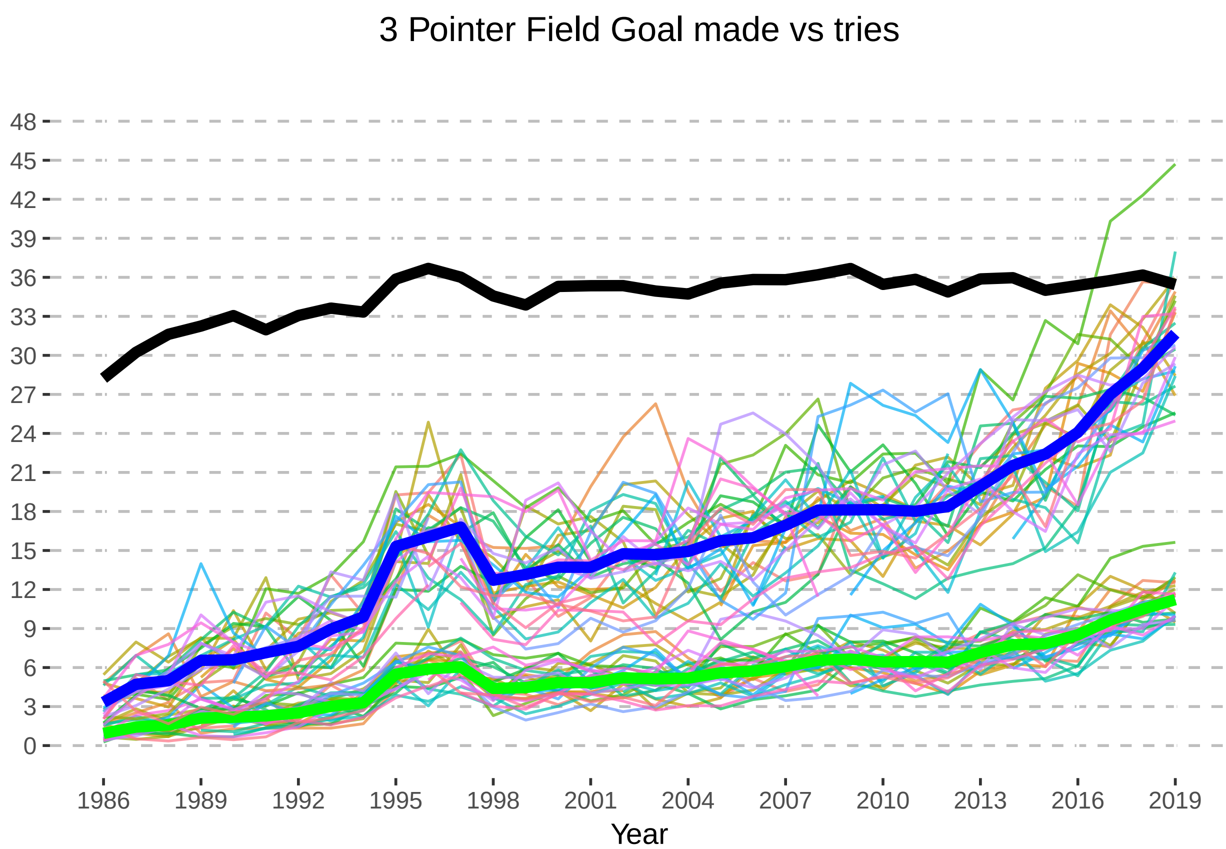
Q1. It seems that players are getting better at making 3-pointers than 20 years ago (both on average and also top 3-pointer shooters vs. top 3-pointer shooters) Is it true?

```
fg3year <- aggregate(dataGameLogsTeam[, 35:36], list(dataGameLogsTeam$yearSeason), sum)
colnames(fg3year)[1] <- "Year"
fg3year <- fg3year %>% filter (Year >= 1986)
fg3year$pctfg3 <- fg3year$fg3mTeam / fg3year$fg3aTeam * 100

fg3yearteam <- aggregate(dataGameLogsTeam[, 35:36], list(dataGameLogsTeam$yearSeason, dataGameLogsTeam$T
colnames(fg3yearteam)[1] <- "Year"
colnames(fg3yearteam)[2] <- "Team"
fg3yearteam <- fg3yearteam %>% filter (Year >= 1986)
fg3yearteam$pctfg3 <- fg3yearteam$fg3mTeam / fg3yearteam$fg3aTeam * 100

xaxisbreaks <- seq(1985, 2019, by=3)
yaxisbreaks <- seq(15, 45, by=5)

Q1 <- ggplot() +
  geom_line(data=fg3yearteam, aes(x=Year, y=pctfg3, colour=Team), size=0.5, show.legend=FALSE, alpha=0.7
  geom_line(data=fg3year, aes(x=Year, y=pctfg3), size=2, colour='black') +
  xlab('Year') +
  ylab(NULL) +
  ggtitle('3 Pointer Field Goal Success Rate') +
  theme(panel.background=element_rect(fill=NA), panel.grid.major.y=element_line(color="grey", linetype=2
        plot.title = element_text(hjust = 0.5)) +
  scale_y_continuous(limits=c(15, 45), breaks=yaxisbreaks, labels=paste(yaxisbreaks,"%")) +
  scale_x_continuous(limits=c(1985,2019), breaks=xaxisbreaks) +
  geom_hline(yintercept=min(fg3year$pctfg3), linetype=2, color="steelblue", size=0.5, alpha=0.9) +
  geom_hline(yintercept=max(fg3year$pctfg3), linetype=2, color="steelblue", size=0.5, alpha=0.9) +
  geom_hline(yintercept=min(fg3yearteam$pctfg3), linetype=3, color="pink", size=0.5, alpha=0.9) +
  geom_hline(yintercept=max(fg3yearteam$pctfg3), linetype=3, color="pink", size=0.5, alpha=0.9) +
  annotate("text", x=1985, y=min(fg3year$pctfg3)+0.6, label=paste(toString(round(min(fg3year$pctfg3), di
  annotate("text", x=1985, y=max(fg3year$pctfg3)+0.6, label=paste(toString(round(max(fg3year$pctfg3), di
  annotate("text", x=1985, y=min(fg3yearteam$pctfg3)+0.6, label=paste(toString(round(min(fg3yearteam$pct
  annotate("text", x=1985, y=max(fg3yearteam$pctfg3)+0.6, label=paste(toString(round(max(fg3yearteam$pct

Q1
```



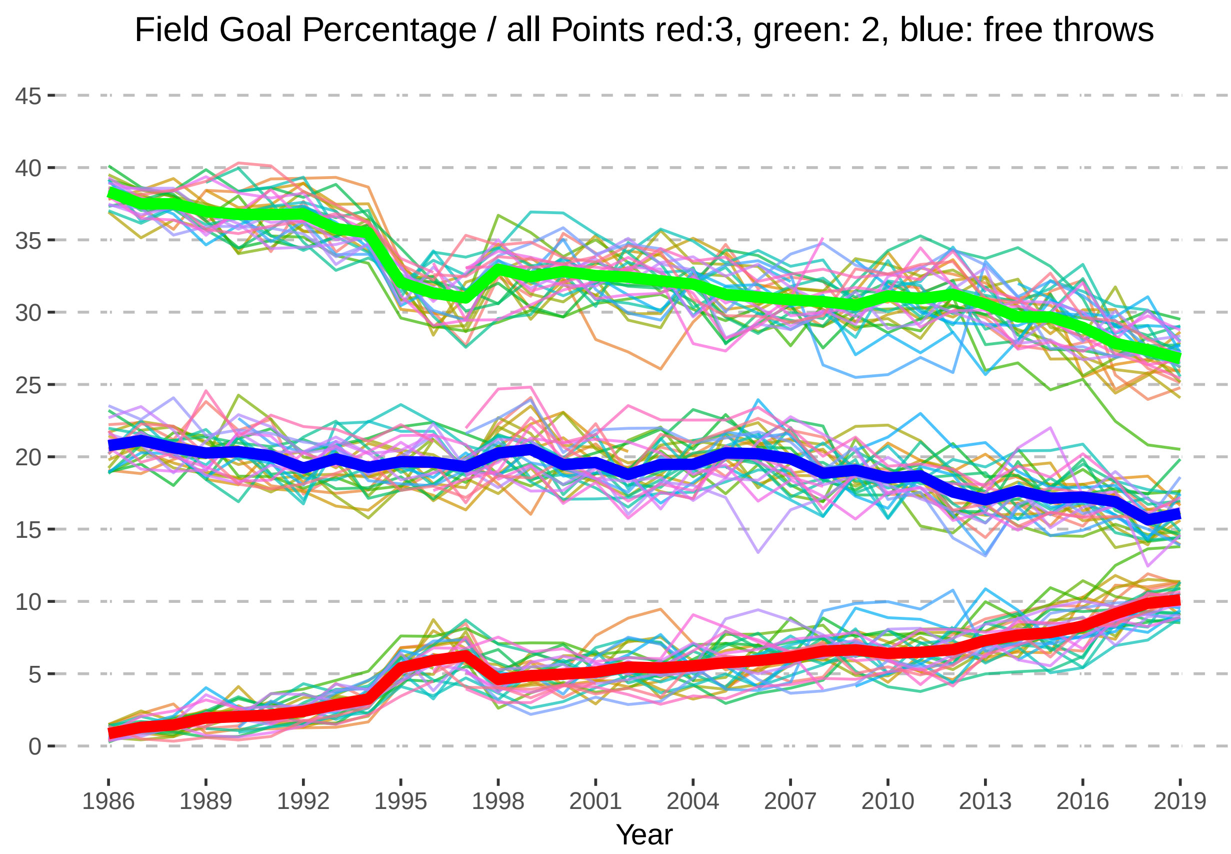

```
fg3yearavg <- aggregate(dataGameLogsTeam[, 35:36], list(dataGameLogsTeam$yearSeason), mean)
colnames(fg3yearavg)[1] <- "Year"
fg3yearavg <- fg3yearavg %>% filter (Year >= 1986)
fg3yearavg$pctfg3 <- fg3yearavg$fg3mTeam / fg3yearavg$fg3aTeam * 100

fg3yearteamavg <- aggregate(dataGameLogsTeam[, 35:36], list(dataGameLogsTeam$yearSeason, dataGameLogsTea
colnames(fg3yearteamavg)[1] <- "Year"
colnames(fg3yearteamavg)[2] <- "Team"
fg3yearteamavg <- fg3yearteamavg %>% filter (Year >= 1986)
fg3yearteamavg$pctfg3 <- fg3yearteamavg$fg3mTeam / fg3yearteamavg$fg3aTeam * 100

xaxisbreaks <- seq(1986, 2019, by=3)
yaxisbreaks <- seq(0, 50, by=3)

Q1_2 <- ggplot() +
  geom_line(data=fg3yearteamavg, aes(x=Year, y=fg3mTeam, colour=Team), size=0.5, show.legend=FALSE, alph
  geom_line(data=fg3yearavg, aes(x=Year, y=fg3mTeam), size=2, colour='green') +
  geom_line(data=fg3yearteamavg, aes(x=Year, y=fg3aTeam, colour=Team), size=0.5, show.legend=FALSE, alph
  geom_line(data=fg3yearavg, aes(x=Year, y=fg3aTeam), size=2, colour='blue') +
  geom_line(data=fg3year, aes(x=Year, y=pctfg3), size=2, colour='black') +
  xlab('Year') +
  ylab(NULL) +
  ggtitle('3 Pointer Field Goal made vs tries') +
  theme(panel.background=element_rect(fill=NA), panel.grid.major.y=element_line(color="grey", linetype=
```

```
      plot.title = element_text(hjust = 0.5)) +
  scale_y_continuous(limits=c(0, 50), breaks=yaxisbreaks) +
  scale_x_continuous(limits=c(1986,2019), breaks=xaxisbreaks)

Q1_2
```




```
fgallyearavg <- aggregate(dataGameLogsTeam[, 29:38], list(dataGameLogsTeam$yearSeason), mean)
colnames(fgallyearavg)[1] <- "Year"
fgallyearavg["plusminusTeam"] = NULL
fgallyearavg["urlTeamSeasonLogo"] = NULL
fgallyearavg["pfTeam"] = NULL
fgallyearavg <- fgallyearavg %>% filter (Year >= 1986)
fgallyearavg$pctpts3 <- fgallyearavg$fg3mTeam / fgallyearavg$ptsTeam * 100
fgallyearavg$pctpts2 <- fgallyearavg$fg2mTeam / fgallyearavg$ptsTeam * 100
fgallyearavg$pctptsft <- fgallyearavg$ftmTeam / fgallyearavg$ptsTeam * 100

fgallyearteamavg <- aggregate(dataGameLogsTeam[, 29:38], list(dataGameLogsTeam$yearSeason, dataGameLogs
colnames(fgallyearteamavg)[1] <- "Year"
colnames(fgallyearteamavg)[2] <- "Team"
fgallyearteamavg["plusminusTeam"] = NULL
fgallyearteamavg["urlTeamSeasonLogo"] = NULL
fgallyearteamavg["pfTeam"] = NULL
fgallyearteamavg <- fgallyearteamavg %>% filter (Year >= 1986)
fgallyearteamavg$pctpts3 <- fgallyearteamavg$fg3mTeam / fgallyearteamavg$ptsTeam * 100
```

```
fgallyearteamavg$pctpts2 <- fgallyearteamavg$fg2mTeam / fgallyearteamavg$ptsTeam * 100
fgallyearteamavg$pctptsft <- fgallyearteamavg$ftmTeam / fgallyearteamavg$ptsTeam * 100

xaxisbreaks <- seq(1986, 2019, by=3)
yaxisbreaks <- seq(0, 45, by=5)

Q1_3 <- ggplot() +
  geom_line(data=fgallyearteamavg, aes(x=Year, y=pctpts3, colour=Team), size=0.5, show.legend=FALSE, alp
  geom_line(data=fgallyearteamavg, aes(x=Year, y=pctpts2, colour=Team), size=0.5, show.legend=FALSE, alp
  geom_line(data=fgallyearteamavg, aes(x=Year, y=pctptsft, colour=Team), size=0.5, show.legend=FALSE, al
  geom_line(data=fgallyearavg, aes(x=Year, y=pctpts3), size=2, colour='red') +
  geom_line(data=fgallyearavg, aes(x=Year, y=pctpts2), size=2, colour='green') +
  geom_line(data=fgallyearavg, aes(x=Year, y=pctptsft), size=2, colour='blue') +
  xlab('Year') +
  ylab(NULL) +
  ggtitle('Field Goal Percentage / all Points red:3, green: 2, blue: free throws') +
  theme(panel.background=element_rect(fill=NA), panel.grid.major.y=element_line(color="grey", linetype=
        plot.title = element_text(hjust = 0.5)) +
  scale_y_continuous(limits=c(0, 45), breaks=yaxisbreaks) +
  scale_x_continuous(limits=c(1986,2019), breaks=xaxisbreaks)

Q1_3
```



Yes, the success rate of 3 point field goal has been increased by about 9% since 1986.

Q2. If true, what could be the reasons for that? - What are the expected average points of 3-pointers and

2-pointers? Show the historical data. - If the expected average point from 3-pointers is getting higher than that of 2-pointers, how should each team's strategy changes

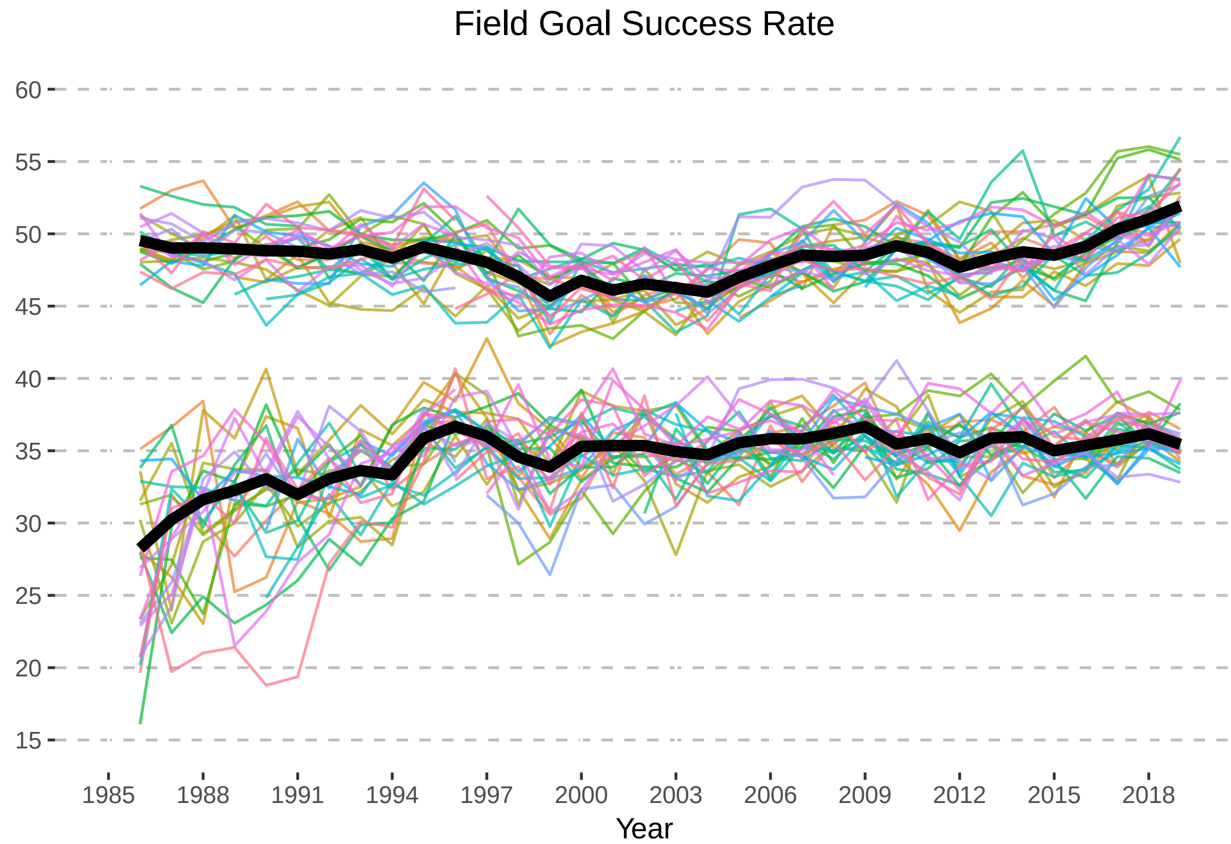
https://www.nytimes.com/2016/01/21/sports/basketball/how-the-nba-3-point-shot-went-from-gimmick-to-game-changer.html

Its debut, in the 1979-80 season, was inauspicious.

There are many reasons for the rise of the 3-point shot, but one may simply be math. It took a while, but coaches finally stopped listening to the traditionalist naysayers and realized that a shot that is worth 50 percent more pays off, even if that shot is a little harder to make.

"Teams have all caught on to the whole points-per-possession argument," Lawrence Frank, the Nets' coach at the time, said in 2009 as the 3 rate began to rapidly increase.

```
fgyear <- aggregate(dataGameLogsTeam[, 35:38], list(dataGameLogsTeam$yearSeason), sum)
colnames(fgyear)[1] <- "Year"
fgyear <- fgyear %>% filter (Year >= 1986)
fgyear$pctfg3 <- fgyear$fg3mTeam / fgyear$fg3aTeam * 100
fgyear$pctfg2 <- fgyear$fg2mTeam / fgyear$fg2aTeam * 100

fgyearteam <- aggregate(dataGameLogsTeam[, 35:38], list(dataGameLogsTeam$yearSeason, dataGameLogsTeam$Te
colnames(fgyearteam)[1] <- "Year"
colnames(fgyearteam)[2] <- "Team"
fgyearteam <- fgyearteam %>% filter (Year >= 1986)
fgyearteam$pctfg3 <- fgyearteam$fg3mTeam / fgyearteam$fg3aTeam * 100
fgyearteam$pctfg2 <- fgyearteam$fg2mTeam / fgyearteam$fg2aTeam * 100

xaxisbreaks <- seq(1985, 2019, by=3)
yaxisbreaks <- seq(15, 60, by=5)

Q2_1 <- ggplot() +
  geom_line(data=fgyearteam, aes(x=Year, y=pctfg3, colour=Team), size=0.5, show.legend=FALSE, alpha=0.7)
  geom_line(data=fgyear, aes(x=Year, y=pctfg3), size=2, colour='black') +
  geom_line(data=fgyearteam, aes(x=Year, y=pctfg2, colour=Team), size=0.5, show.legend=FALSE, alpha=0.7)
  geom_line(data=fgyear, aes(x=Year, y=pctfg2), size=2, colour='black') +
  xlab('Year') +
  ylab(NULL) +
  ggtitle('Field Goal Success Rate') +
  theme(panel.background=element_rect(fill=NA), panel.grid.major.y=element_line(color="grey", linetype=2
        plot.title = element_text(hjust = 0.5)) +
  scale_y_continuous(limits=c(15, 60), breaks=yaxisbreaks, labels=yaxisbreaks) +
  scale_x_continuous(limits=c(1985,2019), breaks=xaxisbreaks)# +
  # geom_hline(yintercept=min(fg3year$pctfg3), linetype=2, color="steelblue", size=0.5, alpha=0.9) +
  # geom_hline(yintercept=max(fg3year$pctfg3), linetype=2, color="steelblue", size=0.5, alpha=0.9) +
  # geom_hline(yintercept=min(fg3yearteam$pctfg3), linetype=3, color="pink", size=0.5, alpha=0.9) +
  # geom_hline(yintercept=max(fg3yearteam$pctfg3), linetype=3, color="pink", size=0.5, alpha=0.9) +
  # annotate("text", x=1985, y=min(fg3year$pctfg3)+0.6, label=paste(toString(round(min(fg3year$pctfg3),
  # annotate("text", x=1985, y=max(fg3year$pctfg3)+0.6, label=paste(toString(round(max(fg3year$pctfg3),
  # annotate("text", x=1985, y=min(fg3yearteam$pctfg3)+0.6, label=paste(toString(round(min(fg3yearteam$
  # annotate("text", x=1985, y=max(fg3yearteam$pctfg3)+0.6, label=paste(toString(round(max(fg3yearteam$
  #
Q2_1
```



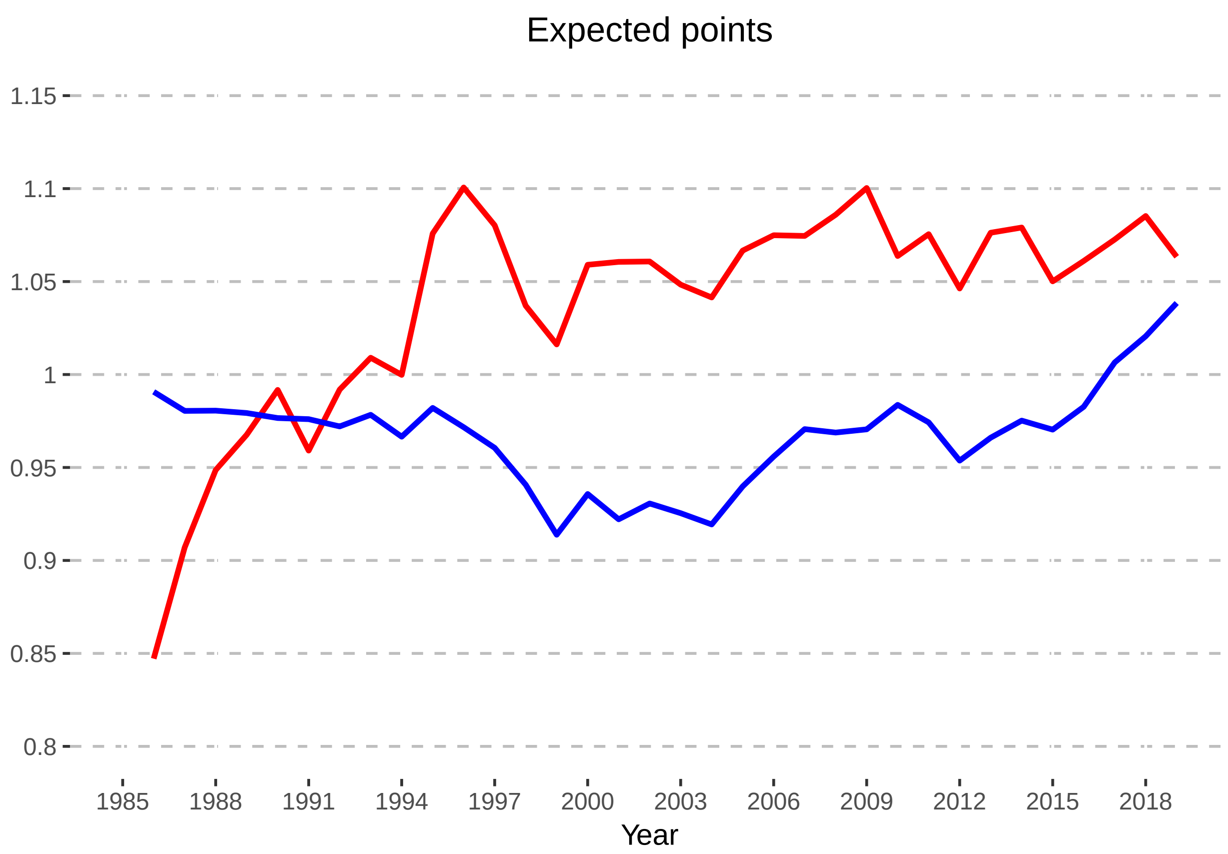
The expected points of 2-point shots in 1986 was ‘r fgyear$pct fg2[1986-1985]/100'*2 =' r f gyear$pctfg2[1986-1985]/100*2’ *The expected points of 3-point shots in 1986 was* ’r fgyear$pct f g3[1986 - 1985]/100' * 3 =' *r f gyearpctfg3[1986-1985]/100*3’

The expected points of 2-point shots in 2019 was ‘r fgyear$pct fg2[2019-1985]/100'*2 =' r f gyear$pctfg2[2019-1985]/100*2’ *The expected points of 3-point shots in 2019 was* ’r fgyear$pct f g3[2019 - 1985]/100' * 3 =' *r f gyearpctfg3[2019-1985]/100*3’

Teams started to focus on 3-point shots after its first introduction in 1979, because the expected points of 3-point shots are higher than that of 2-point shots since early 90’s.

```
fgyear$e2 = fgyear$pctfg2 / 100 * 2
fgyear$e3 = fgyear$pctfg3 / 100 * 3

xaxisbreaks <- seq(1985, 2019, by=3)
yaxisbreaks <- seq(0.8, 1.15, by=0.05)

Q2_2 <- ggplot() +
  geom_line(data=fgyear, aes(x=Year, y=e3), size=1, colour='red') +
  geom_line(data=fgyear, aes(x=Year, y=e2), size=1, colour='blue') +
  xlab('Year') +
  ylab(NULL) +
  ggtitle('Expected points') +
  theme(panel.background=element_rect(fill=NA), panel.grid.major.y=element_line(color="grey", linetype=
        plot.title = element_text(hjust = 0.5)) +
  scale_y_continuous(limits=c(0.8, 1.15), breaks=yaxisbreaks, labels=yaxisbreaks) +
  scale_x_continuous(limits=c(1985,2019), breaks=xaxisbreaks)
```

```
Q2_2
```

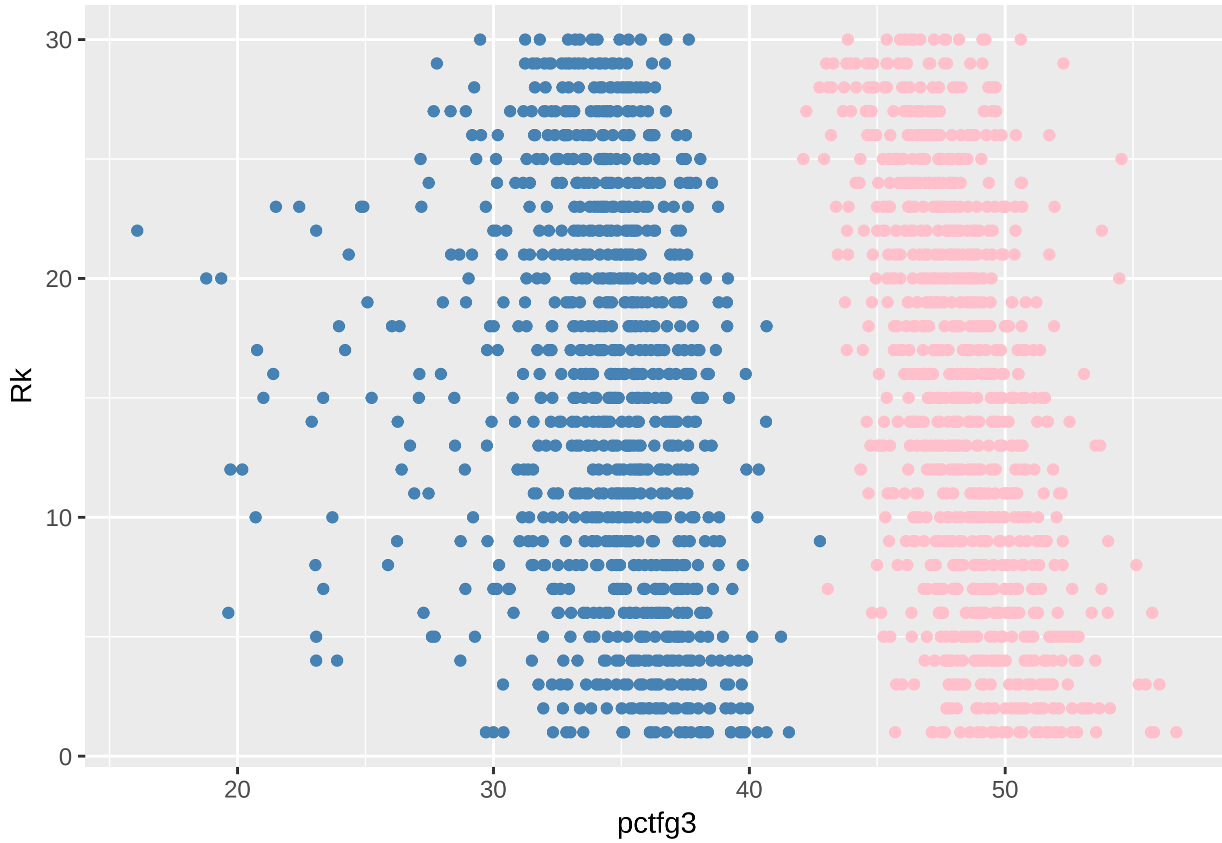


Q3. Teams with more 3-pointers tend to be the better performing teams? - Any insights between standings and 3-pointers?

```
standings <- read_csv("standings.csv")

fgyearteam <- aggregate(dataGameLogsTeam[, 35:38], list(dataGameLogsTeam$yearSeason, dataGameLogsTeam$na
colnames(fgyearteam)[1] <- "Year"
colnames(fgyearteam)[2] <- "nameTeam"
fgyearteam <- fgyearteam %>% filter (Year >= 1986)
fgyearteam$pctfg3 <- fgyearteam$fg3mTeam / fgyearteam$fg3aTeam * 100
fgyearteam$pctfg2 <- fgyearteam$fg2mTeam / fgyearteam$fg2aTeam * 100

standings2 <- left_join(standings, fgyearteam, by=c("Year" = "Year", "Team" = "nameTeam"))

Q3 <- ggplot(standings2) +
  geom_point(aes(x=pctfg3, y=Rk), color="steelblue") +
  geom_point(aes(x=pctfg2, y=Rk), color="pink")
  # geom_line(data=fgyear, aes(x=Year, y=e2), size=1, colour='blue') +
  # xlab('Year') +
  # ylab(NULL) +
  # ggtitle('Expected points') +
  # theme(panel.background=element_rect(fill=NA), panel.grid.major.y=element_line(color="grey", linetyp
  #       plot.title = element_text(hjust = 0.5)) +
```

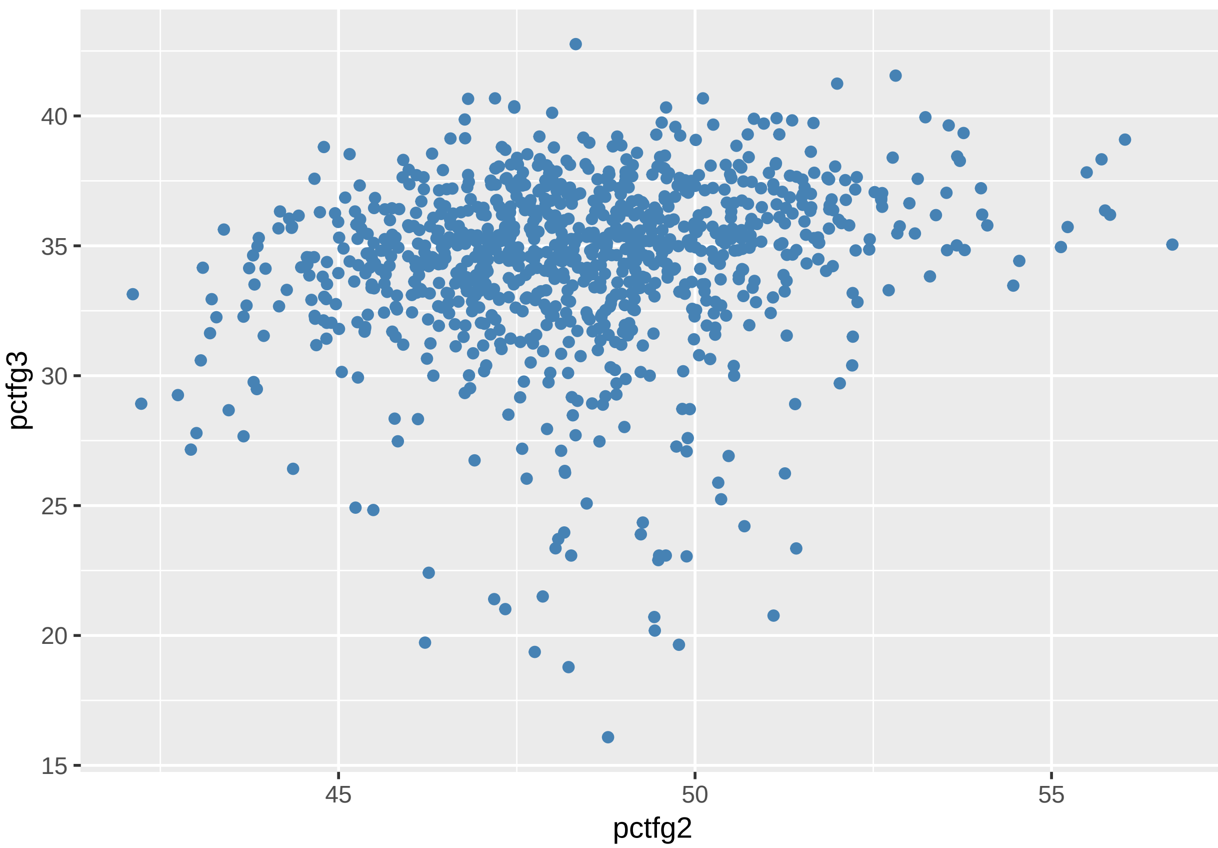
```
# scale_y_continuous(limits=c(0.8, 1.15), breaks=yaxisbreaks, labels=yaxisbreaks) +
# scale_x_continuous(limits=c(1985,2019), breaks=xaxisbreaks)
```

```
Q3
```




```
linearModel <- lm(Rk ~ pctfg3, data=standings2)
tidy(linearModel)
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)   32.6      2.72       12.0  5.33e-31
2 pctfg3        -0.518    0.0787     -6.58 7.74e-11

linearModel2 <- lm(Rk ~ pctfg2, data=standings2)
tidy(linearModel2)
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)   107.       4.97       21.6 2.14e-84
2 pctfg2         -1.91     0.103     -18.6 3.69e-66

linearModel3 <- lm(Rk ~ pctfg3 + pctfg2, data=standings2)
tidy(linearModel3)
# A tibble: 3 x 5
  term        estimate std.error statistic  p.value
```

```
  <chr>           <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)  114.        5.15       22.1  9.52e-88
2 pctfg3         -0.305    0.0694     -4.40 1.23e- 5
3 pctfg2         -1.83     0.103     -17.7  4.80e-61

# plot(linearModel)
# plot(linearModel2)
# plot(linearModel3)

linearModel4 <- lm(pctfg3 ~ pctfg2, data=standings2)
tidy(linearModel4)
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)   22.0       2.29       9.60 6.40e-21
2 pctfg2         0.257     0.0472     5.45 6.57e- 8
# plot(linearModel4)

Q3_2 <- ggplot(standings2) +
  geom_point(aes(x=pctfg2, y=pctfg3), color="steelblue")
Q3_2
```

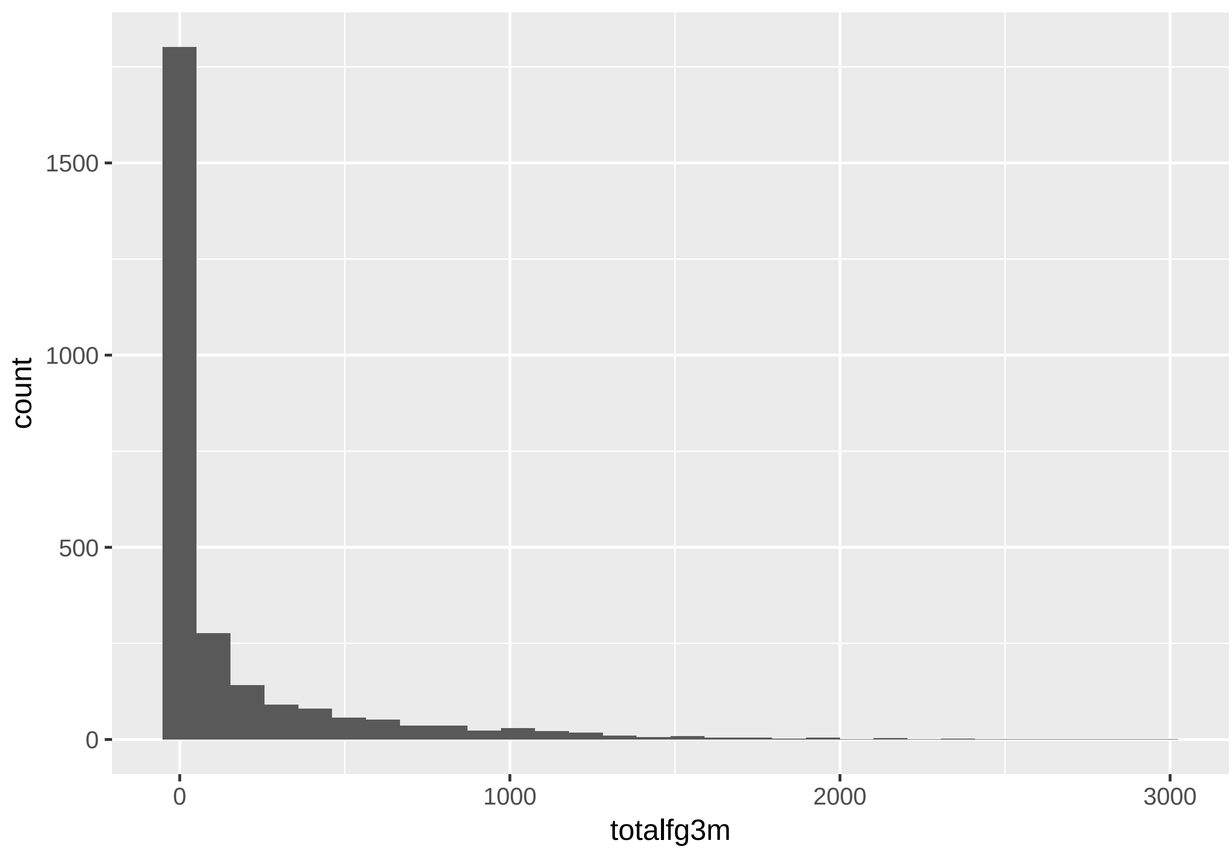


Yes. However, pctfg2 is more relevant than pctfg3

- Focus on three point shooting is a strategy that started fairly recently, we can create a map to show where this strategy initially emerged and how fast it spreaded across the entire country.

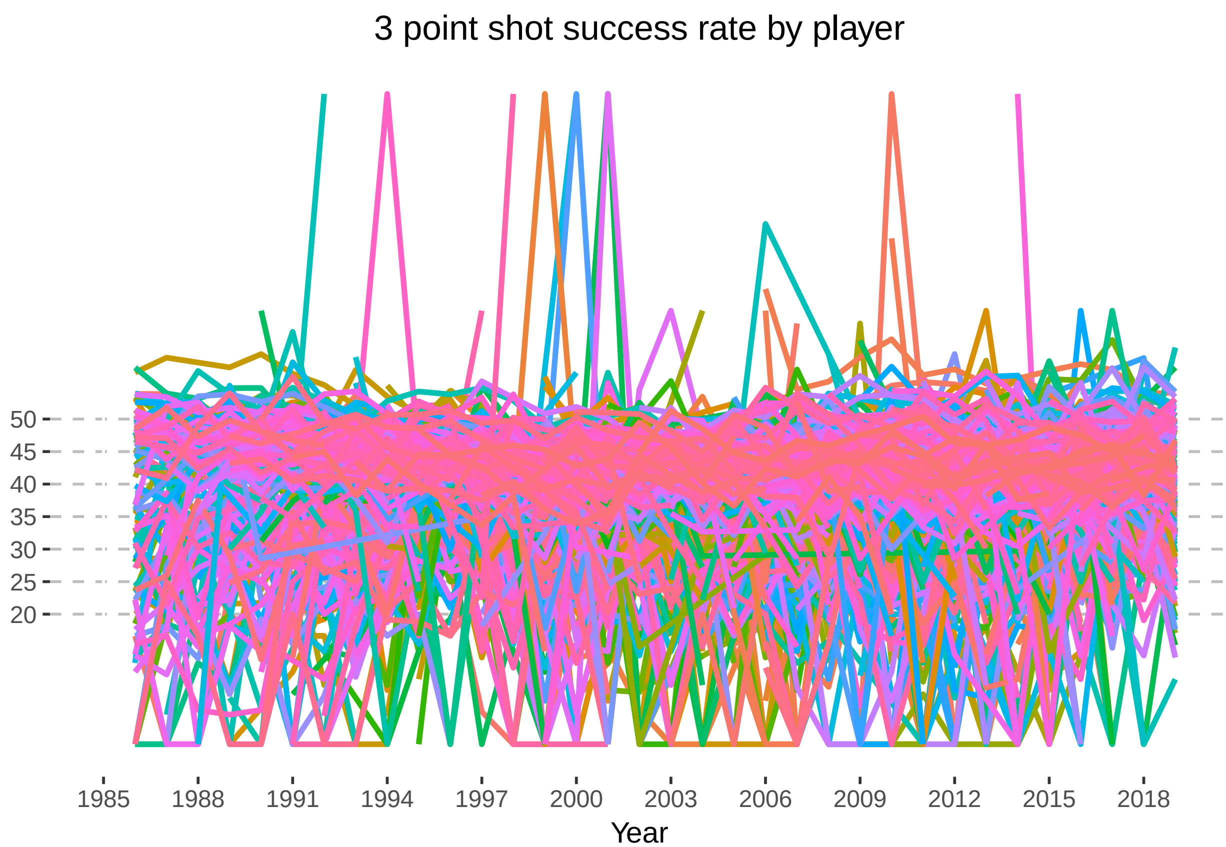
Player level questions

```r
dataGameLogsPlayer1986 <- dataGameLogsPlayer %>% filter(yearSeason >= 1986)

fgyearplayer <- aggregate(dataGameLogsPlayer1986[, 19:26], list(dataGameLogsPlayer1986$yearSeason, data(
colnames(fgyearplayer)[1] <- "Year"
colnames(fgyearplayer)[2] <- "Player"
fgyearplayer$pctFG = NULL
fgyearplayer$pctFG3 = NULL

fgyearplayer$pctfg3 <- fgyearplayer$fg3m / fgyearplayer$fg3a * 100
fgyearplayer$pctfg2 <- fgyearplayer$fgm / fgyearplayer$fga * 100
fgyearplayer$pctft <- fgyearplayer$ftm / fgyearplayer$fta * 100

# Meaningless...
yearplayer <- aggregate(fgyearplayer[,5], list(fgyearplayer$Player), sum)
colnames(yearplayer)[1] <- "Player"
colnames(yearplayer)[2] <- "totalfg3m"
ggplot(yearplayer, aes(totalfg3m)) + geom_histogram()
```




```r
yearplayer100 <- yearplayer %>% filter (totalfg3m>=100)

xaxisbreaks <- seq(1985, 2019, by=3)
yaxisbreaks <- seq(20, 50, by=5)
```

```
fgyearplayer100 <- fgyearplayer %>% filter(Player %in% yearplayer100$Player)
plotYearPlayer <- ggplot() +
  geom_line(data=fgyearplayer100, aes(x=Year, y=pctfg3, colour=Player), size=1, show.legend = FALSE) +
  geom_line(data=fgyearplayer100, aes(x=Year, y=pctfg2, colour=Player), size=1, show.legend = FALSE) +
  xlab('Year') +
  ylab(NULL) +
  ggtitle('3 point shot success rate by player') +
  theme(panel.background=element_rect(fill=NA), panel.grid.major.y=element_line(color="grey", linetype=
        plot.title = element_text(hjust = 0.5)) +
  scale_y_continuous(limits=c(0, 100), breaks=yaxisbreaks, labels=yaxisbreaks) +
  scale_x_continuous(limits=c(1985,2019), breaks=xaxisbreaks)
plotYearPlayer
```



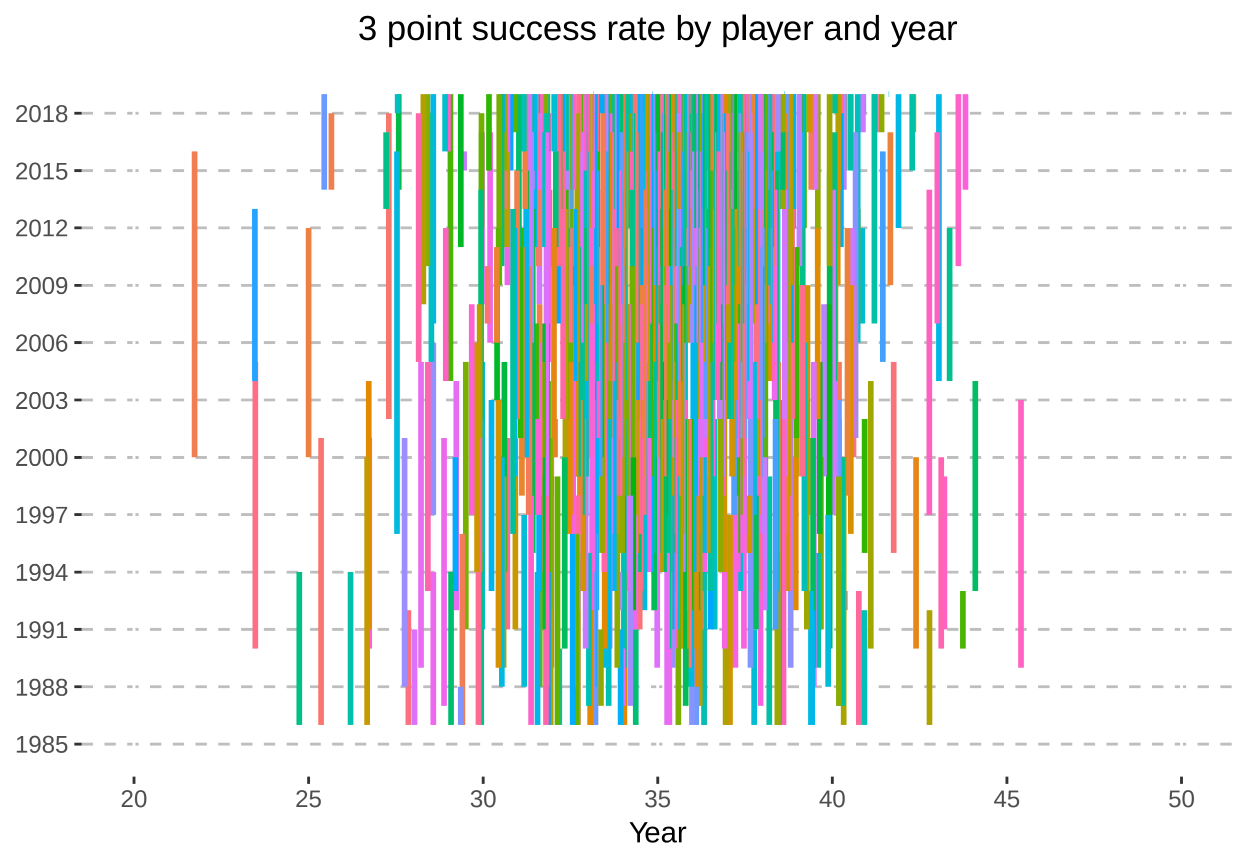

```
# Meaningless...

fgplayer <- aggregate(dataGameLogsPlayer1986[, 19:26], list(dataGameLogsPlayer1986$namePlayer), sum)
colnames(fgplayer)[1] <- "Player"
fgplayer$pctFG = NULL
fgplayer$pctFG3 = NULL

fgplayer$pctfg3 <- fgplayer$fg3m / fgplayer$fg3a * 100
fgplayer$pctfg2 <- fgplayer$fgm / fgplayer$fga * 100
fgplayer$pctft <- fgplayer$ftm / fgplayer$fta * 100

fgplayer <- fgplayer[order(-fgplayer$pctfg3),]
```

```r
fgplayer100 <- fgplayer %>% filter(fg3m >= 100)
```

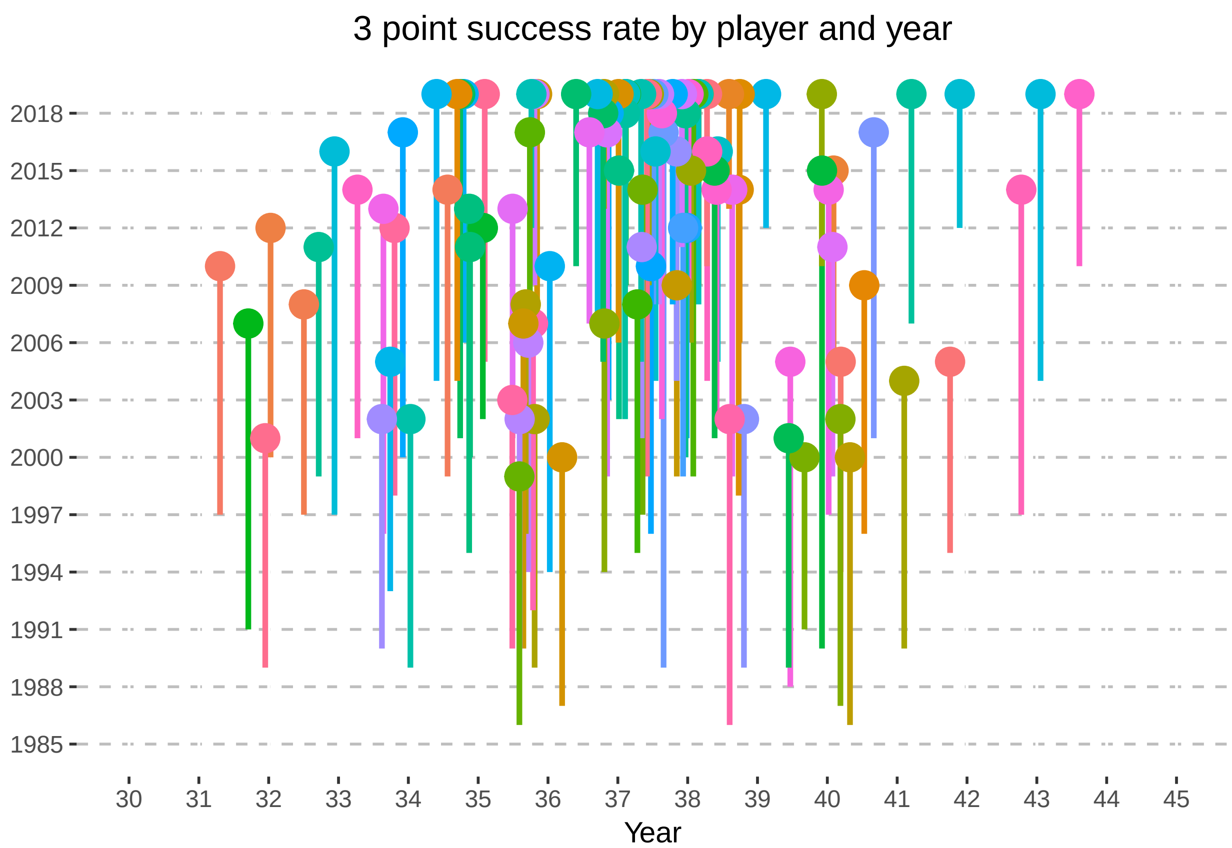
```python
import pandas as pd

fgplayer = r.fgplayer

fgplayer['firstYear'] = 2019
fgplayer['lastYear'] = 1986

print(fgplayer.head(5))
          Player     fgm     fga    ...         pctft  firstYear  lastYear
0     Alvin Sims     4.0    10.0    ...     40.000000       2019      1986
1    Coty Clarke     2.0     4.0    ...           NaN       2019      1986
2     David Pope     9.0    19.0    ...     50.000000       2019      1986
3    Eddy Curry  2578.0  4734.0    ...     64.219474       2019      1986
4  Eric Anderson    12.0    35.0    ...     59.259259       2019      1986

[5 rows x 12 columns]
print(fgplayer.tail(5))
                Player    fgm    fga    ...         pctft  firstYear  lastYear
2720    Winston Crite   34.0   71.0    ...     76.000000       2019      1986
2721       Yinka Dare   86.0  217.0    ...     57.009346       2019      1986
2722      Yvon Joseph    0.0    0.0    ...    100.000000       2019      1986
2723   Zeljko Rebraca  488.0  926.0    ...     79.155673       2019      1986
2724  Zendon Hamilton  176.0  400.0    ...     66.005666       2019      1986

[5 rows x 12 columns]
i=0

for player in fgplayer.values:
  min = player[-2]
  max = player[-1]
  for yp in r.fgyearplayer.values:
    if player[0] == yp[1]:
      if max < yp[0]: max = yp[0]
      if min > yp[0]: min = yp[0]
  fgplayer.iloc[i,-1]=max
  fgplayer.iloc[i,-2]=min
  i += 1

print(fgplayer.head(5))
          Player     fgm     fga    ...         pctft  firstYear  lastYear
0     Alvin Sims     4.0    10.0    ...     40.000000       1999      1999
1    Coty Clarke     2.0     4.0    ...           NaN       2016      2016
2     David Pope     9.0    19.0    ...     50.000000       1986      1986
3    Eddy Curry  2578.0  4734.0    ...     64.219474       2002      2013
4  Eric Anderson    12.0    35.0    ...     59.259259       1993      1994

[5 rows x 12 columns]
print(fgplayer.tail(5))
                Player    fgm    fga    ...         pctft  firstYear  lastYear
2720    Winston Crite   34.0   71.0    ...     76.000000       1988      1989
2721       Yinka Dare   86.0  217.0    ...     57.009346       1995      1998
```

```
2722       Yvon Joseph    0.0    0.0    ...    100.000000       1986      1986
2723    Zeljko Rebraca  488.0  926.0    ...     79.155673       2002      2006
2724   Zendon Hamilton  176.0  400.0    ...     66.005666       2001      2006

[5 rows x 12 columns]
```

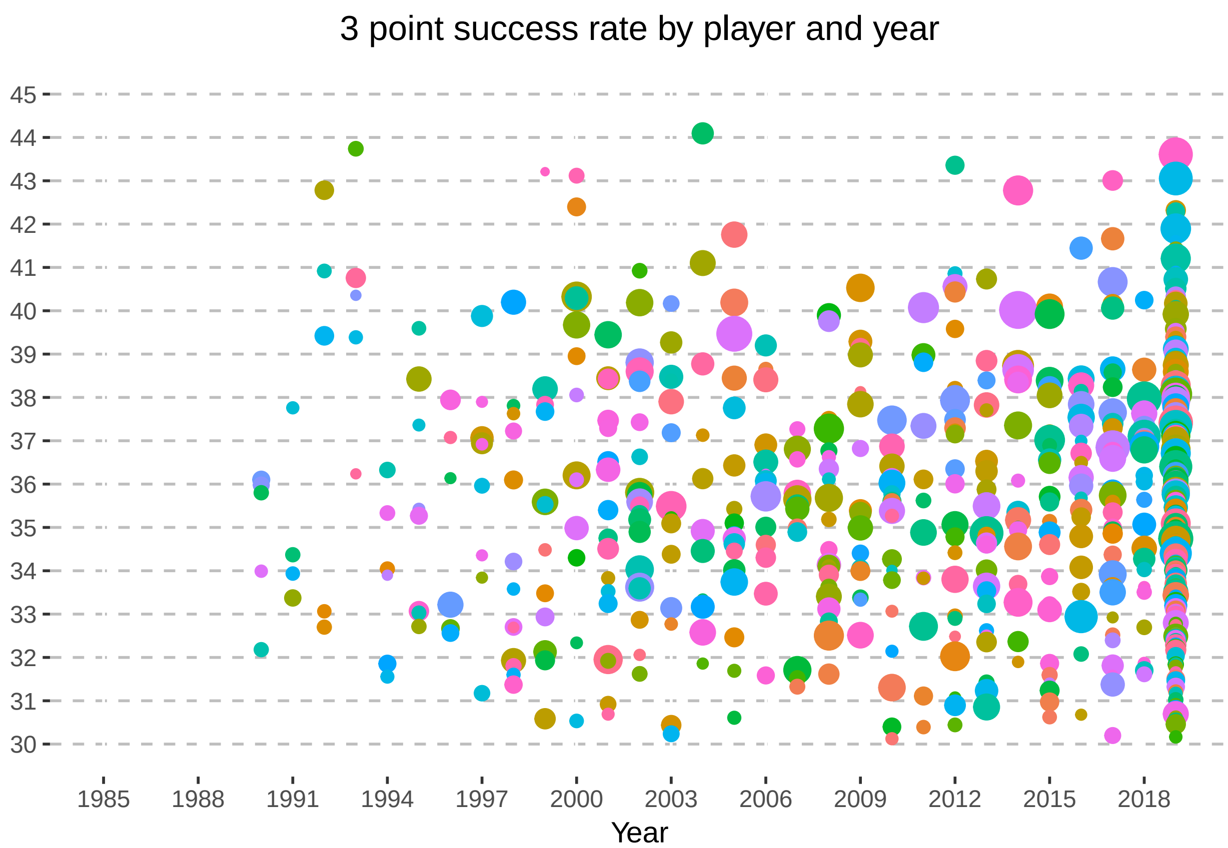
```
fgplayer <- py$fgplayer
fgplayer100 <- fgplayer %>% filter(fg3m >= 100)
fgplayer1000 <- fgplayer100 %>% filter(fg3m >= 1000)
fgplayer2000 <- fgplayer1000 %>% filter(fg3m >= 2000)


xaxisbreaks <- seq(1985, 2019, by=3)
yaxisbreaks <- seq(20, 50, by=5)

plotPlayer100 <- ggplot() +
  geom_linerange(data=fgplayer100, aes(x=pctfg3, y=lastYear, ymin=firstYear, ymax=lastYear, colour=Playe
  # geom_point(data=fgplayer100, aes(x=lastYear, y=pctfg3, colour=Player), size=1, show.legend = FALSE)
  # geom_line
  xlab('Year') +
  ylab(NULL) +
  ggtitle('3 point success rate by player and year') +
  theme(panel.background=element_rect(fill=NA), panel.grid.major.y=element_line(color="grey", linetype=
        plot.title = element_text(hjust = 0.5)) +
  scale_x_continuous(limits=c(20, 50), breaks=yaxisbreaks, labels=yaxisbreaks) +
  scale_y_continuous(limits=c(1985,2019), breaks=xaxisbreaks)
plotPlayer100
```



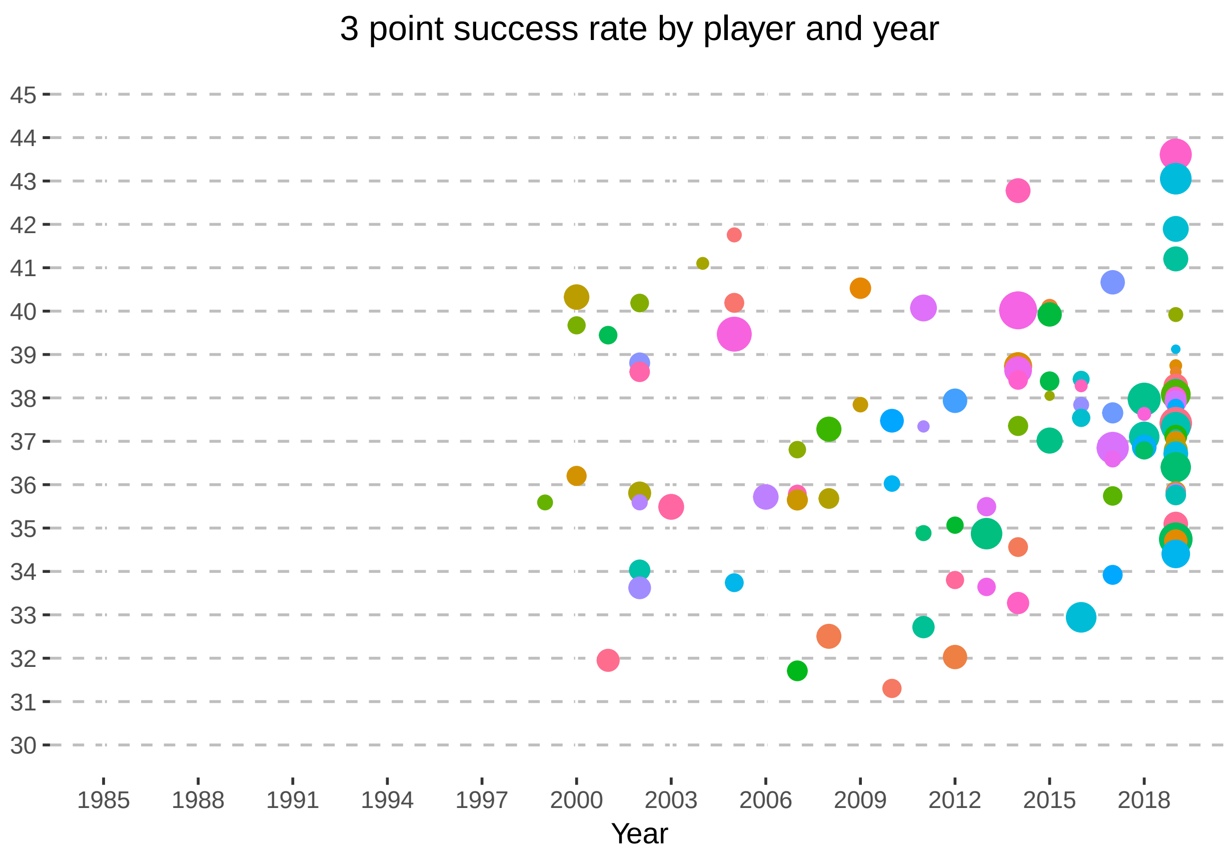

```
xaxisbreaks <- seq(1985, 2019, by=3)
yaxisbreaks <- seq(30, 45, by=1)

fgplayer1000 <- fgplayer100 %>% filter (fg3m >= 1000)
plotPlayer1000 <- ggplot() +
  geom_pointrange(data=fgplayer1000, aes(x=pctfg3, y=lastYear, ymin=firstYear, ymax=lastYear, colour=Pla
  # geom_point(data=fgplayer100, aes(x=lastYear, y=pctfg3, colour=Player), size=1, show.legend = FALSE)
  # geom_line
  xlab('Year') +
  ylab(NULL) +
  ggtitle('3 point success rate by player and year') +
  theme(panel.background=element_rect(fill=NA), panel.grid.major.y=element_line(color="grey", linetype=2
        plot.title = element_text(hjust = 0.5)) +
  scale_x_continuous(limits=c(30, 45), breaks=yaxisbreaks, labels=yaxisbreaks) +
  scale_y_continuous(limits=c(1985,2019), breaks=xaxisbreaks)
plotPlayer1000
```



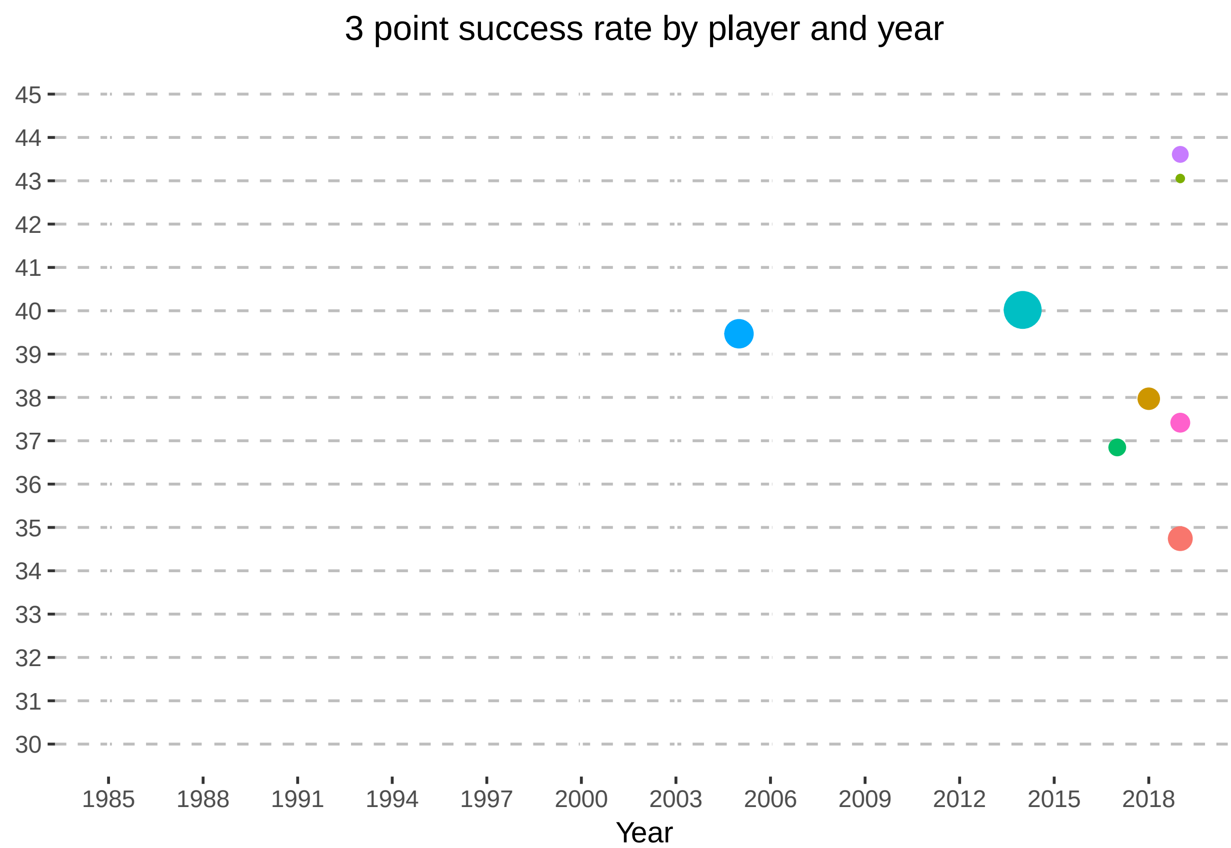

```
xaxisbreaks <- seq(1985, 2019, by=3)
yaxisbreaks <- seq(30, 45, by=1)

plotPlayer100 <- ggplot() +
  geom_point(data=fgplayer100, aes(x=lastYear, y=pctfg3, size=fg3m+fg3a, colour=Player), show.legend = F
  # geom_point(data=fgplayer100, aes(x=lastYear, y=pctfg3, colour=Player), size=1, show.legend = FALSE)
  # geom_line
  xlab('Year') +
  ylab(NULL) +
  ggtitle('3 point success rate by player and year') +
  theme(panel.background=element_rect(fill=NA), panel.grid.major.y=element_line(color="grey", linetype=2
        plot.title = element_text(hjust = 0.5)) +
  scale_y_continuous(limits=c(30, 45), breaks=yaxisbreaks, labels=yaxisbreaks) +
  scale_x_continuous(limits=c(1985,2019), breaks=xaxisbreaks)
plotPlayer100
```




```
plotPlayer1000 <- ggplot() +
  geom_point(data=fgplayer1000, aes(x=lastYear, y=pctfg3, size=fg3m+fg3a, colour=Player), show.legend =
  # geom_point(data=fgplayer100, aes(x=lastYear, y=pctfg3, colour=Player), size=1, show.legend = FALSE)
  # geom_line
  xlab('Year') +
  ylab(NULL) +
  ggtitle('3 point success rate by player and year') +
  theme(panel.background=element_rect(fill=NA), panel.grid.major.y=element_line(color="grey", linetype=2
        plot.title = element_text(hjust = 0.5)) +
  scale_y_continuous(limits=c(30, 45), breaks=yaxisbreaks, labels=yaxisbreaks) +
  scale_x_continuous(limits=c(1985,2019), breaks=xaxisbreaks)
plotPlayer1000
```




```r
plotPlayer2000 <- ggplot() +
  geom_point(data=fgplayer2000, aes(x=lastYear, y=pctfg3, size=fg3m+fg3a, colour=Player), show.legend =
  # geom_point(data=fgplayer100, aes(x=lastYear, y=pctfg3, colour=Player), size=1, show.legend = FALSE)
  # geom_line
  xlab('Year') +
  ylab(NULL) +
  ggtitle('3 point success rate by player and year') +
  theme(panel.background=element_rect(fill=NA), panel.grid.major.y=element_line(color="grey", linetype=
        plot.title = element_text(hjust = 0.5)) +
  scale_y_continuous(limits=c(30, 45), breaks=yaxisbreaks, labels=yaxisbreaks) +
  scale_x_continuous(limits=c(1985,2019), breaks=xaxisbreaks)
plotPlayer2000
```



Above graph shows more players are trying 3 point shots than before. even though the average success rate is similar.

Q4. Players who are good at 3-pointers are also good at 2-pointers or free throws?

By regression.

Players who are good at free throws tend to be good at 3-pointers. However, 2-point field goal success rate is not related with 3-point field goal success rate!!! Why?

```
linearModel <- lm(pctfg3 ~ pctfg2, data=fgplayer100)
tidy(linearModel)
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)  33.7       1.75      19.2   2.81e-67
2 pctfg2        0.0330    0.0400     0.823 4.11e- 1

linearModel2 <- lm(fg3m ~ fgm, data=fgplayer100)
tidy(linearModel2)
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)  184.      19.6         9.41 6.19e-20
2 fgm            0.143    0.00618    23.1  2.24e-89

linearModel3 <- lm(fg3a ~ fga, data=fgplayer100)
tidy(linearModel3)
```

```
# A tibble: 2 x 5
  term        estimate std.error statistic   p.value
  <chr>          <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)  404.     48.0          8.42 1.98e- 16
2 fga            0.197   0.00687     28.6  3.67e-122

linearModel4 <- lm(fg3a ~ fga + fta, data=fgplayer100)
tidy(linearModel4)
# A tibble: 3 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)  276.      47.4         5.82 8.67e- 9
2 fga            0.347    0.0172     20.2  7.38e-73
3 fta           -0.455    0.0481     -9.47 3.52e-20

linearModel5 <- lm(pctfg3 ~ pctft, data=fgplayer100)
tidy(linearModel5)
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)   18.2      1.42        12.8 3.40e-34
2 pctft          0.216    0.0181      11.9 4.54e-30

linearModel6 <- lm(pctfg2 ~ pctft, data=fgplayer100)
tidy(linearModel6)
# A tibble: 2 x 5
  term        estimate std.error statistic   p.value
  <chr>          <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)  41.9       1.42       29.6  4.07e-128
2 pctft         0.0219    0.0180      1.21 2.25e-  1

linearModel7 <- lm(pctfg3 ~ pctfg2 + pctft, data=fgplayer100)
tidy(linearModel7)
# A tibble: 3 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)  17.7       2.10       8.42  1.86e-16
2 pctfg2        0.0136    0.0368     0.370 7.12e- 1
3 pctft         0.216     0.0182    11.9   6.51e-30
```

When we look at all the players, 2-pointers and 3-pointers are reverse-related. Maybe because of dunk shots?

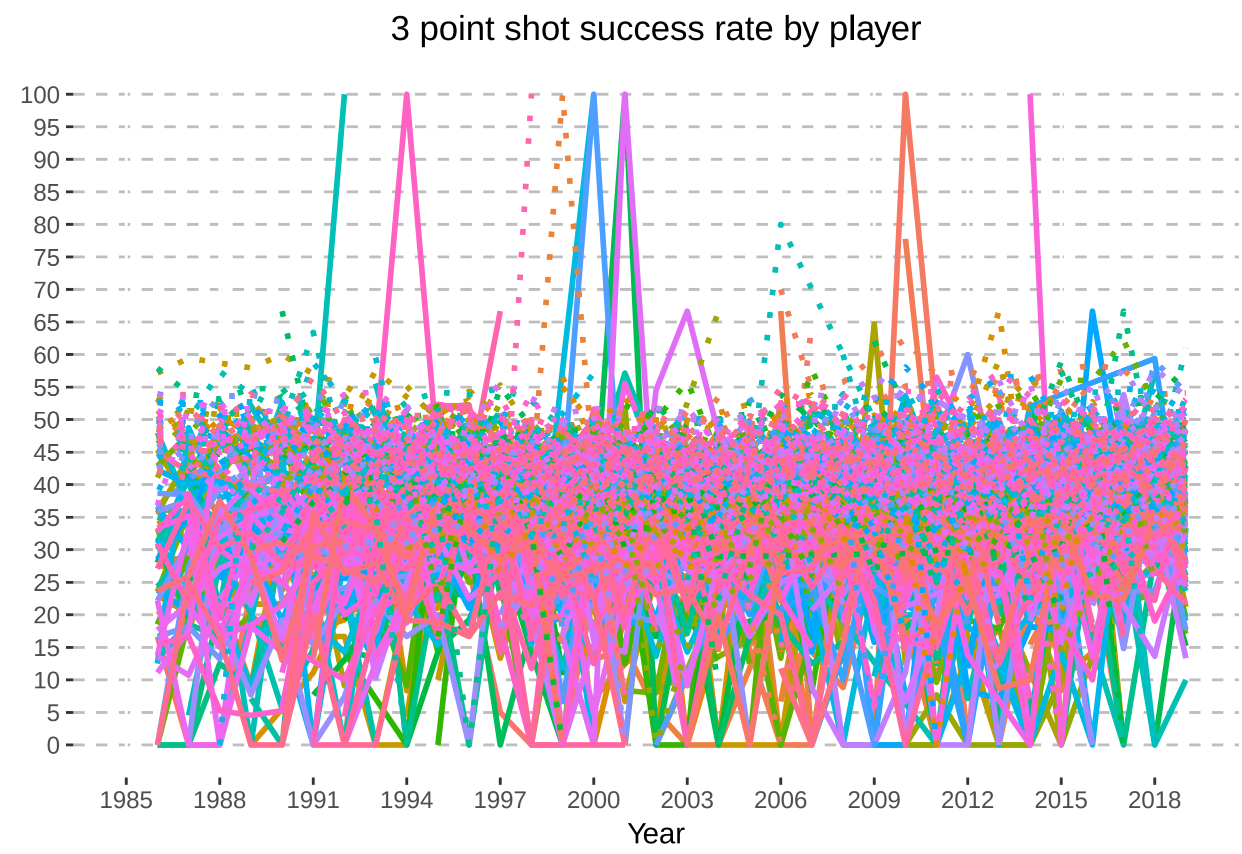
```
linearModel7 <- lm(pctfg3 ~ pctfg2 + pctft, data=fgplayer)
tidy(linearModel7)
# A tibble: 3 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)   3.65      2.52        1.45 1.48e- 1
2 pctfg2       -0.0441    0.0415     -1.06 2.88e- 1
3 pctft         0.329     0.0237     13.9  3.19e-42
```

Best players (more than 1,000 career 3-point field goals) are good at 2-pointers as well!!!

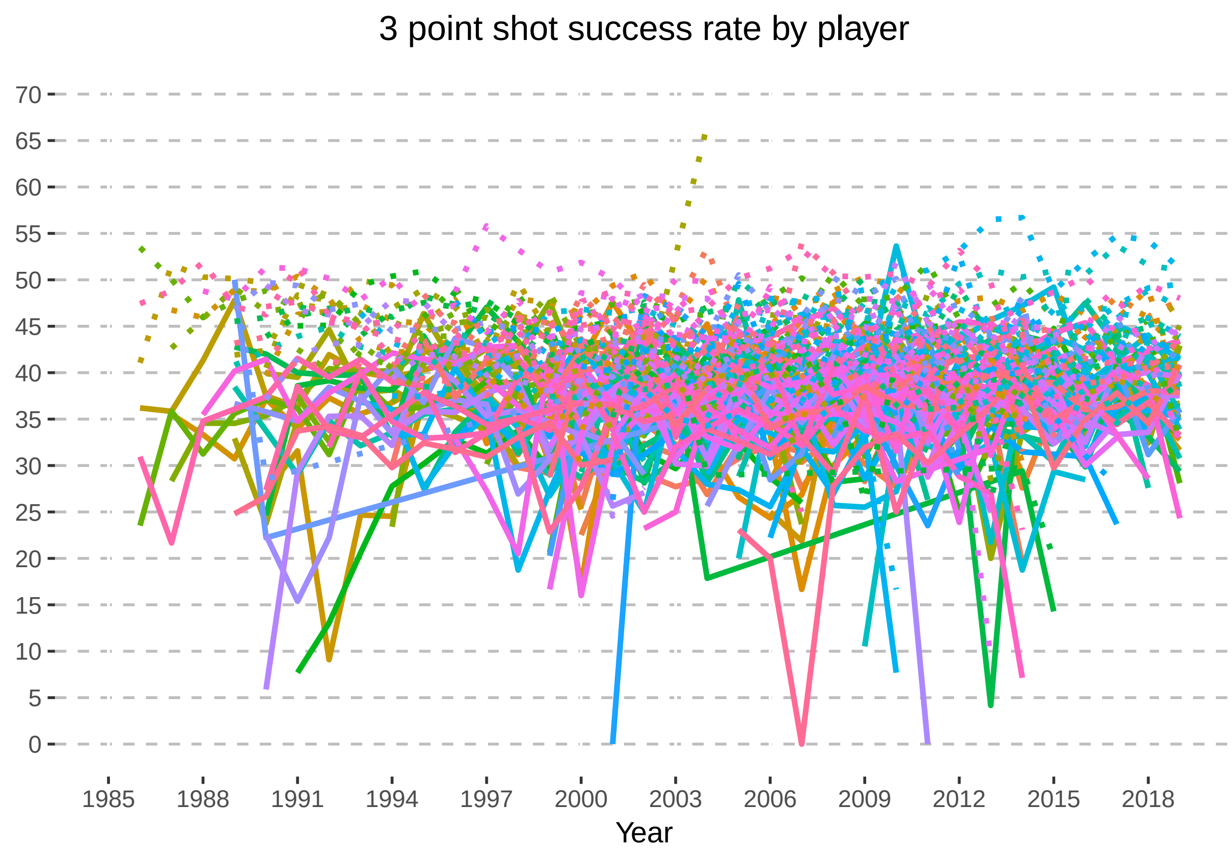
```r
linearModel7 <- lm(pctfg3 ~ pctfg2 + pctft, data=fgplayer1000)
tidy(linearModel7)
# A tibble: 3 x 5
  term        estimate std.error statistic       p.value
  <chr>          <dbl>     <dbl>     <dbl>         <dbl>
1 (Intercept)    3.76     4.06       0.926 0.356
2 pctfg2         0.345    0.0843     4.09  0.0000841
3 pctft          0.226    0.0344     6.58  0.00000000197

linearModel8 <- lm(pctfg3 ~ pctfg2 + pctft, data=fgplayer2000)
tidy(linearModel8)
# A tibble: 3 x 5
  term        estimate std.error statistic p.value
  <chr>          <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)  -21.5      20.1       -1.07   0.334
2 pctfg2         0.799     0.442      1.81   0.131
3 pctft          0.290     0.231      1.26   0.264
```

-. Are there any relationship between players' ages and 3-pointers? Both total and average.

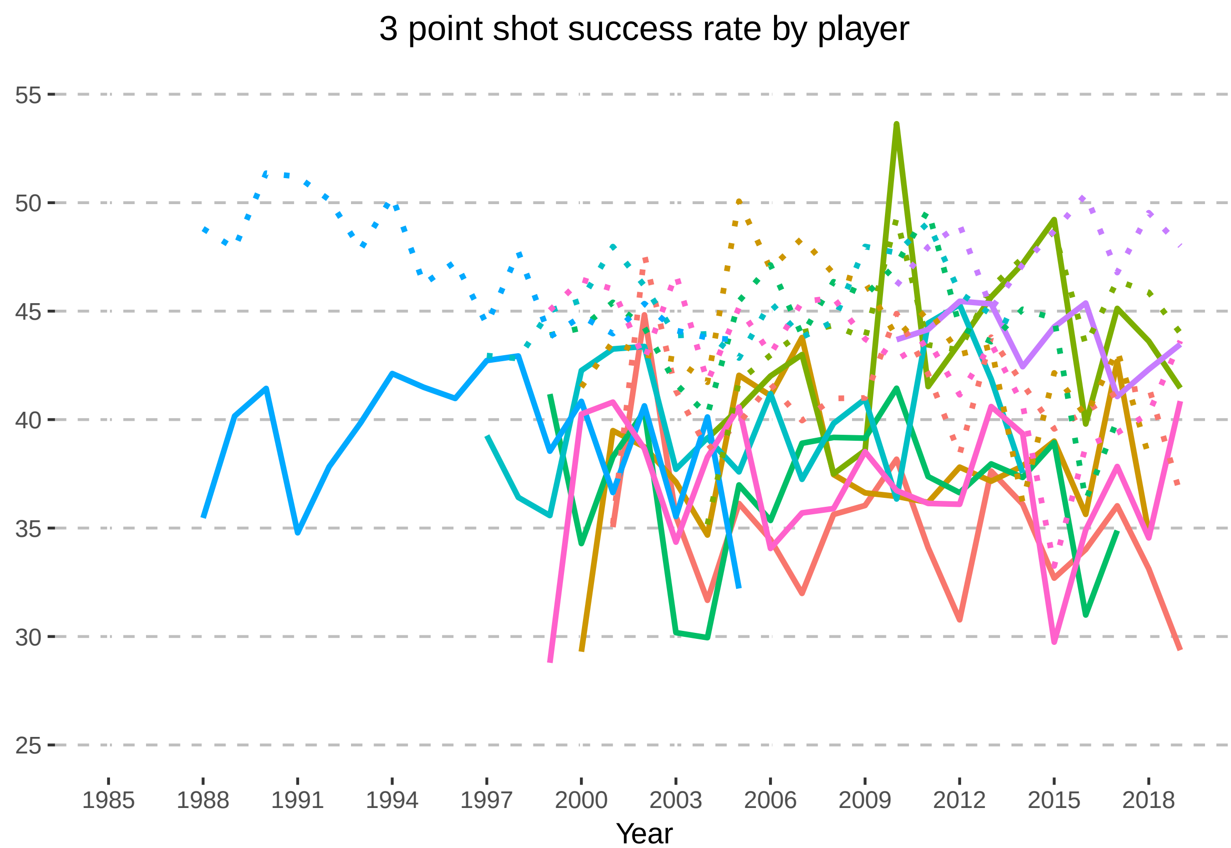
```r
fgyearplayer100 <- fgyearplayer %>% filter(Player %in% fgplayer100$Player)
fgyearplayer1000 <- fgyearplayer100 %>% filter(Player %in% fgplayer1000$Player)
fgyearplayer2000 <- fgyearplayer1000 %>% filter(Player %in% fgplayer2000$Player)

xaxisbreaks <- seq(1985, 2019, by=3)
yaxisbreaks <- seq(0, 100, by=5)

plotYearPlayer100 <- ggplot() +
  geom_line(data=fgyearplayer100, aes(x=Year, y=pctfg3, colour=Player), size=1, show.legend = FALSE) +
  geom_line(data=fgyearplayer100, aes(x=Year, y=pctfg2, colour=Player), size=1, linetype="dotted", show
  xlab('Year') +
  ylab(NULL) +
  ggtitle('3 point shot success rate by player') +
  theme(panel.background=element_rect(fill=NA), panel.grid.major.y=element_line(color="grey", linetype=
        plot.title = element_text(hjust = 0.5)) +
  scale_y_continuous(limits=c(0, 100), breaks=yaxisbreaks, labels=yaxisbreaks) +
  scale_x_continuous(limits=c(1985,2019), breaks=xaxisbreaks)
plotYearPlayer100
```




```
xaxisbreaks <- seq(1985, 2019, by=3)
yaxisbreaks <- seq(0, 70, by=5)

plotYearPlayer1000 <- ggplot() +
  geom_line(data=fgyearplayer1000, aes(x=Year, y=pctfg3, colour=Player), size=1, show.legend = FALSE) +
  geom_line(data=fgyearplayer1000, aes(x=Year, y=pctfg2, colour=Player), size=1, linetype="dotted", sho
  xlab('Year') +
  ylab(NULL) +
  ggtitle('3 point shot success rate by player') +
  theme(panel.background=element_rect(fill=NA), panel.grid.major.y=element_line(color="grey", linetype=
        plot.title = element_text(hjust = 0.5)) +
  scale_y_continuous(limits=c(0, 70), breaks=yaxisbreaks, labels=yaxisbreaks) +
  scale_x_continuous(limits=c(1985,2019), breaks=xaxisbreaks)
plotYearPlayer1000
```




```
xaxisbreaks <- seq(1985, 2019, by=3)
yaxisbreaks <- seq(25, 55, by=5)

plotYearPlayer2000 <- ggplot() +
  geom_line(data=fgyearplayer2000, aes(x=Year, y=pctfg3, colour=Player), size=1, show.legend = FALSE) +
  geom_line(data=fgyearplayer2000, aes(x=Year, y=pctfg2, colour=Player), size=1, linetype="dotted", sho
  xlab('Year') +
  ylab(NULL) +
  ggtitle('3 point shot success rate by player') +
  theme(panel.background=element_rect(fill=NA), panel.grid.major.y=element_line(color="grey", linetype=
        plot.title = element_text(hjust = 0.5)) +
  scale_y_continuous(limits=c(25, 55), breaks=yaxisbreaks, labels=yaxisbreaks) +
  scale_x_continuous(limits=c(1985,2019), breaks=xaxisbreaks)
plotYearPlayer2000
```



Let's regress.

```
fgyearplayerjoined <- left_join(fgyearplayer, fgplayer, by=c("Player" = "Player"))
fgyearplayerjoined$career = fgyearplayerjoined$Year - fgyearplayerjoined$firstYear + 1

fgyearplayerjoined100 <- fgyearplayerjoined %>% filter(Player %in% fgplayer100$Player)
fgyearplayerjoined1000 <- fgyearplayerjoined100 %>% filter(Player %in% fgplayer1000$Player)
fgyearplayerjoined2000 <- fgyearplayerjoined1000 %>% filter(Player %in% fgplayer2000$Player)

linearModel <- lm(pctfg3.x ~ career, data=fgyearplayerjoined2000)
tidy(linearModel)
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)  39.6       0.720      55.0  1.01e-95
2 career       -0.0994    0.0656     -1.51 1.32e- 1
linearModel2 <- lm(pctfg3.x ~ career, data=fgyearplayerjoined1000)
tidy(linearModel2)
# A tibble: 2 x 5
  term        estimate std.error statistic p.value
  <chr>          <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)  35.4       0.281     126.     0
2 career        0.0730    0.0306      2.38  0.0173
linearModel3 <- lm(pctfg3.x ~ career, data=fgyearplayerjoined100)
tidy(linearModel3)
# A tibble: 2 x 5
```

```
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)   31.7      0.208     153.   0.
2 career         0.186    0.0280      6.63 3.63e-11
linearModel4 <- lm(pctfg3.x ~ career, data=fgyearplayerjoined)
tidy(linearModel4)
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)   24.1      0.252       95.5 0.
2 career         0.414    0.0378      11.0 7.90e-28
```

Really good players are not related with ages/career. Average players' success rate is increased by 0.4% in one year. Not bad…?

- Players with high salaries are good at 3-pointers?

2018-2019 season data only

```
nbaInsiderSalaries <- nba_insider_salaries(assume_player_opt_out = T, assume_team_doesnt_exercise = T, 
You got salary data for the Atlanta Hawks
You got salary data for the Boston Celtics
You got salary data for the Brooklyn Nets
You got salary data for the Charlotte Hornets
You got salary data for the Chicago Bulls
You got salary data for the Cleveland Cavaliers
You got salary data for the Dallas Mavericks
You got salary data for the Denver Nuggets
You got salary data for the Detroit Pistons
You got salary data for the Golden State Warriors
You got salary data for the Houston Rockets
You got salary data for the Indiana Pacers
You got salary data for the Los Angeles Clippers
You got salary data for the Los Angeles Lakers
You got salary data for the Memphis Grizzlies
You got salary data for the Miami Heat
You got salary data for the Milwaukee Bucks
You got salary data for the Minnesota Timberwolves
You got salary data for the New Orleans Pelicans
You got salary data for the New York Knicks
You got salary data for the Oklahoma City Thunder
You got salary data for the Orlando Magic
You got salary data for the Philadelphia 76ers
You got salary data for the Phoenix Suns
You got salary data for the Portland Trail Blazers
You got salary data for the Sacramento Kings
You got salary data for the San Antonio Spurs
You got salary data for the Toronto Raptors
You got salary data for the Utah Jazz
You got salary data for the Washington Wizards

fgplayersalary <- left_join(fgplayer, nbaInsiderSalaries, by=c("Player"="namePlayer"))
```

```
fgplayersalary2 <- na.omit(fgplayersalary)
fgplayersalary2$salaryinK = fgplayersalary2$value / 1000
fgplayersalary2$salaryinM = fgplayersalary2$value / 1000000

linearModel <- lm(pctfg3 ~ salaryinM, data=fgplayersalary2)
tidy(linearModel)
# A tibble: 2 x 5
  term        estimate std.error statistic p.value
  <chr>          <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)  29.7       0.450      65.9  0
2 salaryinM     0.0931    0.0343      2.72 0.00671
linearModel2 <- lm(fg3m ~ salaryinM, data=fgplayersalary2)
tidy(linearModel2)
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)     94.5     14.7       6.42 2.10e-10
2 salaryinM       23.1      1.12     20.6  1.38e-79
```

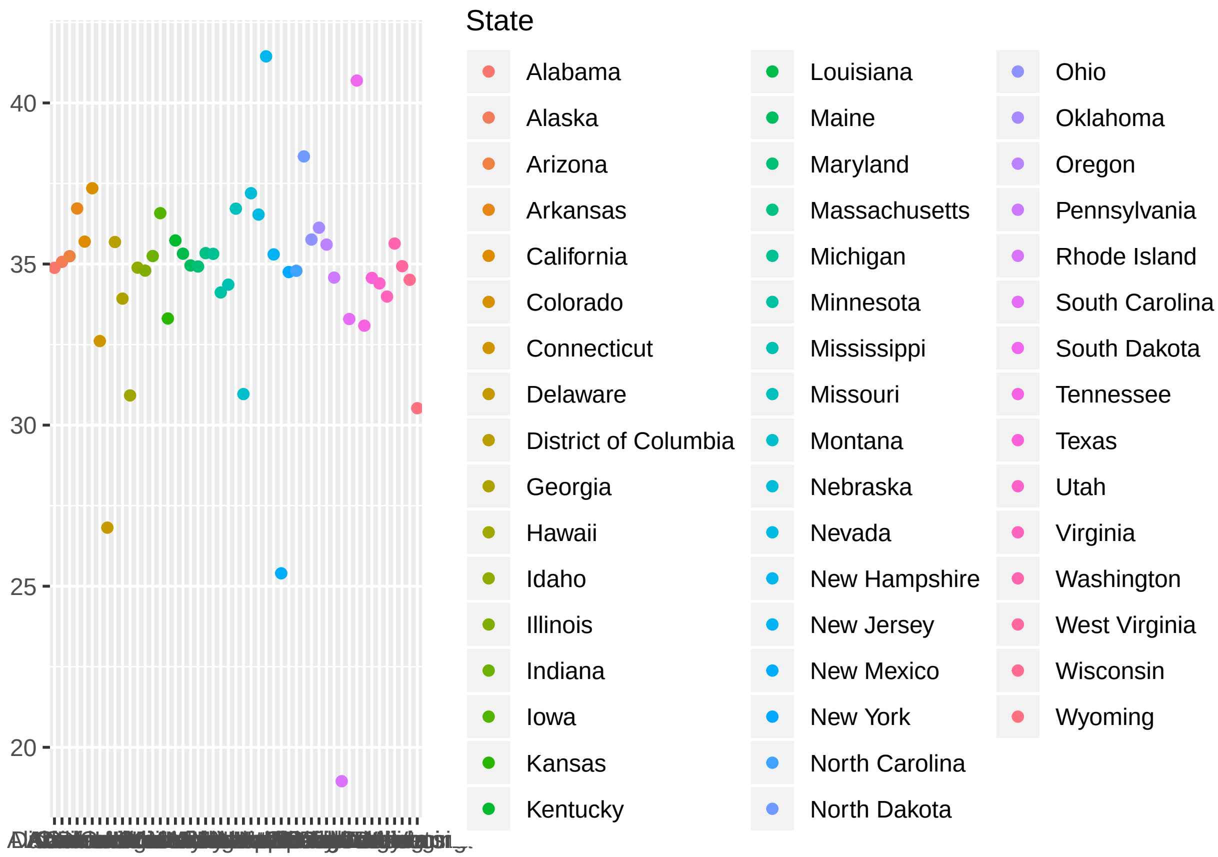
When the salary increases by a million dollar, career success rate of 3-point shots increases by 0.09% only. It's difficult to say that 3-pointer success rate is the most important factor for one's salary.

- We would like to explore the importance of three point shooters in a given team by measuring the share of the team's total salary over time.
- We want to analyze whether players can drastically improve their three point shooting skills over time or the skill is rather something people are borned with.

There is no dramatic increase in 3-pointer success rate. Maybe if we can check the players' data from NCAA or high school league, there might be different insight. However, based on NBA data, no big changes.

- Show the 3-pointer statistics geographically based on players' hometowns. Maybe this help illustrates the different basketball playing style across different regions, both domestic and international.

```
playerHometown <- read_csv("PlayerHometown.csv")

fgplayerhometown <- left_join(fgplayer, playerHometown, by=c("Player"="Player"))
fgplayerhometown <- fgplayerhometown %>% filter(not(is.na(State)))
fgplayerhometown <- na.omit(fgplayerhometown)

fgplayerhometownState <- aggregate(fgplayerhometown[, 2:7], list(fgplayerhometown$State), sum)
colnames(fgplayerhometownState)[1] <- "State"
fgplayerhometownState$pctfg3 <- fgplayerhometownState$fg3m / fgplayerhometownState$fg3a * 100
fgplayerhometownState$pctfg2 <- fgplayerhometownState$fgm / fgplayerhometownState$fga * 100
fgplayerhometownState$pctft <- fgplayerhometownState$ftm / fgplayerhometownState$fta * 100

plotState <- ggplot() +
  geom_point(data=fgplayerhometownState, aes(x=State, y=pctfg3, colour=State)) +
  xlab(NULL) +
  ylab(NULL)
plotState
```


40
35
30
25
20
State
Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
District of Columbia
Georgia
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
Missouri
Montana
Nebraska
Nevada
New Hampshire
New Jersey
New Mexico
New York
North Carolina
North Dakota
Ohio
Oklahoma
Oregon
Pennsylvania
Rhode Island
South Carolina
South Dakota
Tennessee
Texas
Utah
Virginia
Washington
West Virginia
Wisconsin
Wyoming