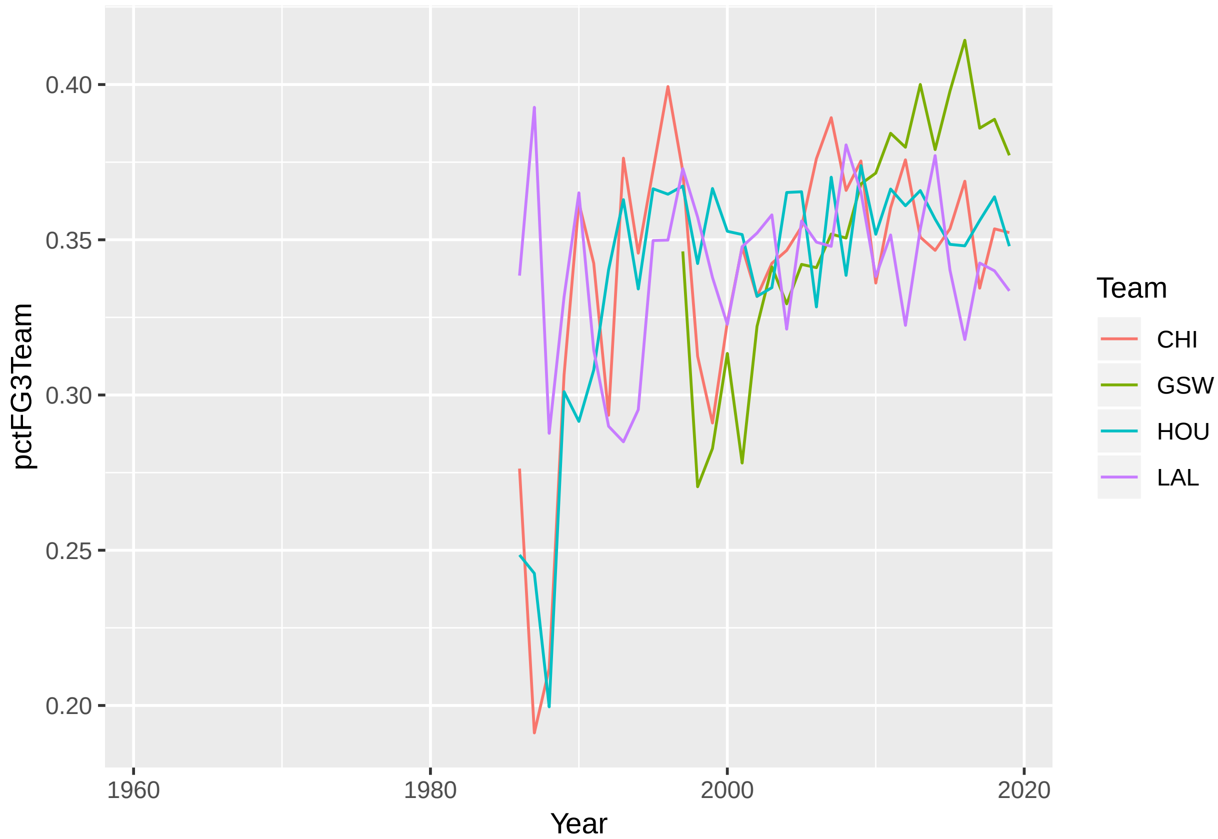
Problem Set 6

Subeom Lee

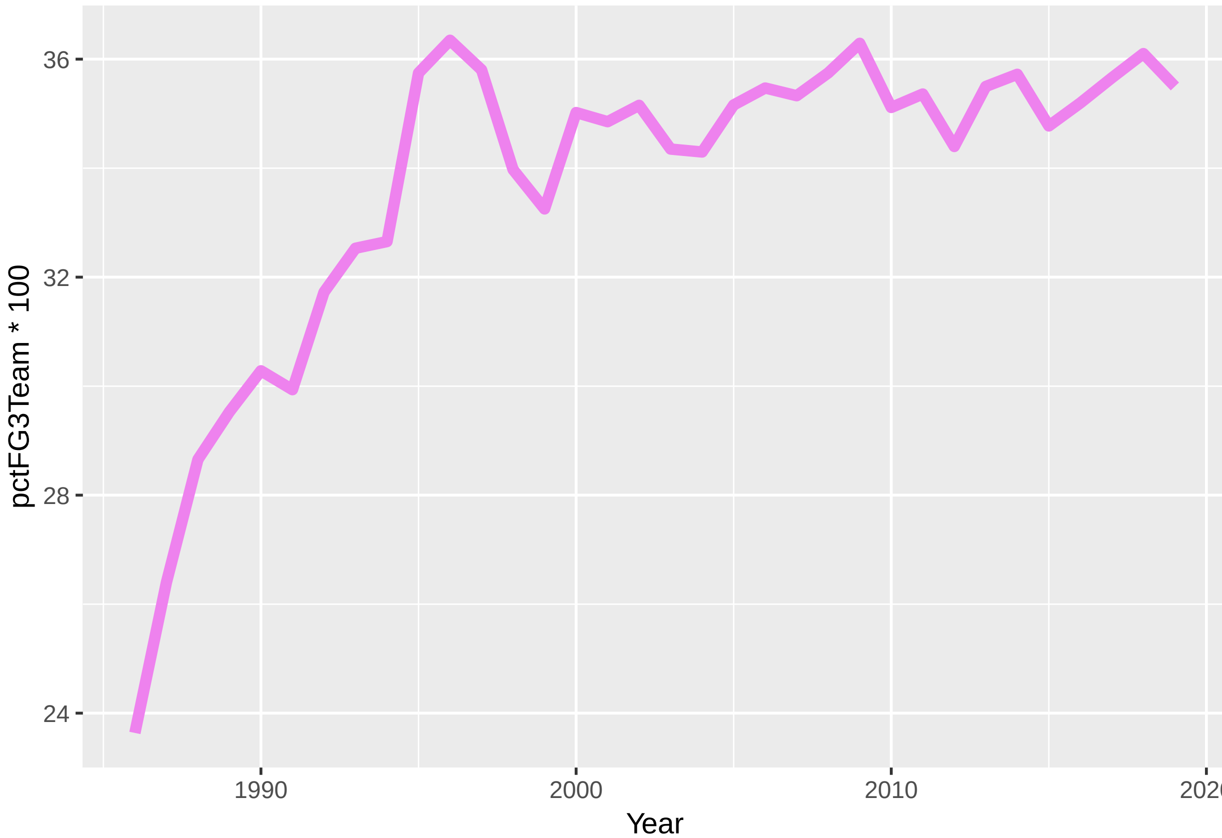
2019-03-09

Questions

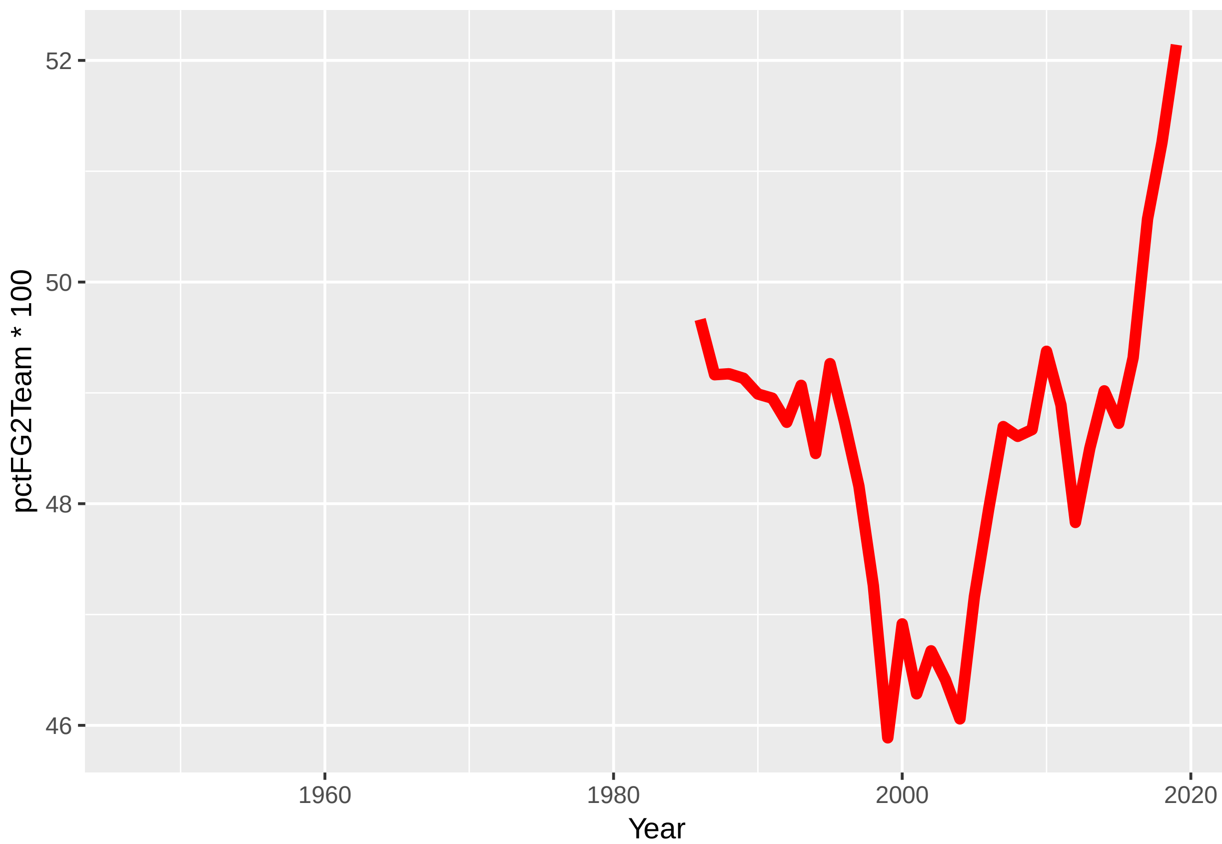
```r
#http://asbcllc.com/nbastatR/index.html

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

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')



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
```



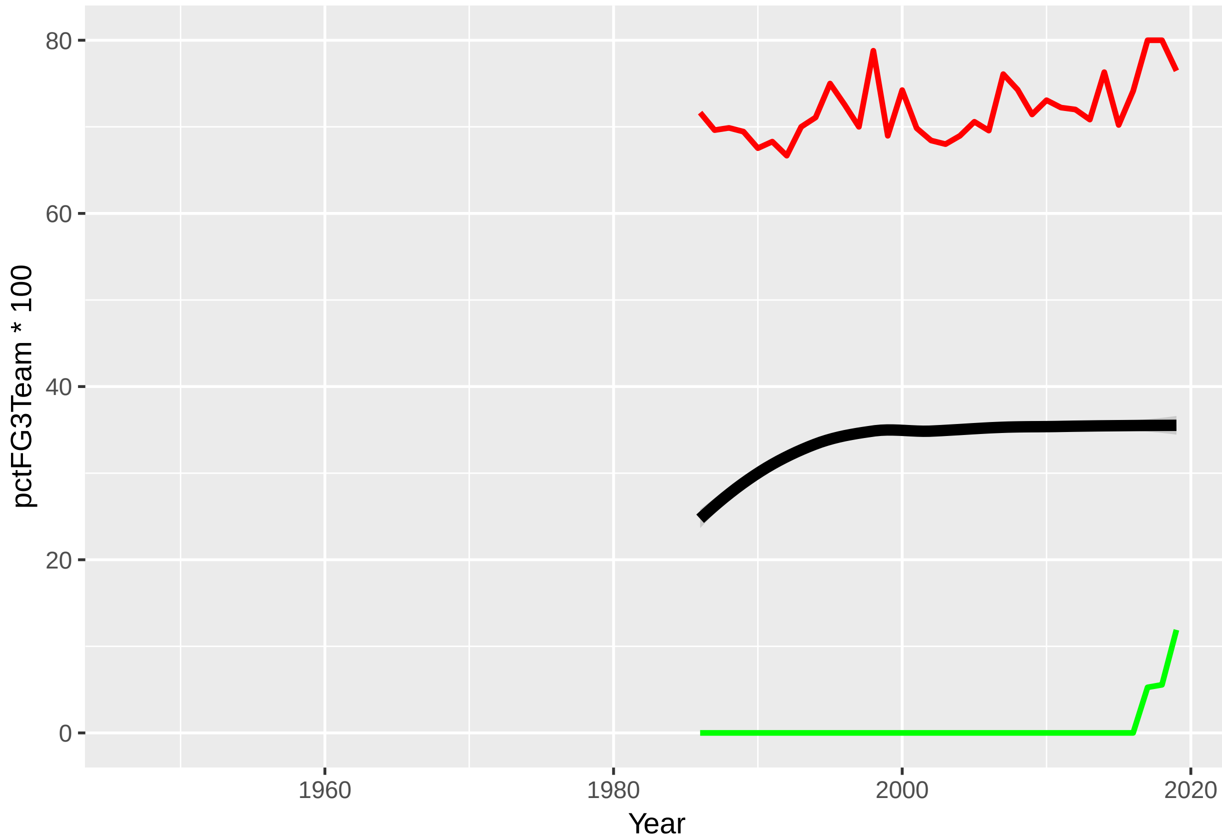

```
avg2 <- aggregate(dataGameLogsTeam[, 24:46], list(dataGameLogsTeam$yearSeason), mean)
colnames(avg2)[1] <- "Year"
min2 <- aggregate(dataGameLogsTeam[, 24:46], list(dataGameLogsTeam$yearSeason), min)
colnames(min2)[1] <- "Year"
max2 <- aggregate(dataGameLogsTeam[, 24:46], list(dataGameLogsTeam$yearSeason), max)
colnames(max2)[1] <- "Year"

avgplot2 <- avg2 %>%
              filter(Year >= 1986) %>%
              ggplot(aes(x=Year, y=pctFG3Team*100)) +
              geom_path(colour='violet', size=2)
avgplot2
```



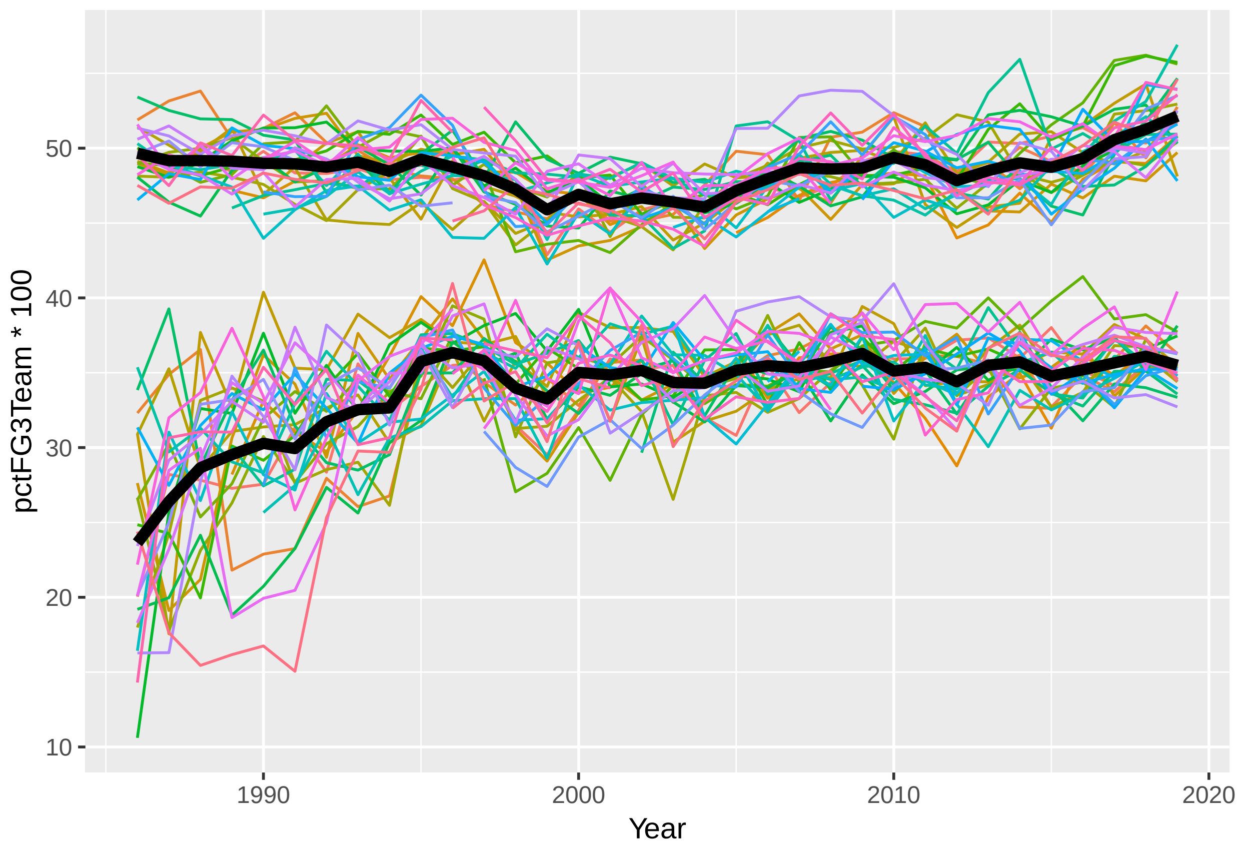

```
avgplot3 <- avg2 %>%
            ggplot(aes(x=Year, y=pctFG2Team*100)) +
            geom_path(colour='red', size=2)
avgplot3
```



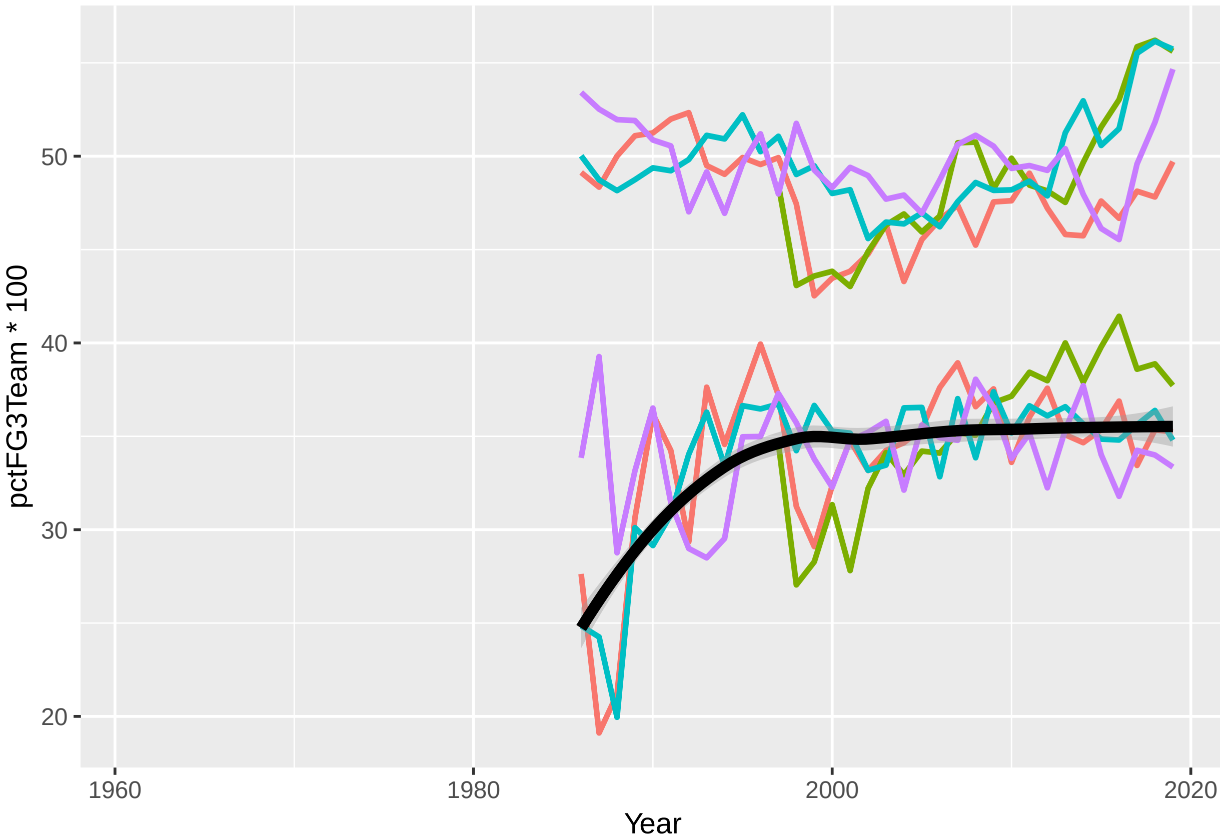

```
avgminmax <- ggplot() +
               geom_line(data=min2, aes(x=Year, y=pctFG3Team*100), size=1, colour='green') +
               geom_line(data=max2, aes(x=Year, y=pctFG2Team*100), size=1, colour='red') +
               geom_smooth(data=avg2, aes(x=Year, y=pctFG3Team*100), size=2, colour='black')
avgminmax
```



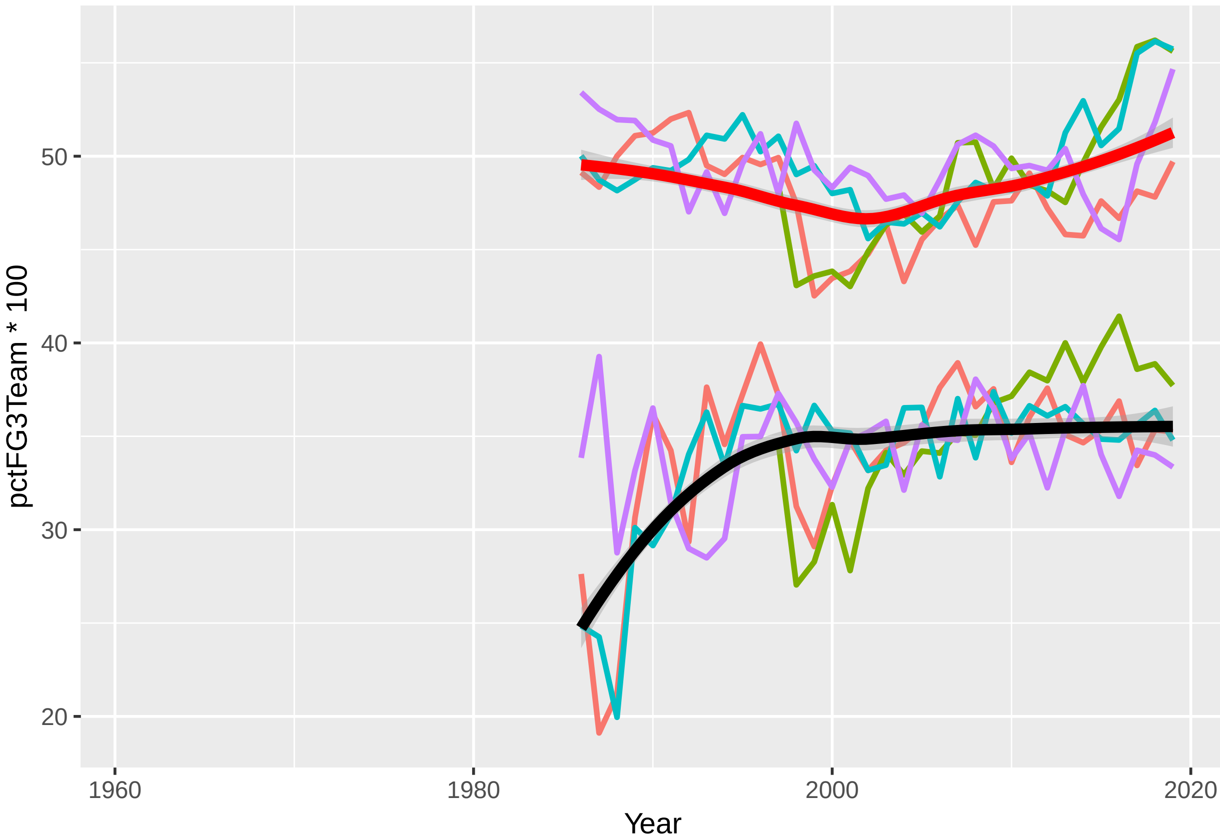

```
avg1986 <- avg %>% filter(Year>=1986)
avg21986 <- avg2 %>% filter(Year>=1986)

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
```



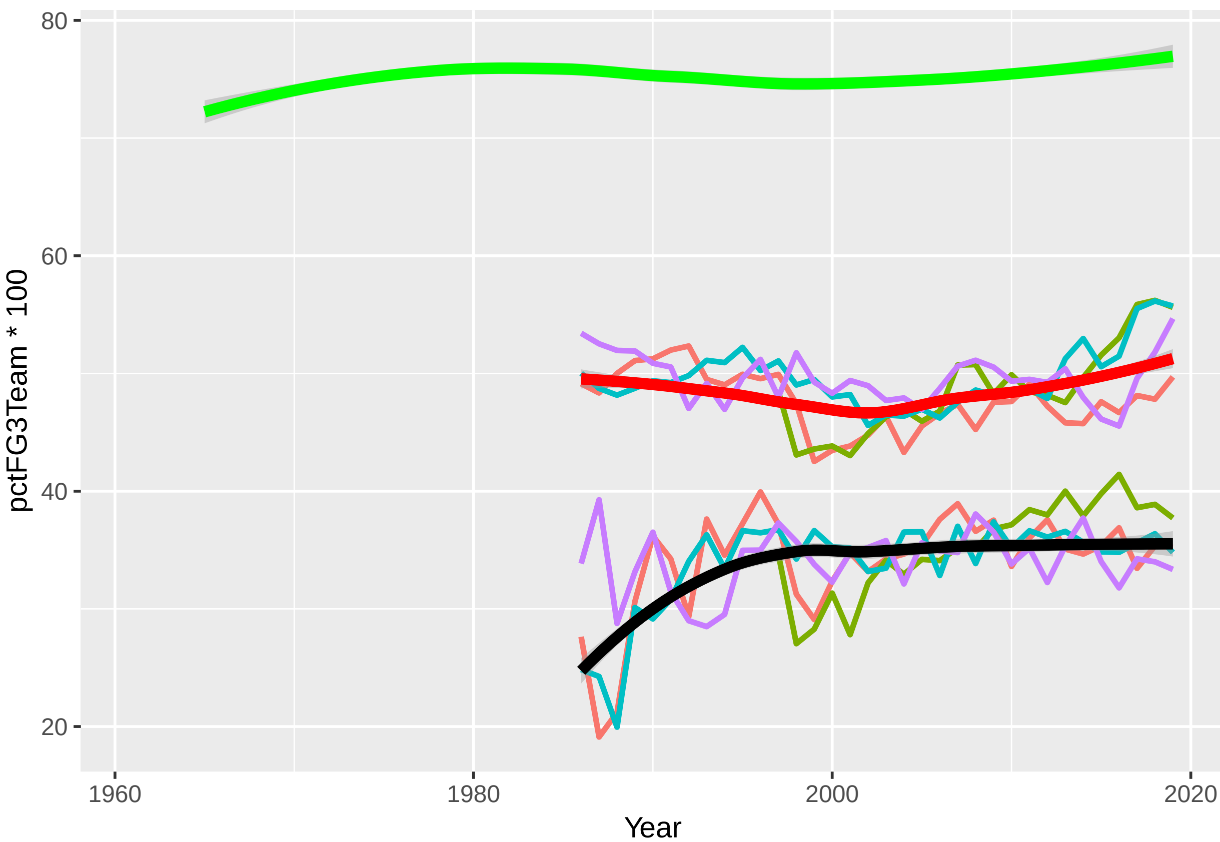

```
avgfiltered <- avg %>% filter(Team %in% c('GSW', 'CHI', 'HOU', 'LAL'))

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
```



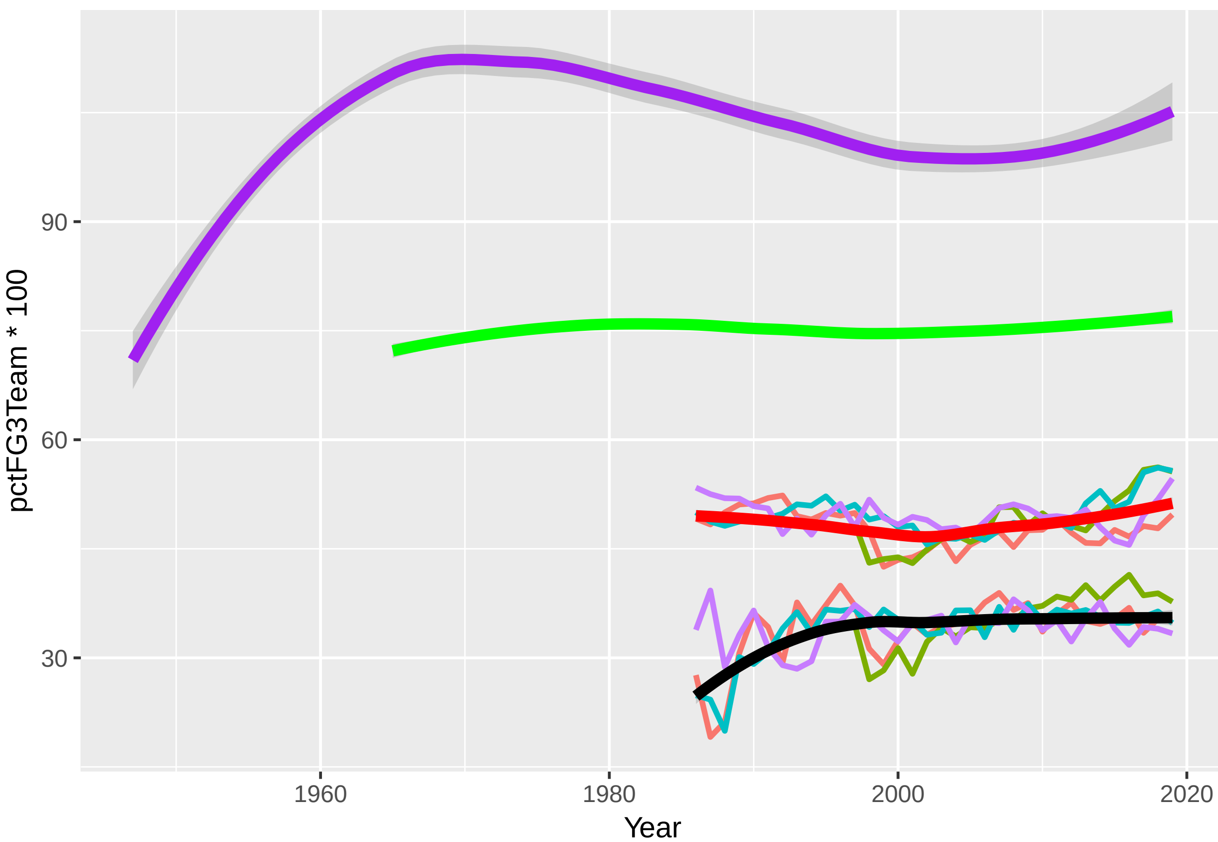

```r
avgcombined2 <- avgcombined +
               geom_smooth(data=avg2, aes(x=Year, y=pctFG2Team*100), size=2, colour='red')
avgcombined2
```




```
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
```

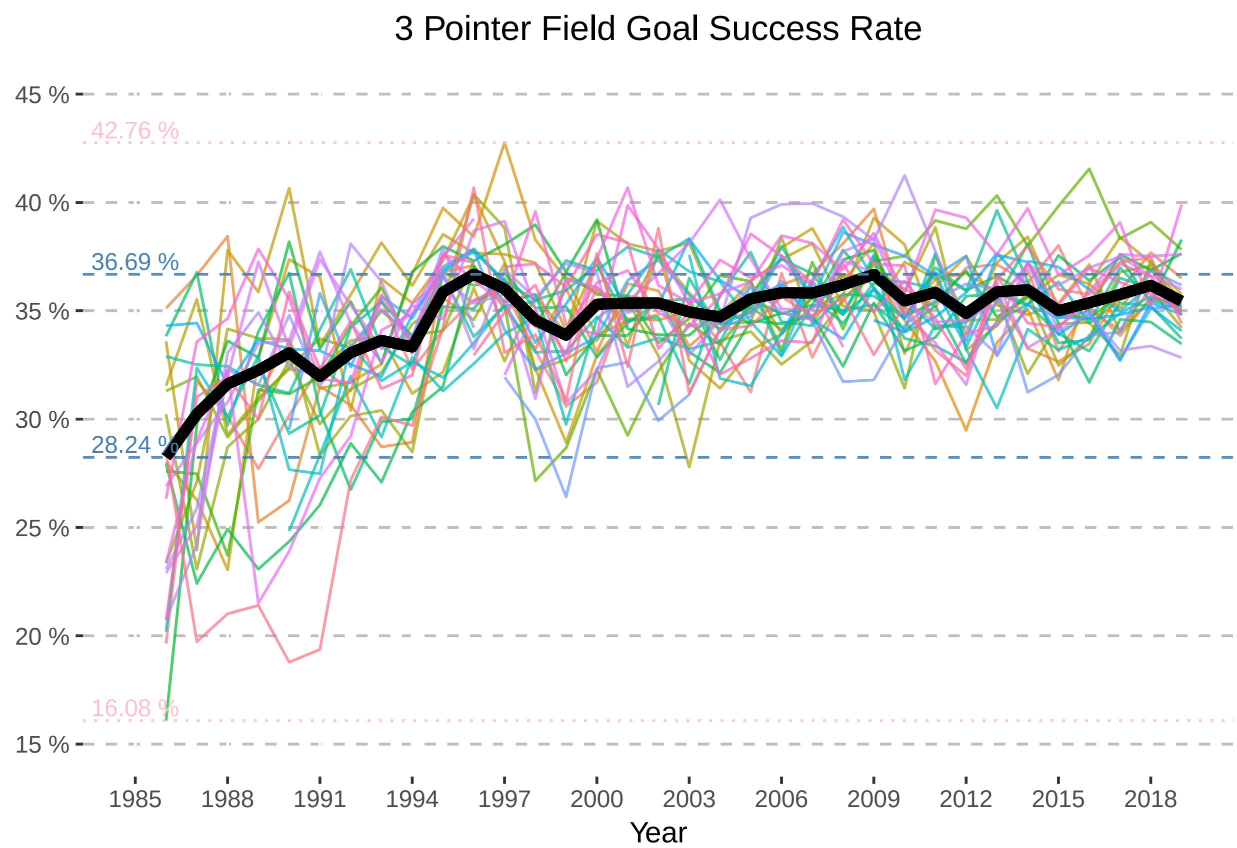



```
# ggplotly(p=ggplot2::last_plot())

# 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$yearSeason, dataGameLogsTeam$
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.
  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=
        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), d
  annotate("text", x=1985, y=max(fg3year$pctfg3)+0.6, label=paste(toString(round(max(fg3year$pctfg3), d
  annotate("text", x=1985, y=min(fg3yearteam$pctfg3)+0.6, label=paste(toString(round(min(fg3yearteam$pc
  annotate("text", x=1985, y=max(fg3yearteam$pctfg3)+0.6, label=paste(toString(round(max(fg3yearteam$pc

Q1
```



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

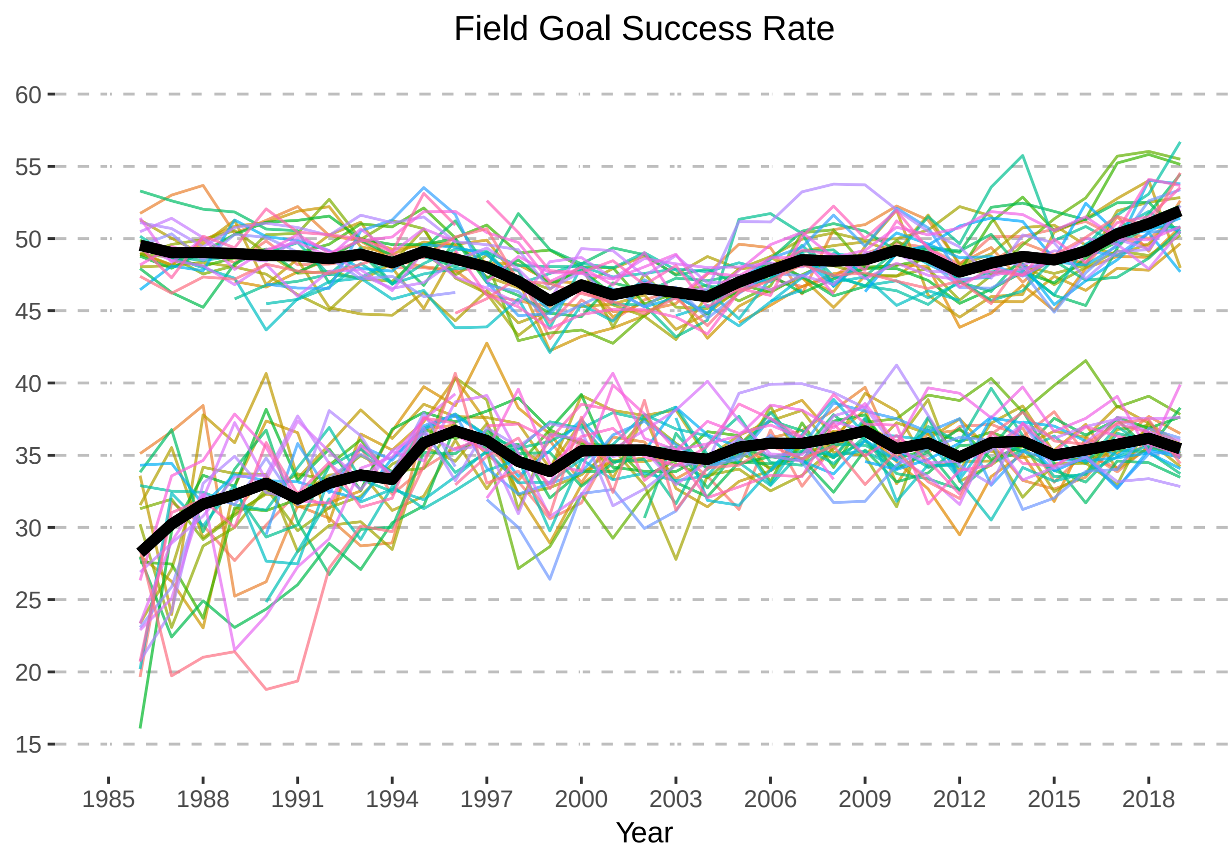
Its debut, in the 1979-80 season, was inauspicious.

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$yearSeason, dataGameLogsTeam$T
colnames(fgyearteam)[1] <- "Year"
colnames(fgyearteam)[2] <- "Team"
fgyearteam <- fgyearteam %>% filter (Year >= 1986)
fgyearteam$pctfg3 <- fgyearteam$fg3mTeam / fgyearteam$fg3aTeam * 100
```

```r
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
```



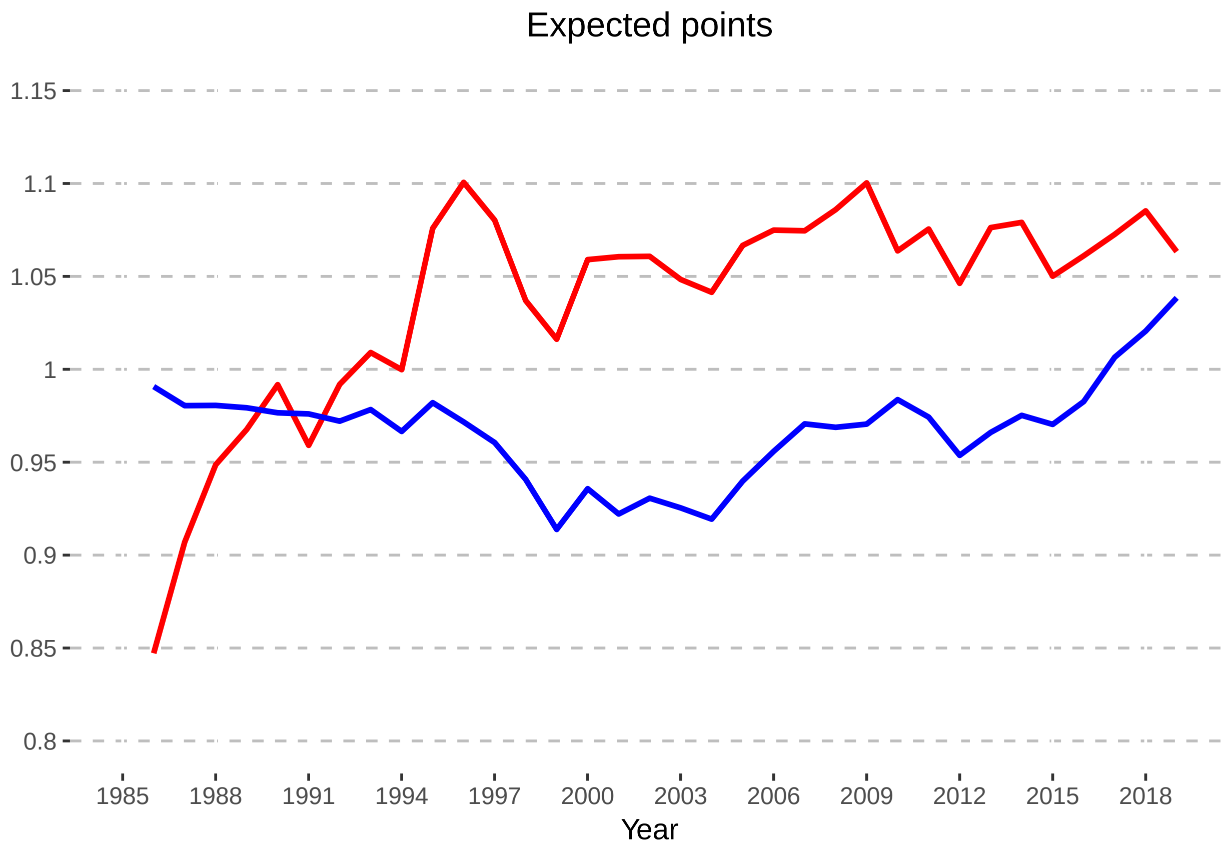
The expected points of 2-point shots in 1986 was ‘r fgyear$pctfg2[1986-1985]/100'*2 =' rfgyear$pctfg2[1986-1985]/100*2’ The expected points of 3-point shots in 1986 was ’r fgyear*pctfg*3[1986 − 1985]/100′ ∗ 3 =′ *rfgyearpctfg3[1986-1985]/1003*’

The expected points of 2-point shots in 2019 was ‘r fgyear$pctfg2[2019-1985]/100'*2 =' rfgyear$pctfg2[2019-1985]/100*2’ The expected points of 3-point shots in 2019 was ’r fgyear*pctfg*3[2019 − 1985]/100′ ∗ 3 =′ *rfgyearpctfg3[2019-1985]/1003*’

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
```

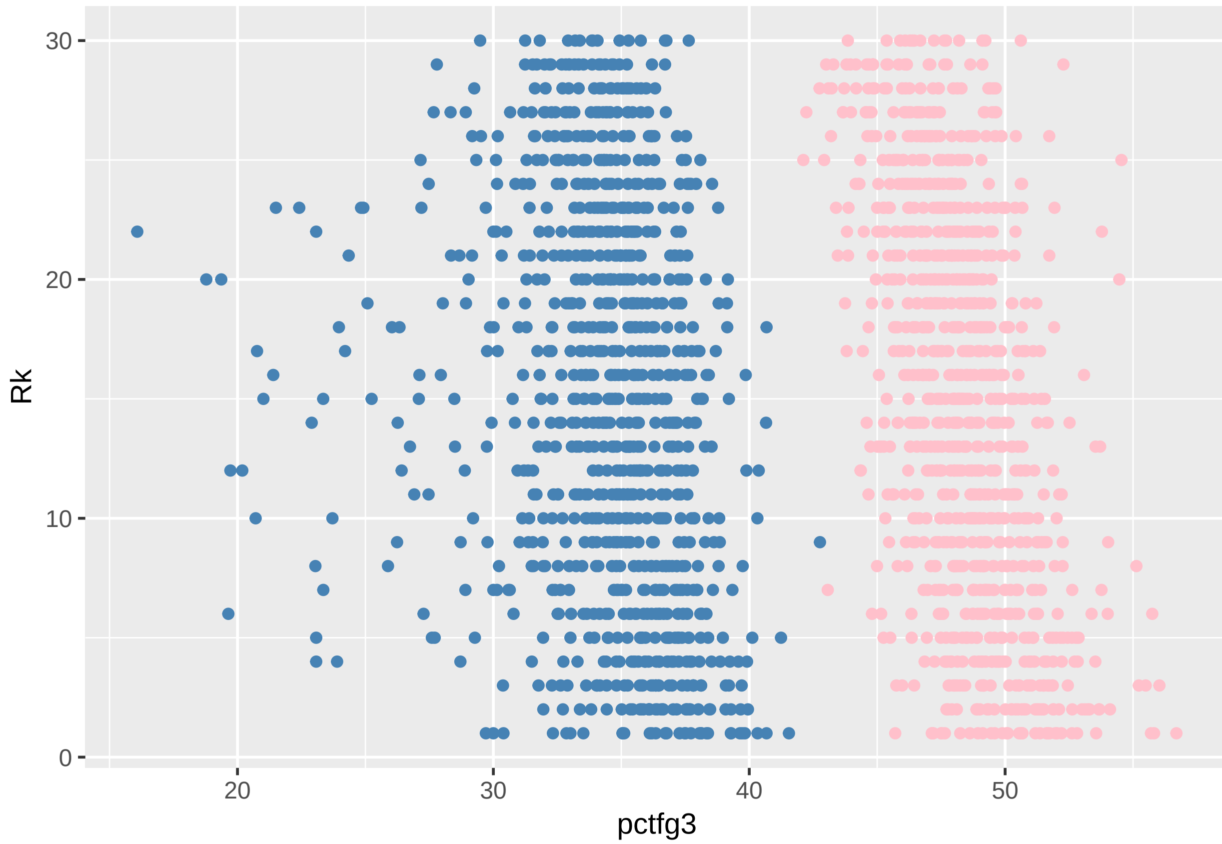


Q3. Teams with more 3-pointers tend to be the better performing teams? - Any insights between standings and 3-pointers?

```
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)) +
```

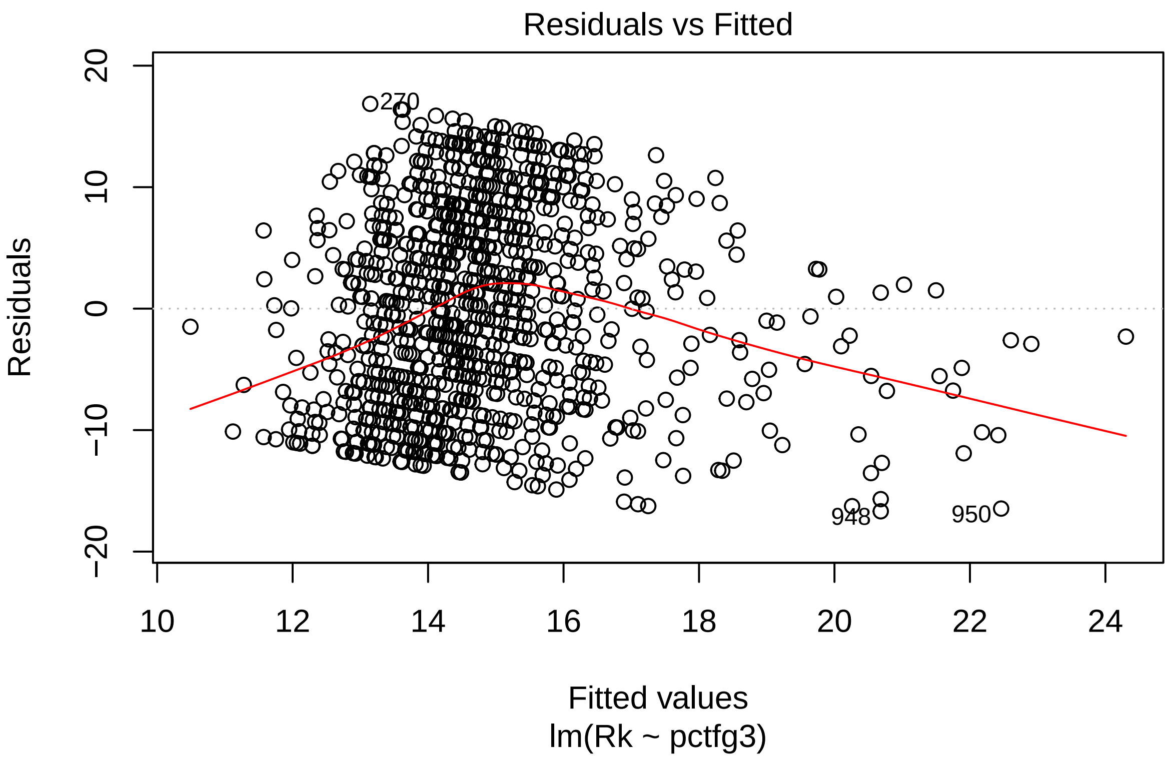
```r
  # scale_y_continuous(limits=c(0.8, 1.15), breaks=yaxisbreaks, labels=yaxisbreaks) +
  # scale_x_continuous(limits=c(1985,2019), breaks=xaxisbreaks)
```

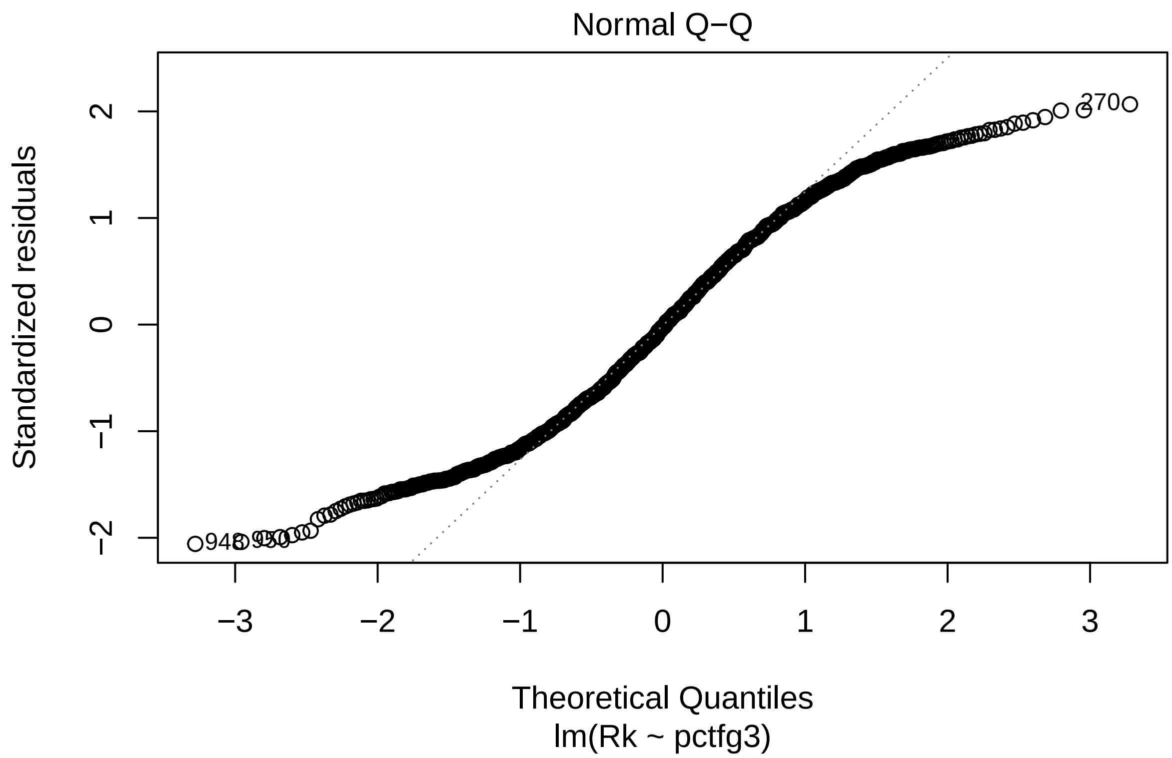
Q3

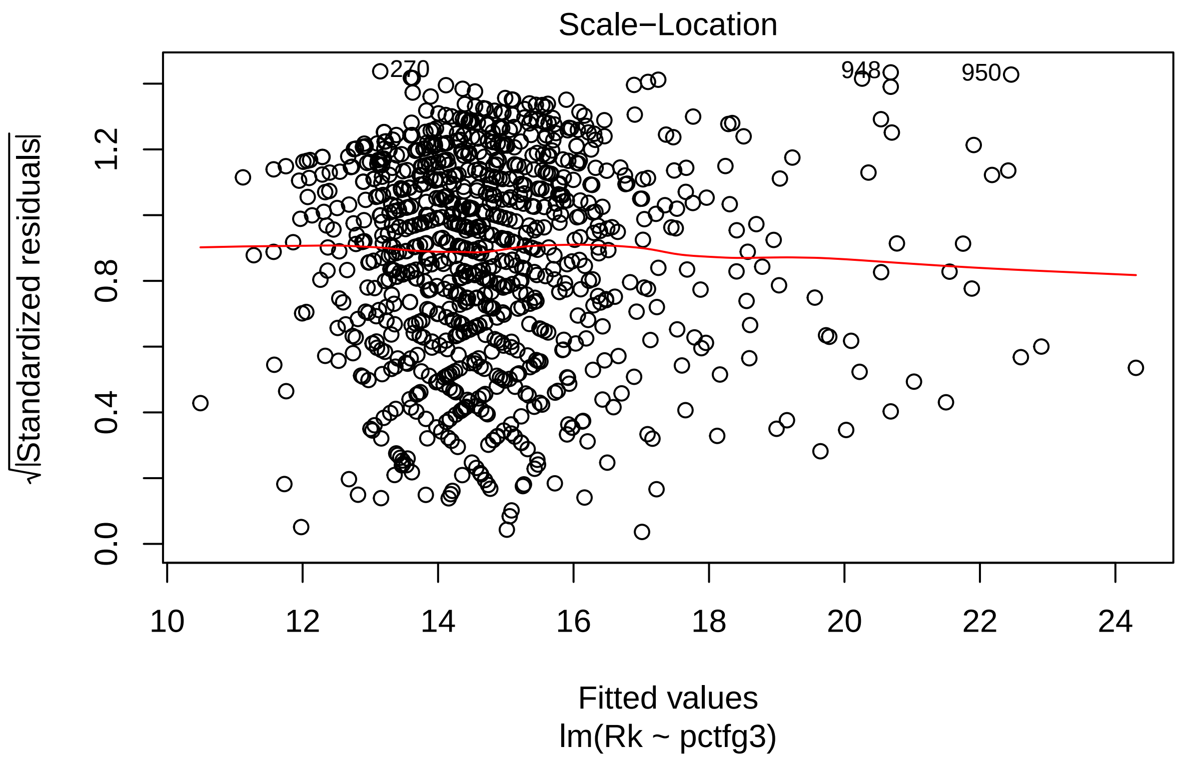


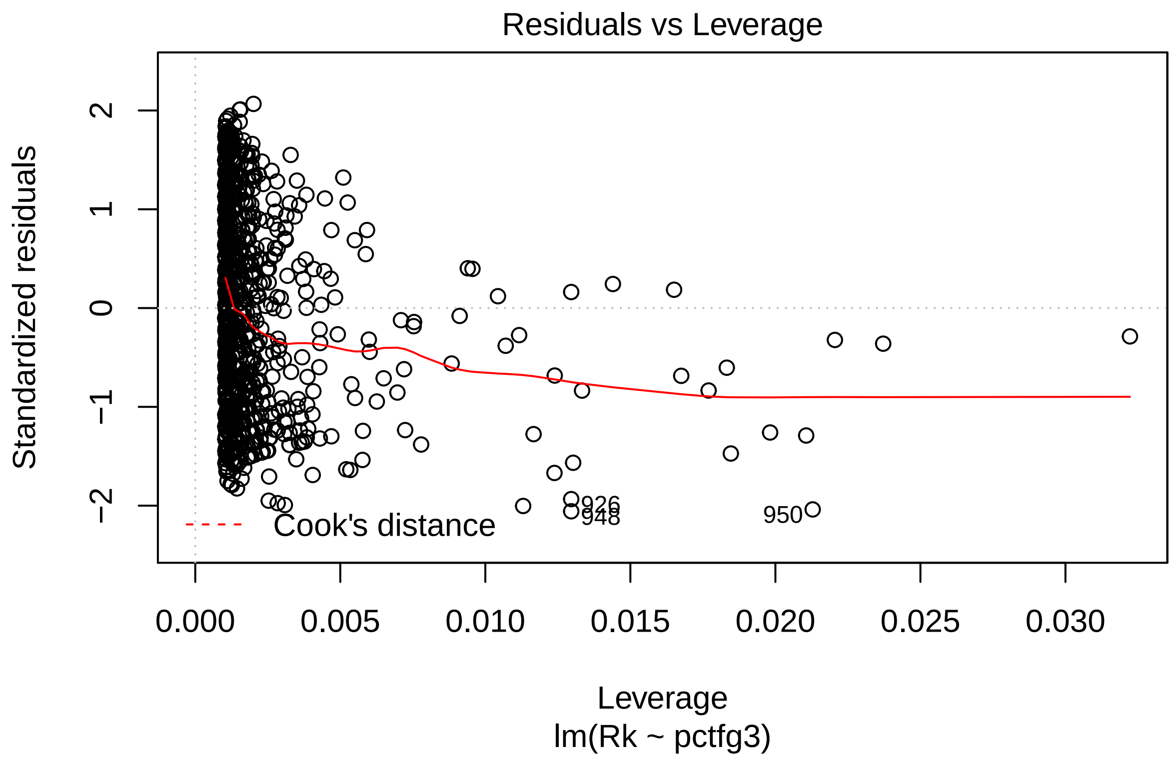

```r
linearModel <- lm(Rk ~ pctfg3, data=standings2)
summary(linearModel)

Call:
lm(formula = Rk ~ pctfg3, data = standings2)

Residuals:
    Min      1Q  Median      3Q     Max 
-16.683  -6.997  -0.212   6.831  16.854 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  32.6295     2.7198   12.00  < 2e-16 ***
pctfg3       -0.5177     0.0787   -6.58  7.7e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.16 on 961 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.0431,    Adjusted R-squared:  0.0421
```

```
F-statistic: 43.3 on 1 and 961 DF,  p-value: 7.74e-11

linearModel2 <- lm(Rk ~ pctfg2, data=standings2)
summary(linearModel2)

Call:
lm(formula = Rk ~ pctfg2, data = standings2)

Residuals:
    Min      1Q  Median      3Q     Max
-18.887  -5.418   0.012   5.334  21.975

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  107.039      4.965    21.6   <2e-16 ***
pctfg2        -1.907      0.103   -18.6   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.16 on 961 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.265, Adjusted R-squared:  0.264
F-statistic:  346 on 1 and 961 DF,  p-value: <2e-16

linearModel3 <- lm(Rk ~ pctfg3 + pctfg2, data=standings2)
summary(linearModel3)

Call:
lm(formula = Rk ~ pctfg3 + pctfg2, data = standings2)

Residuals:
    Min      1Q  Median      3Q     Max
-18.664  -5.402  -0.067   5.285  21.494

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 113.7368     5.1490    22.1  < 2e-16 ***
pctfg3       -0.3049     0.0694    -4.4  1.2e-05 ***
pctfg2       -1.8284     0.1031   -17.7  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.09 on 960 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.279, Adjusted R-squared:  0.278
F-statistic:  186 on 2 and 960 DF,  p-value: <2e-16

plot(linearModel)
```

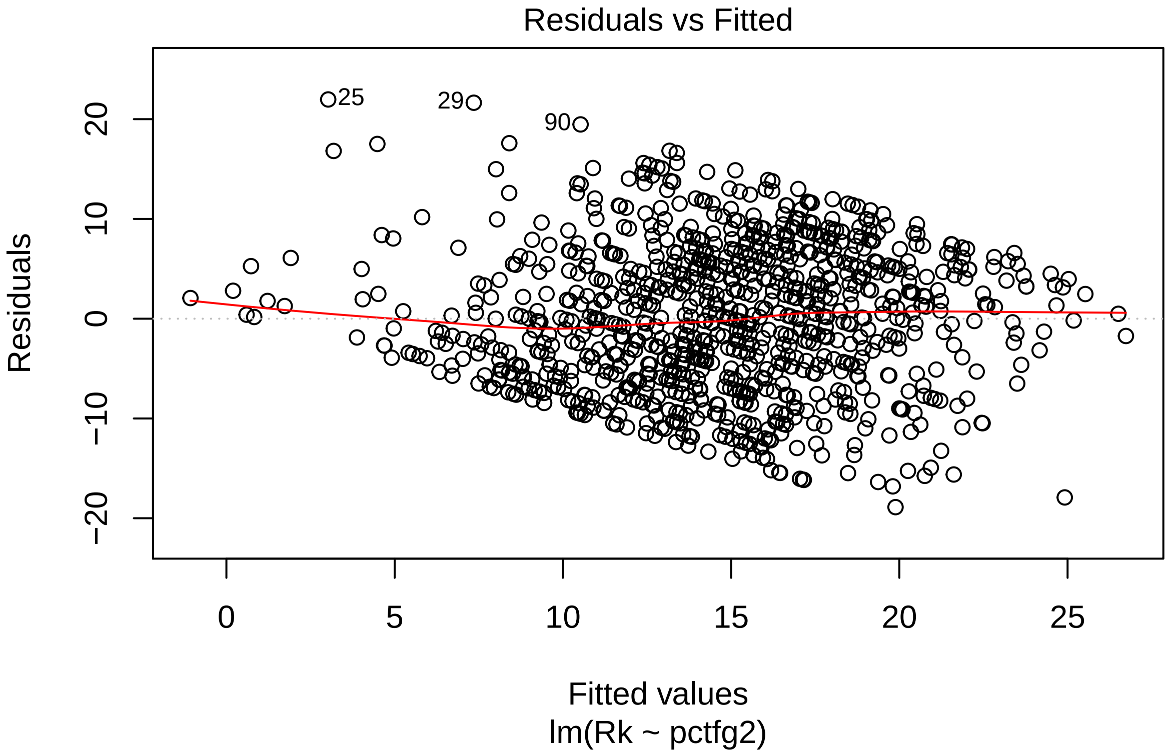

Residuals vs Fitted
Residuals
Fitted values
lm(Rk ~ pctfg3)
270
948
950
20
10
0
−10
−20
10
12
14
16
18
20
22
24

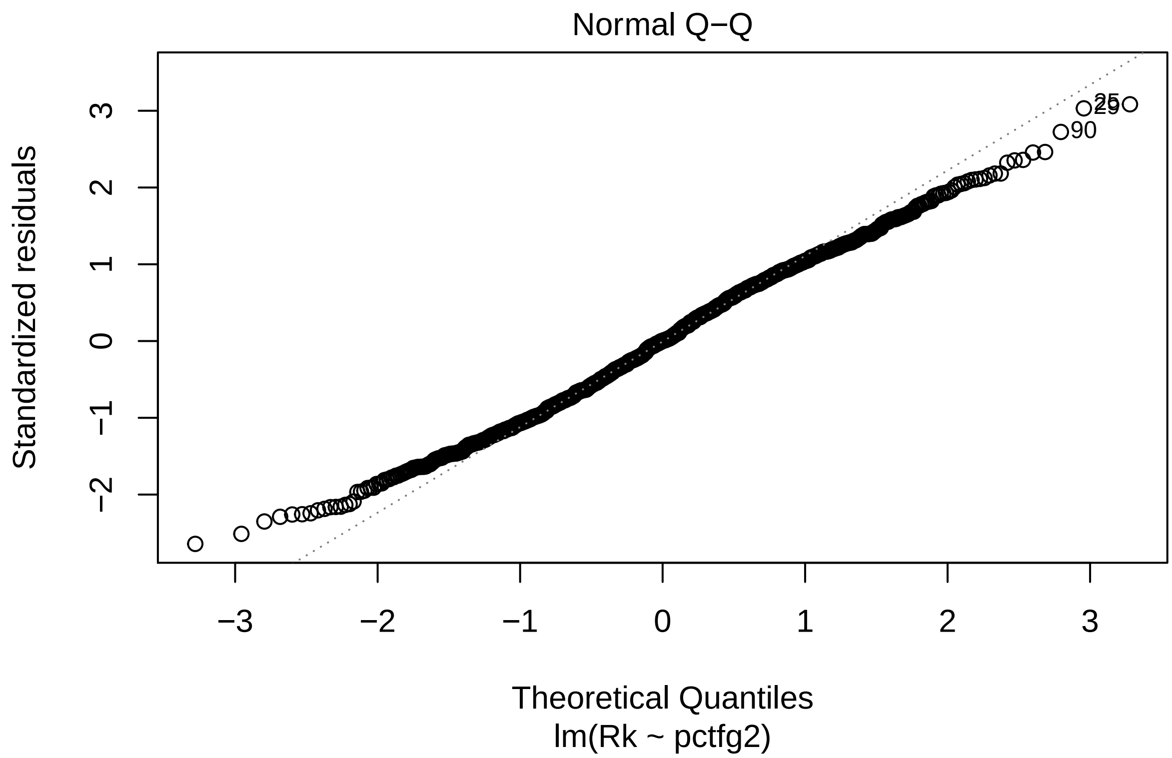

Normal Q−Q
Standardized residuals
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(Rk ~ pctfg3)
270
948

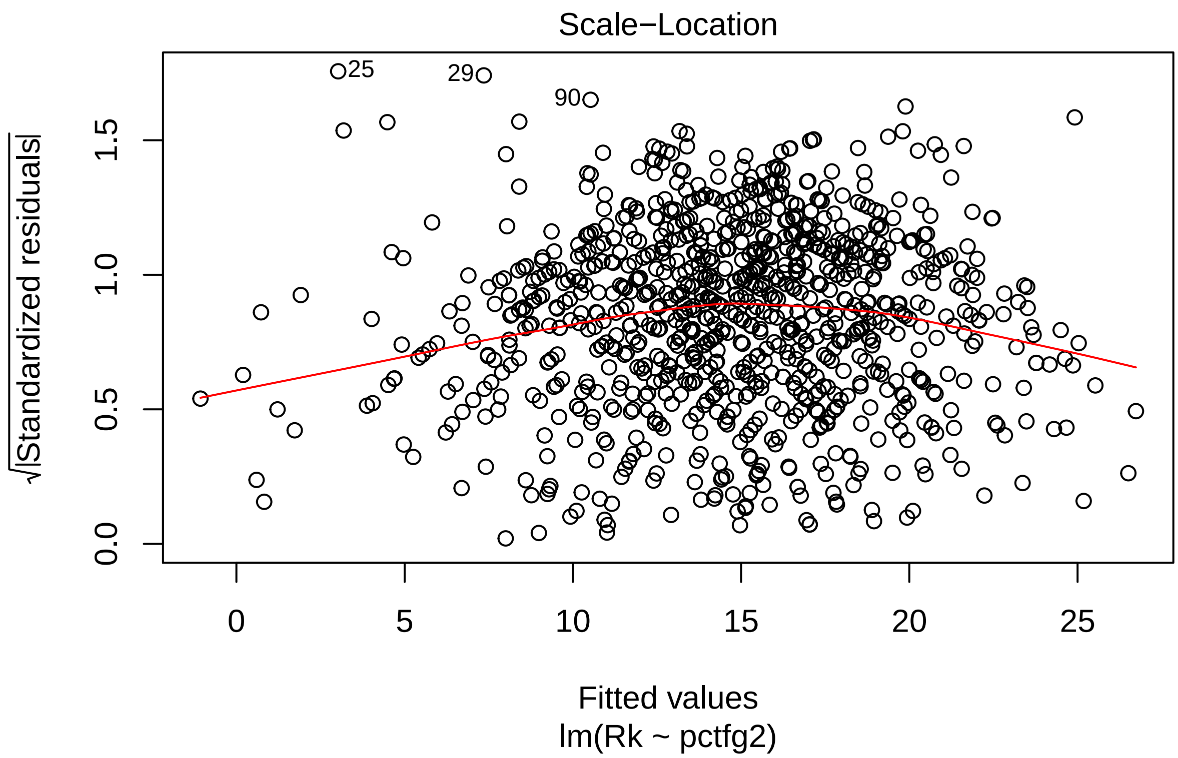

Scale−Location
270
948
950
√|Standardized residuals|
0.0
0.4
0.8
1.2
10
12
14
16
18
20
22
24
Fitted values
lm(Rk ~ pctfg3)

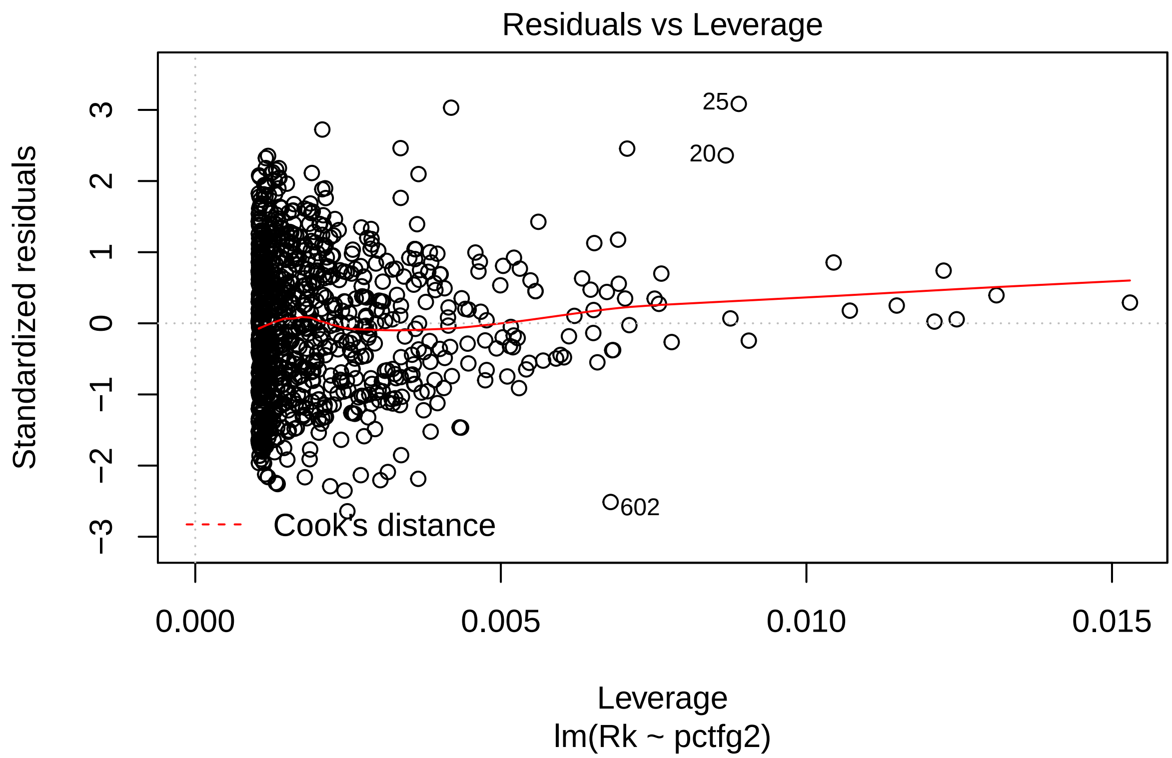



```
plot(linearModel2)
```

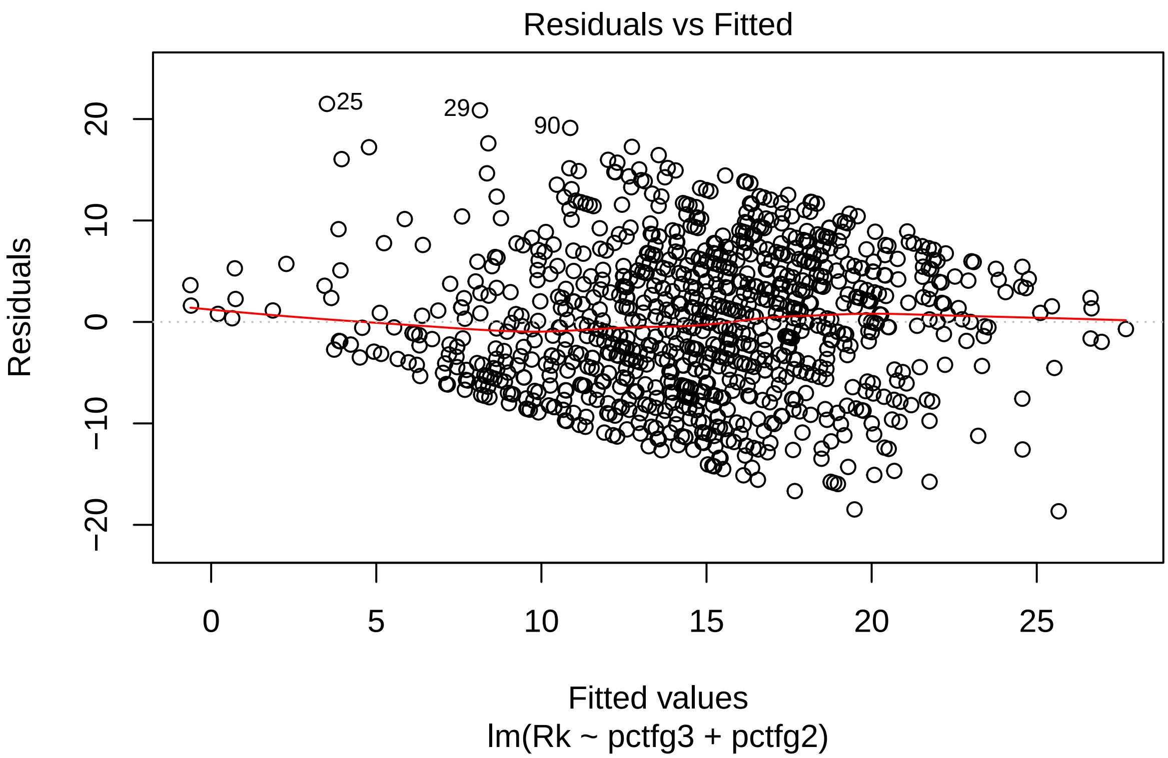

Residuals vs Fitted
Residuals
Fitted values
lm(Rk ~ pctfg2)
25
29
90
20
10
0
−10
−20
0
5
10
15
20
25

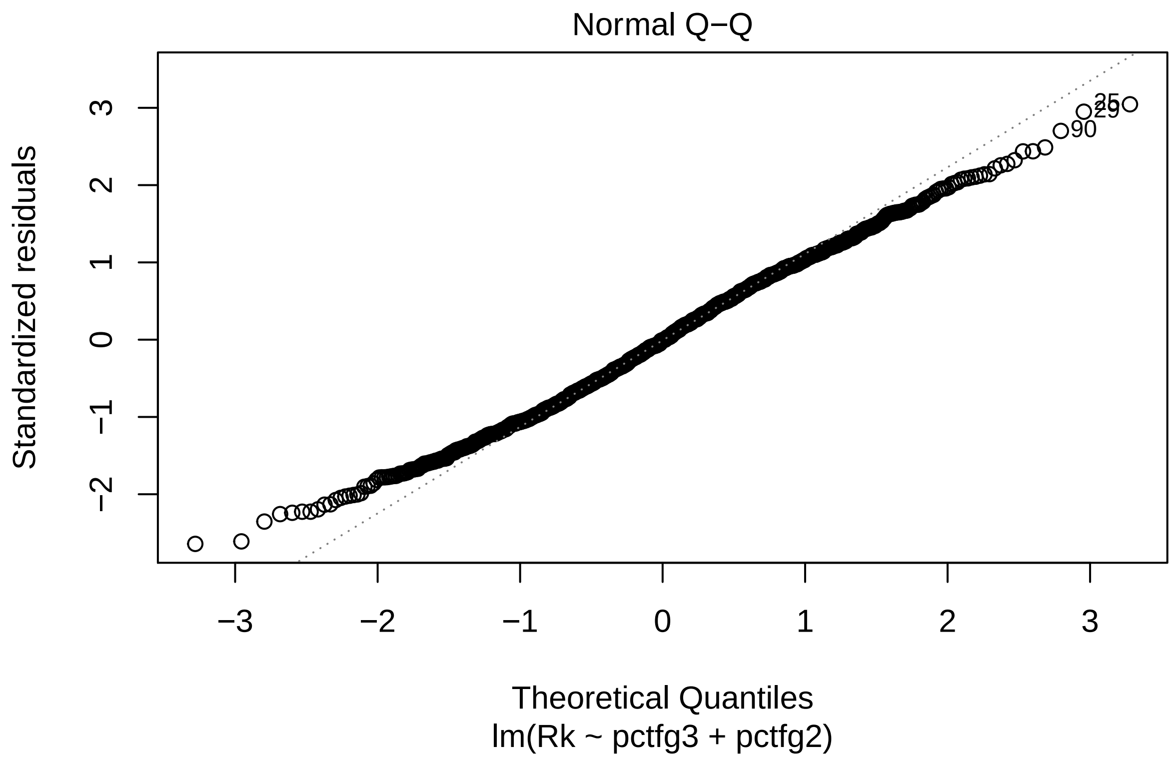

Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
25
29
90
Theoretical Quantiles
lm(Rk ~ pctfg2)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
25
29
90
0
5
10
15
20
25
Fitted values
lm(Rk ~ pctfg2)

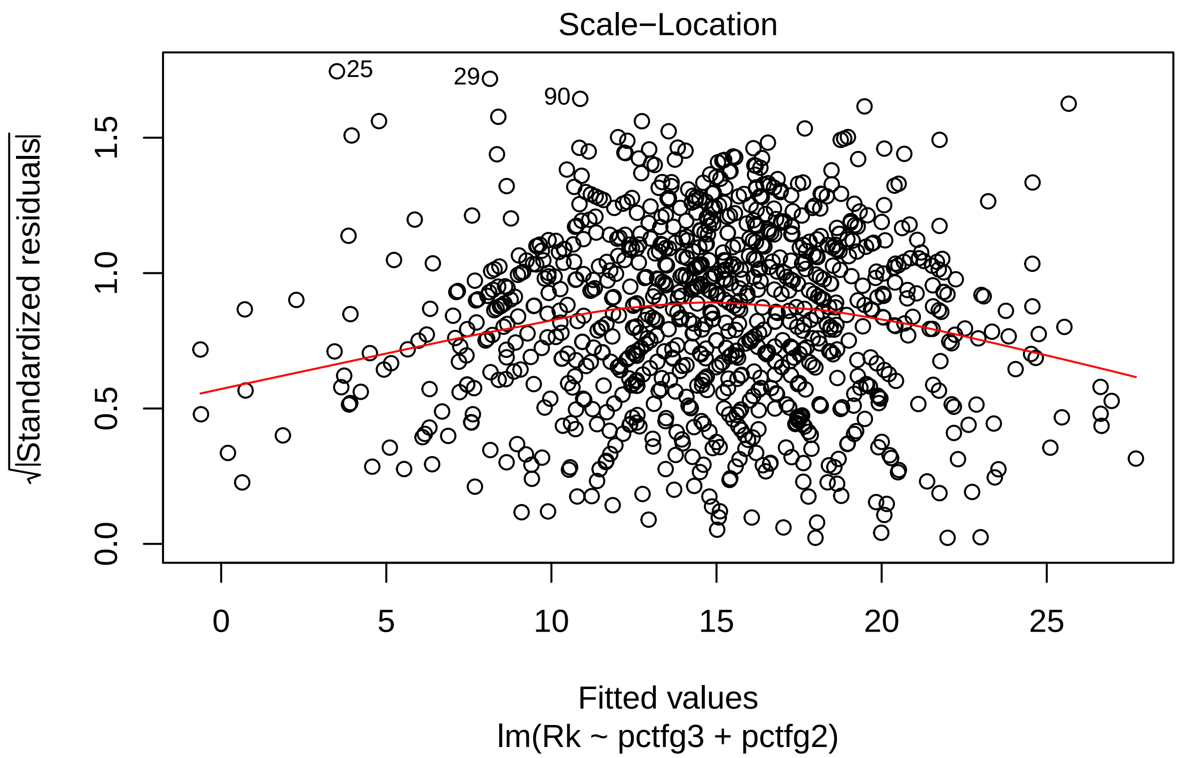



```
plot(linearModel3)
```

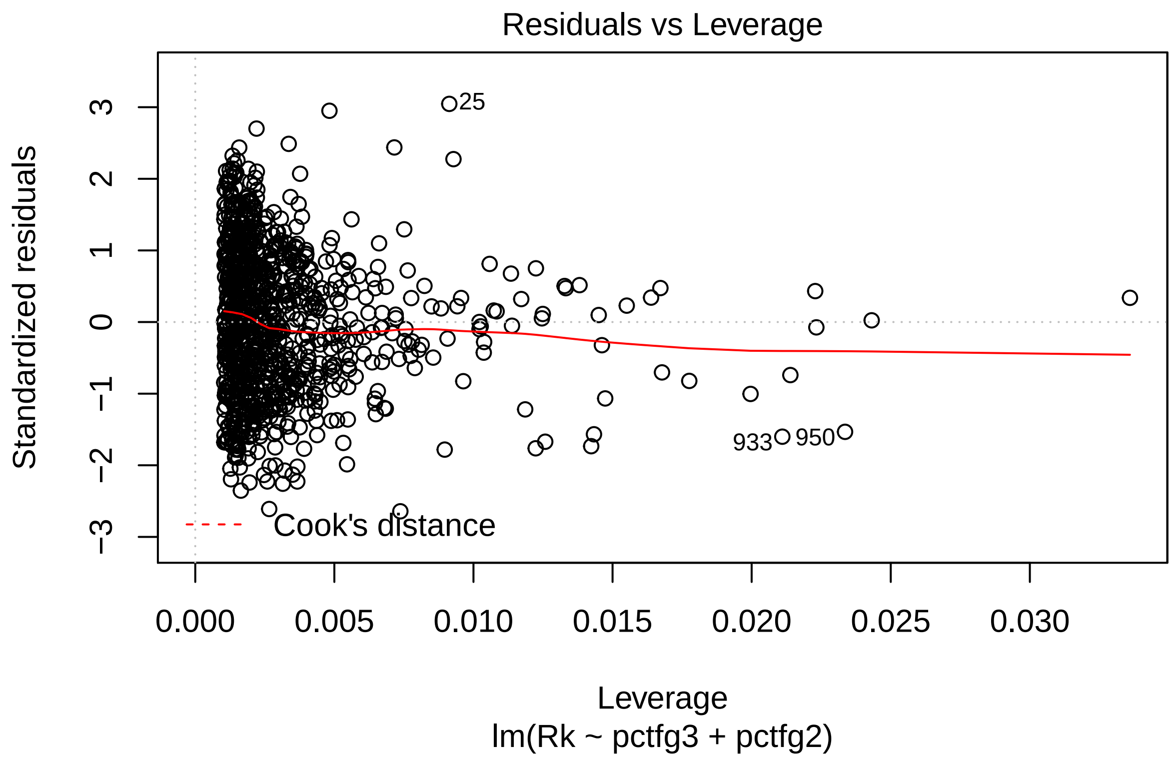

Residuals vs Fitted
25
29
90
20
10
0
−10
−20
Residuals
0
5
10
15
20
25
Fitted values
lm(Rk ~ pctfg3 + pctfg2)



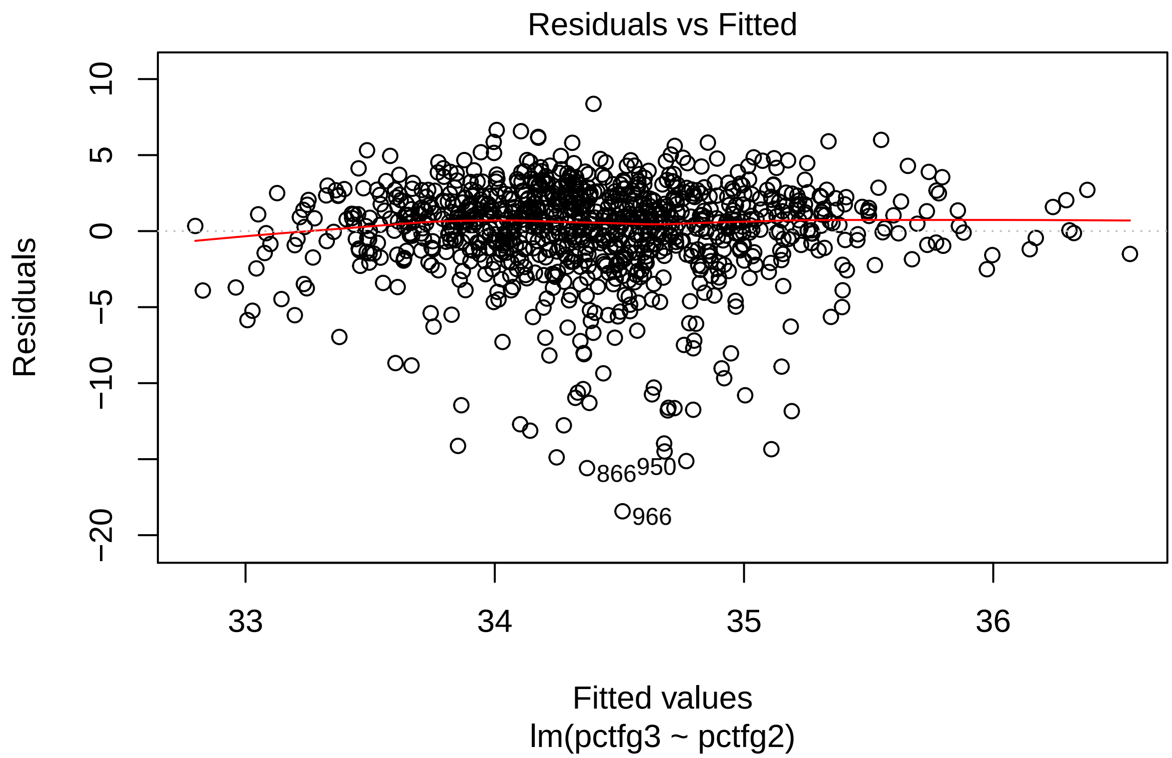
Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
25
29
90
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(Rk ~ pctfg3 + pctfg2)

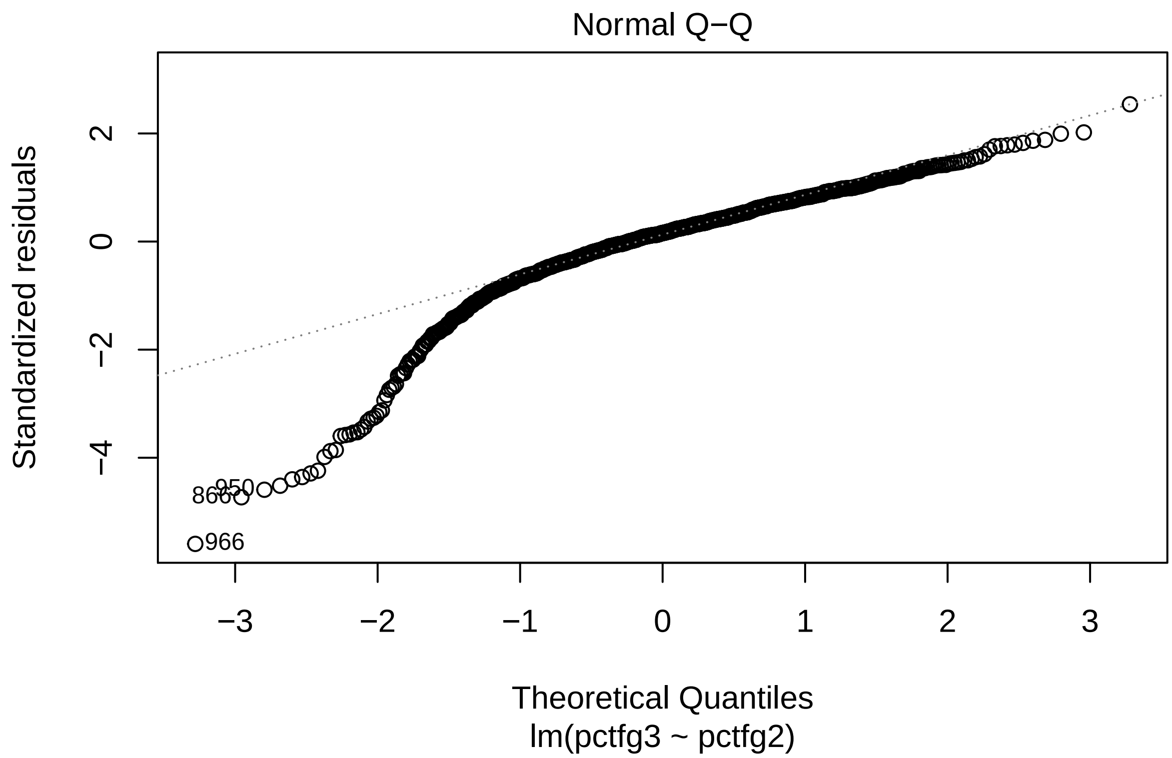


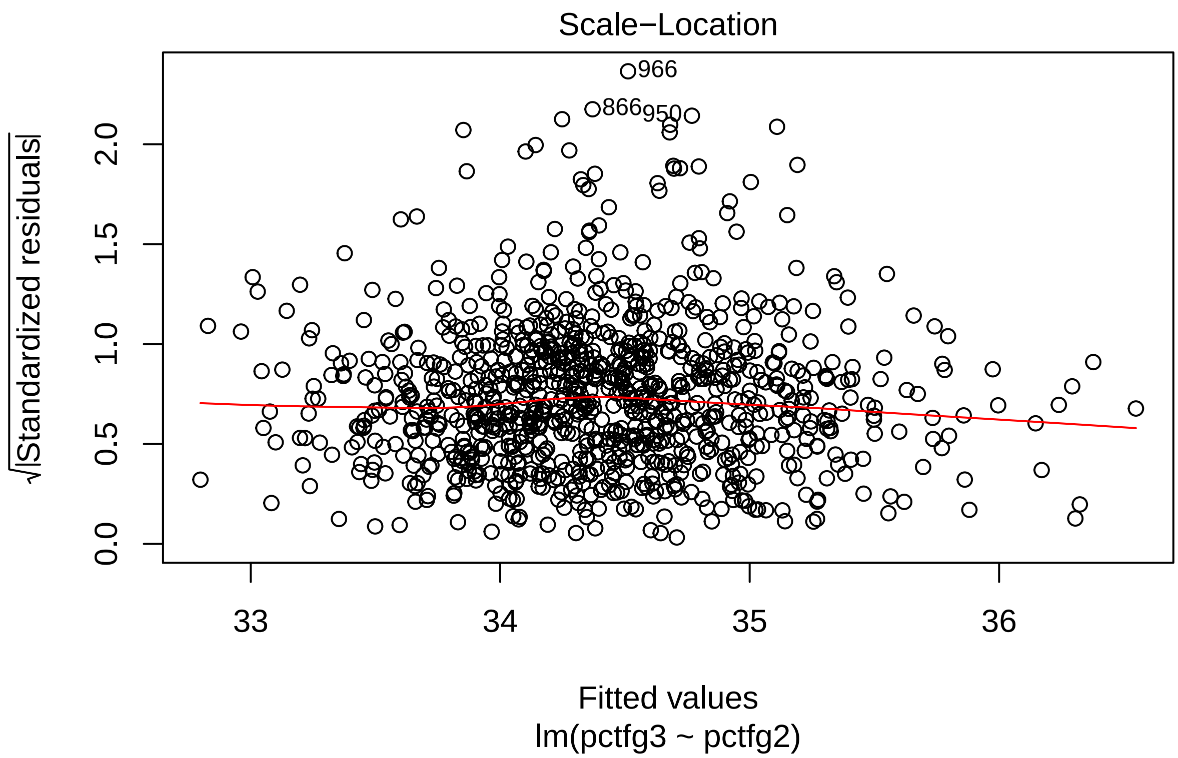
Scale−Location
25
29
90
√|Standardized residuals|
0.0
0.5
1.0
1.5
0
5
10
15
20
25
Fitted values
lm(Rk ~ pctfg3 + pctfg2)

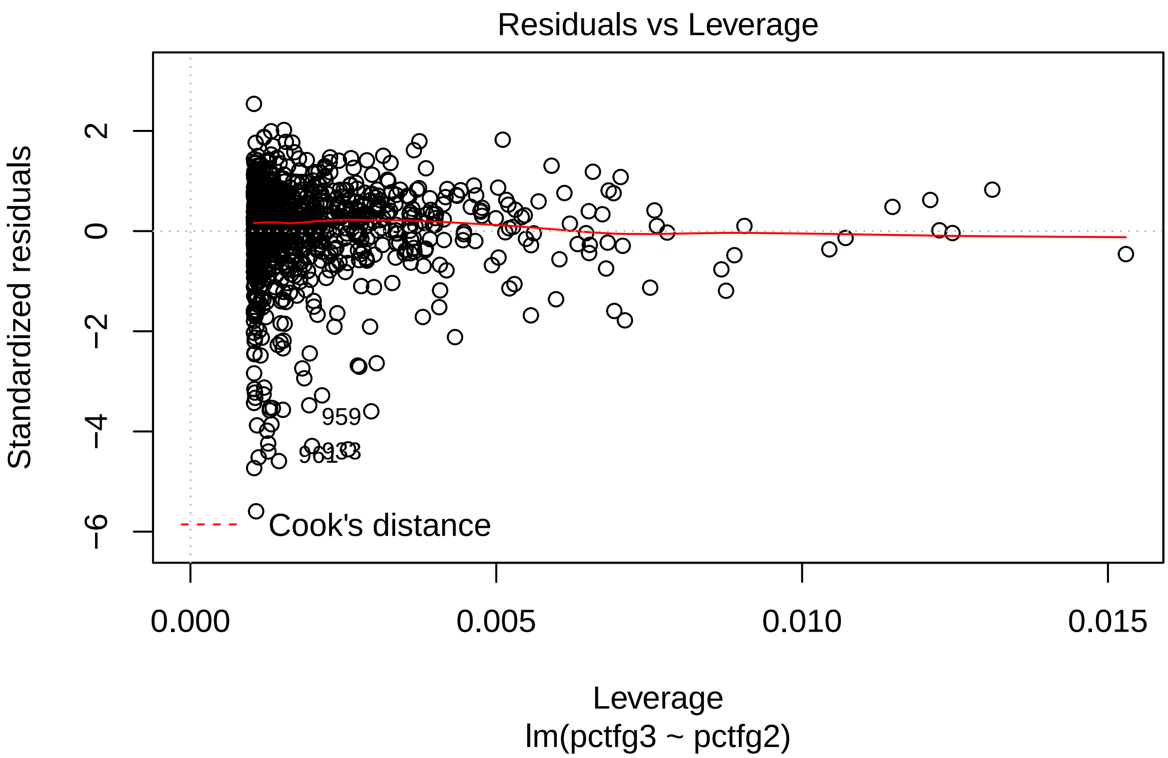



```
linearModel4 <- lm(pctfg3 ~ pctfg2, data=standings2)
summary(linearModel4)

Call:
lm(formula = pctfg3 ~ pctfg2, data = standings2)

Residuals:
    Min       1Q   Median       3Q      Max
-18.429   -1.209    0.517    2.055    8.368

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  21.9658     2.2869    9.60  < 2e-16 ***
pctfg2        0.2572     0.0472    5.45  6.6e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.3 on 961 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.0299,    Adjusted R-squared:  0.0289
F-statistic: 29.7 on 1 and 961 DF,  p-value: 6.57e-08
plot(linearModel4)
```

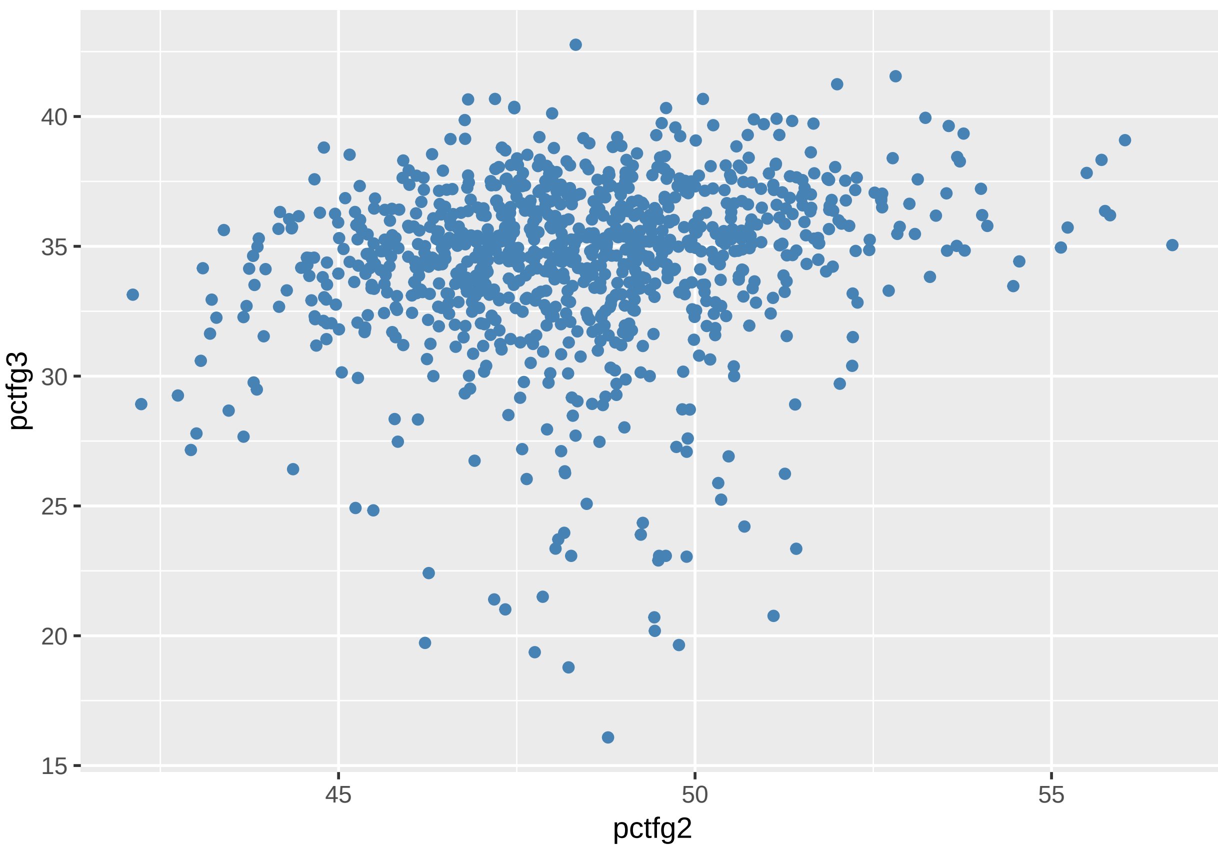

Residuals vs Fitted
10
5
0
−5
−10
−20
Residuals
866
950
966
33
34
35
36
Fitted values
lm(pctfg3 ~ pctfg2)


Normal Q−Q
Standardized residuals
2
0
−2
−4
950
866
966
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(pctfg3 ~ pctfg2)


Scale−Location
966
866
950
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
33
34
35
36
Fitted values
lm(pctfg3 ~ pctfg2)




```
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

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

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

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