Problem Set 6

Subeom Lee

2019-03-15

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

Questions

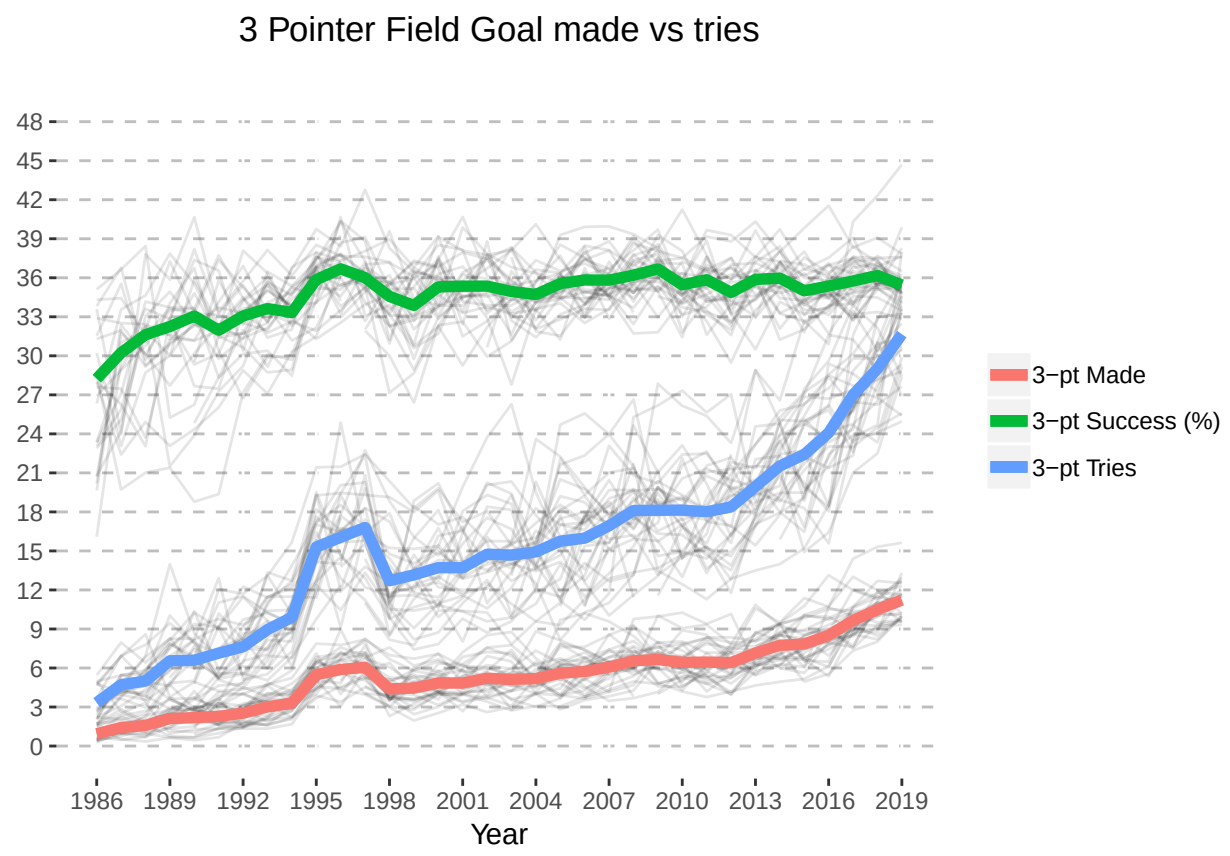
```
load('BaseEnvironment.Rdata')
```

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

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$yearSeason, dataGameLogsTeam$Team), mean)
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, group=Team, alpha=0.5), size=0.5) +
  geom_line(data=fg3yearteamavg, aes(x=Year, y=fg3aTeam, group=Team, alpha=0.5), size=0.5) +
  geom_line(data=fg3yearteamavg, aes(x=Year, y=pctfg3, group=Team, alpha=0.5), size=0.5) +
  geom_line(data=fg3yearavg, aes(x=Year, y=fg3mTeam, colour="3-pt Made", alpha=0.9), size=2) +
  geom_line(data=fg3yearavg, aes(x=Year, y=fg3aTeam, colour="3-pt Tries", alpha=0.9), size=2) +
  geom_line(data=fg3year, aes(x=Year, y=pctfg3, colour="3-pt Success (%)", alpha=0.9), size=2) +
  guides(alpha=FALSE) +
  xlab('Year') +
  ylab(NULL) +
  ggtitle('3 Pointer Field Goal made vs tries') +
  theme(panel.background=element_rect(fill=NA)) +
  theme(panel.grid.major.y=element_line(color="grey", linetype=2)) +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(legend.title=element_blank()) +
```

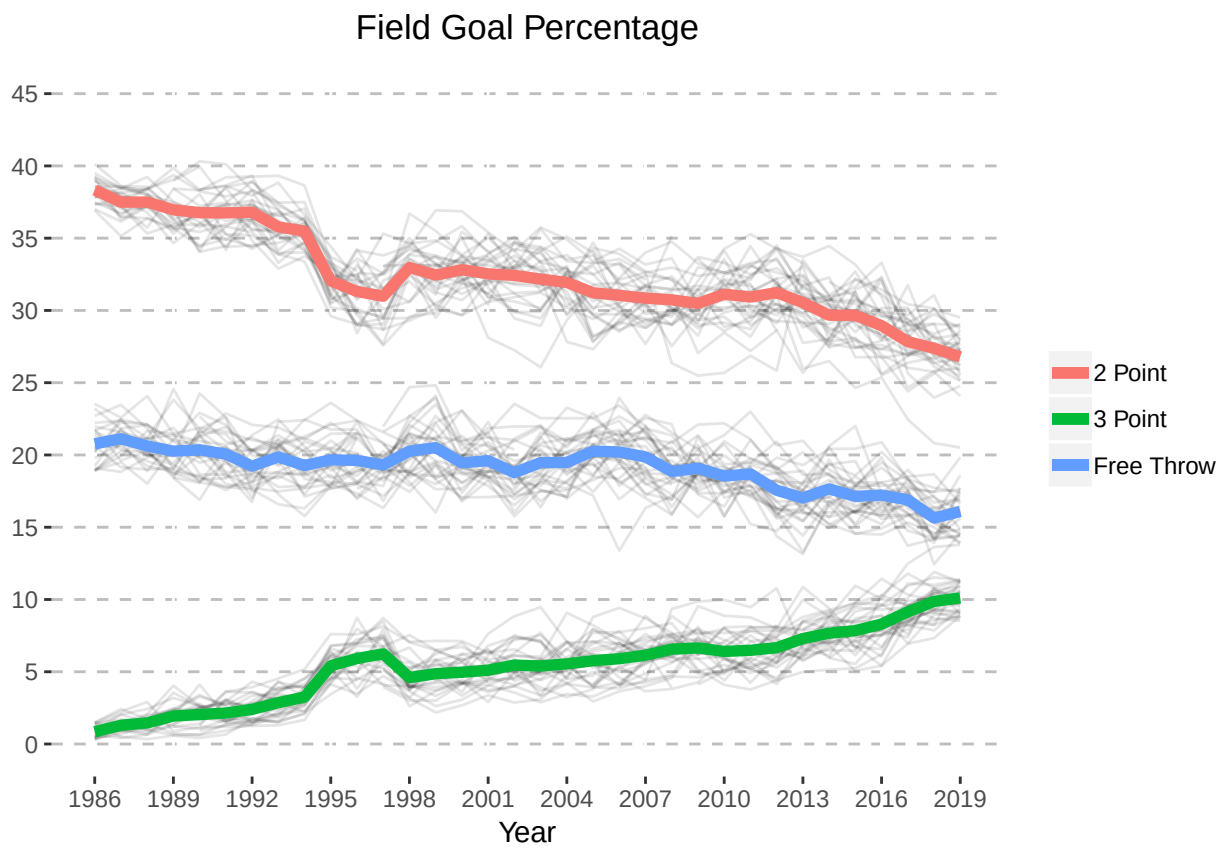
```
  scale_y_continuous(limits=c(0, 50), breaks=yaxisbreaks) +
  scale_x_continuous(limits=c(1986,2019), breaks=xaxisbreaks)

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$yearSeason, dataGameLogsTeam$Team), mean)
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, alpha=0.5, group=Team), size=0.5) +
  geom_line(data=fgallyearteamavg, aes(x=Year, y=pctpts2, alpha=0.5, group=Team), size=0.5) +
  geom_line(data=fgallyearteamavg, aes(x=Year, y=pctptsft, alpha=0.5, group=Team), size=0.5) +
  geom_line(data=fgallyearavg, aes(x=Year, y=pctpts3, colour='3 Point', alpha=0.9), size=2) +
  geom_line(data=fgallyearavg, aes(x=Year, y=pctpts2, colour='2 Point', alpha=0.9), size=2) +
  geom_line(data=fgallyearavg, aes(x=Year, y=pctptsft, colour='Free Throw', alpha=0.9), size=2) +
  guides(alpha=FALSE) +
  xlab('Year') +
  ylab(NULL) +
  ggtitle('Field Goal Percentage') +
  theme(panel.background=element_rect(fill=NA)) +
  theme(panel.grid.major.y=element_line(color="grey", linetype=2)) +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(legend.title=element_blank()) +
  scale_y_continuous(limits=c(0, 45), breaks=yaxisbreaks) +
  scale_x_continuous(limits=c(1986,2019), breaks=xaxisbreaks)
```

```
Q1_3
```



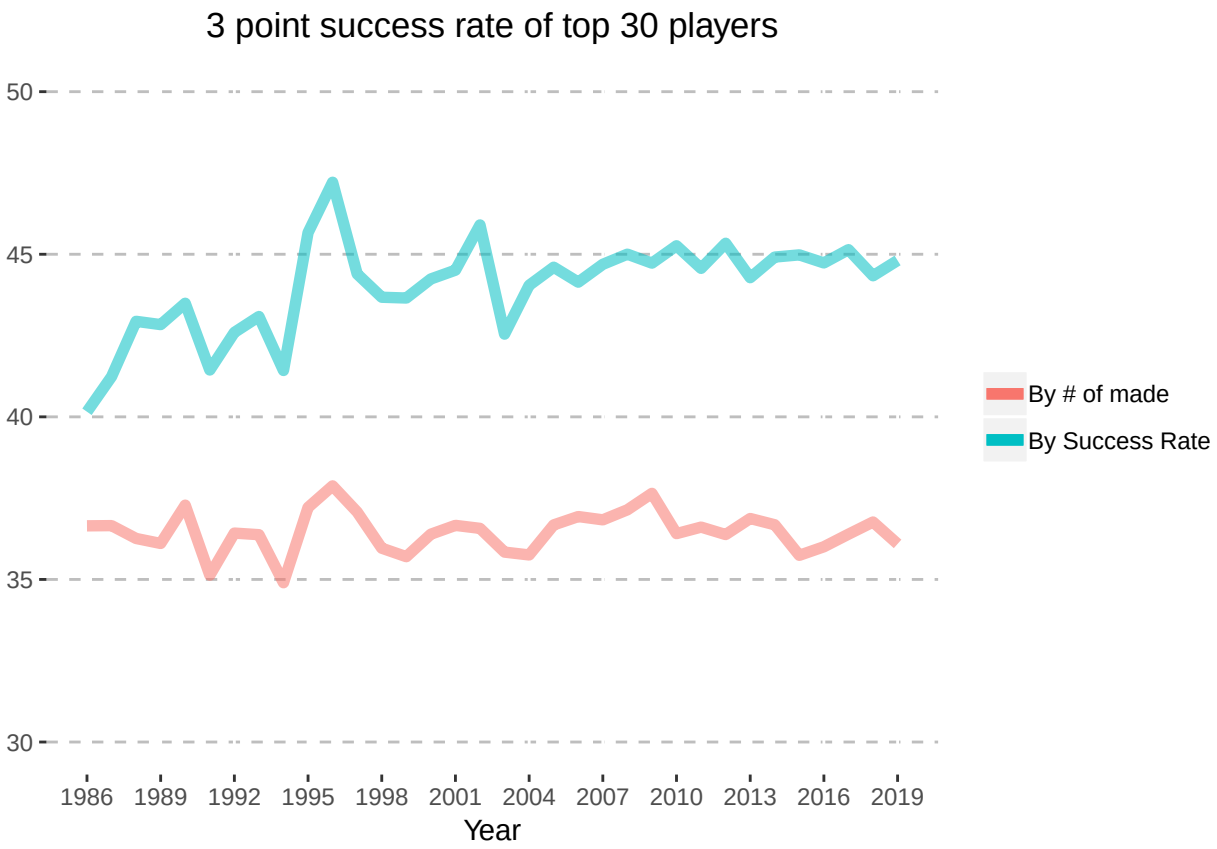
Statistics of top 10 3-point shooters each year

```
summary(fgyearplayer)
      Year         Player               fgm              fga        
 Min.   :1986   Length:14714       Min.   :  0.0   Min.   :  0.0  
 1st Qu.:1995   Class :character   1st Qu.: 49.0   1st Qu.: 115.0  
 Median :2004   Mode  :character   Median : 154.0   Median : 345.0  
 Mean   :2004                      Mean   : 200.3   Mean   : 437.3  
 3rd Qu.:2012                      3rd Qu.: 308.0   3rd Qu.: 670.0  
```

```
 Max.   :2019                    Max.   :1098   Max.   :2279
                                                NA's   :9
      fg3m          fg3a            ftm           fta          pctfg3
 Min.   :  0   Min.   :  0.0   Min.   :  0   Min.   :  0   Min.   :  0.0
 1st Qu.:  0   1st Qu.:  2.0   1st Qu.: 20   1st Qu.: 29   1st Qu.: 16.0
 Median :  5   Median : 21.0   Median : 64   Median : 88   Median : 30.9
 Mean   : 29   Mean   : 82.2   Mean   :101   Mean   :134   Mean   : 26.3
 3rd Qu.: 44   3rd Qu.:128.0   3rd Qu.:145   3rd Qu.:194   3rd Qu.: 36.8
 Max.   :402   Max.   :886.0   Max.   :833   Max.   :972   Max.   :100.0
 NA's   :15    NA's   :16                    NA's   :9     NA's   :2300
     pctfg2          pctft
 Min.   :  0.0   Min.   :  0.0
 1st Qu.: 40.3   1st Qu.: 66.6
 Median : 44.3   Median : 75.0
 Mean   : 44.0   Mean   : 72.4
 3rd Qu.: 48.5   3rd Qu.: 81.4
 Max.   :100.0   Max.   :100.0
 NA's   :63      NA's   :470
```

```
# fgtopplayerbysuccessrate
fgyeartopsr <-
  fgyearplayer %>%
  group_by(Year) %>%
  filter(fg3m >= 29) %>%
  mutate(Rank = order(order(pctfg3, decreasing = TRUE))) %>%
  filter(Rank <= 10) %>%
  arrange(Year, Rank)

fgyeartopsr
# A tibble: 340 x 12
# Groups:   Year [34]
    Year Player   fgm   fga  fg3m  fg3a   ftm   fta pctfg3 pctfg2 pctft
   <int> <chr>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>  <dbl> <dbl>
 1  1986 Craig~   284   568    73   161    75    86   45.3   50    87.2
 2  1986 Trent~   349   739    41    91    79   100   45.1   47.2  79
 3  1986 Larry~   796  1606    82   194   441   492   42.3   49.6  89.6
 4  1986 World~   652  1428    71   169   379   486   42.0   45.7  78.0
 5  1986 Kyle ~   286   592    58   140    73    90   41.4   48.3  81.1
 6  1986 Micha~   274   606    63   163   147   170   38.7   45.2  86.5
 7  1986 Leon ~   184   463    41   112   123   155   36.6   39.7  79.4
 8  1986 Dale ~   193   470    63   174    59    82   36.2   41.1  72.0
 9  1986 Mike ~   252   544    41   114    42    64   36.0   46.3  65.6
10  1986 Brad ~   267   502    32    89   198   228   36.0   53.2  86.8
# ... with 330 more rows, and 1 more variable: Rank <int>

# fgtopplayerbysuccessrateavg
fgyeartopsravg <- aggregate(fgyeartopsr[, 3:8], list(fgyeartopsr$Year), sum)
colnames(fgyeartopsravg)[1] <- "Year"
fgyeartopsravg$pctfg3 <- fgyeartopsravg$fg3m / fgyeartopsravg$fg3a * 100
fgyeartopsravg$pctfg2 <- fgyeartopsravg$fgm / fgyeartopsravg$fga * 100
fgyeartopsravg$pctft <- fgyeartopsravg$ftm / fgyeartopsravg$fta * 100

# fgyeartopplayerbymade
fgyeartopm <-
  fgyearplayer %>%
  group_by(Year) %>%
  filter(fg3m >= 29) %>%
  mutate(Rank = order(order(fg3m, decreasing = TRUE))) %>%
  filter(Rank <= 300) %>%
  arrange(Year, Rank)

fgyeartopm
# A tibble: 4,651 x 12
# Groups:   Year [34]
    Year Player   fgm   fga  fg3m  fg3a   ftm   fta pctfg3 pctfg2 pctft
   <int> <chr>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>  <dbl> <dbl>
 1  1986 Larry~   796  1606    82   194   441   492   42.3   49.6  89.6
```

```
 2  1986 Craig~   284   568    73   161    75    86   45.3   50    87.2
 3  1986 World~   652  1428    71   169   379   486   42.0   45.7  78.0
 4  1986 Dale ~   193   470    63   174    59    82   36.2   41.1  72.0
 5  1986 Micha~   274   606    63   163   147   170   38.7   45.2  86.5
 6  1986 Kyle ~   286   592    58   140    73    90   41.4   48.3  81.1
 7  1986 John ~   365   818    45   146   231   297   30.8   44.6  77.8
 8  1986 Norm ~   403   921    42   121   131   162   34.7   43.8  80.9
 9  1986 Leon ~   184   463    41   112   123   155   36.6   39.7  79.4
10  1986 Mike ~   252   544    41   114    42    64   36.0   46.3  65.6
# ... with 4,641 more rows, and 1 more variable: Rank <int>

# fgyeartopplayerbymadeavg
fgyeartopmavg <- aggregate(fgyeartopm[, 3:8], list(fgyeartopm$Year), sum)
colnames(fgyeartopmavg)[1] <- "Year"
fgyeartopmavg$pctfg3 <- fgyeartopmavg$fg3m / fgyeartopmavg$fg3a * 100
fgyeartopmavg$pctfg2 <- fgyeartopmavg$fgm / fgyeartopmavg$fga * 100
fgyeartopmavg$pctft <- fgyeartopmavg$ftm / fgyeartopmavg$fta * 100

# fgyeartop <- left_join (fgyeartopsr, fgyeartopm, by=c("Year"="Year", "Player"="Player"))

xaxisbreaks <- seq(1986, 2019, by=3)
yaxisbreaks <- seq(30, 50, by=5)

Q1_4 <- ggplot() +
  geom_line(data=fgyeartopsravg, aes(x=Year, y=pctfg3, colour='By Success Rate', alpha=0.5), size=2) +
  geom_line(data=fgyeartopmavg, aes(x=Year, y=pctfg3, colour='By # of made', alpha=0.5), size=2) +
  guides(alpha=FALSE) +
  xlab('Year') +
  ylab(NULL) +
  ggtitle('3 point success rate of top 30 players') +
  theme(panel.background=element_rect(fill=NA)) +
  theme(panel.grid.major.y=element_line(color="grey", linetype=2)) +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(legend.title=element_blank()) +
  scale_y_continuous(limits=c(30, 50), breaks=yaxisbreaks) +
  scale_x_continuous(limits=c(1986,2019), breaks=xaxisbreaks)

Q1_4
```




```
top9597sr <- fgyeartopsr %>% filter(Year>=1995) %>% filter(Year<=1997)
top9597m <- fgyeartopm %>% filter(Year>=1995) %>% filter(Year<=1997)
```

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

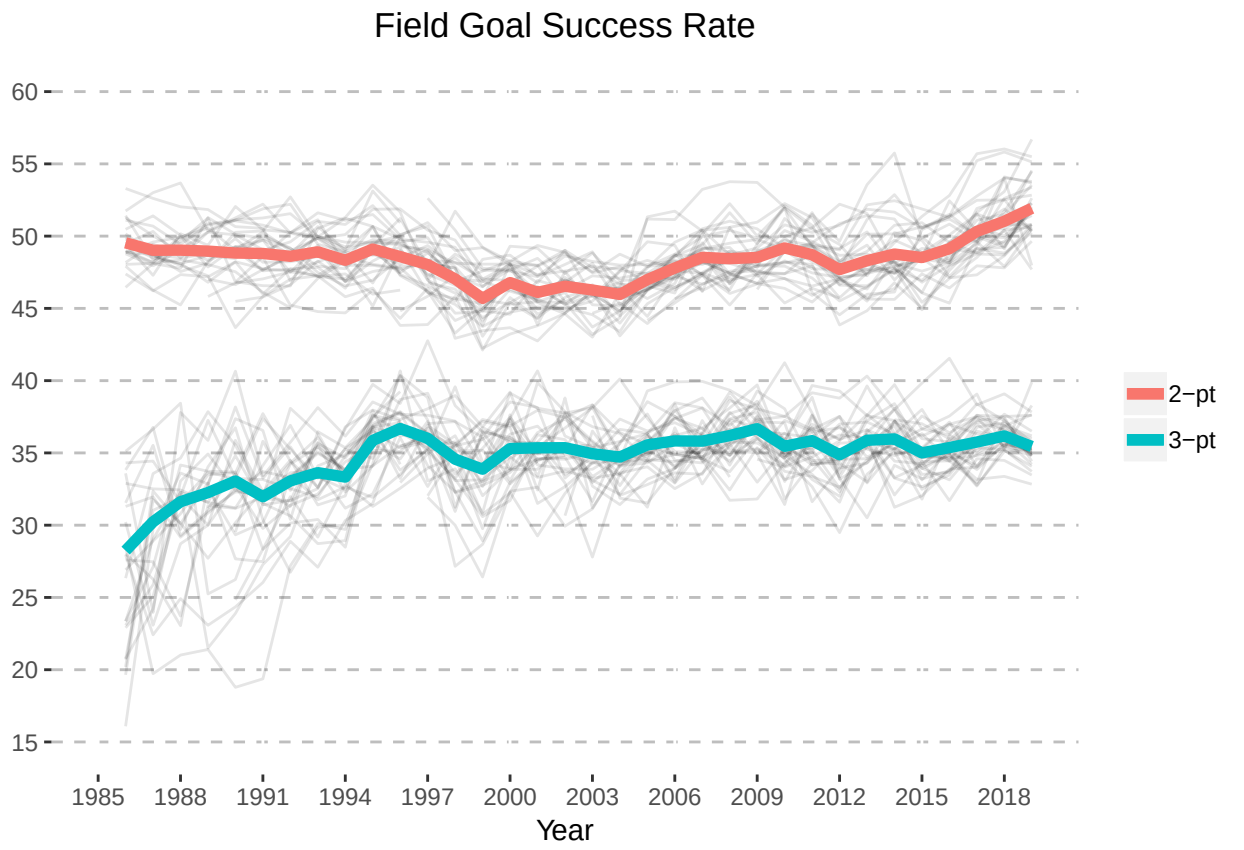
Its debut, in the 1979-80 season, was inauspicious.

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$yearSeason, dataGameLogsTeam$Team), sum)
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, group=Team, alpha=0.5), size=0.5) +
  geom_line(data=fgyearteam, aes(x=Year, y=pctfg2, group=Team, alpha=0.5), size=0.5) +
  geom_line(data=fgyear, aes(x=Year, y=pctfg3, colour="3-pt", alpha=0.9), size=2) +
  geom_line(data=fgyear, aes(x=Year, y=pctfg2, colour="2-pt", alpha=0.9), size=2) +
  guides(alpha=FALSE) +
  xlab('Year') +
  ylab(NULL) +
  ggtitle('Field Goal Success Rate') +
  theme(panel.background=element_rect(fill=NA)) +
  theme(panel.grid.major.y=element_line(color="grey", linetype=2)) +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(legend.title=element_blank()) +
  scale_y_continuous(limits=c(15, 60), breaks=yaxisbreaks, labels=yaxisbreaks) +
  scale_x_continuous(limits=c(1985,2019), breaks=xaxisbreaks)# +

Q2_1
```

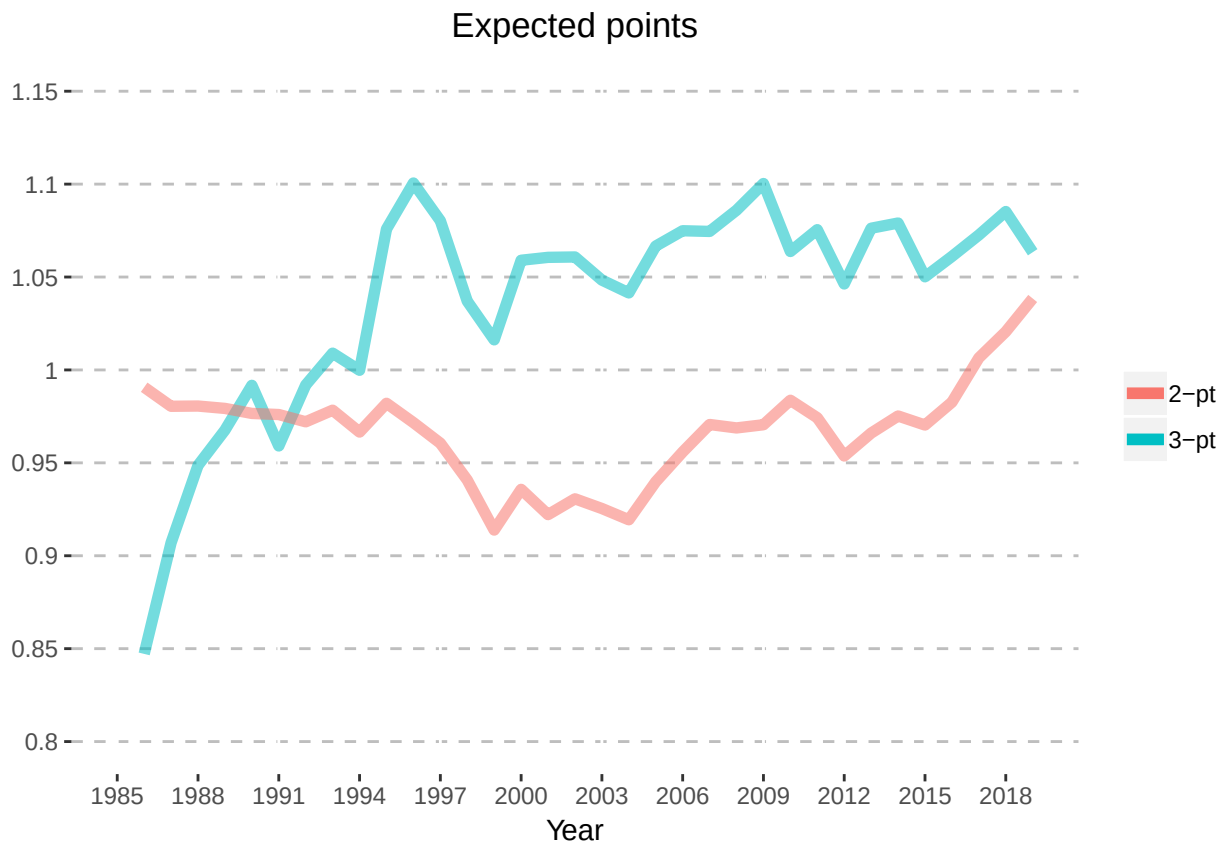


The expected points of 2-point shots in 1986 was 'r fgyear$pctfg2[1986-1985]/100' * 2 =' $rfgyear$pctfg2[1986-1985]/100*2' *The expected points of 3-point shots in 1986 was 'r fgyear$pctfg3[1986-1985]/100' * 3 =' $rfgyear$pctfg3[1986-1985]/100*3*'

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$pctfg3[2019-1985]/100' * 3 =' $rfgyear$pctfg3[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, color="3-pt", alpha=0.9), size=2) +
  geom_line(data=fgyear, aes(x=Year, y=e2, color="2-pt", alpha=0.9), size=2) +
  guides(alpha=FALSE) +
  xlab('Year') +
  ylab(NULL) +
  ggtitle('Expected points') +
  theme(panel.background=element_rect(fill=NA)) +
  theme(panel.grid.major.y=element_line(color="grey", linetype=2)) +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(legend.title=element_blank()) +
  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$yearSeason, dataGameLogsTeam$nameTeam), sum)
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"))

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

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
```

Yes. However, pctfg2 is more relevant than pctfg3

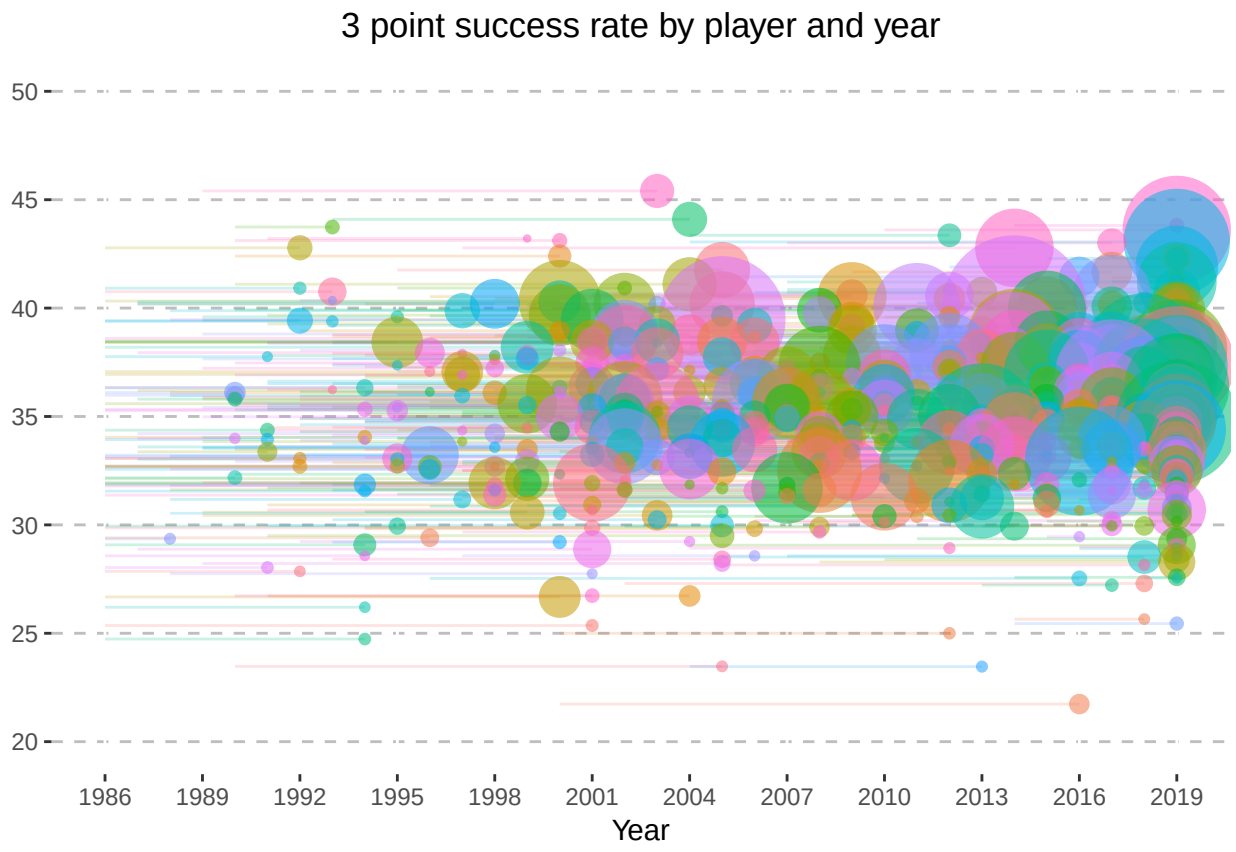
- Focus on three point shooting is a strategy that started fairly recently, we can create a map to show where this strategy initially emerged and how fast it spreaded across the entire country.

Player level questions

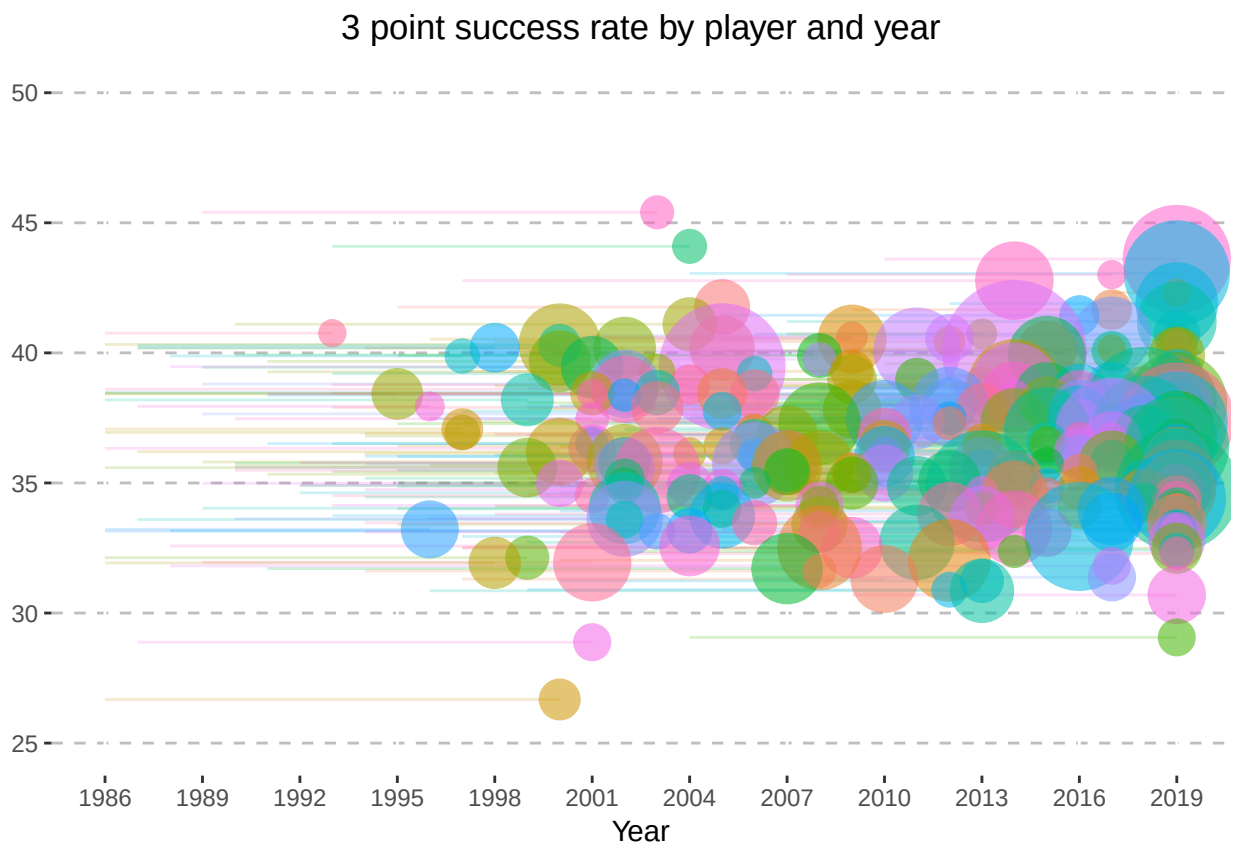
```
dataGameLogsPlayer1986 <- dataGameLogsPlayer %>% filter(yearSeason >= 1986)

fgyearplayer <- aggregate(dataGameLogsPlayer1986[, 19:26], list(dataGameLogsPlayer1986$yearSeason, dataGameLogsPlayer1986$namePl
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

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

```
# fgplayer <- py$fgplayer
load("fgplayer.Rdata")

fgplayer100 <- fgplayer %>% filter(fg3m >= 100)
fgplayer500 <- fgplayer %>% filter(fg3m >= 500)
fgplayer1000 <- fgplayer100 %>% filter(fg3m >= 1000)
fgplayer2000 <- fgplayer1000 %>% filter(fg3m >= 2000)

xaxisbreaks <- seq(1986, 2019, by=3)
yaxisbreaks <- seq(20, 50, by=5)

career100 <- melt(data=fgplayer100, id.var=c("Player","pctfg3"), measure.vars=c("firstYear", "lastYear"))

plotPlayer100 <- ggplot() +
  geom_line(data=career100, aes(x=value, y=pctfg3, color=Player), show.legend=FALSE, alpha=0.2) +
  geom_point(data=fgplayer100, aes(x=lastYear, y=pctfg3, colour=Player, alpha=0.8),
             size=fgplayer100$fg3a/300, show.legend=FALSE) +
  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(20, 50), breaks=yaxisbreaks, labels=yaxisbreaks) +
  scale_x_continuous(limits=c(1986,2019), breaks=xaxisbreaks)
plotPlayer100
```

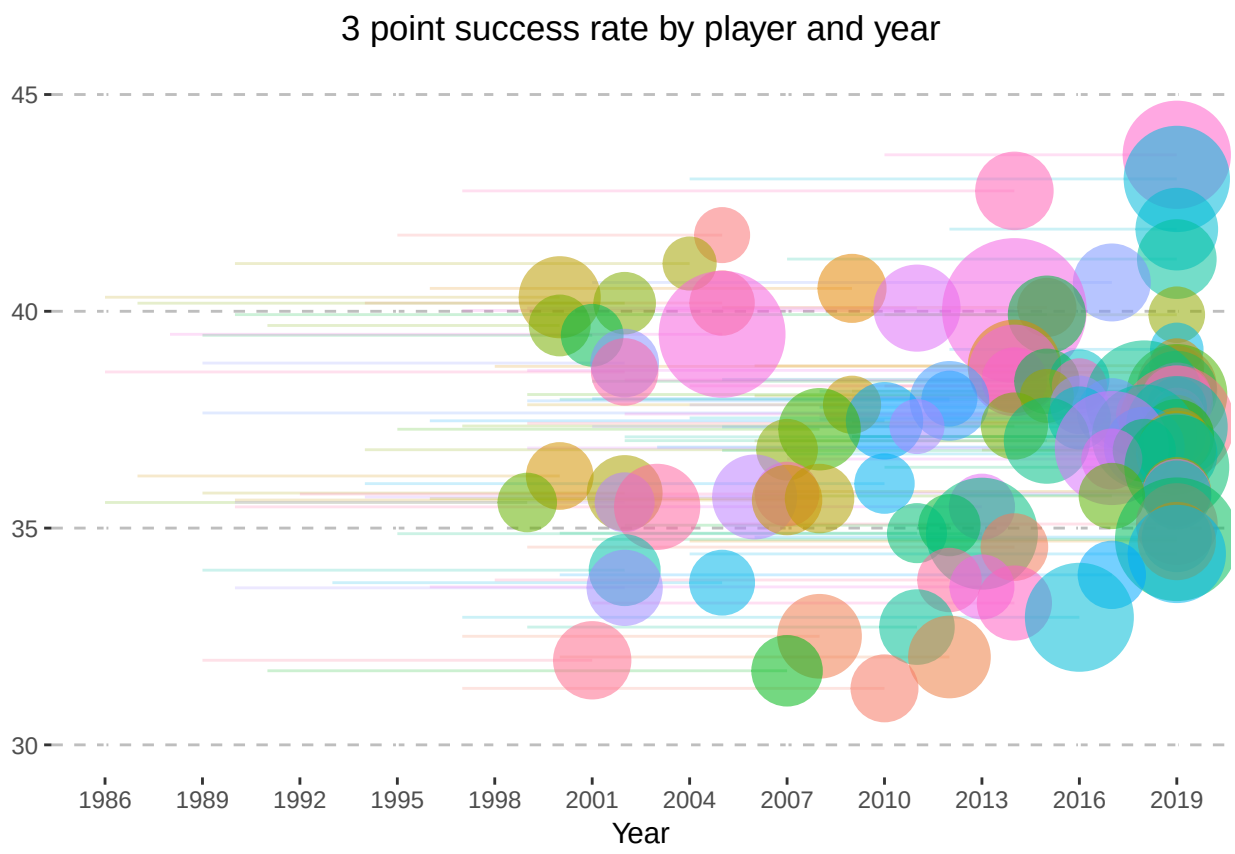


```
xaxisbreaks <- seq(1986, 2019, by=3)
yaxisbreaks <- seq(25, 50, by=5)

career500 <- melt(data=fgplayer500, id.var=c("Player","pctfg3"), measure.vars=c("firstYear", "lastYear"))

plotPlayer500 <- ggplot() +
  geom_line(data=career500, aes(x=value, y=pctfg3, color=Player), show.legend=FALSE, alpha=0.2) +
  geom_point(data=fgplayer500, aes(x=lastYear, y=pctfg3, colour=Player, alpha=0.8),
             size=fgplayer500$fg3a/300, show.legend = FALSE) +
  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(25, 50), breaks=yaxisbreaks, labels=yaxisbreaks) +
  scale_x_continuous(limits=c(1986,2019), breaks=xaxisbreaks) +
  geom_hline()
plotPlayer500
```




```
xaxisbreaks <- seq(1986, 2019, by=3)
yaxisbreaks <- seq(30, 45, by=5)

career1000 <- melt(data=fgplayer1000, id.var=c("Player","pctfg3"), measure.vars=c("firstYear", "lastYear"))

plotPlayer1000 <- ggplot() +
  geom_line(data=career1000, aes(x=value, y=pctfg3, color=Player), show.legend=FALSE, alpha=0.2) +
  geom_point(data=fgplayer1000, aes(x=lastYear, y=pctfg3, colour=Player, alpha=0.8),
             size=fgplayer1000$fg3a/300, show.legend = FALSE) +
  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(1986,2019), breaks=xaxisbreaks)
plotPlayer1000
```



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
```

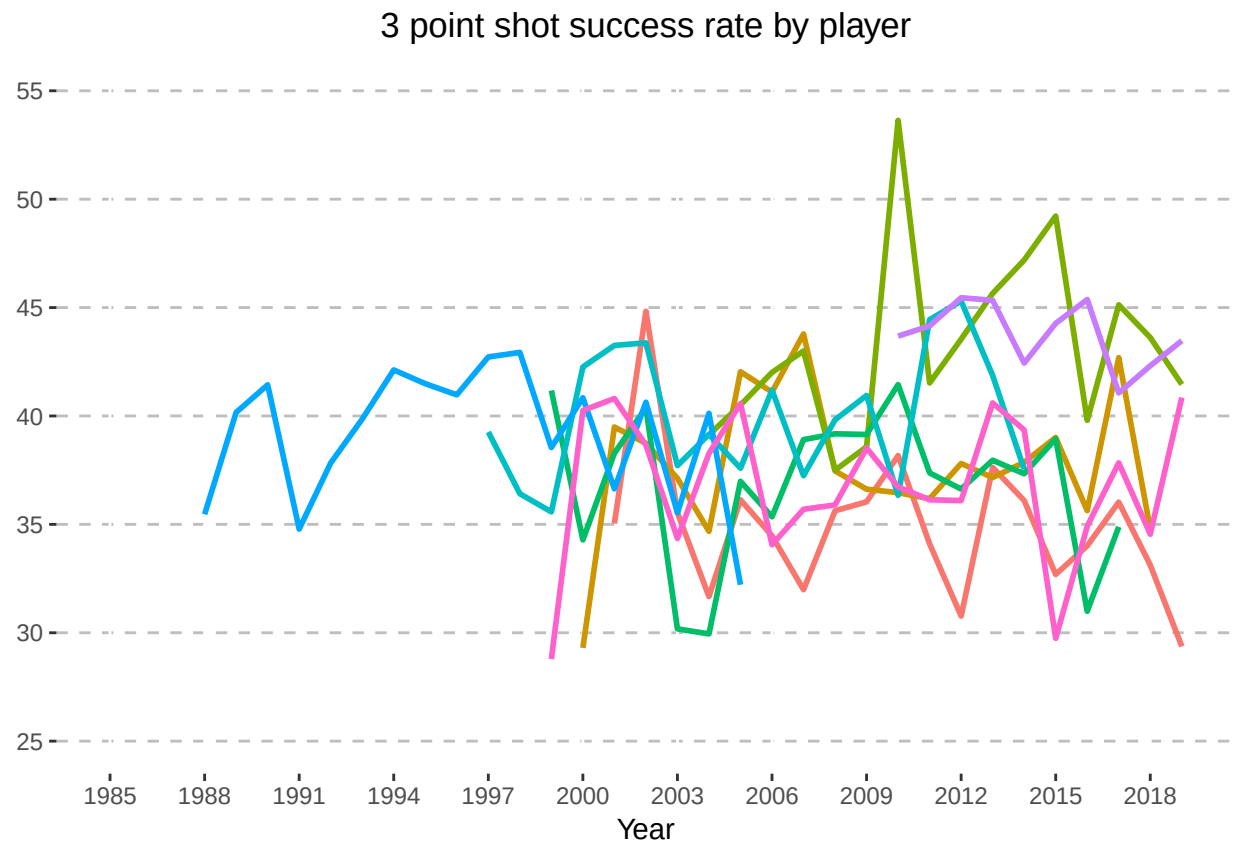
Best players (more than 1,000 career 3-point field goals) are good at 2-pointers as well!!!

```
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.

```
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(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", 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=2),
        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, return_message = TRUE)

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(!(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
```