

FPL points predictor Exploratory Data Analysis

This is a file for conducting explorative data analysis of the premier league players' dataset

Importing the datasets

```
season17 <- read.csv("~/DSI-SRP1/season17.csv", encoding="UTF-8")
season18 <- read.csv("~/DSI-SRP1/season18.csv", encoding="UTF-8")
season19 <- read.csv("~/DSI-SRP1/season19.csv", encoding="UTF-8")
```

Loading the tidyverse package

```
library(tidyverse)

## -- Attaching packages ----- tidyverse 1.3.0 --

## v ggplot2 3.3.0      v purrr  0.3.4
## v tibble  3.0.1      v dplyr  0.8.5
## v tidyr   1.0.3      v stringr 1.4.0
## v readr   1.3.1      v forcats 0.5.0

## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Calculating correlation coefficients between goal scored and total FPL points

```
message("Correlation between goal scored and total fpl points for 2016/17: ",
        with(season17, cor(goals_scored, total_points)))

## Correlation between goal scored and total fpl points for 2016/17: 0.700148307138786

message("Correlation between goal scored and total fpl points for 2017/18: ",
        with(season18, cor(goals_scored, total_points)))

## Correlation between goal scored and total fpl points for 2017/18: 0.679586851190284
```

```
message("Correlation between goal scored and total fpl points for 2018/19: ",
       with(season19, cor(goals_scored, total_points)))
```

```
## Correlation between goal scored and total fpl points for 2018/19: 0.703249889002981
```

Changing the datatype of the position_index to factor

```
season17$position_index <- as_factor(season17$position_index)
season18$position_index <- as_factor(season18$position_index)
season19$position_index <- as_factor(season19$position_index)
```

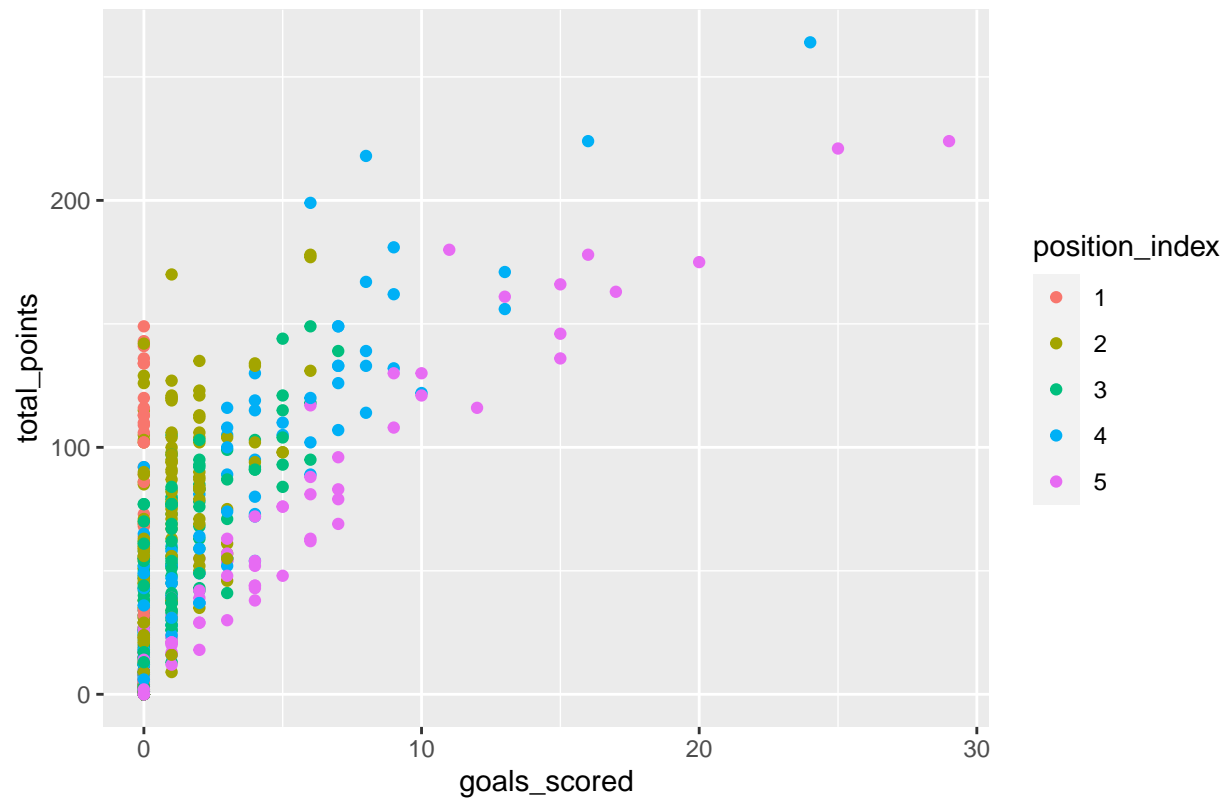
Creating a function to calculate a new metric called total fpl points to game played

```
fpl_to_gaming <- function(df) {
  df %>%
    mutate(fpl_to_game = (total_points/minutes.played)*90)
}
season17 <- fpl_to_gaming(season17)
season18 <- fpl_to_gaming(season18)
season19 <- fpl_to_gaming(season19)
```

Scatter plots graph to highlight how the goals scored vary with fpl points

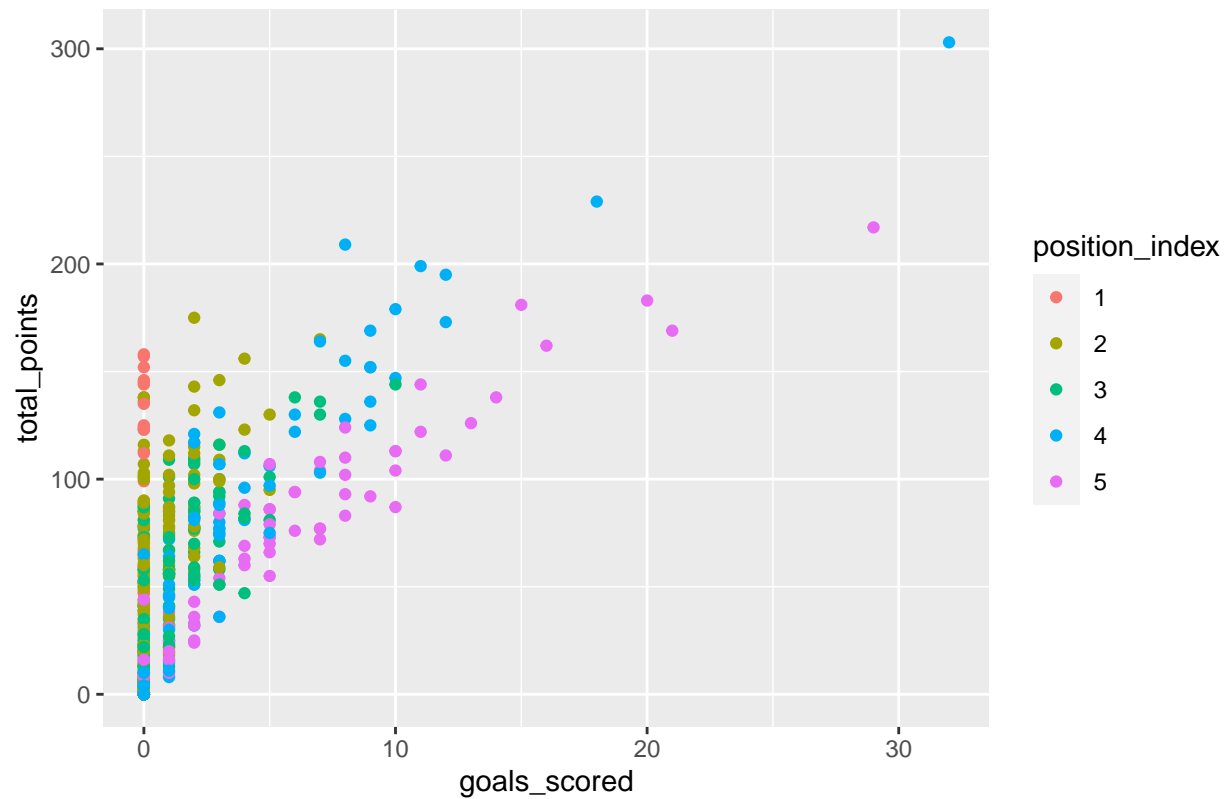
```
ggplot(season17, aes(goals_scored, total_points)) +
  geom_point(aes(color = position_index)) +
  labs(title = "Total FPL points vs Goal Scored in 2016/17")
```

Total FPL points vs Goal Scored in 2016/17



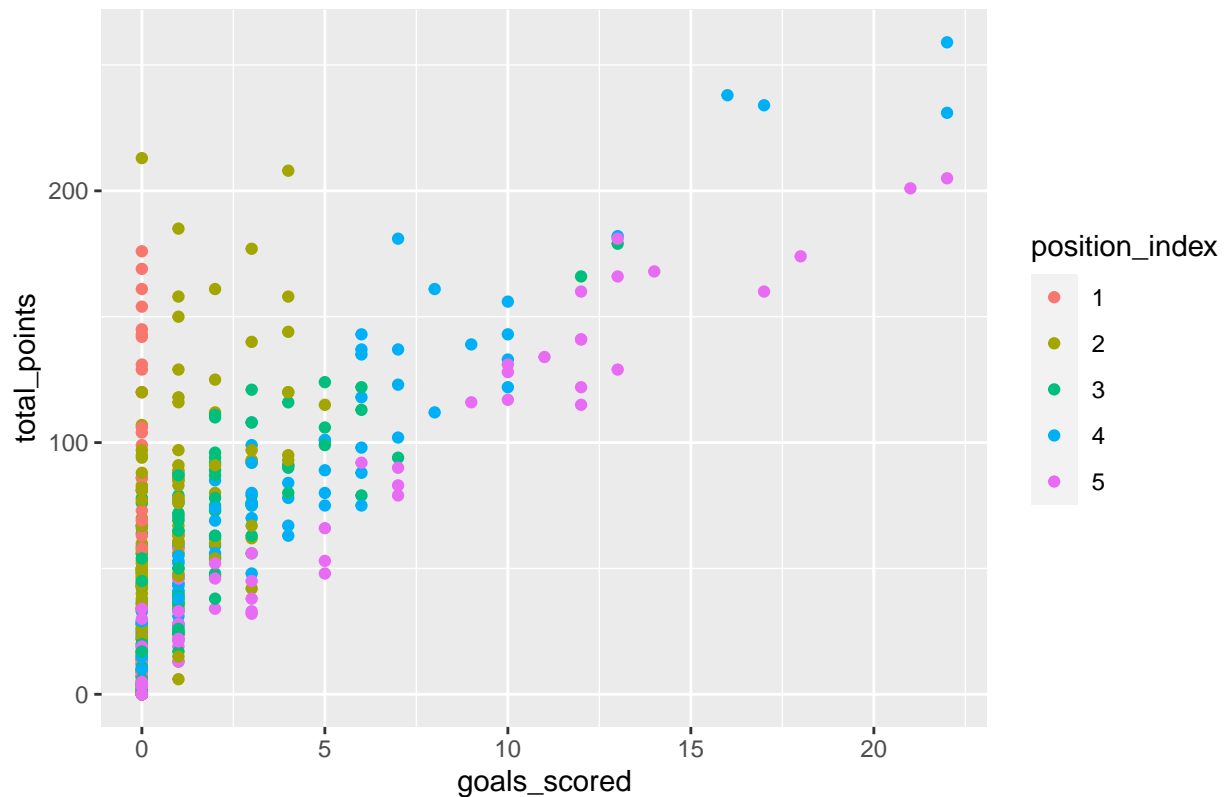
```
ggplot(season18, aes(goals_scored, total_points)) +  
  geom_point(aes(color = position_index)) +  
  labs(title = "Total FPL points vs Goal Scored in 2017/18")
```

Total FPL points vs Goal Scored in 2017/18



```
ggplot(season19, aes(goals_scored, total_points)) +  
  geom_point(aes(color = position_index)) +  
  labs(title = "Total FPL points vs Goal Scored in 2018/19")
```

Total FPL points vs Goal Scored in 2018/19



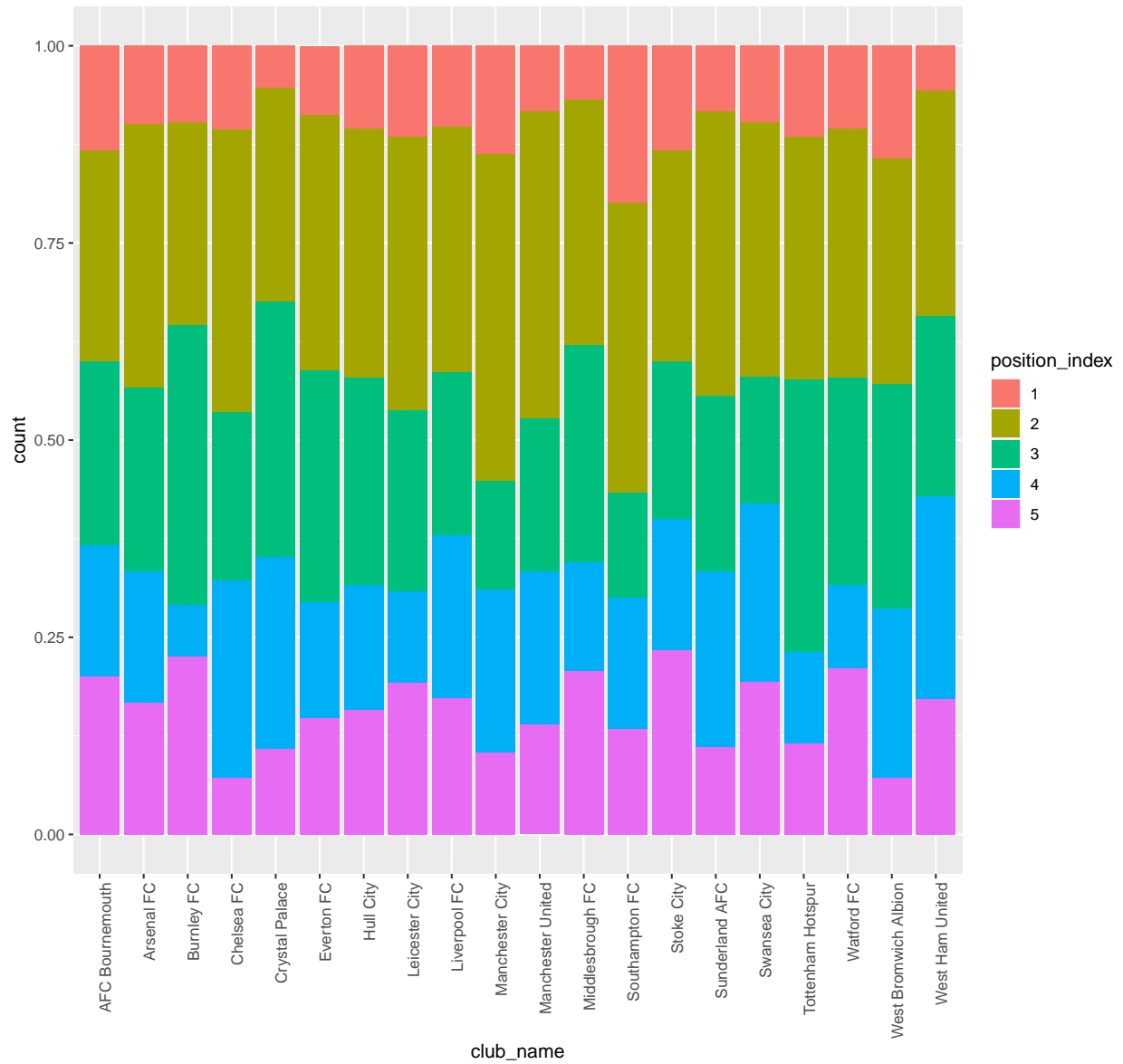
Bar graphs to show how the number of players vary per team with regards to position

```
position_bar <- function(df) {
  df %>%
    ggplot(aes(club_name)) +
    geom_bar(aes(fill = position_index), position = "fill") +
    theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
    labs(title = sprintf("How players' positions varied across teams in
                          %s", unique(df$season)))
}
```

Plotting the bar charts

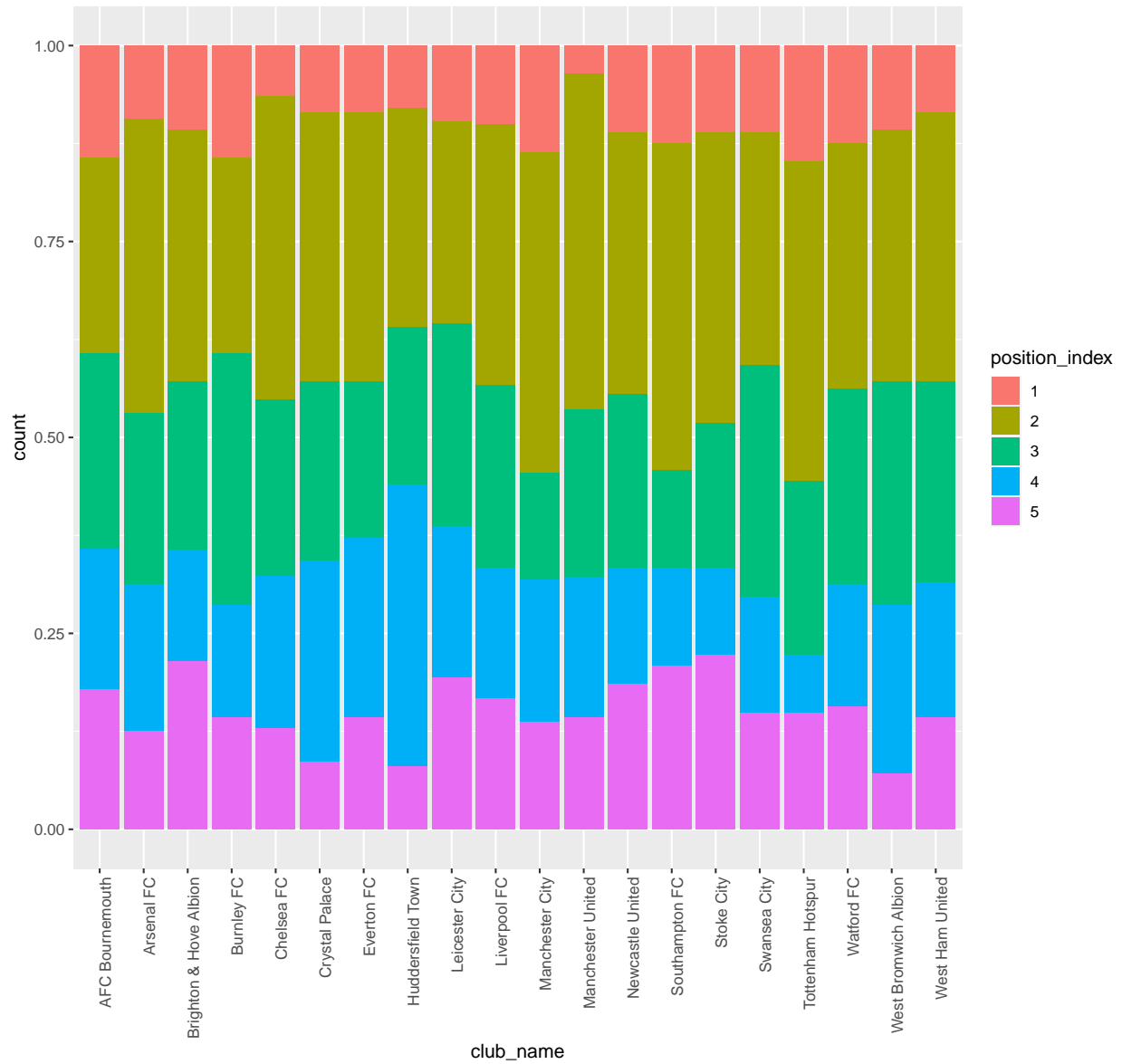
```
position_bar(season17)
```

How players' positions varied across teams in
2016/2017

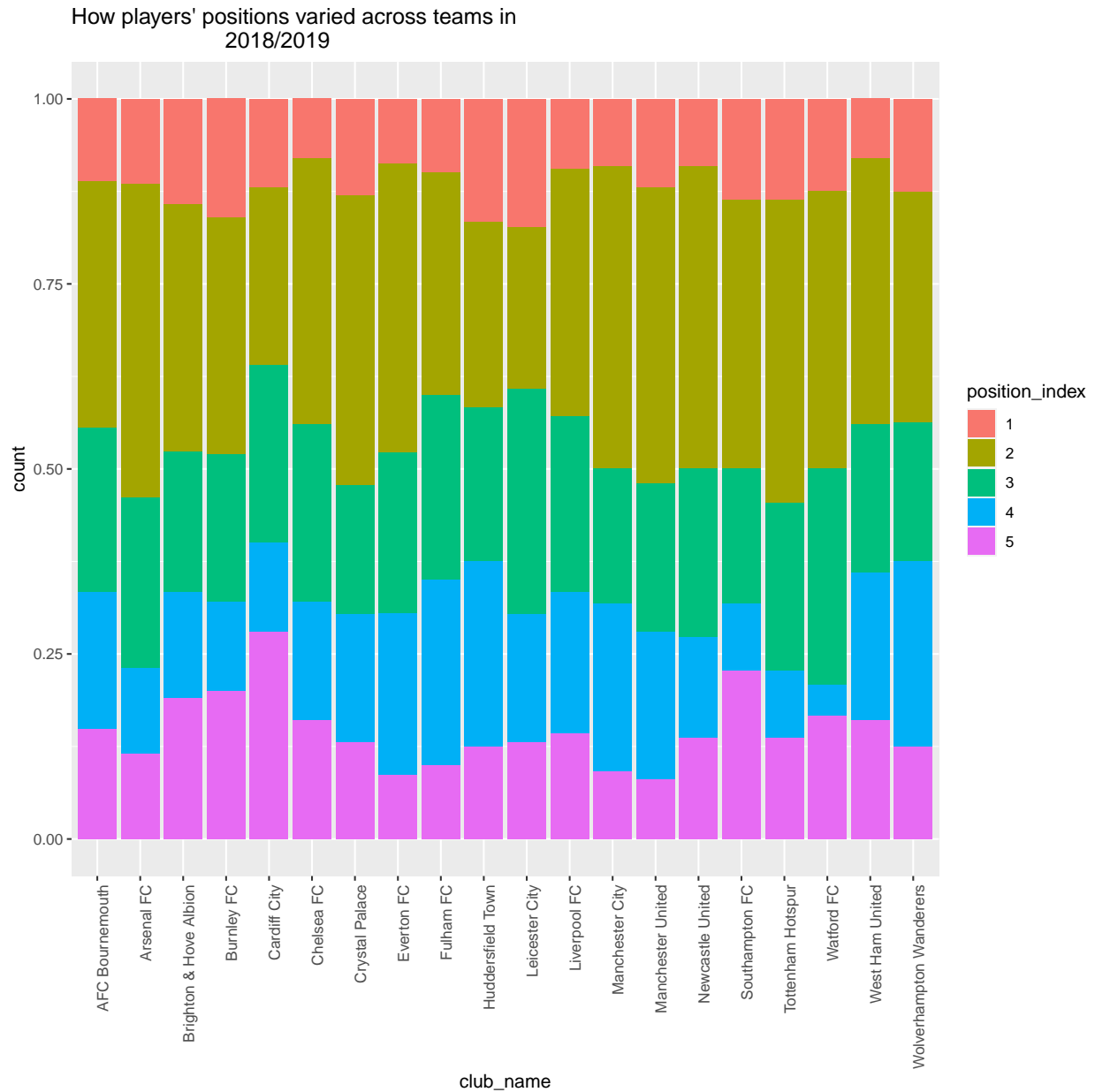


`position_bar(season18)`

How players' positions varied across teams in
2017/2018



`position_bar(season19)`

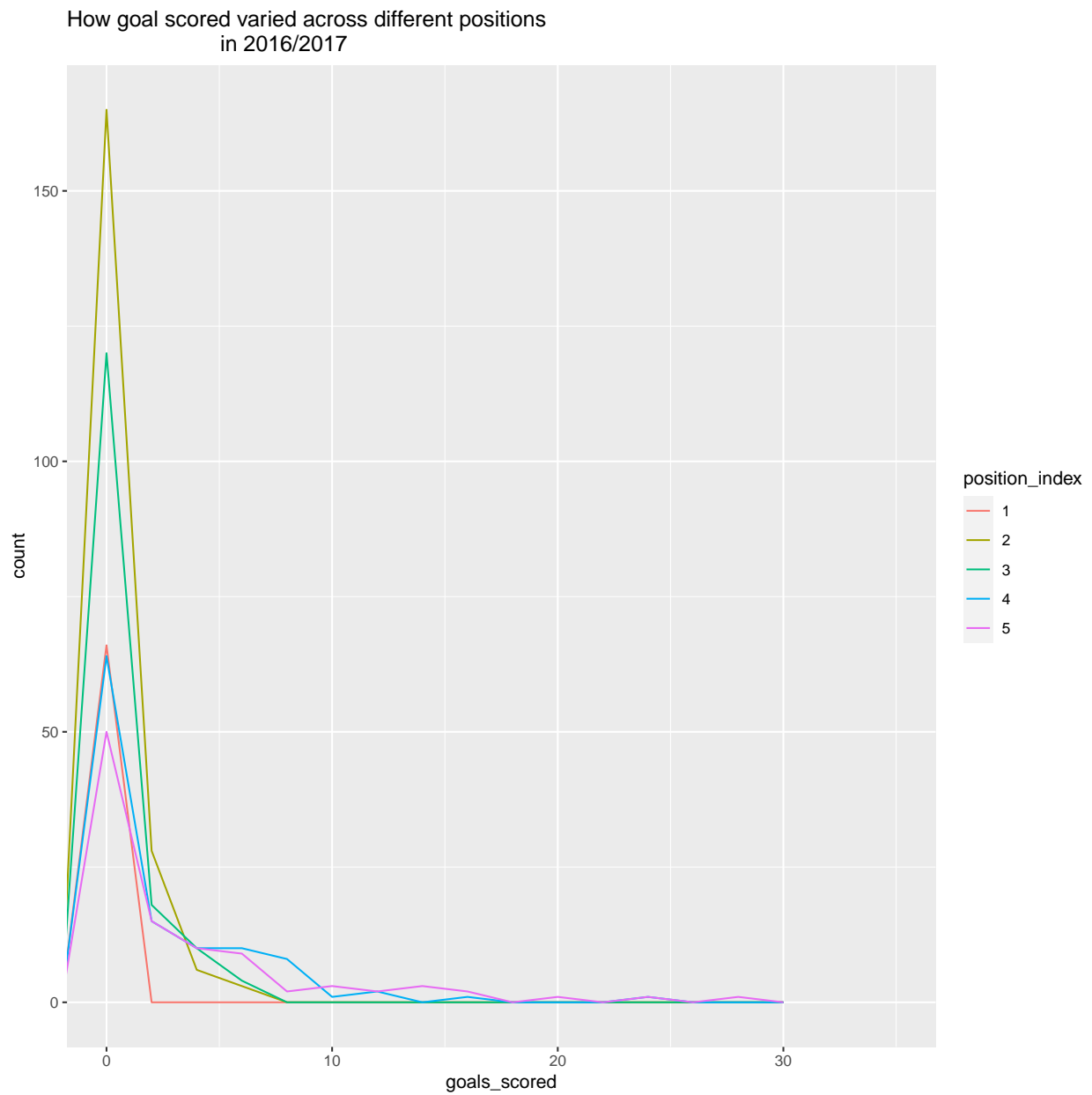


Creating frequency plot for goal scored and how it varies for the players' position

```
goal_freqplot <- function(df) {
  df %>%
    ggplot(aes(x = goals_scored)) +
    geom_freqpoly(aes(color = position_index), binwidth = 2) +
    labs(title = sprintf("How goal scored varied across different positions
                          in %s", unique(df$season))) +
    coord_cartesian(xlim = c(0,35))
}
```

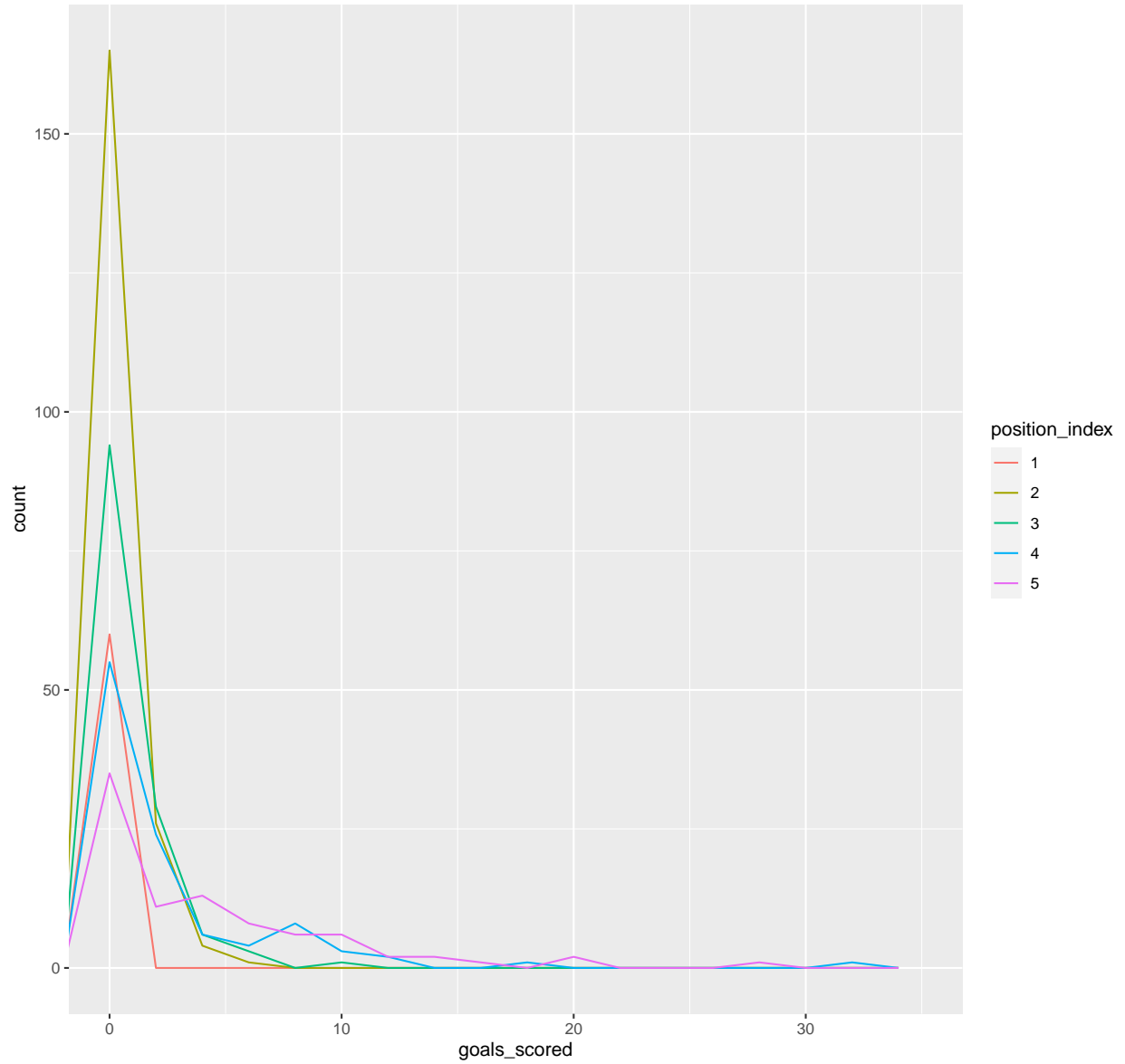

Plotting the graphs

```
goal_freqplot(season17)
```



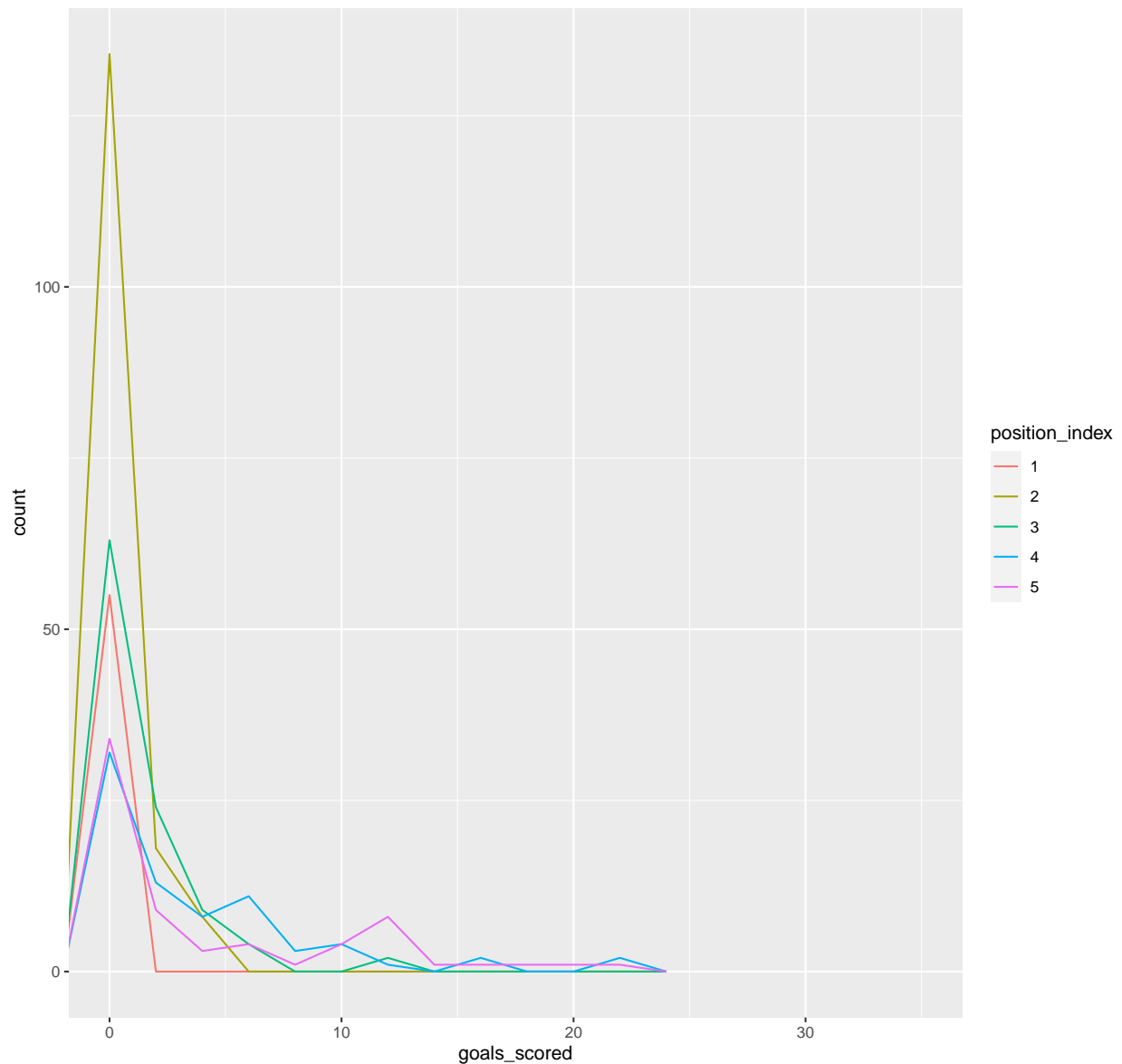
```
goal_freqplot(season18)
```

How goal scored varied across different positions
in 2017/2018



```
goal_freqplot(season19)
```

How goal scored varied across different positions
in 2018/2019

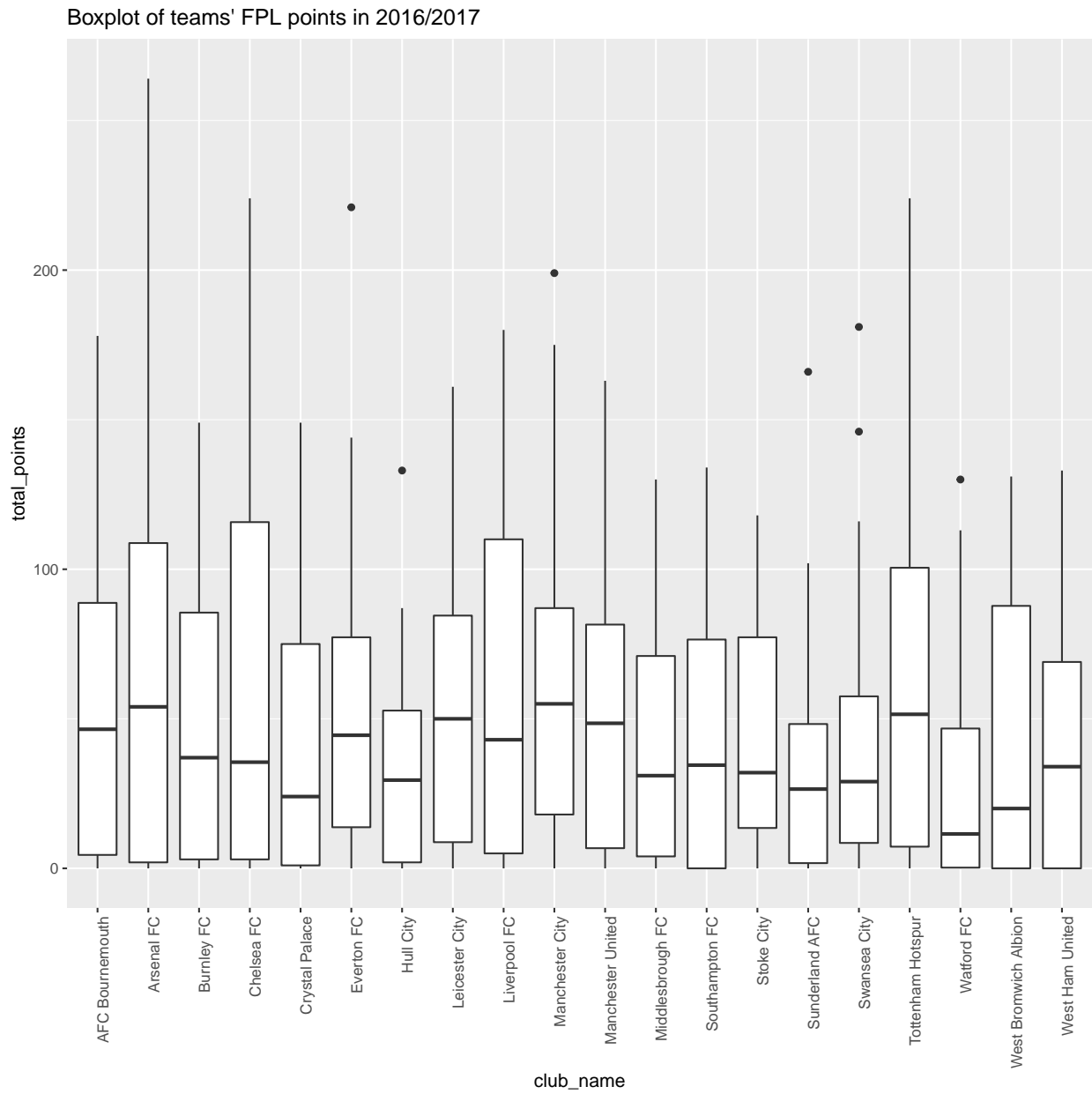


Creating boxplots to show the distribution of the teams' fpl points according to their players

```
boxplot_fpl <- function(df) {
  df %>%
    ggplot() +
    geom_boxplot(aes(x = club_name, y = total_points), na.rm = TRUE) +
    theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
    labs(title = sprintf("Boxplot of teams' FPL points in '%s'",
                        unique(df$season)))
}
```

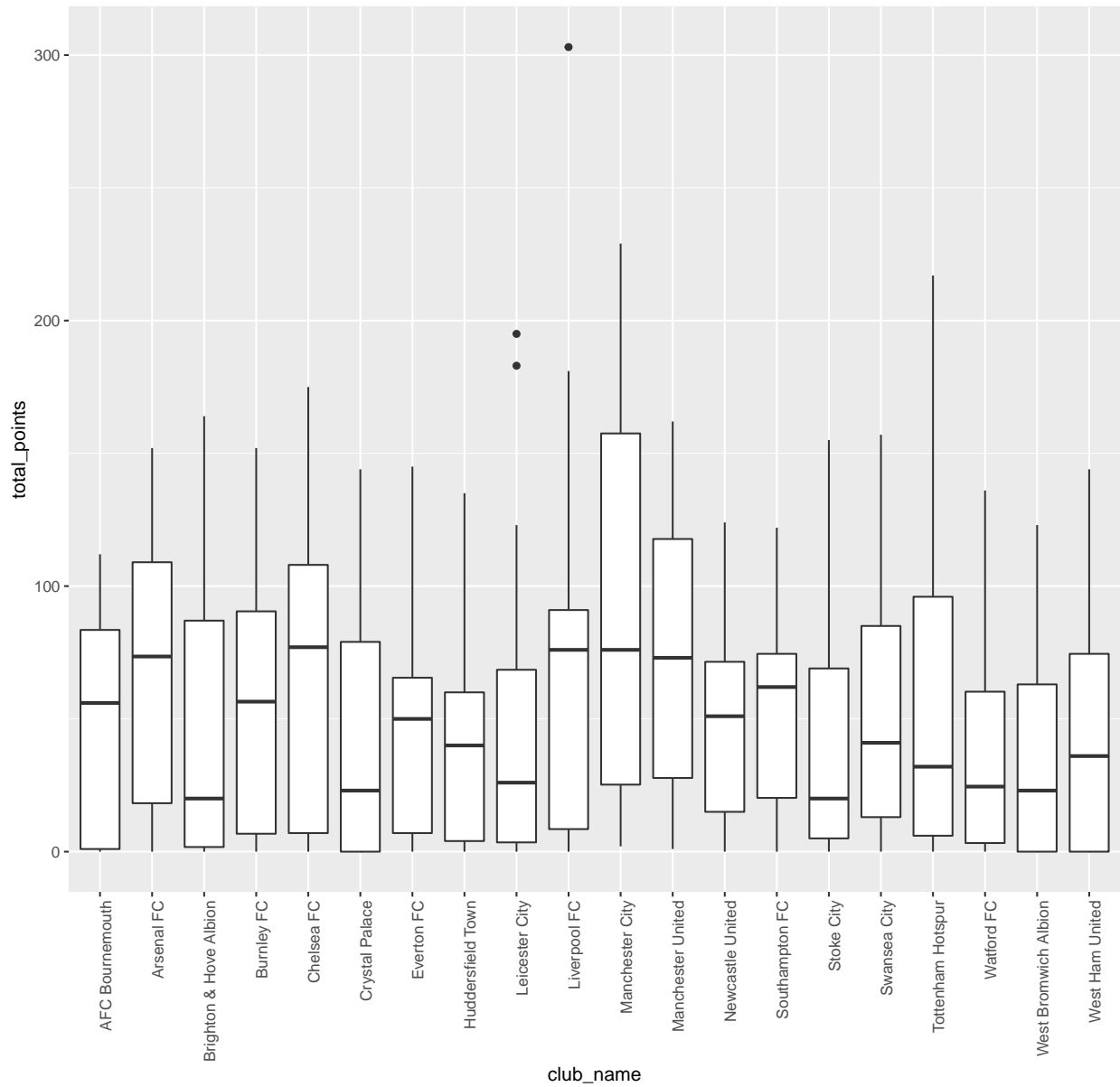
Plotting the graphs

```
boxplot_fpl(season17)
```



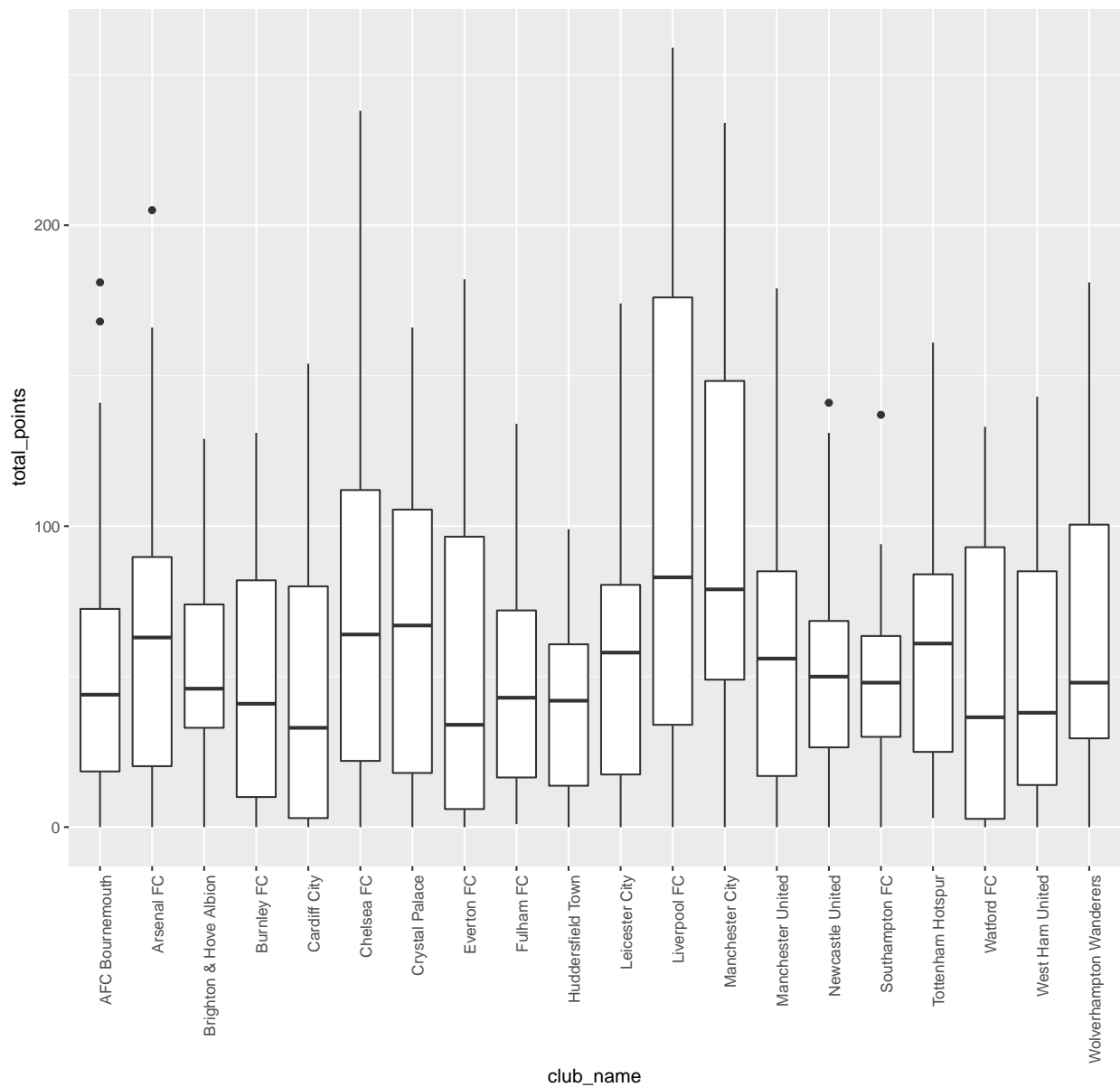
```
boxplot_fpl(season18)
```

Boxplot of teams' FPL points in 2017/2018



```
boxplot_fpl(season19)
```

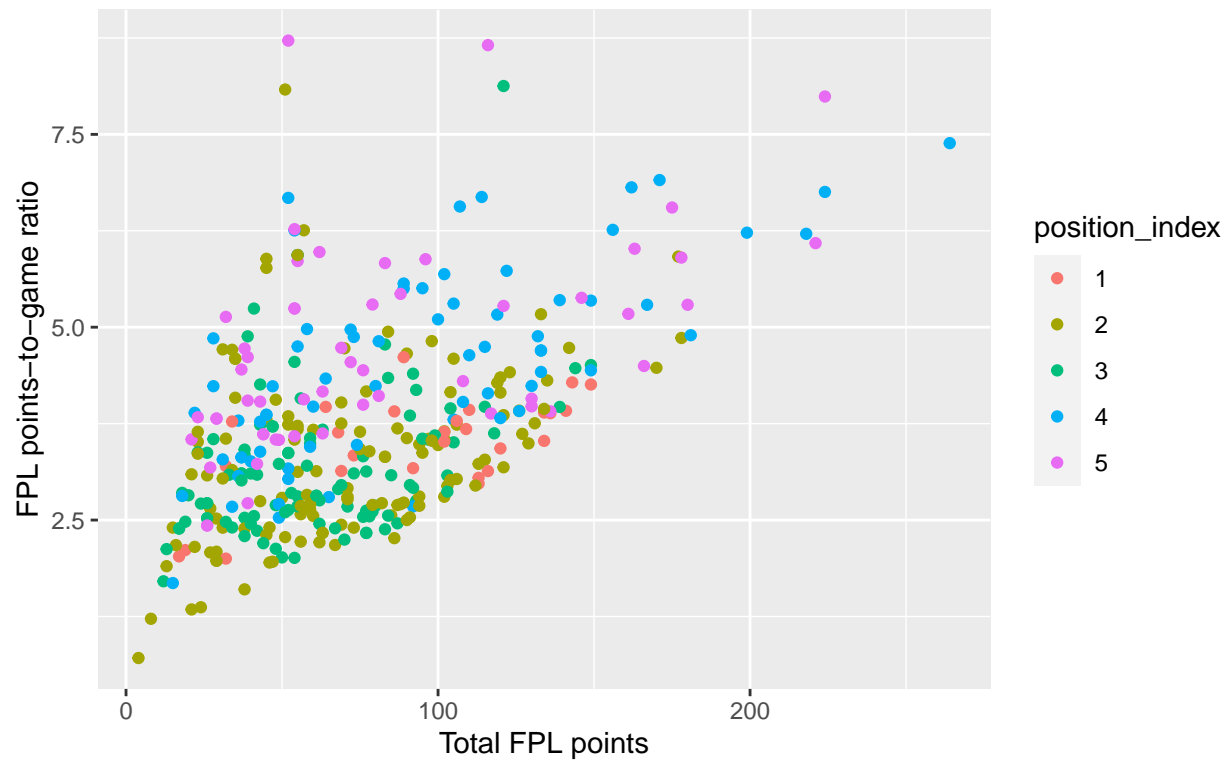
Boxplot of teams' FPL points in 2018/2019



Scatter plots showing the FPL points to game ratio vs the total FPL points for each season

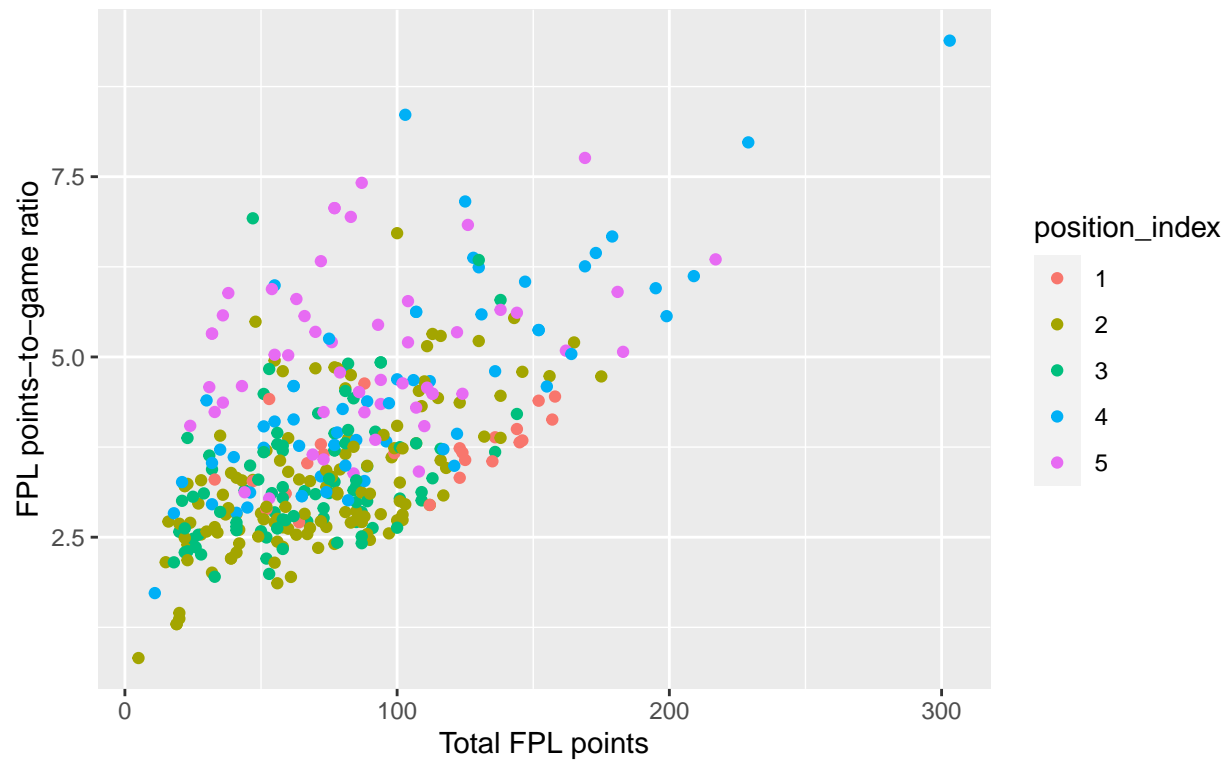
```
season17 %>%
  filter(minutes.played >= 500) %>%
  ggplot() +
  geom_point(aes(total_points, fpl_to_game, color = position_index), na.rm = TRUE) +
  labs(x = "Total FPL points", y = "FPL points-to-game ratio",
  title = title1)
```

FPL points to game ratio vs Total FPL points for players who have played more than 500 minutes in 2016/17



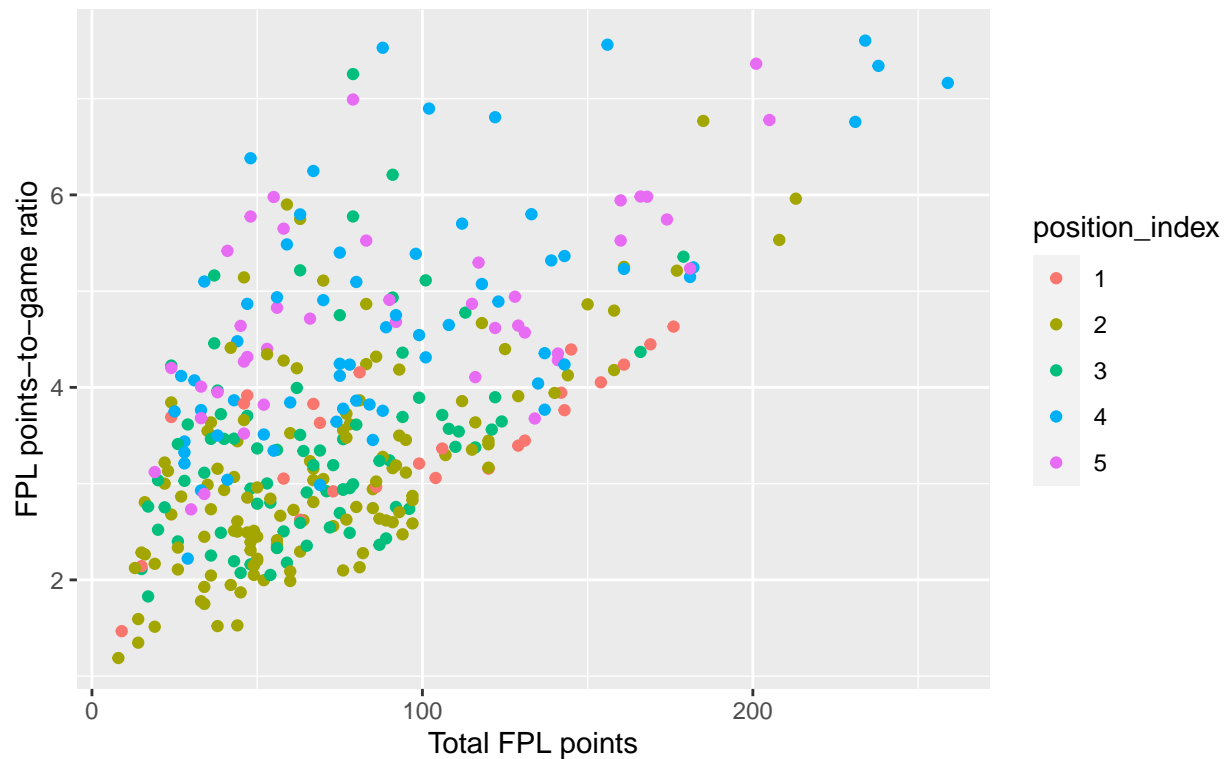
```
season18 %>%  
  filter(minutes.played >= 500) %>%  
  ggplot() +  
  geom_point(aes(total_points, fpl_to_game, color = position_index), na.rm = TRUE) +  
  labs(x = "Total FPL points", y = "FPL points-to-game ratio",  
       title = title2)
```

FPL points to game ratio vs Total FPL points for players who have played more than 500 minutes in 2017/18



```
season19 %>%  
  filter(minutes.played >= 500) %>%  
  ggplot() +  
  geom_point(aes(total_points, fpl_to_game, color = position_index), na.rm = TRUE) +  
  labs(x = "Total FPL points", y = "FPL points-to-game ratio",  
       title = title3)
```


FPL points to game ratio vs Total FPL points for players who have played more than 500 minutes in 2018/19

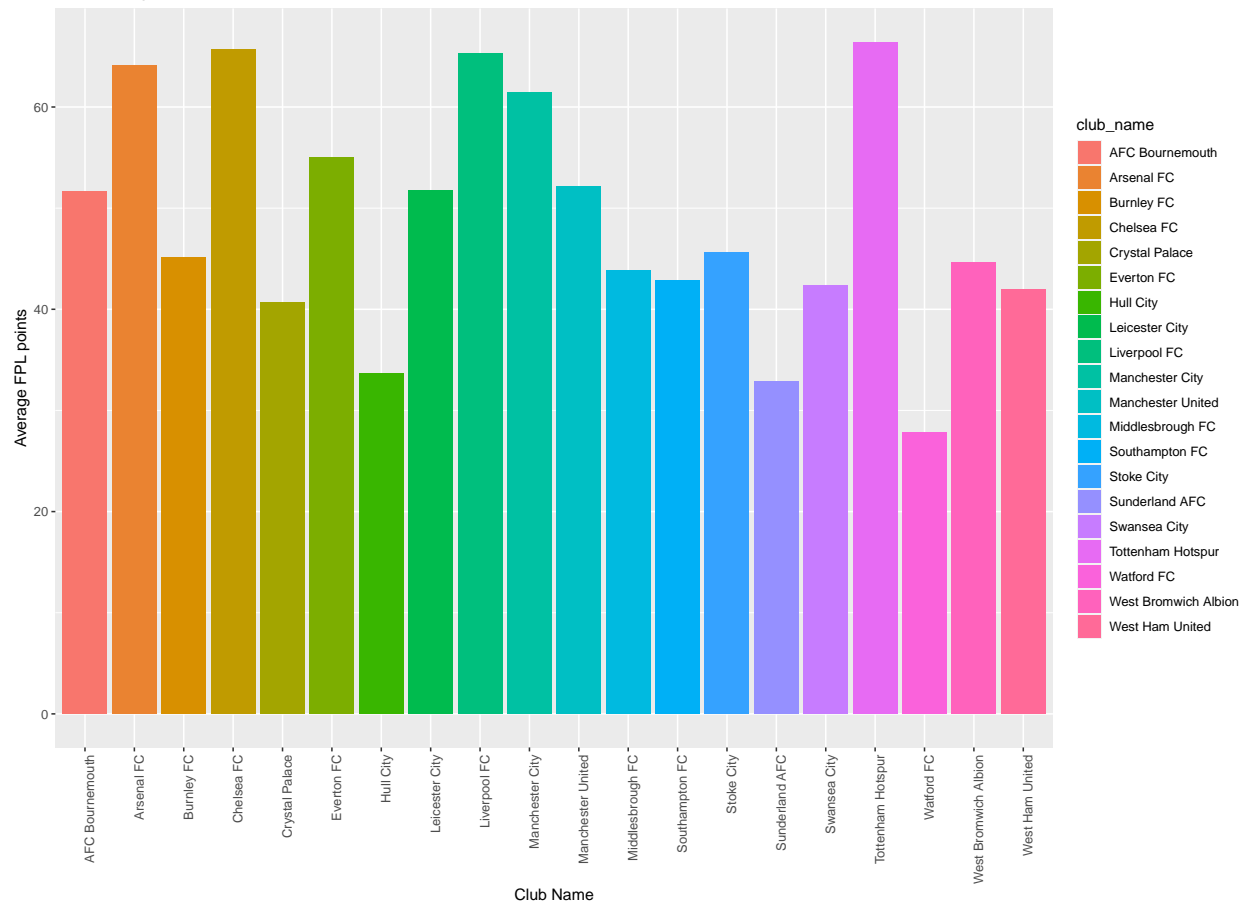


Calculating the average fpl points per team and building a bar chart to show the average fpl points per team

```
avg_fpl_point <- function(df) {
  df %>%
    group_by(club_name) %>%
    summarize(average = mean(total_points, na.rm = TRUE)) %>%
    ggplot() +
    geom_bar(aes(x = club_name, y = average, fill = club_name), stat = "identity") +
    theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
    labs(x = "Club Name", y = "Average FPL points",
    title = sprintf("The average FPL points achieved by each team in the season %s",
    unique(df$season)))
}
```

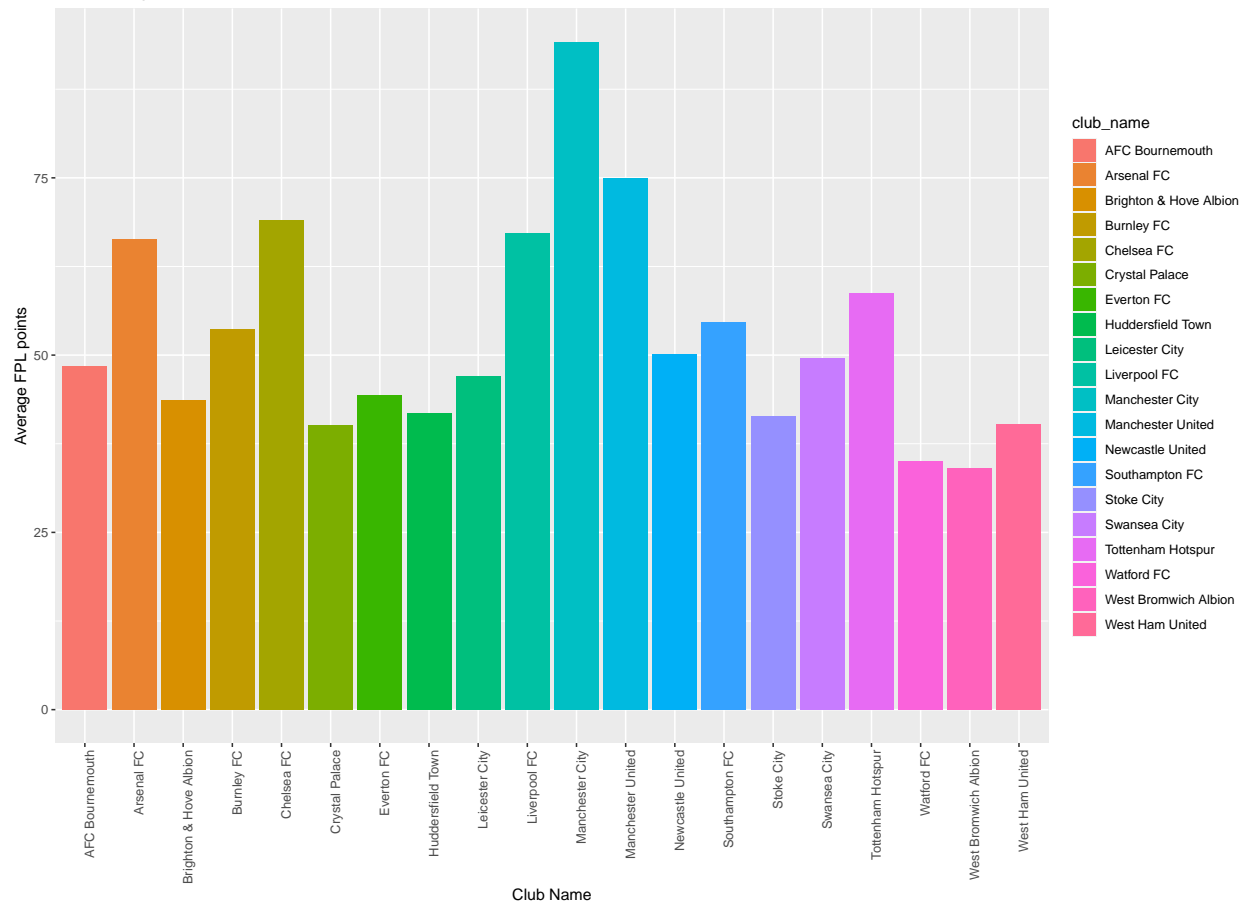
```
avg_fpl_point(season17)
```

The average FPL points achieved by each team in the season 2016/2017

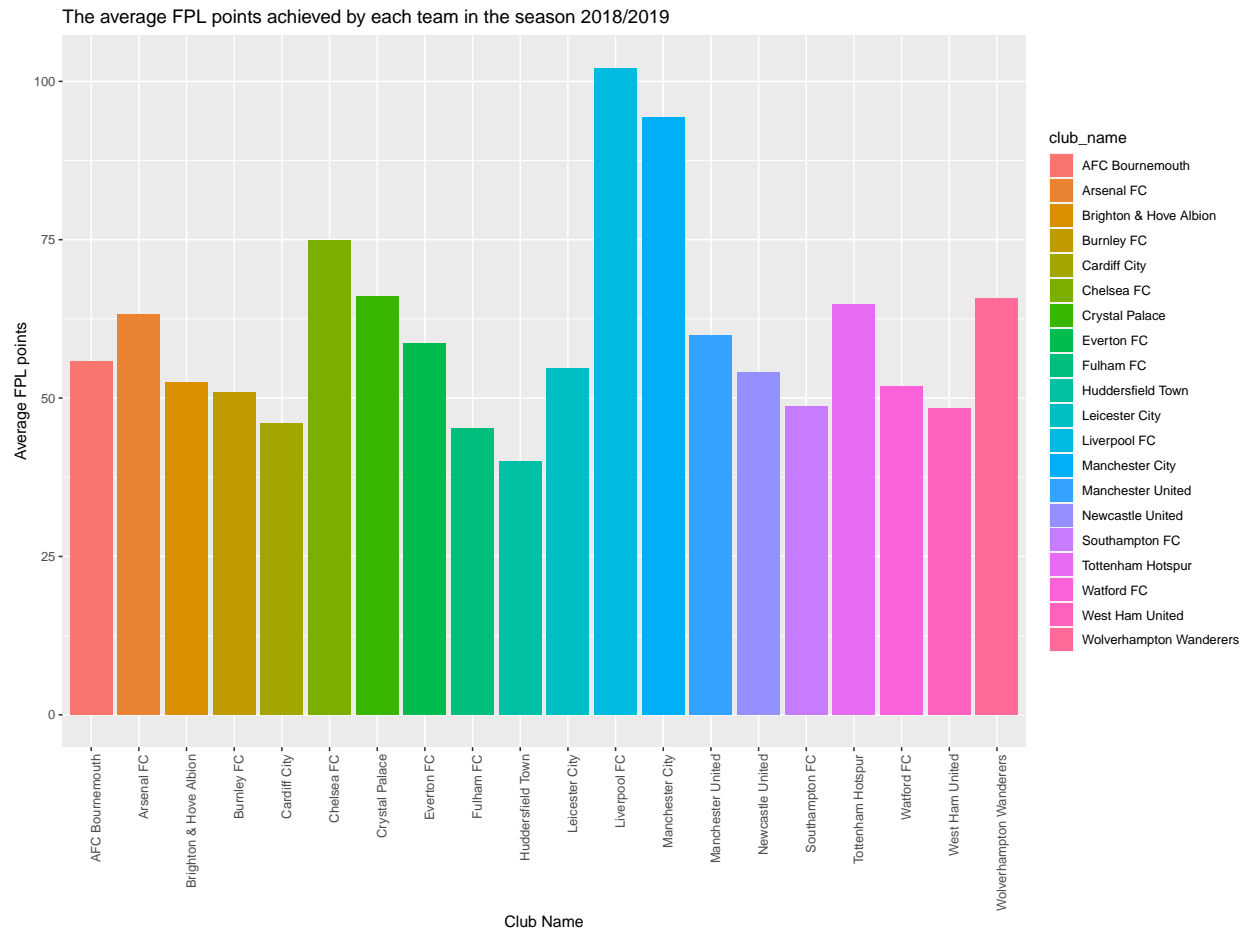


```
avg_fpl_point(season18)
```

The average FPL points achieved by each team in the season 2017/2018



```
avg_fpl_point(season19)
```

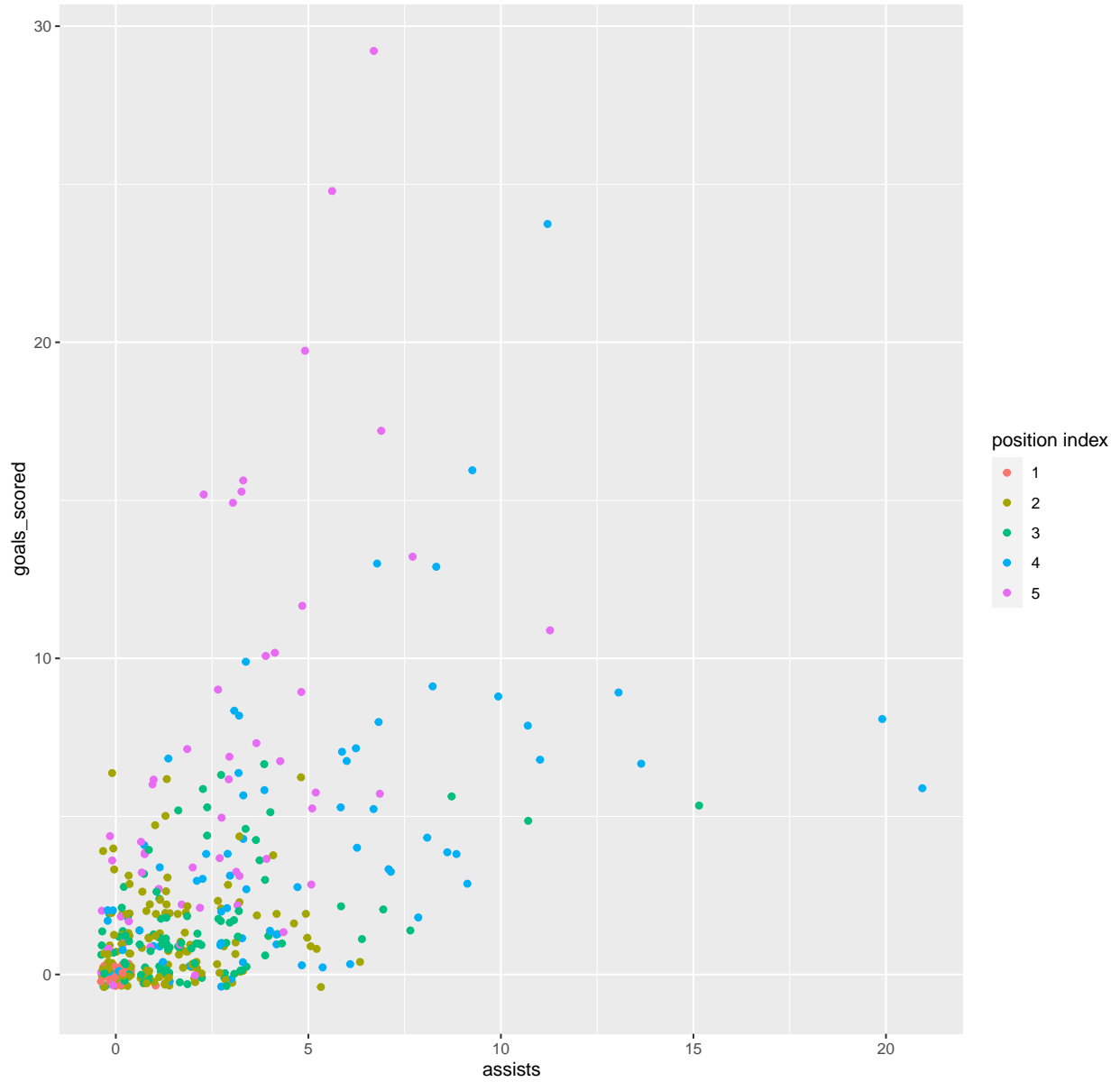


Plotting graphs of goals scored vs assists made for every season

```
goals_vs_assist <- function(df) {
  df %>%
    filter(minutes.played >= 500) %>%
    ggplot(aes(assists, goals_scored)) +
    geom_point(aes(color = position_index), position = "jitter") +
    labs(title = sprintf("Goals scored vs Assists during the %s season",
      unique(df$season)), color = "position index")
}
```

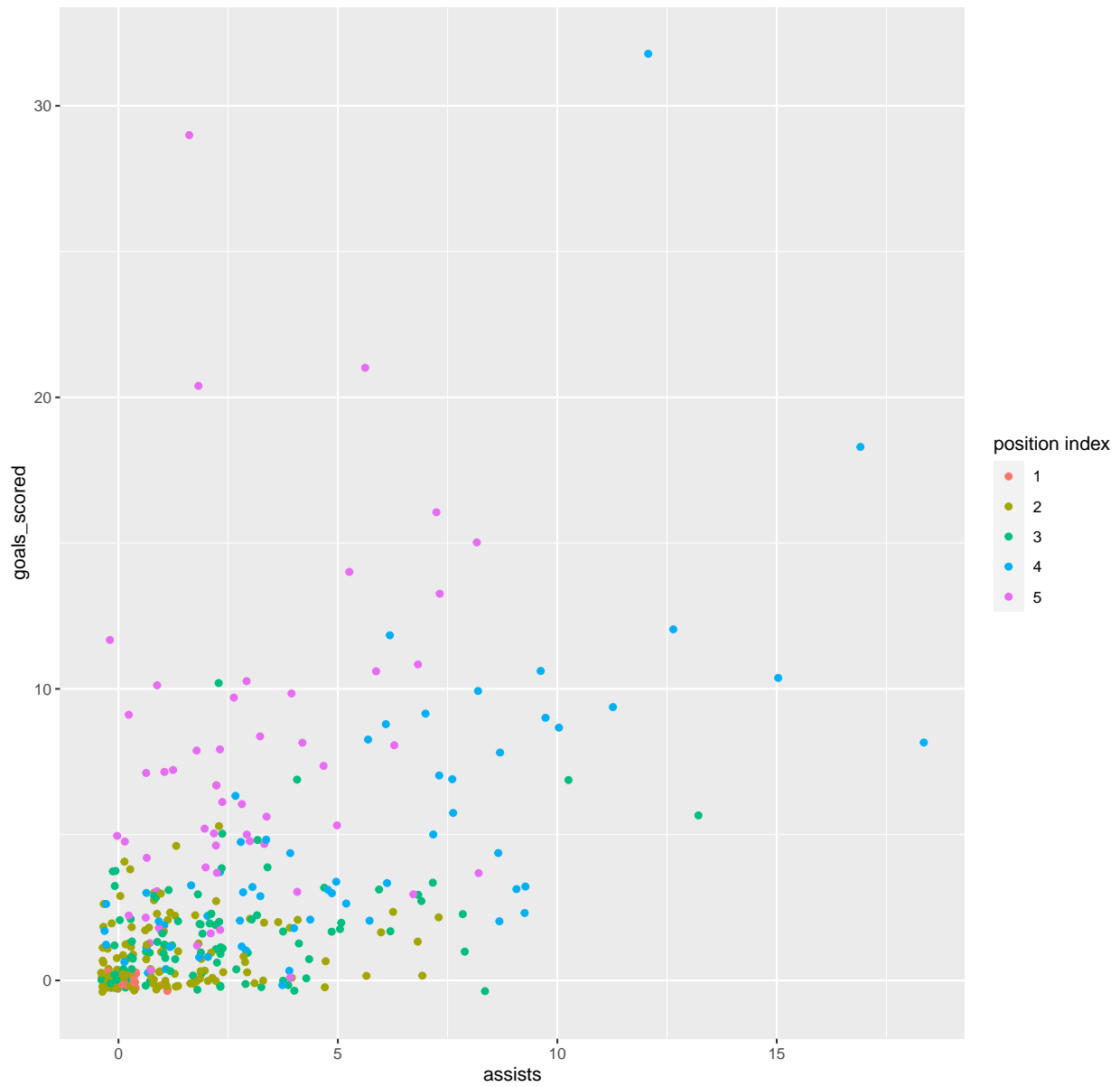
```
goals_vs_assist(season17)
```

Goals scored vs Assists during the 2016/2017 season

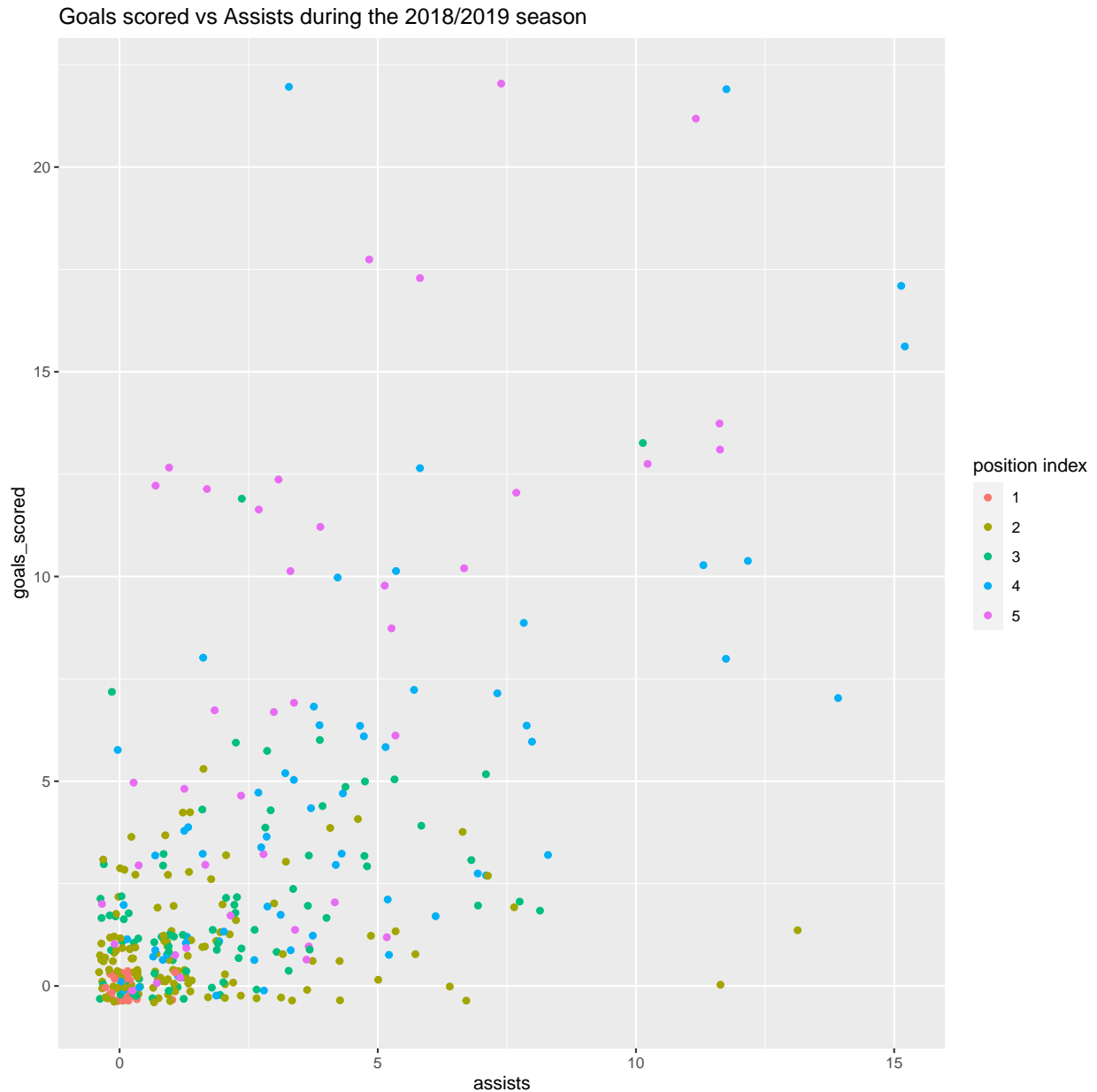


```
goals_vs_assist(season18)
```

Goals scored vs Assists during the 2017/2018 season



```
goals_vs_assist(season19)
```



Ranking the players with the top 10 fpl points per season

```
fpl_point_rank <- function(df) {
  df %>%
    mutate(ranking = dense_rank(desc(total_points))) %>%
    filter(ranking <= 10) %>%
    arrange(desc(total_points)) %>%
    select(ranking, player_name, total_points) %>%
    print() %>%
    ggplot(aes(reorder(player_name, -total_points), total_points, fill = player_name)) +
    geom_bar(stat = "identity", color = "purple") +
    theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
}
```

```

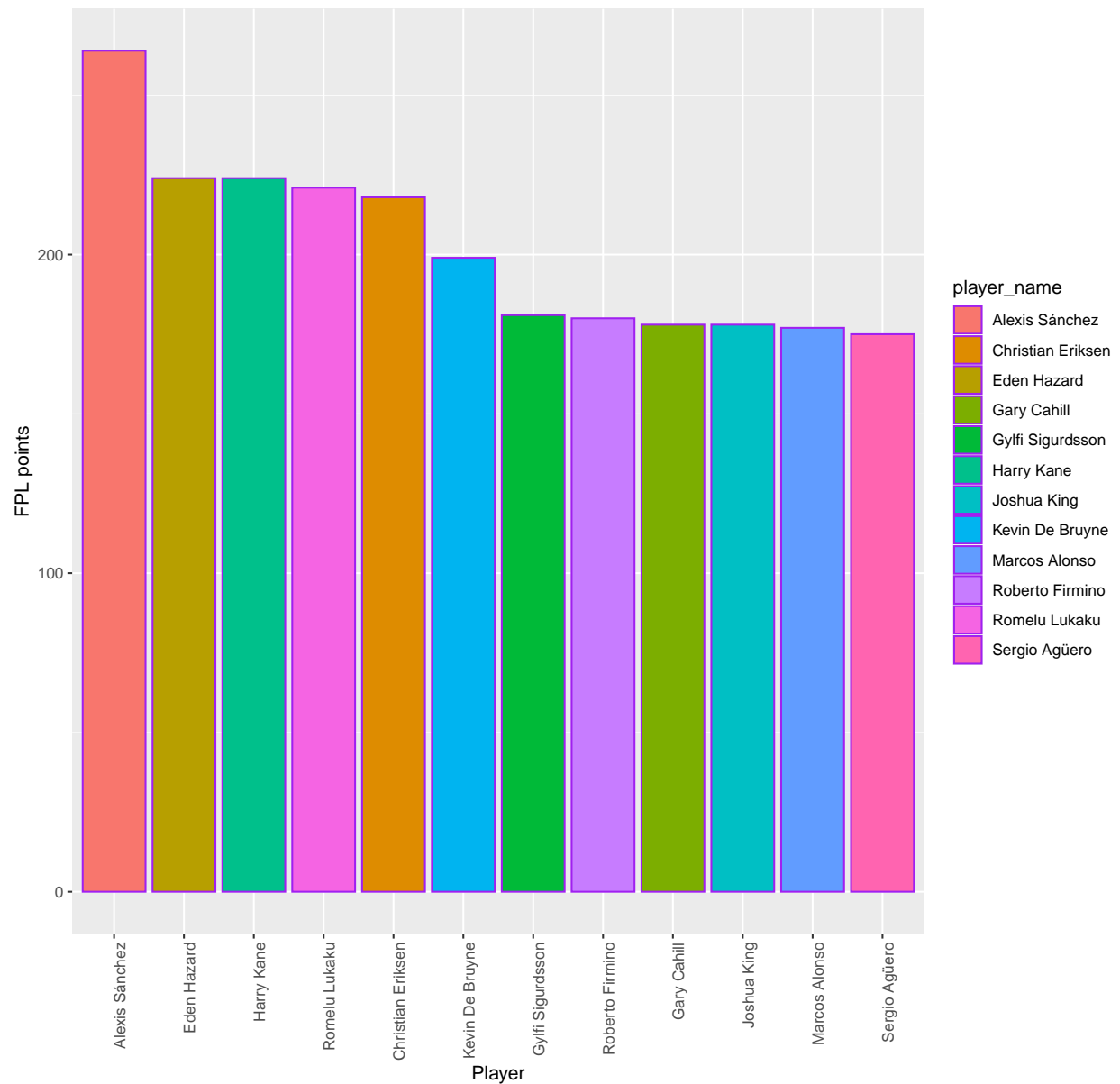
labs(title = sprintf("The top 10 players with the highest FPL points in %s",
unique(df$season)), x = "Player", y = "FPL points")
}

```

```
fpl_point_rank(season17)
```

##	ranking	player_name	total_points
## 1	1	Alexis Sánchez	264
## 2	2	Eden Hazard	224
## 3	2	Harry Kane	224
## 4	3	Romelu Lukaku	221
## 5	4	Christian Eriksen	218
## 6	5	Kevin De Bruyne	199
## 7	6	Gylfi Sigurdsson	181
## 8	7	Roberto Firmino	180
## 9	8	Gary Cahill	178
## 10	8	Joshua King	178
## 11	9	Marcos Alonso	177
## 12	10	Sergio Agüero	175

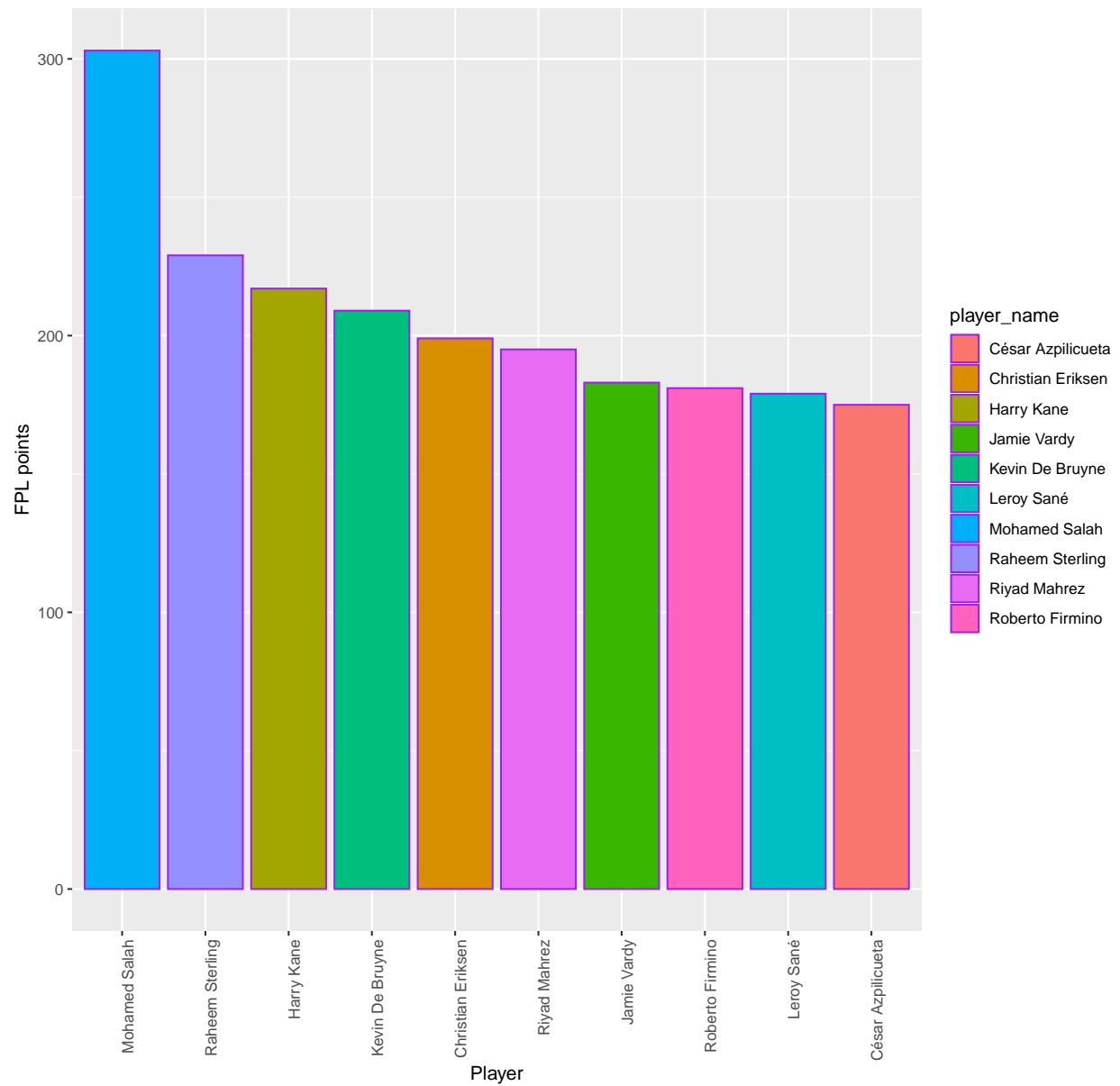
The top 10 players with the highest FPL points in 2016/2017



```
fpl_point_rank(season18)
```

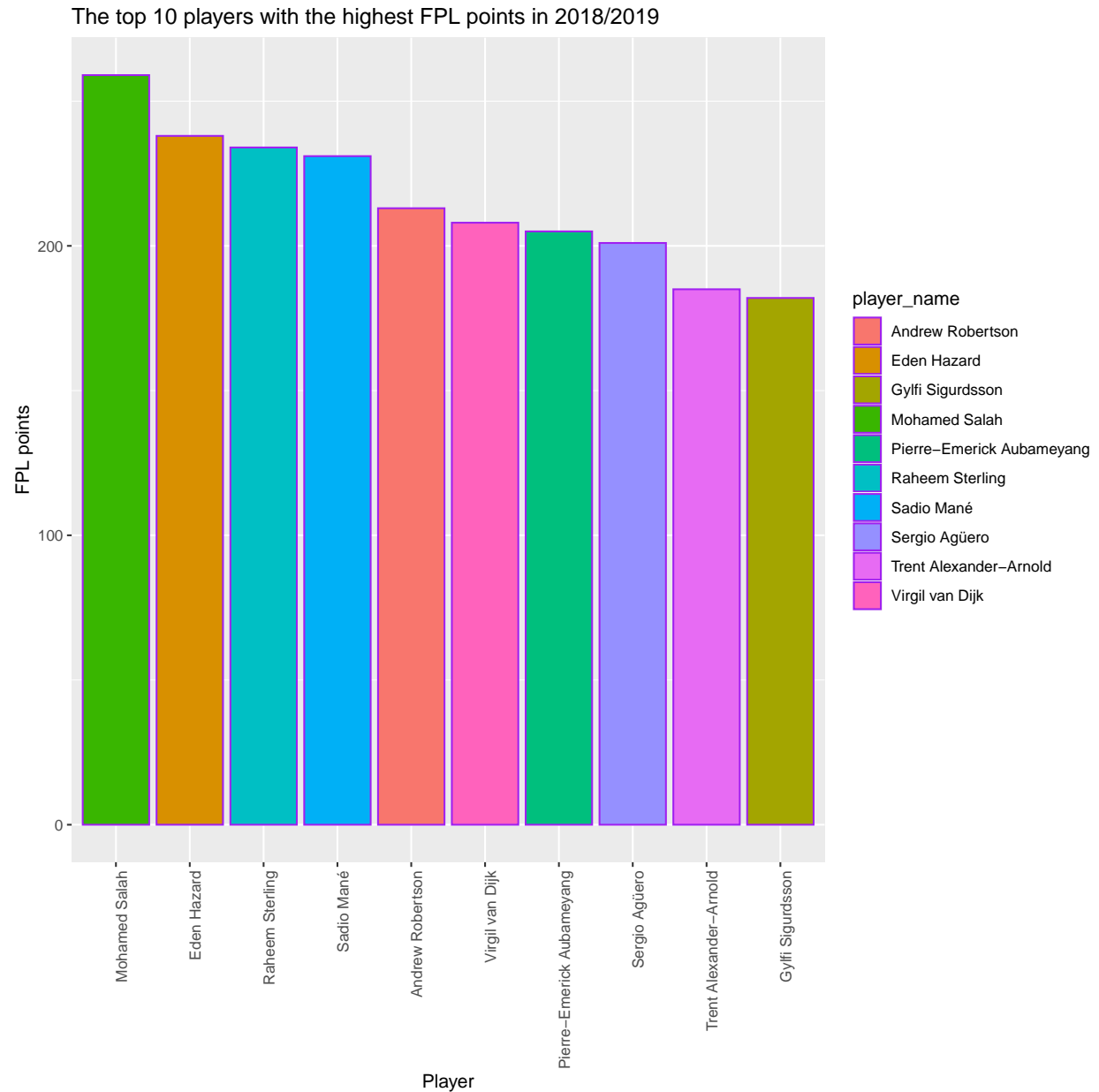
##	ranking	player_name	total_points
## 1	1	Mohamed Salah	303
## 2	2	Raheem Sterling	229
## 3	3	Harry Kane	217
## 4	4	Kevin De Bruyne	209
## 5	5	Christian Eriksen	199
## 6	6	Riyad Mahrez	195
## 7	7	Jamie Vardy	183
## 8	8	Roberto Firmino	181
## 9	9	Leroy Sané	179
## 10	10	César Azpilicueta	175

The top 10 players with the highest FPL points in 2017/2018



```
fpl_point_rank(season19)
```

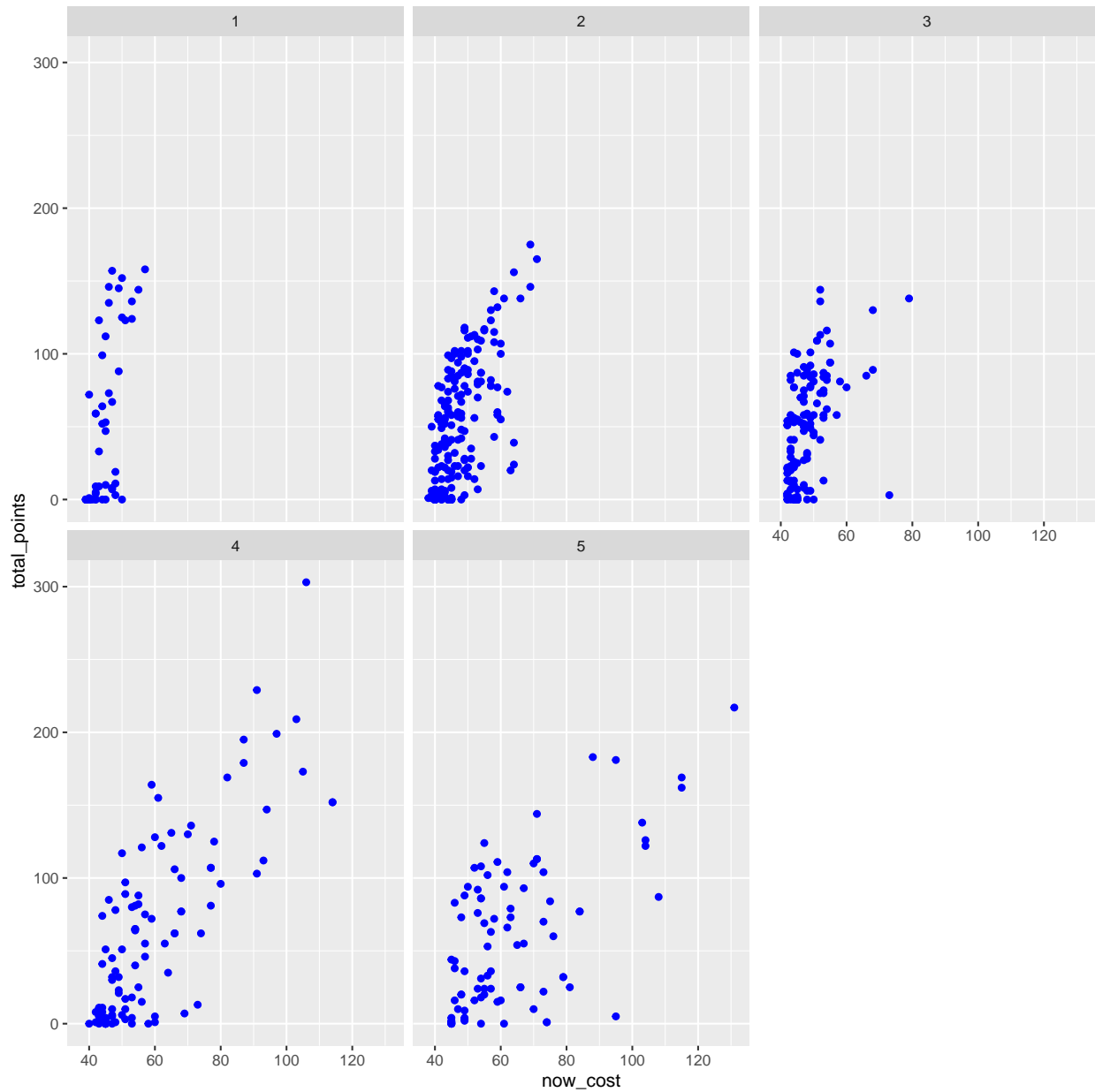
```
## ranking player_name total_points
## 1 1 Mohamed Salah 259
## 2 2 Eden Hazard 238
## 3 3 Raheem Sterling 234
## 4 4 Sadio Mané 231
## 5 5 Andrew Robertson 213
## 6 6 Virgil van Dijk 208
## 7 7 Pierre-Emerick Aubameyang 205
## 8 8 Sergio Agüero 201
## 9 9 Trent Alexander-Arnold 185
## 10 10 Gylfi Sigurdsson 182
```



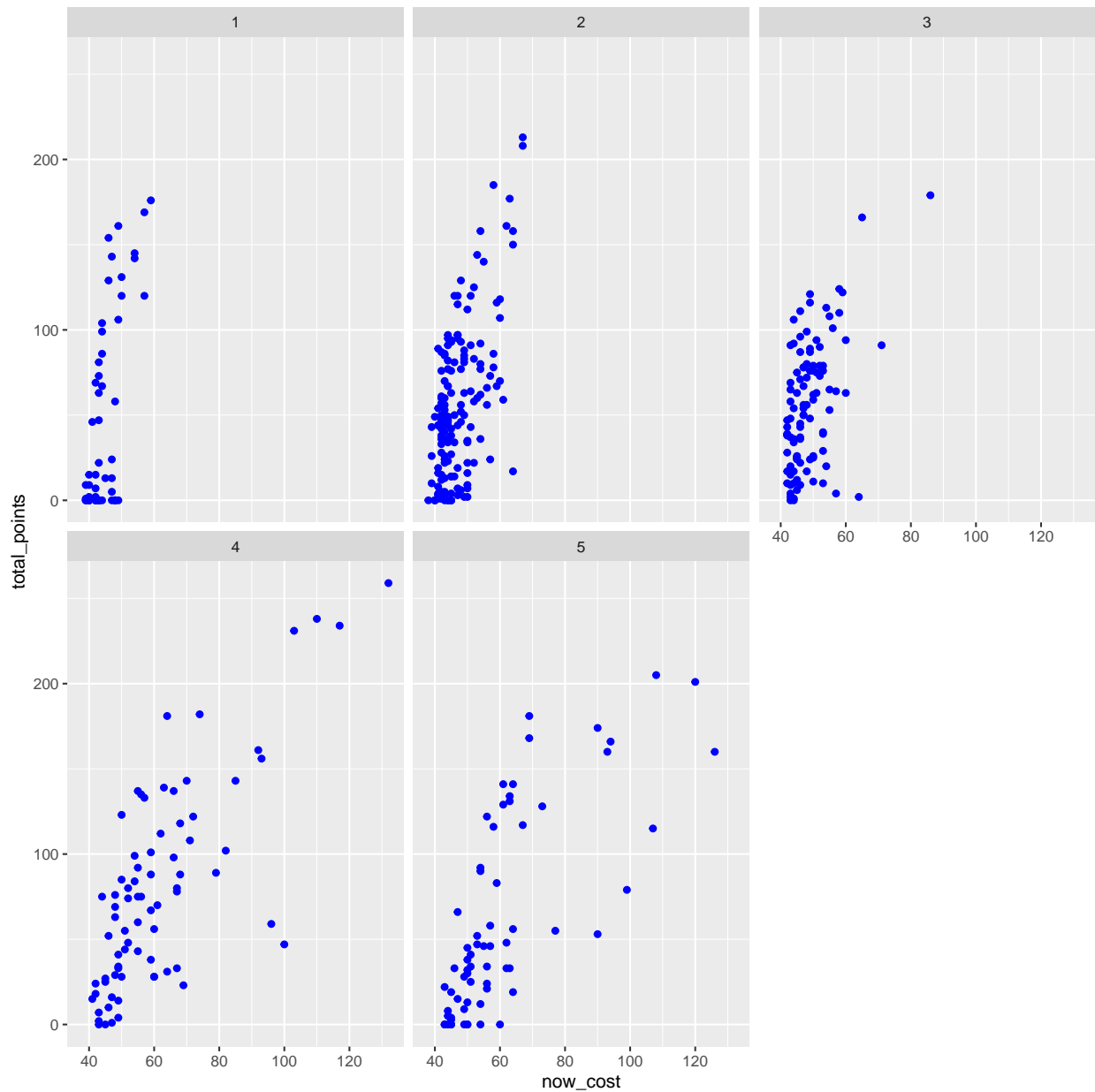
Plotting the total FPL points vs players' cost for every season

```
library(modelr)
options(na.action = na.warn)
value <- function(df) {
  ggplot(df, aes(x = now_cost)) +
    geom_point(aes(y = total_points), color = "blue") +
    facet_wrap(~position_index)
}
```

```
value(season18)
```



```
value(season19)
```

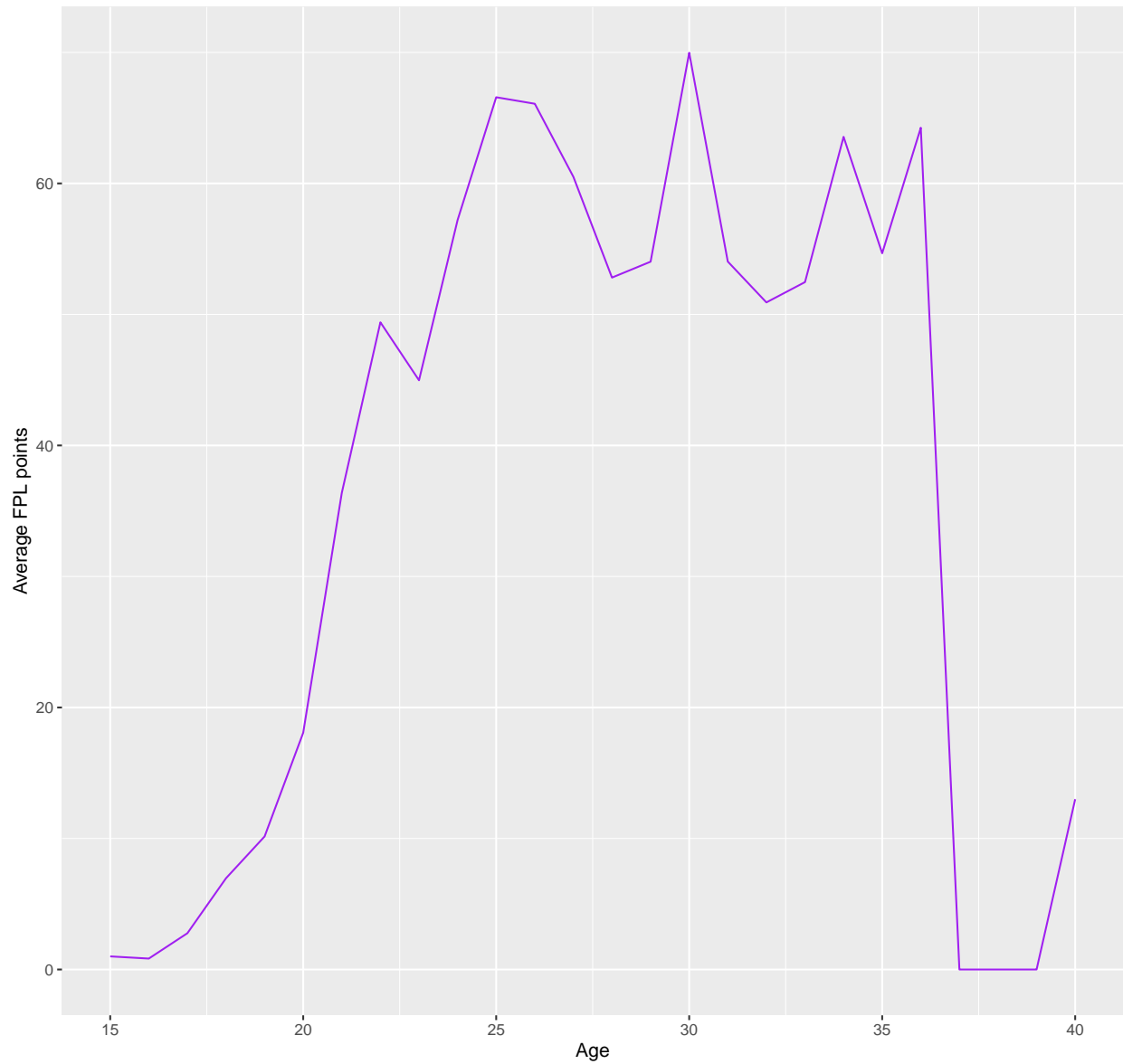


Visualizing how age has an impact on FPL points

```
age_freq <- function(df) {
  df %>%
    group_by(age) %>%
    summarize(mean = mean(total_points, na.rm = TRUE)) %>%
    ggplot(aes(x = age, y = mean)) + geom_freqpoly(stat = "identity", color = "purple") +
    labs(title = sprintf(title, unique(df$season)), x = "Age",
         y = "Average FPL points")
}
```

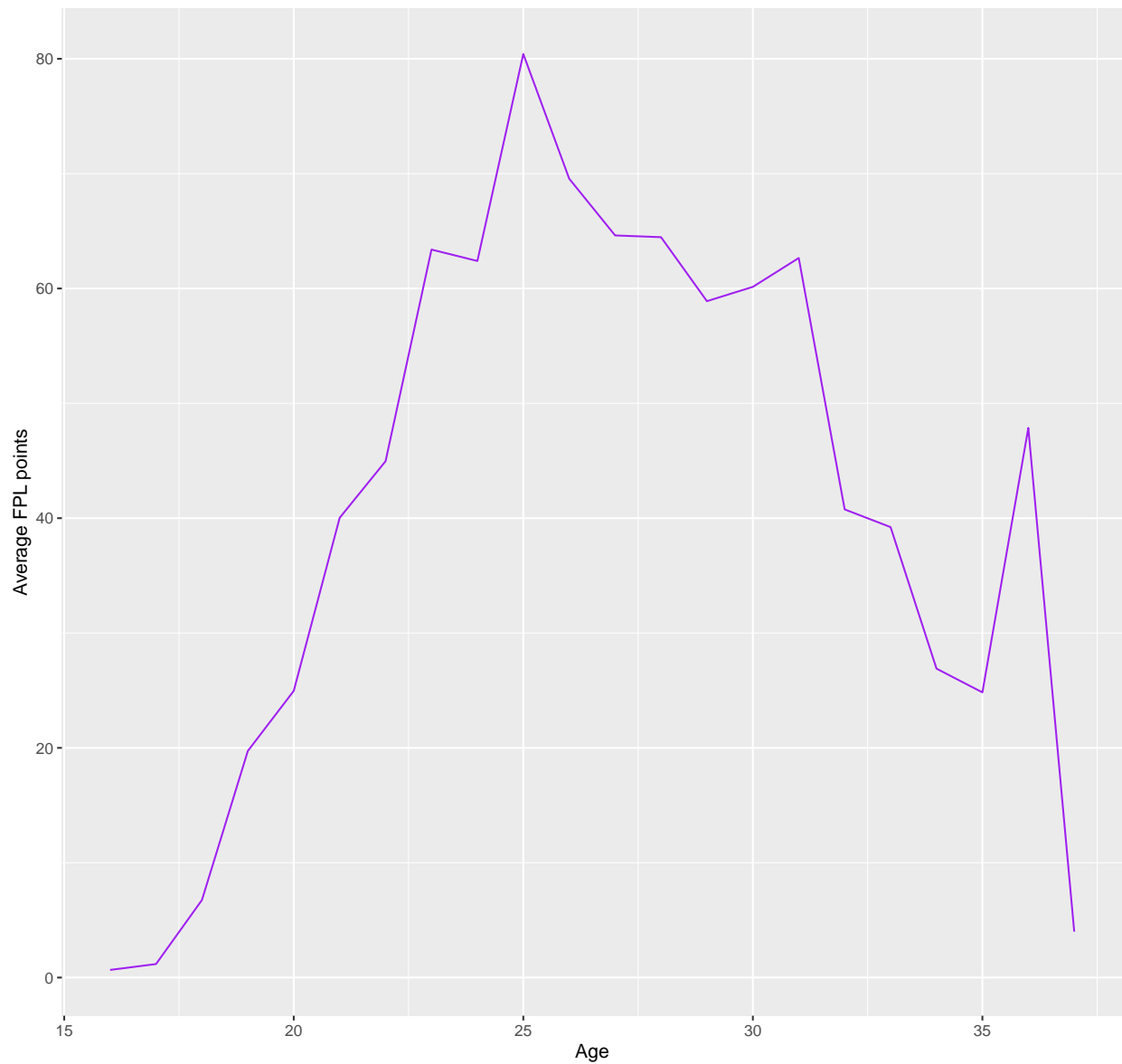
```
age_freq(season17)
```

A frequency polygon graph of the ages of Premier League players with respect to their mean FPL points during the season 2016/2017



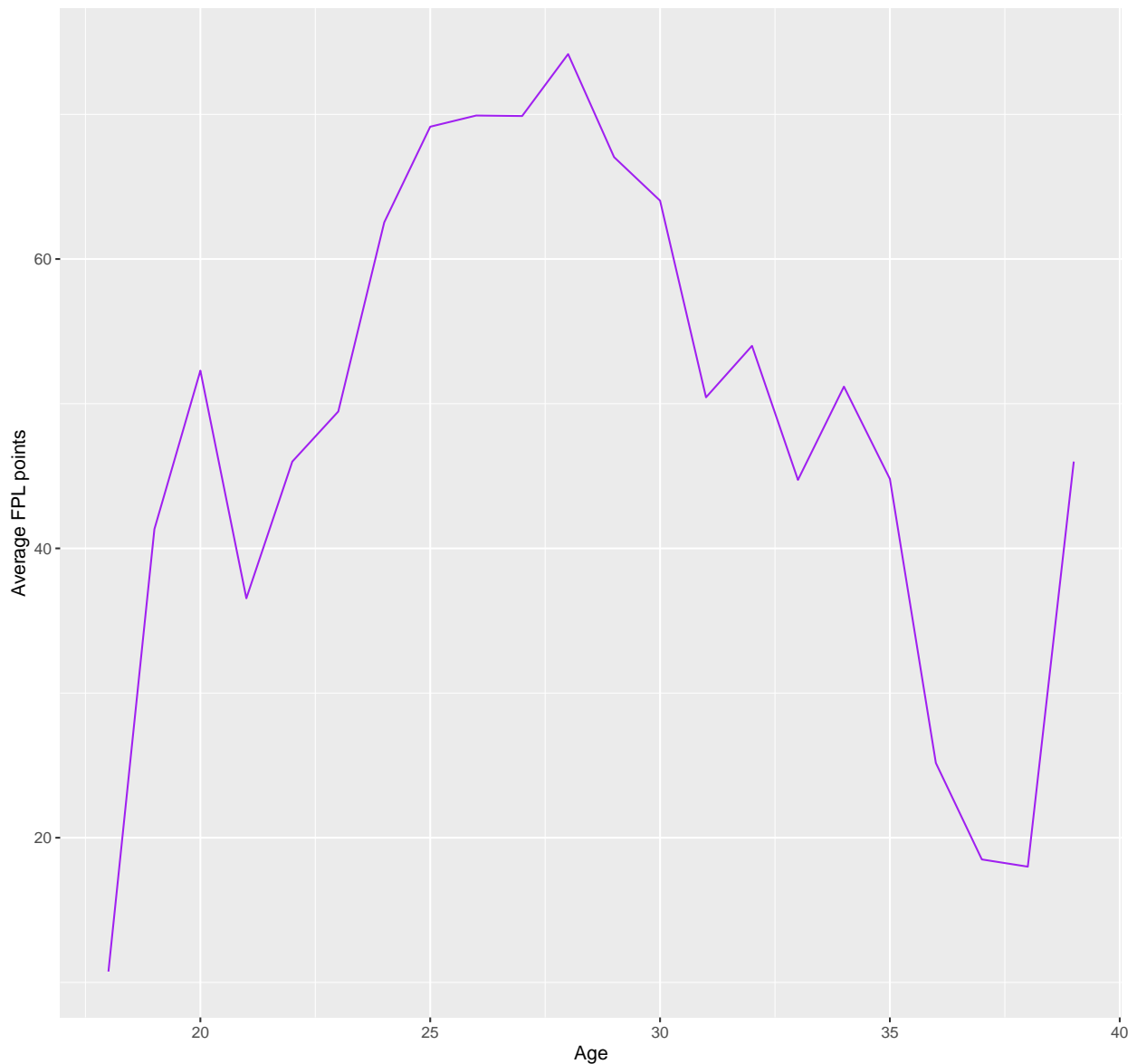
```
age_freq(season18)
```

A frequency polygon graph of the ages of Premier League players with respect to their mean FPL points during the season 2017/2018



```
age_freq(season19)
```

A frequency polygon graph of the ages of Premier League players with respect to their mean FPL points during the season 2018/2019



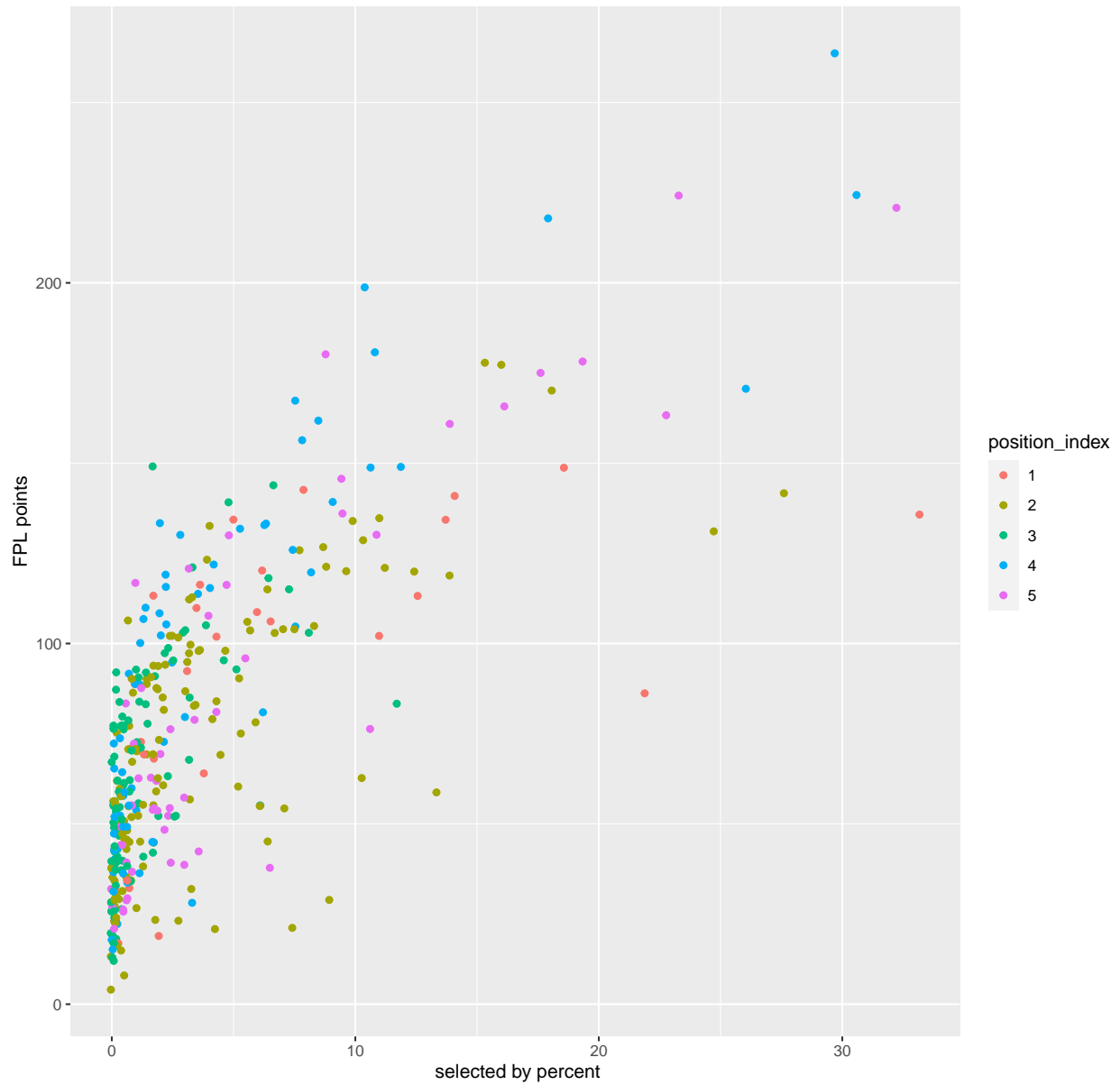
How selected by percent varies with the total FPL points per season

```
sel_fpl <- function(df) {
  df %>%
    filter(minutes.played >= 500) %>%
    ggplot(aes(selected_by_percent, total_points)) +
    geom_point(aes(color = position_index), position = "jitter") +
    labs(
      title = sprintf("A graph of FPL points vs selected by percent in the season %s",
        unique(df$season)), x = "selected by percent", y = "FPL points")
}
```



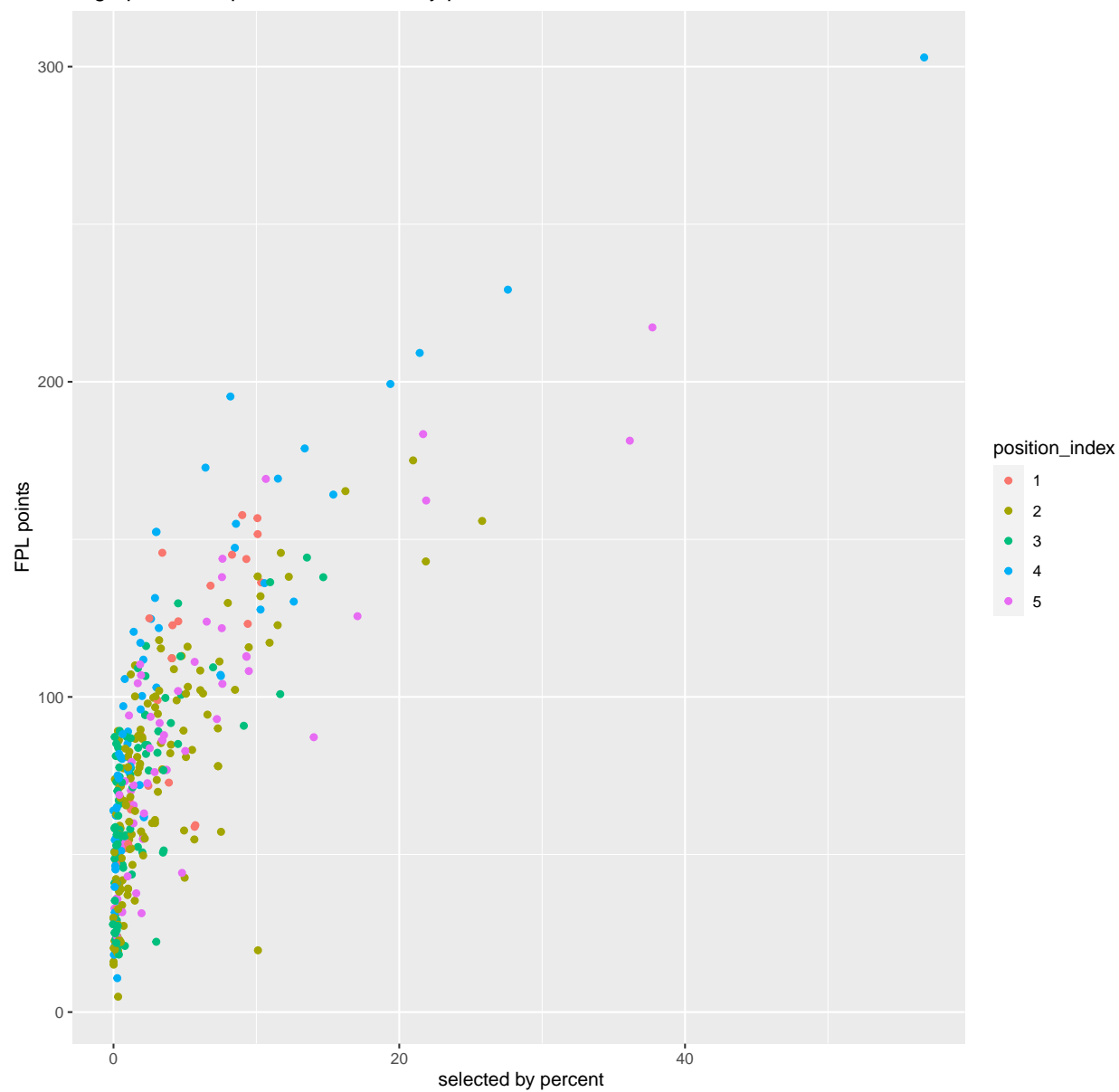
```
sel_fpl(season17)
```

A graph of FPL points vs selected by percent in the season 2016/2017



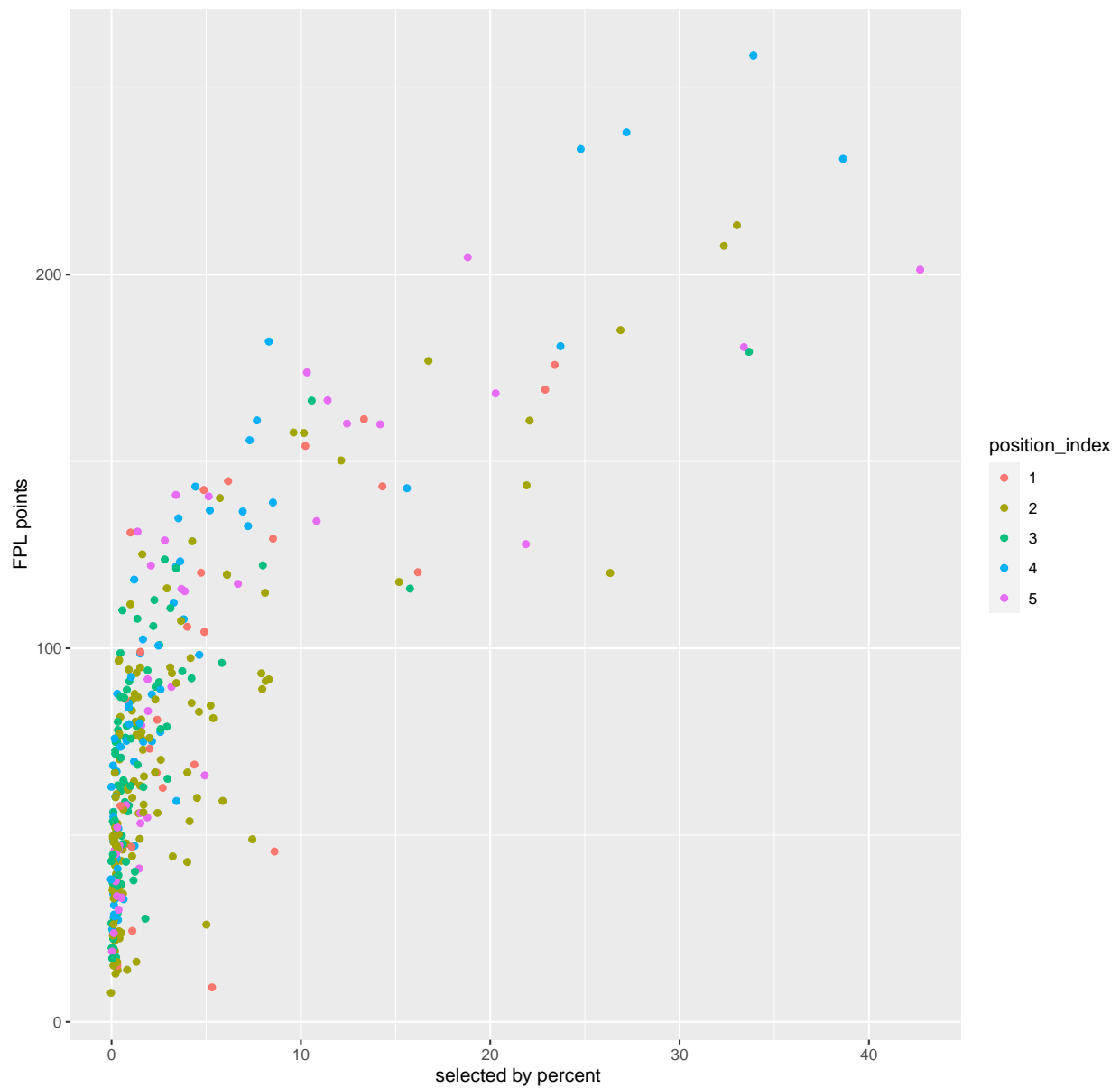
```
sel_fpl(season18)
```

A graph of FPL points vs selected by percent in the season 2017/2018



```
sel_fpl(season19)
```

A graph of FPL points vs selected by percent in the season 2018/2019

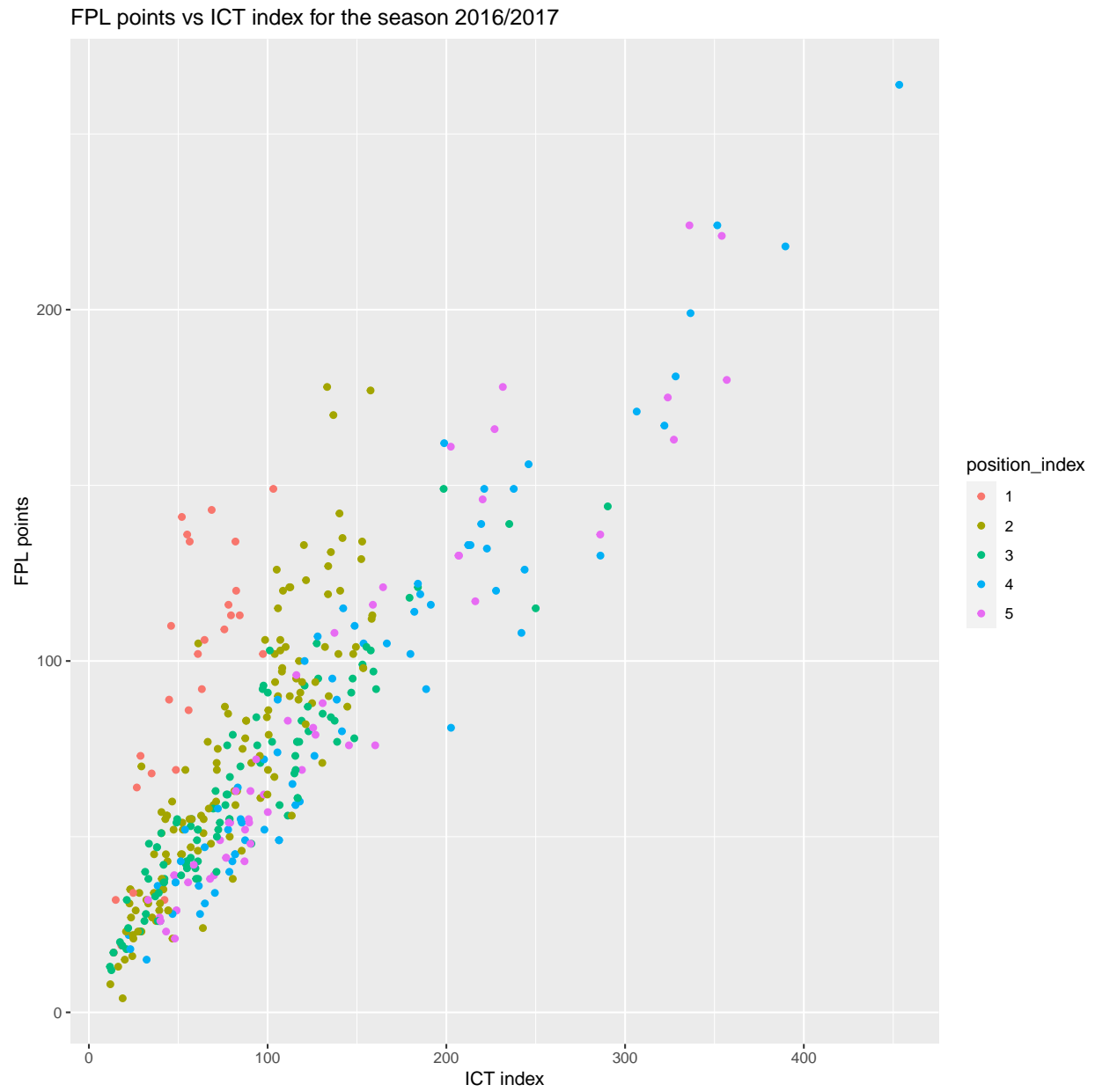


Creating a function that plots FPL points vs ICT index

```
fpl_vs_ict <- function(df) {
  df %>%
    filter(minutes.played >= 500) %>%
    ggplot() +
    geom_point(aes(ict_index, total_points, color = position_index)) +
    labs(title = sprintf("FPL points vs ICT index for the season %s",
                        unique(df$season)), x = "ICT index", y = "FPL points")
}
```

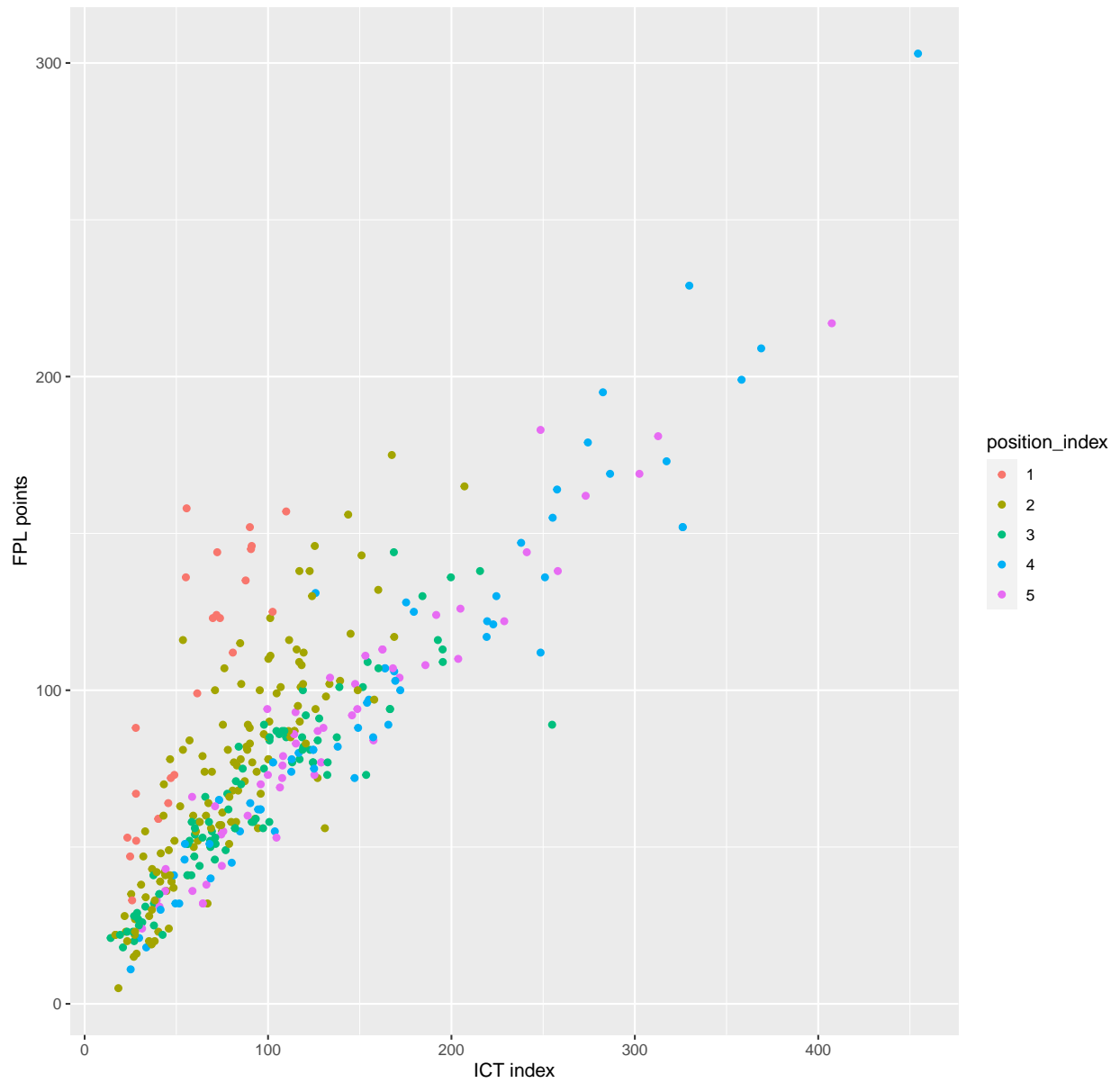
Plotting a graph of FPL points vs ICT index

```
fpl_vs_ict(season17)
```

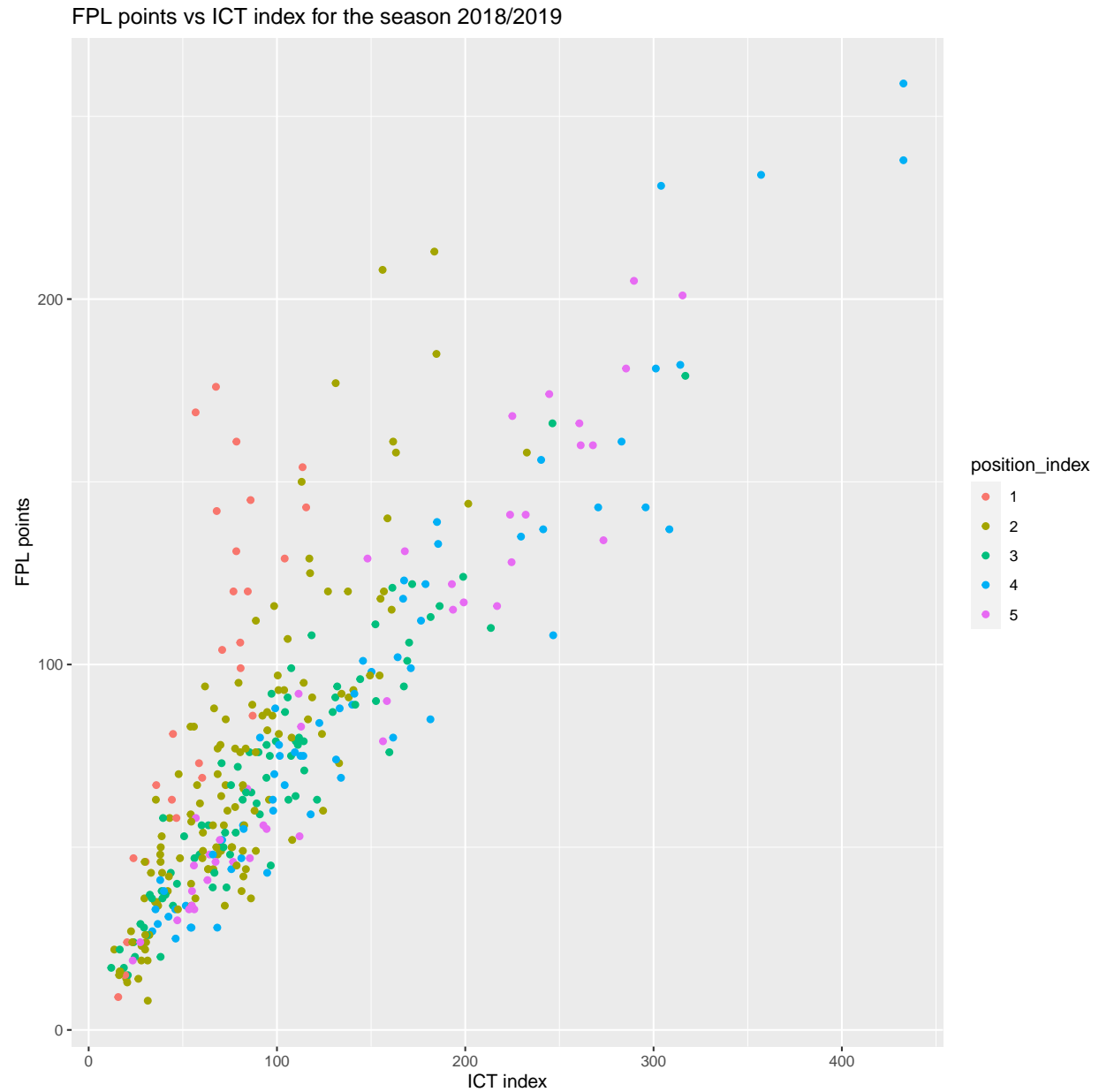


```
fpl_vs_ict(season18)
```

FPL points vs ICT index for the season 2017/2018



```
fpl_vs_ict(season19)
```



Creating a function for printing out the first ten rows alongside selected columns

```
quickview <- function(df) {  
  df1 <- df %>%  
    select(c(player_name,club_name,total_points,  
             minutes.played, fpl_to_game))  
  knitr::kable(df1[1:10,],  
caption = sprintf("The first 10 rows of the season %s dataset with specific columns",  
unique(df$season)))  
}
```

Applying the function to the datasets

```
quickview(season17)
```

Table 1: The first 10 rows of the season 2016/2017 dataset with specific columns

player_name	club_name	total_points	minutes.played	fpl_to_game
Joe Hart	Manchester City	0	0	NaN
Claudio Bravo	Manchester City	73	1968	3.338415
Willy Caballero	Manchester City	64	1452	3.966942
Angus Gunn	Manchester City	0	0	NaN
Nicolás Otamendi	Manchester City	100	2592	3.472222
Vincent Kompany	Manchester City	57	820	6.256098
John Stones	Manchester City	59	2014	2.636544
Eliaquim Mangala	Manchester City	0	0	NaN
Jason Denayer	Manchester City	56	1878	2.683706
Aleksandar Kolarov	Manchester City	95	2535	3.372781

```
quickview(season18)
```

Table 2: The first 10 rows of the season 2017/2018 dataset with specific columns

player_name	club_name	total_points	minutes.played	fpl_to_game
Thibaut Courtois	Chelsea FC	136	3150	3.885714
Willy Caballero	Chelsea FC	11	270	3.666667
Antonio Rüdiger	Chelsea FC	115	2336	4.430651
Andreas Christensen	Chelsea FC	79	2067	3.439768
Gary Cahill	Chelsea FC	74	2082	3.198847
Kurt Zouma	Chelsea FC	87	2887	2.712158
Fikayo Tomori	Chelsea FC	0	0	NaN
Trevoh Chalobah	Chelsea FC	0	0	NaN
Ethan Ampadu	Chelsea FC	1	10	9.000000
Marcos Alonso	Chelsea FC	165	2855	5.201401

```
quickview(season19)
```

Table 3: The first 10 rows of the season 2018/2019 dataset with specific columns

player_name	club_name	total_points	minutes.played	fpl_to_game
Ederson Santana de Moraes	Manchester City	169	3420	4.447368
Claudio Bravo	Manchester City	0	0	NaN
John Stones	Manchester City	83	1761	4.241908
Aymeric Laporte	Manchester City	177	3056	5.212696
Nicolás Otamendi	Manchester City	70	1233	5.109489
Vincent Kompany	Manchester City	58	1220	4.278689

player_name	club_name	total_points	minutes.played	fpl_to_game
Philippe Sandler	Manchester City	0	0	NaN
Benjamin Mendy	Manchester City	59	900	5.900000
Oleksandr Zinchenko	Manchester City	44	1151	3.440487
Kyle Walker	Manchester City	150	2776	4.863112

Writing the files to csv files

```
s17 <- file('season2017.csv', encoding = "UTF-8")
write.csv(season17, file = s17)
s18 <- file('season2018.csv', encoding = 'UTF-8')
write.csv(season18, file = s18)
s19 <- file('season2019.csv', encoding = "UTF-8")
write.csv(season19, file = s19)
```