EDA on Messi and Cristiano Ronaldo Goals Dataset

Introduction

As we matured watching Messi and Ronaldo dominate football, they bestowed us with countless moments of joy. I Kaggled around one hour and found this dataset.

About Dataset

Source : https://www.kaggle.com/datasets/azminetoushikwasi/lionel-messi-vs-cristiano-ronaldo-club-goals?select=data.csv

The data is stored in table form contains 14 columns such as :

- Player: Who? Messi? Ronaldo?
- Season: The season which each goal is scored.
- Competition: The competition which each goal goal is scored.
- Matchday: The match day which each goal goal is scored.
- Date: The date of goals scored.
- Club: Clubs the played for.
- Venue: Where the goal is scored? Home? Away?
- Opponent: Clubs they scored against.
- Result: Result of the match.
- Playing_position: Which position they played when they score that goal.
- Minute: Which minute when the goal is scored.
- At_score: Result of the match after goal is scored.
- Type: How the goal is scored, Penalty? Freekick?
- Goal_assist: Assisted by who?

Tools

- R: Clean,manipulate ,Visualization
- Rmarkdown: Write report (pdf)

Import Libraries

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.1.3
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.1.3
```

```
library(wordcloud)
```

```
## Warning: package 'wordcloud' was built under R version 4.1.3

## Loading required package: RColorBrewer

## Warning: package 'RColorBrewer' was built under R version 4.1.3
```

Import dataset from Kaggle

I downloaded the dataset and store it in a folder called "football project". Now I am going to set the directory folder so that it is easy for me access the file and save the results of this project. Then, I read the file which was saved as .csv and name it as *goal*.

```
setwd("C:/Users/ASUS/Desktop/football project")
goal <- read.csv("messi_vs_cr7.csv", encoding = "UTF-8")
```

Basic Information

Name of each column

```
names(goal)
```

```
##  [1] "Player"           "Season"           "Competition"      "Matchday"
##  [5] "Date"             "Venue"            "Club"             "Opponent"
##  [9] "Result"           "Playing_Position" "Minute"           "At_score"
## [13] "Type"             "Goal_assist"
```

How the data looks like

```
head(goal,5)
```

```
##              Player Season              Competition     Matchday       Date
## 1 Cristiano Ronaldo  02/03            Liga Portugal            6 2002-10-07
## 2 Cristiano Ronaldo  02/03            Liga Portugal            6 2002-10-07
## 3 Cristiano Ronaldo  02/03            Liga Portugal            8 2002-10-26
## 4 Cristiano Ronaldo  02/03 Taca de Portugal Placard Fourth Round 2002-11-24
## 5 Cristiano Ronaldo  02/03 Taca de Portugal Placard  Fifth Round 2002-12-18
##   Venue        Club                Opponent Result Playing_Position Minute
## 1     H Sporting CP           Moreirense FC   3:00               LW     34
## 2     H Sporting CP           Moreirense FC   3:00               LW   90+5
## 3     A Sporting CP             Boavista FC   1:02                     88
## 4     H Sporting CP            CD Estarreja   4:01                     67
## 5     H Sporting CP FC Oliveira do Hospital   8:01                     13
##   At_score              Type    Goal_assist
## 1     2:00          Solo run
## 2     3:00            Header      Rui Jorge
## 3     1:02 Right-footed shot Carlos Martins
## 4     3:00  Left-footed shot   Cesar Prates
## 5     3:00
```

Summary

```
summary(goal)
```

```
##     Player             Season          Competition          Matchday
##  Length:1400        Length:1400        Length:1400        Length:1400
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##      Date               Venue               Club             Opponent
##  Length:1400        Length:1400        Length:1400        Length:1400
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##     Result          Playing_Position      Minute             At_score
##  Length:1400        Length:1400        Length:1400        Length:1400
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##      Type           Goal_assist
##  Length:1400        Length:1400
##  Class :character   Class :character
##  Mode  :character   Mode  :character
```

We can see that there are 1400 rows and all variables have *character* class, but the Date should be set as *date*. So, we need to change its class to date.

Check for Null and Duplication

Check for null in each column:

```
colSums(!is.na(goal) & goal == "")
```

```
##           Player           Season      Competition          Matchday
##                0                0                0                 0
##             Date            Venue             Club         Opponent
##                0                0                0                 0
##           Result Playing_Position           Minute         At_score
##                0               58                0                 0
##             Type      Goal_assist
##               16              455
```

It is understandable that missing values existed in *Type*, *Goal_assist*, and *Playing_Position*. There might be some goals that can not be classified to which type of goal, some goals come from freekick or penalty which do not need anyone to assist, and some match it is hard to define the position of player.

Next, check for duplicated values:

```
sum(duplicated(goal))
```

```
## [1] 0
```

And, there isn't any duplicated row.

Data manipulation

Change class of Date for *character* to *date*

```
goal$Date = as.Date(goal$Date, format ="%Y-%m-%d")
```

Check:

```
class(goal$Date)
```

```
## [1] "Date"
```

Handling the blanks and nulls

Take a look at *Playing_Position* and *Type*

```
# Create separate data frames for Messi and Ronaldo
messi_goals <- subset(goal, Player == "Lionel Messi")
ronaldo_goals <- subset(goal, Player == "Cristiano Ronaldo")
# Apply table to each data frame
table(messi_goals$Type,trimws(messi_goals$Playing_Position))
```

```
##
##                           AM  CF  LW  RW  SS
##                            0   0   0   1   0
##   Chest                    0   0   0   1   0
##   Counter attack goal      0   0   0   1   0
##   Deflected shot on goal   0   0   0   1   1
##   Direct free kick         1  15   0  28   7
##   Header                   0  11   0  12   1
##   Left-footed shot         8 200   1 188  36
##   Long distance kick       0   0   0   0   1
##   Penalty                  3  39   0  38   4
##   Penalty rebound          0   1   0   2   0
##   Right-footed shot        4  41   0  37   3
##   Solo run                 0   2   0   2   0
##   Tap-in                   0   4   0   5   0
```

```
table(ronaldo_goals$Type,trimws(ronaldo_goals$Playing_Position))
```

```
##
##                             CF  LW  RW
##                           9   1   0   5
##   Counter attack goal     1   2   2   0
##   Deflected shot on goal  0   1   1   0
##   Direct free kick        6  10  23   9
##   Header                  7  36  55  14
##   Left-footed shot        7  28  63  13
##   Long distance kick      0   1   7   1
##   Penalty                 5  35  75  14
##   Penalty rebound         1   2   0   0
##   Right-footed shot      20  81 128  22
##   Solo run                1   0   1   0
##   Tap-in                  1   6   6   1
```

We can see that for Messi, most of the goals come from *Left-footed shot* and most of the match he played in the *RW* position. While most of Ronaldo's goals come from *Right-footed shot* and most of the time he played as *CF*.

So, I will fill blanks in *Type* of Ronaldo's goals to *Right-footed shot* and Messi's goals to *Left-footed shot*. The blanks in *Player_Position* of Ronaldo will be replaced by *CF*, while Mess's will be replaced by *RW*.

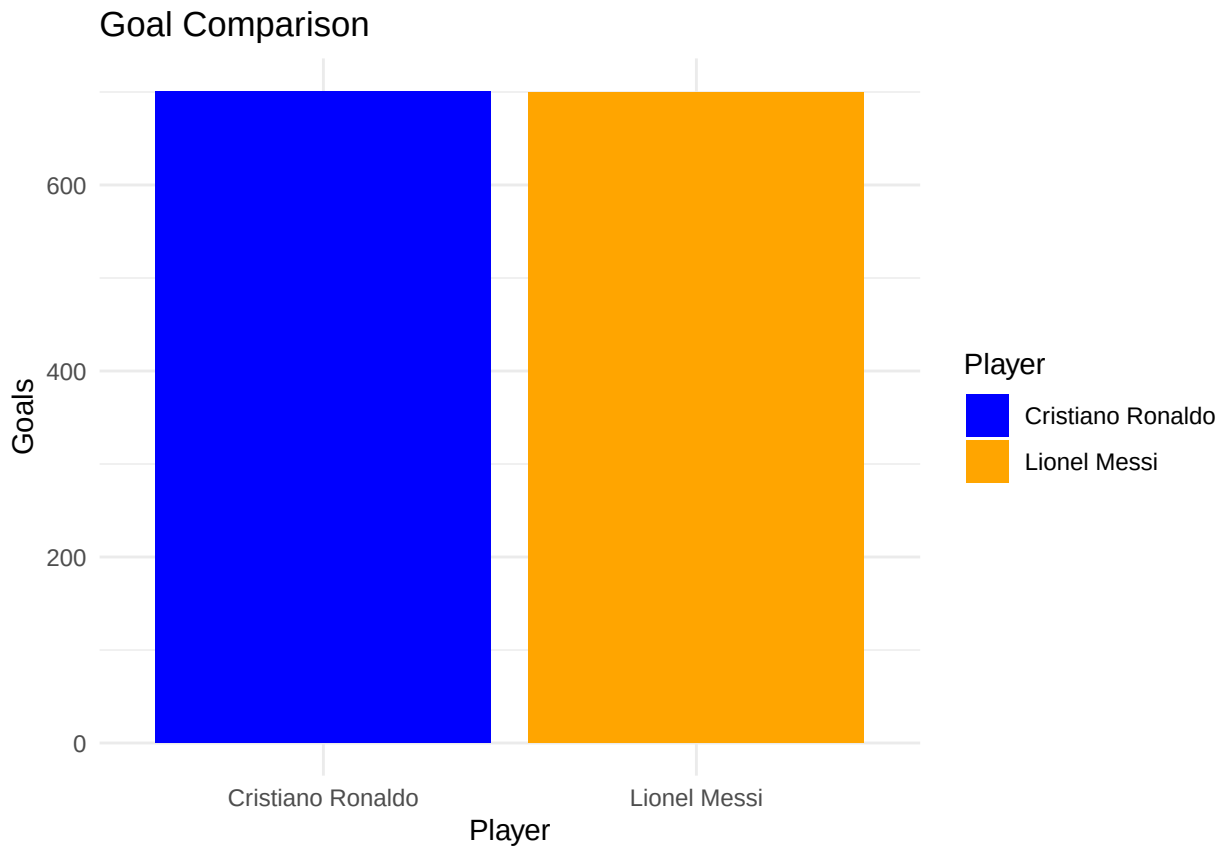
```
# Replace blanks in Type of Ronaldo's goals
goal$Type[goal$Player == "Cristiano Ronaldo" & goal$Type == ""] <- "Right-footed shot"

# Replace blanks in Type of Messi's goals
goal$Type[goal$Player == "Lionel Messi" & goal$Type == ""] <- "Left-footed shot"

# Replace blanks in Playing_Position of Ronaldo's goals
goal$Playing_Position[goal$Player == "Cristiano Ronaldo" & goal$Playing_Position == ""] <- "CF"

# Replace blanks in Playing_Position of Messi's goals
goal$Playing_Position[goal$Player == "Lionel Messi" & goal$Playing_Position == ""] <- "RW"
```

And for the case of *Goal_Assist* I will just leave it empty.

EDA - Exploratory Data Analysis

##Who scored more goals?

```
# count the number of goals for each player
player_counts <- table(goal$Player)

# create a dataframe from the counts
df <- data.frame(Player = names(player_counts), Goals = player_counts)

# create the bar plot
ggplot(df, aes(x = Player, y = Goals.Freq, fill = Player)) +
  geom_bar(stat = "identity") +
  labs(title = "Goal Comparison", x = "Player", y = "Goals") +
  scale_fill_manual(values = c("blue", "orange")) +
  theme_minimal()
```



Who Assisted them most?

```
# create separate dataframes for Ronaldo and Messi
ronaldo_df <- subset(goal, Player == "Cristiano Ronaldo")
messi_df <- subset(goal, Player == "Lionel Messi")
```

```r
# count the number of assists for each player
ronaldo_assists <- table(ronaldo_df$Goal_assist)
messi_assists <- table(messi_df$Goal_assist)

# generate random colors for each player
my_colors <- c("#006400", "#228B22", "#1E90FF", "#0000CD", "#800000", "#A52A2A")
ronaldo_colors <- sample(my_colors, length(ronaldo_assists), replace = TRUE)
messi_colors <- sample(my_colors, length(messi_assists), replace = TRUE)

# create the word clouds for each player
par(mar = c(5, 5, 5, 1))
wordcloud(names(ronaldo_assists), freq = ronaldo_assists, scale=c(8,0.7),
          color=ronaldo_colors, random.order = FALSE)
text(x = 22, y = 1.2, labels = "Cristiano Ronaldo's Goal Assists", col = "#228B22", font = 20, cex = 2)
```

It is apparent that Karim Benzema and Gareth Bale played a significant role in supporting Ronaldo during their prime time at Real Madrid. The collaboration between the three players was logical, given their shared tenure at the club.

```r
par(mar = c(5, 5, 5, 1))
wordcloud(names(messi_assists), freq = messi_assists, scale=c(8,0.5),
          color=messi_colors, random.order = FALSE)
text(x = 0.5, y = 1.2, labels = "Lionel Messi's Goal Assists", col = "#1E90FF", font = 2, cex = 2)
```

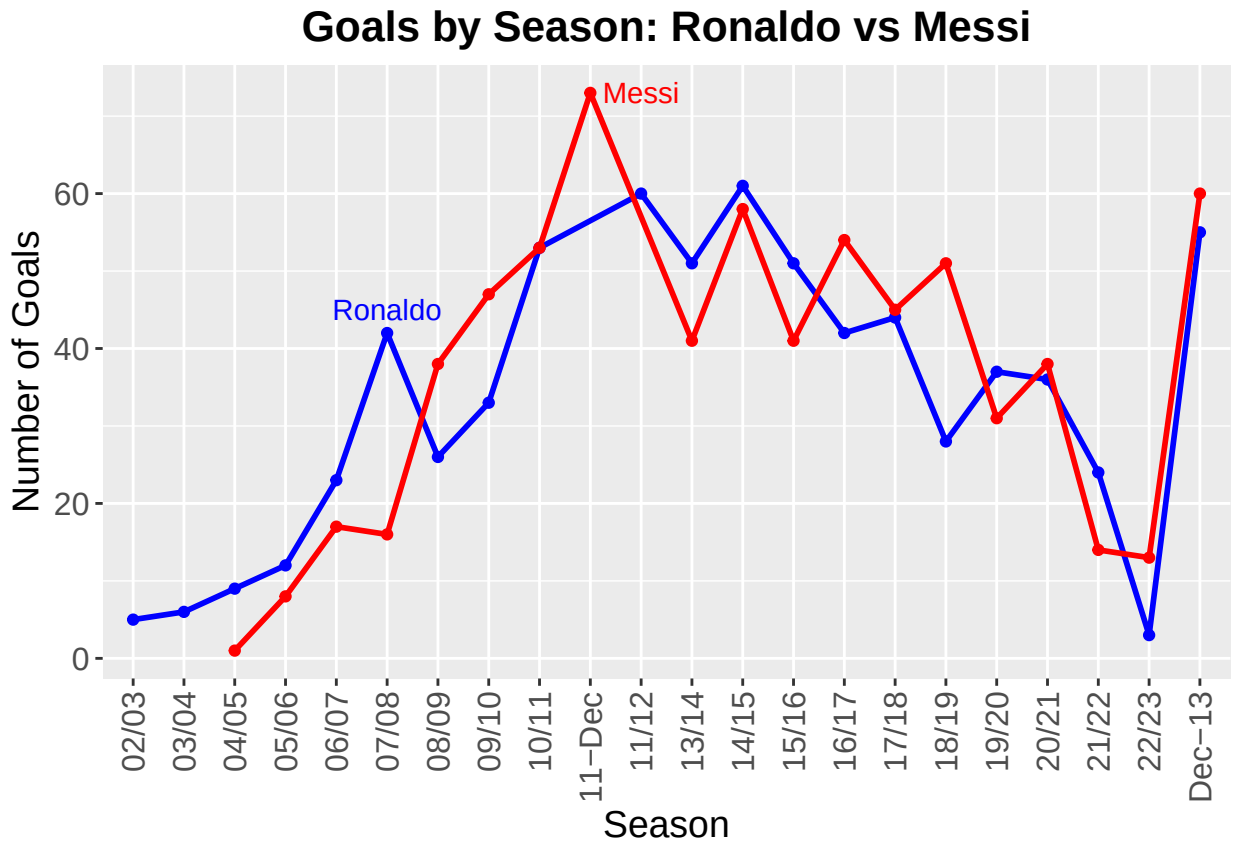
It is indisputable that Luis Suarez, Dani Alves, and Andres Iniesta have all played pivotal roles in the illustrious career of Lionel Messi at Barcelona. Suarez, who shares a close bond with Messi as his best friend, has been a prolific contributor to Messi's success, surpassing all other teammates in terms of assists. Following closely in second place is the dynamic duo of Dani Alves , with the former being widely recognized as the best right-back in Barcelona's history. And last but not least, Iniesta, who had the privilege of playing alongside Messi for more than a decade, has been instrumental in creating countless opportunities for the Argentine superstar to excel on the pitch.

Goals in each season

Line graphs below show goals scored by both in each season.

```
# count goals in each season
ronaldo_seasons <- as.data.frame(table(ronaldo_df$Season))
messi_seasons <- as.data.frame(table(messi_df$Season))

# rename columns
names(ronaldo_seasons) <- c("Season", "Goals")
names(messi_seasons) <- c("Season", "Goals")

# create line plot with points
ggplot() +
  # Ronaldo's line and points
  geom_line(data = ronaldo_seasons, aes(x = Season, y = Goals, group = 1), color = "blue", size = 1) +
  geom_point(data = ronaldo_seasons, aes(x = Season, y = Goals), color = "blue", size = 1.5) +
  # Messi's line and points
```

```
  geom_line(data = messi_seasons, aes(x = Season, y = Goals, group = 1), color = "red", size = 1) +
  geom_point(data = messi_seasons, aes(x = Season, y = Goals), color = "red", size = 1.5 ) +
  # plot labels and theme
  labs(x = "Season", y = "Number of Goals", title = "Goals by Season: Ronaldo vs Messi") +
  theme(plot.title = element_text(size = 16, face = "bold", hjust = 0.5),
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1, size = 12),
        axis.text.y = element_text(size = 12),
        axis.title = element_text(size = 14)) +
  scale_color_manual(values = c("blue", "red"), labels = c("Ronaldo", "Messi")) +
  geom_text(aes(x=11,y=73,label="Messi"),col="red")+
  geom_text(aes(x=6,y=45,label="Ronaldo"),col="blue")
```

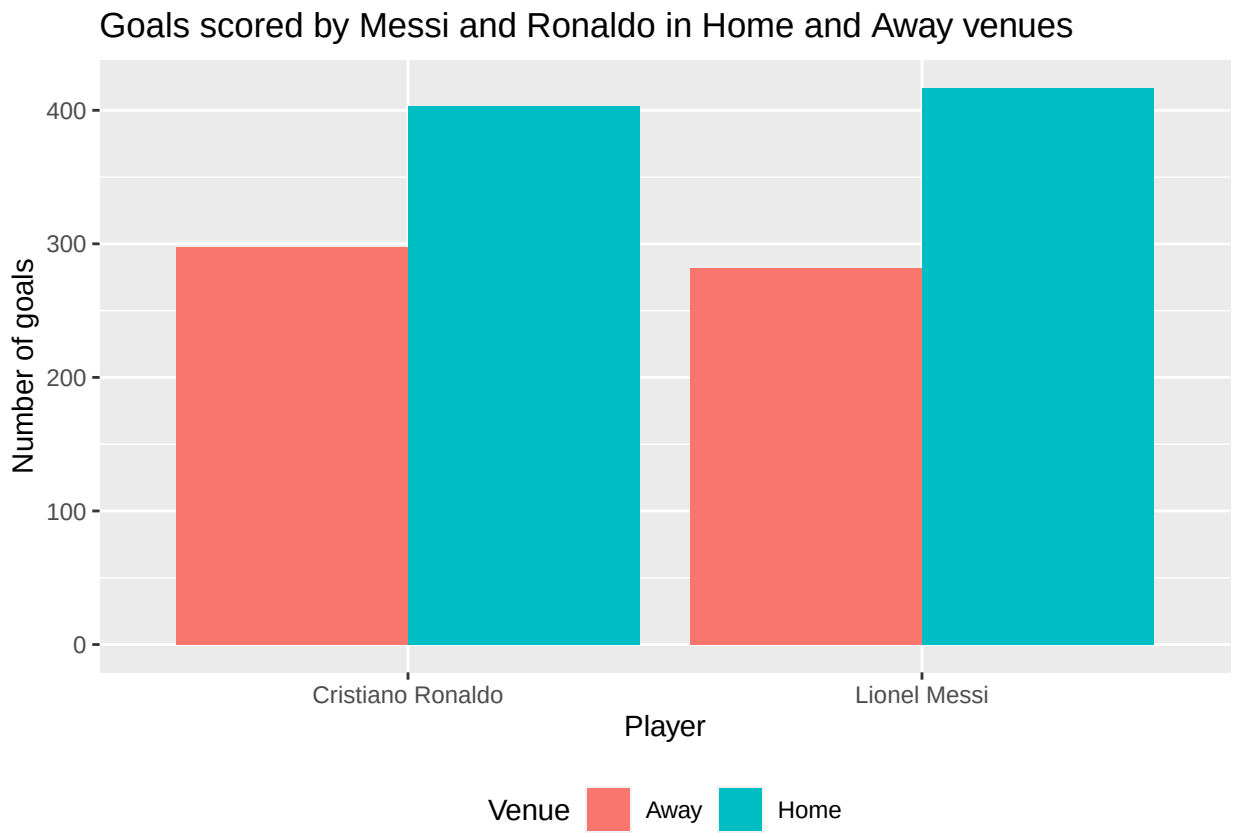


Home Vs Away Games

```
goals_by_player_and_venue <- as.data.frame(goal %>%
  group_by(Player, Venue) %>%
  summarise(Number_of_goals = n()))
```

```
## `summarise()` has grouped output by 'Player'. You can override using the
## `.groups` argument.
```

```
ggplot(data = goals_by_player_and_venue, aes(x = Player, y = Number_of_goals, fill = Venue)) +
  geom_bar(stat = "identity", position = "dodge", linewidth = 0.5) +
  labs(title = "Goals scored by Messi and Ronaldo in Home and Away venues", x = "Player", y = "Number o
  theme(legend.position = "bottom") +
  scale_fill_discrete(labels = c("Away", "Home"))
```

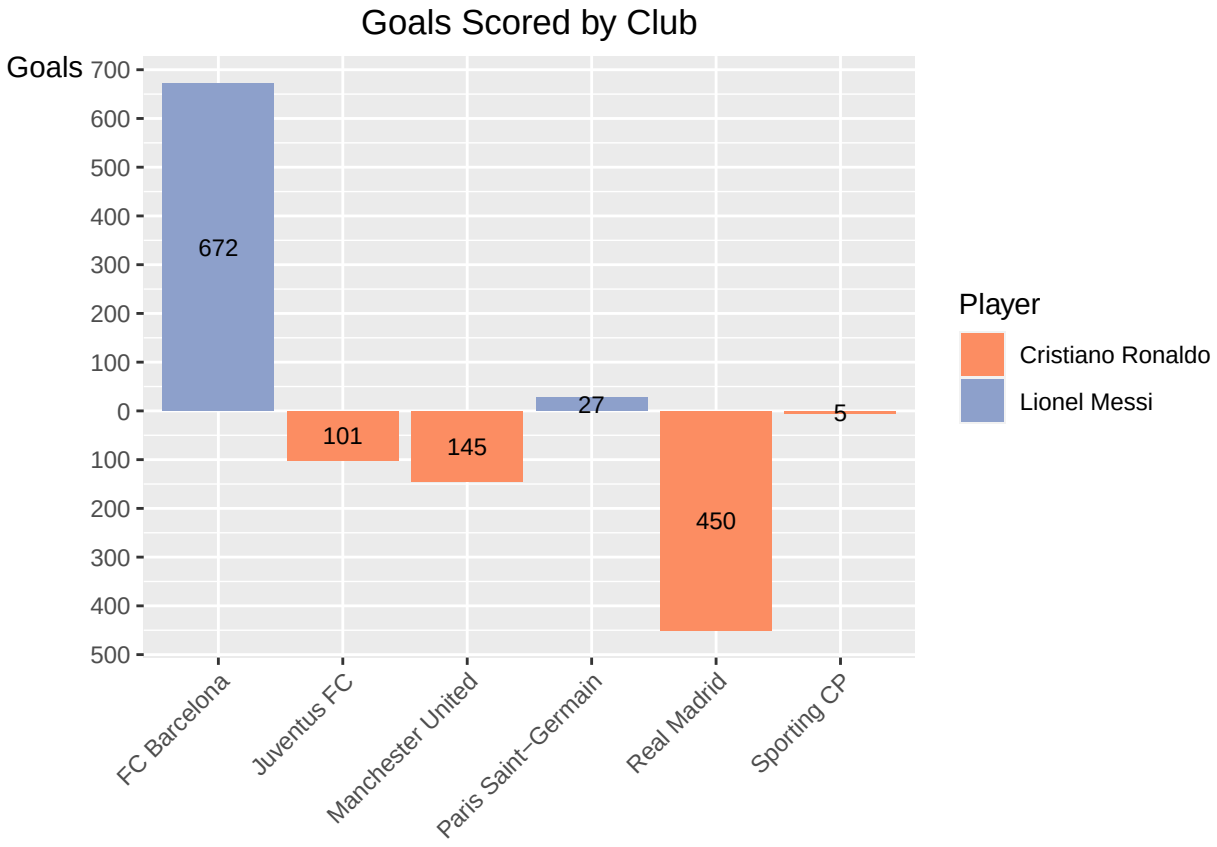


We can see that both of them scored in Home game more than Away game which is understandable.

Goals scored by clubs

```
messi_club_table <- table(subset(goal, Player == "Lionel Messi")$Club)
ronaldo_club_table <- table(subset(goal, Player == "Cristiano Ronaldo")$Club)

df <- data.frame(
  Club = c(names(messi_club_table), names(ronaldo_club_table)),
  Player = rep(c("Lionel Messi", "Cristiano Ronaldo"), times = c(length(messi_club_table), length(ronal
  Goals = c(as.numeric(messi_club_table), as.numeric(ronaldo_club_table))
)

# Create a custom color palette
my_colors <- c("#fc8d62", "#8da0cb", "#e78ac3", "#a6d854", "#ffd92f", "#e5c494")

# Modify y-axis values for Ronaldo bars
df$Goals[df$Player == "Cristiano Ronaldo"] <- -df$Goals[df$Player == "Cristiano Ronaldo"]

# Plot an up-down stacked bar chart with Club names on bars and rotated axis labels
ggplot(df, aes(x = Club, y = Goals, fill = Player)) +
  geom_bar(stat = "identity", position = "stack") +
  geom_text(aes(label = abs(Goals)), position = position_stack(vjust = 0.5), size = 3, color = "black")
  scale_fill_manual(values = my_colors) +
  labs(title = "Goals Scored by Club", x = NULL, y = "Goals") +
```

```
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1),
        axis.title.x = element_blank(),
        axis.title.y = element_text(angle = 0)) +
  scale_y_continuous(labels = function(x) abs(x), breaks = seq(-800, 800, by = 100))
```



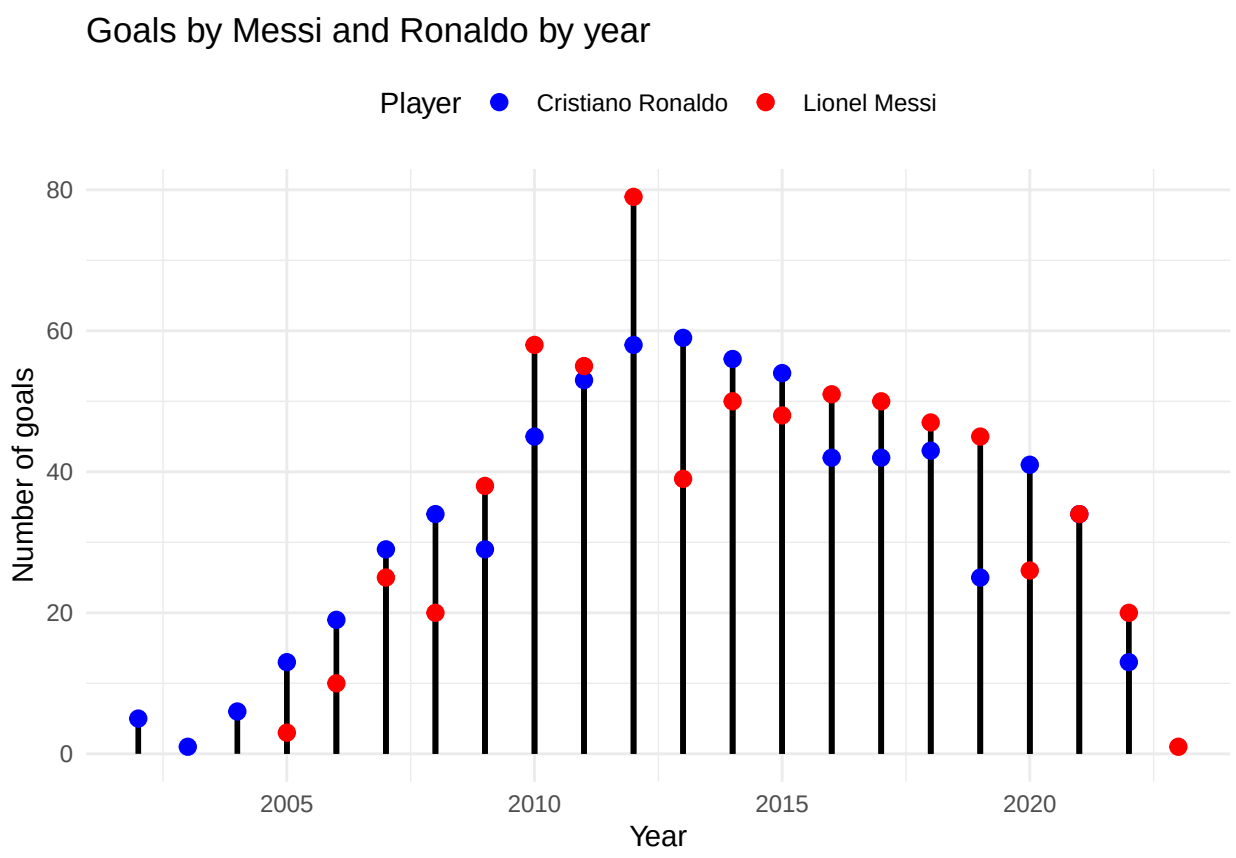
Goals by year

```
year_table <- goal %>%
  filter(Player %in% c("Lionel Messi", "Cristiano Ronaldo")) %>%
  mutate(year = lubridate::year(Date)) %>%
  group_by(Player, year) %>%
  summarise(goals = n())
```

```
## `summarise()` has grouped output by 'Player'. You can override using the
## `.groups` argument.
```

```
ggplot(year_table, aes(x = year, y = goals, group = Player)) +
  geom_segment(aes(xend = year, yend = 0), color = "black", linewidth = 1) +
  geom_point(aes(color = Player), size = 4, shape = 20, fill = "white") +
  scale_color_manual(values = c("Cristiano Ronaldo" = "blue", "Lionel Messi" = "red")) +
  labs(title = "Goals by Messi and Ronaldo by year",
```

```
      x = "Year",
      y = "Number of goals",
      color = "Player") +
  theme_minimal() +
  theme(legend.position = "top")
```



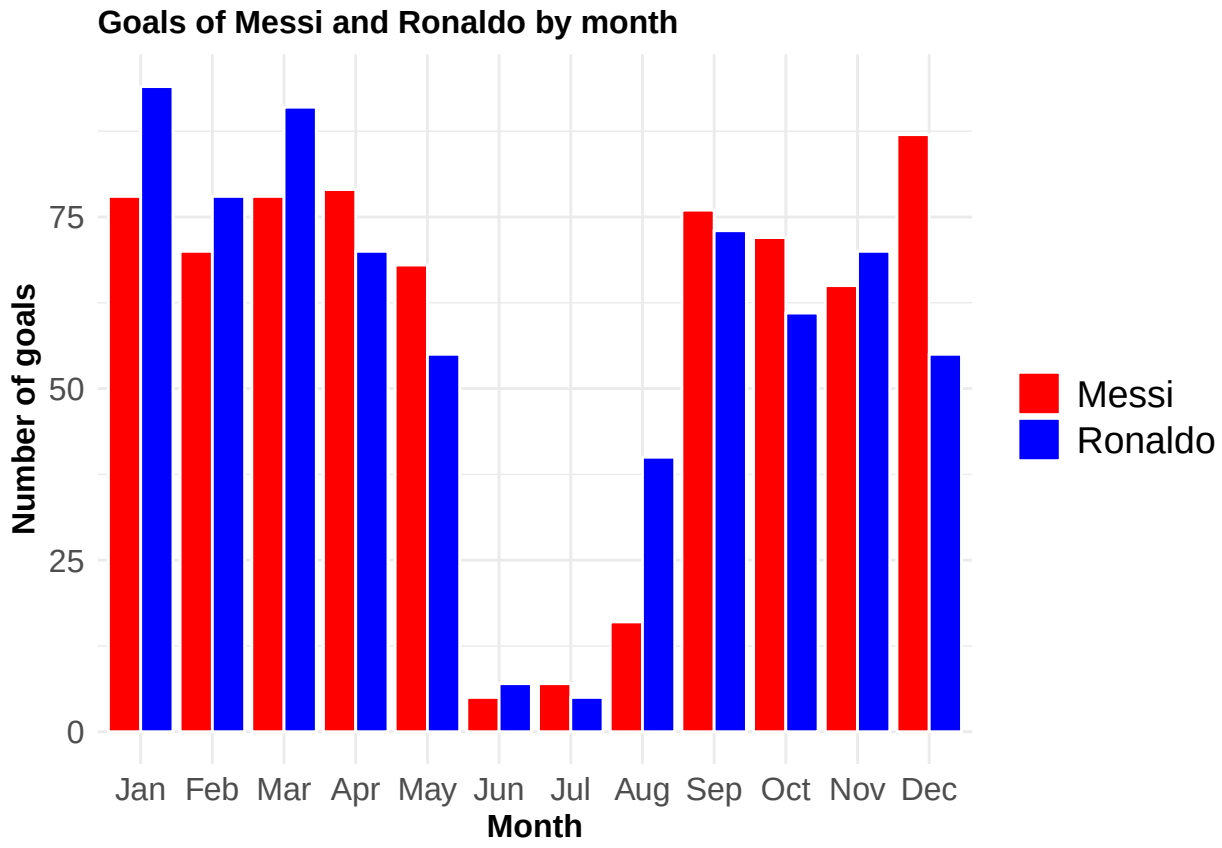
Goals by month

```
month_table <- goal %>%
  filter(Player %in% c("Lionel Messi", "Cristiano Ronaldo")) %>%
  mutate(month = lubridate::month(Date, label = TRUE)) %>%
  group_by(Player, month) %>%
  summarise(goals = n())
```

```
## `summarise()` has grouped output by 'Player'. You can override using the
## `.groups` argument.
```

```
# Set custom colors and fill for the plot
colors <- c("#e74c3c", "#2980b9")
fill <- c("#f5b7b1", "#a9cce3")

# Create the plot
```

```
ggplot(month_table, aes(x = month, y = goals, fill = Player)) +
  geom_col(position = "dodge", color = "white", size = 0.5) +
  scale_x_discrete(labels = month.abb) +
  labs(title = "Goals of Messi and Ronaldo by month",
       x = "Month",
       y = "Number of goals",
       fill = "") +
  scale_fill_manual(values = c("red", "blue"),
                    labels = c("Messi", "Ronaldo")) +
  theme_minimal() +
  theme(plot.title = element_text(size = 12, face = "bold"),
        axis.title = element_text(size = 12, face = "bold"),
        axis.text = element_text(size = 12),
        legend.text = element_text(size = 14),
        legend.title = element_blank())
```

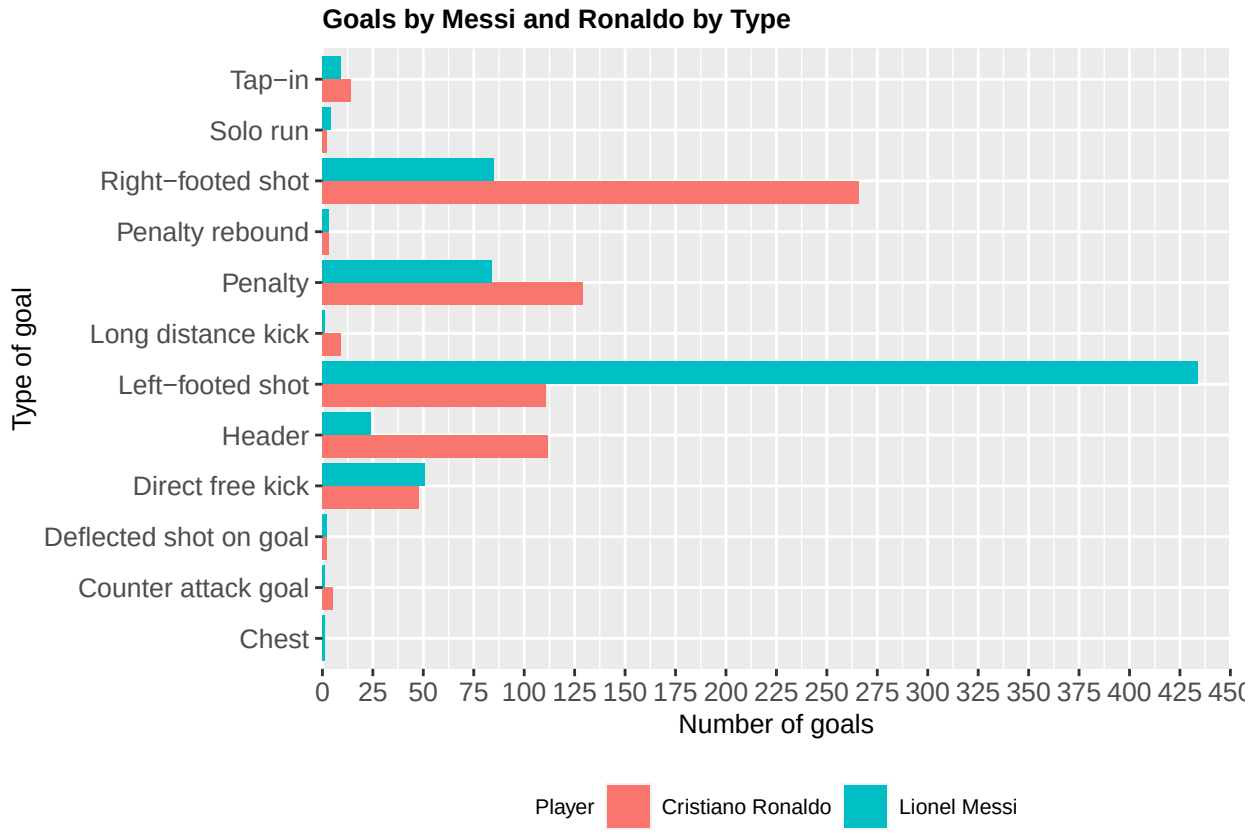


Each Type of goals

```
goal %>%
  filter(Player %in% c("Lionel Messi", "Cristiano Ronaldo")) %>%
  group_by(Player, Type) %>%
  summarise(goals = n()) %>%
```

```
ggplot(aes(x = goals, y = Type, fill = Player)) +
  geom_col(position = "dodge") +
  scale_x_continuous(expand = c(0, 0), limits = c(0, 450), breaks = seq(0, 450, 25)) +
  labs(title = "Goals by Messi and Ronaldo by Type", x = "Number of goals", y = "Type of goal", fill =
  theme(plot.title = element_text(size = 10, face = "bold"),
        axis.title = element_text(size = 10),
        axis.text = element_text(size = 10),
        legend.title = element_text(size = 8),
        legend.text = element_text(size = 8),
        legend.position = "bottom")
```

```
## `summarise()` has grouped output by 'Player'. You can override using the
## `.groups` argument.
```

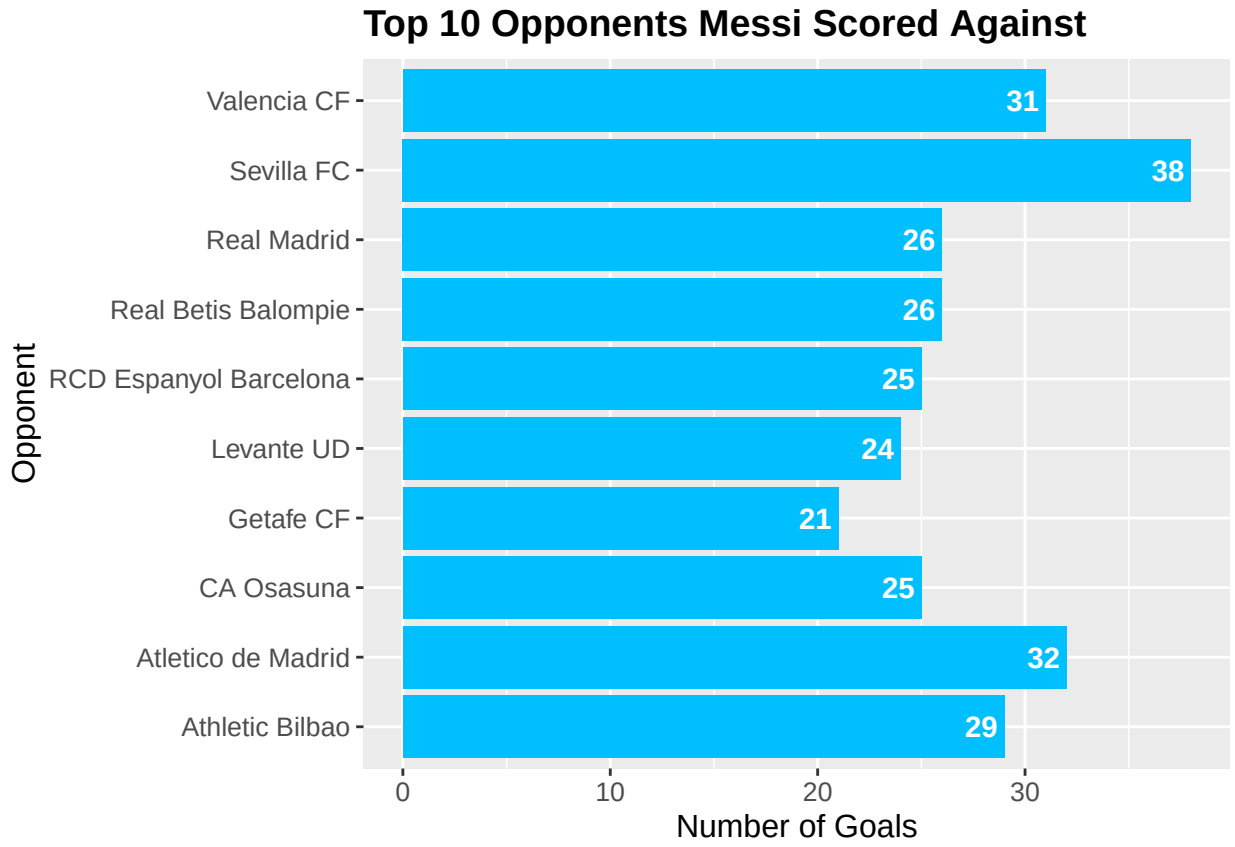


Top 10 oppenents they scored against

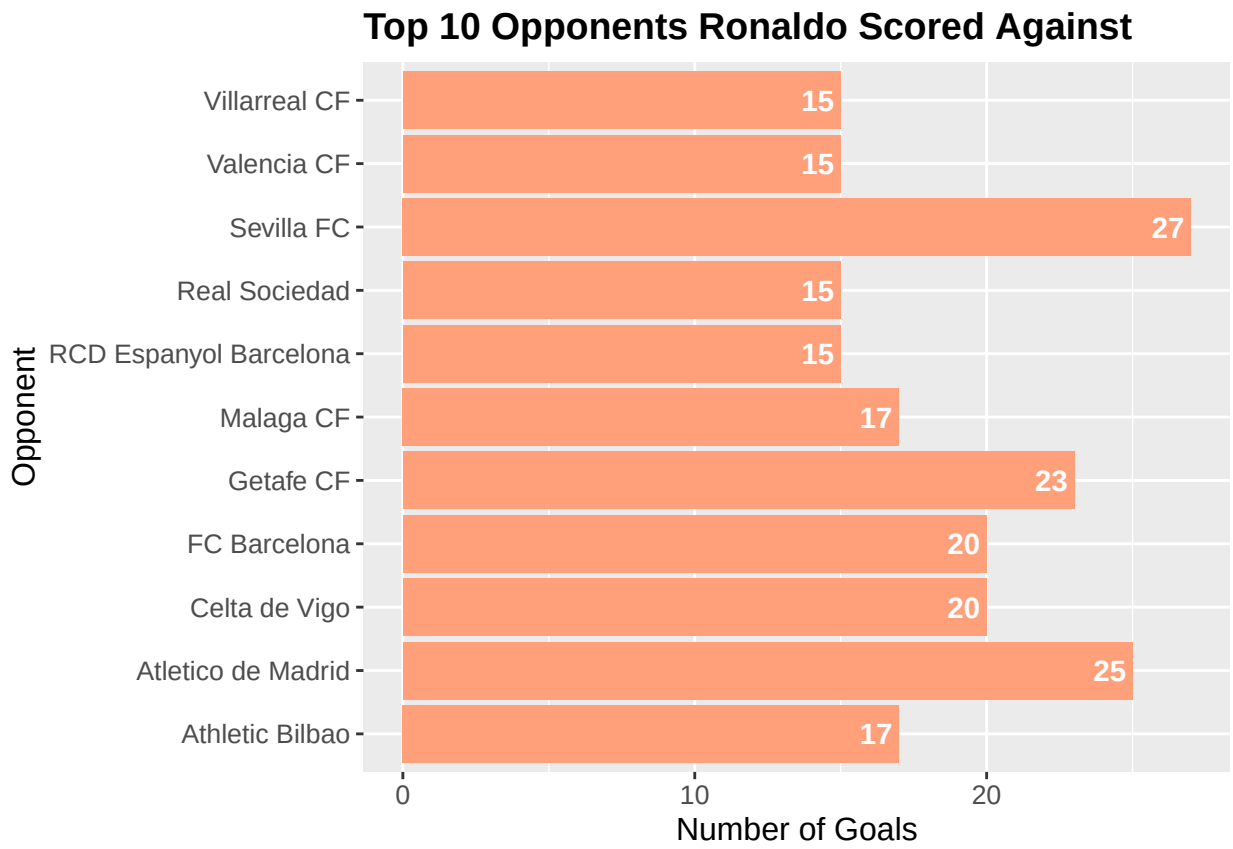
```
top_10_opponents <- goal %>%
  filter(Player %in% c("Lionel Messi", "Cristiano Ronaldo")) %>%
  group_by(Player, Opponent) %>%
  summarise(goals = n()) %>%
  arrange(Player, desc(goals)) %>%
  group_by(Player) %>%
  top_n(10)
```

```
## `summarise()` has grouped output by 'Player'. You can override using the
## `.groups` argument.
## Selecting by goals
```

```
# Create a horizontal bar plot for Messi
ggplot(filter(top_10_opponents, Player == "Lionel Messi"), aes(x = goals, y = Opponent)) +
  geom_bar(stat = "identity", fill = "#00BFFF") +
  labs(title = "Top 10 Opponents Messi Scored Against", x = "Number of Goals", y = "Opponent") +
  theme(plot.title = element_text(size = 14, face = "bold"),
        axis.title = element_text(size = 12),
        axis.text = element_text(size = 10)) +
  geom_text(aes(label = goals), hjust = 1.2, size = 4, fontface = "bold", color = "white")
```




```
# Create a horizontal bar plot for Ronaldo
ggplot(filter(top_10_opponents, Player == "Cristiano Ronaldo"), aes(x = goals, y = Opponent)) +
  geom_bar(stat = "identity", fill = "#FFA07A") +
  labs(title = "Top 10 Opponents Ronaldo Scored Against", x = "Number of Goals", y = "Opponent") +
  theme(plot.title = element_text(size = 14, face = "bold"),
        axis.title = element_text(size = 12),
        axis.text = element_text(size = 10)) +
  geom_text(aes(label = goals), hjust = 1.2, size = 4, fontface = "bold", color = "white")
```



Their Positions

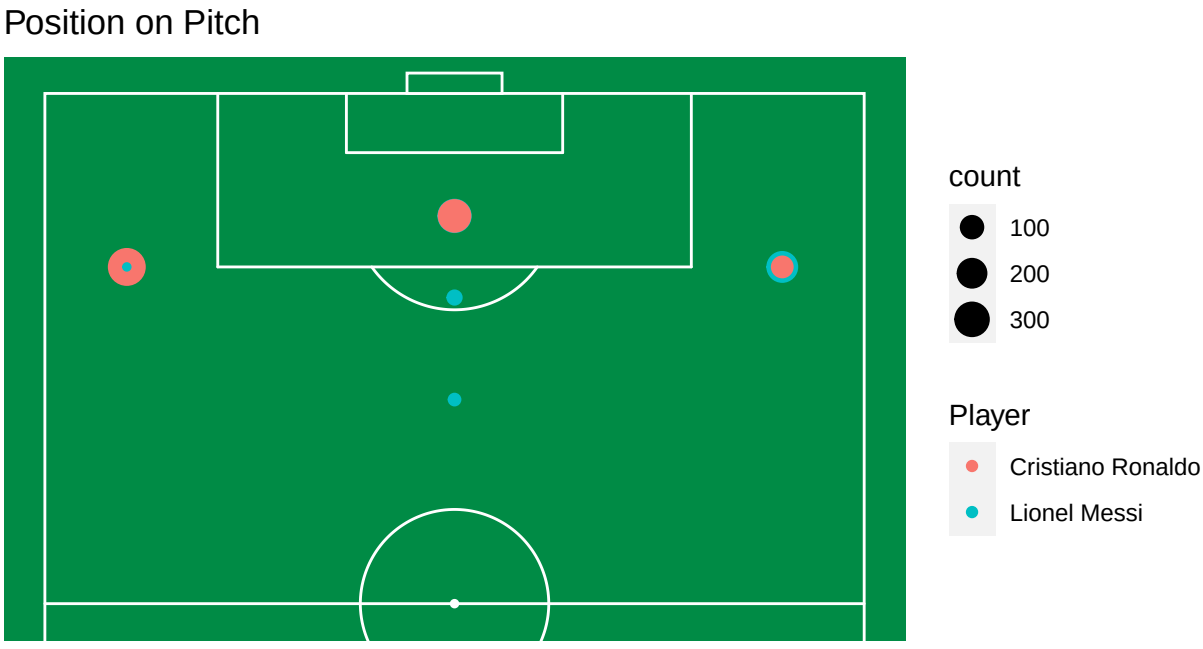
First, I define the location of each position on the pitch.

```
position_count <- as.data.frame(goal %>%
  group_by(Player, Playing_Position) %>%
  summarize(count = n()) %>%
  arrange(desc(count)) %>%
  mutate(x = case_when(
    Playing_Position == "AM" ~ 70,
    Playing_Position == "CF" ~ 88,
    Playing_Position == "LW" ~ 83,
    Playing_Position == "RW" ~ 83,
    Playing_Position == "SS" ~ 80,
    TRUE ~ NA_real_
  ),
  y = case_when(
    Playing_Position == "AM" ~ 50,
    Playing_Position == "CF" ~ 50,
    Playing_Position == "LW" ~ 90,
    Playing_Position == "RW" ~ 10,
    Playing_Position == "SS" ~ 50,
    TRUE ~ NA_real_
  ))
)
```

```
## `summarise()` has grouped output by 'Player'. You can override using the
## `.groups` argument.
```

Then plot it on pitch by using 'ggsoccer' library

```
library(ggsoccer)
ggplot(position_count) +
  annotate_pitch(colour = "white",
                 fill   = "springgreen4",
                 limits = FALSE) +
  geom_point(aes(x = x, y = y,
             colour = Player  ,
             size = count)) +
  theme_pitch() +
  theme(panel.background = element_rect(fill = "springgreen4")) +
  coord_flip(xlim = c(49, 101)) +
  scale_y_reverse() +
  ggtitle("Position on Pitch")
```



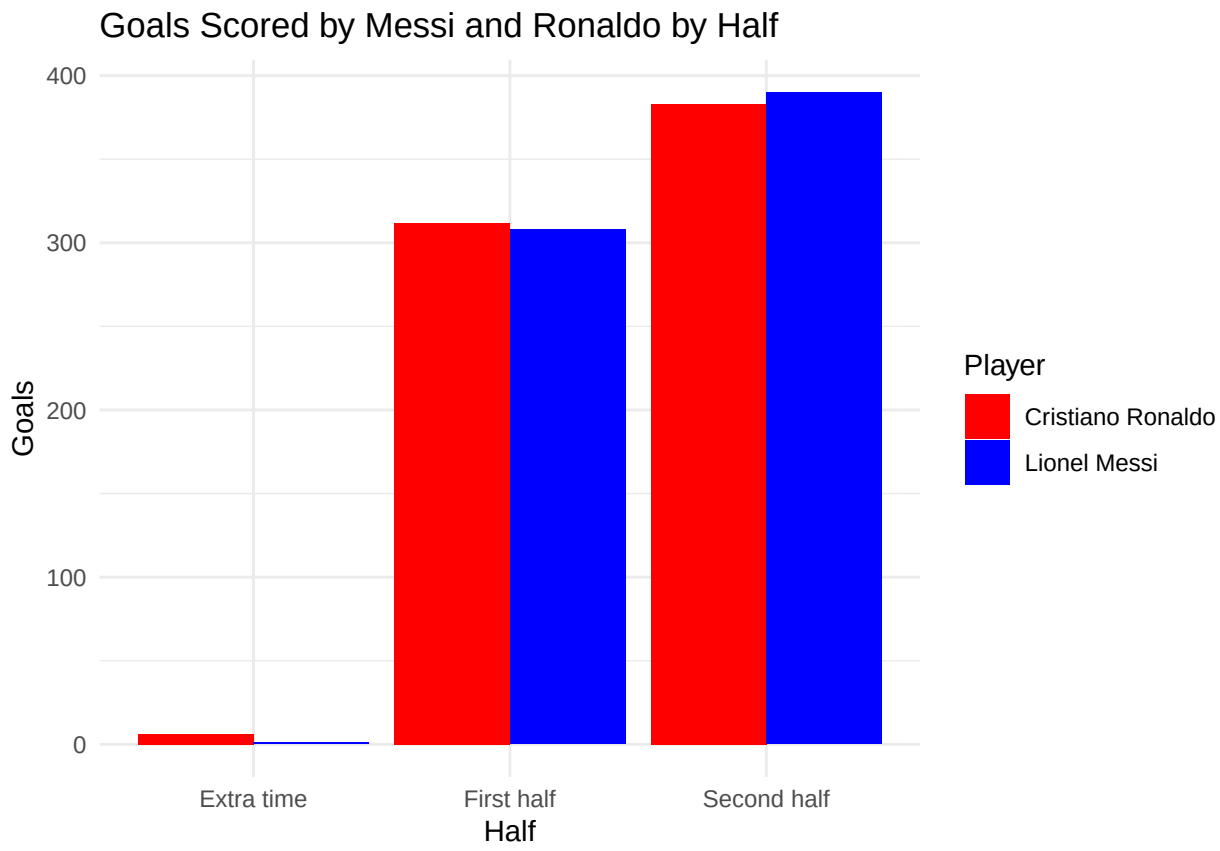
Goal scored in each Half

```
library(stringr)
goal$Minute <- as.character(goal$Minute)
```

```
goal$Minute <- ifelse(str_detect(goal$Minute, "\\+"),
                      str_extract(goal$Minute, "\\d+"),
                      goal$Minute)
goal$Minute <- as.numeric(goal$Minute)
goal <- goal %>%
  mutate(Half = case_when(
    Minute <= 45 ~ "First half",
    Minute >= 46 & Minute <= 90 ~ "Second half",
    TRUE ~ "Extra time"
  ))
half_count <- goal %>%
  filter(Player %in% c("Lionel Messi", "Cristiano Ronaldo")) %>%
  group_by(Player, Half) %>%
  summarize(Goals = n())
```

```
## `summarise()` has grouped output by 'Player'. You can override using the
## `.groups` argument.
```

```
# Create bar chart
ggplot(half_count, aes(x = Half, y = Goals, fill = Player)) +
  geom_bar(stat = "identity", position = "dodge", linewidth=0.2) +
  scale_fill_manual(values = c("red", "blue")) +
  labs(title = "Goals Scored by Messi and Ronaldo by Half",
       x = "Half", y = "Goals") +
  theme_minimal()
```



We can see that most of the time they scored in the second half of the game.