# R Notebook

```
df <- read.csv("Players_Combined_v3.csv", header=TRUE, sep=",")
```

```
library(tidyverse)
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr
```

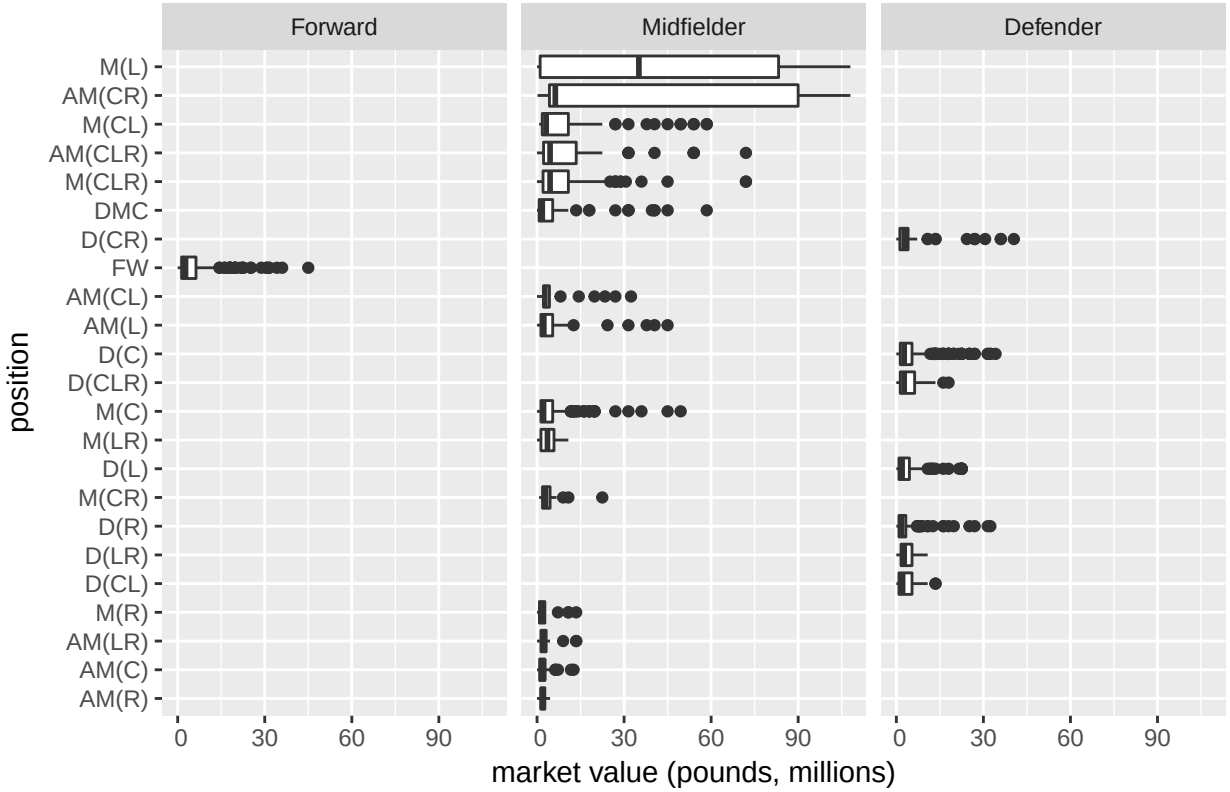
```
## Conflicts with tidy packages ----------------------------------------------
```

```
## filter(): dplyr, stats
## lag():    dplyr, stats
```

```
df3 <- df
df3$position <- trimws(as.character(df3$position))
df3 <- df3 %>% filter(position != "Forward" & position != "Midfielder" & position != "Defender")
df3$pos_type <- factor(df3$pos_type, levels=c("Forward", "Midfielder", "Defender"))

ggplot(df3, aes(x=reorder(position, market_val), y=market_val)) + geom_boxplot() +
  coord_flip() + ggtitle("Figure 1. Boxplot of Market Values by Position") +
  xlab("position") + ylab("market value (pounds, millions)") +
  scale_y_continuous(labels = function(x) format(x/1000000)) + facet_wrap(~pos_type)
```

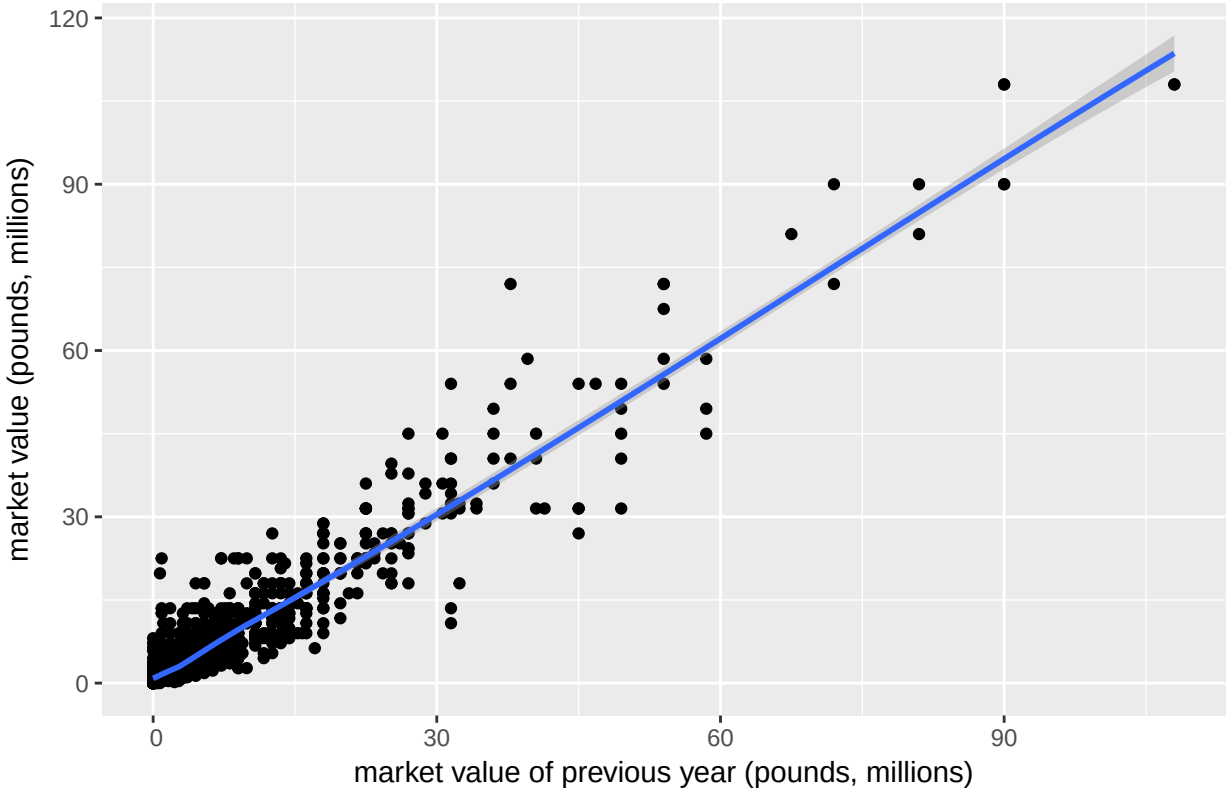
Figure 1. Boxplot of Market Values by Position




Contrary to popular beliefs that forward is the highest paying position in soccer, we can see from Figure 1 that AM(CR) or attacking midfielder center right and M(L) or left midfielder are the highest compensated groups. It is worth noting that superstars such as Lionel Messi and Cristiano Ronaldo can play multiple positions, and they can play either as midfielder or forward depending on the opposing matchup and immediate need of their teams. Therefore the inclusion of those superstars in the aforementioned midfielder categories indeed inflates the overall distribution. And it is clear from the plot that defenders do on average have lower market values than other positions, but there are still plenty of outliers where teams are willing to pay in excess of 10M pounds for a top defender.

```
ggplot(df, aes(x=market_val_prev, y=market_val)) + geom_point() +
  scale_y_continuous(labels = function(x) format(x/1000000)) +
  scale_x_continuous(labels = function(x) format(x/1000000)) +
  geom_smooth(method="loess")  +
  ggtitle("Figure 2. Market Value vs. Market Value of Previous Year") +
  xlab("market value of previous year (pounds, millions)") +
  ylab("market value (pounds, millions)")
```

Figure 2. Market Value vs. Market Value of Previous Year




According to Figure 2, market value of previous year is highly positively correlated with current market value. Looking at predicted market value from actuals, the differences are not substantial. Most players though did see their valuations move up or down in a year, and some see considerable increases or decreases due to possibly a variety of factors, such as injury, breakout season, matching between immediate club need and individual skillsets.

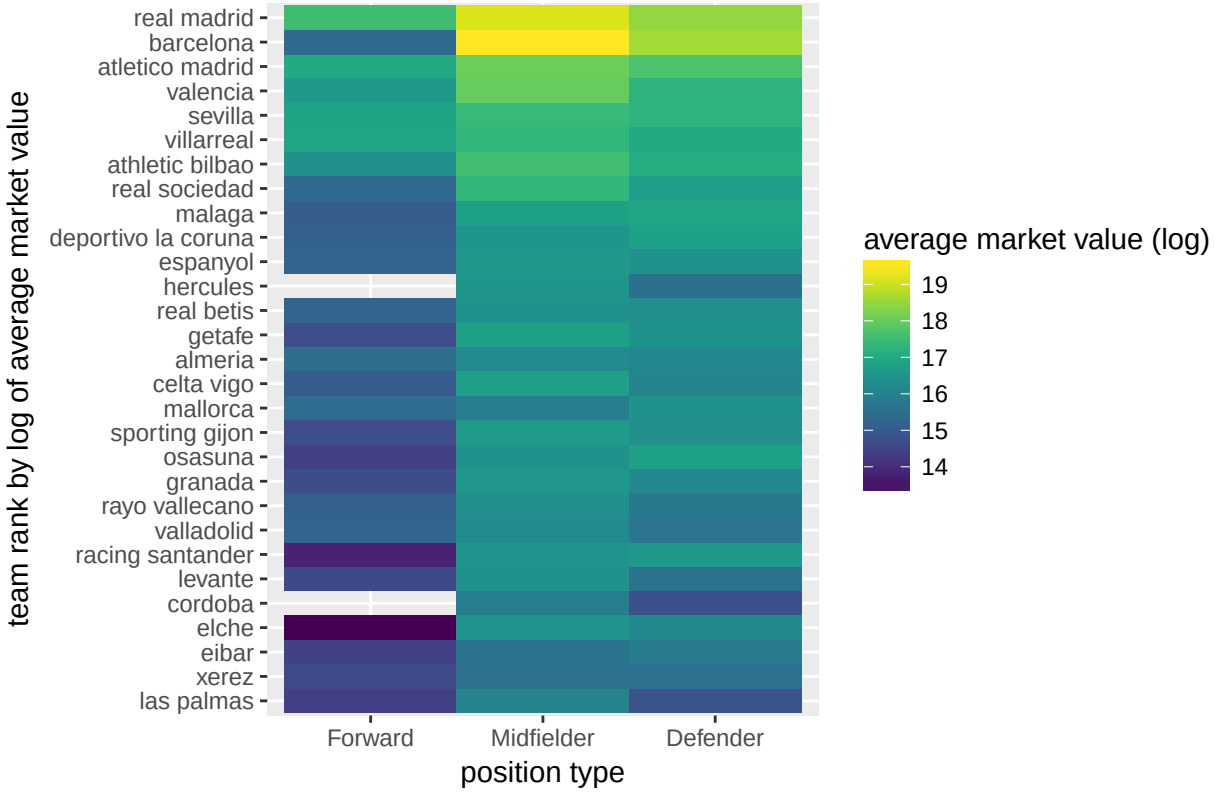
```
library(viridis)
df$pos_type <- factor(df$pos_type, levels=c("Forward", "Midfielder", "Defender"))

df2 <- df %>% group_by(team, pos_type, season) %>%
  summarise(total_mkt_val = sum(market_val, na.rm=TRUE))

df2 <- df2 %>% group_by(team, pos_type) %>%
  summarise(avg_mkt_val_per_season = mean(total_mkt_val, na.rm=TRUE))
df2$avg_mv_log <- log(df2$avg_mkt_val_per_season)

ggplot(df2, aes(y=reorder(team, avg_mv_log), x=pos_type)) +
  geom_tile(aes(fill=df2$avg_mv_log)) +
  scale_fill_viridis() + labs(fill = "average market value (log)") +
  xlab("position type") + ylab("team rank by log of average market value") +
  ggtitle("Figure 3. Heatmap of Market Value By Teams and Position Type")
```

Figure 3. Heatmap of Market Value By Teams and Position Type




Shown by Figure 3 above, the La Liga is highly polarized in terms of club wealth. Four teams stand out in terms of total net worth, Barcelona, Real Madrid, Valencia and Atletico Madrid. While the less wealthy teams can only spend a lot less and fight for survival in the lower rung of league ranking, top teams have the ability to pour lavish sums of cash to acquire megastars and dominate the ranking for years. Examining the graph closely, we can also tell the personnel priority of each team, where Barcelona placed the most emphasis on its midfield lineup, Real Madrid chose to focus on building a formidable frontline.

```
library(gridExtra)
```

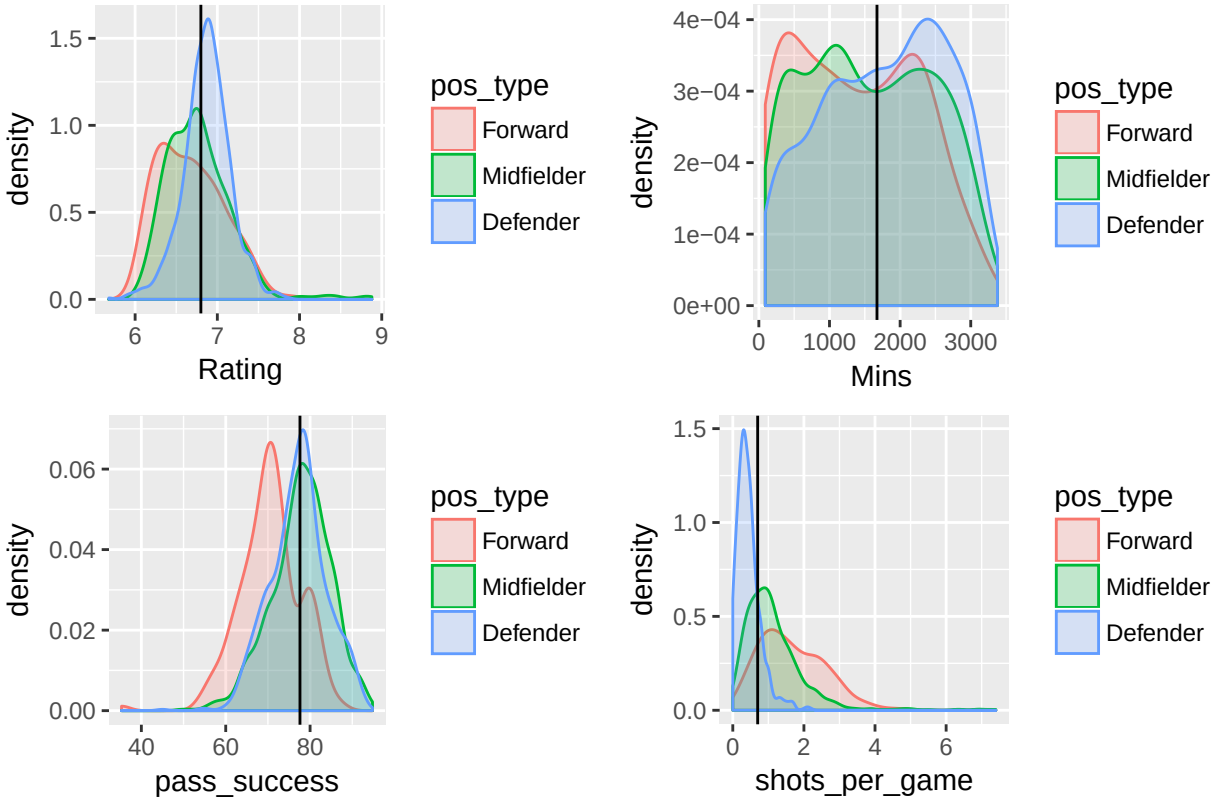
```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
p1 <- ggplot(df, aes(Rating, color=pos_type, fill=pos_type)) +
  geom_density(alpha=0.2) + geom_vline(aes(xintercept=median(df$Rating)))
p2 <- ggplot(df, aes(Mins, color=pos_type, fill=pos_type)) +
  geom_density(alpha=0.2) + geom_vline(aes(xintercept=median(df$Mins)))
p3 <- ggplot(df, aes(pass_success, color=pos_type, fill=pos_type)) +
  geom_density(alpha=0.2) + geom_vline(aes(xintercept=median(df$pass_success)))
p4 <- ggplot(df, aes(shots_per_game, color=pos_type, fill=pos_type)) +
  geom_density(alpha=0.2) + geom_vline(aes(xintercept=median(df$shots_per_game)))

grid.arrange(p1, p2, p3, p4, ncol=2, top = "Figure 4. Density Plot of Performance by Position")
```

Figure 4. Density Plot of Performance by Position




Based on Figure 4, it is somewhat surprising that defenders have the highest median rating, though midfielders who can play multiple positions dominate the higher end of the rating spectrum. Defenders also play the most minutes, where substantial portion of forwards and midfielders play less than 1500 minutes in a season, and the phenomenon is attributed to managers' tendencies to substitute forwards and midfielders late in the game to boost offensive output, especially when the team is down a goal or needs to break the draw. Fowards on average have lower passing success rate than others, as passing accurately is more difficult as ball advances closer to the opponent's goal. But there is a small group of elite forwards can pass at over 80%. Forwards and midfielders understandably also have more shots per game than defenders, thus shouldering the main responsibilities in creating and making shots.