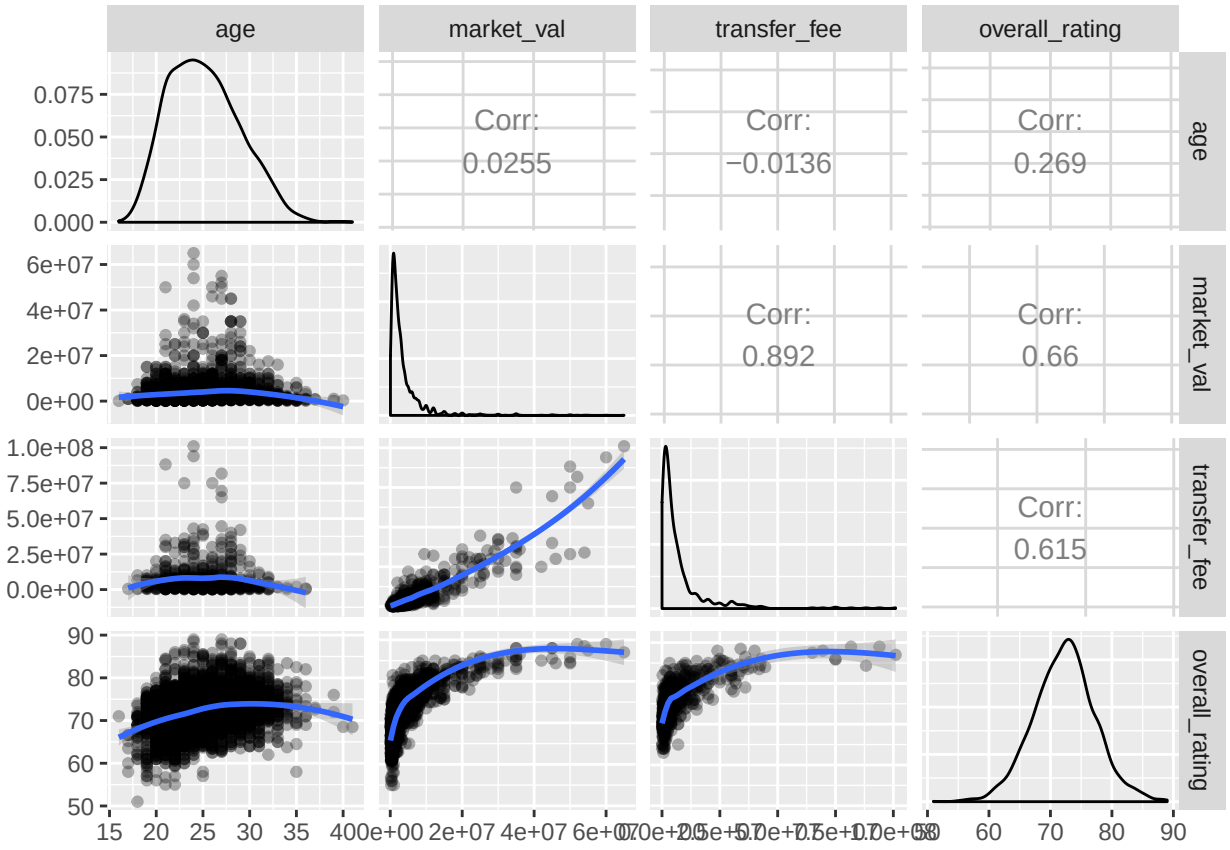
Exploration of Stats and Financial Figures by Player and Season

```
options(warn=-1)

df <- read.csv(file="Main_Data.csv", header=TRUE, sep=",")
df$market_val <- as.numeric(as.character(df$market_val))

get_transfer_type <- function(value) {
  if (!is.na(as.numeric((value)))) {
    return ("Priced transfer")
  }
  else {
    if (value=="Loan"|value=="Loan fee:")
      return ("Loan")
    else if (value=="Free transfer"|value=="End of loan")
      return (value)
    else
      return (NA)
  }
}

df$transfer_fee <- as.character(df$transfer_fee)
df$transfer_type <- mapply(get_transfer_type, df$transfer_fee)
df$transfer_fee <- as.numeric(df$transfer_fee)
df$transfer_type <- as.factor(df$transfer_type)

get_pos_type <- function(value) {
  offense <- c("CF","SS")
  middle <- c("AM", "LW", "RW", "CM", "DM", "RM", "LM")
  defense <- c("CB", "RB", "LB")
  goalkeep <- c("GK")
  if (value %in% offense)
    return ("striker")
  else if (value %in% middle)
    return ("midfielder")
  else if (value %in% defense)
    return ("defender")
  else if (value %in% goalkeep)
    return ("goalkeeper")
  else
    return (NA)
}

df$pos_type <- mapply(get_pos_type, df$pos)
df$pos_type <- as.factor(df$pos_type)
```

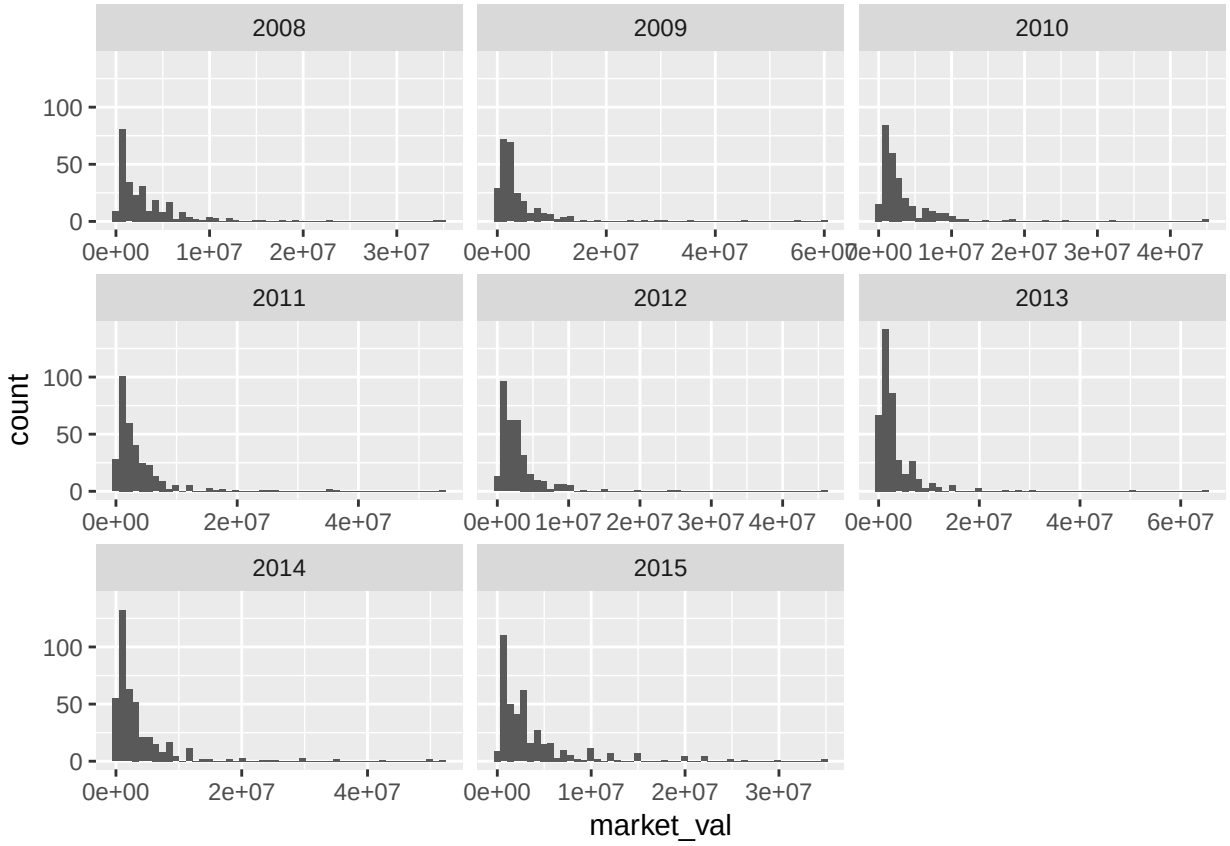
```
library(GGally)
library(ggplot2)

my_fn <- function(data, mapping, ...){
```

```
  p <- ggplot(data = data, mapping = mapping) +
    geom_point(alpha = 0.3) +
    geom_smooth(method = loess)
  p
}

ggpairs(df[c(1,6,12,17)], lower=list(continuous=my_fn))
```



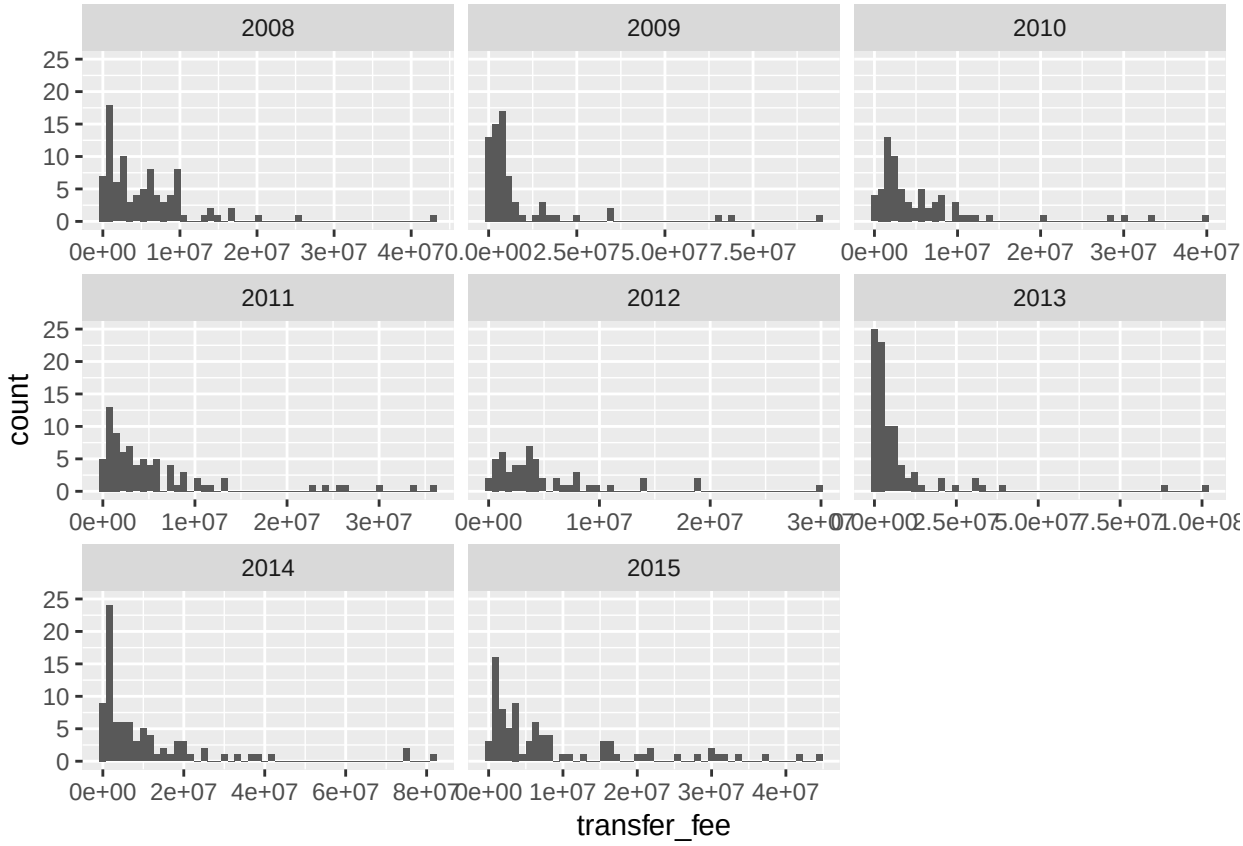
From the scatterplot matrix above, it is as expected that age and ratings follow a roughly Gaussian distribution, while market values and transfer fees are highly skewed to the right. Transfer fee is highly correlated with market value, and rating is also positively associated with both transfer fee and market value. Though it is interesting to see that players of similar ratings, especially towards the high end of the spectrum, can have drastically different valuations, therefore ratings alone can not fully explain valuations. Looking at ratings vs age, we can see that players tend to peak between age 25 and 30, and then their performance level off after that. Age on the other hand is not directly correlated with transfer fee and market value, though some of the highest-figure transactions center around age 25.

```
ggplot(df, aes(x=market_val)) + geom_histogram(bins=50) +
  facet_wrap(~season, scale="free_x")
```



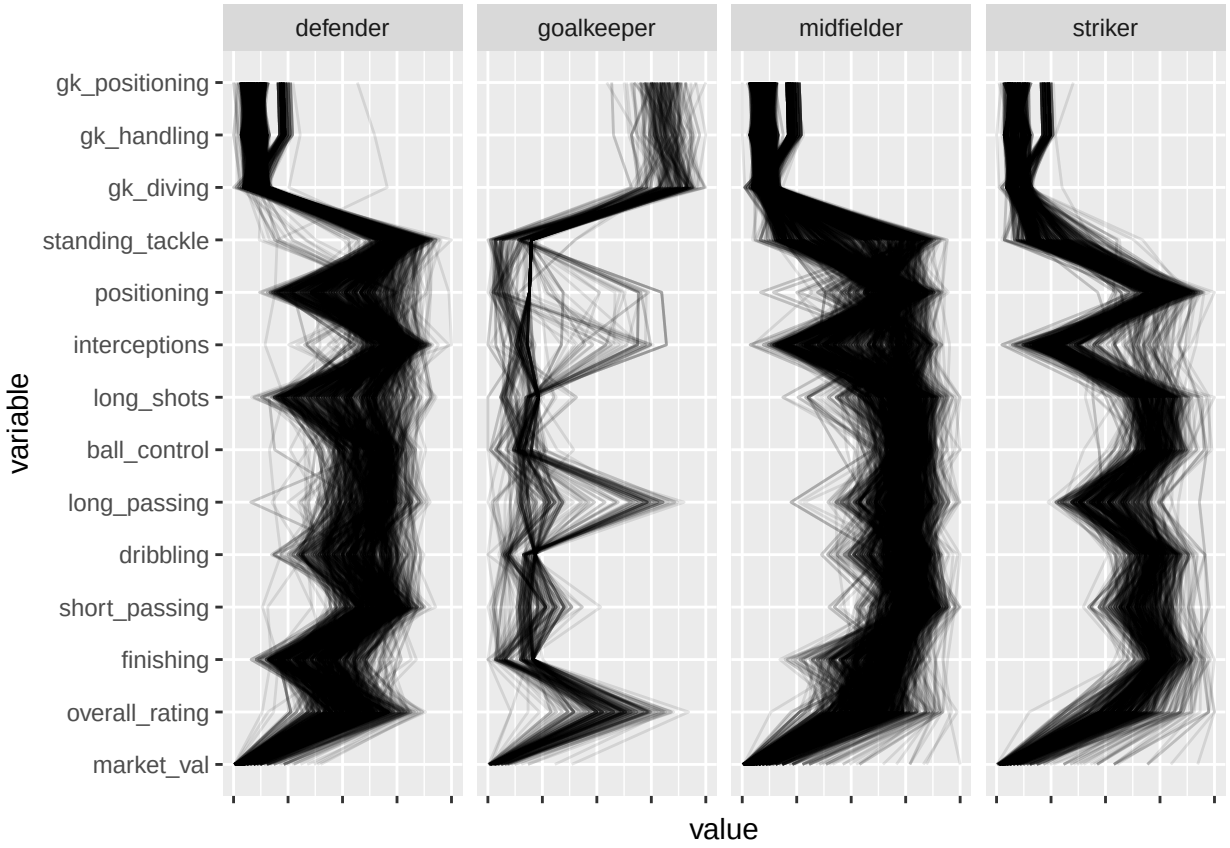
There seems to be progressively more transfers as time goes on, while majority of players among transfers have market values of less than 5M Euros, there are more and more transfers with price tags higher than 10M Euros in the later seasons.

```
ggplot(df, aes(x=transfer_fee)) + geom_histogram(bins=50) +
  facet_wrap(~season, scale="free_x")
```



The 2013-2014 and 2014-2015 season looks like a time of economic slump for the La Liga as most of these transfer congregated near the low-end, but a few record-shattering transactions also appeared in 2013, with values close to 100M Euros. In 2015-2016 season, players' market values rise in tandem with their transfer fees, as the distribution of transfer fees flattens out somewhat comparing to 2008-2009 season.

```
cat_names <- c('1' = "defender", '2' = "goalkeeper", '3' = "midfielder", '4' = "striker")

ggparcoord(data=df, columns=c(6,17,21,23,25,28,29,39,41,42,46,48,49,51),
  scale="uniminmax", alpha=0.1) +
  coord_flip() + theme(axis.text.x=element_blank()) +
  facet_grid(.~pos_type, labeller=as_labeller(cat_names))
```

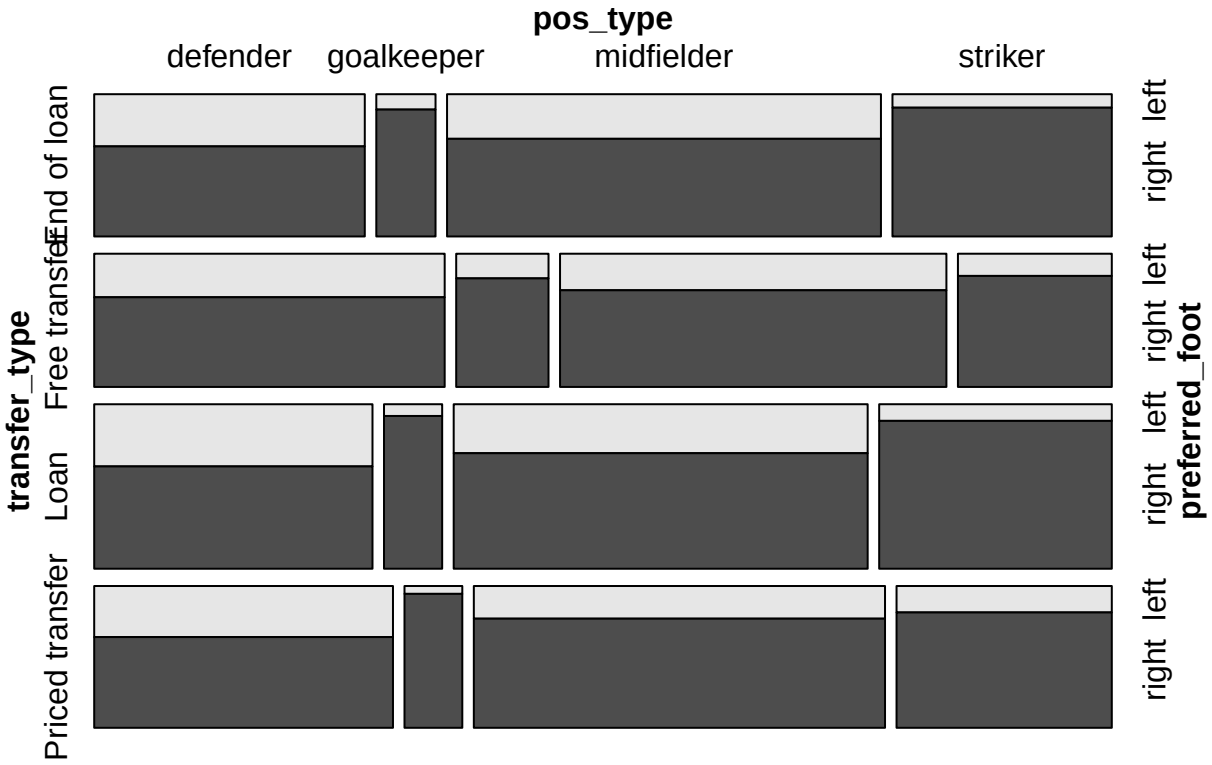


Looking at the parallel coordinate plot above, different metrics matter differently depending on the player's position. Defenders tend to rated highly on standing tackles and interception, though many of them have poor ratings on positioning. Midfielders vary within the group in the ability to intercept and tackle, but overall have decent capabilities in maintaining possession, as measured by ball_control, dribbling and short passing. Strikers, on the other hand, have superior positioning and highly adept at finishing. So here we can see though each metric is relevant to all positions, disparate weights should be applied in forming the overall performance index.

```
library(vcd)
```

```
## Loading required package: grid
```

```
vcd::mosaic(preferred_foot ~ transfer_type + pos_type, data=df )
```



Plotting a couple of categorical variables from the data set, it is interesting to see that defenders have the highest proportion of left-footers, who may play on the left wing. But there are very few left-foot strikers and goalkeepers. In priced transfers, strikers are sough after, usually with high price tags. But forwards are not as popular in free transfers, when contracts tend to expire, defenders and goalkeepers occupy a larger percentage. Overall in any transfer types, midfielders remain the most numerous type, a testament to their versatility in playing a variety of roles.