Esp_football database analysis

Hajdú Márk

2022 03 21

Setting up my environment

Setting up R environment by loading the `tidyverse` and `readxl` packages and setting the working directory

```
library( "readxl" )
library("tidyverse")
setwd("C:/Users/Márk/Documents/okonometria2")
```

Data importing

Importing the excel file which contains the database

```
file<-"c:/Users/Márk/Documents/okonometriajo/ESPfootball.xlsx"
Esp_football<- read_excel( "ESPfootball.xlsx", sheet= 1)
```

Data exploration

Previewing the data

```
str(Esp_football)
```

```
## tibble [260 x 11] (S3: tbl_df/tbl/data.frame)
##  $ Position        : chr [1:260] "Forward" "Forward" "Defender" "Defender"
...
##  $ PricemillionEuro: chr [1:260] "100" "80" "70" "70" ...
##  $ Age             : num [1:260] 33 29 25 27 26 28 23 28 24 28 ...
##  $ YellowCards     : num [1:260] 0.372 0.141 0.254 0 0.308 ...
##  $ RedCards        : num [1:260] 0.0465 0 0 0.0327 0 ...
##  $ Goalsscored     : num [1:260] 0.0465 0.4216 0 0.0653 0.1231 ...
##  $ Foulscommitted  : num [1:260] 1.395 0.506 0.406 0.359 1.692 ...
##  $ Shots           : num [1:260] 0.233 2.417 0.66 0.392 1.231 ...
##  $ Shotsontarget   : num [1:260] 0.093 1.293 0.152 0.131 0.615 ...
##  $ Assists         : num [1:260] 0 0.253 0 0 0.0308 ...
##  $ Passes          : num [1:260] 53.3 39 36.5 55.8 45.6 ...
```

Data cleaning

Changing data types

The Position and the Price are in a bad data type, I change them in order to do the analysis

```
Esp_football$Position<-as.factor(Esp_football$Position)
Esp_football$PricemillionEuro<-as.numeric(Esp_football$PricemillionEuro)
```

Basic statistics about the data

```
summary(Esp_football)
```

```
##        Position   PricemillionEuro       Age         YellowCards
##  Defender  :102   Min.   :    0.5   Min.   :21.00   Min.   :0.0000
##  Forward   : 52   1st Qu.:    4.0   1st Qu.:26.00   1st Qu.:0.1504
##  Midfielder:106   Median :   13.5   Median :29.00   Median :0.2250
##                   Mean   : 9538.8   Mean   :28.47   Mean   :0.2462
##                   3rd Qu.:   60.0   3rd Qu.:31.00   3rd Qu.:0.3291
##                   Max.   :44382.0   Max.   :38.00   Max.   :0.6294
##     RedCards        Goalsscored      Foulscommitted      Shots
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.02851   1st Qu.:0.905   1st Qu.:0.3672
##  Median :0.00000   Median :0.07125   Median :1.240   Median :0.6731
##  Mean   :0.01089   Mean   :0.12722   Mean   :1.338   Mean   :0.9376
##  3rd Qu.:0.00000   3rd Qu.:0.16168   3rd Qu.:1.699   3rd Qu.:1.3514
##  Max.   :0.10989   Max.   :1.19557   Max.   :3.707   Max.   :4.6494
##  Shotsontarget       Assists           Passes
##  Min.   :0.0000   Min.   :0.00000   Min.   :   15.42
##  1st Qu.:0.1180   1st Qu.:0.00000   1st Qu.:   30.86
##  Median :0.2628   Median :0.06396   Median :   40.01
##  Mean   :0.4179   Mean   :0.08537   Mean   :  286.28
##  3rd Qu.:0.6156   3rd Qu.:0.13378   3rd Qu.:   49.95
##  Max.   :2.8893   Max.   :0.43173   Max.   :44377.00
```

Filtering out the wrong data

There are rows that contain unreal, very big vaules, these are probably wrong data. Now I will filter them out

```
Esp_football_filtered<- Esp_football %>% filter(PricemillionEuro<101 &
                          Esp_football$Passes<100)
```

Basic statistics about the filtered database

```
summary(Esp_football_filtered)
```

```
##        Position  PricemillionEuro      Age         YellowCards
##  Defender  :78   Min.   :  0.50   Min.   :21.00   Min.   :0.0000
##  Forward   :46   1st Qu.:  3.00   1st Qu.:25.50   1st Qu.:0.1683
##  Midfielder:79   Median :  9.00   Median :28.00   Median :0.2398
##                  Mean   : 14.62   Mean   :28.18   Mean   :0.2521
##                  3rd Qu.: 20.00   3rd Qu.:31.00   3rd Qu.:0.3322
##                  Max.   :100.00   Max.   :35.00   Max.   :0.6294
##     RedCards         Goalsscored      Foulscommitted       Shots
##  Min.   :0.000000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.000000   1st Qu.:0.01389   1st Qu.:0.9112   1st Qu.:0.3921
##  Median :0.000000   Median :0.07189   Median :1.2398   Median :0.6970
##  Mean   :0.009947   Mean   :0.13615   Mean   :1.3447   Mean   :0.9843
```

```
##  3rd Qu.:0.000000   3rd Qu.:0.17515   3rd Qu.:1.7050   3rd Qu.:1.4200
##  Max.   :0.109890   Max.   :1.19557   Max.   :3.7074   Max.   :4.6494
##  Shotsontarget       Assists            Passes
##  Min.   :0.0000   Min.   :0.00000   Min.   :15.49
##  1st Qu.:0.1304   1st Qu.:0.00000   1st Qu.:29.53
##  Median :0.2797   Median :0.06625   Median :39.85
##  Mean   :0.4480   Mean   :0.08767   Mean   :41.30
##  3rd Qu.:0.6732   3rd Qu.:0.13684   3rd Qu.:50.65
##  Max.   :2.8893   Max.   :0.43173   Max.   :86.32
```

Now there are not any very big outliers, I can start the analysis

Linear regression

I will make a linear regression, the result variable will be the PriceMillionEuro

```
model_1 <- lm(PricemillionEuro ~ Position + Age + YellowCards + RedCards +
Goalsscored +
                Foulscommitted + Shots + Shotsontarget + Assists + Passes,
data = Esp_football_filtered)
summary(model_1)

##
## Call:
## lm(formula = PricemillionEuro ~ Position + Age + YellowCards +
##     RedCards + Goalsscored + Foulscommitted + Shots + Shotsontarget +
##     Assists + Passes, data = Esp_football_filtered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.061  -8.800  -2.943   3.333  77.607
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        47.84699    9.60019   4.984 1.39e-06 ***
## PositionForward    12.67555    3.97551   3.188  0.00167 **
## PositionMidfielder  2.82943    2.69038   1.052  0.29427
## Age                -1.83886    0.31312  -5.873 1.87e-08 ***
## YellowCards        -0.79194   10.56335  -0.075  0.94032
## RedCards           58.29610   50.16596   1.162  0.24666
## Goalsscored        27.30290   13.31812   2.050  0.04173 *
## Foulscommitted     -3.29851    2.27073  -1.453  0.14797
## Shots              -2.67360    4.20811  -0.635  0.52596
## Shotsontarget      -2.59148    9.08496  -0.285  0.77576
## Assists            -1.31150   14.13668  -0.093  0.92618
## Passes              0.45688    0.08004   5.708 4.31e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.87 on 191 degrees of freedom
```

```
## Multiple R-squared:  0.284,  Adjusted R-squared:  0.2428
## F-statistic: 6.888 on 11 and 191 DF,  p-value: 9.608e-10
```

There are a lot of insignificant variables, that means these variables are not in correlation with the result variable. I will apply the backward elimination and I will only show the final model

Final model

```
model_8 <- lm(PricemillionEuro ~ Position + Age
              + Passes, data = Esp_football_filtered)
summary(model_8)

##
## Call:
## lm(formula = PricemillionEuro ~ Position + Age + Passes, data =
Esp_football_filtered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.293  -8.189  -3.156   2.934  78.031
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        40.18263    9.07553   4.428 1.57e-05 ***
## PositionForward    13.07929    3.05332   4.284 2.87e-05 ***
## PositionMidfielder  0.49618    2.40143   0.207    0.837
## Age                -1.71815    0.30761  -5.586 7.62e-08 ***
## Passes              0.47706    0.07775   6.136 4.53e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.98 on 198 degrees of freedom
## Multiple R-squared:  0.2465, Adjusted R-squared:  0.2313
## F-statistic: 16.19 on 4 and 198 DF,  p-value: 1.724e-11
```

All variables are significant, except the midfielder dummy. I leave it in the model, because I do not want to change the reference category. The model has a low R2, it needs a lot of improvement and new variables.

Data visualization

The relationship between the age and the market value

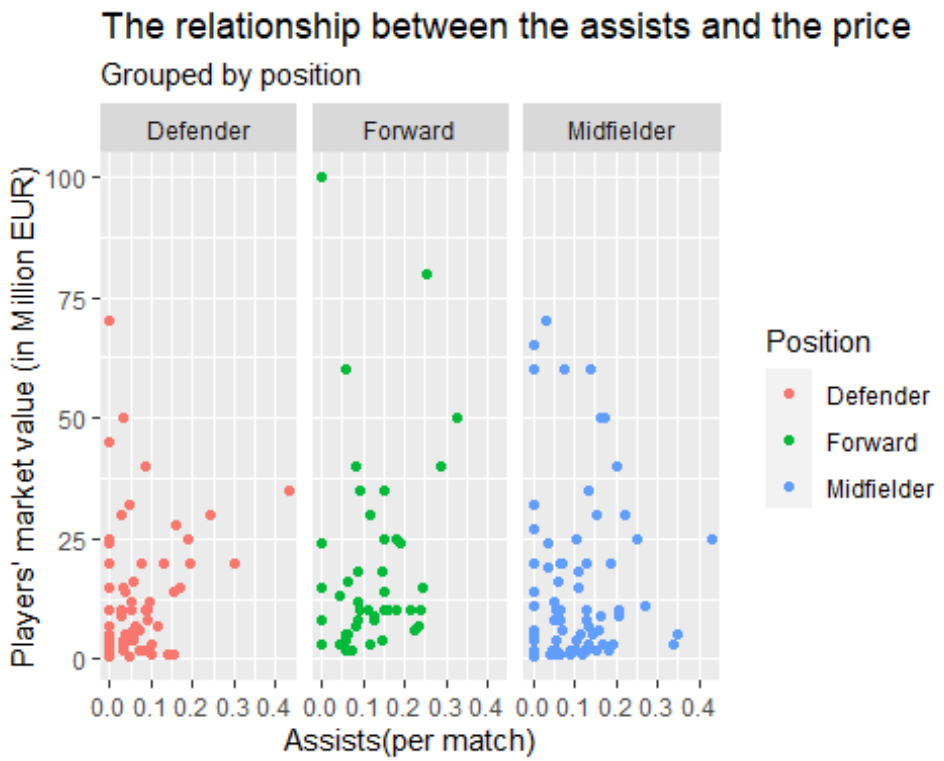
```
library(ggplot2)
ggplot(data = Esp_football_filtered, mapping = aes(x = Age, y =
PricemillionEuro)) +
  geom_point(color="blue")+
  labs(
    title = "The relationship between the age and the price",
    subtitle = "There is a parabola relationship ",
```

```
    x = "Age",
    y = "Players' market value (in Million EUR)")
```

It would be proper to include the square of the age variable also and it would model that parabola line

The relationship between the assists and the market value

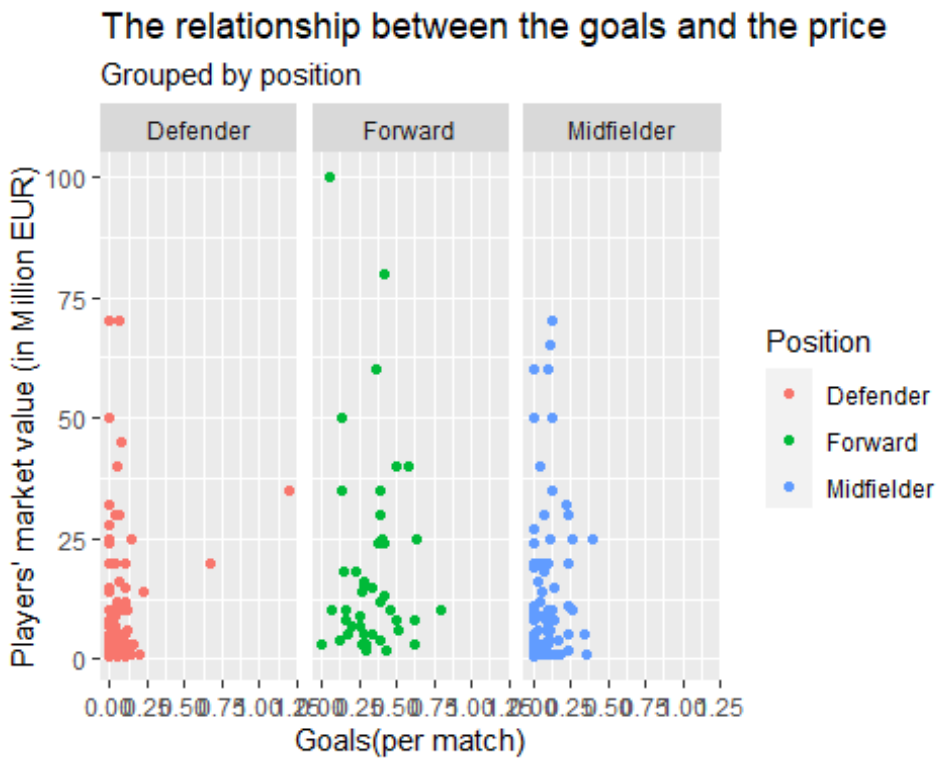
```
ggplot(data = Esp_football_filtered, mapping = aes(x = Assists, y =
PricemillionEuro,color=Position)) + facet_wrap(~Position)+
  geom_point()+
  labs(
    title = "The relationship between the assists and the price",
    subtitle = "Grouped by position",
    x = "Assists(per match)",
    y = "Players' market value (in Million EUR)")
```



There is a surprising negative correlation betewwen the assist and the market value among the midfielders. There is a positive correlation among the forwards. Maybe that is the reason why the assists variable is not significant.

The relationship between the goals and the market value

```
ggplot(data = Esp_football_filtered, mapping = aes(x = Goalsscored, y =
PricemillionEuro,color=Position)) + facet_wrap(~Position)+
  geom_point()+
  labs(
    title = "The relationship between the goals and the price",
    subtitle = "Grouped by position",
    x = "Goals(per match)",
    y = "Players' market value (in Million EUR)")
```

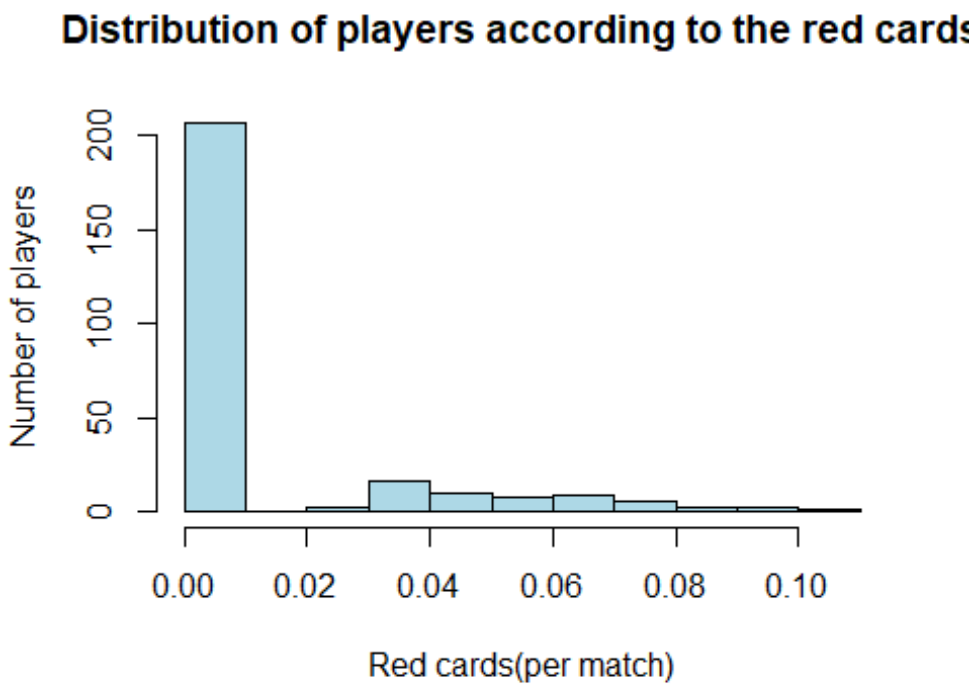


The goals variable is definitely a significant variable among the forwards. It would be worth to group the players by position and do separate linear regressions.

Histogram

```
hist(Esp_football$RedCards,

     main="Distribution of players according to the red cards",
     xlab="Red cards(per match)",
     ylab = "Number of players",
     col = "lightblue")
```



The most players never get a red card. So every player will be an outlier if they get a red card.

Key takeaways

- This final model is a good base model for further analysis
- According to the final model, the position,age and passes are the most important variables if we analyze the players' marekt value
- It would be worth to analyze the players separately, grouped by position
- The square of the age and other new variables might increase the explanatory power