STATISTICS LABORATORY PROJECT

TEAM:

DOMENICO SARCINA

GIUSEPPE NAPOLETANO

VALENTÍN STOYANOV TSVETANOV

ALEX GAYNETDINOV

Introduction

We are going to work on the dataset mentioned below, which consist in a set of Players with the corresponding stats and information for each one.
The variables contained in the dataset and which we will analyze are:

- **ID:** which specify each player
- **Name:** is the First Letter of the name and the surname
- **Age:** is the age of the player
- **OVA:** is the current overall of the player, max is 99
- **Nationality:** is the nationality of the player
- **Club:** id the current club where the player is playing, it's blank if the player doesn't have a club
- **BP:** (Best Position) is the position where the player has the greatest overall, if we put a player in a different position than his natural position, the overall will go down
- **POT:** the best overall that the player can reach in his career, players who are at the end of their career, have already reached their maximum potential
- **Height:** player height, given in feet
- **Weight:** player weight, given in libras
- **Foot:** whether the player is right or left-handed
- **Total.Stats:** sum of all player stats
- **Sprint.Speed:** the maximun speed that a player can reach

This is the URL of the dataset: https://www.kaggle.com/ekrembayar/fifa-21-complete-player-dataset

Random Sample

We use random samples to see how young players have lower overall rankings than players at the top of their career.

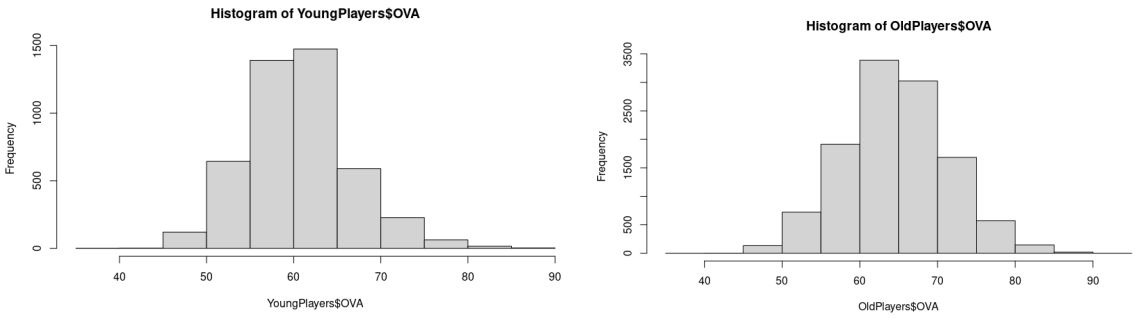
We take players that are younger than 22 years old, and players that are older than 28 years old.

```
YoungPlayers <- subset (Players, Age<22)
OldPlayers <- subset (Players, Age>28)
```

We select young players and old players with the function subset and we make 2 histograms, one for young players and one for old players.

```
hist(YoungPlayers$OVA)
hist(OldPlayers$OVA)
```

The results are:



As we can see, the average overall of Old Players is higher than the average overall of Young Players, this is because young players reach their maximum potential later on.

Overall Average of Old Players more or less is 70, while Overall Average of Young Players more or less is 60.

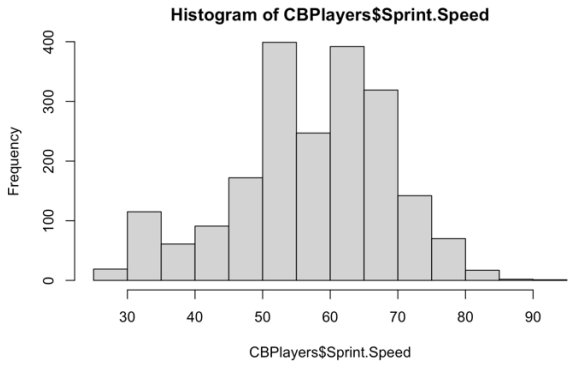
Another random sample is the different of the Sprint Speed between Center Backs (CB) and left and right full backs (RB and LB).

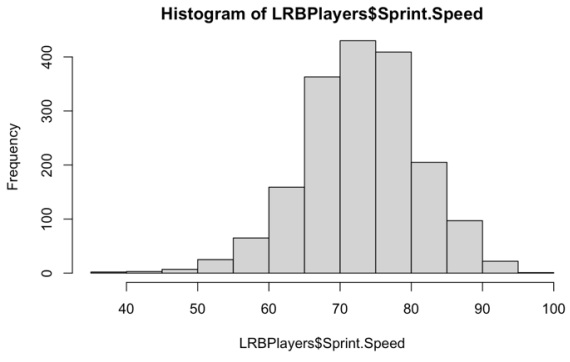
The full backs have to be faster than the central defenders, we can analyze the data.

```
CBPlayers <- subset(Players, Position=='CB')
LRBPlayers <- subset(Players, Position=='LB' | Position=='RB' | Position=='RB RM' | Position
=='LB LM' | Position=='LB LWB' | Position=='RB RWB' | Position=='RB LB')
```

We select the CBs and the LBs and RBs together.

```
hist(CBPlayers$Sprint.Speed)
hist(LRBPlayers$Sprint.Speed)
```





As we can see, CBs are slower than LBs and RBs.

We do a summary to check the average sprint speed of CBs, LBs and RBs.

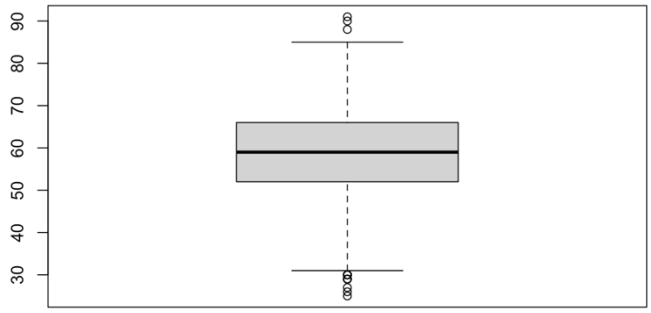
```
summary(CBPlayers$Sprint.Speed)
summary(LRBPlayers$Sprint.Speed)
```

The average Sprint Speed of CBs is 59.

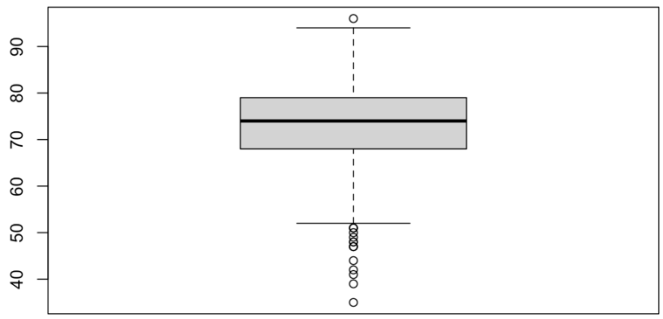
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
25.00	52.00	59.00	57.88	66.00	91.00

The average Sprint Speed of LBs and RBs is 74.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
35.00	68.00	74.00	73.49	79.00	96.00



We can see that the medium is 59, and there are outliners because there are CBs that have more than 90 of sprint speed and less than 30.



We can see that the medium is 74, and also here, there are outliners, so there are LBs and RBs that have much less sprint speed than the average.

Relation between variables

To know if two variables are related between each other, we need to calculate the Pearson's coefficient. The Pearson's coefficient indicates the correlation's degree of two quantitative variables.

- if the degree is close to 0, the relation is weak
- if the degree is close to 1, the relation is direct
- if the degree is close to -1, the relation is inverse

To know the degree, we use the function *cor()* in Rstudio.

We use the function select() of the dplyr R-package, to select only the columns that we want to analyze.

```
install.packages("dplyr")
library(dplyr)
```

We select the 6 basis stats of all the players: PACE, SHOOTING, PASSING, DRIBBLING, DEFENCE and PHYSICAL.

```
PlayersNumeric <- dplyr::select(Players, PAC, SHO, PAS, DRI, DEF, PHY)
```

We use the command cor() to calculate the Pearson's coefficient.

```
cor(PlayersNumeric)
```

```
            PAC         SHO       PAS         DRI        DEF         PHY
PAC   1.0000000  0.30539175 0.2481761  0.52164247 -0.2532565 -0.12633799
SHO   0.3053917  1.00000000 0.6585080  0.76562687 -0.4169566  0.05197273
PAS   0.2481761  0.65850803 1.0000000  0.82321646  0.1423143  0.17275095
DRI   0.5216425  0.76562687 0.8232165  1.00000000 -0.1533617  0.02454016
DEF  -0.2532565 -0.41695655 0.1423143 -0.15336169  1.0000000  0.51321798
PHY  -0.1263380  0.05197273 0.1727509  0.02454016  0.5132180  1.00000000
```

As we can see, for example the correlation between DEFENDING and SHOOTING is low (-0.416), so we can affirm the players who are defenders or with mostly defensive characteristics imply that they have few shooting stats and viceversa.

The correlation between PASSING and DRIBBLING is high (0.823) so we can affirm that players who have mostly PASSING characteristics imply that they have high characteristics about DRIBBLING and viceversa.

And the correlation between PHYSICAL and DRIBBLING is close to 0 (0.024), so we can affirm that the variables are not correlated.

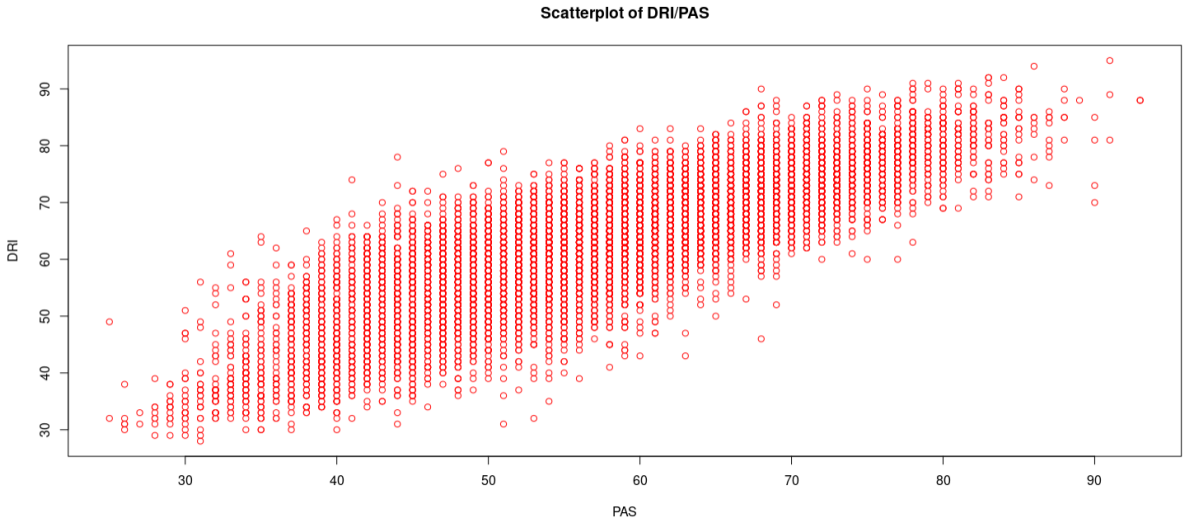
We can do a correlation test between PASSING and DRIBBLING variables.

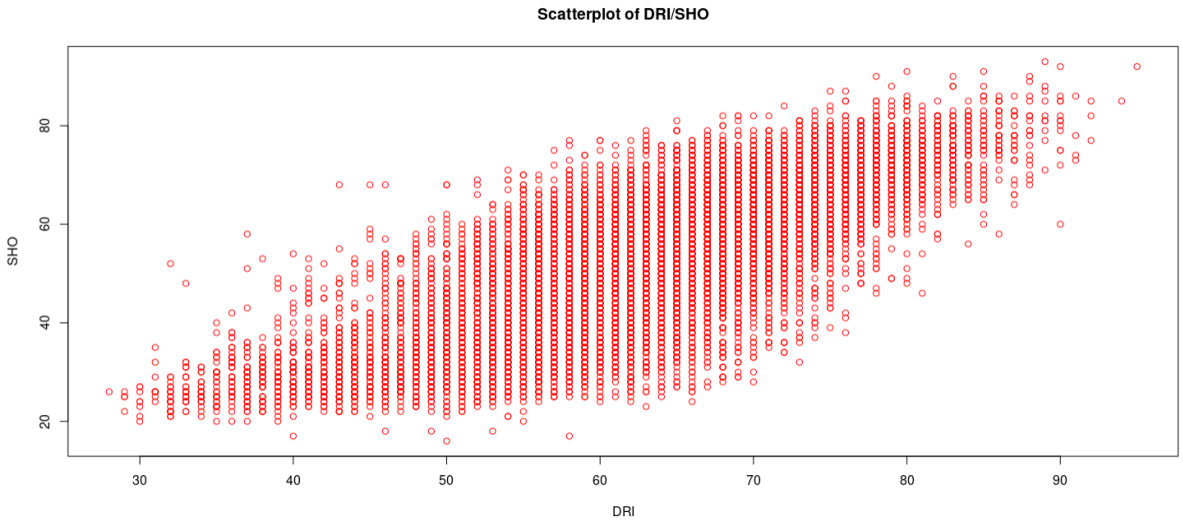
```
        Pearson's product-moment correlation

data:  Players$PAS and Players$DRI
t = 189.74, df = 17123, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8183287 0.8279851
sample estimates:
      cor
0.8232165
```

The p-value is much smaller than the significance level so we can reject the null hypothesis.

Regression model

In this section we are going to perform the linear regression model of the variables previously tested by the correlation test. To start we are going to see by a scatterplot of the variables if the variables have any relation.
The candidates that can have a relation are DRI/PAS and DRI/SHO both with a direct relation because their correlation values are close to one.





We can see in the scatterplot that both variables may have a direct relationship because whenever “x” grows, “y” also does it and vice versa. To test the veracity of this supposition we need to do the linear regression model of the variables.

```
> ModelDRIPAS <- lm(PlayersPASDRI$DRI ~ PlayersPASDRI$PAS)
> summary(ModelDRIPAS)

Call:
lm(formula = PlayersPASDRI$DRI ~ PlayersPASDRI$PAS)

Residuals:
     Min       1Q   Median       3Q      Max
-27.5006  -3.4736   0.0617   3.5533  25.6509

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       17.386294   0.250435   69.42   <2e-16 ***
PlayersPASDRI$PAS  0.794609   0.004188  189.74   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.582 on 17123 degrees of freedom
Multiple R-squared:  0.6777,    Adjusted R-squared:  0.6777
F-statistic: 3.6e+04 on 1 and 17123 DF,  p-value: < 2.2e-16
```

```
> ModelDRISHO <- lm(PlayersDRISHO$DRI ~ PlayersDRISHO$SHO)
> summary(ModelDRISHO)

Call:
lm(formula = PlayersDRISHO$DRI ~ PlayersDRISHO$SHO)

Residuals:
     Min       1Q   Median       3Q      Max
-30.6002  -3.9400   0.2627   4.1778  23.0600

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       34.391676   0.197467   174.2   <2e-16 ***
PlayersDRISHO$SHO  0.542471   0.003483   155.7   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.325 on 17123 degrees of freedom
Multiple R-squared:  0.5862,    Adjusted R-squared:  0.5862
F-statistic: 2.426e+04 on 1 and 17123 DF,  p-value: < 2.2e-16
```

In both linear models we can clearly see that there is a relation between the variables, because the coefficient t value I far away from 0 and a small p-value for the intercept and the slope indicates that we can reject the null hypothesis which allows us to conclude that there is a relationship between both variables.

Inference Statistics

Hypothesis 1

For the first hypothesis we want to test If the mean of the age of the players is 30, so we are going to suppose the next:

$H_0 = p = 30$
$H_1 = p \neq 30$

```
t.test(fifa21_male2$Age,mu=30)

        One Sample t-test

data:  fifa21_male2$Age
t = -125.15, df = 17124, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 30
95 percent confidence interval:
 25.19890 25.34697
sample estimates:
mean of x
 25.27293
```

As the p-value is smaller than the default coincidence interval 95% means that we can reject the null hypothesis, so the mean is not 30.

Hypothesis 2

In the dataset page we can see that 10% of the players have English nationality, 7% are Germans and the remaining 83% have other nationalities. So we have 17,125 players, 10% percent of that are English (1712) lets use a proportional test to see if that is true.

$H_0 = p = 10\%$
$H_1 = p \neq 10\%$

```
prop.test(x = 1712, n = 17125, p = 0.1)
```

```
        1-sample proportions test with continuity correction

data:  1712 out of 17125, null probability 0.1
X-squared = 0, df = 1, p-value = 1
alternative hypothesis: true p is not equal to 0.1
95 percent confidence interval:
 0.09553892 0.10458328
sample estimates:
        p
0.0999708
```

Here we can`t reject the null hypothesis meaning that 10% of the players are English.

Hypothesis 3

Here we want to test if the mean of the players age which are English and non-English are the same.

```
z.test(PlayerEnglish$Age,PlayerNonEnglish$Age,alt = "two.sided", sigma.x = (sd
(PlayerEnglish$Age)),sigma.y = (sd(PlayerNonEnglish$Age)))
```

```
        Two-sample z-Test

data:  PlayerEnglish$Age and PlayerNonEnglish$Age
z = -10.389, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.545246 -1.054749
sample estimates:
mean of x mean of y
 24.10252  25.40252
```

We can see that the mean of both subsets is 24,1 and 25,4.