PRA 2

Junjie Zhu i Antoni Sanchez Teruel

05/06/2021

Contents

Pràctica 2

1 Descripció del dataset. Perquè és important i quina pregunta/problema pretén respondre?

El dataset és obtingut de les estadístiques dels jugadors del videojoc Fifa 2017 (https://www.kaggle.com/artimous/complete-fifa-2017-player-dataset-global) provinents del videojocs desenvolupat per la companyia EA.

Aquest dataset recull tota la informació dels jugadors on part d'aquesta és fictícia degut que el quantifica habilitats del mateix jugadors en valors numèrics, però altres components del dataset són basats en dades reals dels jugadors.

L'objectiu que s'ha triat aquest dataset és poder establir models per mesurar com es relacionen les dades amb la realitat i com es gestionen aquestes diferències en el cas de que existeixin.

Concretament:

- Tenint en compte les dades, construir la probabilitat de que el jugador formi part de la selecció nacional.
- La coherència del ràting basant-nos en les valoracions objectives de les característiques individuals del jugador.

---

2 Integració i selecció de les dades d'interès a analitzar.

2.1 Lectura del fitxer i preparació de les dades

Comencem fent una lectura de les dades i seleccionem les dades que ens interessa pel el nostre anàlisi inicial i pels models.

```
if (!require('ggplot2')) install.packages('ggplot2'); library('ggplot2')
if (!require('dplyr')) install.packages('dplyr'); library('dplyr')
if (!require('GGally')) install.packages('GGally'); library('GGally')
```

```r
if (!require('corrplot')) install.packages('corrplot'); library('corrplot')
if (!require('car')) install.packages('car'); library('car')
if (!require('caret')) install.packages('caret'); library('caret')

fifa <- read.csv("Fifa.csv", header = TRUE, sep = ",")
head(fifa)
```

```
##                Name Nationality National_Position National_Kit           Club
## 1 Cristiano Ronaldo    Portugal                LS           7    Real Madrid
## 2      Lionel Messi   Argentina                RW          10   FC Barcelona
## 3             Neymar      Brazil                LW          10   FC Barcelona
## 4      Luis SuÃ¡rez     Uruguay                LS           9   FC Barcelona
## 5      Manuel Neuer     Germany                GK           1      FC Bayern
## 6             De Gea       Spain                GK           1 Manchester Utd
##   Club_Position Club_Kit Club_Joining Contract_Expiry Rating Height Weight
## 1            LW        7   07/01/2009            2021     94 185 cm  80 kg
## 2            RW       10   07/01/2004            2018     93 170 cm  72 kg
## 3            LW       11   07/01/2013            2021     92 174 cm  68 kg
## 4            ST        9   07/11/2014            2021     92 182 cm  85 kg
## 5            GK        1   07/01/2011            2021     92 193 cm  92 kg
## 6            GK        1   07/01/2011            2019     90 193 cm  82 kg
##   Preffered_Foot Birth_Date Age Preffered_Position       Work_Rate Weak_foot
## 1          Right 02/05/1985  32              LW/ST      High / Low         4
## 2           Left 06/24/1987  29                 RW Medium / Medium         4
## 3          Right 02/05/1992  25                 LW   High / Medium         5
## 4          Right 01/24/1987  30                 ST   High / Medium         4
## 5          Right 03/27/1986  31                 GK Medium / Medium         4
## 6          Right 11/07/1990  26                 GK Medium / Medium         3
##   Skill_Moves Ball_Control Dribbling Marking Sliding_Tackle Standing_Tackle
## 1           5           93        92      22             23              31
## 2           4           95        97      13             26              28
## 3           5           95        96      21             33              24
## 4           4           91        86      30             38              45
## 5           1           48        30      10             11              10
## 6           1           31        13      13             13              21
##   Aggression Reactions Attacking_Position Interceptions Vision Composure
## 1         63        96                 94            29     85        86
## 2         48        95                 93            22     90        94
## 3         56        88                 90            36     80        80
## 4         78        93                 92            41     84        83
## 5         29        85                 12            30     70        70
## 6         38        88                 12            30     68        60
##   Crossing Short_Pass Long_Pass Acceleration Speed Stamina Strength Balance
## 1       84         83        77           91    92      92       80      63
## 2       77         88        87           92    87      74       59      95
## 3       75         81        75           93    90      79       49      82
## 4       77         83        64           88    77      89       76      60
## 5       15         55        59           58    61      44       83      35
## 6       17         31        32           56    56      25       64      43
##   Agility Jumping Heading Shot_Power Finishing Long_Shots Curve
## 1      90      95      85         92        93         90    81
## 2      90      68      71         85        95         88    89
## 3      96      61      62         78        89         77    79
## 4      86      69      77         87        94         86    86
```

```
## 5      52      78      25         25        13         16    14
## 6      57      67      21         31        13         12    21
##   Freekick_Accuracy Penalties Volleys GK_Positioning GK_Diving GK_Kicking
## 1                76        85      88             14         7         15
## 2                90        74      85             14         6         15
## 3                84        81      83             15         9         15
## 4                84        85      88             33        27         31
## 5                11        47      11             91        89         95
## 6                19        40      13             86        88         87
##   GK_Handling GK_Reflexes
## 1          11          11
## 2          11           8
## 3           9          11
## 4          25          37
## 5          90          89
## 6          85          90
```

Les principals variables que farem servir en aquesta activitat són:

Variable	Descripció
Name	Nom del jugador
Nationality	Nacionalitat del jugador
National_Position	Posició de joc a l'equip nacional
National_Kit	Número d'equipació a l'equip nacional
Club	Nom del club
Club_Position	Posició de joc al club
Club_Kit	Número d'equipació al club
Club_Joining	Data que va començar al club
Contract_Expire	Any finalització del contracte
Rating	Valoració global del jugador, entre 0 i 100
Height	Altura
Weight	Pes
Preffered_Foot	Peu preferit
Age	Edat
Preffered_Position	Posició preferida
Work_Rate	valoració qualitativa en termes d'atac-defensa
Weak_foot	valoració d'1 a 5 de control i potència de la cama no preferida
Skill_Moves	valoració d'1 a 5 de l'habilitat en moviments del jugador
La resta de variables fan referència a atributs del jugador.	

Com s'ha explicat anteriorment, en el primer punt, la raó que basat per seleccionar aquestes variables del dataset és en base la informació per classificar els tipus de jugadors basant-nos en el seu físic sense entrar en detall en estadístiques i valors numerics de les seves habilitats com es mostra en el joc. D'aquesta manera es podria extrapolar aquests resulats per possibles avaluacions de jugadors nous i fins i tot aplicar-ho a la realitat en comptes del joc.

Veiem quin tipus de variables són i fem el primer processat si cal.

Variables tipus numèric:

Nagional_Kit, Club_Kit, Contract_Expiry (any), Rating, Height, Weight, Age, Weak_foot, Skill_Moves y la resta de estadístiques del jugador.

Variables tipus categòric:

National_Position, Preffered_Foot, Preffered_Position y Work_Rate. Cal tenir en compte que es podria considerar com a variable categòrica les variables Club i Nationality, segons si jugadors del mateix club apareixen repetidament

2.2 Preparació de les dades

```
# Factgor de variables categòriques
fifa$Nationality <- factor(fifa$Nationality)
fifa$National_Position <- factor(fifa$National_Position)
fifa$Club <- factor(fifa$Club)
fifa$Club_Position <- factor(fifa$Club_Position)
fifa$Preffered_Foot <- factor(fifa$Preffered_Foot)
fifa$Preffered_Position <- factor(fifa$Preffered_Position)
fifa$Work_Rate <- factor(fifa$Work_Rate)


# Height
fifa$Height <- gsub(" [[:alpha:]]*", "", fifa$Height)
fifa$Height <- as.numeric(fifa$Height)

# Weight
fifa$Weight <- gsub(" [[:alpha:]]*", "", fifa$Weight)
fifa$Weight <- as.numeric(fifa$Weight)

head(fifa)
```

```
##                Name Nationality National_Position National_Kit           Club
## 1 Cristiano Ronaldo    Portugal                LS           7    Real Madrid
## 2      Lionel Messi   Argentina                RW          10   FC Barcelona
## 3             Neymar      Brazil                LW          10   FC Barcelona
## 4      Luis SuÃ¡rez     Uruguay                LS           9   FC Barcelona
## 5      Manuel Neuer     Germany                GK           1      FC Bayern
## 6             De Gea       Spain                GK           1 Manchester Utd
##   Club_Position Club_Kit Club_Joining Contract_Expiry Rating Height Weight
## 1            LW        7   07/01/2009            2021     94    185     80
## 2            RW       10   07/01/2004            2018     93    170     72
## 3            LW       11   07/01/2013            2021     92    174     68
## 4            ST        9   07/11/2014            2021     92    182     85
## 5            GK        1   07/01/2011            2021     92    193     92
## 6            GK        1   07/01/2011            2019     90    193     82
##   Preffered_Foot Birth_Date Age Preffered_Position       Work_Rate Weak_foot
## 1          Right 02/05/1985  32              LW/ST      High / Low         4
## 2           Left 06/24/1987  29                 RW Medium / Medium         4
## 3          Right 02/05/1992  25                 LW   High / Medium         5
## 4          Right 01/24/1987  30                 ST   High / Medium         4
## 5          Right 03/27/1986  31                 GK Medium / Medium         4
## 6          Right 11/07/1990  26                 GK Medium / Medium         3
##   Skill_Moves Ball_Control Dribbling Marking Sliding_Tackle Standing_Tackle
## 1           5           93        92      22             23              31
## 2           4           95        97      13             26              28
## 3           5           95        96      21             33              24
## 4           4           91        86      30             38              45
## 5           1           48        30      10             11              10
## 6           1           31        13      13             13              21
```

```
##   Aggression Reactions Attacking_Position Interceptions Vision Composure
## 1         63        96                 94            29     85        86
## 2         48        95                 93            22     90        94
## 3         56        88                 90            36     80        80
## 4         78        93                 92            41     84        83
## 5         29        85                 12            30     70        70
## 6         38        88                 12            30     68        60
##   Crossing Short_Pass Long_Pass Acceleration Speed Stamina Strength Balance
## 1       84         83        77           91    92      92       80      63
## 2       77         88        87           92    87      74       59      95
## 3       75         81        75           93    90      79       49      82
## 4       77         83        64           88    77      89       76      60
## 5       15         55        59           58    61      44       83      35
## 6       17         31        32           56    56      25       64      43
##   Agility Jumping Heading Shot_Power Finishing Long_Shots Curve
## 1      90      95      85         92        93         90    81
## 2      90      68      71         85        95         88    89
## 3      96      61      62         78        89         77    79
## 4      86      69      77         87        94         86    86
## 5      52      78      25         25        13         16    14
## 6      57      67      21         31        13         12    21
##   Freekick_Accuracy Penalties Volleys GK_Positioning GK_Diving GK_Kicking
## 1                76        85      88             14         7         15
## 2                90        74      85             14         6         15
## 3                84        81      83             15         9         15
## 4                84        85      88             33        27         31
## 5                11        47      11             91        89         95
## 6                19        40      13             86        88         87
##   GK_Handling GK_Reflexes
## 1          11          11
## 2          11           8
## 3           9          11
## 4          25          37
## 5          90          89
## 6          85          90
```

Observem que teniem algunes variables (Height i Weight) amb caràcters en comptes de ser únicament numèrics. Pel que s'ha procedit en eliminar aquests caràcters mantenint únicament els valor numèric del camp.

Com per exemple, Height on tenim variable tipus caràcter "183 cm" passa a ser numèric 183.

```
fifa2 <- fifa[,1:20]
```

Finalment tenim una taula amb les 20 variables que ens poden ser d'interés o im totañ de 17588 observacions.

---

3 Neteja de dades

3.1 Valors absents

Seguidament repasarem els valors absents del nostre dataset, cal comentar que en la National_Kit i National_Position **poden** contenir valors buits degut que fan referència a la posició en la selacció nacional del jugador, degut que no tots els jugadors poden ser seleccionats trobarem bastantes observacions buides en aquestes columnes

```
colSums(is.na(fifa2))
```

```
##                Name         Nationality  National_Position        National_Kit
##                   0                   0                   0               16513
##                Club       Club_Position           Club_Kit       Club_Joining
##                   0                   0                   1                   0
##     Contract_Expiry              Rating              Height              Weight
##                   1                   0                   0                   0
##      Preffered_Foot          Birth_Date                 Age Preffered_Position
##                   0                   0                   0                   0
##           Work_Rate           Weak_foot         Skill_Moves        Ball_Control
##                   0                   0                   0                   0
```

```
colSums(fifa2 == "")
```

```
##                Name         Nationality  National_Position        National_Kit
##                   0                   0               16513                  NA
##                Club       Club_Position           Club_Kit       Club_Joining
##                   0                   1                  NA                   1
##     Contract_Expiry              Rating              Height              Weight
##                  NA                   0                   0                   0
##      Preffered_Foot          Birth_Date                 Age Preffered_Position
##                   0                   0                   0                   0
##           Work_Rate           Weak_foot         Skill_Moves        Ball_Control
##                   0                   0                   0                   0
```

```
colSums(fifa2 <= 0)
```

```
##                Name         Nationality  National_Position        National_Kit
##                   0                  NA                  NA                  NA
##                Club       Club_Position           Club_Kit       Club_Joining
##                  NA                  NA                  NA                   1
##     Contract_Expiry              Rating              Height              Weight
##                  NA                   0                   0                   0
##      Preffered_Foot          Birth_Date                 Age Preffered_Position
##                  NA                   0                   0                  NA
##           Work_Rate           Weak_foot         Skill_Moves        Ball_Control
##                  NA                   0                   0                   0
```

Observem que hi ha 1 NA a Club_Kit i 1 NA en Contract_Expiry

```
which(is.na(fifa2$Club_Kit), arr.ind=TRUE)
```

```
## [1] 384
```

```
which(is.na(fifa2$Contract_Expiry), arr.ind=TRUE)
```

```
## [1] 384
```

```
which(fifa2$Club_Position == "", arr.ind=TRUE)
```

```
## [1] 384
```

```
# Eliminem les files amb NA excepte les que tinguin NA únicament a National_Kit
# i National_Position
fifaNet <- fifa2[complete.cases(fifa2[ , -(3:4)]),]
```

Com hem vist sembla que només hi ha un únic jugador que li falta valors fora de les columnes excloses, per tant com observem només hi ha 1 fila de diferència. Podem considerar que el canvi realitzat no afecta de

manera considerable a les dades que tenim.

```
colSums(is.na(fifaNet))
```

```
##               Name        Nationality  National_Position       National_Kit
##                  0                  0                  0              16512
##               Club      Club_Position           Club_Kit       Club_Joining
##                  0                  0                  0                  0
##    Contract_Expiry             Rating             Height             Weight
##                  0                  0                  0                  0
##     Preffered_Foot         Birth_Date                Age Preffered_Position
##                  0                  0                  0                  0
##          Work_Rate          Weak_foot        Skill_Moves       Ball_Control
##                  0                  0                  0                  0
```

```
colSums(fifaNet == "")
```

```
##               Name        Nationality  National_Position       National_Kit
##                  0                  0              16512                 NA
##               Club      Club_Position           Club_Kit       Club_Joining
##                  0                  0                  0                  0
##    Contract_Expiry             Rating             Height             Weight
##                  0                  0                  0                  0
##     Preffered_Foot         Birth_Date                Age Preffered_Position
##                  0                  0                  0                  0
##          Work_Rate          Weak_foot        Skill_Moves       Ball_Control
##                  0                  0                  0                  0
```

```
colSums(fifaNet <= 0)
```

```
##               Name        Nationality  National_Position       National_Kit
##                  0                 NA                 NA                 NA
##               Club      Club_Position           Club_Kit       Club_Joining
##                 NA                 NA                  0                  0
##    Contract_Expiry             Rating             Height             Weight
##                  0                  0                  0                  0
##     Preffered_Foot         Birth_Date                Age Preffered_Position
##                 NA                  0                  0                 NA
##          Work_Rate          Weak_foot        Skill_Moves       Ball_Control
##                 NA                  0                  0                  0
```

El resultat que obtenim és una taula amb 178587 observacions on no disposem de valors absents, buiuts o zeros. A excpeció del número a la posició nacional.

3.2 Transformacions

En aquest apartat procedirem a discretitzar alguns valors com el rating, de manera que segons el rating que tingui el jugador es considera:

Tipus	Rang rating
Excel · lent	90-100
Molt Bo	80-89
Bo	70-79
Regular	50-69
Dolent	40-49
Molt Dolent	0-39

```
fifaNet$clasificacion = cut(fifaNet$Rating, c(0,40,50,70,80,90,101),
                            labels = c("Molt Dolent", "Dolent", "Regular",
                                       "Bo", "Molt Bo", "Excel·lent"),
                            include.lowest = TRUE, right = FALSE)
summary(fifaNet$clasificacion)
```

```
## Molt Dolent      Dolent     Regular          Bo     Molt Bo  Excel·lent
##           0         121       11921        5017         519           9
```

```
fifaNet$clasificacion2 = cut(fifaNet$Rating, c(0,70,101),
                             labels = c("Regular", "Bo"), include.lowest = TRUE,
                             right = FALSE)
summary(fifaNet$clasificacion2)
```

```
## Regular      Bo
##   12042    5545
```

Adicionalment s'ha separat dels jugadors regulars dels bons, on els regulars tenen un ràting entre 0-69 i els bons de 70-100.

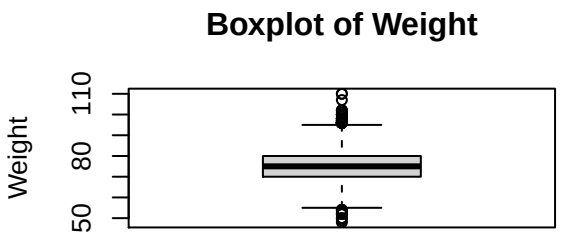
Diferenciem la posició de porter de la resta dels jugadors, ja que aquest normalment són una posició amb característiques molt diferenciats d'un jugador de camp

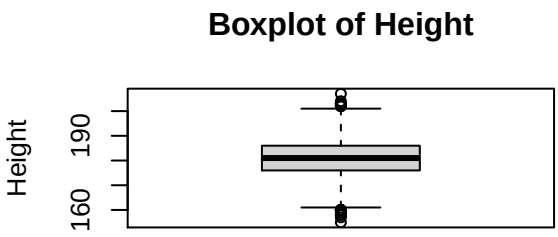
```
# Iterem per les difrents posicions i afegim TRUE o FALSE segons si és
# porter o no, en una nova variable portero
for (i in 1:length(fifaNet$Club_Position)){
  if (fifaNet$Club_Position[i] == "GK"){
    fifaNet$porter[i] = "Porter"
  } else{
    fifaNet$porter[i] = "Camp"
  }
}
fifaNet$porter <- factor(fifaNet$porter)
```

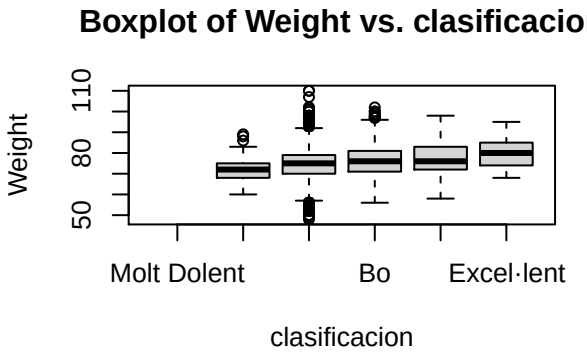
3.3 Valors extrems (outliers)

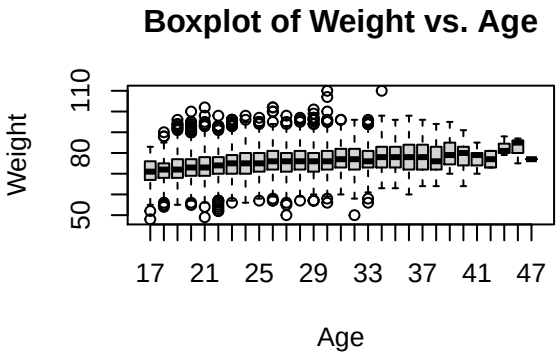
Anem analitzar si tenim valors extrems en el nostre mitjançant les gràfiques boxplot.

```
# Plot boxplot
par(mfrow=c(2,2))
boxplot(fifaNet$Weight, main = "Boxplot of Weight", ylab = "Weight")
boxplot(fifaNet$Height, main = "Boxplot of Height", ylab = "Height")
boxplot(Weight ~ clasificacion , data = fifaNet,
        main = "Boxplot of Weight vs. clasificacion", ylab = "Weight")
boxplot(Weight ~ Age , data = fifaNet,
        main = "Boxplot of Weight vs. Age", ylab = "Weight")
```









En el gràfic de Weight, observem com alguns jugadors pesen més en comparació als altres, i passa el mateix amb el cas de l'alçada dels jugadors.

Weight vs Clasificacion, podem veure com els jugadors considerats pitjors tenen menor pes en comparació als jugadors més ben valorats.

Finalment en el gràfic Wight vs Age, podem observar que tendim a tenir més "outliers" en jugadors amb menys edat.

Però són realment valors extrems? Poden haver jugadors amb més pes que la mitjana?

Anem a revisar-ho:

```
fifaNet %>% arrange_(~ desc(Weight)) %>% head(n = 10)
```

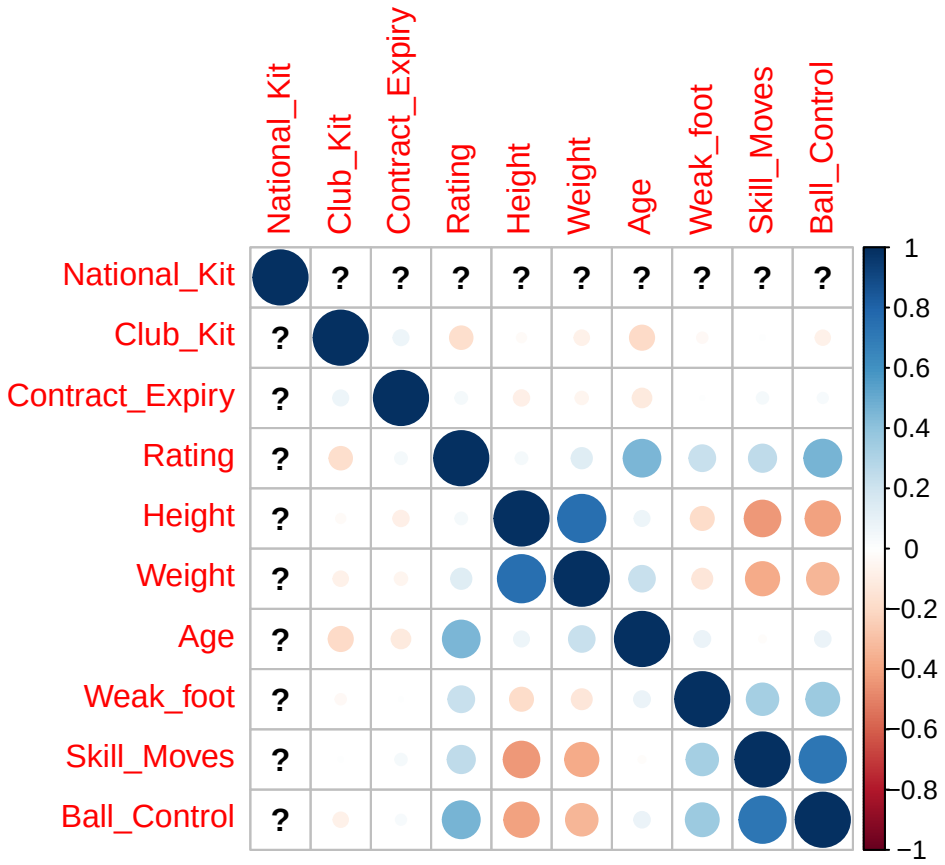
```
##                 Name         Nationality National_Position National_Kit
## 1   Kristof Van Hout             Belgium                              NA
## 2  Adebayo Akinfenwa             England                              NA
## 3        Chris Seitz       United States                              NA
## 4         Bill Hamid       United States                              NA
## 5       Hakeem Araba             England                              NA
## 6         Evan Louro       United States                              NA
## 7     Rene Gilmartin Republic of Ireland                              NA
## 8    Lars Unnerstall             Germany                              NA
## 9         Carl Ikeme             Nigeria                              NA
## 10   Martin PolaÄ\215ek           Slovakia                                 NA
##                Club Club_Position Club_Kit Club_Joining Contract_Expiry Rating
## 1      KVC Westerlo           Sub        1   07/01/2015            2017     67
## 2           Wycombe            ST       20   07/10/2016            2017     64
```

```
## 3         FC Dallas            GK       18   01/01/2010            2020     68
## 4       D.C. United            GK       28   01/01/2009            2021     75
## 5    Falkenbergs FF            LS       18   02/27/2015            2021     62
## 6      NY Red Bulls           Res       45   01/23/2017            2020     53
## 7           Watford           Res       13   08/25/2014            2017     59
## 8    F. DÃ¼sseldorf           Sub       19   07/01/2014            2017     72
## 9             Wolves            GK        1   07/01/2003            2019     70
## 10 ZagÅ‚Ä\231bie Lubin            GK        1   07/10/2015            2018     66
##    Height Weight Preffered_Foot Birth_Date Age Preffered_Position
## 1     207    110          Right 02/09/1987  30                GK
## 2     178    110          Right 05/10/1982  34                ST
## 3     191    107          Right 03/12/1987  30                GK
## 4     191    102          Right 11/25/1990  26                GK
## 5     191    102          Right 02/12/1991  26                ST
## 6     191    102          Right 01/19/1996  21                GK
## 7     197    101          Right 05/31/1987  29                GK
## 8     198    100          Right 07/20/1990  26                GK
## 9     191    100          Right 06/08/1986  30                GK
## 10    199    100          Right 04/02/1990  26                GK
##          Work_Rate Weak_foot Skill_Moves Ball_Control clasificacion
## 1  Medium / Medium         3           1           24       Regular
## 2        Low / Low         3           2           69       Regular
## 3  Medium / Medium         3           1            9       Regular
## 4  Medium / Medium         2           1           20            Bo
## 5  Medium / Medium         2           3           58       Regular
## 6  Medium / Medium         2           1           17       Regular
## 7  Medium / Medium         3           1           22       Regular
## 8  Medium / Medium         1           1           20            Bo
## 9  Medium / Medium         3           1           24            Bo
## 10 Medium / Medium         2           1           19       Regular
##    clasificacion2 porter
## 1         Regular   Camp
## 2         Regular   Camp
## 3         Regular Porter
## 4              Bo Porter
## 5         Regular   Camp
## 6         Regular   Camp
## 7         Regular   Camp
## 8              Bo   Camp
## 9              Bo Porter
## 10        Regular Porter
```

Observem que la majoria de jugadors amb més pes són els dels jugadors més alts, això té sentit a nivell físic ja que l'alçada d'una persona normalment afecta al pes del mateix.

Realitzem un mapa de correlació per confirma la idea plantejada:

```
cor1 <- cor(fifaNet[sapply(fifaNet,is.numeric)])

corrplot(cor1, method="circle")
```



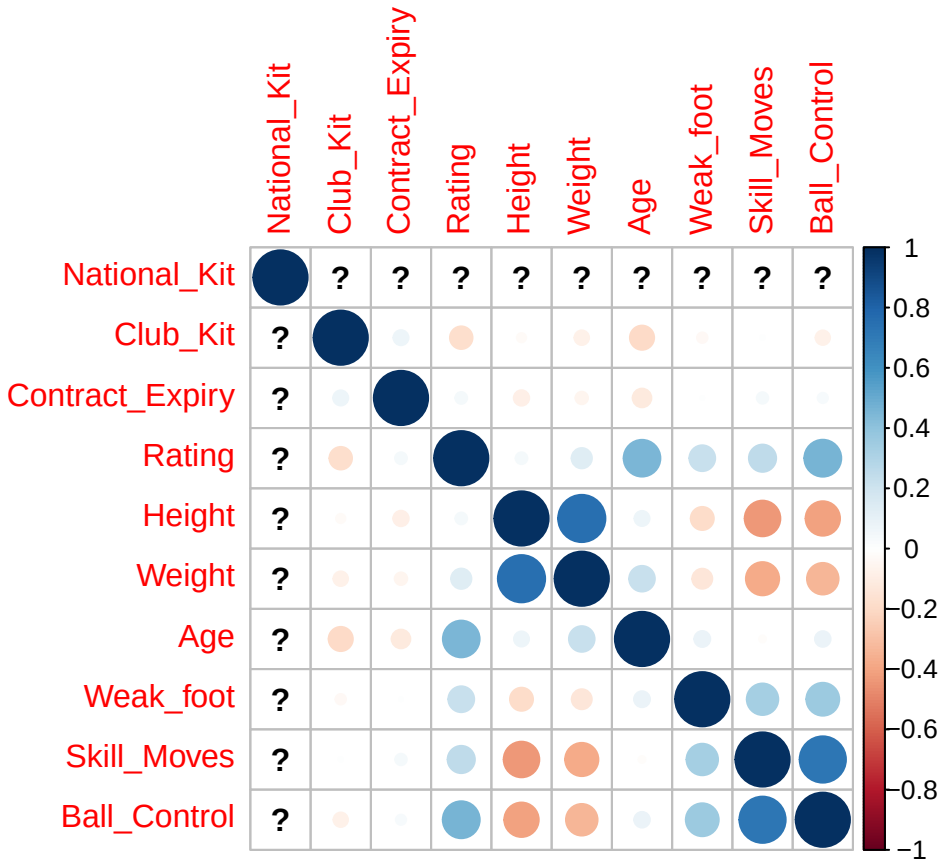
Com observem tenim que Weight i Height tenen una correlació positiva i sembla ser propera a 0.8 això vol dir que com més alçada més pes.

---

4 Anàlisi de les dades

Partint del mapa de correlacions trobat anteriorment:

```
corrplot(cor1, method="circle")
```



Observem que a part de Weight-Height tenim més correlacions interesants entre variables.

Per exemple Age i Rating, amb una correlació positiva la qual ens indica que com més gran es faci un jugador tendeix a tenir un major rating.

Per altra banda, sembla ser que tenim correlacions negatives Skill_Moves-Height o Ball_Control-Height que ens està indicant que jugadors amb més alçada tendeixen a tenir menys habilitat amb la pilota.

4.1 Selecció de dades que es volen utilitzar

Partint de les dades seleccionades al punt 2 i el mapa de correlacions les variables que ens interessen més són les següents:

Variable	Descripció
National_Kit	Número d'equipació a l'equip nacional
Club_Position	Posició de joc al club
Rating	Valoració global del jugador, entre 0 i 100
Height	Altura
Weight	Pes
Preffered_Foot	Peu preferit
Age	Edat
Preffered_Position	Posició preferida
Work_Rate	valoració qualitativa en termes d'atac-defensa
Weak_foot	valoració d'1 a 5 de control i potència de la cama no preferida
Skill_Moves	valoració d'1 a 5 de l'habilitat en moviments del jugador
Ball_Control	valoració d'1 a 5 del control de la bola

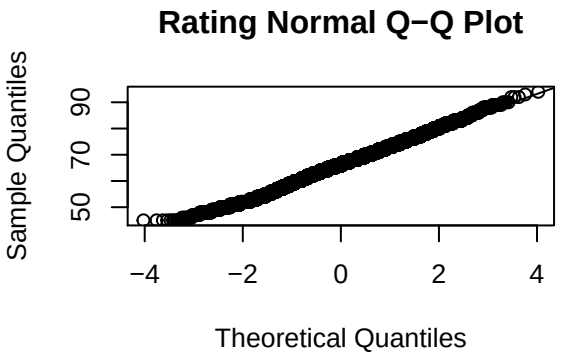
La raó per seleccionar aquestes dades és degut que són informació física del jugadors i són els que donen més informació de la seva habilitat. Addicionalment s'ha seleccionat National_kit ja que ens serà útil per saber si el jugador acaba a la selecció nacional o no. I club_position ens permet diferenciar les diferentes posicions dels jugadors, en el nostra cas el que tindrem en compte és el porter ja que és l'únic que té una gran variació de requeriments i posicó respecte els jugadors de camp (les altres posicions).

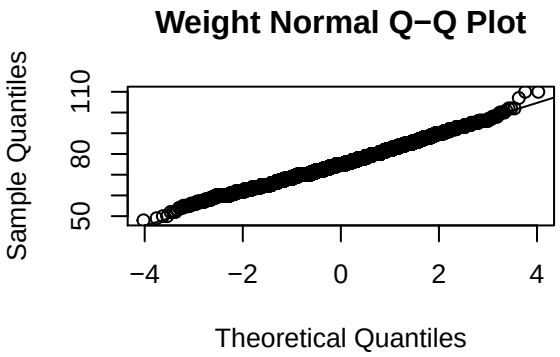
4.2 Comprovació de normalitat i homogeneïtat de la variància

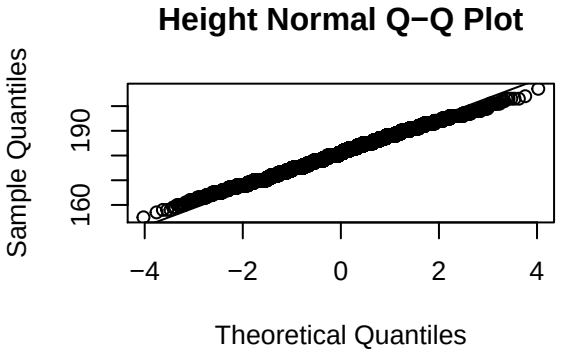
En el nostre dataset els únics valors que podrien estudiar la seva normalitat és el rating, el pes, l'alçada i l'edat ja que són principalment les úniques variables no categoriques del nostre dataset.

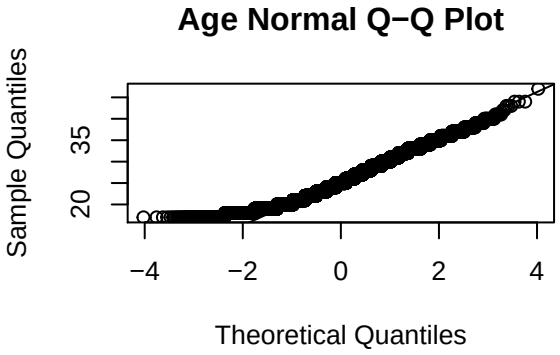
Com que disposem d'una gran quantitat d'observacions mirarem de visualitzar la seva normalitat a través del QQ-plot, generat amb la funció qqnorm(), que presenta els valors de les variables en un eix X, aquesta representació dels valors s'haurien d'aliniar amb la línia diagonal representada amb la funció qqline(). Això ens indicara si aquestes variables tenen una distribució aproximadament normal o no.

```
par(mfrow=c(2,2))
qqnorm(fifaNet$Rating, main='Rating Normal Q-Q Plot')
qqline(fifaNet$Rating)

qqnorm(fifaNet$Weight, main='Weight Normal Q-Q Plot')
qqline(fifaNet$Weight)

qqnorm(fifaNet$Height, main='Height Normal Q-Q Plot')
qqline(fifaNet$Height)

qqnorm(fifaNet$Age, main='Age Normal Q-Q Plot')
qqline(fifaNet$Age)
```









Com observem sembla que totes les 4 varibels tenen una distribució que s'aproxima a la normalitat, tot i així s'ha de comentar que en les edats més baixes no segueixen gaire aquesta normalitat.

Per assegurar-nos realitzarem el test Kolmogorov-Smirnov amb aquestes variables, ja que Shapiro-Wilk funciona quan hi ha un total d'observacions entre 3 i 5000:

Considerem que la nostra hipòtesi nul · la és que la població està distribuida normalment.

```
# Rating
ks.test(fifaNet$Rating, pnorm, mean(fifaNet$Rating), sd(fifaNet$Rating))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  fifaNet$Rating
## D = 0.043971, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
# Weight
ks.test(fifaNet$Weight, pnorm, mean(fifaNet$Weight), sd(fifaNet$Weight))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  fifaNet$Weight
## D = 0.05617, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```r
# Height
ks.test(fifaNet$Height, pnorm, mean(fifaNet$Height), sd(fifaNet$Height))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  fifaNet$Height
## D = 0.048542, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```r
# Age
ks.test(fifaNet$Age, pnorm, mean(fifaNet$Age), sd(fifaNet$Age))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  fifaNet$Age
## D = 0.083494, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

En els resultats de Kolmogorov-Smirnov ens dona que totes les p-values són menors de 0,05, volent dir que s'ha de rebutjar la hipòtesi nul · la degut que p-values és menor que el nivell de significació.

Però com s'ha mostrat en els gràfics qqnorm però tenint en consideració el teorema del límit central podem assumir que hi ha normallitat en la distribució quan existeix un gran nombre de observacions com és el nostre cas amb 17587 observacions.

El teorema ens indica que la mitjana d'una mostra de qualsevol conjunt de dades és cada vegada més normal a mesura que augmentem la quantitat d'observacions.

Seguidament considerem si hi ha homogeneïtat de variància respecte als jugadors que són porters, ja que aquests són els que més es diferèncien entre els jugadors. Apliquem el test de homoscedasticitat amb Levene per dades amb distribució normal:

```r
# Rating
leveneTest(Rating ~ porter, data = fifaNet)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value Pr(>F)
## group     1  3.8863 0.0487 *
##       17585
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Weight
leveneTest(Weight ~ porter, data = fifaNet)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value    Pr(>F)
## group     1   14.97 0.0001096 ***
##       17585
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Height
leveneTest(Height ~ porter, data = fifaNet)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value    Pr(>F)
```

```
## group     1  125.39 < 2.2e-16 ***
##       17585
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Age
leveneTest(Age ~ porter, data = fifaNet)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value Pr(>F)
## group     1  1.3161 0.2513
##       17585
```

Com observem pràcticament tots tenen una p-value menor del nivell de significació (0.05) el que ens esta dient és que es rebutja la hipòtesi nul · la, volent dir que es presenta variàncies estadísticament diferents pels grups de jugadors que són porters.

Per altra banda, veiem que en el cas de l'edat el p-values és major que el nivell de significació, això vol dir que els porters no presenten variànces diferents respecte als jugadors de camp.

Comprovem la homoscedasticitat de les variables en base al tipus de classficació (clasificacion2) que hem creat on separem els jugadors regulars dels bons:

```
# Rating
leveneTest(Rating ~ clasificacion2, data = fifaNet)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value    Pr(>F)
## group     1  519.77 < 2.2e-16 ***
##       17585
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Weight
leveneTest(Weight ~ clasificacion2, data = fifaNet)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value   Pr(>F)
## group     1  9.1628 0.002473 **
##       17585
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Height
leveneTest(Height ~ clasificacion2, data = fifaNet)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value  Pr(>F)
## group     1  5.3889 0.02028 *
##       17585
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Age
leveneTest(Age ~ clasificacion2, data = fifaNet)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value    Pr(>F)
## group     1  129.28 < 2.2e-16 ***
```

```
##       17585
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Com era d'esperar es rebutja la hipòtesi nul · la d'homogeneïtat de variàncies per les variables de Rating, però és interesant que observar que també passa amb l'edat.

Per part del pes i de l'alçada sembla que si presenten homogeneïtat.

4.3 Anàlisis de les dades

4.3.1 Contrast d'hipòtesi Sovint es comenta que els jugadors més joves tenen millor físic en comparació als veterans, pel que es consideren que jugadors joves són millors jugant però és realment així Partint de porter i classificacion2, ens plantegem les preguntes següents:

- Els jugadors són millors quan són més joves?
- Un porter pesa més que un jugador de camp perqué no han de correr pel camp?

  Realitzem un contrast d'hipòtesis de les preguntes plantejades:

  Els jugadors són millors quan són més joves?

- Hipòtesi nul · la (H0): Els jugadors bons **NO** són els més joves
- Hipòtesi alternativa (H1): Els jugadors bons són els més joves

  Comencem a comparar les variances de les dues mostres:

```
df_1<-fifaNet$Age[fifaNet$clasificacion2=="Bo"]
df_2<-fifaNet$Age[fifaNet$clasificacion2=="Regular"]
var.test(df_1,df_2)
```

```
##
##  F test to compare two variances
##
## data:  df_1 and df_2
## F = 0.75159, num df = 5544, denom df = 12041, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7186574 0.7863226
## sample estimates:
## ratio of variances
##          0.7515866
```

El test var.test de R ens mostra un p valor menor de 0.05 pel que no es pot assumir igualtat de variàncies en les dues poblacions. Per tant, apliquem un test sobre la mitjana de dues mostres independents amb variància desconeguda i diferent. És un test unilateral per la dreta.

Procedim a realitzar el contrast d'hipòtesis:

```
t.test(fifaNet$Age~relevel(fifaNet$clasificacion2,ref="Bo"),m=5,var.equal=F,
       alt="greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  fifaNet$Age by relevel(fifaNet$clasificacion2, ref = "Bo")
## t = -28.753, df = 12294, p-value = 1
## alternative hypothesis: true difference in means is greater than 5
## 95 percent confidence interval:
```

```
##  2.910547      Inf
## sample estimates:
##      mean in group Bo mean in group Regular
##              27.52985              24.50623
```

Per un nivell de confiança del 95%, tenim un p-value = 1 major que el nivell de significació, per tant, **no podem rebutjar la hipòtesi nul · la ja que com mostra els resultats, la mitjana d'edat dels jugadors bons és de 27,53 anys i la mitjana d'edat dels jugadors regulars és de 24,51 anys.**

Un porter pesa més que un jugador de camp perqué no han de correr pel camp?

- Hipòtesi nul · la (H0): Els porters **NO** pesen més que els jugadors de camp
- Hipòtesi alternativa (H1): Els porters pesen més que els jugadors de camp

```
df_1<-fifaNet$Weight[fifaNet$porter=="Porter"]
df_2<-fifaNet$Weight[fifaNet$porter=="Camp"]
var.test(df_1,df_2)
```

```
##
##  F test to compare two variances
##
## data:  df_1 and df_2
## F = 0.78916, num df = 631, denom df = 16954, p-value = 7.335e-05
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7073142 0.8857220
## sample estimates:
## ratio of variances
##          0.7891574
```

El test var.test de R ens mostra un p valor menor de 0.05 pel que no es pot assumir igualtat de variàncies en les dues poblacions. Per tant, apliquem un test sobre la mitjana de dues mostres independents amb variància desconeguda i diferent.

Procedim a realitzar el contrast d'hipòtesis:

```
t.test(fifaNet$Weight~relevel(fifaNet$porter,ref="Porter"),m=5,var.equal=F,
       alt="greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  fifaNet$Weight by relevel(fifaNet$porter, ref = "Porter")
## t = 12.446, df = 691.96, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 5
## 95 percent confidence interval:
##  7.639654      Inf
## sample estimates:
## mean in group Porter   mean in group Camp
##             83.00633             74.96408
```

Dels resultats obtinguts tenim que un p-value menor que el nivell de significació (0.05) pel que ens diu que podem rebutjar la hipòtesi nul · la a favor de l'alternativa, per tant, podem dir que **es confirma l'hipòtesi que els porters pesen més que els jugadors de camps, on els porters tenen de mitja 83kg i els jugadors de camp 74kg.** Aquest resultat té sentit ja que el porter no és una posició que necessiti moures constantment pel camp sinó que necessita una bona massa muscular per parar els atacs tenenint les seves cames com a únic punt de suport.

4.3.2 Regressió lineal

4.3.2.1 Model Si el rating es consequencia dels factors de qualitat tindriem com a resultat un model que podria explicar gairabé tot els jugadors de la base de dades. Si per el contrari no es així, tindríem que es fan servir valors ocults i externs a les valoracions numèriques per establir el rating.

```
model <- lm(Rating ~ Skill_Moves + Ball_Control + Age + Height + Weight +
              clasificacion2, data = fifaNet)
summary(model)
```

```
##
## Call:
## lm(formula = Rating ~ Skill_Moves + Ball_Control + Age + Height +
##     Weight + clasificacion2, data = fifaNet)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.8617  -2.5772   0.1026   2.6075  19.6707
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      28.713044   1.058290  27.132  < 2e-16 ***
## Skill_Moves      -0.311565   0.058932  -5.287 1.26e-07 ***
## Ball_Control      0.139320   0.002713  51.356  < 2e-16 ***
## Age               0.354594   0.006794  52.192  < 2e-16 ***
## Height            0.063025   0.007044   8.948  < 2e-16 ***
## Weight            0.091813   0.006783  13.535  < 2e-16 ***
## clasificacion2Bo  8.696913   0.071137 122.255  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.886 on 17580 degrees of freedom
## Multiple R-squared:  0.699,  Adjusted R-squared:  0.6989
## F-statistic:  6804 on 6 and 17580 DF,  p-value: < 2.2e-16
```

El valor de R2 ajustat és 0.6989. És a dir, el model explica un 69.9% de la variància en el ràting dels jugadors. Això indica que la capacitat explicativa del model és bona per a estimar la puntuació dels jugadors. El pvalor del model és menor de 0.05 i, per tant, el conjunt de variables explicatives contribueixen significativament a explicar el ràting dels jugadors.

En relació a l'anàlisi per separat de les variables explicatives, s'observa que totes les variables del model són significatives. Les variables Ball_Control,Age,Height,Weight i clasificacion2 tenen una correlació positiva amb el ràting, indicant que la puntuació del jugador creix si augmenten els valors d'aquestes variables. Per contra, la variable Skill_Moves té una correlació negativa: quan aquesta variable pren mes valor, i la resta de variables prenen el mateix valor, la seva puntuació es redueix en uns -0.311565

4.3.2.2 Predicció. Apliquem el model de regressió per a predir un jugador amb habilitat de moviment 4, control de pilota 70, edat 24, alçada 179, pes 70 i classificació de "Bo"

```
new<-data.frame(Skill_Moves=4 , Ball_Control =70,  Age=24, Height=179,
    Weight=70, clasificacion2= 'Bo')
predict(model,new,type="response")
```

```
##        1
## 72.13482
```

El model prediu que un jugador amb aquestes característiques tindrà un ràting de 72.13. Aquesta predicció s'ha de prendre amb certa cautela ja que el R2 no es del tot elevat

En el cas de que el sistema de valoració de rating fes servir variables no declarades al dataset ens hauríem de trobar que el nostre model predigui una valoració substancialment més baixa de la real en els casos de rating més alt.

Busquem el jugador amb el rating més alt:

```
fifaNet[which.max(fifaNet$Rating),]
```

```
##                Name Nationality National_Position National_Kit        Club
## 1 Cristiano Ronaldo    Portugal                LS            7 Real Madrid
##   Club_Position Club_Kit Club_Joining Contract_Expiry Rating Height Weight
## 1            LW        7   07/01/2009            2021     94    185     80
##   Preffered_Foot Birth_Date Age Preffered_Position  Work_Rate Weak_foot
## 1          Right 02/05/1985  32              LW/ST High / Low         4
##   Skill_Moves Ball_Control clasificacion clasificacion2 porter
## 1           5           93    Excel·lent             Bo   Camp
```

Es Cristiano ROnaldo amb un Rating de 94.

Ara realitzem la predicció del seu ràting tenim en compte les seves característiques:

```
new2<-data.frame(Skill_Moves=5 , Ball_Control =93,  Age=32, Height=185,
    Weight=80, clasificacion2= 'Bo')
predict(model,new2,type="response")
```

```
##        1
## 79.16066
```

Així, trobem que el seu ràting està "sobrevalorat" amb una diferència de 15 comparat amb el model.

Busquem el jugador amb el rating més baix:

```
fifaNet[which.min(fifaNet$Rating),]
```

```
##                Name Nationality National_Position National_Kit          Club
## 17579 Steven Alzate     England                              NA Leyton Orient
##       Club_Position Club_Kit Club_Joining Contract_Expiry Rating Height Weight
## 17579           Res       31   09/21/2016            2017     45    177     65
##       Preffered_Foot Birth_Date Age Preffered_Position       Work_Rate
## 17579          Right 09/08/1998  18                CAM Medium / Medium
##       Weak_foot Skill_Moves Ball_Control clasificacion clasificacion2 porter
## 17579         3           3           46        Dolent        Regular   Camp
```

El jugador amb el ràting més baix es de 45

Ara realitzem la predicció del seu ràting tenim en compte les seves característiques:

```
new2<-data.frame(Skill_Moves=3 , Ball_Control =46,  Age=32, Height=177,
    Weight=18, clasificacion2= 'Regular')
predict(model,new2,type="response")
```

```
##        1
## 58.34221
```

Així, trobem que el seu ràting està "infravalorat" amb una diferencia de 13 comparat amb el model.

Una explicació a aquesta situació podria ser que es valora al alça els jugadors més mediàtics per convertir la diferència emocional en un valor objectiu i es valora a la baixa els jugador menys mediàtics per augmentar la diferència amb els que ho son més enllà dels valors objectius. Així, el joc estaria maximitzant l'espectacle

de la experiència reafirmant biaixos dels espectadors perjudicant la simulació realista i tàctica dels valors mesurables.

4.3.3 Regressió logística

4.3.3.1 Model Finalment, per tal de saber si un jugador té podencial d'anar a la selcció nacional farem servir el seu normero de dorsal nacional "National_Kit" com etiqueta de resultats. Primer de tot creem una columna amb nom internacional on 0 = no és seleccionat per a la selecció i 1 = seleccionat per a la selecció.

```
for (i in 1:length(fifaNet$National_Kit)){
  if (is.na(fifaNet$National_Kit[i])){
    fifaNet$internacional[i] = 0
  } else{
    fifaNet$internacional[i] = 1
  }
}
```

Seguidament creem el model logìstic a partir de les variables clasificacion2, Rating, Age i Work_Rate. La raó d'aquestes varibales és degut que partint del model lineal veiem que "clasificacion2" que seria l'equivalent a una evaluació si el jugador és bo o no és un estimador amb molt pes amb el resultat final; el rating, obtingut del model lineal, és un punt que s'ha cosiderat important a la hora de saber si un jugador anirà a la selecció; L'edat s'ha seleccionat degut que com hem vist al test d'hipòtesi sembla que que els jugadors més ben valorats són els que tenen més edats pel que s'estima que serà un estimador influent; I finalment Work_Rate que similar a la "clasificacion2" també és una evaluació del rendiment del jugador que es podria obtenir sense haver de realitzar un exàmen físic del matiex jugador.

```
# Model logistic
mrl <- glm(internacional ~ clasificacion2 + Rating + Age + Work_Rate,
           family = binomial(link = logit), data = fifaNet)
summary(mrl)
```

```
##
## Call:
## glm(formula = internacional ~ clasificacion2 + Rating + Age +
##     Work_Rate, family = binomial(link = logit), data = fifaNet)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5827  -0.3474  -0.2220  -0.1358   3.6109
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -15.338319   0.602103 -25.475  < 2e-16 ***
## clasificacion2Bo           0.059286   0.116016   0.511 0.609340
## Rating                     0.196367   0.008738  22.473  < 2e-16 ***
## Age                       -0.025564   0.008558  -2.987 0.002815 **
## Work_RateHigh / Low       -0.434751   0.189956  -2.289 0.022097 *
## Work_RateHigh / Medium    -0.489358   0.132643  -3.689 0.000225 ***
## Work_RateLow / High       -0.536066   0.245249  -2.186 0.028830 *
## Work_RateLow / Low        -1.412971   1.068384  -1.323 0.185991
## Work_RateLow / Medium     -0.554730   0.258379  -2.147 0.031797 *
## Work_RateMedium / High    -0.435535   0.149327  -2.917 0.003538 **
## Work_RateMedium / Low     -0.853741   0.202511  -4.216 2.49e-05 ***
## Work_RateMedium / Medium  -0.553934   0.122836  -4.510 6.50e-06 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8091.8  on 17586  degrees of freedom
## Residual deviance: 6457.3  on 17575  degrees of freedom
## AIC: 6481.3
##
## Number of Fisher Scoring iterations: 6
```

```
# Devianza
anova(mrl,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: internacional
##
## Terms added sequentially (first to last)
##
##
##                Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                           17586     8091.8
## clasificacion2  1  1016.56     17585     7075.2 < 2.2e-16 ***
## Rating          1   581.59     17584     6493.7 < 2.2e-16 ***
## Age             1    10.43     17583     6483.2  0.001237 **
## Work_Rate       8    25.90     17575     6457.3  0.001094 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Trobem que Work_Rate és una varibale dicotòmica amb 9 opcions pel que s'ha obtingut un estimadors per cada un de les possibilitats agafant el Work_Rate High/High de referència.
Cal destacar que Work_Rate Low/Low té un estimador negatiu consdirablement alt respecte els altres estimadors, això ens està dient, como és lògic, un jugadora amb un rendiment baix/baix té moltíssimes menys probabilitats de ser seleccionat en compració amb els altres.

Observem que el model es vàlid degut a la devianza residual és menor que la devianza nu · la, on fent anova en veiem que tots els p-values són significatius. Per tant, confirmem que les variables seleccionades tenen pes a la hora d'obtenir el model final.

4.3.3.2 Matriu de confusió Seguidament comprovem la presició del model logístic a paritr del dataset utilitznat la columna internacional com a resultat per comprovar la correcta predicció:

```
# Exercici 5.2
newFifaNet <- fifaNet[, c("clasificacion2", "Rating", "Age", "Work_Rate")]
pred2 <- predict(mrl, newFifaNet ,type ="response")

matriuConf <- matrix(0, nrow = 3, ncol = 3,
                     dimnames = list(c("Internacional", "No Internacional",
                                       "Total"),c("< 50%", ">= 50%", "Total")))
for (i in 1:length(pred2)){
  if (pred2[i] < 0.50){
    if (fifaNet$internacional[i] == 1){
      matriuConf[1,1] <- matriuConf[1,1] + 1
    }
```

```
    else {
      matriuConf[2,1] <- matriuConf[2,1] + 1
    }
  }
  else {
    if (fifaNet$internacional[i] == 1){
      matriuConf[1,2] <- matriuConf[1,2] + 1
    }
    else {
      matriuConf[2,2] <- matriuConf[2,2] + 1
    }
  }
}
# Total, < 50% i >= 50%
matriuConf[3,1] <- matriuConf[1,1] + matriuConf[2,1]
matriuConf[3,2] <- matriuConf[1,2] + matriuConf[2,2]
# Internacional i No Internacional, Total
matriuConf[1,3] <- matriuConf[1,1] + matriuConf[1,2]
matriuConf[2,3] <- matriuConf[2,1] + matriuConf[2,2]
# Total, Total
matriuConf[3,3] <- matriuConf[3,1] + matriuConf[3,2]

matriuConf
```

```
##                   < 50% >= 50% Total
## Internacional       985     90  1075
## No Internacional  16474     38 16512
## Total             17459    128 17587
```

Com observem a la matriu hi ha un total de 16564 jugadors que estàn correctament classificats això significa que el model té una presició del 16564/17587 = 94,18%.

Per altra banda, hi ha un alt nombre de jugadors que tenien una probabilitat de ser internacional menor del 50% i han sigut seleccionats 985 d'questes. De la mateixa manera hi havia 38 casos que tenien una alta probabilitat i no han sigut seleccionats.

4.3.3.3 Predicció. Seguidament realitzem un test de predicció d'un jugador que s'espera que no sigui seleccionat per la selecció nacional i un altre amb altes probabilitats de ser seleccionat:

- **Jugador Bo**: Jugador de 27 anys (mitjana d'edat de jugadors bons trobat al contrast d'hipòtesis), un ràting de 90 punts i una classificació de Work_Rate com High/High
- **Jugador Regular**:Jugador de 24 anys (mitjana d'edat de jugadors regulars trobat al contrast d'hipòtesis), un ràting de 69 punts i una classificació de Work_Rate com Medium/Medium

```
# Jugador bo
pred3 <- predict(mrl, data.frame(clasificacion2 = "Bo", Rating = 90, Age = 27,
                                 Work_Rate = "High / High") ,type ="response")
cat("Probabilitat que el jugador bo sigui seleccionat: ", pred3, "\n")
```

```
## Probabilitat que el jugador bo sigui seleccionat:  0.8460277
```

```
# Jugador regular
pred4 <- predict(mrl, data.frame(clasificacion2 = "Regular", Rating = 69,
                                 Age = 24, Work_Rate = "Medium / Medium"),
                 type ="response")
cat("Probabilitat que el jugador regular sigui seleccionat: ", pred4)
```

```
## Probabilitat que el jugador regular sigui seleccionat:  0.0494321
```

Observe, que hi ha una possibilitat que el primer jugador sigui seleccionat és del 84,60% mentre que el que és més regular té una molt baixa probabilitat de 4,94%

Veiem les possibilitat d'un jugador jove que té potencial:

- **Jugador jove**: Jugador de 21 anys, un ràting de 79 punts i una classificació de Work_Rate com High/Medium

```
# Jugador jove
pred5 <- predict(mrl, data.frame(clasificacion2 = "Bo", Rating = 79,
                                 Age = 21, Work_Rate = "High / Medium"),
                 type ="response")
cat("Probabilitat que el jugador bo sigui seleccionat: ", pred5)
```

```
## Probabilitat que el jugador bo sigui seleccionat:  0.3116895
```

S'ha posat un jugagador jove amb un ràting 79 (Bo) dins de l'escala especificada en el punt 3.2, veiem que la probabilitat de ser seleccionat segueix siguint relativament baix però és molt més alt en comparació un jugador regular.

---

5 Conclusions

5.1 Contrast d'hipòtesis

En aquest anàlisi s'ha realitzat 3 tipus diferents, un test d'hipotèsis amb suposicions de les habilitats d'un jugador amb respecte el seu estat físic, un anàlisi amb model lineal per tal d'obtenir una valoració numèrica del jugador a partir d'informació que es podria extreure visualment i finalment un model logìstic per evaluar les possibilitat que un jugador sigui seleccionat per la selecció nacional.

En el primer cas hem trobat que un jugador jove no sembla ser millors que un més veterà/senior, d'aqui es poden extreure la conclusió que si es voldria resultats de millora immediat s'haura de fitxar jugadors amb edats al voltant de 27 anys que tendeixen a ser més bons, però si es vol millorar un jugador de cara a futur, ja sigui per costos o disponibilitat, val la pena fitxar un jugador al voltant dels 24 ja que tenen potencial a millorar de cara al futur.

Després hem vist que de cara als jugadors la posició del porter és la que tendeix a tenir més pes degut a la necessitat de força per parar xuts dels delanters de l'equip contrari.

5.2 Model lineal

Les variables 'Ball_Control', 'Age','Height', 'Weight', 'clasificacion2' tenen una correlació positiva i significativa amb 'Rating', en canvi, Skill_Moves té una correlació negativa. En conjunt, les variables explicatives expliquen gariaré un 70% de la variància de 'Rating'.

5.3 Model logístic

En aquest apartat hem realitzat un model de regressió logística per tal de poder obtenir la probabilitat que un juegador sigui seleccionat per jugar internacionalment en base de certes característiques. En aquest apartat hem vist que els principals regressors que afecta a la probabilitat de ser seleccionat és si és Bo o Regular, el seu Rating i el Work_Rate.
És interessant destacar, que segons quin nivell de Work_Rate agafem de referència els altres nivells tenen un efecte major o menor, de la mateixa manera un gran efecte és el rating del jugador on naturalment un jugador amb major ràting vol dir que és més bo.

Contribuciones	Firma
Investigació prèvia	Junjie Zhu, Antoni Sanchez Teruel
Redacció de les respostes	Junjie Zhu, Antoni Sanchez Teruel
Desenvolupament codi	Junjie Zhu, Antoni Sanchez Teruel