PRA 2 - Análisis de datos FIFA 17

Autores : Pablo Román-Naranjo Varela y Adrián Vicente Gómez

Enero 2022

Contents

1 Descripción del dataset

El dataset utilizado, fifa.csv, está disponible en Kaggle: https://www.kaggle.com/artimous/complete-fifa-2017-player-dataset-global. Este dataset contiene las **estadísticas de los jugadores en el videojuego Fifa 17**. Contiene 17588 filas con 53 variables diferentes para describir a cada jugador. Entre estas variables nos encontramos con las siguientes:

- Name (Nombre del jugador)
- Nationality (Nacionalidad del jugador)
- National_Position (Posición de juego en equipo nacional).
- National_Kit (Número de equipación en equipo nacional)
- Club (Nombre del club)

- Club_Position (Posición de juego en club)
- Club_Kit (Número de equipación en club)
- Club_Joining (Fecha en la que empezó en el club)
- Contract_Expire (Año finalización del contrato)
- Rating (Valoración global del jugador, entre 0 y 100)
- Height (Altura)
- Weight (Peso)
- Preffered_Foot (Pie preferido)
- Birth_Date (Fecha de nacimiento)
- Age (Edad)
- Preffered_Position (Posición preferida)
- Work_Rate (valoración cualitativa en términos de ataque-defensa)
- Weak_foot (valoración de 1 a 5 de control y potencia de la pierna no preferida)
- Skill_Moves (valoración de 1 a 5 de la habilidad en movimientos del jugador)
- El resto de variables son los distintos atributos que definen a un jugador en este juego.

En la actualidad, el fútbol forma parte de nuestra vida cotidiana y es el **deporte con un mayor impacto social**. Aunque no deja de ser un videojuego, FIFA maneja tantos detalles sobre los jugadores que los ojeadores y los empleados de los clubes lo usan como herramienta para buscar nuevos jugadores que de otra manera no podrían ver. Por ello, realizar un análisis estadístico de los datos de este juego podría permitir arrojar cordura por encima de los sentimientos a la hora de tomar diferentes decisiones. Como ejemplo de estos análisis, **en este documento se planteará**:

- Si en el año 2017, la diferencia entre valoración global entre los jugadores del Real Betis y el Real Madrid era mayor a 5 puntos.
- Obtener un modelo de regresión que nos permita estimar la valoración de un jugador a partir de un subjconjunto de sus atributos.
- Obtener un modelo de regresión que nos permita determinar qué jugadores tienen más probabilidades de ir con la selección española.

2 Intregración y selección de datos de interés a analizar

Se carga el archivo usando la función **read.csv()**, con el separador como coma. Además, se especifica que archivo está **codificado en UTF** para que pueda leer de manera correcta los caracteres especiales como tildes o diéresis.

```
datos = read.csv("../data/fifa.csv", header=T, sep=",", fileEncoding = "UTF-8")

# inspeccionamos el dataset
head(datos)
```

```
##                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 Suárez     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
```

Usando el comando **head()** se inspeccionan las primeras líneas del fichero, y se observa que se ha asignado correctamente el nombre de cada variable.

Para ver qué tipo de datos se asigna a cada variable se utiliza la función **str()**.

```
str(datos)
```

```
## 'data.frame':    17588 obs. of  53 variables:
##  $ Name               : chr  "Cristiano Ronaldo" "Lionel Messi" "Neymar" "Luis Suárez" ...
##  $ Nationality        : chr  "Portugal" "Argentina" "Brazil" "Uruguay" ...
##  $ National_Position  : chr  "LS" "RW" "LW" "LS" ...
##  $ National_Kit       : num  7 10 10 9 1 1 9 11 NA 1 ...
##  $ Club               : chr  "Real Madrid" "FC Barcelona" "FC Barcelona" "FC Barcelona" ...
##  $ Club_Position      : chr  "LW" "RW" "LW" "ST" ...
##  $ Club_Kit           : num  7 10 11 9 1 1 9 11 9 13 ...
##  $ Club_Joining       : chr  "07/01/2009" "07/01/2004" "07/01/2013" "07/11/2014" ...
##  $ Contract_Expiry    : num  2021 2018 2021 2021 2021 ...
##  $ Rating             : int  94 93 92 92 92 90 90 90 90 89 ...
##  $ Height             : chr  "185 cm" "170 cm" "174 cm" "182 cm" ...
##  $ Weight             : chr  "80 kg" "72 kg" "68 kg" "85 kg" ...
##  $ Preffered_Foot     : chr  "Right" "Left" "Right" "Right" ...
##  $ Birth_Date         : chr  "02/05/1985" "06/24/1987" "02/05/1992" "01/24/1987" ...
##  $ Age                : int  32 29 25 30 31 26 28 27 35 24 ...
##  $ Preffered_Position: chr  "LW/ST" "RW" "LW" "ST" ...
##  $ Work_Rate          : chr  "High / Low" "Medium / Medium" "High / Medium" "High / Medium" ...
##  $ Weak_foot          : int  4 4 5 4 4 3 4 3 4 3 ...
##  $ Skill_Moves        : int  5 4 5 4 1 1 3 4 4 1 ...
##  $ Ball_Control       : int  93 95 95 91 48 31 87 88 90 23 ...
##  $ Dribbling          : int  92 97 96 86 30 13 85 89 87 13 ...
##  $ Marking            : int  22 13 21 30 10 13 25 51 15 11 ...
##  $ Sliding_Tackle     : int  23 26 33 38 11 13 19 52 27 16 ...
##  $ Standing_Tackle    : int  31 28 24 45 10 21 42 55 41 18 ...
##  $ Aggression         : int  63 48 56 78 29 38 80 65 84 23 ...
##  $ Reactions          : int  96 95 88 93 85 88 88 87 85 81 ...
##  $ Attacking_Position: int  94 93 90 92 12 12 89 86 86 13 ...
##  $ Interceptions      : int  29 22 36 41 30 30 39 59 20 15 ...
##  $ Vision             : int  85 90 80 84 70 68 78 79 83 44 ...
##  $ Composure          : int  86 94 80 83 70 60 87 85 91 52 ...
##  $ Crossing           : int  84 77 75 77 15 17 62 87 76 14 ...
##  $ Short_Pass         : int  83 88 81 83 55 31 83 86 84 32 ...
##  $ Long_Pass          : int  77 87 75 64 59 32 65 80 76 31 ...
##  $ Acceleration       : int  91 92 93 88 58 56 79 93 69 46 ...
##  $ Speed              : int  92 87 90 77 61 56 82 95 74 52 ...
##  $ Stamina            : int  92 74 79 89 44 25 79 78 75 38 ...
##  $ Strength           : int  80 59 49 76 83 64 84 80 93 70 ...
##  $ Balance            : int  63 95 82 60 35 43 79 65 41 45 ...
##  $ Agility            : int  90 90 96 86 52 57 78 77 86 61 ...
##  $ Jumping            : int  95 68 61 69 78 67 84 85 72 68 ...
##  $ Heading            : int  85 71 62 77 25 21 85 86 80 13 ...
##  $ Shot_Power         : int  92 85 78 87 25 31 86 91 93 36 ...
##  $ Finishing          : int  93 95 89 94 13 13 91 87 90 14 ...
##  $ Long_Shots         : int  90 88 77 86 16 12 82 90 88 17 ...
##  $ Curve              : int  81 89 79 86 14 21 77 86 82 19 ...
##  $ Freekick_Accuracy : int  76 90 84 84 11 19 76 85 82 11 ...
```

```
##  $ Penalties         : int  85 74 81 85 47 40 81 76 91 27 ...
##  $ Volleys           : int  88 85 83 88 11 13 86 76 93 12 ...
##  $ GK_Positioning    : int  14 14 15 33 91 86 8 5 9 86 ...
##  $ GK_Diving         : int  7 6 9 27 89 88 15 15 13 84 ...
##  $ GK_Kicking        : int  15 15 15 31 95 87 12 11 10 69 ...
##  $ GK_Handling       : int  11 11 9 25 90 85 6 15 15 91 ...
##  $ GK_Reflexes       : int  11 8 11 37 89 90 10 6 12 89 ...
```

Como se ve, se dispone de **53 variables** que describen a cada jugador, y el tipo en el que se han cargado.

3 Limpieza de los datos

3.1. Análisis de valores ausentes

Para realizar un análisis de los datos, realizamos un resumen de los mismos con la función **summary()**.

```
summary(datos)
```

```
##      Name           Nationality        National_Position   National_Kit
##  Length:17588       Length:17588       Length:17588       Min.   : 1.00
##  Class :character   Class :character   Class :character   1st Qu.: 6.00
##  Mode  :character   Mode  :character   Mode  :character   Median :12.00
##                                                           Mean   :12.22
##                                                           3rd Qu.:18.00
##                                                           Max.   :36.00
##                                                           NA's   :16513
##      Club           Club_Position         Club_Kit     Club_Joining
##  Length:17588       Length:17588       Min.   : 1.00   Length:17588
##  Class :character   Class :character   1st Qu.: 9.00   Class :character
##  Mode  :character   Mode  :character   Median :18.00   Mode  :character
##                                        Mean   :21.29
##                                        3rd Qu.:27.00
##                                        Max.   :99.00
##                                        NA's   :1
##  Contract_Expiry     Rating         Height             Weight
##  Min.   :2017    Min.   :45.00   Length:17588       Length:17588
##  1st Qu.:2017    1st Qu.:62.00   Class :character   Class :character
##  Median :2019    Median :66.00   Mode  :character   Mode  :character
##  Mean   :2019    Mean   :66.17
##  3rd Qu.:2020    3rd Qu.:71.00
##  Max.   :2023    Max.   :94.00
##  NA's   :1
##  Preffered_Foot      Birth_Date             Age        Preffered_Position
##  Length:17588       Length:17588       Min.   :17.00   Length:17588
##  Class :character   Class :character   1st Qu.:22.00   Class :character
##  Mode  :character   Mode  :character   Median :25.00   Mode  :character
##                                        Mean   :25.46
##                                        3rd Qu.:29.00
##                                        Max.   :47.00
##
##   Work_Rate           Weak_foot      Skill_Moves     Ball_Control
```

```
##  Length:17588       Min.   :1.000   Min.   :1.000   Min.   : 5.00
##  Class :character   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:53.00
##  Mode  :character   Median :3.000   Median :2.000   Median :63.00
##                     Mean   :2.934   Mean   :2.303   Mean   :57.97
##                     3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:69.00
##                     Max.   :5.000   Max.   :5.000   Max.   :95.00
##
##    Dribbling       Marking      Sliding_Tackle  Standing_Tackle   Aggression
##  Min.   : 4.0   Min.   : 3.00   Min.   : 5.00   Min.   : 3.00   Min.   : 2.00
##  1st Qu.:47.0   1st Qu.:22.00   1st Qu.:23.00   1st Qu.:26.00   1st Qu.:44.00
##  Median :60.0   Median :48.00   Median :51.00   Median :54.00   Median :59.00
##  Mean   :54.8   Mean   :44.23   Mean   :45.57   Mean   :47.44   Mean   :55.92
##  3rd Qu.:68.0   3rd Qu.:64.00   3rd Qu.:64.00   3rd Qu.:66.00   3rd Qu.:70.00
##  Max.   :97.0   Max.   :92.00   Max.   :95.00   Max.   :92.00   Max.   :96.00
##
##    Reactions     Attacking_Position Interceptions       Vision
##  Min.   :29.00   Min.   : 2.00      Min.   : 3.00   Min.   :10.00
##  1st Qu.:55.00   1st Qu.:37.00      1st Qu.:26.00   1st Qu.:43.00
##  Median :62.00   Median :54.00      Median :52.00   Median :54.00
##  Mean   :61.77   Mean   :49.59      Mean   :46.79   Mean   :52.71
##  3rd Qu.:68.00   3rd Qu.:64.00      3rd Qu.:64.00   3rd Qu.:64.00
##  Max.   :96.00   Max.   :94.00      Max.   :93.00   Max.   :94.00
##
##    Composure        Crossing       Short_Pass      Long_Pass     Acceleration
##  Min.   : 5.00   Min.   : 6.00   Min.   :10.00   Min.   : 7.0   Min.   :11.00
##  1st Qu.:47.00   1st Qu.:38.00   1st Qu.:52.00   1st Qu.:42.0   1st Qu.:57.00
##  Median :57.00   Median :54.00   Median :62.00   Median :56.0   Median :68.00
##  Mean   :55.85   Mean   :49.74   Mean   :58.12   Mean   :52.4   Mean   :65.29
##  3rd Qu.:66.00   3rd Qu.:64.00   3rd Qu.:68.00   3rd Qu.:64.0   3rd Qu.:75.00
##  Max.   :94.00   Max.   :91.00   Max.   :92.00   Max.   :93.0   Max.   :96.00
##
##      Speed          Stamina         Strength        Balance
##  Min.   :11.00   Min.   :10.00   Min.   :20.00   Min.   :10.00
##  1st Qu.:58.00   1st Qu.:57.00   1st Qu.:57.00   1st Qu.:56.00
##  Median :68.00   Median :66.00   Median :66.00   Median :65.00
##  Mean   :65.48   Mean   :63.48   Mean   :65.09   Mean   :64.01
##  3rd Qu.:75.00   3rd Qu.:74.00   3rd Qu.:74.00   3rd Qu.:74.00
##  Max.   :96.00   Max.   :95.00   Max.   :98.00   Max.   :97.00
##
##     Agility         Jumping         Heading        Shot_Power
##  Min.   :11.00   Min.   :15.00   Min.   : 4.00   Min.   : 3.00
##  1st Qu.:55.00   1st Qu.:58.00   1st Qu.:45.00   1st Qu.:45.00
##  Median :65.00   Median :65.00   Median :56.00   Median :59.00
##  Mean   :63.21   Mean   :64.92   Mean   :52.39   Mean   :55.58
##  3rd Qu.:74.00   3rd Qu.:73.00   3rd Qu.:65.00   3rd Qu.:69.00
##  Max.   :96.00   Max.   :95.00   Max.   :94.00   Max.   :93.00
##
##    Finishing       Long_Shots       Curve       Freekick_Accuracy
##  Min.   : 2.00   Min.   : 4.0   Min.   : 6.00   Min.   : 4.00
##  1st Qu.:29.00   1st Qu.:32.0   1st Qu.:34.00   1st Qu.:31.00
##  Median :48.00   Median :52.0   Median :48.00   Median :42.00
##  Mean   :45.16   Mean   :47.4   Mean   :47.18   Mean   :43.38
##  3rd Qu.:61.00   3rd Qu.:63.0   3rd Qu.:62.00   3rd Qu.:57.00
##  Max.   :95.00   Max.   :91.0   Max.   :92.00   Max.   :93.00
```

```
##
##    Penalties        Volleys      GK_Positioning    GK_Diving
##  Min.   : 7.00   Min.   : 3.00   Min.   : 1.00   Min.   : 1.00
##  1st Qu.:39.00   1st Qu.:30.00   1st Qu.: 8.00   1st Qu.: 8.00
##  Median :50.00   Median :44.00   Median :11.00   Median :11.00
##  Mean   :49.17   Mean   :43.28   Mean   :16.61   Mean   :16.82
##  3rd Qu.:61.00   3rd Qu.:57.00   3rd Qu.:14.00   3rd Qu.:14.00
##  Max.   :96.00   Max.   :93.00   Max.   :91.00   Max.   :89.00
##
##    GK_Kicking     GK_Handling     GK_Reflexes
##  Min.   : 1.00   Min.   : 1.00   Min.   : 1.0
##  1st Qu.: 8.00   1st Qu.: 8.00   1st Qu.: 8.0
##  Median :11.00   Median :11.00   Median :11.0
##  Mean   :16.46   Mean   :16.56   Mean   :16.9
##  3rd Qu.:14.00   3rd Qu.:14.00   3rd Qu.:14.0
##  Max.   :95.00   Max.   :91.00   Max.   :90.0
##
```

Si se observa el resumen realizado, en él es posible estudiar **qué variables disponen de valores ausentes**, marcados como NA's. En este caso, de **las variables numéricas referentes a estadísticas** de jugadores, **ninguna muestra valores ausentes**.

Donde sí se pueden observar **valores ausentes es en las variables National_Kit, National_Position, Club_Kit y Contract_Expiry**. En el caso de las dos primeras variables, National_Kit y National_Position, es normal la existencia de valores nulos, ya que solo encontraremos un valor en estas variables si el jugador forma parte actualmente de la selección de su país. En las otras dos variables, Club_Kit y Contract_Expiry, **solo existe un valor nulo**. Se puede comprobar que **este valor corresponde a un único jugador**.

```
datos[is.na(datos$Club_Kit),]
```

```
##              Name Nationality National_Position National_Kit       Club
## 384 Didier Drogba Ivory Coast                               NA Free agent
##     Club_Position Club_Kit Club_Joining Contract_Expiry Rating Height Weight
## 384                    NA                           NA     81 189 cm  80 kg
##     Preffered_Foot Birth_Date Age Preffered_Position    Work_Rate Weak_foot
## 384          Right 03/11/1978  39                 ST Medium / Low         4
##     Skill_Moves Ball_Control Dribbling Marking Sliding_Tackle Standing_Tackle
## 384           3           80        74      22             29              32
##     Aggression Reactions Attacking_Position Interceptions Vision Composure
## 384         80        80                 81            42     76        80
##     Crossing Short_Pass Long_Pass Acceleration Speed Stamina Strength Balance
## 384       67         60        60           64    64      62       86      56
##     Agility Jumping Heading Shot_Power Finishing Long_Shots Curve
## 384      63      76      85         85        82         79    78
##     Freekick_Accuracy Penalties Volleys GK_Positioning GK_Diving GK_Kicking
## 384                84        84      76              6        10          8
##     GK_Handling GK_Reflexes
## 384          11          14
```

Concretamente el jugador es **Didier Drogba**, siendo además un agente libre. En el caso de esta observación, **dado que no dispone de Posición en el club se eliminará ya que en los próximos apartados se usará dicha variable**.

```
fifaNet = datos[!is.na(datos$Club_Kit),]
```

Por lo tanto, se dispone de unos datos limpios respecto a la ausencia de valores que podrán ser usados sin problemas.

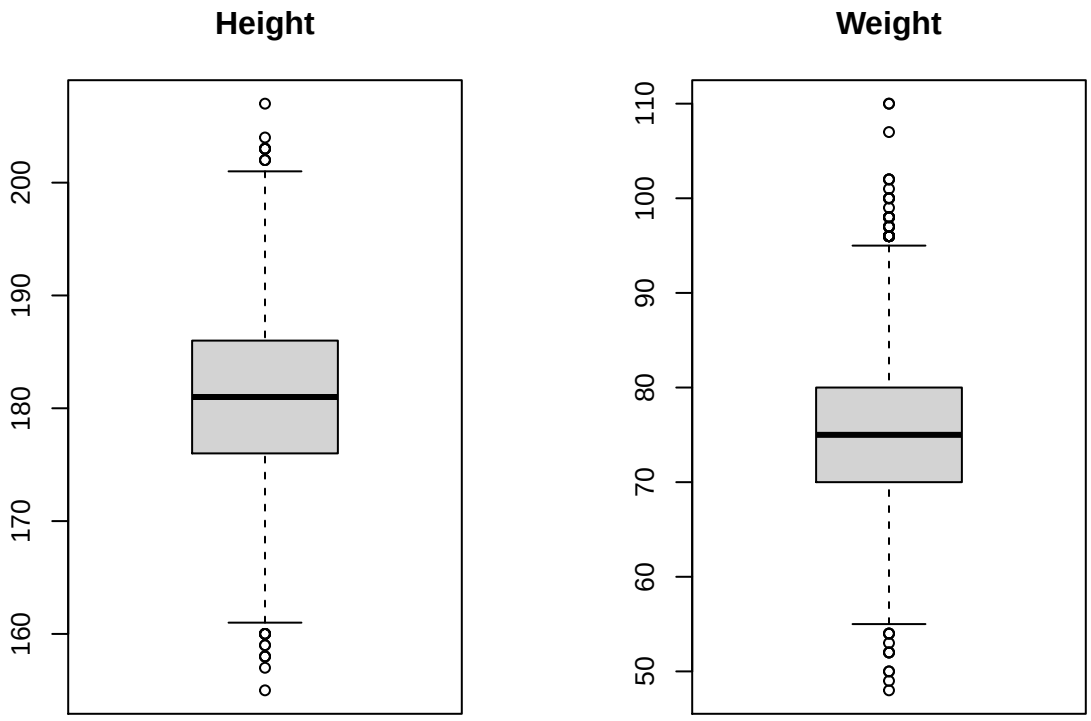
3.2. Valores extremos

Para analizar los **valores extremos** observamos nuevamente el summary de los datos. En este caso, **todas las estadísticas del jugador** se encuentran **entre 0 y 100** que son los **rangos habituales** de estos parámetros en el juego. Conviene realizar la **transformación del peso y la altura a variables numéricas** para así comprobar también si existe algún valor extremos que sea necesario tratar.

```
# Tranformamos ambas variables
fifaNet$Height = as.numeric(gsub("[[:alpha:]]","",fifaNet$Height))
fifaNet$Weight = as.numeric(gsub("[[:alpha:]]","",fifaNet$Weight))
```

En este caso **se ha eliminado cualquier caracter alfabético de las variables**, aunque se podría también haber eliminado las cadenas 'kg' y 'cm' directamente. Comprobamos los boxplot de ambas variables para estas dos variables en búsqueda de valores extremos:

```
par(mai=rep(0.6, 4))
layout(matrix(c(1,1, 2,2), ncol = 2))
boxplot(fifaNet$Height, main = "Height")
boxplot(fifaNet$Weight, main = "Weight")
```



Analizando ambas variables, aunque aparecen puntos fuera de los boxplot, se ve que **se mueven en rangos comunes** para peso y altura, por lo que no se identifican valores extremos.

4 Análisis de los datos

4.1. Selección de los grupos de datos que se quiere analizar/comparar

En esta práctica haremos tres aproximaciones para analizar el dataset elegido. Estos análisis se dividirán en tres apartados:

- **Estadística inferencial**

En este apartado **analizaremos las plantillas del Real Madrid y el Real Betis**. Plantearemos como pregunta de investigación si el *Rating* general de cada una de las plantillas se diferencian en más de 5 puntos. Para realizar este análisis deberemos hacer **dos grupos de datos**: madrid_players y betis_players.

- **Modelo de regresión lineal múltiple**

En el segundo apartado planteamos ajustar un modelo de regresión lineal múltiple para estimar el rating general de los jugadores. En este apartado **usaremos la totalidad del dataset (fifaNet)**, usando como variables explicativas las variables *Age*, *Weight*, *Club_Position*, *Vision*, *Ball_Control*, *Marking*, *Interceptions*, *Freekick_Accuracy*, *Short_Pass*, *Speed* y *Finishing*.

- **Regresión logística**

En tercer y último apartado ajustaremos un modelo predictivo basado en **regresión logística para predecir la probabilidad de que un jugador vaya a la selección de su país**. En este caso, para que la muestra sea más homogénea, hemos seleccionado solo aquellos jugadores nacidos en España. Por tanto, en este apartado usaremos un **subconjunto extraido de fifaNet incluyendo las estadísticas de los jugadores españoles (fifaNet_spain)**.

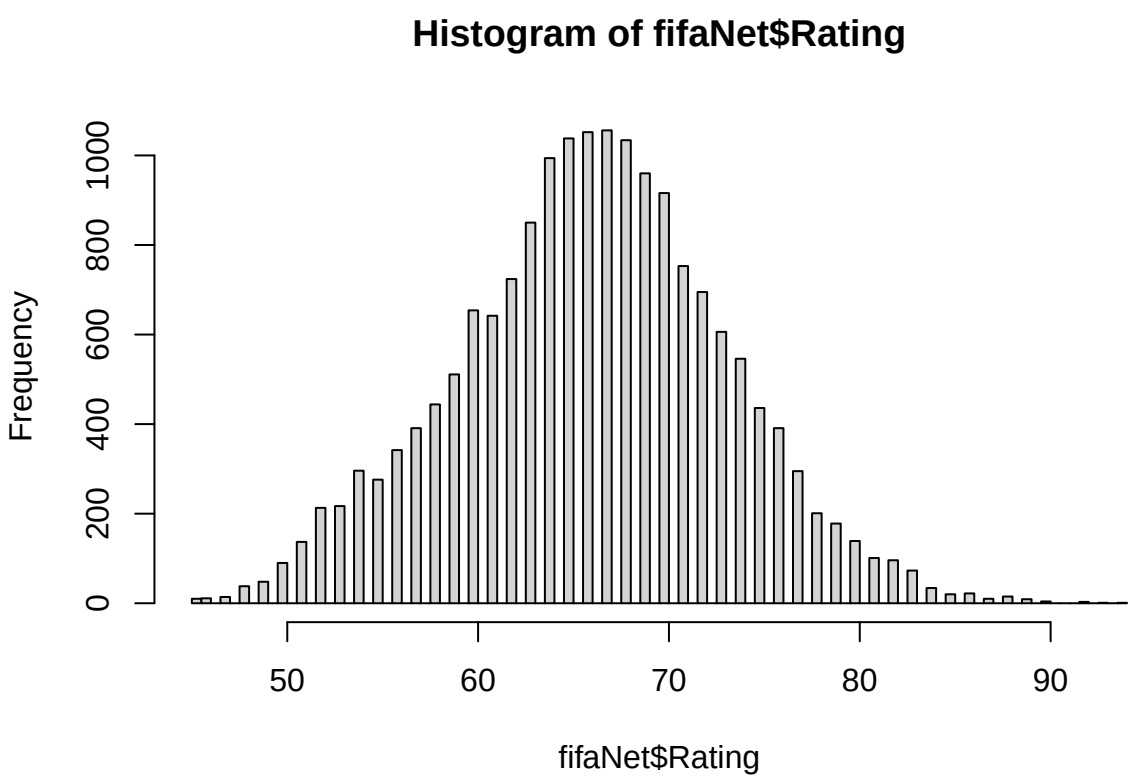
4.2. Estadística inferencial

En este primer análisis se desea saber si, en el año 2017 **la diferencia entre valoración global (*Rating*) entre los jugadores del Real Betis y el Real Madrid era mayor a 5 puntos**. Para ello se llevará a cabo un **contraste de hipótesis**.

4.2.1. Comprobación de normalidad

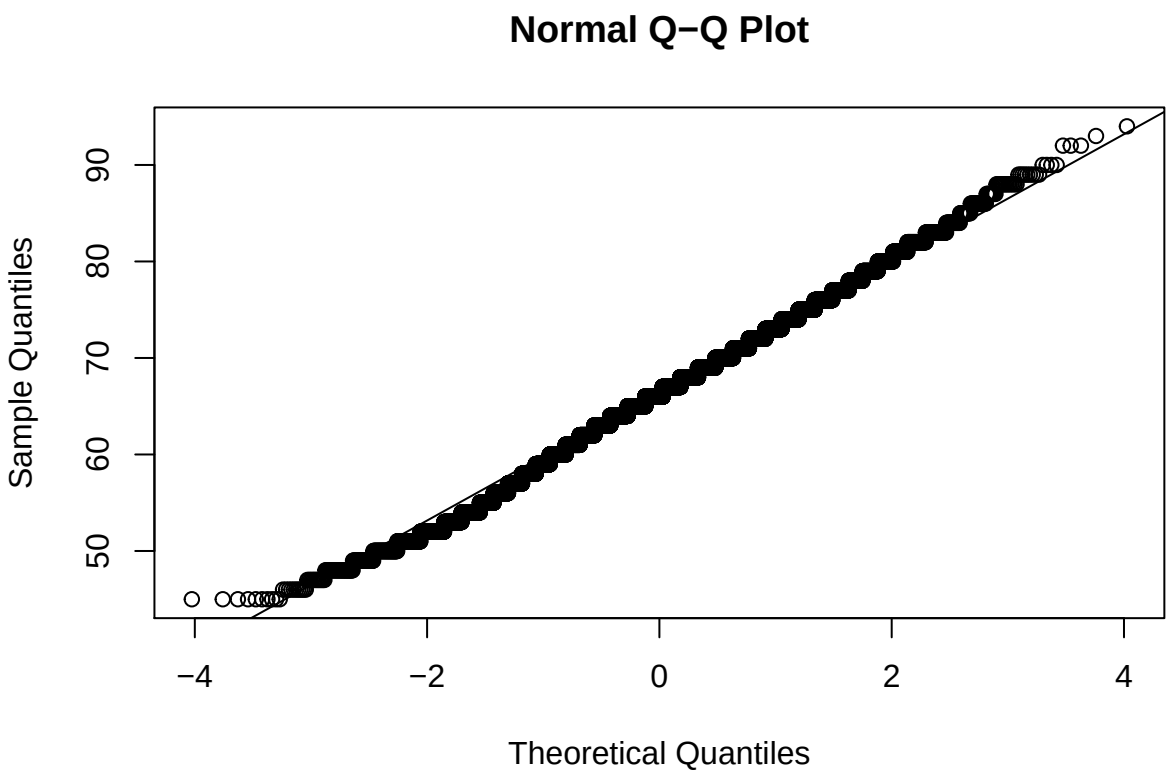
Para comprobar si la variable *Rating* se distribuye de manera normal, se puede realizar su **histograma** y evaluar si se corresponde con una **distribución normal**.

```
# se usa la raiz cuadrada del numero de datos para establecer cuantas columnas usar
hist(fifaNet$Rating, breaks=sqrt(nrow(fifaNet)))
```



A la vista del histograma, **parece que esta variable sigue una distribución aproximadamente normal**, pero con la cola derecha algo menor que la izquierda. Para confirmar la normalidad, se puede realizar un **Q-Q plot** de los valores de *Rating*.

```
qqnorm(fifaNet$Rating)
qqline(fifaNet$Rating)
```



Dado que los puntos se adaptan casi a la perfección a la recta en el QQ plot, y debido al tamaño elevado de la muestra, **asumimos normalidad de la variable *Rating*** por el teorema del límite central.

4.2.2. Contraste de hipótesis para la diferencia de medias

4.2.2.1. Hipótesis nula y la alternativa Se desea comprobar si la **valoración de los jugadores del Real Betis es 5 puntos menor que la de los jugadores del Real Madrid**. Para ello, las hipótesis para este caso se formulan del siguiente modo:

$$\begin{cases} H_0: & \mu_m - \mu_b = 5 \\ H_1: & \mu_m - \mu_b > 5 \end{cases}$$

Siendo μ_m la media de la valoración de los jugadores del Real Madrid y μ_b la media de la valoración de los jugadores del Real Betis.

4.2.2.2. Test a aplicar El test a aplicar sería un **test unilateral sobre la diferencia de medias de dos muestras independientes (varianzas desconocidas)**. Es un **test unilateral por la derecha** porque la pregunta de investigación cuestiona si la media de los jugadores del Real Madrid es mayor a la de los jugadores del Real Betis.

4.2.2.3. Aplicación y comprobación del test Primeramente se deben obtener los datos de los jugadores del Real Madrid y Real Betis.

```
betis_players = fifaNet$Rating[fifaNet$Club=="Real Betis"]
madrid_players = fifaNet$Rating[fifaNet$Club=="Real Madrid"]
length(betis_players)
```

```
## [1] 28
```

```
length(madrid_players)
```

```
## [1] 33
```

Dado que acabamos de comprobar la normalidad de la variable *Rating*, podemos asumir que esta normalidad se trasladará a los dos subconjuntos.

Para concretar más aún el test que se va a realizar, será necesario realizar un **test de homoscedasticidad**, dado que se desconocen las varianzas poblacionales. Con este test se comprobará si las varianzas son iguales o diferentes.

```
var.test(betis_players, madrid_players)
```

```
##
##  F test to compare two variances
##
## data:  betis_players and madrid_players
## F = 0.32757, num df = 27, denom df = 32, p-value = 0.004105
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.1583288 0.6937002
## sample estimates:
## ratio of variances
##          0.3275727
```

Dado que la hipótesis nula implica igualdad de varianzas, y el p-valor obtenido es inferior al nivel de significación α (0.05), se rechaza la hipótesis nula y **se asume que las varianzas son diferentes**.

Por tanto, el test a aplicar será una t de Student con v grados de libertad, también llamado **test de Welch**.

```
t.test(madrid_players,betis_players, var.equal = FALSE,alternative = "greater",
       mu = 5)
```

```
##
##  Welch Two Sample t-test
##
## data:  madrid_players and betis_players
## t = 0.042313, df = 52.248, p-value = 0.4832
## alternative hypothesis: true difference in means is greater than 5
## 95 percent confidence interval:
##  1.952377      Inf
## sample estimates:
## mean of x mean of y
##  78.75758  73.67857
```

4.2.3. Interpretación del test

Se observa que el p-value obtenido es mayor que el nivel de significación 0,05, por tanto, **no se puede rechazar la hipótesis nula**, es decir, **la diferencia entre las plantillas del Betis y el Real Madrid en la temporada 2016/2017 era inferior a 5 puntos**, con una confianza del 95%.

4.3. Modelo de regresión lineal múltiple

Se desea estimar un **modelo de regresión lineal múltiple** que tenga como variables explicativas: Age, Weight, Club_Position, Vision, Ball_Control, Marking, Interceptions, Freekick_Accuracy, Short_Pass, Speed y Finishing; y como **variable dependiente el Rating de los jugadores**. Antes de realizar el modelo será necesario establecer el nivel base de referencia en la variables categórica *Club_Position*.

Para obtener un resultado más fiable, se utilizará una **validación cruzada con 20 iteraciones**.

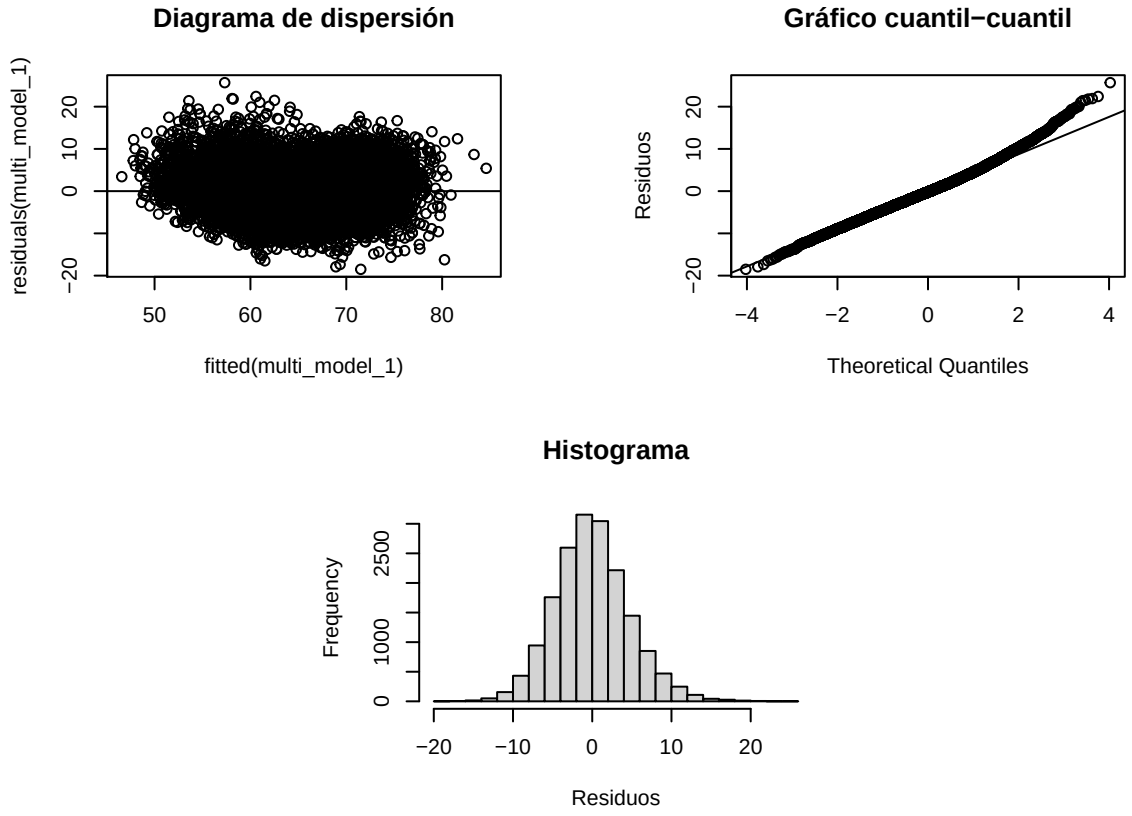
```
library(caret)
# Fijamos como valor referencia el valor mayoritarío para variable cat.
if (!is.factor(fifaNet$Club_Position)){
  max_club = max(fifaNet$Club_Position)
}

fifaNet$Club_Position = relevel(factor(fifaNet$Club_Position), ref=max_club)
# 20-fold cv
fitControl = trainControl(method = "cv", number = 20, savePredictions = T)
# Modelo lineal Multiple
multi_model_1 = train(Rating~Age+Weight+Club_Position+Vision+Ball_Control+Marking
                      +Interceptions+Freekick_Accuracy+Short_Pass+Speed+Finishing,
                      data=fifaNet, method = "lm", trControl = fitControl)
summary(multi_model_1)
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.5109  -3.1013  -0.2009   2.8667  25.6837
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       18.151078   0.590719  30.727  < 2e-16 ***
## Age                0.408174   0.008980  45.456  < 2e-16 ***
## Weight             0.232390   0.006049  38.415  < 2e-16 ***
## Club_PositionCAM   1.127050   0.277844   4.056 5.00e-05 ***
## Club_PositionCB    1.998615   0.525688   3.802 0.000144 ***
## Club_PositionCDM   0.435237   0.441119   0.987 0.323820
## Club_PositionCF   -0.548852   2.363316  -0.232 0.816356
## Club_PositionCM   -0.430434   0.536706  -0.802 0.422568
## Club_PositionGK    9.194994   0.227144  40.481  < 2e-16 ***
```

```
## Club_PositionLAM   1.646652   0.791496   2.080 0.037501 *
## Club_PositionLB    0.505129   0.215851   2.340 0.019286 *
## Club_PositionLCB   2.492531   0.206916  12.046  < 2e-16 ***
## Club_PositionLCM  -0.391225   0.261499  -1.496 0.134649
## Club_PositionLDM  -0.010549   0.299871  -0.035 0.971939
## Club_PositionLF    1.742048   1.365984   1.275 0.202218
## Club_PositionLM    0.746847   0.243029   3.073 0.002122 **
## Club_PositionLS    0.765750   0.341643   2.241 0.025014 *
## Club_PositionLW    3.144144   0.418019   7.522 5.67e-14 ***
## Club_PositionLWB  -0.233501   0.708578  -0.330 0.741755
## Club_PositionRAM   1.184103   0.791637   1.496 0.134733
## Club_PositionRB    0.484535   0.216428   2.239 0.025183 *
## Club_PositionRCB   2.749266   0.207150  13.272  < 2e-16 ***
## Club_PositionRCM  -0.687599   0.261799  -2.626 0.008636 **
## Club_PositionRDM   0.123323   0.300140   0.411 0.681161
## Club_PositionRes  -1.328059   0.104165 -12.750  < 2e-16 ***
## Club_PositionRF    2.359523   1.366237   1.727 0.084181 .
## Club_PositionRM    0.950459   0.242606   3.918 8.97e-05 ***
## Club_PositionRS    0.848001   0.340422   2.491 0.012747 *
## Club_PositionRW    2.736540   0.417351   6.557 5.65e-11 ***
## Club_PositionRWB  -0.463139   0.708291  -0.654 0.513196
## Club_PositionST    2.708894   0.248131  10.917  < 2e-16 ***
## Vision             0.113406   0.004457  25.446  < 2e-16 ***
## Ball_Control       0.105854   0.006864  15.421  < 2e-16 ***
## Marking           -0.000687   0.002791  -0.246 0.805547
## Freekick_Accuracy -0.016020   0.003537  -4.530 5.95e-06 ***
## Short_Pass         0.120790   0.006598  18.308  < 2e-16 ***
## Speed              0.050154   0.003827  13.104  < 2e-16 ***
## Finishing         -0.045555   0.004158 -10.957  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.724 on 17549 degrees of freedom
## Multiple R-squared:  0.5561, Adjusted R-squared:  0.5551
## F-statistic: 594.1 on 37 and 17549 DF,  p-value: < 2.2e-16
```

Para una mejor diagnosis del modelo, se toma el modelo construído y **se realiza un estudio de sus residuos**. Los residuos se definen como **la diferencia entre los valores observados en la muestra y los valores estimados por el modelo**. Los valores que tomen estos residuos determinarán la adecuación del modelo.

```r
# Valores ajustados vs Residuos
par(mai=rep(0.6, 4))
layout(matrix(c(1,1, 2,2, 0, 3,3, 0), ncol = 4, byrow = TRUE))
plot(fitted(multi_model_1), residuals(multi_model_1), main="Diagrama de dispersión")
abline(h=0)
# QQ plot
qqnorm(residuals(multi_model_1), ylab="Residuos", main="Gráfico cuantil-cuantil")
qqline(residuals(multi_model_1))
# Histograma para comprobar normalidad
hist(residuals(multi_model_1), xlab = "Residuos", main="Histograma")
```



4.3.1. Interpretación del modelo

Se obtiene un modelo con un **R2 de 0.56**, lo que significa que el modelo apenas explica el 50% de la variabilidad de los datos, lo cual nos dice que **no se trata de un buen modelo** y que sería necesario introducir más variables. No obstante, en el **diagrama de dispersión**, donde representamos en el 'eje x' los valores de los residuos y en el 'eje y' los valores estimados, podemos comprobar como **la nube de puntos no tiene ninguna estructura ordenada y los valores rondan el valor cero**. Esto es **indicativo de un modelo correcto**. Por otro lado, **en un modelo correcto** donde los datos sigan una distribución normal, **los valores de los residuos deben seguir también una distribución normal**. Esto lo podemos comprobar con un **gráfico cuantil-cuantil y/o un histograma**. Como podemos comprobar en ambos gráficos, **los residuos están normalmente distribuidos**.

De las variables utilizadas todas **resultan signficativas (p<0.05)** a excepción de alguna de las demarcaciones del jugador en su club. Respecto a las demás:

- La edad, el peso, la vision, el control del balón, las interecpciones, pases cortos y velocidad **influyen positivamente** en el rating.
- Algunas posiciones como ser portero influye positivamente en el rating. Esto tiene sentido, ya que no se ha introducido ninguno de los atributos propios de porteros y por tanto, a pesar de tener valores más bajos, obtienen mejores medias.
- El marcaje, la precisión de tiro libre y la finalización **influyen de manera negativa** en el rating.

4.4. Regresión logística

Utilizando diferentes variables explicativas, como *Age* o *Rating*, se ajusta un **modelo predictivo basado en regresión logística** para predecir la **probabilidad de que un jugador vaya a la selección** de su país. Para que la muestra sea más homogénea, este modelo se realizará solo con **jugadores españoles** (Nationality == Spain).

Primero de todo, generaremos y usaremos la **variable dicotómica 'internacional'**. Esta variable tomará el valor **1 si el jugador forma parte de la selección** de su país, y 0 en el caso contrario.

```
fifaNet$internacional[is.na(fifaNet$National_Kit)] = 0
fifaNet$internacional[!is.na(fifaNet$National_Kit)] = 1
```

Para conseguir un mejor modelo, generaremos **otra variable dicotómica, esta vez llamada 'portero'**, para diferenciar los jugadores de campo ('Jugador') y los porteros ('Portero').

```
fifaNet$portero[fifaNet$Club_Position=="GK"] = "Portero"
fifaNet$portero[fifaNet$Club_Position!="GK"] = "Jugador"
fifaNet$portero = as.factor(fifaNet$portero)
```

Filtramos el dataset original (fifaNet) para generar un segundo dataset solo con **jugadores españoles (fifaNet_spain)**. Será con este set de datos con el que se ajuste el modelo predictivo.

```
fifaNet_spain = fifaNet[fifaNet$Nationality == "Spain", ]
```

Se realiza una 20-fold cross-validation mediante la función **trainControl()** del paquete *caret*.

```
# 20-fold cv
fitControl = trainControl(method = "cv", number = 20, savePredictions = T)
# Ajuste del modelo (regresión logistica)
model_fit = train(internacional~Age+Rating+portero+Work_Rate,data=fifaNet_spain,
                  method = "glm", family="binomial", trControl = fitControl)

# Valores predichos y observados
pred_prob = (model_fit$pred)$pred
pred_bin = factor(ifelse(pred_prob >= 0.5, 1, 0))
obs_bin = factor((model_fit$pred)$obs)

# Matriz de confusión
confusionMatrix(table(pred_bin,obs_bin), positive = "1")
```

```
## Confusion Matrix and Statistics
##
##          obs_bin
## pred_bin   0   1
##        0 979   8
##        1   6  15
##
##                Accuracy : 0.9861
##                  95% CI : (0.9768, 0.9924)
```

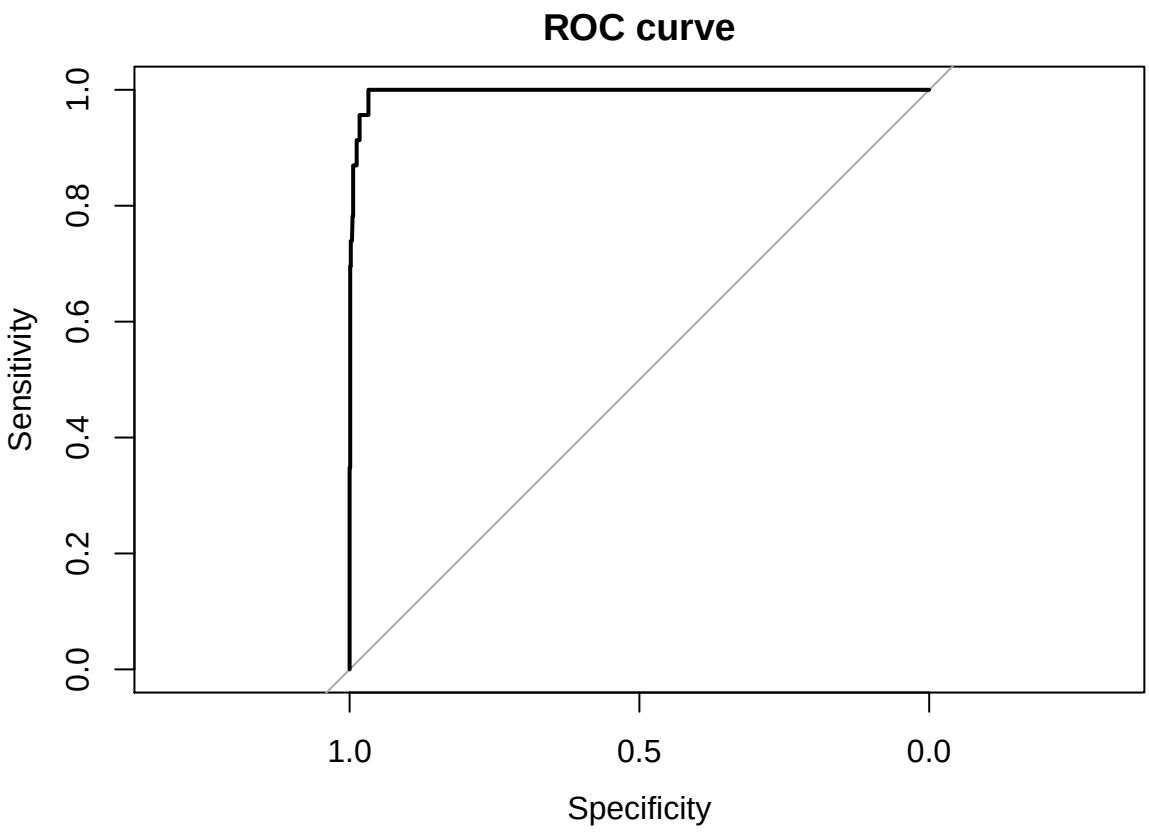
```
##     No Information Rate : 0.9772
##     P-Value [Acc > NIR] : 0.02968
##
##                   Kappa : 0.6747
##
##  Mcnemar's Test P-Value : 0.78927
##
##             Sensitivity : 0.65217
##             Specificity : 0.99391
##          Pos Pred Value : 0.71429
##          Neg Pred Value : 0.99189
##              Prevalence : 0.02282
##          Detection Rate : 0.01488
##    Detection Prevalence : 0.02083
##       Balanced Accuracy : 0.82304
##
##        'Positive' Class : 1
##
```

4.4.1. Interpretación del modelo

Consideramos que hemos conseguido **un buen modelo para predecir si un jugador español es convocado o no por la selección**.

Se observa una estimación de la **precisión del modelo en torno al 99%**, lo cual es una precisión muy buena. Por otra parte, si se analiza la **sensibilidad, se observa que es del 74%**. Aunque está lejos del valor que toma la precisión, la posibilidad de acertar si un jugador seleccionado por su selección es aceptable. Por último, la **especificidad es casi del 100%**, es decir, la posibilidad de detectar realmente los casos en los que no irá a la selección es casi perfecta. Esto se visualiza mediante una **curva ROC**.

```
curve = roc(response=factor(fifaNet_spain$internacional),
            predictor=as.numeric(predict(model_fit)))
plot.roc(curve, main="ROC curve")
```



5 Conclusiones

Las conclusiones obtenidas a partir del análisis de los datos de jugadores de fútbol en el FIFA17 son: * El contraste de hipótesis no nos ha permitido afirmar que existe una diferencia de más de 5 puntos entre los jugadores del Real Madrid y el Real Betis.

- Se obtenido un modelo de regresión lineal para la valoración media de cada jugador que a pesar de ser correcto, no ha mostrado una gran precisión, probablemente debido a que se ha mezclado jugadores de campo con porteros, y eso complica la obtención de la media de cada jugador.
- Se obtenido un modelo de regresión logística para determinar si un jugador tiene posibilidades o no de ir a la selección española. El modelo obtenido presenta una precisión de casi el 99% con una sensibilidad del 74% y una especificidad casi del 100%.

6 Contribuciones

Contribuciones	Firma
Investigación previa	PRNV; AVG
Redacción de las respuestas	PRNV; AVG
Desarrrollo código	PRNV; AVG

7 Fichero final

Se exporta el dataframe limpio a un fichero csv.

```
write.csv(fifaNet,"../data/fifaNet.csv", col.names = T,row.names=F)
write.csv(fifaNet_spain,"../data/fifaNetSpain.csv", col.names = T,row.names=F)
```