PRÁCTICA 2

Arturo Hernández Sánchez y Laia Cebey Ripoll

1 Presentación de la actividad

En esta práctica se elabora un caso práctico orientado a aprender a identificar los datos relevantes para un proyecto analítico y usar las herramientas de integración, limpieza, validación y análisis de las mismas.

Los objetivos concretos de esta práctica son:

- Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinares.
- Saber identificar los datos relevantes y los tratamientos necesarios (integración, limpieza y validación) para llevar a cabo un proyecto analítico.
- Aprender a analizar los datos adecuadamente para abordar la información contenida en los datos.
- Identificar la mejor representación de los resultados para aportar conclusiones sobre el problema planteado en el proceso analítico.
- Actuar con los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.
- Desarrollar las habilidades de aprendizaje que les permitan continuar estudiando de un modo que tendrá que ser en gran medida autodirigido o autónomo.
- Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

2 Resolución de la actividad

Para esta actividad se utilizará el dataset que se puede encontrar en la plataforma kaggle, concretamente en el enlace https://www.kaggle.com/artimous/complete-fifa-2017-player-dataset-global/version/5.

3 Descripción del dataset

El dataset escogido contiene información de el estilo del videojuego de consola Fifa 2017, así como estadísticas reales de los jugadores de futbol. El conjunto de datos contiene más de 17500 registros y 53 variables.

Las principales variables que se usarán en esta actividad son:

- Name (Nombre del jugador)
- Nationality (Nacionalidad del jugador)
- 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)
- Work_Rate (valoración cualitativa en términos de ataque-defensa)

- Ball_Control

La descripción de los atributos se puede consultar en https://www.fifplay.com/encyclopedia. La descripción de las abreviaturas de la posición del jugador en el campo se puede consultar en https://www.dtgre.com/2016/10/fifa-17-position-abbreviations-acronyms.html.

4 Integración y selección de los datos de interés a analizar

Empezamos cargando los datos y seleccionando las columnas que nos interesan.

```
datos <- read.csv('FullData.csv', encoding='UTF-8')
print(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
```

```
datos<- datos[,c('Name', 'Nationality', 'Club_Position', 'Club_Joining','Contract_Expiry',
                 'Rating', 'Height', 'Weight', 'Preffered_Foot','Birth_Date', 'Age',
                 'Work_Rate', 'Ball_Control')]
str(datos)
```

```
## 'data.frame':    17588 obs. of  13 variables:
##  $ Name           : chr  "Cristiano Ronaldo" "Lionel Messi" "Neymar" "Luis Suárez" ...
##  $ Nationality    : chr  "Portugal" "Argentina" "Brazil" "Uruguay" ...
##  $ Club_Position  : chr  "LW" "RW" "LW" "ST" ...
##  $ 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 ...
##  $ Work_Rate      : chr  "High / Low" "Medium / Medium" "High / Medium" "High / Medium" ...
##  $ Ball_Control   : int  93 95 95 91 48 31 87 88 90 23 ...
```

Vemos que tanto las fechas como los campos Height y Weight se han interpretado como carácteres. También hay otros campos que se tienen que podrían poner como factores en vez de únicamente como carácteres como Work_Rate y Preferred_Foot. Los cambios y limpieza de estos campos la haremos en apartados posteriores.

5 Limpieza de los datos

A continuación, vamos a limpiar los datos para poderlos analizar posteriormente.

5.1 Análisis de duplicados

Miramos con la función duplicated que no haya ningún registro que tenga todos los campos iguales.

```
any(duplicated(datos[,c('Name', 'Nationality', 'Club_Joining','Contract_Expiry',
                        'Rating', 'Height', 'Weight', 'Preffered_Foot',
                        'Birth_Date', 'Age', 'Work_Rate' )]))
```

```
## [1] FALSE
```

Vemos que no hay ninguno. Por último, vamos a hacer la comprovación con menos campos. Vamos a considerar que si el nombre, la nacionalidad y la fecha de nacimiento son iguales se trata de un duplicado.

```
any(duplicated(datos[,c('Name', 'Nationality', 'Birth_Date')]))
```

```
## [1] FALSE
```

Vemos que, de nuevo, no encontramos ningún duplicado.

5.2 Normalización de los datos cuantitativos

5.2.1 Rating

El tipo de esta variable ya está bien cargada con el tipo int. Vamos a ver que los valores estén entre 0 y 100.

```
print(min(datos$Rating))
```

```
## [1] 45
```

```
print(max(datos$Rating))
```

```
## [1] 94
```

Vemos que efectivamente, los valores se encuentran entre 0 y 100.

5.2.2 Height

Vemos que todos los registros están en cm.

```
length(datos$Height[str_detect(datos$Height, 'cm') & !is.na(datos$Height)])
```

```
## [1] 17588
```

A continuación eliminamos las unidades y convertimos al tipo numeric el resultado. Por último, cambiamos el tipo de la columna a integer porque no tenemos decimales.

```
# Eliminamos las comas y las reemplazamos con puntos
datos$Height <- str_replace(datos$Height, ',', '.')

# Eliminamos cm de los registros en cm y convertimos a numeric
datos$Altura[str_detect(datos$Height, 'cm') & !is.na(datos$Height)]<-
  as.numeric(str_replace(datos$Height[str_detect(datos$Height, 'cm')
                                      & !is.na(datos$Height)], ' cm', ''))
# Por último, substituimos la columna height por la columna nueva
datos$Height <- as.integer(datos$Altura)
datos$Altura <- NULL
```

5.2.3 Weight

Vemos que todos los registros están en kg.

```
length(datos$Weight[str_detect(datos$Weight, ' kg') & !is.na(datos$Weight)])
```

```
## [1] 17588
```

A continuación eliminamos las unidades y convertimos al tipo numeric el resultado. Por último, cambiamos el tipo de la columna a integer porque no tenemos decimales.

```
# Eliminamos las comas y las reemplazamos con puntos
datos$Weight <- str_replace(datos$Weight, ',', '.')

# Eliminamos kg de los registros en kg y los convertimos a numeric
datos$Peso[str_detect(datos$Weight, ' kg') & !is.na(datos$Weight)] <-
  as.numeric( str_replace(datos$Weight[str_detect(datos$Weight, ' kg')
                                       &  !is.na(datos$Weight) ], ' kg', ''))

# Por último, substituimos la columna weight por la columna
#nueva convertida a entero como se especifica en el enunciado
datos$Weight <- as.integer(datos$Peso)
datos$Peso <- NULL
```

5.3 Normalización de los datos cualitativos

5.3.1 Name y Nationality

Para estas dos columnas eliminamos los espacios en blanco antes y después de su valor (con la función str_trim) y ponemos la primera letra de cada palabra en mayúsculas (con la función str_to_title).

```
datos$Name <- str_to_title(str_trim(datos$Name, side='both'))
datos$Nationality <-  str_to_title(str_trim(datos$Nationality, side='both'))
```

5.3.2 Preferred_Foot

Cambiamos los registros con valor 1 a Left y los registros con valor 2 a Right. Luego convertimos la variable a un factor ya que se trata de un atributo categórico nominal.

```
datos$Preffered_Foot[datos$Preffered_Foot==1] <- 'Left'
datos$Preffered_Foot[datos$Preffered_Foot==2] <- 'Right'
datos$Preffered_Foot <- as.factor(datos$Preffered_Foot)
```

5.3.3 Work_Rate

Empezamos mirando qué valores toma esta variable.

```
print(unique(datos$Work_Rate))
```

```
## [1] "High / Low"      "Medium / Medium" "High / Medium"   "Medium / Low"
## [5] "High / High"     "Medium / High"   "Low / High"      "Low / Medium"
## [9] "Low / Low"
```

Reemplazamos las categorías cortadas con tres letras.

```
datos$Work_Rate <-str_replace(datos$Work_Rate, 'Hig /','High /')
datos$Work_Rate <-str_replace(datos$Work_Rate, 'Med /','Medium /')

datos$Work_Rate <-str_replace(datos$Work_Rate, '/ Hig$','/ High')
datos$Work_Rate <-str_replace(datos$Work_Rate, '/ Med$','/ Medium')

print(unique(datos$Work_Rate))
```

```
## [1] "High / Low"      "Medium / Medium" "High / Medium"   "Medium / Low"
## [5] "High / High"     "Medium / High"   "Low / High"      "Low / Medium"
## [9] "Low / Low"
```

Una vez disponemos de esta información, cambiamos esta variable al tipo ordered de R ya que se trata de un atributo categórico nominal. Se podría argumentar que es un atributo categórico ordinal ya que low es más bajo que medium y medium es más bajo que high. Sin embargo, no se puede saber si por ejemplo low/high es más alto o más bajo que medium/medium. Es por esta razón que se convertirá al tipo de R factor.

```
datos$Work_Rate <-as.factor(datos$Work_Rate)
```

5.4 Posibles inconsistencias y variables tipo fecha

Empezamos cambiando el tipo de las columnas Club_Joining y Birth_Date a fecha.

```
datos$Club_Joining <- as.Date(datos$Club_Joining, "%m/%d/%Y")
datos$Birth_Date <- as.Date(datos$Birth_Date, "%m/%d/%Y")
```

5.4.1 Club_Joining

Para la fecha Club_Joining tenemos que comprovar que está en los rango de 1990 a 2017.

```
print(min(datos$Club_Joining,na.rm = TRUE))
```

```
## [1] "1991-06-01"
```

```
print(max(datos$Club_Joining, na.rm = TRUE))
```

```
## [1] "2017-03-24"
```

Vemos que efectivamente los datos están en el rango correcto.

5.4.2 Contract_Expiry >= Club_Joining?

Comprobamos que no haya registros con contract expiry < club joining.

```
datos[datos$Contract_Expiry < as.integer(format(datos$Club_Joining,"%Y")),]
```

```
##    Name Nationality Club_Position Club_Joining Contract_Expiry Rating Height
## NA <NA>        <NA>          <NA>         <NA>              NA     NA     NA
##    Weight Preffered_Foot Birth_Date Age Work_Rate Ball_Control
## NA     NA           <NA>       <NA>  NA      <NA>           NA
```

5.4.3 Revisar si la edad corresponde a la fecha de nacimiento

```
datos$edades <- as.integer(floor(time_length(as.Date('01/01/2017', "%m/%d/%Y")
                                             - datos$Birth_Date, "years")))
# Vemos para qué registros no coinciden
print(sum(datos$Age != datos$edades))
```

```
## [1] 5561
```

```
# Ponemos el valor correcto en la columna Age
datos$Age <- datos$edades
# Eliminamos la columna edades
datos$edades <- NULL
```

Como vemos que en varios registros la edad no corresponde con la edad calculada con la fecha de nacimiento, así que la corregimos.

5.5 Identificación y tratamiento de ceros o elementos vacíos

```
datos[rowSums(is.na(datos)) > 0, ]
```

```
##              Name Nationality Club_Position Club_Joining Contract_Expiry Rating
## 384 Didier Drogba Ivory Coast                       <NA>              NA     81
##     Height Weight Preffered_Foot Birth_Date Age    Work_Rate Ball_Control
## 384    189     80          Right 1978-03-11  38 Medium / Low           80
```

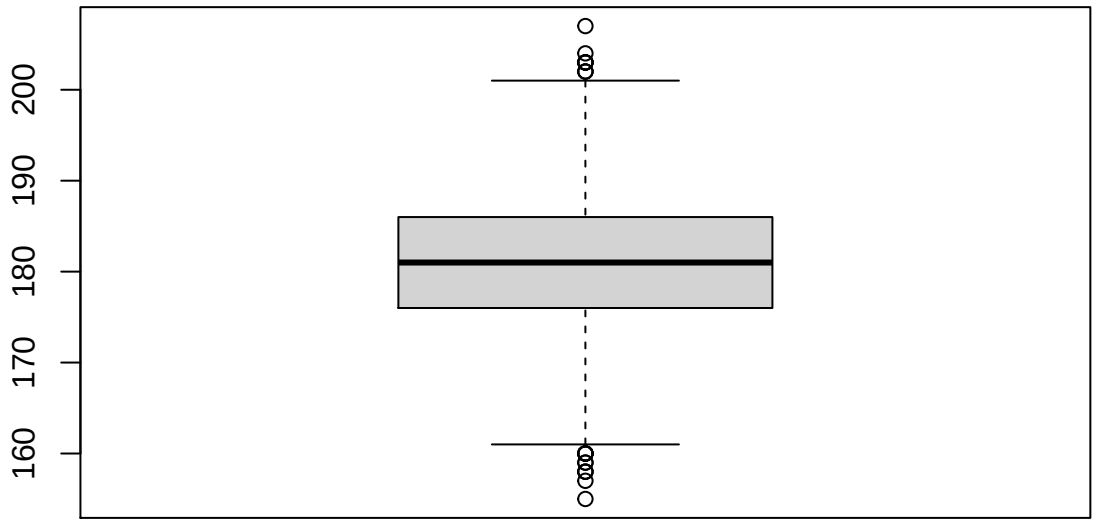
Vemos que únicamente hay una fila con valores vacíos, concretamente faltan las fechas de Club_Joining y Contract_Expiry. Buscando más información, vemos que es debido a que en 2017 no estaba en ningún club porque se retiró en 2014. Por lo tanto, no tiene sentido que rellenemos estos valores.

Si por ejemplo tuviésemos algun valor de altura o de peso vacío podríamos imputar los valores con una regresión lineal porque sabemos que estas dos variables están muy relacionadas.

5.6 Identificación y tratamiento de valores extremos

Empezamos con los valores de altura. Utilizamos el boxplot para ver los valores considerados atípicos.

```
boxplot(datos$Height)
```



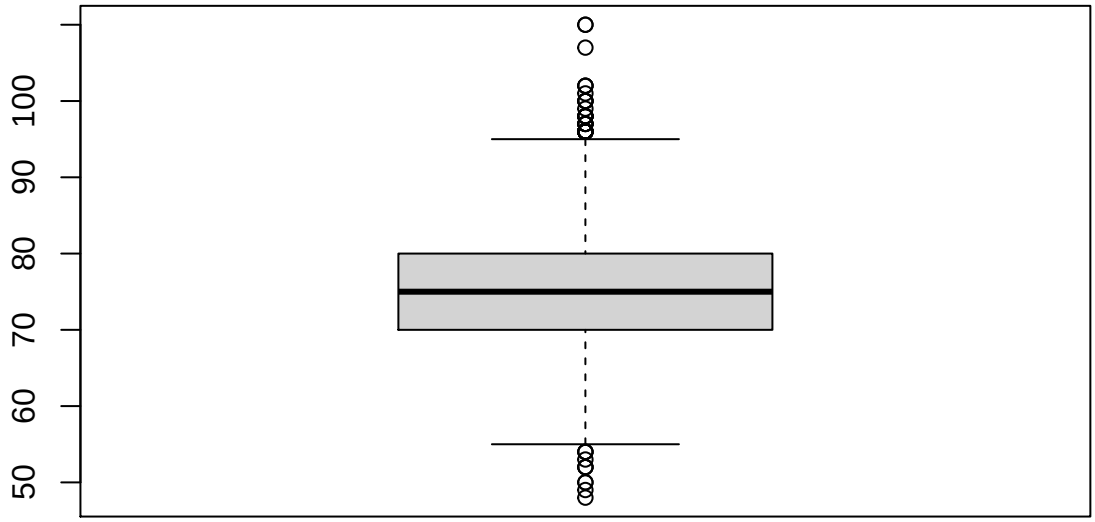

```
sort(boxplot.stats(datos$Height)$out)
```

```
##  [1] 155 157 158 158 158 159 159 160 160 160 160 160 160 202 202 202 203 203 203
## [20] 203 203 204 207
```

Podemos ver que los valores que están fuera del rango interquartílico no se alejan mucho de él y además tienen valores razonables de altura para un jugador de fútbol. Por esta razón, no los consideramos atípicos.

A continuación, repetimos el análisis para el peso.

```
boxplot(datos$Weight)
```




```
sort(boxplot.stats(datos$Weight)$out)
```

```
##  [1]  48  49  50  50  52  52  52  53  54  54  54  96  96  96  96  96  96  96  96
## [20]  96  96  96  96  96  96  96  96  96  96  96  96  96  97  97  97  97  97  98
## [39]  98  98  98  98  98  99 100 100 100 100 101 102 102 102 107 110 110
```

En este caso tenemos más valores fuera del rango intercuartil pero de nuevo no se alejan mucho de éste y tienen valores razonables para un jugador de fútbol.

5.7 Estudio descriptivo de las variables cuantitativas

Por último, hacemos un estudio descriptivo de las variables cuantitativas, que son Rating, Height, Weight y Age. Las medidas de tendencia central que vamos a analizar son la media, la mediana y la moda y las medidas de dispersión,la variana, la desviación estándar, los cuartiles, la simetría y la curtosis.

```
dfEst <- data.frame()
for (col in c("Rating", "Height", "Weight", "Age")){
  min <- min(datos[, col], na.rm = TRUE)
  q1 <- quantile(datos[, col], probs = 0.25, na.rm = TRUE)
  media <- mean.default(datos[, col], na.rm = TRUE)
  mediana <- median.default(datos[, col], na.rm = TRUE)
  moda <- mfv(datos[, col])
  var <- var(datos[, col], na.rm = TRUE)
  desvest <- sd(datos[, col], na.rm = TRUE)
```

```
  q3 <- quantile(datos[, col], probs = 0.75, na.rm = TRUE)
  max <- max(datos[, col], na.rm = TRUE)
  s <- skew(datos[, col])
  c <- kurtosi(datos[, col])
  dfEst <- rbind(dfEst,data.frame( "Mínimo"=min, "Q1"=q1, "Media"=media,
                                   "Mediana"=mediana, "Moda"=moda,"Varianza"=var,
                                   "Desviación Estándar"=desvest, "Q3"=q3,
                                   "Máximo"=max, "Simetría"=s, "Curtosis"=c) )
}
rownames(dfEst) <- c("Rating", "Height", "Weight", "Age")
print(dfEst)
```

```
##        Mínimo  Q1     Media Mediana Moda Varianza Desviación.Estándar  Q3
## Rating     45  62  66.16619      66   67 50.16906            7.083012  71
## Height    155 176 181.10547     181  180 44.55776            6.675160 186
## Weight     48  70  75.25335      75   75 47.58168            6.897948  80
## Age        16  21  25.14413      25   24 21.86727            4.676245  28
##        Máximo    Simetría    Curtosis
## Rating     94 -0.01738810 -0.02853432
## Height    207 -0.02983511 -0.26018505
## Weight    110  0.20701786  0.08613853
## Age        47  0.41933672 -0.42257485
```

En cuanto a las medidas de tendencia central, se puede observar que tanto la media, como la mediana y la moda son muy parecidas en los cuatro atributos de forma que todas son representativas de los atributos.

Un valor de simetría negativo significa que la mayoría de datos son menores que la media, este es el caso de Rating y Height. Contrariamente, un valor positivo de la simetría significa que la mayoría de casos son mayores que la media, que sería el caso de Weight y Age. Para todos los atributos, el valor está muy cerca del 0, lo que significa que están repartidos de manera basante igual a ambos lados de la media.

El valor de la desviación estándar nos sirve para ver cómo de alejados están los puntos de su media. Vemos que el atributo que tiene una desviación estándar más bajo, teniendo en cuenta su media, es Height.

Por último, el valor de curtosis nos indica cómo de concentrados están los valores alrededor de su media. como mayor sea, más cerca de la media se encuentran y como menor, más alejados. Se considera que un valor > 0 una distribución leptocúrtica, un valor $= 0$ una distribución normal y valor < 0 una distribución platicúrtica. En estos casos tenemos valores muy cercanos a 0, por lo que se puede considerar que todos los atributos siguen una distribución normal.

5.8 Archivo datos limpios

Como último paso de la limpieza, guardamos los datos en un archivo csv llamado fifa_clean.csv.

```
write.csv(datos,'fifa_clean.csv')
```

6 Análisis de los datos

6.1 Selección de los grupos que se quieren analizar/comparar

Vamos a comenzar obteniendo las muestras que utilizaremos posteriormente para realizar el contraste correspondiente que dará respuesta la pregunta de investigación que nos plantearemos. El objetivo será obtener una muestra para los jugadores zurdos y otra para los diestros:

```
# Jugadores que no son porteros
datos_filtered <- datos[datos$Club_Position!='GK',]
```

```r
# Obtenemos la muestra para diestros y zurdos:
Left <- datos_filtered[datos_filtered$Preffered_Foot=='Left',]
Right <- datos_filtered[datos_filtered$Preffered_Foot=='Right',]

# Seleccionamos las variables de interés para el contraste:
Left_R <- Left$Rating
Right_R <- Right$Rating

Left_BC <- Left$Ball_Control
Right_BC <- Right$Ball_Control
```

6.2 Comprobación de la normalidad y homogeneidad de la varianza

Comencemos ahora por el contraste de normalidad. Las hipótesis del contraste que realizaremos para cada una de las variables de interés son las siguientes:

$$H_0 : \text{La muestra obtenida proviene de una población que sigue una distribución normal}$$

$$H_1 : \text{La muestra obtenida no proviene de una población que sigue una distribución normal}$$

Realizaremos el contraste de normalidad de Lilliefors:

```r
library(nortest)
lillie.test(Right_BC)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  Right_BC
## D = 0.14621, p-value < 2.2e-16
```

```r
lillie.test(Left_BC)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  Left_BC
## D = 0.11186, p-value < 2.2e-16
```

```r
lillie.test(Right_R)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  Right_R
## D = 0.045066, p-value < 2.2e-16
```

```r
lillie.test(Left_R)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  Left_R
## D = 0.039525, p-value = 8.249e-16
```

Como podemos observar, en todos los casos optenemos un p-valor$< 0.05 = \alpha$, por lo que, para cada uno de los contrastes realizados en cada una de las muestras, rechazamos H_0 con un nivel de significación $\alpha = 0.05$ y concluimos que los datos de todas las muestras no provienen de una distribución normal.

Sin embargo, como el tamaño de las muestras es grande, por el teorema central del límite, podemos asumir que las muestras provienen de una población normal, por lo que podemos aplicar inferencia paramétrica para muestras grandes.

Realicemos ahora un test de homocedasticidad sobre cada par de muestras que consideraremos posteriormente en el contraste para así poder decidir si elegimos un contraste para varianzas desconocidas iguales o diferentes.

Como se ha indicado anteriormente, podemos asumir que las muestras provienen de una población normal, por lo que podemos utilizar el test de homogeneidad *F-test*. Las hipótesis para cada uno de los contrastes realizados serán las siguientes:

$$H_0 : \text{Ambas muestras provienen de poblaciones con misma varianza}$$

$$H_1 : \text{Las muestras provienen de poblaciones con distinta varianza}$$

```
var.test(Right_BC, Left_BC)
```

```
##
##  F test to compare two variances
##
## data:  Right_BC and Left_BC
## F = 1.6922, num df = 12933, denom df = 4021, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.609046 1.778459
## sample estimates:
## ratio of variances
##           1.692198
```

```
var.test(Right_R, Left_R)
```

```
##
##  F test to compare two variances
##
## data:  Right_R and Left_R
## F = 1.1825, num df = 12933, denom df = 4021, p-value = 1.037e-10
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.124363 1.242746
## sample estimates:
## ratio of variances
##           1.182468
```

En ambos contrastes obtenemos un p-valor$< 0.05 = \alpha$, por lo que rechazamos la hipótesis nula H_0 con un nivel de significación $\alpha = 0.05$ y concluimos que, en ambos casos, los pares de muestras considerados provienen de poblaciones con varianza distinta.

Vamos ahora a realizar el test de homogeneidad de Fligner-Killen. A diferencia del anterior, se trata de un test no paramétrico, el cual compara las varianzas basándose en la mediana. Este contraste de homogeneidad es más adecuado cuando no se cumple que las muestras provengan de una población normal. Las hipótesis de contraste serán las mismas que para el caso del *F-test*:

```
fligner.test(list(Right_BC, Left_BC))
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  list(Right_BC, Left_BC)
```

```
## Fligner-Killeen:med chi-squared = 124.74, df = 1, p-value < 2.2e-16
```

```
fligner.test(list(Right_R, Left_R))
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  list(Right_R, Left_R)
## Fligner-Killeen:med chi-squared = 37.411, df = 1, p-value = 9.568e-10
```

En los dos contrastes obtenemos un p-valor$< 0.05 = \alpha$, por lo que, rechazamos la hipótesis nula H_0 con un nivel de significación $\alpha = 0.05$ y de nuevo concluimos que, en ambos casos, los pares de muestras considerados provienen de poblaciones con varianza distinta.

6.3 Aplicación de pruebas estadísticas para comparar los grupos de datos

6.3.1 Comparación de jugadores diestros y zurdos

A partir de las muestras obtenidas en el apartado anteiror para diestros y zurdos, vamos a realizar una serie contrastes que nos permitirán ver si los jugadores zurdos tienen mejor Ball_Control o Rating que los diestros.

6.3.1.1 ¿Los jugadores zurdos tienen mejor Ball_Control que los diestros? Tenemos que realizar un contraste sobre las muestras *Right_BC* y *Left_BC*. Sean $\mu_{\text{Right}_{\text{BC}}}$ y $\mu_{\text{Left}_{\text{BC}}}$ las medias poblacionales correspondientes a las poblaciones asociadas a las muestras *Right_BC* y *Left_BC* respectivamente. Entonces, las hipótesis del contraste a realizar con las siguientes:

$$H_0 : \mu_{\text{Left}_{\text{BC}}} = \mu_{\text{Right}_{\text{BC}}}$$

$$H_1 : \mu_{\text{Left}_{\text{BC}}} > \mu_{\text{Right}_{\text{BC}}}$$

Realizamos por tanto un contraste unilateral de dos muestras independientes sobre la media con varianzas poblacionales desconocidas no iguales, por lo que el estadístico de contraste es el siguiente:

$$t = \frac{\bar{x}_{\text{Left}_{\text{BC}}} - \bar{x}_{\text{Right}_{\text{BC}}}}{\sqrt{\frac{s^2_{\text{Left}_{\text{BC}}}}{n_{\text{Left}_{\text{BC}}}} + \frac{s^2_{\text{Right}_{\text{BC}}}}{n_{\text{Right}_{\text{BC}}}}}}$$

Donde $\bar{x}_{\text{Left}_{\text{BC}}}$ y $\bar{x}_{\text{Right}_{\text{BC}}}$ son las medias muestrales, $s^2_{\text{Left}_{\text{BC}}}$ y $s^2_{\text{Right}_{\text{BC}}}$ son las cuasivarianzas y $n_{\text{Left}_{\text{BC}}}$ y $n_{\text{Right}_{\text{BC}}}$ son los tamaños de las muestras de *Ball_Control* para los zurdos y diestros respectivamente.

Este estadístico sigue una distribución t de Student con υ grados de libertad, donde

$$\upsilon = \frac{\left(\frac{s^2_{\text{Left}_{\text{BC}}}}{n_{\text{Left}_{\text{BC}}}} + \frac{s^2_{\text{Right}_{\text{BC}}}}{n_{\text{Right}_{\text{BC}}}}\right)^2}{\frac{(s^2_{\text{Left}_{\text{BC}}}/n_{\text{Left}_{\text{BC}}})^2}{n_{\text{Left}_{\text{BC}}} - 1} + \frac{(s^2_{\text{Right}_{\text{BC}}}/n_{\text{Right}_{\text{BC}}})^2}{n_{\text{Right}_{\text{BC}}} - 1}}$$

Al ser un test unilateral por la derecha, la zona de aceptación de la hipótesis nula estará en el intervalo $(-\infty, t_{\upsilon,1-\alpha})$, donde α es el nivel de significación del contraste.

```
t.test(Left_BC, Right_BC, alternative="greater", var.equal=FALSE, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  Left_BC and Right_BC
## t = 15.182, df = 8623.8, p-value < 2.2e-16
```

```
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  3.286679      Inf
## sample estimates:
## mean of x mean of y
##  62.16335  58.47727
```

Como podemos observar, obtenemos un p-valor$< 0.05 = \alpha$, por lo que rechazamos la hipótesis nula con un nivel de significación $\alpha = 0.05$, y concluimos que los jugadores zurdos tienen mejor Ball_Control que los diestros.

6.3.1.2 ¿Los jugadores zurdos tienen mejor rating que los diestros?

Tenemos que realizar un contraste sobre las muestras *Right_R* y *Left_R*. Sean $\mu_{\text{Right}_\text{R}}$ y $\mu_{\text{Left}_\text{R}}$ las medias poblacionales correspondientes a las poblaciones asociadas a las muestras *Right_R* y *Left_R* respectivamente. Entonces, las hipótesis del contraste a realizar con las siguientes:

$$H_0 : \mu_{\text{Left}_\text{R}} = \mu_{\text{Right}_\text{R}}$$

$$H_1 : \mu_{\text{Left}_\text{R}} > \mu_{\text{Right}_\text{R}}$$

Realizamos por tanto un contraste unilateral de dos muestras independientes sobre la media con varianzas poblacionales desconocidas no iguales, por lo que el estadístico de contraste es el siguiente:

$$t = \frac{\bar{x}_{\text{Left}_\text{R}} - \bar{x}_{\text{Right}_\text{R}}}{\sqrt{\frac{s^2_{\text{Left}_\text{R}}}{n_{\text{Left}_\text{R}}} + \frac{s^2_{\text{Right}_\text{R}}}{n_{\text{Right}_\text{R}}}}}$$

Donde $\bar{x}_{\text{Left}_\text{R}}$ y $\bar{x}_{\text{Right}_\text{R}}$ son las medias muestrales, $s^2_{\text{Left}_\text{BC}}$ y $s^2_{\text{Right}_\text{BC}}$ son las cuasivarianzas y n_{Left_R} y $n_{\text{Right}_\text{R}}$ son los tamaños de las muestras de *Ball_Control* para los zurdos y diestros respectuvamente.

Este estadístico sigue una distribución t de Student con υ grados de libertad, donde

$$\upsilon = \frac{\left(\frac{s^2_{\text{Left}_\text{R}}}{n_{\text{Left}_\text{R}}} + \frac{s^2_{\text{Right}_\text{R}}}{n_{\text{Right}_\text{R}}}\right)^2}{\frac{(s^2_{\text{Left}_\text{R}}/n_{\text{Left}_\text{R}})^2}{n_{\text{Left}_\text{R}} - 1} + \frac{(s^2_{\text{Right}_\text{R}}/n_{\text{Right}_\text{R}})^2}{n_{\text{Right}_\text{R}} - 1}}$$

Al ser un test unilateral por la derecha, la zona de aceptación de la hipótesis nula estará en el intervalo $(-\infty, t_{\upsilon,1-\alpha})$, donde α es el nivel de significación del contraste.

```
t.test(Left_R, Right_R, alternative="greater", var.equal=FALSE, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  Left_R and Right_R
## t = 5.9338, df = 7218.3, p-value = 1.549e-09
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.5228087       Inf
## sample estimates:
## mean of x mean of y
##  66.58155  65.85820
```

Como podemos observar, obtenemos un p-valor$< 0.05 = \alpha$, por lo que rechazamos la hipótesis nula con un nivel de significación $\alpha = 0.05$, y concluimos que los jugadores zurdos tienen mejor Rating que los diestros.

6.3.2 ¿Qué variables cuantitativas influyen más en el *Rating* del jugador?

Calculamos la matriz de correlación de las variables cuantitativas a utilizar:

```
cor(datos%>%select('Rating', 'Ball_Control', 'Height', 'Weight', 'Age'))
```

```
##                  Rating Ball_Control      Height     Weight        Age
## Rating       1.00000000   0.46328646  0.04707022  0.1397657 0.45603766
## Ball_Control 0.46328646   1.00000000 -0.40247272 -0.3383873 0.08173935
## Height       0.04707022  -0.40247272  1.00000000  0.7582135 0.07647673
## Weight       0.13976567  -0.33838729  0.75821349  1.0000000 0.22225431
## Age          0.45603766   0.08173935  0.07647673  0.2222543 1.00000000
```

Como podemos observar, las variables que presentan una mayor correlación con la variable *Rating* son las variables *Ball_Control* y *Age*, ambas con correlación positiva, aunque no presentan una correlación especialmente fuerte.

Las correlaciones mas bajas vemos que se presentan con las variables *Height* y *Weight*.

No observamos correlación negativa con ninguna de las variables consideradas respecto a la variable *Rating*.

6.3.3 Modelo de regresión lineal

Vamos ahora a ajustar distintos modelos de regresión lineal con la finalidad de poder predecir el valor de la variable *Rating*. Tomaremos distintas variables explicaticas, para así poder discutir cuál podría ser el mejor modelo.

Para las construcción del modelo, vamos a considerar como variables independientes, *Height, Weight, Age, Ball_Control*, y como variable dependiente, *Rating*. Comencemos construyendo el modelo de regresión lineal múltiple considerando todas las variables explicativas:

```
modelo_1 <- lm(datos$Rating~datos$Height+datos$Weight+datos$Age+datos$Ball_Control)
summary(modelo_1)
```

```
##
## Call:
## lm(formula = datos$Rating ~ datos$Height + datos$Weight + datos$Age +
##     datos$Ball_Control)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.189  -3.568  -0.361   3.183  28.658
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        4.163997   1.386390   3.003  0.00267 **
## datos$Height       0.131763   0.009501  13.868  < 2e-16 ***
## datos$Weight       0.146673   0.009213  15.920  < 2e-16 ***
## datos$Age          0.562507   0.008941  62.911  < 2e-16 ***
## datos$Ball_Control 0.223514   0.002621  85.264  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.295 on 17583 degrees of freedom
## Multiple R-squared:  0.4413, Adjusted R-squared:  0.4412
## F-statistic:  3472 on 4 and 17583 DF,  p-value: < 2.2e-16
```

Como podemos observar, el coeficiente de determinación que obtenemos es 0.4413, lo cual indica que el modelo puede explicar el 44.13% de la variabilidad de los datos.

Respecto al contraste para cada cada uno de los coeficientes correspondientes a cada una de las variables del modelo, en todos los casos obtenemos un $p - valor < 0.05$, por lo que para cada uno de los coeficientes β_i, con $i \in \{Height, Weight, Age, BallControl, Intercept\}$ de la ecuación del modelo, si realizamos el contraste:

$$H_0 : \beta_i = 0$$

$$H_1 : \beta_i \neq 0$$

Rechazamos la H_0 con un nivel de significación $\alpha = 0.05$ y concluimos que el contraste es significativo para cada β_i con un nivel de significación $\alpha = 0.05$.

Si realizamos ahora el siguiente contraste:

$$H_0 : \beta_{Intercept} = \beta_{Height} = \beta_{Weight} = \beta_{Age} = \beta_{BallControl} = 0$$

$$H_1 : \exists i \in \{Height, Weight, Age, BallControl, Intercept\} : \beta_i \neq 0$$

Como obtenemos un $p - valor < 0.05$, rechazamos H_0 con un nivel de significación $\alpha = 0.05$ y concluimos que el modelo es globalmente válido.

Ajustemos ahora distintos modelos considerando distintas variables explicativas:

```
modelo_2 <- lm(datos$Rating~datos$Height+datos$Age+datos$Ball_Control)
summary(modelo_2)
```

```
##
## Call:
## lm(formula = datos$Rating ~ datos$Height + datos$Age + datos$Ball_Control)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.3086  -3.6190  -0.3503   3.2282  29.4415
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -5.274560   1.262194  -4.179 2.94e-05 ***
## datos$Height        0.240857   0.006629  36.336  < 2e-16 ***
## datos$Age           0.599795   0.008691  69.014  < 2e-16 ***
## datos$Ball_Control  0.219740   0.002629  83.571  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.333 on 17584 degrees of freedom
## Multiple R-squared:  0.4332, Adjusted R-squared:  0.4331
## F-statistic:  4480 on 3 and 17584 DF,  p-value: < 2.2e-16
```

```
modelo_3 <- lm(datos$Rating~datos$Age+datos$Ball_Control)
summary(modelo_3)
```

```
##
## Call:
## lm(formula = datos$Rating ~ datos$Age + datos$Ball_Control)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.612  -3.712  -0.425   3.301  29.874
##
## Coefficients:
```

```
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        39.672155   0.260369  152.37   <2e-16 ***
## datos$Age           0.637652   0.008946   71.28   <2e-16 ***
## datos$Ball_Control  0.180444   0.002485   72.61   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.529 on 17585 degrees of freedom
## Multiple R-squared:  0.3907, Adjusted R-squared:  0.3906
## F-statistic:  5637 on 2 and 17585 DF,  p-value: < 2.2e-16
```

```
modelo_4 <- lm(datos$Rating~datos$Height+datos$Weight)
summary(modelo_4)
```

```
##
## Call:
## lm(formula = datos$Rating ~ datos$Height + datos$Weight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.5331  -4.5355   0.1209   4.5800  27.2131
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  73.87482    1.62905   45.35   <2e-16 ***
## datos$Height -0.14702    0.01210  -12.15   <2e-16 ***
## datos$Weight  0.25139    0.01171   21.47   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.985 on 17585 degrees of freedom
## Multiple R-squared:  0.0277, Adjusted R-squared:  0.02759
## F-statistic: 250.5 on 2 and 17585 DF,  p-value: < 2.2e-16
```

Realizando los mismos contrastes que para el modelo anterior, en el que hemos considerado todas las variables explicativas, vemos que todos los modelos ajustados son globalmente válidos y todas las variables son significativas.

Para comparar los distintos modelos, nos fijamos en el coeficiente R^2 ajustado. Tanto R^2 como R^2 ajustado informan sobre la bondad del ajuste, sin embargo el primero depende del número de variables explicativas consideradas, es decir, es mayor cuantas más variables explicativas se consideren, aunque estas no sean significativas, mientras que el segundo no depende del número de variables explicativas consideradas, por lo que es más fiable. Cuando más cerano a 1 sea su valor, mejor será el modelo.

Observando el resultado de los modelos ajustados, vemos que tras eliminar la variable *Weight* del primer modelo, obtenemos un resultado muy parecido al obtenido en el modelo con todas las variables explicativas, Vemos un coeficiente R^2 ajustado ligeramente menor, por lo que parece un modelo ligeramente peor.

En el *modelo_3* se consideran las variables explicativas *Age* y *Ball_Control*. Otenemos un coeficiente R^2 ajustado aún menor, por lo que no parece que haya mejoría en modelo.

El último de ellos, es con diferencia el peor de todos los modelos. Obtenemos un coeficiente R^2 ajustado de 0.02759. Es un modelo malo.

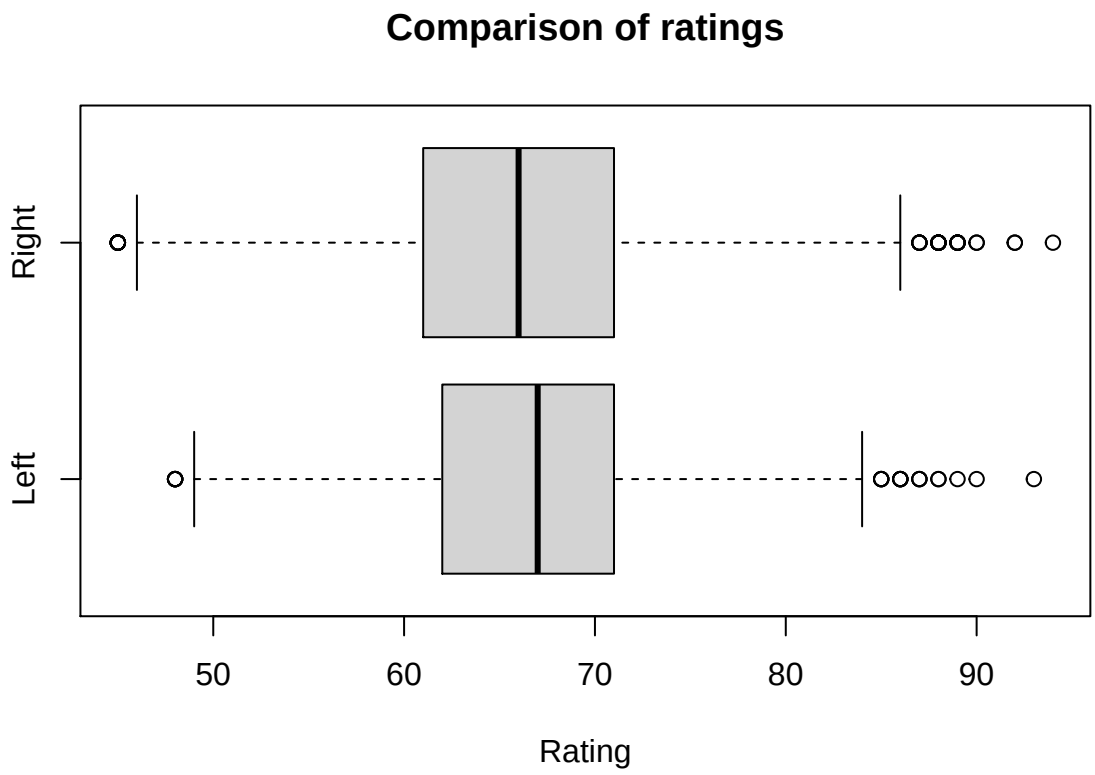
Si observamos la matriz de correlación obtenida en el apartado anterior, vemos como precitamente las variables consideradas en este último modelo (*Height* y *Weight*), son las que presentan menor correlación con la variable *Rating*.

En definitiva, el mejor modelo obtenedo ha sido el *modelo_1*.

7 Representación de los resultados a partir de tablas y gráficas.

Hay varias maneras de visualizar los resultados obtenidos de la comparación de jugadores diestros y zurdos. Una manera sencilla de ver cómo de diferentes o iguales son estas variables en los dos datasets (Rating y Ball_Control), es utilizar un boxplot. Empezamos con rating.

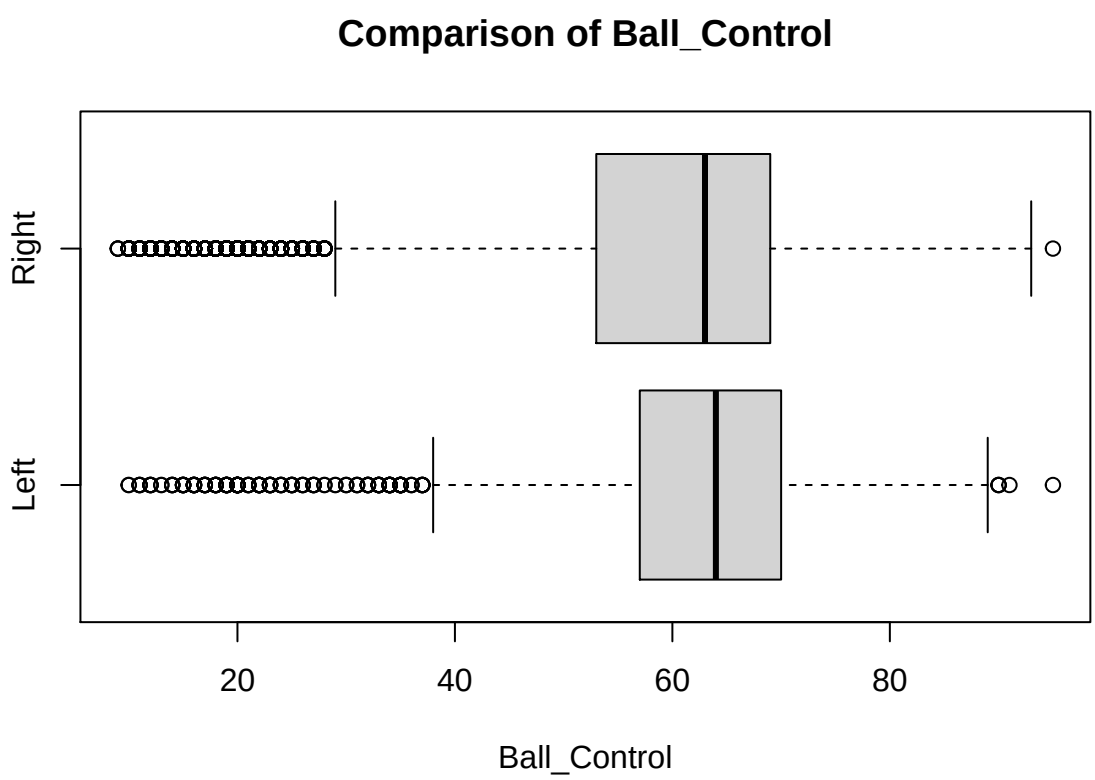
```
boxplot(Left$Rating, Right$Rating, names=c("Left", "Right"),
        horizontal = TRUE, main = "Comparison of ratings", xlab = "Rating")
```



Podemos ver el rating de las dos muestras se parecen bastante, con la mediana y Q1 de los zurdos ligeramente mayores que las de los diestros.

Ahora vamos a visualizar Ball_control.

```
boxplot(Left$Ball_Control, Right$Ball_Control, names=c("Left", "Right"),
        horizontal = TRUE, main = "Comparison of Ball_Control", xlab = "Ball_Control")
```



Podemos observar que para ball control la mediana de los zurdos también está ligeramente por encima de la de los diestros. Además, en este caso la diferencia más clara es que el valor de Q1 de los zurdos es bastante mayor que el de los diestros.

8 Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

Tras realizar distintos constrastes, hemos podido ver como sí que existe una cierta incluencia en las variables *Ball_Control* y *Rating* entre jugadores zurdos y diestros.

También hemos obtenido ciertos modelos de regresión con la finalidad de poder predecir la variable *Rating* en función de ciertas variables cuantiativas del juego de datos, siendo el mejor de ellos el que tiene por variables indipendientes *Height*, *Weight*, *Age* y *Ball_Control*, y el peor de ellos, el que tiene únicamente como variables explicativas *Height* y *Weight*. No obstante, no hemos obtenido un modelo realmente bueno al no tener ninguno de ellos un coeficiente R^2 ajustado especialmente alto.

9 Referencias

- https://rpubs.com/Joaquin_AR/218466
- Vegas, E. (2017). Preprocesamiento de datos. Material UOC.
- Gibergans, J. (2017). Regresión lineal múltiple. Material UOC.
- Rovira, C. (2008). Contraste de hipótesis. Material UOC.
- Test for homogeneity of variances - Lavene's test and the Fligner

10 Contribuciones al trabajo

	Contribuciones	Firma
1	Investigación previa	Arturo Hernández y Laia Cebey
2	Redacción de las respuestas	Arturo Hernández y Laia Cebey
3	Desarrollo del código	Arturo Hernández y Laia Cebey

El vídeo con la presentación de la actividad se puede encontrar en https://drive.google.com/file/d/1_BHrOoazctFoweWs2jXvSvNxhVeuaP2W/view?usp=sharing. El repositorio de github donde se puede encontrar el código es https://github.com/lcebeyuoc/PRA2_Tipologia_UOC.