Universitat Oberta de Catalunya uoc.edu

Práctica 2: Limpieza y validación de los datos

Luis Antón López Sobrado

26/05/2019

1. Descripción del dataset. Por que es importante y que pregunta/problema pretende responder?

Respuesta: El juego de datos escogido sobre el que realizaremos un trabajo descriptivo es "2016 RioOlympics athletes and medals" (https://www.kaggle.com/rio2016/olympic-games). Para empezar a analizar los datos, descargamos el fichero y cargamos los datos en R para comprobar la estructura.

```
# Cargamos el juego de datos

datos <- read.csv('athletes.csv',stringsAsFactors = FALSE, header = FALSE)

# Nombres de los atributos

colnames(datos) <- c("id","name","nacionality","sex","birthDate","height","weight","sport","gold","silver","bronze")

# Verificamos la estructura de los datos cargados

str(datos, vec.len = 2, strict.width = "no", width = 30)
```

```
## 'data.frame':    11539 obs. of  11 variables:
##  $ id         : chr  "id" "736041664" ...
##  $ name       : chr  "name" "A Jesus Garcia" ...
##  $ nacionality: chr  "nationality" "ESP" ...
##  $ sex        : chr  "sex" "male" ...
##  $ birthDate  : chr  "dob" "10/17/69" ...
##  $ height     : chr  "height" "1.72" ...
##  $ weight     : chr  "weight" "64" ...
##  $ sport      : chr  "sport" "athletics" ...
```

```
##  $ gold       : chr  "gold" "0" ...
##  $ silver     : chr  "silver" "0" ...
##  $ bronze     : chr  "bronze" "0" ...
```

Este conjunto de datos consta de las estadísticas oficiales de los 11538 atletas y 306 eventos en los Juegos Olímpicos 2016 en Río de Janeiro. El fichero con los datos (*athletes.csv*) dispone de 11538 registros (filas) y 11 variables (columnas) las cuales describiremos a continuación:

VARIABLE	**DESCRIPCIÓN**
id	Identificador del atleta
name	Nombre del atleta
nacionality	Nacionalidad
sex	Sexo
dob	Fecha de nacimiento
height	Altura
weight	Peso
sport	Deporte
gold	Número de medallas de oro
silver	Número de medallas de plata
bronze	Número de medallas de bronce

Con este juego de datos vamos a realizar un estudio sobre los atletas que participaron en las olimpiadas de Rio 2016, con el que se pretende responder a una serie de preguntas que nos permitan comparar con otros juegos y así ver la evolución, los niveles de participación, mayor número de mujeres o no, la evolución en las diferentes disciplinas, etc.

- Edades de los participantes, el atleta más joven, más mayor o la media de edad.
- El número de participantes de cada sexo.
- El número de participantes de cada sexo en cada disciplina.
- El número de participantes por país.
- Número de medallas entregadas en total, de cada tipo y por país.

2. Integración y selección de los datos de interés a analizar.

Respuesta: Para el proceso de análisis entre las diferentes variables del juego de datos se tienen que tener en cuentas las preguntas que queremos responder. A primera vista se puede observar que las variables id, name, height y weight no influyen en el análisis, y por tanto, las podemos eliminar y realizar la limpieza de los datos para el correspondiente análisis con las variables nacionality, sex, dob sport, gold, silver y bronce.

Para poder realizar el análisis es necesario realizar algunas tranformaciones sobre las siguientes variables:

- dob (Fecha de nacimiento) , necesitamos saber la edad de cada uno de los atletas, para ello se tendrá que a partir e la fecha almacenada, convertirla a numérico y restarle el año actual.
- Gold, silver y bronce. Si observamos esta variables, necesitamos crear una nueva variable llamada "Medallas", para ello sumamos cada una uno de los valores numérico de cada variables y realizando un proceso de discretización donde se crearan 2 categorías (SI o NO) en las que dividiremos las observaciones.

Si realizamos la preparación de los datos y las transformaciones comentadas en R obtenemos el siguiente resultado:

```
# A primera vista podemos ver que los atributos id, name, height y weight, no
influyen en el analisis, y por tanto, las podemos eliminar.

datos[["id"]]=NULL

datos[["name"]]=NULL

datos[["height"]]=NULL

datos[["weight"]]=NULL

# Inicialmente a partir de la fecha de nacimiento calculamos las edades de cada
unos de los cliente. Obtenmos el año de la fecha almacenada, la convertimos a
numerico y restamos el año actual.

options(chron.year.expand =

      function (y, cut.off = 16, century = c(1900, 2000), ...) {

          chron:::year.expand(y, cut.off = cut.off, century = century, ...)

      }
```

```
)

datos$fechaAux <- as.Date(chron(format(as.Date(datos$birthDate, "%m/%d/%Y"), "%m/%d/%y")))

datos$edad <- 2016 - (as.numeric(format(datos$fechaAux, format="%Y")))

datos[["fechaAux"]]=NULL

datos[["birthDate"]]=NULL

# Sumamos todas las medallas obtenidas y las alamacenamos en una varibale auxiliar.

datos$medallasAux <- ((as.numeric(datos$gold)) + (as.numeric(datos$silver)) + (as.numeric(datos$bronze)))

# Si observamos la variable medallasAux, podemos crear un nuevo atributo llamado medallas realizando un proceso de discretización sobre la variable medallasAux donde crearemos 2 categorías en las que dividiremos las observaciones.

datos$medallas[datos$medallasAux > 0] <- "SI"

datos$medallas[datos$medallasAux == 0] <- "NO"

# Eliminamos la variables auxiliares

datos[["medallasAux"]]=NULL

datos[["gold"]]=NULL

datos[["silver"]]=NULL

datos[["bronze"]]=NULL
```

Verificamos la estructura con los datos cargados, donde comprobamos que una vez que los datos han sido transformados disponemos para el análisis 5 variables.

```
str(datos, vec.len = 2, strict.width = "no", width = 30)

## 'data.frame':    11538 obs. of  5 variables:

##  $ nacionality: chr  "ESP" "KOR" ...

##  $ sex        : chr  "male" "female" ...

##  $ sport      : chr  "athletics" "fencing" ...
```

```
##  $ edad       : num  1950 1933 ...
##  $ medallas   : chr  "NO" "NO" ...
```

3. Limpieza de los datos.
    1. Los datos contienen ceros o elementos vacíos? .Como gestionarías cada uno de estos casos?

Respuesta: Para comprobar que los datos no contienen ceros o elementos vacíos tenemos que analizar cada una de las variables por separado, para ello ejecutamos el siguiente comando R para ver si tenemos en alguna de las variables algún elemento nulo:

```
#Comprobamos los valores nulos
colSums(is.na(datos))
```

```
## nacionality         sex       sport         edad    medallas
##           0           0           0           1           0
```

Vemos que la variable edad tiene un valor nulo, por tanto, para el análisis decidimos eliminarlo a partir del siguiente comando:

```
# eliminamos todos los valores vacíos
datos = na.omit(datos)
colSums(is.na(datos))
```

```
## nacionality         sex       sport         edad    medallas
##           0           0           0           0           0
```

A continuación, comprobamos que los valores de cada una de las variables no tengan valores extraños o fuera de rango.

- Edad:

```
table(datos$edad)
```

```
##  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31
##   8   7  45  82 157 325 454 569 842 933 893 920 924 894 829 684 634 505
##  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49
## 413 300 283 190 170  97  73  56  44  34  26  24  18  11  10  18   7   7
##  50  51  52  53  54  55  56  57  58  59  60  61  62
##   4   8  10  10   6   2   2   2   1   1   2   1   2
```

Una vez eliminado el valor nulo, vemos que los valores de edades están entre rangos normales [14,62], tal y como ha conformado la EFE (https://www.informador.mx/Deportes/La-edad-no-importa-en-Juegos-Olimpicos-de-Rio-2016-20160805-0020.html)

- nacionality (nacionalidad):

```
table(datos$nacionality)
```

```
## AFG ALB ALG AND ANG ANT ARG ARM ARU ASA AUS AUT AZE BAH BAN BAR BDI BEL
##   3   6  68   5  26   9 223  32   7   4 431  71  56  30   7  11   9 108
## BEN BER BHU BIH BIZ BLR BOL BOT BRA BRN BRU BUL BUR CAF CAM CAN CAY CGO
##   6   8   2  11   3 124  12  12 485  34   3  50   5   6   6 321   5  10
## CHA CHI CHN CIV CMR COD COK COL COM CPV CRC CRO CUB CYP CZE DEN DJI DMA
##   2  42 404  12  24   4   9 154   4   5  11  88 123  16 104 128   7   2
## DOM ECU EGY ERI ESA ESP EST ETH FIJ FIN FRA FSM GAB GAM GBR GBS GEO GEQ
##  29  38 122  12   8 313  46  38  54  54 410   5   6   4 374   5  40   2
## GER GHA GRE GRN GUA GUI GUM GUY HAI HKG HON HUN INA IND IOA IRI IRL IRQ
## 441  16  93   7  21   5   5   6  10  38  30 154  28 123   9  64  80  26
## ISL ISR ISV ITA IVB JAM JOR JPN KAZ KEN KGZ KIR KOR KOS KSA LAO LAT LBA
##   8  47   7 312   4  57   8 346 103  80  19   3 213   8  11   6  32   7
## LBR LCA LES LIB LIE LTU LUX MAD MAR MAS MAW MDA MDV MEX MGL MHL MKD MLI
##   2   5   8   9   3  67  10   6  49  32   5  23   4 126  43   5   6   6
```

```
## MLT MNE MON MOZ MRI MTN MYA NAM NCA NED NEP NGR NIG NOR NRU NZL OMA PAK
##   7  35   3   6  11   2   7  10   5 249   7  78   6  62   2 208   4   7
## PAN PAR PER PHI PLE PLW PNG POL POR PRK PUR QAT ROT ROU RSA RUS RWA SAM
##  10  11  29  13   6   5   8 242  95  31  40  39  10  98 146 286   7   8
## SEN SEY SIN SKN SLE SLO SMR SOL SOM SRB SRI SSD STP SUD SUI SUR SVK SWE
##  22  10  25   7   4  63   5   3   2 103   9   3   3   6 104   6  51 164
## SWZ SYR TAN TGA THA TJK TKM TLS TOG TPE TTO TUN TUR TUV UAE UGA UKR URU
##   2   7   7   7  54   7   9   3   5  56  32  61 103   1  13  21 205  17
## USA UZB VAN VEN VIE VIN YEM ZAM ZIM
## 567  70   4  88  23   4   3   7  35
```

Los valores disponibles para la variable *nacionality* son 3 caracteres con la abreviación del país al que pertenece el atleta. Se comprueba que no haya registros vacíos y si cada nacionalidad contiene el formato adecuado, tal y como se puede ver en la tabla, y verificamos que todos los registros son correctos.

- sex (sexo):

```
table(datos$sex)
```

```
## female   male
##   5205   6333
```

La variable *sex* se representa con la palabra “male” para los hombres y “female” para las mujeres. Se comprueba que no hay registros vacíos o con caracteres extraños, y verificamos que todos los registros son correctos.

- sport (deporte):

```
table(datos$sport)
```

```
##          aquatics           archery         athletics         badminton
##              1445               128              2363               172
##        basketball            boxing             canoe           cycling
```

```
##              288              286              331              525
##        equestrian          fencing         football             golf
##              222              246              611              120
##        gymnastics         handball           hockey             judo
##              324              363              432              392
## modern pentathlon           rowing      rugby sevens          sailing
##               72              547              300              380
##          shooting      table tennis         taekwondo           tennis
##              390              172              128              196
##         triathlon       volleyball     weightlifting         wrestling
##              110              384              258              353
```

Esta variable contiene la descripción de la disciplina de ese deportista. Se comprueba que no existe ningun valor extraño.

- medallas:

```
table(datos$medallas)
```

```
##   NO   SI
## 9681 1857
```

El campo medalla tenemos que recordar que se formó sumando las variables (gold, silver y bronze) y discretizando la variables en 2 categorías, ara indicar los que tienen medalla (SI) y los que no tienen (NO).

Una vez que terminadas las modificaciones, convertimos las siguientes variables *nacionality*, *sex*, *sport* y *medallas* de continuas en categóricas (discretización). Este proceso se realiza para tener una mejor comprensión de cada uno de los elementos, además de por motivos de almacenamiento.

```
# Una vez que terminadas las modificaciones, convertimos las siguientes variables
de continuas en categóricas (discretización). Este proceso se realiza para tener
```

```
una mejor comprensión de cada uno de los elementos, además de por motivos de
almacenamiento.
datos$nacionality = factor(datos$nacionality)
datos$sex =factor(datos$sex)
datos$sport = factor(datos$sport)
datos$medallas = factor(datos$medallas)
# Verificamos la estructura de los datos cargados
str(datos, vec.len = 2, strict.width = "no", width = 30)
## 'data.frame':    11537 obs. of  5 variables:
##  $ nacionality: Factor w/ 207 levels "AFG","ALB","ALG",..: 60 103 34 120
142 ...
##  $ sex        : Factor w/ 2 levels "female","male": 2 1 2 2 2 ...
##  $ sport      : Factor w/ 28 levels "aquatics","archery",..: 3 10 3 23 8 ...
##  $ edad       : num  47 30 24 25 26 ...
##  $ medallas   : Factor w/ 2 levels "NO","SI": 1 1 2 1 1 ...
```

2. Identificación y tratamiento de valores extremos.

Respuesta: Para identificar y tratar los valores extremos ejecutamos el siguiente comando en R :

```
summary(datos)
##    nacionality       sex             sport          edad       medallas
##  USA    : 567   female:5205   athletics:2363   Min.   :14.0   NO:9680
##  BRA    : 485   male  :6332   aquatics :1445   1st Qu.:23.0   SI:1857
##  GER    : 441                 football : 611   Median :26.0
##  AUS    : 431                 rowing   : 547   Mean   :26.7
##  FRA    : 410                 cycling  : 525   3rd Qu.:30.0
##  CHN    : 404                 hockey   : 432   Max.   :62.0
```

```
##  (Other):8799                  (Other)   :5614
```

Vemos que los datos de las variables *nacionality*, *sex*, *medallas* y *sport* son normales. No obstante, llama la atención los rangos de la variable numérica de la edad, donde el mínimo es 14 y el máximo es 62, a priori parece que los son valores anómalos pero si consultamos la pagina web de la EFE (https://www.informador.mx/Deportes/La-edad-no-importa-en-Juegos-Olimpicos-de-Rio-2016-20160805-0020.html), vemos que esos valores son correctos. En cuanto a la media (26.7) se considerá que es un valor correcto.

4. Análisis de los datos.
    1. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

Respuesta: Una vez transformados y limpiados los datos se procede a analizar los datos que se manejan y así dar respuesta a las preguntas planteadas, para ello, se tiene que tratar:

- nacionality: comprobar el número de atletas por nacionalidad.
- sex: ver el número de hombres frente al de mujeres.
- edad: comprobar la edad de los participantes, el rango mínimo, el máximo y la media.
- sport: comprobar la participación en cada deporte.
- medallas: calcular el total de medallas.

A nivel de grupo:

- Deporte con edad, para determinar las edades que comprende cada deporte.
- Deporte con sexo para ver la participación de hombres y mujeres en cada deporte
- Nacionalidad con medallas, para ver qué país fue el que ha ganado más medallas

2. Comprobación de la normalidad y homogeneidad de la varianza.

Respuesta: Existen muchas soluciones para probar la homogeneidad de la varianza entre los grupos como son: F-test (Compara las variaciones de dos muestras), Prueba de Bartlett (Compara las variaciones de x muestras, donde x puede ser más de dos muestras. Es menos sensible a las desviaciones de la normalidad que la prueba de Levene), Prueba de Levene (compara las

variaciones de x muestras, donde x puede ser más de dos muestras.) y la prueba de Fligner-Killeen (prueba no paramétrica que es muy robusta contra las desviaciones de la normalidad). Para todas estas pruebas tenemos que tener en cuenta que la hipótesis nula es que todas las varianzas de las poblaciones son iguales; y que la hipótesis alternativa es que al menos dos de ellos difieren. Para ello, se comprueba la homogeneidad de la varianza de la edad con el factor del sexo.

```
t.test(datos$edad~datos$sex)
```

```
##  Welch Two Sample t-test
##
## data:  datos$edad by datos$sex
## t = -9.1224, df = 11233, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.1078693 -0.7159718
## sample estimates:
## mean in group female   mean in group male
##             26.19673             27.10865
```

```
var.test(datos$edad~datos$sex)
```

```
##  F test to compare two variances
##
## data:  datos$edad by datos$sex
## F = 0.93734, num df = 5204, denom df = 6331, p-value = 0.01462
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8900203 0.9873089
## sample estimates:
```

```
## ratio of variances
##          0.9373409
```

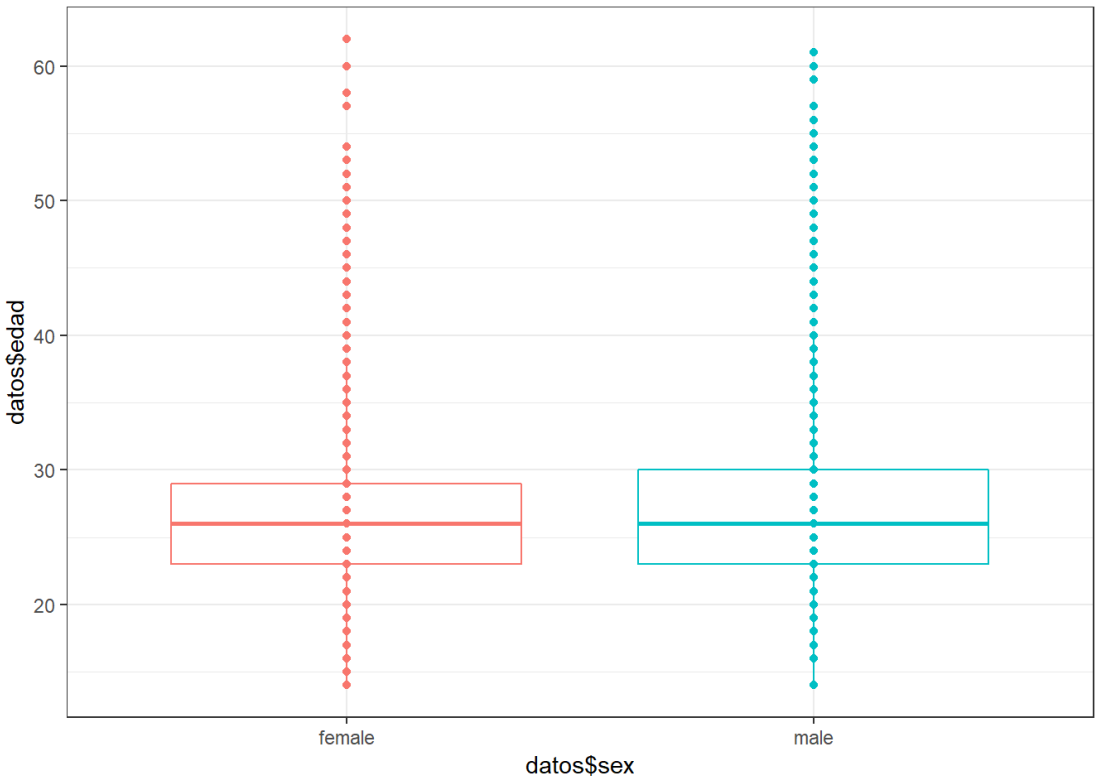
```
res <- bartlett.test(edad ~ sex, data = datos)
res
```

```
##  Bartlett test of homogeneity of variances
##
## data:  edad by sex
## Bartlett's K-squared = 5.9656, df = 1, p-value = 0.01459
```

```
resf <- fligner.test(edad ~ sex, data = datos)
resf
```

```
##  Fligner-Killeen test of homogeneity of variances
##
## data:  edad by sex
## Fligner-Killeen:med chi-squared = 0.20866, df = 1, p-value =
## 0.6478
```

```
gplot(data = datos, aes(x = datos$sex, y = datos$edad, colour = datos$sex)) +
geom_boxplot() + geom_point() + theme_bw() + theme(legend.position = "none")
```



Se comprueba que es homogénea y no varía en función del sexo.

3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

Respuesta: Para la realización de este apartado vamos a analizar cada grupo de datos para determinar los resultado y así poder sacar posibles conclusiones.

- Edad

```
mean(datos$edad)
```

```
## [1] 26.69723
```

```
median(datos$edad)
```

```
## [1] 26
```

```
sd(datos$edad)
```

```
## [1] 5.378814
var(datos$edad)
## [1] 28.93164
quantile(datos$edad,c(0.25,0.5,0.75))
## 25% 50% 75%
##  23  26  30
summary(datos$edad)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    14.0    23.0    26.0    26.7    30.0    62.0
```

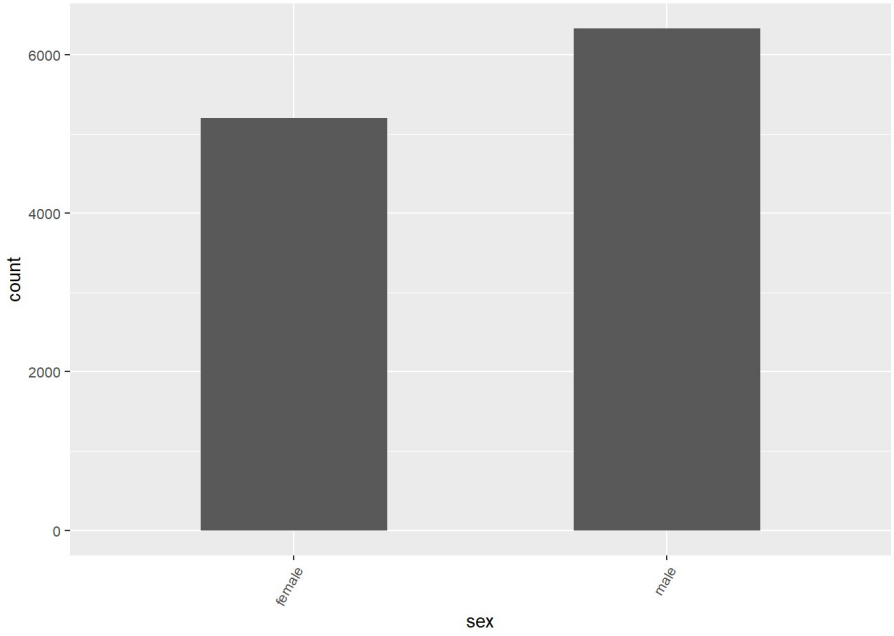
- Sexo , deportes, medallas y nacionalidades.

```
summary(datos)
##   nacionality       sex             sport           edad      medallas
##  USA    : 567   female:5205   athletics:2363   Min.   :14.0   NO:9680
##  BRA    : 485   male  :6332   aquatics :1445   1st Qu.:23.0   SI:1857
##  GER    : 441                 football : 611   Median :26.0
##  AUS    : 431                 rowing   : 547   Mean   :26.7
##  FRA    : 410                 cycling  : 525   3rd Qu.:30.0
##  CHN    : 404                 hockey   : 432   Max.   :62.0
##  (Other):8799                 (Other)  :5614
```

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

Respuesta: A continuación se muestran gráficas y tablas con los resultados obtenidos de combinar varias grupos de datos así como de analizar cada dato por separado:

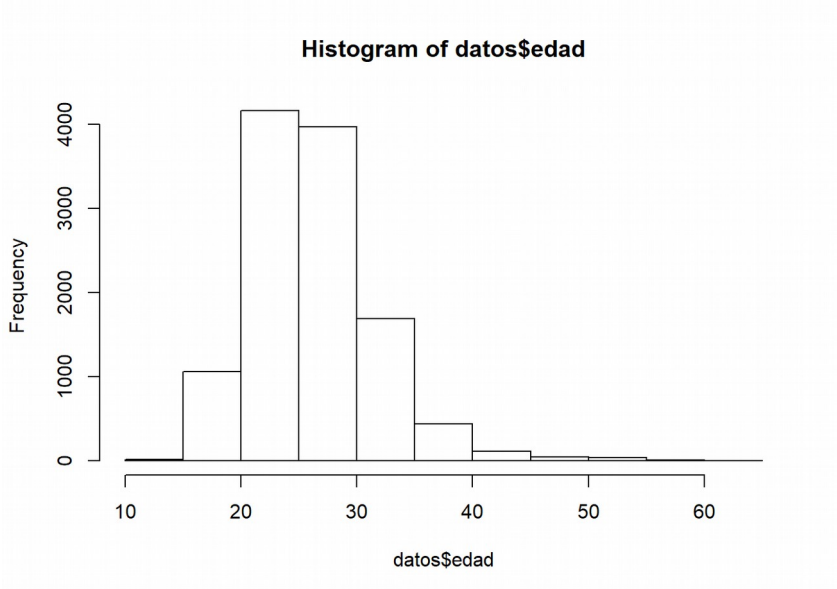
```
ggplot(datos, aes(x = sex)) + geom_bar(width=0.5)  + theme(text =
element_text(size=10),axis.text.x = element_text(angle=60, hjust=1))
```




```
summary(datos$edad)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    14.0    23.0    26.0    26.7    30.0    62.0
```

```
hist(datos$edad)
```



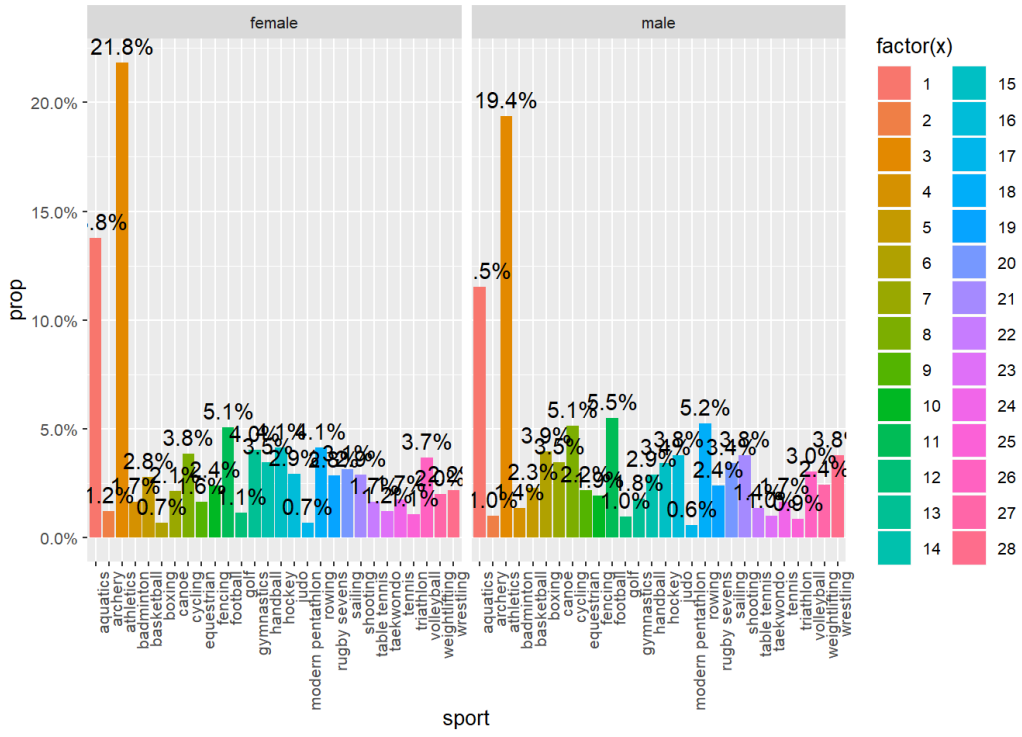
```r
#Matriz de porcentages de frecuencia entre las variables
datosSexoSport<-table(datos$sex,datos$sport)
for (i in 1:dim(datosSexoSport)[1]){
    datosSexoSport[i,]<-datosSexoSport[i,]/sum(datosSexoSport[i,])*100
}
datosSexoSport
```

```
##
##            aquatics    archery  athletics  badminton basketball     boxing
##   female 13.7560038  1.2295869 21.8443804  1.6522574  2.7665706  0.6916427
##   male   11.5129501  1.0107391 19.3619709  1.3581807  2.2741630  3.9481996
##
##               canoe    cycling equestrian    fencing   football       golf
##   female  2.1325648  3.8424592  1.6330451  2.3823247  5.0720461  1.1335255
##   male    3.4744157  5.1326595  2.1636134  1.9267214  5.4801011  0.9633607
##
##          gymnastics   handball     hockey       judo modern pentathlon
##   female  4.0345821  3.4582133  4.1498559  2.9394813         0.6916427
##   male    1.8003790  2.8900821  3.4112445  3.7744788         0.5685407
##
##              rowing rugby sevens    sailing   shooting table tennis
##   female  4.1498559    2.8434198  3.1316042  2.9010567    1.6522574
##   male    5.2274163    2.4005054  3.4112445  3.7744788    1.3581807
##
##           taekwondo     tennis  triathlon volleyball weightlifting
##   female  1.2295869  1.7483189  1.0566763  3.6887608     1.9980788
##   male    1.0107391  1.6582438  0.8686039  3.0322173     2.4320910
##
```

```
##           wrestling
##    female  2.1902017
##    male    3.7744788
```

```
theme_set(theme_gray(base_size = 10))

ggplot(datos, aes(x=sport,  group=sex)) + geom_bar(aes(y = ..prop.., fill =
factor(..x..)), stat="count") + geom_text(aes( label = scales::percent(..prop..),
y= ..prop.. ), stat= "count", vjust = -.5) + facet_grid(~sex) +
scale_y_continuous(labels = scales::percent) + theme(text =
element_text(size=10),axis.text.x = element_text(angle=90, hjust=1))
```




```
table(datos$sport)
```

```
##          aquatics           archery          athletics          badminton
##              1445               128              2363               172
```

```
##        basketball            boxing             canoe           cycling
##               288               286               331               525
##        equestrian           fencing          football              golf
##               222               246               611               120
##        gymnastics          handball            hockey              judo
##               324               363               432               392
## modern pentathlon            rowing      rugby sevens           sailing
##                72               547               300               379
##          shooting      table tennis         taekwondo            tennis
##               390               172               128               196
##         triathlon        volleyball     weightlifting         wrestling
##               110               384               258               353
```

```
tabla<-table(datos$sex,datos$sport)
tabla
```

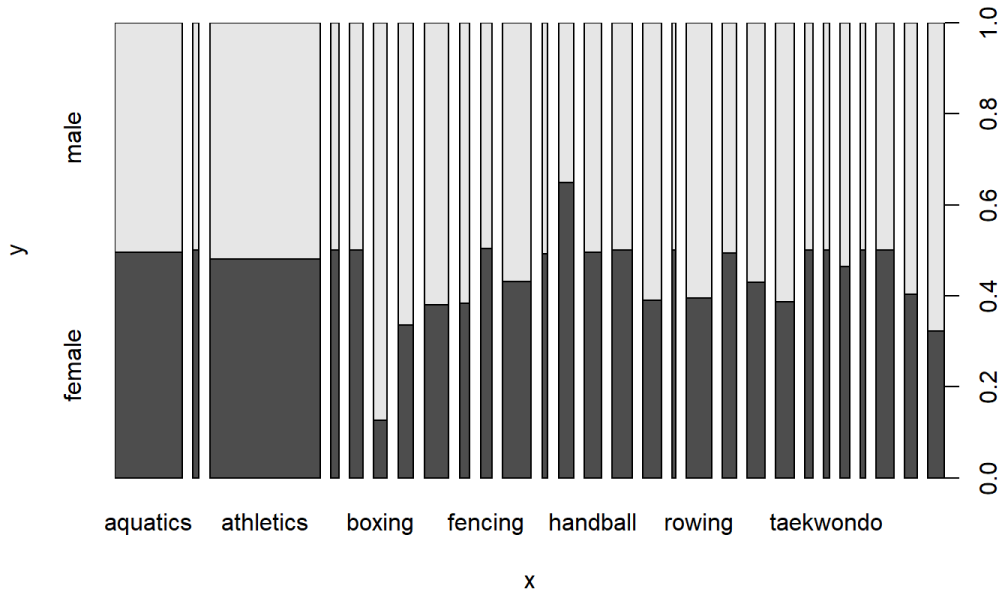
```
##          aquatics archery athletics badminton basketball boxing canoe
##   female      716      64      1137        86        144     36   111
##   male        729      64      1226        86        144    250   220
##
##          cycling equestrian fencing football golf gymnastics handball
##   female     200         85     124      264   59        210      180
##   male       325        137     122      347   61        114      183
##
##          hockey judo modern pentathlon rowing rugby sevens sailing
##   female    216  153                36    216          148     163
##   male      216  239                36    331          152     216
##
##          shooting table tennis taekwondo tennis triathlon volleyball
```

```
##    female      151           86        64      91        55         192
##    male        239           86        64     105        55         192
##
##          weightlifting wrestling
##    female           104       114
##    male             154       239
```

```
reglin<-lm(datos$sport~datos$sex)
reglin
```

```
##
## Call:
## lm(formula = datos$sport ~ datos$sex)
##
## Coefficients:
##    (Intercept)  datos$sexmale
##         10.783          0.469
```

```
plot(datos$sport,datos$sex)
abline(reglin,col=2)
```



```
tabla2<-table(datos$sex,datos$edad)
tabla2
```

```
##           14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29
##   female   7   7  38  61 107 188 218 262 397 399 402 422 432 399 371 289
##   male     1   0   7  21  50 137 236 307 445 534 491 498 492 495 458 395
##
##           30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45
##   female 280 216 151 136 111  75  70  38  25  22  13   9  10   9   6   4
##   male   354 289 262 164 172 115 100  59  48  34  31  25  16  15  12   7
##
##           46  47  48  49  50  51  52  53  54  55  56  57  58  59  60  61
##   female   4   7   3   2   1   3   2   3   1   0   0   1   1   0   1   0
##   male     6  11   4   5   3   5   8   7   5   2   2   1   0   1   1   1
##
##           62
##   female   2
##   male     0
```

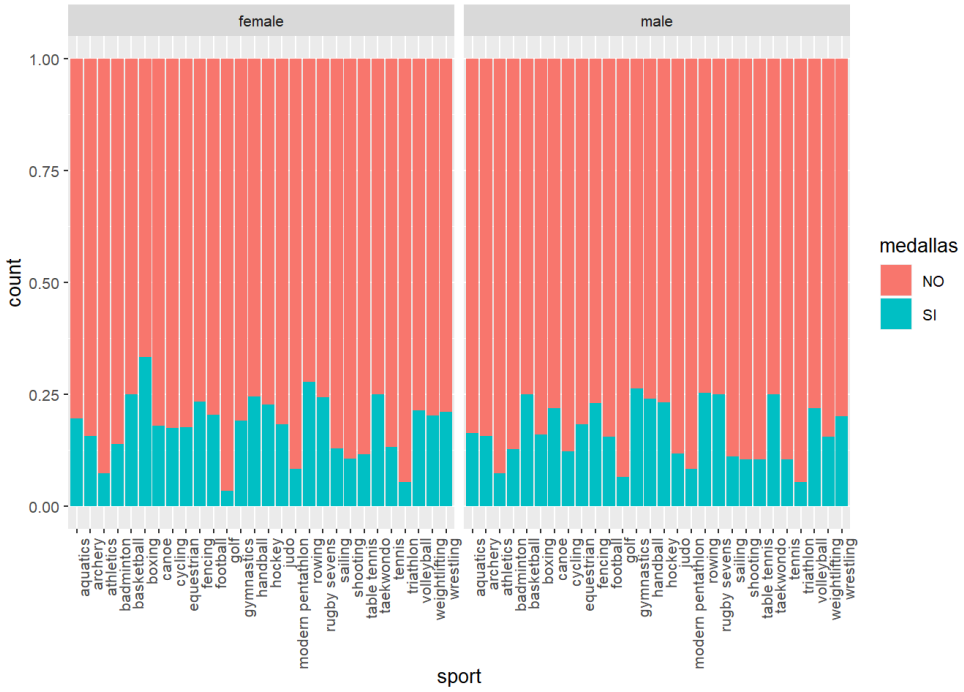
```
tabla5<-table(datos$medallas,datos$nacionality)
tabla5
```

```
##
##      AFG ALB ALG AND ANG ANT ARG ARM ARU ASA AUS AUT AZE BAH BAN BAR BDI
##   NO   3   6  67   5  26   9 201  28   7   4 360  69  38  24   7  11   8
##   SI   0   0   1   0   0   0  22   4   0   0  71   2  18   6   0   0   1
##
##      BEL BEN BER BHU BIH BIZ BLR BOL BOT BRA BRN BRU BUL BUR CAF CAM CAN
##   NO  87   6   8   2  11   3 112  12  12 436  32   3  43   5   6   6 259
##   SI  21   0   0   0   0   0  12   0   0  49   2   0   7   0   0   0  62
```

```
##
##       CAY CGO CHA CHI CHN CIV CMR COD COK COL COM CPV CRC CRO CUB CYP CZE
##    NO   5  10   2  42 304  10  24   4   9 146   4   5  11  64 112  16  90
##    SI   0   0   0   0 100   2   0   0   0   8   0   0   0  24  11   0  14
##
##       DEN DJI DMA DOM ECU EGY ERI ESA ESP EST ETH FIJ FIN FRA FSM GAB GAM
##    NO  89   7   2  28  38 119  12   8 270  42  31  41  53 318   5   6   4
##    SI  39   0   0   1   0   3   0   0  43   4   7  13   1  92   0   0   0
##
##       GBR GBS GEO GEQ GER GHA GRE GRN GUA GUI GUM GUY HAI HKG HON HUN INA
##    NO 244   5  33   2 290  16  87   6  21   5   5   6  10  38  30 139  24
##    SI 130   0   7   0 151   0   6   1   0   0   0   0   0   0   0  15   4
##
##       IND IOA IRI IRL IRQ ISL ISR ISV ITA IVB JAM JOR JPN KAZ KEN KGZ KIR
##    NO 121   7  56  77  26   8  45   7 243   4  33   7 288  86  68  19   3
##    SI   2   2   8   3   0   0   2   0  69   0  24   1  58  17  12   0   0
##
##       KOR KOS KSA LAO LAT LBA LBR LCA LES LIB LIE LTU LUX MAD MAR MAS MAW
##    NO 190   7  11   6  32   7   2   5   8   9   3  60  10   6  48  24   5
##    SI  23   1   0   0   0   0   0   0   0   0   0   7   0   0   1   8   0
##
##       MDA MDV MEX MGL MHL MKD MLI MLT MNE MON MOZ MRI MTN MYA NAM NCA NED
##    NO  22   4 121  41   5   6   6   7  35   3   6  11   2   7  10   5 203
##    SI   1   0   5   2   0   0   0   0   0   0   0   0   0   0   0   0  46
##
##       NEP NGR NIG NOR NRU NZL OMA PAK PAN PAR PER PHI PLE PLW PNG POL POR
##    NO   7  60   5  43   2 173   4   7  10  11  29  12   6   5   8 226  94
```

```
##    SI   0  18   1  19   0  35   0   0   0   0   0   1   0   0   0  16   1
##
##       PRK PUR QAT ROT ROU RSA RUS RWA SAM SEN SEY SIN SKN SLE SLO SMR SOL
##    NO  24  39  38  10  81 124 182   7   8  22  10  24   7   4  59   5   3
##    SI   7   1   1   0  17  22 103   0   0   0   0   1   0   0   4   0   0
##
##       SOM SRB SRI SSD STP SUD SUI SUR SVK SWE SWZ SYR TAN TGA THA TJK TKM
##    NO   2  50   9   3   3   6  93   6  43 138   2   7   7   7  48   6   9
##    SI   0  53   0   0   0   0  11   0   8  26   0   0   0   0   6   1   0
##
##       TLS TOG TPE TTO TUN TUR TUV UAE UGA UKR URU USA UZB VAN VEN VIE VIN
##    NO   3   5  51  31  58  95   1  12  21 192  17 356  57   4  85  22   4
##    SI   0   0   5   1   3   8   0   1   0  13   0 211  13   0   3   1   0
##
##       YEM ZAM ZIM
##    NO   3   7  35
##    SI   0   0   0
```

```
ggplot(data = datos,aes(x=sport,fill=medallas))+geom_bar(position="fill")
+facet_wrap(~sex) + theme(text = element_text(size=10),axis.text.x =
element_text(angle=90, hjust=1))
```



6. Resolución del problema. A partir de los resultados obtenidos, .cuales son las conclusiones? .Los resultados permiten responder al problema?

Respuesta: Las conclusiones obtenidas a partir de la representación de los resultados, y con la intención de dar respuesta a las preguntas formuladas son las siguientes:

- Por un lado el número de hombres en comparación con el de las mujeres. En estas olimpiadas se observa que han participado un total de 11538 atletas, donde 5205 fueron mujeres y 6332 fueron hombres. Por tanto, hubo una participación mayor de hombres que de mujeres.
- De las edades de los participantes, el más joven, el más mayor y la media de edad observamos los siguientes resultados. El atleta de menor edad fue de 14 años y el de mayor edad fue de 62 años (son mujeres). Con una media de edad de 26 años para las mujeres y 27 años para los hombres.
- Si miramos la partición de hombre y mujeres en cada disciplina obtenemos los siguientes datos:

1. Que el atletismo es el deporte con más participantes con 2362 y el pentatlón el que menos, con 12.
2. Todos los deportes a nivel de genero están bastante equilibrados, algunos deporte como el boxeo, la canoa y la lucha donde el género masculino predomina con creces frente al femenino; pero por otro lado en gimnasia el género femenino es el dominante.
3. En cuanto a las edades vemos que las disciplina ecuestre y el tiro contiene los participantes más longevos de 62 años y la natación, con 14 años la persona más joven.

- Si analizamos la participación de los atletas por nacionalidades vemos que América (USA) ha sido la que más participantes ha tenido en los juegos con 567 participantes y las que menos fueron Bolivia y Botswana tan solo con 12 participantes.
- Y por último, en cuento a la cantidad de medallas por país se observa que se han entregado un total de 1857. Donde Estados Unidos, China, Rusia, Gran Bretaña y Alemania fueron los países que más medallas ganaron. En el caso de España, fueron un total de 43 medallas.

Con el análisis realizamos tenemos los datos suficientes de cara a la preparación de los próximos juegos olímpicos.

7. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, analisis y representacion de los datos. Si lo preferís, también podéis trabajar en **Python.**