Tipología y ciclo de vida de los datos II

El objetivo de esta actividad será el tratamiento de un dataset, que puede ser el creado en la práctica 1 o cualquier dataset libre disponible en Kaggle (https://www.kaggle.com). Las diferentes tareas a realizar (y justificar) son las siguientes:

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

El dataset es un conjunto de datos de los atletas que participaron en las olimpiadas de Rio 2016. Se obtiene el dataset de Kaggle
El dataset tiene los siguientes datos:
Id – Identificador del atleta
Name – Nombre del atleta
Nacionality - Nacionalidad
Sex – Sexo
Dob – Fecha de nacimiento
Height - Altura
Weight - Peso
Sport - Deporte
Gold – Numero de medallas de oro
Silver – Número de medallas de plata
Bronze – Número de medallas de bronce

Con estos datos vamos a realizar un estudio sobre los atletas que participaron en estas olimpiadas.
La idea es conseguir la información necesaria para responder a las siguientes preguntas:

- Las edades de los participantes, el atleta más joven, más mayor, la media de edad…
- El número de hombres frente al de las mujeres
- La participación de los hombres y mujeres en cada disciplina
- La participación de los atletas por nacionalidad
- La cantidad total de medallas entregadas. Medallas entregadas por país

Con esta información podremos comparar los datos con los juegos olímpicos anteriores y ver la evolución, la incorporación de más mujeres o no, los niveles de participación, la evolución de la participación en las disciplinas, etc.

2. Limpieza de los datos.

2.1. Selección de los datos de interés a analizar. ¿Cuáles son los campos más relevantes para responder al problema?

Para dar respuesta a las preguntas planteadas, es suficiente con analizar los siguientes datos:

Nacionalidad: nacionalidad del atleta
Sexo: genero
Edad: la edad no la tenemos pero la calculamos a partir de la fecha de nacimiento
Disciplina: el deporte que realiza el atleta
Medallas: indica si el atleta ha obtenido medalla o no. Se obtiene a partir de los campos gold, silver y bronze

En cuanto a la altura y peso, para este estudio en concreto podemos prescindir de ellos. Además contienen muchos registros vacios por lo que los datos no serian del todo reales.

2.2. ¿Los datos contienen ceros o elementos vacíos? ¿Y valores extremos? ¿Cómo gestionarías cada uno de estos casos?

Analizamos los datos seleccionados. Para cada dato comprobamos los siguientes pasos

Nacionalidad: la nacionalidad son 3 caracteres con la abreviación del país al que pertenece el atleta. Se comprueba que no haya registros vacios y si cada nacionalidad contiene el formato adecuado. Como resultado obtenemos que todos los registros son correctos.

Sexo: el sexo se representa con la palabra “male” para los hombres y “female” para las mujeres. Se comprueba que no hay registros vacios y que todos los registros están bien escritos. Como resultado obtenemos que todos los registros son correctos.

Edad: no existe campo edad pero si un campo fecha. Lo que hacemos es calcular la edad con respeto a la fecha de nacimiento y el año de los juegos olímpicos de Río (2016). Previo a esto comprobamos si existen registros vacios o si tienen un formato incorrecto. Se observa que hay un solo atleta que no tiene fecha de nacimiento. Lo que se hace es comprobar de qué atleta se trata, buscamos información de él en Internet y cubrimos este campo con el valor de la fecha. Se decide realizar esto porque se trata de un solo registro. También limitamos la edad de los atletas, solo serán edades válidas para los nacidos entre el 1954 y 2002, aquellos que aparezcan con fecha superior o inferior a este intervalo se descartan. Comprobamos esto y no vemos que ocurra.

Medallas: este campo no existe. Se compone de los campos gold, silver y bronze, si alguno de los campos contiene un valor diferente a cero ponemos una “T” para denotar que ha logrado una medalla o ponemos una “F” en caso de no ganar ninguna medalla.

Deportes: se trata de una cadena con la descripción de la disciplina de ese deportista. Se comprueba si existen registros vacios o si existen deportes mal escritos. No se da el caso

3. Análisis de los datos.

3.1. Selección de los grupos de datos que se quieren analizar/comparar.

Una vez que tenemos los datos limpios y generamos el nuevo dataset con la información útil. Procedemos a analizar la información que contiene este nuevo dataset.

Por lo tanto para dar respuesta a los datos se va analizar:

Nacionalidad: comprobar el número de atletas por nacionalidad
Sexo: ver el número de hombres frente al de mujeres
Edad: comprobar la edad de los participantes, el rango mínimo y máximo, la media, …
Disciplina: comprobar la participación en cada deporte
Medallas: calcular el total de medallas entregadas

Luego se va a estudiar los grupos de datos:

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

3.2. Comprobación de la normalidad y homogeneidad de la varianza. Si es necesario (y posible), aplicar transformaciones que normalicen los datos.

Comprobamos la varianza de la edad con el factor del sexo

```
> #---Varianza edad - sexo
> t.test(edad~sexo)

        Welch Two Sample t-test

data:  edad by sexo
t = -9.126, df = 11233, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.108154 -0.716284
sample estimates:
mean in group F mean in group M
       26.19673        27.10895

> var.test(edad~sexo)

        F test to compare two variances

data:  edad by sexo
F = 0.93747, num df = 5204, denom df = 6332, p-value = 0.01483
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.8901456 0.9874442
sample estimates:
ratio of variances
          0.937471

>
```

Comprobamos que es homogénea y no varía en función del sexo

3.3. Aplicación de pruebas estadísticas (tantas como sea posible) para comparar los grupos de datos.

Vamos a analizar cada grupo de datos para determinar los resultados y así poder sacar posibles conclusiones

Por lo tanto a continuación se muestran los resultados de analizar cada dato de nuestro interés:

Edad:

```
> mean(edad)
[1] 26.69743
> median(edad)
[1] 26
> sd(edad)
[1] 5.378624
> var(edad)
[1] 28.9296
> quantile(edad,c(0.25,0.5,0.75))
25% 50% 75%
 23  26  30
> summary(edad)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   14.0    23.0    26.0    26.7    30.0    62.0
```

Sexo

```
summary(sex)
   F    M
5205 6333

Se analiza total de medallas ganadas
> summary(medal)
   Mode   FALSE    TRUE
logical    9681    1857
```

Disciplinas deportivas

```
> sport<-datosNuevos$sport
> summary(sport)
        aquatics          archery         athletics          badminton       basketball
            1445              128               2363              172              288
          boxing           canoe             cycling         equestrian          fencing
             286              331               525               222              246
        football            golf          gymnastics          handball          hockey
             611              120               324               363              432
           judo modern pentathlon             rowing       rugby sevens          sailing
             392               72               547               300              380
        shooting      table tennis          taekwondo            tennis        triathlon
             390              172               128               196              110
      volleyball     weightlifting          wrestling
             384              258               353
> |
```

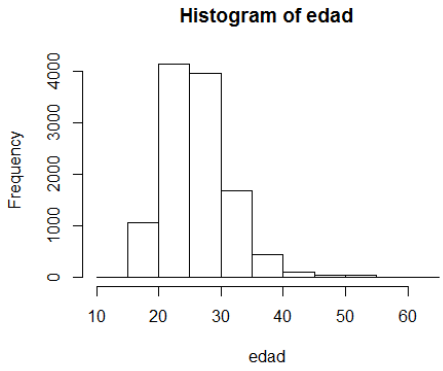
Medallas:

```
summary(medal)
   Mode   FALSE    TRUE
logical    9681    1857
```

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

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

Frecuencias de edad



Frecuencia de sexo

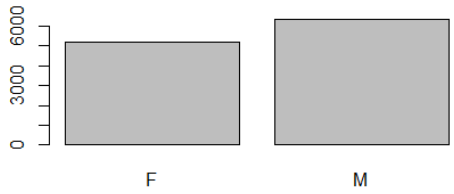
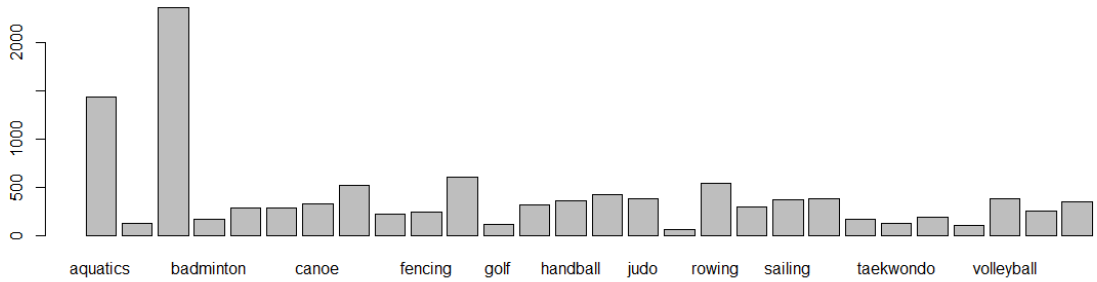


Tabla de deportes



Frecuencia de la nacionalidad

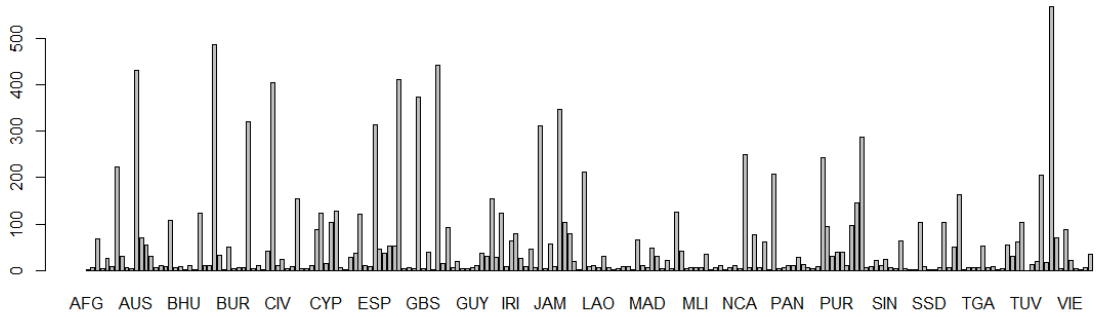
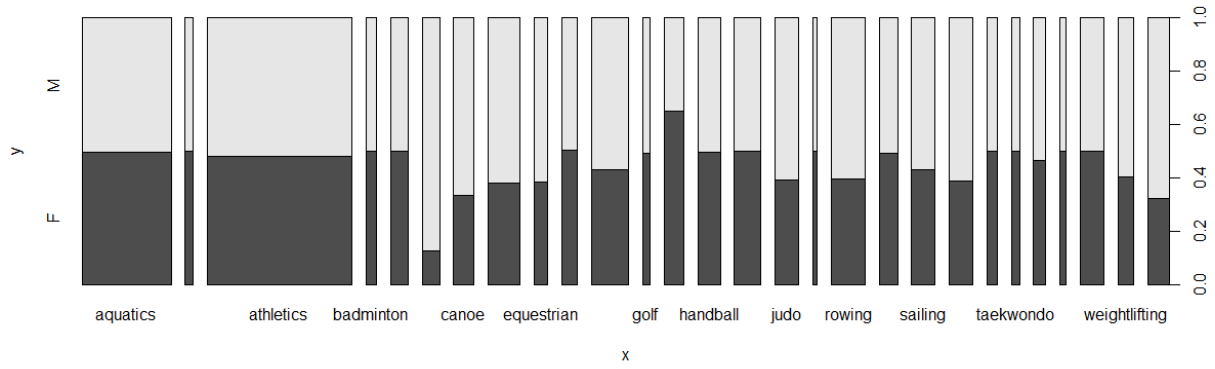
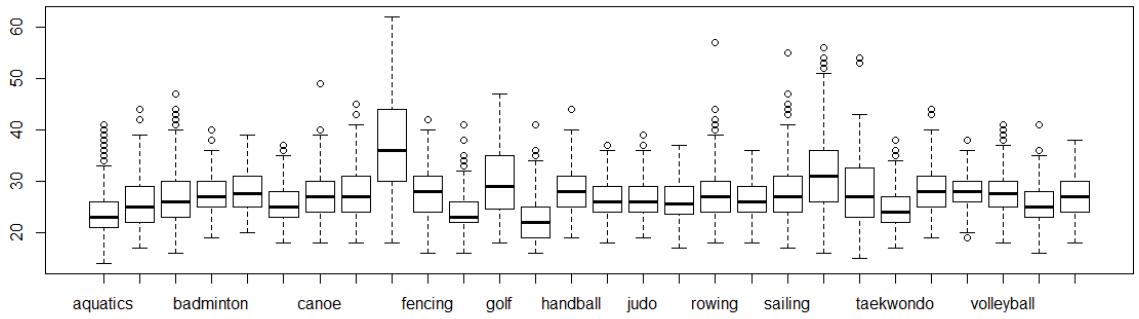


Tabla en función el sexo y deporte

```
> tabla<-table(sexo,sport)
> tabla
    sport
sexo aquatics archery athletics badminton basketball boxing canoe cycling equestrian fencing football golf gymnastics handball
   F      716      64      1137        86        144     36   111      200         85     124      264   59         210      180
   M      729      64      1226        86        144    250   220      325        137     122      347   61         114      183
    sport
sexo hockey judo modern pentathlon rowing rugby sevens sailing shooting table tennis taekwondo tennis triathlon volleyball
   F    216  153                36    216          148     163      151            86         64     91         55         192
   M    216  239                36    331          152     217      239            86         64    105         55         192
    sport
sexo weightlifting wrestling
   F            104       114
   M            154       239
```



Información de las edades en función del deporte



Medallas recibidas por nacionalidad

```
> tabla5<-table(medal,nac)
> tabla5
        nac
medal   AFG ALB ALG AND ANG ANT ARG ARM ARU ASA AUS AUT AZE BAH BAN BAR BDI BEL BEN BER BHU BIH BIZ BLR BOL BOT BRA BRN BRU BUL BUR CAF CAM CAN CAY
  FALSE   3   6  67   5  26   9 201  28   7   4 360  69  38  24   7  11   8  87   6   8   2  11   3 112  12  12 436  32   3  43   5   6   6 259   5
  TRUE    0   0   1   0   0   0  22   4   0   0  71   2  18   6   0   0   1  21   0   0   0   0   0  12   0   0  49   2   0   7   0   0   0  62   0
        nac
medal   CGO CHA CHI CHN CIV CMR COD COK COL COM CPV CRC CRO CUB CYP CZE DEN DJI DMA DOM ECU EGY ERI ESA ESP EST ETH FIJ FIN FRA FSM GAB GAM GBR GBS
  FALSE  10   2  42 304  10  24   4   9 146   4   5  11  64 112  16  90  89   7   2  28  38 119  12   8 270  42  31  41  53 318   5   6   4 244   5
  TRUE    0   0   0 100   2   0   0   0   8   0   0   0  24  11   0  14  39   0   0   1   0   3   0   0  43   4   7  13   1  92   0   0   0 130   0
        nac
medal   GEO GEQ GER GHA GRE GRN GUA GUI GUM GUY HAI HKG HON HUN INA IND IOA IRI IRL IRQ ISL ISR ISV ITA IVB JAM JOR JPN KAZ KEN KGZ KIR KOR KOS KSA
  FALSE  33   2 290  16  87   6  21   5   5   6  10  38  30 139  24 121   7  56  77  26   8  45   7 243   4  33   7 288  86  68  19   3 190   7  11
  TRUE    7   0 151   0   6   1   0   0   0   0   0   0   0  15   4   2   2   8   3   0   0   2   0  69   0  24   1  58  17  12   0   0  23   1   0
        nac
medal   LAO LAT LBA LBR LCA LES LIB LIE LTU LUX MAD MAR MAS MAW MDA MDV MEX MGL MHL MKD MLI MLT MNE MON MOZ MRI MTN MYA NAM NCA NED NEP NGR NIG NOR
  FALSE   6  32   7   2   5   8   9   3  60  10   6  48  24   5  22   4 121  41   5   6   6   7  35   3   6  11   2   7  10   5 203   7  60   5  43
  TRUE    0   0   0   0   0   0   0   0   7   0   0   1   8   0   1   0   5   2   0   0   0   0   0   0   0   0   0   0   0   0  46   0  18   1  19
        nac
medal   NRU NZL OMA PAK PAN PAR PER PHI PLE PLW PNG POL POR PRK PUR QAT ROT ROU RSA RUS RWA SAM SEN SEY SIN SKN SLE SLO SMR SOL SOM SRB SRI SSD STP
  FALSE   2 173   4   7  10  11  29  12   6   5   8 226  94  24  39  38  10  81 124 183   7   8  22  10  24   7   4  59   5   3   2  50   9   3   3
  TRUE    0  35   0   0   0   0   0   1   0   0   0  16   1   7   1   1   0  17  22 103   0   0   0   0   1   0   0   4   0   0   0  53   0   0   0
        nac
medal   SUD SUI SUR SVK SWE SWZ SYR TAN TGA THA TJK TKM TLS TOG TPE TTO TUN TUR TUV UAE UGA UKR URU USA UZB VAN VEN VIE VIN YEM ZAM ZIM
  FALSE   6  93   6  43 138   2   7   7   7  48   6   9   3   5  51  31  58  95   1  12  21 192  17 356  57   4  85  22   4   3   7  35
  TRUE    0  11   0   8  26   0   0   0   0   6   1   0   0   0   5   1   3   8   0   1   0  13   0 211  13   0   3   1   0   0   0   0
>
```

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

A continuación se exponen todas las conclusiones obtenidas en el apartado anterior y dando respuesta a las preguntas propuestas:

- El número de hombres frente al de las mujeres:

  En esas olimpiadas han participado un total de 11538 atletas. De los cuales 5205 fueron mujeres y 6333 hombres. Una participación de 1000 hombres más frente a las mujeres.

- Las edades de los participantes, el atleta más joven, más mayor, la media de edad…

  La edad de los participantes oscila entre los 14 y los 62 años. Con una media de edad de 26 años para las mujeres y 27 años para los hombres.

- La participación de los hombres y mujeres en cada disciplina

En cuanto a las disciplinas el atletismo es el que tiene más participantes con 2363 y pentatlón es el que tiene menos participantes con 12.

Todas los deportes están bastantes equilibrados en lo que se refiere al género. Hay algunos deportes como el boxeo, canoa y lucha donde el género masculino predomina con creces frente al menino. Pero por otro lado tenemos la gimnasia donde prevalece por bastantes puntos el género femenino.

En cuanto a las edades por, comprobamos que la disciplina ecuestre y el tiro contienen participantes más mayores, hasta los 62 años. Y en natación es donde tenemos el participante más joven, tan solo 14 años.

- La participación de los atletas por nacionalidad

  La nacionalidad que más participantes ha tenido fue USA con 567 participantes y la que menos fueron Bolivia y Botswana tan solo con 12 participantes.

- La cantidad total de medallas entregadas. Medallas entregadas por país

  En cuanto a las medallas se han entregado un total de 1857. Estados Unidos, China, Rusia, Gran Bretaña y Alemania fueron los países que más medallas ganaron. En nuestro caso ganamos 43 medallas.

Con todos estos datos ya sabemos cómo fueron estos juegos y determinar las estadísticas en función de estos datos para los próximos juegos olímpicos o eventos deportivos.

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