Actividad Semanal de la clase 3

Dataset

En esta libreta vamos a trabajar con un conjunto de datos que contiene los datos de la FIFA de la copa del mundo. Disponéis de los archivos ya descargados en la sección de recursos. Utilizaremos estos tres ficheros:

- WorldCupMatches.csv: Datos de los partidos disputados
- WorldCupPlayers.csv: Datos de los jugadores que han jugado en cada partido
- WorldCups.csv: Datos de las copas disputadas y los resultados

Introducción

Mediante esta libreta vamos a explorar los datos almacenados en estos tres ficheros y a responder preguntas haciendo operaciones y transformaciones con pandas. Además de completar las celdas de código para responder a los ejercicios, cuando así se indique se deberá incluir texto en *markdown* explicando la información o las conclusiones extraídas.

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Cargamos el conjunto de datos WorldCupMatches
df_partidos <- read.csv("data/WorldCupMatches.csv")
```

1. Exploración

Usar las funcionesheadysummarypara explorar el dataframe `df_partidos`. **Documentar en markdown** brevemente la información almacenada en cada dataframe (columnas y tipo de datos).

No es necesario hacer una documentación exhaustiva, tan solo comentar aspectos principales como qué tipo de información almacena el DataFrame, cuántas filas y columnas tiene, y comentar las columnas que se consideren más relevantes.

Comprobamos que hemos cargado correctamente el dataframe con la función `head`

```
head(df_partidos)
```

```
##   Year             Datetime   Stage        Stadium        City Home.Team.Name
## 1 1930 13 Jul 1930 - 15:00  Group 1        Pocitos Montevideo          France
## 2 1930 13 Jul 1930 - 15:00  Group 4 Parque Central Montevideo             USA
## 3 1930 14 Jul 1930 - 12:45  Group 2 Parque Central Montevideo      Yugoslavia
## 4 1930 14 Jul 1930 - 14:50  Group 3        Pocitos Montevideo         Romania
## 5 1930 15 Jul 1930 - 16:00  Group 1 Parque Central Montevideo       Argentina
## 6 1930 16 Jul 1930 - 14:45  Group 1 Parque Central Montevideo           Chile
```

```
##   Home.Team.Goals Away.Team.Goals Away.Team.Name Win.conditions Attendance
## 1               4               1         Mexico                      4444
## 2               3               0        Belgium                     18346
## 3               2               1         Brazil                     24059
## 4               3               1           Peru                      2549
## 5               1               0         France                     23409
## 6               3               0         Mexico                      9249
##   Half.time.Home.Goals Half.time.Away.Goals               Referee
## 1                    3                   0 LOMBARDI Domingo (URU)
## 2                    2                   0      MACIAS Jose (ARG)
## 3                    2                   0    TEJADA Anibal (URU)
## 4                    1                   0  WARNKEN Alberto (CHI)
## 5                    0                   0    REGO Gilberto (BRA)
## 6                    1                   0  CRISTOPHE Henry (BEL)
##                Assistant.1                Assistant.2 RoundID MatchID
## 1    CRISTOPHE Henry (BEL)        REGO Gilberto (BRA)     201    1096
## 2 MATEUCCI Francisco (URU)      WARNKEN Alberto (CHI)     201    1090
## 3  VALLARINO Ricardo (URU)        BALWAY Thomas (FRA)     201    1093
## 4      LANGENUS Jean (BEL)   MATEUCCI Francisco (URU)     201    1098
## 5     SAUCEDO Ulises (BOL) RADULESCU Constantin (ROU)     201    1085
## 6  APHESTEGUY Martin (URU)        LANGENUS Jean (BEL)     201    1095
##   Home.Team.Initials Away.Team.Initials
## 1                FRA                MEX
## 2                USA                BEL
## 3                YUG                BRA
## 4                ROU                PER
## 5                ARG                FRA
## 6                CHI                MEX
```

Los datos de este dataframe contienen información sobre partidos de fútbol jugados en la copa mundial.

Obtenemos información sobre la dimensión del dataframe:

```
numero_columnas = ncol(df_partidos)
numero_filas = nrow(df_partidos)
print(paste("Número de filas: " , numero_filas))
```

```
## [1] "Número de filas:  4572"
```

```
print(paste("Número de columnas: ", numero_columnas))
```

```
## [1] "Número de columnas:  20"
```

Mostramos los datos estadísticos del dataframe

```
summary(df_partidos)
```

```
##       Year        Datetime            Stage             Stadium
##  Min.   :1930   Length:4572        Length:4572        Length:4572
##  1st Qu.:1970   Class :character   Class :character   Class :character
##  Median :1990   Mode  :character   Mode  :character   Mode  :character
##  Mean   :1985
##  3rd Qu.:2002
##  Max.   :2014
##  NA's   :3720
##      City           Home.Team.Name     Home.Team.Goals  Away.Team.Goals
##  Length:4572        Length:4572        Min.   : 0.000   Min.   :0.000
##  Class :character   Class :character   1st Qu.: 1.000   1st Qu.:0.000
```

```
##  Mode  :character   Mode  :character   Median : 2.000   Median :1.000
##                                        Mean   : 1.811   Mean   :1.022
##                                        3rd Qu.: 3.000   3rd Qu.:2.000
##                                        Max.   :10.000   Max.   :7.000
##                                        NA's   :3720     NA's   :3720
##  Away.Team.Name     Win.conditions       Attendance     Half.time.Home.Goals
##  Length:4572        Length:4572        Min.   :  2000   Min.   :0.000
##  Class :character   Class :character   1st Qu.: 30000   1st Qu.:0.000
##  Mode  :character   Mode  :character   Median : 41580   Median :0.000
##                                        Mean   : 45165   Mean   :0.709
##                                        3rd Qu.: 61374   3rd Qu.:1.000
##                                        Max.   :173850   Max.   :6.000
##                                        NA's   :3722     NA's   :3720
##  Half.time.Away.Goals   Referee          Assistant.1        Assistant.2
##  Min.   :0.000        Length:4572        Length:4572        Length:4572
##  1st Qu.:0.000        Class :character   Class :character   Class :character
##  Median :0.000        Mode  :character   Mode  :character   Mode  :character
##  Mean   :0.428
##  3rd Qu.:1.000
##  Max.   :5.000
##  NA's   :3720
##     RoundID            MatchID          Home.Team.Initials Away.Team.Initials
##  Min.   :     201   Min.   :       25   Length:4572        Length:4572
##  1st Qu.:     262   1st Qu.:     1189   Class :character   Class :character
##  Median :     337   Median :     2191   Mode  :character   Mode  :character
##  Mean   :10661773   Mean   : 61346868
##  3rd Qu.:  249722   3rd Qu.: 43950059
##  Max.   :97410600   Max.   :300186515
##  NA's   :3720       NA's   :3720
```

Comentarios a la exploración

- De los 4572 registros que tiene el dataframe, hay 3720 que están vacíos, por lo tanto será necesario hacer una limpieza de estos registros vacíos
- El número de partidos son 852, la media de asistencia es de 45164 personas por partido, siendo el partido con más asistencia registrada con 173850 personas, y el de menos asistencia 2000 personas.
- El primer mundial del que se tienen datos es el de 1930, y el último de 2014.
- La media de goles por partido para el equipo que juega como local es de 1.81, siendo esta media en la primera parte de 0.7 goles por partido.
- La media de goles por partido para el equipo que juega como visitante es de 1.02, siendo esta media en la primera parte de 0.42 goles por partido.
- El máximo de goles en un partido como local es de 10, y el máximo de goles en un partido como visitante es de 7.

2. Limpieza de datos

Una de las cosas que llama la atención es la alta presencia de valores perdidos en el DataFrame `df_partidos`. Cuando trabajamos con datos obtenidos del mundo real siempre nos toparemos con problemas relacionados con la medición, captura o almacenamiento de dicha información.

Localiza las filas con valores perdido. Analizar y **documentar en markdown** a qué se deben estos valores perdidos. Finalmente usar la función drop_na sobre el DataFrame para eliminar los valores perdidos del

DataFrame, y almacena el resultado en `df_partidos` de nuevo.

```
colSums(is.na.data.frame(df_partidos))
```

```
##                 Year             Datetime                Stage
##                 3720                    0                    0
##              Stadium                 City       Home.Team.Name
##                    0                    0                    0
##      Home.Team.Goals      Away.Team.Goals       Away.Team.Name
##                 3720                 3720                    0
##       Win.conditions           Attendance Half.time.Home.Goals
##                    0                 3722                 3720
## Half.time.Away.Goals              Referee          Assistant.1
##                 3720                    0                    0
##          Assistant.2              RoundID              MatchID
##                    0                 3720                 3720
##   Home.Team.Initials   Away.Team.Initials
##                    0                    0
```

Con esta instrucción podemos ver, de forma concisa, que hay 3720 registros que no tienen valores, a parte hay dos partidos que no tienen asistencia registrada seguramente debido a que no se registro esa información.

Los registros nulos es posible que sea un error en la construcción del csv de datos.

Vamos a eliminar estos registros vacíos, con la

```
df_partidos <- drop_na(df_partidos)
```

Vamos a comprobar ahora si se han eliminado correctamente

```
colSums(is.na.data.frame(df_partidos))
```

```
##                 Year             Datetime                Stage
##                    0                    0                    0
##              Stadium                 City       Home.Team.Name
##                    0                    0                    0
##      Home.Team.Goals      Away.Team.Goals       Away.Team.Name
##                    0                    0                    0
##       Win.conditions           Attendance Half.time.Home.Goals
##                    0                    0                    0
## Half.time.Away.Goals              Referee          Assistant.1
##                    0                    0                    0
##          Assistant.2              RoundID              MatchID
##                    0                    0                    0
##   Home.Team.Initials   Away.Team.Initials
##                    0                    0
```

Por lo tanto, el número de registros final es:

```
nrow(df_partidos)
```

```
## [1] 850
```

3. Cargar datos

Crear las variables `df_jugadores` y `df_copas` que contengan los dataframes correspondientes a la lectura de los csv `WorldCupPlayers` y `WorldCups`.

```
df_jugadores = read.csv("data/WorldCupPlayers.csv")
df_copas = read.csv("data/WorldCups.csv")
```

Mostramos los 5 primeros registros de cada dataframe para comprobar que se ha cargado correctamente:

```
head(df_jugadores)
```

```
##   RoundID MatchID Team.Initials          Coach.Name Line.up Shirt.Number
## 1     201    1096           FRA CAUDRON Raoul (FRA)       S            0
## 2     201    1096           MEX    LUQUE Juan (MEX)       S            0
## 3     201    1096           FRA CAUDRON Raoul (FRA)       S            0
## 4     201    1096           MEX    LUQUE Juan (MEX)       S            0
## 5     201    1096           FRA CAUDRON Raoul (FRA)       S            0
## 6     201    1096           MEX    LUQUE Juan (MEX)       S            0
##        Player.Name Position Event
## 1      Alex THEPOT       GK
## 2  Oscar BONFIGLIO       GK
## 3 Marcel LANGILLER           G40'
## 4     Juan CARRENO           G70'
## 5  Ernest LIBERATI
## 6     Rafael GARZA        C
```

```
head(df_copas)
```

```
##   Year     Country     Winner     Runners.Up   Third     Fourth GoalsScored
## 1 1930     Uruguay    Uruguay      Argentina     USA Yugoslavia          70
## 2 1934       Italy      Italy Czechoslovakia Germany    Austria          70
## 3 1938      France      Italy        Hungary  Brazil     Sweden          84
## 4 1950      Brazil    Uruguay         Brazil  Sweden      Spain          88
## 5 1954 Switzerland Germany FR        Hungary Austria    Uruguay         140
## 6 1958      Sweden     Brazil         Sweden  France Germany FR         126
##   QualifiedTeams MatchesPlayed Attendance
## 1             13            18    590.549
## 2             16            17    363.000
## 3             15            18    375.700
## 4             13            22  1.045.246
## 5             16            26    768.607
## 6             16            35    819.810
```

4. Rango temporal

El dataframe `df_copas` contiene datos de todos los mundiales disputados. ¿Cuál es el año del mundial más antiguo disputado? ¿Y el año del mundial más reciente?

Pista: recuerda que podemos usar funciones de agregación (`min` `mean`, etc.) directamente sobre columnas. Por ejemplo, el siguiente código nos muestra el mayor número de goles marcado en un mundial.

El ejemplo del enunciado estaba mal, lo he corregido.

```
max(df_copas$GoalsScored)
```

```
## [1] 171
```

El año del mundial más antigüo es:

```
min(df_copas$Year)
```

```
## [1] 1930
```

El año del mundial más reciente es:

```
max(df_copas$Year)
```

```
## [1] 2014
```

Opcional: Una vez localizados los años de interés (más antiguo y más reciente) visualizar las filas completas correspondientes a cada año usando la función `filter`.

```
df_copas %>%
  filter(df_copas$Year == min(df_copas$Year))
```

```
##   Year Country  Winner Runners.Up Third     Fourth GoalsScored QualifiedTeams
## 1 1930 Uruguay Uruguay  Argentina   USA Yugoslavia          70             13
##   MatchesPlayed Attendance
## 1            18    590.549
```

```
df_copas %>%
  filter(df_copas$Year == max(df_copas$Year))
```

```
##   Year Country  Winner Runners.Up       Third Fourth GoalsScored QualifiedTeams
## 1 2014  Brazil Germany  Argentina Netherlands Brazil         171             32
##   MatchesPlayed Attendance
## 1            64  3.386.810
```