Tarea 1

Adriana Palacio y Jorge Arteaga

18/11/2021

Ejercicio 1

Cree un nuevo dataframe que sea un subconjunto del dataframe original de dfFires. El subconjunto debe contener todos los incendios del Estado de Idaho y las columnas deben ser limitadas para que sólo estén presentes las columnas YEAR_, CAUSE y TOTALACRES. Cambie el nombre de las columnas. Agrupe los datos por CAUSE y YEAR_ y luego resuma por el total de acres quemados. Trazar los resultados.

Primero, cargamos las librerias necesarias

```
library(readr) #leer archivos
library(dplyr) #manipular dataframe
library(ggplot2) #graficar
```

Procedemos a cargar el archivo "StudyArea.csv" que contiene datos de incendios forestales de los años 1980-2016 para los estados de California, Oregón, Washington, Idaho, Montana, Wyoming, Colorado, Utah, Nevada, Arizona y Nuevo México.

```
dfFires = read_csv("DataSets/StudyArea.csv", col_types = list(UNIT = col_character()),
col_names = TRUE)
head(dfFires)
```

```
## # A tibble: 6 x 14
##     FID ORGANIZATI UNIT  SUBUNIT SUBUNIT2        FIRENAME CAUSE YEAR_ STARTDATED
##   <dbl> <chr>      <chr> <chr>   <chr>           <chr>    <chr> <dbl> <chr>
## 1     0 FWS        81682 USCADBR San Diego Bay ~ PUMP HO~ Human  2001 1/1/01 0:~
## 2     1 FWS        81682 USCADBR San Diego Bay ~ I5       Human  2002 5/3/02 0:~
## 3     2 FWS        81682 USCADBR San Diego Bay ~ SOUTHBAY Human  2002 6/1/02 0:~
## 4     3 FWS        81682 USCADBR San Diego Bay ~ MARINA   Human  2001 7/12/01 0~
## 5     4 FWS        81682 USCADBR San Diego Bay ~ HILL     Human  1994 9/13/94 0~
## 6     5 FWS        81682 USCADBR San Diego Bay ~ IRRIGAT~ Human  1994 4/22/94 0~
## # ... with 5 more variables: CONTRDATED <chr>, OUTDATED <chr>, STATE <chr>,
## #   STATE_FIPS <dbl>, TOTALACRES <dbl>
```

```
nrow(dfFires)
```

```
## [1] 439362
```

Para el análisis solo se necesitan el año, la causa y las acres quemadas de los incendios del estado de Idaho, por lo que se filtrarán las columnas y los datos y se aprovechará para establecer nombres más amigables.

```
dfFires = filter(dfFires, STATE=="Idaho") %>%
  select("YR" = "YEAR_", "CAUSE", "ACRES" ="TOTALACRES")

nrow(dfFires)
```

```
## [1] 36510
```

Con el proceso anterior, pasamos de tener 439.362 registros y 14 variables, a 36.510 registros y 3 variables. Ahora, procedemos a agrupar los incendios forestales del estado de Idaho por causa y año.

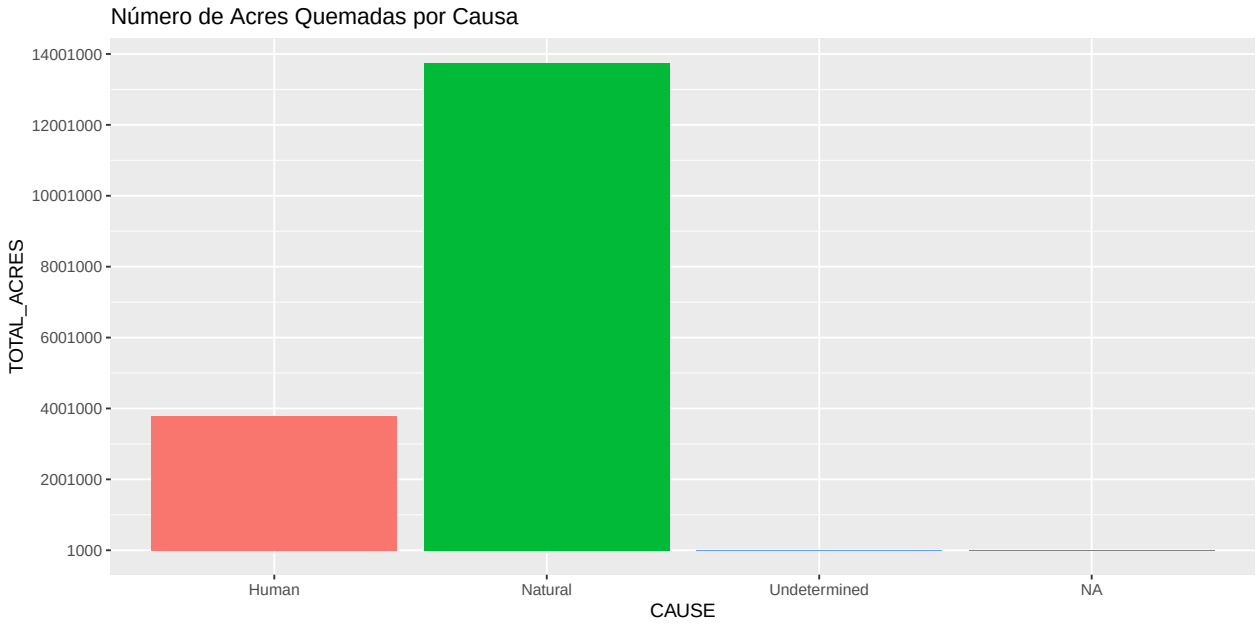
```
sm = group_by(dfFires, YR, CAUSE) %>%
  summarize(sum(ACRES))
```

```
names(sm) <- c("YEAR", "CAUSE", "TOTAL_ACRES")
knitr::kable(head(sm,10))
```

YEAR	CAUSE	TOTAL_ACRES
1980	Human	71974.7
1980	Natural	42938.2
1980	NA	50.0
1981	Human	219362.4
1981	Natural	276593.3
1981	NA	43.0
1982	Human	34016.2
1982	Natural	42156.9
1982	NA	4.0
1983	Human	48242.5

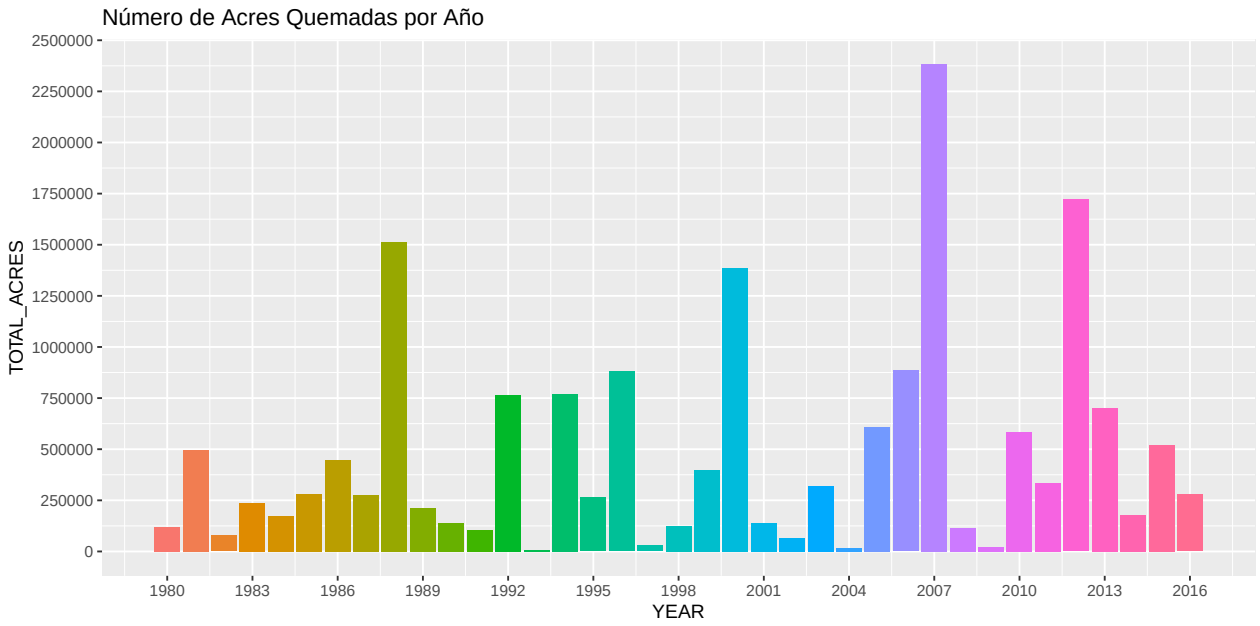
El mayor número de acres quemadas registrados en los años 1980-2016 se tiene cuando el incendio es por causa natural con un valor alrededor de los 14 millones y es muy bajo cuando las causas son indeterminadas, por debajo de 4 mil.

```
ggplot(data=sm, aes(x=CAUSE, y=TOTAL_ACRES, fill=CAUSE))+
  geom_bar(stat="identity")+
  ggtitle ("Número de Acres Quemadas por Causa")+
  scale_y_continuous(breaks=seq(1000, 14001000, 2000000))+
  theme(legend.position = "none")
```



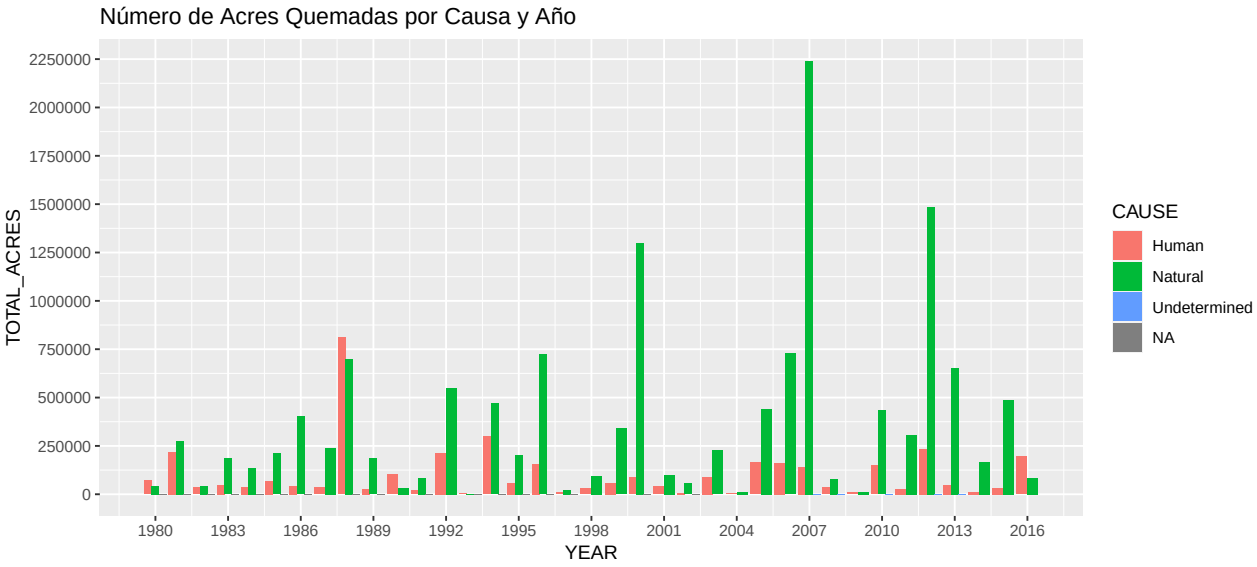
Si revisamos los años, vemos que hay un valor significativamente alto en el 2007 (alrededor de 2.3 millones de acres quemadas), seguido del año 2012 (1.7 millones) y valores bajos en los años 1993, 1997, 2004 y 2009 (por debajo de 31 mil acres).

```
ggplot(data=sm, aes(x=YEAR, y=TOTAL_ACRES, fill=factor(YEAR)))+
  geom_bar(stat="identity")+
  ggtitle ("Número de Acres Quemadas por Año")+
  scale_x_continuous(breaks=seq(1980, 2016, 3))+
  scale_y_continuous(breaks=seq(0, 2500000, 250000))+
  theme(legend.position = "none")
```



Por último, realizando un análisis conjunto entre el comportamiento de los años y la causa, en cuanto a las acres quemadas, vemos que lo usual es que los valores más altos se den cuando el incendio es por causa natural en lugar de por causa humana, excepto para los años 1980, 1988, 1990 y 2016 donde la tendencia es inversa.

```
ggplot(data=sm, aes(x=YEAR, y=TOTAL_ACRES, fill = CAUSE))+
  ggtitle ("Número de Acres Quemadas por Causa y Año")+
  scale_x_continuous(breaks=seq(1980, 2016, 3))+
  scale_y_continuous(breaks=seq(0, 2500000, 250000))+
  geom_bar(stat="identity", position="dodge")
```



Ejercicio 2

Identificar los cinco deportes más importantes según el mayor número de medallas otorgadas en el año 2016 y luego realizar el siguiente análisis:

1. Genere un gráfico que indique el número de medallas concedidas en cada uno de los cinco principales deportes en 2016.
2. Trace un gráfico que represente la distribución de la edad de los ganadores de medallas en los cinco principales deportes en 2016.
3. Descubre qué equipos nacionales ganaron el mayor número de medallas en los cinco principales deportes en 2016.
4. Observe la tendencia del peso medio de los atletas masculinos y femeninos ganadores en los cinco principales deportes en 2016.

Trabajaremos con el conjunto de datos de 120 años de historia olímpica, para esto primero cargamos las librerías necesarios y posteriormente el archivo de datos:

```
library(readr) #leer archivos
library(dplyr) #manipular dataframe
library(ggplot2) #graficar
df = read_csv("DataSets/athlete_events.csv",col_names = TRUE)
```

```
head(df)
```

```
## # A tibble: 6 x 15
##      ID Name    Sex     Age Height Weight Team   NOC   Games   Year Season City
##   <dbl> <chr>   <chr> <dbl>  <dbl>  <dbl> <chr>  <chr> <chr>  <dbl> <chr>  <chr>
## 1     1 A Diji~ M        24    180     80 China  CHN   1992 ~  1992 Summer Barc~
## 2     2 A Lamu~ M        23    170     60 China  CHN   2012 ~  2012 Summer Lond~
## 3     3 Gunnar~ M        24     NA     NA Denma~ DEN   1920 ~  1920 Summer Antw~
## 4     4 Edgar ~ M        34     NA     NA Denma~ DEN   1900 ~  1900 Summer Paris
## 5     5 Christ~ F        21    185     82 Nethe~ NED   1988 ~  1988 Winter Calg~
## 6     5 Christ~ F        21    185     82 Nethe~ NED   1988 ~  1988 Winter Calg~
## # ... with 3 more variables: Sport <chr>, Event <chr>, Medal <chr>
```

```
nrow(df)
```

```
## [1] 271116
```

Solo nos interesa los registros correspondientes a los deportistas que ganaron una medalla en el año 2016, entonces procedemos a extraerlo del conjunto de datos:

```
unique(df$Medal)
```

```
## [1] NA       "Gold"   "Bronze" "Silver"
```

```
df_2016 = filter(df, Year == 2016, !is.na(Medal))
nrow(df_2016)
```

```
## [1] 2023
```

Hemos seleccionado 13.688 registros de 2.023. Para identificar los 5 deportes más importantes, se hace necesario contar cuantas medallas fueron ganadas por deporte y ordenar de manera descendente por número de medallas.

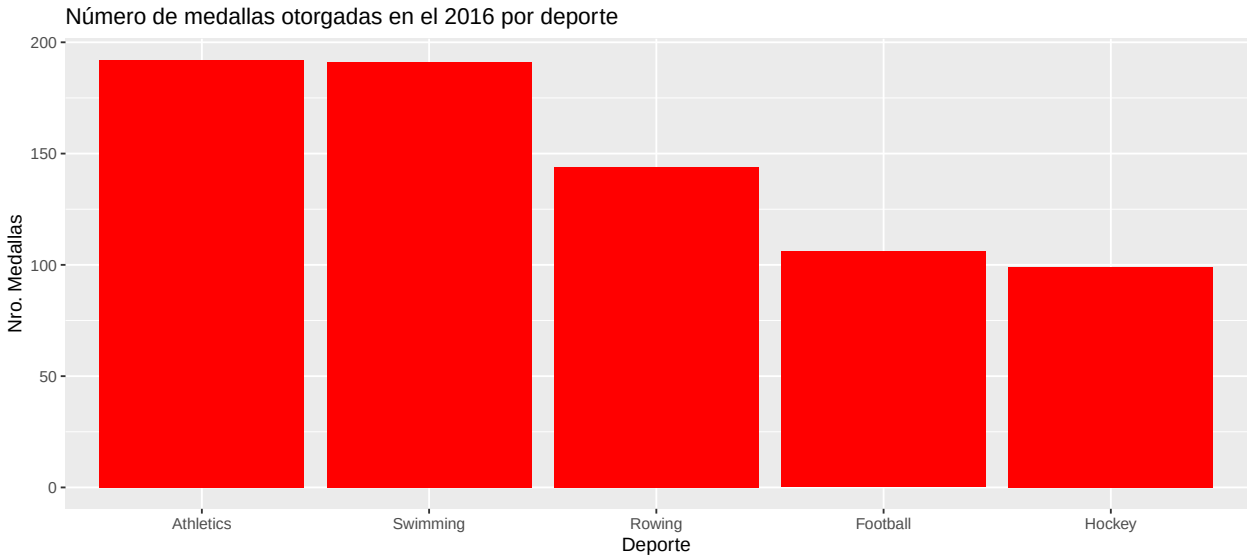
```
df_2016 %>%
  group_by(Sport) %>%
  summarise(n = n()) %>%
  arrange(desc(n)) -> df_top5

df_top5 = df_top5[1:5,]
```

```
knitr::kable(head(df_top5))
```

Sport	n
Athletics	192
Swimming	191
Rowing	144
Football	106
Hockey	99

Entonces, los cinco deportes más importantes de acuerdo al número de medallas otorgadas en el 2016 son: Atletismo (Athletics), Natación (Swimming), Remo (Rowing), Fútbol (Football) y Hockey.

```
ggplot(data = df_top5) +
  geom_col(aes(x=reorder(Sport, -n), y=n), fill="red")+
  xlab("Deporte") + ylab("Nro. Medallas") +
  ggtitle("Número de medallas otorgadas en el 2016 por deporte")
```



De nuestro conjunto de ganadores de medallas de 2016 (df_2016), tomamos solo aquellos correspondiente a los mejores 5 deportes, que corresponden a 732 registros. Para estos, vemos que el ganador con menor edad tenía 16 y el de mayor edad 40. La edad promedio de ganadores de medallas, es de 25-26 años, el 50% de los jugadores tienen 25 años o menos y solo el 25% tienen una edad mayor a 29 años y existe un dato extremo en 40 años.

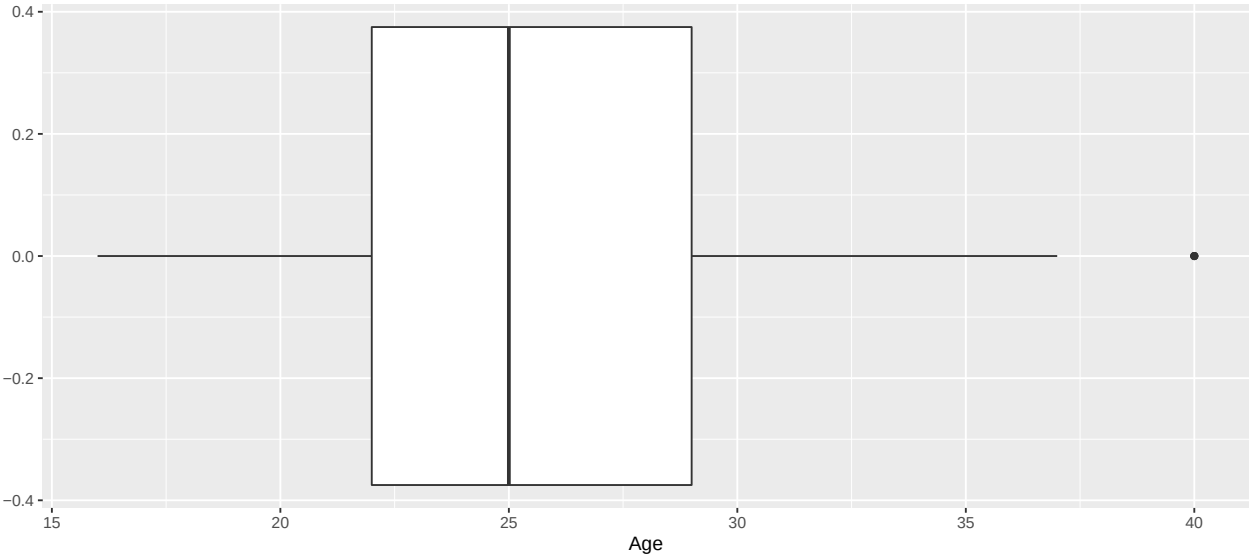
```
df_best = filter(df_2016, Sport %in% df_top5$Sport)
nrow(df_best)
```

```
## [1] 732
```

```
summary(df_best$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   16.00   22.00   25.00   25.58   29.00   40.00
```

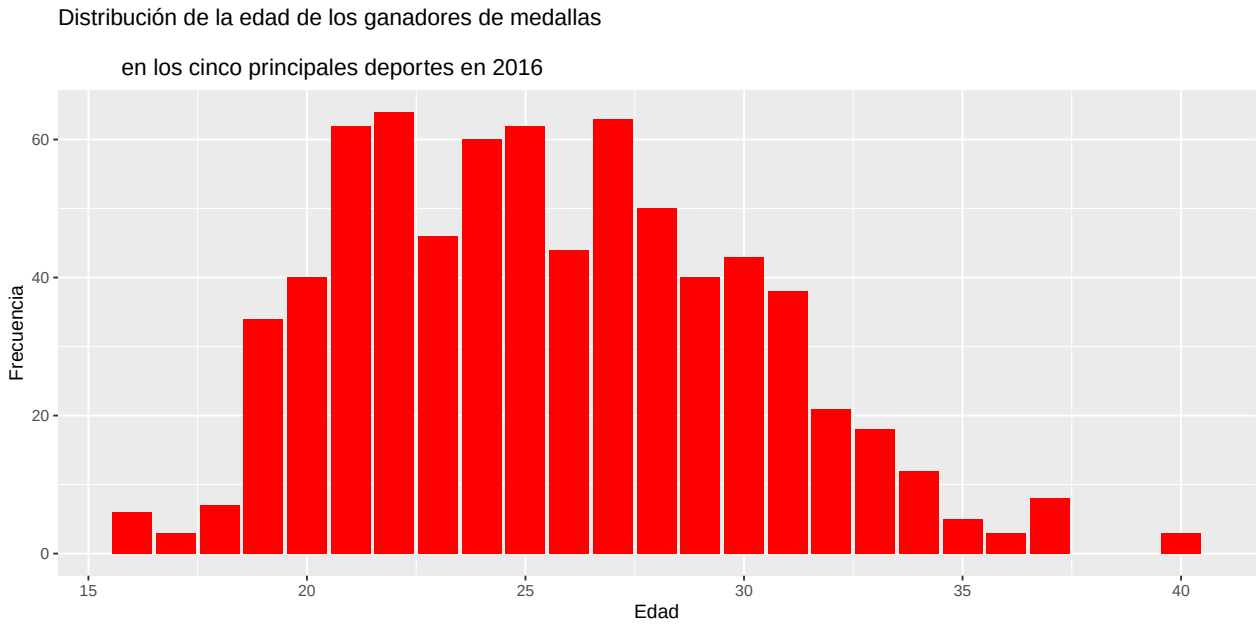
```
df_best %>%
  ggplot(mapping = aes(x = Age)) + geom_boxplot()
```



En el gráfico siguiente, se puede apreciar la distribución de la edad de los ganadores de medallas en los cinco

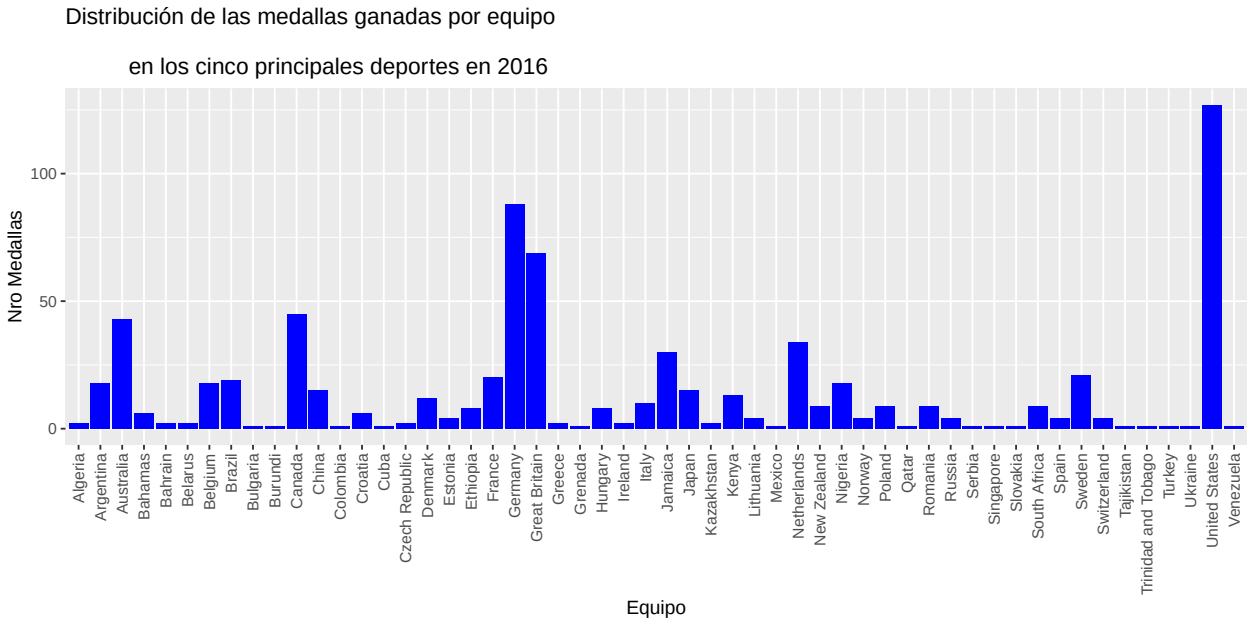
principales deportes en 2016. Los datos parecen estar distribuidos simetricamente sin sesgo aparente, teniendo en cuenta que la media y la mediana son muy cercanas.

```
ggplot(data = df_best) +
  geom_bar(aes(x=Age), fill="red")+
  xlab("Edad") + ylab("Frecuencia") +
  ggtitle("Distribución de la edad de los ganadores de medallas \n
          en los cinco principales deportes en 2016")
```



Ahora bien, si revisamos la distribución de medallas ganadas por equipo en los cinco deportes principales en 2016, tenemos que el mayor número de medallas lo ganó Estados Unidos con 127, seguido de Alemania con 88 y Gran Bretaña con 69.

```
df_best %>%
  group_by(Team) %>%
  ggplot(mapping = aes(x=Team)) +
  geom_bar(fill="Blue")+
  theme(axis.text.x = element_text(angle = 90, vjust=0.5, hjust=1))+
  xlab("Equipo") + ylab("Nro Medallas") +
  ggtitle("Distribución de las medallas ganadas por equipo \n
          en los cinco principales deportes en 2016")
```



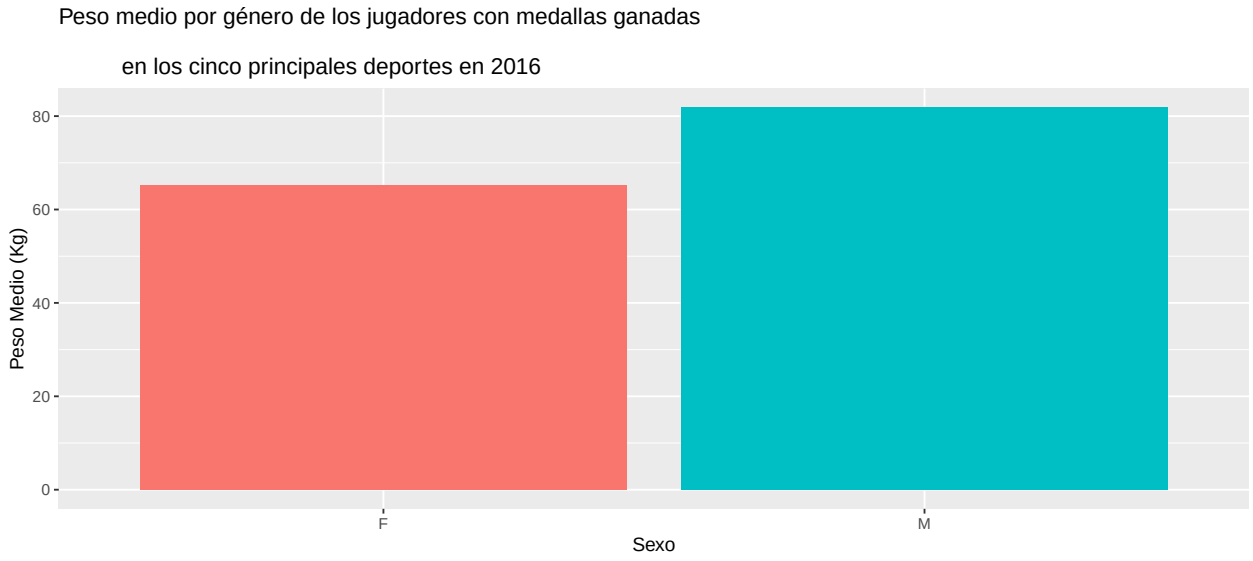
Para el caso de Colombia, solo Caterine Ibargüen Mena ganó medallas en los cinco principales deportes en 2016, su medalla fue de oro en el evento de salto triple para mujer.

```
knitr::kable(head(filter(df_best, Team == 'Colombia')))
```

ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
51514	Caterine Ibargn Mena	F	32	185	70	Colombia	COL	2016 Summer	2016	Summer	Rio de Janeiro	Athletics	Athletics Women's Triple Jump	Gold

Por último cuando analizamos los pesos medios, tenemos que las mujeres que ganaron medallas tienen un peso medio menor que los hombres, con 65.25 y 81.81 Kg respectivamente.

```
df_best%>%
  group_by(Sex) %>%
  summarise(m = mean(Weight, na.rm = TRUE)) %>%
  ggplot(aes(x=Sex, y=m, fill=factor(Sex))) +
  geom_bar(stat="identity")+
  xlab("Sexo") + ylab("Peso Medio (Kg)") +
  ggtitle("Peso medio por género de los jugadores con medallas ganadas \n
          en los cinco principales deportes en 2016")+
  theme(legend.position = "none")
```



Si revisamos está misma distribución del peso teniendo en cuenta los 5 principales deportes en 2016 encontramos que el promedio de pesos en las mujeres es menor que en los hombres. Adicionalmente, el peso de las mujeres varía entre 62.58 y 71.76 kg pero se mantiene muy similar (63kg) en los deportes de atletismo, fútbol y Hockey y para los hombres el peso varía entre 75.72 y 88.83 Kg.

```
df_best%>%
  group_by(Sport, Sex) %>%
  summarise(m = mean(Weight, na.rm = TRUE))%>%
  ggplot(aes(x=Sport, y=m, fill=Sex)) +
  geom_bar(stat="identity", position="dodge")+
  xlab("Sexo") + ylab("Peso Medio (Kg)") +
  ggtitle("Peso medio por deporte y género de los jugadores con medallas ganadas
          \n en los cinco principales deportes en 2016")+
  theme()
```

```
## `summarise()` has grouped output by 'Sport'. You can override using the `.groups` argument.
```