Prueba Pasantia JEP

Fabian Castellanos

2024-07-23

En primer lugar se cargan las librerias a utilizar

```
### Librerias
library(readxl)
library(dplyr)
```

```
##
## 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
```

```
library(knitr)
library(ggplot2)
library(scales)
```

Se cargan las bases de datos requeridas para el analísis

```
### Carga de bases

A <- read_excel("C:/Users/fabia/OneDrive - Universidad Nacional de Colombia/UNAL/JEP_Pureba/Prueba_Tecni
B <- read_excel("C:/Users/fabia/OneDrive - Universidad Nacional de Colombia/UNAL/JEP_Pureba/Prueba_Tecni
```

1. Se añade la columna fuente y ademas se identifican las personas que estan en las dos fuentes de la siguiente manera.

```
# Agregar Columna "fuente"

A$fuente<- "A"
B$fuente<- "B"

# Unir los Data frame
General<-rbind(A,B)
str(General)
```

```
## tibble [20 x 12] (S3: tbl_df/tbl/data.frame)
##  $ NOMBRE1          : chr [1:20] "JOSE" "ISMAEL" "CARLOS" "EFREM" ...
##  $ NOMBRE2          : chr [1:20] "JAVIER" "ALBERTO" "ALFONSO" NA ...
##  $ APELLIDO1        : chr [1:20] "HURTADO" "OROZCO" "FLOREZ" "SALDARRIAGA" ...
##  $ APELLIDO2        : chr [1:20] "VERGARA" "RAMIREZ" NA NA ...
##  $ DEPARTAMENTO     : chr [1:20] "CAQUETA" "RISARALDA" "CAQUETA" "ANTIOQUIA" ...
##  $ MUNICIPIO        : chr [1:20] "CARTAGENA DEL CHAIRA" "PEREIRA" "CARTAGENA DEL CHAIRA" "PUERTO BERRI
##  $ FECHA_HECHOS     : POSIXct[1:20], format: "1998-03-03" "1996-08-08" ...
##  $ COD_MUNICIPIO    : num [1:20] 18150 66001 18150 5579 5842 ...
##  $ NUMERO_DOCUMENTO: num [1:20] 1e+07 1e+07 1e+07 1e+07 1e+07 ...
##  $ EDAD             : num [1:20] 22 19 21 NA 25 NA 20 29 NA NA ...
##  $ SEXO             : chr [1:20] "HOMBRE" "HOMBRE" "MUJER" "HOMBRE" ...
##  $ fuente           : chr [1:20] "A" "A" "A" "A" ...
```

Al identicar las personas que se encuentran en las dos fuentes, estas se pueden identificar en la base de datos como "TRUE" en la columna "Identificador".

```
# Crear una columna que indique si el Número de documeto  está repetido
General$Identificador<- duplicated(
  General$NUMERO_DOCUMENTO) | duplicated(General$NUMERO_DOCUMENTO,
                                         fromLast = TRUE)
str(General)
```

```
## tibble [20 x 13] (S3: tbl_df/tbl/data.frame)
##  $ NOMBRE1          : chr [1:20] "JOSE" "ISMAEL" "CARLOS" "EFREM" ...
##  $ NOMBRE2          : chr [1:20] "JAVIER" "ALBERTO" "ALFONSO" NA ...
##  $ APELLIDO1        : chr [1:20] "HURTADO" "OROZCO" "FLOREZ" "SALDARRIAGA" ...
##  $ APELLIDO2        : chr [1:20] "VERGARA" "RAMIREZ" NA NA ...
##  $ DEPARTAMENTO     : chr [1:20] "CAQUETA" "RISARALDA" "CAQUETA" "ANTIOQUIA" ...
##  $ MUNICIPIO        : chr [1:20] "CARTAGENA DEL CHAIRA" "PEREIRA" "CARTAGENA DEL CHAIRA" "PUERTO BERRI
##  $ FECHA_HECHOS     : POSIXct[1:20], format: "1998-03-03" "1996-08-08" ...
##  $ COD_MUNICIPIO    : num [1:20] 18150 66001 18150 5579 5842 ...
##  $ NUMERO_DOCUMENTO: num [1:20] 1e+07 1e+07 1e+07 1e+07 1e+07 ...
##  $ EDAD             : num [1:20] 22 19 21 NA 25 NA 20 29 NA NA ...
##  $ SEXO             : chr [1:20] "HOMBRE" "HOMBRE" "MUJER" "HOMBRE" ...
##  $ fuente           : chr [1:20] "A" "A" "A" "A" ...
##  $ Identificador    : logi [1:20] TRUE FALSE TRUE FALSE FALSE FALSE ...
```

2. Analísis

Se muestra una tabla de contigencia, donde vemos la relación entre Número de identificacion de la persona reportada y la fuente, se indetica que el Número de identificación 10001916 ha sido reportado dos veces en la fuente B, revisando a profundidad las bases de datos, esto se da porque la persona se reporto en dos municipios diferentes del departamento.

```
### Tabla de contingencia

kable(table(General$NUMERO_DOCUMENTO,General$fuente))
```

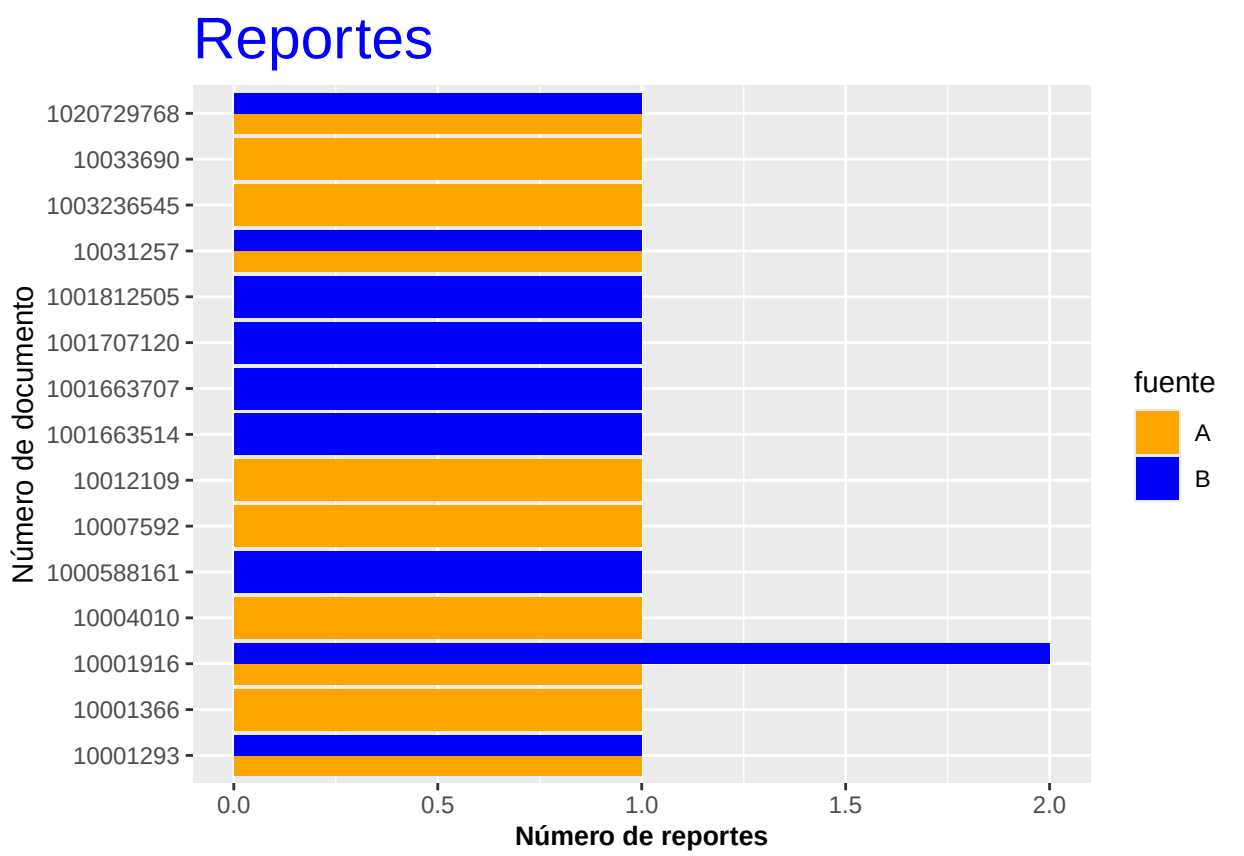
	A	B
10001293	1	1

	A	B
10001366	1	0
10001916	1	2
10004010	1	0
10007592	1	0
10012109	1	0
10031257	1	1
10033690	1	0
1000588161	0	1
1001663514	0	1
1001663707	0	1
1001707120	0	1
1001812505	0	1
1003236545	1	0
1020729768	1	1

En el siguiente gráfico se puede ver de una forma mas detallada, la tabla de contingencia antes presentada.

```
g3 = ggplot(General, aes(as.character(NUMERO_DOCUMENTO), fill=fuente) ) +
  labs(title = "Reportes")+ylab("Número de reportes")+xlab("Número de documento") +
  theme(plot.title = element_text(size = rel(2), colour = "blue"))+ coord_flip()

g3+geom_bar(position="dodge") + scale_fill_manual(values = alpha(c("orange", "blue"), 1)) +
  theme(axis.title.x = element_text(face="bold", size=10))
```



Se genera un gráfico de barras para identificar la frecuencia de reportes por fuente y departamento, el que tiene mas reportes es el departamento de Caqueta.

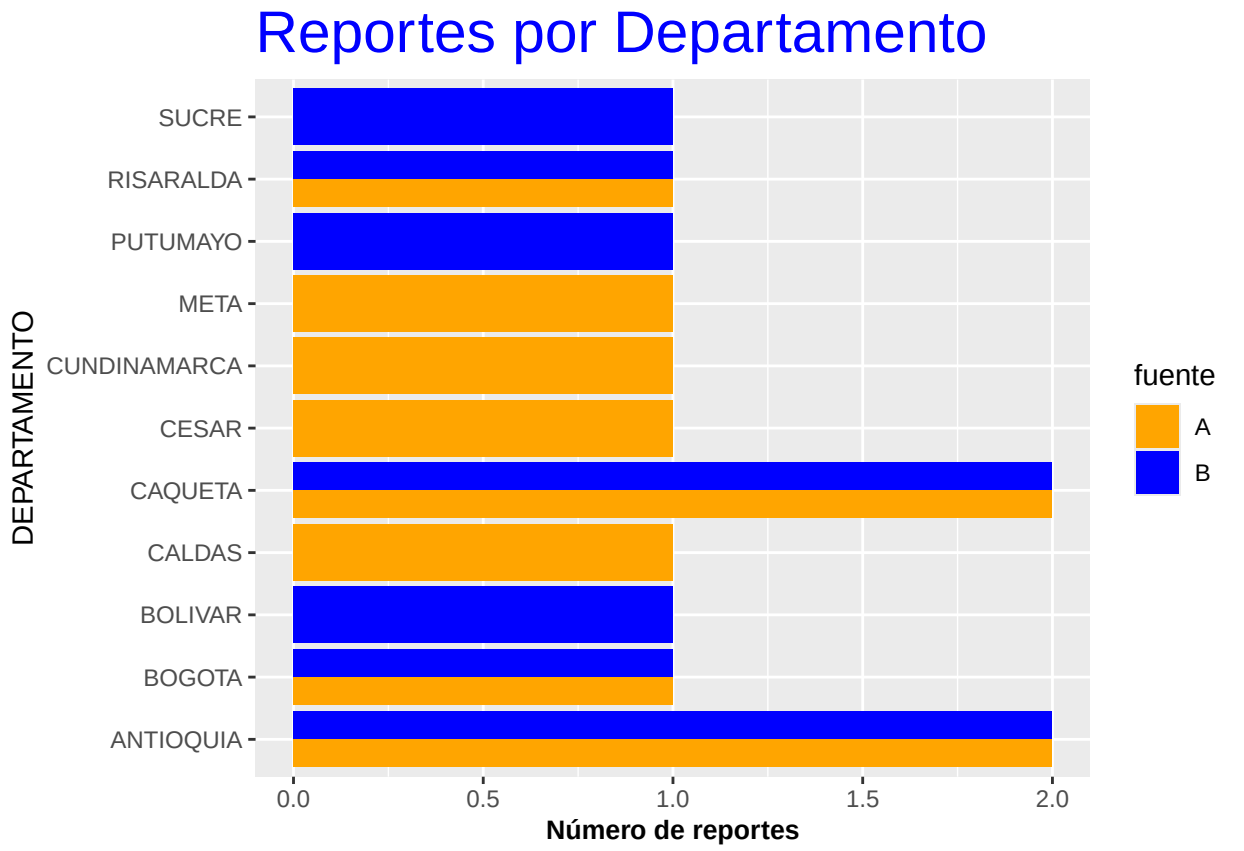
Se elimina el numero de documeto 10001916 una vez en la fuente para que no interfiera en el analisis por departamento.

```
# Departamento


A_sin_duplicados <- A[!duplicated(A$NUMERO_DOCUMENTO), ]
B_sin_duplicados <- B[!duplicated(B$NUMERO_DOCUMENTO), ]
General_dep<-rbind(A_sin_duplicados ,B_sin_duplicados)


g = ggplot(General_dep, aes(DEPARTAMENTO, fill=fuente) ) +
  labs(title = "Reportes por Departamento")+ylab("Número de reportes") +
  theme(plot.title = element_text(size = rel(2), colour = "blue"))+ coord_flip()

g+geom_bar(position="dodge") + scale_fill_manual(values = alpha(c("orange", "blue"), 1)) +
  theme(axis.title.x = element_text(face="bold", size=10))
```



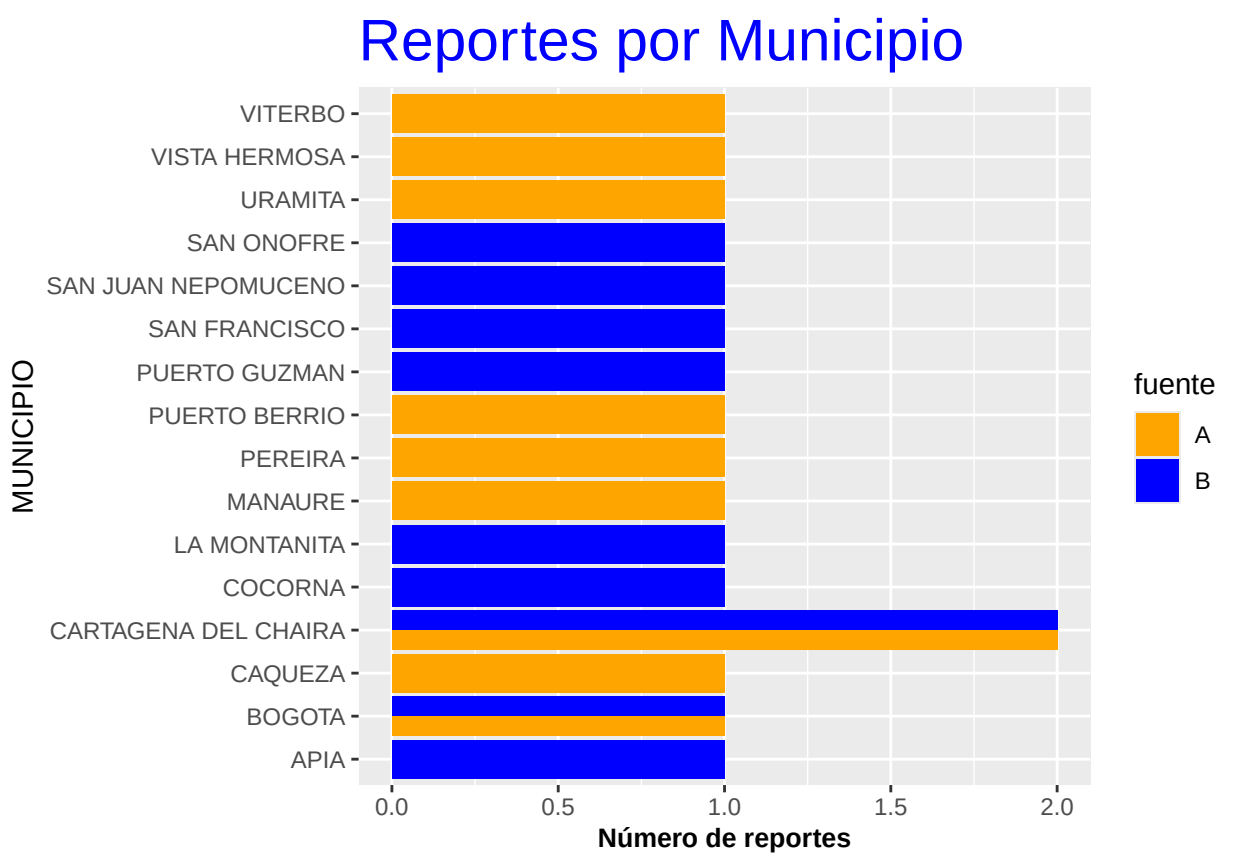
Se genera un gráfico de barras para identificar la frecuencia de reportes por fuente y municipio, el que tiene mas reportes es el municipio de Cartagena del Chaira.

```
# Municipio

g1 = ggplot(General, aes(MUNICIPIO, fill=fuente) ) +
```

```
  labs(title = "Reportes por Municipio")+ylab("Número de reportes") +
  theme(plot.title = element_text(size = rel(2), colour = "blue"))+ coord_flip()

g1+geom_bar(position="dodge") + scale_fill_manual(values = alpha(c("orange", "blue"), 1)) +
  theme(axis.title.x = element_text(face="bold", size=10))
```

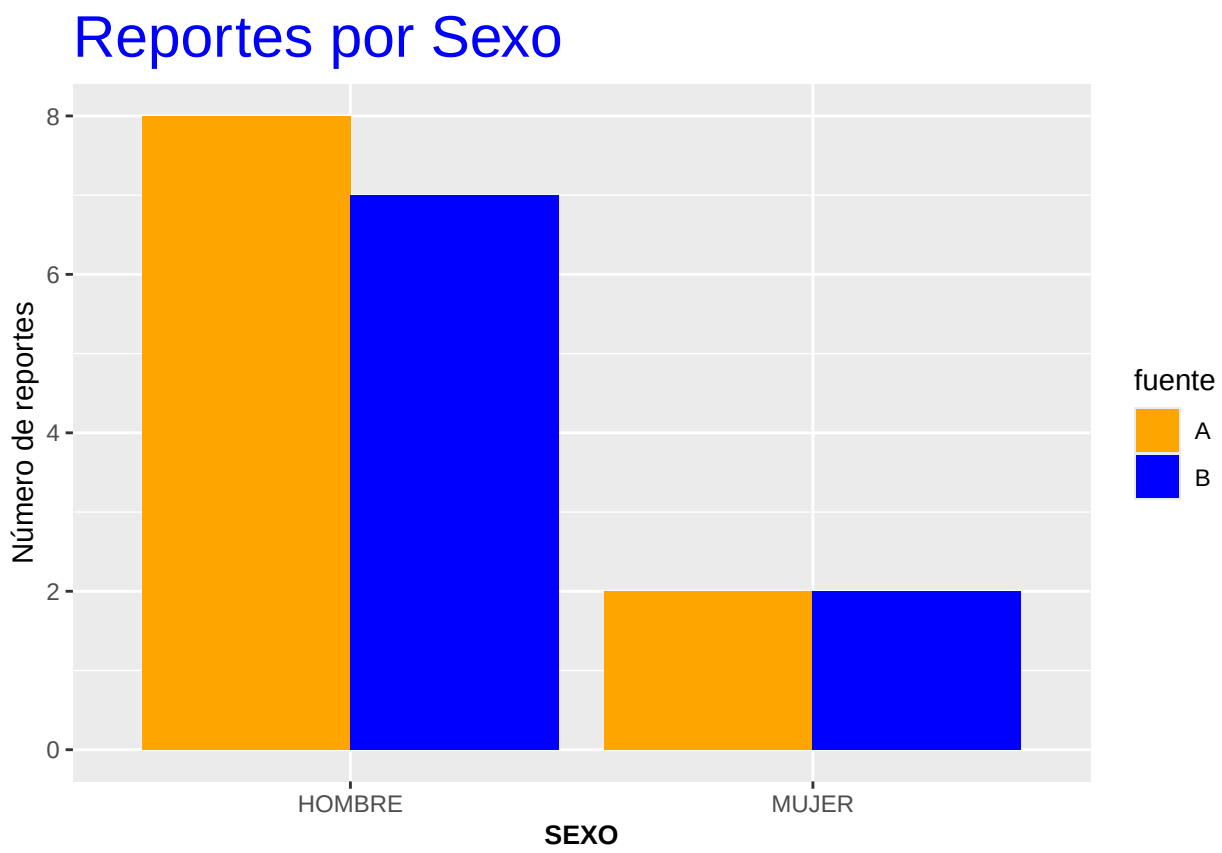


Se genera un gráfico de barras para identificar la frecuencia de reportes por fuente y sexo, el que tiene mas reportes es elsexo masculino.

```
# Sexo

g2 = ggplot(General_dep, aes(SEXO, fill=fuente) ) +
  labs(title = "Reportes por Sexo")+ylab("Número de reportes") +
  theme(plot.title = element_text(size = rel(2), colour = "blue"))

g2+geom_bar(position="dodge") + scale_fill_manual(values = alpha(c("orange", "blue"), 1)) +
  theme(axis.title.x = element_text(face="bold", size=10))
```


Reportes por Sexo
Número de reportes
8
6
4
2
0
HOMBRE
MUJER
SEXO
fuente
A
B