Laboratorio 8

Grupo 6

El objetivo principal de este laboratorio consiste en realizar un análisis exploratorio sobre las importaciones de motos, con el fin de representar toda esta información por medio de gráficos estáticos o dinámicos que faciliten la comprensión y los descubrimientos de la información.

Librerías a utilizar

```
library("tools")
library("lubridate")
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
library(stringr)
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
library(ggplot2)
library(RColorBrewer) #Para colorcitos de graficas
library(wesanderson)
library(dplyr)
```

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

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

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

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

```
theme_set(
  theme_minimal() +
    theme(legend.position = "right")
  )
```

Obtención de los datos

Para este apartado, se descargaron los archivos directamente de internet y se comprimen en la carpeta creada de *datos* en un archivo csv. La primera parte indica todos los enlaces de todos los archivos de importaciones, desde enero 2011 hasta febrero 2019.

```r
#C:/Users/Home/Documents/Tercer año/Sexto Semestre/Data Science/Laboratorio8DS
#/Users/odalisrg/Downloads/datos
#/Users/quiebres/Downloads/datos
setwd("/Users/odalisrg/Downloads/datos")
dataset <- read.csv("importacionesVehiculosSAT.csv",TRUE, ";")
accidentes <- read.csv("accidentesMotos.csv",TRUE, ",")
```

Limpieza de datos

No se considera que sea necesaria una limpieza de datos al haber revisado el archivo .csv generado.

Análisis exploratorio

Para comenzar con el análisis exploratorio se hizo una partición del dataset completo para poder analizar solamente las importaciones y exportaciones de motocicletas.

```r
class(dataset)
```

```
## [1] "data.frame"
```

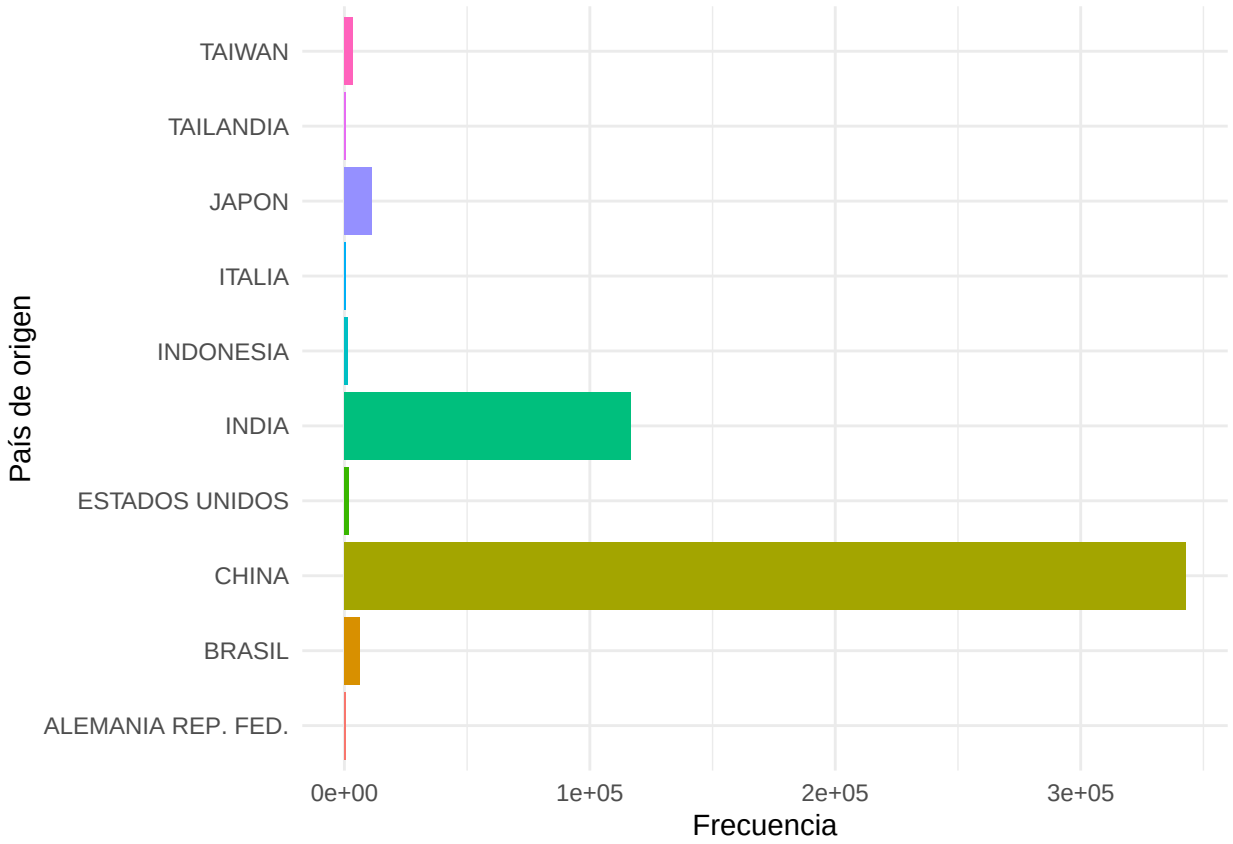
```r
motos <- subset(dataset, Tipo.de.Vehiculo == "MOTO") #Ahora trabajaremos solo con motos
class(motos)
```

```
## [1] "data.frame"
```

Análisis de variables cualitativas.

Vamos a analizar que país es el que más exporta motocicletas hacia Guatemala.

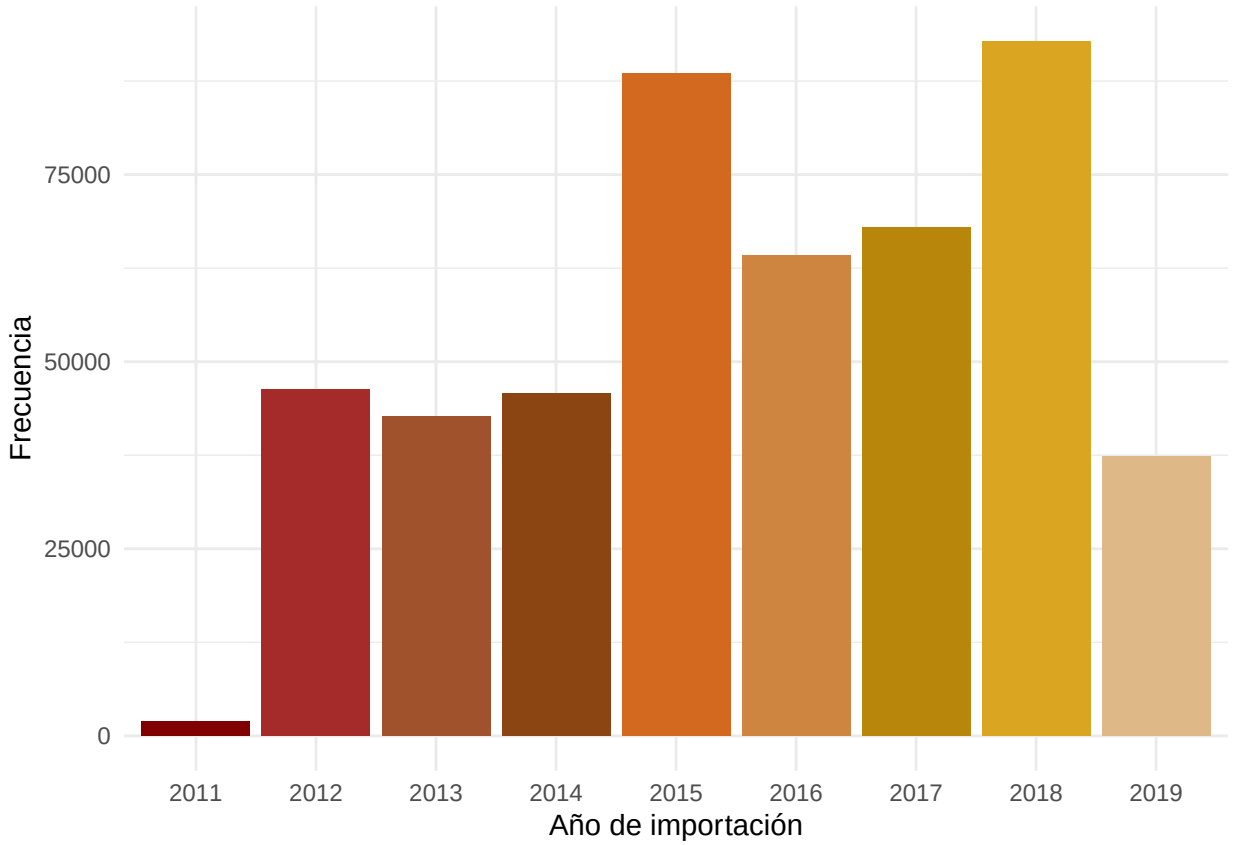
```r
freqPais <- as.data.frame(table(motos$Pais.de.Proveniencia))
 freqPais<- freqPais[order(freqPais$Freq, decreasing = TRUE),]

freqPais <- as.data.frame(head(freqPais,10))

ggplot(freqPais, aes(x=Var1, y=Freq, fill = Var1)) +
  geom_bar(stat = "identity") +
  coord_flip() + xlab("País de origen") +
  ylab("Frecuencia") +
  theme(legend.position = "top") +
  theme(legend.position = "none")
```



Vemos que la mayoría de motos proviene de China con más de 600,000 importaciones en total, seguido por India.

Nos interesa conocer sobre el año en el cual se realizaron mayor cantidad de importaciones, ende se prosiguió de la misma manera:

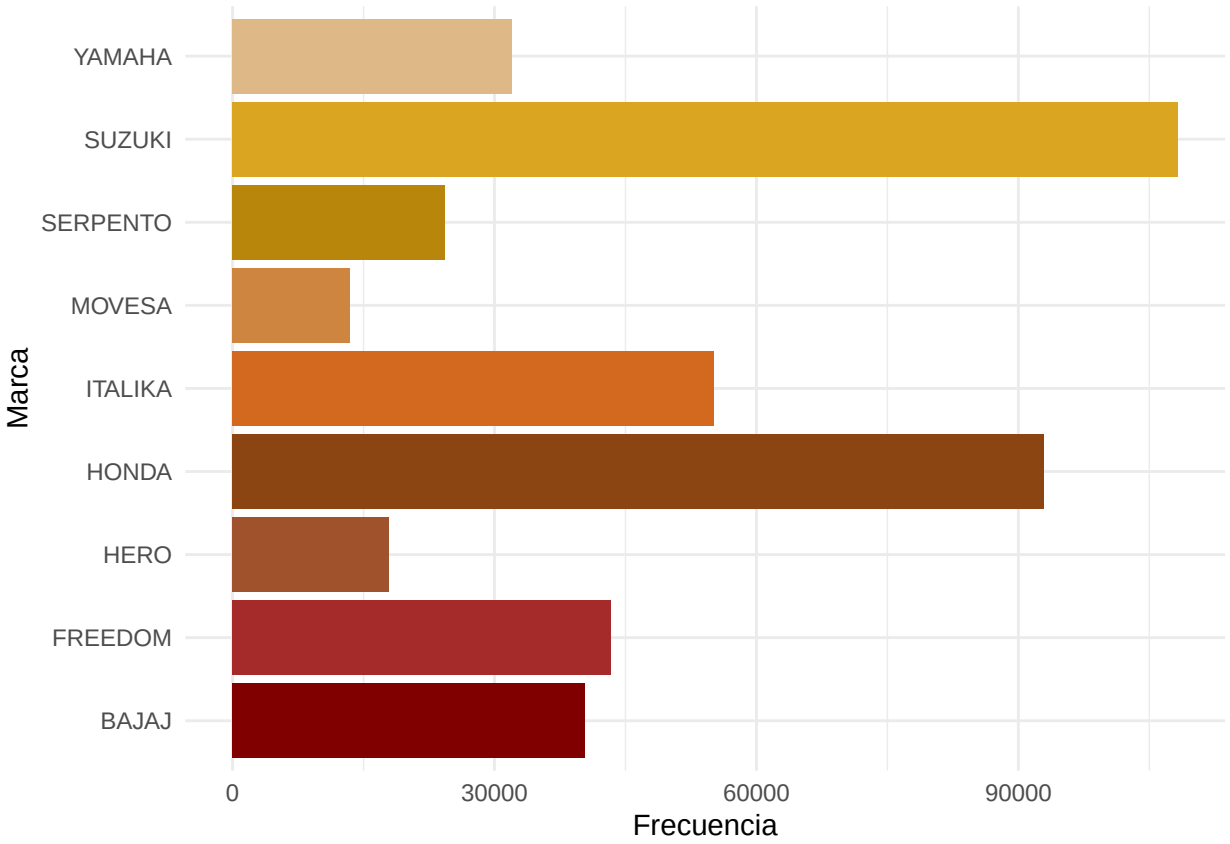
```
freqAnio <- as.data.frame(table(motos$Anio))
 freqAnio<- freqAnio[order(freqAnio$Freq, decreasing = TRUE),]




ggplot(freqAnio, aes(x=Var1, y=Freq, fill = Var1)) +
  geom_bar(stat = "identity") + xlab("Año de importación")+
  ylab("Frecuencia") +
  scale_fill_manual(values = c("#800000", "#A52A2A","#A0522D","#8B4513",
                               "#D2691E", "#CD853F","#B8860B","#DAA520",
                               "#DEB887","#F5DEB3")) +
  theme(legend.position = "none")
```



Se tiene que el año en el que se tuvo mayor cantidad general de registros así el del 2018, seguido para el del 2017.

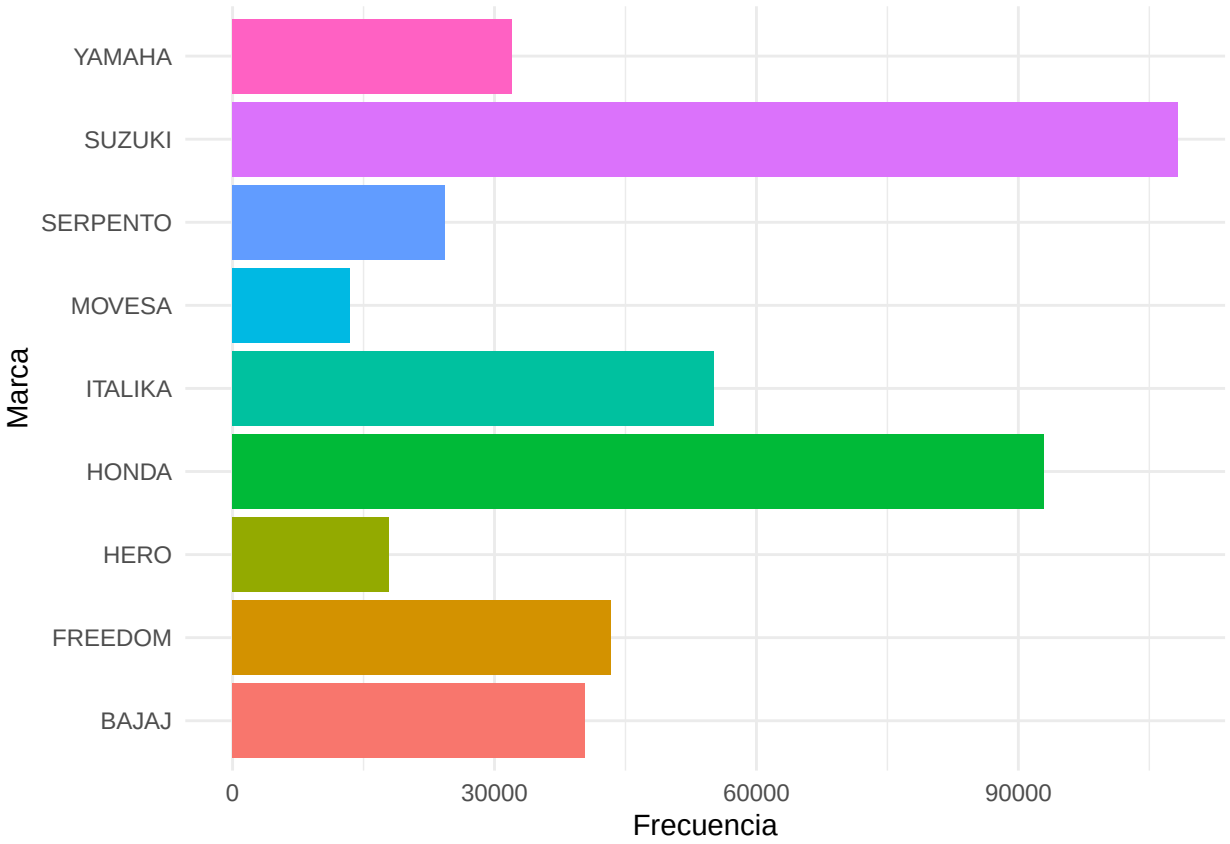
También consideramos importante ver cuál es la marca que predomina en las importaciones.

```
freqMarca <- as.data.frame(table(motos$Marca))
 freqMarca<- freqMarca[order(freqMarca$Freq, decreasing = TRUE),]
freqMarca <- head(freqMarca,9)

ggplot(freqMarca, aes(x=Var1, y=Freq, fill = Var1)) + geom_bar(stat = "identity") +
  coord_flip()+ xlab("Marca") + ylab("Frecuencia") +
    scale_fill_manual(values = c("#800000", "#A52A2A", "#A0522D",
                                 "#8B4513", "#D2691E","#CD853F",
                                 "#B8860B","#DAA520","#DEB887")) +
  theme(legend.position = "none")
```



Dado que es una gran cantidad de marcas, solo analizamos las 10 que poseen mayor cantidad. Observamos que las marcas que predominan son Honda y Suzuki. Ahora, vamos a explorar las marcas que tienen menor cantidad de importaciones.

```
 freqMarca<- freqMarca[order(freqMarca$Freq, decreasing = FALSE),]
freqMarca <- head(freqMarca,20)

ggplot(freqMarca, aes(x=Var1, y=Freq, fill=Var1)) + geom_bar(stat = "identity") +
  coord_flip()+ xlab("Marca")+ ylab("Frecuencia") +
  theme(legend.position = "top") +
  theme(legend.position = "none")
```

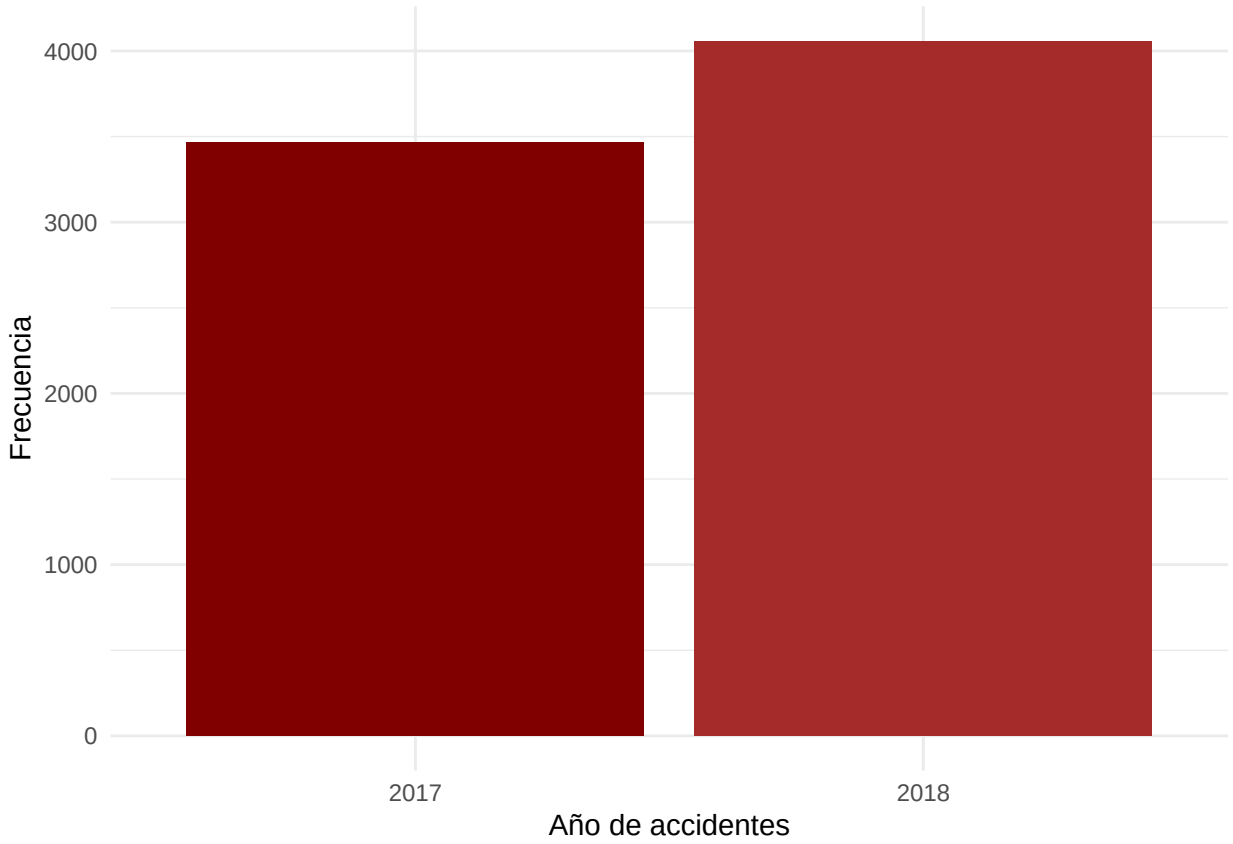


Vemos que hay marcas que poseen una sola importación. Tiene sentido lo que encontramos, dado que estas marcas casi nunca se ven en el tránsito en Guatemala.

Análisis de accidentes de motos

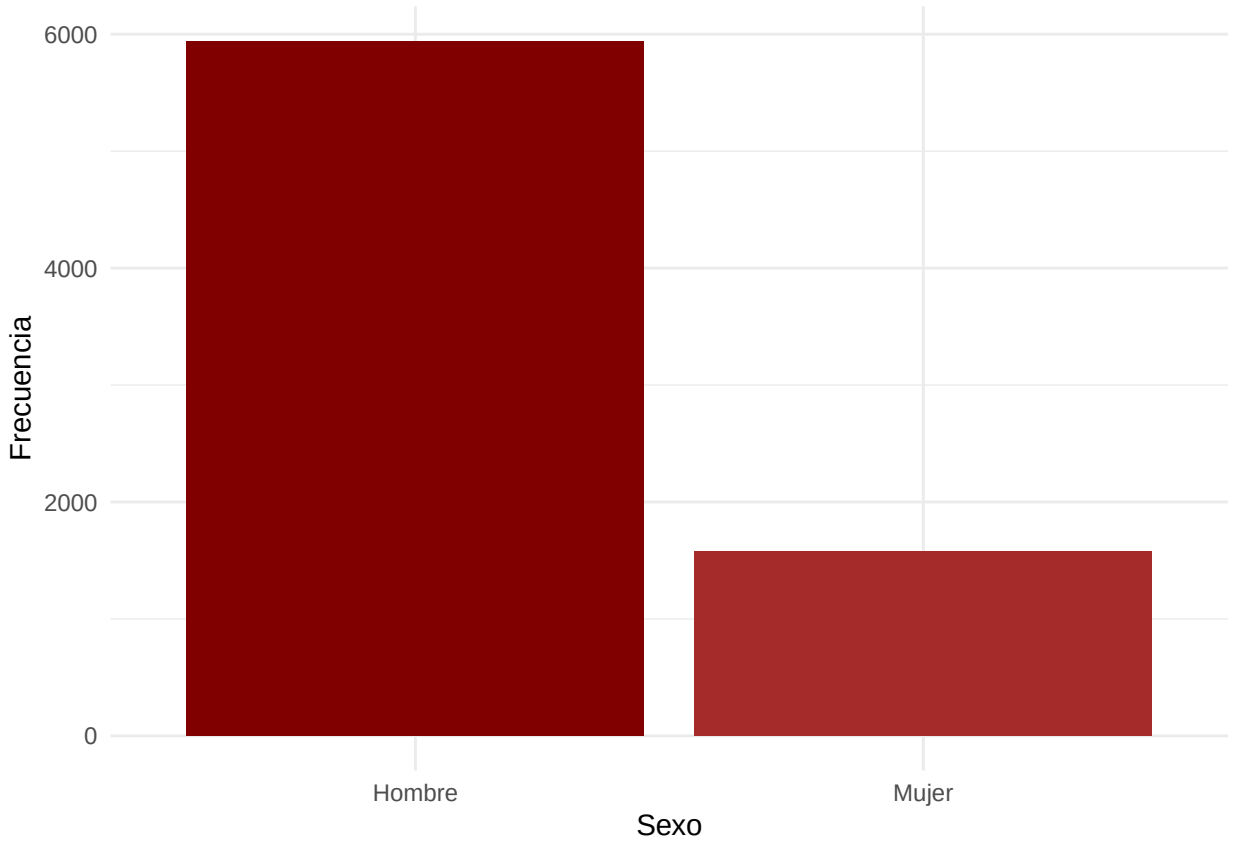
El data solo proporcionaba información del año 2017 y 2018, por lo que se decidió utilizar esa información a pesar de ser pequeña ya que indica informaciéon reciente de los accidentes y lesiones de las motos. Vemos que la diferencia entre estos dos años no es mucha; sin embargo, sigue siendo una cantidad alta.

```
freqAnioOcu <- as.data.frame(table(accidentes$Anio))
 freqAnioOcu<- freqAnioOcu[order(freqAnioOcu$Freq, decreasing = TRUE),]




ggplot(freqAnioOcu, aes(x=Var1, y=Freq, fill = Var1)) +
  geom_bar(stat = "identity") + xlab("Año de accidentes")+
  ylab("Frecuencia") +
  scale_fill_manual(values = c("#800000", "#A52A2A","#A0522D","#8B4513",
                               "#D2691E", "#CD853F","#B8860B","#DAA520",
                               "#DEB887","#F5DEB3")) +
  theme(legend.position = "none")
```



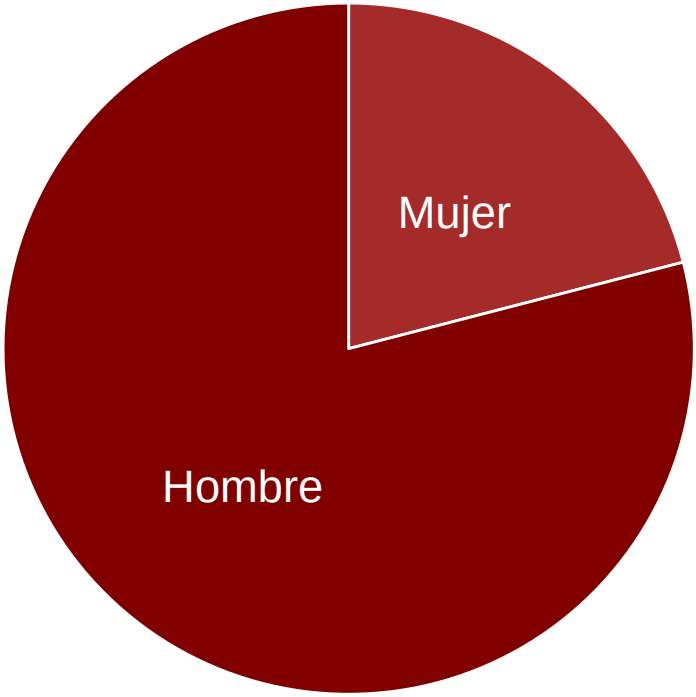
Asimismo, decidimos ver el sexo que más sufría de estos accidentes y vemos que los hombres sobrepasan esta proporción por más de 2,500 accidentes. Esto se puede deber a que la mayoría de hombres que trabajan no solo en la ciudad sino que en el interior se transporta por medio de motocicletas. Sin embargo, vemos que el uso de motocicletas por una mujer es demasiado pequeño, por cual muestra menos accidentes por año.

```
freqSexo <- as.data.frame(table(accidentes$Sexo))
 freqSexo<- freqSexo[order(freqSexo$Freq, decreasing = TRUE),]
freqSexo$prop <- freqSexo$Freq/sum(freqSexo$Freq)



ggplot(freqSexo, aes(x=Var1, y=Freq, fill = Var1)) +
  geom_bar(stat = "identity") + xlab("Sexo")+
  ylab("Frecuencia") +
  scale_fill_manual(values = c("#800000", "#A52A2A","#A0522D","#8B4513",
                               "#D2691E", "#CD853F","#B8860B","#DAA520",
                               "#DEB887","#F5DEB3")) +
  theme(legend.position = "none")
```



Se decidió representar la información anterior en forma de un pie chart para visualizar de mejor manera la proporción de accidentes.

```
freqSexo <- freqSexo %>%
  arrange(desc(Var1)) %>%
  mutate(prop = Freq / sum(freqSexo$Freq) *100) %>%
  mutate(ypos = cumsum(prop)- 0.5*prop )

# Basic piechart
ggplot(freqSexo, aes(x="", y=prop, fill=Var1)) +
  geom_bar(stat="identity", width=1, color="white") +
  coord_polar("y", start=0) +
  theme_void() +
  theme(legend.position="none") +

  geom_text(aes(y = ypos, label = Var1), color = "white", size=6) +
  scale_fill_manual(values = c("#800000", "#A52A2A","#A0522D","#8B4513",
                               "#D2691E", "#CD853F","#B8860B","#DAA520",
                               "#DEB887","#F5DEB3"))
```

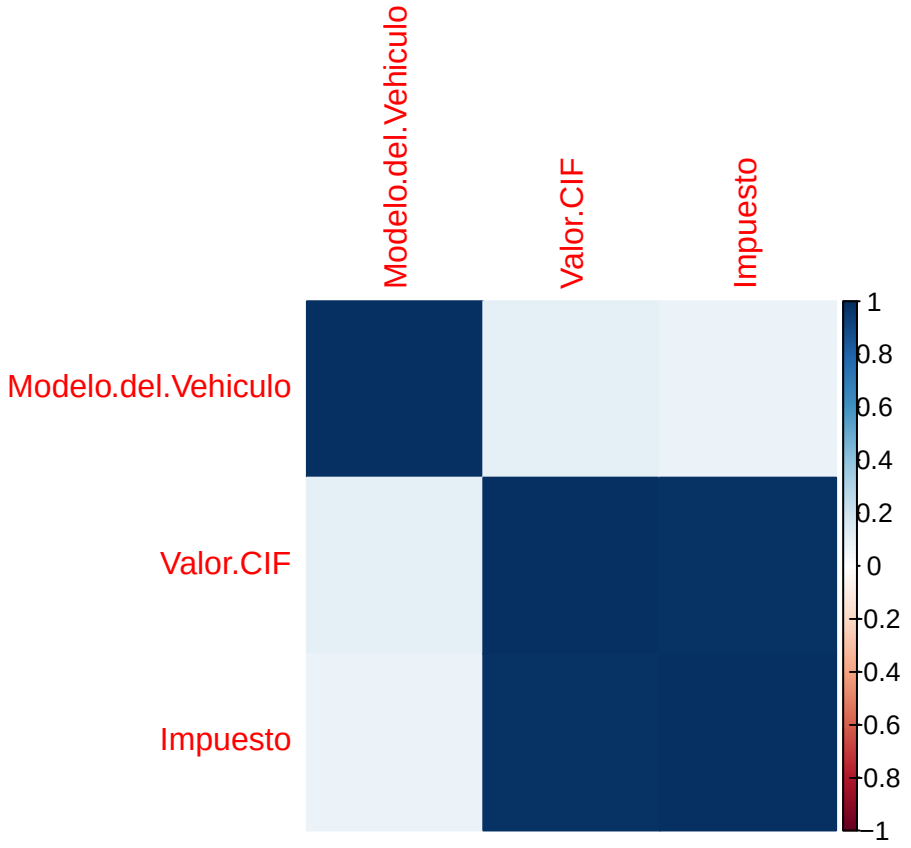



```
accidentes<- accidentes %>% mutate(quantile = ntile(Edad, 3))
```

Análisis de variables cuantitativas.

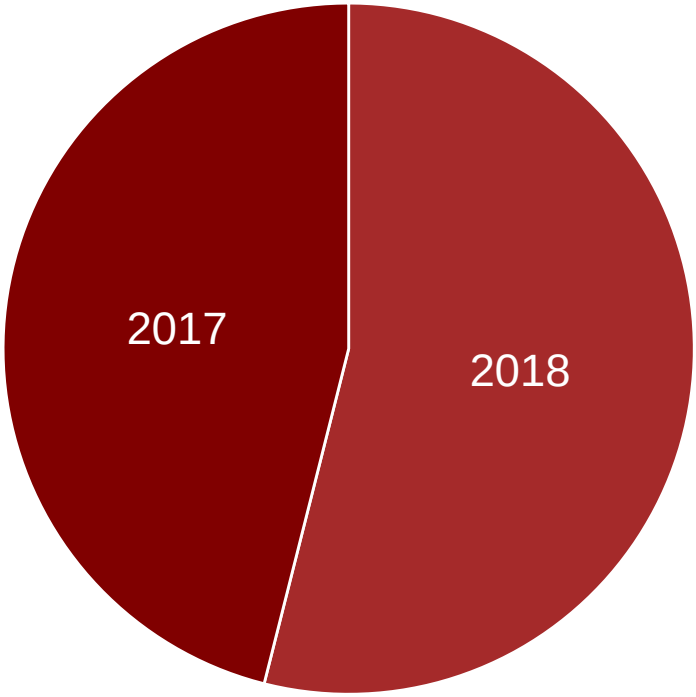
Se procedió con realizar una partición del dataset con contuviera las variables cuantitativas de interés, en este caso fueron: modelo del vehículo, valor CIF, impuesto.

```
motoN <- motos[,c(5,16,17)]
motoN$Modelo.del.Vehiculo <- as.numeric(as.character(motoN$Modelo.del.Vehiculo))
motoN$Valor.CIF <- as.numeric(as.character(motoN$Valor.CIF))
motoN$Impuesto <- as.numeric(as.character(motoN$Impuesto))
M <- cor(motoN)
corrplot(M, method = "color")
```



Se nota que las variables Valor CIF Y el impuesto están altamente correlacionadas. Otro hecho importante es que no sen encuentra una anticorrelación entre ninguna de las variables. En conclusión no se hizo ningún análisis factoreal debido a las pocas variables que se tienen.

```
freqAnioOcu <- as.data.frame(table(accidentes$Anio))
 freqAnioOcu<- freqAnioOcu[order(freqAnioOcu$Freq, decreasing = TRUE),]
freqAnioOcu$prop <- freqAnioOcu$Freq/sum(freqAnioOcu$Freq)



freqAnioOcu <- freqAnioOcu %>%
  arrange(desc(Var1)) %>%
  mutate(prop = Freq / sum(freqAnioOcu$Freq) *100) %>%
  mutate(ypos = cumsum(prop)- 0.5*prop )

# Basic piechart
ggplot(freqAnioOcu, aes(x="", y=prop, fill=Var1)) +
  geom_bar(stat="identity", width=1, color="white") +
  coord_polar("y", start=0) +
  theme_void() +
  theme(legend.position="none") +

  geom_text(aes(y = ypos, label = Var1), color = "white", size=6) +
  scale_fill_manual(values = c("#800000", "#A52A2A","#A0522D","#8B4513",
                               "#D2691E", "#CD853F","#B8860B","#DAA520",
                               "#DEB887","#F5DEB3"))
```


2017
2018