Proyecto - CO3321

Baudilio Velasquez. 18-10665 Juan Cuevas. 19-10056 Anya Marcano. 19-10336

2024-07-04

Parte 1

1. Realice un análisis descriptivo de las variables.

- Brand

Como Brand es una variable cualitativa, procedemos a visualizar la frecuencia de las marcas de los carros usando la función table y un gráfico de barras.

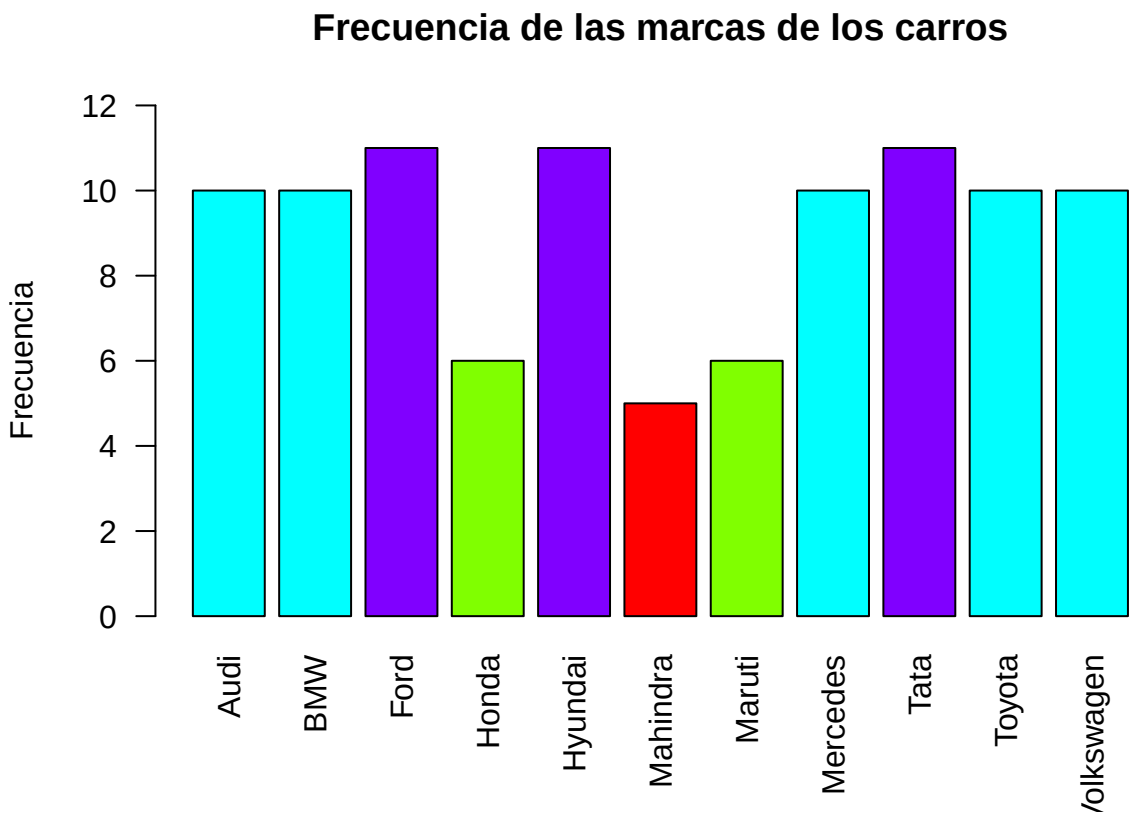
```
table(Brand)
```

```
## Brand
##       Audi        BMW       Ford      Honda    Hyundai   Mahindra     Maruti
##         10         10         11          6         11          5          6
##   Mercedes       Tata     Toyota Volkswagen
##         10         11         10         10
```

Para facilitar el análisis de la variable Brand procedemos a generar dos funciones que nos permitan visualizar la frecuencia de las marcas de los carros en un gráfico de barras con colores iguales para las marcas que tengan la misma frecuencia

```
frecuencias <- table(Brand)
colores_unicos <- rainbow(length(unique(frecuencias)))
mapeo_colores <- colores_unicos[as.factor(frecuencias)]

barplot(table(Brand), col = mapeo_colores,
        main = "Frecuencia de las marcas de los carros", ylab = "Frecuencia",
        las=2, ylim=c(0, 12))
```



Análisis: A partir del gráfico de barras y del resumen proporcionado por la función table podemos apreciar que de las 11 opciones posibles para la variable Brand, las marcas con mayor frecuencia son: Ford, Hyundai y Tata, cada una de ellas con una frecuencia de 11, seguidas por Audi, BMW, Mercedes, Toyota y Wolksvagen con una frecuencia de 10, mientras que por otro lado, la marca con menor frecuencia corresponde con Mahindra con una frecuencia de 5, seguida por las marcas Maruti y Honda con una frecuencia de 6.

- Model

Como Model es una variable cualitativa, procedemos a visualizar la frecuencia de los modelos de los carros usando la funcion table y un grafico de barras.

```
table(Model)
```

```
## Model
##      3 Series      5 Series      7 Series            A3            A4
##             2             1             2             2             1
##            A5            A6        Altroz         Ameo        Aspire
##             2             2             3             2             2
##         BR-V       C-Class         Camry          City         Civic
##             2             3             2             1             1
##       Corolla         Creta       E-Class      EcoSport       Elantra
##             1             1             2             2             2
##     Endeavour        Ertiga         Figo      Fortuner           GLA
##             1             1             1             2             2
```

```
##           GLC           GLE       Harrier        Innova Innova Crysta
##             1             2             1             1             2
##       Mustang         Nexon        Passat         Polo            Q3
##             3             1             2             1             1
##            Q7        Ranger       S-Cross        Safari        Santro
##             2             2             2             2             2
##       Scorpio        Sonata         Swift        T-Roc         Thar
##             1             1             1             2             2
##         Tiago         Tigor        Tiguan        Vento         Venue
##             2             2             1             2             2
##         Verna        Vitara         WR-V            X1            X3
##             3             2             2             1             2
##            X5        XUV300         Yaris
##             2             2             2
```

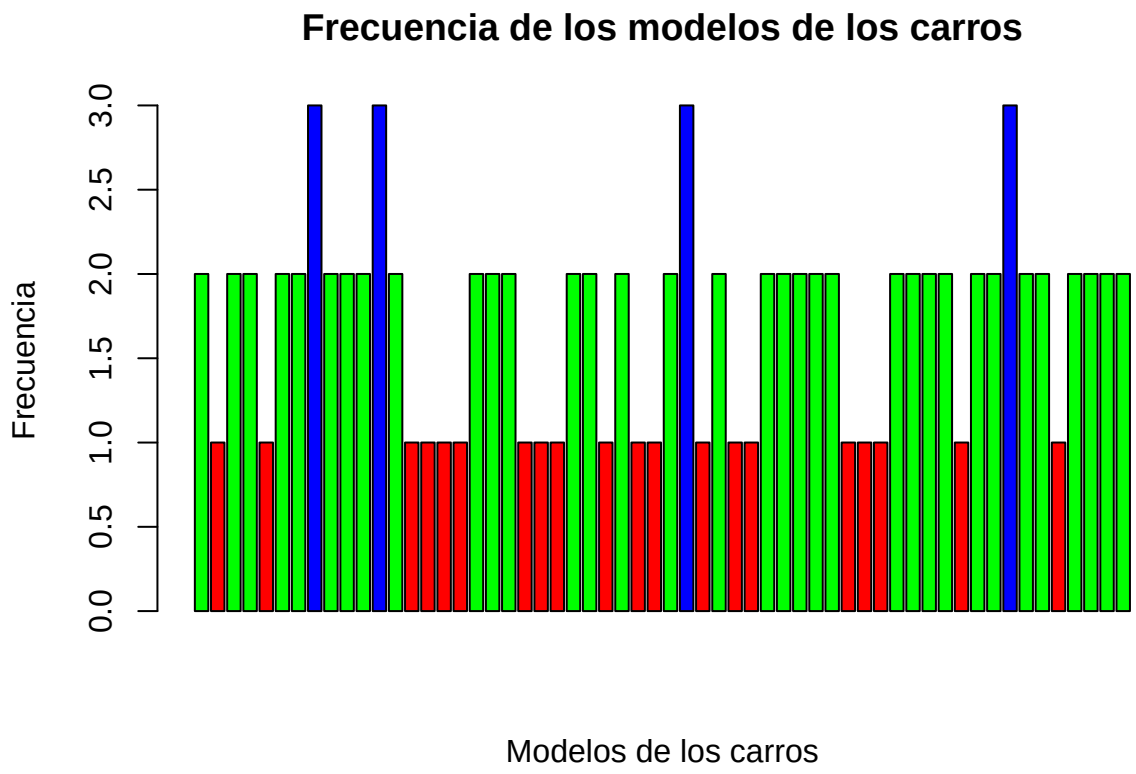
Igual que como hicimos en el análisis anterior, vamos a generar dos funciones que nos permitan visualizar la frecuencia de los modelos de los carros en un gráfico de barras con colores iguales para los modelos que tengan la misma frecuencia:

```r
frecuencias <- table(Model)
length(frecuencias)
```

```
## [1] 58
```

```r
colores_unicos <- rainbow(length(unique(frecuencias)))
mapeo_colores <- colores_unicos[as.factor(frecuencias)]

barplot(table(Model), col = mapeo_colores,
        main = "Frecuencia de los modelos de los carros", ylab = "Frecuencia",
        xlab = "Modelos de los carros", names.arg="")
```



Se decidió no mostrar los nombres de los modelos en el eje X, ya que al ser 58 modelos distintos se dificulta poder mostrar el nombre de cada uno de ellos asociado a su respectiva barra, por esto se aclarara que sucede con cada uno de estos modelos en el análisis de la variable Model.

Análisis: A partir del grafico de barras y del resumen proporcionado por la funcion table vemos que para esta variable hay 58 modelos distintos, de los cuales la mayor frecuencia corresponde con loz modelos Altroz, C-Class, Mustang y Verna teniendo cada uno de ellos una frecuencia de 3, seguidos por los modelos: Yaris, XUV300, X5, x3, Vento, Venue, Vitara, WR-V, T-Roc, Thar, Tiago, Tigor, Ranger, S-Cross, Safari, Santro, Passat, Q7, GLA, GLE, Innova Crysta, EcoSport, Elantra, Fortuner, Camry, E-Class, A6, Ameo, Aspire, BR-V, 3 Series, 7series, A3 y A5 los cuales tienen una frecuencia de 2, mientras que los modelos: 5 Series, A4, City, Civic, Corolla, Creta, Endeavor, Ertiga, Figo, Harrier, Innova, Nexon, Polo, Q3, Scorpio, Sonata, Swift, Tiguan y X1 tienen una frecuencia de 1.

En resumen, la mayor frecuencia la tiene 4 modelos, seguidos por 34 modelos con una frecuencia de 2 y los restantes 20 modelos con una frecuencia de 1.

- Year

Como Year es una variable cuantitativa, procedemos a realizar un análisis descriptivo de la variable Year, para ello usamos las funciones summary, sd, IQR y quantile para obtener un resumen de la variable y calcular la media, desviación estándar, rango intercuartílico y cuartiles respectivamente; así mismo generaremos un histograma y un boxplot para visualizar la distribución de los datos.

```
# Resumen de la variable Year:
summary(Year)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2016    2018    2018    2018    2019    2021
```

```r
sd(Year)
```

```
## [1] 1.17116
```

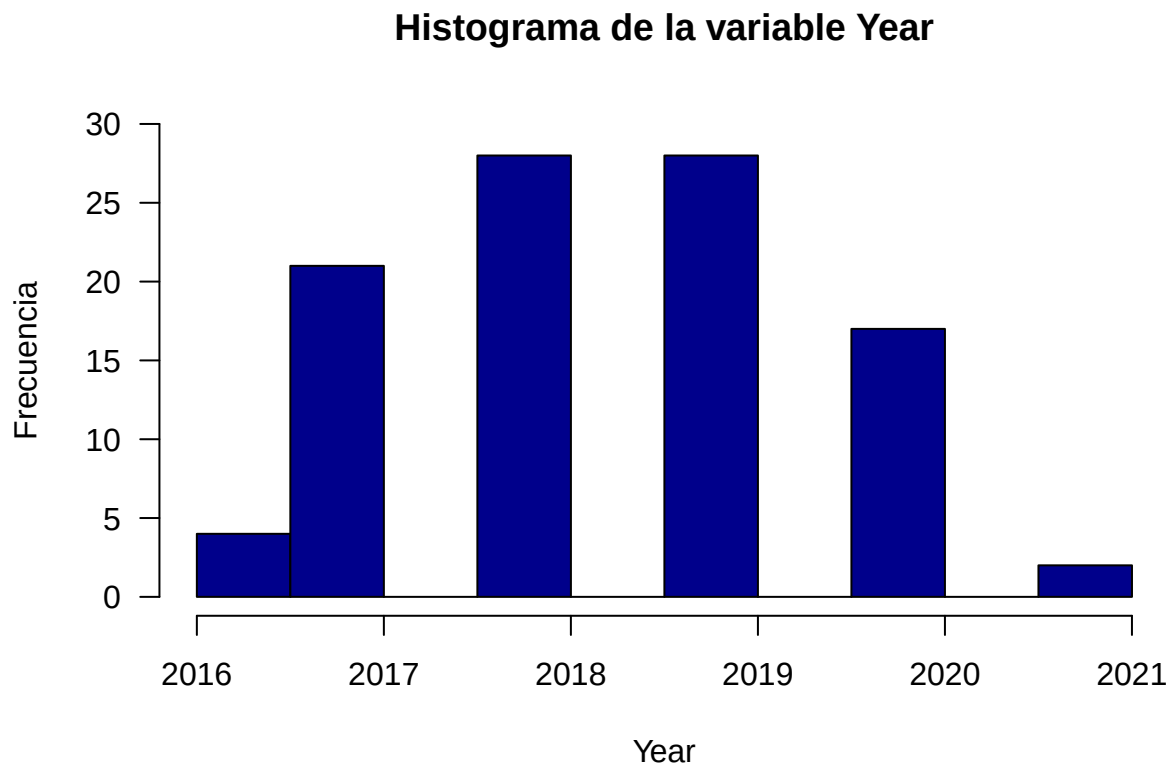
```r
IQR(Year)
```

```
## [1] 1.25
```

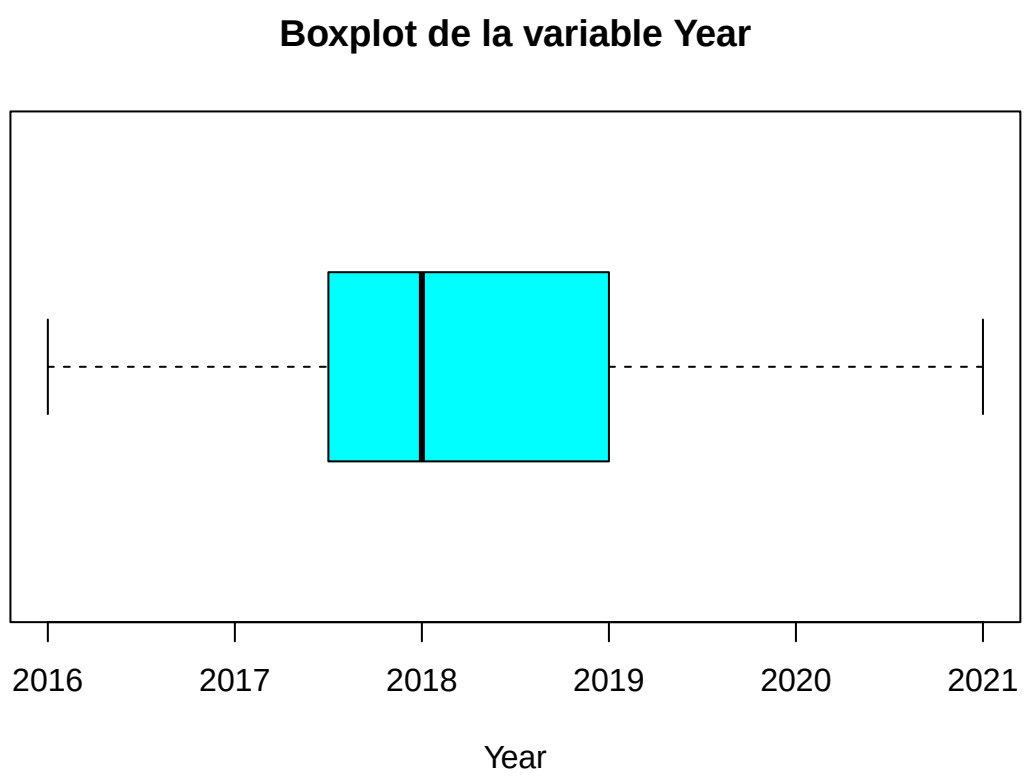
```r
quantile(Year, c(0.25, 0.5, 0.75))
```

```
##     25%     50%     75%
## 2017.75 2018.00 2019.00
```

```r
# Gráficas de la variable Year:
hist(Year, main = "Histograma de la variable Year", xlab = "Year",
     ylab = "Frecuencia", col = "darkblue", las=1, ylim = c(0, 30))
```




```r
boxplot(Year, main = "Boxplot de la variable Year", xlab = "Year",
        col = "cyan", horizontal=TRUE)
```



Análisis: El año promedio de los datos es 2018, con una desviación estándar de 1.17116 y una mediana también de 2018.

El año mínimo presente en la muestra es 2016, mientras que el año máximo es 2021. En cuanto al rango intercuartílico, este es de 2 años, siendo el 25% de los datos menores a 2017, el 50% menores a 2018 y el 75% menores a 2019.

Notándose que la mayor parte de los datos se encuentran en el rango de 2017 a 2019, siendo los años 2018 y 2019 los que tienen la mayor frecuencia de aparición en la muestra.

En cuando a la distribución de los datos, tanto en el histograma como en el boxplot se observa que los datos están distribuidos mostrando una asimetría desplazada hacia la izquierda, esto claramente evidenciable en el diagrama de caja, ya que la parte más grande de misma es la que se encuentra superior a la mediana. Así mismo, podemos apreciar que no hay presencia de valores atípicos en el boxplot de la muestra para esta variable.

Finalmente, tanto en el histograma como en el boxplot se observa cierta dispersión de los datos, los cuales se encuentran más dispersos entre el segundo y tercer cuartil cuartil, y más concentrados entre el primer y segundo cuartil.

- Kilometers_Driven

Como Kilometers_Driven es una variable cuantitativa, procedemos a realizar un análisis descriptivo de la variable Kilometers_Driven, para ello haremos un análisis parecido al que planteamos para la variable Year, haciendo un resumen de los datos y generando un histograma y un boxplot para visualizar la distribución de los mismos.

```r
# Resumen de la variable Kilometers_Driven:
summary(Kilometers_Driven)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10000   22000   27000   28150   32000   60000
```

```r
sd(Kilometers_Driven)
```

```
## [1] 9121.376
```

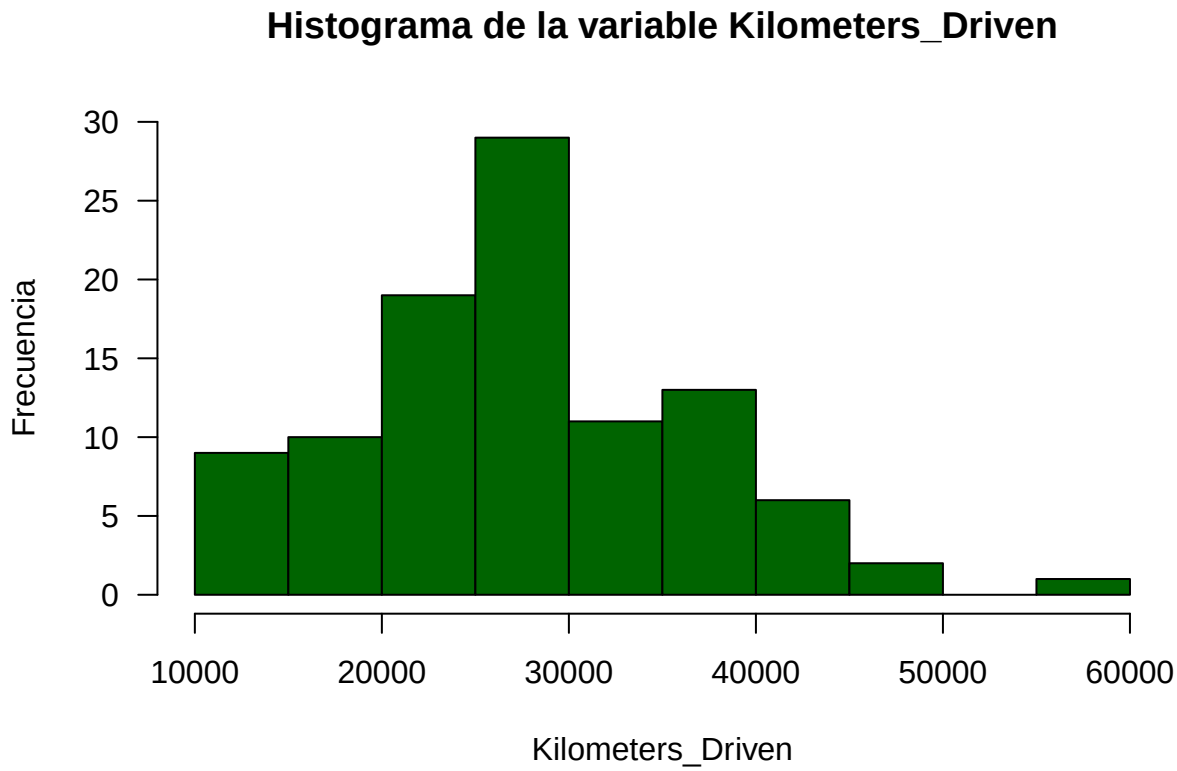
```r
IQR(Kilometers_Driven)
```

```
## [1] 10000
```

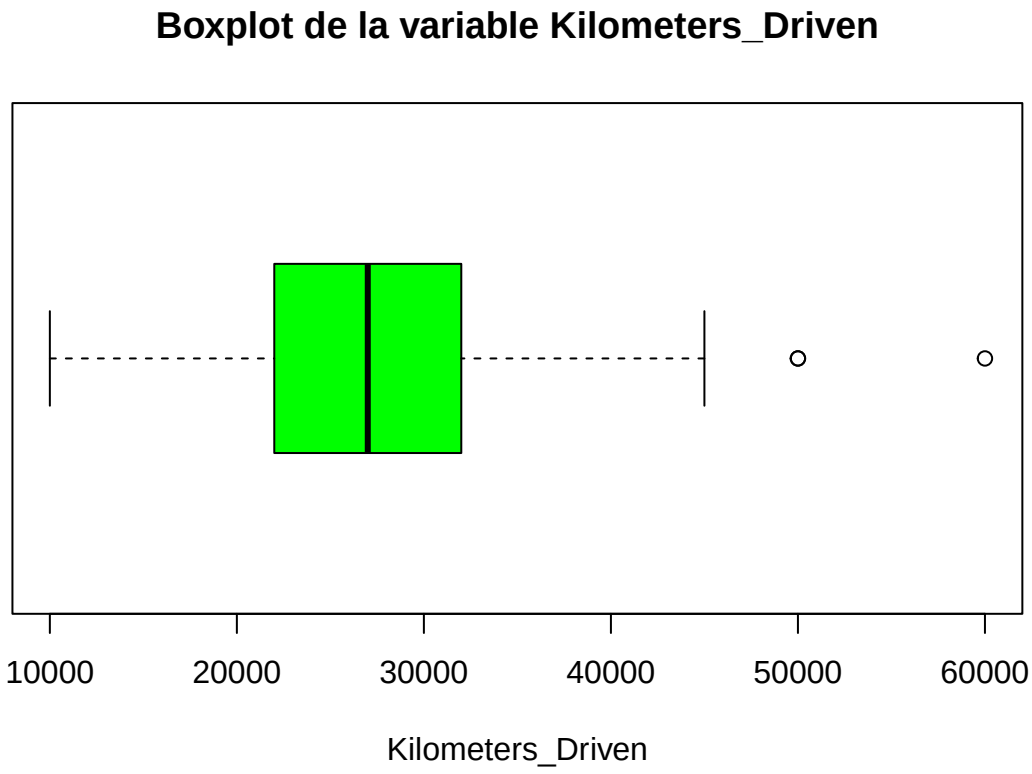
```r
quantile(Kilometers_Driven, c(0.25, 0.5, 0.75))
```

```
##   25%   50%   75%
## 22000 27000 32000
```

```r
# Gráficas de la variable Kilometers_Driven:
hist(Kilometers_Driven, main = "Histograma de la variable Kilometers_Driven",
     xlab = "Kilometers_Driven", ylab = "Frecuencia", col = "darkgreen",
     las=1, ylim = c(0, 30))
```



```r
boxplot(Kilometers_Driven, main = "Boxplot de la variable Kilometers_Driven",
        xlab = "Kilometers_Driven", col = "green", horizontal=TRUE)
```



Análisis: El promedio de kilometros recorridos por los carros es de 28150km, con una desviación estándar de 9121.376 km y una mediana de 27000 km

El kilometraje minimo presente en la muestra es de 10000 km, mientras que el kilometraje maximo es de 60000.

En cuanto al rango intercuartílico, este es de 10000 km, siendo el 25% de los datos menores a 22000 km, el 50% menores a 27000 km y el 75% menores a 32000 km. Notándose que la mayor parte de los datos se encuentran en el rango de 22000 a 32000 km.

En cuando a la distribución de los datos, tanto en el histograma como en el boxplot se observa que los datos están distribuidos aparentemente de forma simétrica, teniéndose la presencia de dos valores atípicos, los cuales son claramente evidenciados en el boxplot y que corresponden con kilometrajes de 50000 km y 60000 km, los cuales se encuentran por encima del tercercuartil, donde cabe mencionar que el kilometraje de 60000 km es el valor máximo de la muestra observado una sola vez en el Car_ID 5, mientras que el kilometraje de 50000 km es el segundo valor máximo de la muestra observado dos veces, una en el Car_ID 1 y otra en el Car_ID 21.

Finalmente, tanto en el histograma como en el boxplot se puede observar que la distribución de los datos es casi simétrica, y esto lo podemos afirmar, ya que a pesar de que visualmente el boxplot muestra que la mediana divide en dos partes iguales la caja, lo cierto es que la simetría implica que el valor de la mediana debería coincidir con el de la media, sin embargo, esto no es así, existe una sutil diferencia entre ellas siendo la mediana de 27000 km y la media de 28150 km, es decir, hay un margen de 1150 km entre ambas medidas;

no obstante, esta diferencia puede deberse precisamente a la presencia de valores atípicos mencionados anteriormente, ocasionan que la media se vea ligeramente desplazada y no coincida con la mediana.

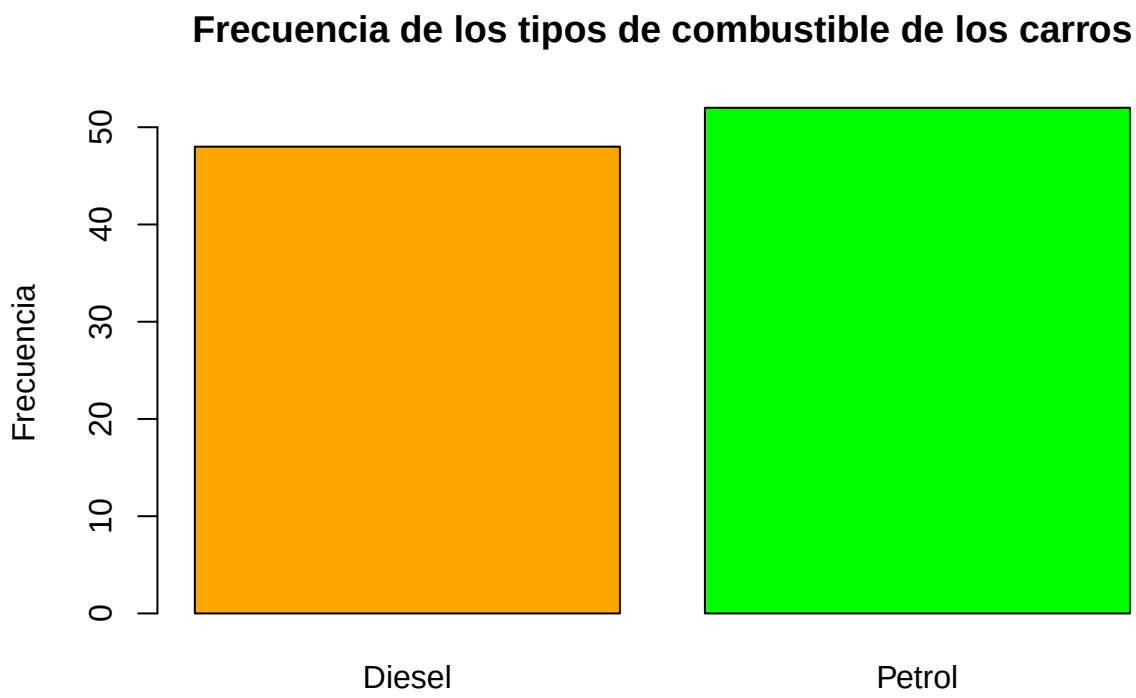
- Fuel_Type

Como Fuel_Type es una variable cualitativa, procedemos a visualizar la frecuencia de los tipos de combustible de los carros usando la funcion table y un grafico de barras.

```
table(Fuel_Type)
```

```
## Fuel_Type
## Diesel Petrol
##     48     52
```

```
barplot(table(Fuel_Type), col = c("orange", "green"),
        main = "Frecuencia de los tipos de combustible de los carros",
        ylab = "Frecuencia")
```



Análisis: A partir del gráfico de barras y del resumen proporcionado por la función table podemos decir que, en el caso de la variable Fuel_Type, se tienen un total de 2 tipos de combustible, siendo el Diesel el tipo de combustible con menor frecuencia de aparición en la muestra, con una frecuencia de 48 autos, mientras que su alternativa para este estudio, el Petrol, tuvo una frecuencia sutilmente mayor, con 52 autos, teniéndose una diferencia de tan solo 4 autos entre ambos tipos de combustible.

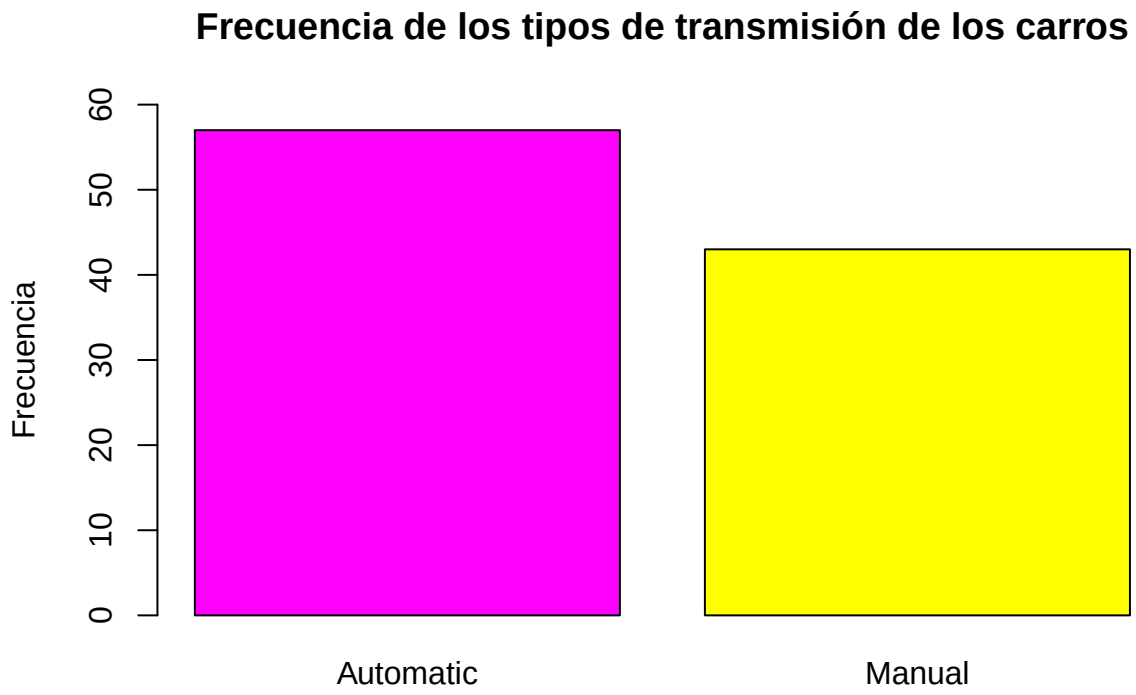
- Transmission

Como Transmission es una variable cualitativa, procedemos a visualizar la frecuencia de los tipos de transmisión de los carros usando la función table y un gráfico de barras.

```
table(Transmission)
```

```
## Transmission
## Automatic    Manual
##        57        43
```

```
barplot(table(Transmission), col = c("magenta", "yellow"),
        main = "Frecuencia de los tipos de transmisión de los carros",
        ylab = "Frecuencia", ylim = c(0,60))
```



Análisis: A partir del gráfico de barras y del resumen proporcionado por la función table podemos decir que, en el caso de la variable Transmission, se tienen un total de 2 tipos de transmisión, siendo la Automática la transmisión con mayor frecuencia de aparición en la muestra, con una frecuencia de 57 autos, mientras que su alternativa para este estudio, la Manual, tuvo una frecuencia sutilmente menor, con 43 autos, teniéndose una diferencia de tan solo 14 autos entre ambos tipos de transmisión.

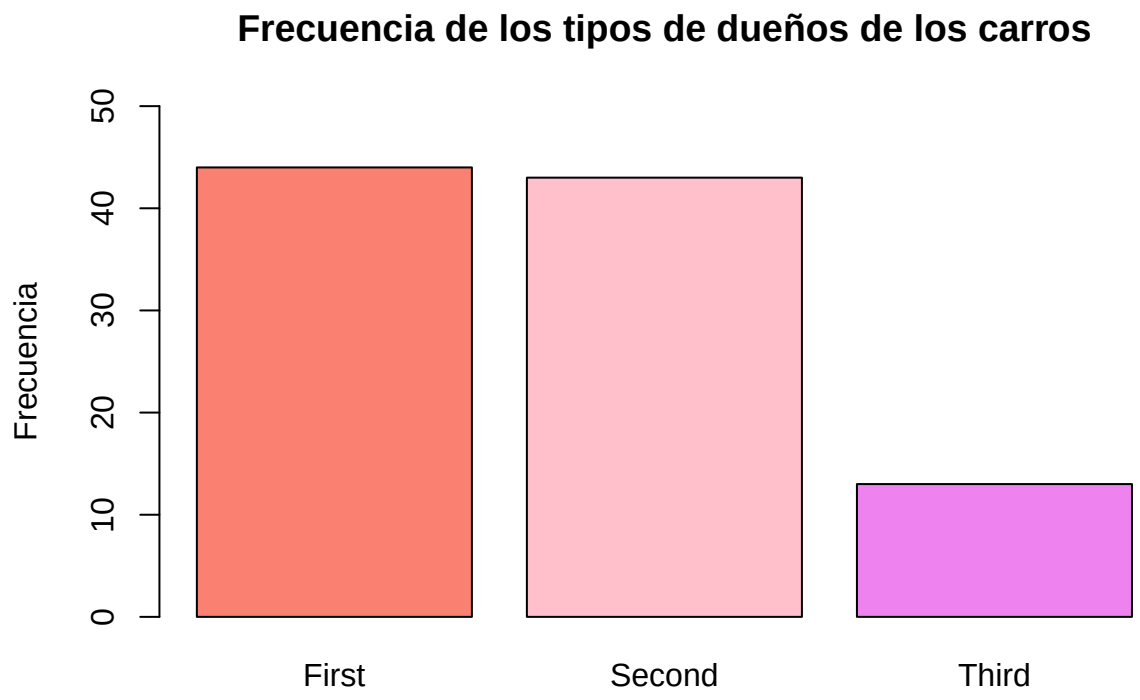
- Owner_Type

Como Owner_Type es una variable cualitativa, así como en los casos anteriores procedemos a generar, la tabla y el gráfico de barras para visualizar la frecuencia de los tipos de dueños de los carros.

```r
table(Owner_Type)
```

```
## Owner_Type
##  First Second  Third
##     44     43     13
```

```r
barplot(table(Owner_Type), col = c("salmon", "pink", "violet"),
        main = "Frecuencia de los tipos de dueños de los carros",
        ylab = "Frecuencia", ylim = c(0,50))
```



Análisis: A partir del gráfico de barras y del resumen proporcionado por la función table podemos decir que, en el caso de la variable Owner_Type, se tienen un total de 3 tipos de dueños, siendo el First Owner el tipo de dueño con mayor frecuencia de aparición en la muestra, con una frecuencia de 44; seguidas inmediatamente por la categoría Second Owner, la cual tiene una frecuencia de 43 autos, es decir, solamente 1 punto por debajo que la categoría First Owner, mientras que la tercera posibilidad contemplada para esta variable, la de Third Owner, tuvo una frecuencia de tan solo 13 autos, siendo la que menos frecuencia tuvo en toda la muestra, teniendo una diferencia de 31 y 30 puntos con respecto a las categorías First y Second Owner respectivamente.

- Mileage

Como Mileage es una variable cuantitativa, procedemos a realizar un análisis descriptivo de la variable Mileage, para ello haremos un analisis parecido al que planteamos para la variable Year, haciendo un resumen de los datos y generando un histograma y un boxplot para visualizar la distribución de los mismos.

```r
# Resumen de la variable Mileage:
summary(Mileage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.00   15.00   17.00   17.21   19.00   25.00
```

```r
sd(Mileage)
```

```
## [1] 3.309902
```

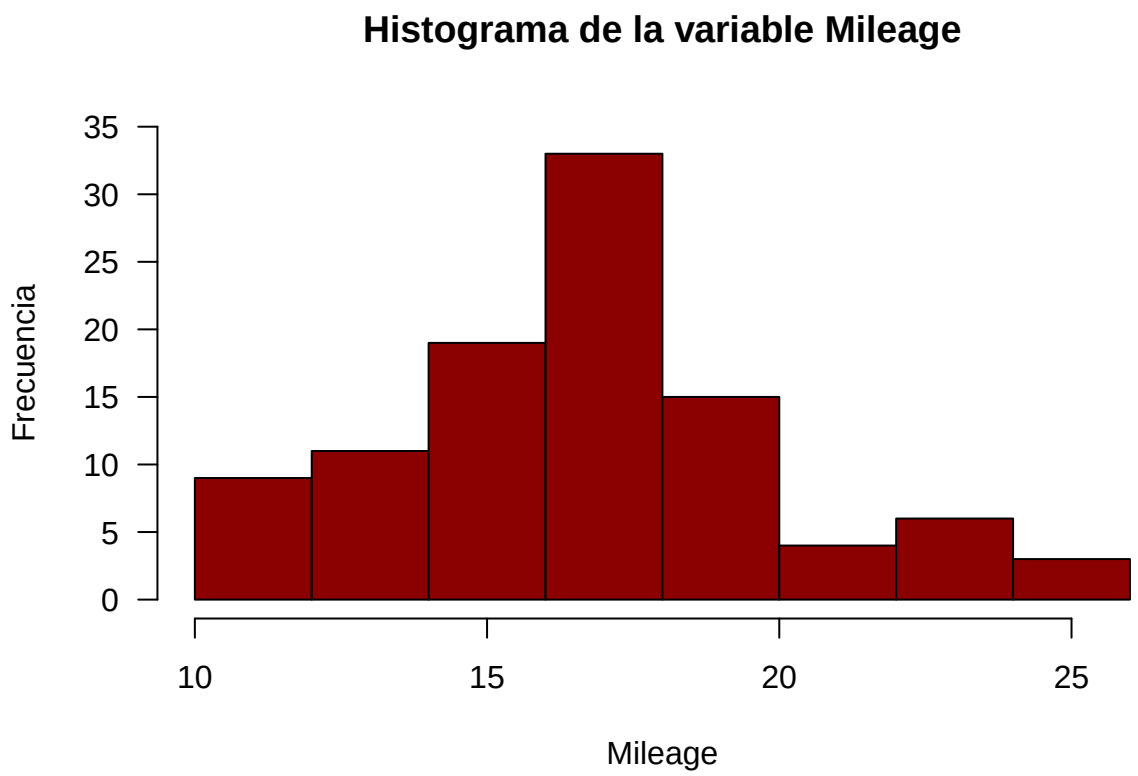
```r
IQR(Mileage)
```

```
## [1] 4
```

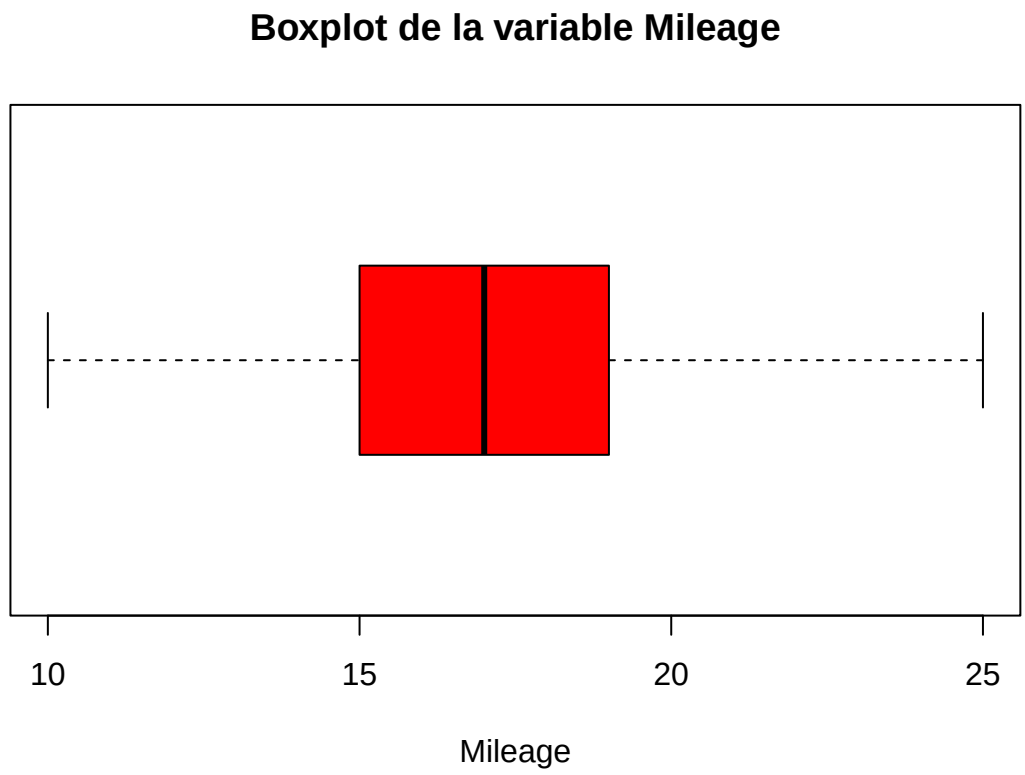
```r
quantile(Mileage, c(0.25, 0.5, 0.75))
```

```
## 25% 50% 75%
##  15  17  19
```

```r
# Gráficas de la variable Mileage:
hist(Mileage, main = "Histograma de la variable Mileage", xlab = "Mileage",
     ylab = "Frecuencia", col = "darkred", las=1, ylim = c(0, 35))
```



```r
boxplot(Mileage, main = "Boxplot de la variable Mileage", xlab = "Mileage",
        col = "red", horizontal=TRUE)
```



Análisis: El promedio de millas por galón de los carros es de 17.21, con una desviación estándar de 3.309902 y una mediana de 17.00, siendo el valor mínimo de 10.00 y el valor máximo de 25.00.

En cuanto al rango intercuartílico, este es de 4 millas por galón, siendo el 25% de los datos menores a 15.00, el 50% menores a 17.00 y el 75% menores a 19.00. Notándose que la mayor parte de los datos se encuentran en el rango de 15.00 a 19.00 millas por galón.

En cuando a la distribución de los datos, tanto en el histograma como en el boxplot se observa que los datos están distribuidos aparentemente de forma simétrica, sin presencia de valores atípicos en el boxplot de la muestra para esta variable.

Finalmente, tanto en el histograma como en el boxplot se puede observar que la distribución de los datos es casi simétrica, y esto lo podemos afirmar, ya que a pesar de que visualmente el boxplot muestra que la mediana divide en dos partes iguales la caja, lo cierto es que de ser una dsitribución simétrica en toda regla, el valor de la mediana debería coincidir con el de la media, sin embargo, esto no es así en este caso, existe una sutil diferencia entre ellas siendo la mediana de 17.00 km y la media de 17.21 km, es decir, hay un margen de 0.21 km entre ambas medidas, sin embargo esta diferencia es tan pequeña que no afecta la interpretación de la simetría de la distribución, además de que en este caso, a diferencia de la variable Kilometers_Driven, no se observa la presencia de valores atípicos en la muestra, lo que hace que la distribución de los datos sea más homogénea y simétrica.

- Engine

Como Engine es una variable cuantitativa, procedemos a realizar un análisis descriptivo de la variable Engine, para ello haremos un analisis parecido al anterior:

```
# Resumen de la variable Engine:
summary(Engine)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     999    1462    1774    1855    2143    4951
```

```
sd(Engine)
```

```
## [1] 631.3115
```

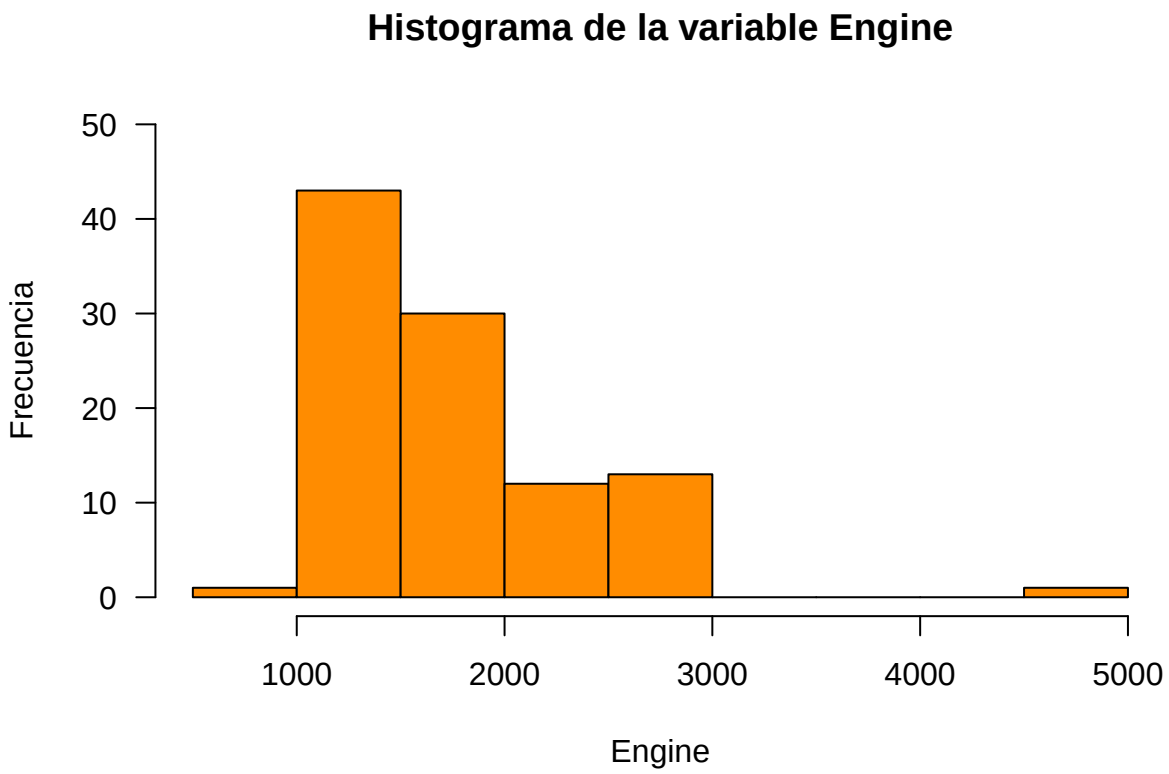
```
IQR(Engine)
```

```
## [1] 681
```

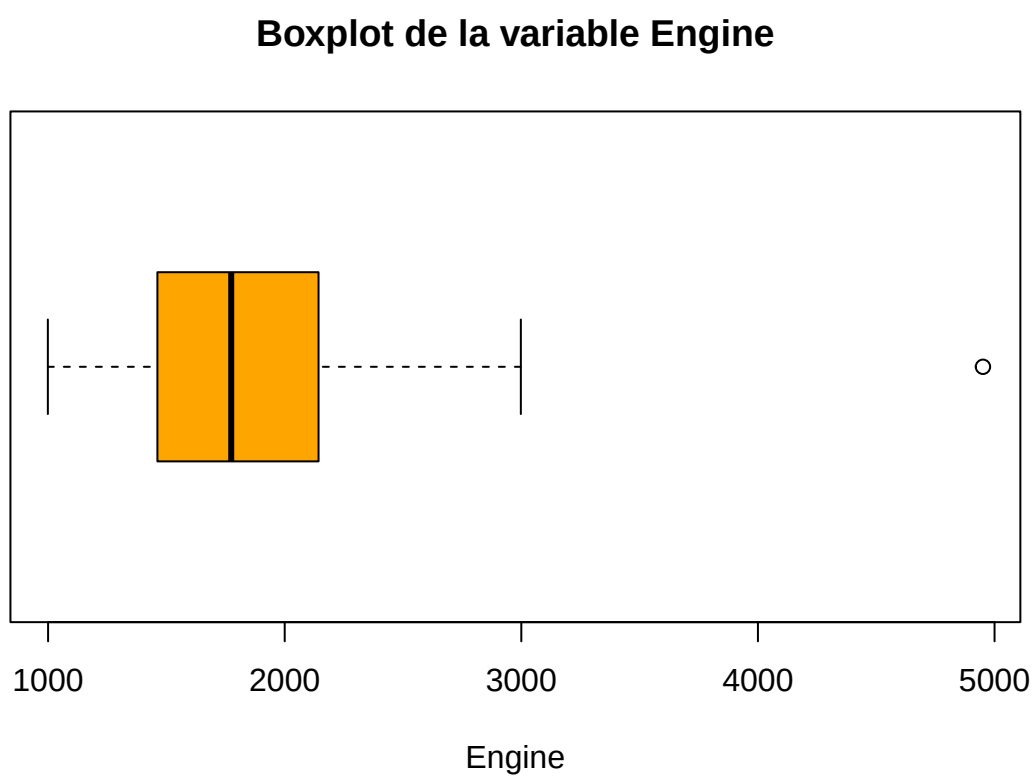
```
quantile(Engine, c(0.25, 0.5, 0.75))
```

```
##  25%  50%  75%
## 1462 1774 2143
```

```
# Gráficas de la variable Engine:
hist(Engine, main = "Histograma de la variable Engine", xlab = "Engine",
     ylab = "Frecuencia", col = "darkorange", las=1, ylim = c(0, 50))
```




```
boxplot(Engine, main = "Boxplot de la variable Engine", xlab = "Engine",
        col = "orange", horizontal=TRUE)
```



Análisis: El promedio de la cilindrada de los carros es de 1855 cc, con una desviación estándar de 631.3115 cc y una mediana de 1774 cc, siendo el valor mínimo de999 cc y el valor máximo de 4951 cc.

En cuanto al rango intercuartílico, este es de 681 cc, siendo el 25% de los datos menores a 1462 cc, el 50% menores a 1774 cc y el 75% menores a 2143 cc. Notándose que la mayor parte de los datos se encuentran en el rango de 1462 a 2143 cc.

En cuando a la distribución de los datos, tanto en el histograma como en el boxplot se observa que los datos están distribuidos de forma asimétrica, notándose un ligero desplazamiento de la mediana a la izquierda del boxplot de la muestra para esta variable, lo que indica que la distribución de los datos es asimétrica positiva, es decir, que los datos se encuentran más dispersos entre el segundo y tercer cuartil cuartil, y más concentrados entre el primer y segundo cuartil. Además, podemos evidenciar que la distribución de los datos no es homogénea, ya que se observa la presencia de un valor atípico el cual corresponde con una cilindrada de 4951 cc, siendo el valor más alto de la muestra observado una sola vez en el Car_ID 3.

Finalmente, podemos mencionar que a raíz de la asimetría de la distribución existe una sutil diferencia entre los valores de la media y la mediana, siendo la mediana de 1774 cc y la media de 1855 cc, es decir, hay un margen de 81 puntos entre ambas medidas.

- Power

Esta variable también es cuantitativa, por lo que procedemos a realizar un análisis descriptivo por medio de summary, sd, IQR y quantile, así como a generar un histograma y un boxplot para visualizar la distribución de los datos.

```r
# Resumen de la variable Power:
summary(Power)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    68.0   103.0   148.0   158.1   187.0   396.0
```

```r
sd(Power)
```

```
## [1] 76.96814
```

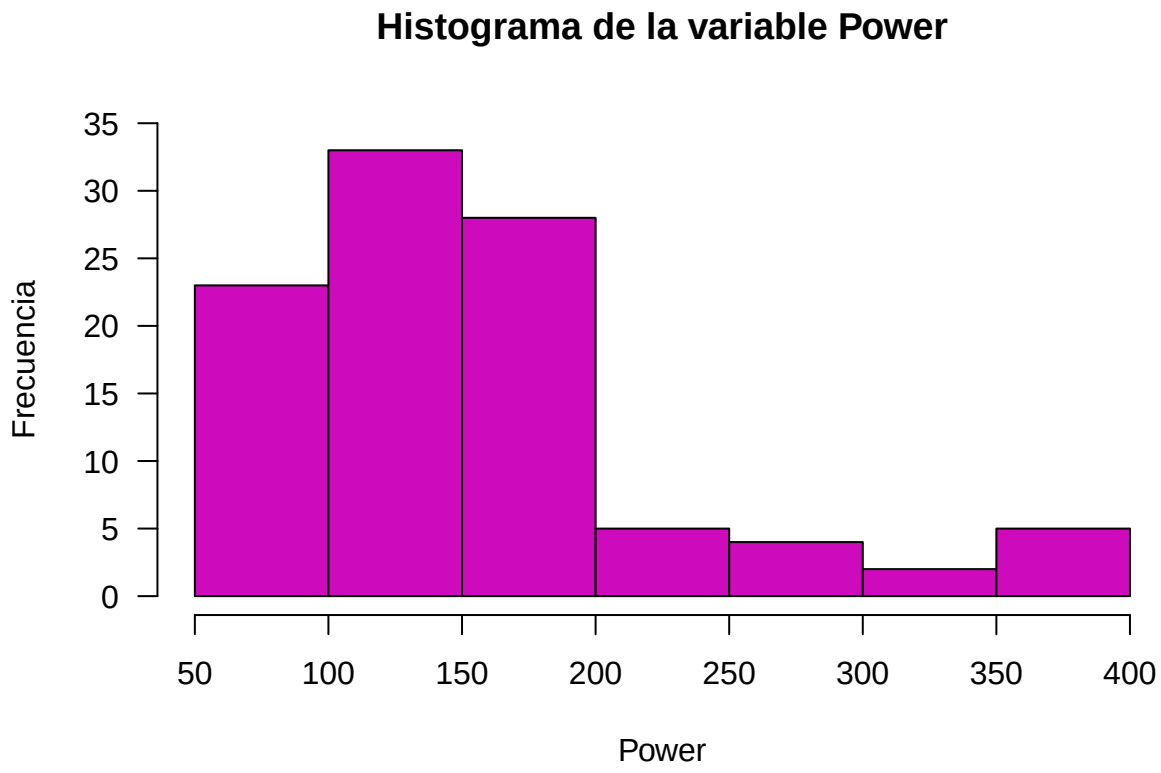
```r
IQR(Power)
```

```
## [1] 84
```

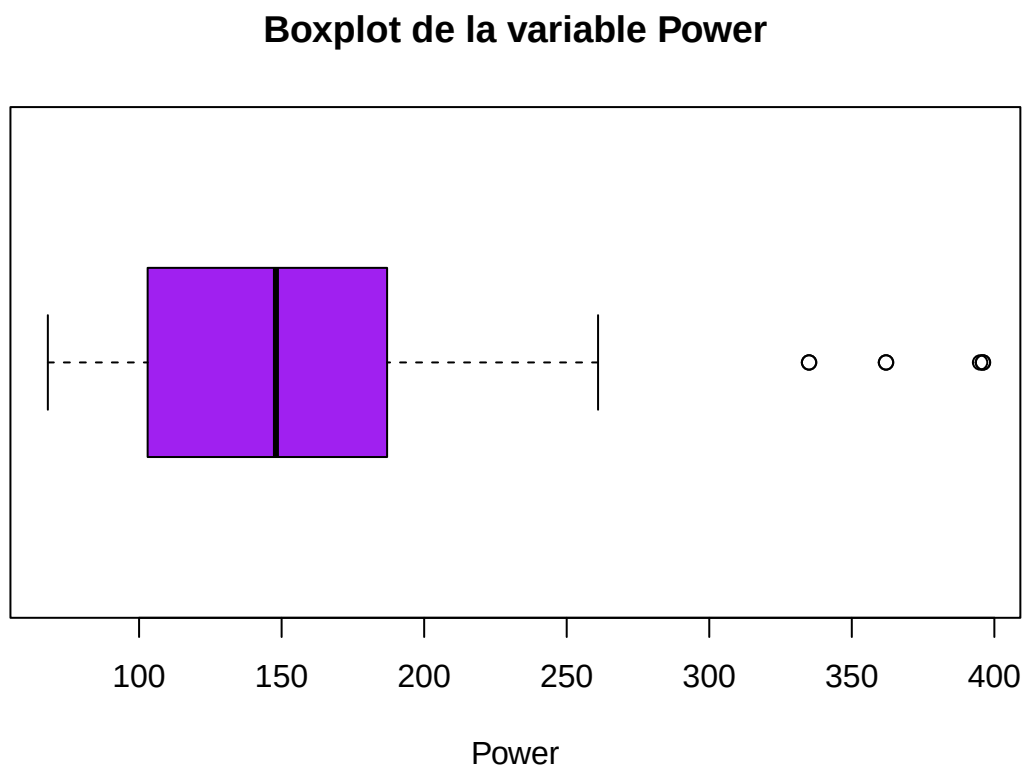
```r
quantile(Power, c(0.25, 0.5, 0.75))
```

```
## 25% 50% 75%
## 103 148 187
```

```r
# Gráficas de la variable Power:
hist(Power, main = "Histograma de la variable Power", xlab = "Power",
     ylab = "Frecuencia", col = "301934", las=1, ylim = c(0, 35))
```



```r
boxplot(Power, main = "Boxplot de la variable Power", xlab = "Power",
        col = "purple", horizontal=TRUE)
```



Análisis: La potencia promedio de los carros de la muestra es de 158.1 hp, con una desviación estándar de 76.96814 hp y una mediana de 138.1 hp, siendo el valor mínimo de 68.0 hp y el valor máximo de 396.0 hp.

En cuanto al rango intercuartílico, este es de 84 hp, siendo el 25% de los datos menores a 103.00 hp, el 50% menores a 148 hp y el 75% menores a 187 hp. Notándose que la mayor parte de los datos se encuentran en el rango de 103 a 187 hp.

En cuando a la distribución de los datos, tanto en el histograma como en el boxplot se observa que los datos están distribuidos de forma asimétrica, notándose un ligero desplazamiento de la mediana a la derecha del boxplot, es decir, que se tiene una distribución asimétrica negativa, o sesgada a la izquierda, lo que indica que los datos se encuentran más dispersos entre el primer y segundo cuartil cuartil, y más concentrados entre el segundo y tercer cuartil. Además, podemos evidenciar que la distribución de los datos no es homogénea, ya que se observa la presencia de varios valores atípicos a la derecha del bigote superior del boxplot, los cuales corresponden con potencias sobre los 300 hp, siendo el valor más alto de la muestra el 396 hp, el cual fue observado en dos oportunidades, para el Car_ID 51 y el 89.

Finalmente, podemos mencionar que a raíz de la asimetría de la distribución existe una sutil diferencia entre los valores de la media y la mediana, siendo la primera de 158.1 hp y la segunda de 148.0 hp, es decir, hay un margen de 10.1 hp entre ambas medidas.

- Seats

Como Seats es una variable cuantitativa, procedemos a realizar un análisis descriptivo de la variable Seats, para ello haremos un análisis parecido al que planteamos para la variable anterior:

```
# Resumen de la variable Seats:
summary(Seats)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.00    5.00    5.00    5.23    5.00    7.00
```

```
sd(Seats)
```

```
## [1] 0.7501515
```

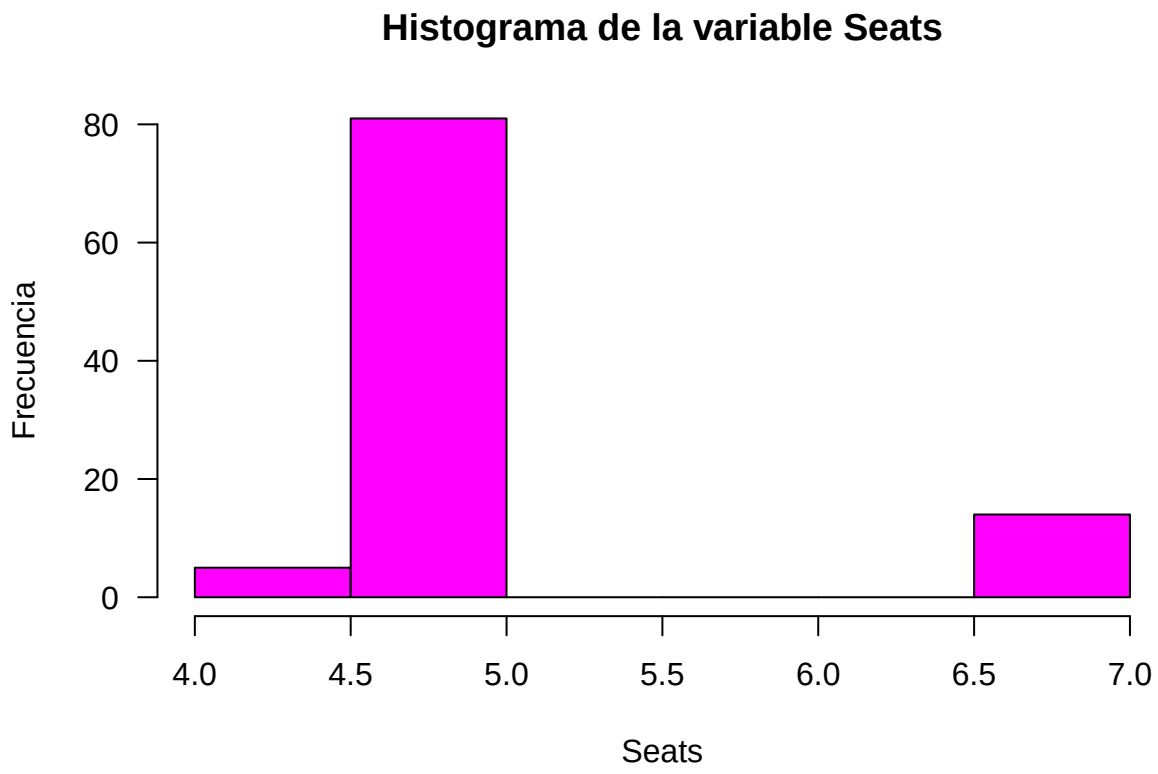
```
IQR(Seats)
```

```
## [1] 0
```

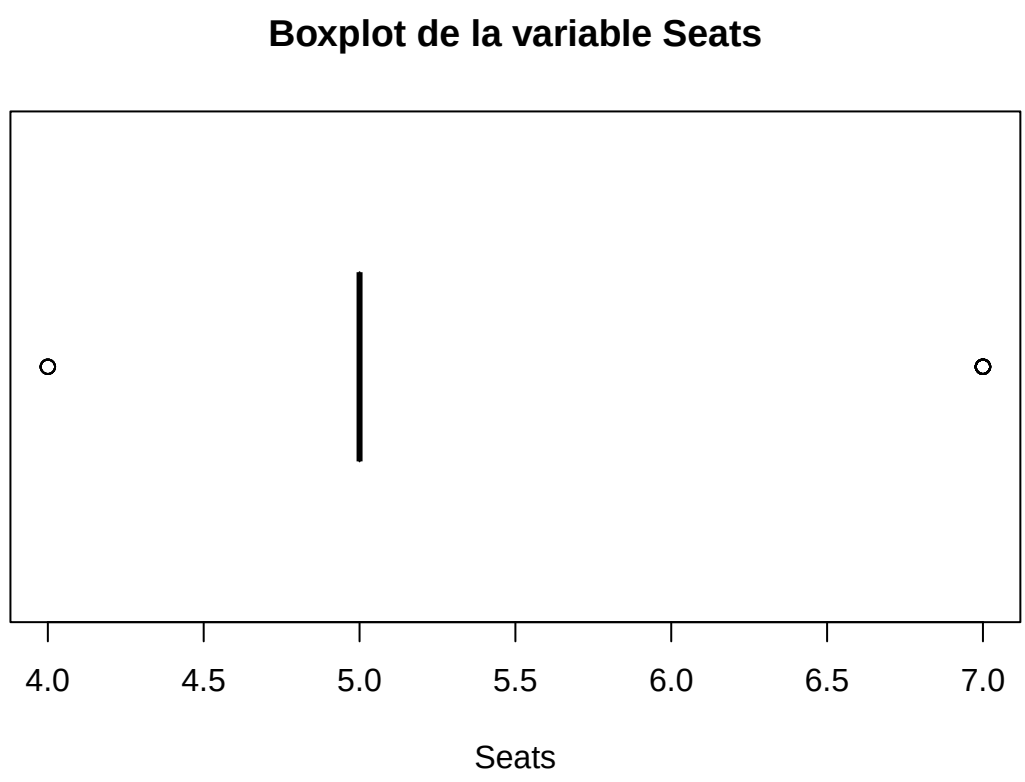
```
quantile(Seats, c(0.25, 0.5, 0.75))
```

```
## 25% 50% 75%
##   5   5   5
```

```
# Gráficas de la variable Seats:
hist(Seats, main = "Histograma de la variable Seats", xlab = "Seats",
     ylab = "Frecuencia", col = "magenta", las=1, ylim = c(0, 80))
```




```
boxplot(Seats, main = "Boxplot de la variable Seats", xlab = "Seats",
        col = "violet", horizontal=TRUE)
```



Análisis: El promedio de asientos de los carros de la muestra es de 5.23 asientos, con una desviación estándar de 0.7501515 asientos y una mediana de 5 asientos, siendo el valor mínimo de 4 asientos y el valor máximo de 7 asientos.

En cuanto al rango intercuartílico, podemos apreciar que este es de 0, ya que para este caso en particular tenemos que los datos están extremadamente agrupados, es decir, que los mismos están casi por completo agrupados en el valor de 5, con muy pocos datos en otros valores, lo cual lo podemos ver evidenciado en que la caja resultante del boxplot es una "caja" que no es visible ya que es extremadamente delgada, apreciándose únicamente la mediana ubicada en 5, así mismo, el histograma nos hace hacernos una idea de porqué está pasando esto, ya que vemos la diferencia que existe entre los autos que presentan 5 asientos y el resto de los autos, siendo estos últimos una clara minoría de la muestra.

En cuando a la distribución de los datos, vemos que a consecuencia de la agrupación de los datos que se mencionó anteriormente, tenemos que el otro dato relevante que puede sacarse del boxplot es la presencia de valores atípicos, los cuales corresponden con autos 4 observaciones que poseen 4 asientos, y 14 autos que poseen 7 asientos, siendo estos valores atípicos debido a que el resto de los autos de la muestra tienen exactamente 5 asientos lo cual puede ser un indicio de que esta variable presenta una distribución uniforme.

- Price

Como Price es una variable cuantitativa, procedemos a realizar un análisis descriptivo de la variable Price, para ello haremos un analisis parecido al que planteamos para la variable anterior:

```
# Resumen de la variable Price:
summary(Price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  450000  700000 1300000 1574000 2500000 4000000
```

```
sd(Price)
```

```
## [1] 1000265
```

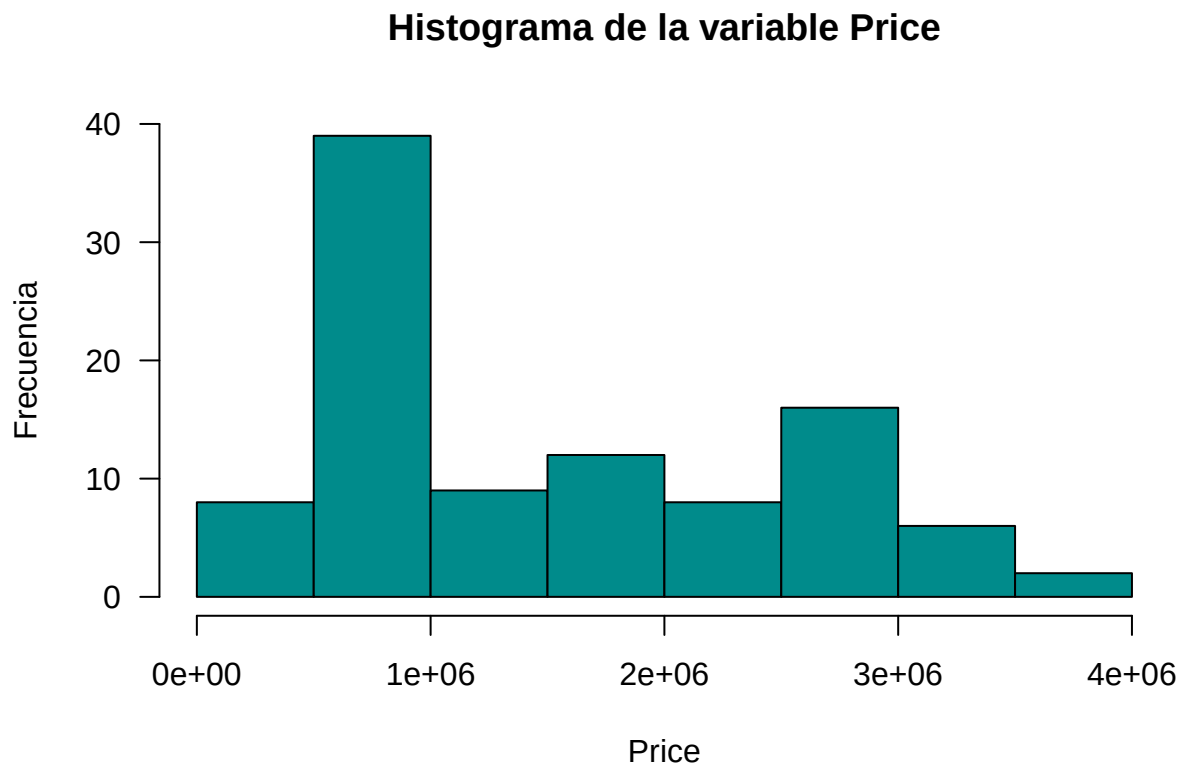
```
IQR(Price)
```

```
## [1] 1800000
```

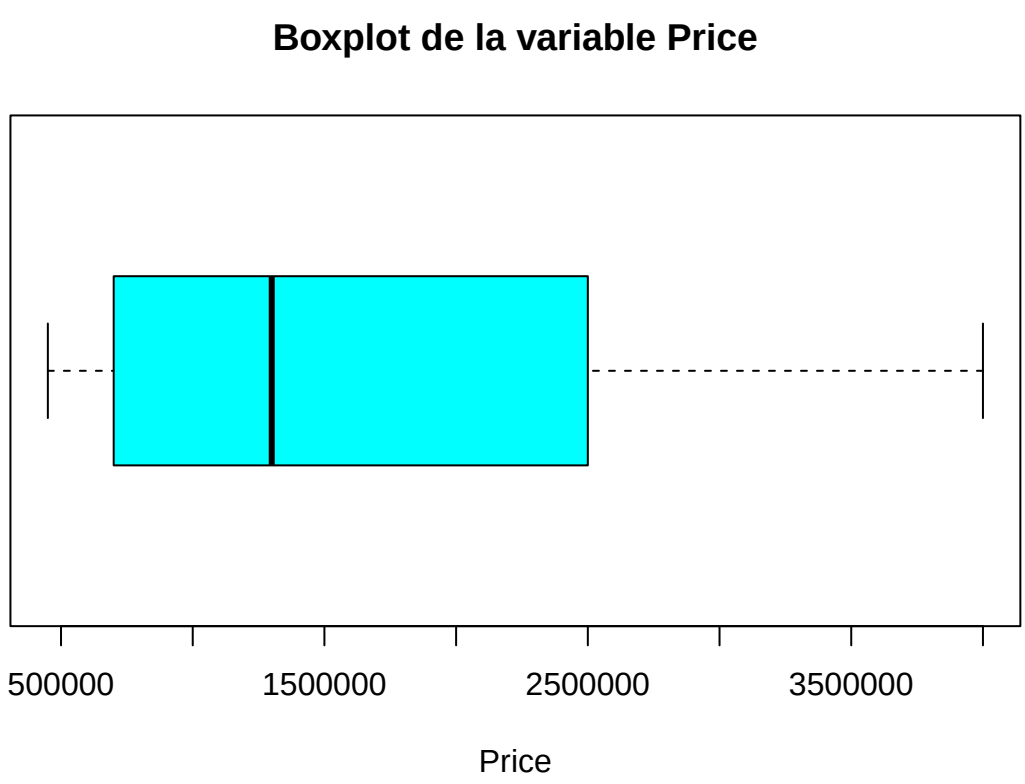
```
quantile(Price, c(0.25, 0.5, 0.75))
```

```
##     25%     50%     75%
##  700000 1300000 2500000
```

```
# Gráficas de la variable Price:
hist(Price, main = "Histograma de la variable Price", xlab = "Price",
     ylab = "Frecuencia", col = "darkcyan", las=1, ylim = c(0, 40))
```




```
boxplot(Price, main = "Boxplot de la variable Price", xlab = "Price",
        col = "cyan", horizontal=TRUE)
```



Análisis: El precio promedio de los carros de la muestra es de 1,574,000 unidades, con una desviación estándar de 1000265 unidades y una mediana de 1,300,000 unidades, siendo el valor mínimo de 450,000 unidades y el valor máximo de 4,000,000 unidades.

En cuanto al rango intercuartílico, este es de 1,800,000 unidades, siendo el 25% de los datos menores a 700,000 unidades, el 50% menores a 1,300,000 unidades y el 75% menores a 2,500,000 unidades. Notándose que la mayor parte de los datos se encuentran en el rango de 700,000 a 2,500,000 unidades.

En cuando a la distribución de los datos, tanto en el histograma como en el boxplot se observa que los datos están distribuidos de forma asimétrica, notándose un ligero desplazamiento de la mediana a la izquierda del boxplot, lo que significa que la distribución de los datos es asimétrica positiva, o sesgada a la derecha, lo que indica que los datos se encuentran más dispersos entre el segundo y tercer cuartil cuartil, y más concentrados entre el primer y segundo cuartil. Además, podemos evidenciar que la distribución de los datos es homogénea, ya que no se observa la presencia de valores atípicos en la muestra para esta variable.

Finalmente, podemos mencionar que a raíz de la asimetría de la distribución existe una sutil diferencia entre los valores de la media y la mediana, siendo la primera de 1,574,000 unidades y la segunda de 1,300,000 unidades, es decir, hay un margen de 274,000 unidades entre ambas medidas.

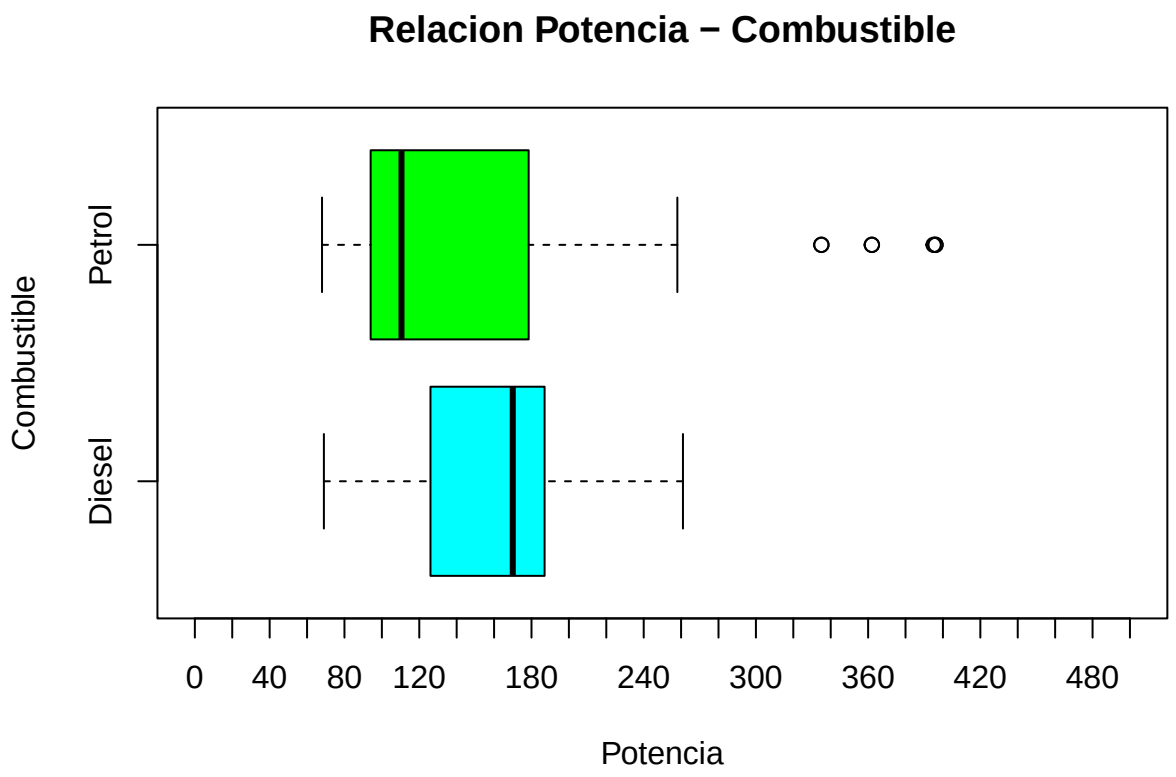
2. ¿Es cierto que la potencia promedio del carro es diferente si utiliza gasoil o gasolina? (Realice previamente un análisis descriptivo y luego la prueba de hipótesis adecuada. Apóyese en gráficos para su interpretación).

Análisis Descriptivo

Para responder a esta pregunta, primero realizaremos un análisis descriptivo de la variable Power, separando los datos en dos grupos, uno para los carros que utilizan gasolina y otro para los carros que utilizan gasoil, para ello usaremos la función subset para separar los datos y luego generaremos un histograma y un boxplot para visualizar la distribución de los datos.

```
# Boxplot para la relación potencia - combustible:
boxplot(split(Datos_car$Power, Datos_car$Fuel_Type), main =
          "Relacion Potencia - Combustible", xlab="Potencia", ylab="Combustible",
        col = c("cyan", "green", "purple"), horizontal = T, ylim = c(0, 500),
        xaxt = "n")

axis(side = 1, at = seq(0, 500, by = 20), labels = seq(0, 500, by = 20))
```




```
# Datos específicos para cada combustible en relación a la potencia:

# Gasolina (Petrol):
Potencia_Gasolina = Datos_car$Power[Datos_car$Fuel_Type=="Petrol"]

# Resumen de la variable:
summary(Potencia_Gasolina)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    68.0    94.0   110.5   157.5   177.2   396.0
```

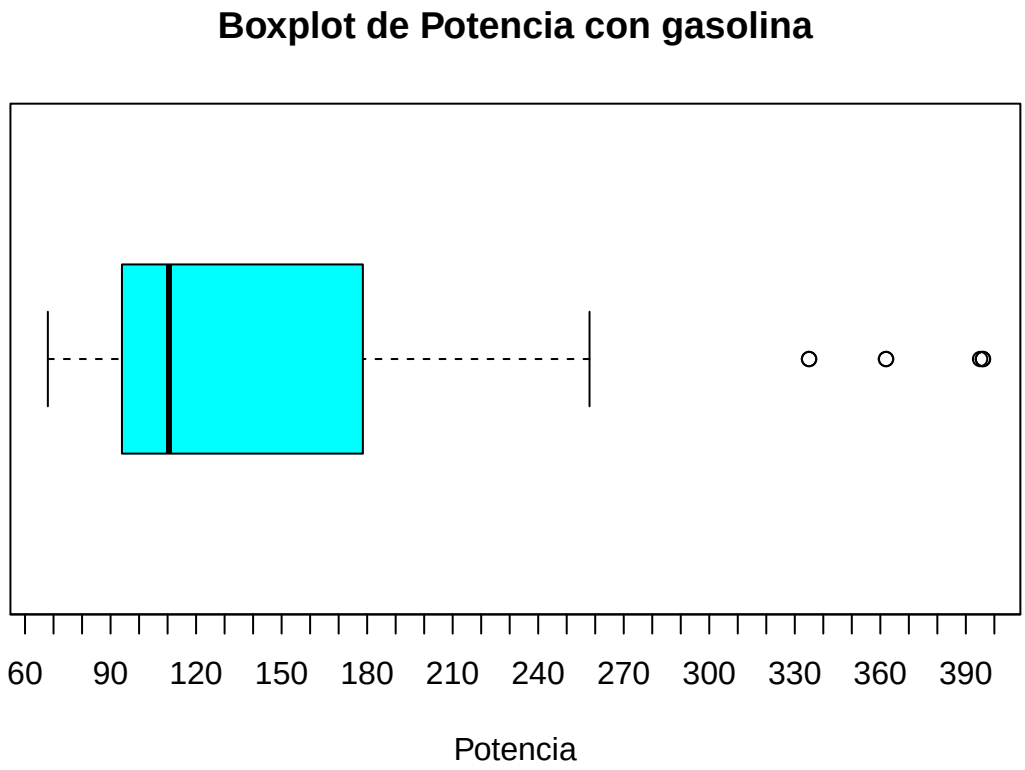
```
sd(Potencia_Gasolina)
```

```
## [1] 97.67082
```

```
quantile(Potencia_Gasolina, c(0.25, 0.5, 0.75))
```

```
##    25%    50%    75%
##  94.00 110.50 177.25
```

```
# Boxplot Individual de la variable
boxplot(Potencia_Gasolina, main = "Boxplot de Potencia con gasolina",
        xlab = "Potencia", col = "cyan", horizontal = T, xaxt = "n")
axis(side = 1, at = seq(0, 500, by = 10), labels = seq(0, 500, by = 10))
```



Análisis Potencia - Gasolina: Vemos que, en cuanto a la relación que existe entre la potencia y el uso de gasolina como combustible, tenemos principalmente que en este caso la potencia promedio es de 157.5 con una desviación estándar de 97.67082 y una mediana de 110.5

La potencia mínima bajo estas condiciones es de 68, mientras que la máxima es de 396. El 25% de los autos que usaron gasolina tuvieron una potencia de menos de 94, el 50% tuvieron su potencia menor a 110.50 y el 75% tuvo una potencia inferior a 177.25, con lo cual, la mayor parte de los datos se encuentran en el rango de 94 a 177.25 de potencia.

Así mismo, gracias al boxplot podemos evidenciar claramente la presencia de al menos 4 datos atípicos de entre los autos que fueron tomados en cuenta en la muestra analizada, los cuales corresponden con valores de potencia de: 396, 395, 362 y 335. Además, el gráfico también nos permite evidenciar con facilidad que existe una distribución asimétrica de los datos, siendo sesgada a la derecha, lo cual indica que la mayoría de los datos se encuentran más dispersos en el rango de 110.5 a 396 de potencia y más concentrados en el rango de 68 a 110.5 de potencia.

```
# Gasoil (Diesel):
Potencia_Gasoil = Datos_car$Power[Datos_car$Fuel_Type=="Diesel"]

# Resumen de la variable
summary(Potencia_Gasoil)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    69.0   126.0   170.0   158.8   187.0   261.0
```

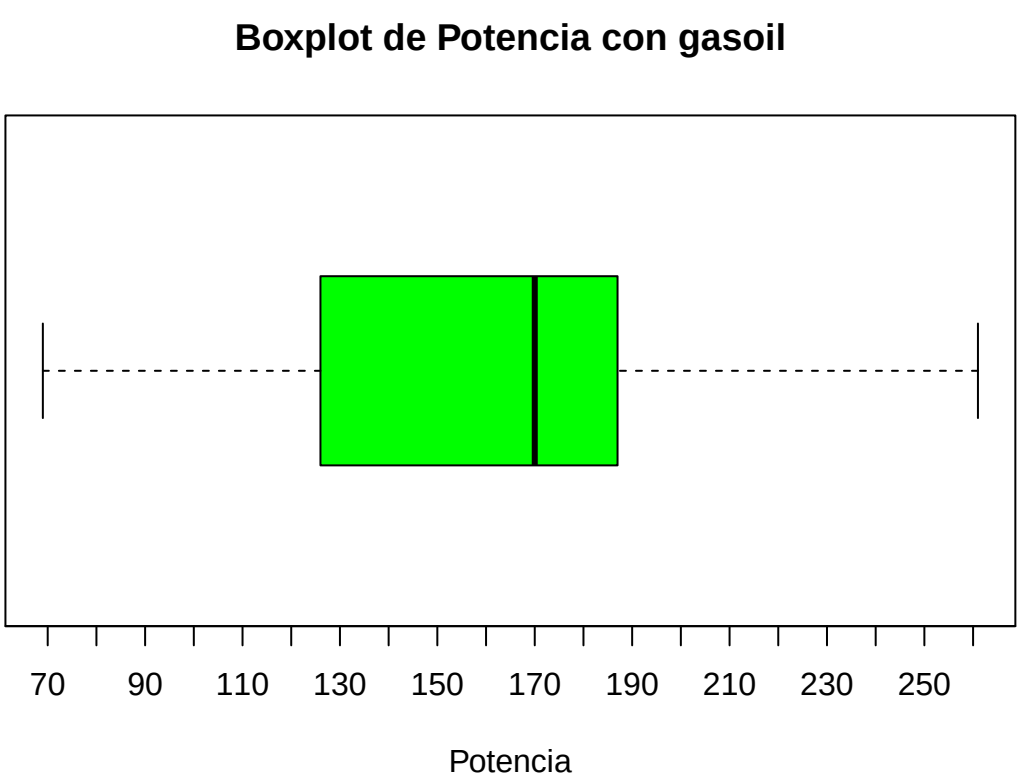
```
sd(Potencia_Gasoil)
```

```
## [1] 46.10887
```

```
quantile(Potencia_Gasoil, c(0.25, 0.5, 0.75))
```

```
## 25% 50% 75%
## 126 170 187
```

```
# Boxplot Individual de la variable
boxplot(Potencia_Gasoil, main = "Boxplot de Potencia con gasoil",
        xlab = "Potencia", col = "green", horizontal = T, xaxt = "n")
axis(side = 1, at = seq(0, 500, by = 10), labels = seq(0, 500, by = 10))
```



Análisis Potencia - Gasoil/Diesel: Vemos que en cuanto a la relación que existe con la potencia y el uso de gasoil/Diesel como combustible, tenemos principalmente que la potencia promedio fue de 158.8 con una desviación estándar de 46.10887 y una mediana de 170.00.

En este caso, la potencia mínima fue de 69 mientras que la máxima fue de 261, lo cual representa un rango de potencia menor que el de los autos que usaron gasolina como combustible.

Aquí, 25% de los autos tuvieron una potencia de menos de 126, el 50% tuvo su potencia menor a 170 y el 75% tuvo una potencia inferior a 187, con lo cual, la mayor parte de los datos se encuentran en el rango de 126 a 187 de potencia, es decir, en este caso tenemos un rango intercuartílico de 61, lo cual es menor que el rango intercuartílico de los autos que usaron gasolina, el cual fue de 83.

Adicionalmente, vemos que a diferencia de lo ocurrido con la gasolina, usando gasoil no hay presencia de datos atípicos en el boxplot manteniéndose también una distribución asimétrica pero en este caso sesgada a la izquierda, lo cual indica que la mayoría de los datos se encuentran más dispersos en el rango de 126 a 170 de potencia y más concentrados en el rango de 170 a 187 de potencia.

Conclusión Preliminar: A partir de los análisis descriptivos realizados, podemos decir que la potencia promedio de los carros que utilizan gasolina es de 157.5 hp, mientras que la potencia promedio de los carros que utilizan gasoil es de 158.8 hp, por lo que, en efecto la potencia promedio de los carros que utilizan gasoil y los que utilizan gasolina es diferente, sin embargo, para poder afirmar esto con certeza, es necesario realizar una prueba de hipótesis que nos permita determinar si esta diferencia es significativa o no.

Prueba de Hipótesis

Para realizar la prueba de hipótesis, planteamos las siguientes hipótesis:

- H0: La potencia promedio de los carros que utilizan gasolina es igual a la potencia promedio de los carros que utilizan gasoil.
- Ha: La potencia promedio de los carros que utilizan gasolina es diferente a la potencia promedio de los carros que utilizan gasoil.

Para realizar la prueba de hipótesis, usaremos la función t.test, la cual nos permitirá realizar una prueba t de dos muestras independientes, para ello usaremos las muestras de potencia de los carros que utilizan gasolina y los que utilizan gasoil, y estableceremos un nivel de significancia del 1%, es decir, un nivel de confianza del 99%, lo cual nos permitirá rechazar la hipótesis nula si el p-valor es menor a 0.01

```
# Prueba de Hipótesis:
t.test(Potencia_Gasolina, Potencia_Gasoil, alternative = "two.sided",
       mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.99)
```

```
##
##  Welch Two Sample t-test
##
## data:  Potencia_Gasolina and Potencia_Gasoil
## t = -0.086971, df = 73.923, p-value = 0.9309
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  -41.21354  38.58854
## sample estimates:
## mean of x mean of y
##  157.5000  158.8125
```

Conclusión: A partir de los resultados obtenidos de la prueba de hipótesis, podemos decir que, como el p-valor obtenido es de 0.9309, el cual es mayor al nivel de significancia establecido, el cual fue de 0.01, no tenemos suficiente evidencia para rechazar la hipótesis nula, por lo que no podemos afirmar que la potencia promedio de los carros que utilizan gasolina es diferente a la potencia promedio de los carros que utilizan gasoil. Esto indica que no hay evidencia suficiente para afirmar que existe una diferencia significativa en las medias de potencia entre los carros que usan gasolina y aquellos que usan gasoil, incluso con un nivel de confianza del 99%, por lo que podemos decir que, desde el punto de vista estadístico, las potencias medias de ambos podrían considerarse equivalentes con el nivel de confianza utilizado.