Precios de Autos Entrega 1

Autor: Rodolfo Sandoval Matricula: A01720253 Desc: Momento de Retroalimentación: Módulo 1 Técnicas de procesamiento de datos para el análisis estadístico y para la construcción de modelos (Portafolio Análisis) Una empresa automovilística china aspira a entrar en el mercado estadounidense. Desea establecer allí una unidad de fabricación y producir automóviles localmente para competir con sus contrapartes estadounidenses y europeas. Contrataron una empresa de consultoría de automóviles para identificar los principales factores de los que depende el precio de los automóviles, específicamente, en el mercado estadounidense, ya que pueden ser muy diferentes del mercado chino. Esencialmente, la empresa quiere saber:

Qué variables son significativas para predecir el precio de un automóvil Qué tan bien describen esas variables el precio de un automóvil

#Carga de librerias y del conjunto de datos desde un archivo csv

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.1.3
```

```
## corrplot 0.92 loaded
```

```
library(MASS)
library(moments)
```

```
## Warning: package 'moments' was built under R version 4.1.3
```

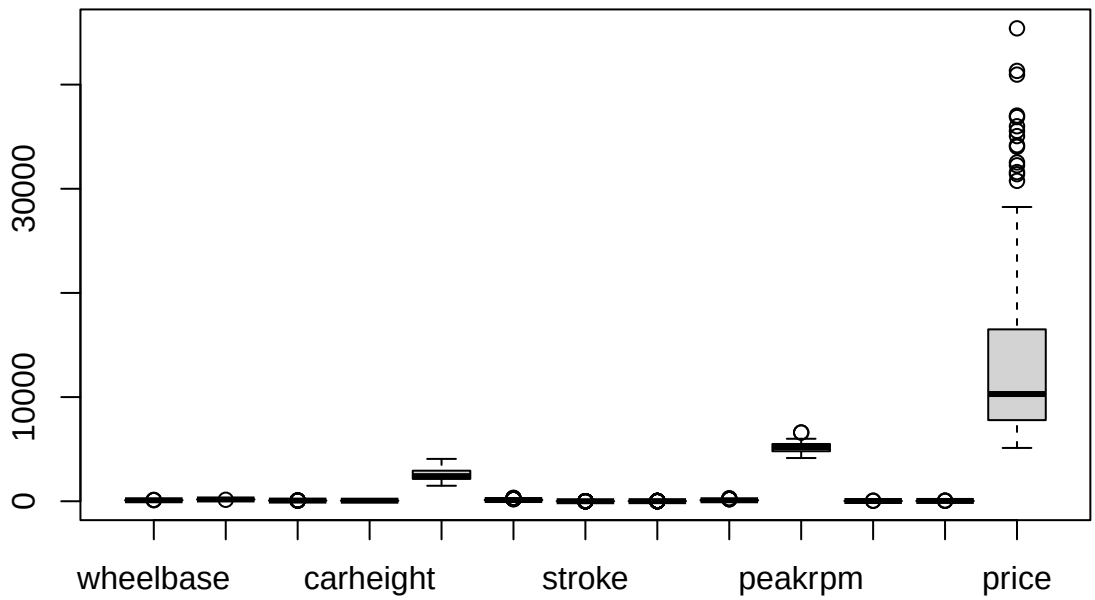
```
data = read.csv("precios_autos.csv")
```

#Exploracion de la base de datos

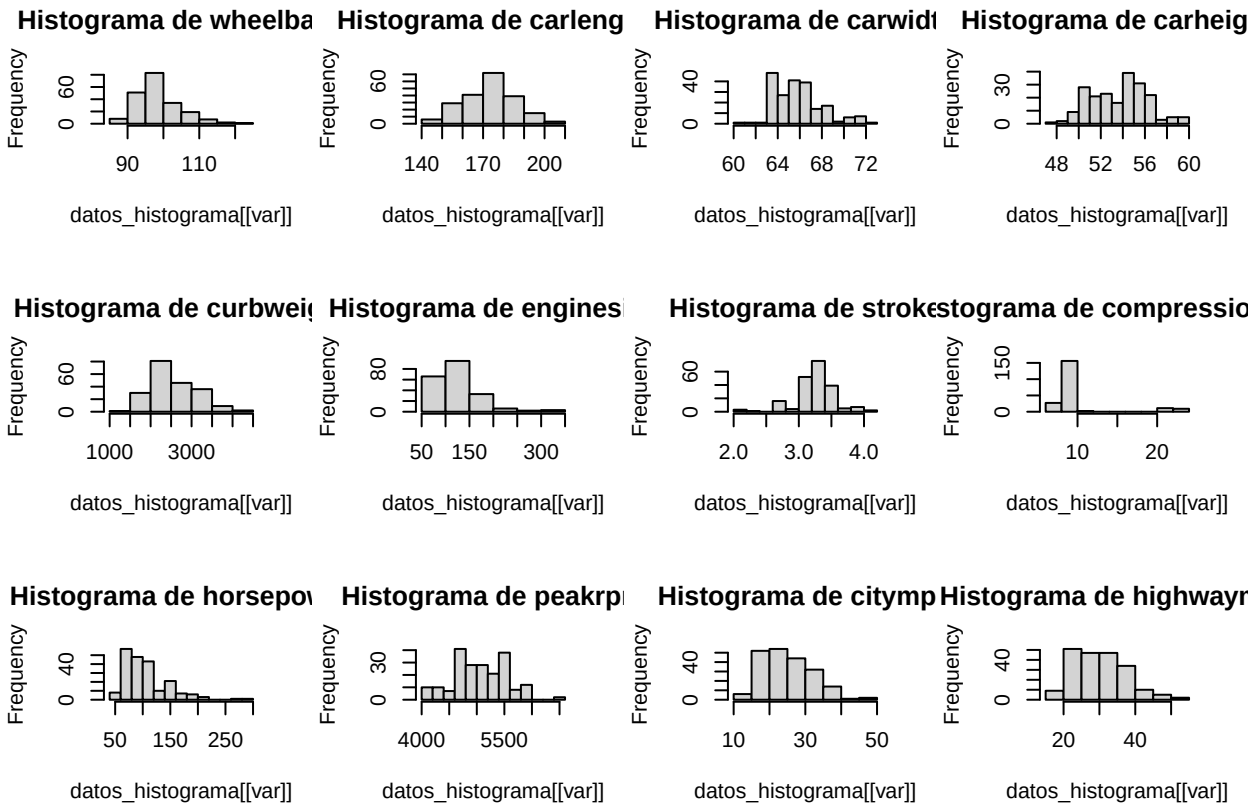
```
valor_cuantitativo = summary(data[, sapply(data, is.numeric)])
valor_categorico = summary(data[, sapply(data, is.character)])
```

#Visualizar variables cuantitativas

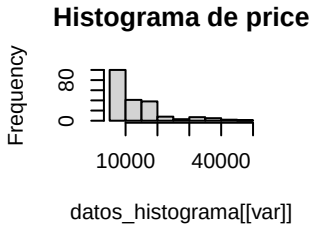
```
variables_cuantitativas = c("wheelbase", "carlength", "carwidth", "carheight", "curbweight", "enginesize
                "stroke", "compressionratio", "horsepower", "peakrpm", "citympg", "highwaympg", "price"

datos_cuantitativos = data[, variables_cuantitativas]

# Boxplots
boxplot(datos_cuantitativos, main="Boxplots de Variables Cuantitativas")
```




```
# Histogramas solo para las variables cuantitativas
datos_histograma = datos_cuantitativos
par(mfrow=c(3, 4))
for (var in colnames(datos_histograma)) {
  hist(datos_histograma[[var]], main=paste("Histograma de", var))
}
```

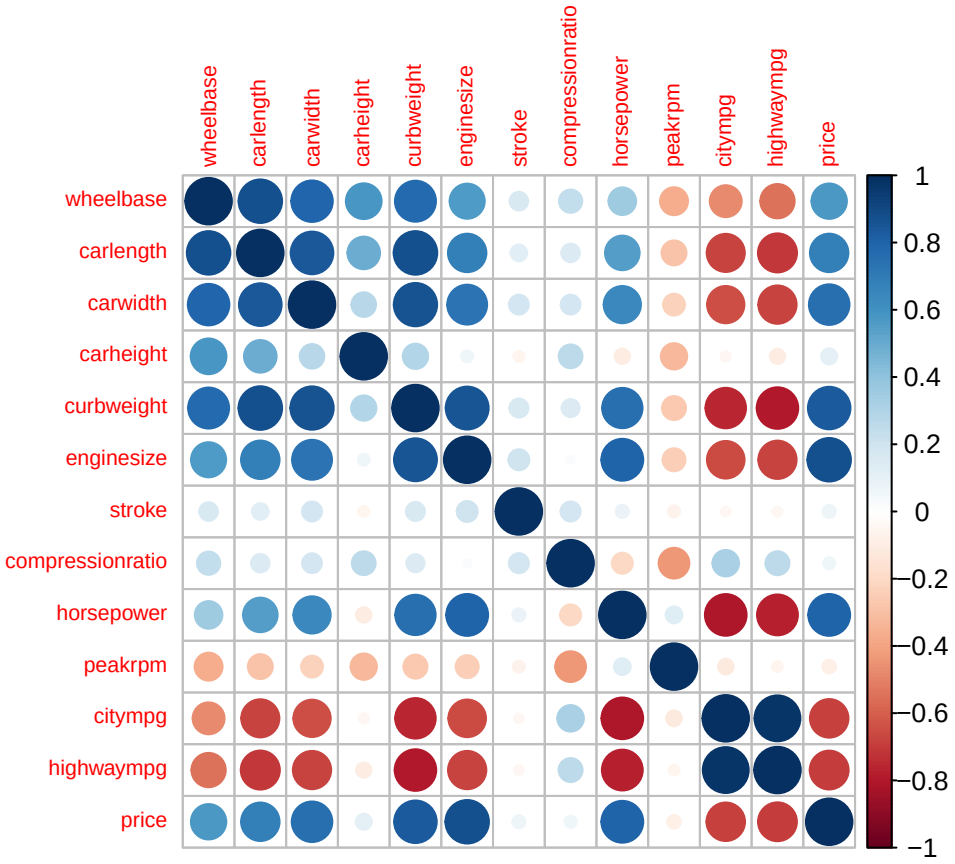



```
par(mfrow=c(1, 1))
```



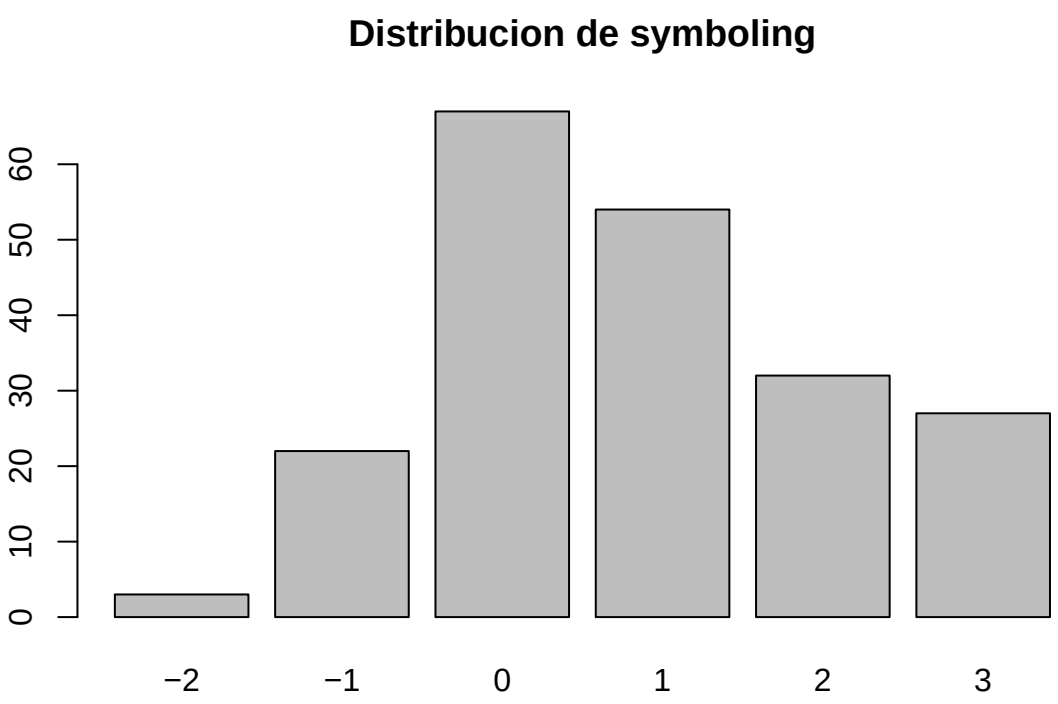
#Correlacion y dispersion

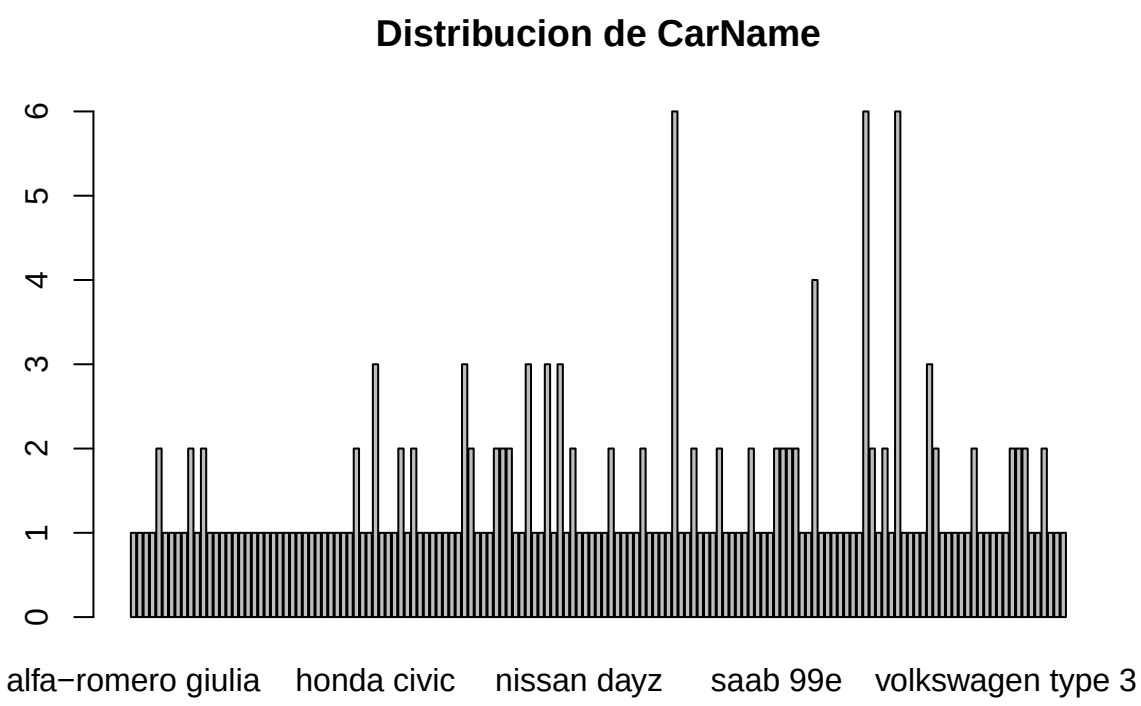
```
matriz_correlacion = cor(datos_cuantitativos)
corrplot(matriz_correlacion, method="circle", type="full", tl.cex=0.7)
```

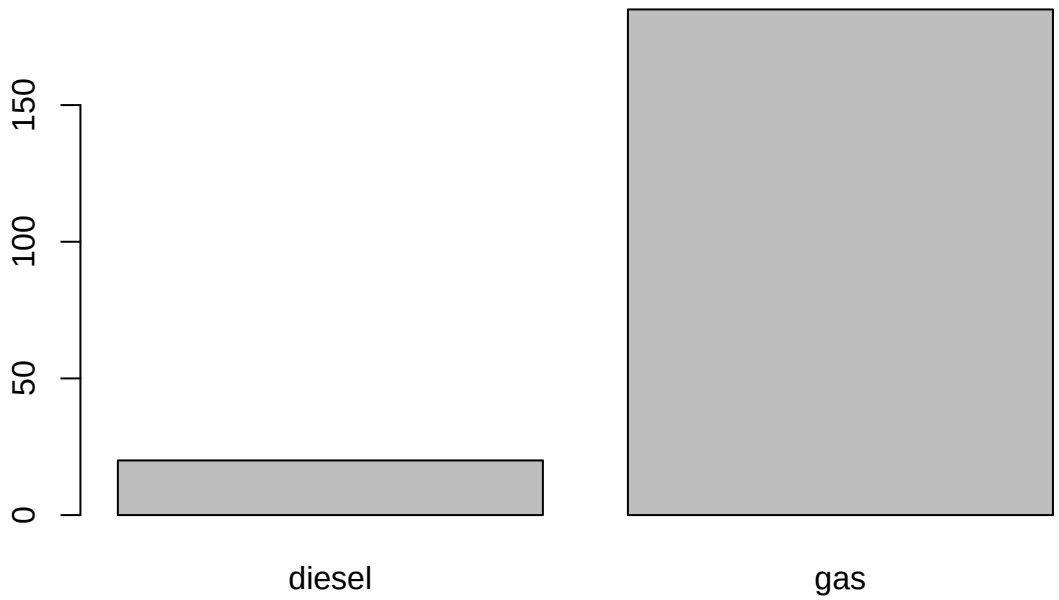


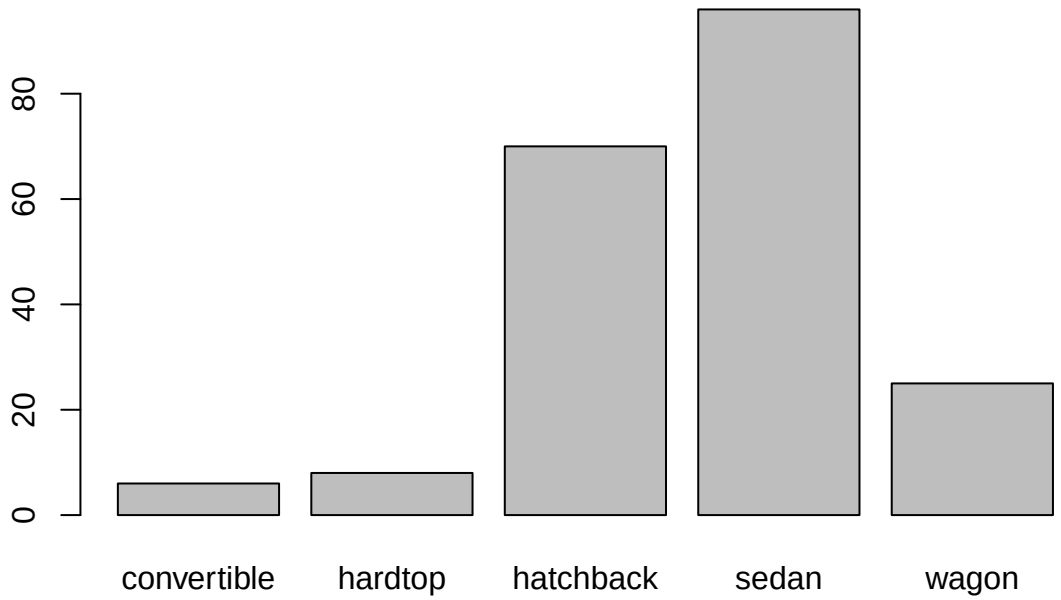
#Visualizar variables categoricas ##Distribucion de los datos

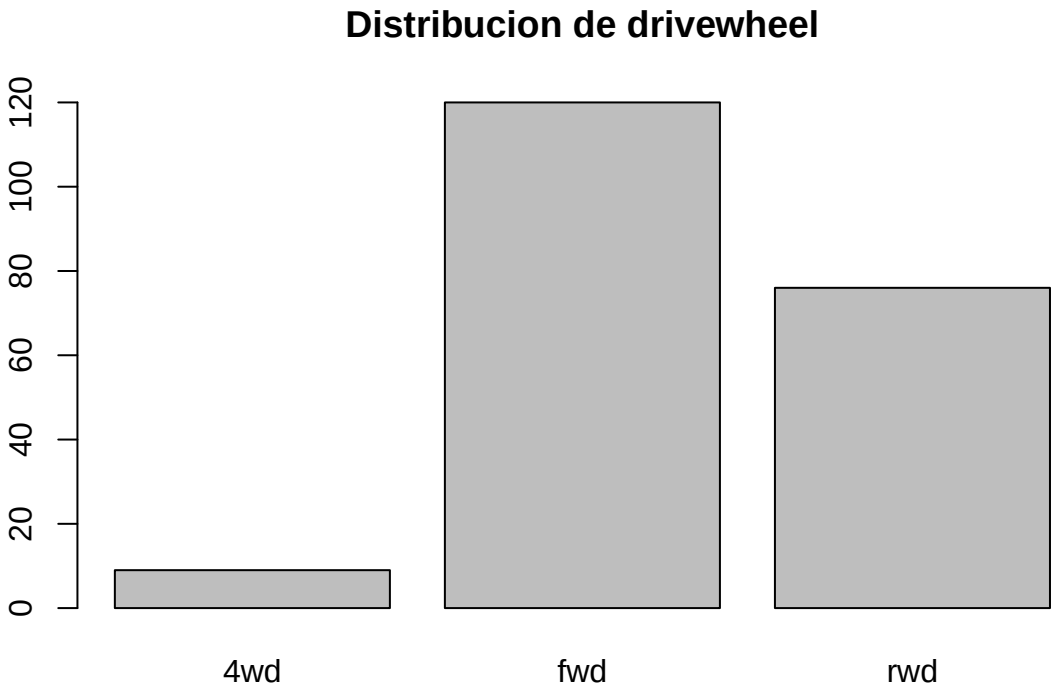
```
variables_categoricos = c("symboling", "CarName", "fueltype", "carbody", "drivewheel", "enginelocation"

for (var in variables_categoricos) {
  barplot(table(data[[var]]), main=paste("Distribucion de", var))
}
```

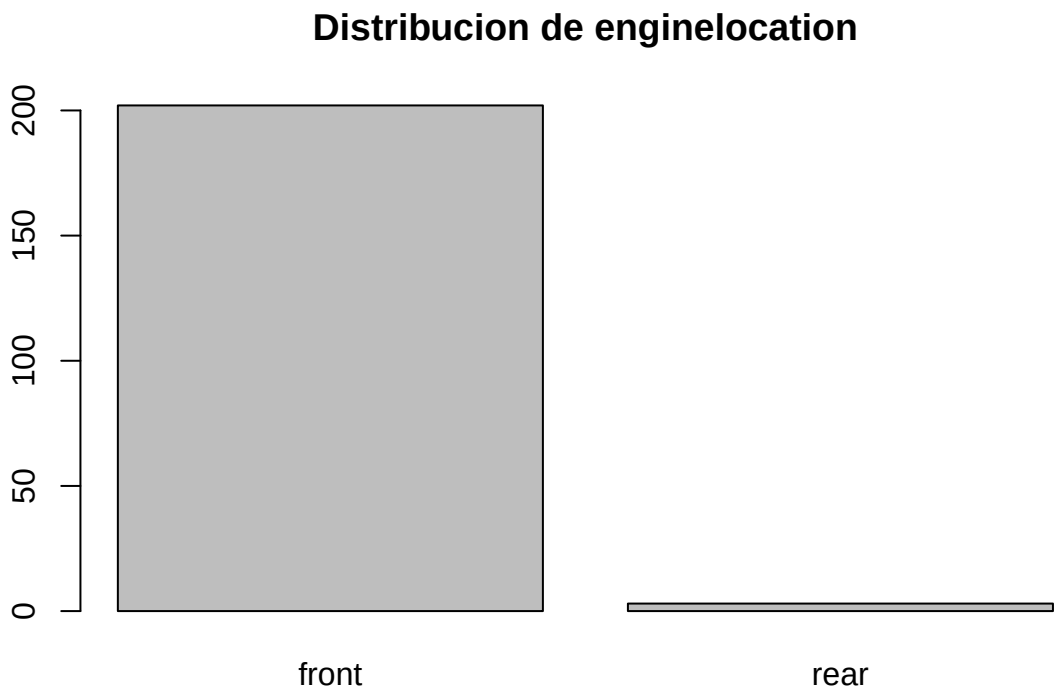

Distribucion de symboling
0
10
20
30
40
50
60
−2
−1
0
1
2
3

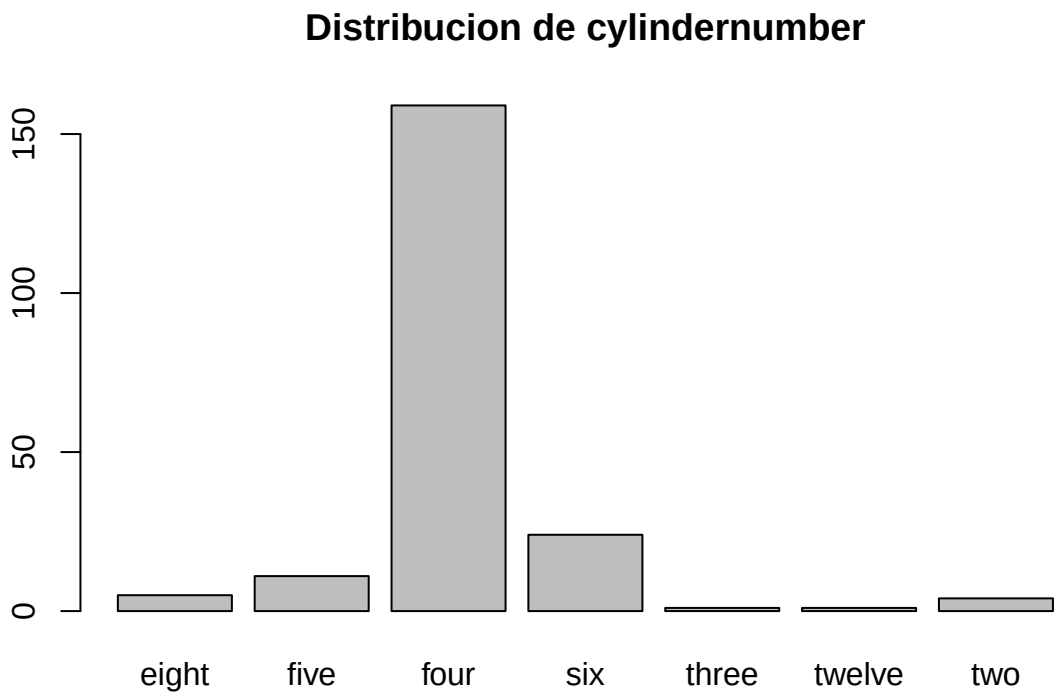

Distribucion de CarName
0
1
2
3
4
5
6
alfa−romero giulia
honda civic
nissan dayz
saab 99e
volkswagen type 3


Distribucion de fueltype
0
50
100
150
diesel
gas


Distribucion de carbody
0
20
40
60
80
convertible
hardtop
hatchback
sedan
wagon

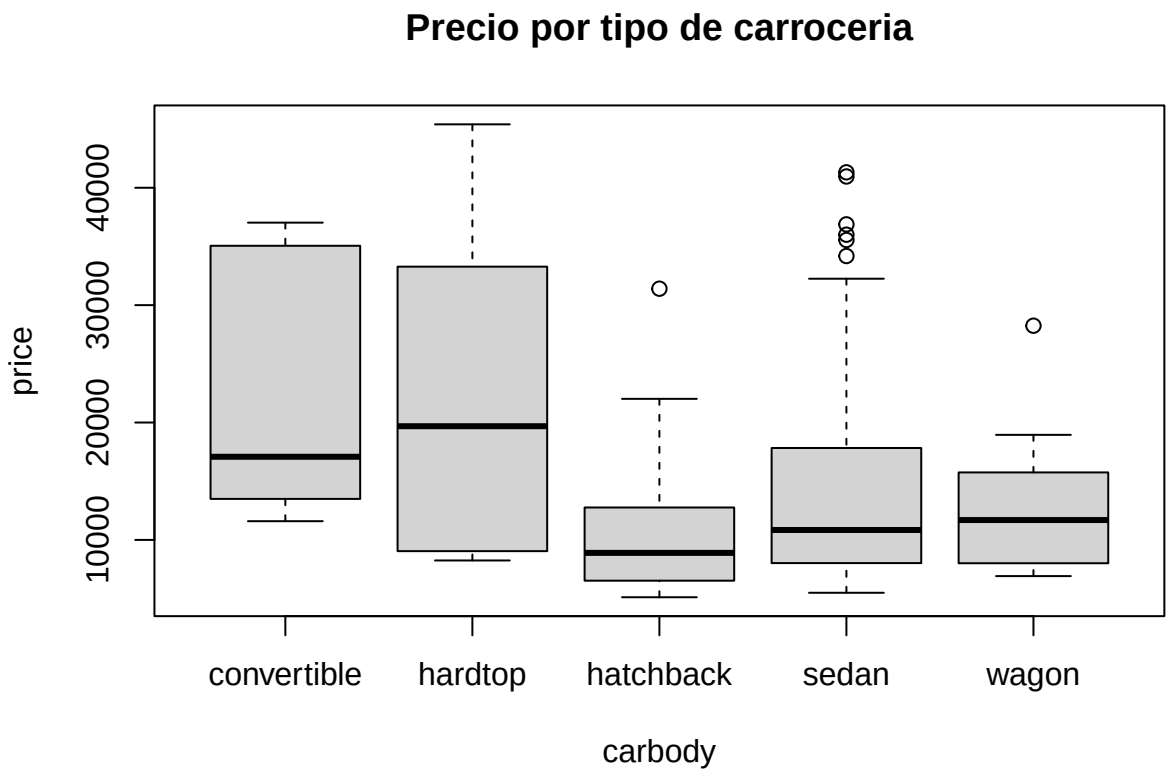

Distribucion de drivewheel
0
20
40
60
80
100
120
4wd
fwd
rwd


Distribucion de enginelocation
0
50
100
150
200
front
rear



##Colinealidad

```
boxplot(price ~ carbody, data, main="Precio por tipo de carroceria")
```



#Seleccionar variables para el analisis

```
variables = data[, c("wheelbase", "carlength", "carwidth", "curbweight", "enginesize", "horsepower", "pr
```

Preparacion de los datos

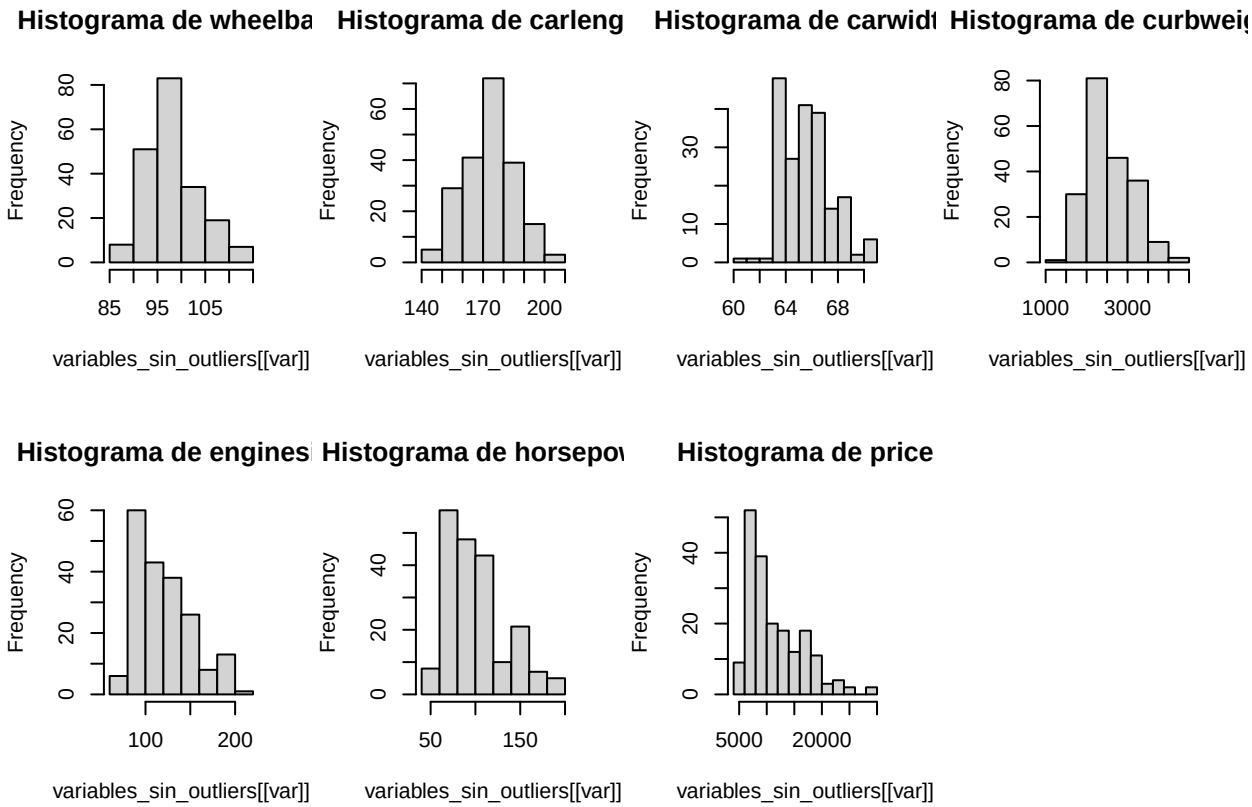
Transformacion

```
# Eliminar valores atipicos y ceros en las variables seleccionadas
umbral_atipico = 1.5
umbral_cero = 1e-6

variables_sin_outliers <- variables
for (var in colnames(variables)) {
  q1 = quantile(variables[[var]], 0.25, na.rm = TRUE)
  q3 = quantile(variables[[var]], 0.75, na.rm = TRUE)
  iqr = q3 - q1
  upper_limit = q3 + umbral_atipico * iqr
  lower_limit = q1 - umbral_atipico * iqr
  variables_sin_outliers[[var]][variables[[var]] > upper_limit] = NA
  variables_sin_outliers[[var]][variables[[var]] < lower_limit] = NA
  variables_sin_outliers[[var]][abs(variables[[var]]) < umbral_cero] = NA
}
```

```r
# Eliminar filas con valores faltantes
datos_limpiados <- na.omit(variables_sin_outliers)
```

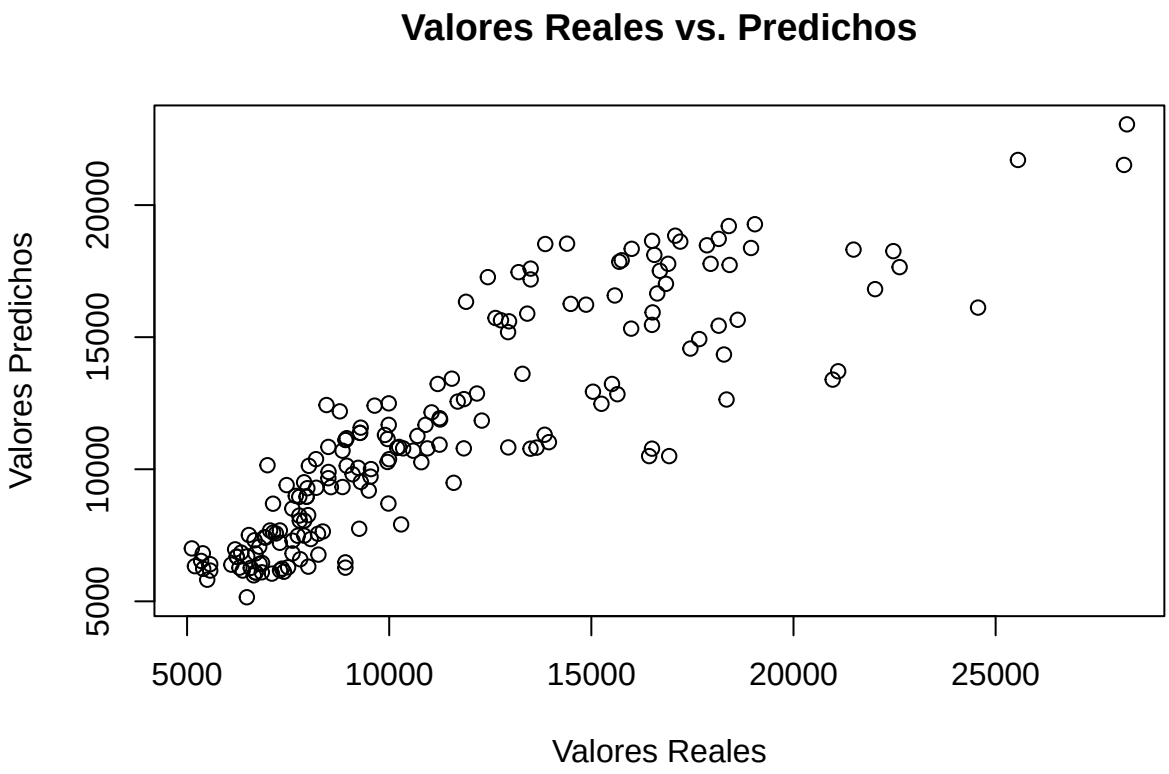
#Graficamos las nuevas distribuciones con los datos seleccionados

```r
par(mfrow=c(2, 4))
for (var in colnames(variables_sin_outliers)) {
  hist(variables_sin_outliers[[var]], main=paste("Histograma de", var))
}
par(mfrow=c(1, 1))
```

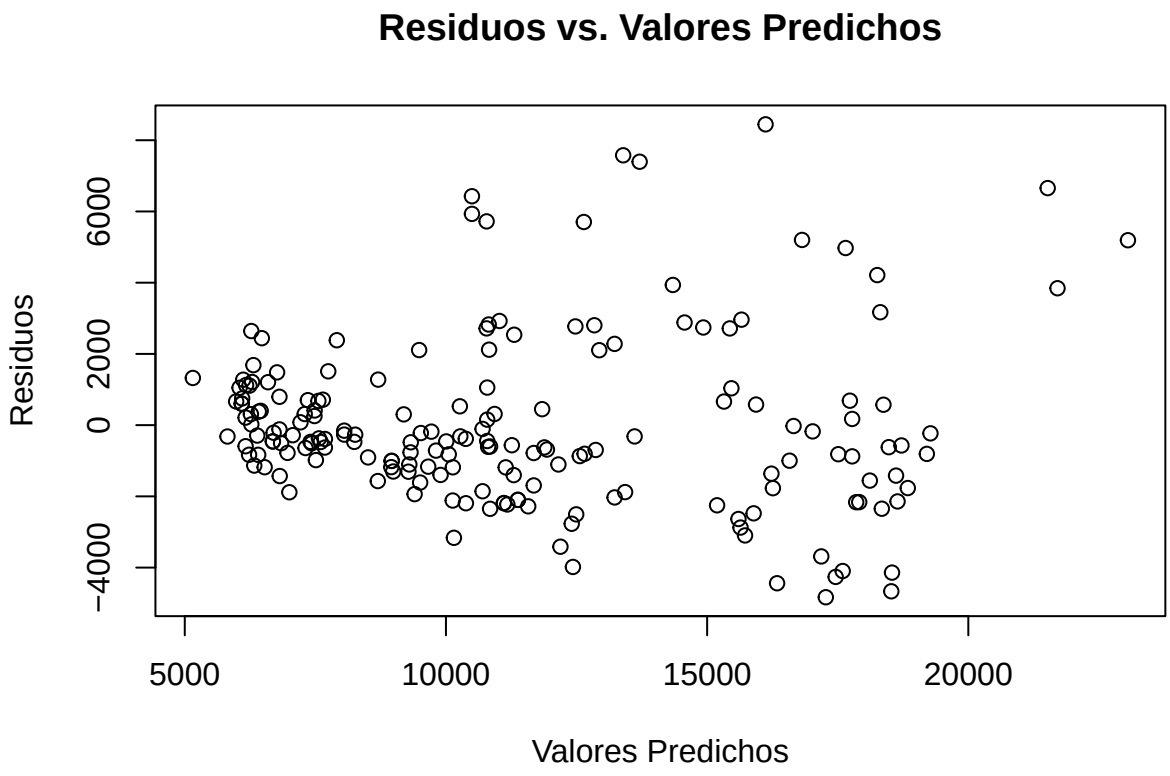


#Graficamos el analisis de los datos seleccionados con regresion lineal

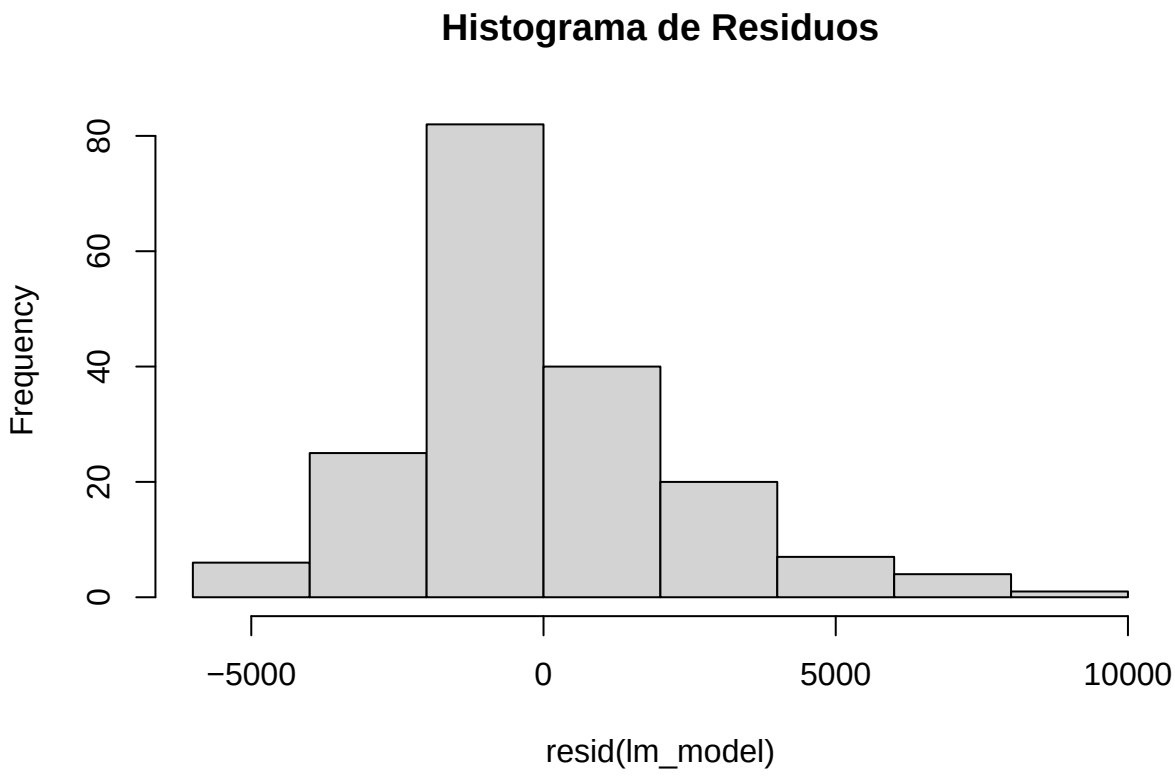
```r
# Ajustar el modelo de regresión
lm_model = lm(price ~ ., data = datos_limpiados)

# Gráfico de dispersión de los valores reales vs. predichos
plot(datos_limpiados$price, predict(lm_model), main="Valores Reales vs. Predichos", xlab="Valores Reales
```




```r
# Gráfico de residuos vs. valores predichos
plot(predict(lm_model), resid(lm_model), main="Residuos vs. Valores Predichos", xlab="Valores Predichos
```




```r
# Gráfico de histograma de los residuos
hist(resid(lm_model), main="Histograma de Residuos")
```



#Sacamos kurtosis y sesgo para cada variable

```
valor_kurtosis = sapply(datos_limpiados, kurtosis)
valor_sesgo = sapply(datos_limpiados, skewness)
valor_kurtosis
```

```
##  wheelbase  carlength   carwidth curbweight enginesize horsepower      price
##   3.986458   2.730903   2.830602   2.428882   2.810534   2.781552   3.781622
```

```
valor_sesgo
```

```
##  wheelbase  carlength   carwidth curbweight enginesize horsepower      price
## 1.03854988 0.01717055 0.60831427 0.52939034 0.74363938 0.77853891 1.07954608
```