Actividad Integradora 2

Nallely Serna

2024-09-06

```r
# Carga de datos
M <- read.csv("precios_autos.csv", stringsAsFactors = TRUE)

# Exploración inicial
head(M)
```

```
##   symboling                  CarName fueltype     carbody drivewheel
## 1         3       alfa-romero giulia      gas convertible        rwd
## 2         3      alfa-romero stelvio      gas convertible        rwd
## 3         1 alfa-romero Quadrifoglio      gas   hatchback        rwd
## 4         2              audi 100 ls      gas       sedan        fwd
## 5         2               audi 100ls      gas       sedan        4wd
## 6         2                 audi fox      gas       sedan        fwd
##   enginelocation wheelbase carlength carwidth carheight curbweight
enginetype
## 1          front      88.6     168.8     64.1      48.8       2548
dohc
## 2          front      88.6     168.8     64.1      48.8       2548
dohc
## 3          front      94.5     171.2     65.5      52.4       2823
ohcv
## 4          front      99.8     176.6     66.2      54.3       2337
ohc
## 5          front      99.4     176.6     66.4      54.3       2824
ohc
## 6          front      99.8     177.3     66.3      53.1       2507
ohc
##   cylindernumber enginesize stroke compressionratio horsepower peakrpm
citympg
## 1           four        130   2.68              9.0        111    5000
21
## 2           four        130   2.68              9.0        111    5000
21
## 3            six        152   3.47              9.0        154    5000
19
## 4           four        109   3.40             10.0        102    5500
24
## 5           five        136   3.40              8.0        115    5500
18
## 6           five        136   3.40              8.5        110    5500
19
##   highwaympg price
```

```
## 1         27 13495
## 2         27 16500
## 3         26 16500
## 4         30 13950
## 5         22 17450
## 6         25 15250

summary(M)

##    symboling                 CarName      fueltype          carbody
##  Min.   :-2.0000   peugeot 504   :  6   diesel: 20   convertible: 6
##  1st Qu.: 0.0000   toyota corolla:  6   gas   :185   hardtop    : 8
##  Median : 1.0000   toyota corona :  6                hatchback  :70
##  Mean   : 0.8341   subaru dl     :  4                sedan      :96
##  3rd Qu.: 2.0000   honda civic   :  3                wagon      :25
##  Max.   : 3.0000   mazda 626     :  3
##                    (Other)       :177
##  drivewheel enginelocation   wheelbase        carlength
carwidth
##  4wd:  9    front:202      Min.   : 86.60   Min.   :141.1   Min.
:60.30
##  fwd:120    rear :  3      1st Qu.: 94.50   1st Qu.:166.3   1st
Qu.:64.10
##  rwd: 76                   Median : 97.00   Median :173.2   Median
:65.50
##                            Mean   : 98.76   Mean   :174.0   Mean
:65.91
##                            3rd Qu.:102.40   3rd Qu.:183.1   3rd
Qu.:66.90
##                            Max.   :120.90   Max.   :208.1   Max.
:72.30
##
##    carheight       curbweight   enginetype  cylindernumber
enginesize
##  Min.   :47.80   Min.   :1488   dohc : 12   eight :  5     Min.   :
61.0
##  1st Qu.:52.00   1st Qu.:2145   dohcv:  1   five  : 11     1st Qu.:
97.0
##  Median :54.10   Median :2414   l    : 12   four  :159     Median
:120.0
##  Mean   :53.72   Mean   :2556   ohc  :148   six   : 24     Mean
:126.9
##  3rd Qu.:55.50   3rd Qu.:2935   ohcf : 15   three :  1     3rd
Qu.:141.0
##  Max.   :59.80   Max.   :4066   ohcv : 13   twelve:  1     Max.
:326.0
##                                 rotor:  4   two   :  4
##      stroke      compressionratio   horsepower       peakrpm
##  Min.   :2.070   Min.   : 7.00    Min.   : 48.0   Min.   :4150
##  1st Qu.:3.110   1st Qu.: 8.60    1st Qu.: 70.0   1st Qu.:4800
```

```
##  Median :3.290   Median : 9.00    Median : 95.0   Median :5200
##  Mean   :3.255   Mean   :10.14    Mean   :104.1   Mean   :5125
##  3rd Qu.:3.410   3rd Qu.: 9.40    3rd Qu.:116.0   3rd Qu.:5500
##  Max.   :4.170   Max.   :23.00    Max.   :288.0   Max.   :6600
##
##     citympg        highwaympg        price
##  Min.   :13.00   Min.   :16.00   Min.   : 5118
##  1st Qu.:19.00   1st Qu.:25.00   1st Qu.: 7788
##  Median :24.00   Median :30.00   Median :10295
##  Mean   :25.22   Mean   :30.75   Mean   :13277
##  3rd Qu.:30.00   3rd Qu.:34.00   3rd Qu.:16503
##  Max.   :49.00   Max.   :54.00   Max.   :45400
##
```

```
str(M)
```

```
## 'data.frame':    205 obs. of  21 variables:
##  $ symboling       : int  3 3 1 2 2 2 1 1 1 0 ...
##  $ CarName         : Factor w/ 147 levels "alfa-romero giulia",..: 1 3
2 4 5 9 5 7 6 8 ...
##  $ fueltype        : Factor w/ 2 levels "diesel","gas": 2 2 2 2 2 2 2
2 2 2 ...
##  $ carbody         : Factor w/ 5 levels "convertible",..: 1 1 3 4 4 4
4 5 4 3 ...
##  $ drivewheel      : Factor w/ 3 levels "4wd","fwd","rwd": 3 3 3 2 1 2
2 2 2 1 ...
##  $ enginelocation  : Factor w/ 2 levels "front","rear": 1 1 1 1 1 1 1
1 1 1 ...
##  $ wheelbase       : num  88.6 88.6 94.5 99.8 99.4 ...
##  $ carlength       : num  169 169 171 177 177 ...
##  $ carwidth        : num  64.1 64.1 65.5 66.2 66.4 66.3 71.4 71.4 71.4
67.9 ...
##  $ carheight       : num  48.8 48.8 52.4 54.3 54.3 53.1 55.7 55.7 55.9
52 ...
##  $ curbweight      : int  2548 2548 2823 2337 2824 2507 2844 2954 3086
3053 ...
##  $ enginetype      : Factor w/ 7 levels "dohc","dohcv",..: 1 1 6 4 4 4
4 4 4 4 ...
##  $ cylindernumber  : Factor w/ 7 levels "eight","five",..: 3 3 4 3 2 2
2 2 2 2 ...
##  $ enginesize      : int  130 130 152 109 136 136 136 136 131 131 ...
##  $ stroke          : num  2.68 2.68 3.47 3.4 3.4 3.4 3.4 3.4 3.4 3.4
...
##  $ compressionratio: num  9 9 9 10 8 8.5 8.5 8.5 8.3 7 ...
##  $ horsepower      : int  111 111 154 102 115 110 110 110 140 160 ...
##  $ peakrpm         : int  5000 5000 5000 5500 5500 5500 5500 5500 5500
5500 ...
##  $ citympg         : int  21 21 19 24 18 19 19 19 17 16 ...
##  $ highwaympg      : int  27 27 26 30 22 25 25 25 20 22 ...
##  $ price           : num  13495 16500 16500 13950 17450 ...
```

##Exploración de la base de datos ##Calcula medidas estadísticas apropiadas para las variables:

```
# Medidas estadísticas para las variables cuantitativas
mean(M$carheight)

## [1] 53.72488

sd(M$carheight)

## [1] 2.443522

quantile(M$carheight)

##   0%  25%  50%  75% 100%
## 47.8 52.0 54.1 55.5 59.8

summary(M$carheight)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   47.80   52.00   54.10   53.72   55.50   59.80

mean(M$carwidth)

## [1] 65.9078

sd(M$carwidth)

## [1] 2.145204

quantile(M$carwidth)

##   0%  25%  50%  75% 100%
## 60.3 64.1 65.5 66.9 72.3

summary(M$carwidth)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   60.30   64.10   65.50   65.91   66.90   72.30

mean(M$price)

## [1] 13276.71

sd(M$price)

## [1] 7988.852

quantile(M$price)

##    0%   25%   50%   75%  100%
##  5118  7788 10295 16503 45400

summary(M$price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5118    7788   10295   13277   16503   45400
```

```
# Análisis de frecuencias para la variable 'carbody'
table(M$carbody)
```

```
##
## convertible     hardtop   hatchback       sedan       wagon
##           6           8          70          96          25
```

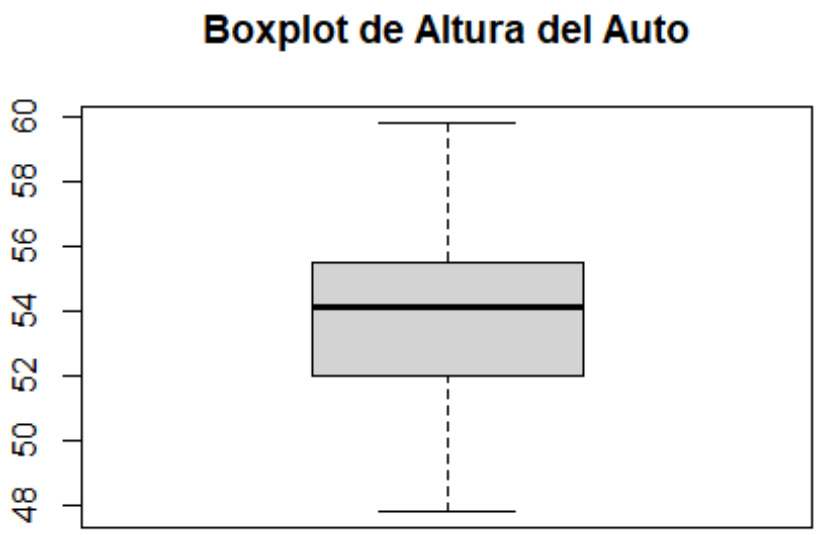
```
prop.table(table(M$carbody))
```

```
##
## convertible     hardtop   hatchback       sedan       wagon
##  0.02926829  0.03902439  0.34146341  0.46829268  0.12195122
```

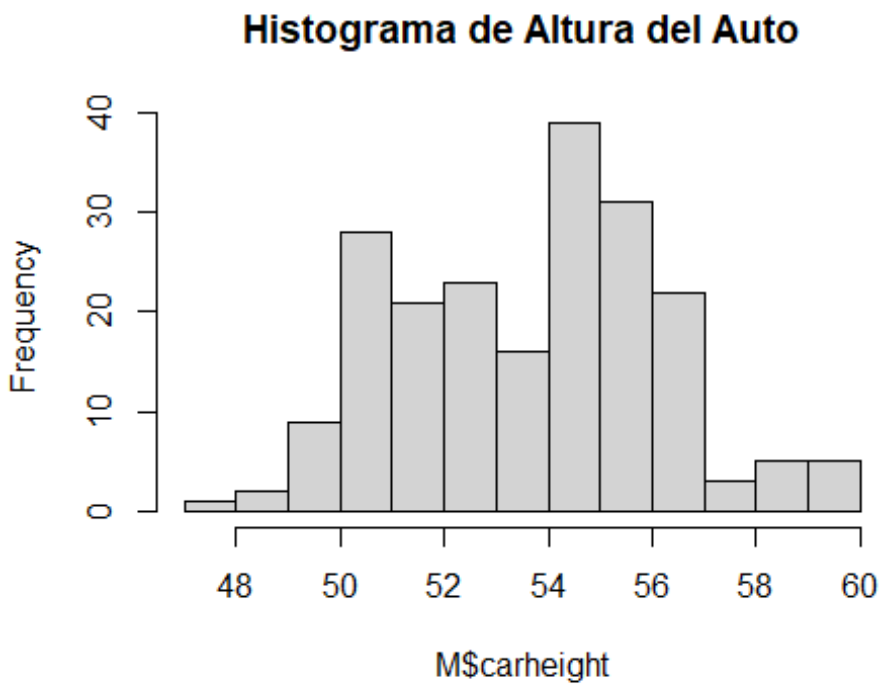
```
# Matriz de correlación entre las variables cuantitativas
cor(M[, c("carheight", "carwidth", "price")], use = "complete.obs")
```

```
##           carheight  carwidth     price
## carheight 1.0000000 0.2792103 0.1193362
## carwidth  0.2792103 1.0000000 0.7593253
## price     0.1193362 0.7593253 1.0000000
```

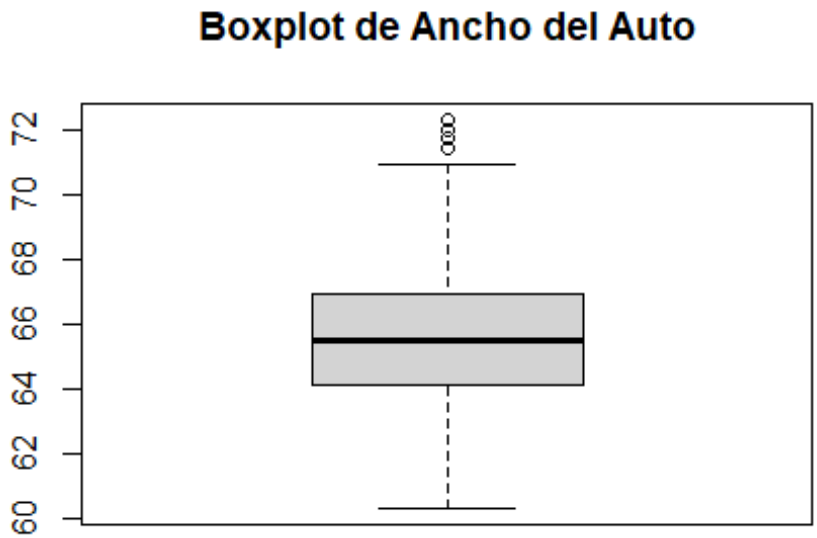
```
# Boxplot y Histogramas para 'carheight' y 'carwidth'
boxplot(M$carheight, main = "Boxplot de Altura del Auto")
```



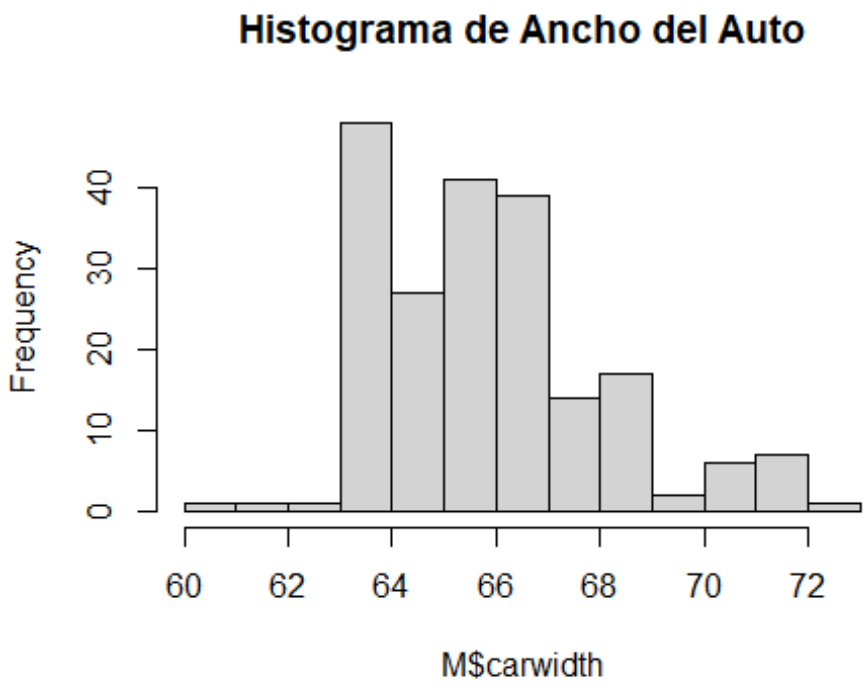

```
hist(M$carheight, main = "Histograma de Altura del Auto")
```



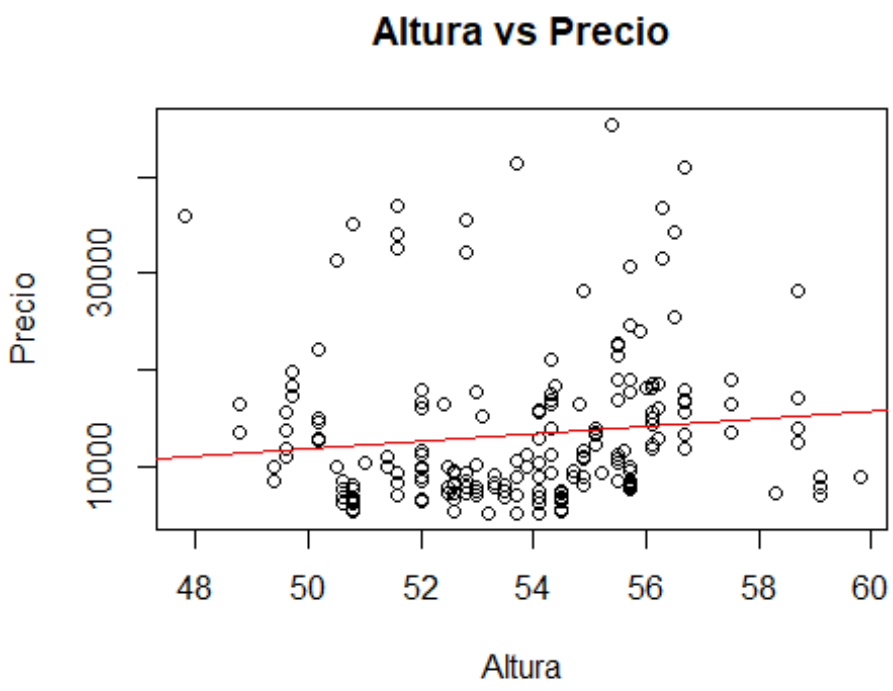

```
boxplot(M$carwidth, main = "Boxplot de Ancho del Auto")
```



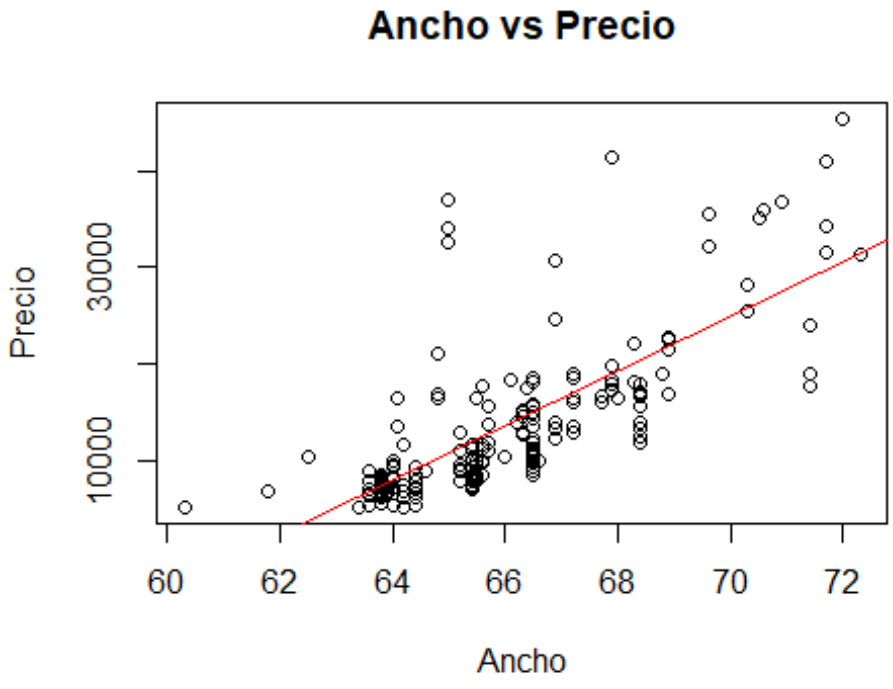

```
hist(M$carwidth, main = "Histograma de Ancho del Auto")
```




```r
# Scatterplots entre 'carheight', 'carwidth' y 'price' con recta de mejor ajuste
plot(M$carheight, M$price, main = "Altura vs Precio", xlab = "Altura", ylab = "Precio")
abline(lm(price ~ carheight, data = M), col = "red")
```

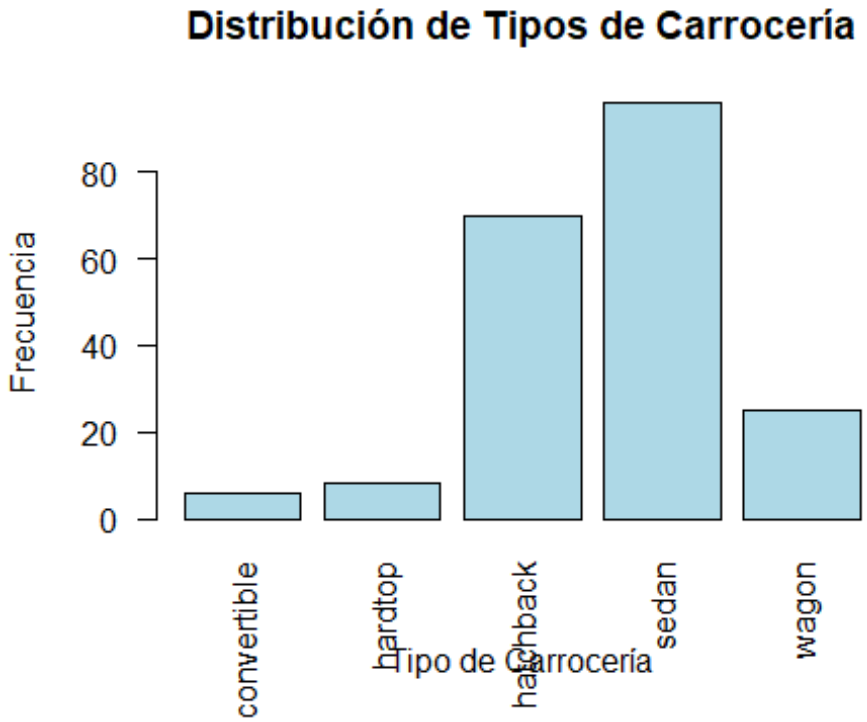



```
plot(M$carwidth, M$price, main = "Ancho vs Precio", xlab = "Ancho", ylab
= "Precio")
abline(lm(price ~ carwidth, data = M), col = "red")
```



```r
# Gráfico de barras para la variable 'carbody'

barplot(table(M$carbody),
        main = "Distribución de Tipos de Carrocería",
        xlab = "Tipo de Carrocería",
        ylab = "Frecuencia",
        col = "lightblue",
        las = 2)
```



La variable carwidth (ancho del coche) tiene una correlación positiva y significativa con el precio del vehículo. Esto se refleja en el modelo de regresión lineal, donde por cada aumento en el ancho del coche, el precio aumenta en aproximadamente 2932 unidades monetarias. Esta relación es muy significativa, con un valor p menor a 0.001.

Por otro lado, la variable carheight (altura del coche) tiene una relación negativa con el precio, aunque menos pronunciada. El coeficiente de carheight indica que por cada aumento en la altura, el precio disminuye en 328.6 unidades monetarias. Sin embargo, la significancia es menor (p = 0.034), lo que indica que la influencia de la altura es estadísticamente débil pero relevante.

##Modelación y verificación del modelo

```r
# Modelo de regresión lineal múltiple
modelo1 <- lm(price ~ carheight + carwidth, data = M)
summary(modelo1)
```

```
##
## Call:
## lm(formula = price ~ carheight + carwidth, data = M)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -11022  -2951  -1196   1156  25715
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -162328.7    12212.4 -13.292   <2e-16 ***
## carheight      -328.6      154.2  -2.132   0.0342 *
## carwidth       2932.3      175.6  16.699   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5166 on 202 degrees of freedom
## Multiple R-squared:  0.5859, Adjusted R-squared:  0.5818
## F-statistic: 142.9 on 2 and 202 DF,  p-value: < 2.2e-16
```

```
# Cálculo del VIF para verificar la multicolinealidad
vif(modelo1)
```

```
## carheight  carwidth
##   1.08455   1.08455
```

```
# Validación de la significancia del modelo 1
alpha <- 0.04
summary(modelo1)
```

```
##
## Call:
## lm(formula = price ~ carheight + carwidth, data = M)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -11022  -2951  -1196   1156  25715
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -162328.7    12212.4 -13.292   <2e-16 ***
## carheight      -328.6      154.2  -2.132   0.0342 *
## carwidth       2932.3      175.6  16.699   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5166 on 202 degrees of freedom
## Multiple R-squared:  0.5859, Adjusted R-squared:  0.5818
## F-statistic: 142.9 on 2 and 202 DF,  p-value: < 2.2e-16
```

```r
# Validación de la significancia de los coeficientes
coef_significance <- summary(modelo1)$coefficients[, 4] < alpha
print(coef_significance)
```

```
## (Intercept)   carheight    carwidth
##        TRUE        TRUE        TRUE
```

```r
# Porcentaje de variación explicada por el modelo
r_squared <- summary(modelo1)$r.squared
cat("Porcentaje de variación explicada por el modelo 1: ", r_squared *
100, "%\n")
```

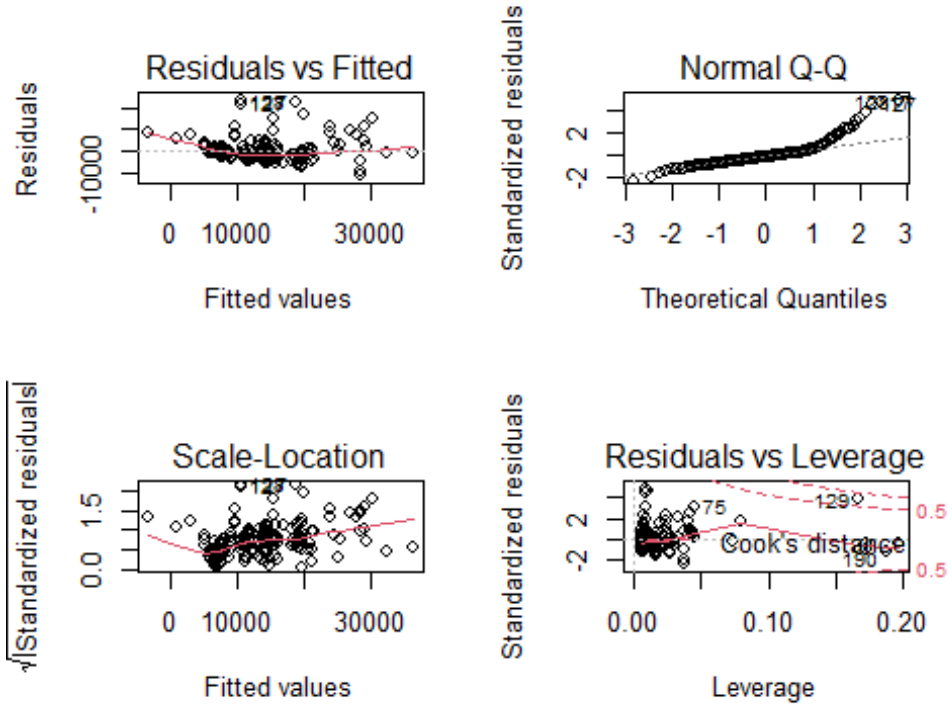
```
## Porcentaje de variación explicada por el modelo 1:  58.58898 %
```

```r
# Convertir 'carbody' en variable binaria (convertible o no)
M$convertible <- ifelse(M$carbody == "convertible", 1, 0)

# Modelo de regresión incluyendo variable binaria 'convertible'
modelo2 <- lm(price ~ carheight + carwidth + convertible, data = M)
summary(modelo2)
```

```
##
## Call:
## lm(formula = price ~ carheight + carwidth + convertible, data = M)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10880.4  -2612.0   -956.7   1065.8  23205.9
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -167451.0    11724.6 -14.282  < 2e-16 ***
## carheight      -219.5      149.3  -1.471    0.143
## carwidth       2916.9      167.8  17.381  < 2e-16 ***
## convertible    9330.3     2073.7   4.499 1.15e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4937 on 201 degrees of freedom
## Multiple R-squared:  0.6238, Adjusted R-squared:  0.6182
## F-statistic: 111.1 on 3 and 201 DF,  p-value: < 2.2e-16
```

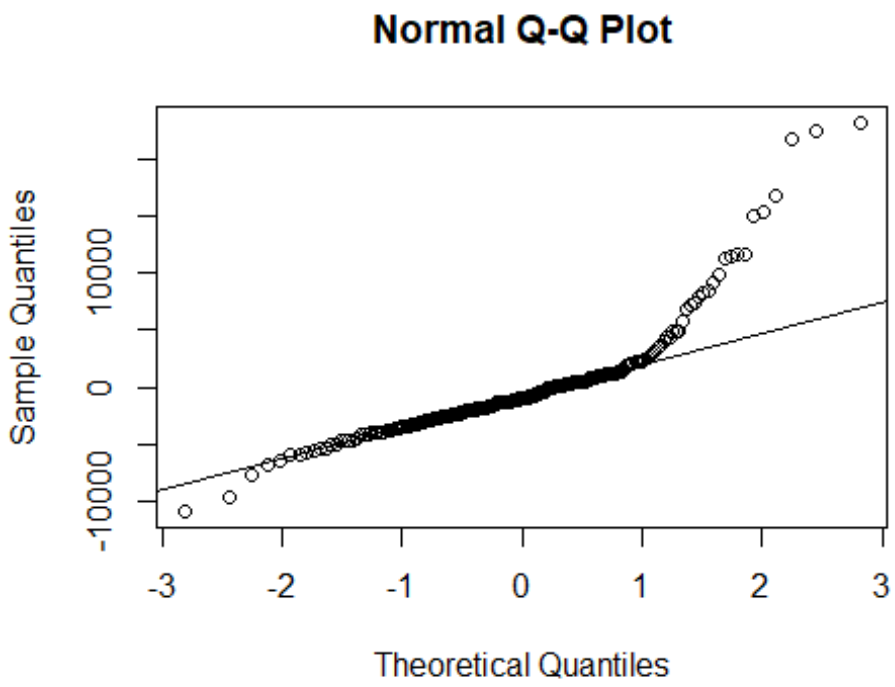
```r
# Graficar residuos del modelo
par(mfrow = c(2, 2))
plot(modelo2)
```




```
# Prueba de normalidad de los residuos
ad.test(resid(modelo2))  # Anderson-Darling test

##
##  Anderson-Darling normality test
##
## data:  resid(modelo2)
## A = 10.657, p-value < 2.2e-16

# QQ plot para los residuos del modelo 2
qqnorm(resid(modelo2))
qqline(resid(modelo2))
```




```r
# Verificación de media cero
mean(resid(modelo2))
```

```
## [1] 1.770313e-13
```

```r
# Verificación de homocedasticidad
bptest(modelo2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  modelo2
## BP = 6.9477, df = 3, p-value = 0.07359
```

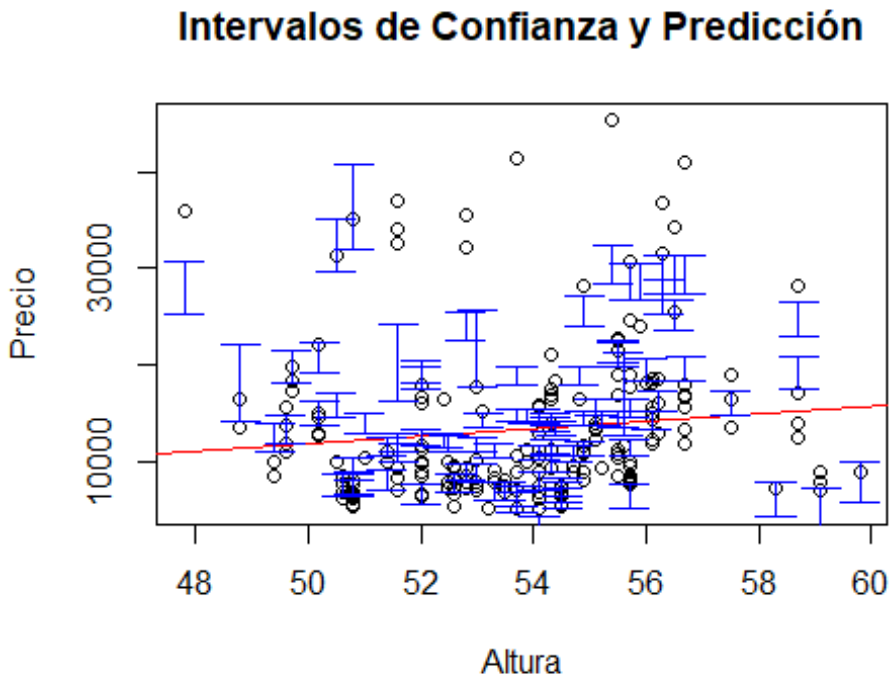
```r
# Intervalos de confianza para el modelo 2
confint(modelo2, level = 0.95)
```

```
##                     2.5 %        97.5 %
## (Intercept) -190569.9695 -144331.96129
## carheight      -513.8972      74.83847
## carwidth       2586.0144    3247.85566
## convertible    5241.2226   13419.32054
```

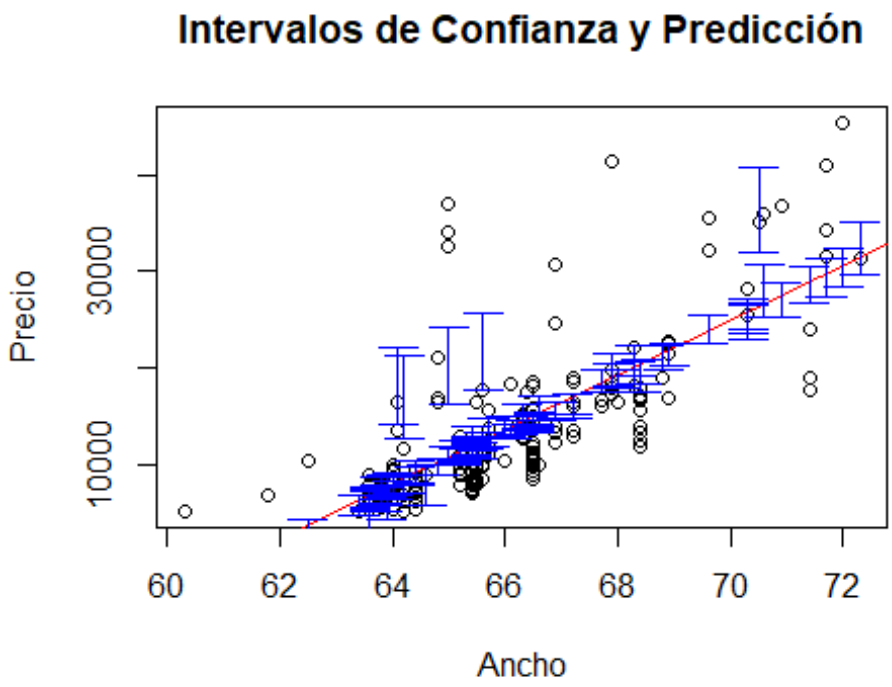
```r
# Intervalos de predicción para nuevas observaciones
new_data <- data.frame(carheight = 50, carwidth = 60, convertible = 1)
predict(modelo2, newdata = new_data, interval = "predict", level = 0.95)
```

```
##        fit       lwr      upr
## 1 5918.938 -4744.111 16581.99
```

```r
# Gráfico de los intervalos de predicción y confianza para el modelo 2
# Predicciones y límites de intervalo
predictions <- predict(modelo2, interval = "confidence", level = 0.95)
plot(M$carheight, M$price, main = "Intervalos de Confianza y Predicción",
xlab = "Altura", ylab = "Precio")
abline(lm(price ~ carheight, data = M), col = "red")
# Añadir intervalos de confianza y predicción
arrows(M$carheight, predictions[, "lwr"], M$carheight, predictions[,
"upr"], angle = 90, code = 3, length = 0.1, col = "blue")
```




```r
# Gráfico de los intervalos de predicción y confianza para el modelo 2
con 'carwidth'
# Predicciones y límites de intervalo
predictions <- predict(modelo2, interval = "confidence", level = 0.95)
plot(M$carwidth, M$price, main = "Intervalos de Confianza y Predicción",
xlab = "Ancho", ylab = "Precio")
abline(lm(price ~ carwidth, data = M), col = "red")
# Añadir intervalos de confianza y predicción
arrows(M$carwidth, predictions[, "lwr"], M$carwidth, predictions[,
"upr"], angle = 90, code = 3, length = 0.1, col = "blue")
```



El test de normalidad Anderson-Darling sobre los residuos muestra un valor de p muy bajo (< 2.2e-16), lo que indica que los residuos del modelo no siguen una distribución normal, lo que podría afectar la robustez de los resultados.

El test de heterocedasticidad de Breusch-Pagan arroja un valor de p = 0.07359, cercano al límite de significancia de 0.05, lo que sugiere que podría haber algún grado de heterocedasticidad (varianza no constante) en el modelo. Aunque no es un problema crítico, se debe tener en cuenta para la interpretación de los resultados.

##CONCLUSION FINAL:

Los resultados sugieren que para los consumidores, el ancho del vehículo y el hecho de que sea convertible son factores clave que influyen en el precio de los automóviles. Sin embargo, la altura del vehículo parece tener una influencia negativa menor, y podría no ser tan importante para los consumidores como el ancho.

El modelo proporciona una buena aproximación para entender los factores que afectan el precio de los automóviles, pero se recomienda explorar otros factores que podrían mejorar la capacidad predictiva del modelo, como el tipo de combustible, potencia del motor o el número de cilindros.

##Más allá - Existe una fuerte correlación positiva entre carwidth (ancho del auto) y el precio (0.76), lo que indica que a mayor ancho del vehículo, el precio tiende a aumentar.

- La correlación entre carheight y precio es más débil (0.12).

Propuesta de nuevas agrupaciones

- Segmentación según el tipo de carrocería:

El análisis muestra que el tipo de carrocería tiene una relación importante con el precio, particularmente los convertibles tienen un impacto significativo. Se podría considerar agrupar o segmentar los vehículos según su tipo de carrocería para análisis adicionales o ajustes de modelos específicos para cada tipo.

- Análisis por rango de ancho del auto:

El ancho (carwidth) tiene una relación tan fuerte con el precio, podría ser útil crear agrupaciones por rangos de carwidth. Esto permitiría identificar patrones de precio más específicos según la amplitud del vehículo.

- Dado que la altura del vehículo no resulta ser significativa en el modelo, se podría intentar agregar otras variables como el tipo de motor (enginetype), número de cilindros (cylindernumber), o eficiencia de combustible (citympg, highwaympg) para ver si mejoran la capacidad explicativa del modelo.