Actividad Integradora 2

Oscar Gutierrez

2024-09-06

Descripción

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 Con base en varias encuestas de mercado, la consultora ha recopilado un gran conjunto de datos de diferentes tipos de automóviles en el mercado estadounidense que presenta en el siguiente archivo Download archivo. Las variables recopiladas vienen descritas en el diccionario de términos Download diccionario de términos. Por un análisis de correlación, la empresa automovilistica tiene interés en analizar las variables agrupadas de la siguiente forma para hacer el análisis de variables significativas:

Primer grupo. Distancia entre los ejes (wheelbase), tipo de gasolina que usa y caballos de fuerza **Segundo grupo. Altura del auto, ancho del auto y si es convertible o no.** Tercer grupo. Tamaño del motor (ensinesize), carrera o lanzamiento del pistón (stroke) y localización del motor en el carro Selecciona uno de los tres grupos analizados (te será asignado por tu profesora) y analiza la significancia de las variables para predecir o influir en la variable precio. ¿propondrías una nueva agrupación a la empresa automovilísitica?

Exploración de la base de datos

Importar dataset y seleccionar las variables asignadas.

```
M1 = read.csv('precios_autos.csv')
M <- M1[, c("carheight", "carwidth", "carbody", 'price')]
# cambiar la variable carbody a convertible
M$convertible <- ifelse(M$carbody == "convertible", 1, 0)
M$carbody <- NULL

head(M)
```

```
##   carheight carwidth price convertible
## 1      48.8     64.1 13495           1
## 2      48.8     64.1 16500           1
## 3      52.4     65.5 16500           0
## 4      54.3     66.2 13950           0
## 5      54.3     66.4 17450           0
## 6      53.1     66.3 15250           0
```

Se seleccionaron las variables de interés y se cambió la variable carbody a convertible para saber si un auto es convertible o no

```
summary(M)
```

```
##    carheight        carwidth         price        convertible
##  Min.   :47.80   Min.   :60.30   Min.   : 5118   Min.   :0.00000
##  1st Qu.:52.00   1st Qu.:64.10   1st Qu.: 7788   1st Qu.:0.00000
##  Median :54.10   Median :65.50   Median :10295   Median :0.00000
##  Mean   :53.72   Mean   :65.91   Mean   :13277   Mean   :0.02927
##  3rd Qu.:55.50   3rd Qu.:66.90   3rd Qu.:16503   3rd Qu.:0.00000
##  Max.   :59.80   Max.   :72.30   Max.   :45400   Max.   :1.00000
```

```
sd(M$carwidth)
```

```
## [1] 2.145204
```

```
sd(M$carwidth)
```

```
## [1] 2.145204
```

```
sd(M$price)
```

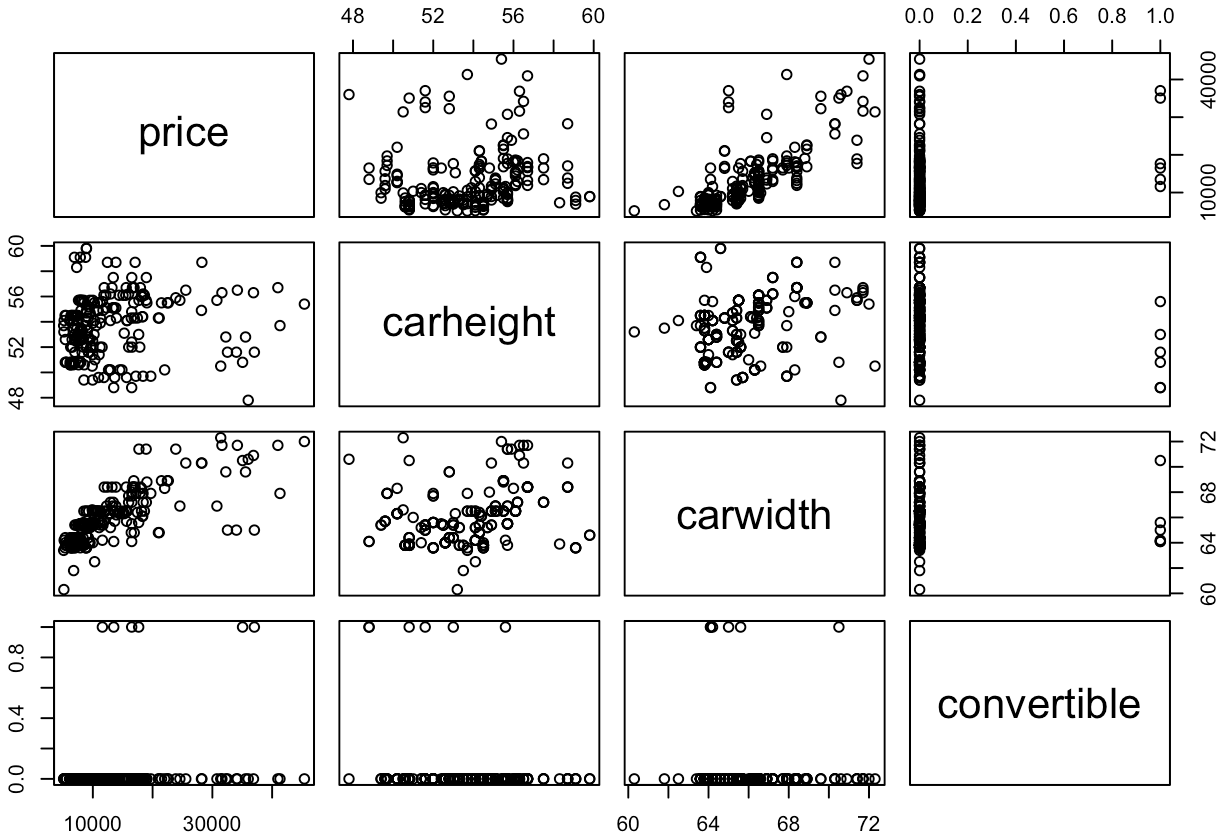
```
## [1] 7988.852
```

```
table(M$convertible)
```

```
##
##   0   1
## 199   6
```

Solo hay 6 carros convertibles

```
pairs(~ price + carheight + carwidth + convertible, data = M)
```

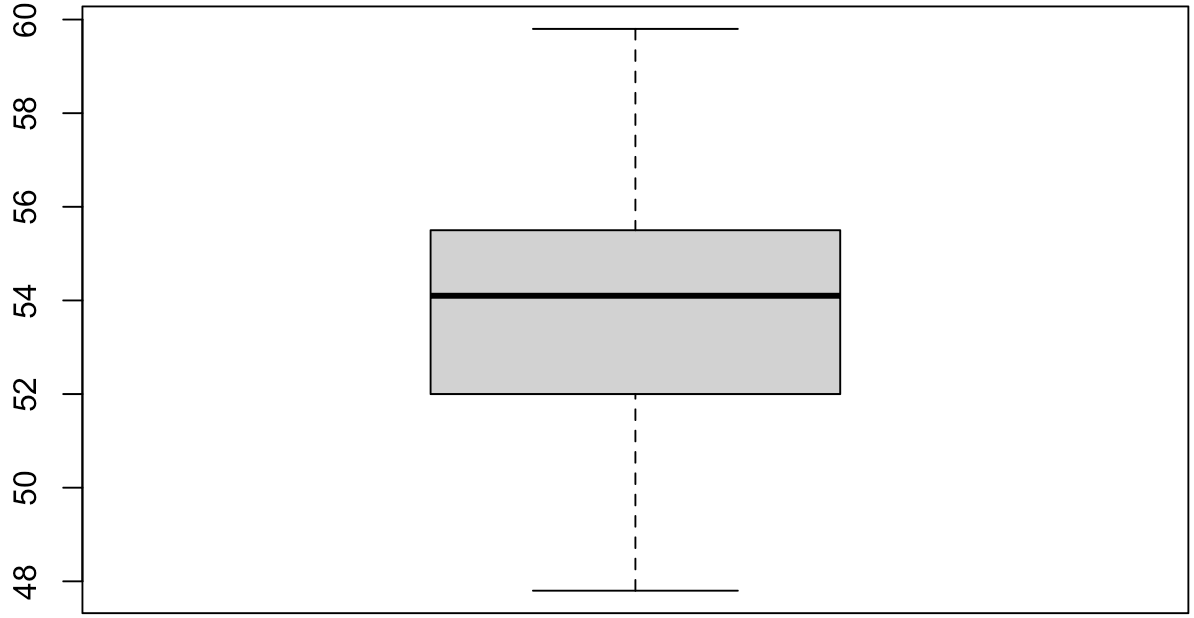



```
cor(M)
```

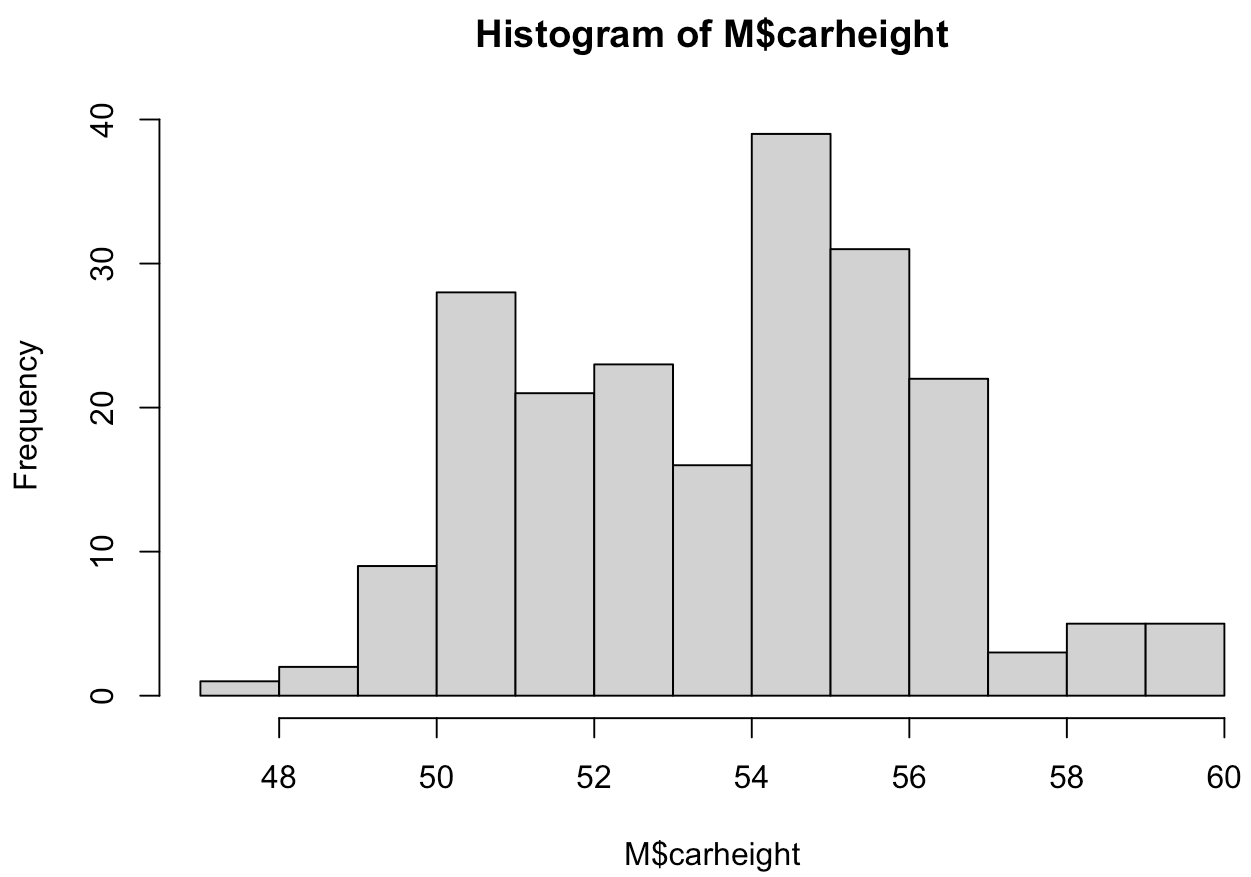
```
##              carheight    carwidth     price convertible
## carheight    1.0000000  0.27921032 0.1193362 -0.16323866
## carwidth     0.2792103  1.00000000 0.7593253 -0.02632807
## price        0.1193362  0.75932530 1.0000000  0.18768121
## convertible -0.1632387 -0.02632807 0.1876812  1.00000000
```

Hay poca correlación entre variables, la mayor es la el precio con el ancho, teniendo 0.76 de correlación.

```
boxplot(M$carheight)
```

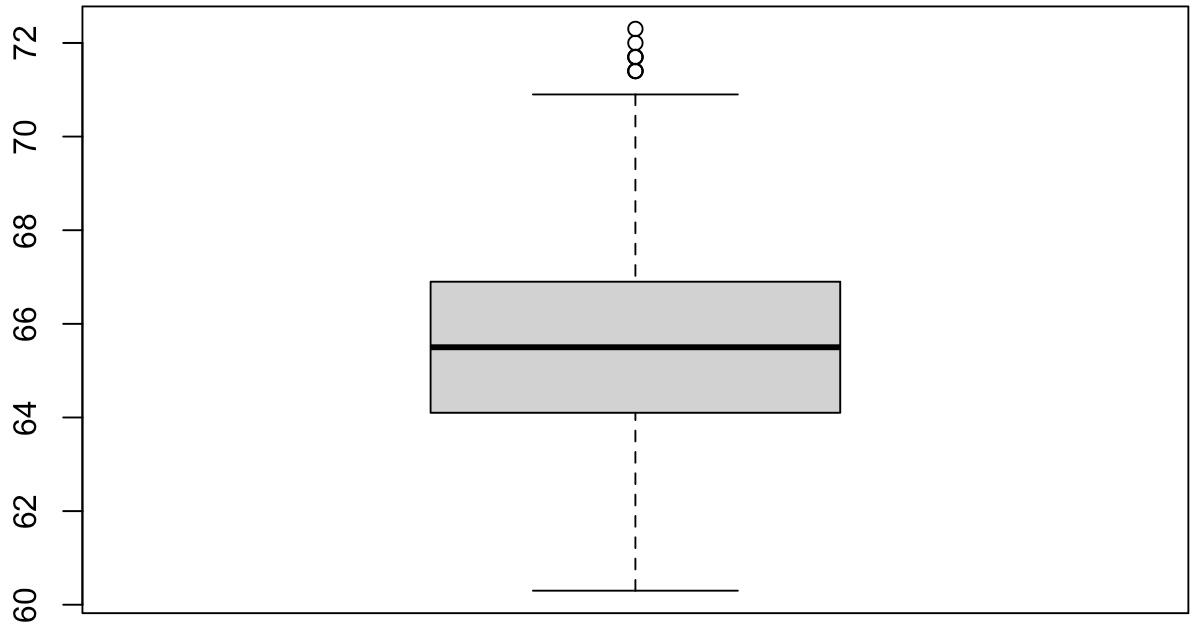



```
hist(M$carheight)
```

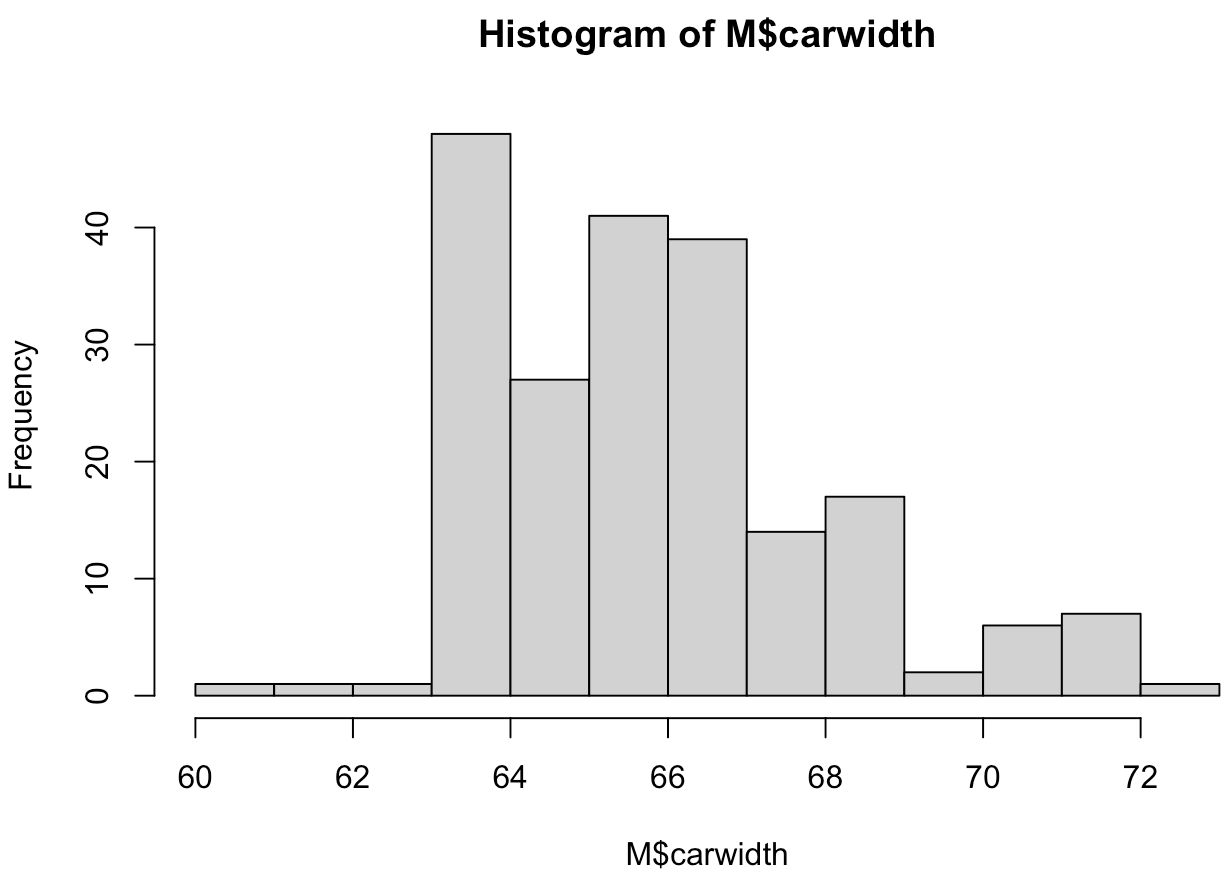


No hay valores atípicos, no se cuenta con una distribución en particular.

```
boxplot(M$carwidth)
```




```
hist(M$carwidth)
```



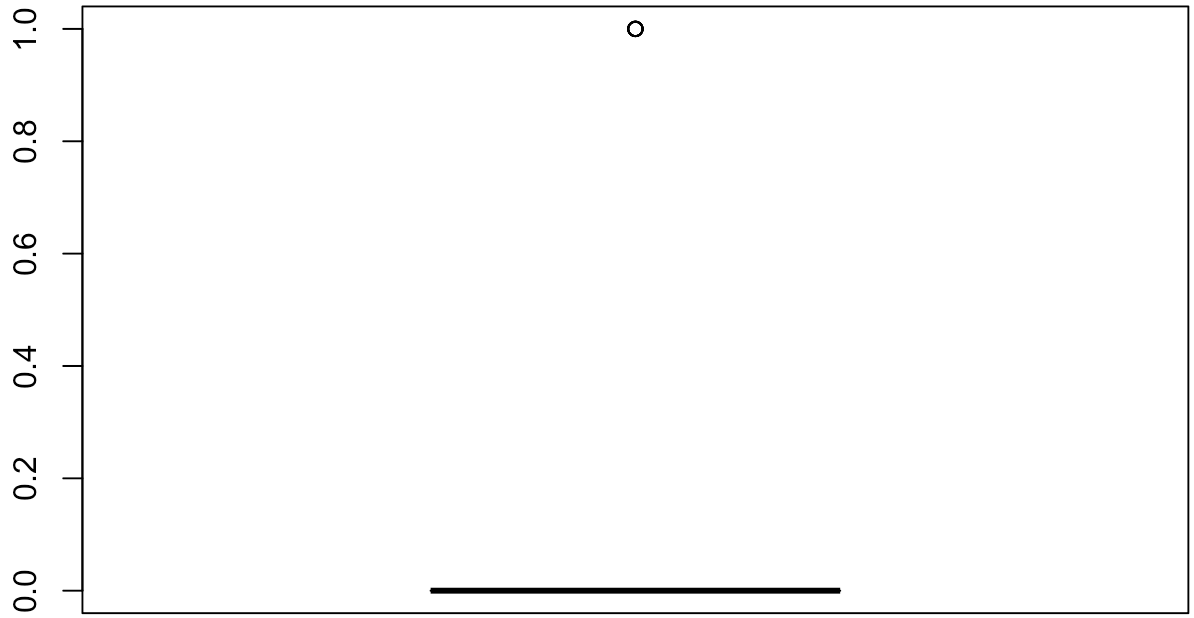
Hay unos cuantos valores atípicos

```
Q1 <- quantile(M$carwidth, 0.25)
Q3 <- quantile(M$carwidth, 0.75)
IQR <- Q3 - Q1

lower_bound <- Q1 - 1.5 * IQR
upper_bound <- Q3 + 1.5 * IQR

outliers <- M1[M$carwidth < lower_bound | M$carwidth > upper_bound, ]

print(outliers)
```

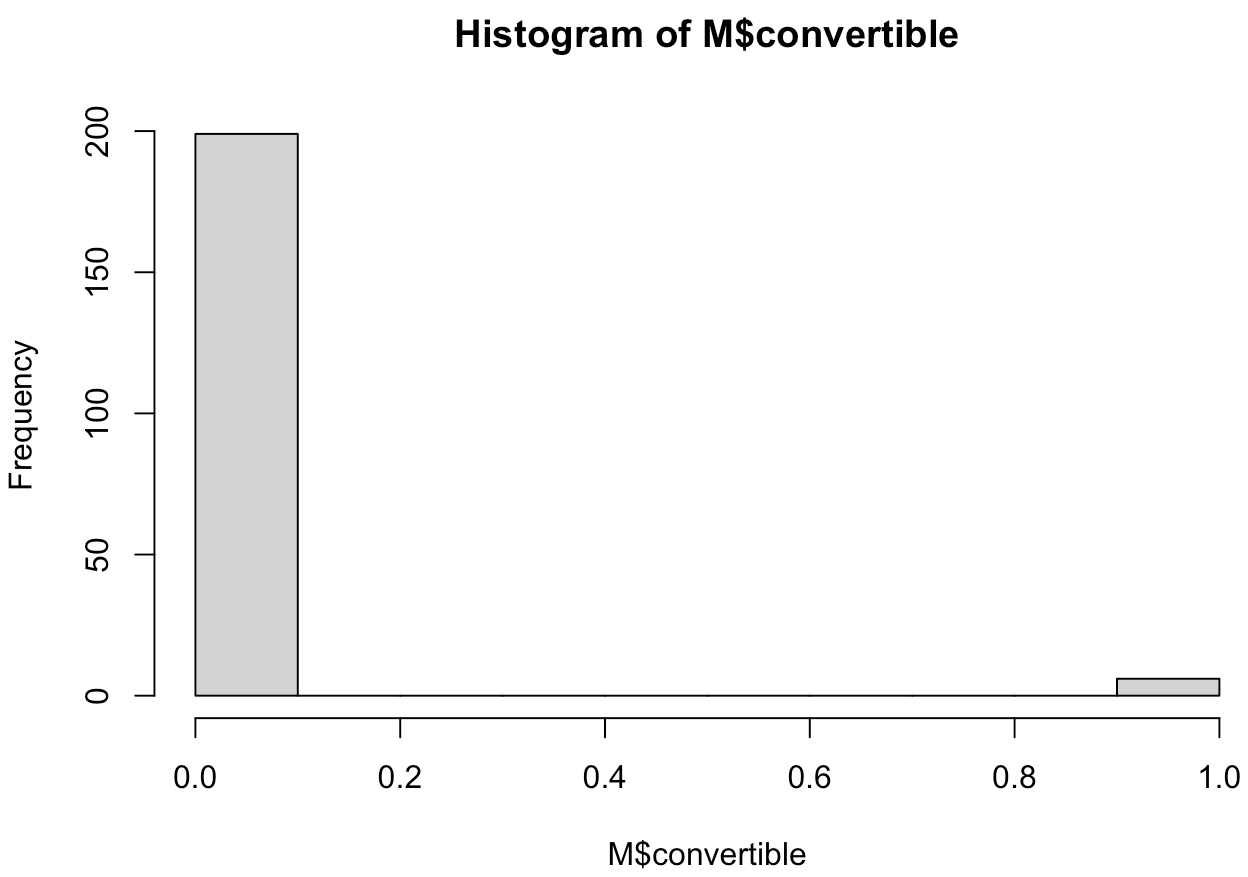
```
##     symboling                        CarName fueltype   carbody drivewheel
## 7           1                     audi 100ls      gas     sedan        fwd
## 8           1                      audi 5000      gas     wagon        fwd
## 9           1                      audi 4000      gas     sedan        fwd
## 71         -1                   buick skyhawk   diesel     sedan        rwd
## 72         -1         buick opel isuzu deluxe      gas     sedan        rwd
## 74          0           buick century special      gas     sedan        rwd
## 75          1 buick regal sport coupe (turbo)      gas   hardtop        rwd
## 130         1                 porsche cayenne      gas hatchback        rwd
##     enginelocation wheelbase carlength carwidth carheight curbweight enginetype
## 7            front     105.8     192.7     71.4      55.7       2844        ohc
## 8            front     105.8     192.7     71.4      55.7       2954        ohc
## 9            front     105.8     192.7     71.4      55.9       3086        ohc
## 71           front     115.6     202.6     71.7      56.3       3770        ohc
## 72           front     115.6     202.6     71.7      56.5       3740       ohcv
## 74           front     120.9     208.1     71.7      56.7       3900       ohcv
## 75           front     112.0     199.2     72.0      55.4       3715       ohcv
## 130          front      98.4     175.7     72.3      50.5       3366      dohcv
##     cylindernumber enginesize stroke compressionratio horsepower peakrpm
## 7             five        136   3.40              8.5        110    5500
## 8             five        136   3.40              8.5        110    5500
## 9             five        131   3.40              8.3        140    5500
## 71            five        183   3.64             21.5        123    4350
## 72           eight        234   3.10              8.3        155    4750
## 74           eight        308   3.35              8.0        184    4500
## 75           eight        304   3.35              8.0        184    4500
## 130          eight        203   3.11             10.0        288    5750
##     citympg highwaympg   price
## 7        19         25 17710.0
## 8        19         25 18920.0
## 9        17         20 23875.0
## 71       22         25 31600.0
## 72       16         18 34184.0
## 74       14         16 40960.0
## 75       14         16 45400.0
## 130      17         28 31400.5
```

Los valores atípicos de carwidth corresponden a camionetas o coches grandes, lo cual tiene sentido.

```
boxplot(M$convertible)
```

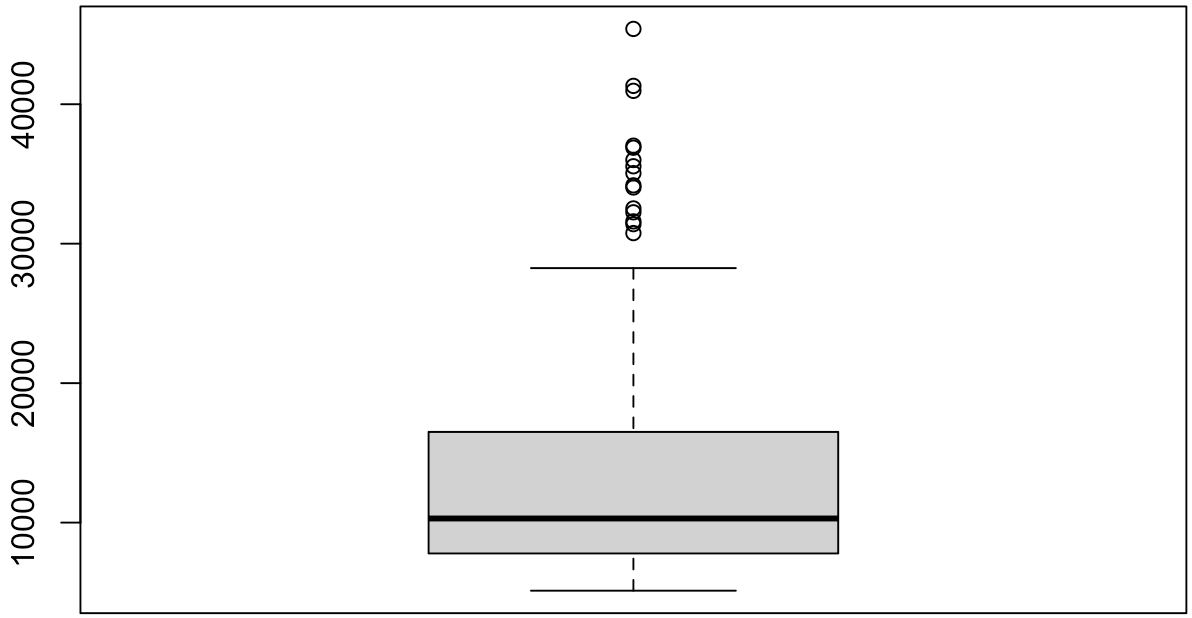



```
hist(M$convertible)
```

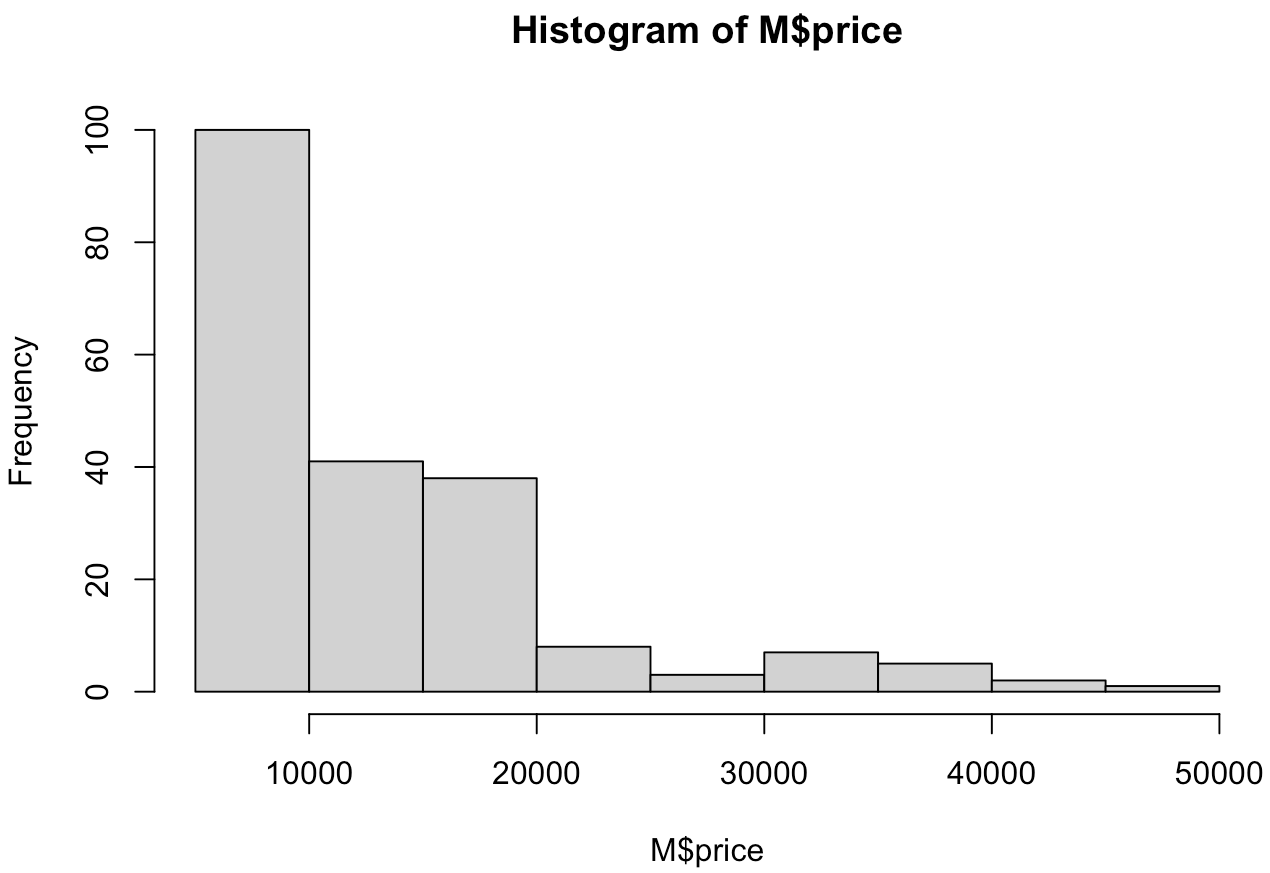


La proporción de carros que no es convertible es mucho mayor a los que si son convertibles.

```
boxplot(M$price)
```




```
hist(M$price)
```



Tiene sentido que haya valores atípicos en precio debido a que hay coches que son de marcas de lujo como Porsche.

Primer modelo

Hipótesis de variables $H_0 : \beta_i = 0 \; H_1 : \beta_i \neq 0$

Hipótesis de modelo $H_0 : \beta_1 = \beta_2 = \cdots = \beta_n = 0 \; H_1 : \text{Al menos un } \beta_i \neq 0$

Con $\alpha = 0.04$

En este modelo se consideran todas las variables

```
Modelo1 = lm(price~convertible+carwidth+carheight, M)
Modelo1
```

```
##
## Call:
## lm(formula = price ~ convertible + carwidth + carheight, data = M)
##
## Coefficients:
## (Intercept)  convertible     carwidth    carheight
##   -167451.0       9330.3       2916.9       -219.5
```

```
model_summary <- summary(Modelo1)
model_summary
```

```
##
## Call:
## lm(formula = price ~ convertible + carwidth + carheight, 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 ***
## convertible    9330.3     2073.7   4.499 1.15e-05 ***
## carwidth       2916.9      167.8  17.381  < 2e-16 ***
## carheight      -219.5      149.3  -1.471    0.143
## ---
## 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
```

```
df <- model_summary$df[2]
alpha <- 0.04

critical_value <- qt(1 - alpha / 2, df)

cat('Valor frontera variables =',critical_value)
```

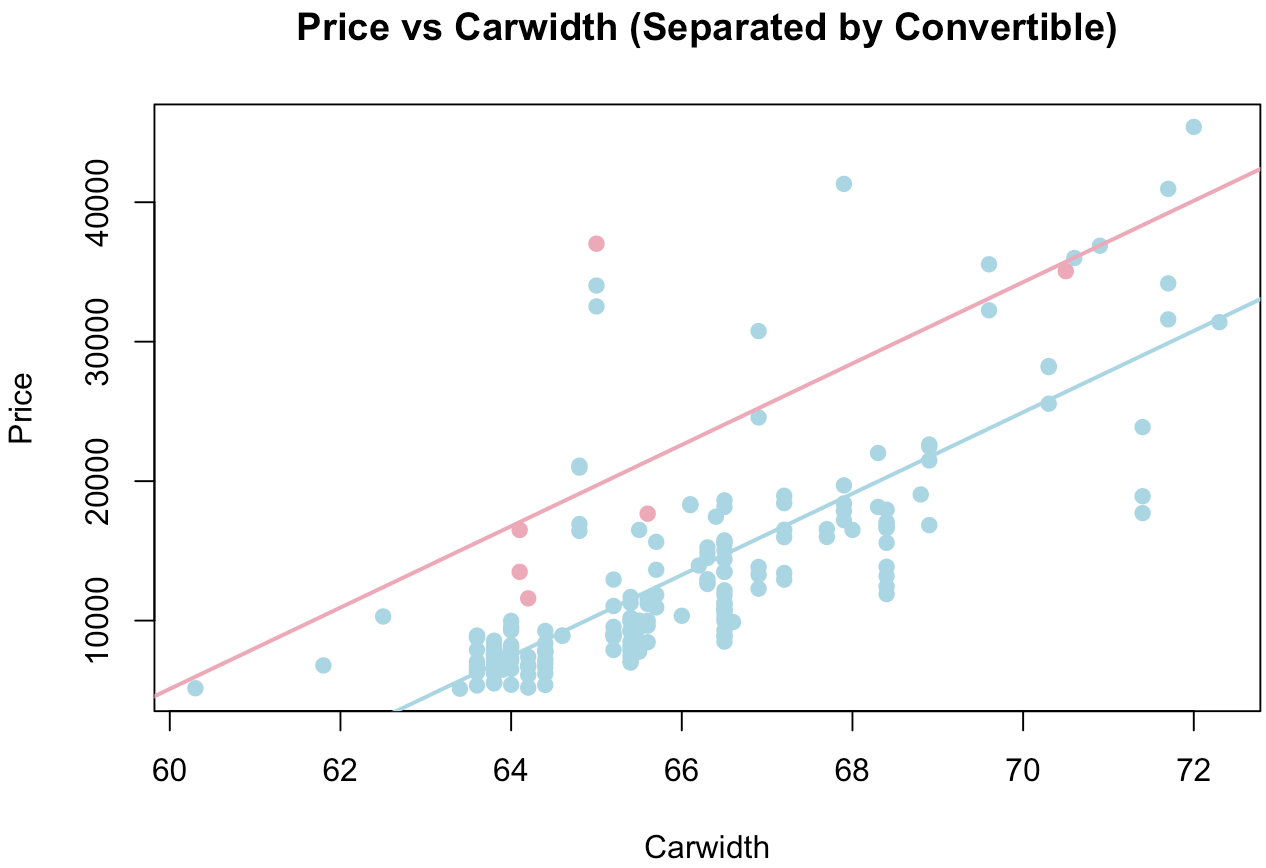
```
## Valor frontera variables = 2.067162
```

```
cat('\nValor frontera modelo =',qf(0.96,3,201))
```

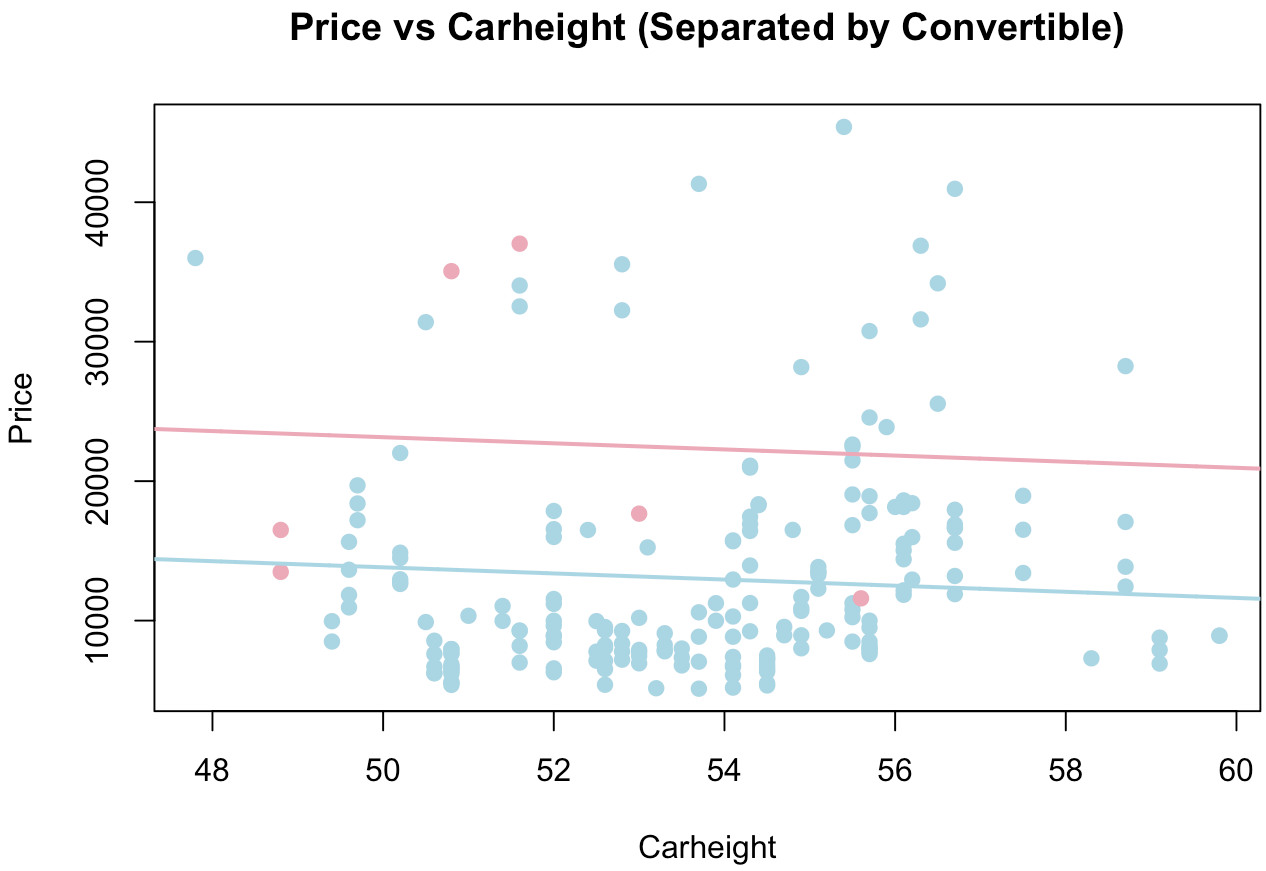
```
##
## Valor frontera modelo = 2.82134
```

Con el summary se puede observar que la altura del coche no es significativa para predecir el precio ya que tiene un valor p mayor a alpha. El modelo logra explicar un 61.82% de la variación en la variable de interés. El modelo es significativo ya que el F-statistic es mucho mayor al valor frontera.

```
b0 = Modelo1$coefficients[1]
b1 = Modelo1$coefficients[2]
b2 = Modelo1$coefficients[3]
b3 = Modelo1$coefficients[4]

Y_convertible = function(x, y){b0 + b1*1 + b2*x + b3*y}
Y_non_convertible = function(x, y){b0 + b1*0 + b2*x + b3*y}

colores = c('lightblue', 'pink2')

plot(M$carwidth, M$price, main="Price vs Carwidth (Separated by Convertible)",
     xlab="Carwidth", ylab="Price", pch=19, col=colores[factor(M$convertible)])

x = seq(min(M$carwidth)*0.9, max(M$carwidth)*1.1, 0.01)

mean_carheight <- mean(M$carheight)

lines(x, Y_convertible(x, mean_carheight), col = "pink2", lwd = 2)  # For convertible =
1
lines(x, Y_non_convertible(x, mean_carheight), col = "lightblue", lwd = 2)  # For conver
tible = 0
```



```
plot(M$carheight, M$price, main="Price vs Carheight (Separated by Convertible)",
     xlab="Carheight", ylab="Price", pch=19, col=colores[factor(M$convertible)])


y = seq(min(M$carheight)*0.9, max(M$carheight)*1.1, 0.01)


mean_carwidth <- mean(M$carwidth)

lines(y, Y_convertible(mean_carwidth, y), col = "pink2", lwd = 2)  # For convertible = 1
lines(y, Y_non_convertible(mean_carwidth, y), col = "lightblue", lwd = 2)  # For convert
ible = 0
```



En los plots, una de las variables se define como constante, de lo contrario tendriamos gráficas en 3 dimensiones, sin embargo, aún asi se logra ver el comportamiento del modelo contra los datos. Se puede observar que hay mucha variabilidad en la variable altura, por esta razón el modelo marcó esta variable como no significativa. Por otro lado, se observa un mejor comportamiento cuando se grafica el precio contra el ancho, donde la recta de mejor ajuste sí va de acuerdo a los datos.

Modelo 2

Hipótesis de variables $H_0 : \beta_i = 0 \; H_1 : \beta_i \neq 0$

Hipótesis de modelo $H_0 : \beta_1 = \beta_2 = \cdots = \beta_n = 0 \; H_1 : \text{Al menos un } \beta_i \neq 0$

Con $\alpha = 0.04$

En este modelo se descarta el carheight ya que no era significativo y connvertible ya que la proporción es muy diferente.

```
Modelo2 = lm(price~carwidth+carheight, M)
Modelo2
```

```
##
## Call:
## lm(formula = price ~ carwidth + carheight, data = M)
##
## Coefficients:
## (Intercept)     carwidth    carheight
##   -162328.7       2932.3       -328.6
```

```
model_summary <- summary(Modelo2)
model_summary
```

```
##
## Call:
## lm(formula = price ~ carwidth + carheight, 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 ***
## carwidth       2932.3      175.6  16.699   <2e-16 ***
## carheight      -328.6      154.2  -2.132   0.0342 *
## ---
## 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
```

```
df <- model_summary$df[2]
alpha <- 0.04

critical_value <- qt(1 - alpha / 2, df)

cat('Valor frontera variables =',critical_value)
```

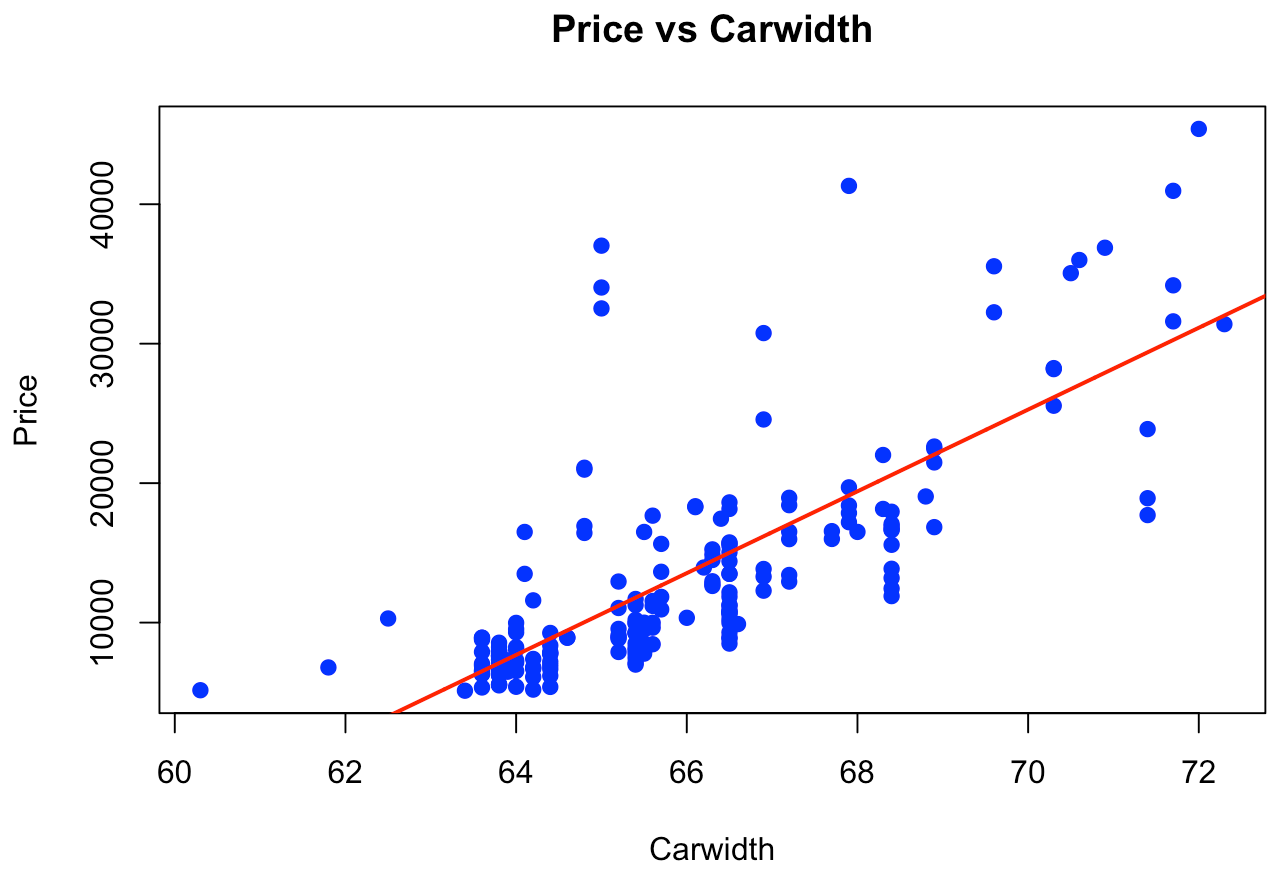
```
## Valor frontera variables = 2.067096
```

```
cat('\nValor frontera modelo =',qf(0.96,2,202))
```

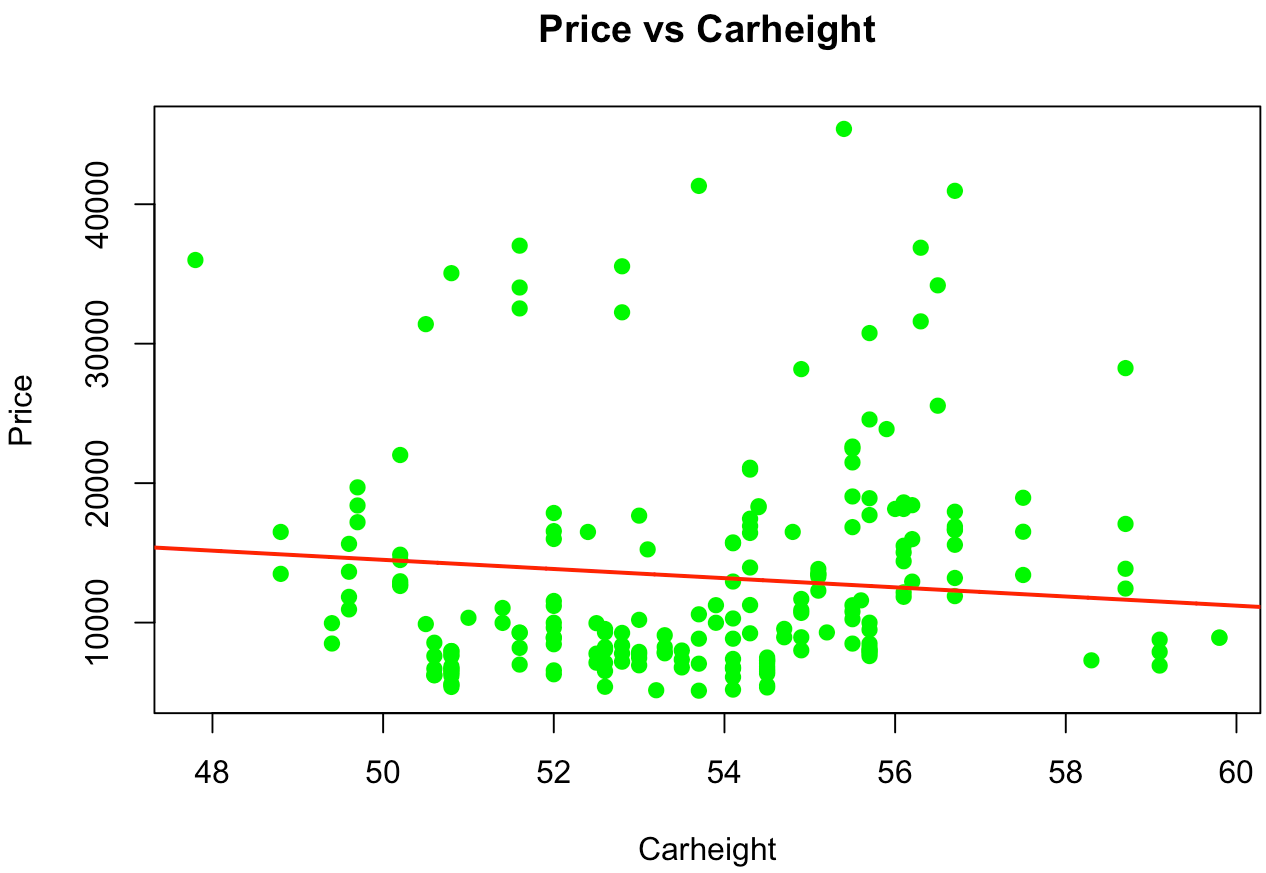
```
##
## Valor frontera modelo = 3.270718
```

El modelo es significativo debido a que el estadístico f es mayor al valor frontera, además, todas las variables son significativas, el modelo logra explicar el 58% de la variación en la variable de interés.

```
# Extract coefficients from Modelo2
b0 = Modelo2$coefficients[1]  # Intercept
b1 = Modelo2$coefficients[2]  # Coefficient for carwidth
b2 = Modelo2$coefficients[3]  # Coefficient for carheight

Y = function(x, y){b0 + b1*x + b2*y}

# Create the scatter plot for carwidth vs price
plot(M$carwidth, M$price, main="Price vs Carwidth",
     xlab="Carwidth", ylab="Price", pch=19, col="blue")

mean_carheight <- mean(M$carheight)

x = seq(min(M$carwidth)*0.9, max(M$carwidth)*1.1, 0.01)

lines(x, Y(x, mean_carheight), col = "red", lwd = 2)
```



```
plot(M$carheight, M$price, main="Price vs Carheight",
     xlab="Carheight", ylab="Price", pch=19, col="green")

mean_carwidth <- mean(M$carwidth)

y = seq(min(M$carheight)*0.9, max(M$carheight)*1.1, 0.01)

lines(y, Y(mean_carwidth, y), col = "red", lwd = 2)
```



Hay mucha similitud entre este modelo y el otro, existe mucha variabilidad entre los datos.

Normalidad de residuos

H_0 : Los residuos se distribuyen normalmente H_1 : Los residuos no se distribuyen normalmente

```
library(nortest)
ad.test(Modelo1$residuals)
```

```
##
##  Anderson-Darling normality test
##
## data:  Modelo1$residuals
## A = 10.657, p-value < 2.2e-16
```

```
ad.test(Modelo2$residuals)
```

```
##
##  Anderson-Darling normality test
##
## data:  Modelo2$residuals
## A = 10.319, p-value < 2.2e-16
```

Los residuos no se distribuyen normalmente en ninguno de los dos modelos, se tiene evidencia para rechazar la hipótesis nula.

Comprobar media = 0

$H_0 : \mu = 0$ $H_1 : mu \neq 0$

```
t.test(Modelo1$residuals)
```

```
##
##  One Sample t-test
##
## data:  Modelo1$residuals
## t = 6.0928e-16, df = 204, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -674.7772  674.7772
## sample estimates:
##    mean of x
## 2.085183e-13
```

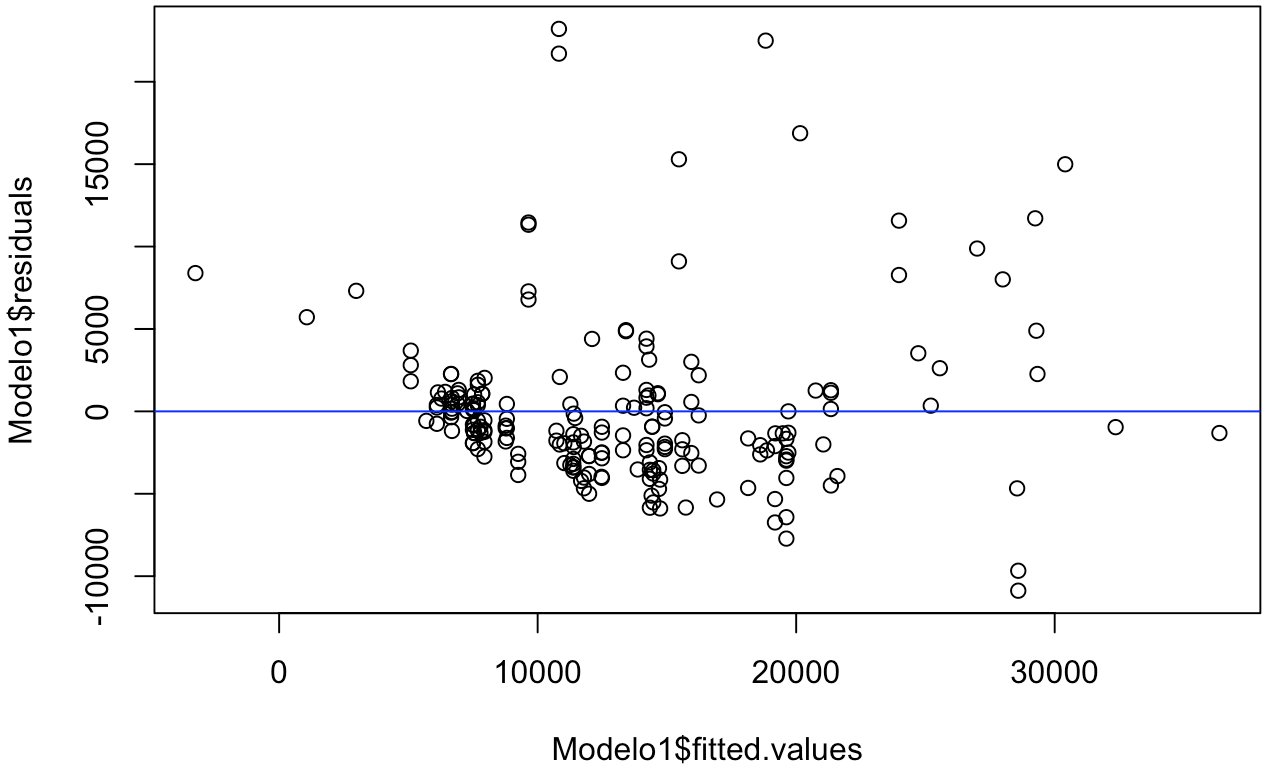
```
t.test(Modelo2$residuals)
```

```
##
##  One Sample t-test
##
## data:  Modelo2$residuals
## t = -1.4796e-15, df = 204, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -707.942  707.942
## sample estimates:
##    mean of x
## -5.31278e-13
```

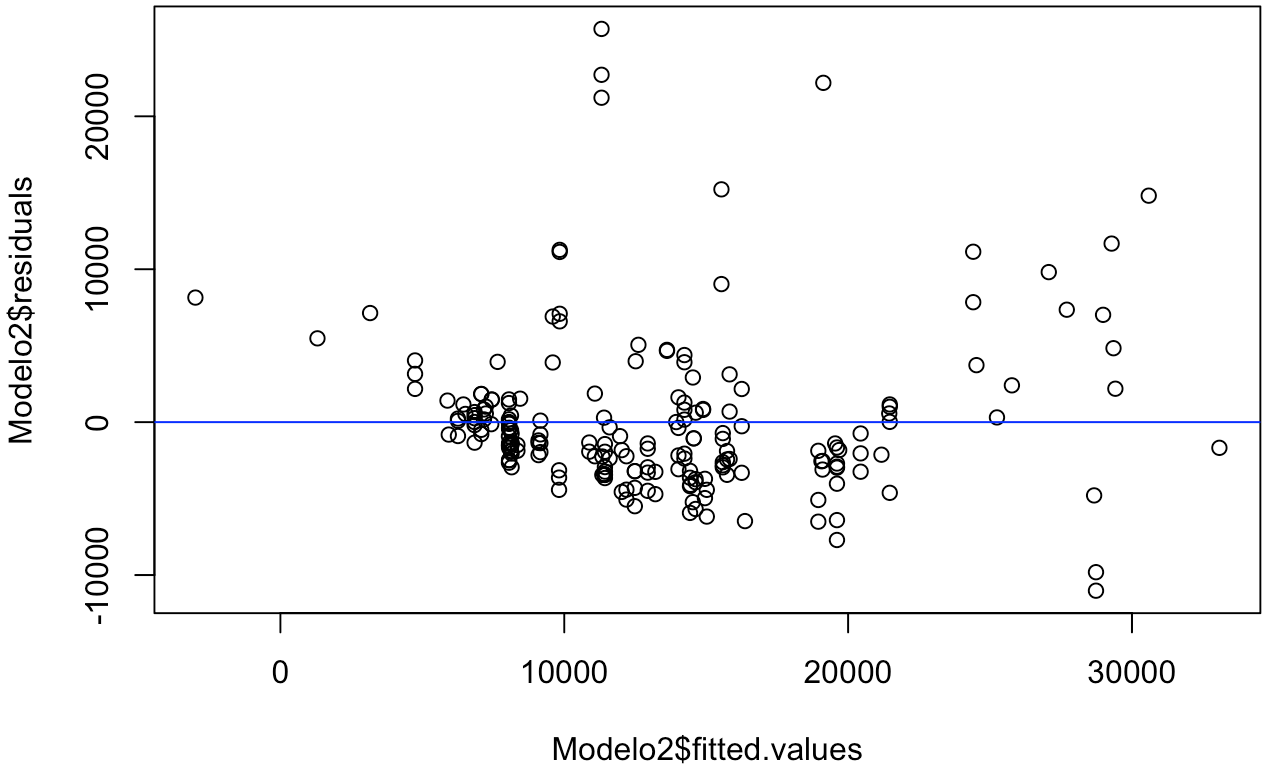
No se tiene evidencia para rechazar la hipótesis nula, los residuos tienen media 0.

Homocedasticidad e independencia

```
# Modelo 1
plot(Modelo1$fitted.values, Modelo1$residuals)
abline(h=0, col= 'blue')
```




```
# Modelo 2
plot(Modelo2$fitted.values, Modelo2$residuals)
abline(h=0, col= 'blue')
```



No hay homocedasticidad, parece haber un patrón en la región inferior, no hay varianza constante.

Prueba de independencia

H_0: Los errores no están autocorrelacionados. H_1: Los errores están autocorrelacionados.

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
dwtest(Modelo1)
```

```
##
##  Durbin-Watson test
##
## data:  Modelo1
## DW = 0.6719, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

```
bgtest(Modelo1)
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  Modelo1
## LM test = 92.308, df = 1, p-value < 2.2e-16
```

```
dwtest(Modelo2)
```

```
##
##  Durbin-Watson test
##
## data:  Modelo2
## DW = 0.67299, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

```
bgtest(Modelo2)
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  Modelo2
## LM test = 92.131, df = 1, p-value < 2.2e-16
```

Los errores sí están autocorrelacionados ya que el valor p es menor a alpha, por lo que se rechaza la hipótesis nula.

Conclusión

El mejor modelo entre estos dos sería el primero, logra explicar un poco más de la variabilidad y tiene un poco menos de error, sin embargo, es necesario mencionar que ninguno de los dos modelos cumple con las suposiciones de la regresión lineal por lo que es posible que las predicciones no sean precisas. Es posible que haya algún tipo de relación no lineal con la variable dependiente, por lo que un modelo linal no es ideal para intentar realizar predicciones.

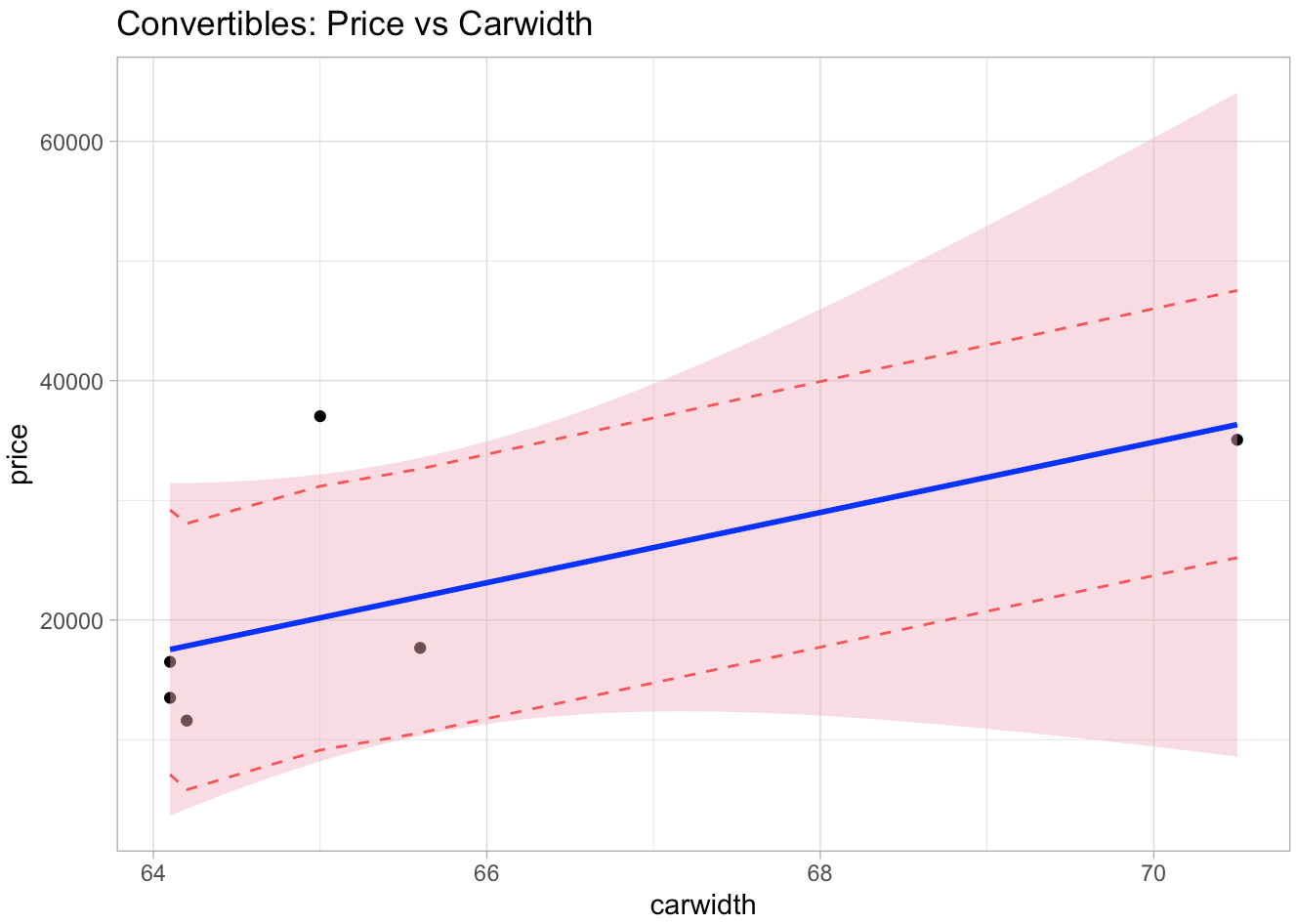
De acuerdo con el primer modelo, las variables significativas son la de convertible y ancho, la variable convertible ajusta el intercept dependiendo si tiene valor de 1 o 0 y el ancho afecta la pendiente de la recta.

Intervalos de confianza y predicción

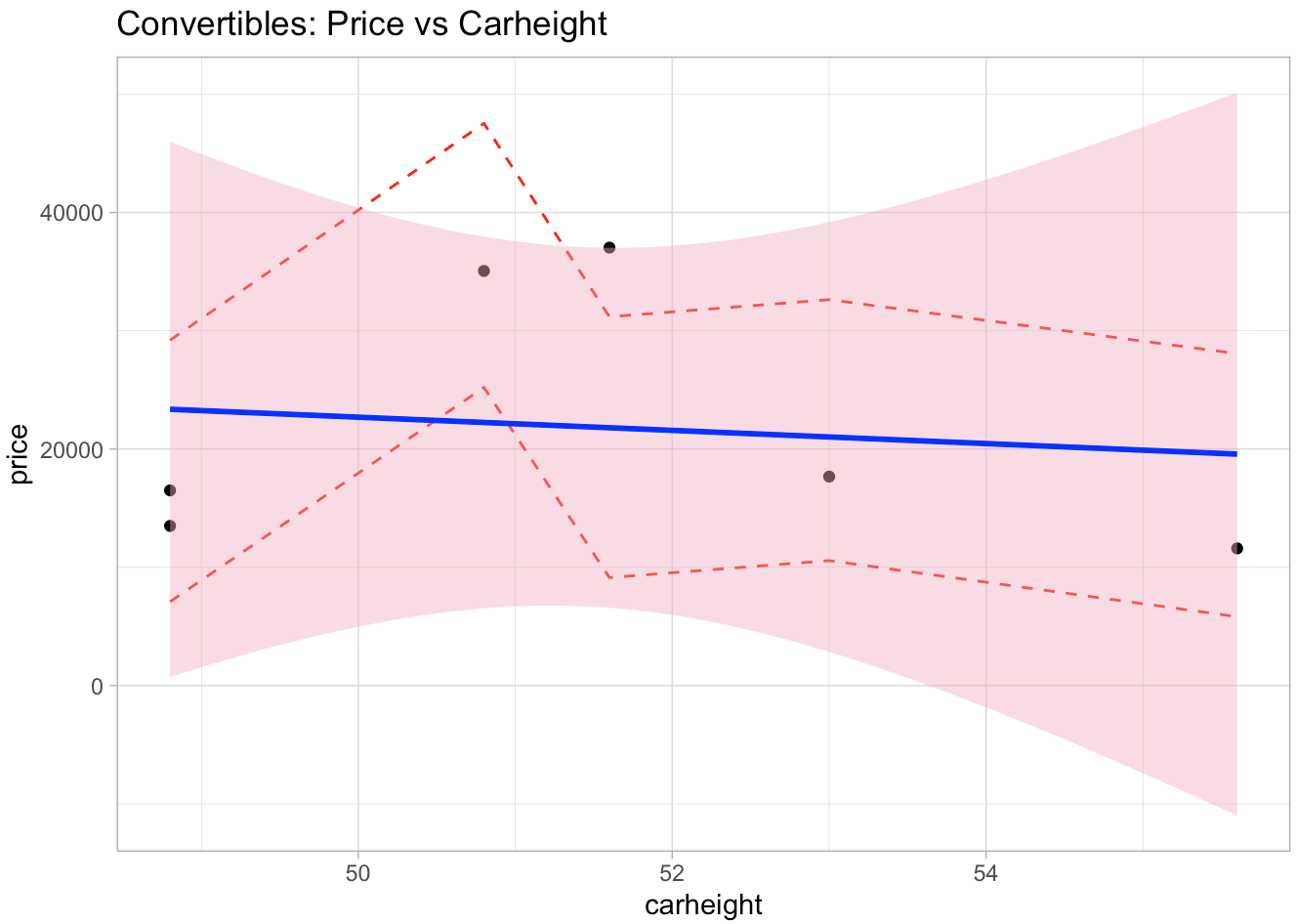
```
Ip <- predict(object = Modelo1, interval = "prediction", level = 0.96)
```

```
## Warning in predict.lm(object = Modelo1, interval = "prediction", level = 0.96): predi
ctions on current data refer to _future_ responses
```

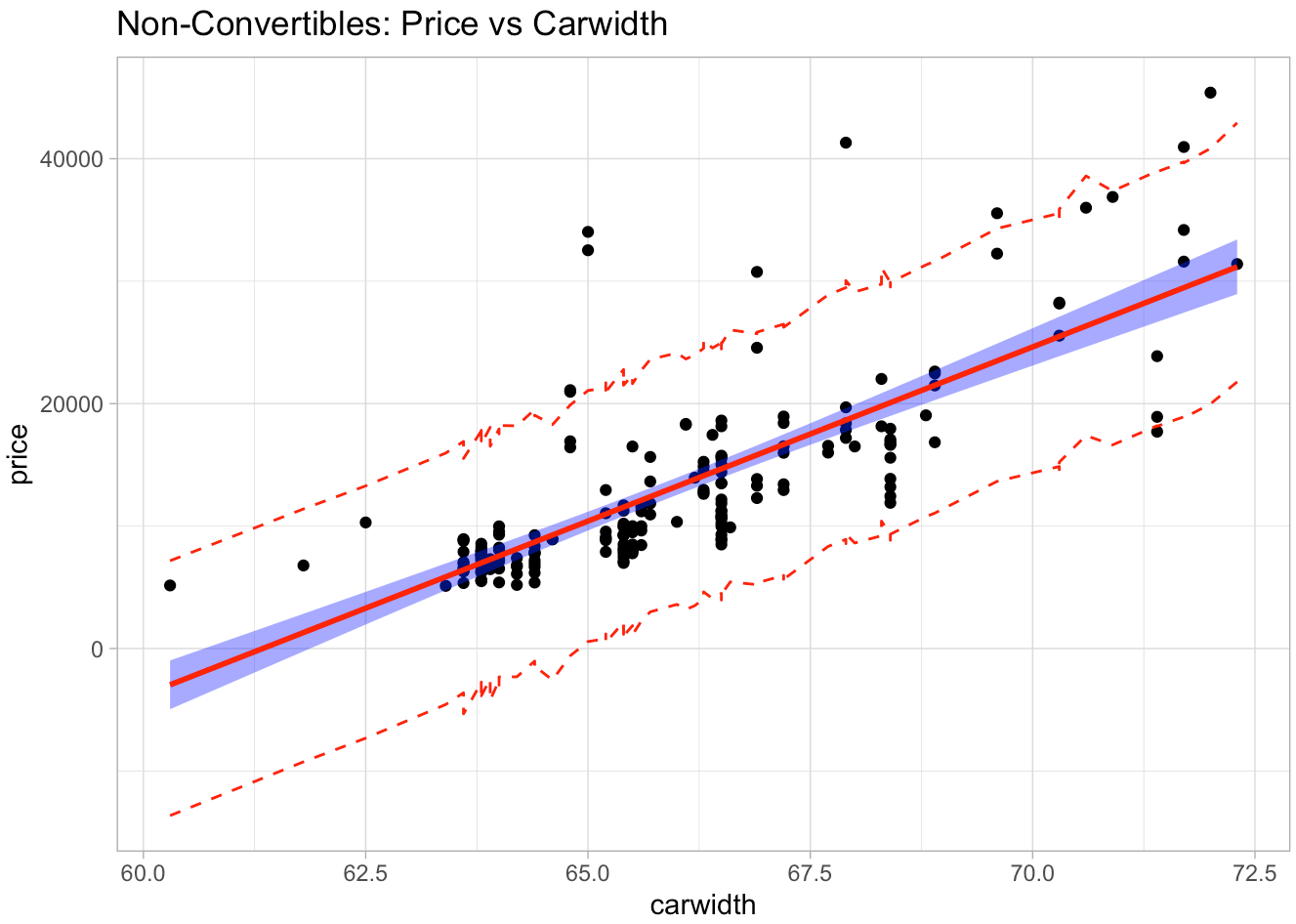
```
Ic <- predict(object = Modelo1, interval = "confidence", level = 0.96)


datos1 <- cbind(M, Ip)

MM <- subset(datos1, convertible == 1)
MH <- subset(datos1, convertible == 0)

library(ggplot2)

ggplot(MM, aes(x = carwidth, y = price)) +
  ggtitle("Convertibles: Price vs Carwidth") +
  geom_point() +
  geom_line(aes(y = lwr), color = "red", linetype = "dashed") +
  geom_line(aes(y = upr), color = "red", linetype = "dashed") +
  geom_smooth(method = lm, formula = y ~ x, se = TRUE, level = 0.96, col = "blue", fill
= "pink2") +
  theme_light()
```



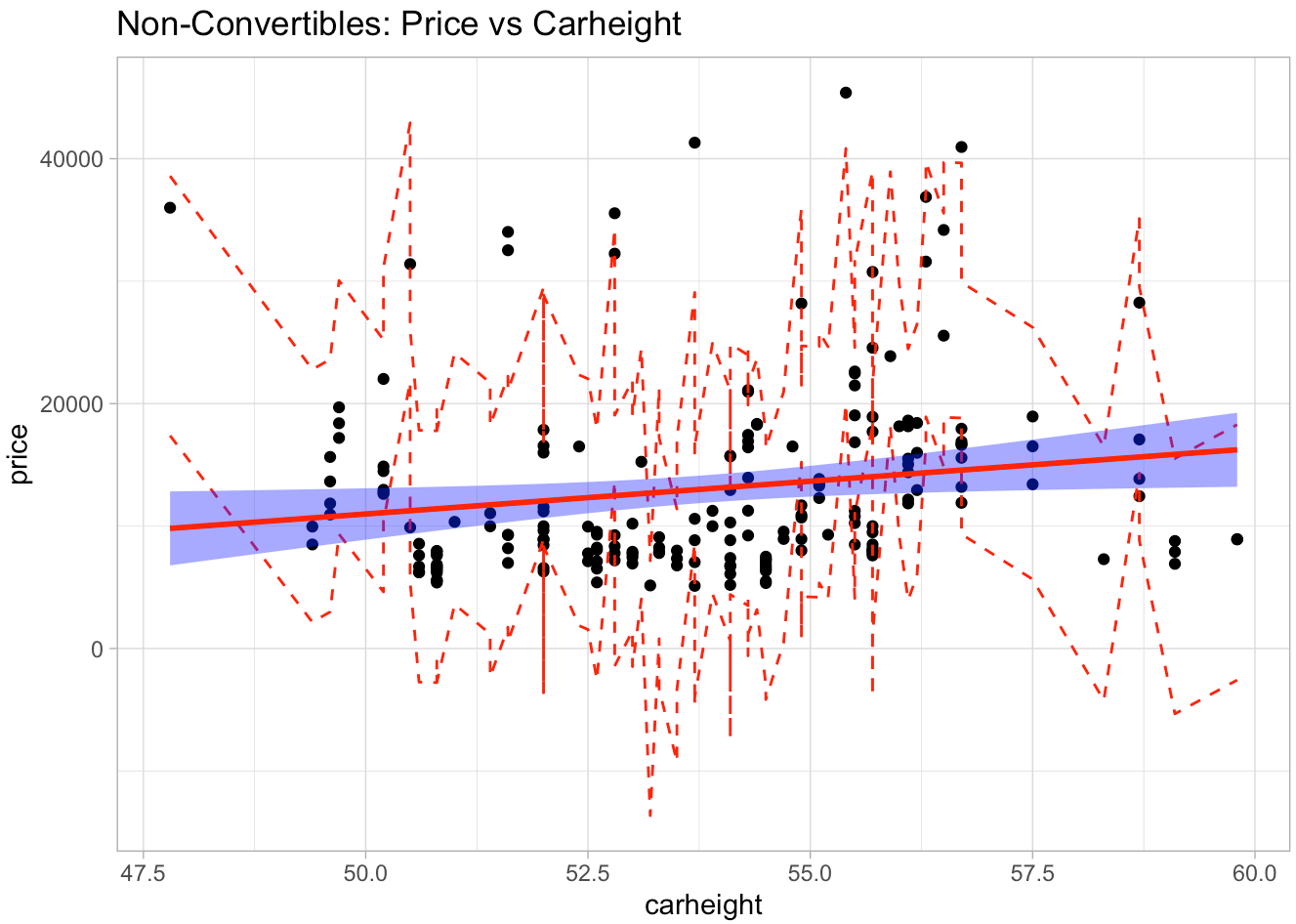

```
ggplot(MM, aes(x = carheight, y = price)) +
  ggtitle("Convertibles: Price vs Carheight") +
  geom_point() +
  geom_line(aes(y = lwr), color = "red", linetype = "dashed") +  # Lower bound
  geom_line(aes(y = upr), color = "red", linetype = "dashed") +  # Upper bound
  geom_smooth(method = lm, formula = y ~ x, se = TRUE, level = 0.96, col = "blue", fill
= "pink2") +
  theme_light()
```




```
ggplot(MH, aes(x = carwidth, y = price)) +
  ggtitle("Non-Convertibles: Price vs Carwidth") +
  geom_point() +
  geom_line(aes(y = lwr), color = "red", linetype = "dashed") +  # Lower bound
  geom_line(aes(y = upr), color = "red", linetype = "dashed") +  # Upper bound
  geom_smooth(method = lm, formula = y ~ x, se = TRUE, level = 0.96, col = "red", fill =
"blue") +
  theme_light()
```




```
ggplot(MH, aes(x = carheight, y = price)) +
  ggtitle("Non-Convertibles: Price vs Carheight") +
  geom_point() +
  geom_line(aes(y = lwr), color = "red", linetype = "dashed") +  # Lower bound
  geom_line(aes(y = upr), color = "red", linetype = "dashed") +  # Upper bound
  geom_smooth(method = lm, formula = y ~ x, se = TRUE, level = 0.96, col = "red", fill =
"blue") +
  theme_light()
```



En este ejemplo es mucho más relevante el grupo en el cual no se cuentan con los autos convertibles puesto que estos representan la mayor proporción de los datos disponibles. La cantidad de datos que se tienen de autos convertibles no es suficiente para realizar un análisis significativo.

Más allá

Si tuviera que realizar un modelo escogiendo cualquiera de las variables, utilizaría las que tienen mas correlación con el precio.

```
cor_matrix <- cor(M1[sapply(M1, is.numeric)])

price_corr <- cor_matrix[, "price"]

sorted_corr <- sort(abs(price_corr), decreasing = TRUE)
top_features <- names(sorted_corr)[2:12]

top_features_corr <- price_corr[top_features]
print(top_features_corr)
```

```
##  enginesize  curbweight  horsepower    carwidth  highwaympg     citympg
##  0.87414480  0.83530488  0.80813882  0.75932530 -0.69759909 -0.68575134
##   carlength   wheelbase   carheight     peakrpm   symboling
##  0.68292002  0.57781560  0.11933623 -0.08526715 -0.07997822
```

```
summary(M1[, c(top_features, "price")])
```

```
##    enginesize      curbweight     horsepower       carwidth       highwaympg
##  Min.   : 61.0   Min.   :1488   Min.   : 48.0   Min.   :60.30   Min.   :16.00
##  1st Qu.: 97.0   1st Qu.:2145   1st Qu.: 70.0   1st Qu.:64.10   1st Qu.:25.00
##  Median :120.0   Median :2414   Median : 95.0   Median :65.50   Median :30.00
##  Mean   :126.9   Mean   :2556   Mean   :104.1   Mean   :65.91   Mean   :30.75
##  3rd Qu.:141.0   3rd Qu.:2935   3rd Qu.:116.0   3rd Qu.:66.90   3rd Qu.:34.00
##  Max.   :326.0   Max.   :4066   Max.   :288.0   Max.   :72.30   Max.   :54.00
##     citympg        carlength       wheelbase        carheight
##  Min.   :13.00   Min.   :141.1   Min.   : 86.60   Min.   :47.80
##  1st Qu.:19.00   1st Qu.:166.3   1st Qu.: 94.50   1st Qu.:52.00
##  Median :24.00   Median :173.2   Median : 97.00   Median :54.10
##  Mean   :25.22   Mean   :174.0   Mean   : 98.76   Mean   :53.72
##  3rd Qu.:30.00   3rd Qu.:183.1   3rd Qu.:102.40   3rd Qu.:55.50
##  Max.   :49.00   Max.   :208.1   Max.   :120.90   Max.   :59.80
##     peakrpm       symboling          price
##  Min.   :4150   Min.   :-2.0000   Min.   : 5118
##  1st Qu.:4800   1st Qu.: 0.0000   1st Qu.: 7788
##  Median :5200   Median : 1.0000   Median :10295
##  Mean   :5125   Mean   : 0.8341   Mean   :13277
##  3rd Qu.:5500   3rd Qu.: 2.0000   3rd Qu.:16503
##  Max.   :6600   Max.   : 3.0000   Max.   :45400
```

```
cor(M1[, c(top_features, "price")])
```

```
##             enginesize curbweight  horsepower   carwidth  highwaympg
## enginesize  1.00000000  0.8505941  0.80976865  0.7354334 -0.67746991
## curbweight  0.85059407  1.0000000  0.75073925  0.8670325 -0.79746479
## horsepower  0.80976865  0.7507393  1.00000000  0.6407321 -0.77054389
## carwidth    0.73543340  0.8670325  0.64073208  1.0000000 -0.67721792
## highwaympg -0.67746991 -0.7974648 -0.77054389 -0.6772179  1.00000000
## citympg    -0.65365792 -0.7574138 -0.80145618 -0.6427043  0.97133704
## carlength   0.68335987  0.8777285  0.55262297  0.8411183 -0.70466160
## wheelbase   0.56932868  0.7763863  0.35329448  0.7951436 -0.54408192
## carheight   0.06714874  0.2955717 -0.10880206  0.2792103 -0.10735763
## peakrpm    -0.24465983 -0.2662432  0.13107251 -0.2200123 -0.05427481
## symboling  -0.10578971 -0.2276906  0.07087272 -0.2329191  0.03460600
## price       0.87414480  0.8353049  0.80813882  0.7593253 -0.69759909
##                citympg  carlength  wheelbase   carheight     peakrpm
## enginesize -0.65365792  0.6833599  0.5693287  0.06714874 -0.24465983
## curbweight -0.75741378  0.8777285  0.7763863  0.29557173 -0.26624318
## horsepower -0.80145618  0.5526230  0.3532945 -0.10880206  0.13107251
## carwidth   -0.64270434  0.8411183  0.7951436  0.27921032 -0.22001230
## highwaympg  0.97133704 -0.7046616 -0.5440819 -0.10735763 -0.05427481
## citympg     1.00000000 -0.6709087 -0.4704136 -0.04863963 -0.11354438
## carlength  -0.67090866  1.0000000  0.8745875  0.49102946 -0.28724220
## wheelbase  -0.47041361  0.8745875  1.0000000  0.58943476 -0.36046875
## carheight  -0.04863963  0.4910295  0.5894348  1.00000000 -0.32041072
## peakrpm    -0.11354438 -0.2872422 -0.3604687 -0.32041072  1.00000000
## symboling  -0.03582263 -0.3576115 -0.5319537 -0.54103820  0.27360625
## price      -0.68575134  0.6829200  0.5778156  0.11933623 -0.08526715
##              symboling       price
## enginesize -0.10578971  0.87414480
## curbweight -0.22769059  0.83530488
## horsepower  0.07087272  0.80813882
## carwidth   -0.23291906  0.75932530
## highwaympg  0.03460600 -0.69759909
## citympg    -0.03582263 -0.68575134
## carlength  -0.35761152  0.68292002
## wheelbase  -0.53195368  0.57781560
## carheight  -0.54103820  0.11933623
## peakrpm     0.27360625 -0.08526715
## symboling   1.00000000 -0.07997822
## price      -0.07997822  1.00000000
```

A partir de estas, eliminaría las que tengan mucha relación con otras variables que deberían ser independientes para evitar multicolinearidad entre variables.