Actividad Integradora 2

Ricardo Salinas

2024-09-06

1. Exploración de la base de datos

```
#1.Exploración de la base de datos
library(readr)
autop = read_csv("Downloads/precios_autos.csv")

## Rows: 205 Columns: 21
## ── Column specification
────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (7): CarName, fueltype, carbody, drivewheel, enginelocation,
enginetype...
## dbl (14): symboling, wheelbase, carlength, carwidth, carheight,
curbweight, ...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this
message.

autos = autop[, c('carheight', 'carwidth', 'carbody', 'price')]

#  Calcula medidas estadísticas apropiadas para las variables:
#cuantitativas (media, desviación estándar, cuantiles, etc)

summary(autos)

##    carheight        carwidth       carbody              price
##  Min.   :47.80   Min.   :60.30   Length:205         Min.   : 5118
##  1st Qu.:52.00   1st Qu.:64.10   Class :character   1st Qu.: 7788
##  Median :54.10   Median :65.50   Mode  :character   Median :10295
##  Mean   :53.72   Mean   :65.91                      Mean   :13277
##  3rd Qu.:55.50   3rd Qu.:66.90                      3rd Qu.:16503
##  Max.   :59.80   Max.   :72.30                      Max.   :45400

#cualitativas: cuantiles, frecuencias (puedes usar el comando table o
prop.table)

cual1 = table(autos$carbody)
cual2 = table(autos$carheight)
cual3 = table(autos$carwidth)

print(cual1)
```

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

```
print(cual2)
```

```
##
## 47.8 48.8 49.4 49.6 49.7 50.2 50.5 50.6 50.8   51 51.4 51.6   52 52.4 52.5
52.6
##    1    2    2    4    3    6    2    5   14    1    2    7   12    1    3
7
## 52.8   53 53.1 53.2 53.3 53.5 53.7 53.9 54.1 54.3 54.4 54.5 54.7 54.8 54.9
55.1
##    6    6    1    1    4    3    5    2   10    8    2   10    2    1    6
5
## 55.2 55.4 55.5 55.6 55.7 55.9   56 56.1 56.2 56.3 56.5 56.7 57.5 58.3 58.7
59.1
##    1    1    9    1   12    1    1    7    3    2    2    8    3    1    4
3
## 59.8
##    2
```

```
print(cual3)
```

```
##
## 60.3 61.8 62.5 63.4 63.6 63.8 63.9   64 64.1 64.2 64.4 64.6 64.8   65 65.2
65.4
##    1    1    1    1   11   24    3    9    2    6   10    2    4    3    7
15
## 65.5 65.6 65.7   66 66.1 66.2 66.3 66.4 66.5 66.6 66.9 67.2 67.7 67.9   68
68.3
##    8    6    4    1    2    1    6    1   23    1    5    6    2    5    1
2
## 68.4 68.8 68.9 69.6 70.3 70.5 70.6 70.9 71.4 71.7   72 72.3
##   10    1    4    2    3    1    1    1    3    3    1    1
```

```
#Analiza la correlación entre las variables (analiza posible colinealidad
entre las variables)
```

```
correlacion = cor(autos$carheight, autos$carwidth)
print("La correlacion entre car height y car width es: ")
```

```
## [1] "La correlacion entre car height y car width es: "
```

```
print(correlacion)
```

```
## [1] 0.2792103
```

```
print("La correlacion entre car height y car width es baja")
```

```
## [1] "La correlacion entre car height y car width es baja"
```

```r
correlacion1 = cor(autos$price, autos$carheight)
print("La correlacion entre price y car height es: ")
```

```
## [1] "La correlacion entre price y car height es: "
```

```r
print(correlacion1)
```

```
## [1] 0.1193362
```

```r
print("La correlacion entre car height y price es baja")
```

```
## [1] "La correlacion entre car height y price es baja"
```

```r
correlacion2 = cor(autos$price, autos$carwidth)
print("La correlacion entre price y car width es: ")
```

```
## [1] "La correlacion entre price y car width es: "
```

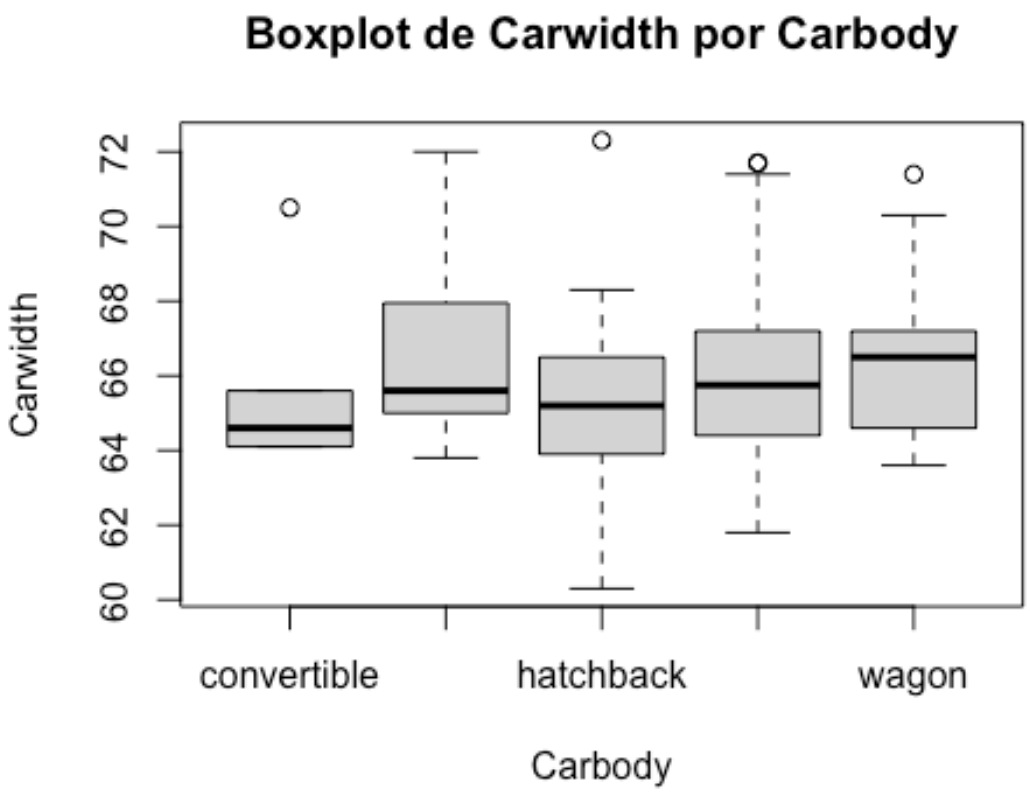
```r
print(correlacion2)
```

```
## [1] 0.7593253
```

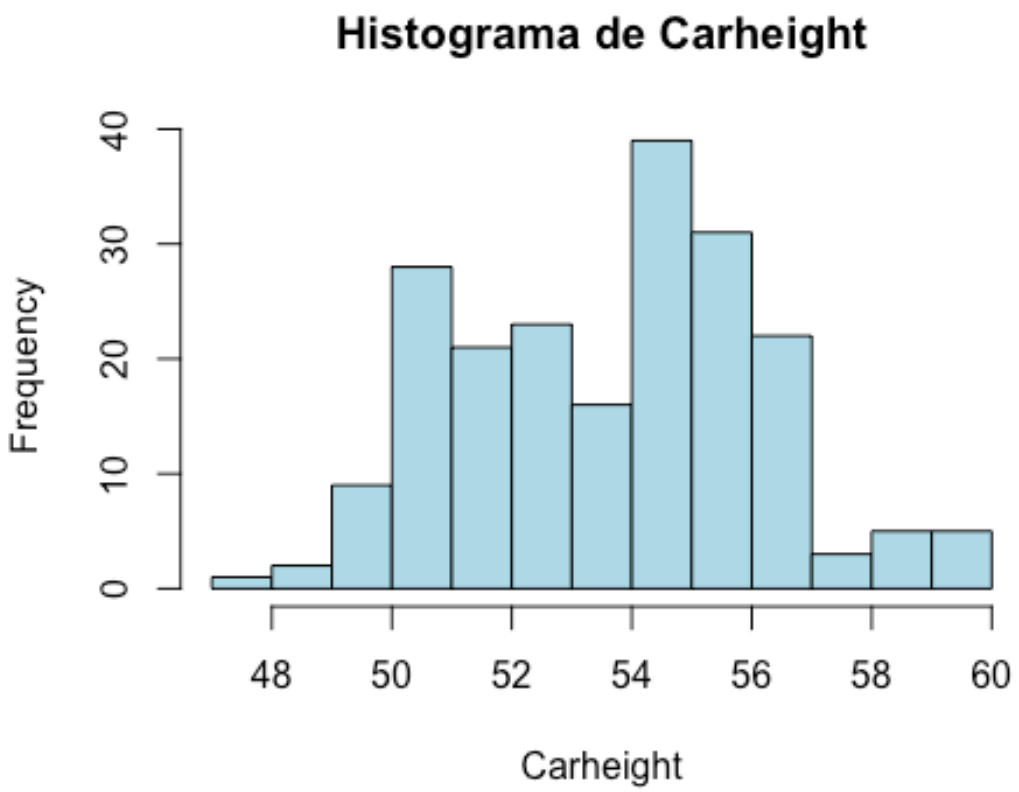
```r
print("La correlacion entre price y car width es alta")
```

```
## [1] "La correlacion entre price y car width es alta"
```

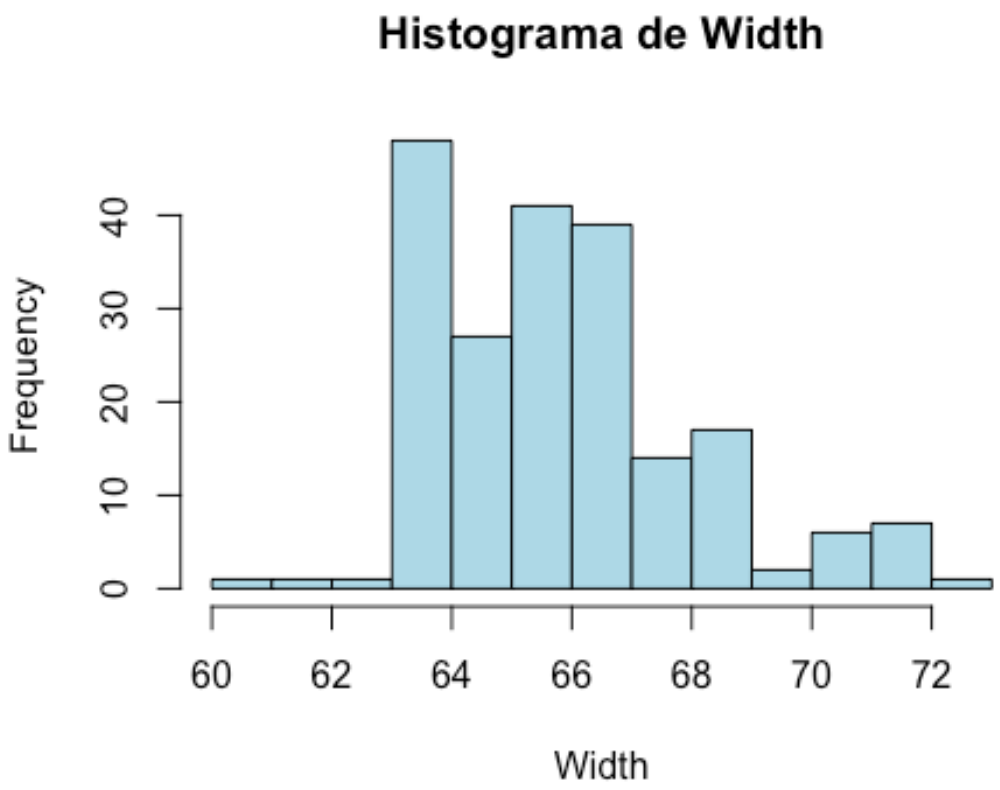
```r
#Explora los datos usando herramientas de visualización (si lo consideras
necesario):
#Variables cuantitativas:

#Boxplot (visualización de datos atípicos)

boxplot(autos$carwidth ~ autos$carbody, main = "Boxplot de Carwidth por
Carbody", xlab = "Carbody", ylab = "Carwidth")
```



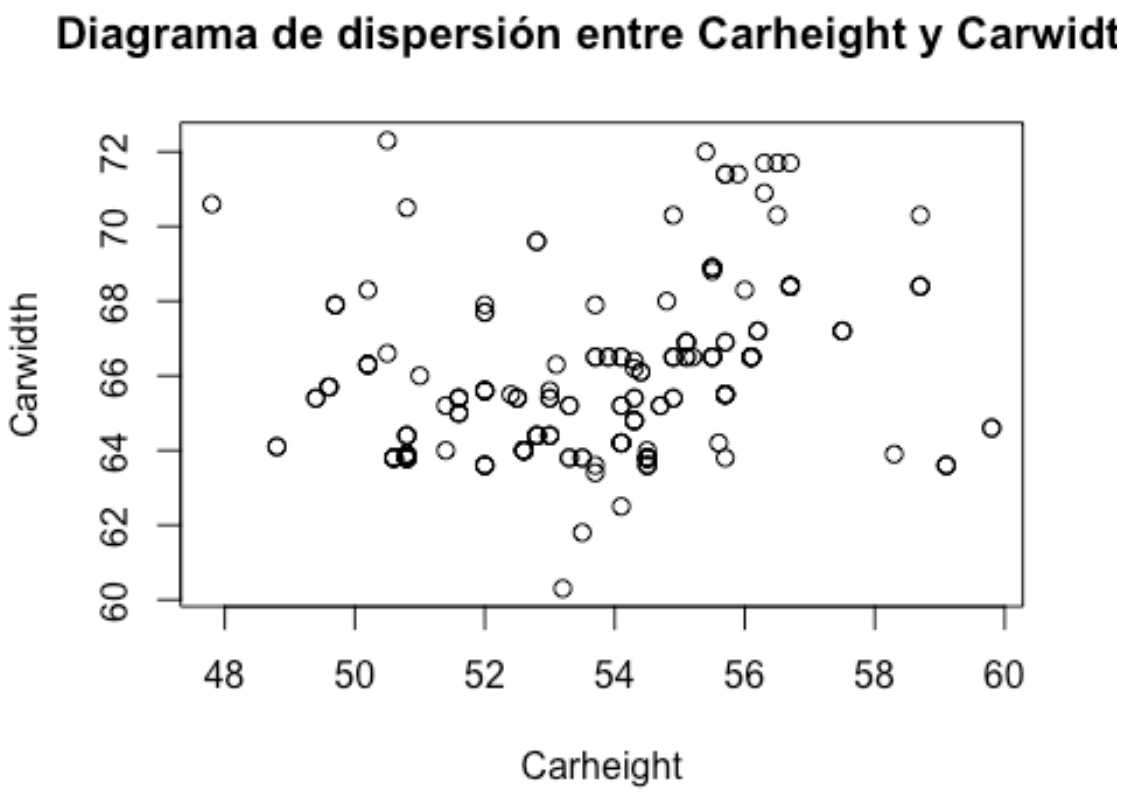

```r
#Histogramas
hist(autos$carheight, main = "Histograma de Carheight", xlab = "Carheight",
col = "lightblue")
```



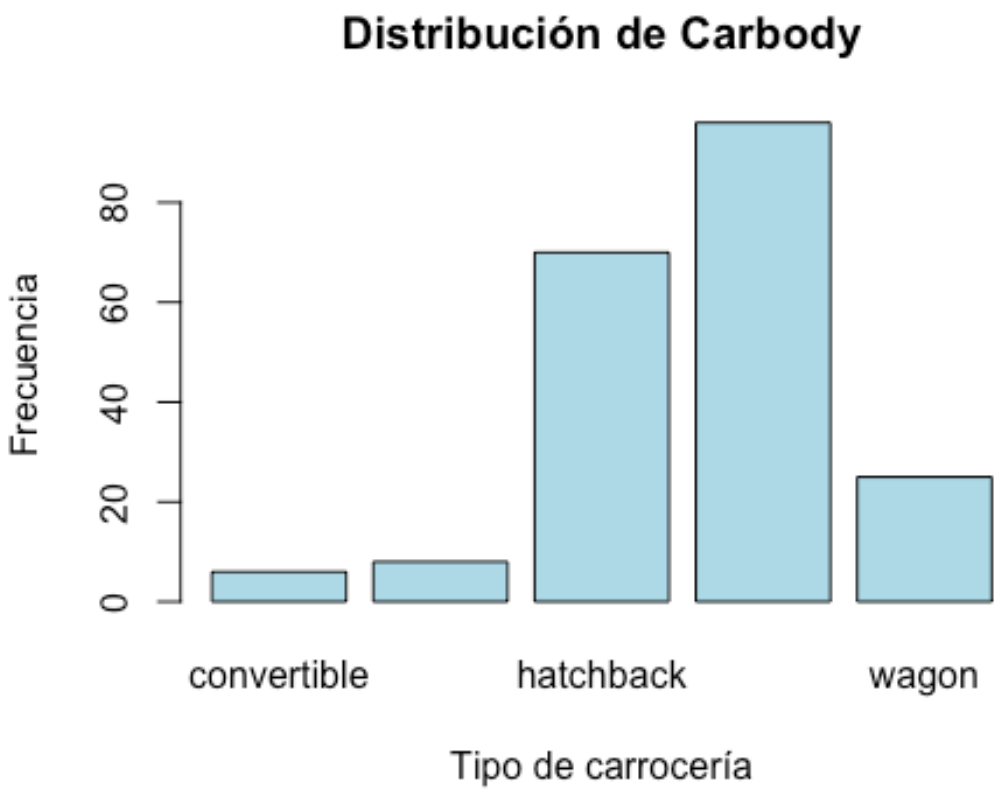

```r
hist(autos$carwidth, main = "Histograma de Width", xlab = "Width", col = "lightblue")
```



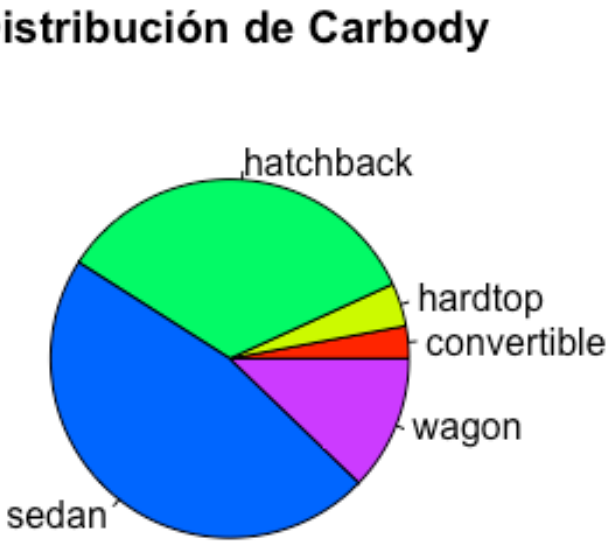

```r
#Diagramas de dispersión y correlación por pares

plot(autos$carheight, autos$carwidth, main = "Diagrama de dispersión entre
Carheight y Carwidth", xlab = "Carheight", ylab = "Carwidth")
```



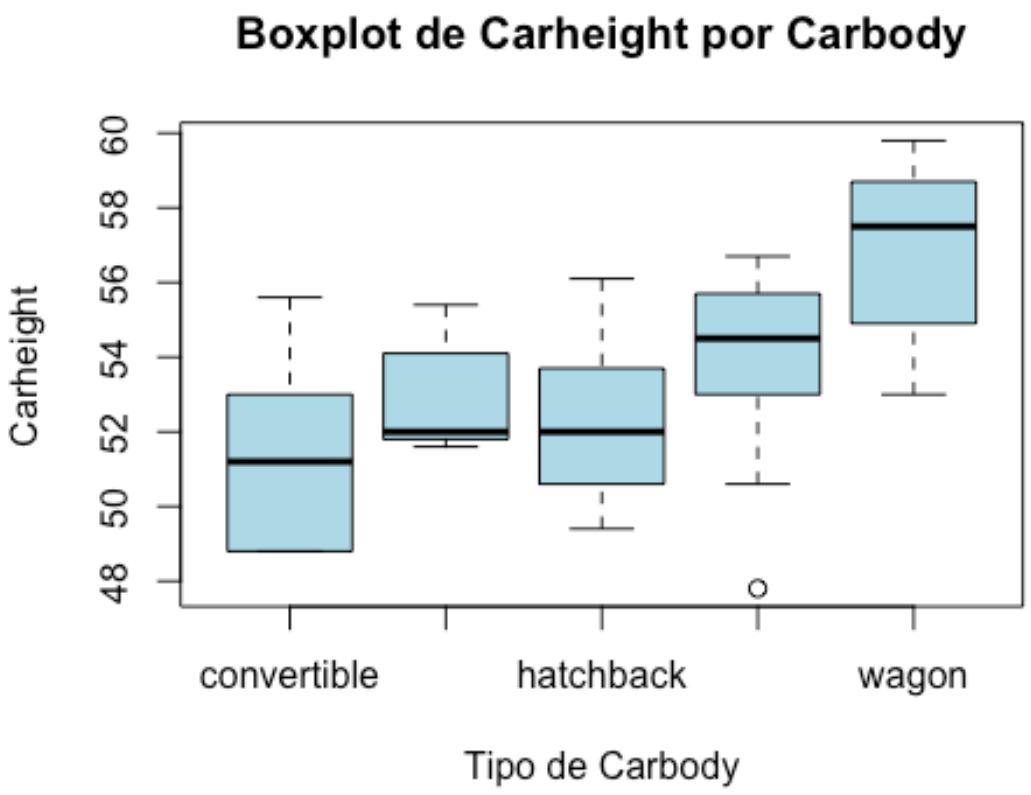

```r
#Variables categóricas
barplot(table(autos$carbody), main = "Distribución de Carbody", xlab = "Tipo de carrocería", ylab = "Frecuencia", col = "lightblue")
```



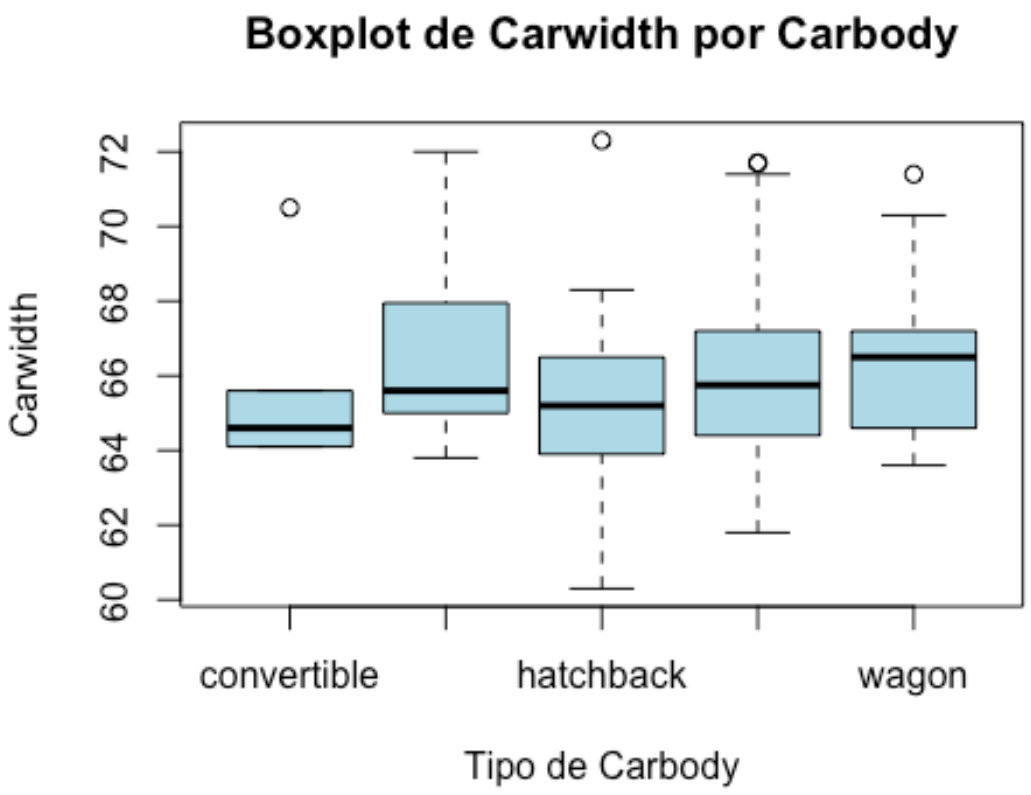

```r
#Distribución de los datos (diagramas de barras, diagramas de pastel)
pie(table(autos$carbody), main = "Distribución de Carbody", col =
rainbow(length(table(autos$carbody))))
```




```r
#Boxplot por categoría de las variables cuantitativas
boxplot(autos$carheight ~ autos$carbody, main = "Boxplot de Carheight por
Carbody", xlab = "Tipo de Carbody", ylab = "Carheight", col = "lightblue")
```




```r
boxplot(autos$carwidth ~ autos$carbody, main = "Boxplot de Carwidth por
Carbody", xlab = "Tipo de Carbody", ylab = "Carwidth", col = "lightblue")
```



2. Modelación y verificación del modelo

```
#Encuentra la ecuación de regresión de mejor ajuste. Propón al menos 2
modelos de ajuste para encontrar la mejor forma de ajustar la variable
precio.

#El primer modelo sera (price ~ carwidth + carheight)
#El segundo modelo sera (price ~ carwidth + carheight + carbody)

#Para cada uno de los modelos propuestos:
#Realiza la regresión entre las variables involucradas

modelo1 = lm(price ~ carwidth + carheight, data = autos)

modelo2 = lm(price ~ carwidth + carheight + carbody, data = autos)

#Analiza la significancia del modelo:
```

H0 = 0, el modelo no es significativo H1 ≠ 0, el modelo si es significativo

```
#Valida la significancia del modelo con un alfa de 0.04 (incluye las
hipótesis que pruebas y el valor frontera)
summary1 = summary(modelo1)
p_value_modelo1 = summary1$fstatistic[1]
```

```r
p_value = pf(summary1$fstatistic[1],
summary1$fstatistic[2],summary1$fstatistic[3], lower.tail = FALSE)
print(paste("El p-value del modelo 1 es:", p_value))
```

```
## [1] "El p-value del modelo 1 es: 2.13162386719804e-39"
```

```r
print("El valor de p-value del primer modelo no es igual a cero, se rechaza
la hipotesis")
```

```
## [1] "El valor de p-value del primer modelo no es igual a cero, se rechaza
la hipotesis"
```

```r
summary2 = summary(modelo2)
p_value_modelo2 = summary2$fstatistic[1]
p_value1 = pf(summary2$fstatistic[1],
summary2$fstatistic[2],summary2$fstatistic[3], lower.tail = FALSE)
print(paste("El p-value del modelo 2 es:", p_value1))
```

```
## [1] "El p-value del modelo 2 es: 3.21737716060082e-44"
```

```r
print("El valor de p-value del segundo modelo no es igual a cero,, se rechaza
la hipotesis")
```

```
## [1] "El valor de p-value del segundo modelo no es igual a cero,, se
rechaza la hipotesis"
```

```r
#Valida la significancia de βî con un alfa de 0.04 (incluye las hipótesis que
pruebas y el valor frontera de cada una de ellas)
print("El valor de p-value del primer modelo no es igual a cero,, se aprueba
H1 y se demuestra que si es significativo.")
```

```
## [1] "El valor de p-value del primer modelo no es igual a cero,, se aprueba
H1 y se demuestra que si es significativo."
```

```r
print("El valor de p-value del segundo modelo no es igual a cero,, se aprueba
H1 y se demuestra que si es significativo.")
```

```
## [1] "El valor de p-value del segundo modelo no es igual a cero,, se
aprueba H1 y se demuestra que si es significativo."
```

```r
#Indica cuál es el porcentaje de variación explicada por el modelo.

ve = summary1$r.squared
print("La variacion explicada de Modelo 1 es: ")
```

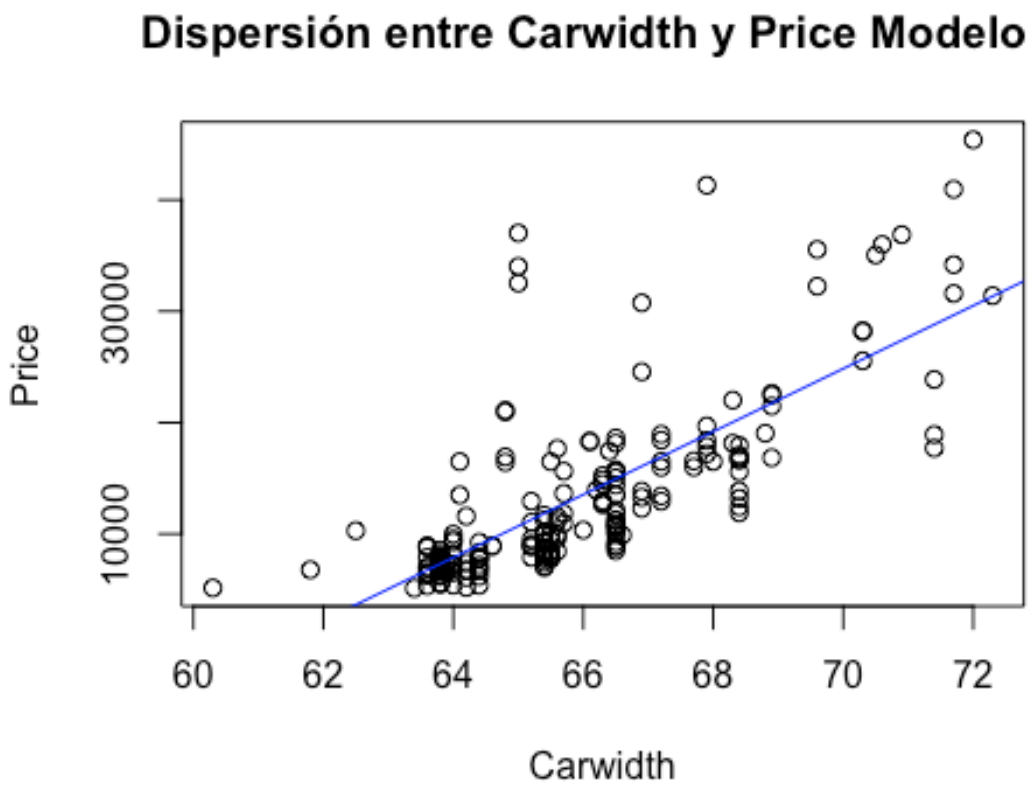
```
## [1] "La variacion explicada de Modelo 1 es: "
```

```r
print(ve)
```

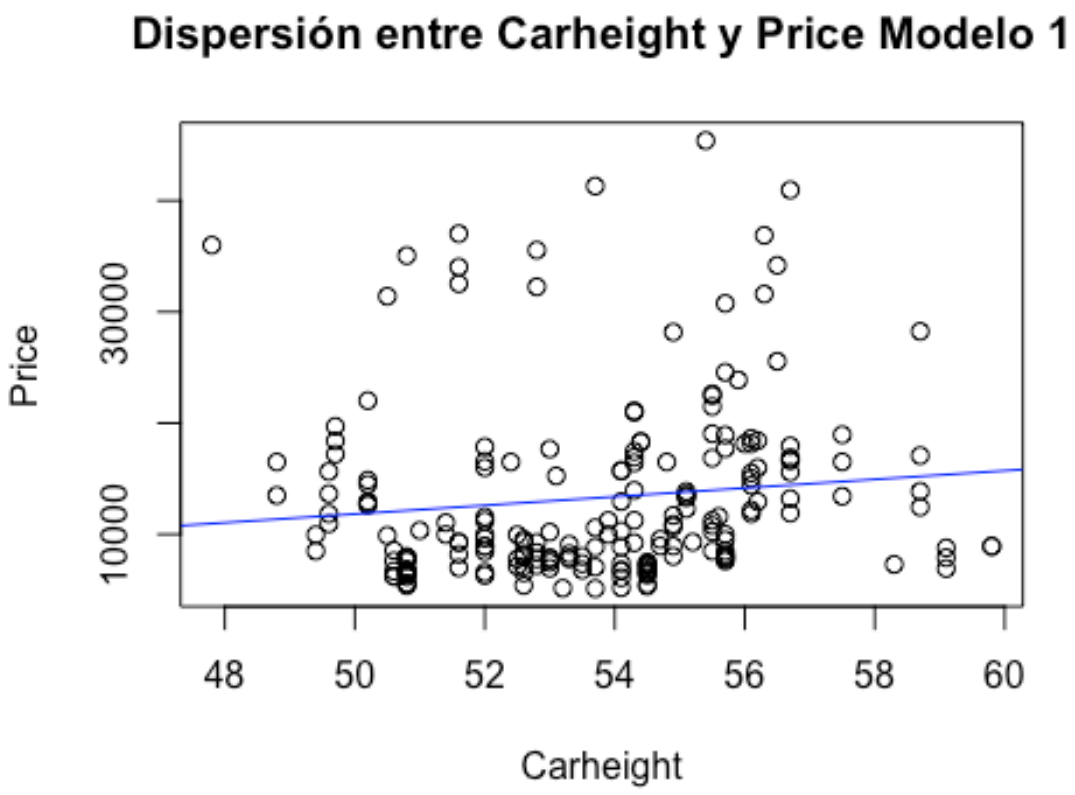
```
## [1] 0.5858898
```

```r
ve1 = summary2$r.squared
print("La variacion explicada de Modelo 2 es: ")
```

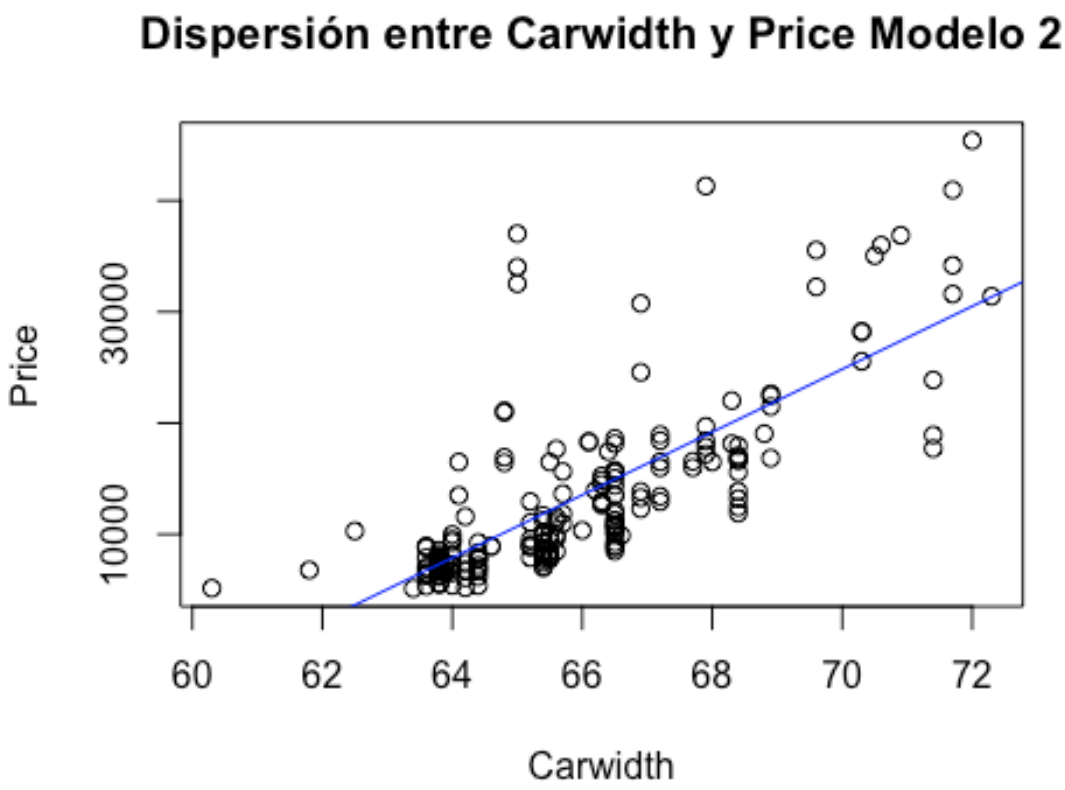
```
## [1] "La variacion explicada de Modelo 2 es: "

print(ve1)

## [1] 0.6636217

#Dibuja el diagrama de dispersión de los datos por pares y la recta de mejor
ajuste.
#Modelo 1
plot(autos$carwidth, autos$price, main = "Dispersión entre Carwidth y Price
Modelo 1", xlab = "Carwidth", ylab = "Price")
abline(lm(price ~ carwidth, data = autos), col = "blue")
```



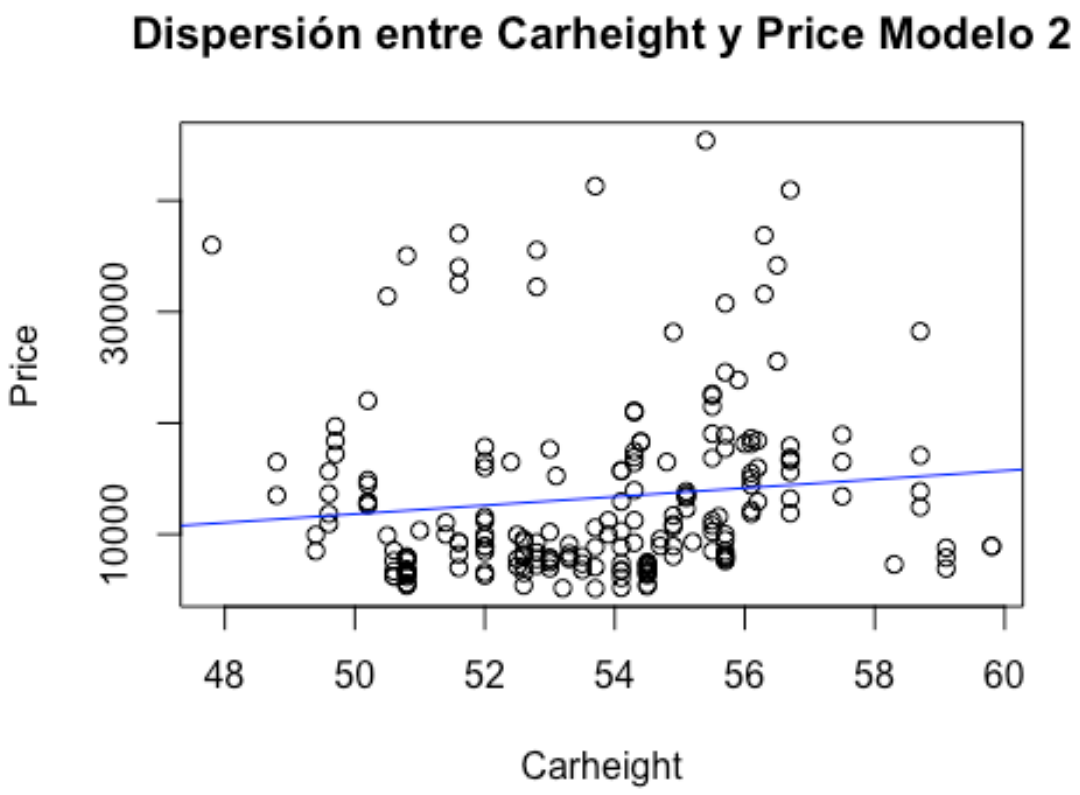

```
plot(autos$carheight, autos$price, main = "Dispersión entre Carheight y Price
Modelo 1", xlab = "Carheight", ylab = "Price")
abline(lm(price ~ carheight, data = autos), col = "blue")
```




```r
#Modelo 2

carbodyc = c("sedan" = "blue", "hatchback" = "red", "convertible" = "green",
"wagon" = "purple", "hardtop" = "orange")

plot(autos$carwidth, autos$price, main = "Dispersión entre Carwidth y Price
Modelo 2", xlab = "Carwidth", ylab = "Price")
abline(lm(price ~ carwidth, data = autos), col = "blue")
```




```r
plot(autos$carheight, autos$price, main = "Dispersión entre Carheight y Price
Modelo 2", xlab = "Carheight", ylab = "Price")
abline(lm(price ~ carheight, data = autos), col = "blue")
```




```
plot(autos$carwidth, autos$price, col = carbodyc[autos$carbody], pch = 19,
xlab = "Carwidth", ylab = "Price", main = "Dispersión entre Carwidth y Price,
diferenciando Carbody")
abline(lm(price ~ carwidth, data = autos), col = "black", lwd = 2)
legend("topleft", legend = names(carbodyc), col = carbodyc, pch = 19)

#Interpreta en el contexto del problema cada uno de los análisis que hiciste.
print("Con los analisis hechos podemos ver que el segundo modelo es mejor,
esto puede ser debido a que el carbody si llega a tener influencia en el
precio del carro, lo cual deja a que el modelo 1 no sea tan bueno como el
segundo.")
```

```
## [1] "Con los analisis hechos podemos ver que el segundo modelo es mejor,
esto puede ser debido a que el carbody si llega a tener influencia en el
precio del carro, lo cual deja a que el modelo 1 no sea tan bueno como el
segundo."
```

```
#Analiza la validez de los modelos propuestos:
#Normalidad de los residuos
library(nortest)
ad.test(modelo1$residuals)
```

```
##
##  Anderson-Darling normality test
```

```
##
## data:  modelo1$residuals
## A = 10.319, p-value < 2.2e-16
```

```
ad.test(modelo2$residuals)
```

```
##
##  Anderson-Darling normality test
##
## data:  modelo2$residuals
## A = 6.3689, p-value = 1.103e-15
```

```
#Verificación de media cero
```

```
t.test(modelo1$residuals)
```

```
##
##  One Sample t-test
##
## data:  modelo1$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
```

```
t.test(modelo2$residuals)
```

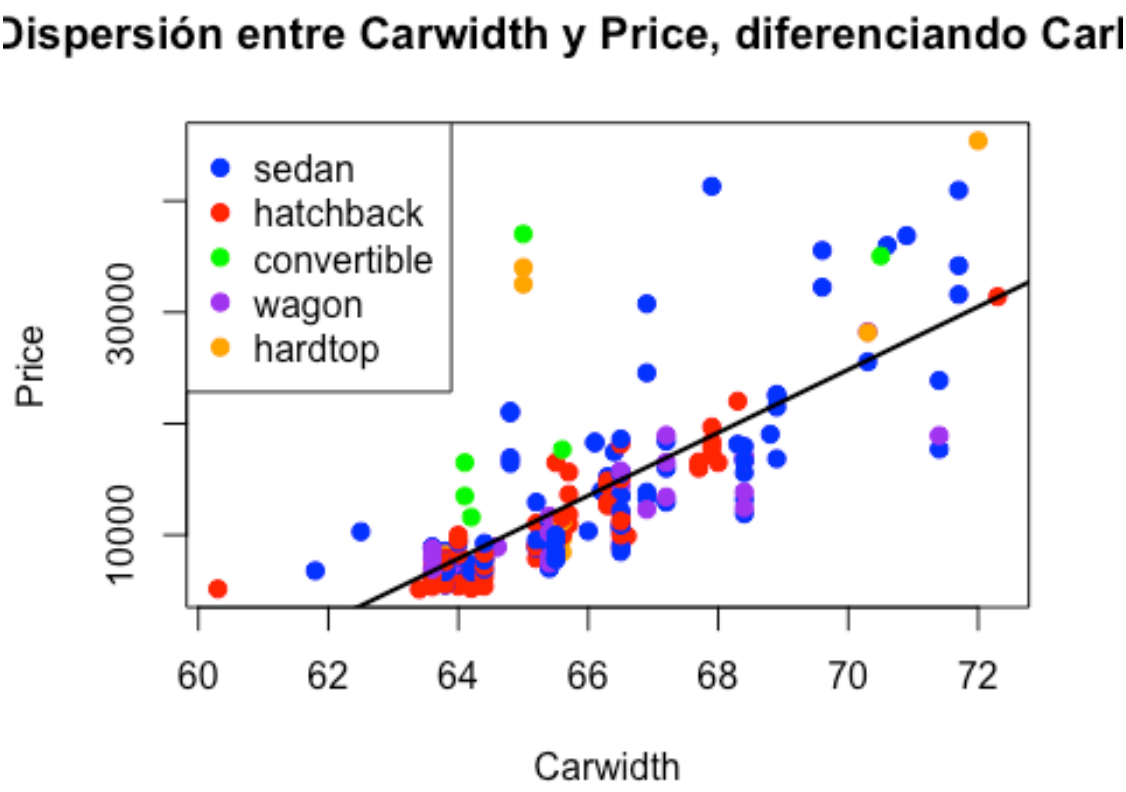
```
##
##  One Sample t-test
##
## data:  modelo2$residuals
## t = 3.9415e-16, df = 204, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -638.0485  638.0485
## sample estimates:
##    mean of x
## 1.275511e-13
```

```
#Homocedasticidad, linealidad e independencia
#Independencia
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 = 0.67299, 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 = 92.131, df = 1, p-value < 2.2e-16

dwtest(modelo2)
```

```
##
##  Durbin-Watson test
##
## data:  modelo2
## DW = 0.76974, 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 = 80.703, df = 1, p-value < 2.2e-16
```

```
#Homcedasticidad
bptest(modelo1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  modelo1
## BP = 4.6072, df = 2, p-value = 0.0999
```

```
gqtest(modelo1)
```

```
##
##  Goldfeld-Quandt test
##
## data:  modelo1
## GQ = 0.8351, df1 = 100, df2 = 99, p-value = 0.815
## alternative hypothesis: variance increases from segment 1 to 2
```

```
bptest(modelo2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  modelo2
## BP = 37.966, df = 6, p-value = 1.141e-06
```

```
gqtest(modelo2)
```

```
##
##  Goldfeld-Quandt test
##
## data:  modelo2
## GQ = 0.67139, df1 = 96, df2 = 95, p-value = 0.9736
## alternative hypothesis: variance increases from segment 1 to 2
```

```
#Linealidad
dwtest(modelo1)
```

```
##
##  Durbin-Watson test
##
## data:  modelo1
## DW = 0.67299, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

```
dwtest(modelo2)
```

```
##
##  Durbin-Watson test
##
## data:  modelo2
## DW = 0.76974, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

```
#Interpreta cada uno de los analisis que realizaste
print("En los análisis se puede ver que el segundo modelo es mejor,
probablemente porque carbody sí influye en el precio del carro, asi podemos
ver que el primer modelo no es tan efectivo como el segundo, ya que no
considera esa variable importante.")
```

```
## [1] "En los análisis se puede ver que el segundo modelo es mejor,
probablemente porque carbody sí influye en el precio del carro, asi podemos
ver que el primer modelo no es tan efectivo como el segundo, ya que no
considera esa variable importante."
```

```
#Emite una conclusión final sobre el mejor modelo de regresión lineal y
contesta la pregunta central:
#Concluye sobre el mejor modelo que encontraste y argumenta por qué es el
mejor
print("El mejor modelo definitivamente es el segundo modelo, ya que este
tiene un p value de 3.217e-44, lo cual lo hace significativo, y tambien este
modelo explica el 66% de la variacion de los precios de los carros, valor
mucho mas alto que el primer modelo, por lo cual se toma el segundo modelo
como mejor opcion")
```

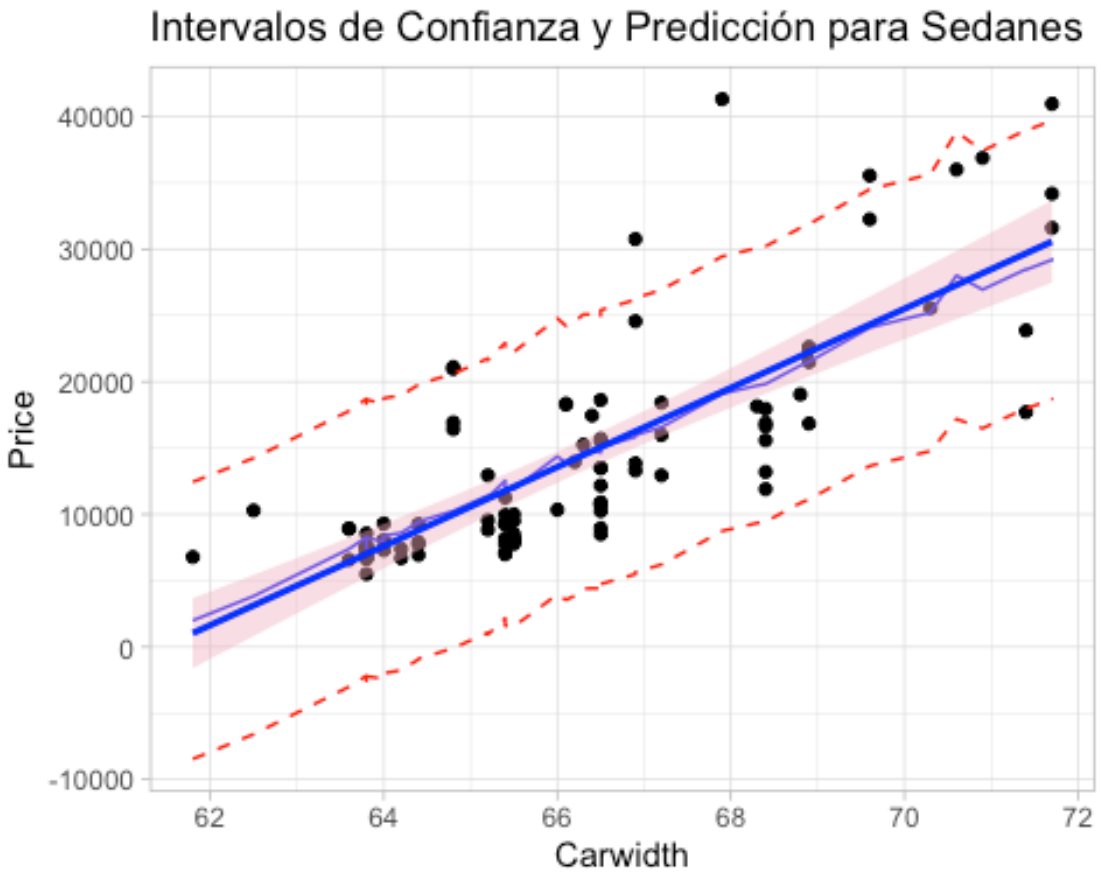
```
## [1] "El mejor modelo definitivamente es el segundo modelo, ya que este
tiene un p value de 3.217e-44, lo cual lo hace significativo, y tambien este
modelo explica el 66% de la variacion de los precios de los carros, valor
mucho mas alto que el primer modelo, por lo cual se toma el segundo modelo
como mejor opcion"
```

```
#¿Cuáles de las variables asignadas influyen en el precio del auto? ¿de qué
manera lo hacen
print("Las variables que influyen en el precio del auto son carwidth,
carheight, y carbody, carwidth tiene un impacto fuerte y positivo en el
precio, lo que significa que autos más anchos suelen ser más caros y
carheight también afecta el precio, pero de forma menos marcada. Tambien
carbody resulta ser muy importante, ya que el tipo de carrocería puede hacer
que el precio varíe considerablemente entre diferentes modelos de autos.")
```

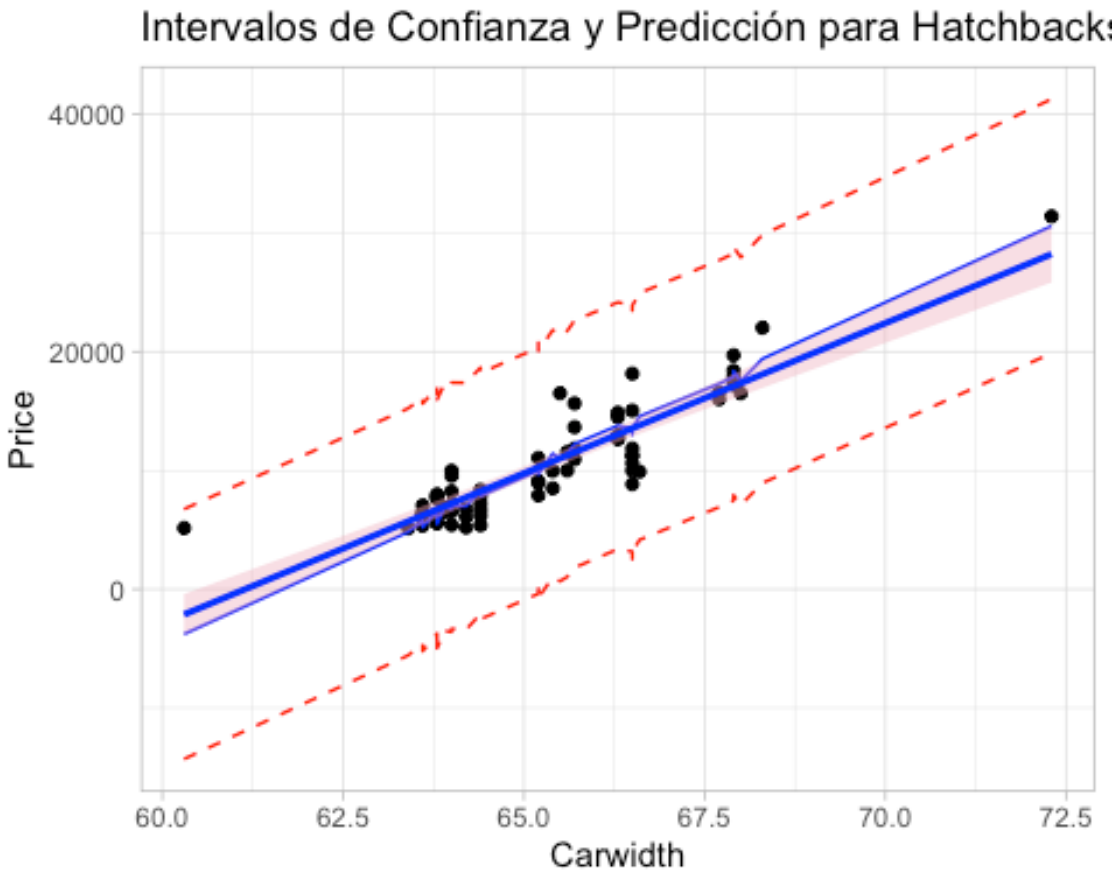
```
## [1] "Las variables que influyen en el precio del auto son carwidth,
carheight, y carbody, carwidth tiene un impacto fuerte y positivo en el
precio, lo que significa que autos más anchos suelen ser más caros y
carheight también afecta el precio, pero de forma menos marcada. Tambien
carbody resulta ser muy importante, ya que el tipo de carrocería puede hacer
que el precio varíe considerablemente entre diferentes modelos de autos."
```

3. Intervalos de predicción y confianza

```
#Con los datos de las variables asignadas construye la gráfica de los
intervalos de confianza y predicción para la estimación y predicción del
precio para el mejor modelo seleccionado:
#Calcula los intervalos para la variable Y
Ip = predict(object = modelo2, interval = "prediction", level = 0.97)

## Warning in predict.lm(object = modelo2, interval = "prediction", level =
0.97): predictions on current data refer to _future_ responses

datos1 = cbind(autos, Ip)
#Selecciona la categoría de la variable cualitativa que, de acuerdo a tu
análisis resulte la más importante, y separa la base de datos por esa
variable categórica.
autos_sedan = subset(datos1, datos1$carbody == "sedan")

#Grafica por pares de variables numéricas
library(ggplot2)
ggplot(autos_sedan, aes(x = carwidth, y = price)) +
geom_point() +
geom_line(aes(y = fit), color = "blue") +
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 = 0.97, col =
"blue", fill = "pink2") +
theme_light() +
labs(title = "Intervalos de Confianza y Predicción para Sedanes", x =
"Carwidth", y = "Price")
```




```
#Puedes hacer el mismo análisis para otra categoría de la variable
cualitativa, pero no es necesario, bastará con que justiques la categoría
seleccionada anteriormente.
autos_hatch = subset(datos1, datos1$carbody == "hatchback")

ggplot(autos_hatch, aes(x = carwidth, y = price)) +
geom_point() +
geom_line(aes(y = fit), color = "blue") +
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 = 0.97, col =
"blue", fill = "pink2") +
theme_light() +
labs(title = "Intervalos de Confianza y Predicción para Hatchbacks", x =
"Carwidth", y = "Price")
```




```
#Interpreta en el contexto del problema
print("En ambos gráficos, se ve que carwidth afecta positivamente al precio,
a mayor ancho del auto, mayor es el precio, la línea azul muestra esta
tendencia, mientras que las áreas rosas y las líneas rojas indican los
intervalos de confianza y predicción, y aunque hay variabilidad en los
precios de autos individuales, la relación es clara en sedanes y hatchbacks,
lo que confirma que el ancho del auto es un factor clave en su precio.")
```

```
## [1] "En ambos gráficos, se ve que carwidth afecta positivamente al precio,
a mayor ancho del auto, mayor es el precio, la línea azul muestra esta
tendencia, mientras que las áreas rosas y las líneas rojas indican los
intervalos de confianza y predicción, y aunque hay variabilidad en los
precios de autos individuales, la relación es clara en sedanes y hatchbacks,
lo que confirma que el ancho del auto es un factor clave en su precio."
```

4. Más allá:

```
#Contesta la pregunta referida a la agrupación de variables que propuso la
empresa para el análisis: ¿propondrías una nueva agrupación de las variables
a la empresa automovilísitica?
print("Haria dos propuestas principales, una enfocada en la capacidad del
carro en cuanto al motor, tomando en cuenta horsepower y engine size, y la
segunda propuesta se basaria en las dimensiones del carro, tomando en cuenta
las variables car height, width y wheelbase.")
```

```
## [1] "Haria dos propuestas principales, una enfocada en la capacidad del
carro en cuanto al motor, tomando en cuenta horsepower y engine size, y la
segunda propuesta se basaria en las dimensiones del carro, tomando en cuenta
las variables car height, width y wheelbase."

#Retoma todas las variables y haz un análisis estadístico muy leve (medias y
correlación) de cómo crees que se deberían agrupar para analizarlas.

media_motor = colMeans(autop[, c("horsepower", "enginesize")], na.rm = TRUE)
print("Medias de capacidad del moto:")

## [1] "Medias de capacidad del moto:"

print(media_motor)

## horsepower enginesize
##   104.1171   126.9073

media_tamaño = colMeans(autop[, c("carheight", "carwidth", "wheelbase")],
na.rm = TRUE)
print("Medias de dimensiones del carro:")

## [1] "Medias de dimensiones del carro:"

print(media_tamaño)

## carheight  carwidth wheelbase
##  53.72488  65.90780  98.75659

cor_motor = cor(autop$horsepower, autop$enginesize, use = "complete.obs")
print("Correlación entre horsepower y enginesize:")

## [1] "Correlación entre horsepower y enginesize:"

print(cor_motor)

## [1] 0.8097687

cor_tamaño <- cor(autop[, c("carheight", "carwidth", "wheelbase")], use =
"complete.obs")
print("Correlación entre carheight, carwidth y wheelbase:")

## [1] "Correlación entre carheight, carwidth y wheelbase:"

print(cor_tamaño)

##           carheight  carwidth wheelbase
## carheight 1.0000000 0.2792103 0.5894348
## carwidth  0.2792103 1.0000000 0.7951436
## wheelbase 0.5894348 0.7951436 1.0000000
```