Actividad Integradora 2

Daniela Jiménez Téllez

2024-09-06

Problema

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. 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?

En este archivo se hará uso del segundo grupo.

Importación de datos

```
datos <- read.csv("precios_autos.csv")

# Segundo grupo

altura <- datos$carheight # Cuantitativa

ancho <- datos$carwidth # Cuantitativa

convertible <- datos$carbody # Cualitativa

precio <- datos$price

nuevo_df <- datos[, c("carheight", "carwidth", "carbody", "price")]
```

Instrucciones

1. Exploración de la base de datos

- Exploración de la base de datos
    - Calcula medidas estadísticas apropiadas para las variables:
        - cuantitativas (media, desviación estándar, cuantiles, etc)
        - cualitativas: cuantiles, frecuencias (puedes usar el comando table o prop.table)

```
# CUANTITATIVAS

# Media

mean_altura <- mean(altura)
mean_ancho <- mean(ancho)

cat("La media de la altura es:", mean_altura, "\n")
```

```
## La media de la altura es: 53.72488
```

```
cat("La media del ancho es:", mean_ancho, "\n\n")
```

```
## La media del ancho es: 65.9078
```

```
# Desviación estándar

sd_altura <- sd(altura)
sd_ancho <- sd(ancho)

cat("La desviación estándar de la altura es:", sd_altura, "\n")
```

```
## La desviación estándar de la altura es: 2.443522
```

```
cat("La desviación estándar del ancho es:", sd_ancho, "\n\n")
```

```
## La desviación estándar del ancho es: 2.145204
```

```
# Cuantiles

quan_altura <- quantile(altura, probs = c(0.25, 0.5, 0.75))
quan_ancho <- quantile(ancho, probs = c(0.25, 0.5, 0.75))

cat("Los cuantiles de la altura son: \n")
```

```
## Los cuantiles de la altura son:
```

```
print(quan_altura)
```

```
##  25%  50%  75%
## 52.0 54.1 55.5
```

```
cat("Los cuantiles del ancho son: \n")
```

```
## Los cuantiles del ancho son:
```

```
print(quan_ancho)
```

```
##  25%  50%  75%
## 64.1 65.5 66.9
```

```
cat("\n\n")
```

```
# CUALITATIVAS

# Frecuencia/ cuantiles

freq_conv <- table(convertible)

cat("La frecuencia de si es convertible o no es: \n")
```

```
## La frecuencia de si es convertible o no es:
```

```
print(freq_conv)
```

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

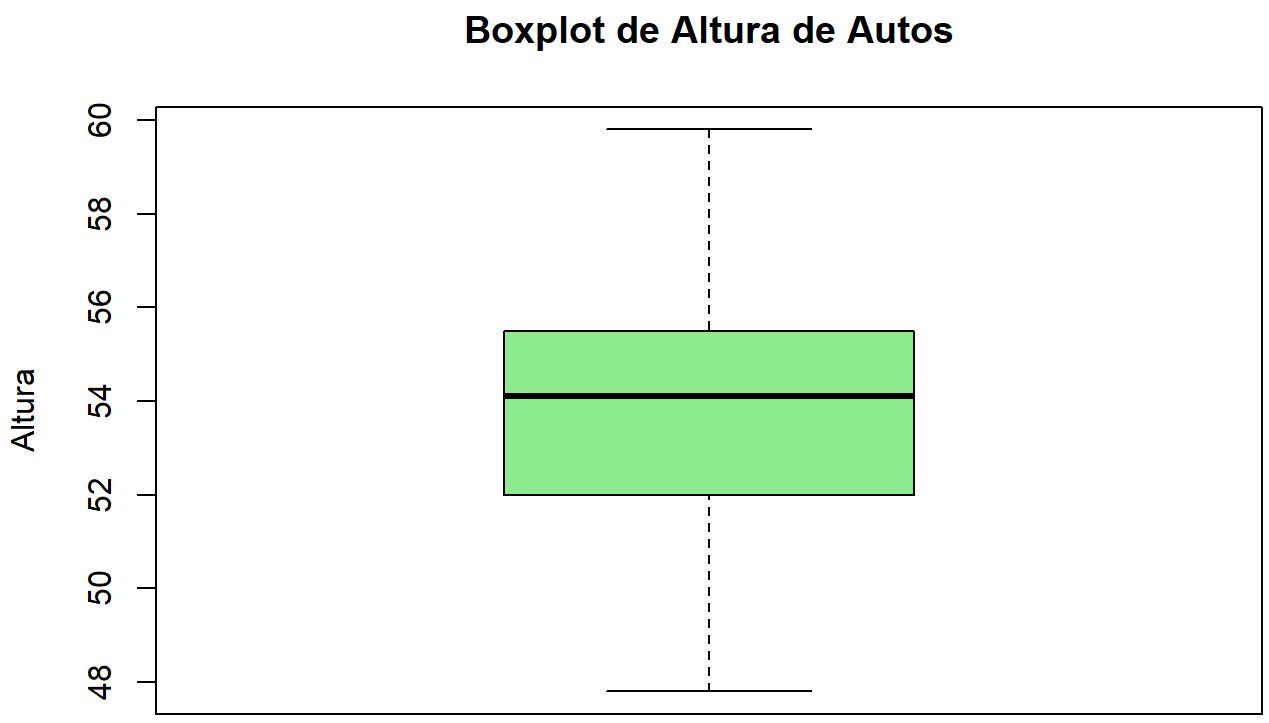
- Analiza la correlación entre las variables (analiza posible colinealidad entre las variables)

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

print(cor_matrix)
```

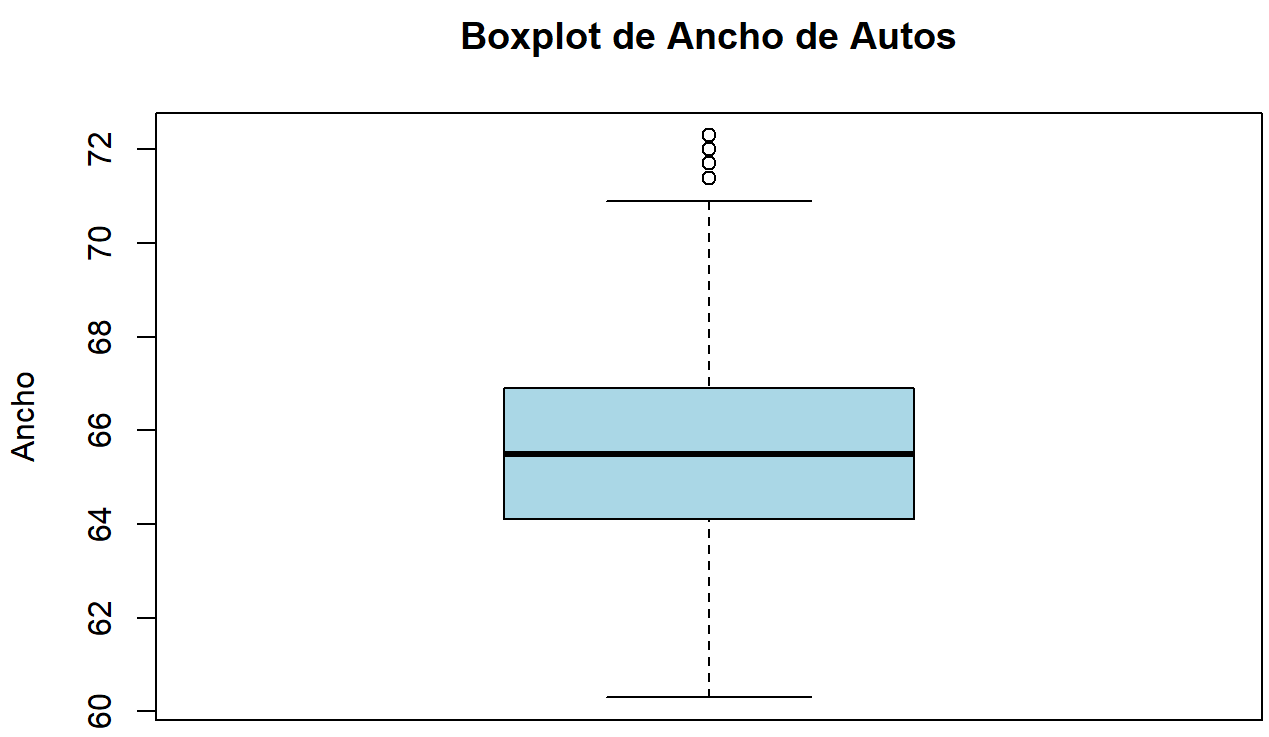
```
##           carheight  carwidth     price
## carheight 1.0000000 0.2792103 0.1193362
## carwidth  0.2792103 1.0000000 0.7593253
## price     0.1193362 0.7593253 1.0000000
```

- Explora los datos usando herramientas de visualización (si lo consideras necesario):
    - Variables cuantitativas:
        - Boxplot (visualización de datos atípicos)
        - Histogramas
        - Diagramas de dispersión y correlación por pares
    - Variables categóricas:
        - Distribución de los datos (diagramas de barras, diagramas de pastel)
        - Boxplot por categoría de las variables cuantitativas

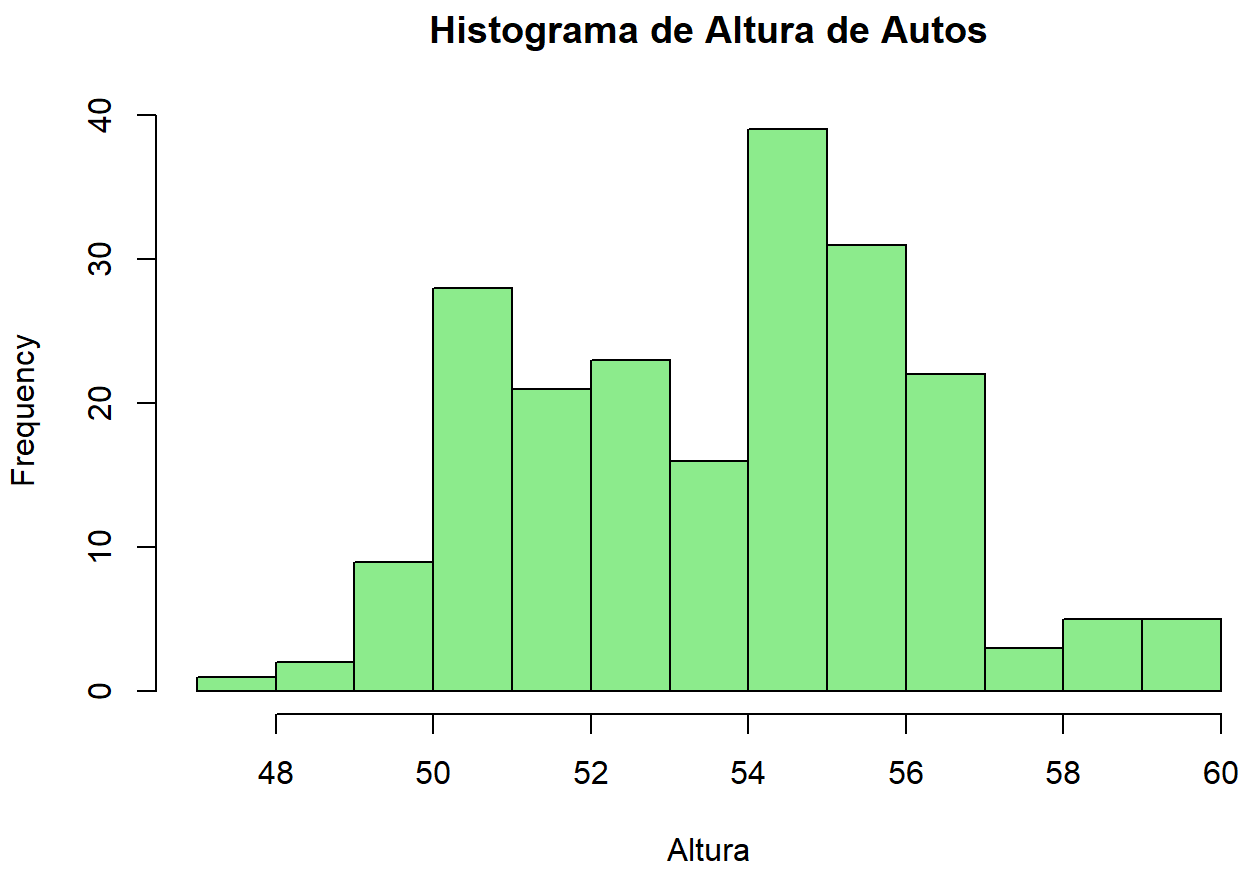
```
# CUANTITATIVAS

boxplot(altura, main = "Boxplot de Altura de Autos", col = "lightgreen", ylab = "Altura")
```



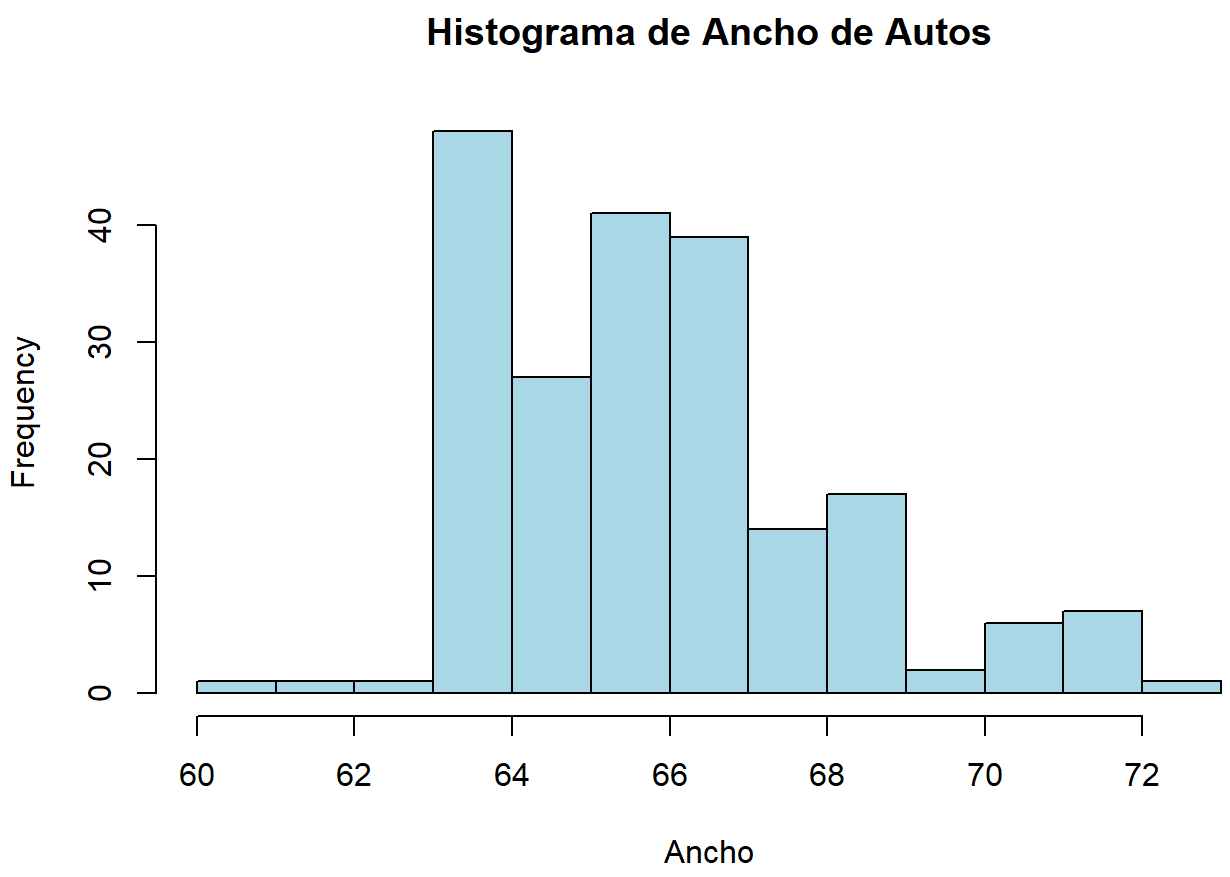

```
boxplot(ancho, main = "Boxplot de Ancho de Autos", col = "lightblue", ylab = "Ancho")
```



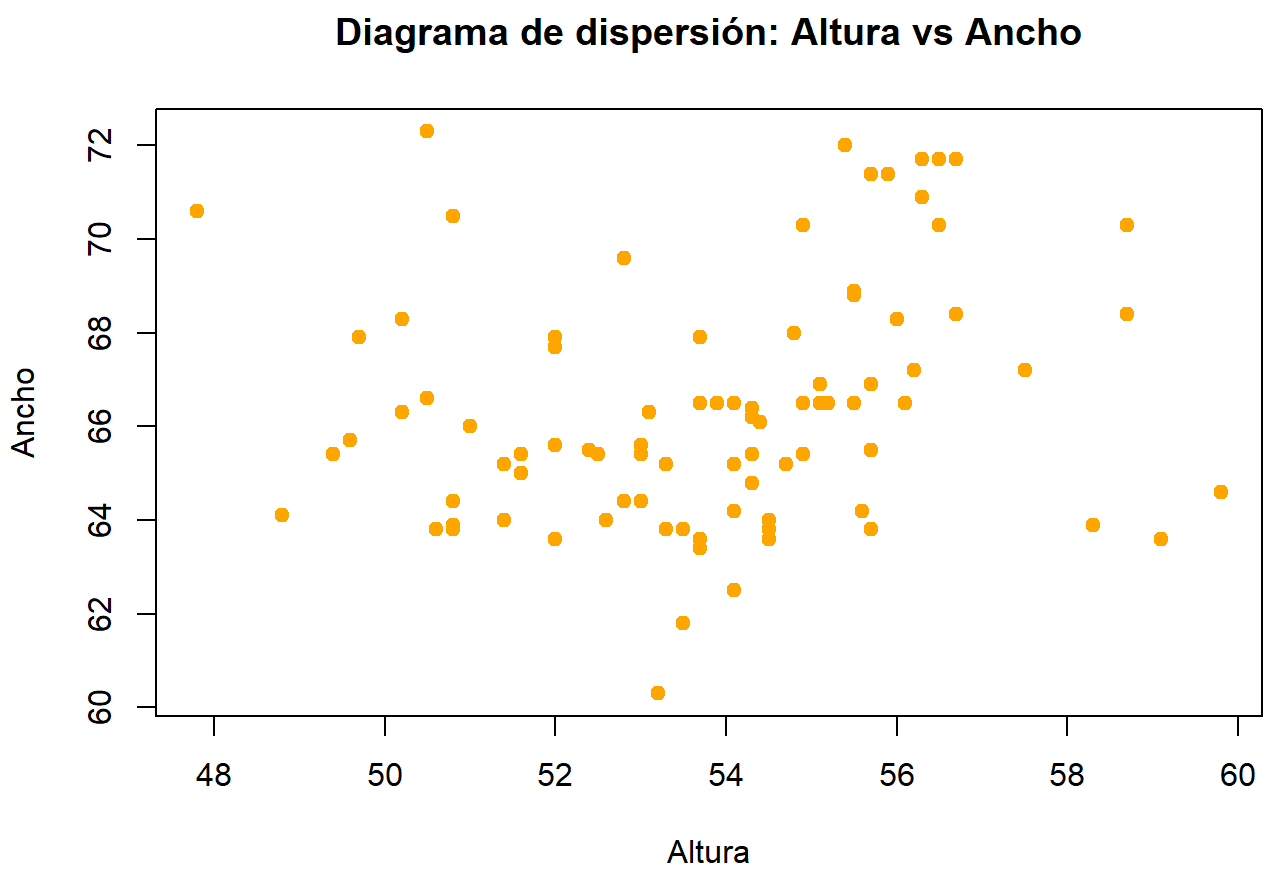

```
hist(altura, main = "Histograma de Altura de Autos", xlab = "Altura", col = "lightgreen")
```



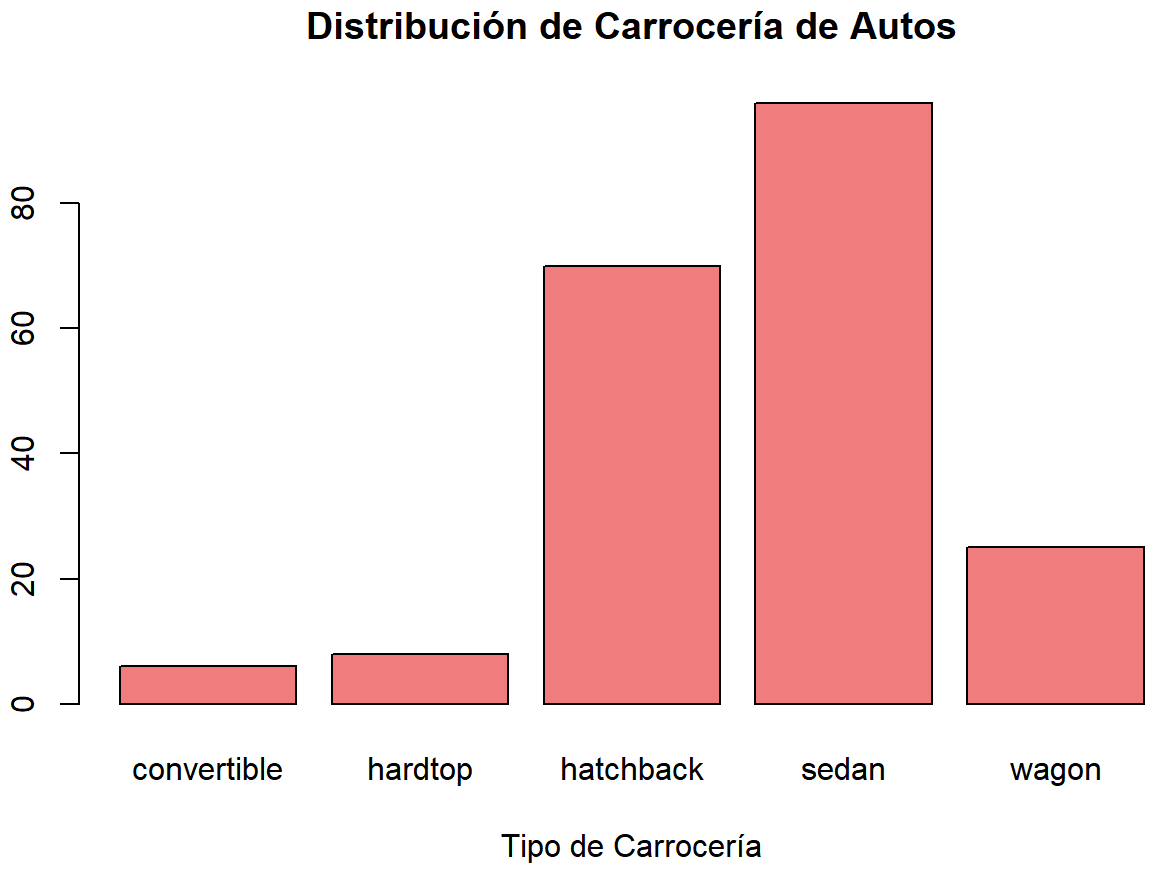

```
hist(ancho, main = "Histograma de Ancho de Autos", xlab = "Ancho", col = "lightblue")
```



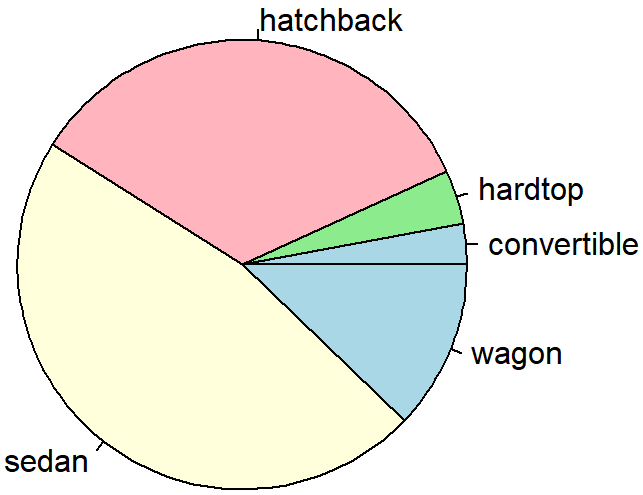

```
plot(altura, ancho, main = "Diagrama de dispersión: Altura vs Ancho", xlab = "Altura", ylab = "A
ncho", col = "orange", pch = 19)
```



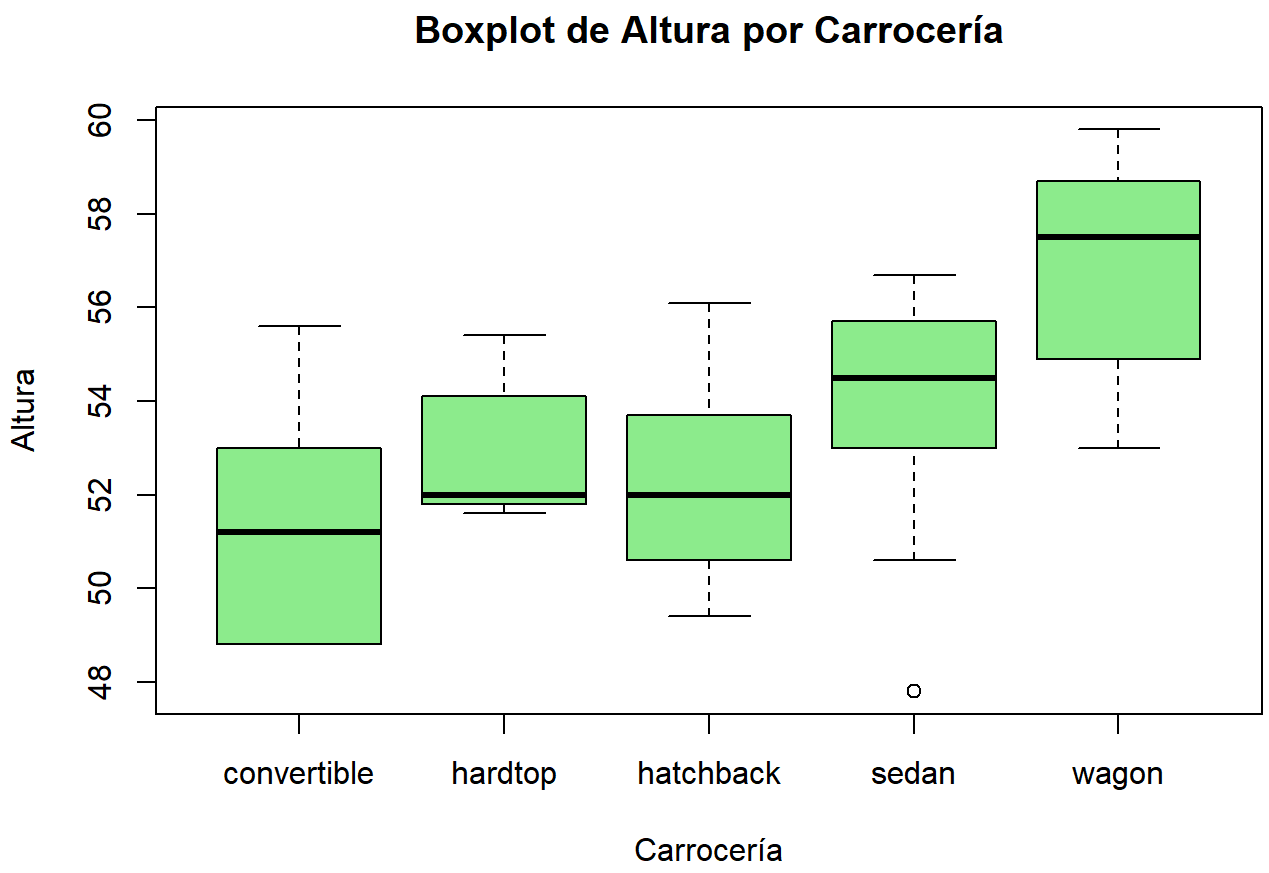

```
# CATEGÓRICAS

barplot(table(convertible), main = "Distribución de Carrocería de Autos", xlab = "Tipo de Carroc
ería", col = "lightcoral")
```



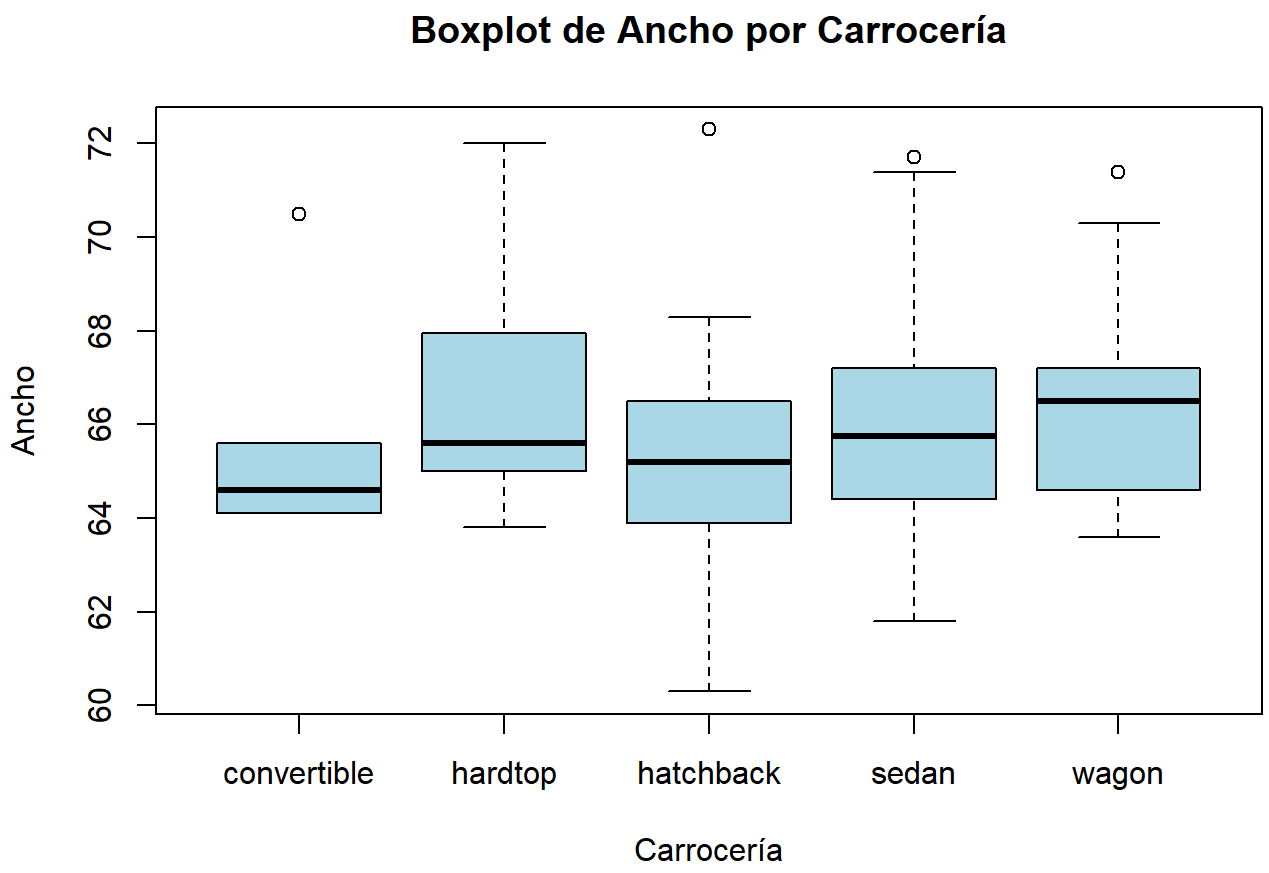

```
pie(table(convertible), main = "Distribución de Carrocería de Autos", col = c("lightblue", "ligh
tgreen", "lightpink", "lightyellow"))
```




```
boxplot(altura ~ convertible, main = "Boxplot de Altura por Carrocería", ylab = "Altura", xlab =
"Carrocería", col = "lightgreen")
```




```
boxplot(ancho ~ convertible, main = "Boxplot de Ancho por Carrocería", ylab = "Ancho", xlab = "C
arrocería", 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.

En este caso se usarán los modelos: con interacción y sin interacción.

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

```
# Modelo sin interacción

modelo_SI <- lm(precio ~ altura + ancho + convertible)

cat("Modelo sin interacción: \n")
```

```
## Modelo sin interacción:
```

```
summary(modelo_SI)
```

```
##
## Call:
## lm(formula = precio ~ altura + ancho + convertible)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11103.9  -2404.6   -657.1   1430.6  22217.3
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -150934.9    12535.9 -12.040  < 2e-16 ***
## altura                  -225.0      177.7  -1.266    0.207
## ancho                   2811.7      161.7  17.388  < 2e-16 ***
## convertiblehardtop     -2256.9     2554.2  -0.884    0.378
## convertiblehatchback  -10416.6     2005.7  -5.194 5.10e-07 ***
## convertiblesedan       -8796.5     2042.1  -4.307 2.60e-05 ***
## convertiblewagon      -10218.4     2328.8  -4.388 1.86e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4703 on 198 degrees of freedom
## Multiple R-squared:  0.6636, Adjusted R-squared:  0.6534
## F-statistic:  65.1 on 6 and 198 DF,  p-value: < 2.2e-16
```

```
cat("\n\n")
```

```
# Modelo con interacción

modelo_CI <- lm(precio ~ altura * ancho * convertible)

cat("Modelo con interacción: \n")
```

```
## Modelo con interacción:
```

```
summary(modelo_CI)
```

```
##
## Call:
## lm(formula = precio ~ altura * ancho * convertible)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11721.4  -2229.1   -534.1   1197.2  20838.2
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -9.338e+05  5.428e+06  -0.172 0.863584
## altura                           1.506e+04  1.069e+05   0.141 0.888123
## ancho                            1.517e+04  8.441e+04   0.180 0.857535
## convertiblehardtop               2.757e+07  7.165e+06   3.848 0.000164 ***
## convertiblehatchback             1.693e+04  5.461e+06   0.003 0.997530
## convertiblesedan                 2.439e+05  5.439e+06   0.045 0.964281
## convertiblewagon                 1.486e+06  5.497e+06   0.270 0.787195
## altura:ancho                    -2.413e+02  1.662e+03  -0.145 0.884724
## altura:convertiblehardtop       -5.093e+05  1.371e+05  -3.715 0.000269 ***
## altura:convertiblehatchback     -3.312e+02  1.076e+05  -0.003 0.997547
## altura:convertiblesedan         -6.049e+03  1.071e+05  -0.056 0.955024
## altura:convertiblewagon         -2.678e+04  1.080e+05  -0.248 0.804406
## ancho:convertiblehardtop        -4.133e+05  1.100e+05  -3.758 0.000229 ***
## ancho:convertiblehatchback      -9.038e+02  8.491e+04  -0.011 0.991519
## ancho:convertiblesedan          -4.038e+03  8.457e+04  -0.048 0.961970
## ancho:convertiblewagon          -2.367e+04  8.546e+04  -0.277 0.782075
## altura:ancho:convertiblehardtop  7.626e+03  2.105e+03   3.623 0.000376 ***
## altura:ancho:convertiblehatchback 1.444e+01  1.672e+03   0.009 0.993117
## altura:ancho:convertiblesedan    9.608e+01  1.665e+03   0.058 0.954046
## altura:ancho:convertiblewagon    4.244e+02  1.679e+03   0.253 0.800675
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4298 on 185 degrees of freedom
## Multiple R-squared:  0.7375, Adjusted R-squared:  0.7105
## F-statistic: 27.35 on 19 and 185 DF,  p-value: < 2.2e-16
```

- Analiza la significancia del modelo:
    - Valida la significancia del modelo con un alfa de 0.04 (incluye las hipótesis que pruebas y el valor frontera).

```
# SI

cat("PARA EL MODELO SIN INTERACCIÓN: \n")
```

```
## PARA EL MODELO SIN INTERACCIÓN:
```

```
fstat_SI <- summary(modelo_SI)$fstatistic
p_valor_SI <- pf(fstat_SI[1], fstat_SI[2], fstat_SI[3], lower.tail = FALSE)

cat("Modelo sin interacción - P-valor del Test F: ", p_valor_SI, "\n")
```

```
## Modelo sin interacción - P-valor del Test F:  3.217377e-44
```

```
if (p_valor_SI < 0.04) {cat("El modelo sin interacción es significativo con un alfa de 0.04\n
\n")} else {cat("El modelo sin interacción NO es significativo con un alfa de 0.04 \n\n")}
```

```
## El modelo sin interacción es significativo con un alfa de 0.04
```

```
# CI

cat("PARA EL MODELO CON INTERACCIÓN: \n")
```

```
## PARA EL MODELO CON INTERACCIÓN:
```

```
fstat_CI <- summary(modelo_CI)$fstatistic
p_valor_CI <- pf(fstat_CI[1], fstat_CI[2], fstat_CI[3], lower.tail = FALSE)

cat("Modelo con interacción - P-valor del Test F: ", p_valor_CI, "\n")
```

```
## Modelo con interacción - P-valor del Test F:  9.584862e-44
```

```
if (p_valor_CI < 0.04) {cat("El modelo con interacción es significativo con un alfa de 0.04\n")}
else {cat("El modelo con interacción NO es significativo con un alfa de 0.04\n")}
```

```
## El modelo con interacción es significativo con un alfa de 0.04
```

- Valida la significancia de βi con un alfa de 0.04 (incluye las hipótesis que pruebas y el valor frontera de cada una de ellas).

```
# SI

cat("PARA EL MODELO SIN INTERACCIÓN: \n")
```

```
## PARA EL MODELO SIN INTERACCIÓN:
```

```
p_valores_coef_SI <- summary(modelo_SI)$coefficients[, 4]
significativos_SI <- p_valores_coef_SI < 0.04
cat("Coeficientes significativos en el modelo sin interacción: \n")
```

```
## Coeficientes significativos en el modelo sin interacción:
```

```
print(significativos_SI)
```

```
##          (Intercept)               altura                anco
##                 TRUE                FALSE                TRUE
##   convertiblehardtop convertiblehatchback     convertiblesedan
##                FALSE                 TRUE                TRUE
##     convertiblewagon
##                 TRUE
```

```
cat("\n\n")
```

```
# CI

cat("PARA EL MODELO CON INTERACCIÓN: \n")
```

```
## PARA EL MODELO CON INTERACCIÓN:
```

```
p_valores_coef_CI <- summary(modelo_CI)$coefficients[, 4]
significativos_CI <- p_valores_coef_CI < 0.04
cat("Coeficientes significativos en el modelo con interacción: \n")
```

```
## Coeficientes significativos en el modelo con interacción:
```

```
print(significativos_CI)
```

```
##                       (Intercept)                            altura
##                             FALSE                             FALSE
##                             ancho                convertiblehardtop
##                             FALSE                              TRUE
##              convertiblehatchback                  convertiblesedan
##                             FALSE                             FALSE
##                  convertiblewagon                      altura:ancho
##                             FALSE                             FALSE
##         altura:convertiblehardtop       altura:convertiblehatchback
##                              TRUE                             FALSE
##           altura:convertiblesedan           altura:convertiblewagon
##                             FALSE                             FALSE
##          ancho:convertiblehardtop        ancho:convertiblehatchback
##                              TRUE                             FALSE
##            ancho:convertiblesedan            ancho:convertiblewagon
##                             FALSE                             FALSE
##   altura:ancho:convertiblehardtop altura:ancho:convertiblehatchback
##                              TRUE                             FALSE
##     altura:ancho:convertiblesedan     altura:ancho:convertiblewagon
##                             FALSE                             FALSE
```

- Indica cuál es el porcentaje de variación explicada por el modelo.

```
# SI

r2_ajustado_SI <- summary(modelo_SI)$adj.r.squared
cat("R² ajustado del modelo sin interacción: ", r2_ajustado_SI, "\n\n")
```

```
## R² ajustado del modelo sin interacción:  0.6534284
```

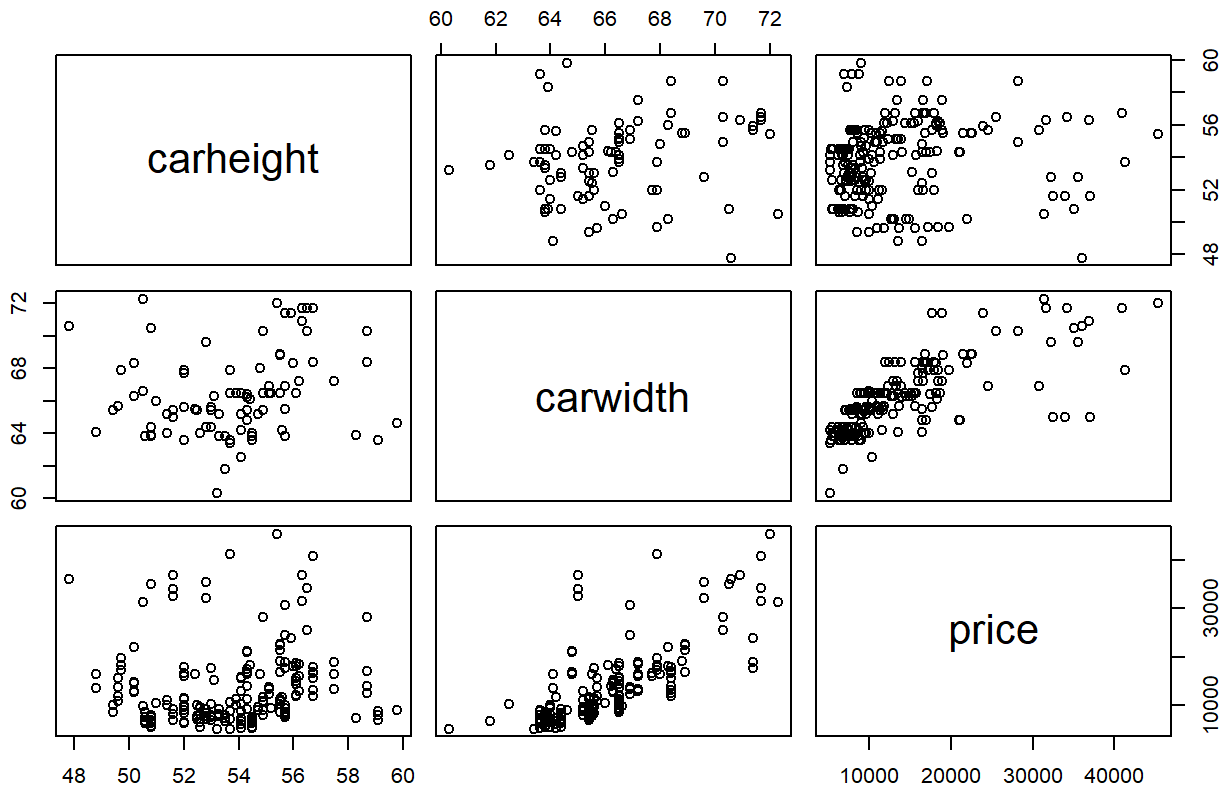
```
# CI

r2_ajustado_CI <- summary(modelo_CI)$adj.r.squared
cat("R² ajustado del modelo con interacción: ", r2_ajustado_CI, "\n")
```

```
## R² ajustado del modelo con interacción:  0.7105209
```

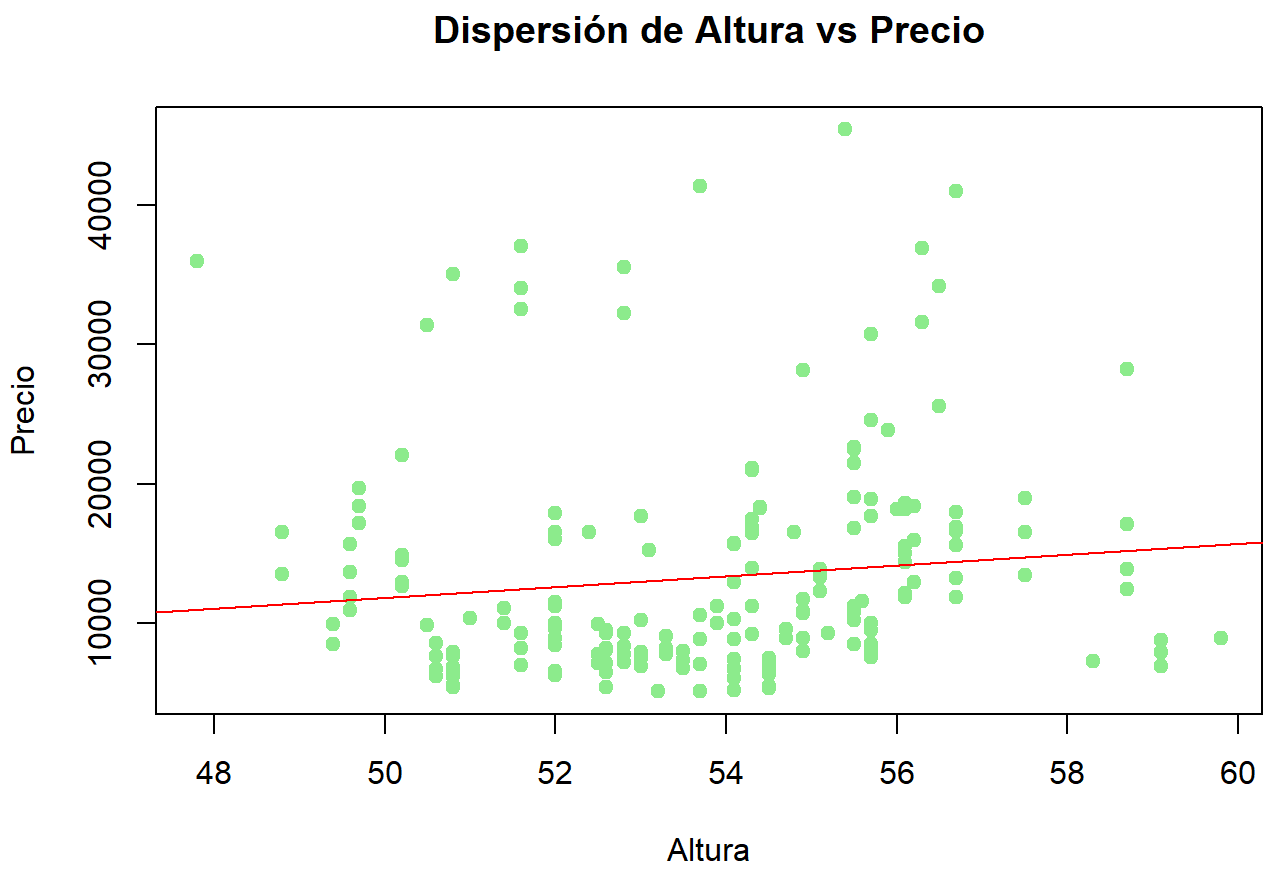
- Dibuja el diagrama de dispersión de los datos por pares y la recta de mejor ajuste.

```
library(ggplot2)
```

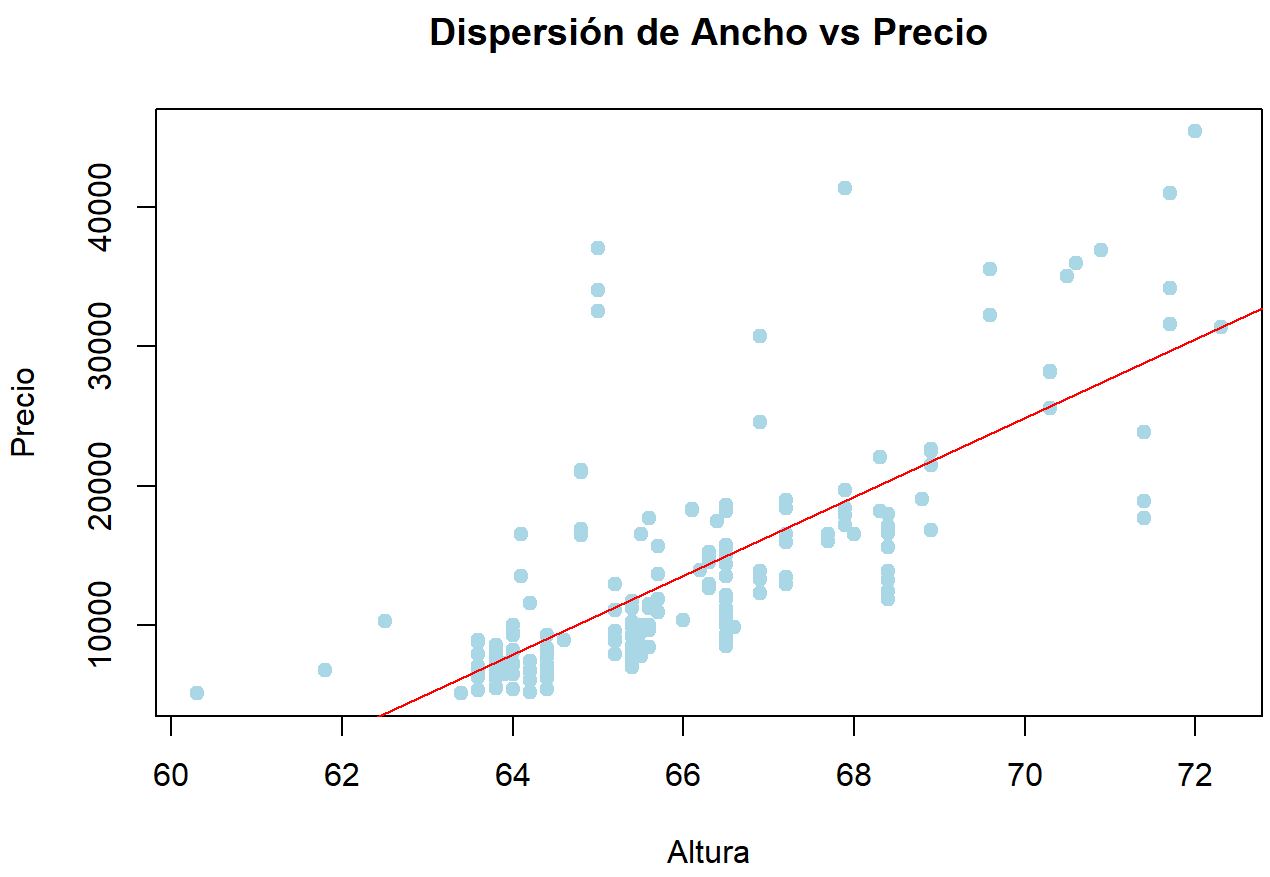
```
# Dispersión por pares

pairs(datos[, c("carheight", "carwidth", "price")], main = "Dispersión por Pares")
```



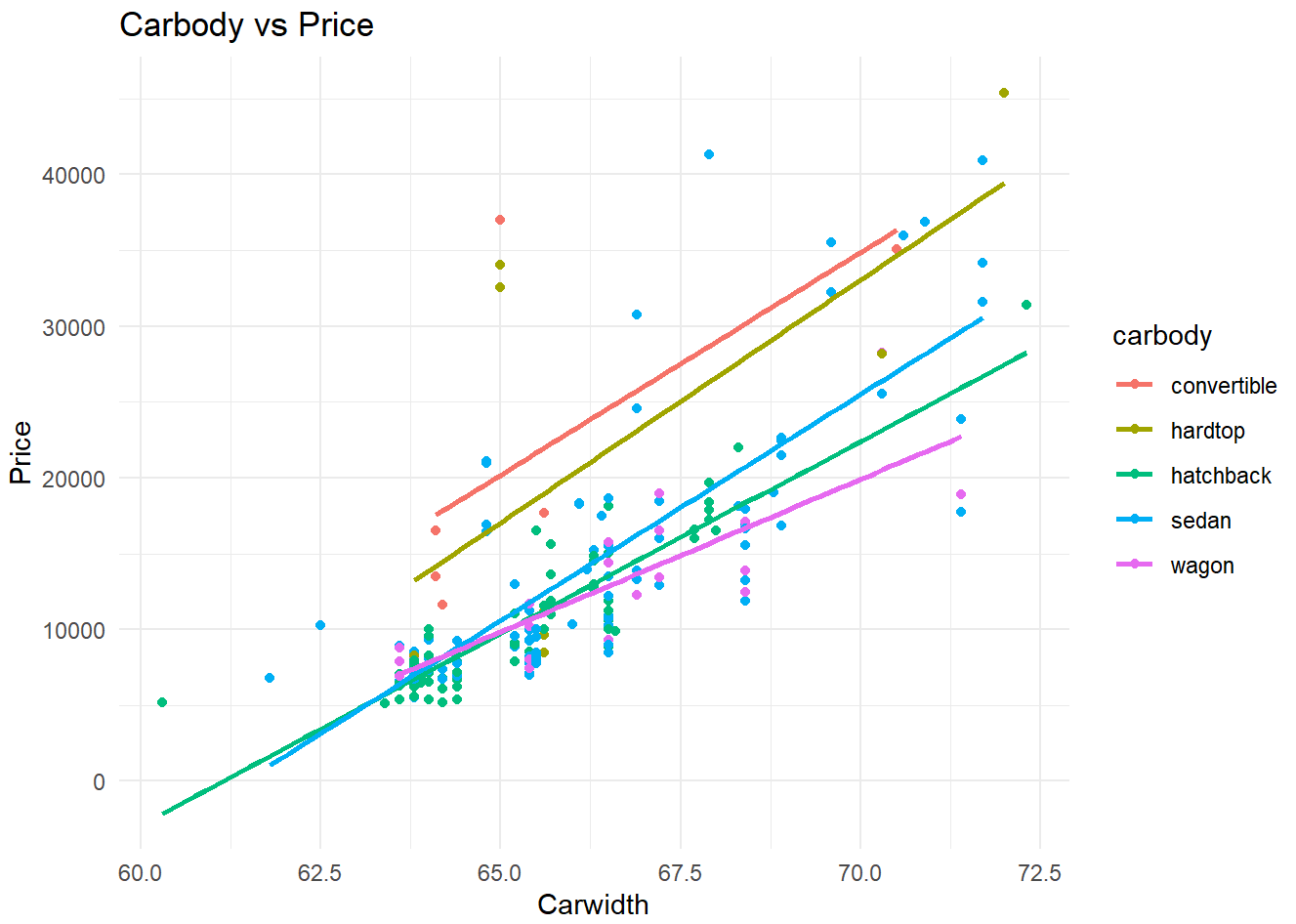

```
# Altura

plot(altura, precio, main = "Dispersión de Altura vs Precio", xlab = "Altura", ylab = "Precio",
col = "lightgreen", pch = 19)
abline(lm(precio ~ altura), col = "red")
```




```
# Ancho

plot(ancho, precio, main = "Dispersión de Ancho vs Precio", xlab = "Altura", ylab = "Precio", co
l = "lightblue", pch = 19)
abline(lm(precio ~ ancho), col = "red")
```




```
# Convertible

p3 <- ggplot(nuevo_df, aes(x = carwidth, y = price, color = carbody)) + geom_point() + geom_smoo
th(method = "lm", formula = y ~ x, se = FALSE) + labs(title = "Carbody vs Price", x = "Carwidt
h", y = "Price") + theme_minimal()

print(p3)
```



- Interpreta en el contexto del problema cada uno de los análisis que hiciste.

En estos análisis se puede observar que el modelo que mejor se va ajustando a los datos es el modelo con interacción ya que tiene un mejor R^2, además de que cumple con los demás estadísticos de prueba. Por otro lado, podemos observar cómo es que en el diagrama de dispersión por pares se nota una mayor correlación entre el ancho y el precio, a comparación de la altura y el precio.

- Analiza la validez de los modelos propuestos:
    - Normalidad de los residuos
    - Verificación de media cero
    - Homocedasticidad, linealidad e independencia
    - Interpreta cada uno de los analisis que realizaste

```
library(lmtest)
```

```
## Cargando paquete requerido: zoo
```

```
##
## Adjuntando el paquete: 'zoo'
```

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

```
# SIN INTERACCIÓN

# Normalidad

normSI <- shapiro.test(resid(modelo_SI))
cat("Los resultados de la normalidad para el modelo sin interacción usando el test de Shapiro-Wi
lk son: \n")
```

```
## Los resultados de la normalidad para el modelo sin interacción usando el test de Shapiro-Wilk
son:
```

```
print(normSI)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(modelo_SI)
## W = 0.89026, p-value = 4.299e-11
```

```
cat("\n\n")
```

```
# Media 0

medSI <- t.test(resid(modelo_SI))
cat("El resultado del test de media cero para el modelo sin interacción es: \n")
```

```
## El resultado del test de media cero para el modelo sin interacción es:
```

```
print(medSI)
```

```
##
##  One Sample t-test
##
## data:  resid(modelo_SI)
## t = 1.1555e-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
## 3.739394e-14
```

```
cat("\n\n")
```

```
# Homocedasticidad, linealidad e independencia

hom1SI <- bptest(modelo_SI)
hom2SI <- gqtest(modelo_SI)

cat("Los resultados de los tests de homocedasticidad para el modelo sin interacción son: \n")
```

```
## Los resultados de los tests de homocedasticidad para el modelo sin interacción son:
```

```
print(hom1SI)
```

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

```
print(hom2SI)
```

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

```
cat("\n\n")
```

```
ind1SI <- dwtest(modelo_SI)
ind2SI <- bgtest(modelo_SI)

cat("Los resultados de los tests de independencia para el modelo sin interacción son: \n")
```

```
## Los resultados de los tests de independencia para el modelo sin interacción son:
```

```
print(ind1SI)
```

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

```
print(ind2SI)
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  modelo_SI
## LM test = 80.703, df = 1, p-value < 2.2e-16
```

```
cat("\n\n")
```

```
linSI <- resettest(modelo_SI)

cat("Los resultados del test de linealidad para el modelo sin interacción son: \n")
```

```
## Los resultados del test de linealidad para el modelo sin interacción son:
```

```
print(linSI)
```

```
##
##  RESET test
##
## data:  modelo_SI
## RESET = 6.8362, df1 = 2, df2 = 196, p-value = 0.001349
```

```
cat("\n\n")
```

```
cat("------------------------------------------------------------------------------ \n\n")
```

```
## ------------------------------------------------------------------------------
```

```
cat("CON INTERACCIÓN \n\n")
```

```
## CON INTERACCIÓN
```

```
cat("------------------------------------------------------------------------------ \n\n")
```

```
## ------------------------------------------------------------------------------
```

```
# CON INTERACCIÓN

# Normalidad

normCI <- shapiro.test(resid(modelo_CI))
cat("Los resultados de la normalidad para el modelo con interacción usando el test de Shapiro-Wi
lk son: \n")
```

```
## Los resultados de la normalidad para el modelo con interacción usando el test de Shapiro-Wilk
son:
```

```
print(normCI)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(modelo_CI)
## W = 0.88107, p-value = 1.237e-11
```

```
cat("\n\n")
```

```
# Media 0

medCI <- t.test(resid(modelo_CI))
cat("El resultado del test de media cero para el modelo con interacción es: \n")
```

```
## El resultado del test de media cero para el modelo con interacción es:
```

```
print(medCI)
```

```
##
##  One Sample t-test
##
## data:  resid(modelo_CI)
## t = 6.1354e-16, df = 204, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -563.6624  563.6624
## sample estimates:
##    mean of x
## 1.753995e-13
```

```
# Homocedasticidad, linealidad e independencia

hom1CI <- bptest(modelo_CI)
hom2CI <- gqtest(modelo_CI)

cat("Los resultados de los tests de homocedasticidad para el modelo sin interacción son: \n")
```

```
## Los resultados de los tests de homocedasticidad para el modelo sin interacción son:
```

```
print(hom1CI)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  modelo_CI
## BP = 26.748, df = 19, p-value = 0.1107
```

```
print(hom2CI)
```

```
##
##  Goldfeld-Quandt test
##
## data:  modelo_CI
## GQ = 0.22582, df1 = 83, df2 = 82, p-value = 1
## alternative hypothesis: variance increases from segment 1 to 2
```

```
cat("\n\n")
```

```
ind1CI <- dwtest(modelo_CI)
ind2CI <- bgtest(modelo_CI)

cat("Los resultados de los tests de independencia para el modelo sin interacción son: \n")
```

```
## Los resultados de los tests de independencia para el modelo sin interacción son:
```

```
print(ind1CI)
```

```
##
##  Durbin-Watson test
##
## data:  modelo_CI
## DW = 0.96514, p-value = 2.164e-15
## alternative hypothesis: true autocorrelation is greater than 0
```

```
print(ind2CI)
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  modelo_CI
## LM test = 69.113, df = 1, p-value < 2.2e-16
```

```
cat("\n\n")
```

```
linCI <- resettest(modelo_CI)

cat("Los resultados del test de linealidad para el modelo con interacción son: \n")
```

```
## Los resultados del test de linealidad para el modelo con interacción son:
```

```
print(linCI)
```

```
##
##  RESET test
##
## data:  modelo_CI
## RESET = 15.416, df1 = 2, df2 = 183, p-value = 6.494e-07
```

- 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
    - ¿Cuáles de las variables asignadas influyen en el precio del auto? ¿de qué manera lo hacen?

El mejor modelo de este análisis es el modelo con interacción, ya que basándonos en diferentes métricas, como lo es R^2, podemos observar que el ajuste es mejor por su valor de 0.7105209 a comparación del 0.6534284 del modelo sin interacción. Igualmente, aunque se pudieron observar algunos problemas en el test de residuos, el modelo logra capturar relaciones más difíciles que hay entre los datos.

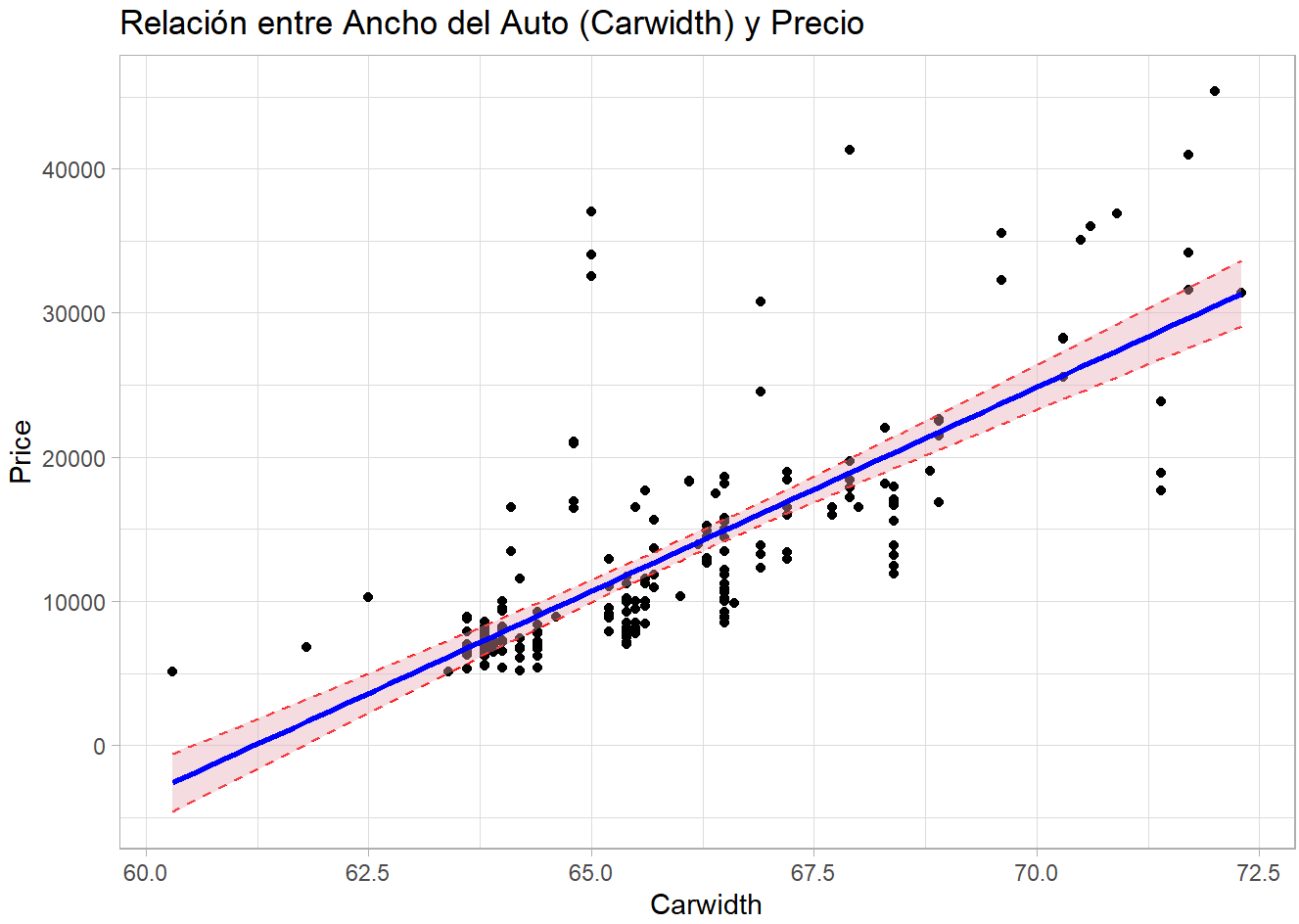
Igualmente, observando la gráfica de pares, una de las variables sí influye en el precio, y otra no tanto. En el caso del ancho, medida que este aumenta, también tiende a aumentar el precio. Por otro lado, en el caso del alto, podemos observar que hay más dispersión en los datos, lo que nos dice que la altura no tiene un impacto directo y lineal sobre el precio del auto.

3. Intervalos de predicción y confianza

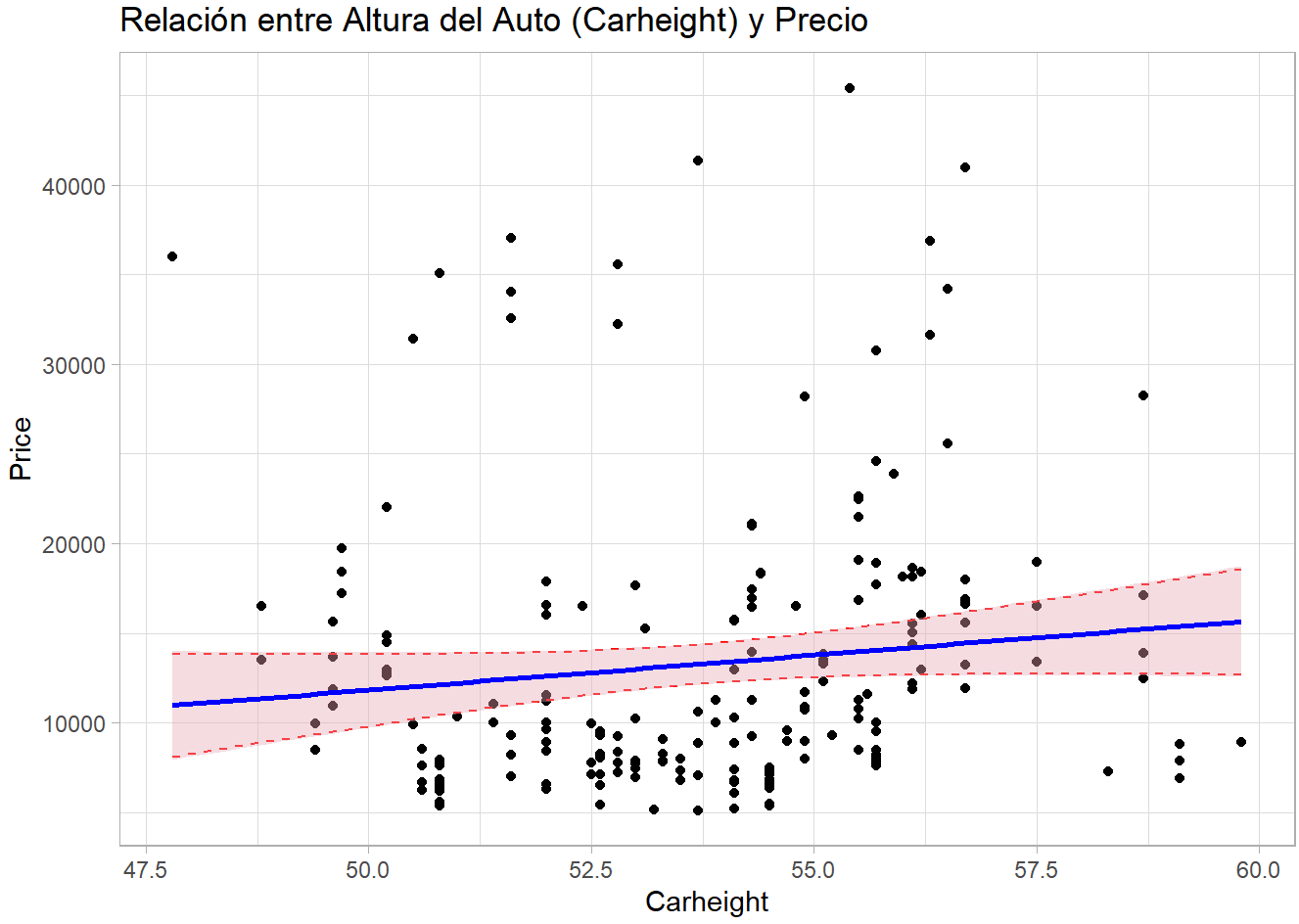
- Con los datos de las variables asignadas construye la gráfica de los intervalos de confianza y predicción para la estimación y predicción del precio para el mejor modelo seleccionado:
    - Calcula los intervalos para la variable Y
    - Selecciona la categoría de la variable cualitativa que, de acuerdo a tu análisis resulte la más importante, y separa la base de datos por esa variable categórica.

- Grafica por pares de variables numéricas

```
# Carwidth

modelo_carwidth <- lm(price ~ carwidth, data = nuevo_df)
new_data <- data.frame(carwidth = nuevo_df$carwidth)
predicciones <- predict(modelo_carwidth, newdata = new_data, interval = "confidence")

nuevo_df$fit <- predicciones[, "fit"]
nuevo_df$lwr <- predicciones[, "lwr"]
nuevo_df$upr <- predicciones[, "upr"]

ggplot(nuevo_df, aes(x = carwidth, y = price)) +
  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 = "pin
k2") +
  theme_light() +
  labs(title = "Relación entre Ancho del Auto (Carwidth) y Precio", x = "Carwidth", y = "Price")
```



```
# Carheight

modelo_carheight <- lm(price ~ carheight, data = nuevo_df)
new_data <- data.frame(carheight = nuevo_df$carheight)
predicciones <- predict(modelo_carheight, newdata = new_data, interval = "confidence")

nuevo_df$fit <- predicciones[, "fit"]
nuevo_df$lwr <- predicciones[, "lwr"]
nuevo_df$upr <- predicciones[, "upr"]

ggplot(nuevo_df, aes(x = carheight, y = price)) +
  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 = "pin
k2") +
  theme_light() +
  labs(title = "Relación entre Altura del Auto (Carheight) y Precio", x = "Carheight", y = "Pric
e")
```



```
# Carbody y sus diferentes categorías

carbody_sedan <- subset(nuevo_df, carbody == "sedan")
carbody_hatchback <- subset(nuevo_df, carbody == "hatchback")
carbody_wagon <- subset(nuevo_df, carbody == "wagon")
carbody_convertible <- subset(nuevo_df, carbody == "convertible")
carbody_hardtop <- subset(nuevo_df, carbody == "hardtop")

modelo_sedan <- lm(price ~ carwidth, data = carbody_sedan)
modelo_hatchback <- lm(price ~ carwidth, data = carbody_hatchback)
modelo_wagon <- lm(price ~ carwidth, data = carbody_wagon)
modelo_convertible <- lm(price ~ carwidth, data = carbody_convertible)
modelo_hardtop <- lm(price ~ carwidth, data = carbody_hardtop)

pred_sedan <- predict(modelo_sedan, interval = "prediction", level = 0.97)
```

```
## Warning in predict.lm(modelo_sedan, interval = "prediction", level = 0.97): predictions on cu
rrent data refer to _future_ responses
```

```
pred_hatchback <- predict(modelo_hatchback, interval = "prediction", level = 0.97)
```

```
## Warning in predict.lm(modelo_hatchback, interval = "prediction", level = 0.97): predictions o
n current data refer to _future_ responses
```

```
pred_wagon <- predict(modelo_wagon, interval = "prediction", level = 0.97)
```

```
## Warning in predict.lm(modelo_wagon, interval = "prediction", level = 0.97): predictions on cu
rrent data refer to _future_ responses
```

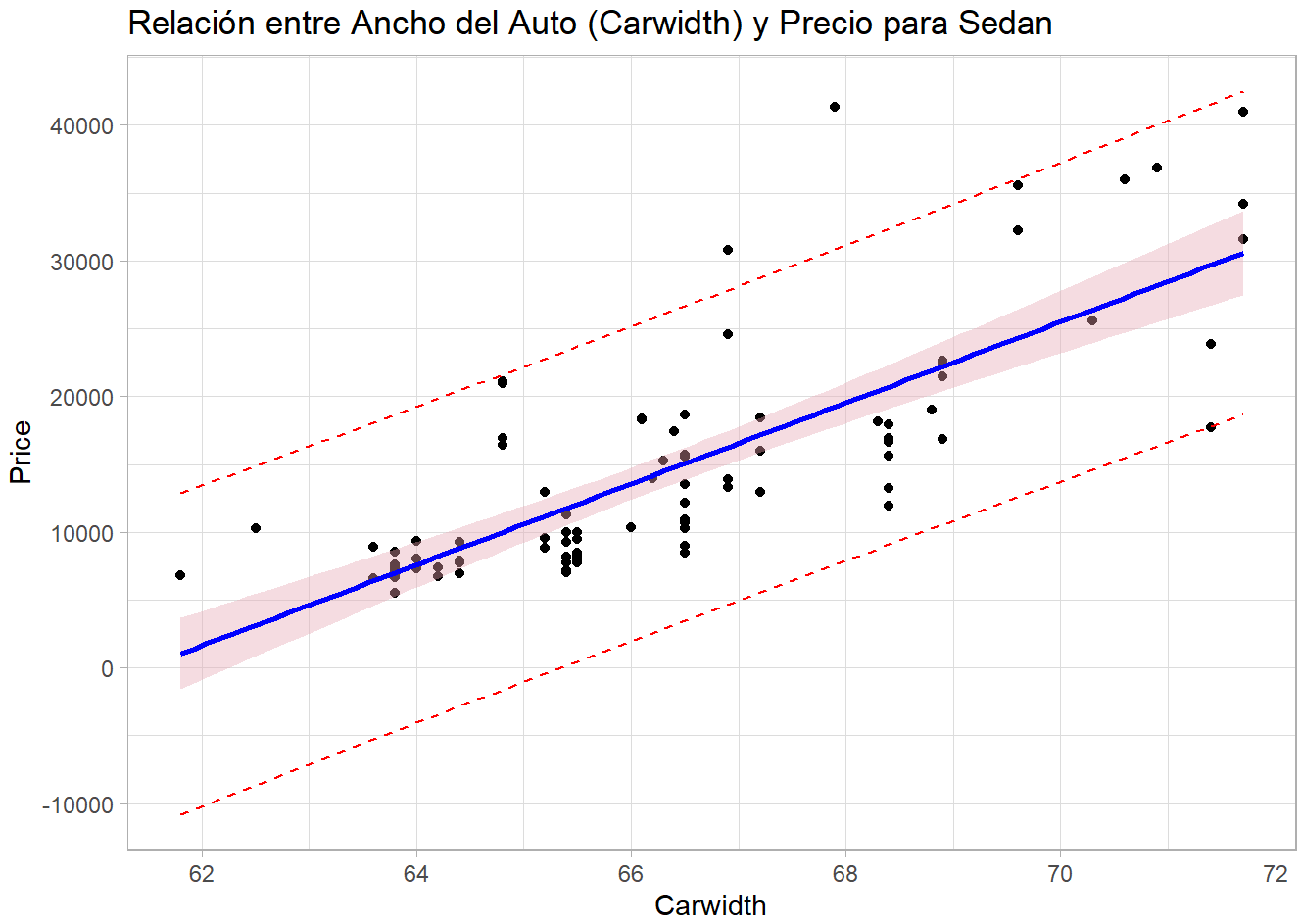
```
pred_convertible <- predict(modelo_convertible, interval = "prediction", level = 0.97)
```

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

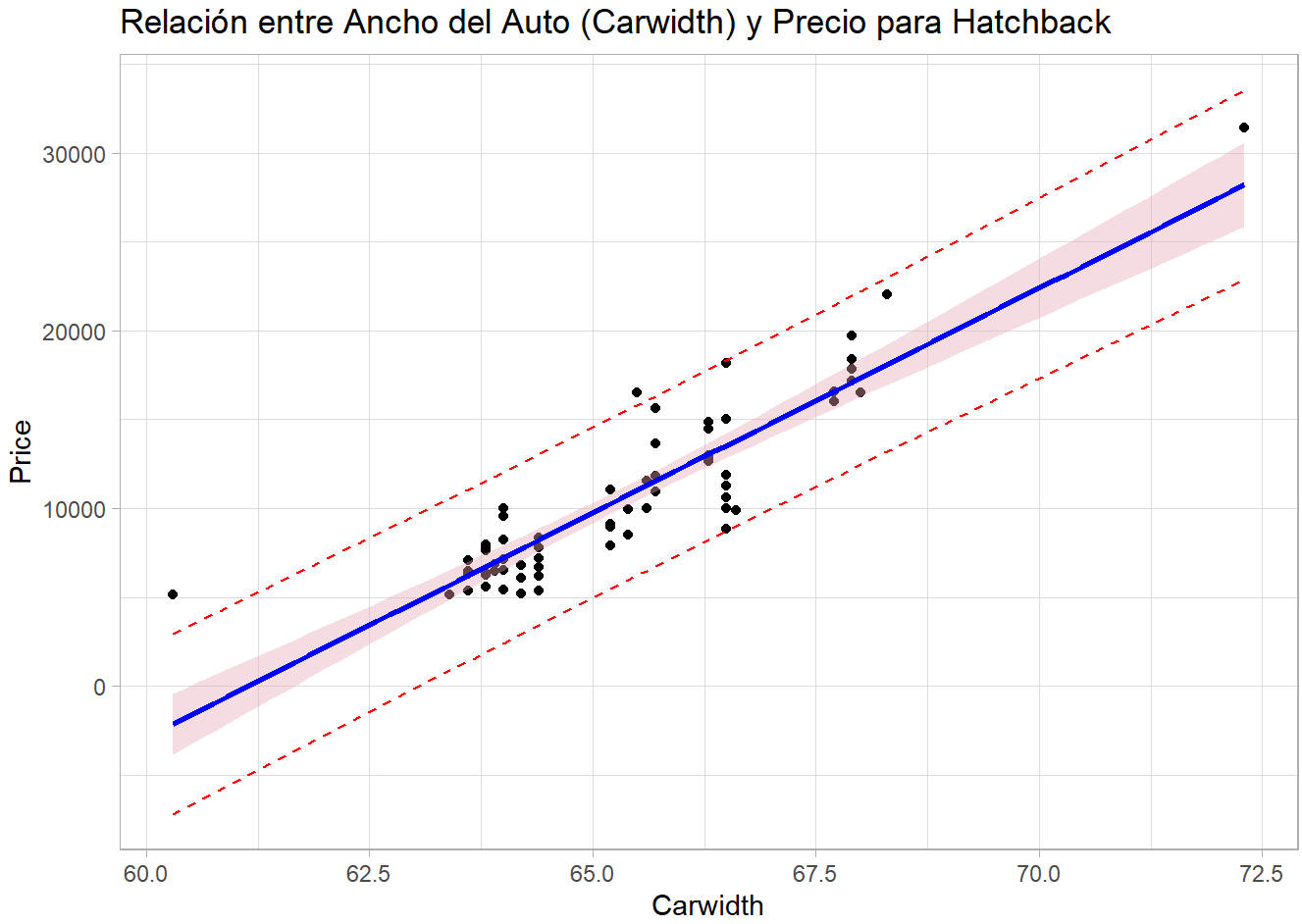
```
pred_hardtop <- predict(modelo_hardtop, interval = "prediction", level = 0.97)
```

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

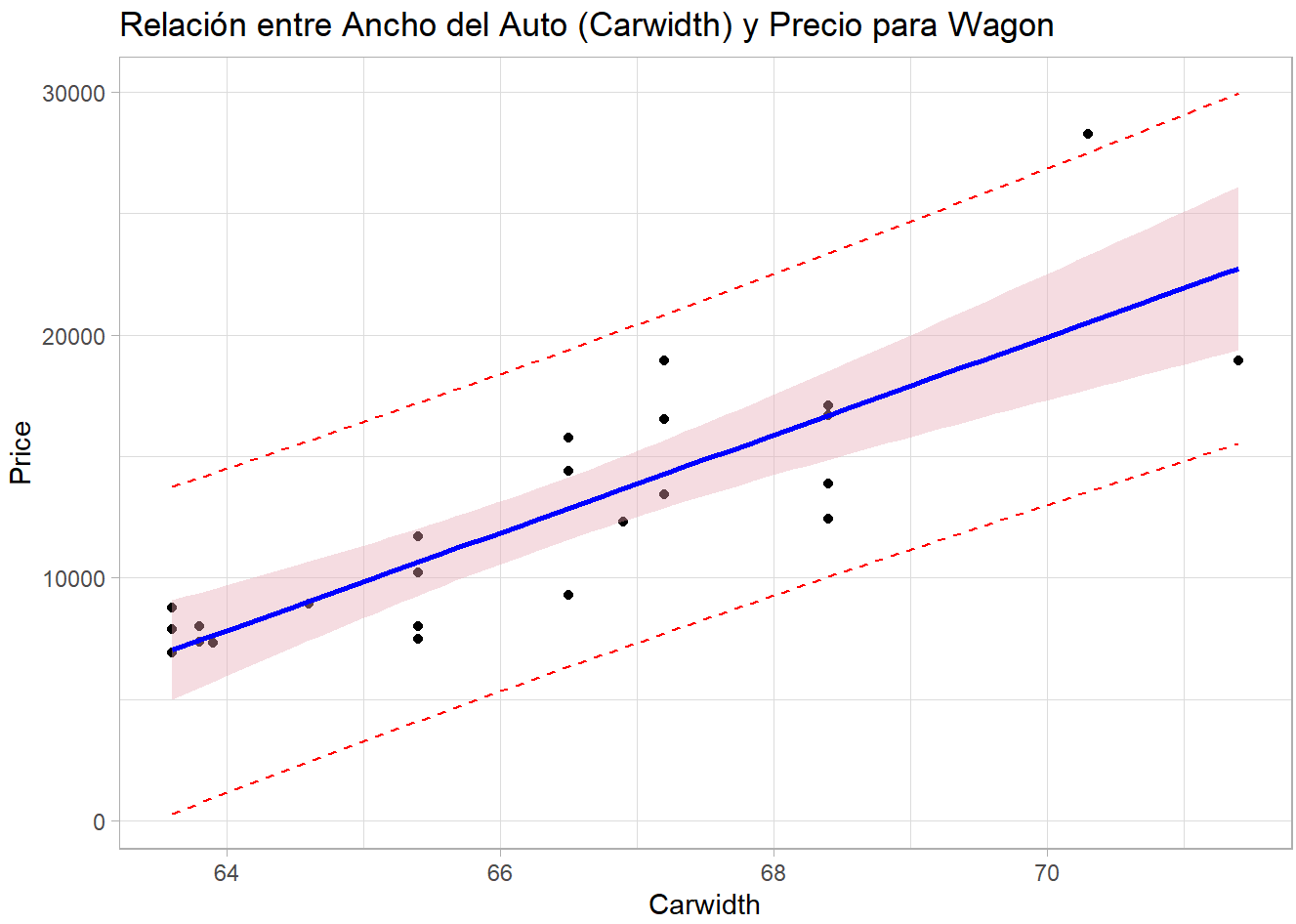
```
carbody_sedan <- carbody_sedan[ , !names(carbody_sedan) %in% c("fit", "lwr", "upr")]
carbody_hatchback <- carbody_hatchback[ , !names(carbody_hatchback) %in% c("fit", "lwr", "upr")]
carbody_wagon <- carbody_wagon[ , !names(carbody_wagon) %in% c("fit", "lwr", "upr")]
carbody_convertible <- carbody_convertible[ , !names(carbody_convertible) %in% c("fit", "lwr",
"upr")]
carbody_hardtop <- carbody_hardtop[ , !names(carbody_hardtop) %in% c("fit", "lwr", "upr")]

carbody_sedan <- cbind(carbody_sedan, pred_sedan)
carbody_hatchback <- cbind(carbody_hatchback, pred_hatchback)
carbody_wagon <- cbind(carbody_wagon, pred_wagon)
carbody_convertible <- cbind(carbody_convertible, pred_convertible)
carbody_hardtop <- cbind(carbody_hardtop, pred_hardtop)

ggplot(carbody_sedan, aes(x = carwidth, y = price)) +
  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 = 0.97, col = "blue", fill = "pin
k2") +
  theme_light() +
  labs(title = "Relación entre Ancho del Auto (Carwidth) y Precio para Sedan", x = "Carwidth", y
= "Price")
```



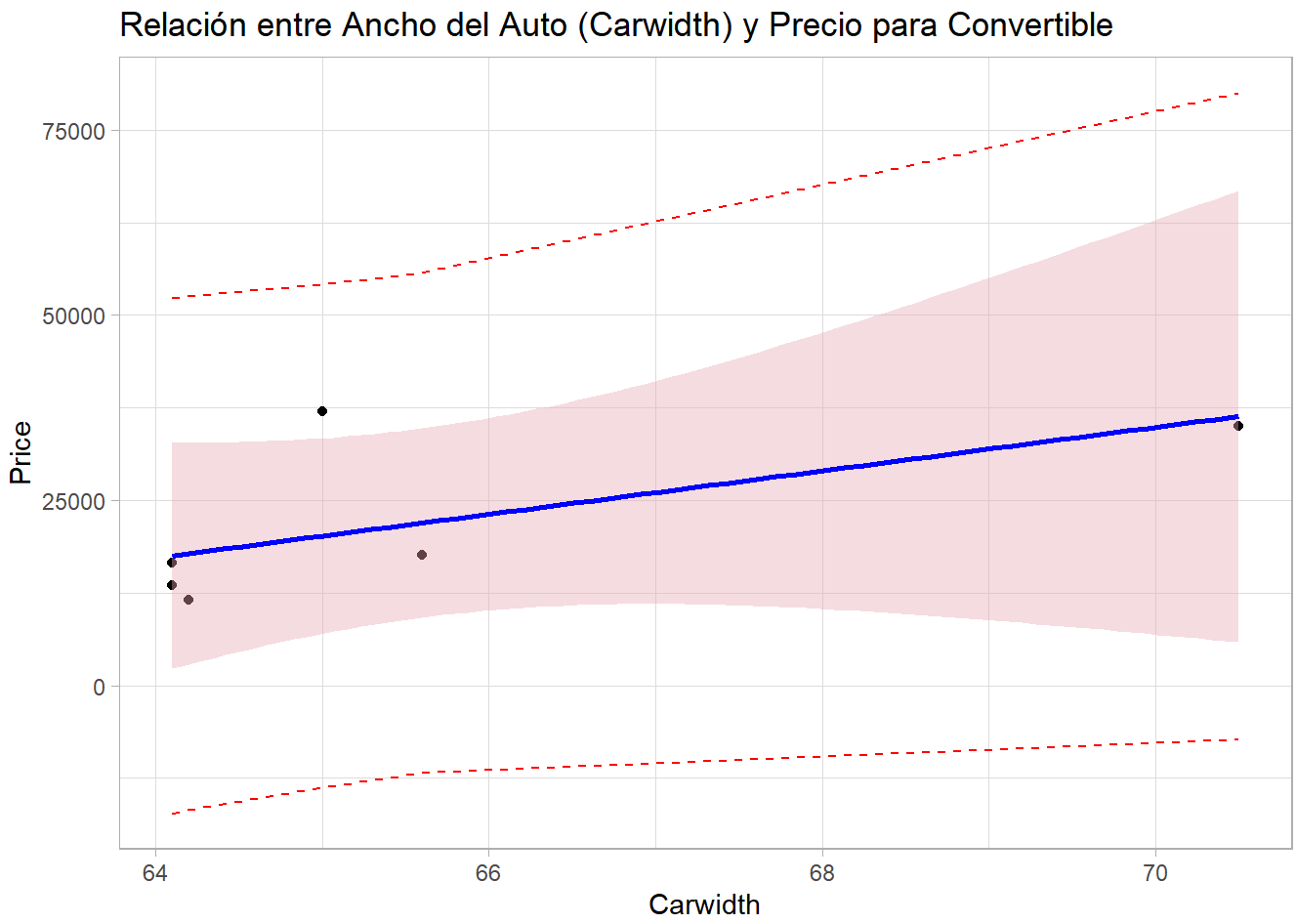
```
ggplot(carbody_hatchback, aes(x = carwidth, y = price)) +
  geom_point() +
  geom_line(aes(y = lwr), color = "red", linetype = "dashed") +
  geom_line(aes(y = upr), color = "red", linetype = "dashed") +
  geom_smooth(method = "lm", formula = y ~ x, se = TRUE, level = 0.97, col = "blue", fill = "pin
k2") +
  theme_light() +
  labs(title = "Relación entre Ancho del Auto (Carwidth) y Precio para Hatchback", x = "Carwidt
h", y = "Price")
```



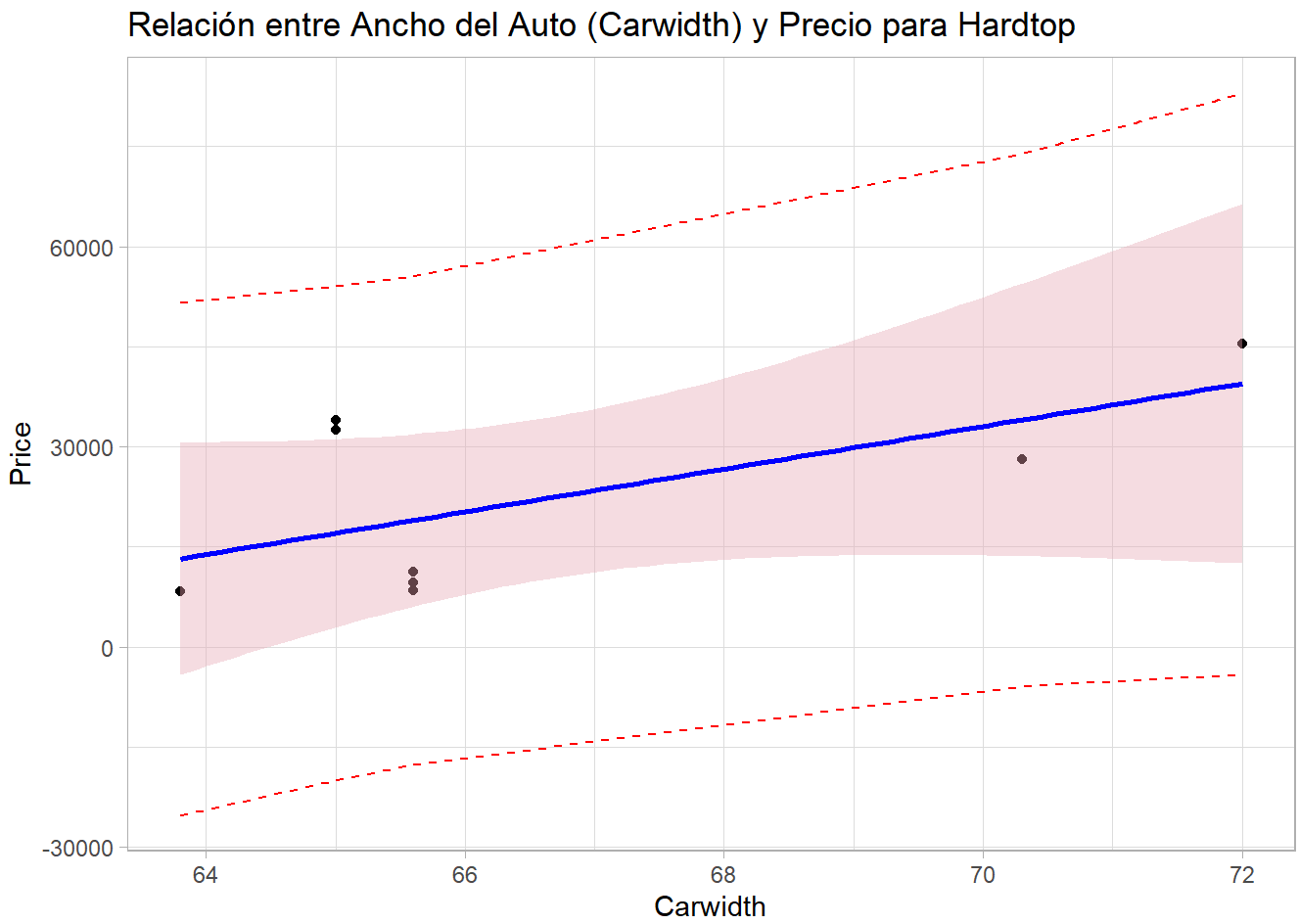

```
ggplot(carbody_wagon, aes(x = carwidth, y = price)) +
  geom_point() +
  geom_line(aes(y = lwr), color = "red", linetype = "dashed") +
  geom_line(aes(y = upr), color = "red", linetype = "dashed") +
  geom_smooth(method = "lm", formula = y ~ x, se = TRUE, level = 0.97, col = "blue", fill = "pin
k2") +
  theme_light() +
  labs(title = "Relación entre Ancho del Auto (Carwidth) y Precio para Wagon", x = "Carwidth", y
= "Price")
```




```
ggplot(carbody_convertible, aes(x = carwidth, y = price)) +
  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 = 0.97, col = "blue", fill = "pin
k2") +
  theme_light() +
  labs(title = "Relación entre Ancho del Auto (Carwidth) y Precio para Convertible", x = "Carwid
th", y = "Price")
```




```
ggplot(carbody_hardtop, aes(x = carwidth, y = price)) +
  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 = 0.97, col = "blue", fill = "pin
k2") +
  theme_light() +
  labs(title = "Relación entre Ancho del Auto (Carwidth) y Precio para Hardtop", x = "Carwidth",
y = "Price")
```



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?

Sí propondría una nueva agrupación de variables. Observando el dataset, hay ciertas características que van muy bien juntas, como lo son:

1. Wheelbase (Distancia entre los ejes), Carheight (Altura del auto), Carwidth (Ancho del auto), Carlength (Longitud del auto), las cuales podrían describirse como las dimensiones del auto.
2. Enginesize (Tamaño del motor), Stroke, Horsepower (Caballos de fuerza), Compressionratio (Relación de compresión), Peakrpm (Revoluciones máximas), Enginetype (Tipo de motor), Cylindernumber (Número de cilindros), los cuales se describen como las características del motor.
3. Citympg (Consumo en ciudad), Highwaympg (Consumo en carretera), Fueltype (Tipo de gasolina), Drivewheel (Tracción), los cuales pueden llamarse como el rendimiento del auto.

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

```
grupo1 <- c("wheelbase", "carheight", "carwidth", "carlength")
grupo2 <- c("enginesize", "stroke", "horsepower", "compressionratio", "peakrpm")
grupo3 <- c("citympg", "highwaympg")

# Medias

mean_grupo1 <- colMeans(datos[, grupo1], na.rm = TRUE)
mean_grupo2 <- colMeans(datos[, grupo2], na.rm = TRUE)
mean_grupo3 <- colMeans(datos[, grupo3], na.rm = TRUE)

cat("Media del grupo 1: ", mean_grupo1, "\n")
```

```
## Media del grupo 1:  98.75659 53.72488 65.9078 174.0493
```

```
cat("Media del grupo 2: ", mean_grupo2, "\n")
```

```
## Media del grupo 2:  126.9073 3.255415 104.1171 10.14254 5125.122
```

```
cat("Media del grupo 3: ", mean_grupo3, "\n\n")
```

```
## Media del grupo 3:  25.21951 30.75122
```

```
# Matriz de correlación

correlacion_grupo1 <- cor(datos[, grupo1], use = "complete.obs")
correlacion_grupo2 <- cor(datos[, grupo2], use = "complete.obs")
correlacion_grupo3 <- cor(datos[, grupo3], use = "complete.obs")

cat("Matriz de correlación del grupo 1: \n")
```

```
## Matriz de correlación del grupo 1:
```

```
print(correlacion_grupo1)
```

```
##           wheelbase carheight  carwidth carlength
## wheelbase 1.0000000 0.5894348 0.7951436 0.8745875
## carheight 0.5894348 1.0000000 0.2792103 0.4910295
## carwidth  0.7951436 0.2792103 1.0000000 0.8411183
## carlength 0.8745875 0.4910295 0.8411183 1.0000000
```

```
cat("\n\n")
```

```
cat("Matriz de correlación del grupo 2: \n")
```

```
## Matriz de correlación del grupo 2:
```

```
print(correlacion_grupo2)
```

```
##                   enginesize      stroke  horsepower compressionratio
## enginesize        1.00000000  0.20312859  0.80976865       0.02897136
## stroke            0.20312859  1.00000000  0.08093954       0.18611011
## horsepower        0.80976865  0.08093954  1.00000000      -0.20432623
## compressionratio  0.02897136  0.18611011 -0.20432623       1.00000000
## peakrpm          -0.24465983 -0.06796375  0.13107251      -0.43574051
##                      peakrpm
## enginesize       -0.24465983
## stroke           -0.06796375
## horsepower        0.13107251
## compressionratio -0.43574051
## peakrpm           1.00000000
```

```
cat("\n\n")
```

```
cat("Matriz de correlación del grupo 3: \n")
```

```
## Matriz de correlación del grupo 3:
```

```
print(correlacion_grupo3)
```

```
##             citympg highwaympg
## citympg    1.000000   0.971337
## highwaympg 0.971337   1.000000
```