M1_AI2

Sofia Cantu

2024-09-06

Exploración de la base de datos

```
M = read.csv("~/Downloads/ArchivosCodigos/precios_autos.csv")
```

Calcula medidas estadísticas apropiadas para las variables

```
# Estadísticas de síntesis de las variables cuantitativas
summary(M[, c("carwidth", "carheight", "price")])
```

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

```
sd(M$carwidth, na.rm = TRUE)
```

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

```
sd(M$carheight, na.rm = TRUE)
```

```
## [1] 2.443522
```

```
sd(M$price, na.rm = TRUE)
```

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

```
# Frecuencia para variables categóricas
table(M$carbody)
```

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

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

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

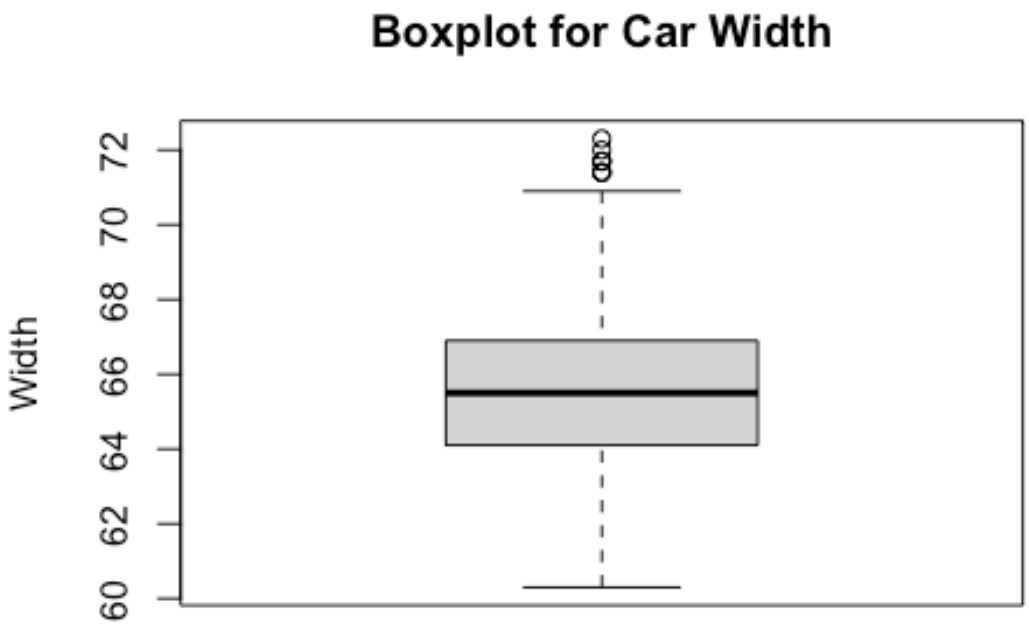
Analiza la correlación entre las variables (analiza posible colinealidad entre las variables)

```
cor(M[, c("carwidth", "carheight", "price")], use = "complete.obs")
```

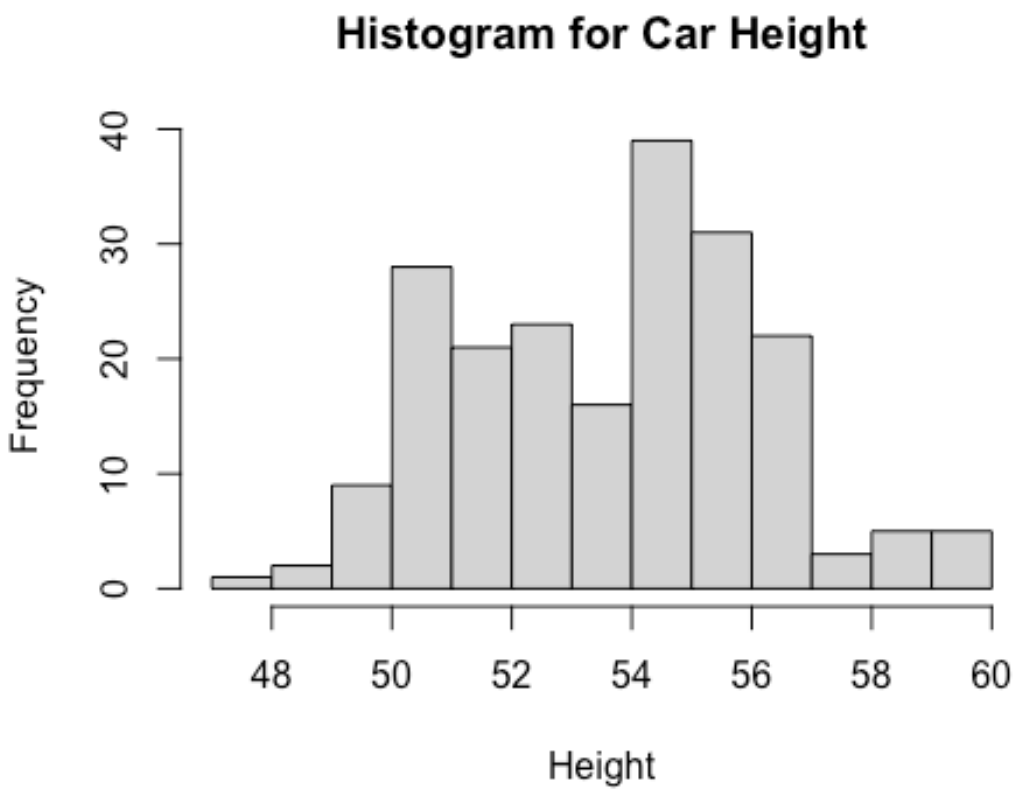
```
##             carwidth carheight     price
## carwidth  1.0000000 0.2792103 0.7593253
## carheight 0.2792103 1.0000000 0.1193362
## price     0.7593253 0.1193362 1.0000000
```

Explora los datos usando herramientas de visualización

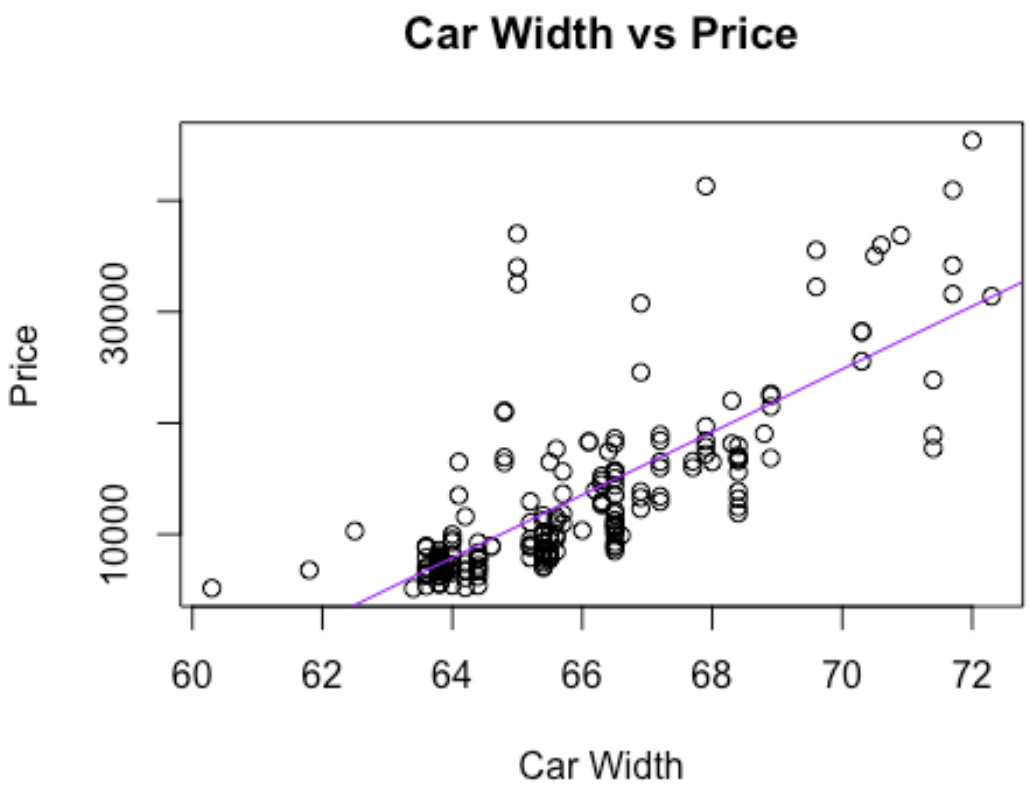
```
# Boxplot para anchura de cabina
boxplot(M$carwidth, main="Boxplot for Car Width", ylab="Width")
```



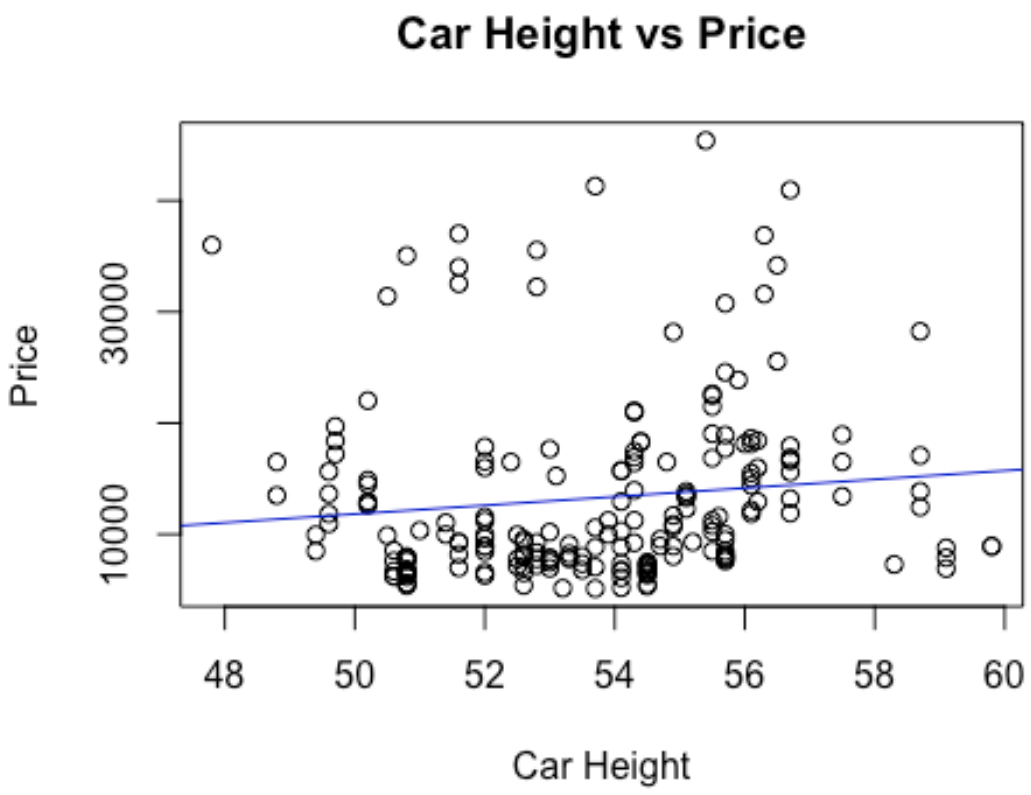

```
# Histograma de la altura de la cabina
hist(M$carheight, main="Histogram for Car Height", xlab="Height", breaks=10)
```



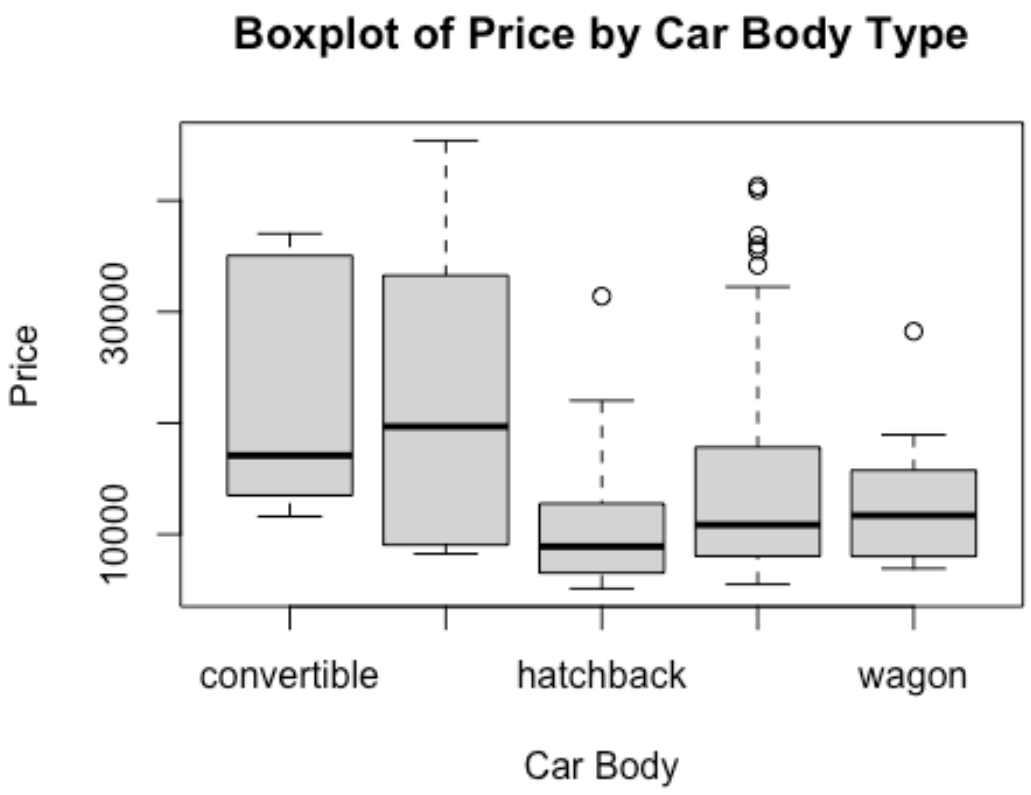

```r
# Gráfico de dispersión del precio en función de la anchura del vehículo
plot(M$carwidth, M$price, main="Car Width vs Price", xlab="Car Width", ylab="
Price")
abline(lm(M$price ~ M$carwidth), col="purple")
```




```r
# Gráfico de dispersión del precio en función de la altura del coche
plot(M$carheight, M$price, main="Car Height vs Price", xlab="Car Height", yla
b="Price")
abline(lm(M$price ~ M$carheight), col="blue3")
```




```r
# Boxplot por carrocería (variable cualitativa)
boxplot(M$price ~ M$carbody, main="Boxplot of Price by Car Body Type", ylab="
Price", xlab="Car Body")
```



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.

Mis propuestas son las siguientes: - Modelo 1: Regresión simple con carwidth - Modelo 2: Regresión múltiple con carwidth y carheight

```
modelo1 <- lm(price ~ carwidth, data = M)
summary(modelo1)

##
## Call:
## lm(formula = price ~ carwidth, data = M)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11097.4  -2690.0   -857.3    798.7  26318.4
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -173095.2    11215.6  -15.43   <2e-16 ***
## carwidth       2827.8      170.1   16.63   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5211 on 203 degrees of freedom
## Multiple R-squared:  0.5766, Adjusted R-squared:  0.5745
## F-statistic: 276.4 on 1 and 203 DF,  p-value: < 2.2e-16
```

```
modelo2 <- lm(price ~ carwidth + carheight, data = M)
summary(modelo2)
```

```
##
## 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
```

Analisis del primer modelo y sus coeficientes

```
# Nivel de significancia alfa = 0.04
summary(modelo1)
```

```
##
## Call:
## lm(formula = price ~ carwidth, data = M)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11097.4  -2690.0   -857.3    798.7  26318.4
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -173095.2    11215.6  -15.43   <2e-16 ***
## carwidth       2827.8      170.1   16.63   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5211 on 203 degrees of freedom
## Multiple R-squared:  0.5766, Adjusted R-squared:  0.5745
## F-statistic: 276.4 on 1 and 203 DF,  p-value: < 2.2e-16
```

```
# Extraer el valor p para validar si el modelo es significativo
anova(modelo1)
```

```
## Analysis of Variance Table
##
## Response: price
##            Df     Sum Sq    Mean Sq F value    Pr(>F)
## carwidth    1 7506797404 7506797404  276.42 < 2.2e-16 ***
## Residuals 203 5512841958   27156857
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Analisis del segundo modelo y sus coeficientes

```
summary(modelo2)
```

```
##
## 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
```

```
anova(modelo2)
```

```
## Analysis of Variance Table
##
## Response: price
##            Df     Sum Sq    Mean Sq  F value  Pr(>F)
## carwidth    1 7506797404 7506797404 281.2491 < 2e-16 ***
## carheight   1  121276268  121276268   4.5437 0.03425 *
## Residuals 202 5391565690   26690919
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Porcentaje de variación explicada por cada modelo

```
cat("R-squared modelo1:", summary(modelo1)$r.squared, "\n")
```

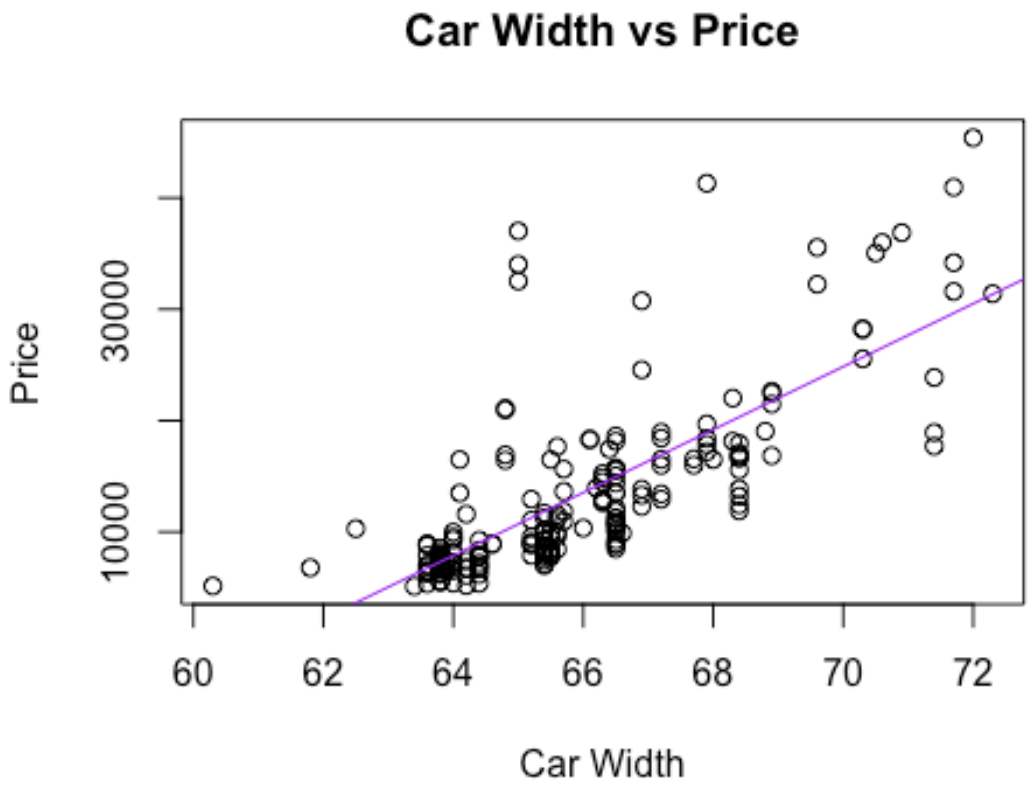
```
## R-squared modelo1: 0.5765749
```

```
cat("R-squared modelo2:", summary(modelo2)$r.squared, "\n")
```

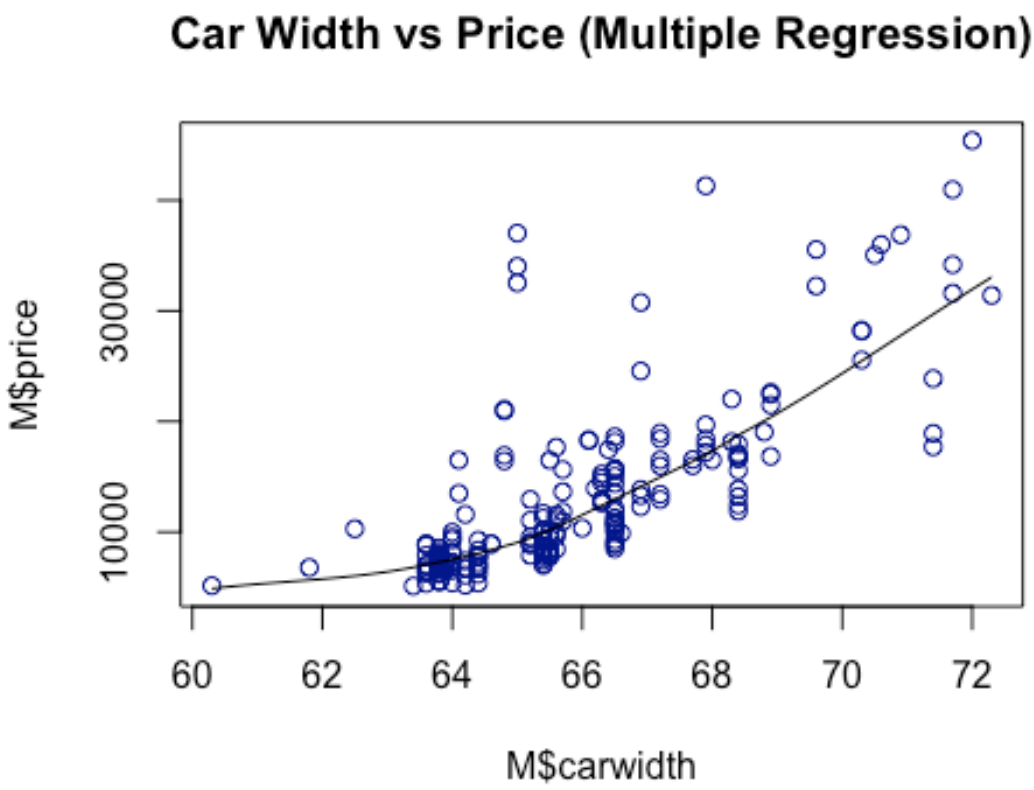
```
## R-squared modelo2: 0.5858898
```

Visualización de los modelos

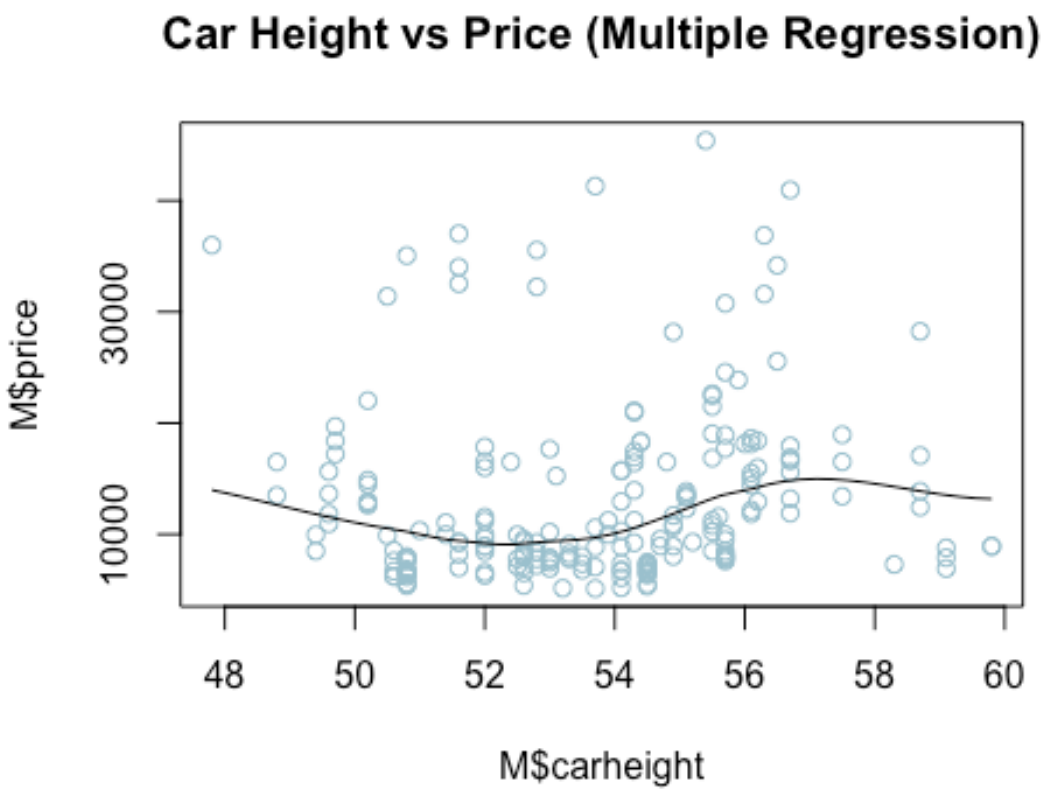
```
# Gráfico de dispersión con línea de mejor ajuste para el primer modelo
plot(M$carwidth, M$price, main="Car Width vs Price", xlab="Car Width", ylab="
Price")
abline(modelo1, col="purple")
```




```
# Gráfico de dispersión para el modelo múltiple (carwidth y carheight)
# Predicciones del modelo múltiple
pred <- predict(modelo2)
scatter.smooth(M$carwidth, M$price, main="Car Width vs Price (Multiple Regres
sion)", col="blue4")
```




```
scatter.smooth(M$carheight, M$price, main="Car Height vs Price (Multiple Regr
ession)", col="lightblue3")
```



##

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

Modelo 1: Regresión Simple con carwidth

Ecuación del modelo: - Precio = − 173095.2 + 2827.8 × Carwidth

Interpretación: - El coeficiente de carwidth (2827.8) indica que por cada unidad de aumento en el ancho del automóvil (carwidth), se espera que el precio aumente en aproximadamente $2827.8, manteniendo todo lo demás constante.

Significancia: - El modelo es altamente significativo, con un valor p menor que 0.001. - El coeficiente de carwidth también es altamente significativo, lo que indica que tiene una relación fuerte con el precio.

R^2: - El R^2 del modelo es 0.5766, lo que indica que aproximadamente el 57.66% de la variabilidad en el precio del automóvil puede explicarse mediante el ancho del auto (carwidth). Para ser un modelo de una sola variable, es un porcetaje muy bueno.

Modelo 2: Regresión Múltiple con carwidth y carheight

Ecuación del modelo: - Precio = −162328.7 + 2932.3 × Carwidth − 328.6 × Carheight

Interpretación: - El coeficiente de carwidth (2932.3) sugiere que el precio aumenta en $2932.3 por cada unidad de incremento en el ancho del auto, manteniendo constante la altura. - El coeficiente de carheight (-328.6) sugiere que por cada unidad de incremento en la altura del automóvil, el precio disminuye en aproximadamente $328.6, lo que podría indicar que los autos más altos tienden a tener precios ligeramente más bajos, manteniendo el ancho constante.

Significancia: - Tanto el modelo completo como los coeficientes son significativos con un nivel de confianza del 96% (α=0.04). - El coeficiente de carheight es significativo con un valor p de 0.0342. Indica que, aunque más débil que carwidth, tiene un impacto relevante en el precio.

R^2: - El R^2 del modelo múltiple es 0.5859, lo que indica que este modelo explica un 58.59% de la variación en el precio. Es un aumento pequeño con respecto al modelo simple, lo que sugiere que la adición de carheight no mejora drásticamente la capacidad predictiva del modelo.

Validación de Significancia y Coeficientes:

Pruebas de hipótesis: Para el modelo completo: - Hipótesis nula (H0): El modelo no es significativo ($R^2 = 0$). - Hipótesis alternativa (H1): El modelo es significativo ($R^2 > 0$).

Con un valor p menor a 0.001 en ambos modelos, rechazamos la hipótesis nula, indicando que los modelos son significativos. Para los coeficientes (βi): - H0: El coeficiente no es significativamente diferente de 0. - H1: El coeficiente es significativamente diferente de 0.

Para ambos modelos, los coeficientes de carwidth y carheight son significativos a un nivel alfa de 0.04.

Visualización y Gráficos

Modelo Simple: - El gráfico de dispersión muestra una relación positiva clara entre carwidth y el precio, con la línea de regresión ajustada en morado. La mayoría de los puntos están relativamente cerca de la línea, lo que indica un buen ajuste.

Modelo Múltiple: - El gráfico de dispersión de carwidth vs. precio en el modelo múltiple muestra una tendencia similar, con un ajuste visualmente comparable. - El gráfico de dispersión de carheight vs. precio muestra una relación mucho más débil, con una curva leve y menos ajuste alrededor de la línea de regresión, lo que confirma que carheight tiene menos impacto en el precio comparado con carwidth.

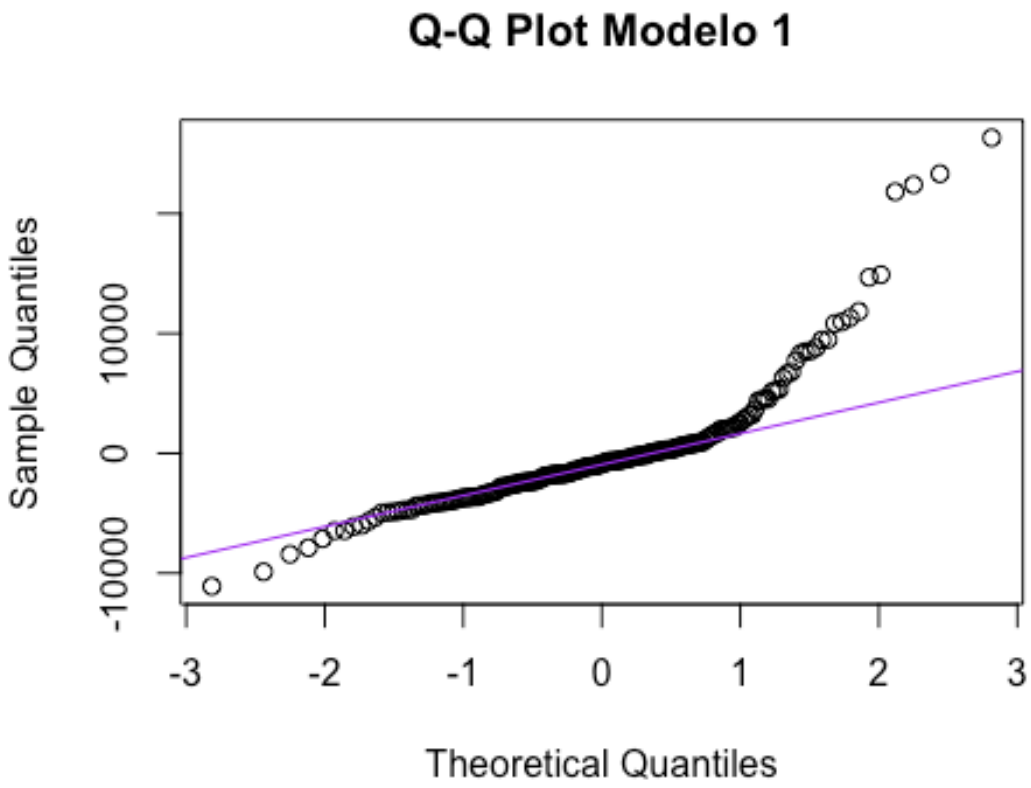
Concusion de esta sección

El modelo basado únicamente en carwidth ya explica una cantidad considerable de la variación en los precios de los automóviles (57.66%). Al añadir carheight, la capacidad explicativa del modelo solo mejora ligeramente (58.59%), indicando que carheight tiene un impacto menor sobre el precio en comparación con carwidth.

Analiza la validez de los modelos propuestos:

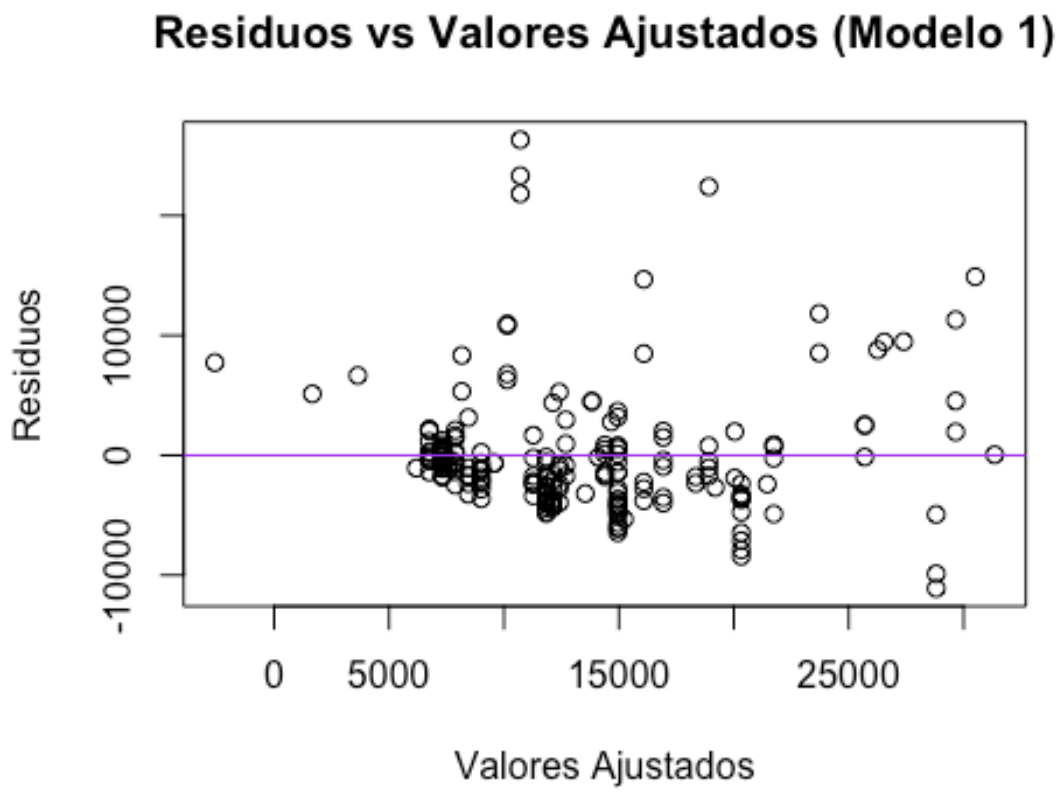
Modelo 1

```
modelo1 <- lm(price ~ carwidth, data = M)
residuos_modelo1 <- residuals(modelo1)
valores_ajustados_modelo1 <- fitted(modelo1)

qqnorm(residuos_modelo1, main="Q-Q Plot Modelo 1")
qqline(residuos_modelo1, col="purple")
```




```
shapiro.test(residuos_modelo1)

##
##  Shapiro-Wilk normality test
##
## data:  residuos_modelo1
## W = 0.80313, p-value = 2.324e-15

mean(residuos_modelo1)

## [1] -9.760431e-13

plot(valores_ajustados_modelo1, residuos_modelo1,
     main="Residuos vs Valores Ajustados (Modelo 1)",
```

```
     xlab="Valores Ajustados", ylab="Residuos")
abline(h = 0, col = "purple")
library(lmtest)

## Loading required package: zoo

##
## Attaching package: 'zoo'

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

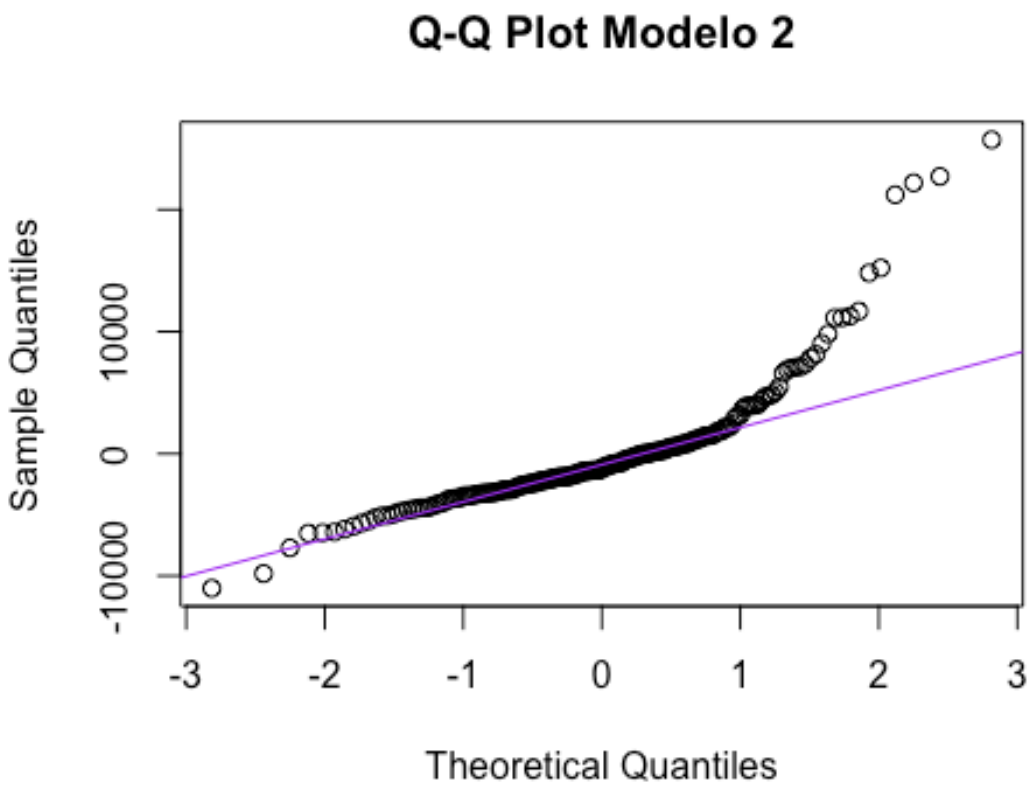



```
bptest(modelo1)

##
##  studentized Breusch-Pagan test
##
## data:  modelo1
## BP = 4.0726, df = 1, p-value = 0.04358

dwtest(modelo1)

##
##  Durbin-Watson test
```

```
##
## data:  modelo1
## DW = 0.63823, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

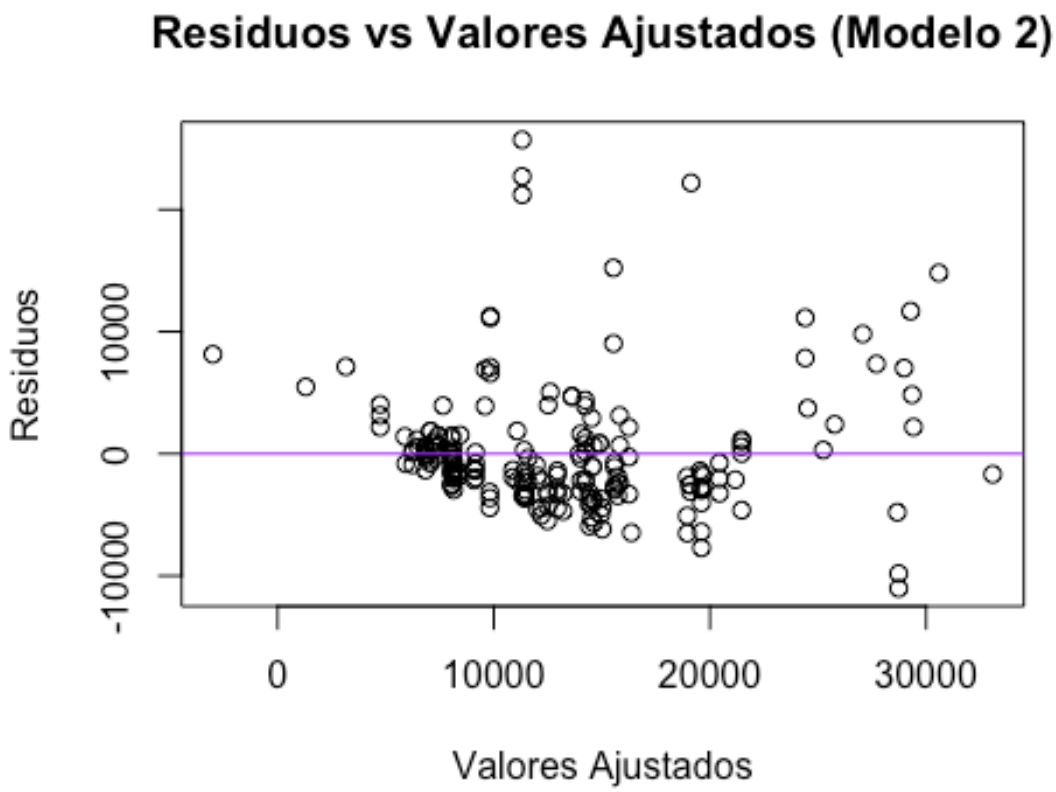
Modelo 2

```
modelo2 <- lm(price ~ carwidth + carheight, data = M)
residuos_modelo2 <- residuals(modelo2)
valores_ajustados_modelo2 <- fitted(modelo2)

qqnorm(residuos_modelo2, main="Q-Q Plot Modelo 2")
qqline(residuos_modelo2, col="purple")
```




```
shapiro.test(residuos_modelo2)

##
##  Shapiro-Wilk normality test
##
## data:  residuos_modelo2
## W = 0.80882, p-value = 3.97e-15

mean(residuos_modelo2)

## [1] -5.31278e-13
```

```r
plot(valores_ajustados_modelo2, residuos_modelo2,
     main="Residuos vs Valores Ajustados (Modelo 2)",
     xlab="Valores Ajustados", ylab="Residuos")
abline(h = 0, col = "purple")
```




```r
bptest(modelo2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  modelo2
## BP = 4.6072, df = 2, p-value = 0.0999
```

```r
dwtest(modelo2)
```

```
##
##  Durbin-Watson test
##
## data:  modelo2
## DW = 0.67299, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

Interpretación de cada uno de los analisis

Modelo 1: Regresión con carwidth

Normalidad de los residuos: - Q-Q Plot: Observamos que los puntos se desvían de la línea roja en las colas, lo que indica que los residuos no siguen completamente una distribución normal. Hay evidencia de que los valores extremos no se ajustan a una distribución normal. - Prueba de Shapiro-Wilk: El valor p es extremadamente pequeño (2.324e-15), lo que indica que podemos rechazar la hipótesis nula de normalidad. Por lo tanto, los residuos no son normales.

Media cero de los residuos: - La media de los residuos es muy cercana a cero (-9.760431e-13), lo que indica que el modelo cumple con este supuesto.

Homocedasticidad (Varianza constante): - Gráfico de residuos vs valores ajustados: El gráfico muestra una dispersión que parece crecer a medida que aumentan los valores ajustados, lo que indica una posible heterocedasticidad (la varianza de los residuos no es constante). - Prueba de Breusch-Pagan: El valor p es de 0.04358, menor que 0.05, lo que indica que hay suficiente evidencia para rechazar la hipótesis de homocedasticidad. Esto sugiere que existe heterocedasticidad.

Independencia de los residuos: - Prueba de Durbin-Watson: El valor de la prueba es 0.63823, lo que es significativamente menor que 2. Esto indica que existe autocorrelación positiva entre los residuos, violando el supuesto de independencia.

Modelo 2: Regresión con carwidth y carheight

Normalidad de los residuos: - Q-Q Plot: Similar al Modelo 1, los puntos se desvían de la línea roja en las colas, lo que indica que los residuos no son completamente normales. - Prueba de Shapiro-Wilk: El valor p es 3.97e-15, lo que significa que podemos rechazar la hipótesis de normalidad. Por lo tanto, los residuos no son normales.

Media cero de los residuos: - La media de los residuos es cercana a cero (-5.31278e-13), lo que cumple con el supuesto de media cero.

Homocedasticidad (Varianza constante): - Gráfico de residuos vs valores ajustados: Similar al Modelo 1, se observa una dispersión creciente a medida que aumentan los valores ajustados, lo que indica posible heterocedasticidad. - Prueba de Breusch-Pagan: El valor p es 0.0999, lo que significa que no podemos rechazar la hipótesis de homocedasticidad con un nivel de significancia de 0.05. Sin embargo, está cerca del límite de significancia, por lo que se sugiere precaución.

Independencia de los residuos: - Prueba de Durbin-Watson: El valor de la prueba es 0.67299, indicando que también existe autocorrelación positiva entre los residuos, lo que viola el supuesto de independencia.

Concusion de esta sección

Estos resultados indican que, aunque los modelos ajustan bien en términos de coeficientes y R^2, la validez de los supuestos puede estar en cuestión, especialmente en relación con la normalidad, homocedasticidad e independencia de los residuos. Podría ser útil considerar transformaciones de las variables o usar modelos más robustos que no dependan tanto de estos supuestos.

Concusión General

Mejor modelo

El Modelo 2, que incluye tanto carwidth como carheight como variables predictoras, es el mejor modelo de regresión lineal entre los dos que se analizaron. Esto se debe a las siguientes razones: - Mejor capacidad explicativa: El R^2 ajustado del Modelo 2 es ligeramente mayor que el del Modelo 1 (0.5818 frente a 0.5745), lo que indica que el Modelo 2 explica una mayor proporción de la variación en el precio del automóvil. - Significancia estadística: Ambos coeficientes en el Modelo 2 (carwidth y carheight) son estadísticamente significativos con un nivel alfa de 0.04, lo que significa que ambas variables tienen un impacto significativo en el precio. - Validez de los supuestos: Aunque ambos modelos presentan problemas de heterocedasticidad y autocorrelación de los residuos, el Modelo 2 presenta una mejor homocedasticidad (según la prueba de Breusch-Pagan) en comparación con el Modelo 1. Además, la inclusión de la variable carheight mejora ligeramente la capacidad explicativa sin deteriorar sustancialmente la validez de los supuestos.

Variables asignadas que influyen en el precio del auto

Carwidth (ancho del auto): - El coeficiente positivo de carwidth (2932.3 en el Modelo 2) indica que el ancho del automóvil tiene una influencia positiva en el precio.

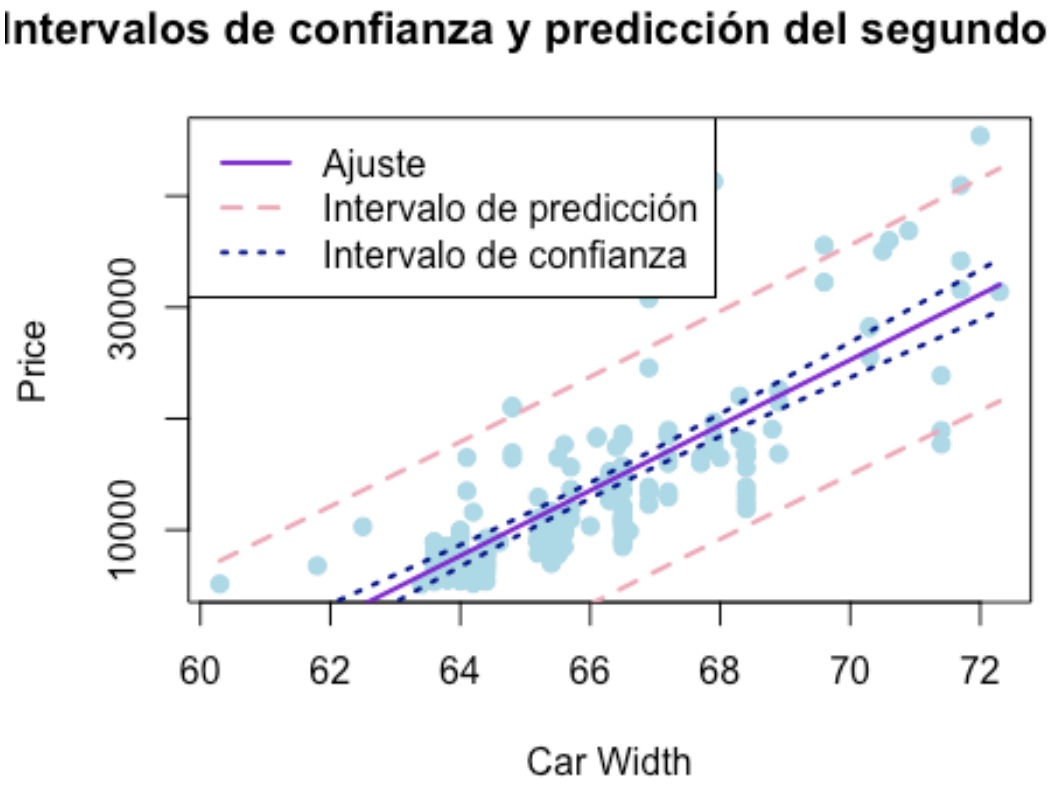
Carheight (altura del auto): - El coeficiente negativo de carheight (-328.6 en el Modelo 2) indica que la altura del automóvil tiene una influencia negativa en el precio. Por cada unidad adicional de altura, el precio disminuye en aproximadamente $328.6.

Esto se púede deber al mercado objetivo prefiere auntos bajos y anchos, como lo son los deportivos.

Intervalos de predicción y confianza

```
nuevos_datos <- data.frame(
  carwidth = seq(min(M$carwidth), max(M$carwidth), length.out = 100),
  carheight = mean(M$carheight) # Mantener carheight fijo como su media para
simplificar la gráfica
)
predicciones <- predict(modelo2, nuevos_datos, interval = "prediction")
confianza <- predict(modelo2, nuevos_datos, interval = "confidence")
```

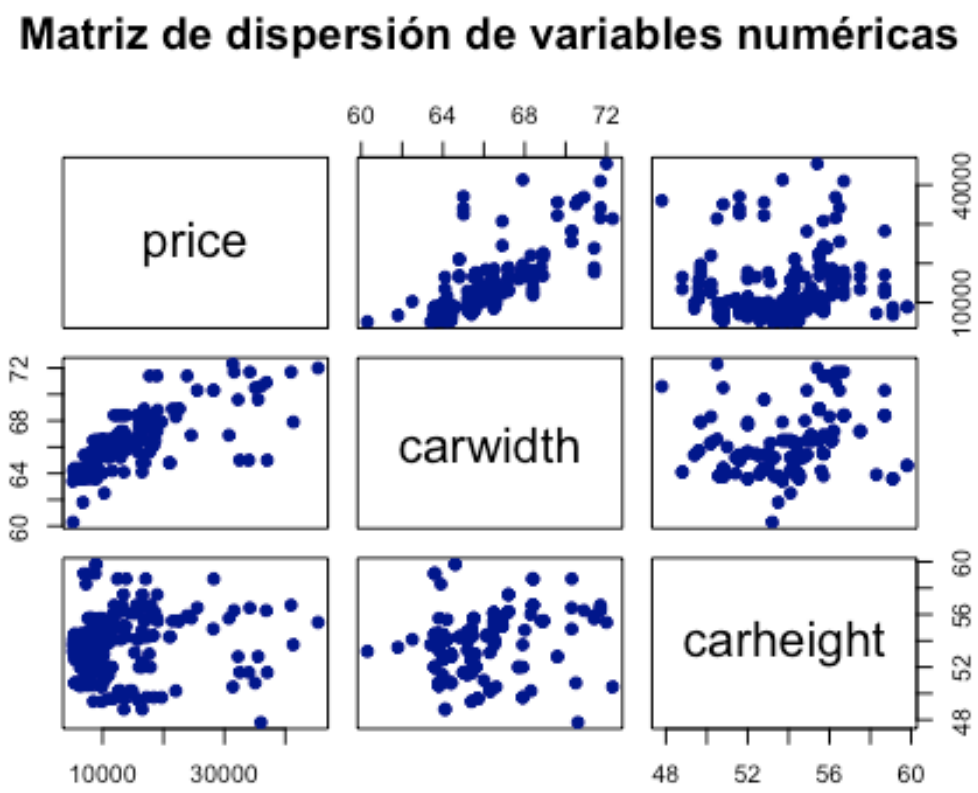
```r
plot(M$carwidth, M$price, main = "Intervalos de confianza y predicción del se
gundo modelo",
     xlab = "Car Width", ylab = "Price", pch = 19, col = "lightblue")
lines(nuevos_datos$carwidth, predicciones[, "fit"], col = "purple3", lwd = 2)
lines(nuevos_datos$carwidth, predicciones[, "lwr"], col = "pink2", lty = 2, l
wd = 2)
lines(nuevos_datos$carwidth, predicciones[, "upr"], col = "pink2", lty = 2, l
wd = 2)
lines(nuevos_datos$carwidth, confianza[, "lwr"], col = "blue4", lty = 3, lwd
= 2)
lines(nuevos_datos$carwidth, confianza[, "upr"], col = "blue4", lty = 3, lwd
= 2)

legend("topleft", legend = c("Ajuste", "Intervalo de predicción", "Intervalo
de confianza"),
       col = c("purple3", "pink2", "blue4"), lty = c(1, 2, 3), lwd = 2)
```




```r
# Calcular los intervalos para la variable Y
intervalos <- predict(modelo2, interval = "prediction")

## Warning in predict.lm(modelo2, interval = "prediction"): predictions on cu
rrent data refer to _future_ responses

summary(intervalos)
```

```
##       fit              lwr              upr
##  Min.   :-2994   Min.   :-13382   Min.   : 7393
##  1st Qu.: 8146   1st Qu.: -2086   1st Qu.:18386
##  Median :12486   Median :  2255   Median :22716
##  Mean   :13277   Mean   :  3016   Mean   :23538
##  3rd Qu.:15585   3rd Qu.:  5312   3rd Qu.:25858
##  Max.   :33080   Max.   : 22528   Max.   :43632
```

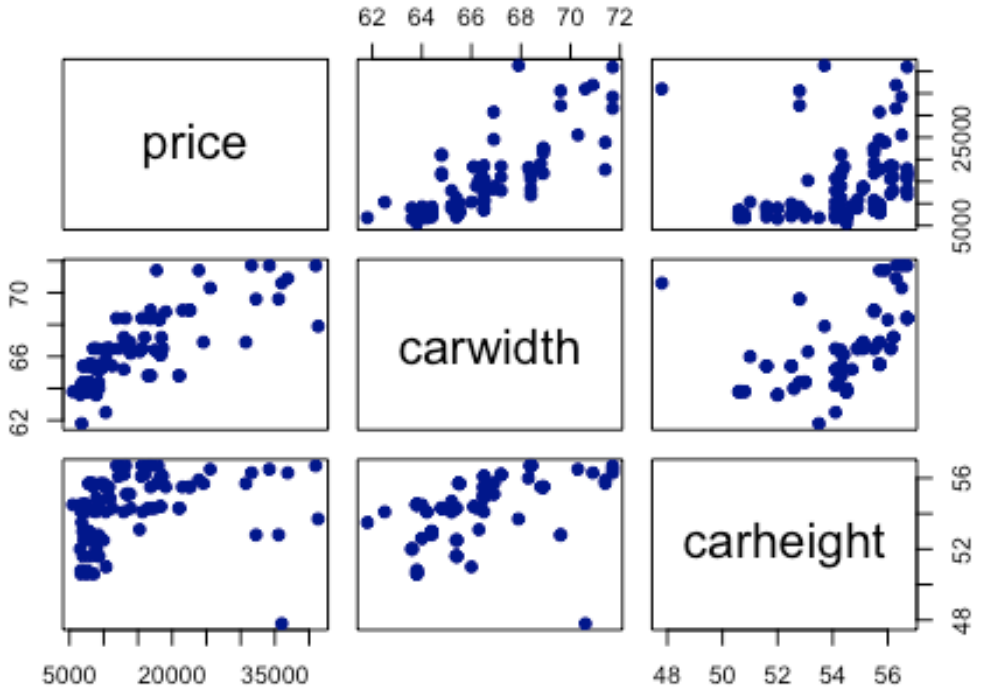
```r
# Gráfica de pares de las variables numéricas más importantes
pairs(~ price + carwidth + carheight, data = M,
      main = "Matriz de dispersión de variables numéricas",
      pch = 19, col = "blue4")
```



Variavle cualitativa adicional

```r
M_sedan <- subset(M, carbody == "sedan")
pairs(~ price + carwidth + carheight, data = M_sedan,
      main = "Matriz de dispersión de variables numéricas (Carrocería Sedan)"
,
      pch = 19, col = "blue4")
```

z de dispersión de variables numéricas (Carrocería S



Interpreta en el contexto del problema

Gráfico de Intervalos de Confianza y Predicción: - En el gráfico que muestra los intervalos de predicción y confianza para carwidth (ancho del auto), observamos que los intervalos de predicción (líneas punteadas rosadas) son más amplios que los intervalos de confianza (líneas punteadas azules), lo que es esperado. Los intervalos de confianza representan la certeza del valor esperado del precio, mientras que los intervalos de predicción consideran la variabilidad en los datos individuales, por lo que son más amplios. - A medida que aumenta el ancho del auto (carwidth), el precio también aumenta de manera consistente, lo que refuerza que carwidth es una variable significativa y de fuerte influencia positiva en el precio del auto.

Matriz de Dispersión: - La matriz de dispersión de las variables price, carwidth y carheight muestra una clara relación positiva entre carwidth y price, como ya se había observado en los modelos de regresión. - La relación entre carheight (altura del auto) y price es más débil, mostrando una mayor dispersión. Esto sugiere que la altura del auto tiene un menor impacto en el precio, e incluso un impacto negativo, como vimos en el análisis anterior.

Análisis por Categoría de Carrocería (Sedan): - Al segmentar por la categoría sedan, los patrones observados son consistentes con los datos globales. Se observa que para los autos tipo sedan, carwidth sigue siendo un fuerte predictor del precio, mientras que

carheight tiene un impacto menor. - La variabilidad dentro de los sedanes es menor, lo cual es razonable porque los sedanes tienden a tener características más homogéneas que otras carrocerías (como SUVs o hatchbacks).

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?

Propongo los siguientes grupos

Grupo 1: Dimensiones físicas del auto: - Carwidth (ancho del auto) - Carheight (altura del auto) - Curbweight (peso del auto) - CarLength (longitud del auto) Estar variables ayudan a determinar la categoría del vehículo (compacto, sedán, SUV, etc.) y, por ende, su precio.

Grupo 2: Desempeño del motor: - Enginesize (tamaño del motor) - Horsepower (potencia del motor) - Compressionratio (relación de compresión) - Peakrpm (revoluciones máximas del motor) son variables asociadas al desempeño del motor, y también tienen un impacto significativo en el precio, ya que autos con motores más potentes suelen ser más costosos.

Grupo 3: Consumo y eficiencia: - Citympg (millas por galón en ciudad) - Highwaympg (millas por galón en carretera) Para ciertos mercados como son los ubers de todo el mundo, o ciertos países como el estaunidense. La eficiencia tiene un impacto directo al costo que te da a largo plazo el vehículo y por ende puede ser muy atrativo para cierto sector, ayudandonos a predecir el precio al conocer el mercado a un nivel mayor.

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

```
# Cálculo de medias
summary(M[c("carwidth", "carheight", "curbweight", "enginesize", "horsepower"
, "price")])
```

```
##     carwidth       carheight       curbweight     enginesize      horsepow
er
##  Min.   :60.30   Min.   :47.80   Min.   :1488   Min.   : 61.0   Min.   : 4
8.0
##  1st Qu.:64.10   1st Qu.:52.00   1st Qu.:2145   1st Qu.: 97.0   1st Qu.: 7
0.0
##  Median :65.50   Median :54.10   Median :2414   Median :120.0   Median : 9
5.0
##  Mean   :65.91   Mean   :53.72   Mean   :2556   Mean   :126.9   Mean   :10
4.1
##  3rd Qu.:66.90   3rd Qu.:55.50   3rd Qu.:2935   3rd Qu.:141.0   3rd Qu.:11
6.0
##  Max.   :72.30   Max.   :59.80   Max.   :4066   Max.   :326.0   Max.   :28
```

```
8.0
##      price
##  Min.   : 5118
##  1st Qu.: 7788
##  Median :10295
##  Mean   :13277
##  3rd Qu.:16503
##  Max.   :45400
```

```
# Matriz de correlación
cor(M[c("price", "carwidth", "carheight", "curbweight", "enginesize", "horsep
ower", "citympg", "highwaympg")])
```

```
##                 price   carwidth   carheight curbweight  enginesize horsep
ower
## price       1.0000000  0.7593253  0.11933623  0.8353049  0.87414480  0.808
1388
## carwidth    0.7593253  1.0000000  0.27921032  0.8670325  0.73543340  0.640
7321
## carheight   0.1193362  0.2792103  1.00000000  0.2955717  0.06714874 -0.108
8021
## curbweight  0.8353049  0.8670325  0.29557173  1.0000000  0.85059407  0.750
7393
## enginesize  0.8741448  0.7354334  0.06714874  0.8505941  1.00000000  0.809
7687
## horsepower  0.8081388  0.6407321 -0.10880206  0.7507393  0.80976865  1.000
0000
## citympg    -0.6857513 -0.6427043 -0.04863963 -0.7574138 -0.65365792 -0.801
4562
## highwaympg -0.6975991 -0.6772179 -0.10735763 -0.7974648 -0.67746991 -0.770
5439
##                citympg highwaympg
## price      -0.68575134 -0.6975991
## carwidth   -0.64270434 -0.6772179
## carheight  -0.04863963 -0.1073576
## curbweight -0.75741378 -0.7974648
## enginesize -0.65365792 -0.6774699
## horsepower -0.80145618 -0.7705439
## citympg     1.00000000  0.9713370
## highwaympg  0.97133704  1.0000000
```

Esta agrupación permite un análisis más estructurado y profundo de las diferentes características del vehículo que afectan el precio. Mantendría las agrupaciones mencionadas en el punto anterior.