Actividad 2

Catherine Rojas

2024-09-06

Actividad Integradora 2

Descripción

Una empresa automovilística china aspira a entrar en el mercado estadounidense. Desea establecer allí una unidad de fabricación y producir automóviles localmente para competir con sus contrapartes estadounidenses y europeas. Contrataron una empresa de consultoría de automóviles para identificar los principales factores de los que depende el precio de los automóviles, específicamente, en el mercado estadounidense, ya que pueden ser muy diferentes del mercado chino. Esencialmente, la empresa quiere saber:

Qué variables son significativas para predecir el precio de un automóvil?

Qué tan bien describen esas variables el precio de un automóvil?

Con base en varias encuestas de mercado, la consultora ha recopilado un gran conjunto de datos de diferentes tipos de automóviles en el mercado estadounidense que presenta en el siguiente archivo. Las variables recopiladas vienen descritas en el diccionario de términos 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:

Segundo grupo. Altura del auto, ancho del auto y si es convertible o no.

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?

Parte 1: Exploración de la base de datos

```
library(readr)
```

```
## Warning: package 'readr' was built under R version 4.3.3
```

```
data <- read_csv("Concentracion Estadistica/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.

# data
```

Calcula medidas estadísticas apropiadas para las variables:

cuantitativas (media, desviación estándar, cuantiles, etc)

Variables cuantitativas: carheight (Altura del auto), carwidth (Ancho del auto), price (Precio del Auto)

```
# Medidas estadísticas para las variables cuantitativas (altura y ancho)
altura_stats <- summary(data$carheight)
altura_sd <- sd(data$carheight)
ancho_stats <- summary(data$carwidth)
ancho_sd <- sd(data$carwidth)
precio_stats <- summary(data$price)
precio_sd <- sd(data$price)

# Imprimir estadísticas
cat("Altura del auto:\n")

## Altura del auto:

altura_stats

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

cat("Desviación estándar:", altura_sd, "\n")

## Desviación estándar: 2.443522

cat("Ancho del auto:\n")

## Ancho del auto:

ancho_stats

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

cat("Desviación estándar:", ancho_sd, "\n")

## Desviación estándar: 2.145204
```

```r
cat("Preciodel auto:\n")
```

```
## Preciodel auto:
```

```r
precio_stats
```

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

```r
cat("Desviación estándar:", precio_sd, "\n" )
```

```
## Desviación estándar: 7988.852
```

Una observación importante aquí es que, el precio promedio es 13,277 con una desviación estándar alta de 7,988.85, lo que refleja una gran variabilidad en los precios, desde 5,118 hasta 45,400, sugiriendo así que los autos en el conjunto de datos varían mucho en costo.

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

Variable cualitativa: Si el auto es convertible o no, que la identificaremos a partir de la columna carbody, donde uno de los valores es “convertible”.

```r
# Variable cualitativa: Si el auto es convertible o no
# Crear una nueva columna que indique si el auto es convertible o no
data$is_convertible <- ifelse(data$carbody == "convertible", "Sí", "No")

# Frecuencia y proporción de autos convertibles
convertible_freq <- table(data$is_convertible)
convertible_prop <- prop.table(convertible_freq)

# Imprimir frecuencias y proporciones
cat("Frecuencias de convertibles:\n")
```

```
## Frecuencias de convertibles:
```

```r
print(convertible_freq)
```

```
##
##  No  Sí
## 199   6
```

```r
cat("Proporciones de convertibles:\n")
```

```
## Proporciones de convertibles:
```

```r
print(convertible_prop)
```

```
##
##         No         Sí
## 0.97073171 0.02926829
```

Como interpretación de estos resultados, se puede decir que los autos convertibles son muy poco comunes en este conjunto de datos, representando una pequeña proporción del total de autos disponibles, lo que podría tener un impacto limitado en el análisis, dado que hay muchos más autos no convertibles que convertibles.

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

```
# Crear una nueva columna que indique si el auto es convertible o no
data$is_convertible <- ifelse(data$carbody == "convertible", 1, 0)

# Seleccionar las columnas de interés: altura, ancho y si es convertible
relevant_columns <- data[, c("carheight", "carwidth",  "price",
"is_convertible")]

# Calcular la matriz de correlación
correlation_matrix <- cor(relevant_columns, use = "complete.obs")

# Imprimir la matriz de correlación
correlation_matrix
```

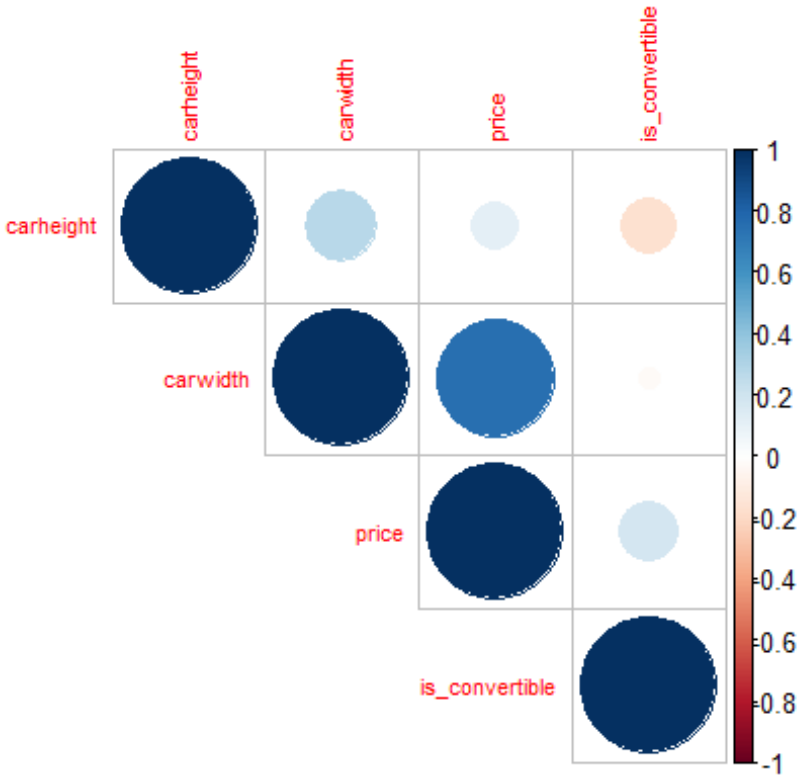
```
##                carheight    carwidth     price is_convertible
## carheight       1.0000000  0.27921032 0.1193362    -0.16323866
## carwidth        0.2792103  1.00000000 0.7593253    -0.02632807
## price           0.1193362  0.75932530 1.0000000     0.18768121
## is_convertible -0.1632387 -0.02632807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

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.3.3
```

```
## corrplot 0.94 loaded
```

```
# Graficar la matriz de correlación
corrplot(correlation_matrix, method = "circle", type = "upper", tl.cex =
0.7)
```



* El ancho del auto es el factor más relacionado con el precio con una correlación de 0.759, por lo que parece ser un predictor importante.

- La altura del auto y si es convertible o no tienen una correlación débil con el precio, por lo que son menos relevantes.
- Se puede decir que, el ancho del auto es un factor clave para predecir el precio, mientras que las otras variables tienen menos peso en este sentido.

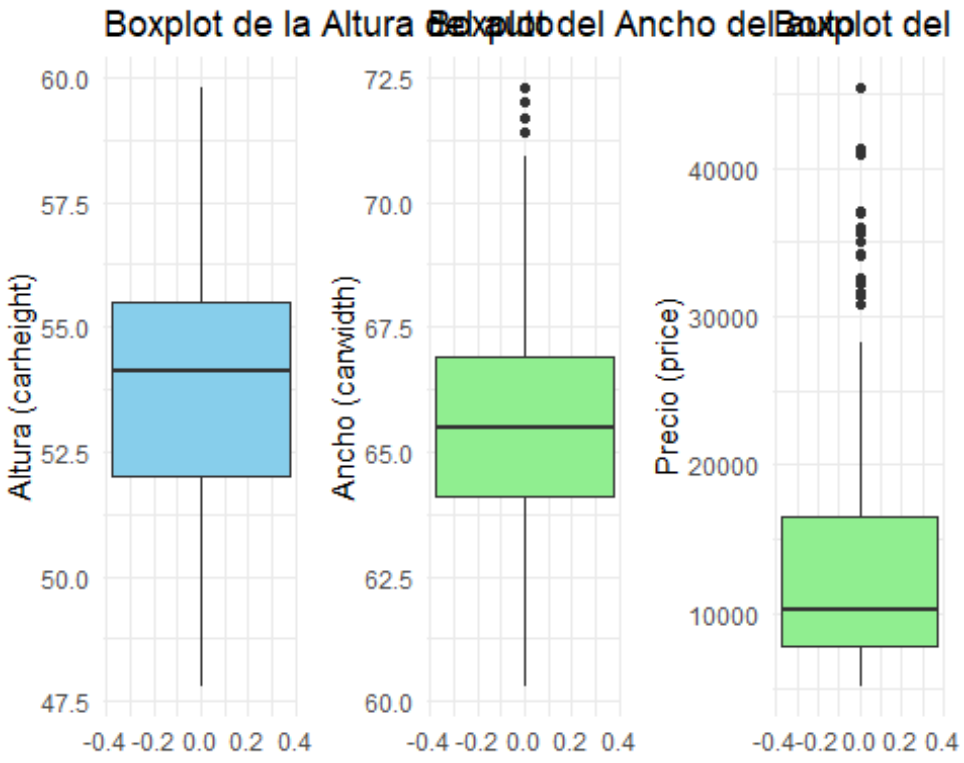
Explora los datos usando herramientas de visualización (si lo consideras necesario):

Variables cuantitativas:

Boxplot (visualización de datos atípicos)

```
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.3.3

library(GGally)

## Warning: package 'GGally' was built under R version 4.3.3

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```r
library(gridExtra)

## Warning: package 'gridExtra' was built under R version 4.3.3

# Boxplot de la Altura del auto
plot1 <- ggplot(data, aes(y = carheight)) +
  geom_boxplot(fill = "skyblue") +
  labs(title = "Boxplot de la Altura del auto", y = "Altura (carheight)")
+
  theme_minimal()

# Boxplot del Ancho del auto
plot2 <- ggplot(data, aes(y = carwidth)) +
  geom_boxplot(fill = "lightgreen") +
  labs(title = "Boxplot del Ancho del auto", y = "Ancho (carwidth)") +
  theme_minimal()

# Boxplot del Precio del auto
plot3 <- ggplot(data, aes(y = price)) +
  geom_boxplot(fill = "lightgreen") +
  labs(title = "Boxplot del Precio del auto", y = "Precio (price)") +
  theme_minimal()


grid.arrange(plot1, plot2, plot3, nrow = 1)
```

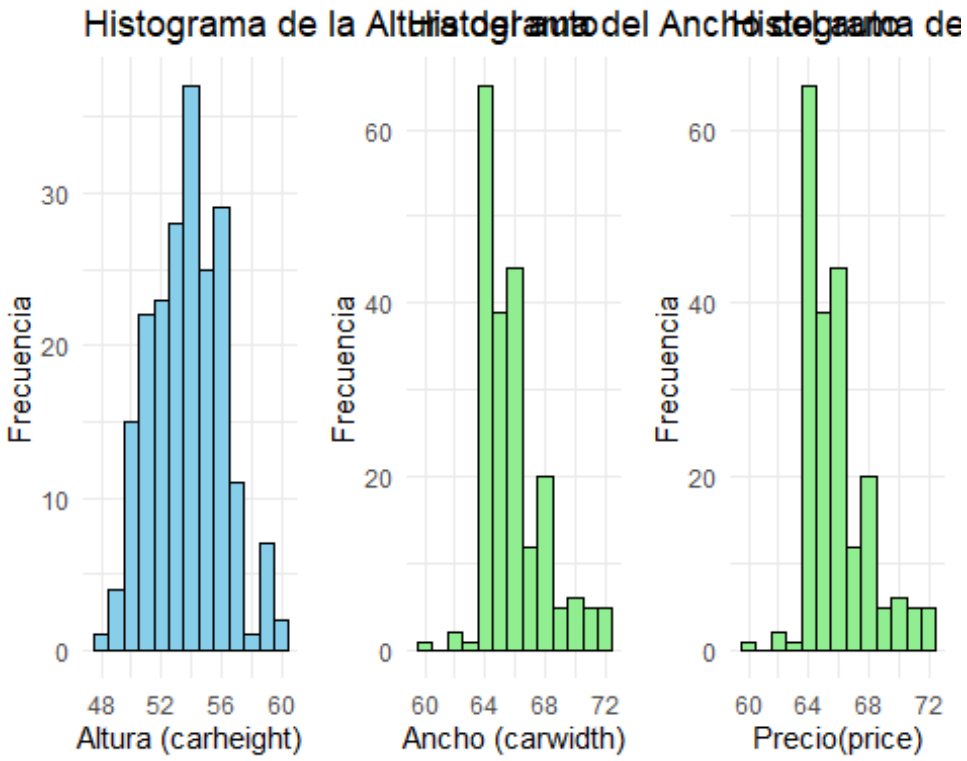


El precio de los autos tiene una mayor variabilidad que las otras dos variables, con varios valores

atípicos presentes. El ancho también presenta algunos autos inusualmente anchos, mientras que la altura muestra una distribución más compacta y sin valores extremos.

Histogramas

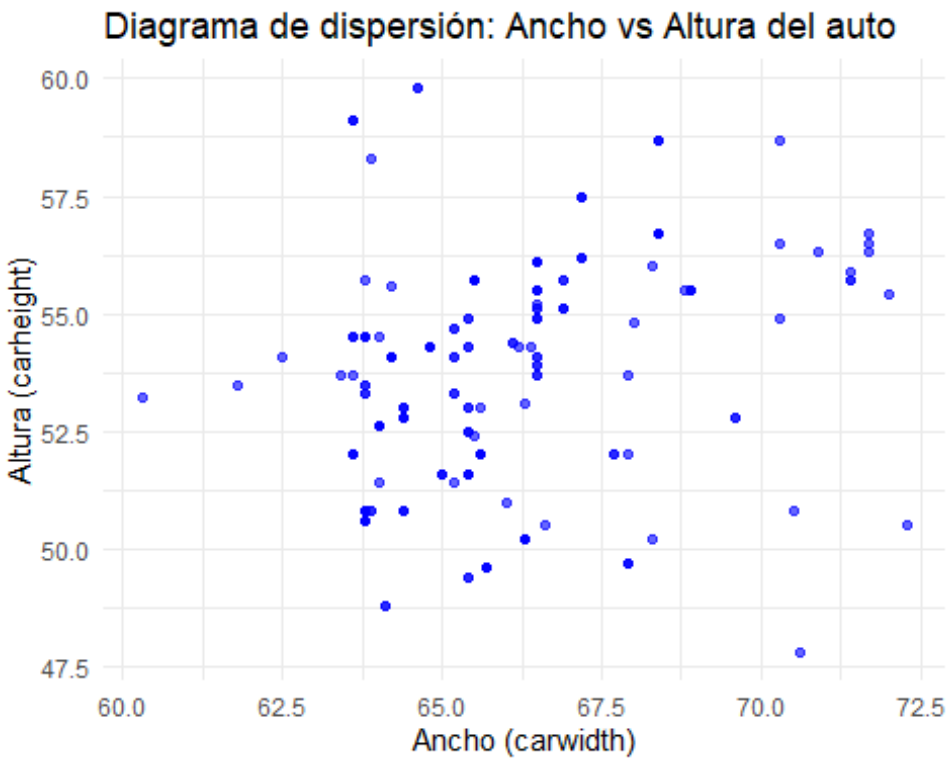
```
# Histograma de la Altura del auto
h1 <- ggplot(data, aes(x = carheight)) +
  geom_histogram(binwidth = 1, fill = "skyblue", color = "black") +
  labs(title = "Histograma de la Altura del auto", x = "Altura
(carheight)", y = "Frecuencia") +
  theme_minimal()

# Histograma del Ancho del auto
h2 <- ggplot(data, aes(x = carwidth)) +
  geom_histogram(binwidth = 1, fill = "lightgreen", color = "black") +
  labs(title = "Histograma del Ancho del auto", x = "Ancho (carwidth)", y
= "Frecuencia") +
  theme_minimal()

# Histograma del Precio del auto
h3 <- ggplot(data, aes(x = carwidth)) +
  geom_histogram(binwidth = 1, fill = "lightgreen", color = "black") +
  labs(title = "Histograma del Preciodel auto", x = "Precio(price)", y =
"Frecuencia") +
  theme_minimal()

grid.arrange(h1, h2, h3, nrow = 1)
```



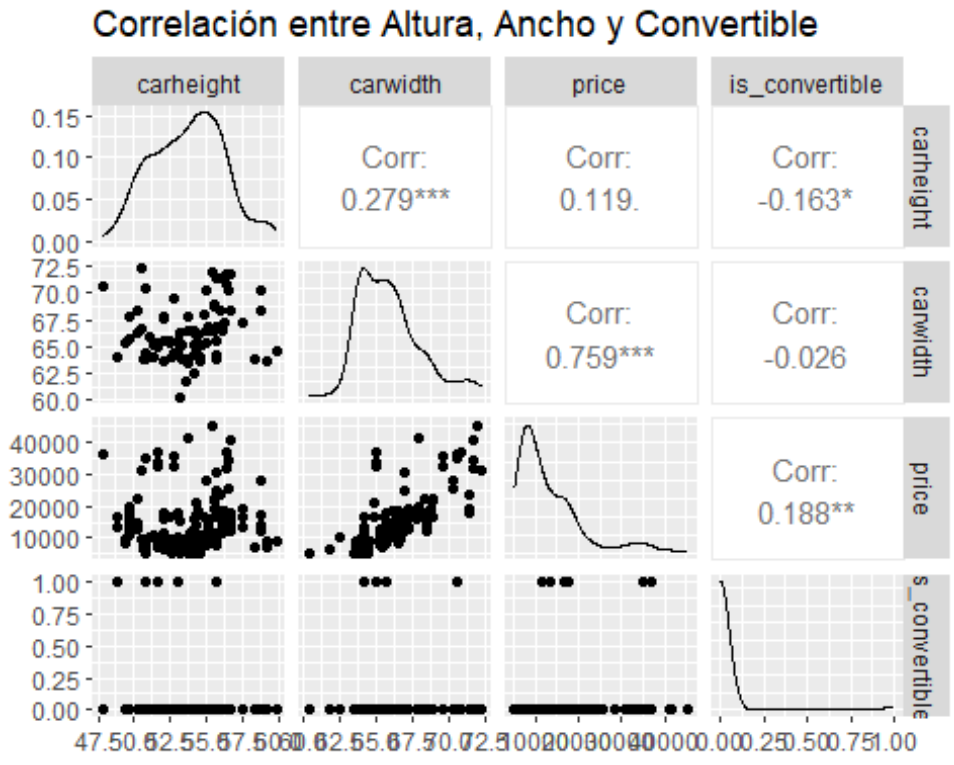
La altura del auto tiene una distribución más equilibrada y simétrica, mientras que tanto el ancho como el precio presentan distribuciones asimétricas hacia la derecha, con unos pocos autos más anchos y costosos que generan colas en esas variables. Estas distribuciones sugieren que la mayoría de los autos en el conjunto de datos son de tamaños y precios promedio, con pocos autos que se desvían mucho de esas características.

Diagramas de dispersión y correlación por pares

```
# Diagrama de dispersión entre Altura y Ancho del auto
ggplot(data, aes(x = carwidth, y = carheight)) +
  geom_point(color = "blue", alpha = 0.6) +
  labs(title = "Diagrama de dispersión: Ancho vs Altura del auto",
       x = "Ancho (carwidth)", y = "Altura (carheight)") +
  theme_minimal()
```




```
# Pares de correlación (scatter plot) entre Altura, Ancho y si es convertible
ggpairs(relevant_columns, title = "Correlación entre Altura, Ancho y Convertible")
```



* El ancho del auto está más relacionado con el precio, mientras que la altura tiene una relación débil con las demás variables.

- Es convertible no influye mucho en el ancho o altura, pero está débilmente relacionado con precios más altos.
- La dispersión de los datos indica que el ancho es un factor clave en la predicción del precio del automóvil.

Variables categóricas

Distribución de los datos (diagramas de barras, diagramas de pastel)

```
library(dplyr)
```

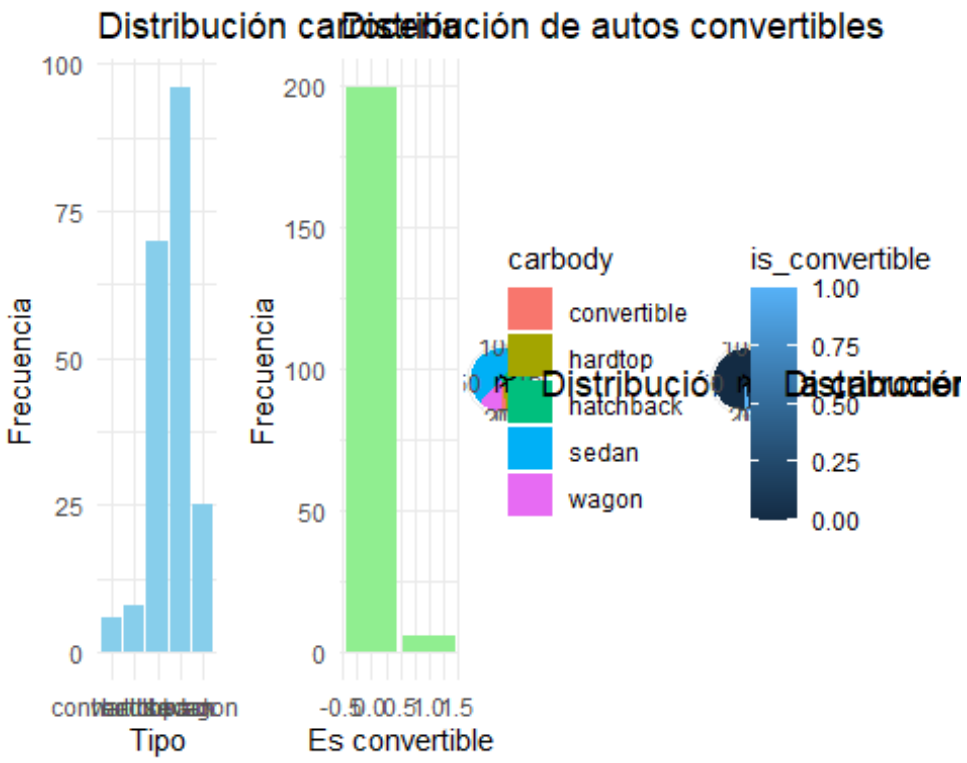
```
## Warning: package 'dplyr' was built under R version 4.3.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:gridExtra':
##
##     combine
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

# Diagramas de barras para la variable categórica "carbody"
d1 <- ggplot(data, aes(x = carbody)) +
  geom_bar(fill = "skyblue") +
  labs(title = "Distribución carrocería", x = "Tipo", y = "Frecuencia") +
  theme_minimal()

# Diagramas de barras para la variable "is_convertible" (convertible o
no)
d2 <- ggplot(data, aes(x = is_convertible)) +
  geom_bar(fill = "lightgreen") +
  labs(title = "Distribución de autos convertibles", x = "Es
convertible", y = "Frecuencia") +
  theme_minimal()

# Diagrama de pastel para la variable "carbody"
carbody_count <- data %>%
  count(carbody)

d3 <- ggplot(carbody_count, aes(x = "", y = n, fill = carbody)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start = 0) +
  labs(title = "Distribución de la carrocería") +
  theme_minimal()

# Diagrama de pastel para la variable "is_convertible"
convertible_count <- data %>%
  count(is_convertible)

d4 <- ggplot(convertible_count, aes(x = "", y = n, fill =
is_convertible)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start = 0) +
  labs(title = "Distribución de autos convertibles") +
  theme_minimal()

grid.arrange(d1, d2, d3, d4, nrow = 1)
```



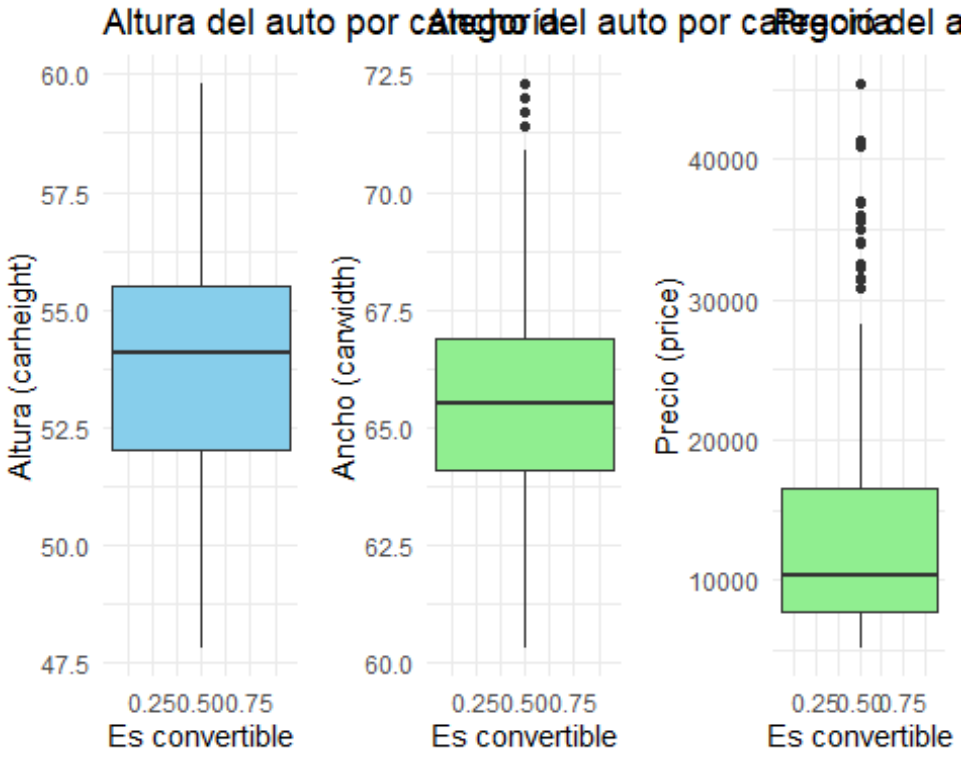
* La mayoría de los autos en este conjunto de datos son sedanes o hatchbacks, y los convertibles representan una fracción muy pequeña del total.

- Se confirma nuevamente que los autos convertibles no son una característica predominante, por lo que es probable que la variable “convertible” tenga una influencia limitada en los análisis posteriores, dado que es poco representativa.

Boxplot por categoría de las variables cuantitativas

```
# Boxplot de Altura del auto por categoría (si es convertible o no)
b1 <- ggplot(data, aes(x = is_convertible, y = carheight)) +
  geom_boxplot(fill = "skyblue") +
  labs(title = "Altura del auto por categoría", x = "Es convertible", y =
"Altura (carheight)") +
  theme_minimal()

# Boxplot de Ancho del auto por categoría (si es convertible o no)
b2 <- ggplot(data, aes(x = is_convertible, y = carwidth)) +
  geom_boxplot(fill = "lightgreen") +
  labs(title = "Ancho del auto por categoría", x = "Es convertible", y =
"Ancho (carwidth)") +
  theme_minimal()

# Boxplot del Precio del auto por categoría (si es convertible o no)
b3 <- ggplot(data, aes(x = is_convertible, y = price)) +
  geom_boxplot(fill = "lightgreen") +
  labs(title = "Precio del auto por categoría", x = "Es convertible", y =
```

```
"Precio (price)") +
  theme_minimal()

grid.arrange(b1, b2, b3, nrow = 1)

## Warning: Continuous x aesthetic
## ℹ did you forget `aes(group = ...)`?
## Continuous x aesthetic
## ℹ did you forget `aes(group = ...)`?
## Continuous x aesthetic
## ℹ did you forget `aes(group = ...)`?
```



* La variable convertible no parece estar fuertemente relacionada con la altura o el ancho del auto, pero tiene cierta relación con el precio, ya que los autos convertibles tienden a ser más costosos, como se refleja en los valores atípicos en el gráfico de precios.

- El comportamiento de los precios sugiere que ser convertible está asociado con algunos autos de gama alta, aunque en su mayoría los autos tienen precios más bajos.

Parte 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.

Para cada uno de los modelos propuestos:

Realiza la regresión entre las variables involucradas

```
# Crear un modelo de regresión lineal simple: precio en función del ancho del auto (carwidth)
modelo1 <- lm(price ~ carwidth, data = data)
summary(modelo1)
```

```
##
## Call:
## lm(formula = price ~ carwidth, data = data)
##
## 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
```

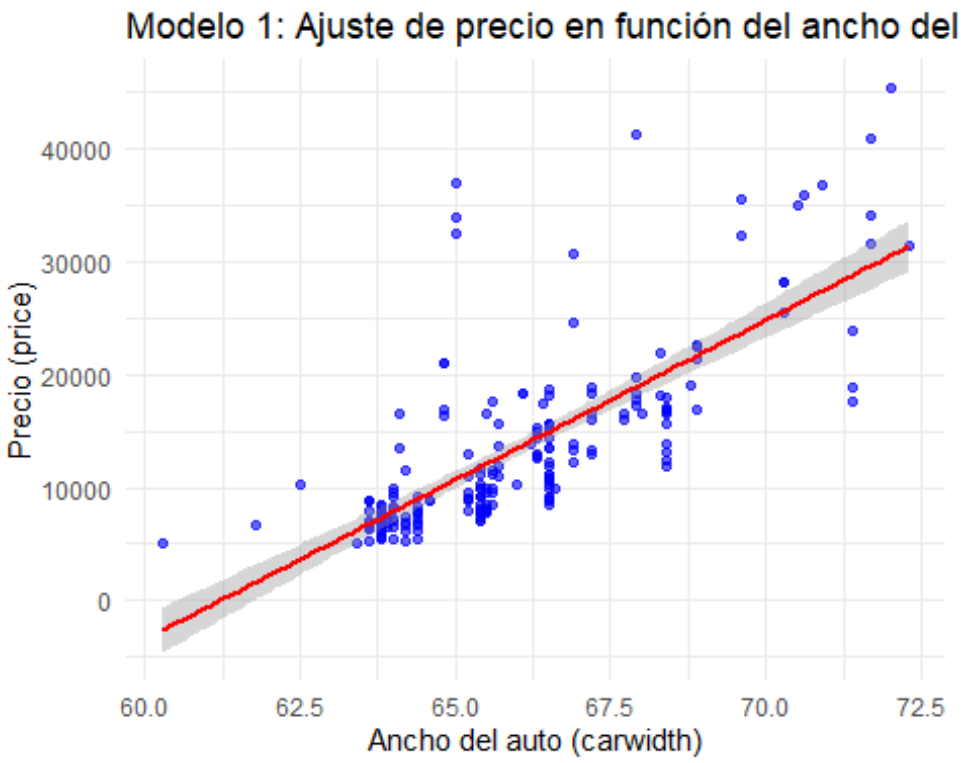
```
# Crear un modelo de regresión lineal múltiple: precio en función del ancho del auto (carwidth) y si es convertible (is_convertible)

data$is_convertible <- ifelse(data$carbody == "convertible", 1, 0)

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

```
##
## Call:
## lm(formula = price ~ carwidth + is_convertible, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10921.7  -2505.8   -756.8    971.2  23624.4
##
```

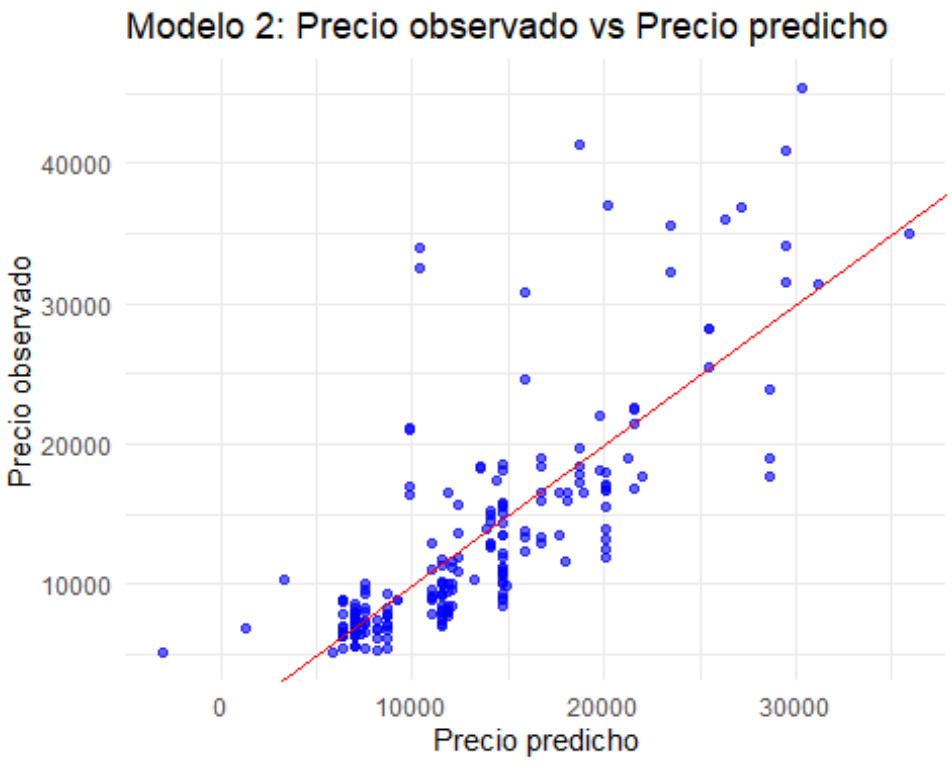
```
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -174725.7    10660.4 -16.390  < 2e-16 ***
## carwidth          2848.1      161.6  17.621  < 2e-16 ***
## is_convertible    9825.5     2052.1   4.788 3.25e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4951 on 202 degrees of freedom
## Multiple R-squared:  0.6197, Adjusted R-squared:  0.616
## F-statistic: 164.6 on 2 and 202 DF,  p-value: < 2.2e-16
```

- El Modelo 2 es preferible ya que explica mejor la variabilidad en el precio de los autos al incluir tanto el ancho del auto como si el auto es convertible. Ambos factores tienen un impacto significativo en el precio.

```
# Comparar los modelos utilizando el AIC y R-squared
cat("Modelo 1 - AIC:", AIC(modelo1), "\n")

## Modelo 1 - AIC: 4094.769

cat("Modelo 2 - AIC:", AIC(modelo2), "\n")

## Modelo 2 - AIC: 4074.731

cat("Modelo 1 - R-squared:", summary(modelo1)$r.squared, "\n")

## Modelo 1 - R-squared: 0.5765749

cat("Modelo 2 - R-squared:", summary(modelo2)$r.squared, "\n")

## Modelo 2 - R-squared: 0.6197328

# Verificar los ajustes con visualización gráfica de ambos modelos
# Gráfico del ajuste del modelo 1
ggplot(data, aes(x = carwidth, y = price)) +
  geom_point(color = "blue", alpha = 0.6) +
  geom_smooth(method = "lm", formula = y ~ x, color = "red") +
  labs(title = "Modelo 1: Ajuste de precio en función del ancho del
auto",
       x = "Ancho del auto (carwidth)", y = "Precio (price)") +
  theme_minimal()
```




```r
# Gráfico del ajuste del modelo 2 (regresión múltiple)
# Utilizamos predicciones del modelo 2 para visualizarlo
data$predicted_price <- predict(modelo2)

ggplot(data, aes(x = predicted_price, y = price)) +
  geom_point(color = "blue", alpha = 0.6) +
  geom_abline(slope = 1, intercept = 0, color = "red") +
  labs(title = "Modelo 2: Precio observado vs Precio predicho",
       x = "Precio predicho", y = "Precio observado") +
  theme_minimal()
```



* Mejor Modelo: El Modelo 2 es mejor que el Modelo 1 según el valor más bajo del AIC y el R^2 más alto. El modelo que incluye tanto el ancho del auto como si el auto es convertible ofrece una mejor explicación de la variabilidad del precio.

- Aunque el ancho del auto es un buen predictor del precio, incluir si el auto es convertible o no mejora significativamente el ajuste del modelo, lo que se refleja en el aumento del R^2 y la reducción del AIC.

##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)

Significancia del modelo

H_0:El modelo no es significativo (no hay relación lineal entre las variables predictoras y la variable depe
\
H_1:El modelo es significativo (al menos una variable predictora tiene relación lineal con la variable depe

Coeficientes individuales $H_0: \beta_i =$ 0 (El coeficiente de la variable i es cero, no tiene un impacto significativo) \ $H_1: \beta_i \neq$ 0 (El coeficiente de la variable i no es cero, tiene un impacto significativo)

```
# Definir el nivel de significancia
alpha <- 0.04

# Función para validar la significancia del modelo
validate_model_significance <- function(model, alpha) {
```

```r
  # Resumen del modelo
  model_summary <- summary(model)

  # Valor p global (F-statistic)
  f_statistic <- model_summary$fstatistic
  p_value_global <- pf(f_statistic[1], f_statistic[2], f_statistic[3],
lower.tail = FALSE)

  # Validar la significancia global del modelo
  if (p_value_global < alpha) {
    cat("El modelo es globalmente significativo (p-value global =",
p_value_global, ")\n")
  } else {
    cat("El modelo NO es globalmente significativo (p-value global =",
p_value_global, ")\n")
  }

  # Ver los p-values de los coeficientes individuales
  cat("P-values de los coeficientes individuales:\n")
  print(model_summary$coefficients[, "Pr(>|t|)"])

  # Validar la significancia de cada coeficiente individual
  significant_vars <- model_summary$coefficients[, "Pr(>|t|)"] < alpha
  if (any(significant_vars)) {
    cat("Las siguientes variables son significativas a nivel", alpha,
":\n")
    print(names(significant_vars)[significant_vars])
  } else {
    cat("Ninguna variable es significativa a nivel", alpha, "\n")
  }
}

# Validar la significancia del modelo 1
cat("Validación del Modelo 1:\n")
```

```
## Validación del Modelo 1:
```

```r
validate_model_significance(modelo1, alpha)
```

```
## El modelo es globalmente significativo (p-value global = 9.627438e-40
)
## P-values de los coeficientes individuales:
##  (Intercept)     carwidth
## 4.580237e-36 9.627438e-40
## Las siguientes variables son significativas a nivel 0.04 :
## [1] "(Intercept)" "carwidth"
```

```r
# Validar la significancia del modelo 2
cat("\nValidación del Modelo 2:\n")
```

```
##
## Validación del Modelo 2:

validate_model_significance(modelo2, alpha)

## El modelo es globalmente significativo (p-value global = 3.881277e-43
)
## P-values de los coeficientes individuales:
##    (Intercept)       carwidth is_convertible
##   5.859464e-39   1.042510e-42   3.251192e-06
## Las siguientes variables son significativas a nivel 0.04 :
## [1] "(Intercept)"    "carwidth"       "is_convertible"
```

- Ambos modelos son globalmente significativos, lo que significa que las variables incluidas explican de manera significativa la variación en el precio.

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

$$H_0: \beta_i = 0 (\text{el coeficiente no es significativo})$$

$$H_1: \beta_i \neq 0 (\text{el coeficiente es significativo})$$

```
# Nivel de significancia
alpha <- 0.04

# Función para validar la significancia de cada coeficiente
validate_coeff_significance <- function(model, alpha) {
  coef_summary <- summary(model)$coefficients
  p_values <- coef_summary[, "Pr(>|t|)"]

  cat("\nValidación de los coeficientes con alfa =", alpha, "\n")

  for (i in 1:length(p_values)) {
    if (p_values[i] < alpha) {
      cat("El coeficiente", rownames(coef_summary)[i], "es significativo
con un valor p =", p_values[i], "\n")
    } else {
      cat("El coeficiente", rownames(coef_summary)[i], "NO es
significativo con un valor p =", p_values[i], "\n")
    }
  }

  # Calcular el valor frontera (t-critical)
  df <- summary(model)$df[2]
  t_critical <- qt(1 - alpha / 2, df = df)
  cat("\nValor frontera (t-critical) para alfa =", alpha, "con", df,
"grados de libertad es t =", t_critical, "\n")
}
```

```r
# Validar la significancia de los coeficientes del Modelo 1
cat("Modelo 1: Regresión lineal simple con interacción\n")
```

```
## Modelo 1: Regresión lineal simple con interacción
```

```r
validate_coeff_significance(modelo1, alpha)
```

```
##
## Validación de los coeficientes con alfa = 0.04
## El coeficiente (Intercept) es significativo con un valor p =
4.580237e-36
## El coeficiente carwidth es significativo con un valor p = 9.627438e-40
##
## Valor frontera (t-critical) para alfa = 0.04 con 203 grados de
libertad es t = 2.067029
```

```r
# Validar la significancia de los coeficientes del Modelo 2
cat("\nModelo 2: Regresión lineal múltiple\n")
```

```
##
## Modelo 2: Regresión lineal múltiple
```

```r
validate_coeff_significance(modelo2, alpha)
```

```
##
## Validación de los coeficientes con alfa = 0.04
## El coeficiente (Intercept) es significativo con un valor p =
5.859464e-39
## El coeficiente carwidth es significativo con un valor p = 1.04251e-42
## El coeficiente is_convertible es significativo con un valor p =
3.251192e-06
##
## Valor frontera (t-critical) para alfa = 0.04 con 202 grados de
libertad es t = 2.067096
```

Modelo 1: * Tanto el intercepto como el coeficiente de carwidth son significativos, lo que indica que el ancho del auto es un predictor importante para el precio del auto.

Modelo 2: * Incluir la variable is_convertible en el modelo mejora la capacidad del modelo para explicar el precio, ya que ambos coeficientes son estadísticamente significativos.

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

```r
# Extraer el R-squared de ambos modelos
r_squared_modelo1 <- summary(modelo1)$r.squared
r_squared_modelo2 <- summary(modelo2)$r.squared

# Calcular el porcentaje de variación explicada por ambos modelos
porcentaje_variacion_modelo1 <- r_squared_modelo1 * 100
porcentaje_variacion_modelo2 <- r_squared_modelo2 * 100

# Mostrar el porcentaje de variación explicada para cada modelo
```

```r
cat("El porcentaje de variación explicada por el Modelo 1 es:",
porcentaje_variacion_modelo1, "%\n")
```

```
## El porcentaje de variación explicada por el Modelo 1 es: 57.65749 %
```

```r
cat("El porcentaje de variación explicada por el Modelo 2 es:",
porcentaje_variacion_modelo2, "%\n")
```

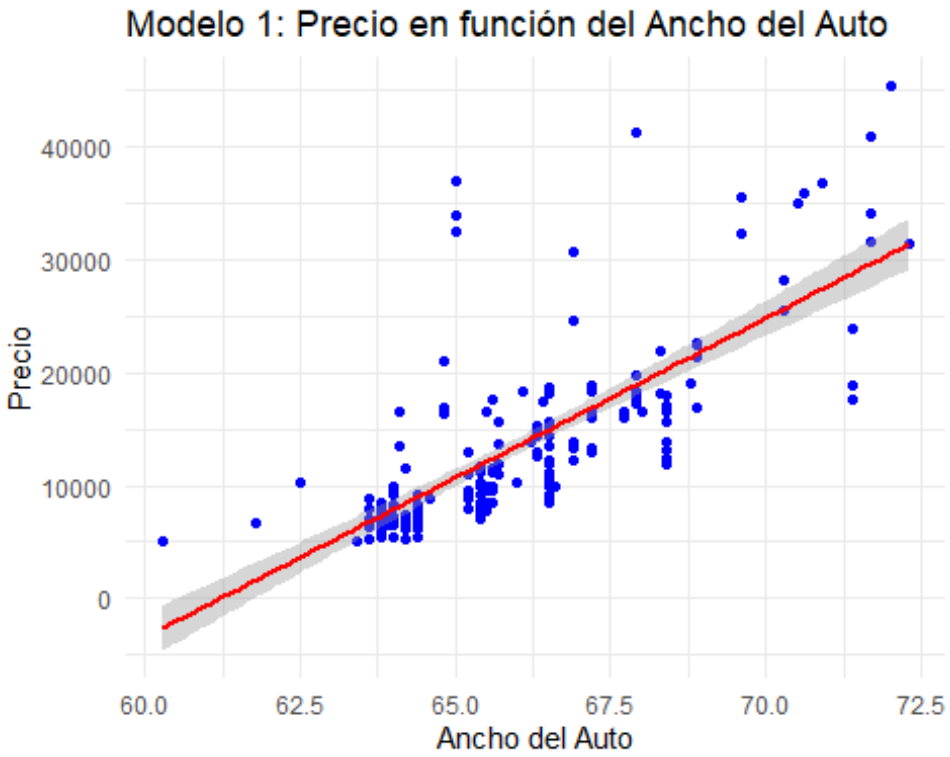
```
## El porcentaje de variación explicada por el Modelo 2 es: 61.97328 %
```

- El Modelo 2 explica un mayor porcentaje de la variación en el precio (casi un 62%) al considerar tanto el ancho del auto como si el auto es convertible, lo que lo convierte en un mejor modelo para predecir el precio del automóvil en comparación con el Modelo 1.

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

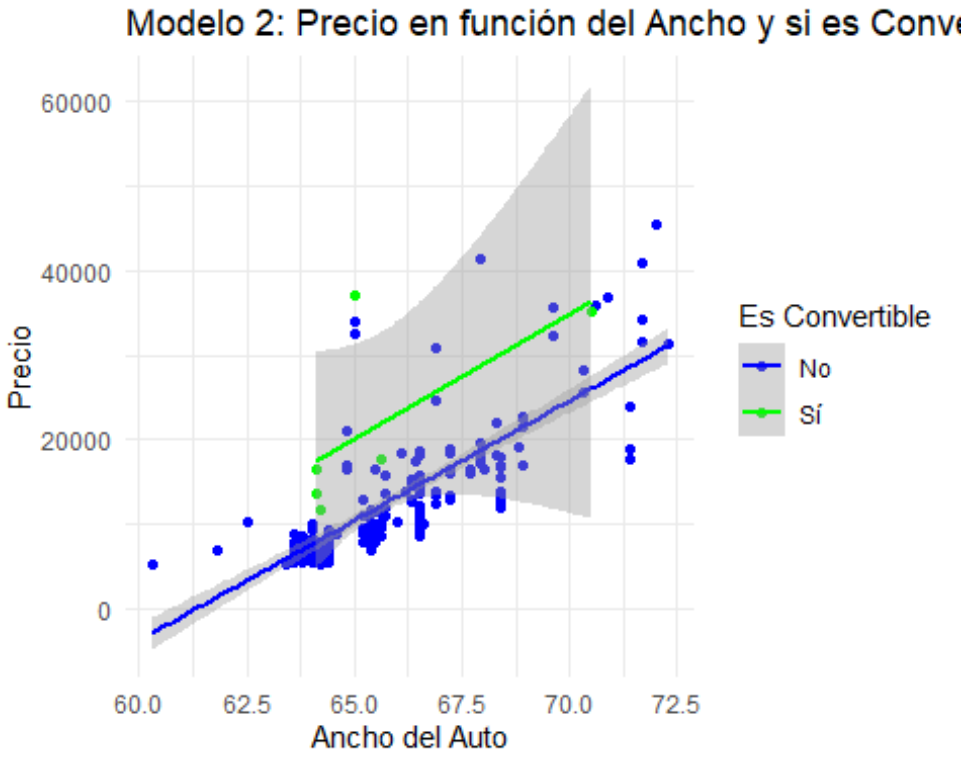
```r
# Diagrama de dispersión y recta de mejor ajuste para el Modelo 1
ggplot(data, aes(x = carwidth, y = price)) +
  geom_point(color = "blue") +  # Puntos de los datos
  geom_smooth(method = "lm", color = "red", se = TRUE) +  # Recta de
mejor ajuste
  labs(title = "Modelo 1: Precio en función del Ancho del Auto", x =
"Ancho del Auto", y = "Precio") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



```r
# Diagrama de dispersión y recta de mejor ajuste para el Modelo 2
ggplot(data, aes(x = carwidth, y = price, color =
as.factor(is_convertible))) +
  geom_point() +  # Puntos de los datos
  geom_smooth(method = "lm", se = TRUE) +  # Recta de mejor ajuste
  labs(title = "Modelo 2: Precio en función del Ancho y si es
Convertible",
       x = "Ancho del Auto",
       y = "Precio",
       color = "Es Convertible") +
  scale_color_manual(values = c("blue", "green"), labels = c("No", "Sí"))
+
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



* Ambos modelos muestran que el ancho del auto es un predictor significativo del precio, y que los autos convertibles tienen precios generalmente más altos que los no convertibles. Sin embargo, el Modelo 2 es más robusto al incluir la variable convertible, lo que mejora la capacidad de predicción del precio.

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

- En el contexto del problema, se están analizando dos modelos para identificar qué variables son significativas al predecir el precio de un automóvil, y cómo estas variables explican la variación en el precio.

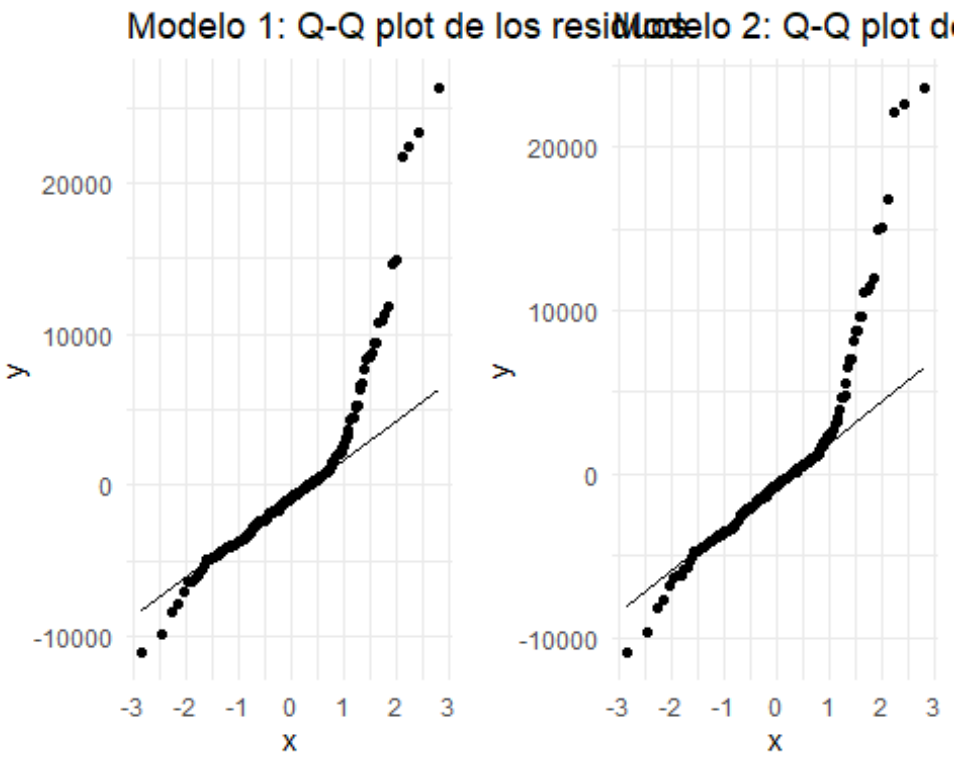
- El ancho del auto es un factor importante para predecir el precio de un automóvil en el mercado estadounidense. La empresa debería considerar ofrecer vehículos más anchos, ya que tienden a tener precios más altos.
- El hecho de que un auto se aconvertible es un atributo de lujo que incrementa significativamente el precio de los autos.
- Aunque ambos modelos explican una parte importante de la variabilidad en los precios, queda claro que hay otros factores adicionales que también podrían estar afectando el precio, como el tipo de motor, el rendimiento, entre otros.

Analiza la validez de los modelos propuestos:

Normalidad de los residuos

H_0: Los datos provienen de una población normal H_1: Los datos no provienen de una población normal.

```
# Calcular los residuos de ambos modelos
residuos_modelo1 <- resid(modelo1)
residuos_modelo2 <- resid(modelo2)

# Gráfico Q-Q para la normalidad de los residuos del Modelo 1
q1 <- qq_plot_modelo1 <- ggplot(data.frame(residuos_modelo1), aes(sample
= residuos_modelo1)) +
  stat_qq() +
  stat_qq_line() +
  labs(title = "Modelo 1: Q-Q plot de los residuos") +
  theme_minimal()


# Gráfico Q-Q para la normalidad de los residuos del Modelo 2
q2 <- qq_plot_modelo2 <- ggplot(data.frame(residuos_modelo2), aes(sample
= residuos_modelo2)) +
  stat_qq() +
  stat_qq_line() +
  labs(title = "Modelo 2: Q-Q plot de los residuos") +
  theme_minimal()

grid.arrange(q1, q2, ncol = 2)
```



* Ambos modelos muestran que los residuos no siguen una distribución normal de manera adecuada, especialmente en los valores más extremos (outliers).

```
# Prueba de Shapiro-Wilk para normalidad de los residuos
shapiro_test_modelo1 <- shapiro.test(residuos_modelo1)
shapiro_test_modelo2 <- shapiro.test(residuos_modelo2)

# Imprimir los resultados de las pruebas de Shapiro-Wilk
cat("Prueba de Shapiro-Wilk para el Modelo 1:\n")
```

```
## Prueba de Shapiro-Wilk para el Modelo 1:
```

```
print(shapiro_test_modelo1)
```

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

```
cat("\nPrueba de Shapiro-Wilk para el Modelo 2:\n")
```

```
##
## Prueba de Shapiro-Wilk para el Modelo 2:
```

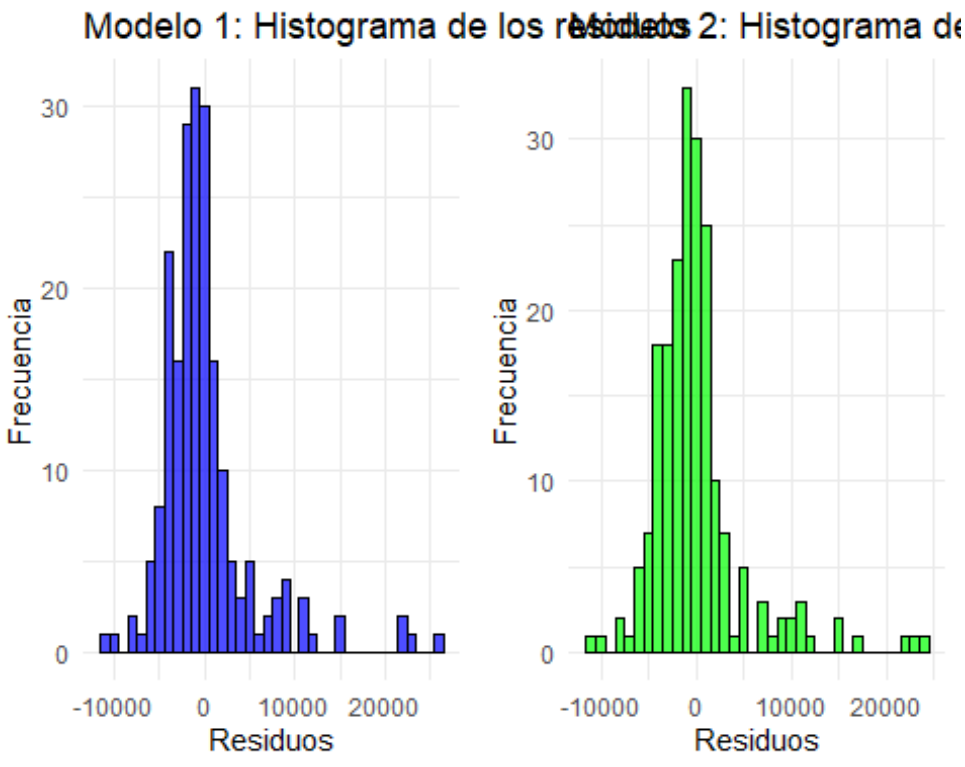
```
print(shapiro_test_modelo2)
```

```
##
##  Shapiro-Wilk normality test
```

```
##
## data:  residuos_modelo2
## W = 0.81033, p-value = 4.587e-15
```

- Ambos modelos, según la prueba de Shapiro-Wilk, no cumplen con la suposición de normalidad de los residuos. Esto refuerza la interpretación de los gráficos Q-Q, que también mostraban un comportamiento no normal.

```
# Histograma de los residuos del Modelo 1
h1 <- ggplot(data.frame(residuos_modelo1), aes(x = residuos_modelo1)) +
  geom_histogram(binwidth = 1000, fill = "blue", color = "black", alpha =
0.7) +
  labs(title = "Modelo 1: Histograma de los residuos", x = "Residuos", y
= "Frecuencia") +
  theme_minimal()

# Histograma de los residuos del Modelo 2
h2 <- ggplot(data.frame(residuos_modelo2), aes(x = residuos_modelo2)) +
  geom_histogram(binwidth = 1000, fill = "green", color = "black", alpha
= 0.7) +
  labs(title = "Modelo 2: Histograma de los residuos", x = "Residuos", y
= "Frecuencia") +
  theme_minimal()

grid.arrange(h1, h2, ncol = 2)
```



* Ambos modelos presentan una concentración de residuos cerca de cero, lo cual es un buen signo, pero los histogramas revelan asimetría y outliers. Esto confirma que los

residuos no siguen una distribución normal, lo que concuerda con los resultados previos de los Q-Q plots y las pruebas de Shapiro-Wilk.

Verificación de media cero

$H0$: $\mu = 0$ $H1$: $\mu \neq 0$

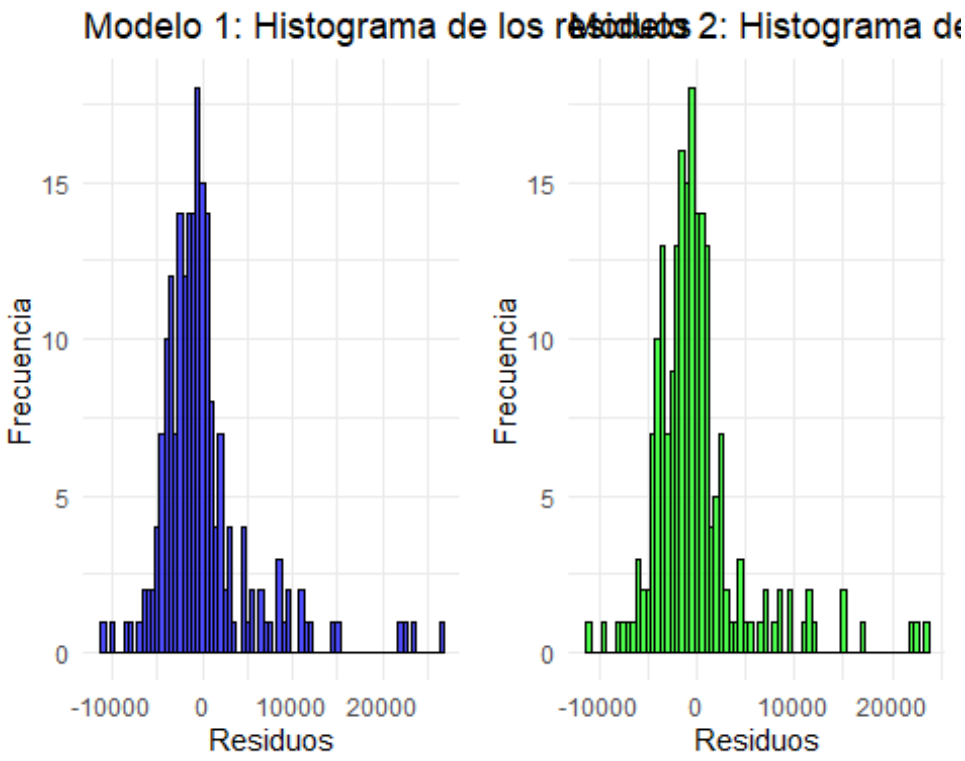
```
# Calcular la media de los residuos de ambos modelos
media_residuos_modelo1 <- mean(residuos_modelo1)
media_residuos_modelo2 <- mean(residuos_modelo2)

# Mostrar las medias de los residuos
cat("La media de los residuos del Modelo 1 es:", media_residuos_modelo1,
"\n")
```

```
## La media de los residuos del Modelo 1 es: -3.551338e-13
```

```
cat("La media de los residuos del Modelo 2 es:", media_residuos_modelo2,
"\n")
```

```
## La media de los residuos del Modelo 2 es: -8.874635e-14
```

```
# Graficar los residuos del Modelo 1
m1 <- ggplot(data.frame(residuos_modelo1), aes(x = residuos_modelo1)) +
  geom_histogram(binwidth = 500, fill = "blue", color = "black", alpha =
0.7) +
  labs(title = "Modelo 1: Histograma de los residuos", x = "Residuos", y
= "Frecuencia") +
  theme_minimal()

# Graficar los residuos del Modelo 2
m2 <- ggplot(data.frame(residuos_modelo2), aes(x = residuos_modelo2)) +
  geom_histogram(binwidth = 500, fill = "green", color = "black", alpha =
0.7) +
  labs(title = "Modelo 2: Histograma de los residuos", x = "Residuos", y
= "Frecuencia") +
  theme_minimal()

grid.arrange(m1, m2, ncol = 2)
```



* Aunque los residuos tienen una media cercana a cero, lo cual es deseable, la distribución no es normal y existen valores extremos. Esto sugiere que los modelos están capturando parte de la relación entre las variables, sin embargo podrían mejorarse con la inclusión de variables adicionales o una transformación de los datos.

Homocedasticidad, linealidad e independencia

H_0: La varianza de los errores es constante (homocedasticidad) H_1: La varianza de los errores no es constante (heterocedasticidad)

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

```
library(readr)
library(car)          # Para la prueba de Breusch-Pagan y Durbin-Watson
```

```
## Warning: package 'car' was built under R version 4.3.3
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.3.3
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```r
library(lmtest)       # Para la prueba de homocedasticidad
```

```
## Warning: package 'lmtest' was built under R version 4.3.3
```

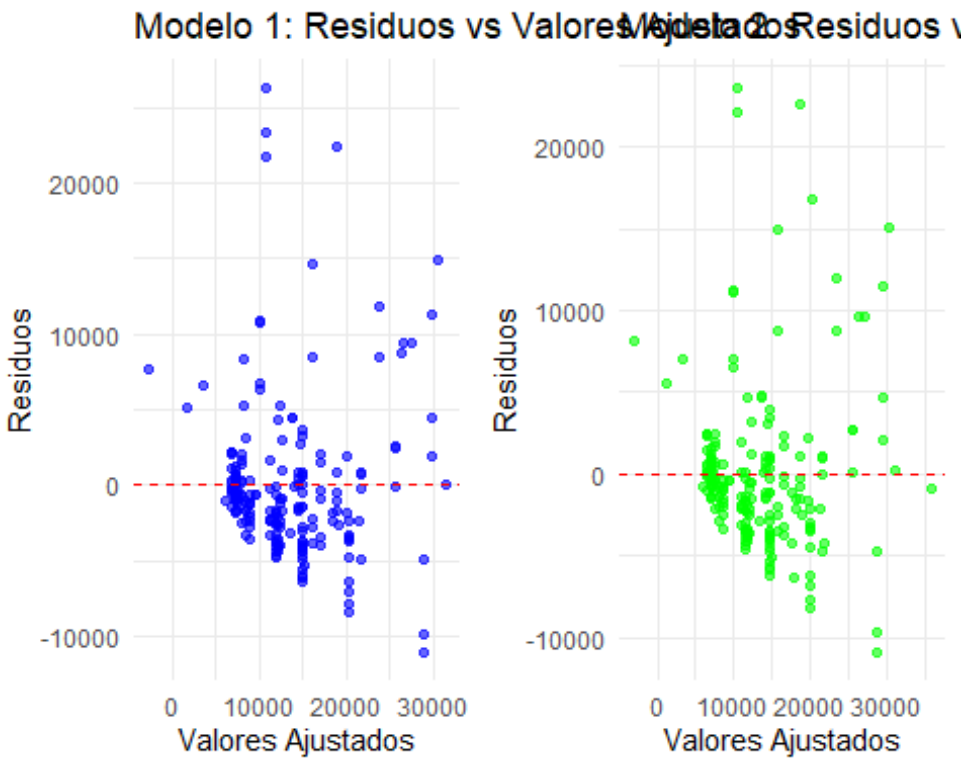
```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 4.3.3
```

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

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

```r
# 1. Evaluación de Homocedasticidad y Linealidad
# Gráfico de residuos vs valores ajustados para Modelo 1
e1 <- ggplot(data, aes(x = fitted(modelo1), y = resid(modelo1))) +
  geom_point(color = "blue", alpha = 0.6) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(title = "Modelo 1: Residuos vs Valores Ajustados", x = "Valores
Ajustados", y = "Residuos") +
  theme_minimal()

# Gráfico de residuos vs valores ajustados para Modelo 2
e2 <- ggplot(data, aes(x = fitted(modelo2), y = resid(modelo2))) +
  geom_point(color = "green", alpha = 0.6) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(title = "Modelo 2: Residuos vs Valores Ajustados", x = "Valores
Ajustados", y = "Residuos") +
  theme_minimal()

grid.arrange(e1, e2, ncol = 2)
```



* Ambos modelos muestran indicios de heterocedasticidad (la varianza de los residuos no es constante), especialmente en los valores ajustados más altos.

- En términos de linealidad, no se observan patrones claros de curvatura, lo que indica que la relación entre los valores ajustados y los residuos es aproximadamente lineal. Sin embargo, la presencia de outliers y la dispersión en los residuos sugiere que los modelos pueden no estar capturando completamente la relación entre las variables predictoras y el precio del automóvil.

```
# 2. Prueba de homocedasticidad (Breusch-Pagan)
cat("Prueba de Breusch-Pagan para Modelo 1:\n")

## Prueba de Breusch-Pagan para Modelo 1:

bptest(modelo1)

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

cat("\nPrueba de Breusch-Pagan para Modelo 2:\n")

##
## Prueba de Breusch-Pagan para Modelo 2:

bptest(modelo2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  modelo2
## BP = 7.2341, df = 2, p-value = 0.02686
```

- Ambos modelos presentan heterocedasticidad, lo que significa que la varianza de los residuos no es constante.

```
# 3. Prueba de independencia de los errores (Durbin-Watson)
cat("\nPrueba de Durbin-Watson para Modelo 1:\n")

##
## Prueba de Durbin-Watson para Modelo 1:

durbinWatsonTest(modelo1)

##  lag Autocorrelation D-W Statistic p-value
##    1       0.6782376     0.6382283       0
##  Alternative hypothesis: rho != 0

cat("\nPrueba de Durbin-Watson para Modelo 2:\n")

##
## Prueba de Durbin-Watson para Modelo 2:

durbinWatsonTest(modelo2)

##  lag Autocorrelation D-W Statistic p-value
##    1       0.6705619     0.6551122       0
##  Alternative hypothesis: rho != 0
```

- Tanto el Modelo 1 como el Modelo 2 presentan una fuerte autocorrelación positiva en los residuos, sugieriendo así que, los residuos no son independientes.

Interpreta cada uno de los analisis que realizaste

- Normalidad: Ninguno de los dos modelos cumple con la suposición de residuos normalmente distribuidos.
- Heterocedasticidad: En ambas pruebas de Breusch-Pagan, se rechazó la hipótesis nula de homocedasticidad, lo que indica que la varianza de los residuos no es constante (heterocedasticidad) para ambos modelos.
- Independencia: La prueba de Durbin-Watson mostró que ambos modelos presentan autocorrelación positiva significativa en los residuos, lo que sugiere que los errores no son independientes.
- Linealidad: En los gráficos de residuos vs valores ajustados, no se observaron patrones claros que violen la suposición de linealidad. Sin embargo, los residuos mostraron dispersión a lo largo de los valores ajustados, lo que refuerza la evidencia de heterocedasticidad. La linealidad parece estar

razonablemente bien representada en los modelos, aunque la presencia de heterocedasticidad y autocorrelación sugiere que el modelo no captura completamente la relación entre las variables.

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

- El Modelo 2 es el mejor modelo porque explica más variabilidad en los precios de los automóviles y considera una variable adicional significativa (si es convertible o no), lo que mejora la precisión en las predicciones de precio. Aunque ambos modelos presentan ciertos problemas con la homocedasticidad y la autocorrelación, el Modelo 2 ofrece un ajuste superior y mayor capacidad explicativa.

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

- Las variables significativas que mejor predicen el precio de un automóvil en el mercado estadounidense, de acuerdo a este análisis, son:
- Ancho del automóvil (carwidth): Existe una relación positiva significativa entre el ancho del automóvil y su precio. A mayor ancho, mayor es el precio del auto.
- Si el automóvil es convertible o no (is_convertible): Los automóviles convertibles tienen un precio significativamente mayor en comparación con los autos no convertibles.

Parte 3. Intervalos de predicción y confianza

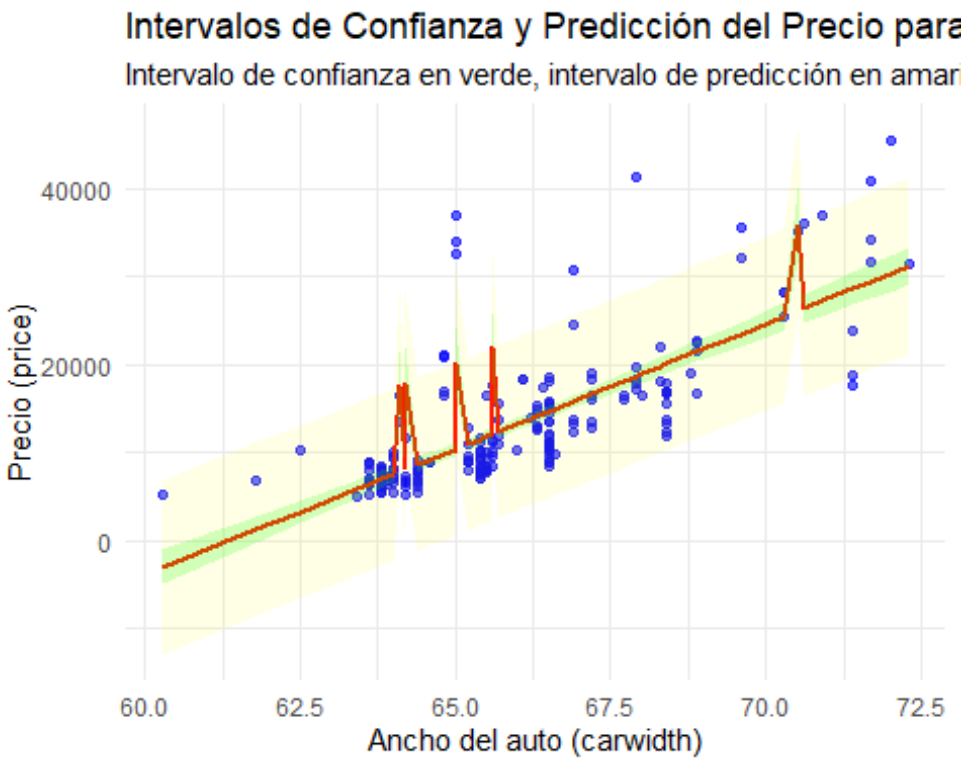
Con los datos de las variables asignadas construye la gráfica de los intervalos de confianza y predicción para la estimación y predicción del precio para el mejor modelo seleccionado:

Calcula los intervalos para la variable Y

```
# Predicción con intervalos de confianza y predicción
# Confianza del 95%
predicciones <- predict(modelo2, interval = "confidence", level = 0.95)
predicciones_pred <- predict(modelo2, interval = "prediction", level =
0.95)

## Warning in predict.lm(modelo2, interval = "prediction", level = 0.95):
predictions on current data refer to _future_ responses
```

```r
# Combinar los datos reales con las predicciones y los intervalos
data$predicted_price <- predicciones[, "fit"]
data$conf_lwr <- predicciones[, "lwr"]
data$conf_upr <- predicciones[, "upr"]

data$pred_lwr <- predicciones_pred[, "lwr"]
data$pred_upr <- predicciones_pred[, "upr"]

# Gráfica de los intervalos de confianza y predicción
ggplot(data, aes(x = carwidth, y = price)) +
  geom_point(color = "blue", alpha = 0.6) +
  geom_line(aes(y = predicted_price), color = "red", size = 1) +
  geom_ribbon(aes(ymin = conf_lwr, ymax = conf_upr), alpha = 0.2, fill =
"green") +  # Intervalo de confianza
  geom_ribbon(aes(ymin = pred_lwr, ymax = pred_upr), alpha = 0.1, fill =
"yellow") +  # Intervalo de predicción
  labs(title = "Intervalos de Confianza y Predicción del Precio para el
Modelo 2",
       subtitle = "Intervalo de confianza en verde, intervalo de
predicción en amarillo",
       x = "Ancho del auto (carwidth)", y = "Precio (price)") +
  theme_minimal()
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2
3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning
was
## generated.
```



* El Modelo 2 proporciona una buena predicción general del precio de los automóviles en función del ancho del auto y si es convertible o no. Los intervalos de confianza (verde) son ajustados, lo que indica que el modelo tiene precisión en las estimaciones promedio. Sin embargo, los intervalos de predicción (amarillo) muestran más dispersión, lo que refleja la variabilidad de los precios individuales.

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.

```
# Separar la base de datos por la variable is_convertible
convertibles <- subset(data, is_convertible == 1)  # Autos convertibles
no_convertibles <- subset(data, is_convertible == 0)  # Autos no convertibles

# Imprimir el número de observaciones en cada categoría
cat("Número de autos convertibles:", nrow(convertibles), "\n")
```

```
## Número de autos convertibles: 6
```

```
cat("Número de autos no convertibles:", nrow(no_convertibles), "\n")
```

```
## Número de autos no convertibles: 199
```

```
# Visualizar las primeras filas de la base de datos de autos convertibles
#head(convertibles)

# Visualizar las primeras filas de la base de datos de autos no convertibles
```

```r
#head(no_convertibles)

# Análisis adicional para cada categoría, ajustar modelos por separado
# Modelo solo para autos convertibles
modelo_convertibles <- lm(price ~ carwidth, data = convertibles)
summary(modelo_convertibles)
```

```
##
## Call:
## lm(formula = price ~ carwidth, data = convertibles)
##
## Residuals:
##     1     2     3     4     5     6
## -4040 -1035 -1271 16850 -4270 -6234
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -170682     112130  -1.522    0.203
## carwidth        2936       1709   1.718    0.161
##
## Residual standard error: 9487 on 4 degrees of freedom
## Multiple R-squared:  0.4247, Adjusted R-squared:  0.2809
## F-statistic: 2.953 on 1 and 4 DF,  p-value: 0.1608
```

```r
# Modelo solo para autos no convertibles
modelo_no_convertibles <- lm(price ~ carwidth, data = no_convertibles)
summary(modelo_no_convertibles)
```

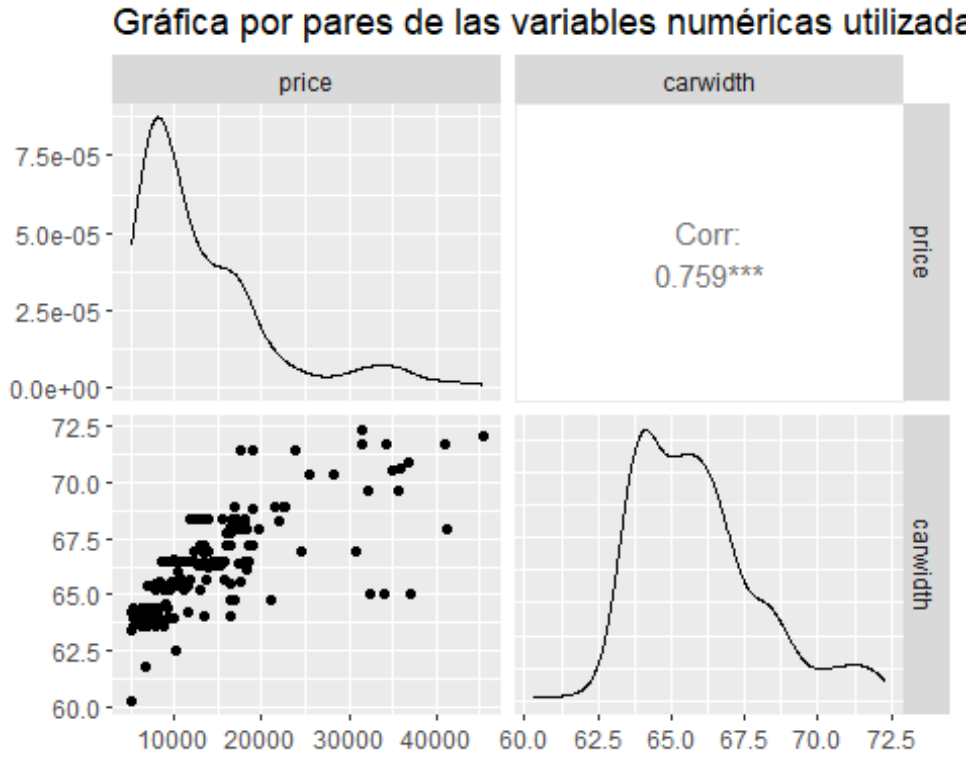
```
##
## Call:
## lm(formula = price ~ carwidth, data = no_convertibles)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10905.3  -2489.2   -735.2    966.2  23621.7
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -174528.3    10569.5  -16.51   <2e-16 ***
## carwidth       2845.1      160.3   17.75   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4827 on 197 degrees of freedom
## Multiple R-squared:  0.6154, Adjusted R-squared:  0.6134
## F-statistic: 315.2 on 1 and 197 DF,  p-value: < 2.2e-16
```

- El modelo para autos no convertibles es claramente más fuerte debido al mayor número de observaciones y a la significancia estadística de los coeficientes.

- El modelo para autos convertibles no es confiable debido a la pequeña muestra y la falta de significancia de las variables. Esto indica que no se puede concluir con confianza que el ancho del auto predice de manera efectiva el precio de los autos convertibles.

Grafica por pares de variables numéricas

```
# Seleccionar solo las variables numéricas que se están utilizando en el modelo (price, carwidth)
variables_numericas <- data[, c("price", "carwidth")]

# Graficar por pares utilizando ggpairs
ggpairs(variables_numericas, title = "Gráfica por pares de las variables numéricas utilizadas en el modelo")
```



* La gráfica muestra que el ancho del auto es un factor que se correlaciona fuertemente con el precio del automóvil. Aunque la correlación es alta, la variabilidad dentro de ciertos rangos del ancho indica que otros factores también juegan un papel importante en la determinación del precio.

Puedes hacer el mismo análisis para otra categoría de la variable cualitativa, pero no es necesario, bastará con que justiques la categoría seleccionada anteriormente.

- La categoría is_convertible fue seleccionada porque:
- Tiene significancia estadística en la predicción del precio.

- Muestra una diferencia clara en los precios entre autos convertibles y no convertibles.
- Permite una diferenciación de mercado, lo que justifica el análisis por separado.
- Es relevante desde el punto de vista tanto analítico como comercial, lo que valida su elección como una variable categórica clave en el análisis del precio de los automóviles.

Interpreta en el contexto del problema

El ancho del auto parece ser un factor mucho más relevante para predecir el precio de los autos no convertibles en comparación con los convertibles. Para la empresa automovilística china, esto indica que, si bien el ancho del auto tiene un impacto significativo en los precios de autos no convertibles, no es un factor determinante. Dado que la mayoría de los autos en el conjunto de datos no son convertibles, el análisis sugiere que deben centrarse en ajustar el ancho del auto para competir en este segmento en el mercado estadounidense.

Parte 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?

Con base en el análisis realizado, y en la agrupación de variables propuesta por la empresa (que incluye **altura del auto**, **ancho del auto**, y **si el auto es convertible o no**), sugeriría una nueva agrupación de variables que pueda proporcionar un mejor entendimiento y predicción del precio del automóvil. La agrupación actual ha mostrado que **el ancho del auto** y **si el auto es convertible** son variables significativas, mientras que **la altura del auto** no ha resultado relevante para predecir el precio, por l que se descarto en la decisón del desarrollo de los modelos.

Propuesta de nueva agrupación de variables

1. **Ancho del automóvil (carwidth)**: Mantendría esta variable, ya que ha demostrado ser un predictor sólido del precio, con una correlación alta y significativa. Los autos más anchos tienden a ser percibidos como más espaciosos y de mayor calidad, lo que está alineado con los resultados obtenidos.
2. **Convertibilidad (is_convertible)**: Esta variable ha sido significativa y debería conservarse, ya que afecta el precio de manera positiva, especialmente para los autos convertibles, que tienden a ser más caros debido a su diseño y percepción de exclusividad.

3. **Potencia o rendimiento del motor (horsepower)**: Incorporaría una nueva variable relacionada con el **rendimiento del motor**, como la **potencia del motor** o el **tamaño del motor**. Estas características suelen estar fuertemente vinculadas con el precio del automóvil, ya que los consumidores en mercados como el estadounidense suelen valorar la potencia y el rendimiento al momento de tomar decisiones de compra.

4. **Tipo de combustible (fuel type)**: Incluiría una variable que indique si el auto es **eléctrico, híbrido o de combustión interna**. La tendencia creciente hacia vehículos más eficientes y ecológicos en el mercado estadounidense puede influir significativamente en el precio de los autos. Este tipo de variable podría aportar información valiosa para entender mejor las preferencias del mercado.

- La agrupación original de variables propuesta por la empresa puede mejorarse incorporando variables que tengan un mayor impacto en el mercado estadounidense. Variables como **potencia del motor** y **tipo de combustible** pueden agregar un valor significativo al modelo predictivo y proporcionar una comprensión más precisa de los factores que influyen en el precio del automóvil.

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

```
# Seleccionar solo las variables numéricas
variables_numericas <- data[, sapply(data, is.numeric)]

# Calcular las medias de las variables numéricas
medias <- colMeans(variables_numericas, na.rm = TRUE)
cat("Medias de las variables numéricas:\n")

## Medias de las variables numéricas:

medias

##        symboling        wheelbase        carlength         carwidth
##     8.341463e-01     9.875659e+01     1.740493e+02     6.590780e+01
##        carheight       curbweight       enginesize           stroke
##     5.372488e+01     2.555566e+03     1.269073e+02     3.255415e+00
## compressionratio       horsepower          peakrpm          citympg
##     1.014254e+01     1.041171e+02     5.125122e+03     2.521951e+01
##       highwaympg            price   is_convertible  predicted_price
##     3.075122e+01     1.327671e+04     2.926829e-02     1.327671e+04
##         conf_lwr         conf_upr         pred_lwr         pred_upr
##     1.226107e+04     1.429235e+04     3.444735e+03     2.310869e+04

# Calcular la matriz de correlación entre las variables numéricas
correlacion <- cor(variables_numericas, use = "complete.obs")

# Imprimir la matriz de correlación
cat("\nMatriz de correlación:\n")
```

```
##
## Matriz de correlación:
```

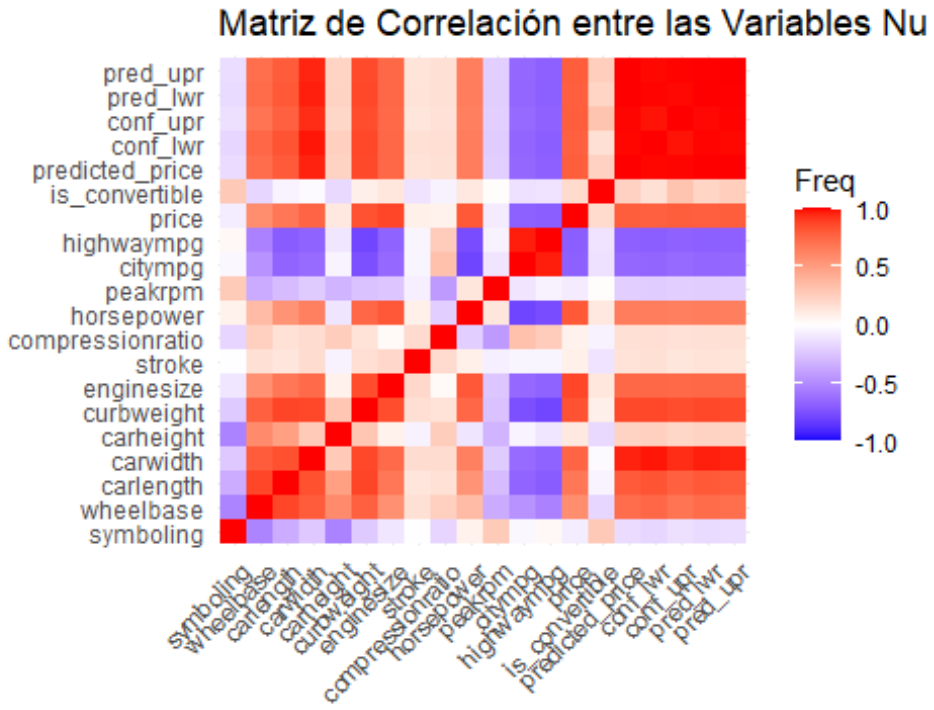
```
print(correlacion)
```

```
##                     symboling  wheelbase   carlength    carwidth
carheight
## symboling         1.000000000 -0.5319537 -0.35761152 -0.23291906 -
0.54103820
## wheelbase        -0.531953682  1.0000000  0.87458748  0.79514364
0.58943476
## carlength        -0.357611523  0.8745875  1.00000000  0.84111827
0.49102946
## carwidth         -0.232919061  0.7951436  0.84111827  1.00000000
0.27921032
## carheight        -0.541038200  0.5894348  0.49102946  0.27921032
1.00000000
## curbweight       -0.227690588  0.7763863  0.87772846  0.86703246
0.29557173
## enginesize       -0.105789709  0.5693287  0.68335987  0.73543340
0.06714874
## stroke           -0.008735141  0.1609590  0.12953261  0.18294169 -
0.05530667
## compressionratio -0.178515084  0.2497858  0.15841371  0.18112863
0.26121423
## horsepower        0.070872724  0.3532945  0.55262297  0.64073208 -
0.10880206
## peakrpm           0.273606245 -0.3604687 -0.28724220 -0.22001230 -
0.32041072
## citympg          -0.035822628 -0.4704136 -0.67090866 -0.64270434 -
0.04863963
## highwaympg        0.034606001 -0.5440819 -0.70466160 -0.67721792 -
0.10735763
## price            -0.079978225  0.5778156  0.68292002  0.75932530
0.11933623
## is_convertible    0.279439620 -0.1750710 -0.05172208 -0.02632807 -
0.16323866
## predicted_price  -0.152513635  0.7262677  0.80349439  0.96455196
0.22816095
## conf_lwr         -0.174185525  0.7505040  0.82629503  0.98153921
0.24424342
## conf_upr         -0.131512892  0.6988407  0.77683914  0.94209762
0.21176103
## pred_lwr         -0.158517091  0.7330649  0.80897254  0.96941511
0.23249095
## pred_upr         -0.146548387  0.7193013  0.79779612  0.95939696
0.22381315
##                   curbweight  enginesize       stroke compressionratio
## symboling        -0.22769059 -0.10578971 -0.008735141      -0.17851508
## wheelbase         0.77638633  0.56932868  0.160959047       0.24978585
```

```
## carlength         0.87772846  0.68335987  0.129532611       0.15841371
## carwidth          0.86703246  0.73543340  0.182941693       0.18112863
## carheight         0.29557173  0.06714874 -0.055306674       0.26121423
## curbweight        1.00000000  0.85059407  0.168790035       0.15136174
## enginesize        0.85059407  1.00000000  0.203128588       0.02897136
## stroke            0.16879004  0.20312859  1.000000000       0.18611011
## compressionratio  0.15136174  0.02897136  0.186110110       1.00000000
## horsepower        0.75073925  0.80976865  0.080939536      -0.20432623
## peakrpm          -0.26624318 -0.24465983 -0.067963753      -0.43574051
## citympg          -0.75741378 -0.65365792 -0.042144754       0.32470142
## highwaympg       -0.79746479 -0.67746991 -0.043930930       0.26520139
## price             0.83530488  0.87414480  0.079443084       0.06798351
## is_convertible    0.08227216  0.12648287 -0.117717599      -0.05299033
## predicted_price   0.86404249  0.74786466  0.146652625       0.16197823
## conf_lwr          0.87247999  0.74617764  0.161812646       0.16760021
## conf_upr          0.85015925  0.74411104  0.131678634       0.15566213
## pred_lwr          0.86664309  0.74834645  0.150006613       0.16376090
## pred_upr          0.86115216  0.74710359  0.143296398       0.16016211
##                   horsepower     peakrpm     citympg  highwaympg
price
## symboling         0.07087272  0.27360625 -0.03582263  0.03460600 -
0.07997822
## wheelbase         0.35329448 -0.36046875 -0.47041361 -0.54408192
0.57781560
## carlength         0.55262297 -0.28724220 -0.67090866 -0.70466160
0.68292002
## carwidth          0.64073208 -0.22001230 -0.64270434 -0.67721792
0.75932530
## carheight        -0.10880206 -0.32041072 -0.04863963 -0.10735763
0.11933623
## curbweight        0.75073925 -0.26624318 -0.75741378 -0.79746479
0.83530488
## enginesize        0.80976865 -0.24465983 -0.65365792 -0.67746991
0.87414480
## stroke            0.08093954 -0.06796375 -0.04214475 -0.04393093
0.07944308
## compressionratio -0.20432623 -0.43574051  0.32470142  0.26520139
0.06798351
## horsepower        1.00000000  0.13107251 -0.80145618 -0.77054389
0.80813882
## peakrpm           0.13107251  1.00000000 -0.11354438 -0.05427481 -
0.08526715
## citympg          -0.80145618 -0.11354438  1.00000000  0.97133704 -
0.68575134
## highwaympg       -0.77054389 -0.05427481  0.97133704  1.00000000 -
0.69759909
## price             0.80813882 -0.08526715 -0.68575134 -0.69759909
1.00000000
## is_convertible    0.12126730  0.01211972 -0.12557064 -0.12009393
0.18768121
```

```
## predicted_price   0.65448528 -0.21054301 -0.65753736 -0.68962161
0.78723110
## conf_lwr          0.65713877 -0.21613743 -0.66642761 -0.69737229
0.77815839
## conf_upr          0.64740451 -0.20390781 -0.64470261 -0.67760670
0.78998662
## pred_lwr          0.65528380 -0.21190252 -0.65909667 -0.69120974
0.78635037
## pred_upr          0.65344843 -0.20912460 -0.65575080 -0.68779438
0.78779549
##                  is_convertible predicted_price   conf_lwr   conf_upr
## symboling            0.27943962      -0.1525136 -0.1741855 -0.1315129
## wheelbase           -0.17507100       0.7262677  0.7505040  0.6988407
## carlength           -0.05172208       0.8034944  0.8262950  0.7768391
## carwidth            -0.02632807       0.9645520  0.9815392  0.9420976
## carheight           -0.16323866       0.2281609  0.2442434  0.2117610
## curbweight           0.08227216       0.8640425  0.8724800  0.8501592
## enginesize           0.12648287       0.7478647  0.7461776  0.7441110
## stroke              -0.11771760       0.1466526  0.1618126  0.1316786
## compressionratio    -0.05299033       0.1619782  0.1676002  0.1556621
## horsepower           0.12126730       0.6544853  0.6571388  0.6474045
## peakrpm              0.01211972      -0.2105430 -0.2161374 -0.2039078
## citympg             -0.12557064      -0.6575374 -0.6664276 -0.6447026
## highwaympg          -0.12009393      -0.6896216 -0.6973723 -0.6776067
## price                0.18768121       0.7872311  0.7781584  0.7899866
## is_convertible       1.00000000       0.2384067  0.1606573  0.3081768
## predicted_price      0.23840675       1.0000000  0.9959785  0.9966038
## conf_lwr             0.16065733       0.9959785  1.0000000  0.9852184
## conf_upr             0.30817685       0.9966038  0.9852184  1.0000000
## pred_lwr             0.21976473       0.9998034  0.9975084  0.9948217
## pred_upr             0.25665685       0.9998097  0.9940910  0.9979747
##                    pred_lwr   pred_upr
## symboling        -0.1585171 -0.1465484
## wheelbase         0.7330649  0.7193013
## carlength         0.8089725  0.7977961
## carwidth          0.9694151  0.9593970
## carheight         0.2324910  0.2238132
## curbweight        0.8666431  0.8611522
## enginesize        0.7483464  0.7471036
## stroke            0.1500066  0.1432964
## compressionratio  0.1637609  0.1601621
## horsepower        0.6552838  0.6534484
## peakrpm          -0.2119025 -0.2091246
## citympg          -0.6590967 -0.6557508
## highwaympg       -0.6912097 -0.6877944
## price             0.7863504  0.7877955
## is_convertible    0.2197647  0.2566568
## predicted_price   0.9998034  0.9998097
## conf_lwr          0.9975084  0.9940910
## conf_upr          0.9948217  0.9979747
```

```
## pred_lwr           1.0000000  0.9992262
## pred_upr           0.9992262  1.0000000

# Visualización de la matriz de correlación
ggplot(data = as.data.frame(as.table(correlacion)), aes(Var1, Var2, fill
= Freq)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
midpoint = 0, limit = c(-1, 1)) +
  theme_minimal() +
  labs(title = "Matriz de Correlación entre las Variables Numéricas", x =
"", y = "") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



* Propuesta de agrupación para la predicción del precio: Para un modelo basado en dimensiones físicas: Variables: carwidth, carheight, carlength, curbweight, wheelbase

- Para un modelo basado en características del motor y rendimiento: Variables: enginesize, horsepower, compressionratio, stroke
- Para un modelo basado en eficiencia de combustible: Variables: citympg, highwaympg
- Para el modelo basado en características especiales: Variables: is_convertible, symboling