

Universidad Nacional De Colombia

Sede Medellín



Facultad de Ciencias

Departamento de Estadística

Proyecto - 1 y 2 Entrega

Daniel Felipe Villa Rengifo

Luis David Hernández Pérez

Juan Gabriel Carvajal Negrete

Modelos de Regresión

Enero, 2025

Contexto de los datos

Planteamiento de Problema

Una empresa automotriz china, Geely Auto, tiene la intención de ingresar al mercado de Estados Unidos estableciendo una planta de fabricación allí y produciendo vehículos localmente para competir con sus contrapartes estadounidenses y europeas.

Para ello, han contratado una consultora automotriz para comprender los factores que influyen en la fijación de precios de los vehículos. Específicamente, desean entender los factores que afectan el precio de los autos en el mercado estadounidense, ya que estos pueden ser muy diferentes a los del mercado chino. La empresa desea conocer:

- Qué variables son significativas para predecir el precio de un automóvil.
- Qué tan bien estas variables describen el precio de un automóvil.

A partir de diversas encuestas de mercado, la consultora ha recopilado un conjunto de datos <https://www.kaggle.com/datasets/hellbuoy/car-price-prediction/data> sobre diferentes tipos de vehículos en el mercado estadounidense.

Objetivo Empresarial

Se requiere modelar el precio de los automóviles utilizando las variables independientes disponibles. Este modelo será utilizado por la gerencia para comprender cómo varían exactamente los precios en función de las variables independientes. De esta manera, podrán ajustar el diseño de los vehículos, la estrategia empresarial, entre otros, para alcanzar ciertos niveles de precio. Además, el modelo será una herramienta útil para que la gerencia entienda la dinámica de precios de un nuevo mercado.

Descripción de las variables

- **Car_ID**: ID único de cada observación (entero).
- **Symboling**: Clasificación del riesgo de seguro; un valor de +3 indica que el automóvil tiene alto riesgo y un valor de -3 indica que probablemente es seguro (categórico).
- **fueltype**: Tipo de combustible del automóvil, por ejemplo, gasolina o diésel (categórico).
- **aspiration**: Tipo de aspiración utilizado en el automóvil (categórico).
- **doornumber**: Número de puertas del automóvil (categórico).
- **carbody**: Tipo de carrocería del automóvil (categórico).
- **drivewheel**: Tipo de tracción (ruedas motrices) del automóvil (categórico).

- **engineloaction:** Ubicación del motor del automóvil (categórico).
- **wheelbase:** Distancia entre los ejes del automóvil (numérico).
- **carlength:** Longitud del automóvil (numérico).
- **carwidth:** Ancho del automóvil (numérico).
- **carheight:** Altura del automóvil (numérico).
- **curbweight:** Peso del automóvil sin ocupantes ni equipaje (numérico).
- **enginetype:** Tipo de motor del automóvil (categórico).
- **cylindernumber:** Número de cilindros del motor (categórico).
- **enginesize:** Tamaño del motor del automóvil (numérico).
- **fuelsystem:** Sistema de combustible del automóvil (categórico).
- **boreratio:** Relación de diámetro del cilindro (numérico).
- **stroke:** Carrera o volumen dentro del motor (numérico).
- **compressionratio:** Relación de compresión del motor (numérico).
- **horsepower:** Potencia del motor en caballos de fuerza (numérico).
- **peakrpm:** Revoluciones máximas por minuto (RPM) del motor (numérico).
- **citympg:** Rendimiento de combustible en ciudad, medido en millas por galón (numérico).
- **highwaympg:** Rendimiento de combustible en carretera, medido en millas por galón (numérico).
- **price:** Precio del automóvil, considerado como la variable dependiente (numérico).

El conjunto de datos está formado por 205 registros y 26 variables, sin valores ausentes en las variables.

Limpieza de los datos

En la variable `car_name` podemos observar que los valores almacenan tanto el nombre de la empresa como el nombre del coche por lo cual hay 147 categorías distintas. Por tanto limpiaremos esa variable separando los nombres de las empresas de carros de la variable `car_name`. Por lo tanto crearemos una variable llamada `company_name` la cual tendrá solo el nombre de la empresa o compañía a la cual pertenece el carro.

Vemos que hay algunas categorías de la variable `company_name` están mal escritas como:

- `maxda = mazda`
- `Nissan = nissan`
- `porsche = porcshee`
- `toyota = toyouta`
- `vokswagen = volkswagen = vw`

Reemplazaremos los nombres incorrectos con el nombre correcto de la empresa.

Análisis descriptivo de los datos.

Para el análisis exploratorio de datos, y dado el gran número de variables disponibles, se seleccionarán aquellas que, según la investigación previa, podrían ser relevantes para explicar el comportamiento del precio. En este análisis se evaluará la correlación entre las variables numéricas, así como su relación con la variable precio. Para las variables categóricas, se analizará su interacción con el precio, buscando patrones o asociaciones significativas, por tanto las variables que probablemente sean más importantes para predecir el precio de un automóvil:

- **Dimensiones del vehículo:** Las variables `wheelbase`, `carlength`, `carwidth`, y `carheight` podrían estar correlacionados con el precio porque un automóvil más grande o más espacioso tiende a ser más caro.
- **Especificaciones del motor:** Variables como `enginesize`, `horsepower`, y `compressionratio` están directamente relacionados con el rendimiento del automóvil y podrían influir significativamente en el precio.
- **Peso:** La variable `curbweight` puede ser un buen indicador del tipo y tamaño del vehículo, y suele correlacionarse con el precio.
- **Eficiencia de combustible:** Las variables `citympg` y `highwaympg` podrían influir en el precio, ya que los automóviles más eficientes suelen tener precios diferentes según el segmento de mercado.

Las variables categóricas como `carbody`, `drivewheel`, `fueltype`, `enginetype` y `company_name` suelen ser indicadores del tipo de vehículo y su mercado objetivo.

Variable respuesta

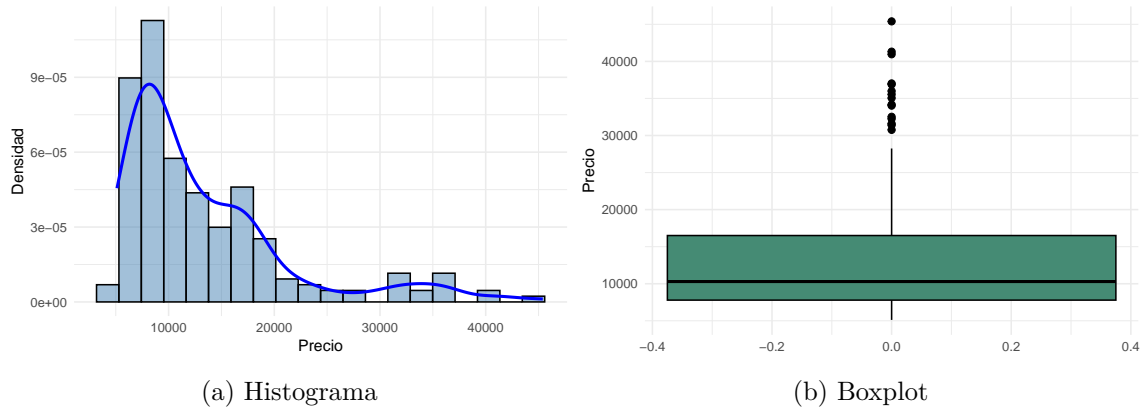


Figura 1: Distribución de la variable respuesta (Precio)

De la Figura 1 podemos decir que los precios de los carros tienen una distribución asimétrica, concentrándose en valores bajos, pero con una minoría de carros significativamente más caros, debido a la distribución puede existir problemas de normalidad. Los valores atípicos en el rango superior deben considerarse, ya que pueden representar carros de lujo o especiales por tanto analizaremos ahora la variable que corresponde al nombre de la empresa o compañía a la cual pertenece el carro para ver que relación existe con el precio.

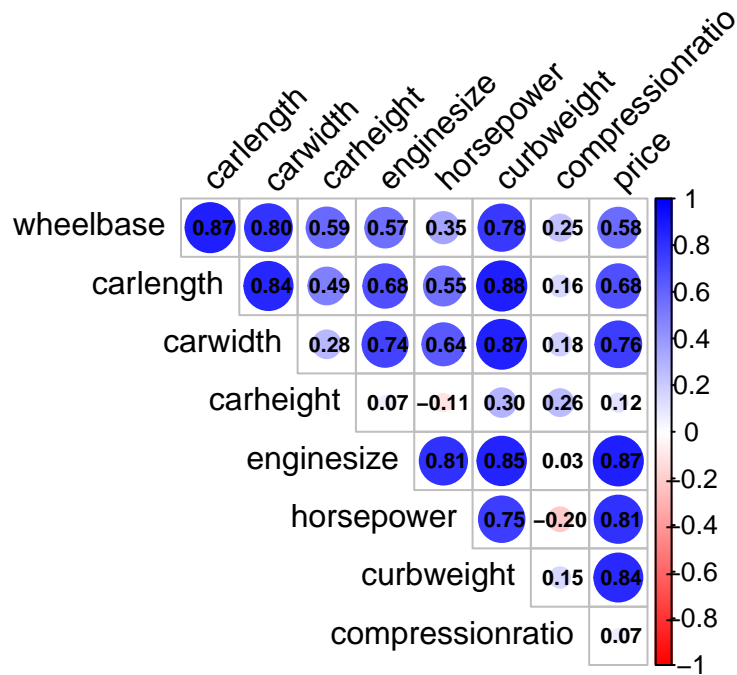


Figura 2: Grafico de Correlación

De la Figura 2 se pueden extraer las siguientes conclusiones:

- Las variables que tienen una mayor relacion lineal con el precio son : **enginesize** ($r = 0.87$), **curbweight** ($r = 0.84$) , **horsepower** ($r = 0.81$) y **carwidth** ($r = 0.76$).
- Hay pares de variables que tienen correlacion alta como **Enginesize** y **horsepower**: 0.8098 ,**Wheelbase** y **carlength**: 0.8746 , **Carlength** y **curbweight**: 0.8777 por lo que posiblemente no sea útil introducir algunas pares de variables en el modelo para evitar multicolinealidad.

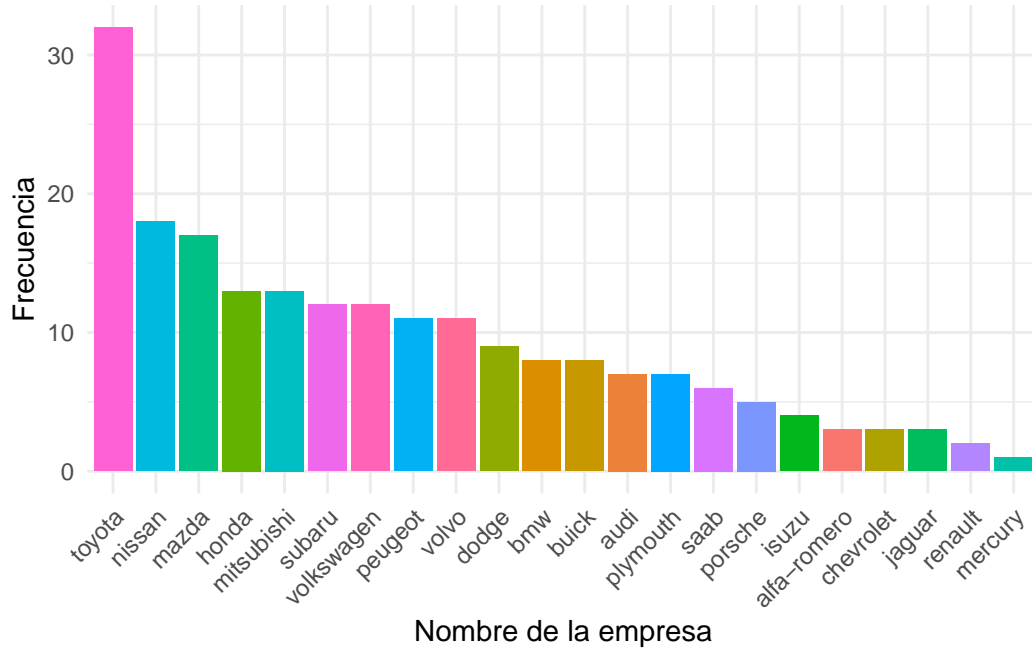


Figura 3: Distribución de los nombre de la empresa

De la Figura 3, observamos que la mayoría de los carros en este conjunto de datos está dominado por unas pocas marcas, especialmente Toyota, que supera ampliamente a las demás, por lo tanto, podemos decir que Toyota es la empresa preferida de los clientes.

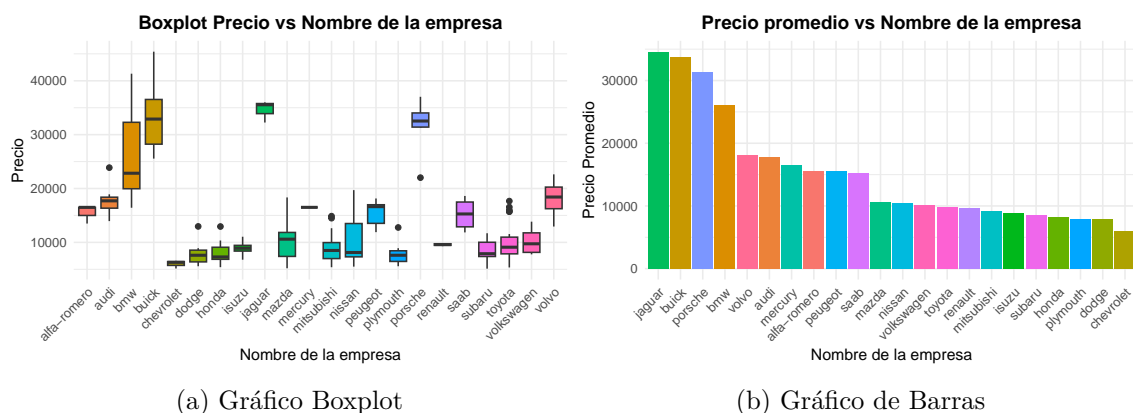


Figura 4: Distribución del Precio vs Nombre de la empresa

A partir de la Figura 4, observamos una diferencia significativa en el precio de los carros según la empresa a la que pertenecen. **Jaguar** y **Buick** parecen ofrecer los carros con las gamas de precios más altas.

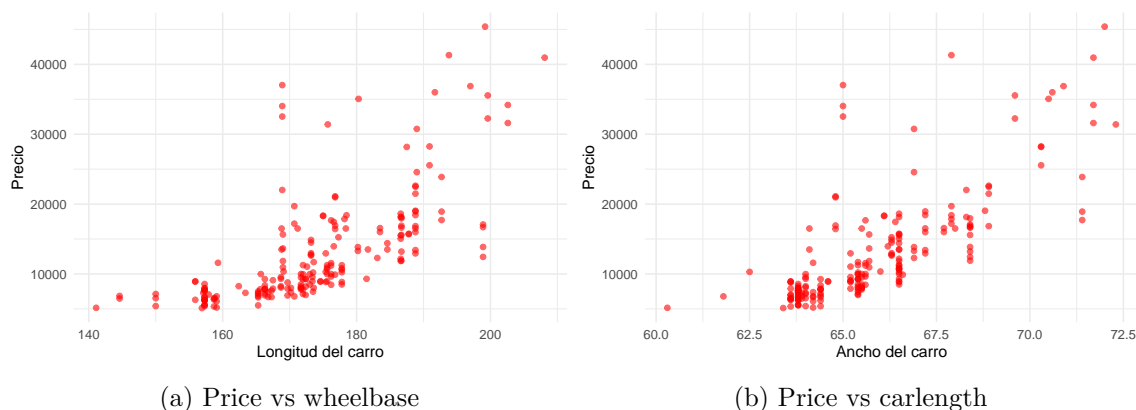


Figura 5: Gráficos de dispersión

De la Figura 5 se observa claramente que la longitud y la anchura del carro están estrechamente relacionadas con su precio. A medida que aumentan la longitud y la anchura del carro, también tiende a aumentar su precio. Sin embargo, no es posible hacer inferencias claras basadas únicamente en la relación entre la longitud y el precio, debido a la alta dispersión de los datos. La altura del carro no parece tener un impacto significativo en el precio.

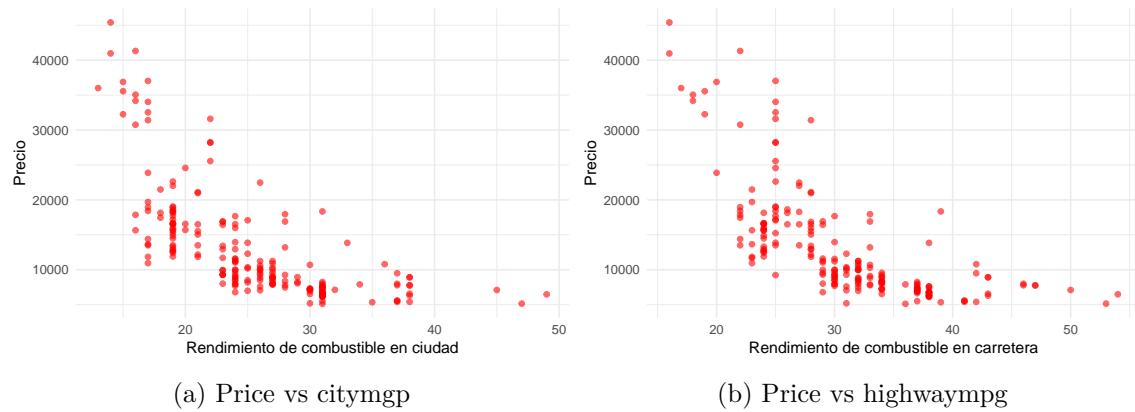


Figura 6: Gráficos de dispersión

De la Figura 6 se observa claramente que `citympg` y `highwaympg` tienen una correlación negativa con el precio del carro. A medida que aumentan los valores de `citympg` y `highwaympg`, el precio del carro tiende a disminuir. Dado que ambas características están relacionadas con el precio de manera significativa, `citympg` y `highwaympg` son características útiles para predecir el precio de los carros.

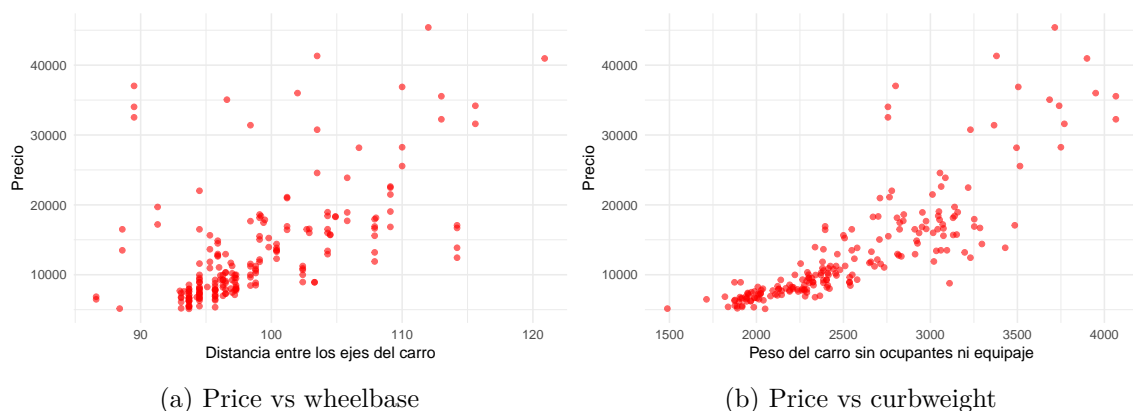


Figura 7: Gráficos de dispersión

De la Figura 7 se observa claramente que el peso del carro en vacío tiene una alta correlación (0.84) con el precio del carro. A medida que aumenta el peso en vacío, el precio del carro también incrementa de manera significativa. Aunque la distancia entre ejes y el precio no presentan una correlación tan alta, todavía existe una relación positiva. Por lo tanto, un aumento en la distancia entre ejes también está asociado con un incremento en el precio del carro.

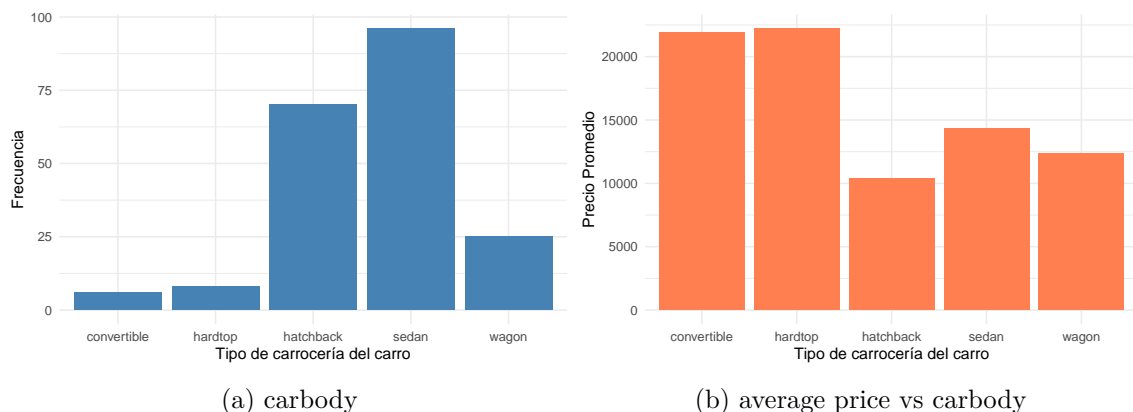


Figura 8: Gráfico de barras

De la Figura 8 se observa que los carros con carrocería sedán son los más vendidos, seguidos por los hatchback. Por otro lado, los descapotables y los de techo rígido tienen menores ventas. Estos últimos son también los más caros, seguidos de los descapotables. Es importante señalar que los descapotables y los carros con techo rígido se venden menos debido a su alto costo, lo que los hace menos atractivos para la mayoría de los clientes. Aunque la carrocería sedán ocupa el tercer lugar en términos de precio, sigue siendo la más popular,

lo que sugiere que los clientes prefieren carros de gama media.

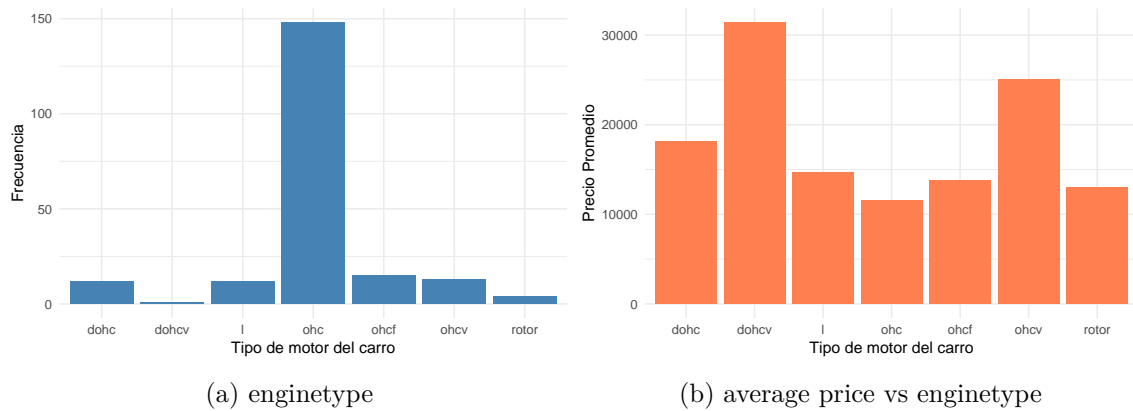


Figura 9: Gráfico de barras

De la Figura 9 observamos que la mayoría de los carros vendidos tienen motores de árbol de levas en cabeza (OHC). Solo se ha vendido un carro con motor DOHCV, y existen muy pocos datos disponibles para los motores DOHCV y de rotor. Los carros con motores DOHCV son, en su mayoría, más caros. Por otro lado, los carros con motores OHC son los menos costosos.

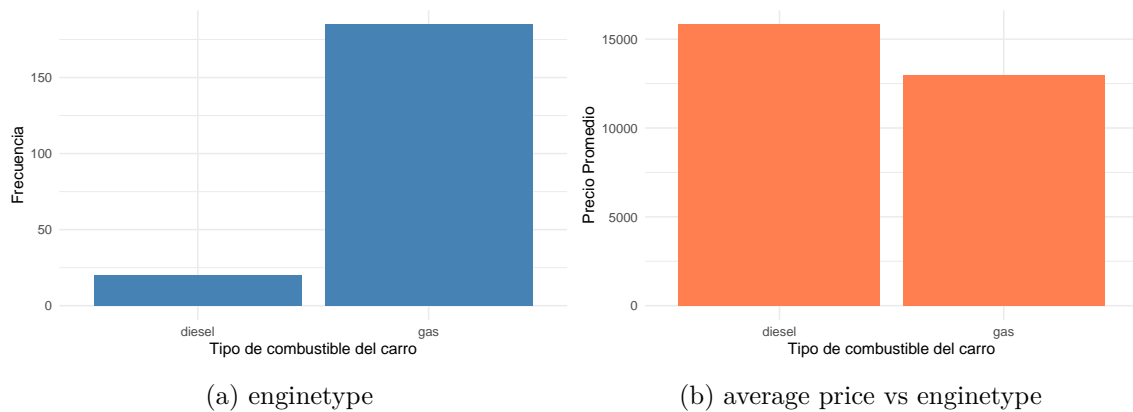


Figura 10: Gráfico de barras

De la Figura 10 podemos deducir que los carros con sistema de combustible a gas son los más preferidos, además se observa que el precio medio de los carros a gasolina es inferior al de los carros a diésel. Por lo tanto, podemos inferir que los clientes tienden a preferir carros que consumen menos combustible.

Conclusión

Después de realizar el análisis descriptivo, observamos que algunas variables parecen tener una mayor influencia en la explicación del precio de los automóviles, mientras que otras presentan una relación menos significativa. Además, identificamos que, en ciertas variables categóricas, algunas categorías tienen un número reducido de observaciones, lo que podría limitar su representatividad en los análisis posteriores. Por ello, será importante evaluar cómo estas características afectan la interpretación y robustez de los resultados.

Segunda Entrega

Punto 1

Ajuste un modelo de regresión lineal múltiple únicamente con las covariables continuas, muestre la tabla de parámetros ajustados y escriba la ecuación ajustada. Calcule la Anova del modelo ¿Es significativo el modelo? ¿Que proporción de la variabilidad total de la respuesta es explicada por el modelo? Opine sobre esto ultimo.

```
# seleccion de variables continuas
var_continuas <- datos %>%
  select(wheelbase, carlength, carwidth, carheight, boreratio, stroke,
         compressionratio, price)
# Ajuste del modelo con las variables continuas
mod_cont <- lm(price ~., data = var_continuas)
```

Tabla 1: Parámetros ajustados

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-130372.229	17959.665	-7.259	0.000
wheelbase	-120.339	141.310	-0.852	0.395
carlength	181.660	72.666	2.500	0.013
carwidth	2142.931	342.170	6.263	0.000
carheight	-480.859	201.996	-2.381	0.018
boreratio	3782.874	1691.110	2.237	0.026
stroke	-1181.080	1189.200	-0.993	0.322
compressionratio	-23.424	94.432	-0.248	0.804

El modelo ajustado es:

$$\hat{Y} = -130372.23 - 120.34X_1 + 181.66X_2 + 2142.93X_3 - 480.86X_4 + 3782.87X_5 - 1181.08X_6 - 23.42X_7$$

```
MiAnova(mod_cont)
```

Analysis of Variance Table

Response: price

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Model	7	8137665735	1162523676	46.911	< 2.2e-16 ***
Residuals	197	4881973627	24781592		

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

Del resultado de la tabla anova podemos concluir que

- El modelo es altamente significativo globalmente ($p < 2.2 \times 10^{-16}$).
- Las variables significativas individuales son `carlength`, `carwidth`, `carheight` y `boreratio`.
- El modelo explica el 62.48% de la variabilidad en los precios, lo que lo convierte en un modelo razonablemente efectivo, aunque con espacio para mejorar.

Punto 2

Calcule los coeficientes de regresión estandarizados y concluya acerca de cual de las variables aporta mas a la respuesta según la magnitud en valor absoluto de tales coeficientes (cuidado, no confunda esto con la significancia de los coeficientes de regresión).

Tabla 2: Coeficientes estimados y Coeficientes estimados estandarizados

	Estimacion	Coef.Std
(Intercept)	-130372.22875	0.0000000
wheelbase	-120.33897	-0.0907082
carlength	181.66039	0.2805405
carwidth	2142.93078	0.5754298
carheight	-480.85887	-0.1470786
boreratio	3782.87419	0.1282497
stroke	-1181.08015	-0.0463625
compressionratio	-23.42355	-0.0116461

De la Tabla 2 podemos concluir con respecto a los coeficientes estandarizados.

- La variable `carwidth` es, con diferencia, la variable que tiene el mayor aporte a la respuesta, según la magnitud absoluta.
- Las variables `compressionratio` y `stroke` aportan muy poco a la respuesta.

Punto 3

Pruebe la significancia individual de cada uno de los parámetros del modelo (excepto intercepto), usando la prueba t, establezca claramente la prueba de hipótesis y el criterio de decisión.

La hipótesis para la prueba t de la significancia individual de cada uno de los parámetros β_j (excepto el intercepto β_0): esta dada por:

$$H_0 : \beta_j = 0 \quad vs \quad H_1 : \beta_j \neq 0$$

Mediante el siguiente estadístico de prueba con su distribución bajo H_0 y criterios de rechazo:

$$\text{Estadístico de prueba: } T_0 = \frac{\hat{\beta}_j}{\sqrt{\text{MSE } C_{jj}}} \stackrel{H_0}{\sim} t_{n-k-1}$$

Rechazo con valor P: si $P(|t_{n-k-1}| > |T_0|)$ es pequeño;

Rechazo con región crítica a un nivel de significancia α : si $|T_0| > t_{\alpha/2, n-k-1}$.

De la Tabla 1 observamos las variables `carlength`, `carwidth`, `carheight` y `boreratio` son significativas.

Punto 4

Teniendo en cuenta los resultados anteriores, realice una prueba con sumas de cuadrados extras con test lineal general; especifique claramente el modelo reducido y completo, estadístico de la prueba, su distribución, cálculo de valor P, decisión y conclusión a la luz de los datos. Justifique la hipótesis que desea probar en este numeral.

Teniendo en cuenta el resultado anterior tenemos que la hipótesis a probar es:

- H_0 : Las variables significativas identificadas en la Tabla 1 (`carlength`, `carwidth`, `carheight` y `boreratio`) no contribuyen significativamente al modelo.
- H_1 : Al menos una de estas variables contribuye significativamente al modelo.

El modelo completo(MF):

$$Y_i = \beta_0 + \beta_1 \text{wheelbase} + \beta_2 \text{carlength} + \beta_3 \text{carwidth} + \beta_4 \text{carheight} + \beta_5 \text{boreratio} + \beta_6 \text{stroke} + \beta_7 \text{compressionratio} + \epsilon$$

El modeloreducido(MR):

$$Y_i = \beta_0 + \beta_1 \text{wheelbase} + \beta_6 \text{stroke} + \beta_7 \text{compressionratio} + \epsilon$$

Especificación de la prueba:

- El estadístico de prueba utilizado es el estadístico F , definido como:

$$F = \frac{(\text{RSS_reducido} - \text{RSS_completo}) / (p_completo - p_reducido)}{\text{RSS_completo} / (n - p_completo)}$$

donde:

- RSS_reducido : Suma de cuadrados residuales del modelo reducido.
- RSS_completo : Suma de cuadrados residuales del modelo completo.
- p : Número de parámetros en el modelo.
- n : Número de observaciones.

- El estadístico F sigue una distribución F de Fisher con:

$$gl_numerador = p_completo - p_reducido = 4, \quad gl_denominador = n - p_completo = 197$$

```
# Ajuste del modelo completo
MF <- lm(price ~ wheelbase + carlength + carwidth + carheight +
          boreratio + stroke + compressionratio,
          data = var_continuas)

linearHypothesis(MF, c("carlength=0", "carwidth=0", "carheight=0",
                      "boreratio=0"))
```

Linear hypothesis test:

```
carlength = 0
carwidth = 0
carheight = 0
boreratio = 0
```

Model 1: restricted model

Model 2: price ~ wheelbase + carlength + carwidth + carheight + boreratio +
stroke + compressionratio

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	201	8591783382				
2	197	4881973627	4	3709809756	37.425	< 2.2e-16 ***

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

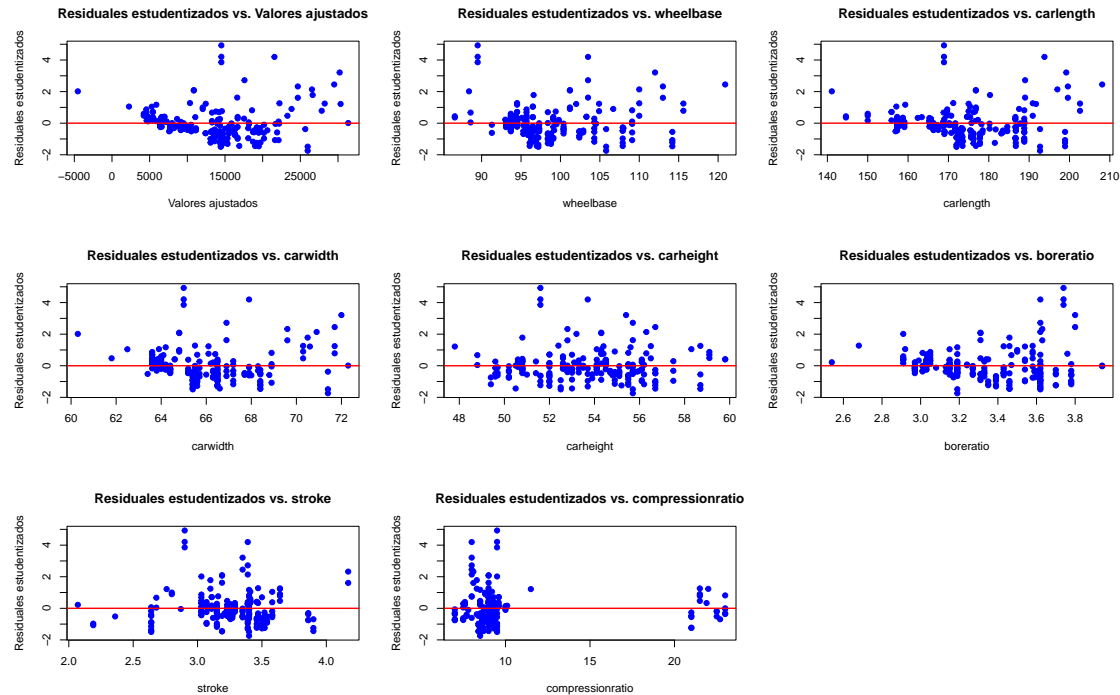
De la salida anterior podemos concluir que:

Dado que el valor p es mucho menor que 0.05, rechazamos la hipótesis nula H_0 . Esto significa que al menos una de las variables `carlength`, `carwidth`, `carheight` o `boreratio` tiene un impacto significativo en la variable dependiente (`price`).

Las variables `carlength`, `carwidth`, `carheight` y `boreratio` contribuyen significativamente al modelo. Por lo tanto, deben incluirse en el modelo para explicar adecuadamente la variabilidad de la variable dependiente (`price`).

Punto 5

Construya y analice gráficos de los residuales estudentizados vs. Valores ajustados y contra las variables de regresión utilizadas. ¿Que información proporcionan estas gráficas?



Del gráfico anterior podemos decir que

- Los gráficos sugieren que podría haber problemas de varianza no constante (heterocedasticidad), especialmente en los extremos de los valores ajustados y en variables como `boreratio` y `compressionratio`.
- En variables como `stroke`, el patrón observado podría indicar que una relación no lineal no está siendo capturada.
- Hay algunos puntos extremos (residuales estudentizados vs. Valores ajustados) que podrían estar influyendo en el modelo.

Punto 6

Construya una gráfica de probabilidad normal para los residuales estudentizados. ¿Existen razones para dudar de la hipótesis de normalidad sobre los errores en este modelo?

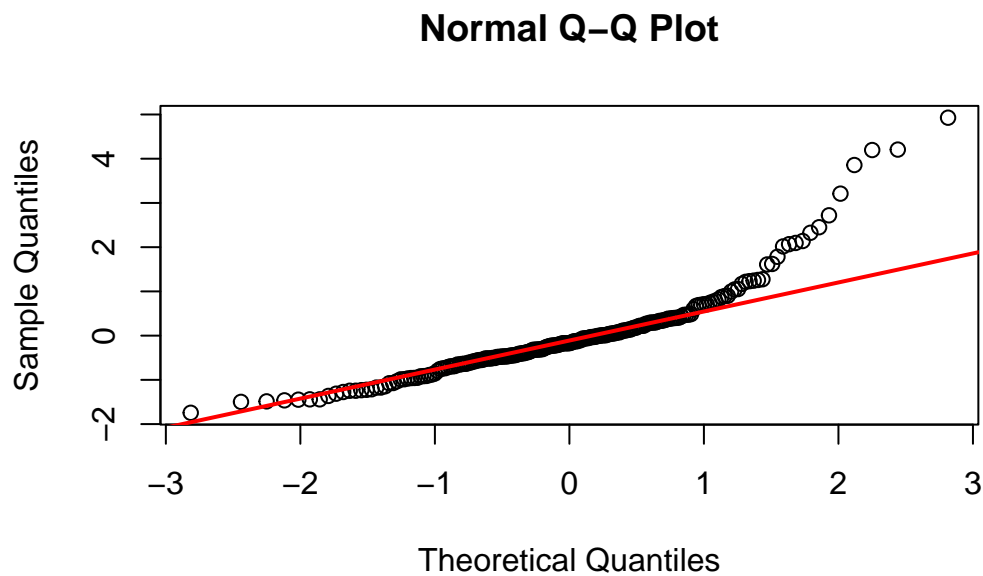


Figura 11: Gráfico Q-Q Residuales Estudentizados

De la Figura 11 observamos posibles indicios de que los errores podrían no seguir perfectamente una distribución normal debido a la desviación de algunos puntos en los extremos.

Punto 7

Diagnostique la presencia de observaciones atípicas, de balanceo y/o influyentes y concluya.

Para esto utilizaremos la función `influencePlot()`.

```
influencePlot(mod_cont)
```

Tabla 3: Observaciones potencialmente influyentes, de balanceo y/o atípiacs

Index	StudRes	Hat	CookD
28	4.20659283	0.0584620	1.266128e-01
29	4.92766935	0.0584620	1.685452e-01
30	0.01077858	0.1803456	3.211570e-06
35	0.22152729	0.2372586	1.917390e-03

Tabla 4: Valores de umbrales (Threshold values)

Threshold values	dfbeta	dffit	cov.r	Cook.d	hat	StudRes
1	0.1396861	0.3950918	0.1170732	0.02030457	0.07804878	2

De la Tabla 3 podemos concluir que:

- Las observaciones 28 y 29 son outliers claros según sus residuales estandarizados y podrían requerir mayor análisis o tratamiento.
- Las observaciones 30 y 35 aunque no son outliers extremos, tienen valores de leverage moderadamente altos, lo que indica que podrían tener un impacto significativo en el modelo.

Punto 8

Suponga que algunas de las observaciones que presentaron problemas en el numeral anterior son por problemas de digitación, elimínelas (no mas de 10) y ajuste el modelo de regresión sin dichas observaciones. Presente solo la tabla de parámetros ajustados resultante ¿Cambian notoriamente las estimaciones de los parámetros, sus errores estandar y/o la significancia? ¿Que concluye al respecto? Evalué el gráfico de normalidad para los residuales estudentizados para este ajuste ¿mejoro la normalidad? Concluya sobre los efectos de estas observaciones.

```
# Crear un nuevo conjunto de datos sin las filas problemáticas
data_clean <- var_continuas[-c(28, 29, 30, 35), ]
# Modelo ajustado sin las observaciones problemáticas
mod_cont1 <- lm(price ~., data = data_clean)
```

Tabla 5: Parámetros nuevo ajuste

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-129745.892	18185.817	-7.134	0.000
wheelbase	-137.316	143.452	-0.957	0.340
carlength	193.546	74.691	2.591	0.010
carwidth	2133.955	346.107	6.166	0.000
carheight	-499.919	207.262	-2.412	0.017
boreratio	3901.620	1715.652	2.274	0.024
stroke	-1137.553	1221.120	-0.932	0.353
compressionratio	-19.354	95.794	-0.202	0.840

De la Tabla 1 y Tabla 5 podemos concluir:

Las variables más afectadas por el nuevo ajuste son:

- **carlength**: Aumentó en magnitud y significancia, indicando que las observaciones eliminadas podrían haber reducido su importancia en el modelo original.
- **boreratio**: Mantuvo su relevancia y mostró un leve aumento en magnitud y significancia.

En general, no hubo cambios drásticos en los parámetros, lo que indica que el modelo era relativamente robusto frente a las observaciones eliminadas.

- Los errores estándar se mantuvieron similares, lo que sugiere que las eliminaciones no afectaron de manera significativa la estabilidad del modelo.
- No se perdieron variables significativas tras el ajuste. De hecho, **carlength** mostró una mayor relevancia en el nuevo modelo.
- Las observaciones eliminadas tenían un impacto limitado en el modelo general, aunque influían ligeramente en la significancia y magnitud de algunas variables clave, como **carlength** y **boreratio**.

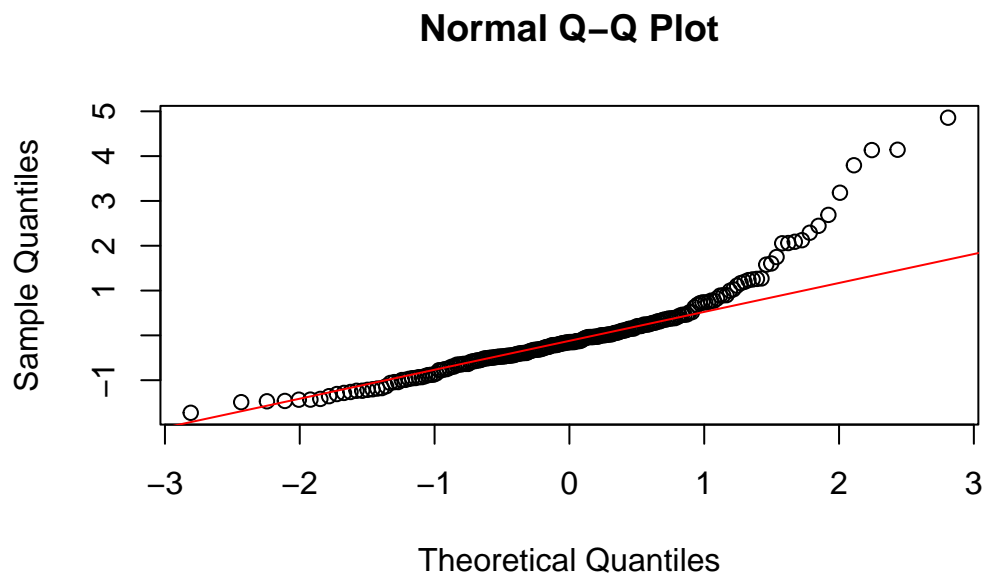


Figura 12: Gráfico Q-Q Residuales Estudentizados (nuevo ajuste)

De la Figura 12 vemos que todavía hay algunos puntos que se alejan significativamente de la línea recta (en particular en los extremos), lo cual podemos decir de que no hay mejoría en la normalidad, por lo tanto se podría optar por eliminar esas observaciones.

Tercera Entrega

Punto 1

Matriz de correlación de las variables predictoras