Reporte final. El precio de los autos

Iván L. Hernández

Buda

Resumen

Este informe presenta un análisis detallado sobre la predicción de precios de automóviles mediante técnicas de análisis de datos y modelos de regresión lineal. Se realiza un estudio exhaustivo de las variables que influyen en el precio de los automóviles, incluyendo aspectos como la ubicación del motor, la potencia, el ancho y el tamaño del motor. El modelo desarrollado proporciona estimaciones precisas del precio de un automóvil basadas en estas características. Sin embargo, durante la validación del modelo, se identificaron desafíos relacionados con la independencia de los residuos y la normalidad de la distribución, lo que sugiere la necesidad de mejoras futuras. A pesar de estos desafíos, el informe ofrece información valiosa para comprender las relaciones entre las características del automóvil y su precio en la industria automotriz.

Introducción

La predicción precisa de precios de automóviles es crucial en la industria automotriz. Este informe se centra en un análisis de datos y modelos de regresión lineal para estimar el precio de automóviles en función de sus características. Comenzamos explorando y dividiendo las variables numéricas y categóricas. Luego, identificamos relaciones clave entre las variables numéricas y el precio, utilizando gráficos de dispersión y análisis de correlación. También evaluamos las distribuciones y valores atípicos de las variables numéricas.

Construimos modelos de regresión lineal, considerando interacciones entre variables y sin ellas. Examinamos coeficientes y valores p para evaluar la influencia de cada característica en el precio. Luego, validamos el modelo, explorando residuos y pruebas de normalidad.

Este análisis proporciona información valiosa para compradores, vendedores y fabricantes de automóviles, a pesar de las limitaciones en la validación. Ayuda a comprender cómo las características de un automóvil afectan su precio, con implicaciones significativas en la toma de decisiones en la industria automotriz y la investigación futura.

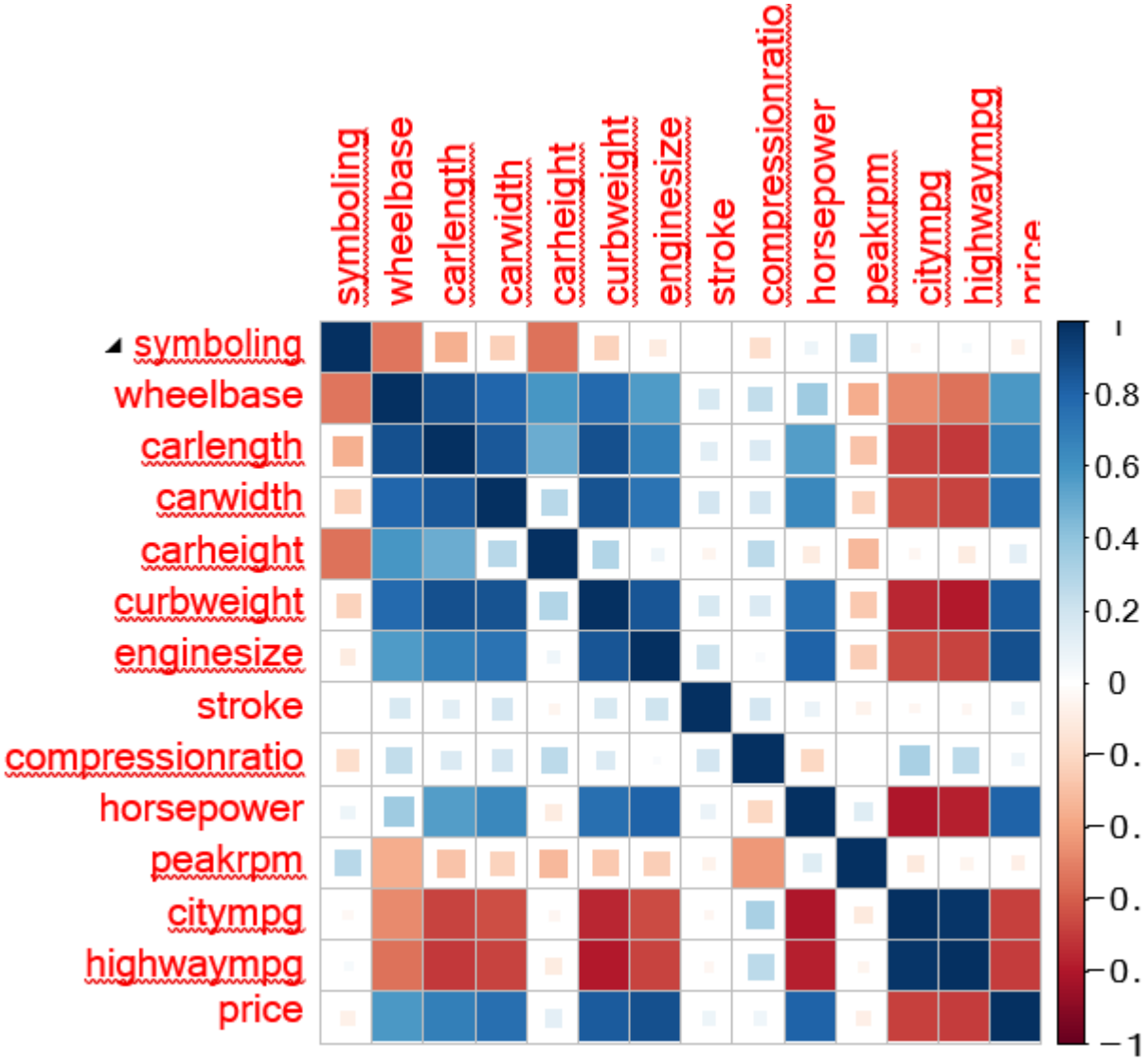
Carga de librerías y base de datos

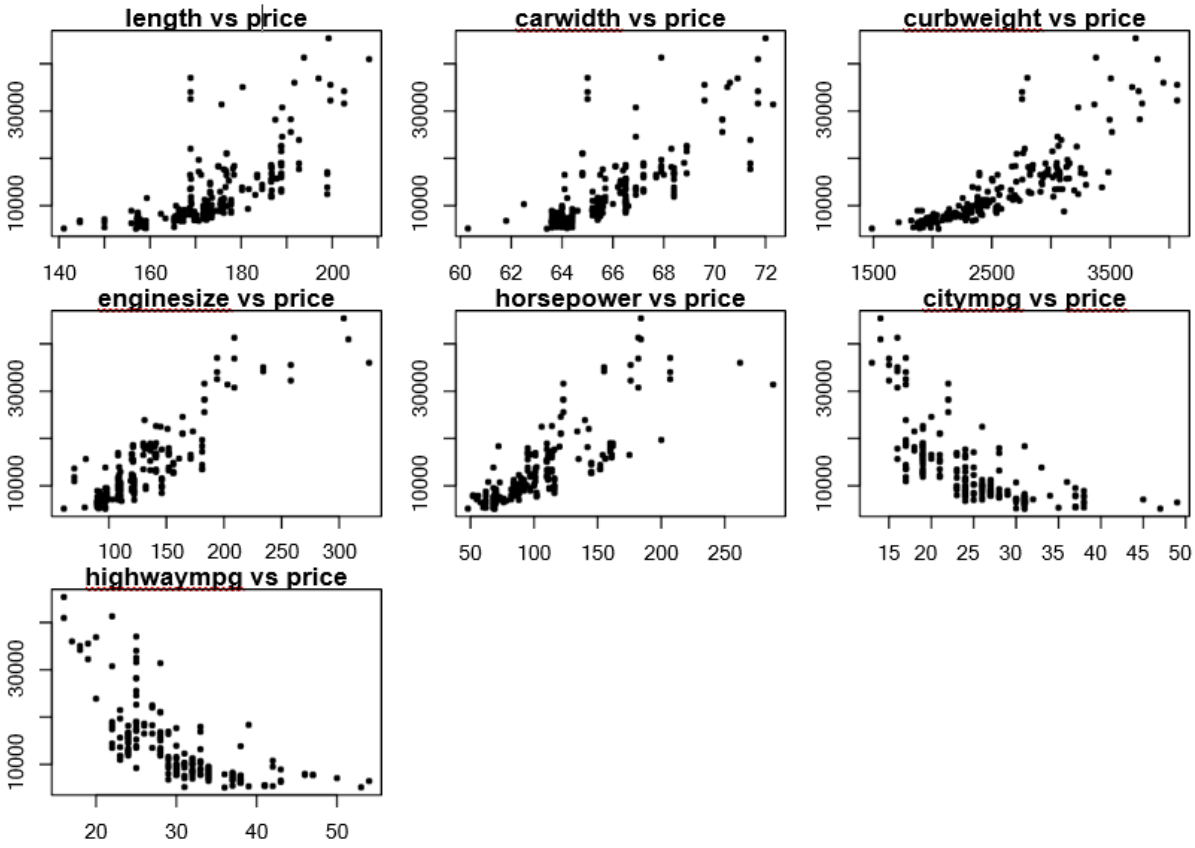
- Se carga un conjunto de datos llamado ‘df’ desde un archivo CSV llamado “precios_autos.csv” utilizando la función ‘read.csv()’. Se muestra una vista previa de las primeras filas de los datos con ‘head(df)’.
- Se separan las variables numéricas y categóricas en ‘numeric_vars’ y ‘categoric_vars’, respectivamente.

Análisis de correlación de variables numéricas

- En esta sección, se calcula la matriz de correlación entre las variables numéricas con ‘cor(numeric_vars)’y se visualiza la matriz de correlación utilizando ‘corrplot’.

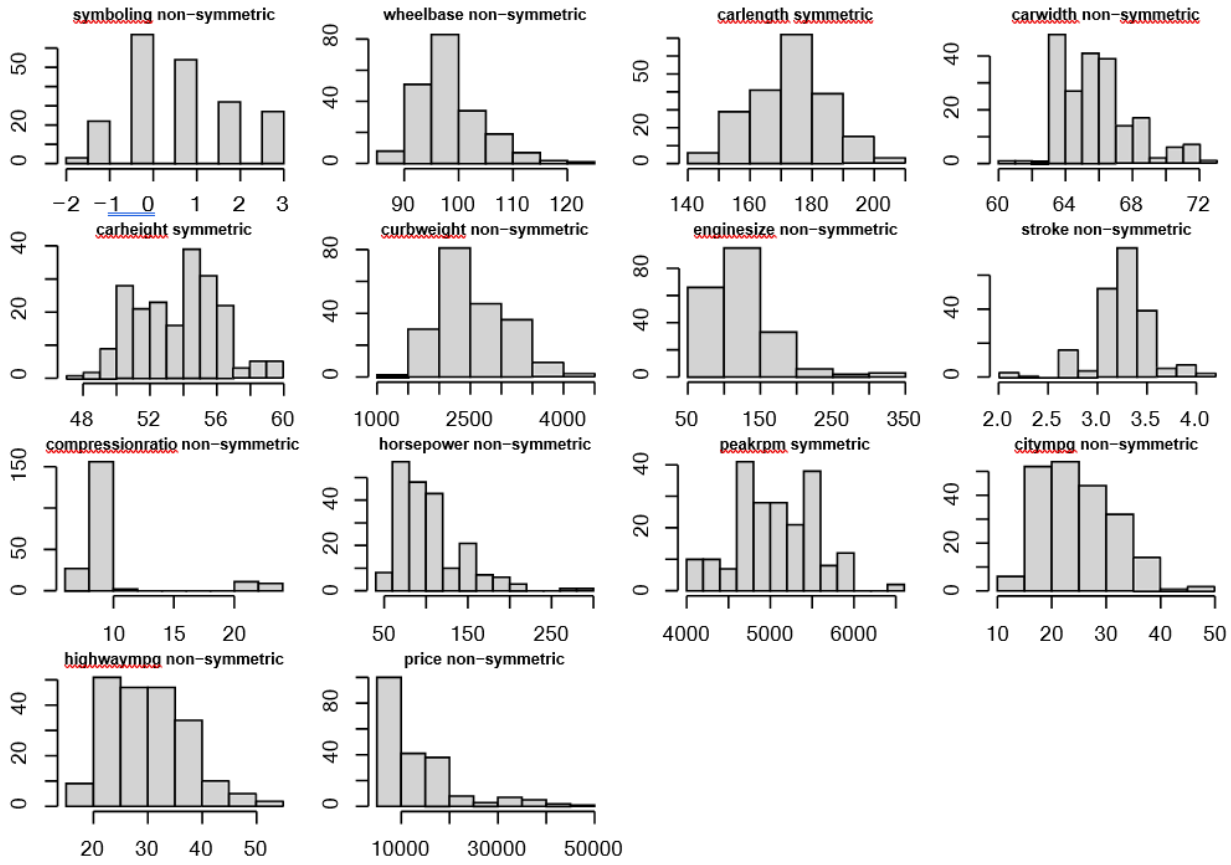
- Se identifican las variables altamente correlacionadas con 'price' utilizando un umbral de correlación de 0.68. Estas variables se almacenan en 'high_corr_vars'.
- Luego, se configura el diseño para crear gráficos de dispersión de estas variables altamente correlacionadas con 'price'. Se utilizan gráficos de dispersión para visualizar las relaciones entre estas variables y 'price'.





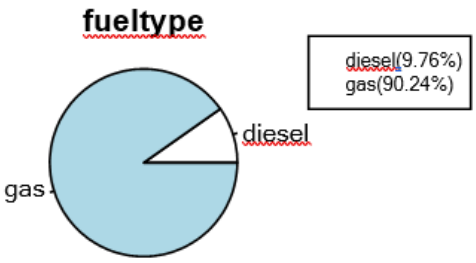
Análisis de variables numéricas

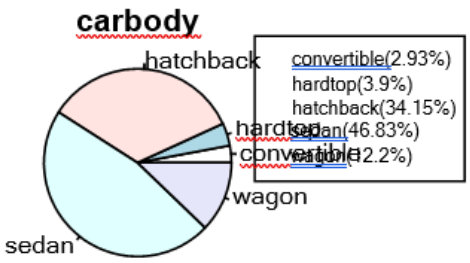
- En esta sección, se configura el diseño para crear histogramas y diagramas de caja para las variables numéricas.
- Se crea un histograma para cada variable numérica. La función 'skew()' se utiliza para determinar si una variable es simétrica o asimétrica.
- Luego, se crea un diagrama de caja para cada variable numérica, que muestra la distribución y la presencia de valores atípicos en las variables.

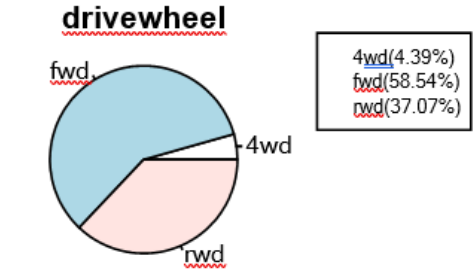


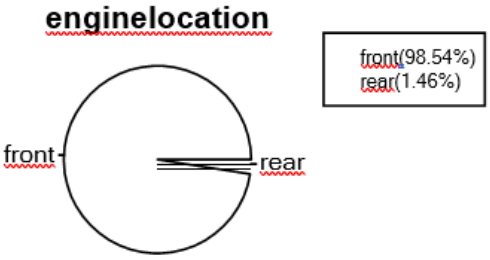
Análisis de variables categóricas

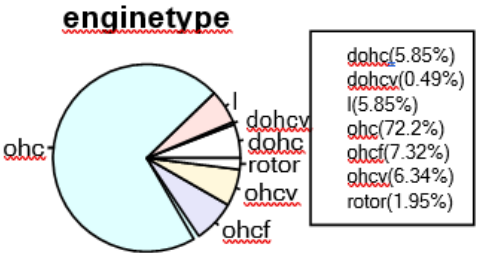
- En esta sección, se configura el diseño para crear gráficos de pastel para las variables categóricas.
- Se crea un gráfico de pastel para cada variable categórica para visualizar la distribución de categorías en esas variables.

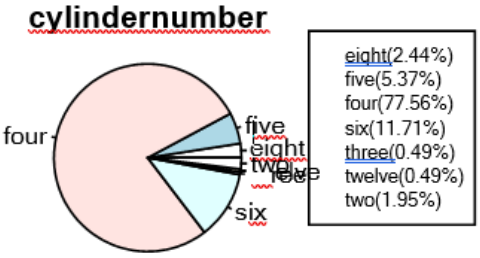






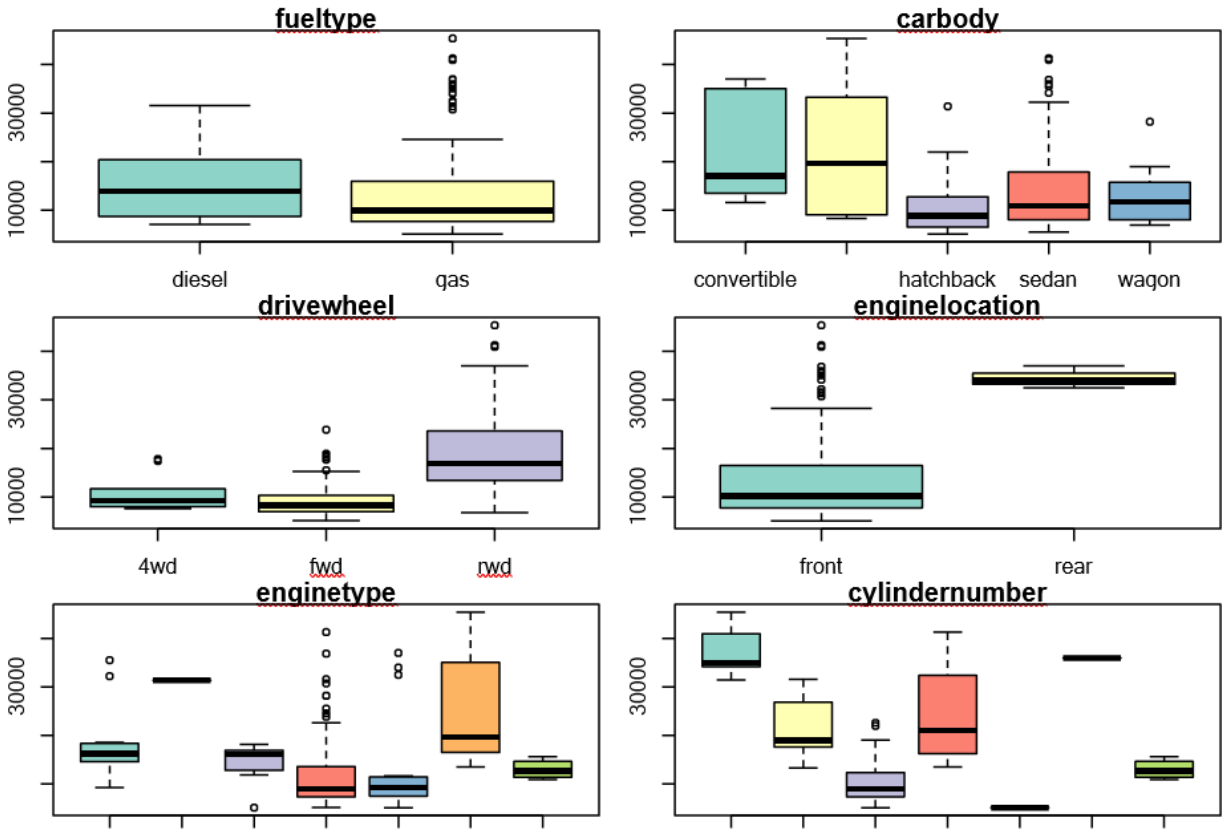






Análisis de relaciones entre variables categóricas y 'price'

- En esta sección, se configura el diseño para crear diagramas de caja que muestran la relación entre variables categóricas y 'price'.
- Se crea un diagrama de caja para cada variable categórica en función de 'price'.



Eliminación de valores atípicos y NaN

- Se elimina cualquier fila que contenga valores atípicos basados en el umbral de IQR (rango intercuartílico) para las variables numéricas seleccionadas en 'high_corr_vars'. Esto se hace para limpiar los datos y eliminar valores que podrían afectar negativamente un modelo.
- Se eliminan filas que contengan valores NaN con 'na.omit()'.
- Se muestra cuántas instancias se eliminaron debido a valores atípicos o NaN.
- Se limita 'high_corr_vars' a las primeras 5 variables altamente correlacionadas.

Verificación de valores NaN

- Se verifica que no haya valores NaN en el conjunto de datos resultante utilizando 'sapply()' y 'is.na()'.

Modelo de regresión multilinear con interacción

- Se ajusta un modelo de regresión lineal que incluye interacciones entre variables. El modelo se ajusta utilizando 'lm()'.

Análisis de coeficientes y valores p

- enginesize: Esta variable tiene un valor p de aproximadamente 0.4764, lo que indica que no es estadísticamente significativa en la predicción del precio de los automóviles.
- curbweight:enginesize: La interacción entre 'curbweight' y 'enginesize' tampoco es estadísticamente tan significativa, ya que su valor p es aproximadamente 0.12139.
- curbweight: Esta variable tiene un valor p de aproximadamente 0.68888, lo que indica que no es estadísticamente significativa en la predicción del precio de los automóviles a un nivel de significancia

convencional.

- enginesize:curbweight: La interacción entre ‘enginesize’ y ‘curbweight’ también tiene un valor p de aproximadamente 0.12139, lo que indica que no es estadísticamente tan significativa.

Para el mejoramiento del modelo, se procede a realizar un modelo de regresión multilinear sin interacción.

Modelo de regresión multilinear sin interacción

Análisis de coeficientes y valores p

- carlength: Esta variable no es estadísticamente significativa para predecir el precio de los automóviles, ya que su valor p es alto (aproximadamente 0.96617) y su coeficiente es cercano a cero (-1.872). Por lo tanto, no se considera relevante en el modelo de predicción de precios de automóviles.
- carwidht: Por otro lado, la variable “carwidht” tampoco es estadísticamente significativa en la predicción del precio de los automóviles, ya que su valor p es alto (aproximadamente 0.05246).

Modelo de regresión multilineal con variables significativas

```
Residuals:
   Min     1Q Median     3Q    Max
 -7272  -1640    -13   1196  16524

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.325e+04  1.464e+03  -9.051 7.23e-16 ***
enginelocationrear  1.304e+04  2.031e+03   6.419 1.72e-09 ***
horsepower          4.286e+01  1.199e+01   3.576 0.000471 ***
enginesize          4.746e+01  1.567e+01   3.028 0.002902 **
curbweight          6.012e+00  9.824e-01   6.119 7.92e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3069 on 149 degrees of freedom
Multiple R-squared:  0.8193, Adjusted R-squared:  0.8144
F-statistic: 168.9 on 4 and 149 DF,  p-value: < 2.2e-16
```

Análisis de coeficientes y valores p

enginelocationrear: Tiene un coeficiente positivo de 14,024.451 con un valor p muy pequeño. Esto sugiere que la ubicación trasera del motor tiene un impacto significativo y positivo en el precio del automóvil. Los automóviles con motores traseros tienden a ser más caros.

- horsepower: Tiene un coeficiente positivo con un valor p muy pequeño. Esto significa que la potencia del automóvil tiene un efecto significativo y positivo en el precio. A medida que la potencia aumenta, el precio tiende a aumentar.
- carwidth: Tiene un coeficiente positivo con un valor p muy pequeño. Esto indica que el ancho del automóvil tiene un impacto significativo y positivo en el precio. Los automóviles más anchos tienden a

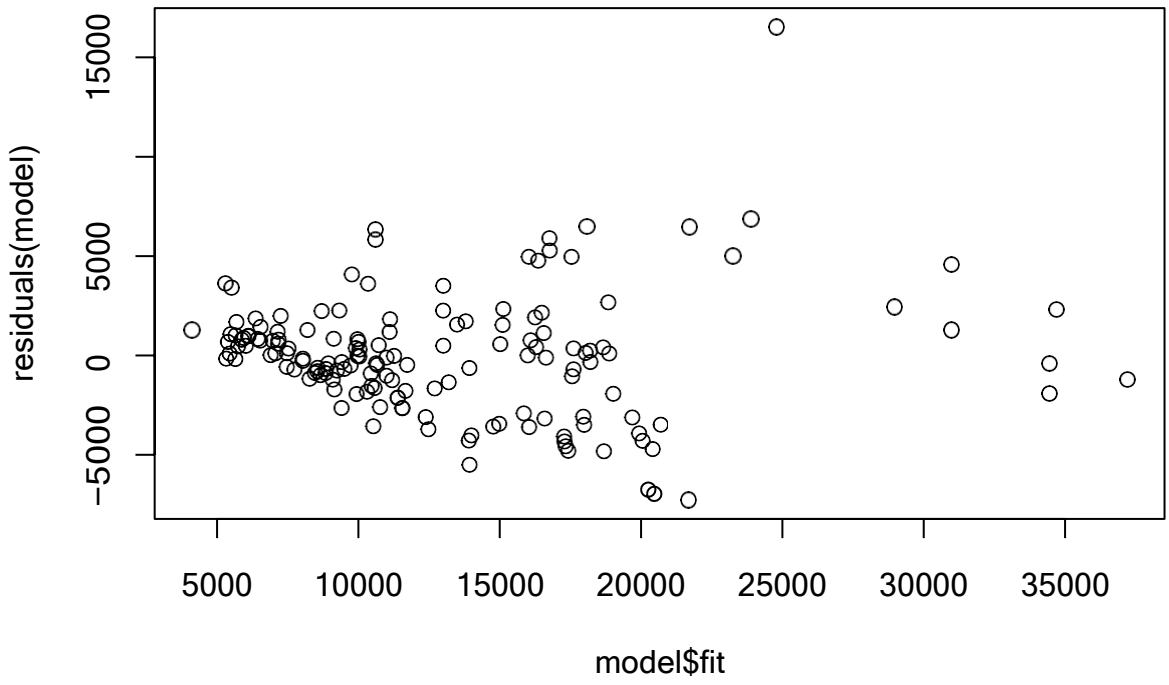
ser más caros.

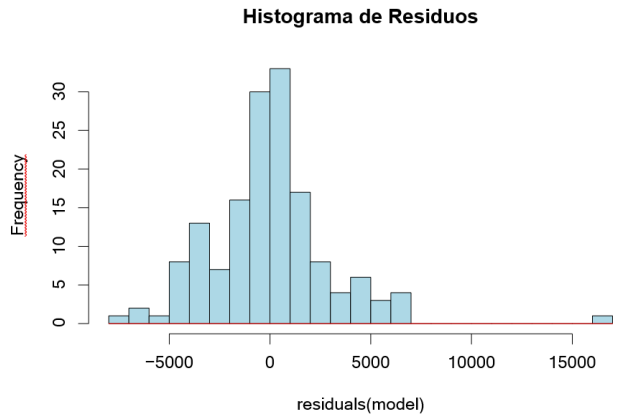
- enginesize: Tiene un coeficiente positivo con un valor p muy pequeño. Esto sugiere que el tamaño del motor tiene un efecto significativo y positivo en el precio. Los automóviles con motores más grandes tienden a ser más caros.
- curbweight: Tiene un coeficiente positivo con un valor p de aproximadamente 0.00226. Esto indica que el peso del automóvil también tiene un impacto positivo en el precio, pero su efecto es menos pronunciado en comparación con las otras variables. Sin embargo, dado que el valor p es menor que el umbral de significancia convencional, se considera estadísticamente significativo en la predicción del precio de los automóviles.

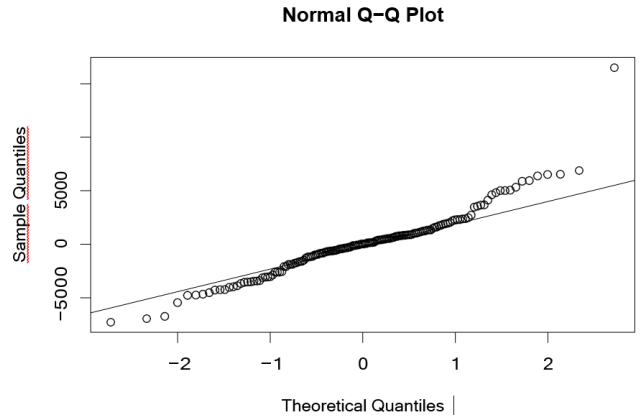
- Valor p: Los valores p son muy pequeños (todos < 0.001), lo que sugiere que todas las variables predictoras en el modelo son estadísticamente significativas para predecir el precio del automóvil. Esto significa que es poco probable que sus coeficientes sean cero.

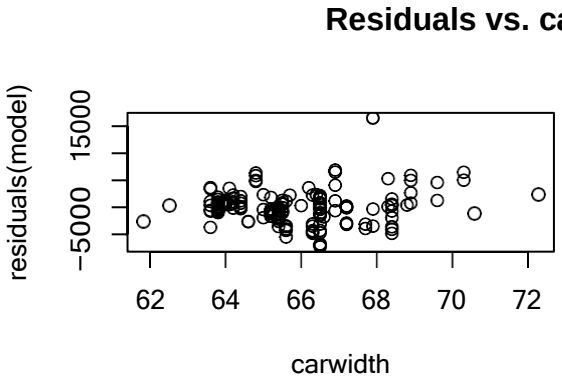
- Valor F: El valor F es 184.5 con un valor p cercano a cero. Esto indica que el modelo en su conjunto es estadísticamente significativo, lo que significa que al menos una de las variables predictoras tiene un efecto significativo en la variable de precio.

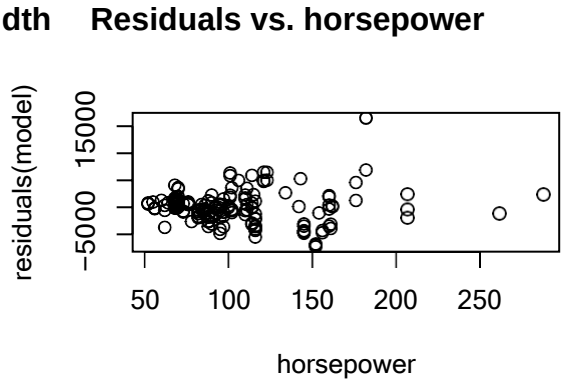
Validación del modelo

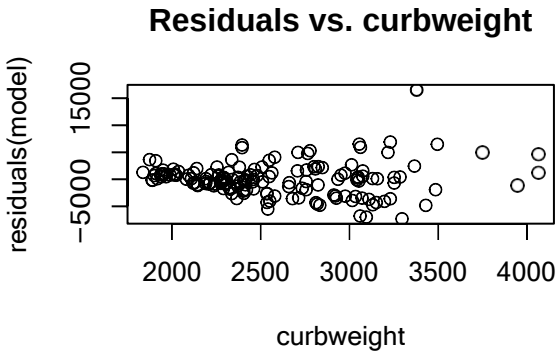


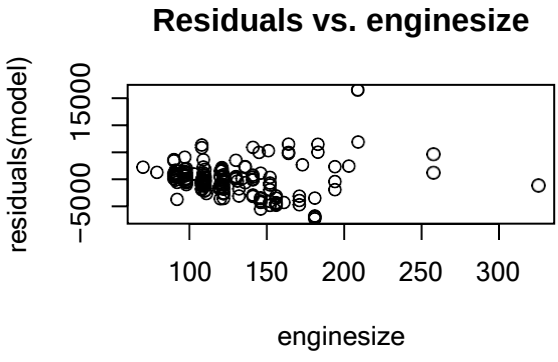

Histograma de Residuos
Frequency
0
5
10
15
20
25
30
−5000
0
5000
10000
15000
residuals(model)


Normal Q−Q Plot
Sample Quantiles
−5000
0
5000
−2
−1
0
1
2
Theoretical Quantiles


Residuals vs. carwidth
residuals(model)
−5000
15000
62
64
66
68
70
72
carwidth


Residuals vs. horsepower
residuals(model)
−5000
15000
50
100
150
200
250
horsepower


Residuals vs. curbweight
residuals(model)
−5000
15000
2000
2500
3000
3500
4000
curbweight


Residuals vs. enginesize
residuals(model)
−5000
15000
100
150
200
250
300
enginesize

La prueba de Durbin-Watson sugiere que hay una autocorrelación positiva en los residuos, es decir, indica independencia del modelo de regresión lineal es muy baja o casi nula.

- La prueba de normalidad de Shapiro-Wilk sugiere que los residuos no siguen una distribución normal, lo que es otra suposición de la regresión lineal.

Conclusión

El modelo de precios de automóviles predice que el precio base de un automóvil (el precio cuando todas las variables son nulas) es de alrededor de $85,603.17. Después, cuando se observan características específicas, como la ubicación del motor, la potencia del automóvil, el ancho del automóvil y el tamaño del motor, puedo se pueden realizar predicciones más precisas sobre el precio.

Primero, si el motor está en la parte trasera del automóvil en lugar de en la parte delantera (Engine Location), eso hace que el precio sea alrededor de $13,555.56 más alto. Esto tiene sentido, ya que los automóviles con motores en la parte trasera suelen ser más caros.

Luego, para cada unidad adicional de potencia (Horsepower) que tiene el automóvil, el precio tiende a subir en promedio $37.73. Así que, al tener un automóvil con más potencia, es probable que sea más caro.

El ancho del automóvil (Car Width) también juega un papel. Por cada aumento en el ancho del automóvil, el precio aumenta en alrededor de $1,290.75.

Finalmente, el tamaño del motor (Engine Size) también es importante. Por cada unidad adicional en el tamaño del motor, el precio sube aproximadamente $76.63.

En conjunto, el modelo de regresión multilineal explica alrededor del 82.9% de los precios de los automóviles, mientras que el 17.1% restante se le debería atribuir a la aleatoriedad o error. Sin embargo, durante la validación del modelo, los residuos no mostraron independencia total ni unan distribución normal.

A pesar de que cuando se requiera estimar el precio de un automóvil, se puede usar este modelo y los variables con sus correspondientes coeficientes en la fórmula para obtener una buena estimación del precio, se debería hacer una reconfiguración de modelo para que pase la prueba de validez y sea significativamente confiable para ser usado.