Tecnológico de Monterrey

Instituto Tecnológico y de Estudios Superiores de Monterrey

Reporte Final. de “El Precio de los Autos”

TC3006C.101 Inteligencia artificial avanzada para la ciencia de datos I

Profesores:

Ivan Mauricio Amaya Contreras

Blanca Rosa Ruiz Hernandez

Antonio Carlos Bento

Frumencio Olivas Alvarez

Hugo Terashima Marín

Alumno:

Alberto H Orozco Ramos - A00831719

11 de Septiembre de 2023

Reporte Final: "El Precio de los Autos"

Resumen de la problemática

Para este entregable se plantea el problema de una empresa automovilística china que se enfrenta al desafío de ingresar al mercado estadounidense y competir con sus contrapartes locales y europeas. Para abordar esta problemática, se realizó un análisis exhaustivo de los factores que influyen en el precio de los automóviles en el mercado estadounidense. Se utilizaron métodos y técnicas estadísticas, incluyendo la regresión lineal múltiple y la imputación de datos faltantes utilizando el paquete "mice" en R.

Los resultados clave del análisis son los siguientes:

- Se identificaron variables significativas que afectan el precio de los automóviles en el mercado estadounidense, incluyendo la categoría del automóvil, la recaudación mundial bruta, el porcentaje de presupuesto recuperado y otros factores relevantes.
- Se evaluó la capacidad de estas variables para describir el precio de los automóviles y se encontró que algunas variables tienen un impacto significativo en la predicción del precio, mientras que otras tienen una influencia limitada.

Este análisis proporciona a la empresa automovilística china una comprensión más profunda de los factores críticos que afectan el precio de los automóviles en el mercado estadounidense, lo que les permitirá tomar decisiones estratégicas informadas al ingresar a este mercado competitivo.

Introducción

En una era de globalización, dónde la mayor parte de los mercados y negocios aspiran a expandir su presencia más allá de las fronteras, la industria automotriz se ha convertido en el claro ejemplo de una competición internacional de este calibre. Conforme los mercados han evolucionado y las preferencias de los consumidores por igual, las compañias automotrices han buscado adaptarse a los nuevos panoramas, especialmente cuando se trata de apuntar a mercados extranjeros. Este reporte busca profundizar en esta problemática crítica que encara una compañia automotriz china que aspira a establecerse en el altamente competitivo mercado estadounidense.

Abordamiento del problema

La problemática en mano gira alrededor de una pregunta en cuestión de entender los determinantes de precios de los automóviles en el mercado estadounidense. Esta investigación asume primordial importancia en el cliente, la firma automotriz china, como busca competir con los mercados ya establecidos de América y Europa.

Relevancia de la problemática

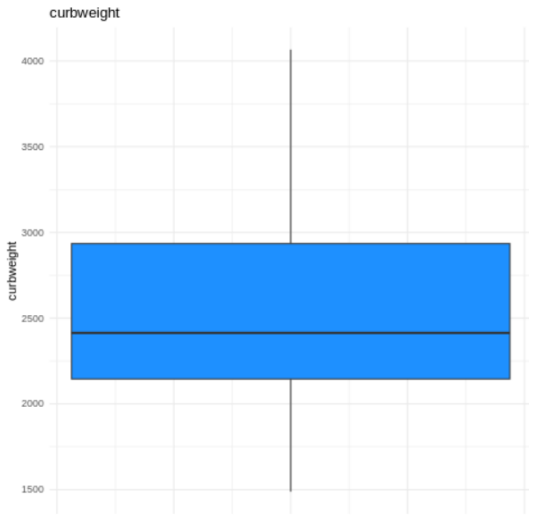
La situación esta subrayadada por las nada estables, dinámicas del mercado global. Para denotar su relevancia, es escencial considerar que la literatura creciente estrategias de entrada al mercado y las dinámicas de precios para negocios internacionales. Entre más nos adentremos en este reporte, emplearemos métodos estadísticos, analíticos y predictivos para decifrar la compleja interacción de factores que impactan el precio de los automóviles en losEstados Unidos.

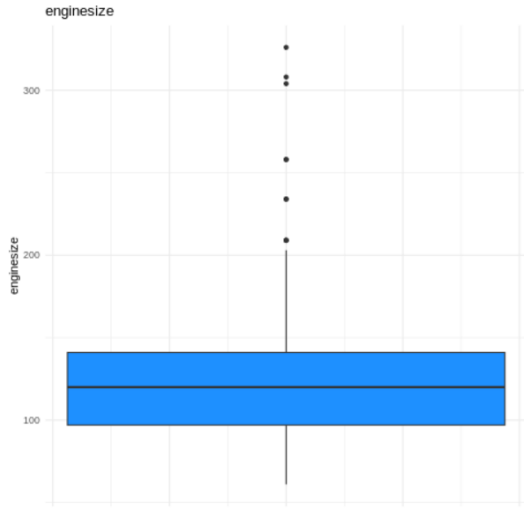
Exploración de la Base de Datos

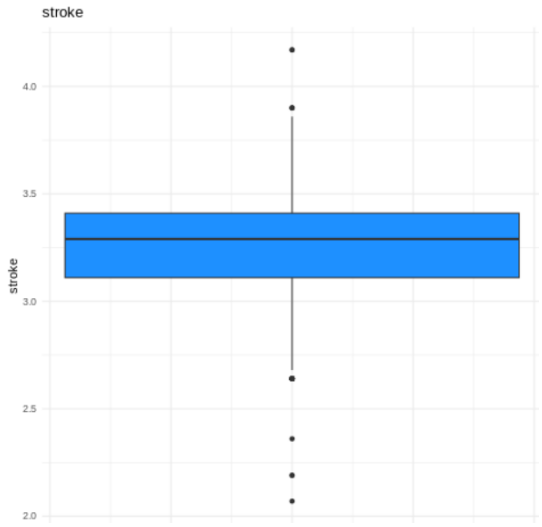
Visualización de Variables Cuantitativas

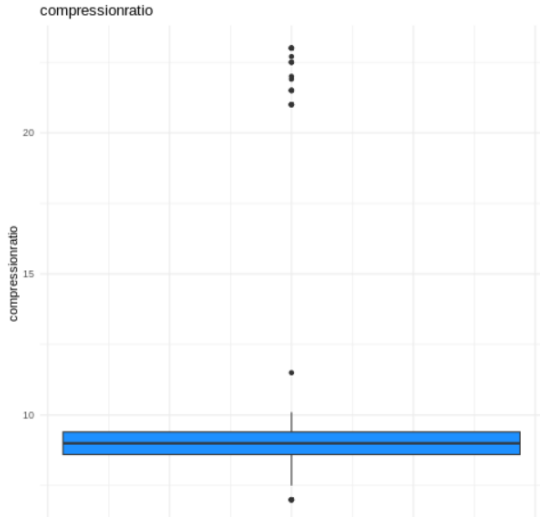
Bloxplot y Valores Atípicos

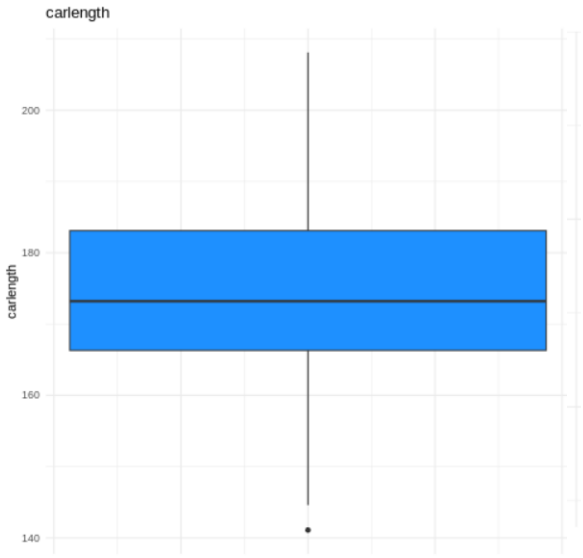
Representaremos los valores cuantitativos en forma de gráficas de caja y bigotes o *boxplots*, que nos permitirán conocer el rango en el cual se encuentra cada variable, la tendencia central y la variabilidad de los datos, así como visualizar los datos atípicos:

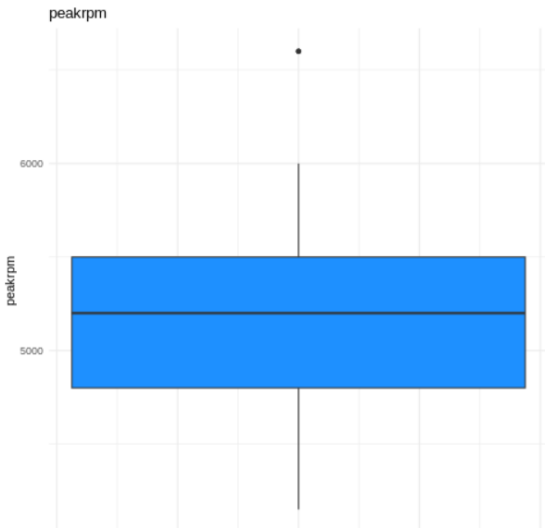


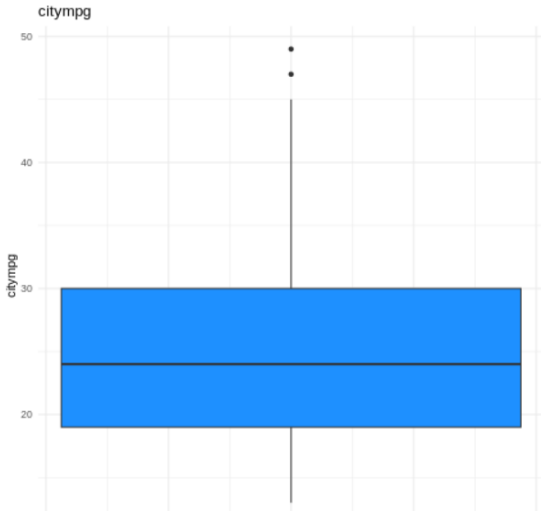


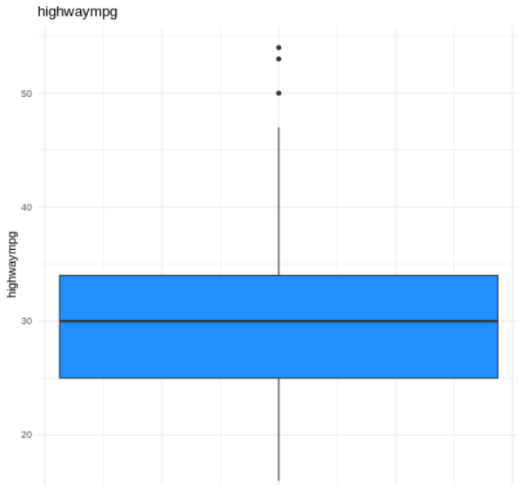


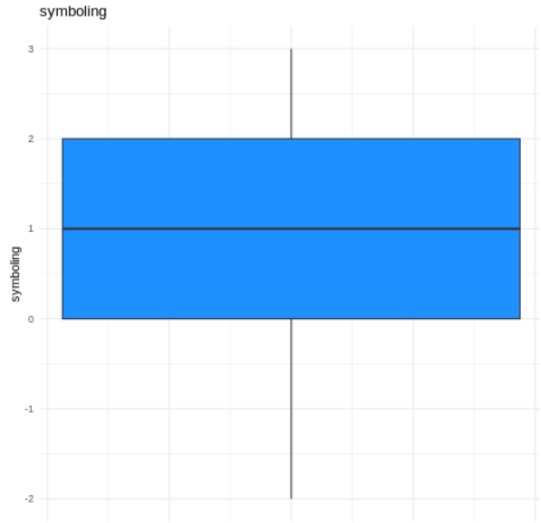


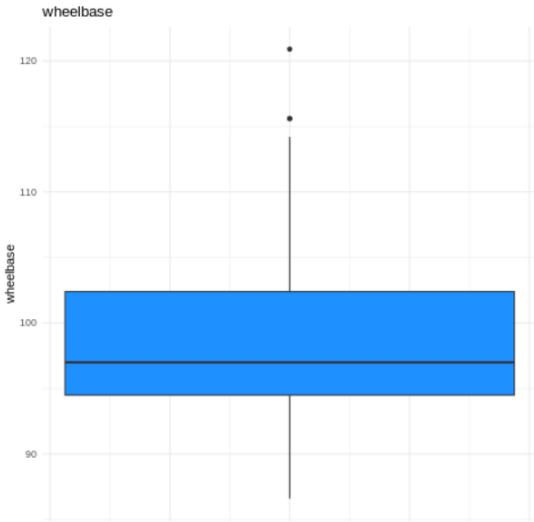

carlength
200
180
160
140
carlength

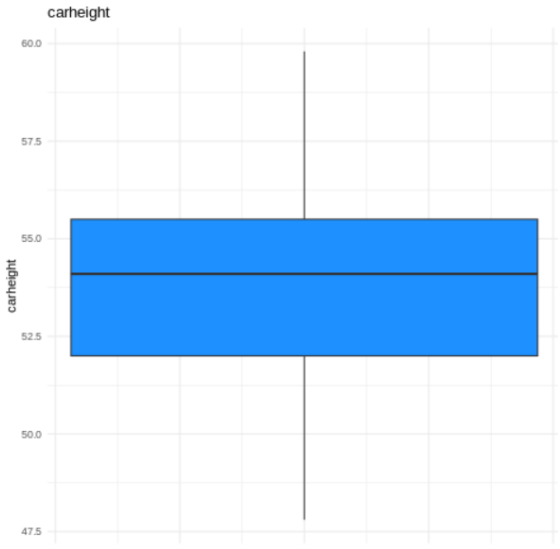

peakrpm
6000
5000
peakrpm

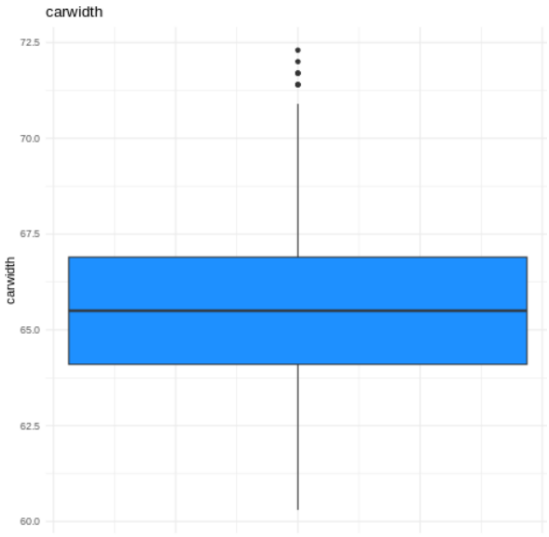

citympg
50
40
30
20
citympg


highwaympg
50
40
30
20
highwaympg


symboling
3
2
1
0
-1
-2
symboling


wheelbase
120
110
100
90
wheelbase





En base a los resultados obtenidos, podemos determinar que *symboling*, *carheight*, *curbweight*, *peakrm* y *carlenght* poseen casi o nada de datos atípicos, mientras que el resto de datos posee una cantidad moderada o mayor de este tipo de datos, pero los que más sufren de esto son las variables *carwidth*, *enginesize*, *stroke*, *compressionratio* y *horsepower*.

Con ayuda de la línea que representa a la mediana, podemos determinar qué tan sesgados se encuentran los datos de forma que se puede estimar si existe una distribución normal o no. Variables como *symboling*, *carwidth*, *enginesize*, *compressionratio*, *horsepower* y *highwaympg* son valores que tienden a una normalidad debido a que la línea media en cada una de sus boxplots se aproxima o esta muy cerca de la mitad, mientras que otras variables sufren demasiado de sesgo hacia valores bajos como *curbweight* o *wheelbase*, y altos como *carheight* o *stroke*.

A pesar de que estas variables contengan datos atípicos y algunos no presenten un comportamiento normal según estas gráficas, además de que en algunas se presente mayor variabilidad que en otras debido al tamaño de sus cajas, esto no nos indica que son completamente descartables o no, simplemente este recurso nos ayuda a comprender ciertas caracterísitcas que son relevantes para posteriormente, en conjunto con otros análisis, gráficas y valores, se seleccionarán las variables que mejor se adapten al modelo a implementar, por lo que aún es muy pronto para desechar variables cuantitativas.

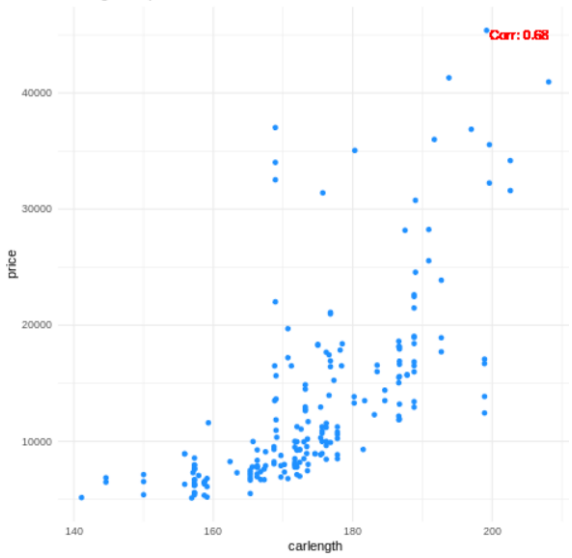
Diagramas de Dispersión

Desplegaremos todas las posibles combinaciones que podamos relizar comparando todas las variables con todas, con el fin de evaluar las posibles relaciones que exisan entre las variables. Para ello, necesitamos tomar en cuenta los posibles patrones que podamos encontrar, ya sean positivos, negativos o simplemente no exista alguna relación entre dichas variables. Entre otras cosas que podremos identificar están los posibles datos atípicos, la concentración y forma de ajustar de los datos, correlaciones que peuden variar entre -1 (correlación negativa perfecta) a 1 (correlación positiva perfecta) y 0 (no existe correlación lineal).

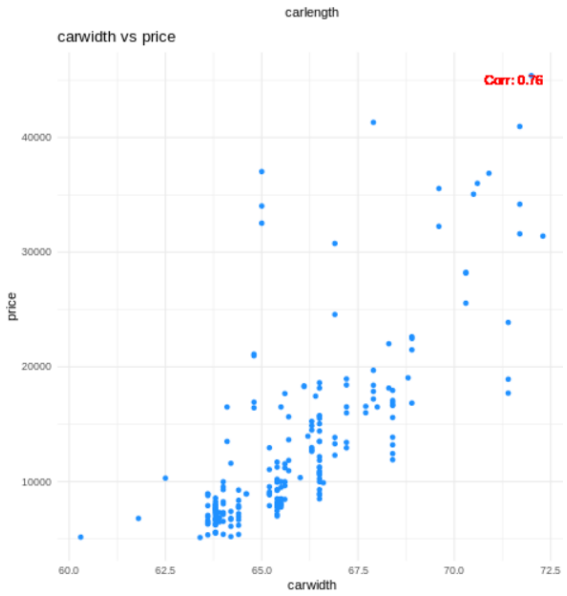
Tomando en cuenta lo mencionado anteriormente, se identificarán, detectarán, evaluarán y explorarán las relaciones, tendencias y ajustes de las variables con el fin de definir de forma precisa cuáles son las variables influyentes que vale la pena tomar en cuenta para el posterior análisis con respecto a la variable "precio":

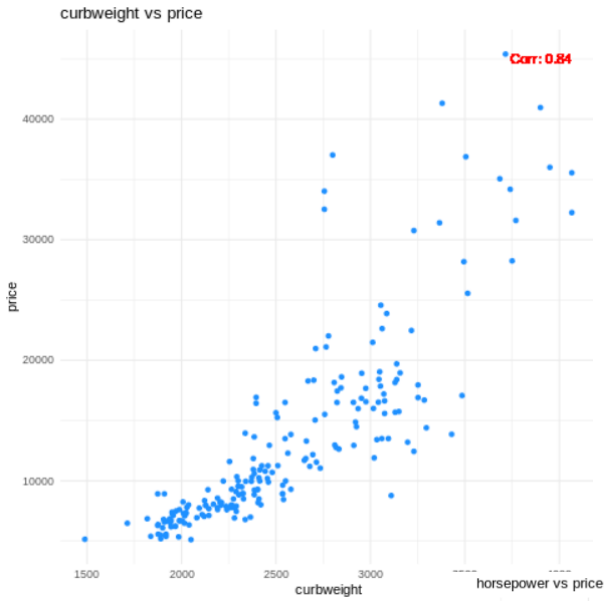
---

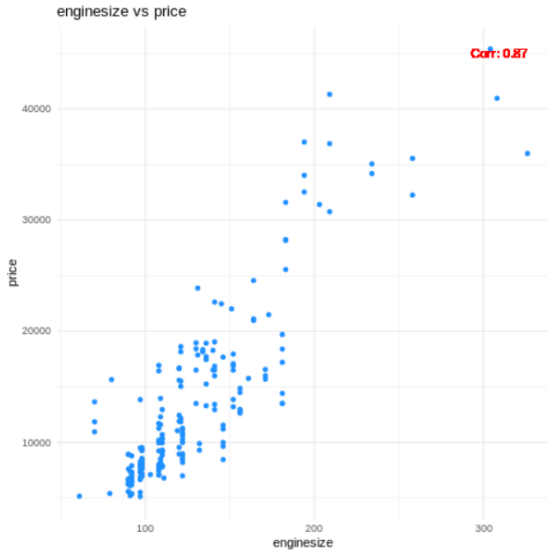
wheelbase



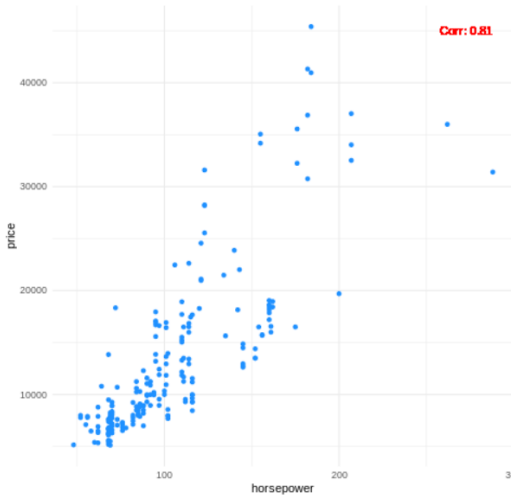
carlength

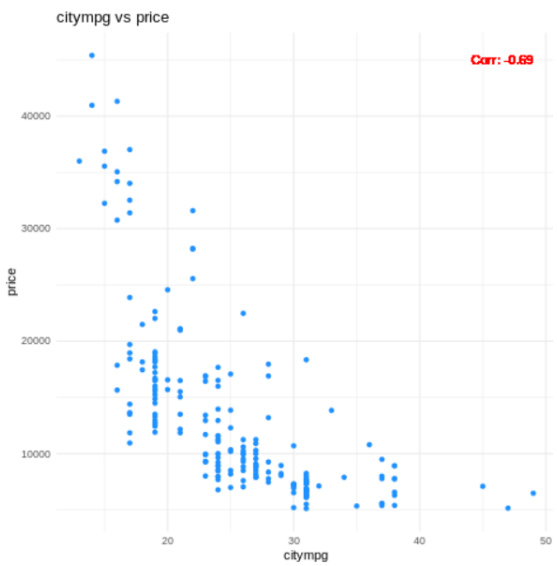


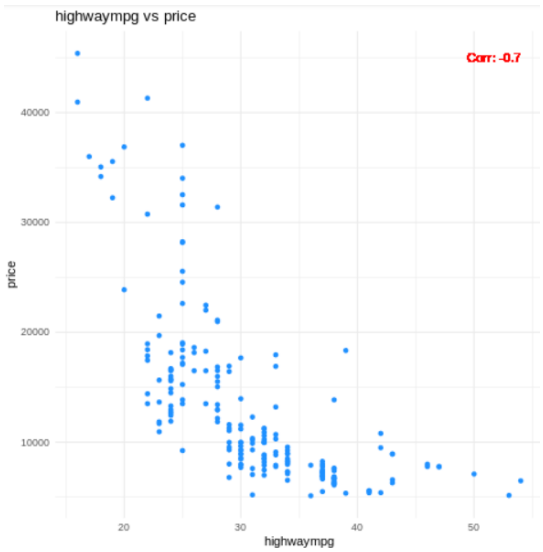




horsepower vs price







Análisis de Correlación

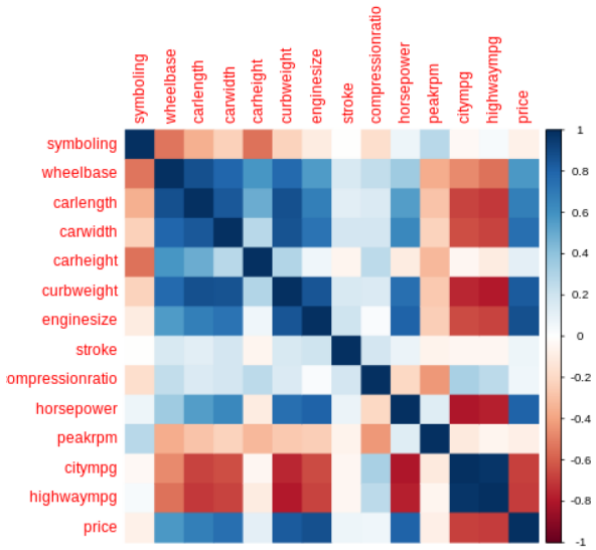
Como podemos observar, muchas de las gráficas generadas no demuestran tener algún tipo de relación directa con respecto al resto de las variables, inclusive hay variables que presentan comportamientos bastante peculiares pero nada de liniealidad ni presentan algún tipo de correlación entre ellas. Aún así, podemos destacar otras variables que sí se ajustan entre ambas y demustran tener poca o mucha linealidad. Por ejemplo, podemos destacar gráficas como:

- wheelbase vs price (coeff 0.58)
- carlenght vs price (coeff 0.68)
- carwidth vs price (coeff 0.76)
- curbweight vs price (coeff 0.84)
- enginesize vs price (coeff 0.87)
- horsepower vs price (coeff 0.81)
- citympg vs price (coeff -0.69)
- highwaympg vs price (coeff -0.7)

De las 14 variables cuantitativas, solo 9 son las que mejor correlación presentan con respecto al precio de los autoa. Aún así, analizaremos en las siguientes secciones si es necesario excluir otras variables o quedarnos con el conjunto elegido hasta el momento.

Matriz de Correlación

Esta gráfica es un complemento más para lo visto en la sección anterior, se busca ahondar un poco más en la correlación de estas variables con respecto al precio:

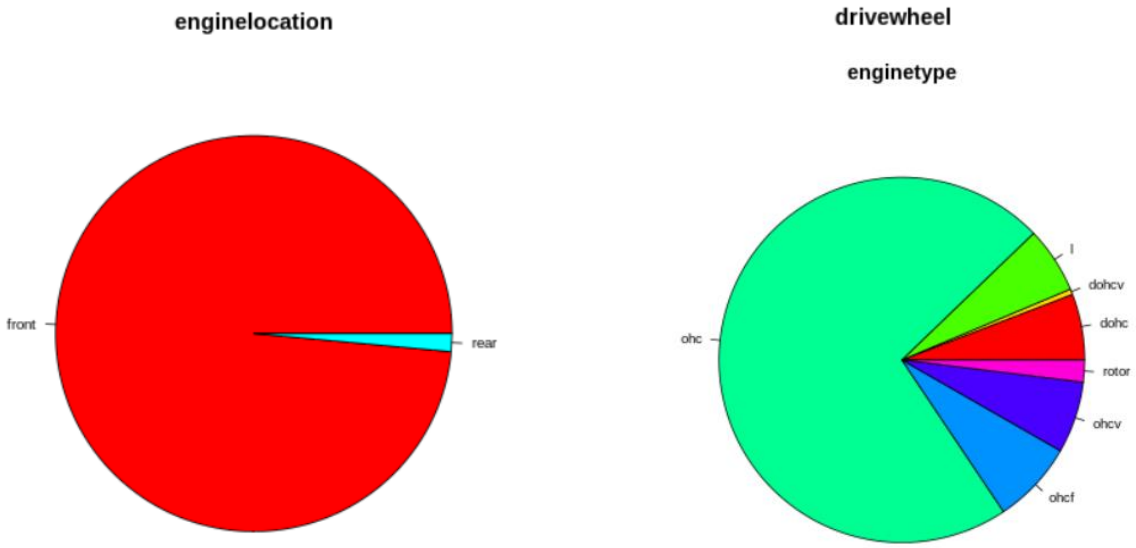


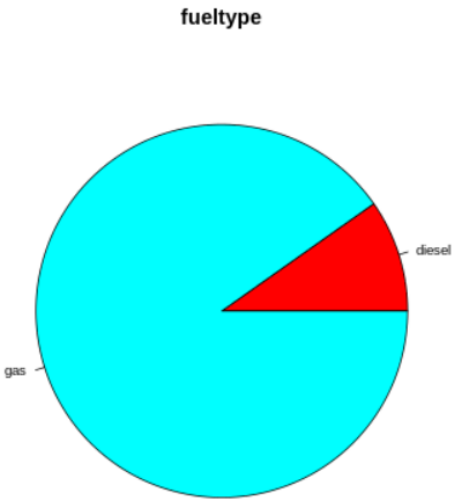
Visualización de Variables Cualitativas (Categóricas)

Diagrama de Pastel

Los diagramas de pastel a elaborar consideran el cálculo de tablas de frecuencia para cada una de las variables categóricas. Utiliza el dataset "plot_data" y agrupa los datos por valores categóricos y sus valores únicos. Este tipo de procedimientos se realizan para entender de mejor manera la composición de las variables categóricas y sus frecuencias relativas dentro del set de datos:

Gráfica de Barras





Con respecto a las gráficas obtenidas, podemos observar que los agrupamientos de las gráficas con mayor variedad de tipo de datos son:

- *Carname*
- *cardbody*
- *cylindernumber*
- *enginetype*

Y con respecto al resto de variables categóricas, estas son las que menos variedad de datos tienen:

- *drivewheel*
- *enginelocation*
- *fueltype*

Ahora que tenemos los resultados de estas gráficas de frecuencias y agrupamientos relativos de datos, podemos entender claramente la distribución de los datos con sus distintos grupos; esto nos puede mostrar perspectivas de cuáles categorías son las más o menos predominantes dentro del dataset.

El hecho de poder conocer la distribución de lso datos categóricos resulta muy útil para enfocarnos en las variables que tienden a tener un impacto significativo en el valor objetivo, en este caso es *price*. Además, este procedimiento nos ayuda a identificar potenciales retos, como lo es el desequilibrio de clases, el cual puede afectar el rendimiento del modelo. Evaluar este tipo de problemas durante el desarrollo del modelo estadístico asegura estimaciones más certeras y precisas. Siendo que las variables categóricas pueden variar entre pocos o muchos subconjuntos de datos, puede indicarnos cómo es que influye est variación con el precio de los autos.

Identificando Problemas de Calidad

Con base en el análisis realizado utilizando la base de datos de los precios de autos, podemos concluir que variables
como *wheelbase*, *carlenght*, *carwidth*, *curbweight*, *enginesize*, *horsepower*, *citympg* y *highwaympg* presentan una

correlación muy estrecha con la variable objetivo que es *price* de entre todas las variables cuantitativas, lo cual se ve reflejado claramente tanto en la matriz de correlaciones como en las gráficas de dispersión.

En cuanto a valores atípicos y como se mencionó en la sección de *boxplots*, se puede apreciar gracias a las gráficas de caja y bigotes que variables como *symboling*, *carheight*, *curbweight*, *peakrm* y *carlenght* presentan valores atípicos mínimos, casi o completamente nulos, lo cuál sugiere que su impacto dentro de inferencias estadísticas será menor si nos referimos a que podrían afectar el análisis de los datos. Por otro lado, variables como *carwidth*, *enginesize*, *stroke*, *compressionratio* y *horsepower* presentan más irregularidades en cuanto a sus datos. Estos datos atípicos pueden presentar bastantes distorsiones para posibles análisis de datos, claro que se podrían excluir con métodos como transformaciones o métodos estadísitcos robustos para regularizar este comportamiento, con el fin de no descartar por completo dichas variables y además obtener resultados precisos para posibles pruebas e investigaciones a futuro.

Ahora, con respecto a los datos faltantes dentro de la base de datos proporcionada, encontré un par de irregularidades que se presentan más dentro de las variables categóricas que en las cuantitativas. Algunas de estas son errores de dedo, datos mal escritos, incompletos o bien que no te dicen nada sobre el valor de la variable, pero son desperfectos mínimos, casi no se presentan casos de este tipo dentro de los datos. Si hablamos de los datos cuantitativos, es más difícil determinar si los números son datos erróneos a simple vista, de hecho las posibles irregularidades (datos atípicos) que existan se pueden identificar con el análisis explicado anteriormente, pero si hablamos de datos faltantes no encontré algún campo vacío o bien datos no numéricos.

Selección de Variables a Utilizar

Tomando en cuenta las conclusiones planteadas anteriomente, he determinado que las variables cuantitativas más útiles para un análisis más profundo que mejore la solidez de las conclusiones que podamos formular son las siguientes:

- *Highwaympg*
- *Citympg*
- *Horsepower*
- *Enginesize*
- *Curbweight*
- *Carwidth*
- *Carlength*
- *Wheelbase*

Si hablamos de variables categóricas, estás serían las seleccionadas para aportar al análisis por cuestión tanto de lógica como de los resultados realizados previamente:

- *fueltype*
- *cardbody*
- *drivewheel*
- *enginelocation*
- *enginetype*

- *cylindernumber*

Además de ser datos que por lógica de relación consideraríamos como relevantes para su posterior análisis, también el hecho de que se haya analizado estos datos por medio de tablas y gráficas nos ayuda a entender de mejor forma su comportamiento, distribución y cantidad de agrupamientos existentes, y con ello determinar si en realidad aportan o se ajustan al contexto con el que deseamos trabajar.

Preparación de la Base de Datos

Filtrado y Limpieza de los Datos Categóricos

En primera instancia verificaremos si las variables categóricas presentan valores faltantes o NAs en su defecto:

 # Filtramos y verificamos si existen campos vacíos en las variables categóricas

Variables categóricas faltantes: 0

Básicamente se tomaron las columnas de interés para su posterior uso y análisis, y con ello obtenemos los siguientes datos:

 # Seleccionamos solamente las variables categóricas del dataset
categorical_vars <- data %>% select_if(~ !is.numeric(.))

 # Creamos un dataframe solamente incluyendo las variables categóricas seleccionadas para nuestro anál isis

---

Filtrado y Limpieza de los Datos Cuantitativos

Dentro de esta sección se realiza un procemiento muy similar al anterior pero con lo valores cuantitativos. Al preprocesar los datos, se seleccionan y se extrae el conjunto de interés, se filtran los datos atípicos que se encuentren fuera del rango que se considera respecto a los intercuantiles (Q1 y Q3) de forma que tenemos un rango [Q1 - 1.5 * IQR, Q3 + 1.5 * IQR], excluyendo los valores que se encuentren fuera del mismo.

 # Aplicamos la función de filtrado para cada una de las columnas en la variables "quantitative_vars"

 # Seleccionamos solamente las columnas deseadas para guardarlas dentro de "quantitative_vars_cleane d"

Unión de los Datos Cuantitativos con los Categóricos

Tomamos tanto los valores cuantitativos como los categóricos y los fusionamos en un mismo conjunto de datos

---

Limpiamos Posibles Datos Faltantes (NA)

Este paso ya se había realizado con los datos categóricos, sin embargo para tomar precauciones y evitar interferencias con el correcto análisis de los datos, se buscarán campos vacíos dentro del conjunto de datos completo.

Regresión Lineal Múltiple

Fórmula de la regresión lineal

Dentro de este apartado implementaremos la fórmula de regresion lineal múltiple que considera a *price* como la variable dependiente de las variables independientes que elegimos en las secciones anteriores.

Cabe resalta que en un principio abordaremos este análisis sin realizar algún tipo de transformación o uso de otra herramienta estadística que busque normalidad de momento, con el fin de observar cómo se comportan los datos sin estos recursos.

$$\begin{aligned} \boldsymbol{price} = {} & \boldsymbol{\beta 0} + \boldsymbol{\beta 1} \times \boldsymbol{Highway\ MPG} + \boldsymbol{\beta 2} \times \boldsymbol{City\ MPG} + \boldsymbol{\beta 3} \times \boldsymbol{Horsepower} + \boldsymbol{\beta 4} \times \boldsymbol{Engine\ Size} \\ & + \boldsymbol{\beta 5} \times \boldsymbol{Car\ Width} + \boldsymbol{\beta 6} \times \boldsymbol{Car\ Length} + \boldsymbol{\beta 7} \times \boldsymbol{Wheelbase} + \boldsymbol{\beta 8} \times \boldsymbol{Curb\ Weight} \\ & + \boldsymbol{\beta 9} \times \boldsymbol{Fuel\ Type} + \boldsymbol{\beta 10} \times \boldsymbol{Car\ Body} + \boldsymbol{\beta 11} \times \boldsymbol{Drive\ Wheel} + \boldsymbol{\beta 12} \times \boldsymbol{Engine\ Location} \\ & + \boldsymbol{\beta 13} \times \boldsymbol{Engine\ Type} + \boldsymbol{\beta 14} \times \boldsymbol{Cylinder\ Number} \end{aligned}$$

Evaluación de las variables más apropiadas para realizar la regresión lineal

En base al modelo de regresión lineal creado por la función *lm()*, se realizará un ajuste a las variables consideradas para ser parte del modelo. Ya sea que se quiten o se agreguen variables con el procedimiento de selección del modelo de regresión paso a paso.

Residual standard error: 2450 on 177 degrees of freedom

Multiple R-squared: 0.7724

Adjusted R-squared: 0.7376

F-statistic: 22.24 on 27 and 177 DF

p-value: < 2.2e-16

Los resultados de la evaluación del modelo revelan importantes conocimientos sobre su desempeño. El error estándar residual (RSE), que mide la variabilidad residual, es de aproximadamente 2450. Un RSE más bajo implica un mejor ajuste del modelo. El valor de R cuadrado múltiple de 0,7724 significa que alrededor del 77,24% de la variación de los precios se explica por variables predictivas. Aunque el R cuadrado ajustado (0,7376) se ajusta a la complejidad del modelo, sugiere que es posible que los predictores adicionales no mejoren significativamente el rendimiento. Además, el modelo muestra significación estadística en general, con un estadístico F de 22,24 y un valor p muy bajo (< 2,2e-16). En resumen, el modelo demuestra un poder explicativo decente con un R cuadrado alto, pero la redundancia potencial en algunos predictores (por ejemplo, mpg en ciudad y longitud del automóvil) exige una consideración cuidadosa de la relevancia de las variables y un mayor refinamiento.

Volvemos a declarar la función de Regresión Lineal

Esto se realiza con la finalidad de incluir el resultado proporcionado por la función *step()*, que evaluó diferentes modelos mediante la adición o supresión de posibles variables predictivas para nuestro modelo *A*. Las variables que se consideraron conservar en el modelo son:

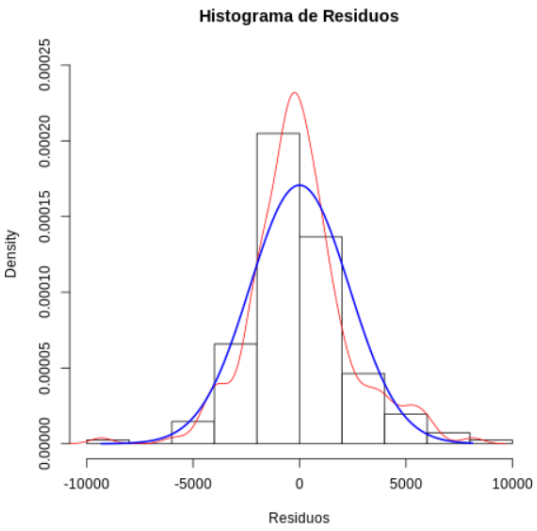
1. **Variables Cuantitativas Conservadas:**
    - *Citympg*
    - *Horsepower*
    - *Enginesize*
    - *Carwidth*
    - *Wheelbase*
    - *Curbweight*
2. **Variables Categóricas Conservadas:**
    - *Fueltype*
    - *Carbody*
    - *DriveWheel*
    - *Enginelocation*
    - *Cylindernumber*

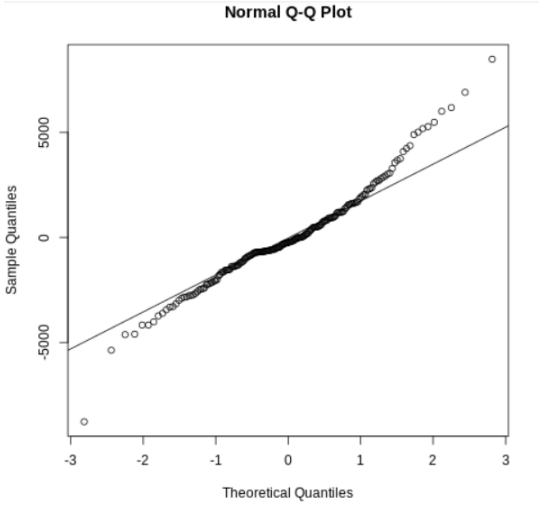
Con respecto a las variables cuantitativas se suprimió a *highwaympg*, *carlenght* y *enginetype*, y en cuanto a las variables categóricas se suprimió solamente a *drivewheel*, por lo que estaremos trabajando con estas variables independientes para posteriormente evaluar este modelo.

$$\boldsymbol{price = \beta 0 + \beta 1 \times citympg + \beta 2 \times horsepower + \beta 3 \times enginesize + \beta 4 \times carwidth + \beta 5 \times wheelbase + \beta 6 \times curbweight + \beta 7 \times fueltype + \beta 8 \times carbody + \beta 9 \times drivewheel + \beta 10 \times enginelocation + \beta 11 \times cylindernumber}$$

Normalidad de los Residuos

Utilizando herramientas de análisis estadístico, podemos observar cómo la gráfica de histograma nos muestra un comportamiento bastante cercano a una normal debido a su distribución y comportamiento, a pesar de ello, la gráfica de QQPlot nos indica que los datos de residuo no se comportan como una normal debido a que ciertos datos están esparcidos por ambos extremos de la distribución de los datos. Lo que se prentede realizar en base a estos resultados es ajustar lo mejor posible los datos, ya sea excluyendo/filtrando variables, normalizando los datos o bien haciendo transformaciones, esto con el fin de obtener un dataset confiable del cual podamos confiar y nos ayude a establecer modelos con mejores resultados.



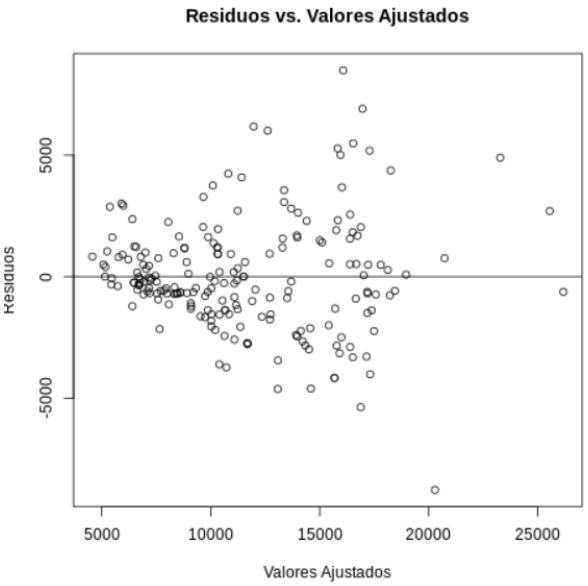


Homocedasticidad

En cuanto a la gráfica de dispersión que proyecta homocedasticidad de los datos, en realidad no demuestra homocedasticidad casi para nada. Existen ciertos datos que no existe un rango consistente de los residuos con respecto a los valores ajustados.

Otro indicio muy importante a destacar es que dichos valores poseen una forma muy peculiar, podríamos decir que es como una especie de cono, lo cual no se relaciona para nada con homocedasticidad, lo que demuestra nuevamente que los datos no tienen un comportamiento constante a través del rango de valores ajustados.

Dichas observaciones pueden interpretarse de forma que, tanto los p-values, como los coeficientes, y posibles intervalos de confianza que calculemos o bien ya tenemos calculados, no nos sirven para nada, ya que esto se trata más bien de un caso de heterocedasticidad más que otra cosa.



Prueba de Hipótesis

Para esta situación en concreto, he definido una función de hipótesis en la cual para cada una de las variables se debe cumplir que:

- H0:p−value<0.05
- H1:p−value>=0.05

Esto se hace con el fin de determinar qué tipo de variables son convenientes dejar en el modelo y cuales deben ser removidas por su carencia de relación con respecto a la variable "precio". El t-value de cada uno de estos valores ayudará a determinar de mejor forma qué variable no es estadísticamente significativa para el contexto del problema. Entre más grande sea el t-value y menor su distancia con respecto a la media, la variable resultará más significativa al modelo, por el contrario, entre más pequeño sea el t-value y mayor sea su distancia con respecto a la media, menor influencia y más irrelevante se vuelve la variable.

Residuos:

- El residual mínimo es -8776,3 y el residual máximo es 8486,9.
- Los residuos tienen un valor mediano de -249,4, lo que indica que, en promedio, el modelo subestima en esta cantidad.
- El rango intercuartil (RIC) de los residuos es de -1214,2 (Q1) a 1158,4 (Q3).

Coeficientes:

- Los coeficientes representan los efectos estimados de cada variable predictiva sobre la variable de respuesta (precio).
- La columna "Estimación" muestra los valores de coeficiente estimados.
- El "Error estándar" representa el error estándar de las estimaciones de los coeficientes.
- El "valor t" es el estadístico t para probar la hipótesis nula de que el coeficiente es igual a cero.
- "Pr(>|t|)" es el valor p asociado con la prueba t para cada coeficiente. Los valores de p más bajos indican una mayor significancia.

Error Estándar Residual:

- Los códigos de significancia indican el nivel de significancia estadística de cada coeficiente: '' (altamente significativo), '' (significativo), '' (marginalmente significativo), '.' (insignificante).

Rendimiento del modelo:

- El R cuadrado ajustado del modelo es 0,7376, lo que indica que aproximadamente el 73,76% de la varianza en la variable de respuesta (precio) se explica por las variables predictoras.
- El estadístico F prueba si alguna de las variables predictivas tiene un efecto general significativo en la respuesta. El valor p es cercano a cero, lo que indica que al menos un predictor es significativo.

Error estándar residual:

- El error estándar residual es 2450, lo que representa la desviación estándar de los residuos. Mide el error promedio entre los valores observados y predichos.

Conclusiones sobre la prueba de hipótesis

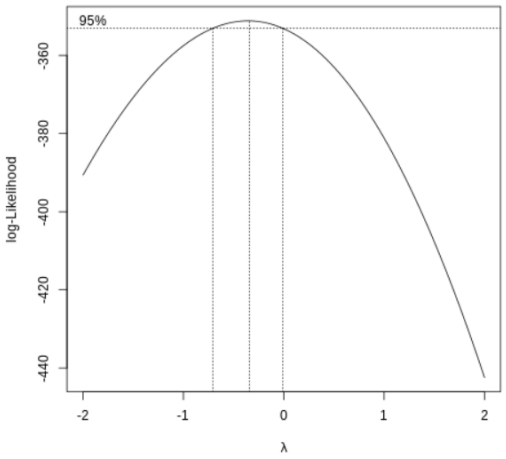
Teniendo en cuenta la prueba de hipótesis que se realizó acerca de los t-values de cada una de las variables y en términos de la misma, se buscaba rechazar esta función de hipótesis, con el fin de descartar las variables que no plasmaran una significancia relevante para el contexto del problema, y esto nos ayuda a filtrar de mejor manera cada una de las variables que influyen en el modelo. En base a esta tabla proporcionada por la función de summary() podemos destacar a las variables citympg, horsepower, curbweight, fueltypediesel, carbodyconvertible y carbodysedan como potenciales variables significativas para predecir con mejor precisión a price. Otro valor que da soporte a la significancia de estas variables es el t-value, por ejemplo, en cuanto a la variable "horsepower" podemos identificar que este cuenta con un t-value de 2.648, lo que nos indica que este coeficiente se encuentra 2.648 veces lejano a el error estándar lejos de 0, esto también lo podemos observar con citympg, aunque su coeficiente negativo nos indica una relación inversa con respecto a price, esto además de tener un sentido estadístico es también lógico, ya que al aumentar

las millas por galón de un auto, su valor decrese. Si hablamos de curbweight, su coeficiente estimado de 4.526 nos dice que en promedio, el valor estimado del precio del automóvil aumenta aproximadamente $4,526; fueltypediesel nos muestra que los vehículos con motor diesel tienden a ser más caros que los vehículos motorizados por gasolina.

Otras características a mencionar son que el modelo en su totalidad es altamente significante, por lo menos con una variable, debido a que existe un p-value: < 2.2e-16 menor comparado a la estadística F. En cuanto al valor de la r^2, podemos determinar que el 80.97% explica la proporción de variabilidad en la variable de respuesta del modelo.

Aplicamos una transformación de datos

Esta transformación que realizaremos es un recurso bastante útil para encontrar una distribución normal de los datos. Para ello, necesitaremos recurrir a la transformación de Box-Cox. Para ello necesitamos tomar las variables que destacaron una correlación fuerte con el precio.

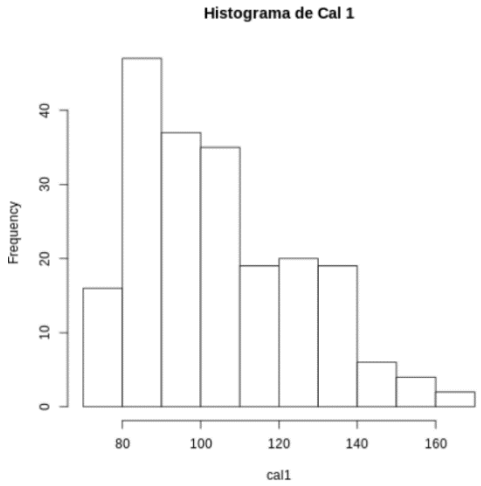


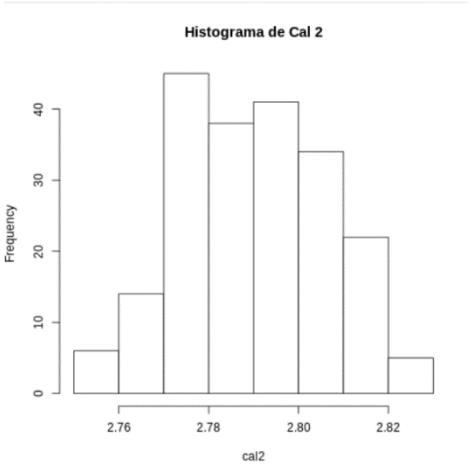
La gráfica de Box-Cox nos permite identificar qué valor de lambda es el que maximiza la función de verosimilitud (o bien minimiza alguna otra medida de no linealidad). Debido a que el valor de lambda arrojado por la función es negativo, esto indica que con respecto a los valores de la variable dependiente *price* se requiere aplicar una transformación recíproca para ajustar estos datos a una normalidad.

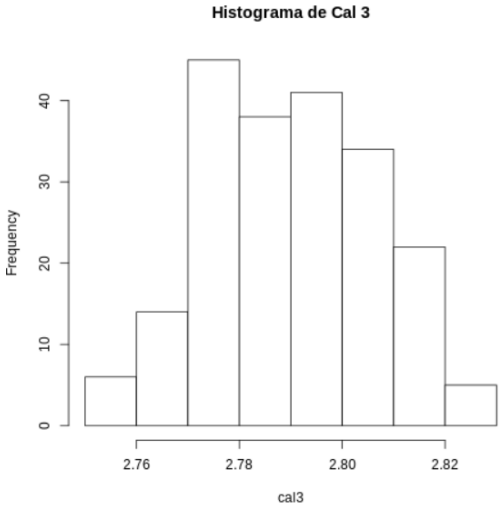
Para transformar los valores de forma correcta, implementaremos 3 tipos de transformaciones, transformación de raíz cuadrada, transformación con valor exacto y transformación de Yeo-Johnson. Teniendo estos resultados, determinaremos la mejor transformación y procederemos a utilizar la que mejor resultados arroje.

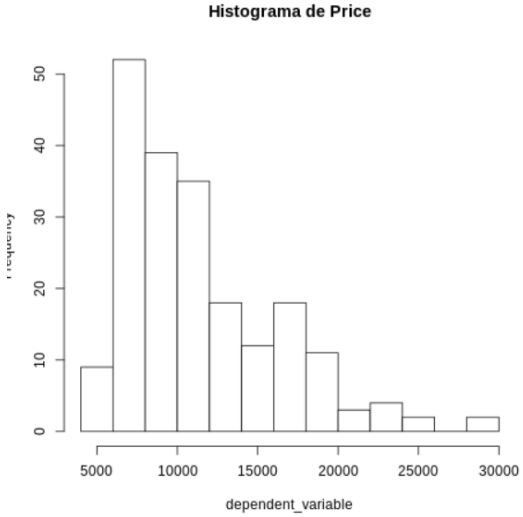
Aplicamos las 3 transformaciones

Realizamos las 3 transofmraciones, además de realizar pruebas de normalidad para corroborar que dichas pruebas arrojaron buenos resultados









En base a los resultados recibidos por la variable dependiente original son que la prueba de Anderson-Darling rechaza fuertemente la hipótesis nula de que la variable dependiente sigue una distribución normal. El valor p es demasiado pequeño (casi 0), lo que da indicios de que los datos se desvían significativamente de la normalidad. En cuanto a su histograma, podemos observar que dicha gráfica presenta un sesgo muy marcado hacia la derecha y una curtosis moderada, presentando pocos valores extremos comparandola con una distribución normal, aunque se aleja bastante de ser una distribución cercana a la normalidad.

En cuánto a los resultados de la transformación de raíz cuadrada, tenemos que es muy similar al comportamiento de la variable dependiente original. La prueba de Anderson-Darling revela un valor p muy pequeño cercano a cero, por lo que similar al caso original, por lo que se rechaza la hipótesis nula que tiene *cal1* respecto a una distribución normal. Su histograma se encuentra un poco más cercano a algo normal, sin embargo los picos que posee son bastnate pronunciados como para comportarse con normalidad.

El caso de transformación con valor exacto arroja un resultado menos concluyente, comparado con las otras 2 transformaciones. Mientras que el valor p es relativamente pequeño, es más largo que en los casos anteriores. Esto sugiere que *cal2* puede desviarse de una distribución normal hasta cierta medida, pero la evidencia para no normalidad no es tan sólida. Y será cierto que su histograma tiende a comportarse más como una normal que el

par de histogramas correspondientes a cal1 y el original respectivamente, sin embargo esto no lo hace suficiente como para considerar una buena normalidad.

El último modelo que respecta al de Yeo-Johnson no es muy distinto de *cal2*, prácticamente es el mismo modelo, muestra el mismo nivel de normalidad, que no es mucho pero es lo más cercano que se ha obtenido hasta ahora, posee el mismo valor p que *cal2*, así que sufre de la misma cantidad de desviación con respecto a una distribución normal.

Cabe resaltar que no se espera obtener un valor una distribución de los datos perfecta, puesto que los datos no necesariamente deben estar hechos para que cualquier herramienta estadística arroje los mejores resultados posibles, debido a que no siempre se trabajará con datos que se ajusten lo mejor posible a normalidad, homocedasticidad, dispersiones o correlaciones, esto a pesar de que se intente filtrar y organizar los datos de la mejor forma posible. Es por ello que se trabajará con el modelo, ya sea *cal2* o *cal3*, ya que son los que mejores resultados arrojaron de todas las transformaciones realizadas.

Aplicamos nuevamente la Regresión Lineal Múltiple

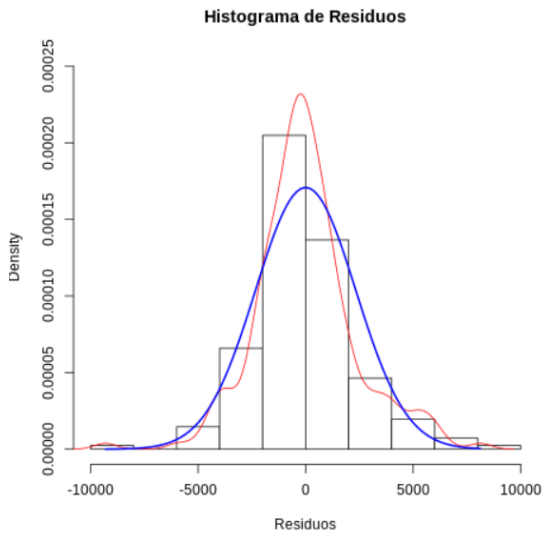
$$\boldsymbol{A = lm(price \sim citympg + yeo.johnson(horsepower, optimal\_lambda) + yeo.johnson(enginesize, optimal\_lambda) + yeo.johnson(carwidth, optimal\_lambda) + wheelbase + yeo.johnson(curbweight, optimal\_lambda) + fueltype + carbody + drivewheel + enginelocation + cylindernumber, data = final\_data)}$$

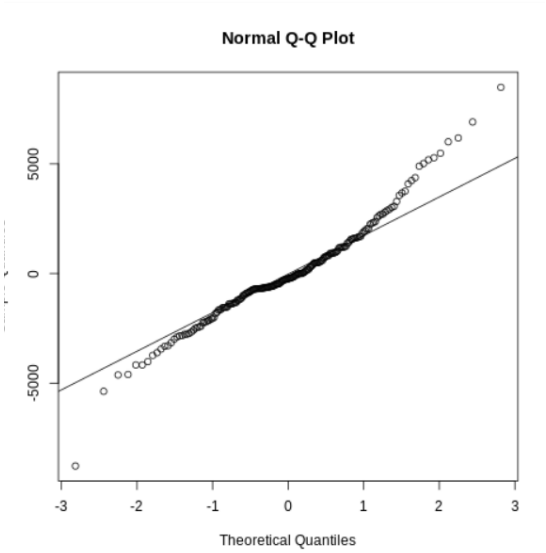
Teniendo los resultados de las transformaciones, y considerando las mejores variables independientes que mostraron una correlación fuerte con la variable dependiente *price*, se implementarán transformaciones en estas variables:

Variables Cuantitativas a Transformar

- *Horsepower*
- *Enginesize*
- *Carwidth*
- *Curbweight*

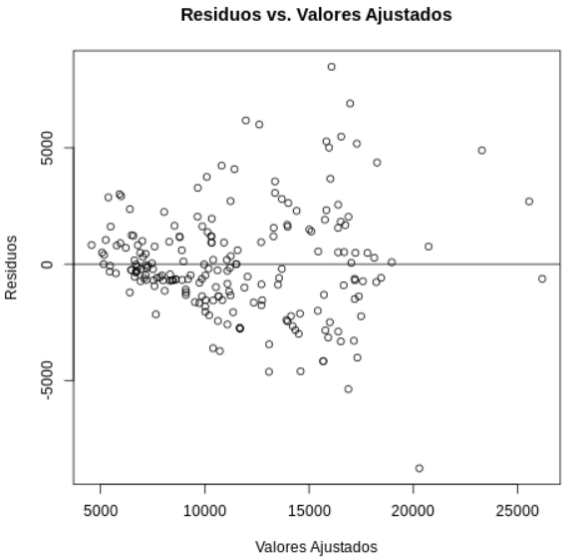
Normalidad de los Residuos





Por lo visto en las gráficas, no existe una mejora significativa con respecto a la QQPlot o el histograma mostrado, y si ha de existir algún tipo de cambio sería mínimo. El QQPlot sigue sin adaptarse a la línea recta diagonal del todo, a pesar de que en cierto punto logra de alguna manera acoplarse a la normalidad, los extremos siguen muy desfasados con respecto a la pendiente, por lo que no se logra una distribución normal de forma correcta.

Homocedasticidad



La gráfica de dispersión demuestra poca variación con respecto a la del modelo original. Es cierto que existe cierto cambo con respecto al comportamiento de los residuos de la anterior gráfica a la generada considerando una transformación de datos, aun así no logra una distribución uniforme alrededor de la línea horizontal y persiste en formar una figura en forma de cono o embudo, por ende, se puede concluir que existe mayor variabilidad en los valores más altos de la variable precio/*price* estimado.

Conclusiones

En este proyecto, nos embarcamos en un proceso integral a través del ámbito del análisis de regresión, con el objetivo de construir un modelo predictivo para un problema empresarial crítico. Nuestro objetivo era desarrollar un modelo que pudiera estimar de manera confiable el precio de los automóviles basándose en un conjunto de variables predictivas. En el camino, nos encontramos con varios desafíos, suposiciones y técnicas estadísticas complejas, que exploramos y abordamos meticulosamente.

Comenzamos con la exploración y el preprocesamiento de datos, donde limpiamos y transformamos el conjunto de datos para prepararlo para el modelado. Esta fase implicó el manejo de valores faltantes, detección de valores atípicos y selección de características. Profundizamos en las complejidades de la visualización de datos para obtener información sobre las relaciones entre variables y comprender su impacto en la variable objetivo, el "precio".

A continuación, nos aventuramos probando diversas herramientas que nos proporcionaron una concepción más clara de cómo se comportaban las variables y los datos en general. Aplicamos boxplots para conocer el rango, la tendencia central y la variabilidad de cada una de las variables, así como determinar cuáles y qué tantos datos atípicos poseen. Aplicamos también diagramas de dispersión, los cuales nos ayudaron a conocer la correlación que podía existir por cada una de las variables cuantitativas con respecto a la variable dependiente “precio”. Este paso fue clave para determinar cuáles de las variables valía la pena considerar para el análisis. Aunado a esto, tenemos una matriz de correlación que soporta este último punto. También, elaboramos gráficas de pastel y de barras con el fin de identificar la cantidad de agrupamientos que poseen cada una de las variables categóricas y como es que se distribuyen sus frecuencias.

Por último, para esta sección de explorar la base de datos, identificamos problemas de calidad tentativos dentro de la base de datos, ya sea datos atípicos, datos faltantes o campos vacíos, datos que requerían ser convertidos, o que presentaban algún error en su escritura, además de englobar lo ya mencionado acerca de la correlación de los datos y el ajuste encontrado con respecto a la variable objetivo. Finalmente, se seleccionaron las variables que se utilizarían para construir el modelo estadístico con el fin de llevar a cabo el análisis profundo del mismo, esto en base a la interpretación del análisis previo realizado con la base de datos.

A través del proceso de regresión por pasos, seleccionamos un subconjunto de variables predictivas que contribuyeron significativamente al poder predictivo del modelo y al mismo tiempo simplificaron su complejidad. Examinamos la normalidad de los residuos, probamos la multicolinealidad y evaluamos los supuestos del modelo para garantizar su confiabilidad.

Para mejorar el rendimiento del modelo, nos embarcamos en una serie de transformaciones y ajustes. Exploramos las transformaciones de Box-Cox y Yeo-Johnson para abordar cuestiones de no normalidad y heterocedasticidad en los datos. Empleamos herramientas de diagnóstico como gráficos residuales, gráficos QQ y gráficos de apalancamiento para identificar valores atípicos y observaciones influyentes.

Si bien hemos logrado avances significativos en la mejora de nuestro modelo inicial, reconocemos que modelar fenómenos del mundo real es un esfuerzo continuo. Puede haber factores y complejidades adicionales que no hemos considerado, y el rendimiento del modelo podría beneficiarse de una mayor exploración y refinamiento.

En conclusión, este proyecto sirve como testimonio del poder de la toma de decisiones basada en datos, pero también subraya los desafíos e imperfecciones inherentes al modelado de sistemas complejos del mundo real. A pesar de nuestros mejores esfuerzos, reconocemos que nuestro modelo no es perfecto y es posible que nunca alcance la elusiva distribución normal. Sin embargo, esto refleja la realidad que enfrentan los científicos y analistas de datos en su búsqueda por representar situaciones de la vida real, donde persisten irregularidades y complejidades.

Al concluir este proyecto, alentamos una mayor investigación y experimentación para construir sobre las bases establecidas aquí. Nuestra esperanza es que el conocimiento adquirido no solo sirva de base para las estrategias de fijación de precios en la industria automotriz, sino que también inspire esfuerzos futuros en el apasionante campo de la ciencia de datos y el análisis predictivo.