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:

Iván Mauricio Amaya Contreras

Blanca Rosa Ruiz Hernández

Antonio Carlos Bento

Frumencio Olivas Álvarez

Hugo Terashima Marín

Alumno:

Alberto H Orozco Ramos - A00831719

12 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 lenguaje 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 compañías 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 compañía 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

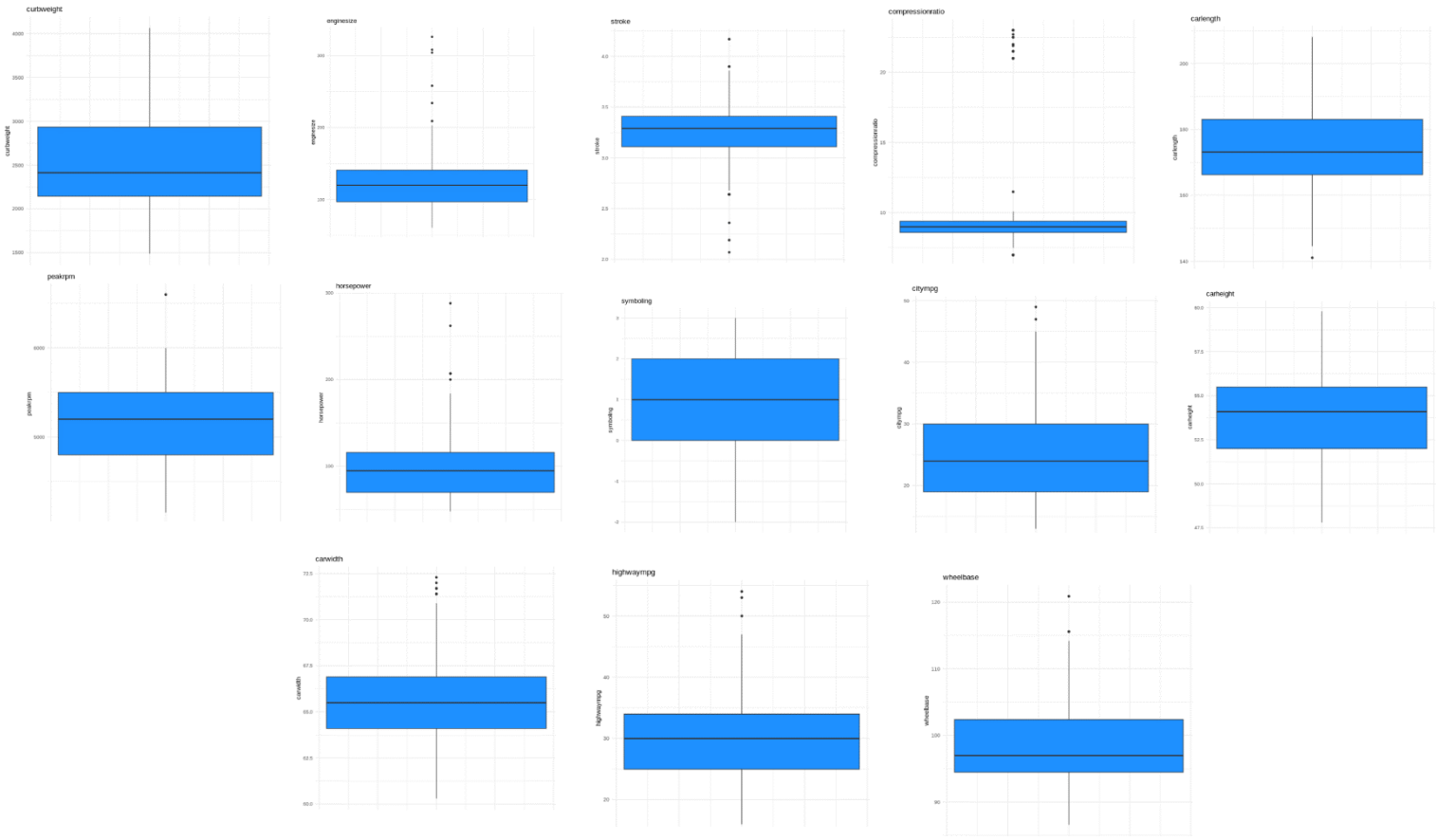
La situación esta subrayada por las nada estables, dinámicas del mercado global. Para denotar su relevancia, es esencial considerar que la literatura creciente estrategias de entrada al mercado y las dinámicas de precios para negocios internacionales. Entre más nos adentremos en este reporte, emplearemos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

Exploración de la Base de Datos

Visualización de Variables Cuantitativas

Bloxplot y Valores Atípicos

Se utilizó el análisis de boxplots para evaluar la variabilidad y presencia de datos atípicos en las variables cuantitativas. Se observó que algunas variables como symboling, carheight y curbweight tenían pocos datos atípicos, mientras que otras como carwidth, enginesize y horsepower presentaban más datos atípicos. No obstante, todas las variables se consideran relevantes en esta etapa inicial.



Diagramas de Dispersión

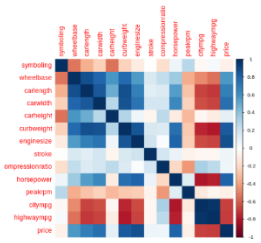
Se realizaron diagramas de dispersión para explorar las relaciones entre variables cuantitativas y su impacto en el precio de los automóviles. Se identificaron nueve variables con correlación significativa con el precio, incluyendo carwidth, enginesize y curbweight.

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 linealidad ni presentan algún tipo de correlación entre ellas. Aun así, podemos destacar otras variables que sí se ajustan entre ambas y demuestran tener poca o mucha linealidad. Por ejemplo, podemos destacar gráficas como:

Comparación	Coeficiente de Relación
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 autos. Aun así, analizaremos en las siguientes secciones si es necesario excluir otras variables o quedarnos con el conjunto elegido hasta el momento.

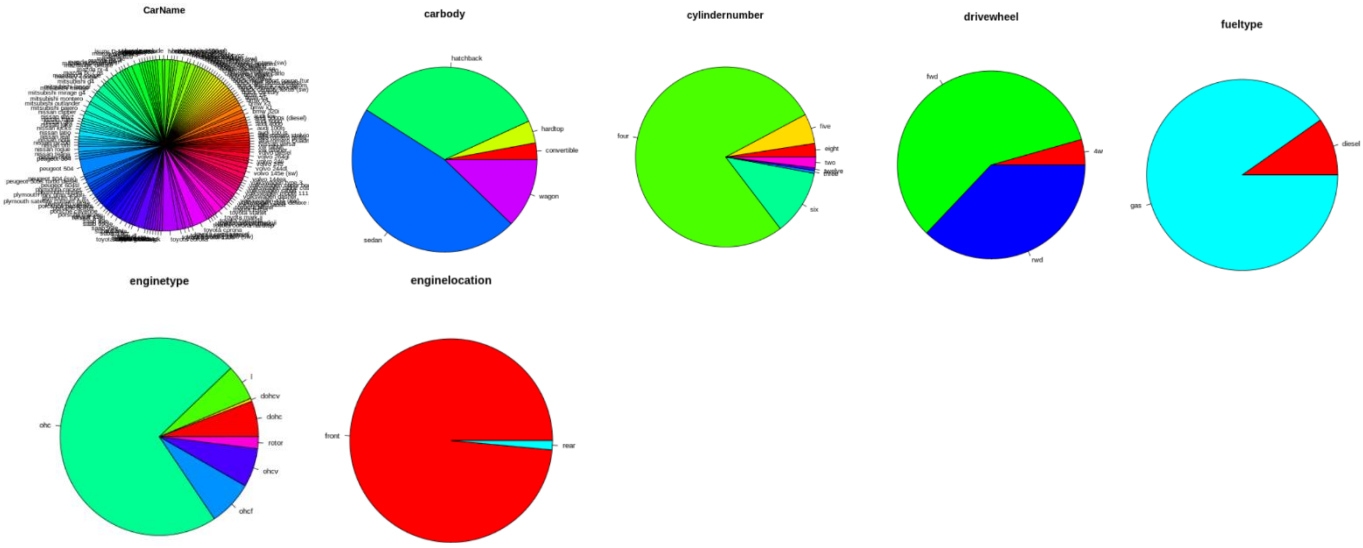


Matriz de Correlación

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)

Se emplearon diagramas de pastel y gráficos de barras para analizar las variables categóricas. Se destacaron variables como carname, carbody y cylindernumber por su variedad de datos. Esta exploración ayudó a comprender la distribución de las categorías y su posible influencia en el precio de los autos.



Gráfica de Barras

Con respecto a las gráficas obtenidas, p podemos observar que los agrupamientos de las gráficas con mayor variedad de tipo de datos son *carname*, *cardbody*, *cylindernumber* y *enginetype.*

Y con respecto al resto de variables categóricas, estas son las que menos variedad de datos tienen: *drivewheel*, *enginelocation* y *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 los 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 esta variación con el precio de los autos.

Identificando Problemas de Calidad

Se señalaron las variables con mayor correlación con el precio y aquellas con valores atípicos. Se consideró que las irregularidades podrían abordarse mediante transformaciones o métodos estadísticos robustos.

Selección de Variables a Utilizar

Se seleccionaron ocho variables cuantitativas y seis categóricas para un análisis más profundo basado en su relevancia y comportamiento observado.

Esta síntesis reduce la sección de exploración de la base de datos sin perder los puntos clave de análisis y selección de variables:

Variables Cuantitativas

Highwaympg, Citympg, Horsepower, Enginesize, Curbweight, Carwidth, Carlength y *Wheelbase.*

Variables Categóricas

Fueltype, cardbody, drivewheel, enginelocation, enginetype, cylindernumber

---

Preparación de la Base de Datos

Filtrado y Limpieza de los Datos Categóricos

Se realizó una limpieza de los datos categóricos para eliminar valores nulos y asegurar la calidad de los datos**.** Básicamente se tomaron las columnas de interés para su posterior uso y análisis, y con ello obtenemos los siguientes datos:

Filtrado y Limpieza de los Datos Cuantitativos

Se aplicó un procedimiento similar al anterior, pero enfocado en los datos cuantitativos. Se eliminaron valores atípicos que estaban fuera del rango definido por los cuartiles.

Unión de los Datos Cuantitativos con los Categóricos

Los datos cuantitativos y categóricos se combinaron en un solo conjunto de datos para su posterior análisis.

Regresión Lineal Múltiple

Fórmula de la regresión lineal

Se implementó un modelo de regresión lineal múltiple con "price" como variable dependiente y las variables seleccionadas como independientes. Se realizó este análisis sin transformaciones para evaluar el comportamiento inicial de los datos.

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

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

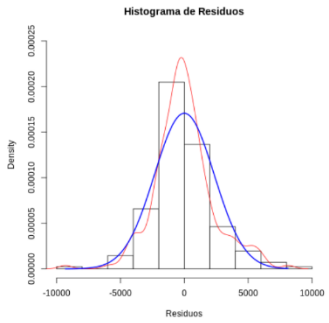
Con respecto a las variables cuantitativas se suprimió a *highwaympg*, *carlenght* y *enginetype*, y en cuanto a las variables categóricas se suprimió solamente a *drivewheel*, por lo que estaremos trabajando con estas variables independientes para posteriormente evaluar este modelo.

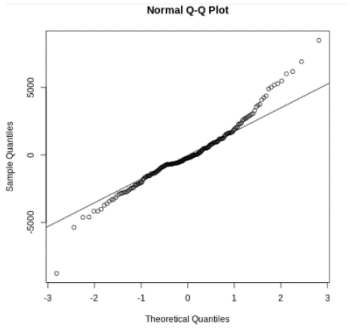
$$\boldsymbol{price = \beta 0 + \beta 1 \times citympg + \beta 2 \times horsepower + \beta 3 \times enginesize + \beta 4 \times carwidth + \beta 5 \times wheelbase + \beta 6 \times curbweight + \beta 7 \times fueltype + \beta 8 \times carbody + \beta 9 \times drivewheel + \beta 10 \times enginelocation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

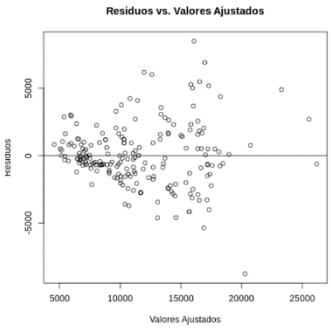
Normalidad de los Residuos

Se examinó la normalidad de los residuos utilizando gráficos de histograma y QQPlot. A pesar de la distribución aparentemente normal en el histograma, el QQPlot indicó que los residuos no siguen una distribución normal, ya que algunos datos se desvían de la línea de normalidad. Se consideraron posibles transformaciones y ajustes para

obtener un conjunto de datos más confiable.







Homocedasticidad

Se evaluó la homocedasticidad de los datos mediante gráficos de dispersión. Los resultados no mostraron homocedasticidad significativa, ya que algunos residuos mostraron un comportamiento inusual y no constante a lo largo del rango de valores ajustados. Esto planteó preocupaciones sobre la validez de las pruebas estadísticas realizadas.

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

En resumen, el modelo de regresión lineal múltiple mostró un poder explicativo decente con un R cuadrado alto. Sin embargo, se señaló la redundancia potencial en algunas variables y se destacó la necesidad de una consideración cuidadosa de la relevancia de las variables. Se sugirió un mayor refinamiento del modelo y posibles ajustes en función de los resultados de las pruebas estadísticas y la normalidad de los residuos.

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 Inter cuartil (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:

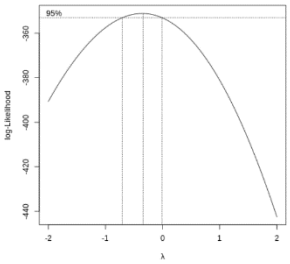
- El R cuadrado ajustado del modelo es 0,7376, lo que indica que aproximadamente el 73,76% de la varianza en la variable de respuesta (precio) se explica por las variables predictoras.
- El estadístico F prueba si alguna de las variables predictivas tiene un efecto general significativo en la respuesta.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

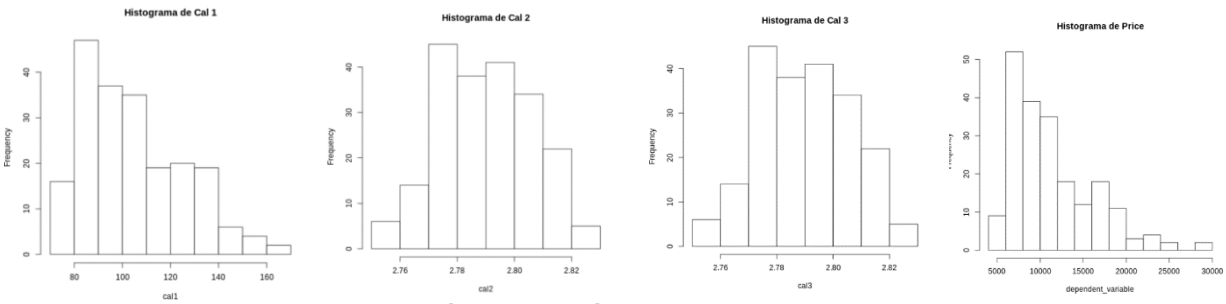
La prueba de hipótesis realizada en función de los t-values de las variables tuvo como objetivo identificar las variables más significativas para predecir con precisión el precio. Se destacan las variables citympg, horsepower, curbweight, fueltypediesel, carbodyconvertible y carbodysedan como potenciales variables significativas. El t-value de la variable "horsepower" se encuentra a 2.648 veces el error estándar lejos de cero, lo que indica su relevancia. El modelo en su conjunto resultó altamente significativo, con un valor p muy bajo y un R cuadrado del 80.97%.



Aplicamos una transformación de datos

Se aplicaron transformaciones de datos para buscar una distribución normal. La transformación de Box-Cox se utilizó en variables como horsepower, enginesize, carwidth y curbweight. Sin embargo, los resultados indicaron que ninguna de las transformaciones logró una distribución normal adecuada. El histograma y QQPlot aún mostraron desviaciones significativas de la normalidad.

Aplicamos las 3 transformaciones



Aplicamos nuevamente la Regresión Lineal Múltiple

Se implementó un modelo de regresión lineal múltiple que incluye las transformaciones realizadas en las variables cuantitativas mencionadas. Aunque se realizaron transformaciones, los resultados de normalidad de los residuos

no mostraron una mejora significativa. El QQPlot todavía no se ajustó completamente a la línea diagonal, y los extremos seguían desviados de la pendiente.

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

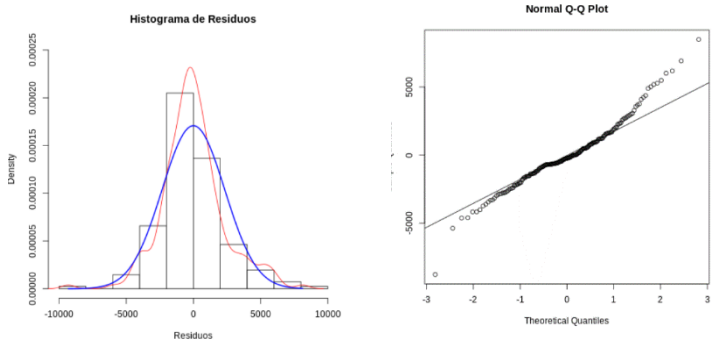
En resumen, a pesar de los esfuerzos por aplicar transformaciones para lograr una distribución normal de los datos, los resultados aún no indican una normalidad adecuada. Es importante considerar que los datos reales pueden no cumplir con todas las suposiciones de normalidad, y se trabajará con el modelo que arroje los mejores resultados de entre las transformaciones aplicadas.

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:

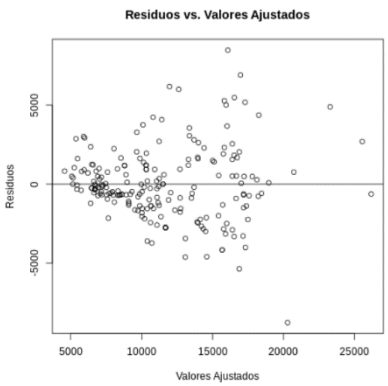
Variables Cuantitativas a Transformar

- *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 sumergimos en el análisis de regresión para construir un modelo predictivo que estime el precio de los automóviles basado en un conjunto de variables predictoras. Realizamos una exhaustiva exploración y preprocesamiento de datos, abordando problemas como valores faltantes, valores atípicos y selección de características. Utilizamos diversas herramientas de visualización para comprender las relaciones entre las variables y su impacto en el precio.

Luego, aplicamos un enfoque de regresión por pasos para seleccionar un subconjunto de variables predictoras significativas y simplificar el modelo. Evaluamos los supuestos del modelo y abordamos cuestiones de normalidad y heterocedasticidad mediante transformaciones de datos, como Box-Cox y Yeo-Johnson. A pesar de los avances, reconocemos que modelar fenómenos del mundo real es un esfuerzo continuo con desafíos inherentes.

En conclusión, este proyecto destaca la toma de decisiones basada en datos y subraya la complejidad del modelado de situaciones del mundo real. A pesar de las imperfecciones, alentamos la investigación futura en ciencia de datos y análisis predictivo para construir sobre estos cimientos y contribuir al campo y a la industria automotriz.

Links a GitHub

- Link anexo al portafolio de Análisis: TC3006_Portafolio_Analisis/final/M1_Statistics/Técnicas de Procesamiento de Datos para el Análisis Estadístico at main · 4lb3rt0r/TC3006_Portafolio_Analisis (github.com)
- Link anexo al portafolio de Implementación: TC3006_Portafolio_Implementacion/final/M1_Statistics/Técnicas de Procesamiento de Datos para el Análisis Estadístico at main · 4lb3rt0r/TC3006_Portafolio_Implementacion (github.com)