Instituto Tecnológico y de Estudios Superiores de Monterrey
Campus Monterrey.

Reporte El precio de los autos.

Alumno: Diego Alberto Baños Lopez
A01275100

Monterrey, N.L, a 11 de septiembre del 2023.

Resumen

Una empresa automovilística quiere reducir sus precios para ser mas competitivos, se entrego un conjunto de datos en donde primeramente realizamos una exploración inicial en donde se identificaron datos atípicos y transformamos los datos acordes a lo que se pide para primeramente identificar las variables que más influyeron dentro del precio, luego con ello normalizamos los datos y modelamos usando regresión lineal múltiple, una vez hecho ello se obtiene un modelo con una precisión del 82%

Introducción

En el mundo contemporáneo para poder prosperar dentro de un nuevo mercado una empresa debe de ser competitiva para poder posicionarse como una opción viable, para que una empresa sea competitiva tiene que optimizar precios y gestionar los gastos que tiene y para realizar ello una de las formas es comprender cuales son los factores que componen el precio del servicio o producto que se ofrece al cliente, que en este caso estamos hablando del precio de los automóviles de una marca, y ver qué tanta relación se tiene dentro del mismo, de ser contestada estas dos preguntas la empresa puede obtener una decisión más informada acerca de cómo construir sus coches para así optimizar costes, en este caso estamos hablando de una empresa automovilística de origen chino pueda entrar el mercado estadounidense y pueda hacerle frente a la competencia estadounidense, para resolver dichos dilemas se requiere de la realización de un análisis estadístico, y en base al análisis poder observar que son los factores más determinantes para el precio, para este caso, se seleccionaran 6 variables que se seleccionaran dentro de los datos promocionados por la misma empresa, y en base a ello contestar dichas cuestiones que se pueden ser resumidas en dos ¿Qué variables son significativas para predecir el precio de un automóvil? Y ¿Qué tan bien describen esas variables el precio de un automóvil?, para este ejercicio se decidió utilizar la herramienta de Python en su implementación de Jupyter Notebook, ya que estas herramientas nos ayudan a hacer el análisis estadístico de una forma mas comprensible, de igual forma se usarán librerías como pandas para leer el contenido del notebook, scikit-learn para el modelado, y seaborn y matplotlib para ayudarnos a hacer los gráficos

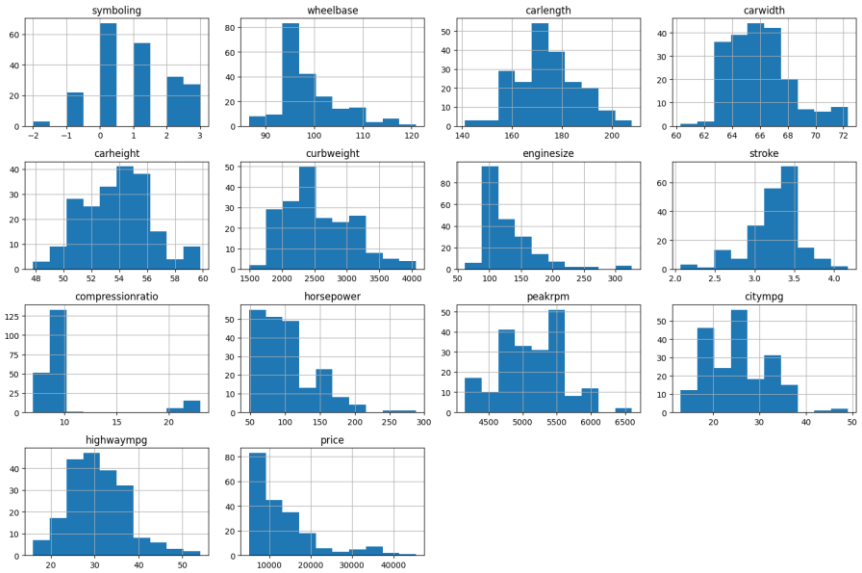
Análisis de los resultados:

Carga de los datos y observaciones iniciales:

Primeramente, se cargaron los datos con pandas, en el cual se cargo como un dataframe y usando pandas podemos observar que identificamos 205 registros y 26 características. Algunas de estas características eran numéricas, mientras que otras eran categóricas, es importante denotar ello para poder tratar a las categóricas más adelante, una vez hecho ello con nuestra herramienta estadística, se decidió que para entender mas los datos que nos arroja el código debemos de hacer gráficos y hacer más observaciones.

Análisis descriptivo y visualización:

Primeramente, decidimos hacer histogramas para las variables cuantitativas y observa la distribución de estas

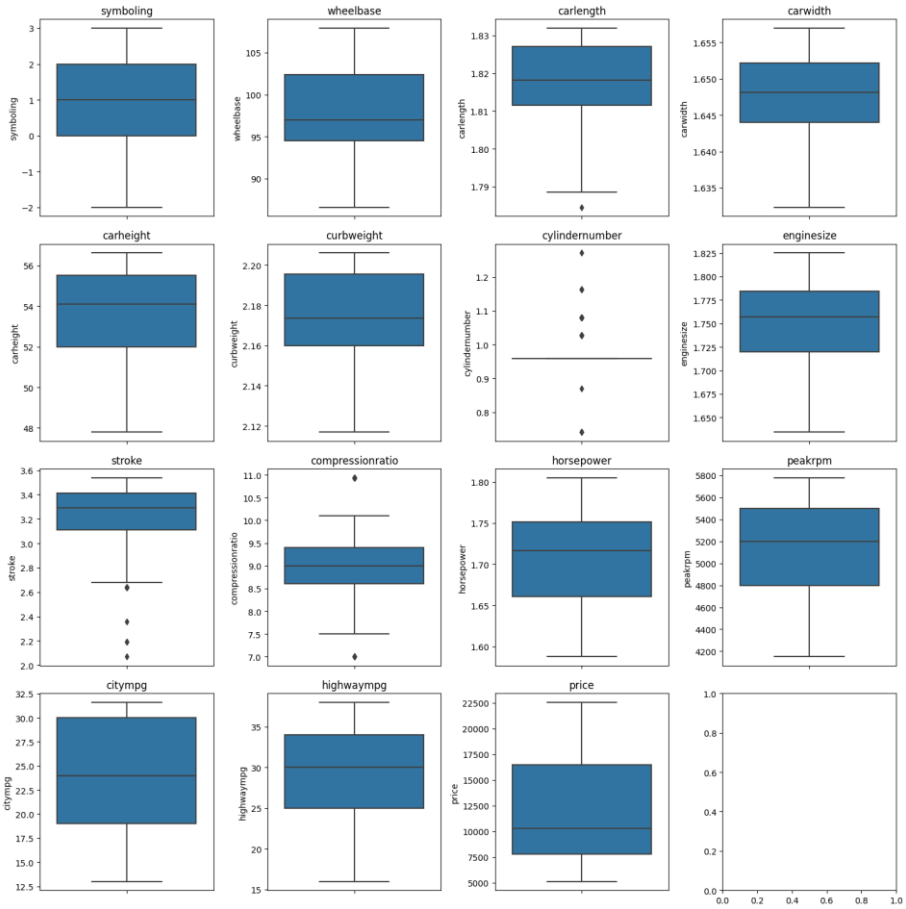


De aquí se puede interpretar lo siguiente de estas variables:

1. **symboling**: Distribución con múltiples picos, con la mayoría de los autos teniendo un riesgo de seguro neutral o ligeramente elevado.
2. **normalized-losses**: Distribución sesgada a la derecha con un pico notorio alrededor de 160.
3. **wheel-base**: La mayoría de los autos tienen una distancia entre ejes en el rango de 90 a 100.
4. **length**: Distribución bimodal, sugiriendo dos categorías distintas de autos en términos de longitud.
5. **width**: Distribución aproximadamente normal con un ligero sesgo a la derecha.
6. **height**: Distribución uniforme con un pico alrededor de 54.
7. **curb-weight**: Sesgo a la derecha, con la mayoría de los autos en el rango de 2000 a 3000 libras.
8. **engine-size**: Sesgo a la derecha, con la mayoría en el rango de 60 a 140.
9. **bore**: Distribución aproximadamente normal con un pico alrededor de 3.6.
10. **stroke**: Distribución bimodal con picos alrededor de 3.3 y 3.4.
11. **compression-ratio**: Mayoría en el rango de 8 a 10, con algunos autos teniendo una relación mucho más alta.
12. **horsepower**: Sesgo a la derecha, indicando la presencia de autos de alto rendimiento.

13. **peak-rpm**: Distribución aproximadamente normal centrada en 5500.
14. **city-mpg**: Sesgo a la izquierda, con la mayoría entre 19 y 31 mpg.
15. **highway-mpg**: Similar a city-mpg, sesgo a la izquierda con la mayoría entre 25 y 37 mpg.
16. **price**: Sesgo a la derecha, con la mayoría de los autos en el rango de $5000 a $18000, pero con algunos autos con precios más altos.

Luego se usan diagramas de caja para poder visualizar la mediana, los cuartiles y los datos atípicos:

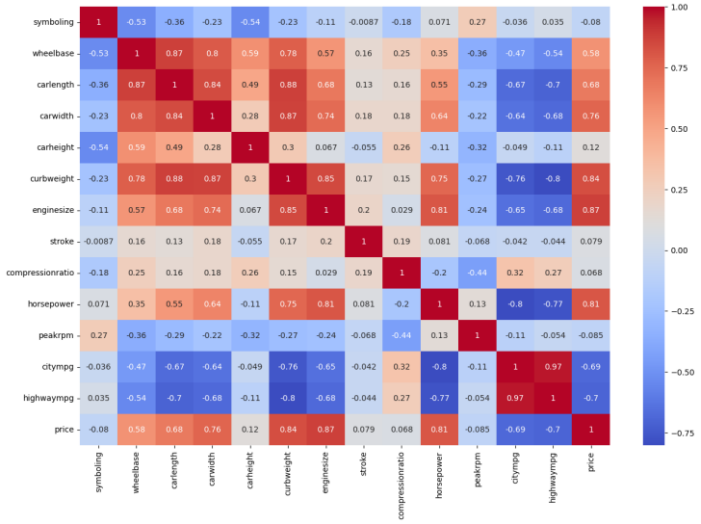


Con ello podemos observar lo siguiente de los diagramas de caja:

1. **symboling**: Mediana cerca de 1; valores atípicos en el extremo inferior.
2. **normalized-losses**: Mediana alrededor de 115; valores atípicos superiores.
3. **wheel-base**: Mediana en el rango de 95-100; autos más grandes como valores atípicos.
4. **length**: Mediana cerca de 170; autos más largos como valores atípicos superiores.
5. **width**: Mediana cerca de 66; autos más anchos como valores atípicos.
6. **height**: Distribución uniforme con mediana alrededor de 54.

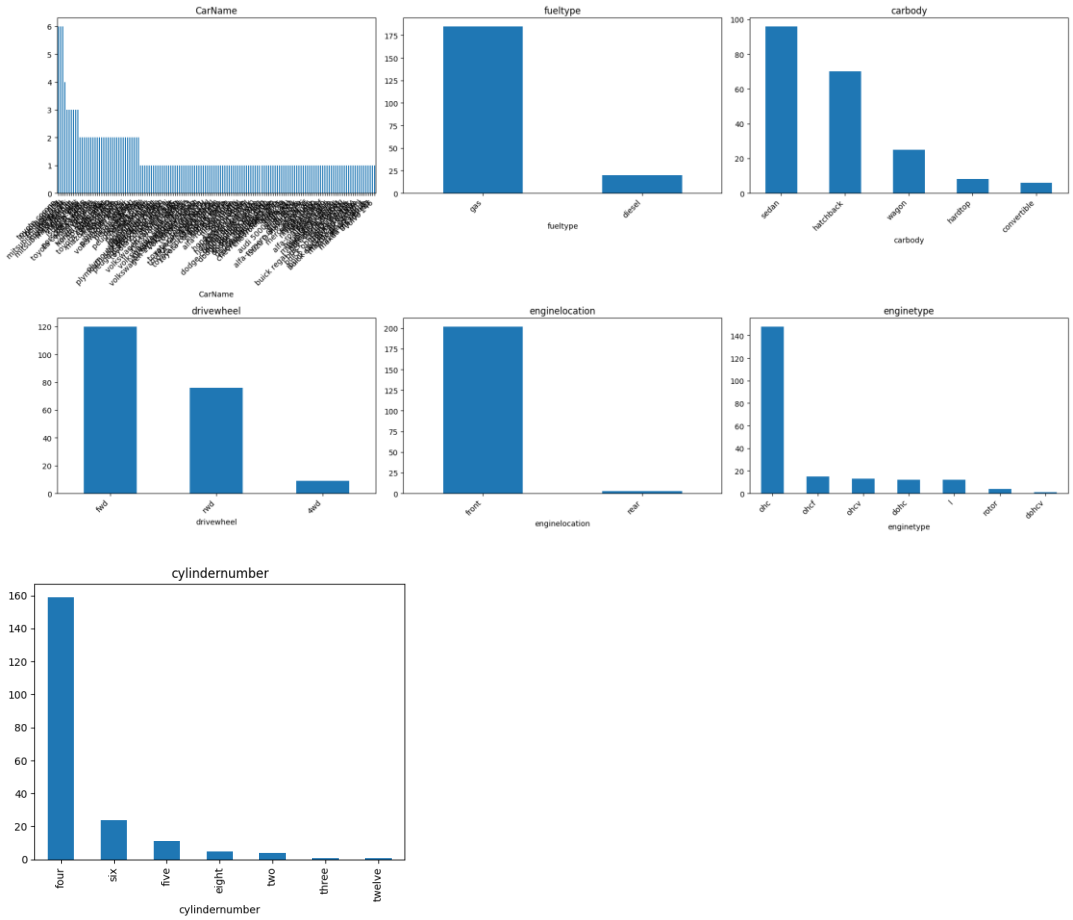
7. **curb-weight**: Mediana alrededor de 2400; autos más pesados como valores atípicos.
8. **engine-size**: Mediana cerca de 120; motores grandes como valores atípicos.
9. **bore**: Mediana alrededor de 3.6; valores atípicos en ambos extremos.
10. **stroke**: Mediana cerca de 3.3; valores atípicos en ambos extremos.
11. **compression-ratio**: Mediana cerca de 9; altas ratios como valores atípicos.
12. **horsepower**: Mediana cerca de 95; alta potencia como valores atípicos.
13. **peak-rpm**: Mediana cerca de 5200; valores atípicos en ambos extremos.
14. **city-mpg**: Mediana cerca de 24; alta eficiencia como valores atípicos.
15. **highway-mpg**: Mediana cerca de 28; alta eficiencia como valores atípicos.
16. **price**: Mediana cerca de $10,000; autos de mayor coste como valores atípicos superiores.

Una vez realizado ello continuamos para poder observar la matriz de correlación que es como un mapa de calor en el cual se puede observar las correlaciones que tiene cada variable numérica con el precio



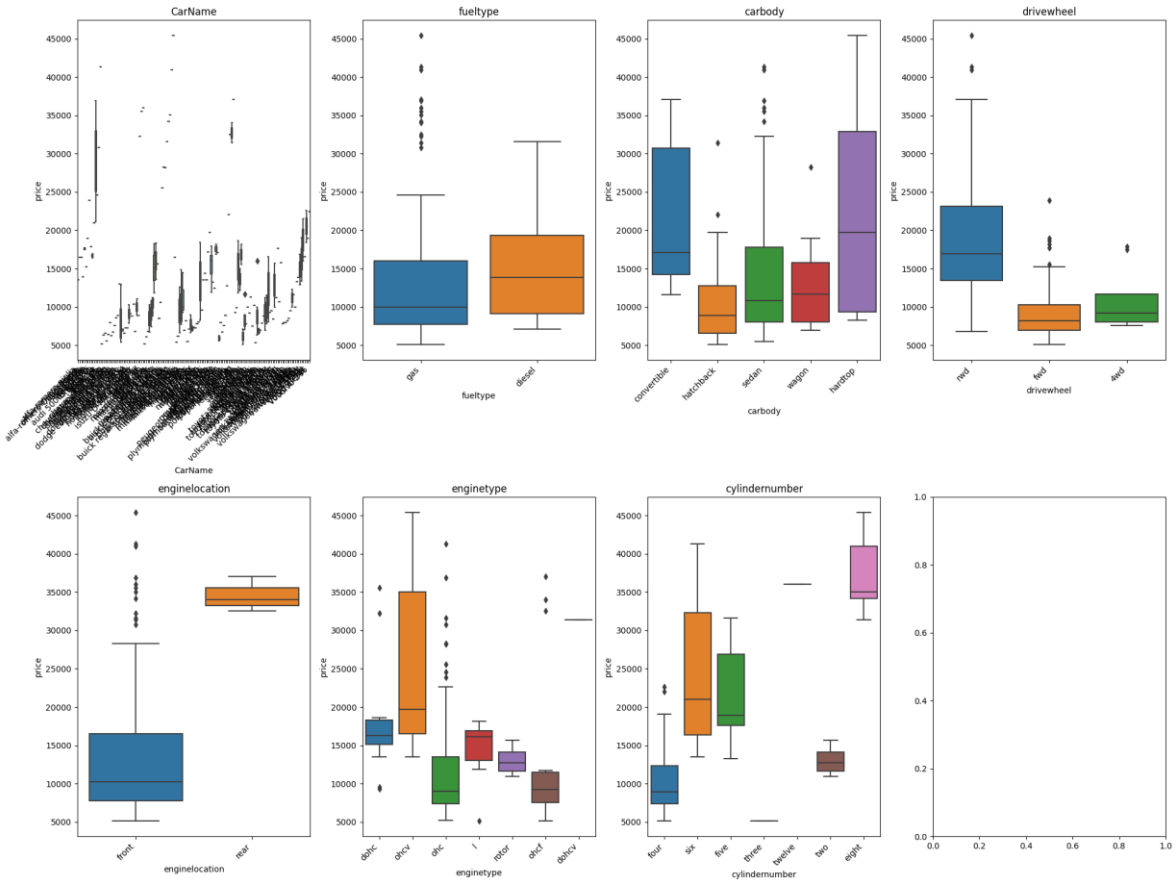
Esto nos ayudara a elegir nuestras 6 variables más adelante.

Para visualizar los datos de las variables categóricas primero se usó una gráfica de barras para poder observar cual es el resultado más repetido dentro de las variables y en cierta manera ver de mejor manera lo que contienen, dichos gráficos están en la siguiente pagina

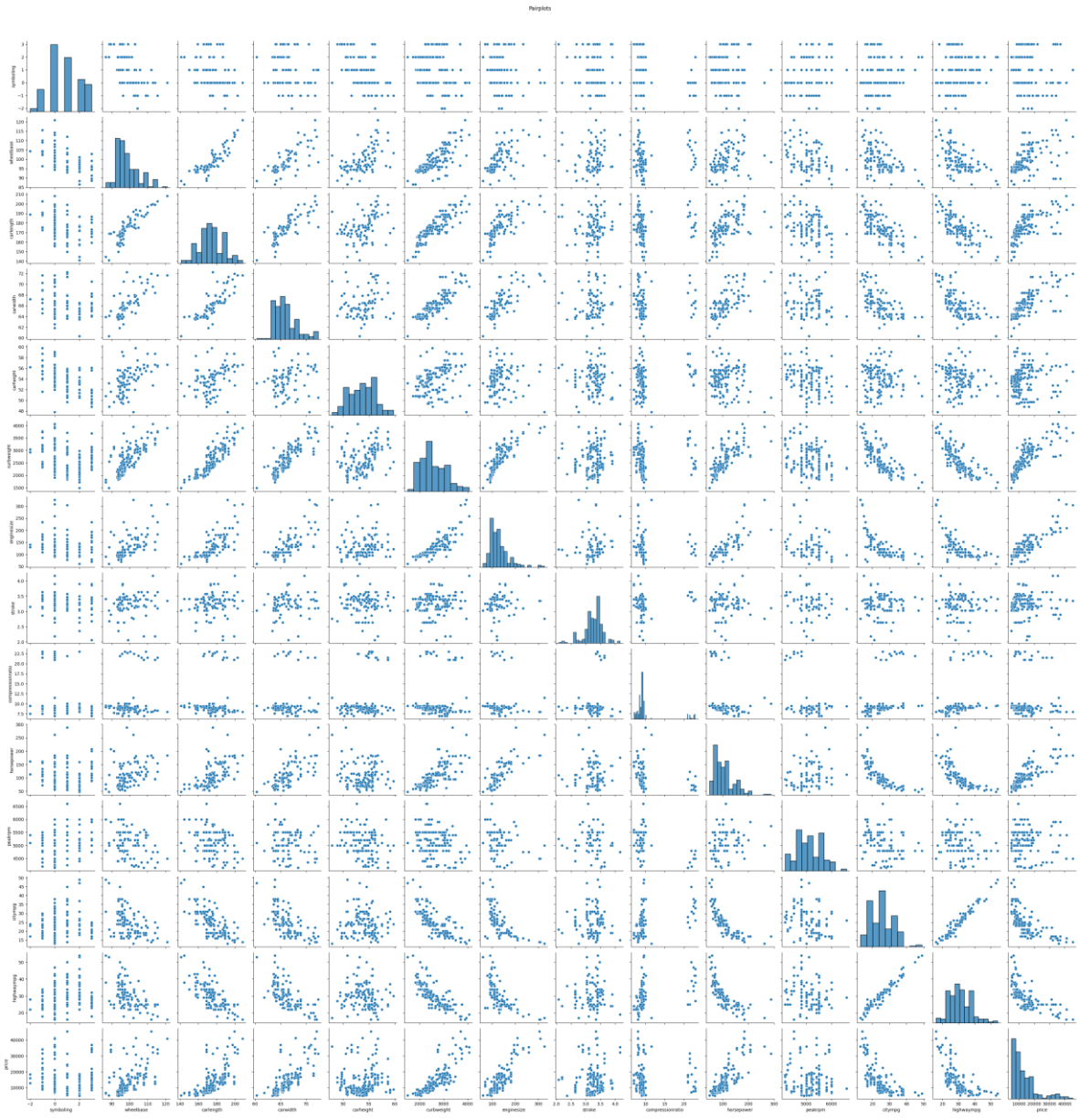


Parece que carname es una variable que contiene muchos resultados diferentes, esto hace que sea inviable poder usarlo o tratarlo como por ejemplo convertir a variables dummy, además de que el nombre a opinión no parece tener un impacto en el mundo real de cuanto cuesta un coche, si no más bien por las características que tenga un coche y otros factores como la compañía que fabrica dichos coches.

Una vez dicho ello procedemos a graficar los diagramas de caja de las variables cuantitativas para poder visualizar mediana, cuartiles y datos atípicos:



Al final hacemos varios diagramas de dispersión para pares de variables, esto con el objetivo de poder visualizar las relaciones entre variables:



Selección de variables, identificación y Tratamiento de Valores Atípicos

En base a lo que podemos observar, junto con una verificación breve de si había algún valor faltante que afortunadamente no hubo ninguno en ninguna de las variables se decidió elegir las siguientes variables:

1. **enginesize**: El tamaño del motor es una característica crucial que suele influir en el precio del automóvil. Un motor más grande suele asociarse con vehículos de mayor rendimiento y, por lo tanto, más caros.

2. **curbweight**: El peso del automóvil puede ser un indicativo de su robustez y, posiblemente, de la calidad de los materiales utilizados, lo que afecta el precio.

3. **horsepower**: La potencia del motor es una métrica de rendimiento. Los autos con más caballos de fuerza suelen ser más caros.

4. **carwidth**: La anchura del automóvil puede estar relacionada con características de lujo y espacio, lo que podría influir en el precio.

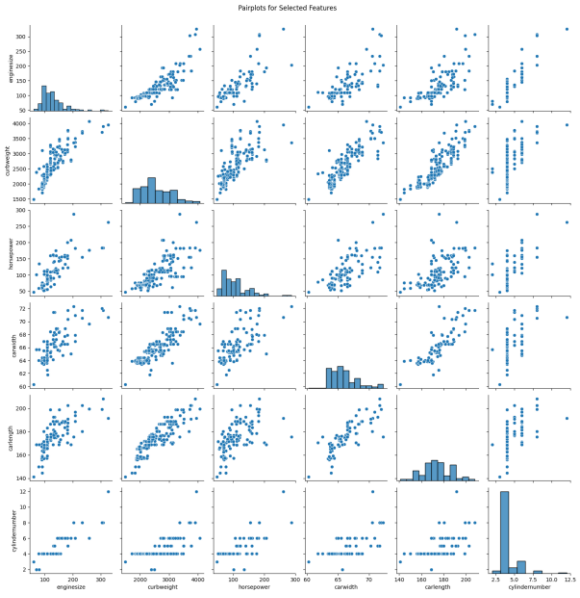
5. **carlength**: Similar a la anchura, la longitud del automóvil puede indicar un vehículo más grande y espacioso, lo que podría justificar un precio más alto.

6. **cylindernumber**: El número de cilindros es determinante para la potencia del vehículo lo que podría justificar un precio más alto.

Estas variables fueron seleccionadas principalmente por su alta correlación con el precio y su relevancia en el contexto de la industria automotriz.

La mayor parte de nuestras variables son numéricas a excepción de cylindernumber, en el cual al ver que sus datos categóricos son simplemente números escritos en ingles se decidió a convertir los datos directamente a números para trabajar de una forma más fácil con ello.

Una vez hecho ello hacemos diagramas de dispersión de las variables seleccionadas para poder observar la relación de estas variables:



En base a todo lo anterior y viendo como se comportan las variables seleccionadas se decidió realizar un tratamiento de los outliers o de los valores atípicos, de los cuales se decidió que todo valor que supere el límite superior (Que es el percentil 99) será cambiado al valor de este último, esto con el objetivo de poder eliminar valores extremos que puedan afectar al modelado y solución de este problema.

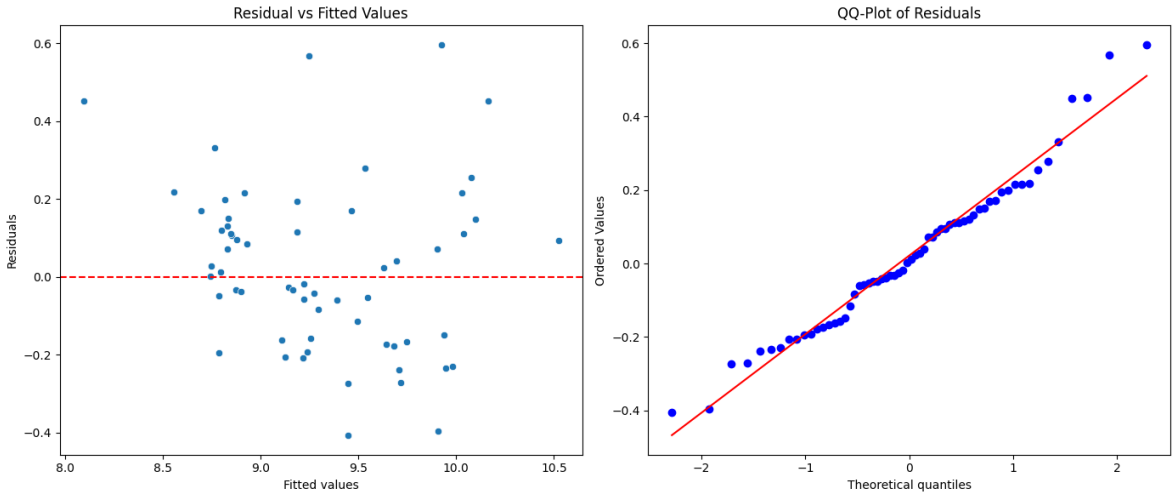
Una vez realizado ello y en base a las distribuciones de los datos, así como el supuesto de que los datos recolectados en la vida real no suelen seguir una distribución ideal se decidió aplicar una transformación logarítmica, esto con el propósito de que siga una distribución mas normal y que se pueda mejorar varias características de los datos como seria la variabilidad.

Modelado y verificación del modelo

En el modelado se decidió usar un modelo de regresión lineal múltiple en el cual se modela la relación entre una variable dependiente y múltiples variables independientes, esto nos resultara útil para predecir los precios en función de las 6 variables.

Primeramente, se procedió a dividirlos en conjuntos de entrenamiento y prueba. Este paso es crucial para validar el rendimiento del modelo en datos no vistos. Luego, se utilizó el conjunto de entrenamiento para "enseñar" al modelo la relación entre las características y el precio. El modelo resultante logró un $R2$ de 0.8266, una métrica que indica que el modelo pudo explicar el 82.66% de la variabilidad en el precio. Es un resultado bastante positivo, pero como cualquier modelo, tiene margen de mejora.

Para validar sen usaron dos gráficas, una de residuos contra los valores ajustados y una QQ-plot, en el cual se observa lo siguiente



- **Gráfico de Residuos vs Valores Ajustados**: Parece que los residuos se distribuyen aleatoriamente alrededor de la línea horizontal, lo que es una buena señal. Sin embargo, se puede observar cierta estructura en los residuos, lo que podría sugerir que el modelo podría mejorarse.
- **QQ-Plot:** La mayoría de los puntos se alinean cerca de la línea diagonal, lo que sugiere que los residuos tienen una distribución aproximadamente normal. Sin embargo, hay algunos puntos que se desvían de la línea, especialmente en los extremos, lo que indica la presencia de algunos residuos atípicos.

Conclusión:

En general se concluye que las variables elegidas para esta empresa ayudaran a poder predecir el precio de sus automóviles con una precisión del 82.66%, el porcentaje restante se deberá a otros factores que dicho modelo no puede contemplar, no obstante esto puede ayudar de gran manera a

que se posicione dentro del mercado estadounidense, a planificar las características de sus coches y en general a tomar mejores decisiones respecto al costo-beneficio de sus producto, dando el terreno y la información para que poco a poco la empresa tome mejores decisiones respecto a su posicionamiento, así como la planeación de futuros vehículos

Referencias bibliográficas:

López, M. (2023). La competitividad empresarial en base a los precios y costos. Recuperado de https://1library.co/article/competitividad-empresa-precios-costos.yng5mg0z

Anexos:

Este análisis, incluyendo el Jupyter Notebook, así como el conjunto de datos original, están disponibles en el siguiente repositorio bajo la carpeta Final/Evidencia2_Modulo1

El diccionario del conjunto de datos esta disponible en el mismo repositorio bajo la carpeta Final/Evidencia1_Modulo1

5100-chap/TC3006C_Portafolio: Portafolio de los avances del reto de la materia TC3006C (github.com)