Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Monterrey

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

Reporte final de “El precio de los autos”

Marcelo Márquez A01720588

11 sept 2023

Resumen

El problema radica en que una empresa China quiere entrar al mercado estadounidense y debe determinar si los autos que ofrece son rentables dentro de dicho mercado. Para esto, se tiene una base de datos con las características y precios de una variedad de automóviles que se utilizarán como aproximación del precio de mercado de vehículos en Estados Unidos.

Para abordar la problemática, primero que nada se realizó el proceso de estadística descriptiva de los datos, tomando en cuenta las características generales de los mismos y obteniendo un conocimiento *a priori* de la media, la desviación estándar, la mediana y el rango de datos que se está manejando para cada variable. Se determina a partir de una visualización y examinación de los datos que se utilizará una Regresión Lineal Múltiple para la modelación y se realiza una verificación del modelo con base a los supuestos de Mínimos Cuadrados Ordinarios

Se obtuvo como resultado un modelo con una R^2 ajustada de 0.85 con el uso de las variables enginesize, horseporwer, highwaympg, cylindernumber, carwidth y curveweight para determinar el precio. De estas, únicamente highwaympg resultó no ser significante para el modelo.

Introducción

Una empresa automovilística china aspira entrar al mercado estadounidense, quieren agregar su propia fábrica para competir contra las compañías estadounidenses y europeas. Contrataron una empresa de consultoría de automóviles con el fin de identificar cuáles son los factores principales que definen el precio final de un automóvil.

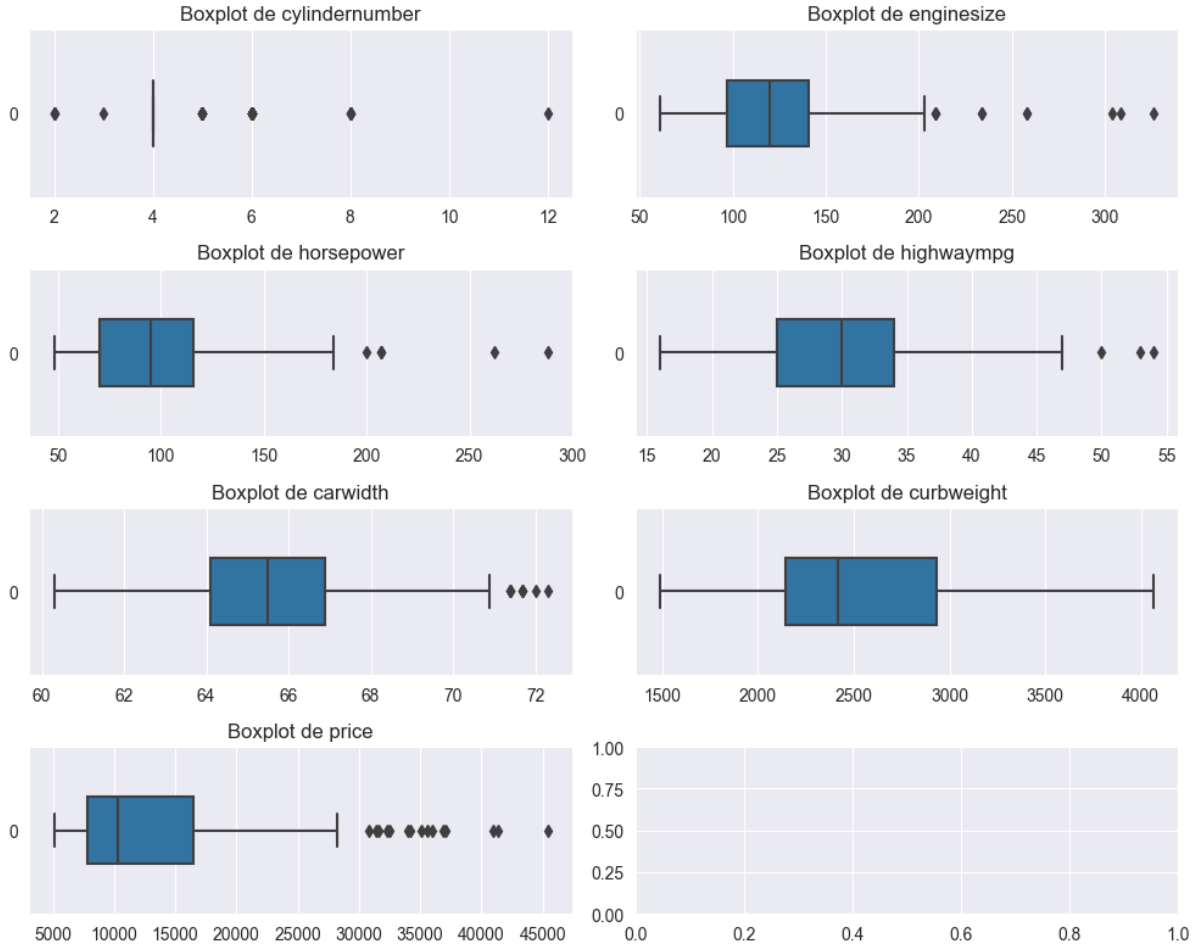
El problema es esencialmente un paso importante en la toma de decisiones de cualquier corporativo ya que entrar a un mercado implica muchos riesgos para la empresa los cuales se pueden reflejar en pérdidas de millones de dólares. Por lo tanto, el análisis presenta una propuesta importante para las empresas a la hora de decidir si expandirse a un nuevo país o, por el contrario, mantenerse en los mercado en los que ya tiene presencia.

Análisis de Resultados

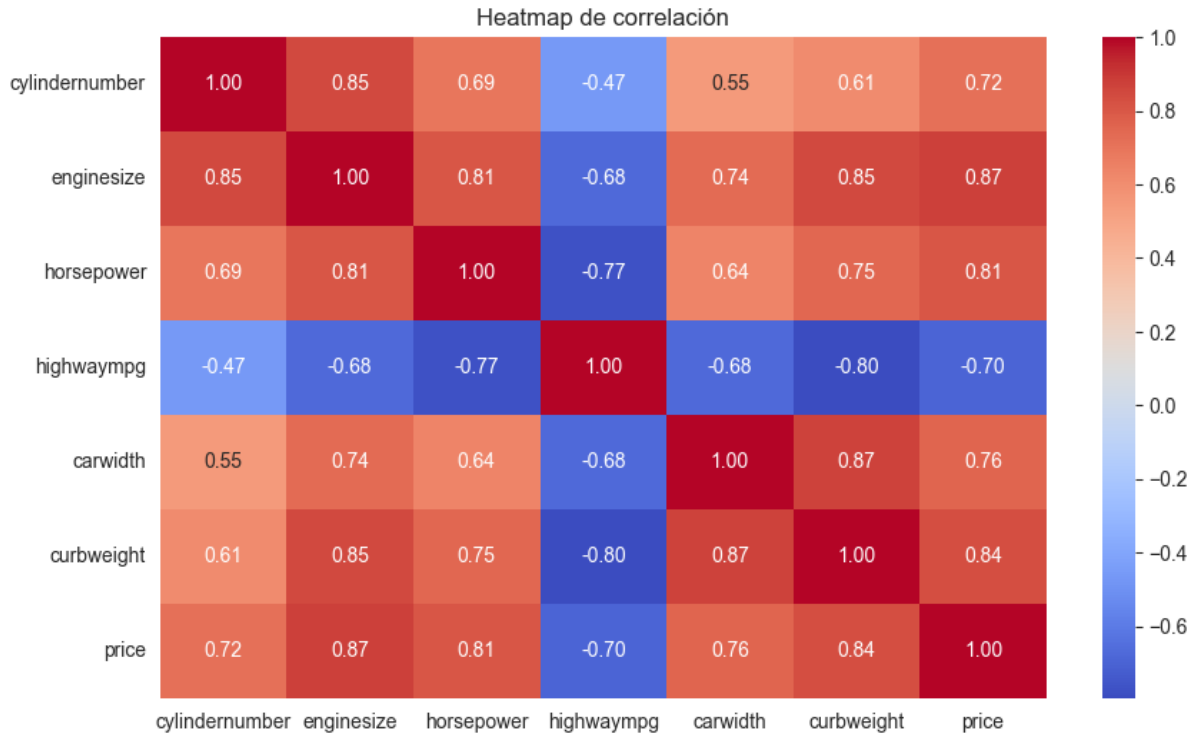
Para poder encontrar esos valores con más significancia hacia el precio final de un automóvil en el mercado estadounidense, se implementaron análisis utilizando modelos de regresión lineal simples y pruebas de hipótesis.

Los automóviles son bienes de lujo que tienen altos costos de producción pero que, a la vez, pueden tener un gran margen de ganancia. Por lo tanto, al haber encontrado un modelo que pueda

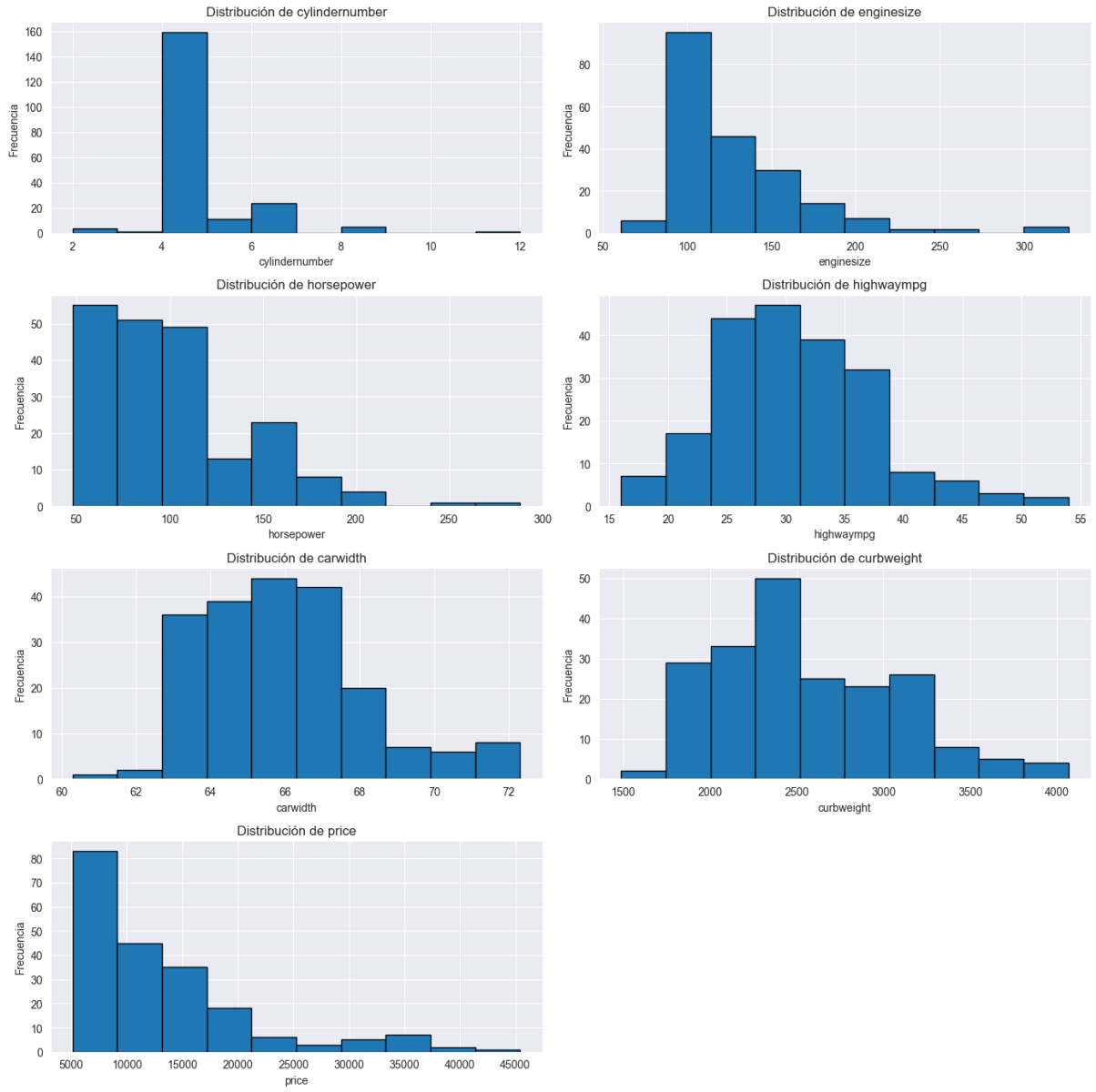
Gráficos de caja y bigote



Análisis de Correlación



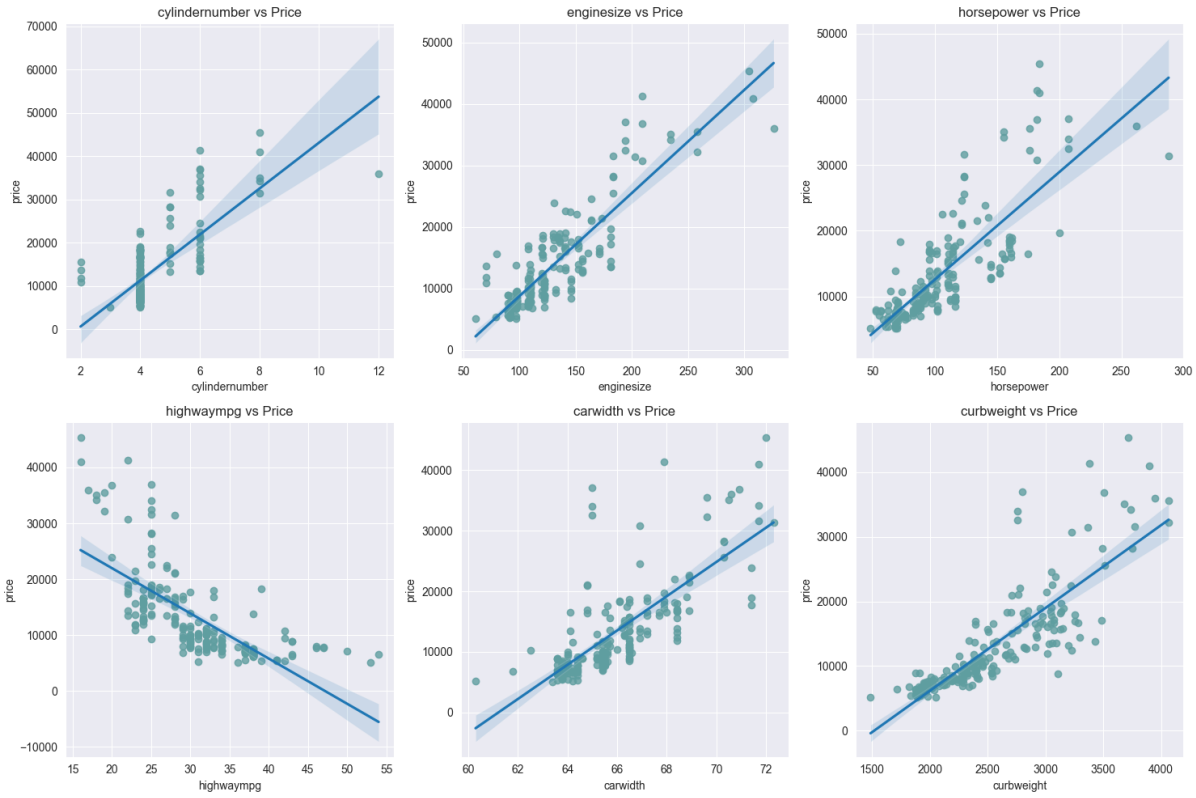
Distribución de los datos



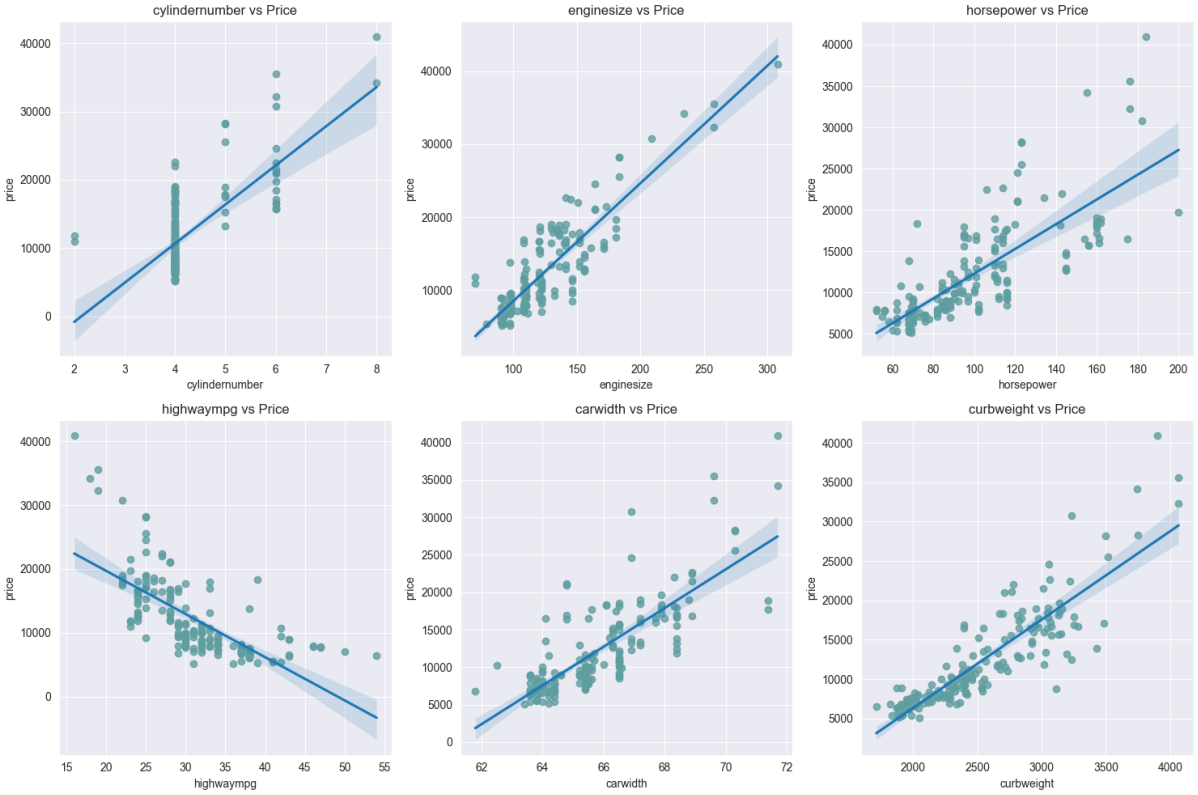
Herramienta 1 - Regresión Lineal Simple

Con base al comportamiento de los datos, se decidió utilizar una regresión lineal múltiple observando que las variables independientes tienen una relación lineal respecto la variable dependiente. No obstante, se tuvieron que realizar pruebas para verificar que se estuvieran cumpliendo los supuestos de Mínimos Cuadrados Ordinarios (OLS por sus siglas en inglés) y realizar ajustes a los datos para poder verdaderamente llevar a cabo el análisis con esta herramienta estadística.

Con datos influyentes



Sin datos influyentes:



La tabla observada a continuación nos presenta un resumen del desempeño del modelo donde se puede verificar el cumplimiento de los supuestos de Mínimos Cuadrados Ordinarias respectivamente. De acuerdo a los resultados obtenidos se determina a un 95% de confianza que el modelo puede explicar el 85.1% de los datos

de la variable dependiente y que todas las variables son significativas a excepción de la variable "highwaympg".

Dado lo anterior se consigue que las variables enginesize, carwidth, curbweight, cylindernumber y horsepower son buenos estimadores en conjunto para que la empresa China realice una predicción del vehículo en cuestión y determine si puede entrar al mercado estadounidense y determinar su rentabilidad por consecuencia.

La R^2 indica que las variables describen adecuadamente el comportamiento del precio y por lo tanto la empresa podrá tomar decisiones a partir de la ecuación que emite el mismo modelo.

Herramienta 2 - Pruebas de hipótesis:

Las pruebas de hipótesis permiten determinar qué variables deben permanecer en el modelo y, por el contrario, cuales no están aportando nada al mismo. Al establecer el valor de significancia (alpha=0.05) podemos obtener un nivel de confianza sobre los resultados que se están presentando. Al utilizar esta herramienta se pudieron hacer ajustes sobre las variables seleccionadas en el modelo y es por eso que se consideró pertinente su uso.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.820
Model:                            OLS   Adj. R-squared:                  0.815
Method:                 Least Squares   F-statistic:                     150.4
Date:                Tue, 12 Sep 2023   Prob (F-statistic):           6.11e-71
Time:                        22:47:52   Log-Likelihood:                -1956.7
No. Observations:                 205   AIC:                             3927.
Df Residuals:                     198   BIC:                             3951.
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -4.787e+04   1.33e+04     -3.605      0.000   -7.41e+04   -2.17e+04
cylindernumber   113.1838    462.886      0.245      0.807    -799.634    1026.002
enginesize        79.5642     18.498      4.301      0.000      43.085     116.043
horsepower        52.2147     12.383      4.217      0.000      27.795      76.634
highwaympg        32.2545     67.455      0.478      0.633    -100.768     165.277
carwidth         568.9752    226.241      2.515      0.013     122.824    1015.126
curbweight         2.5909      1.422      1.822      0.070      -0.213       5.395
==============================================================================
Omnibus:                       27.523   Durbin-Watson:                   0.773
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               57.555
Skew:                           0.639   Prob(JB):                     3.18e-13
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.851
Model:                            OLS   Adj. R-squared:                  0.846
Method:                 Least Squares   F-statistic:                     172.7
Date:                Tue, 12 Sep 2023   Prob (F-statistic):           3.79e-72
Time:                        22:47:52   Log-Likelihood:                -1728.2
No. Observations:                 188   AIC:                             3470.
Df Residuals:                     181   BIC:                             3493.
Df Model:                           6
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const         -4.54e+04   1.09e+04     -4.167      0.000   -6.69e+04   -2.39e+04
cylindernumber 1414.7844    396.847      3.565      0.000     631.743    2197.826
enginesize       39.5202     14.980      2.638      0.009       9.963      69.078
horsepower       21.5190     11.691      1.841      0.067      -1.549      44.587
highwaympg       -3.9825     54.549     -0.073      0.942    -111.616     103.651
carwidth        492.5788    182.428      2.700      0.008     132.620     852.538
curbweight        4.8866      1.070      4.566      0.000       2.775       6.998
==============================================================================
Omnibus:                        8.038   Durbin-Watson:                   0.957
Prob(Omnibus):                  0.018   Jarque-Bera (JB):                8.076
Skew:                           0.418   Prob(JB):                       0.0176
...
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.58e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Análisis de los Coeficientes

Modelo con Todos los Datos

- **cylindernumber**: El coeficiente es 113.1838, pero tiene un p-valor muy alto (0.807), lo que indica que no es estadísticamente significativo al nivel de confianza del 95%.
- **enginesize**: Es significativo con un coeficiente de 79.5642 y un p-valor muy bajo.
- **horsepower**: También es significativo con un coeficiente de 52.2147.
- **highwaympg**: No es significativo con un p-valor de 0.633.
- **carwidth**: Es significativo con un coeficiente de 568.9752.
- **curbweight**: Tiene un p-valor de 0.070, lo que indica que está en el límite de la significancia.

Modelo sin Datos Influyentes

- **cylindernumber**: Se volvió significativo con un coeficiente más alto de 1414.7844.
- **enginesize**: Aunque el coeficiente disminuyó a 39.5202, sigue siendo significativo.
- **horsepower**: El coeficiente disminuyó a 21.5190 pero sigue siendo significativo.
- **highwaympg**: Continúa siendo no significativo con un p-valor muy alto.

- **carwidth**: Aunque el coeficiente disminuyó ligeramente, sigue siendo significativo.
- **curbweight**: Aumentó su significancia con un coeficiente de 4.8866.

Análisis de las Estadísticas de los Residuos y Métricas de Ajuste del Modelo

Ambos Modelos

- **R-squared**: El modelo sin datos influyentes tiene un R-squared más alto (0.851) en comparación con el modelo con todos los datos (0.820), lo que indica que explica una mayor proporción de la variabilidad en el precio.
- **Adjusted R-squared**: Similar al R-squared, es más alto en el modelo sin datos influyentes, lo que indica un mejor ajuste del modelo.
- **F-statistic**: Es más alto en el modelo sin datos influyentes, lo que indica que el modelo es más significativo estadísticamente.
- **Prob (F-statistic)**: Muy bajo en ambos modelos, lo que indica que los modelos son significativos.
- **Log-Likelihood, AIC, y BIC**: Estos indicadores muestran que el modelo sin datos influyentes es preferible ya que tiene un Log-Likelihood más alto y valores más bajos de AIC y BIC.

Conclusiones

- **Multicolinealidad**: En ambos modelos, el número de condición es bastante alto, lo que indica problemas potenciales de multicolinealidad. Esto sugiere que algunas de las variables independientes están correlacionadas, lo que puede afectar la precisión de los coeficientes estimados.
- **Normalidad de los Residuos**: Las pruebas Omnibus y Jarque-Bera indican que los residuos no están normalmente distribuidos, especialmente en el modelo con todos los datos.
- **Autocorrelación**: La estadística Durbin-Watson está más cerca de 2 en el modelo sin datos influyentes, lo que indica menos autocorrelación en comparación con el modelo con todos los datos.
- Para mejorar el ajuste del modelo con todos los datos, se podría borrar las variables no significativas (columnas de 'cylindernumber' y 'highwaympg' del DataFrame).
- Para mejorar el ajuste del modelo sin datos influyentes, se podría borrar la variable no significativa (columna de 'highwaympg' del DataFrame).

Las variables utilizadas en el modelo son de utilidad para que la empresa tome una decisión acerca de entrar al mercado estadounidense. Al poder realizar una predicción del precio de un automóvil, la empresa China podrá estimar el márgen que obtendrá por automóvil vendido.

Los métodos utilizados para realizar el análisis fueron seleccionados con base a las características de los datos así como la correlación que existe tanto entre las variables independientes como con respecto a la variable dependiente.

Link al portafolio de análisis:
https://github.com/MarceloM123/TC3006C.101_A01720588_Portafolio_Analisis/tree/main/Final/Modulo1-Estadistica

Link al portafolio de implementación:
https://github.com/MarceloM123/TC3006C.101_A01720588_Portafolio_Implentacion/tree/main/Final/Modulo1-Estadistica