Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Monterrey

"Yo, como integrante de la comunidad estudiantil del Tecnológico de Monterrey, soy consciente de que la trampa y el engaño afectan mi dignidad como persona, mi aprendizaje y mi formación, por ello me comprometo a actuar honestamente, respetar y dar crédito al valor y esfuerzo con el que se elaboran las ideas propias, las de los compañeros y de los autores, así como asumir mi responsabilidad en la construcción de un ambiente de aprendizaje justo y confiable"

"Inteligencia artificial avanzada para la ciencia de datos I"

Técnicas de procesamiento de datos para el análisis estadístico y para la construcción de modelos

Alumna:

Frida Cano Falcón A01752953

Profesores:

Ivan Mauricio Amaya Contreras

Antonio Carlos Bento

Frumencio Olivas Alvarez

Blanca Rosa Ruiz Hernandez

Hugo Terashima Marín

Fecha de entrega: 11 de septiembre de 2023

Resumen

En este proyecto se busca obtener un modelo de predicción que nos permita determinar las principales variables que determinan el precio de un automóvil para una empresa china que busca cotizar sus automóviles en Estados Unidos. Para ello se preparó y analizó un dataset que contiene características de automóviles que se encuentran en la industria estadounidense y europea. Todo esto haciendo un análisis estadístico impulsado con las herramientas de las librerías de cálculo estadístico de Python (Pandas, Seaborn, Numpy, Matplotlib, Statsmodels y Scikit Learn).

Introducción

En un mundo cada vez más globalizado, las empresas buscan expandirse a nuevos mercados para alcanzar un crecimiento sostenible. En este contexto, una empresa automovilística china ha fijado su mirada en el mercado estadounidense como un destino estratégico para su expansión. Su objetivo es establecer una unidad de fabricación en los Estados Unidos y competir con éxito en un mercado altamente competitivo, donde las preferencias de los consumidores y los factores que determinan el precio de los automóviles pueden ser muy distintos a los del mercado chino.

Con la determinación de tomar decisiones informadas y estratégicas, esta empresa ha contratado los servicios de una destacada firma de consultoría en la industria automotriz. El encargo de la consultoría es identificar los factores clave que influyen en el precio de los automóviles en el mercado estadounidense. En esencia, se busca responder a dos preguntas fundamentales:

1. ¿Cuáles son las variables más significativas para predecir el precio de un automóvil en el mercado estadounidense?
2. ¿En qué medida estas variables describen de manera precisa y efectiva el precio de un automóvil en dicho mercado?

Para llevar a cabo esta tarea, la consultora ha recopilado un extenso conjunto de datos, que incluye información detallada sobre diferentes tipos de automóviles presentes en el mercado estadounidense. Este conjunto de datos se encuentra documentado en el "Diccionario de Términos" proporcionado. La información recopilada permitirá a la consultora realizar un análisis profundo y riguroso con el objetivo de proporcionar a la empresa automovilística china una base sólida para la toma de decisiones estratégicas.

Este informe se estructura en tres etapas bien definidas: la primera se enfoca en la exploración y preparación de los datos, la segunda aborda la modelación y verificación del modelo, y finalmente, la tercera etapa culmina con la presentación de la versión final de ambas partes del problema: el análisis y la implementación de soluciones. Cada etapa representa un paso crucial hacia la consecución del objetivo final de este proyecto, que es

permitir a la empresa automovilística china ingresar con éxito en el mercado estadounidense y competir de manera efectiva con sus contrapartes estadounidenses y europeas.

Exploración y preparación de la base de datos

Durante esta fase se vio el panorama de los datos y se hizo la respectiva limpieza para trabajar con los valores contextualmente lógicos en el problema. Cabe mencionar que el siguiente análisis fue realizado con ayuda de las librerías de análisis estadístico de Python (Pandas, Seaborn, Numpy, Matplotlib, Statsmodels y Scikit Learn) y su implementación está desarrollada en el archivo anexo ***ProcesamientoDeDatos.ipynb.***

El primer paso fue identificar si la base de datos contenía variables categóricas y numéricas, y si las variables categóricas contenían errores en sus clasificaciones. Se encontró que la variable que describe los nombres de los autos contenía errores de sintaxis en algunos nombres, se identificaron los errores y se reemplazaron por los nombres correctos (ejemplo: *mazda* se cambió por *mazda*).

En seguida se dividieron en dos datasets diferentes las variables categóricas de las numéricas para analizar adecuadamente ambos casos (***df_cuantitativo*** y ***df_cualitativo***).

Para las variables cuantitativas se calcularon las métricas pertinentes (media, máximos, mínimos, cuartiles, etc). Para identificar si las métricas obtenidas eran congruentes se plotearon gráficas de distribución (histogramas) y de proporción (boxplots).

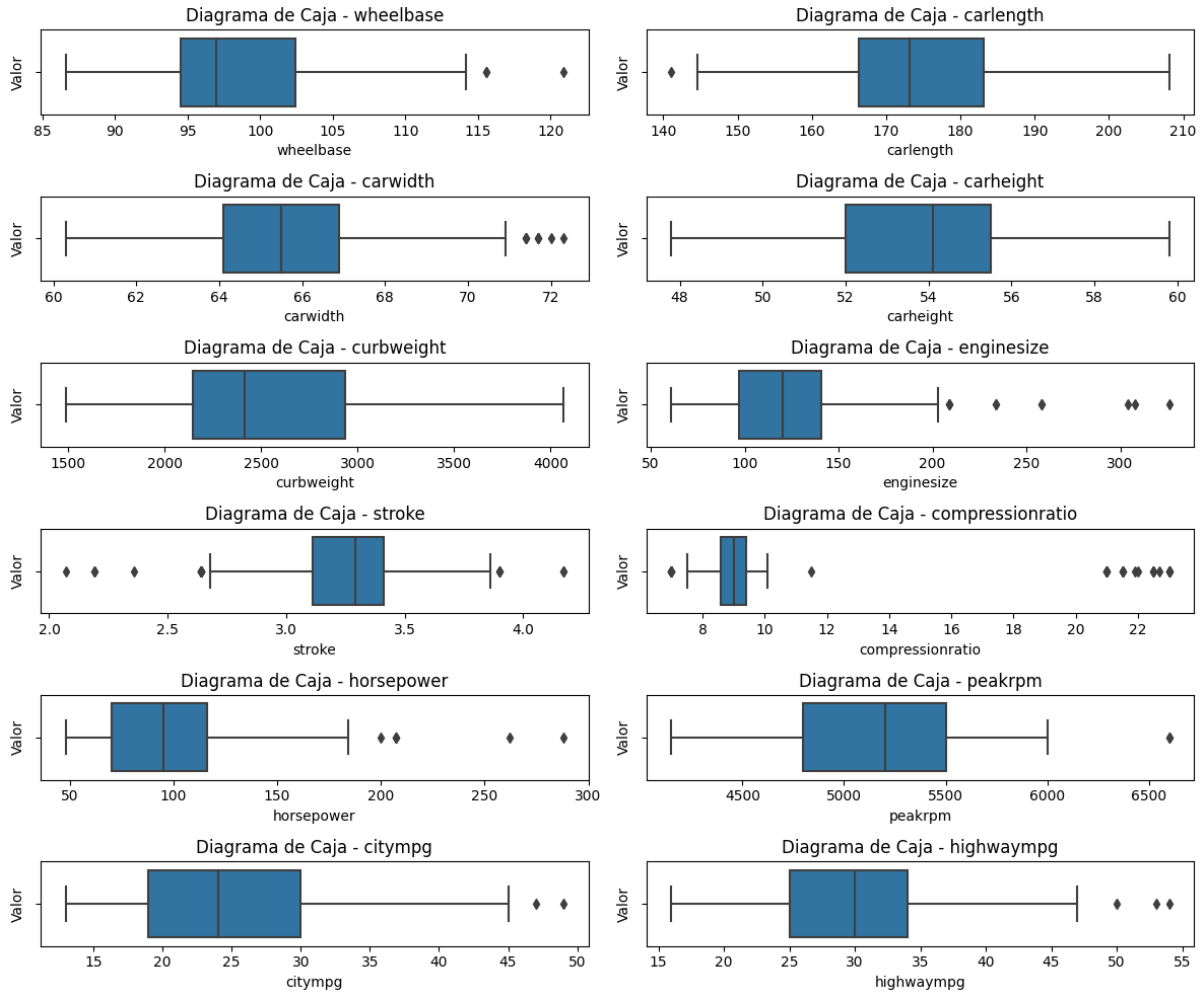


Fig 1. Boxplots de variables cuantitativas

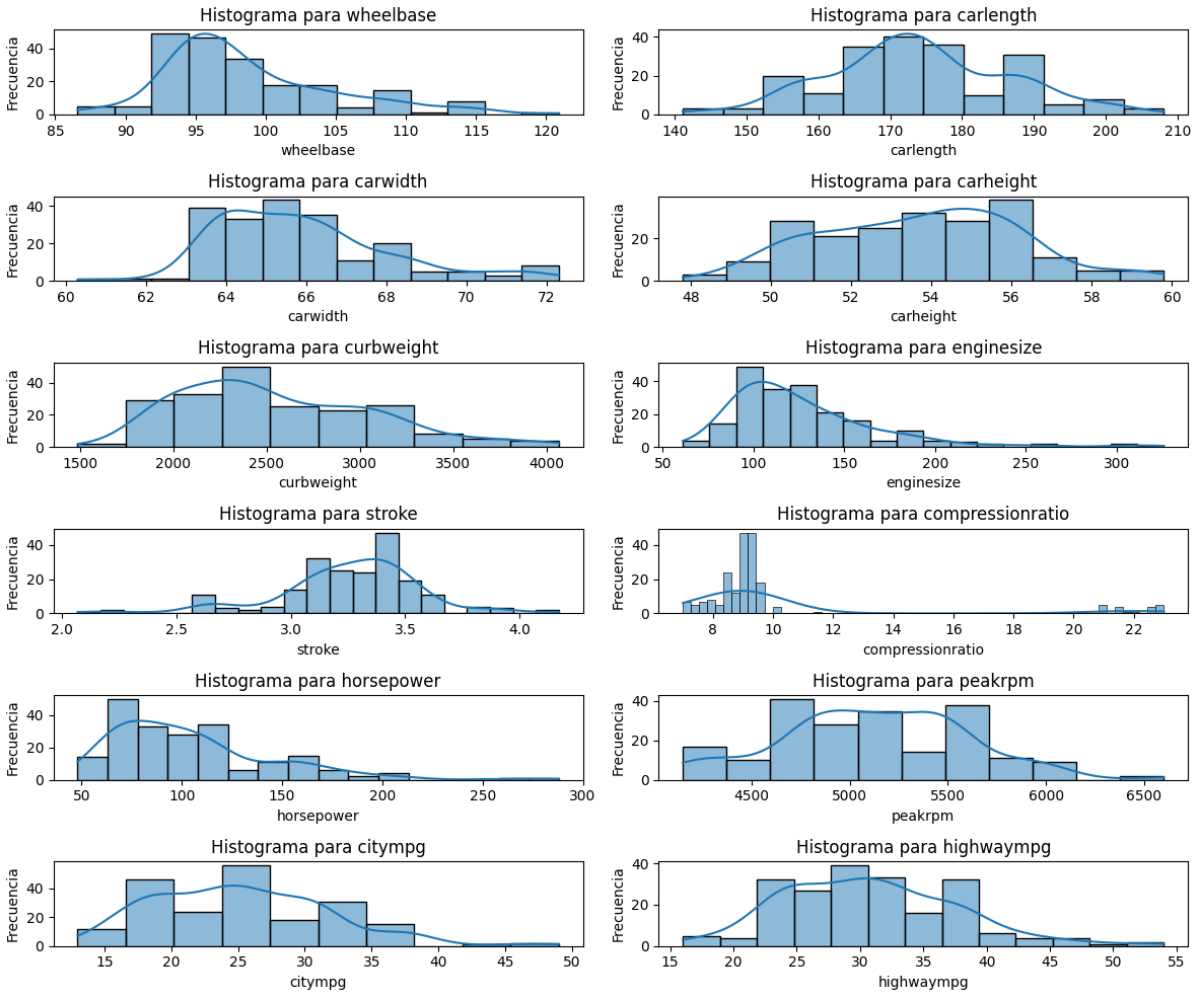


Fig 2. Distribución de variables cuantitativas

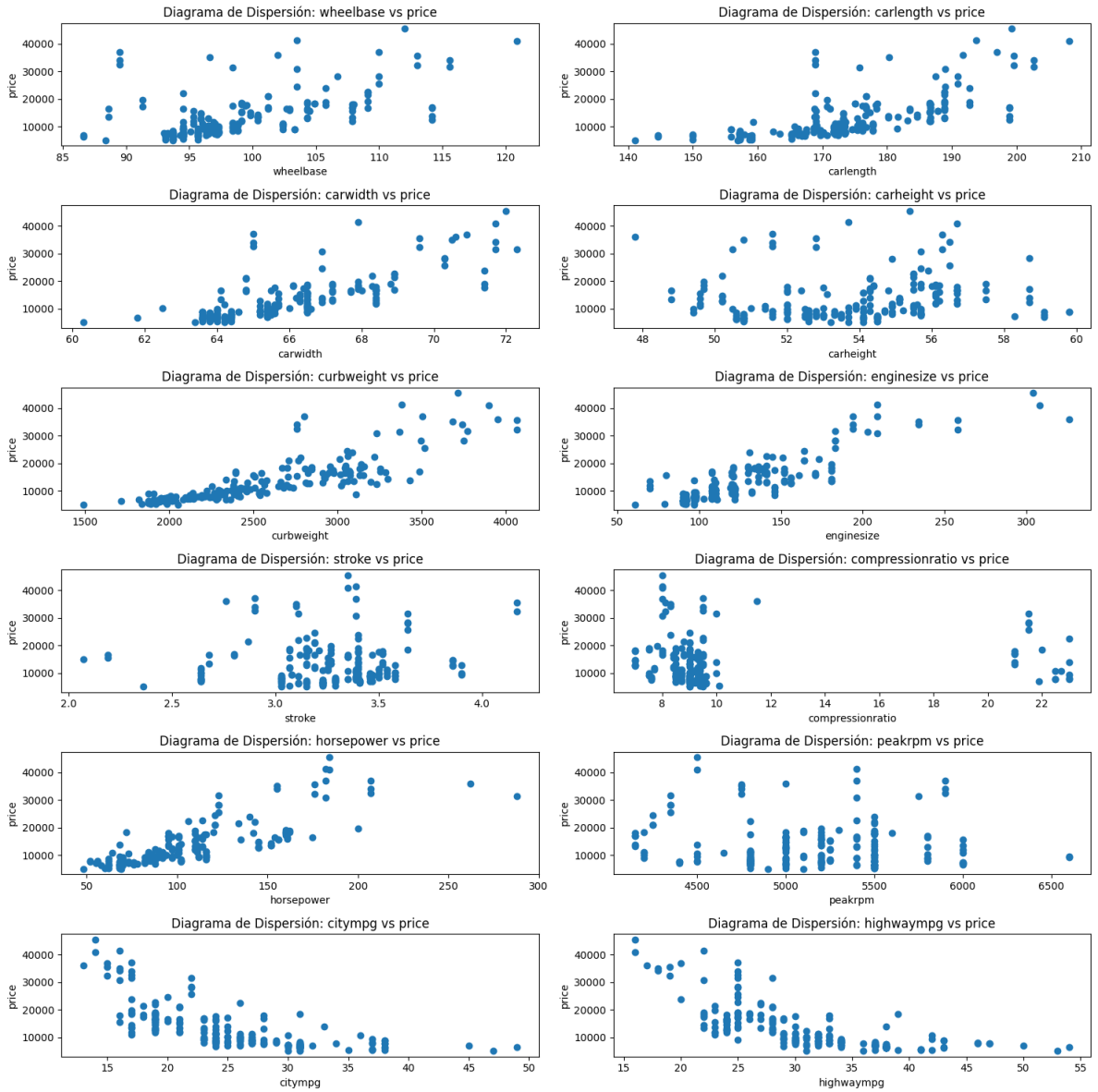


Fig 3. Dispersión de variables cuantitativas contra el precio

Esto nos brindó un panorama de datos que pueden generar ruido en nuestro modelo, así como identificar las variables que pueden tener un impacto en el precio final del carro. De primer ojo se identifica que existen valores que tienen precios similares por una misma característica (posibles variables dependientes del precio), o en donde la variación de una misma característica obtiene valores diferentes de precios (posibles variables independientes del precio). Así mismo, se ve que existe sesgo en alguna de las variables, tales como ***enginesize, compressionratio, highwaympg.***

Para acercarnos a la comprobación de la hipótesis sobre variables que tuvieran relación con el precio, realizamos una matriz de correlación.

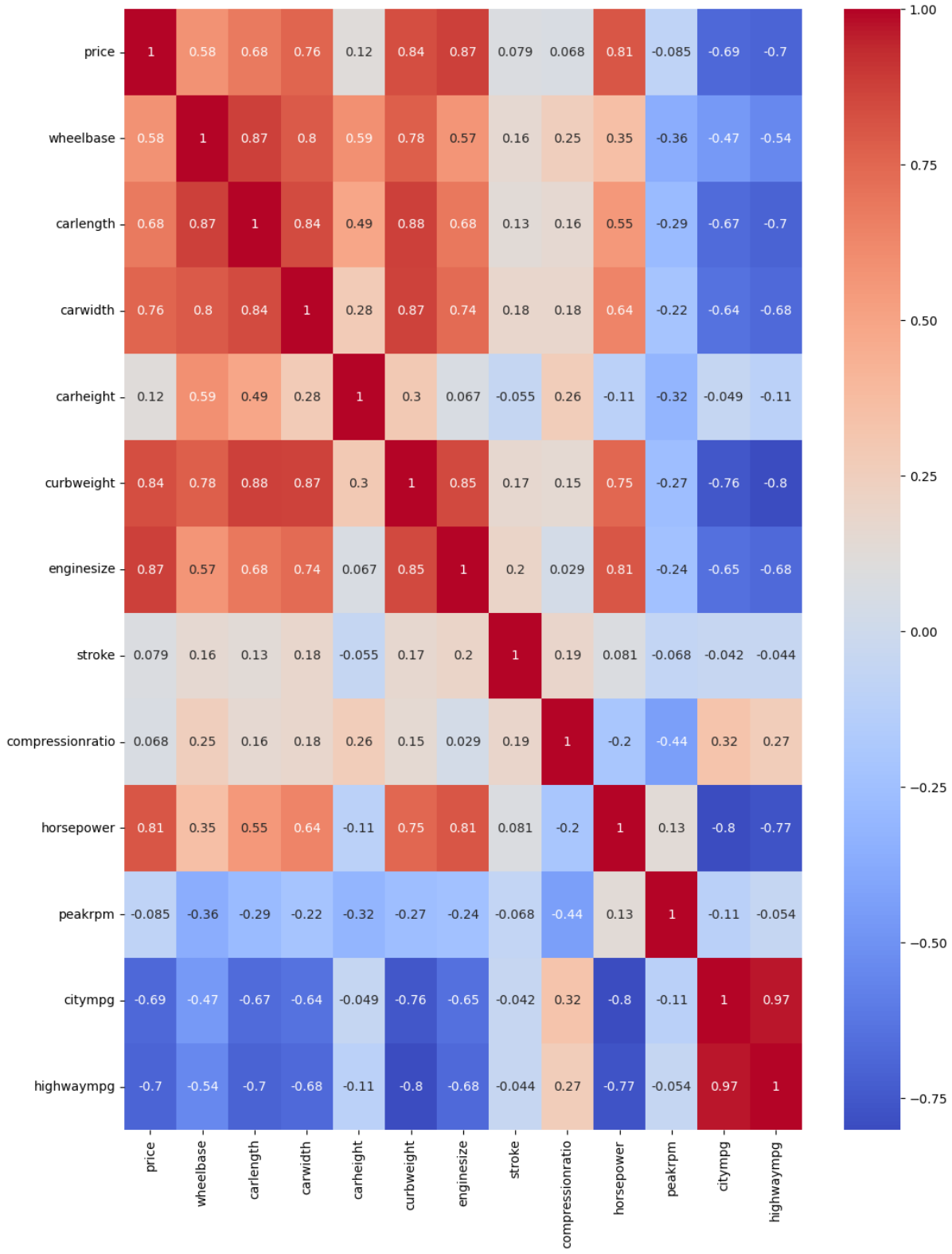


Fig 4. Correlación de las variables con respecto al precio

De acuerdo con la matriz, las variables que tienen una alta relación ($coef \geq 0.8$) con el precio son: ***curbwaight, enginesize, horsepower***.

Variable	Correlación
curbweight	0.84
enginesize	0.87
horsepower	0.81

En seguida se buscó analizar las variables categóricas de la base de datos, se desplegaron gráficas de barras que nos ayudaron a visualizar la frecuencia de cada clase de cada característica. después se compararon dichas características contra el precio.

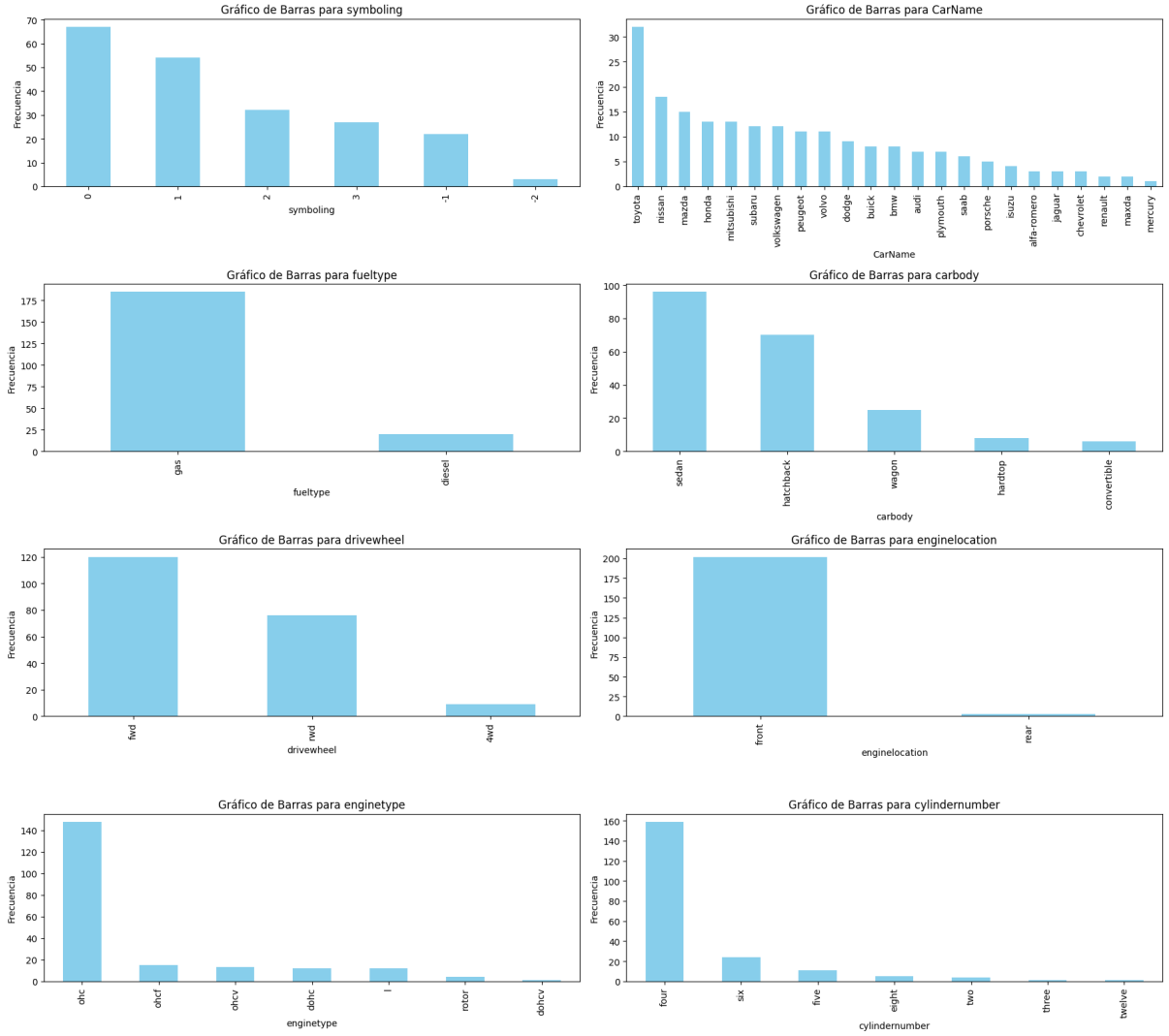


Fig 5. Frecuencia de las variables con respecto al precio

Se identificó que existe sesgo en las variables de ***carname,*** identificando que la variable que engloba gran proporción de las muestras son los autos **Toyota** por lo que se puede decir que el siguiente análisis aplica para su mayoría a carros **Toyota,** es decir, este análisis no resulta tan genérico**.**

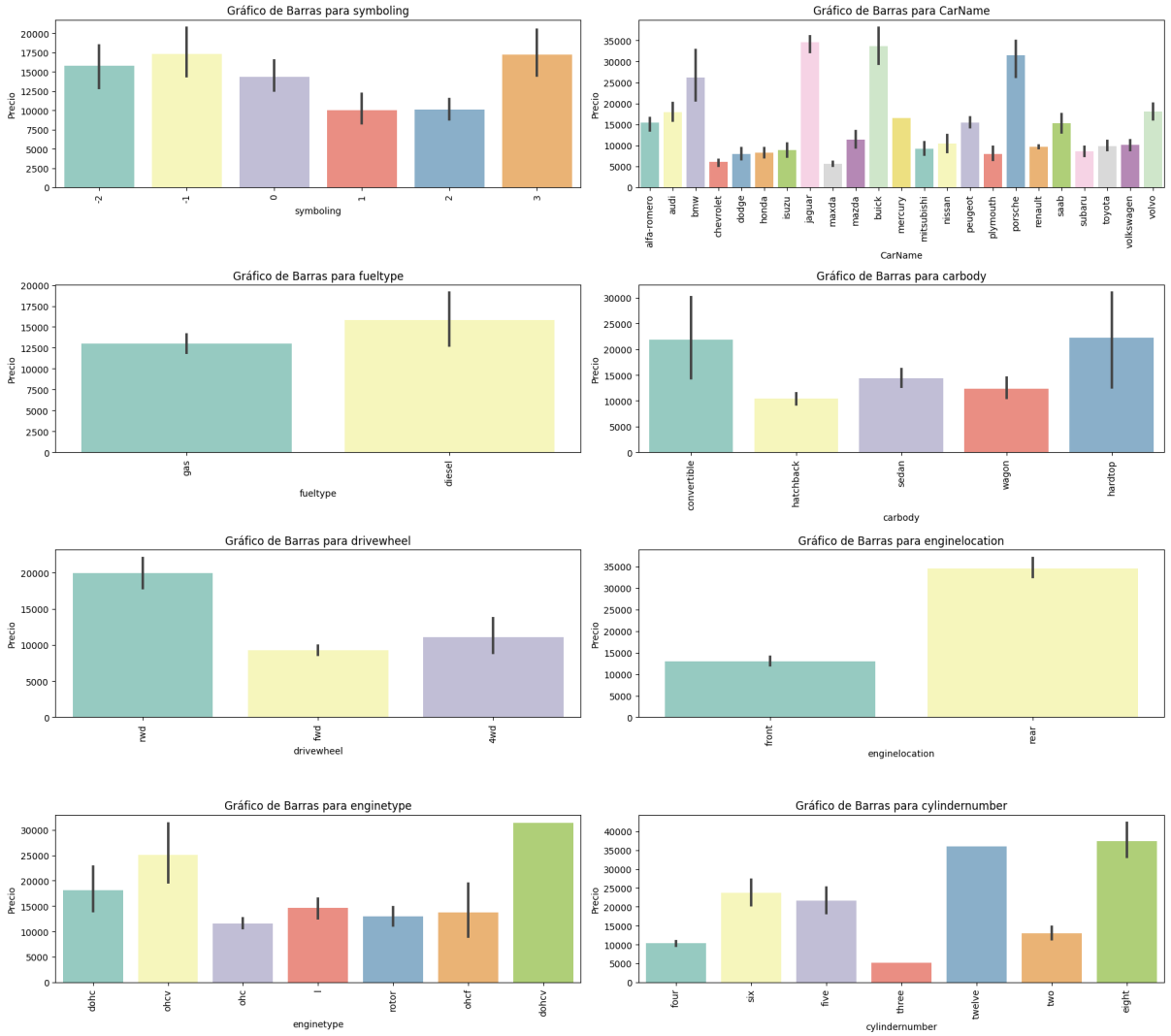


Fig 6. Relación de las variables con respecto al precio

De acuerdo con la gráfica de barras, se denota que las variables en donde el precio se ve como un gran diferenciador son ***carname, symboling, fueltype.***

Gracias a este panorama podemos identificar seis posibles variables que se relacionan directamente con el precio, escogemos tres cuantitativas y tres cualitativas: ***carname, symboling, fueltype, urbwaight, enginesize, horsepower*..**

Una vez identificadas dichas variables se revisó que las variables cuantitativas no tuvieran valores nulos y se verificaron los valores atípicos, se eliminaron dichos valores para evitar tener ruido en nuestros resultados. Esto redujo nuestra muestra de 205 valores a 190.

En cuanto a las variables categóricas, estas se transformaron en variables dummies y las categorías obtenidas se concatenaron al dataset.

	curbweight	enginesize	horsepower	symboling	cylinder_eight	cylinder_five	cylinder_four	cylinder_six	cylinder_three	cylinder_twelve	cylinder_two	fueltype_diesel	fueltype_gas
0	0.143050	0.431717	0.484293	1.784802	0.0	-0.247896	0.441552	-0.29277	-0.072739	0.0	-0.146647	-0.342997	0.342997
1	0.143050	0.431717	0.484293	1.784802	0.0	-0.247896	0.441552	-0.29277	-0.072739	0.0	-0.146647	-0.342997	0.342997
2	0.751562	1.252351	1.953536	0.150525	0.0	-0.247896	-2.264737	3.41565	-0.072739	0.0	-0.146647	-0.342997	0.342997
3	-0.323845	-0.351616	0.176777	0.967664	0.0	-0.247896	0.441552	-0.29277	-0.072739	0.0	-0.146647	-0.342997	0.342997
4	0.753775	0.655526	0.620967	0.967664	0.0	4.033947	-2.264737	-0.29277	-0.072739	0.0	-0.146647	-0.342997	0.342997

Fig 7. Head del dataset final a evaluar

Modelación y verificación del modelo

Teniendo un panorama de los datos y habiendo reducido nuestras variables dependientes, buscamos implementar dos modelos estadísticos para determinar la relación de dichas variables con el precio. Buscamos tener una mayor cobertura de modelos, por lo que se seleccionaron modelos lineales (Regresión Lineal, Mínimos Cuadrados Ordinarios) y no lineales (Modelo Gradiente Descendiente).

Regresión Lineal

Para esto se estandarizaron los datos y se dividió el dataset en 80% de entrenamiento y 20% de testeo. Se aplicó un modelo de regresión lineal que nos devolvió un $R^2 = 0.825$.

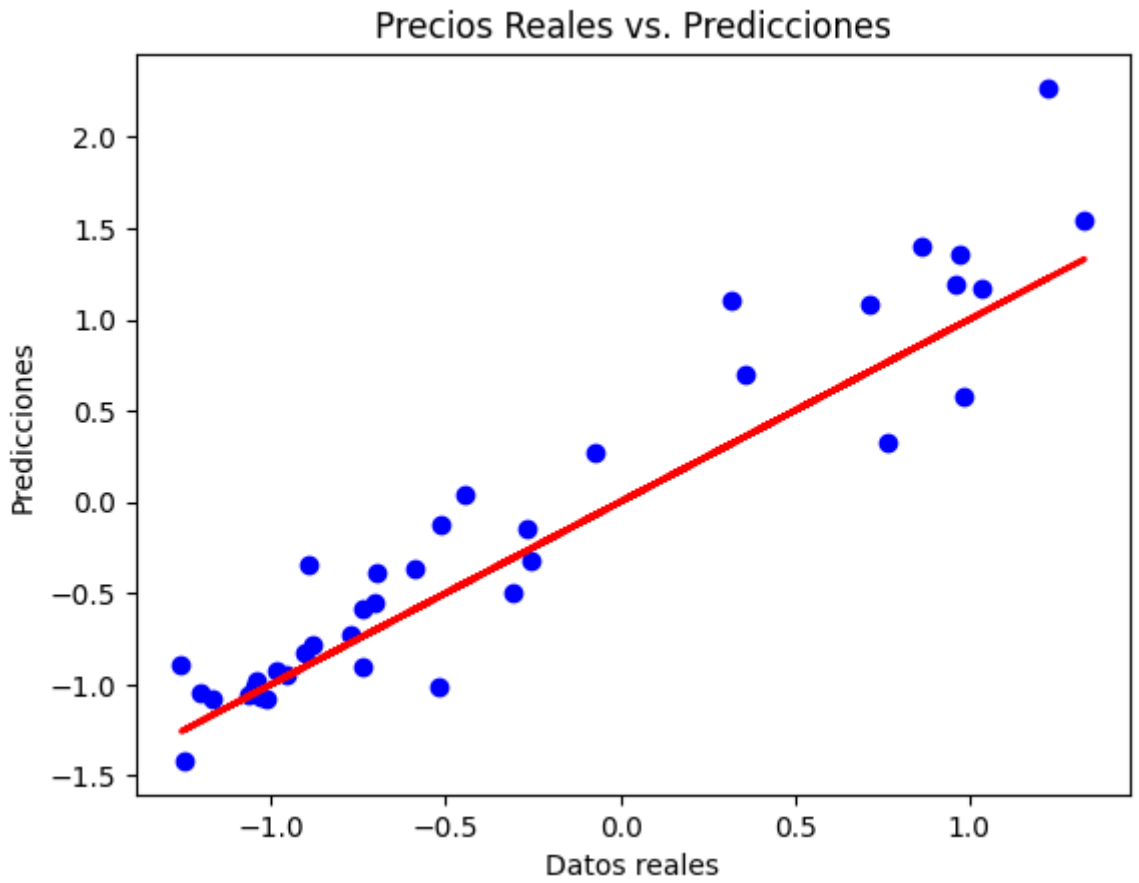


Fig 8. Verificación del modelo

En esta gráfica podemos ver los valores reales (línea roja) y los valores que predijo el modelo (puntos azules), su aproximación es buena de acuerdo con el coeficiente de R^2 ya que es superior a 0.80.

Mínimos cuadrados ordinarios (OLS)

En este modelo buscamos un valor de R cuadrada adecuado y lo que nos devuelve el modelo es también los valores de P, que podemos comparar con un alpha = 0.05, para determinar la relación entre la variable con el precio.

En la primera prueba ponemos a comparar las seis variables escogidas y se obtienen los siguientes resultados.

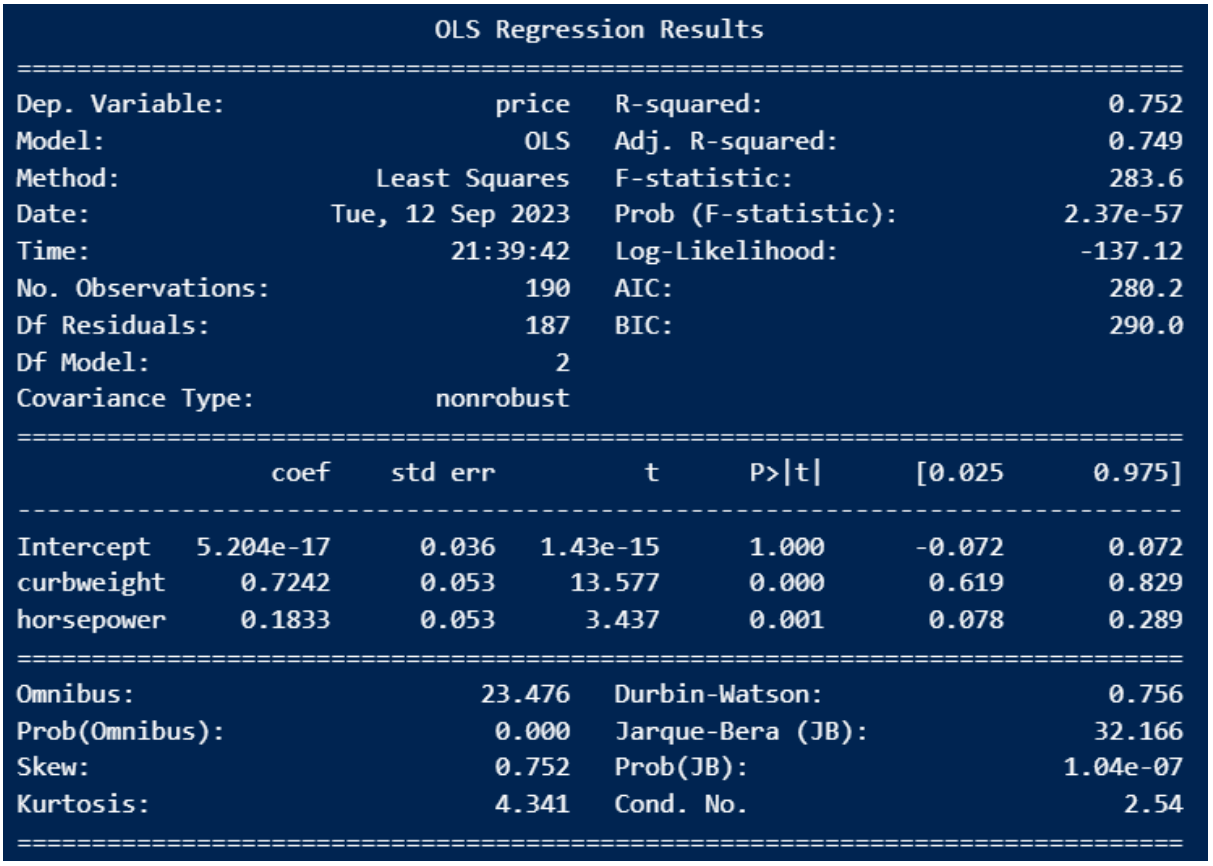
OLS Regression Results

Dep. Variable:	price	R-squared:	0.809
Model:	OLS	Adj. R-squared:	0.799
Method:	Least Squares	F-statistic:	84.53
Date:	Tue, 12 Sep 2023	Prob (F-statistic):	7.14e-60
Time:	21:39:42	Log-Likelihood:	-112.49
No. Observations:	190	AIC:	245.0
Df Residuals:	180	BIC:	277.5
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.245e-17	0.033	1.92e-15	1.000	-0.064	0.064
curbweight	0.5459	0.082	6.685	0.000	0.385	0.707
enginesize	-0.0264	0.089	-0.295	0.768	-0.203	0.150
horsepower	0.2670	0.070	3.839	0.000	0.130	0.404
symboling	-0.0134	0.038	-0.354	0.723	-0.088	0.061
cylinder_eight	2.475e-17	8.59e-17	0.288	0.773	-1.45e-16	1.94e-16
cylinder_five	0.1408	0.078	1.811	0.072	-0.013	0.294
cylinder_four	-0.1327	0.105	-1.260	0.209	-0.341	0.075
cylinder_six	-0.0222	0.097	-0.228	0.820	-0.214	0.170
cylinder_three	0.0087	0.038	0.230	0.818	-0.066	0.083
cylinder_twelve	0	0	nan	nan	0	0
...						

Fig 9. Verificación del modelo OLS primera prueba

Gracias al valor P (siendo este inferior al alpha) identificamos que las variables con más relación con el precio son: ***curbweight*** y ***horsepower***. En la siguiente prueba ponemos a evaluar en el modelo estas dos variables para ver la aproximación de R cuadrada y vemos que no baja demasiado.



OLS Regression Results

Dep. Variable:	price	R-squared:	0.752
Model:	OLS	Adj. R-squared:	0.749
Method:	Least Squares	F-statistic:	283.6
Date:	Tue, 12 Sep 2023	Prob (F-statistic):	2.37e-57
Time:	21:39:42	Log-Likelihood:	-137.12
No. Observations:	190	AIC:	280.2
Df Residuals:	187	BIC:	290.0
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	5.204e-17	0.036	1.43e-15	1.000	-0.072	0.072
curbweight	0.7242	0.053	13.577	0.000	0.619	0.829
horsepower	0.1833	0.053	3.437	0.001	0.078	0.289

Omnibus:	23.476	Durbin-Watson:	0.756
Prob(Omnibus):	0.000	Jarque-Bera (JB):	32.166
Skew:	0.752	Prob(JB):	1.04e-07
Kurtosis:	4.341	Cond. No.	2.54

Fig 10. Verificación del modelo OLS segunda prueba

Esto nos ayuda a determinar que efectivamente, las variables con más impacto en el precio son: ***curbweight*** y ***horsepower***. En seguida se realizan más pruebas agregando y quitando diferentes variables del modelo pero se llega a la misma conclusión.

Modelo Gradiente Descendiente

Para este modelo lineal se utilizo el mismo dataset dividido en entrenamiento y testeo y se obtuvo la siguiente ecuación.

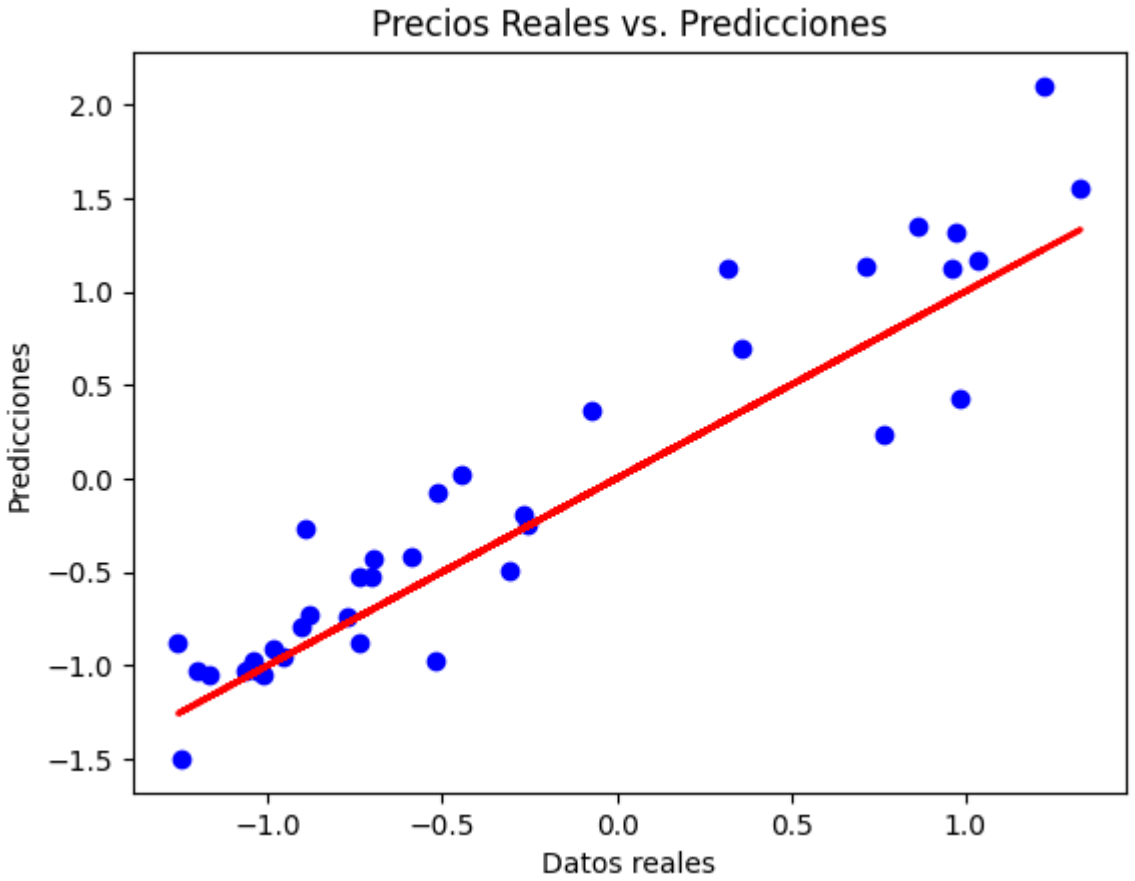


Fig 11. Verificación del modelo Gradiente descendiente

Conclusión

Podemos concluir que las variables que escogimos sí tienen un impacto significativo en el precio sin embargo hay algunas que se pueden descartar ya que hay otras que superan el impacto o la importancia. Gracias a los modelos estadísticos se nos permitió concluir esto y aún hay bastantes áreas de oportunidad en donde se pueden buscar otros modelos no lineales que se ajusten mejor a los datos que se tienen.

Anexos

Link de **Google Colab**, en donde se realizó el procesamiento y análisis estadístico de la base de datos. • ***ProcesamientoDeDatos.ipynb***	https://github.com/FCANOF/PortafolioAnalisis_TE3006_101_FridaCanoFalcon_A01752953/tree/main/final/M1_Estadistica/Precio%20de%20Autos