1.

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

Alumno:

Guillermo Romeo Cepeda Medina A01284015

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 el presente trabajo se busca brindar un apoyo a una empresa nueva en el mercado automotriz por medio del análisis estadístico de un dataset de automóviles con sus respectivas características y precio. En el análisis se encuentran dos tipos de variables entre las cuales encontramos significancia estadística en la predicción del precio de un automóvil según las mismas, como lo es el "Horsepower", el tamaño del motor y el peso del vehículo así como también se encontró que dos variables cualitativas tienen una alta capacidad de describir el precio del automóvil como lo es el nombre de la compañía del vehículo y a lo que le llaman "Symboling". Concluimos que estas variables son importantes para tomar en cuenta al tratarse de una empresa que quiere entrar al mercado automotriz y que usarlas a su favor podría aumentar el rendimiento de la empresa.

Introducción

En el presente trabajo se tiene como objetivo analizar de manera estadística el dataset "precios autos" con motivo de resolver una problemática en la cual una empresa nos pide responder las preguntas de ¿Qué variables son significativas para predecir el precio de un automóvil? ¿Qué tan bien describen esas variables el precio de un automóvil? Para esto se llevó a cabo un análisis estadístico exhaustivo de las variables involucradas, así como la detección de valores atípicos en los vehículos. Este análisis tiene valor debido a que plantea una solución importante al momento de que entra una nueva empresa al mercado. Si se desea entrar a un mercado existente, en el que el precio de los vehículos tiene una pauta por ciertas características, es importante tenerlas en cuenta para que los principios de la marca se encuentren bajo fundamentos sólidos y puedan ser competitivos contra otras compañías de autos.

El análisis que realice se enfocó principalmente en la separación de variables cuantitativas y cualitativas, esto me permitió utilizar las características de cada una de las variables para encontrar las de mayor significancia y entender que tan bien describen dichas variables el precio del automóvil. Para este reporte seleccionamos 3 variables cualitativas y 3 variables cuantitativas El dataset con el que contamos tiene las siguientes características:

Variables cuantitativas: wheelbase, carlength, carwidth, carheight, curbweight, enginesize, stroke, compressionratio, horsepower, peakrpm, citympg y highwaympg

Variables cualitativas: symboling, CarName, fueltype, carbody, drivewheel, enginelocation, enginetype, cylindernumber

Utilizando estas variables se puede llegar a la conclusión de cuáles de ellas son significativas para la prediccion del precio y que tan buenas son para describir el precio del automovil.

Selección de las variables:

Con el objetivo de entender mejor las variables se obtuvieron los siguientes parámetros estadísticos de las variables cuantitativas:

- Media
- Desviación estándar
- Cuartiles 1,2,3
- Rango Intercuartil
- Límite superior
- Límite inferior
- Número de “Outliers”
- Correlación entre la variable y el precio

Se ordenaron las variables cuantitativas de mayor a menor correlación y se seleccionaron las 3 con mayor correlación para nuestro análisis. Estas variables son: enginesize, horsepower, curbweight.

Para seleccionar las variables cualitativas, primero se hizo un análisis de frecuencias, asi como un análisis anova de la varianza entre el precio y las variables, para seleccionar las que tengan un mayor estadístico F. Las variables con mayor estadístico F fueron: CarName, Symboling y carbody.

Análisis de resultados

En la siguiente sección se mostrarán los resultados obtenidos en el análisis estadístico que se realizó con el dataset

Para las variables cuantitativas con mayor correlación encontramos los siguientes gráficos:

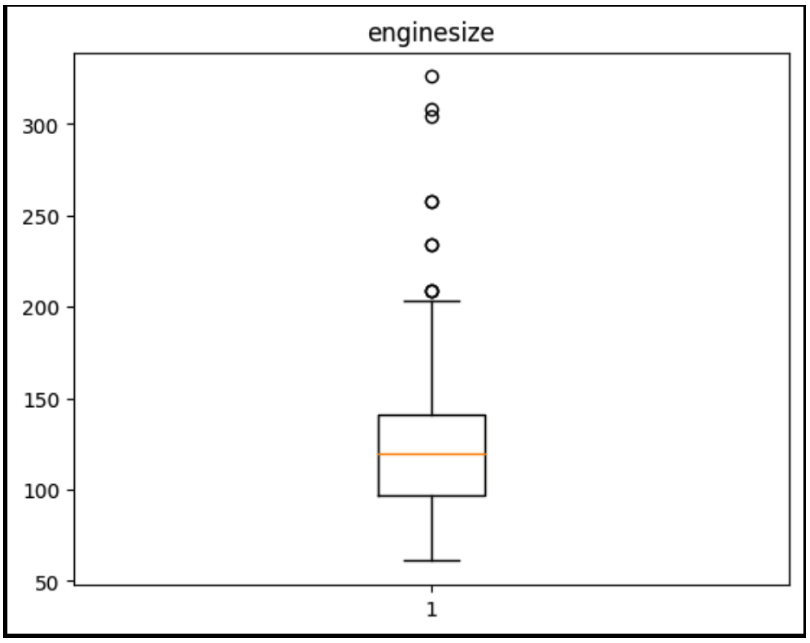

Imagen 1.1 Box-Plot enginesize, outliers de 200 a 400 en las unidades correspondientes

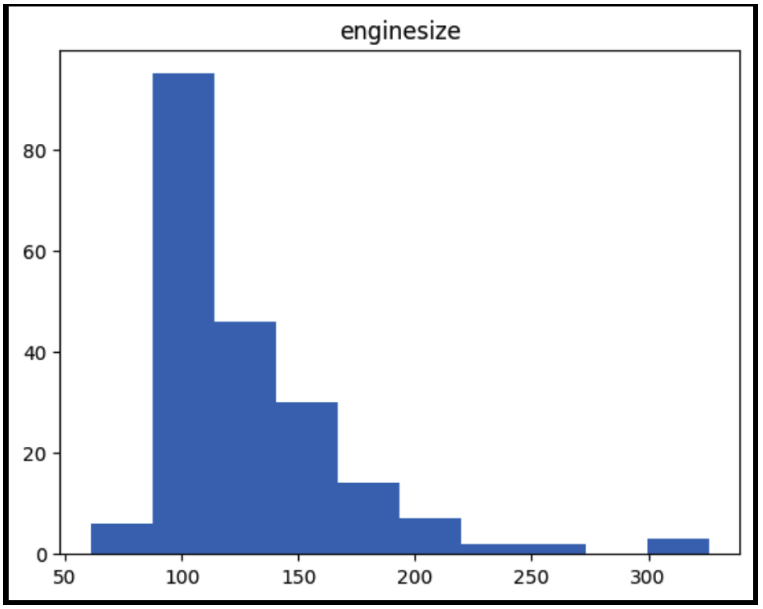

Imagen 1.2 histograma enginesize, los outliers se encuentran mejor definidos después de las 300 unidades

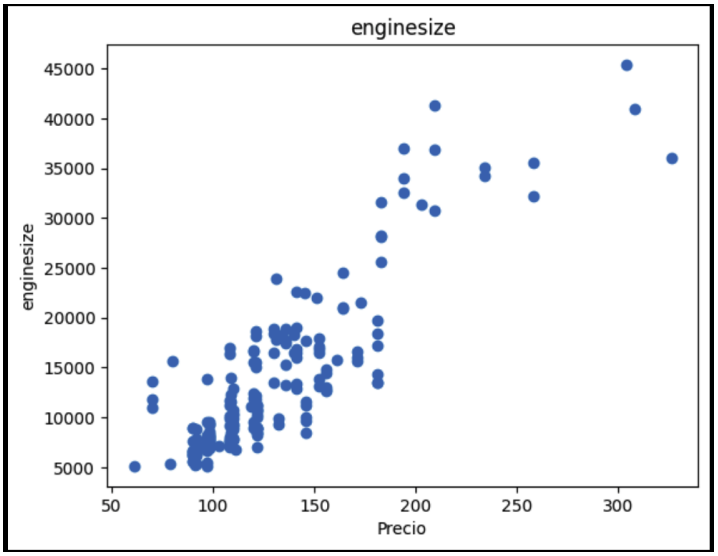

Imagen 1.3: Gráfica de precio vs enginesize en milimetros, se aprecia la correlación en todo momento

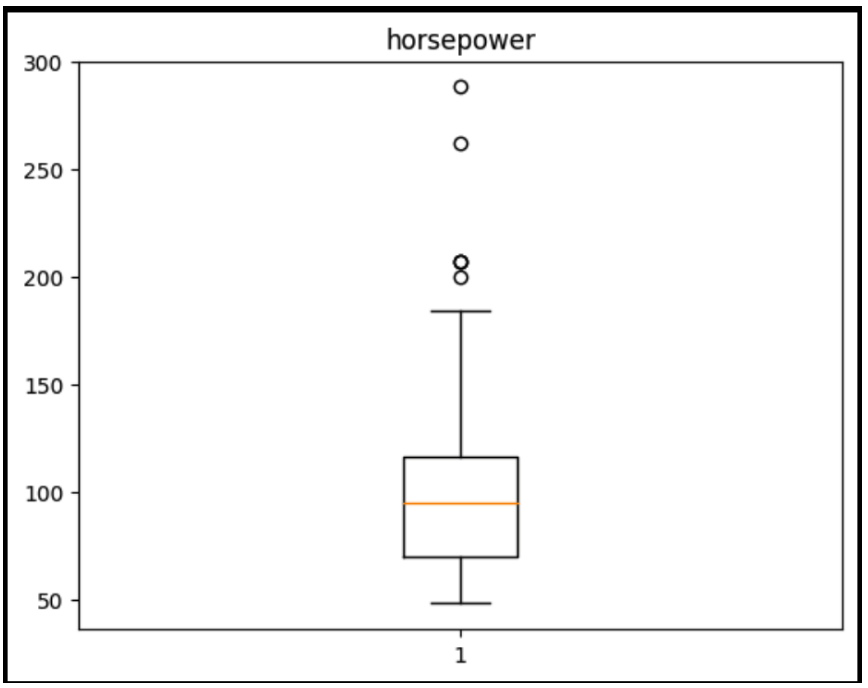

Imagen 2.1 Box-Plot horsepower, outliers de por encima de 200 unidades

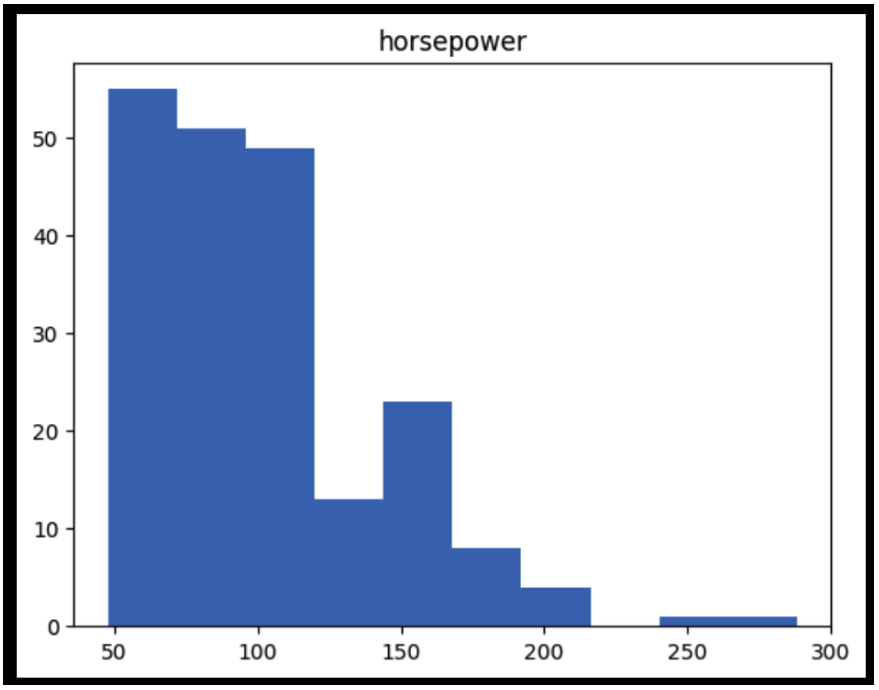

Imagen 2.2 histograma horsepower, los outliers se encuentran mejor definidos después de las 250 unidades

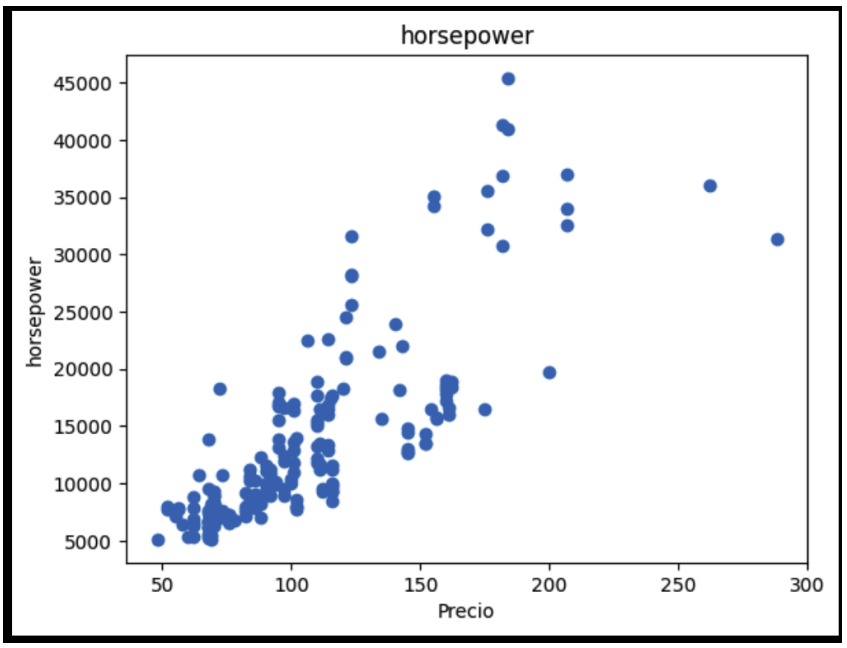

Imagen 2.3: Gráfica de precio vs horsepower, se aprecia la correlación en todo momento

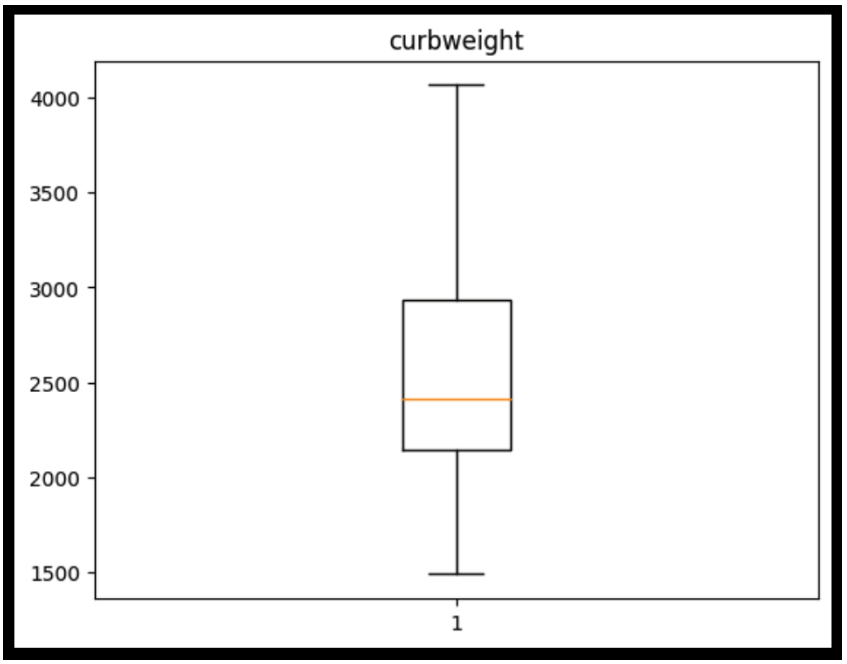

Imagen 3.1 Box-Plot curbweight, no se avistan outliers

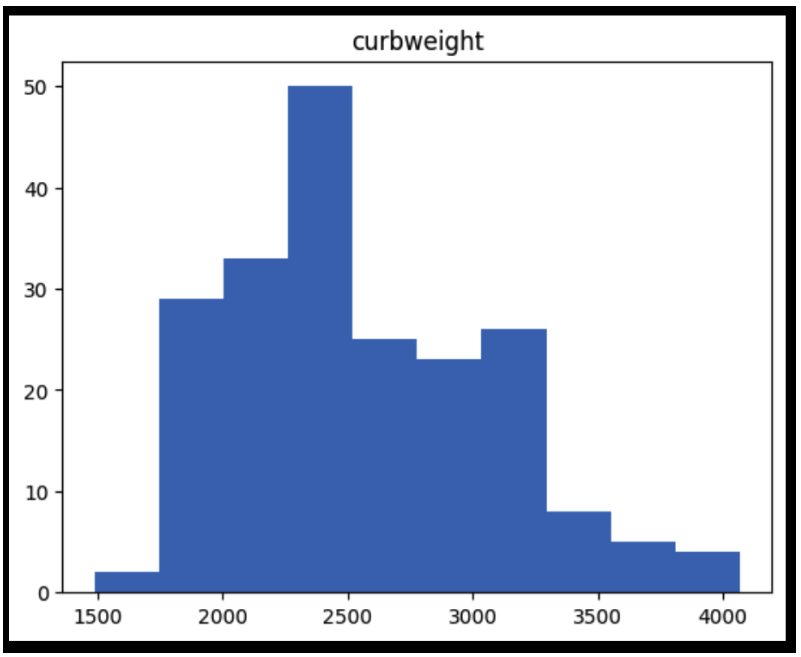

Imagen 3.2 histograma curbweight

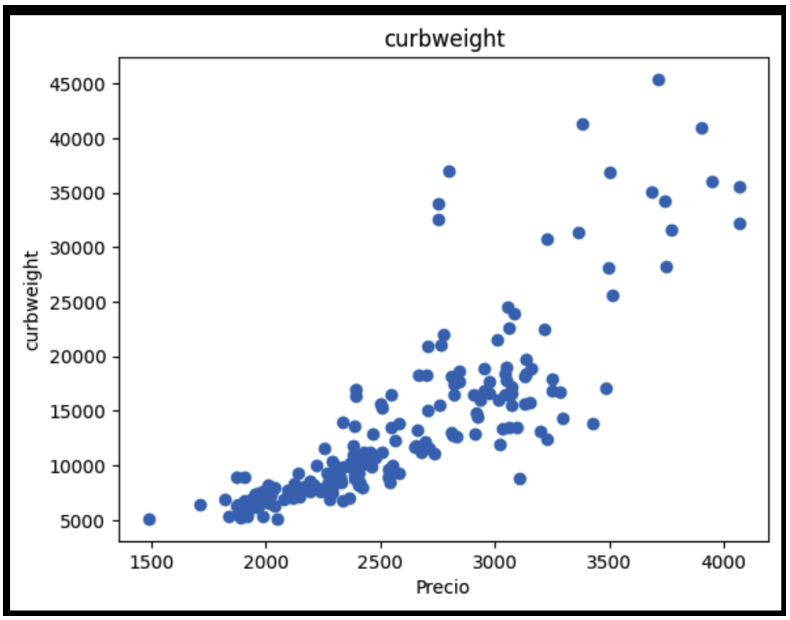

Imagen 3.3: Gráfica de precio vscurbweight, se aprecia la correlación en todo momento

Utilizando una regresión lineal podemos identificar qué tan buenas son las variables para predecir el precio del automóvil.

Tras haber normalizado los datos y utilizando una regresión lineal simple con estas tres variables obtuvimos los siguiente resultados:

Error cuadrático medio: 14145734.53
Coeficiente de determinación: 0.82
Score del modelo: 0.82
Statistics=0.922, p=0.008
Los datos no siguen una distribución normal (se rechaza H0)

Statistics=0.999, p=0.318
Las varianzas son iguales (no se rechaza H0)

Tenemos un error cuadrático medio muy grande, además de que nuestros datos no siguen una distribución normal. Sin embargo, contamos con un coeficiente de correlación cercano a 1, lo cual indica que nuestro modelo es bueno para explicar la varianza de los datos. Se podría investigar si esto se debe a los llamados “Detalles” que brindan las demás características del dataset para describir el precio de los automóviles.

Variables cualitativas

A continuación mostraré las gráficas de histogramas, así como comparación contra el precio de las variables cualitativas.

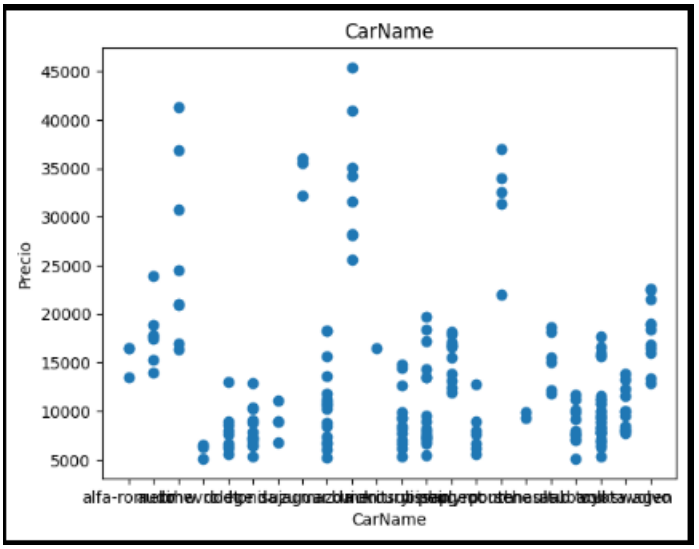

Figura 4.1 CarName vs Precio

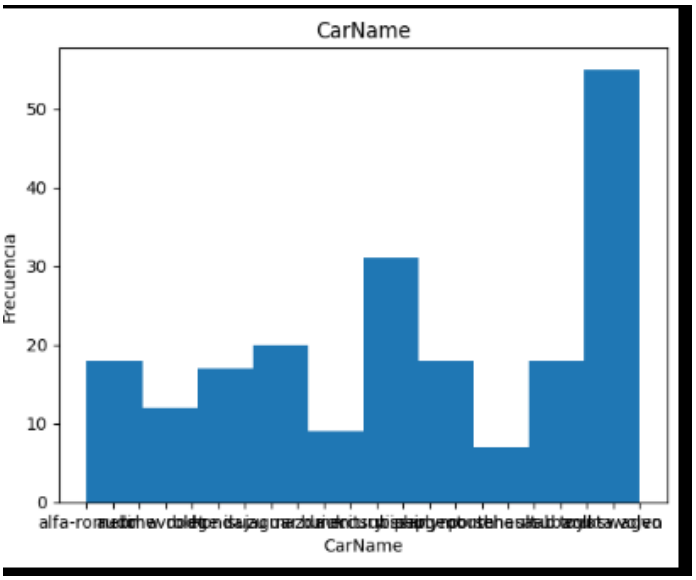

Figura 4.1 Frecuencia CarName

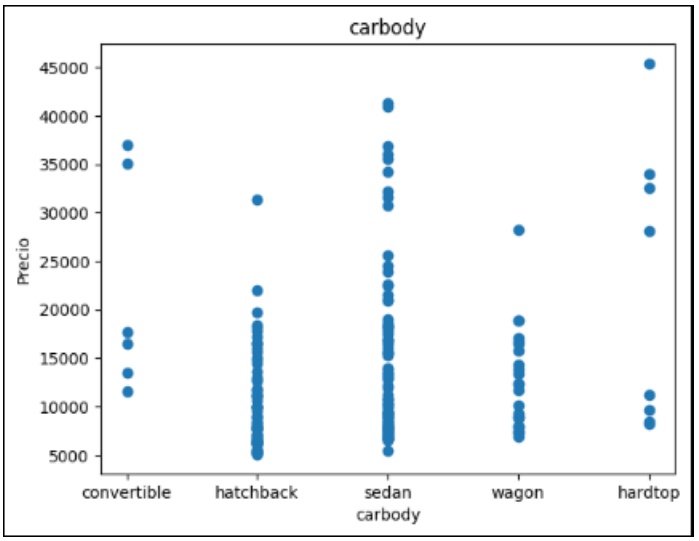

Figura 5.1 CarBody vs Precio

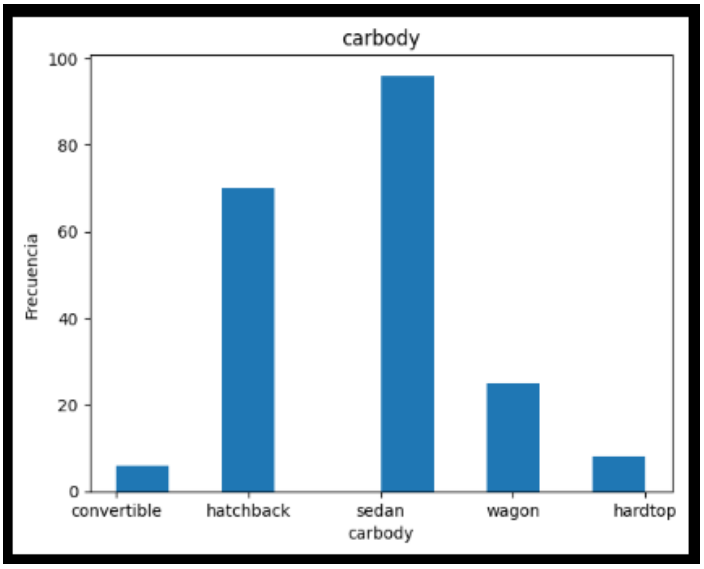

Figura 4.1 Frecuencia de CarBody

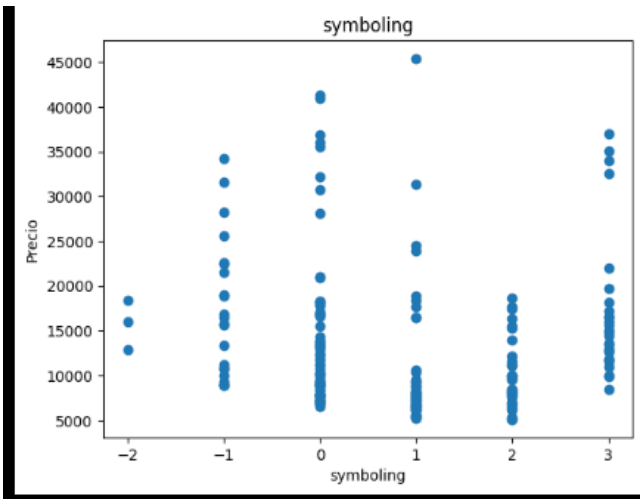

Figura 6.1 Symboling vs Precio

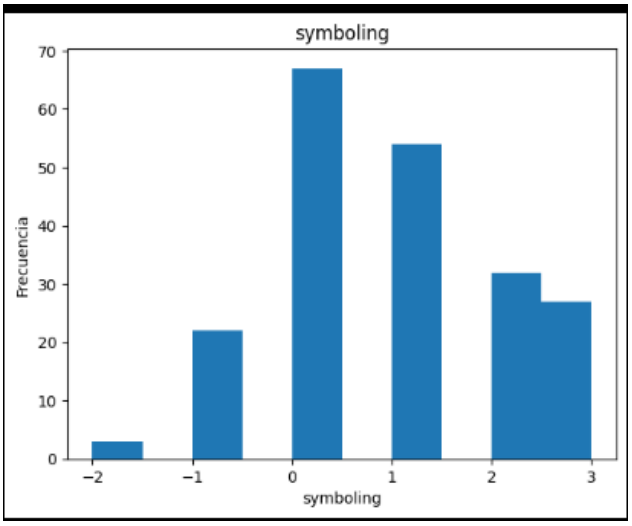

Figura 6.2 Frecuencia Symboling

Análisis ANOVA
Debido a los resultados obtenidos se puede concluir que:

Variable 'symboling':

- Estadístico F: 562.18
- Valor p (p-value): 9.34e-79

Interpretación:

El valor extremadamente bajo del p-value (9.34e-79) indica que hay evidencia significativa para rechazar la hipótesis nula. Esto sugiere que la variable 'symboling' tiene un impacto significativo en el precio del automóvil. En otras palabras, las diferentes categorías de 'symboling' están asociadas con diferencias en el precio.
Variable 'CarName':

Estadístico F: 564.98
Valor p (p-value): 5.19e-79

Interpretación:

Similar al caso anterior, el valor extremadamente bajo del p-value (5.19e-79) indica que la variable 'CarName' también tiene un impacto significativo en el precio del automóvil. Las diferentes marcas de automóviles están asociadas con diferencias en el precio.

La variable carbody entregó un F estadístico muy bajo y un p-value indefinido, por lo que parece que el análisis ANOVA no fue adecuado para esta categoría.

Conclusiones:

Con base en los resultados obtenidos en el análisis estadístico del conjunto de datos de precios de automóviles, se pueden extraer conclusiones valiosas que pueden ser de gran utilidad para una empresa que está ingresando al mercado automotriz.

En primer lugar, se ha determinado que las variables cualitativas 'CarName' y 'Symboling' son adecuadas para describir el precio de los automóviles. Esto se basa en el análisis de varianza (ANOVA), que indica que las diferentes marcas de automóviles están fuertemente relacionadas con diferencias en el precio. Por otro lado, en el análisis de regresión lineal, se identificó que las variables cuantitativas 'CurbWeight', 'EngineSeize' y 'HorsePower' son altamente significativas para predecir el precio de los automóviles en el mercado. A pesar de un error cuadrático medio relativamente alto, el coeficiente de determinación (R-squared) de 0.82 indica que el modelo de regresión lineal es capaz de explicar el 82% de la variabilidad en los precios. Esto sugiere que estas tres variables tienen un poder explicativo notable sobre el precio de los automóviles.

Estos hallazgos son cruciales para una nueva empresa que busca comprender mejor el mercado automotriz y comercializar sus productos de manera más efectiva. Se recomienda que la empresa considere la influencia de las marcas de automóviles ('CarName') y las categorías de 'Symboling' en su estrategia de precios y segmentación de mercado. Además, las variables cuantitativas, como el peso del vehículo ('CurbWeight'), el tamaño del motor ('EngineSeize') y la potencia del motor ('HorsePower'), deben ser tenidas en cuenta al establecer los precios y diseñar estrategias de marketing. Si bien estos resultados proporcionan una base sólida para la toma de decisiones, se sugiere que la empresa continúe investigando y refinando su estrategia a medida que adquiera más información sobre el mercado y sus clientes. En conjunto, este análisis estadístico ofrece una valiosa orientación para una empresa emergente en el competitivo mercado automotriz.