Instituto Tecnológico y de Estudios Superiores de Monterrey

Escuela de Ingeniería y Ciencias

Ingeniería en Ciencia de Datos y Matemáticas

Inteligencia artificial avanzada para la ciencia de datos I

Módulo 1: Estadística para ciencia de datos

Paola Sofia Reyes Mancheno A00831314

Supervisado por:

Dra. Blanca R. Ruiz Hernández

Monterrey, Nuevo León. Fecha, 11 de septiembre de 2023

1. Resumen

2. Introduction

A lo largo del mundo, los mercados para las diferentes industrias cambian dependiendo de la región donde se encuentren. Esto no es una diferencia en la industria automovilística, en donde dependiendo de la región del mundo donde te encuentres, las regulaciones, materiales, marcas e incluso la importancia de las características de un auto, son completamente diferentes. Este es el escenario de la situación problema que se trabaja en este proyecto.

En el año 2022, la industria automotriz China fue la que más carros produjo, incluso sobrepasando 3 veces la producción de carros estadounidense. [1] Esto posiciona a China y sus empresas como las más fuertes en el mercado actualmente, por lo que no es sorpresa su objetivo de expandirse en diferentes mercados diferentes al oriental. En esta situación problema, una empresa China desea expandirse específicamente al mercado estadounidense, por lo que previamente desean realizar una investigación de dicho mercado.

En esta investigación se tiene como objetivo identificar las características o factores que influyen en el precio de automóviles en este país. Este objetivo nace ya que, el precio de los autos no solo se fija por el precio de sus partes o la calidad de ellas, sino también sobre como el mercado de la región valora unas u otras partes/ características del vehículo.

Para poder cumplir este objetivo, se han planteado dos preguntas que ayuden a analizar la base de datos de vehículos en venta en Estados Unidos con todas sus características y precios. Por un lado, se necesita definir qué variables son significativas para predecir el precio del automóvil; y por otro, qué tanto esas variables llegan a describir este precio. [2]

Estas dos cuestiones se van a resolver mediante un análisis estadístico descriptivo y predictivo donde en primer lugar se conozca de manera general el comportamiento de todas las variables, así como su correlación con respecto a la variable independiente, el precio. Posterior a la sección descriptiva del análisis, se procede a definir 6 variables que posiblemente sean significativas para el precio de los vehículos. Finalmente, se evaluan las variables categóricas a través de análisis ANOVA, donde se defina si las categorías de cada variable son estadísticamente diferentes con respecto al precio de los autos y también se realiza una regresión lineal múltiple que tiene como meta generar un modelo que ayude describir los precios de autos y posteriormente evaluar qué tan bien lo hacen las variables escogidas.

3. Análisis de Resultados

3.1. Exploración de la base de datos

La base de datos utilizada en este proyecto, tiene como nombre "precios_autos", la cual está en un documento tipo .csv. En total, el dataset cuenta con 21 variables, las cuales 13 son cuantitativas (incluyendo el precio) y 8 categóricas.

3.1.1. Estadística descriptiva

Variables Cuantitativas

Para analizar la distribución de los datos cuantitativos, se hace el uso de histogramas en

la figura 1. Para estas 13 variables, gráficamente las distribuciones están repartidas entre las siguientes: normal, sesgada a la derecha (la mayoría de las variables), e incluso exponencial. Específicamente hablando de la variable objetivo, "price", esta se muestra como una distribución exponencial que tiene ciertos valores influyentes al final de su curva. Esto significa que, posterior a la elección de variables, será importante considerar la normalización de esta variable, así como la eliminación de outliers para una mejor calidad del modelo.

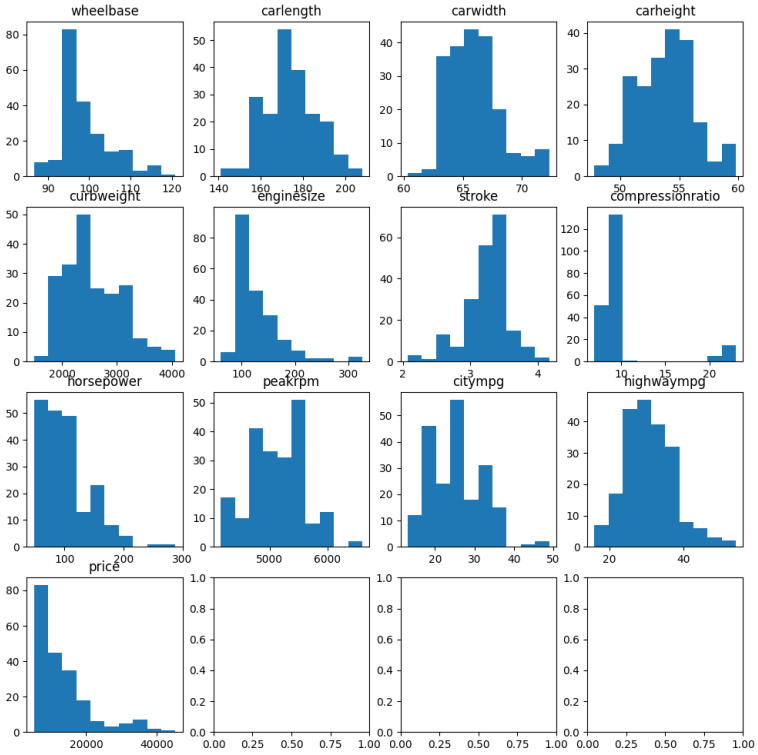


Figura 1: Histogramas de distribución variables cuantitativas

Por otro lado, al tener una gran cantidad de variables cualitativas, es importante realizar un análisis de correlación entre ellas. Esto con el objetivo de observar si es que existen variables que se relacionan entre sí, y por ende ser repetitivas al describir al precio de los autos.

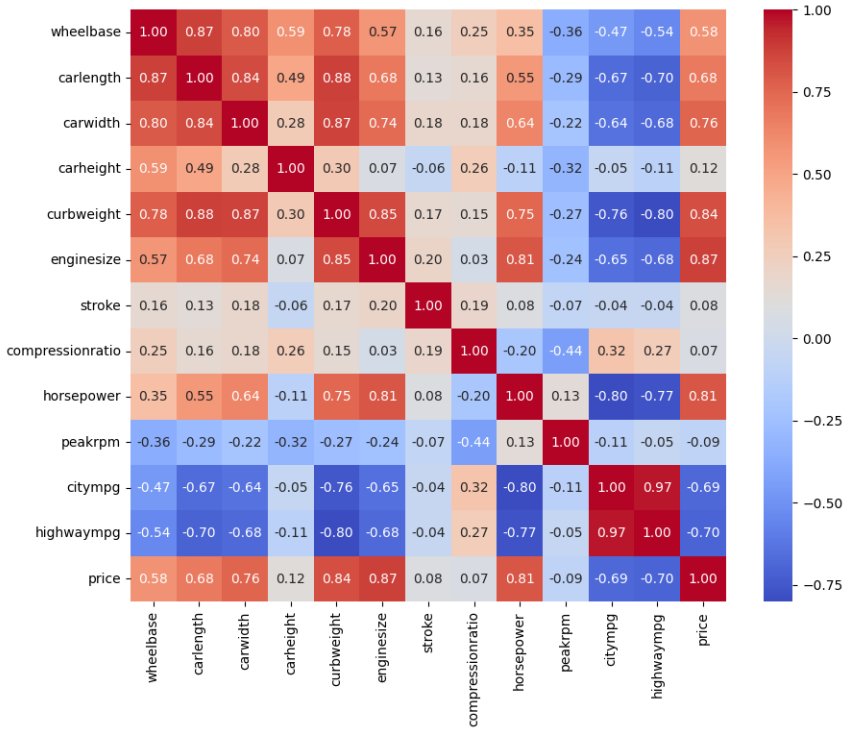


Figura 2: Heatmap de correlación entre variables cuantitativas

La matriz de correlación de Pearson, que se encuentra en la figura 2, permite observar que algunos pares de variables tienen gran correlación entre ellos, lo que nos da la posibilidad de identificar qué variables no deberían escogerse juntas para el modelo para evitar multicolinealidad. Algunas de ellas son, CityMPG con HighwayMPG (al tratarse ambas de las millas por galón), WheelBase con CarWidth,CarWeight con Curve Weight, etc.

Ahora bien, la matriz de correlación de Pearson también ayuda a definir qué variables ayudan a explicar la variable objetivo, al tener una correlación absoluta mayor al 0.5. En este caso, la variable "price"tiene una correlación mayor al 0.5 con 6 variables (wheelbase, carlength, carwidth, curbweight, enginesize, horsepower), y correlación negativa significativa con CityMPG y HighwayMPG. Esta matriz de correlación, brinda ya una idea general de qué variables describen de mejor manera el precio de los autos, por lo que posteriormente se realizará una comparación más exhausta para escoger las variables.

Variables Categóricas

Ahora bien, al hablar de variables categóricas, estas deben ser analizadas de forma diferente que las cuantitativas. En este caso, la distribución de las variables cumple la función de entender las características más o menos usadas en los automóviles en Estados Unidos.

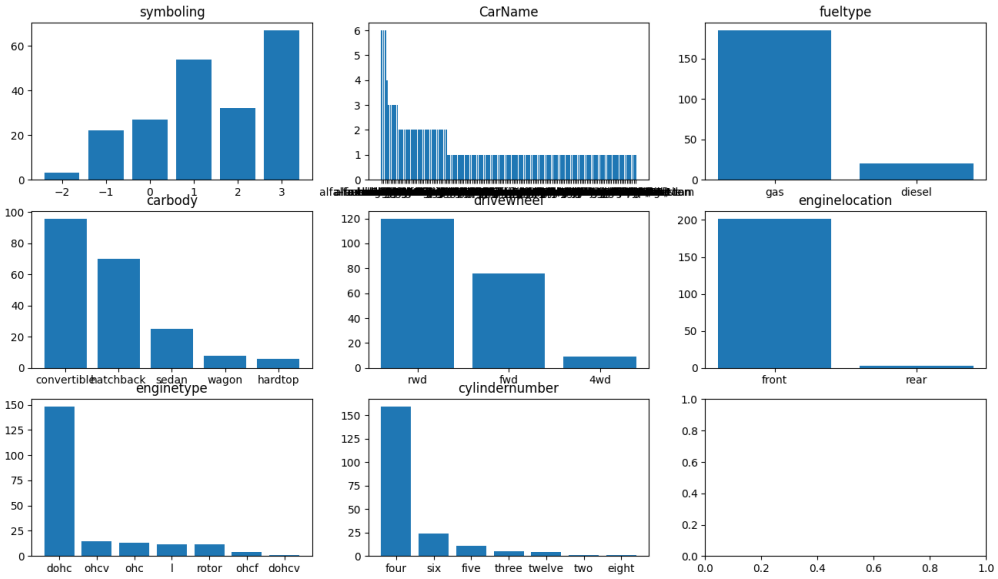


Figura 3: Distribución de categorías de variables categóricas

En la figura 2, la distribución de los datos no es cercano a uniforme en ninguna de estas. Tanto en FuelType, EngineLocation, EngineType y CilinderNumber, una de sus categorías reúne a la mayoría de los registros. No obstante, debido a que no existe una distribución uniforme en ninguna de las variables, no existe una buena representación de todas las categorías. Esto puede resultar en que al aplicar un modelo de regresión lineal multiple, exista heterocedasticidad en los residuos, pues la varianza de estos puede llegar a ser diferente para cada una de las categorías.

Por otro lado, es importante analizar la asociación de las categorías con la variable dependiente. En la figura 4 los diagramas de caja y bigote se muestran las variables categóricas con relación al precio resaltando su mediana, cuartiles y outliers. En la mayoría de ellas se observa una gran diferencia en las medidas de dispersión de las categorías, por lo que los boxplots no abarcan el mismo rango de precios. Por ejemplo, en la variable de enginetype el tipo ohcv cubre precios desde 1300 hasta 4500 aproximadamente, mientras que el resto de tipos tiene un rango de valores más corto. La variable de fueltype es la única en donde los boxplots se comportan similarmente, pero aun así hay bastantes outliers en el primer tipo.

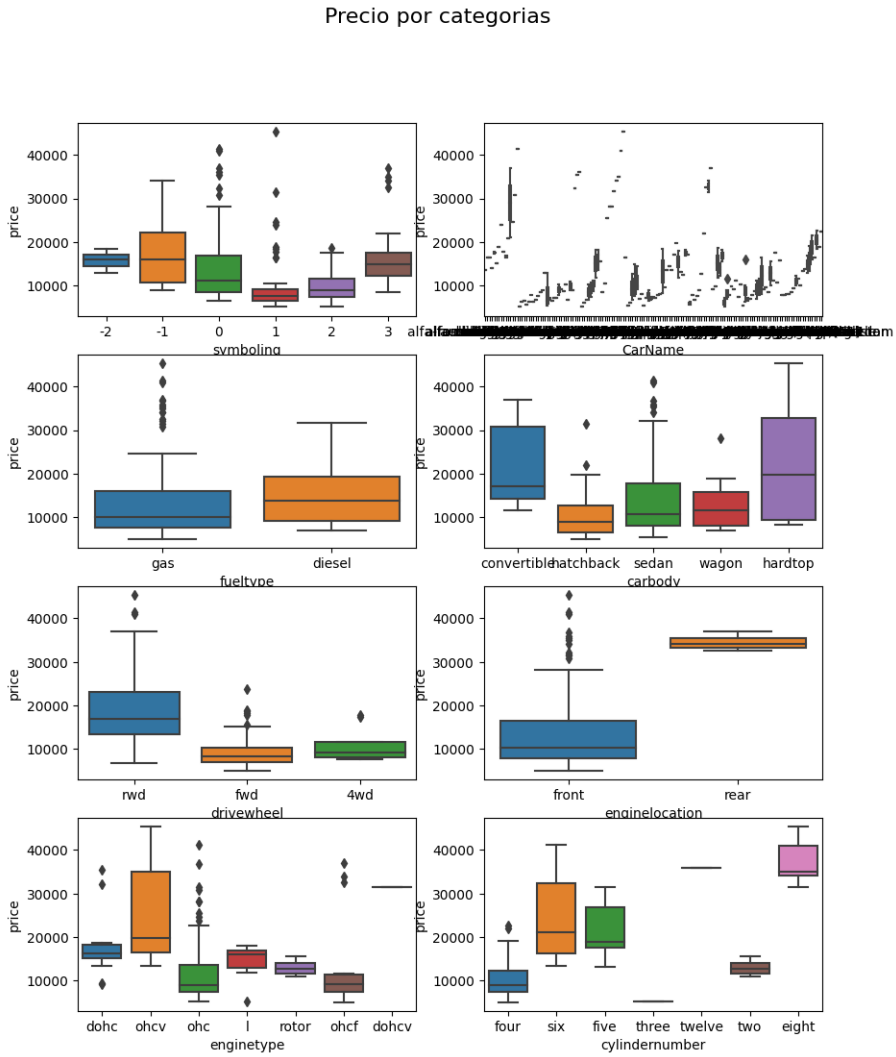


Figura 4: Diagrama de caja y bigote de Precio por Categoría

3.1.2. Calidad de datos

Se pudo observar que de las 205 observaciones, ninguna de ellas tiene un valor nulo. Con respecto a los outliers, se utilizó el rango intercuartil o IQR para calcular los valores atípicos de las variables cuantitativas; en el caso de la variable dependiente, los outliers representan el 7 %. Debido a esto, se decidió realizar la construcción del modelo de regresión lineal múltiple que se presentará posteriormente con y sin los 15 registros atípicos, lo cual resultó en un mejor modelo sin ellos y por ende se decidió eliminarlos.

3.2. Selección de variables

En total, se escogen 6 variables independientes (y 1 dependiente, price), las cuales, posterior a la preparación de los datos, serán 7. Estas variables fueron escogidas según el siguiente criterio:

1. A través de conocimiento cultural y por medio de investigación se preseleccionaron las variables más populares al promocionar la venta de un carro, y lo que las personas suelen buscar.

2. Para las variables preseleccionadas numéricas, se analiza según la matriz de correlación con respecto al precio, así como que no exista colinealidad entre dichas variables. Además, se busca que las variables tengan menos del 10 % de outliers.

3. En el caso de que las variables preseleccionadas son categóricas, se analizan los diagramas de caja del precio según estas variables y se seleccionaron las variables que mostraron diferencia de comportamiento dependiendo de las categorías.

El listado de variables es el siguiente:

- **Symboling:** *categórica.* Va de -2 a 3 y es la calificación del riesgo según el riesgo que representa el vehículo. Fue escogida ya que la distribución de los precios por categoría muestran una posible diferencia del precio, tomando a las categorías 0,1 y 2 con las de menor precio en comparación a las otras.

- **Car Body:** *categórica.* Tiene 5 categorías (Convertible, Hatchback, Sedan, Wagon y Hardtop) y representa el tipo de modelo del auto. Esta se escogió de igual manera por la distribución de los precios según los modelos, pues tanto el Hatchback como el Wagon tienen un menor rango de precios a diferencia del resto por lo que parece que puede describirse el precio de los autos a través de estas categorías. Además, el tipo de vehículo retoma características implícitamente como el alto, largo y ancho del auto.

- **Engine Type:** *categórica.* Esta variable tiene 7 categorías las cuales se ven muy diferentes entre sí en el diagrama de caja y bigotes de la figura 4. De igual manera, culturalmente esta es una característica importante para el público.

- **Wheel Base:** *numérica.* Es la distancia entre las llantas delanteras y traseras. Esta se escogió debido a una correlación moderada con "price.ªdemás de que no tiene colinealidad con el resto de variables escogidas. Es importante mencionar que esta variable tiene solamente 2 valores atípicos.

- **Cylinder Number:** *categórica.* Esta variable es el número de cilindros que posee el carro.

- **Horsepower:** *numérica.* Los caballos de fuerza, además de ser una característica que culturalmente el mercado está interesado en saber, es una de las variables con mayor correlación con la variable dependiente, lo que significa que ayuda a describirla.

- **City y Highway Miles Per Galon:** *numérica.* Estas variables se escogieron en un principio, con el fin de realizar un promedio de ambas debido a su correlación negativa con la variable dependiente. No obstante, se optó por eliminarlas al encontrar que existía una fuerte colinealidad con la variable Horsepower.

3.3. Herramientas estadísticas para responder la pregunta base

Posterior a haber escogido las variables que mejor describen a la variable objetivo, se procedió a preparar la base de datos al codificar a las variables categóricas, eliminar los registros con datos atípicos en "price", la discretización de las variables cuantitativas y la normalización de estas últimas. En este último paso se realizó tres pruebas diferentes, donde se probaban las herramientas estadísticas sin aplicar normalización, aplicando normalización a las 3 variables numéricas y solamente normalizando a "price". Con la comparación de dichos resultados, se tomó como la mejor opción el normalizar mediante Yeo-Johnson a la variable objetivo, pues de esta forma se encontraron los mejores resultados.

3.3.1. Análisis ANOVA para variables categóricas

Tomando en cuenta las dos preguntas que partieron del objetivo de este proyecto, se decidió realizar un análisis ANOVA para las diferentes variables categóricas que se escogieron en el modelo. El objetivo de esto es poder identificar si las diferentes categorías dentro de cada variable son estadísticamente diferentes entre sí con respecto al precio en los autos, y así poder saber cuáles de estas variables ayudan a describir la diferencia de precios de los autos.

Para ralizar las pruebas de hipótesis del análisis ANOVA, se utilizó una distribución muestral normal, debido a que la variable objetivo ya fue normalizada y es la que se utiliza normalmente para realizar esta prueba.

Con las hipótesis siguientes y la regla de decisión:

Hipótesis: H_0 : No hay diferencia significativa en el precio de los carros dependiendo de las categorías de la variable*

H_1 : Hay una diferencia significativa en el precio de los carros dependiendo de las categorías de la variable*

**entiéndase como variable a cualquiera de las variables categóricas escogidas.*

Regla de decisión: Se utilizará un $\alpha = 0{,}05$. El valor P debe ser menor a α para rechazar la hipótesis nula.

Al aplicar el análisis ANOVA para todas las variables, se obtuvo la siguiente tabla:

	P-Valor
Car body	$7{,}62 \times 10^{-3}$
Engine Type	$6{,}04 \times 10^{-6}$
Cylinder Number	$6{,}11x10^{-14}$
Symboling	$1{,}18x10^{-11}$

En la figura 3.3.1 están todos los p-valores obtenidos de los análisis ANOVA realizados. Tomando en cuenta el alpha de la regla de decisión, se rechaza la hipótesis nula para todas las variables. Con esto se puede afirmar que todas las variables categóricas describen correctamente la diferencia de precios en automóviles a través de sus categorias individuales, pues las medias del precio por categoría son estadísticamente diferentes entre sí; lo que responde a la primera pregunta. Además, apelando a la segunda pregunta de la situación problema, se observa que

las variables que mejor describen al precio de los autos son Symboling y Cylinder Number al tener p-valores bastante pequeños. En el caso de Car Body y Engine Type igual lo describen correctamente, pero la diferencia de precios entre categorias es menor a comparación con las otras dos variables mencionadas.

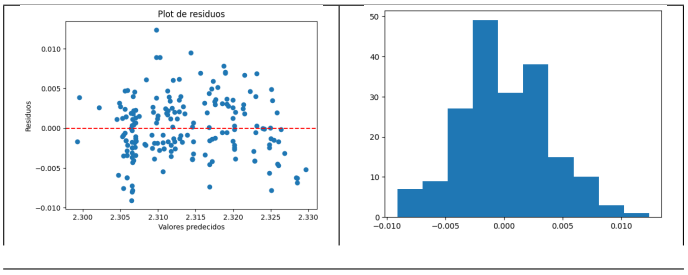
3.3.2. Regresión lineal múltiple con todas las variables.

Posterior al análisis ANOVA, se realizó una regresión lineal múltiple con las variables escogidas. Esta regresión se hará tanto con variables categóricas como numéricas y la meta es poder tener un modelo que ayude a describir el modelo y posteriormente evaluar qué tan bien lo hace y así poder responder la segunda pregunta planteada en el caso de las variables numéricas. Este proceso, sirve para evaluar con ayuda de las métricas del modelo, cómo y qué tanto describen las variables independientes a la objetivo.

En este caso, se pudo obtener un modelo con las variable escogidas, el cual tiene una r cuadrada ajustada del 0.76, lo cual es bastante aceptable para un modelo de regresión. Asimismo, el valor del MAPE es de 13.25 % lo que, según la tabla de interpretación, es un buen modelo para predecir el precio de los carros. No obstante, es importante observar que dos p-valores, específicamente de las variables carbody (0.78), engine type (0.8) y cylinder number (0.71) son mayores al alpha de 0.05 y por ende estas dos variables no son significativas para describir la variable "price".

Por otro lado, es importante resaltar que, al calcular los valores de VIF para las variables del modelo, se encuentra que no existe colinealidad entre ellas, pero un valor que alerta es el VIF de la constante del modelo. Esto puede indicar otros problemas aparte de la multicolinealidad entre variables. Para comprobar esto, se realizaron los cuatro supuestos de la regresión.

En primer lugar, la linealidad se cumple en base a la figura 2, donde se observa que las variables se relacionan de forma lineal con la variable objetivo. Por otro lado, es hizo una prueba de Breush Pagan para comprobar homocedasticidad en el modelo, en donde no se rechazó la hipótesis nula al no haber evidencia de heterocedasticidad. Finalmente, en la figura 3.3.2 se puede observar que los residuos se distribuyen aproximadamente como una normal, con la media muy cercana a cero, por lo que se podría decir que se cumple la normalidad de los errores; y la independencia puede cumplirse ya que no existe ninguna tendencia en el scatter plot de los residuos.



Por último, se puede observar en la figura 5 que a grandes rasgos existe una linea recta que significa que los valores predecidos son muy cercanos a los reales. Sin embargo, a pesar de que los supuestos fundamentales se cumplen, la r ajustada es decente para el tipo de modelo realizado, y que el modelo se ajusta de manera general, como tal este no es fiable debido al VIF alto en

la constante, el cual puede estar apareciendo porque una regresión lineal múltiple no es el mejor acercamiento para resolver este problema.

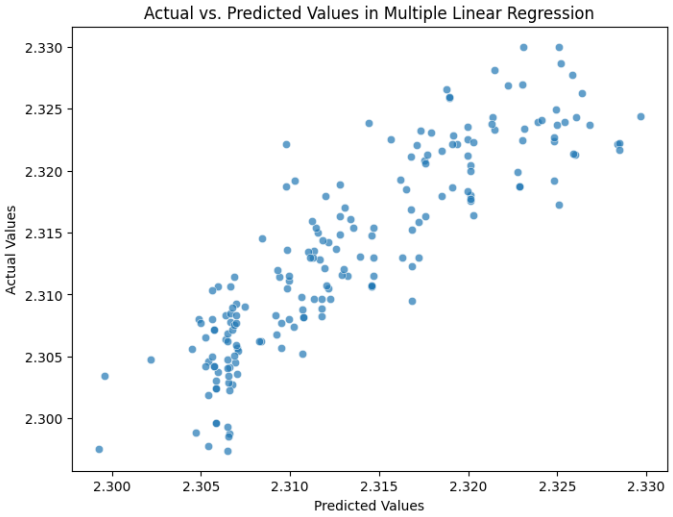


Figura 5: Predecidos vs. Reales

4. Conclusión

Después de este análisis descriptivo y predictivo de la base de datos sobre los precios de los carros, se puede llegar a varias conclusiones. Aunque el modelo se ve como uno muy bueno, y ciertas métricas muestran que no puede ser del todo fiable, igual se cumplió con el objetivo de responder las dos preguntas planteadas en un principio. Esto ya que, tanto el análisis ANOVA y los p-valores encontrados en el análisis de la regresión lineal múltiple, brindan la información suficiente para encontrar las variables que describen el precio de los vehículos.

En primer lugar, con el ANOVA se observó que las 4 variables independientes (Car body, Engine Type, Cylinder Number y Symboling) pueden describir una diferencia de precios según las categorías de cada una de ellas. Como se había mencionado en el análisis de resultados. A pesar de que todas las variables cumplen que el p-valor es menor al alpha, la magnitud de este representa inversamente la fuerza con la que estas variables describen al precio. Por lo que se tiene que Carbody explica a la variable objetivo en un nivel moderado, mientras que el resto de variables lo hacen fuertemente.

Ahora bien, hablando del modelo, es importante recalcar que carbody, engine type, y cylinder number salieron en el modelo como si no fueran significativas, pero esto se dio a que no fueron tratadas de la forma correcta, pues se codificaron pero no se transformaron a variables "dummies.º ficticias. Esto resultó en que en el modelo no puedan impactar de la forma correcta en la que debían. Por otra parte, si se analizan las variables numéricas del modelo, estas tienen p-valores bastante bajos, lo que significa que describen al precio de los autos de forma fuerte. Estos significa que las variables escogidas en un principio dan respuesta al objetivo de la empresa china de la situación problema, dando contexto suficiente sobre las características de los autos que marcan los precios en Estados Unidos.

Para cerrar, es importante mencionar que existen ciertos puntos a mejorar, como el trato de las variables categóricas para posteriormente poderlas aplicar a un modelo predictivo; así como que no hubo una experimentación exhausta con otras variables aparte de las escogidas, lo que puede implicar que alguna de las que no se escogieron, sea otra característica que pueda describir al precio significativamente. Así también, es importante reconocer que las diferentes pruebas que se realizaron de los modelos, ya sea con los datos con y sin normalizar, o también con o sin outliers, permitieron tener un mejor entendimiento del comportamiento del dataset así como de las herramientas estadísticas utilizadas.

5. Bibliografía

[1] Mena, M. (2023). *China siguió encabezando la producción mundial de vehículos en 2022.* Recuperado de: https://es.statista.com/grafico/29576/principales-paises-productores-del-sector-de-la-automocion-segun-el-numero-de-vehiculos-fabricados/

[2] Ruiz, B. (2023). *Momento de Retroalimentación: Módulo 1 Técnicas de procesamiento de datos para el análisis estadístico y para la construcción de modelos (Portafolio Análisis.* Recuperado de: https://experiencia21.tec.mx/courses/406127/assignments/13327365

6. Anexos

Liga a Drive con actividades del módulo:

https://drive.google.com/drive/folders/1C_nRM7p67AhBeAtcXKhd0OZGtRjDtqOK?usp=sharing