# Reporte final "El precio de los autos" Inteligencia Artificial Avanzada: Módulo de Estadística

Elías Garza Valdés

A01284041

*Monterrey, Nuevo León, México*

13 de septiembre de 2023

# Índice

**Resumen**

En este reporte se encuentra la solución estadística de un estudio de mercado. Se intenta conocer las principales variables o factores que se deben de considerar al momento de decidir el precio de un automóvil en el mercado estadounidense. Esto se hizo utilizando diferentes herramientas estadísticas como regresiones y análisis de factores. Se concluye que uno de los principales factores es el tamaño del coche, así como si este es de lujo o no.

## Código fuente

Este proyecto fue desarrollado principalmente en `python` y `R`. El código fuente puede encontrarse en el siguiente repositorio:

`https://github.com/EliasGarzaV/PortafolioImplementacionClaseIA`

# 1. Introducción

## 1.1. Descripción del problema

Una empresa automovilística china aspira a entrar en el mercado estadounidense. Desea establecer allí una unidad de fabricación y producir automóviles localmente para competir con sus contrapartes estadounidenses y europeas. Contrataron una empresa de consultoría de automóviles para identificar los principales factores de los que depende el precio de los automóviles, específicamente, en el mercado estadounidense, ya que pueden ser muy diferentes del mercado chino. Esencialmente, la empresa quiere saber:

- ¿Qué variables son significativas para predecir el precio de un automóvil?
- ¿Qué tan bien describen esas variables el precio de un automóvil?

Con base en varias encuestas de mercado, la consultora ha recopilado un gran conjunto de datos de diferentes tipos de automóviles en el mercado estadounidense los cuales se pueden encontrar en el repositorio .

## 1.2. Importancia del problema

Este tipo de análisis es bastante importante ya que entrar en un nuevo mercado implica inversiones fuertes y de hacerse de forma incorrecta puede impactar negativamente a la empresa.

Por ejemplo, existe una diferencia considerable en los patrones de consumo y culturales que implican la compra de un automóvil. Por ejemplo, Estados Unidos solo representa más de la mitad del mercado de pick-ups a nivel mundial a pesar de tener una población que representa tan solo el 3.75 %. 3. Asimismo, es posible que los estadounidenses consideren cosas distintas al momento de decidir el precio de un automóvil Por ejemplo puede que no estén interesados en un carro de un tamaño pequeño o que gaste mucha gasolina lo cual puede contrastar con otros mercados como el europeo en donde hay calles de menor tamaño.

# 2. Desarrollo

## 2.1. Limpieza y preparación de datos

Tenemos un total de 205 registros los cuales no son demasiados pero lo que puede llegar a complicar el problema y es que tenemos un total de 20 variables independientes de las cuales 13 son numéricas y 7 categóricas.

Algo importante que revisar es si existen valores atípicos en algunas de las variables y esto se hizo considerando como atípicos con el criterio del rango intercuartil. Si un valor esta 1.5 veces el rango intercuartil más abajo del percentil 25 o más arriba del 75 lo consideramos atípico. Esto se puede ver en un diagrama de caja y bigotes como el siguiente:

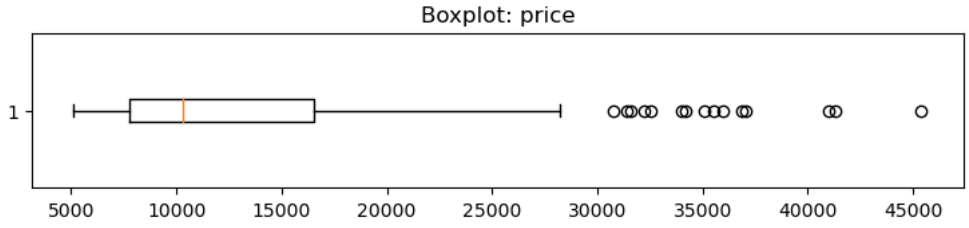



Figura 1: Diagrama para la variable de precio.

Los diagramas de cajas y bigotes de todas las variables se pueden encontrar en el código.

Es común retirar este tipo de datos dependiendo del modelo que queramos realizar pero en este caso no lo haremos ya que se encontró que todos los casos atípicos que hay son a la derecha de la mediana y es muy probable que correspondan a automóviles de lujo los cuales tienen motores más poderosos y también un precio muy elevado. No los quitaremos del modelo ya que este puede ser uno de los factores que más afecten al precio de un automóvil por lo que es algo que queremos analizar.

También, otra cosa a considerar es que se hicieron pruebas de normalidad de Anderson-Darling para todas las variables numéricas pero lamentablemente no fue posible establecer que alguna de ellas fuera normal.

## 2.2. Modelación

Primero se hizo un cálculo para revisar la correlación de las variables numéricas y el resultado es el siguiente:

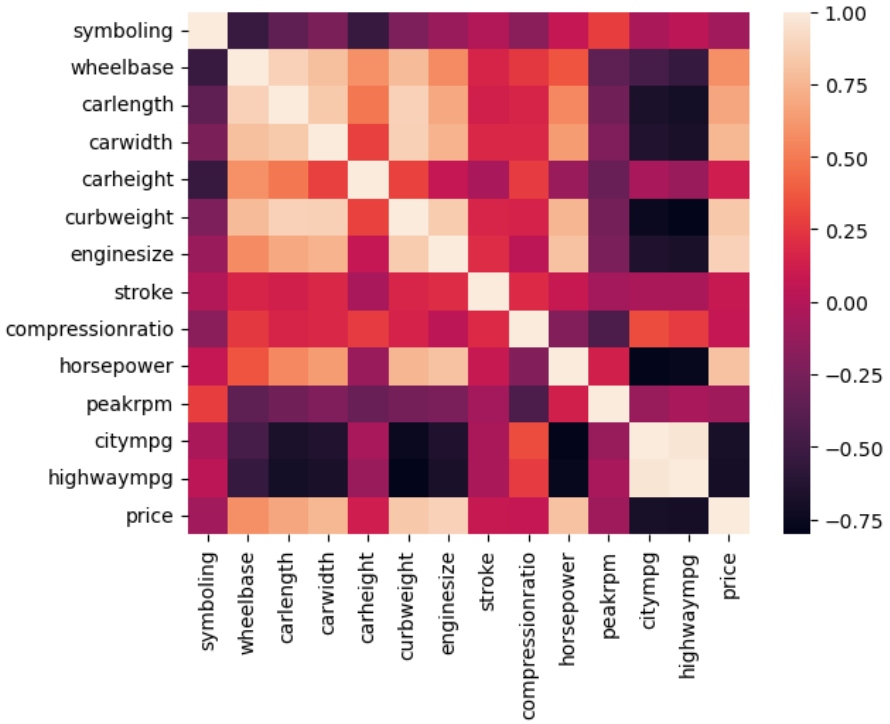



Figura 2: Matriz de correlación de variables

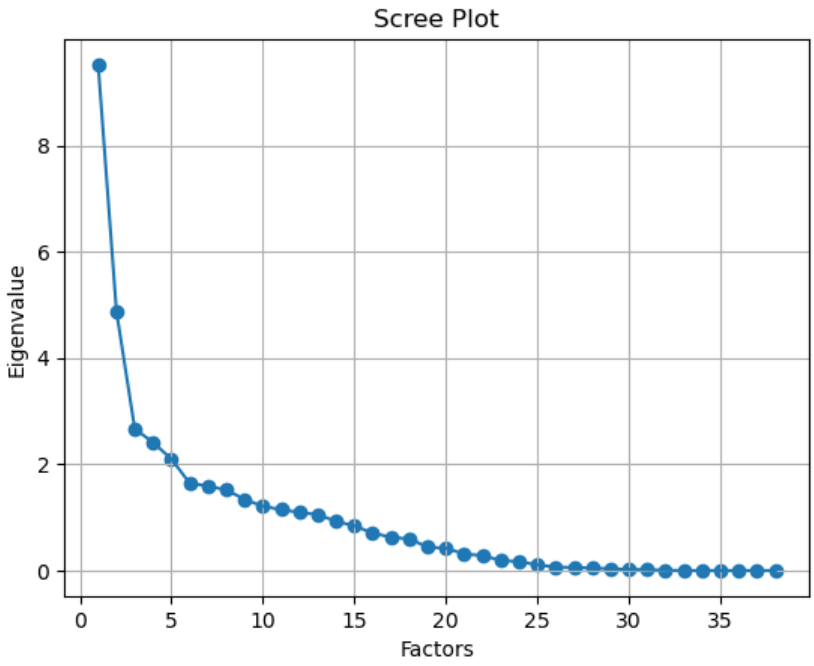
De aquí podemos ver que existen varias variables que tienen una alta correlación entre ellas lo cual no es ideal para generar un modelo porque tendríamos información redundante. Es por esto que sospechamos que nuestros datos son candidatos a realizar un analisis de factores.

### 2.2.1. Analisis de Factores

Antes que nada revisamos con la prueba de esfericidad de Barlett que nuestros datos sean viables para hacer este análisis pero afortunadamente nos da un valor-p de 0.0004 por lo que podemos continuar con este analisis.

A continuación mostramos el diagrama de scree representando la varianza que explican

nuestros vectores. De este tomamos la decisión de tomar 6 factores para hacer el análisis.




Posteriormente, construimos los factores utilizando el método de componentes principales (para evitar tener los supuestos de normalidad de máxima verosimilitud) los cuales se pueden ver en la siguiente gráfica.

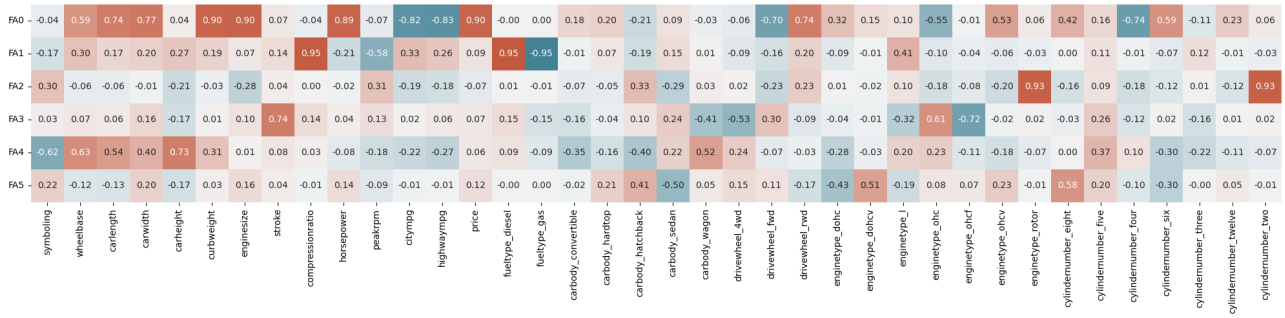



Figura 3: Factores obtenidos por el análisis

De aqui el factor que consideramos más importante es el primero y justo las variables importantes de este son:

- Car lenght
- Car width
- Car weight
- Engine size

- Horsepower
- City mpg *este es un factor negativo*
- Road mpg *también un factor negativo*
- Price

Esto nos indica que uno de los factores importantes tiene que ver con el tamaño y poder de los carros. Justo las camionetas grandes como las pick-ups tienen motores grandes con bastantes caballos de fuerza y son poco eficientes en cuestión de gasolina (las variables negativas indican este rendimiento).

### 2.2.2. Regresión Lineal

Ahora que ya hicimos un análisis preliminar continuaremos con otra alternativa con la cual veremos de manera más particular cómo se involucra el precio con las otras variables.

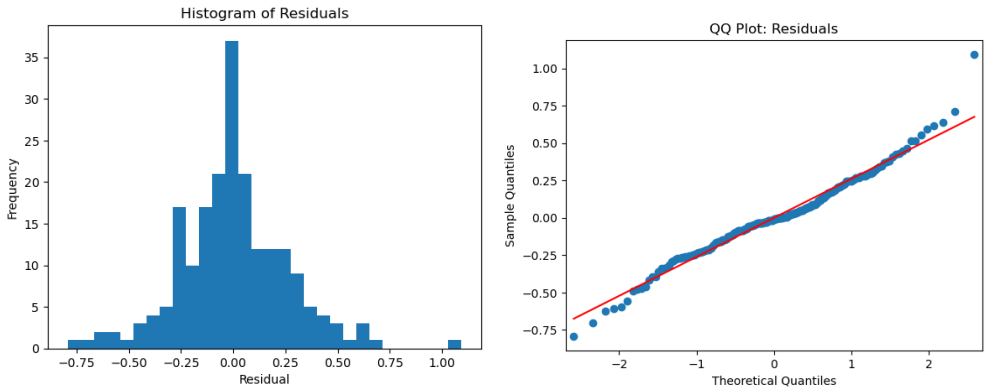
Antes que nada vamos a eliminar la escala de las variables estandarizando todas las columnas de nuestro dataframe. Posteriormente realizamos un modelo lineal con nuestras variables por mínimos cuadrados para obtener la relación entre estas variables y el precio y obtenemos los siguientes coeficientes:

| Vars | Coeficients |
|---:|---:|
| enginesize | 0.509522 |
| curbweight | 0.229140 |
| horsepower | 0.225637 |
| cylindernumber_eight | 0.213201 |
| highwaympg | 0.194691 |
| carwidth | 0.163136 |
| enginetype_ohc | 0.140213 |
| peakrpm | 0.137938 |
| fueltype_diesel | 0.114427 |
| cylindernumber_six | 0.083914 |

Figura 4: Los primeros coeficientes del modelo (están ordenados por tamaño)

Esto confirma nuestra hipótesis de que el tamaño del carro tiene un gran efecto en el precio ya que las variables que más aportan a este modelo son justo las que observamos en el análisis factorial.

Ahora vamos a validar el modelo observando los residuos los cuales deberían seguir una distribución normal.




(a) Histograma de Residuos (b) QQ-plot residual

Figura 5: Gráficas de análisis residual

Podemos observar que se parecen bastantes a una distribución normal con algunos problemas en las colas.

Sin embargo, al hacer el diagrama para revisar la homoestacisidad descubrimos cierto error en los residuos conforme crece la predicción.

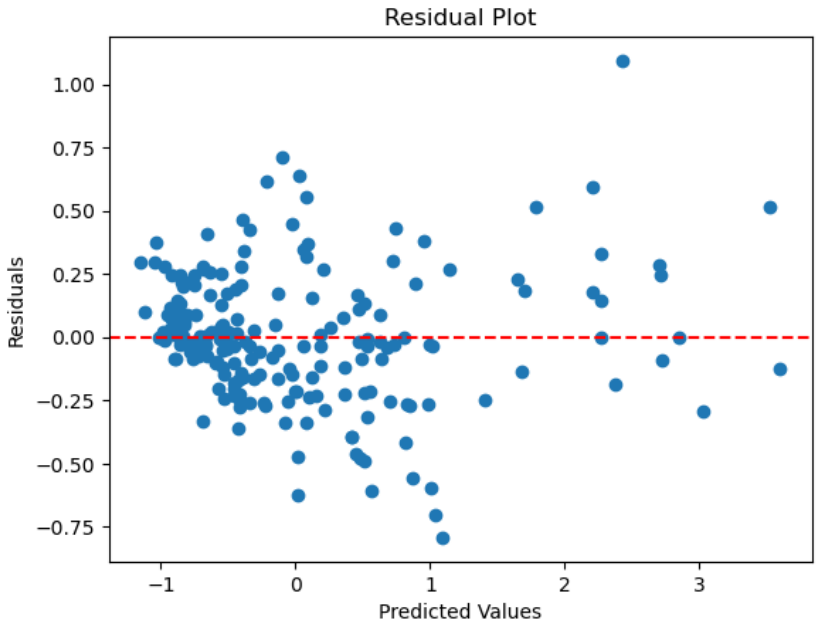



Figura 6: Diagrama de residuos

# 3. Conclusiones

Concluimos este reporte con el conocimiento de que las variables que más afectan el precio de un automóvil en el mercado estadounidense son las que tienen relación con el tamaño y poder

del carro como los caballos de fuerza y el peso. Esto es importante ya que es altamente probable que en China esto no ocurra debido al tamaño de las calles y la estructura de las ciudades por lo que es algo que la empresa debe considerar.

Llegamos a esta conclusión de dos formas distintas. Primero haciendo un análisis de factores en donde descubrimos que el factor principal de los datos es el que involucra esta idea del tamaño del carro y motor.

Posteriormente, generamos un modelo lineal con las variables estandarizadas y observamos que las mismas variables que habíamos observado antes son las que tenían un peso más grande al momento de predecir el precio del auto. Asimismo, validamos nuestro modelo observando el análisis de residuos. El único problema con el que nos quedamos fue un poco de heteroestasicidad la cual requeriría realizar un modelo más complejo para resolverla.

# Referencias

Statista. (2023). *Pickup Trucks - Worldwide | Statista Market Forecast*. Recuperado el 12 de septiembre de 2023, de https://www.statista.com/outlook/mmo/passenger-cars/pickup-trucks/worldwide#unit-sales