Instituto Tecnológico y de Estudios Superiores de Monterrey

Reporte final, "El precio de los autos"

Francisco Castorena Salazar, A00827756

Inteligencia artificial avanzada para la ciencia de datos I (Gpo 101)

TC3006C.101

Dra. Blanca R. Ruiz Hernández

10 de Septiembre del 2023

Resumen

Este informe se enfoca en identificar los principales factores que afectan los precios de los automóviles en Estados Unidos. Utilizamos métodos estadísticos y técnicas como Box Cox, análisis exploratorio, correlaciones y modelos de regresión para seleccionar las variables clave. Posteriormente, aplicamos pruebas y análisis para validar nuestras hipótesis. Los resultados ayudan a comprender mejor la influencia de las variables en los precios de los automóviles en el mercado estadounidense para así llegar a las conclusiones que se mostrarán al final del documento.

1. Introducción

1.1. Sobre la problemática

La problemática a resolver se define como sigue:

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 contra partes 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 que se presenta como la base de datos que se compartió a los miembros de esta clase, así como un diccionario de términos que describe brevemente cada variable, que es y el tipo de datos que contiene (categóricos, numéricos continuos). Para consultar la base de datos y el diccionario de términos se puede acceder dado clic en el siguiente enlace: ENLACE.

1.2. Sobre la importancia de la problemática

El hacer análisis estadístico sobre estos datos es crítico para abordar con éxito esta problemática, al proporcionar una base cuantitativa sólida para la toma de decisiones estratégicas, se le permite a la empresa comprender y responder a las dinámicas del mercado de forma informada y efectiva. La aplicación de este tipo de análisis permite tener un punto de vista más amplio amplio, el cual se considera que es necesario en un contexto global y de constante cambio.

2. Análisis de los resultados

El análisis hecho sobre la base de datos se realizó utilizando R en un archivo de tipo R Markdown, si se desea consultar el notebook en formato pdf ó Rmd con el código se puede dar clic en el siguiente enlace: ENLACE.

2.1. Carga y limpieza de la base de datos

La base de datos sobre la cual se realizó el análisis tenía 205 instancias (sin contar headers) y 21 columnas, de las cuales 7 fueron del tipo categóricas y 13 de tipo numéricas tanto con únicamente números enteros como numéricas continuas, a continuación se muestran los nombres de las variables y su clasificación.

- Categóricas: symboling, fueltype, carbody, drivewheel. enginelocation, enginetype, cylindernumber.
- Numéricas: wheelbase, carlength, carwidth, carheight, curbweight, enginesize, stroke, compressionratio, horsepower, peakrpm, citympg, highwaympg, price.

Se buscó por valores nulos en la base de datos por medio de la función $colSums(is.na(df))$, no se encontró ningun valor nulo en la base de datos.

2.2. Análisis de distribución de datos para variables numéricas

Para analizar si las variables numéricas se comportaban acorde a la distribución normal, primeramente se realizaron histogramas sobre todas estas variables. Como ejemplo de los histogramas arrojados en el notebook véase la figura 1.

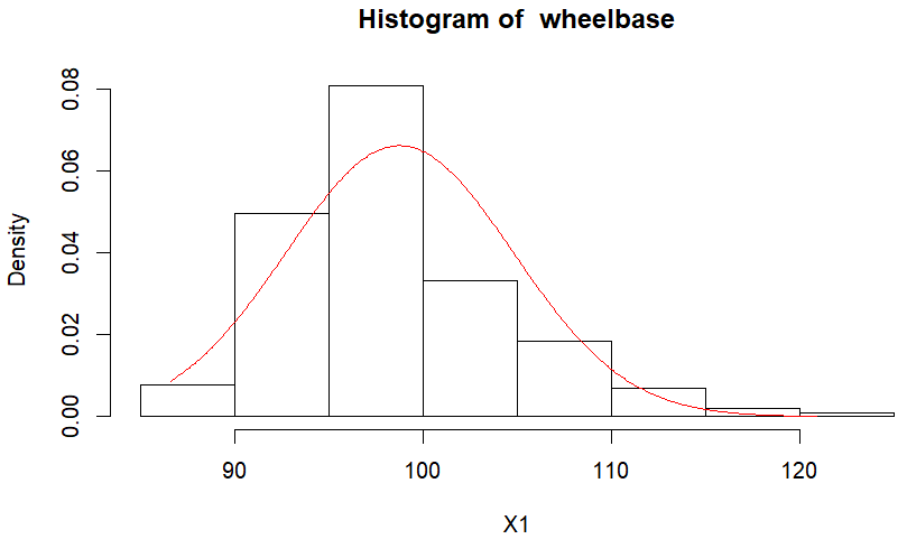

Figura 1: Histograma de la variable wheelbase con su función de densidad de probabilidad normal correspondiente a su media y desviación estándar.

Si se desea consultar un histograma en específico, véase el notebook en formato pdf compartido en en enlace anterior.

Después de ver los histogramas de las variables numéricas se concluyo que todas "encajaban" de forma correcta sobre una distribución normal para ser consideradas como normales, sin embargo había presencia de distintos grados de sesgo para ciertas variables, como la variable price (véase figura 2) que es de gran

importancia debido a que buscamos determinar como se ve afectada por otro tipo de variables que podrían tener otro tipo de distribución.

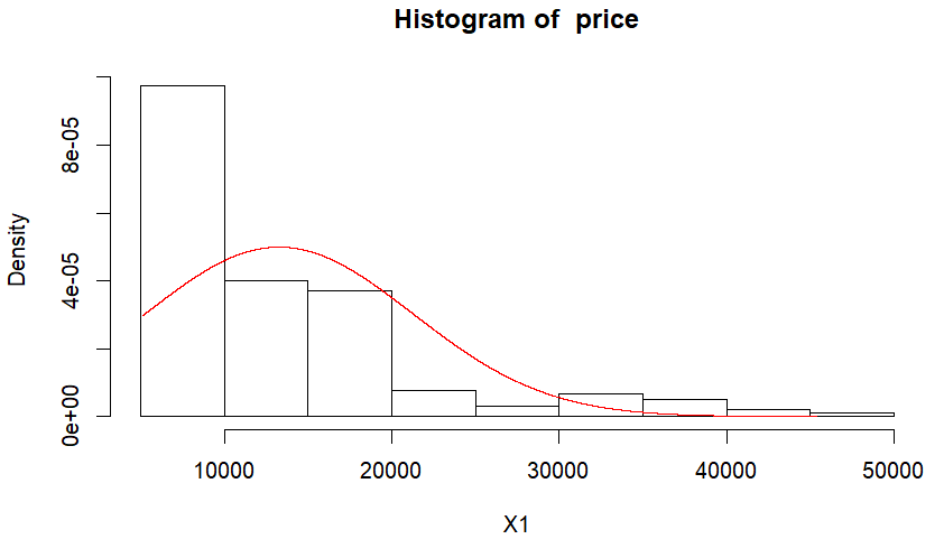

Figura 2: Histograma de la variable price con su función de densidad de probabilidad normal correspondiente a su media y desviación estándar.

Por otra parte se analizaron medidas estadísticas sobre las variables cuantitativas como lo son la media, cuantiles, mediana y moda. A excepción de las variables price, horsepower y compressionratio que son las variables que presentan mayor sesgo en sus datos originales, se pudo observar que las distancias entre mínimos y máximos de la mediana eran hasta cierto punto simétricas, lo cual pueden indicar un sesgo bajo, por otra parte la media y la mediana en estas variables eran muy parecidas, todos estos factores mencionados pueden ser indicios de un comportamiento acorde a la distribución normal.

2.2.1. Observaciones sobre diagramas de caja de variables cuantitativas

Para cada una de las variables cuantitativas se realizó su respectivo diagrama de caja, para observar la distribución de datos en cuartiles e identificar valores atípicos. Véase figura 3 como ejemplo de los diagramas de caja obtenidos.

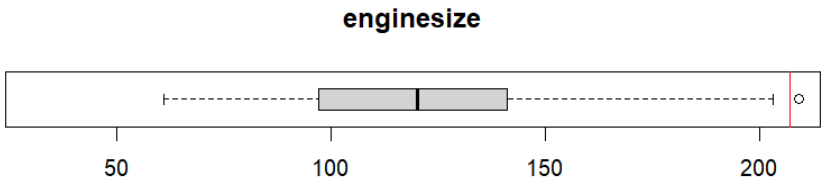

Figura 3: Diagrama de caja de la variable enginesize.

En los boxplots se puede observar como los valores están mayormente agrupados entre el cuartil 1 y el cuartil 3, lo cual es un indicio de normalidad ya que esto nos dice que los datos suelen estar agrupados alrededor de la mediana.

Para identificar valores extremos se definía un valor el cual era 1.5 veces la distancia entre el cuartil 1 y el cuartil 3 y se sumaba al cuartil 3, si el valor de algún dato era mayor a esta, se consideraba como valor extremo. Cabe mencionar que se tuvieron muy pocos datos atípicos, se tenía 1 o ningún dato atípico en todas las variables cuantitativas, lo cual es bueno para poder implementar modelos lineales, ya que la recta se ajusta mejor sobre las zonas donde hay mayor densidad de datos.

En primer momento no se eliminan los datos atípicos, se analizará el comportamiento de las variables y las pruebas de significancia con todos los datos, sin embargo en partes posteriores del documento se explica que si serán tratados para que encajen mejor a una distribución normal.

2.3. Análisis de medidas estadísticas sobre variables cualitativas (cuantiles y frecuencias)

Se analizó la frecuencia y distribución de datos de las variables cuantitativas, se pudo observar que las frecuencias de aparición de los datos respecto al total no era balanceado entre clases, como se puede observar en la figura 5, todas las variables tenían alguna clase que predominaba sobre las demás, por otra parte las variables cuantitativas tenían entre 2 y 7 valores por columna, lo cual considero que es un buen número para hacer análisis sobre grupos, no se tienen demasiados grupos por variable, lo cual hace que no se pierda valor por grupo y estos puedan ser diferenciados más fácilmente por sus características.

2.3.1. Diagramas de barras y de pastel

Se realizaron diagramas de barras y pie charts para analizar de forma visual la distribución de los datos por las frecuencias de sus distintos valores. Véase figura 4 y 5 como ejemplos de los tipos de gráficas que se obtuvieron en esta sección del análisis.

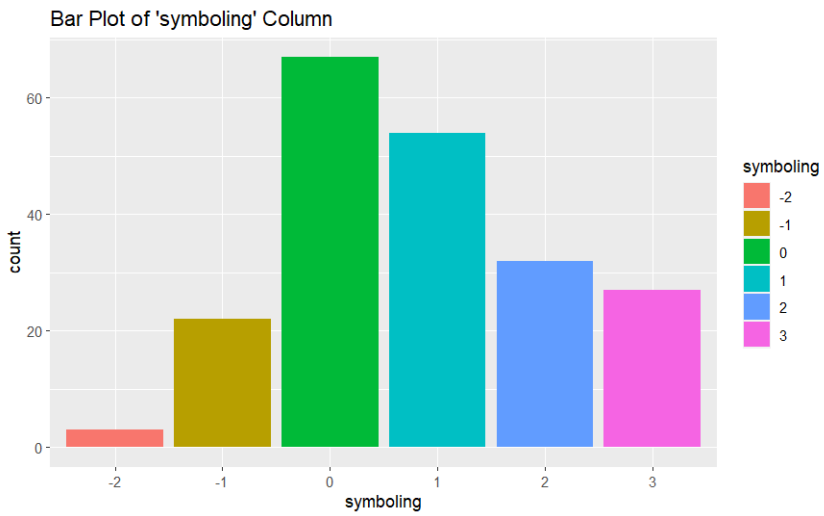

Figura 4: Distribución por tipo de datos, pordría considerarse en este caso como un histograma.

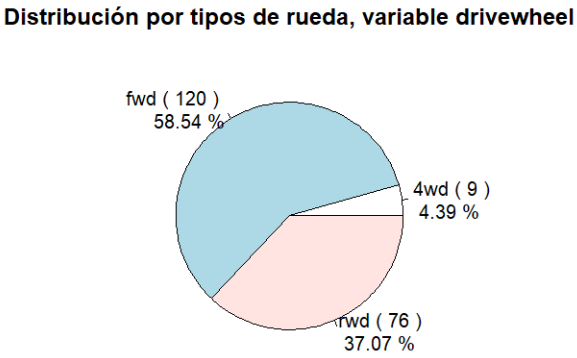

Figura 5: Diagrama de pastel de la variable drivewheel, se pueden observar los porcentajes de cada tipo de valor que tomaba la variable.

2.4. Análisis de correlación

Para identificar aquellas variables que estaban mayormente relacionadas con la variable price, así como identificar aquellas que estaban más relacionadas entre sí, se realizó una matriz de correlación de coeficientes de Pearson con las variables numéricas de la base de datos. Debido a que se obtuvo una matriz de correlación de dimensiones 13x13, era algo difícil identificar aquellas variables que solían representar correlaciones altas con una o varias variables, por lo que primeramente se realizó un heatmap, que sirviera como apoyo visual para esto. Véase figura 6.

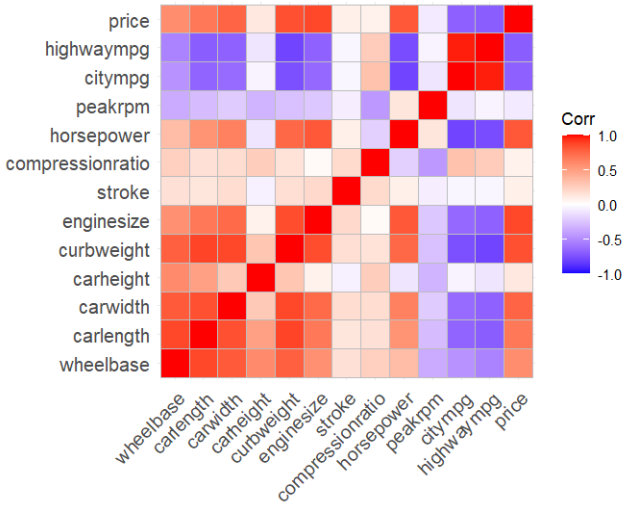

Figura 6: Heatmap de correlación entre variables numéricas del conjunto de datos, podemos ver que aquellos recuadros con colores rojos y azules intensos son aquellos con mayor correlación.

Para la selección de variables se tiene que tener en cuenta aquellas variables que tengan mayor correlación con la variable price, sin embargo tenemos que procurar que la correlación entre variables predictoras sea bajo, esto para evitar colinealidad, aún y bajo el supuesto de que dos variables tienen una correlación alta con la variable price, y también alta entre ellas, puede que la varianza explicada de la variable price por ambas variables en un modelo multilineal sea casi el mismo al que explican por separado, debido a que suelen comportarse de la misma manera ante el cambio de la variable price.

Por lo tanto, para identificar y seleccionar aquellas variables que pueden ser significativas sobre la variable price, se tomará en cuenta que tengan una alta correlación con la variable price procurando evitar colinealidad, la siguiente subsección del documento nos ayudará a identificar aquellas variables con alta colinealidad.

2.5. Análisis por Componentes Principales (PCA)