

Universidad Nacional De Colombia
Sede Medellín



Facultad de Ciencias

Departamento de Estadística

Proyecto - 1 Entrega

Daniel Felipe Villa Rengifo

Luis David Hernández Pérez

Juan Gabriel Carvajal Negrete

Modelos de Regresión

Diciembre, 2024

Contexto de los datos

Planteamiento de Problema

Una empresa automotriz china, Geely Auto, tiene la intención de ingresar al mercado de Estados Unidos estableciendo una planta de fabricación allí y produciendo vehículos localmente para competir con sus contrapartes estadounidenses y europeas.

Para ello, han contratado una consultora automotriz para comprender los factores que influyen en la fijación de precios de los vehículos. Específicamente, desean entender los factores que afectan el precio de los autos en el mercado estadounidense, ya que estos pueden ser muy diferentes a los del mercado chino. La empresa desea conocer:

- Qué variables son significativas para predecir el precio de un automóvil.
- Qué tan bien estas variables describen el precio de un automóvil.

A partir de diversas encuestas de mercado, la consultora ha recopilado un conjunto de datos <https://www.kaggle.com/datasets/hellbuoy/car-price-prediction/data> sobre diferentes tipos de vehículos en el mercado estadounidense.

Objetivo Empresarial

Se requiere modelar el precio de los automóviles utilizando las variables independientes disponibles. Este modelo será utilizado por la gerencia para comprender cómo varían exactamente los precios en función de las variables independientes. De esta manera, podrán ajustar el diseño de los vehículos, la estrategia empresarial, entre otros, para alcanzar ciertos niveles de precio. Además, el modelo será una herramienta útil para que la gerencia entienda la dinámica de precios de un nuevo mercado.

Descripción de las variables

- **Car_ID**: ID único de cada observación (entero).
- **Symboling**: Clasificación del riesgo de seguro; un valor de +3 indica que el automóvil tiene alto riesgo y un valor de -3 indica que probablemente es seguro (categórico).
- **fueltype**: Tipo de combustible del automóvil, por ejemplo, gasolina o diésel (categórico).
- **aspiration**: Tipo de aspiración utilizado en el automóvil (categórico).
- **doornumber**: Número de puertas del automóvil (categórico).
- **carbody**: Tipo de carrocería del automóvil (categórico).
- **drivewheel**: Tipo de tracción (ruedas motrices) del automóvil (categórico).
- **engine location**: Ubicación del motor del automóvil (categórico).
- **wheelbase**: Distancia entre los ejes del automóvil (numérico).
- **carlength**: Longitud del automóvil (numérico).

- `carwidth`: Ancho del automóvil (numérico).
- `carheight`: Altura del automóvil (numérico).
- `curbweight`: Peso del automóvil sin ocupantes ni equipaje (numérico).
- `enginetype`: Tipo de motor del automóvil (categórico).
- `cylindernumber`: Número de cilindros del motor (categórico).
- `enginesize`: Tamaño del motor del automóvil (numérico).
- `fuelsystem`: Sistema de combustible del automóvil (categórico).
- `boreratio`: Relación de diámetro del cilindro (numérico).
- `stroke`: Carrera o volumen dentro del motor (numérico).
- `compressionratio`: Relación de compresión del motor (numérico).
- `horsepower`: Potencia del motor en caballos de fuerza (numérico).
- `peakrpm`: Revoluciones máximas por minuto (RPM) del motor (numérico).
- `citympg`: Rendimiento de combustible en ciudad, medido en millas por galón (numérico).
- `highwaympg`: Rendimiento de combustible en carretera, medido en millas por galón (numérico).
- `price`: Precio del automóvil, considerado como la variable dependiente (numérico).

El conjunto de datos está formado por 205 registros y 26 variables, sin valores ausentes en las variables.

Limpieza de los datos

En la variable `car_name` podemos observar que los valores almacenan tanto el nombre de la empresa como el nombre del coche por lo cual hay 147 categorías distintas. Por tanto limpiaremos esa variable separando los nombres de las empresas de carros de la variable `car_name`. Por lo tanto crearemos una variable llamada `company_name` la cual tendrá solo el nombre de la empresa o compañía a la cual pertenece el carro.

Vemos que hay algunas categorías de la variable `company_name` están mal escritas como:

- `maxda` = mazda
- `Nissan` = nissan
- `porsche` = porcshe
- `toyota` = toyouta
- `volswagen` = volkswagen = vw

Reemplazaremos los nombres incorrectos con el nombre correcto de la empresa.

Análisis descriptivo de los datos.

Para el análisis exploratorio de datos, y dado el gran número de variables disponibles, se seleccionarán aquellas que, según la investigación previa, podrían ser relevantes para explicar

el comportamiento del precio. En este análisis se evaluará la correlación entre las variables numéricas, así como su relación con la variable precio. Para las variables categóricas, se analizará su interacción con el precio, buscando patrones o asociaciones significativas, por tanto las variables que probablemente sean más importantes para predecir el precio de un automóvil:

- **Dimensiones del vehículo:** Las variables `wheelbase`, `carlength`, `carwidth`, y `carheight` podrían estar correlacionados con el precio porque un automóvil más grande o más espacioso tiende a ser más caro.
- **Especificaciones del motor:** Variables como `enginesize`, `horsepower`, y `compressionratio` están directamente relacionados con el rendimiento del automóvil y podrían influir significativamente en el precio.
- **Peso:** La variable `curbweight` puede ser un buen indicador del tipo y tamaño del vehículo, y suele correlacionarse con el precio.
- **Eficiencia de combustible:** Las variables `citympg` y `highwaympg` podrían influir en el precio, ya que los automóviles más eficientes suelen tener precios diferentes según el segmento de mercado.

Las variables categóricas como `carbody`, `drivewheel`, `fueltype`, `enginetype` y `company_name` suelen ser indicadores del tipo de vehículo y su mercado objetivo.

Variable respuesta

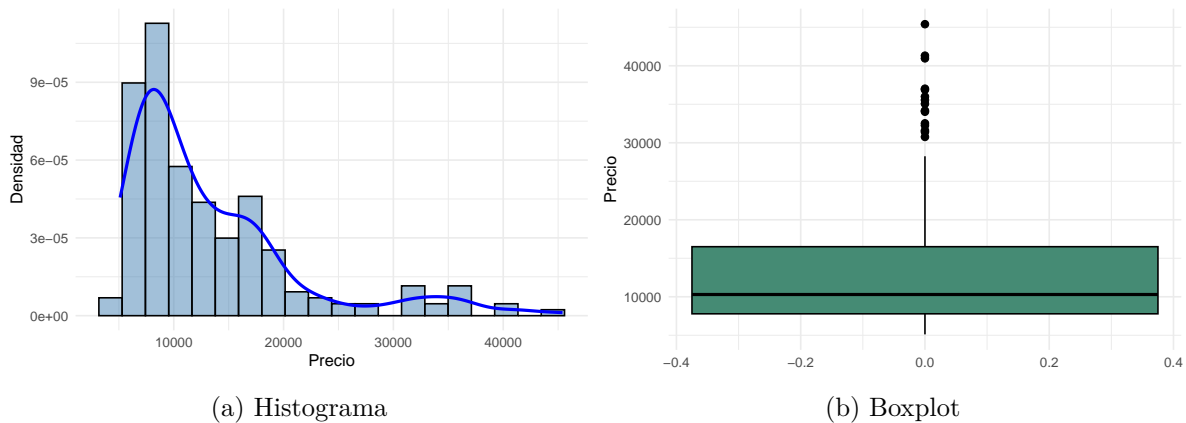


Figura 1: Distribución de la variable respuesta (Precio)

De la Figura 1 podemos decir que los precios de los carros tienen una distribución asimétrica, concentrándose en valores bajos, pero con una minoría de carros significativamente más caros, además el precio de la mayoría de los carros está por debajo de 14,000. Existe una diferencia significativa entre el valor medio y la mediana y los valores atípicos en el rango superior deben

considerarse, ya que pueden representar carros de lujo o especiales por tanto analizaremos ahora la variable que corresponde al nombre de la empresa o compañía a la cual pertenece el carro para ver que relación existe con el precio.

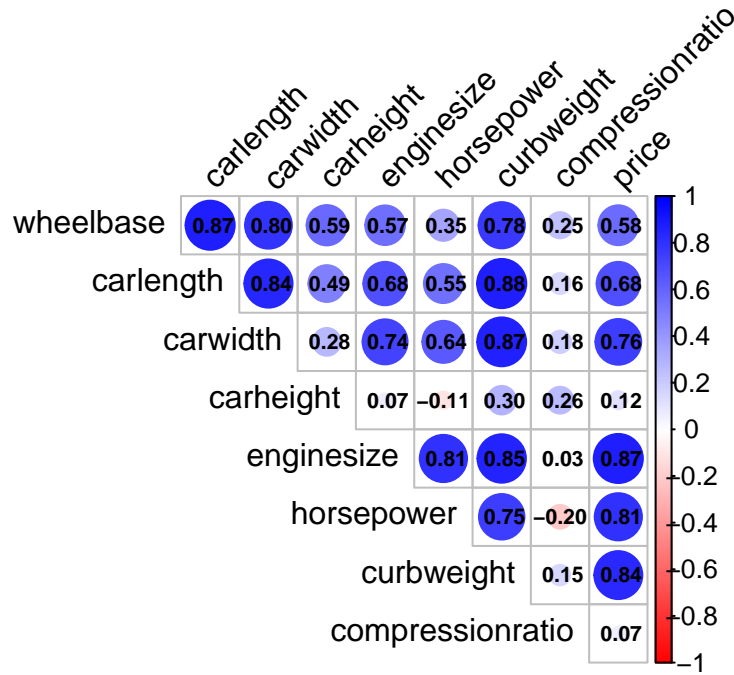


Figura 2: Grafico de Correlación

De la Figura 2 se pueden extraer las siguientes conclusiones:

- Las variables que tienen una mayor relacion lineal con el precio son : **enginesize** ($r = 0.87$), **curbweight** ($r = 0.84$) , **horsepower** ($r = 0.81$) y **carwidth** ($r = 0.76$).
- Hay pares de variables que tienen correlacion alta por lo que posiblemente no sea útil introducir algunas pares de variables en el modelo.

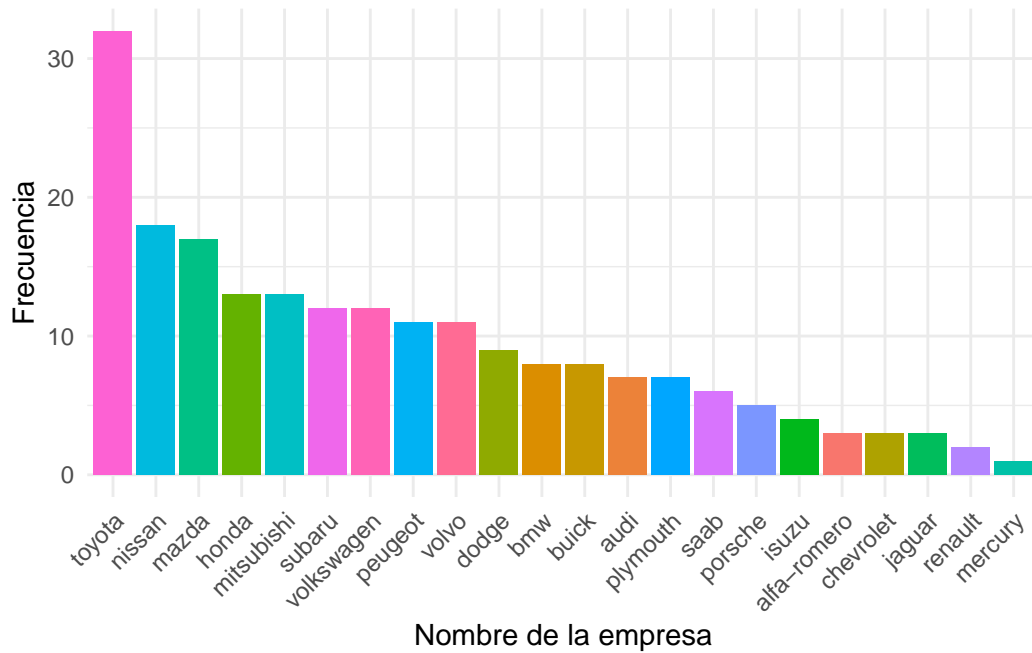


Figura 3: Distribución de los nombre de la empresa

De la Figura 3, observamos que la mayoría de los carros en este conjunto de datos está dominado por unas pocas marcas, especialmente Toyota, que supera ampliamente a las demás, por lo tanto, podemos decir que Toyota es la empresa preferida de los clientes.

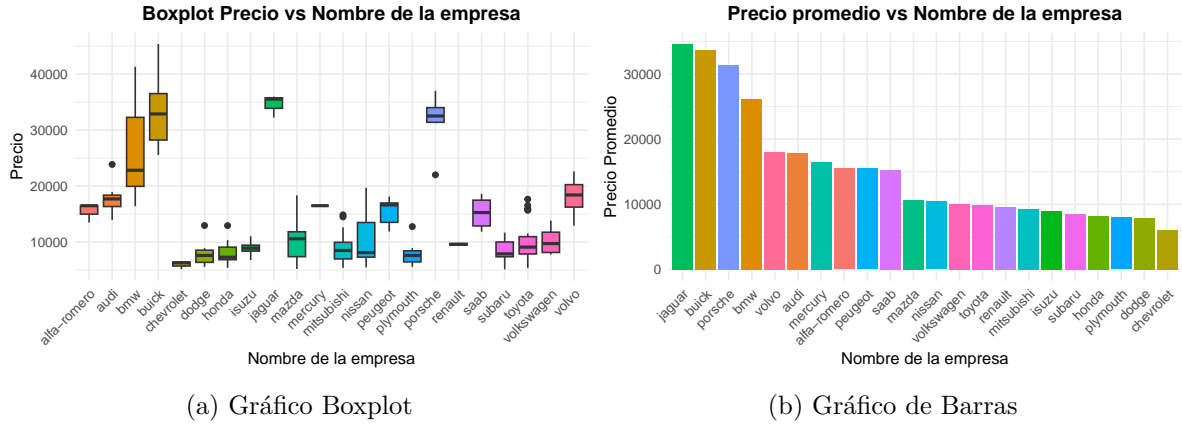


Figura 4: Distribución del Precio vs Nombre de la empresa

A partir de la Figura 4, observamos una diferencia significativa en el precio de los carros según la empresa a la que pertenecen. **Jaguar** y **Buick** parecen ofrecer los carros con las gamas de precios más altas.

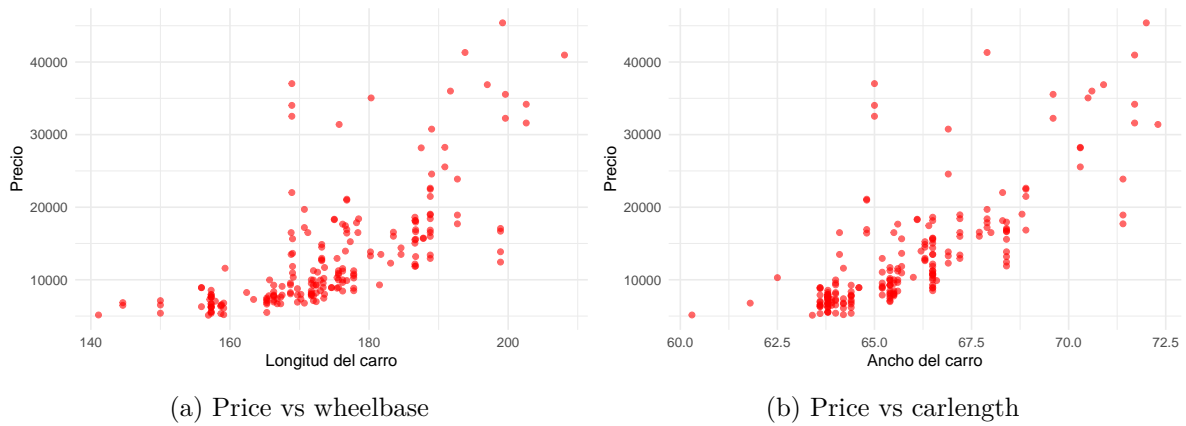


Figura 5: Gráficos de dispersión

De la Figura 5 se observa claramente que la longitud y la anchura del carro están estrechamente relacionadas con su precio. A medida que aumentan la longitud y la anchura del carro, también tiende a aumentar su precio. Sin embargo, no es posible hacer inferencias claras basadas únicamente en la relación entre la longitud y el precio, debido a la alta dispersión de los datos. La altura del carro no parece tener un impacto significativo en el precio.

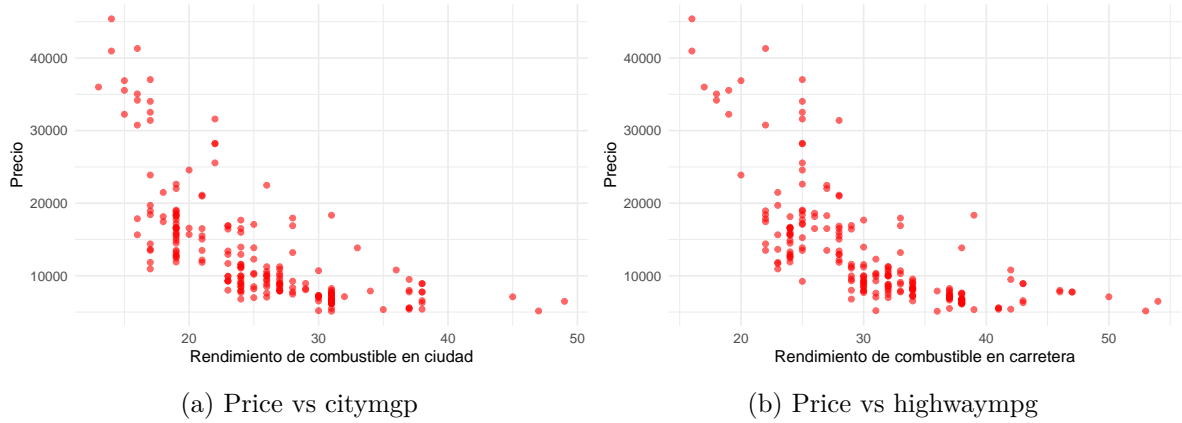


Figura 6: Gráficos de dispersión

De la Figura 6 se observa claramente que `citympg` y `highwaympg` tienen una correlación negativa con el precio del carro. A medida que aumentan los valores de `citympg` y `highwaympg`, el precio del carro tiende a disminuir. Dado que ambas características están relacionadas con el precio de manera significativa, `citympg` y `highwaympg` son características útiles para predecir el precio de los carros.

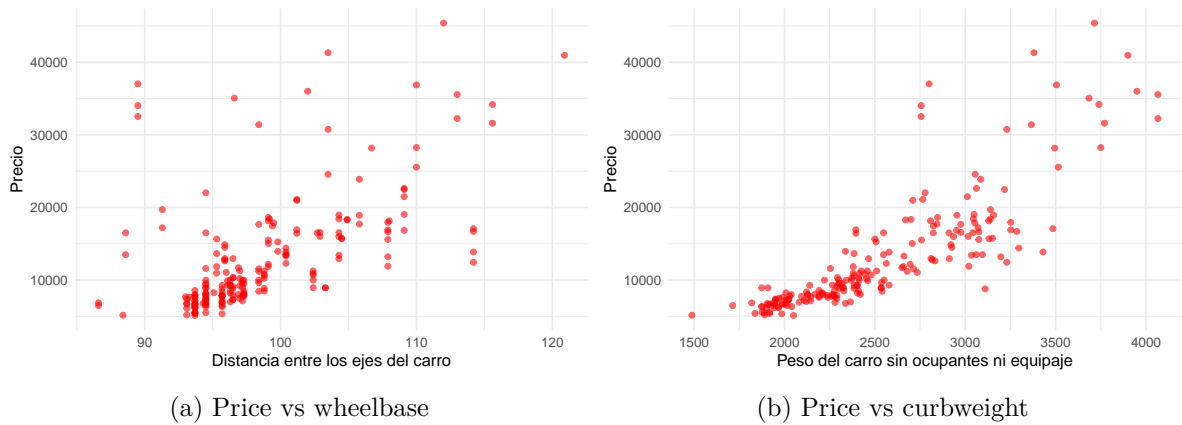


Figura 7: Gráficos de dispersión

De la Figura 7 se observa claramente que el peso del carro en vacío tiene una alta correlación (0.84) con el precio del carro. A medida que aumenta el peso en vacío, el precio del carro también incrementa de manera significativa. Aunque la distancia entre ejes y el precio no presentan una correlación tan alta, todavía existe una relación positiva. Por lo tanto, un aumento en la distancia entre ejes también está asociado con un incremento en el precio del carro.

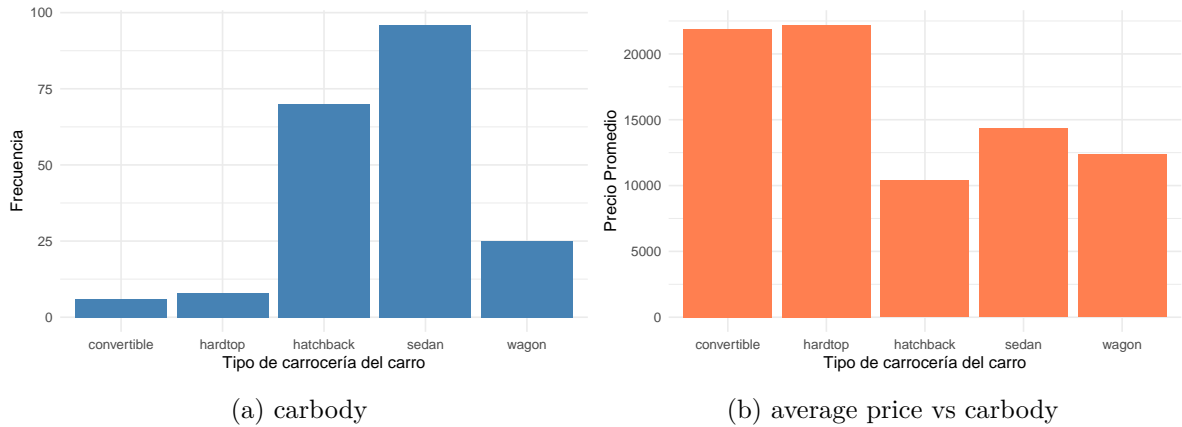


Figura 8: Gráfico de barras

De la Figura 8 se observa que los carros con carrocería sedán son los más vendidos, seguidos por los hatchback. Por otro lado, los descapotables y los de techo rígido tienen menores ventas. Estos últimos son también los más caros, seguidos de los descapotables. Es importante señalar que los descapotables y los carros con techo rígido se venden menos debido a su alto costo, lo que los hace menos atractivos para la mayoría de los clientes. Aunque la carrocería sedán ocupa el tercer lugar en términos de precio, sigue siendo la más popular, lo que sugiere que los clientes prefieren carros de gama media.

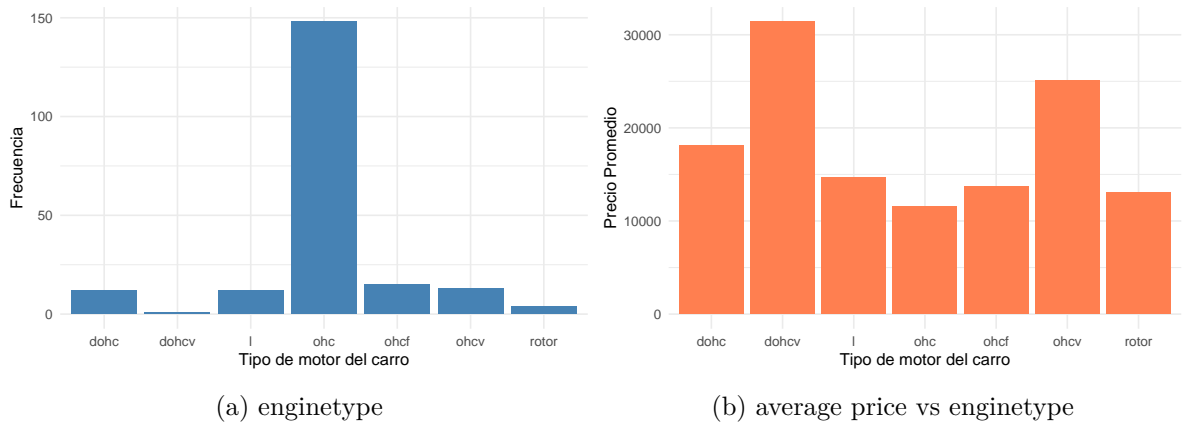


Figura 9: Gráfico de barras

De la Figura 9 observamos que la mayoría de los carros vendidos tienen motores de árbol de levas en cabeza (OHC). Solo se ha vendido un carro con motor DOHCV, y existen muy pocos datos disponibles para los motores DOHCV y de rotor. Los carros con motores DOHCV son, en su mayoría, más caros. Por otro lado, los carros con motores OHC son los menos costosos.

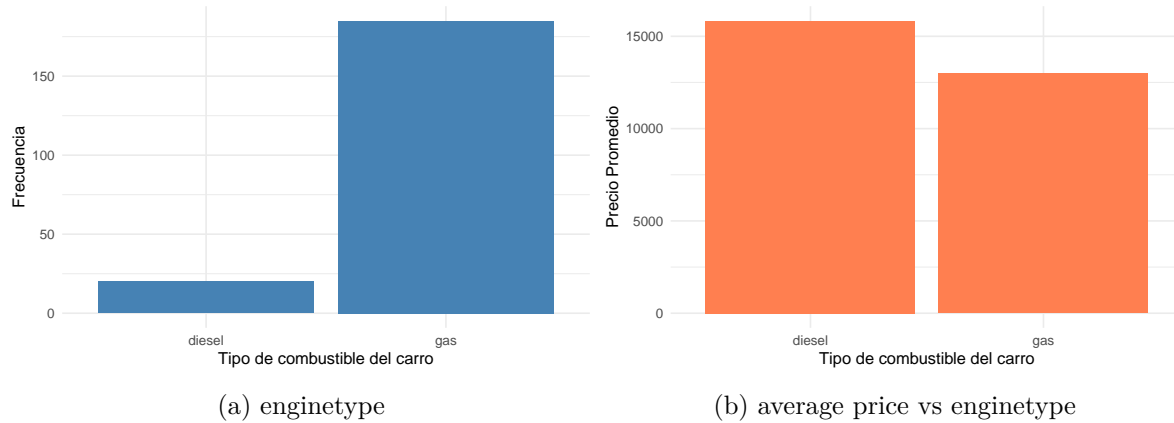


Figura 10: Gráfico de barras

De la Figura 10 podemos deducir que los carros con sistema de combustible a gas son los más preferidos, además se observa que el precio medio de los carros a gasolina es inferior al de los carros a diésel. Por lo tanto, podemos inferir que los clientes tienden a preferir carros que consumen menos combustible.

Conclusión

Después de realizar el análisis descriptivo, observamos que algunas variables parecen tener una mayor influencia en la explicación del precio de los automóviles, mientras que otras presentan una relación menos significativa. Además, identificamos que, en ciertas variables categóricas, algunas categorías tienen un número reducido de observaciones, lo que podría limitar su representatividad en los análisis posteriores. Por ello, será importante evaluar cómo estas características afectan la interpretación y robustez de los resultados.

Segunda Entrega

Punto 1

Ajuste un modelo de regresión lineal múltiple únicamente con las covariables continuas, muestre la tabla de parámetros ajustados y escriba la ecuación ajustada. Calcule la Anova del modelo. ¿Es significativo el modelo? ¿Qué proporción de la variabilidad total de la respuesta es explicada por el modelo? Opine sobre esto último.