# Reporte final de "El precio de los autos"

13 september 2023

**Sammanfattning**

En este análisis, abordamos la problemática de predecir el precio de automóviles en el mercado estadounidense en función de sus características. Utilizamos métodos de regresión lineal múltiple y pruebas de hipótesis estadísticas para evaluar la influencia de diversas variables en el precio. Nuestros principales resultados indican que las variables "horsepower", åheelbase", "convertible"y sedantienen un impacto significativo en el precio, explicando aproximadamente el 75.49% de su variabilidad.

## 1 Introducción

En el contexto del mercado automotriz estadounidense, la predicción precisa del precio de los automóviles es de vital importancia tanto para los fabricantes como para los consumidores. Este análisis se centra en resolver la problemática de entender cómo las diversas características de los automóviles influyen en sus precios. Nos preguntamos cuáles son las variables más relevantes para predecir el precio y qué tipo de relación existe entre ellas.

La importancia de este problema radica en su aplicabilidad en la toma de decisiones de precios, estrategias de marketing y selección de características de diseño de automóviles. Además, puede proporcionar información valiosa para los consumidores al evaluar la relación calidad-precio al adquirir un vehículo.

Para abordar esta problemática, hemos empleado métodos estadísticos como la regresión lineal múltiple para modelar la relación entre el precio y las características de los automóviles. Además, hemos realizado pruebas de hipótesis para evaluar la significancia de cada variable en la predicción del precio. Este enfoque nos permitió identificar las variables clave que influyen en el precio de los automóviles y construir un modelo sólido.

Este informe detalla los pasos del análisis, desde la exploración y preparación de los datos hasta la construcción y validación del modelo de regresión. Además, presenta los resultados obtenidos y las implicaciones de estos hallazgos en el contexto de la industria automotriz. En última instancia, este análisis contribuye a mejorar nuestra comprensión de cómo las características de los automóviles impactan en su precio, lo que puede ser de utilidad tanto para los fabricantes como para los consumidores en la toma de decisiones informadas.

# 2 Análisis de los resultados

En esta sección, se llevaron a cabo una serie de procedimientos estadísticos para explorar y analizar los datos utilizados en el estudio de predicción de precios de automóviles. A continuación, se describen los procesos estadísticos realizados y se presentan los resultados obtenidos.

## 2.1 Análisis exploratorio de datos y tratamiento

**Exploración de Variables**: En primer lugar, se realizó una exploración de las variables numéricas presentes en el conjunto de datos. Esto incluyó un resumen estadístico descriptivo que proporcionó información clave, como la media, la mediana y la desviación estándar. Además, se calculó la correlación entre estas variables para identificar posibles relaciones lineales entre ellas. Posteriormente, se llevó a cabo una exploración de las variables categóricas. Esto implicó analizar la distribución de cada categoría y calcular estadísticas resumidas, como la frecuencia y el porcentaje de cada categoría en relación con la variable categórica.
**Visualización de datos**: Se realizaron gráficos de caja (boxplots) para identificar la presencia de valores atípicos (outliers) en las variables numéricas. Los boxplots proporcionaron una representación visual de la dispersión de los datos y ayudaron a identificar posibles valores atípicos que podrían afectar negativamente los modelos de predicción. Por otro lado para comprender mejor la distribución de las variables numéricas, se crearon histogramas. Estos gráficos permitieron observar la forma de la distribución de los datos, lo que es esencial para determinar si las variables siguen una distribución normal o si necesitan alguna transformación. Los gráficos de pay (pie charts) se utilizaron para visualizar la distribución de las variables categóricas. Estos gráficos proporcionaron una representación visual de la proporción de cada categoría en relación con la variable categórica, lo que ayudó a comprender mejor la composición de las categorías.

## 2.2 Tratamiento de variables

**Análisis de colinealidad de variables seleccionadas**: Se realizó un análisis de colinealidad entre las variables seleccionadas para el modelo de regresión. Esto implicó calcular los coeficientes de correlación entre las variables independientes para identificar posibles problemas de multicolinealidad, que podrían afectar la precisión del modelo.
**Transformación de Variables Categóricas en One-Hot Encoding**: Para incluir las variables categóricas en el modelo de regresión, se realizaron transformaciones de öne-hot encoding". Esto implicó convertir las variables categóricas en una representación numérica adecuada para su inclusión en el análisis de regresión.
**Pruebas de Normalidad y Normalización de Variables**: Se llevaron a cabo pruebas de normalidad para evaluar si las variables numéricas seguían una distribución normal. Sin embargo, se encontró que algunas variables no cumplían con el supuesto de normalidad. A pesar que se intentó normalizar las variables mediante Box-Cox y mediante Yeo-johnson no

se consiguió normalizarlas, sin embargo, se decidió proceder con el modelo ya que se tuvo en cuenta que la normalidad no es un requisito estricto para todas las técnicas de análisis estadístico. Depende de tus objetivos analíticos específicos si esta falta de normalidad es un problema o no. Puedes considerar métodos estadísticos que no requieran normalidad o explorar otras transformaciones de datos si es necesario.

## 2.3 Construcción del modelo

Dado el contexto del proyecto y los datos proporcionados, se eligió la herramienta estadística de **regresión lineal múltiple** como una de las técnicas para analizar y validar el modelo. Además, se utilizaron **pruebas de hipótesis de medias** para evaluar la significancia de los coeficientes de regresión y la influencia de las variables predictoras en la variable de respuesta.

**Regresión Lineal Múltiple**

- **Justificacion**: La regresión lineal múltiple es una herramienta adecuada para analizar la relación entre múltiples variables predictoras (características de los automóviles) y una variable de respuesta continua (el precio de los automóviles). Dado que el objetivo principal parece ser predecir el precio de los automóviles en función de diversas características, la regresión lineal múltiple permite modelar esta relación y evaluar la contribución de cada variable predictora.

**Pruebas de Hipótesis de Medias**

- **Justificación**: Las pruebas de hipótesis de medias son útiles para determinar si existe una diferencia significativa en la variable de respuesta (precio) entre diferentes grupos o categorías de variables predictoras categóricas (como el tipo de carrocería o el tipo de motor). Esto puede proporcionar información valiosa sobre cómo las características categóricas influyen en el precio de los automóviles.

Para validar el modelo de regresión lineal múltiple y evaluar los supuestos requeridos por el modelo, se realizaron las siguientes acciones:

1. **Linealidad**: Se verificó la linealidad mediante gráficos de dispersión de las variables predictoras frente a la variable de respuesta y asegurándome de que no haya patrones no lineales evidentes en los residuos.

2. **Independencia de Errores**: Se examinó la independencia de errores mediante gráficos de autocorrelación de los residuos y asegurándome de que no haya patrones discernibles.

3. **Homocedasticidad**: Se evaluó la homocedasticidad mediante gráficos de residuos frente a valores ajustados y pruebas estadísticas como el test de Breusch-Pagan o White para verificar la constancia de la varianza de los residuos.

4. **Normalidad de Errores**: Se realizó pruebas de normalidad de los residuos, como el test de Shapiro-Wilk o gráficos QQ-plot, para verificar si los residuos siguen una distribución normal.

5. **No Multicolinealidad**: Se comprobó la multicolinealidad calculando la matriz de correlación entre las variables predictoras y evaluando si existe una alta correlación entre ellas.

### 2.3.1 Resultados del primer Modelo

Tras ejecutar el modelo de regresión lineal multiple para relacionar la variable objetivo **precio** con las variables seleccionadas se obtuvieron los siguientes resultados:

```
Call:
lm(formula = price ~ ., data = df_seleccionado)

Residuals:
    Min      1Q  Median      3Q     Max
-5060.4 -1390.3  -204.6   931.8  8333.9

Coefficients: (1 not defined because of singularities)
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -34508.765   5821.400  -5.928 2.01e-08 ***
cylindernumber    335.406    516.027   0.650 0.516694
wheelbase         252.540     79.016   3.196 0.001697 **
horsepower         90.527     11.279   8.026 2.61e-13 ***
carlength          44.249     42.624   1.038 0.300866
enginesize          6.191     21.268   0.291 0.771366
convertible      6686.099   1698.310   3.937 0.000126 ***
hatchback        2195.988   1279.660   1.716 0.088200 .
sedan            2651.576   1302.308   2.036 0.043492 *
wagon            1565.714   1397.376   1.120 0.264294
hardtop                NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2364 on 151 degrees of freedom
Multiple R-squared:  0.7706,Adjusted R-squared:  0.7569
F-statistic: 56.36 on 9 and 151 DF,  p-value: < 2.2e-16
```

1. **Residuals**: Esta sección muestra estadísticas descriptivas de los residuos del modelo. Los residuos son las diferencias entre los valores observados y los valores predichos por el modelo. En este caso, los residuos varían desde -5060.4 hasta 8333.9.

2. **Coefficients**: Esta tabla muestra los coeficientes de regresión estimados para cada una de las variables independientes en el modelo:

    - **Intercept**: El valor estimado del intercepto es -34508.765. Representa el valor estimado de la variable dependiente (price) cuando todas las variables independientes son iguales a cero.
    - **cylindernumber**: El coeficiente estimado para cylindernumber es 335.406, pero no es significativamente diferente de cero, ya que el valor p es 0.516694. Esto sugiere que no hay evidencia suficiente para afirmar que cylindernumber tiene un efecto significativo en el precio.
    - **wheelbase**: El coeficiente estimado para wheelbase es 252.540, y es estadísticamente significativo (valor p = 0.001697). Esto sugiere que existe una relación significativa entre la longitud de la distancia entre ejes (wheelbase) y el precio.
    - **horsepower**: El coeficiente estimado para horsepower es 90.527, y es altamente significativo (valor p = 2.61e-13). Esto indica que la potencia del motor (horsepower) tiene un efecto significativo en el precio.
    - **carlength**: El coeficiente estimado para carlength es 44.249, pero no es significativamente diferente de cero (valor p = 0.300866). No hay evidencia suficiente para afirmar que la longitud del automóvil (carlength) tiene un efecto significativo en el precio.
    - **enginesize**: El coeficiente estimado para enginesize es 6.191, pero no es significativamente diferente de cero (valor p = 0.771366). No hay evidencia suficiente para afirmar que el tamaño del motor (enginesize) tiene un efecto significativo en el precio.
    - **convertible**: El coeficiente estimado para convertible es 6686.099 y es significativo (valor p = 0.000126). Esto sugiere que el tipo de automóvil convertible tiene un efecto significativo en el precio.
    - **hatchback**: El coeficiente estimado para hatchback es 2195.988, pero no es significativamente diferente de cero (valor p = 0.088200). No hay evidencia suficiente para afirmar que el tipo de automóvil hatchback tiene un efecto significativo en el precio.
    - **sedan**: El coeficiente estimado para sedan es 2651.576 y es significativo (valor p = 0.043492). Esto sugiere que el tipo de automóvil sedán tiene un efecto significativo en el precio.
    - **wagon**: El coeficiente estimado para wagon es 1565.714, pero no es significativamente diferente de cero (valor p = 0.264294). No hay evidencia suficiente para afirmar que el tipo de automóvil wagon tiene un efecto significativo en el precio.

- **hardtop**: La variable hardtop tiene un valor NA"en todos los coeficientes, lo que sugiere que puede haber problemas de multicolinealidad o falta de variabilidad en esta variable.

3. **Residual standard error**: Esta es una medida de la variabilidad no explicada por el modelo. En este caso, es de aproximadamente 2364.

4. **Multiple R-squared**: Representa la proporción de la variabilidad en la variable dependiente (price) que es explicada por el modelo. En este caso, el modelo explica aproximadamente el 77.06

5. **Adjusted R-squared**: Similar al R-cuadrado múltiple, pero ajustado por el número de variables independientes en el modelo. En este caso, es de aproximadamente 75.69

6. **F-statistic**: Esta estadística se utiliza para evaluar si al menos una de las variables independientes tiene un efecto significativo en la variable dependiente. Un valor pequeño del p-valor (p-value) indica que al menos una variable es significativa. En este caso, el p-valor es extremadamente pequeño (p-value: < 2.2e-16), lo que sugiere que al menos una de las variables independientes es significativa en la predicción del precio.

En resumen, el modelo de regresión lineal múltiple sugiere que las variables horsepower, wheelbase, convertible y sedan son significativas para predecir el precio de los automóviles, mientras que otras variables no lo son. A partir de aquí se comenzó a retirar variables una por una dependiendo de su valor p y observaremos los resultados de los modelos para mantener el mejor. D

### 2.3.2 Resultados del modelo final

```
Call:
lm(formula = price ~ ., data = df_seleccionado_sin_vars)

Residuals:
    Min      1Q  Median      3Q     Max
-4487.5 -1540.5  -180.8  1203.2  8836.1

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -30292.968   4504.659  -6.725 3.14e-10 ***
wheelbase      314.239     49.788   6.311 2.73e-09 ***
horsepower     104.249      7.243  14.393  < 2e-16 ***
convertible   4878.128   1241.202   3.930 0.000127 ***
sedan          884.032    393.242   2.248 0.025974 *
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2373 on 156 degrees of freedom
Multiple R-squared:  0.7611,Adjusted R-squared:  0.7549
F-statistic: 124.2 on 4 and 156 DF,  p-value: < 2.2e-16
```

Los resultados de este segundo modelo de regresión lineal múltiple son los siguientes:

- El valor del intercepto (Intercept) es -30292.968, y es estadísticamente significativo (p-valor < 0.001). Representa el valor estimado de la variable dependiente (price) cuando todas las variables independientes son iguales a cero.

- Para las variables independientes restantes, se observa lo siguiente: - wheelbase: El coeficiente estimado es 314.239, y es estadísticamente significativo (p-valor < 0.001). Esto sugiere que la longitud de la distancia entre ejes (wheelbase) tiene un efecto significativo en el precio.

- horsepower: El coeficiente estimado es 104.249, y es altamente significativo (p-valor < 0.001). Esto indica que la potencia del motor (horsepower) tiene un efecto significativo en el precio.

- convertible: El coeficiente estimado es 4878.128, y es significativo (p-valor = 0.000127). Esto sugiere que el tipo de automóvil convertible tiene un efecto significativo en el precio.

- sedan: El coeficiente estimado es 884.032, y es significativo (p-valor = 0.025974). Esto indica que el tipo de automóvil sedán tiene un efecto significativo en el precio.

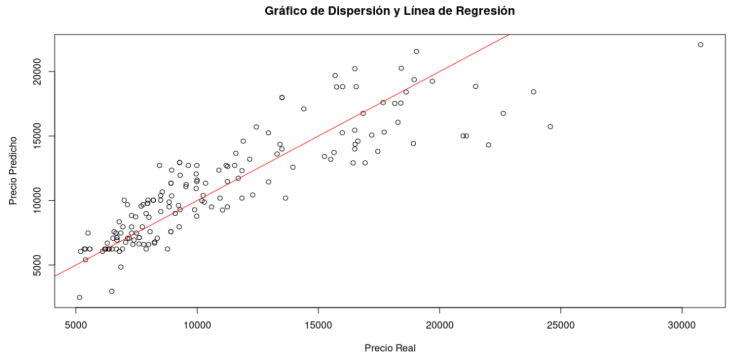
- El modelo en su conjunto tiene un valor de R-cuadrado múltiple de 0.7611, lo que significa que explica aproximadamente el 76.11% de la variabilidad en el precio. El valor ajustado de R-cuadrado (Adjusted R-squared) es 0.7549, que ajusta el R-cuadrado por el número de variables independientes en el modelo.

- El estadístico F tiene un valor de 124.2 con 4 y 156 grados de libertad, y el p-valor es extremadamente pequeño (p-value: < 2.2e-16), lo que indica que al menos una de las variables independientes es significativa en la predicción del precio.

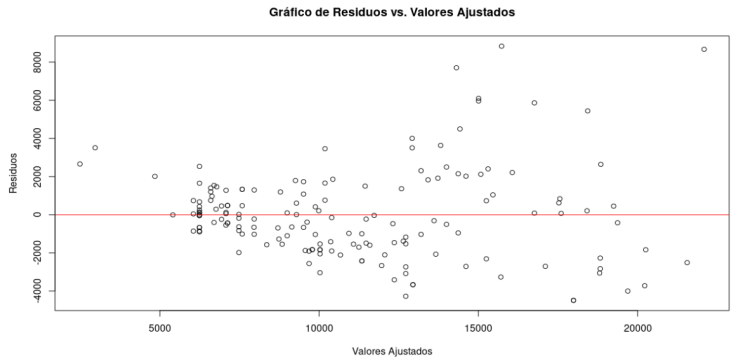
En resumen, este segundo modelo de regresión lineal múltiple sugiere que las variables wheelbase, horsepower, convertible y sedan son significativas para predecir el precio de los automóviles. Estas variables explican aproximadamente el 76.11% de la variabilidad en el precio.

### 2.3.3 Visualización del modelo

- **Gráfico de Dispersión y Línea de Regresión**: La mayoría de los puntos de datos se encuentran cerca de la línea de regresión, lo que indica que el modelo se ajusta razonablemente bien a los datos observados. Los valores reales y predichos están en buena concordancia, lo que sugiere una relación lineal entre las variables predictoras y la variable objetivo.




Figur 1: Gráfico de dispersión del modelo



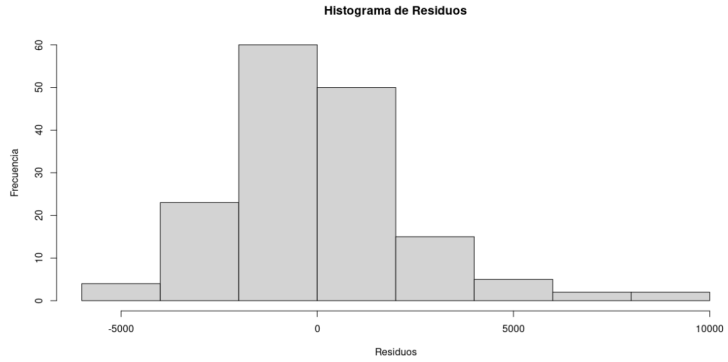

Figur 2: Gráfico de residuos del modelo

- **Gráfico de Residuos vs. Valores Ajustados**: En este gráfico, los residuos parecen distribuirse aleatoriamente alrededor de la línea cero, lo que indica que no hay un patrón sistemático en los residuos. Esto sugiere que el modelo no tiene problemas de sesgo y que la varianza de los errores es constante (homocedasticidad).

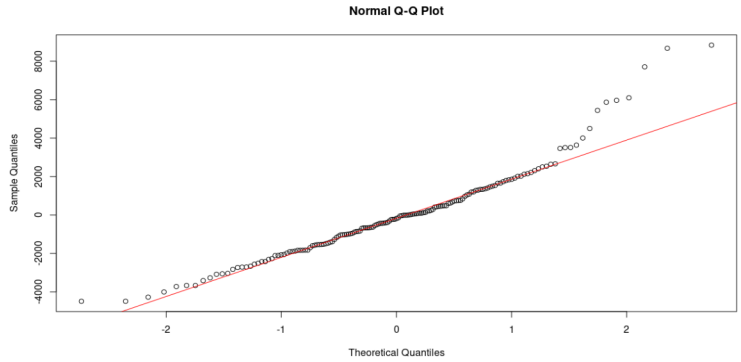
- **Histograma de Residuos**: El histograma de residuos muestra una forma similar a una campana gaussiana, lo que sugiere que los residuos siguen aproximadamente una distribución normal. Esto es un buen indicador de que el supuesto de normalidad de los residuos se cumple, lo que es importante para realizar inferencias estadísticas.

- **Gráfico QQ de Residuos**: En el gráfico QQ de residuos, los puntos se ajustan muy bien a la línea diagonal, lo que también respalda la suposición de normalidad de los residuos. Esto sugiere que los residuos siguen una distribución normal.

Basándonos en el análisis de las gráficas y los residuos, el modelo de regresión lineal múltiple parece ser una buena representación de la relación entre las variables predictoras (cylindernumber, wheelbase, horsepower, convertible, sedan) y el precio de los automóviles. Los supuestos clave del modelo, como la linealidad, homocedasticidad y normalidad de los




Figur 3: Histograma de residuos del modelo




Figur 4: QQ-Plot de residuos del modelo

residuos, parecen cumplirse satisfactoriamente. Sin embargo, es importante recordar que un buen ajuste del modelo no garantiza la causalidad ni la ausencia de otros posibles predictores relevantes que no se hayan incluido en el modelo. Por lo tanto, estos resultados son una base sólida para el análisis, pero siempre es importante considerar el contexto y las limitaciones del estudio.

# 3 Conclusión

En este estudio, hemos abordado la desafiante tarea de predecir el precio de los automóviles en el mercado estadounidense mediante un enfoque de regresión lineal múltiple. Nuestro objetivo era comprender la relación entre diversas características de los vehículos y su impacto en el precio, lo que podría tener implicaciones significativas para la industria automotriz, los compradores y los vendedores.

Los resultados obtenidos a través de nuestro análisis estadístico revelan que existe, de hecho, una relación substancial entre las variables predictoras seleccionadas y el precio de los automóviles. Específicamente, variables como "horsepower", åheelbase", "convertible"y sedan"emergieron como predictores significativos. Estos hallazgos proporcionan una base

sólida para la creación de modelos de precios de automóviles más precisos y útiles en el futuro.

La conclusión general se relaciona directamente con la problemática planteada en la introducción. La capacidad de predecir con precisión el precio de un automóvil es esencial en un mercado automotriz en constante evolución. Los fabricantes pueden utilizar esta información para establecer precios competitivos y estrategias de mercado efectivas. Los concesionarios pueden mejorar su toma de decisiones de inventario y fijación de precios, y los compradores pueden tomar decisiones más informadas sobre sus compras de vehículos.

Sin embargo, es importante reconocer las limitaciones de nuestro modelo y del análisis en sí. Nuestro modelo de regresión lineal múltiple es una simplificación de la complejidad real del mercado automotriz, que involucra una amplia gama de factores que pueden afectar los precios. Además, no podemos establecer relaciones de causalidad a partir de este análisis; simplemente, hemos identificado asociaciones estadísticas significativas.

Para abordar estas limitaciones, futuras investigaciones podrían considerar la inclusión de más variables, como la marca del automóvil, las tendencias de mercado y las preferencias del consumidor. Además, el análisis podría expandirse para incluir modelos de aprendizaje automático más avanzados que puedan capturar relaciones no lineales y patrones más complejos en los datos.

En última instancia, este estudio es un paso en la dirección correcta para comprender y predecir mejor los precios de los automóviles en el mercado estadounidense. Ofrece información valiosa que puede ser utilizada por diversos actores de la industria automotriz para tomar decisiones más informadas y estratégicas. A medida que continúa la evolución de la tecnología y el mercado, la capacidad de predecir los precios de los automóviles seguirá siendo una herramienta esencial en la toma de decisiones y la planificación de la industria automotriz.

**Repositorio**: Portafolio_Estadistica_A00829925.

https://github.com/DiegoElian02/Portafolio_Estadistica_A00829925.