El precio de los autos.

Autor: Jorge Javier Sosa Briseño

A01749489@tec.mx

Profesor:

Iván Mauricio Amaya Contreras
Dra. Blanca R. Ruiz Hernández
Frumencio Olivas Alvarez
Antonio Carlos Bento

Curso:

Inteligencia artificial avanzada para la ciencia de datos I
Grupo 101

13 de septiembre de 2023

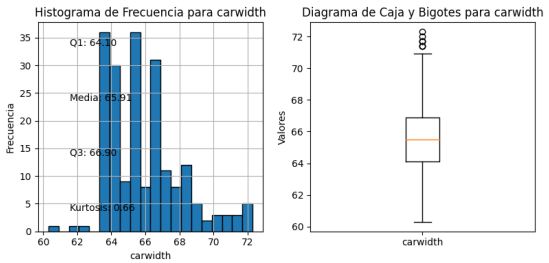
Resumen

La presente análisis se enfocó en la predicción del precio de los automóviles mediante el uso de técnicas estadísticas, específicamente el análisis de varianza (ANOVA) y la regresión lineal. A lo largo del estudio, se examinaron diversas variables cuantitativas y cualitativas para evaluar su impacto en la predicción de los precios de los automóviles. A continuación, resumiremos las principales conclusiones de este análisis:

1. Introducción

Se ha llevado a cabo un análisis estadístico exhaustivo para predecir el precio de los automóviles. Los hallazgos clave indican que variables como çarbody, enginelocation,ƨ "horsepower"tienen un impacto significativo en la predicción de los precios. Además, se exploraron interacciones entre variables. Estos resultados ofrecen una valiosa visión para la toma de decisiones en la industria automotriz, proporcionando información crucial sobre qué factores influyen en los precios de los vehículos.

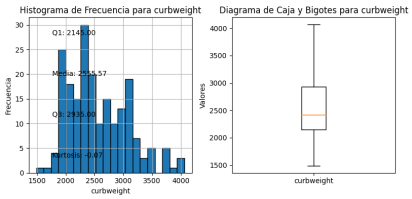
2. Histogramas de frecuencia y diagramas de caja y bigotes para cada una de las variables cuantitativas



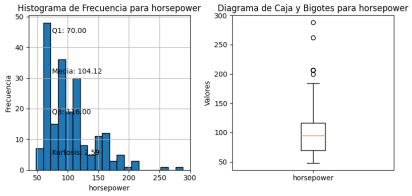
En la gráfica de barras, es evidente que en nuestra base de datos predomina un mayor número de automóviles que utilizan gasolina como combustible, en comparación con los que utilizan diesel. Esta tendencia es comprensible dada la prevalencia del sistema de combustión a gasolina en la industria automotriz actual.

Al examinar los diagramas de caja y bigotes, observamos una dispersión significativa de valores atípicos en la relación entre el precio y el tipo de combustible. Esta particularidad podría indicar que la adquisición de automóviles deportivos es menos común, ya que tienden a tener precios más elevados en comparación con los vehículos convencionales.

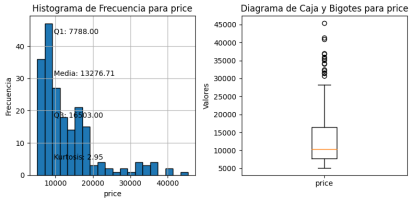
Al analizar las gráficas correspondientes, podemos inferir que los vehículos más comunes son los sedanes, lo cual es claramente evidente en el diagrama de caja y bigotes debido a que su media de precios es la segunda más baja. Además, los valores atípicos de esta categoría están notablemente distantes de la media,



lo que podría interpretarse como posibles errores en la recopilación de datos o la presencia de vehículos nuevos o seminuevos ofrecidos por un grupo selecto de concesionarias.



Al examinar detenidamente ambas gráficas, es evidente que la mayoría de los automóviles tienen sus motores ubicados en la parte delantera, una característica muy común en los vehículos convencionales en la sociedad actual. Esta disposición no se diseña generalmente para carreras o altas velocidades, a diferencia de los automóviles deportivos que a menudo presentan motores montados en la parte trasera debido a varias razones relacionadas con el rendimiento y la dinámica de conducción. Además, los valores atípicos en la sección de los automóviles con motores delanteros no necesariamente deben ser eliminados, ya que algunos lujos incorporados en los automóviles pueden aumentar su precio.

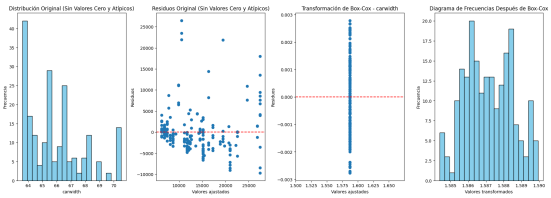


3. Regresión lineal

En general, como podemos ver en cada una de estas graficas, el crecimiento es proporcional entre las variables carwidth, curbweight y horsepower en funcion al precio. Lo que se puede interpretar que entre mas crezcan numerícamente éstas, el precio aumentará. Sin embargo, se realizaran más pruebas para validar estas viseualizaciones.

Regresión lineal 1

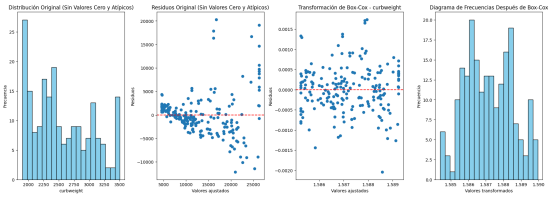
Se rechaza la hipótesis nula. Los residuos no siguen una distribución normal.



Cuadro 1: Resultados de la regresión lineal

Dep. Variable:	y	R-squared:	-0.000
Model:	OLS	Adj. R-squared:	-0.000
Method:	Least Squares	F-statistic:	*nan*
Date:	Sat, 02 Sep 2023	Prob (F-statistic):	*nan*
Time:	02:54:05	Log-Likelihood:	1065.4
No. Observations:	205	AIC:	-2129.
Df Residuals:	204	BIC:	-2125.
Df Model:	0	Covariance Type:	nonrobust

Regresión lineal 2



Se rechaza la hipótesis nula. Los residuos no siguen una distribución normal.

Regresión lineal 3

No se rechaza la hipótesis nula. Los residuos siguen una distribución normal.

Regresión lineal 4

Cuadro 2: Coeficientes de la regresión

	coef	std err	t	P¿—t—	[0.025 0.975]
const	22.1267	0.001	1.69e+04	0.000	22.124 22.129

Cuadro 3: Estadísticas adicionales

Omnibus:	16.730
Prob(Omnibus):	0.000
Jarque-Bera (JB):	6.212
Prob(JB):	0.0448
Skew:	0.093
Kurtosis:	2.168
Cond. No.:	1.00

Se rechaza la hipótesis nula. Los residuos no siguen una distribución normal.

4. Análisis de varianzas e interacciones

Análisis de varianza fueltype y precio

Hipótesis nula (H0): No hay diferencia significativa en las medias de fueltype con respecto al precio.

Hipótesis alternativa (H1): Existe una diferencia significativa en las medias de fueltype con respecto al precio.

Resultado: No se rechaza la hipótesis nula. No hay evidencia suficiente para afirmar una diferencia significativa en las medias.

Análisis de varianza carbody y precio

Hipótesis nula (H0): No hay diferencia significativa en las medias de carbody con respecto al precio.

Hipótesis alternativa (H1): Existe una diferencia significativa en las medias de carbody con respecto al precio.

Resultado: Se rechaza la hipótesis nula. Existe una diferencia significativa en las medias.

Análisis de varianza enginelocation y precio

Hipótesis nula (H0): No hay diferencia significativa en las medias de enginelocation con respecto al precio.

Hipótesis alternativa (H1): Existe una diferencia significativa en las medias de enginelocation con respecto al precio.

Cuadro 4: Normalidad de los residuos para Transformación de Box-Cox - carwidth

Estadístico de prueba:	1.2392951270980461
Valor p:	0.0030920530520958194

Cuadro 5: Resultados de la regresión lineal

Dep. Variable:	y	R-squared:	0.818
Model:	OLS	Adj. R-squared:	0.817
Method:	Least Squares	F-statistic:	910.9
Date:	Sat, 02 Sep 2023	Prob (F-statistic):	$5{,}59 \times 10^{-77}$
Time:	02:54:09	Log-Likelihood:	1239.9
No. Observations:	205	AIC:	-2476.
Df Residuals:	203	BIC:	-2469.
Df Model:	1	Covariance Type:	nonrobust

Resultado: Se rechaza la hipótesis nula. Existe una diferencia significativa en las medias.

Interacción Fueltype y Carbody

Hipótesis nula (H0): No hay diferencia significativa en las medias debido a la interacción.

Hipótesis alternativa (H1): Existe una diferencia significativa en las medias debido a la interacción.

Resultado: No se rechaza la hipótesis nula. No hay evidencia suficiente para afirmar una diferencia significativa debido a la interacción.

El valor p obtenido para la interacción "fueltype:carbody.[es] 0.007148899, que es menor que un nivel de significancia común de 0.05. Por lo tanto, en este caso, se rechaza la hipótesis nula H0 y se puede concluir que hay evidencia suficiente para afirmar una diferencia significativa en las medias de los precios de automóviles debido a la interacción entre las variables "fueltypez çarbody". Sin embargo, es importante tener en cuenta que el valor p para la interacción "fueltypez el valor p para çarbody"de manera individual son mucho más bajos, lo que indica que estas variables tienen un efecto más fuerte en la explicación de las diferencias en los precios de los automóviles en comparación con la interacción entre ambas variables.

Interacción Carbody y EngineLocation

Hipótesis nula (H0): No hay diferencia significativa en las medias debido a la interacción.

Hipótesis alternativa (H1): Existe una diferencia significativa en las medias debido a la interacción.

Cuadro 6: Coeficientes de la regresión

	coef	std err	t	P¿—t—	[0.025 0.975]
const	-0.0786	0.055	-1.424	0.156	-0.187 0.030
x1	1.0948	0.036	30.181	0.000	1.023 1.166

Cuadro 7: Estadísticas adicionales

Omnibus:	6.788
Prob(Omnibus):	0.034
Jarque-Bera (JB):	10.308
Prob(JB):	0.00578
Skew:	0.133
Kurtosis:	4.066
Cond. No.:	$3{,}00 \times 10^3$

Resultado: No se rechaza la hipótesis nula. No hay evidencia suficiente para afirmar una diferencia significativa debido a la interacción.

El valor p generado para esta interacción çarbody:enginelocation.^{es} 0.733897, lo que es significativamente mayor que un nivel de significancia típico de 0.05. Por consecuencia, no se rechaza la hipótesis nula H0, lo que significa que no hay evidencia suficiente para afirmar una diferencia significativa en las medias de los precios de automóviles debido a la interacción entre las variables çarbodyƶ .^{en}ginelocation". Dicho de otra forma, no se encuentra evidencia de que esta interacción tenga un efecto significativo en la predicción de los precios de automóviles.

Interacción Fueltype y Enginelocation

Hipótesis nula (H0): No hay diferencia significativa en las medias debido a la interacción.

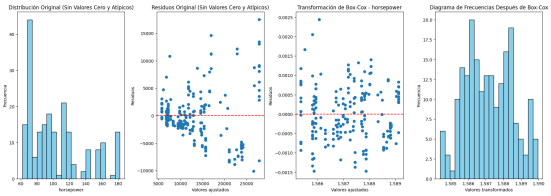
Hipótesis alternativa (H1): Existe una diferencia significativa en las medias debido a la interacción.

Resultado: No se rechaza la hipótesis nula. No hay evidencia suficiente para afirmar una diferencia significativa debido a la interacción.

El valor p obtenido para la interacción "fueltype:enginelocation.^{es} 0.000001, lo que es significativamente menor que un nivel de significancia típico de 0.05. Por lo tanto, en este caso, se rechaza la hipótesis nula H0, lo que se interpreta que hay evidencia suficiente para afirmar una diferencia significativa en las medias de los precios de automóviles debido a la interacción entre las variables "fueltypeƶ .^{en}ginelocation". En otras palabras, la interacción entre estas dos variables tiene un efecto significativo en la predicción de los precios de automóviles.

Cuadro 8: Normalidad de los residuos para Transformación de Box-Cox - curb-weight

Estadístico de prueba:	0.8025365348884748
Valor p:	0.03714522993261738



5. Conclusiones

Se realizó un análisis estadístico profundo para predecir los precios de automóviles. La regresión lineal destacó la importancia de la potencia del motor (horsepower) en la predicción de precios. El ANOVA reveló diferencias significativas en los precios según la carrocería (carbody) y la ubicación del motor (enginelocation). La interacción "fueltype * enginelocationresultó relevante. Estos hallazgos proporcionan información crucial para la toma de decisiones y estrategias en la industria automotriz, permitiendo ajustes precisos en estrategias de precios y marketing.

6. ANEXO

Cuadro 9: Resultados de la regresión lineal

Dep. Variable:	y	R-squared:	0.723
Model:	OLS	Adj. R-squared:	0.722
Method:	Least Squares	F-statistic:	529.8
Date:	Sat, 02 Sep 2023	Prob (F-statistic):	$1{,}71 \times 10^{-58}$
Time:	02:54:11	Log-Likelihood:	1197.0
No. Observations:	205	AIC:	-2390.
Df Residuals:	203	BIC:	-2383.
Df Model:	1	Covariance Type:	nonrobust

Cuadro 10: Coeficientes de la regresión

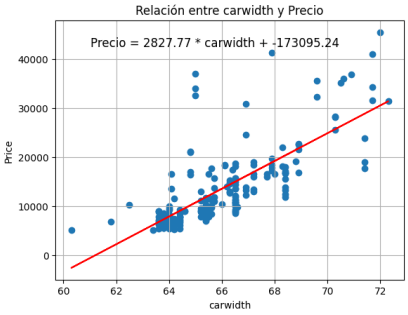
	coef	std err	t	P¿—t—	[0.025 0.975]
const	1.4629	0.005	270.843	0.000	1.452 1.474
x1	0.0906	0.004	23.017	0.000	0.083 0.098

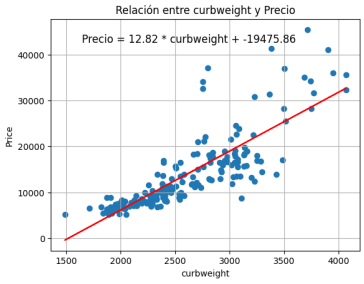
Cuadro 11: Estadísticas adicionales

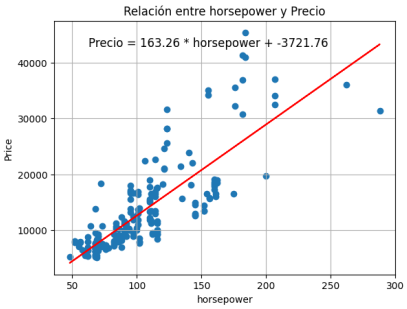
Omnibus:	3.239
Prob(Omnibus):	0.198
Jarque-Bera (JB):	3.213
Prob(JB):	0.201
Skew:	0.304
Kurtosis:	2.917
Cond. No.:	229.

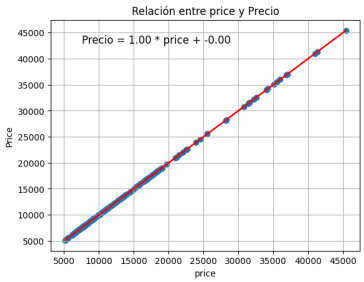
Cuadro 12: Normalidad de los residuos para Transformación de Box-Cox - horsepower

Estadístico de prueba:	0.7300200507943089
Valor p:	0.05616457723038909









Cuadro 13: Resultados de la regresión lineal

Dep. Variable:	y	R-squared:	1.000
Model:	OLS	Adj. R-squared:	1.000
Method:	Least Squares	F-statistic:	$7{,}934 \times 10^{20}$
Date:	Sat, 02 Sep 2023	Prob (F-statistic):	0.00
Time:	02:54:14	Log-Likelihood:	5640.1
No. Observations:	205	AIC:	-1.128e+04
Df Residuals:	203	BIC:	-1.127e+04
Df Model:	1	Covariance Type:	nonrobust

Cuadro 14: Coeficientes de la regresión

	coef	std err	t	P¿—t—	[0.025-0.975]
const	$-9{,}193 \times 10^{-14}$	$4{,}92 \times 10^{-11}$	-0.002	0.999	$-9{,}72 \times 10^{-11}$
$9{,}7 \times 10^{-11}$					
x1	1.0000	$3{,}55 \times 10^{-11}$	$2{,}82 \times 10^{10}$	0.000	1.000
1.000					

Cuadro 15: Estadísticas adicionales

Omnibus:	60.710
Prob(Omnibus):	0.000
Jarque-Bera (JB):	106.372
Prob(JB):	$7{,}97 \times 10^{-24}$
Skew:	-1.705
Kurtosis:	3.907
Cond. No.:	$5{,}43 \times 10^{03}$

Cuadro 16: Normalidad de los residuos para Transformación de Box-Cox - price

Estadístico de prueba:	57.42941735924313
Valor p:	0.0

Factor	Suma de cuadrados	df	Valor F	Valor p
fueltype	1.454053e+08	1.0	2.292741	0.131536
Residual	1.287423e+10	203.0	NaN	NaN

Factor	Suma de cuadrados	df	Valor F	Valor p
carbody	1.801997e+09	4.0	8.031976	0.000005
Residual	1.121764e+10	200.0	NaN	NaN

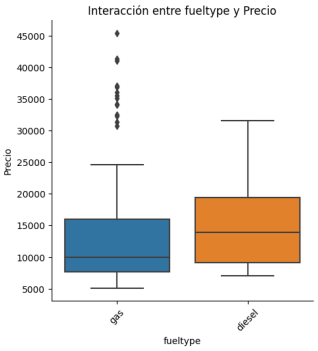


Figura 1: Gráfica 1

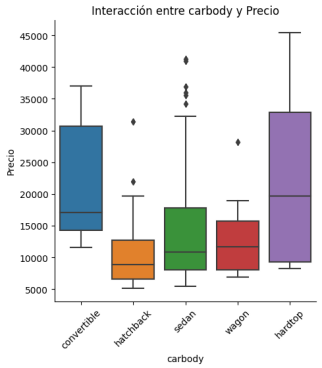


Figura 2: Gráfica 2

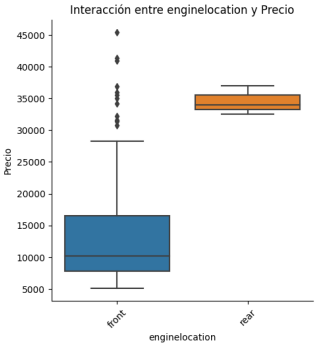


Figura 3: Gráfica 3

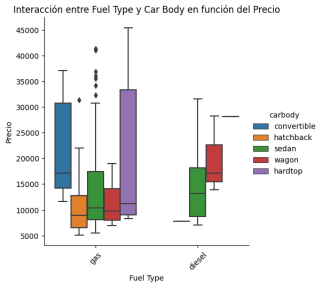


Figura 4: Gráfica 4

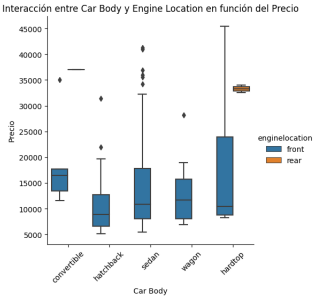


Figura 5: Gráfica 5

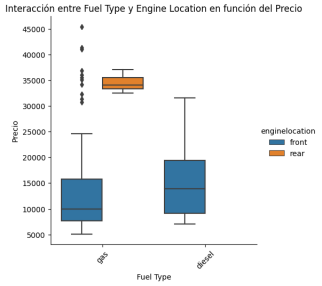


Figura 6: Gráfica 6

Factor	Suma de cuadrados	df	Valor F	Valor p
enginelocation	1.374973e+09	1.0	23.96974	0.000002
Residual	1.164467e+10	203.0	NaN	NaN

Factor	Suma de cuadrados	df	Valor F	Valor p
carbody	8.038503e+07	4.0	0.37482	0.541091
enginelocation	NaN	1.0	NaN	NaN
carbody:enginelocation	2.485407e+07	4.0	0.11589	0.733897
Residual	1.061591e+10	198.0	NaN	NaN

Factor	Suma de cuadrados	df	Valor F	Valor p
carbody	8.038503e+07	4.0	0.37482	0.541091
enginelocation	NaN	1.0	NaN	NaN
carbody:enginelocation	2.485407e+07	4.0	0.11589	0.733897
Residual	1.061591e+10	198.0	NaN	NaN

Factor	Suma de cuadrados	df	Valor f	Valor p
fueltype	1.454053e+08	1.0	2.562784e+00	0.110969
enginelocation	-7.126619e-08	1.0	-1.256074e-15	1.000000
fueltype:enginelocation	1.413309e+09	1.0	2.490972e+01	0.000001
Residual	1.146093e+10	202.0	-	-