modelacion

September 2, 2023

[1]:
```
import pandas as pd
```

[2]:
```
df = pd.read_csv(r'E:
↪\Github\Portafolio_Implementacion\EstadisticoBase\autos_prepared.csv')
```

[10]:
```
df.head()
```

[10]:

	carwidth	carlength	curbweight	enginesize	horsepower	price	fueltype
0	0.170668	0.415927	0.342192	0.258442	0.380165	0.232964	0
1	0.170668	0.415927	0.342192	0.258442	0.380165	0.317370	0
2	0.323440	0.457547	0.471404	0.353205	0.657800	0.317370	0
3	0.399825	0.551193	0.243051	0.167988	0.322056	0.245744	0
4	0.421650	0.551193	0.471874	0.284287	0.405992	0.344054	0

[12]:
```
import statsmodels.api as sm

X = df[['carwidth', 'carlength', 'curbweight', 'enginesize', 'horsepower']]
y = df['price']

X = sm.add_constant(X)

model = sm.OLS(y, X).fit()

model_summary = model.summary()
print(model_summary)
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.828
Model:                            OLS   Adj. R-squared:                  0.824
Method:                 Least Squares   F-statistic:                     191.4
Date:                Fri, 01 Sep 2023   Prob (F-statistic):           5.27e-74
Time:                        19:22:05   Log-Likelihood:                 198.82
No. Observations:                 205   AIC:                            -385.6
Df Residuals:                     199   BIC:                            -365.7
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
```

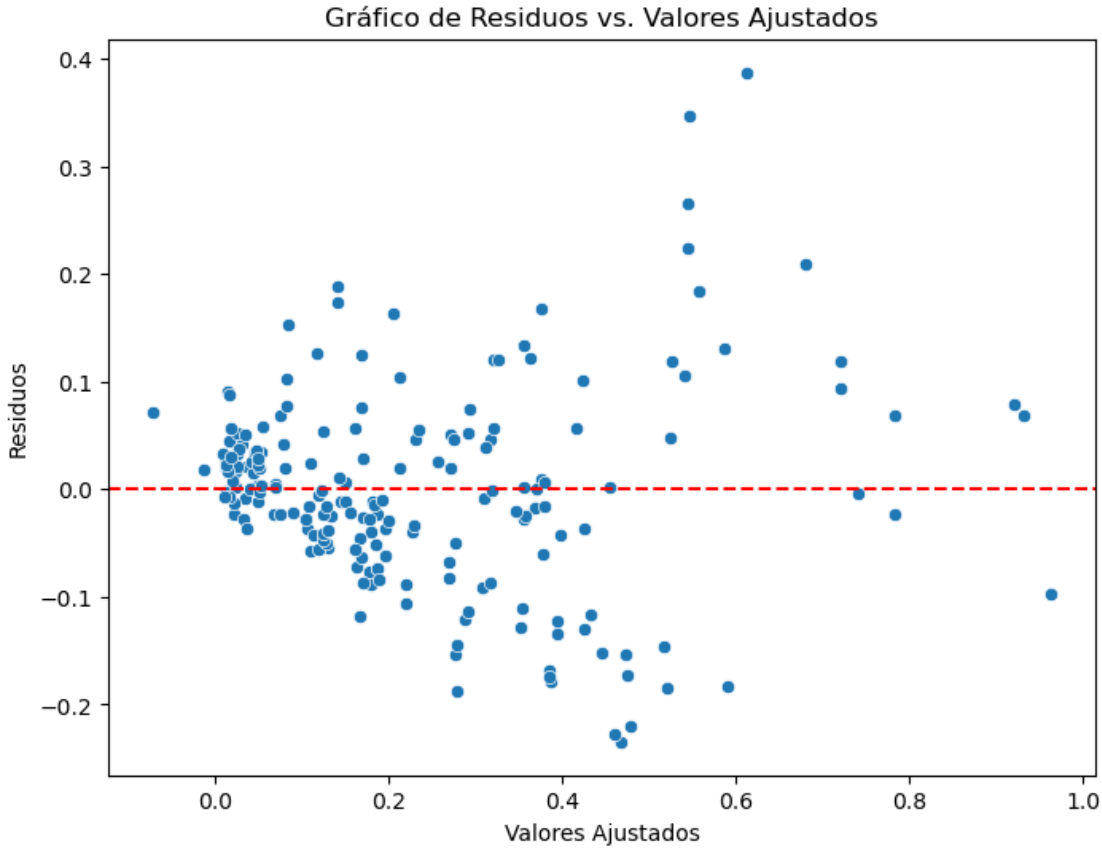
```
------------------------------------------------------------------------------
const         -0.0708      0.021     -3.397      0.001      -0.112      -0.030
carwidth       0.1900      0.062      3.085      0.002       0.069       0.312
carlength     -0.0744      0.071     -1.048      0.296      -0.214       0.066
curbweight     0.1602      0.090      1.782      0.076      -0.017       0.337
enginesize     0.5201      0.082      6.366      0.000       0.359       0.681
horsepower     0.2467      0.049      5.042      0.000       0.150       0.343
==============================================================================
Omnibus:                       22.599   Durbin-Watson:                   0.701
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               49.820
Skew:                           0.500   Prob(JB):                     1.52e-11
Kurtosis:                       5.198   Cond. No.                         22.9
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

Justificacion:

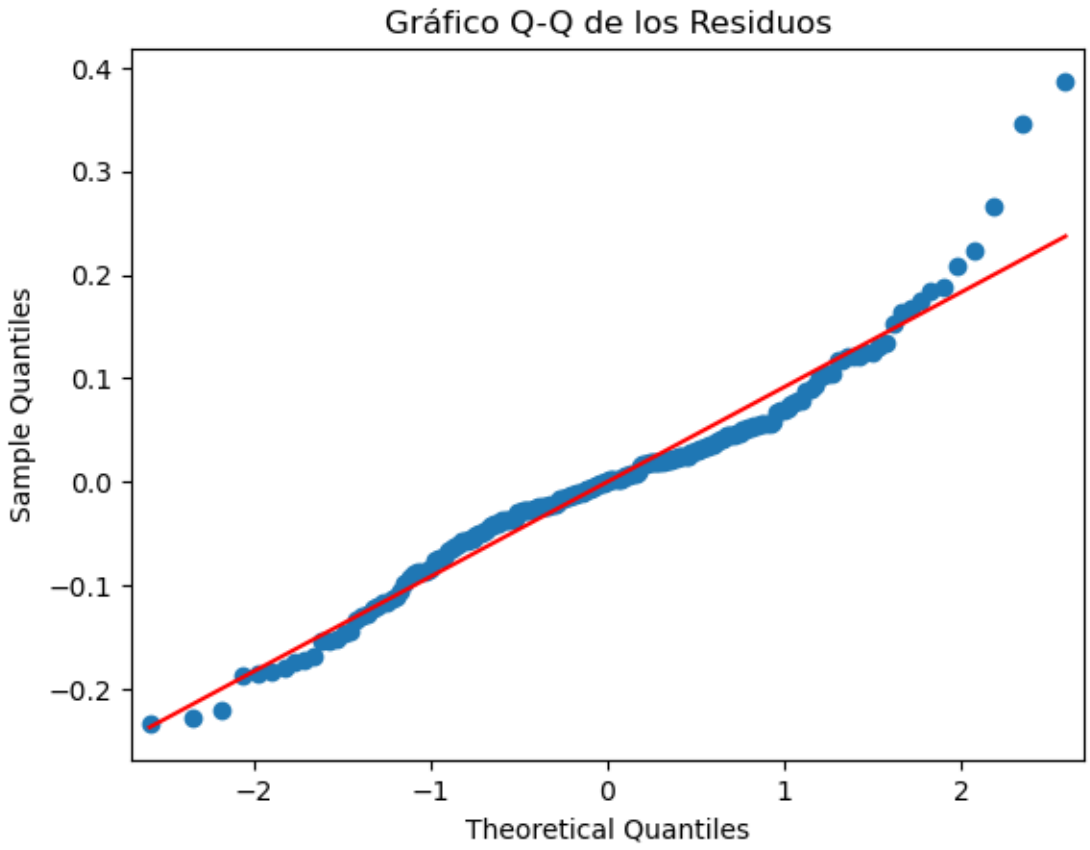
- Naturaleza de los Datos: Los datos contienen múltiples características de automóviles (como ancho, longitud, peso, tamaño del motor, y caballos de fuerza) y un objetivo cuantitativo (precio). La regresión lineal es adecuada para entender cómo estas características múltiples juntas pueden influir en el precio.
- Interpretabilidad: Una de las ventajas de la regresión lineal es que es altamente interpretable. Nos proporciona coeficientes para cada variable, que indican el cambio esperado en la variable dependiente por un cambio unitario en la variable independiente, manteniendo constantes las demás variables. Esto es útil para entender cuáles características tienen el mayor impacto en el precio.
- Adaptabilidad: La regresión lineal puede ser fácilmente extendida o adaptada. Si descubrimos que la relación no es estrictamente lineal, podemos agregar términos polinómicos o de interacción.

[13]:
```python
import matplotlib.pyplot as plt
import seaborn as sns
from statsmodels.graphics.gofplots import qqplot
from statsmodels.stats.outliers_influence import variance_inflation_factor

plt.figure(figsize=(8, 6))
sns.scatterplot(x=model.fittedvalues, y=model.resid)
plt.axhline(y=0, color='r', linestyle='--')
plt.title('Gráfico de Residuos vs. Valores Ajustados')
plt.xlabel('Valores Ajustados')
plt.ylabel('Residuos')
plt.show()

plt.figure(figsize=(8, 6))
qqplot(model.resid, line='s')
```

```
plt.title('Gráfico Q-Q de los Residuos')
plt.show()

vif_data = pd.DataFrame()
vif_data["Variable"] = X.columns
vif_data["VIF"] = [variance_inflation_factor(X.values, i) for i in range(X.
 ↪shape[1])]

vif_data
```




```
<Figure size 800x600 with 0 Axes>
```



[13]:

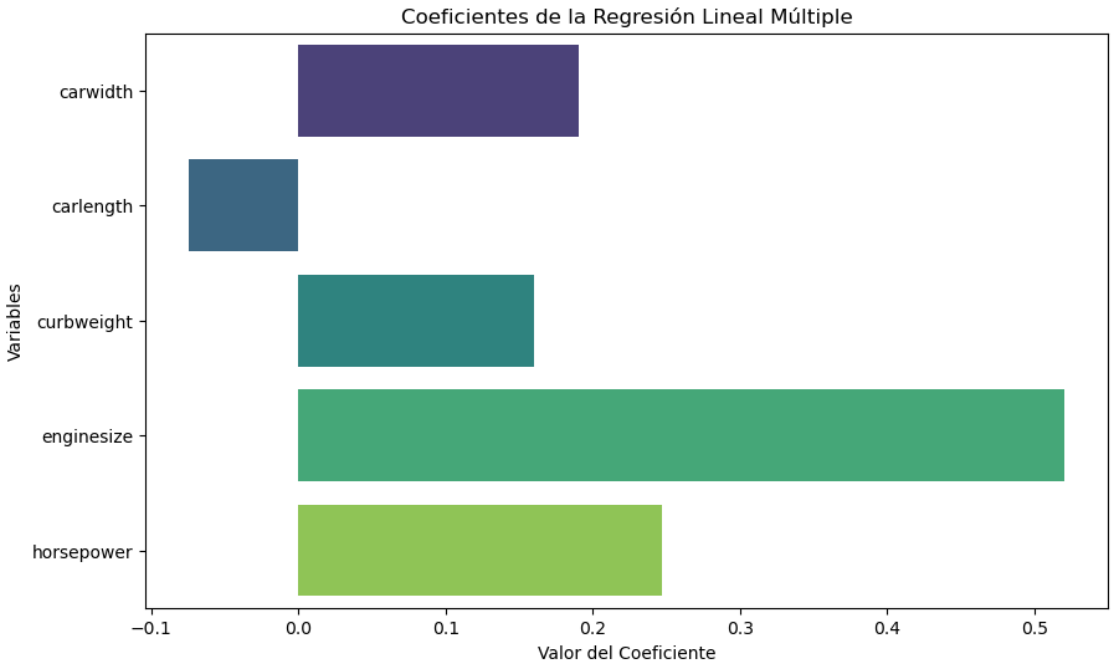
	Variable	VIF
0	const	10.272981
1	carwidth	4.702674
2	carlength	5.317722
3	curbweight	11.074636
4	enginesize	4.870335
5	horsepower	3.234256

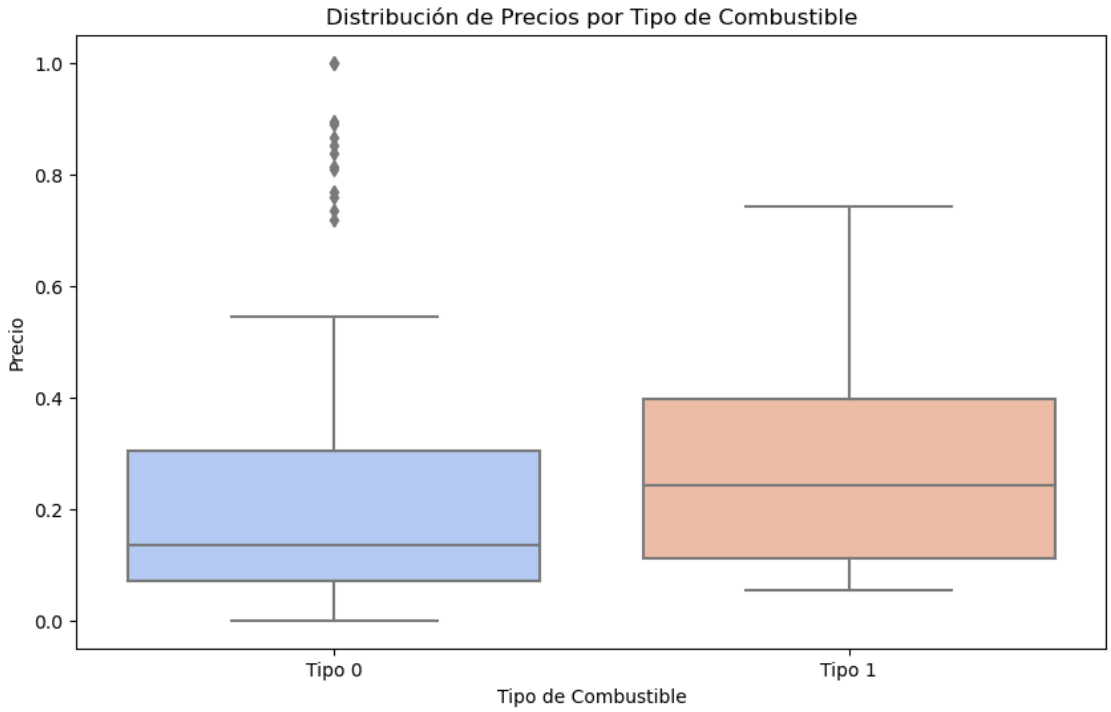
Justificacion:

- Naturaleza de los Datos: El conjunto de datos tiene una variable binaria fueltype, lo que sugiere una comparación natural entre dos grupos. Una prueba de hipótesis para medias es adecuada para determinar si hay una diferencia significativa en el precio medio entre estos dos grupos.
- Objetivo de Análisis: Queríamos determinar si el tipo de combustible tiene un impacto significativo en el precio del coche. Una prueba t para comparar medias es una herramienta estándar para este tipo de análisis.
- Claridad y Sencillez: Las pruebas de hipótesis ofrecen resultados claros y concluyentes en términos de p-values, lo que facilita la toma de decisiones.

0.0.1 Graficacion de resultados:

[15]:
```
coefficients = model.params[1:]
coef_labels = coefficients.index

plt.figure(figsize=(10, 6))
sns.barplot(x=coefficients, y=coef_labels, palette='viridis')
plt.title('Coeficientes de la Regresión Lineal Múltiple')
plt.xlabel('Valor del Coeficiente')
plt.ylabel('Variables')
plt.show()

plt.figure(figsize=(10, 6))
sns.boxplot(x='fueltype', y='price', data=df, palette='coolwarm')
plt.title('Distribución de Precios por Tipo de Combustible')
plt.xlabel('Tipo de Combustible')
plt.ylabel('Precio')
plt.xticks(ticks=[0, 1], labels=['Tipo 0', 'Tipo 1'])
plt.show()
```

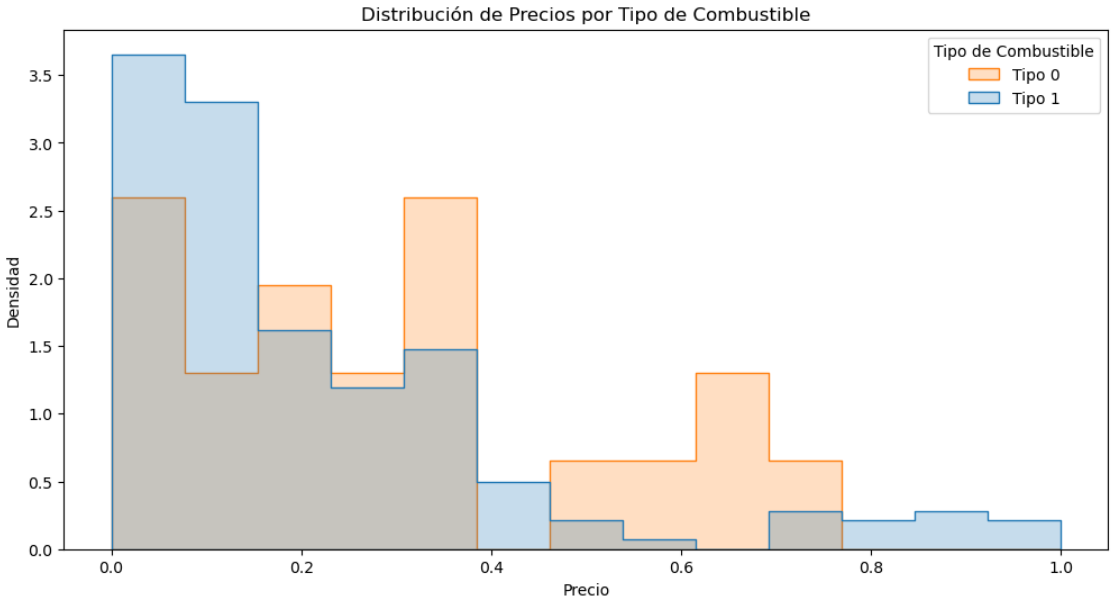




Visualizacion de resultados:

- Coeficientes de la Regresión Lineal Múltiple: Esta gráfica muestra el valor de los coeficientes para cada variable en la regresión. Un coeficiente positivo indica que hay una relación positiva entre esa característica y el precio, mientras que un coeficiente negativo indica una relación negativa. La magnitud del coeficiente nos da una idea de la fuerza de esa relación.
- Distribución de Precios por Tipo de Combustible: Este boxplot muestra la distribución de precios para los dos tipos de combustible. Las líneas centrales en las cajas representan las medianas, mientras que las cajas en sí muestran el rango intercuartil (IQR) y los bigotes se extienden hasta 1.5 veces el IQR. Los puntos individuales fuera de los bigotes son valores atípicos.

[16]:
```
plt.figure(figsize=(12, 6))
sns.histplot(df, x='price', hue='fueltype', element='step', stat='density',␣
 ↪common_norm=False)
plt.title('Distribución de Precios por Tipo de Combustible')
plt.xlabel('Precio')
plt.ylabel('Densidad')
plt.legend(title='Tipo de Combustible', labels=['Tipo 0', 'Tipo 1'])
plt.show()
```



El histograma nos permite visualizar cómo se distribuyen los precios de los coches para cada tipo de combustible y nos da una idea de la superposición y las diferencias entre las dos distribuciones.

Interpretacion de resultados:

Regresión Lineal Múltiple:

- Coeficientes: Los coeficientes nos indican cómo cambia el precio del coche con un cambio unitario en cada característica, manteniendo todo lo demás constante.
  - Por ejemplo, el coeficiente positivo para enginesize sugiere que, en promedio, un aumento en el tamaño del motor está asociado con un aumento en el precio del coche. Esto tiene sentido ya que autos con motores más grandes suelen ser más caros.
  - De manera similar, el coeficiente positivo para horsepower indica que los coches con más caballos de fuerza tienden a ser más caros, lo que es coherente con la idea de que los coches más potentes suelen ser más caros.
- Significancia de los coeficientes:
  - Algunos coeficientes no eran estadísticamente significativos (como carlength), lo que sugiere que, después de considerar las otras características, la longitud del coche no tiene un impacto significativo en su precio.

Prueba de Hipótesis sobre el tipo de combustible:

- La prueba t que realizamos buscaba determinar si hay una diferencia significativa en el precio medio de los coches según el tipo de combustible.
- Nuestro resultado no mostró una diferencia estadísticamente significativa en el precio entre los dos tipos de combustible. Esto indica que, basándonos en los datos proporcionados, no podemos afirmar que un tipo de combustible es más caro que el otro.

0.0.2 Conclusiones:

El análisis de los datos del mercado automovilístico revela patrones interesantes sobre cómo ciertas características de los automóviles influyen en su precio. A través de la regresión lineal múltiple, se identificó que el tamaño del motor (enginesize) y la potencia (horsepower) son factores significativos que influyen positivamente en el precio de un coche. Específicamente, los coches con motores más grandes y mayor potencia tienden a ser más caros en el mercado.

Por otro lado, la longitud del coche (carlength) no demostró tener un impacto significativo en el precio cuando se consideran otras características. Esto podría indicar que los compradores no valoran tanto la longitud del coche como el tamaño y potencia del motor cuando deciden cuánto están dispuestos a pagar por un vehículo.

En cuanto al tipo de combustible, a pesar de las variaciones observables en las distribuciones de precios entre los dos grupos, el análisis estadístico concluyó que no hay una diferencia significativa en el precio medio entre los coches basada en su tipo de combustible. Esto sugiere que otros factores, más allá del tipo de combustible, desempeñan un papel más importante en la determinación del precio.

Visualmente, las distribuciones de precios y los coeficientes de regresión respaldan estas conclusiones, ofreciendo una representación clara de las tendencias y relaciones en los datos.

En conjunto, este análisis proporciona una comprensión valiosa para diversas partes interesadas, como fabricantes, concesionarios y compradores, sobre los factores que más influyen en el precio de un coche. Las decisiones de producción, comercialización y compra pueden beneficiarse de estos hallazgos para maximizar el valor y la eficiencia en el mercado automovilístico.