EXTRACCIÓN DE CONOCIMIENTOS DE BASES DE DATOS
ANÁLISIS DE DATOS. - AUTOS

ARANDA GONZÁLEZ LUCÍA DEL CARMEN
9ADGS

DOMINGO 23 DE JUNIO DE 2024

ROL DEL AUTOMÓVIL EN EL MUNDO

El automóvil ha desempeñado un papel fundamental en la configuración de la sociedad moderna, transformando radicalmente la forma en que nos transportamos día a día. El automóvil, desde su crecimiento a finales del siglo XIX, ha avanzado para transformarse en un componente esencial de la vida diaria in muchos lugares del mundo, impactando en las dinámicas sociales, la economía global y el diseño urbano. La revolución industrial ayudo la evolución del automóvil, pasando de ser un tipo de transporte de mayor capacidad y con un funcionamiento un tanto diferente en que se utilizaba vapor, posteriormente los combustibles y la electricidad, siendo más eficiente cada vez, y cumpliendo la función de transporte no solo para las personas sino también para cosas, cada vez en un menor tiempo. La globalización y el libre comercio son factores importantes influyendo en su desarrollo, transformado la sociedad, economía y países.

En este contexto, el análisis de datos de automóviles genera una cierta relevancia. Utilizando métodos sofisticados de análisis de datos y visualización, el estudio presentado explora varios aspectos de los vehículos. Este método ayuda a desentrañar las complejas relaciones entre las características de los automóviles y sus precios, lo que proporciona información útil sobre las tendencias del mercado y las preferencias de los consumidores. Este tipo de análisis es importante porque puede informar a fabricantes, consumidores y legisladores sobre los elementos que afectan el diseño, la producción y la comercialización de automóviles. Comprender las correlaciones entre el tamaño del motor, el consumo de combustible y el precio es esencial para el desarrollo de vehículos más sostenibles y accesibles en un mundo cada vez más consciente del impacto ambiental y la eficiencia energética.

El análisis de factores como el estilo de la carrocería, la ubicación del motor y el tipo de tracción ofrece una perspectiva integral de cómo las diversas configuraciones de los automóviles afectan su valor en el mercado. Estos conocimientos son útiles para la industria automotriz y clientes al tomar decisiones informadas al comprar un vehículo.

OBJETIVO DEL ANÁLISIS

Analizar los datos del archivo auto.csv, en la herramienta web JupyterLite, con las librerías de python pandas, matplotlib.pylab, piplite, seaborn y numpy, para generar gráficas e identificar la correlación entre datos.

DESARROLLO DEL ANÁLISIS

Se importan las bibliotecas necesarias, tales como pandas y numpy por el momento, el primero para analizar los datos y manipular las estructuras, y el segundo para realizar cálculos numéricos y operacionales matriciales. Cargamos el archivo CSV, indicando su ruta dentro de path, y se crea el DataFrame de auto, identificándolo como 'df', esto usando la función de pandas 'read_csv()'.

[1]:
```
import pandas as pd
import numpy as np
path='auto.csv'
df=pd.read_csv(path)
df
```

[1]:

	Unnamed: 0	symboling	normalized-losses	make	aspiration	num-of-doors	body-style	drive-wheels	engine-location	wheel-base	...	compression-ratio	horsepower	peak-rpm	city-mpg	highway-mpg	price	city-L/100km	horsepower-binned	diesel	gas
0	0	3	122	alfa-romero	std	two	convertible	rwd	front	88.6	...	9.0	111.0	5000.0	21	27	13495.0	11.190476	Medium	0	1
1	1	3	122	alfa-romero	std	two	convertible	rwd	front	88.6	...	9.0	111.0	5000.0	21	27	16500.0	11.190476	Medium	0	1
2	2	1	122	alfa-romero	std	two	hatchback	rwd	front	94.5	...	9.0	154.0	5000.0	19	26	16500.0	12.368421	Medium	0	1
3	3	2	164	audi	std	four	sedan	fwd	front	99.8	...	10.0	102.0	5500.0	24	30	13950.0	9.791667	Medium	0	1
4	4	2	164	audi	std	four	sedan	4wd	front	99.4	...	8.0	115.0	5500.0	18	22	17450.0	13.055556	Medium	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
196	196	-1	95	volvo	std	four	sedan	rwd	front	109.1	...	9.5	114.0	5400.0	23	28	16845.0	10.217391	Medium	0	1
197	197	-1	95	volvo	turbo	four	sedan	rwd	front	109.1	...	8.7	160.0	5300.0	19	25	19045.0	12.368421	High	0	1
198	198	-1	95	volvo	std	four	sedan	rwd	front	109.1	...	8.8	134.0	5500.0	18	23	21485.0	13.055556	Medium	0	1
199	199	-1	95	volvo	turbo	four	sedan	rwd	front	109.1	...	23.0	106.0	4800.0	26	27	22470.0	9.038462	Medium	1	0
200	200	-1	95	volvo	turbo	four	sedan	rwd	front	109.1	...	9.5	114.0	5400.0	19	25	22625.0	12.368421	Medium	0	1

Mostramos las primeras cinco líneas del DataFrame, obteniendo una vista de los datos cargados de manera limitada.

[2]:
```
df.head(5)
```

[2]:

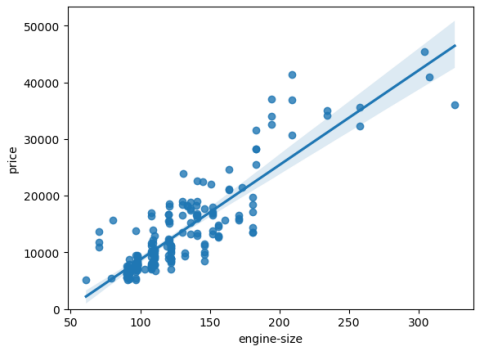
	Unnamed: 0	symboling	normalized-losses	make	aspiration	num-of-doors	body-style	drive-wheels	engine-location	wheel-base	...	compression-ratio	horsepower	peak-rpm	city-mpg	highway-mpg	price	city-L/100km	horsepower-binned	diesel	gas
0	0	3	122	alfa-romero	std	two	convertible	rwd	front	88.6	...	9.0	111.0	5000.0	21	27	13495.0	11.190476	Medium	0	1
1	1	3	122	alfa-romero	std	two	convertible	rwd	front	88.6	...	9.0	111.0	5000.0	21	27	16500.0	11.190476	Medium	0	1
2	2	1	122	alfa-romero	std	two	hatchback	rwd	front	94.5	...	9.0	154.0	5000.0	19	26	16500.0	12.368421	Medium	0	1
3	3	2	164	audi	std	four	sedan	fwd	front	99.8	...	10.0	102.0	5500.0	24	30	13950.0	9.791667	Medium	0	1
4	4	2	164	audi	std	four	sedan	4wd	front	99.4	...	8.0	115.0	5500.0	18	22	17450.0	13.055556	Medium	0	1

Importamos la biblioteca de matplotlib para crear gráficos y visualizaciones en las líneas siguientes. Importamos igualmente el módulo piplite y el paquete seaborn con funciones similares a la biblioteca primera, sin embargo, muestra resultados más atractivos y legibles. Y configuramos matplotlib con inline para mostrar las gráficas

dentro del notebook en vez de abrir una ventana emergente para que se muestre cerca de la línea.

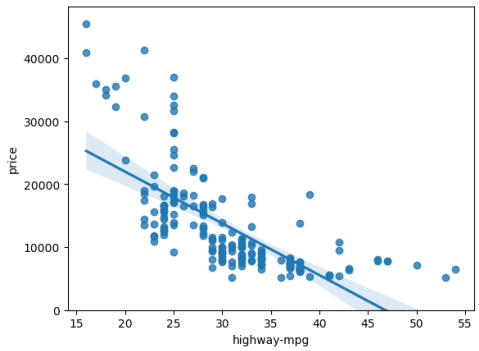
En el siguiente bloque se encuentra la creación del grafico de dispersión (scatter plot) entre dos variables que se encuentran en le DataFrame, que son ‘engine-size’ (en eje x) y ‘price’ (en eje y). Con la función de regplot se ajusta la línea de regresión lineal a los datos, mostrando la tendencia general entre el tamaño del motor y el precio de los automóviles, no sin antes establecer el límite inferior del eje y en 0 para no mostrar valores bajo esa cantidad.

```
[3]: import matplotlib.pylab as plt
     import piplite
     await piplite.install('seaborn')
     import seaborn as sns
     %matplotlib inline

[4]: sns.regplot(x = 'engine-size', y='price', data=df)
     plt.ylim(0)

[4]: (0.0, 53329.40897842941)
```



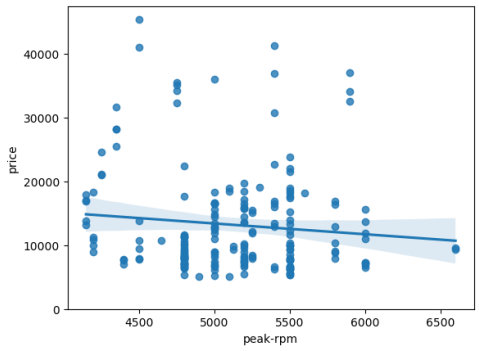
En la primera línea se crea el segundo gráfico de dispersión, ahora en las variables ‘highway-mpg’ en el eje ‘x’ y ‘price’ en el eje y. Con regplot se ajusta en automático una línea de regresión mostrando la tendencia general entre las millas por galón y el precio. En la segunda línea igual que el anterior, se mantiene el límite de 0.

```
[5]: sns.regplot(x = 'highway-mpg', y='price', data=df) #millas por galón
     plt.ylim(0)

[5]: (0.0, 48185.97243481936)
```



Se crea el tercer gráfico de dispersión, que además de la variable del precio, se coloca en el eje x la revolución por minuto o ‘peak-rpm’, se ajusta la línea y se limita a 0.

```
[6]: sns.regplot(x = 'peak-rpm', y='price', data=df) #revoluciónes por minuto
     plt.ylim(0)

[6]: (0.0, 47414.1)
```



Se calcula la matriz de correlación entre las columnas dentro del DataFrame df ‘engine-size’ y ‘price’, en que corr() calcula la correlación entre todas las columnas del DataFrame, excluyendo valores NA o nulos. Al considerar solo esa cantidad de columnas, la matriz resulta de 2x2.

```
[7]: df[['engine-size','price']].corr()
```

[7]:

	engine-size	price
engine-size	1.000000	0.872335
price	0.872335	1.000000

Se calcula la segunda matriz de correlación entre las columnas dentro del DataFrame df ‘highway-mpg’ y ‘price’, en que corr() calcula la correlación entre todas las columnas del DataFrame, excluyendo valores NA o nulos. Al considerar solo esa cantidad de columnas, es de 2x2.

```
[10]: df[['highway-mpg','price']].corr()
```

[10]:

	highway-mpg	price
highway-mpg	1.000000	-0.704692
price	-0.704692	1.000000

Se calcula la tercera matriz de correlación entre las columnas dentro del DataFrame df ‘peak-rpm’ y ‘price’, en que corr() calcula la correlación entre todas las columnas del DataFrame, excluyendo valores NA o nulos. Al considerar solo esa cantidad de columnas, la matriz resulta de 2x2.

```
[11]: df[['peak-rpm','price']].corr()
```

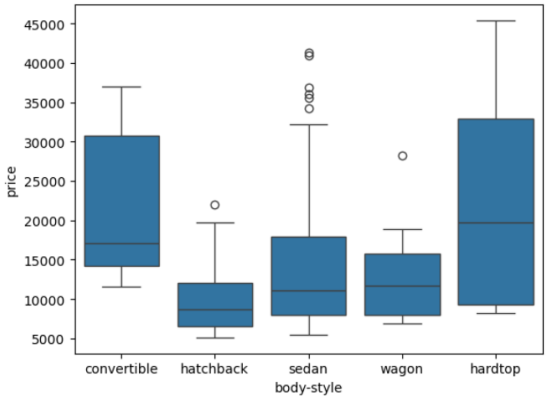
[11]:

	peak-rpm	price
peak-rpm	1.000000	-0.101616
price	-0.101616	1.000000

Con la biblioteca Seaborn se crea un gráfico de caja ‘boxplot’, en que muestra datos cuantitativos en función de categorías o niveles de una variable categórica, tomando como referencia ‘body-style’ o el estilo de los automóviles en x, y ‘price’ en y.

```
[12]: sns.boxplot(x='body-style', y='price', data=df)
```

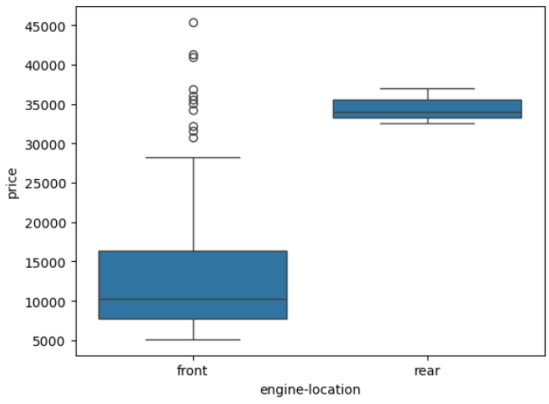
[12]: <AxesSubplot:xlabel='body-style', ylabel='price'>



Con la biblioteca Seaborn se crea el segundo gráfico de caja ‘boxplot’, en que muestra datos cuantitativos en función de categorías o niveles de una variable categórica, tomando como referencia ‘engine-location’ o la ubicación del motor en el auto, es decir, si se encuentra por la parte delantera o trasera, en el eje x, y ‘price’ en el eje y.

```
[13]: sns.boxplot(x='engine-location', y='price', data=df) #motor al frente o atrás
```

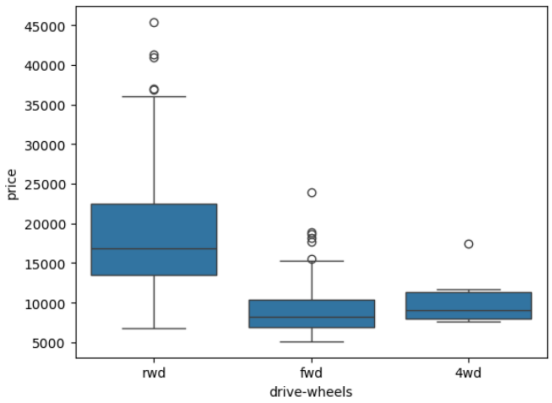
[13]: <AxesSubplot:xlabel='engine-location', ylabel='price'>



Con la biblioteca Seaborn se crea el tercer gráfico de caja ‘boxplot’, en que muestra datos cuantitativos en función de categorías o niveles de una variable categórica, tomando como referencia ‘drive-wheels’, y ‘price’ en y.

```
[14]: sns.boxplot(x='drive-wheels', y='price', data=df) #la trasera es más costosa
```

```
[14]: <AxesSubplot:xlabel='drive-wheels', ylabel='price'>
```



RESULTADOS

La gráfica de dispersión entre 'engine-size' y 'price' muestra una correlación positiva clara. A medida que aumenta el tamaño del motor, tiende a incrementar el precio del vehículo. Esto sugiere que los consumidores generalmente pagan más por automóviles con motores más grandes, posiblemente asociados con mayor potencia o lujo.

El gráfico que relaciona 'highway-mpg' (millas por galón en carretera) con el precio muestra una tendencia negativa. Los vehículos con mayor eficiencia de combustible tienden a ser menos costosos. Esto podría indicar que los automóviles de gama alta o de lujo suelen ser menos eficientes en términos de consumo de combustible.

La relación entre 'peak-rpm' y precio parece ser más débil o menos clara que las anteriores. Esto sugiere que las RPM máximas del motor no son un factor tan determinante en el precio como el tamaño del motor o la eficiencia de combustible.

Las matrices de correlación confirman estas observaciones visuales.

La correlación entre 'engine-size' y 'price' es fuertemente positiva (0.87), mientras que 'highway-mpg' y 'price' muestran una correlación negativa significativa (-0.70). La correlación entre 'peak-rpm' y 'price' es débilmente negativa (-0.10).

El boxplot de 'body-style' vs 'price' indica que ciertos estilos de carrocería, como los convertibles y los hatchback, tienden a tener rangos de precios más amplios y medianas más altas que otros estilos como los sedanes o wagons.

La gráfica de 'engine-location' vs 'price' muestra que los vehículos con motor trasero tienden a ser significativamente más caros que aquellos con motor delantero, aunque hay menos variedad en los primeros.

Finalizando, con los vehículos con tracción en las cuatro ruedas (4wd) tienden a ser más caros, seguidos por los de tracción trasera (rwd), mientras que los vehículos con tracción delantera (fwd) suelen ser los más económicos, según el boxplot de "drive-wheels" versus "price".

DISCUSIONES

Los resultados obtenidos en este análisis ofrecen valiosas perspectivas sobre la relación entre diversas características de los automóviles y sus precios, lo cual tiene implicaciones significativas para la industria automotriz y los consumidores.

La fuerte correlación positiva entre el tamaño del motor y el precio del vehículo (0.87) sugiere que los consumidores están dispuestos a pagar más por motores más grandes, posiblemente asociándolos con mayor potencia o lujo. Esta tendencia plantea interrogantes sobre la sostenibilidad a largo plazo, considerando la creciente preocupación global por la eficiencia energética y la reducción de emisiones. La industria automotriz podría enfrentar el desafío de reconciliar esta preferencia del consumidor con la necesidad de desarrollar vehículos más ecológicos.

Por otro lado, la correlación negativa entre la eficiencia de combustible (highway-mpg) y el precio (-0.70) indica una paradoja interesante: los vehículos más eficientes tienden a ser menos costosos. Esto podría reflejar una segmentación del mercado donde los vehículos de lujo priorizan el rendimiento sobre la eficiencia, mientras que los modelos más económicos se centran en el ahorro de combustible. Esta dinámica presenta una oportunidad para los fabricantes de desarrollar vehículos de alto rendimiento que

también sean eficientes en consumo, potencialmente creando un nuevo segmento de mercado.

La débil correlación entre las RPM máximas y el precio sugiere que este factor técnico tiene menos influencia en la percepción de valor por parte del consumidor. Esto podría indicar que los compradores se centran más en características más tangibles o fácilmente comprensibles al evaluar el valor de un vehículo.

Los resultados relacionados con el estilo de carrocería, la ubicación del motor y el tipo de tracción revelan preferencias de mercado específicas. La tendencia de precios más altos para convertibles y hatchbacks podría reflejar una prima por estilos de vida o funcionalidad específicos. La notable diferencia de precio entre vehículos con motor trasero y delantero sugiere una segmentación clara del mercado, posiblemente relacionada con vehículos deportivos o de alto rendimiento. Asimismo, la preferencia de precios más altos para vehículos con tracción en las cuatro ruedas indica una valoración de la versatilidad y el rendimiento en diferentes condiciones de manejo.

Estos hallazgos tienen implicaciones importantes para la estrategia de producto y marketing en la industria automotriz. Sugieren que existe un mercado diversificado con diferentes segmentos de consumidores que valoran características específicas. Los fabricantes podrían utilizar esta información para ajustar sus líneas de productos y estrategias de precios, potencialmente desarrollando vehículos que combinen características altamente valoradas (como motores potentes) con mayor eficiencia energética.

Además, estos resultados plantean cuestiones interesantes sobre la evolución futura del mercado automotriz en el contexto de la creciente conciencia ambiental y las regulaciones gubernamentales más estrictas sobre emisiones. La industria podría necesitar innovar para satisfacer las preferencias del consumidor mientras se adapta a un entorno regulatorio cambiante.