INFORME DE ANÁLISIS EXPLORATORIO DE DATOS (EDA)

NUCLIO DIGITAL SCHOOL – Data Science

Fecha	**:**	4 de enero de 2024	
Proyecto	**:**	Data Engineering: BMW pricing dataset	
Realizado por	**:**	Jorge Guiberteau Sánchez	Pedro Rincón Fernández
Grupo 6		Esther Sanz Llorente	Antonio Vargas Gómez

1. Análisis inicial

Este proyecto trata el dataset *'bmw_pricing_v3.csv*', que contiene 4846 registros cada uno relacionado con la venta de un automóvil de la marca BMW, de segunda mano, con detalles del modelo, el tipo de combustible, los kilómetros recorridos, los extras, las fechas de registro y de venta y el precio de la venta. El objetivo, es realizar la limpieza y el preprocesamiento de los datos con el fin de generar un modelo de predicción del precio de venta.

2. Columnas eliminadas

Se comprueba que no existen registros duplicados. Se eliminan las siguientes columnas:

- Columna **marca**: Sus valores son 'BMW' o Nulo (20%). No aporta información.
- Columna **fecha_registro:** Tiene un 50% de nulos y no hay manera objetiva de imputarle datos fiables.
- Columna **asientos_traseros_plegables:** Tiene un 70% de nulos no imputables
- Columna **fecha_venta:** Las fechas no tienen continuidad y no aportan información relevante
- Columna **gps:** Analizando la correlación de la variable frente al target (precio), se observa que esta columna no es significativa.

3. Tratamiento de datos nulos

- Columna **color**: 9% nulos no imputables. Se imputa el valor '*sin_info*'.
- Columna **tipo_coche:** 30% nulos. La columna es significativa, pero hay demasiada variabilidad para imputar valores a los registros nulos. Se imputa el valor *'sin_info'*.
- Columna **aire_acondicionado:** 10% nulos. Se imputa el valor '*sin_info'*.
- Columna **bluetooth**: 15% nulos. Se imputa el valor '*sin_info*'.
- Columna **alerta_lim_velocidad:** 15% nulos. Se imputa el valor '*sin_info'*.
- Columna **modelo**: 3 registros nulos. Se imputan manualmente valores para los registros 174 (*modelo 318)* y 4766 (*modelo X1*), basándonos en la tendencia del dataset. Como el registro 4802 no tiene una imputación manual clara, se elimina.
- Columna **km**: 2 registros nulos. Se eliminan.
- Columna **potencia**: 1 registro nulo. Se imputa manualmente el valor *'160'.*

- Columna **tipo_gasolina:** 5 registros nulos. Se imputa el valor *'diesel'* después de analizar la probabilidad para cada registro respecto al dataset.
- Columna **volante_regulable:** 4 registros nulos no imputables. Se eliminan.
- Columna **camara_trasera**: 2 registros nulos. Se imputa el valor *False.*
- Columna **elevalunas_electrico:** 2 registros nulos. Se imputa el valor *True*.
- Columna **precio**: 6 registros nulos. Como son pocos y es el target, se eliminan.

4. Análisis univariable

Las variables **volante_regulable, camara_trasera** y **elevalunas_electrico** son tipo *object,* pero sus valores son *True* y *False*. Se convierten a tipo *boolean*.

Variables numéricas:

- **km**: Se elimina el valor negativo y los outliers por encima de 400.000. Se convierte a *int32* para optimizar el cálculo y almacenamiento.
- **potencia**: Se elimina el registro con valor 0 y los outliers por encima de 400. Se convierte a *int16*.
- **precio (Target)**: Hay 3 registros por encima de 70.000, que se consideran outliers y se eliminan.

Variables categóricas:

- **modelo:** Se agrupan los modelos por su categoría general (por ejemplo *318 Gran Turismo* pasa a ser *318*). Los modelos con una representación inferior al 0.2% se agrupan en la categoría 'otros'. A partir de los 75 modelos iniciales, se obtienen 28 categorías relevantes.
- **tipo_gasolina:** Se agrupan las categorías *'Diesel'* y *'diesel'*. Se eliminan las categorías *'hybrid_petrol'* y *'electro'* ya que tienen muy pocos registros (8 y 3). Finalmente, se obtienen solamente dos categorías: *'diesel'* y *'petrol'*.
- **color:** No se realizan modificaciones. 11 categorías (10 colores y '*sin_info'*).
- **tipo_coche:** No se realizan modificaciones. 9 categorías (8 tipos y *'sin_info'*).
- **aire_acondicionado, bluetooth y alerta_lim_velocidad:** Son categóricas porque se inputó el valor '*sin_info'* a los registros nulos, pero sus valores son *True*, *False* y '*sin_info'.* No se realizan modificaciones.

5. Análisis de correlación inicial

Se observa que las variables con mayor influencia sobre el **precio** son la **potencia** y el **kilometraje**. La relación de la potencia con el target es directa, la del kilometraje inversa.

Las variables de extras **volante_regulable**, **camara_trasera** y **elevalunas_electrico** también tienen cierta relación directa con el precio de venta.

6. Análisis Variable vs Target

- **modelo:** Distribución única para cada modelo. Variable significativa.
- **tipo_gasolina:** Diferencias mínimas entre *diesel* y *petrol.*
- **color:** Distribución bastante uniforme. Los colores que se alejan de la tendencia general tienen muy pocos registros.
- **tipo_coche:** Al igual que modelo, parece ser una variable significativa.
- **potencia y km:** Se aprecia claramente la relación de estas variables con el precio de venta, directa e indirecta respectivamente.
- **Extras:** Se observa como el precio tiende a ser mayor cuando el vehículo tiene algún tipo de extra que cuando no lo tiene.

7. Transformación de variables categóricas

Se transforman las variables categóricas (**modelo**, **tipo_gasolina, color**, **tipo_coche, aire_acondicionado**, **bluetooth** y **alerta_lim_velocidad**) a booleanas utilizando el método **One_Hot-Encoding.** Esto hace que se obtenga una columna por cada valor único de las variables categóricas representado por valor 0 (False) o 1 (True) para cada registro. Así se puede procesar el dataset para la construcción del modelo.

8. Normalización de variables numéricas

Se normalizan las variables numéricas (km y potencia) utilizando **MinMaxScaler**. De esta forma, se ponderan los valores de las columnas numéricas asignando valores comprendidos entre 0 y 1, siendo 0 el valor mínimo de la muestra y 1 el valor máximo.

9. Análisis de correlación final

- Se reafirma que **potencia** y **km** tienen gran influencia en el precio de venta.
- También se observa cómo afecta directa o inversamente al precio que el vehículo tenga **extras** o no los tenga.
- Se observa la diversa relación de cada **modelo** y cada **tipo de coche** con el target. Destacan la tendencia a precios más altos del modelo *X5* y el tipo de coche *suv*, y la tendencia a precios más bajos del modelo *116*.
- Por otro lado, se observa que el **tipo de combustible** y el **color** del vehículo tienen una relación mínima con el precio, por lo tanto, se podría eliminar estas variables.
- Hay valores que muestran una relación significativa entre ellos (*modelo 218* con *tipo de coche van,* modelo *X3* con tipo de coche *suv* o **alerta_lim_velocidad** con **potencia)**. Esto se podría considerar en el análisis univariable para imputar o agrupar valores.

ANEXO:

Tras la limpieza y preprocesamiento, se ha obtenido un DataFrame con un tamaño de 4805 registros y 65 columnas. Las columnas se describen a continuación:

	columna	tipo_dato		columna	tipo_dato
0	km	float64	33	modelo_otros	bool
1	potencia	float64	34	tipo_gasolina_diesel	bool
2	volante_regulable	bool	35	tipo_gasolina_petrol	bool
3	camara_trasera	bool	36	color_beige	bool
4	elevalunas_electrico	bool	37	color_black	bool
5	precio	float64	38	color_blue	bool
6	modelo_114	bool	39	color_brown	bool
7	modelo_116	bool	40	color_green	bool
8	modelo_118	bool	41	color_grey	bool
9	modelo_120	bool	42	color_orange	bool
10	modelo_218	bool	43	color_red	bool
11	modelo_316	bool	44	color_silver	bool
12	modelo_318	bool	45	color_sin_info	bool
13	modelo_320	bool	46	color_white	bool
14	modelo_325	bool	47	tipo_coche_convertible	bool
15	modelo_330	bool	48	tipo_coche_coupe	bool
16	modelo_335	bool	49	tipo_coche_estate	bool
17	modelo_420	bool	50	tipo_coche_hatchback	bool
18	modelo_435	bool	51	tipo_coche_sedan	bool
19	modelo_518	bool	52	tipo_coche_sin_info	bool
20	modelo_520	bool	53	tipo_coche_subcompact	bool
21	modelo_525	bool	54	tipo_coche_suv	bool
22	modelo_530	bool	55	tipo_coche_van	bool
23	modelo_535	bool	56	aire_acondicionado_False	bool
24	modelo_640	bool	57	aire_acondicionado_True	bool
25	modelo_730	bool	58	aire_acondicionado_sin_info	bool
26	modelo_740	bool	59	bluetooth_False	bool
27	modelo_M550	bool	60	bluetooth_True	bool
28	modelo_X1	bool	61	bluetooth_sin_info	bool
29	modelo_X3	bool	62	alerta_lim_velocidad_False	bool
30	modelo_X4	bool	63	alerta_lim_velocidad_True	bool
31	modelo_X5	bool	64	alerta_lim_velocidad_sin_info	bool
32	modelo_X6	bool			

A continuación, se muestran los cinco primeros registros del DataFrame final:

5 primeros vehíclulos del DataFrame

Columnas	0	1	2	3	4
km	0,352	0,034	0,460	0,321	0,243
potencia	0,135	1,000	0,215	0,275	0,375
volante_regulable	True	True	False	True	True
camara_trasera	False	False	False	False	False
elevalunas_electrico	True	False	True	True	False
precio	11300	69700	10200	25100	33400
modelo_114	False	False	False	False	False
modelo_116	False	False	False	False	False
modelo_118	True	False	False	False	False
modelo_120	False	False	False	False	False
modelo_218	False	False	False	False	False
modelo_316	False	False	False	False	False
modelo_318	False	False	False	False	False
modelo_320	False	False	True	False	False
modelo_325	False	False	False	False	False
modelo_330	False	False	False	False	False
modelo_335	False	False	False	False	False
modelo_420	False	False	False	True	False
modelo_435	False	False	False	False	False
modelo_518	False	False	False	False	False
modelo_520	False	False	False	False	False
modelo_525	False	False	False	False	False
modelo_530	False	False	False	False	False
modelo_535	False	False	False	False	False
modelo_640	False	False	False	False	False
modelo_730	False	False	False	False	False
modelo_740	False	False	False	False	False
modelo_M550	False	False	False	False	False
modelo_X1	False	False	False	False	False
modelo_X3	False	False	False	False	False
modelo_X4	False	False	False	False	False
modelo_X5	False	False	False	False	False
modelo_X6	False	False	False	False	False
modelo_otros	False	True	False	False	True
tipo_gasolina_diesel	True	False	True	True	True
tipo_gasolina_petrol	False	True	False	False	False
color_beige	False	False	False	False	False
color_black	True	False	False	False	False
color_blue	False	False	False	False	False
color_brown	False	False	False	False	False

color_green	False	False	False	False	False
color_grey	False	True	False	False	False
color_orange	False	False	False	False	False
color_red	False	False	False	True	False
color_silver	False	False	False	False	True
color_sin_info	False	False	False	False	False
color_white	False	False	True	False	False
tipo_coche_convertible	False	True	False	True	False
tipo_coche_coupe	False	False	False	False	False
tipo_coche_estate	False	False	False	False	False
tipo_coche_hatchback	False	False	False	False	False
tipo_coche_sedan	False	False	False	False	False
tipo_coche_sin_info	True	False	True	False	True
tipo_coche_subcompact	False	False	False	False	False
tipo_coche_suv	False	False	False	False	False
tipo_coche_van	False	False	False	False	False
aire_acondicionado_False	False	False	True	False	False
aire_acondicionado_True	True	True	False	True	True
aire_acondicionado_sin_info	False	False	False	False	False
bluetooth_False	False	False	True	False	False
bluetooth_True	False	True	False	True	True
bluetooth_sin_info	True	False	False	False	False
alerta_lim_velocidad_False	False	False	True	False	False
alerta_lim_velocidad_True	False	True	False	False	True