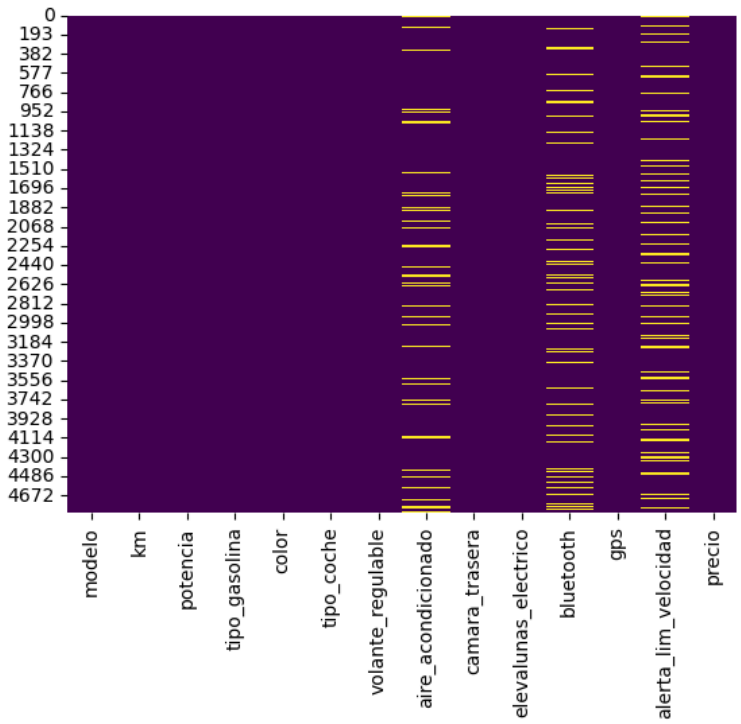
1.Data Engineering

Tras importar las librerías necesarias y cargar mi csv comienzo con el la limpieza de datos.
Reviso la estructura de los datos mediante los métodos .info() y .describe()
Se comprueba que no existen duplicados en el dataset.
Detectamos los nulos que tenemos en nuestro dataset. Y saco el porcentaje por cada columna. La columna marca no es relevante, ya que el mismo dato se repite para todos los registros ya que todos los vehículos son de la misma marca. Por ello, se elimina esta columna.
La columna fecha_registro tiene un porcentaje de nulos equivalente a 50.03%. Por lo que a pesar de ser una columna determinante para el precio, ya que si un coche es más nuevo tendrá mayor precio, procedo a eliminarla ya que caso contrario, tendría un dataset con la mitad de datos rellenos por la mediana(ya que sería mi elección) lo cual podría sesgar los datos.
De la misma manera, elimino la columna asientos_traseros_plegables ya que tiene un alto porcentaje de nulos (70.01%) y además no considero que esta columna sea determinante para el precio. La columna fecha_venta, también es eliminada ya que es un dato más bien informativo pero no influye en el precio de venta.
Saco el porcentaje de nulos resultante tras eliminar las columnas antes mencionadas y tengo las columnas 'modelo', 'km', 'potencia', 'tipo_gasolina', 'volante_regulable', 'camara_trasera', 'elevalunas_electrico', 'precio' cuyo porcentaje de nulos es menor al 1% por lo que procedo a eliminar los nulos.

Para los nulos de las columnas cuyo porcentaje ronda el 15% realizo un gráfico de nulos para ver si es conveniente eliminarlos o rellenarlos . El gráfico me sirve para saber si los nulos están distribuidos o no.



 En este caso, tras observar que están distribuidos y que no están sesgados hacia ninguna característica en particular, procedo a rellenarlos con la moda. Ya sin nulos en mi dataset procedo a hacer el análisis univariable.
Reviso primero las columnas numéricas. Observo outliers en la columna km por lo que procedo a eliminar aquellos inferiores a 0 y superiores a 400.000. Además observo que tras eliminar estos outliers la columna de km tiene una distribución normalizada. En el caso de la columna potencia elimino los outliers,

aquí selecciono aquellos que tienen una potencia inferior a 50 y superior a 400.La distribución de la columna muestra que hay más coches con potencia entre 100 y 150 de potencia. Y que hay pocos coches con más potencia, aquellos con más de 200 de potencia. Más adelante se analizará su relación con el precio.
Elimino los outliers de la columna precio, eliminando aquellos cuyo valor sea inferior o igual a 1300 y que tengan menos de 200.000km. Así mismo elimino aquellos cuyo importe sea superior a 150.000. Luego, convierto la columna precio a escala logarítmica para su posterior análisis y observo una distribución normal.
Convierto el dataframe en un archivo pickle para continuar con el preprocesamiento. Las columnas de tipo bool las convierto a númericas(int64), para después poder analizar la correlación.
Comienzo el análisis de las variables categóricas. Al revisar la columna tipo_gasolina veo que se han creado dos variables para el tipo diesel por ello las unifico, pasando a minusculas las de tipo “Diesel”.
Hago el análisis de correlación inicial donde puedo ver que a medida que aumentan los km se reduce el precio y que a más potencia el precio aumenta.Por otro lado vemos que extras como alerta_lim_velocidad, o elevalunas eléctrico también aumentan el valor del coche.
Voy a analizar las variables por ello imprimo los valores de las variables categóricas y genero un histograma de aquellas de tipo numérico.
Luego realizo un violinplot de las columnas de tipo numérico. La columna de volante_regulable me muestra que no es factor determinante en el precio, ya que no hay una diferencia significativa. Sin embargo, al revisar la relación entre el aire_acondicionado y el precio, vemos que el valor aumenta significativamente si se cuenta con ese extra. Analizando la relación entre el precio y la camara_trasera también vemos un aumento del importe del coche significativo. Respecto a la columna de elevalunas_electricos veo que el precio aumenta ligeramente si disponen de este extra. La columna bluetooth también aumenta ligeramente el precio de los coches que tienen incorporado este extra. Así mismo podemos observar que los coches que cuentan con gps tienen un precio significativamente superior a los que no lo tienen.Por último comprobamos la relación del precio respecto a alerta_lim_velocidad y vemos que el precio aumenta ligeramente si el coche incluye este extra. Así mismo, se comprueba mediante el gráfico scatterplot que los coches con menos km son significativamente más caros que aquellos que tienen más km . Ocurre lo mismo con la potencia, aquellos coches que tienen mayor potencia tienen mayor valor. Y estos coches más caros y con más potencias son menos frecuentes, algo lógico debido al precio.
Clasifico las columnas según su categoría, quedandome con 3 listas. Una con valores numéricos, otros de tipo bool y la última de tipo object a la que llamaremos categóricas.
Para las variables categóricas antes de convertirlas a numéricas observo que hay etiquetas que tienen menos del 1% de frecuencia. Estas etiquetas son eliminadas para no generar una columna de las mismas.

1.Data Engineering

Mediante el método get_dummies() genero las columnas numéricas de la lista a la que hemos llamado categóricas. Las columnas originales han sido eliminadas del dataset, las cuales eran modelo, color, tipo_coche y tipo_gasolina
Por último usando el método de minMaxScaler convierto las variables de tipo numéricas y observo la correlación de las columnas. Hay columnas que tienen correlación pero no la suficiente para ser eliminadas. Como es el caso del modelo X5 con el tipo de coche suv, ya que estos coches pertenecen a esa categoría.
Finalmente convierto el dataset ya limpio a csv y a xlsx.
Las columnas del dataset final se incluyen en el anexo.

ANEXOS

```
#   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   km                              4084 non-null
float64
 1   potencia                        4084 non-null
float64
 2   volante_regulable               4084 non-null   int64
 3   aire_acondicionado              4084 non-null   int64
 4   camara_trasera                  4084 non-null   int64
 5   elevalunas_electrico            4084 non-null   int64
 6   bluetooth                       4084 non-null   int64
 7   gps                             4084 non-null   int64
 8   alerta_lim_velocidad            4084 non-null   int64
 9   LOG_PRECIO                      4084 non-null   int64
 10  modelo_116                      4084 non-null   int64
 11  modelo_118                      4084 non-null   int64
 12  modelo_316                      4084 non-null   int64
 13  modelo_318                      4084 non-null   int64
 14  modelo_318 Gran Turismo         4084 non-null   int64
 15  modelo_320                      4084 non-null   int64
 16  modelo_320 Gran Turismo         4084 non-null   int64
 17  modelo_518                      4084 non-null   int64
 18  modelo_520                      4084 non-null   int64
 19  modelo_525                      4084 non-null   int64
 20  modelo_530                      4084 non-null   int64
 21  modelo_X1                       4084 non-null   int64
 22  modelo_X3                       4084 non-null   int64
 23  modelo_X5                       4084 non-null   int64
 24  tipo_gasolina_diesel            4084 non-null   int64
 25  tipo_gasolina_petrol            4084 non-null   int64
 26  color_beige                     4084 non-null   int64
 27  color_black                     4084 non-null   int64
 28  color_blue                      4084 non-null   int64
 29  color_brown                     4084 non-null   int64
 30  color_green                     4084 non-null   int64
 31  color_grey                      4084 non-null   int64
 32  color_orange                    4084 non-null   int64
 33  color_red                       4084 non-null   int64
 34  color_silver                    4084 non-null   int64
 35  color_sin_color                 4084 non-null   int64
 36  color_white                     4084 non-null   int64
 37  tipo_coche_estate               4084 non-null   int64
 38  tipo_coche_hatchback            4084 non-null   int64
```

```
 39  tipo_coche_sedan                    4084 non-null   int64
 40  tipo_coche_subcompact               4084 non-null   int64
 41  tipo_coche_suv                      4084 non-null   int64
 42  tipo_coche_tipo_coche_desconocido   4084 non-null   int64
dtypes: float64(2), int64(41)
```

1.Data Engineering

Al hacer un .head(5) este es el resultado:

```
df_bmw10.head(5)
```

✓ 0.0s

	km	potencia	volante_regulable	aire_acondicionado	camara_trasera	elevalunas_electrico	bluetooth	gps	alerta_lim_velocidad	LOG_PRECIO	...	color_red	color_silver	color_sin_color	color_white	tipo_coche_estate	tipo_coche_hatchback	tipo_coche_sedan
0	0.351757	0.100	1	1	0	1	0	1	1	4	...	0	0	0	0	0	0	0
2	0.459665	0.200	0	0	0	1	0	1	0	4	...	0	0	0	1	0	0	0
30	0.489398	0.125	1	1	0	0	1	1	0	3	...	1	0	0	0	0	0	0
56	0.826712	0.150	1	0	0	0	1	1	1	3	...	0	1	0	0	0	0	0
85	0.508768	0.250	0	0	0	1	0	1	1	3	...	0	1	0	0	0	0	0

5 rows × 43 columns

tipo_coche_subcompact	tipo_coche_suv	tipo_coche_tipo_coche_desconocido
0	0	1
0	0	1
0	0	1
0	0	1
0	0	1