Entregable 1: Pre-procesamiento Dataset BMW

Alumno: Juan Miguel Coll González

1. Columnas eliminadas

Las columnas eliminadas directamente del DataFrame de origen son:

- Marca (Todo el dataset es de BMW)
- Tipo_gasolina (El 96% del dataset era diesel)

De manera indirecta se eliminaron:

- Fecha_registro_DIA (Dia 1 siempre)
- Fecha_registro_NOMBREDIA (Al eliminar el anterior no aporta info.)
- Fecha_venta_DIA (IDEM)
- Fecha_venta_NOMBREDIA (IDEM)
- Fecha_venta_AÑO (Todas las ventas son en 2018)
- Fecha_registro_AÑO (La correlación con la antigüedad es muy alta)

2. Tratamiento de Nulos

Primeramente, se han eliminado los nulos del Target y a continuación se ha realizado un resumen de los nulos y diferenciado en 3 grupos según el porcentaje que suponen los registros nulos sobre el resto.

- Relevancia Baja: Columnas con % nulos < 1%.
- Relevancia Alta: Columnas con % nulos > 80%.
- Relevancia Media: Columnas con % nulos entre 1% y 80%.

Los de baja relevancia se eliminan directamente ya que no merece la pena realizar un estudio en profundidad.

Para los de alta relevancia se eliminaría la columna, no ha habido casos.

Para los de relevancia media se realiza la limpieza en el análisis univariable.

3. Análisis univariable

Columna Modelo

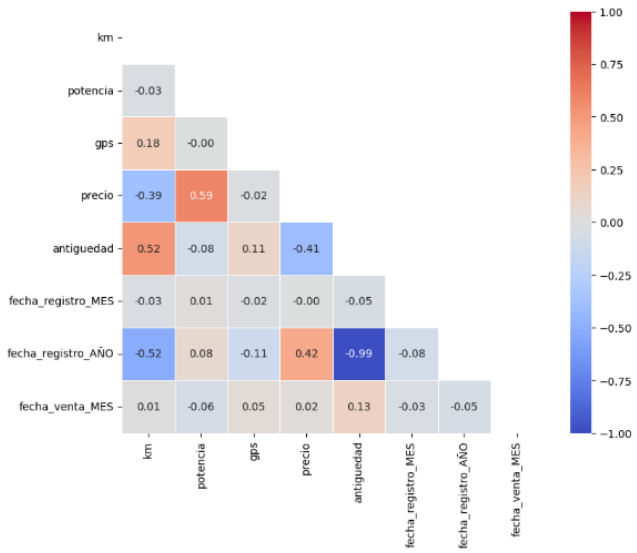
En cuanto al análisis univariable se encuentra que la columna 'modelo' cuenta con demasiados valores de poca relevancia (pocas apariciones) entonces se decide agrupar todos los elementos cuya aparición es menor al 2% en una categoría llamada 'Others'.

Varias columnas con nulos

Para las siguientes columnas, que aun contienen nulos, se añade una categoría 'Unknown'.

- Color
- Tipo_coche
- Aire_acondicionado
- Asientos_traseros_plegables
- Bluetooth
- Alerta_lim_velocidad.

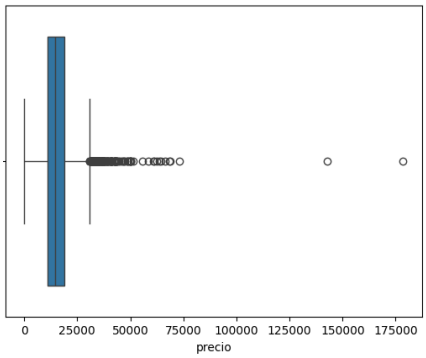
4. Correlaciones



El triángulo de correlaciones muestra una correlación muy fuerte entre la 'antigüedad' (en días) y 'fecha_registro_AÑO' (en años).

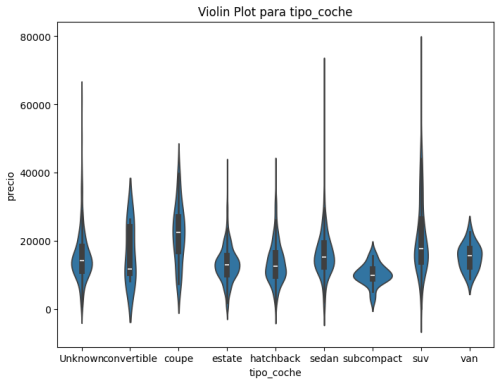
Se decide eliminar la columna 'fecha_registro_AÑO' porque la antigüedad aporta más información al encontrarse en días.

5. Análisis de variables VS Target



- **Análisis del Target:**

El target presenta dos outliers destacables. Estos se eliminan para mantener una distribución normal.



- **Observación 1:**

Si el precio supera los 50.000 probablemente el tipo de coche sea tipo 'sedan' o 'suv'.

- **Observación 2:**

El precio aumenta si en las columnas de True/False el valor es True.

6. Transformación de variables

Para la transformación de variables categóricas a numéricas se ha utilizado la técnica de One Hot Encoder. Las columnas a codificar son:

- Modelo
- Color
- Tipo_coche
- Volante_regulable
- Aire_acondicionado
- Cámara_trasera
- Asientos_traseros_plegables
- Elevalunas_electrico
- Bluetooth
- Alerta_lim_velocidad

7. Correlación final

Realizando la correlación de nuevo se obtienen muchas correlaciones superiores al 0.8. Se eliminan las siguientes columnas:

- color_orange
- color_green
- tipo_coche_convertible
- tipo_coche_van
- tipo_coche_coupe
- color_beige
- color_red
- tipo_coche_subcompact
- modelo_118
- modelo_530
- modelo_X5
- modelo_316
- modelo_525
- modelo_X1
- color_silver
- modelo_116
- modelo_X3
- asientos_traseros_plegables_True
- color_brown

Entregable 1: Anexo

Lista de columnas del result_df

```
Data columns (total 38 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   km                                    2310 non-null   float64
 1   potencia                              2310 non-null   float64
 2   gps                                   2310 non-null   object
 3   precio                                2310 non-null   float64
 4   antiguedad                            2310 non-null   float64
 5   fecha_registro_MES                    2310 non-null   float64
 6   fecha_venta_MES                       2310 non-null   float64
 7   modelo_318                            3526 non-null   bool
 8   modelo_320                            3526 non-null   bool
 9   modelo_520                            3526 non-null   bool
 10  modelo_Others                         3526 non-null   bool
 11  color_Unknown                         3526 non-null   bool
 12  color_black                           3526 non-null   bool
 13  color_blue                            3526 non-null   bool
 14  color_grey                            3526 non-null   bool
 15  color_white                           3526 non-null   bool
 16  tipo_coche_Unknown                    3526 non-null   bool
 17  tipo_coche_estate                     3526 non-null   bool
 18  tipo_coche_hatchback                  3526 non-null   bool
 19  tipo_coche_sedan                      3526 non-null   bool
 20  tipo_coche_suv                        3526 non-null   bool
 21  volante_regulable_False               3526 non-null   bool
 22  volante_regulable_True                3526 non-null   bool
 23  aire_acondicionado_False              3526 non-null   bool
 24  aire_acondicionado_True               3526 non-null   bool
 25  aire_acondicionado_Unknown            3526 non-null   bool
 26  camara_trasera_False                  3526 non-null   bool
 27  camara_trasera_True                   3526 non-null   bool
 28  asientos_traseros_plegables_False     3526 non-null   bool
 29  asientos_traseros_plegables_Unknown   3526 non-null   bool
 30  elevalunas_electrico_False            3526 non-null   bool
 31  elevalunas_electrico_True             3526 non-null   bool
 32  bluetooth_False                       3526 non-null   bool
 33  bluetooth_True                        3526 non-null   bool
 34  bluetooth_Unknown                     3526 non-null   bool
 35  alerta_lim_velocidad_False            3526 non-null   bool
 36  alerta_lim_velocidad_True             3526 non-null   bool
 37  alerta_lim_velocidad_Unknown          3526 non-null   bool
dtypes: bool(31), float64(6), object(1)
memory usage: 299.7+ KB
```

```
result_df.head()
```

0.0s

	km	potencia	gps	precio	antiguedad	fecha_registro_MES	fecha_venta_MES	modelo_318	modelo_320	modelo_520	...	asientos_traseros_plegables_False	asientos_traseros_plegables_Unknown	elevalunas_electrico_False	ele
0	140411.0	100.0	True	11300.0	2161.0	2.0	1.0	False	False	False	...	False	True	False	
1	183297.0	120.0	True	10200.0	2132.0	4.0	2.0	False	False	False	...	True	False	False	
2	132025.0	135.0	True	21700.0	1461.0	3.0	3.0	False	False	False	...	False	True	True	
3	77061.0	135.0	True	36300.0	943.0	8.0	3.0	False	True	False	...	False	True	False	
4	174631.0	120.0	True	10500.0	3377.0	1.0	4.0	False	False	False	...	True	False	False	

5 rows × 38 columns

elevalunas_electrico_True	bluetooth_False	bluetooth_True	bluetooth_Unknown	alerta_lim_velocidad_False	alerta_lim_velocidad_True	alerta_lim_velocidad_Unknown
True	False	False	True	False	False	True
True	False	True	False	False	False	True
False	True	False	False	False	False	True
True	True	False	False	True	False	False
True	True	False	False	True	False	False

Juan Miguel Coll González