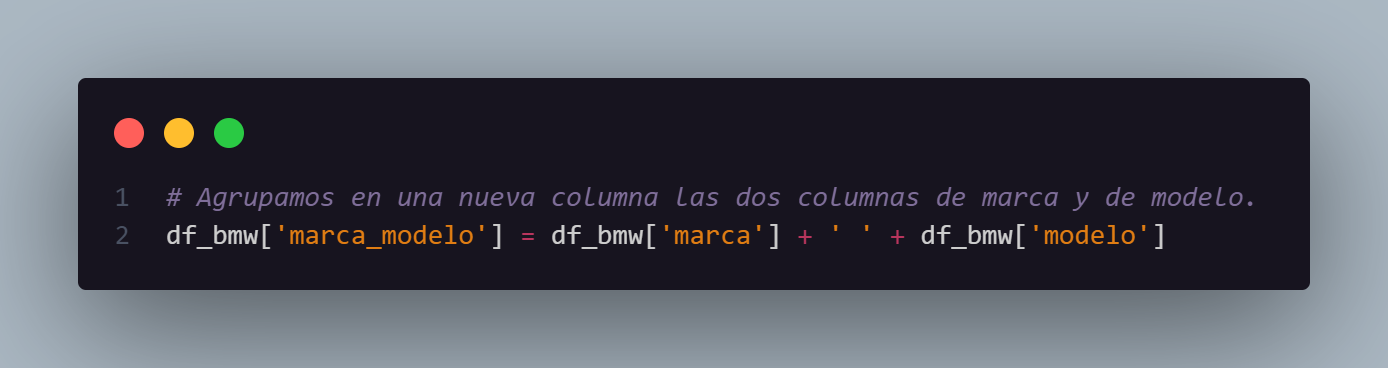
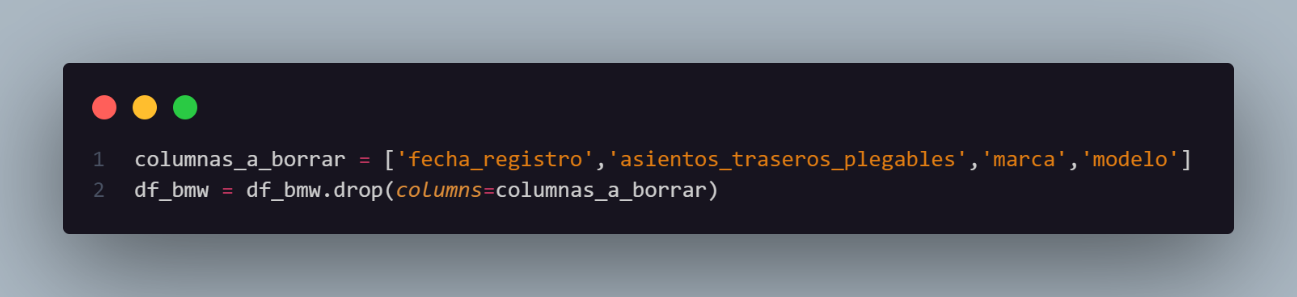
- Desde la primera columna vimos que había nulos y comenzamos a reemplazarlos por la marca BMW, puesto que estaba completamente llena de la marca en sí. A continuación, aplicamos la estrategia de aplicar la moda para reemplazar valores nulos de las columnas en poca cantidad.

- Agrupamos marca y modelo en una nueva columna llamada ‘marca\_modelo’:



* Consideramos quitar los outliers de la columna de potencia, ya que nos parecían datos irreales de potencia. También vimos que hay outliers en la columna de km, como valores negativos. La columna de precios, si hemos mantenido los valores outliers, ya que consideramos que hay que estudiar con una mayor profundidad esos valores disparados de precios, con su modelo y los extras que tiene.
* Eliminamos estas columnas, ya que vimos que tienen el 50% o más de valores nulos. Las otras dos, ya hicimos una unión de ellas en una nueva columna, por lo tanto, los consideramos irrelevantes.

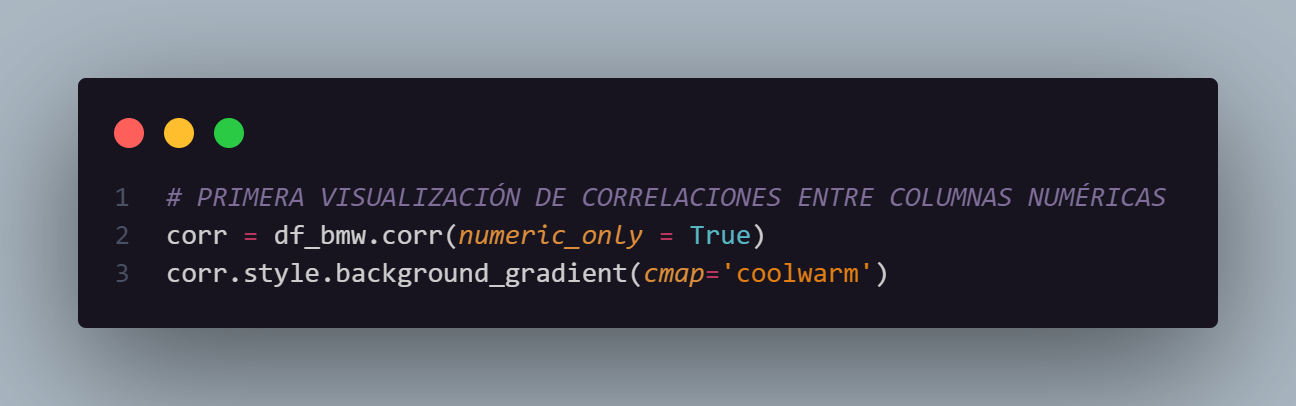


* Hemos hecho algunas agrupaciones como el promedio de km y potencia según el tipo de coche, la media de km según la marca\_modelo:

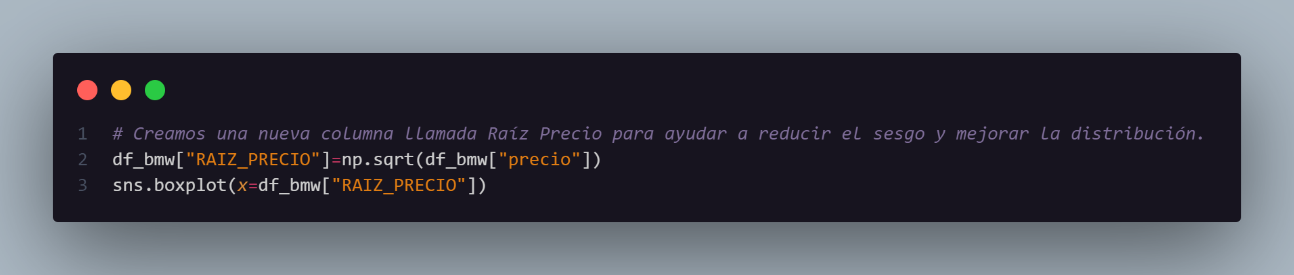


**ANÁLISIS UNIVARIABLE Y CORRELACIÓN INICIAL**

* Comenzamos con un análisis univariable inicial para ver la correlación de las columnas respecto al target.



* A continuación, hacemos diferentes gráficas para poder ver todas las columnas, y dependiendo del tipo de dato que sean, int float o object, que se representen con una gráfica en concreto. Imprimimos su frecuencia, y en otra gráfica relacionamos con el target. Esto lo hacemos con la función que hemos creado de visualize\_data.
* Realizamos un 'Boxplot' de la columna 'precio', donde observamos que el resultado está desviado hacia la parte izquierda de la gráfica, con puntos fuera de la media, lo que indica que el rango de precios tiene bastante variación. Para intentar reducir el sesgo y mejorar la distribución de la gráfica, probamos aplicar el logaritmo de 10, pero notamos que la variación se mantiene, aunque reflejada por el otro extremo. Por lo tanto, decidimos probar mediante la raíz cuadrada como una tercera opción.



**Transformación de Datos Categóricos y Numéricos**

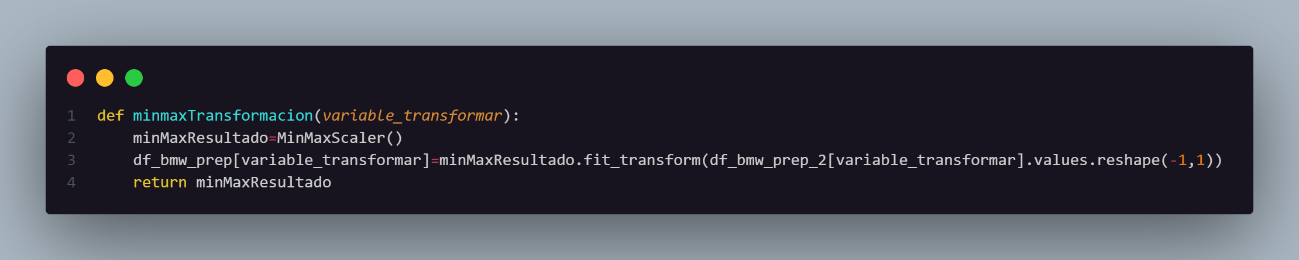
* Utilizamos una función para clasificar las columnas del dataset en numéricas, booleanas y categóricas. Para las categóricas, utilizamos el método de 'one-hot encoding', que es el proceso de convertir las variables de tipo 'object' en valores binarios (0 y 1). Este proceso separa cada valor de la columna en columnas individuales.



* "Convertimos las columnas booleanas 'volante\_regulable', 'aire\_acondicionado', 'camara\_trasera', 'elevalunas\_electrico', 'bluetooth', 'gps' , 'alerta\_limite\_velocidad' a tipo 'int' para transformarlas en valores numéricos, ya que originalmente están en un formato no numérico.

**Normalización de las Variables**

* Realizamos la normalización de la lista numérica creada anteriormente, transformando los valores de esas columnas en un rango específico entre 0 y 1.



**Correlación Final**

* En esta última parte, creamos una tabla final de correlación con el dataset modificado hasta ahora para analizar los valores resultantes y sus correlaciones.

