Instituto Politécnico Nacional

Escuela Superior de Cómputo

Práctica 1 y 2

Análisis exploratorio, limpieza de datos, transformación de datos, e ingeniería de características.

Alumno: López Fabián Jesús Manuel
Profesor: Flores Estrada Ituriel Enrique
Materia: Análitica y visualización de datos
Grupo: 5AV1

16 de marzo de 2024

1. Introducción

2. Análisis exploratorio de datos (EDA, Exploratory Data Analysis).

2.1. Datos generales.

1. Indicar todas las dimensiones del dataset, representadas por columnas.

Las dimensiones presentes en el dataset son las siguientes:

- S.No.
- Name
- Location
- Year
- Kilometers_Driven
- Fuel_Type
- Transmission
- Owner_Type
- Mileage
- Engine
- Power
- Seats
- New_Price
- Price

2. Desplegar los primeros 10 (top 10) registros del dataset.

	S.No.	Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	New_Price	Price
0	0	Maruti Wagon R LXI CNG	Mumbai	2010	72000	CNG	Manual	First	26.6 km/kg	998 CC	58.16 bhp	5.0	NaN	1.75
1	1	Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	Manual	First	19.67 kmpl	1582 CC	126.2 bhp	5.0	NaN	12.50
2	2	Honda Jazz V	Chennai	2011	46000	Petrol	Manual	First	18.2 kmpl	1199 CC	88.7 bhp	5.0	8.61 Lakh	4.50
3	3	Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	First	20.77 kmpl	1248 CC	88.76 bhp	7.0	NaN	6.00
4	4	Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Diesel	Automatic	Second	15.2 kmpl	1968 CC	140.8 bhp	5.0	NaN	17.74
5	5	Hyundai EON LPG Era Plus Option	Hyderabad	2012	75000	LPG	Manual	First	21.1 km/kg	814 CC	55.2 bhp	5.0	NaN	2.35
6	6	Nissan Micra Diesel XV	Jaipur	2013	86999	Diesel	Manual	First	23.08 kmpl	1461 CC	63.1 bhp	5.0	NaN	3.50
7	7	Toyota Innova Crysta 2.8 GX AT 8S	Mumbai	2016	36000	Diesel	Automatic	First	11.36 kmpl	2755 CC	171.5 bhp	8.0	21 Lakh	17.50
8	8	Volkswagen Vento Diesel Comfortline	Pune	2013	64430	Diesel	Manual	First	20.54 kmpl	1598 CC	103.6 bhp	5.0	NaN	5.20
9	9	Tata Indica Vista Quadrajet LS	Chennai	2012	65932	Diesel	Manual	Second	22.3 kmpl	1248 CC	74 bhp	5.0	NaN	1.95

Figura 1: Primeros 10 registros del dataset.

3. Indicar por cada dimensión si son numéricas o categóricas, lo que representan, y el tipo de dato utilizado para tal representación.

Nombre	Tipo	Descripción	Tipo de dato
S.No.	Numérico	Identificador único del registro	Int
Name	Categórico	Nombre del vehículo	String
Location	Categórico	Ubicación del vehículo	String
Year	Numérico	Año de fabricación	Int
Kilometers_Driven	Numérico	Kilómetros recorridos	String
Fuel_Type	Categórico	Tipo de combustible	String
Transmission	Categórico	Tipo de transmisión (automática o manual)	String
Owner_Type	Categórico	Tipo de dueño (primero, segundo, etc.)	String
Mileage	Numérico	Consumo de combustible (km/kg o kmpl)	String
Engine	Numérico	Cilindrada del motor expresada en centímetros cúbicos	String
Power	Numérico	Potencia del motor expresada en caballos de fuerza (bhp)	String
Seats	Nuérico	Número de asientos disponibles en el vehículo	Float
New_Price	Numérico	Precio del vehículo como nuevo (Lakh)	String
Price	Numérico	Precio de venta (Lakh)	String

Tabla 1: Diccionario de datos.

Podemos observar que hay dimensiones que representan tipos numéricos, pero están almacenadas como "String"(cadenas de caracteres). Por lo tanto, necesitamos realizar un tratamiento para convertir estas dimensiones a un tipo de dato numérico, lo cual realizaremos más adelante.

2.2. Análisis numérico.

1. Por cada dimensión, obtener la cantidad de observaciones, la media, la desviación estándar, y sus cuartiles (valor mínimo, 25 %, 50 %, 75 % y valor máximo)

En la Figura 2 se muestra, por cada dimensión numérica, el número de observaciones, la media, la desviación estándar, el valor mínimo, el primer cuartil, el segundo cuartil, el tercer cuartil y el valor máximo, respectivamente.

	S.No.	Year	Kilometers_Driven	Seats	Price
count	7253.000000	7253.000000	7.253000e+03	7200.000000	6019.000000
mean	3626.000000	2013.365366	5.869906e+04	5.279722	9.479468
std	2093.905084	3.254421	8.442772e+04	0.811660	11.187917
min	0.000000	1996.000000	1.710000e+02	0.000000	0.440000
25%	1813.000000	2011.000000	3.400000e+04	5.000000	3.500000
50%	3626.000000	2014.000000	5.341600e+04	5.000000	5.640000
75%	5439.000000	2016.000000	7.300000e+04	5.000000	9.950000
max	7252.000000	2019.000000	6.500000e+06	10.000000	160.000000

Figura 2: Estadísticos de las dimensiones numéricas.

2. Identificar por cada dimensión la cantidad y porcentaje de valores nulos.

La Figura 3 muestra la cantidad de valores nulos por dimensión, podemos observar 6 de nuestras 14 dimensiones tienen presencia de valores nulos. La Figura 4 muestra el porcentaje de valores nulos por dimensión.

	Nombre dimensión	Cantidad de valores nulos
0	S.No.	0
1	Name	0
2	Location	0
3	Year	0
4	Kilometers_Driven	0
5	Fuel_Type	0
6	Transmission	0
7	Owner_Type	0
8	Mileage	2
9	Engine	46
10	Power	46
11	Seats	53
12	Price	1234
13	New_Price	6247

Figura 3: Estadísticos de las dimensiones numéricas.

	Nombre dimensión	Porcentaje de valores nulos (%)
0	S.No.	0.00
1	Name	0.00
2	Location	0.00
3	Year	0.00
4	Kilometers_Driven	0.00
5	Fuel_Type	0.00
6	Transmission	0.00
7	Owner_Type	0.00
8	Mileage	0.03
9	Engine	0.63
10	Power	0.63
11	Seats	0.73
12	Price	17.01
13	New_Price	86.13

Figura 4: Estadísticos de las dimensiones numéricas.

3. Identificar por cada dimensión la cantidad valores duplicados..

Podemos observar en la Figura 5 que la única dimensión que no tiene valores duplicados es "S.No.", ya que representa un identificador único por cada observación en el conjunto de datos.

	Dimension	Num_duplicados
0	S.No.	0
4	Kilometers_Driven	3593
1	Name	5212
13	Price	5879
12	New_Price	6627
8	Mileage	6802
10	Power	6866
9	Engine	7102
3	Year	7230
2	Location	7242
11	Seats	7243
5	Fuel_Type	7248
7	Owner_Type	7249
6	Transmission	7251

Figura 5: Cantidad de duplicados por cada dimensión.

2.3. Análisis gráfico.

1. Grafique la distribución de cada una de las dimensiones numéricas.

Como se menciono anteriormente, contamos con dimensiones numéricas pero están almacenadas como "Stringz contienen caracteres, por lo que aun no podemos considerarlas para graficar su distribución. No se graficará la dimensión 'S.No.' ya que al ser un identificador único para cada observación no brinda información relevante. Hasta este punto las variables numéricas son las siguientes:

- Year
- Kilometers_Driven
- Seats

- Price

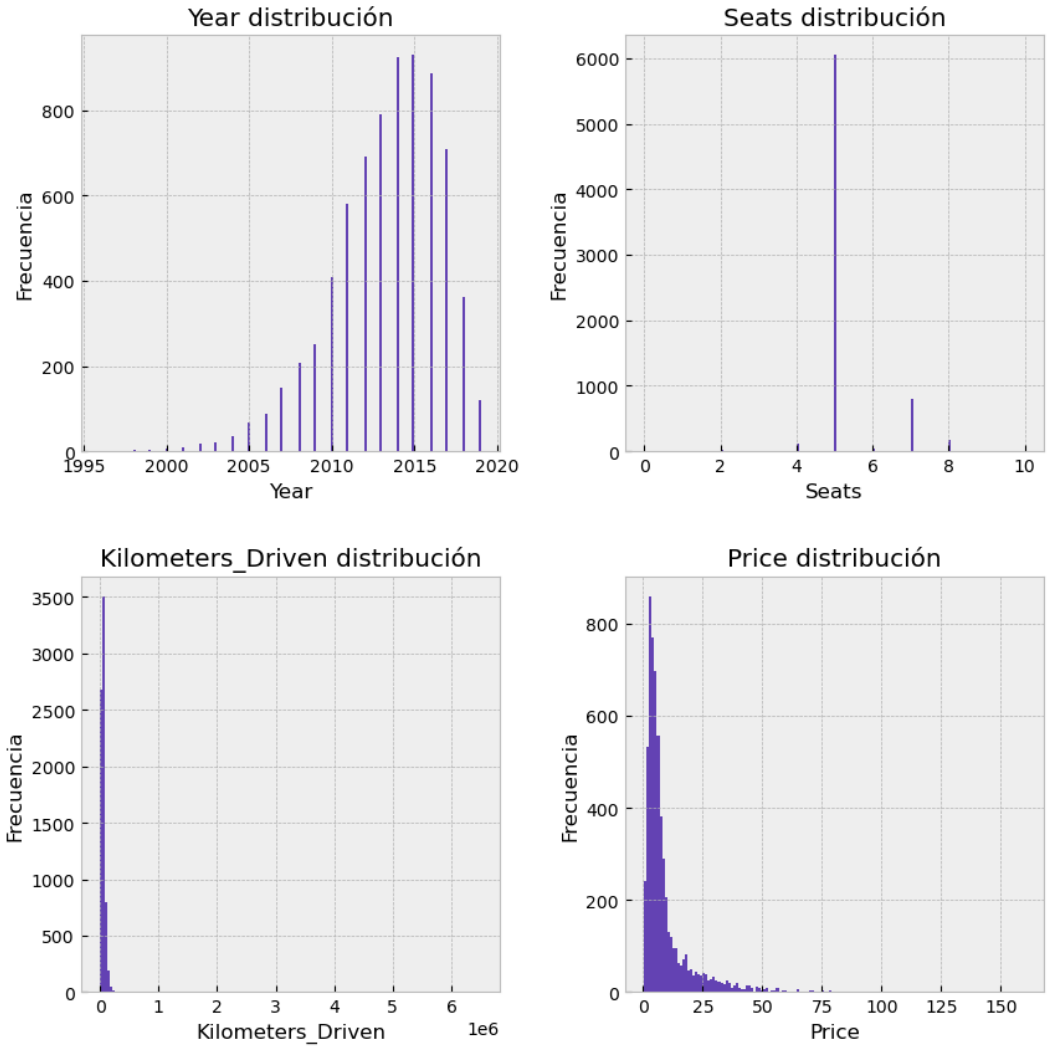

Figura 6: Distribución de dimensiones numéricas.

Al observar las distribuciones de las variables podemos notar lo siguiente:

- De la gráfica "Year distribución"que los años de fabricación más frecuentes de los automóviles son de 2013 a 2017.
- La gran mayoría de los automóviles tienen 5 asientos disponibles.
- Gran parte de los automóviles han recorrido menos de un millón de kilómetros.
- La mayoría de los automóviles tienen un precio menor de 25 Lakh.

2. Compare entre sí cada una de las dimensiones numéricas (análisis bivariado) con un gráfico del tipo "pairplot".

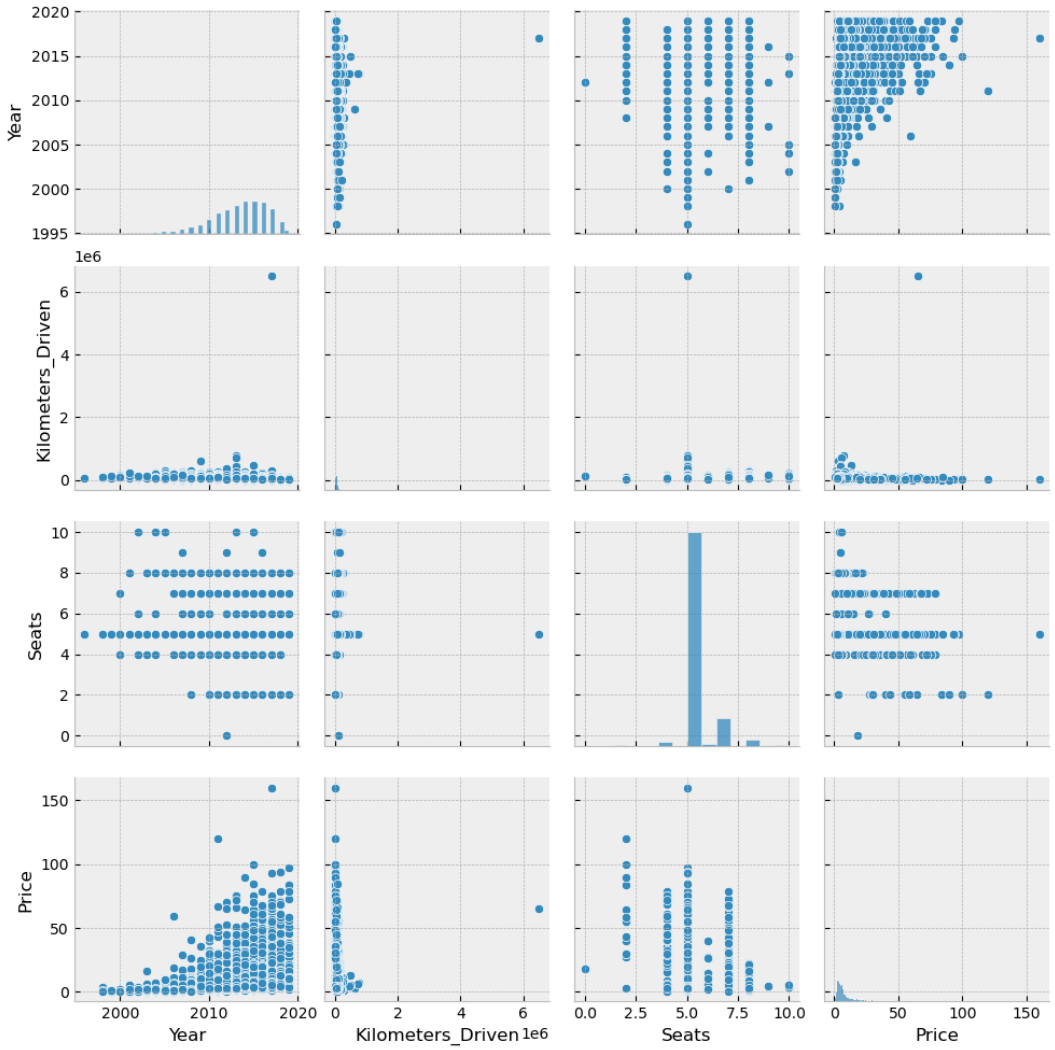

Figura 7: Graficos dispersión.

El gráfico de dispersión (Figura 7) nos ofrece la posibilidad de visualizar la existencia de alguna correlación (negativa, positiva o nula) entre diferentes pares de dimensiones. Observamos que las dimensiones 'Seats' y 'Kilometers_Driven' aparentan tener una correlación nula, ya que no se percibe ningún patrón que indique el comportamiento de la variable dependiente en relación con el aumento de la variable independiente.

Por el contrario, las dimensiones 'Year' y 'Price' parecen estar altamente correlacionadas, ya que al haber un aumento en la variable independiente (Year) aumenta la variable dependiente (Price). Esto indica una correlación positiva.

3. Realice un mapa de calor para identificar la correlación entre todas las variables (análisis multivariado).

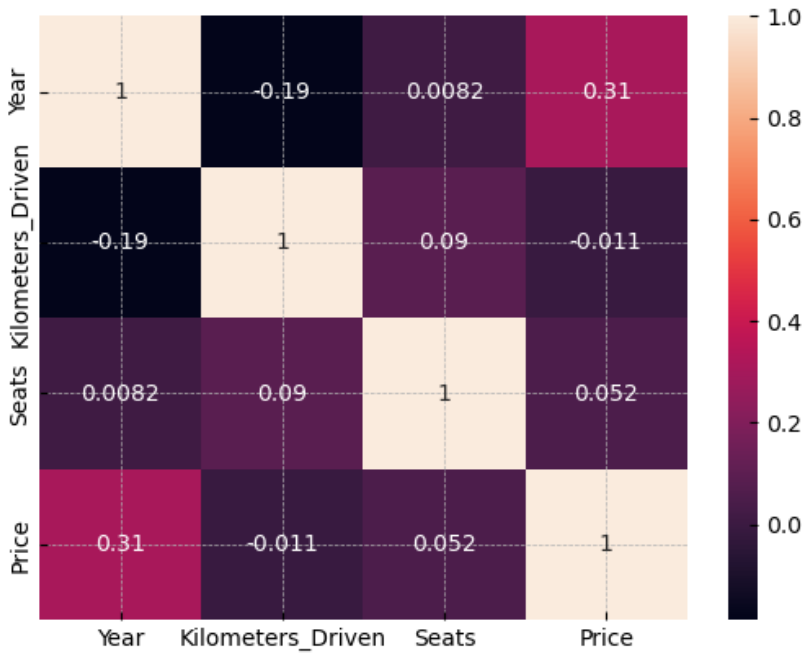

Figura 8: Mapa de calor para correlación.

Con este gráfico de calor, podemos visualizar los coeficientes de correlación de Pearson entre cada par de dimensiones. Como se mencionó en el punto anterior, las dimensiones 'Seats' y 'Kilometers_Driven' tienen una correlación muy cercana a 0, lo cual indica una correlación positiva muy débil. Por otro lado, las dimensiones 'Year' y 'Price' son las que tienen un coeficiente de correlación positivo más fuerte, mientras que 'Year' y 'Kilometers_Driven' tienen el coeficiente de correlación negativo más fuerte.

3. Grafique la distribución de cada una de las dimensiones categóricas

No se considero a la dimensión 'Name', ya que al tener muchos nombres de vehiculos diferentes se generaran muchas barras con muy pocos valores cada una.

Las variables categoricas consideradas fueron:

- Location
- Transmission
- Fuel_Type
- Owner_Type

La frecuencia de cada clase para cada dimensión se muestra en la Figura 9.

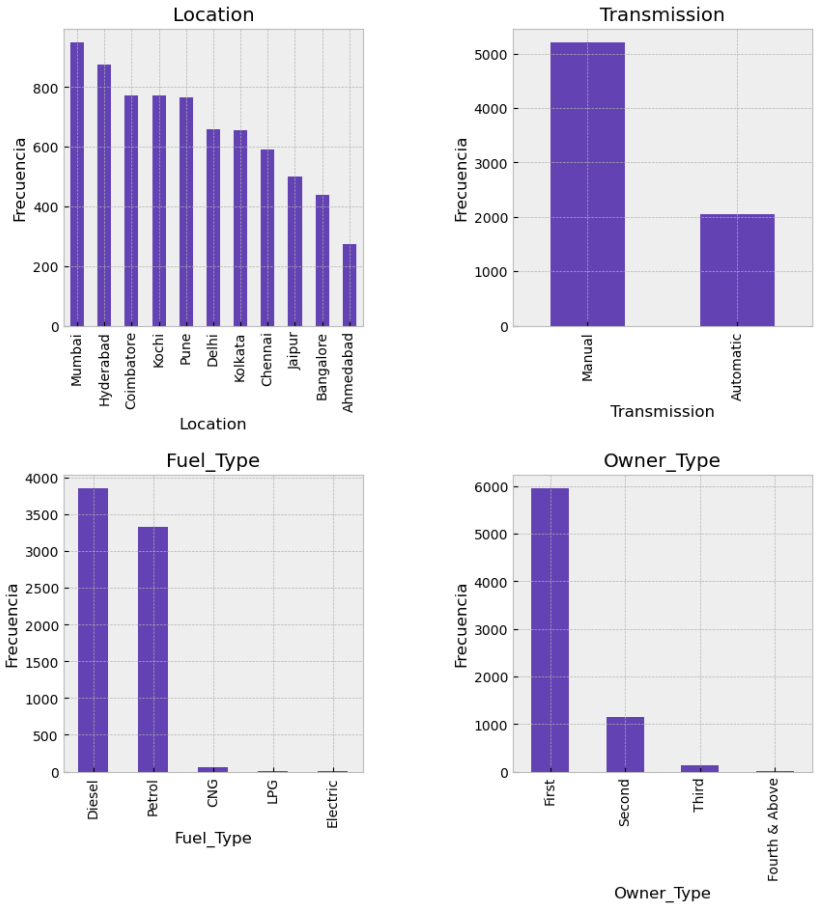

Figura 9: Grafico de barras para variables categóricas.

Al observar las distribuciones de las clases de cada dimensión (Figura 9) podemos notar lo siguiente:

- La ubicación más frecuente de los automóviles es Mumbai.
- La ubicación menos frecuente de los automóviles es Ahmedabad.
- En nuestros registros hay más automóviles con transmisión automática que estándar.
- El combustible mas frecuente de los automóviles es diesel.
- La mayor parte de los automóviles solo han tenido un solo dueño.

4. Compare cada una de las dimensiones categóricas (eje X) contra la dimensión “Precio” (eje y). Para ello debe primero agrupar por categoría, después obtener su media, y finalmente ordenar los valores de mayor a menor.

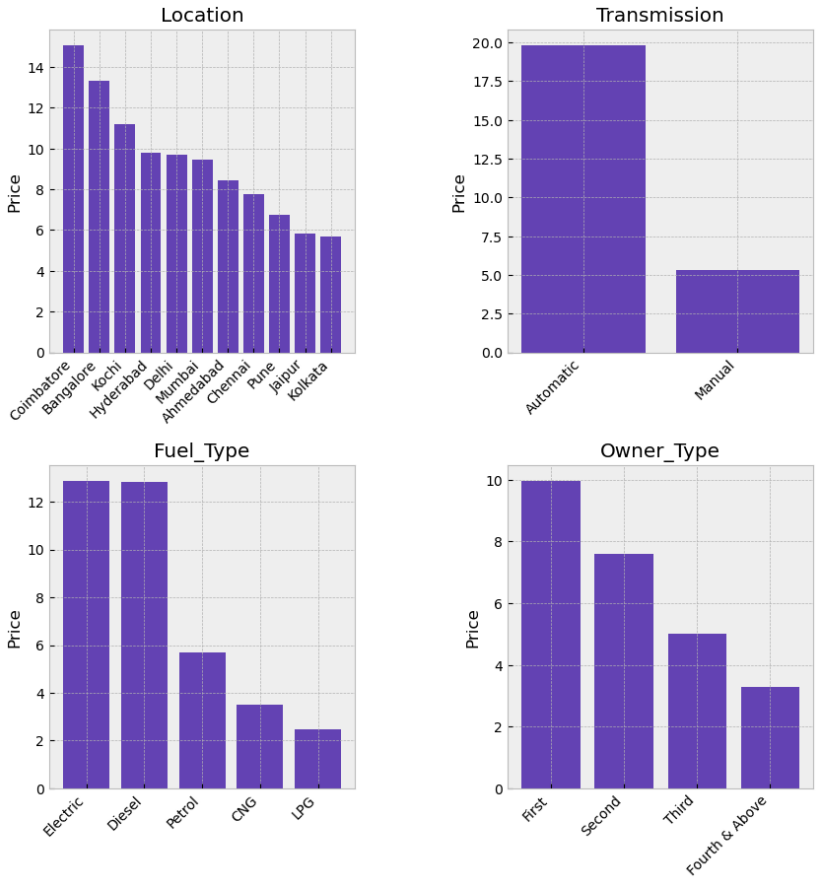

Figura 10: Comparación de dimensiones categóricas contra 'Precio'

Podemos notar lo siguiente:

- Los automóviles con transmisión automática en promedio son mas caros que los que tienen transmisión manual.
- Los automóviles que solo han tenido un dueño tienen un precio mas elevado.
- Los automóviles que usan Diesel o son eléctricos en promedio son los más caros y tienden a tener un precio similar.

3. Limpieza de datos

1. Elimine las dimensiones "S. No." Y "New_Price".

Como mencionamos anteriormente, "S. No." no proporciona información relevante, ya que se trata de un identificador único para cada observación en el dataset. Por otro lado, el 86.13 % de las observaciones en la dimensión 'New_Price' representan valores nulos, por lo que eliminaremos estas dos dimensiones.

	S.No.	Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	New_Price	Price
0	0	Maruti Wagon R LXI CNG	Mumbai	2010	72000	CNG	Manual	First	26.6 km/kg	998 CC	58.16 bhp	5.0	NaN	1.75
1	1	Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	Manual	First	19.67 kmpl	1582 CC	126.2 bhp	5.0	NaN	12.50
2	2	Honda Jazz V	Chennai	2011	46000	Petrol	Manual	First	18.2 kmpl	1199 CC	88.7 bhp	5.0	8.61 Lakh	4.50
3	3	Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	First	20.77 kmpl	1248 CC	88.76 bhp	7.0	NaN	6.00
4	4	Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Diesel	Automatic	Second	15.2 kmpl	1968 CC	140.8 bhp	5.0	NaN	17.74

Figura 11: Primeras 5 observaciones antes de eliminar las dimensiones.

	Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	Price
0	Maruti Wagon R LXI CNG	Mumbai	2010	72000	CNG	Manual	First	26.6 km/kg	998 CC	58.16 bhp	5.0	1.75
1	Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	Manual	First	19.67 kmpl	1582 CC	126.2 bhp	5.0	12.50
2	Honda Jazz V	Chennai	2011	46000	Petrol	Manual	First	18.2 kmpl	1199 CC	88.7 bhp	5.0	4.50
3	Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	First	20.77 kmpl	1248 CC	88.76 bhp	7.0	6.00
4	Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Diesel	Automatic	Second	15.2 kmpl	1968 CC	140.8 bhp	5.0	17.74

Figura 12: Primeras 5 observaciones después de eliminar las dimensiones.

Nuestro nuevo conjunto de datos (Figura 12) cuenta con 12 dimensiones.

2. Sustituya el nombre en inglés de cada dimensión por su traducción en español.

La Tabla 2 muestra los nombres de las columnas antes y después de su traducción. Se evito evito usar caracteres con acentos y 'ñ'.

Antes	Después
Location	Ubicacion
Year	Anio
Kilometers_Driven	Kilometros_Recorridos
Fuel_Type	Tipo_Combustible
Transmission	Transmision
Owner_Type	Tipo_Propietario
Mileage	Millaje
Engine	Motor
Power	Potencia
Seats	Asientos
Price	Precio

Tabla 2: Traducción del nombre de las dimensiones.

	Nombre	Ubicacion	Año	Kilometros_Recorridos	Tipo_Combustible	Transmision	Tipo_propietario	Millaje	Motor	Potencia	Asientos	Precio
0	Maruti Wagon R LXI CNG	Mumbai	2010	72000	CNG	Manual	First	26.6 km/kg	998 CC	58.16 bhp	5.0	1.75
1	Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	Manual	First	19.67 kmpl	1582 CC	126.2 bhp	5.0	12.50
2	Honda Jazz V	Chennai	2011	46000	Petrol	Manual	First	18.2 kmpl	1199 CC	88.7 bhp	5.0	4.50
3	Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	First	20.77 kmpl	1248 CC	88.76 bhp	7.0	6.00
4	Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Diesel	Automatic	Second	15.2 kmpl	1968 CC	140.8 bhp	5.0	17.74

Figura 12: Primeras 5 observaciones después de traducir los nombres de las dimensiones.

3. Elimine todas las observaciones que únicamente tengan valores nulos en la dimensión "Asientos". Es decir, hay observaciones con valores nulos en dicha dimensión, en la dimensión "Potencia" y en otras más. Estos casos no deben ser eliminados.

Al aplicar una función *'eliminar_nulos_unicamente_dimension'* que elimina una observación solo si tiene un valor nulo en 'Asientos' se observa que no hay un cambio antes (Figura 13) y después (Figura 14) de hacer esta tarea. Por lo que no hay observación que tenga valores nulos únicamente en la dimensión 'Asientos'.

	Dimension	Cantidad nulos
0	Nombre	0
1	Ubicacion	0
2	Anio	0
3	Kilometros_Recorridos	0
4	Tipo_Combustible	0
5	Transmision	0
6	Tipo_Propietario	0
7	Millaje	2
8	Motor	46
9	Potencia	46
10	Asientos	53
11	Precio	1234

Figura 13: Valores nulos por dimensión antes de aplicar la función.

	Dimension	Cantidad nulos
0	Nombre	0
1	Ubicacion	0
2	Anio	0
3	Kilometros_Recorridos	0
4	Tipo_Combustible	0
5	Transmision	0
6	Tipo_Propietario	0
7	Millaje	2
8	Motor	46
9	Potencia	46
10	Asientos	53
11	Precio	1234

Figura 14: Valores nulos por dimensión después de aplicar la función.

4. Sustituya todos los valores nulos en la dimensión “Millaje” por la media de la dimensión.

Para llevar a cabo esta tarea, necesitamos modificar el tipo de dato en la dimensión 'Millaje', ya que contiene caracteres. Esto implica eliminar los caracteres presentes en la dimensión y convertirlos a un valor *'float'*. Sin embargo, dado que hay diferentes unidades que miden el consumo de combustible, necesitamos transformar todas las observaciones de esta dimensión a una misma unidad. Las dos unidades diferentes son km/kg (kilómetro por kilogramo) y kmpl (kilómetro por litro).

Para convertir a kmpl, utilizamos la siguiente fórmula:

$$kmpl = \frac{km/kg}{densidad\ del\ combustible(kg/l)}$$

En las observaciones que tenemos unidades de kg/kg solo tenemos dos posibles combustibles, CNG y LPG, las densidades de dichos combustibles se encuentran en la Tabla 3:

Combustible	Densidad
CNG	0.13 kg/l
LPG	0.58 kg/l

Tabla 3: Densidades de combustibles.

	Nombre	Ubicacion	Anio	Kilometros_Recorridos	Tipo_Combustible	Transmision	Tipo_Propietario	Millaje	Motor	Potencia	Asientos	Precio
0	Maruti Wagon R LXI CNG	Mumbai	2010	72000	CNG	Manual	First	204.615385	998 CC	58.16 bhp	5.0	1.75
1	Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	Manual	First	19.6	1582 CC	126.2 bhp	5.0	12.50
2	Honda Jazz V	Chennai	2011	46000	Petrol	Manual	First	18.2	1199 CC	88.7 bhp	5.0	4.50
3	Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	First	20.7	1248 CC	88.76 bhp	7.0	6.00
4	Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Diesel	Automatic	Second	15.2	1968 CC	140.8 bhp	5.0	17.74

Figura 15: Primeras 5 observaciones después de transformar la dimensión 'Millaje'.

Una vez que la dimensión ya no tiene caracteres y es numérica, ya podemos hacer la imputación.

	Dimension	Cantidad nulos
0	Nombre	0
1	Ubicacion	0
2	Anio	0
3	Kilometros_Recorridos	0
4	Tipo_Combustible	0
5	Transmision	0
6	Tipo_Propietario	0
7	Millaje	2
8	Motor	46
9	Potencia	46
10	Asientos	53
11	Precio	1234

Figura 16: Valores nulos por dimensión antes de imputar media de la dimensión en 'Millaje'.

	Dimension	Cantidad nulos
0	Nombre	0
1	Ubicacion	0
2	Anio	0
3	Kilometros_Recorridos	0
4	Tipo_Combustible	0
5	Transmision	0
6	Tipo_Propietario	0
7	Millaje	0
8	Motor	46
9	Potencia	46
10	Asientos	53
11	Precio	1234

Figura 17: Valores nulos por dimensión después de imputar media de la dimensión en 'Millaje'.

Podemos notar que antes de hacer la imputación (Figura 16), tenemos dos valores nulos en la dimensión 'Millaje', despues de imputar la media (Figura 17) ya no tenemos presencia de nulos en la dimensión.

5. Sustituya todos los valores nulos en la dimensión “Motor” por la moda de la dimensión.

	Dimension	Cantidad nulos
0	Nombre	0
1	Ubicacion	0
2	Anio	0
3	Kilometros_Recorridos	0
4	Tipo_Combustible	0
5	Transmision	0
6	Tipo_Propietario	0
7	Millaje	0
8	Motor	46
9	Potencia	46
10	Asientos	53
11	Precio	1234

Figura 18: Valores nulos por dimensión antes de aplicar la función.

	Dimension	Cantidad nulos
0	Nombre	0
1	Ubicacion	0
2	Anio	0
3	Kilometros_Recorridos	0
4	Tipo_Combustible	0
5	Transmision	0
6	Tipo_Propietario	0
7	Millaje	0
8	Motor	0
9	Potencia	46
10	Asientos	53
11	Precio	1234

Figura 19: Valores nulos por dimensión antes de aplicar la función.

Notamos que pasamos de 46 valores nulos en la dimensión 'Motor' (Figura 18) a cero (Figura 19).

4. Transformación de datos

1. Sustituya los registros que contengan las siguientes palabras en la dimensión "Nombre" por la respectiva cadena sustituto: ÏSUZU"por Ïsuzu", "Mini"por "MiniCooper", y "Land"por "LandRover".

Al visualizar las observaciones que tenían contenido la palabra 'Mini' en su nombre (Figura 20), se noto que había dos valores posibles: 'Mini' y 'Mini Cooper'. Entonces, esos valores serán sustituidos por 'MiniCooper'.

```
Mini Countryman Cooper D
Mini Cooper Convertible S
Mini Clubman Cooper S
Mini Cooper Convertible 1.6
Mini Cooper Convertible S
Mini Cooper 5 DOOR D
Mini Cooper S
Mini Cooper 5 DOOR D
Mini Cooper 5 DOOR D
Mini Cooper Countryman D
Mini Cooper 3 DOOR D
Mini Cooper Countryman D High
Mini Cooper 5 DOOR D
Mini Cooper Countryman D High
Mini Cooper Convertible S
Mini Cooper 3 DOOR S
Mini Cooper Convertible S
Mini Cooper Convertible S
Mini Cooper S Carbon Edition
Mini Cooper 5 DOOR D
Mini Cooper 3 DOOR D
Mini Cooper 3 DOOR D
Mini Cooper Countryman D
Mini Cooper S Carbon Edition
Mini Cooper S Carbon Edition
```

Figura 20: Algunos nombres de vehículos con la palabra 'Mini' contenida.

	Nombre	Ubicacion	Anio	Kilometros_Recorridos	Tipo_Combustible	Transmision	Tipo_Propietario	Millaje	Motor	Potencia	Asientos	Precio
13	LandRover Range Rover 2.2L Pure	Delhi	2014	72000	Diesel	Automatic	First	12.7	2179 CC	187.7 bhp	5.0	27.00
14	LandRover Freelander 2 TD4 SE	Pune	2012	85000	Diesel	Automatic	Second	0.0	2179 CC	115 bhp	5.0	17.50
176	MiniCooperCooper Countryman Cooper D	Jaipur	2017	8525	Diesel	Automatic	Second	16.6	1998 CC	112 bhp	5.0	23.00
191	LandRover Range Rover 2.2L Dynamic	Coimbatore	2018	36091	Diesel	Automatic	First	12.7	2179 CC	187.7 bhp	5.0	55.76
228	MiniCooperCooper Convertible S	Kochi	2017	26327	Petrol	Automatic	First	16.8	1998 CC	189.08 bhp	4.0	35.67
...	...	...	...	...	...	...	...	...	...	...	...	...
6919	Isuzu D-MAX V-Cross 4X4	Jaipur	2017	290000	Diesel	Manual	First	12.4	2499 CC	134 bhp	5.0	NaN
7132	MiniCooperCooper Clubman Cooper S	Pune	2017	2890	Petrol	Manual	First	13.8	1998 CC	192 bhp	5.0	NaN
7157	LandRover Range Rover 2.2L Pure	Hyderabad	2015	49000	Diesel	Automatic	Second	12.7	2179 CC	187.7 bhp	5.0	NaN
7160	MiniCooperCooper Countryman D	Hyderabad	2013	50000	Diesel	Automatic	First	23.8	1998 CC	112 bhp	5.0	NaN
7198	LandRover Discovery 4 TDV6 Auto Diesel	Hyderabad	2012	147202	Diesel	Automatic	First	11.8	2993 CC	241.6 bhp	7.0	NaN

Figura 21: Nombres corregidos.

Estos son algunos registros (Figura 21) con los nombres corregidos.

2. Elimine los caracteres de las dimensiones “Millas por Galón”, “Motor”, y “Potencia”.

Los caracteres de la dimensión 'Millas por Galón' ya fueron eliminados en el paso 2.4.

	Nombre	Ubicacion	Anio	Kilometros_Recorridos	Tipo_Combustible	Transmision	Tipo_Propietario	Millaje	Motor	Potencia	Asientos	Precio
0	Maruti Wagon R LXI CNG	Mumbai	2010	72000	CNG	Manual	First	204.615385	998	58.16	5.0	1.75
1	Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	Manual	First	19.600000	1582	126.2	5.0	12.50
2	Honda Jazz V	Chennai	2011	46000	Petrol	Manual	First	18.200000	1199	88.7	5.0	4.50
3	Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	First	20.700000	1248	88.76	7.0	6.00
4	Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Diesel	Automatic	Second	15.200000	1968	140.8	5.0	17.74
5	Hyundai EON LPG Era Plus Option	Hyderabad	2012	75000	LPG	Manual	First	36.379310	814	55.2	5.0	2.35
6	Nissan Micra Diesel XV	Jaipur	2013	86999	Diesel	Manual	First	23.000000	1461	63.1	5.0	3.50
7	Toyota Innova Crysta 2.8 GX AT 8S	Mumbai	2016	36000	Diesel	Automatic	First	11.300000	2755	171.5	8.0	17.50
8	Volkswagen Vento Diesel Comfortline	Pune	2013	64430	Diesel	Manual	First	20.500000	1598	103.6	5.0	5.20
9	Tata Indica Vista Quadrajet LS	Chennai	2012	65932	Diesel	Manual	Second	22.300000	1248	74	5.0	1.95

Figura 22: Primeros 10 registros después de eliminar caracteres de las dimensiones 'Motor' y 'Potencia'.

3. Redondee al número entero más cercano los valores de las dimensiones “Millas por Galón”, “Motor” y “Potencia”.

	Nombre	Ubicacion	Anio	Kilometros_Recorridos	Tipo_Combustible	Transmision	Tipo_Propietario	Millaje	Motor	Potencia	Asientos	Precio
0	Maruti Wagon R LXI CNG	Mumbai	2010	72000	CNG	Manual	First	205.0	998	58.16	5.0	1.75
1	Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	Manual	First	20.0	1582	126.2	5.0	12.50
2	Honda Jazz V	Chennai	2011	46000	Petrol	Manual	First	18.0	1199	88.7	5.0	4.50
3	Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	First	21.0	1248	88.76	7.0	6.00
4	Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Diesel	Automatic	Second	15.0	1968	140.8	5.0	17.74
5	Hyundai EON LPG Era Plus Option	Hyderabad	2012	75000	LPG	Manual	First	36.0	814	55.2	5.0	2.35
6	Nissan Micra Diesel XV	Jaipur	2013	86999	Diesel	Manual	First	23.0	1461	63.1	5.0	3.50
7	Toyota Innova Crysta 2.8 GX AT 8S	Mumbai	2016	36000	Diesel	Automatic	First	11.0	2755	171.5	8.0	17.50
8	Volkswagen Vento Diesel Comfortline	Pune	2013	64430	Diesel	Manual	First	20.0	1598	103.6	5.0	5.20
9	Tata Indica Vista Quadrajet LS	Chennai	2012	65932	Diesel	Manual	Second	22.0	1248	74	5.0	1.95

Figura 22: Primeros 10 registros después de redondear la dimensiones.

4. Multiplique por mil los valores en la dimensión “Precio”.

	Nombre	Ubicacion	Anio	Kilometros_Recorridos	Tipo_Combustible	Transmision	Tipo_Propietario	Millaje	Motor	Potencia	Asientos	Precio
0	Maruti Wagon R LXI CNG	Mumbai	2010	72000	CNG	Manual	First	205.0	998	58.16	5.0	175.0
1	Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	Manual	First	20.0	1582	126.2	5.0	1250.0
2	Honda Jazz V	Chennai	2011	46000	Petrol	Manual	First	18.0	1199	88.7	5.0	450.0
3	Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	First	21.0	1248	88.76	7.0	600.0
4	Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Diesel	Automatic	Second	15.0	1968	140.8	5.0	1774.0
5	Hyundai EON LPG Era Plus Option	Hyderabad	2012	75000	LPG	Manual	First	36.0	814	55.2	5.0	235.0
6	Nissan Micra Diesel XV	Jaipur	2013	86999	Diesel	Manual	First	23.0	1461	63.1	5.0	350.0
7	Toyota Innova Crysta 2.8 GX AT 8S	Mumbai	2016	36000	Diesel	Automatic	First	11.0	2755	171.5	8.0	1750.0
8	Volkswagen Vento Diesel Comfortline	Pune	2013	64430	Diesel	Manual	First	20.0	1598	103.6	5.0	520.0
9	Tata Indica Vista Quadrajet LS	Chennai	2012	65932	Diesel	Manual	Second	22.0	1248	74	5.0	195.0

Figura 23: Primeros 10 registros después de multiplicar por mil la dimension 'Precio'.

5. Obtenga el logaritmo de las dimensiones “Precio” y “Kilómetros”.

Se aplico a ambas dimensiones un logaritmo base 10.

	Nombre	Ubicacion	Anio	Kilometros_Recorridos	Tipo_Combustible	Transmision	Tipo_Propietario	Millaje	Motor	Potencia	Asientos	Precio
0	Maruti Wagon R LXI CNG	Mumbai	2010	72000	CNG	Manual	First	205.0	998	58.16	5.0	175.0
1	Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	Manual	First	20.0	1582	126.2	5.0	1250.0
2	Honda Jazz V	Chennai	2011	46000	Petrol	Manual	First	18.0	1199	88.7	5.0	450.0
3	Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	First	21.0	1248	88.76	7.0	600.0
4	Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Diesel	Automatic	Second	15.0	1968	140.8	5.0	1774.0
5	Hyundai EON LPG Era Plus Option	Hyderabad	2012	75000	LPG	Manual	First	36.0	814	55.2	5.0	235.0
6	Nissan Micra Diesel XV	Jaipur	2013	86999	Diesel	Manual	First	23.0	1461	63.1	5.0	350.0
7	Toyota Innova Crysta 2.8 GX AT 8S	Mumbai	2016	36000	Diesel	Automatic	First	11.0	2755	171.5	8.0	1750.0
8	Volkswagen Vento Diesel Comfortline	Pune	2013	64430	Diesel	Manual	First	20.0	1598	103.6	5.0	520.0
9	Tata Indica Vista Quadrajet LS	Chennai	2012	65932	Diesel	Manual	Second	22.0	1248	74	5.0	195.0

Figura 24: Primeros 10 registros después de aplicar logaritmo base 10 a ambas dimensiones.

5. Ingeniería de características (Feature Engineering)

1. Crear una nueva dimensión llamada antigüedad, la cual se debe calcular a partir del año de fabricación del auto ("Year").

	Nombre	Ubicacion	Anio	Kilometros_Recorridos	Tipo_Combustible	Transmision	Tipo_Propietario	Millaje	Motor	Potencia	Asientos	Precio	Antiguedad
0	Maruti Wagon R LXI CNG	Mumbai	2010	4.857332	CNG	Manual	First	205.0	998	58.16	5.0	3.243038	14
1	Hyundai Creta 1.6 CRDi SX Option	Pune	2015	4.612784	Diesel	Manual	First	20.0	1582	126.2	5.0	4.096910	9
2	Honda Jazz V	Chennai	2011	4.662758	Petrol	Manual	First	18.0	1199	88.7	5.0	3.653213	13
3	Maruti Ertiga VDI	Chennai	2012	4.939519	Diesel	Manual	First	21.0	1248	88.76	7.0	3.778151	12
4	Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	4.609274	Diesel	Automatic	Second	15.0	1968	140.8	5.0	4.248954	11
5	Hyundai EON LPG Era Plus Option	Hyderabad	2012	4.875061	LPG	Manual	First	36.0	814	55.2	5.0	3.371068	12
6	Nissan Micra Diesel XV	Jaipur	2013	4.939514	Diesel	Manual	First	23.0	1461	63.1	5.0	3.544068	11
7	Toyota Innova Crysta 2.8 GX AT 8S	Mumbai	2016	4.556303	Diesel	Automatic	First	11.0	2755	171.5	8.0	4.243038	8
8	Volkswagen Vento Diesel Comfortline	Pune	2013	4.809088	Diesel	Manual	First	20.0	1598	103.6	5.0	3.716003	11
9	Tata Indica Vista Quadrajet LS	Chennai	2012	4.819096	Diesel	Manual	Second	22.0	1248	74	5.0	3.290035	12

Figura 25: Primeros 10 registros después de aplicar agregar la dimensión antigüedad.

2. Crear dos nuevas dimensiones llamadas "Marca" y "Modelo" a partir de la dimensión "Nombre". El contenido de "Marca" debe ser la primera palabra de la dimensión "Nombre"; por su parte, "Modelo" debe ser el resultado de concatenar sin espacios en blanco la segunda y tercera palabra de la dimensión "Nombre".

Por la transformación que se realizo en el paso 3.1 tenemos valores con solo 2 palabras, por lo que para esos casos, se considero que la 'Marca' sera la primera letra y el 'Modelo' la segunda.

	Nombre	Ubicacion	Anio	Kilometros_Recorridos	Tipo_Combustible	Transmision	Tipo_Propietario	Millaje	Motor	Potencia	Asientos	Precio	Antiguedad	Marca	Modelo
0	Maruti Wagon R LXI CNG	Mumbai	2010	4.857332	CNG	Manual	First	205.0	998	58.16	5.0	3.243038	14	Maruti	WagonR
1	Hyundai Creta 1.6 CRDi SX Option	Pune	2015	4.612784	Diesel	Manual	First	20.0	1582	126.2	5.0	4.096910	9	Hyundai	Creta1.6
2	Honda Jazz V	Chennai	2011	4.662758	Petrol	Manual	First	18.0	1199	88.7	5.0	3.653213	13	Honda	JazzV
3	Maruti Ertiga VDI	Chennai	2012	4.939519	Diesel	Manual	First	21.0	1248	88.76	7.0	3.778151	12	Maruti	ErtigaVDI
4	Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	4.609274	Diesel	Automatic	Second	15.0	1968	140.8	5.0	4.248954	11	Audi	A4New
5	Hyundai EON LPG Era Plus Option	Hyderabad	2012	4.875061	LPG	Manual	First	36.0	814	55.2	5.0	3.371068	12	Hyundai	EONLPG
6	Nissan Micra Diesel XV	Jaipur	2013	4.939514	Diesel	Manual	First	23.0	1461	63.1	5.0	3.544068	11	Nissan	MicraDiesel
7	Toyota Innova Crysta 2.8 GX AT 8S	Mumbai	2016	4.556303	Diesel	Automatic	First	11.0	2755	171.5	8.0	4.243038	8	Toyota	InnovaCrysta
8	Volkswagen Vento Diesel Comfortline	Pune	2013	4.809088	Diesel	Manual	First	20.0	1598	103.6	5.0	3.716003	11	Volkswagen	VentoDiesel
9	Tata Indica Vista Quadrajet LS	Chennai	2012	4.819096	Diesel	Manual	Second	22.0	1248	74	5.0	3.290035	12	Tata	IndicaVista

Figura 26: Primeros 10 registros después de agregar las dimensiones 'Marca' y 'Modelo'.

6. Análisis exploratorio de datos posterior a etapas de limpieza, transformación, e ingeniería de características

6.1. Análisis numérico.

1. Por cada dimensión, obtener la cantidad de observaciones, la media, la desviación estándar, y sus cuartiles (valor mínimo, 25 %, 50 %, 75 % y valor máximo)

	Anio	Kilometros_Recorridos	Millaje	Motor	Potencia	Asientos	Precio	Antiguedad
count	7253.000000	7253.000000	7253.000000	7253.000000	7078.000000	7200.000000	6019.000000	7253.000000
mean	2013.365366	4.673433	18.968427	1613.912450	112.768014	5.279722	3.792629	10.634634
std	3.254421	0.311097	12.776516	594.328359	53.482602	0.811660	0.379599	3.254421
min	1996.000000	2.232996	0.000000	72.000000	34.000000	0.000000	2.643453	5.000000
25%	2011.000000	4.531479	15.000000	1197.000000	75.000000	5.000000	3.544068	8.000000
50%	2014.000000	4.727671	18.000000	1462.000000	94.000000	5.000000	3.751279	10.000000
75%	2016.000000	4.863323	21.000000	1968.000000	138.000000	5.000000	3.997823	13.000000
max	2019.000000	6.812913	205.000000	5998.000000	616.000000	10.000000	5.204120	28.000000

Figura 27: Estadísticos de las dimensiones numéricas.

En la Figura 27 notamos que hay un cambio de escala en la dimensión 'Kilometros_Recorridos', ahora los valores son más pequeños, resultado de aplicar logaritmo a la dimensión. También contamos con dos nuevas dimensiones numéricas: 'Millaje' y 'Antiguedad'.

2. Identificar por cada dimensión la cantidad y porcentaje de valores nulos.

	Nombre dimensión	Cantidad de valores nulos
0	Nombre	0
1	Ubicacion	0
2	Anio	0
3	Kilometros_Recorridos	0
4	Tipo_Combustible	0
5	Transmision	0
6	Tipo_Propietario	0
7	Millaje	0
8	Motor	0
9	Antiguedad	0
10	Marca	0
11	Modelo	0
12	Potencia	46
13	Asientos	53
14	Precio	1234

Figura 28: Cantidad de nulos por dimensión después de la transformación.

	Nombre dimensión	Porcentaje de valores nulos (%)
0	Nombre	0.00
1	Ubicacion	0.00
2	Anio	0.00
3	Kilometros_Recorridos	0.00
4	Tipo_Combustible	0.00
5	Transmision	0.00
6	Tipo_Propietario	0.00
7	Millaje	0.00
8	Motor	0.00
9	Antiguedad	0.00
10	Marca	0.00
11	Modelo	0.00
12	Potencia	0.63
13	Asientos	0.73
14	Precio	17.01

Figura 29: Porcentaje de nulos por dimensión después de la transformación.

Notamos que contamos con apenas presencia de valores nulos, con una mayor frecuencia en la dimensión de precio con apenas 17.01 %. Mientras que antes de la transformación la dimensión con mayor índice de nulos tenia un porcentaje del 86.13 %.

3. Identificar por cada dimensión la cantidad valores duplicados.

	Dimension	Num_duplicados
3	Kilometros_Recorridos	3593
0	Nombre	5212
11	Precio	5879
14	Modelo	6520
9	Potencia	6866
8	Motor	7103
7	Millaje	7212
13	Marca	7221
2	Anio	7230
12	Antiguedad	7230
1	Ubicacion	7242
10	Asientos	7243
4	Tipo_Combustible	7248
6	Tipo_Propietario	7249
5	Transmision	7251

Figura 30: Cantidad de duplicados por cada dimensión después de la transformación.

6.2. Análisis gráfico.

1. Grafique la distribución de cada una de las dimensiones numéricas.

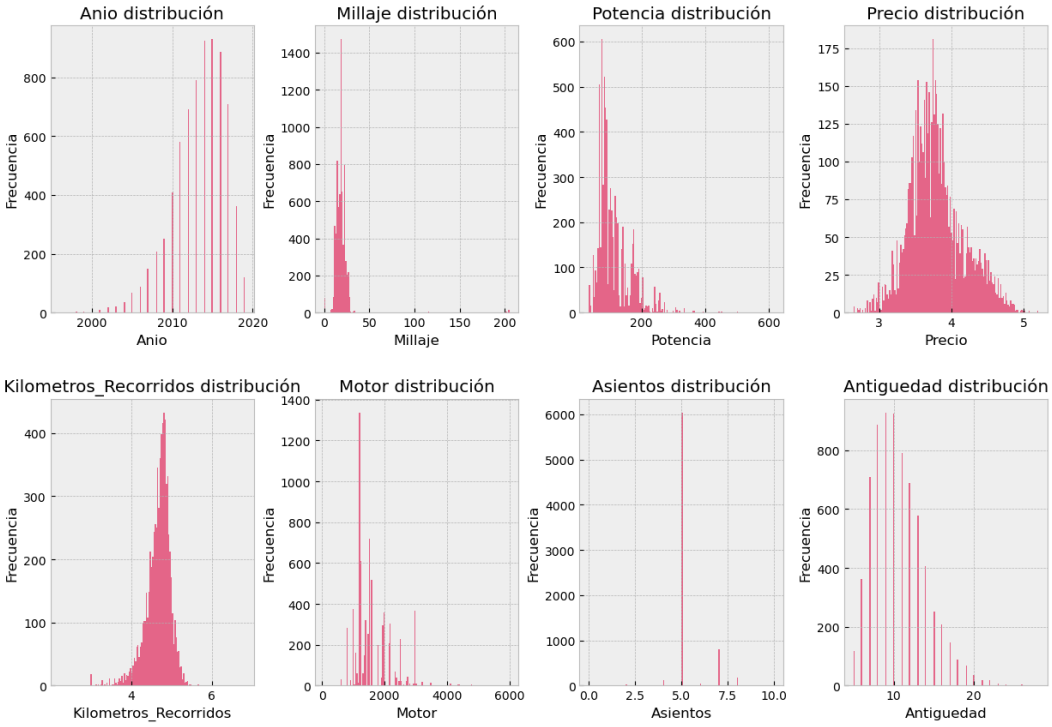

Figura 31: Distribución de dimensiones numéricas después de transformación.

Podemos ver que las dimensiones 'Kilometros_Recorridos' y 'Precio' (Figura 31) son las dimensiones que tuvieron un cambio más significativo en su distribución después de la transformación realizada en los pasos anteriores. Esto es debido a la aplicación del logaritmo, ya que redujo la escala de los datos.

2. Compare entre sí cada una de las dimensiones numéricas (análisis bivariado) con un gráfico del tipo "pairplot".

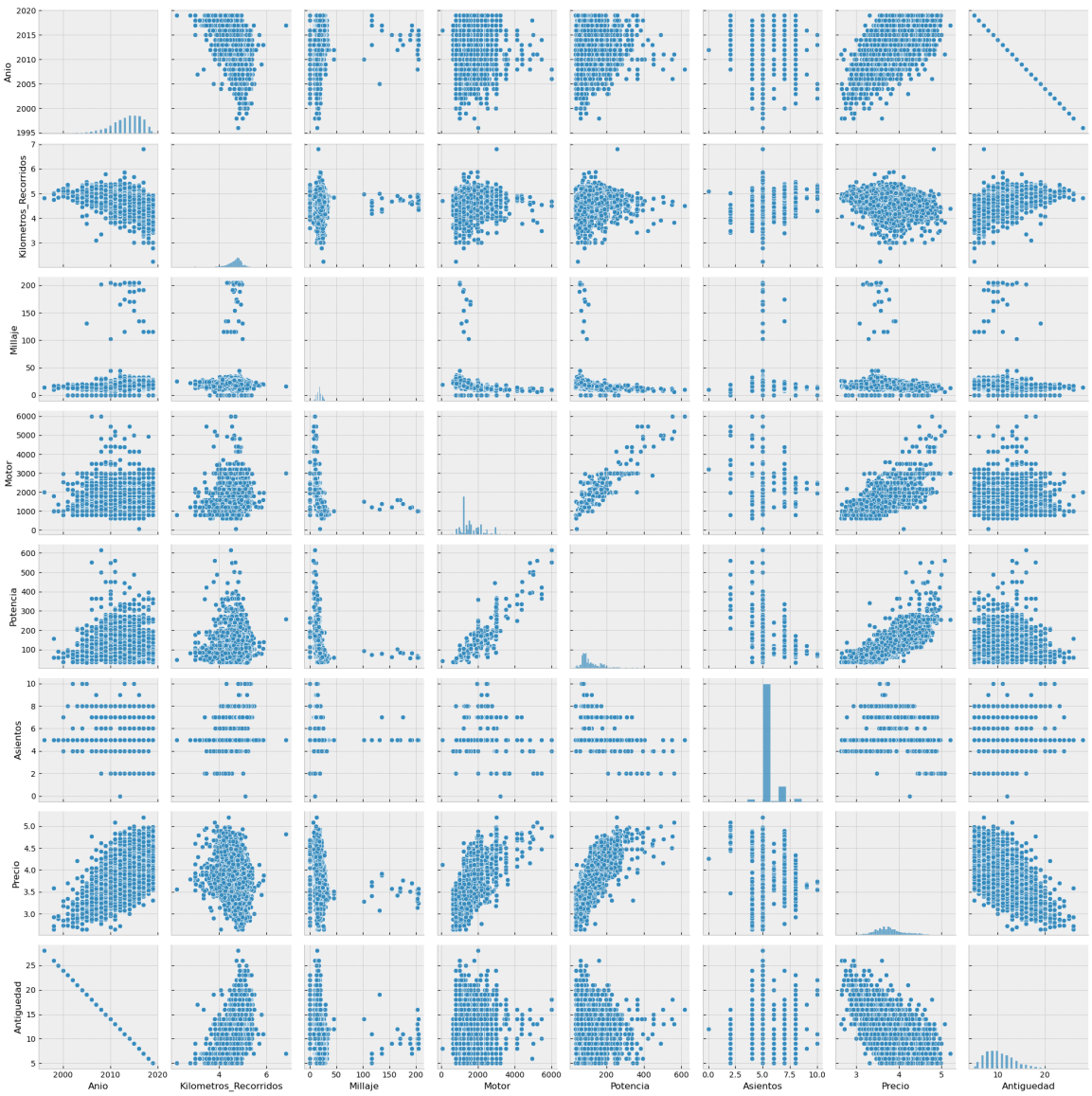

Figura 32: Graficos dispersión.

Despues de la transformación en el grafico de dispersión (Figura 32) de las dimensiones aparecieron nuevos graficos. Uno de los más interesantes es el es 'Potencia' y 'Motor', se puede apreciar que estas dos variables tienen una correlación positiva alta. Antes de la transformación esta relación no era visible.

3. Realice un mapa de calor para identificar la correlación entre todas las variables (análisis multivariado).

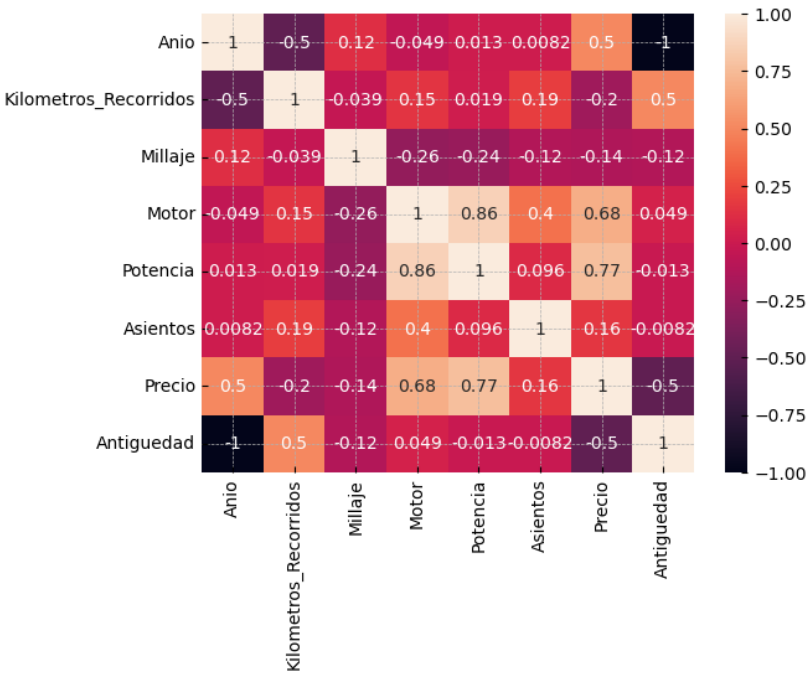

Figura 33: Mapa de calor para correlación después de transformación.

Con este gráfico de calor, podemos visualizar los coeficientes de correlación de Pearson entre cada par de dimensiones. Observamos que ahora contamos con mas correlaciones altas, como 'Potencia' con 'Motor', 'Potencia' con 'Precio' y 'Motor' con 'Precio'.

7. Conclusiones.

En conclusión, el proceso de procesamiento de datos, limpieza, imputación e ingeniería de características emerge como un componente crucial en el análisis de datos y la construcción de modelos predictivos efectivos. Estas etapas no solo garantizan la calidad y la integridad de los datos, sino que también permiten extraer información significativa y relevante que impulsa la toma de decisiones fundamentadas en diversos campos. Al invertir tiempo y recursos en estas fases fundamentales, se establece una base sólida para obtener resultados precisos y confiables, lo que subraya su importancia ineludible en cualquier proyecto de análisis de datos.

Además, se observó que al realizar estas mismas etapas con los datos ya preprocesados, pudimos rescatar nueva información, como correlaciones muy fuertes o cambios en las distribuciones de dimensiones existentes. También, gracias a la ingeniería de características, logramos rescatar información presente en el conjunto de datos pero que no se encontraba a simple vista, ampliando así el potencial de análisis y la capacidad predictiva de nuestros modelos.