Tecnológico de Monterrey

Actividad 2 - Preprocesamiento de datos

Análisis de Ciencia de Datos

TC2004B.101

Gpo. 101

Raúl Correa Ocañas

A01722401

Docentes:

Mtro. Rafael Martínez García Peña

Dr. Rasikh Tariq

Monterrey, Nuevo León, 21 de febrero 2024

Importación de datos

La importación de los datos comenzó con la carga de las bases de datos como DataFrames mediante la librería pandas en Python. Se empleó la función read_csv() para cargar los archivos CSV en los DataFrames cars1 y cars2.

Comparación de estructuras

Para determinar si las bases de datos tenían la misma estructura, se imprimieron las primeras filas de ambos DataFrames y se verificaron sus formas (número de filas y columnas). Se observó que las columnas data1 y data2 estaban presentes solo en cars1, mientras que no estaban en cars2, indicando que las estructuras no eran idénticas.

	mpg	cylinders	displacement	horsepower	weight	acceleration	model	origin	car	data1	data2
0	18.0	8	307	130	3504	12	70	1	chevrolet chevelle malibu	NaN	NaN
1	15.0	8	350	165	3693	11.5	70	1	buick skylark 320	NaN	NaN
2	18.0	8	318	150	3436	11	70	1	plymouth satellite	NaN	NaN
3	16.0	8	304	150	3433	12	70	1	amc rebel sst	NaN	NaN
4	17.0	8	302	140	3449	10.5	70	1	ford torino	NaN	NaN

	mpg	cylinders	displacement	horsepower	weight	acceleration	model	origin	car
0	33.0	4	91	53	1795	17.4	76	3	honda civic
1	20.0	6	225	100	3651	17.7	76	1	dodge aspen se
2	18.0	6	250	78	3574	21	76	1	ford granada ghia
3	18.5	6	250	110	3645	16.2	76	1	pontiac ventura sj
4	17.5	6	258	95	3193	17.8	76	1	amc pacer d/l

Figura 1. El DataFrame uno tiene shape (198, 11), mientras que el segundo tiene un shape de (200, 9). Se concluye que no tienen la misma estructura.

Eliminación de columnas con valores NaN

Posteriormente, se identificaron y eliminaron las columnas que contenían sólo valores NaN (Not a Number) utilizando el método drop() de pandas. Removiendo

estas, ahora tienen la misma estructura y se prosigue a hacer concatenación de los Dataframes.

```
# En la base de datos aparecen columnas que contienen sólo valores NaN (Not a Number). Remueve estas columnas.
cars1.drop(['data1','data2'], axis=1, inplace=True)
cars1
```

✓ 0.0s Python

	mpg	cylinders	displacement	horsepower	weight	acceleration	model	origin	car
0	18.0	8	307	130	3504	12	70	1	chevrolet chevelle malibu
1	15.0	8	350	165	3693	11.5	70	1	buick skylark 320
2	18.0	8	318	150	3436	11	70	1	plymouth satellite
3	16.0	8	304	150	3433	12	70	1	amc rebel sst
4	17.0	8	302	140	3449	10.5	70	1	ford torino
...	...	...	...	...	...	...	...	...	...
193	24.0	6	200	81	3012	17.6	76	1	ford maverick
194	22.5	6	232	90	3085	17.6	76	1	amc hornet
195	29.0	4	85	52	2035	22.2	76	1	chevrolet chevette
196	24.5	4	98	60	2164	22.1	76	1	chevrolet woody
197	29.0	4	90	70	1937	14.2	76	2	vw rabbit

198 rows × 9 columns

Figura 2. El DataFrame 'cars1' ahora tiene las mismas columnas que 'cars2', tienen los mismos nombres y sus tipos de datos son los mismos.

Reemplazo de valores faltantes

Se identificaron las columnas que contenían valores faltantes representados por '?'. El código utilizado para encontrarlo se muestra a continuación:

```
print("Columnas con valor ?")
for column in cars.columns:
    count = cars[column].isin(['?']).value_counts().get(True)
    print("{column}: {count}".format(column=column, count=count))
```

✓ 0.0s Python

```
Columnas con valor ?
mpg: None
cylinders: None
displacement: None
horsepower: 9
weight: None
acceleration: 7
model: None
origin: None
car: None
```

Figura 3. Código realizado para identificar cuántas '?' hay en cada columna. Se encuentran 9 valores en la columna 'horsepower' y 7 en la columna 'acceleration'.

Para las columnas ‘horsepower’ y 'acceleration’, se reemplazaron los valores '?' con la mediana y la media de las columnas, respectivamente, utilizando el método .loc[] de Pandas.

Figura 4. Para cada columna con valores ‘?’, convertir todos los valores que no sean ‘?’ a un tipo entero o flotante. Después, reemplazar los valores utilizando la mediana y media. Por último, volver a hacer casting de los valores para asegurar que todos los datos sean el tipo correcto. Se recibe de output dos listas vacías, indicando que ya no hay valores ‘?’.

Resumen de los datos

Finalmente, se proporcionó un resumen de los datos utilizando diferentes métodos de pandas. Se utilizó info() para obtener información general sobre los datos, describe() para obtener estadísticas descriptivas y select_dtypes() junto con describe() para obtener resúmenes de las variables categóricas.