Ian de Jesús Bonilla Díaz

Materia: Ciencia de datos

Nombre del profesor: Jaime Alejandro Sierra Romero

Fecha de entrega:20 de octubre de 2025

Git Hub: https://github.com/LansFacha/Proyecto

Limpieza de datos

Para preparar la base de datos antes del análisis se realiza un proceso de limpieza de la información.

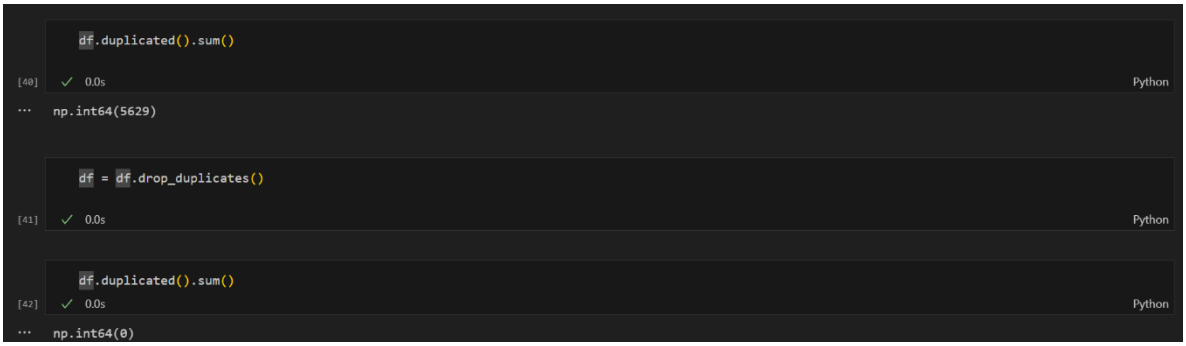
Se eliminaron valores nulos, se corrieron valores atípicos y se transformaron los tipos de datos según la conveniencia.

Este proceso nos permitió tener una base de datos completa, limpia y confiable, listo para su análisis.

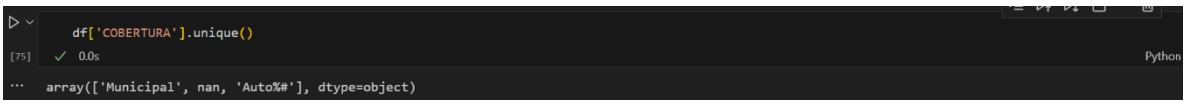
Proceso de limpieza.

Revisión de datos faltantes.

Detecciones y manejo de duplicados.



Detección de valores atípicos



Borrar datos Nan

En las mayorias de las columnas se uso el fillna para rellenar los valores Nan con valor desconcido para evitar generar datos falsos y en algunos se rellena con el

promedio o la moda ya que esas columnas no son tan necesarias para un analisi profundo y nos funcionan correctamente.

```
df["COBERTURA"]=df["COBERTURA"].fillna("Municipal")
```
0.0s Python

```
df['ESTADO'] = df['ESTADO'].fillna('Desconocido')
```
[83] 0.0s Python

```
df['MES'] = df['MES'].fillna('Desconocido')
```
0.0s Python

```
df['ESTADO'] = df['ESTADO'].fillna('Desconocido')
```
[83] 0.0s Python

```
df['MUNICIPIO'] = df['MUNICIPIO'].fillna('Desconocido')
```
0.0s Python

```
df['PRODUCTO'] = df['PRODUCTO'].fillna('Desconocido')
```
0.0s Python

```
df['GRUPO_PRODUCTO'] = df['GRUPO_PRODUCTO'].fillna('Desconocido')
```
0.0s Python

```
moda = df['UNIDAD_MEDIDA'].mode()[0]
```
0.0s Python

```
df['UNIDAD_MEDIDA'] = df['UNIDAD_MEDIDA'].fillna(moda)
```
0.0s Python

```
moda1= df['VOLUMEN'].mode()[0]
df['VOLUMEN'] = df['VOLUMEN'].fillna(moda1)
```
0.0s Python

```
moda2= df['ESTATUS'].mode()[0]
df['ESTATUS'] = df['ESTATUS'].fillna(moda2)
```
0.0s Python

En el tema de los años se organizo de 2001 a 2023 para que df saliera ordenada y no afectara en el analisis.

```
df = df.sort_values(by='ANIO').reset_index(drop=True)
```
0.0s Python

```
df['ANIO'] = pd.to_numeric(df['ANIO'], errors='coerce')
```
0.0s Python

Borrar o corregir palabras raras.

Se detecto que en la base de datos habia un valor atipico y raro que era Auto%# y para corregirlo se hizo lo siguiente.

```
Mes={'Auto%#':'Desconocido'}
```
0.0s Python

```
df['MES'] = df['MES'].replace(Mes)
```
0.0s Python

```
df['COBERTURA']=df['COBERTURA'].replace(Mes)
```
0.0s Python

```
df['GRUPO_PRODUCTO']=df['GRUPO_PRODUCTO'].replace(Mes)
```
0.0s Python

```
df['PRODUCTO']=df['PRODUCTO'].replace(Mes)
```
0.0s Python

```
df['UNIDAD_MEDIDA']=df['UNIDAD_MEDIDA'].replace(Mes)
```
0.0s Python

Y se confirma que la base esta limpia con

```
df.isnull().sum()
```

✓ 0.0s Python

```
COBERTURA          0
ANIO               0
MES                0
ESTADO             0
MUNICIPIO          0
GRUPO_PRODUCTO     0
PRODUCTO           0
UNIDAD_MEDIDA      0
VOLUMEN            0
ESTATUS            0
```

Y ordenada

	COBERTURA	ANIO	MES	ESTADO	MUNICIPIO	GRUPO_PRODUCTO	PRODUCTO	UNIDAD_MEDIDA	VOLUMEN	ESTATUS
0	Municipal	2001.0	Enero	Durango	Centro	Metales Industriales No Ferrosos	Zinc	Toneladas	1.0	Cifras Definitivas.
1	Municipal	2001.0	Desconocido	Durango	Concepción del Oro	Metales Preciosos	Oro	Kilogramos	0.1	Cifras Definitivas.
2	Municipal	2001.0	Enero	Durango	Concepción del Oro	Metales Preciosos	Plata	Kilogramos	0.0	Cifras Definitivas.
3	Municipal	2001.0	Enero	Durango	Concepción del Oro	Metales Industriales No Ferrosos	Plomo	Toneladas	5.0	Cifras Definitivas.
4	Auto%#	2001.0	Enero	Durango	Aquila	Metales Preciosos	Oro	Kilogramos	0.6	Cifras Definitivas.
...	...	...	...	...	...	...	...	...	...	...
89586	Municipal	2023.0	Diciembre	Durango	Mazapil	Metales Preciosos	Oro	Kilogramos	168.0	Cifras Preliminares.
89587	Municipal	2023.0	Abril	Sonora	Jiménez del Teul	Metales Industriales No Ferrosos	Zinc	Toneladas	1669.0	Cifras Preliminares.
89588	Municipal	2023.0	Febrero	Aguascalientes	Agua Prieta	Metales Industriales No Ferrosos	Cobre	Toneladas	48.0	Cifras Preliminares.
89589	Municipal	2023.0	Diciembre	Durango	Mazapil	Metales Preciosos	Desconocido	Kilogramos	13285.0	Cifras Preliminares.
89590	Municipal	2023.0	Diciembre	Durango	San Luis Potosí	Metales Preciosos	Desconocido	Kilogramos	829.4	Cifras Preliminares.

Conclusiones.

- Problemas principales:
  La base presentaba valores nulos en columnas clave como MES, ESTADO y PRODUCTO, duplicados exactos en varios registros y inconsistencias en nombres de estados, municipios y productos. Además, algunas columnas numéricas (VOLUMEN) tenían valores atípicos o incorrectos.
- Técnicas aplicadas:
  Se realizaron imputaciones de valores nulos (Desconocido para texto, promedio para numéricos), se eliminaron duplicados exactos, se corregieron errores ortográficos mediante diccionarios y se ajustaron tipos de datos para asegurar consistencia en la base.
- Aprendizaje del proceso:
  Aprendí a identificar y corregir inconsistencias en datos reales, a mantener la integridad de la información durante la limpieza y a preparar una base confiable para análisis y visualizaciones precisas.