Ejercicio de Laboratorio 6: ETL – P1

Instituto Politécnico Nacional.
Escuela Superior de Cómputo.
Licenciatura en ciencia de datos.
Bases de Datos Avanzadas

Emiliano López Méndez.

Introduccion

En esta práctica, trabajaremos con un archivo de datos en formato CSV que contiene información sobre ventas, y nuestro objetivo es aplicar técnicas de limpieza y manipulación de datos utilizando la biblioteca pandas en Python. En particular, nos enfocaremos en la eliminación de filas con valores nulos en la columna state y en el almacenamiento del DataFrame procesado en un archivo XML. Este ejercicio es fundamental para comprender los procesos de transformación de datos que se utilizan en proyectos de ciencia de datos y análisis de datos, donde la integridad y la organización de la información son claves para obtener resultados precisos y confiables.

A lo largo de esta práctica, no solo aplicaremos habilidades técnicas de manejo de archivos CSV y XML, sino que también comprenderemos la importancia de los procesos de ETL (Extract, Transform, Load), específicamente en la fase de transformación, en la que se asegura que los datos estén en el formato adecuado antes de ser utilizados en análisis posteriores o cargados en sistemas de almacenamiento de datos.
.

Desarrollo de la Actividad:

Carga de los Datos: En primer lugar, importamos la biblioteca pandas y utilizamos la función read_csv para cargar el archivo sales_data.csv en un DataFrame. Esta función nos permite leer el archivo CSV y tener la información en memoria en una estructura tabular, lo que facilita su manipulación.

Limpieza de Datos: Identificamos y eliminamos las filas en las que la columna state contiene valores nulos o vacíos. Para ello, utilizamos el método dropna, que elimina las filas con valores faltantes en una columna específica. Este paso es crucial, ya que los valores nulos pueden afectar el análisis de los datos y llevar a resultados incorrectos si no se manejan adecuadamente.

Almacenamiento en Formato XML: Una vez que los datos están limpios, guardamos el DataFrame resultante en un archivo XML utilizando el método to_xml. Este archivo se almacena en una ubicación diferente a la del archivo original para mantener una organización clara de los datos procesados y para que el archivo limpio esté disponible para futuras consultas o análisis.

Parte 1 Carga de archivo

En esta sección se procedemos a cargar el archivo csv como data frame para poder manejar con mucha facilidad y poder modifcarlo mejor

```
ruta = r'C:\Users\emlom\OneDrive\Documents\4semestre\4to-Semestre-Ciencia-de-Datos\Bases_de_Datos_Avanzadas\sales_data.csv'
df_sales = pd.read_csv(ruta, encoding='ISO-8859-1')
display(df_sales)
```

0.0s Python

	ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	ORDERDATE	STATUS	QTR_ID	MONTH_ID	YEAR_ID	...	PHONE	ADDRESSLINE1	ADDRESSLINE2	CITY	STATE	POSTALCODE	COUNTRY	TERRITORY	CONTACTLASTNAME
0	10107	30	95.70	2	2871.00	2/24/2003 0:00	Shipped	1	2	2003	...	2125557818	897 Long Airport Avenue	NaN	NYC	NY	10022	United States	NaN	Yu
1	10121	34	81.35	5	2765.90	5/7/2003 0:00	Shipped	2	5	2003	...	26.47.1555	59 rue de l'Abbaye	NaN	Reims	NaN	51100	France	EMEA	Henriot
2	10134	41	94.74	2	3884.34	7/1/2003 0:00	Shipped	3	7	2003	...	+33 1 46 62 7555	27 rue du Colonel Pierre Avia	NaN	Paris	NaN	75508	France	EMEA	Da Cunha
3	10145	45	83.26	6	3746.70	8/25/2003 0:00	Shipped	3	8	2003	...	6265557265	78934 Hillside Dr.	NaN	Pasadena	CA	90003	USA	NaN	Young
4	10159	49	100.00	14	5205.27	10/10/2003 0:00	Shipped	4	10	2003	...	6505551386	7734 Strong St.	NaN	San Francisco	CA	NaN	United States	NaN	Brown
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2818	10350	20	100.00	15	2244.40	12/2/2004 0:00	Shipped	4	12	2004	...	(91) 555 94 44	C/ Moralzarzal, 86	NaN	Madrid	NaN	28034	Spain	EMEA	Freyre
2819	10373	29	100.00	1	3978.51	1/31/2005 0:00	Shipped	1	1	2005	...	981-443655	Torikatu 38	NaN	Oulu	NaN	90110	Finland	EMEA	Koskitalo
2820	10386	43	100.00	4	5417.57	3/1/2005 0:00	Resolved	1	3	2005	...	(91) 555 94 44	C/ Moralzarzal, 86	NaN	Madrid	NaN	28034	Spain	EMEA	Freyre
2821	10397	34	62.24	1	2116.16	3/28/2005 0:00	Shipped	1	3	2005	...	61.77.6555	1 rue Alsace-Lorraine	NaN	Toulouse	NaN	31000	France	EMEA	Roule
2822	10414	47	65.52	9	3079.44	5/6/2005 0:00	On Hold	2	5	2005	...	6175559555	8616 Spinnaker Dr.	NaN	Boston	MA	51003	USA	NaN	Yoshido

2823 rows × 24 columns

Parte 2 Consultar espacios nulos en State y quitarlos

En esta etapa, se realizan consultas sobre los datos nulos de la columna STATE y los vamos a quitar

Vemos que columans hay

```
df_sales.iloc[0]
```

[22] ✓ 0.0s

```
ORDERNUMBER                10107
QUANTITYORDERED               30
PRICEEACH                   95.7
ORDERLINENUMBER                2
SALES                     2871.0
ORDERDATE          2/24/2003 0:00
STATUS                   Shipped
QTR_ID                         1
MONTH_ID                       2
YEAR_ID                     2003
PRODUCTLINE           Motorcycles
MSRP                           95
PRODUCTCODE              S10_1678
CUSTOMERNAME     Land of Toys Inc.
PHONE                  2125557818
ADDRESSLINE1    897 Long Airport Avenue
ADDRESSLINE2                  NaN
CITY                          NYC
STATE                          NY
POSTALCODE                 10022
COUNTRY              United States
TERRITORY                     NaN
CONTACTLASTNAME                Yu
CONTACTFIRSTNAME             Kwai
Name: 0, dtype: object
```

```
df_sales = df_sales.dropna(subset=['STATE'])
display (df_sales)
```

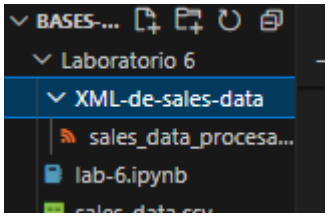
[23] ✓ 0.0s Python

	ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	ORDERDATE	STATUS	QTR_ID	MONTH_ID	YEAR_ID	...	PHONE	ADDRESSLINE1	ADDRESSLINE2	CITY	STATE	POSTALCODE	COUNTRY	TERRITORY	CONTACTLASTN
0	10107	30	95.70	2	2871.00	2/24/2003 0:00	Shipped	1	2	2003	...	2125557818	897 Long Airport Avenue	NaN	NYC	NY	10022	United States	NaN	
3	10145	45	83.26	6	3746.70	8/25/2003 0:00	Shipped	3	8	2003	...	6265557265	78934 Hillside Dr.	NaN	Pasadena	CA	90003	USA	NaN	Y
4	10159	49	100.00	14	5205.27	10/10/2003 0:00	Shipped	4	10	2003	...	6505551386	7734 Strong St.	NaN	San Francisco	CA	NaN	United States	NaN	B
5	10168	36	96.66	1	3479.76	10/28/2003 0:00	Shipped	4	10	2003	...	6505556809	9408 Furth Circle	NaN	Burlingame	CA	94217	USA	NaN	H
8	10201	22	98.57	2	2168.54	12/1/2003 0:00	Shipped	4	12	2003	...	6505555787	5557 North Pendale Street	NaN	San Francisco	CA	NaN	United States	NaN	Mu
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2809	10248	23	65.52	9	1506.96	5/7/2004 0:00	Cancelled	2	5	2004	...	2125557818	897 Long Airport Avenue	NaN	NYC	NY	10022	USA	NaN	
2810	10261	29	50.78	7	1472.62	6/17/2004 0:00	Shipped	2	6	2004	...	(514) 555-8054	43 rue St. Laurent	NaN	Montr□al	Qu□bec	H1J 1C3	Canada	NaN	Fres
2812	10283	33	51.32	12	1693.56	8/20/2004 0:00	Shipped	3	8	2004	...	(604) 555-4555	23 Tsawassen Blvd.	NaN	Tsawassen	BC	T2F 8M4	Canada	NaN	Li
2817	10337	42	97.16	5	4080.72	11/21/2004 0:00	Shipped	4	11	2004	...	2125558493	5905 Pompton St.	Suite 750	NYC	NY	10022	USA	NaN	Herna
2822	10414	47	65.52	9	3079.44	5/6/2005 0:00	On Hold	2	5	2005	...	6175559555	8616 Spinnaker Dr.	NaN	Boston	MA	51003	USA	NaN	Yo

1337 rows × 24 columns

Parte 3 Guardamos el archivo como xml

En esta parte del laboratorio cambiamos el archivo csv a xml y guardarlo en otra carpeta nueva



Conclusión

En esta práctica, logramos procesar y limpiar un conjunto de datos utilizando herramientas fundamentales de pandas en Python, enfocándonos en la eliminación de valores nulos y el almacenamiento en distintos formatos. A través de la eliminación de las filas con valores nulos en la columna state, comprendimos la importancia de asegurar la calidad de los datos, ya que los valores faltantes pueden introducir errores en análisis y decisiones posteriores.

Además, el proceso de conversión del DataFrame a formato XML destacó la versatilidad de pandas y nos brindó una experiencia práctica en la manipulación y transformación de datos para prepararlos para almacenamiento o integración en otros sistemas. Resolver los problemas técnicos, como la codificación y la instalación de bibliotecas adicionales, nos permitió afianzar habilidades para manejar situaciones comunes al trabajar con archivos de distintos orígenes y formatos.