* Cleaning Data using Pandas:-

→ import pandas as pd
import numpy as np

people = {
"First": ["Essam", "Jannet", "John", np.nan, None, "NA"],
"Last": ["Shenhab", "Doe", "Doe", np.nan, np.nan, "Missing"],
"age": [20, 21, 22, None, None, "Missing"]
}

df = pd.DataFrame(people)

df →

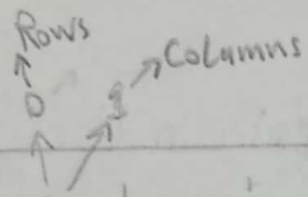
	First	Last	age
0	Essam	Shenhab	20
1	Jannet	Doe	21
2	John	Doe	22
3	NaN	NaN	None
4	None	NaN	None
5	NA	Missing	Missing

Removing Rows/columns with missing values:-

→ df.dropna() → يحذف الـ Rows التي بها NaN تلقائيًا

→

	First	Last	age
0	Essam	Shenhab	20
1	Jannet	Doe	21
2	John	Doe	22
5	NA	Missing	Missing



→ df.dropna(axis=0, how='...', inplace=True)
- any: drop Rows if any of the values is missing
- all " " " all " " - " ade "

※ Dropping Rows with missing values in a certain Column

→ df.dropna(axis=0, how='any', subset=['last']

→

	First	last	age
0	Essam	Shenhab	20
1	Jannet	Doe	21
2	John	Doe	22
5	NA	Missing	Missing

→ df.dropna(axis=0, how='all', subset['last', 'age']

※ Customizing standard missing values to Remove missing values Like(NA, missing)

→ df.replace('NA', np.nan, inplace=True)
df.replace('Missing', np.nan, inplace=True)
← كل قيمة بـ NA أو Missing ستصبح NAN ليسهل حذفها

※ identifying missing values using isna()

→ df.isna() → True و أي شيء آخر بـ False يستبدل NAN

✳ Filling missing values

→ df.Fillna('MISSING') → MISSING بـ NaN يستبدل

و لكن مفيد أكثر لإستبدال الـ numerical data بمثلاً الـ mean أو حتى

→ df['age'].mean() → 21 عشان Error

✳ Converting missing values to appropriate datatypes

→ df['age']=df['age'].astype(float)

→ df['Years Code'].unique() → gives us all the unique of that column

✳ Removing duplicates

→ df.drop_duplicates(inplace=True)

✳ STrip

left ← الحروف التي سيتم حذفها

→ df['Lastname']=df["Lastname"].str.Lstrip('...')

✳ STandarizing Phone numbers

→ df['PhoneNumber']=df['PhoneNumber'].str.replace('[^a-zA-Z0-9]',' ')

→ df['PhoneNumber'].apply(Lambda x: x[0:3]+'-'+x[3:6]+'-'+x[6:10])

→ df['PhoneNumber'].apply(Lambda x: str(x))

✳Splitting Columns

→ df[['Street_Address', 'State', 'zip_code']] = df['Address'].str.split(',', 2, expand=True)

✳Filtering Rows of Data

→ for x in df.index:
    if df.loc[x, 'PhoneNumber'] == '':
        df.drop(x, inplace = True)

df → هيحذف كل الـRows اللي بهاتفات فاضيه

→ df=df.dropna(subset = "PhoneNumber"), inplace=True)

✳Reseting index

→ df=df.reset_index(drop_index)

✳Check For Null values

→ df.isnull().any()

→ df[df['Phone Number'].isna()==1]

✳Filling missing value equal to the value after it

→ df.ffill
    ↳ infront of it

→ df.bfill
    ↳ behind of it