Objective: Clean and prepare a raw dataset (with nulls, duplicates, inconsistent formats). Tools: Excel / Python (Pandas)

Deliverables: Cleaned dataset + short summary of changes Hints / Mini Guide: Identify and handle missing values using .isnull() in Python or filters in Excel. Remove duplicate rows using .drop\_duplicates() or Excel’s “Remove Duplicates”. Standardize text values like gender, country names, etc. Convert date formats to a consistent type (e.g., dd-mm-yyyy). Rename column headers to be clean and uniform (e.g., lowercase, no spaces). Check and fix data types (e.g., age should be int, date as datetime). Dataset names from Kaggle suitable for Task 1: Customer Personality Analysis Medical Appointment No Shows Mall Customer Segmentation Data Netflix Movies and TV Shows Sales Data

***Answer by step by step***

*This is a solid plan for dataset cleaning! Since you're already practicing data cleaning with Excel and Python (Pandas), here's a structured breakdown of how you can approach this task efficiently:*

*Step-by-Step Guide for Cleaning & Preparing the Dataset*

*1. Handling Missing Values*

* *Excel: Use filters or conditional formatting to find missing values. Fill them using formulas like IF(ISBLANK(A2), "Unknown", A2), or remove them based on relevance.*
* *Python (Pandas): Use .isnull().sum() to check missing values. Fill using .fillna(value) or drop with .dropna().*

*2. Removing Duplicates*

* *Excel: Use the "Remove Duplicates" option in the "Data" tab.*
* *Pandas: Use df.drop\_duplicates(inplace=True) to eliminate duplicate rows.*

*3. Standardizing Text Values*

* *Excel: Use functions like UPPER(), LOWER(), or TRIM() to clean inconsistent text values.*
* *Pandas: Apply df['column'] = df['column'].str.lower().str.strip() to normalize text formats.*

*4. Formatting Dates Consistently*

* *Excel: Use TEXT(A2, "DD-MM-YYYY") for formatting dates.*
* *Pandas: Convert dates using pd.to\_datetime(df['date\_column'], format='%d-%m-%Y').*

*5. Renaming Columns*

* *Excel: Modify headers manually or use "Find & Replace."*
* *Pandas: Rename columns using df.rename(columns={'OldName': 'new\_name'}, inplace=True).*

*6. Fixing Data Types*

* *Excel: Use the "Format Cells" option to adjust types (number, date, text).*
* *Pandas: Convert types using df['age'] = df['age'].astype(int) or df['date'] = pd.to\_datetime(df['date']).*

*Deliverables*

* *Cleaned Dataset: Ensure the cleaned version is saved as .csv or .xlsx for easy access.*
* *Summary of Changes: Document the changes made, including missing values handled, duplicates removed, standardization applied, and any transformations performed.*