Dataset Used: Netflix Movies and TV Shows Dataset

✅ Objective:

To clean and preprocess a raw dataset using Python (Pandas) by identifying and fixing common data issues such as:

Missing values

Duplicates

Inconsistent formats

Data type mismatches

Column renaming

🔧 Tools Used:

Python

Pandas

Jupyter Notebook

🧼 Cleaning Steps Performed

1. Loaded the dataset from a .csv file using Pandas

2. Checked the shape (number of rows and columns)

3. Inspected column names and overall dataset structure

4. Checked for missing values using isnull().sum()

5. Removed duplicate rows using drop\_duplicates()

6. Standardized column headers (e.g., lowercase, replacing spaces)

7. Converted date format (date\_added) into proper datetime format

8. Extracted added\_year and added\_month from the date\_added column

9. Verified data types and corrected them where needed

10. Exported the cleaned dataset as netflix\_titles\_cleaned.csv

🧠 Interview Questions & Answers

1. What are missing values and how do you handle them?

Missing values are empty or null entries in a dataset. They can be handled by removing rows (dropna()), filling them with suitable values (fillna()), or using domain knowledge to infer them.

2. How do you treat duplicate records?

Duplicate records can be dropped using drop\_duplicates() in Pandas to ensure data integrity.

3. Difference between dropna() and fillna() in Pandas?

dropna() removes rows or columns with missing values.

fillna() replaces missing values with a specific value (e.g., mean, median, or a fixed string).

4. What is outlier treatment and why is it important?

Outlier treatment involves detecting and handling extreme values that may distort analysis. It's crucial to improve model performance and accuracy.

5. Explain the process of standardizing data.

Standardizing involves formatting values uniformly (e.g., all lowercase, consistent date formats) to avoid redundancy and errors during analysis.

6. How do you handle inconsistent data formats (e.g., date/time)?

By converting data types using functions like pd.to\_datetime() and formatting dates uniformly (e.g., dd-mm-yyyy).

7. What are common data cleaning challenges?

Missing/incomplete data

Duplicates

Wrong data types

Inconsistent formatting

Outliers

8. How can you check data quality?

By checking for null values, duplicates, invalid formats, and data consistency using functions like .info(), .isnull().sum(), .duplicated(), and .describe().

📌 Outcome

This task helped in building strong foundational skills in real-world data cleaning using Python and Pandas, which is a crucial step before data analysis or visualization.