**1. Dataset Overview**

The dataset used for this project is the **Netflix Movies and TV Shows dataset** from Kaggle, containing over 8,000 titles.

Columns include:

* **show\_id**
* **type**
* **title**
* **director**
* **cast**
* **country**
* **date\_added**
* **release\_year**
* **rating**
* **duration**
* **listed\_in** (genres)
* **description**

**Dataset Inspection**

* Used df.info() to check column data types and null values.
* Used df.head(3) to view sample records for understanding structure.

**2. Data Cleaning Steps (Pandas)**

To prepare the dataset for analysis, the following cleaning operations were implemented:

1. **Removing Duplicates**
2. df.drop\_duplicates(inplace=True)

This ensured that no repeated entries existed in the dataset.

1. **Handling Missing Values**
   * Filled missing director, cast, and country with "Unknown".
   * Filled missing rating with "Not Rated".
2. df['director'].fillna("Unknown", inplace=True)
3. df['cast'].fillna("Unknown", inplace=True)
4. df['country'].fillna("Unknown", inplace=True)
5. df['rating'].fillna("Not Rated", inplace=True)
6. **Checking Remaining Nulls**  
   Verified null values after filling:
7. df.isnull().sum()
8. **Saving Cleaned Dataset**  
   Saved the processed dataset for further analysis:

df.to\_csv("/Volumes/workspace/default/netflix\_dataset/netflix\_cleaned.csv", index=False)

**3. Conclusion:**

The dataset has been successfully cleaned using Pandas, ensuring that duplicates and missing values were handled appropriately.  
The cleaned dataset now provides a strong foundation for further **exploratory data analysis (EDA)**, **clustering**, and **dashboard development** to uncover trends in Netflix’s content strategy.