***DataSet Review***

* The file name is netflix\_titles.csv
* There are 8000+ rows and 12 columns
* Columns names are

• show\_id

• type

• title

• director

• cast

• country

• date\_added

• release\_year

• rating

• duration

• listed\_in (genres)

• description

* There are many nulls values in this dataset

***Dataset Cleaning Process using PySpark***

* In the first we cast the data type of each columns into correct one

Eg. Date\_added is in String type So converting it into date Data Type

* Using Fillna we fill the null value containing column to something so that we can’t simply drop all rest important data from that same row
* Now we use Drop\_Duplicates to remove all the duplicates from the dataset
* So now removing all the rows which has nulls

***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