

**FACULTY OF COMPUTER SCIENCE AND ENGINEERING  
Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Topi  
  
DATA SCIENCE REPORT**

**Instructor: MISS SAFIA**

**Assignment 3**

|  |
| --- |
| **Muhammad Kumail Hasan rizvi 2021307** |
|  |
|  |

**Netflix Data Analysis Report**

**Problem Statement**

The goal of this report is to perform **Exploratory Data Analysis (EDA)** on the Netflix dataset using **PySpark**. The analysis includes cleaning the data, handling missing values, analyzing the content distribution, and visualizing the insights gained.

Netflix, as a leading streaming platform, provides a massive amount of content in the form of Movies and TV Shows. Our objectives are:

1. Identify the distribution of Movies and TV Shows.
2. Analyze popular genres, top-producing countries, and content trends over time.
3. Clean the dataset and resolve missing values.

**Tools and Libraries Used**

1. **Google Colab**: For running the analysis on a cloud-based environment.
2. **Apache Spark**: Used for distributed data processing and handling large datasets.
3. **PySpark**: Python library for Spark to perform SQL-like operations and analysis.
4. **Matplotlib**: Library for creating visualizations.

**Spark Configuration**:

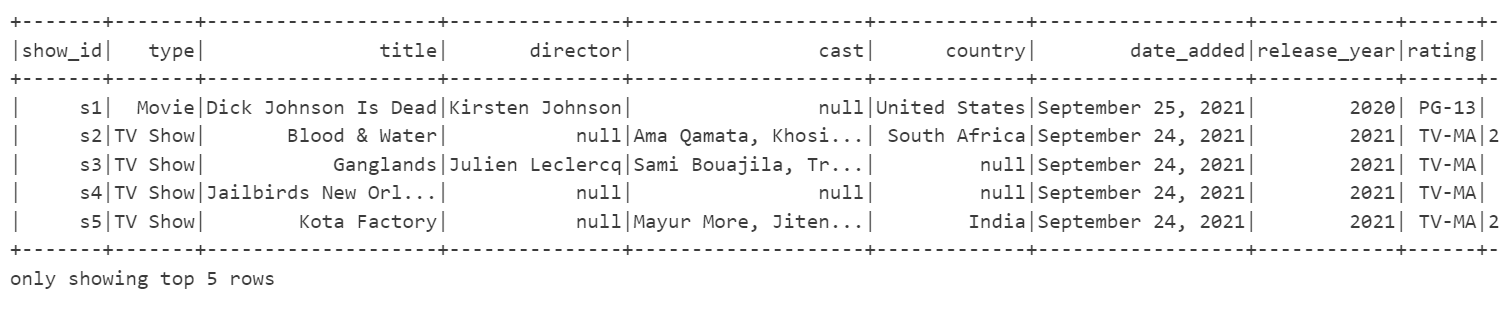
* Spark version: 3.3.0
* Java version: 8
* Libraries: findspark, pyspark.sql, matplotlib.

**Dataset Description**

The dataset used is the **Netflix Titles Dataset**, containing information about Movies and TV Shows available on Netflix.

* **File Name**: netflix\_titles.csv
* **Columns**:
  + show\_id, type, title, director, cast, country, date\_added, release\_year, rating, duration, listed\_in, description.

The dataset contains missing values in multiple columns, which were handled during the cleaning phase.



**What We Have Done in This Report**

**1. Data Setup and Initialization**

* Installed **Java 8**, **Spark 3.3.0**, and **findspark**.
* Configured Spark environment in Google Colab.
* Initialized a Spark session to load and process the dataset.

**2. Data Loading and Cleaning**

* Loaded the dataset into a **Spark DataFrame**
* Handled **missing values**:
  + Numerical columns: Replaced with 0.
  + String columns: Replaced with "Unknown".

A close-up of a white paper

Description automatically generated

**3. Exploratory Data Analysis (EDA)**

**a) Movies vs TV Shows Count**

**b) Popular Genres**

* Exploded the listed\_in column to analyze genres:

**Result**:

* Drama is the most common genre.

A screenshot of a computer

Description automatically generated

**c) Content Distribution Over Years**

* Grouped the data by release\_year:

**Result**:

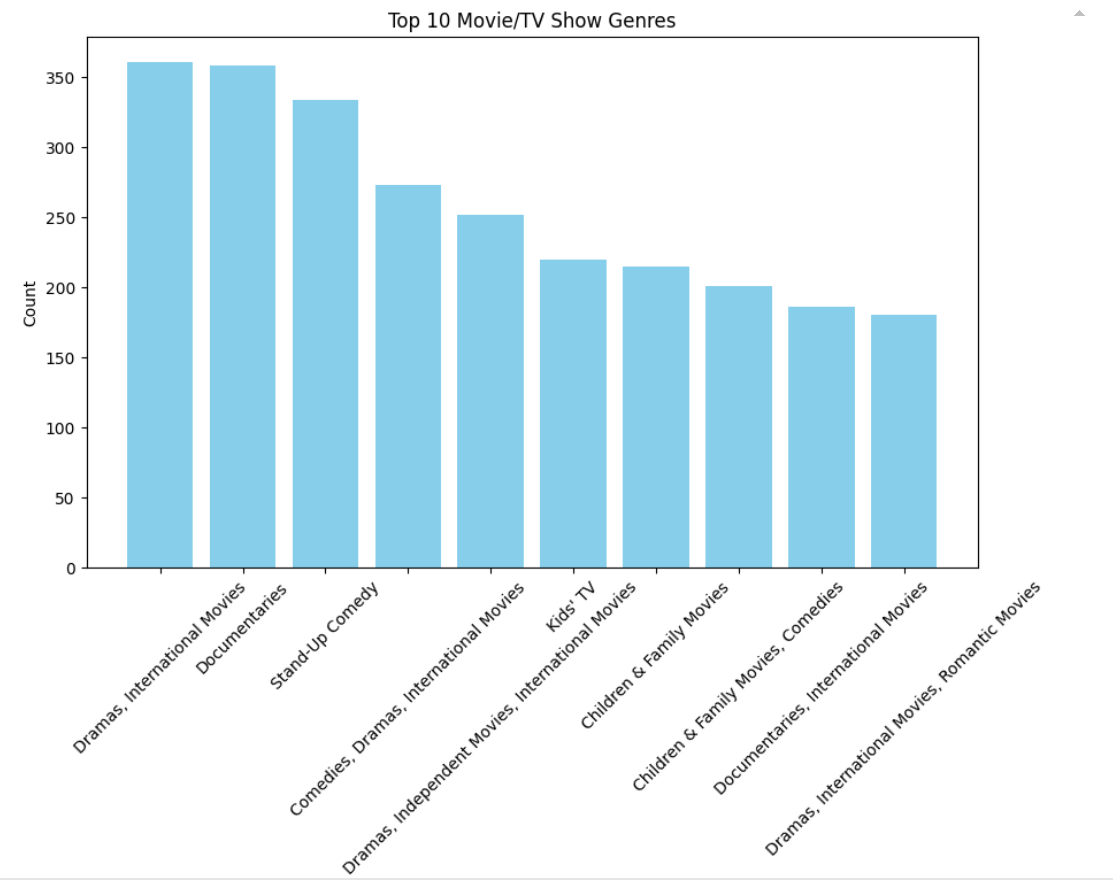
* Content production significantly increased after the 2000s.
* A screenshot of a computer

  Description automatically generated

**4. Visualizations**

We created simple visualizations to present the analysis clearly:

1. **Movies vs TV Shows Count**  
   Bar graph showing the distribution of Movies and TV Shows.



1. **Top 5 Countries with Most Content**  
   Bar graph highlighting the countries producing the most content.

A graph of blue bars with black text

Description automatically generated

1. **Top 5 Years with Most Releases**

A graph of orange bars

Description automatically generated

**Link:**

**https://github.com/Kumalhasan2001/netflix-spark-eda.git**