```
!pip install pyspark
```

```
Requirement already satisfied: pyspark in /usr/local/lib/python3.11/dist-packages (3.5.1)
 Requirement already satisfied: py4j==0.10.9.7 in /usr/local/lib/python3.11/dist-packages (from pyspark) (0.10.9.7)
```

```
from pyspark.sql import SparkSession from pyspark.sql.functions
import col, count, desc, explode, split
```

```
spark = SparkSession.builder.appName("Netflix Big Data Analysis").getOrCreate()
```

```
from google.colab import files
uploaded = files.upload()
```

Choose Files netflix_titles.csv

- **netflix_titles.csv**(text/csv) - 3399671 bytes, last modified: 7/11/2025 - 100% done

Saving netflix_titles.csv to netflix_titles.csv

```
df = spark.read.csv("netflix_titles.csv", header=True,

inferSchema=True) df.show(5) df.printSchema()
```

```
+-------+-------+--------------------+--------------+--------------------+-------------+------------------+------------+------+--------
|show_id|   type|               title|      director|                cast|      country|        date_added|release_year|rating| duratio
+-------+-------+--------------------+--------------+--------------------+-------------+------------------+------------+------+--------
|     s1|  Movie|Dick Johnson Is Dead|Kirsten Johnson|                NULL|United States|September 25, 2021|        2020| PG-13|   90 mi
|     s2|TV Show|       Blood & Water|          NULL|Ama Qamata, Khosi...| South Africa|September 24, 2021|        2021| TV-MA|2 Season
|     s3|TV Show|           Ganglands|Julien Leclercq|Sami Bouajila, Tr...|         NULL|September 24, 2021|        2021| TV-MA| 1 Seaso
|     s4|TV Show|Jailbirds New Orl...|          NULL|                NULL|         NULL|September 24, 2021|        2021| TV-MA| 1 Seaso
|     s5|TV Show|        Kota Factory|          NULL|Mayur More, Jiten...|        India|September 24, 2021|        2021| TV-MA|2 Season
+-------+-------+--------------------+--------------+--------------------+-------------+------------------+------------+------+-------
only showing top 5 rows

root
 |-- show_id: string (nullable = true)
 |-- type: string (nullable = true)
 |-- title: string (nullable = true)
 |-- director: string (nullable = true)
 |-- cast: string (nullable = true)
 |-- country: string (nullable = true)
 |-- date_added: string (nullable = true)
 |-- release_year: string (nullable = true)
 |-- rating: string (nullable = true)
 |-- duration: string (nullable = true)
 |-- listed_in: string (nullable = true)
 |-- description: string (nullable = true)
```

```
df.groupBy("type").count().show()

df.groupBy("country").count().orderBy(desc("count")).show(5)

df.groupBy("release_year").count().orderBy(desc("release_year")).show(10)

genres_df = df.withColumn("genre", explode(split(col("listed_in"), ", ")))
genres_df.groupBy("genre").count().orderBy(desc("count")).show(5)
```

```
+-------------+-----+
|         type|count|
+-------------+-----+
|         NULL|    1|
|      TV Show| 2676|
|        Movie| 6131|
|William Wyler|    1|
+-------------+-----+
+--------------+-----+
|       country|count|
+--------------+-----+
| United States| 2805|
|         India|  972|
|          NULL|  832|
|United Kingdom|  419|
|         Japan|  245| +--
------------+-----+ only
showing top 5 rows

+-----------------+-----+
|     release_year|count|
+-----------------+-----+
|    United States|    1|
|    June 12, 2021|    1|
| January 15, 2021|    1|
| January 13, 2021|    1|
|December 15, 2020|    1|
|  August 13, 2020|    1|
|           40 min|    1|
|             2021|  589|
|             2020|  952|
|             2019| 1026| +--
---------------+-----+ only
showing top 10 rows

+--------------------+-----+
|               genre|count|
+--------------------+-----+
|International Movies| 2748|
|              Dramas| 2419|
|            Comedies| 1670|
|International TV ...| 1350|
|       Documentaries|  866| +--
------------------+-----+ only
showing top 5 rows
```

```
from pyspark.sql.functions import regexp_extract df = df.withColumn("release_year_clean",

regexp_extract(col("release_year"), r"\b(19|20)\d{2}\b", 0)) df_clean =

df.filter(col("release_year_clean") != "")

df_clean.groupBy("release_year_clean").count().orderBy(desc("release_year_clean")).show(10)
```

```
+------------------+-----+
|release_year_clean|count|
+------------------+-----+
|              2021|  592|
|              2020|  954|
|              2019| 1026|
|              2018| 1145|
|              2017| 1030|
|              2016|  901|
|              2015|  559|
|              2014|  352|
|              2013|  288|
|              2012|  237| +--
----------------+-----+ only
showing top 10 rows
```