# databricks Big-Data-Assignment-Netflix-Ads-Platform

(https://databricks.com)

```
# read csv file from dbfs'
file_path = "/FileStore/Netflix_TV_Shows_and_Movies.csv"  # Adjust the path accordingly

# Read CSV into a DataFrame
df = spark.read.option("header", "true").csv(file_path)
```

```
df.show()
```

```
+-----+--------+--------------------+-----+--------------------+--------------------+--------------------+--------------------+--
------------------+----------+----------+
|index|      id|               title| type|         description|        release_year|   age_certification|             runtime|
imdb_id|imdb_score|imdb_votes|
+-----+--------+--------------------+-----+--------------------+--------------------+--------------------+--------------------+--
------------------+----------+----------+
|    0| tm84618|         Taxi Driver|MOVIE|A mentally unstab...|                1976|                   R|                 113|
tt0075314|       8.3|    795222|
|    1|tm127384|Monty Python and ...|MOVIE|"King Arthur, acc...|                1975|                  PG|                  91|
tt0071853|       8.2|    530877|
|    2| tm70993|       Life of Brian|MOVIE|Brian Cohen is an...|                1979|                   R|                  94|
tt0079470|         8|    392419|
|    3|tm190788|        The Exorcist|MOVIE|12-year-old Regan...|                1973|                   R|                 133|
tt0070047|       8.1|    391942|
|    4| ts22164|Monty Python's Fl...| SHOW|A British sketch ...|                1969|               TV-14|                  30|
tt0063929|       8.8|     72895|
|    5| tm14873|         Dirty Harry|MOVIE|When a madman dub...|                1971|                   R|                 102|
tt0066999|       7.7|    153463|
|    6|tm185072|        My Fair Lady|MOVIE|A snobbish phonet...|                1964|                   G|                 170|
tt0058385|       7.8|     94121|
|    7| tm98978|     The Blue Lagoon|MOVIE|Two small childre...|                1980|                   R|                 104|
```

```
df.printSchema()
```

```
root
 |-- index: string (nullable = true)
 |-- id: string (nullable = true)
 |-- title: string (nullable = true)
 |-- type: string (nullable = true)
 |-- description: string (nullable = true)
 |-- release_year: string (nullable = true)
 |-- age_certification: string (nullable = true)
 |-- runtime: string (nullable = true)
 |-- imdb_id: string (nullable = true)
 |-- imdb_score: string (nullable = true)
 |-- imdb_votes: string (nullable = true)
```

```
df.count()
```

```
Out[24]: 5364
```

```
# Display top-rated titles based on IMDb score
df.orderBy("imdb_score", ascending=False).select("title", "imdb_score").show(10)
# This query helps identify the top-rated titles based on IMDb scores, which can be useful for promoting high-quality content.
```

```
+--------------------+----------+
|               title|imdb_score|
+--------------------+----------+
|              Losers| tt9817218|
|           revisions| tt9522354|
|   Wedding Unplanned| tt9519642|
|            Puerta 7| tt9170386|
|             Gormiti| tt9077350|
|ReMastered: The L...| tt9046576|
|        Chaman Bahar| tt8747450|
|Dance Dreams: Hot...| tt8741182|
|Derren Brown: Mir...| tt8599562|
|Dragon Pilot: His...| tt8528256|
+--------------------+----------+
only showing top 10 rows
```

```
# Analyze the distribution of age certifications
df.groupBy("age_certification").count().orderBy("count", ascending=False).show()
# Understanding the distribution of age certifications can help tailor content to specific audience segments.
```

```
+-----------------+-----+
|age_certification|count|
+-----------------+-----+
|             null| 2275|
|            TV-MA|  760|
|                R|  521|
|            PG-13|  412|
|            TV-14|  398|
|               PG|  228|
|            TV-PG|  158|
|                G|  102|
|            TV-Y7|   98|
|             TV-Y|   90|
|             TV-G|   64|
|            NC-17|   12|
|             2017|   11|
|             2018|   11|
|             2019|   10|
|             2016|    8|
|             2020|    7|
|             2015|    6|
```

```
# Analyze user engagement based on IMDb votes
df.groupBy("title").agg({"imdb_votes": "sum"}).orderBy("sum(imdb_votes)", ascending=False).show(10)
# Identifying titles with high user engagement, as measured by IMDb votes, can guide content promotion strategies.
```

```
+--------------------+---------------+
|               title|sum(imdb_votes)|
+--------------------+---------------+
|        Forrest Gump|      1994599.0|
|        Breaking Bad|      1727694.0|
|    Django Unchained|      1472668.0|
| Saving Private Ryan|      1346020.0|
|     Stranger Things|       989090.0|
|    The Walking Dead|       945125.0|
|         Taxi Driver|       795290.0|
|  The Imitation Game|       748654.0|
|   Full Metal Jacket|       723306.0|
|How to Train Your...|       719717.0|
+--------------------+---------------+
only showing top 10 rows
```

```
# Identify titles with low IMDb scores for potential improvement or removal
df.filter(df["imdb_score"] < 6).select("title", "imdb_score").show()
# Identifying titles with low IMDb scores allows for a review of content quality and potential strategies for improvement.
```

```
+--------------------+----------+
|               title|imdb_score|
+--------------------+----------+
|     The Blue Lagoon|       5.8|
|             Dostana|       2.1|
|              Bandie|       4.2|
|         Khoon Khoon|       5.1|
|        Endless Love|       4.9|
|In Defense of a M...|       5.6|
|       Aakhri Adaalat|       5.1|
|A Stoning in Fulh...|       5.8|
|  Muqaddar Ka Faisla|       4.8|
|                Jaal|       5.2|
|              Mujrim|       5.4|
|  3 Ninjas Kick Back|       4.5|
|     Johnny Mnemonic|       5.6|
|I Know What You D...|       5.7|
|          Hollow Man|       5.8|
|            Godzilla|       5.4|
|               Spawn|       5.2|
|   Dennis the Menace|       5.6|
```

```
# Writing some more complex queries to gain insights


# Identify titles with high IMDb votes for each age certification
from pyspark.sql.functions import dense_rank

window_spec = Window().partitionBy("age_certification").orderBy(col("imdb_votes").desc())

df.withColumn("rank", dense_rank().over(window_spec)).filter(col("rank") == 1).orderBy("age_certification").show()
# This query identifies titles with the highest IMDb votes within each age certification category.
```

```
+--------------------+--------------------+--------------------+-----------------+--------------------+--------------------+-----
---------------+--------------------+--------------------+--------------------+--------------------+----+
|               index|                  id|               title|             type|         description|        release_year|   ag
e_certification|             runtime|             imdb_id|          imdb_score|          imdb_votes|rank|
+--------------------+--------------------+--------------------+-----------------+--------------------+--------------------+-----
---------------+--------------------+--------------------+--------------------+--------------------+----+
|                4381|            tm816432|         Bypass Road|            MOVIE|The story revolve...|                2019|     
null|                 137|           tt9176260|                 5.4|                 999|   1|
|                 645|             tm28024| A Very Special Love|            MOVIE|"Laida Magtalas i...| the youngest mem...| ""Ba
chelor"". In...| Laida revels wor...|                2008|                  PG|                 105|   1|
|                2232|             ts79409|A.I.C.O. -Incarna...|             SHOW|"In Japan in the ...| spawning an out-...| 15-y
ear-old Aiko...| who lost her fam...| learns something...| a new student at...| and the answer t...|   1|
|The series has be...| Ancient Ruler Di...| which is made by...| 2007. As of 2008| an English adapt...| but moved to The...|
2008.|                null|                null|                null|                null|   1|
|                 983|             tm35895|             Buddies|            MOVIE|"A road movie tha...| they decide to r...| Anin
ha looks for...| they embark on s...|                2013|                null|                  94|   1|
|It is broadcast i...| in over twelve l...|   including Spanish|           French|             Italian|               Greek|
Arabic|                Thai|             Finnish|              Hebrew|          Portuguese|   1|
|                2358|            tm429377|Bert Kreischer: S...|            MOVIE|"Comedian Bert Kr...| Bert Kreischer: ...| Bert
regales the...| his daughter pra...| and upstaging ex...|                2018|                null|   1|
|                1307|             ts35987|Transformers: Rob...|             SHOW|"Years after the ...| Strongarm (an El...| Grim
```

```
+--------------------+--------------------+--------------------+----------+-----------+------------+-----------------+---------+-
-------+--------------------+------------+-------------------+
|               index|                  id|               title|      type|description|release_year|age_certification|  runtime|
imdb_id|          imdb_score|  imdb_votes|running_total_votes|
+--------------------+--------------------+--------------------+----------+-----------+------------+-----------------+---------+-
-------+--------------------+------------+-------------------+
|,2010,TV-14,58,tt...|                null|                null|      null|       null|        null|             null|     null|
null|                null|        null|               null|
|Being the mid-yea...| it was the first...|                null|      null|       null|        null|             null|     null|
null|                null|        null|               null|
| In the fall of 2008| Upper Deck Compa...|                null|      null|       null|        null|             null|     null|
null|                null|        null|               null|
|It aired in Spain...| on Telecinco for...|                null|      null|       null|        null|             null|     null|
null|                null|        null|               null|
|It was first broa...| on Sundays at 22...|                null|      null|       null|        null|             null|     null|
null|                null|        null|               null|
|Autumn's Concerto...|             Taiwan.|                null|      null|       null|        null|             null|     null|
null|                null|        null|               null|
|Fated to Love You...| Taiwan. It was a...|                null|      null|       null|        null|             null|     null|
null|                null|        null|               null|
|The series premie...| on Mediacorp Cha...|                null|      null|       null|        null|             null|     null|
```