ASSIGNMENT 5
QUERYING THE DATASET USING "SPARK"

Download small MovieLens dataset from this link: small_movielense_dataset.zip

Develop an Apache Spark application to answer the following questions. Do not limit your imagination and understanding:

- Programming languages can be Scala, Python, or any as per your choice
- We expect you to work with RDD.
- Develop the application as you would do in a **real-life project** (use engineering best practices for structuring and implementation; think about the performance, scalability, and maintenance).

Feel free to include documents with your application, to explain how you understand the problem, your approach to the question, caveats & limitations.

Questions:

1. How many "Drama" movies (movies with the "Drama" genre) are there?
2. How many unique movies are rated, how many are not rated?
3. Who gave the most ratings, how many rates did he make?
4. Compute min, average, max rating per movie.
5. Output dataset containing users that have rated a movie but not tagged it.
6. Output dataset containing users that have rated AND tagged a movie.
7. *Output dataset showing the number of movies per Genre per Year (movies will be counted many times if it's associated with multiple genres).

   *Solving the first six questions is compulsory. The seventh question is a bonus question.

ALL THE BEST!