## MUBI Platform Analysis with PySpark

### Introduction:

MUBI is a global streaming platform, production company and film distributor available in over 190 countries. This platform produces and theatrically distributes films by emerging and established filmmakers, which are exclusively available on its platform. The catalogue consists of world cinema films, such as arthouse, documentary and independent films. MUBI platform could be described as the “Netflix for art movies”.

In this project we use PySpark to explore five datasets from MUBI to better understand user activity, film preferences, and the relationship between lists and ratings.

Our analysis is based on the following datasets:

- mubi_lists_data.csv: Contains data from MUBI lists for users who didn’t set their profile in private mode.
- mubi_lists_user_data.csv: Contains user data related to the created lists for users who didn’t set their profile in private mode.
- mubi_movie_data.csv: Contains data from all movies registered on MUBI.
- mubi_ratings_data.csv: Contains data from rating on MUBI for users who didn’t set their profile in private mode.
- mubi_ratings_user_data.csv: Only the user information related to the last rating for a specific day is stored in this table.

### Problem Definition:

We are trying to understand how users interact with the MUBI platform by analyzing their movies lists, ratings and general activity. Specifically, we want to know which films are the most popular, who the most active users are, and how users behave when it comes to ratings and the lists they create.

Questions We Intend to Answer

1. Which movies are most frequently included in user lists?
2. Rating frequency by months
3. Movies with high rating difference (standard deviation)
4. Which movies have the highest ratings?
5. Which users are the most active on MUBI? Are they list creators or movie reviewers?
6. Most popular lists
7. Which movies show up most in popular lists?

**Methodology**:

To explore user behaviour on MUBI, we are using PySpark functions and Spark SQL inside Google Colab to analyze the five CSV datasets from Kaggle. Our goal is to answer some key research questions to resolve the problem. Here is how we approached the analysis:

Tools and Environment:

- Google Colab: It is a free, cloud based platform that allows users to write and execute Python code in a browser based environment, acting as a hosted Jupyter Notebook. It’s particularly useful for machine learning, data science, and education.

- Kaggle: Used to download the MUBI datasets.

- PySpark: It is an open source, distributed computing framework designed for big data processing and analytics. PySpark provides a robust set of tools and libraries for various data processing tasks, including data manipulation, SQL queries, machine learning, and stream processing.

  PySpark Operations Implemented:

  - ❖ Filtering
  - ❖ New column creation
  - ❖ Aggregate functions
  - ❖ Grouping
  - ❖ Sorting
  - ❖ Joins
  - ❖ Window functions
  - ❖ Aggregate window functions

  Each operation will help us answer the research questions listed above. Code implementation will be done in Google Colab using PySpark DataFrames and Spark SQL.

Why These Questions Matter:

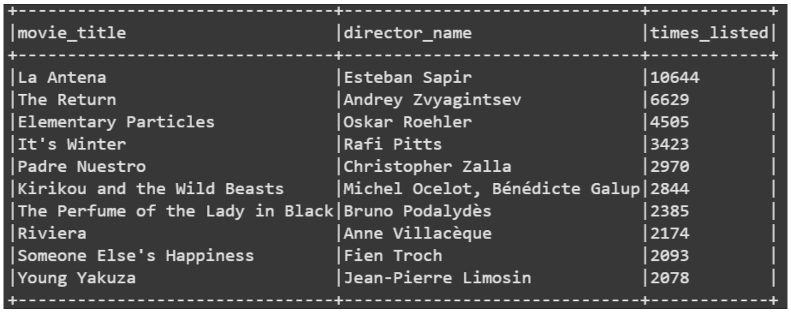
- Most listed and most rated movies show us the community favorites and cultural resonance.
- Rating frequency by month uncovers engagement patterns over time.
- High rating variance figures out how diverse users’ tastes are.
- Top rated films reflect collective critical taste.
- User activity analysis helps identify which users are more into leaving reviews.
- Popular lists and their movie content show how users drive visibility and shared experience.

Workflow:

- Importing and inspecting each dataset
- Cleaning and unifying schemas
- Filtering, joining, and transforming data to extract insights

**Results:**

<u>Question 1:</u> Which movies are most frequently included in user lists?

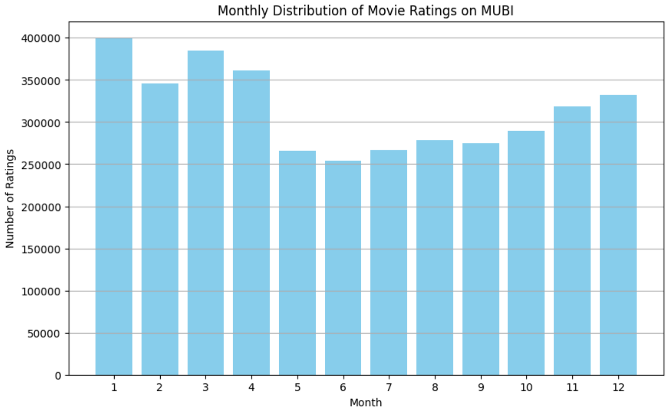


```
+--------------------------------+------------------------------+------------+
|movie_title                     |director_name                 |times_listed|
+--------------------------------+------------------------------+------------+
|La Antena                       |Esteban Sapir                 |10644       |
|The Return                      |Andrey Zvyagintsev            |6629        |
|Elementary Particles            |Oskar Roehler                 |4505        |
|It's Winter                     |Rafi Pitts                    |3423        |
|Padre Nuestro                   |Christopher Zalla             |2970        |
|Kirikou and the Wild Beasts     |Michel Ocelot, Bénédicte Galup|2844        |
|The Perfume of the Lady in Black|Bruno Podalydès               |2385        |
|Riviera                         |Anne Villacèque               |2174        |
|Someone Else's Happiness        |Fien Troch                    |2093        |
|Young Yakuza                    |Jean-Pierre Limosin           |2078        |
+--------------------------------+------------------------------+------------+
```

These results reveal that MUBI users are not primarily drawn to globally recognized blockbusters or mainstream titles. Instead, the most frequently included films are lesser known, often regionally or thematically specific works. For example:

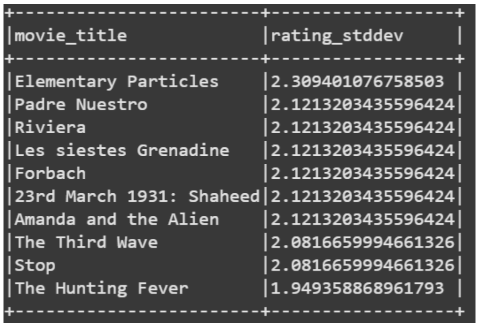
- La Antena is an Argentina surrealist film known for its silent era aesthetics and political symbolism.
- The return is a Russian drama, explores family dynamics and masculinity.
- Elementary Particles is a German controversial novel, touching on existential and social themes.

<u>Question 2:</u> Rating frequency by months



January had the highest rating activity (399,676 ratings), followed by March and December. The lowest activity occurred in June (254,154 ratings) and May. This suggests that MUBI users are most active during the start and end of the year, possibly due to holidays or film challenges.

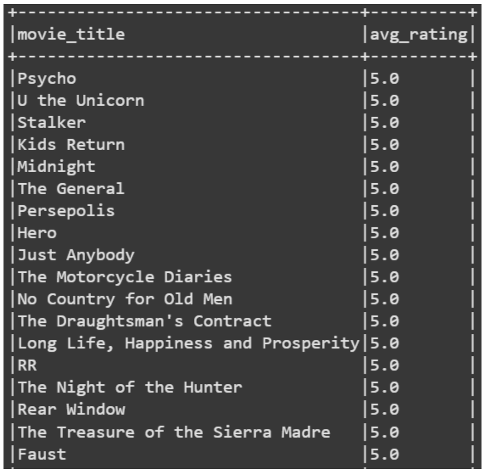
Question 3: Movies with high rating difference (standard deviation)



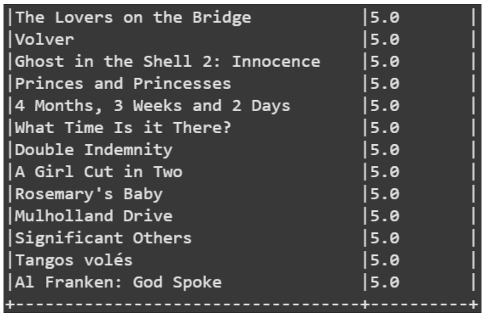
```
+------------------------+------------------+
|movie_title             |rating_stddev     |
+------------------------+------------------+
|Elementary Particles    |2.309401076758503 |
|Padre Nuestro           |2.1213203435596424|
|Riviera                 |2.1213203435596424|
|Les siestes Grenadine   |2.1213203435596424|
|Forbach                 |2.1213203435596424|
|23rd March 1931: Shaheed|2.1213203435596424|
|Amanda and the Alien    |2.1213203435596424|
|The Third Wave          |2.0816659994661326|
|Stop                    |2.0816659994661326|
|The Hunting Fever       |1.949358868961793 |
+------------------------+------------------+
```

The film Elementary Particles showed the highest rating variability, with a standard deviation of 2.31, followed closely by Padre Nuestro and Riviera. This suggests that these movies sparked mixed or polarized reactions among users likely due to controversial themes or unconventional storytelling.

Question 4: Which movies have the highest ratings?



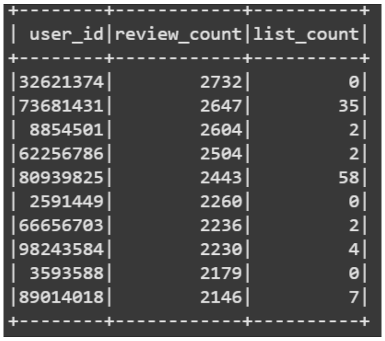
```
+-----------------------------------+----------+
|movie_title                        |avg_rating|
+-----------------------------------+----------+
|Psycho                             |5.0       |
|U the Unicorn                      |5.0       |
|Stalker                            |5.0       |
|Kids Return                        |5.0       |
|Midnight                           |5.0       |
|The General                        |5.0       |
|Persepolis                         |5.0       |
|Hero                               |5.0       |
|Just Anybody                       |5.0       |
|The Motorcycle Diaries             |5.0       |
|No Country for Old Men             |5.0       |
|The Draughtsman's Contract         |5.0       |
|Long Life, Happiness and Prosperity|5.0       |
|RR                                 |5.0       |
|The Night of the Hunter            |5.0       |
|Rear Window                        |5.0       |
|The Treasure of the Sierra Madre   |5.0       |
|Faust                              |5.0       |
```



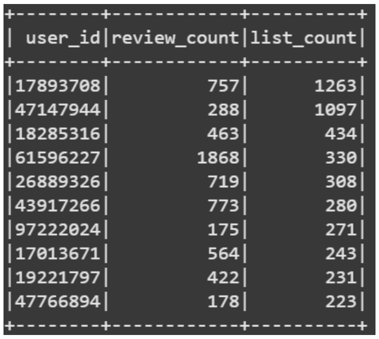
```
|The Lovers on the Bridge           |5.0       |
|Volver                             |5.0       |
|Ghost in the Shell 2: Innocence    |5.0       |
|Princes and Princesses             |5.0       |
|4 Months, 3 Weeks and 2 Days       |5.0       |
|What Time Is it There?             |5.0       |
|Double Indemnity                   |5.0       |
|A Girl Cut in Two                  |5.0       |
|Rosemary's Baby                    |5.0       |
|Mulholland Drive                   |5.0       |
|Significant Others                 |5.0       |
|Tangos volés                       |5.0       |
|Al Franken: God Spoke              |5.0       |
+-----------------------------------+----------+
```

A total of 30 movies received a perfect average rating of 5.0, including critically acclaimed titles like Psycho, Stalker, Persepolis, and No Country for Old Men. This indicates a strong consensus among users in favor of cinematic classics, visually striking films, and emotionally resonant stories. It reflects MUBI's audience preference for high quality cinema.

Question 5: Which users are the most active on MUBI? Are they list creators or movie reviewers?



```
+--------+------------+----------+
| user_id|review_count|list_count|
+--------+------------+----------+
|32621374|        2732|         0|
|73681431|        2647|        35|
| 8854501|        2604|         2|
|62256786|        2504|         2|
|80939825|        2443|        58|
| 2591449|        2260|         0|
|66656703|        2236|         2|
|98243584|        2230|         4|
| 3593588|        2179|         0|
|89014018|        2146|         7|
+--------+------------+----------+
```

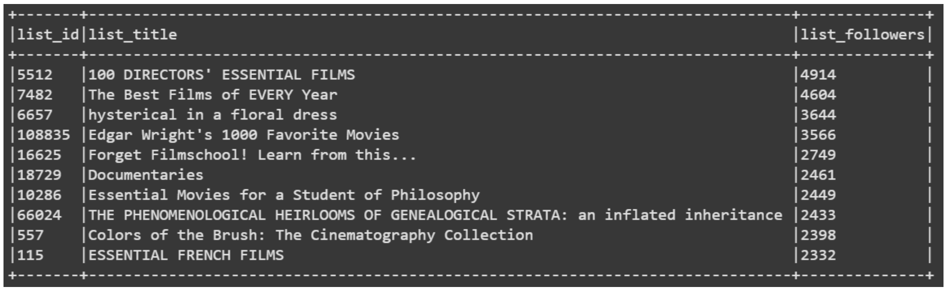


```
+--------+------------+----------+
| user_id|review_count|list_count|
+--------+------------+----------+
|17893708|         757|      1263|
|47147944|         288|      1097|
|18285316|         463|       434|
|61596227|        1868|       330|
|26889326|         719|       308|
|43917266|         773|       280|
|97222024|         175|       271|
|17013671|         564|       243|
|19221797|         422|       231|
|47766894|         178|       223|
+--------+------------+----------+
```

The most active reviewers, such as user 32621374 with over 2,700 reviews, primarily focused on rating films but created no lists. In contrast, users like 17893708 created over 1,200 lists, showing a strong preference for curating rather than rating.

Only a few users were highly active in both areas, and the overall correlation between review count and list count was just 0.22, indicating that reviewers and curators tend to be different types of users. This reinforces the idea that MUBI hosts distinct user behaviors, some prioritize watching and rating, while others focus on organizing and sharing film collections.

Question 6: Most popular lists



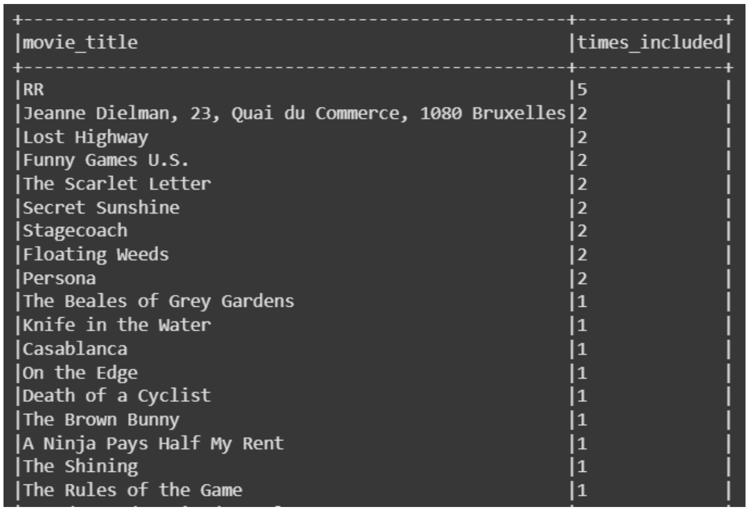
```
+-------+-------------------------------------------------------------------------------+--------------+
|list_id|list_title                                                                     |list_followers|
+-------+-------------------------------------------------------------------------------+--------------+
|5512   |100 DIRECTORS' ESSENTIAL FILMS                                                 |4914          |
|7482   |The Best Films of EVERY Year                                                   |4604          |
|6657   |hysterical in a floral dress                                                   |3644          |
|108835 |Edgar Wright's 1000 Favorite Movies                                            |3566          |
|16625  |Forget Filmschool! Learn from this...                                          |2749          |
|18729  |Documentaries                                                                  |2461          |
|10286  |Essential Movies for a Student of Philosophy                                   |2449          |
|66024  |THE PHENOMENOLOGICAL HEIRLOOMS OF GENEALOGICAL STRATA: an inflated inheritance |2433          |
|557    |Colors of the Brush: The Cinematography Collection                             |2398          |
|115    |ESSENTIAL FRENCH FILMS                                                         |2332          |
+-------+-------------------------------------------------------------------------------+--------------+
```

The most followed lists on MUBI are:

- 100 DIRECTORS' ESSENTIAL FILMS (4,914 followers)
- The Best Films of EVERY Year (4,604 followers)
- Edgar Wright's 1000 Favorite Movies (3,566 followers)

These titles suggest that users are especially drawn to lists that offer expert guidance, historical overviews, or personal curation from well known filmmakers. The popularity of lists centered on film education, auteur cinema, and visual aesthetics shows that users value lists as tools for discovery, learning, and taste refinement.

Question 7: Which movies show up most in popular lists?



```
+-------------------------------------------------+--------------+
|movie_title                                      |times_included|
+-------------------------------------------------+--------------+
|RR                                               |5             |
|Jeanne Dielman, 23, Quai du Commerce, 1080 Bruxelles|2          |
|Lost Highway                                     |2             |
|Funny Games U.S.                                 |2             |
|The Scarlet Letter                               |2             |
|Secret Sunshine                                  |2             |
|Stagecoach                                       |2             |
|Floating Weeds                                   |2             |
|Persona                                          |2             |
|The Beales of Grey Gardens                       |1             |
|Knife in the Water                               |1             |
|Casablanca                                       |1             |
|On the Edge                                      |1             |
|Death of a Cyclist                               |1             |
|The Brown Bunny                                  |1             |
|A Ninja Pays Half My Rent                        |1             |
|The Shining                                      |1             |
|The Rules of the Game                            |1             |
```

The most popular movie in the lists are:

- RR (Railroad) (5 times)

These films indicate a strong overlap between personal curation and community influence. Their repeated presence suggests that MUBI users not only value these titles individually but also recognize them as essential recommendations for others.

The films range from international dramas to experimental animation, highlighting that diversity, originality, and emotional depth are highly appreciated across the platform’s most impactful lists.

**Problems:**

What problems did you encounter?

1. Finding a suitable topic with compatible datasets: One of the biggest initial challenges was identifying a topic that had publicly available datasets large enough to meet the project's 1GB requirement and that also contained multiple CSV files. This significantly limited the range of topics we could explore.

How did you solve them?

1. To solve the dataset issue, we used Kaggle’s advanced filters to search by file size, number of files, and format type, which helped us discover the MUBI dataset, a rich, structured, and thematically interesting resource.

**Project Summary:**

This project gave us a strong understanding of how MUBI users engage with content. The most rated films are often international, emotionally resonant, or visually bold, showing that users are highly intentional in their film discovery.

We also learned the value of segmenting users into list creators vs. reviewers, and saw how list popularity helps define which films gain broader community traction.

From a technical perspective, the project strengthened our ability to work with real world, messy data using PySpark, and showed how filtering, joins, and aggregations come together to support real insights at scale.

**References:**

- https://en.wikipedia.org/wiki/Mubi_(streaming_service)#:~:text=Mubi%20
- https://mubi.com/en/ca
- https://www.kaggle.com/datasets/clementmsika/mubi-sqlite-database-for-movie-lovers?select=mubi_db.sqlite
- https://colab.google/#:~:text=Google%20Colaboratory,Blog
- https://domino.ai/data-science-dictionary/pyspark
- https://www.theguardian.com/culture/2008/may/16/worldcinema.drama
- https://en.wikipedia.org/wiki/RR_(film)
- https://www.imdb.com/title/tt0430051/