Phase 1 – Project Idea Submission

Project Title: Web Scraping & Analysis of IMDb's Top 250 Movies

1. Website to Scrape

For our project, we've chosen to scrape data from the IMDb Top 250 Movies page. This list includes the highest-rated movies on IMDb and provides key information about each film, making it ideal for analysis.

---

2. What Data Will Be Collected

From the IMDb Top 250 page and individual movie links, we'll collect the following details:

- **Movie Title**
- **Genre**
- **Release Year**
- **IMDb Rating**
- **Main Cast**
- **Number of Ratings**
- **Director Name**

This dataset will allow us to explore various aspects of popular cinema and audience preferences.

---

3. Project Objective

The goal of this project is to analyze the top-rated movies on IMDb to discover:

- Which **genres**, **actors**, and **directors** appear most frequently among the top 250 movies.
- How audience preferences have changed **across decades** (e.g., genre trends, rating patterns).
- Which genres or creative individuals are linked to **consistently high ratings** and **viewer engagement** (number of ratings).

Our insights could be helpful for film students, researchers, or even studios looking to understand what types of films tend to succeed with audiences.

---

4. Data Cleaning & Processing

Once the raw data is extracted, we'll clean and organize it using **Pandas**. We'll convert the data into a structured **DataFrame**, then save it as a **JSON file** for easier use later.

We'll also apply **Regular Expressions (Regex)** to handle text cleaning tasks like:

- Removing unwanted characters or formatting.
- Extracting and cleaning email/text fields (if any).
- Standardizing date formats.

The result will be a clean and ready-to-analyze dataset.

---

5. Data Analysis & Visualization

Using **Seaborn** and **Matplotlib**, we'll perform both statistical and visual analysis to uncover meaningful trends. Our focus will be on answering questions like:

- What are the most **popular genres** among the top 250?
- Which **actors and directors** appear most often in highly rated movies?
- How do movie ratings and genres vary across **different decades**?
- Are there differences between movies with **high ratings vs. high engagement**?

Planned Visuals:

- **Bar Charts** for top genres, directors, and actors.
- **Line Graphs** showing rating or genre trends over time.
- **Heatmaps** to explore correlations between genres, ratings, and number of ratings.
- **Pie Charts** to show genre distributions.
- **Box Plots** to highlight rating spreads across decades or genres.

---

6. Data Storage

After processing, we'll save the final dataset as a **JSON file**. This format is easy to work with, portable, and can be reused in future projects, web apps, or databases like MongoDB if needed.

---

7. (Optional) Web App Deployment with Streamlit

If time allows, we'd like to build a simple **Streamlit app** to showcase our findings interactively. Users could:

- Browse visual summaries.
- Filter results by decade or genre.
- Explore trends in a more dynamic way.

This would be a bonus step to improve the presentation and make the project more engaging.
