Milestone 1 Report: Data Collection, Preprocessing, and Exploratory Data Analysis (EDA)

1. Project Objective

The goal of this project is to develop a **data science tool** that facilitates **data-driven decision-making** by analyzing multiple datasets and presenting insights interactively. The tool will take the form of a **dashboard** that allows users to explore relationships, trends, and key patterns within the data.

The tool will focus on **movie-related datasets**, integrating data from platforms such as IMDb, Rotten Tomatoes, and Netflix. The dashboard will assist users in understanding key factors that influence movie success, audience engagement, and critical reception.

2. Data Used

Three datasets have been selected for analysis:

1. **Rotten Tomatoes Movies Dataset** – Contains metadata about movies, including ratings, genres, and production details.
2. **Rotten Tomatoes Reviews Dataset** – Includes audience and critic reviews for various movies.
3. **Netflix Movies Dataset** – Provides information about movies available on Netflix, such as release year, runtime, and genre.
4. **IMDb Movies Dataset** – Contains extensive details about movies, including user ratings, director information, and box office earnings.

These datasets were obtained from **Kaggle** and other public repositories.

3. Technology Stack

The project is implemented using the following tools and technologies:

- **Programming Language:** Python
- **Data Manipulation & Analysis:** Pandas, NumPy
- **Machine Learning & Statistical Analysis:** Scikit-learn, SciPy
- **Visualization:** Matplotlib, Seaborn, Plotly Dash
- **Interactive Dashboard Development:** Streamlit
- **Jupyter Notebook** for exploratory data analysis and iterative testing

4. Project Timeline

The project follows the **CRISP-DM methodology** and consists of three key milestones:

Milestone	Task Description	Start Date	End Date
Milestone 1	Data collection, preprocessing, exploratory data analysis (EDA)	**Feb 5, 2025**	**Feb 21, 2025**
Milestone 2	Feature engineering, feature selection, model development	**Feb 21, 2025**	**Mar 21, 2025**
Milestone 3	Model evaluation, tool development, final presentation	**Mar 24, 2025**	**Apr 23, 2025**

5. Exploratory Data Analysis (EDA) Report

5.1 Data Preprocessing Steps

- **Handling Missing Data:**
  - Numerical columns were imputed using mean/median values.
  - Categorical missing values were filled using mode or removed if the missing percentage exceeded a threshold.

- **Outlier Detection & Treatment:**
  - Z-score thresholding and IQR methods were applied to remove extreme outliers in numerical columns.

- **Feature Scaling & Encoding:**
  - Continuous features were normalized using **Min-Max Scaling**.
  - Categorical variables were **one-hot encoded** or **label encoded**, depending on cardinality.

5.2 Key Insights from EDA

Descriptive Statistics

We analyzed the datasets to understand the distribution and characteristics of key numerical features. Below are some key insights from the descriptive statistics:

- **Rotten Tomatoes Movies Dataset:**
    - Contains **15,625 movies**, with an average tomatometer rating of **60%** (std = 28.2%).
    - The **audience rating** has a mean of **59.8%** (std = 20.5%).
    - The number of **critics' reviews** per movie is relatively low, with an average tomatometer count of **15.9%** of the maximum observed value.
    - The dataset covers movies from various **genres, directors, and production companies**, with significant variability.
    - **Runtime distribution** is relatively uniform, with most movies ranging from **~90 to 120 minutes**.

- **Rotten Tomatoes Reviews Dataset:**
    - Contains **over 1.13 million reviews**, covering thousands of movies and critics.
    - The **average review score** is **474.5** (std = 288.6), with a wide range.
    - Review dates span a long period, with some older and newer movies included.

- **Netflix Movies and TV Shows Dataset:**
    - Covers **8,590 titles**, including both movies and TV shows.
    - The **average movie duration** is **~89 minutes**, though TV show episodes introduce variability.
    - The **release years** range widely, with a mix of older and recent content.
    - A diverse set of **countries and genres** is represented.

- **IMDb Top 1000 Movies Dataset:**
    - Contains **809 top-rated movies** based on IMDb rankings.
    - The **average IMDb rating** is **~6.3**, with some highly rated movies reaching **9+**.
    - The **meta score** has a mean of **62.9**, indicating mostly well-received movies.
    - The **gross revenue** varies significantly, with an average of **~$397 million** but high deviation.

5.2.2 Data Visualization Findings

1. Correlation Matrix

- The correlation matrix reveals strong relationships among various features.
- `tomatometer_rating` and `audience_rating` have a moderate positive correlation, indicating that movies well-rated by critics also tend to be liked by audiences.
- `tomatometer_count` and `tomatometer_top_critics_count` show high correlation, suggesting that the number of critic reviews is proportional to the number of top critics who review the movie.
- `audience_count` and `audience_rating` have a weak correlation, implying that popularity (measured by count) does not strongly influence audience ratings.

2. Histograms of Numerical Features

- `tomatometer_rating` is skewed toward higher values, suggesting that most movies receive favorable critic ratings.
- `audience_count`, `tomatometer_count`, and `tomatometer_top_critics_count` exhibit right-skewed distributions, meaning a small number of movies receive an exceptionally high number of reviews.
- `runtime` follows a roughly normal distribution, with most movies having an average runtime.

3. Pairplot (Scatter Matrix)

- Some scatter plots show distinct patterns, such as `audience_rating` vs. `tomatometer_rating`, where a weak but visible positive trend exists.
- Several relationships appear non-linear, indicating that transformations might be necessary for better modeling.
- Outliers are present, particularly in review count variables (`tomatometer_count`, `audience_count`), which might require further investigation.

4. Box Plots of Numerical Features

- `tomatometer_count`, `audience_count`, and `tomatometer_top_critics_count` contain numerous outliers, suggesting that some movies receive an unusually high number of reviews.
- `tomatometer_rating` and `audience_rating` have a relatively symmetric distribution but show some extreme values.
- `runtime` has a few outliers, indicating a small number of movies that are significantly longer or shorter than average.

6. Next Steps

Following the completion of Milestone 1, we will proceed to:

- **Feature Engineering (Milestone 2):**
  - Select the most relevant features using statistical methods.
  - Create new derived features (e.g., revenue-to-budget ratio, engagement scores).
  - Train and evaluate machine learning models for predicting movie success.
- **Tool Development (Milestone 3):**
  - Design an interactive dashboard using **Streamlit** or **Plotly Dash**.
  - Implement model-driven insights for user interaction.
  - Finalize reports and prepare the final presentation.