Anime Dataset Analysis Report

1. Project Overview

The purpose of this project is to analyze the anime.csv dataset, which contains information about various anime, including their type, genres, number of episodes, ratings, and popularity. The dataset contains 1,864 records and provides insights into trends, relationships, and key characteristics of anime series.

---

2. Dataset Description

- **Total Records:** 1,864
- **Columns:** anime_id, name, genre, type, episodes, rating, members
- **Data Types:**

Column	Data Type
anime_id	float64
name	object
genre	object
type	object
episodes	object
rating	float64
members	float64

Column Information:

- anime_id: Unique identifier for each anime.
- name: Name/title of the anime.
- genre: One or more genres of the anime.
- type: Type of anime (TV, Movie, OVA, ONA, Special).
- episodes: Number of episodes (some unknown).
- rating: Average user rating.
- members: Number of members/users who have added the anime to their list.

---

3. Data Cleaning and Preparation

Before performing analysis, it is essential to **clean and prepare the dataset**. Raw data often contains inconsistencies, missing values, or incorrect data types. For this project, the following steps were performed:

1. **Convert episodes to numeric:**
   - The episodes column initially contained string values, including "Unknown" or missing values.
   - These were converted to numeric (int) values, and any unknown entries were treated as null to avoid errors during analysis.
   - This step ensures that **aggregations, correlations, and visualizations** using the episodes field are accurate.
2. **Convert numeric columns (anime_id, rating, members) to appropriate types:**
   - Some numeric columns were initially read as strings due to CSV formatting.
   - Converting them to numeric types (float) ensures that operations like **averages, sorting, and correlations** are correctly computed.
3. **Split genre into lists:**
   - Many animes have multiple genres listed together separated by commas.
   - Splitting the genre column into lists allows for **genre-level analysis** such as counting the frequency of each genre, finding top genres, and visualizing rating trends per genre.
4. **Handle missing values:**
   - Any missing or null values in key columns like rating, episodes, or members were carefully handled to avoid skewed results.
   - For example, animes with unknown episodes were excluded from certain visualizations like scatter plots involving episodes.

Result:
The dataset became **clean, structured, and ready for analysis**, enabling accurate computations and meaningful visualizations.

---

4. Analysis and Visualizations

After data cleaning, the dataset was analyzed using **descriptive statistics, relationship exploration, and correlations**.

4.1 Descriptive Statistics

- **Average rating:** ~7.0, indicating most animes are moderately well-rated by users.
- **Average number of episodes:** ~26, showing that most animes are short or mid-length series.

- **Popularity measured by members:** Varies widely; some animes have millions of members, while others have only a few hundred.
- This step gives a **general overview** of the dataset and helps identify patterns and outliers.

4.2 Relationship Analysis

- **Members vs Rating:**
    - Scatter plots show **weak correlation**, meaning highly rated animes are not always the most popular.
    - Popular animes often have **moderate to high ratings**, suggesting a balance between quality and mass appeal.
- **Episodes vs Rating:**
    - Most animes have <50 episodes, but some **long-running series** show high ratings.
    - This helps identify series that are both **long and well-received**.
- **Rating by Type:**
    - Boxplots show that **TV series** have a more diverse rating distribution, while **Movies and OVAs** tend to cluster around higher ratings.
    - This indicates that TV shows vary more in quality, whereas movies are often rated more uniformly.
- **Distribution of Episodes:**
    - Histograms reveal that short series (<50 episodes) dominate the dataset.
    - Very long series (>100 episodes) are rare but often popular.
- **Top Genres:**
    - Action, Comedy, and Adventure are the most common genres.
    - Genre analysis helps understand **trends and user preferences**.
- **Top Animes by Members:**
    - Identified the **most popular anime series** based on member counts.
    - This shows which series have the largest fanbase.

4.3 Correlation

- Correlation analysis was performed on numeric columns: episodes, rating, and members.
- **Members vs Rating:** Weak positive correlation (~0.15), indicating **popularity is not strongly determined by rating**.
- A **heatmap** visually displays correlations, helping identify patterns and relationships between numeric fields.

4.4 Top Insights

- **Highest rated anime:** Identified the anime with the maximum rating.
- **Most popular anime:** Anime with the highest number of members.
- **TV type anime dominates** the dataset, followed by Movies and OVAs.
- **Popular genres** include Action, Comedy, and Adventure, showing trends in anime preferences.

---

5. Charts and Graphs Produced

To visualize the analysis, the following charts were produced:

1. **Scatter plot: Members vs Rating** – shows popularity vs quality.
2. **Histogram: Episodes distribution** – shows how many animes fall into different episode ranges.
3. **Boxplot: Rating by Type** – compares rating variations among TV, Movie, and OVA.
4. **Heatmap: Correlation between numeric fields** – shows relationships between episodes, rating, and members.
5. **Bar chart: Top 10 anime genres** – highlights the most common genres in the dataset.
6. **Bar chart: Top 10 animes by rating (with >1000 members)** – identifies the best-rated animes with significant popularity.
7. **Scatter plot: Episodes vs Members** – shows how series length relates to popularity.

Purpose of Charts:

- Provide **visual understanding** of trends, distributions, and relationships.
- Help **identify outliers**, such as extremely popular or highly rated animes.
- Support insights and conclusions with **clear graphical evidence**.

---

6. Conclusion

The analysis of the anime.csv dataset revealed:

- TV anime are the most common type and show diverse ratings.
- Action, Comedy, and Adventure are dominant genres.
- Members count is only weakly correlated with ratings, suggesting popularity is not solely determined by quality.
- Most anime have fewer than 50 episodes, while long-running series are fewer but often highly rated.

This project demonstrates how PySpark and Python visualization tools (matplotlib, seaborn) can be used to explore, clean, and analyze structured datasets, producing both quantitative insights and visual representations.