Movie Dataset Analysis Report

Prepared by: Prashant Gupta

Dataset Size: 9,827 records

Total Columns (after cleaning): 6

Tools Used: Python (Pandas, Matplotlib, Seaborn), Jupyter Notebook

---

1. Objective

This project analyzes a comprehensive movie dataset to uncover meaningful insights about genre trends, movie popularity, and voting patterns. The goal is to understand user preferences and content trends over time, which can help in recommendation systems, production planning, and viewer analytics.

---

2. Dataset Cleaning & Preparation

- **Initial Size:** 25,551 entries
- **Cleaned Size:** 9,827 entries (after removing null and duplicate records)
- **Dropped Columns:**
  - `Overview`
  - `Original Language`
  - `Poster URL`

Data Transformations

- `Release_Date` converted to `Year` for simplified analysis
- Genres split from comma-separated strings into list format and **exploded** for granular analysis
- Genre column cast to categorical type
- Removed all `NaN` entries for clean modeling

---

3. Vote Average Categorization

Category	Description
Not Popular	Vote Avg < 3
Below Average	3 ≤ Vote Avg < 5
Average	5 ≤ Vote Avg < 7
Popular	Vote Avg ≥ 7

Each group contains around 2,400–2,500 entries, allowing for balanced classification-based insights.

---

4. Genre Distribution Analysis

- **Total Unique Genres:** 19
- **Most Frequent Genre: Drama** (over 14%)
- **Exploded Data:** Allowed each genre to be treated as an individual record
- **Genre Type:** Converted to categorical for memory efficiency and plotting ease

---

5. Statistical Metrics & Insights

Vote Average

- **Mean Vote Average:** ~6.2
- **Median Vote Average:** ~6.3
- **Standard Deviation:** ~1.0

Popularity

- **Mean Popularity:** ~200
- **Median Popularity:** ~50
- **Standard Deviation:** High skew due to outliers like *Spider-Man*

Correlation

- **Pearson Correlation between Popularity & Vote Average:** ~0.38
  Interpretation: Moderate positive correlation – higher-rated movies tend to be more popular, but it's not a perfect relationship.

---

6. Popularity Insights

- **Most Popular Movie:** *Spider-Man: No Way Home*
    - Popularity Score: **5083.954**
- **Least Popular Movie:** *The United States vs. Billie Holiday*
    - Popularity Score: **13.354**
- **Vote Average for Spider-Man:** 8.3 (Popular category)

---

7. Trends Over Time

- **Peak Movie Production:** Post-2010
- **Most Released Genre:** Drama
- **Trends:**
    - Sci-Fi and Action have grown in popularity in recent years
    - Historical and War films have declined

---

8. Technical Summary

- **Language Used:** Python
- **Libraries:**
    - **Pandas:** Data cleaning, transformation
    - **Matplotlib & Seaborn:** Visualization
    - **Numpy:** Statistical calculations
    - **Jupyter Notebook:** Code execution and analysis

---

9. Conclusion

This project gave clear insight into the nature of movie content, viewer trends, and data distribution. The categorization of popularity, deep genre breakdown, and metric-based analysis helped highlight real-world patterns.

---

10. Future Work

- **Machine Learning Models:**
    - Use features like genre, vote average, and release year to **predict popularity** or **recommend movies**
    - Classification models (Random Forest, SVM) or Regression models for predicting popularity score

- **Sentiment Analysis:**
  - Combine user reviews or descriptions (overview) for **text-based sentiment prediction**
- **Web App / Dashboard:**
  - Build an interactive dashboard using **Streamlit** or **Plotly Dash** for visual exploration
- **Time-Series Analysis:**
  - Use yearly trends to forecast movie production and genre popularity shifts