Netflix Movie Data Analysis

1. Project Overview

This project focuses on analyzing a Netflix movie dataset containing 9,827 records and 9 key attributes, including release dates, movie titles, overviews, popularity scores, votes, languages, genres, and poster links. Using Python, we performed exploratory data analysis (EDA) to understand the structure of the dataset, identify patterns, and uncover meaningful insights. The analysis includes cleaning the data, examining distributions, and creating visualizations that help explain trends in movie genres, popularity, votes, and release years. The goal is to better understand what types of movies dominate the platform and how viewers engage with them.

2. Dataset Summary

- Rows: 9827 - Columns: 9 - Key Features:
- Release Date, Title, Overview, Popularity, Vote Count, Vote Average, Original Language, Genre, Poster Url.
- Our dataset looks a bit tidy with no NaNs nor duplicated value.

3. Exploratory Data Analysis using Python

We began with data preparation and cleaning in Python:

- **Data Loading:** Imported the dataset using pandas.

- **Initial Exploration:** Used df.info() to check structure and .describe() for summary statistics.

```
RangeIndex: 9827 entries, 0 to 9826
Data columns (total 9 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Release_Date       9827 non-null   object
 1   Title              9827 non-null   object
 2   Overview           9827 non-null   object
 3   Popularity         9827 non-null   float64
 4   Vote_Count         9827 non-null   int64
 5   Vote_Average       9827 non-null   float64
 6   Original_Language  9827 non-null   object
 7   Genre              9827 non-null   object
 8   Poster_Url         9827 non-null   object
dtypes: float64(2), int64(1), object(6)
```

	Popularity	Vote_Count	Vote_Average
count	9827.000000	9827.000000	9827.000000
mean	40.326088	1392.805536	6.439534
std	108.873998	2611.206907	1.129759
min	13.354000	0.000000	0.000000
25%	16.128500	146.000000	5.900000
50%	21.199000	444.000000	6.500000
75%	35.191500	1376.000000	7.100000
max	5083.954000	31077.000000	10.000000

- **Missing Data Handling:** Checked for null values and our dataset looks a bit tidy with no NaNs nor duplicated value.

- **Feature Engineering:**

    ○ Release_Date column needs to be casted into date time and to extract only the year value.

    ○ Overview, Original_Language and Poster_Url wouldn't be so useful during analysis, so we have drop them.

- **Data Consistency Check:**

    ○ Vote_Average better be categorised for proper analysis. So, We have cut the Vote_Average values and made 4 categories: popular, average, below avg, not popular to describe it more using categorize col() function.

    ○ Genre column has comma separated values and white spaces that needs to be handled and casted into category.

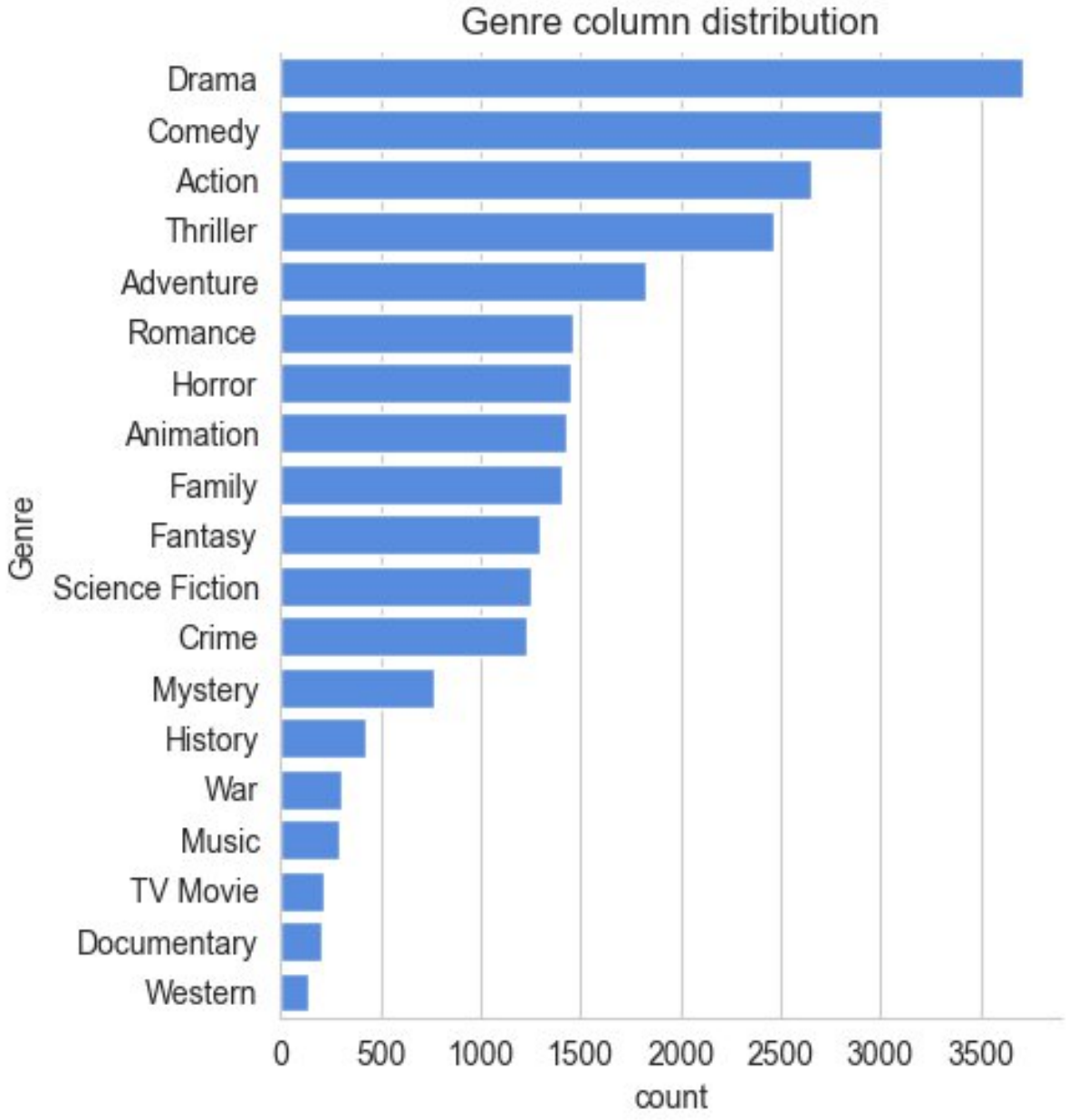
    ○ We have split genres into a list and then exploded our dataframe to have only one genre per row for each movie.

4. Data Visualization using Python

We performed structured visualization in Python to answer key business questions:

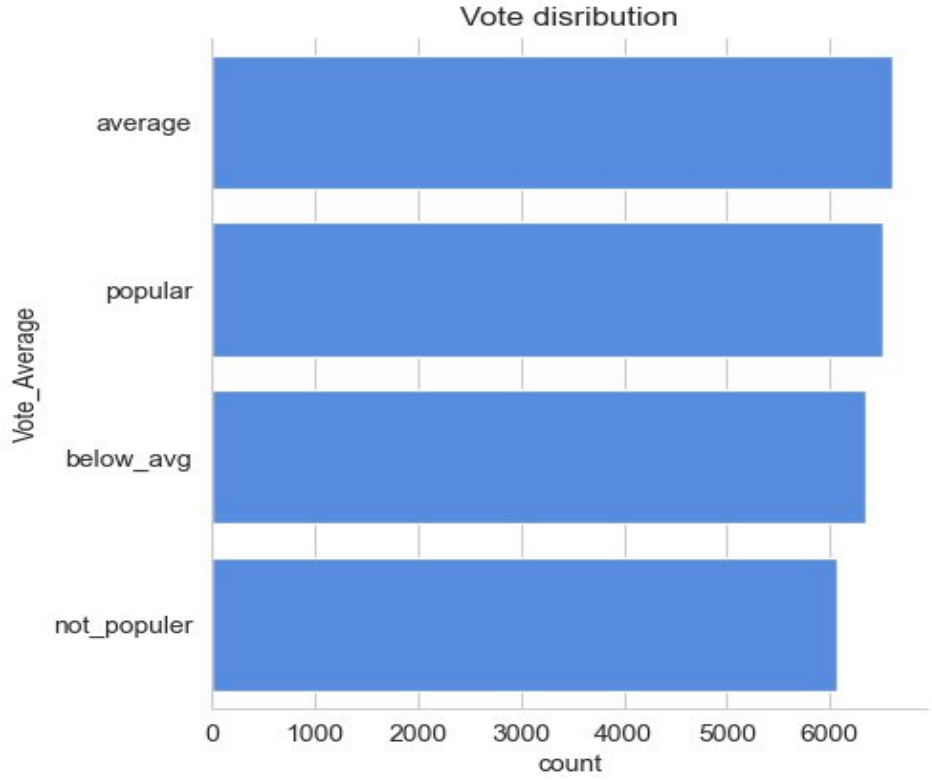
1. What is the most frequent genre of movies released on Netflix?

à Drama is the most frequent genre in our dataset and has appeared more than 14% of the times among 19 other genres.



2. What genre has the highest votes?

à We have 25.5% of our dataset with popular vote (6520 rows). Drama again gets the highest popularity among fans by being having more than 18.5% of movies popularities.



3. What movie got the highest popularity? what's it's genre?

à Spider-Man: No way Home has the highest popularity rate in our dataset and it has genres of Action, Adventure and Science Fiction.

	Release_Date	Title	Popularity	Vote_Count	Vote_Average	Genre
0	2021	Spider-Man: No Way Home	5083.954	8940	popular	Action
1	2021	Spider-Man: No Way Home	5083.954	8940	popular	Adventure
2	2021	Spider-Man: No Way Home	5083.954	8940	popular	Science Fiction

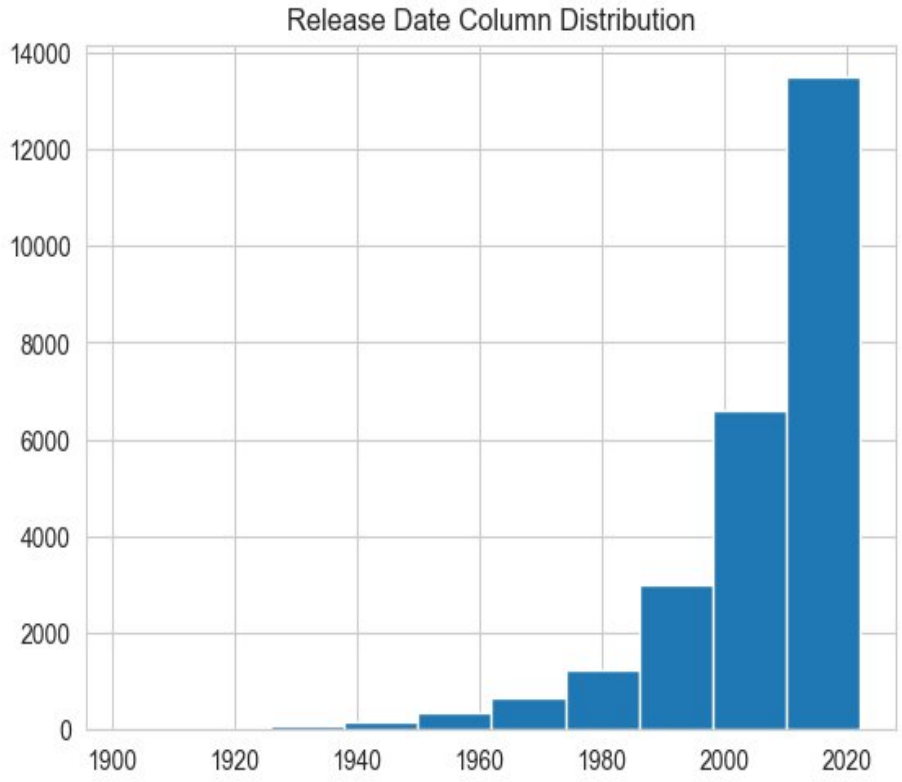
4. What movie got the lowest popularity? What's it's genre?

à The United States, Threads has the lowest popularity rate in our dataset and it has genres of music, drama, war, sci-fi and history.

	Release_Date	Title	Popularity	Vote_Count	Vote_Average	Genre
25546	2021	The United States vs. Billie Holiday	13.354	152	average	Music
25547	2021	The United States vs. Billie Holiday	13.354	152	average	Drama
25548	2021	The United States vs. Billie Holiday	13.354	152	average	History
25549	1984	Threads	13.354	186	popular	War
25550	1984	Threads	13.354	186	popular	Drama
25551	1984	Threads	13.354	186	popular	Science Fiction

5. Which year has the most filmed movies?

à Year 2020 has the highest filming rate in our dataset.



5. Business Recommendations

- **Boost Subscriptions** – Promote exclusive benefits for subscribers.
- **Recommendations** – Recommend the movie genres based on the highest popularity for maximum audience engagement.
- **Review Discount Policy** – Balance subscription boosts with margin control.
- **Movie Positioning** – Highlight Highest voted genre and most popular movies.
- **Targeted Marketing** – Focus efforts on high-revenue age groups and highly engaging users.