**Netflix Movie Data Analysis Project**

**Problem Statement**

Netflix relies on data-driven insights to optimize its content strategy and recommendation algorithms. With a dataset of 9,000+ movies, this project addresses critical business questions to identify trends in genres, popularity, audience engagement, and production timelines. The goal is to uncover actionable insights that inform content acquisition and personalized recommendations.

**Objective**

To analyze Netflix movie data and answer:

1. **Content Trends**: Identify the most frequent genres and yearly production patterns.

2. **Audience Preferences**: Determine genres/movies with the highest votes and popularity.

3. **Strategic Insights**: Highlight underperforming genres and peak production years.

**Dataset Overview**

* **Initial Size**: 9,827 rows × 9 columns.
* **Cleaned Data**: 9,827 rows × 6 columns after preprocessing.
* **Key Columns**: `title`, `genre`, `release\_year`, `popularity`, `vote\_average`, ‘Vote\_count’

**Tools & Technologies**

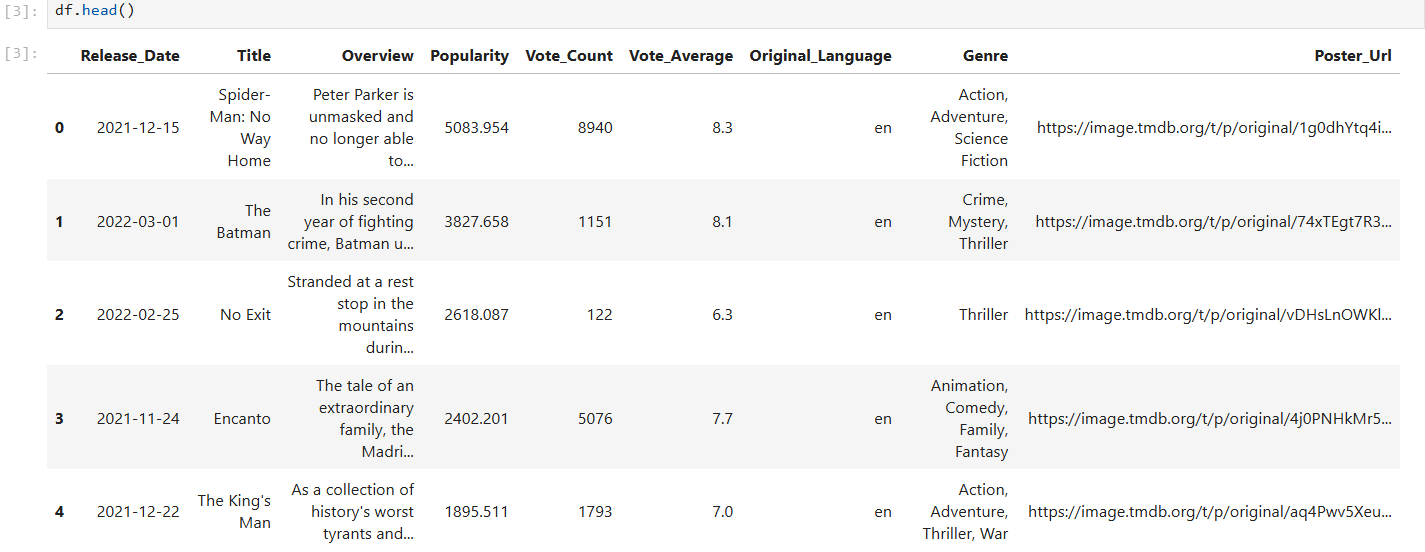
**Languages**: Python

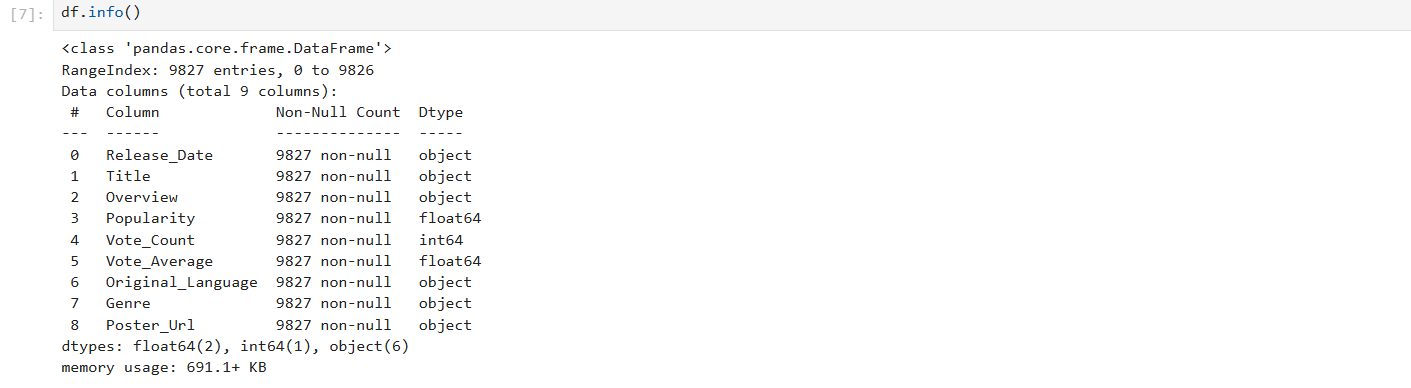
**Libraries**:

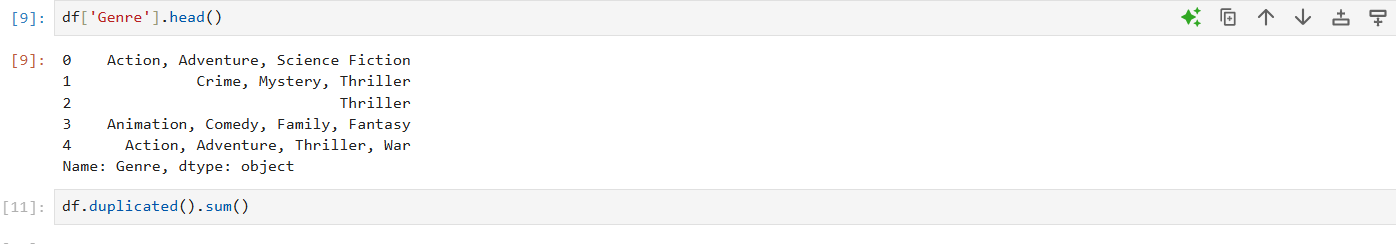
* `Pandas` (data manipulation)
* -`Matplotlib`/`Seaborn` (visualization)

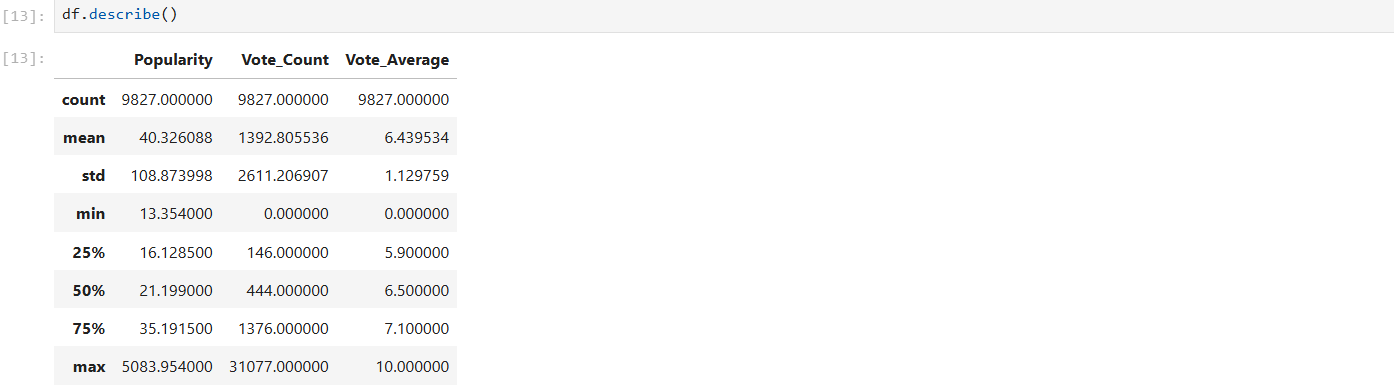
**Steps involved:**

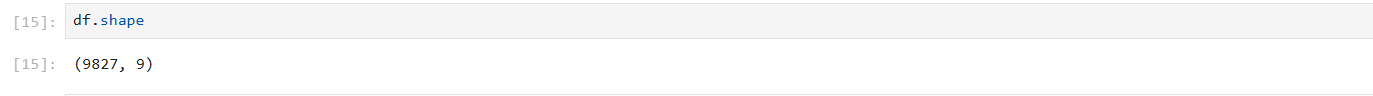






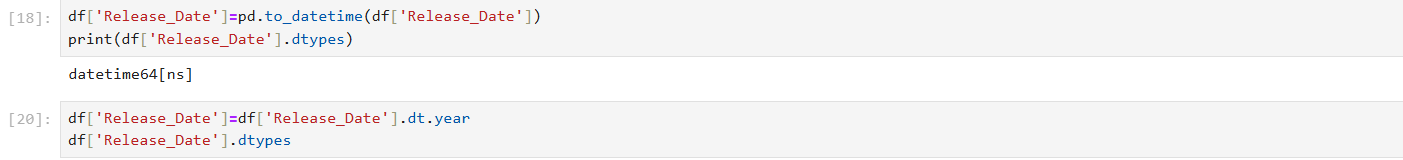


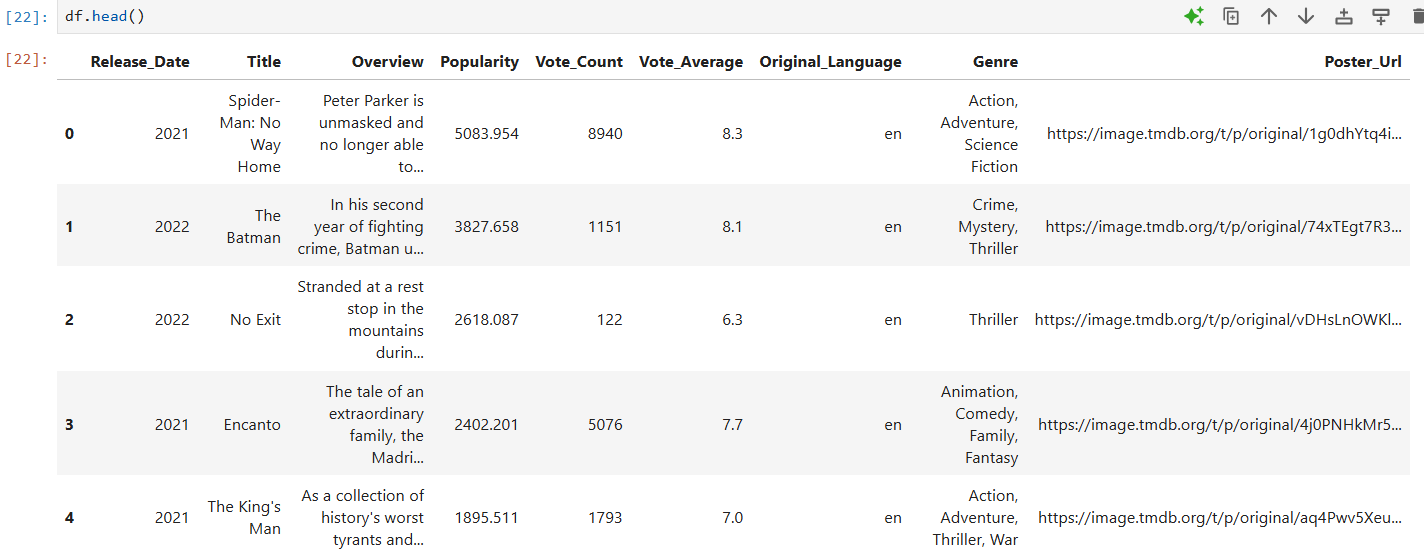




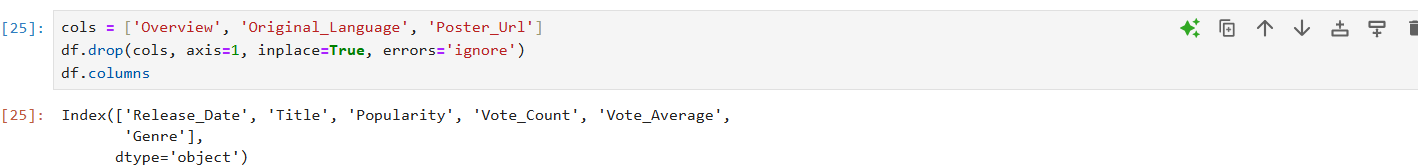
**Exploration summary:**

* **we have a dataframe consisting of 9827 rows and 9 columns.**
* **our dataset looks a bit tidy with no NaNs nor duplicated values**
* **release date column needs to be casted into date time and to extract only year value.**
* **overview, original\_language and poster-url wouldn't be useful during analysis, so we'll drop them .**
* **there is noticable outliers in Popularity column**
* **Vote\_average better be categorised for proper analysis. genre columm seperated values and white spaces that needs to be handled and casted into category.**



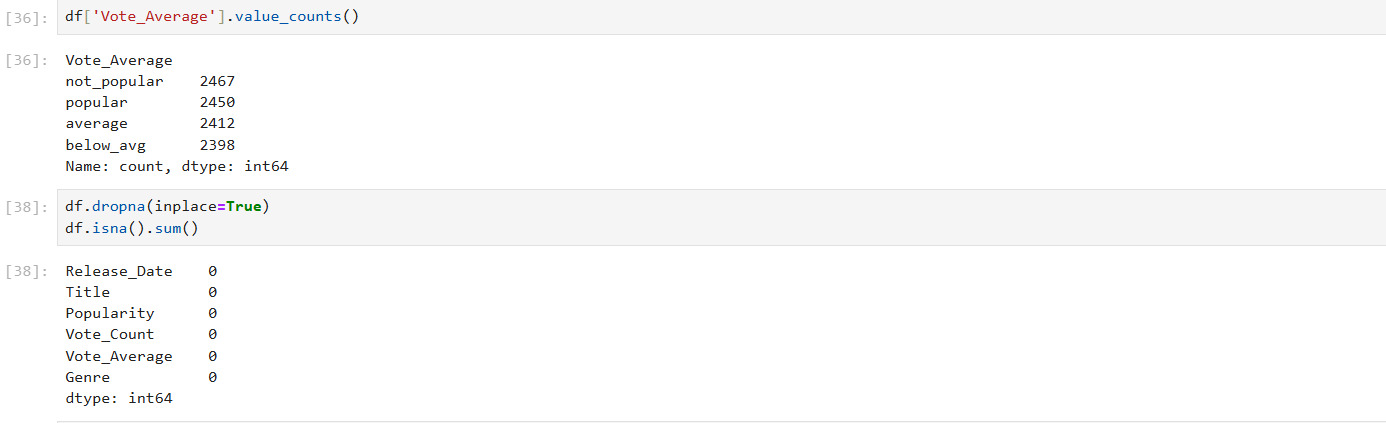


**Dropping the columns**

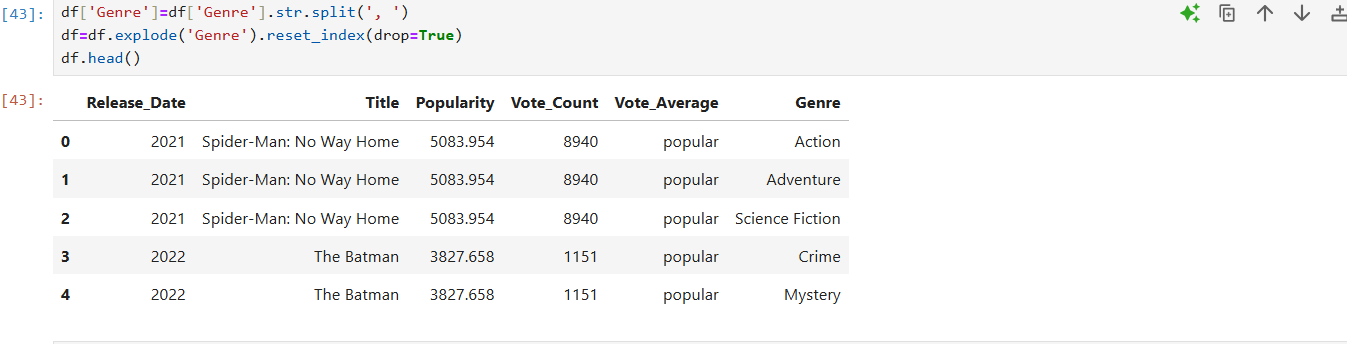


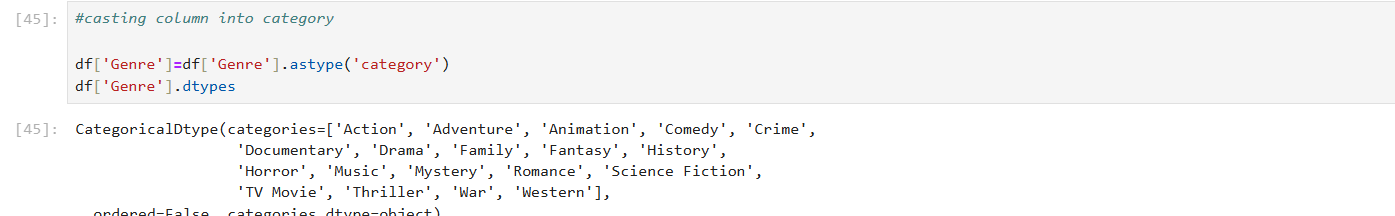
**Categorizing the Vote\_Average column**





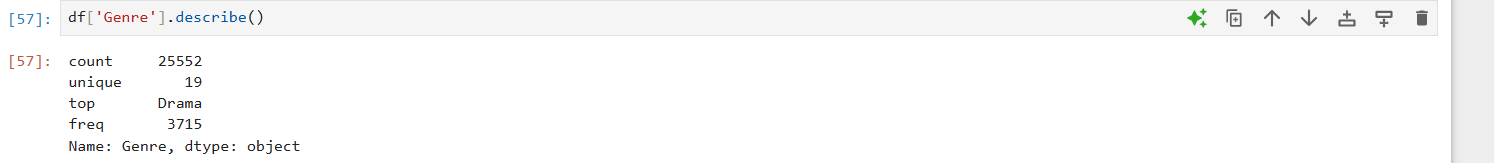
**we'd split genres into a list and then explode our dataframe to have only one genre per row fro each movie**

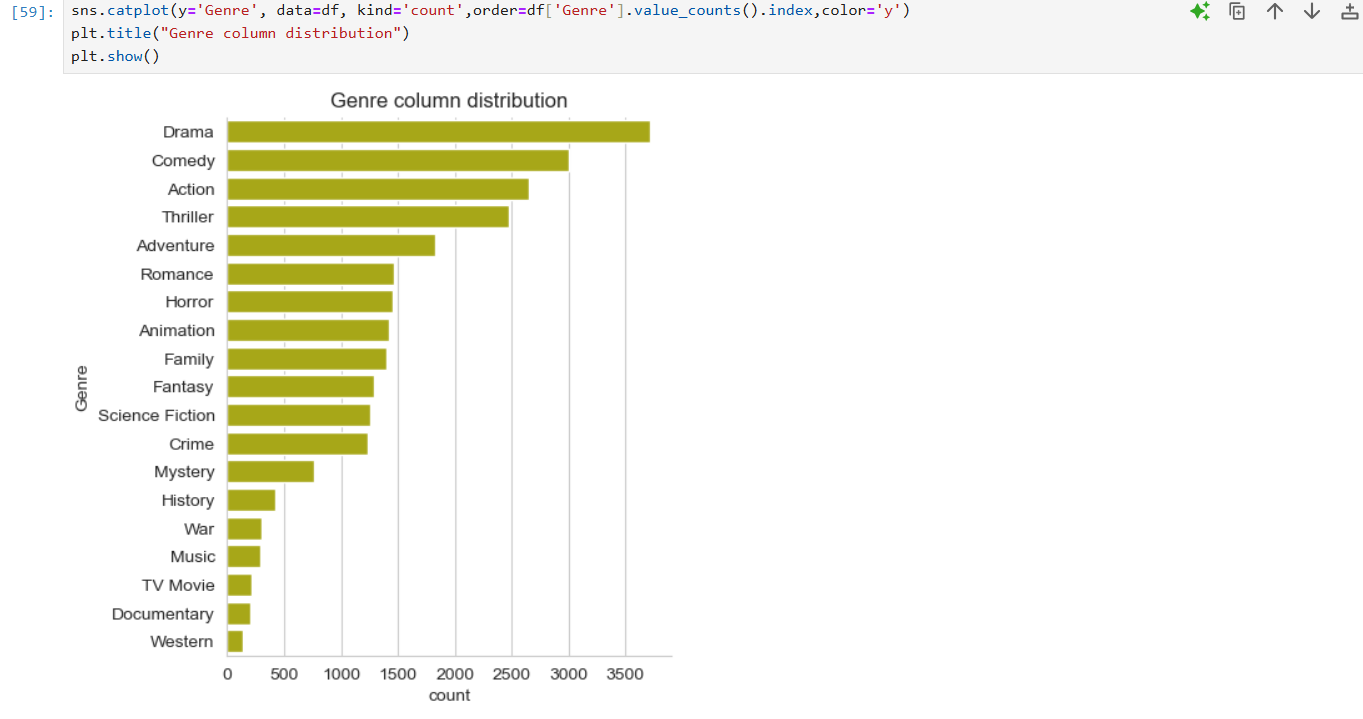




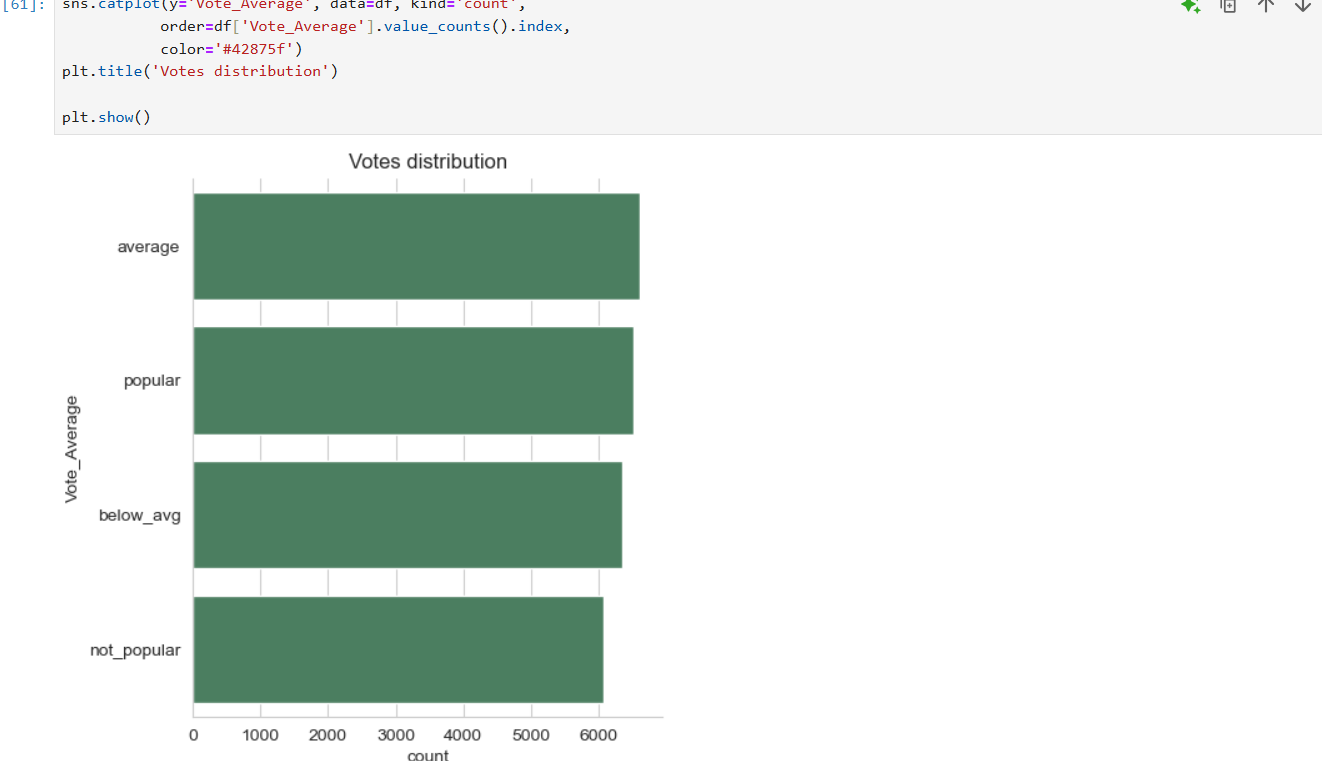
**Data Visualization:**

**what is the most frequent genre of movies released on Netflix?**

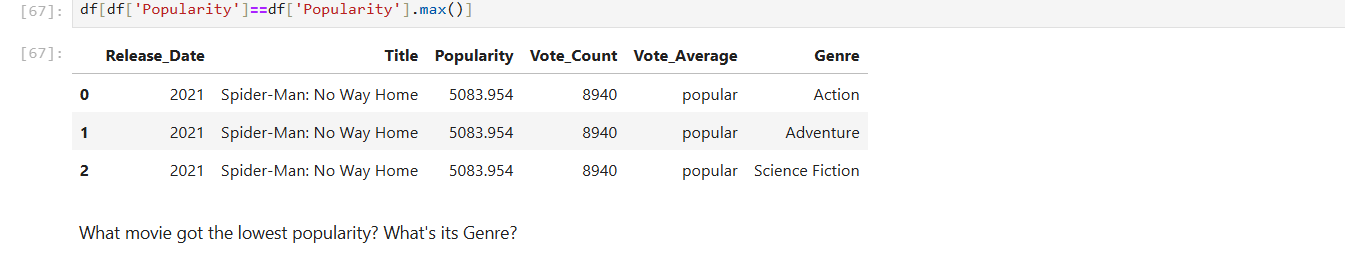




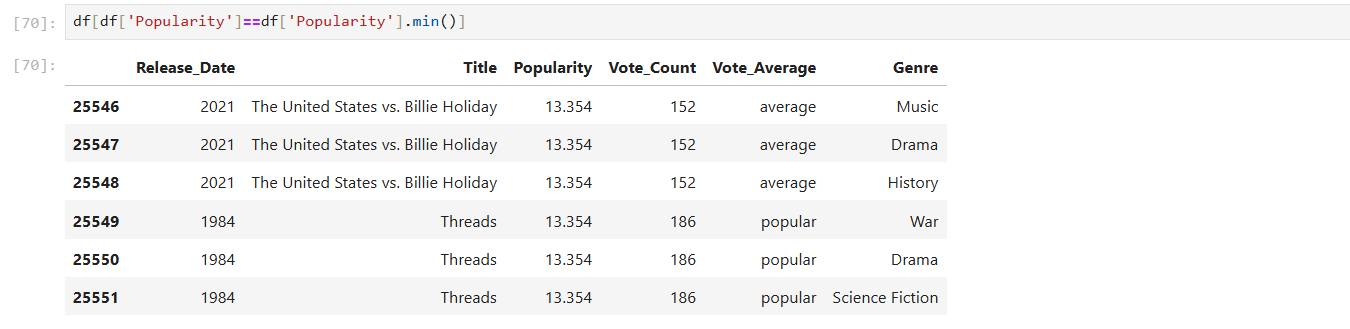
**Which has highest votes in avg column?**



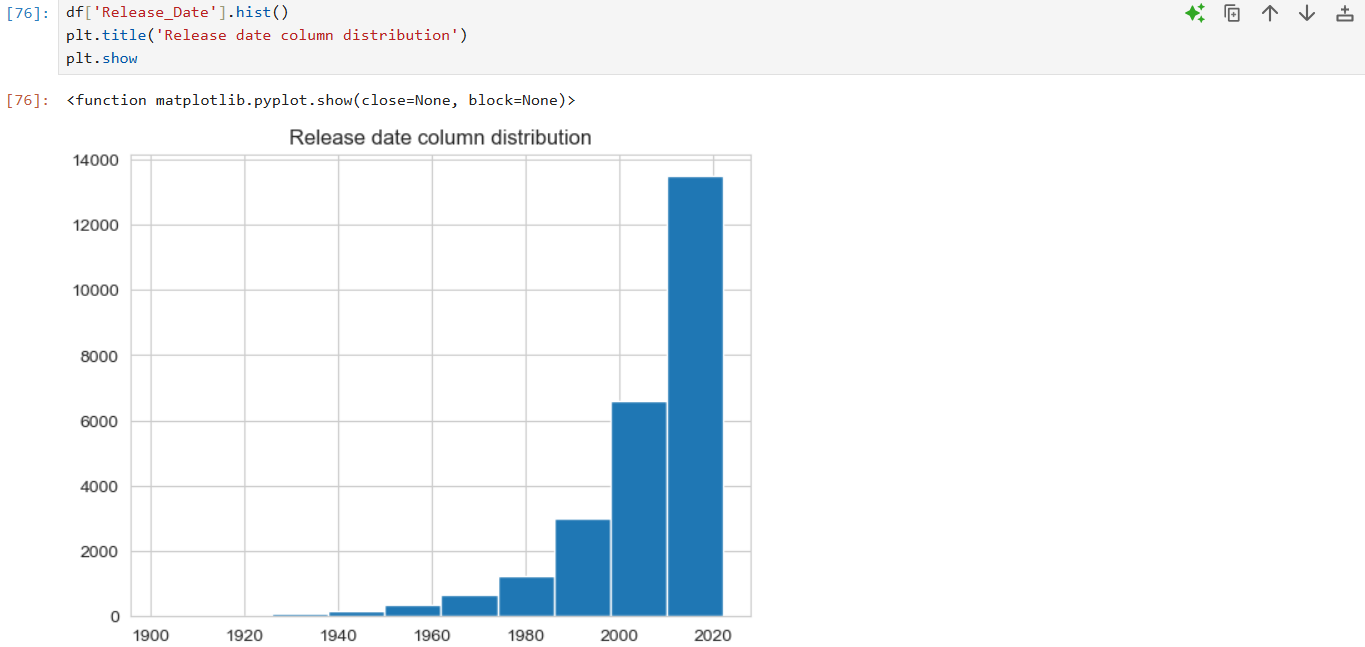
**What movie got the highest popularity? whats its genre?**



**What movie got the lowest popularity? What's its Genre?**



**Which year has the most filmmed movie?**



**Conclusion:**

**Q1: What is the most frequent genre in the dataset?**

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

**Q2: What genres has highest votes?**

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 .

**Q3: What movie got the highest popularity? what's its genre?**

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

**Q4: What movie got the lowest popularity? what's its genre?**

The united states, ‘thread' has the lowest popularity in our dataset and it has genres of music, drama, 'war', 'sci-fi' and history.

**Q5: Which year has the most filmmed movies?**

Year 2020 has the highest filmming rate in our dataset.