TMDV Movie Data Analysis

NAME : SAKSHI

MENTOR : JAYA PANDEY

Project Introduction :

Movies that cost over $100 Million can still fail, why so? Movie lovers might have different interests.

A production company wants to analyze a movie dataset to identify what kinds of movies perform well in cinemas, which genres they belong to, and so on. It will help the company predict if a movie will be a commercial success, if the movie will be highly rated, etc.

Description of the dataset

budget: The budget of the movie.

genres: A list of genres associated with the movie.

homepage: The URL of the movie's official homepage.

id: The unique identifier of the movie.

keywords: A list of keywords associated with the movie.

original_language: The original language of the movie.

original_title: The original title of the movie.

overview: A brief overview or synopsis of the movie.

popularity: The popularity score of the movie.

production_companies: A list of production companies involved in the movie.

production_countries: A list of production countries associated with the movie.

release_date: The release date of the movie.

revenue: The revenue generated by the movie.

runtime: The duration of the movie in minutes.

spoken_languages: A list of spoken languages in the movie.

status: The status of the movie (e.g., Released, In Production).

tagline: The tagline or slogan of the movie.

title: The title of the movie.

vote_average: The average rating of the movie.

vote_count: The number of votes received by the movie.

Brief On Implementation Of The Capstone :

The project consisted of several tasks:

1. Load the movie dataset in the python notebook. Display the numbers of rows and columns in the dataset. Display the titles and genres of the first 50 movies.

2. Identify the columns that have null values and perform the null value treatment. (choose the imputation method based on the type of data in the columns of interest).

3. Display the movie categories that have a budget greater than $220,000.

4. Display the movie categories where the revenue is greater than $961,000,000.

5. In the dataset, there are some movies for which the budget and revenue columns have the value 0, which mean unknown values. Remove the rows with value 0 from both the budget and revenue columns.

6. List the top 10 movies with the highest revenues and the top 10 movies with the least budget.

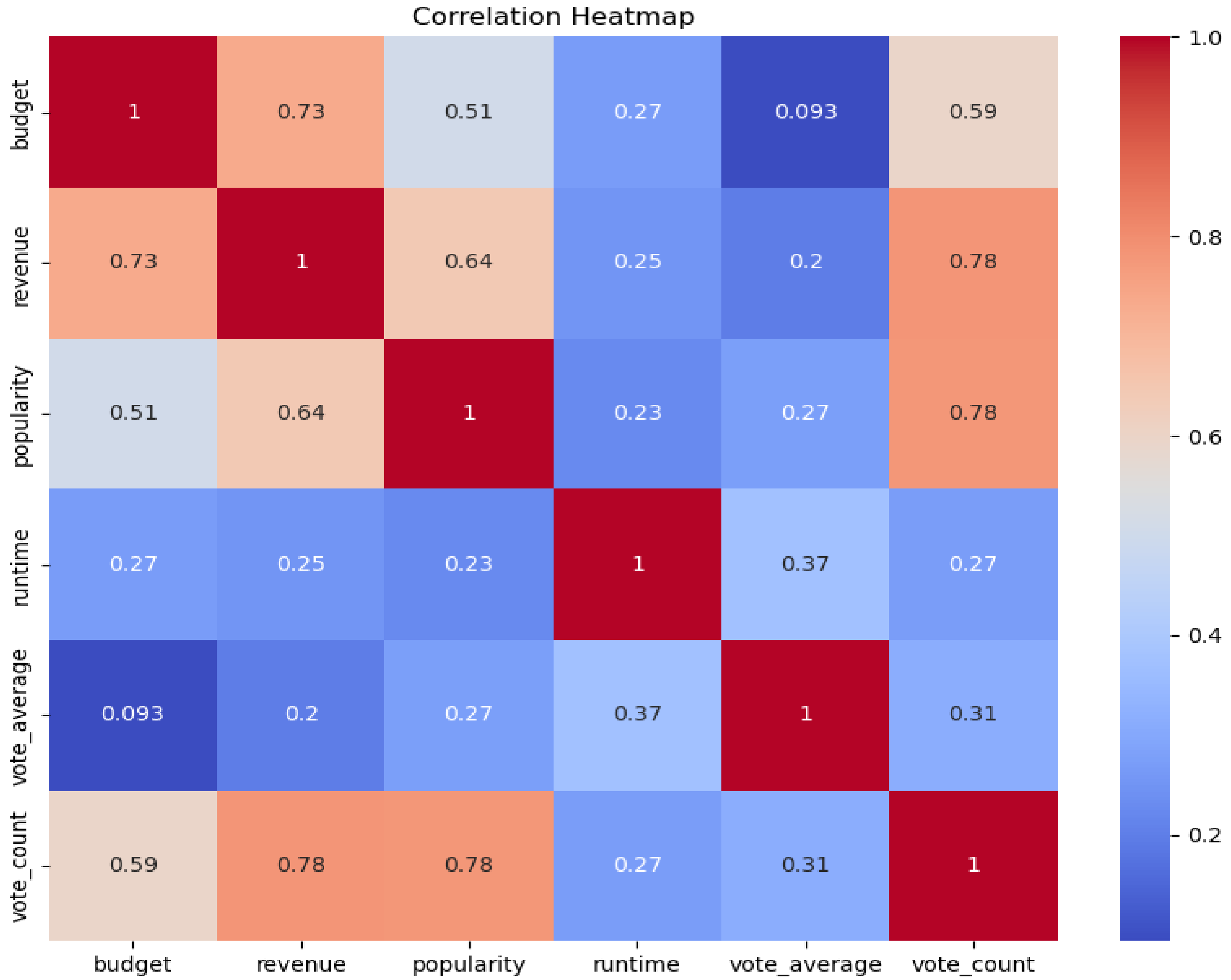
7. How are popularities of movies related with the movie budgets? Are they correlated or totally uncorrelated with each other? Write the interpretation of your analysis.

8. Identify and display the names of all production companies along with the number of times they appear in the dataset.

9. Display the names of 25 production companies based on the number of movies they have produced in descending order of the number of movies produced.

10. Sort the data in descending order based on revenue and filter the top 500 movies. Find the measures of tendency for the following columns using the filtered data: budget, revenue, runtime. Perform outlier analysis for the above three columns using box plots.

11. Identify and display the names of the movies along with their run times for those movies that have above average runtime, using the data from the previous tasks.



Heat Map

# Interpretation :

* 'budget' and 'revenue' have a strong positive correlation of 0.73.

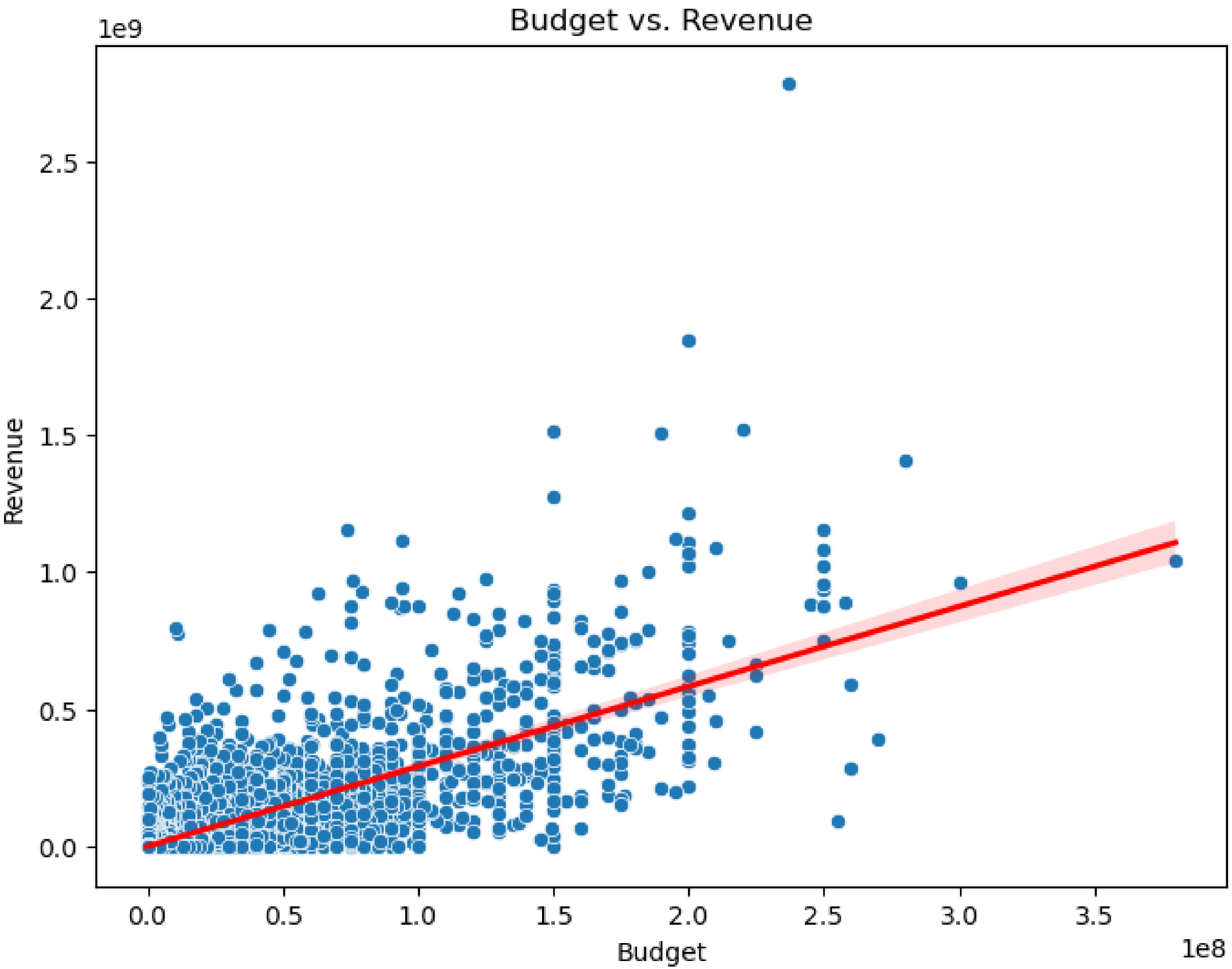
* 'revenue' and 'popularity' also have a strong positive correlation of 0.64.

* 'popularity' and 'vote count' exhibit a strong positive correlation of 0.78.

* The correlation between 'budget' and 'popularity' is 0.51,

* There is a relatively weak positive correlation between 'runtime' and 'vote average' (0.37) and between 'popularity' and 'vote average' (0.27).

* The correlation between 'vote average' and 'vote count' is 0.31.

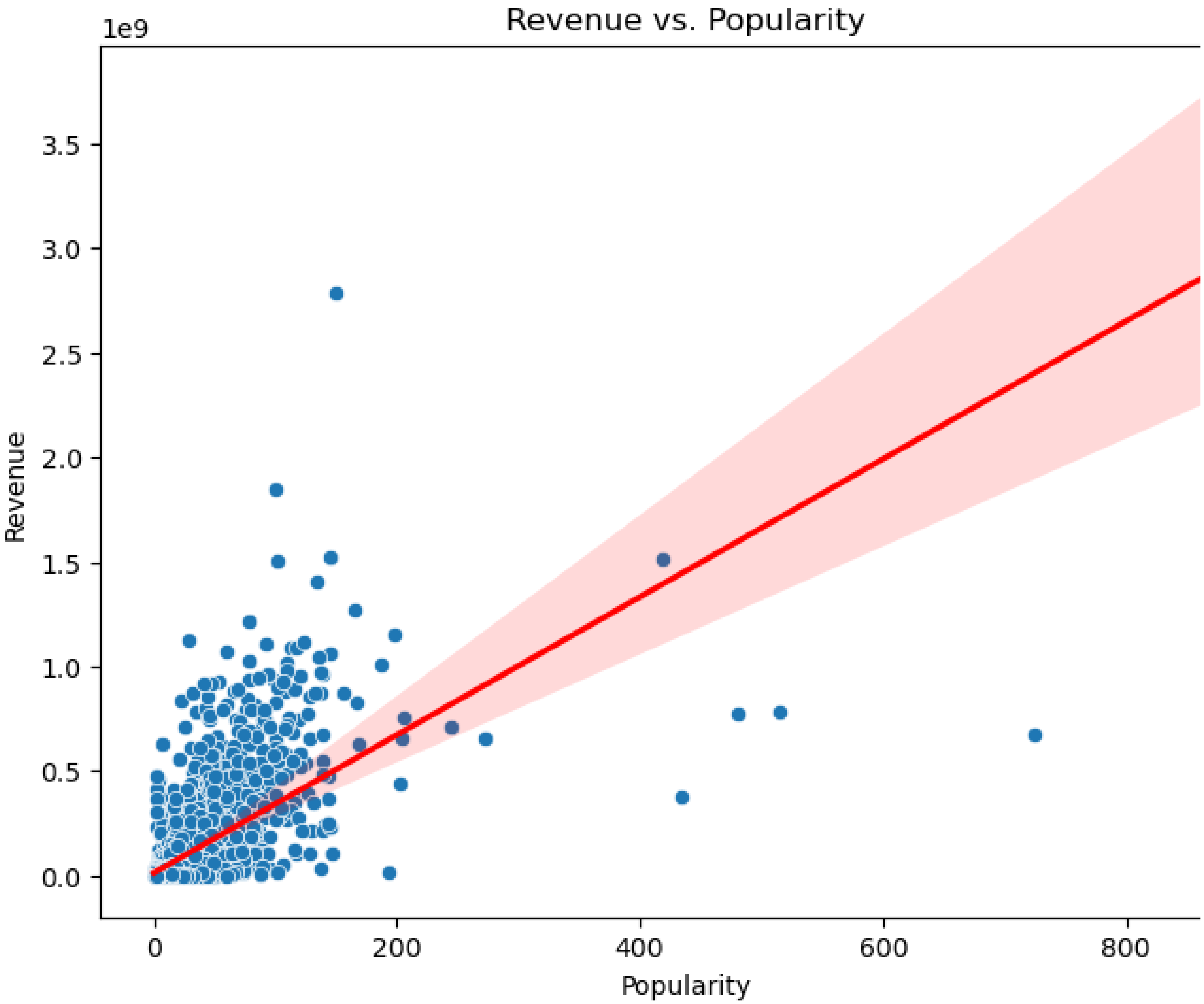


Interpretation :

- The trend line in the graph has a positive slope, that means, movies with higher budgets tend to generate higher revenues.
- The data points on the scatter plot are spread out, suggesting some variability , Actual revenue generated by a movie can vary even within a similar budget range.
- That outliers represent movies that achieved unexpectedly high or low revenues compared to their budgets.

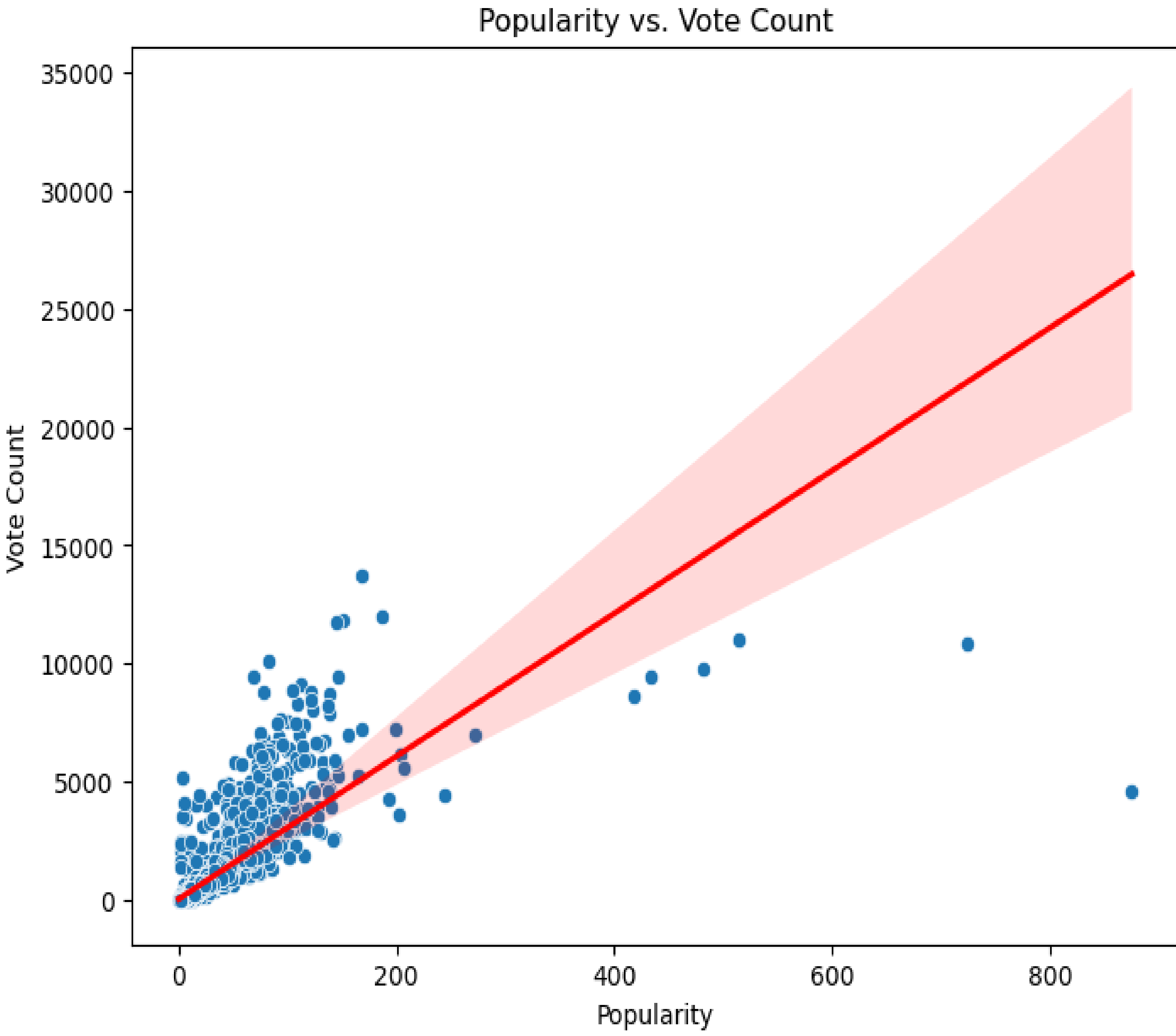
Interpretation :

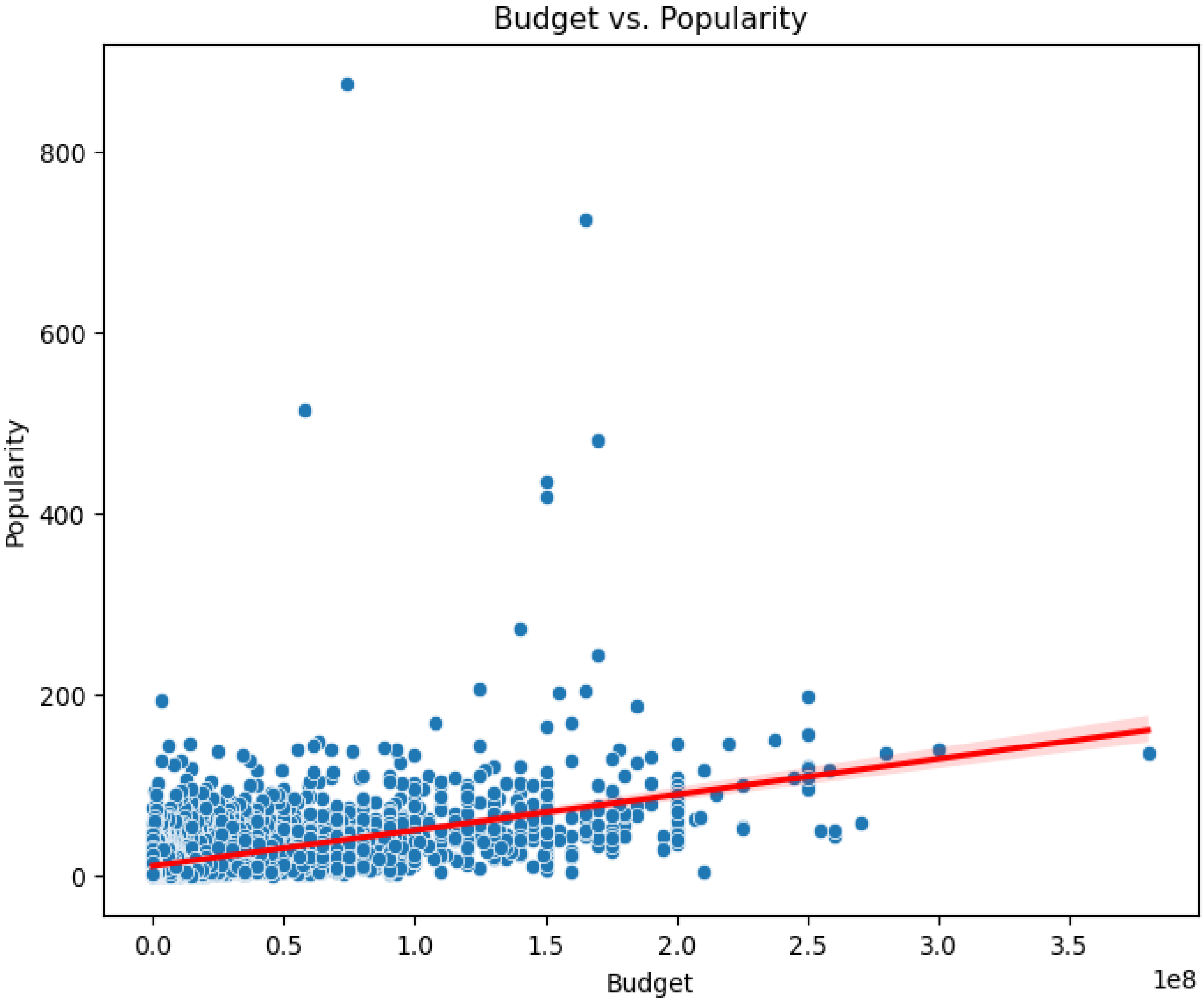
- The trend line has a positive slope ,this means that movies with higher popularity tend to generate higher revenues.
- The outliers represent movies that achieved unusually high or low revenues compared to their popularity levels.



Interpretation :

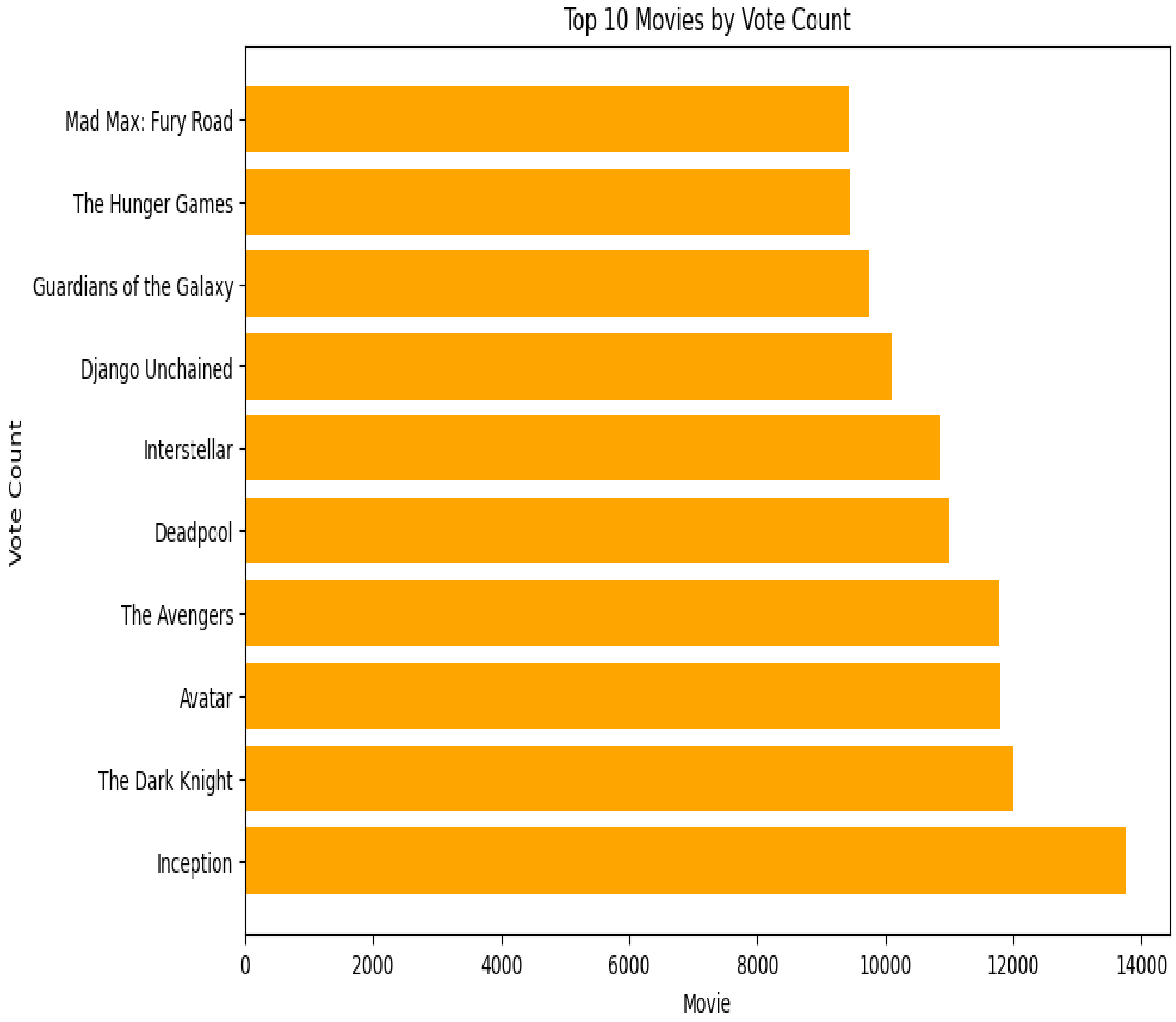
- The trend line has a positive slope, it indicates a positive relationship between popularity and vote count. This means that movies with higher popularity tend to receive a higher number of votes.
- The outliers represent movies that achieved unexpectedly high or low vote counts compared to their popularity levels.

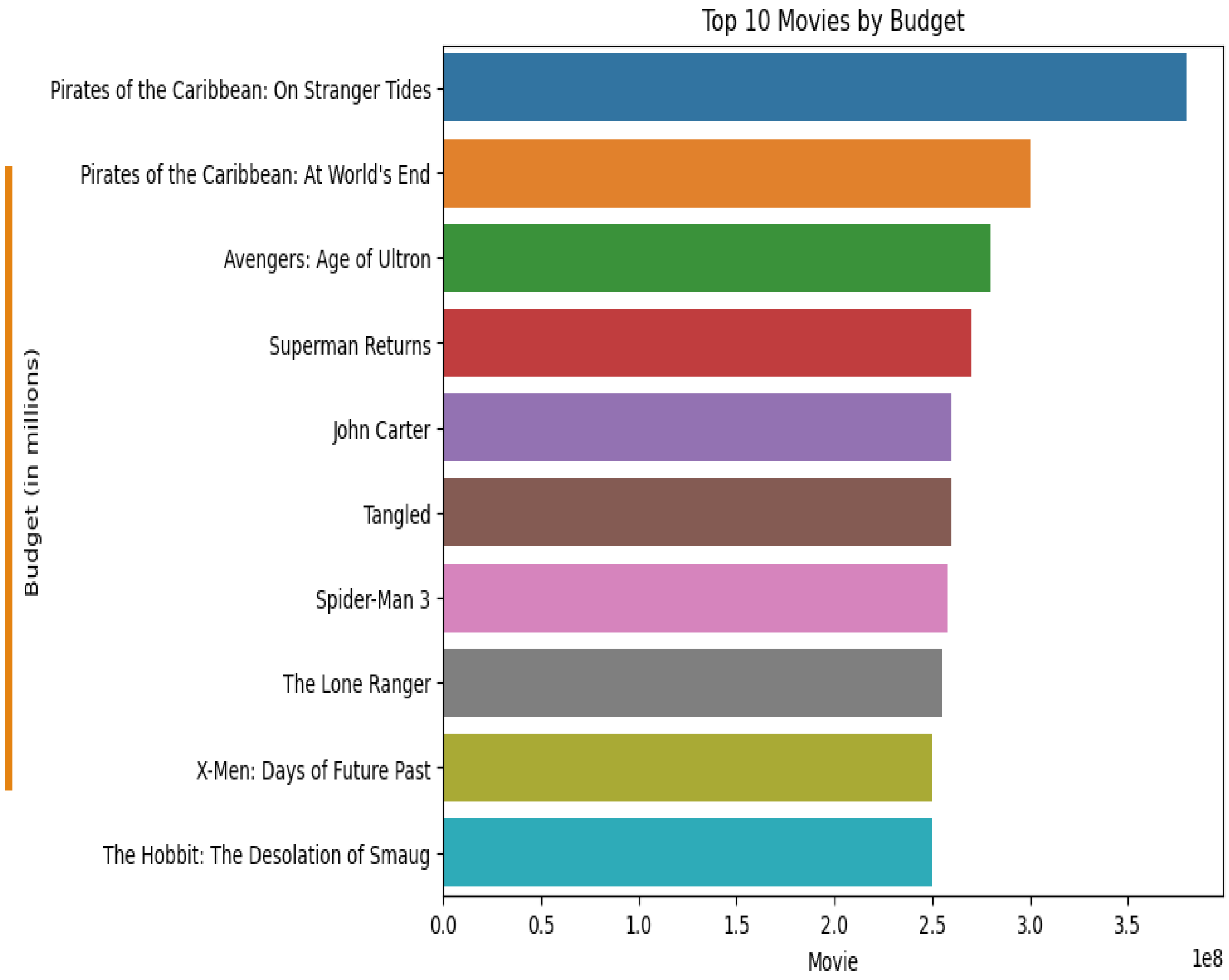




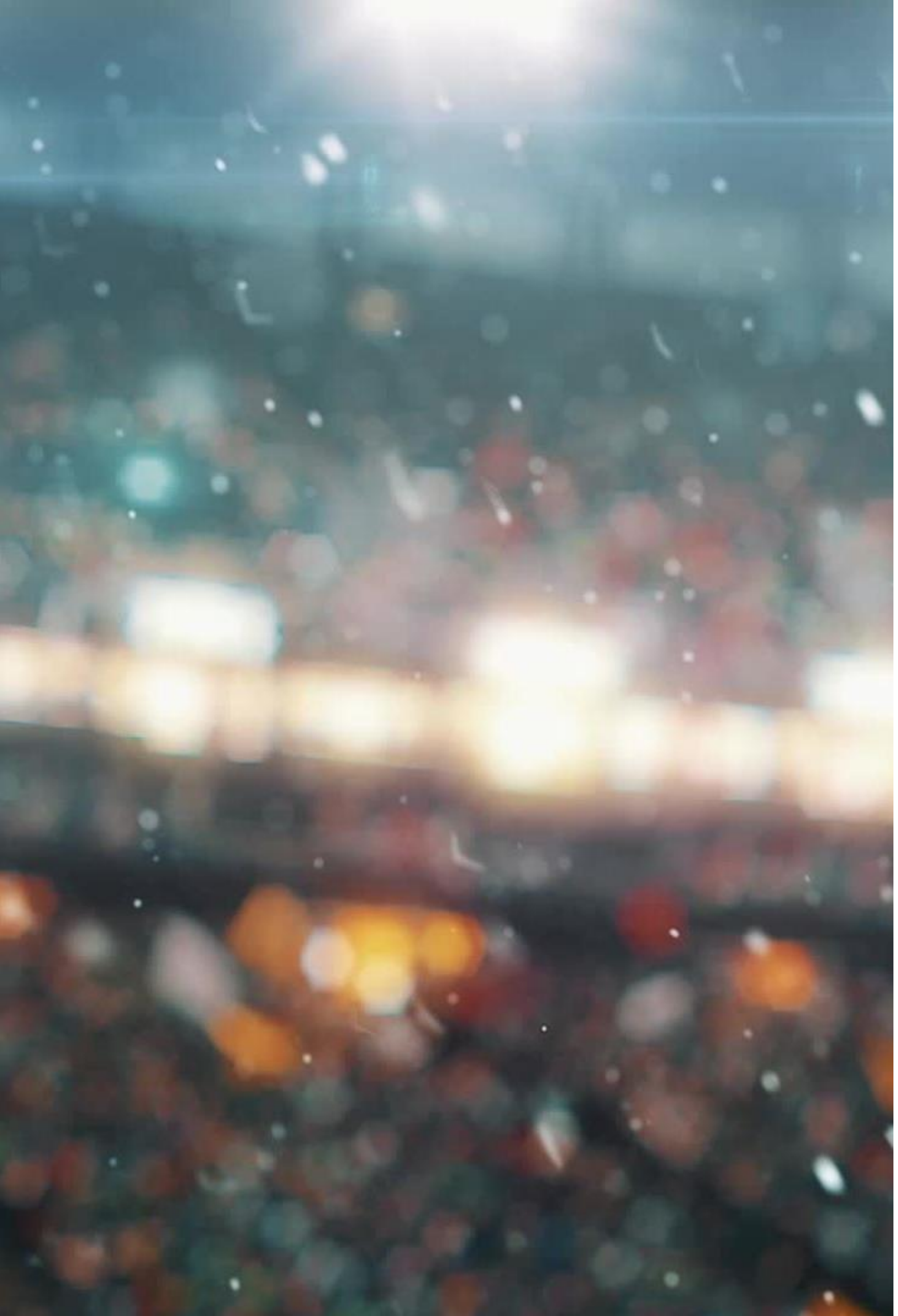
Interpretation :

- The scatter plot shows a general positive trend.
- This suggests that movies with larger budgets have a higher likelihood of attracting attention and interest from audiences.
- Some data points deviate from the overall trend , these outliers may indicate unique cases where factors other than budget contribute significantly to a movie's popularity.





Interpretation : These movies must be stocked up in the inventory to increase the sales.



Conclusion :

- Movies with higher budgets, popularity, vote count, tend to generate higher revenues. So, the rental store must stock up movies with higher budgets, popularity and vote count to generate higher revenue.
- ***Insights :***
- In terms of budget : Pirates Of The Caribbean, Avengers, Superman Returns, movies must be stocked up.
- In terms of vote count : Inception, The Dark Knight, Avatar, movies must be stocked up.