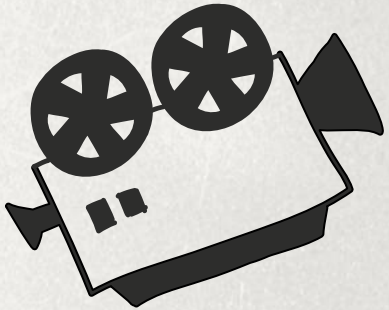
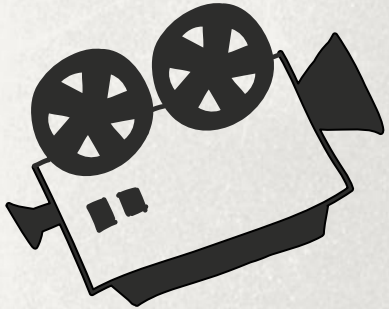
IMDb 2024 Movies Data Visualization Project



Jayna Clark

Data Set and Variables



IMDb Movies and TV Shows 2024

- I accessed this data set from Kaggle (linked above in title)
- Data set includes 501 movies and TV shows from the year 2024 sourced from IMDb
- There are 16 variables in the dataset

Variables (1)

Variable Name	Variable Type	Description
Home_Page	Character	Link to IMDb webpage for movie
Movie_Name	Character	Title of movie or TV show
Genres	Character	List of Genres associated with movie/TV
Overview	Character	Brief description of movie plot
Cast	Character	Brief list of main actors
Original_Language	Character	List of languages movie/TV was originally made in
Storyline	Character	Detailed description of plot
Production_Company	Character	Companies involved in the production

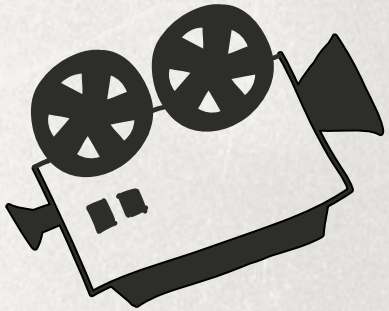
Variables (2)

Variable Name	Variable Type	Description
Release_Date	Date	Release date in format YYYY-MM-DD
Tagline	Character	Movie/TV promotional slogan
Vote_Average	Numeric	Average IMDb rating (0-10)
Vote_Count	Numeric	Number of votes on rating received (in thousands)
Budget_USD	Numeric	Budget in millions of USD
Revenue_$	Numeric	Revenue in millions of USD
Run_Time_Minutes	Numeric	Duration in minutes
Release_Country	Character	Country where title was first released

Data Cleaning Process

- When read into R some variables not in proper data format
  - I ensured dates were date variables, numbers were numeric, and character variables were strings or what was appropriate to the variable
- I used a “clean_numeric” function to remove extra characters from numeric variables
- In cases where a data value was 0 and that was impossible to be the value- I changed these values to NA
    - Example: runtime of 0 is not possible

Questions of Interest



Questions of Interest

How are the movie ratings distributed?

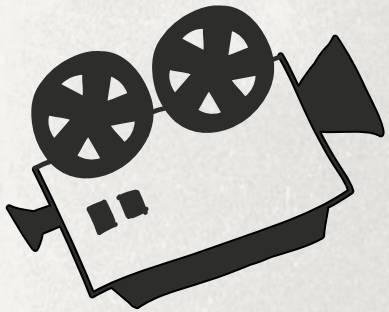
Is there a correlation between the numeric variables (budget, revenue, average votes, average rating)?

How are the movies release date distributed by month?

Do variables follow any pattern over the time of the dataset?

How are the runtimes distributed?

Visualizations and Data Summaries



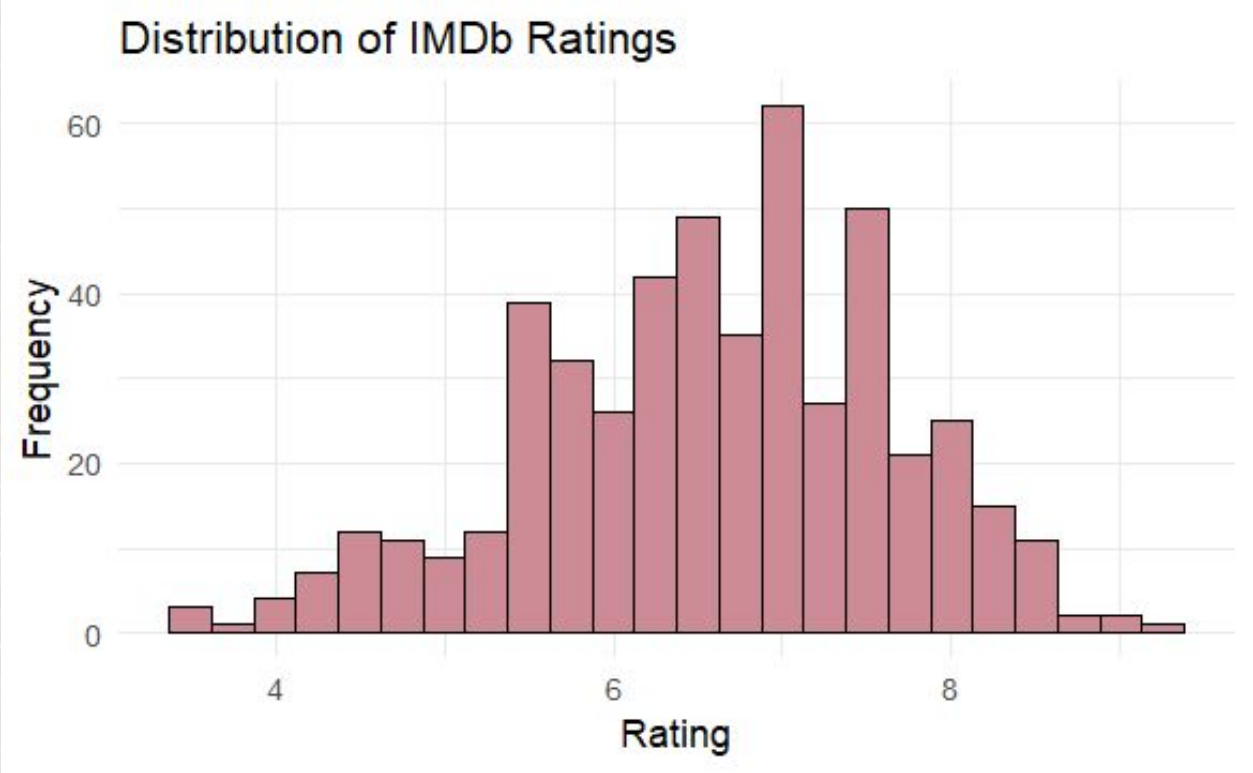
How are the movie ratings distributed?

Mean: 6.58

Median: 6.70

SD: 1.07

Distribution is slightly skewed to the left.



5-number summary and boxplot of movie ratings

Min: 3.4

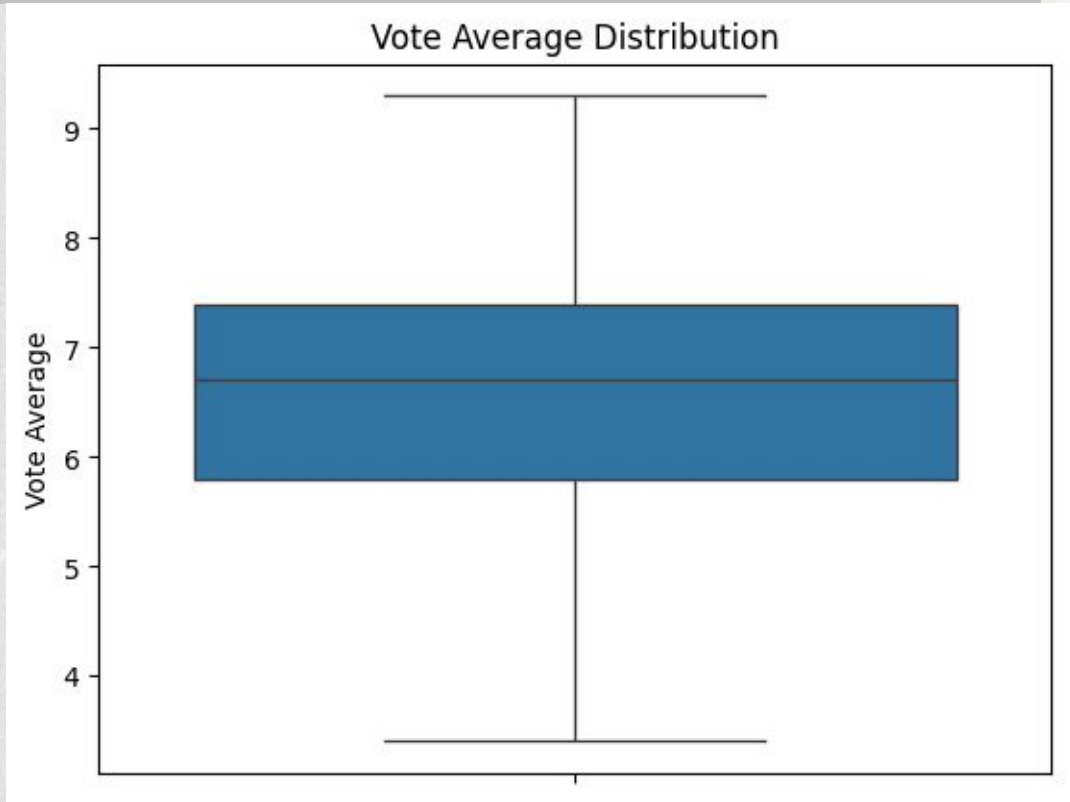
Q1: 5.8

Median: 6.7

Q3: 7.4

Max: 9.3

There is quite a large spread (range = 5.9) of movie ratings (out of 10)



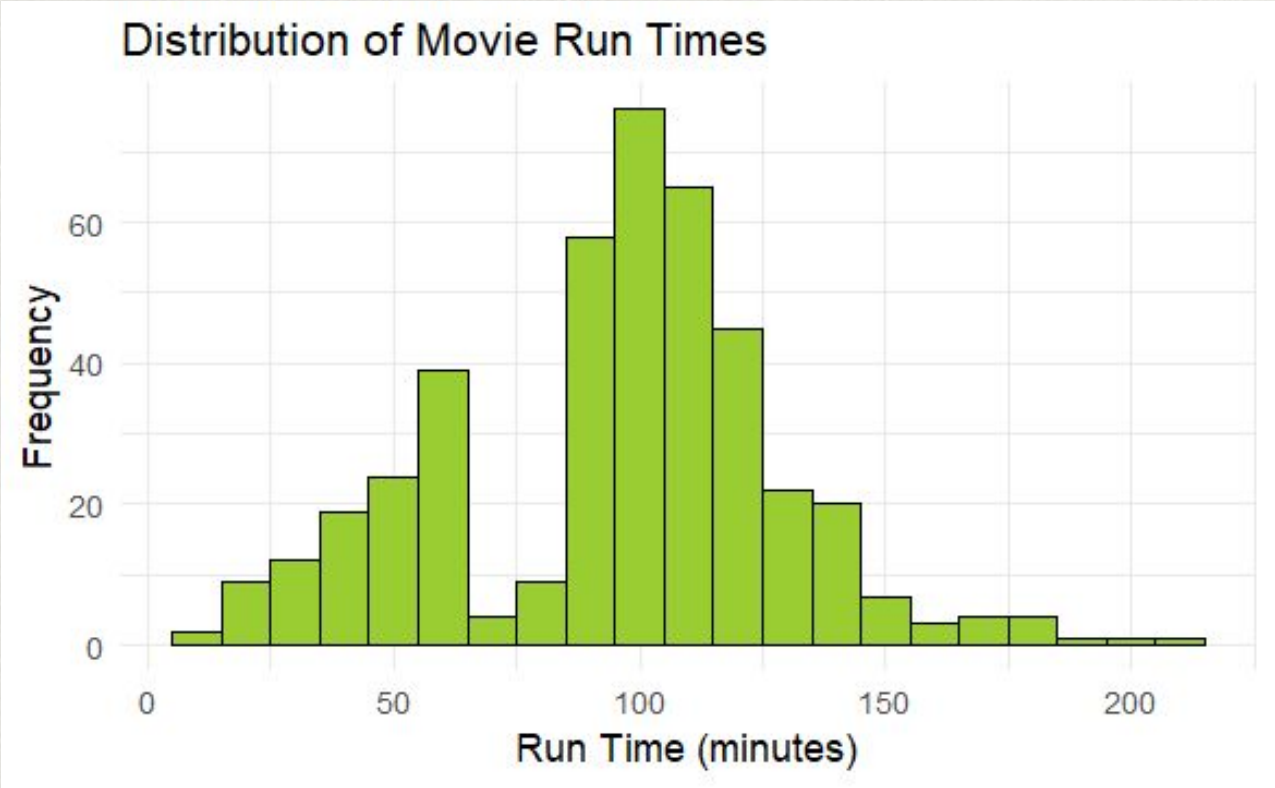
How are movie run times distributed?

Mean: 95.71

Median: 100

SD: 34.14

Bimodal distribution peaks about 60 and 100



Boxplot of Movie Run Times

Min: 13.0

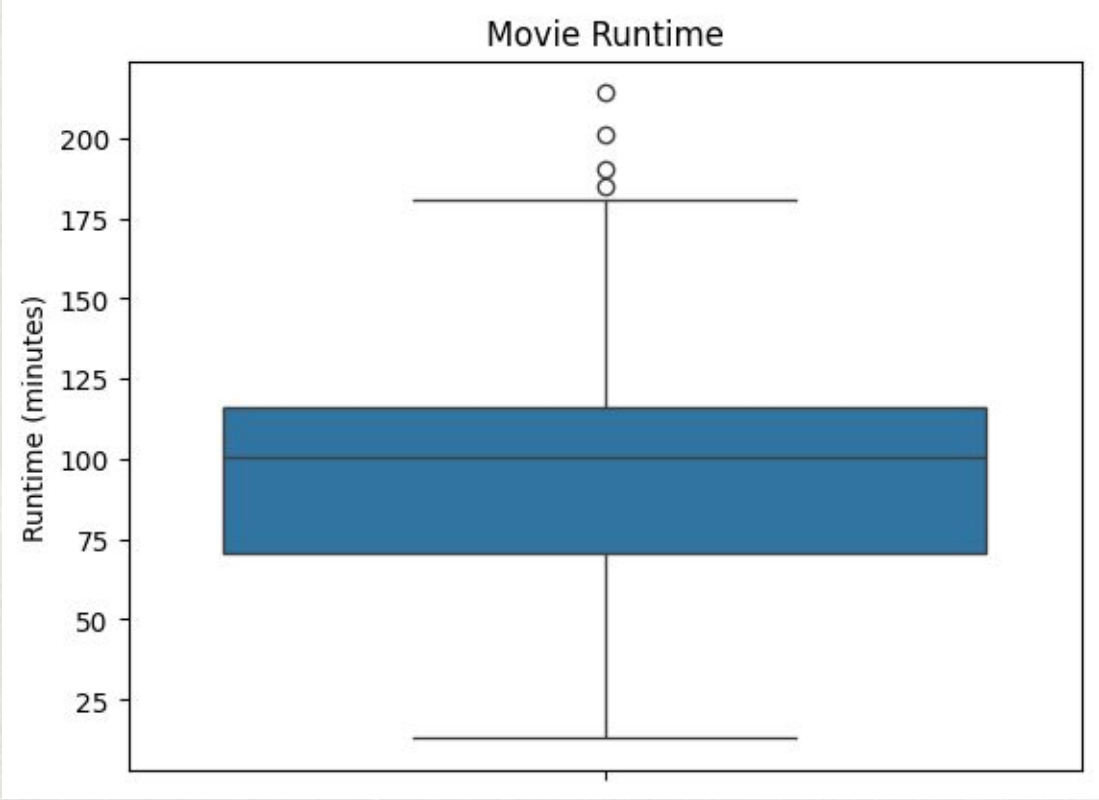
Q1: 70.75

Median: 100.0

Q3: 116.0

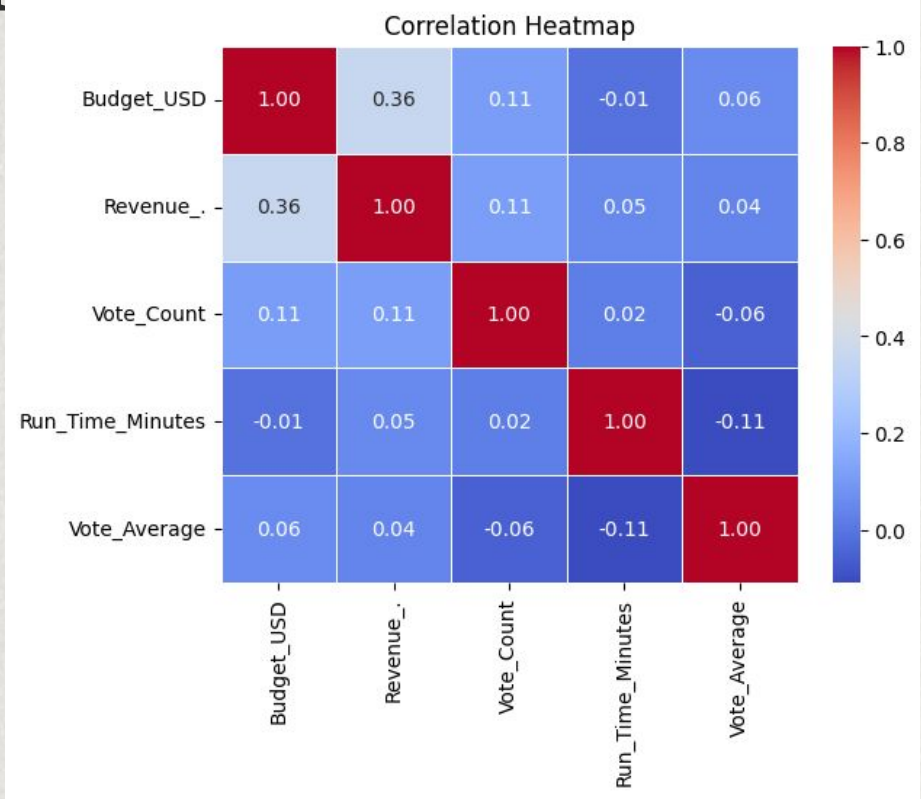
Max: 214.0

There is a very large range in runtimes (range = 201 minutes)

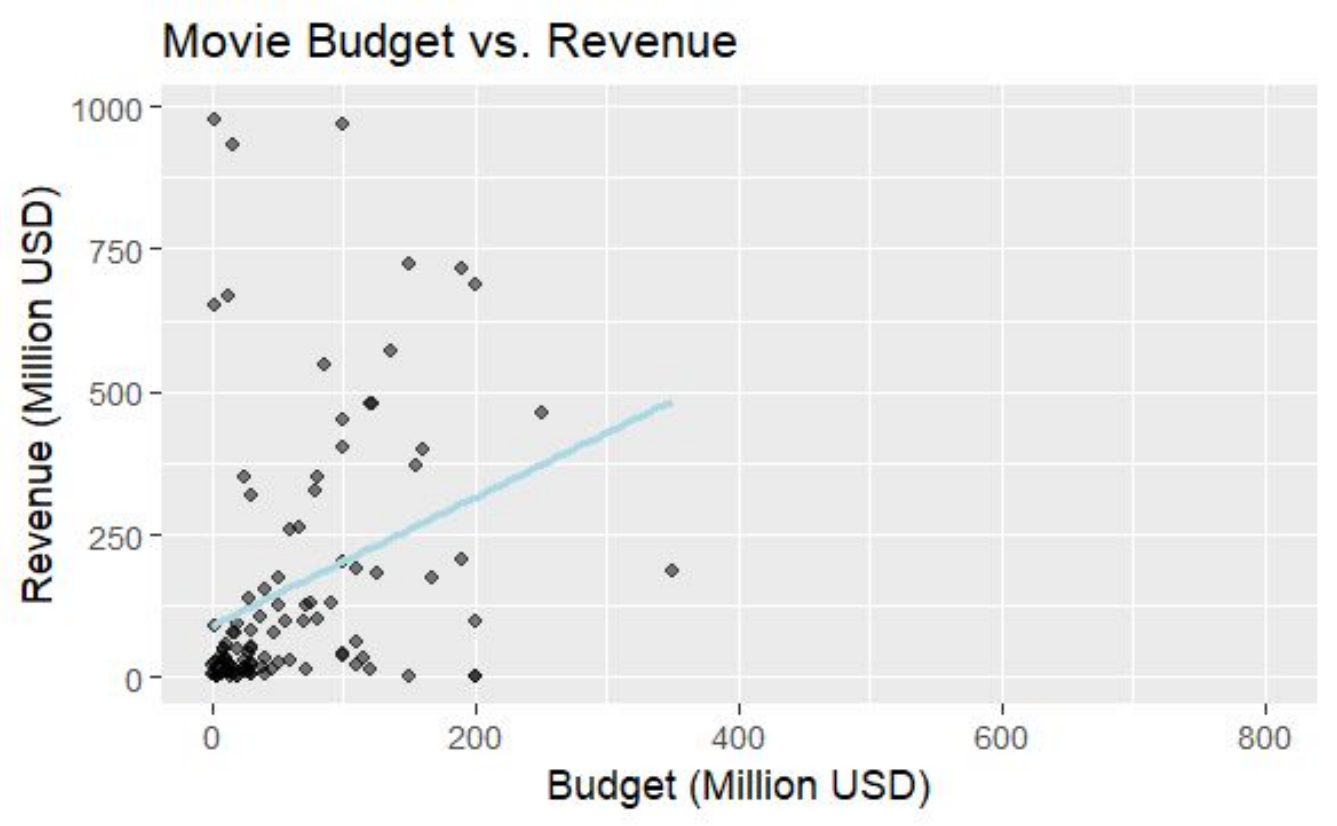


Is there correlation between numeric variables?

- Largest correlation: budget and revenue (r=0.36)
- All of the following have a correlation of r=|0.11|
  - Budget and vote count
  - Revenue and vote count
  - Runtime and vote count (-0.11)
- Vote average (rating) and vote count have a correlation of



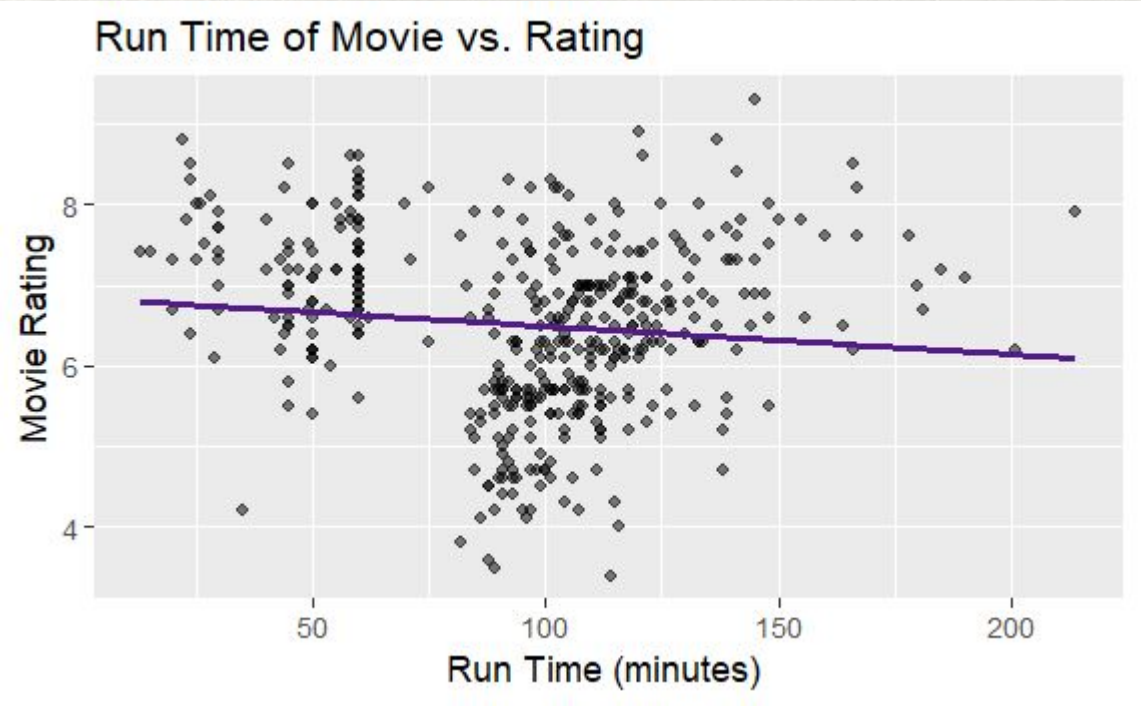
Budget and Revenue



r = 0.36

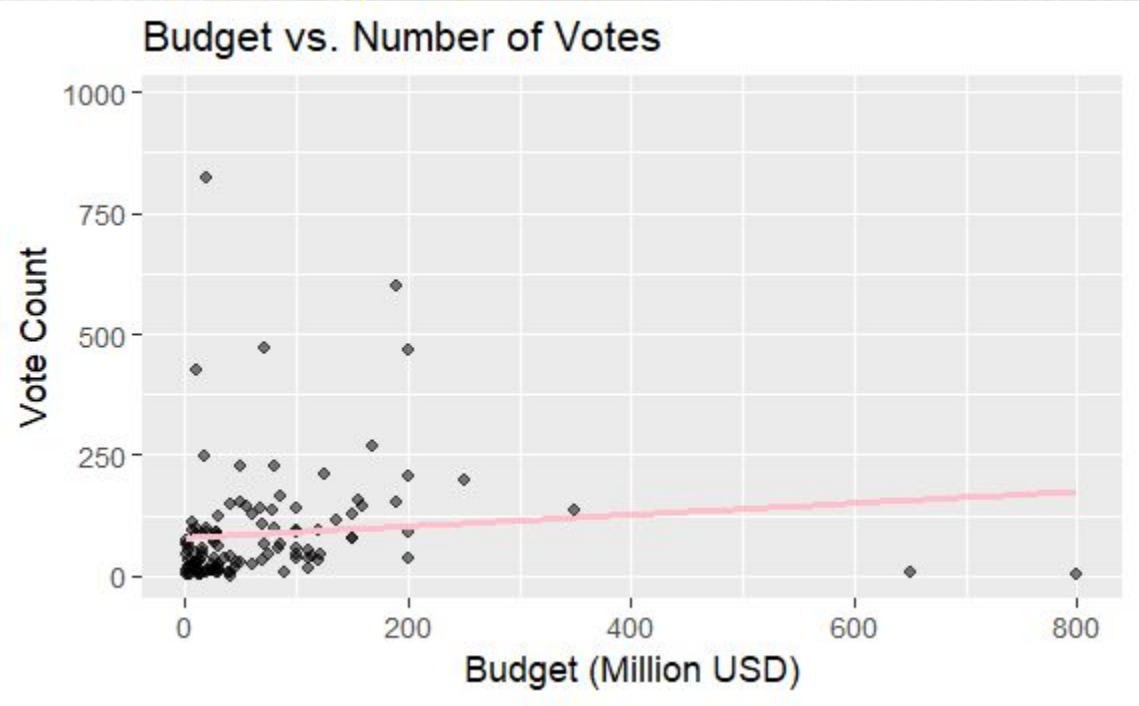
Strongest linear relationship between variables

Movie Run Time and Rating



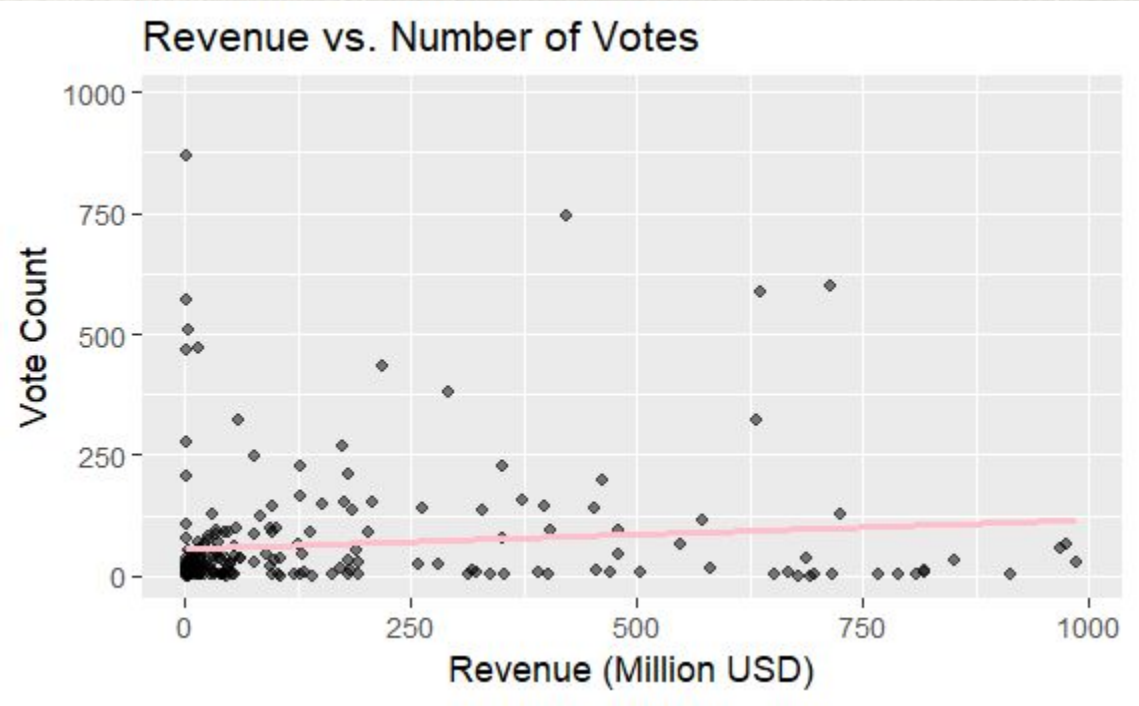
r = -0.11

Budget and Vote Count



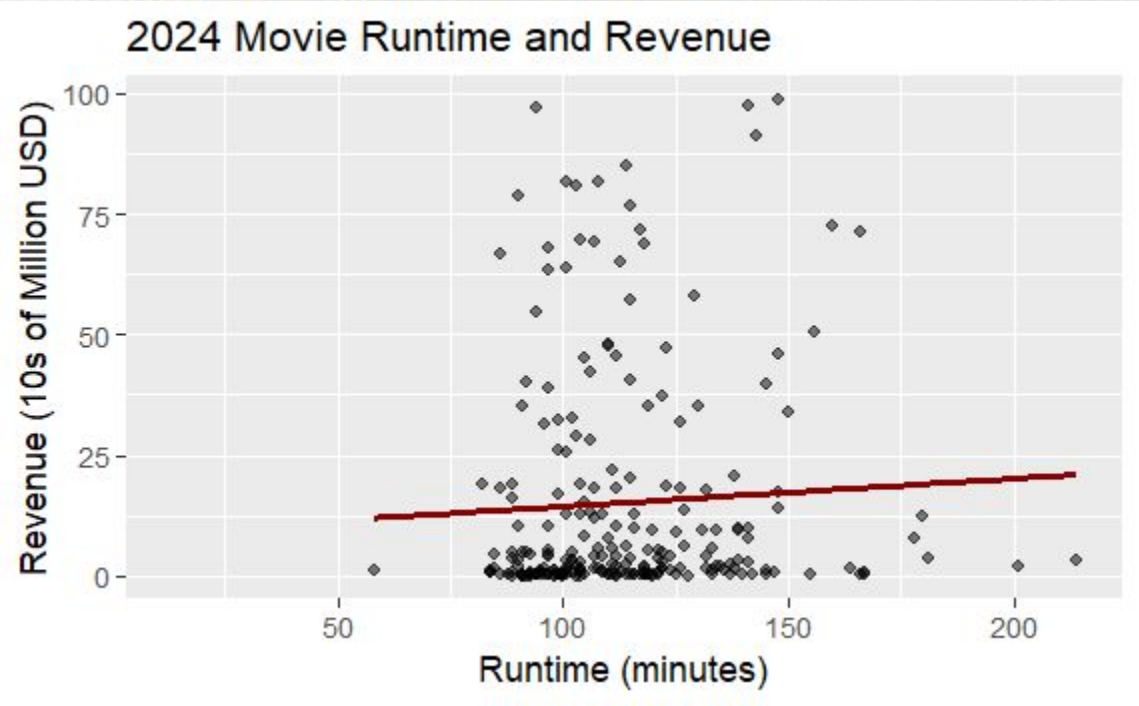
r = 0.11

Revenue and Vote Count



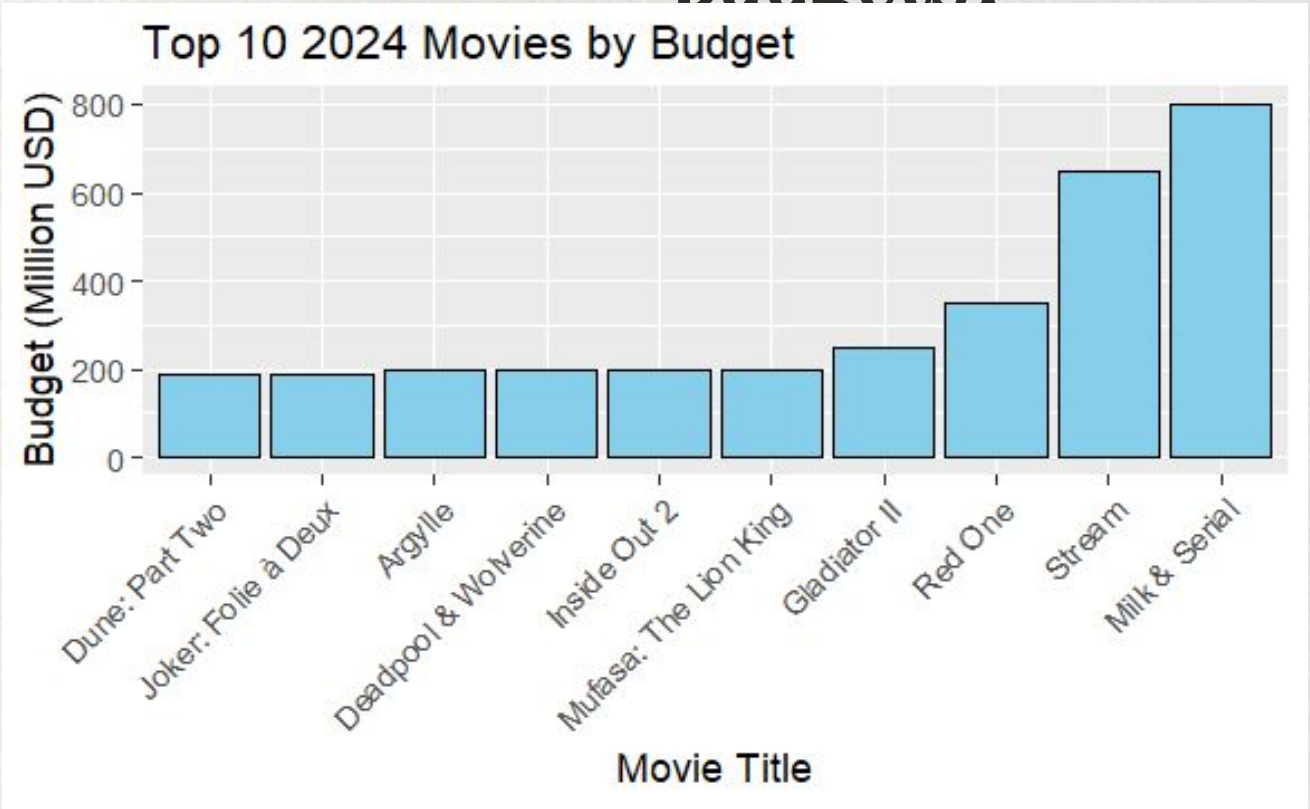
r = 0.11

Movie Run Time and Revenue



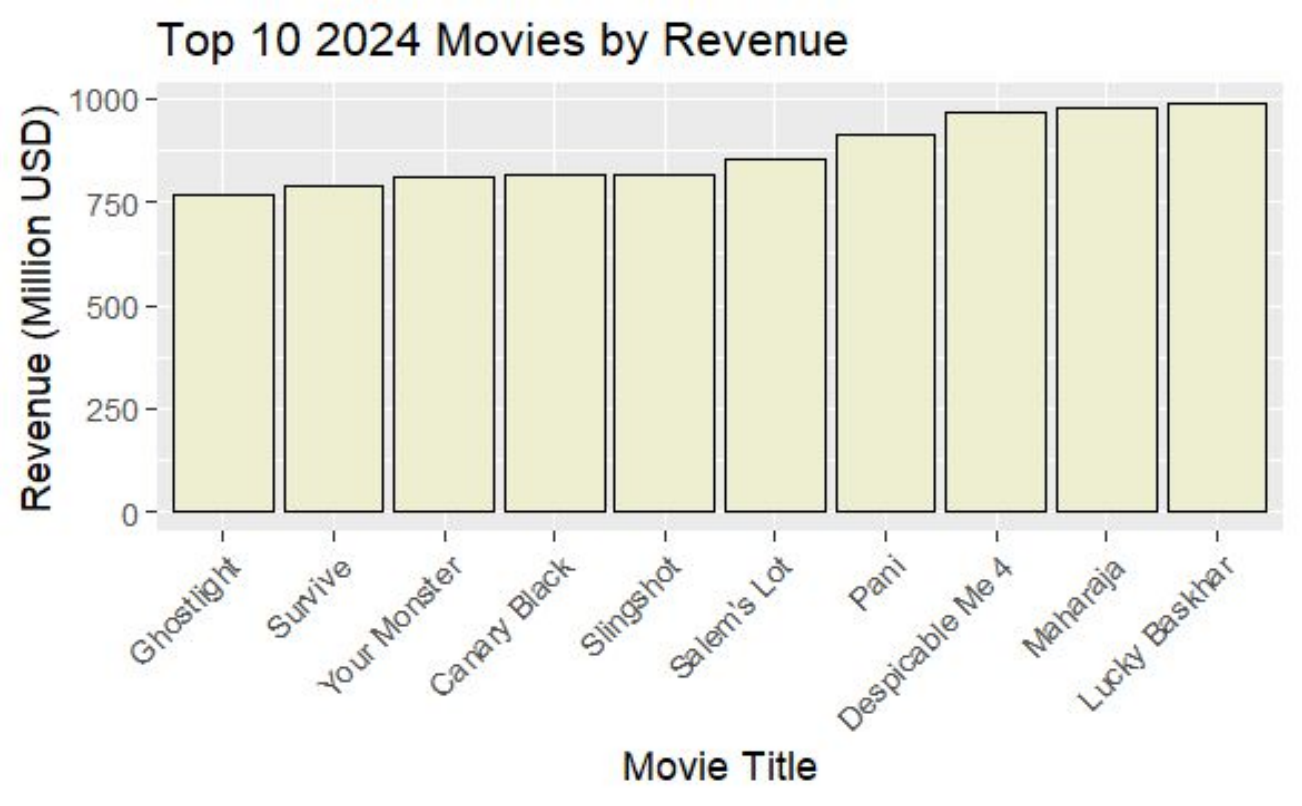
r = 0.054

What are the 2024 Movies with Highest Budgets?



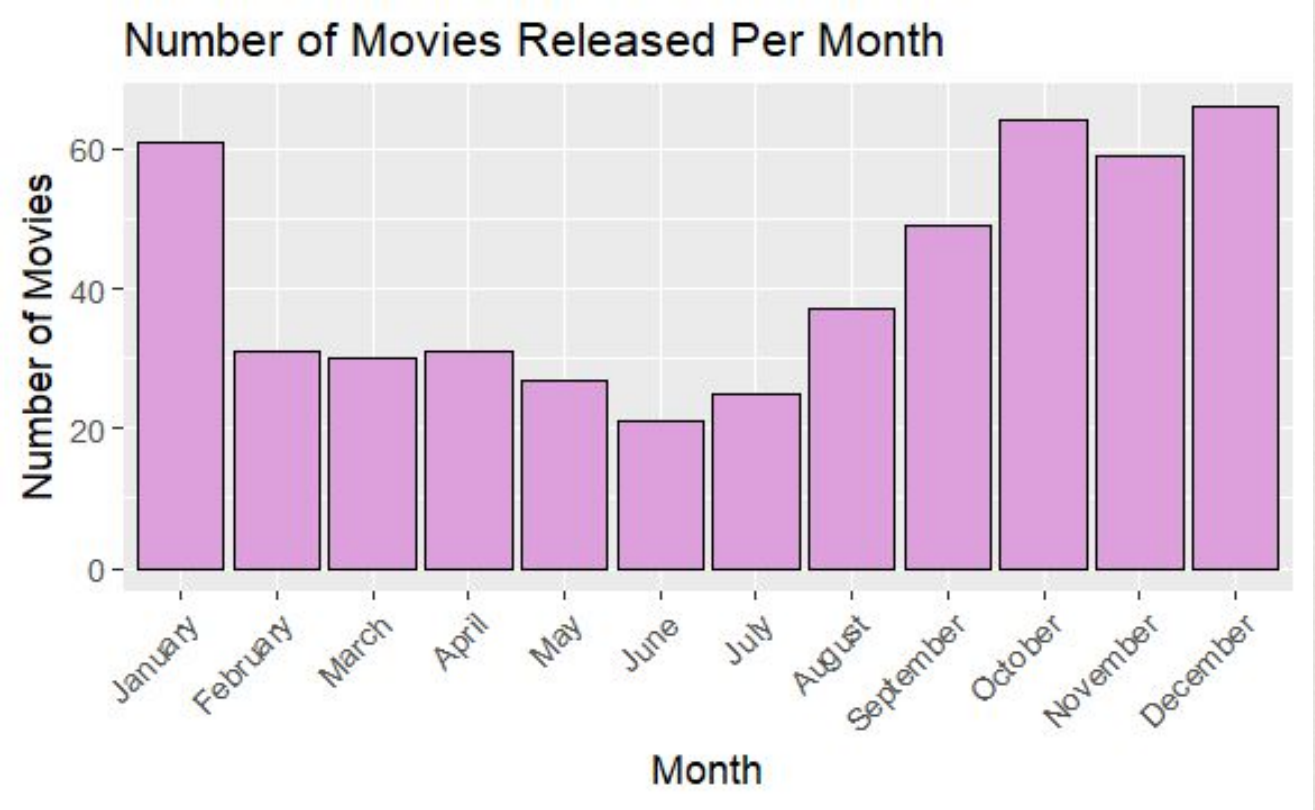
Red One, Stream, and Milk & Serial had the largest budgets

What are the 2024 Movies with Highest Revenues?



Despicable Me 4, Maharaja, and Lucky Basknar had the 3 highest revenues

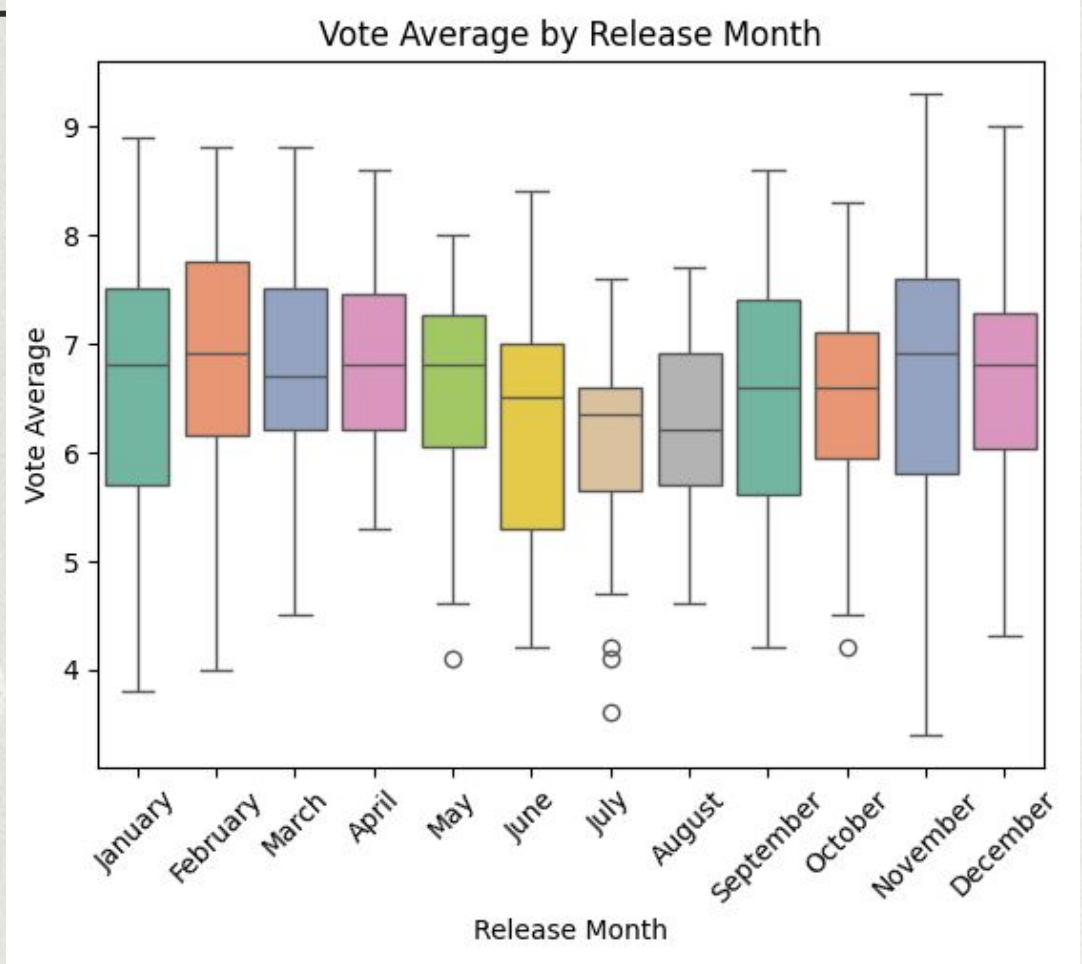
Which months had the largest number of movies released?



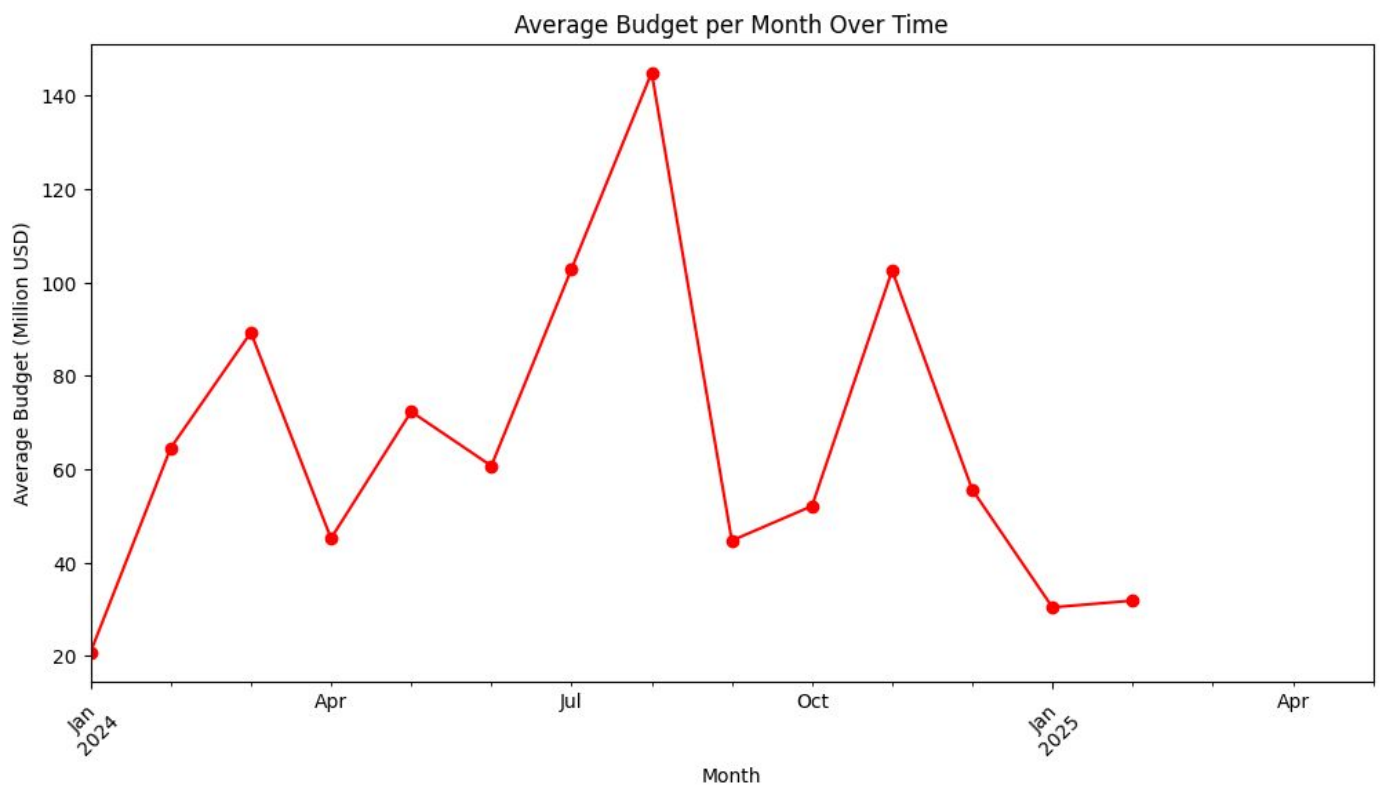
September through January had the most movies released (Fall and Winter)

Do ratings have any pattern/correlation with release month?

The median ratings were lower in the summer months- this is also when less movies are released as shown before

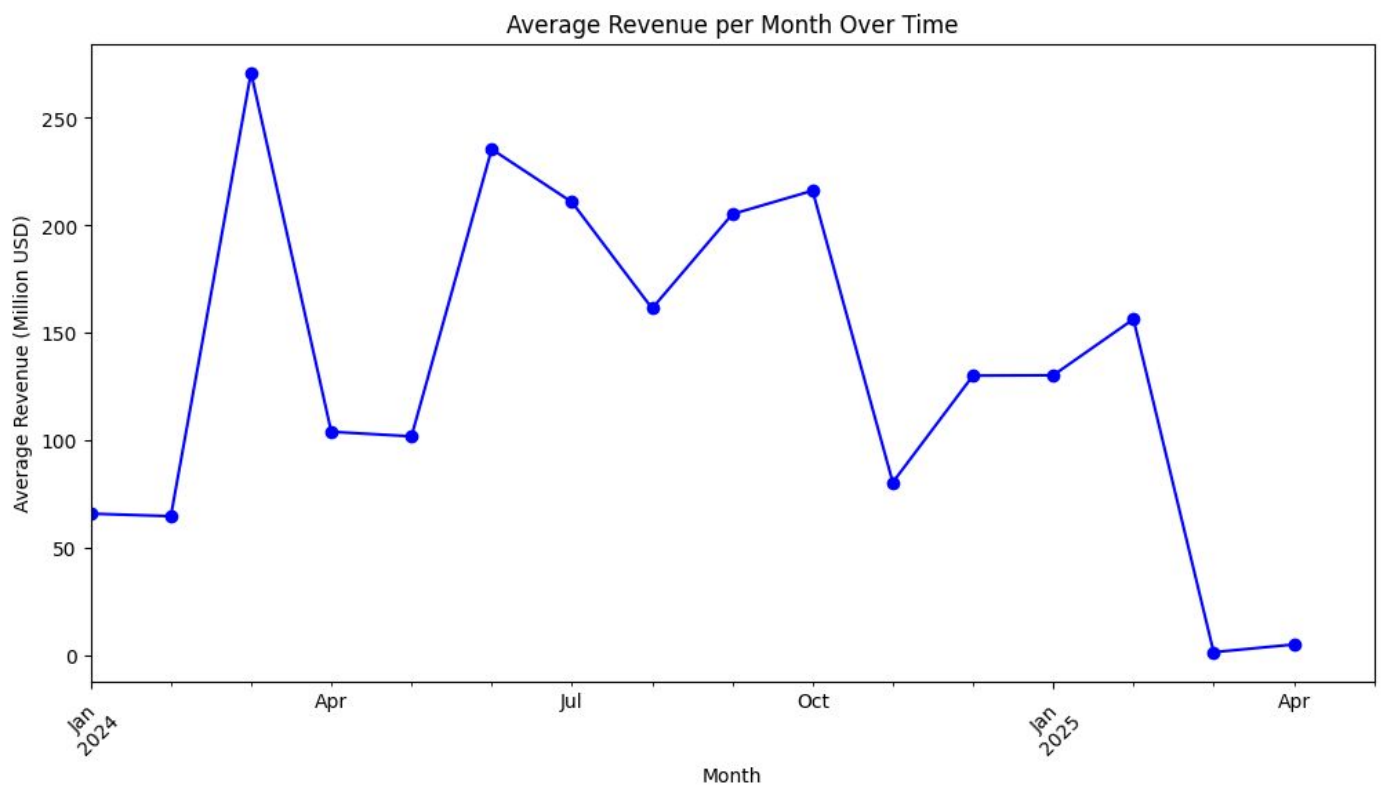


Rating by Release Month



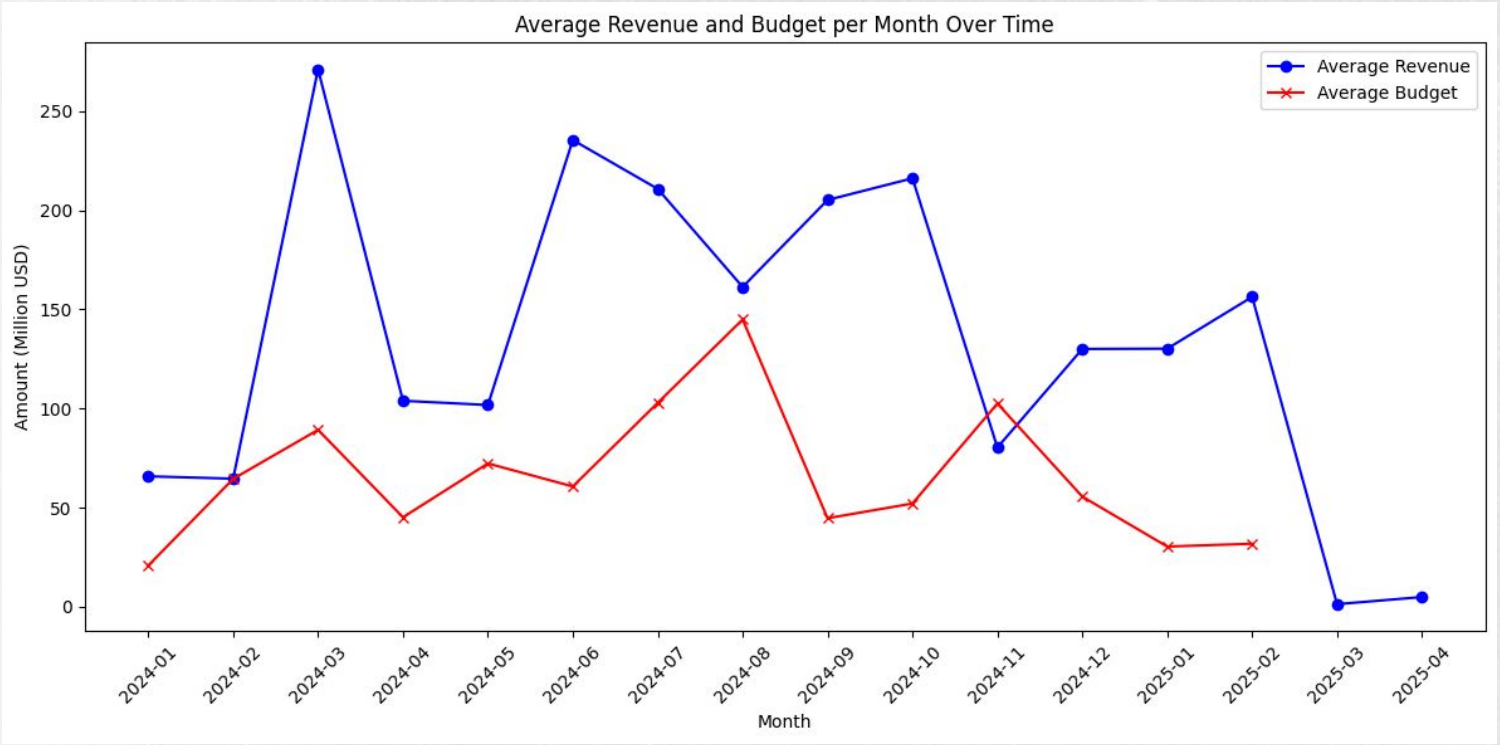
The mean budget increases and decreases without an obvious pattern- having a peak in August 2024

Revenue by Release Date



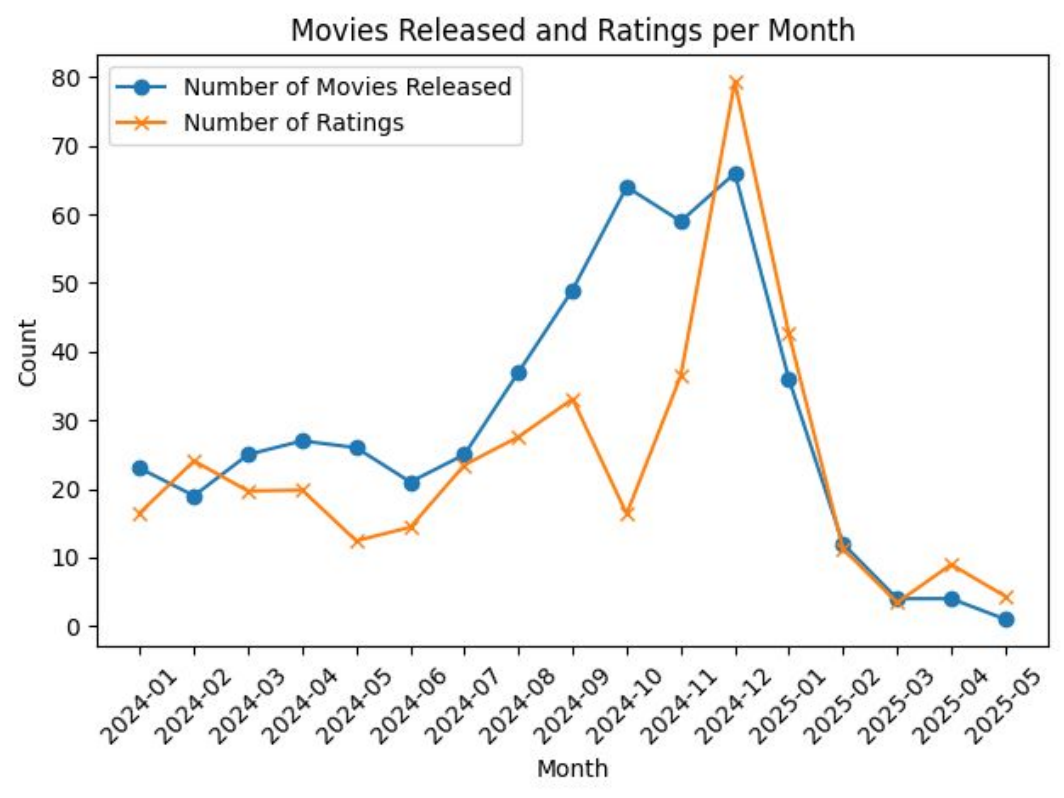
The mean revenue seems to sporadically increase and decrease- overall decreasing over time into 2025

How do revenue and budget compare over the same time period?



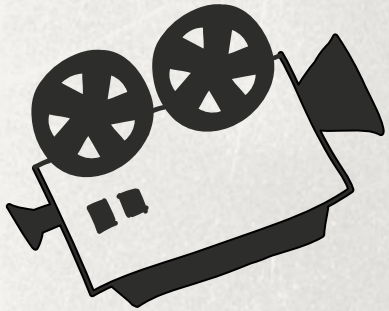
There does not seem to be a correlation between budget and revenue over time- the only connection is a decrease of both in 2025 months

How do number of movies released and number of ratings given relate over the same time period?



Number of movies released and number of ratings follow approximately the same trends- the have the same peak in december 2024 and the same low in the 2025 months

Conclusions and Findings



Conclusion

There was not a strong correlation between any of the numeric variables in this data set. The strongest (still weak) was between movie budget and revenue

Movie ratings (0-10 possibility) are roughly left skewed and have a mean of 6.58/10

Run times had a bimodal distribution with a mean of 95.71.

There seem to be larger numbers of movies released in Fall and Winter months

Number of ratings and number of movies released in a month follow the same pattern