1. Datasets

I choose TMDb movie data to analyze. It was chosen based on my preference for movies.

This data set contains information about 10,000 movies collected from The Movie Database (TMDb), including information about:

+ id

+ imdb_id

+ poplularity

+ budget

+ revenue

+ original_title

+ cast

+ homepage

+ director

+ tagline

+ keyworks

+ overview

+ runtime

+ genres

+ production_companies

+ genres

+ nies

+ release_date

+ vote_count

+ vote_average

+ release_year

+ budget_adj

+ revenue_adj

2. Asking questions:

- For the first question, I want to know what is the average length of movies.
- Second, What is the average budget for making films per year?
- Third, How many movies are made each year?
- And last, I am curious about the ratio between revenue and budget. What is the difference between revenue - budget? And how many movies have revenue - budget > budget? Does this ratio depend on the release_year or not?

3. Cleaning data:

I do four steps to clean data before analysis:

- Drop Extraneous Columns With the questions I want to find answers, I will focus on some main columns: budget, revenue, release_year, and genres:

I will keep them and some more necessary pieces of information: id, imdb_id, original_title, cast, runtime, and production_companies.

- Drop Nulls
  In this step, I will check seen have line is NULL or not. If rows are NULL, we need to drop them.
- Drop any duplicate rows in both datasets:
  Remove rows duplicated here
- Clean data Os in revenue, budget, and runtime
- Check types of data:
  When read from .csv file. Some records will have the type is 'object'. We need change to the string.

4. Descriptions of what I do to investigate:

- what is the average length of movies?

For this question, I just need to calculate the average of the run_time colums.

- What is the average budget for making films per year?

For this question, first, I need to use the group by method to group the data of release year and average budget (get by mean()).

After that, I draw bar graphs to see visual results. It's a really good way to see the relative average of the budget.

- How many movies are made each year?

Similar to y above, I still use group by method to group the data for each year. With this question, I use .count() to count the number.

The difference with the above question, I want to choose line graphs here. Because I want to see the growth in number of movies.

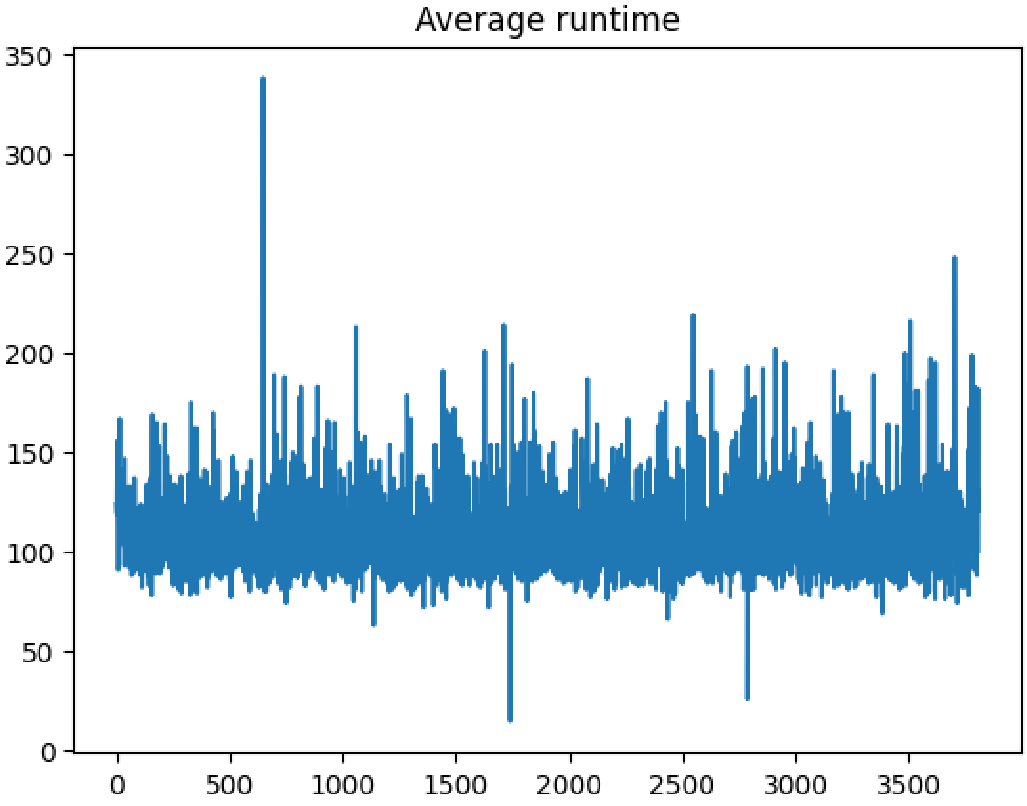
- What is the difference between revenue - budget? And how many movies have revenue - budget > budget? Does this ratio depend on the release_year or not?

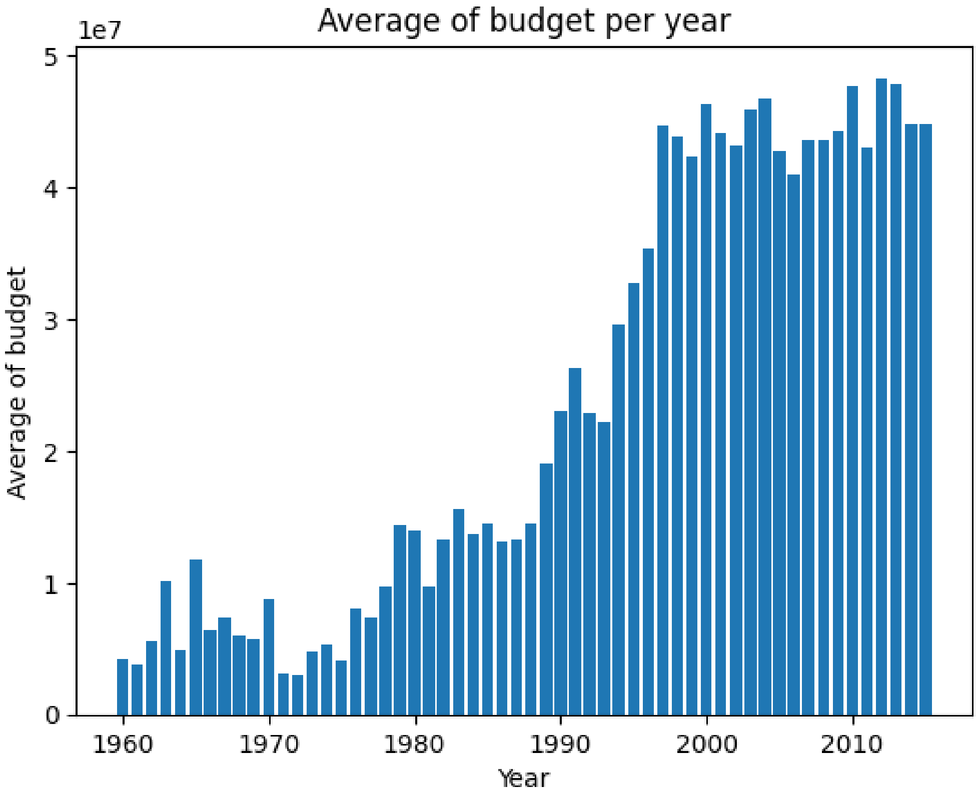
  First, I calculate the difference between revenue and budget.

  Second, I will compare this result with the base budget of movies.

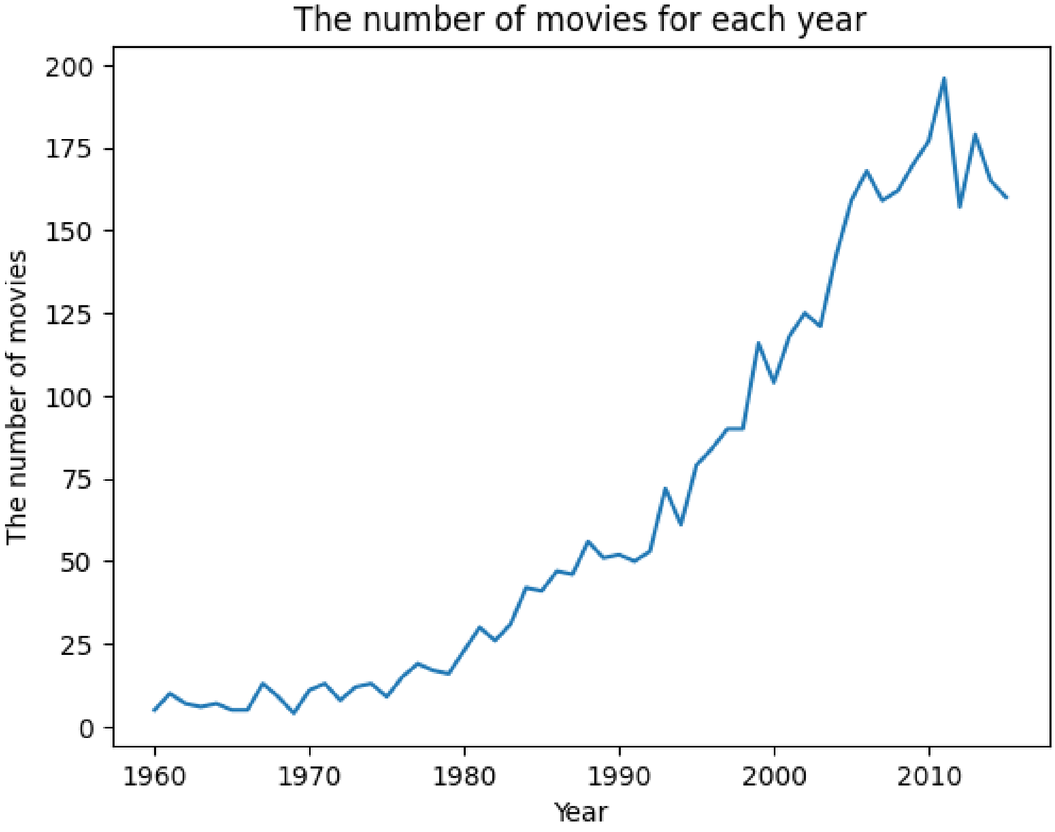
  And last, I choose to scatter graphs to draw visualizations. Because the scatter allows researchers to identify anomalies in the data more easily.

5. Visualizations

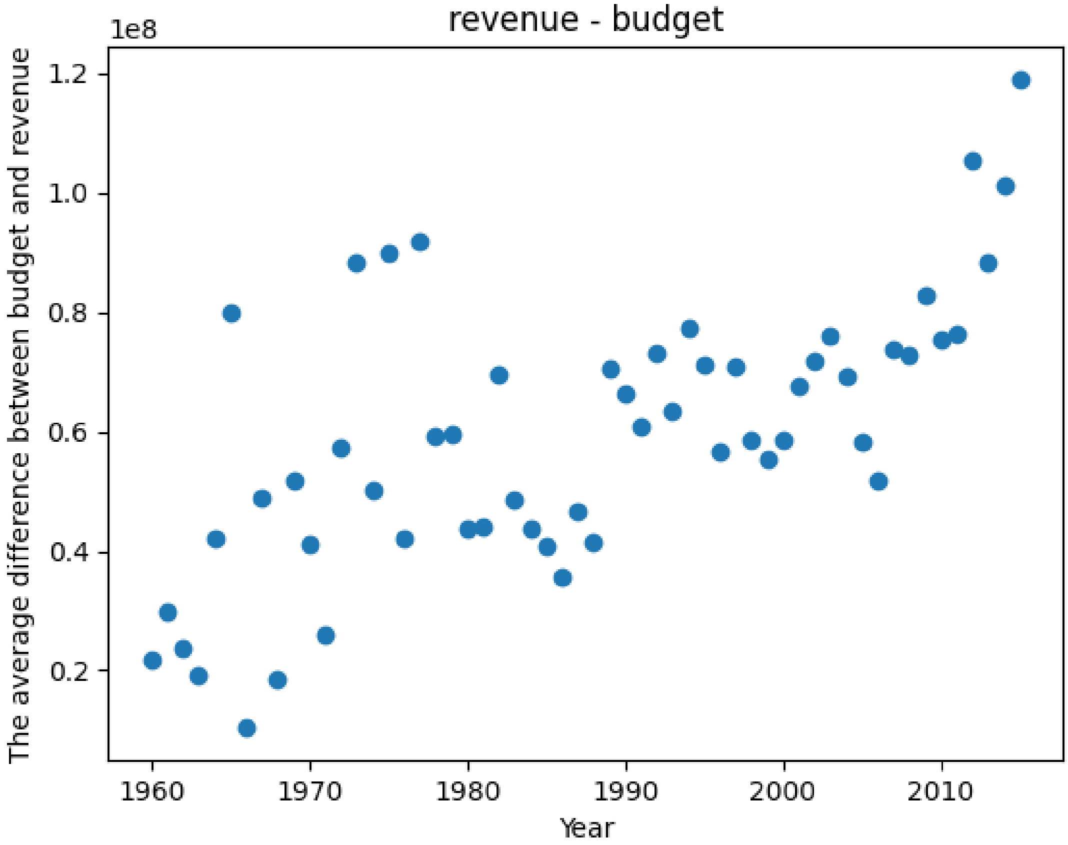




Average of budget per year



The number of movies for each year



6. Conclusions:

- The average time is almost 2 hours. It's been a pretty good time for movies.
- After 1995s, the film industry was the most developed (Because the budget was very high and the difference between revenue and budget was also high).

- In general, the film industry has markedly developed. It also brings more profit (based on the difference between revenue and budget) – because of the slope of the graph to the left.
- The number of movies is increasing day by day.
- The budget has slowed down in recent years but it's still at a high.

Shortcomings/factors limiting this analysis:

- This analysis points to the number of movies per year but doesn't care about the average budget for each movie.
- This analysis doesn't care about the genres of movies, It's also important part that affects the revenue.