



taingocbui /  
phase2\_project



<> Code

Issues

Pull requests

Actions

Projects

Wiki

Security

Insights

Settings



☆ 0 stars    🍴 0 forks    👁 1 watching    🌿 1 Branch    🏷 0 Tags    ↗ Activity

🌐 Public repository

main

1 Branch

0 Tags








Go to file

Go to file

+

Add file

Code

 taingocbui	include presentation.pdf	1 minute ago
	photos	update README
	.gitignore	update gitignore
	Project.ipynb	include notebook pdf
	ProjectNotebook.pdf	include notebook pdf
	README.md	update README file
	presentation.pdf	include presentation.pdf

## README



# Movie Data Exploratory Analysis

Author: Tai Ngoc Bui

Date Completion: March 24th, 2024

## 1.Business Understanding

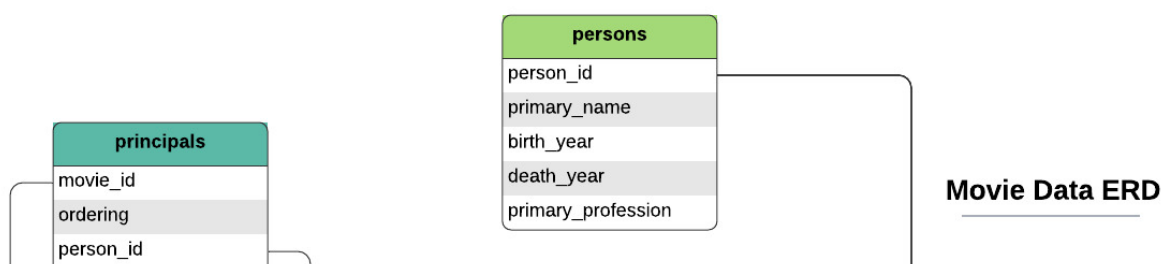
This data exploratory project focuses on analyzing characteristics of successful movies at the box office to support the company's strategic investment in the movie industry. The goal is to gain valuable insights contributing to movie production success and provide meaningful recommendations to major stakeholders of the investments.

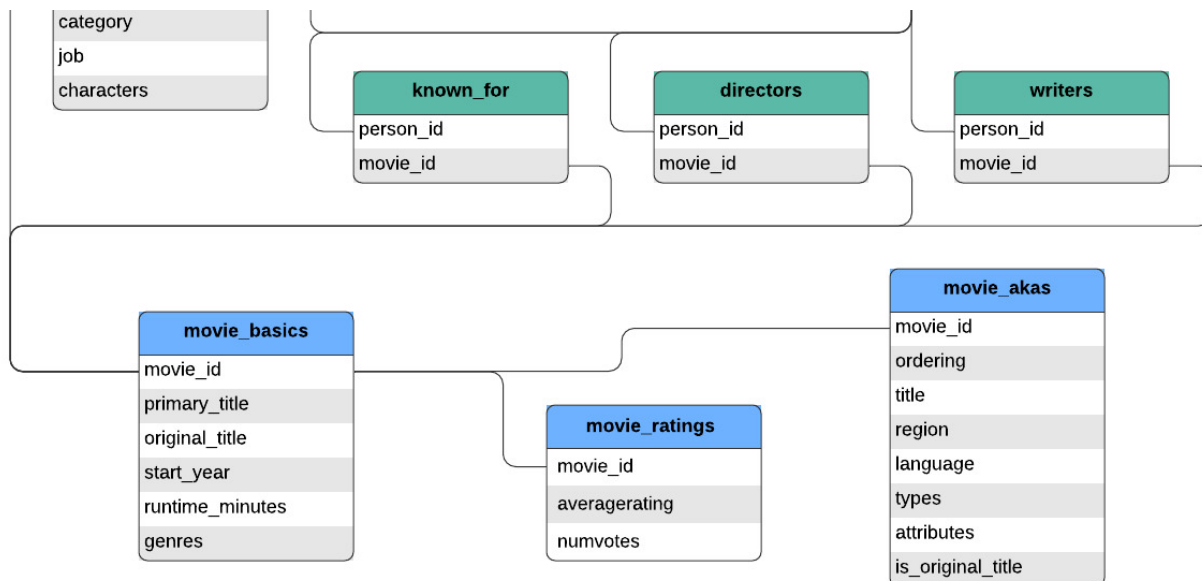
## 2.Data Understanding

The dataset used in this analysis extracted from various different movie sources including:

- [IMDb](#)
- [The Numbers](#)
- [TheMovieDB](#)

As these datasets were collected from different sources, they have different formats. While The Movie DB (TMDb) and the Numbers datasets are compressed as CVS files, the IMDb, the largest dataset among the three, is stored within a SQLite database.





These datasets not only contain movies' information on their casts, directors, budget, and revenues, etc. but also the public opinions regarding the movies' success such as ratings and votes.

### 3. Objectives

- Explore and analyze three different datasets to identify general trends in the movie industry, and the major factors that contribute to the success of a movie.
- Provide data-driven evidence to support confidence in the strategic investment and the best strategies to minimize the overall investment risks.
- Identify a list of potential directors based on their track records to ensure the success of stakeholders' investment in the movie industry.

### 4. Key Questions

With this project, I will address specific questions such as:

- What specific measure should be used to compare the success of movies and their production team?
- Would it be possible to reduce investment risk if we prioritizing certain movies' genres or release month?
- Who should be prioritized to be our studios' lead directors?
- What should we learn from the success of past blockbusters?

### 5. Data Cleaning

As data is extracted from different sources, each dataset requires different cleaning techniques.

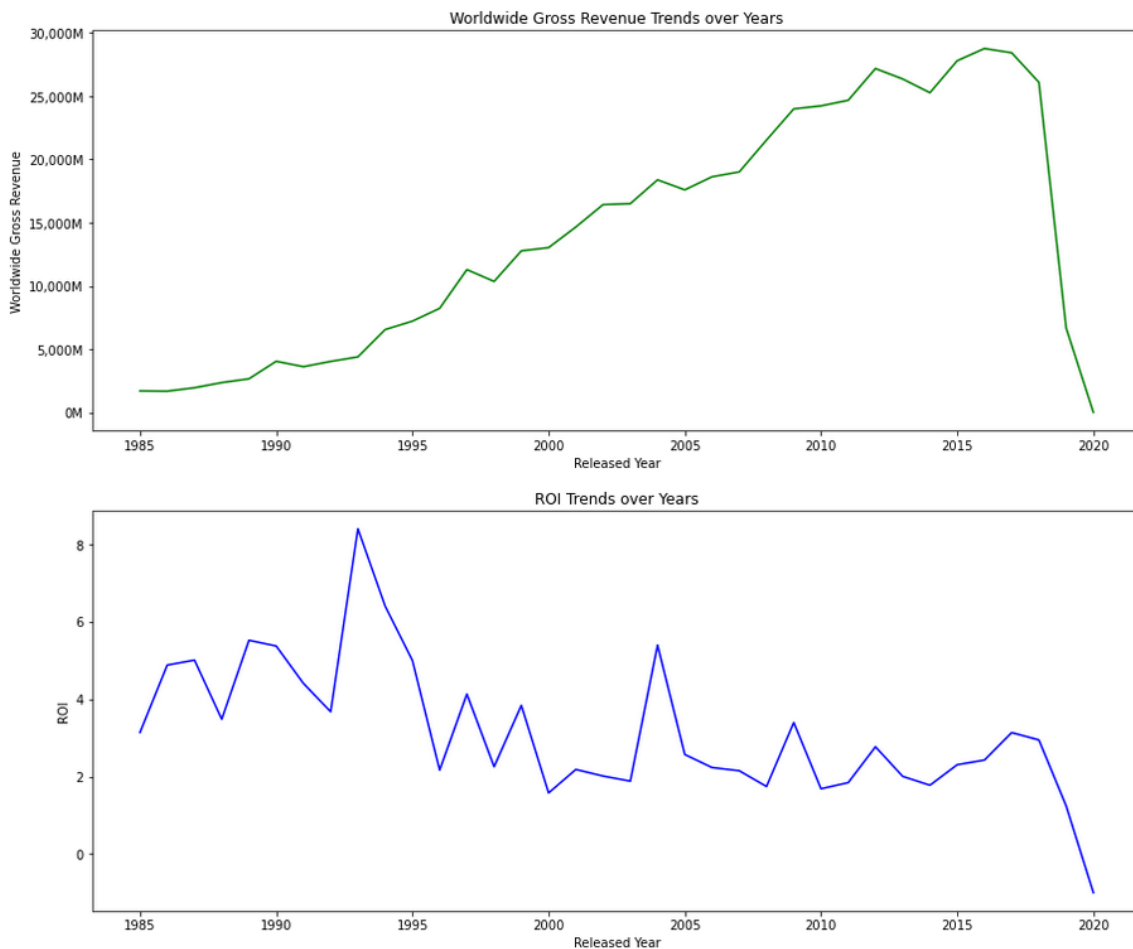
- IMDb records with directors who are still alive (the category column is not Nan) are filtered.
- With IMDb's quantitative features such as average ratings and number of votes, Nan values are replaced with zero, while the runtime minute's Nan values are replaced with the column's mode.
- The Numbers' and the Movie DB's datasets do not have any Nan values or duplicates.
- All quantitative features in the Numbers' dataset such as production budget, and gross revenue are all formatted as string type. To convert these features into float type, all special characters \$ sign had to be removed from these strings.

removed from these strings.

- The Movie DB's genre id code needs to be converted to actual genres' names. These genre ids are splitted into multiple columns. Each column's headers is later converted to an actual genre name.
- If the movie is assigned with a certain genre, that genre type will show as 1 for that column, else will be 0.

## 6. Exploratory Data Analysis

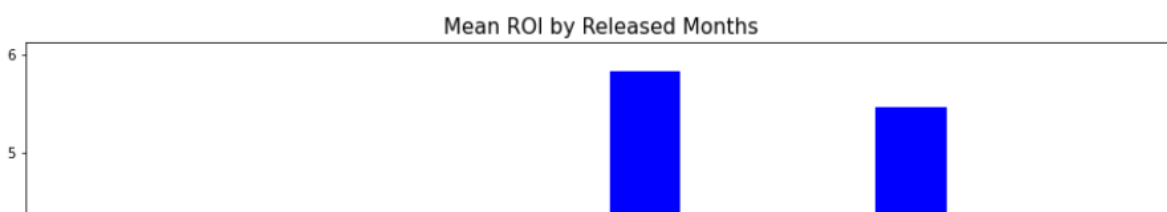
For this project, I decided to use Return on Investment (ROI) as the core feature to determine the level of sucess for various movies. This ROI measure is simply a rate between the gross profit (difference between production budget and gross revenue) and the production budget. The reason I use this ratio is that it normalize the difference in scales among movies' production budget and gross revenue. The impact of normalization can see clear in the below visualization.

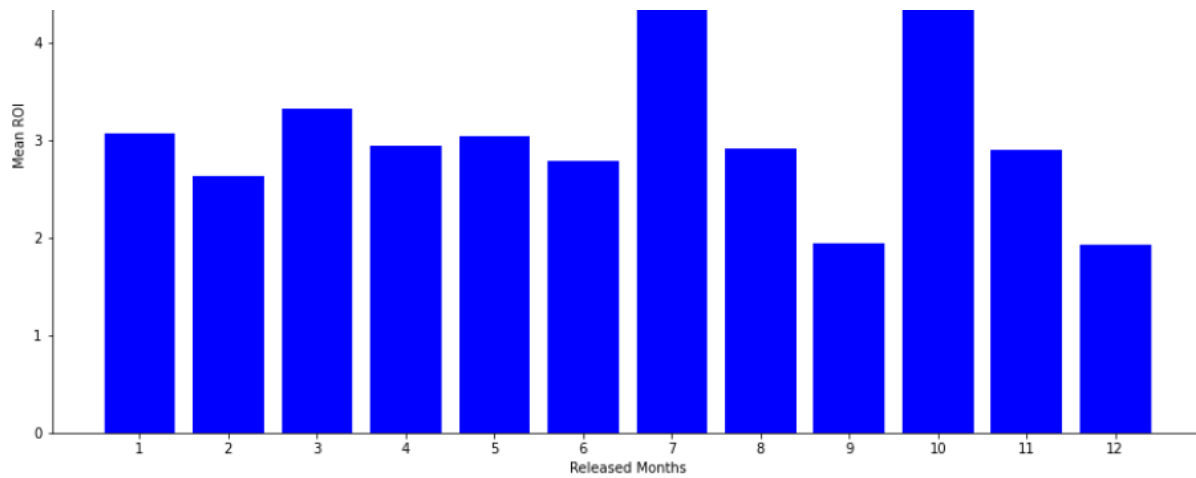


While gross revenue from movie production had steadily increased during the period between 1985 and 2000, return on investment in the industry had been on a slight down trend since.

Based on ROI, this exploratory analysis answers three questions to our stakeholders

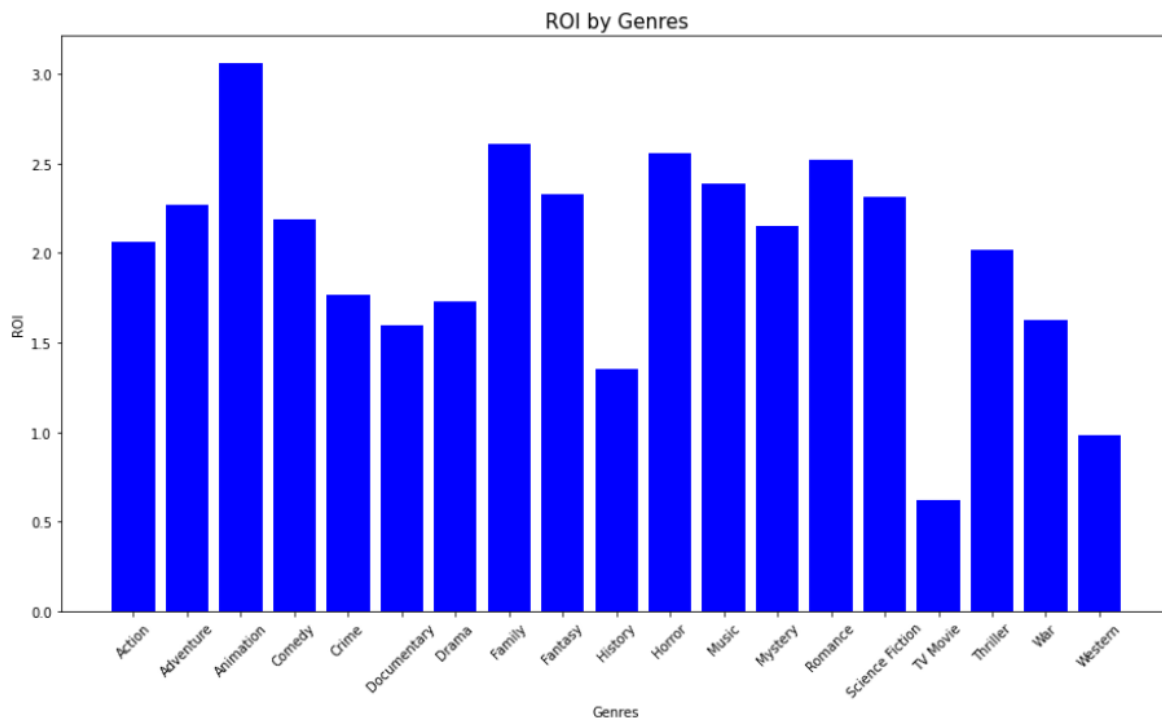
### 1. Which month is best for movie release?





The chart above shows that July and October are the two months with the highest ROI for movies to be released. However, as July tends to be the month in which blockbuster movies premiered according to budget production analysis, I recommend the stakeholders to prioritize releasing movies in October.

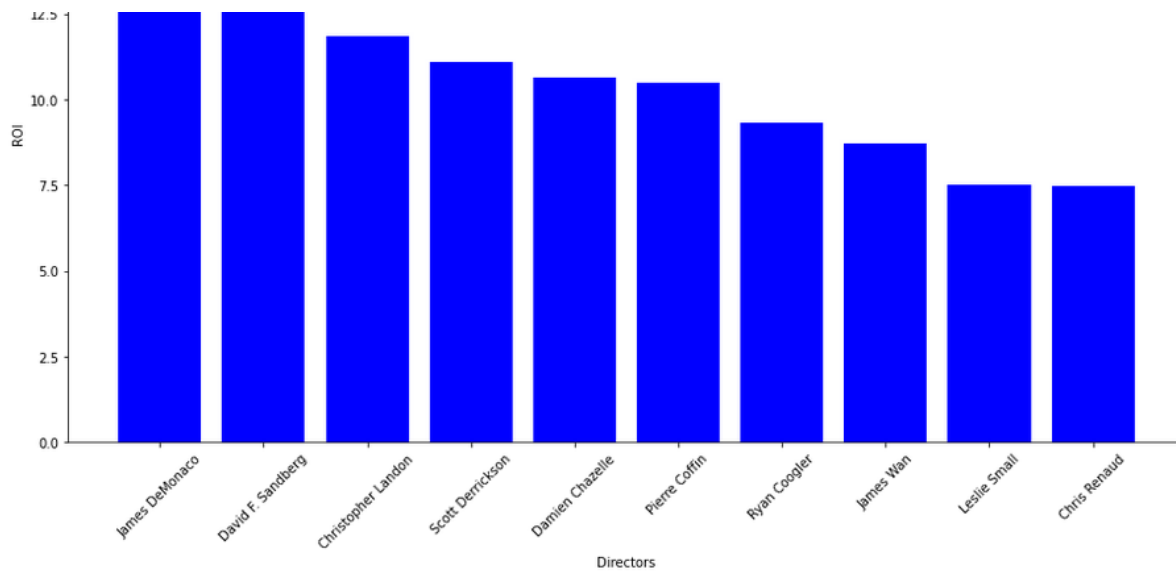
## 2. Which movie genres should the stakeholders invest in?



The ROI analysis shows that animation, horror, romance, and family movies are movie genres in which investors can receive great returns for their investments. They not only have the highest returns, but also lowest production budget compared to other genres.

## 3. What directors should the stakeholders select to direct their movies?





The chart above shows a ranking of top directors based on their movies' ROI. Though James DeMonaco, David Sandberg and Christopher Landon are the top 3 directors based on their movies' ROI, David Sandberg is not included in my recommendation to stakeholders. Unlike the other 2 directors, David's movies have a much higher production budget. Thus, selecting David may lead to a higher production budget, deeming a greater risk to the investment. My final recommendation for this question based on not only ROI but also the directors' ability to utilize small budget.

## 7. Conclusion

Based on my exploratory analysis, I have three recommendations for the stakeholders:

- Prioritize releasing movie in October of the year. Not only we can avoid compete directly with major movies, but also have a higher chance of hitting higher ROI rates.
- Prioritize producing animation, horror, romance, and family movies due to their general low production cost relative to gross revenue and high ROI rate.
- Prioritize inviting James DeMonaco, Christopher Landon, and Damien Chazelle to be the directors of the company's movies. They not only have great experience in the movie industry but also good track record of turning small budget movies into great returns.



### Releases

No releases published

[Create a new release](#)

Packages

No packages published  
[Publish your first package](#)

---

Languages

● Jupyter Notebook 100.0%