



# Movie Metadata Analysis

---

*By Nataliya Pivnitskaya*

pivnitskaya.natali@gmail.com



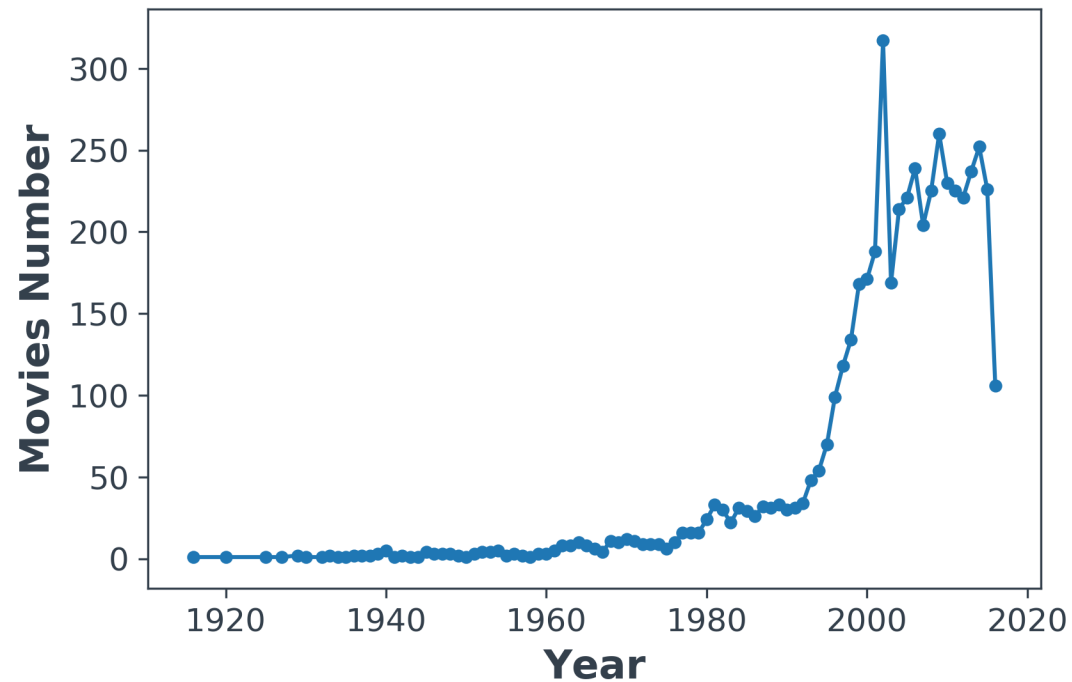
1.

**What are your observations based on exploration of this data?**

2



## Number of Movies Over Time

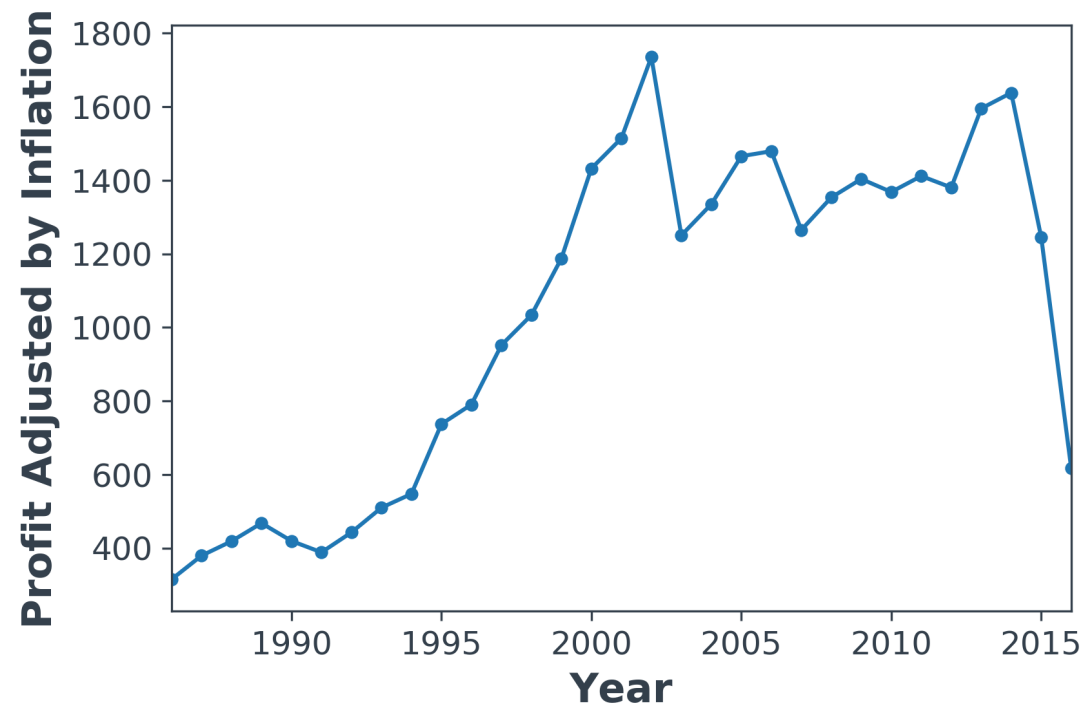
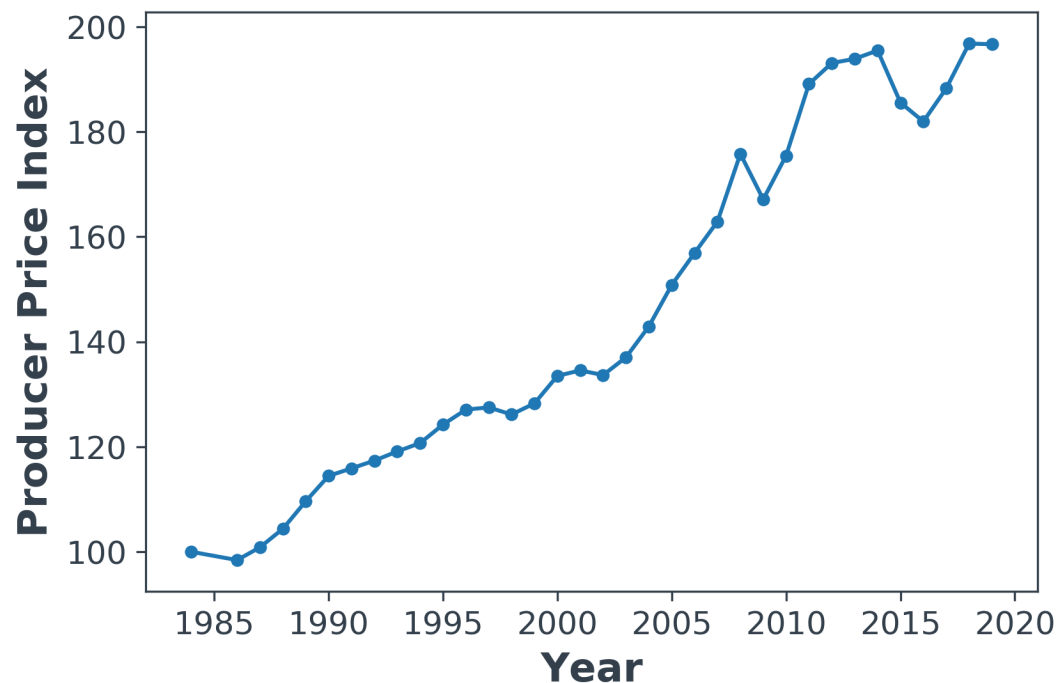


- Note 1. Before 1995 the number of produced movies remain very low, at the beginning of 2000 we can observe a sharp increase in produced movies number.
- Note 2. The latest point significant drop most likely correspond to the lack of observations for the last year in data set so we don't need to take it seriously as drop due movie domain overall.
- Note 3. It might be difficult to features connected with 'likes', for example, '*number of facebook likes*' since people start to use Facebook not so long time ago\*.
- Note 4. We can reasonably delete very old movies without any loss since they don't provide much information for the future implementation (due to low number and protentional old style)

\* <https://www.statista.com/>

## Let's take inflation into consideration

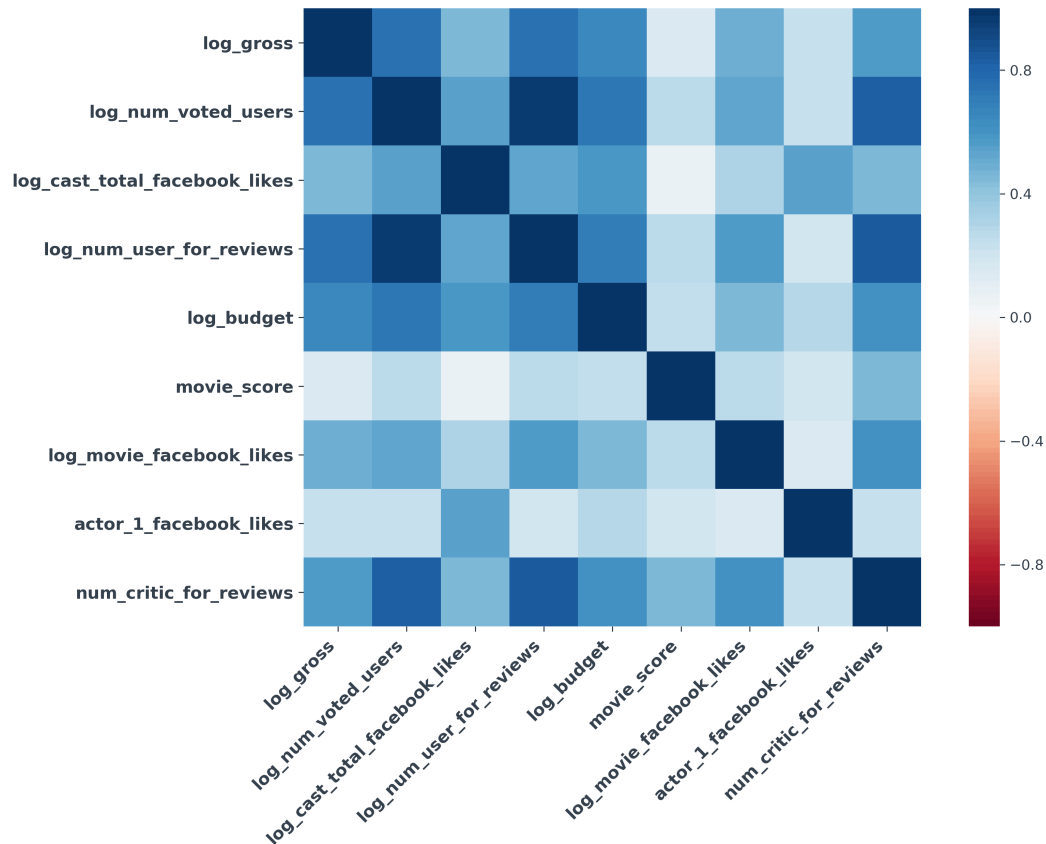
To make analysis more fair I would like to adjust quantitative variables connected with the level of price (budget, gross, profit\*) by inflation impact \*\*



\* Profit = Gross - Budget. I also analyzed relative measure  $(\text{Gross} - \text{Budget}) / \text{Budget}$ , but results on this presentation are given in absolute profit value.

\*\* Producer Price Index was taken from <https://inflation>.

# How quantitative features are correlated?

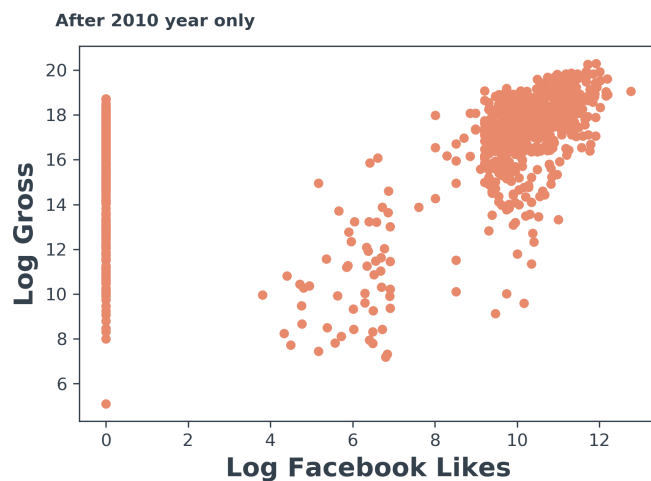
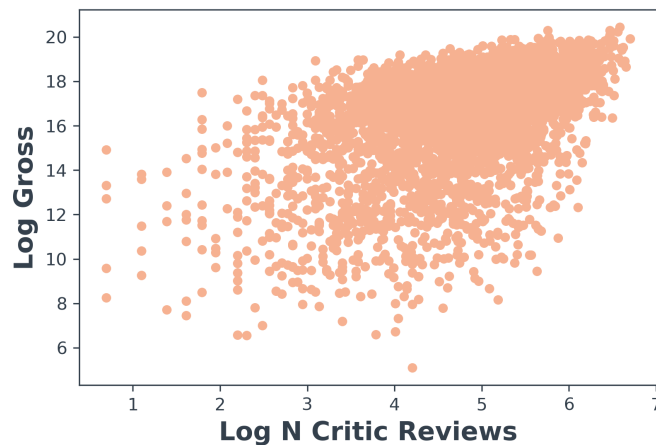
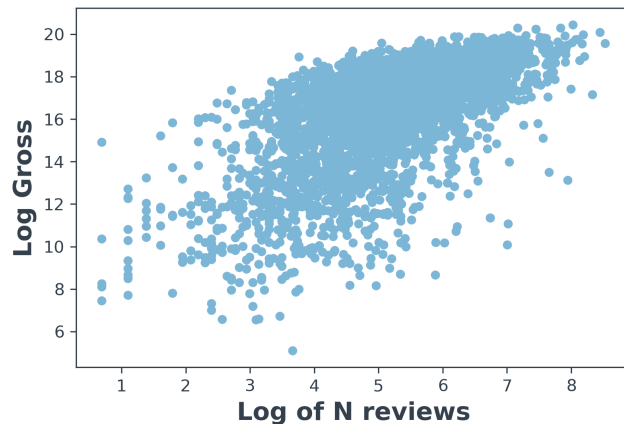
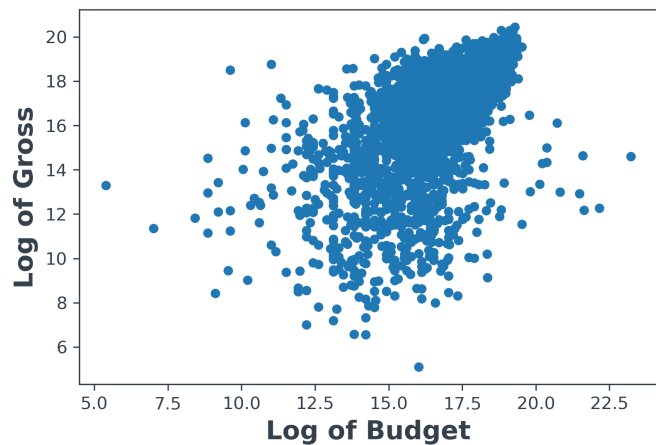


## Take-aways:

- `log_gross` correlated with `Log number of votes`, `Log Number of users who gave a review` and `Budget`
- `'log_num_votes_users'`, `'log_num_users_for_reviews'`, `'log_num_critic_for_reviews'` are highly correlated – no need to analyze all of them
- No high correlation between gross and any sort of likes (cast and movie)

\*Here and later were used only movies that were produced after 2012, because before 2012 Facebook wasn't so popular: <https://www.statista.com/>.

# Scatter Plots



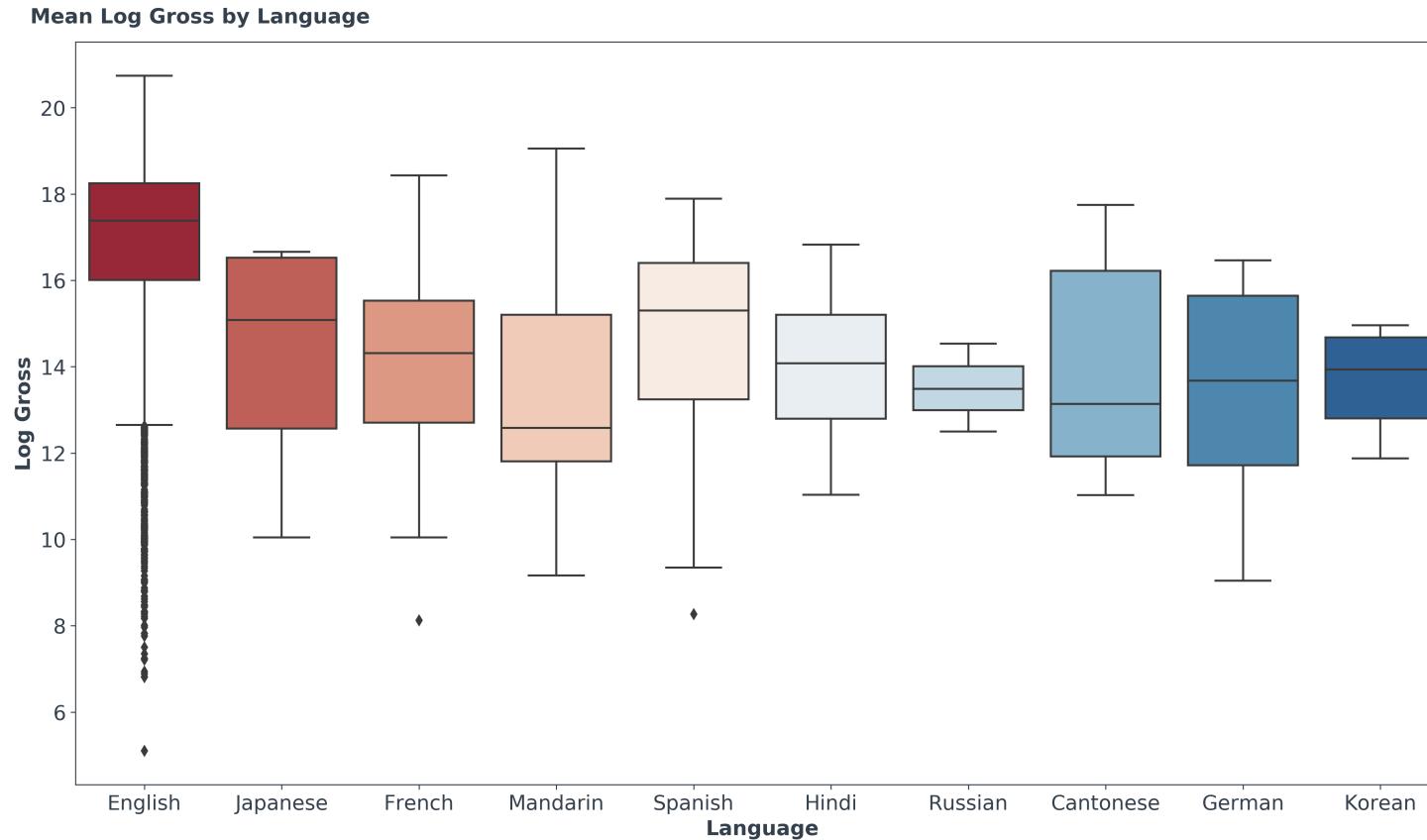
## Take-away:

We haven't found the correlation between 'Gross' and 'Number of Likes' on the previous slide (heatmap), but based on the scatterplot they actually have some sort of dependency. But still for most of the movies people don't react on the Facebook with likes.

# Exploring Gross by Language

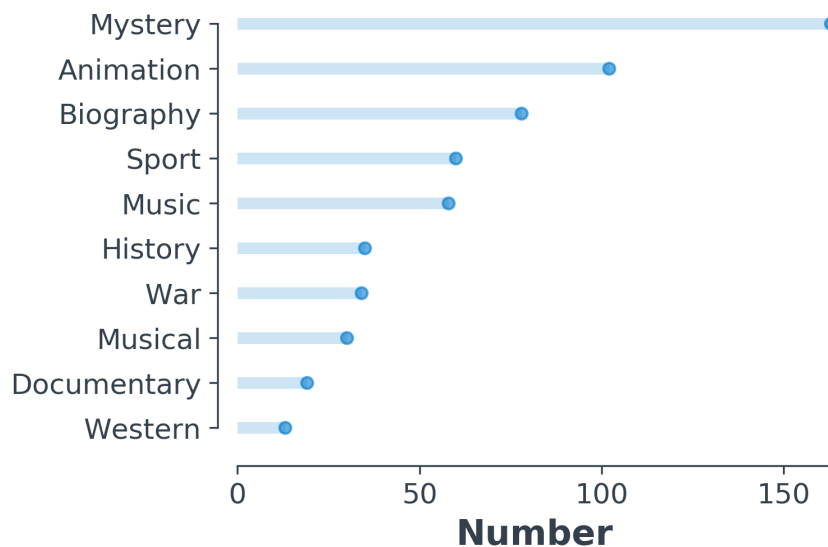
## Take-aways:

- English movies have the highest mean of 'Log Gross'
- A lot of outliers in 'Log Gross' of English movies
- Most of the movies in the data set are in English (93%). So this feature is not so important.

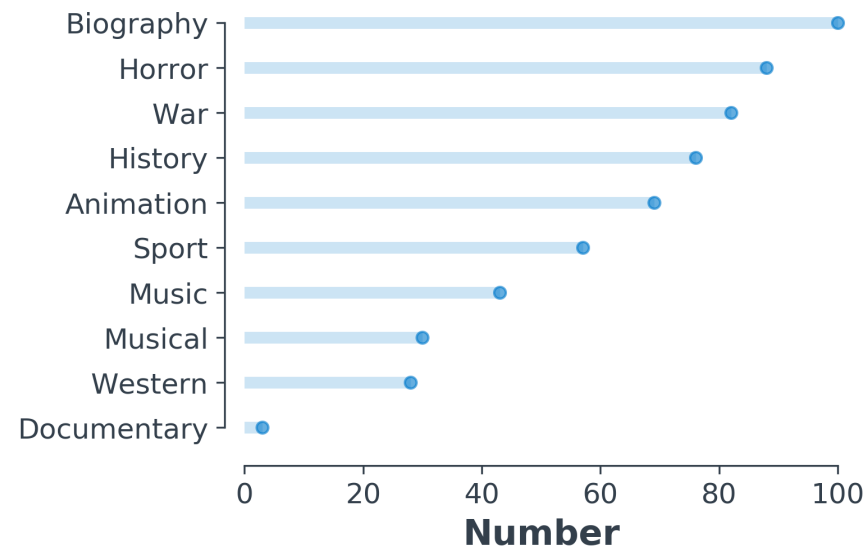


# Exploring Genres

## Most Profitable Genres



## Least Profitable Genres

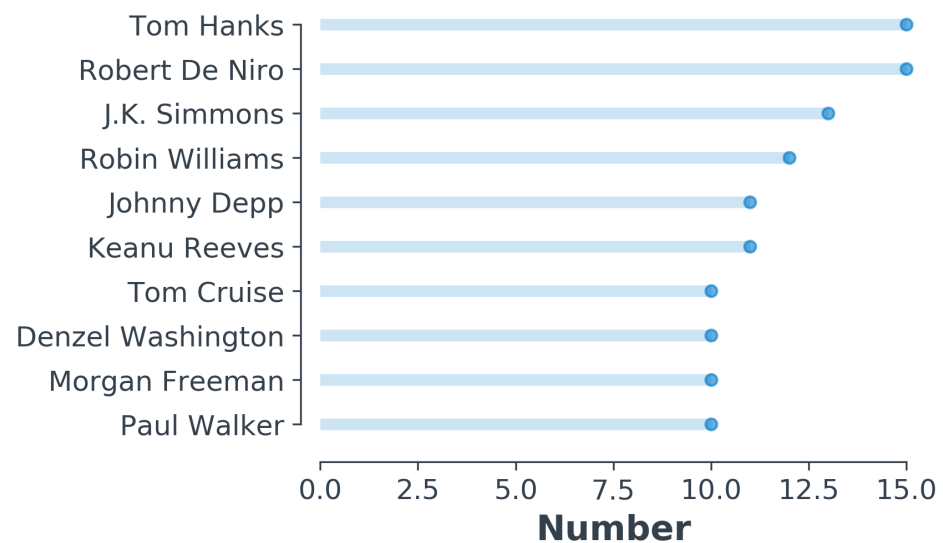


Biography, Horror, War Genres specialize on specific auditory, whereas Mystery has very wide auditory – being loved by kids and adults

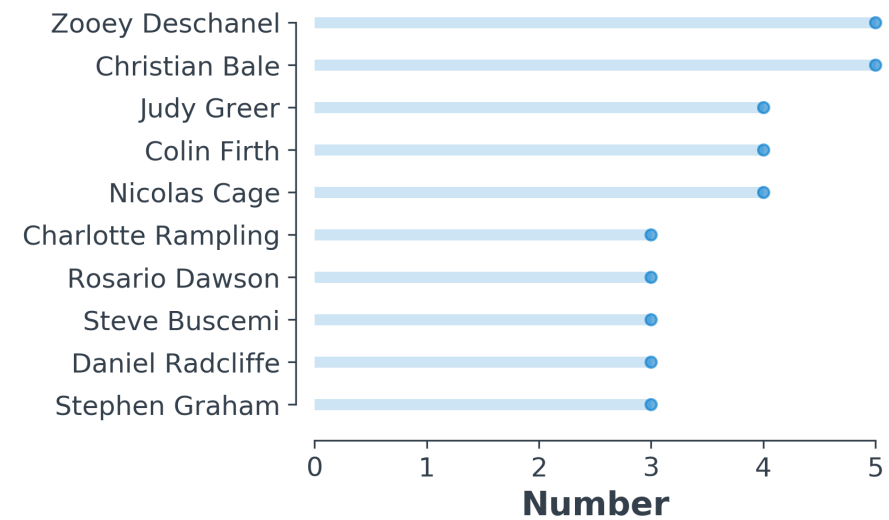


# Exploring Actors

## Most Profitable Actors



## Least Profitable Actors



As expected we can often see most popular actors in the list of most profitable movies.



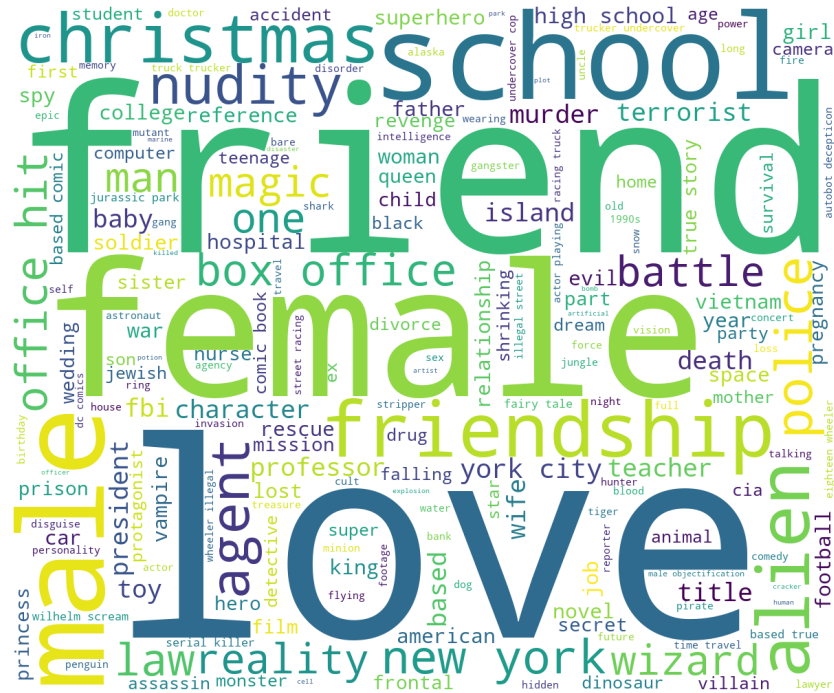
# Exploring

Most Pro

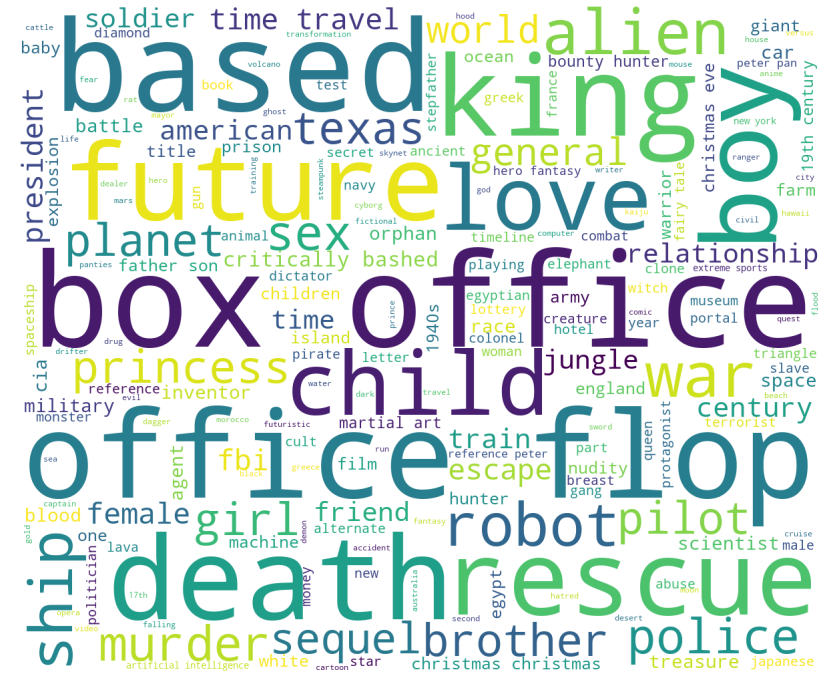
student • doctor accident

# christma

## Most Profitable Movies



## Least Profitable Movies



**Note.** Words Cloud helps to get some understanding on what key words present in profitable and not profitable movies. Looks like most profitable more centered around the idea of friendship, love and other positive aspects of life (again this topic captures wider auditory). Whereas on the least profitable movies side we can see some more specific words, for example, 'king' might be connected with the historical movie, some other examples: office, war, future, etc.



## Exploring Other Features

After some analysis some features turned out not to have any impact on profitability of movies much, for example:

● Duration

● Color

● Face number at Poster

● Movie link



2.

**What is the recipe to make a blockbuster, profitable movie?  
Share your hypothesis and insights based on the data.**

## Recipe to make a blockbuster: hypothesis

A surprisingly large number of hit films are **based on the myth genre**. Why? Because myth beats cultural age and national boundaries. This is why 'Batman' is as popular globally as it is in the U.S.

But you should keep in mind that blockbusters often combine the myth form with at least one other major genre. The other genres help to modernize the myth form, and also overcome many of the weaknesses inherent to this tricky genre. The trick is **to find the one or two best forms** that will bring out the 'gold' in your idea.

Hollywood is also said to be incapable of making an inexpensive film. These films have **massive budgets**, so they are expected to make huge revenues, as we also observed in analysis. They go through the trouble to cast **big-name actors** in their films.



# Thank you for your attention

**GitHub** repository with code can be found [here](#)

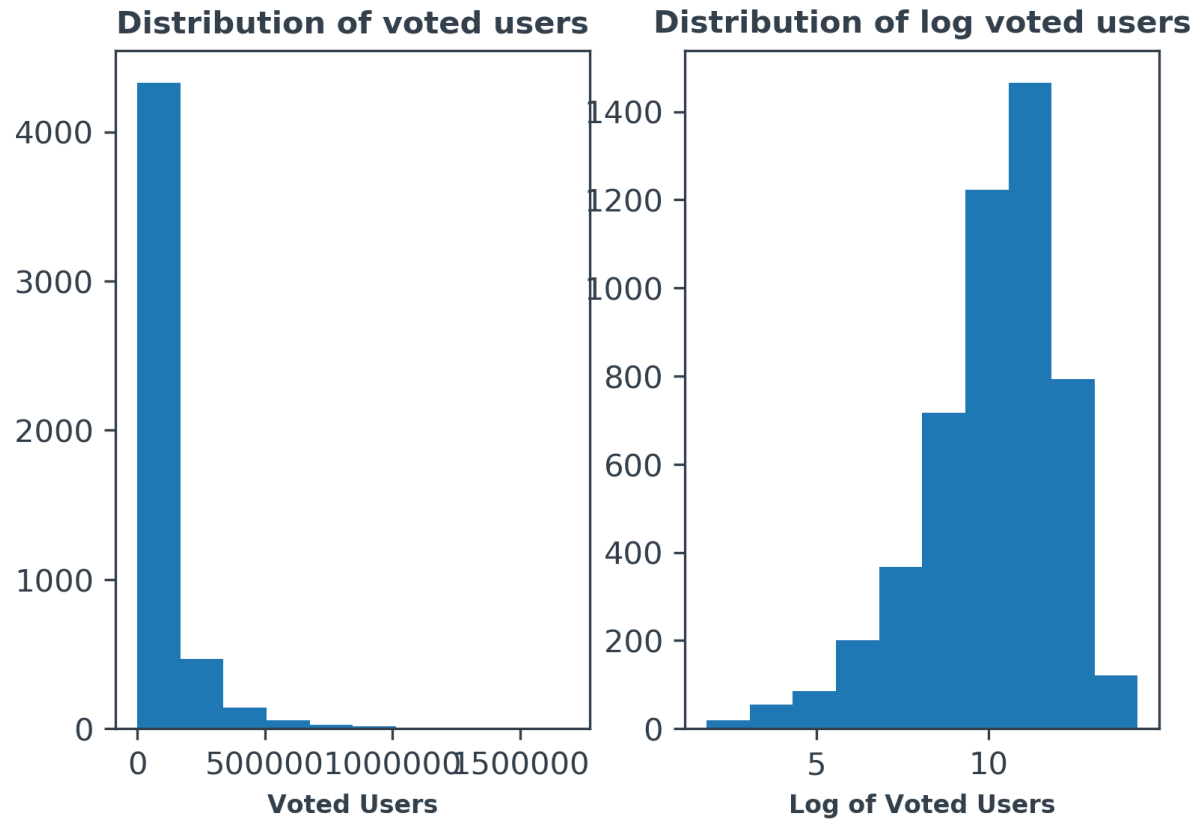




# Appendix



# Features Transformation



Note:

In current analysis log transformed features often used since log transformation help to make feature more normally distributed