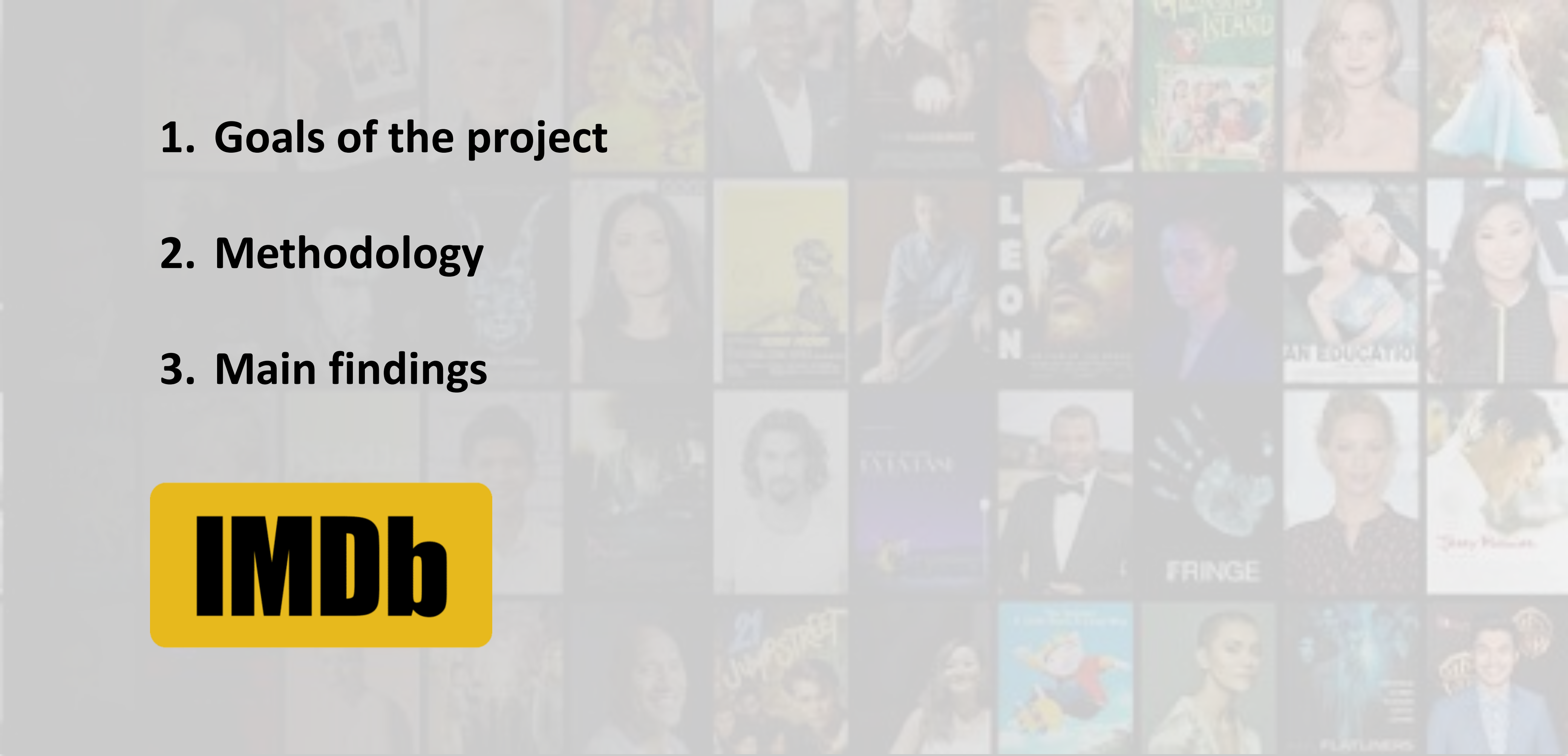
IRON HACK

The movie industry and the influential factors of IMDb movie ratings

Mafalda Fragoso

Data Analytics Bootcamp
27/03/2020

Table of contents



1. Goals of the project
2. Methodology
3. Main findings

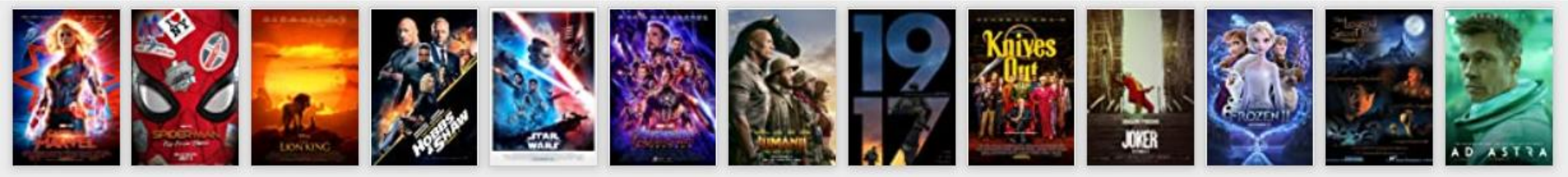
IMDb

Goals of the project

Research purpose

1. Exploratory analysis on the movie industry (2019)

2. Understand the influential factors of IMDd movie ratings:
   - Is the movie rating variable correlated with the duration of the movie, genre of the movie, number of votes and revenue?



Methodology

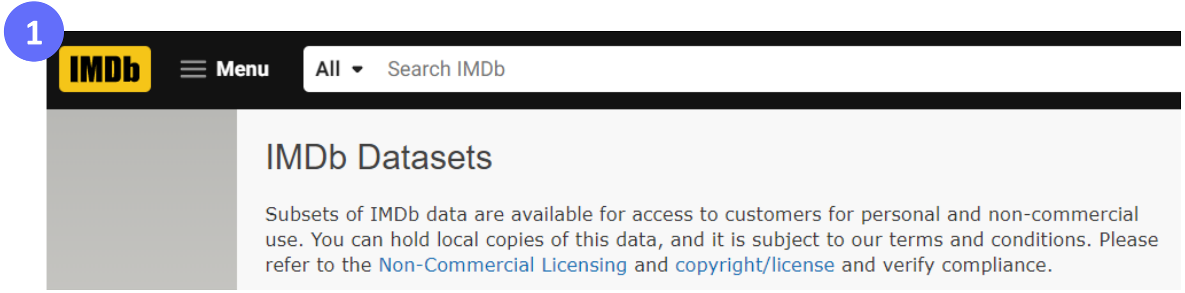
1. Data gathering

- IMDb » open datasets
- The Movie Database » API

2. Data cleaning

- Joined several tables from the IMDb open datasets with the information needed for the study
- Joined the final IMDb dataset with the data gathered from the API of The Movie Database (revenue and budget variables)
- Data cleaning included checking for missing values, number of duplicates, low variance and data types
- Sample of 1640 movies after cleaning the data

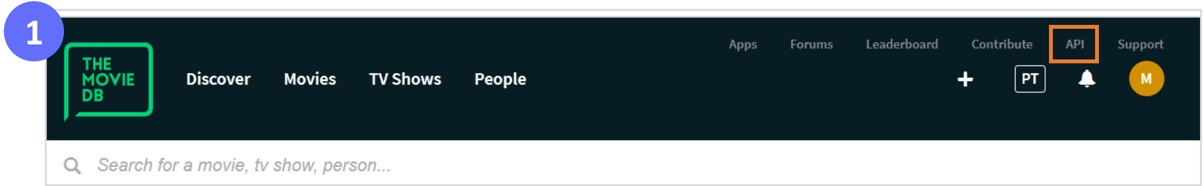
1



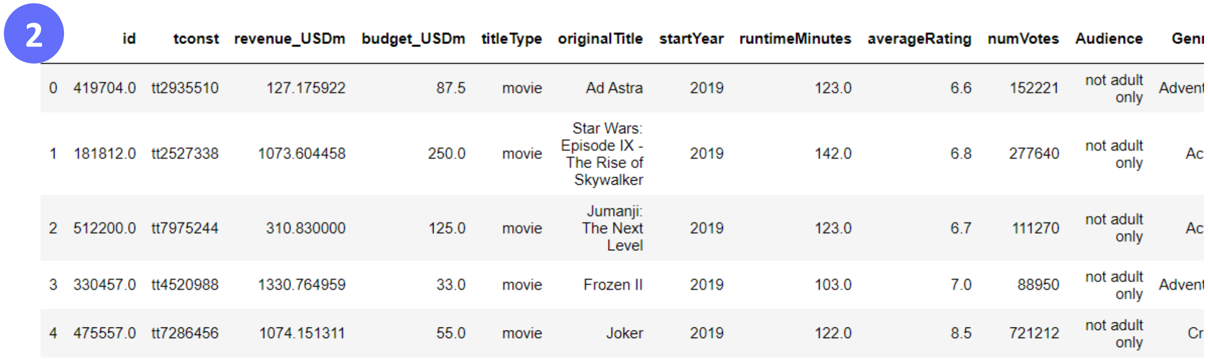
IMDb Datasets

Subsets of IMDb data are available for access to customers for personal and non-commercial use. You can hold local copies of this data, and it is subject to our terms and conditions. Please refer to the Non-Commercial Licensing and copyright/license and verify compliance.

1



2



	id	tconst	revenue_USDm	budget_USDm	titleType	originalTitle	startYear	runtimeMinutes	averageRating	numVotes	Audience	Geni
0	419704.0	tt2935510	127.175922	87.5	movie	Ad Astra	2019	123.0	6.6	152221	not adult only	Advent
1	181812.0	tt2527338	1073.604458	250.0	movie	Star Wars: Episode IX - The Rise of Skywalker	2019	142.0	6.8	277640	not adult only	Ac
2	512200.0	tt7975244	310.830000	125.0	movie	Jumanji: The Next Level	2019	123.0	6.7	111270	not adult only	Ac
3	330457.0	tt4520988	1330.764959	33.0	movie	Frozen II	2019	103.0	7.0	88950	not adult only	Advent
4	475557.0	tt7286456	1074.151311	55.0	movie	Joker	2019	122.0	8.5	721212	not adult only	Cr

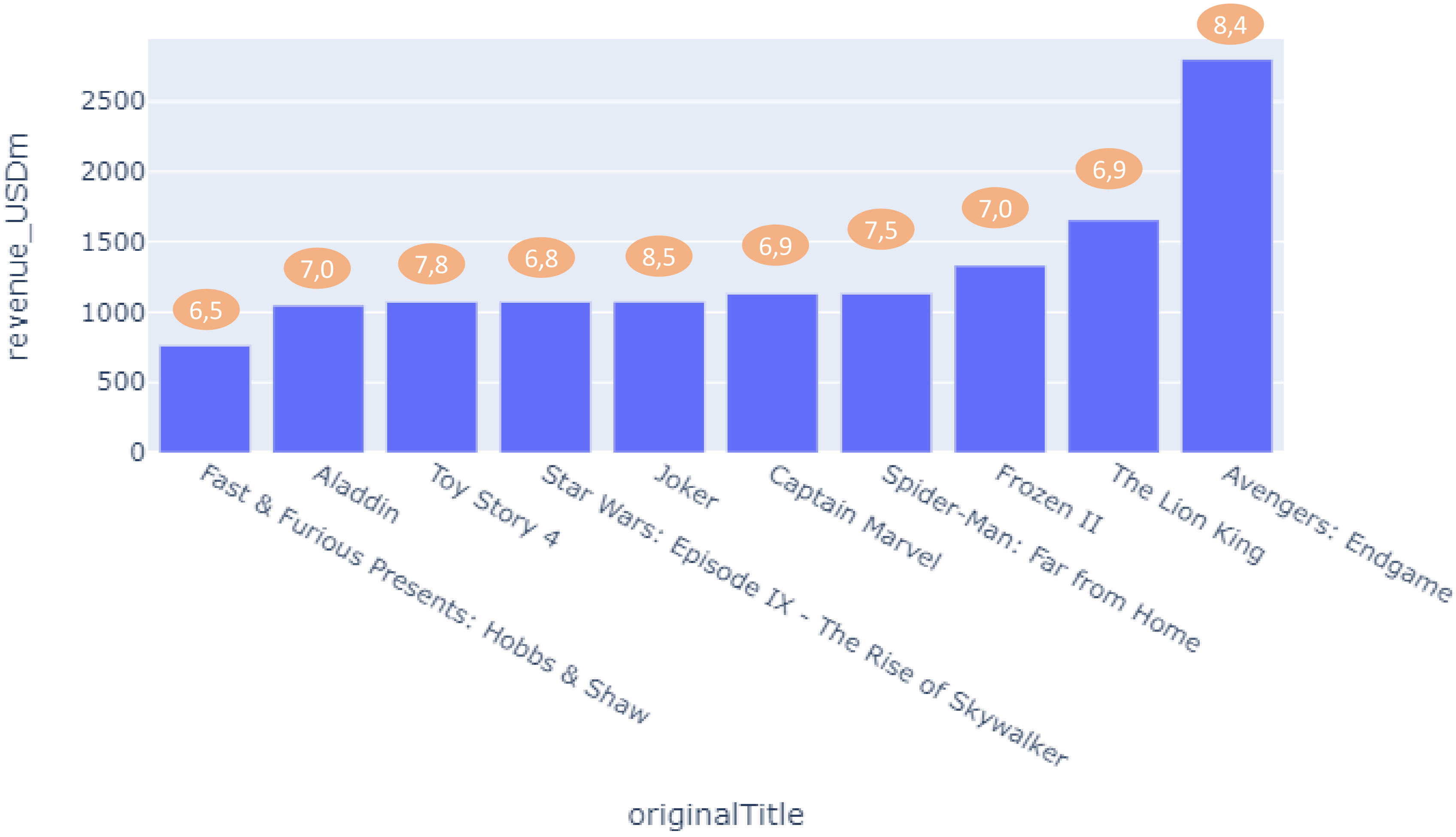
Main findings

The top movies released in 2019 based on revenue

(revenue in UDS millions - Cumulative Worldwide Gross)

● IMDb rating

■ Revenue

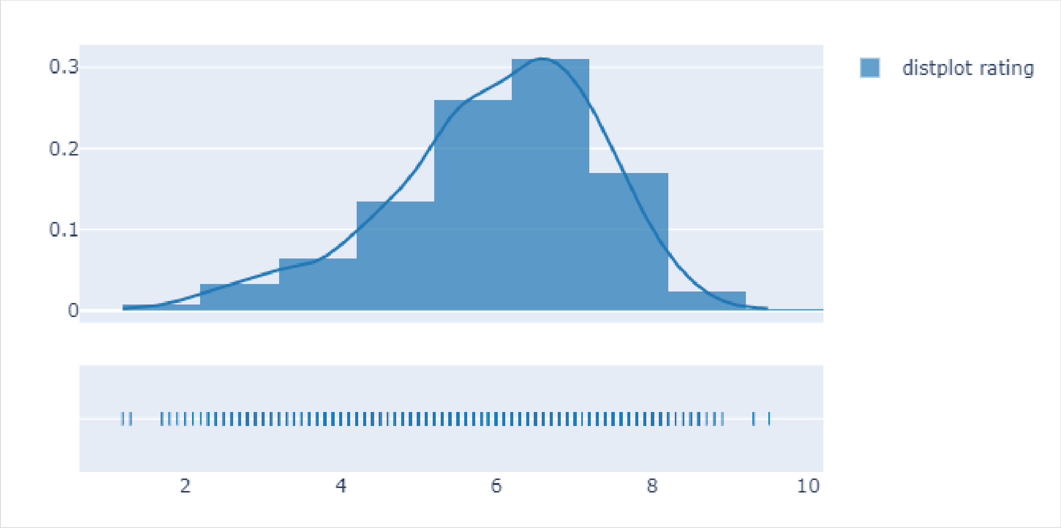


Main findings

Rating distribution

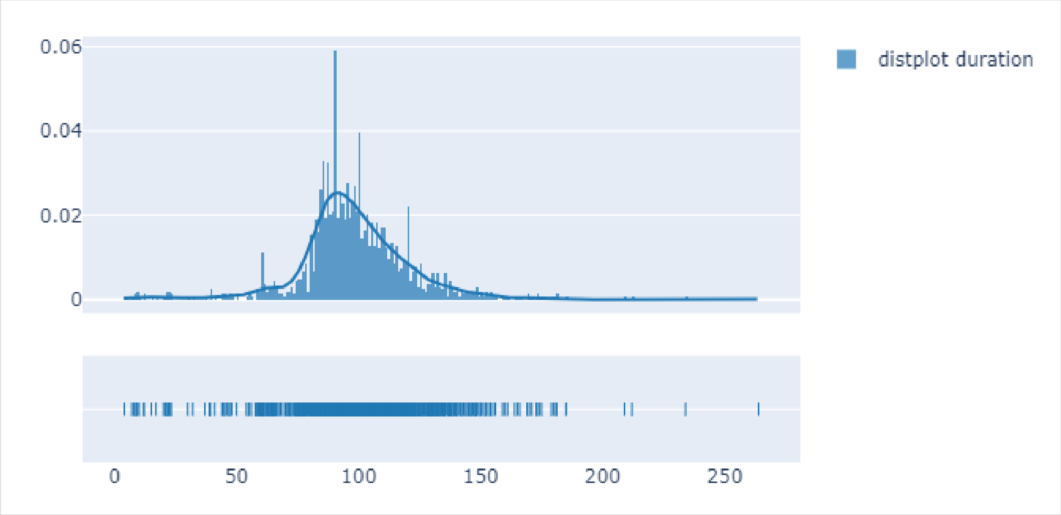
The average rating variable follows a normal distribution with a mean of 5,97

(rating scale from 1 to 10)



Duration of the movie distribution

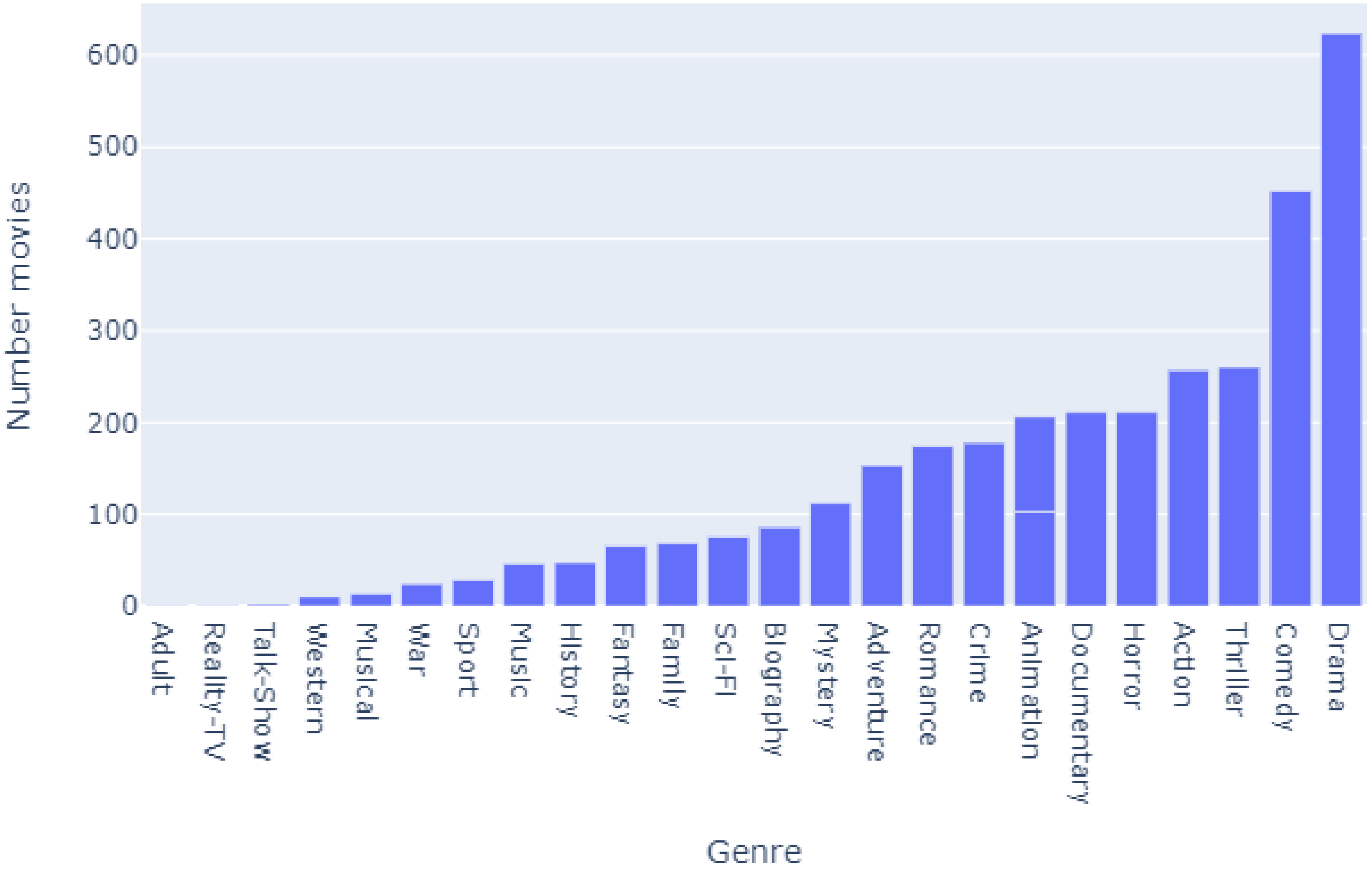
The duration of the movie variable follows a normal distribution with a mean of 98 minutes



Main findings

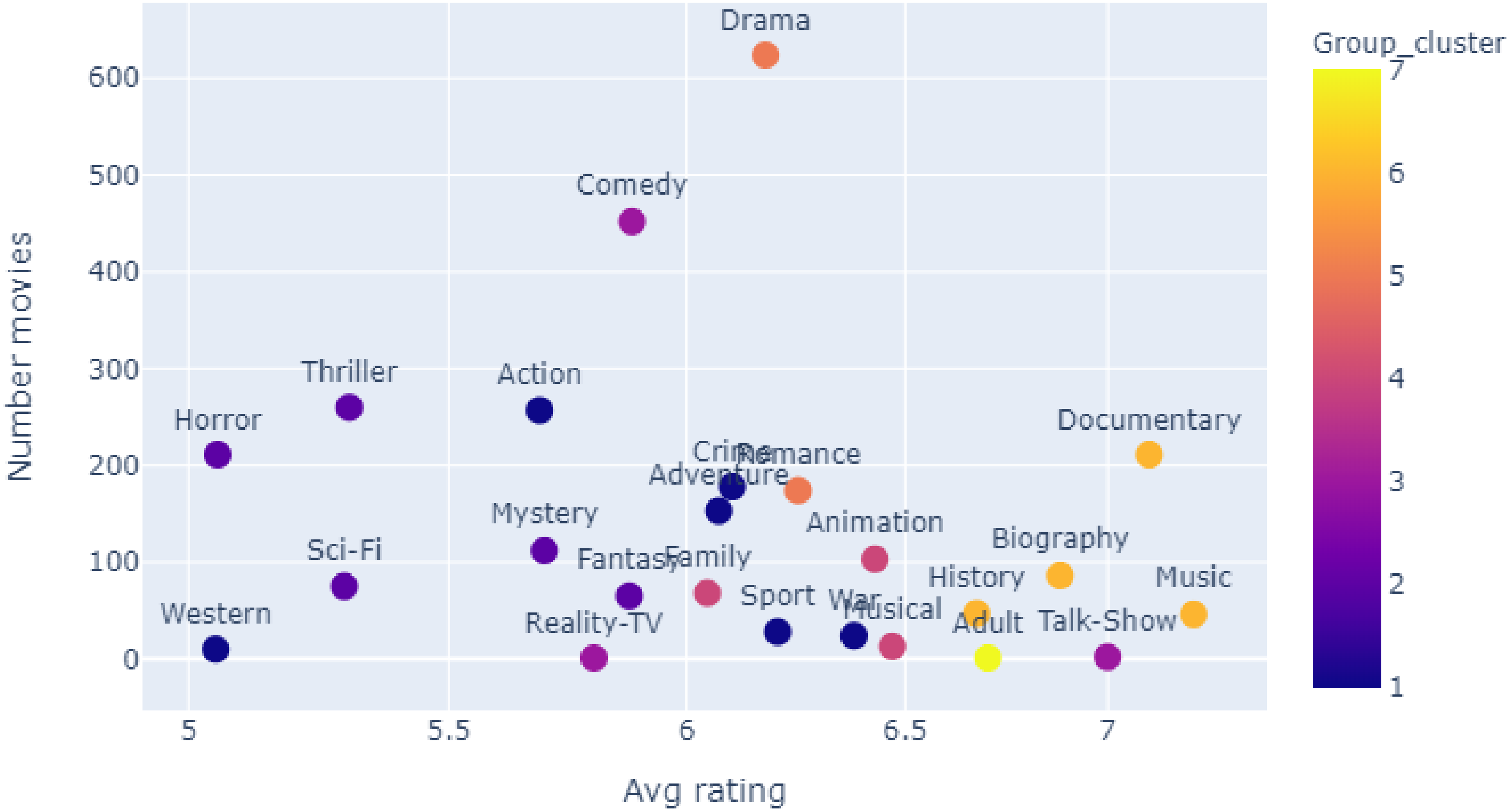
The most representative genres in my sample are Drama, Comedy, Thriller and Action

(note: each movie can have until 3 genres classification)



Main findings

Distribution of average rating vs number of movies per gender

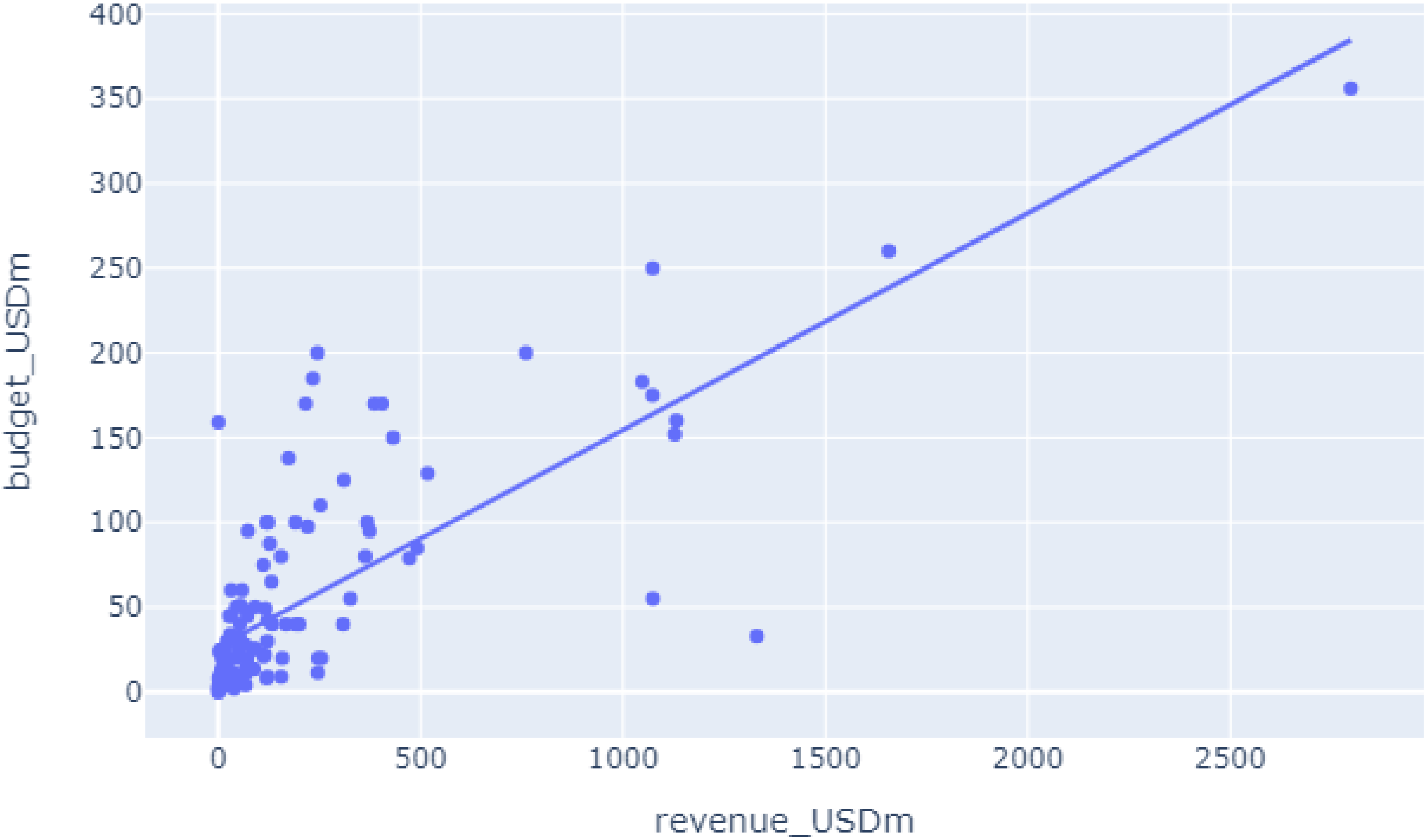


- There are less movies classified as technical/informative (documentary, Biography, History) but they are more likely to have higher ratings
- Movies in the cluster of action, horror, sci.fi are more likely to have lower ratings

Main findings

Strong positive linear relationship between revenue and budget

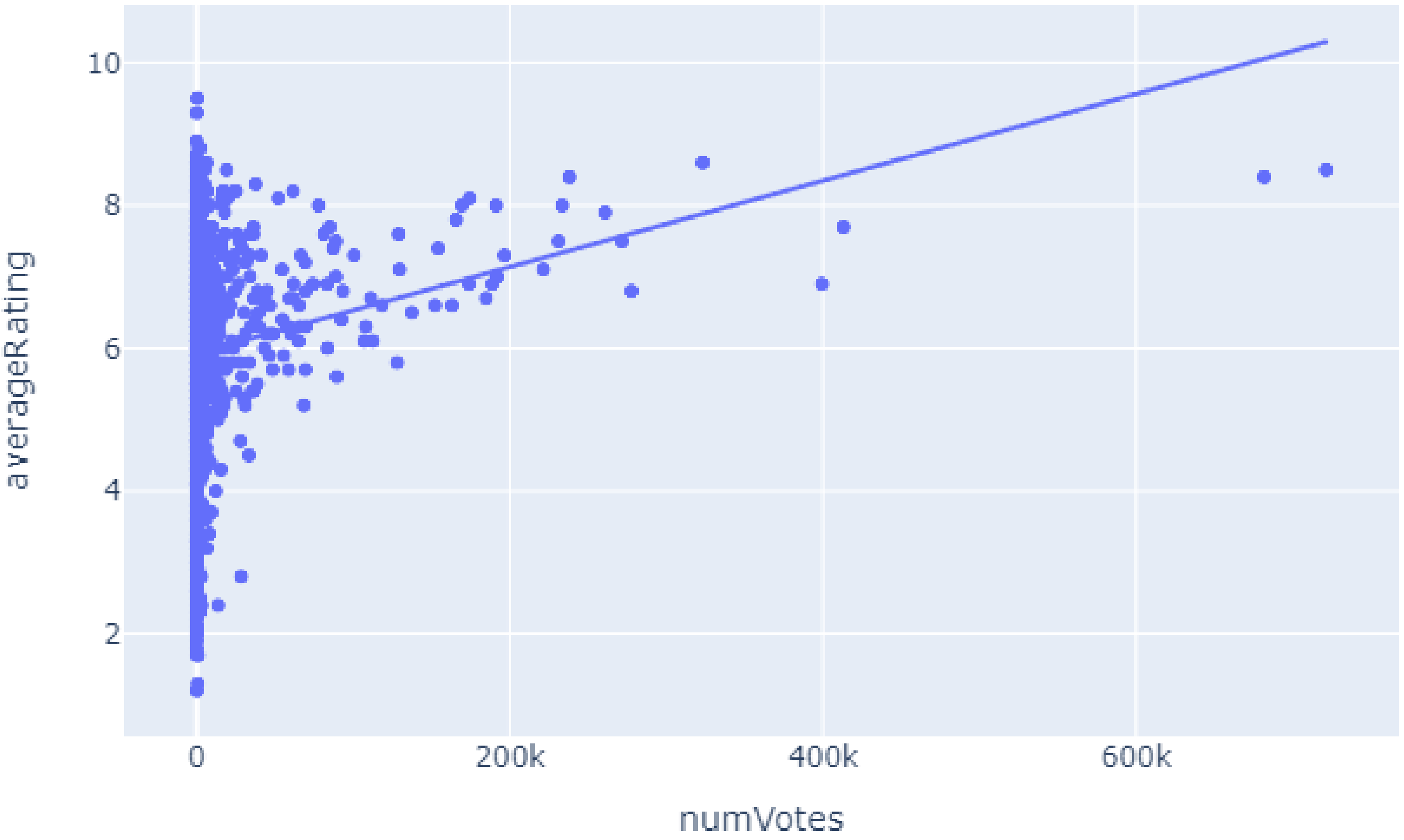
(correlation coefficient **0.76**)



More investment results in more payback

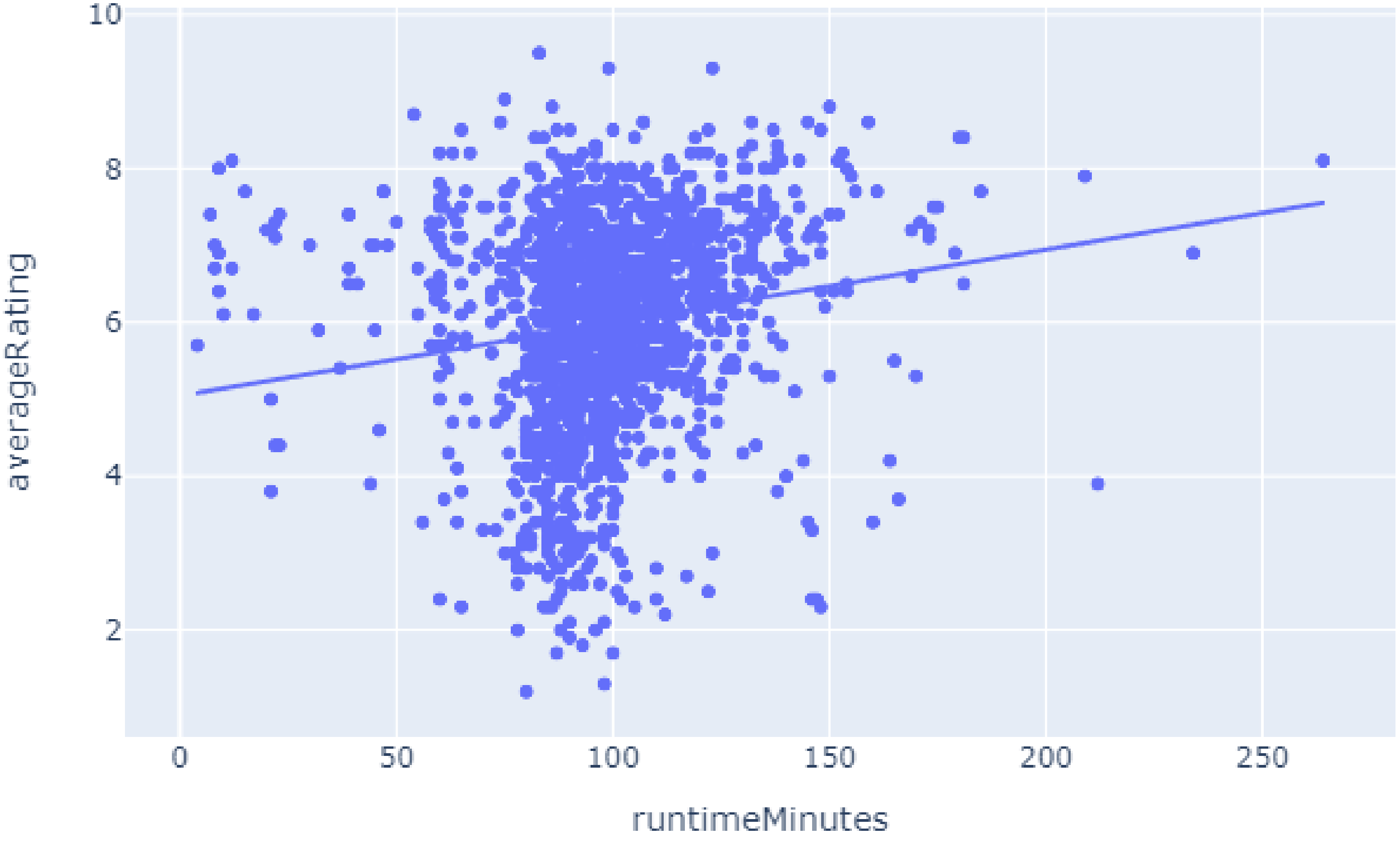
Main findings

No relationship found between average rating and number of votes
(correlation coefficient **0.17**)



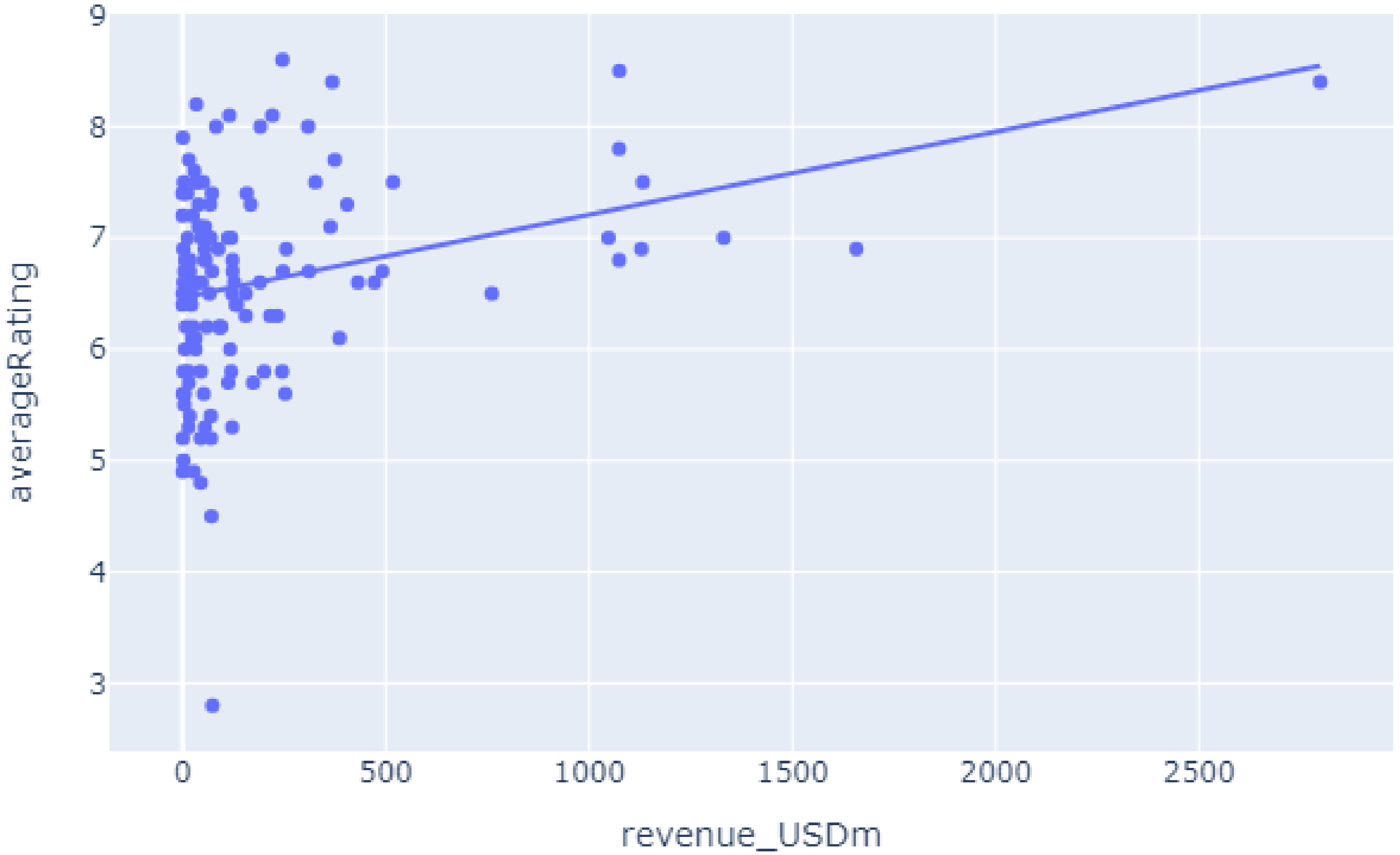
Main findings

No relationship found between average rating and duration of the movie
(correlation coefficient **0.16**)



Main findings

No relationship found between average rating and revenue of the movie
(correlation coefficient **0.3**)



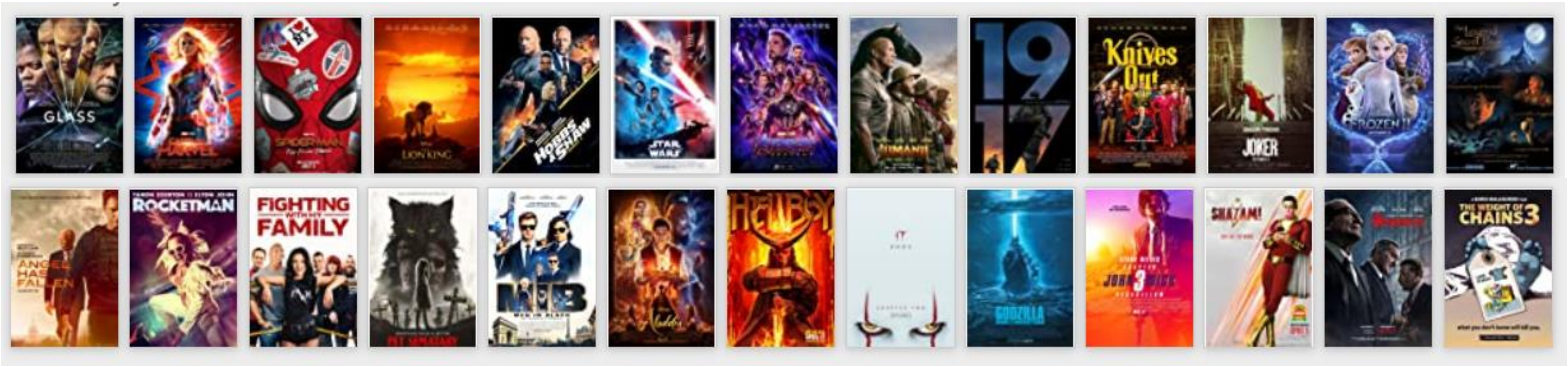
Higher rating doesn't necessarily correlate with higher profits for a movie

Conclusions

- The movie rating variable is not correlated with the duration of the movie, number of votes and revenue

- Technical/informative type of movies (documentary, Biography, History) are more likely to have higher ratings while movies in the cluster of action, horror, sci.fi are more likely to have lower ratings

- Strong positive linear relationship between revenue and budget meaning that more investment results in more payback



IRON HACK

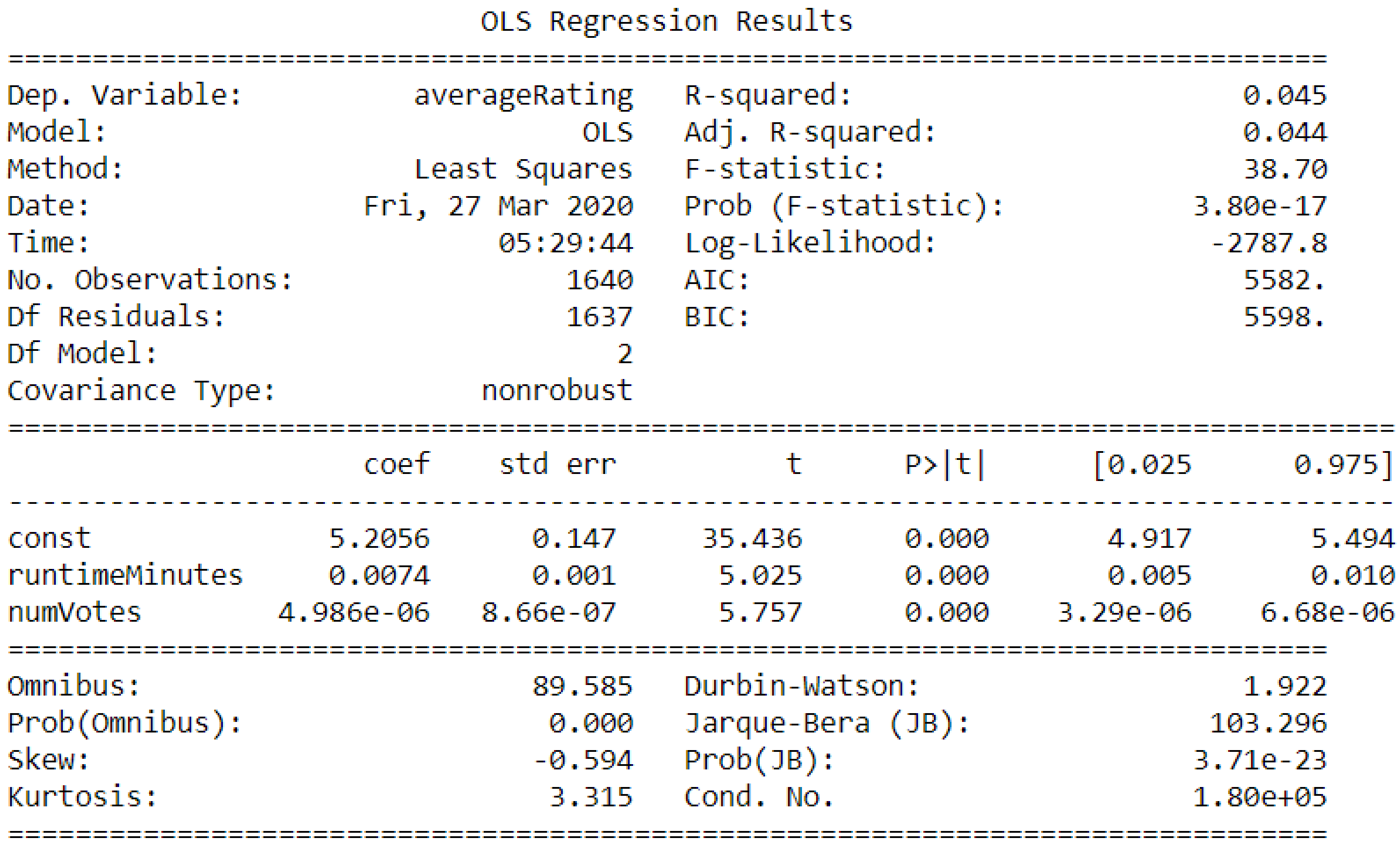
Thank you.

Mafalda Fragoso

Data Analytics Bootcamp
27/03/2020

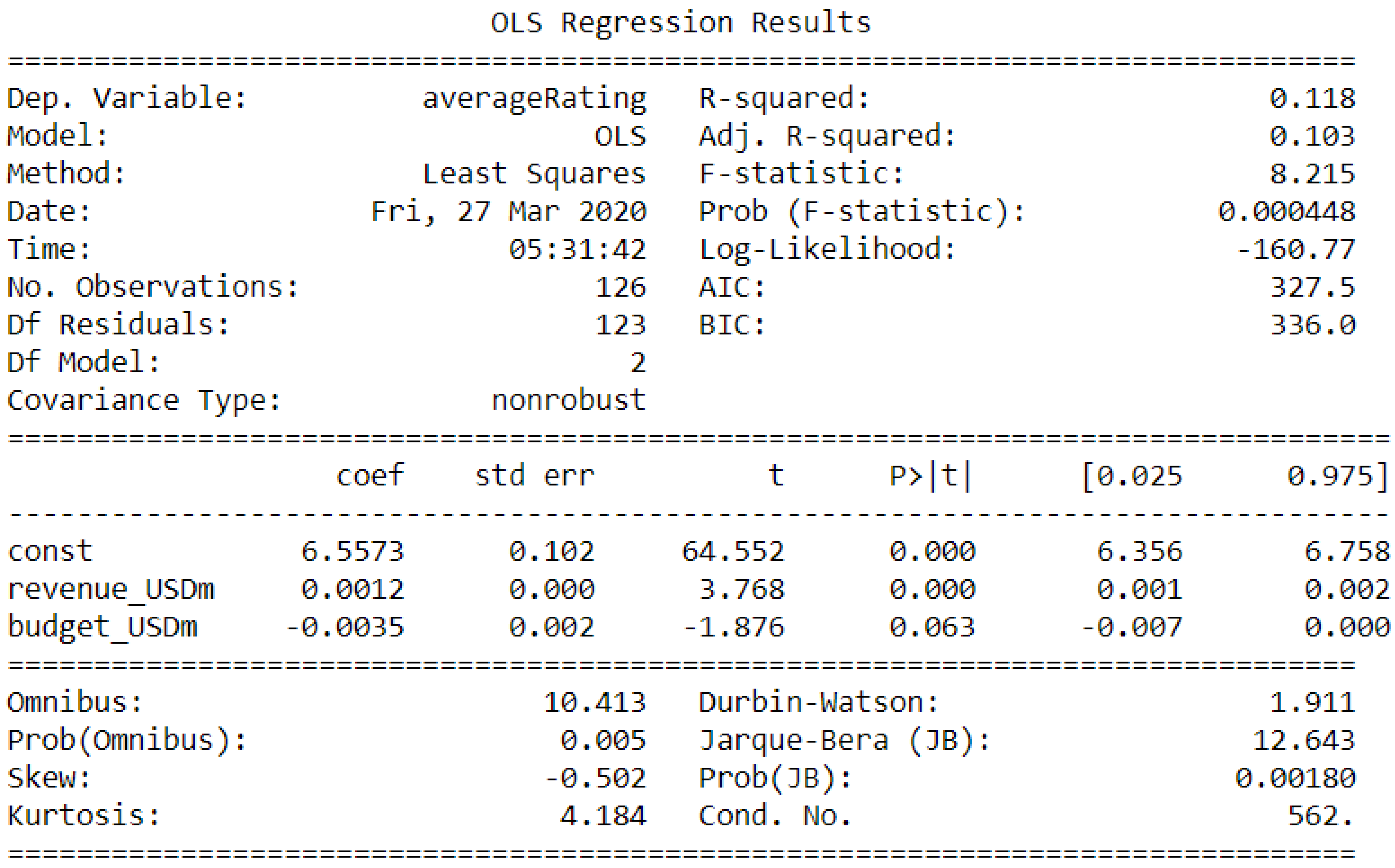
Appendix

OLS Regression Results



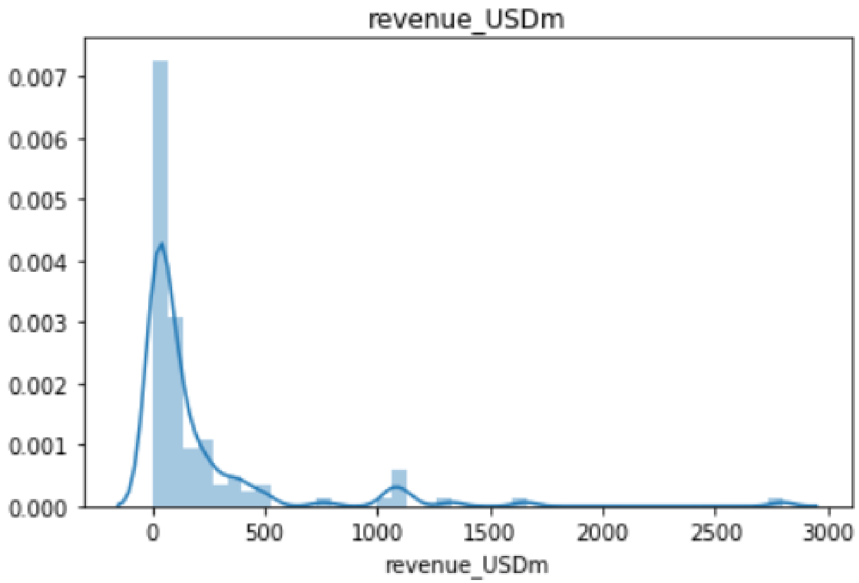
```
                            OLS Regression Results
==============================================================================
Dep. Variable:          averageRating   R-squared:                       0.045
Model:                            OLS   Adj. R-squared:                  0.044
Method:                 Least Squares   F-statistic:                     38.70
Date:                Fri, 27 Mar 2020   Prob (F-statistic):           3.80e-17
Time:                        05:29:44   Log-Likelihood:                -2787.8
No. Observations:                1640   AIC:                             5582.
Df Residuals:                    1637   BIC:                             5598.
Df Model:                           2
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const             5.2056      0.147     35.436      0.000       4.917       5.494
runtimeMinutes    0.0074      0.001      5.025      0.000       0.005       0.010
numVotes       4.986e-06   8.66e-07      5.757      0.000    3.29e-06    6.68e-06
==============================================================================
Omnibus:                       89.585   Durbin-Watson:                   1.922
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              103.296
Skew:                          -0.594   Prob(JB):                     3.71e-23
Kurtosis:                       3.315   Cond. No.                     1.80e+05
==============================================================================
```

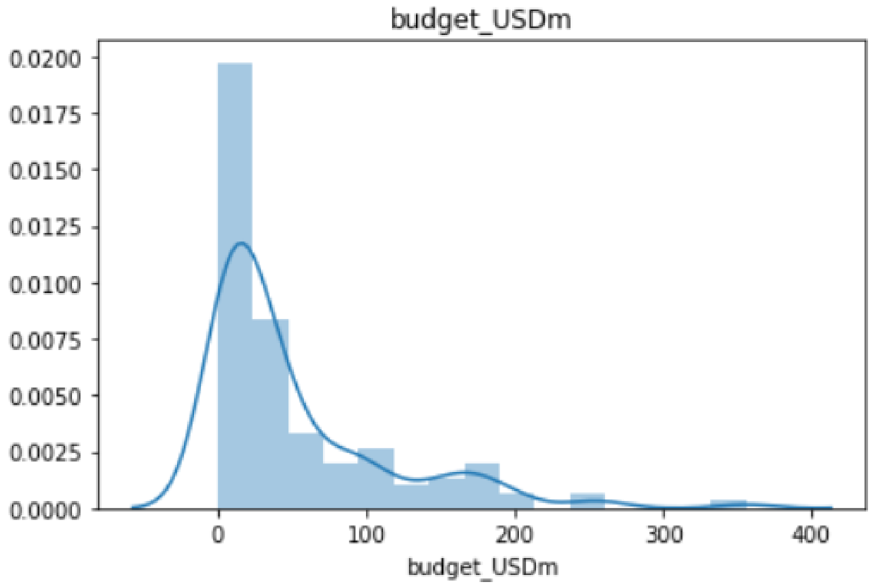
OLS Regression Results



```
                            OLS Regression Results
==============================================================================
Dep. Variable:          averageRating   R-squared:                       0.118
Model:                            OLS   Adj. R-squared:                  0.103
Method:                 Least Squares   F-statistic:                     8.215
Date:                Fri, 27 Mar 2020   Prob (F-statistic):           0.000448
Time:                        05:31:42   Log-Likelihood:                -160.77
No. Observations:                 126   AIC:                             327.5
Df Residuals:                     123   BIC:                             336.0
Df Model:                           2
Covariance Type:            nonrobust
================================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const           6.5573      0.102     64.552      0.000       6.356       6.758
revenue_USDm    0.0012      0.000      3.768      0.000       0.001       0.002
budget_USDm    -0.0035      0.002     -1.876      0.063      -0.007       0.000
==============================================================================
Omnibus:                       10.413   Durbin-Watson:                   1.911
Prob(Omnibus):                  0.005   Jarque-Bera (JB):               12.643
Skew:                          -0.502   Prob(JB):                      0.00180
Kurtosis:                       4.184   Cond. No.                         562.
==============================================================================
```

Revenue and budget distribution


revenue_USDm
revenue_USDm


budget_USDm
budget_USDm