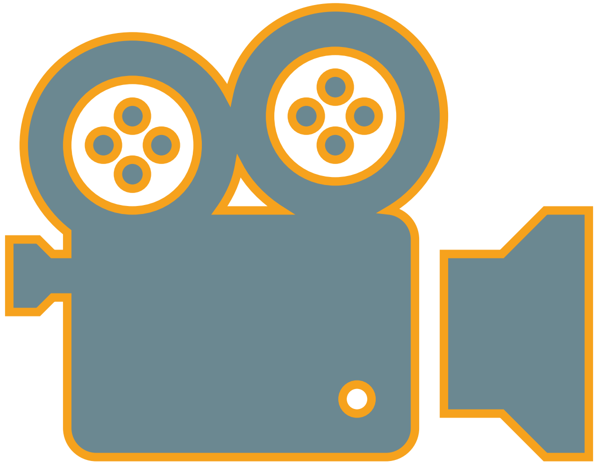
RECOMMENDATIONS FOR MOVIES PRODUCTION COMPANIES AND PREDICTING THE PROFITS OF THE MOVIES

Halah Banuqaytah || T5 bootcamp project

PROJECT GOALS:

1- Provide the movies production companies with recommendation to have more profits from movies streamed in cinema to keep up with the new direction streaming which is streaming platforms like Netflix.

2- Use machine learning algorithms to predict the profits of a movie before releasing it.



PERSONAL MOTIVATION:

This is my first ever data science project, so I thought about making the project about something I really love and enjoy which is movies and especially movies streamed only in cinema.

DATASET

DATASET INFORMATION

- **Source of dataset:**

  Kaggle website under the name 'Movie industry'

- **Number of columns and rows:**

  The dataset contains 7668 rows and 15 columns

- **The data types of columns:**

  9 objects ,5 float ,1 int.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7668 entries, 0 to 7667
Data columns (total 15 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   name      7668 non-null   object 
 1   rating    7591 non-null   object 
 2   genre     7668 non-null   object 
 3   year      7668 non-null   int64  
 4   released  7666 non-null   object 
 5   score     7665 non-null   float64
 6   votes     7665 non-null   float64
 7   director  7668 non-null   object 
 8   writer    7665 non-null   object 
 9   star      7667 non-null   object 
 10  country   7665 non-null   object 
 11  budget    5497 non-null   float64
 12  gross     7479 non-null   float64
 13  company   7651 non-null   object 
 14  runtime   7664 non-null   float64
dtypes: float64(5), int64(1), object(9)
memory usage: 898.7+ KB
```

Fig1: Dataset Information

DATA CLEANING

1- Null values: Dropping and Non dropping

2-Changing some data types

3- Rename columns

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 7409 entries, 0 to 7659
Data columns (total 15 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   movie_name         7409 non-null   object
 1   rating             7409 non-null   object
 2   genre              7409 non-null   object
 3   imdb_score         7409 non-null   float64
 4   imdb_votes         7409 non-null   int64
 5   director           7409 non-null   object
 6   writer             7409 non-null   object
 7   main_star          7409 non-null   object
 8   country            7409 non-null   object
 9   budget             7409 non-null   int64
 10  profits            7409 non-null   int64
 11  prod_company       7409 non-null   object
 12  runtime            7409 non-null   float64
 13  releasing_date     7409 non-null   datetime64[ns]
 14  releasing_country  7409 non-null   object
dtypes: datetime64[ns](1), float64(2), int64(3), object(9)
memory usage: 926.1+ KB
```

Fig2: Dataset Information after Cleaning

EXPLORATORY DATA ANALYSIS AND FINDINGS

MOVIES GENRE

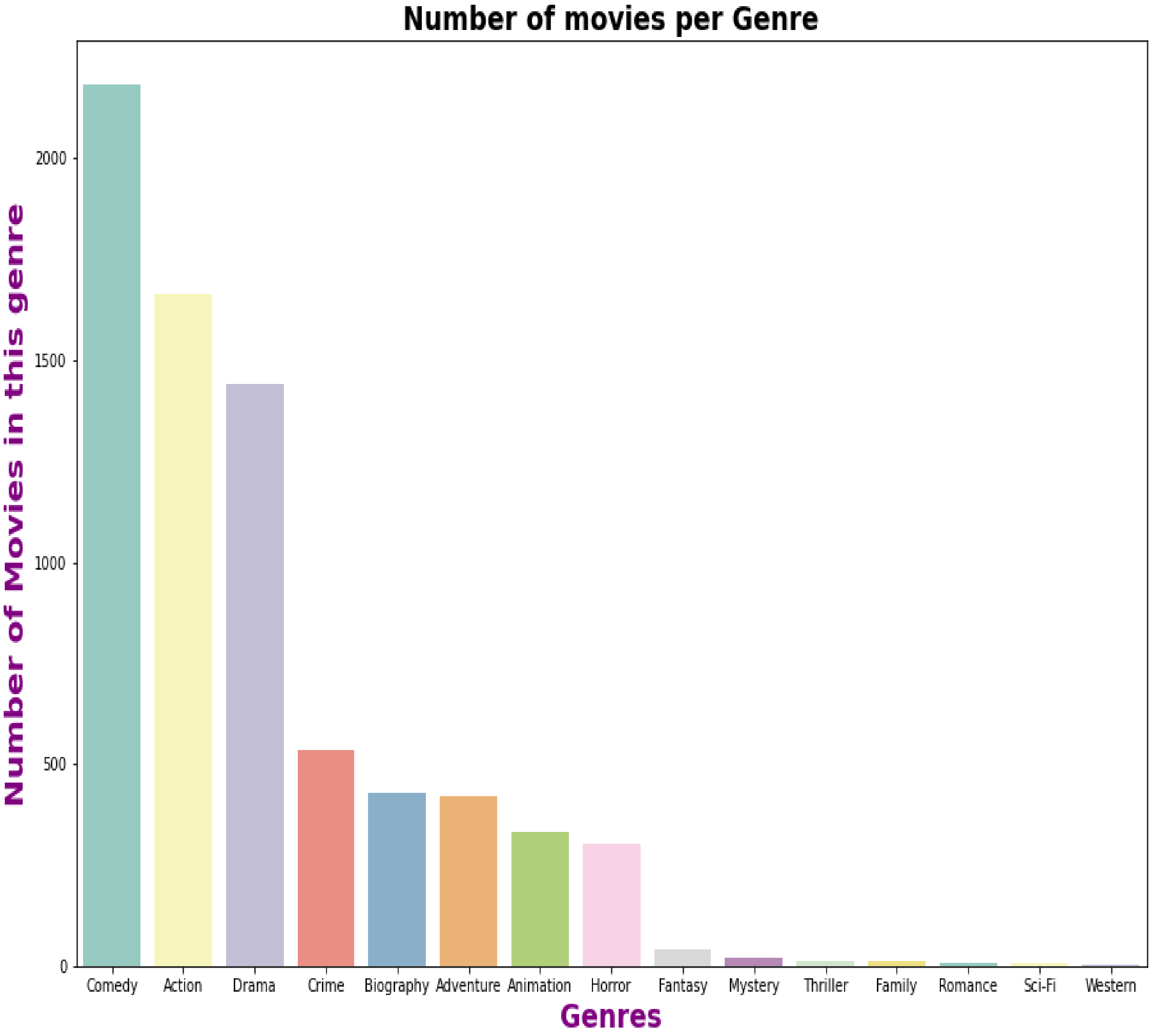
- **15 different genre which are:**

Drama, Adventure, Action, Comedy, Horror , Biography ,Crime ,Fantasy ,Family ,Animation , Romance, Western, Thriller, Sci-Fi ,Mystery

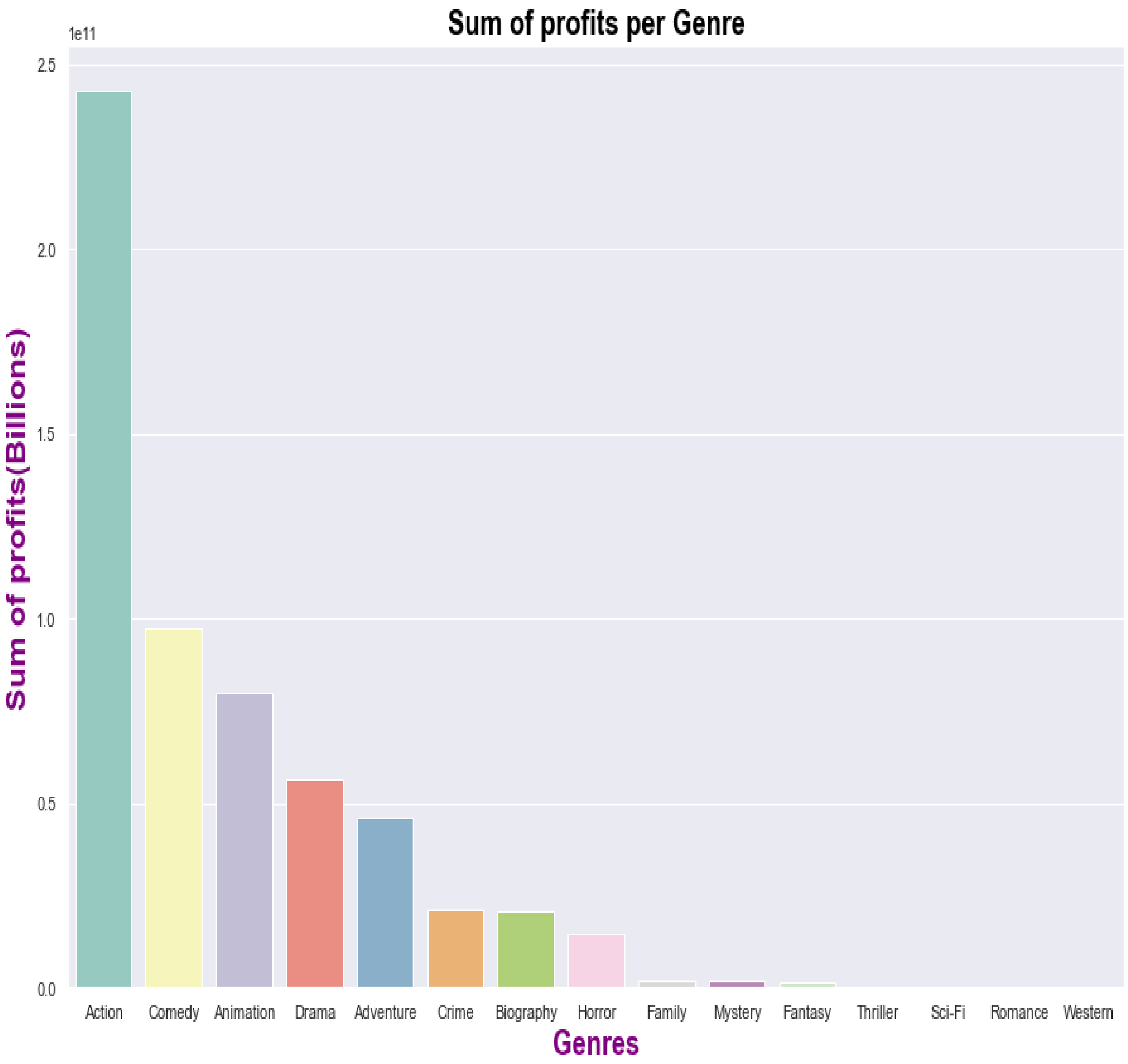
- **For this feature, I have the following assumptions:**

1. Most of the movies under the fantasy genre.
2. The action genre is the most profitable.

MOST OF THE MOVIES UNDER THE COMEDY GENRE.



THE ACTION GENRE IS THE MOST PROFITABLE.



MOVIES MAIN STARS

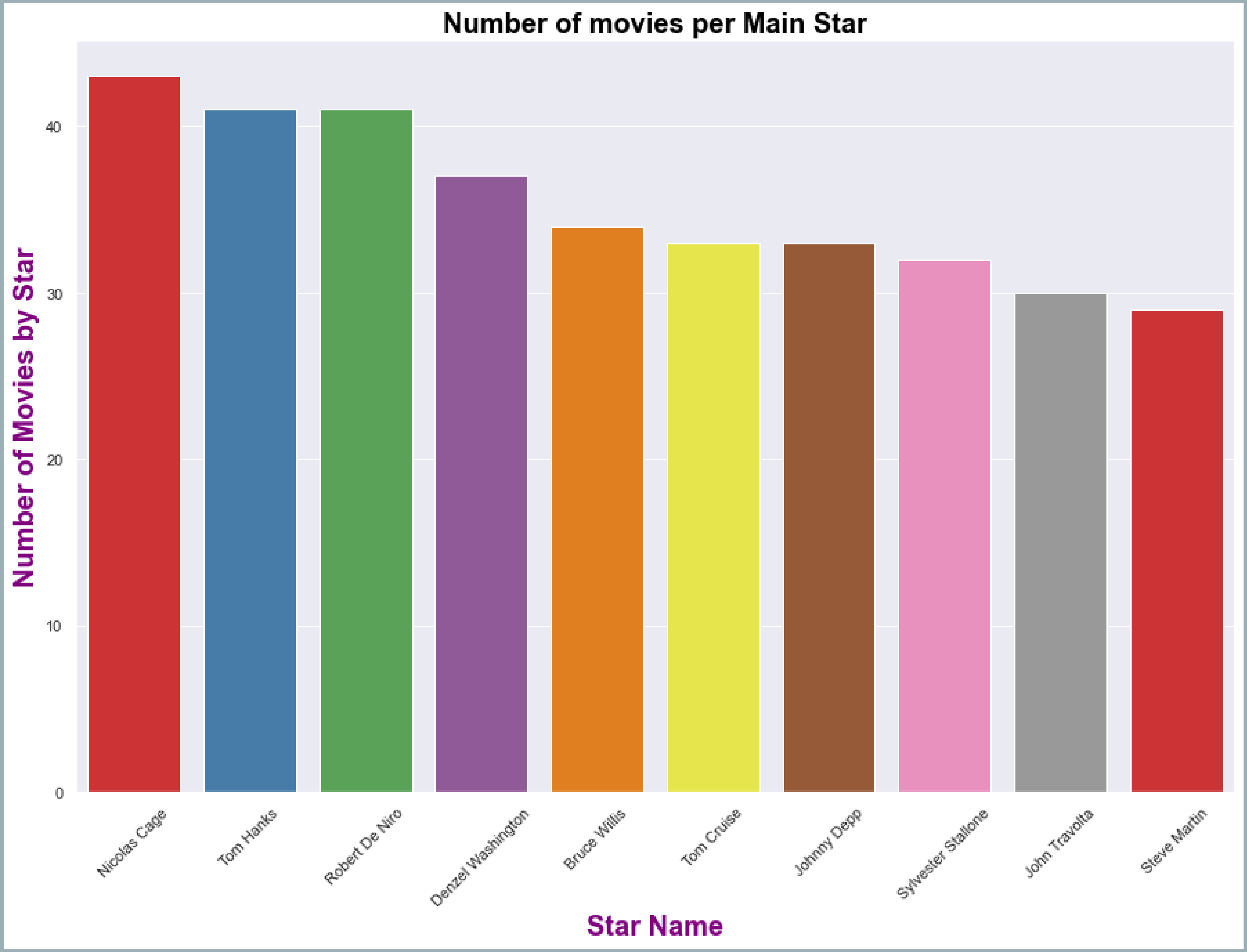
- **The dataset has:**

1. 2649 Main star

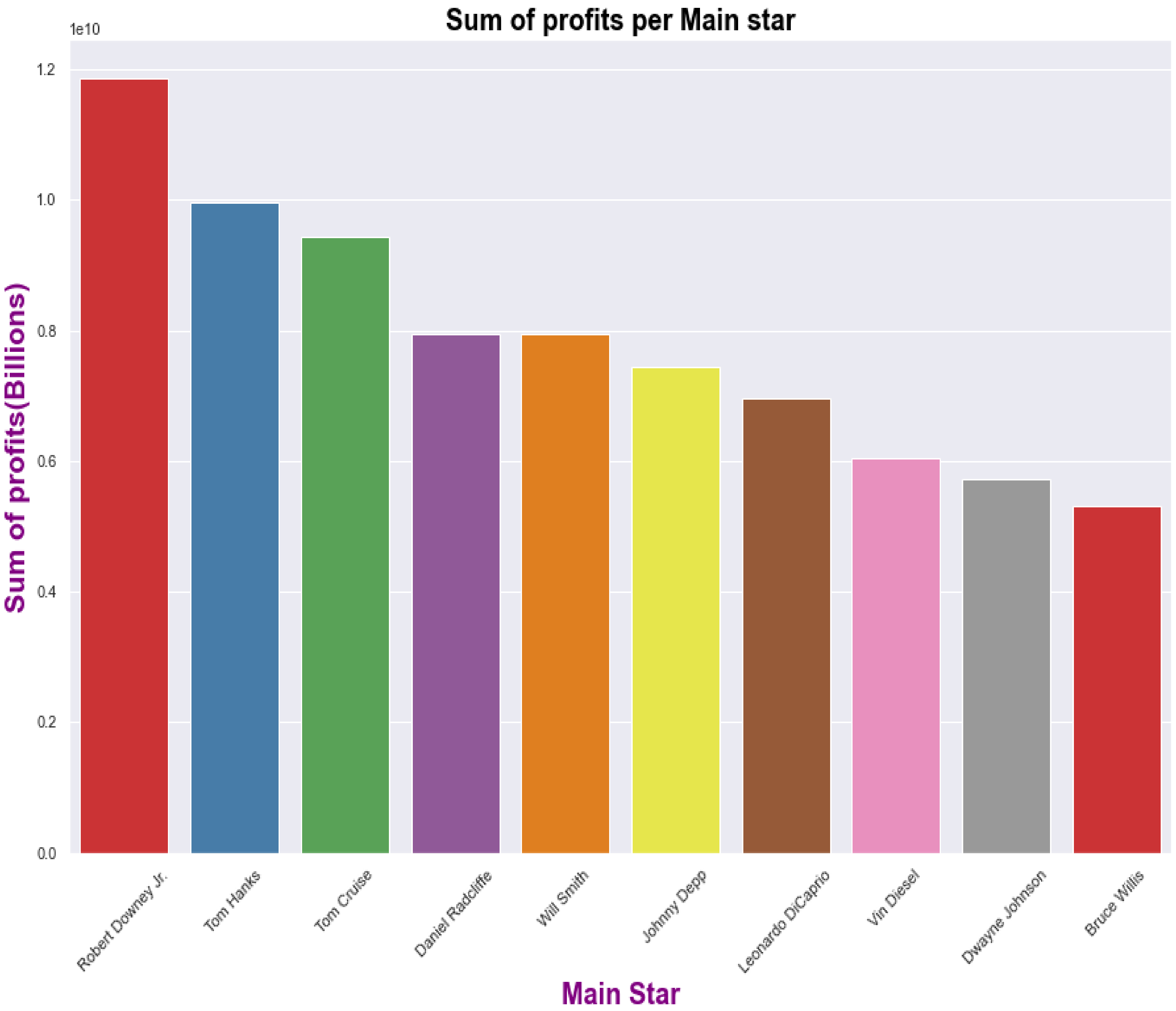
- **For this feature, I have the following assumption:**

Main Stars with more movies have the highest profits.

TOP10 MAIN STARS WITH MORE MOVIES



TOP10 MAIN STARS WITH HIGHEST PROFITS

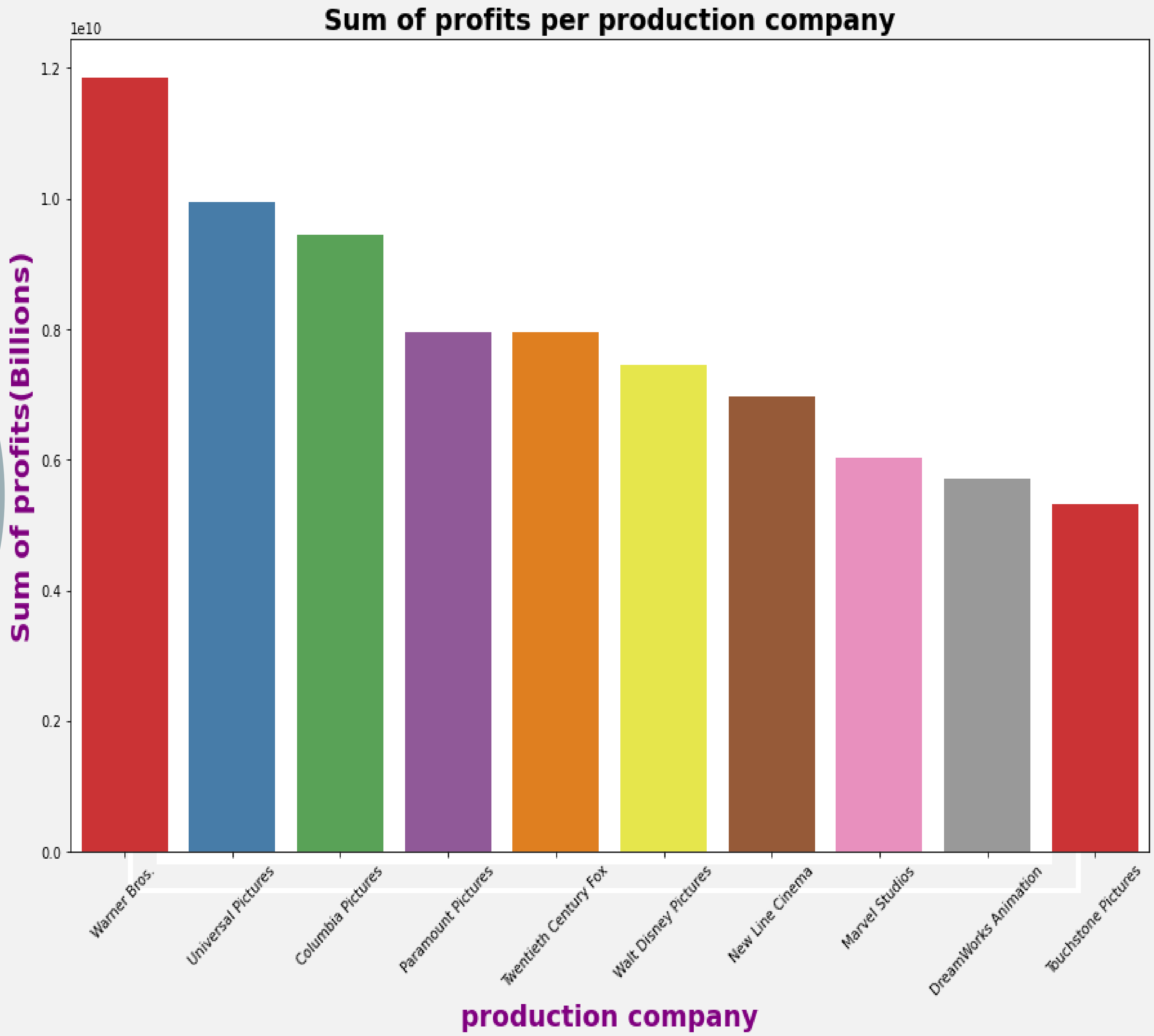


MOVIES PRODUCTION COMPANIES

- **2239 different production company.**
- **For this feature, I have the following assumption:**

Marvel production company is the most profitable.

TOP10 PRODUCTION COMPANIES WITH HIGHEST PROFITS



RELEASING DATE

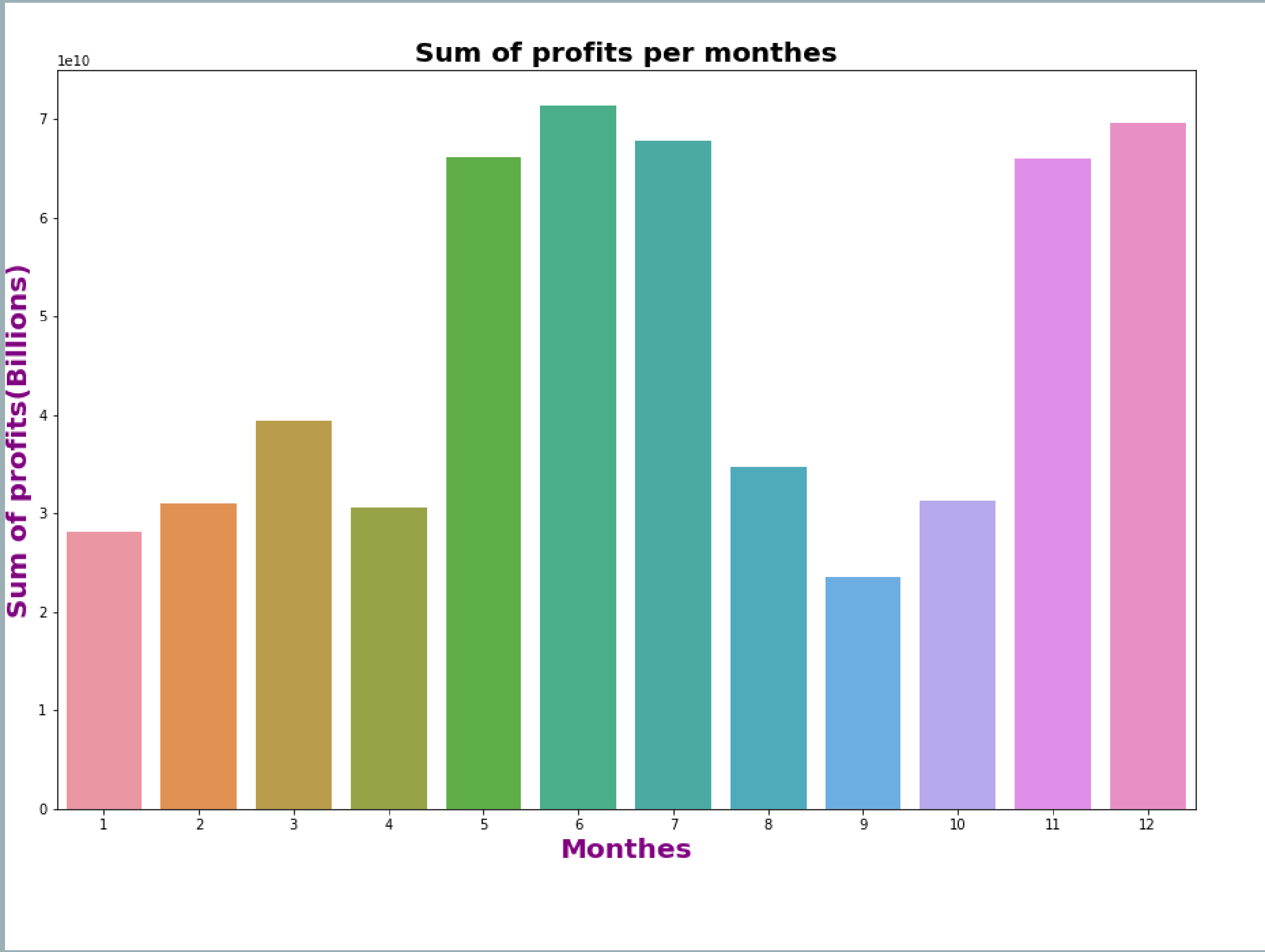
Month	Number of movies
January	584
February	584
March	666
April	626
May	558
June	547
July	570
August	716
September	634
October	724
November	628
December	572

Number of movies for each month in the past 40 years.

- **For this feature, I have the following assumption:**

Movies in December has the highest profits compared to other months.

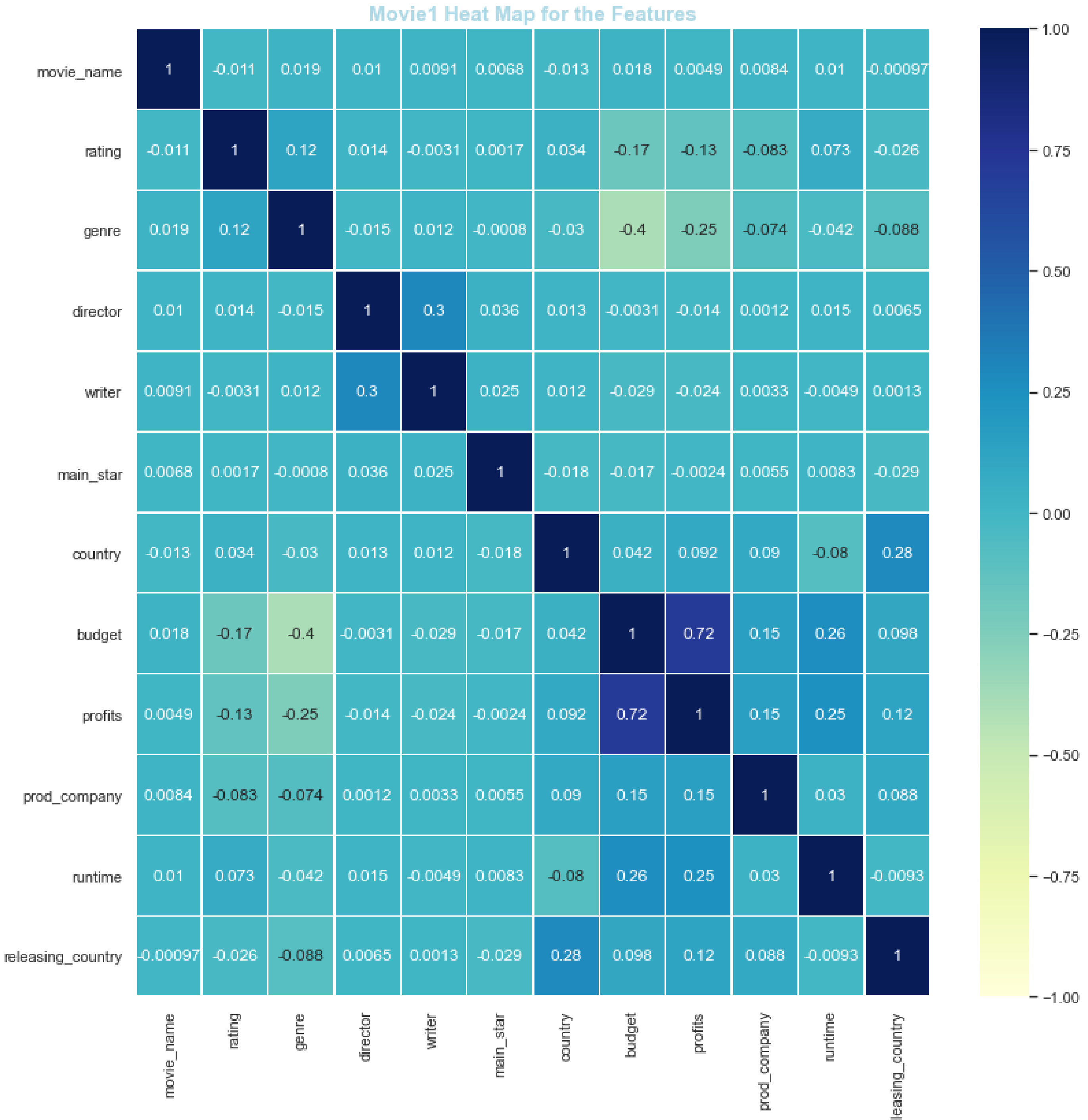
JUNE IS THE BEST MONTH TO RELEASE A MOVIE



MACHINE LEARNING ALGORITHM

FEATURE SELECTION

From the heatmap, we can see that budget is the most correlated feature with the target.



LINEAR REGRESSION MODEL

R squared value for Linear Regression Table.

	Dropping budget null values	Filling budget null values
Training set	0.54	0.51
Test set	0.56	0.51

RANDOM FOREST REGRESSOR

R squared value for Random Forest Regression Table.

	Dropping budget null values	Filling budget null values
Training set	0.63	0.67
Test set	0.60	0.55

CONCLUSION

- Movies production companies should focus more on action movies rather than any other genre.
- The best time to release a movie will be in June or December
- There is high correlation between the profits and the budget, so the production companies need to be generous with the budget to grantee the success of the movie.

FUTURE WORK

Improve the model using deep learning

Try different regression models to see if the accuracy will be higher

Adding more features from another dataset like movie's poster and plot.