REPORT

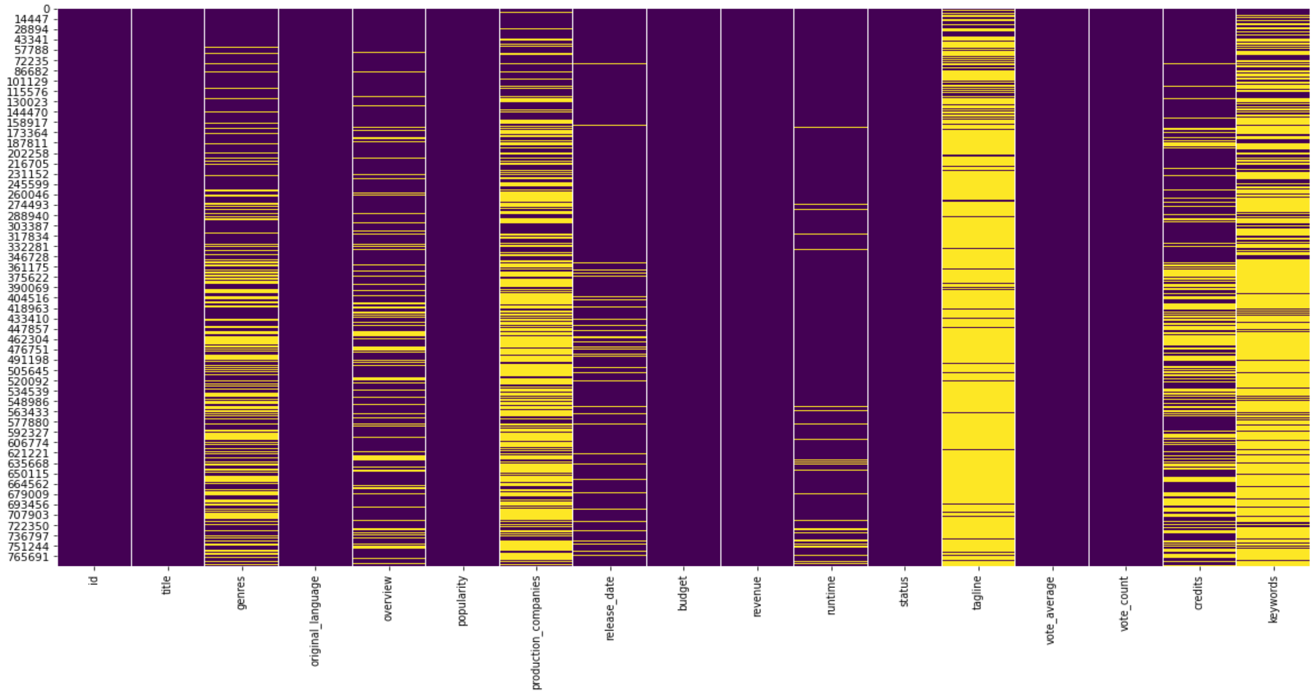
Exploratory Data Analysis on Movies Dataset

Submitted By:

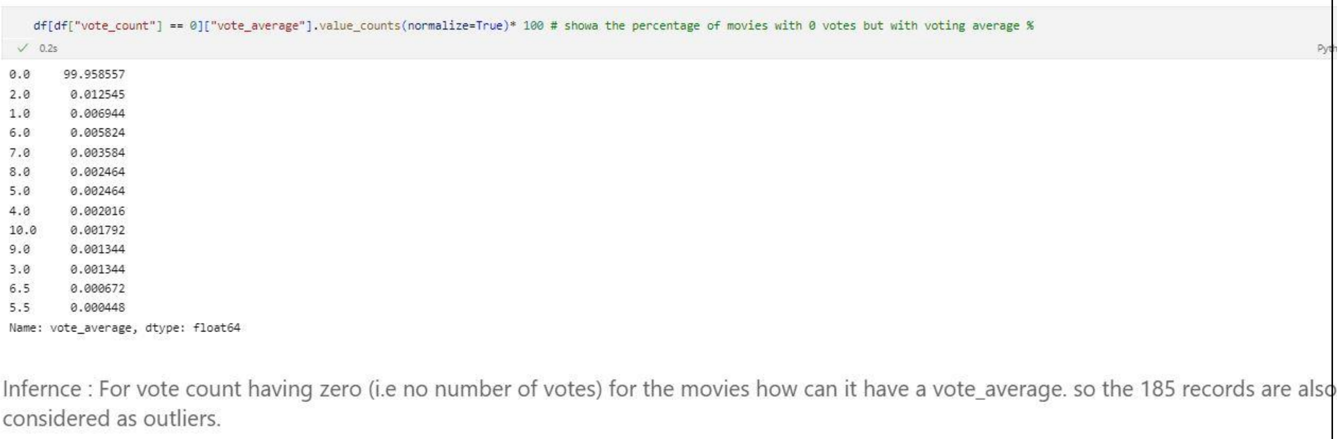
V.MURALIDHARAN

Exploratory data analysis:

In this exploratory data analysis, we find the missing values in the dataset. This is the visual representation of missing data in the dataset by each columns.



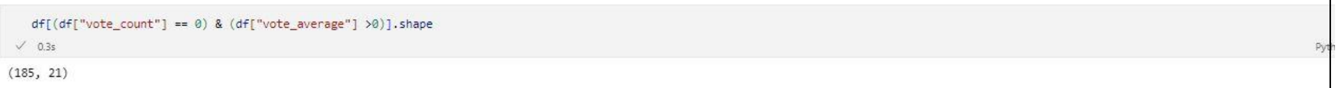
After this we get to find out the misclassified records.



```
df[df["vote_count"] == 0]["vote_average"].value_counts(normalize=True)* 100 # showa the percentage of movies with 0 votes but with voting average %
```

```
0.0     99.958557
2.0      0.012545
1.0      0.006944
6.0      0.005824
7.0      0.003584
8.0      0.002464
5.0      0.002464
4.0      0.002016
10.0     0.001792
9.0      0.001344
3.0      0.001344
6.5      0.000672
5.5      0.000448
Name: vote_average, dtype: float64
```

Infernce : For vote count having zero (i.e no number of votes) for the movies how can it have a vote_average. so the 185 records are also considered as outliers.



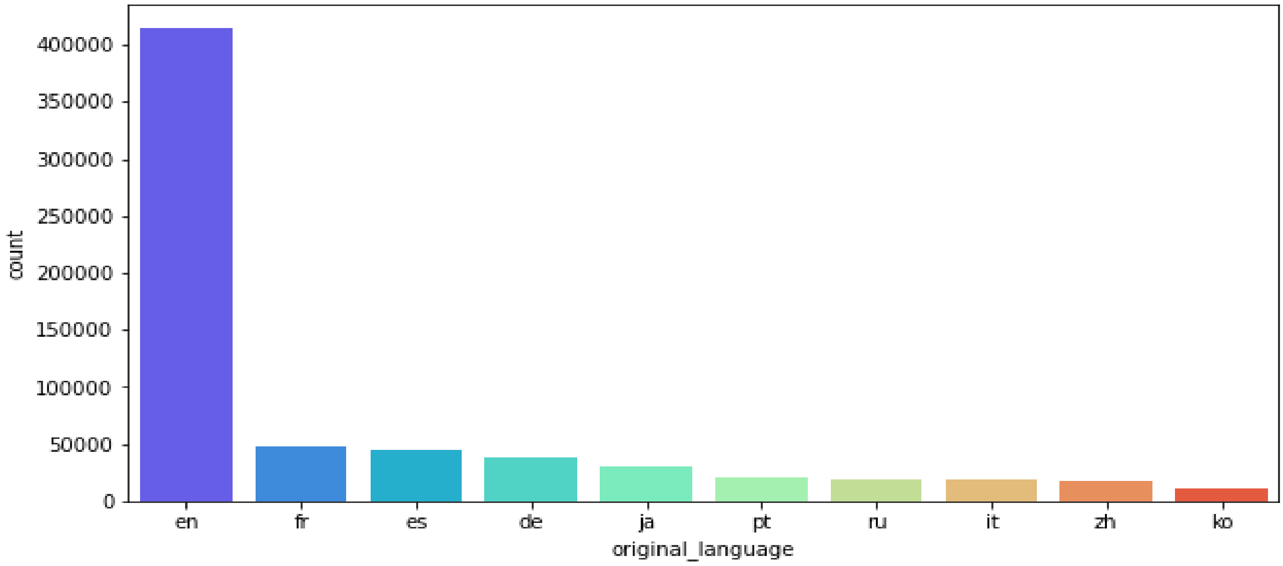
```
df[(df["vote_count"] == 0) & (df["vote_average"] >0)].shape
```

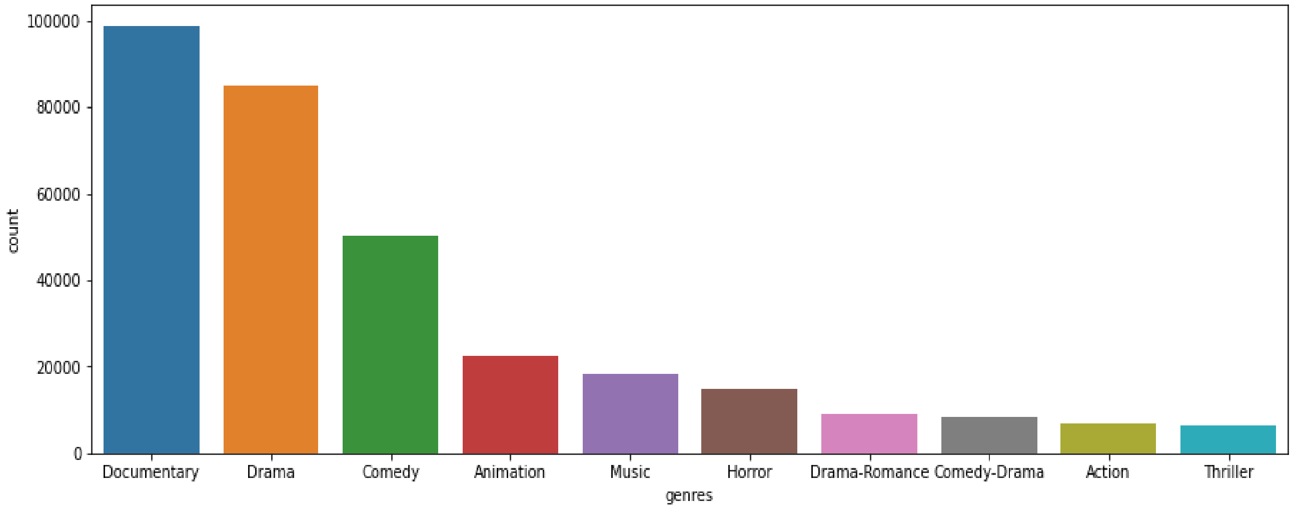
```
(185, 21)
```

Here the inference is for vote count having zero (i.e no number of votes) for the movies how can it have a vote average. So the 185 records are also considered as outliers.

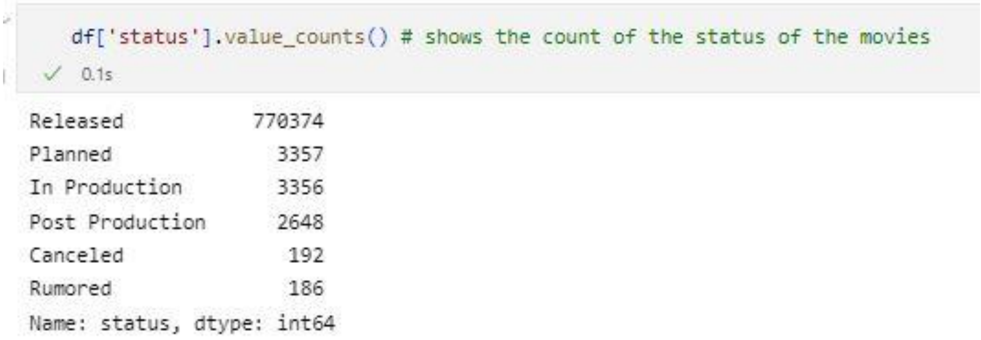
Uni-variate analysis:

The below graph shows that count of English (en) movies is very much higher when compared to other language movies.





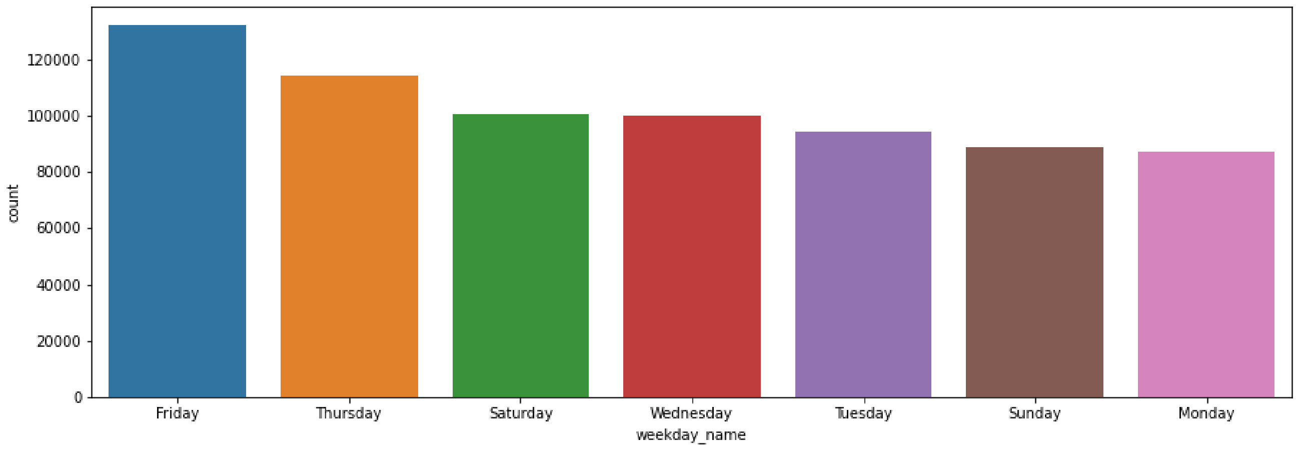
The above graph shows that the top 10 genres of all the movies in which documentary, drama, comedy are the top 3 genres of movies.



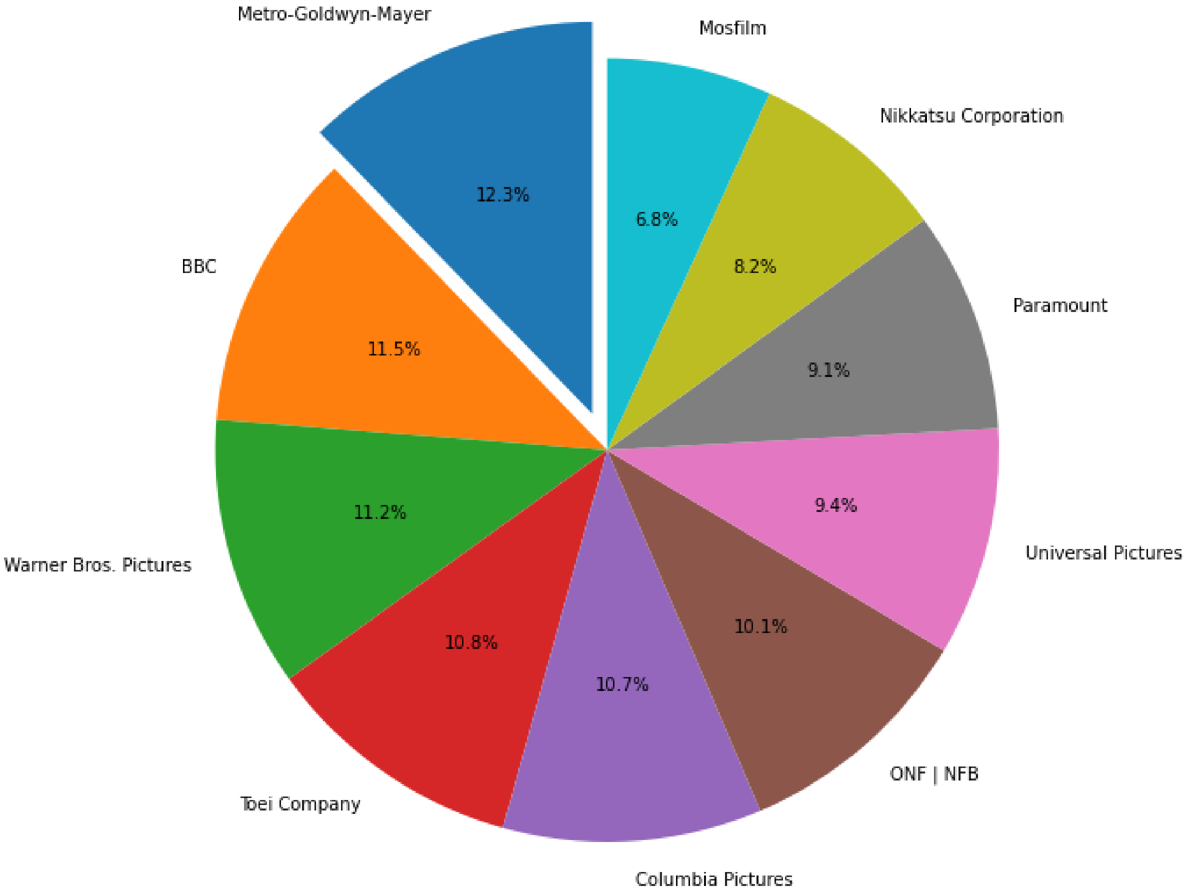
```
df['status'].value_counts() # shows the count of the status of the movies
✓ 0.1s

Released           770374
Planned              3357
In Production        3356
Post Production      2648
Canceled              192
Rumored               186
Name: status, dtype: int64
```

The above result shows the total count values of the status column in the dataset.

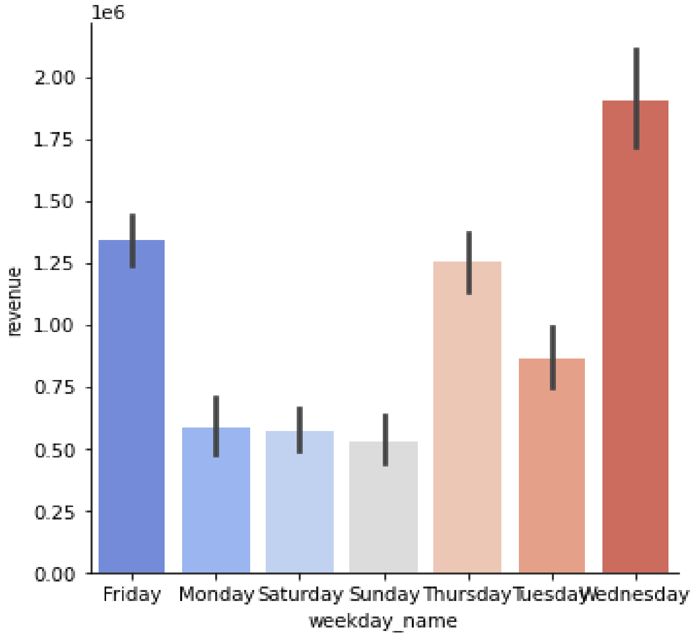


This graph provides the information of on which weekdays the movies counts were higher.

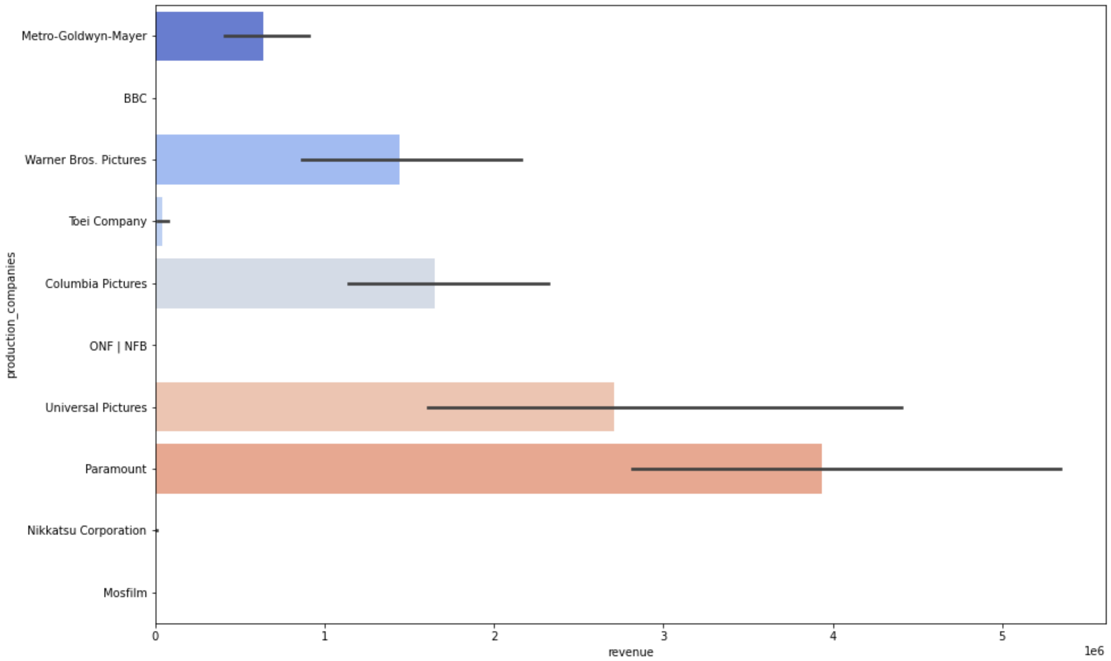


The above pie chart shows that the top 10 production company made how much of movies in percentage.

Bivariate analysis:

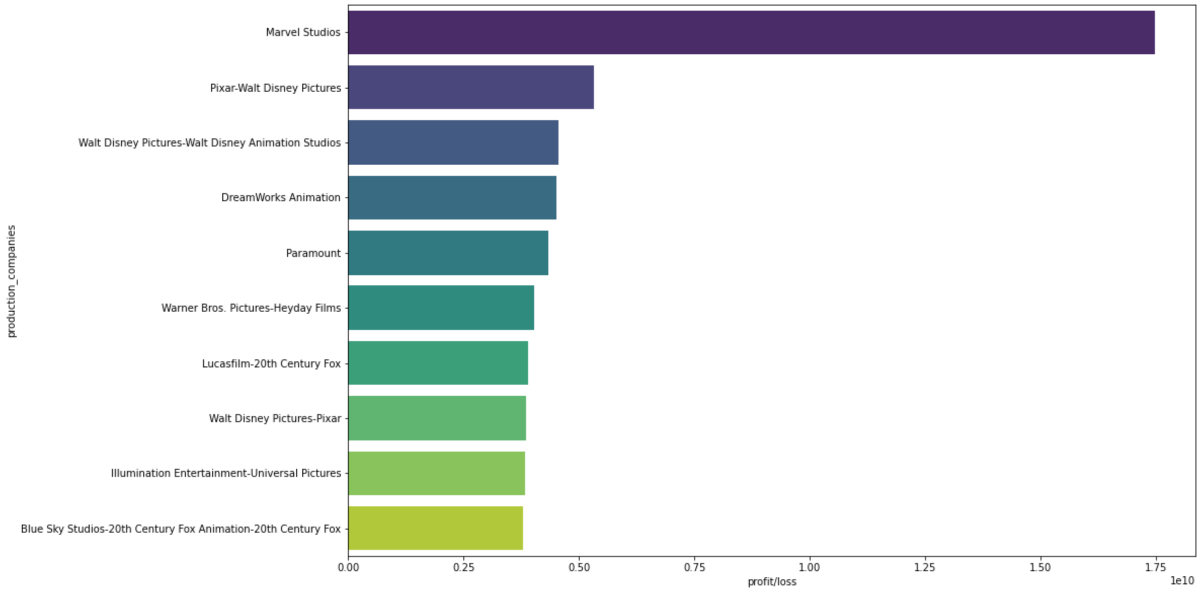


The above graph shows that on which week days the total revenue of the movies were higher.

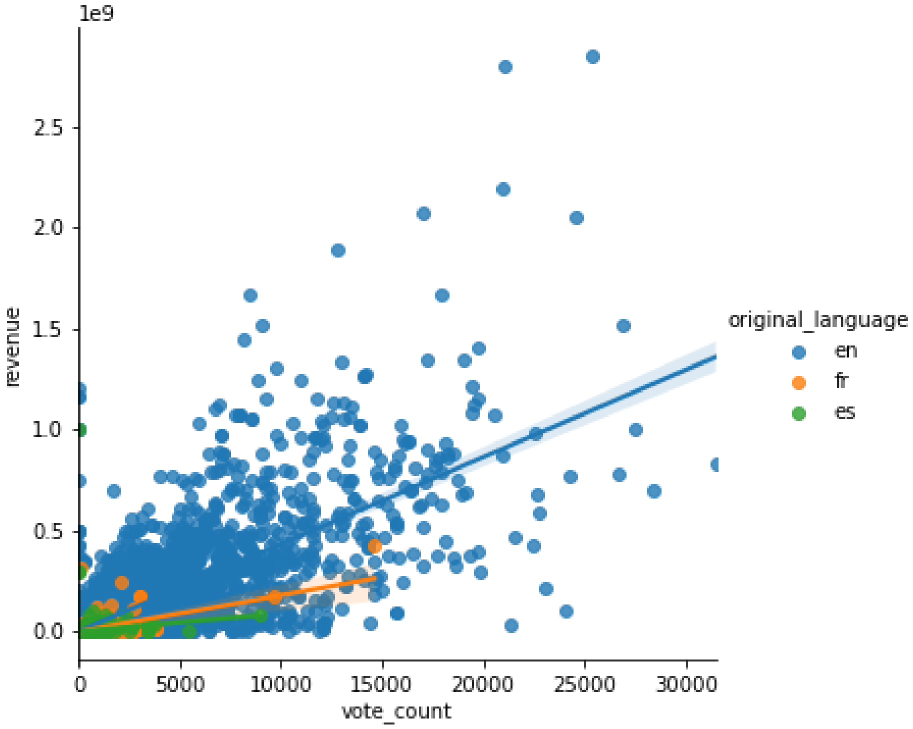


The below graph shows that the revenue made by the top 10 production companies. This shows that the top 7th production company made bigger profit when compared to other top production companies.

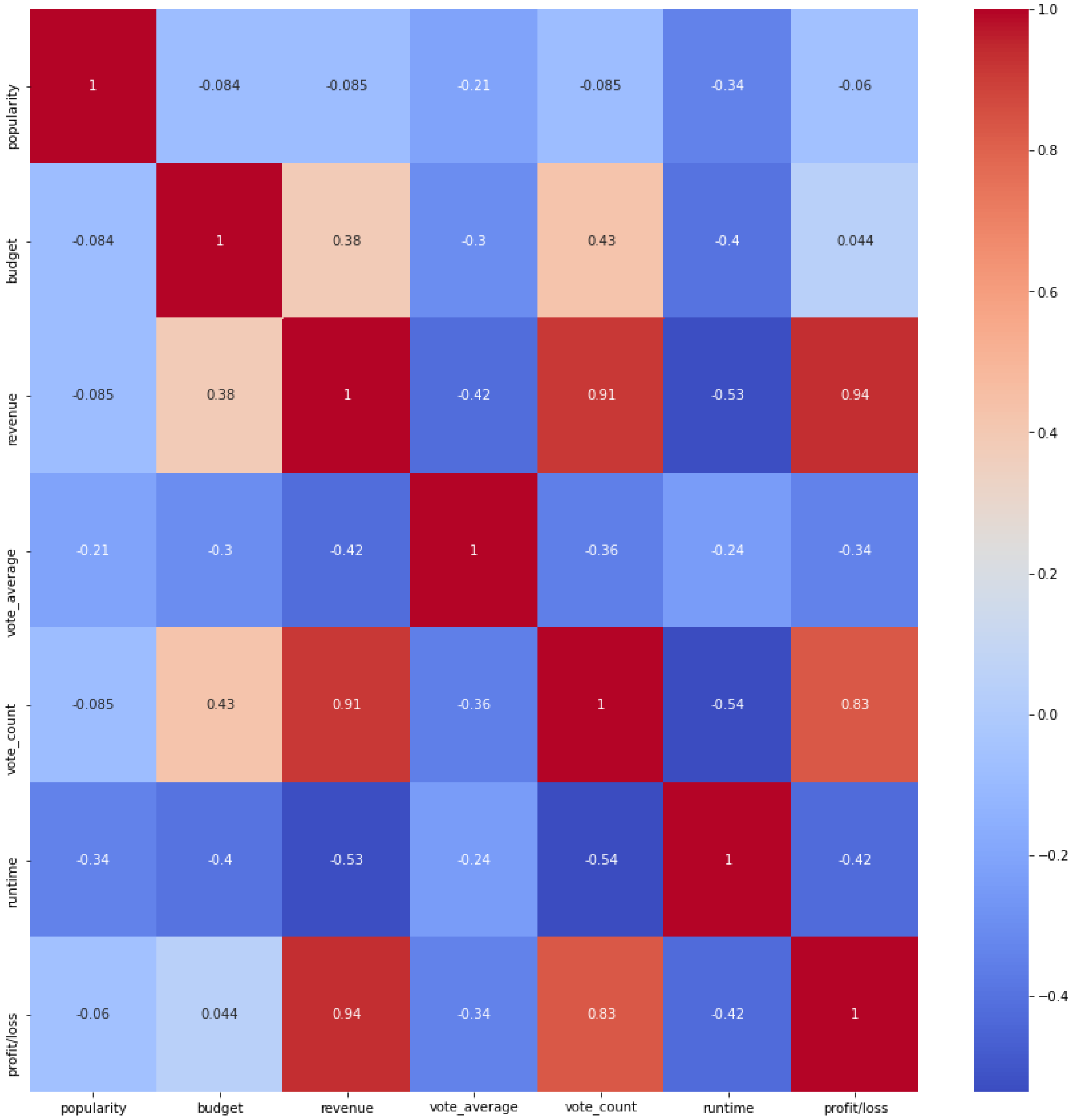
The below graph shows that the marvel studios production company made high profit compared to other production companies



Multivariate analysis:



The above graph shows an increasing trend over the vote count when compared to revenue on the top 3 languages of the movies.



	popularity	budget	revenue	vote_average	vote_count	runtime	profit/loss
popularity	1	-0.084	-0.085	-0.21	-0.085	-0.34	-0.06
budget	-0.084	1	0.38	-0.3	0.43	-0.4	0.044
revenue	-0.085	0.38	1	-0.42	0.91	-0.53	0.94
vote_average	-0.21	-0.3	-0.42	1	-0.36	-0.24	-0.34
vote_count	-0.085	0.43	0.91	-0.36	1	-0.54	0.83
runtime	-0.34	-0.4	-0.53	-0.24	-0.54	1	-0.42
profit/loss	-0.06	0.044	0.94	-0.34	0.83	-0.42	1

The above map shows the correlation between numeric variables of the dataset. This shows that vote count and revenue has high correlation.