BLL514E: WHAT MAKES A MOVIE SUCCESSFUL

AHMET DROBI

OUTLINE:

Introduction	• Introduce the subject of the study
Collecting Data	• Dataset description
Statistical Analysis	• Study of the overall dataset • Testing Hypotheses
ML Models	• ML predictions • Accuracy of ML model
Summary	• Final Findings

Introduction:

- This study was based on movies data-set which includes 8,000 movies meta-data.
- The original data-set had missing values, in order to do feature engineering, we had to clean the data-set.
- The second step was statistical analysis the data so we can test 9 hypothesis with the aim of finding what effects the revenue of a movie.
- The last step was creating a machine model to predict the revenue based on multiple factors.

Collecting Dataset

- To create the dataset we collected the movie data from 2 sources:
  - The Movie Database API
  - OMDb API
- From The Movie Database API we used an endpoint that allowed us to receive a random list of 10,000 movies from the top 100,000 by revenue.
- Then we used OMDb API to fill the meta-data of those movies.
- The resulting dataset columns were as follows:

  1) Title 2) Budget 3) Revenue 4)Run-time
  5) Motion Picture Association of America rating 6) Genre
  7) Top100 Actors in Movie 8) Language 9) Country
  10) Meta-score 11) IMDb Rating 12) IMDb-Votes
  13) Production-Company 14) Year of Release

Dataset before any alterations:

Title	Budget	Revenue	Date	Runtime	Year	Rated	Genre	Director	Writer	Actors	Language	Country	Awards	Metascore	imdbRatir	imdbVote	Production
Avengers:	3.56E+08	2.8E+09	########	181	2019	PG-13	Action, Ac	Anthony F	Christoph	Robert Do	English, Ja	USA	N/A	78	8.5	589,503	Marvel Studios
Avatar	2.37E+08	2.79E+09	########	162	2009	PG-13	Action, Ac	James Car	James Car	Sam Wort	English, Sp	USA	Won 3 Osc	83	7.8	1,064,516	20th Century Fox
Star Wars:	2.45E+08	2.07E+09	########	136	2015	N/A	News	N/A	Scott Bron	J.J. Abram	English	USA	N/A	N/A	6	8	N/A
Avengers:	3E+08	2.05E+09	########	149	2018	PG-13	Action, Ac	Anthony F	Christoph	Robert Do	English	USA	N/A	68	8.5	717,106	Walt Disney Pictures
Titanic	2E+08	1.85E+09	########	194	1997	PG-13	Drama, Ro	James Car	James Car	Leonardo	English, Sv	USA	Won 11 Os	75	7.8	978,848	Paramount Pictures
Jurassic W	1.5E+08	1.67E+09	6/6/2015	124	2015	PG-13	Action, Ac	Colin Trev	Rick Jaffa	Chris Pratt	English	USA	14 wins &	59	7	549,286	Universal Pictures
The Lion K	2.6E+08	1.65E+09	########	118	1994	G	Animatior	Roger Alle	Irene Mec	Rowan Atl	English, Sv	USA	Won 2 Osc	88	8.5	869,323	Buena Vista
The Aveng	2.2E+08	1.52E+09	########	143	2012	PG-13	Action, Ac	Joss Whec	Joss Whec	Robert Do	English, Ru	USA	Nominate	69	8	1,204,899	Walt Disney Pictures
Furious 7	1.9E+08	1.51E+09	4/1/2015	137	2015	PG-13	Action, Ac	James Wa	Chris Mor	Vin Diesel	English, Th	USA, Chin	Nominate	67	7.2	341,279	Universal Pictures
Avengers:	2.8E+08	1.41E+09	########	141	2015	PG-13	Action, Ac	Joss Whec	Joss Whec	Robert Do	English, Ko	USA	7 wins & 4	66	7.3	685,405	Walt Disney Pictures
Black Pant	2E+08	1.35E+09	########	134	2018	PG-13	Action, Ac	Ryan Coog	Ryan Coog	Chadwick	English, Sv	USA	14 nomina	88	7.3	545,565	Marvel Studios
Harry Pott	1.25E+08	1.34E+09	7/7/2011	130	2011	PG-13	Adventure	David Yate	Steve Klov	Ralph Fier	English	USA, UK	Nominate	87	8.1	704,107	Warner Bros. Picture
Star Wars:	2E+08	1.33E+09	########	152	2008	N/A	Short, Act	N/A	N/A	Taylor Cla	English	USA	N/A	N/A	8.3	87	N/A
Jurassic W	1.7E+08	1.3E+09	6/6/2018	129	2018	PG-13	Action, Ac	J.A. Bayor	Derek Cor	Chris Pratt	English, Ru	USA	N/A	51	6.2	231,869	Universal Pictures

Cleaning data-set:

The original data-set contained 10,000 movies, after filtering the data-set It reduced to 8,000 movies, because of the problems in the following features:

1-) The Year column : the OMDB API provided a lot of missing values , whereas the data from “The Movie Database” API was almost complete, so we extracted only the year and then renamed it into “Year of Release”

2-) The Language column: we changed the missing values of this column to English.

Cleaning data-set: (cont'd)

3-) The Country column: the missing value in this columns were assigned the value USA

4-) The IMDB Rating and Votes columns: we assigned Zero to the missing values, so the missing values will not effect data-set

5-) The Actors column: The main goal of this column to check how many famous actors/actress in these movies, so we generated a list that contains 100 famous actor/actress, and compered each movie with this list.

Cleaning data-set: (cont'd)

6) The Production column: The missing values in this columns has been changed from null to others, and there were some production companies that only produced a few movies we also added them to others

7) The Run-time column: we assigned the mean to the missing value in this column.

8) The Genre column: we only took the first value in this column to be able to use label encoding on it.

Cleaning data-set: (cont'd)

9) The Awards column: Had a lot of null/missing values and also unusable values, so we found a better alternative in a data-set called "academy-awards" from "Kaggle" and cross referenced our movies data-set with it and counted how many awards each movie have and added it to a column called "Number of Awards" in our data-set.

10) The Released Outside USA column: we add this column to indicate if the movies were released inside USA only or not.

Collecting Dataset:

- Some Notes on the dataset:
  - Top 100 Actors in Movie column holds the number of top 100 actors (according to IMDb) that acted in the movie.
  - Motion Picture Association of America rating column is the age rating of the movie. We used this matric because it’s the most used metric in the world.

Filtered movies data-set

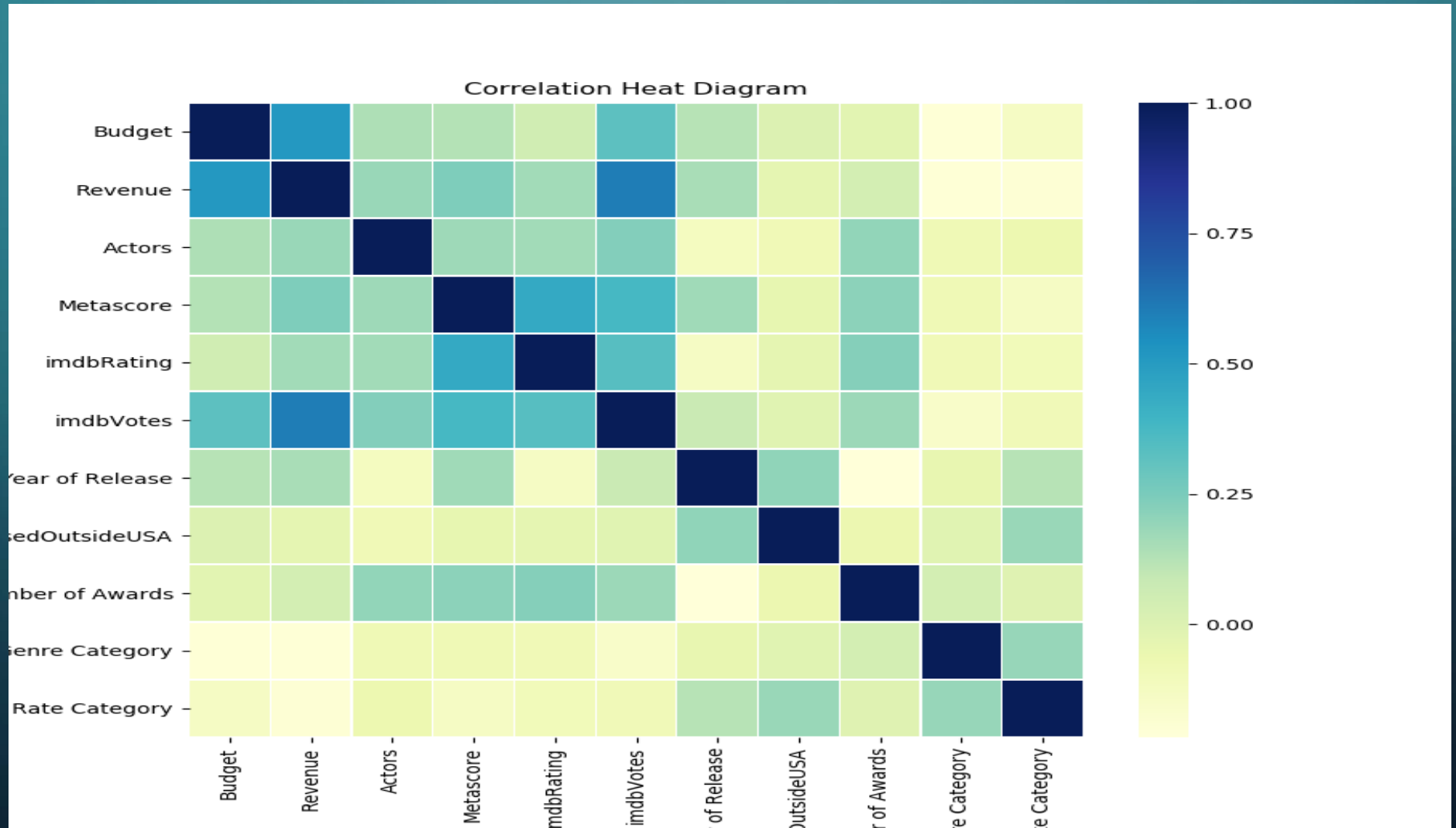
Title	Budget	Revenue	Runtime	Rated	Genre	Actors	Language	Country	Metascore	imdbRating	imdbVotes	Production	Year of Release	ReleasedOutsic	Number of Awards
Avengers:	356000000	2797800564	181	PG-13	Action	3	English	USA	78	8.5	589503	Marvel Studios	2019	0	0
Avatar	237000000	2787965087	162	PG-13	Action	0	English	USA	83	7.8	1064516	20th Century Fox	2009	0	0
Star Wars:	245000000	2068223624	136	Unrated	News	0	English	USA	0	6	8	Others	2015	0	0
Avengers:	300000000	2046239637	149	PG-13	Action	3	English	USA	68	8.5	717106	Walt Disney Pictu	2018	0	0
Titanic	200000000	1845034188	194	PG-13	Drama	2	English	USA	75	7.8	978848	Paramount Pictur	1997	0	2
Jurassic W	150000000	1671713208	124	PG-13	Action	0	English	USA	59	7	549286	Universal Pictures	2015	0	0
The Lion K	260000000	1649676757	118	G	Animatior	0	English	USA	88	8.5	869323	Buena Vista	2019	0	0
The Aveng	220000000	1519557910	143	PG-13	Action	3	English	USA	69	8	1204899	Walt Disney Pictu	2012	0	0
Furious 7	190000000	1506249360	137	PG-13	Action	0	English	USA	67	7.2	341279	Universal Pictures	2015	1	0
Avengers:	280000000	1405403694	141	PG-13	Action	3	English	USA	66	7.3	685405	Walt Disney Pictu	2015	0	0
Black Pant	200000000	1346739107	134	PG-13	Action	0	English	USA	88	7.3	545565	Marvel Studios	2018	0	0
Harry Pott	125000000	1342000000	130	PG-13	Adventure	0	English	USA	87	8.1	704107	Warner Bros. Pict	2011	1	0
Star Wars:	200000000	1332459537	152	Unrated	Short	0	English	USA	0	8.3	87	Others	2017	0	0
Jurassic W	170000000	1303459585	129	PG-13	Action	0	English	USA	51	6.2	231869	Universal Pictures	2018	0	0

Statistical Analysis:

- Firstly we found the statistical data for the numerical features as follows:
  - For IMDb Votes the mean is: 79283.1397 vote
  - For Production Company the mode is ”Others” which indicates that no one production company had their movie be in top 100,000 movie by revenue.
  - For Revenue, the mean is: 7.099426e+07.
- Then using the correlation heat diagram Figure 1 we noticed the features that had a likely significant relation to revenue. Those features were: the budget and IMDb votes.

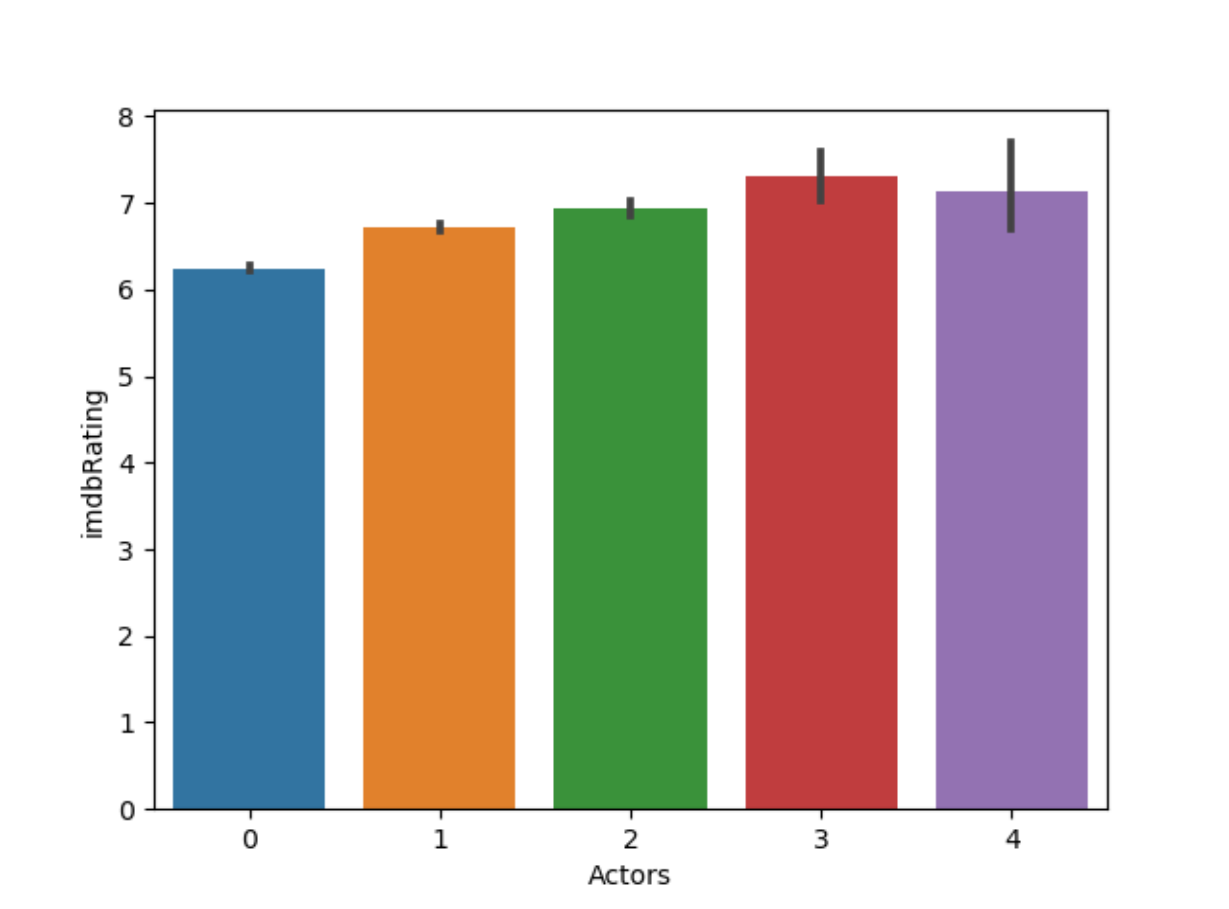
Statistical Analysis: Correlation Heatmap

- Figure 1



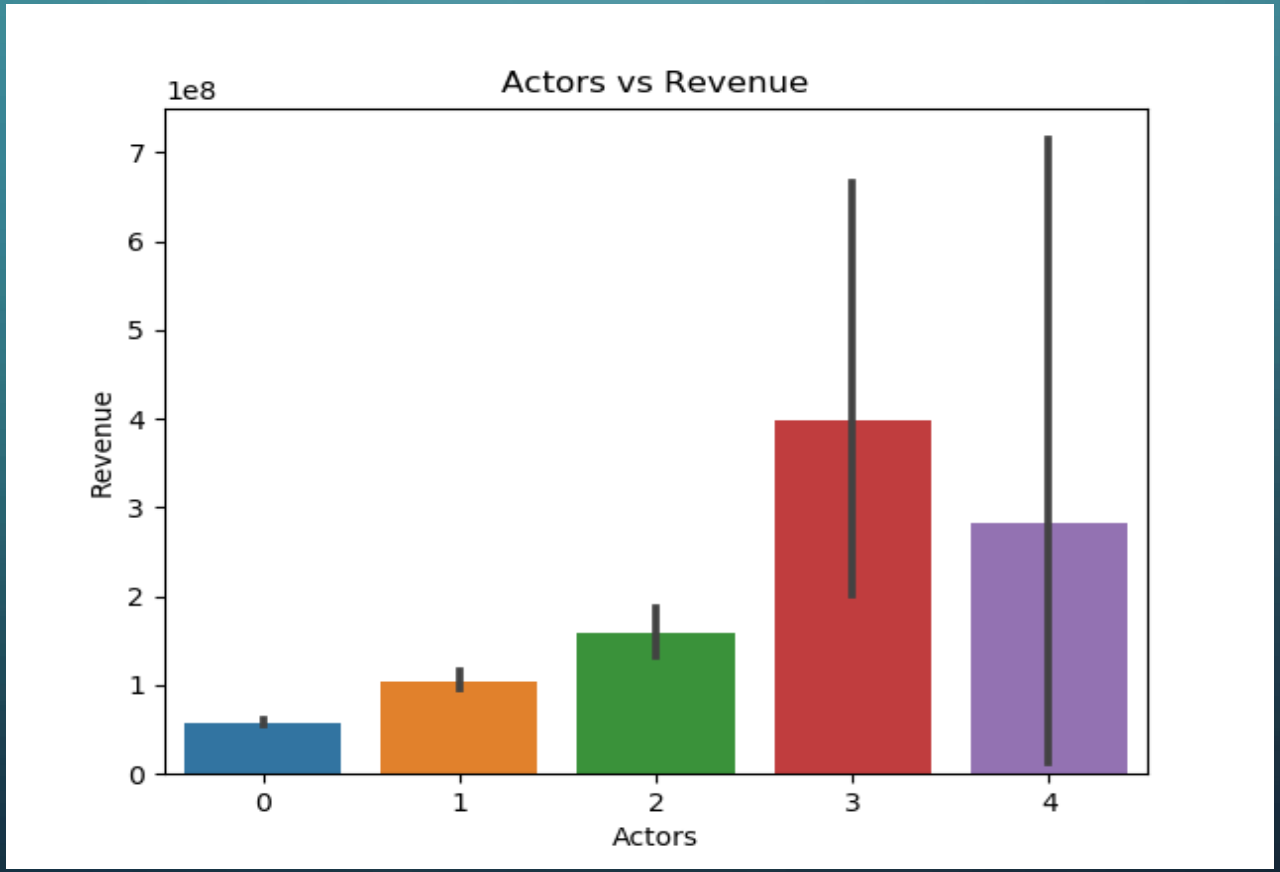
Statistical Analysis: Some Diagrams

- Figure 2: Actors vs IMDB rating
  - This shows that generally the more famous actors in a movie the higher the rating



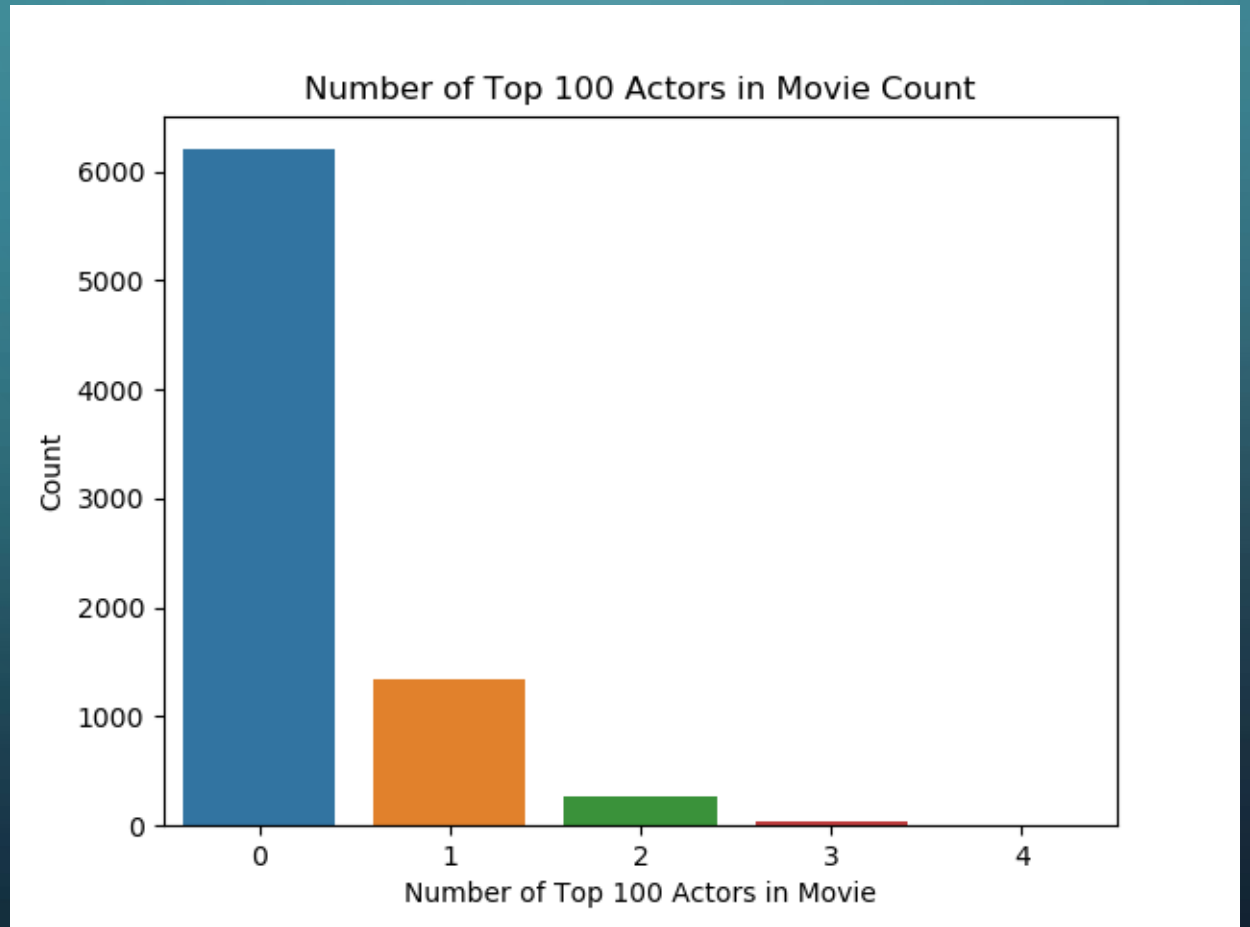
Statistical Analysis: Some Diagrams

- Figure 3: Actors vs Revenue
  - This shows that generally the more famous actors in a movie the higher the revenue is (note however that it goes down at 4 which could be interesting if we higher values for actors to check for a trend)



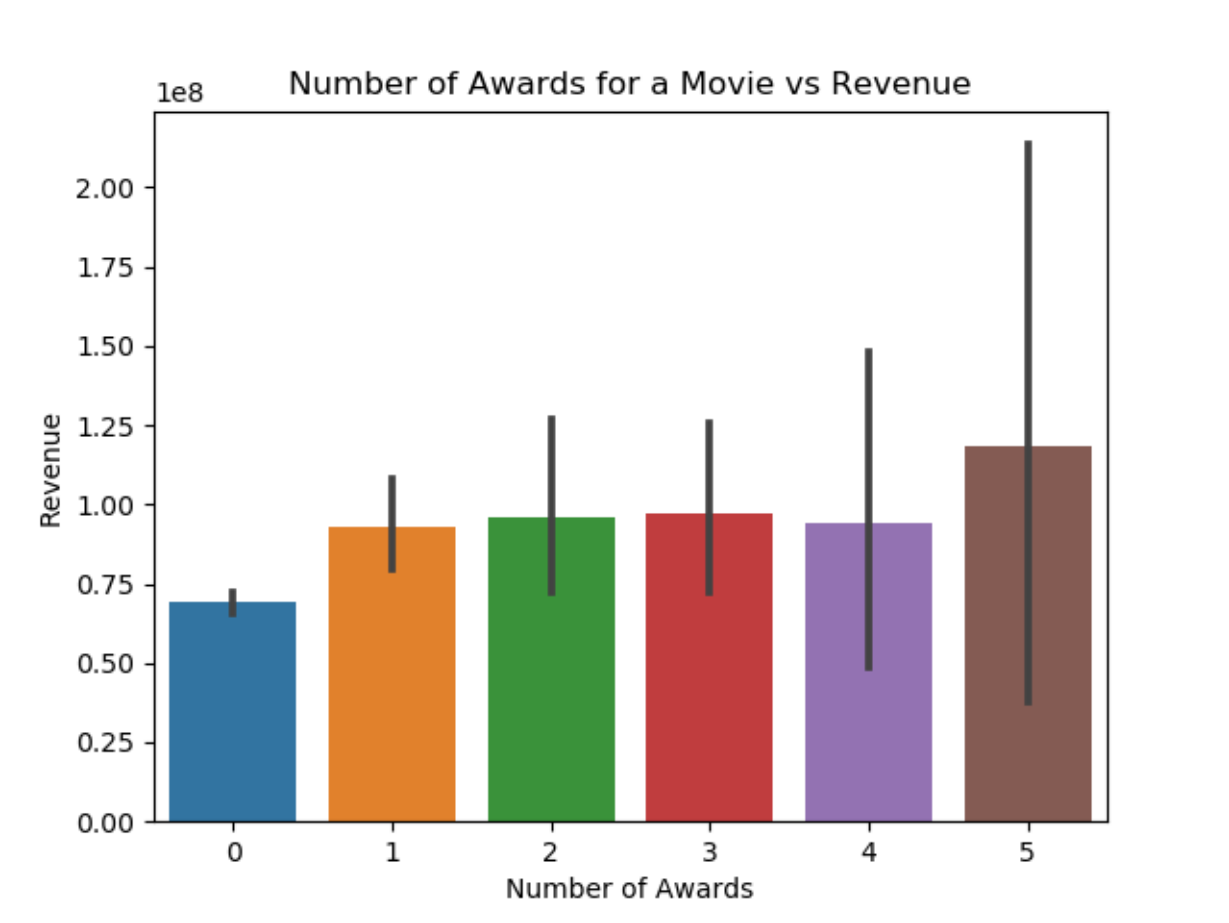
Statistical Analysis: Some Diagrams

- Figure 4: Count of movies depending on number of famous actors in it
  - This shows that there are almost no movies with more than 1 famous actors in them



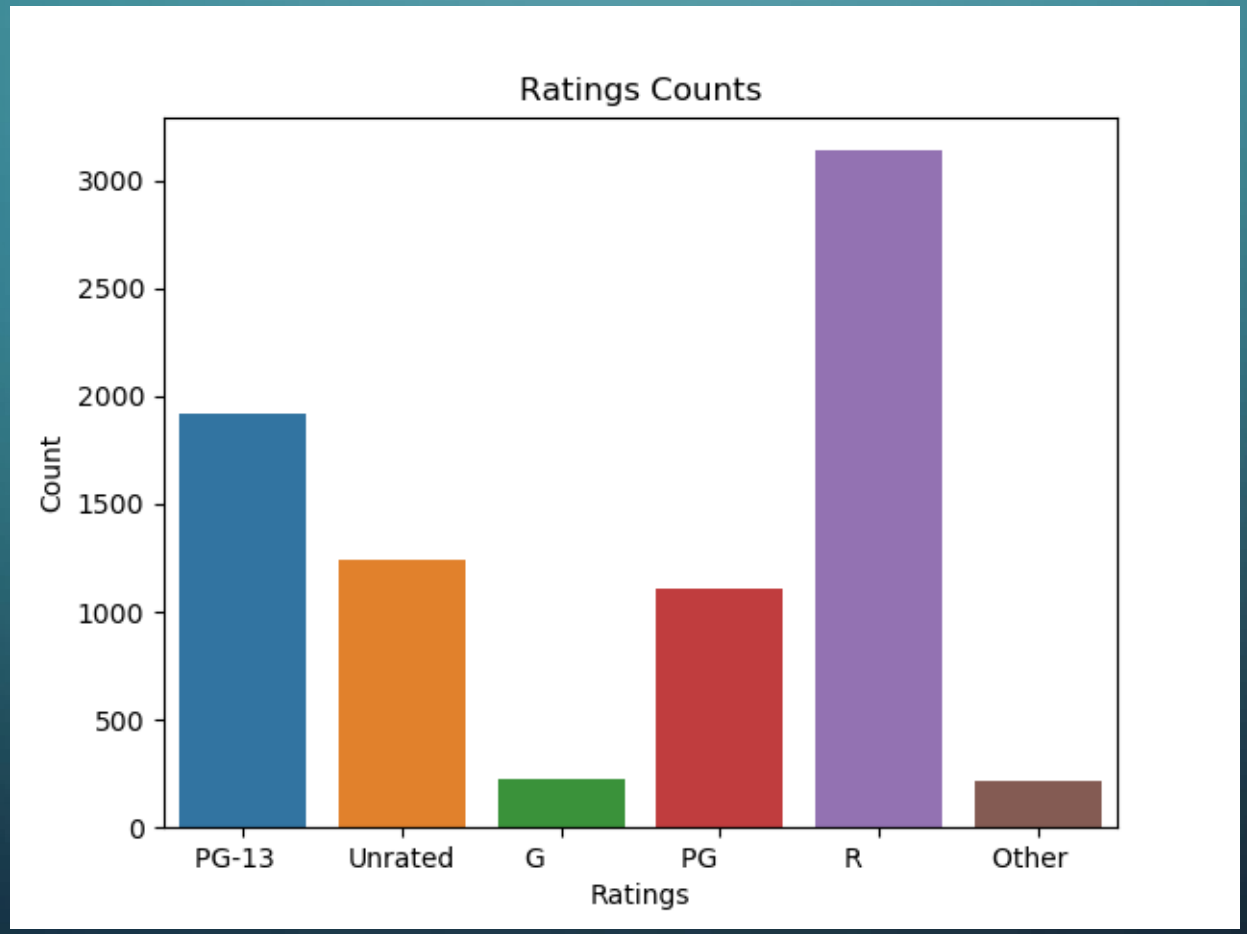
Statistical Analysis: Some Diagrams

- Figure 5: Awards vs Revenue
  - This shows that generally number of awards doesn't effect the revenue much



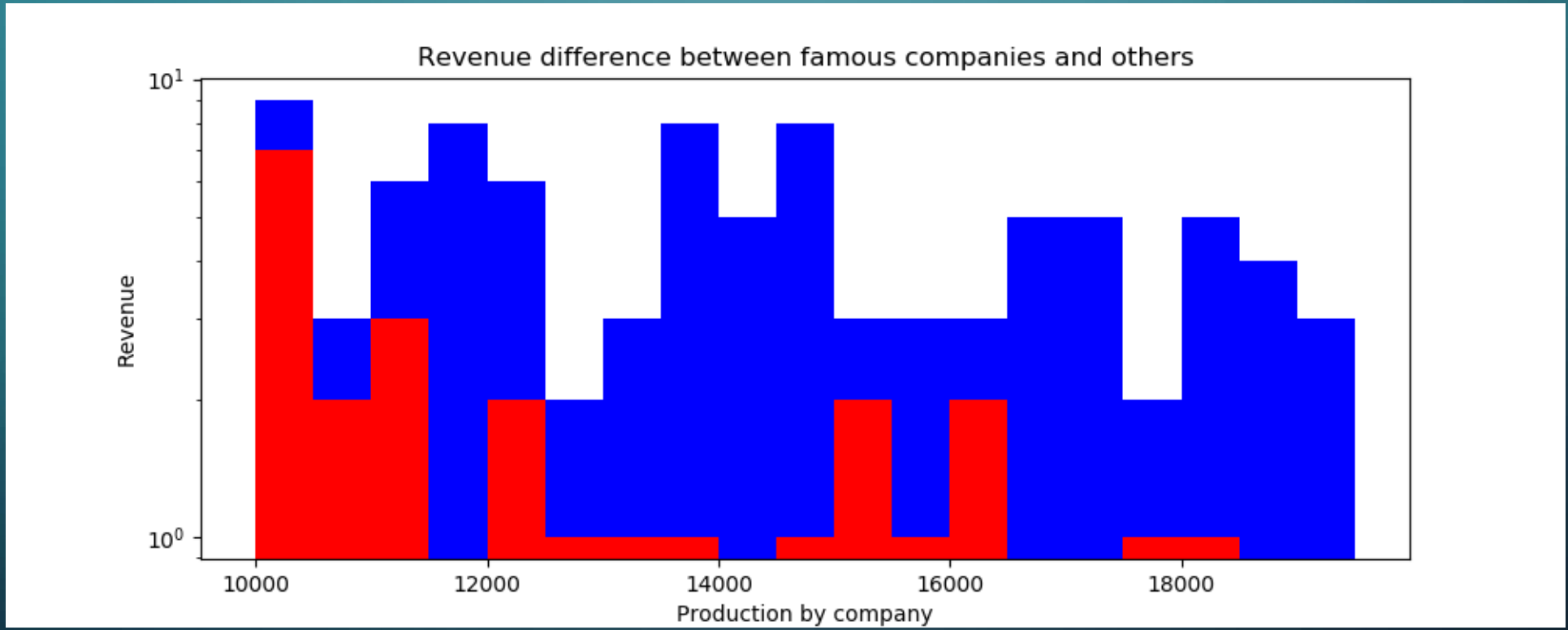
Statistical Analysis: Some Diagrams

- Figure 6: Count of movies depending on rating



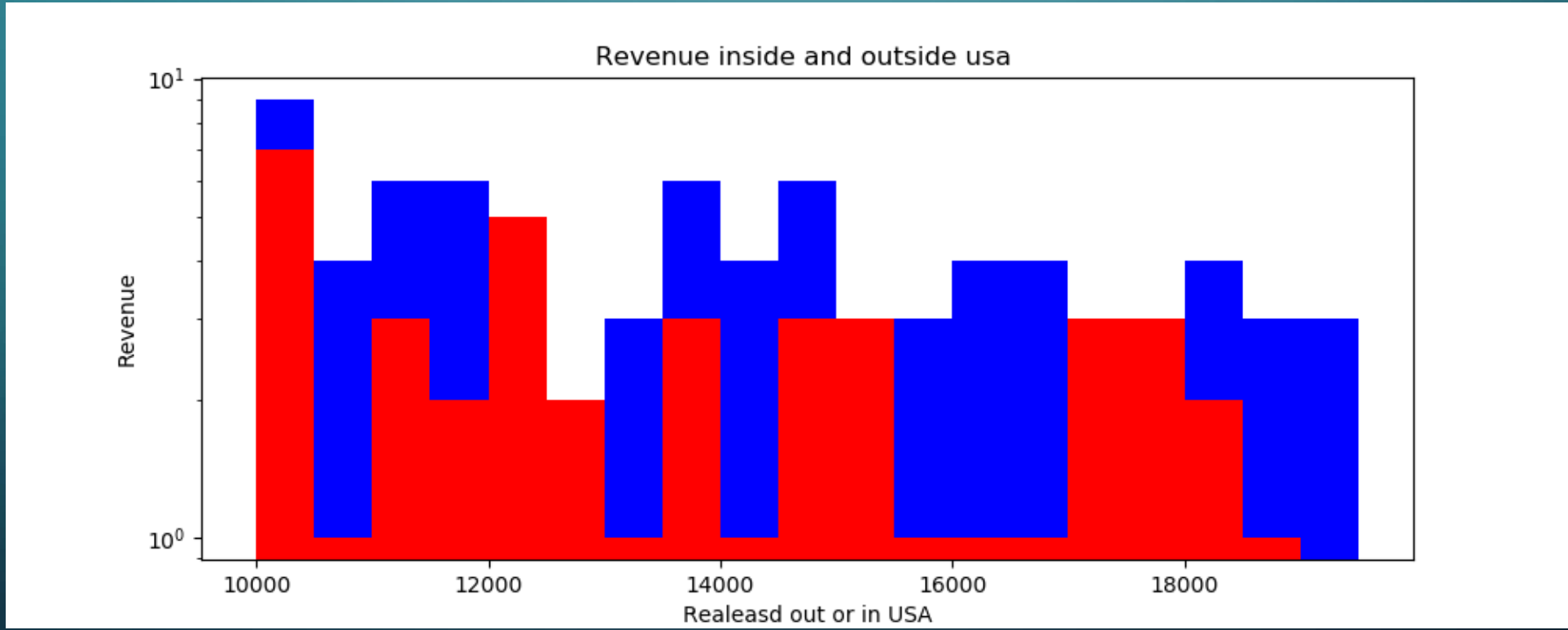
Statistical Analysis: Some Diagrams

- Figure 7: Revenue difference between named companies and others.
  - Red is others
  - Blue is named companies



Statistical Analysis: Some Diagrams

- Figure 8: Revenue difference between movies released inside USA only or world-wide
  - Red is inside of USA only
  - Blue is outside of USA too.



Statistical Analysis: Hypotheses Time

- we tested 9 hypotheses which are (each with its NULL hypothesis, method of testing and result): Hypothesis 1:

H0	• The lower the number of famous actors in a movie, the lower the revenue is
H1	• The higher the number of famous actors in a movie, the higher the revenue is.
Method of Testing	• using bar plot and checking the Peterson's Correlation Coefficient which was 0.187.
Result	• We rejected the alternative hypothesis.

Statistical Analysis: Hypotheses Time

- we tested 9 hypotheses which are (each with its NULL hypothesis, method of testing and result): Hypothesis 2:

H0	• The budget of a movie doesn't effect the number of famous actors in it.
H1	• The budget of a movie does effect the number of actors in it.
Method of Testing	• the Peterson's Correlation Coefficient which was 0.141.
Result	• We rejected the alternative hypothesis.

Statistical Analysis: Hypotheses Time

- we tested 9 hypotheses which are (each with its NULL hypothesis, method of testing and result): Hypothesis 3:

H0	• If a movie's genre is Action, it has a lower average revenue, than if it was of another genre.
H1	• If a movie's genre is Action, it has a higher average revenue, than if it was of another genre.
Method of Testing	• calculating the overall mean and comparing it with the mean of action genre movies.
Result	• The alternative hypothesis passed, and we reject the null hypothesis

Statistical Analysis: Hypotheses Time

- we tested 9 hypotheses which are (each with its NULL hypothesis, method of testing and result): Hypothesis 4:

H0	• If a movie is released outside of USA, it's revenue will be lower than revenue if released in USA only.
H1	• If a movie is released outside USA, it's revenue will be higher than revenue if released in USA only.
Method of Testing	• the Mann-Whitney U test with p-value = 2.44e-10
Result	• We reject the alternative hypothesis

Statistical Analysis: Hypotheses Time

- we tested 9 hypotheses which are (each with its NULL hypothesis, method of testing and result): Hypothesis 5:

H0	• The Higher the budget of a movie the Lower the revenue is.
H1	• The Higher the budget of a movie the Higher the revenue is.
Method of Testing	• Method of testing: the Peterson's Correlation Coefficient which was 0.51.
Result	• The alternative hypothesis passed, and we reject the null hypothesis

Statistical Analysis: Hypotheses Time

- we tested 9 hypotheses which are (each with its NULL hypothesis, method of testing and result): Hypothesis 6:

H0	• The Higher number of famous actors in a movie, the Lower its rating is.
H1	• The Higher the number of famous actors in a movie, the Higher its rating is.
Method of Testing	• bar plot.
Result	• The alternative hypothesis passed, and we reject the null hypothesis

Statistical Analysis: Hypotheses Time

- we tested 9 hypotheses which are (each with its NULL hypothesis, method of testing and result): Hypothesis 7:

H0	• The Higher the number of people who voted for a movie, the higher its rating is.
H1	• The Higher the number of people who voted for a movie, the lower its rating is.
Method of Testing	• the Peterson's Correlation Coefficient which was 0.34.
Result	• We reject the alternative hypothesis.

Statistical Analysis: Hypotheses Time

- we tested 9 hypotheses which are (each with its NULL hypothesis, method of testing and result): Hypothesis 7:

H0	• If a movie is produced by "others", then it will have higher than average revenue.
H1	• If a movie is produced by "others", then it will have lower than average revenue.
Method of Testing	• the Mann-Whitney U test with p-value = 2.44e-10.
Result	• We reject the alternative hypothesis.

Statistical Analysis: Hypotheses Time

- we tested 9 hypotheses which are (each with its NULL hypothesis, method of testing and result): Hypothesis 7:

H0	• If a movie has more than 1 award, it has a lower revenue.
H1	• If a movie has more than 1 award, it will have a higher revenue.
Method of Testing	• using bar plot and checking the Peterson's Correlation Coefficient which was 0.04.
Result	• We reject the alternative hypothesis.

Machine Learning Model: Logistic Regression

For the machine learning we conclude that a movie is successful if the revenue is double the budget.
For the string type columns (Genre , Country,…) we changed them to category type and used label encoding

```
data["Genre"] = data["Genre"].astype('category')
data['Genre_cat'] = data["Genre"].cat.codes
```

We used Logistic Regression to build the model. Firstly, we started with all the features to see how much accuracy they give us, and we kept iterating the model until we got an acceptable accuracy. The features we were using at that point were the budget and imdb-votes.
The final accuracy of the model was 0.77

Machine Learning Model: Checking for Overfitting

We used cross validation to try to avoid overfitting problem.

```
accuracy = cross_val_score(logreg,x,y,scoring="accuracy", cv=10)
print(accuracy)
print("Accuracy of the model with cross validation is: ", accuracy.mean() * 100)# => 76.099
```

The resulting accuracy is 76.099, which indicates that it’s likely we didn’t fall in overfitting trap.

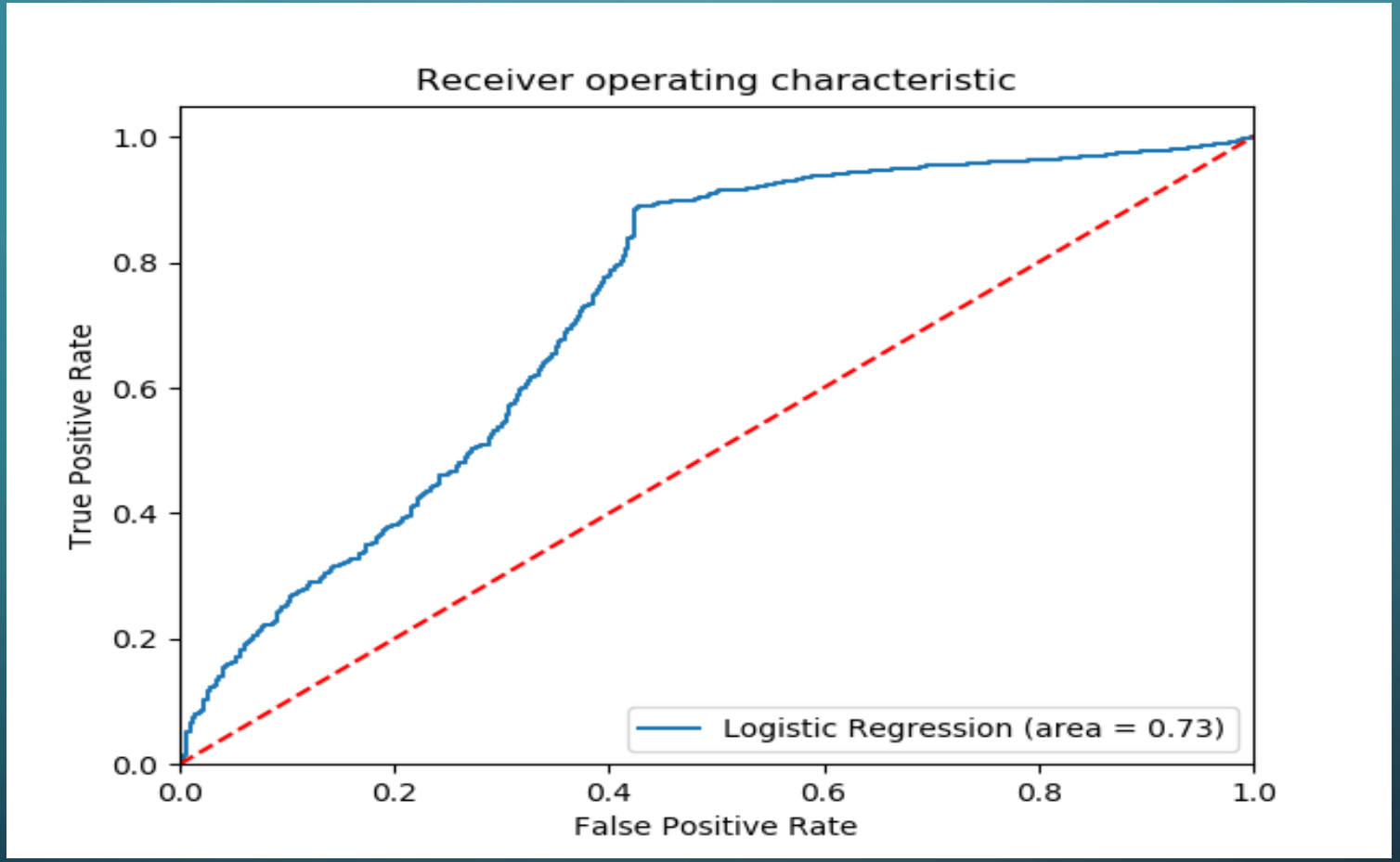
Logistic Regression Confusion Matrix

```
Accuarcy of logistic regression classifier on test set: 0.77
[[ 497  366]
 [ 173 1318]]
              precision    recall  f1-score   support

         0.0       0.74      0.58      0.65       863
         1.0       0.78      0.88      0.83      1491

    accuracy                           0.77      2354
   macro avg       0.76      0.73      0.74      2354
weighted avg       0.77      0.77      0.76      2354
```

Logistic Regression ROC Curve



Conclusion

We set out in this study to answer the question of what affects the revenue of a movie and to build a ML model to classify/predict whether a movie is successful or not.
We found out that the budget and the number of IMDB-votes of a movie are the main effecting forces on its revenue. The budget in this case is predictable, however, the IMDB-votes could indicate that the more successful the movie is the more the audience are inclined To vote for it on IMDB – needs further studying.

For the ML model we used Logistic Regression to create a model that can classify/predict whether a movie is successful with an accuracy of 0.73 using 2 features which are budget and IMDB-votes.

THANKS FOR LISTING