TMDB Box Office Prediction

COMP9417 Machine Learning Project
Shu Yang (z5172181), Yue Qi (z5219951), Tim Luo (z5115679), Yixiao Zhan (z5210796)

1. Introduction:

The booming film industry nowadays not only brings revenue, but also has a positive impact on the global economy. The global box office was worth $41.7 billion in 2018 (Dave, 2019). Researchers and industry have drawn great attention to movie box office revenue prediction (Ramesh et al., 2006).

This project topic was chosen from a past Kaggle competition which concluded on May 30, 2019 and had 19,034 entries. The purpose of this project is to build a model to predict the expected revenue for a movie given the information such as budget, production company, cast, crew etc. Our stacked model is built on multiple regression models including Linear Regression, K-Nearest-Neighbour, Ridge Regression, Support Vector Regression, Elastic Net as well as ensemble methods including Cat Boost, Random Forests and XGBoost. Linear regression was chosen as the stacked regressor to generate the output.

2. Related Work:

The majority of Kaggle competitors tackled this task using boosting models (LightGBM, XGBoost, CatBoost) or Random Forest. Ensemble methods provides a robust tool to combine the strengths of a collection of simpler base models (Hastie et al., 2009), and therefore is also the main approach we adopt. However, rather than picking a single boosting model, we additionally include regression models and perform substantial numbers of diverse experiments (feature extraction, model hyperparameters tuning, etc).

Our approach differentiates from the previous work in two ways:

- A broader selection of ML methods and the analysis of their performance on an individual basis, compared to the competition on Kaggle which have mostly stuck with one gradient boosted model.
- An extra step to use a two-layer stacked model based on the combination of these selected models and to demonstrate comparable results.

3. Implementation:

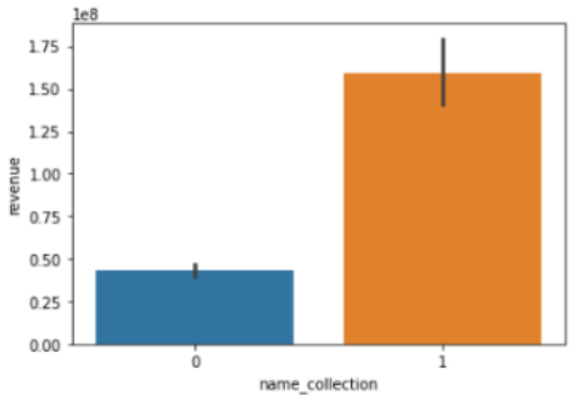
The Movie Database (TMDB) box office prediction software used in this project is coded in Python 3.7 and the source code is contained in the provided files. The training data and testing data are provided by Kaggle platform. There was significant data pre-processing involved, and the data had to be separated and processed into new attributes, as some of them are in dictionary and json form. For example, the genres feature has 'name' and 'id' under them. Hot one encoding was performed on some categorical features and all missing values in the dataset will be filled with either a zero or their true value. Some features were dropped entirely because of the difficulty in processing them, like the path to each movie's poster. We assumed a skewed nature in the dataset and performed log transformations across all non-Boolean data.

This data would be saved and then passed to the stacked model, and the results of the stacked model and each underlying model in the stacked model would be evaluated. Further evaluation of each model would be performed by modifying the script to include slightly altered replications of each model.

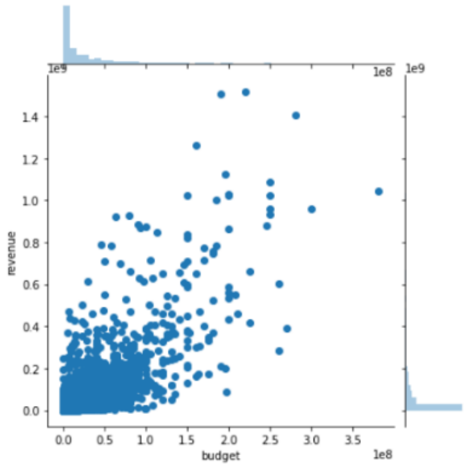
4. Experimentation:

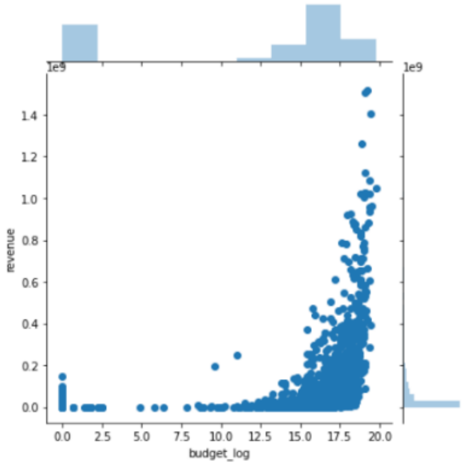
Data Pre-processing

The first step for the project is to check the relationship between different attributes and movie revenue. The following chart shows that movies with a collection attribute will gain more revenue than those without.

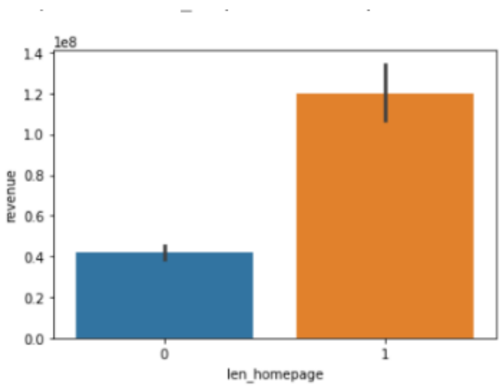


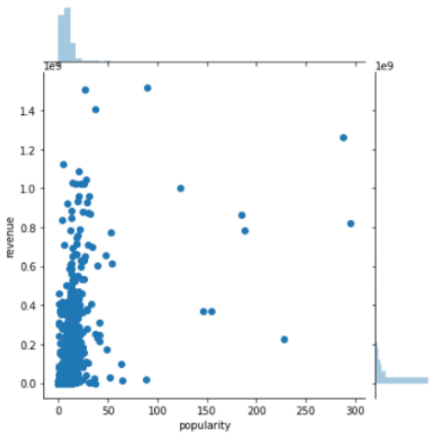
One movie can earn more if it belongs to one of the collections. After that, we can see that the budget also has a positive relationship with revenue. The next chart describes the movie revenue according to budget.





McKenzie (2012) believes that the relationship between budget and movie success is of critical importance. We use the logarithm to transform the original budget and get a new chart. This new graph at least gives a better visual indication that the more of a budget you have, the greater chance you have of having better than average revenue. The next step is to check the genres. There are numerous genre types and each movie may have more than one genre. After that, we explore how the homepage may have some effect on revenue:





This relationship might be because customers who have access to more information about the movie will be more attracted to that movie.

The relationship between popularity and revenue is not straight forward. From the above graph on the right, a movie can have lots of revenue even when its popularity is at a low level. Since the database does not explain how popularity is calculated, the attribute remains. The production companies feature is also another difficult feature to evaluate. Some movies will have multiple production companies in production. This requires further statistical analysis which we are not skilled at.

The following just describe pre-processing performed on some features, with little to no evaluation on their results. They were merely added as new features to the dataset:

- Next, the production countries attribute contains dictionary values, and pre-processing is taken to extract the production countries out of each value. We examined the relationship between the number of production countries, and which production countries are related with the revenue and got the result shown in Appendix (Figure 1 and 2).
- Release date has format dd/mm/yy or dd/mm/yyyy, and pre-processing was taken to extract information about each movie's release month, day, and year. We then examine the relationship for each of these attributes vs revenue and also which day, month or year produces the most revenue and we got the result shown in Appendix (Figure 3,4 and 5).
- We then examined the distribution of runtime for each movie, and it seems like movies with runtime 75-175 mins have the most revenue. The result is shown in the Appendix (Figure 6).
- For spoken language, we extracted information about the total number of spoken languages as well as which languages are spoken in the movie. The result shown in the Appendix (Figure 7 and 8).
- There are two status for movies, released or rumoured, and we examined the mean revenue for each status. Obviously, released movies have a much higher revenue than rumoured movies, therefore we add a binary attribute "isReleased" to our data. The result is shown in the Appendix (Figure 9).
- For the keywords feature, we extracted information about movie keywords as well as number of keywords for each movie. The result is shown in the Appendix (Figure 10 and 11).

Pre-processing of cast and crew were also performed by summing the revenues of the movies that each person participated in, divided by their frequency of occurrence. Then, each movie was assigned a cast score and a crew score, which is the sum of the individual cast scores / individual crew scores. These new features; cast score and crew score had high correlation scores (0.7 and 0.88 respectively) to revenue.

Feature Extraction, Models and Training

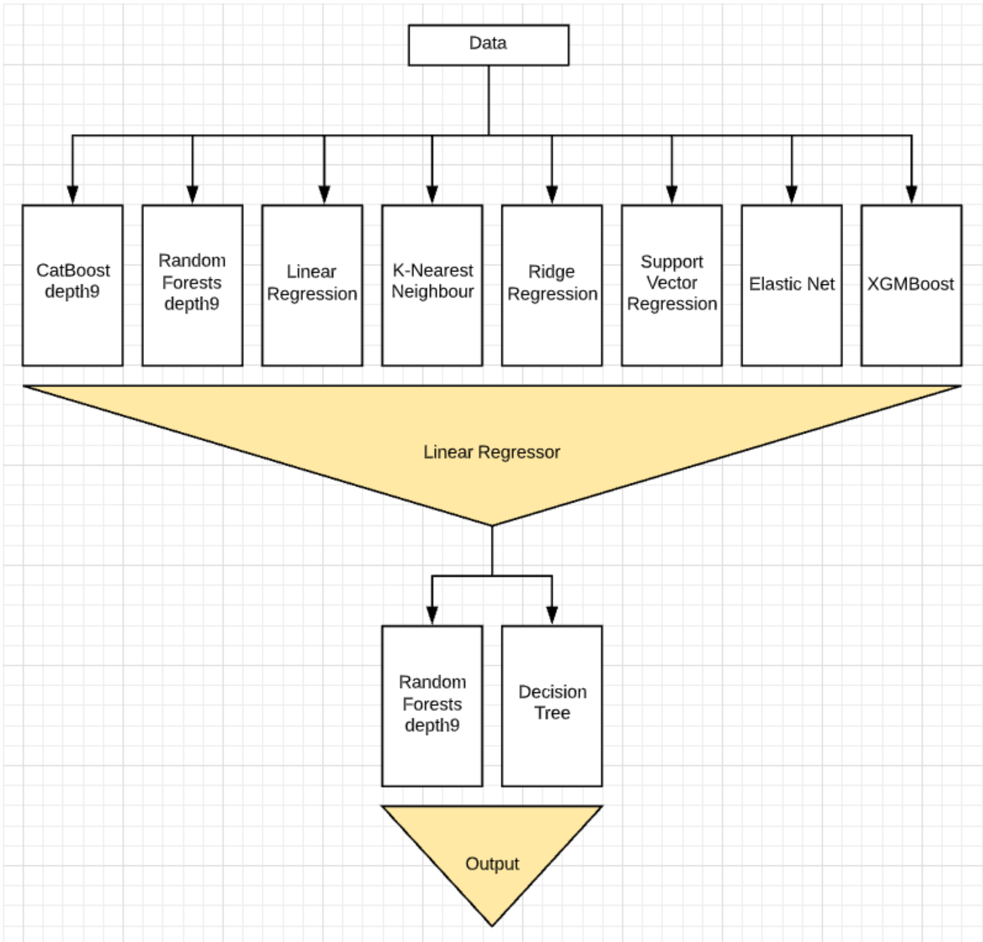
Feature extraction occurred by picking the top 50 features via the use of the chi2 function (X^2), which allows us to rank categorical data based on expected E and observed frequency O. The top three features ended up being cast, crew and budget:

$$X_c^2 = \frac{(O_i - E_i)^2}{E_i}$$

We then trained multiple models with these chosen features using 5-fold validation, and also used stacking of such models. Linear regression was used here instead of merely taking the averaging of each model output, as there was significant variation between some models. The output at each stacked level can be represented as:

$$Y_{level} = \sum x_i w_i$$

Where the output is a linear combination of the outputs of each model x and their respective weights w. The result was a marginally better model than the best individual model based on their RMSE score.



Training was done with RMSE as opposed to RMSLE, as the logged error was negative for many data points. The competition on Kaggle also used RMSE as opposed to RMSLE as their training metric for the same reason, although they submitted for RMSLE.
A study into the effectiveness of each model can be broken up into the two categories: models which had boosting, and other methods of regression. The RMSE of each individual model, and the stacked model are shown in the next table.

Model	RMSE
CatBoost	1.8057692568734254
Random Forests	1.8526249072546879
Linear Regression	1214848788.2093296
K Nearest Neighbours	2.430380245387095
Ridge Regression	2.006801850717835
SVR	2.291424765672057
Elastic Net	2.2433106029805914
XGBoost	1.9093307954392251
Decision Tree	2.113053342461618
Stacked Model	1.7807357845309713

It can be seen that the stacked model performed slightly better than the cat-boosted model. However, it could even be argued that the cat-boosted model was better, as k-fold validation showed that it had a lower standard of deviation of scores.

Cat Boost, XGBoost, Random Forest, Decision Tree

This group of models performed significantly better than the linear regression models. This is expected, as the solution is non-linear. The key difference between each of these models is the way they handle overfitting and gradient descent.

The traditional random forest and decision tree lack gradient boosting. Surprisingly however, the random forest model performed just as well as the boosted models. This may be because of our small training set of 3000 samples, which may end up reducing the effectiveness of gradient boosting. Hence, data hungry gradient boosted models did not perform significantly better than a random forest. Pulling more data from external data sources may solve this issue, however it would've been an extremely expensive operation. Experimentation on larger values of k produced better results across all models as shown in the next table.

Model	k=2	k=5	k=10	% Improvement (k=2 to k=10)
CatBoost	1.8027	1.784	1.766	2.07
XGBoost	1.9786	1.887	1.854	6.72
Random Forest	1.8393	1.812	1.794	2.52
Decision Tree	2.1060	2.113	2.049	2.78

The decision tree model was expected to perform worse than the other tree models. This is because no randomness was utilised, and so the utilisation of the entire data set was more prone to overfitting.

Despite CatBoost and XGBoost being limited by the small sample size, it is worth exploring the small improvement over the Random Forest by CatBoost. Looking at how each model deals with overfitting we can see that XGBoost allows the user to declare a minimum child weight, whereas CatBoost has L2 leaf regularisation.

CatBoost L2 Leaf Coefficient	Result	XGBoost Minimum Child Weight	Result
2	1.772	1	1.888
3	1.784	2	1.907
4	1.789	3	1.947

Clearly, increasing any of the parameters which help regulate overfitting is not effective here, so they should be kept to a minimum. Taking CatBoost as the best of the boosting models, let's observe what happens when we change the parameter early stopping rounds at different iteration numbers:

Early Stopping Rounds	Iterations = 1000	Iterations = 2000
100	1.784	1.779
200	1.784	1.779

No effect was made by increasing early stopping rounds or iterations which means that the overfitting detector had no effect. This explains why Random Forests performed nearly as good as Cat Boost, as there was no need for the extra parameters to account for overfitting. Finally, let's look at the effect of the depth of the tree:

Model	Depth = 3	Depth = 5	Depth=7
Cat Boost	1.832	1.797	1.785
Random Forest	1.959	1.827	1.806

Increasing the depth made a reasonable improvement to the model, but Cat Boost still performed better due to the added gradient boosting. The Cat Boosted model alone could have been used for the submission of this competition.

Linear Regression, Ridge Regression, Support Vector Regression, Elastic Net

Linear regression performed the worst with an RMSE of 1214848788. This is expected as the 50 features being used were certainly not going to be able to be fit linearly.

SVR with an RBF kernel followed with an RMSE of 2.29. The kernel is meant to base the maximal margin on a radial basis, based on the projection of data onto higher dimensions. This performance is decent despite the fact that the data used is heterogeneous and most of the original data is completely independent of each other.

The remaining regression models can be broken down into three categories: linear, lasso and ridge regression. Lasso and ridge regression are regularisations of linear regression with the aim of punishing large coefficients. Ridge regression is aimed at reducing the complexity of the model by adding a penalty parameter that is equivalent to the square of the magnitude of the coefficients. Lasso regression also aims to reduce the complexity of the model by adding a penalty parameter which limits the sum of the absolute values of the model coefficients. Elastic Net performed the worst out of the remaining models with an RMSE of 2.24, which means a Lasso regularisation did not work as well as Ridge Regression (Ridge regularisation) with an RMSE of 2.01.

Despite the effectiveness of these models when it comes to high dimensionality, these models did not perform as well as the gradient boosted trees.

K-Nearest Neighbours

KNN, which performs poorly with high dimensional data due to the curse of high dimensionality, seems like a terrible idea, but very surprisingly it achieved a result comparable to SVR, elastic net and ridge regression models. Further experimentation produced the results (with the inverse of the distance taken) in the next table.

Number of Nearest Neighbours	Euclidean Distance	Manhattan Distance	Minkowski (p = 2)	Minkowski (p = 5)
5	2.221	2.209	2.221	2.224
10	2.174	2.162	2.174	2.180
15	2.151	2.145	2.151	2.160
20	2.144	2.136	2.143	2.147
25	2.143	2.130	2.143	2.146
30	2.142	**2.124**	2.142	2.149

A K-NN of larger nearest neighbours works better, with the Manhattan distance metric. However, these results are nowhere near as well as Cat Boost and it can be concluded that perhaps the data has been processed in a way that in fact reduced the dimensionality of the data set and it ended up being more akin to linear regression.

5. Conclusion - Result of Stacked Model & Cat Boosted Model:

Our final stacked model achieved a RMSE of 1.781 gave us a RMSLE of 2.66627:

Submitting our Cat Boosted model with an RMSE of 1.784 gave us an RMSLE of 2.59963:

Our results place us in the top two thirds of the competition, if we were to compete in it. I believe that we could have improved significantly if this project did not have such a heavy emphasis on data pre-processing and statistical analysis, to which we are not as well-trained in. Despite being outperformed by simpler models used by other competitors such as a single random tree model with a RMSLE of 1.71, we are still happy with our efforts.

References:

Hastie, T., Tibshirani, R. and Friedman, J (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). pp.605-624. Springer.

Ramesh Sharda, Dursun Delen (2006) Predicting box-office success of motion pictures with neural networks. *Expert Systems with Applications*, Vol. 30, pp.243–254.

McKenzie, J 2012, 'The Economics of Movies: A Literature Survey.' *Journal of Economic Surveys*, vol.26 (1), pp.42–70.

McNary, Dave (2019). *2018 Worldwide Box Office Hits Record as Disney Dominates* https://variety.com/2019/film/news/box-office-record-disney-dominates-1203098075/

Appendix:

```
TOP 10 revenue by production countries count

Movie produced by 4 countries has mean revenue 86812511.32
Movie produced by 2 countries has mean revenue 86128791.22
Movie produced by 3 countries has mean revenue 70720933.22
Movie produced by 5 countries has mean revenue 63699051.14
Movie produced by 1 countries has mean revenue 63105182.26
Movie produced by 8 countries has mean revenue 16756372.0
Movie produced by 0 countries has mean revenue 4090428.27
Movie produced by 6 countries has mean revenue 2957964.0
```

Figure 1: Top 10 revenue by production countries count

```
TOP 10 revenue by production countries

Movie produced from Czech Republic,United Arab Emirates,United States of America has mean revenue 694713380.0
Movie produced from New Zealand,United States of America has mean revenue 607134808.86
Movie produced from Czech Republic,Germany,Italy,United Kingdom,United States of America has mean revenue 599045960.0
Movie produced from Germany,New Zealand,United States of America has mean revenue 550000000.0
Movie produced from Canada,Hong Kong,Taiwan,United States of America has mean revenue 532950503.0
Movie produced from Malta,United States of America has mean revenue 531865000.0
Movie produced from Czech Republic,Poland,Slovenia,United States of America has mean revenue 419651413.0
Movie produced from Australia,Canada,China,Hong Kong,United States of America has mean revenue 331957105.0
Movie produced from Australia,Canada,France,Germany has mean revenue 312242626.0
Movie produced from Czech Republic,United States of America has mean revenue 300257475.0
```

Figure 2: Top 10 revenue by production countries

```
TOP 10 revenue by release year

Movie produced on 2017 has mean revenue 181403935.1
Movie produced on 2015 has mean revenue 103854185.98
Movie produced on 1975 has mean revenue 90480379.5
Movie produced on 2002 has mean revenue 87773835.64
Movie produced on 2012 has mean revenue 86166013.78
Movie produced on 2005 has mean revenue 81908092.21
Movie produced on 2008 has mean revenue 80945077.79
Movie produced on 2004 has mean revenue 80308074.69
Movie produced on 2003 has mean revenue 78921195.15
Movie produced on 1999 has mean revenue 77762278.52
```

Figure 3: Top 10 revenue by release year

```
TOP 10 revenue by release month

Movie produced in June has mean revenue 127886388.25
Movie produced in December has mean revenue 96462381.08
Movie produced in July has mean revenue 94040208.18
Movie produced in May has mean revenue 90487720.1
Movie produced in November has mean revenue 77024607.19
Movie produced in March has mean revenue 72241452.53
Movie produced in April has mean revenue 58314160.82
Movie produced in February has mean revenue 57788894.47
Movie produced in October has mean revenue 42042643.07
Movie produced in August has mean revenue 41510704.6
```

Figure 4: Top 10 revenue by release month

```
TOP 10 revenue by release day of week

Movie produced on Wednesday has mean revenue 114644613.81
Movie produced on Tuesday has mean revenue 94524296.19
Movie produced on Thursday has mean revenue 75184812.39
Movie produced on Monday has mean revenue 75174747.1
Movie produced on Saturday has mean revenue 49078893.58
Movie produced on Friday has mean revenue 45778331.97
Movie produced on Sunday has mean revenue 45516473.78
```

Figure 5: Top 10 revenue by release day of week

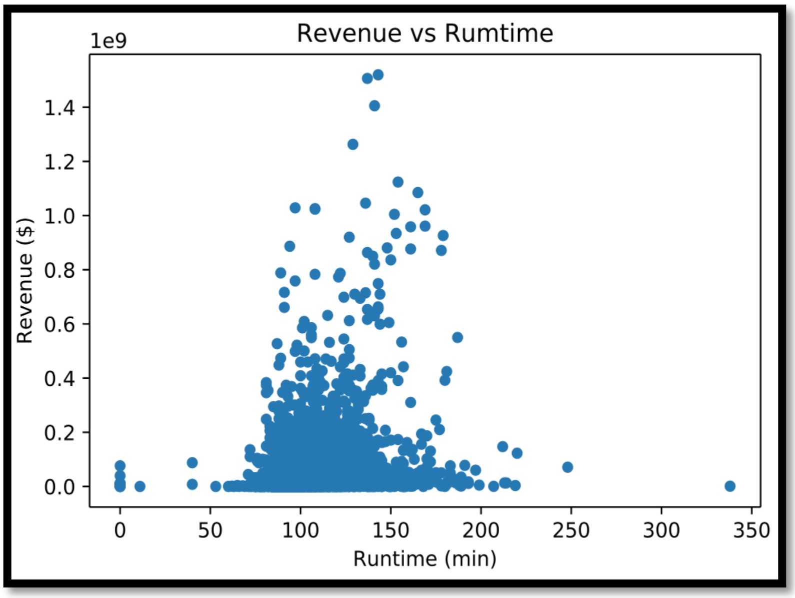


Figure 6: Revenue vs Runtime graph

```
TOP 10 revenue by spoken languages

Movie of spoken language Deutsch,English,Español,Français,Italiano has mean revenue 733382668.0
Movie of spoken language English,Français,Pусский,svenska,العربية has mean revenue 694713380.0
Movie of spoken language Latin,עִבְרִית has mean revenue 611899420.0
Movie of spoken language English,Français,Italiano,日本語 has mean revenue 559852396.0
Movie of spoken language English,Español,Français,العربية,日本語 has mean revenue 544272402.0
Movie of spoken language ,English,Italiano,日本語 has mean revenue 461983149.0
Movie of spoken language English,Français,Latin has mean revenue 457363168.0
Movie of spoken language Deutsch,English,Français,Latin,ελληνικά,العربية has mean revenue 441306145.0
Movie of spoken language Deutsch,English,Español,Italiano,Íslenska,广州话 / 廣州話,한국어/조선말 has mean revenue 431971116.0
Movie of spoken language English,العربية,ภาษาไทย has mean revenue 407778013.0
```

Figure 7: Top 10 revenue by spoken languages

```
TOP 10 revenue by spoken languages

Movie of 5 spoken language  has mean revenue 195729621.78
Movie of 6 spoken language  has mean revenue 140989896.5
Movie of 7 spoken language  has mean revenue 126276621.33
Movie of 4 spoken language  has mean revenue 78345438.85
Movie of 3 spoken language  has mean revenue 73720764.0
Movie of 2 spoken language  has mean revenue 70674970.35
Movie of 1 spoken language  has mean revenue 63441481.86
Movie of 9 spoken language  has mean revenue 14624826.0
Movie of 0 spoken language  has mean revenue 7348542.9
Movie of 8 spoken language  has mean revenue 572461.5
```

Figure 8: Top 10 revenue by spoken languages count

```
Released movie mean revenue = 66810292.01301736
Rumored movie mean revenue = 3480198.75
```

Figure 9: Revenue by movie status

```
TOP 10 revenue by keywords

Movie of 33 Keywords  has mean revenue 352927224.0
Movie of 20 Keywords  has mean revenue 209592217.75
Movie of 27 Keywords  has mean revenue 158019845.42
Movie of 16 Keywords  has mean revenue 155621054.57
Movie of 21 Keywords  has mean revenue 144255409.13
Movie of 19 Keywords  has mean revenue 121454306.36
Movie of 18 Keywords  has mean revenue 116418103.62
Movie of 9 Keywords  has mean revenue 112406881.81
Movie of 14 Keywords  has mean revenue 111617902.95
Movie of 32 Keywords  has mean revenue 107989703.0
```

Figure 10: Top 10 revenue by keywords count

```
TOP 10 revenue by keywords

Movie of Keywords aftercreditsstinger,alien invasion,based on comic,duringcreditsstinger,marvel cinematic universe,marvel comic,new york,shield,superhero,superhero team
has mean revenue 1519557910.0
Movie of Keywords car,car race,muscle car,race,revenge,speed,suspense has mean revenue 1506249360.0
Movie of Keywords 3d,based on comic,duringcreditsstinger,marvel cinematic universe,marvel comic,sequel,superhero,superhero team,vision has mean revenue 1405403694.0
Movie of Keywords 18th century,3d,anthropomorphism,beast,castle,creature,curse,fairy tale,france,gothic,held captive,magic,musical has mean revenue 1262886337.0
Movie of Keywords alien planet,based on cartoon,bodyguard,commando,duringcreditsstinger,giant robot,moon,sabotage,spacecraft,traitor,transformers,word domination has me
an revenue 1123746996.0
Movie of Keywords batman,burglar,cat burglar,catwoman,cover-up,crime fighter,criminal underworld,dc comics,destruction,flood,gotham city,hostage drama,imax,secret ident
ity,superhero,terrorism,terrorist,time bomb,tragic hero,vigilante,villainess has mean revenue 1084939099.0
Movie of Keywords 3d,aftercreditsstinger,battle,captain,duke,mermaid,mutiny,pirate,prime minister,sailing,sea,ship,silver,soldier,swashbuckler,sword has mean revenue 10
45713802.0
Movie of Keywords amnesia,animation,anthropomorphism,fish,sequel,talking animal,underwater has mean revenue 1028570889.0
Movie of Keywords 3d,alice in wonderland,based on novel,fantasy,fantasy world,fictional place,queen has mean revenue 1025491110.0
Movie of Keywords 3d,animals,anthropomorphism,conspiracy,discrimination,female protagonist,fox,injustice,missing person,prejudice,rabbit,rookie cop,stereotype,urban has
mean revenue 1023784195.0
```

Figure 11: Top 10 revenue by keywords