TMDB Box Office Prediction: Group 6

Abhinav Pateria, Ravi Gunjan, Saurabh Kumar Gupta, Vipul Kumar, Yash Panchal

07/11/2019

Introduction

In this project we will predict how much revenue a movie makes at the box office. During this process we will be going through:

1 Exploratory data analysis

2 Feature engineering

3 Treating missing values

4 Machine learning using random forest.

Read the Data

Read in the train and test data sets and then bind the two sets using bind_rows() from the DPLYR package. We will do all feature engineering and data preparation on both data sets and then divide our data into train and test sets again later before creating our model.

With the help of glimpse of data we can categorize data into two types.

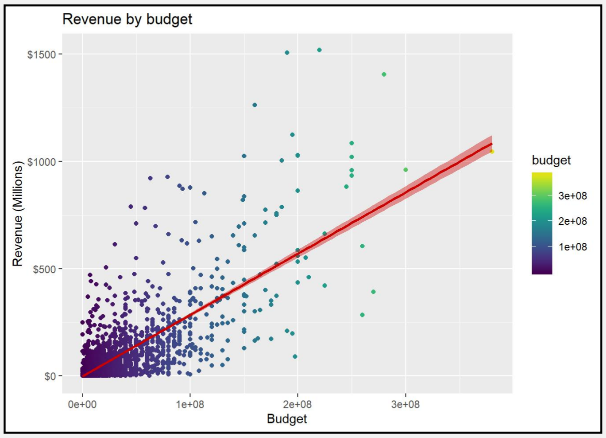
a) One which is less messier, requires little or no cleaning-i..id, budget, homepage, imdb_id, original_language, original_title, overview, popularity, poster_path, release_date, runtime, status, tagline, title, revenue.

b) Attribute which look quite messy, we will extract appropriate information from them before using in our model - belongs_to_collection, genres, production_companies, production_countries, spoken_languages, Keywords, cast,crew.

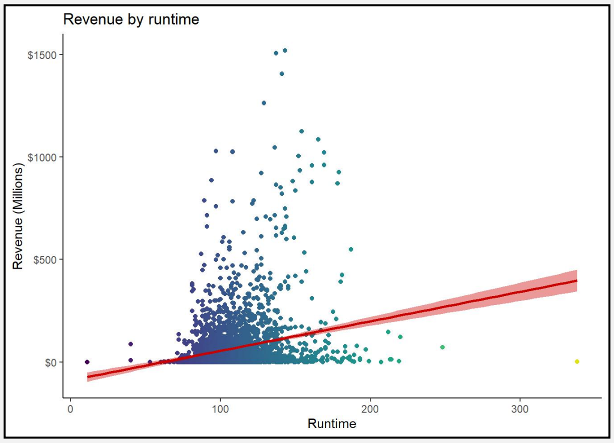
1. Exploratory Data Analysis

Lets begin by plotting our existing variables budget, runtime, and popularity in order to see their relation to the variable we are trying to predict, revenue.

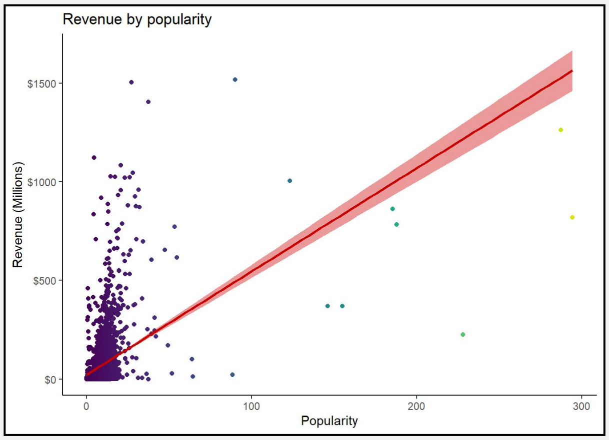
Budget



Run-time



Popularity



From the above graphs we can see that an increase in the budget and the popularity tend to lead to higher revenue. Run-time also show an increasing trend with revenue but not as strong as budget and popularity.

Now, we will do feature engineering to create features for our machine learning algorithm.

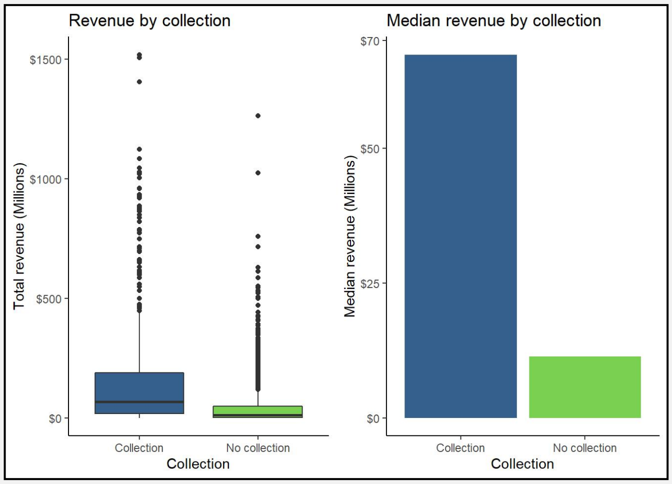
2. Feature Engineering Belongs to collection

Attribute belongs_to_collection are messy and unnecessary information is present which we do not need. To handle this, we will use regular expressions to extract the collection names from the strings in belongs_to_collection.

After the extraction of the collection names, we will check for the biggest collections.

From the above table we can see that in each collection movie count is fairly small, so we will engineer a new variable that consist of either ‘being in a collection’ or ‘not being in a collection’.

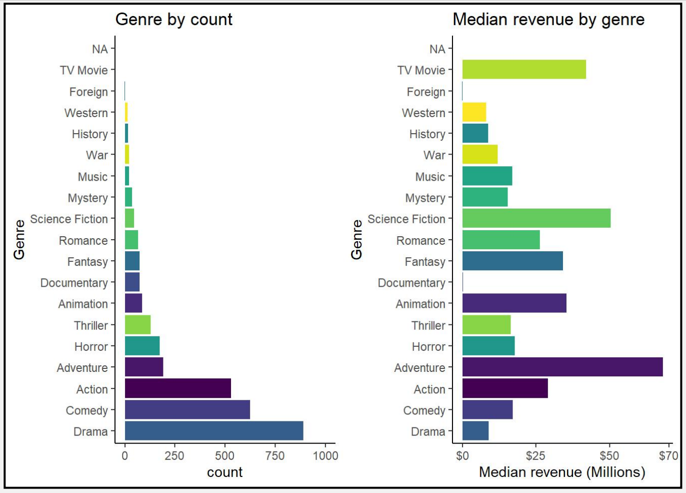
Now lets plot our Collection variable to visualize how the two levels differ on revenue.



On average, movies that are in collections seem to be getting higher revenues as we can see by looking at the box plot and bar plot.

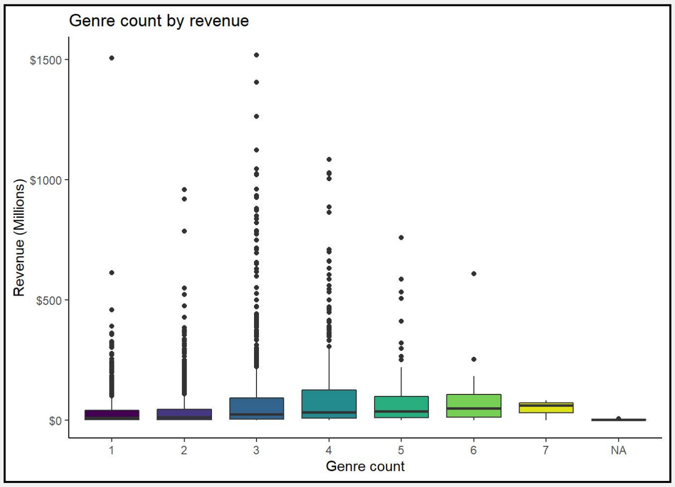
Main genre

We now want to extract the first genre from the genres strings to get the main genre for each movie. First, we will create a vector with the genres we want to extract. Next, we will extract the genres and add them to a new variable called main_genre.



Here we can see that different genres seem to be making different revenues. Adventure movies seem to have the highest median revenue, followed by science fiction. One thing to note is that the median revenue for genres with few counts, such as TV Movie, might be over-underestimations due to small sample sizes.

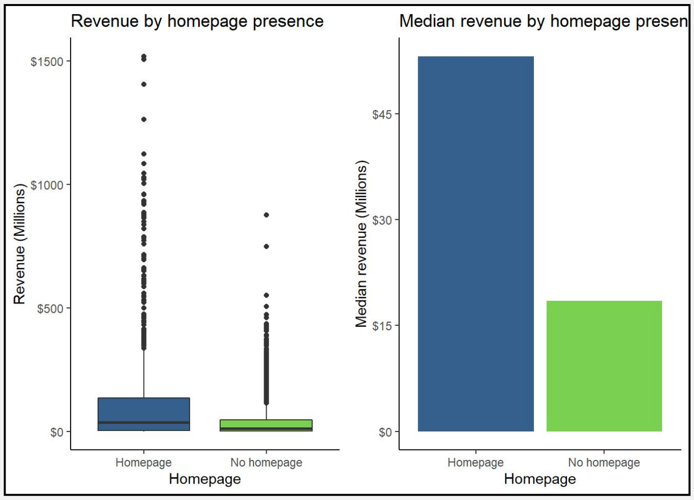
Genre Count



The trend in genre count increases from 1 to 4, reaches its peak and then shows a decreasing tendency. We can say that movies with 3-4-5-6 genre seems to generate a good revenue.

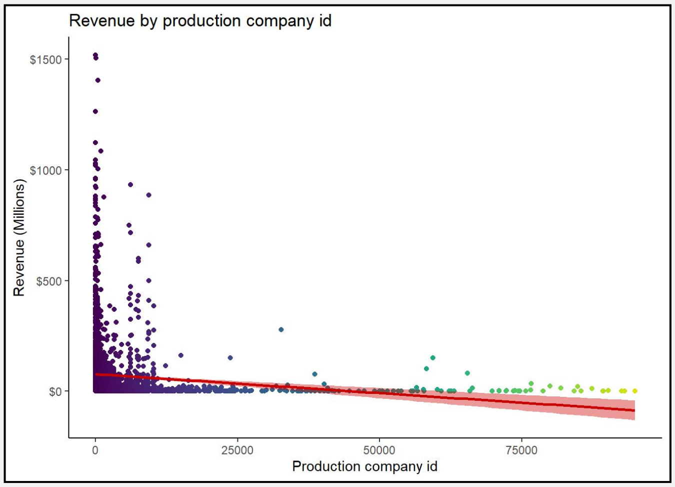
The correlation between Revenue and number of Genre a movie contains is 0.1636539.

Homepage



Movies with Homepage generates higher revenue than movies with no Homepage. Movies with homepages seem to be making on average 3 times as much as movies without a homepage.

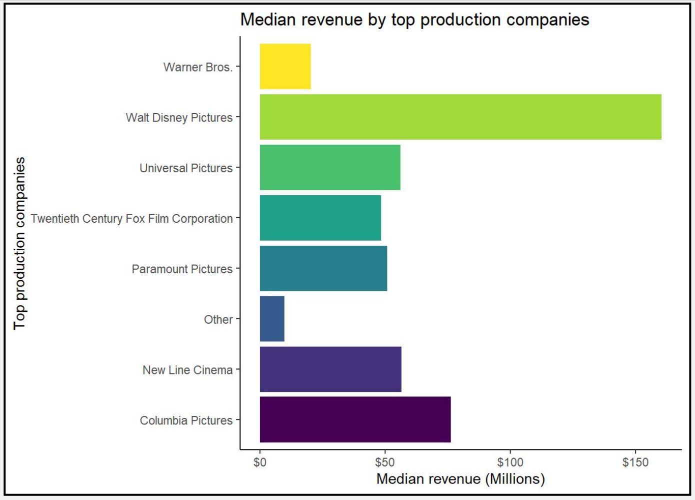
Production Company ID



Production companies with lower numbered id's seem to be making more revenue compared to the ones with higher id's. There is small negative correlation present (Correlation: -0.1282278)

Top production companies

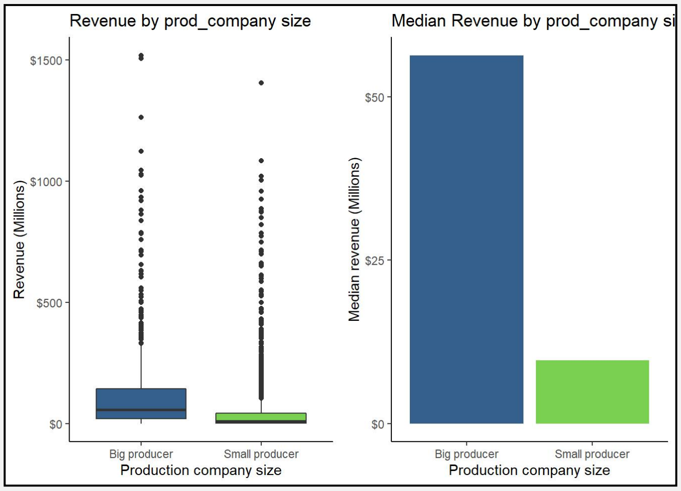
Lets extract the main production company name from production_companies. Separate into top production countries (criteria: 100+ movies) and 'other'. We will create a new variable called top_prod_comp (top production companies). We will create a separate category for each production company that has produced at least 60 movies that are present in our data set. All other production companies, including NAs, get put into an 'other' category.



Here we can see that the average revenue for a lot of the top production companies is higher than the 'other' production companies.

Production company size

Now, lets move on to create prod_comp_size (production company size – big producer v. small producer).

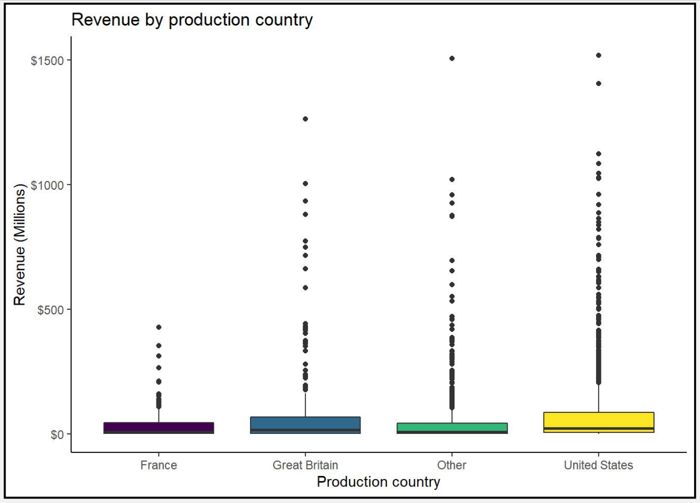


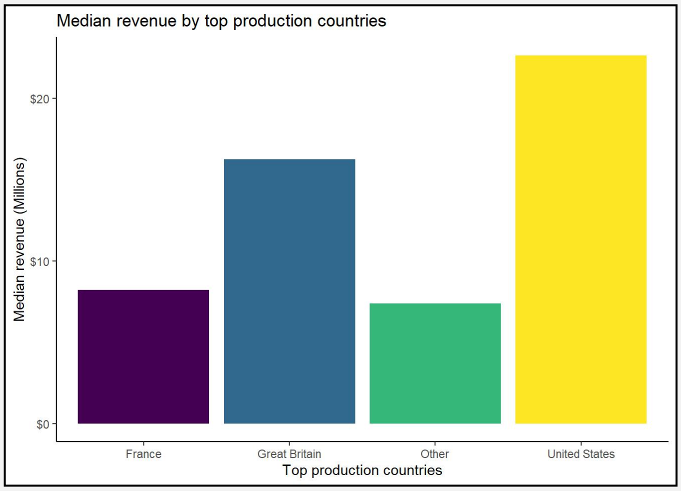
We assigned the production companies that have at least 60 movies each as big producers and all the rest as small producers. We will assume that all NAs are small producers.

Again, we can see that the big production companies are, on average, making more the smaller production companies.

Top production countries

We will separate into top production countries (criteria: 100+ movies) and 'other'.





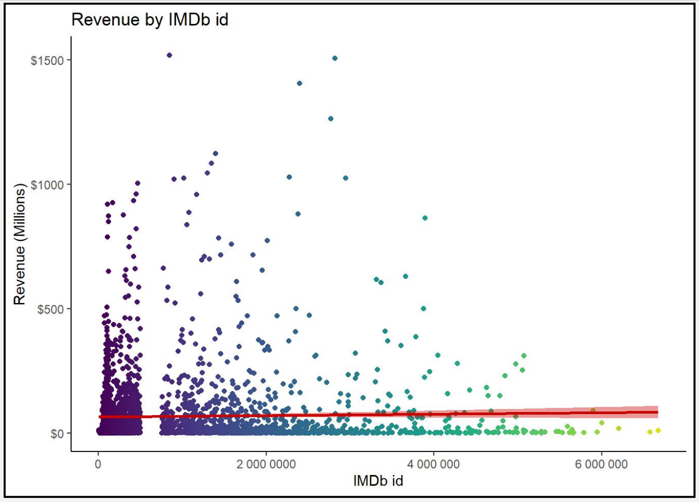
The U.S. and Great Britain seem to, on average, be getting more revenue than the countries that are not among the top production countries.

IMDB id

We will now extract the IMDb number from the IMDb_id string in order to see if this variable affects revenue. There will likely not be any correlation with this and revenue, but we will plot and explore this to make sure.

The format of the extracted value from imdb_id is in string. To create it's scatter plot we will parse the string values as in Integer.

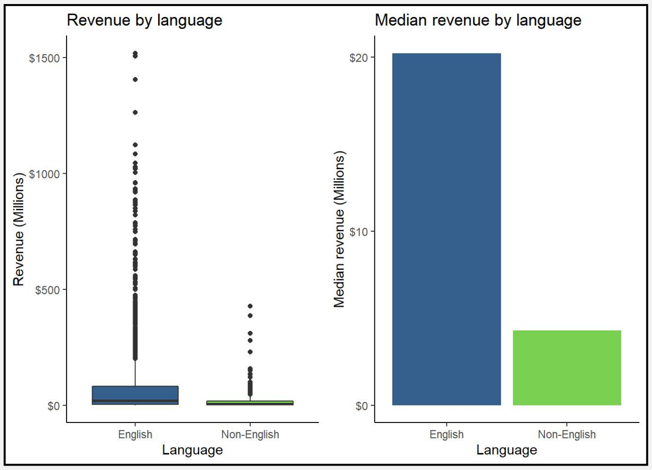
Create scatter plot of revenue by imdb_id_2



The correlation is very low (say threshold be 0.1. Here the correlation is 0.02428) This confirms that there is next to no correlation and that it is probably best to not include this variable in our prediction model.

Language

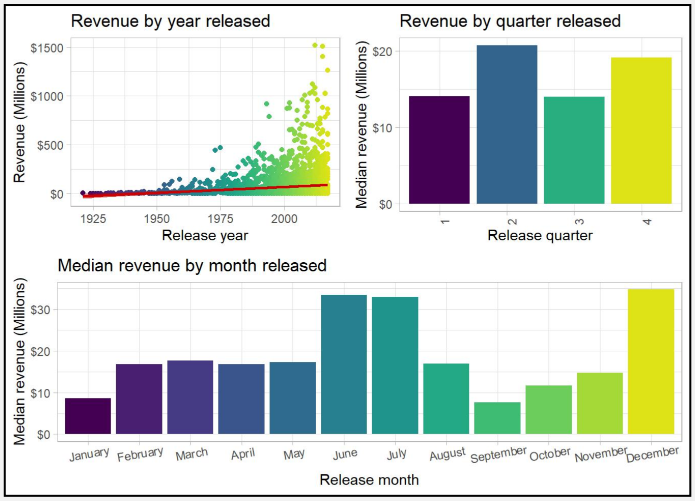
Absolute majority of the movies are English, we will create the variable language with levels English versus Non-English.

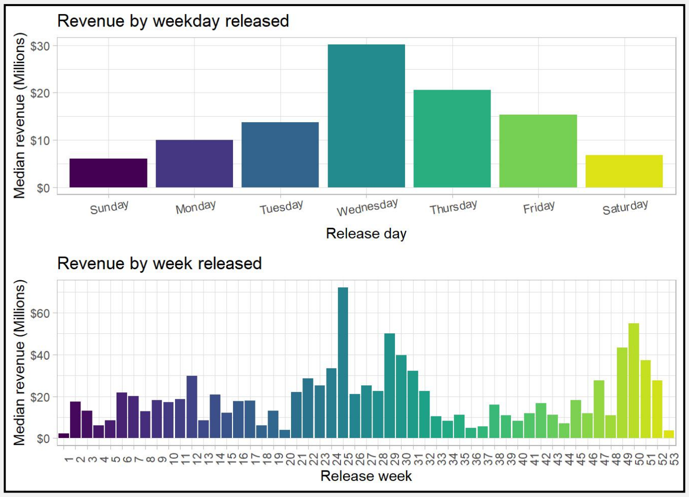


Seems like English-language movies make on average about 5 times the revenue of non-English language movies.

DATE

Now we will create 5 new variables: (1) year_released, (2) quarter_released, (3) month_released, (4) week_released, and (5) weekday_released.

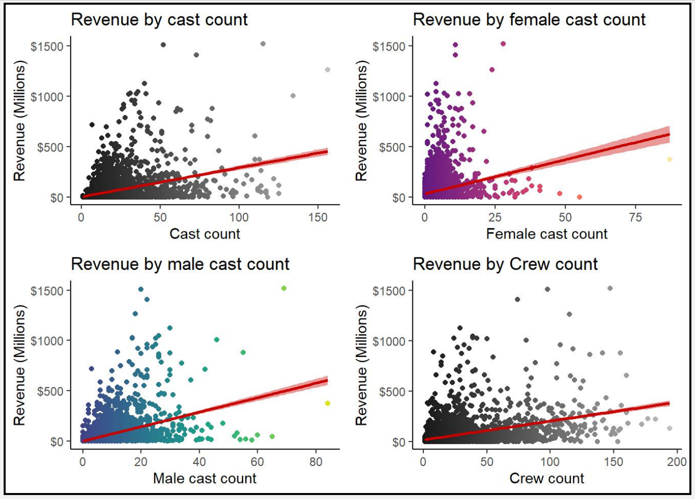


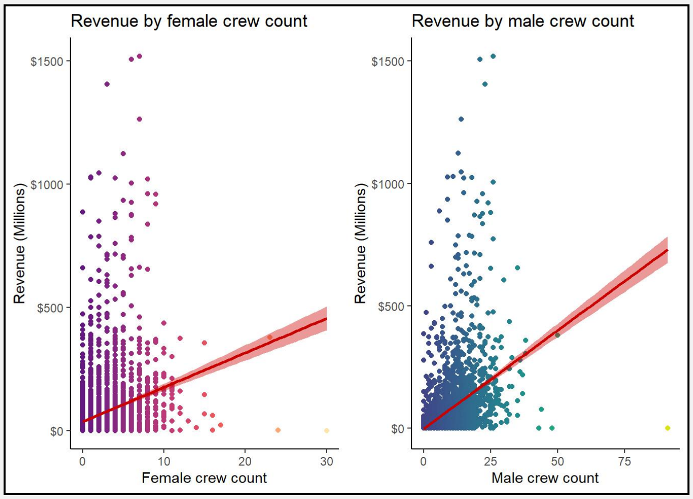


The year plot seems to indicate revenue has been increasing over the years. Movies being released in June, July and December seem to be getting higher revenues. This is in line with what one would believe as a lot of blockbuster movies are released during the summer, while a lot of movies that are trying to compete for the Oscars are released in December. Movies that are released on Wednesdays seem to be getting somewhat higher revenues as well.

Gender of cast & crew

We will now create new variables to see how gender of cast and crew affect revenue



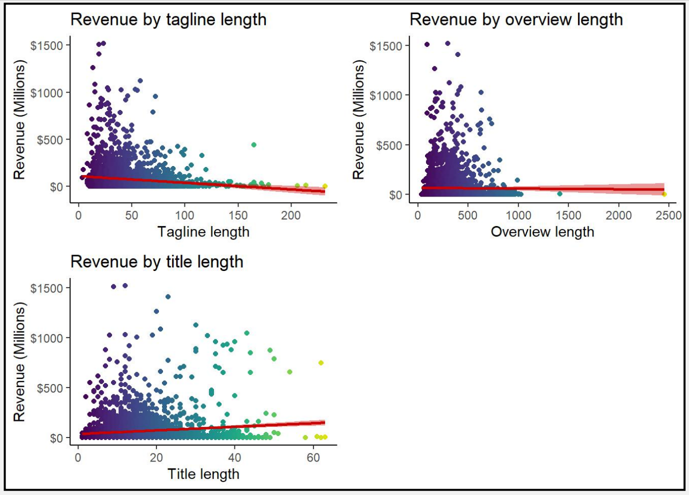


The distribution in revenue by gender for cast and crew. There seems to be a quite clear trend that the more cast and crew the movie has, the higher the revenue.

The more genres a movie has, the higher the median revenue. The more production companies a movie has, the higher the revenue, up to 6 production companies. A higher number than that seems to have more volatile results. This might be explained by smaller sample sizes. There seems to be no clear trend between number of production countries and revenue. There seems like there is no clear trend for number of spoken languages either. There is a trend between more keywords and higher revenue.

Length of title_length, overview_length and tagline

We will now create 3 additional variables, (1) title_length, (2) overview_length, and (3) tagline_length by extracting the lengths of the strings of the variables.



The correlation between these variables and revenue seem small. Lets take a look at what the actual correlations are. Revenue, Title_length: 0.1087646

Revenue, Tag-line_length: -0.1206457

Revenue, Overview_length: -0.008616267

So, there is a weak correlation between title length and tag-line length and revenue. There is no correlation between overview length and revenue so it is probably best to not include the variable in our model.

Machine Learning Model

Model 1: Linear Regression

We have splitted data in 80-20 ratio as Train data and Validation data.

Linear Regression for numerical attributes: Popularity, Runtime, Budget to predict revenue (ModelLR1)

- Residual standard error: 1.049 on 2396 degrees of freedom
- Multiple R-squared: 0.3984, Adjusted R-squared: 0.3976
- F-statistic: 528.9 on 3 and 2396 DF, p-value: < 2.2e-16

Now we have introduced some more attributes (prod_comp_size, prod_comp_id, collection, male_crew) along with popularity, runtime, budget to. (ModelLR2)

Check whether adjusted r squared value will increase or decrease by considering more attributes to our Linear Regression model. We got following values:

- Residual standard error: 0.9956 on 2392 degrees of freedom
- Multiple R-squared: 0.459, Adjusted R-squared: 0.4574
- F-statistic: 289.9 on 7 and 2392 DF, p-value: < 2.2e-16

Since the value of adjusted R squared value has increased from 0.3976 to 0.4574. So we have decided to consider all attributes after removing attributes having correlation value greater than 0.1 with revenue (ModelLR3). We got following:

- Residual standard error: 0.9638 on 2333 degrees of freedom
- Multiple R-squared: 0.5055, Adjusted R-squared: 0.4915
- F-statistic: 36.14 on 66 and 2333 DF, p-value: < 2.2e-16

Predicting Revenue for validation test using ModelLR3 and calculating RMSE value

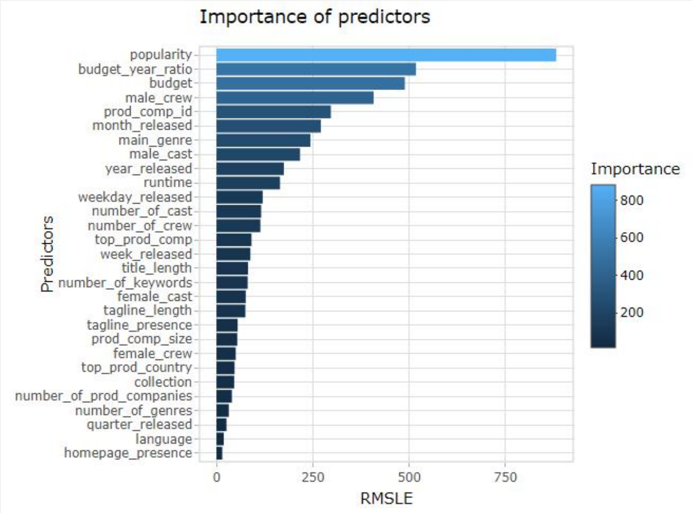
- RMSE Value for Linear Regression model (Model3) = 0.9133

Model 2: Random Forest

randomForest(formula = revenue ~ ., data = train, ntree =

- Number of trees: 501
- No. of variables tried at each split: 9
- Mean of squared residuals: 0.8591925
- % Var explained: 51.38

Importance of variables using Random Forest



Conclusion

The MSE of Random forest model (0.859) which is less than residual error using linear regression model (0.963). So we will use Random forest with bootstrap (taking samples with replacement).

Using Random forest machine learning model we calculated importance of predictors. Popularity, budget year ratio, budget,male crew number are the top important attributes to predicting the revenue of the Box Office movie.

After training and checking the performance of the model using validation data, we used test data to predict the revenue of the unknown movie. The corresponding revenue prediction of 4397 test data was saved in the “Box_office_prediction.csv”.