**1st Capstone Project Final Project: Movie Revenue Prediction**

1. **Problem statement**

I’m interested in figuring out whether those stereotypical thinkings are right or not based on a statistician or a data scientist angles. Can famous actors, directors, or both bring the highest revenue? How social media and networking service company ex: Facebook influence the movie industry? Is a high IMDB score a good sign for movie companies that they are gonna make money for sure? Is it guaranteed that high budget will lead to high revenue? Which genre is the revenue promise? Is it important to pick a good timing (month) to the movie’s announcement?

All in all, I want to know what are the most important features of a successful movie? Can we actually create a model to predict the profit for a movie? Therefore, the movie company can use the model to estimate its strategy in producing a movie.

1. **Data collection and wrangling summary**

Data collection

This dataset is from the Data World website, although it was originally posted on the Kaggle competition.

1. Data wrangling

[Cleaning step]

* 1. Checking the percentage of missing values in each variable (column) and observation (row)
     1. It tells me how to prioritize the recovery steps
  2. Duplicates removal

[Categorical variables]

* 1. Proofread ‘’movie title” column
     1. Remove unnecessary words and spaces
  2. Manually fix “color”, “country”, “language” columns
     1. Fill up NaN values
     2. Use one hot encoding
  3. “content\_rating” column
     1. Remove TV series
     2. Fill in NaN by web scraping (However, scraped data shows most of them are TV-series or Not rated. I would just skip the fill-in)
     3. Group them into 4 and dummify them
  4. Dummify “genres” column
  5. Replace “Actor\_name” and “director\_name” columns into frequency

[Numeric variables]

### Fill in "title\_year" column by web-scraped data and subgroup it

* 1. Fill in "budget" column with web-scraped data
  2. Fill in "gross" column” with web-scraped data
  3. Add “month” column by web-scraped data

### Impute "num\_critic\_for\_reviews", "director\_facebook\_likes", "actor\_3\_facebook\_likes", "actor\_1\_facebook\_likes", "facenumber\_in\_poster", "actor\_2\_facebook\_likes", "aspect\_ratio", "duration", "num\_user\_for\_reviews" columns with median

[Final steps]

* 1. Remove “movie\_imdb\_link” column
  2. Remove all rows with NaN
  3. Save it to ‘final\_wrangle.csv’

1. Data preprocessing

[Prepare target variable: revenue]

* 1. Create a new column called ‘revenue’ by ‘gross’ - ‘budget’
  2. Change its unit to 1 million

[Outliers]

* 1. Use seaborn.pairplot to get histograms of all predictive variables
  2. Check target variable

[Correlation]

* 1. Create a correlation heatmap
  2. Identify high positive and low negative correlation between variables
  3. Remove the variable which is highly related to the other variable positively or negatively

[Save the preprocessed data into final\_pre.csv]

1. **Exploratory data analysis**

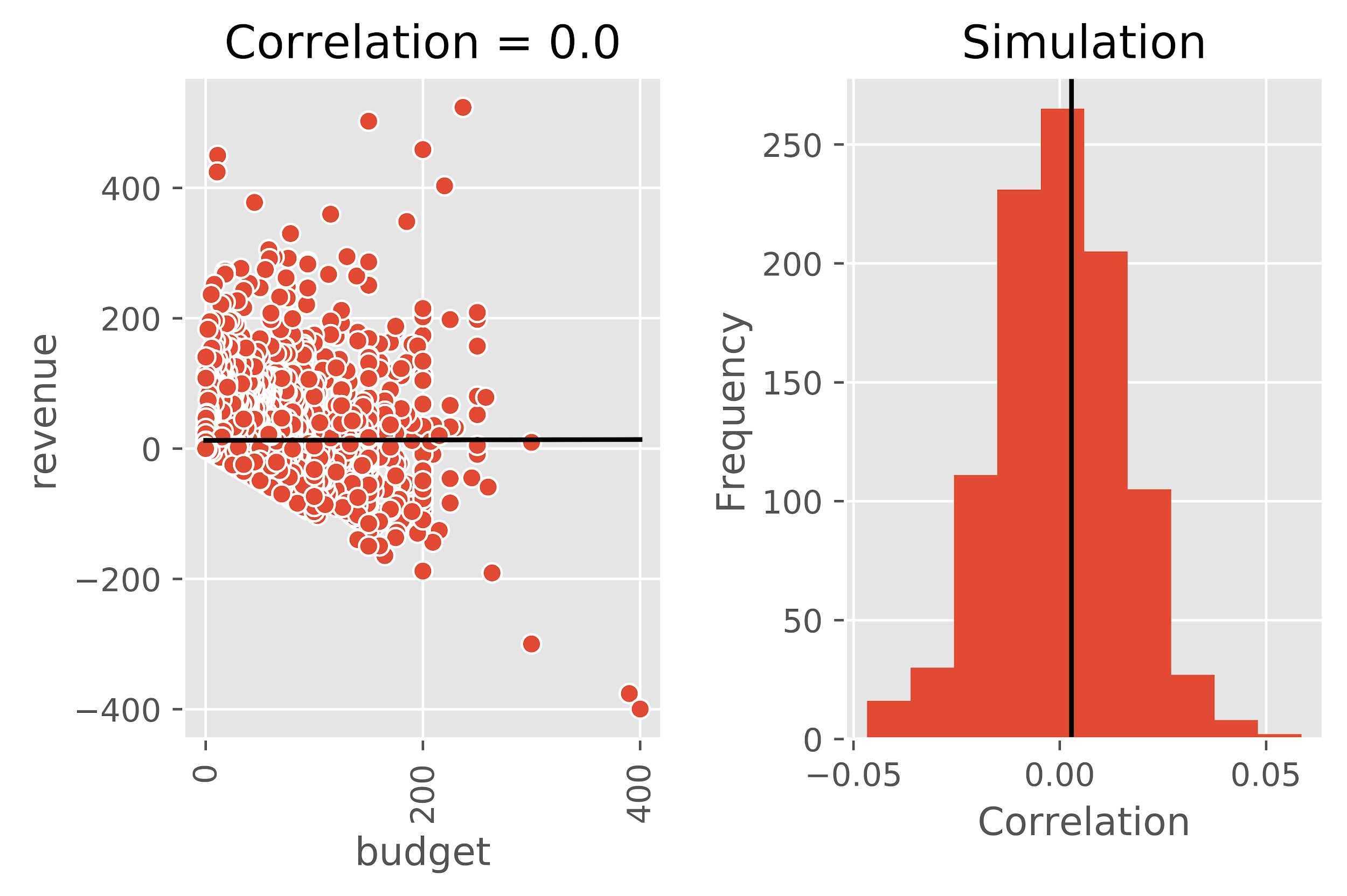
In the following exploration, I set up two hypothesis tests for correlation and mean difference for the inferential analysis purpose. For the correlation, I use the permutation technique to simulate a null hypothesis repeated at least 1000 times by bootstrapping technique. The simulated data with correlations form a histogram distribution. The significance is defined to compare the distribution and p-value, which is the fraction of the correlations from simulated data sets are at least as extreme as actual correlation. For the mean difference, I calculate the actual mean difference between two groups. Then, I use the permutation technique to simulate a null hypothesis repeated at least 1000 times by bootstrapping technique. The simulated data with mean difference form a histogram distribution. The significance is defined to compare the distribution and p-value, which is the fraction of the mean difference from simulated data sets are at least as extreme as actual correlation.

1. Load pre-processed data and functions
2. Check the correlations between predictive variables and revenue (target variable)

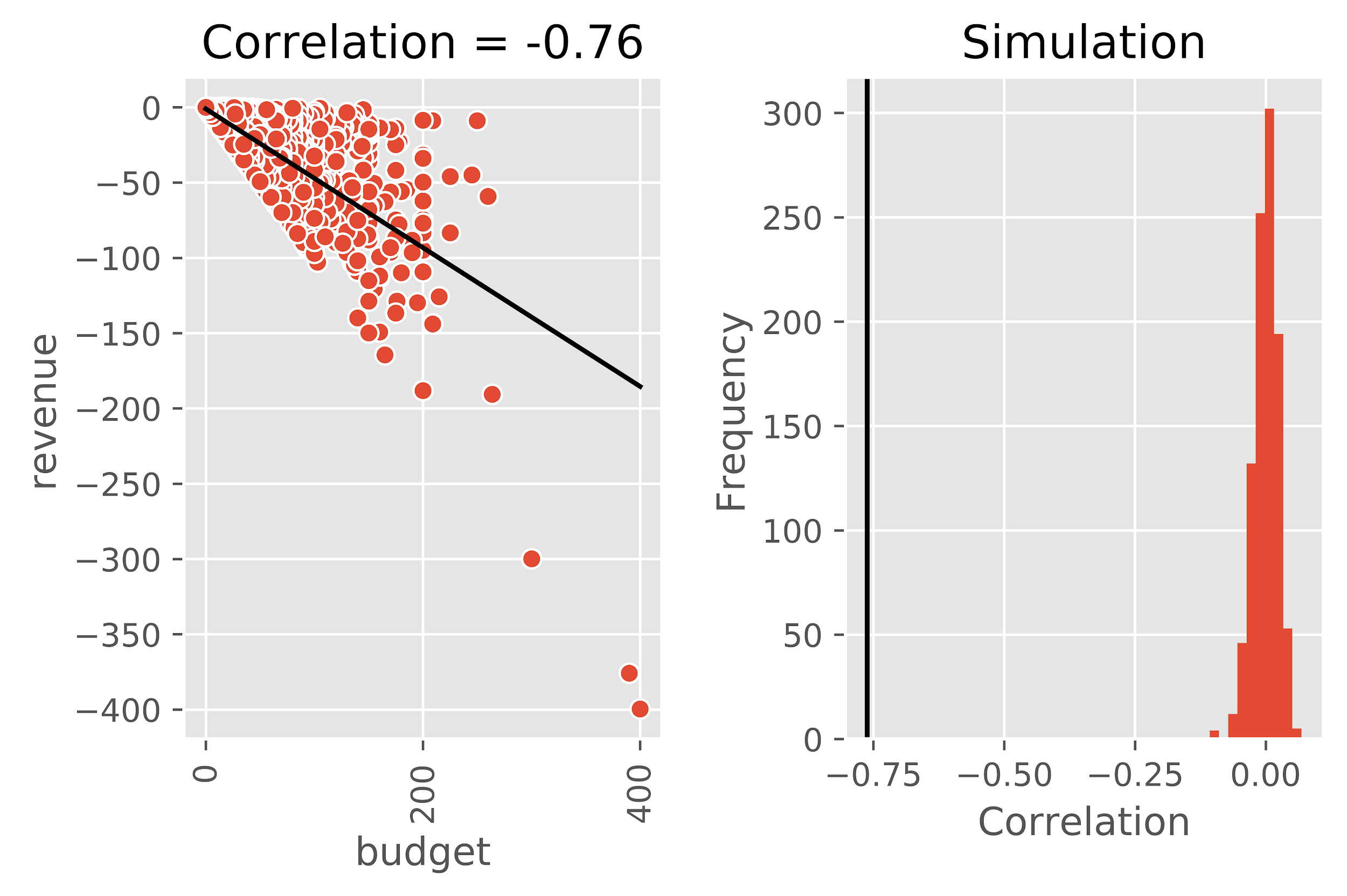
**[IMDB score]**

The correlation between imdb\_score and revenue is 0.24 (left scatterplot). In order to check whether this correlation is significant, I compare it (black vertical line) to the simulated null hypothesis (red histogram) in the right histogram. The plot shows that the simulated correlations are not as extreme as 0.24 (p-value < 0.05) so the null hypothesis is rejected.

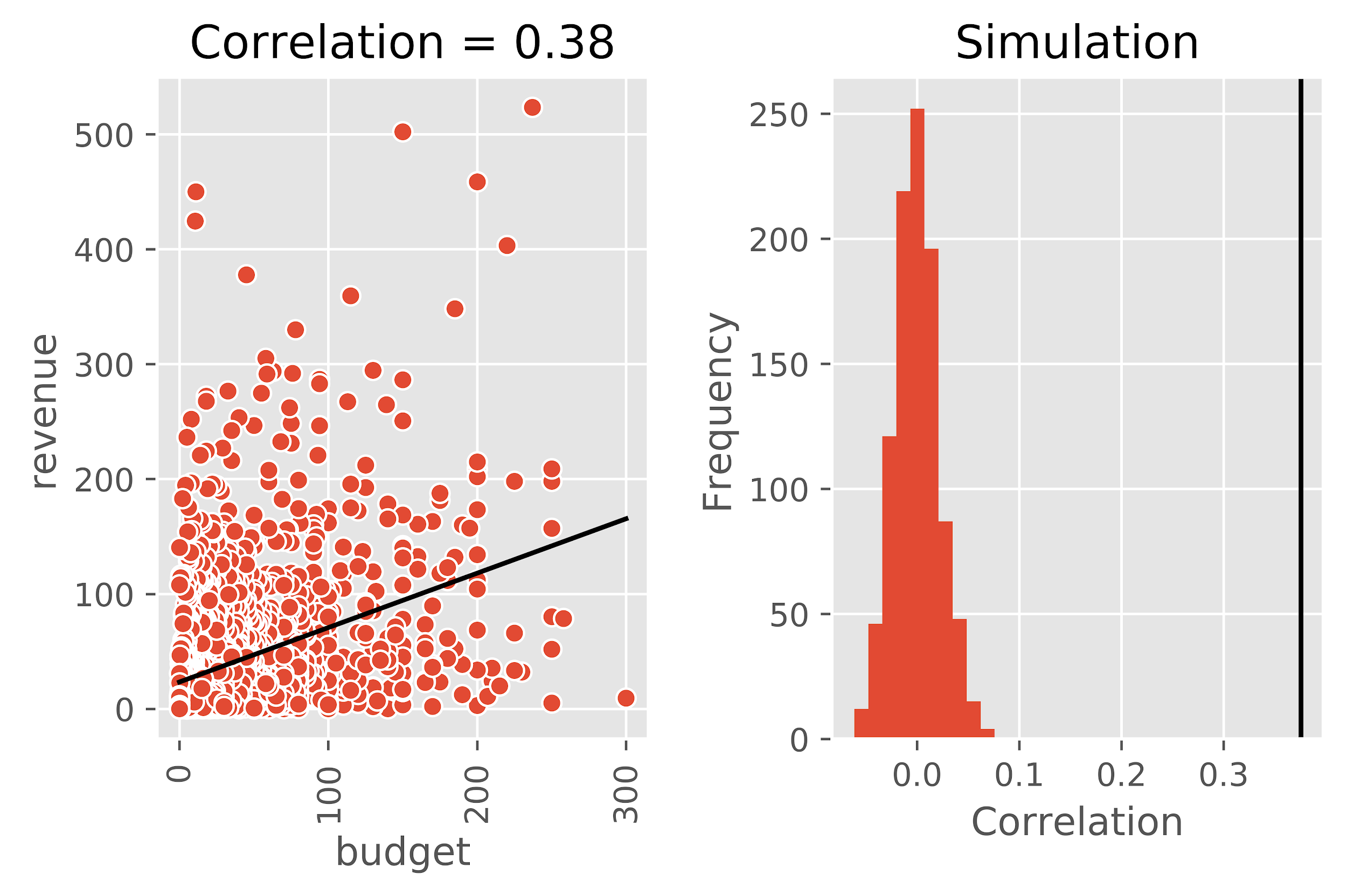
**[Gross and Budget]**

It makes sense that both of them are highly related to the revenue since it is calculated from gross and budget. However, there is an interesting pattern in the budget and revenue that it seems to show a negative correlation. 

I further separate the samples population into two parts: the positive revenue and the negative revenue. Then I compare them to the budget, respectively. The result shows that there is a ~76% correlation between revenue and budget in movies with a negative revenue.



On the other hand, there is a ~35% correlation between revenue and budget in movies with a positive revenue. Both of them are statistically significant. I would conclude that it is not necessary to make more money when investors invest more money. The budget has to be in an adequate range for the movie's success (HIGH REVENUE).



**[Directors and actors name frequency]**

Among ~4000 movies, it is making sense to think that popular actors or directors can contribute a higher movie revenue and popularity is related to how often they are in the movies. However, there are no obvious correlations based on the scatter plots.

**[Facebook likes]**

The correlation between different types of Facebooks likes and movie revenue is around 0.11 to 0.23. It is not a really high correlation, but they are all significant in the hypothesis tests. Based on here, it would not be a not great indicator as I thought. It is possible that Facebook likes only show how you care about a movie instead of actually watching it. It might also interact with other variables.

**[Critics and voted\_users]**

This category is similar to Facebook likes category and shows how people care about certain movies. However, the hypothesis test shows that correlations are high (0.24 - 0.49) and significant. I presumed that this category is more representative than the social media because it requires devoted people to actually leaving a review or voting online.

**[Others]**

Here includes duration, aspect ratio, and poster’s face-number. Apparently, the

correlation is low and not significant.

**[Movie information]**

I include color, language, and country variables with 0 and 1 codes.

Color variable: 0 is black-and-white movies; 1 is color movies

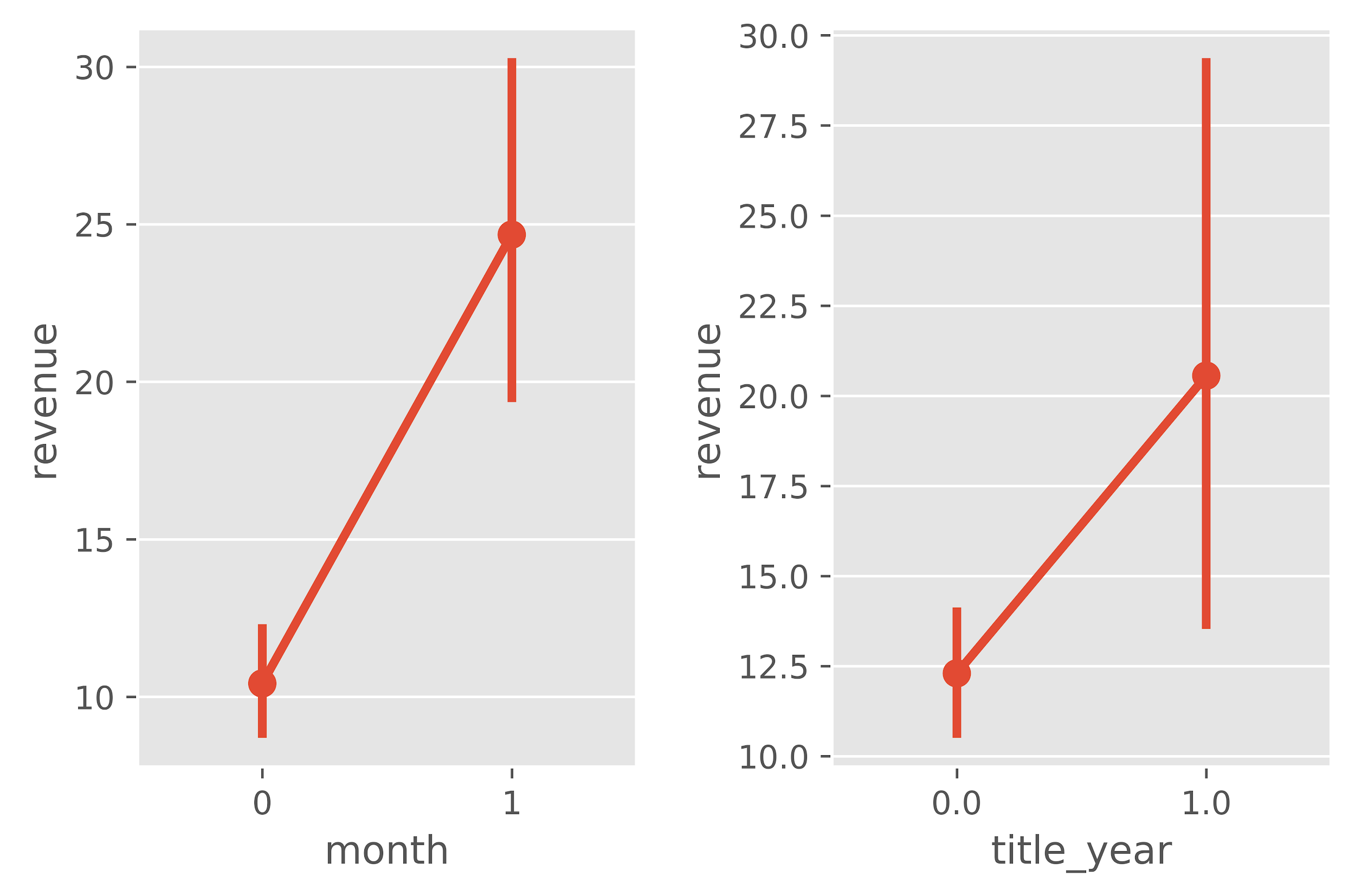
Language : 0 is Non-English; 1 is English

Country: 0 is movies published by non-USA; 1 is movies published by USA

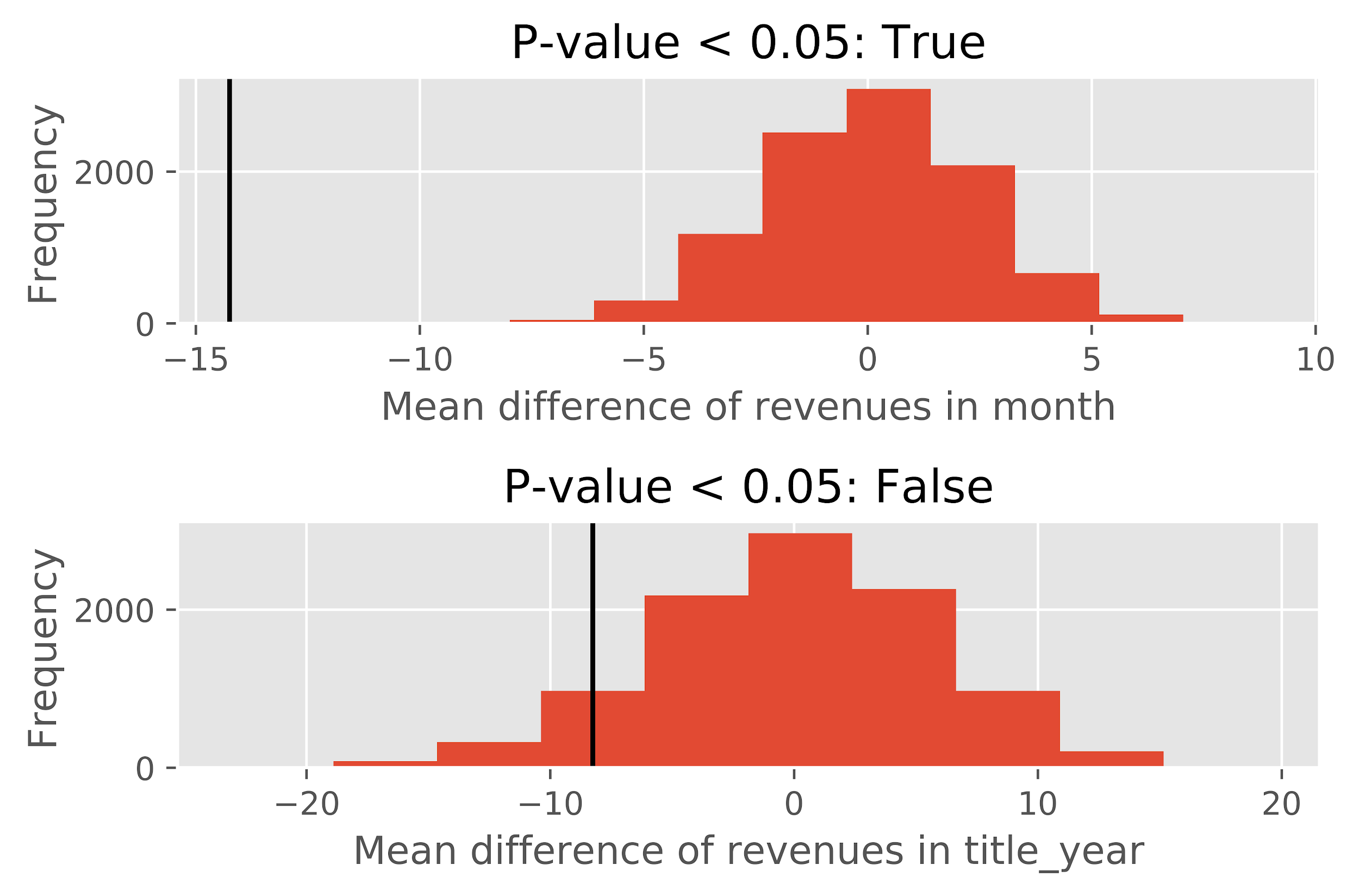
First, I check their distribution by box plots. The distribution is showing that there are more black and white (0) movies with higher revenues, more movies in English with higher revenues, and more movies produced by the United States with higher revenues. Then, I use point plots to show the mean and confidence interval. Interestingly, the mean revenue of black and white movies is higher than the mean revenue of color movies, but the confidence interval is wide. It is hard to tell whether it is significant or not. I use permutated hypothesis tests to prove the significance in all of them.

**[Time]**

Here I want to check the effect of the seasonality on movie revenue. For the month variable, 0 represents non-June or December and 1 represents Jun or December. For the title\_year variable, 0 represents movies after 1966 and 1 represents movies before 1966. Although there is no obvious difference in the distribution, I find out the mean differences in revenue are quite different in both month variable and title\_year variable (Points plots below).



In the permuted hypothesis tests, the result even shows that mean difference of revenue in the month variable is significantly different but not title\_year variable. It is quite reasonable that there will be more people in movie theaters in June and December (month variable: 1).



**[Genres]**

Following previous steps, I check each genre’s distribution in box plots and mean with the confidence interval in point plots. Then, I calculate the p-values of the genre’s mean difference and collect the group with p < 0.05. The result shows that PG-13, R, Adventure, Animation, Comedy, Crime, Drama, Family, Fantasy, History, Thriller, and War are all significantly different in mean difference. It means they can represent a lower or higher movie revenue for the future prediction.

1. Remove an unnamed column, budget, and gross columns
2. Save the file after EDA: final\_eda.csv
3. **Results and machine learning**

The most straightforward idea is to predict positive or negative revenue with a regression model. I began with the simplest linear regression model. However, the accuracy score is not really great. Before I dive into more linear regression tuning, I decided to use the classification model to test whether it is the model’s problem or the dataset’s problem.

To prepare the dataset for the classification model, I separated the revenue (target variable) column into two groups (1: positive revenue; 0: negative revenue). Then, I began to test the IMDB movie dataset with the following classification tools.

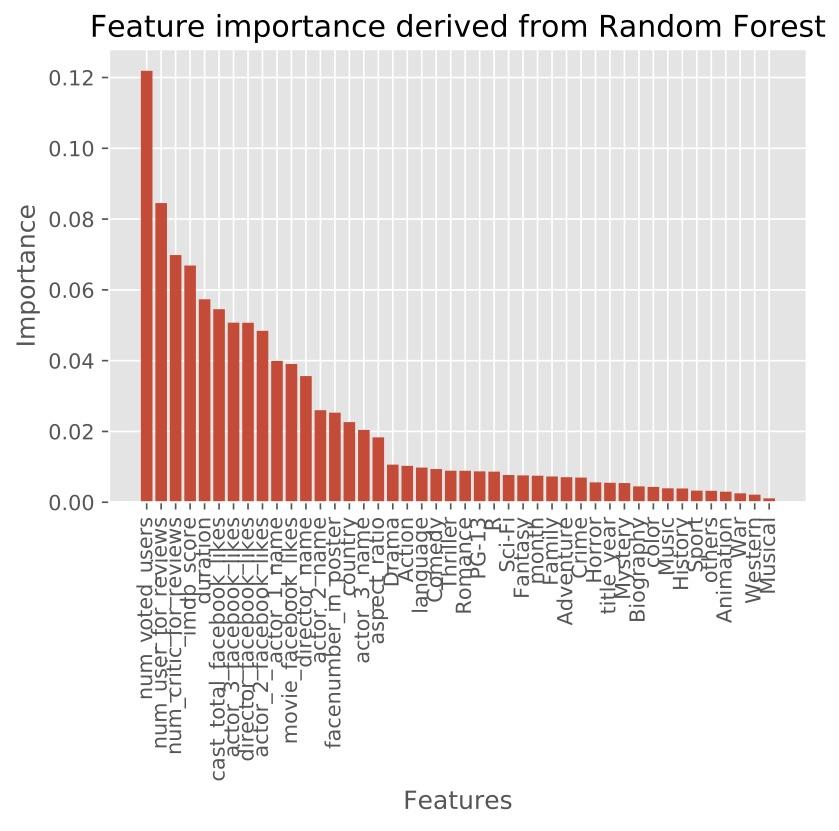
1. Linear regression

I tried a basic linear regression and got a pretty low accuracy score of 0.32. Therefore, I am going to use the classification model to deal with this data.

1. Prepare data for clasification

Separate target variable into two groups -> Split training and testing dataset -> Standardize training datasets.

1. Check feature importance by Randomforest classifier



1. Classification

**[Logistic regression model]**

In order to get an idea of the accuracy rate, I start with the most basic classification model, logistic regression, for my problem. I got an accuracy score around 0.688. Now, I wanted to improve the score by hyperparameter tuning C. It represents the inverse of regularization strength (Smaller = Stronger regularization). The result shows that C=1 can improve the accuracy to 0.689, although the improvement is not really dramatic.

**[Logistic regression with polynomial features]**

Polynomial features change hyperparameter C from 1 to 0.01, but it does not improve the accuracy score.

**[SVM + classification]**

Another great tool for classification model is the support vector machine for classification (SVC). It also can be applied with Gaussian kernel function for non-linear classification. Therefore, I can confirm whether the data points require a higher dimensional separation. The same strategy here is to tune different C hyperparameter and check the accuracy score. The result shows that C=0.1 gives the best score of 0.698. Apparently, the score is not really satisfying even it is a bit higher than the logistic regression model. Next step is to use SVC with kernel function by tuning C and gamma hyperparameters. Gamma hyperparameter is a cut-off parameter for the boundary of Gaussian sphere (larger = tighter boundary; also can be overfitting). Unfortunately, the result shows a worse accuracy score instead of a better one. It only proves that the issue is not about the dimensions.

**[A Lazy learner: KNN]**

Next approach that I want to try is the K-nearest neighbors (KNN), which is fundamentally different from other learning algorithms. KNN belongs to a subcategory of nonparametric models called instance-based learning. Maybe it is a better approach from a non-classical algorithm. Manhattan and Euclidean are two major distance metrics in SciKit-Learn’s KNN model. I tried both of them and the result shows that the accuracy score of Manhatten metric is 0.678 higher than the score analyzed by Euclidean.

**[Ensemble methods]**

The last but the least trial is the famous ensemble learnings, Random forest (RF). It is a robust model based on decision tree and free of hyperparameter tuning. The randomness from the bootstrapped samples and the number of features has counteracted to the noises and variances. Before I dig into RF, I tried the basic decision tree model out of curiosity and compare it to a raw RF model. The accuracy of the decision tree is 0.636 and of the RF model is 0.716. A further strategy is to use RandomizedSearchCV to get a base of parameter combination first. Then, I will apply the result to GridSearchCV with the different combinations as my second search. However, accuracy is still similar to 0.716.

Gradient boosting also belongs to ensemble learnings. However, it uses a boosting technique instead of bagging technique in the RF model. The difference is that the gradient boosting will start with a weak predictor and sequentially improve it to a better predictor. A basic model without any tuning shows a lower accuracy score (0.69) compared to the score from the RF model. With GridSearchCV strategy, I improve the score to a similar score (0.714)

1. Conclusion

Let us think back the questions I want to ask in my problem statement and check whether we know the answer from my analysis.

* Can famous actors, director, or both bring the highest revenue?

Here I define famous actors and directors by their name frequency showed up in the movies between 1906 and 2006. I would say yes to this question since actor\_1\_name, actor\_2\_name, and director name are all in the top 20 of feature importance plot. It means that they all contribute to the movie revenue prediction.

* How social media and networking service company ex: Facebook influence the movie industry?

In the feature importance plot, social network-related features are all in top 15. More importantly, critics and voter are the most important indicators of the movie’s revenue. They are top 3 important feature for the prediction. Also, they all pass the hypothesis testing significantly.

* Is a high IMDB score a good sign for movie companies that they are gonna make money for sure?

IMDB score is no doubt a great indicator of movie revenue prediction. The rank is 4 in the feature importance plot and it significantly correlates to movie’s revenue.

* Is it guaranteed that high budget will lead to high revenue?

Not really! From the correlation plot, it even shows a high budget might lead to a lower movie’s revenue.

* Which genre is the revenue promise?

In my analysis, it is not so significant for differentiating the movie’s revenue. Although some of them pass the hypothesis tests, all of them are evaluated as low important features in the prediction.

* Is it important to pick a good timing (month) to the movie’s announcement?

YES and we all know that already. I still like to use statistics and machine learning to prove this point. However, it is not counted as a great feature for revenue prediction even the mean difference between the popular season and the non-popular season is significantly different.

In conclusion, a great feature for future prediction might not be the same from a statistic based on the previous history. There is a lot of correlations between different features that might produce leaking effect and they are hard to find out.