DSCI 510 - Final Project

Correlation Between Movie Budgets and Ratings

Description

This project aims to find out the correlation between movie budgets and ratings, i.e. whether and how a movie's budget affects its ratings. The project performs analysis on a dataset containing budgets and Metascores of 777 movies from 2012 to 2015 collected from Kaggle and Rapid API. To examine the impact of budget on Metascore, this project will compare the Metascores of large-budget and small-budget movies, and build a simple linear regression model to determine the relationship between these two variables.

Dependencies

pandas
requests
json
matplotlib.pyplot
seaborn
numpy
statsmodels.formula.api
statsmodels.stats.outliers_influence
OLSInfluence

Note: There is no specific required version of each library.

Running the Project

My code is a Jupyter notebook, and it can be run using Cell->Run All.

Github Repository

Here is the link to my Github repository:
https://github.com/SiweiC/final_project_siwei_cheng

Data Collection

The final dataset was collected from two data source and contains IMDB ID, budget and Metasocre of 777 movies from 2012-2015. I saved the final dataset as “budget_metascore.csv" in the ‘data’ folder, which has 777 rows and 3 columns.

I. Movies (2012-2015) Budget
Data Source:
Kaggle: https://www.kaggle.com/datasets/juzershakir/tmdb-movies-dataset

I downloaded the TMDb Movies Dataset from Kaggle, which is in CSV format, from Kaggle and saved it as “tmdb_movies_data.csv” in the ‘data’ folder. The original dataset contains 21 attribtues of 10866 movies, but for this project I only selected imdb_id and budget attributes of movies from 2012 to 2015, and dropped rows with null value of budget.

II. Movies (2012-2015) Rating
Data Source:
RapidAPI: https://rapidapi.com/rapidapi/api/movie-database-alternative

I used the imbd_id collected from the first data source to collect the corresponding Metascore from the Movie Database Alternative API from RapidAPI and created a Dataframe that contains imbd_id, budget, and Metascore. Then, I dropped rows with null value of Metascore and saved the Dataframe as “budget_metascore.csv” in the ‘data’ folder, which is the final dataset that I used in performing following analyses and visualization.

budget_metascore

	imdb_id	budget	Metascore
0	tt0369610	150000000	59.0
1	tt1392190	150000000	90.0
2	tt2908446	110000000	42.0
3	tt2488496	200000000	80.0
4	tt2820852	190000000	67.0
5	tt1663202	135000000	76.0
6	tt1340138	155000000	38.0
7	tt3659388	108000000	80.0
8	tt2293640	74000000	56.0
9	tt2096673	175000000	94.0
10	tt2379713	245000000	60.0

Sample of the final dataset

Changes from original plan:
In the original plan, I did not take into account that there are movies whose Metascores cannot be fetched from the API. Therefore, when processing the data, I had to drop movies with empty Metascore.

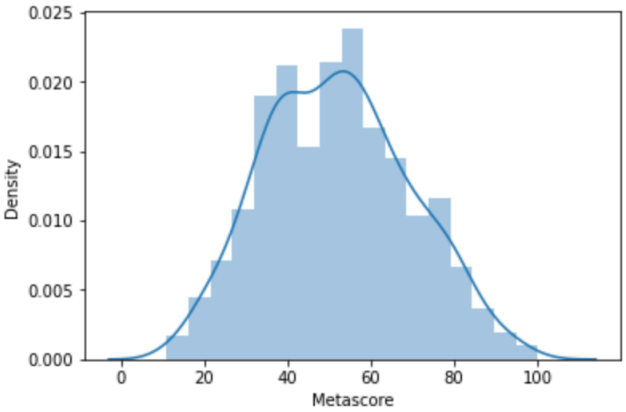
Analyses and Visualization

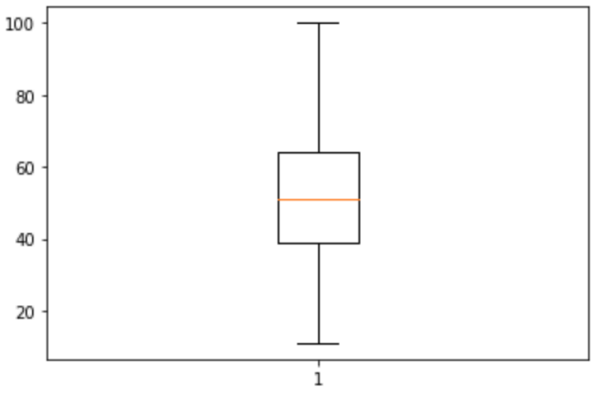
- **Summary of the whole dataset**

I find the mean, median, min, max, first and third quartiles, and standard deviation of budgets and Metascores. It can seen from the data frame below that Metascore values range from 11 to 100, and the median budget is about $20 million.

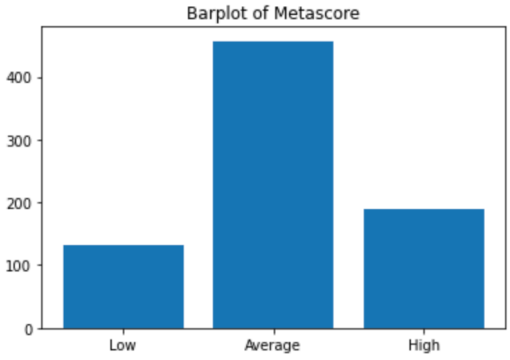
	budget	Metascore
count	7.770000e+02	777.000000
mean	4.104488e+07	51.927928
std	5.260989e+07	17.552864
min	1.000000e+01	11.000000
25%	7.700000e+06	39.000000
50%	2.000000e+07	51.000000
75%	5.000000e+07	64.000000
max	2.800000e+08	100.000000

I plot histogram, densities and box plot of Metascores. It can be seen from the figures below that the Metascore is close to normal distribution, such that most movies are rated between 35 and 65 and few movies are rated extremely high or low.





According to the distribution of Metascores, I divide movies' ratings into 3 levels: ratings from 0 to 34 are 'low', ratings from 35 to 64 are 'average', ratings from 65 to 100 are 'high'. Then, I plot a bar chart based on the 3 levels. There are 131 movies have low Metascores, 457 have average Metascores, and 189 movies have high Metascores.



	class	count
0	Low	131
1	Average	457
2	High	189

- **Compare Metascores of 10 largest-budget films and 10 smallest-budget films**

I find out 10 largest-budget films and 10 smallest-budget films from the dataset.

10 largest-budget movies

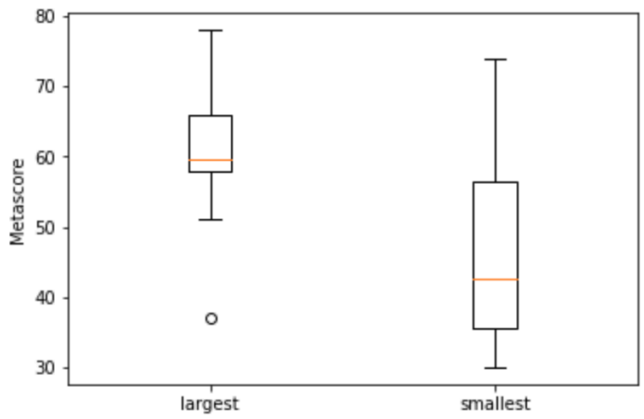
	imdb_id	budget	Metascore	class
14	tt2395427	280000000	66.0	High
432	tt0401729	260000000	51.0	Average
641	tt1210819	255000000	37.0	Average
194	tt2310332	250000000	59.0	Average
203	tt1877832	250000000	75.0	High
388	tt1345836	250000000	78.0	High
392	tt0903624	250000000	58.0	Average
571	tt1170358	250000000	66.0	High
10	tt2379713	245000000	60.0	Average
406	tt1409024	225000000	58.0	Average

10 smallest-budget movies

	imdb_id	budget	Metascore	class
366	tt2556874	10	32.0	Low
758	tt3074732	10	74.0	High
748	tt1828959	15	40.0	Average
769	tt1699755	15	52.0	Average
543	tt1710396	25	44.0	Average
766	tt2991296	89	34.0	Low
561	tt2129928	110	66.0	High
775	tt2187884	650	58.0	Average
554	tt2149360	8000	41.0	Average
162	tt2309260	100000	30.0	Low

Among the 10 largest-budget films, 4 films have high Metascores and no films have low Metascores, while among the 10 smallest_budget films, there are 3 films have low Metascores. This observation suggests that a sufficient budget may reduce the likelihood of a movie getting a low rating. Comparing the ratings of the ten movies with the biggest budgets and the ten movies with the smallest budgets

I also plot a boxplot to show the difference in the distribution of Metascores between large-budget movies and small-budget movies.



- **Simple Linear Regression Model**

I build a simple linear regression model to find out the correlation between budget and Metascore.

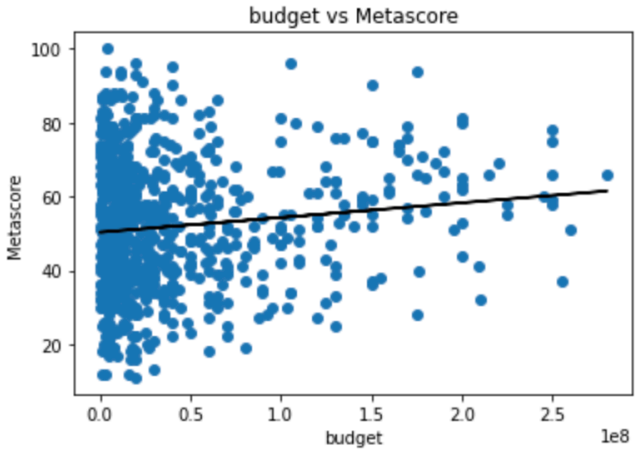
```
                            OLS Regression Results
==============================================================================
Dep. Variable:              Metascore   R-squared:                       0.014
Model:                            OLS   Adj. R-squared:                  0.013
Method:                 Least Squares   F-statistic:                     11.19
Date:                Wed, 07 Dec 2022   Prob (F-statistic):           0.000861
Time:                        12:47:17   Log-Likelihood:                -3322.7
No. Observations:                 777   AIC:                             6649.
Df Residuals:                     775   BIC:                             6659.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     50.2939      0.794     63.368      0.000      48.736      51.852
budget      3.981e-08   1.19e-08      3.346      0.001    1.65e-08    6.32e-08
==============================================================================
Omnibus:                       17.016   Durbin-Watson:                   1.922
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               13.015
Skew:                           0.221   Prob(JB):                      0.00149
Kurtosis:                       2.546   Cond. No.                     8.46e+07
==============================================================================
```

In this model, the p-value (0.001) of budget is quite small (less than 0.05) which means that there is a statistically significant association between the value of budget and Metascore. In addition, it can be seen from the positive coefficient (3.981e-08) of budget that there is a linear relationship between budget and Metascore, such that the value of Metascore increases as the value of budget increases. However, the pearson correlation between budget and Metascore is about 0.12, which suggests that the relationship between these two variables is not strong.

```
# calculate pearsons correlation between Metascore and budget
corr = np.corrcoef(dataset['Metascore'], dataset['budget'])[0, 1]
print(f'The Pearson correlation betweeen Metascore and Budget is {corr}')
```

```
The Pearson correlation betweeen Metascore and Budget is 0.1193190880712784
```

The plot below shows how the linear regression model fits the dataset.



Conclusion

In conclusion, movie's budget and Metascore are positive correlated, such that a movie with large budget tend to have higher rating than the movie with small budget. However, since the pearson correlation between budget and Metascore is only 0.12, the impact of budget on Metascore is not significant.

Future Work

Given more time, I hope to take this project further in two directions. First of all, the dataset I use now only contains movie data from 2012 to 2015. However, from 2015 to now, the movie industry has developed rapidly. Thus, in order for my analysis to better reflect the current state of the film market, I plan to collect more recent data in the future. Secondly, at the current stage, I only analyzed the correlation between budget and rating. In the future, I hope to expand my project by adding more factors that may affect ratings for analysis and modeling, such as movie duration, movie category, and so on.