TDS 2101 FUNDAMENTAL DATA SCIENCE

Individual Project

High Revenue Movie Prediction Using Linear Regression as an Algorithm

NUR ANIS NABILA BT MOHD ROMZI

(1211303587)

(MDM SYUHAIDAH AZNI)

Table of Contents:

1.0 Project Description

As a movie producer, we provide money to make movies and when the movies make money, we make money and when the movies lose money, we lose money. Imagine, we have done everything including sold our car or mortgaged our house to fund this film, is it worth it? Hence, we need to predict the future. We can measure the success of the movie depending on the actors or the script but overall, budget is maybe the most related because we need to pay the good actors, special effects and marketing. Therefore, my project is to investigate the association of movie budgets that develop high revenue.

2.0 Objectives

- To determine the relationship between movie budget and the revenue
- Analysing train dataset using Linear Regression as an algorithm
- Understanding statistical analysis to support algorithm used

3.0 Expected Outputs

- Provide result from the algorithm used
- Accept the hypothesis which is movie budget related the most in gaining profit for the film.

4.0 Exploratory Data Analysis

4.1 Dataset

	Release Date	Movie	Production Budget	Domestic Gross	Worldwide Gross
1	Dec 16, 2022	Avatar: The Way of Water	$460,000,000	$684,075,767	$2,320,003,887
2	Apr 26, 2019	Avengers: Endgame	$400,000,000	$858,373,000	$2,794,731,755
3	May 20, 2011	Pirates of the Caribbean: On Stranger Tides	$379,000,000	$241,071,802	$1,045,713,802
4	May 1, 2015	Avengers: Age of Ultron	$365,000,000	$459,005,868	$1,395,316,979
5	May 19, 2023	Fast X	$340,000,000	$145,084,410	$712,161,071
6	Dec 18, 2015	Star Wars Ep. VII: The Force Awakens	$306,000,000	$936,662,225	$2,064,615,817
7	May 24, 2007	Pirates of the Caribbean: At World's End	$300,000,000	$309,420,425	$960,996,492
8	Jun 30, 2023	Indiana Jones and the Dial of Destiny	$300,000,000	$7,200,000	$7,602,944
9	Nov 6, 2015	Spectre	$300,000,000	$200,074,175	$879,077,344
10	Apr 27, 2018	Avengers: Infinity War	$300,000,000	$678,815,482	$2,048,359,754
11	Nov 17, 2017	Justice League	$300,000,000	$229,024,295	$655,945,209
12	Jul 12, 2023	Mission: Impossible Dead Reckoning Part One	$290,000,000	$0	$0
13	Dec 20, 2019	Star Wars: The Rise of Skywalker	$275,000,000	$515,202,542	$1,072,767,997
14	May 25, 2018	Solo: A Star Wars Story	$275,000,000	$213,767,512	$393,151,347
15	Mar 9, 2012	John Carter	$263,700,000	$73,058,679	$282,778,100
16	Mar 25, 2016	Batman v Superman: Dawn of Justice	$263,000,000	$330,360,194	$872,395,091
17	Dec 15, 2017	Star Wars Ep. VIII: The Last Jedi	$262,000,000	$620,181,382	$1,331,635,141
18	Nov 24, 2010	Tangled	$260,000,000	$200,821,936	$583,777,242
19	Jul 19, 2019	The Lion King	$260,000,000	$543,638,043	$1,647,733,638
20	May 4, 2007	Spider-Man 3	$258,000,000	$336,530,303	$894,860,230

figure 1 : Dataset from website The Numbers

Rank	Release Date	Movie Title	Production Budget ($)	Worldwide Gross ($)	Domestic Gross ($)
5293	8/2/1915	The Birth of a Nation	$110,000	$11,000,000	$10,000,000
5140	5/9/1916	Intolerance	$385,907	$0	$0
5230	12/24/1916	20,000 Leagues Under the Sea	$200,000	$8,000,000	$8,000,000
5299	9/17/1920	Over the Hill to the Poorhouse	$100,000	$3,000,000	$3,000,000
5222	1/1/1925	The Big Parade	$245,000	$22,000,000	$11,000,000
4250	12/30/1925	Ben-Hur	$3,900,000	$9,000,000	$9,000,000
4630	12/8/1927	Wings	$2,000,000	$0	$0
5141	1/2/1929	The Broadway Melody	$379,000	$4,358,000	$2,800,000
4240	1/1/1930	Hell's Angels	$4,000,000	$0	$0
5043	12/31/1931	Mata Hari	$558,000	$900,000	$900,000
5017	7/4/1933	King Kong	$672,000	$10,000,650	$10,000,000
5234	9/2/1933	She Done Him Wrong	$200,000	$2,000,000	$2,000,000
5120	9/3/1933	42nd Street	$439,000	$2,281,000	$1,438,000
5154	1/1/1934	It Happened One Night	$325,000	$2,500,000	$2,500,000
5025	6/9/1935	Top Hat	$609,000	$3,202,000	$1,782,000
4738	5/2/1936	Modern Times	$1,500,000	$165,049	$163,245
4768	6/26/1936	San Francisco	$1,300,000	$5,273,000	$2,868,000
4814	10/20/1936	Charge of the Light Brigade, The	$1,200,000	$0	$0
4756	12/21/1937	Snow White and the Seven Dwarfs	$1,488,000	$184,925,485	$184,925,485
4570	1/1/1938	Alexander's Ragtime Band	$2,000,000	$4,000,000	$4,000,000
4693	1/1/1938	You Can't Take It With You	$1,644,000	$4,000,000	$4,000,000

figure 2 : Dataset uploaded in csv file before cleaning

	A	B	C
1	Movie Title	production_budget_usd	worldwide_gross_usd
2	American Hero	1000000	26
3	The Rise and Fall of Miss Thang	10000	401
4	The Dark Hours	400000	423
5	Destiny	750000	450
6	Bang	10000	527
7	Ed and his Dead Mother	1800000	673
8	The Jimmy Show	1000000	703
9	Perrierâ€™s Bounty	6600000	828
10	In Her Line of Fire	1000000	884
11	The Mongol King	7000	900
12	To Be Frank, Sinatra at 100	2000000	926
13	Childless	1000000	1036
14	B-Girl	700000	1160
15	The Trials of Darryl Hunt	200000	1217
16	Skin Trade	9000000	1242
17	Grip: A Criminal's Story	12000	1336
18	Return to the Land of Wonders	5000	1338
19	Keeping it Real: The Adventures of Greg Walloch	100000	1358
20	Eddie: The Sleepwalking Cannibal	1400000	1632
21	The Looking Glass	300000	1711
22	Detention of the Dead	500000	1778

figure 3 : Dataset after cleaning in csv file

4.2 Data Cleaning

As soon as I upload the data, I have done a few things before modelling.

- **Delete Missing values**

  To maintain the integrity and quality of the dataset, I found and deleted approximately 400 from 5400 rows that do not have complete attributes. Deleting missing values could prevent bias that may arise from the overall dataset.

- **Remove unnecessary columns**

  For my predictive model, I decided to find what contributed to the higher revenue and I found the most associated with my objectives are only production budget and worldwide gross columns so I dropped the remaining columns. I also included title columns for me not getting confused while reading the dataset.

- **Data Normalisation**

  Machine learning algorithm that I'm going to build is designed to work with statistical and numerical datasets. Hence, I removed the special characters such as dollar sign and comma. This also ensures consistency and standardisation while I train the data.

4.3 Data Visualisation

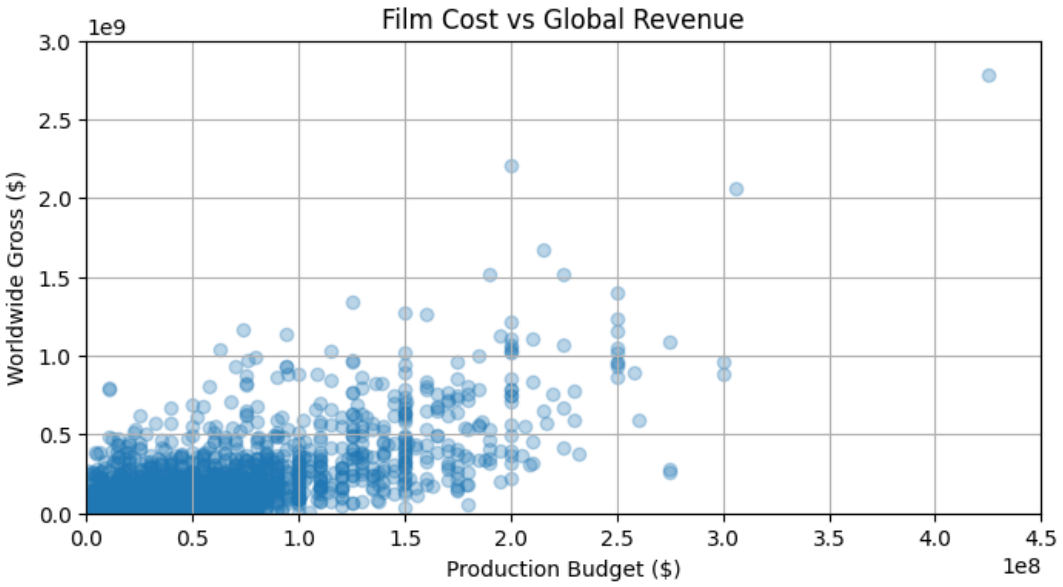

figure 4 : Scatter plot showing trends between production budget vs worldwide gross

From the plot, it shows an upward trend where we can simplify higher production budgets resulting in higher revenue. The outliers above also made an enormous production cost and made a lot of money in return. The title of the movie is Avatar. From here, we also can understand the potential relationship between movie budget and its revenue. Most films actually spent less than one million in making a film.

4.4 Relationship between variables in heatmap

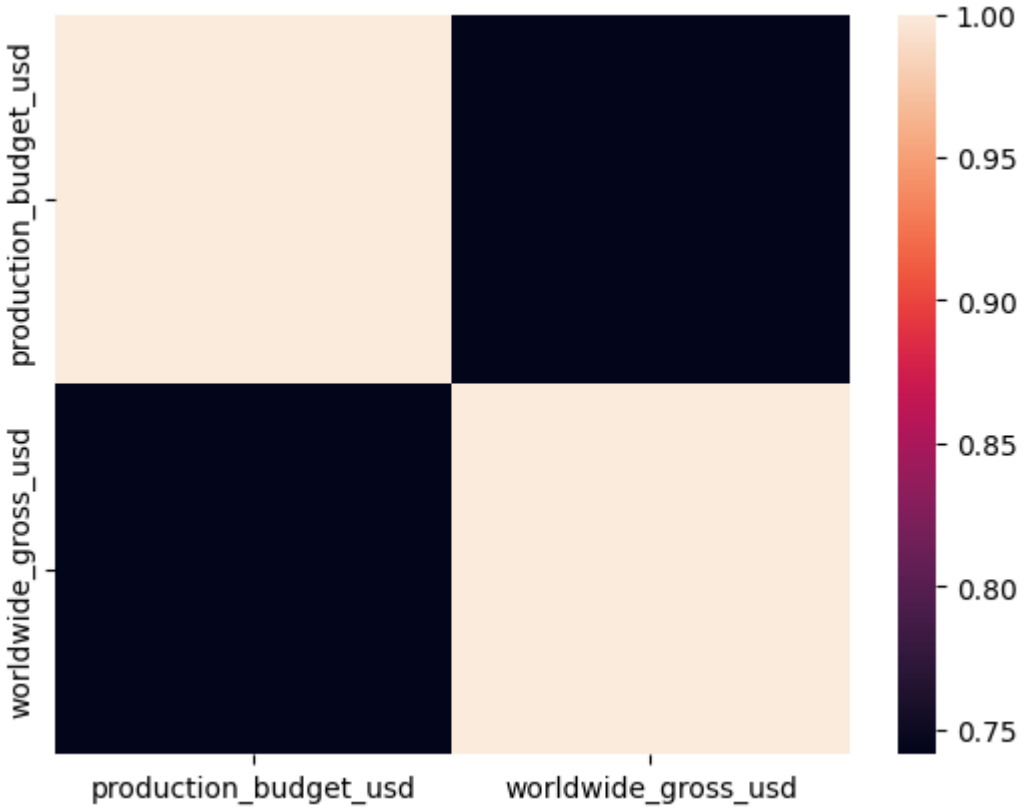

figure 5: relationship between production budget and worldwide gross

	production_budget_usd	worldwide_gross_usd
production_budget_usd	1.000000	0.741383
worldwide_gross_usd	0.741383	1.000000

figure 6: correlation between production budget and worldwide gross

The correlation between "production_budget_usd" and "worldwide_gross_usd" is 0.741383. This indicates a moderately strong positive correlation between the two variables. As the production budget increases, there is a tendency for the worldwide gross to also increase. The value of 0.741383 suggests a fairly strong positive linear relationship between the production budget and worldwide gross. From the heatmap, we can see the strength of the correlation between the two variables mentioned. We clearly observed that the warm colour in production budget indicates the lesser the budget, the lesser revenue and while the colour is becoming darker it represents the higher the production budget, the higher the movie revenue.

5.0 Feature Selection

5.1 Univariate feature selection by using ANOVA Test

Feature selection for univariate feature selection used in this project is by using ANOVA Test.

Null Hypothesis (Ho): There is no significant difference between mean of production budget and worldwide gross.

Alternative Hypothesis (Ha): There is a difference between the mean of production budget and worldwide gross.

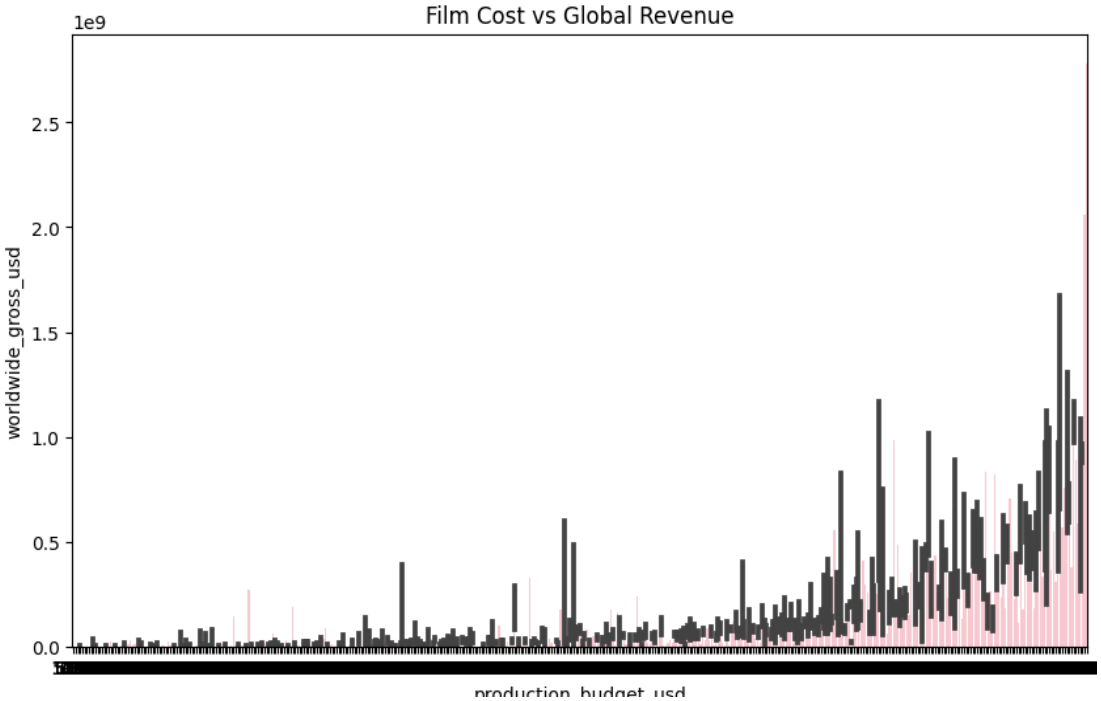

figure 7: Error bars of two group overlapping

Error bars of two groups in a bar plot below overlap by more than 25%, thus, it suggests that the difference between the means or values represented by the bars is not statistically significant.

Result obtained for F-statistic measured that ratio of variation between production budget and worldwide gross is 619.597

For p-value, we obtain 8.97e-133 which is near to zero.

Thus, since p-value is near to zero, it suggests strong evidence that Ho is true where there is no significant difference between mean of production budget and worldwide gross.

Therefore, the ANOVA test result indicates strong evidence that the independent variables(production budget) is significantly related to the dependent variable(worldwide gross).

6.0 Model Construction and Comparison

From our conclusion above, we need a proof to support our hypothesis that says a strong positive linear relationship between the production budget and worldwide gross. Hence, I build a predictive model using linear regression as an algorithm.

6.1 Using Linear Regression as an Algorithm

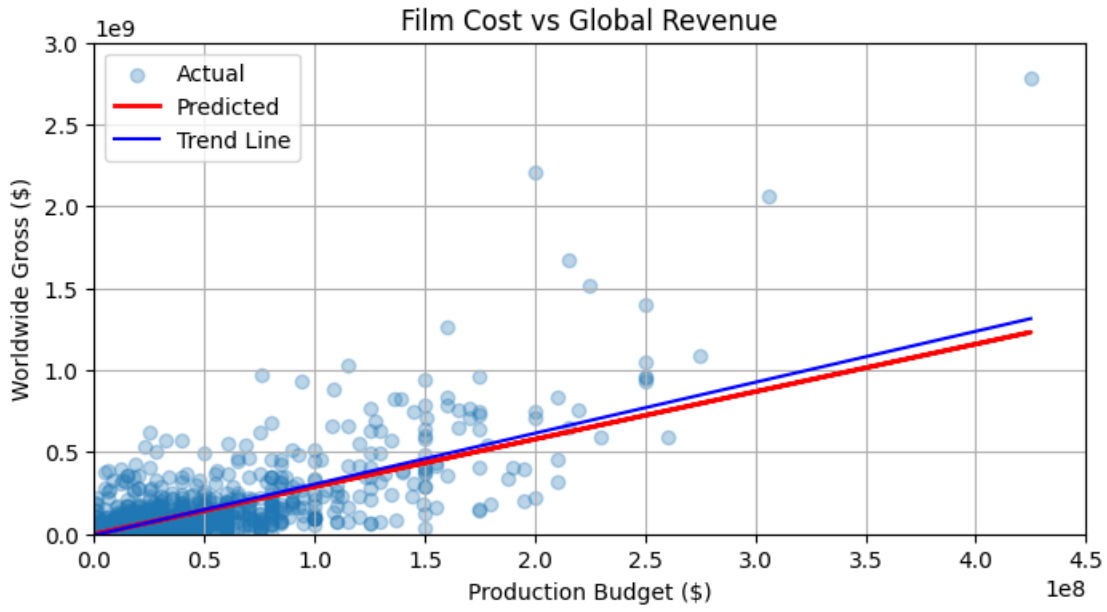

figure 8: line predicted using linear regression prediction

For the algorithm, I do Train-Test Split where I split the data into a training set and a separate test set. I also train the linear regression model using the training set and evaluate its performance on the test set using appropriate metrics such as mean squared error (MSE), mean absolute error (MAE), or R-squared. The result I got for mse is 1.696816032823077e+16 and R-squared is 0.577403562365628.

In the regression models above, a predicted line is a straight line that represents the estimated relationship between the independent variable(production budget) and the dependent variable(worldwide gross). The predicted line obtained by fitting the model to the training data and then using the trained model to predict the target variable for new or unseen data points. The trend line visualises the general direction or relationship between variables. Trend lines are in the graph used to identify long-term trends, patterns, or tendencies in the data.

6.2 Supporting algorithm by statistical calculation and analysis

For the slope coefficient, the result is 3.11150918. From the result obtained, positive coefficient means when production budget increases, the worldwide gross also increases. In the other word, the result shows a positive relationship between budget and revenue. Each dollar we spend on producing a movie, we should get around 3.1 dollar revenue in return.

Interception value from the line is -7236192.72913958 which means movies that have zero budget will lose over 7 million dollars. In the other words, the movie did not generate enough ticket sales or other sources of income to cover its production and marketing cost.

R-squared can help assess best fit in a regression model. The final result for R-squared is 0.5496485356985729.

Overall, the result suggests that the film's budget can explain approximately 55% of the variation observed in worldwide earnings. It indicates that there is a moderate relationship or association between the film's budget and its revenue.

However, it does not provide a comprehensive understanding of all factors that contribute to the film's earnings. Other variables and factors, such as marketing, genre, release date, and audience reception, may also play a significant role in determining the film's success.

7.0 Deployment

7.1 How to connect to the workspace

The purpose of my deployment is to predict the association between production budget and worldwide gross in overall budget specifically to the producer who wants to produce movies but is thinking about what is related the most in gaining high revenue for their projects.

For the deployment platform, you can use google colab or jupyter notebook to perform the project. All codes can be accessed in the “ Predict_Movie_Box.ipynb ” single file. You can run all the code in one time. For the dataset file, do use the “ cost_revenue_clean ” file before running the code as the dataset has been cleaned in the csv file earlier. Finally, you can see the plotting graph, how the algorithm works and expected outputs. It also can be downloaded easily and open online because it will automatically be saved in your google drive cloud.

7.2 Improvement project idea

In the upcoming project, I would love to explore what other attributes that may affect revenue of the movies. I'm thinking of including deep learning models such as neural networks because it can learn complex patterns and relationships in the data, making them suitable for capturing the nuances of movie revenue prediction. Lastly, I decided to discover the dataset by real-time prediction. For the visualisation, I would try to make it more interactive and useful for people to easily access and understand it.