MICROSOFT'S GENERATION OF MOVIES PROJECT.

1 BUSINESS UNDERSTANDING

Link to Jupyter Notebook:

Link to GitHub Repository:

1.1 UNDERSTANDING THE PROBLEM.

Microsoft wants to start making their own movies and it would like to make an informed decision before they begin the process of movie making. As is any business, the main aim is profit making, that means that we have to shun away from factors that might lead to losses. And to do that we need to scrutinize as much data as we can from the various potential competitors.

With the data that we have from various sources, we can analyze the data and make informed decisions.

1.1 PROBLEM STATEMENT

Microsoft is planning on expanding its borders by venturing into film making. They are very green in this area and therefore Microsoft needs to investigate which films are the most productive, that is they have large profit margins.

2 DATA UNDERSTANDING

2.1 DATA COLLECTION.

The data was collected from https://www.boxofficemojo.com/, TheMovieDB, and https://www.the-numbers.com/ .

2.2 DATA DESCRIPTION

Column	Description
Movie	The name of the movie
Production budget	The budget used for the production of the movie
Domestic gross	The amount of revenue generated by the movie in the country of origin
Worldwide gross	The amount of revenue generated by the movie globally
Original language	The language in which the movie was originally made.
Vote average	It is the average rating on a scale of 0 - 10
Vote count	The number of people who rated the movie
Release date	The date in which the movie was released.

2.3 DESCRIBING THE QUESTION.

2.3.1 SPECIFYING THE QUESTION.

Microsoft wants to venture into film making, but they know nothing about the film industry. Therefore, for them to start making films they need to know which films are going to be easy to penetrate in the market, and give the organization huge profit margins.

2.3.2 DEFINING THE METRIC FOR SUCCESS.

The movies with the highest profit margins and also average ratings.

2.3.3 EXPERIMENTAL DESIGN

- Loading Datasets
- Preparing the Data
- Cleaning data
- Exploratory Data Analysis
- Conclusions and Recommendation

3 DATA PREPARATION.

3.1 SELECTING DATA

We will use columns of the two csv files: tmdb.movies.csv and tn.movie_budgets.csv as they are relevant to the study.

3.2 DATA CLEANING.

This was done to ensure the Validity, Accuracy, Completeness, Consistency and Uniformity of the Data.

We checked for missing and null values and then we checked for duplicates. We found a couple of duplicates and decided to drop all the duplicates leaving the first value only.

We also dropped the ‘$’ and the ‘,’ from the domestic gross and production budget.

4 DATA ANALYSIS.

4.1 EXPLORATORY DATA ANALYSIS

- Categorical Data

Determine the columns with the highest values of domestic gross and production budget.

Determine the columns with the highest number of average ratings and also has high vote counts.

.

Summary statistics

	popularity	vote_average	vote_count	id_y
Count	2316.000000	2316.000000	2316.000000	2316.000000
Mean	10.685747	6.206261	1712.143351	50.842832
Min	0.600000	0.000000	1.000000	1.000000
Std	8.244311	1.178079	2704.578576	28.359866
25%	5.487000	5.600000	65.00000	27.000000
50%	9.506000	6.300000	615.000000	51.000000
75%	14.423750	6.900000	2100.500000	
max	80.773000	10.000000	22186.000000	100.000000

5 CONCLUSION.

Ratings do not really determine the revenue that a film might gain. Therefore, we should be able to make a conclusion just by how much revenue(profit margin) the film got.

6 RECOMMENDATION.

According to the movies with the highest revenue, Microsoft can produce movies similar to ‘Christopher Robin’ and ‘Olympus Has Fallen’ which have the highest domestic gross.

In terms of the profit margin, movies similar to ‘Black Panther’ , ‘Jurassic World’ and ‘Incredibles 2’ would be the best choice.