TMDB MOVIE DATASET

Table of Contents

A. Project Overview .................................................................................................................................... 2

A1. Research Question or Organizational Need........................................................................................... 2

A2. Context and Background ......................................................................................................................2

A3. Summary of Published Works............................................................................................................... 2

A4. Summary of Data Analytics Solution .................................................................................................... 3

A5. Benefit to Organization and Decision-Making Process......................................................................... 3

B. Data Analytics Plan................................................................................................................................. 3

B1. Goals, Objectives and Deliverables...................................................................................................... 3

B2. Scope of Project ....................................................................................................................................3

B3. Standard Methodology..........................................................................................................................4

B4. Timeline and Milestones…………….

B5. Resources and Costs..............................................................................................................................4

B6. Criteria for Success................................................................................................................................ 5

C. Design of Data Analytics Solution........................................................................................................ 5

C1. Hypothesis...........................................................................................................................................5

C2. Analytical Method................................................................................................................................ 5

C2a. Justification of Analytical Method..................................................................................................... 5

C3. Tools and Environments of Solution.................................................................................................... 5

C4. Methods and Metrics to Evaluate Statistical Significance ...................................................................6

C5. Practical Significance...........................................................................................................................6

C6. Visual Communication ........................................................................................................................ 6

D. Description of Datasets......................................................................................................................... 6

D1. Source of Data .....................................................................................................................................6

D2. Appropriateness of Dataset..................................................................................................................6

D3. Data Collection Methods..................................................................................................................... 7

D4. Data Quality......................................................................................................................................... 7

D5. Data Governance, Privacy and Security, Ethical, Legal, and Regulatory Compliance ....................... 7

D5a. Precautions.........................................................................................................................................7

E. Sources.................................................................................................................................................. 8

**A. Project Overview**

**A1. Research Question or Organizational Need**

This project will explore various data from the TMDM movie dataset and do various data analysis. We will use statistics and python to see the co-relation between various data such as the movie budget and the movie revenue.

**A2. Context and Background**

This dataset contains information about the 10000 movies from the movie database. It contains various data such as the budget, revenue, genre, rating popularity etc. We will be doing data cleaning and exploratory data analysis to find out meaningful insights.

These insights will help the producer and the movie production company to decide what movie genre they should make. It will also provide them the revenue estimate and will give them idea on how much budget should the movie have. These insights can also help the general audience who wants to see the list of successful and unsuccessful movies.

**A3. Summary of Published Works**

According to the research done by vocal. Media website, the larger the movie budget, the higher the revenue will be. Also, it found out that higher budget movies have bigger stars attached to the movie.

In another study done in medium.com website, it found out a clear correlation between movie genre and its box office revenue. The action and adventure genre turned out to be more profitable genre than others.

In the blog posted by Lisa charlotte in <https://blog.datawrapper.de>, she researched whether the movie with high budget is more successful. Through her analysis, she found out that high budget movies have always earned more than their production budget. So, high budget movies are most likely to be successful.

**A4. Summary of Data Analytics Solution**

My data analysis process will include data cleaning, exploratory data analysis, statistics and plotting the findings in the graph to forecast future data. I will be using python and statistics to draw a correlation among various data such as movie budget and its revenue.

**A5. Benefit to Organization and Decision-Making Process**

With this data analysis, it will be easier to make a decision for the movie producers. The movie production company will know which genre is most likely to succeed, how much budget they need, how long the movie should be etc.

**B. Data Analytics Plan**

**B1. Goals, Objectives and Deliverables**

The goal of this project is to find out the relation between the various data in the TMDM movie dataset.

The objectives of this goal are:

* larger budget movies are likely to produce the higher revenues.

The deliverable of this objective is to return the revenue of high budget movies.

* The movie’s budget has increased over the years.

The deliverable for this objective is to return the movie’s budget over

different years.

* The movie’s revenue increased over the years?

The deliverable for this objective is to return the movie’s revenue over the years.

**B2. Scope of Project**

The scope of the project is to find out answers to the above objectives through data analysis. The analysis will only include data from the given dataset. It will not use any other published works or data to find out the insights.

B3. Standard Methodology

For this project, I will be using ADDIE methodology. This methodology consists

of five major steps in a specific order. These five steps include Analysis, Design,

Development, Implementation and Evaluation. These steps are explained below:

* Analysis: The first step in this method is Analysis. In this step, the goal of the project is clearly defined. I will define the objective of this analysis- which is to find out the correlation between movie budget and revenue.
* Design: In this step, I will identify and develop strategies to meet my goal. I will design a present a line graph to show the trend of movie budget and revenue over the years.
* Development: After designing, I will be collecting all the pieces created during the design. I will develop a storyboard to tell the story about the movie budget and its revenue over the years.
* Implementation: I will summarize the whole project in this section. It is now ready to be implemented and the movie producers and the audience can access to the data easily.
* Evaluation: This is the final section where I will ask audience about their opinions and feedback about this project.

**B4. Timeline and Milestones**

|  |  |  |  |
| --- | --- | --- | --- |
| Milestone | Projected start date | Projected End date | Duration(days/hrs) |
| Establish requirements  for analytics process | 04/10/2022 | 04/10/2022 | 1 |
| Data Cleaning and Wrangling | 04/11/2022 | 04/11/2022 | 1 |
| Exploratory Data Analysis | 04/12/2022 | 04/12/2022 | 1 |
| Data Visualization | 04/13/2022 | 04/13/2022 | 1 |
| Data Report in a pdf format | 04/14/2022 | 04/16/2022 | 2 |

**B5. Resources and Costs**

Following is the breakdown of costs associated with this project:

|  |  |
| --- | --- |
| Technology or infrastructure | Cost |
| Computer | $1000 |
| Dataset | Free (Kaggle.com) |
| Python Development Environment | Free |
| Work Hours (60 hours) | N/A |

**B6. Criteria for success**

My criteria for success will be meeting my objectives. The project will be deemed successful if I am able to answer all my questions using the proper methods.

Measurable criteria: The project will be considered successful if I am able to show the data of correlation between movie budget and its revenue. To be able show the trend of movie budget and its revenue over the years is my criteria for success of this project.

**C. Design of Data Analytics Solution**

**C1. Hypothesis**

The higher budget movies are likely to be more successful because they get to cast the popular stars and also get to deploy the best technologies in the movie.

**C2. Analytical Method**

The method I will use for this data analysis is a Time series analysis. It is a method

used to identify trends and cycles over time. I will be using this method to find out

how the movie budget and revenue has fared over the various years.

**C2a. Justification of Analytical Method**

This method will measure movie budget and its revenue at different point in time.

Choosing this method will allow me to forecast how much revenue it can expect to

produce based on its budget.

**C3. Tools and Environments of Solution**

I will use python to do the data wrangling and exploratory data analysis. Python will allow me to do the data analysis lot quicker than other programs by using its built-in packages.

**C4. Methods and Metrics to Evaluate Statistical Significance**

I will be using t-test: paired two sample for means as a method to evaluate statistical significance. The metric I will be using will be the p-value and the t stat value. If the p value is less than the alpha, I can reject the null hypothesis and say that the revenue is higher than the budget.

**C4a. Justification of Methods and Metrics**

T-test method will allow to compare the mean of movie budget and movie revenue. It will be used to determine whether the amount of movie budget will have any effect on its revenue.

The metric of p-value and t stat will determine whether to accept or reject the null hypothesis.

**C5. Practical Significance**

If there is a positive co-relation between the movie budget and its revenue, it will be easier for the production company to decide how much budget they can invest in the movie based on the revenue forecast.

**C6. Visual Communication**

I will be using line graph in python to communicate my data findings. The line

graph will allow me to illustrate the trend of movie budget and revenue over

various years. It will help me achieve my objective my comparing the movie

budget to its revenue. The changes and trends will be easily visible and

understandable through the help of line graph.

**D. Description of Datasets**

**D1. Source of Data**

The source of my data is Kaggle.com. Here is the link for the data source:

<https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata>

**D2. Appropriateness of Dataset**

The dataset is appropriate because it provides all the required variables for a proper data analysis. It is also in csv format which makes it easier to import data into python and do the needed analysis.

**D3. Data Collection Methods**

To collect this data, I downloaded the csv file from the Kaggle.com website. Then, I imported that csv file in to the jupyter notebook. The csv file was then cleaned using proper data wrangling method.

**D4. Data Quality**

The data quality was pretty good as it was in csv file which made it easier to

import. Only thing needed was some data wrangling. Some of the data were

irrelevant to my project. So, I had to correct those data.

I had to drop some of the columns like

(‘imdb\_id’, ‘overview', ‘homepage’, ‘tagline’, ‘budget\_adj’,’revenue\_adj’, ‘keywords’)

Budget and Revenue had lot of number of rows with 0 as their value. So, dropping them is not a good choice so, I decided to keep them and replace 0 with Null values. All those changes can be found in the attached codes.py file.

**D5. Data Governance, Privacy and Security, Ethical, Legal, and Regulatory Compliance**

Since I used the dataset from Kaggle.com, I didn’t have any problem with data privacy or legality.

**D5a. Precautions**

I didn’t have to use any precautions as Kaggle.com is trusted by some of the largest data science companies in the world.

**Sources**

Lisa Charlotte Muth (May 6th, 2021) Does a higher budget

make a movie more successful? Retrieved from

https://blog.datawrapper.de/higher-budget-more-successful-

movies/

Florence Ordesh (April 12th ,2021) Predicting Movie Box Office

Revenue According to Budget and Runtime. Retrieved from

https://vocal.media/geeks/predicting-movie-box-office-revenue-

according-to-budget-and-runtime

Ben Niu (April 14th, 2019) Movie Genre and Profits: Which

factors lead to a more profitable movie? Retrieved from

https://medium.com/@ordosnb/movie-genre-and-profits-

913a3142a0fb