Investigate The Movies DataBase (TMDB) Report

By Aya Ghorabah

February 2021

This report is an illustration for the data investigation of TMDB.

In this project, I will investigate the (The Movie Database (TMDb)). This dataset contains different pieces of information on films like the main actor, director, budget, ...etc.

After a quick view of the dataset on Microsoft Excel, I made some questions that their answers would interest the viewer.

Here are these questions:

1- which film has the largest budget and which one has the lowest?

2- which film has the largest revenue and which one has the lowest?

3- which film has the highest rating?

4- which year has the largest number of produced films and which one has the lowest?

5- Which year has the highest total budget for all the produced films and which one has the lowest?

6- Which year has the highest total revenue for all the produced films and which one has the lowest?

7- For the year 2015, What is the relation between the budget and revenue? Is more budget means more revenue?

8- For the year 2015, what is the relation between the rating and popularity? Is a higher rate means higher popularity?

The Steps I followed in my investigation were:

A) Data Gathering:

This data was given by the Udacity instructor in CSV format.

B) Data Assessing:

This Assessing the data for quality and tidiness issues. It contains two main steps:

1- Visual assessment using Microsoft Excel.
2- Programmatic assessment using Jupyter notebook and different python libraries.

The results are:

a) Quality Assessment

1- Some columns have inappropriate column names like (budget_adj, revenue_adj).

2- One row is duplicated.
3- Release_date column has a type of object.

b) Tidiness Assessment

1. For fair comparison I will use the columns (budget_adj, revenue_adj) instead of (budget, revenue) as the two columns ending with "_adj" show the budget and revenue of the associated movie in terms of 2010 dollars, accounting for inflation over time. drop the columns (budget, revenue)
2. In my investigation, some columns are not needed like (budget, revenue, tagline, keywords, overview, cast, homepage, imdb_id)

C) Data Cleaning:

This is a three stages process:

a) Define: specify the problem.
b) Code: fix the problem.
c) Test: is the problem fixed?

First, create a copy for the dataframes to apply the changes to.

Quality / Tidiness	Problem	Solution
Quality	Some column names not appropriate	Replace with appropriate ones
	Duplicated row	Drop the duplicate
	The "Release_date" type is an object.	Converted to date-time format.
Tidiness	Some columns not needed in my investigation	Drop these columns.

Conclusion

After investigating the data. I built some insights as follows.

1. Film (The Warrior's Way) released in 2010-12-02 has the highest budget in our dataframe with 425,000,000 Dollars.

2. Film (Love, Wedding, Marriage) released in 2011-06-03 has the lowest budget in our dataframe with 0.97 Dollars.
3. Film (Avatar) released in 2009-12-10 has the highest revenue in our dataframe with 2,827,123,750.41 Dollars.
4. Film (Shattered Glass) released in 2003-11-14 has the lowest revenue in our dataframe with 02.37 Dollars.
5. Film (The Shawshank Redemption) released in 1994-09-10 has the highest rate in our dataframe with an average vote rate of 8.4.
6. Film (Foodfight!) released in 2012-06-15 has the lowest rate in our dataframe with an average vote rate of 2.2.
7. The year 2011 has the largest number of released films with 199 films.
8. The year 1969 has the lowest number of released films with 4 films.
9. The year 2010 has the highest total film budget in our dataframe with 8,463,138,439.00 Dollars.
10. The year 192 has the lowest total film budget in our dataframe with 123,398,694.38 Dollars.
11. The year 2015 has the highest total film budget in our dataframe with 24,106,678,369.98 Dollars.
12. The year 1966 has the lowest total film budget in our dataframe with 569,262,321.68 Dollars.
13. more budget doesn't mean more revenue.
14. popular films are not necessary to have higher rates.

Limitations

During my investigation, I noticed that this database has some missing data which will help me get more specific results like:

1- Missing the budget and/or revenue for some films result in dropping their rows. Which limits the results found.
2- Having mixed genre film is normal but it would be more reliable if there is a column for the main genre it will help get the best result for the most popular genre.
3- Also, I think if the database had a column for the main actor or actress would help me figure out the most popular one.
4- Dropping NAN rows made a lot of key data lost.

My References:

1- https://carlyhochreiter.files.wordpress.com/2018/05/investigating-movie-dataset.pdf
2- https://github.com/leogovan/investigate-a-dataset/blob/master/investigate-a-dataset-tidyup-version.ipynb
3- https://github.com/aghorabah/Udacity-Nano-degree-Project-Wrangle-and-Analyze-Data
4- https://www.shanelynn.ie/bar-plots-in-python-using-pandas-dataframes/

5- https://stackoverflow.com/questions/46794373/make-a-bar-graph-of-2-variables-based-on-a-dataframe
6- https://stackoverflow.com/questions/3899980/how-to-change-the-font-size-on-a-matplotlib-plot
7- https://stackoverflow.com/questions/11640243/pandas-plot-multiple-y-axes