Movies: Budgets, Popularity and Revenue

Final Project

*Team Slightly Late for Dinner*

Devan-Sunshine Grey

Michael Farrell

IST 652 Scripting for Data Analysis

Professor Deborah Landowski

Syracuse University

June 2021

Table of Contents

[Data and its source, including any preprocessing 1](#_Toc74227011)

[About the Data 1](#_Toc74227012)

[Moviegoer Demographics 1](#_Toc74227013)

[Description of your data exploration and data cleaning steps 2](#_Toc74227014)

[Clearly stated question(s) that describes whether it is a summary or comparison question and what fields are being used in the data 5](#_Toc74227015)

[Brief description of the program 5](#_Toc74227016)

[Description of the output files, which should include an analysis of your results along with a final conclusion about your overall results. 6](#_Toc74227017)

[Conclusions 14](#_Toc74227018)

[Appendix A – Sample Data 1](#_Toc74227019)

[Appendix B: Group Member Tasks & Roles 1](#_Toc74227020)

# Data and its source, including any preprocessing

## About the Data

The datasets used were downloaded from the IMDb website. The separate datasets contained data about movies to include release date, title, budget, genres, production companies, revenue, profit, popularity, vote average, and vote count. An example of the data is in Appendix A.

The dataset includes 503 rows and 10 columns. There is a headings row, and the first column has row id numbers. The data is a mix of numeric and string.

The data dictionary

|  |  |
| --- | --- |
|  | Definition |
| release\_date | Date of release |
| title | Official movie title |
| budget | Movie budget |
| genres | Movie genre: e.g., Action, Thriller, Music, Adventure, Horror, Drama, Comedy, Fantasy, History, Science Fiction, Crime, Mystery, Family, Romance, Animation |
| production\_companies | Name of production company |
| revenue | Box office gross |
| profit | Movie profit |
| popularity | How many people went to see it |
| vote\_average | Average up-vote count from 0 to 10, 10 the highest |
| vote\_count | Count of viewers rating the movie |

## Moviegoer Demographics

While not used in the data analysis, the following is good information to know when evaluating movie popularity ratings. [[1]](#footnote-1)

Graphical user interface, application, website

Description automatically generated

# Description of your data exploration and data cleaning steps

Pandas was able to recognize that the first row of data was the column names, this was helpful because when we downloaded the csv file after getting the information, our csv did not include proper column names, instead the column names were in the first row of our data. The package pd.read\_csv correctly identified that the first row of our file contained the column names.

Pandas recognized that the first row was the column names and moved the first row to column names, however, it left the first row blank and that row was removed.

Rows were removed that contained missing values using the pandas’s dropna() function. The new data frame replaced the original one.

Movie budget information is important for the analysis, so any movies that have a 0 in the budget column will be removed. To do this we are going to get a list of index names for movies that have a budget equal to 0, again revising the data frame by removing these movies.

Similarly, it is necessary to remove movies that have a 0 revenue, as it appears that this data just has not be collected, and therefore is not accurate. These entries were removed using the same method that was used to drop the movies that had a 0 in the budget column.

The data type used for the release data needed to be changed to date type. This was done using the pd\_to\_datetime function, allowing extraction of each element of the date. The month, day, and year were extracted, and new columns created named month, day and year.

It is interesting to note

* the lowest budget for a movie is only $258,157
* the maximum budget is $300,000,000.
* the lowest revenue is $4,537.
* the maximum revenue is $2,046,240,000. This is a massive revenue.
* the minimum profit is - $60,477,350 and
* the maximum profit is $1,746,240,000. This is a massive profit.
* there is a large range in popularity scores.
  + The maximum popularity score is 71.54 and
  + the minimum is 3.54.
  + The mean popularity is 18.4 and
  + 75% of the movies have a popularity score less than 21.2.
* The vote\_average column has a range of 4 - 8.4, with an average of 6.5.
* The vote\_count has a maximum of 14,913.
* The max value in this column, also appears to be an outlier, as 75% of the movies have less than 2151 votes and the average vote count is 1701

This makes one wonder

* is the maximum popularity score an error? Or
* does it correspond to the movie with the highest profit?
* does the maximum vote count reference the same movie that had the largest popularity?

This project will briefly discuss those and other questions.

We have decided to aggregate a percent profit column as well, attempting to normalize the data. To do this we are diving the profit column by the budget column and multiply the result by 100 and saving it in a new column named percent\_profit.

In order to answer some of the questions, some of the data needs to be discretized.

The first column we are discretizing is the budget column. The budget column is being discretized into four groups: extremely\_low, low, high and extremely\_high. To do this, we first need to create a list of the categories: "extremely\_low", "low", "high", "extremely\_high". The same was done for the revenue, profit, popularity, vote average, vote count, and percent profit using the same categories.

What follows are the categories.

**Budget**. Budget is a continuous variable. In order to use budget in our analysis, it needs to be converted to categorical data. As stated above, we chose the values based on the quartiles. The categories created are:

* extremely low budgets - budgets less than $13,000,000,
* low budgets - budgets between $13,000,000 and $30,000,000,
* high budgets - budgets between $30,000,000 and $62,192,550, and
* extremely\_high budgets - budgets between $62,192,550 and $300,000,000.

**Revenue**.

* extremely\_low revenue - revenues less than $21,458,200,
* low revenues - revenues between $21,458,200 and $62,954,020,
* high revenues - revenues between $62,954,020 and $187,976,900, and
* extremely\_high revenues - revenues between $187,976,900 and $2,046,240,000.

**Profit**.

* Negative profit are profits less than $0,
* low profits are profits between $0 and $29,314,900,
* high profits are profits between $29,314,900 and $140,784,100, and
* extremely\_high profits between $140,784,100 and $1,746,240,001.

**Popularity**.

* extremely\_low popularity are popularities less than 12.442,
* low popularities are popularities between 12.442 and 15.7405,
* high popularity are popularities between 15.7405 and 21.23025 and
* extremely\_high popularity between 21.23025 and 71.538.

**Vote average**.

* extremely\_low vote\_average is vote averages less than 6,
* low are between 6 to 6.5,
* high between 6.5 and 7 and
* extremely\_high 7 and 8.5.

**Vote count**.

* extremely\_low vote counts are vote counts less than 440,
* low vote counts are between 440 and 1151,
* high vote counts are between 1151 and 2522 and
* extremely\_high vote counts are between 2522 and 14913.

**Percent profit**.

* extremely\_low are percent profits between -100 and 0,
* low between 6.5 and 108,
* high between 108 and 436 and
* extremely\_high between 436 and 6527.

**Date**. Date was also discretized. We are setting new categories for the day column by creating a new column for week. The categories are

* week\_1 is the first 7 days of the month,
* week\_2 is days 8 - 14,
* week\_3 is days 15 ‑21, and
* week\_4 is the rest of the days of a month.

# Clearly stated question(s) that describes whether it is a summary or comparison question and what fields are being used in the data

1. How are the amounts of percent\_profits distributed across budget levels? We want to compare the budget category percentage make up for each percent\_profit level. To do this we need to get the count for each budget level, the count for each percent\_profit level by budget level and then divide the count of the percent\_profit/count of budget level and multiply by 100. We must do this for each budget level and level of percent\_profits.
2. Do big name production companies impact the percent profit? We want to compare the production company percentage make up for each percent\_profit level.
3. Does time of the month the movie is released affect percent profit? We want to compare the percent\_profit level percentage make up for each time of month.
4. Do "Good" Movies Make Money? -- We're defining "Good" as vote average. This is a comparison question using profit and vote average.
5. Does Popularity = Profit? This is a comparison question using profit and popularity
6. How does budget impact vote average? This is a comparison question using budget and vote average
7. How does budget impact popularity? This is a comparison question using budget and popularity
8. Is there a relationship between "Above Average Movies" and Budget/Price? This is a summary question using a created field of above average movies and the ratio of budget/price
9. aka *The* Big *Question*: What role do production companies play in the entertainment industry?
   1. Is there a relationship between production studio and average vote? This is a comparison question using production studio and average vote
   2. Production studio and budget? This is a comparison question using production studio and budget
   3. Production studio and percent profit? This is a comparison question using production studio and calculated percent profit

# Brief description of the program

Python was used as the programming language. The modules used

bs4: BeautifulSoup

pandas

requests

time

re

json

csv

urllib.parse: quote

numpy

matplotlib.pyplot

matplotlib.patches

# Description of the output files, which should include an analysis of your results along with a final conclusion about your overall results.

1. How are the amounts of percent\_profits distributed across budget levels?

Chart, bar chart

Description automatically generated

This graph proved very interesting. Movies with an extremely low budget have the highest percentage make-up of making an extremely high percent profit. Movies with an extremely high budget are the most likely to be profitable overall, being that they are the least likely to have a negative profit, with only 5.9% of the movies classified as having an extremely high budget in our dataset made a negative profit. Movies with a low or high budget only make an extremely high percent profit less than 17.1% and 15.6% of the time respectively. They also have the highest chance of making a low or negative profit out of all the budget categories. Based, on this analysis, percent profits are not uniformly distributed across budget levels. Movies with an extremely high budget are the least likely to have a negative percent profit. Movies with an extremely low budget are the most likely to have an extremely high percent profit. Our recommendation to studios, would be to either have a extremely low or extremely high budget and to veer away from productions with an extremely low or high budget. Further analysis for tighter recommendations is needed.

1. Do big name production companies impact the percent profit?

A screenshot of a computer

Description automatically generated with medium confidence

This graph provides some insights, however, most of our movies have more than one main production company and only one production company is being shown. For example, DreamWorks and Universal had a movie named First Man and it was profitable. However, based on the way that we assigned a main production company, only Universal was given credit for that movie.

1. Does time of the month the movie is released affect percent profit?

Chart, bar chart

Description automatically generated

This is interesting in that it shows that movies released within the first two weeks of the month tend to be more profitable. We would like to look at a breakdown of month to percent profit for further analysis

1. Chart, scatter chart

   Description automatically generatedDo "Good" Movies Make Money? -- We're defining "Good" as vote average above 6.5

From the data it appears that profit is not based upon whether a movie is good or bad.

1. Chart, scatter chart

   Description automatically generated Does Popularity = Profit?

The data indicates that popularity does positively influence profit to a certain extent, although there were many popular movies that did not earn a large profit. The sweet spot appears to be extremely high popularity.

1. How does budget impact vote average?

Chart, scatter chart

Description automatically generatedLooking at the above chart, big budget movies do get an average vote over 6.5, more lower budget movies got higher vote averages.

1. Chart, scatter chart

   Description automatically generated How does budget impact popularity?

Popularity reaches a peak at about 35. The majority of lower budget movies stayed at 20 or below.

1. Chart, scatter chart

   Description automatically generated Is there a relationship between "Above Average Movies" and Budget/Price?

High budget movies with above average votes earn the most profit.

1. Is there a relationship between "Budget, Percent Profit and Average Vote?

Chart, scatter chart

Description automatically generated

The data indicates that t is a mixed bag. It can be misleading as big budget movies profit percentage can be lower but still earn multi-millions of dollars.

1. aka *The* Big *Question*: What role do production companies play in the entertainment industry?
   * + 1. Is there a relationship between production studio revenue and budget?

Chart, scatter chart

Description automatically generated

It is no wonder that Disney has some of the biggest budget movies. That said, Fox has some of the better earning films. That said, this graph provides some insights; however, most movies have more than one main production company, and only one production company is being shown. For example, DreamWorks and Universal had a movie named First Man, and it was profitable. However, based on the way that we assigned the main production company, only Universal was given credit for that movie

* + - 1. Production studio and percent profit?

Chart, scatter chart

Description automatically generated

Universal and Paramount garnered the best percent profit from a few of their films. Most films earned between no profit and 400% profit. It should be noted that the film industry has a unique way of reporting profit, which tends to understate profits, at times by a large margin.

## Conclusions

Based on this analysis

* + Percent profits are not uniformly distributed across budget levels.
  + Movies with an extremely high budget are the least likely to have a negative percent profit.
  + Movies with an extremely low budget are the most likely to have an extremely high percent profit.

Higher popular movies earned a larger percentage for the studios. The most popular movies are Disney and Warner movies. Disney has the Marvel and Star Wars franchises – all money makers. Warner Brothers has the DC comic franchises. In both cases, action adventure and comic books coming to life make money at the box office.

Our recommendation to studios would be to follow Warner Brothers and Disney - produce movies with an extremely high budget and veer away from productions with an extremely low budget. Further analysis for tighter recommendations is needed.

project provided an interesting view into

| **release\_date** | | **title** | | **budget** | | **genres** | **production\_companies** | | **revenue** | | **profit** | | **popularity** | | **vote\_average** | | | **vote\_count** | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| NaN | NaN | | NaN | | NaN | | | NaN | | NaN | | NaN | | NaN | | NaN | NaN | |
| 2018-12-07 | Aquaman | | 160000000.0 | | [{'id': 28, 'name': 'Action'}, {'id': 12, 'nam... | | | [{'id': 429, 'logo\_path': '/2Tc1P3Ac8M479naPp1... | | 1.143689e+09 | | 9.836892e+08 | | 28.789 | | 6.8 | 6781.0 | |
| 2019-08-29 | Night Hunter | | 0.0 | | [{'id': 53, 'name': 'Thriller'}, {'id': 80, 'n... | | | [{'id': 109529, 'logo\_path': None, 'name': 'Ar... | | 0.000000e+00 | | 0.000000e+00 | | 48.191 | | 0.0 | 13.0 | |
| 2018-12-06 | Spider-Man: Into the Spider-Verse | | 90000000.0 | | [{'id': 28, 'name': 'Action'}, {'id': 12, 'nam... | | | [{'id': 5, 'logo\_path': '/71BqEFAF4V3qjjMPCpLu... | | 3.754504e+08 | | 2.854504e+08 | | 39.780 | | 8.4 | 4808.0 | |
| 2018-10-24 | Bohemian Rhapsody | | 52000000.0 | | [{'id': 18, 'name': 'Drama'}, {'id': 10402, 'n... | | | [{'id': 3281, 'logo\_path': '/8tMybAieh64uzvm8k... | | 8.940275e+08 | | 8.420275e+08 | | 35.534 | | 8.1 | 8357.0 | |
| 1985-07-13 | Live Aid | | 0.0 | | [{'id': 10402, 'name': 'Music'}, {'id': 99, 'n... | | | [{'id': 3065, 'logo\_path': '/sJ919bmS8B53hujwb... | | 0.000000e+00 | | 0.000000e+00 | | 2.065 | | 7.8 | 23.0 | |
| 2018-04-25 | Avengers: Infinity War | | 300000000.0 | | [{'id': 12, 'name': 'Adventure'}, {'id': 28, '... | | | [{'id': 420, 'logo\_path': '/hUzeosd33nzE5MCNsZ... | | 2.046240e+09 | | 1.746240e+09 | | 71.537 | | 8.3 | 14913.0 | |

# **Appendix A – Sample Data**

# Appendix B: Group Member Tasks & Roles

**Both team members were involved in all stages of the project. Devan did the initial programming with Michael reviewing it for clarity. Michael drafted the reports with Devan reviewing for clarity.**

**Both members created questions and reviewed how the data answered those questions.**

**Devan drafted the presentation, Michael reviewed it and both presented.**

**Bottom line: It was a balanced team effort.**

1. Graphic from Statistica, https://www.statista.com/statistics/251466/us-movie-theater-audience-by-age/ [↑](#footnote-ref-1)