rsajac / dsc-phase-1-project-v2-3 Public

forked from learn-co-curriculum/dsc-phase-1-project-v2-3

View license

☆ 0 stars   48 forks

☆ Star   Watch ▾

<> Code   Pull requests   Actions   Projects   Wiki   Security   Insights

master ▾   ...

This branch is 6 commits ahead of learn-co-curriculum:master.   Contribute ▾   Fetch upstream ▾

rsajac Finished Project ...   1 minute ago   14

View code

README.md

Title

Authors: Ryan Sajac

Overview

Microsoft is venturing into the exciting world of movie making. While a first movie is a wonderful opportunity to diversify the company, movie making can be a risky investment. Microsoft needs to think about the cost of production, first time needs, and the safest way to expand into the market. Microsoft wants to target the largest possible audience, or at least generate the most revenue.

Using the genre, runtime, budgets, and gross earnings, we recommend that the best way for Microsoft to safely generate a substantial profit is to direct a movie under the animation genre with a runtime between 95 and 105 minutes as supported by the findings in our comparisons of genre vs. total gross. Our findings are detailed below.

Business Problem

Microsoft must be most interested in two things, how to get a large return on investment, and how to limit risk in accomplishing that task.

For these two pain points, the questions we need to answer in our data analysis are:

1. What type of genres are most likely to be most profitable?
2. What runtimes are most likely to produce higher profit?
3. Which genre limits the risk on return of investment?

From a business perspective, these questions are important to answer because Microsoft needs to establish itself as a key player in the market early on. It can either do that by taking a big risk and hitting it out of the park (question 3), or by playing it safe and producing a movie that is likely to succeed, and continuing to build safely by producing 'safe' movies and diversifying by targeting those that are 'riskier' in terms of return on investment.

Data Understanding

Target Variable - Profit

Microsoft's goal is to make money in this venture.

Data is from Rotten Tomatoes, and IMDB. These sites contain information about gross, budget, runtimes, and genres of movies. We need to correlate this information to determine what are recommendations will be for Microsoft.

I included the following files:

- imdb_title_basics_csv_gz - has genres, runtime minutes, and title
- bom_movie_gross_csv_gz - has the domestic and foreign gross, and title
- tn.movie_budgets.csv.gz - has the budget, and title

I can join these three tables to get relationships among the data to find answers to our questions. Once joined, the dataframe has 1001 movie entries with 17 columns. The most relevant columns for our analysis were:

- Production Budget
- Runtime Minutes
- Worldwide Gross

I created two further columns:

- Profit
- Gross-to-cost-Ratio

While profit is our target variable, Gross-to-cost-ratio helped to indicate the risk of a certain genre.

With more time we could have narrowed down a short list of directors based on titles that have done well for Microsoft to choose from.

Methods

First, we looked at a list of possibly relevant files and loaded them into dataframes and checked to see what information in those dataframes could be analyzed to tackle our problems. We chose three relevant dataframes and combined them with a merge. We had to drop duplicate results, search for missing information and either replace or delete without compromising the data. We either found and inserted relevant data, or deleted rows. In both cases this effected less than .1% of the data. Eventually, we did need to eliminate 4 genres in our data because I did not feel confident recommending a genre of which there were 9 or less movies in our data.

Given the data and problem at hand, we want to target profit. We want to target limiting risk. We needed information about type of movie, profit of movies, and runtimes of movies all based upon their genre.

After combining our dataframe, we could remove extraneous columns for our data, to present a more simplified dataframe to target our business problem. So that we could easily group our information by genre, we exploded out our original dataframe genre so that each movie would have a row for each individual genre tag. We showed bar graphs of the max, min, and median profit and gross-to-cost-ratio sorted by genre with ascending worldwide gross to make a determination of both likely profit and less risk.

We also made several scatter plots to determine likely range of movie to generate revenue.

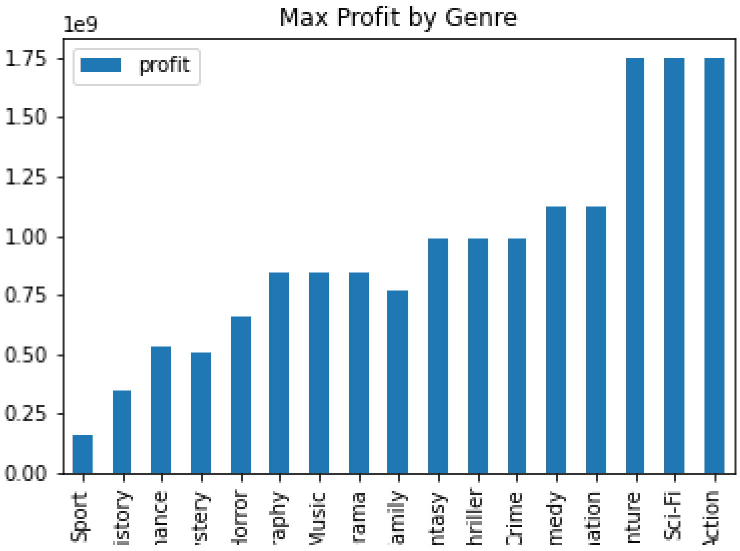
Results

The model fits the data well. I am confident that the results would generalize in most cases, but one cannot predict the success of a movie based solely on genre. I am confident that the genre and runtime that I ultimately conclude that Microsoft should go with would benefit the business if they also filled all the needed roles to make this project happen with people who know the industry well.

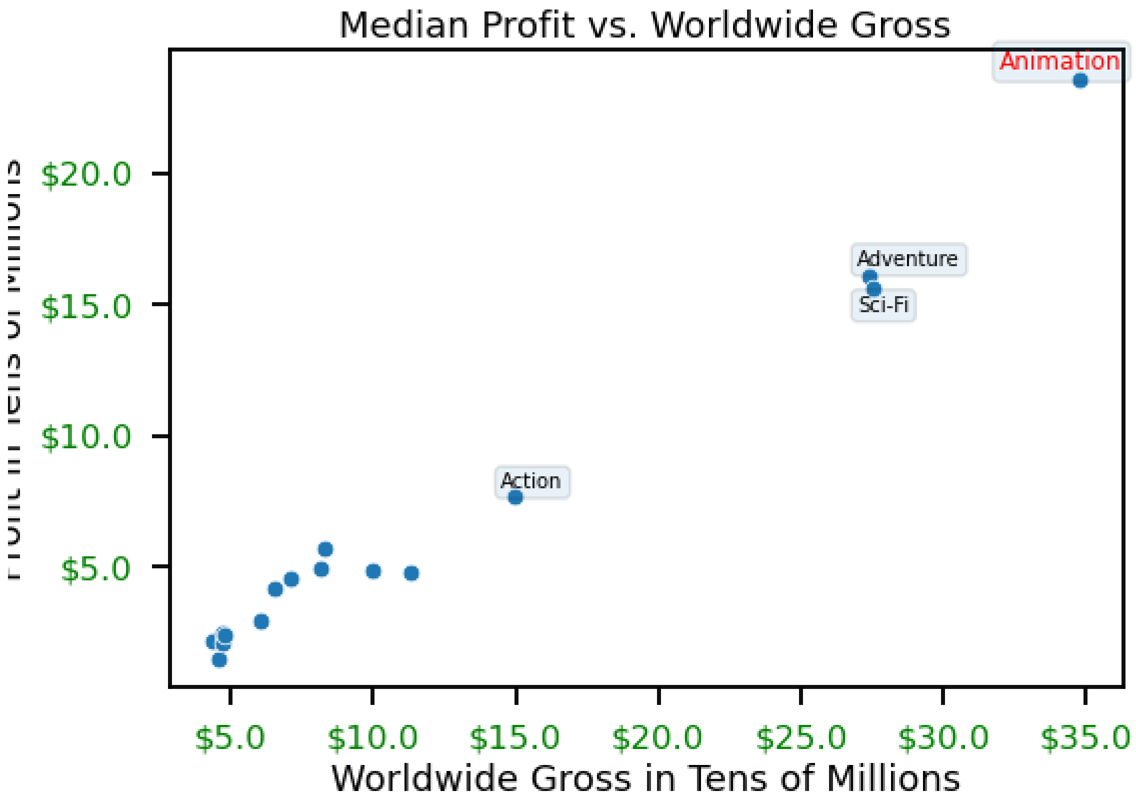
My work solves the business problem of which movie type Microsoft should invest in creating in order to make a significant amount of money while limiting the risk.

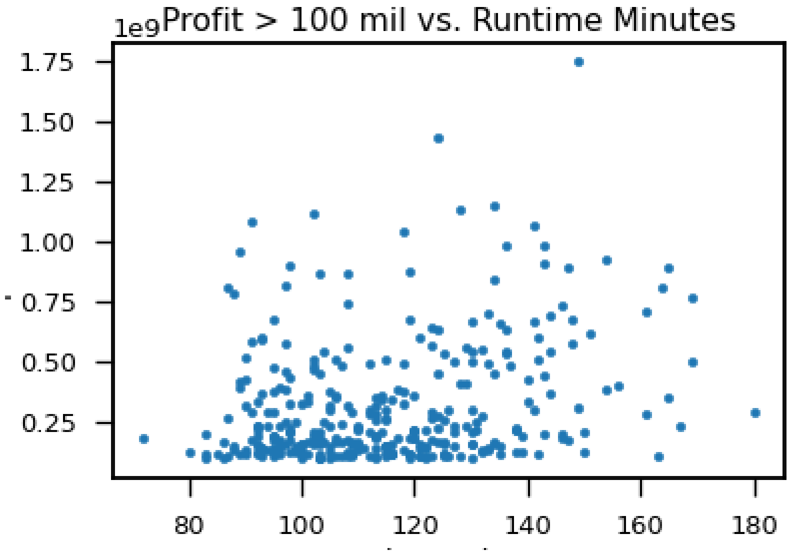
Visual

graph1

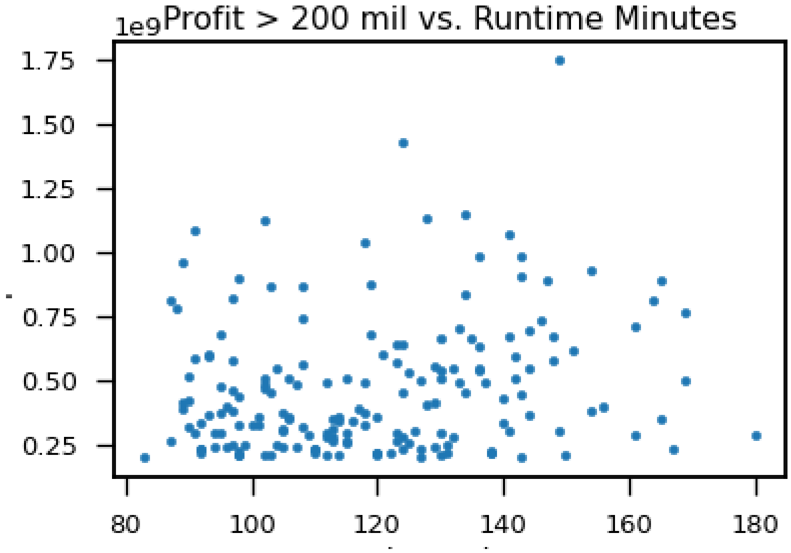


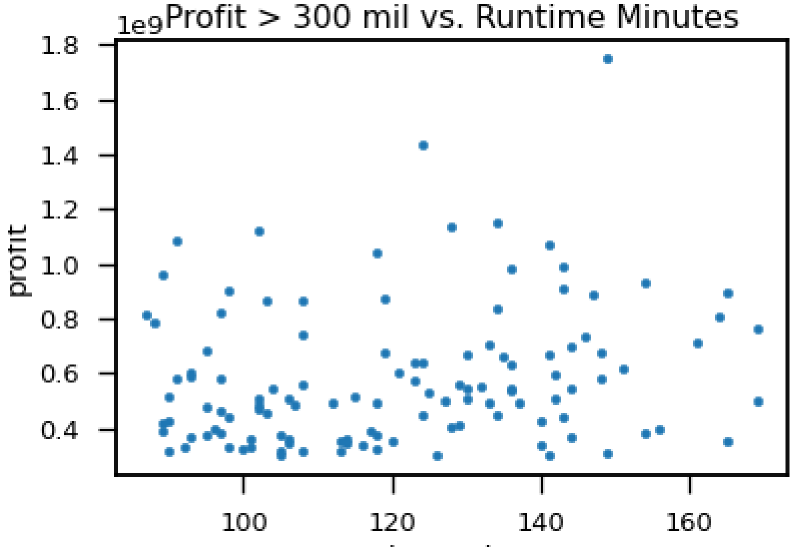
graph3

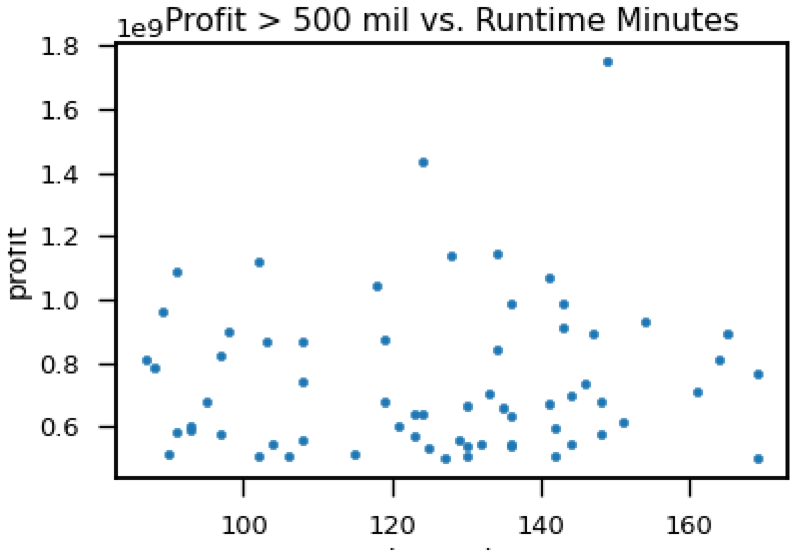


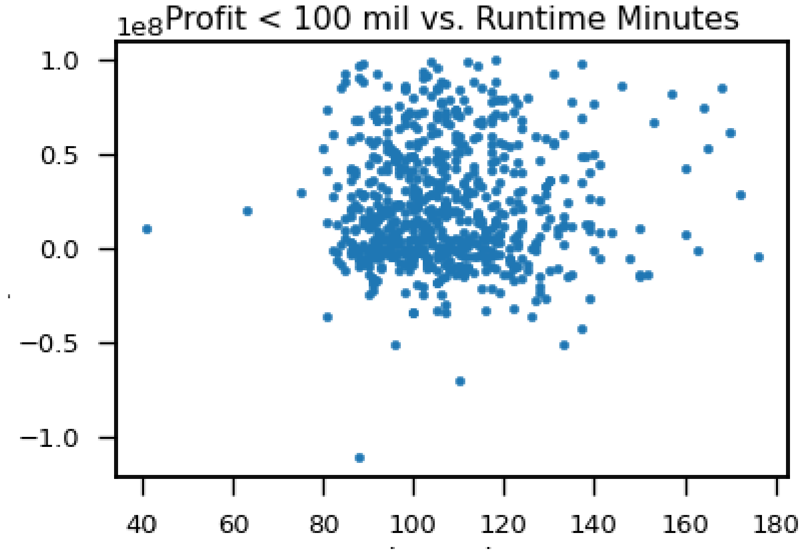

1e9
Profit > 100 mil vs. Runtime Minutes
1.75
1.50
1.25
1.00
0.75
0.50
0.25
80
100
120
140
160
180

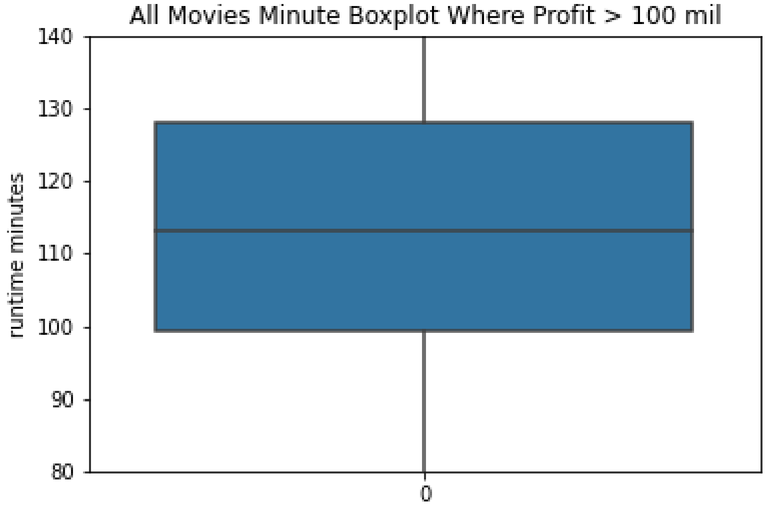
graph5

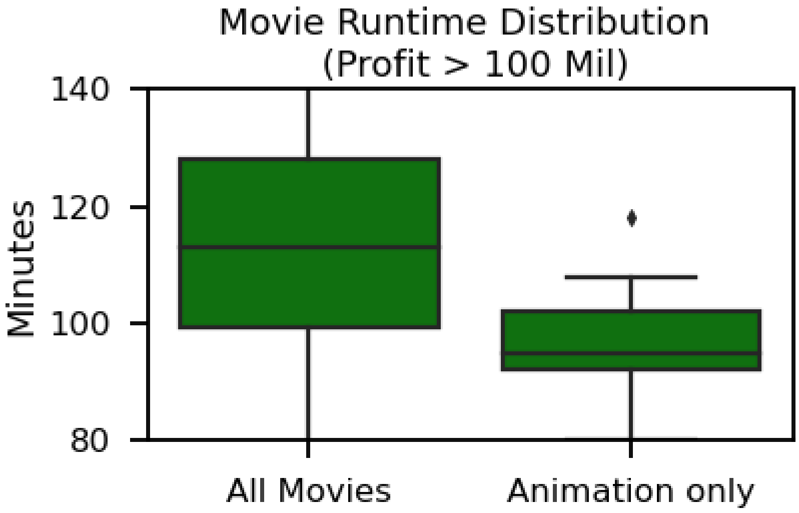

1e9
Profit > 200 mil vs. Runtime Minutes
1.75
1.50
1.25
1.00
0.75
0.50
0.25
80
100
120
140
160
180


1e9
Profit > 300 mil vs. Runtime Minutes
profit
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
100
120
140
160


Profit > 500 mil vs. Runtime Minutes
1e9
profit
1.8
1.6
1.4
1.2
1.0
0.8
0.6
100
120
140
160


Profit < 100 mil vs. Runtime Minutes
1e8
1.0
0.5
0.0
−0.5
−1.0
40
60
80
100
120
140
160
180





Conclusions

Based on the results, my top recommendation for Microsoft as a company is to start with an animation movie. We can tall by both the max genre and the median genre that both profit and gross to cost ratio are high. From the min graph, we can see that its flops will lose only 65 - 70% making animation less risky than most alternatives and a safe venture with likely profit and low downside.

Microsoft also has software and a team that can easily transition into the field of animation. This would likely cut startup cost.

Based on the profit vs runtime minute graphs, movies that have a profit of more than 100 million mostly fall between 90 and 150 minutes with a median of 112. However, the animation median is 95 minutes, so the recommendation is to follow to this median and set a animation movie length between 90 and 105 minutes.

Profit is most correlated with budget. So we recognize Microsoft as a company that could potential have an almost unlimited budget and for success in the movie we recommend an approximate 10-15 million dollar budget for your movie as this is the mean budget of an animation film.

We are limited in our results, and for future movies would be able to provide more specific recommendations.

Limiting factors in our results included throwing out Westerns, War, Documentaries, and Musicals with too few data points. We have a small sample size of movies within an approximate 10 year team period and they may suffer from recency bias. We did not connect success (profit) with directors, authors, or crew, and we could create a short list of those to make a determination of whom to hire. Most importantly however, we did not look at the effect that streaming services have had on the industry and whether Microsoft should target Theater Releases, or Streaming Releases.

Suggestions for our future exploration include:

1. Expand our data to more than the most recent 10 years.
2. Do a similar data analysis but separate movies by streaming and by theater release.
3. Connect author, producer, and director to our data and analyze the importance of those people into our selection.

I have confidence in my initial suggestion, but would love to explore further into this world to continue giving recommendations as Microsoft develops its studio.

For More Information

Please review our full analysis in our Jupyter Notebook or our presentation.

For any additional questions, please contact **Ryan Sajac, rsajac.gmail.com**

Repository Structure

```
├── README.md                           <- The top-level README for
reviewers of this project
├── dsc-phase1-project.ipynb            <- Narrative documentation of
analysis in Jupyter notebook
├── DS_Project_Presentation.pdf         <- PDF version of project
presentation
├── zippedData                          <- Both sourced externally and
generated from code
```

```
└── images                                <- Both sourced externally and
generated from code
```

Releases

No releases published

Create a new release

Packages

No packages published

Publish your first package

Languages

● **Jupyter Notebook** 100.0%