johnlocke333 / **movie_analysis**

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

movie_analysis Public

1 Branch 0 Tags Go to file t Go to file Add file Code

johnlocke333 resetting kernel so code line starts at 1

57b143c · 20 hours ago

images	made adjustments to th...	2 days ago
.gitignore	initial commit	last week
README.md	typos	4 days ago
movie_analysis_p...	made adjustments to th...	2 days ago
notebook.ipynb	resetting kernel so code...	20 hours ago

About

No description, website, or topics provided.

Readme
Activity
0 stars
1 watching
0 forks

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

README

Movie Studio Data Analysis

Author: Jack Locke

- **Jupyter Notebook** 100.0%

Overview

1. Business Understanding
2. Data Understanding
3. Data Preparation
    - a. Merging Datasets
    - b. Filtering Repeated Rows
    - c. Dropping Unnecessary Columns
    - d. Primary Columns Information
    - e. Cleaning Primary Columns
    - f. Creating Additional Columns for Analysis
4. Exploratory Data Analysis
    - a. Genre vs. ROI
    - b. Released Month vs. ROI
    - c. Director vs. ROI
5. Conclusions
    - a. Limitations
    - b. Recommendations
6. Next Steps

Business Understanding

The business stakeholder is a company creating a new movie studio because they want to break into the film industry. My project analyzes films based on box office performance data. The aim is to find patterns and trends within the data in order to provide recommendations for what filmmakers should focus on when creating films for their new movie studio. I will focus on recommending what genres, release dates, and directors will best suit the stakeholder's business when compared to their respective ROIs (return on investment). The recommendations will help the business create films that will lead to profit for the company. My analysis will use the CRISP-DM (Cross-industry standard process for data mining) methodology.

Data Understanding

For this project, I am working with datasets from two different resources. One dataset is from IMDB's relational database. I will gather data about movie genres and their directors from this database. The genre and director information will come from two different tables within the relational database. The final dataset contains information on roughly 160,000 movies.

The other dataset comes from a CSV file called Movie Budgets. This dataset contains information on movie release dates and finances, such as production budget and gross revenue. The dataset contains information on roughly 5,800 movies.

The link attached will guide you to a github account containing the IMDB database and Movie Budgets CSV file with IMDB's respective ERD (Entity Relationship Diagram), if you wish to clone the repository and view the data: https://github.com/learn-co-curriculum/dsc-phase-2-project-v3/tree/main/zippedData

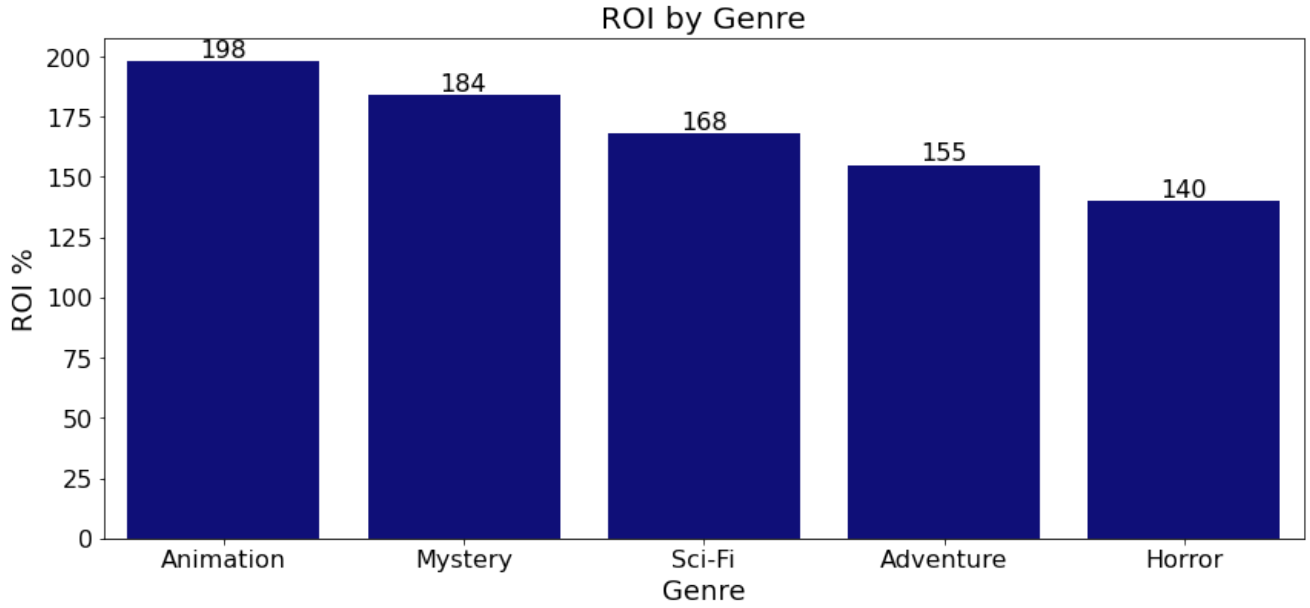
Data Preparation

- Merging Datasets together into one combined dataframe
- After merging the dataset we will deal with repeated rows
- Drop the unnecessary columns from our dataframe
- Get information on our primary columns
- Clean the dataframe
- Create columns needed for our EDA

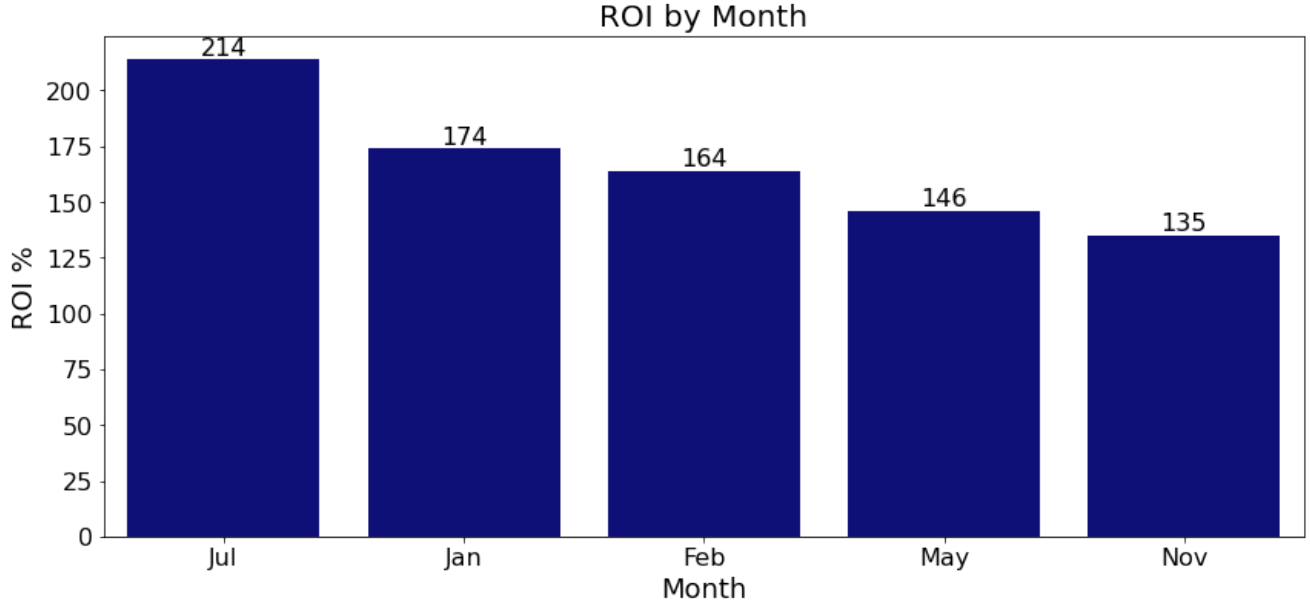
Exploratory Data Analysis

As mentioned in previous sections, I will explore what the stakeholders can control with my created measurement of success. I will examine which genres, months, and directors are the most highly represented. I will need additional data engineering to show my results graphically. I will view the relationships with bar graphs as this will best show us the most highly represented values. I will be looking at the median gross ROI values. Extreme outliers are present in the data, so using the mean will lead to highly skewed results. I will use the median to accurately measure the central tendency to reflect the data.

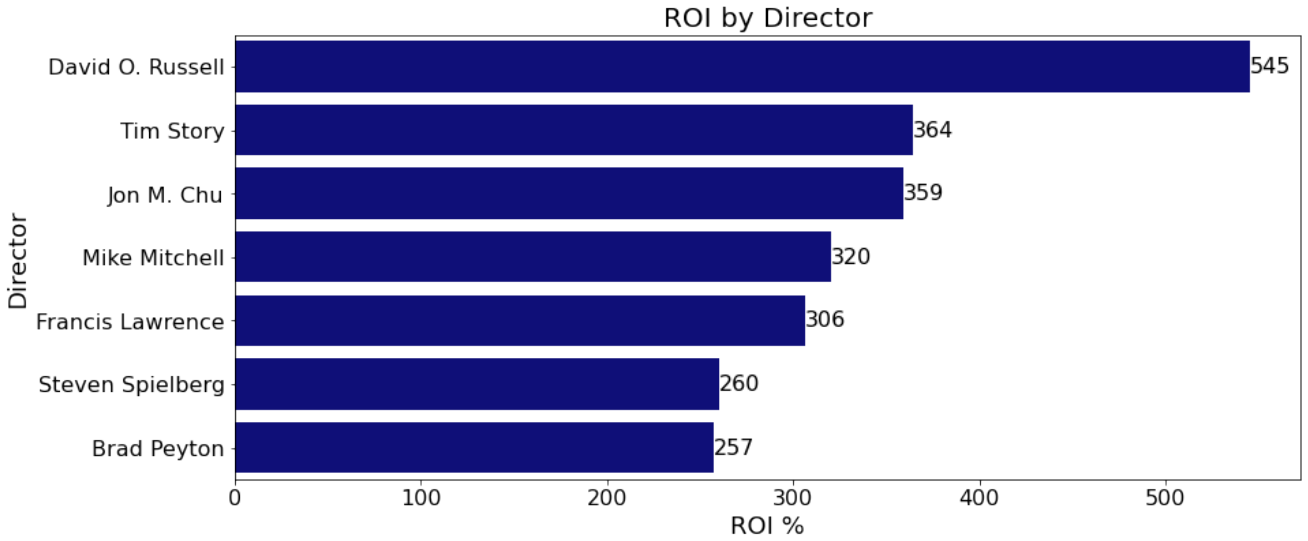
Genre vs. ROI


ROI by Genre
198
184
168
155
140
200
175
150
125
100
75
50
25
0
ROI %
Animation
Mystery
Sci-Fi
Adventure
Horror
Genre

Month vs. ROI


ROI by Month
214
174
164
146
135
200
175
150
125
100
75
50
25
0
ROI %
Jul
Jan
Feb
May
Nov
Month

Director vs. ROI



Conclusions

My Exploratory Data Analysis presents three business findings. The findings will help the business stakeholders figure out specific attributes of movies to help release successful films for their new studio, which, in turn, will lead to a profitable increase for their company. Given these findings and data enrichment in the future, I can build on my recommendations and know where to proceed next.

Limitations

The data possesses many limitations at this stage. I do not have information on every movie release date, genre, and director in the film industry. Additionally, this is gross financial data, so I do not have data on every financial cost that could impact my ROIs. Due to this missing data, I cannot confidently say which movie attributes best suit the company. I can only determine which movie attributes are the most highly represented compared to my measurement of success. Representation will help show us trends and patterns in the data. From there, I can recommend what actions need to be taken in the future to confirm my suspicions on which movie attributes have the potential to lead to the most profit.

Recommendations

I recommend investigating these three recommendations further as we enrich the data and become more confident in our results.

1. **Genre vs. ROI**: I assessed each genre with median ROI to show the five most highly represented genres.
    - Animation
    - Mystery
    - Sci-Fi
    - Adventure
    - Horror
2. **Released Month vs. ROI**: I assessed each release month with median ROI to show the five most highly represented months.
    - July
    - January
    - February

    - May
    - November
3. **Director vs. ROI**: I assessed each director with median ROI to show the seven most highly represented directors.
    - David O. Russell
    - Tim Story
    - Jon M. Chu
    - Mike Mitchell
    - Francis Lawrence
    - Steven Spielberg
    - Brad Peyton

Next Steps

Data enrichment is crucial to help us confirm our business recommendations. As mentioned, without more information on movie genres, release dates, directors, and additional financial costs, I can't say with absolute certainty what movie attributes are best. ROIs could change as more financial costs are presented, leading to different results for our most highly represented genres, release dates, and directors. However, as the data increases, I can start to confirm or deny our suspicions.

In the next steps, I would suggest finding more information on the following:

- More movies → additional information on our highly represented genres, release dates, and directors
- Additional financial costs not accounted for or that occured after the box office release → transition from gross ROI to net ROI
    - Additional post-production costs
    - Additional revenue generated

- Streaming costs

For More Information

See the full analysis in the Jupyter Notebook or review this presentation.

For additional info, contact Jack Locke at jackdlocke@gmail.com

Repository Structure

```
├── images
├── README.md
├── presentation.pdf
└── airplane_safety.ipynb
```

Citations