📘 **adamburstyn** / **Final_Project** ⬭ Public

Flatiron Capstone Project

⭐ **1** star   ⑂ **0** forks

⭐ Star	▾	👁 Unwatch ▾

`<>` **Code**   ⊙ Issues   ⑂ Pull requests   ▷ Actions   ⊞ Projects   📖 Wiki   ⚠ Security   ⬚ Insig

⑂ main ▾   ⋯

Your main branch isn't protected

Protect this branch from force pushing, deletion, or
require status checks before merging.

Protect this branch   Dismiss

🖼 **tflanagan7** Merge pull request #140 from adamburstyn/Trevor ⋯   17 seconds ago   🕙 291

View code

☰ **README.md**   ✎

Box Office Analysis

Flatiron Capstone Project

Python Fever - Danish Ali, Adam Burstyn, Trevor Flanagan, Ian Haas, Hope Miller

Project Overview

To discover insights into movie box office revenue from current industry data that will be
used to make recomendations to Computing Vision in order to maximize the box office
performance as well as popularity of their new movie production venture.

Specifically focus on genres of movie that have:

The highest gross revenue.

The highest net profit.

The highest popularity according to total internet interaction.

Provided data included domestic as well as international reveneus. Domestic revenue was used for analysis after linear regression confirmed they were highly correlated, and any recommendations should be applicable to both foriegn and domestic audiences.

Mean revenue values will be used for comparison so that high outlier values are counted since a high outlier, ie high grossing, film is desirable.

Data Structure

Helper functions stored in helper_functions.py.

/helper_functions.py:

Data that should be unzipped into a folder called /Data in order to run the notebook locally.

/zippedData:

Gross Revenue CSV: bom.movie_gross.csv

IMDB movie database: im.db

Rotten Tomatoes movie details: rt.movie_info.tsv

Rotten Tomatoes movie review data: rt.reviews.tsv

The Movies Database details CSV: tmdb.movies.csv

The Numbers movie budget details: tn.movie_budgets.csv

Notebooks with indendant exploration and further research by team members.

/Individual Notebooks:

Final work flow that calls the get_clean_df() function.

/Python Fever Final Notebook.ipynb:

- Run get_clean_df() from helper function in Python Fever Final Notebook to automate data cleaning and organization.
- Shows additional data aggregation
- Shows statistical analysis and visualizations
- Shows conclusions and business recomendations.

Data Understanding

The team created a singular data frame, using Pandas to join all the tables on the relevent data allowing us to efficiently explore, ask questions and form hypotheses. The core data comes from the IMDB data base giving us movie titles, revenues, ratings, production budgets and allowed us to merge supplemental details from all other sources and prepare a clear overview of box office information and start to hypothesise about what helps movies perform well.

In order to narrow our recommendations we tested the correlation between foriegn and domestic gross revenue and found that they were highly correlated. So all further data was filtered to domestic information since it can be reasonably assumed that performance will be similar.
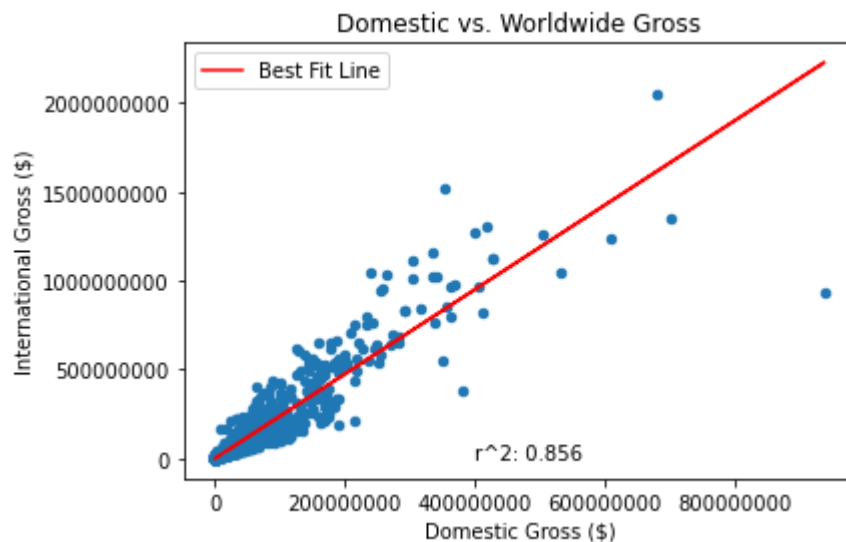
The aggregated, domestic data was then transformed using Panda's functions to seperate the combined genre column that described each movie with multiple tags, e.g. Action, Adventure, Comedy. Now each movie entry could be grouped by genre. Allowing us to test whether one genre tag or combination of genre tags could result in higher grossing, more popular movies.

Goals:

Discover the highest grossing movie genre from the data provided.

Discover the genre with the highest average net revenue from the data provided.

Discover the most popular genre from the data provided.

Gross Box Office by genre

The team first decided to investigate which genre of movie has the highest average gross revenue from the data provided. Used Python string access function explode() to seperate the genre column where many movie entries had many genre tags seperated by commas.
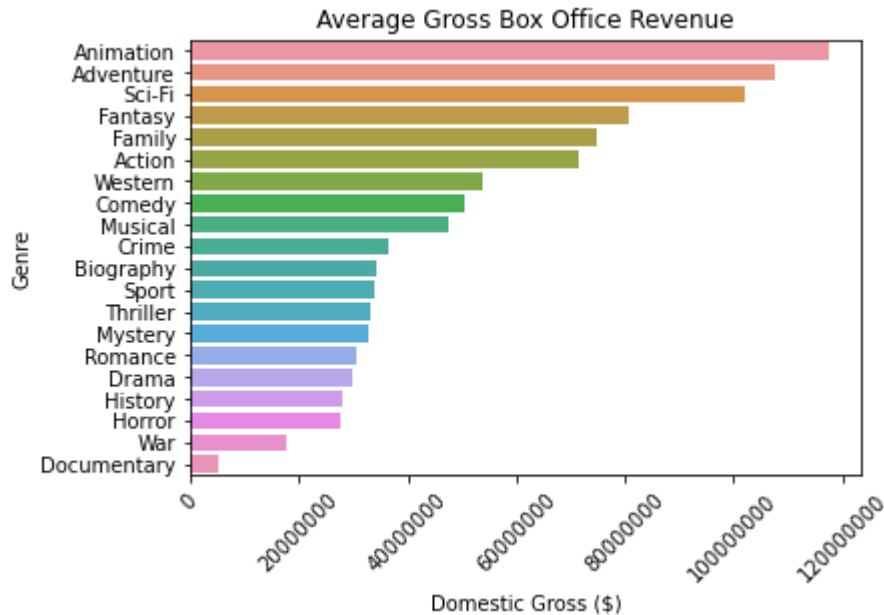
Then we ordered these categories by gross revenue and discovered the Animation genre had the highest box office revenue. t test compared to the next two highest average grossing genres, Adventure and Sci-Fi, showed a statistically significant difference between these and the fourth highest revenue. **The team can confidently make the recomendation that movies in the Animation, Adventure and Sci-Fi genre might make high box office revenues.**

Net revenue by genre

Using the average gross revenue and subtracting the budget column from the IMDB database that describes the amount of money budgeted to make the film, the team aggregated and calculated the average net revenue.

Mean Net Box Office Profit
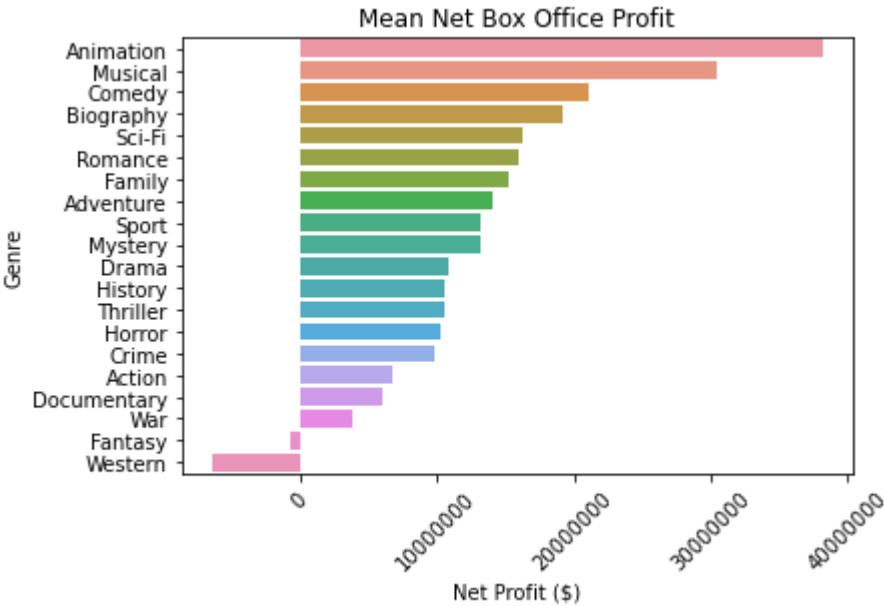
Animation, Musical, and Comedy genre tags have the top three average net revenues. _**We make the recomendation that, for highest profit margins, Computing Vision should produce movies that can be lables as Animation, Musical, and Comedy.**

Popularity by genre

Using the 'popularity' column from the "The Movies" database, the team investigated what genres were the highest in popular opinion. The team found Adventure is the most popular genre, while Documentaries received the highest vote score on average, but concluded the Documentaries vote score was an outlier, having few votes all with high scores and skewing the average. **We recommend that, for a goal of high popularity and engagement, a movie with the Adventure tag would perform well.**

Movie Genres by Popularity (Based on TMDB Metric)
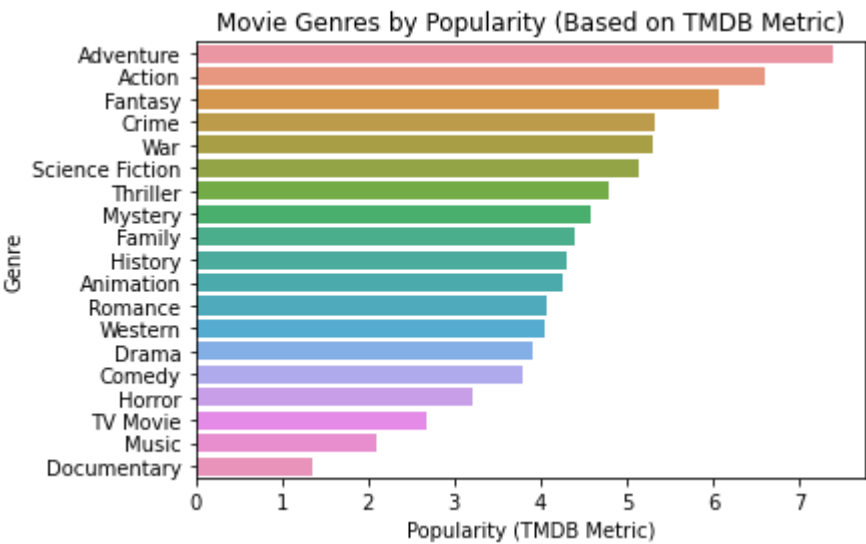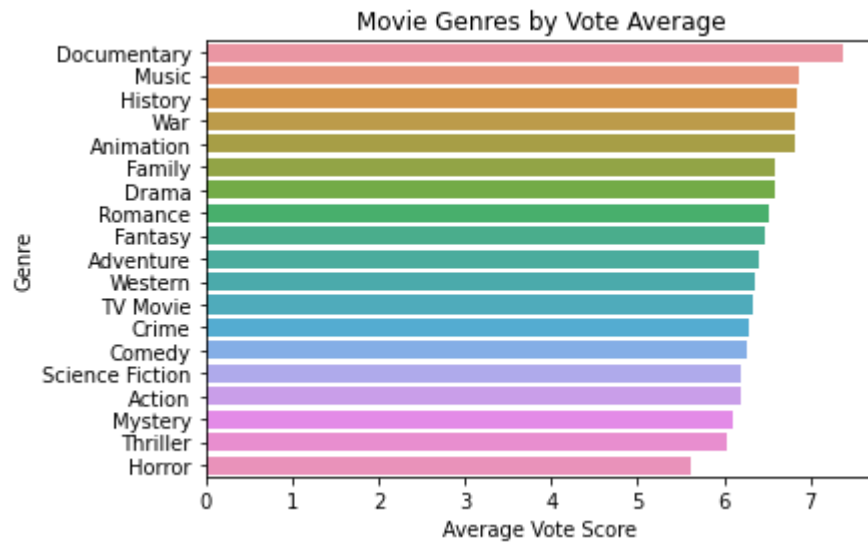
In conclusion our recommendations to Computing Vision are:

Make movies that can be labled as Animation, Adventure, and/or Sci-fi to achieve the highest box office performance.

Make movies that can be labled as Animation to achieve the highest net revenue, ie profit.

Make movies that can be labled as Adventure to achieve maximum popularity and public appeal.

Next Steps

Preliminary data exploration has already begun to discover which writers, directors, and actors are associated with high performing movies, and who Computing Vision may seek to involve in the creative process in the future.

- See the supplemental presentation material from Trevor
- See the optimization model from Adam

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Contributors 5

Languages

- 🔴 **Jupyter Notebook** 99.0%   🔵 **Python** 1.0%