

reboyce


Update README.md

History

3 contributors

94 lines (67 sloc) | 5.28 KB

Gilbert Crew Capstone Project



Produced by: Gilbert Crew

Ryan Burger, Hannah Bowman, Telma Ashton, Reece Boyce, Chris Christensen

Summary

In this project we were charged with exploring what types of films are currently doing the best at the box office using different samples of available data. Through our analysis we translated our findings into actionable insights that the head of Computing Vision's new movie studio can use to help decide what type of films to create.

After exploring the Imdb and Rotten Tomatoes database we were able to come to 3 recommendations based of data analysis on genre, runtime, actors, directors/writers, revenue and competitor studios.

Data Science Steps

Obtaining

We obtained data from sources obtained from the client and moved them into our repository. Below we list the data sets:

- [IMDB](#)
- [Rotten Tomatoes Movie Information](#)
- [Rotten Tomatoes Reviews](#)
- ['The Numbers'](#)
- [TMDB Movies](#)

Cleaning

There was some initial data cleaning that needed to be done. There were multiple null values that needed to be changed to 0 or dropped completely from the datasets so that the data could be analyzed correctly. For most of our analysis we decided to drop null values, as they were determined to be integral in the foundation of a movie. For example, movie runtime had a few null values throughout the IMDB set and we decided to drop them instead of making them an average, or 0. This allowed for proper analysis without compromising the datasets we worked with.

Exploring

Basic exploration was conducted to gain a better understanding of the data we were given. This helped in putting us into the right direction to formulate the recommendations.

Modeling

After cleaning and exploring the data we were able to create tables and visualizations that led us to be able to interpret the data.

Interpeting

Here we came to a final conclusions of our three recommendations gathered in the rest of the data science steps

Sources

Notebooks

All of our data files from individual exploration, version changes and final notebook can be found on the link below:

- [Notebooks link](#)

Presentation

Our presentation shows a summary of our data analysis using visualizations from Matplotlib and Seaborn libraries, our findings from the analysis and statistical tests, and our 3 final recommendations.

- [Presentation link](#)

Navigating Repository

Instructions on navigating repository:

- Click [here](#) to access our repository. The link will go to the main page of the repository
- Go to Notebooks folder and inside the folder you will find:
 - Files folder with all the data explored for our analysis
 - Image folder with the ER diagram for the Imdb data files
 - Ryan's Notebook
 - Reece's Notebook
 - Hannah's Notebook
 - Telma's Notebook
 - Chris's Notebook
 - [Final Notebook](#)
- Back to the main on repository link you will find each team member individual branch with a copy of the Notebooks file from main

Libraries

Below we have all the libraries we used for our code (also referenced on the repository final notebook):

- Pandas
- Numpy
- Matplotlib
- Seaborn
- Scipy.stats
- Math from Scipy.stats
- Warnings Library

Recommendations

During our analysis for Computing Vision we were striving to provide 3 recommendations that help their firm succeed in creating a new film studio. Our first recommendation is film genre selection. The least successful film Genre's include History, Music/Musical, Sports, and Western films based off of ratings and revenue. We can confidently advise that Computing Vision strive to create movies in the Action or Adventure Genre's and stay clear of the least successful ones. Our second recommendation is personnel. After analysing the most successful writers, directors, and actors by net revenue we can suggest the following options for a successful film.

Actors/Actresses: - Scarlett Johansson - Chris Evans - Steve Carell

Writers: - Chico Paul - Suzanne Collins - Neal Purvis

Directors: - Pierre Coffin - Francis Lawrence - Sam Mendez

Our final recommendation is movie runtime. Using a one tail T-Test we tested our hypothesis that there was a statistically significant increase in runtime for movies with the highest ratings and revenue. After we tested the data we concluded that the top 100 movies by revenue and ratings had significantly higher runtime than average films and that Computing Vision should keep movie runtimes between 118 minutes -130 minutes to reap the most benefit.