SteveNdirangu / dsc-phase-1-project-v2-4 Public

forked from learn-co-curriculum/dsc-phase-1-project-v2-4

View license

0 stars 439 forks

Star Watch

Code Pull requests Actions Projects Wiki Security Insights Se

master

This branch is 8 commits ahead of learn-co-curriculum:master. Contribute Sync fork

SteveNdirangu Finished Project … 2 minutes ago **26**

View code

README.md

Phase 1 Project

Project Overview

Business Problem

Microsoft sees all the big companies creating original video content and they want to get in on the fun. They have decided to create a new movie studio, but they don’t know anything about creating movies. You are charged with exploring what types of films are currently doing the best at the box office. You must then translate those findings into actionable insights that the head of Microsoft's new movie studio can use to help decide what type of films to create.

The Data

In the folder `zippedData` are movie datasets from:

- Box Office Mojo
- IMDB
- Rotten Tomatoes
- TheMovieDB
- The Numbers

I used datasets from these 3

Box Office Mojo IMDB The Numbers

Loading the datasets

tn_df = pd.read_csv("zippedData/tn.movie_budgets.csv.gz", index_col=0) imdb_df = pd.read_sql("""SELECT * FROM movie_basics JOIN movie_ratings USING(movie_id);""", conn) bom_df = pd.read_csv("zippedData/bom.movie_gross.csv.gz")

Data Cleaning

IMDB Dataset

imdb_df["genres"].fillna("missing", inplace=True) imdb_df.drop(columns=["movie_id","original_title"], inplace=True)

Box office Mojo Dataset

bom_df.drop(columns=["foreign_gross"], inplace=True) bom_df = bom_df.rename(columns={'title': 'movie'})

The Numbers Dataset

This was how I hanged the values to number type, and got rid of Dollarsigns and Commas

tn_df['domestic_gross'] = tn_df['domestic_gross'].str.replace('$', '').str.replace(',', '') tn_df['production_budget'] = tn_df['production_budget'].str.replace('$', '').str.replace(',', '') tn_df['worldwide_gross'] = tn_df['worldwide_gross'].str.replace('$', '').str.replace(',', '')

tn_df["domestic_gross"]=pd.to_numeric(tn_df["domestic_gross"])

Combining the datasets with Merge

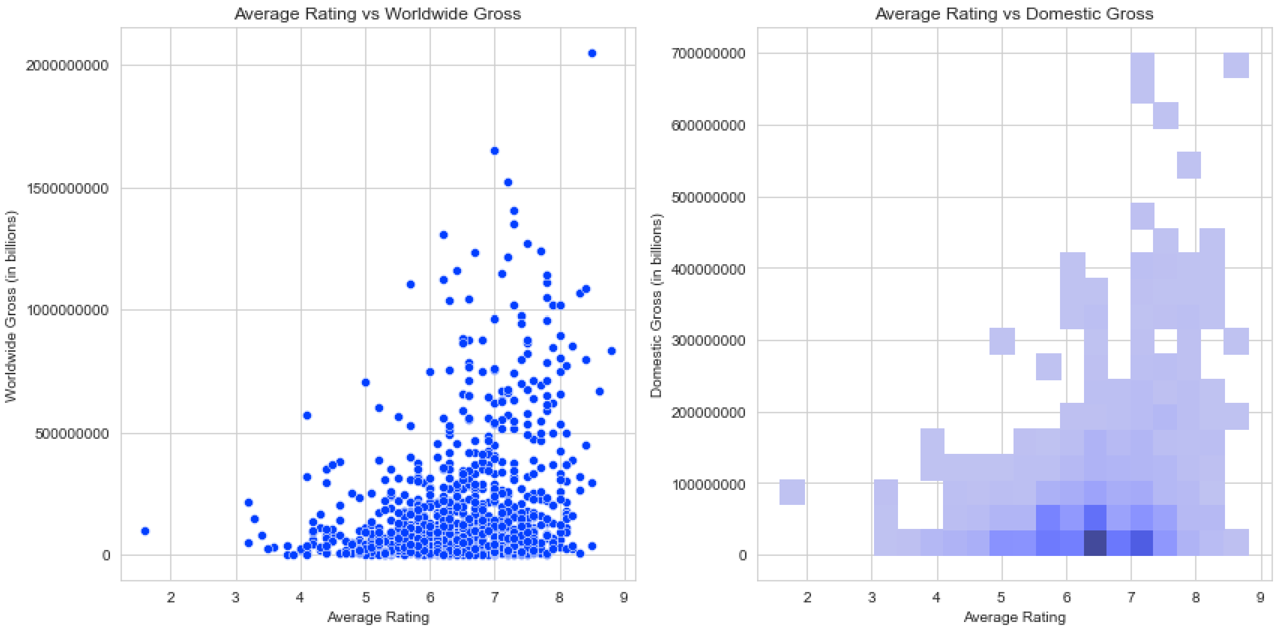
merged_df = imdb_df.merge(bom_df, on=['movie', 'year']).merge(tn_df, on=['movie', 'year'])

creating a grouped dataset to compare monthly data

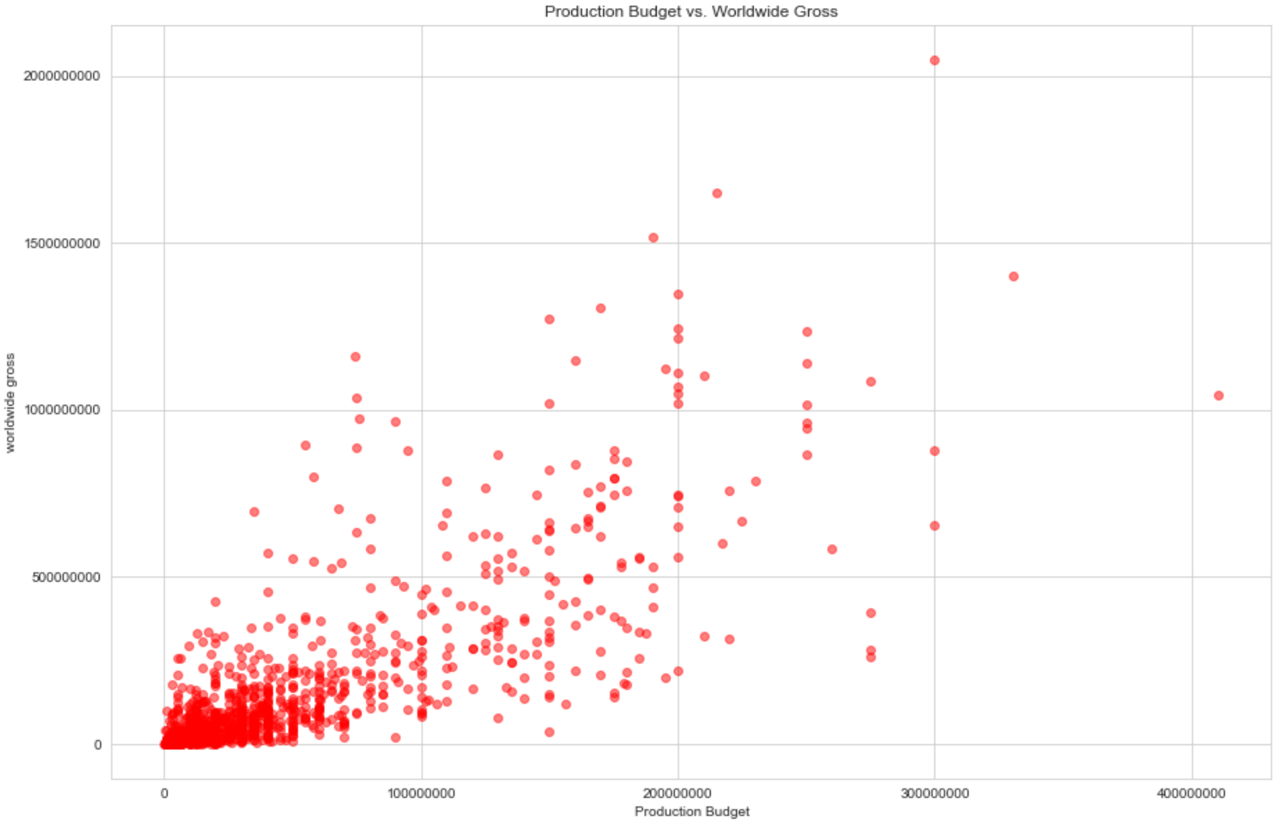
merged_df['month'] = merged_df['release_date'].dt.month grouped_df = merged_df.groupby(['year', 'month'])["worldwide_gross"].mean()

Visualizations

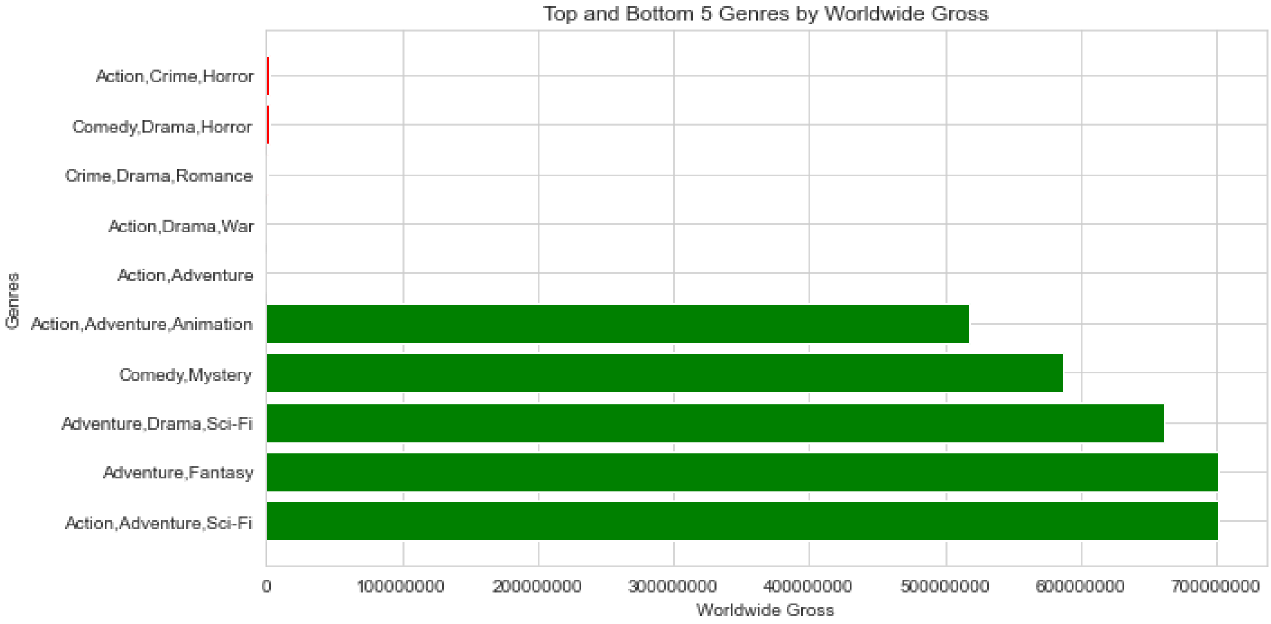
Scatter plot and Histogram to compare the Average rating and Worldwide Gross as well as Domestic



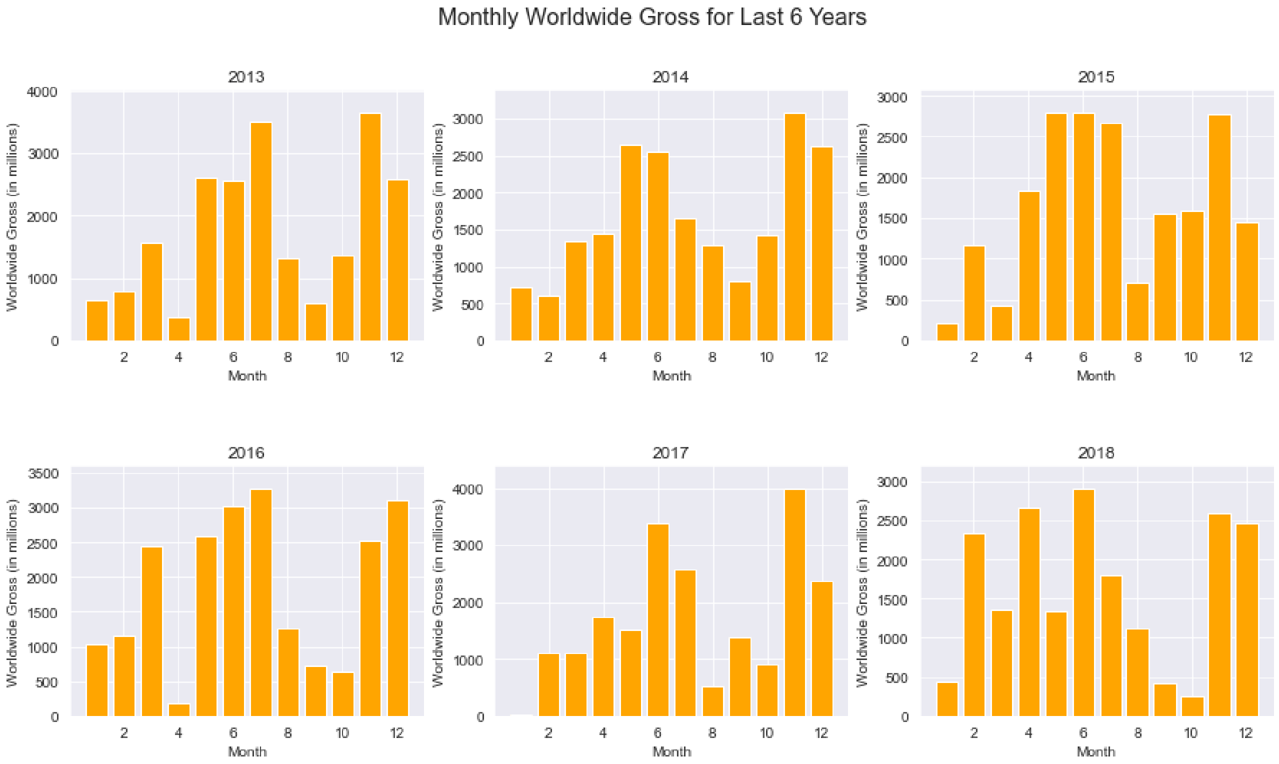
Checking whether Production Budget has a relation with the worldwide gross



Top and Bottom 5 Genres with worldwide gross returns in the box office



Per year look at the worldwide Gross mean of each month in the past 6 years



Monthly Worldwide Gross for Last 6 Years

2013 2014 2015 2016 2017 2018

Findings

Budget

The production budget has very strong positive correlation with the domestic and worldwide grosses

Average rating

-movies rated higher did much better in the box office,both for domestic and worlwide audiences, and in fact, movies rated lower than 5 did very poor

-curiously, the rating isn't really affected by increased budget, probably due to the fact that many things go into a production

-it would seem the average rating has a weak positive correlation with both domestic and worldwide gross,

Genre

Generally genres dont really affect a rating, but the top 5 that make money worldwide are anything of the "adventure genre" involving scifi, animation and comedy

Runtime

-Rating has a strong positive correlation with the runtime of the movie, but closer investigation shows the dirstibution is clustered within a value range of 80 to 140 minutes

Month of release

-it would seem from the per-year and mean of 5 years bar plots, that in the middle of the year and in November, these are the best times to release movies

Conlusions an Recommendations

1. Mirosoft should invest in the Genres of Adventure, with scifi,comedy,and/or animation, with main projects being of the genre "Adventure,Drama,Scifi" as it gets high reviews as well as high grossing worldwide

2. A high budget in these genres will give better returns worldwide, somewhere above the 100 million mark

3)A runtime of Between 80 minutes and 140 minutes is the most consistent at good ratings

4. Releasing between months 4-6, and at the tail end of the year might also give good worldwide grosses...probably because these are the months in which holidays occur eg easter, christmas

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

● **Jupyter Notebook** 100.0%