Cinema Revenue Analysis

Marriah Lewis

IST 652 Scripting for Data Analysis
Professor Deborah Landowski
Syracuse University // Summer 2021

Table of Contents

Introduction

Pre-pandemic the movie industry was old enough, and the box office numbers were stable enough for blockbusters, which seemed to be a good criterion for judging the performance of movies. The COVID19 pandemic has had a major impact on the film industry in 2020 and reflects its impact on all arts. Globally, cinemas and movie theaters are closed to varying degrees, film festivals are cancelled or postponed, and movie release dates are postponed or postponed indefinitely. In 2020, the U.S. box office plummeted by 80%, and global revenue plummeted by 71% because of the pandemic. The goal of the study is to examine the budget and gross income across thousands of movies throughout North America to determine the distribution of profit across production companies. This study also examines the average votes and popularity to assess if a movie will equate to profits.

Data Sources

Kaggle, Statista, and Wikipedia were used for analysis.

Data source 1: **TMDB 5000 Movie Dataset**
Columns:

- homepage
- id
- original_title
- overview
- popularity
- production_companies
- production_countries
- release_date
- spoken_languages
- status
- tagline
- vote_average

Data source 2: **IMDb movies extensive dataset**
Columns:

- imdb_title_id
- title
- original_title
- year
- date_published
- genre
- duration
- country
- director
- production_company
- actors
- avg_vote

- votes
- budget
- usa_gross_income
- worldwide_gross_income
- metascore

Data source 3: Statista: Global-box-office-revenue-COVID and AMC-Theatres-revenue-2006-2020

- 2020 fiscal year, AMC Theatres reported annual revenues of 1.24 billion U.S. dollars, a dramatic decrease from previous years because of the COVID-19 pandemic.
- The impact of the COVID pandemic on worldwide box office revenues has reduced the estimated figure from 44.5 billion U.S. dollars to 16.3 billion for the year 2020.

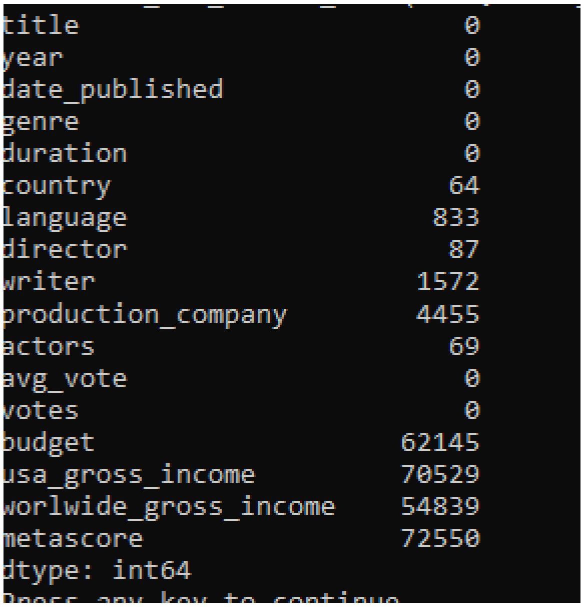
Data source 4: Wikipedia: Film series
URL: https://en.wikipedia.org/wiki/Film_series
Columns:

- Franchise
- Total gross
- Movie count
- Average gross
- Highest-grossing film
- Gross
- First

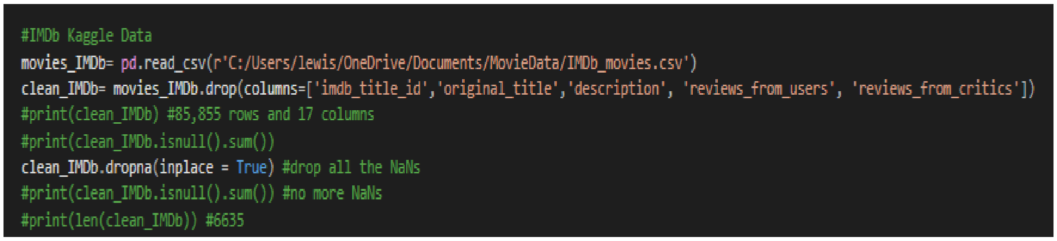
Pre-processing

IMDB and TMBD was downloaded from the Kaggle. The highest gross film franchises based on box office gross was web scraped from Wikipedia. COVID19 revenue prediction and AMC revenue dataset was downloaded from Statista. For each dataset, NaNs and zeros were removed using the following code below. Originally, the IMDB dataset had 85,855 rows and 17 columns after the NaNs were removed, there were 6,635 rows and 17 columns due the column metascore having 72,550 NaNs shown below in Figure 1.

Figure 1: IMDB NaNs and the code used to remove the NaNs and zeros



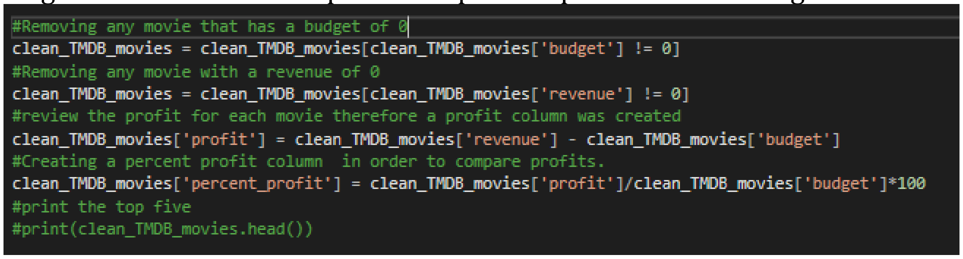
```
title                      0
year                       0
date_published             0
genre                      0
duration                   0
country                   64
language                 833
director                  87
writer                  1572
production_company      4455
actors                    69
avg_vote                   0
votes                      0
budget                 62145
usa_gross_income       70529
worlwide_gross_income  54839
metascore              72550
dtype: int64
Press any key to continue
```



```
#IMDb Kaggle Data
movies_IMDb= pd.read_csv(r'C:/Users/lewis/OneDrive/Documents/MovieData/IMDb_movies.csv')
clean_IMDb= movies_IMDb.drop(columns=['imdb_title_id','original_title','description', 'reviews_from_users', 'reviews_from_critics'])
#print(clean_IMDb) #85,855 rows and 17 columns
#print(clean_IMDb.isnull().sum())
clean_IMDb.dropna(inplace = True) #drop all the NaNs
#print(clean_IMDb.isnull().sum()) #no more NaNs
#print(len(clean_IMDb)) #6635
```

The same code was used to remove the NaNs and zeros from the TMDB dataset. For the TMDB and IMDB dataset unnecessary/unused columns were removed. Each dataset was saved in a pandas dataframe. A profit column was created in the clean TMDB dataframe. After the profit column was created, a percent profit column was added as well to compare profit.

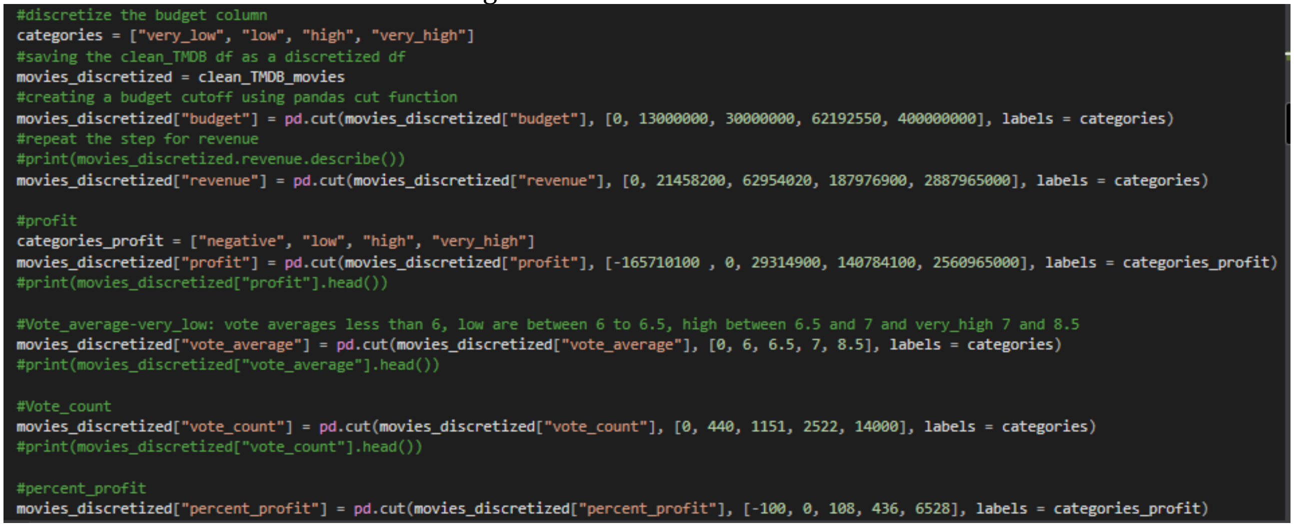
Figure 2: Creation of the profit and percent profit column using basic math



```
#Removing any movie that has a budget of 0
clean_TMDB_movies = clean_TMDB_movies[clean_TMDB_movies['budget'] != 0]
#Removing any movie with a revenue of 0
clean_TMDB_movies = clean_TMDB_movies[clean_TMDB_movies['revenue'] != 0]
#review the profit for each movie therefore a profit column was created
clean_TMDB_movies['profit'] = clean_TMDB_movies['revenue'] - clean_TMDB_movies['budget']
#Creating a percent profit column  in order to compare profits.
clean_TMDB_movies['percent_profit'] = clean_TMDB_movies['profit']/clean_TMDB_movies['budget']*100
#print the top five
#print(clean_TMDB_movies.head())
```

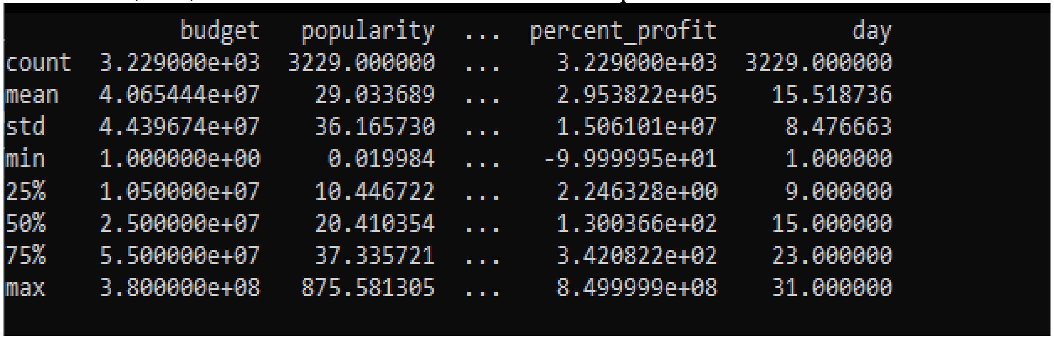
The release_date in the TMDB dataframe was convert to date/time and separated by month, day, and year. The month data type was change from int to ordered category. The budget, revenue, profit, vote_average, vote_count, percent_profit, and the days column was discretized using the cut function from pandas.

Figure 3: Discretized Columns



```
#discretize the budget column
categories = ["very_low", "low", "high", "very_high"]
#saving the clean_TMDB df as a discretized df
movies_discretized = clean_TMDB_movies
#creating a budget cutoff using pandas cut function
movies_discretized["budget"] = pd.cut(movies_discretized["budget"], [0, 13000000, 30000000, 62192550, 400000000], labels = categories)
#repeat the step for revenue
#print(movies_discretized.revenue.describe())
movies_discretized["revenue"] = pd.cut(movies_discretized["revenue"], [0, 21458200, 62954020, 187976900, 2887965000], labels = categories)

#profit
categories_profit = ["negative", "low", "high", "very_high"]
movies_discretized["profit"] = pd.cut(movies_discretized["profit"], [-165710100 , 0, 29314900, 140784100, 2560965000], labels = categories_profit)
#print(movies_discretized["profit"].head())

#Vote_average-very_low: vote averages less than 6, low are between 6 to 6.5, high between 6.5 and 7 and very_high 7 and 8.5
movies_discretized["vote_average"] = pd.cut(movies_discretized["vote_average"], [0, 6, 6.5, 7, 8.5], labels = categories)
#print(movies_discretized["vote_average"].head())

#Vote_count
movies_discretized["vote_count"] = pd.cut(movies_discretized["vote_count"], [0, 440, 1151, 2522, 14000], labels = categories)
#print(movies_discretized["vote_count"].head())

#percent_profit
movies_discretized["percent_profit"] = pd.cut(movies_discretized["percent_profit"], [-100, 0, 108, 436, 6528], labels = categories_profit)
```

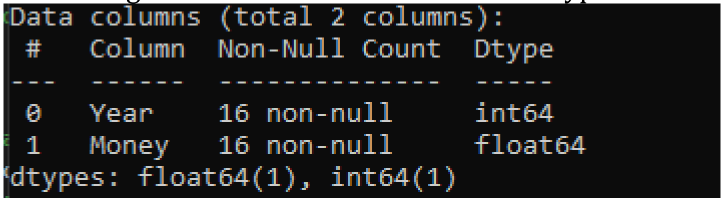
The budget column was discretized into four groups: very_low, low, high, very_high by creating a list of categories. Very low budget are budgets less than 13,000,000, low are budgets between 13,000,000 and 30,000,000, high are budgets between 30,000,000 and 62,192,550, and very_high are budgets between 62,192,550 and 400,000,000. The values are based on the quartiles.



```
            budget   popularity  ...  percent_profit          day
count  3.229000e+03  3229.000000  ...    3.229000e+03  3229.000000
mean   4.065444e+07    29.033689  ...    2.953822e+05    15.518736
std    4.439674e+07    36.165730  ...    1.506101e+07     8.476663
min    1.000000e+00     0.019984  ...   -9.999995e+01     1.000000
25%    1.050000e+07    10.446722  ...    2.246328e+00     9.000000
50%    2.500000e+07    20.410354  ...    1.300366e+02    15.000000
75%    5.500000e+07    37.335721  ...    3.420822e+02    23.000000
max    3.800000e+08   875.581305  ...    8.499999e+08    31.000000
```

The revenue column was created the same way with alternative numbers as shown in Figure 3. The profit column differs in categories due to the minimum showing a negative (categories= ['negative', 'low', 'high', 'very_high'], ordered=True). Pertaining to the day column another column was created for weeks. Week_1 is the first seven days of the month, week_2 is days eight through fourteen, week_3 is days fifteen through twenty-one, and week_4 is the remaining days. The Statista was a clean dataset the only cleaning/prep was renaming the columns. No datatypes needed to be converted for the Statista dataset.

Figure 4: AMC Theatre revenue datatypes



```
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   Year    16 non-null     int64  
 1   Money   16 non-null     float64
dtypes: float64(1), int64(1)
```

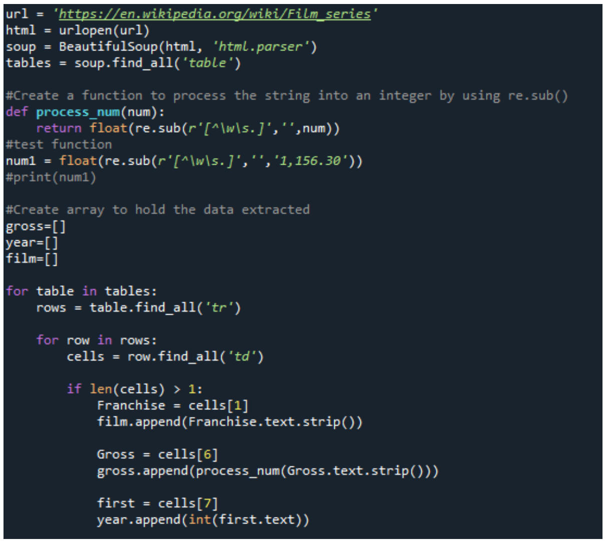
Web Scaping

Libraries used:

- from bs4 import BeautifulSoup
- import pandas as pd
- from urllib.request import urlopen
- import re

Using BeautifulSoup, Wikipedia highest gross film franchises based on box office gross was scaped. Using this dataset, I wanted to explore what year had the most movies release and by which production which in turn lead to the secondary question, do major production companies have an impact on the percent of profit (profit margin)? One of the first steps was creating a function to process the string into an integer by using re.sub(). After creating the function, arrays were created to hold the data that was extracted shown in Figure 6. The data was then placed in a pandas dataframe.

Figure 6: Wikipedia Web Scaping



```
url = 'https://en.wikipedia.org/wiki/Film_series'
html = urlopen(url)
soup = BeautifulSoup(html, 'html.parser')
tables = soup.find_all('table')

#Create a function to process the string into an integer by using re.sub()
def process_num(num):
    return float(re.sub(r'[^\w\s.]','',num))
#test function
num1 = float(re.sub(r'[^\w\s.]','','1,156.30'))
#print(num1)

#Create array to hold the data extracted
gross=[]
year=[]
film=[]

for table in tables:
    rows = table.find_all('tr')

    for row in rows:
        cells = row.find_all('td')

        if len(cells) > 1:
            Franchise = cells[1]
            film.append(Franchise.text.strip())

            Gross = cells[6]
            gross.append(process_num(Gross.text.strip()))

            first = cells[7]
            year.append(int(first.text))
```

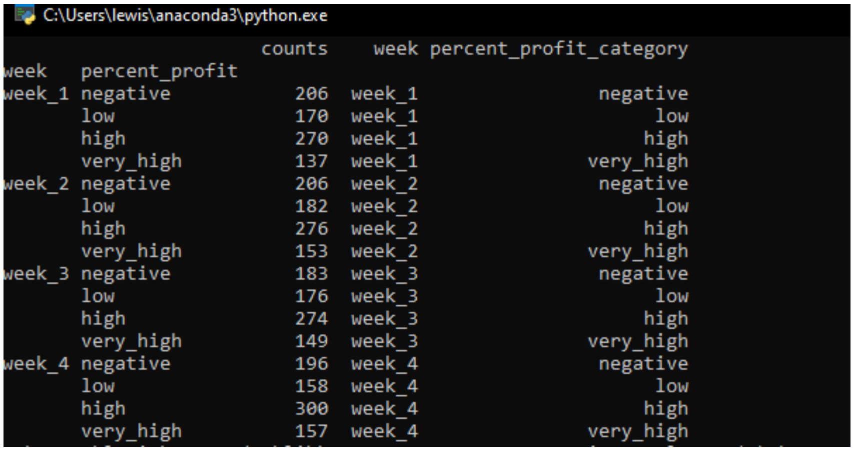
Method of Analysis

Topic Questions:

- Do major production companies have an impact the profit margin?
- Is it true that the month in which a film is released has an impact on its profit margin?
- How does budget impact vote average?
- In which year most movies were released?

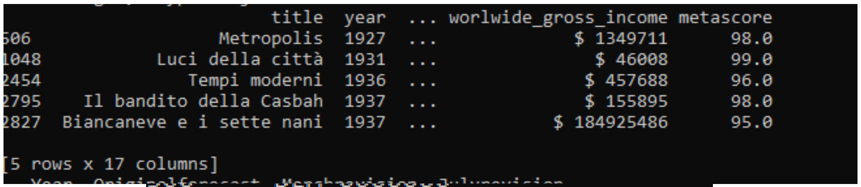
What fields were used:

- TMDB columns used in analysis:
  - Budget
  - Revenue
  - Percent_profit
  - Production_company
  - Week

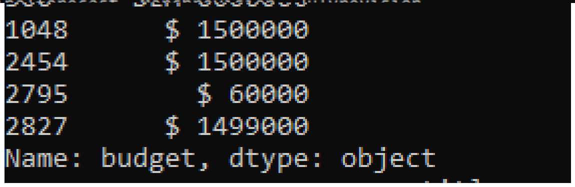


```
                          counts    week percent_profit_category
week   percent_profit
week_1 negative             206  week_1                negative
       low                  170  week_1                     low
       high                 270  week_1                    high
       very_high            137  week_1               very_high
week_2 negative             206  week_2                negative
       low                  182  week_2                     low
       high                 276  week_2                    high
       very_high            153  week_2               very_high
week_3 negative             183  week_3                negative
       low                  176  week_3                     low
       high                 274  week_3                    high
       very_high            149  week_3               very_high
week_4 negative             196  week_4                negative
       low                  158  week_4                     low
       high                 300  week_4                    high
       very_high            157  week_4               very_high
```

- IMDB columns used in analysis:
  - Budget
  - Avg_vote (only used this column)
  - Percent_profit (only used this column)
  - Worldwide_gross_income

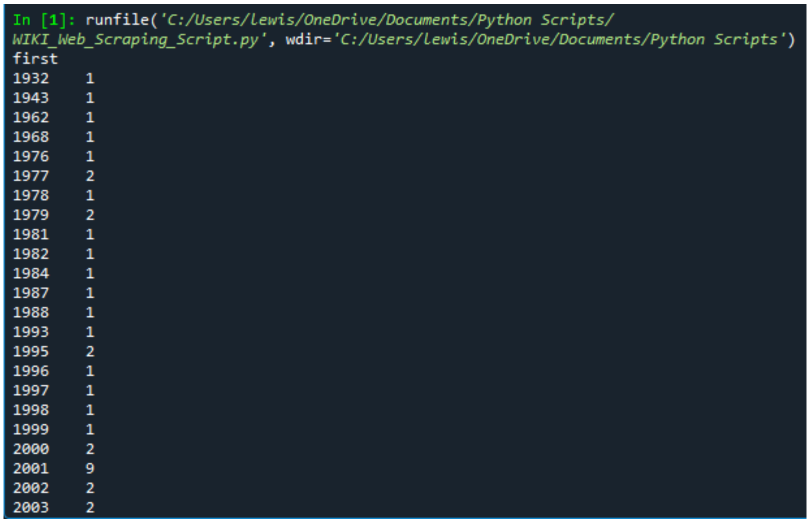


```
                          title  year  ... worlwide_gross_income metascore
506                  Metropolis  1927  ...             $ 1349711      98.0
1048           Luci della città  1931  ...               $ 46008      99.0
2454              Tempi moderni  1936  ...              $ 457688      96.0
2795     Il bandito della Casbah  1937  ...              $ 155895      98.0
2827  Biancaneve e i sette nani  1937  ...           $ 184925486      95.0

[5 rows x 17 columns]
```



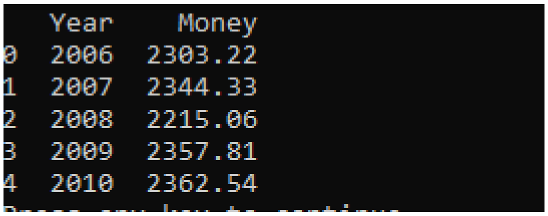
```
1048       $ 1500000
2454       $ 1500000
2795         $ 60000
2827       $ 1499000
Name: budget, dtype: object
```

- Wikipedia
    - Gross
    - First (year)
    - Franchise (film)



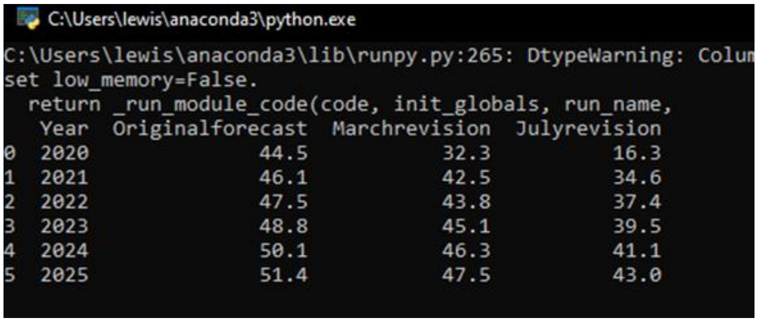
```
In [1]: runfile('C:/Users/lewis/OneDrive/Documents/Python Scripts/
WIKI_Web_Scraping_Script.py', wdir='C:/Users/lewis/OneDrive/Documents/Python Scripts')
first
1932    1
1943    1
1962    1
1968    1
1976    1
1977    2
1978    1
1979    2
1981    1
1982    1
1984    1
1987    1
1988    1
1993    1
1995    2
1996    1
1997    1
1998    1
1999    1
2000    2
2001    9
2002    2
2003    2
```

- Statista
    - AMC Theatre dataset
        - Year
        - Money



```
   Year    Money
0  2006  2303.22
1  2007  2344.33
2  2008  2215.06
3  2009  2357.81
4  2010  2362.54
```

-
    - COVID revenue impact 2020-2025
        - Originalforecast
        - Marchrevision
        - Julyrevision



```
C:\Users\lewis\anaconda3\python.exe
C:\Users\lewis\anaconda3\lib\runpy.py:265: DtypeWarning: Colu
set low_memory=False.
  return _run_module_code(code, init_globals, run_name,
   Year  Originalforecast  Marchrevision  Julyrevision
0  2020              44.5           32.3          16.3
1  2021              46.1           42.5          34.6
2  2022              47.5           43.8          37.4
3  2023              48.8           45.1          39.5
4  2024              50.1           46.3          41.1
5  2025              51.4           47.5          43.0
```

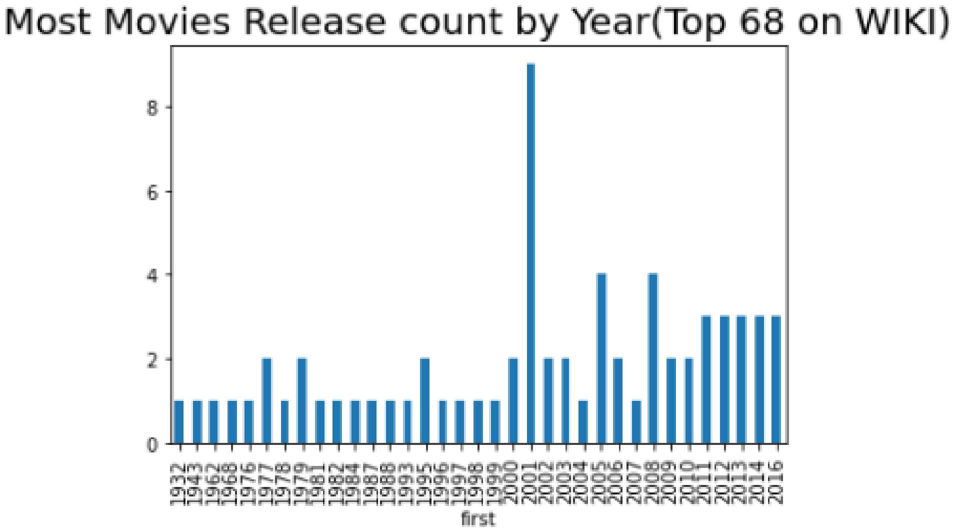
Used count for each budget level, the count for each percent_profit level by budget level and divided the count of the percent_profit/count budget level and multiply by 100.

Program Description

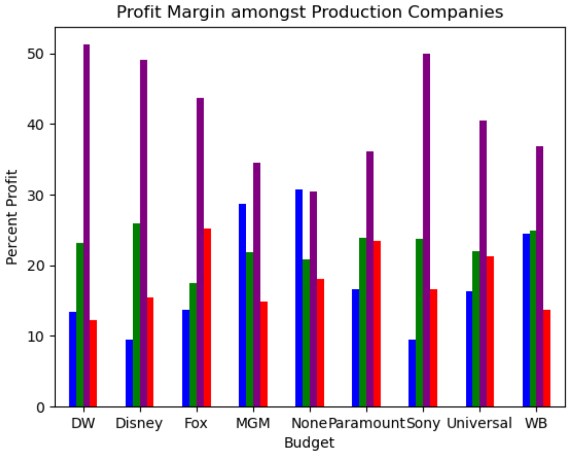
This program was used to understand what attributes contribute to the movie industry and its success. Is it based on budget, popularity, or is it just random luck? How much did COVID impact the movie revenue? This program will answer all those questions by considering average votes, budget, revenue, and percent profit.

Analysis & Interpretation

Question: In which year most movies were released?



Question: Do major production companies have an impact on profit margin?



On the graph, the color indicates the following:

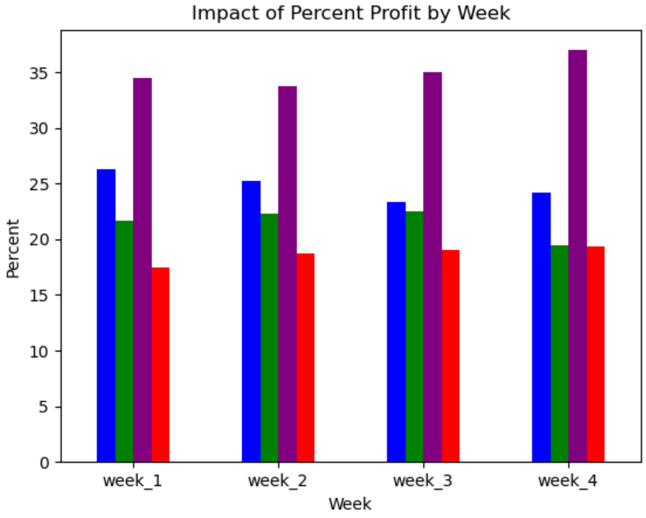
Blue: very low

Green: low

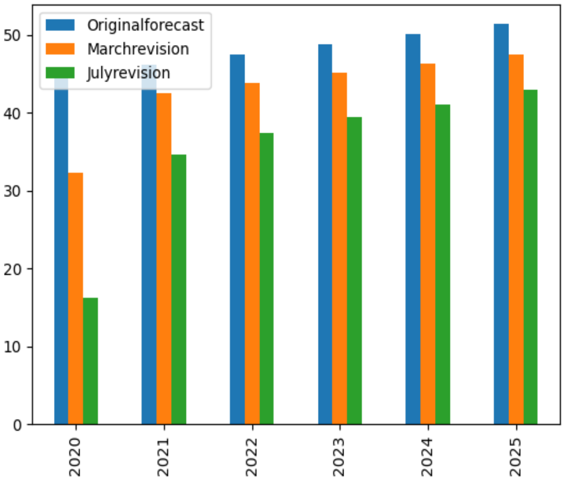
Purple: high

Red: very high

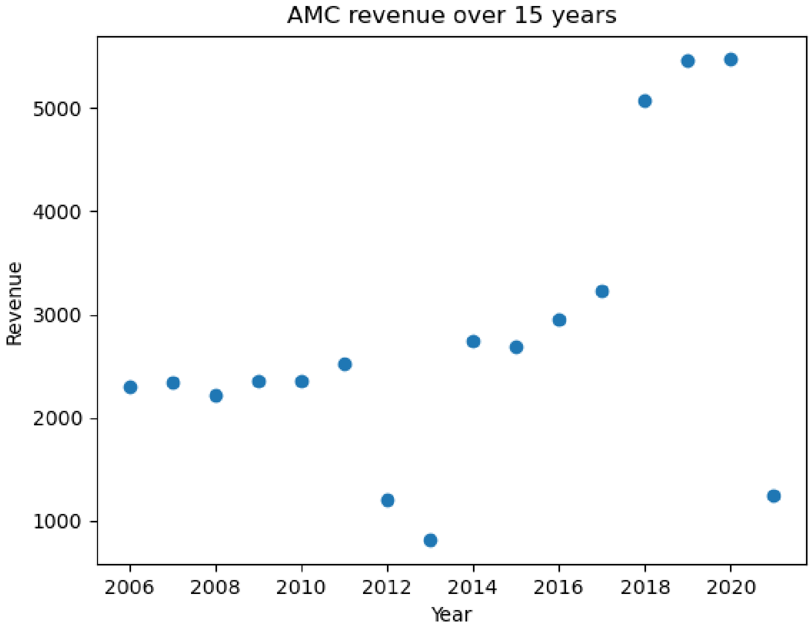
Question: Does it matter when the movie was released?



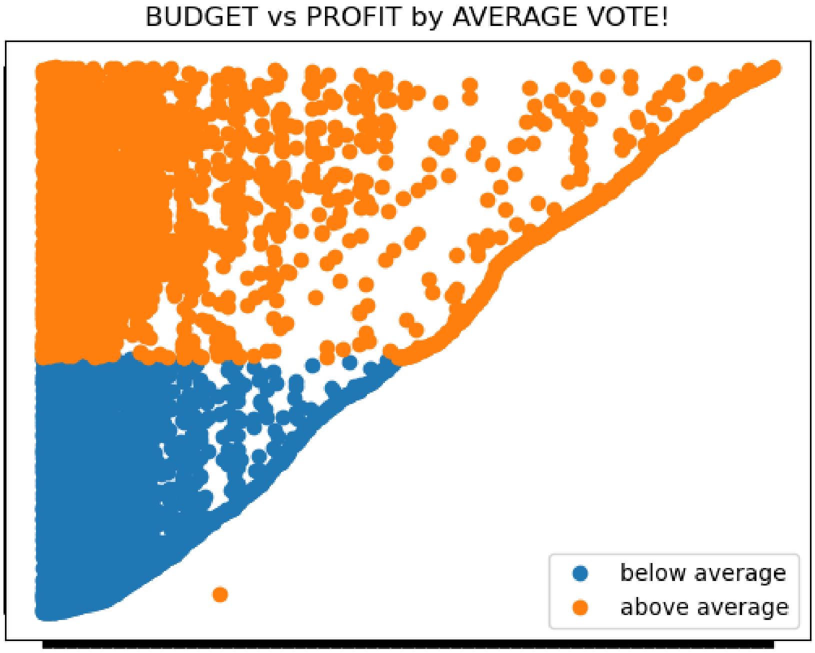
COVID REVENUE PREDICTION

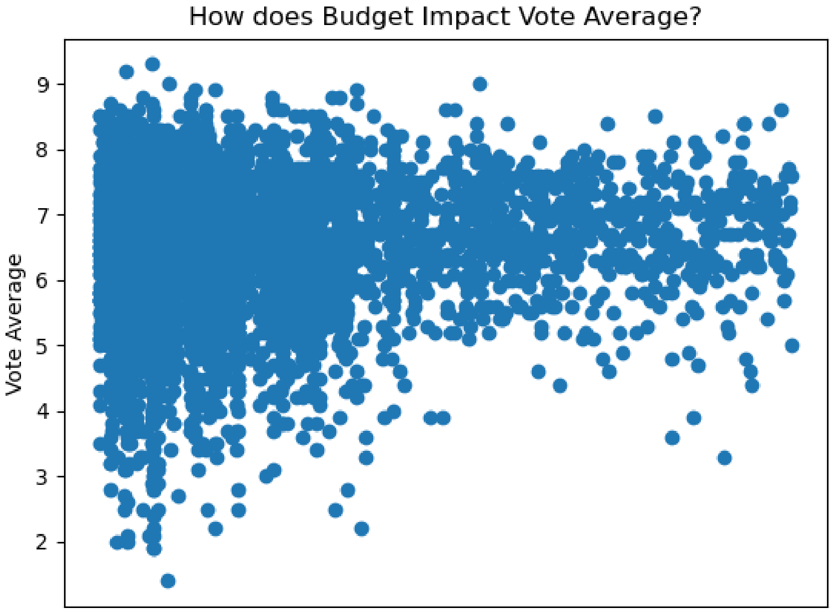


AMC Theatre Revenue over 15-year period



Does budget impact vote average?





Conclusions

The initial investigation came from the web scraping from Wikipedia where in 2001 the top gross movies were released but by who was the question and how did these production companies generate this much revenue to be the top grossing movies. It's worth noting that the lowest budget for a film in the df is $1, while the most budget is $380,000,000. The smallest profit is $5. The greatest amount of money that can be made is $2,787,965,000. This is a huge sum of money. A profit column was created, with a minimum profit of $-165,710,100 and a maximum profit of $2,550,965,000. The popularity scores span a wide range. The highest popularity score is 875.581305, with the lowest being 0.019984. The average popularity score is 29, and 75% of the films have a score of less than 37.3. This raises the question of whether the highest popularity score is an error or if it corresponds to the top-grossing film. The range of the vote average column is 0 to 8.5, with an average of 6.3. The maximum number of votes is 13,752. This could be a reference to the same movie that was the most popular. This column's maximum value appears to be an outlier as well. The major production companies graph provides insights, Dreamworks(DW) and Sony have high profit value whereas Paramount and Fox have very high profit percentage but these companies have extremely high budgets. Another intriguing observation was that in weeks 3 and 4 the percent of profit increased could this because customers wait sometime after the premier to avoid crowds or is it due to social media buzz. For example, the newly released Shang Chi movie is still at the top of the box office after it was released September 3.

Pandemic impact

The impact of the COVID outbreak on global box office sales has cut the predicted sum for the year 2020 from 44.5 billion dollars to 16.3 billion dollars. The July forecasts show a more severe impact than was predicted in March 2020, when revenue was estimated to fall to only $32.3 billion. Cinemas were shut down all around the world in Q2 2020, and the revenue loss is predicted to endure for the following five years, while moderate annual increase is still expected in 2021. According to the graph, movie revenue will level out by 2025 but the movie industry will never be the same.

AMC has about 11 thousand screens worldwide, with the vast bulk of them in the United States. Most of the company's revenue comes from ticket sales, but food and beverage sales have been a steady source of income over the years. As shown in the graph, there was a serious decline in revenue 2012-2014 and then another decline during the peak of the pandemic.

Limitation of the Study/Future Analysis

Due to the small sample size and the exploration nature of the study, there needs to be additional data collect to get a more accurate revenue prediction and percent profit. Also, a confusion matrix was conducted but did not show any significance therefore was eliminated. SVM analysis would be another option to use to produce a more cumulative accuracy score. Along with mining tweets from Twitter or mining Facebook post to understand how weeks 3 and 4 generated more revenue than weeks 1 and 2 could it be due to social media buzz.

Appendix A: References

U.S. Box Office Plummets 80%, Global Revenue Drops 71% in 2020 Amid Pandemic. (2021). Variety. https://variety.com/2021/film/box-office/box-office-final-revenues-2020-coronavirus-pandemic-1234879082/