NW Data Science Bootcamp (Fall 2022)

Exploratory Data Analysis
Top 500 Movies (1991-2022)

By: Geetha Kandukuri, Sandhya Datla, Frank Lindwall

Introduction

The following is a report of our team's exploratory analysis of the 500 most expensive movies to make in cinematic history. Included is a description of the data, the analytical questions that guided our analysis, our findings, and potential limitations of our project.

Data

Source of our dataset was Kaggle. There were top 500 movies by production cost ranging from 91 million USD to 400 million USD with release years from 1991 through 2022. Dataset was in csv format and layout is described in the reference section. (Layout of dataset)

A couple of calculated fields were added to the dataset.

- **Add-ons**
  IMDB rating from OMDB API call
  Adjusted values for production cost and gross revenue (domestic and worldwide) values using CPI index from Nasdaq API call.
  Month from release date
- **Adjusted value calculation formula**
  Adjusted value = unadjusted value * (Latest CPI/CPI of month of Release date)
- **Cleanup**
  There were many missing values from the base dataset. Basic cleaning was done to eliminate rows missing release date, title, production cost, domestic gross and release year having values greater than current year.
  Note: Original row count of 500 was reduced to 470 after cleanup. Vital information was lost due to missing values.
- **Final dataset**
  Base dataset from csv was read into dataframe, Year+Month from release date (after cleanup) added as a new column. CPI return csv was read into a new dataframe and Year+Month added as a new column and the two dataframes were merged using the YearMonth column. New columns for adjusted values, month of release and IMDB rating added.

Analysis

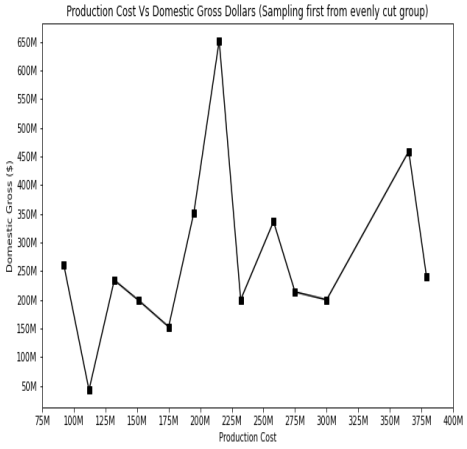
Our team analyzed the following:

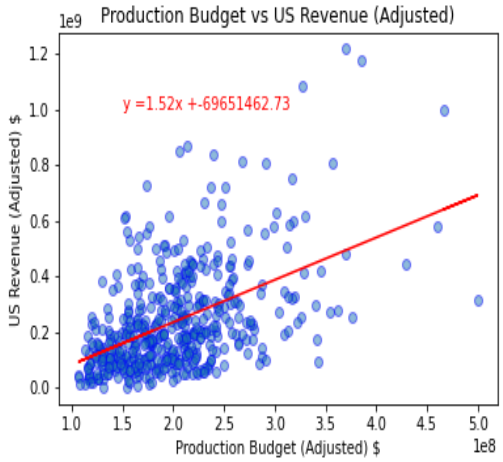
- Is there a correlation between Production cost Vs Gross Margin?
- Review IMDB rating and gross margin.
- Is there a correlation between runtime and domestic gross?
- Which month of release stands out for highest gross margin?
- How to adjust cost and gross values for inflation?

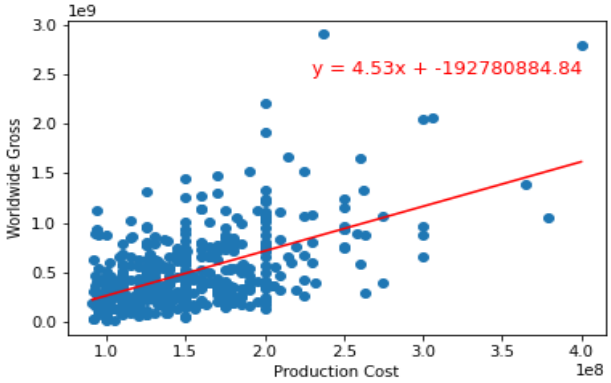
1. **Is there a correlation between Production cost Vs Gross Margin?**

   An eyeball analysis from a sample cut at equal interval showed a very slight positive correlation. But the sample was too small and could not be relied on. Scatter plot with regression of domestic gross and worldwide gross against production cost showed a moderate positive correlation with r value close to 0.5

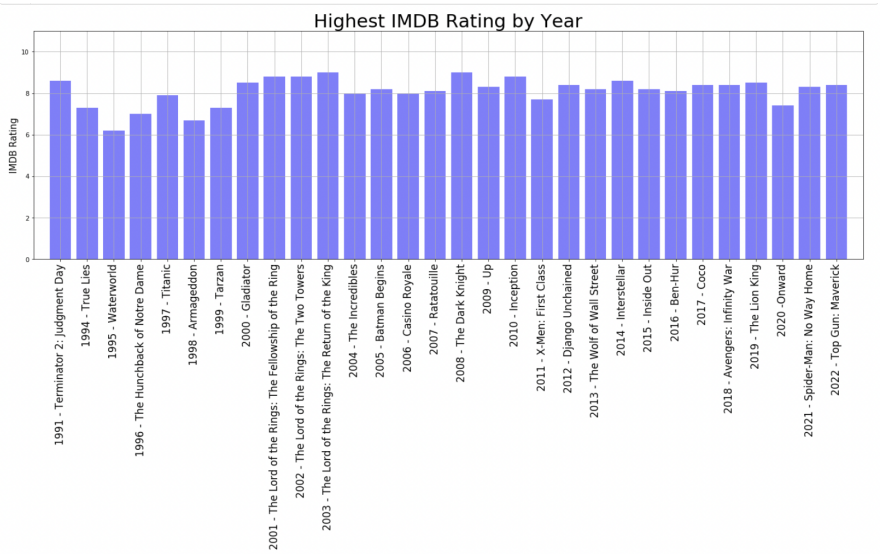
   Backed by positive correlation, it makes sense that higher production cost might be for new features and sets etc. but analyzing some movies that had low gross, there have been box office disasters even for higher production costs. When it comes to the strength of the model, it is relatively weak.







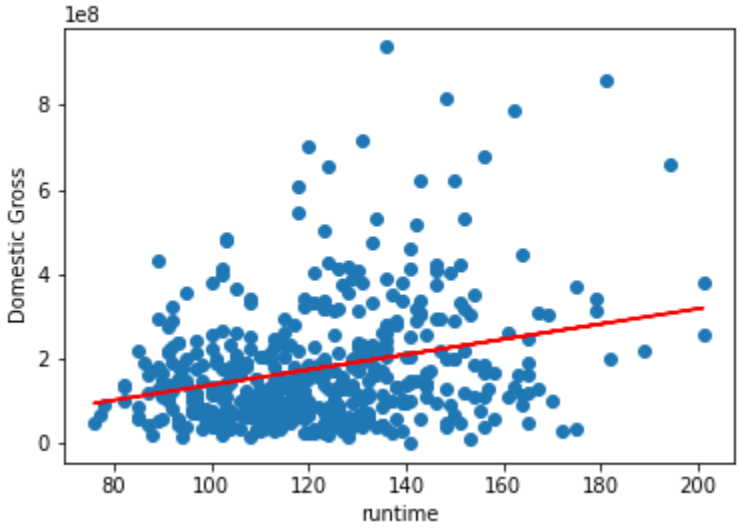
2. **Review the list of highly rated movies every year**



The above graph shows the movie with the highest IMDB rating of its respective year of the movies in our dataset. Of the thirty movies shown, only eight of them have an IMDB rating under eight. It's interesting to note that six of these movies with a rating under 8 were made before 2000.

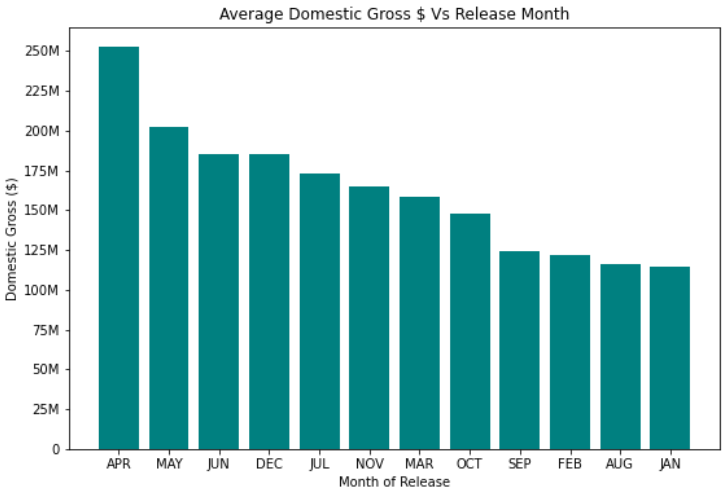
3. **Is there a correlation between runtime and domestic gross revenue?**

   There is some correlation between runtime and domestic gross revenue from the plot below. However, audiences do not specifically choose a longer movie over a shorter one as a sole determinant. Longer movies have significantly less screenings compared to the shorter movies per day. Which means box office numbers can be affected. Also ticket price for longer movies is certainly higher which is also an important factor. Biggest hits of all time had longer screening times and higher ticket price per person. Since we didn't have that specific data available. It was not taken into account. Run time alone cannot affect the domestic gross revenue significantly.



4. **Which month of release stands out for highest domestic gross margin?**

   Average of domestic gross grouped by month column was used here. Dataset was cleaned to have values in both columns to ensure that mean value is not affected. Result of bar plot listed below. Discussion on why **April** showed as the highest profitable month, made us come to the conclusion on spring break as one of the reasons. May, June and December were next in line.

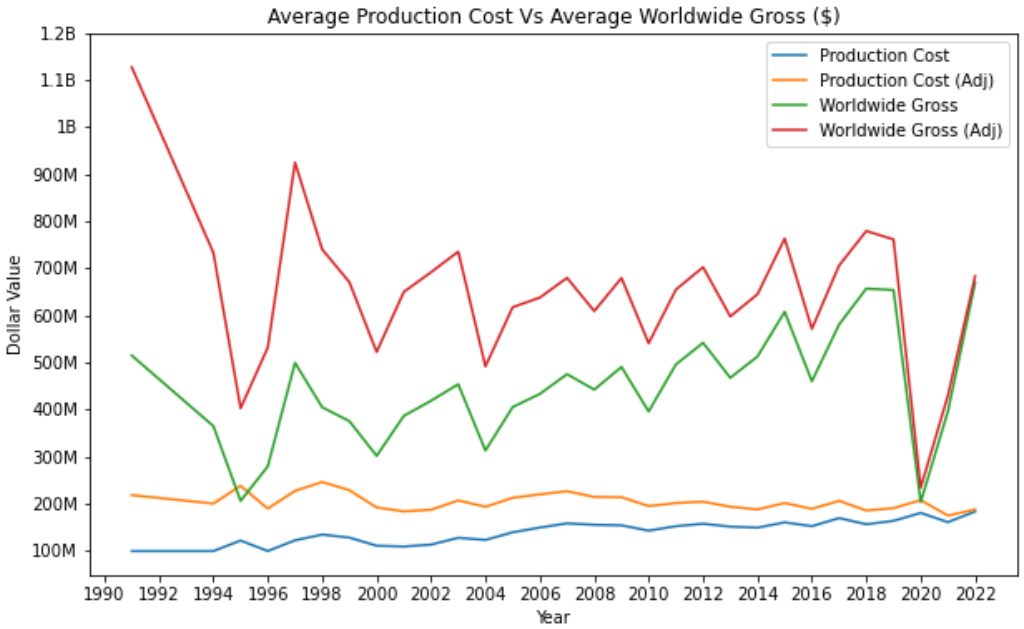


5. How to adjust cost and gross values for inflation?

The original dataset covering the top 500 movies by production cost over a 22 year span did not have values adjusted for inflation. Domestic gross was adjusted for inflation using both the CPI from the month of the release year and October 2022. The Nasdaq API was used to get the CPI values from 1991 through 2022. The top 10 movies ordered by adjusted vs non-adjusted show a variance in topper list.

Title	Domestic Gross $
Star Wars Ep. VII: The Force Awakens	936662225
Avengers: Endgame	858373000
Spider-Man: No Way Home	814108407
Avatar	785221649
Top Gun: Maverick	715270567
Black Panther	700059566
Avengers: Infinity War	678815482
Titanic	659363944
Jurassic World	652306625
The Avengers	623357910

Title	Domestic Gross (Adj)
Titanic	1218216787
Star Wars Ep. VII: The Force Awakens	1180156783
Avatar	1083614530
Avengers: Endgame	1001007460
Spider-Man: No Way Home	870202059
Star Wars Ep. I: The Phantom Menace	850902577
Black Panther	837886315
Jurassic World	814602879
Avengers: Infinity War	807417238
The Avengers	807389171



Conclusion

- **Limitations**
    - The missing values in the dataset cause a loss of vital statistics
    - A comparison between genre and gross margin was not possible because the majority of the movies in the list were labeled Action/Adventure.
    - It's important to keep in mind that this is only an analysis of the top 500 most expensive movies to make, so we can't extrapolate trends we see in this dataset to all movies. For example, we can't generalize a correlation that we found between IMDB rating and gross revenue to all movies ever made.
- **Foresight**
    - Review a couple different datasets for a given topic before finalizing.
    - Be more diligent with consistent graph labels and standards for measuring data.

Works Cited

- Dataset
    - Kaggle.com
- Websites
    - StackOverflow
    - https://dfrieds.com (function to format ticks for large numbers)
- APIs
    - OMDB API for IMDB Rating
    - Nasdaq API for csv of CPI values

References

Movie dataset layout (Top 500 movies by Production cost, ranging from 400M to 91M)

rank	Number in ascending order of rank (400M ranked as #1)
release_date	Release date of movie in USA format
title	Title of movie
url	Url of movie on The Number site
production_cost	Production cost for the movie
domestic_gross	Gross revenue from theaters in USA (in USD)
worldwide_gross	Gross revenue from theaters worldwide in USD (includes domestic)
opening_weekend	Gross revenue of opening weekend
mpaa	Motion Picture Association Rating
genre	Genre of the movie
theaters	Number of theaters the movie was screened
runtime	Movie runtime in minutes
year	Year of release

Nasdaq API resultant csv layout (1 row per month-end for the date range in API call)

Date	Date in ISO format (month-end)
Value	CPI value (rounded to 3 decimals)