<https://www.kaggle.com/datasets/sanjeetsinghnaik/top-1000-highest-grossing-movies>

I would say before covid I was definitely a movie buff. I regularly went to the movies and interested in cinematography. I choose to analyze the highest grossing Hollywood films because I probably watched most of these films before. Immediately my first question was is there any trends in the top grossing films? Can I identify any obvious trend amongst these movies? If by genre, rating, release data, etc. can there be a quantifiable explanation as to certain movies success. Without even looking at data I already knew studios tend to release their biggest films of the year during the summer months. Would this data support studios reasoning for releasing during those months? Also looking at total sales per movie based on genre would be intriguing. Would it be possible to show what genre occurs the most often? Which genre makes the most money? I had a lot of questions about the data before even looking at it.

The selected dataset is the “1000” (918) highest grossing Hollywood films. It includes movie title, description, distributor, release date, sales (domestic, international, and world), genre, movie runtime, and license (G, PG, PG-13, R). This is an excellent dataset for the questions I had. Has multiple columns that I could cross analyze to determine if the top movies have anything in common. The publisher of the data used web scrapping from multiple sites (IMDB, rotten tomatoes, etc.) to gather the information and added some additional data.

I wrongly assumed cleaning this data was going to be easy when I first found the dataset. I believed I was just going to fill in some missing NaN values and be ready for EDA. Firstly, the dataset came with an extra column ‘Unnamed:0’ that was just another index. Also, the data came with ‘Movie Info’. While that could be useful for some NLP questions about the movie descriptions, I wasn’t really interested in looking into it. Both columns were dropped. I then needed to rename each sales column. This was to make it easier in further code when referencing those specific columns. I then needed to further investigate the NaN values present in ‘Release Date’ and ‘License’. With ‘License’ I wasn’t comfortable filling the missing values because no category was predominant. I was hoping maybe PG-13 would be enough of the column that I could fill in the rest, but it wasn’t. As for ‘Release Date’ my idea was to just pull the year out of the title and not worry about month and day for these entries. ‘Movie Runtime’ and ‘Release Date’ needed their data types to be changed for easier analysis. ‘Genre’ was the category that I truly didn’t know how to approach at first. I had multiple ideas, and each was going to make the dataset a mess. It would involve splitting the genres because most of the movies had 3+ genres attached to them. Lastly, I investigated distributors and noticed some names came up multiple times. I renamed the studios that were directly tied to one another. If I had renamed studios by who they were owned by about 10 distributors would have been left.

Univariate analysis wasn’t going to be interesting for this dataset. The average runtime across 900+ films with 20+ distributors over almost 100 years of cinema wasn’t going to be insightful as to what was going on. None the less the average runtime was just under two hours. I was also interested in seeing which month was the most frequent. Not a surprise to see it was June. Or, that PG-13 was the highest license since it is the safest bet for studio to release movies as it garners the widest audience range.

My biggest challenge with multivariate analysis was having to deal with genre. I ended up splitting each genre labeled and constructing a new data frame because I wanted to know which genre was the most present before I started using that information. Once I saw which genres were the most prevalent, I wanted to bring in the sales data and start comparing. Seeing which genres made the most money total and which genre was the most profitable. No surprise the genre that appeared the most often made the most money. But since there was differences in the count and sum data, I wanted to see which for the most profitable. I was genuinely surprised when Sci-Fi returned as the most profitable movie genre. Moving away from genre I wanted to keep looking at the sales data but concentrate on the distributors. First, I did some simple graphs that showed Disney had the highest world sales. The graph below was a step above that since it includes domestic and international which totals the world sales for the highest earning distributors. This graph also clearly shows the relation between domestic and international sales. Often when a movie does well domestically it should do well internationally also.

A picture containing bar chart

Description automatically generated

The following graph shows total world sales across the years in the data set. I wanted to make this table just to see the peak right before the inevitable crash that was caused by covid. Crazy to see how low it reached in 2020 and stayed that low for 2021. This data shows that world sales are most likely to not return to their former peak for a couple of years.

Chart, line chart

Description automatically generated

The last graph I was interested in the year most prevalent in the data and how many movies made it on the list. 2014, 2013, and 2010 all had over 40 movies from each year on that list which is crazy to think about. That represents over 10% of the data. Next step I should have done was calculate world sales for those three years and compare to the total. I wouldn’t be surprised if it was near 30% of total world sales.

Chart, histogram

Description automatically generated

Take Away:

* PG-13 is the most common license
* Movie runtime across entire data set is 1 hour and 56 minutes
* June is the most common month
* Adventure is the most common genre
* Sci-Fi is the most profitable genre
* Disney has the highest world sales amongst all distributors
* Warner Bros. and Disney have the most numbers of movies in the data
* Covid had a catastrophic impact on world sales
* 2014, 2013, and 2010 are the years with the most movies in the data

Future Work:

1. Interesting data to add would be director and production cost of the films. I could see how much these movies made to how much they cost to make. Or see if there are common directors that regularly made it on the list over the years.
2. I would like to come back to this data when I learn more about graphing in python. I had ideas for complex graphs that I think would have been insightful, but I struggled just getting that stacked bar graph to print.