Natalia Edelson

Jones Morgan

June 2, 2022

Setting up Microsoft Studios for Success

Data Science, Project 1.

**Introduction**

Now that the Covid-19 pandemic is under control and people are resuming their usual activities, I am excited to return to one of my favorite places: the cinema. Previously, I would meet my friends on a regular basis to escape from our hectic lives through the magic portal of the so-called Big Screen. Within the sacred space of approximately two hours, I would travel to faraway lands, meet bizarre characters, laugh, cry, and cringe as we surfed the spectrum of human emotions until we finally emerged, ready to regain ourselves, our jobs, and our families at home.

The movie industry, which has always been an integral part of American identity, has suffered immensely during the Covid-19 crisis. The majority of its income having come from box office sales, studios struggled to adapt their content to meet the needs and desires of in-home streaming. Even the heavy-hitters, such as Walt Disney, Warner Brothers, and Sony, had to compete with platforms such as Netflix and Hulu who have dominated this market for the past two decades. However, now that theaters are back in business, what will it take to urge audiences from the living room sofa and out into the dark temple of the Big Screen? In this project, I complement my own personal experience as an avid movie-goer with data from the pre-pandemic era to predict the movie trends that will bring viewers back into theaters and thus revive the industry that has, in many ways, defined America.

**Purpose**

Microsoft is establishing a new movie studio. This project uses data analysis to advise Microsoft Studio (“MS”) toward a business plan that will increase gross domestic revenue, the total amount of money spent on buying movie tickets in the United States and Canada.

**Method**

We have determined that the best predictors of profitability for a given film are budget, release month, genre, cast and crew, movie length, and studio. We have collected data regarding these elements from the three largest available online databases – *IMBd, The numbers,* and *Mojo Box office* – from the years 2010-2018. During this time period, each one of these websites accumulated more than 100 billion dollars, a testament to the growing public interest in film culture and industry.

The analysis was coded using Python - focusing on Pandas, Seaborn, and

Matplotlib visualization. The code can be found on GitHub: [ offer link]

**Data and Analysis**

1. **Budget**

As expected, the budget is typically a good predictor of gross domestic, but only up to a certain point. Movies with a budget of less than $200 million show a direct relationship to the increase in gross domestic. However, a budget of more than $200 million did not correlate with higher figures.

[Do you have a graph for this??]

1. **Release Month**

The best months to release movies are June and November, followed by December, May and July. During these months, the gross domestic was the highest – three times more than the months of Jan, Sep or Oct. and twice as much as March and August.

Chart, bar chart

Description automatically generated

1. **Genre**

Of the top 20 % gross domestic performers, Adventure is by far the dominant genre, followed by Action, Comedy, and Sci Fi.

Analysis of the average domestic gross yielded slightly different results. Again, Adventure and Action came first, followed by Animation and Sci-Fi. From the budget perspective, Adventure, Animation Sci-Fi, and Fantasy are the dominant genres.

Chart, bar chart

Description automatically generated

1. **Cast and Crew**

Cast and crew members also determine the profitability of a production. The following performers appeared most in movies within the top 20% of gross domestic.

A picture containing table

Description automatically generated

1. **Movie Length**

The length of the movie concerns the budget and affects the audience’s willingness to watch it. Of the top 20% performing movies produced from 2010 to 2018, most ran between 90 and 140 minutes.

Chart, histogram

Description automatically generated

1. **Studio**

Studios play an important role with their established reputation in movie productions. Walt Disney, Universal, Fox, WB, and Sony are consistently among the 30 top performing studios.

Chart, histogram

Description automatically generated

**Conclusions:** Recommendations for Microsoft Studio

1. **Budget**: Budgets as low as $15 to $35 million show an increase in the movie’s gross domestic, but a budget above $200 million does not show the same rate of return. Do not exceed $200 million.
2. **Release Month:** Aim to release the movie in June or November; if not, then in December, July, May. June and November bring twice as much gross domestic as the other months.
3. **Genres:** Prioritize the production of Adventure, Action, Comedy, and Sci Fi movies, as these genres are the top performers.

1. **Cast and Crew:** Hire from the top performing cast and crew to increase chances for success. Certain names are associated with the most successful movies.
2. **Movie Length:** Produce movies that would range from 90 minutes to 140 minutes. Movies within this length of time are the most profitable.
3. **Studios:** Partner with established studios to produce movies, at least for the first year. This is a long-term investment that can lead to fruitful returns and greater independence in the future.

**Comments on the Coding Notebook**

Collecting Data

Setting Microsoft Studios up for Success

I fetch relevant data using SQL. Joining three tables: Principals, MovieBasics and Movie Ratings.

I check the size of the data. From the 3 tables I are getting around 60 thousand rows and 5 columns.

I load the other relevant data: Movie Gross and Movie Budget. I join the three data frames to create one master data frame which will allow me to examine multiple variables.

I load the data with the gross and budget numbers. I use inner join to join the three data frames. It joins only when it sees the match and then drops the rest. I am left with a little over 13,000 rows and 17 columns.

Cleaning the Data

I convert the titles to capital letters to make it look cleaner using lambda function.

I check the Null values in the data. Foreign\_gross column has 1,933 null values. I will narrow my focus analyzing gross domestic and therefore can drop the ‘Forign\_gorss’ all at ones. I drop the rows of the 'Null’ values since there are not that many of them and they won’t impact our data.

I drop the columns analysis

I visualize the Null values using heat map to get a better sense on the overall data.

I look closely into the stats to determine whether there are any outliers and understand the range of statistics. As of now, there are no outliers or place holders.

I drop duplicates.

# I plot the data looking at each year, The data is largely distributed uniformly.

I convert the units to $MM. Production budget’s data type shows as “object”. We need to clean the symbols that might be causing it.

I plot histogram using Seaborn to visualize the distribution for domestic gross.

The distribution is skewed to the left, where the mean is $60 million, the 25 percentile is $7 million, the median is $33 million, and the 75 percentile is $74 million. The maximum is $700 million. The majority of the data of gross domestic lies in the range of $10 million to $70 million.

Exploratory Data Analysis

Correlation

I plot the correlation using Seaborn heatmap to examine any relationships of the variables in our movies data.

#Corr represents the correlation data. The darker the blue hue color the higher the correlation.

The darker the blue hue color the higher the correlation.

I sum up the correlations that are higher than 0.65. Production Budget and Domestic Gross show a higher correlation.

Correlation table gives us a precise number which supports the heatmap.

What impact did the production budget have on domestic gross?

What impact does production budget have on domestic gross?

I examine closely the correlation among budget and domestic gross.

I plot a scatter diagram and noticed that the data is concentrated in the lower part of the graph (0-$150 million). I split the data into two parts to check if this strong correlation still holds.

I use regplot to compare the direction to which the line is headed. Below $200 MM the correlation is .66 whereas above $200 it is at .22.

Calculating the correlation of the two variables discussed.

Data where production is less than $200 million dollars.

Calculating the correlation of the split data – less than $ 200 million dollars.

I plot the

During what month(s) should Microsoft Studio release their movies?

#I created a lamba function to extract the month into a new column.

I compare the months to the gross domestic figured: July and November are the most profitable months to release a movie.

# I build a function where the create values, indices and palette in order to show the max value in a darker color.

Do rating show any relationship to Gross Domestic?

I used a box plot to compare rating versus gross domestic. I see that average raring of 6 and above has a small increase with gross domestic.

What Genres do best at the box office?

The genres column has data combined with different types of genres. I would need to separate them into each individual row to allow each genre to be considered separately.

I first create a list of the genres in each entry by using lamb function. The lamba function loops through each row.

I use the explode() method to move each individual genre into a different row. I check the count for each genre.

I took top 20 % domestic gross performing movies out of 13,376 and check what genre appeared the most.

I examine to see which genre appears in the top domestic gross and the top production budget.

Who should be part of Microsoft movie production?

I examine closely the cast and crew to determine whose names appear the most in the top 20 % gross movies.

What is the ultimate length of the movie.

I plotted a line plot.

There was no particular trend. Beside a few spikes/outliers, movies within 90 min to 2 hours stayed close to $100 million gross domestic. I saw an increase from two hours to three house the domestic gross hovered across $350 million gross domestic.

I focus then on the top 20% gross domestic and the data shows that the majority of the movies range from 90 mi to 150 min.

I show the below top performing studios.

The top three performing studios are Walt Disney, Warner Brothers and Sony.