Marc, Kevin, Destin, Ricardo, John

Data Mining

12 May 2021

Movie Data

Abstract: Summarizing the goals and approach of the work.

We will pull data off the website boxofficemojo.com about the top 100 grossing movies from the years 2015 through 2019. The unusual circumstances of 2020 had an outsized effect on movies. Theaters were forced to close, and digital streaming was used heavily, so movies in 2020 will not be evaluated in our data set. Our goal of this project is to find out what factors influence the total gross of a movie through the use of data analysis and data training. We will evaluate multiple attributes about each movie including genre, release date, number of theaters the movie was released in, total gross, distributor, and IMDB rating. Then we will find which of these attributes contributes to total gross.

The reason that we want to look into this data is that movies is one of the largest industries in the world. By analyzing this data, we will be able to get insights into which movies generate the most gross sales. Then we could predict based on genre or time of year or other characteristics, how well a particular movie would perform in theaters. This would be useful for business executives at movie production studios to make data-backed predictions on how well a particular movie would perform when it was released. They could also choose to release or not release a certain movie and strategically choose the number of theaters to release it in or the time of year to release it to optimize performance. All of these would be helpful in achieving the highest possible return on investment.

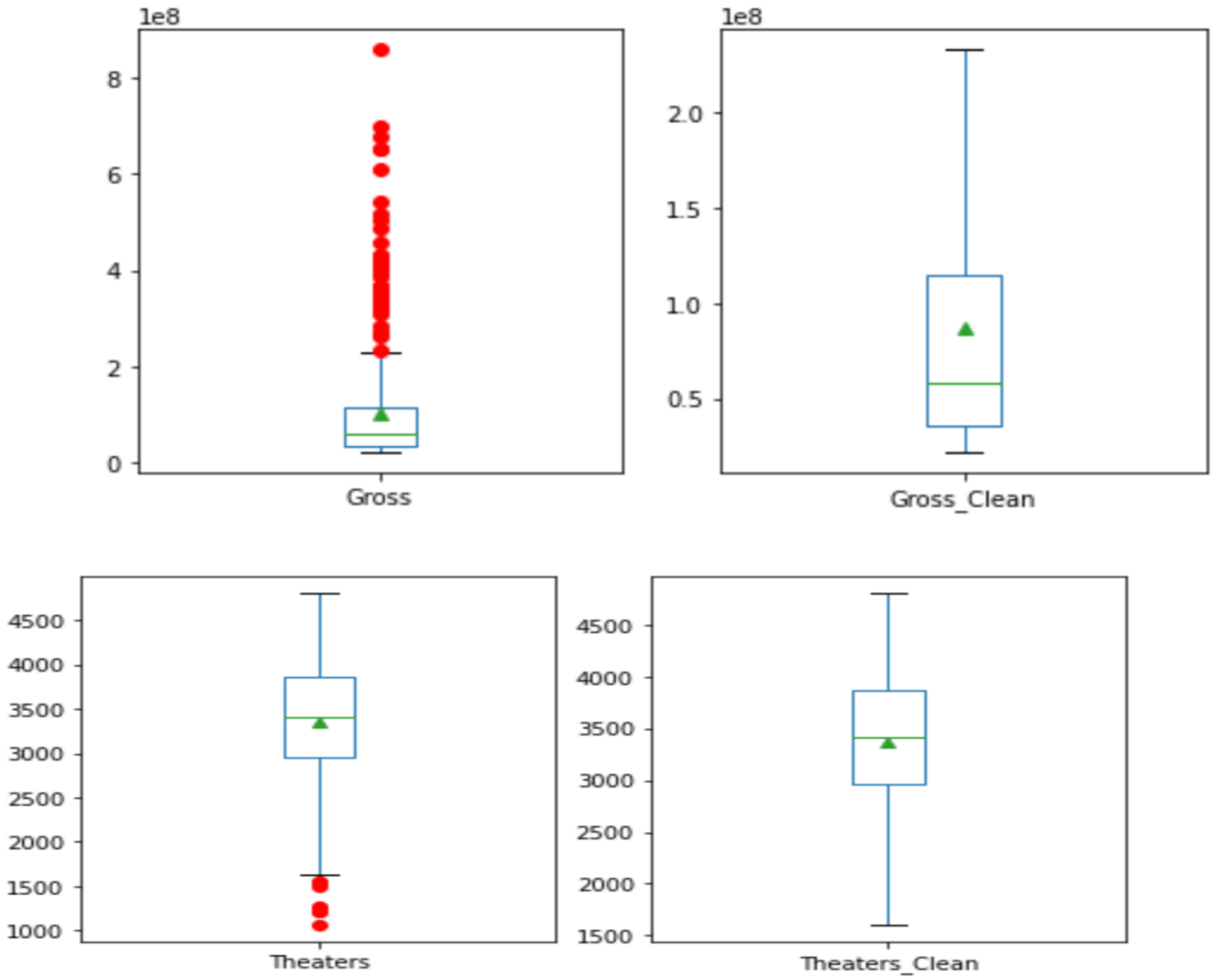
Our data from boxofficemojo.com was very high quality and there were no values missing in the attributes that we were looking for and we were not forced to delete any entire rows from our dataset. The values also seemed to be accurate and there were no instances where the data provided was clearly incorrect. Even though the data was high quality, we did have to deal with outliers from movies that may have performed extraordinarily well. Movies that perform extremely well may be part of a franchise such as Star Wars or Hunger Games, and that by itself would skew our data. Their gross would be very high because of anticipation from customers who have been waiting for the release, regardless of theaters, release date, etc.

We hypothesize that overall, the movies that finish with the highest total gross will be the movies that are released more recently than others. Movies that are released later on are more likely to gross more because the popularity of movies continues to increase over time as more people become part of the target age of audience. People also continue to look for more and more entertainment options. In addition, the government allows inflation to increase at a rate of approximately two percent per year so over a period of five years, this increase could add up to a fair amount of increased revenues for movies released more recently. Secondly, the movies that are released in the later months of the year will perform better than other movies because movies released late in the year have a bigger audience due to families being together during the holiday season. Lastly, movies that earn the highest total gross will be movies that are released in more theaters. Being released in more theaters would help any particular movie reach a wider customer base. The movie would be more easily accessible and potentially lead to more ticket sales and therefore gross revenues.

Experiments: Describing the experiments and the experimental setup, including the data sets description, the evaluation metrics, the data mining methods used, and any other details related to the experiments.

We cleaned and binned our data in a Google Sheet and then moved it into Google Colab. One of the first things we did was to create multiple copies of our data to use when analyzing and performing calculations on the data. This ensured that our original data would always be there and unchanged in case a mistake happened. We used Python code and the Python libraries numpy, sklearn, pandas, and matplot to examine our data and find our results. We trained a decision tree and drew graphs, bar charts, and box plots, to easily visualize our data. Below are some examples.

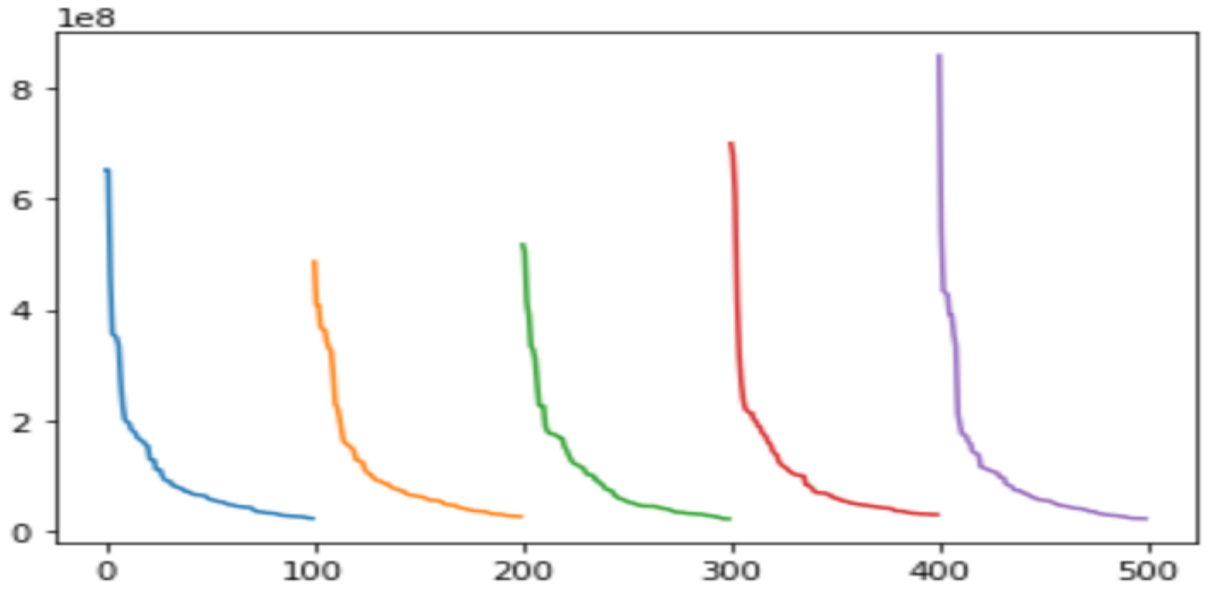
Here we can see that originally when we drew our box plot there were many outliers and that our interquartile range was quite small. Similarly, our upper and lower quartiles were close together. However, once we cleaned our data, we can see that we fixed our plot to expand the minimum and maximum values to fix our outliers. This also expanded our upper and lower quartiles and expanded our range of the box which allowed for our data to become more consistent and uniform. We performed a similar calculation on the number of theaters that a movie was released in because we found a similar phenomenon albeit, not so drastic compared to the gross.

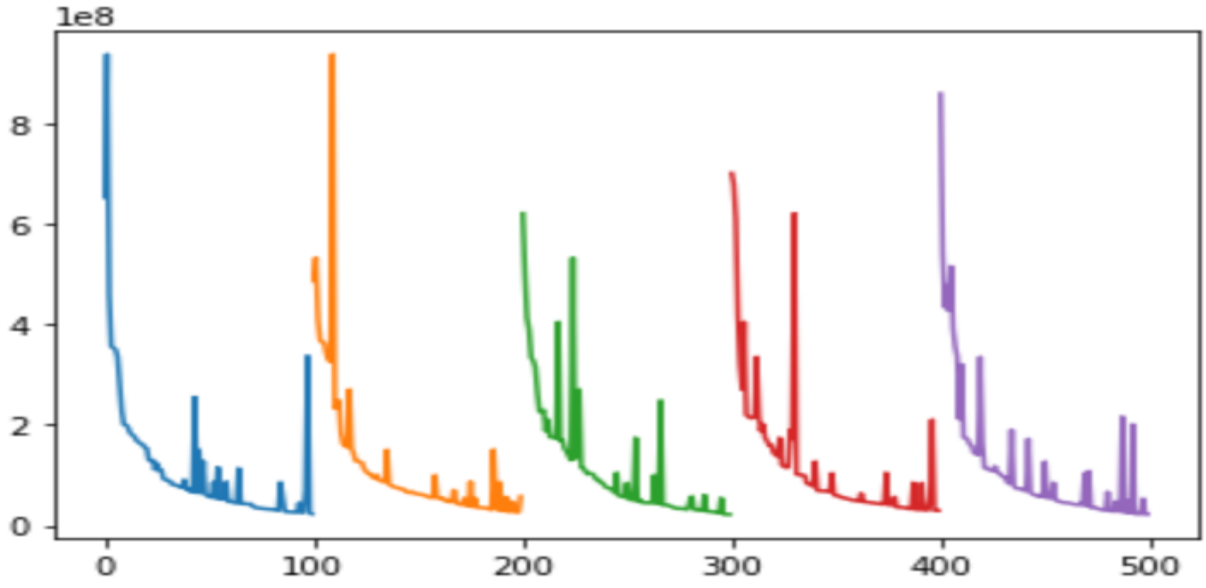


Results: Including the experiment results and discussion of the results.

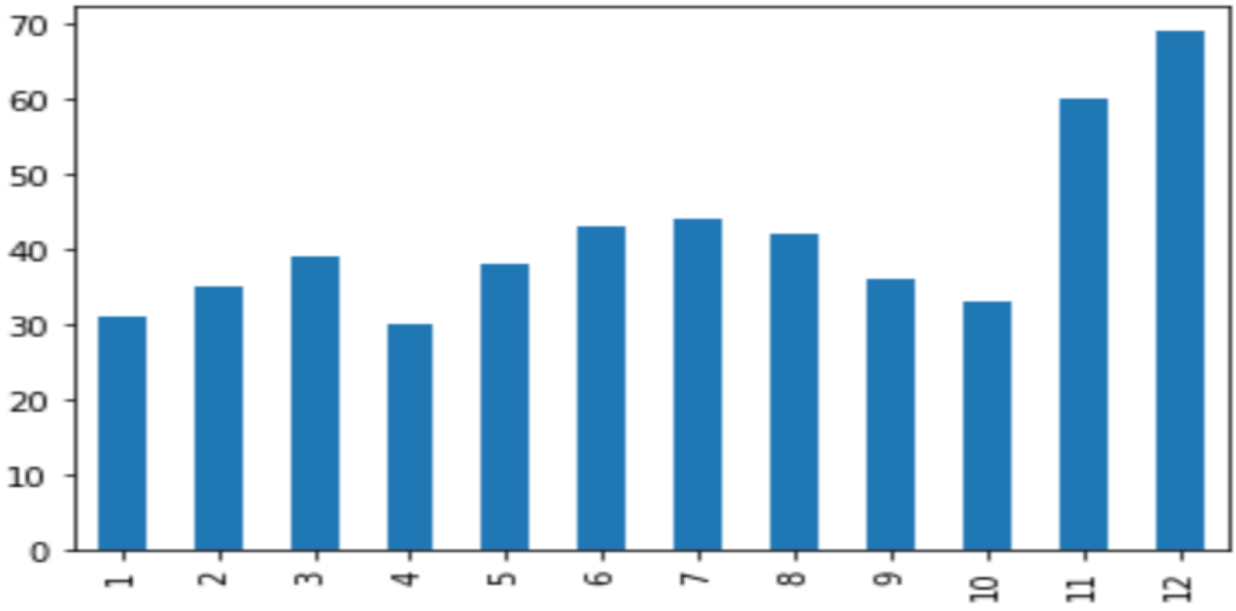
As we can see in our graph of the total gross and year, the overall trend is that movie revenue is trending upwards year over year. This is regardless of other factors that may influence the results, but in general each new year will produce higher revenue. We can also see in our next graph that the performance of a movie in the United States may be different than the performance of a movie internationally. Some movies that may not be received well in the United States but still produce a reasonable number for total gross by generating sales abroad. Our graph shows that many movies appear to follow this pattern. In general, there is a slight

bump in the total gross revenues of movies that are ranked near the bottom of US gross which implies that some movies may not be targeting American audiences. They may be looking for international sales and releasing movies in the US secondarily. Whereas the movies near the top of the rankings are releasing movies with a focus in the United States and less focus internationally.

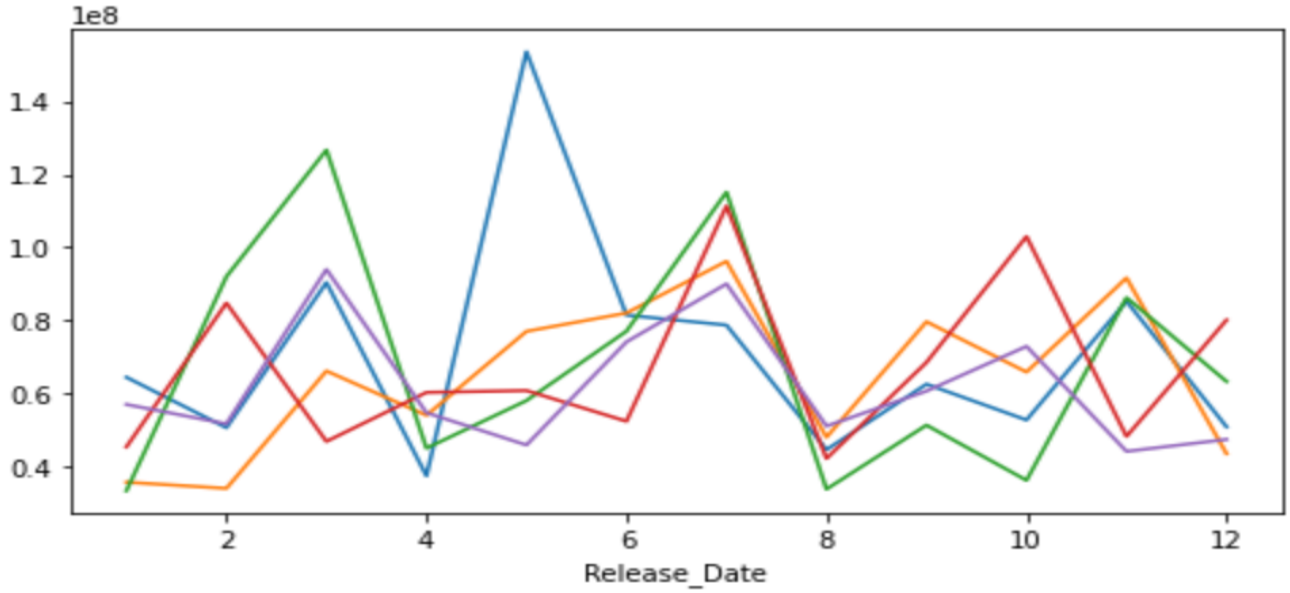




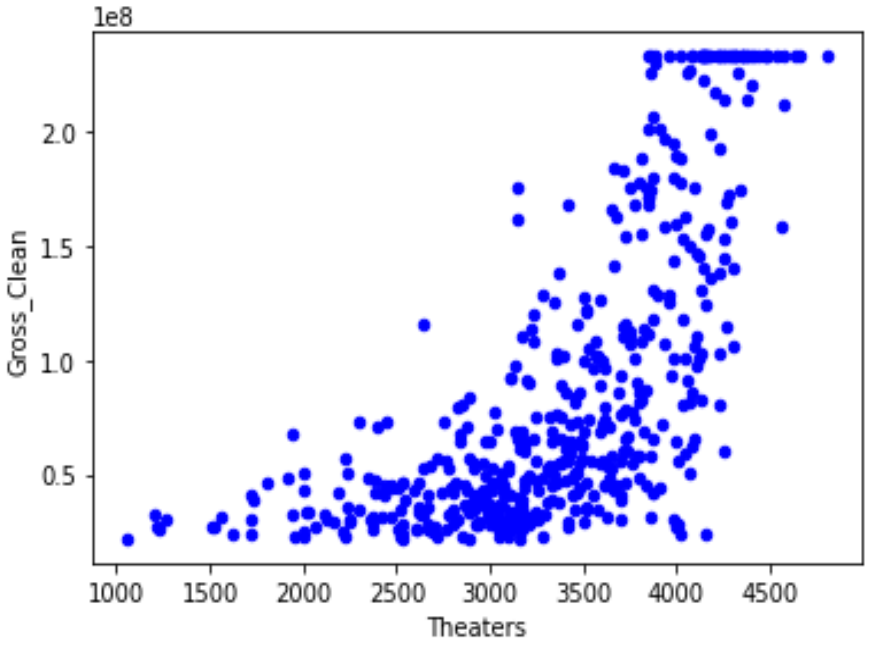
In the next bar chart, we can see that there is clearly a jump in the number of movies that are released late in the year, specifically in November and December. This is in line with our hypothesis that more movies would be released later on in the year and that movies released late would gross more than other movies. The spike for movies released in those two months seems to show a clear intention by studios to release films during the holiday months. These are typically times where families are more likely to be together and do things in groups. It is a time when children have more days off school than normal, and adults are more likely to take off days from work surrounding the New Year period. Studios seem to be anticipating customers will have time to go see new releases during this holiday season. The next jump in the number of movies released seems to be in the summer months. However, there is no particular one or two month period to target because summer can be varying lengths and times depending on the region. But in general we still see higher than normal number of movies being released in June and July versus the average.



However, we can see in the following chart, that the month that the movie is released in, actually does not correlate with the gross revenues of the movie. The following chart shows gross revenues by month and if we look at December, it is actually one of the worst performing months. This is actually contrary to what we had hypothesized which was that late year releases would perform better because families were spending more time with each other and children had days off school. We could possibly attribute this to the cold winter months where people are less likely to want to go outside.



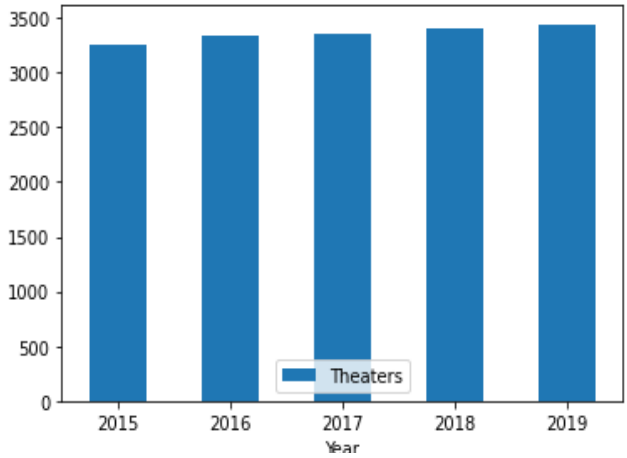
We also collected data to see whether the number of theaters a filmed opened in had a major impact on revenue.

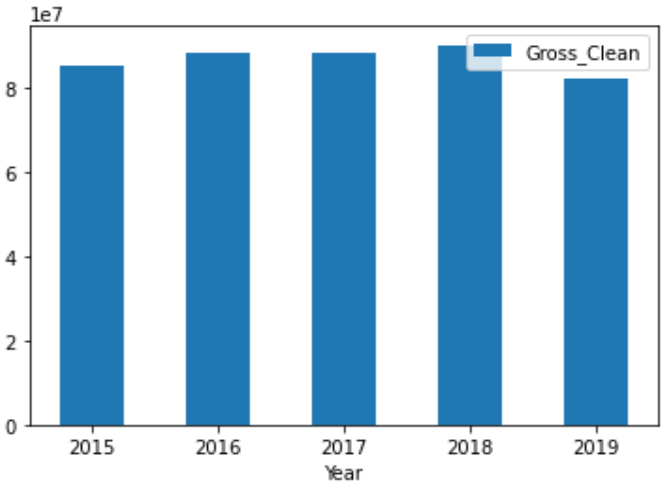


In the above scatter plot, each data point represents a movie. The X axis represents the amount of theaters opened in (getting larger as we go to the right side), and y-axis shows "Gross_Clean" (getting larger as we move vertical). As we can see from the scatter plot, it appears that movies releasing in <3000 theaters have very limited opportunity to achieve breakout revenue success comparatively to many of the other films studied. However, releasing in over 3000 theaters does not guarantee success -- but it does open up the possibility of achieving a more successful release by quite a bit. This is probably why there are more movies choosing to open up in over 3000 theaters than movies that choose to not. We still see some movies with over 4000 theaters released in who perform just as poorly or worse than movies releasing in around 2000 theaters (so, again, more theaters does not directly equal more success), but there are more movies opening up with at least 3500 theaters and achieving a higher revenue

than anything that opens up in under 2500 theaters than there are movies opening up in under 2500 period. In summary, to expand the possibility of success, it is recommended to open up in at least 3000 theaters.

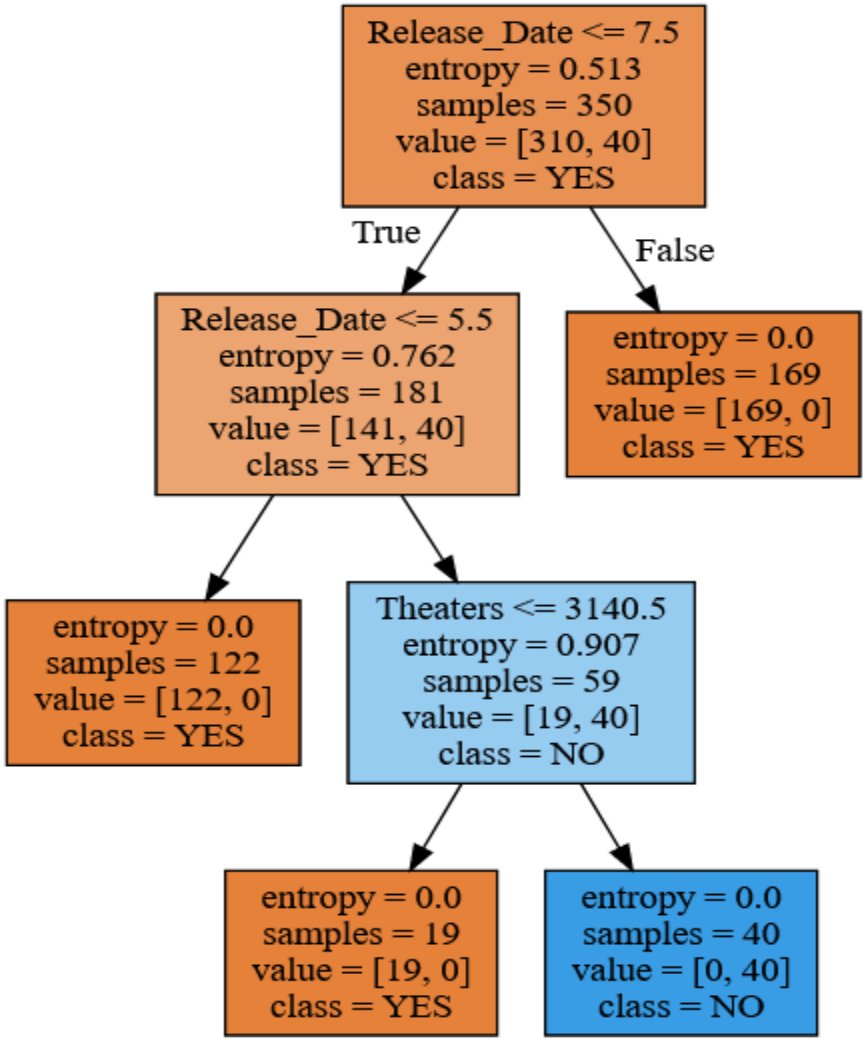
We also wanted to compare overall trends in industry growth, so we thought it would be a good idea to track changes in average theaters a movie opens up in against revenue to see if there was a correlation.





As we can see, the top graph represents avg theaters opened up in for a movie across the tracked years. Year-over-year, there is a slight increase in theaters opened up in on average, with may be a signal that theaters are opening up and certainly represents a form of industry growth. We would expect generally to see the average revenue also increase with more theaters being released in, so we graphed the bottom chart to show average revenue for films in a given year. To start off, our hypothesis seems generally true for every year before 2019. However, we see average revenue drop off by a somewhat small but definitely noticeable margin. This surprised us, as movies were on average being released in more theaters. Still, these are very slight variations from year to year in the grand scheme of things, and the correlation here isn't pronounced enough to be able to draw any major conclusions from.

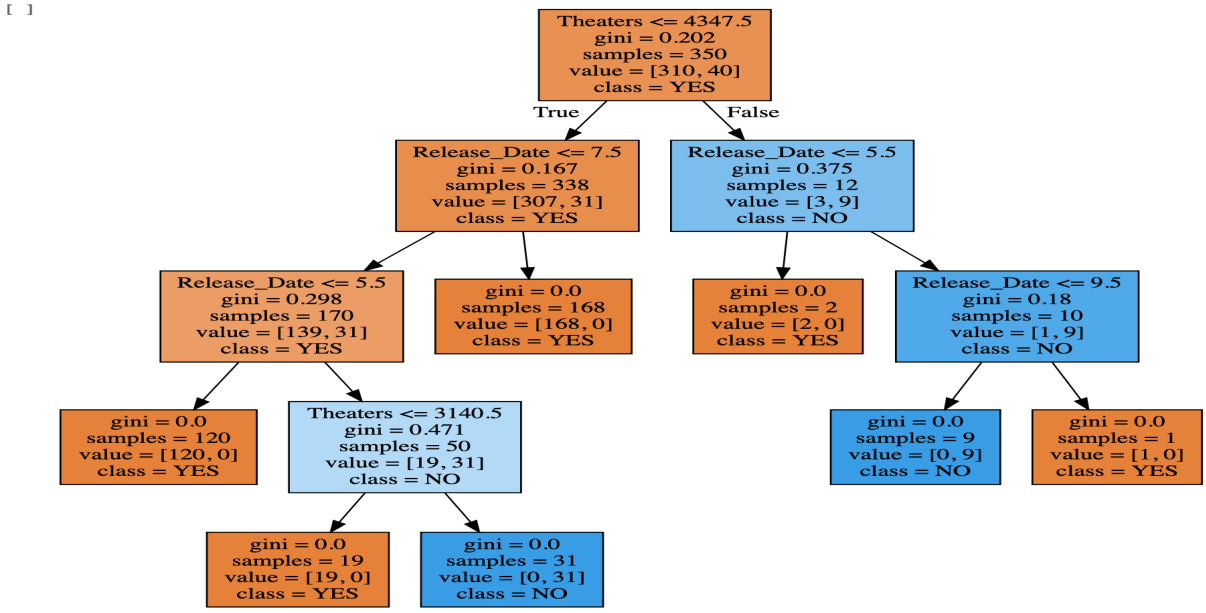
Based on our hypothesis that movies would do better if released in summer months and in a larger amount of theaters (>3000), we trained a decision tree to show if a relatively successful movie (gross >$50 million) fit the criteria of our hypothesis, The decision tree below shows that successful movies do, indeed, release in summer months (June/July) and in greater than 3140 theaters. To begin training the decision tree, we had to process the dataset first by removing all unnecessary columns. Then we isolated the boolean hypothesis columns that the model would be trained on.


Release_Date <= 7.5
entropy = 0.513
samples = 350
value = [310, 40]
class = YES
True
False
Release_Date <= 5.5
entropy = 0.762
samples = 181
value = [141, 40]
class = YES
entropy = 0.0
samples = 169
value = [169, 0]
class = YES
entropy = 0.0
samples = 122
value = [122, 0]
class = YES
Theaters <= 3140.5
entropy = 0.907
samples = 59
value = [19, 40]
class = NO
entropy = 0.0
samples = 19
value = [19, 0]
class = YES
entropy = 0.0
samples = 40
value = [0, 40]
class = NO

```
test_df.head(10)
```

	Release_Date	Gross_Clean	Theaters	class	pre	error	NO	YES	p_score
0	4.0	60311495.0	3418.0	NO	NO	0	1.0	0.0	1.0
1	7.0	110212700.0	3171.0	NO	YES	1	0.0	1.0	1.0
2	6.0	22105643.0	3162.0	NO	YES	1	0.0	1.0	1.0
3	2.0	175750384.0	4088.0	NO	NO	0	1.0	0.0	1.0
4	4.0	24801212.0	2215.0	NO	NO	0	1.0	0.0	1.0
5	3.0	45729221.0	2866.0	NO	NO	0	1.0	0.0	1.0
6	4.0	32149404.0	1203.0	NO	NO	0	1.0	0.0	1.0
7	12.0	37921265.0	2978.0	NO	NO	0	1.0	0.0	1.0
8	6.0	80001807.0	4224.0	YES	YES	0	0.0	1.0	1.0
9	3.0	62524260.0	3492.0	NO	NO	0	1.0	0.0	1.0

We trained two trees, one using entropy (above) and one using gini (below). We prefer the first tree because we believe it has more accurate and concise results that are easier to convey to people who may not be experts in data mining. The entropy score was also .9733 while the gini score was only .9666, so the higher score would indicate better results.



	Release_Date	Gross_Clean	Theaters	class	pre	error	NO	YES	p_score
0	4.0	60311495.0	3418.0	NO	NO	0	1.0	0.0	1.0
1	7.0	110212700.0	3171.0	NO	YES	1	0.0	1.0	1.0
2	6.0	22105643.0	3162.0	NO	YES	1	0.0	1.0	1.0
3	2.0	175750384.0	4088.0	NO	NO	0	1.0	0.0	1.0
4	4.0	24801212.0	2215.0	NO	NO	0	1.0	0.0	1.0
5	3.0	45729221.0	2866.0	NO	NO	0	1.0	0.0	1.0
6	4.0	32149404.0	1203.0	NO	NO	0	1.0	0.0	1.0
7	12.0	37921265.0	2978.0	NO	NO	0	1.0	0.0	1.0
8	6.0	80001807.0	4224.0	YES	YES	0	0.0	1.0	1.0
9	3.0	62524260.0	3492.0	NO	NO	0	1.0	0.0	1.0

In addition to a decision tree, we trained our hypothesis on a Naive Bayes model. We chose this model because it is good for determining the probability that a hypothesis is correct. The model yielded positive results with a 99% confidence level that our hypothesis was correct. For this model, the data was converted from integers and floats to categorical data (figure to the left) and used the BernoulliNB functions from the sklearn.naive_bayes library in python. To begin the process, we created an X and Y dataframe, X being a boolean table of which attributes a row has and Y being a boolean column of rows where the hypothesis is true. We then calculated the probability of the hypothesis being correct using the built in library functions.

Gross	Theaters	Total_Gross	N
>150	>4000	>160	
>150	>4000	>160	
>150	>4000	>160	
>150	>4000	>160	
>150	>4000	>160	
...	...	...	
<30	3200-3599	<35	
<30	<2800	<35	
<30	<2800	35-55	
<30	2800-3199	<35	
<30	2800-3199	<35	

```
[ ]  NB_B = BernoulliNB()
```

```
[ ]  scores = cross_val_score(NB_B, Computer['data'], Computer['target'], cv=5, scoring='accuracy')
     scores

     array([1.  , 1.  , 0.99, 0.99, 0.98])
```

```
[ ]  #mean and 99% confidence level
     print("Accuracy: %0.2f (+/- %0.2f)" % (scores.mean(), scores.std() * 2))

     Accuracy: 0.99 (+/- 0.01)
```

```
[ ]  NB_M = MultinomialNB()
     scores = cross_val_score(NB_M, Computer['data'],Computer['target'], cv=5, scoring='accuracy')
     scores

     array([1.  , 1.  , 0.99, 0.99, 0.98])
```

```
[ ]  #mean and 99% confidence level
     print("Accuracy: %0.2f (+/- %0.2f)" % (scores.mean(), scores.std() * 2))

     Accuracy: 0.99 (+/- 0.01)
```

We can say with some certainty that as time goes on, the highest grossing films released in a given year will earn more revenue than the highest grossing films in the previous few years. We have seen this result play out in most of the years that we analyzed and see a clear and obvious trendline. This would imply that movie studio executives can be confident that if the movie they plan to release is of similar quality to previous years, they can reasonably expect that the new release will perform at least slightly better than the last release.

Some issues that we ran into during our analysis was with drawing our diagrams accurately and in the way that we wanted. Sometimes we wanted to create a chart with two attributes and had a problem with the intervals on the axes. Unlike in a spreadsheet where we can more easily change the intervals, on Google Colab, we had a hard time adjusting the scale to the way we wanted. In addition, we could not name our axis with the proper title such as “Total Gross” if that is what we were graphing. Eventually we were able to get our axis and chart title named properly on most of our graphs. Another issue we had was with charting the median gross revenue per month on a graph of the number of movies released per month in a particular year.

When graphed separately, we had no issue, but when we tried to overlay it, we found that the median gross per month line had shifted by one month to the right. This resulted in the chart starting in February instead of January.