Team-9 Final Project Report

IS115 | Katy Reese | Winter 2021

Team Members: Taylor Sabin, Danie Allen, Tyler Curtis, & Jeff Toronto

Python Analytics using Data from IMDB and Box Office Mojo

Table of Contents

Introduction: 3
- Finding a Dataset and the Benefits of Good Data 3
- Troubles with using our Dataset 4
- Data Cleaning 4
- About our Dataset 5
- How we organized this report 6

Solving for Skewness and Kurtosis 7
- The problem 7
- The solution 7

Univariate 8
- Is there a difference between audience and critics' scores? 8
- What is the average domestic gross? Is it normally distributed? 10

Bivariate: Num/Num 12
- Is there a statistically significant relationship between critics' score and box office performance? 12
- Exploring visualizations 13

Bivariate: Cat/Num 15
- Is there a relationship between distributor and domestic gross? 15
- In which month(s) should distributors release movies, and which should they avoid? 15
- Other factors to consider 18

Bivariate: Cat/Cat 18
- Is there a relationship between the month/time of year when a movie is released and its success ranking for that year? 19
- Do certain distributors have more Top 10 movies than others? 21
- Have the Top 10 movie distributors of the last 10 years increased their movie release output? 22

Multivariate: Numeric and Categorical 25
- What factors determine annual domestic gross ranking? 25
- What patterns are there between release date, domestic gross, and rating? 27
- What patterns are there between ratings, top ten ranking status, and quarter of release? 28

Conclusion 29

Introduction:

Finding a Dataset and the Benefits of Good Data

As a team, we originally began looking for datasets that compared Audience Scores and Critics' Scores on Rotten Tomatoes. We found a dataset on Kaggle.com called "Rotten Tomatoes movies and critics reviews dataset," which seemed to be exactly what we were looking for. Upon further examination, however, we identified several issues that we would run into if we decided to use it. (1) The movies in question were given labels based on their movie link id on RottenTomatoes.com, which wasn't difficult to work with, but would require the extra step of looking up which movie was which after we would have filtered the data. (2) There were other labels that seemed unnecessary or unrelated to our desired topic of research, such as "critic_name" or "review_content," the latter being a snippet quote from an individual critic's review of the movie in question. There was another field that represented the audience score, but only had two criteria: "Fresh" and "Rotten," "Fresh" indicating that the movie had received 64% or higher, and "Rotten" indicating that the movie received a score of 36% or lower, thus limiting our use of audience scores to those two values. Nevertheless, it was still usable and there were labels that would be useful in answering certain questions, and we could still work with those. (3) The dataset was *enormous*-- 232 MB worth! Though in one sense, this was very good, in that we would be able to have more than enough data to work with and validate our research questions. On the other hand, we predicted that a dataset this size would take a fairly long time to sort through. (4) There was a specific field ("review_score") that had a combination of numeric and categorical values which were input with different scoring ranges and types. For example, some critics' scores were input as fractions out of inconsistent denominators, sometimes five, ten, four, etc. (e.g. "2.5/5" or "8.5/10"), which, though possible to solve by using some input scrubbing and ranges, would be a significant amount of work. Unfortunately, as mentioned previously, there were also categorical entries using letter-grades (e.g. "A-", "B," "C+"). Though not impossible to work with these issues, we felt it best to find a cleaner, smaller, and more evenly categorized dataset.

In our search for a better dataset, we found one on Kaggle entitled "Movie Ratings 2010-2019: average Joe vs critics." Though not associated with data from rotten tomatoes, this dataset was a combination of smaller datasets from both BoxOfficeMojo.com and IMDB.com, and provided much clearer and

cleaner values. This dataset was significantly smaller (only 4 MB) and covered 100 movies from each year between the decades 2010 to 2019. We will get into the specific details of this dataset in a later section.

Troubles with using our Dataset

Having found a good set of data to work with, we attempted to run basic queries using https://colab.research.google.com/ to look at the data (ex. using the command df.head()), but came across some difficulty in running the queries. After some closer inspection, we found that the .csv files containing the data were using semicolons as separators rather than the standard comma. We solved this problem by adding this code to filter the data frame: *sep=";"*, which solved our issue (see fig.1).

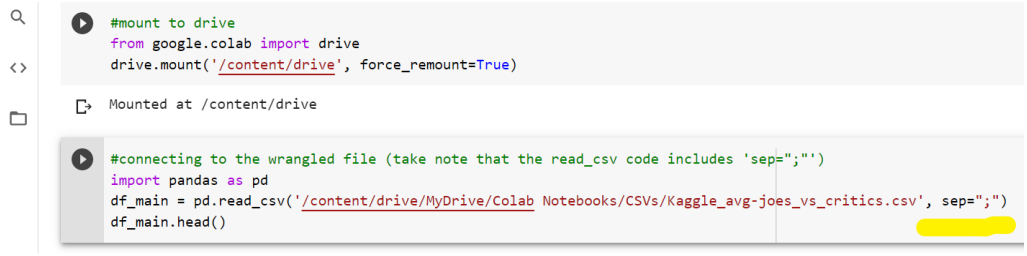


Fig.1

Data Cleaning

Our dataset that we selected for this project was very clean, but after some reviewing, there did happen to be one significant issue. Some of the data that was collected from Metacritic uses a comma as a decimal separator, which is a common formatting in various countries around the world (see fig. 2). Yet in this case, this comma formatting renders the data unusable because python does not interpret float data types with commas as a decimal separator. Thus, performing calculations and analyses of the data in this form simply is not possible.

Metacritic_critics_mean_score	Metacritic_critics_median_score	Metacritic_critics_weighted_score	Distance_critics_user
83,85714286	88	83	5,857142857
90,33333333	90	92	8,333333333
58,63157895	60	53	7,368421053
60,075	61,5	57	10,925
61,18421053	61,5	58	7,184210526

fig.2

In order to clean the data so that we could use it in our analyses, we performed a replace method on the columns that contained this formatting, by simply replacing the comma with a period for values that contained a comma (see fig. 3). This process formatted the data into a usable data type that can be more easily interpreted and analyzed in python, thus allowing us to proceed with the analysis portion of this project.

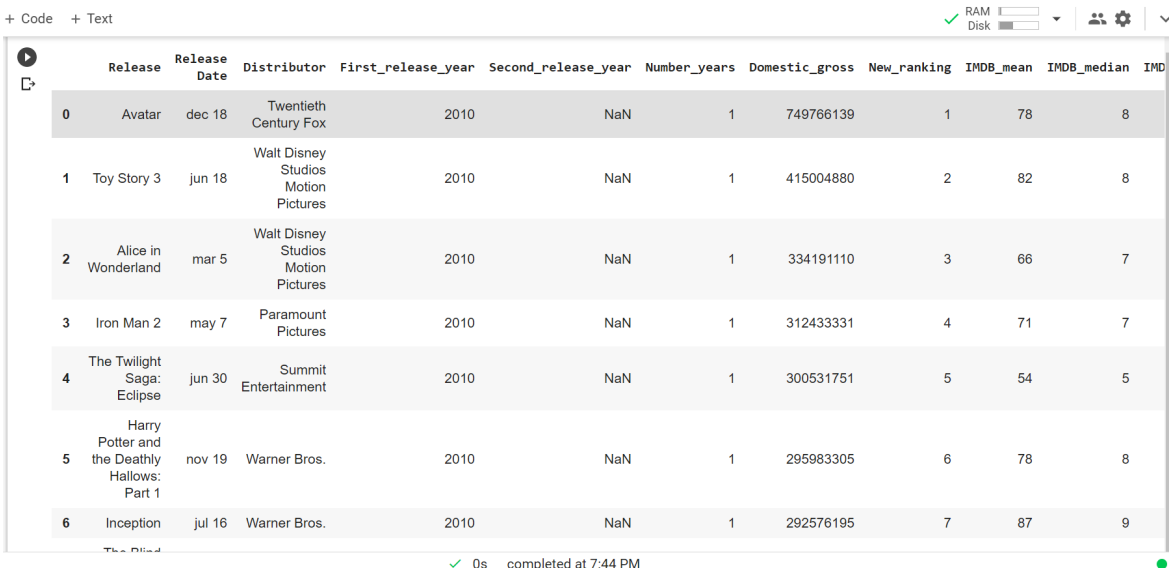
```
df_main['Metacritic_critics_mean_score'] = (df_main['Metacritic_critics_mean_score'].replace(',','.', regex=True).astype(float))
df_main['Distance_critics_user'] = (df_main['Distance_critics_user'].replace(',','.', regex=True).astype(float))
df_main['Metacritic_critics_mean_score_US'] = (df_main['Metacritic_critics_mean_score_US'].replace(',','.', regex=True).astype(float))
df_main['Distance_critics_user_US'] = (df_main['Distance_critics_user_US'].replace(',','.', regex=True).astype(float))
df_main['Metacritic_critics_median_score'] = (df_main['Metacritic_critics_median_score'].replace(',','.', regex=True).astype(float))
```

fig.3

About our Dataset

As mentioned previously, this dataset had a much more manageable size and organization. It focuses on the top 100 highest grossing movies (based on the domestic box office records from Box Office Mojo) for each year, from the years 2010 to 2019. The data were organized by year, starting from 2010. The entries were further organized by ranking each movie of that year by its gross box office, beginning with the highest grossing movie of that year to the lowest (see fig.4). Also from Box Office Mojo, you will see that the dataset includes the movie's distributor, release month, and whether or not the movie was out for a second year.

In addition to this information from Box Office Mojo, the dataset also includes a substantial amount of critical and demographic information from IMDB.com, such as “IMDB_mean” (which represents the audience score on IMDB.com), “Metacritic_critics_weighted_score” (the metacritic score you see on IMDB.com), and a variety of related measurements, including the IMDB metacritic score for the US and the varying metacritic scores among males and females. It’s clear to see that this dataset has the potential of answering a large variety of questions.



+ Code + Text

RAM Disk

	Release	Release Date	Distributor	First_release_year	Second_release_year	Number_years	Domestic_gross	New_ranking	IMDB_mean	IMDB_median	IMD
0	Avatar	dec 18	Twentieth Century Fox	2010	NaN	1	749766139	1	78	8	
1	Toy Story 3	jun 18	Walt Disney Studios Motion Pictures	2010	NaN	1	415004880	2	82	8	
2	Alice in Wonderland	mar 5	Walt Disney Studios Motion Pictures	2010	NaN	1	334191110	3	66	7	
3	Iron Man 2	may 7	Paramount Pictures	2010	NaN	1	312433331	4	71	7	
4	The Twilight Saga: Eclipse	jun 30	Summit Entertainment	2010	NaN	1	300531751	5	54	5	
5	Harry Potter and the Deathly Hallows: Part 1	nov 19	Warner Bros.	2010	NaN	1	295983305	6	78	8	
6	Inception	jul 16	Warner Bros.	2010	NaN	1	292576195	7	87	9	

0s completed at 7:44 PM

fig.4

How we organized this report

This report is organized by the various questions we sought to answer using this dataset, separating these questions based on the type of statistical information we were analyzing. We begin by addressing an issue that we needed to research a bit on our own (something that was outside the scope of the class): solving for skewness. The next section discusses different questions we answered using univariate statistics. The third focuses on questions we answered using tests involving bivariate numeric/numeric data. The fourth does the same, only using categorical/numeric data. The last two sections address questions and tests involving bivariate categorical/categorical data and multivariate numeric and categorical data, respectively. In each section we will include our findings and conclusions as to what the data tells us concerning each question.

Solving for Skewness and Kurtosis

The problem

Towards the beginning of our data analysis, we created some univariate visualizations (addressed in the following section) to check to see if our data was normally distributed. As you will see in the following section, most of the data we would be working with fell within the “normal” range, except for one key column: “Domestic_gross.” The data for Domestic gross show a significant positive skewness (a skew value of 3.01) and a very high kurtosis (a kurt value of 12.53). What did this mean for our further analysis? For the majority of the questions we explored, this meant very little, because we were simply tracking the record Domestic gross rather than checking for correlation and statistically significant relationships. However, there were a few questions where we did seek to test for said relationships (we will make mention of these moments when appropriate) and would need to normalize this data to reliably use and analyze the results.

The solution

After researching this topic in our textbook (ch. 22 was skipped over this semester, hence, this being our “above and beyond the scope of this class” portion of the project), we found a way to normalize “Domestic_gross” using the natural log of each entry (also shown and discussed in the following section), we discovered that normalizing these data would make it difficult to interpret in terms of record count for many of our questions, thus over-complicating our research. You will see in figure 9 that domestic gross, previously measured in “hundred-millions,” had narrowed down to fit between a range of 17 to 21. We decided that the best way to solve this problem was to simply use our code to normalize “Domestic_gross,” but to then include the normalized data into a new column *in addition* to the original “Domestic_gross” column, calling the new column “Normalized_Domestic_Gross,” which we could reference for appropriate tests (without changing the readability of the “Domestic_gross” entries. In the next section, we show two histograms for “Domestic_gross”: (1) one with the original values (figure 9), and (2) one after the data have been normalized (figure 10). Below, we have included the code that we used to normalize “Domestic_gross” and create a new column in the DataFrame (fig. 5).

```
from matplotlib import pyplot as plt
import numpy as np
#code to normalize domestic gross
normalized_gross = np.log(df_main.Domestic_gross)
df_main.insert(7, 'Normalized_Domestic_Gross', np.log(df_main.Domestic_gross))
#df_main.drop(['Normalized_Domestic_Gross'], axis=1, inplace=True) #<--- This is just to run the cell again
df_main.head()
```

	Release	Release Date	Distributor	First_release_year	Second_release_year	Number_years	Domestic_gross	Normalized_Domestic_Gross	New_ranking	IMD
0	Avatar	dec 18	Twentieth Century Fox	2010	NaN	1	749766139	20.435272	1	
1	Toy Story 3	jun 18	Walt Disney Studios Motion Pictures	2010	NaN	1	415004880	19.843801	2	
2	Alice in Wonderland	mar 5	Walt Disney Studios Motion Pictures	2010	NaN	1	334191110	19.627224	3	
			Paramount							

✓ 0s completed at 2:58 PM

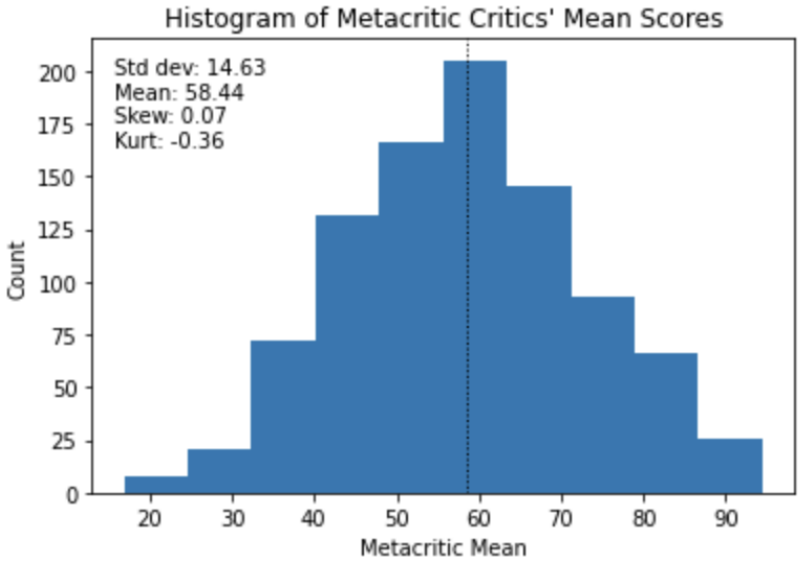
fig.5

Univariate

Is there a difference between audience and critics' scores?

For the univariate portion of our analysis, we chose to explore a few of the data points that we deemed would be most significant for the majority of this project, and those that tell the greatest "story" of our dataset. We first explored the distribution of scores for both the IMDB Audience and the Metacritic Critics' scores, as well as their respective mean scores. It is important to note that our dataset includes additional data types for weighted mean scores for both critics and audience, but due to their abnormal distribution and ambiguity in how these weighted averages were calculated, we chose to focus our research on the un-weighted average data types. The two histograms that we generated for IMDB and Metacritic average scores appear below, which include statistics integrated into their plots to determine their normality in addition.

fig.6



Through these two histograms, we found that both are distributed normally because of their skew and kurtosis values falling within the normal range of 1 and -1, meaning that they can be used in further analysis. Additionally, we can see that there are notable differences in mean scores and standard deviations between the audience and critics' scores. These findings led us to question: what is the average difference between the audience and critics' scores?

With these two plots in mind, there is an additional data type from our data set that records the average difference between audience and critics' scores. The following histogram takes these differences between the critics and audience and finds the mean difference, as well as the appropriate statistics to determine normality:

fig . 7

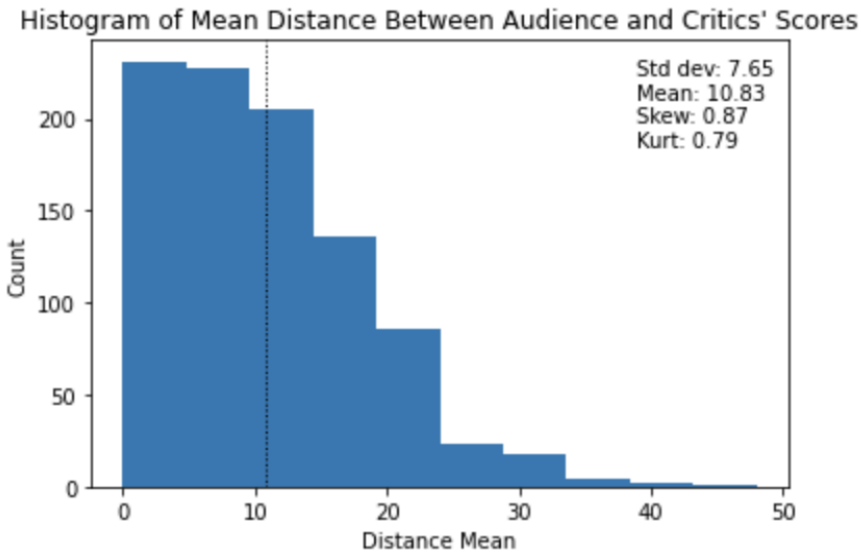
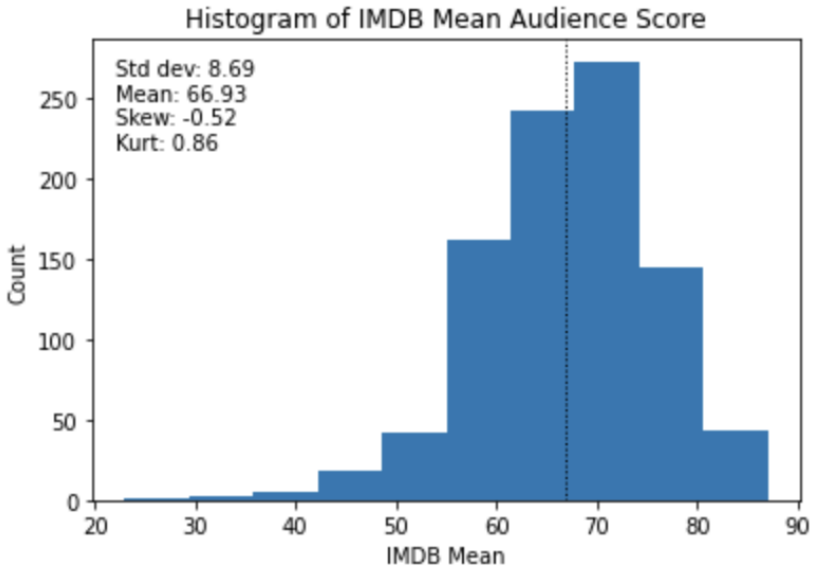


Fig. 8

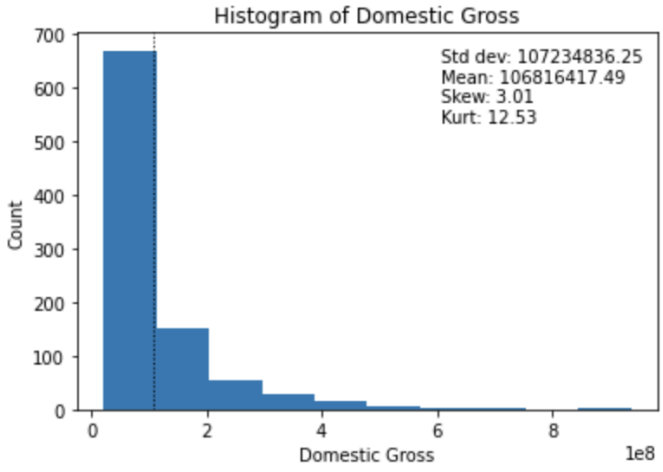


From this we can see that the average difference between the audience and critics' scores is greater than 10 points. This appears to be an incredibly large discrepancy and given that the rating scale is out of 100, we can see that on average, that this difference is greater than 10 percent. The positive mean value signals that the difference is indeed positive, and that the audience score is greater than the critics' score on average. Thus we can conclude that on average, the audience will score any given movie a score greater than 10% larger than that of the critics.

What is the average domestic gross? Is it normally distributed?

The other question we sought to answer is what is the average domestic gross and is it normally distributed? Similar to the previous univariate tests, we created a histogram of domestic office and these were the results:

fig.9

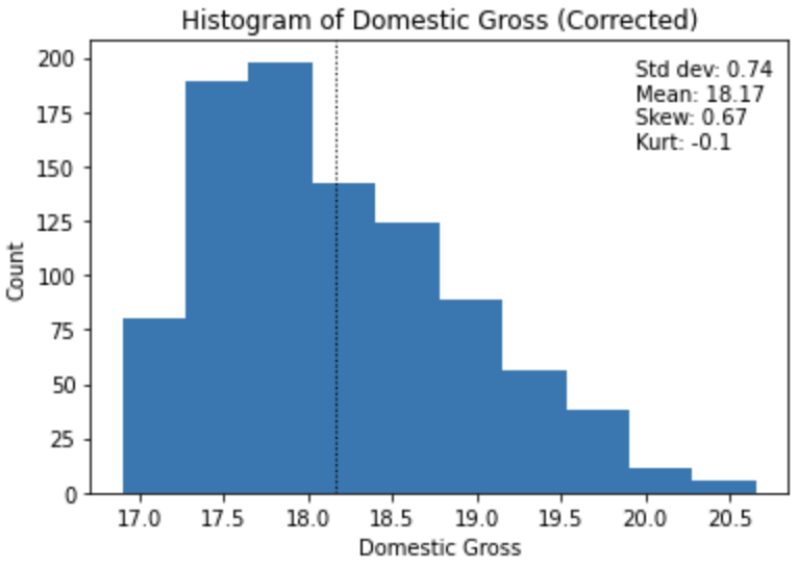


From this plot, we can see that domestic gross is not distributed normally because both its skew and kurtosis values are outside of the normal range of -1 to 1. Because of this, we determined that we must adjust the data to normalize both the skew and kurtosis in order to be able to use it in further analysis. After some testing, we found that using the natural log method produced the most normal normal data and the following line of code was used to do so:

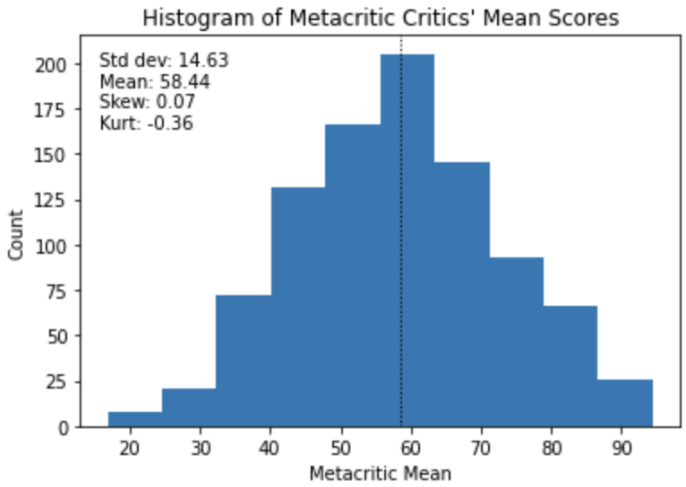
```
Domestic_grossNormalized = np.log(df_main.Domestic_gross) # Correcting for high skew and kurt values
```

With skew and kurtosis now accounted for the, following is the corrected histogram of domestic gross:

fig.10



Now with the domestic gross data corrected for normality, we can see that it is no longer right-skewed as heavily before and its new skew and kurtosis fall within normality. With domestic gross now accounted for, we were able to proceed with further exploration.



Bivariate: Num/Num

Is there a statistically significant relationship between critics' score and box office performance?

For this question, we wanted to explore whether or not the relationship between a movie's average critics' score (the Metacritic score in a colored box on IMDB.com, represented in this dataset under the column "Metacritic_critics_weighted_score") and its domestic box office performance. The answer to this question was fairly simple to find, and we spent most of the time exploring different ways to represent this relationship. Below, we have included our simple analysis of this question, and our thought process and comments with each visualization.

One important preliminary step, as we have referenced in our section about solving for skewness and kurtosis, was to use the normalized data from the new column we created ("Normalized_Domestic_Gross") with the "Metacritic_critics_weighted_score" to run a correlation and here is what we found:

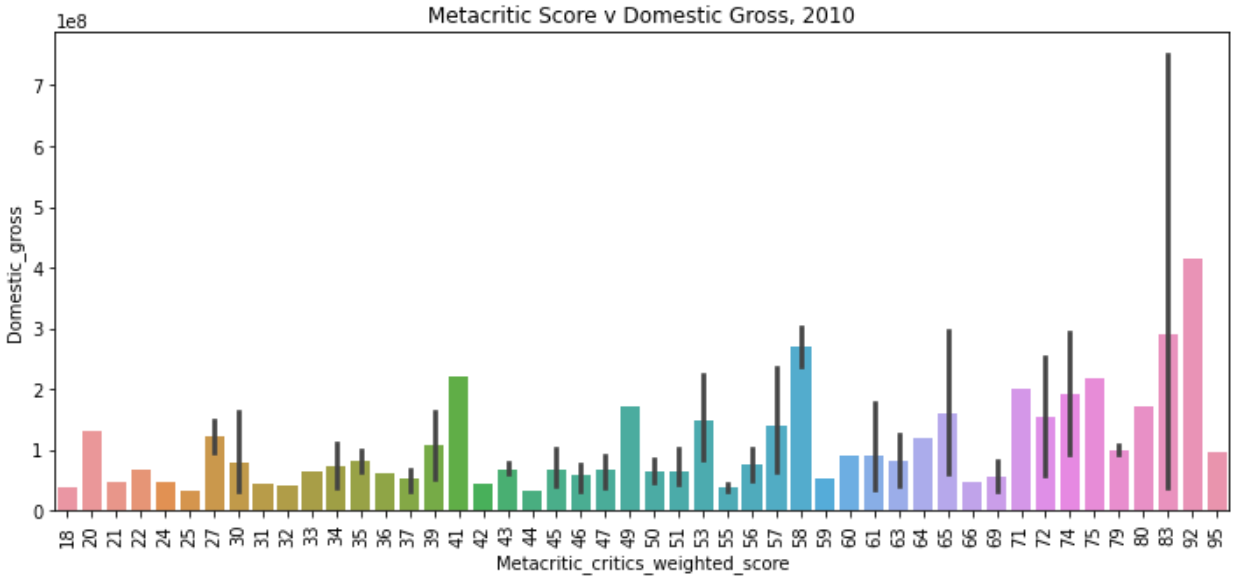
```
from scipy import stats
import pandas as pd

corr = stats.pearsonr(df_Tyler.Metacritic_critics_weighted_score, df_Tyler.Normalized_Domestic_Gross)
corr
print('r: \t' + str(round(corr[0], 4)))
print('p-value:' + str(round(corr[1], 4)))
```

```
r:      0.3216
p-value:0.0
```

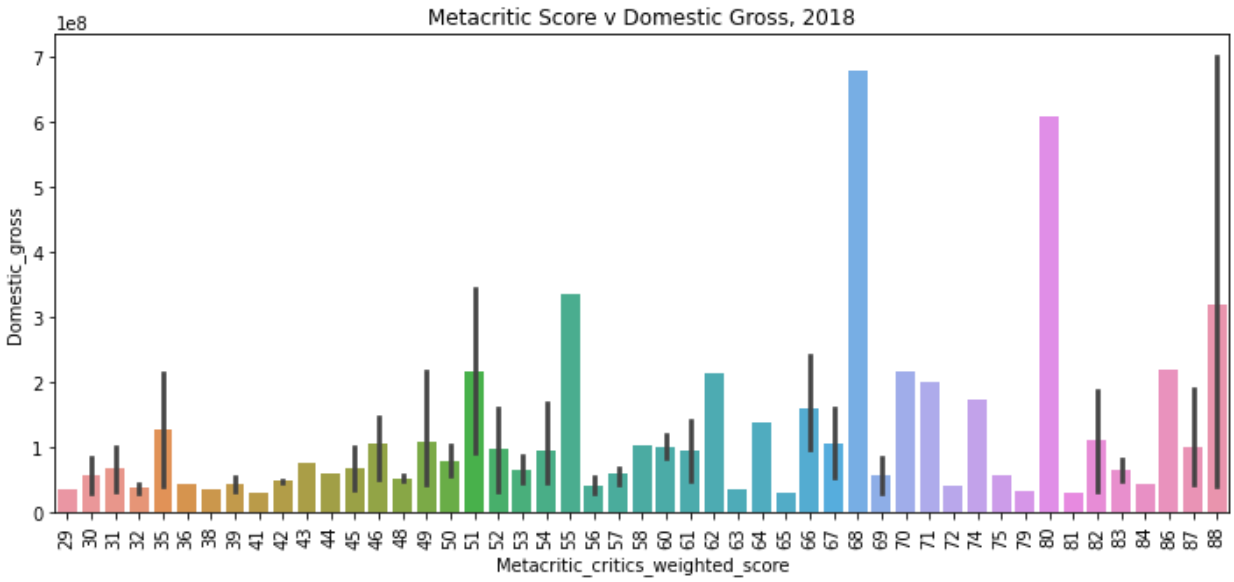
As we see from the p-value, the relationship between the two *is* statistically significant, with a value much less than 0.5, and we can see that there is a weak-to-moderate positive correlation between the two. We see from the r-value that there is a 32% chance that a unit increase to one will result in a unit increase in the other. This information is useful in that, though there might be a more positively correlated relationship for both critics' score and domestic gross, we *can* say as a general rule, that striving to increase your performance among critics will likely aid your performance in the box office.

Exploring visualizations

Here are two visualizations of the aforementioned relationship. Both are barplots, and compare critics' score to domestic box office performance. The first barplot is of the releases during the year 2010, and the second is for the year 2018.



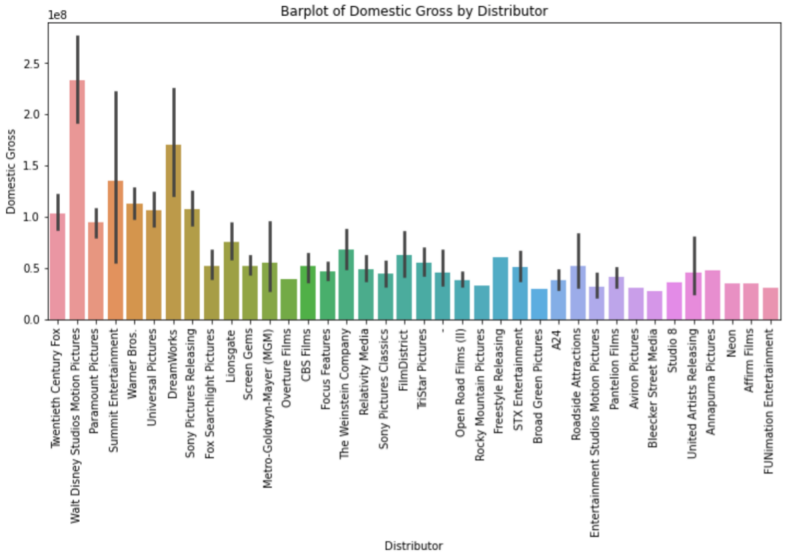
From this first barplot, right away your eyes are drawn to the gray bar over the IMDB score "83." Looking at the preview of the dataset, you can tell right away that this large spike is a representation of the movie Avatar, which was pretty cool to recognize. What we can say about this barplot, is that, though there are exceptions in the lower scores, there's an obvious trend showing that the higher the score, the higher the domestic box office, which supports our correlation test.



Unlike the previous barplot, this shows something a little different. First off, let's address the 3 highest box office scores: "68", "80", and "88", which represent Avengers: Infinity War, The Incredibles: 2, and Black Panther, respectively -- which are the top 3 domestic box office scores. The first takeaway is that it's very cool that we can use python and data tools to visualize data in a way that we can relate to it a little better. The second takeaway from this visualization is that unlike the correlation and first barplot, this bar plot shows a slightly different trend. The most successful movies (with exception to the unique case of *Black Panther*) are found between 55 and 80. Now, since this is one year, and not the whole decade of data being represented, we could say that this is an anomaly of 2018. A third and final takeaway, is that overall domestic gross increased between these two years. This could be due to an increase in ticket sales due to inflation over time.

Concerning further development of this question, the next step would be to take the decade as a whole and bin up the Metacritic scores into groups of 10: 0-9, 10-19, 20-29, and so on, to make the barplot look cleaner. With this larger perspective, we might have a better understanding of where the right balance between critics' score and domestic gross lies.

Bivariate: Cat/Num



ANOVA
F: 7.14
p-value: 0.0

Sig. comparisons (Bonferroni-corrected)
Twentieth Century Fox-Walt Disney Studios Motion Pictures: t=-6.1778, p=0.0
Twentieth Century Fox-Screen Gems: t=4.2213, p=0.0
Walt Disney Studios Motion Pictures-Paramount Pictures: t=6.0613, p=0.0
Walt Disney Studios Motion Pictures-Warner Bros.: t=5.9388, p=0.0
Walt Disney Studios Motion Pictures-Universal Pictures: t=6.1826, p=0.0
Walt Disney Studios Motion Pictures-Sony Pictures Releasing: t=5.1572, p=0.0
Walt Disney Studios Motion Pictures-Fox Searchlight Pictures: t=4.527, p=0.0
Walt Disney Studios Motion Pictures-Lionsgate: t=8.3899, p=0.0
Walt Disney Studios Motion Pictures-Screen Gems: t=6.6321, p=0.0
Walt Disney Studios Motion Pictures-Focus Features: t=5.8106, p=0.0
Walt Disney Studios Motion Pictures-The Weinstein Company: t=4.4381, p=0.0
Walt Disney Studios Motion Pictures-Relativity Media: t=4.9791, p=0.0
Walt Disney Studios Motion Pictures-TriStar Pictures: t=4.7733, p=0.0
Walt Disney Studios Motion Pictures-Open Road Films (II): t=5.2635, p=0.0
Walt Disney Studios Motion Pictures-STX Entertainment: t=5.4161, p=0.0
DreamWorks-Lionsgate: t=4.3497, p=0.0
DreamWorks-Screen Gems: t=6.4785, p=0.0
DreamWorks-Focus Features: t=6.4373, p=0.0
DreamWorks-Relativity Media: t=5.3404, p=0.0
DreamWorks-TriStar Pictures: t=5.036, p=0.0
DreamWorks-Open Road Films (II): t=6.7582, p=0.0
DreamWorks-STX Entertainment: t=5.3118, p=0.0
Sony Pictures Releasing-Lionsgate: t=4.2691, p=0.0
Sony Pictures Releasing-Screen Gems: t=4.6435, p=0.0
Sony Pictures Releasing-Focus Features: t=4.4743, p=0.0
Sony Pictures Releasing-Open Road Films (II): t=4.4639, p=0.0

Is there a relationship between distributor and domestic gross?

This barplot was generated in order to compare domestic office (which has been normalized) to the distributors. From these results, we can see that there is a statistically significant difference between each distributor and domestic office, because the p-value is lower than the 0.05 threshold. In addition, we can see from the barplot that most of the distributors have relatively similar domestic gross values, yet a few distributors, such as Disney and Dreamworks, have domestic gross values that are substantially greater than the others.

In which month(s) should distributors release movies, and which should they avoid?

We approached this question thinking of executives of a film distribution company, such as Twentieth Century Fox, and tailored our suggestions (based on our analysis) to real-life decisions that we believe they should consider. To answer this question and make these suggestions, we first wanted to see which release month resulted in the best domestic box office performance over the course of the past

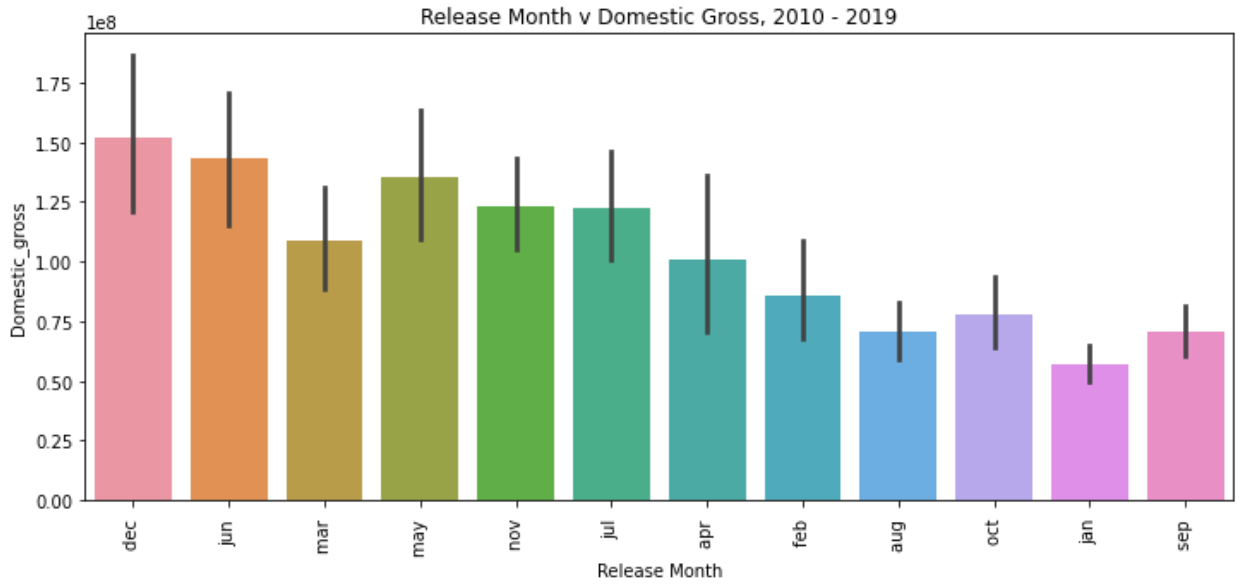
decade (2010-2019). To better understand this, we need to explain the column labeled “Release Date,” and how “Domestic_gross'' ties into it. The column “Release Date” simply refers to which day that particular movie was released to theaters. “Domestic_gross” shows the total domestic gross of that movie for its entire duration of being in theaters. What we wanted to know was which release month resulted in the most successful movies. For example, if we tracked all the movies released during May and all the movies released during September, which group would have the highest overall box office?

To begin, we added a new column called “Release Month,” by filtering through the original DataFrame (df_main) and slicing the first three letters of the column “Release Date” for each entry, as shown here:

```
#create empty list
release_month_list = []

#loop through the release date column and take the first 3 letters
for date in df_Tyler['Release Date']:
  release_month_list.append(date[0:4])
release_month_list

#make a column that is populated by the values of the list; move new column right after "Release Date"
df_Tyler.insert(2, 'Release Month', release_month_list)
#df_Tyler.drop(['Release Month'], axis=1, inplace=True) #<--- This is just to run the cell again
df_Tyler.head()
```

(Note: ‘df_Tyler’ is a copy of ‘df_main,’ used for the sake of avoiding accidentally tampering with the original DataFrame)

From there, we were able to sort a barplot by each unique value of “Release Month,” and used “Domestic_gross” as the y-axis. After running this code, we see the following barplot:



From this barplot, we can clearly see that the top 5 most-successful release months were December, June, May, and November/July. This result is not surprising. December is likely high because children and families tend to have school/work off for the holidays, making it more likely that individuals, and more importantly families will go see a movie released during the end of the year. This logic also holds true for June, when most public schools end their school year, and the Summer begins, again, bringing more families to the theater, rather than individuals. As for May, our speculation is that the school year is ending, and movies released at the end of the month likely receive high family traffic starting with memorial day weekend.

Looking at the lowest 5 months, February, August, October, September and January, we see that they are all times of the year when school starts up again, and movies released at this time will likely not carry over into a major holiday season. We are not saying school schedules are directly correlated with our barplot, rather, we are simply making the case that during periods of the year when school is out, it is *more likely* that a family will be attending the movies together, and more frequently (however, it would be interesting to add this particular data to this dataset, to see how true this is and how closely correlated the two categories would be).

From this analysis, our recommendations to distributors would be as follows: to increase your chances of performing well in the box office, we recommend releasing movies at the beginning of longer holiday periods, especially Christmas break, and the Summer. Based on our analysis of release

months and domestic box office trends, you should focus on releasing your movies during the months of December, the end of November, June, July, and the end of May.

Other factors to consider

There are some factors of this question, and others that we will explore in this report, that prevent us from seeing the full picture. First, as we already mentioned, we don't know if the relationship between release month for movies and release months for school are statistically significant. This will require more data and further analysis. Second, it is very important to note that "Domestic_gross" is tracking the *domestic* gross, not worldwide. Why is this important? While some movies outperform domestically, others often outperform internationally. For example, from our dataset, in 2010, you will see that the domestic gross for the movie *Avatar* was nearly 800 million dollars. This was outperformed a few years later by *Star Wars VII*, which made nearly 1 billion dollars. However, if you look online at the total gross for the movie *Avatar* (still using BoxOfficeMojo.com), it made over 1 billion dollars internationally, and is in fact the highest grossing movie of all time. Perhaps then, there is more to consider. What trends are true for the United States, may not be true for China, as an example. There may be some *other* lucrative factor that is more important to distributors than the month they release their movies. Nevertheless, we do believe that based on our data analysis, and the data we have, that the month a distributor releases their movie does relate to that movie's likely box office performance, and to increase chances of higher *domestic* box office performance, the data would suggest waiting till the start of the Summer and Winter holidays.

Bivariate: Cat/Cat

For our Bivariate analysis of two categorical variables, we found that this analysis was best suited to compare the performance of different movie distributors and display the influence of the release date on a movie's domestic success. We also found that much of the information discovered in this section was very applicable to movie distributors seeking to get an edge in their oversaturated and competitive market. This section also supports our findings in the previous section, and you will find that our recommendations for distributors remain consistent.

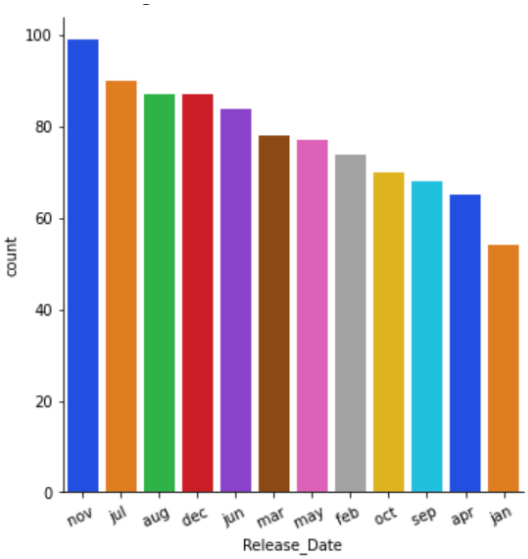
Is there a relationship between the month/time of year when a movie is released and its success ranking for that year?

Of all the bivariate analyses comparing two categorical variables, this first question required the most data cleaning to obtain clear visualizations. Specifically, we replaced individual rankings 1-100 for each year with grouping labels such as “Top 10”, “11-20”, etc. Then we spent a lot of time researching until we found a simple piece of code that would let us remove individual dates from the movies so that all the release dates were organized by month.

```
# Removing individual days from the release date so that I can display by month

df_JeffRank['Release_Date'] = df_JeffRank['Release_Date'].str.replace('\d+', '')
```

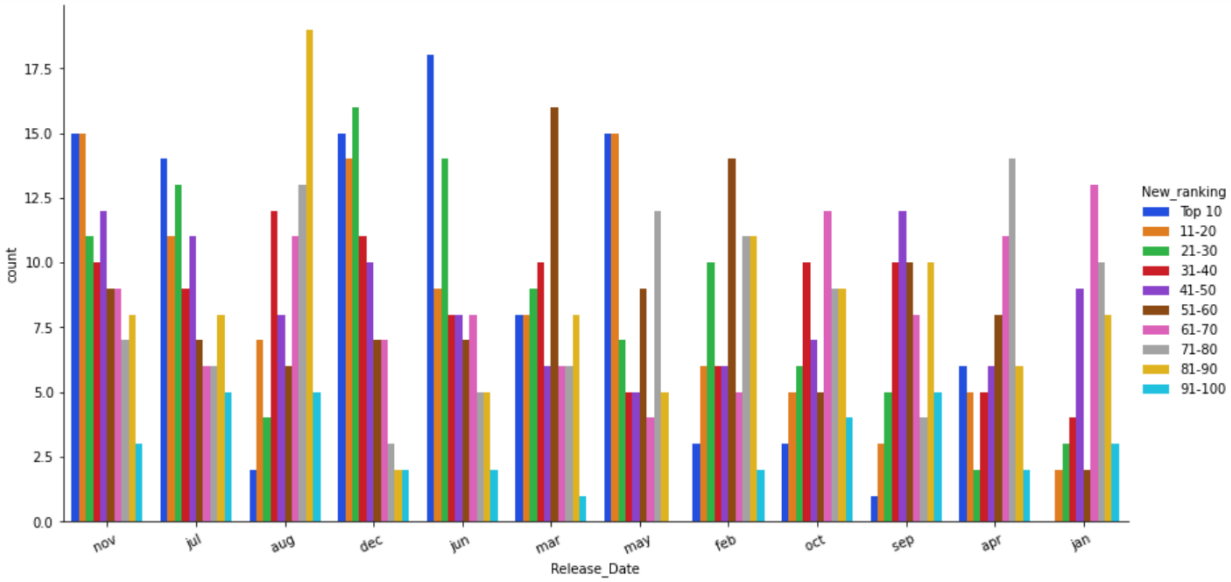
With this housekeeping done, we proceeded with the visualizations. The first visualization shown below is a bar chart that counts the number of Top 100 movies released by month from 2010 to 2019. This chart is useful in showing us that most movies are released during the summer months or winter holidays (all in the top 5). Not surprisingly, the Christmas movie releases are followed by a dearth of releases in January, allowing these holiday hits to gobble up movie goers spending money for the next 31 days.

From a practical standpoint, what if a smaller movie distributor wanted to make a big splash and release a successful movie at a time with less competition from some of the bigshot movie producers? To answer this question, we created a second bar graph that splits up the movies released in each month by groupings related to their rankings. The story this bar graph told was interesting. While, of course, the winter holidays and summer months had plenty of Top 20 films released during those times, May was an interesting anomaly. Coming in at number 7 on the list for release volume, May still had just as many Top 20 movie releases in ten years as November (the highest volume month) did.

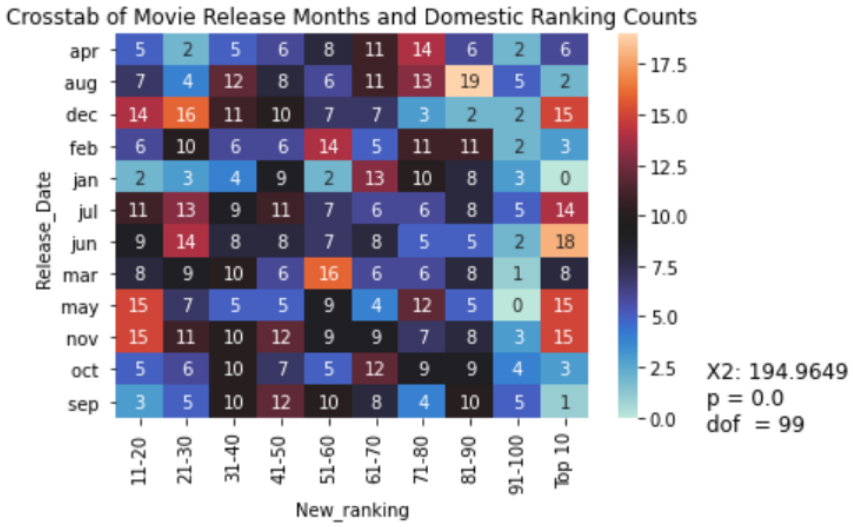


Top 100 movie release count by month in descending order (ABOVE)

Top 100 movie release count by month AND ranking group (BELOW)

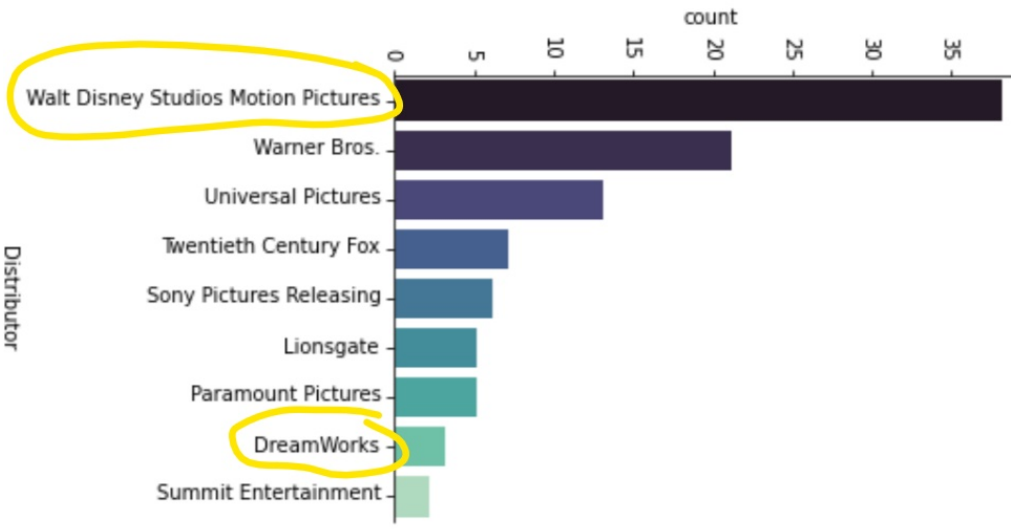


An opportunistic mid-sized distributor would probably find the information in the graphs above intriguing, but can they trust what their eyes are seeing? To validate this information, we took the same data from the second graph and calculated a Chi-square statistic to determine the correlation between release date and domestic ranking. Our analysis showed a Chi-square of ~195 with a p-value of 0. We concluded that we could confidently recommend May as a release date with a strategic advantage for mid-sized distributors trying to make a big splash with a hit movie.



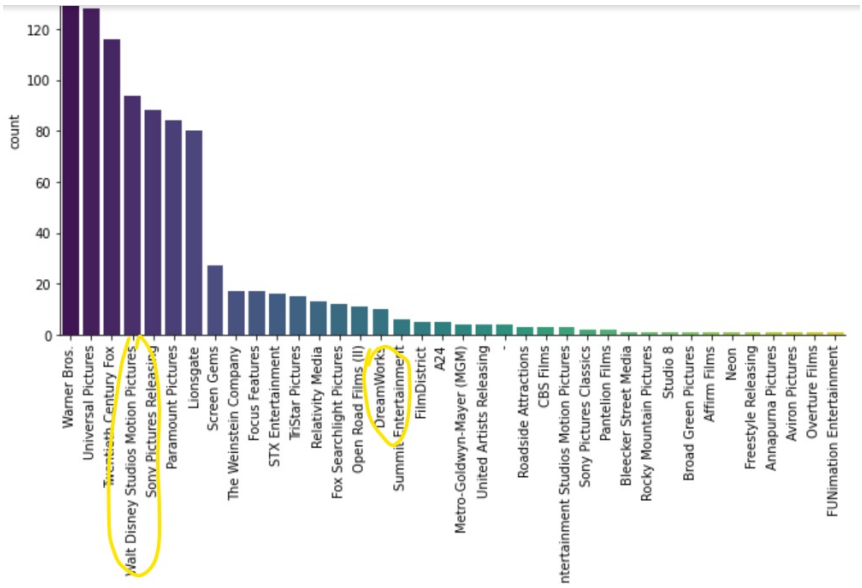
Do certain distributors have more Top 10 movies than others?

Our group's next question was to find out a little more about the Top 10 movie group. Filtering for Top 10 movies, we created a bar graph that would display every distributor who had a Top 10 movie in the 2010-2019 era with the count shown on the y-axis. We were surprised to find that out of 38 distributors in the data set, nine distributors accounted for all 100 of the Top 10 movies. When we compared this data to the total movie counts chart we had created, two distributors stuck out to us.



Top 10 movie count by distributor (ABOVE)

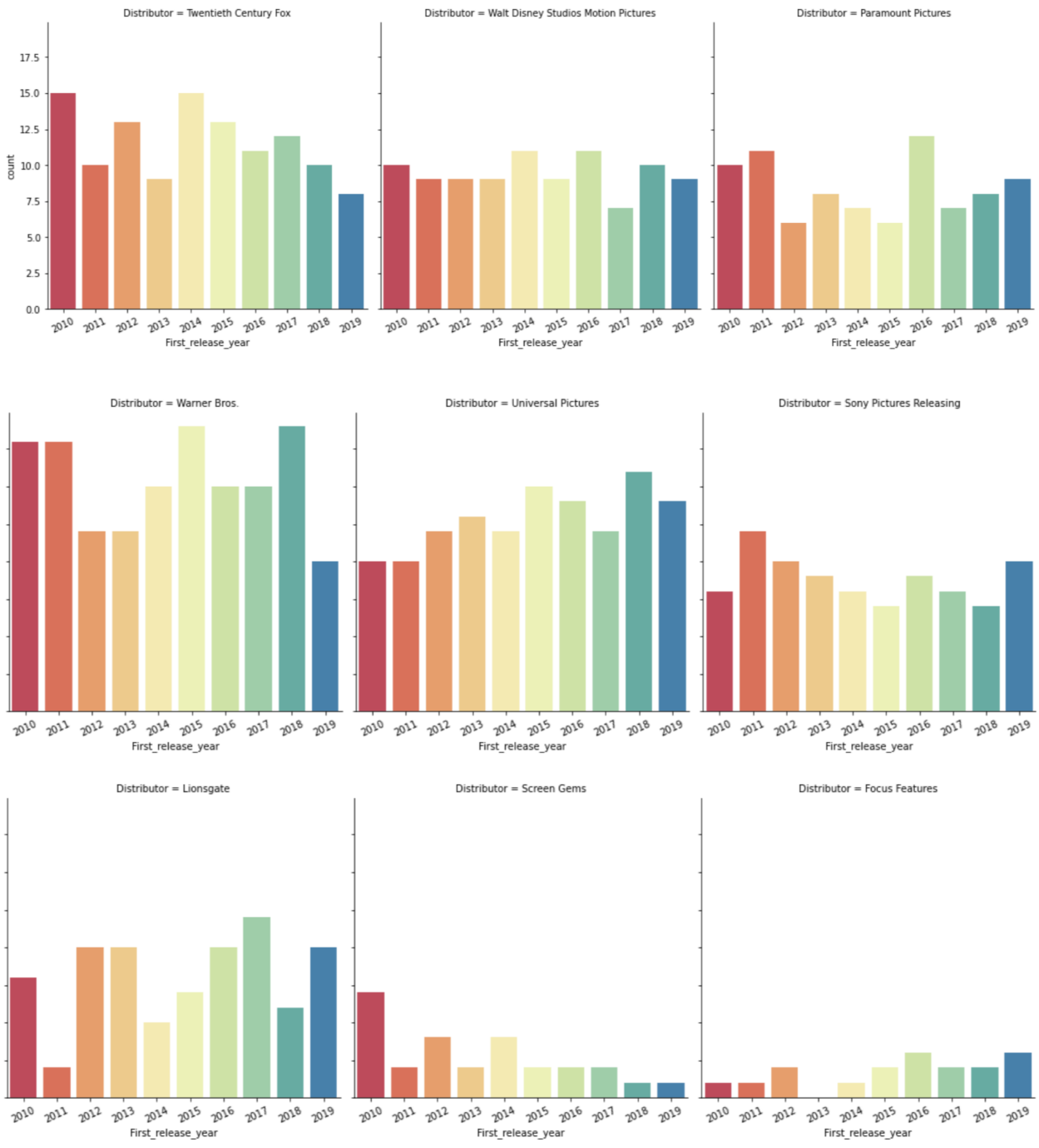
Top 100 movie count by distributor (BELOW)

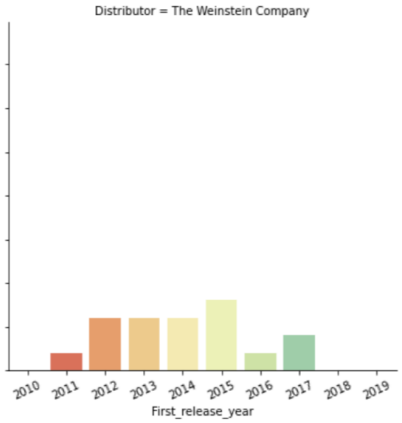


Both Walt Disney Studios and Dreamworks exhibited impressive ratios of Top 10 movies per Top 100 movies (almost 1:2 in both cases). It seems that if a young distributor company wants a model for excellent returns on investment, it should focus on studying the methods of Dreamworks.

Have the Top 10 movie distributors of the last 10 years increased their movie release output?

This final question is meant to explore trends in the output of the Top 10 movie producers by total volume. To our surprise, the visualizations show only a few interesting trends. Disney seemed to remain stable and almost formulaically steady with their release volume, and Universal Pictures definitely had an upward trend in their release volume. These results contradicted our initial hypothesis that release volumes would be increasing consistently across the board as entertainment becomes more important to Americans. However, seeing the lack of apparent growth, we believe that rising prices and online streaming platforms are probably the culprits of the missing market share.


Distributor = Twentieth Century Fox
Distributor = Walt Disney Studios Motion Pictures
Distributor = Paramount Pictures
Distributor = Warner Bros.
Distributor = Universal Pictures
Distributor = Sony Pictures Releasing
Distributor = Lionsgate
Distributor = Screen Gems
Distributor = Focus Features
count
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
First_release_year


Distributor = The Weinstein Company
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
First_release_year

Multivariate: Numeric and Categorical

What factors determine annual domestic gross ranking?

Obviously there are many determinants of how much revenue a movie brings in, such as publicity and actor recognition. In this analysis, numerous critiquing metrics will be held as explanatory variables to determine which of them actually influence domestic gross relative to other movies released that year. The multivariate numeric regression results are as follows:

OLS Regression Results

= OLS Regression Results

Dep. Variable:	New_ranking	R-squared:	0.308
Model:	OLS	Adj. R-squared:	0.296
Method:	Least Squares	F-statistic:	25.46
Date:	Thu, 15 Apr 2021	Prob (F-statistic):	1.57e-62
Time:	03:10:26	Log-Likelihood:	-4227.5
No. Observations:	933	AIC:	8489.
Df Residuals:	916	BIC:	8571.
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
IMDB_mean	-0.1841	0.873	-0.211	0.833	-1.897	1.529
IMDB_median	1.5487	2.445	0.633	0.527	-3.249	6.347
IMDB_weighted_mean	5.8985	1.561	3.779	0.000	2.835	8.962
Metacritic_critics_mean_score	1.1671	0.948	1.231	0.219	-0.693	3.027
Metacritic_critics_median_score	0.0746	0.259	0.288	0.774	-0.434	0.583
Metacritic_critics_weighted_score	-0.6701	0.672	-0.997	0.319	-1.989	0.648
Distance_critics_user	0.1457	0.417	0.349	0.727	-0.672	0.964
IMDB_mean_score_US	-0.1277	0.653	-0.196	0.845	-1.408	1.153
Metacritic_critics_mean_score_US	-0.4633	0.600	-0.772	0.440	-1.642	0.715
Distance_critics_user_US	0.2964	0.443	0.668	0.504	-0.574	1.167
IMDB_mean_score_males	-1.3345	1.005	-1.328	0.185	-3.307	0.638
IMDB_mean_score_females	-3.0720	0.487	-6.307	0.000	-4.028	-2.116
IMDB_mean_score_30-45	-1.8624	1.161	-1.603	0.109	-4.142	0.417
IMDB_mean_score_45+	0.0466	0.461	0.101	0.919	-0.858	0.952
IMDB_general_sample	-6.003e-05	4.55e-06	-13.205	0.000	-6.9e-05	-5.11e-05
New Date	-0.3393	0.229	-1.484	0.138	-0.788	0.109
const	83.1177	10.287	8.080	0.000	62.929	103.307

Omnibus:	28.635	Durbin-Watson:	0.558
Prob(Omnibus):	0.000	Jarque-Bera (JB):	20.824
Skew:	0.261	Prob(JB):	3.01e-05
Kurtosis:	2.487	Cond. No.	3.92e+06

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.92e+06. This might indicate that there are strong multicollinearity or other numerical problems.

Based upon the results of this regression, there are only a few rating metrics that have a statistically significant effect on annual domestic gross rating. At a significance level of $\alpha = 0.05$, the metrics that have a proven impact on annual ranking are IMDB weighted score, IMDB mean score of females, and the sample size of IMDB voters.

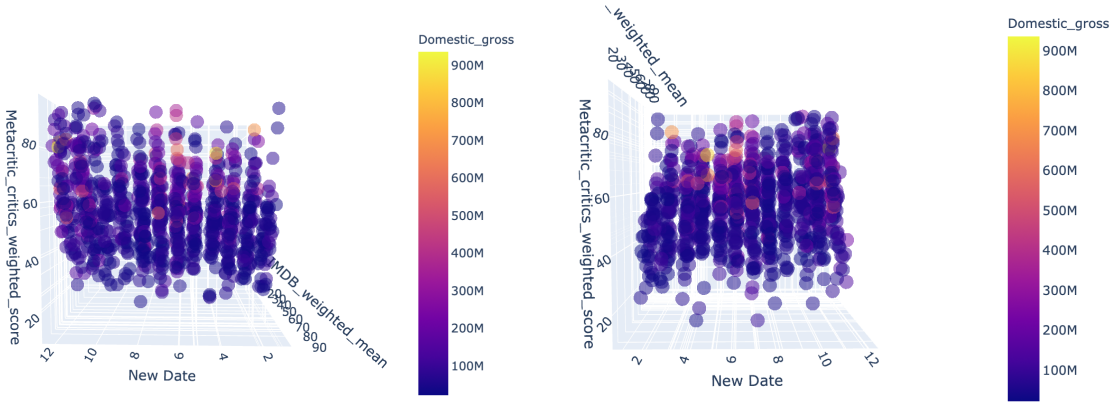
The implications of these estimates are fascinating. Firstly, the IMDB weighted score is the number displayed on the movie's website. This raises the question of simultaneity between ranking and score. Perhaps the relationship is solely based upon potential viewers seeing the rating and then deciding to buy a ticket because of the rating they saw. Or, the rating is contrived from domestic gross, giving already more successful movies better ratings, and thus, more viewers, further driving up both the ranking and score. While this is mainly conjecture, conducting rigorous questioning of even a statistically significant estimator is imperative to a more informed understanding of the relationship.

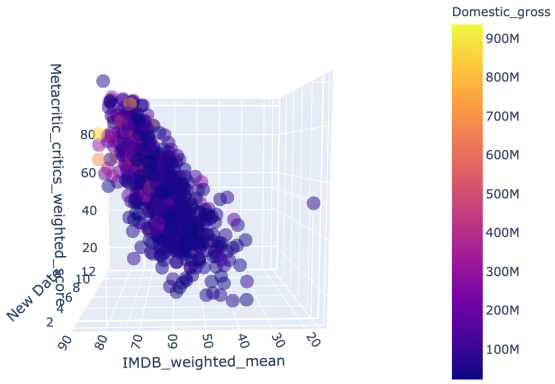
Another intriguing statistically significant metric is the average score of actresses in the movie. According to this regression, there is a negative effect on ranking from this score. This is a puzzling finding, because it ultimately suggests that movies earn more domestic gross when they have lower "quality" actresses. Perhaps more famous actresses, i.e. those in top-performing films, are critiqued more harshly because of their reputation. Perhaps movies making larger profits feature less females, biasing this estimator. While it is not surprising that female ratings have a statistically significant effect on ranking, it is, however, unexpected for the coefficient to be negative.

The final statistically significant explanatory variable is the number of voters contributing to the IMDB rating. It is a very minor coefficient, but the results of the hypothesis test warrant further discussion. Once again, the multivariate model determined a negative effect on ranking. Personally, this is baffling and calls for more intense evaluation to determine why and how this seemingly insignificant metric is one of the only statistically significant variables in the regression.

What patterns are there between release date, domestic gross, and rating?

Before creating the 3D model of these variables, a numeric date variable was generated as a float. The figure below shows IMDB weighted mean score, date, Metacritic weighted score, and domestic gross.



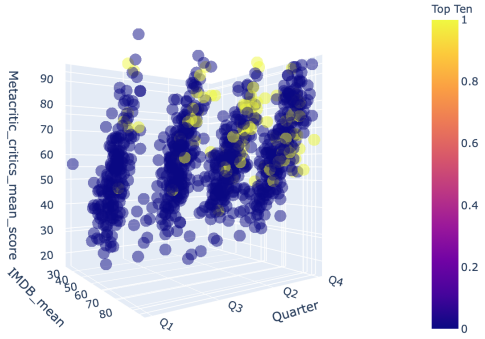


The observed patterns from this figure are not necessarily surprising. Metacritic and IMDB ratings tend to be consistent, and films with higher scores bring in more money. However, something to note is the higher domestic gross earlier in the year. Multiple movies might receive stellar ratings, but it is the ones released during the first half of the year that earn the highest relative domestic gross.

What patterns are there between ratings, top ten ranking status, and quarter of release?

Before constructing this figure, two categorical variables were generated. The first one, “Top Ten”, indicates whether a variable was ranked higher than 10 in the annual domestic gross rankings. The second one, “Quarter”, simply states in what quarter of the year the movie was released. For example, a movie labeled “Q1” was released between January 1 and March 31 of its respective year.

The figure is displayed below, with Top Ten movies in yellow.



The most fascinating observation is the lower Metacritic rating for Top Ten movies at the end of the year. In the first three quarters, the Top Ten movies receive high Metacritic ratings, but in the last quarter, they are noticeably lower. The IMDB score remained somewhat constant across the year.

Conclusion

After coming together as a team and truly diving into this data set, the discoveries and relationships that we found far exceeded our prior expectations for this project. From basic univariate analyses of audience and critic scores to multivariate analyses of domestic gross, ratings, and release date, we often had difficulty settling on which questions to investigate and explore. We were able to discover trends that were initially obvious, but more importantly, trends that were harder to find, buried in the numbers. For example, we discovered that audience and critics' average scores did vary greatly from one another, that Walt Disney manages to produce the greatest box office success among all distributors, and that December and June have historically been the best months to release a new film. These findings are truly interesting, but furthermore, these findings have very significant implications. Not just for filmmakers, but for casual viewers too. After discovering these surprising trends, we learned how to appropriately and effectively model our findings.

This project has shown us how surprisingly unpredictable the data can be by producing results contrary to the notions that we previously held. That real world data can be messy in its purest form, but after organizing it and manipulating it to meet our needs, we can glean truly profound results. That working together as a team can be difficult, but can also be extremely rewarding and thought-provoking. That real-world analysis can have incredibly complex sets of variables and factors to account for. We believe that our data speaks for itself and that we found it to be an invaluable experience starting with a simple inquiry for research, to opening up doors to innumerable opportunities for discovery and exploration.