Final Report

TItle and author(s):
Data Analysis of COVID's effect on box office performance and box office score prediction.

Jun Yong Shin

Summary of research questions:

1. **Before the covid outbreak, was there any correlation between properties (genre, budget, IMDB rating) and box office performance?**

   After analyzing and visualizing the data, I could conclude that there was a correlation between genre and box office performance, budget and box office performance, and IMDB rating and box office performance. However, there some showed more correlations than others.

2. **Can we predict box office performance if there wasn't a covid outbreak?**

   By using the machine learning regression model, I could get the predicted performance data. The detail will be explained later in this report.

3. **How did covid affect the box office performance?**

   By comparing the numbers we get from the ML model and actual performance as well as visualization, I could see there was quite a huge difference between the two. Also, it showed drastic changes in well-performed genres.

4. **What was the actual difference between actual earnings and the expected performance of the movie Black Widow?**

   After training the ML model using data before 2020, I passed the data of Black Widow to get predicted data. There was quite a difference between actual and expected earnings.

Motivation:

Many people enjoy watching movies and films. I have seen so many blockbuster movies score big and make so much money at the box office. However, since the covid outbreak in early 2020, many industries have suffered, such as the film industry. One aspect is movie theaters took a huge hit during COVID since they had to shut down. This has impacted many movie box offices as this is what is considered the “success” of a movie.

So many people have been affected due to movie theaters shutting down. For example, during the pandemic, the movie Black Widow was released, and there has been controversial litigation between Disney and actress Scarlett Johansson. Johansson accused Disney of lowering box office performance by releasing the movie on the Disney Plus platform, which affected her box office score-based income. Such problems posed a question of whether we can predict the box office performance if there wasn’t the COVID outbreak and what will be the difference between actual and predicted values.

I wanted to research what movies did well at the box office and how without COVID. How it would have shaped the box office of such movies. How is box office related to a movie’s properties? These questions are extremely important as we are trying to see if there is a shift in this large industry and a potential shift in the entertainment industry that we all love.

Dataset:

IMDB dataset:
https://www.imdb.com/interfaces/

Movies Budget
https://www.the-numbers.com/movie/budgets/all

Method:

1. After loading IMDB file into dataframe using Pandas, I filtered columns to reduce datasize and leave only important data. Also, I used apply method to convert data more convenient form for further analysis (e.g. string value into an integer value.)
2. To get budget and performance gross data, I used pandas.read_html method. However, since there was 60 pages of information, I made a for loop and stored links in a list. (I could use ‘for loop’ since there were certain patterns in terms of link address. For example, the first page ends with 101 and the next page ends with 201, and so on.) By making a def function, I got dataframe from each link address and stored them in a list. Finally, I used pandas concat to combine all dataframe saved in a list.
3. Combined IMDB data and budget data using merge method.
4. Combined data was separated into two using filter. One is before covid data (years between 2000 and 2019), and the other is after covid data (years 2020 and 2021).

5. Budget-genre, budget-gross, log budget - log gross, and rating - log gross graphs were plotted to see the correlation between them.
6. After finding correlation exists, ML model was trained using data before the COVID.
7. Used model to predict data after COVID outbreak and compared the difference.
8. Used ‘Black Widow’ movie data and compared predicted and actual data.

Work plan:

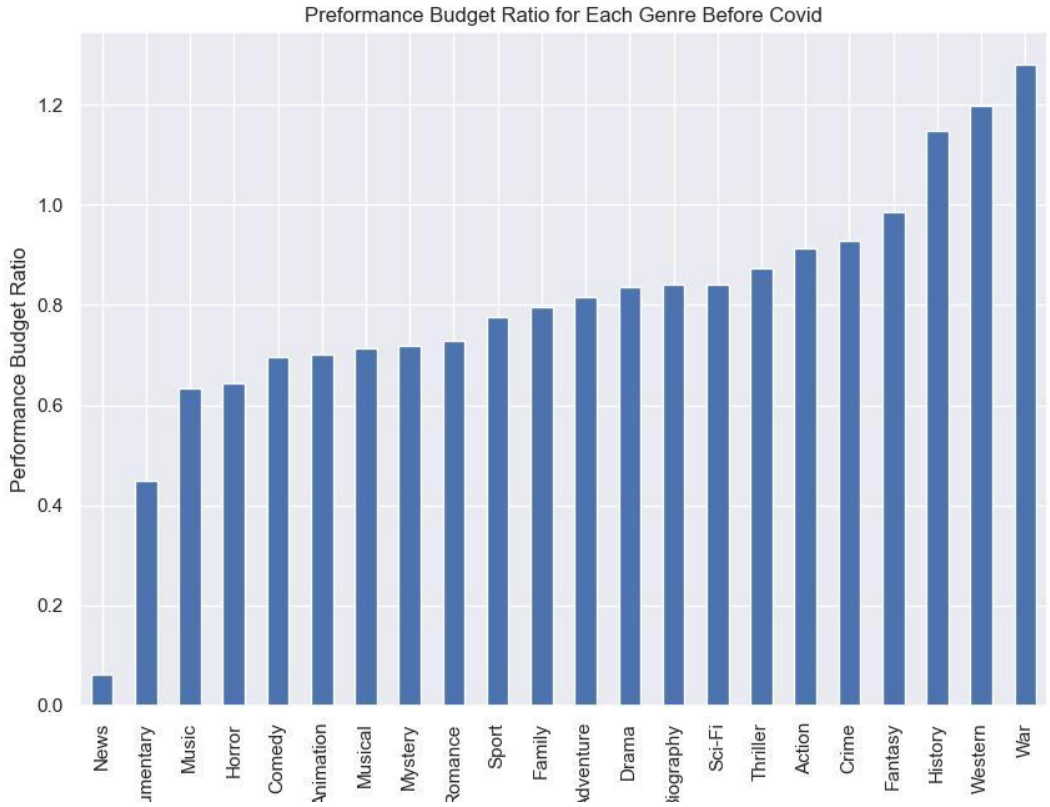
We will primarily be working through Google Colab to write our programs that will help answer our research questions. Any extra files we may need will be shared on Google Drive. Communication and collaboration will be done on Discord and in person. Since we need to combine and work with the same datasets, we will be collectively working on the programs and research questions rather than split up the responsibility. Since we’d be working on the same program, if one of us runs into trouble it can easily be shared with another group member for help. We will split up the report so each of us completes some sections and the video presentation will be done together. We will have the analysis done by March 4th so we have a week to write up the report. We expect that collecting data into a usable form and combining it will take around 4 hours and writing code to answer the research questions will take an additional 10 hours.

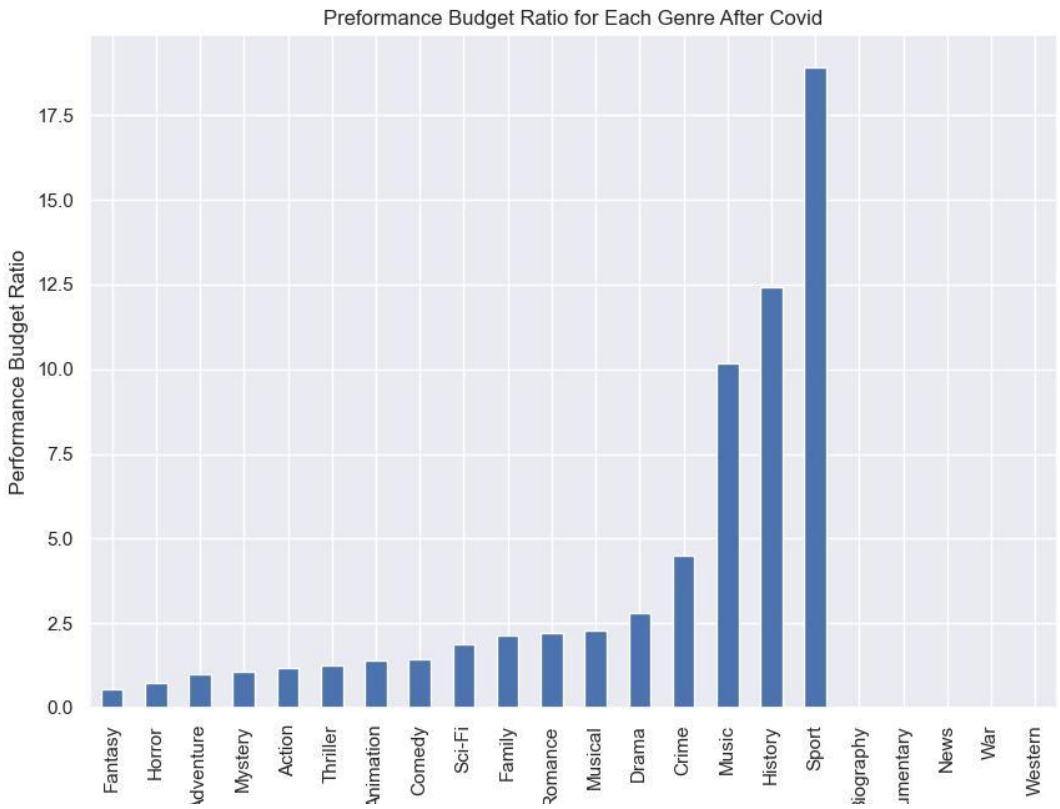
Results

- Present and discuss your research results. Treat each of your research questions separately. Focus in particular on the results that are most interesting, surprising, or important. Discuss the consequences or implications.
- Interpret the results. If the answers are unexpected, try to offer an explanation. A good report not only presents the results, but provides an argument or interpretation based on the data analysis
- Include any visualizations you have made. In general, these should be generated programmatically as part of your project code. If you plotted by hand, explain why it was not possible to create the plot you wanted in Python

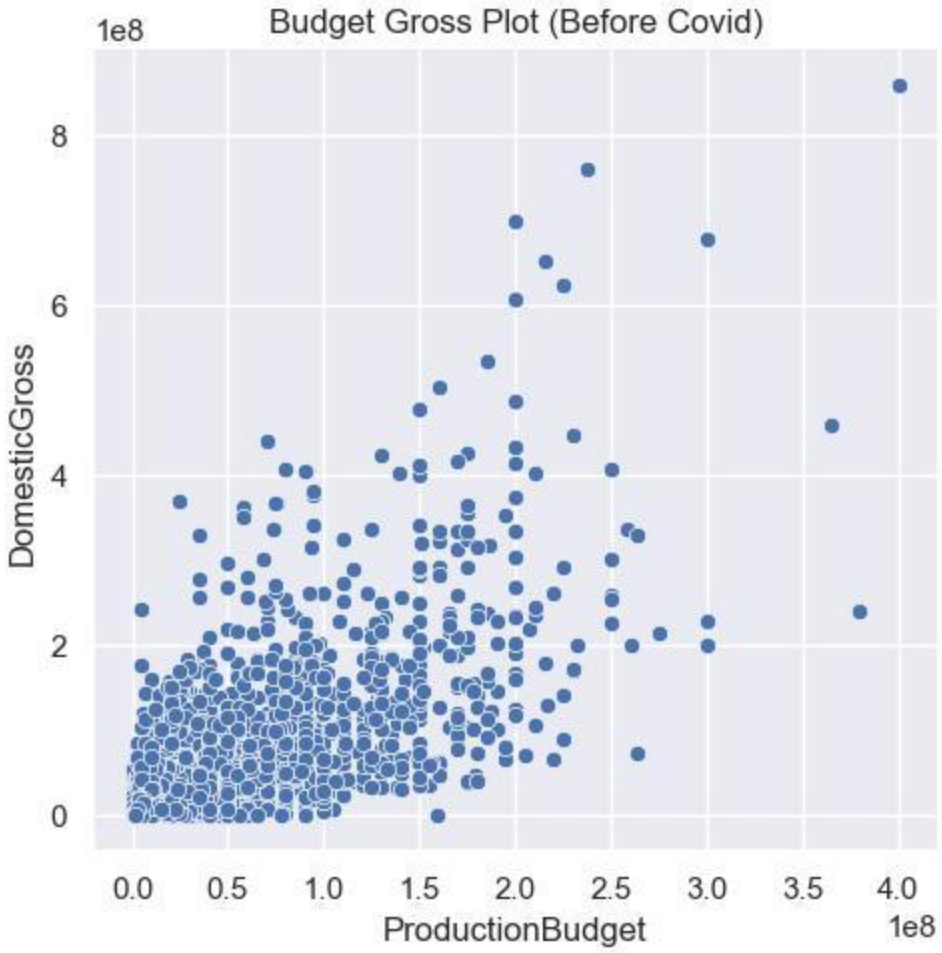
As I visualize the data, it showed that war, western, and history genres made revenue more than 100% compared to their budgets (more than the value of 1 in performance per unit budget data), which was not what I expected at beginning of the research. My expectations were comedy and actions or sci-fi will have the highest values since they are usually ranked top in terms of box office score or number of viewers. This result tells us that those genres also spent much more on the budget than

other genres. In addition to the difference between genres, there was a drastic change between before COVID data and after COVID data in the performance budget ratio graph. However, this difference is believed to come from small amounts of data. Since the movies released in 2020 and 2021 were quite small compared to data before, it can be skewed and hard to normalize due to the small size of data. Therefore, I decided to use only before data to find correlations between movie properties and domestic gross as well as ML training.

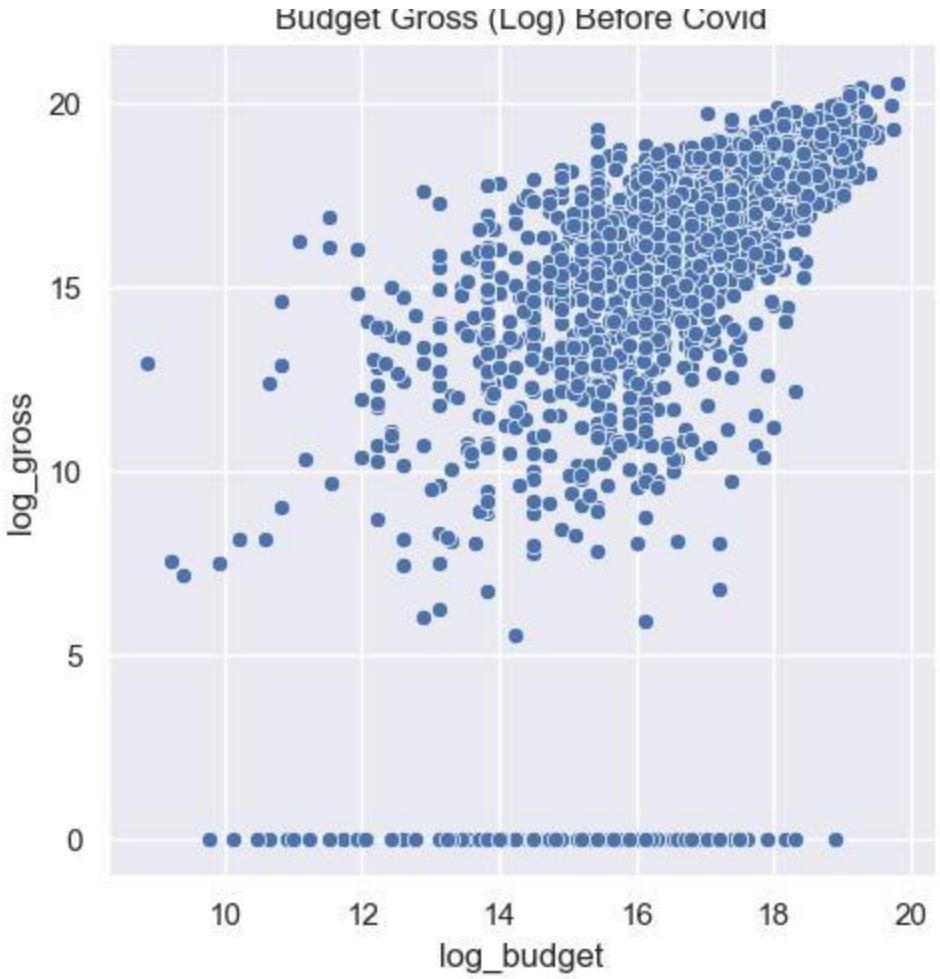




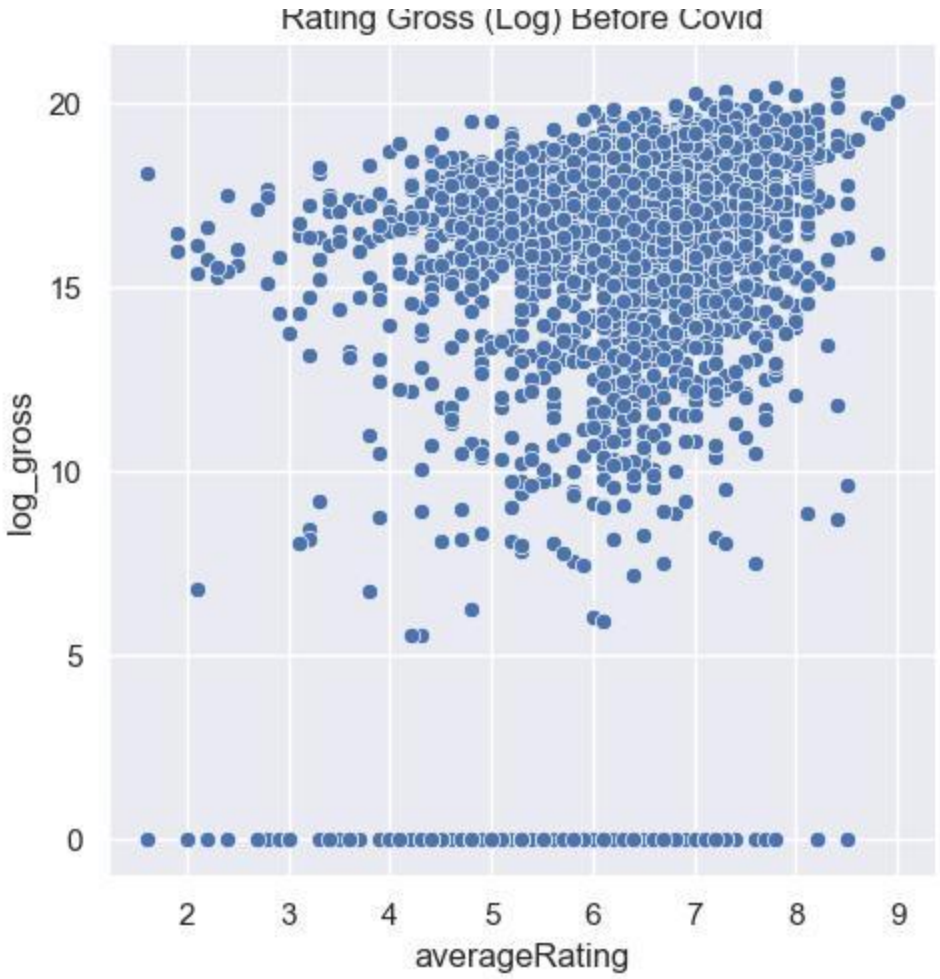
After comparing differences between genres, I plotted budget and domestic gross.



However, as we can see from the graph, they were clustered at the lower left of the graph since the numbers were too big. To lower the values, I changed this data into log scale.

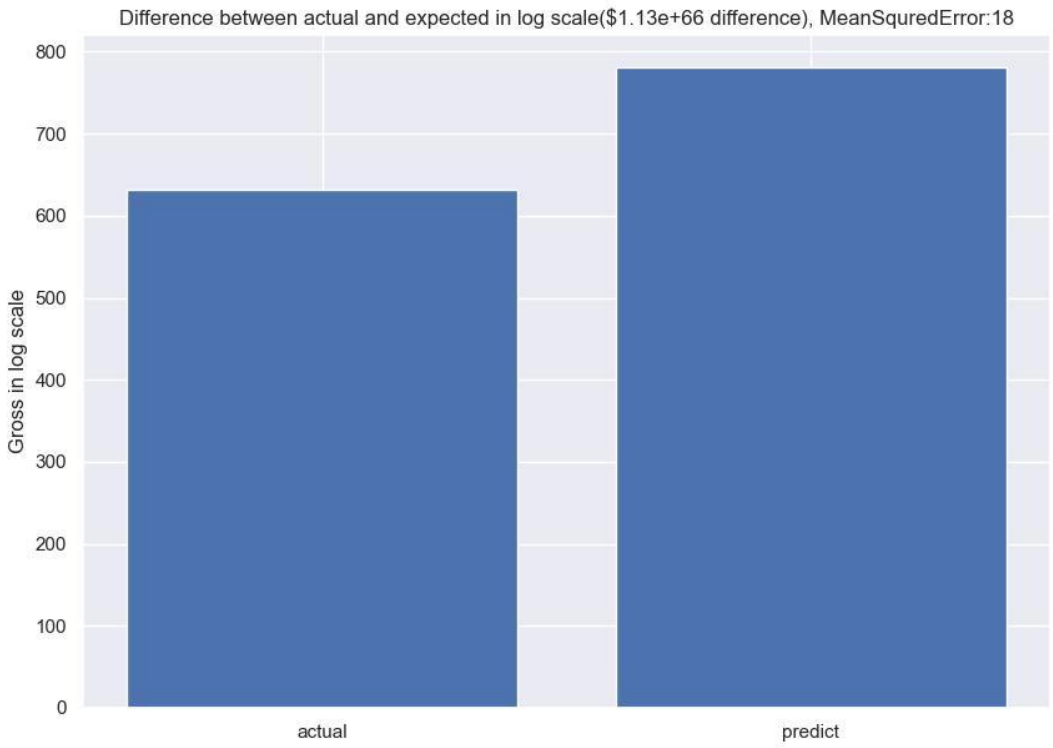


And it was clearer that the graph has a tendency to head right top corner. I could conclude that there is a correlation between budget and domestic gross and correlation is strong as budget increases since the spread data points get narrower as budget increases.



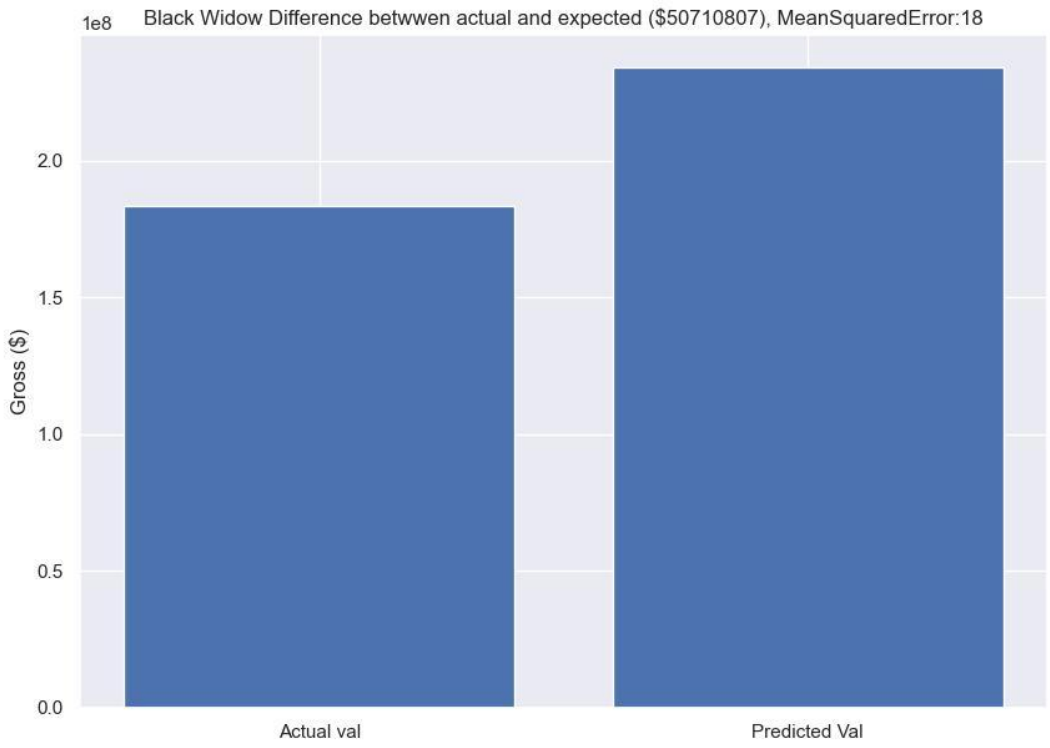
I could find correlation between rating score and domestic gross, but the correlation was weaker than budget gross data. From this data, I could learn that the rating score doesn't reflect domestic gross (domestic box office score) proportionally. For example, we can see some data points with a quite low rating score but have high gross, while some have almost 8 to 9 rating scores with low gross scores.

After finding the relevance between movie features, I made an ML regression model using data before the COVID outbreak. This model has a train set mean square error of 0.03 and a test set mean square error of 18.



The difference between actual and predicted domestic gross in entire movies released after 2020 was $6.61e+64. Even though this value is a huge amount, it is only 1.78e-208 percent of the actual value, which means almost 0 percent. It was quite a surprising result to me that there was not much of an effect due to COVID even though theaters were closed and the box office scores were expected to be lower.

I also tried to check if this result is only true on a large scale or in a single movie too. By using the model I built before, I passed the movie ‘Black Widow’ data to get the predicted value.



In this graph, result showed that the difference between the predicted and actual value is about 27.6% of the actual value.

As a result, it seems like some of the individual movies had quite of an effect due to COVID while the entire movie industry's value was similar to the predicted value(little effect on the industry due to COVID). However, there were certainly some limitations on this result and research which will be explained next.

Impact and Limitations

The effect of COVID based on the domestic gross difference between expected and actual was quite high as expected. However, when it was converted into a percent value, it was small enough to be ignored. On the other hand, there was a quite huge difference, about 27.6 percent, between predicted and actual data of Black Widow movie gross. This result tells us that even though the entire movie industry's gross was about the same as before, the individual movies could have been affected by COVID. It can be used for movie producers who are not sure whether they should delay the release of their movies due to COVID and to predict how their movie will perform at the beginning of the release using rating scores they received.

However, there were certain limitations too on this result and research. As we can see on the performance budget ratio vs genre graph, there was a huge difference between before and after COVID in terms of genre performance, which makes it difficult to use a prediction model. There was quite a deviation between those two data and the ML model also made the mean squared error of 18 on its trainset data. Also, compared to data before COVID, the data size of movies released after 2020 was too small to compare and generalize the result. And finally, the effects due to streaming services were not considered in this result and that can cause a huge difference. Based on those limitations, I believe that more study about how those errors should be interpreted and the impact of streaming services should be considered before using this conclusion.

Challenge goals:

1. **Messy Data:** While processing the dataset, I found that some of the data had different types than other data on the same column. For example, on the Release year column, one of the data had the format of '10-05-2019' (month/date/year) while most of the other formats had '2019' (year) format. I need to check those data impurities and need to use the 'apply' method to filter out those differences. Also, while I was trying to use a different dataset, I found it did not include all the fields I need and had to change datasets after analyzing different datasets.

2. **Multiple Datasets**: Data about the movie such as release date, genre, and IMDB ratings did not come in the same dataset as box office performance and budget. I had already found multiple datasets each with relevant information, so I had to combine those datasets to utilize for further analysis.

3. **New Features**

Since not all of my data is conveniently organized, I needed to collect data from different places. To collect the data table from the webpage, I had to use pandas.read_html features as well as request. Also pd.concat was used to combine datasets.

4. **Machine Learning:**
   In addition to analyzing past data on movies and how Covid-19 may have affected their commercial success, I used ML regression model to predict and compare data before and after the covid outbreak.

Work Plan Evaluation Evaluate your proposed work plan.

At first, it was expected to take about 4 hours to complete filtering and merging data. However, as I face impurities and errors on datasets, it took more than 15 hours to understand datasets and combine them. I could follow and conduct the work plan I had, but the time required to fix one small problem or bug was much longer than I expected. At the end of the project, at least a total of 27 hours were spent to finish this project. I used google Colab to test and see the result at first as planned and then moved to PyChamr to write '.py' files. Overall plan and steps were close to my expectation but the time spent was way far from reality.

Testing Describe how you tested your code.

Since the dataset was too huge to analyze or make small dataset for it, I used Jupyter notebook to test the result using df.head().

```
rating_filtered = rating_data.loc[:,['tconst', 'averageRating']]
rating_filtered.head()
```

	tconst	averageRating
0	tt0000001	5.7
1	tt0000002	6.0
2	tt0000003	6.5
3	tt0000004	6.0
4	tt0000005	6.2

```
movie_merged.head()
```

	tconst	primaryTitle	originalTitle	startYear	genres	averageRating
0	tt0011216	Spanish Fiesta	La fête espagnole	2019	Drama	6.9
1	tt0035423	Kate & Leopold	Kate & Leopold	2001	Comedy,Fantasy,Romance	6.4
2	tt0062336	The Tango of the Widower and Its Distorting Mi...	El Tango del Viudo y Su Espejo Deformante	2020	Drama	6.3
3	tt0063351	Summer in Narita	Nihon Kaiho sensen: Sanrizuka no natsu	2012	Documentary	7.2
4	tt0069049	The Other Side of the Wind	The Other Side of the Wind	2018	Drama	6.7

Also, I used made test file and tested whether it includes unwanted year values as well as to check functions.

Collaboration State

While I was working on this project, I used Stack Overflow as a resource to get help as well as libraries documentation.