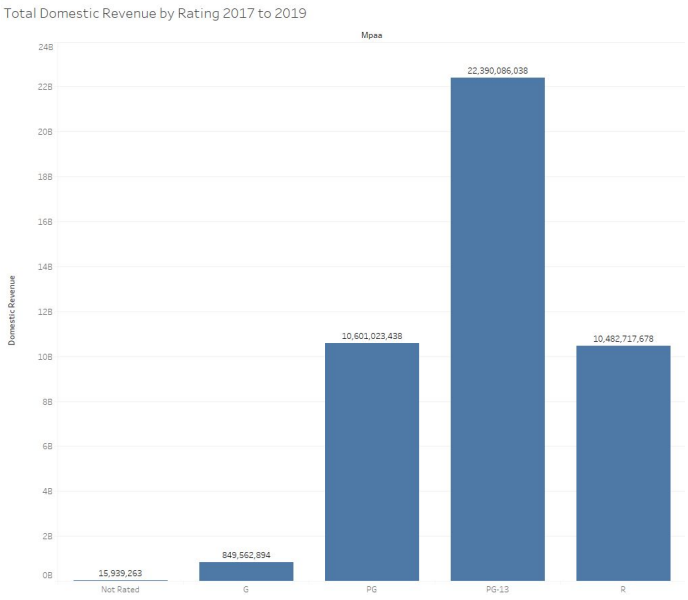
Movies Pre-Covid (2017-2019) by CovidCinema

Rich Finley, Yi-Ching Hsieh, Lanny Xu, Zhong Xie, Robert Kaszubski

The movie industry has seen a rapid decline due to the Covid-19 Pandemic. Theaters are on the verge of bankruptcy. Studios are in the midst of confusion, not knowing what to do with their already shot productions. It was just last year that the industry saw a record breaking year. Our dataset features every single film released in theaters in the United States between the years of 2017-2019 and the financial information associated with each. This includes the film's title, domestic box office gross, worldwide gross, opening weekend gross, as well as other factors such as the theater count, the number of days in release, and the film's budget. This dataset was found on Kaggle but was sourced from BoxOfficeMojo - a website that contains just about every piece of information relating to a film's box office. We are primarily exploring this dataset looking at the relationship between the big studios capable of releasing the majority of films, both big and small budget, in relation to the smaller studios that are not. Everyone may be familiar with Disney, Warner Brothers, and Sony Pictures, but there are around two hundred and fifty different studios releasing films here. Many of these films, most have not heard of before as they are overshadowed by their blockbuster counterparts. We're going to be looking at those differences between the two.

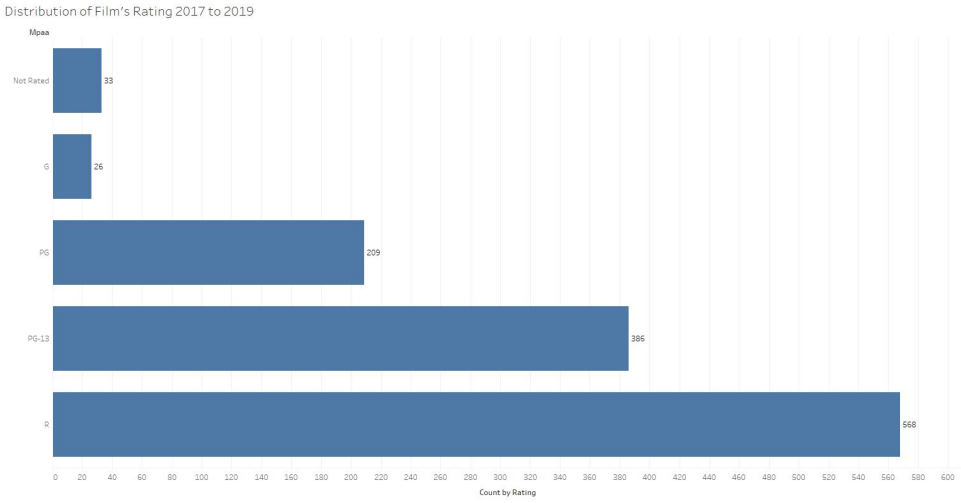
Exploratory Analysis

Our dataset gave us many different possibilities on what to focus on. We wanted to explore as many different angles as we could before settling on our topic. Our focus was always going to be on the financial side of movies because it is a box office dataset, but there are numerous factors to consider that influence a movies box office. We started our exploratory analysis by each picking two different aspects to explore and create some basic visualizations off. These included looking at the relationship between a movie's financials and the movie's rating, theater count, days in release, as well as opening weekends.



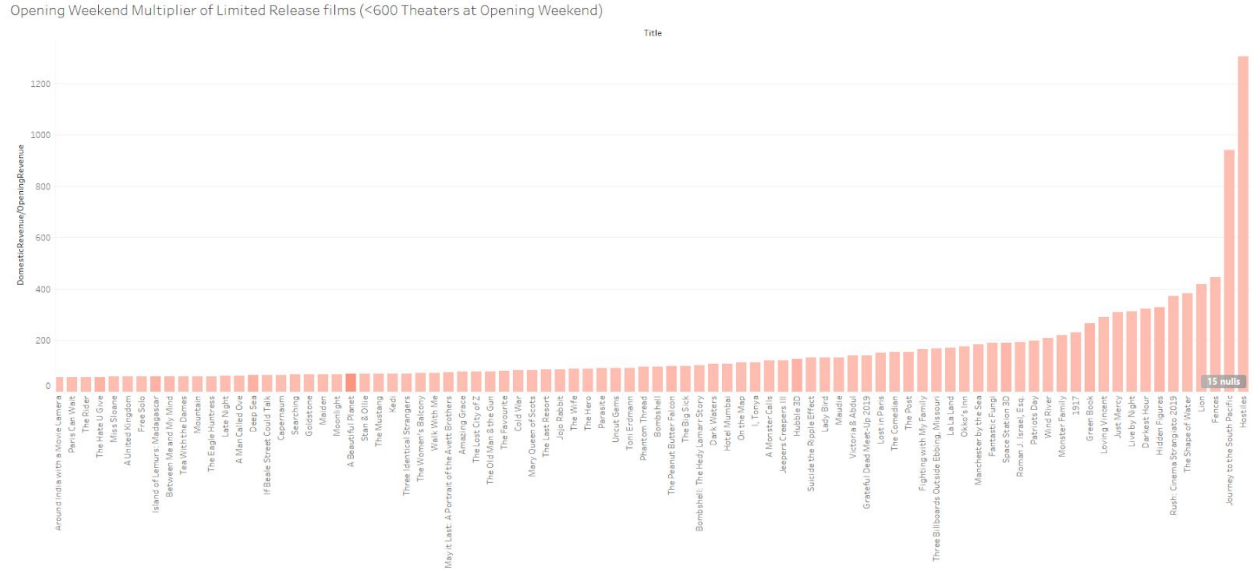
This chart was made to look at the differences in revenue made based on the rating a film has received. You can see that PG-13 movies reign supreme which is no surprise considering the bulk of big budget blockbusters aim for this rating in order to appeal to the widest possible crowd. PG and R are about equal and each less than half of PG-13. G is a very low number likely due to most films for children leaning closer to the PG rating.

The not rated category are films with low budgets and very small theatrical runs that likely couldn't afford being submitted to the MPAA to be rated. Our assumptions here were confirmed by our other rating related graph showing the distribution of films present in our dataset by their ratings.



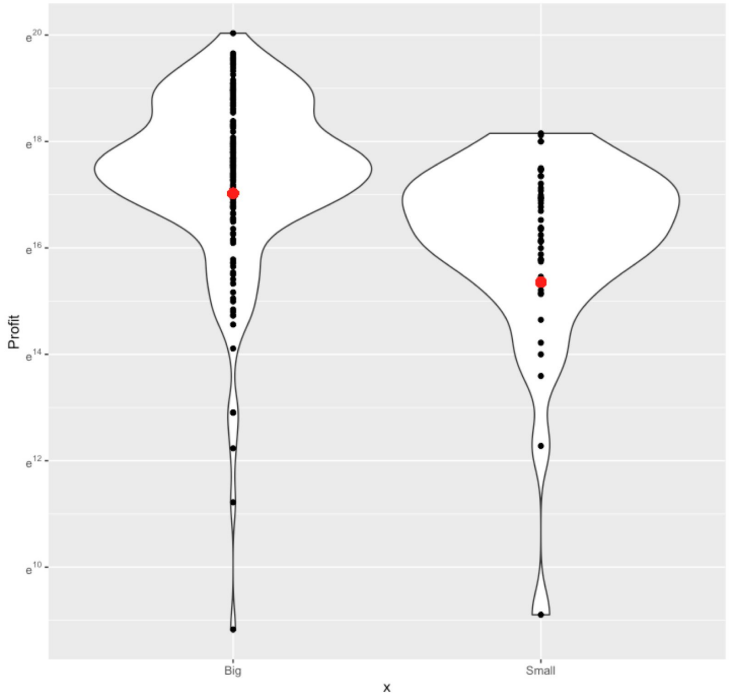
Interestingly enough most films were rated R yet PG-13 still more than doubled its overall gross.

We also looked into distributions based on budget. To get a good idea of how many small and big budget films are included. Another one of the other aspects we looked into that didn't quite make it into our finished work at all was the breakdown of domestic openings weekends and what is known as the opening weekend multiplier - a ratio of total domestic revenue divided by the opening weekend revenue. As expected, films that released in a low number of theaters typically had higher multipliers, at least those that saw a wide release later on.



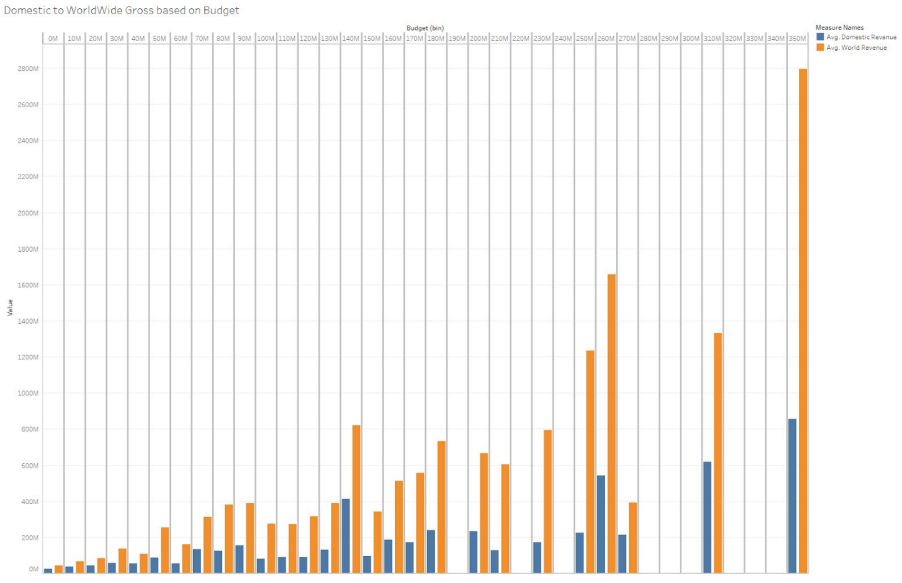
The rest of our exploratory work came from looking at big studios vs small studios which ended up being the story we wanted to focus on. This way we were also able to incorporate our findings about budgetary differences and other variables featured in our data that we felt explained the disparity in behaviour between the big studios and the small studios. The vast

majority of films released each year are by a select few studios meanwhile there are hundreds of smaller studios fighting for a single theater to release their films in.



This chart for instance shows the distribution of profit for the big studios versus the small studios giving us a good idea of the trends we should expect.

We also immediately looked into the relationship between domestic and worldwide grosses. We knew that both of these were key variables that we would absolutely have to use as they are concrete and numerically represent a film's and studio's success. The bottom line of any production is money after all. We saw a common pattern of the average domestic revenue typically representing less than half of the average worldwide gross no matter what the budget range was. However that ratio seemed to narrow the higher the budget became. We felt as this made sense given that most every big budget blockbusters played worldwide while miniscule budget productions likely never expanded to foreign markets. We felt that this was an interesting aspect to depict in our visualizations and worked perfectly for the story we wanted to tell in showcasing the differences in behaviour among studios of different scales.

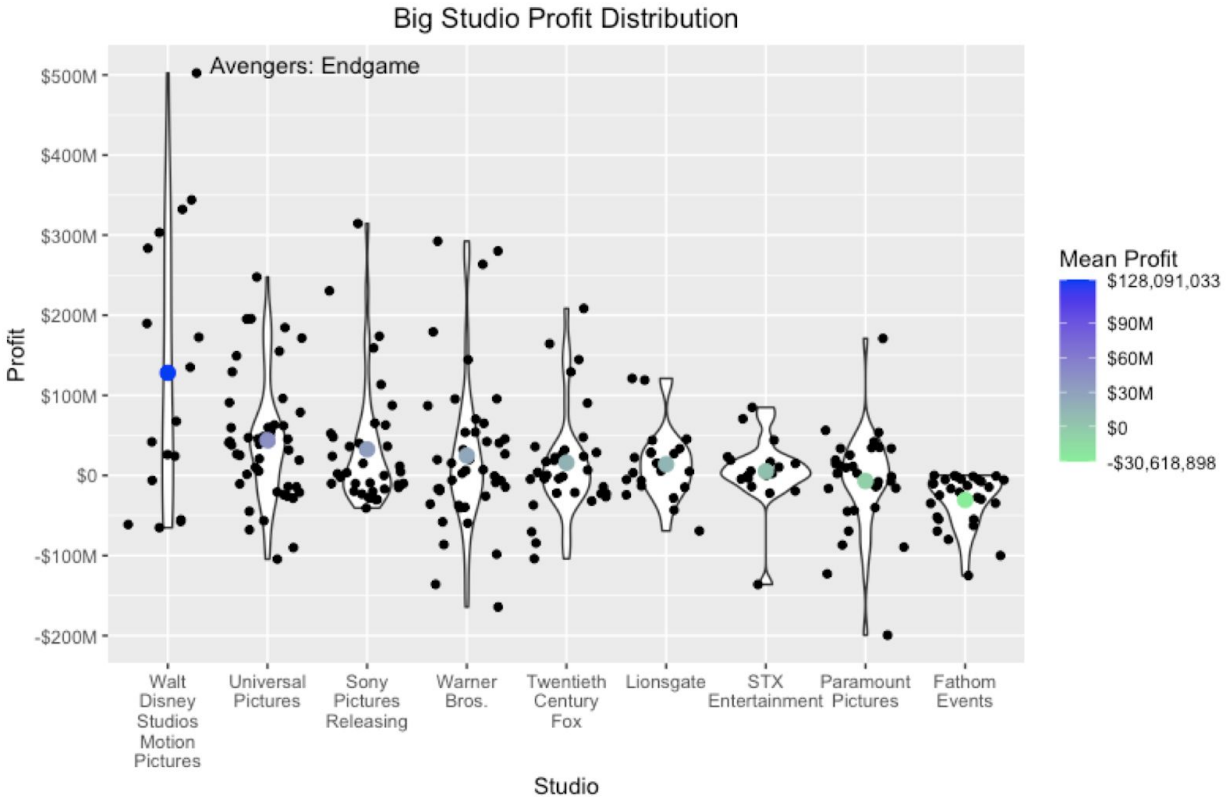


Visualizations

Big Studio Profit Distribution:

To begin this visualization, I created a filter on the data for studios that made 17 or more movies, and grouped them by Studio. My first iteration of the final visualization I was creating for this project involved the profits of each movie grouped by studio, and plotted using a violin plot with jitters depicting the specific profits for each movie within each studio. I also included the mean profit for each studio across all movies using a Red dot on each violin plot. This version was used in our Milestone 3 report.

After feedback on Milestone 3 and the presentation, I modified my visualization accordingly. The final iteration of the plot has the profits from each movie grouped by studio, using a violin plot to show the distribution, and jitters to show the specific profit of each movie. The visualization also includes the mean profit of all movies for each studio. The Mean Profits are plotted using colored dots, with the highest profit margin in Blue, and lowest in Light Green. I also sorted the plot by Mean Profit margin for each big studio. The profit tick marks themselves are set in "hundreds of millions", as displayed with each tick mark. I also included a legend on the right-hand side to show how each color variant corresponds to each profit level. I set the legend tick marks and labels to include the minimum and maximum Mean Profits across the studios:



We believe this is an effective plot at comparing the highs and lows for the "big" movie studios, classified as making 17 or more movies. As you can see from the visualization, Disney has the highest profit margins from their movies. Their mean profit is well above $100 million, and their highest profited movie they made is just above $500 million (Avengers: Endgame, which I have labeled on the plot). Even if you disregard the Avengers: Endgame, they still have the second and third highest profited movies at around $350 million. Sony is really the only studio that is close to having the 3rd highest profited movie. Most of Disney's movies have profited, while all the other studio's distributions are showing quite a bit of movies that lost money.

The mean profits for all the other studios do not exceed $50 million, and only Paramount Pictures and Fathom Events operated at a Mean loss across all their movies. Fathom Events is a bit of an outlier, as most of its movies were re-releases of older movies. So the profit was calculated with its original budget, and the revenue it made upon re-release. So naturally they will have a lower profit margin.

The other interesting takeaway that we can see from using a violin pot with jitters, is that while Sony has a decent Mean Profit, the jitters show that most of its movies have lost money. Its Mean Profit is being brought up a bit by a couple movies that have profited over $300M and $200M respectively.

Overall, this visualization succeeds in showing how well each studio performed in terms of Profit for all the movies they released from 2017 to 2019. The big takeaway from this is Disney having such high profits. This makes sense, given that they own Marvel and Star Wars, two franchises with some of the most widespread followings.

Domestic vs International Revenue of Big Studios

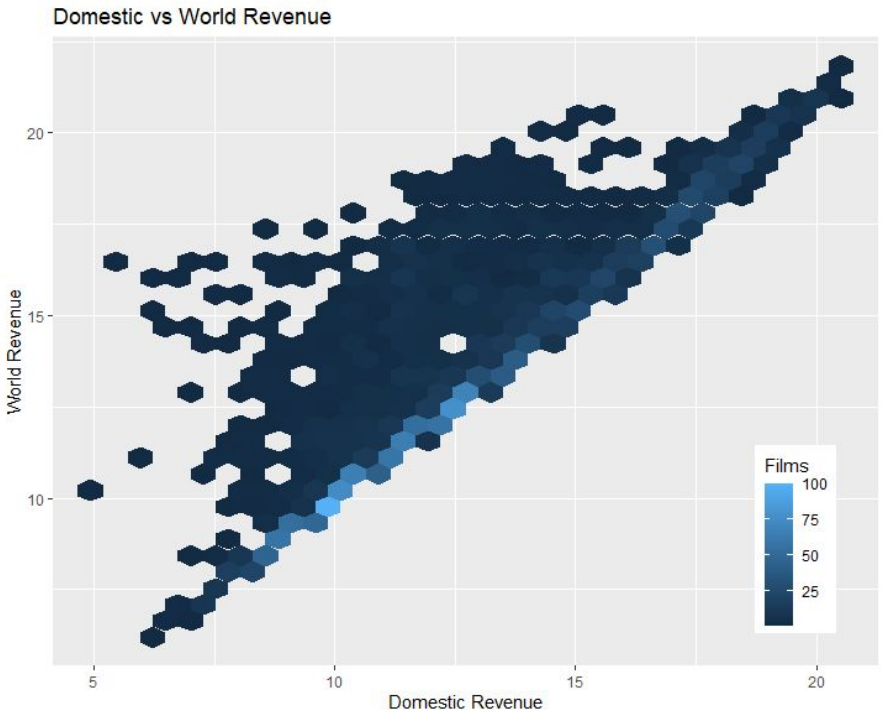
This visualization was made using Tableau. It depicts the domestic and international revenue of what we considered to be the big studios (filtered out the same way as the previous visualization). Below this is another chart that better pictures every single film those studios had released during this time frame. As our dataset didn't contain an international gross revenue variable, a new calculated field had to be created subtracting domestic revenue from worldwide revenue. From there the domestic revenue and created international revenue variables had to be pivoted in order to create this treemap with this particular hierarchical structure. The structure goes domestic/international to the studio then to each movie of that studio. This visualization saw numerous iterations with different chart types such as bubble charts and bar graphs. Careful consideration was put into placement and positioning of each studio's square on the tree maps as well as the coloring. The colors were meant to distinguish the studios without necessarily drawing immediate attention to any of them hence the use of slightly more subdued or muted colors rather than bright and radiant.

We believe this is good visualization to show a few different things. We can instantly see the dominance by Walt Disney Studios, leading in both domestic and international revenue. The studio released heavy hitters such as Avengers: Endgame, Avengers: Infinity War, and Star Wars: The Last Jedi, during this time frame. Big blockbusters like this were what catapulted Disney as the leading studio. The biggest takeaway from this graph which we believe the second chart really emphasizes is the quantity of films released versus the revenue made. As you can see, Disney released the second smallest number of films among the big studios yet led in terms of revenue by quite the margin. Interestingly enough, Fathom Events, the smallest studio in terms of revenue depicted here, released the highest number of films by a very wide margin.

It is also worth looking at the smaller squares inside each studio's box. The way this tree map is structured, each square represents an individual movie and the size of each square is correlated to the film's revenue. The largest squares representing films with the largest revenues of that studio are situated in the top left corner. The squares get progressively smaller as you move left to right and top to bottom. The bottom right corner consists of numer tiny square representing those films that made a lesser amount of revenue. Interestingly enough Disney doesn't have this happening, at least not to the extent of every other studio. Warner Bros. for instance has its share of high grossers but has a plethora of smaller films making up a sizable chunk of their revenue. This really highlights the difference in strategies that these studios utilize. Walt Disney prefers to focus on this blockbuster franchise ignoring smaller scale films almost entirely while Warner Bros. has a balanced bit of both.

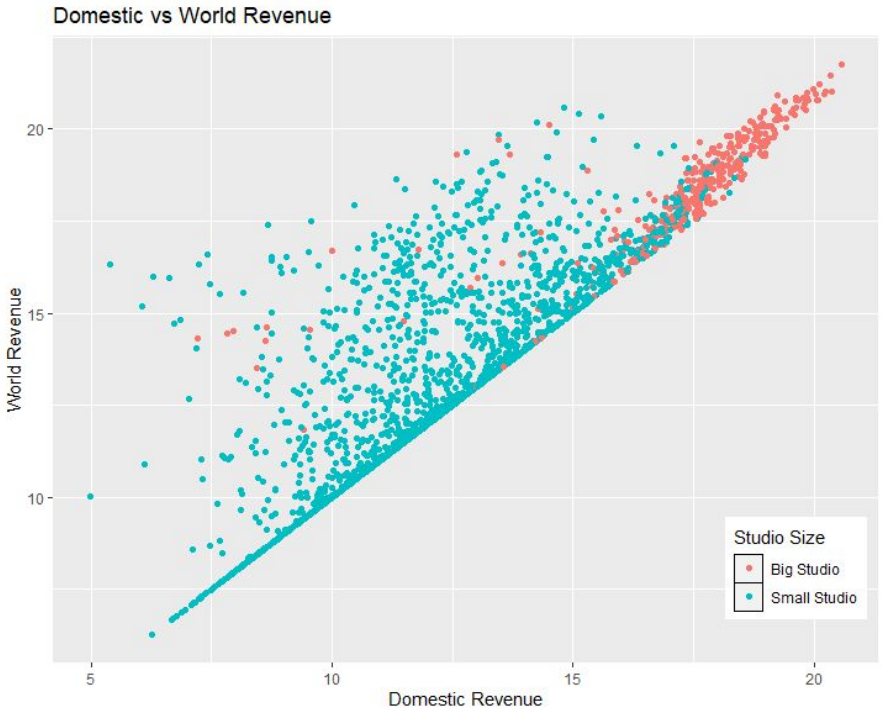
Another takeaway from this visualization is the differences between the domestic and international markets. For the most part studios scale proportionally here between the two, with international typically winning out. It shows how much bigger the foreign market has been when compared to the domestic market for Hollywood releases.

Domestic/World Revenue by Studio



The hex bin density chart was created using the hexbin library, which was integrated into the ggplot. In order to have a better visualization, the domestic and world revenue was transformed into log base e to reflect the percentage change. It also eliminated skewness as some films tended to have over a billion dollar of box office.

The chart focuses on the number of films in a set box office range. Creating a density heatmap representing how many films lie under the same box office range between world and domestic. We can see from this chart that films between domestic and world tended to have equal box office. Only a mirror portion of the films has a better box office in the international market, and most tended to be in the range between 6 to 17.



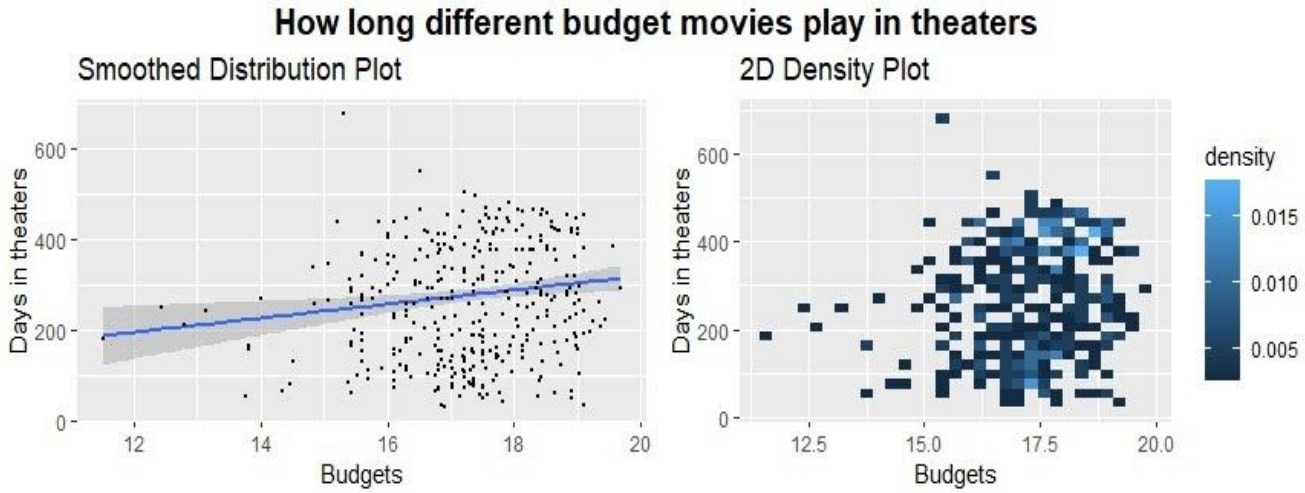
The chart is a scatter plot using the exact data from the first chart except for studio size. The studio size measures the distributor's operation

sizes, the number of film releases per year, and the average revenue range. Big studios tended to produce mainstream films with an extensive theatrical release. Small Studios is an indie studio that focuses on smaller budget films and often releases less than 5 films per year.

The scatter plot produces a better image of outliers or anomaly. We see that by adding big Studio and small Studio to the equation, Small studios tend to perform better on the international stage compared to big Studio. The graph also shows that the big studios often produce the highest box office. If they do not perform well domestically, they often get better internationally.

Combining both charts, we see a trend of small studios might focus solely on the international market or perform considerably better than the domestic market. It is possible that the film was produced for the international market but was imported to the US market, which explained why big studios have the tendency to have equal revenue on both the world and domestic.

How long different budget movies play in theaters



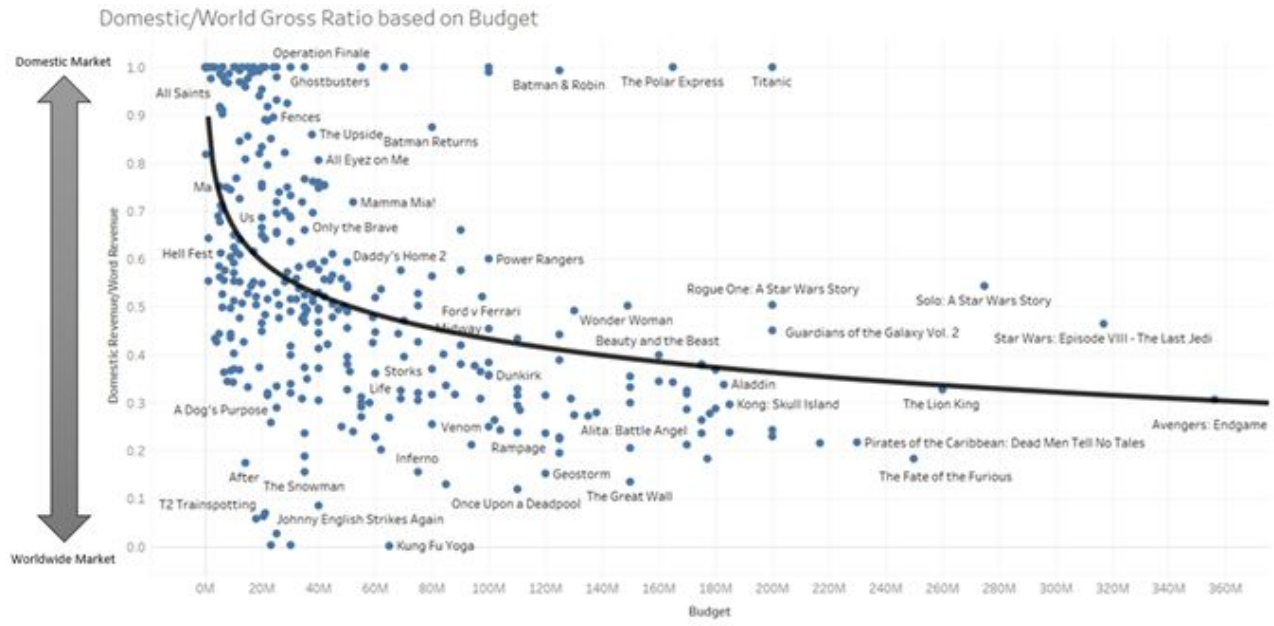
The number of release days in this dataset means how many days the movie is played in the movie theaters. The budget includes all costs relating to the development, production, and post-production of a film, in this dataset, we are just comparing the numbers.

To create the Smoothed Distribution Plot and 2D density plot, I removed all the NAs with special characters in budgets and release_days, and transformed them to be numeric in r. Also did the data transformation to make the Budgets to be log of Budgets, because they tend to be huge, and the log transformation removes the skewness of the original data. Also made some modifications based on the presentation feedback. The Smoothed Distribution Plot uses Smoothing in ggplot, which is as simple as adding a smoothed layer. I used geom_smooth to provide model-based smoothing, which includes LOESS and a linear line. The 2D density plot is useful to study the relationship between 2 numeric variables, here as the Budgets and ReleaseDays. To avoid overlapping, it divides the plot area in a multitude of squares and calls

using the geom_bin_2d() function. This function offers a bins argument that controls the number of bins to display. Their ranges seem close to each other, and variables are likely to fall near each other. I have also created a grid layout with ggplot by combining both of the plots together by using the 'gridExtra' library for better comparison

Looking at both plots, it is clear to see big-budgets films tend to have longer days in the theaters than the small-budgets films, and the budgets and days in theatres of different movies seem to be synchronized and concentrated.

Domestic/World Gross Ratio based on Budget



This scatterplot shows the relationship between the gross ratio of the domestic gross and the worldwide gross in relation to the budget of the film. The ratio is calculated by Domestic gross/Worldwide gross. If the ratio is near 1, this means that the film has more domestic gross than the worldwide gross, as the ratio is near 0, meaning more worldwide gross than domestic gross. There are several movies that have a domestic to worldwide ratio of 1 in the lower budget area, which means they made their profits domestically. This shows that most of the films are receiving limited budget to film and have less chance to get exposure overseas. And as the budget becomes bigger, the ratio has distributed mostly between 0.20 to 0.45, which means most of their profits come from worldwide. This shows us that with more budget to create a film, the chances that the film will be exposed to the world will be slightly bigger. This also reflects that the international movie markets are growing over the years, movie makers will eventually promote their work to the world if they have more budget to spare for more profit. There are also some films who have bigger budgets and the ratio of 1. It may suggest that those big budget

films didn't promote their work worldwide, but as these films are considered outliers to the data, they could misrepresent the data and can be excluded from the graph.

After feedback on Milestone 3 and the presentation, I added the thickness of the trend line to make it more conspicuous to the audience. And I added the label of the movies to make it more clear about what exact movie the graph wants to show. I added the two-headed arrow to give the audience a better knowledge of the ratio instead of mentioning them next to the chart. There are several interesting findings after the adjustment.

The outliers that have big budgets but only promote domestically turns out to be one of the most famous films, which are "Titanic", "The Polar Express", and "Batman and Robin". These are all rereleases of old films during our time frame. It's funny to find out that big movies can misrepresent the whole movie market trend. Another interesting finding is that Big Studios with Big Budget films balanced the Domestic/World gross ratio very well. Big films like "Star Wars", "Wonder Woman", and "Guardians of the Galaxy 2" have a 0.5 ratio, which means big budget films can choose to make profits from both of the markets due to their dominance of the movie industry. Overall, the graph shows that most films with smaller budgets aim their target on the domestic market, and as the budget gets bigger, the worldwide market will play an important role for making more profits.

Conclusion

We found a lot of interesting patterns and trends. Some of these may not be too surprising, the big studios make more money than the smaller studios. But this is a lot more nuanced than that which is shown in our visualizations. Comparing the different big studios we can see studios' release strategies, the film's they produce, how many films they produce versus how much revenue they actually bring in. For instance, we have Walt Disney Studios which doesn't produce as many films as most of their primary competitors yet dominate in terms of revenue. We look at Disney's profit distribution and see little change or an almost uniform plot unlike other studios where most of their profit is similar movie to movie with a few outliers. We can also see the relationship of domestic versus worldwide gross among all the movies in our dataset to observe the reliance of foreign markets as the budget of a film increases highlighting the fact that bigger budget productions are capable of securing international releases. Possibly the biggest surprise was the much lower difference in days in release between films of differing budgets. While bigger budget films do stay in theaters longer than low budget films, the difference is not nearly as wide as one might expect.

It would be nice to create a few more visualizations incorporating the other variables in our dataset. We had some great exploratory work done particularly regarding film rating, and the opening weekend revenues. Those are some factors that could be potentially added to our existing visualizations or appended on as brand new visualization. This could be done while still maintaining our story and theme of analysing the behaviour of films released by big and small studios.

Appendices:

Individual Reports:

Rich Finley

Big Studio Profit Distribution - For the final group project, I was assigned the role of pulling data and creating visualizations detailing and analyzing the revenues and profits of the different studios who released movies between 2017 and 2019 from the dataset. To begin, I created a new profit field that takes the budget of each movie, and subtracts it from the revenue for each movie. I then grouped movies together by studio, and filtered out studios that only made 1 movie, and calculated a mean profit for each studio. This gave me an overview of how profits compared across all movie studios. For the purposes of this project, we wanted to analyze the various profits across the different studio sizes. To do this, I first filtered the data by studios that made 17 or more movies. We classified studios that made 17 or more movies as the "Big Studios" and studios that made less than that the "Small Studios". I separated the studios into the "Big Studio" and "Small Studio" categories, because I did not want to skew the data by comparing the profits of a studio that made 25 movies against a studio that made 3 movies. For the first iteration of the visualization, I plotted the Total Profit of each studio, to show the highest Total Profited studios, compared to the average and lowest total profited studios. I then plotted the Total Profits across all movies grouped by studio for studios that made under 17 movies. I used a univariate scatter plot to show the profits for each of these visualizations. Through these 2 visualizations, it became clear to me and the group that it would be most interesting to compare the profits among the bigger movie studios, analyzing the studios that performed at the highest and lowest profit margins between 2017 and 2019.

To compare these, I decided to use a violin plot for each of the Big Studios with jitters showing their movie's profits, and a colored dot on each violin depicting the Mean Profit for each studio. The visualization went through many iterations where I was refining different portions of the visualization based on how it appeared from my perspective, as well as feedback from milestone 3 and the presentation. I go into more detail about these iterations in the "Visualizations" section of this report, so I will not repeat myself here.

What I learned from this project - One of the biggest takeaways from this project and this course in general, is how much of an iterative process it is to create an effective, informative visualization. The process involves a lot of drafts, refining different portions of each draft. I learned how there is no one "correct" way to create a visualization, and that there are many routes you can take, as long as it aligns with the data, audience, and message you are creating the visualization for. I also learned how to create a visualization that is clutter free, and does not take away from the visualization itself, and the intended message for the specified audience. And

lastly, I learned which different visualization techniques, plots, and graphs are effective for different types of datasets, audiences, and intended messages.

Yi-Ching Hsieh

Domestic/World Gross Ratio based on Budget - For the final group project, I was assigned to do the visualization of the high and low budget movie. But the results of the chart are very easy, high budget movies tend to get higher profits and low budget movies have the lesser chance to be on the same level. Then I googled for more inspiration about the budget of making movies. One article points out Oscars has lots of international movies rewarded these few years, which can lend to the growth of the international movie market. I was very curious about the truth of this article, and I could get answers by comparing the domestic gross to world gross. To begin, we have to create a new calculated field called Domestic Gross/World Gross by calculating the two fields "Domestic Revenue" and "World Revenue". After creating the field, we put the "Budget" field on the column and put the "Domestic Gross/World Gross" field on the row. Now you will get the scatterplot of the gross ratio based on budget. Right click the mouse and select "Trend Lines", then click on "Show Trend Lines" to put the trend line on the chart. Then, put the "Title" field on the Mark section, and click on "Label" to show the label of each movie. From the chart I can surely tell that most low budget movies will make their profit from the local market. There are still some examples of high budget films earning their money domestically, but after all they are extreme examples to the industry and could be seen as exceptions. And the higher the budget goes, the more proportion of world gross there is. A good example of big budget movies having more world gross proportion is Walt Disney studios. Big movies like "Avengers", "Aladdin", "The Lion King", and "Star Wars" have below 0.5 ratio, which means they make more money at the international market. So the article is telling us about the real truth about the recent movie market.

What I learned from this project - the biggest acquisition from this project and this course is that it takes much time to create a straight-to-the-point and informative visualization. The process of making it really requires a lot of thinking and implementation. Another takeaway is that I learned very many visualization techniques from this course. Knowing the right way to show the true meaning of a dataset is a powerful tool to use nowadays, and telling a good story is always a useful tool for us too!

Lanny Xu

How long different budget movies play in theaters - For the group project, we split up and each created visualizations representing the 10 topics we came up with regarding our data. I

chose the role of pulling data and creating visualizations detailing and analyzing the release days of the different budget movies and distribution of revenue between 2017 and 2019 from the dataset. As time goes by, I decided to mainly focus on how long different budget movies play in theaters based on what I learned from class. This topic has more to show and it is more relevant to our project's subject.

To create the Smoothed Distribution Plot and 2D density plot shows in this final report, I removed all the NAs with special characters in budgets and release_days, and transformed them to be numeric in r. Also did the data transformation to make the Budgets to be log of Budgets, because they tend to be huge, and the log transformation removes the skewness of the original data. The Smoothed Distribution Plot uses Smoothing in ggplot, which is as simple as adding a smoothed layer. I used geom_smooth to provide model-based smoothing, which includes LOESS and a linear line. The 2D density plot is useful to study the relationship between 2 numeric variables, here as the Budgets and ReleaseDays. To avoid overlapping, it divides the plot area in a multitude of squares and calls using the geom_bin_2d() function. This function offers a bins argument that controls the number of bins to display. Their ranges seem close to each other, and variables are likely to fall near each other. I have also created a grid layout with ggplot by combining both of the plots together by using the 'gridExtra' library for better comparison.

What I learned from this project - I have learned a lot from the course and the project, to create a great visualization, I will always keep in mind to choose the right graphs for different job, know that I want to say, pick the right color for clear data stories and design for audiences' eye, carefully and intentionally apply texts, avoid chart junks and more. I used R more frequently in the second half of the class, and all the work I have done for this final project is solely produced using R. I used Tableau a lot in my work, that's why the study and use of RStudio in this class and for this project is even more valuable to me, because R can solve some tricky problems which Tableau is not able to.

Zhong Xie

Domestic/ World Revenue by Studio - For the final group project, my task is to explore the domestic and world revenue relationship by studios. Creating visualization that reflects the relationship and supporting our finding on studio size may impact how they produce the film, theatrical releases, and the budget of the film. First, I have to sort out the dataset by introducing two new attributes that help formulate the visualization; transforming the domestic/ world revenue into a log base of e, and categorize and classify the distributor (studio) into big studio and small studio. The classification was done using Python, iterating the entire dataset into a dictionary, cataloging each studio's number distributions and average revenue. Using the collected information, we classify each film's studio and add a new row, big studio, or small studio.

At first, I decided a scatter plot would show the relationship between big studio and small studio revenue trends by plotting the world vs. domestic revenue. My initial thought of the plot is too crowded and challenging for the audience to visualize the relationship. So, I decided to use the same dataset but in a hex bin plot. Hex bin is a mixture of heatmaps with a scatter plot. It retained the scatter plot information with an added heatmap, which helps the overplotting issue. The heatmap is the count of films in the same bin range. It tells the relationship that most films have equal revenue on both world and domestic. But switching to the scatter plot, the plot displays that the smaller studio will often have better world performance than a big studio. It either tells that small studios focus on the international market, or the film was imported to the US, which explains the low revenue on the domestic level.

If you had more time, where might you have liked to have developed your visualizations further? - I would focus on solely using R to sort and classify the dataset without the help of another programming language. Also, integrating interactivity into the plot could help the audience explore further and strengthen our theory. It is more memorable when exploring the data and seeing some of the films they might have seen in the last three years.

Robert Kaszubski

Domestic vs International Revenue of Big Studios - I made the treemap visualization looking at the relationship between the domestic revenue and international revenue between what we labeled the big studios. I also included a second chart below that represented the number of films each studio had released during this time period. I think that really drove the message home and allowed the audience to easily infer and compare the differing strategies these studios use when it comes to their releases. These visualizations were a long and iterative process particularly for the tree map. I had originally tried out other graph types to represent this information such as bubble charts, which although they looked cool, didn't show as much information nor were they easy to read and compare. What I like about the tree map is it's simplicity, the areas are relatively easy to compare but we also encode information about each specific movie through the size of each square. Not only does the reader see the big picture, but can also see what contribution each individual movie made toward that studio's total. In order to create the treemap, the international revenue had to first be created, then both the domestic and international revenue variables were pivoted. Careful consideration was taken in deciding the layout of the squares of the treemap. I didn't want them to appear too stacked up, so they could still be easily compared to between the domestic and international, but also didn't want to make the treemap too long. The colors weren't too difficult to choose. I didn't want any colors that were too vibrant that they only drew your eyes to them. Color is only meant to represent the studios here. The great thing with the color is that it matches in the two charts, so only the

second one has labels instead of including a separate legend for the treemap, this was a great suggestion a few people pointed out after the presentations.

Besides this visualization, I also worked on much of the exploratory work - two of my visualizations from that are featured above: the opening weekend multiplier plot and the domestic/worldwide bar graph. There are a few more of my exploratory visualizations featured in our Milestone 3. The domestic/worldwide bar graph served in many ways as the foundation of both this visualization and the other ones featuring this topic. Additionally I spent a good amount of time looking for this dataset when we were still forming groups. Zhong and I had the idea to do something relating to the entertainment industry and specifically movies. There was a surprisingly small number of quality datasets that included financial information about movies. We got lucky finding this great one on Kaggle, many others had a random assortment of movies, were short, or included only information like the actors, director, etc.

I had never worked much with data visualization besides making a few simple graphs and charts in my other classes so this entire class was a big learning experience. I learned a lot about just how much thought and consideration has to go into every single aspect of a visualization in order for it to be successful. I was also surprised at just how long of a process it can be to create a good visualization. You have to go through numerous adjustments and changes before you finally land on one that's effective. I think one of the most valuable skills that this class taught me was being able to look at visualizations and criticizing them. I know exactly what to look for in a visualization that might mean it is misleading in some way. I was always a little skeptical of graphs with a lot of chart junk and I'm glad that we prioritized making clean and clutter-free graphs. I do feel as though now I will try and be more careful in examining the visualizations featured around us instead of simply trusting that they are accurate. Also it was great to work with both Tableau and R. Tableau was fairly straightforward but still had a small learning curve to figure out certain things. I'm glad I was able to learn R as it seems like it's a more powerful tool in terms of the different kinds of graphs it can create. This class was a great experience.

Code:

Big Studio Profit Distribution Visualization (Rich Finley):

```
BoxOffice$Profit <- (BoxOffice$domestic_revenue - BoxOffice$budget)

studio_means <-BoxOffice %>%
  group_by(distributor) %>%
  summarise(Mean_Dom_Rev = mean(domestic_revenue),
            Mean_Budget = mean(budget),
            Mean_Profit = (Mean_Dom_Rev - Mean_Budget),
            Mean_Open_Rev = mean(opening_revenue),
            Open_Percentage_of_Total = percent(Mean_Open_Rev/Mean_Dom_Rev),
            Total_Rev = sum(domestic_revenue),
            Total_Budget = sum(budget),
            Total_Profit = sum(Profit),
            Num_Movies = count(distributor))

studio_means_fitler_17 <- studio_means %>%
  filter(Num_Movies > 16)
studio_means_fitler_17

big_studios <- studio_means_fitler_17["distributor"]
big_studios

BO_BS <- BoxOffice %>%
  filter(distributor %in% unlist(big_studios, use.names=FALSE))
BO_BS

in_mil <- c("-$200M", "-$100M", "$0", "$100M", "$200M", "$300M", "$400M", "$500M",
"$600M")

profit_dist_graph_big_studio <- ggplot(data=BO_BS, aes(x=reorder(distributor, -Profit),
y=Profit)) +
  ggtitle("Big Studio Profit Distribution")+
  theme(plot.title = element_text(hjust = 0.5))+
  geom_violin() + geom_jitter() +
geom_text(aes(label=ifelse(Profit>500000000,title,''),hjust=-.2, vjust=0))+
  scale_color_gradient(low="light green",high="blue",
                       limits = c(-30618899, 128091034),
```

```
                breaks = c(-30618898, 0, 30000000, 60000000, 90000000, 128091033),
                labels = c("-$30,618,898", "$0", "$30M", "$60M", "$90M", "$128,091,033"),
                name = "Mean Profit")+
  scale_x_discrete(labels = function(x) str_wrap(x, width = 10))+
  scale_y_continuous(breaks=seq(-200000000, 600000000, 100000000), labels = in_mil)+
  xlab("Studio")+
  geom_point(data=studio_means_fitler_17,
          aes(x=distributor,y=Mean_Profit, color=Mean_Profit),
          size=3)
profit_dist_graph_big_studio
```

How long different budget movies play in theaters (Lanny Xu):

```
na.omit(boxoffice2017_2019)
library(dplyr)
library(ggplot2)
library(gridExtra)
library(grid)

boxoffice = boxoffice2017_2019 %>% select(budget, release_days)

boxoffice$Budget <- as.numeric(gsub("^\\s*[(]", "-", gsub("[$),]", "",
boxoffice$budget)))
boxoffice$ReleaseDays <- as.numeric(as.character(boxoffice$release_day
s))
boxoffice = na.omit(boxoffice)

plt <- ggplot(boxoffice, aes(x=log(Budget), y=ReleaseDays))

plt + stat_bin2d(aes(fill=..density..))+theme_grey()+
    ggtitle("2D Density Plot of How long different budget movies play i
n theaters") +
    xlab("Budgets") + ylab("Days in theaters")

plt +
  stat_density2d(aes(colour=..level..)) +  geom_smooth(method=lm)+
  geom_point()+theme_bw()

plt +
  stat_density2d(aes(fill=..density..),
                 geom="raster",
                 contour=FALSE)+  geom_smooth(method=lm)
plt + geom_smooth(method=lm)

plt + geom_smooth(method=lm)+ geom_point(col="grey")+
  ggtitle("Smoothed Distribution Plot") +
  xlab("Budgets") + ylab("Days in theaters")

p1 <- plt + geom_smooth(method=lm)+geom_point(size = 0.5)+
  ggtitle("Smoothed Distribution Plot") +
  xlab("Budgets") + ylab("Days in theaters")

p2<- plt + stat_bin2d(aes(fill=..density..))+theme_grey()+
  ggtitle("2D Density Plot") +
  xlab("Budgets") + ylab("Days in theaters")

tg <- textGrob('How long different budget movies play in theaters', gp
= gpar(fontsize = 16, fontface = 'bold'))
grid.arrange(p1,p2,nrow=2, ncol=2,top = tg)
```

Domestic/ World Revenue by Studio (Zhong Xie):

```
library(ggplot2)
library(dplyr)
library(magrittr)
library(hexbin)

bo <- boxoffice2017_2019
bo <- bo %>% mutate(domestic_revenue.log = log(domestic_revenue),
            world_revenue.log = log(world_revenue))

# hex bin
ggplot(bo, aes(domestic_revenue.log, world_revenue.log)) +
  geom_hex() +
  ggtitle('Domestic vs World Revenue') +
  ylab('World Revenue') + xlab('Domestic Revenue') + labs(fill = 'Films') +
  theme(legend.position = c(.90, .20))


#scatter
ggplot(bo, aes(domestic_revenue.log, world_revenue.log, color=studio_size)) +
  geom_point() +
  ggtitle('Domestic vs World Revenue') +
  ylab('World Revenue') + xlab('Domestic Revenue') +
  labs(color = 'Studio Size') +
  theme(legend.position = c(.90, .15),legend.key = element_rect(colour = "black"))
```