**ISM 6562 Data Visualization**

**Group Project**

**Group 5**

Movie Data Analysis: Studying the factors of success and other meaningful patterns in the world of commercial cinema

April 19, 2022

Nico Mora

Narottam Jajodia

Sathvik Ganta

Sage Ramsammy

Denise Collins

**1. Abstract**

The International Movie Database (IMDb) is an open source platform which contains data on all movies released globally. The information available includes Title, Runtime, Censor rating, User score, Cast & Crew, Budget, Revenue, Technical specs, Trivia, etc. As millennials, our team was naturally passionate about movies. Also, as data analysis students, this massive data was attractive to us. The power of data analysis is evident by how corporations and governments use people’s data to create high-level strategies. We wanted to use the movie data and perform visual analytics to find interesting patterns, correlations, potential causal relationships, etc. that could be easily missed if one looks only at the raw data. Data, within itself, often hides immense information and insight. Successful data analysis is that process which brings out meaningful value from raw data, just like a sculptor brings out a beautiful image from an ugly rock. The insights we may find, could be leveraged by production companies, media bodies, and researchers for various purposes.

**2. Introduction**

**2.1 Project Description**

Through Tableau visualization, our project aims to answer multiple questions on factors that relate to movies and their production. Questions regarding which factors determine a movie’s success, how genres change over time, movie release timing, and the country’s production of movies and profit can be answered. Analyzing movies from within IMDB’s database from 1980 to 2020 trends and relationships can then be seen. The project consists of 10 worksheets, 2 dashboards, and 1 story, to see these relationships accordingly. Through these visualizations, studios, bloggers, and researchers can derive further information and stories to understand how to create a more profitable sustainable movie industry.

**2.2 Project Purpose**

Each year hundreds of movies are made by countries all around the world. The first long-form feature film, *The Story of the Kelly Gang,* was created in 1906, and since then movies like *Gone with the Wind, Avatar,* and *Three Idiots,* have captivated the world. With so many movies, directors, genres, and budgets, this project set out to find trends and information through IMDB information. Through visualizing this data, people can see how different variables affect the profit, rating, and release dates. Understanding these variables becomes pivotal for production studios to appropriately craft and release their stories to the world.

**3. Data Description and Previous Work**

We researched the movie data from a free source website (Kaggle): <https://www.kaggle.com/datasets/danielgrijalvas/movies>

This dataset has been extracted by the Kaggle user from the renowned global movie database platform: **IMDb**

The dataset contains the following variables: name of the movie, censor rating, genre, year released, date released, IMDb user rating, number of votes, director, writer, star, country, budget, gross revenue, production company, and runtime.

We believe that this data is very useful for our intended project objective. The dataset has all the right variables we need to analyze themes like - factors of success, time series analysis, release trends, and other patterns. Certain visual analysis that we intended to perform, required the use of profits made by movies. We used Microsoft Excel to create a new variable – profit. This was calculated by deducting the movie’s budget from its gross revenue.

**4. Design Process**

**4.1 Design steps and interpretation**

**Question 1**

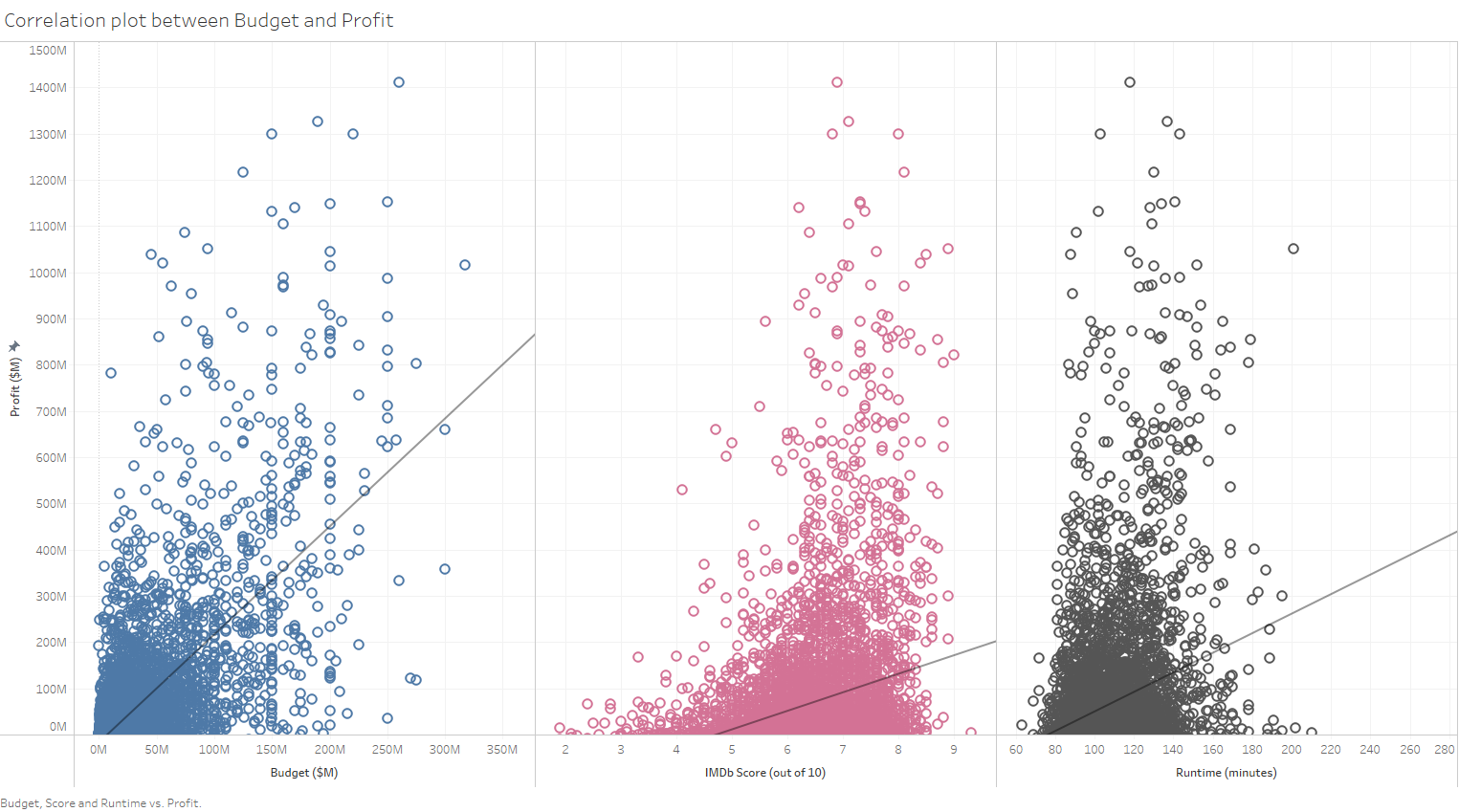
Which factors determine a movie’s success?

Visualization 1 : *‘Study Finds Big-Budget Movies Make Greater Profits’ (The Wrap, 2009).*

We begin with studying the relationship between profits made by a movie and factors like: Budget, IMDb rating, and Runtime. The scatter plots may help us identify if any of these factors has a significant relationship with Profit. This relationship could also be causal in nature. This analysis could help producers focus on certain aspects to ensure maximum returns.

Steps: We added Profit under ‘Rows’ and Budget, Score and Runtime under ‘Columns’. This automatically yielded three scatter plots. We then added a trend line from the ‘Analysis’ tab. Finally we changed the colors of the scatter plot, to differentiate between the 3 X-axis variables, from the ‘Color’ button under the ‘Marks’ section. We also edited the X and Y axis labels to be more descriptive.

**Visualization 1**

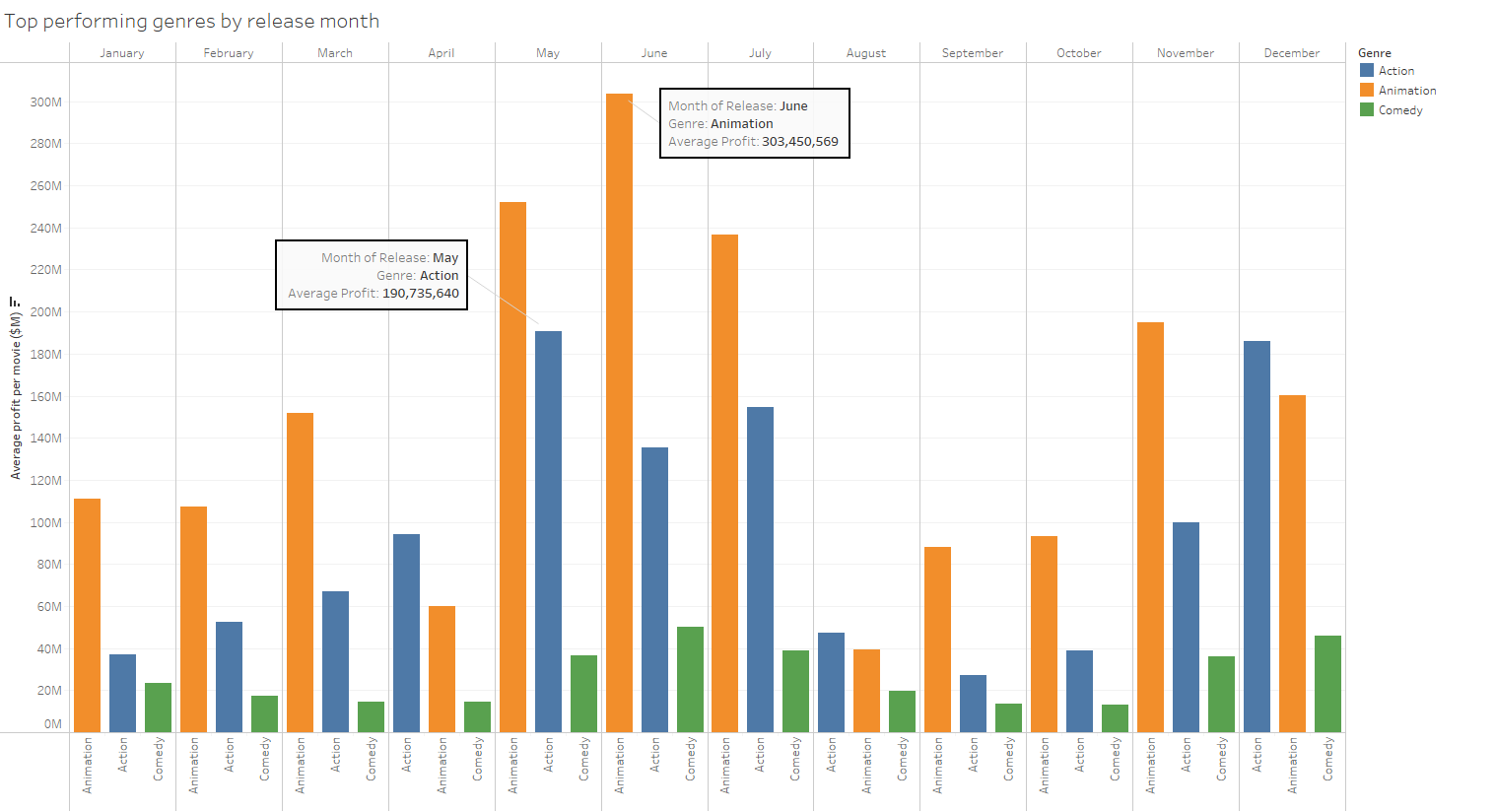


Interpretation: Budget seems to have the strongest positive correlation with Profits, as compared with other two plots. However, the concentrated distribution near the origin and the sparse distribution at the top tells us that most movies, upto $100M budget, made profits under $200M. Movies that earned very high profits (over $300M) range from budgets $5M to $317M. This means that factors, other than budget, also impact a film’s profitability. Speaking about the other 2 scatter plots, the highest profits fall between IMDb score of 6-9 and runtime of 80-160 mins. Though not very conclusive, we can tell that people may prefer watching movies with a high IMDb score and a certain runtime range.

Visualization 2 : We, next, look at the impact that certain genres have on profits, based on their release month. This may help us understand if certain genres attract higher profits if released in certain months of the year.

Steps: We created a set of top 3 genres by sum of profits and added Released date (selected only ‘month’ from drop-down), Genre Set, and Genre under ‘Columns’. We created a calculated field to calculate average profit by count of movies, and added this field under ‘Rows’. This automatically yielded a bar plot. We then changed the color palette to differentiate between the 3 genres and also edited the Y axis label to be more descriptive. Finally, we added annotations for the top 2 genres.

**Visualization 2**

****

Interpretation: *Adults like animation too. (The Hindu, 2015).* Animation movies seem to make the most profits throughout the year. This could be because animation movies generally require low budget, but attract children & adults equally. However, the animation genre is most prominent in May-July and least in August. Action movies make the most profits in May and December.

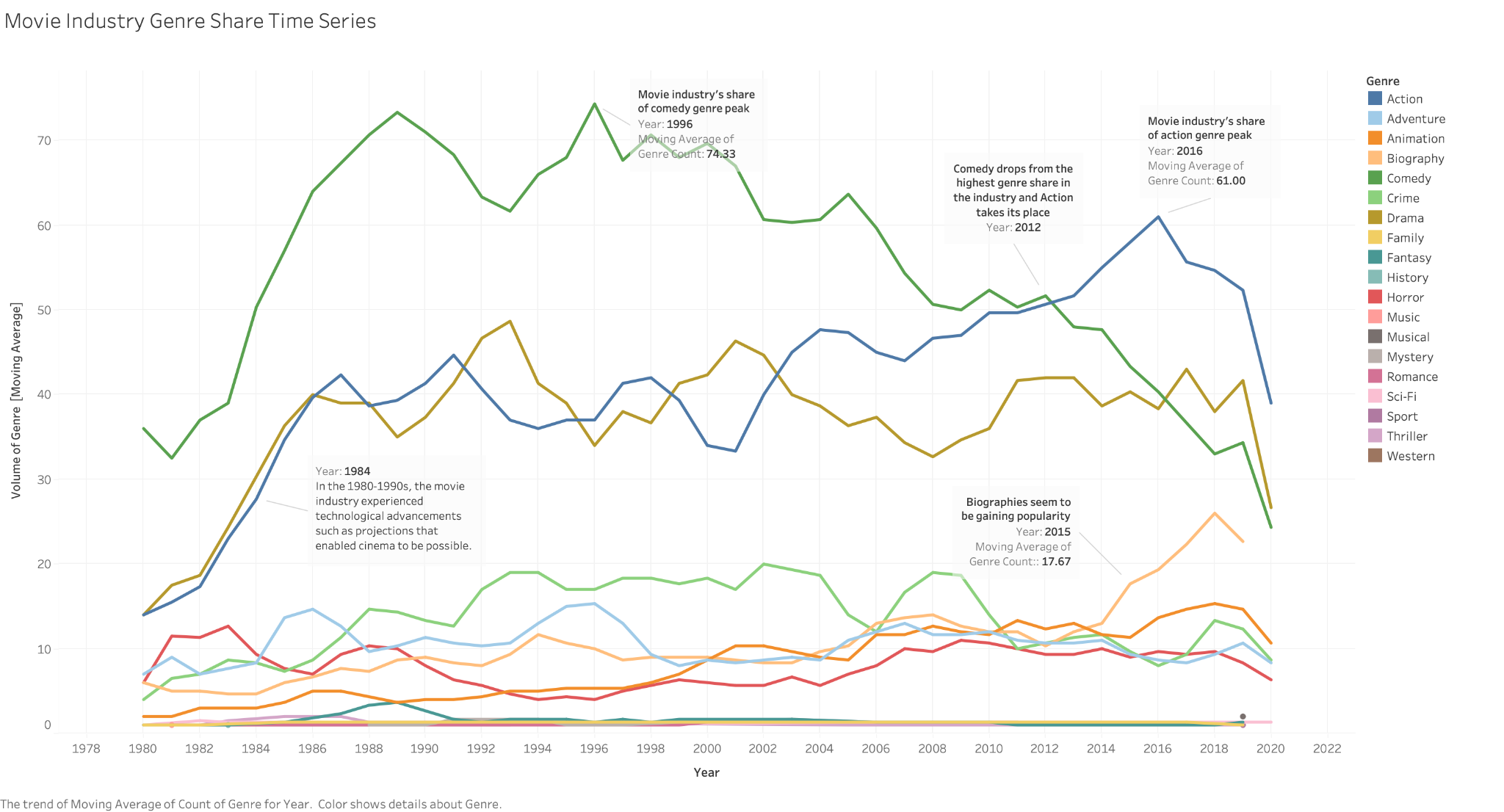
**Question 2**

How does the share of genres/ratings change over time?

Visualization 1 : In order to understand and visualize the history of the movie industry, we decided to create separate time series for genres and ratings. First, we will explain the design process of the Movie Industry Genre Share Times Series.

Steps: In order to create this time series, we put the variable Genre into Rows and the variable Year into Columns and selected a continuous line chart. In order to plot every Genre, we dragged the Genre variable to Marks > Color. We then clicked on Rows > Genre > Measure > Count in order to get the volume of ratings per year. In order to smooth our time series, we clicked on Rows > Genre > Quick Table Calculation > Moving Average. To clean up our chart, we edited the Title to “Movie Industry Genre Share Time Series” and used Filters to filter out Null Genres in our chart. Last, to point out key findings in our visualization, we right-clicked on different points on our chart > Annotate > Mark to write out explanations. In conclusion,

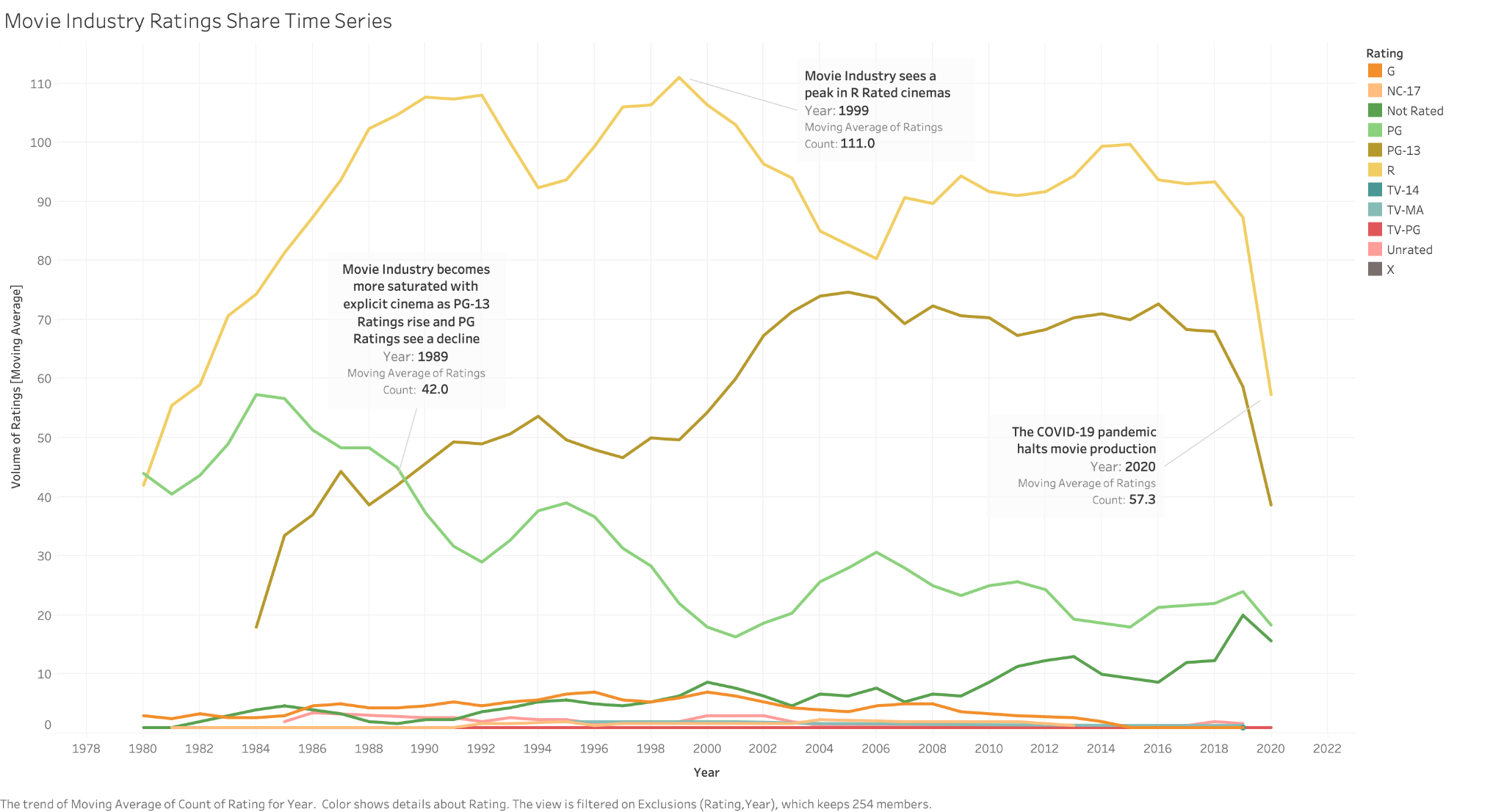
**Visualization 1**



Interpretation: This chart begins by showing us rapid growth in the movie industry during the 1980s through the 1990s due to technological advancements such as movement dissection and film cameras. Moving towards 2012, we see the top Genre for the movie industry change from Comedy to Action. Last, in 2014 we also saw a spike in production for Biographies within the movie industry.

Visualization 2: Once we created the time series for Genres, our next task was to re-create this time series but for Ratings instead.

**Visualization 2**



Steps: To start, we dragged the variable Ratings into Rows and Year into Columns and selected a continuous line chart. We aimed to plot every Rating as the chart before, so we dragged the Ratings variable to Marks > Color. We then clicked on Rows > Genre > Measure > Count in order to get the volume of Genres per Year. Now that we had a rough idea of what the chart looked like, we wanted to smooth our times series like before, so we clicked on Rows > Ratings > Quick Table Calculation > Moving Average to get a better line for each Rating continuously for each Year. To create a title, we right-clicked the automated title > Edit Title and changed it to “Movie Industry Ratings Share Time Series”. We used Filters to filter out Null Ratings to get rid of these values on our chart. Last, as before we wanted to point out key findings so we right-clicked on various points on our chart > Annotate > Mark and wrote out explanations.

Interpretation: Starting in 1989 through 2004, the movie industry became saturated with PG-13 movies as high-grossing cinemas like *Batman Returns* and *Titanic* released during this time. This chart also shows that Rated R movies have led the movie industry since the 1980s. Last, we can visualize the COVID-19 pandemic lockdowns lowering the count of movies in 2020.

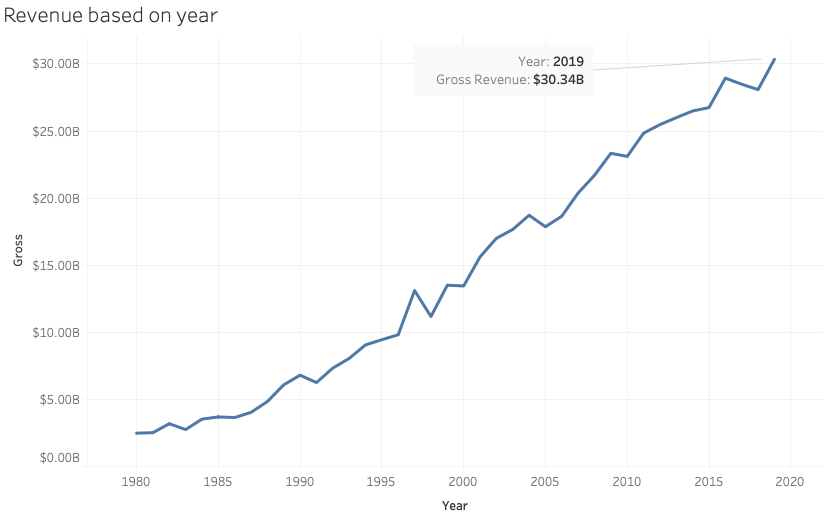
**Question 3**

Do movie releases cluster within certain periods of a year?

Visualization 1: To fully understand whether movies cluster, we first decided to see the data as a whole. Through this, we could detect any consistent markings over time. A line graph depicting the revenue based on the year released was first implemented.

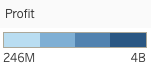
Steps: First, the variable Year will be dragged into Columns while making sure it is a continuous dimension. Keeping the variable as a whole number rather than a date data type makes it easier to filter out any year after 2019 (to avoid any null value). The SUM(Gross) will then be dragged into the rows. The y-axis should have the numbers labeled as a currency, so we right clicked SUM(GROSS) in the rows > format > Axis > Numbers to change this. After clicking on the highest point > Annotate > Mark to emphasize this datapoint. A title was then added, by right clicking the default title > Edit Title and renamed into “Revenue based on year”.

**Visualization 1**

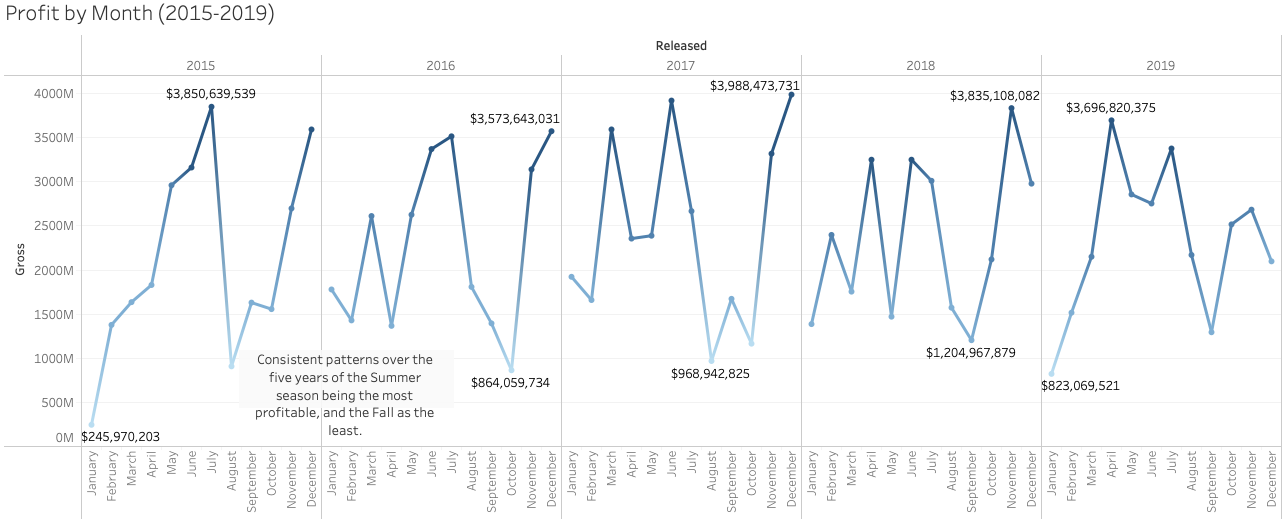


Interpretation: The revenue for the movie industry is increasing steadily. However, it was evident that there were no visible patterns over the time frame of the data aside from this. The following step was to create a visualization focused on a smaller set of years in order to see seasonal patterns in where movies have been released.

Visualization 2: This visualization focuses on years 2015-2019. The years 2020 and 2021 were filtered out due to the amount of nulls. Here, we will be able to see the most popular months and seasons. We chose to look at the data by month to see granular trends.

Steps: After dragging SUM(GROSS) to rows, “Released” must be converted into a date data type and dragged into columns. We’ll only expand to YEAR(Released) and MONTH(Released). A filter is placed onto the year to avoid any null values by right clicking YEAR(Released) > Edit Filter > Exclude 2020. SUM(Gross) is then added on the Marks pane into Color. The color will now appear on the top right segment of the visualization as a legend. To better differentiate the color differences on the graph we clicked the upside down triangle on the legend > Edit Colors > Automatic > Stepped Color 4 Steps. We now want to show the mins and max points of each year, so we click the Label in the Marks pane. Then we click Show mark labels > Min/Max > Pane > Allow labels to overlap other marks. A text box describing the seasonal patterns in the five years was made by clicking an empty space in the graph > Annotate > Area. A title was then added, by right clicking the default title > Edit Title.

**Visualization 2**



Interpretation: This visualization was a better choice in order to see if movie releases cluster during certain periods of the year. The seasons most movies are released are the Summer and Winter, and the least amount being Fall. Profitability seasons vary by year, with all five years having the Summer or Winter season as their most profitable and the Fall as the least.

**Question 4:**

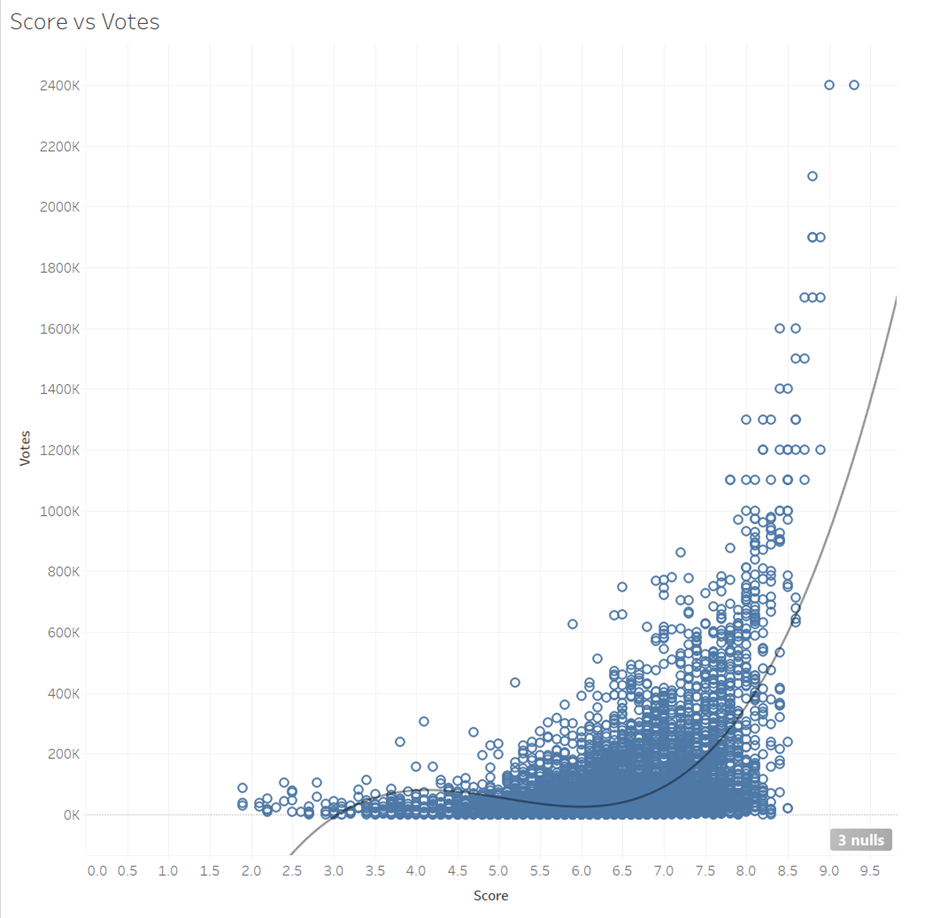
Company / Director relevancy in the movie industry based on scores?

Visualization 1**:** In this questionwe will be observing how a company can choose a director and vice versa based on the IMDB scores they receive. Firstly, in order to understand the relation between them we need to understand how the scoring occurs based on the number of people voting for a movie. To understand score vs votes relation, we plot score and votes (in the plot below).

Steps**:** This was done by dragging the score to columns and then we drag Votes to rows. We use the show me function and choose a scatter plot. The plot shows until score 5 is linear and after 5 we can observe a polynomial relation. To show a clear relation we added a trend line (Analytics>Trend Line> Polynomial) as seen in the figure below.

Interpretation**:** From this plot we can see that the scores between 1.5 and 7.5 received almost less than half a million votes and from the trend line we can observe that as the number of votes increases in the later half, the score also increases.

**Visualization 1**

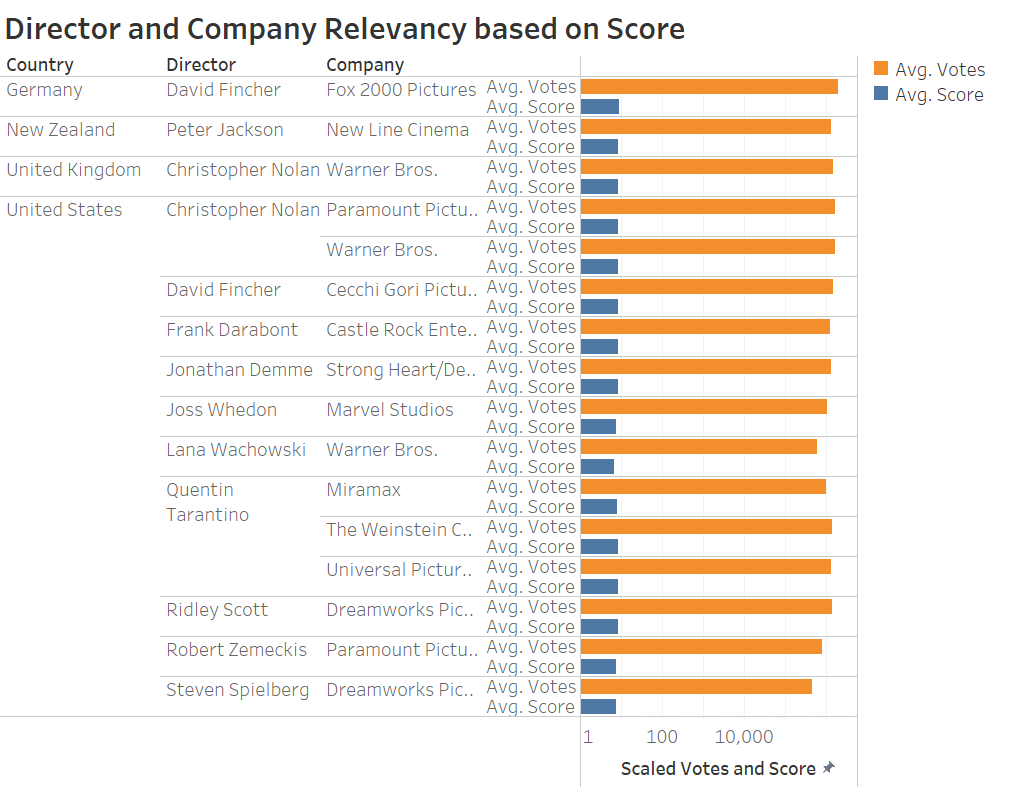


Visualization 2**:** Once we understand the relation between score and votes our next task was to see how it would be relevant to the director and industry. In this visualization we focus on how a director and a production company are relevant based on their votes and score.

Steps:For this we started with making a measure values field adding the average of score and votes as a director can make more than one movie with the same production company. Later we drag this to the column and then we drag director and company as we aim to understand the variation of the rating between the director and production company combination. In the show me section, we choose the side by side bar to represent the data. Later, we observed the combination receiving the highest votes was 2,400,000.

To minimize the effects of temporal bias, we look at the distributions for sets of films within maximum half of the highest votes received. We set a filter for the director and company combination receiving at least half of the votes by dragging votes to filter > votes> measure> maximum> at least> changing the range. This filter’s out the plot and then we use a similar method to select movies with score more than 7, by dragging score to filter > score> measure> maximum> at least> changing the range. To distinguish between the two side by side bars we drag the measure values to mark>color>select color>edit colors, in the select data item assign the color to score and votes. We study the logarithm of these quantities because their values span several orders of magnitude. The next step would be dragging countries to the row to understand which countries have the most relevant combination.

**Visualization 2**

****

Interpretation: This visualization shows us the relevant directors and company combination by country. As Hollywood is the largest film industry (The Hollywood Report, 2020) we see that the combination of a director Frank Darabont backed by Castle Rock Entertainment and Christopher Nolan as director backed by Warner Brothers have the highest votes and the highest rating. Some examples of their combinations are Batman trilogy, Inception, Dunkirk, Interstellar, and the Shawshank Redemption which were some of the highest rated movies (The Washington Post, 2019). On the other side we see as the votes reduce the score also goes down.

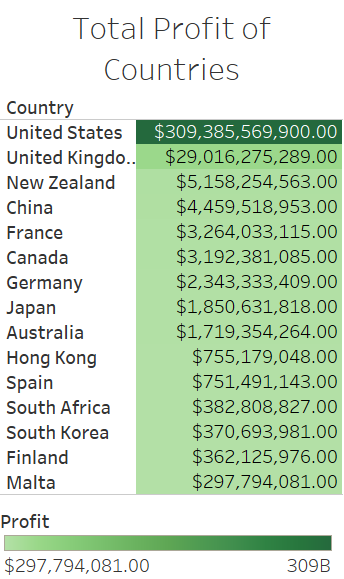
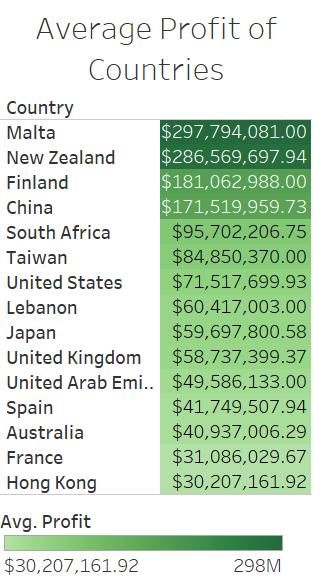
**Question 5**

Which country has the greatest profits?

Visualization 1 and 2

Description: The primary purpose of the first two visualizations is to visualize the average profits and total profits of the top fifteen countries of the dataset. Understanding this information provides a better context for which countries have the largest film industries. Tables can be the most effective visualization to display certain data, and this was chosen as the best option for the data provided.

Visualization 1 Visualization 2

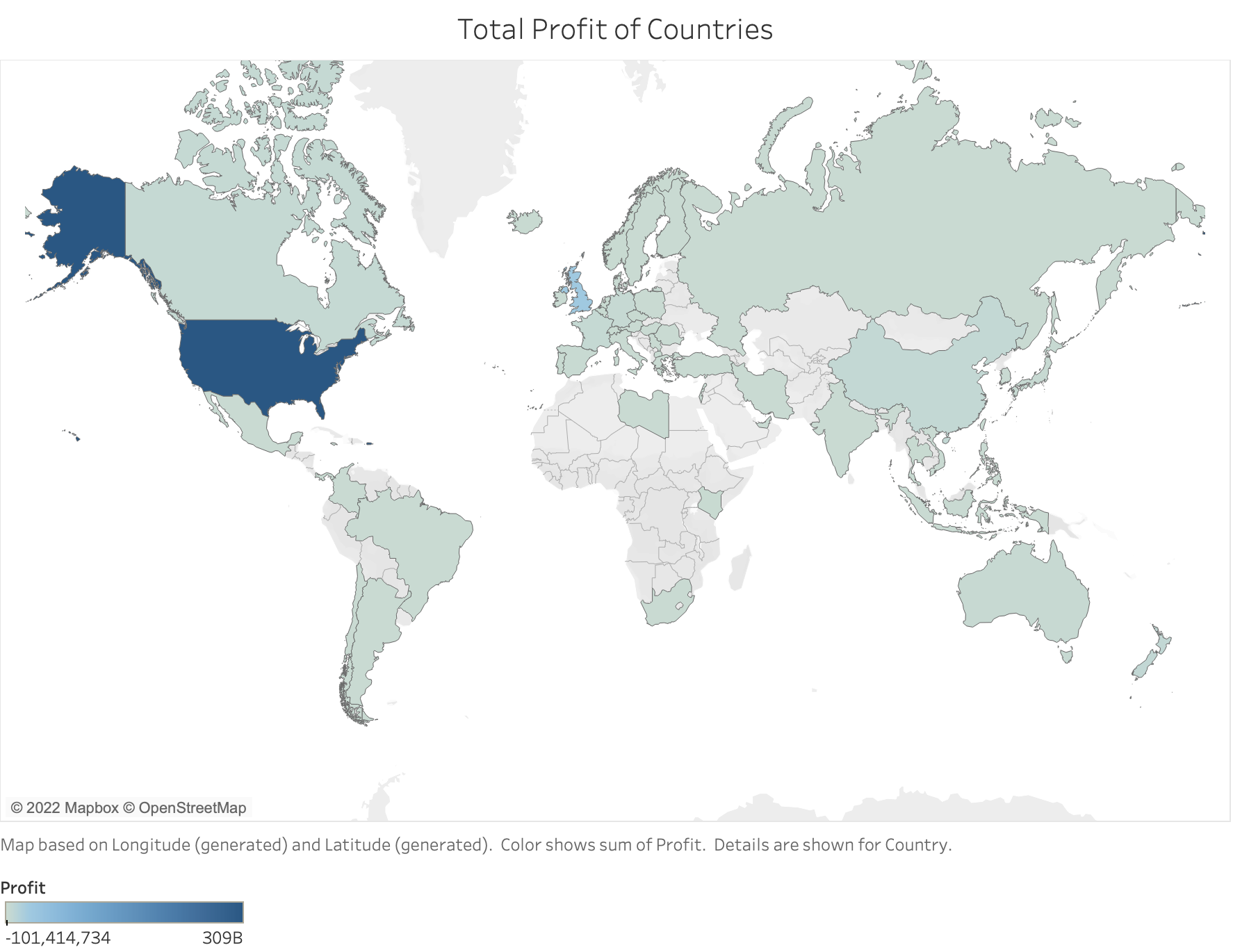
 

Steps: The country variable was dragged into the rows column and the highlight table was chosen as the visualization to see the rows of information with colors to see the scale. For visualization 1 the profit was dragged to both the color and label of the marks. For both the color and label, the sum measure must be selected. From there the information was placed in descending order by clicking the option at the top of the screen. Next the information was filtered by the top fifteen countries by dropping the country into the filter area and editing it to display the top fifteen. After filtering the information, this concludes the process for visualization 1. As for visualization 2, the process remains almost identical except for the change in the profit variable. Instead of choosing the sum measure like the first visualization, the average measure is chosen. Apart from this one step, the process remains the same.

Interpretation: The two visualizations display slightly different results as evident in the tables. The total profits of countries within visualization 1 are led by the United States by a huge margin. There is a very large difference between the total profits of the United States and the next country. It is interesting to see that the top five countries for total profit are on four different continents. For visualization 2 it is interesting to see that the list changes a lot. The United States drops to the sixth rank and Malta takes the top result. China and New Zealand both remain in the top five rankings the countries. These insights are helpful to see which countries are leading in the movie industry.

Visualization 3:

Description: The visualization aims to display the profits of countries around the world and see which areas have the greatest concentration of total profits. A global heat map was chosen to see where in the world, the movie industries are most profitable.

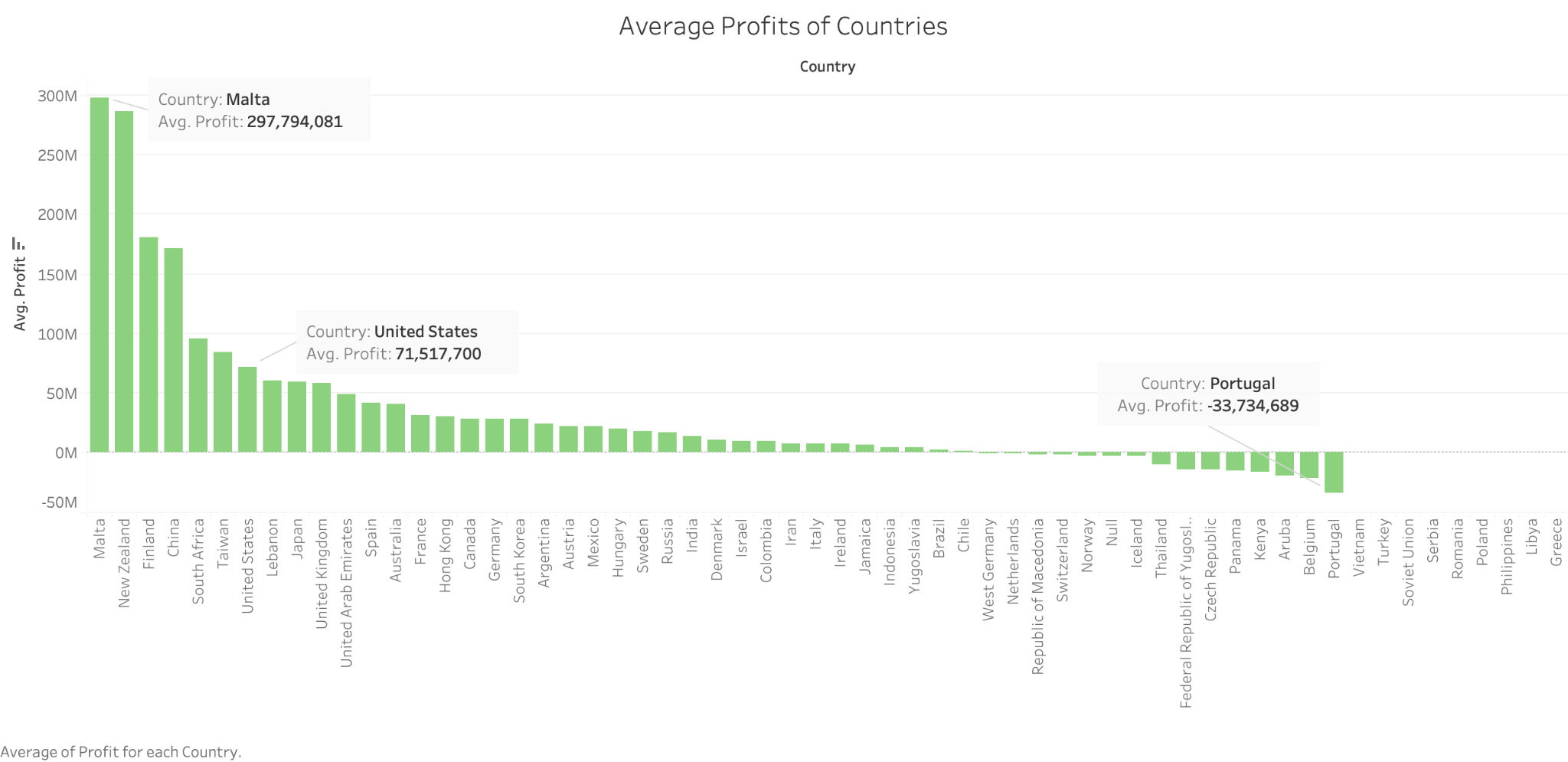


Steps: The process of the visualization is somewhat similar to the first two visualizations with the data usage. The country variable was dragged into the rows column and the global heat map was chosen as the visualization to see the regions of information with colors to see the scale. For visualization 3 the profit was dragged to the color marks with the sum measure being selected. The green-blue diverging was selected from the color marks, seen below, to show distinction in scale between countries.

Interpretation: From the global map it is evident that just like the first table, the United States is the most profitable. The visualization also displays that most of Africa has either little profit or no data available. Further interaction with this visualization would zoom into the region of Europe to see the profits of the United Kingdom in comparison to other countries. This visualization is helpful in showing how large of a difference the United States is in comparison to other countries.

Visualization 4

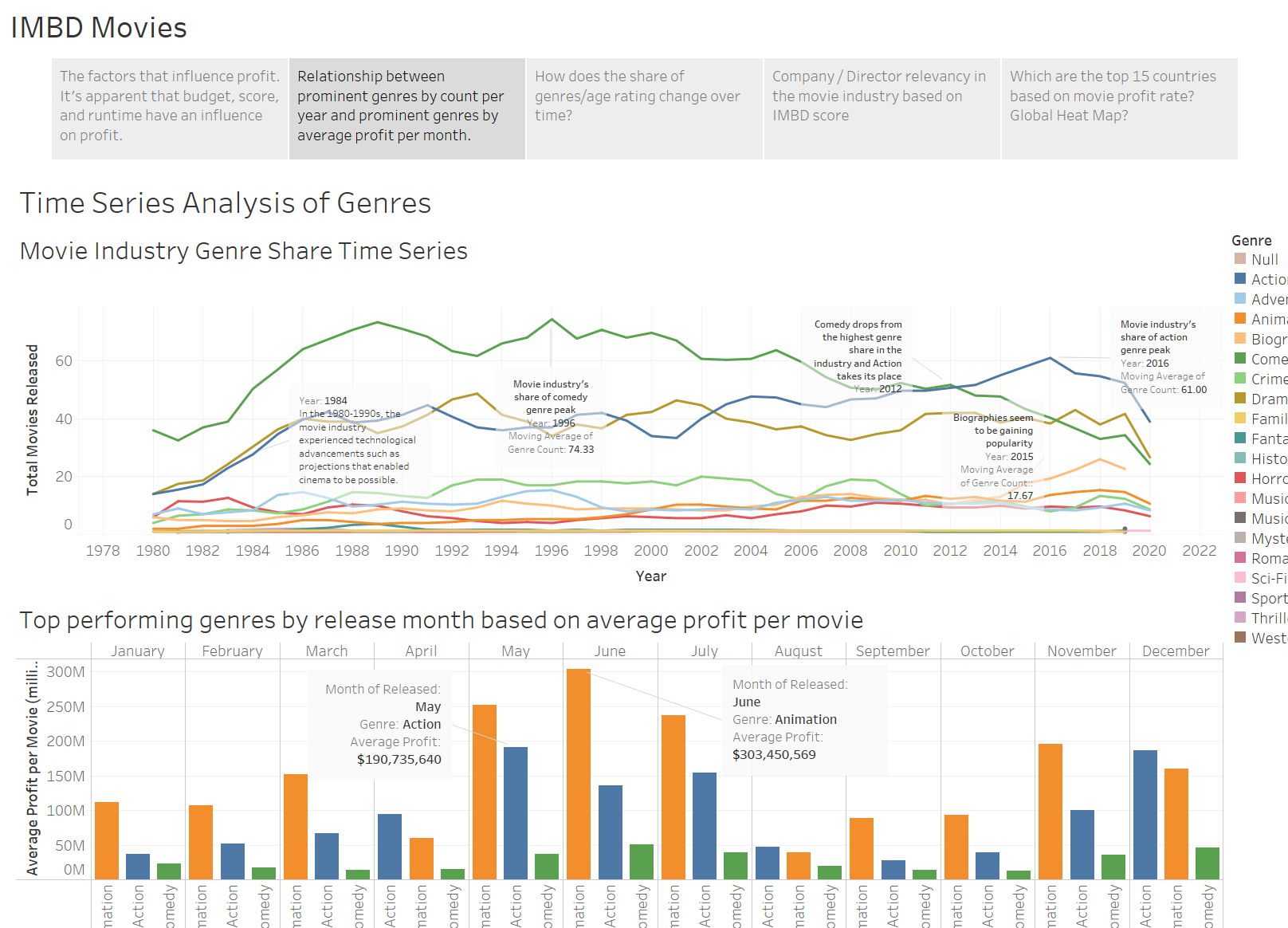
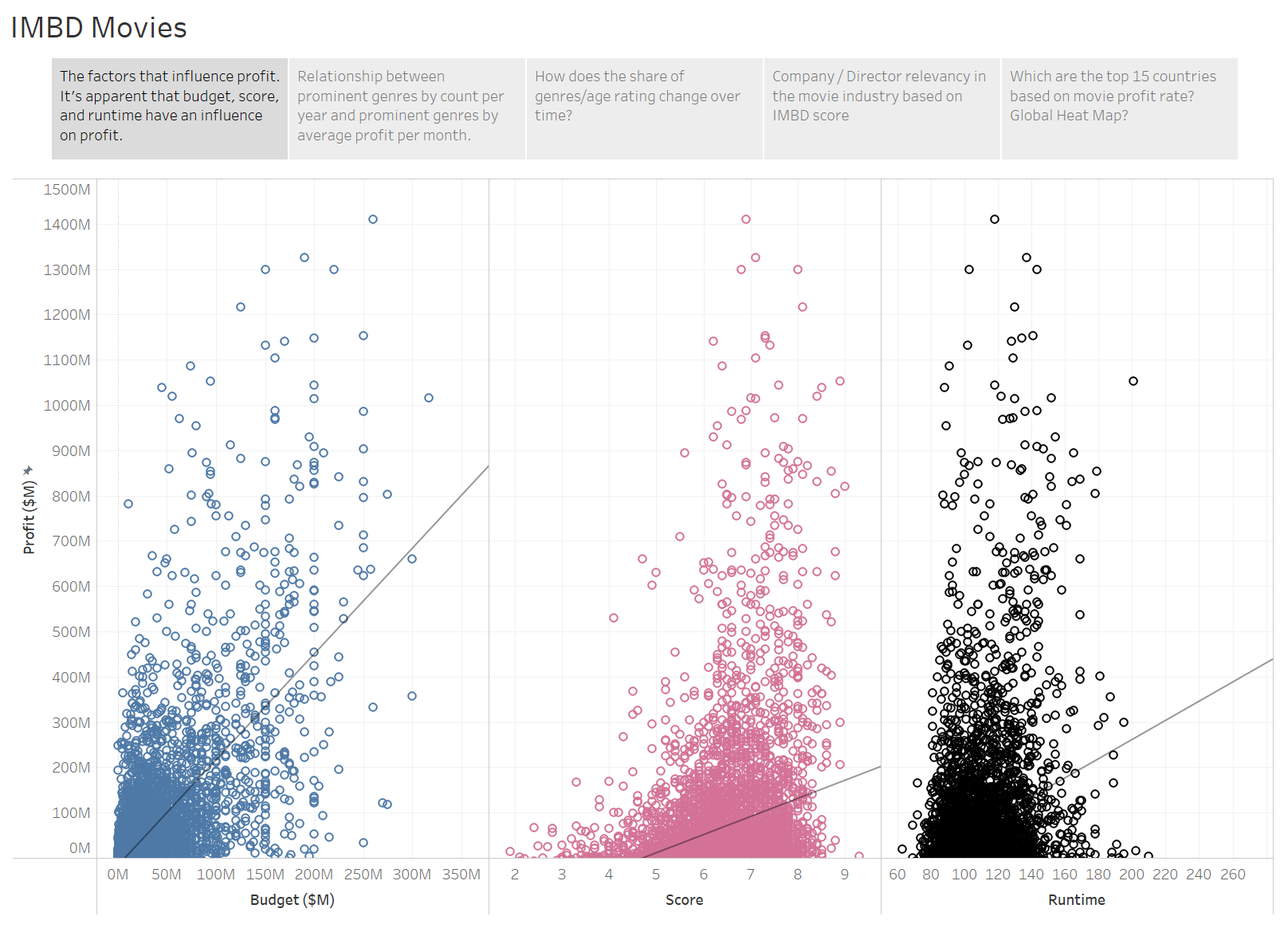
Description: The aim of this visualization is to see the difference in the average profits in all of the countries compared to one another. A bar graph was chosen to most easily compare categorical and continuous variables.

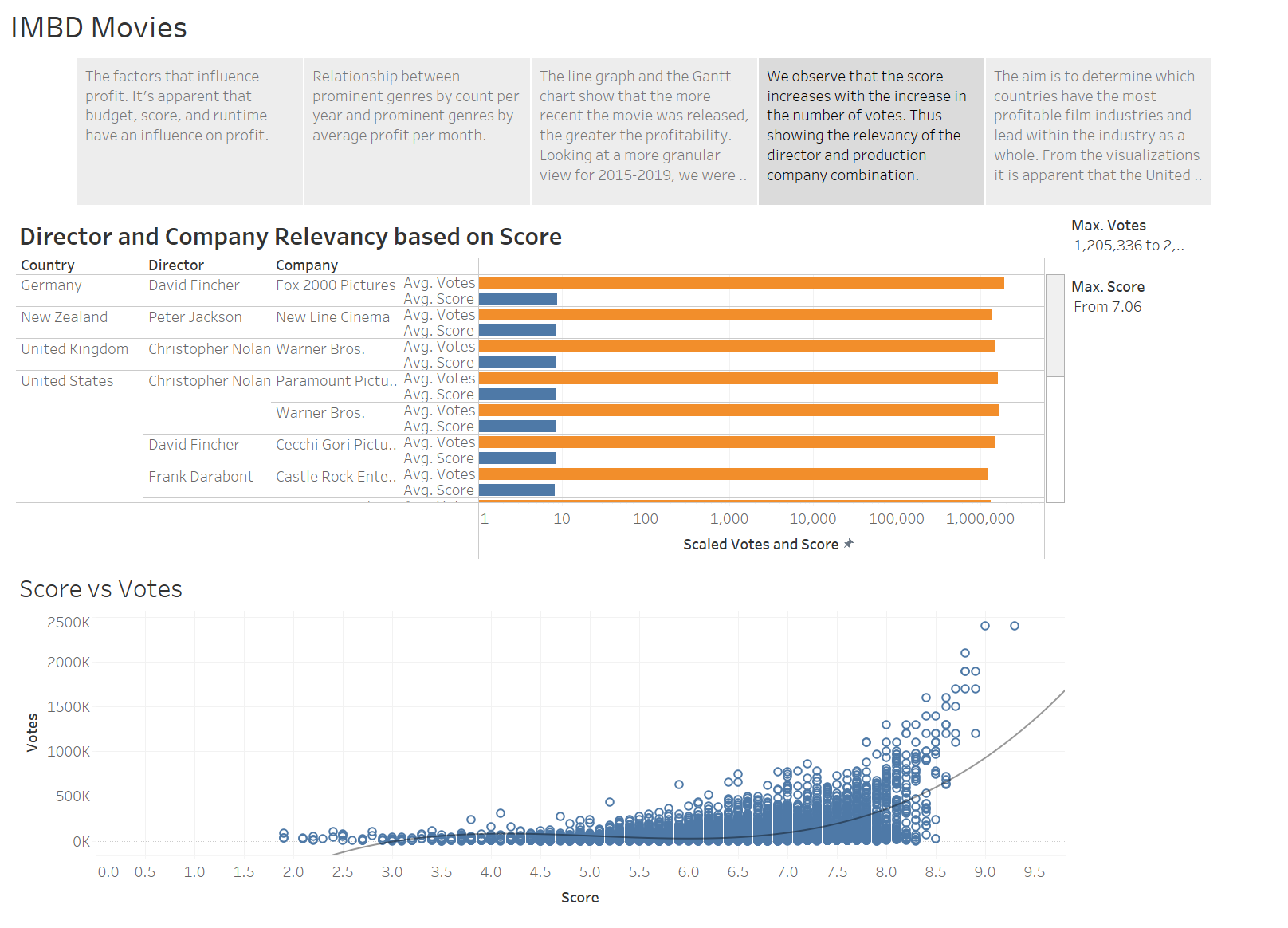
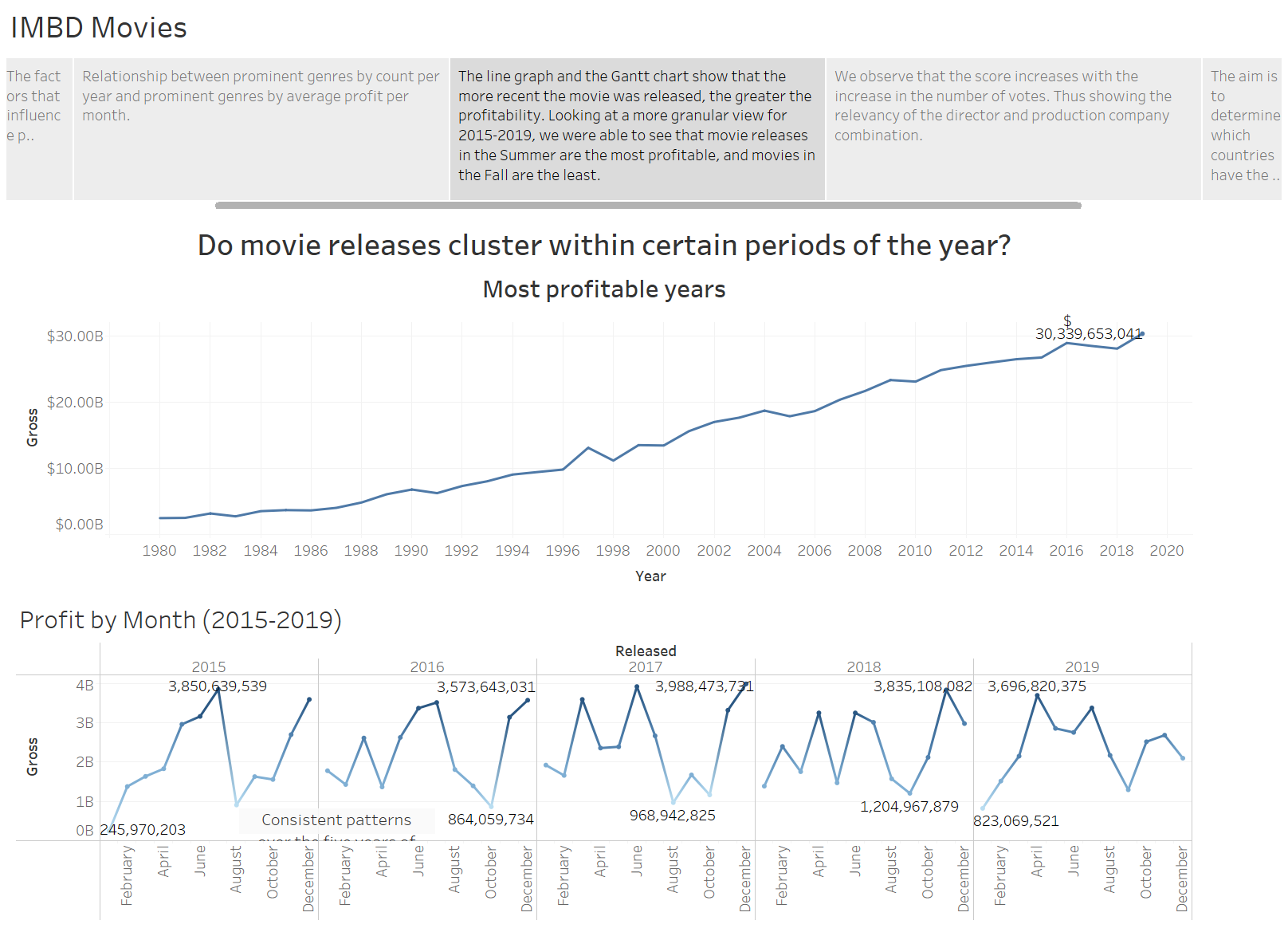


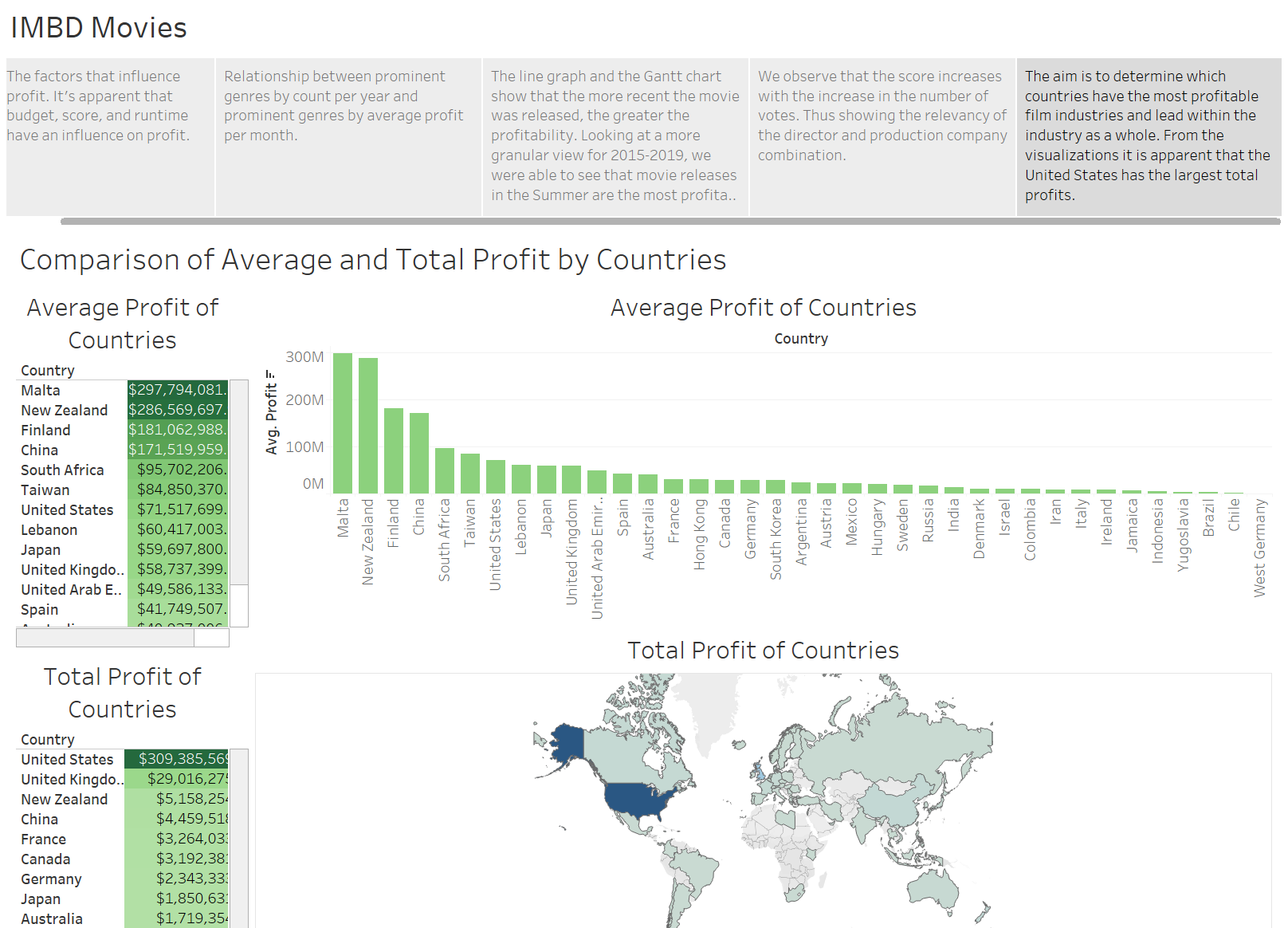
Steps: The process of the visualization is somewhat similar to the first three visualizations with the data usage. The country variable was dragged into the rows column and the bar graph was chosen as the visualization to compare the countries alongside one another. For visualization 4 the profit was dragged to the color marks with the average measure being selected. The variables were then sorted by descending value and green was chosen as the color within the color marks edit. Annotations were added to display the information for Malta, the United States, and Portugal.

Interpretation: The visualization displays that Malta has the highest average profit, while Portugal has a negative average profit. Many of the countries in the data set have less than 50 million dollars in average profit. It is interesting to note that while the United States has the greatest total profit, it is ranked seventh, most likely do to a large total profit, but a great count in the number of movies made.

**4.2 Storyboard**

****

****

****

We created 2 dashboards by selecting ‘New Dashboard’ from the Dashboard tab. For dashboard 1, we added 2 graphs one below the other - ‘Movie Industry Genre Share Time Series’ and ‘Top performing genres by release month based on average profit per movie’. This shows us a relationship between prominent genres by count per year and prominent genres by average profit per month.

For Dashboard 2, we placed 2 charts one below the other - ‘Director and Company Relevancy based on Score’ and ‘Score vs Votes’. This shows the connection between the score-votes relationship and top performing director-company pairs. For example, a movie could attract high scores when its directed by Steven Spielberg and produced by Dreamworks Pictures.

**4.3 Key findings**

1. Production investment for a movie does seem to have some impact on profits. However, higher spendings alone does not promise returns. Movies with an IMDb user rating between 6 and 9 seem to be the most profitable. Also, a movie runtime of 80-160 mins apparently attracts most viewership.
2. Animation is the top genre based on highest average profits per movie. The genre is most prominent in May-July, which could be due to summer vacations at educational institutions. Action and Comedy are the next two top genres by the same measure.
3. We discovered the technological boom in the movie industry between the 1980s and 1990s. There was a rapid growth in the movie industry due to advancements in movement dissection, film cameras, and many more tools.
4. Starting in 1989 through 2004, the movie industry became saturated with PG-13 movies as high-grossing cinemas like *Batman Returns* and *Titanic* released during this time. Movies such as this influenced many more production companies to follow similar narratives.
5. Comparing revenue based on year released is not the best indicator to tell whether or not movies cluster during certain periods. It’s evident that movies that have released more movies have a higher profit. The yearly rate of movies being released has also increased since 1980.
6. After looking at the common patterns within 2015-2019, the Summer and Winter are the seasons with the most movies released, also making them the most profitable. Meanwhile 2019 is the year with the highest overall profit in the line graph, 2017 had the highest number of movies released and also had the highest earning point in the five years.
7. We find that directors and production companies in the USA receive more votes than films do in the n on-US countries. These findings suggest that there is both a language bias and a temporal bias in the distribution of user votes in IMDb. This tells us that the movies released in the country of the US have the most popularity compared to movies released in non-US groups.
8. After analyzing the visualizations it is apparent that the United States had the largest film industry of all the countries for the total profit. It is surprising that Malta has the highest average profit for the industry, however, this is mainly due to only one movie from the country. It is insightful to see that for the total profits of countries there is a great spread between the regions of where these countries are located.

**5. Limitations and Further Work**

1. Data Limitations: The dataset utilized displayed elements of data bias towards certain countries and did not display movies from every country. For example, India should display a much larger film industry then the data revealed. IMDB as well as the dataset require further efforts to reduce further bias within the data analysis. Additionally, 2020 only contained partial data and lacked all of the movies from the year.
2. Further Work: IMDB could be webscraped for more variables and detailed analysis to look at new trends and influences of the film industry. For instance, attempting to see the impact of COVID on the film industry with a larger dataset including 2020-2021. It would be insightful to continue at a greater depth with outside data manipulation and analysis to create future projections within the film industry. This projection data could then be used to visualize future trends and where the industry is heading.

**6. Acknowledgements**

The visualization project was a pleasure to complete due to the support, encouragement, and participation of each team member. Each individual was instrumental in the success and accomplishments of the project. Searching and finding a theme that would intrigue its members was an enjoyable task that allowed each member to play a role in. This work could not have been completed without IMDb’s extensive data on the film industry. Each member played an integral role within this project and created stellar visuals and new data findings.

Professor Dogan was instrumental in providing not just the purpose of good data visualization, but the enjoyment of it as well. His teachings have been pivotal in the success of this project, as well as the many future uses of data visualization. His class was always an enjoyable space to be a part of due to his great taste in music, off-road vehicle adventures, and authenticity in sharing his life through iPad photos. This could not have been done without his help.

**7. References**

1. https://www.hollywoodreporter.com/news/general-news/2019-global-box-office-hit-record-425b-4-percent-plunge-us-1268600/
2. https://www.washingtonpost.com/lifestyle/style/the-unlikely-greatness-of-the-shawshank-redemption-25-years-later/2019/08/29/e832b066-c806-11e9-be05-f76ac4ec618c\_story.html
3. https://www.thewrap.com/study-finds-big-budget-movies-make-greater-profits-9021/
4. https://www.thehindu.com/features/metroplus/animation-movies-meant-for-adults-as-well/article7170709.ece

**Link to the Presentation:** https://youtu.be/-KhG-SDd\_80