**MOVIE\_GROSS\_RATING FINAL**

**BACKGROUND**

I always wondered if there was any correlation between the quality of a movie and how much gross revenue that particular movie made. This project can give us insight on answering the question “Is there any correlation between gross revenue and rating?”

**PART I: FIND A DATASET**

I found a movie\_gross\_rating dataset on a CSV file at Kaggle.com to assist me with this project.

**PART II: EXPLAIN THE MODEL**

The movie\_gross\_rating dataset consists of the top 20 movies from every year, spanning 25 years, starting in 1989, and ending in 2014. The “top 20” term has multiple implications such as movies that grossed the highest revenue, movies that received critical acclaim, and the number of nominations and awards associated with the movie. The original dataset contained the following variables and data types:

* Index (Integer)- Uniquely identifying each movie with its own number.
* MovieID (Integer)- Uniquely identifying each movie with its own number.
* Title (String)- The name of the movie. 509 movies in total.
* MPAA Rating (String)- The Motion Picture Association of America rating of the movie. The ratings are:
  + G
  + PG
  + PG-13
  + R
* Budget (Integer)- The budget of the movie in US dollars.
* Gross (Integer)- The gross revenue of the movie in US dollars.
* Release Date (Date)- The date the movie was released. The release date is broken down into:
  + Year
  + Quarter
  + Month
  + Day
* Genre (String)- The genre of the movie. The genres listed are:
  + Action
  + Adventure
  + Animation
  + Comedy
  + Crime
  + Drama
  + Family
  + Fantasy
  + History
  + Horror
  + Mystery
  + Romance
  + Science fiction
  + Thriller
  + War
  + Western
* Runtime (Integer)- The length of the movie in minutes.
  + The runtime ranges between 79-201 minutes.
* Rating (Decimal)- The average rating of the movie based on a 10-point scale.
* Rating Count (Integer)- The number of ratings the movie has received.

The data seemed thoroughly comprehensive and didn’t have any missing values. However, I did find that having an “Index” column *and* a “MovieID” column was a little redundant, so I deleted the “Index” column (both columns contained the *exact* same information). Looking over the dataset, I did see some outliers but I kind of expected it given the nature of the dataset. I expected vast differences when it came to gross, rating count, budget, etc. given that these numbers depend on numerous variables such as the amount of movies in this database, the year these movies were made and released, the difference in the amount of critics throughout the years, the budget allotment due to studio expectations for each movie, etc.

These are the ranges for the following columns:

* Budget- Range $379,940,000 (Minimum $60,000, Maximum $380,000,000)
* Gross- Range $2,787,912,087 (Minimum $53,000, Maximum $2,787,965,087)
* Rating-Range 4.9 (Minimum 4.10, Maximum 9.0)
* Rating Count-Range 2,126,254 (Minimum 974, Maximum 2,127,228)
* Runtime-Range 122 minutes (Minimum 79 minutes, Maximum 201 minutes)

**PARTS III & IV: CREATE A VISUALIZATION THAT LEADS TO INSIGHTS/PROVIDE FILTERING CAPABILITIES TO YOUR DASHBOARD:**

I created three different pages (the first two have Slicers for filtering capabilities) that are broken down as such:

1. **Movie Gross Rating Page**- This page is the page that answers the initial question “Is there any correlation between gross revenue and rating?”
   1. This page consists of a Line Chart with the movie “Titles” listed along the X-axis, the “Gross Revenue” along they Y-axis, and the “Ratings” along the Y-axis as well. This chart is sorted by “Gross Revenue.”
   2. This page has two Cards: one displaying the “Average Gross Revenue” and one displaying “Average Rating.”
   3. This page contains four Slicers:
      1. Genre, Title
      2. Date Hierarchy (Year, Quarter, Month, Day)
      3. MPAA Rating
      4. Runtime
2. **Further Insights Page**- This page really digs deep and can give you a number of combinations to get very specific insights.
   1. This page consists of a Line and Stacked Column Chart so we can better visualize the “Budget” versus the “Gross Revenue” amount as well as the rating for each “Title.” This chart is sorted by “Gross Revenue.”
   2. This page consists of seven Multi-Row Cards that display the following insights for the Budget, Gross, and Rating columns:
      1. Minimum
      2. Maximum
      3. Average
      4. Standard Deviation
      5. Median
      6. Variance
      7. Sum
   3. This page consists of five Slicers:
      1. Genre, Title
      2. Date Hierarchy (Year, Quarter, Month, Day)
      3. MPAA Rating
      4. Runtime
      5. Rating
3. **Numbers Summary**- A summary of all of the columns that contain numerical data.
   1. This page consists of seven Multi-Row Cards that display the following insights for Budget, Gross, Rating, Rating Count, and Runtime.
      1. Minimum
      2. Maximum
      3. Average
      4. Standard Deviation
      5. Median
      6. Variance
      7. Sum

**PART V: PREPARE A SUMMARY**

**DASHBOARD:**

**“Movie/Gross/Rating” Page**

The name of the first page in this report is “Movie\_Gross\_Rating.” I used this name because it sums up what this project is about without getting overly complex. I used a darker background for my visual because the constant staring at a bright computer screen can become detrimental to a person’s vision, so I thought I’d create a visual that was little less intense.

I created two Cards; one displaying the “Average Gross Revenue” and the other displaying the “Average Rating.” I did this because the entire point of this project is to see if there is any correlation between revenue and rating and I felt like getting the average of these particular variables was the most solid foundation for this investigation. They are displayed at the top left of the dashboard so that the numbers are big and bold because of their importance (also in the United States we read left to right).

I created a Line Chart displaying the “Titles” along the X-axis and “Gross Revenue” and “Rating” along the Y-axis. I used a Line Chart because I had two numerical variables and one categorial variable that I wanted to display for my Multivariate analysis. This chart gave me an accurate visual of what this project represents: the possible correlation between gross revenue and ratings when pertaining to movies. I placed it at the bottom of the dashboard because I had other visuals that took up space on the top and I wanted the Line Chart to fully engulf the entire width of the dashboard.

I created four Slicers (Genre/Title, Date Hierarchy (Year, Quarter, Month, Day), MPAA Rating, and Runtime) so that users can filter out specifics if they so desire based on these variables. Clicking on multiple data points from multiple slicers simultaneously can give you a really detailed experience.

**“Further Insights” Page**

After creating the initial page, I felt like creating another page just in case a user wants to get extremely intricate with the data. The “Further Insights” page goes beyond the scope of the finding out the average of the gross revenue and ratings of movies; this page allows you to see if there is any correlation with the budget alongside the revenue and ratings of the movies, as well as statistical information based on filtered variables.

I created a Line and Stacked Column Chart as the centerpiece visual for this page. This chart contains the “Title” variable along the X-axis, the “Gross Revenue” and “Budget” variables on the Column Y-axis, and the “Rating” variables on the Line Y-axis. The I did this so you can easily see how the budget and gross variables correlate within one another while the line that represents the rating variable is separate yet distinctive. This chart can give you a conclusive Multivariate visual analysis of these respective variables.

I created seven Multi-Cards displaying the “Budget,” “Gross,”, and “Rating” of the following statistics: minimum, maximum, average, standard deviation, median, variance, and sum. This can shed some detailed statistical information if a user wants to get more technical with their data. These visuals are located directly beneath the Line and Stacked Column Chart as this information is secondary to the visual and may or may not be needed for further insight.

I created five Slicers (Genre/Title, Date Hierarchy (Year, Quarter, Month, Day), MPAA Rating, Rating, and Runtime). These filters were created to not only assist filtering data in the Line and Stacked Column Chart visual, but also assist in filtering data for the Multi-Cards that are computing the statistical data. The slicers are located directly beneath the Multi-Cards as they may or may not be needed for further insight.

It might seem like some of the variable usage on this particular page is redundant, for instance, using “Budget” in the Line and Stacked Column Chart *and* inside of the Multi-Cards. If someone wanted to know the average budget for a particular genre, for example, they can click on “Comedy” in the “Genre/Title” Slicer and see that the average budget for the comedy genre is $48,698,404.26 inside of the Multi-Card that displays averages. You wouldn’t be able to get that information just looking at the Line and Stacked Column Chart. This page allows you to really get as specific as you want outside of the original scope of this project if need be.

**“Numbers Summary” Page**

The third and final page of this project is the “Numbers Summary” page. This is an overall look at all of the numerical data from the entire dataset. There are seven Multi-Cards displaying the “Budget,” “Gross,” “Rating,” “Rating Count,” and “Runtime” for the following statistical data: minimum, maximum, average, standard deviation, median, variance, and sum. Below is a chart displaying the numbers that were retrieved from this page:

**Numerical Data Summary**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Stat Type | Budget | Gross | Rating | Rating Count | Runtime |
| Minimum | $60,000 | $53,000 | 4.10 | 974 | 79 |
| Maximum | $380,000,000 | $2.79B | 9.0 | 2,127,228 | 201 |
| Average | $84,299,050.07 | $383.20M | 6.92 | 337,926 | 117.76 |
| Standard Deviation | $59,894,151.19 | $227,281,623.37 | 0.89 | 321,088.95 | 22.66 |
| Median | $70,500,000 | $310,071,633 | 6.90 | 239,865 | 115 |
| Variance | $3.5Q | $76Q | 0.79 | 103B | 513.31 |
| Sum | $42,992,515,538 | $195,431,345,457 | 3,527.30 | 172,347120 | 60,057 |

Ranges for reference for the following columns:

* Budget- Range $379,940,000 (Minimum $60,000, Maximum $380,000,000)
* Gross- Range $2,787,912,087 (Minimum $53,000, Maximum $2,787,965,087)
* Rating-Range 4.9 (Minimum 4.10, Maximum 9.0)
* Rating Count-Range 2,126,254 (Minimum 974, Maximum 2,127,228)
* Runtime-Range 122 minutes (Minimum 79 minutes, Maximum 201 minutes)

**Insights**

“Is there any correlation between gross revenue and rating?” Speaking in a general sense, I believe so. Here are some of the insights that helped me with my conclusion:

* According to the data, the average gross revenue is $383.20M and the average rating is 6.92 (and the median is 6.90) on a scale of 1-10. This tells me that the average film rating is above average, almost a 7, for movies that grossed an average amount of $383.20M (a substantial amount).
* The minimum rating was 4.10 and the maximum rating was a 9.0. This tells me that even though the lowest rating was below average, it wasn’t extremely below average, it was just below a 5.
* The highest grossing movie was “Avatar,” grossing $2.79B, and rated at a 7.8. This tells me that the highest grossing movie has a relatively high rating.
* The lowest movie rating belonged to “The Last Airbender,” rated at 4.10. This is below the average rating of 6.92. This movie grossed $318,502,923.00. This is below the average of $383.20M. Even though this movie grossed a significant amount of money, it was still below average and it correlates with its below average rating.

**Miscellaneous Insights:**

* The lowest grossing movie was “Edward Scissorhands,” grossing $53M. It was rated a 7.90. This is telling me that even though this movie was the lowest grossing, it still maintained a relatively high rating.
* The highest rated movie was “The Dark Knight,” which is rated at 9.0. This movie grossed $1B. This tells me that even though this movie didn’t gross the most, it did substantial numbers while maintaining the highest rating.

These miscellaneous insights tell me that the highest grossing amount doesn’t necessarily translate to being the highest rated and vice versa.

**Insights Conclusion:**

Overall, given these insights, I think it is safe to say that there is a correlation between rating and gross revenue. It might not be accurate in every context, but it does apply in a general sense.

**Limitations and/or Challenges While Working with the Data**

I didn’t find any real limitations or challenges in this particular instance. All of the data seemed to be very comprehensive for the nature of the dataset and it was easily uploaded into Power BI. However, I will notate the following:

* When importing the data, I noticed that the data types for the columns were different: MovieID (Whole Number), Title (Text), MPAA Rating (Text), Budget (Whole Number), Gross (Whole Number), Release Date (Date), Genre (Text), Runtime (Whole Number), Rating (Decimal), Rating Count (Whole Number).
  + I changed “Budget” and “Gross” from whole number data types to currency data types with two decimal places so that the data can be read in US currency.
  + I added a coma format to the “Rating” column so that the data can be read in a more comprehensible manner.

**Future Steps to Continue Exploring the Data**

Once I started to get my initial analysis, my mind started to wonder about all of the possible ways I can use these variables to get so many other insights about this dataset, beyond the correlation of the gross revenue and ratings of movies. I can get as specific as the day of the week to really get in-depth insights and start correlating that data with other variables. I feel like there’s dozens of ways this data can be used for specifics. I would use the all of the slicers and statistical cards with all of their respective variables in the “Future Insights” page I created to continue exploring the data. Clicking on the variety of options between slicers and seeing how they measure up statistically can give me a very thorough way of exploring the data. Maybe adding different filters, changing the orders around from ascending to descending, trying out different graphs and charts can assist in further exploration as well.