## Table of Contents

## Abstract

As the rapid growth of data, visual analytics is becoming a demand to be used in business. The implementation of visual analytics enables users to analyse the large complex data and visualize the information visually. Adopting the analytics in the film industry would provide insights for movie producers to predict movie trends, increasing the success rate of the movie and understand the movie audience’s interest in movies. **Objectives:** The main purpose of this paper is to present a dashboard to the movie producers. The dashboard was created to present the insights obtained from the IMDB dataset. **Methodology:** This report utilizes the PowerBI tool to create the dashboard. The development of the dashboard was mentioned to explain how the dashboard is formed. **Findings:** Two dashboard screens included in this report. The first dashboard screen is IMDB movie analysis by rating. The second dashboard screen is IMDB movie industry financial analysis. Overall, we conclude the insights from the movies released within the year 2018, 2019 and 2020.

## Introduction

In this new era of technology, many companies have implemented visual analytics to improve their business performance. By using visual analytics, we can gain meaningful insights from the real time data. This helps us in making data-driven decision making. Therefore, a dashboard is created to present the data visually to our viewer. The dashboard would grab the viewer's attention as it is presented in a way that we can easily understand the data and it encourages viewer’s interactivity with the data. To define the term, a data dashboard is an information management tool that analyse and display the data visually (DataConsulting, 2018). It is efficient to use because it can be customised based on the customer's needs. Moreover, making a dashboard can present the insights by summarizing the information with different visual techniques such as bar chart and line graph. As we had drawn dashboard sketches previously, we will be using PowerBI to create a dashboard with IMDB movie dataset. Also, the viewers of our dashboard are from the film industry, we therefore used movie data to display the relevant content that would be useful to them ("Reasons to use dashboards for data analytics", 2020).

# Methodology

## Data Exploration

For the IMDb extensive movies dataset we have chosen from kaggle, we only took the IMDb movies.csv file as we are only interested in the average weighted user rating and the average revenue for each title. In this dataset, there are variables such as IMDb title ID, title, original title, year, date published, genre, duration, country, language, and director.

## Rating

All IMDb registered users are allowed to give a vote ranging from one to ten for each and every released title in the IMDb database. Each and every user's votes are accumulated and condensed into a single IMDb rating which is shown on the main page of the title. What IMDb considers as a "released title" is that the movie or show has been screened in public at least once. Each and every user is allowed to recast their vote as frequently as desired and their old vote will be overwritten by their latest vote of the same title. This means that IMDb only allows one vote for each user for each title. IMDb uses all registered user ratings to compute and determine a single rating. They also show the weighted average on the title's page but do not use the arithmetic mean. The IMDb weighted average is updated a great many times a day instead of updated based on each new vote casted. For TV series, users vote individually for the rating of the TV series as a whole on every title's series page instead of taking the weighted average ratings for each episode. Distinct measures are taken such as using a different weighting calculation to be able to maintain the reliability of the IMDb system when faced with detected unusual voting activity.

IMDb does not plainly compute the average rating by simple math of summing up all votes and dividing it by total number of votes. They do not disclose information about how the weighted ratings are computed in order to halt abuse and reduce attempts to control the outcome by users. In infrequent cases, IMDb might spot attempts to sway the user rating for a title by using clump voting campaigns and other alike methods. Without going against their policy to not eliminate valid votes, IMDb computes weighted user averages by applying a mechanism to counterbalance the ramifications of efforts to unnaturally increase or decrease a title's rating ("IMDb | Help", 2020). For these reasons, we have decided to choose user ratings as one of our indicators of a successful movie.

## Revenue

Movie revenue is calculated in terms of number of tickets sold or the sum of money made from ticket sales. The analysis and computation of these revenue is crucial for the creative industries and frequently a source of interest for fans. There are a number of ways to calculate the revenue of a movie.  The first one is the revenue obtained from receipts received by the distributors itself .The second one is the revenue obtained from the title's first weekend gross. The third is the revenue obtained from the total length of the theatrical run which is the one we have chosen as one of our indicators of a successful movie. The reason for this is because a title may advance at a slow pace through art houses or by being screened week after week in tens of hundreds of multiplexes which would indicate a title having a long run (*Cinematic Success Criteria and Their Predictors: The Art and Business of the Film Industry*, 2020).

## Data Transformation

To make a dashboard, we have to make sure that the dataset that we are using is clean and well formatted. To perform data cleaning, we can use various softwares including MS excel but we have to export the data into another software to visualize. For this reason we used Power BI to clean and visualize the data all in one place. Once the dataset is imported into the Power BI software, a query editor tab will appear. This tab contains tools that can help in removing, filtering, adding, splitting and formatting rows and columns. The query editor also allows us to add more data sources and create connections between datasets. For our dashboard, the dataset that we have acquired alone contains sufficient information to create informative dashboards.

While observing the dataset inside the query editor, we can see that there is a lot of data that can be difficult to process due to several reasons. All the columns were also categorised under general data type. Firstly, we removed columns that we were sure would not be used under any circumstances during the making of the dashboards. The columns we removed were review_from_critics, review_from_users, metascore, votes, writers, actors, directors, description, Language and title. Some of these attributes do contain data that is useful and can be analysed and visualised such as writers, actors and directors but since we were not planning on using them, we removed them to increase efficiency of the software. Some other attributes such as Production company and US gross income, that we did not use in the dashboards but were not sure of removing were hidden after applying the query changes. Once the columns were removed, we filtered rows that could decrease the accuracy and efficiency of the analysis like blank values and null values. Another particular and significant thing that we did was keeping values under budget, US gross profit and World gross profit that started with “$” only.  This was to make sure that there was no confusion with the currency values of other countries. Values without any currency symbols were also not taken as it may refer to any currency’s monetary value.

One of the most important transformations that we made that was crucial to the dashboard was to the genres variable. The genres variable contained more than one genre that a movie may relate to ranked from the most related to the least. This will cause problems when wanting to visualize a report that contains the genre variable. For that reason, we split the genre column by the delimiter as comma(,) then removing every column that was splitted only leaving the column with the most relevant genre. Another important transformation that we performed was creating a new column called Duration Range which was a conditional column created from the variable Duration. The purpose of this variable is to categorise the every minutes in the duration variable into 3 ranges, 0-90mins,90-120mins,>120mins so that it will be easy to include into the visualization. The condition was based on the “IF” function on the different ranges of minutes.

We have also created several measures to help our visualization. The Average vote, Average Budget, Average Revenue are all measures created by calculating the average of variables avg_vote, budget, Worldwide_gross_revenue respectively. The profit measure was created by calculating the difference between Average Revenue and Average Budget.

## Visualization

For the first page of Dashboards, there are a total of 6 visuals present on the report screen. The first visual is the Average vote variable shown in a card visual. This visual will clearly present the average vote in the dataset when filters are added. The visual is added selecting the Average Vote measure and clicking on the card visual on the visualization panel. The visual beside the Average Vote card is a pie chart that contains the percentage of distribution of the dataset throughout the three years. This is created by dragging the date published variable to the legends fields and dragging the IMdb_title_id variable to the values field while the pie chart visual is selected in the visualization panel. The visual next to it is a horizontal clustered bar chart that shows the distribution of movies through different duration ranges. This is created by dragging the duration range variable to the axis field and IMdb_title_id variable to the values field while the horizontal clustered bar chart is selected in the visualization. The visual below the Horizontal clustered bar chart is a vertical clustered bar chart similar to the one to the left of it. It contains the average ratings for each of the duration ranges. It is created by dragging the duration range variable to the axis field and Average vote variable to the values field while the vertical clustered bar chart visual being selected. Similarly the bar chart beside it is created in the same way except the value field contains the date_published variable. Since the date published  is a variable that contains multiple hierarchies of data (Year, month, day), A hierarchy is automatically created and the drill options are shown above the visual. When we drill down the hierarchy, the data will show average ratings for every month and even dates. The final visual in this dashboard is located on top right corner beside the title as a slicer. The slicer contains the Genres variable and works as a filter.

The second page of Dashboard screen contains a total of 6 visuals. Two of them being card visuals, two being horizontal clustered charts, one vertical clustered chart and one slicer. The two cards show the Profit Variable and IMdb_title_id counts. The horizontal clustered charts both have the Genre variable as the axis. One of them has Average revenue as Values and the other has Average Budge as Values. The vertical clustered bar chart is created with the axis containing duration Range the values containing Average Revenue. Finally the slicer contains the date published variable which allows in multi hierarchical filtering.

## Literature Review

The dashboard is formed according to the viewer's needs. This is because we need to know what content on the dashboard is to be shown. Thus the dashboard's message is successfully conveyed. Besides, the dashboards are visual displays. We'd like to observe every dashboard screen so that doesn't overfit one visual display unit for easier viewing. For instance, the layout of the dashboard may be adjusted within the dashboard tools. Additionally, the dashboard is monitored at a look. This implies that the dashboard audiences should be ready to easily understand the data displayed within the dashboard. Moreover, the dashboard often acts as a communication tool. It can provide data storytelling to our audiences. Hence, maintaining our viewer’s attention in looking at the dashboards.Thus, delivering the message of the dashboard effectively to the viewers. (Park, Y., & Jo, I, 2019 ). There are many guidelines for developing a successful dashboard framework with respect to the dashboard design literature. Summarize the best practices of dashboard design. In a few words and they are, make the complex simple, say a story about it, the visualization of data needed to be right, represent the information and disclose detail as needed (no more, no less). There are some design guidelines, such as"choosing your final goals before implementing the design elements, choosing your layout and colors, prioritizing simplicity, and using interactive elements," to name a few. Different representations of numbers, such as using secondary statistics to endorse a variable in a dashboard view to give it a better meaning. For instance, instead of only reporting that revenues have increased by 25,000, a dashboard may also demonstrate that it has risen by 25% since last month. The "secondary statistics" that provide a better meaning to raw sales data are the 25 percent growth percentage. Displaying small charts by the side of the numeric data improves the visualization. (Yap, 2020).

## Findings

IMDB movie analysis by ratings.

This dashboard displays the IMDB movie analysis by ratings. The following result is obtained without selecting any specific movie categories. First, the average vote of movies. The value is displayed using the card visualization. From Figure 1.0, the average vote of all categories of movies is 6.26. Secondly, a pie chart is used to view the count of imdb_title_id by year, months and days. We can view the counts of imdb_title_id by months and days through pressing the drill down function. It can be observed that 15.41% of the movies produced in year 2018, 38.66% of the movies produced in year 2019, and 45.94% of the movies produced in year 2020. Based on Figure 1.1, 9.24% of the movies produced in January, 8.68% of the movies produced in February, 11.76% of the movies produced in March, 7.28% of the movies produced in April, 7.28% of the movies produced in May, 6.72% of the movies produced in June, 7% of the movies produced in July, 10.64% of the movies produced in August, 7.56% of the movies produced in September, 11.2% of the movies produced in October, 7.84% of the movies produced in November, 4.76% of the movies produced in December. In Figure 1.2, 4.76% for day 1, 1.96% for day 2, 2.8% for day 3, 1.4% for day 4, 5.32% for day 5, 3.08% for day 6, 2.52% for day 7, 3.64% for day 8, 2.24% for day 9, 1.68% for day 10, 3.92% for day 11, 2.8% for day 12, 3.36% for day 13, 5.6% for day 14, 3.08% for day 15, 2.8% for day 16, 3.64% for day 17, 2.8% for day 18, 3.08% for day 19, 2.52% for day 20, 4.2% for day 21, 2.8% for day 22, 1.96% for day 23, 3% for day 24, 3.92% for day 25, 3.36% for day 26, 3.08% for day 27, 4.48% for day 28, 3.64% for day 29, 1.96% for day 30, 4.48% for day 31. Thirdly, the horizontal bar chart is used to display the count of imdb_title_id by duration_range. The highest count of movies is 223 movies by 90 to 120 mins duration of movie. Following, 86 movies for more than 120 mins duration of movie and 48 movies for 0 to 90 mins duration of movie. The fourth chart is a bar chart showing the average vote by year, months and day. 6.23 average vote for year 2018, 6.39 average vote for year 2019, 5.96 average vote for year 2020. For average vote in months (Figure 1.3), 6.72 for January, 5.99 for February, 6.06 for March, 6.07 for April, 6.19 for May, 6.18 for June, 6.06 for July, 6.12 for August, 6.27 for September, 6.48 for October, 6.43 for November, 6.65 for December. Continuing by days(Figure 1.4), 6.27 for day 1, 5.70 for day 2, 6.52 for day 3, 5.98 for day 4, 6.69 for day 5, 5.89 for day 6, 6.77 for day 7, 5.72 for day 8, 5.98 for day 9, 6.4 for day 10, 5.83 for day 11, 6.18 for day 12, 6.06 for day 13, 6.25 for day 14, 6.46 for day 15, 6.09 for day 16, 6.23 for day 17, 6.07 for day 18, 6.44 for day 19, 6.09 for day 20, 6.22 for day 21, 6.31 for day 22, 6.79 for day 23, 6.85 for day 24, 6.39 for day 25, 6.08 for day 26, 6.24 for day 27, 6.25 for day 28, 6.26 for day 29, 6.33 for day 30, 6.44 for day 31. The fifth chart is the average vote by duration_range. More than 120 min duration of movie has the highest average vote which are 6.99. 90 to 120 mins duration of movies has average vote of 6.16, and 0 to 90 mins duration of movies has average vote of 5.40.

When the category of movies such as action, adventure, animation, or biography is specified, the dashboard will have changes based on the chosen movie category. If we clicked on the action movie category, the analysis would be based on the action movies. Thus, we will be able to obtain results as the above analysis.

IMDb Movie Industry Financial Analysis

In this dashboard screen, it shows the financial analysis of the year 2018, 2019 and 2020's profit. Without selecting anything specific yet, figure 2.1 shows the average profit of all 3 years with $124.61 million by 357 titles. It also shows the average revenue by duration of titles by all 3 years with over $342 million for more than 120 minutes, over $125 million for between 90 to 120 minutes, and slightly over $61 million for less than 90 minutes. Based on this, it shows that the longer the title duration, the higher the revenue with each longer title's duration range being at least double its revenue. Other than that, figure 2.1 also shows the average revenue by genre of all 3 years with animation being the highest followed respectively by war, action, adventure, crime, drama, comedy, horror, biography, fantasy, romance and music being the lowest. Furthermore, figure 2.1 shows the average budget by genre of titles by all 3 years with animation being the highest followed respectively by war, adventure, action, family, crime, comedy, biography, drama, fantasy, horror, romance,  and music being the lowest. In figure 2.2, each year can be selected individually as well as each individual month of each year to show the financial analysis of that particular time. Now, by selecting 2018 in figure 2.2, it shows the financial analysis of the year 2018. Based on the figure 2.3, it shows that in the year 2018, the average profit is $139.59 million by 138 titles. The average revenue by duration of titles by year 2018 is over $354 million for more than 120 minutes, over $150 million for between 90 to 120 minutes, and slightly over $78 million for less than 90 minutes. Based on this it also shows that the longer the duration, the higher the revenue with each longer title's duration range being at least double its revenue. Other than that it shows that the average revenue by genre of year 2018 with animation being the highest followed respectively by action, adventure, horror, biography, comedy, drama, crime, family being the lowest. Furthermore, figure 2.3 shows the average budget by genre of titles by year 2018 with animation being the highest followed respectively by action, adventure, comedy, family, biography, crime, drama, and horror being the lowest. Now by selecting 2019 in figure 2.2, it shows the financial analysis of the year 2019. Based on figure 2.4, it shows that in the year 2019, the average profit is $142.88 million by 164 titles. The average revenue by duration by year 2019 is slightly over $382 million for more than 120 minutes, over

$128 million for between 90 to 120 minutes, and very slightly over $63 million for less than 90 minutes. Based on this it also shows that the longer the duration, the higher the revenue with each longer title’s duration range being at least double its revenue. Other than that it shows that the average revenue by genre of year 2019 with animation being the highest followed respectively by action, adventure, crime, comedy, drama, horror, biography, fantasy, and romance being the lowest. Furthermore, figure 2.4 shows the average budget by genre of titles by year 2019 with animation being the highest followed respectively by action, adventure, crime, comedy, drama, biography, fantasy, horror, and romance being the lowest. Now by selecting 2020 in figure 2.2, it shows the financial analysis of the year 2020. Based on figure 2.5, it shows that in the year 2020, the average profit is $32.54 million by 55 titles. The average revenue by duration by year 2020 is over $126 million for more than 120 minutes, over $59 million for between 90 to 120 minutes, and slightly over $15 million for less than 90 minutes. Based on this it also shows that the longer the duration, the higher the revenue with each longer title’s duration range being at least double its revenue. Other than that it shows that the average revenue by genre of year 2020 with war being the highest followed respectively by adventure, drama, action, crime, horror, biography, comedy, animation, fantasy, and music being the lowest. Furthermore, figure 2.5 shows the average budget by genre of titles by year 2020 with animation being the highest followed respectively by adventure, war, action, drama, biography, crime, comedy, and horror being the lowest.

## Recommendation

The recommendations that can improve in a movie production industry is by developing movies that are longer than 120 minutes because as it was shown in the dashboard. Movies that are longer than 120 minutes have a higher revenue compared to movies that have a lower time duration. Longer films are more likely than other films to be scored Fresh on Rotten Tomatoes. Since evaluating the 1,431 movies that have had a wide theatrical release since 2010, and separating them into four categories, the film and TV rating aggregator came to this conclusion: under 100 minutes, 100-120 minutes, 120-140 minutes, and over 140 minutes. Therefore, If movie producers would like to earn more profit, we would recommend producing movies which are longer than 120 minutes. The following recommendation will be developing more animation movies. According to the  dashboard, animation movies in all the 3 years have the most revenue among all the other movie genres. Animation movies are important because it allows them to tell stories and express feelings and thoughts in a distinctive, easy-to-perceive manner that can be grasped by both young children and adults. Animation has continued to bind audiences around the world in a manner that can often not be written and live-action movies. Therefore, movie producers will be able to earn more profit when they produce more animation movies. The last recommendation will be movies that movie producers can develop on a low budget. According to the dashboard, it was stated that the lowest movie budget genre is drama movies. Generally speaking, dramas don't need huge budgets in order to be successful. There are many examples of dramas that have been able to achieve more for fewer, surpassing even the most cynical observers' standards.Therefore, If movie producers that have a tight budget and would like to make a movie with a low cost then the drama movie genre is the best choice in doing so. They also might not know that the drama movie that they are going to develop, depending on the content, might bloom with popularity and could earn millions of profit.

## Conclusion

This study aimed to propose a methodology to analyze and visualize the IMDB movie datasets and also a recommendation on how movie production industries can improve their efficiency and decision making on producing movies. We then constructed a dashboard with the IMDB movie analysis by their ratings, votes, revenues and budgets. Thus, concluding that most movies among the 3 years which are 2018, 2019 and 2020, year 2019 has the most movies produced. Year 2019 has also the most votes among all 3 and majority of the movies produced within the 3 years have a movie duration range between 90 to 120 minutes. The movies data set is also analyzed by the average revenue by movie duration range and average revenue by genre with animation being the highest and the duration range more than 120 minutes as the highest. Our analysis shows that animation movies have the highest budget for production and drama being the lowest.

## Appendix

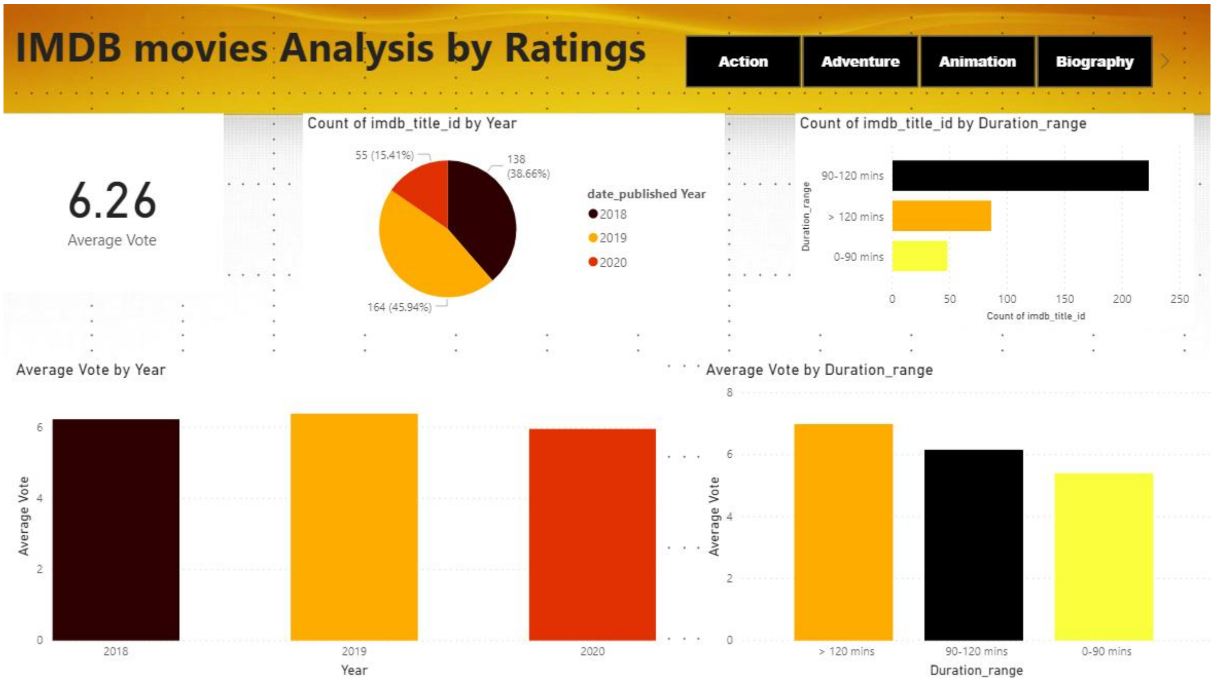


Figure 1.0

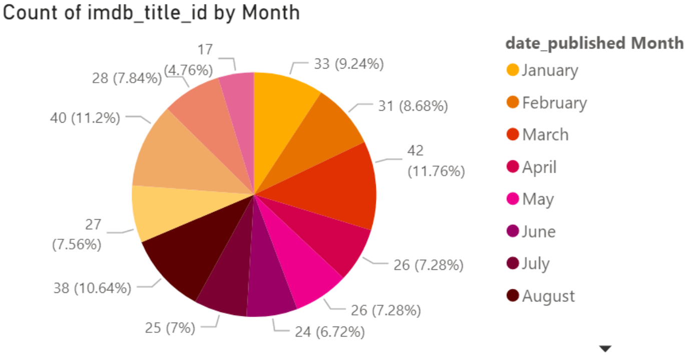


Figure 1.1

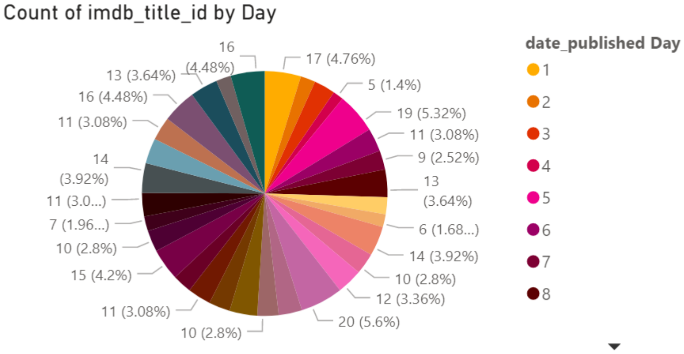


Figure 1.2

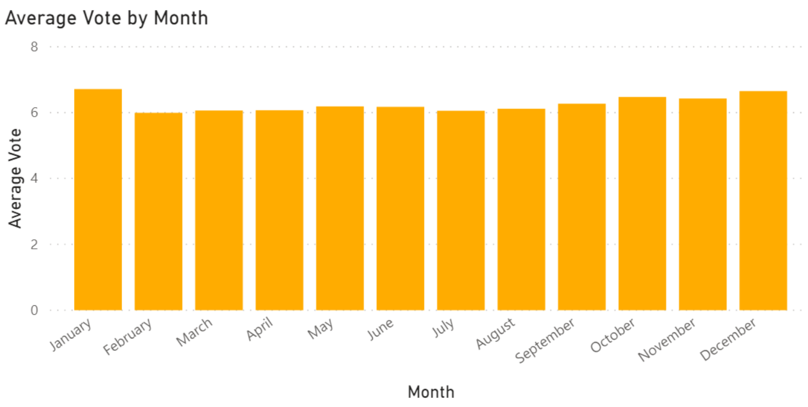


Figure 1.3

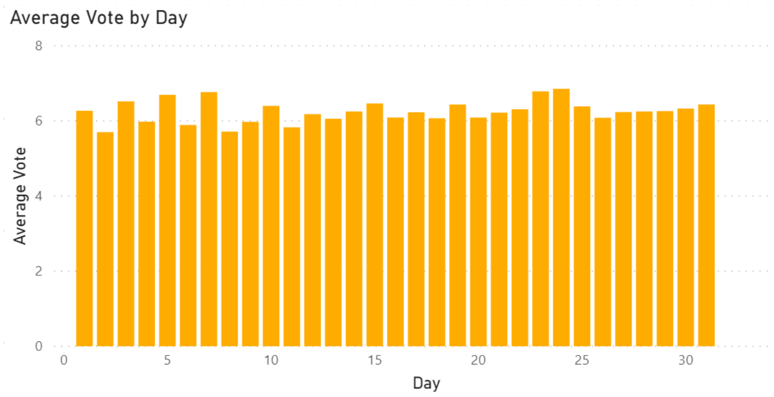


Figure 1.4

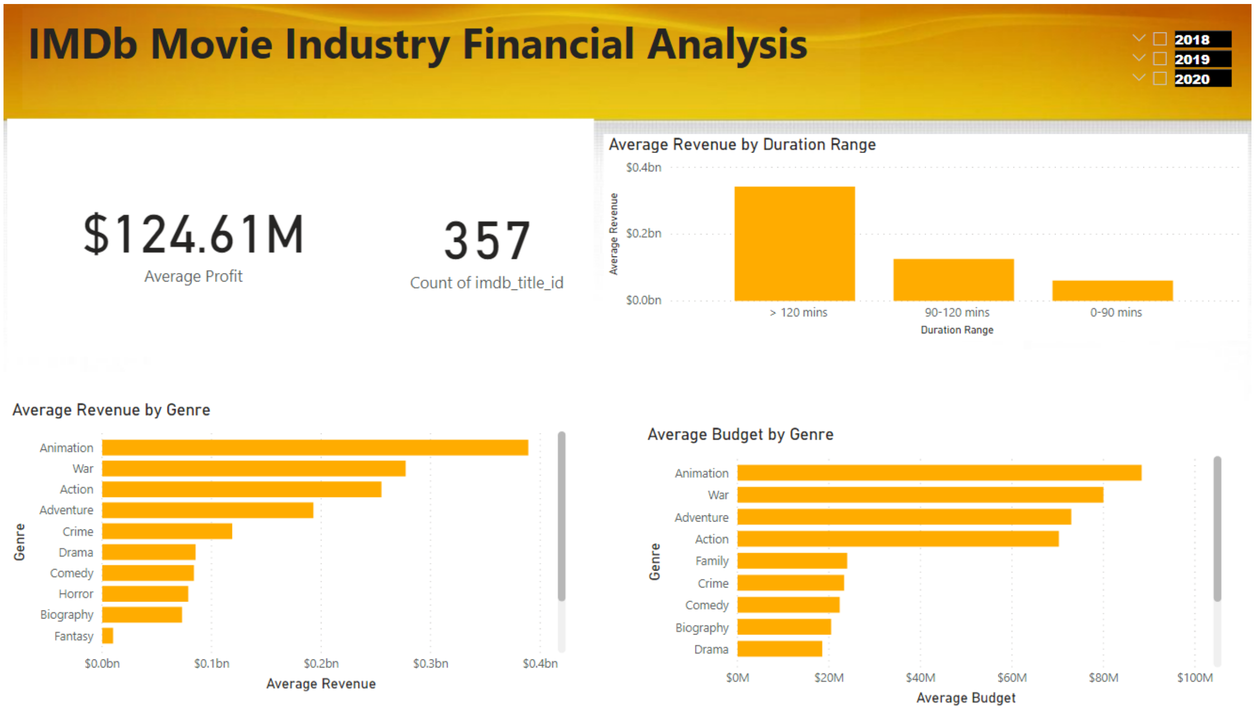
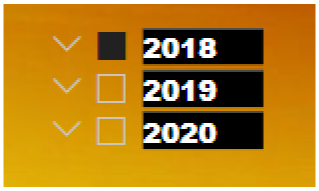


Figure 2.1

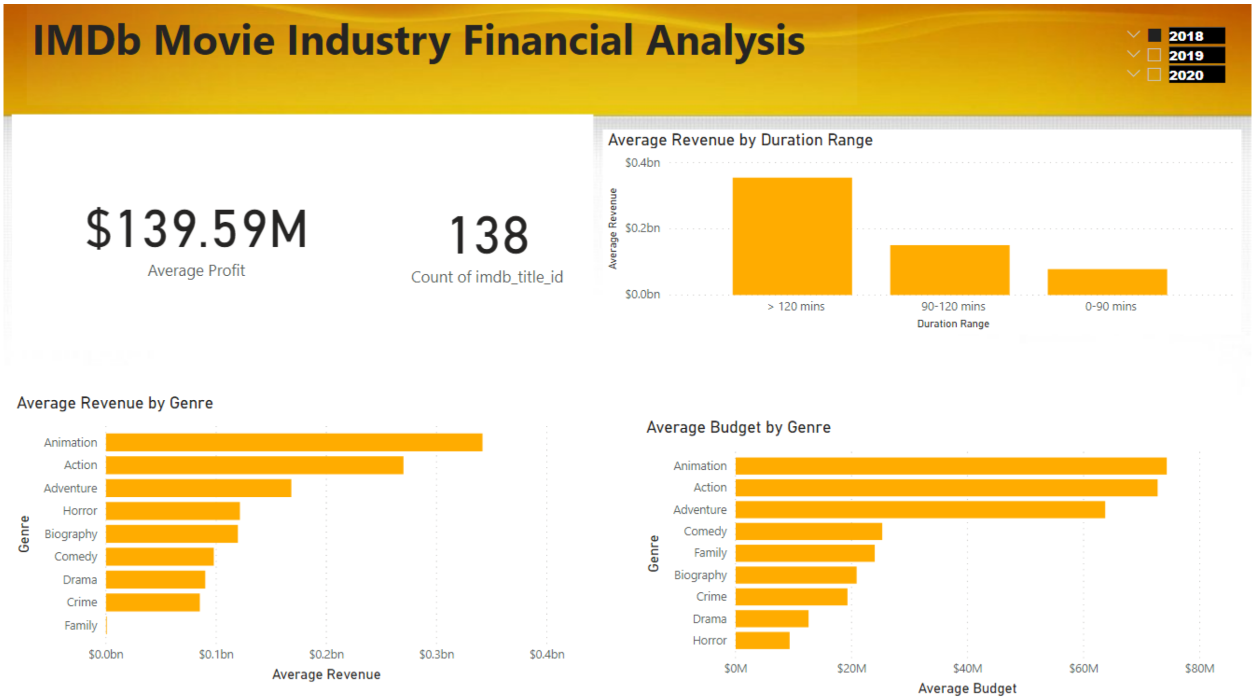


■ 2018

□ 2019

□ 2020

Figure 2.2



IMDb Movie Industry Financial Analysis

■ 2018

□ 2019

□ 2020

$139.59M

Average Profit

138

Count of imdb_title_id

Figure 2.3

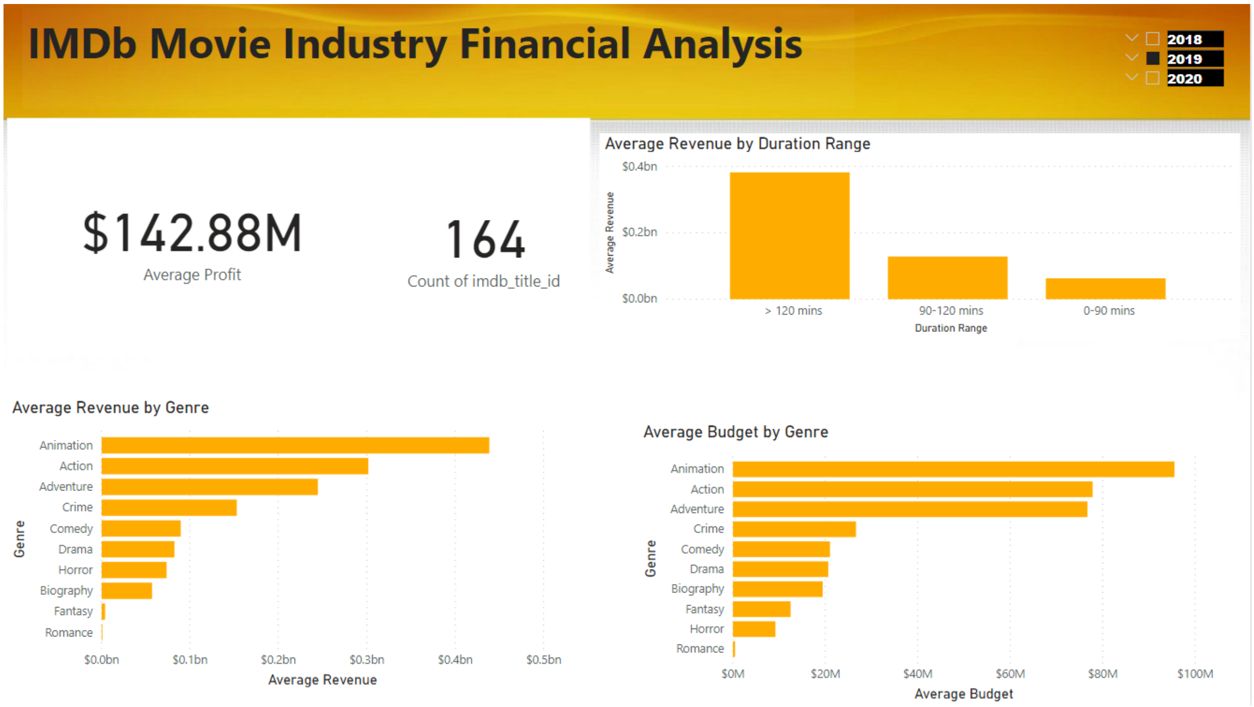


Figure 2.4

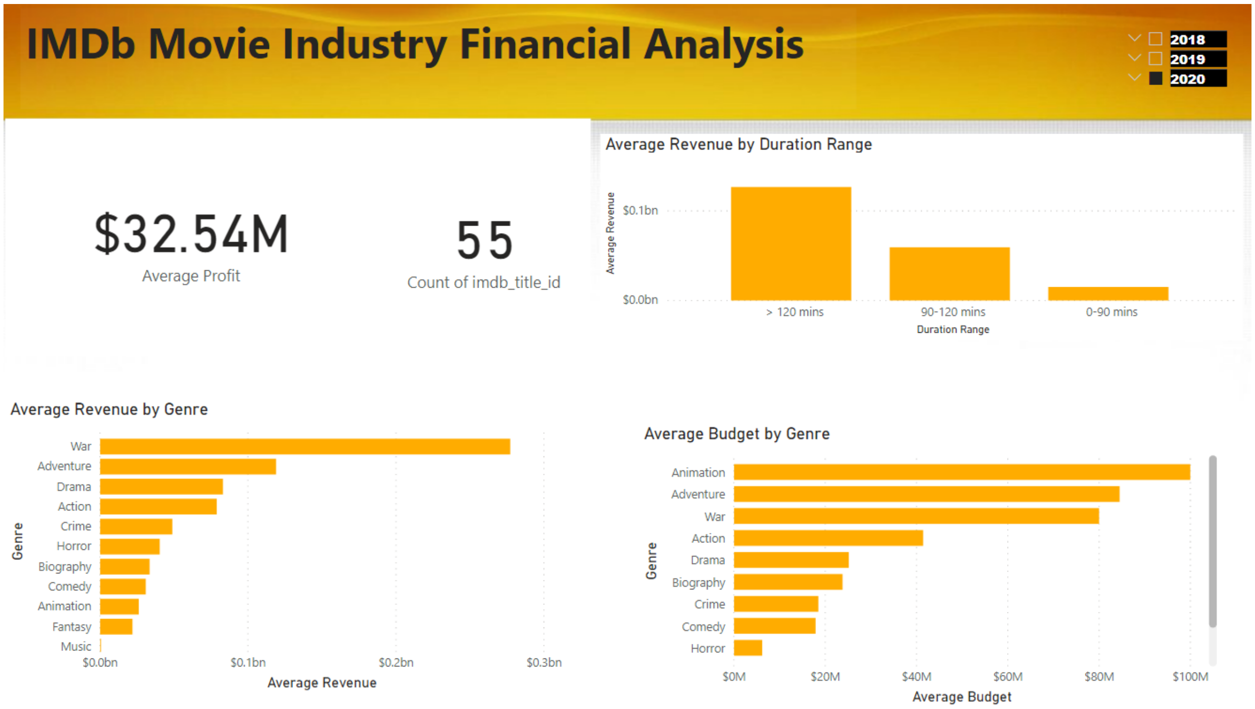


Figure 2.5

## References

*Reasons to use dashboards for data analytics*. Dataconsulting.co.uk. (2020). Retrieved 9 December 2020, from http://www.dataconsulting.co.uk/what-is-a-data-dashboard/.

*IMDb* | *Help*. Help.imdb.com. (2020). Retrieved 9 December 2020, from https://help.imdb.com/article/imdb/track-movies-tv/ratings-faq/G67Y87TFYYP6TWAV#.

Dean Keith Simonton. (2020). *Cinematic Success Criteria and Their Predictors: The Art and Business of the Film Industry* [Ebook]. Retrieved 9 December 2020, from https://onlinelibrary.wiley.com/doi/pdf/10.1002/mar.20280.

Park, Y., & Jo, I. (2019). Factors that affect the success of learning analytics dashboards. Educational Technology Research and Development, 1-25, from https://eds-b-ebscohost-com.ezproxy.sunway.edu.my/eds/detail/detail?vid=2&sid=03081472-d1a6-4689-9515-91662dc7cdc6%40sessionmgr101&bdata=JnNpdGU9ZWRzLWxpdmUmc2NvcGU9c2l0ZQ%3d%3d#AN=140064319&db=asn

Yap, A. (2020). *Creating Business Analytics Dashboard Designs using Visualization Methodologies: Case Methods for Innovative Analytics Pedagogy* [Ebook]. Department of Business Education. Retrieved 9 December 2020, from https://files.eric.ed.gov/fulltext/EJ1258142.pdf.

Sandy Schaefer. (2019). Longer Movies Are Better (According To Rotten Tomatoes) from
https://screenrant.com/longer-movies-fresh-ratings-rotten-tomatoes/

Dylan Gibson. (2019). Why Animation is Important from https://stonesoup.com/post/why-animation-is-important/#:~:text=Animation%20is%20important%20because%20it,and%20live%2Daction%20films%20cannot.

David Chiodaroli. (2019). 10 Highest-Grossing Low-Budget Drama Movies Ever Made
https://screenrant.com/highest-grossing-low-budget-dramas/