EN.605.662 – Data Visualization – Final Project

Tanner Amundsen

**Abstract**

This paper proposes a new dashboard created in Tableau that visualizes a dataset of 5000 movies from The Movie Database (TMDB) for the specific purpose of understanding the factors that contribute to a movie’s commercial success at the box office. In the end, three dashboards were created that relate revenue to things like budget, popularity, rating, runtime, release date, genre, production house, production company, keywords, and cast. These dashboards were used to quickly and easily identify how these factors maximize or minimize revenue.

**Introduction**

What makes a movie profitable can be a difficult thing to identify. There are many factors that can influence the success of any given film – factors like budget, runtime, genre, cast, release date, and name recognition can all play into the total revenue a film can earn. This paper proposes a series of dashboards to help interested parties such as studio executives understand what makes a movie profitable.

There is no shortage of data in this field. Databases like the Internet Movie Database (IMDB), The Movie Database (TMDB), and Box Office Mojo’s datasets are all well-formed, regularly updated databases that compile movie data. The problem with these databases is that they are too big for most people to sift through. A successful visualization of this data would be one that allows executives to draw meaningful conclusions around what makes a film profitable. Some analytical questions that users of these dashboards might want to answer are:

1. Which genre of film is most profitable?
2. What is the relationship between film budget and film revenue?
3. What is the average runtime of the highest earning movies?
4. When should I release a film (time of year) to maximize revenue?
5. Which production companies have produced the highest earning films?
6. How important is filming location in the box office success of a movie?
7. What do the most successful films have in common in terms of plot/script/content?
8. Which actors bring in the most revenue??

In addition to helping people answer these questions, a good visualization is also an interactive one. These dashboards should include data tips and filters that users than toggle to gain more insights. For example, in a scatterplot showing film budget by film revenue with potentially thousands of datapoints, a user should be able to hover over a marker to see which title that marker represents. Additionally, in summary plots that are showing something like average revenue by genre, a user should be able to hover over a genre to see the top 5 highest-earning films in that genre. In summary, these dashboards will help decision makers make data-driven decisions in picking films to produce and understand what factors make a film successful at the box office.

**Background**

**Dataset Background:** Several datasets have been compiled on movie data. The Movie Database (TMDB) is a popular “community-built movie and TV database” with over 800,000 movies [TMDB]. This database was started in 2008 and has grown steadily since then. There is a website that visualizes this dataGraphical user interface, application, website

Description automatically generated. It shows titles that default to release date as the order of appearance. Figure 1 is a dashboard that allows the user to sort and filter the titles using various interactive features such as checkboxes, buttons, calendars, sliders, and search bars. Notably, this dashboard does not include any data on revenue.

Figure - TMDB website visualization [source: <https://www.themoviedb.org/movie?language=en-US>]

Graphical user interface, application

Description automatically generated IMDB also includes a variety of visualizations on their website that attempt to digest their vast database. These charts include “Box Office”, “Most Popular Movies”, “Top 250 Movies”, “Most Popular TV Shows”, “Top 250 TV Shows”, “Top Rated Indian Movies”, “Lowest Rated Movies”, and “Popular TV Show and Movie Genres.” This list is similar to the TMDB dashboard in that it includes the poster for the movie and its title. This list includes the weekend and gross box office revenue along with the number of weeks that it has been in theaters.

Figure - IMDB Top Box Office (US) visualization for the weekend of April 21 - 23, 2023 [source: <https://www.imdb.com/chart/boxoffice/?ref_=nv_ch_cht>]

A picture containing text

Description automatically generatedChart, bubble chart

Description automatically generated**Literature Review:** There have been several attempts to create dashboards of either the IMDB, TMDB, or other movie database. One was detailed in “An exploration of great cinema using information visualization” [Blecksmith 2008]. They created a visual analysis of the IMDB top 100 rated movies using Many Eyes and Tableau. Through their work, they were able to identify a few factors that have consistent influence on commercial success and rating. The purpose of their visualization was oriented more towards an audience that cares about critical and popular success rather than commercial success. For example, Figure 4 shows the top 100 rated movies on IMDB sized by the number of Oscar nominations that film received. Through their work, they were able to identify genre and release date as two factors that had a high degree of correlation with critical success.

Figure - A visualization of the IMDB co-star network. Top chart is most connected actors and bottom chart is minor actors [source: Haughton 2014]

Figure - IMDB Top 100 movies sizes by the number of Oscar nominations they received [source: Blecksmith et al 2008].

Another popular application of data visualization to these datasets is visualization of the co-star network. The co-star network is a graphical representation of which actors and actresses have co-starred in films together. One reason to visualize a graph like this is to identify clusters of actors that tend to work together often or to identify other common attributes within a cluster. For example, one cluster might consist of actors that tend to star in mostly action films. Haughton et al [Haughton 2014] proposes a novel visualization of this network that seeks to overcome the hurdle of visualizing such a large network (over 2.6 million actors) with such a high degree of connectedness. Using the k-core approach, they were able to create a visualization that reveals the clear presence of clusters. This visualization does not immediately aid in analyzing revenue related data. Future work could be done to add a size or color element that encodes the average revenue of the films this actor has starred in.

Diagram, schematic

Description automatically generated Zooming in to the actual content of these movies, data visualization of plot, script, and content is a far less researched field. Genre is the closest attempt to summarize the content of a movie. Drucker discusses the use of data visualization in digital humanities research, including its potential for generating new insights and visualizing complex patterns in data [Drucker 2011]. However, the author also points out the limitations in visualizing humanistic expression and emphasizes that written reflection of the digital humanities still reigns in importance. Rajamanickam et al discuss how to create interactive data visualizations to help people understand and discuss non-linear storylines in films [Rajamanickam 2021]. The visualization in Figure 5 shows the timeline of the movie *500 Days of Summer*. The horizontal axis is a timeline ranging from day #0 to day #500. The circle markers represent scenes, sized by length of scene, colored by the relationship status of the two main characters at the time of that scene. The thin arcs connecting the scene markers show the order of appearance of these scenes in the actual film. This is just one example of a humanistic approach to film data visualization.

Figure - A visualization of the film 500 Days of Summer. [source: Rajamanickam 2021]

**Approach**

**Dataset Preparation:** The dataset used to create this visualization is called TMDB 5000 Movie Dataset and was uploaded to Kaggle in 2017 from The Movie Database [TMDB]. The dataset was compiled to answer questions about what makes a movie successful and includes data like budget, genres, keywords, popularity, production company, release date, revenue, and TMDB vote average. The creators of the dataset were not explicit in how they chose these 5000 movies in particular, but they do offer a great diversity in genre, popularity, and revenue. And given that the stated purpose of this visualization was to allow exploration of movie success, it is reasonable to assume that this sample of 5000 movies are representative of the larger TMDB database.

After importing the data, some data cleaning needed to be done. There were several columns that were saved in the string representation of a JSON file. These were converted into comma separated strings using a Python script so that it would be easier to split them in the Tableau Flow creation step. This data cleaning resulted in two new CSV files.

These two new CSV files were loaded in Tableau Prep Builder. There were several columns that contained entire lists, meaning that we needed to create joining tables for each of the many-to-many relationships. Using Tableau Prep Builder, the following joining tables were created:

1. A picture containing diagram

   Description automatically generatedfilm\_id 🡪 keywords
2. film\_id 🡪 actors
3. film\_id 🡪 genres
4. film\_id 🡪 production\_companies
5. film\_id 🡪 production\_countries

Diagram

Description automatically generatedThese resulted in a total of 6 Tableau .hyper files that served as the data sources for the Tableau dashboard.

Figure - Screenshot from Tableau Prep

In Tableau Desktop, these data sources were linked using the many-to-many cardinality joined on the “id” field.

Figure - Screenshot from Tableau Desktop 'Data Source' tab.

**Tableau Dashboard Creation:**

[***Dashboard #1: Revenue by Quantitative Metrics***](https://public.tableau.com/app/profile/tanner.amundsen/viz/AmundsenEN_605_662FinalProject-Dashboard1/Revenuevs_QuantitativeMeasurement)– This dashboard is a simple summary dashboard that relates various quantitative metrics to revenue using bar plots, scatterplots, and line charts. This dashboard is designed to give the user a very quick understanding of the sign and strength of the correlation between revenue and any other quantitative variable of interest. For example, this dashboard shows a weak to moderate, positive correlation between film budget and film revenue. See screenshot below:

Chart, histogram

Description automatically generated

Figure - Dashboard #1: Revenue vs. various quantitative metrics.

Chart, scatter chart

Description automatically generatedDashboard #1 is interactive in that users can hover over any marker for more information on that marker. For example, Figure 9 shows a user hovering over a marker in the Budget vs. Revenue scatterplot to reveal a data tip indicating that this marker represents the film *Avengers: Age of Ultron*.

Figure - Screenshot of user interacting with 'Budget vs. Revenue' scatterplot in Dashboard #1

[***Dashboard #2: Revenue by Keywords and Cast***](https://public.tableau.com/app/profile/tanner.amundsen/viz/AmundsenEN_605_662FinalProject-Dashboard2/RevenuebyActorandKeyword) – This dashboard shows the relationship between film content, cast, and revenue. The TMDB dataset included a list of keywords for each film which gives a glimpse into the actual content of the film. These keywords are represented in a Graphical user interface, text

Description automatically generated**Graphical user interface, application

Description automatically generated with medium confidence**word cloud, sized by count, and colored by average revenue of movies Graphical user interface, text, application

Description automatically generated**A picture containing graphical user interface

Description automatically generated**containing that keyword. Actors are represented in a similar way.

Figure - Dashboard #2 - Revenue by Keyword and Cast. Note the dashboard on the left that allows for global filtering on genre, title, keyword, and actor. Note also the “Count of \_\_\_\_” slider filters below the Keywords and Actors word clouds.

Figure - Dashboard #2 filtered by genre. This is showing only the revenue, keywords, and actors that were in animated films.

Figure - Dashboard #2 filtered by films containing the keyword "family". Note that adding filters may require changes to the slider filters for word count.

Figure – Dashboard #2 filtered by highest earning films as selected in the Revenue bar chart.

**Chart, funnel chart

Description automatically generated**This dashboard has several options for filtering and interaction as seen in Figures 11-13. It was also important to provide the user insight into which films were associated with any given keyword or actor. To that end, a data tip was added so that users can hover over a given keyword or actor and see the average revenue, count of that word appearing in the data, and the top 5 highest earning films associated with that keyword or actor (See Figure 14).

Figure - User hovering over the keyword "princess" to reveal the data tip.

**Graphical user interface, chart

Description automatically generatedChart

Description automatically generatedChart

Description automatically generatedChart, bubble chart

Description automatically generated**

Figure - Dashboard #3 filtered by the genre "Animation"

Figure - Dashboard #3 filtered by films starring Julia Roberts

Figure - Dashboard #3 - Revenue by Categorical Data. Note the sidebar on the left that allows the user to filter by genre, title, keyword, or actor.

**Table

Description automatically generated with medium confidence**[***Dashboard #3: Revenue by Categorical Data***](https://public.tableau.com/app/profile/tanner.amundsen/viz/AmundsenEN_605_662FinalProject-Dashboard3/Revenuevs_CategoricalMetrics) – The third dashboard (Figure 15) is designed to help the users quickly understand the relationship between film revenue and various categorical variables like genre, production house, and production country. Note that there is a many-to-many relationship between each of these variables. This dashboard contains the same filtering options as Dashboard #2 (see Figures 16-18) as well as the same data tips (see Figure 19). The data tips work for the genre chart, the production company chart, and the production country chart.

Figure - Dashboard #3 filtered by highest earning films.

Figure - Screenshot of a user hovering over the Warner Bros. square in the Avg. Revenue by Production Company chart to see more information.

**Results**

This dashboard identified several factors that contribute to the amount of revenue a film will earn. From dashboard 1, we can deduce that popularity and budget have a moderate, positive correlation with revenue. We can also deduce that revenue tends to increase uniformly with audience rating and runtime. However, for high extreme values of both rating and runtime, this pattern is broken and tend to make less money than values closer to the mean. Finally, dashboard #1 reveals that the best time to release a movie is late May/early June and late November/December and the worst time is late winter and early fall.

Dashboard #2 reveals that the highest earning films in terms of content tend to be a part of larger franchises specifically in the sci-fi, fantasy, and comic book realm. The keywords associated with the highest earning films were “3D”, “IMAX”, “marvel cinematic universe”, “sequel”, and “magic” and the actors associate with these films were Stan Lee, Hugo Weaving, Orlando Bloom, and others famously known to be involved in the MCU, Lord of the Rings, or Pirates of the Caribbean. These trends can change when the dashboard is filtered by genre.

Dashboard #3 revealed that the highest earning genre on average is animation followed by adventure and fantasy. The lowest earning genres on average were foreign films and documentaries. The highest earning production houses on *average* are the companies who produced franchises like *Pirates*, *Transformers*, *Fast & Furious,* and *Avatar* though these companies tend to be small and specialized. The biggest earners *overall* are Warner Bros, Universal Pictures, Paramount, Twentieth Century Fox, and Columbia Pictures. Lastly, dashboard #3 revealed that the production countries where revenue was highest on average were the Bahamas, Dominican Republic, Jamaica (all 3 for *Pirates*), Malta (*The Davinci Code, World War Z*), and New Zealand (*The Lord of the Rings*).

**Conclusion**

Many parties have an interest in analyzing what factors contribute to a movie’s commercial success. While many vast databases exist that compile data on movies, there are limited existing visualizations geared toward predicting revenue. These three Tableau dashboards give users a way to understand the relationship between revenue and various other metrics of interest. This dashboard was used to answer each of the sample analytical questions in the background section. This dashboard is also flexible and interactive, including many different kinds of filters and data tips to suit the user’s specific needs.

**References**

A. Blecksmith, M. Mazzola, K. Pellegrino, and J. Schmidt, “An exploration of great cinema using information visualization ...,” *Drexel.Edu*. [Online]. Available: http://cluster.ischool.drexel.edu/~cchen/courses/INFO633/07-08/g4.pdf. [Accessed: 10-Apr-2023].

D. Haughton, M. -D. McLaughlin, K. Mentzer and Changan Zhang, "Movie analytics: Visualization of the co-starring network," 2014 IEEE 4th Symposium on Large Data Analysis and Visualization (LDAV), Paris, France, 2014, pp. 115-116, doi: 10.1109/LDAV.2014.7013216.

Drucker, J. (2011). Data visualization in digital humanities. Digital Humanities Quarterly, 5(1).

M. Mestyán, T. Yasseri, and J. Kertész, "Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data," PLoS One, vol. 8, no. 8, p. e71226, Aug. 2013. [Online]

Rajamanickam, V., Ramey, J., & Chatham, C. (2021). Using interactive data visualization to explore non-linear movie narratives. In Proceedings of the 54th Hawaii International Conference on System Sciences (pp. 3916-3925). IEEE. doi: 10.24251/HICSS.2021.477

The Movie Database (TMDB). 2017 Apr. “TMDB 5000 Movie Dataset”, v.2. Retrieved 2023 Apr. 29 from <https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata?select=tmdb_5000_movies.csv>