Movies: A Closer Look

Mayra Weidner

Introduction

My goal with this project was to analyze movie data from every genre and explore profitability, budget, revenue, popularity, actors, and directors. I used Tableau to create all my visuals and dashboards and Excel to clean up and merge the dataset. A visual approach is most appropriate for this type of analysis as it allows users to explore topics of interest and share their findings in a quick, easy to understand format. The target users for this analysis are lay people, movie executives, and content providers (i.e., video subscriptions, TV, etc.,). I expect that a lay person would use this analysis as a form of entertainment and strictly be interested in the visuals for their own enjoyment and curiosities. Movie executives could use this analysis to determine if the market would be accepting of the movie they wish to produce. Content providers such as Netflix or Hulu could use this analysis to determine which movies are the most popular and would attract the most users to their platform.

Dataset

I combined two different datasets to create a master movie file in Excel which I imported to Tableau. I downloaded The Movies Dataset and American Film Institute, Sight and Sound Critic's Survey Dataset from www.Kaggle.com.

The Movies Dataset is approximately 300 MB and includes five CSV files with metadata for 45,000 movies released on or before July 2017. The CSV files include data points such as cast, crew, plot, keywords, budget, revenue, etc. This dataset is an ensemble of data collected from The Movie Database and GroupLens. The American Film Institute, Sight & Sound Critic's Survey Dataset contains a sample of over 1,000 movies considered “the greatest films” and “films to see before you die”.

I used the Movies Metadata file as a starting point as this file contains the most attributes and pulled in additional fields from the other CSV files as needed. I chose to exclude the Adult, Belongs to Collection, Homepage, Poster Path, Video, and Spoken Language attributes as these did not add value to my analysis. I cleaned up several formatting issues and split some fields as some of these columns included more than one set of data in a field (i.e., several genres or actors in one cell for a single movie). Of note, some attributes include null values or non-sensical entries which are evident in some of filters available across the visualization system. These entries are few and far between and do not affect the integrity of the data overall.

My master movie file is a flat static table, and each row or item consists of the data for one movie. There are no duplicated movies in this file. Each movie has its own set of Movies and IMBd identification numbers which I used to pull in attributes from the various CSV files. Most of the attributes in the file mean exactly what they are titled (i.e., budget is budgeted dollars). The Title attribute is one of the few fields that can be left open to interpretation. This attribute means the "official" title of the movie. There are both categorical (i.e., genre) and ordered attributes in the data. Ordered attributes include both quantitative (average rating) and ordinal (i.e., popularity) data.

Tasks

- *Present:*

  A user of my visualization system can use the results of their analysis to present a story to a third party. For example, a movie executive can present the Budget Versus Revenue visual to investors to determine whether it is a wise financial investment to follow up a movie with a sequel. Content subscription providers could also use my visualization system to determine if there are specific genres that their subscribers would be more interested in or in movies that were highly rated and present their findings to decision makers.
- *Enjoy:*

  Users of these visuals may be driven by curiosity rather than a need to verify or generate a hypothesis. These users can utilize my visualization system to deep dive into actors, directors, and movies they like or have interest in. Some of the dashboards include a Movie Details table that a user can utilize to obtain additional movie details once they have filtered the various options. Users may also get ideas for which movies to watch next based on their exploration.
- *Trend:*

  The Genre Profitability Over Time Dashboard shows increases, decreases, and peaks in profitability by genre through the years. Users of this visual can explore what drives these changes. For example, if a user is interested to see what drives the spike in the action genre in 2015, the user can click on the spike and the visual will pull up a bar chart that shows every movie that was released in that year and the profit made.

- *Compare:*

  The Revenue Versus Budget Dashboard shows a comparison that users can deep dive by using filters. The visual also provides additional movies details to allow users to further explore specific movies.
- *Search:*

  All visuals include one or more filters that allow users to drill-down into the data. For example, the Director Analysis Dashboard and Genre Profitability Over Time Dashboard provide further drill-down capabilities using multiple views.
- *Lookup:*

  The Genre Profitability Over Time Dashboard and the Director Analysis Dashboard allow users to further explore points of interest by using the lookup features included in the dashboards.

Visual Design

Genre Profitability Over Time Dashboard:

I selected the line chart idiom to illustrate changes in profitability by movie genre over time. I believe a line chart is the best visual idiom to show one quantitative value attribute (profit) and one ordered key attribute (release year) augmented by the color channel in a clean, readable way. The x-axis shows the release year and the y-axis shows profit which was calculated based on the budget and revenue figures in the dataset. I used the line mark and color channel to illustrate the trend and individual genres are color coded using Tableau's 20 color palette. I selected this color palette because it has 20 very distinct colors. I included multiple value dropdown list filters for genre and release year that users can utilize to drill down into specific years of interest or to limit the amount of data shown. The dataset includes 20 different genres as shown in the genre legend. Of note, nine genres make up about 89% of the total ratings vote count and 90% of total profit in the dataset. These genres are drama, comedy, action, adventure, crime, fantasy, horror, animation, and science fiction. It is likely users of this visual would be more interested in filtering the data with only these genres as it captures most of the dollars and would reduce clutter and occlusion in the visual. The visual allows users to not only filter the data for genres and years of interest, but to also zoom and pan using Tableau's built-in View Toolbar. With this tool, users can zoom and pan on the entire chart or select a zoom area to further explore. Users can also hover over lines and points to see total yearly profit.

To help users further explore what movies are driving spikes (i.e., one point selection) or what movies make up a specific section of the line chart (i.e., multiple point selections), I created a multiple view dashboard that includes the visual described above and a horizontal bar chart that shows what movies were released each year. I selected the bar chart for this purpose as it shows movies and their profits in a clean list format that is ordered from highest to lowest profit which is essential when users wish to explore spikes each year. The x-axis shows profit, and the movie titles are on the y-axis. By selecting a point or several points in the line chart, users can filter the bar chart automatically to see what movies were released. Users can also hover over each bar in the bar chart to see the budget, revenue, and profit figures. To reset the charts, users can click anywhere in the white space of the line chart or the ESC button on their keyboard. Because of the size of the visualization system and user machine limitations, there may be a slight delay while the visual resets. I also included a Reset Filter button so that users can reset filters with a click of a button.

Budget Versus Revenue Dashboard:

I selected the side-by-side bar chart idiom to show a comparison between movie budget and revenue. In my opinion, the side-by-side bar chart is the best visual idiom to show a comparison between two quantitative value attributes (budget and revenue) and many marks in a simple, readable way that does not cause confusion. I used the bar mark and color channel to represent budget in blue and revenue in orange which are listed along the x-axis. I selected blue and orange as these are two very distinct colors that will likely not be confused by users. The visual also includes a legend. The y-axis shows the values in millions and the top header shows movie titles. In the labels, I included movie title, budget, and revenue that users can view when they hover over each bar. Users can change the visual by using the dropdown list filters for genres, directors, and actors and a slider filter for IMBd ratings. All filters will accept multiple selections and can be reset when users click on the Filter Reset button.

To provide users with additional details for each movie, I created a multiple view dashboard where users can click on any bar, and it will generate a table with movie title, genre, plot overview, main actors, director, and IMBd rating. Users can also hover over the last field in this table to see all the information in a single view. Users can continue to make selections by clicking on other bars or reset their view by clicking anywhere in the whitespace above the bar chart or hitting ESC. I also included a Reset Filter button so that users can easily reset filters.

Movie Popularity:

I selected the tree map idiom to illustrate movie popularity. I used the square mark to represent each movie. The size and color channels are driven by total revenue. The larger and darker the square, the higher the revenue and vice versa. I used a blue to teal color palette with five stepped colors to reduce any confusion caused by color blending. I believe the tree map idiom is the best choice for this type of analysis as a tree map allows users to quickly see which movies are the most popular as these squares will be bigger and a darker colored versus less popular movies that are represented by a smaller, lighter square. I included IMBd ratings and revenue in the label mark that users can see when they hover over each square. I also included multiple value dropdown list filters for actor, director, release year, and genre. Users can also use the slider filter to select which IMBd ratings they'd like to include. Higher IMBd ratings would be representative of more popular movies. Realistically, users can use this visual to see the least popular movies by filtering the data for low IMBd ratings.

I created a multiple view dashboard with this tree map and linked it to the Budget Versus Revenue and Movie Details visuals discussed in the section above. Users can filter the tree map as they wish. Once the tree map generates, users can click on individual squares to pull up the Budget Versus Revenue bar chart and the Movie Details table that will automatically populate. Users can continue making selections in this fashion or reset their view by clicking anywhere in the whitespace above the tree map or ESC on their keyboard. Users can also reset all filters by clicking on the Reset Filter button.

Director Analysis:

I selected the scatterplot idiom to show an analysis of budget versus revenue by director. I used the circle mark to represent each director. The color channel is driven by average IMBd ratings, and the size channel is based on number of movies by director. Budget figures are listed along the x-axis and the revenue figures are on the y-axis. I used a gold to green five stepped color palette. The stepped color palette helps reduce readability issues caused by similar colors blending. The scatterplot idiom is the best visual to show this type of analysis because a scatterplot encodes two quantitative value variables (i.e., revenue and budget) using both the vertical and horizontal spatial position channels, and the mark type is a point (i.e., director). I included a director filter that will accept multiple selections. There are also slider filters for IMBd rating, total revenue, total budget, and number of movies made that allow users to drill-down. Since each row in the dataset represents one movie, I used the Director field to create a calculated field that counts the number of movies for each director. The circle size is based on this calculated field. I also included legends for color and circle size.

To increase interactivity with this scatterplot, I created a multiple view dashboard. Users can use this dashboard to zoom and pan the scatterplot to zero in on areas of interest that are difficult to read due to occlusion. Users can either select one point or several points using Tableau's built-in selection tool. The Movie Details table previously discussed is linked to this dashboard and will provide additional background. I also linked a Movie Listing visual which I created using the horizontal bar chart idiom. This bar chart shows movies listed in descending order by revenue. I selected the horizontal bar chart idiom because it is easy to read and provides the perfect list view without overcomplicating this multiple view dashboard. These two visuals will update based on user selection in the scatterplot. Users can reset their view by clicking anywhere in the whitespace above the scatterplot or by hitting ESC on their keyboard. Users can also reset all filters by clicking on the Reset Filter button.

Results

Which genres have been profitable over time?

I filtered the Genre Profitability Over Time Dashboard to exclude all null values for both genre and release year. Action movies have seen a big increase in profits since 2006. Big spikes in profits occurred in 2009, 2013, and 2015 with the release of Avatar, Iron Man 3, and Star Wars: The Force Awakens which made $2.5, $1.01, and $1.8 billion in profits respectively. After 2015, profits for action movies start to drop off. Adventure movies have seen a steady increase in profits since 1998. In 2001, Adventure movies saw a spike in profits with the release of Harry Potter and the Philosopher's Stone and Lord of the Rings: The Fellowship of the Ring each earning $851 and $778 million respectively. Profits have remained consistent until 2016 when Captain America: Civil War was released. The most profitable movie in this dataset appears to be Avatar with a budget of $237 million and profits of $2.5 billion. Followed by Star Wars: The Force Awakens and Titanic with budgets of $245 and $215 million and profits of $2 and $1.8 billion respectively.

How does budget compare to revenue?

The Budget Versus Revenue Dashboard was designed for exploration and comparisons, so I filtered the data for genres, directors, and actors that were of interest to me. The highest rated movie with an IMBd rating of 8.8 in the comedy genre is Forest Gump. The budget for Forest Gump was $55 million and it made $677 million in revenue. Forest Gump was also the highest grossing movie in the comedy genre. Jaws was the highest grossing movie in the Horror genre with a budget of $7 million and revenue of $471 million. Christopher Nolan's The Dark Knight was his highest grossing movie with a budget of $185 million and revenue of $1 billion. This is followed by Inception with a budget of $160 million and revenue of $826 million. Alfred Hitchcock's Rear Window had a budget of $1 million and made $37 million followed by Psycho with a budget of $807,000 and revenue of $32 million. The highest grossing movie that Bruce Willis starred in was The Sixth Sense which made $673 million and had a budget of $40 million.

What are the most popular movies?

I filtered the Movie Popularity Dashboard to include movies with IMBd rating of at least 8.0. Per www.IMBd.com, the average movie rating on IMBd is 7.0. The tree map shows that The Lord of the Rings: The Return of the King with an IMBd rating of 8.9 and revenue of $1.1 billion is the highest rated and most profitable movie in this dataset. To get additional information, I clicked on the movie square and the Budget Versus Revenue bar chart, and the Movie Details table updated to show only this movie. The second most popular and profitable movie in this dataset is the Dark Knight with an IMBd rating of 9.0 and revenue of $1 billion. The smallest square in the tree map with an IMBd rating of 8.7 and $270,000 in revenue is Seven Samurai.

How do directors compare?

Alfred Hitchcock directed the most movies in this dataset. He directed 19 movies which include Rear Window, Psycho, and Vertigo. The average IMBd rating for all of Alfred Hitchcock's movies is 7.8 and total revenue is approximately $200 million. Alfred Hitchcock earned the most revenue for Rear Window which grossed $36.8 million.

James Cameron made the most revenue at approximately $5.4 billion. The dataset includes five movies directed by James Cameron with an average IMBd rating of 8.1. His highest rated movie is Terminator 2: Judgment Day with Aliens as a close second. James Cameron's highest grossing movie was Avatar which earned $2.8 billion followed by Titanic at $200 million.

Christopher Nolan is the director whose movies earned the highest average IMBd rating at 8.7. Christopher Nolan's highest rated and grossing movie is The Dark Knight with an IMBd rating of 9.0 and $1 billion in revenue.

Martin Scorsese is the director that had the highest budget at $755 million. He directed 13 movies with an average IMBd rating of 7.8. Martin Scorsese's movies earned approximately $1.4 billion. His most popular movies with an IMBd rating of 8.7 and 8.5 respectively are Goodfellas and The Departed.