INFO 250 Project 2C

Jerry Li - jl4533@drexel.edu

Lixiao Yang - ly364@drexel.edu

Mengyang Xu - mx73@drexel.edu

This is the link to our video:

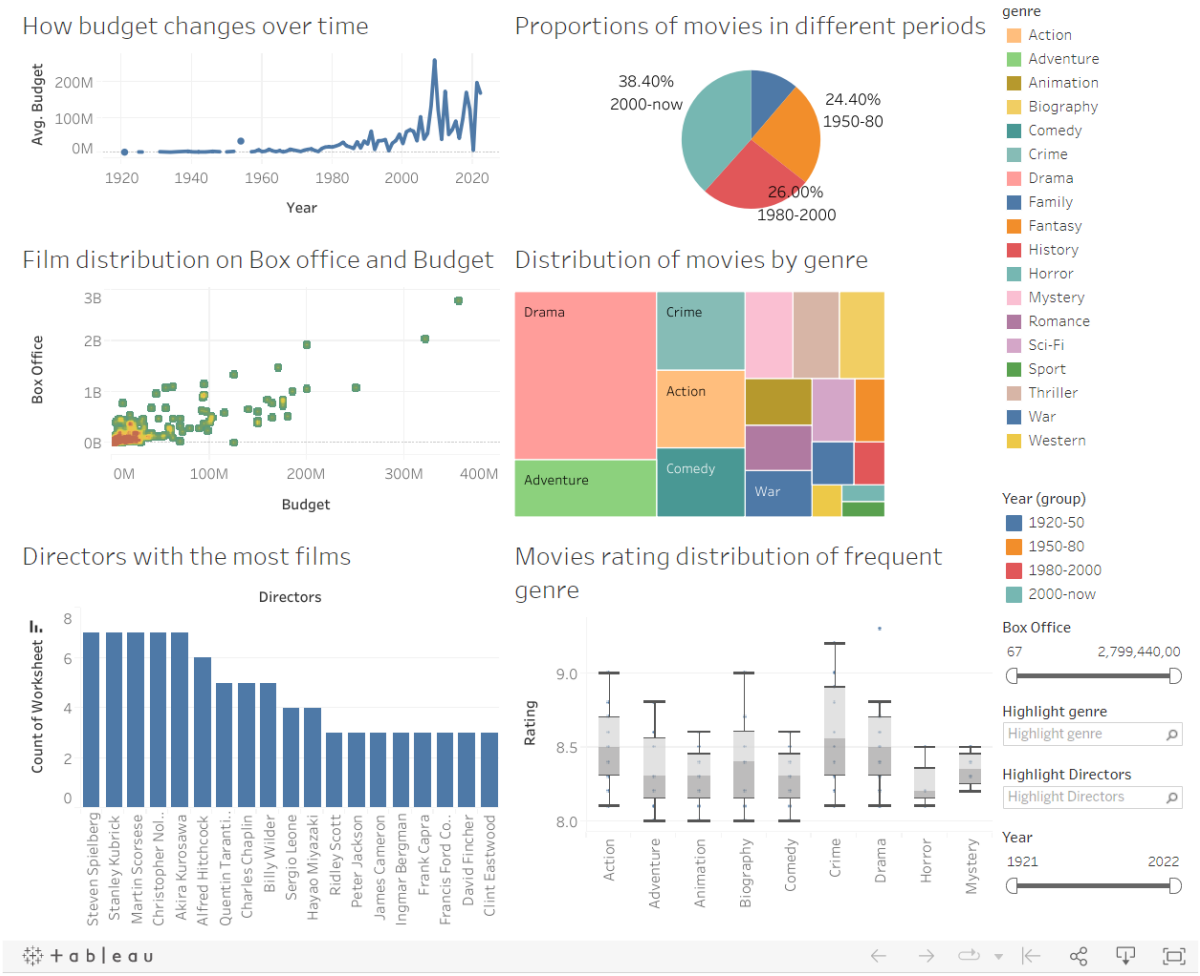
https://www.canva.com/design/DAFdTE4gAE4/Fipx1vkgKoG0jPf_RQz-5Q/view?utm_content=DAFdTE4gAE4&utm_campaign=designshare&utm_medium=link&utm_source=recording_view

Motivating questions

As movie fans, we all like to use ratings to determine whether a movie is good or not. As one of the largest movie review sites in the world, imdb undoubtedly has pretty accurate ratings. We want to analyze the relationship among a series of attributes such as film budget, box office, shooting date and director through the top 250 films provided by imdb. Through this relationship, we can better grasp the development process of the film industry, and for a film without a score, we can infer the score of the film through a series of relevant attributes. For film industry workers and fans, this can do a lot to help them.

- Movie makers and producers
    - Movie producers can know from our dataset that this dataset does not just evaluate the quality of a movie from the perspective of the audience. It also includes data that producers and manufacturers are more concerned about, such as the theme of the movie, the production cost of the movie, and the box office of the movie, which allows them to better analyze the direction and investment of their own movie production. For example, for budget and box_office in this data set, movie manufacturers can count the production costs and box office of movies of various themes to determine their own production direction and effect evaluation.
- Moviegoers
    - Moviegoers can know from the topic of our dataset that the movies in our dataset are all top-ranked movies in IMDB. Therefore, movie viewers should know that all the movies in our dataset are either with good reputation or high quality. Movie viewers may be most concerned about the rating and subject matter data in the data set. For example, viewers can find some movies with higher ratings among the subjects they are interested in from these movies.

Design decisions and justification



Design decisions

This is our final dashboard design. The idea is to reflect as much of the relationship between the attributes as possible. In our dashboard, we use a series of intuitive charts such as line charts, density charts, pie charts, and tree charts to show the relationship between different attributes.

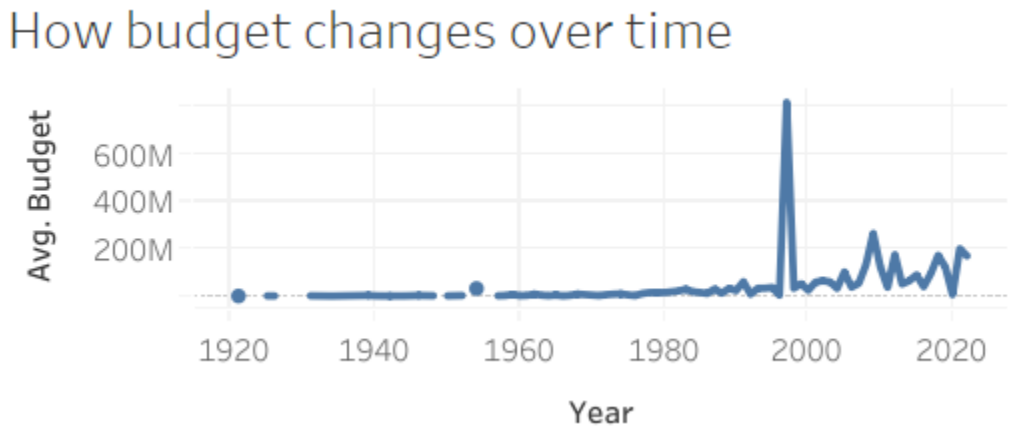
- The first chart we used is the line chart, which shows the trend of film budgets over time. We can see that over time, the budget of the film is going up.Through this line chart, we can experience the rise and fall of the film industry.
- In our second chart, a pie chart is chosen to show the proportion of films from different decades in the data set. We can see that films from the 21st century occupy the largest proportion, and at the same time, as time goes on, the proportion of films closer to the present increases. From this pie chart, we can find that the number of high score movies increases with time.
- In the third chart, we chose the density graph to show the relationship between film budget and box office. As can be seen, there is basically a positive correlation between budget and box office, which means that in most cases, the higher the budget of a movie, the higher the box office.
- For our fourth chart, we chose the tree map. We select the category of movies as input to the tree, and the number of movies corresponding to the category maps to the size of the graph. We can see that drama types occupy the largest size in the tree. From this we can infer that drama categories are relatively more likely to get high marks.
- Our fifth chart chooses a bar chart to show the number of high score works of the director. As can be seen from the bar chart, the top five directors each have seven films on the list. For those directors whose works are listed more, there is no intention to have better film creation ability.
- In the sixth chart, we chose a boxplot to show the relationship between different kinds of movies and ratings. As we can see, crime films have the highest average score of all films, but also the largest fluctuation in ratings. The mystery genre, on the other hand, has a lower draw score, but the range of scores is more stable.

Justifications

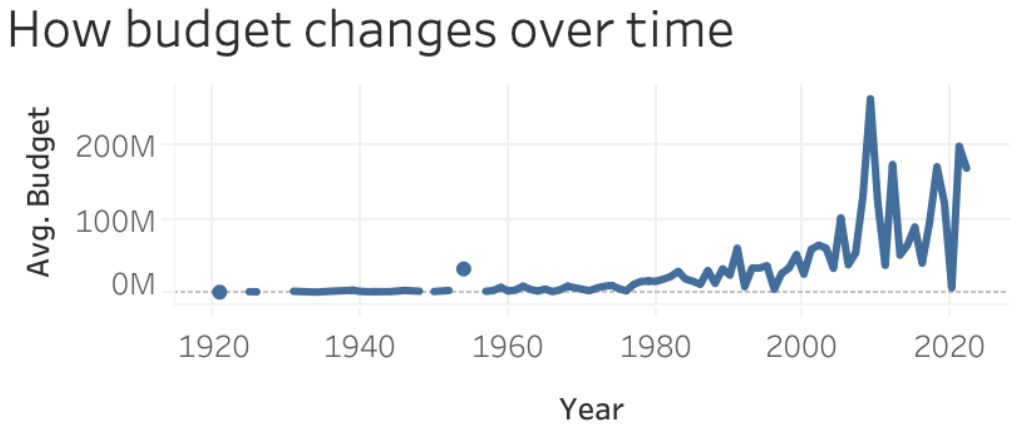
- In the process of turning data to a chart, we found some abnormal data in the data set. For example, the unit of some data in the budget was not USD, which led to the inaccuracy of the whole table. Therefore, we unified the units so that the final table can clearly reflect the content of the data.

Preliminary visualization and revised visualization

For our dashboard, our biggest correction was the change in currency units.



In the pre-revision line chart, we can clearly see that there was a peak towards the end of the 20th century. This makes little sense for the film's budget, which is unlikely to increase significantly in a given period of time. We then checked the data provided by the website and found that the budget unit of the film in the imdb database was yen instead of US dollars, which caused such an anomaly. In response, we modified the units of the data to make the resulting line graph more accurate.



Lessons learned and conclusion

Lessons learned

During the completion of this project, we went through the whole process of data visualization, from the selection of data at the beginning to the cleaning of data and drawing in the middle, which were all completed by ourselves. Through this project, we have a general understanding of the flow of data visualization. In the process of data cleaning, the biggest problem we encounter is the problem of the unit. We realized that a well-cleaned data set is very important for data visualization. Therefore, in the following courses, we will learn some knowledge related to data cleaning, so that we can better complete the visualization of data later.

Limitations

Although we work with a lot of numerical data, we can't do anything with some data that contains text. We were regret that the original data set contained part of the data about tagline, which was eventually abandoned because it was not easy to visualize. If we can keep this part of the data, it's better to keep the information that the data set wants to convey.

At the same time, due to the limitation of tableau's functions, we did not complete some functions in the end. Because we wanted to have a full dashboard, we eventually gave up using tools like ggplot. Perhaps it would be better if we could integrate visualizations from different platforms into one dashboard.

Conclusion

In this project, we understood the process of data visualization, and also clearly realized our own shortcomings. We recognize that data visualization is a complex process that takes into account many factors, not just turning data into images. In the process of data visualization, a series of factors such as story, information and readers should be fully considered. Only when various factors and steps affecting visualization are fully considered, the results of visualization can achieve the expected purpose.