INFO 250 Project 2B

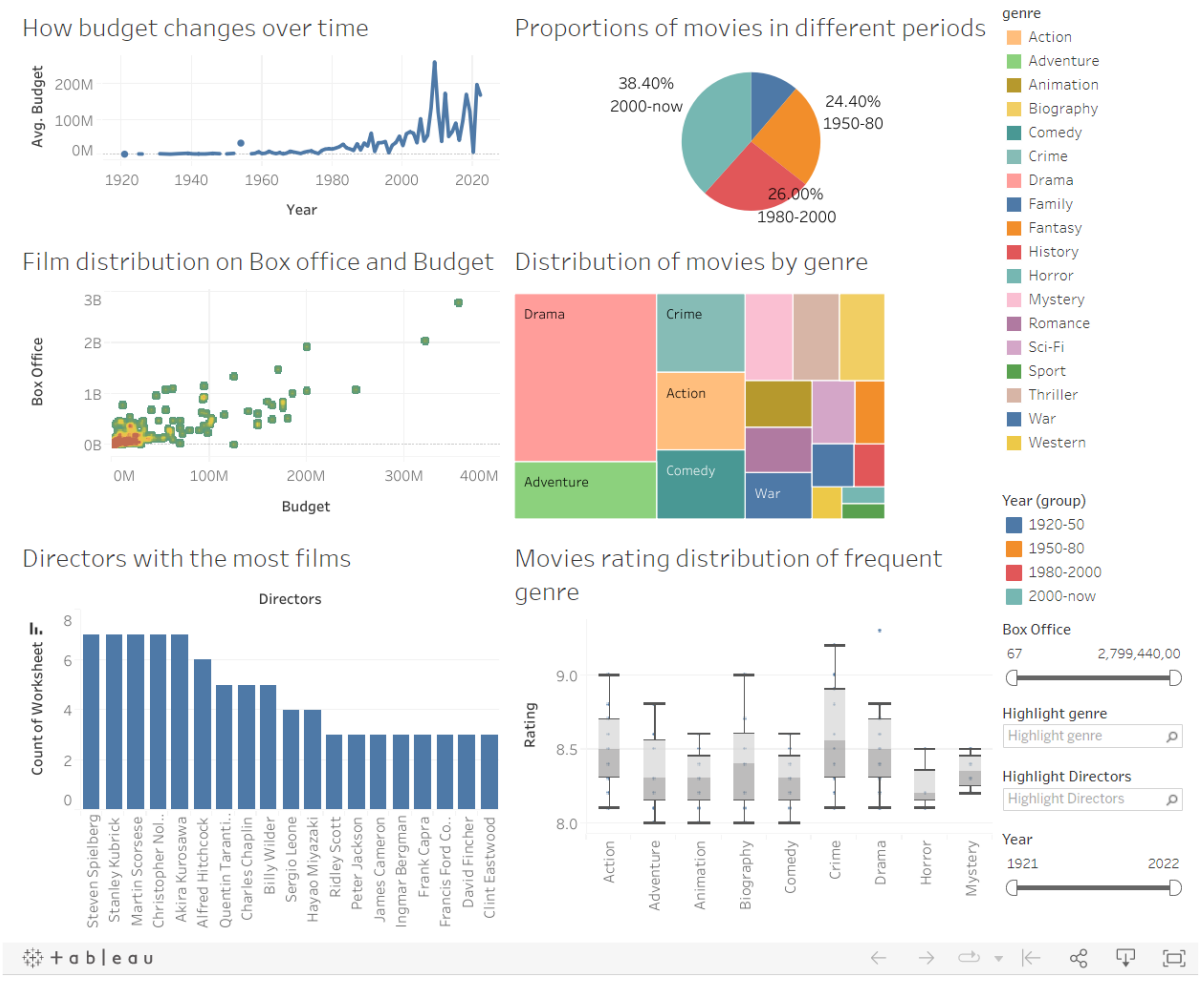
Jerry Li - jl4533@drexel.edu

Lixiao Yang - ly364@drexel.edu

Mengyang Xu - mx73@drexel.edu

Tableau Public Link:

https://public.tableau.com/views/dashboard_16778922729570/Dashboard1?:language=en-US&:display_count=n&:origin=viz_share_link



- Your visualization as they are implemented according to your plan in Project 2A. If you have to deviate from your original plan, please include an explanation of the change of plan and justify the new plan accordingly. **5 points**
    - For the question “How film budgets/length have trended in the last 100 years?”. We ultimately chose a density graph over a scatter plot because we believe it can give people a clearer understanding of the intensity of the data distribution.
    - For the question “Do longer movies have better box office collections?” The line chart's upward and downward oscillations are extremely erratic, and we discovered that there is no clear relationship between the two variables. We changed this question to "What are the proportions of movies for different times" in order to lessen user misunderstanding. Using a pie chart to answer this, showing the proportion of movies as time progresses and not being repeated in other dashboard displays.

- Outline the intended messages of the visualizations. Highlight patterns and insights that the visualizations evidently convey. Elaborate any messages, if any, that may be overlooked or not obvious from your visualization. **5 points**
    - Patterns and insights
        - For the imdb dataset, we created a dashboard consisting of six charts to represent the patterns and insights of the dataset as graphically as possible.
        - 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.
    - Questions we have faced and solved
        - The biggest problem we have is data cleansing. 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. At the same time, we also 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.