COMP6234 - Data Visualisation

Report for “What kinds of movies are more likely to get high score?” Data Story

Hang Su

ID: 30005019

hs1a18@soton.ac.uk

I. Data Story Summary

In our daily life, there are lots of movies released around the world every day. However, it is hard to know if it of high quality or low quality unless it has been released for a long time, this can cost one to two years sometimes. But to watch the excellent movies the first time can be the best experience. In this data story, it analyses 5000 movie data from IMDB and tries to address the following questions:

A. What genre of movies gets high scores most?

B. Which countries produce high score English movies most and whether only countries using English as mother tongue can produce high score English movies?

C. Figure out preferences for movie genres of female and male audiences of different age groups.

D. Try to explore if there is relationship between budget and score of a movie.

People always feel disappointed when they choose a movie of low quality and it is a waste of our time. Thus this data story is intended to summarize some common features of high score English movies released between 2010-2016, which may help people to predict the quality of a movie suitable for the taste of a specific group of people. It is also worth to mention that all of the charts provided in this story are designed to be interactive.

II. Dataset Summary

A. Dataset

I used two datasets in this coursework:

I used two datasets in this coursework:

The first dataset is “Top rated English movie from 2010-2016 from IMDB”, it is retrieved from data.world website. For my project, this dataset mainly provides data of average score of different movie genre given by female and male audiences of four different age groups. And this dataset has filtered the English movie data between 2010 and 2016, this is quite important. I only want to focus on analysing movies in English cause there exist big gap between tastes of different culture. As for the reason narrowing the time in 2010-2016, it is because the preferences of audiences change with time passing, only analysing data gathered merely a decade before this year is reasonable. Meanwhile, I also notice that the score of movies released in 2017 and 2018 still need time to be stable.

The second dataset is “IMDB 5000 movie dataset” downloaded from Kaggle. This dataset contains more movies and gives more features of every movie, such as director name and country. I use this dataset to add some features I want to analyse but absent in the first dataset.

B. Data Processing

In this part, I mainly used pandas and collections packages of Python to clean and process the data.

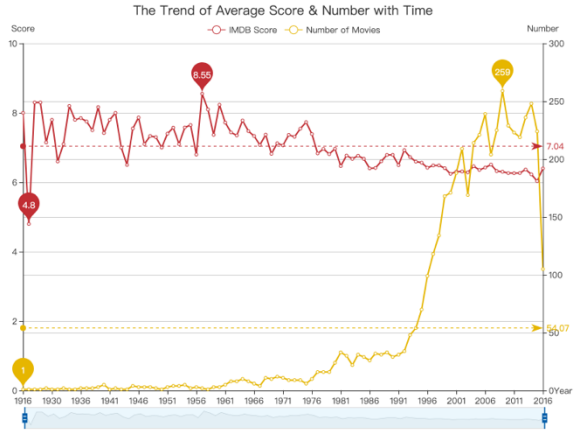
When answering the question C and D, I need to link these two datasets. The common feature in them is the “movie_title” column, but the movie title in the first dataset followed by a time string, thus I need to pick out the movie title using regulation expression and then merge them to get a dataset includes all the features provided by these two datasets. Next, I delete some features I will not use in this project to minimize the size of the dataset as well as rename these columns with my own habit.

But the most difficult part is reorganizing the data to make every row contains only one genre and store them to JSON files when addressing sub-questions A, C and D. Thus I used defaultdict of collections package of Python to store each row of data.

Finally, I got three JSON files for questions A, B and D and three JSON files for question C, they are all stored in the data folder.

III. Visualisations

A. The Trend of Average Score&Number with Time



1) Description

The line graph above illustrates the decreasing trend of movie score and the increasing trend of number with time suitably, especially after 1960. Animation of the chart enhances the visual experience when the user drags the bottom toolbar.

2) Justification

In this part, I intended to show the trend of average IMDB score and the number of movies with time. The data I will use is time series data, line chart is always the best choice to illustrate this. Besides, because the time range is in a large scale, thus I add a toolbar which can set the time range of data to display on the chart to help the audience to observe the trend in a short term.

3) Narrative Design Patterns

This graph uses silent data pattern to emphasise the decreasing trend of the data. What's more, the toolbar makes the chart use exploration design pattern, allowing audience explore trend in different time periods.

4) Strengths and Weaknesses

Strengths: Line graph is the best choice to show data over time. I zoomed in the y-axis because the data I used starts above zero [3]. Allow audience zoom in or out the graph to highlight part of the data trend.

Weaknesses: I only showed the trend based on the data of IMDB without comparing data from other movie platforms. Maybe the decreasing trend of score only exists in IMDB.

5) Improvements

To improve this chart, I can grab movie data of other movie platforms of the same time period and add lines of them to test if the decline of the quality of movies is common in any platform. If I do this, the chart will use the repetition design pattern.

B. Number of Genres

1) Description

In this word cloud graphic, audiences can easily recognize different genres without a legend. Because I set a suitable size of the minimum word and set different word of different colours to make every word clearly on the panel.

What's more, when users put their mouse on a specific word, there will occur a light yellow shadow to emphasise the genre they select as well as a tooltip box to tell them the certain number of this genre in the dataset.

2) Justification

The intention of this word cloud graphic is to show the proportion of each movie genre in this movie dataset. In general, people use pie charts to compare the proportion of different parts. But in this project, as you can see, there are 19 genres to be displayed, which means there will be too many slices to recognize each genre clearly. Thus word cloud is a better choice to visualise the frequencies of movie genres.

3) Narrative Design Patterns

In this graphic, I used comparison design patterns. Because I plan to compare the proportion of every genre in the dataset and highlight the genres who take the most part of the movies, thus I choose to use comparison in this chart.

4) Strengths and Weaknesses

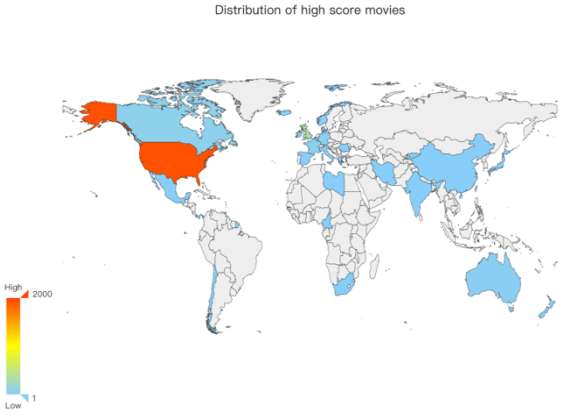
Strengths: Word cloud is a simple and intuitive visualization technique. It is often used to show the most frequent words of a text as a weighted list [2]. All these characters word cloud has is suitable for what I want to show of the movie genres.

Weaknesses: The human brain is not sensible to size or area. For example, look at Comedy and Biography in this word cloud graphic, the word biography looks longer than comedy when there is only a little difference between their sizes, but the number of comedy is more than that of biography in this case.

5) Improvements

Add a tool which can transform the word cloud to a stacked bar chart. It will allow users to choose the way to explore the data themselves.

C. Distribution of High Score Movies



1) Description

In this part, I want to show which countries produced high score English movies most in a direct way. Display these data on a world map is the most familiar way for people.

2) Justification

About this graphic, what I want to show are geographical information and count details. For spatial data in this dataset, showing it on a map is the most familiar way for people. As for showing the count details, we usually use circles or colours to illustrate, but some data in this case is quite small and a lot of data is centralized in Europe, which means the circles will overlap and cannot be distinguished clearly by audiences, thus I choose colour hue to represent the count of movies of each country.

3) Narrative Design Patterns

I use silent data pattern to emphasize that the majority of high-quality movies are produced by America and UK. The

map also can be said using the familiarisation pattern because world map is a common elements in our daily life.

4) Strengths and Weaknesses

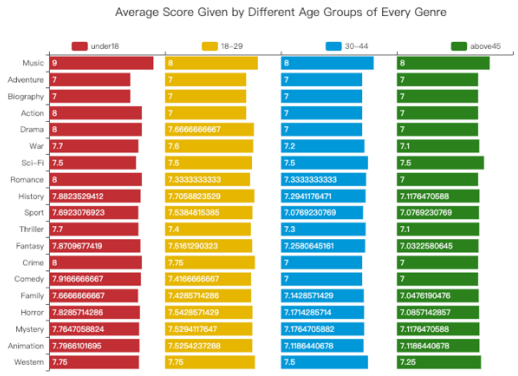
Strengths: Showing the geographical information of high score movie in a way audience familiar with. And the color is a pattern human sensitive to. Also, this map will highlight the area and prompt a tool-tip box of the country that audiences want to check, which increase the sense of engagement.

Weaknesses: A large part of the data are centralized in a small range, which leads to a result that lots of countries have a similar hue on the map. It is hard for audiences to separate them.

5) Improvements

The most important issue needed to be resolved is finding a more suitable visualisation chart which can distinguish movie data in a small range. Besides, there will throw an exception when users put the mouse on a country does not contain data, I will try to fix this problem in the future.

D. Average Score Given by Different Age Groups of Every Genre / Average Score Given by Female of Every Genre / Average Score Given by Male of Every Genre



1) Description

In these three bar charts, each rows represents average score of one movie genre given by four different age groups, each column stand for the average scores of all genres given by a certain age group. The label on the x-axis is designed as a filter to make it convenient to hide some data of a age group.

2) Justification

In this part, I plan to design three charts which can allow the audience to compare scores of all genres within an age group as well as to compare them of all age groups within a movie genre. Thus two-dimensional bar chart can help me achieve this purpose.

3) Narrative Design Patterns

These three bar charts mainly use comparison pattern, the only purpose in this part is to draw some conclusions by comparing the data horizontally and vertically.

4) Strengths and Weaknesses

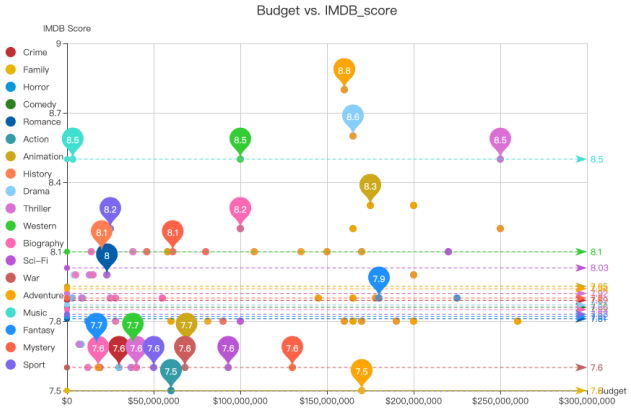
Strengths: The interaction functions are the main strengths of them, because these interactions can help audience compare the data more conveniently.

Weaknesses: The difference between each two scores is small, thus the lengths of the strips are similar.

5) Improvements

I will learn the tools I used in this project more detailed to try to change the scale of these three charts to make it easier to observe the difference.

E. Budget vs IMDB Score



1) Description

This scatter graph shows the distribution of data points stands for the combination of information of budget and score of each high score English movie from 2010-2016. Meanwhile, it provides necessary tools for audience to explore this data.

2) Justification

In general, scatter plot is used to check the potential relationship between two variables, thus I generate a scatter chart to show the relationship between budget and score of every genre. After displaying all the data points and marking the average score, maximum score and minimum score of every genre at the same time, it is not easy to look all these data clearly, thus I add zoom tool and set the legend as a filter to help audience to explore these data points more clearly.

3) Narrative Design Patterns

This scatter graph uses users-find-themselves pattern because the chart will be too complicated if I display the regression lines. Let users find the answer themselves is a better choice in this situation.

4) Strengths and Weaknesses

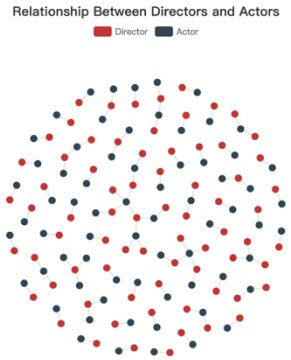
Strengths: This chart shows all of the information about the budget and score one time and adds useful tools to avoid the potential problems. "Overview first, zoom and filter, then details on demand." [4]. According to this rule, I add necessary tools in this graph.

Weaknesses: It is too complicated when I display the linear or polynomial regression lines of every genre, I still do not know how to use the legend as a filter to hide some of these lines, thus I did not display them in the end.

5) Improvements

Try to resolve the problem that the legend filter cannot filter the regression lines.

F. Relationship Between Directors and Actors



1) Description

This relationship force graph shows the collaborative relationship between directors and actors.

2) Justification

I tried circular relationship graph at first and found that the connection lines are all crossed together. That's why I replace it with force graph.

3) Narrative Design Patterns

This force graph uses users-find-themselves pattern. Relation chart itself can show links of nodes clearly enough, what I need to do is giving necessary information when audience interact with the graph.

4) Strengths and Weaknesses

Strengths: It is easy to find the optimal combination of directors and actors.

Weaknesses: Recognizing people only by their name is a little bit annoying.

5) Improvements

Try to use "people behind it " pattern in the future, for instance, replacing the nodes with the photo of the directors and actors.

IV. Technologies

1) Echarts: A data visualisation tool developed by Baidu, a China technology company. I used this tool to generate the charts in my project.
2) Python, Pandas: I use them to clean and process the dataset to generate the data I will use in each chart.
3) JQuery: I use it to load the data files cleaned.
4) HTML+CSS: I use HTML and CSS to develop and beautify the frontend story pages.

Fullpage.js: This is quite suitable to show a story step by step, thus I use it to strengthen the feeling of users and add some animations to my pages. To be honest, it needs to get GNU GPL license to use this tool in open source project. I have pushed my project to GitHub and I will resolve the license problem after submitting this coursework.

V. Conclusion

In summary, this data story use one line chart to give the background information to prove the value to do this analysis. Then using word cloud to show the movie genres having most occurrence, displaying the countries producing most high-quality English movies on a map, concluding different preferences of different genders and various age groups through three two-dimensional bar charts and trying to explore the relationship between budget and score. Finally, I think I can find more common features which can label high-quality movies such as different combinations of director and actors and so on.

References

[1] B. Bach, M. Stefaner, J. Boy, S. Drucker, L. Bartram, J. Wood, P. Ciuccarelli, Y. Engehardt, U. Köppen, and B Tversky. "Narrative design patterns for data-driven storytelling." In Data-Driven Storytelling, N. H. Riche, C. Hurter, N. Diakopoulos, and S. Carpendale, Eds. CRC Press, USA, 2018, ch. 5, pp. 107–134.

[2] Lohmann, S., Heimerl, F., Bopp, F., Burch, M., & Ertl, T. (2015, July). *Concentri cloud: Word cloud visualization for multiple text documents*. In Information Visualisation (iV), 2015 19th International Conference on (pp. 114-120). IEEE.

[3] Nathan Yau, *Data Points: Visualization that means something*. 1st ed. Wiley, 2013.

[4] Shneiderman, B., 1996, September. *The eyes have it: A task by data type taxonomy for information visualizations*. In Visual Languages, 1996. Proceedings., IEEE Symposium on (pp. 336-343). IEEE.

[5] Li, D., Mei, H., Shen, Y., Su, S., Zhang, W., Wang, J., ... & Chen, W. (2018). *ECharts: A declarative framework for rapid construction of web-based visualization*. In Visual Informatics