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](mailto:hs1a18@soton.ac.uk)

# Data Story Summary

In our daily life, there are thousands 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 try to address the following questions:

1. What genre of movies get high scores most?
2. Which countries produce high score English movies most and whether only countries using English as mother tongue can produce high score English movies?
3. Figure out preferences for movie genres of female and male audiences of different age groups.
4. 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 for specific people. It is also worth to mention that all of the charts provided in this story are designed to be interactive.

# Dataset Summary

Here you should briefly summarise the data sources you used, describe where you retrieved them from, describe how you linked together multiple datasets (if applicable), and describe any processing you needed to do to the data.

## Dataset

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.

## Data Processing

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

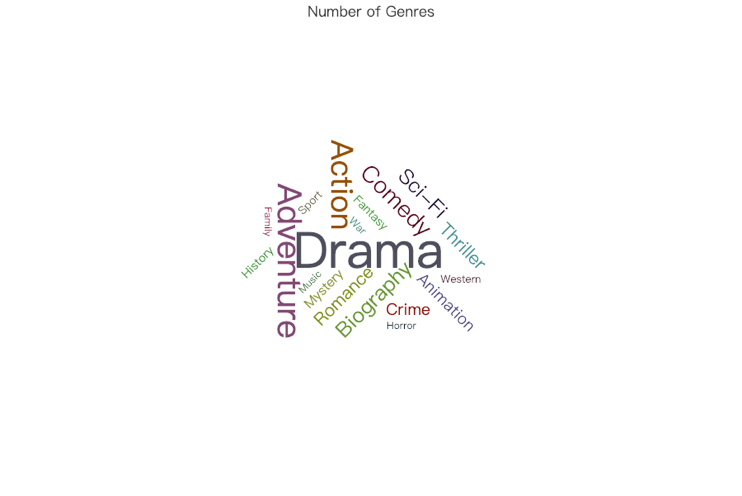
When answering the question C and D, I need to link these two datasets. The common feature in them is the “movie\_title” column, but the movie title the first dataset followed by a time string, thus I need to pick out the movie title using regulation expression and then merge them to get a dataset includes all the features provided in them. 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.

# Visualisations

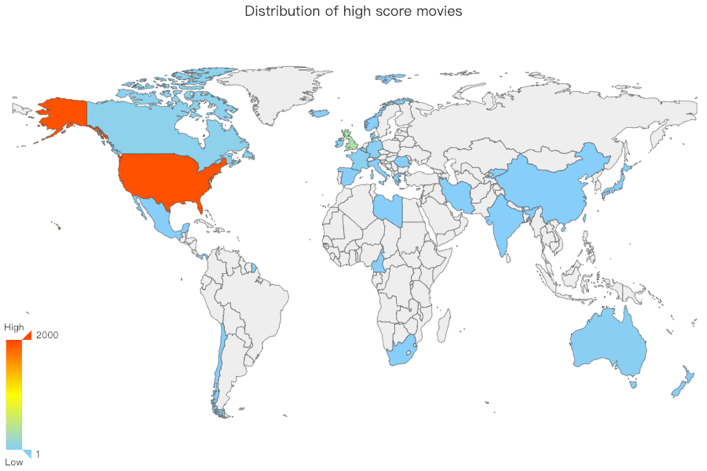
## Number of Genres



### Description

Provide a brief description of the visualisation you have used (include an image or screenshot). Describe how interaction has been used appropriately.

In this word cloud graphic, audiences can easily recognize different genres without a legend. Because I set 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.

### Justification

Discuss why the chosen visualisation is suitable for your narrative, and the data. Describe how the visualisation makes the data readable by the intended audience, avoids bias or confusion, and emphasises the key data. You should justify why (and how) you are using each visualisation with regard to concepts covered in the course; make sure you cite any relevant papers.

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.

### Narrative Design Patterns

Describe which narrative design patterns [1] your visualisation uses, describe why you chose to use them, and how they support your story.

In this graphic, I used comparison and exploration 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.

### Strengths and Weaknesses

Describe the strengths and weaknesses of your visualisation; again, make reference to concepts covered in the course (making sure you cite any relevant papers).

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 to 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.

### Improvements

Describe how you could modify your visualisations to make them even better suited to their task, and how they could tell your story better. What additional data sources could be incorporated to improve the story?

## Distribution of High Score Movies

### 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.

### Justification

About this graphic, what I want to show are geographical information and count details. For geographical data, 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 can not distinguished clearly by audiences, thus I choose hue to represent the count of movies of each country.

### Narrative Design Patterns

Exploration, familiarisation

### Strengths and Weaknesses

Strengths:

Weaknesses:

### Improvements

## 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

### Description

### Justification

### Narrative Design Patterns

Silent data, Compare

### Strengths and Weaknesses

### Improvements

## Budget vs IMDB Score

### Description

### Justification

### Narrative Design Patterns

Users-find-themselves

### Strengths and Weaknesses

### Improvements

# Conclusion

Conclude your report by summarising your data story, the key features you tried to get across to your audience, and how your visualisations accomplish that. You can also include any lessons you feel you have learned through completing this course.

##### References

References should be cited in text as consecutive numbers, within brackets (i.e. [1], [2], [3], etc.). References should be in [standard IEEE style](https://www.bath.ac.uk/publications/library-guides-to-citing-referencing/attachments/ieee-style-guide.pdf). Make sure to properly cite **all** work that is not your own. An example reference section is show below.

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.