Info 5110 Project 3

Written Description

Data source:

Dataset 1: http://www.the-numbers.com/movie/budgets/all
Dataset 2: http://www.omdbapi.com/

Data collection and organization:

The movie budgets data is acquired from the first website. However, since every piece of the data is encapsulated in the html table, it is not in a suitable format for us to utilize. Therefore, the BeautifulSoup Python module was adopted to format the data and generate corresponding csv files. Because the movie budgets data somehow lacks other relevant information about a movie that are very essential for the purpose of this project, such as genre and rating, we've found a hosting website of open movie database which basically provides an API to request its database. The API enables search by movie titles. Once a movie title has been entered through invoking the API call, the returned JSON object will cover every aspect of information that we need about a certain movie. The standard API call is in the format of URL request, therefore, it's pretty handy to invoke the API call in our Python code. And since the response is in JSON format, we can convert it in a Python data structure called dictionary, and then extract whatever information we want from the dictionary. Since the volume of movie records is relatively huge, we optimized the Python code to fully explore the principle of parallelization by invoking the API call concurrently to maximize efficiency. The resulting csv files include the ones each labeled by year, and the one that contains all the movie records from the year 2010 to 2015, which can be further utilized by the D3 Javascript code.

Visualization 1:

- **Data use:**

The first visualization is expected to have return rate (raw rate and the logarithm of the raw rate) as the y-axis and the year as the x-axis. The original dataset obtained from the webpage contains 5122 observations.

Two criteria were applied to further clean up the data. As the purpose of the Viz 1 is to show the trend of investment return rate over years, movies with budgets but zero “worldwide gross” (because they have not been released) were eliminated from the dataset. Secondly, when we were trying to draw the regression line, eight observations with extremely high return rate (>180) were eliminated because the inclusion of them will significantly skew the regression.

4737 observations entered the final dataset. Two new variables were created: the raw rate, “yearMean” is the division of the variable “worldwidegross” by “budget”; the variable “logYearMeanRate” is the logarithm transformation of “yearMean”.

- **Mapping from data to visual elements**:

Initially, time scale was used for the x-axis since x-axis is supposed to display time information, but that later turned out to be a big challenge when we were trying to draw the regression line. Regression function requires data of numerical type, but “time” is not. To circumvent that, “year” was extracted from the variable “release date”. Once we have “year” and the “return rate” in numbers, the regression model was successfully implemented. Both axises used linear scale in mapping.

Visualization 2:

- **Data use:**

The variables used here are the initialData, which has all the movies of all genres and years in it. The data is then reformated in initialData and stored in json format where the key is the top rating number, eg- top5, top15, top50 . In our dataset, there are few entries of the movie that have just released or not released hence the imdb rating of those movies were “N/A” or “None”, we had to filter out such data from our dataset.

- **Mapping from data to visual elements:**

We wanted to know the trend for the return rate for the top rated movies and wanted to check out if we observe any trend. Hence, we have a filter over the genres which will plot a graph of budget vs the worldwide gross income. Initially, for this visualization you would see all the top 50 rated movies from all the genres.

The x-axis represents the budget of the movie and the y-axis is the worldwide gross income acquired. We used logarithmic scale for both x and y axis. The movies are represented by circles and the size of the circle indicates the imdb rating (max: 10) of the movie. If the rating of the movie is higher, the radius of the circle is more and vice versa. The color of the circles represents the genre of that movie.

We have added a dropdown menu which you would select out of the ten genres - Action, Adventure, Animation, Comedy, Crime, Drama, Horror, Romance, Thriller, Documentary.

The data points of the movie can be filtered to show the movies of particular release year. At the same time, we can show the movies that are in the top 50 ratings. That is, if we have selected Top-50, it will show us only 50 movies that are high rated in the chosen genre. If we have selected top-5, we will get highly rated top 5 movies. There are chances that after applying filter, the number of movies among the filtered option is very small. Example, if you select documentary genre for year 2010, we have just few movies of that category for that year. So, hence even if we select top 50, it will show us all the points for the movie which are less than 50.

The next element in this visualization is that when you click on a circle, it will give you more information about that movie. When you click on the circle, that circle also itself gets highlighted.

The story:

We were trying to explore the investment return rate in movie industry. The first visual shows a decreasing trend in return rate over the years. Especially when modern entertainment starts to take off around the 1980s, the return rate of movies seems to be quite stable: around 10 or less.

In the second visual, we focused on the highly rated movies in different genres, and our assumption is that these should be quite profitable movies since they are highly rated by the audiences.

What surprised us was the top rated “horror” movies seem to be the safest bet in terms of return rate. All the “horror” movies are crowded towards the upper and left side in the Cartesian plane. “Romance”, “Documentary” appear to risky choices in investment. The data points seem to be quite dispersed, suggesting there is a high chance of low return rate despite high imdb ratings.