Information Visualization

Blockbusters

University of Aveiro
Bruno Assunção – 89010
Rodrigo Ferreira – 104737

Abstract

This is an application for visual data exploration, containing multiple visualizations using data regarding blockbusters (extremely successful movies), its goal is to help a possible investor make an informed decision on which movie distributor to invest in, and to show in what areas different distributors and genres find success.

I. MOTIVATION AND OBJECTIVES

The movie industry is a very interesting subject and since we found such a good dataset, we thought it would be a good idea to use it and make it the subject of our project.

Our goal is to visually explore and demonstrate various phenomena.

II. USERS AND QUESTIONS

It was a bit difficult to find a target user in the beginning, but we settled on making our application for those interested in investing in the movie industry, though the visualizations can be used for other purposes, to satisfy the curiosity of those interested in the subject.

Characterization of the users and their context

Our target users are those considering investing in the industry, mainly in a distributor, by analysing the public's preferences and distributor's capabilities.

Questions to Answer

With our data, and investors in mind, we tried to think of what questions would a new investor in the business have. Deciding where to put money in such a big industry can be a very overwhelming decision to make, so we tried to help by asking simple but important questions and providing straight-forward, easy to interpret visualizations.

With the main questions being:

- How do the public's genre preferences change over the years?
- Which distributor is better at which genres?
- How much money does each distributor make on average worldwide and how does it relate to its average budget (maximizing money earned while minimizing budget)?
- Which genres are most profitable for each audience rating (*MPAA*).

III. DATASET

We found our dataset on Kaggle (https://www.kaggle.com/narmelan/top-ten-blockbusters-20191977). It contains 430 entries, each representing one of the 10 highest grossing films per year, from 1977 to 2019.

It contains a lot of useful data, though we did not use all of its attributes in our visualizations.

The most important attributes for our project were: Title; release year; distributor; main genre; second genre; worldwide gross; budget; *IMDB* rating and *MPAA* rating.

IV. VISUALIZATION SOLUTION

Since we had quite a few questions to answer, we wanted to visualize some different phenomena, so we came up with a lot of ideas for charts, made a low fidelity application prototype (simply with pen and paper) and made our decisions considering other colleagues' feedback.

Low fidelity prototype and user feedback

We started with 5 visualizations in mind, 4 for the main purpose of aiding the investors and 1 just to bring clarity over the data at hand, so we decided to have 2 pages, one with the main charts and another to allow the user to explore the dataset.

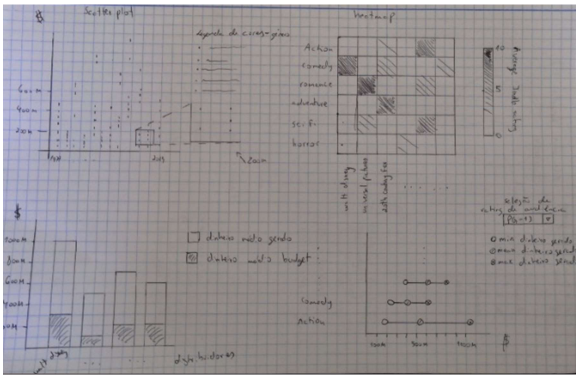

Figure 1- Prototype of the main charts

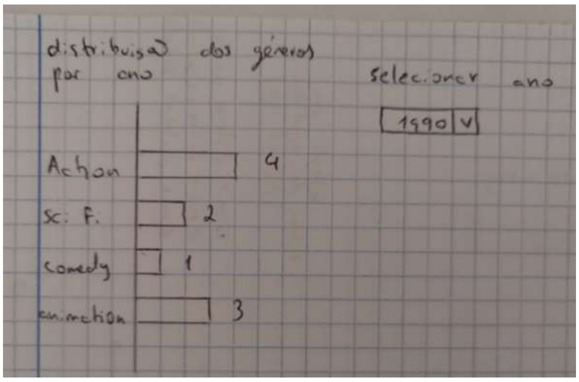

Figure 2 - Prototype of the data exploration part.

In our initial design, figure 2 would represent the data exploration section of our application and figure 1 would be the main chart, investor section.

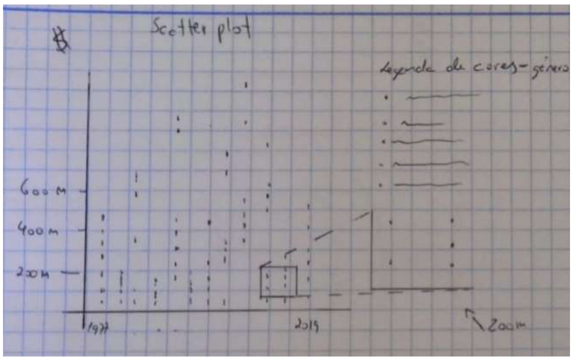

Figure 3 – Scatter plot

The scatter plot of figure 3 maps every film in the dataset within a chart where the x axis represents the movie's release year, and the y axis its worldwide gross in millions of dollars. It is meant to show the user how the blockbuster market evolved over the years, the difference in revenue values both across the years and within each year as well. Along with a different colour representation for each cell, depending on the movie's genre, we can clearly see how the public's genre preferences have changed over time.

There is also a zoom feature to facilitate the visualization where there might be a lot of cells cluttered.

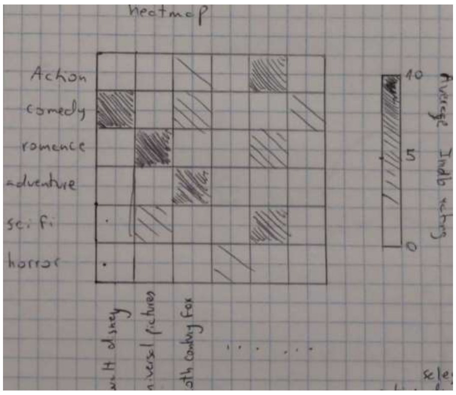

Figure 4 - Heatmap

The heatmap of figure 4 displays the average *IMDB* rating for each (genre, distributor) combination, displaying higher values with darker tones.

Its goal is to show the user which distributors excel at which genres to facilitate the investor's decision based on their preferences.

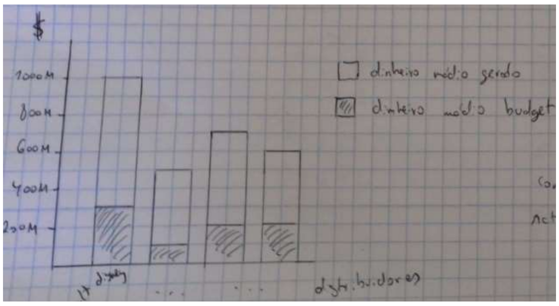

Figure 5 - Stacked bar chart.

Figure 5's stacked bar chart displays for each distributor both its average worldwide gross and its average budget (across all its movies on the dataset) to give the user a better idea on how each distributor spends their money and the average money generated, it answers questions such as:

- Which distributor earns more money on average?
- Which distributor spends more money on average?
- Which distributor has a better budget to gross ratio (good use of money)?

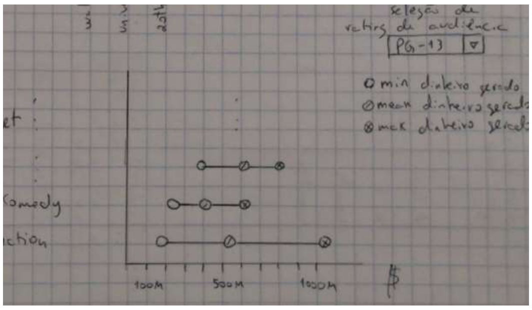

Figure 6 - Modified lollipop chart.

Finally, to finish the main charts, we have figure 6's modified lollipop chart.

While researching visualizations, we came across a modification of the lollipop chart called Cleveland dot plot [1] which we found interesting, it added another circle for each entry, to compare 2 different values, we modified it even further to include 3 different values for each entry.

The user selects a *MPAA* audience rating, and then, for that rating we will plot the minimum, average and maximum worldwide gross for each genre.

This visualization seeks to show the user how much money different genres generate worldwide for the different target audiences (*MPAA* ratings). Allowing the user to choose whichever genre maximizes his gains for his target audience of interest or vice-versa.

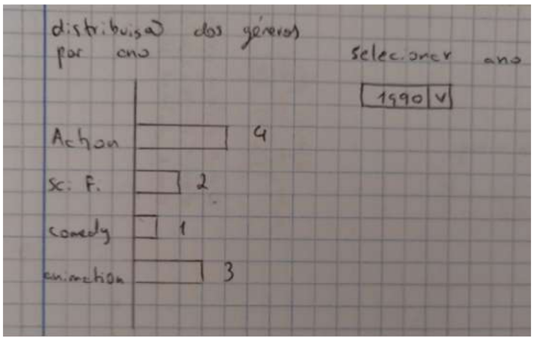

Figure 7 - Horizontal bar chart.

The data's section horizontal bar chart (figure 7) is there mainly to allow the user to explore the data.

When a year is selected in a dropdown box, the chart will update to show display the distribution of genres of the 10 highest grossing films of the selected year.

The user feedback we received proved to be very useful, and we decided to implement many of its ideas.

We will start from the scatter plot of figure 3. We asked our colleagues some questions about it and it seemed like most of them got the point of it, and thought that once coloured, it would be easy to identify genre trends. One interesting idea came up, a colleague suggested that on the legend, to the right of the chart, by hovering over the coloured circle, all other genres would disappear from the chart. This would facilitate the process of analysing the evolution of each individual genre, and so, we decided to implement it.

The heatmap of figure 4 also proved to be easy to digest, and most people could tell which genre distributor X was best/worst at. It was also mentioned that a tooltip of some sort to display the actual rating and number of movies for each combination on hovering would be useful, we added this feature as well.

Most users also answered correctly to our questions regarding the chart of figure 5, but here it was also suggested that a tooltip should be added to improve clarity of the values, such as displaying the budget/gross ratio and number of movies for each bar (distributor).

Some people had mild difficulty understanding the chart of figure 6, but after some time they could correctly answer our questions.

Figure 7 is perhaps the easiest to understand chart in the whole application so there were no issues there.

And in the end, there was another interesting suggestion, this time not regarding a chart, but a functionality, it was suggested that we added a search bar to the data exploration section, to allow the users to search for individual movies and see their most relevant attributes.

Functional prototype

Once we had received the feedback, we started implementing a basic version of our application so we could later perform some usability tests to evaluate our application's structure, visualizations and features.

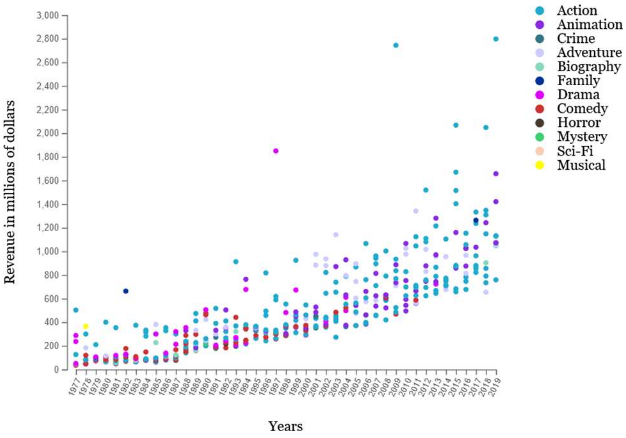

Figure 8 - Scatter plot implementation

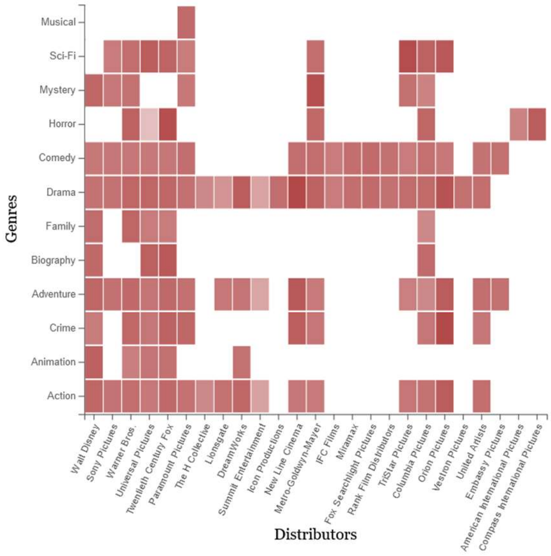

Figure 9 - Heatmap implementation.

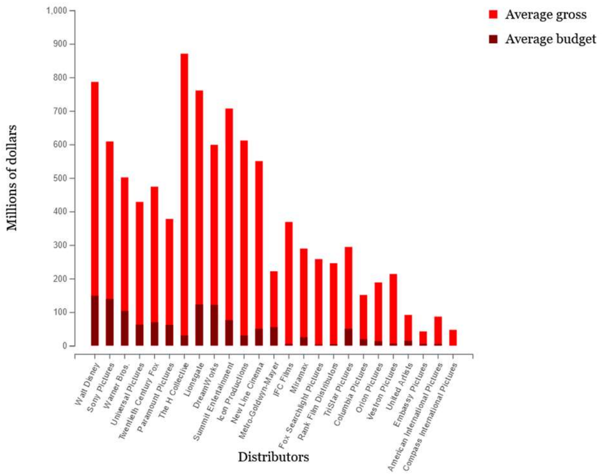

Figure 10 - Stacked bar chart implementation.

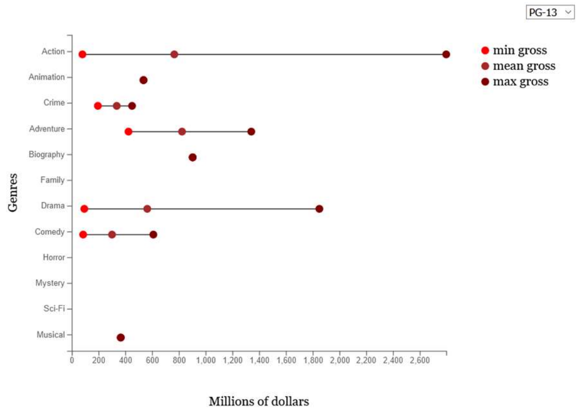

Figure 11 - Modified lollipop chart implementation

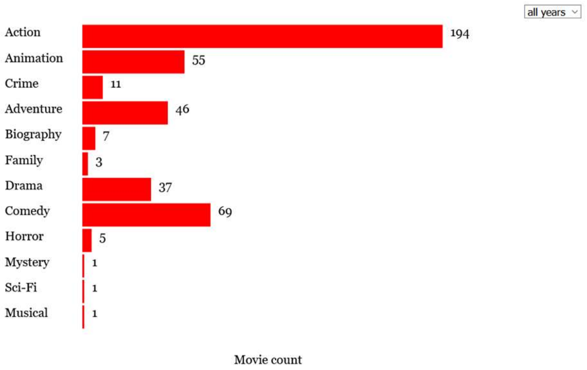

Figure 12 - Data exploration bar chart implementation.

Implementation challenges

Given that we were completely new to d3js, the implementation was very difficult in the beginning and we needed some guidance, we got it mainly from the classes and the basic tutorials at the d3-graph-gallery website [2].

We found that the most difficult parts occurred sometimes in the simplest of charts, for instance, it took a lot of time before genre chart represented in figure 7 could update its bars when selecting different years, and there were some problems with updating the axis in the lollipop chart as well.

We found that we had to take into consideration the film's subgenre in the heatmap calculation, otherwise there would be too many blank cells, thus the average value for a genre = g and distributor = d combination takes into consideration movies from distributor d that might have g not as their main genre but as their subgenre.

But by far, the most difficult chart was the scatter plot. Implementing the chart and the colour filtering was easy, but it took a lot of effort to get the zoom feature working, even referring to relevant tutorials on the subject [3], since it was only zooming on the x axis and we needed to zoom on both.

Regarding other code used, we modified some template code for the CSS, concerning d3js, we took some guidance from d3-graph-gallery's interactivity tutorial [3] for the zoom part, though as previously stated, it was heavily modified to zoom on both axes. We also used the same website's heatmap [4] and scatter plot [5] tutorial as the basis for our heatmap and scatter plot visualization.

The code that produces the main visualizations is in the investors.html file, while the data exploration part is in the data.html file, both in the pages folder.

Evaluation and changes in the prototype

As previously stated, in order to evaluate our functional prototype's usability, we got 4 colleagues to complete some tasks using our application. We asked questions and checked whether they could find the appropriate pages, visualizations and answers, and how difficult of a process it was.

We asked simple questions such as:

- How has the public's preference for genre G changed?
- Which genre is the distributor D better at?
- What is distributor D's average budget and how does it relate to their average revenue?
- Which genre is the most profitable on average for the target audience M (*MPAA*)?
- How many movies of genre G were there in the year Y's top 10?

The results were similar to our previous feedback session as expected, since we did not make any major changes and only added some features, all 4 subjects completed all the tasks, mostly with ease. Thus, we proceeded to polish the application to its final state.

V. CONCLUSION AND FUTURE WORK

In the end, we think we managed to make an interesting application, and even though, going into it, our knowledge on JavaScript, D3js and visualization techniques was limited, we came out of it much more educated, and interested, both on the coding aspect as well as visualizations in general.

REFERENCES

[1] Cleveland (Dot Plot) Rocks!
https://spencerbaucke.com/2019/09/12/cleveland-dot-plot-rocks/

[2] The D3 Graph Gallery
https://www.d3-graph-gallery.com/

[3] Brushing in d3.js
https://www.d3-graph-gallery.com/graph/interactivity_brush.html#brushingforzoom

[4] Basic heatmap in d3.js
https://www.d3-graph-gallery.com/graph/heatmap_basic.html

[5] Basic scatterplot in d3.js
https://www.d3-graph-gallery.com/graph/scatter_basic.html