Process Book

COM-480 - Data visualization - Project milestone 3

Authors: Pablo Pfister - Luis de Lima Carvalho - Stanislas Furrer

Dataset

The dataset selected for this project is a daily record of the top trending YouTube videos. It proposes numerous insights such as title, channel name, category, views, likes, dislikes, comments, publish time, tags, etc. from different regions of the planet (USA, Great Britain, Germany, Canada, France, Russia, Mexico, South Korea, Japan and India).
It is provided by Kaggle (https://www.kaggle.com/datasnaek/youtube-new/) and offers information for more than 200k top trending videos.

Path

We started by exploring the data and processing some basic statistics in python. Once we had a clear overview of the dataset we started by looking at existing projects on YouTube, and we were able to build our problematic.

We create a website that provides essential information on how to optimize the publishing of videos. You can explore the website and get an overview of meaningful data. Select your country or any particular category and be able to catch the essential metadata that top trending video are using

Let's put a context:
You live in the USA and you want to create a YouTube sport channel.
We will inform you about

- The current number of sport channels in the USA
- Example of Sport YouTube channel
- The trending days and hours to upload your video
- The trending length of a title
- The trending tags used

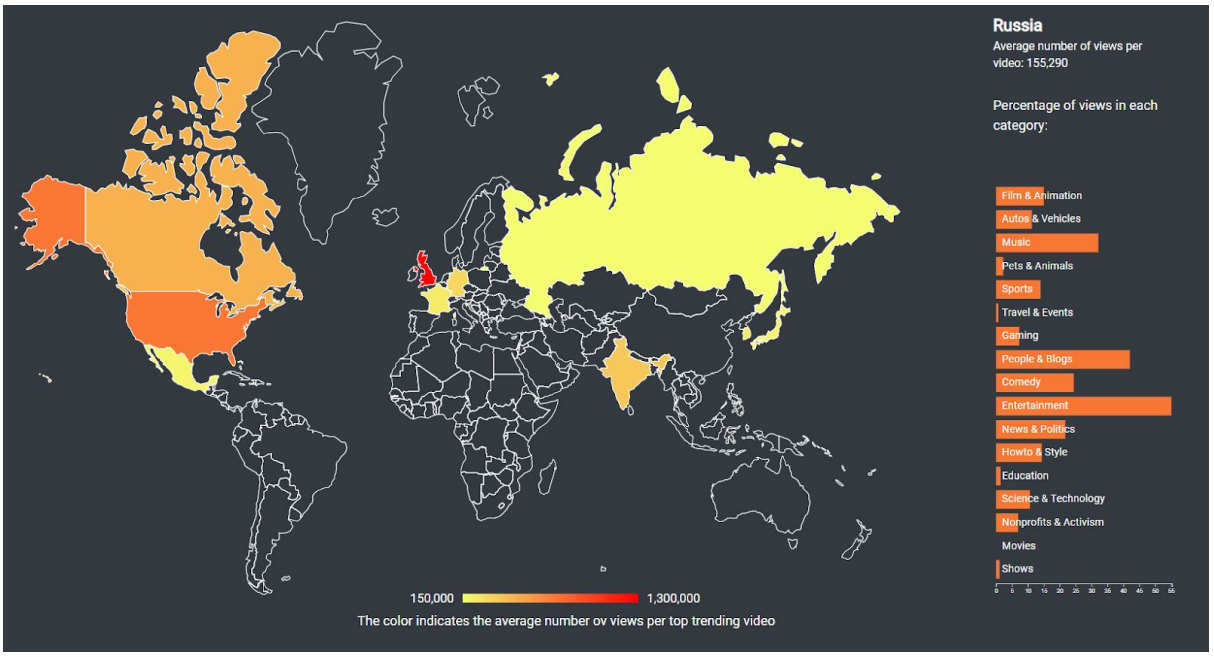
This website is fully interactive and ensures an easy way to get some statistics and make some comparisons according to what one wants.

Visualization

Our exploratory analysis on the data set shows us the most meaningful variable to optimize.

We started their visualization by making some sketches, plans and looking at some existing templates. Our website is divided in three distinct visualization :

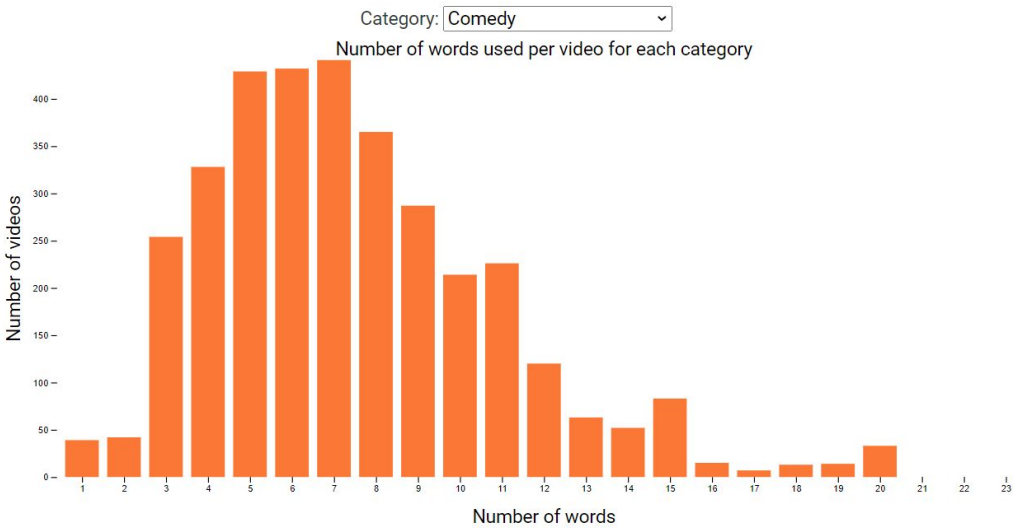
Country



Probably our favorite graph, the map shows the different countries of our database and the user is able to mouseover a country to get the percentage of views in each category for the given country. As a basic statistic, we also have the average number of views per each country that we used to represent the graph as a choropleth map. This visualization helps the user have an idea of how popular trending videos are in their country (if available) and also most importantly, which categories are trending in the country. It also gives a good overall view of that dataset as a whole. We were thinking about a map graph from the start, but we haven't managed to design one that pleased us until now.

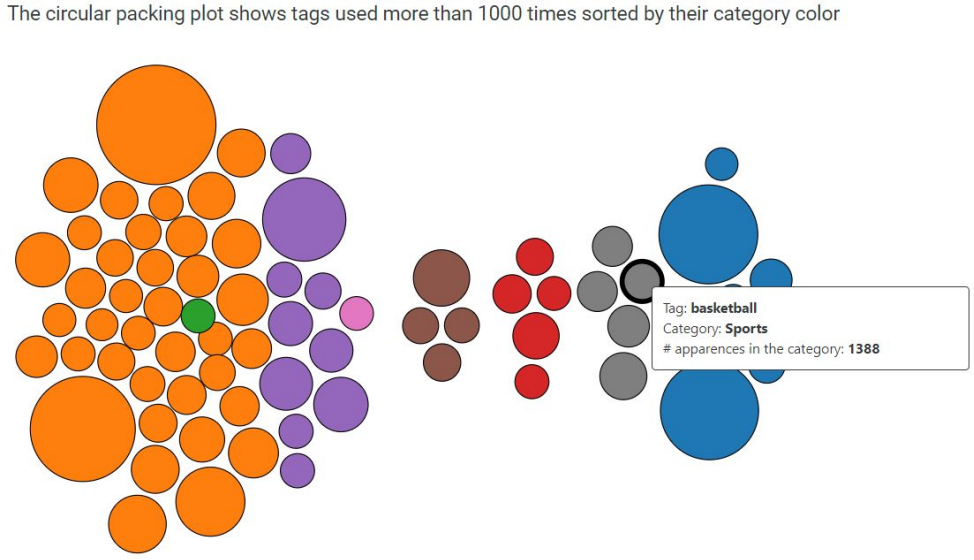
Title Analysis

The multiple category selection bar plot in the title analysis shows in its x-axis the number of words used in the title and the y-axis, the number of titles with x total words. This is done for each category and it shows that some of them have more descriptive titles than others which is interesting to take on account when choosing a title for a video. Prior to this milestone, we had some different representations such as title character length and a word cloud, but the cloud we saw in class that it has various disadvantages such as an unstable layout and the number of words proved more reliable, so we decided to keep it and expand it to every category and all of them together too.



Tags analysis I

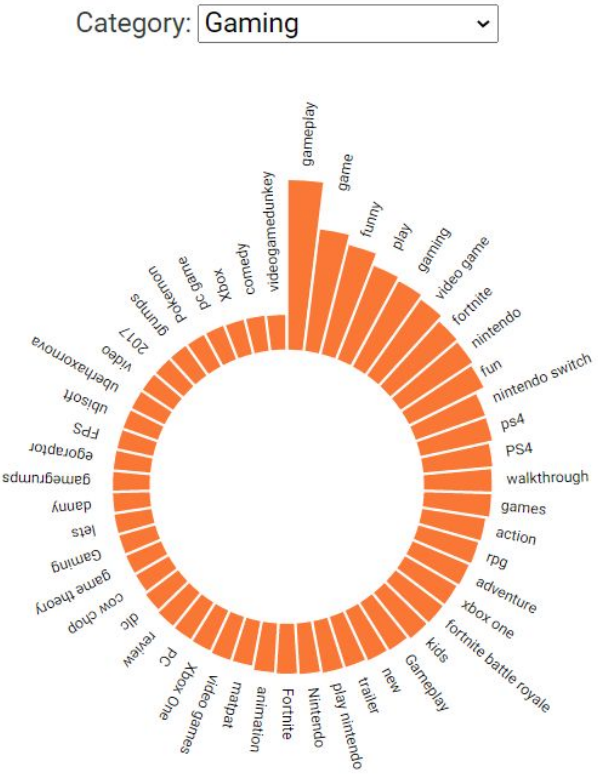
The circular packing plot shows tags used more than 1000 times sorted by their category color. Its result shows the most dominant tags over all videos and this is interesting to see the categories containing the most used tags which means that for some categories, certain tags are seen as a staple of their content. To make the visualization clear and interactive, the user can mouse over the nodes to see the tag in bold, the color category and also how many times it was used. The downside is that we cannot see the most important tags for each category and this is why at first we wanted to have a homogenous distribution of nodes for each category, but we would miss the difference of the importance of tags in each category. To compensate for this downside, we have the next plot focused on the top tags of each category



Tags analysis II

The circular bar plot shows in a stylish way the most used tags of each category. This helps understand what kind of content is trending in a certain category, for example in sports, we

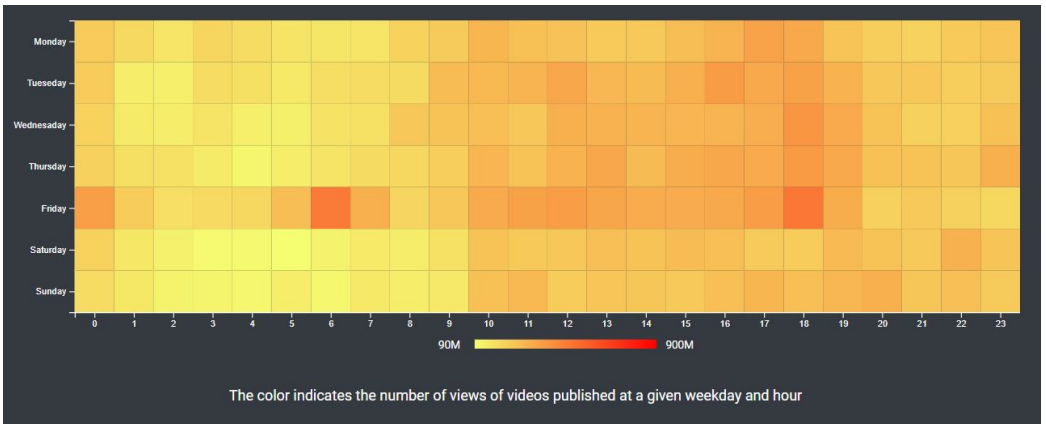
can identify the most popular sports in our database. If the user aims to have a trending video he/she might consider the type of content and the corresponding tags to do so easier statistically. At first, it was a simple bar plot, the formatting as a circular bar plot adds a bit of style specially with the data sorted as it creates a spiral effect that pops out the relevance of the tags.



Timing analysis

Last but not least, to appear among the trending videos, it is important to have videos posted at an appropriate time. We decided to use here two different heat maps : the first one shows which publishing time of a day gives the most views and the second one use hour of the day too but extended throughout a whole month, so the user will also have an idea on which seasons have more views. We had a timing analysis prior to this milestone but the visualization was not as clean and interesting as it is now, indeed, we think that the heat

map really helps pop out the impact of timing.

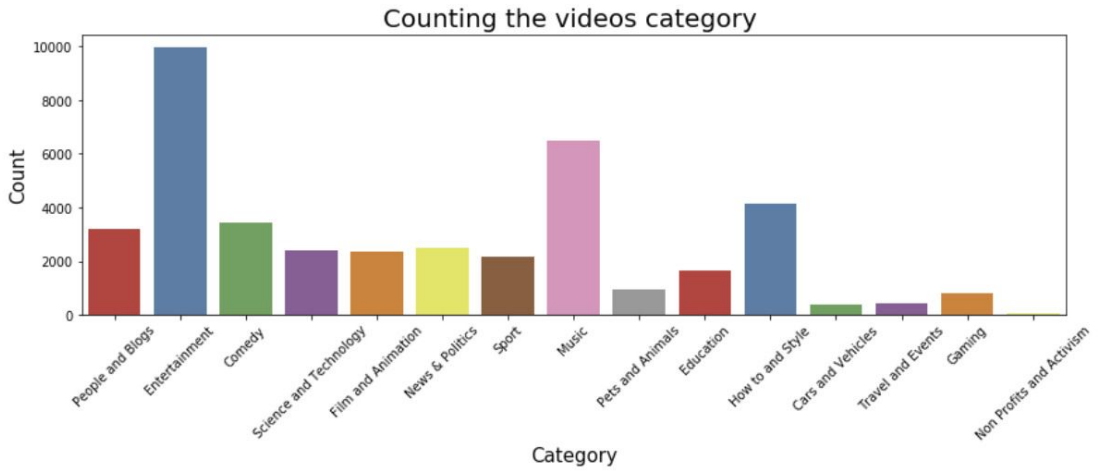


Challenges

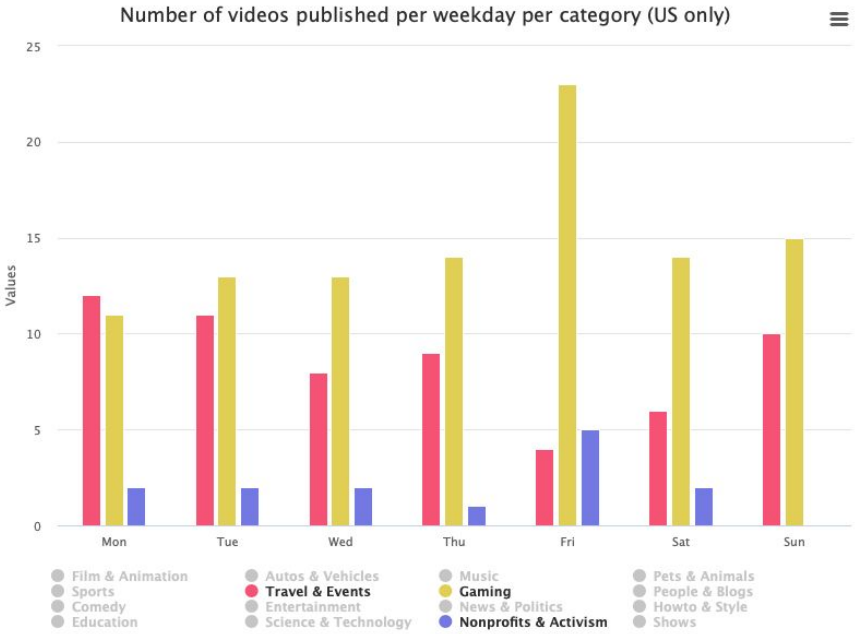
During this process we faced many challenges but the most important ones were choosing the most relevant data pieces, implementing visualization, mixing everything we know and present the whole in a meaningful and enjoyable way. First, our dataset has a lot of different stats from YouTube videos which means there are many ways to approach it and to pick one we needed to see far ahead to not end up needing to start it all over again. Since our videos are already trending videos, we did not approach the number of likes or views directly, but we used the total views as a tool of comparison in other analysis. The goal is to show what trending videos have in common so that the user can make a video just like the top ones.
The next challenge is a technical one. Being new to the enormous world of JavaScript and D3.js, it is already a challenge to catch up but also in current situations of the coronavirus COVID-19, things got a bit more complicated when working in groups and following online classes. We tried our best to fuse different concepts together and use meaningful good-looking visualizations.
Finally, after preparing the data and setting up their visualization, the whole project needs a presentation. This is not easy because a good presentation is not simply to show what we did but to make a data story in a fun, short, interesting and meaningful way all at the same time. Fortunately, our idea from the beginning of analyzing what trending videos have in common, brought us to the idea of presenting the data story as how to create the next trending video in a country and category that the user is interested in.

Comparison with the first two milestones

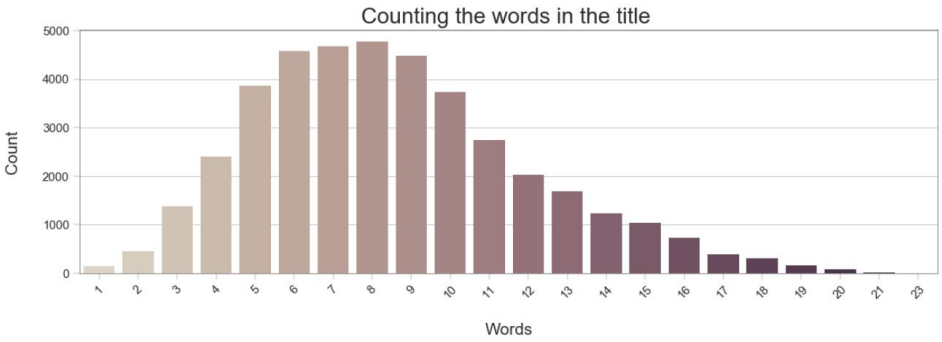
We are going to take some of our plots from Milestone 1 and 2 and explain what we decided to do with them.



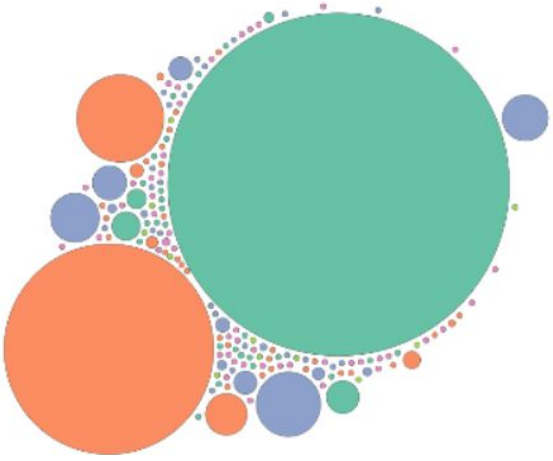
The plot above was interesting at first to have an idea of the data we are using but in the end it does not contribute in helping our user create a trending video, it simply shows that some categories will have more information and be more reliable.



This next plot is interesting because we wanted an analysis on the publishing time but ended up transforming this categorical bar plot into two heat maps that actually give us an easy meaningful representation of the best hours in a day and or month. Also, we noticed that timing is not categorically dependent, which means that we did not need to select them.



The word count plot was actually interesting for us, but we wanted it to be different for each category because some of them have more descriptive titles than others which means that the title word count varies from each category.
As briefly mentioned in the challenge section, we decided that the trending videos are already successful, so we did not continue our direct analysis on number of likes and views since this can vary for special reasons like famous YouTubers making simple videos but that are already active from years now like PewDiePie or movie trailers which, depending on the movie, are already massively expected to be seen.
We also had a sketch for a network graph that would link tags to their categories but also the other categories using it. While this may seem very exciting, not only it would take an enormous space and effort to make it as clean as possible, we did not think the results, based on prior data frame analysis on Python, would justify it. We are not even including the sketch from milestone 2 since it is unreadable.



The circular packing graph above was meant to show the top tags of a category but as mentioned in Visualization, we decided it would be clearer and more interactive to be able to have the circular bar plot for each category and a single circular packing to show the dominance of tag and also the categories with most used tags.
Lastly, we had a word cloud but as seen in the text viz class, they have a very bad and inaccurate visual encoding, so we decided to drop it.

Peer assessment

Firstly, we would like to state that everyone has given its best and the workload was evenly distributed, some parts took more time and effort than others but overall we do think we all worked together as a team. At first, we broke down the analysis into title (Stan), timing (Pablo) and views/likes (Luis). After receiving feedback, the split became title (Stan), timing (Pablo), tags (Luis), countries (Pablo), page setup (Pablo), process book (each one their analysis part, Stan and Luis with the rest) and finally the video (all together).

How to improve ?

There is always room for improvement, the question is how ? First, our dataset had some missing values, some unbalanced distributions between categories and not many countries were included, which means our site has a few limitations that displease us. It was difficult with the distance to prepare a nice video presentation and with more video editing skills it could have been more interesting to watch. Finally, we think we might have been able to make some more unique analysis using more complex visualizations to take full advantage of the D3.js library.

Conclusion

This project helped us discover the various visual possibilities available through JavaScript and D3.js, we also learned how important and impactful it can be to have a visualization that catches people’s attention, in the end, we all think that we understand better that a good data analysis needs a powerful visualization to reach the others and transfer the information.