Visualization of Youtube Trend

Francesco Angiulli 897415, Alessandro Castagna 898057, Matteo Pasotti 901810

Data Visualization, Gennaio 2023

In this report we will analyse an infographic based on the theme of YouTube trends in the last 3 years. From data analysis to infographic analysis.
(https://public.tableau.com/app/profile/alessandro.castagna8105/viz/DataViz_16734642491290/Story)

INDEX

1 Introduction

At a time when the internet was not yet at the centre of the world and the first social media platforms were just being born, the idea that ordinary people could see and share videos was an innovation. No one would have imagined using their computers as televisions, and even better, using them to share their ideas, passions and knowledge with others. To watch a documentary or a cartoon, you had to wait for it to air on television, and to hear a song, you had to buy the CD.

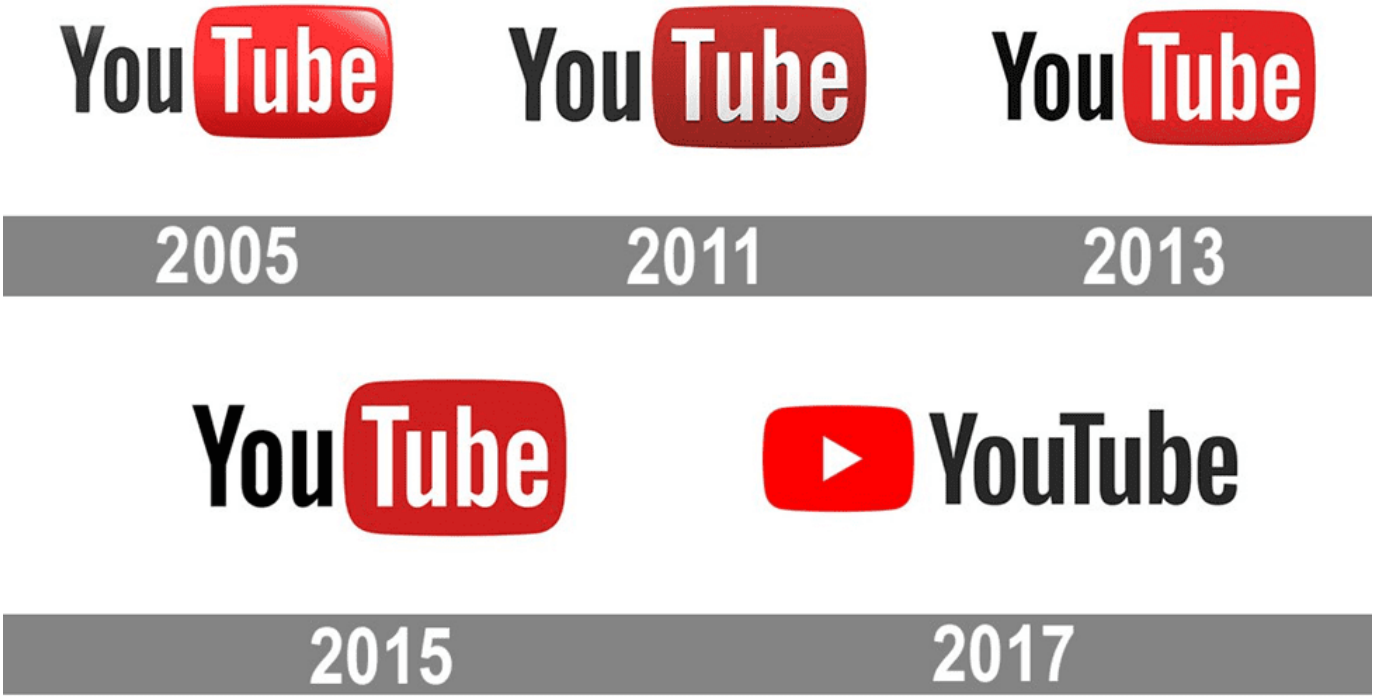
YouTube was founded with the aim of allowing ordinary people to upload their videos and allowing everyone to watch them whenever they want. Today, this platform, which now has more than 2 billion active users every month, hosts a wide variety of videos. YouTube is the second most popular social platform, and we aim to carry out an analysis of its trends.
The analysis will begin with a view of the countries that use Youtube the most and how much of the population actually uses it within the country. Next, we will look at which categories are the most watched for each country over the 3-year period.
The two most important categories will be looked at specifically : we will look at their trends from 2020 to 2022 on a quarterly level and then which videos in these categories are the most viewed, considering the selected month, year, and country.
Going further into specifics, we will also look at how many days it takes for a video to enter Trend, compared to its publication date; consequently we will also take a look at which are the days where the most videos are viewed and published, always considering month and year.
The analysis continues with a look at the Likes x Views ratio of videos above a certain viewing threshold, they also consider the category they belong to. In conclusion, for each country, we will be able to see the Likes x Dislikes ratio as the years go on.
The result will be a careful analysis of Youtube usage for each country as years go by. This can lead to knowing a country's preferences and acting accordingly to bring in more relevant content.

1.1 Domain Analysis:YouTube

YouTube is one of the largest and most popular video distribution platforms on the internet. It has more than 4 billion hours of video viewed every month, and an estimated 500 hours of video content are uploaded to YouTube every minute. Since its inception in 2005, YouTube has evolved from a platform for amateur videos to one that distributes original content. It has also enabled the creation of an entirely new profession – YouTube content creator – which can be a very profitable career for some YouTubers around the world.

History: YouTube was founded on Valentine's Day in 2005. It was the brainchild of Chad Hurley, Steve Chen, and Jawed Karim, who were all former employees of PayPal. According to its founders, the idea was born at a dinner party in San Francisco, about a year earlier, in 2004. The trio was frustrated by how hard it was, at the time, to find and share video clips online.

The Growth Of YouTube: One of the main factors that helped boost YouTube's popularity in its early days was the proliferation of viral videos, including "Charlie Bit My Finger," "Evolution of Dance," and Justin Bieber's "Baby."



Source:https://1000logos.net/wp-content/uploads/2019/08/YouTube-logo-history.jpg

Figure 1.1:YouTube logo history

YouTube's success has been unprecedented, with the platform becoming a multi-billion dollar business and benefiting numerous media markets and YouTube personalities. By 2018, YouTube was generating more revenue than any major TV network. The platform's popularity has also been boosted by lockdowns due to the COVID-19 pandemic, leading to an increase in video publishing and viewing in the most affected countries in 2020 and 2021. However, this trend seems to be falling off in 2022.

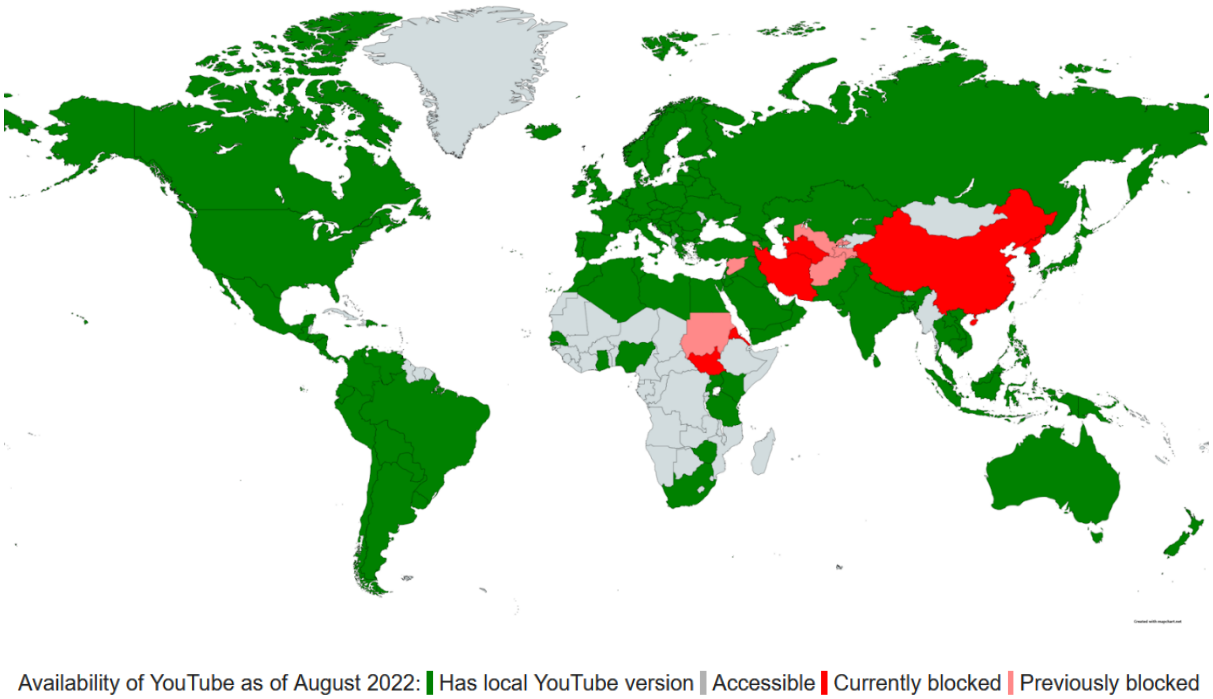
1.2 How Does YouTube Make Money?

Before it was purchased by Google, YouTube had reportedly earned a profit of around $15 million. The primary way that YouTube makes money is through

advertising. The platform enters into contracts with companies that want to advertise their products globally through the site. In exchange for providing advertising services, the company pays YouTube for the service. Another source of YouTube's income is through monthly subscription services, where individuals and companies pay to become official partners of the platform. Additionally, YouTube earns revenue from the content creators' video uploads, taking a 45% cut of the total earnings from the user's channels.

1.3 Territorial Expansion

According to the company's press page, YouTube is now available in more than 75 countries and in 61 languages, with hundreds of hours of video content being uploaded every minute. Censorship of the platform has occurred and continues to occur to varying degrees in most countries around the world.



Source:https://upload.wikimedia.org/wikipedia/commons/thumb/c/c7/YouTube_Availability.png/1920px-YouTube_Availability.png

Figure 1.2: Censorship of YouTube

There are several reasons why YouTube may be blocked, including:

- Preventing criticism of rulers, governments, government officials, religion, or religious leaders;
- Preventing videos promoting racism;
- Violating national laws, such as copyright and intellectual property protection laws, hate speech laws, or laws based on ethics or morality;
- National security legislation;
- Preventing access to videos deemed inappropriate for youth;
- Businesses, schools, government agencies, and other private institutions often block social media sites, including YouTube, due to bandwidth limitations and the potential for distraction.

In some countries YouTube is completely blocked, either through a long-term standing ban or for more limited periods of time such as during periods of unrest, the run-up to an election, or in response to upcoming political anniversaries. In other countries, access to the website as a whole remains open, but access to specific videos is blocked due to many reasons including orders from country jurisdiction. In both cases, a VPN is usually deployed to bypass geographical restrictions. In cases where the entire site is banned due to one particular video, YouTube will often agree to remove or limit access to that video in order to restore service.

2 Description of the database

The database used for this report collects data on YouTube trends from 2020 to the present. In this section, we will describe this database and discuss any changes that have been made.

2.1 Database YouTube

The database that we have chosen to base our infographic on contains information about the popularity of videos on YouTube. This database, provided by Kaggle, is constantly updated starting from 2020.

The dataset in question includes the following information:

- **Video_ID:** an identification ID for each video
- **Title:** The full title of the video
- **Date of pubblication:** The publication date of the video
- **Channel ID:** The ID of the channel that published the video
- **Channel Title:** The title of the channel that published the video
- **Category ID:** The ID of the category of that video
- **Date of trending:** The date that the video became trending
- **Number of days before the trend:** A value representing the ratio of the publication date to the trend date number of views: The total number of views for the video
- **Likes:** Total likes for the video
- **Dislike:** Total dislikes for the video
- **Category Name:** The name of the category of that video
- **Country1:** The name of the country where data for that video was collected
- **Year:** The year in which data for that video is available
- ChannelViews Video: The ratio of total views for that video to the number of videos uploaded

In Figure 2, it is possible to see a small extract of the database described above.In detail, information is collected about 1,857,817 videos that are associated with 11 different countries and 3 time periods.

Abc YT_official.csv Channel Title	YT_official.csv Country	YT_official.csv Date of trending	YT_official.csv Date of pubblication	Abc YT_official.csv Title	Abc YT_official.csv Video Id
HYBE LABELS	France	16/10/2021 00:00:00	12/10/2021 08:56:11	ENHYPEN (엔하이픈) 'Tamed...	6IRsW_TkyVA
Timal - Topic	France	16/10/2021 00:00:00	14/10/2021 22:03:32	Sales idées	T4rJOSZULfA
Timal - Topic	France	16/10/2021 00:00:00	14/10/2021 22:09:26	Jeunes et ambitieux	sRhSi4m61EU
StromaeVEVO	France	16/10/2021 00:00:00	15/10/2021 16:00:11	Stromae - Santé (Official Mu...	P3QS83ubhHE
Blackswan Official	France	16/10/2021 00:00:00	14/10/2021 09:00:11	[BLACKSWAN] 'Close to Me' ...	NnW22WrN7g0
PLK TV	France	16/10/2021 00:00:00	15/10/2021 15:39:06	PLK - A la base (clip officiel)	8ZIYDQRkPO4
Graya - Topic	France	16/10/2021 00:00:00	14/10/2021 10:08:43	Truc de malade (feat. Jul)	d4rpPU1iHTg
Jessi	France	16/10/2021 00:00:00	12/10/2021 15:30:00	Jessi (제시) - Cold Blooded (...	1JHOI9CSmXk
TriBeVEVO	France	16/10/2021 00:00:00	12/10/2021 09:00:10	TRI.BE - WOULD YOU RUN	VC81SgWh86U
Driks Officiel	France	16/10/2021 00:00:00	12/10/2021 17:05:35	Driks - Plan feat. Abou Debei...	yvVCG11nMPo

Figure 2.1: Database preview

2.2 Database Country

We have integrated a database containing population information for 11 states for the years 2020, 2021, and 2022. The database consists of four columns:

- **Country:** Country under consideration
- **Pop 2020:** Population in 2020
- **Pop 2021:** Population in 2021
- **Pop 2022:** Population in 2022

These information have been necessary for the data processing, as many calculated fields are the ratio between video information and population.

2.3 Data integration

The integration of the two different databases for the creation of the infographic was carried out using Tableau. In particular, it was decided to merge the two databases by comparing the states and keeping all occurrences of both databases. Various calculated fields were added to the main database, based on the population data for each year and in each country.

3. Infographic analysis

The goal of this infographic is to introduce less experienced users to the world of YouTube through explanatory and illustrative graphs of the main aspects of the videos uploaded to the platform every day.
The goal of this inforgraphic is to give relevance, from the analysis of past years, to certain aspects that would lead a country to wanting to bring more YouTube consumption, based on total usage, top categories, highest usability days, and number of likes/dislikes.
The explanation provided is simple but detailed, in order to facilitate learning.

3.1 YouTube trend

To examine this topic, the following dashboard analyzes the behaviors of the countries in question during this historical period.

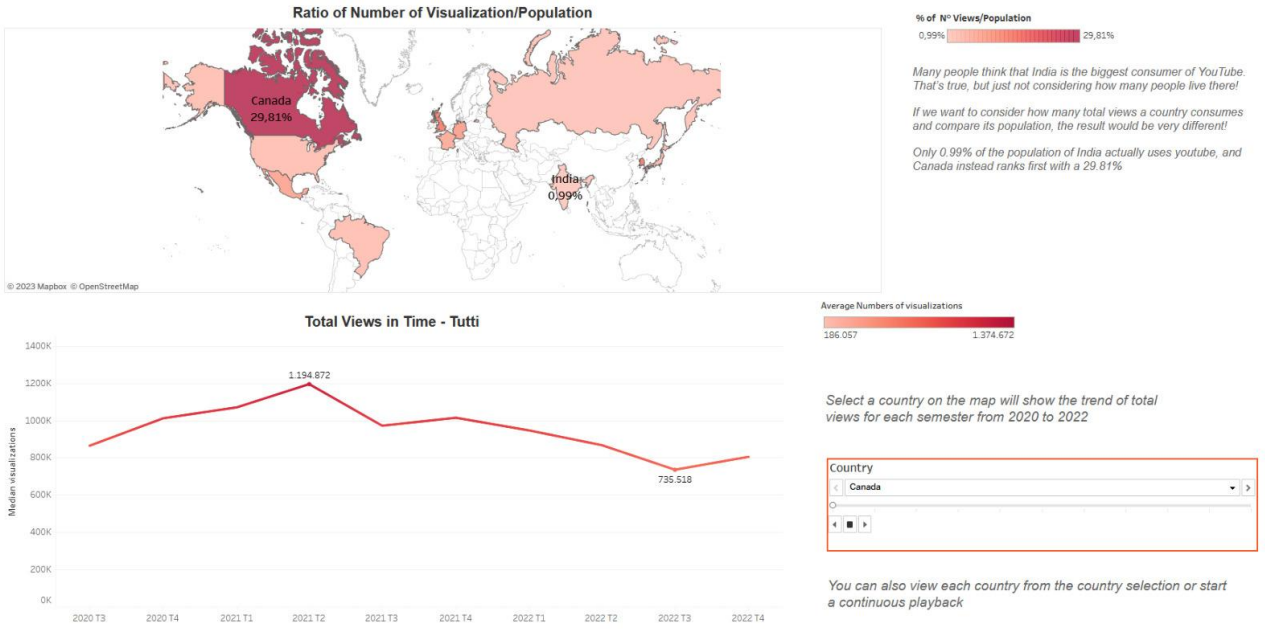


Figure 3.1: Dashboard 1

As the image above shows, the first data visualization is a geographical map that represents and compares all the countries considered in this study. The analysis is based on the total number of views for each country, normalized based on the total population. Without this normalization, heavily populated countries such as India would automatically be among the top positions, while this method allows for a fair analysis.

The chart reveals two percentage values, which represent the maximum and minimum of the analyzed attribute:

- Canada has the highest value of views in relation to the size of its population, with 29.81% of the total values for each country.
- India, on the other hand, has the lowest value with 0.99% of the total.

To facilitate visual understanding by users, a "value" has been added, which is a monochromatic scale (in this case red) with saturation variation proportional to the value of the analyzed attribute.

The aim of the second visualization is to analyze each country more closely by displaying the behavior of their total number of views per quarter over the three years considered (2020, 2021, and 2022), highlighting the maximum and minimum values for each country.

3.2 Most viewed category

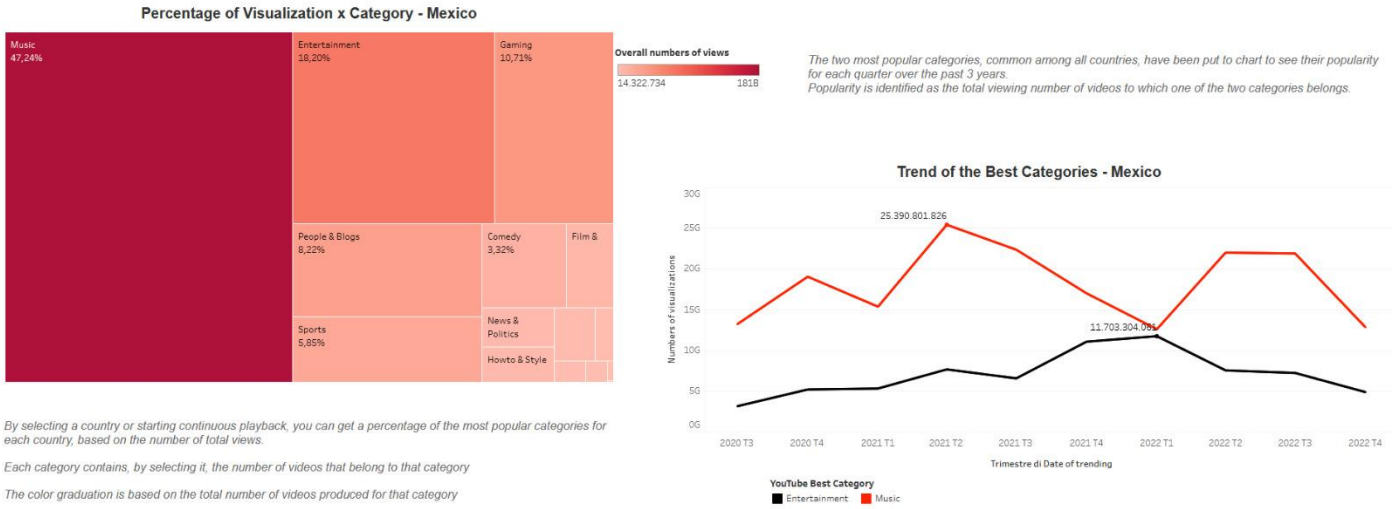


Figure 3.2: Dashboard 2

As can be seen from the above image, the goal of this dashboard is to dig into the theme of "views". To this end, a Heat Map has been created, which is a graphical representation of how views are divided within the country of interest. In fact, it is possible to select the country on which you want to perform a more in-depth analysis and see how the number of views is divided among the 15 categories used by YouTube to classify videos. This division is highlighted by two visual indications: value and size. The more intense the value(in this case, red), the higher the percentage of the total, and the larger the size, the higher the percentage. By checking the data for all states, it can be noted that the order of categories is almost always the same: the first two categories with the highest number of views are always Music and Entertainment. To better analyze this phenomenon, as shown in the second graph, particular attention has been paid to the trend of categories over time (2020-2022). To facilitate their visual comparison, they are represented with two distinct colors: Music with the color red and Entertainment with the color black. In addition, the highest values for each time line are visible in each graph.

3.3 Most viewed videos

In the following dashboard, there is further focus on the two most popular categories for each state: Music and Entertainment. The dashboard presents two graphs, one for the Music category and the other for the Entertainment category, which show all the videos that make them up. Each video is represented by a lollipop chart, which is a variation of a bar chart where the bars are replaced with circles ("lollipops") on a stick. The length of the stick represents the number of views for that specific video and the circle marks the end of the stick.
In addition, another aspect that indicates the number of views is the value (in this case, red); the more intense the color, the more that video is characterized by a high number of views, and the clearer the color, the fewer views that video will have.
In both graphs, it is possible to apply filters: it is possible to select the month or months of interest, the year or years of interest, and also on which countries the comparison must be made. By applying some filter, the graph will show that particular situation by changing the dimension and the number of the lollipops and the dimension of the two axes.

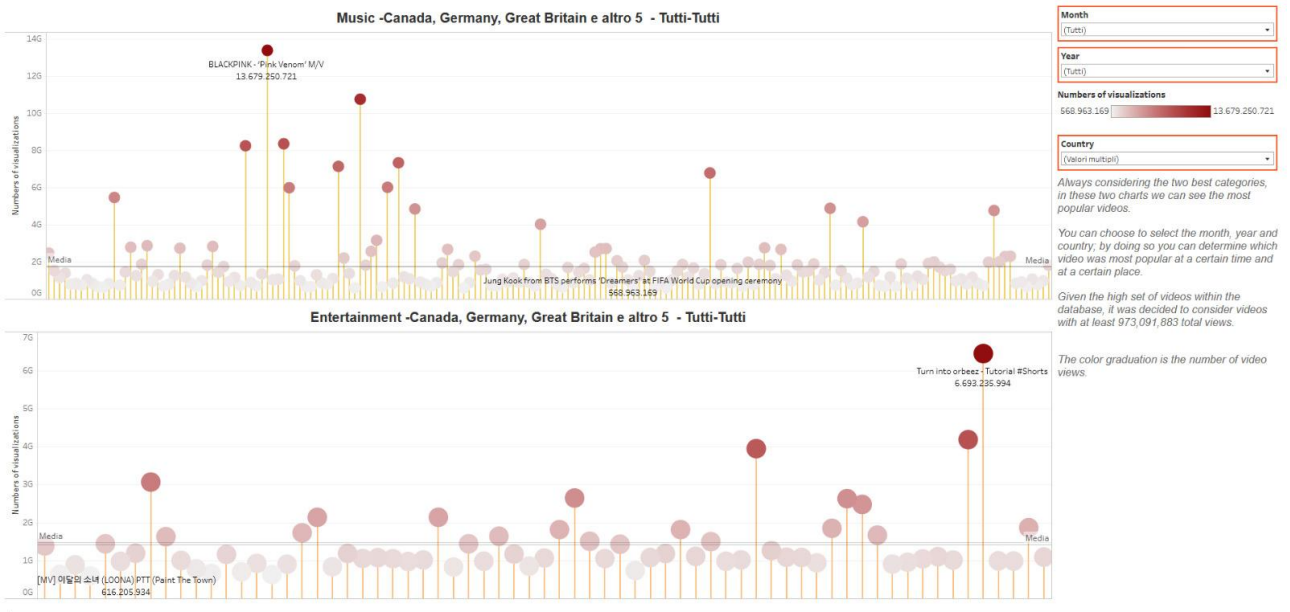


Figure 3.3: Dashboard 3

3.4 Trend analysis

The following dashboard aims to analyze YouTube trends from 2020 to 2022.

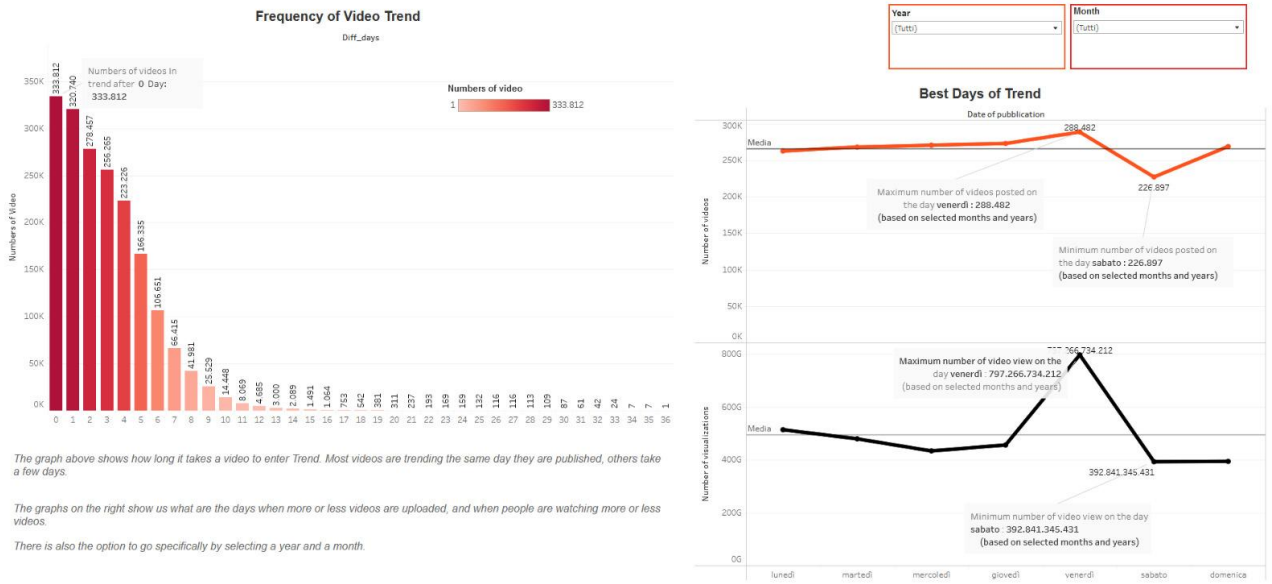


Figure 3.4: Dashboard 4

The visualization on the left of this dashboard represents the time to trend, or the time it takes for a video to become viral. This histogram has the number of days on the x-axis, representing the time (in days) it takes for a video to become trending, and the frequency of the number of videos that become trending on the y-axis. So, if we consider the first horizontal bar, it indicates that there are 333,812 videos that became viral on the same day they were published. The trend of the graph shows how the number of videos that become viral tends to decrease more and more as the number of days since publication increases, until it reaches zero after 36 days. To highlight this decrease even more, the value, or the same colour with different saturation, has been applied to each horizontal bar: the lower the number of days, the more intense the colour will be, and the higher the number of days, the lighter the colour will be.

The second chart on this dashboard deals with the topic of Trends, focusing on the number of videos published and the number of views based on a particular day of the week. This way, it is possible to discover which are the most frequent days on

YouTube. In addition, it is possible to focus on the year and month of interest by selecting them in the filter below the chart.

3.5 Video statistics

The last dashboard of this infographic provides information about some YouTube statistics for the period 2020-2022.

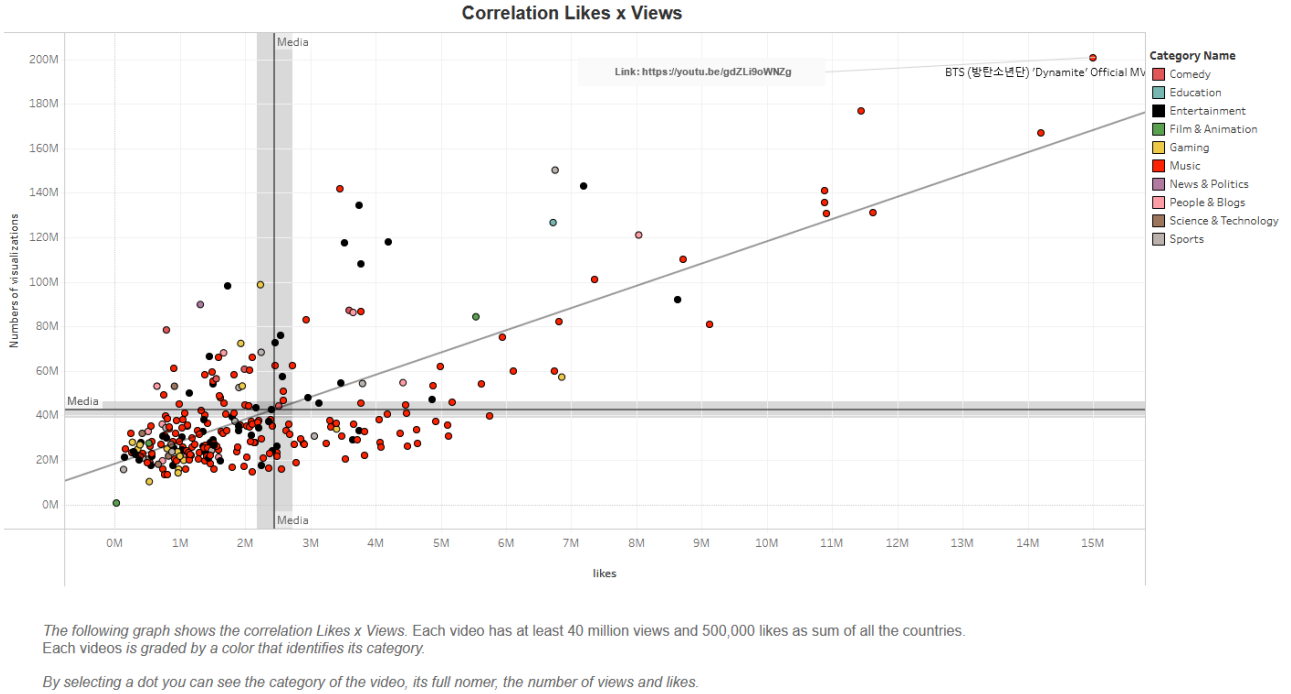


Figure 3.5: Dashboard 5

The graph, a scatter plot, displays the relationship between likes and views for all videos with at least 40 million views and at least 500,000 likes. This filter on the videos has been applied to show how the most popular and well-known videos have behaved in the past 3 years. Each point represents a specific video and its color represents its belonging to a specific category as shown in the color legend on the right. Within the graph, there is a linear regression line to highlight a correlation between the two variables: as can be seen from the line, there is a positive correlation between the two variables and this means that as one variable increases, so does the other. To facilitate understanding, two lines representing the

average number of likes and the average number of views with their respective confidence intervals have been introduced; these divide the representation into four quadrants:

- In the bottom right: high number of likes and low number of views
- In the bottom left: low number of likes and low number of views
- In the upper left: low number of likes and high number of views
- In the upper right: high number of likes and high number of views

This graph also highlights the two most extreme videos, the one with the most views and likes and the one with the least views and likes, indicating them by their names.

3.6 Like vs Dislike

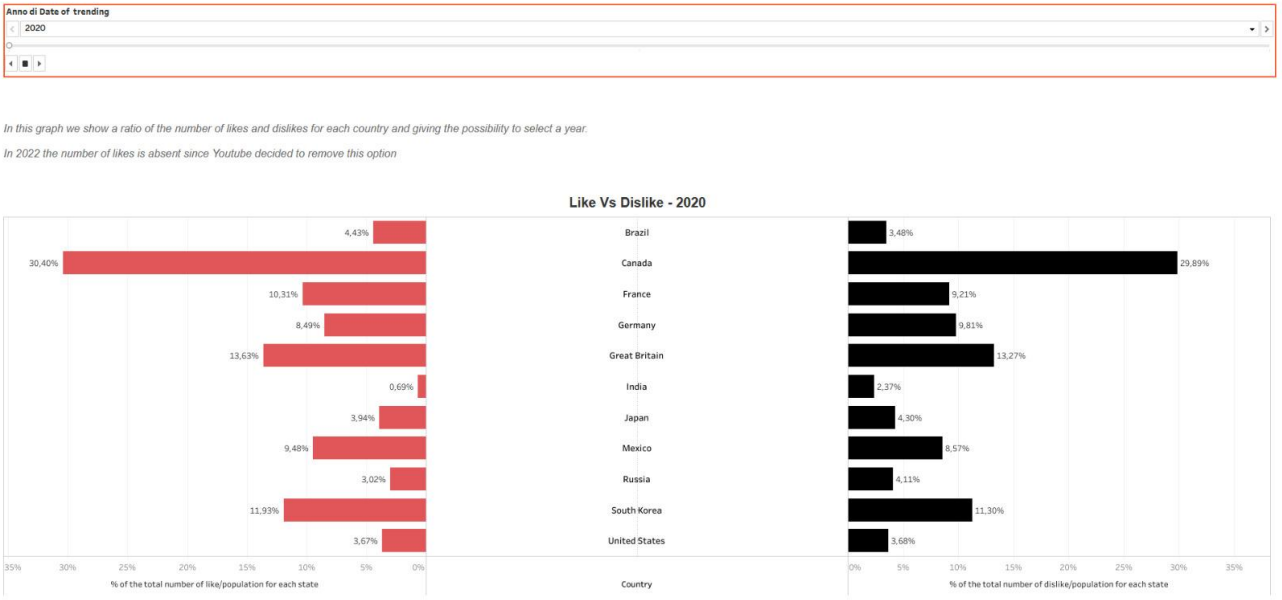


Figure 3.6: Dashboard 6

Finally, the last but not least dashboard is the representation of likes and dislikes for each country. In this butterfly chart, each bar represents the total number of likes or dislikes, compared to its own population value. This ratio follows the same reasoning seen in the first dashboard (Chapter 3.1), it serves as a normalizer between countries, not allowing populous countries like India to automatically rise to

the top positions, but allowing for a fair comparison. In this graph, it is possible to see the variations of these values over the past 3 years, which allows us to see how in 2022 the number of dislikes is 0 for every country. This particular data is due to the new YouTube policy introduced in January 2022, which removes the ability to view the number of dislikes under each individual video.

4 Data Visualization Evaluation

4.1 Heuristic Evaluation

The Heuristic Evaluation involved 4 users. We asked them to interact with our data visualizations without any restrictions and to comment aloud during every single stage in order to improve our visualizations and identify any potentially misleading information. We analysed their behaviours and comments for several minutes and detected a few issues, which are listed in the table below:

	HEURISTIC EVALUATION
USER 1	- In the map about the first Dashboard , the colour difference between countries with similar values isn't relevant. - The Tree Maps into Second Dashboard Choose better colour to highlight the difference between categories.
USER 2	- In the first Dashboard Explain better the colour difference between lines in the first graph. - Into Second Dashboard Depict values for all the top two categories
USER 3	-In the first Dashboard more Interactive graph , if it selected a Country on the Map, the line chart below should depict, automatically, the selected country
USER 4	- The user asked to define a coloring more appropriate to the context

Table 4.1: Description of the issues about data viz

4.1.1 Editing Data viz

Based on the findings from the Heuristic Evaluation, we made a few modifications to simplify the data visualization and make it easier to understand.

One modification was made to Dashboard 1. After the Heuristic Evaluation, we increased the number of bins for the colour value (shades of blue) on the map graph to better highlight the ratio of views to population for each country. Another issue was revealed during the Heuristic Evaluation about country selection on the map. We added a condition on both the map graph and the line chart to make a correlation between them. Now when a country is selected, the line chart automatically depicts the trend of views for that country over each trimester, as shown below in the figure:

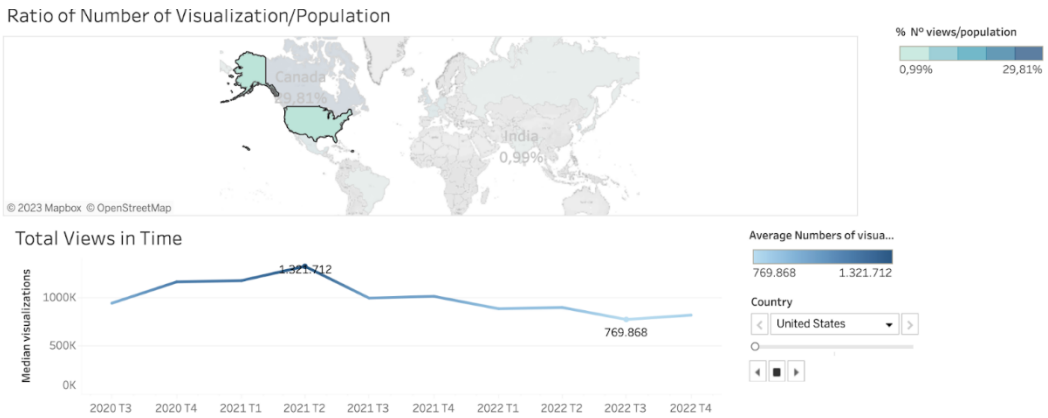


Figure 4.2: Selected US Country and Line Chart shown a US trend

An issue was revealed during the Heuristic Evaluation for Dashboard 2 regarding the use of colours to distinguish the categories on YouTube. Initially, categories were distinguished by different colours, but we decided to change to a colour gradient (from green to yellow) where green represents the largest category and yellow represents the smallest category. This modification made it easier to immediately understand the most popular YouTube categories. Overall, the users found our infographic easy to understand and meaningful, and they stated that the other dashboards had well-explained visualizations.

Based on User 4’s observation, the colour base line has been changed with a more appropriate YouTube colour palette.This variation can be seen in the previous

paragraph (Chapter 3), where it shows that all the dashboard have a black and red colour palette.

4.2 User Test

After addressing the issues revealed in the Heuristic Evaluation, we conducted a User Test with 10 subjects. We gave the users 3 tasks and measured the time it tooks them to complete each task to evaluate efficiency, as well as the number of tasks completed to evaluate effectiveness. The tasks were of medium complexity, designed for users with a good knowledge of data visualization. The tasks are listed in the table below:

	USER TEST
TASK 1	Find the most viewed video in Japan in 2021 for the Music category and for the Entertainment category
TASK 2	What is the day when fewer videos were released in December 2020?
TASK 3	The lowest mean value of views in France

Table 4.3: User test's tasks

The following rank level is asked for each user tested:

B = Beginner

I = intermediate

E = Expert

	RANK
USER 1	B
USER 2	B
USER 3	I
USER 4	E
USER 5	E
USER 6	B
USER 7	I
USER 8	E
USER 9	E
USER 10	B

Table 4.4: User test's ranks

4.2.1 User test's results

To evaluate the efficiency, we timed the execution times of each task performed by the user. The tables below show the times expressed in seconds as the unit of measurement: table 4.6 shows the mean and the medium value for each task and for all the values, while table 4.7 shows mean and medium value for each task and rank.

	TASK 1 EXECUTION TIME	TASK 2 EXECUTION TIME	TASK 3 EXECUTION TIME
USER 1	74"	121"	59"

USER 2	80"	117"	64"
USER 3	65"	89"	39"
USER 4	51"	66"	20"
USER 5	56"	60"	29"
USER 6	110"	134"	67"
USER 7	61’’	75’’	35’’
USER 8	45’’	59’’	23’’
USER 9	55’	63’’	25’’
USER 10	79’’	108’’	49’’

Table 4.5: User test’s execution time

	TASK 1	TASK 2	TASK 3	OVERALL
MEAN	67,6’’	89,2’’	41’’	65,9’’
MEDIAN	63’’	82’’	37’’	62’’

Table 4.6: User test’s mean execution time

	TASK 1	TASK 2	TASK 3
Rank B	85,75’’	120’’	59,75’’
Rank I	63’’	82’’	37’’
Rank E	51,75”	62”	24,25”

Table 4.7: User test’s mean execution time x rank type

From the table above (Table 4.7) as can be seen, from the task execution time, the ease which the expert users solved assigned tasks compared to the other users. Expert users achieved the goal of the task with a lower time of execution than the other users, followed by the intermediate one and last the beginner one.

The following violin plot depicts the execution time for each task.

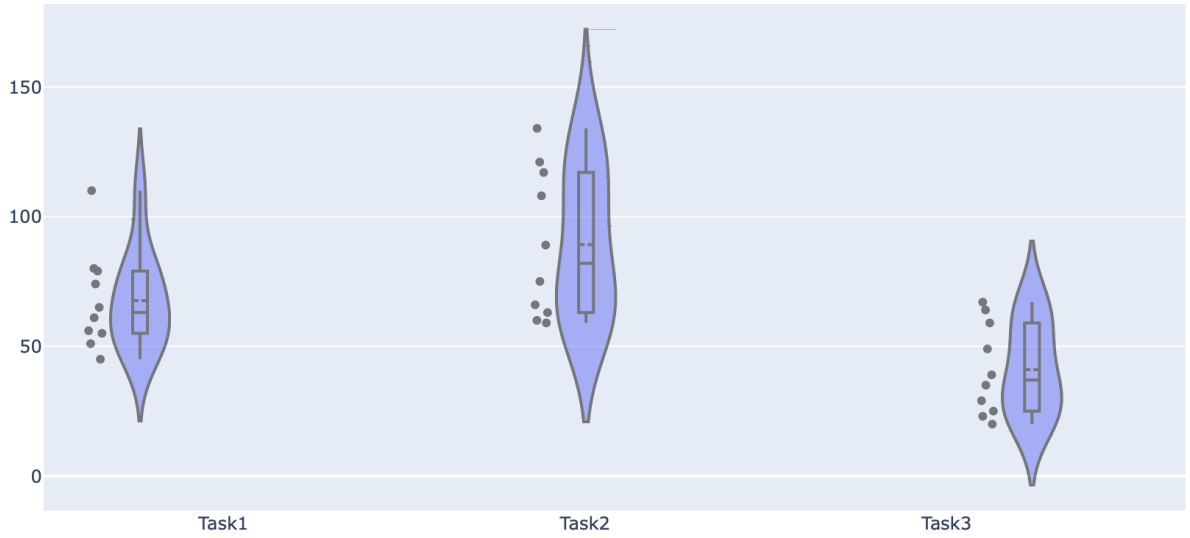


Figure 4.8: User test's violin plot

The graph is composed of a violin plot, box plot, and dot plot. These three graphs depict how values are distributed: the violin plot allows you to quickly visualize data distributions; the dot plot highlights the value distribution, particularly the execution time for each subject for each task; the box plot allows you to see where the maximum and minimum values are, indicated by the length of the whiskers. Values above the whiskers are called outliers, while the middle line is the median value. To evaluate effectiveness, we calculated the error rate for each task. The table below shows the results:

	TASK 1	TASK 2	TASK 3
USER 1 B	74	121	59
USER 2. B	80	117	64
USER 3. I	65	89	39
USER 4. E	51	66	20
USER 5. E	56	60	29
USER 6. B	110	134	67
USER 7 I	61	75	35
USER 8 E	45	59	23
USER 9 E	55	63	25
USER 10 B	79	108	49

Table 4.9: User test's results

If the label is coloured green, it means that the users completed the task without any suggestions, indicating that they completed the task in total autonomy.
If the label is coloured red, it means that the users completed the task with some tips or suggestions.
It can be seen, that the majority of users who had difficulty solving the tasks are beginner users.

To demonstrate the effectiveness of our infographic, the error rate for each task is shown (a confidence interval of 5% was chosen).

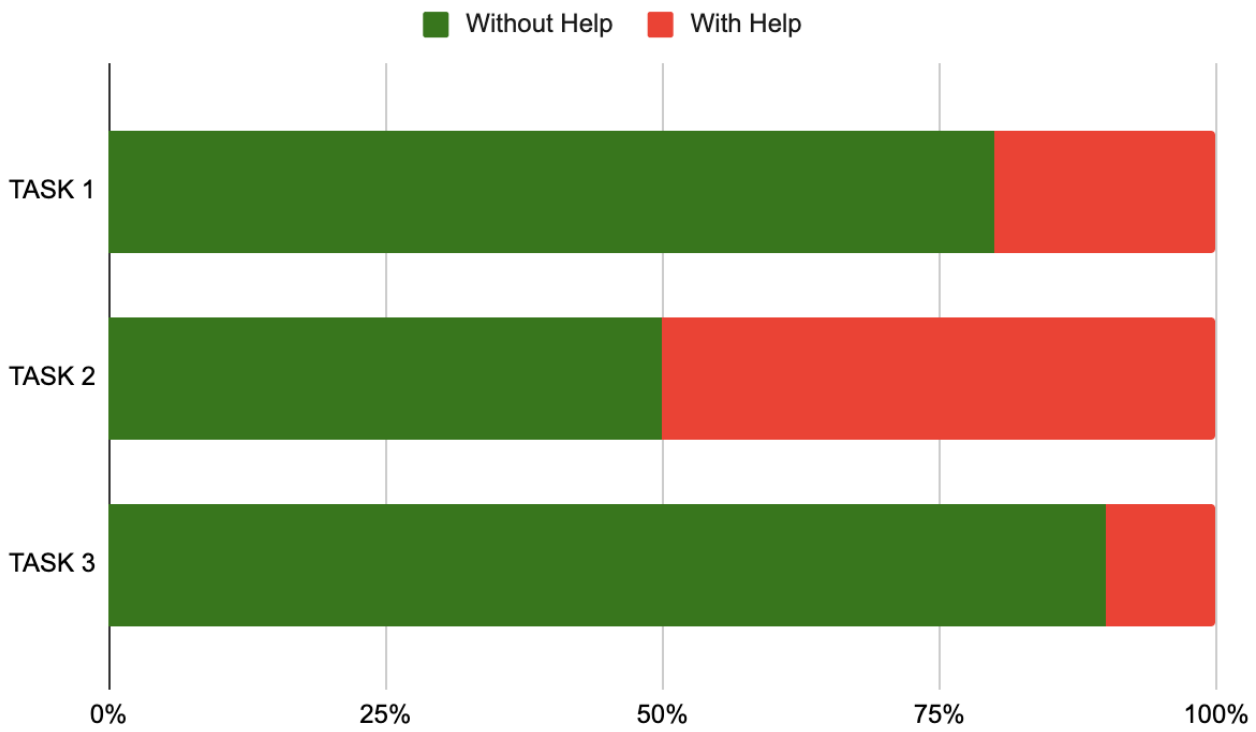


Table 4.10: User test's horizontal stacked bar chart

According to the graph above, task 2 was the most difficult to complete. In fact, half of the users requested tips to successfully complete the task. On the other hand, task 3 was the easiest to resolve without help, with almost 85% of users succeeding.

4.3 Psychometric Questionnaire

The Psychometric Questionnaire involved 32 participants. It is a quantitative test that asks users to rate their satisfaction with six aspects of a product or service on a scale of 1 to 6. The six aspects are:

- Useful
- Clear
- Informative
- Beautiful
- Intuitive

This questionnaire is based on the Cabitza-Locoro Psychometric questionnaire. The higher the score, the greater the level of satisfaction with each aspect.

In the table below, the overall value achieved from the survey for each feature is displayed.

	1	2	3	4	5	6
Useful	0	1	2	8	15	6
Clear	0	4	5	4	11	8
Informative	2	5	2	7	10	6
Beautiful	0	0	3	9	12	8
Intuitive	1	0	6	4	13	8
Overall	0	1	3	7	16	5

Table 4.11: Psychometric questionnaire's results

Based on the answers provided, it is possible to calculate various statistical metrics.

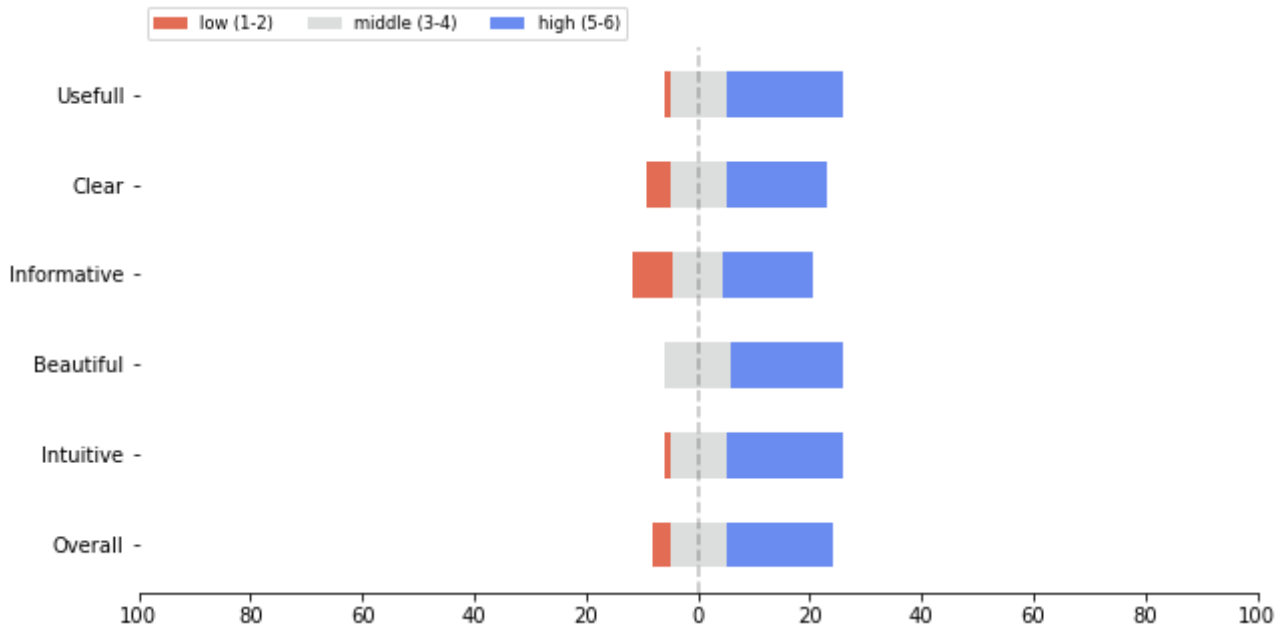


Table 4.12: Psychometric questionnaire's divergent bar chart

As illustrated in the image, the graph is a divergent stacked bar chart that divides the answers into three categories: low, middle, and high:

- Low category includes all answers with a value between 1 and 2
- Middle category includes all answers with a value between 3 and 4 (considered uncertain)
- High category includes all answers with a value between 5 and 6.

Corrplot:

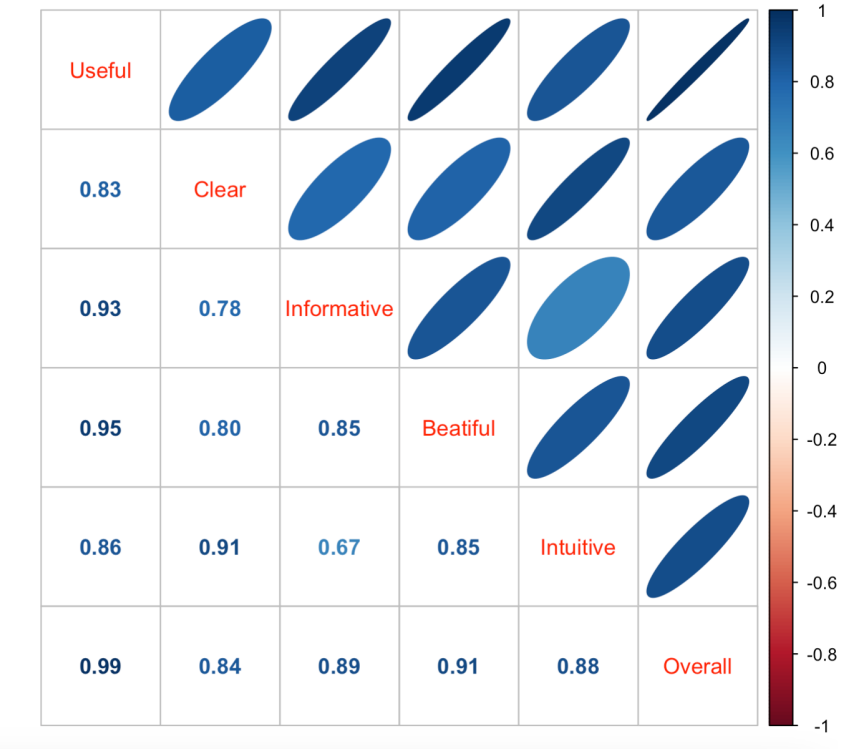


Table 4.13: Psychometric questionnaire's corrplot

The graph shown above was plotted to highlight the Pearson correlation between different features in order to assess the validity of our infographic. The most notable correlation is between the "Useful" and "Overall" features, which has a value of 0.99. The corrplot also indicates that most features have a high correlation with each other.

The lowest value among the correlations is 0.67, which is the correlation between the "Informative" and "Intuitive" features.

5. Conclusions and future developments.

The infographic we created was appreciated by those interviewed: they found it particularly informative, rich in content, and very interesting. Users found the analysis of countries with the most views in relation to the national population of particular interest. In the creation of this infographic, we dedicated ourselves to the analysis of a domain known to everyone and the number of data on this topic is far beyond imaginable; even discovering some "curious" sides of the platform. Knowing the trend of the various categories and YouTube videos could be the introduction to a potential future development that could lead both the platform and its creators to develop and trend certain videos. At the end of the project, we consider ourselves satisfied with the hard work we have done. We believe that the project has also allowed us to grow professionally, as we have been able to learn the use of a new tool such as Tableau, but above all to delve into a new field for us: data analysis and representation. We have learned techniques for identifying the most appropriate graph to represent a particular information using more datasets (since one alone is not always enough to meet requirements) aimed at representing a particular information as coherently as possible. All these efforts have also allowed us to improve existing interpersonal relationships and strengthen teamwork skills.

FIGURE AND TABLE INDEX

Source

YouTube's History and Its Impact on the Internet :

https://interestingengineering.com/culture/youtubes-history-and-its-impact-on-the-internet

https://www.businessinsider.com/history-of-youtube-in-photos-2015-10?r=US&IR=T

Censorship of Youtube:

https://en.wikipedia.org/wiki/Censorship_of_YouTube