Project Report for the Course Visual Analytics

Analysis of "Trending Videos" on YouTube from Nov. 2017 - Jun. 2018

Link to Dashboard:
https://public.tableau.com/app/profile/david.trotzky/viz/Dashboard_Group_09/Dashboard

Group: David Trotzky (149871), Konrad Schulte (149872), Luca Ludwig (149890), Niklas Steinfurth (149901)

Copenhagen, 03.01.2022

Pages: 25

Characters including Spaces: 44,993

Table of Content

1. Introduction

1.1 Project Description

As part of the "Visual Analytics" course at Copenhagen Business School a dataset on videos featured in "Trending on YouTube" is examined, transformed and visualized in the form of a dynamic dashboard that is intended to provide insights for content creators and advertisers.

"Trending on YouTube" is a list of trending videos selected by an algorithm that are displayed in a separate section on YouTube. The list is updated every 15 minutes, with videos being added and removed from the list (YouTube, 2021a). "Trending on YouTube" tries to capture the general mood of current activities on the platform by selecting videos that have the following characteristics: They "are appealing to a wide range of viewers"; they "are not misleading, clickbaity or sensational"; they "capture the breadth of what's happening on YouTube and in the world"; they "showcase a diversity of creators" and they are ideally "surprising or novel" (YouTube, 2021a). To identify videos fulfilling these criteria the algorithm considers the following, not exhaustive, signals: "View count"; "how quickly the video is generating views (i.e. "temperature"); "where views are coming from, including outside of YouTube"; "the age of the video" and "how the video performs compared to other recent uploads from the same channel". Trending videos differ between countries, but are identical within a country, i.e. the videos are not personalized to the individual YouTube visitor (YouTube, 2021a). Moreover, it is not possible for channels to pay in order to be featured in "Trending on YouTube" (YouTube, 2021a)**.**

1.2 Target Audience

With decreasing reach levels of 2-3% globally each year, linear TV is facing increasing competition from over-the-top content players such as YouTube (Nielsen, 2021). YouTube is the world's biggest online video platform, gathering "over 2 billion logged in users" on its platform each month (YouTube, 2021b). Its advertising revenue is expected to almost equal Netflix" revenue in 2021 assuming that the current growth trajectory applies (Elias, 2021). This projected USD 29 - 30 billion advertising revenue does not include YouTube's subscription revenues from "YouTube Premium" which might also be of significant size, since the offer is expected to have several million subscribers allocating 50 million subscribers of YouTube Premium and YouTube Music in total (Clark, 2021).

These numbers suggest that YouTube is becoming increasingly relevant for different stakeholders on the platform. In order to provide an overview of who might be particularly interested in an analysis of

YouTube’s most trending videos, the most relevant stakeholders were summed into two groups and their motivations described.

Content Creators / “Classical YouTubers”

Most content creators intend to increase their audience and number of monthly subscribers. The more regular viewers they have, the higher the remuneration they receive from YouTube.

Hence, current, or soon-to-be content creators could be interested in knowing which video categories are popular in their target countries. Existing content creators could use the data when considering diversifying their YouTube channel by adding videos of a different category to it. Soon-to-be content creators are even more flexible in their category choice since they may not already have narrowed down their target audience and may not have committed to a certain category of videos.

Advertisers

As the above mentioned metrics on YouTube’s advertising revenues show, the platform is becoming an increasingly relevant marketing channel. “YouTube Marketing” has become his own field in the domain of online marketing. Researchers investigate, for instance, the “virality”, the “advertising effectiveness” or the “sales effect” of YouTube videos (Tafesse, 2020).

Advertisers on YouTube could be companies, non-profit-organizations or private individuals. They might start a YouTube campaign on their own or consult specialized agencies that help with, e.g., the channel’s strategy, video conception, video production, or search engine optimization. Advertisers in general intend to optimize the effectiveness and efficiency of their marketing spend. In order to do so, a general understanding of the target audience including a high-level understanding of the target country, may be beneficial.

This analysis and the dashboard as its final product, caters to both above mentioned target groups. The dashboard is intended to provide a high-level overview of YouTube’s trending videos that can be used to derive insights for different markets and different categories.

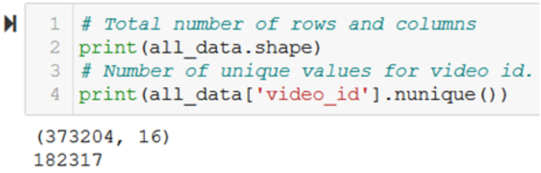
2. Methodology

2.1 Data Description

The dataset used in this project was published on “Kaggle” by a software developer named Mitchell Jolly (Jolly, 2017). Kaggle, which is owned by Google, is an online platform for data scientists and machine learning operators, where users can publish datasets. Since the users are private persons who - unlike companies or public authorities - are not subject to any obligations to comply with data quality, the trustworthiness of the data must be questioned particularly critically in this case. One advantage of this dataset is that the creator has published its source code and thus made the creation of the dataset more transparent. The data itself was collected using the YouTube API and the dataset is updated regularly (as of 28.12.2021: version 115). Kaggle itself offers two rating systems for an easier quality comparison of data sets: On the one hand, the users can give upvotes to datasets themselves. Depending on the number of votes, a dataset then receives a certain medal: From 5 upvotes a Bronze Medal, from 20 a Silver Medal and from 50 a Gold Medal (Kaggle, n.d.a). The "Trending YouTube Video Statistics" dataset used in this project is awarded with a Gold Medal. Furthermore, with 4360 upvotes, it is one of the most upvoted datasets on the Kaggle platform (8th place out of 124,392 datasets, as of 12/28/2021 (Kaggle, n.d.b). Thus, the dataset is very popular from the user's point of view. Nevertheless, trustworthiness cannot be directly inferred from this, as the voting process is not transparent (just because a user upvotes a dataset, he or she does not necessarily have to check it for data quality or trustworthiness, but may also have liked it for other reasons). The other metric for comparing quality is Kaggle's own "usability score": This mainly includes factors that serve user-friendliness and is 7.9 out of 10 points for the dataset at hand (as of 28.12.2021). This score is also above average for the Kaggle community and yet it does not include an assessment of data trustworthiness.

The data quality is therefore assessed as moderate to good due to the high popularity and transparency, contrary to the sparse information on trustworthiness.

To validate the data quality, the dataset was then checked even further:



```
# Total number of rows and columns
print(all_data.shape)
# Number of unique values for video id.
print(all_data['video_id'].nunique())
```

```
(373204, 16)
182317
```

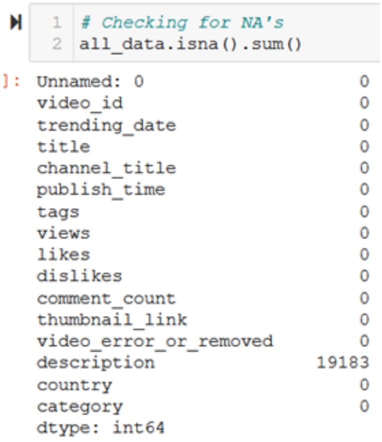
Figure 1: Data Exploration 1

The final dataset contains 16 columns (including index columns) and 373.204 rows.

At first glance, it is somewhat irritating that there are significantly fewer video ID's than the total number of entries (only about half). After checking the data, however, it turned out that this is due to the fact that some videos appear in YouTube Trends over a longer period of time, i.e. on several days, and therefore the same ID's can occur over several entries.

In addition, the dataset was checked for missing values (NA's). The result was that "description" was the only category with missing values, but this does not lead to a reduction in quality (since the mere absence of a description itself is not a problem for the subsequent analysis).

Upon further analysis of the data, it became apparent that "Nonprofits & Activism", "Movies" and "Trailers" are not represented in every country. Since the date range of the dataset starts and ends in the middle of a month (14.11.17 - 14.06.18), the dates of the dataset were checked. The number of videos per full month differed between the months. April had the lowest number of videos (45,518), whereas March (58,350) and May (57,403) had the highest number of videos.

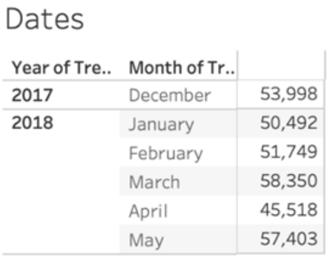


```
# Checking for NA's
all_data.isna().sum()
```

```
]: Unnamed: 0                    0
   video_id                      0
   trending_date                 0
   title                         0
   channel_title                 0
   publish_time                  0
   tags                          0
   views                         0
   likes                         0
   dislikes                      0
   comment_count                 0
   thumbnail_link                0
   video_error_or_removed        0
   description               19183
   country                       0
   category                      0
   dtype: int64
```

Figure 2: Data Exploration 2

After breaking down the number of videos per month on a country level, it was observed that the number of videos also differed between different countries. Additionally, there was no data included for Japan for December 2017 and January 2018. Due to these irregularities, it was decided not to do any analysis on trending dates throughout the project.



Dates

Year of Tre..	Month of Tr..	
2017	December	53,998
2018	January	50,492
	February	51,749
	March	58,350
	April	45,518
	May	57,403

Figure 3: Data Exploration 3

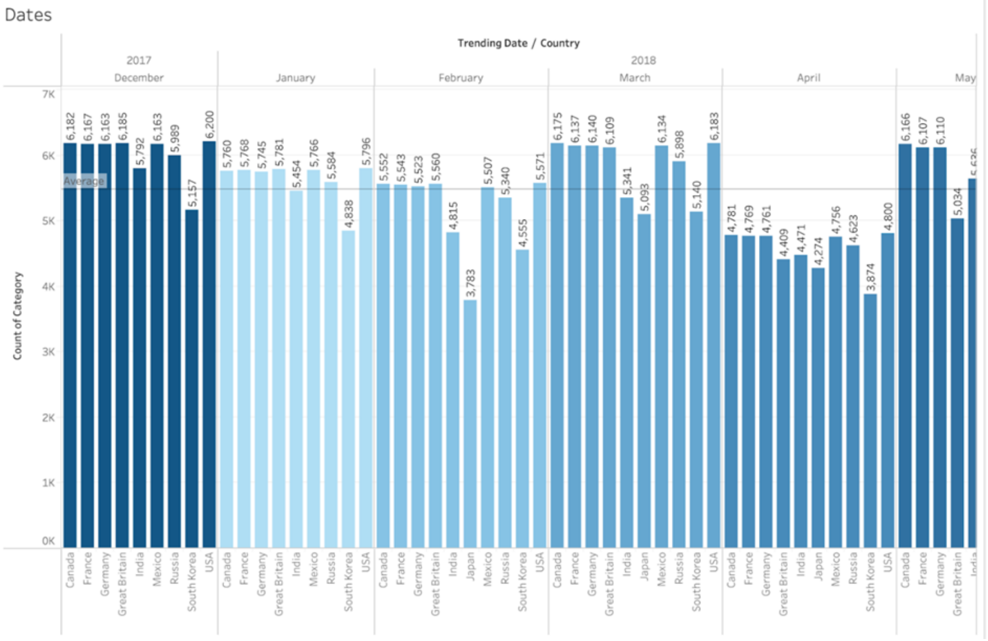


Figure 4: Data Exploration 4

Since the other dimensions of the dataset are considered reliable, the dataset was still used for the project.

2.2 Data Pipeline

Initially, all existing files were downloaded from Kaggle. The entire data package comprised 20 files, 10 of which were CSV files with the data of the respective countries and 10 JSON files with the different categories and their corresponding IDs (depending on the country, the IDs vary for the same category). In the next step, the CSV files were cleaned and transformed using the Python data analysis library "Pandas" so that the result was an entire dataset that could be used for the subsequent visualization. This process is explained in greater detail in the following chapter. With the interactive data visualization software "Tableau" the different data was visualized, and an interactive dashboard was created. Some of the findings from this project can be found in this Tableau (online) dashboard. In addition, the most important findings were recorded in this document (the written report).

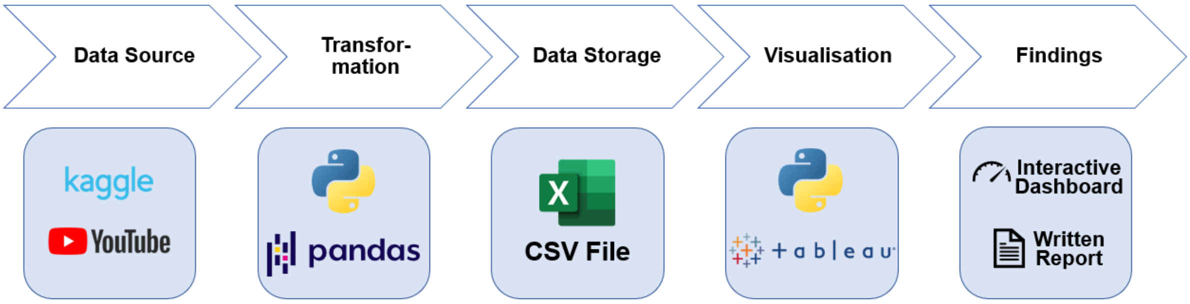


Figure 5: Data Pipeline

2.3 Data Transformation

To be able to visualize the existing data and thus derive real benefit from this dataset, the raw data first had to be cleaned and transformed:

1. As already mentioned in the previous chapter, the following files were downloaded at the beginning: 10 CSV files of the different countries (each consisting of 16 different columns and 20k to 45k rows) and 10 JSON files with information about the respective category IDs. Using the Python library "Pandas" all 10 different CSV files of the respective countries were loaded.
2. To ensure that the individual videos or entries can still be assigned to the countries later, a column with the name of the respective country was then added to each dataset of a country.

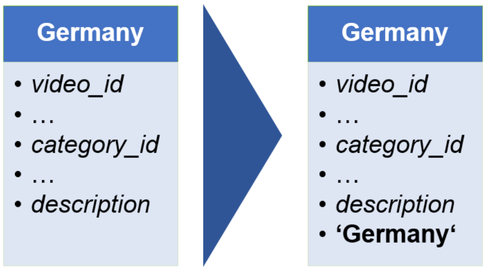


Figure 6: Data Transformation 1

3. As outlined in chapter 1.4, all 10 CSV files contained a column with the various category IDs, but these IDs differed from country to country. In order to make the data of all countries comparable, all country tables had to be merged with the respective category tables. To do so, the 10 JSON files were loaded in Python using "Pandas" and converted into a dataframe. Then all CSV files of the countries were merged with their respective category dataframes, so that the dataset of the countries at the end no longer contained an ID, but now the respective name of a category. In addition to the improved comparability, this also had the advantage that the category names (instead of the not always intuitively understandable category id's) significantly facilitate the subsequent handling of the dataset for the end user.

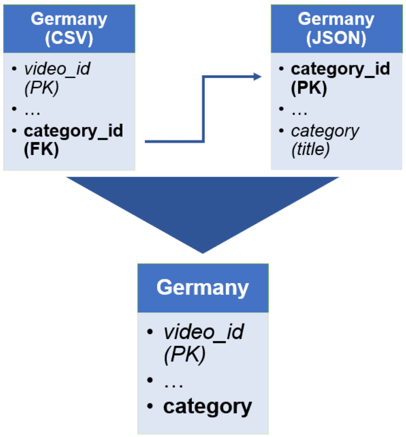


Figure 7: Data Transformation 2

4. It was checked for all countries whether the merge also worked without generating NA's or nulls. After every dataframe had passed this test, the 10 different datasets could now be merged using the "concat" function of Pandas.
5. To facilitate the later handling of the data, the column "trending_date" was converted into a "year-month-day" format. In addition, the columns "comments_disabled" and "ratings_disabled" were removed from the dataframe, as they contained negligible information for later analysis (since a check showed that ratings or comments always equal 0 (only) if they were disabled, this information was classified as redundant).

6. Altogether 10 different CSV files with 15 columns and 20k to 45k rows were transformed to one big dataset with 15 columns (index column excluded) and 373k rows and exported as CSV file in the last step.

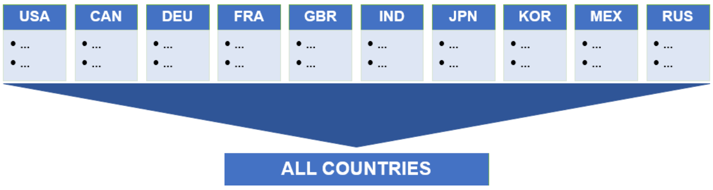


Figure 8: Data Transformation 3

2.4 Data Visualization

2.4.1 Dashboard Design

Goal of Visualization

The goal of our dashboard is to present content creators and advertising agencies various insights into YouTube's trending videos. This should enable them to quickly recognize important categories and countries when planning a new video or advertising campaign. Ultimately, our dashboard aims to support decision making for our target audience by displaying important diagrams. We present the number of videos uploaded and the number of views in the respective categories. The fractional number of categories of a country's total number of videos uploaded is also shown to emphasize the relevance of certain categories in the respective countries. We display how well specific categories are perceived in countries by displaying the like-dislike ratio. Our dashboard also gives insights about user engagement which we measure by comments per views in each category. This section explains the process of developing our dashboard, including different tools, iterations, and the reason behind the steps. Section 2.4.2 will explain the design choices and functionalities of our final dashboard.

Overview

We started the dashboard creation process with the tool Python Dash as we were interested in learning the tool. Python Dash is a Python library which allows the creation of visualizations and dashboards through code. The created visualizations can then be easily run on a localhost, making them accessible through any browser. While user centric tools as Tableau are more intuitive and easier to use, Python Dash allows a more detailed configuration and a closer confrontation with the underlying data. Therefore, much of the inspiration for the insight section comes from the conception and implementation of the dashboard using Python Dash.

We came to the realization that the well-known business intelligence tool Tableau, which we already got to know during our lectures, is better suited for the purpose of this project (see 3.2.2 for further insights on our choice of tools). Since we invested significant time to explore Python Dash, we included the dashboard creation process with Python Dash in this report, additionally to the process of our final dashboard in Tableau.

Design Process

A dashboard is an interactive tool that is perceived differently by individuals which makes following an iterative process particularly useful. This section describes the most important steps that led to our final dashboard.

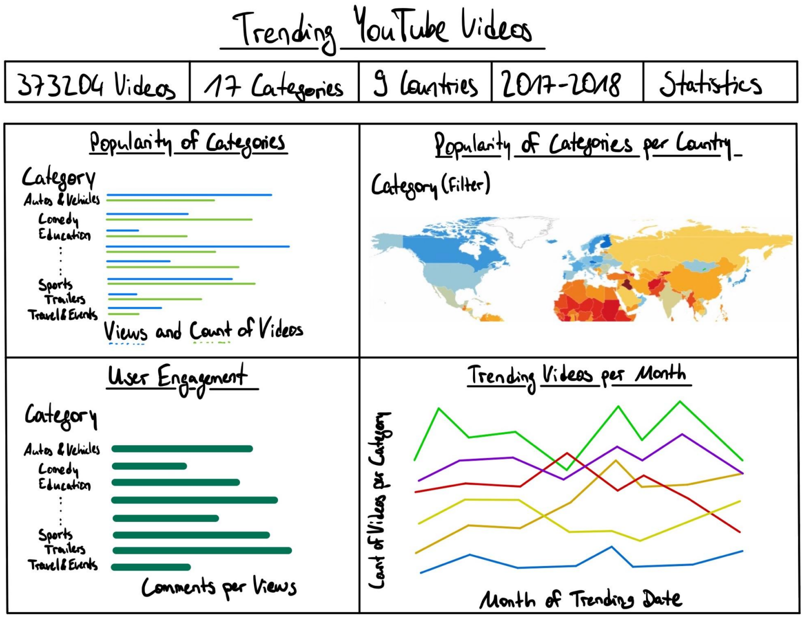
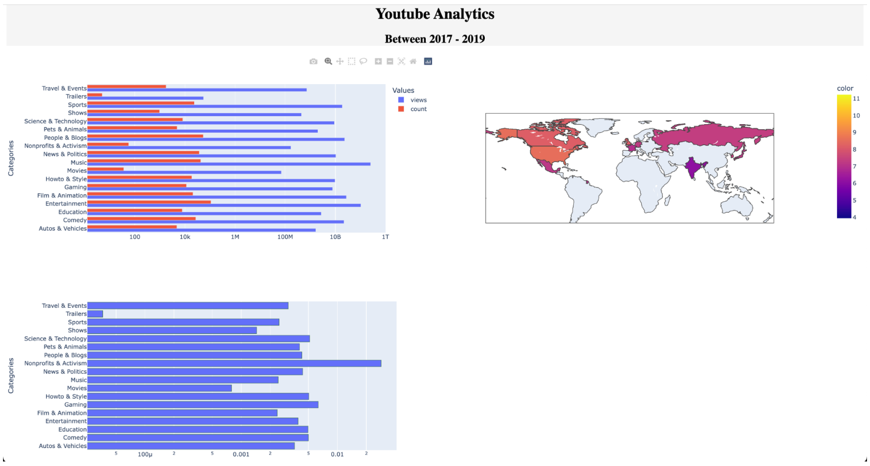


Figure 9: Initial Sketch; Date: 13.11.2021

As part of mandatory assignment 3 of this course, we made an initial sketch of a dashboard after briefly exploring the dataset. Based on the data, we defined four diagrams, three of which remained in our final dashboard.



We initially decided to implement our dashboard with Python Dash before going with Tableau. This was our first draft without any functionalities but with three diagrams which we took from our initial sketch.

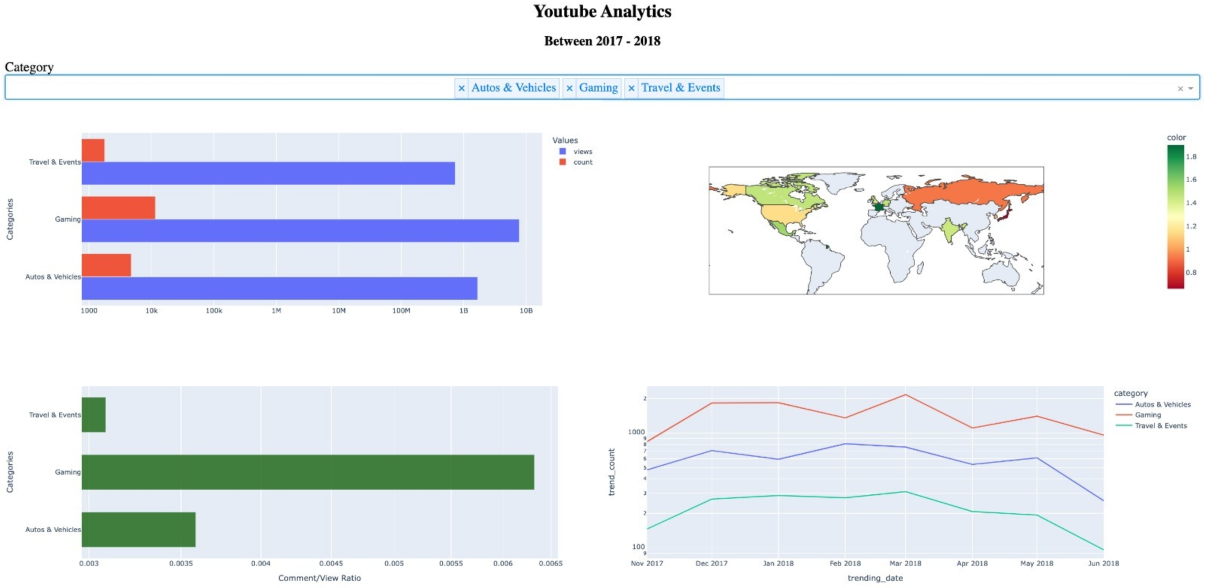


Figure 11: Second Draft with Python Dash; Date: 28.11.2021

Building on our first draft with Python Dash, we implemented the fourth diagram and the function to filter for categories. We also experimented with color schemes to help us in finding our final dashboard design. This dashboard can be found here: http://visualanalytics.pythonanywhere.com

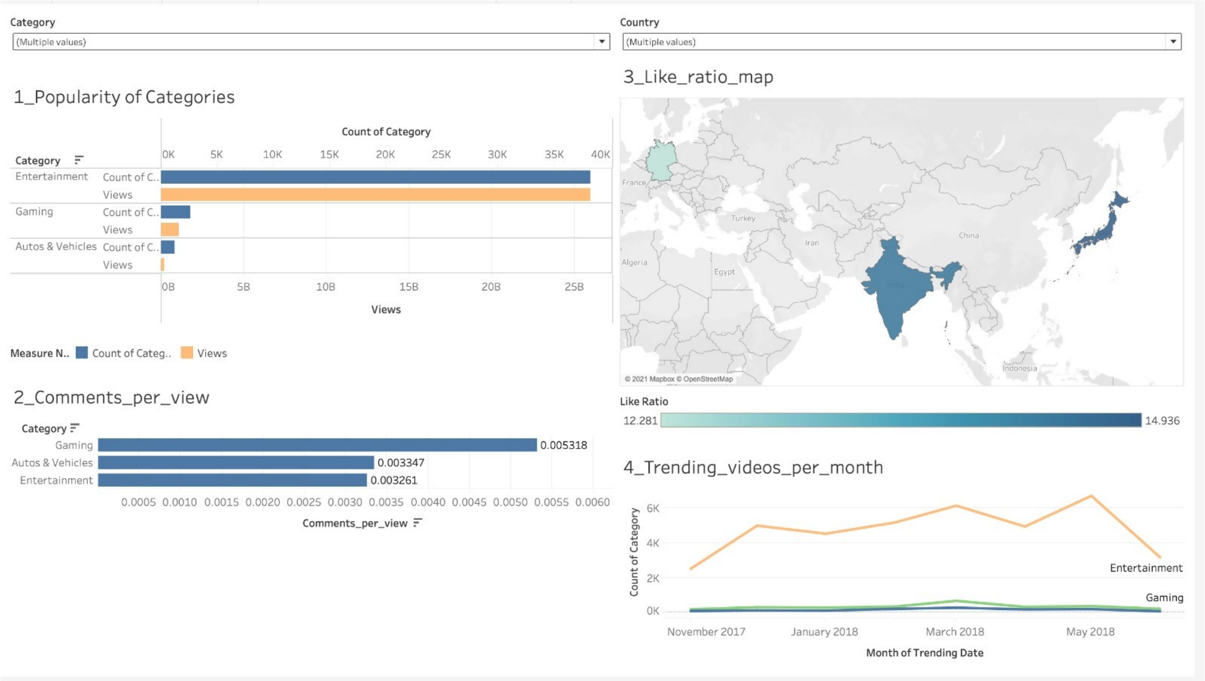


Figure 12: First draft in Tableau; Date: 03.12.2021

When we realized that it would take more time and effort to customize the Python Dash Dashboard for our needs, we decided to implement our final dashboard in Tableau. This is our first draft in Tableau in which we tried to copy our progress from Python Dash to Tableau. We noticed the more user-friendly interface and implemented a second filter option which allowed us to filter by country.

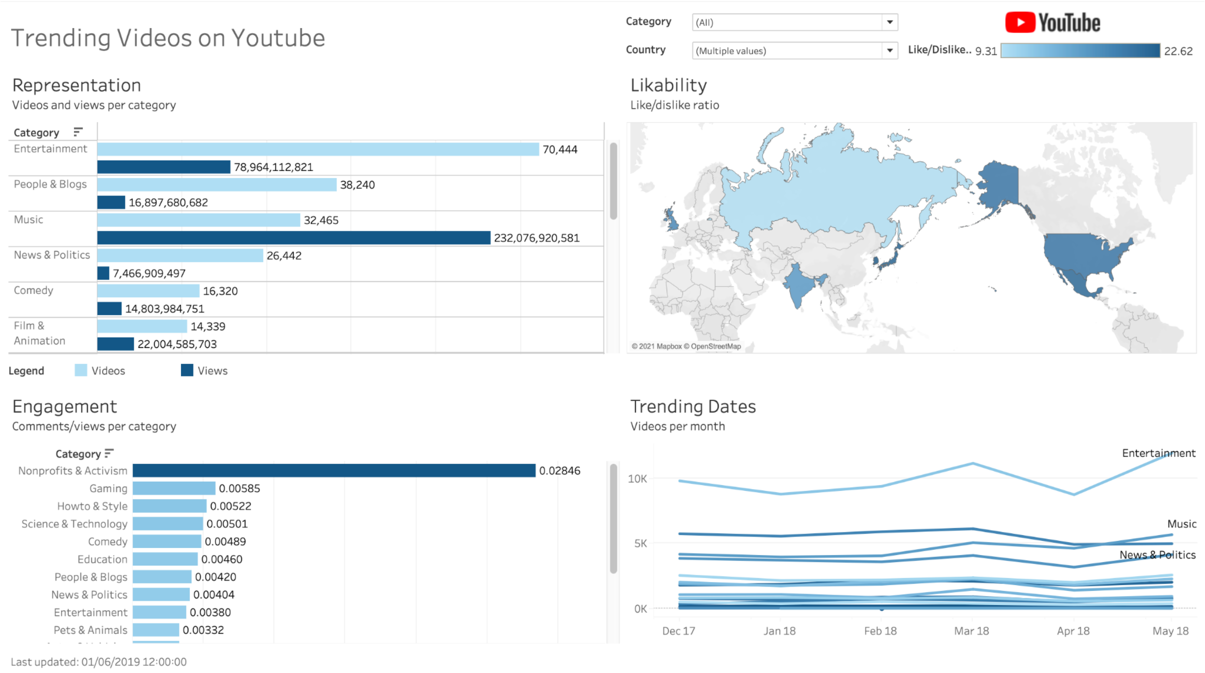


Figure 13: Second draft in Tableau; Date: 08.12.2021

We made a lot of progress with our second draft in Tableau. We added a dashboard title, changed the placement of our filters, and put the YouTube logo in the top right. We also changed the color scheme to be equal for all visualizations and generally made it much better to look at for the user.

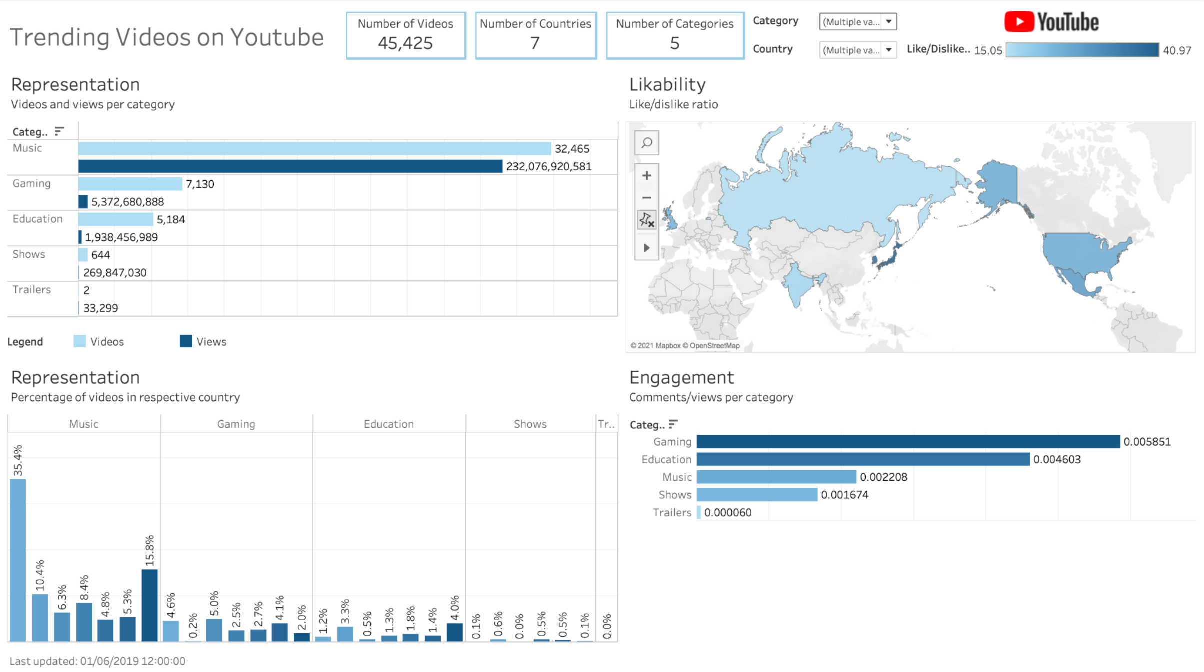


Figure 14: Third Draft in Tableau; Date: 15.12.2021

We replaced the Trending Dates visualization with a second Representation visualization that shows the percentage that each category makes up of the total amount of a country's videos. The Engagement visualization was put to the right to allow the placement of both Representation visualizations beneath each other. We also added summary text tiles that display the number of videos, countries, and categories selected. The numbers reflect the current filter selection.

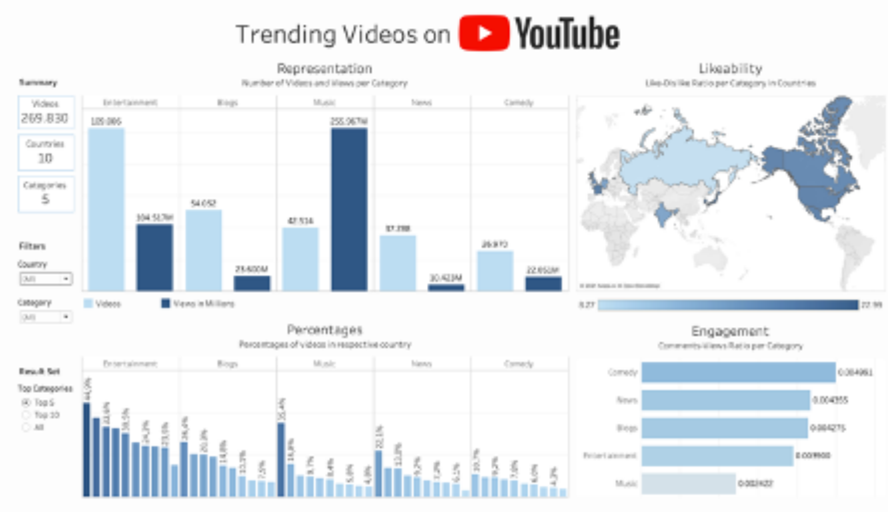


Figure 15: Final Dashboard in Tableau, Date 20.12.2021

Our final dashboard, again, saw significant changes. We changed the columns and rows for the Representation visualization to be able to fit more categories onto our dashboard. The second Representation diagram was renamed to Percentages to prevent confusing the user. We also changed the color scheme and sorted the visualization from high to low, changing the colors depending on the values. On the left, we created a toolbar which shows our summary tiles, filters, and the new function to quickly filter the result set for top categories. The intention to create a toolbar was to make the title cleaner and stand out from the rest of the dashboard. The next section explains our design choices and functionalities of the final dashboard in more detail. The final dashboard can be found under the link: https://public.tableau.com/app/profile/david.trotzky/viz/Dashboard_Group_09/Dashboard

Dashboard creation in Python Dash

The dashboard creation in Python Dash can be summarized into five main steps:

1. Data Preparation:
   To enable an efficient runtime, the necessary steps of the data preparation were executed in a separate script and saved as five different CSV files. These five tables are based on the results of the data cleaning process and have been adapted for the different graph types. In the case of the map graph, the countries had to be joined with a table from the United Nations Statistics Division, since Python Dash requires this data format for display (United Nations Statistics Division, n.d.). After the data was formatted, grouped and static KPIs were calculated, the individual files were exported to the working directory.
2. Data Connection:
   The prepared CSV files are imported via the Pandas library as data frames and assigned variables for the upcoming script.
3. View Creation:
   Five diagrams are generated via the Plotly library from the prepared data sources: An indicator bar, a singular bar chart, a multiple bar chart, a map chart and a line chart. This includes, among other things, the assignment of values to the x and y axes, color and label selection, the information given on hovering, or the configuration of a logarithmic axis.
4. Dashboard Creation:
   In the next step, the individual views are merged using the Dash library, which creates a dashboard. Attributes such as alignment, size or font are determined in the process. Through the implementation of the Dash callback functionality, a category filtering on all graphs could be enabled. This requires defining the behavior of every attribute after the selection of one or more categories. Thus, each of the underlying data sources is filtered according to the categories selected by the user. The individual graphs are then redefined with

the filtered data source. The mapping to step 4 is done by defining an id for each graph. This allows all parts of the dashboard to be dynamically filtered, calculated and adjusted.

5. Publishing:
   Running the designed dashboard on a local host through the Dash "run.server" method.
   In order to be able to use the dashboard externally without effort, an online presence was set up via Pythonanywhere after the final iteration was reached. For that purpose, the required libraries had to be installed via bash terminal, a file structure set up and a configuration file written. While the prepared data sources were stored locally in the use case of Jupyter Notebook, they were uploaded to the server provided by Pythonanywhere when the online presence was set up and thus integrated into the file structure.

Dashboard creation in Tableau

The dashboard creation in Tableau can be summarized into four main steps:

1. Data Connection:
   We established a live connection to the csv-files created in our data transformation process.
2. View Creation:
   We first created the diagrams in Tableau on their own. To display the information we wanted, we had to create calculated fields in Tableau. These fields included "Comments_per_view" and "Like/Dislike ratio" which were used in our dashboard. Other calculated fields were created for analytic purposes. We chose three bar charts ("Representation", "Percentages", "Engagement") and one map chart ("Likeablitity") for our dashboard.
3. Dashboard Creation:
   After we finished each single view, we created the dashboard and made everything interact with each other. Filters used in a single view were made globally to affect the entire dashboard. We created a toolbar which holds the global filters "Category", "Country", and "Result Set". While and after building the interactive functionality of our dashboard, we made sure to maintain a clean and simple design which is further described in section 2.4.2.
4. Publishing:
   We published our dashboard on Tableau Public. To do this, we changed the data source connection from a live connection to an extract connection as it would not function otherwise. After changing the data connection type, the dashboard including the data was successfully uploaded on Tableau Public.

2.4.2. Design Choices and Functionalities

Design Choices

Since our dashboard is mainly intended for content creators and advertisers, it should allow them to quickly derive the main information of the underlying data. Therefore, our dashboard presents easy to read diagrams which we present through simple visualizations. We chose to make the dashboard title stand out from the rest in making the font size greater than the rest and exclude any distracting elements to the left or right. Those steps were taken to reduce the level of complexity in the dashboard and improve factors such as engagement and memorability (Harrison, Reinecke and Chang, 2015). The dashboard is split into three parts that differ in size. The left part serves as a tool bar containing summary tiles representing the number of videos, countries, and categories, depending on the selected number of categories and countries. We provide two filters to select single or multiple countries and categories. It is also possible to quickly choose the result set by selecting either the top five, top ten, or all categories. We display the top five categories by default. This is done to prevent an information overload to the user and maintain a clean, easy to look at dashboard. The top categories are calculated by the number of videos in the respective category in the dataset. As we faced some space related issues to display the data for all 18 categories at once, the top five or top ten options provide a good compromise between maintaining a clean dashboard and keeping as much important information as possible. It is still possible to select additional categories with our filter function. The center and right part of our dashboard display the four diagrams which are described in the next section. We put the Representation diagram above the Percentage diagram as both visualizations use the number of videos uploaded and display information regarding the representation of videos in each category. The Likeability and Engagement diagrams are on the right side because their information differs from the center part and aims to display user related information in terms of perception and user engagement behavior. We make use of Tableau's tooltip function in all diagrams. This is done to display additional information that is not as important as the initial information shown but still of great use to the user. We also make use of aliases in Tableau to save space in the diagrams. The country "Great Britain" is renamed to "UK". We also rename the following categories: "Howto & Style" to "Style", "News & Politics" to "News", "Nonprofits & Activism" to "Activism", "People & Blogs" to "Blogs", "Pets & Animals" to "Animals", "Science & Technology" to "Science", "Travel & Events" to "Events".

Choice of Diagrams:

Representation

We display the number of videos uploaded and the number of views for each category. This information is simple but very important for our target audience as it allows to quickly determine the most popular categories for both measures. We use two colors to represent the number of videos uploaded and the number of views, respectively, as a categorical representation. The tooltip of this view displays additional information in breaking down the exact number of videos and views per country in the respective category.

Percentages

This diagram relates to the Representation KPI regarding the number of videos uploaded. It provides important information in setting the videos uploaded in each category into relation to the total number of videos uploaded in this country. We display this information by using percentages. Our intention behind the diagram is to enable the user to identify which categories are most represented in each country in terms of the fractional number of videos uploaded. The tooltip gives information about which bar displays which country. Considering the space of the dashboard, we chose to remove labels for each bar. Displaying labels would bring no net benefit to the user as it would be an information overload and harm our dashboard's simple and clean visuals, especially when selecting more than five categories.

Likeability

This chart is intended to display how well categories are perceived in each country. We measure this perception or likeability by calculating the ratio between the sum of likes and dislikes per category for each country. A lower value means that the number of likes was not much higher than the number of dislikes and, therefore, the videos in this category may not be perceived very positively by the viewers. A higher value would represent that the videos in this category are liked more by the viewers. If multiple categories are selected, we calculate the average value for the categories in each country and display this value. The tooltip for each country provides the ratios for every selected category and the overall ratio for the specific country. We use the common "Mercator" projection for this map and display the entire world to fit every country in our dataset onto the map (Futuremaps, 2019). We think the relation between likes and dislikes is of greater use to our target audience than the total number of likes and dislikes. Our reasoning is that total values can easily distort the perception while a ratio instantly provides the correct relation between the likes and dislikes. We do acknowledge that some creators or advertisers might sometimes only be interested in the highest possible traffic for their videos and, therefore, may not care as much about a positive perception.

Engagement

We measure user engagement by calculating a ratio between comments and views. We display the values per category from highest to lowest. The tooltip contains country specific information as it displays the comments-views ratio for the selected category for each country. When creating a video or advertisement, the total number of views or the perception of it might sometimes not be as relevant as the interaction that it creates with the users. Therefore, we chose this diagram to enable our target audience to determine the categories that attract the highest level of interaction in terms of comments per view. We also considered other KPI's such as comments per video. However, this measure is more prone to errors in our opinion. For example, when looking at comments per video, there might be two videos A and B with the same number of comments. Video A might have 1 million views and B might have 50,000 views. In that case both videos would have the same value for comments per video, but different values for comments per view. Video B would have a higher ratio of comments per view reflecting the higher user engagement it generated.

Color Choice:

Our dashboard uses the same color scheme for all visualizations. We do this deliberately to maintain a clean and understandable look. We opted for a blue color palette to make our dashboard accessible to the majority of users. Green and red were deliberately excluded to accommodate the most common type of color blindness, namely Deuteranopia (Domingues, 2021). Even for users with a blue related color blindness, our dashboard should still be accessible as we only use one color. Therefore, the user does not have to distinguish between blue and affected colors such as yellow. We use no diverging color scheme because our data has no zero values or extreme baselines that would justify the use of diverging color schemes (Adobe, n.d.). The allocation of our colors is value based, giving higher values a darker shade and lower values a lighter shade of blue (Muth, 2021). We do this as all our diagrams use numerical measures that can easily be sorted. The only exception is the Representation diagram which uses categorical colors for the number of videos and the number of views in millions. We allocate a darker blue to views in millions and a lighter blue to the number of videos because the number of views in millions is numerically higher than the number of videos. We do not use categorical color schemes in the rest of the visualizations as our data contains 18 categories. Using colors that continuously represent a specific category would bring no benefit to the user and make the dashboard harder to understand because of the high number of categories (Adobe, n.d.).

Mobile View:

Regarding accessibility, we chose to create a mobile view of our dashboard which is optimized for the Tableau mobile app. Especially considering our target group, it makes sense to enable users to view our dashboard on a phone. Content creators typically use their phone a lot and will profit from being able to retrieve our provided information while on the go.

Functionalities

Our dashboard provides various functionalities. As described above, the left side of our dashboard serves as a toolbar. The summary tiles display the number of videos, countries, and categories according to the set filters. The numbers change depending on how many categories and countries are selected. We provide two filters to select single and multiple countries and categories. The selection will affect every visualization on our dashboard as each diagram depends on the set filters. This allows for exploring specific countries and categories in more detail and presents the opportunity to compare specific countries or categories. We make it possible to quickly change the result set by selecting the top categories. The options are to select the top five, top ten, or all categories. It is still possible to add other categories by using the category filter. Our Likeability diagram also serves as a filter. It is possible to click on single or multiple countries which will then change the country filter and affect the other diagrams and the summary tiles. Clicking on a non-selectable point in the map will remove the current filter. We use the tooltip function in Tableau a lot to provide additional information. Tooltips can be viewed by hovering over a specific data point in a diagram. The Representation and Engagement diagrams use tooltips to display country specific information which is not visible in the diagram at first glance. The Percentages diagram uses tooltips to show which data point belongs to which country because of the space related issues mentioned above. The Likeability map uses tooltips to display category specific information and provides the user with a detailed breakdown of the values.

3. Discussion

3.1 Insights and Recommendations for Target Audience

3.1.1 Insights

To derive insights from the dataset, the created dashboard was applied. To increase the readability of this report additional visualizations were created. Although these may not be part of the dashboard itself, most of the respective insights could be derived through the application of the dashboard as well.

Representation and Percentages

The metrics “count of videos per category in total”, “views” and “percentage of videos in respective country” could be used to evaluate the representation of a certain category in a country.

The category “Entertainment” had by far the most videos in the dataset (ca. 109,000), followed by “People & Blogs” (ca. 54,000) and “Music” (ca. 42,500). “Nonprofits & Activism” (57), “Movies” (36) and “Trailers” (5) had the fewest number of videos in the dataset.

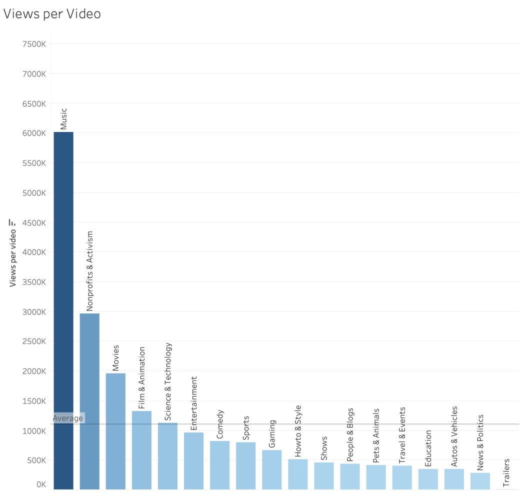


Figure 16: Views per Video

“Music” videos had the most views in total (ca. 256 bn views), followed by “Entertainment” (ca. 105 bn views) and “Film and Animation” (ca. 27 bn views). When dividing the total views by the number of videos per category, the metric “views per video” was obtained, shown in figure 16. “Music” videos had the most views per video, followed by “Nonprofits & Activism” and “Movies”. “Autos & Vehicles”, “News & Politics” and “Trailers” had the lowest views per video ratios. Out of the 10 countries in the dataset, Great Britain had the most video views followed by the USA. That is surprising given that the number of inhabitants of the United States is more than four times the amount of Great Britain. As can be seen in the percentage of videos in respective country, a possible reason for that might be the high amount of “Music” videos placed in the trending videos.

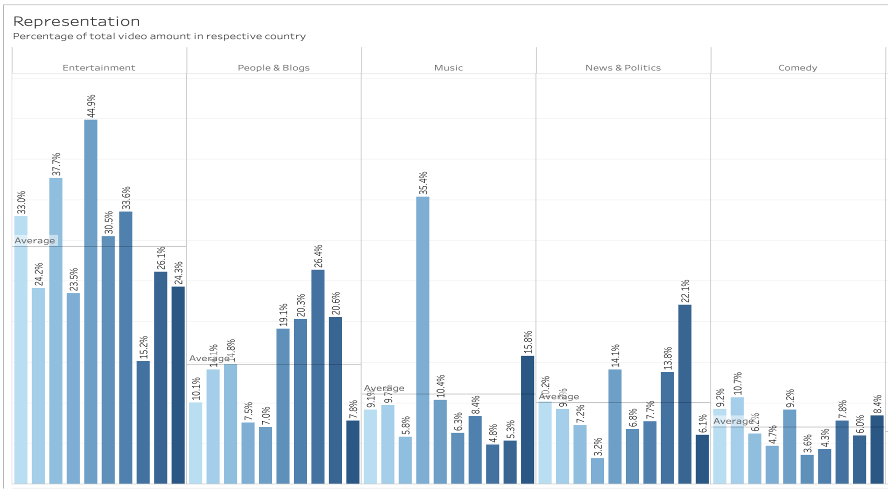


Figure 17: Percentage of videos in respective country for the five biggest categories by video amount

Figure 17 shows the percentage of videos in respective country for the five biggest categories by video amount. The shares of the different categories per country differed. Thus, some categories were represented stronger in certain countries than in other countries. The outliers in both directions, i.e. a significant higher or lower representation of a category, for the five biggest categories by video count were especially worth mentioning: The category "Entertainment", was relatively underrepresented in Russia. Only 15.2% of the most trending videos in Russia fell into the category "Entertainment". In India, on the other hand, "Entertainment" videos made up for 44.9% of all trending videos. Videos in the category "People & Blogs" were highly represented in Russia (26.4%), but low represented in Great Britain (7.5%), India (7%) and USA (7.8%). One of the strongest outliers could be observed in the category "Music". In Great Britain 35.4% of all videos were "Music" videos, which was 24.3%-points above the average representation for the Music category (11.1%). "News & Politics" were featured often in South Korea (22.1%) but were low represented in Great Britain (3.2%). The category "Comedy" was rather equally distributed across the different countries.

Likeability

The KPI "Like/dislike ratio" was used to describe the "likeability" of a category.

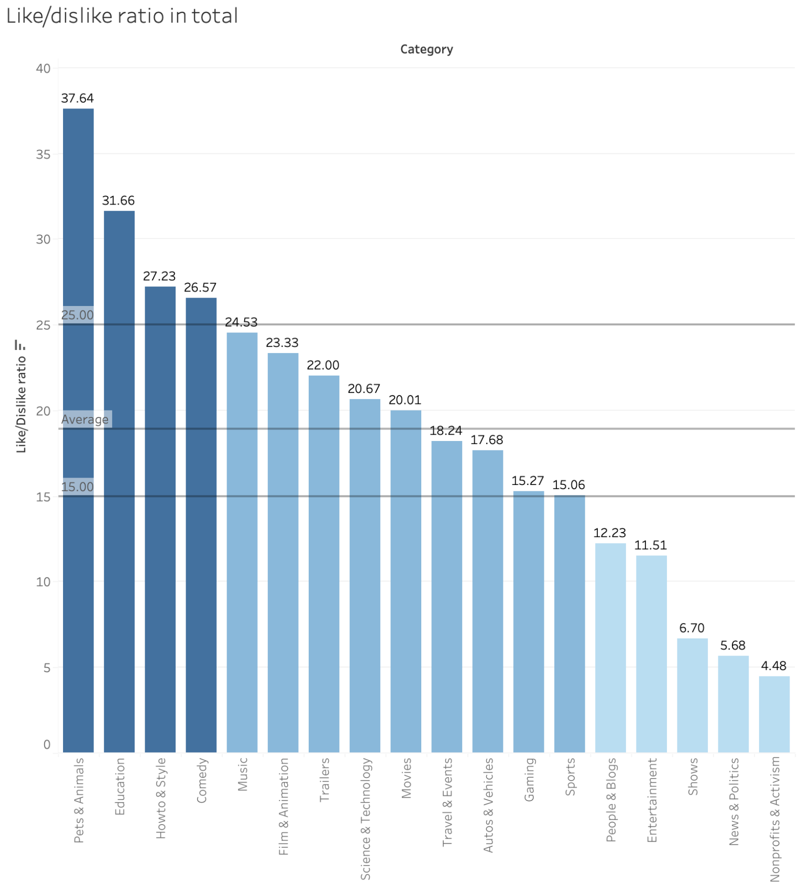


Figure 18: Like/dislike ratio per Category

The categories “Pets and Animals” (37.64), “Education” (31.66) and “Howto & Style” (27.23) had the highest like/dislike ratio overall, as shown in figure 18. “Shows” (6.70), “News & Politics (5.68) and “Nonprofits & Activism” (4.48) had the lowest like/dislike ratio. Of the three biggest categories by videos, “Music” had the highest like ratio (24.53), which was above the average. “People & Blogs” (12.23) and “Entertainment” (11.51), on the other hand, had rather low like/dislike ratios which lies below the average.

Going into more detail, the like/dislike ratios were split on a country-level. This revealed the like/dislike ratios were rather equally distributed across the five biggest categories, displayed by figure 19. In contrast to that, the like/dislike ratios of the smallest categories showed some outliers in figure 20. “Shows” in Great Britain, for instance, had the highest like/dislike ratio of the dataset (130.8). In France, “Travel & Events” had an unusually high like/dislike ratio (79.4).

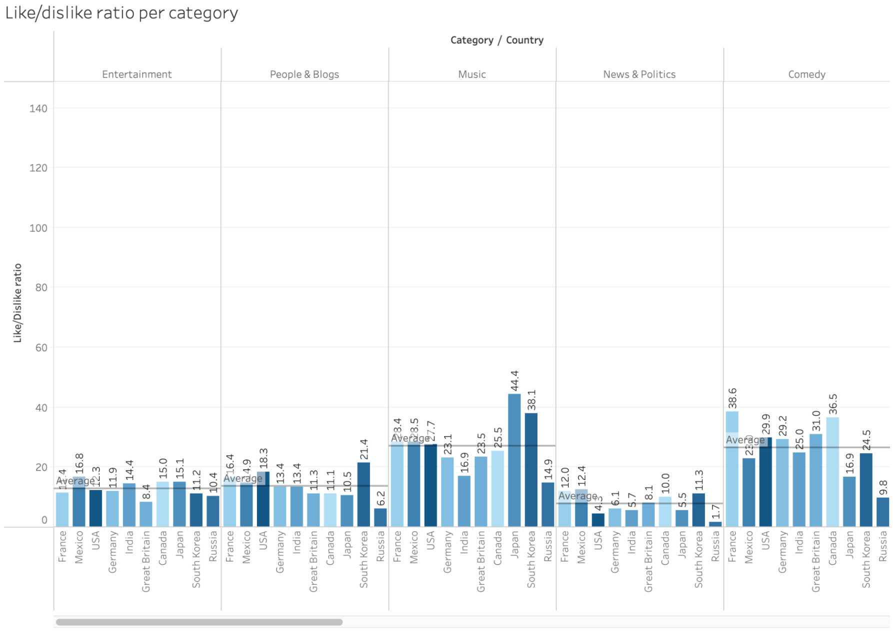


Figure 19: Like/dislike ratio for five biggest categories

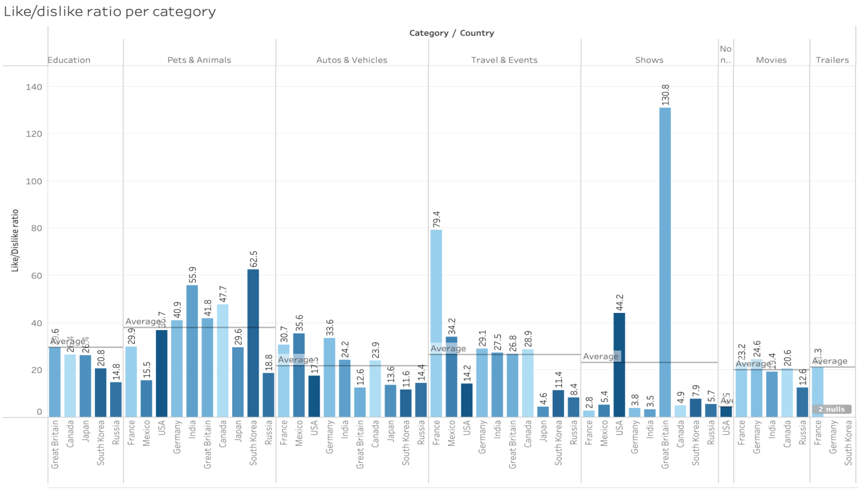


Figure 20: Like/dislike ratio for five smallest categories

Engagement

When evaluating the engagement of the different categories, the metric "Comments per views" was looked at. "Nonprofits and Activism" had the highest comment per view ratio (0.02848 comments per view). Reversely calculated this accounted for 35 views per comment, which means, a video in that category received a comment with every 35th click. As shown above the category did only include 57 videos in total, with all of them coming from the USA-charts. In addition, all of the top 8 videos were from the YouTube channel "Logan Paul Vlogs". Thus, the seemingly high engagement observed in this category may not be representative.

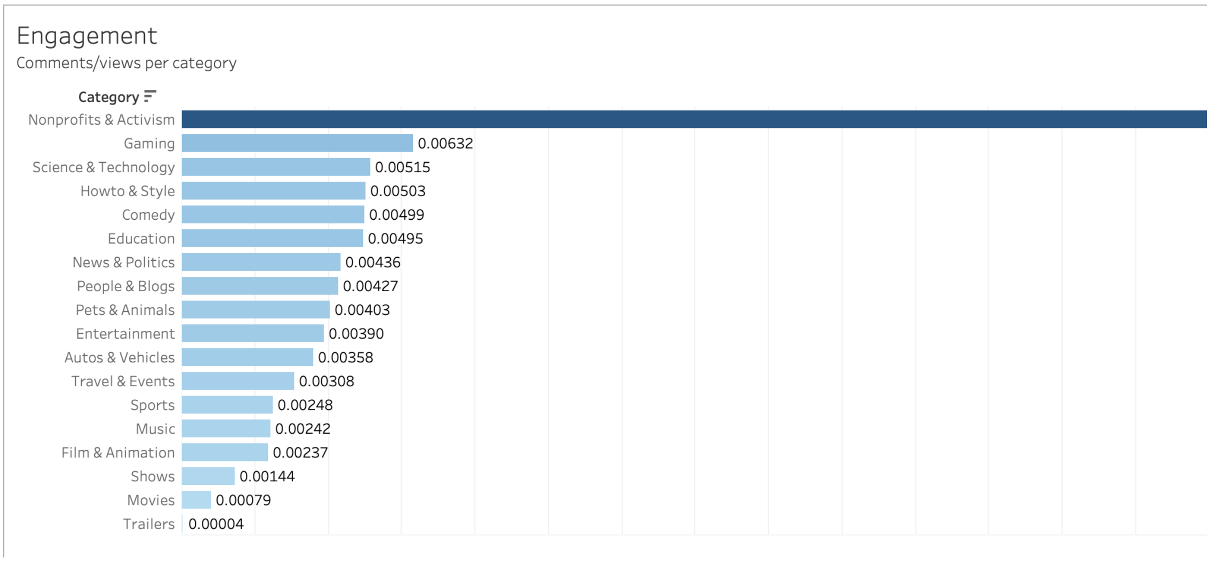


Figure 21: Comments per views per category

The categories with the highest user engagement after "Nonprofits and Activism" were: "Gaming" (0.00632), "Science and Technology" (0.00515) and "Howto & Style" (0.00503). When only looking at the three biggest categories by video count, "People & Blogs" had the highest comments/views ratio (0.004275), followed by "Entertainment" (0.003900) and "Music" (0.002422). Overall "Shows", "Movies" and "Trailers" had the lowest comments/views ratio but were also very small in size.

3.1.2 Recommendations

In 3.1.1 a rather broad overview of the dataset was given. The three main topics of the dashboard, "Representation and Percentages", "Likeability" and "Engagement" were elaborated in greater detail. Specific recommendations to the target groups (Content Creators/YouTubers and Advertisers) would depend on the respective use case. In general, it is recommended to "play around" with the dashboard to discover insights that fit the individual needs.

Nevertheless, the user should keep several things in mind, when using the dashboard:
When using the dashboard all graphs should be seen as a collective. When looking at the user engagement metric of a certain category, for instance, the respective size and relevance of that category in a certain country should be considered. The user should also keep in mind that when selecting multiple categories and countries, some KPI's are aggregated. When selecting multiple countries, the respective numbers of videos and views are added for "videos and views per category" and averaged for "Engagement". When selecting several countries, the like/dislike ratio is averaged as well.

3.2 Tool Choice

3.2.1 Pandas vs. Alteryx

In this project, the Python data analysis library "Pandas" was used for the data transformation. Another option would have been to use the "Alteryx" program. Both variants have their advantages and disadvantages, which are explained in more detail in the following section.

A major advantage of Alteryx is its ease of use. Due to the intuitive user interface, it is possible to perform data transformation even without extensive programming knowledge. In addition, the workflows in Alteryx can be saved and transferred. Especially in the case of merging all datasets of the countries with their respective category lists, this iterative process would have been useful and would have saved some time.

However, Python with its Pandas library is free, while an Alteryx license can be quite expensive. Nevertheless, Pandas offers at least the same, if not more possibilities and a strong helpful community (e.g. StackOverflow), which has already solved most problems and answered questions. But what ultimately led to the decision for Pandas and against Alteryx was the realization that by programming yourself you get a deeper insight into the background of the data transformation processes and thus a better understanding of the data transformation as such, especially its pitfalls and hurdles.

3.2.2 Dash vs. Tableau

At first glance, Tableau seems like the logical choice for creating a dashboard of this scope. Columns are quickly added, graph types selected, and colors adjusted. What led us to first take a different path? We started out with the Python Plotly library to create the individual graphs and the Python Dash library to integrate the different visualizations into a single interactive dashboard. In short, the technology will be referred to as Dash in the following for simplification reasons. The motivation to start out with Dash was derived from a motivation of practicing a new technology. In fact, the mentioned libraries offer a wide field of applications and are useful in the career of a future Data Scientist. In addition, Dash has a wide range of capabilities compared to Tableau. Especially in the conception phase and at the beginning of the creation, a wide range of configuration options seemed particularly attractive. Thus, a tool like Dash also drives creative inspiration, while Tableau contains many preset options. While these are efficient and helpful, they can also limit creativity to some extent.

After the initial dashboard was in place and we added the category filter, it slowly became clear that linking multiple filters in Dash required an increased effort. Furthermore, resetting filters is not implemented by default. Finally, the feature changing views in a set of connected views also took a workaround and therefore more time than initially expected. Thus, during the project we decided that Tableau's standard toolbox was sufficient for our needs.

In addition to the aforementioned functionalities, Tableau has a consistent design standard that must be configured individually in Dash. Especially in terms of accessibility, this offered us an immense advantage, allowing us to focus more on the underlying literature and its implications. Lastly, as our Dash dashboard gained functionality over time, the underlying code became more complex. The more code we had in Python Dash the harder it was to co-work on the dashboard. Along with this came the code from *Pythonanywhere*, which allows us to host the dashboard online. Tableau, on the other hand, provided a user-friendly interface at all stages of development, so that all group members could modify the dashboard at any time without the need to introduce the change made to the rest of the group.

4. Summary

In the context of the course "Visual Analytics" at Copenhagen Business School, a dataset on trending YouTube videos was chosen to be visualized.
First, the dataset was described and validated. Due to some irregularities with regard to the number of videos per month and per country, it was decided not to do any analyses that focus on the "Trending Date" of videos. All other dimensions were considered reliable. After the validation, the dataset, which included separate files for countries and category names, was cleaned and merged into one dataframe using Python's Pandas library. The final dataset was then exported as a CSV-file, ready to be used for visualization purposes.
At the beginning of the visualization process the target audience of the dashboard was defined as two groups: Content Creators/"Classical YouTubers" and Advertisers. It was assumed that both groups would be interested in a dashboard that gives a broad overview of all available video categories in all available countries. After having defined the target audience, the KPI choice was discussed. Once the

first KPI’s were defined, the dashboard was sketched with pen and paper and afterwards implemented with Python’s Dash library, since the group wanted to learn to use the tool. After realizing some disadvantages of Python Dash another dashboard was built using the software “Tableau”, which facilitated co-working on the dashboard. The final dashboard was finished in Tableau. It focused on the three dimensions “Representation and Percentages”, “Likeability” and “User Engagement” of trending YouTube videos, which could be globally filtered by category and country. The tooltip allowed the comparison of categories or countries for aggregated KPI’s. Furthermore, accessible colors were chosen and an uncluttered mobile view was created. Lastly, the final dashboard was published in Tableau Public to make it accessible to a wider audience and provide benefit to our target group.

After having finished the final version of the dashboard, the dashboard was applied to provide insights and recommendations. Additional Tableau views were created for the purpose of this report and a brief qualitative overview of the dataset was given. Since the purpose of the dashboard was to give a rather high-level overview of the respective categories and countries and to be applicable to many different use cases, it was avoided to point out specific recommendations.

References

Adobe. (n.d.). *Color*. Retrieved 15.12.2021 from https://spectrum.adobe.com/page/color/

Clark, M. (2021, September 2). *YouTube reports having 50 million Premium and Music subscribers*. The Verge. https://www.theverge.com/2021/9/2/22654318/youtube-50-million-premium-music-subscribers-streaming-services

Domingues, M. (2021, September 9). *How to Color your Data Wisely*. https://www.biztory.com/blog/how-to-color-your-data-wisely

Elias, J. (2021, April 28). *YouTube is a media juggernaut that could soon equal Netflix in revenue*. CNBC. https://www.cnbc.com/2021/04/27/youtube-could-soon-equal-netflix-in-revenue.html

Futuremaps. (2019, August 31). *Top 10 World Map Projections*. https://futuremaps.com/blogs/news/top-10-world-map-projections

Harrison, L., Reinecke, K., & Chang, R. (2015). *Infographic Aesthetics: Designing for the first Impressiong*. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems.

Jolly, M. (2017, November 13). *Trending YouTube Video Statistics*. Retrieved 10.11.2021 from https://www.kaggle.com/datasnaek/youtube-new

Kaggle. (n.d.a). *Progression*. Retrieved 10.11.2021 from https://www.kaggle.com/progression/

Kaggle. (n.d.b). *Datasets*. Retrieved 10.11.2021 from https://www.kaggle.com/datasets?sort=votes

Muth, L. C. (2021, March 16). *When to use sequential and when to use diverging color scales*. https://blog.datawrapper.de/diverging-vs-sequential-color-scales/

Nielsen. (2021). *Tracking the Evolution of Global TV Viewing*. https://www.nielsen.com/us/en/insights/article/2021/tracking-the-evolution-of-global-tv-viewing

Tafesse, W. (2020). *YouTube marketing: How marketers" video optimization practices influence video views*. Internet Research, *30*(6), 1689–1707. https://doi.org/10.1108/INTR-10-2019-0406

YouTube. (2021a). *Trending on YouTube—YouTube Help*.

https://support.google.com/youtube/answer/7239739?hl=en

YouTube. (2021b). *YouTube Playbook for Small Business*. https://services.google.com/fh/files/misc/youtube_smallbusiness_playbook_v2.pdf