

**Tecnológico de Monterrey**

[**Data Analytics Boot Camp**](https://bootcamp.tec.mx/data/)

**Project 1**

*“Youtube Analysis”*

**Team members:**

Loncho Lozano

Cesar Cruz

Maria Jose Garcia

Marcela Maldonado

**Introduction**

The most popular videos on YouTube, the well-known video sharing website, are kept on a list. According to Variety magazine, "YouTube employs a combination of metrics, including analyzing users interactions, to select the year's top-trending videos" (number of views, shares, comments and likes). Keep in mind that they are not the videos with the most views overall for the year. Music videos (such the notoriously lewd "Gangam Style"), celebrity and/or reality TV performances, and the ubiquitous "random dude with a camera" viral videos are among the top performers on the YouTube trending list. This dataset keeps track of the most popular YouTube videos on a daily basis.

The goal of this study was to analyze a YouTube Dataset. This information is used to examine trends in YouTube video usage, including the most popular video categories, top-performing channels, and top viewing nations. The information helps content producers better understand their audience and the kinds of videos that are most likely to be successful by offering insightful information from the YouTube video ecosystem. Due to the very fascinating results that could be gained from nations with diverse languages, the researchers chose to focus on Germany, France, and Great Britain from that large dataset with 270k data points.

**Questions**

For this project, the following questions were established to perform the data analysis:

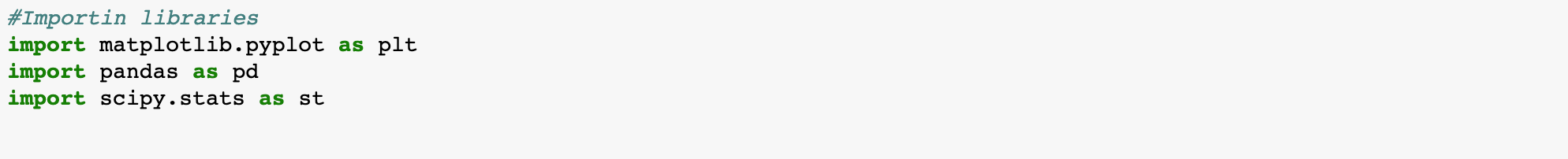
* What are the differences in the revenue according to each country?
* What are the factors that contribute to a video's popularity, such as the number of likes or dislikes, the publication date, or the tags used?
* What are the most common topics or categories of videos on YouTube?
* Are there any correlations between the number of views, likes, and dislikes for a video?
* What is the distribution of views across different video categories, such as music, comedy, or gaming?
* How does the language used in the video title and description affect the number of views or engagement metrics?

**Kaggle**

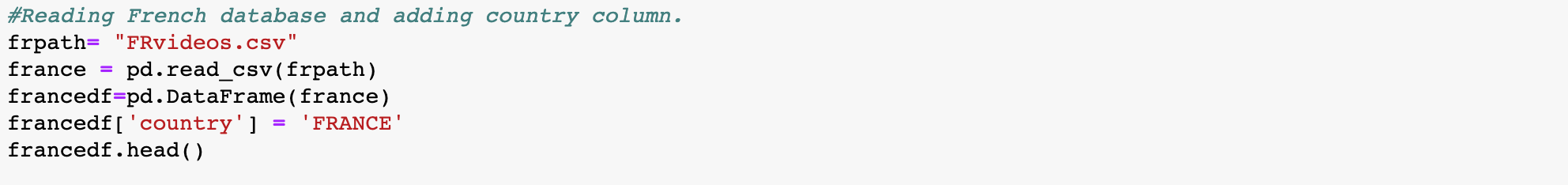
To begin with, the dataset was obtained from Kaggle1. Kaggle is an online community platform where you can find an incredible amount of datasets. It was founded in 2010 and acquired by Google in 2017. You can upload your datasets and download other users' (over 8 million) datasets. Kaggle has competitions where you can score points and win medals as proof of your progress, and starting earning points gives you the chance to enter a liver leaderboard. Shining in Kaggle gives you the opportunity to find new job opportunities.

**Clean up data**

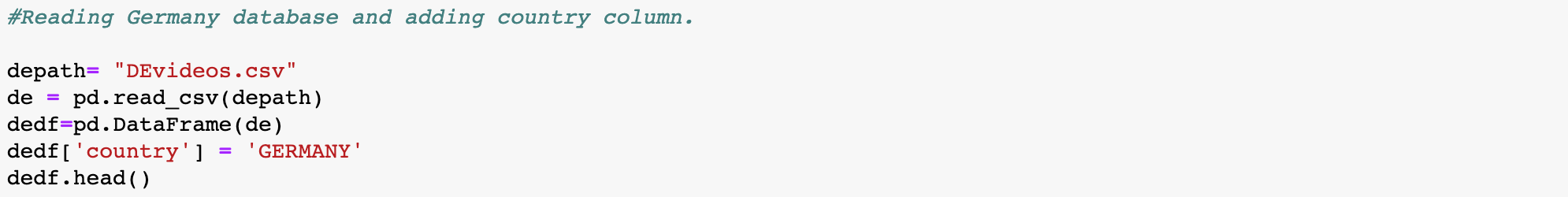
To clean up the dataset first we import the libraries necessaries for that task:



We read the dataset from France, Germany and Great Britain, put them in a dataframe and add a column to each dataframe with the name of the Country (that will help us later when we merge the dataframes).



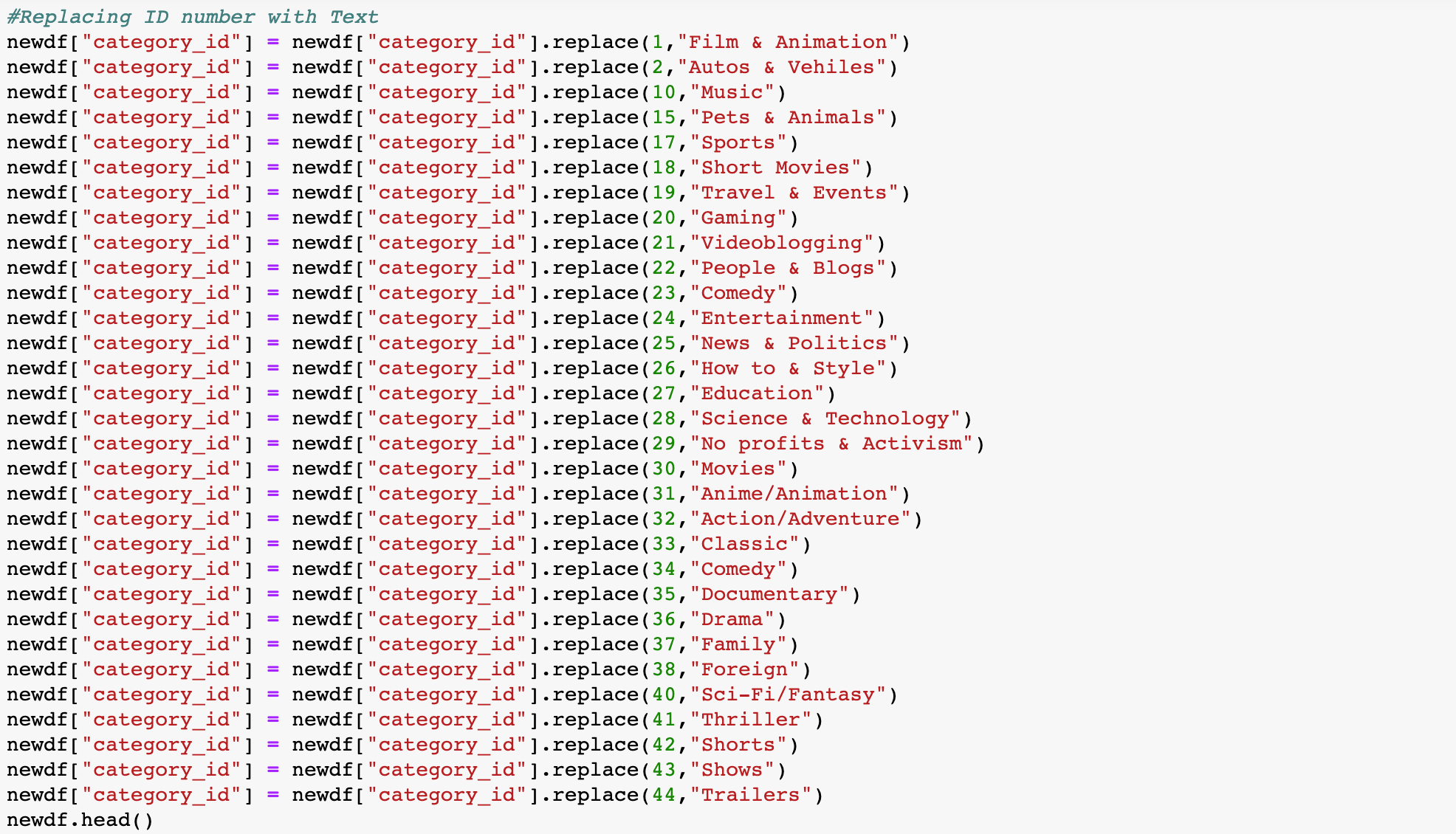




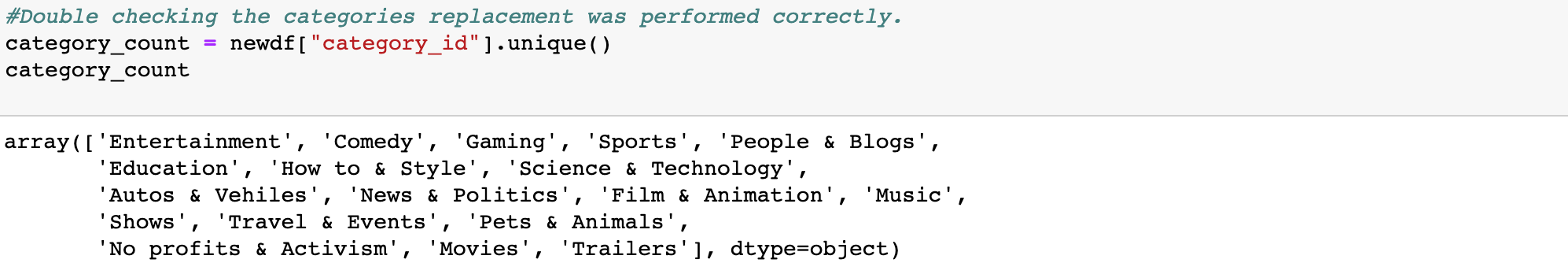
The next step is merging the German, French and British dataframes in a new one with the concat function and selecting the useful columns to have a clean dataset.



The team chose to replace the category ID number with the category name, that makes it easier to comprehend, analyze and graph.



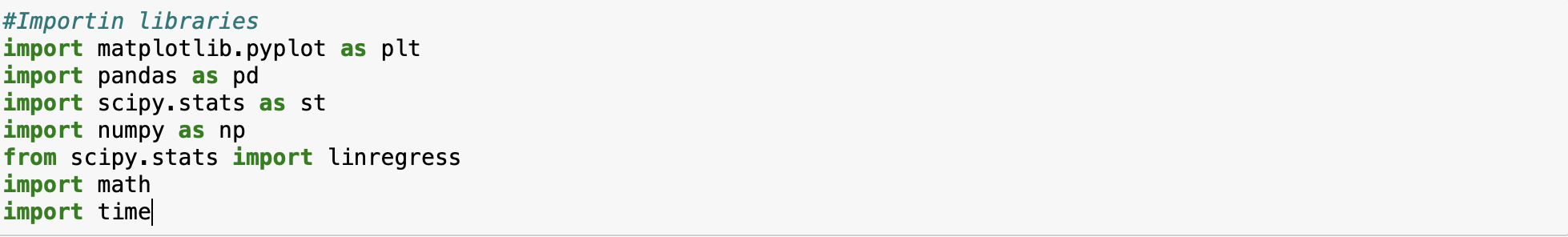
We verify that the categories was replaced correctly:



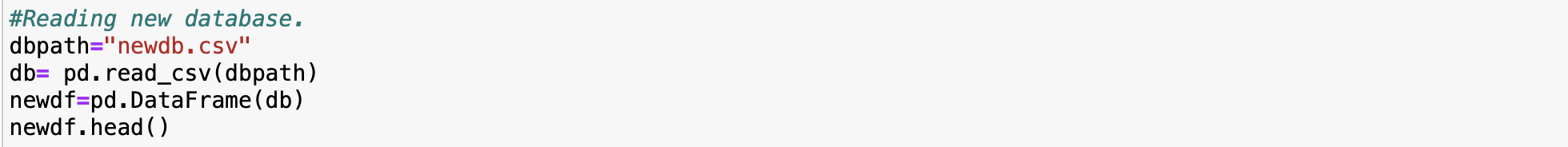
Finally export the dataset to a new CSV document:



After cleaning up the dataset, the libraries matplotlib.pyplot, pandas, scipy.stats, numpy, scipy.stats, linregress, and math were imported.



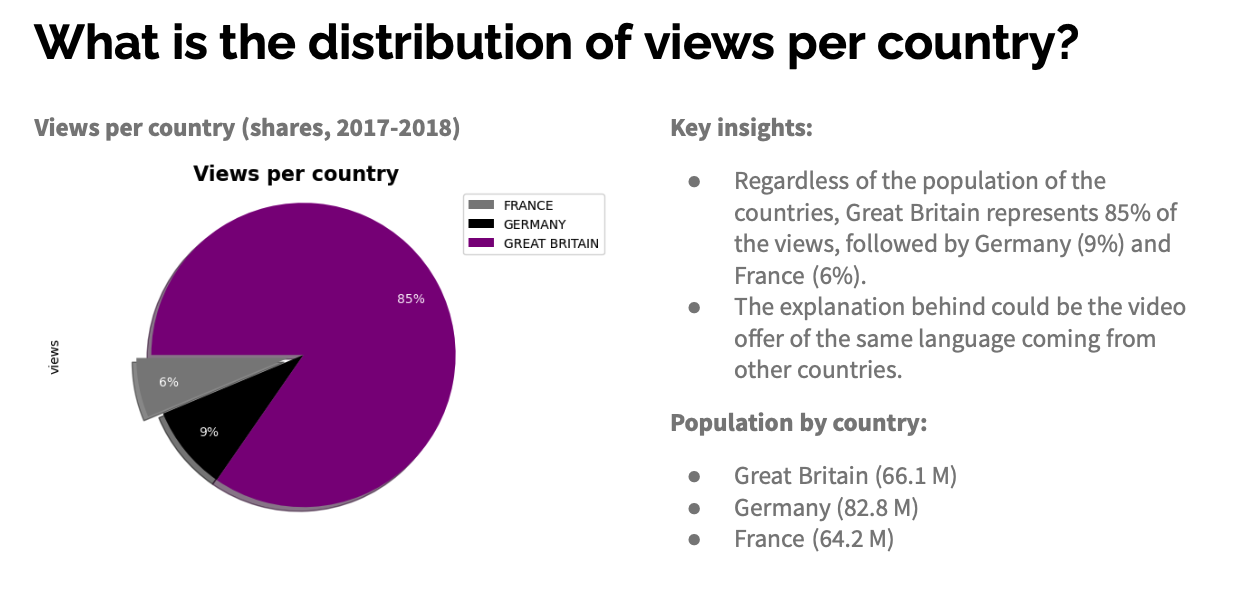
The team defined "newdb.csv", which is the CSV file, as dbpath. Then the code reads the CSV file. After that, we created a Pandas dataframe to create a new Dataframe that has the same data as db. Finally, the information was printed.

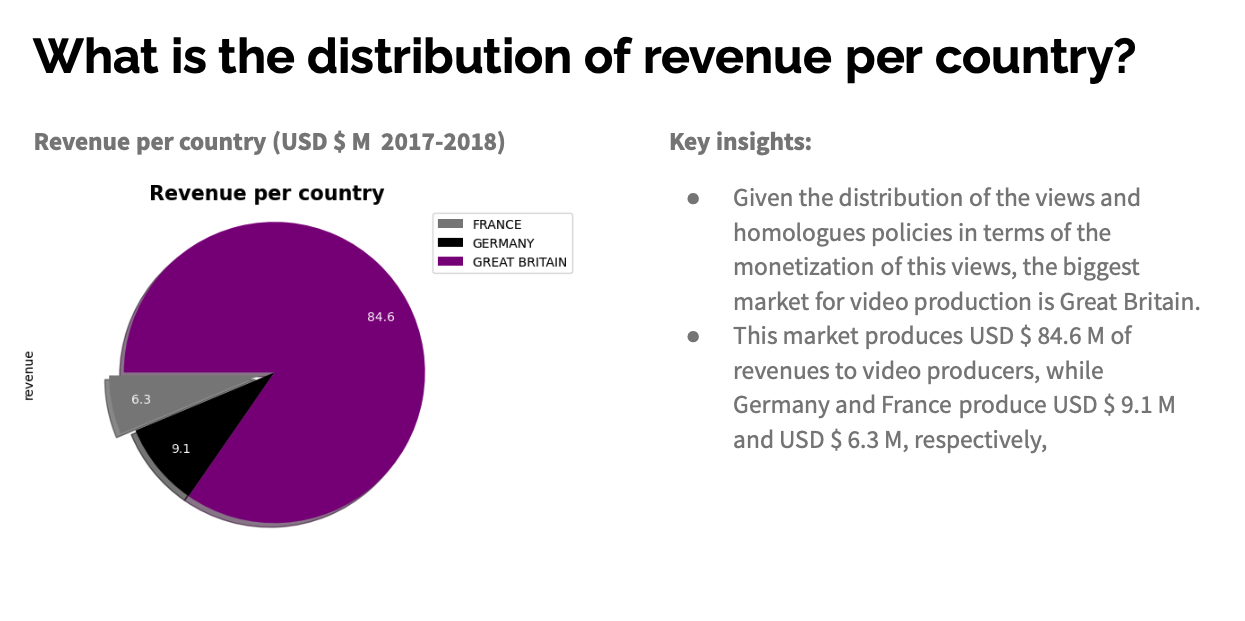


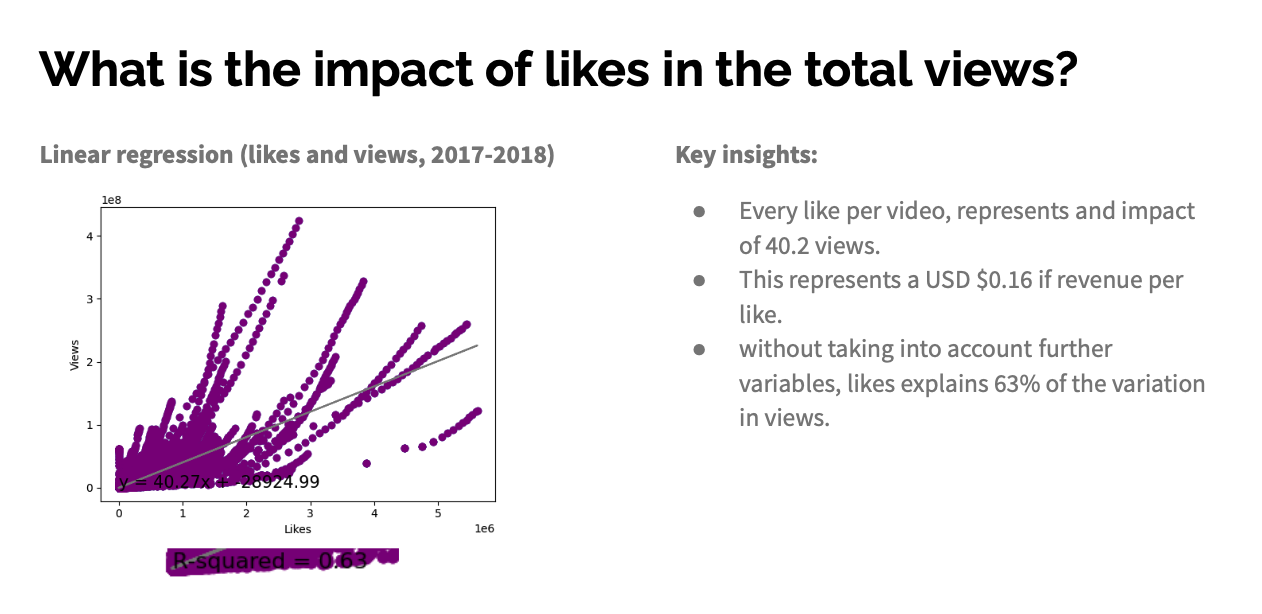
Afterwards, the team decided to create a new column to calculate the revenue, which was obtained by multiplying the number of views times 0.004 (which is the amount of money earned per video). Finally, the natural logarithm from the views and likes were obtained Beta from the regression.

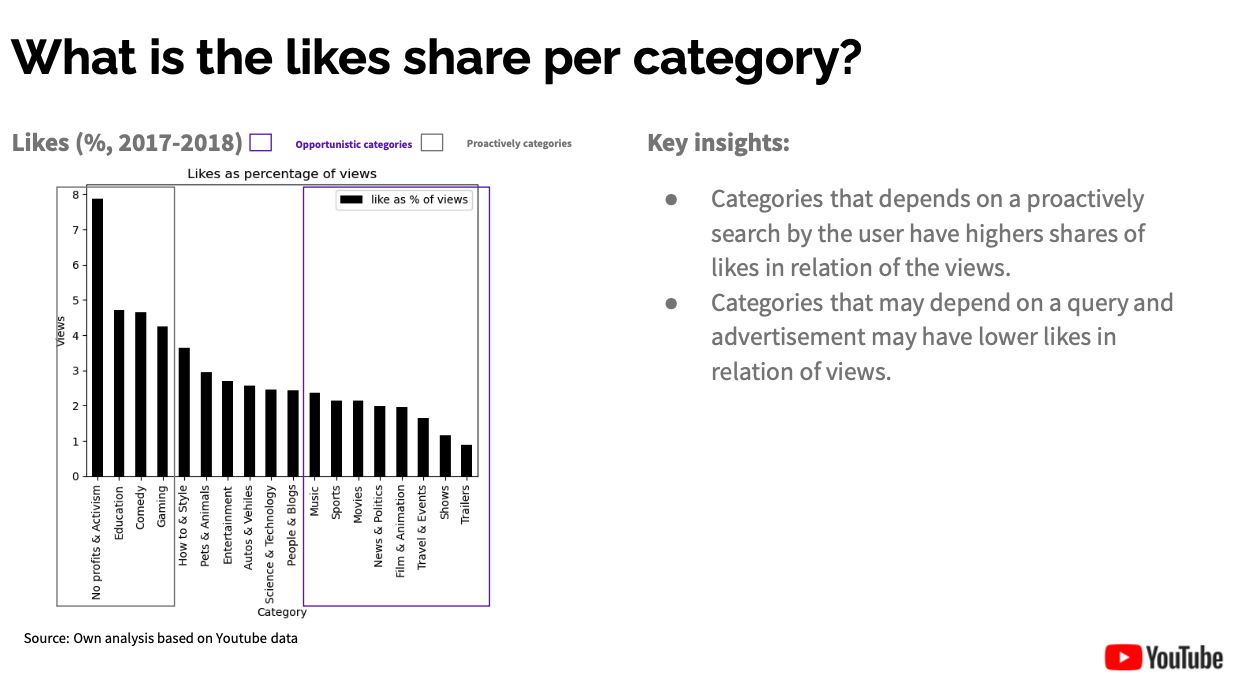
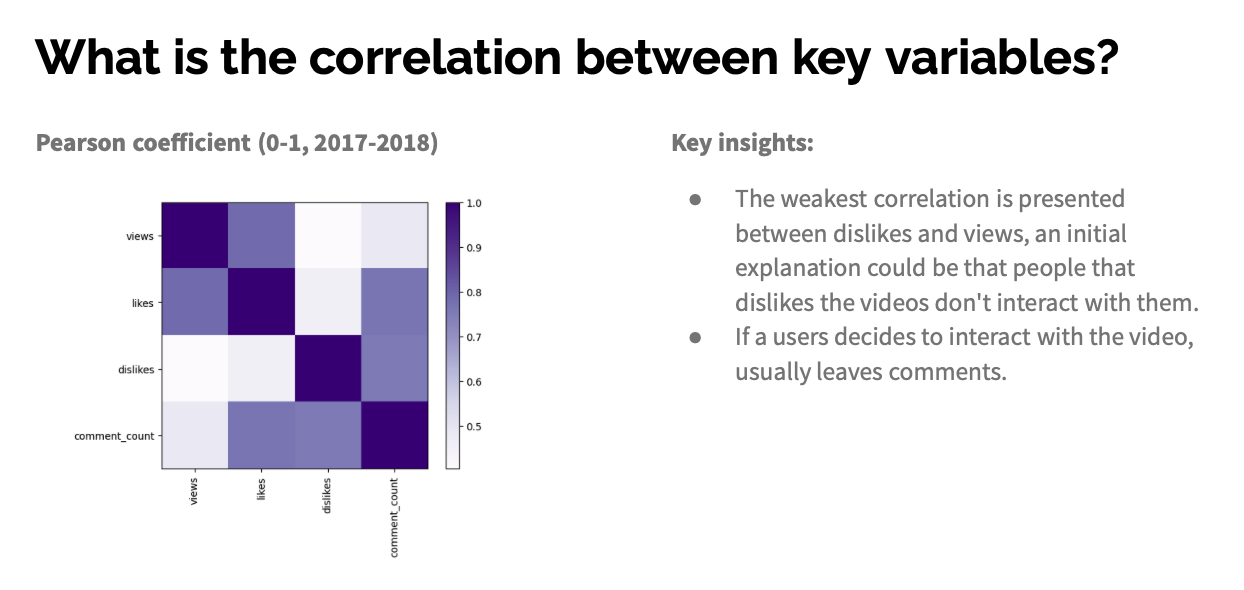
After doing this, the team started plotting different information to answer the questions mentioned before and the result and analysis obtained was the following:

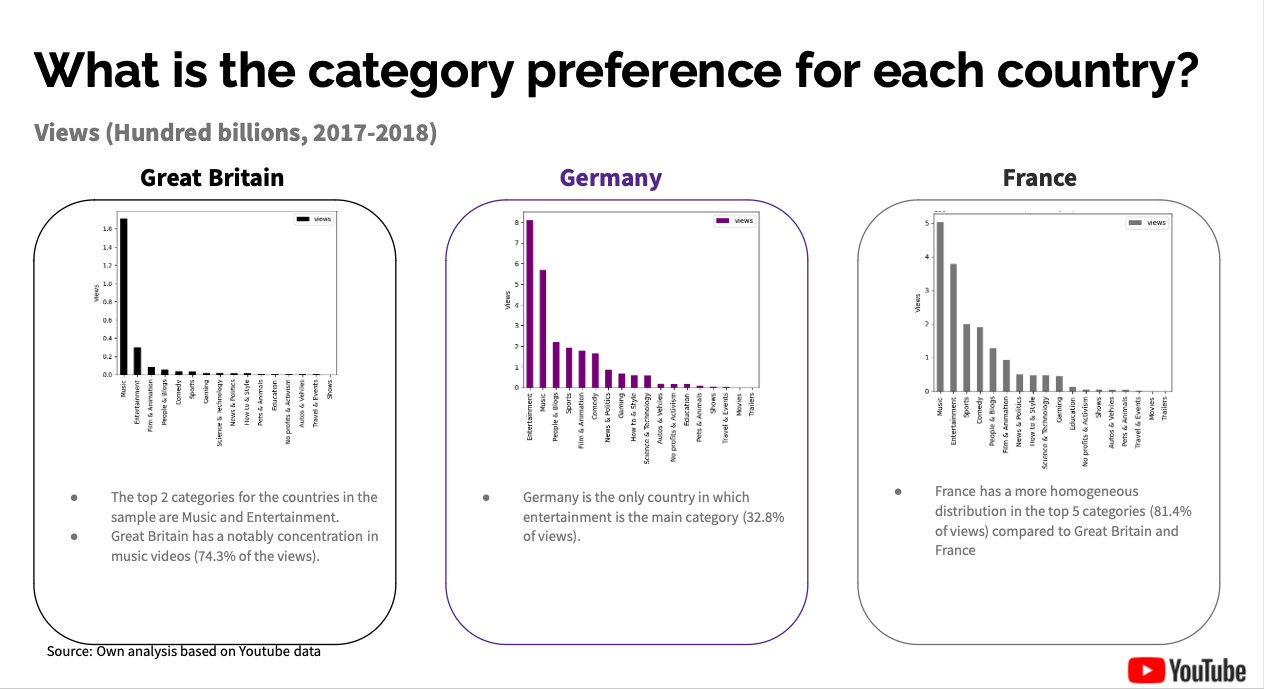
**Graphics and key insights:**







****

****

**Conclusion**

After analysing the information of more than 700k data points related to Trending YouTube Video Statistics1 the conclusions were the following:

Great Britain represents the biggest market for video production in terms of both views and revenues. It has been observed that 63% of the variation in views can be explained by the variation in likes. Specifically, one like is estimated to result in 40.2 views and USD $0.16 of revenue. When users interact with a video, they tend to leave comments. Categories that are initiated by users tend to have higher like-to-view ratios when compared to categories that rely on queries and ads.

In terms of video preferences, France and Great Britain tend to prefer music videos, while Germany watches more entertainment videos.

**References**

1. Trending YouTube Video Statistics. (2019, 3 junio). Kaggle. <https://www.kaggle.com/datasets/datasnaek/youtube-new?resource=download>
2. M. (2022, 24 agosto). How Much Do YouTubers Make? Facts and Figures for 2022. MintLife Blog. https://mint.intuit.com/blog/relationships/how-much-do-youtubers-make/