Hand over report

Project title

**YouTube trending video analysis**

**Big Data Engineering**

Written by

**Declan Stockdale**

Contents

[1. Project/Executive Summary 3](#_Toc97892894)

[2. Purpose 3](#_Toc97892895)

[2.1 Definitions 3](#_Toc97892896)

[2.2 Scope 3](#_Toc97892897)

[2.3 Constraints 4](#_Toc97892898)

[2.4 Assumptions 4](#_Toc97892899)

[2.5 Outstanding issues 4](#_Toc97892900)

[2.6 Risks 4](#_Toc97892901)

[3. Operational Handover 5](#_Toc97892902)

[3.1 Documents for handover 5](#_Toc97892903)

[3.2 Set up and Analysis 5](#_Toc97892904)

[3.3 Data Loading 6](#_Toc97892905)

[4. Data Cleaning 6](#_Toc97892906)

[5. Data Analysis 7](#_Toc97892907)

[5.1 Question 1. 7](#_Toc97892908)

[5.2 Question 2. 7](#_Toc97892909)

[5.3 Question 3. 8](#_Toc97892910)

[5.4 Question 4. 8](#_Toc97892911)

[5.5 Question 5. 9](#_Toc97892912)

[6. Business question 10](#_Toc97892913)

[6.1 Yearly analysis 10](#_Toc97892914)

[6.2 Entire dataset 12](#_Toc97892915)

[7. Final Recommendation 14](#_Toc97892916)

[8. References 15](#_Toc97892917)

[9. Appendix 16](#_Toc97892918)

1. Project/Executive Summary

The powers that be wish to create a new YouTube channel to drive interaction between viewers and our company. The YouTube channel will primarily be focused on generating a strong reliable viewership base which can then generate advertisement and paid partnership/sponsored based revenue streams.

To achieve the initial goal of generating views for our channel, the top trending videos over a period from August 2020 to December 2021 were analyzed across major international regions (Brazil, Canada, Germany, France, Great Britain, India, Japan, South Korea, Mexico, Russia, and the USA). This dataset is freely available on Kaggle [1]. It should be noted that the dataset is updated daily. The exact data records used will be explained under Constraints.

1. Project Overview
   1. Purpose

The goal of this project is to build alternative revenue streams by creation of a YouTube channel. The aim is to create engaging content which will then gather a viewership which can be leveraged for advertising or paid promotions.

* 1. Definitions

Azure Cloud – Microsoft based cloud computing service

Snowflake – Cloud computing-based warehouse company which is used to run SQL queries

SQL – Programming language used to analyze the datasets

UUID – Universally Unique Identifier, gives each row in database a unique value

* 1. Scope

The scope is limited to analyzing only trending videos.

* 1. Constraints

This project worked with outdated files

* 1. Assumptions

There are several assumptions in this project.

The first is that all seasonal trends within the dataset balance out as the regions are globally widespread.

The second assumption is that this dataset is indicative of the future e.g., no significant change in category popularity within the near term.

The third assumption concerns YouTube itself. Various content creators vocalized concerns during mid 2021 about how the YouTube algorithm was influencing the recommendation of their videos. It’s assumed that there was no change throughout the dataset although that may be disputed.

The last assumption concerns copyright infringement and that no trending videos were “claimed” as infringing on copyright either through valid claims or extraneously by bad faith actors. This doesn’t appear in the dataset and may have a major impact on various categories involving movies/television, gaming, or music. As the goal is to generate advertisement revenue, copyright must be considered into how videos are created.

* 1. Outstanding issues

As of writing this, all analysis has been performed in the Snowflake cloud. This has a free 30-day trial which may have elapsed for future maintainers of this project. The SQL query codes have been saved and are included however they may need to be rewritten for implementation in other SQL code versions.

* 1. Risks

No one can predict whether a video will trend or go “viral”, and this project may be a misuse of resources if the goal is to get trending videos rather than a consistent but potentially lower viewership.

1. Operational Handover
   1. Documents for handover

The most up to date versions of the dataset can be found online [1]. The datasets used for this project are already loaded into the Azure Cloud (more detail in Maintenance and support). They can also be found in the references [2,3].

16 SQL files have also been attached which include various data cleaning and exploratory analysis of the dataset (Appendix table 1).

If access to the relevant Azure Cloud storage is required, a request can be made to the relevant department (IT). Alternatively, the files can be loaded into a new azure cloud account following the process detailed in “Set up and Analysis”.

* 1. Set up and Analysis

The exact methodology of uploading the data into Azure and loading into Snowflake is detailed in these documents [4,5]. It assumes that the maintainer of this project has an Azure account and a Snowflake account.

A brief rundown of the process is explained as follows

1. The datasets should be downloaded online [1] for the most up to date dataset or from here [2,3] which contains trending videos and category data.
2. In Azure, create a new container with a name of your choice (storage\_name) and upload the dataset(s) – used in step 7
3. Click the settings icon and find the Azure directory Id – save for step 7
4. If not already done, create an account on Snowflake
5. Create a new database with a name of your choice (CREATE DATABASE <chosen\_name>;)
6. Use the database (USE DATABASE <chosen\_name>;)
7. Create a storage integration between Snowflake and Azure

(CREATE STORAGE INTEGRATION azure\_<chosen\_name>

TYPE = EXTERNAL\_STAGE

STORAGE\_PROVIDER = AZURE

ENABLED = TRUE

AZURE\_TENANT\_ID = (See step 3)

STORAGE = AZURE

STORAGE\_ALLOWED\_LOCATIONS = ('azure:// (step 2 container).blob.core.windows.net/(step 5 database name '))

1. Get the consent URL (DESC STORAGE INTEGRATION azure\_<chosen\_name>;) and paste into new window and click accept
2. In Azure storage account assign yourself a role under “Storage Blob Data Owner”
3. Upload dataset(s) into Azure container
4. Stage the uploaded files into required format (csv, json, etc.)
5. Create external tables

If using alternative cloud storage, please seek relevant documentation.

* 1. Data Loading

The downloaded files came in two formats, csv for video information which was missing category title and json for which had category id and category title. These were loaded into separate tables (table\_youtube\_trending and table\_youtube\_category) before being combined as a final table on category ID which was shared across both tables with 1,174,255 rows (table\_youtube\_final).

1. Data Cleaning

There were 1,223 videos which occurred twice. A UUID string was applied as a primary key and the duplicate video with a higher value was removed. This resulted in 1223 less rows.

The table\_youtube\_category contained category id’s not in the table\_youtube\_trending (horror, foreign etc.) including two numerical ids for the ‘Comedy’ category (23 and 34). Only 23 exists in table\_youtube\_category. As the tables are merged on common ID values, no extra steps have to be performed to remove these.

The category\_id 29 was present in both tables however only the category name of ‘Nonprofits & Activism’ was provided in the USA category data with other country data giving a null name. This was rectified by updating the table so that all null category name with category id of 29 were set to ‘Nonprofits & Activism’.

There was one video\_id (‘9b9MovPPewk’) with a missing channel title which belonged to ‘Juvis Productions’)

The way the trending videos were collected sometimes resulted in two records for a trending video within the same day and country. We are only interested in the video with the higher view count. A new duplicate table of the records was created by partitioning by Country, trending date and video id and keeping the vales with a lower view count. The lower view count duplicates were removed from the original table resulting in resulting in a final count of 1,123,846 records.

1. Data Analysis

Various analysis has been performed including answers to various questions proposed. Each question has an associated csv file attached to this report.

* 1. Question 1.

A snapshot of the top 3 viewed videos for each country in the ‘Sports’ section that trended on 2020-10-17 can be seen below (Fig 1), with the entire table available in “part\_3\_q1.csv”.

Table

Description automatically generated

Figure 1.Snapshot of top 3 videos for sports on 2020-10-17 for each country

The most common sports appear to be soccer, boxing, and Formula 1 within the results.

* 1. Question 2.

A snapshot of the number of unique videos containing the phrase (‘BTS’) can be seen below ranked by count (CT) (Fig 2). The videos were converted to uppercase for this search. The full results can be found in part\_3\_q2.csv

Clearly BTS (a Korean pop group) is most popular in Korea and has a strong following worldwide with a high number of trending videos across all regions.

Table

Description automatically generated

Figure 2. Snapshot of unique videos containing 'BTS'

* 1. Question 3.

Converting the trending date from YYYY-MM-DD to YYYY-MM we can look at trends over the months. As the table is large, the snapshot is available in the appendix (Appendix Figure 1). Full results in part\_3\_q3.csv.

From the results it appears the top videos are dominated by music videos across all regions largely from K-POP groups such as BTS, BLACKPINK, Black eyed peas etc.

* 1. Question 4.

Which category has the most videos for each country?

In the snapshot below (figure 3), Entertainment is by far the category with the highest number of videos for all countries except for Russia where it’s the ‘People & Blogs’ category. The full result is in the part\_3\_q4.csv. Entertainment videos make anywhere from 20% in the US, up to almost 40% in India for all trending videos (figure 4).

Table

Description automatically generated

Figure 3. Snapshot of category with most videos as a percentage

Chart

Description automatically generated

Figure 4. Entertainment videos percentage by country

* 1. Question 5.

Which channel title has produced the most trending videos?

Colors TV has created 809 distinct trending videos over the time of the dataset (figure 5). This number seems extremely high however checking their YouTube channel <https://www.youtube.com/colors/videos> there are numerous videos with less than 24 hours old with over 100,000 views.

Chart, histogram

Description automatically generated

Figure 5. Colors TV YouTube channel number of trending videos per month

1. Business question

This project was implemented to assess which category of video we should start investing in creating. A quote from the tubics.com website [6], “As of 2022, there are more than 51 million YouTube channels out there. The number of channels is growing strong: last year it grew by 36%. People all around the world are creating a YouTube channel and uploading 500 hours of video every minute”. Due to this we need to carefully analyze the datasets to identify trends and maximize out channel growth.

The requirements state that Music and Entertainment should be avoided. After investigating the ‘Nonprofits & Activism’ category which contains largely political videos, we can avoid that category as well.

* 1. Yearly analysis

The first consideration is which categories are seeing growth and which are in decline. We’ll look at the period from 2020-12-01 to 2021-12-01 which is the latest complete year in the dataset. Any seasonal trends should even out over the chosen period.

Over the year from 2020-12-01 to 2021-12-01 the following trends appear.

Gaming has seen a significant increase in the number of unique channels possibly due to the pandemic where more people can game and create content from home with minimal effort. Almost all other categories saw a decline (figure 6).

Chart

Description automatically generated

Figure 6. Change in number of unique channels for each category over the period from December 2020 to December 2021

The number of unique videos was again significantly higher in gaming due to reasons discussed above (figure 7). Travel and events saw an increase possibly due to increases in tourism videos promoting travel after lockdown. Education may have risen due to an increase in content being made available for online learning. All other categories except for People and blogs decreased over the period.

Chart, bar chart

Description automatically generated

Figure 7. Change in unique number of videos for each category over the period from December 2020 to December 2021

Interestingly the number of views and likes followed different trends with substantial increases in most categories, namely travel, education, pets, how to and film ana animation (figure 8). How much of this is due to lockdown which may drive views is unknown. We could compare to the year before, but any other trends may disappear.

Chart, bar chart

Description automatically generated

Figure 8. Change in number of likes and comments for each category over the period from December 2020 to December 2021

* 1. Entire dataset

Looking over the entire dataset, it’s clear that Science and Technology has the highest average view count per video followed closely by Gaming and Comedy (figure 9). These categories also feature in the top 4 of categories ranked by average comments per category (figure 10).

Chart, funnel chart

Description automatically generated

Figure 9. Average view count for each channel for each category

Chart, funnel chart

Description automatically generated

Figure 10. Number of comments for each category

Science and Technology also has a quite low number of unique channels compared to other categories (figure 11).

Chart

Description automatically generated

Figure 11. Number of unique channels for category in each country

1. Final Recommendation

This recommendations from this report are to start a new channel in the Science and Technology category. This is due to the high average view count and comments numbers per video along with the low number of unique channels may mean less competition and more immediate impact of videos reaching a trending status. From the year from 2020 -2021 the category saw relatively small changes in average viewership and unique channel numbers meaning the category is relatively stable and not likely to experience wild trends over the foreseeable future.

The categories by average view count can be seen in figure 12. Science and Technology makes a sizeable proportion of each country’s average viewership by category, and from figure 13 we can see that it’s ranked in the top 4 in every country by average views.

Chart

Description automatically generated

Figure 12. Category makeup by average views for each country

Graphical user interface, application

Description automatically generated

Figure 13. Ranking of Science and Technology for each country ranked by average views

1. References
2. Kaggle YouTube dataset

<https://www.kaggle.com/rsrishav/youtube-trending-video-dataset>

1. Trending videos <https://utsbde.blob.core.windows.net/assignment1/youtube-trending.zip>
2. Category data <https://utsbde.blob.core.windows.net/assignment1/youtube-category.zip>
3. Lab 2- Exercise 1 https://colab.research.google.com/drive/1642q21u9C1USS5\_UDLM1pKkj8iaU7BwV
4. Lab 2 – Exercise 2 <https://colab.research.google.com/drive/1653tmz3fXVprh09NVWIw7JUDhLfa_2gS>
5. Matthew Funk, 2022, *tubics GmbH,* “How Many YouTube Channels Are There?”, retrieved 9th March 2022 from https://www.tubics.com/blog/number-of-youtube-channels
6. Appendix

Appendix Table 1. Snapshot of result from question 3. Ranking of most viewed videos for each country for each month.

|  |  |
| --- | --- |
| SQL Filename | Brief Description |
| part\_1 | Loading all data into snowflake and creating tables |
| part\_2\_q1 | Find duplicate category title/id |
| part\_2\_q2 | Find missing category titles |
| part\_2\_q3 | Find missing category ids |
| part\_2\_q4 | Update table for missing category title |
| part\_2\_q5 | Find missing channel titles |
| part\_2\_q6 | Delete rows with #NAME? |
| part\_2\_q7 | Find duplicate values and create new table |
| part\_2\_q8 | Delete duplicates from original table |
| part\_2\_q9 | Count number of records |
| part\_3\_q1 | Most viewed videos by country |
| part\_3\_q2 | Count videos containing ‘BTS’ for each country |
| part\_3\_q3 | Most viewed and liked videos per country |
| part\_3\_q4 | Find category with most views for each country |
| part\_3\_q5 | Channel title with most views |
| part4 | Query to find best category to start a channel in |

Graphical user interface, application, table

Description automatically generated with medium confidence

Appendix Figure 1. Larger version of answer to query 3. Ranking of countries by highest videos for each Year-Month