Everyone uses Youtube across all ages. From kids looking up paw patrol, women searching for hair tutorials, and guys debating sports clips, youtube has become a common place of interest that attracts people from all over. With so many people coming to one place, that leaves a lot of data available that can open doors to new possibilities. That is why I’ve chosen to explore a dataset that contains over 2,000 youtube videos and comments to see if anything interesting can be found or atleast used to help others.

**ASK**

* What are the most commented-upon videos? Or the most liked?
* How many total views does each category have? How many likes?
* What are the most-liked comments?

Measurable questions can help us gain a numeric answer, and help establish the popularity of different ideas.

* What is the average sentiment score in each keyword category?
* How many times do company names (i.e., Apple or Samsung) appear in each keyword category?

Gaining insight to a customer's mindset can explain their interests and give them useful recommendations.

**Prepare**

We will be using a [youtube dataset](https://www.kaggle.com/datasets/advaypatil/youtube-statistics) that is available to the public with a no copyright license (CC0 1.0 universal public domain). It spans over 2 data tables one detailing data on the comments and the other consisting of stats. Some of the metrics are likes, views, comments, and keyword categories. Exploring this information can help in collecting data to support the direction a company wants to go whether it be in marketing, business, or anyone wanting to see the public's view.

**Process**

Excel, tablea, and SQL are the tools I wanted to practice with in exploring the dataset for I felt they could help give me the most insight. First I used excel to check for duplicates, and empty spaces which it was free of. The tricky part came down to the comments because it had spelling errors, but I chose to leave them be since people communicate differently. Learning their phrases, symbols, and code words is very important to understanding your audience. In order to translate the symbols and emojis I had to reformat the unicode back into excel because the first upload had them untranscribable. I also had to partition the datasets to count for quotations and symbols in the comments in order to import it into SQL google bigQuery. Appending the data table and quoting newlines allowed me to begin my analysis.

In SQL I wanted to answer the questions that I asked above in order to gain an understanding of my explorations. Using commands such as order by, group by, where, and left join helped me sift through different information to see what exactly was going on. Joining the tables using a column they both had allowed me to see them side by side at once and allowed for my code to be shorter and easier to read. Using where allowed me to identify specifics, and count allowed me to get a number that could identify the popularity of certain metrics. More information on my thought process can be found [**here**](https://docs.google.com/document/d/1QWX0NaisyoPjB5jhTBWTAVvtVy9_8IlFeIDcKw9uW_k/edit?usp=sharing).

**ANALYZE**

The main idea I wanted to understand was the popularity of the youtube videos because they told me what people were willing to spend their time on. Observing the views, likes, and comments was a direct response from the consumers and the data there holds answers to numerous possibilities. Knowing the most viewed showed what caught people's attention, the likes was an instant positive feedback, and the comments were more in depth about the thought process in viewing the video. The keywords are what people typed in the search bar that they identified the video with the most. Seeing this I found out that El Chombo - Dame Tu Cosita feat. Cutty Ranks (Official Video) [Ultra Music], ranked number 1 in views, comments, and likes. The most viewed keyword was google, the most liked/commented was mrbeast, and lofi had the highest sentiment.

I also noticed that the top liked comments had a keyword of mrbeast so that makes me believe that mrbeast has a heavy presence and strong following on the internet. He could be a beneficial partner for companies to work with. In my analysis I noticed the highest keyword count representing a video was 50, and some well known companies were in this range.

**VISUALIZE**

Link to dashboard made with Tableau

<https://public.tableau.com/views/YoutubeDashboard_16856611459230/Dashboard1?:language=en-US&:display_count=n&:origin=viz_share_link>

**ACT**

* What are the most commented-upon videos? Or the most liked?

The most commented videos where El Chombo - Dame Tu Cosita feat. Cutty Ranks (Official Video) [Ultra Music], I Opened A Restaurant That Pays You To Eat At It, $456,000 Squid Game In Real Life!, One Direction - History (Official Video), and history of the entire world, i guess with a combined total of almost 3,00,000 comments

The most liked videos where El Chombo - Dame Tu Cosita feat. Cutty Ranks (Official Video) [Ultra Music], $456,000 Squid Game In Real Life!, Martin Garrix - Animals (Official Video), Powfu - death bed (coffee for your head) (Official Video) ft. beabadoobee, The Weeknd - Save Your Tears (Official Music Video) with almost 50,000,000 likes.

* How many total views does each category have? How many likes?

The most keyword views came from google with over 46,000,000 views alone and Mr Beast had the most likes that surpassed 1 billion.

* What are the most-liked comments?

The most liked comments came from the Mr beast category and the number 1 was ‘Like I said in the video, subscribe if you haven’t already and you could win $10,000!’ coming from Mr beast on the video ‘$456,000 Squid Game In Real Life!’.

* What is the average sentiment score in each keyword category?

Lofi has the highest sentiment score with 1.832, machine learning 1.745, music 1.739, asmr 1.739, and tutorial 1.7088.

* How many times do company names (i.e., Apple or Samsung) appear in each keyword category?

For the companies category cnn had 50, marvel 50, nintendo 48, google 45, and apple 42.

This EDA helped me analyze this youtube dataset by summarizing its main characteristics, notice patterns, and visualized data in an easier format to observe. It helped reveal the popularity of mr beast, popular genres of music, and what consumers go to keywords for their searches are. Hopefully this analysis will help others with questions about youtube to easily solve their problems.