FIT5147 DATA EXPLORATION AND VISUALIZATION

DATA EXPLORATION REPORT

"Trending Youtube Videos Analysis"

Name: Gayatri Aniruddha
Student ID: 30945305

TABLE OF CONTENTS:

1. Introduction

YouTube is the most popular and most used video platform in the world today. YouTube has a list of **trending videos** that is updated constantly.

The project that I have chosen revolves around - "**Trending Youtube videos".** This contains data about **more than `40,000` trending videos**. I have analyzed this data to get insights into YouTube's trending videos, to see what is common between these videos. This dataset was collected over 200 days and contains data about the trending videos pertaining to that day.
The dataset that I have used was obtained from Kaggle. It contains data about trending videos for many countries including USA, Canada, France and India. Here, I have **Python, Tableau and R** to analyze the dataset of the USA. Those insights can also be used by people who want to increase the popularity of their videos on YouTube.

I have done in **depth analysis, visualisation and exploration** to answer three main questions which are mentioned below.

1) What is the relationship between the various trending videos and their publishing times?
2) How are the various trending videos related to their captions and video lengths?
3) Which category and channels have the largest number of trending videos? Also, what are the common words found in their title?

2. Data Wrangling

Description of Data Sources used:

I have used multiple and large datasets. Datasets of countries include -

1. India
2. Canada
3. USA
4. France
5. Great Britain

The data available was in a tabular format with 16 Columns x 40k Rows. The various columns were titles as follows - video title, channel title, publish time, tags, views, likes, dislikes, description and comment count. This dataset has both text data and numeric data.

Links of the data sources used:

https://www.kaggle.com/datasnaek/youtube
https://www.kaggle.com/datasnaek/youtube-new

Tools Used:

1. Python
2. Excel
3. R Studio
4. Tableau Public

Steps used in Data Wrangling:- Cleaning and Transformations

My dataset had around 16 columns and more than 40,000 entries. It was a **sufficiently large dataset** and it was difficult to detect and clean it using excel alone. As a result, significant cleaning and wrangling was required. Hence, I did my Data Wrangling using Python in Jupyter Notebook.

Transformation of the trending_date into the "date" format.
In my given dataset, the various trending dates of the videos were in this format - "yy.dd.mm".
I have **cleaned** this data and converted the dates into this format - "yyyy-mm-dd".
Example - 14th November, 2017 was mentioned as - "17.14.11" and now is mentioned as "2017-11-14".

```
df['trending_date'] = pd.to_datetime(df['trending_date'], format='%y.%d.%m').dt.date
```

Cleaning out the publish_time into publish_date and publish_time_only.
Here, the publish time consists of the date and time of the video release. Here, while **cleaning** the data, I am slicing out the date and time from the publish_time. This will help us in analysing, visualising and solving my first question.

```
df["publishing_day"] = df["publish_time"].apply(
    lambda x: datetime.datetime.strptime(x[:10], "%Y-%m-%d").date().strftime('%a'))

df["publishing_hour"] = df["publish_time"].apply(lambda x: x[11:13])
```

Creation of a new Column - days_to_trending.
In order to get a better idea and picture of the data, I have created a new variable i.e column named days_to_trending which tells us the actual number of days it took for the video to become a trend after it was released.

```
# Create New Variable Counting Days to Achieving Trending Status
df['days_to_trending'] = (df.trending_date - df.publish_date).dt.days
```

Creating a meaningful index for the dataset - trending_date, video_id
Here, while **cleaning**, we have arranged the videos according to the dates. Thus, for a given date, we have listed out all the videos which were trending on that day.

3. Data Checking

Here, after sufficiently cleaning the dataset, I further performed some basic checks to see - blank and duplicate values. All the data checking was done using **Python** in Jupyter Notebook.

Error 1 : NaN Values
Checking NaN Values - Replacing the NaN values with blank spaces.
After loading the data, I observed that there were 40,949 entries in the dataset. I also observed that all columns in the dataset were complete (i.e. they have 40,949 non-null entries) except the description column which had some **null values** which were denoted by NaN. So to do some sort of data cleaning, and to get rid of those null values, we put an empty string in place of each null value in the description column.

Error 2 : Un-reliable values
Checking un-reliable observations - and removing them.
Here, we have a column named - **video_error_or_removed.** Thus, when this value is True, it means that the video was removed or has some error. Hence, we need to **clean** this up and remove the rows where this value is true!

```
df = df[~df.video_error_or_removed]
```

Error 3 : Duplicate values
Checking duplicate values and their removal.
Along with the above cleaning and transformations, I further removed **duplicate video_id's** and sorted the data according to the number of views. After removing the duplicate ids, there were comparatively lesser rows to analyse and visualise.

Error 4 : Insufficient data
Checking insufficient data and their removal.
Here, we have a column named ratings_disabled. Thus, when this value is True, it means that we don't really know the number of likes and dislikes this video has received. Hence, we need to **clean** this up and remove the rows where this value is True as this data is of no use to us while analysing the trends of our youtube videos. This further reduced around 150 records which were of no use for us in our exploration.

4. Data Exploration

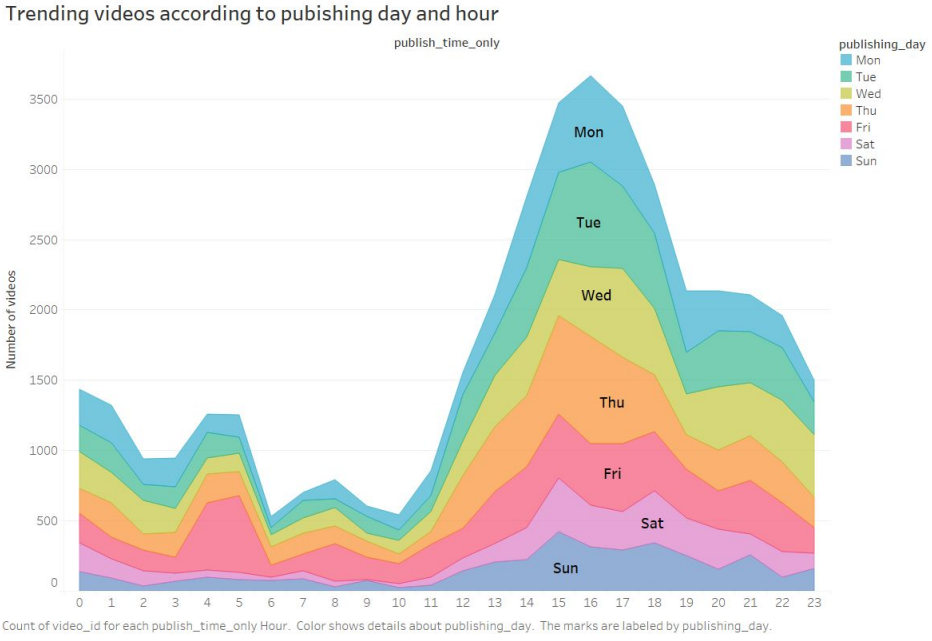
Questions:

1) What is the relationship between the **various trending videos** and their **publishing times?**

Answer:

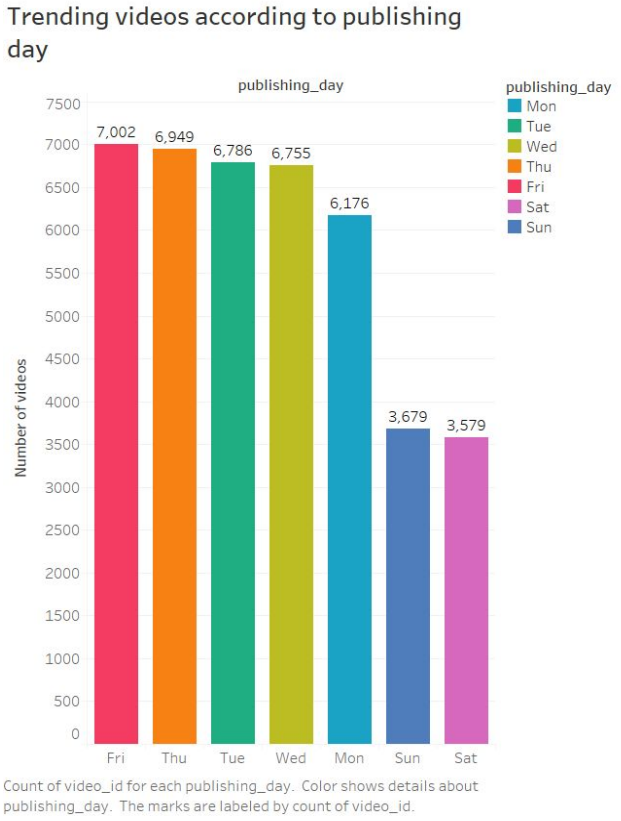
During the process of data cleaning, I saw that the publishing time was a mix of date and time. I separated it out in **two columns** - which specified the **publishing day** and the **publishing hour** for easier visualisation and analysis of our dataset. I have plotted the sum of various trending videos against their publishing days of the weeks and publishing hours.

Relationship between videos with respect to publishing days and hours:
Figure 1:



Relationship between trending videos and publishing days :

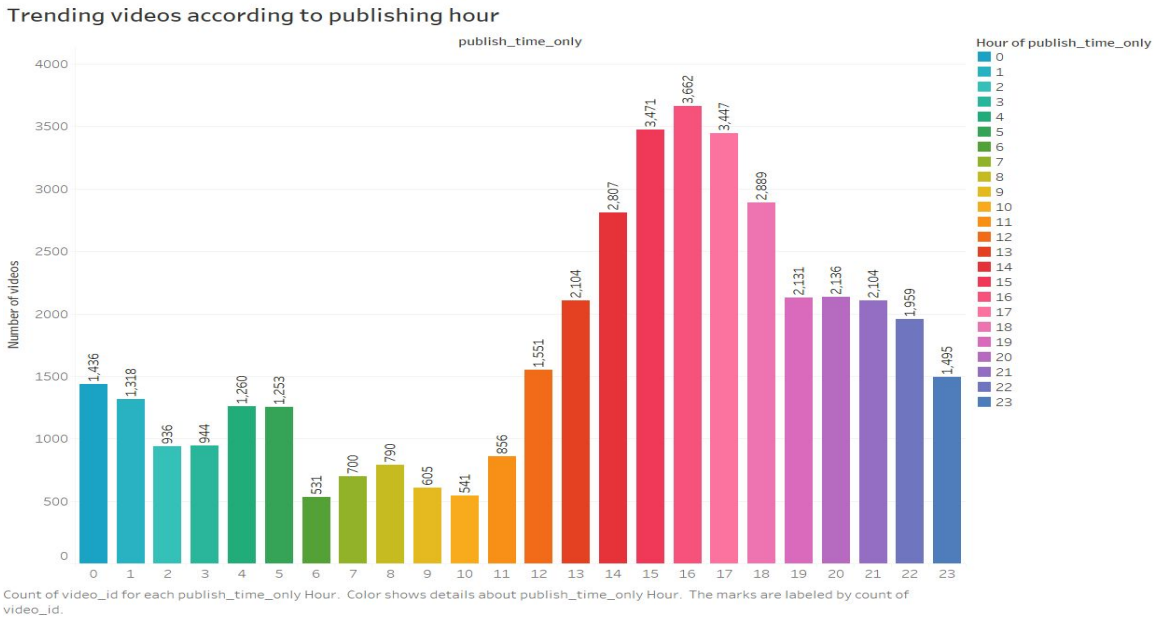
Figure 2:



- We can observe the **trend** that the number of trending videos published on Sunday and Saturday are noticeably less than the number of trending videos published on other days of the week.
- Clearly, a video which is released on a Friday has the highest chances of becoming viral and come in the trending category.
- This **pattern** is because - Friday is the beginning of the weekend. Most people are in their best spirits and are relaxed when compared to the weekdays. They are more active on social media and social sites. Hence, more people are likely to watch videos and this increases the video's chance of gaining more public attention and hence trend!
- However, on sundays, people become lazy again and they have the entire week waiting ahead of them! People are busy preparing for the activities of the week. Hence, the amount of time spent by people on their phones and tabs is less. As a result, videos released on the weekends are not able to reach a larger audience.

Relationship between trending videos and publishing hour :

Figure 3:



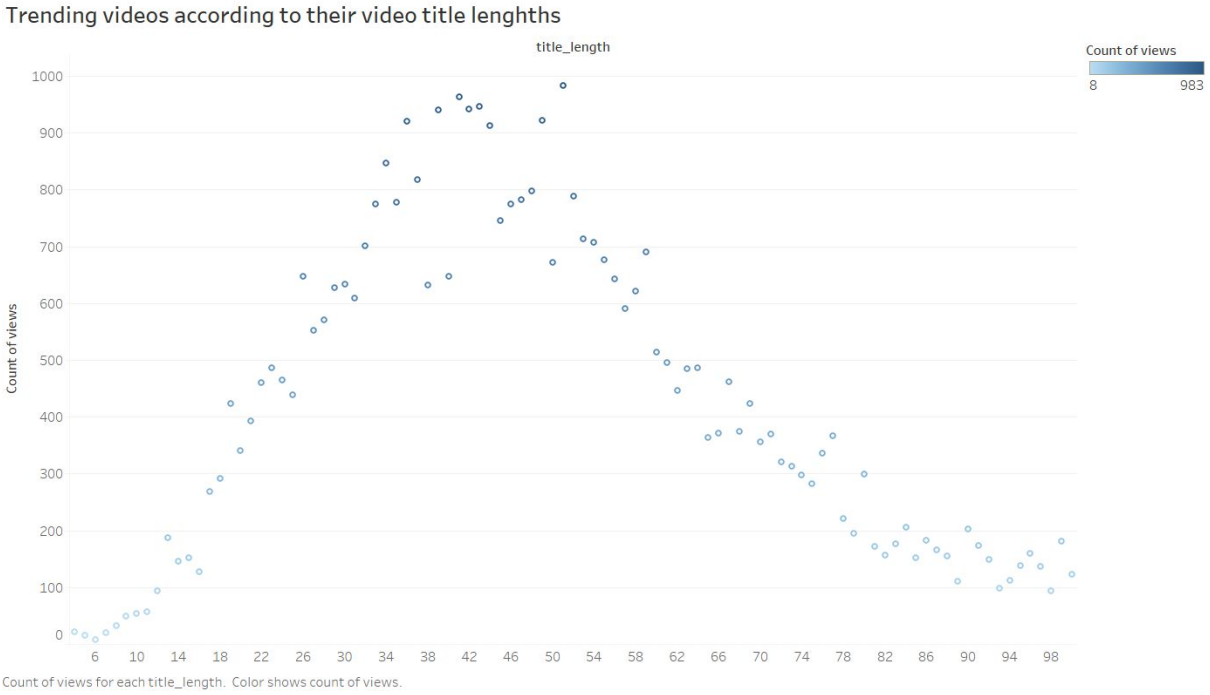
- From the visualisation, it is clear that the period between 2PM and 7PM, peaking between 4PM and 5PM, has the largest number of trending videos. We notice also that the period between 12AM and 1PM has the smallest number of trending videos.
- This **trend i**s because people publish a lot more videos between 2PM and 7PM. Also, this is the time when people are wrapping up their chores for the day and are comparatively relaxed. Especially, 4 PM to 5 PM is the time when people are travelling back home and are most likely to use their phones after a tiring day at school, college or work.
- These values drop after 5 PM as people get back home and are immersed in home chores, prepping up for the next day, busy doing assignments or spending time with their family. The same **pattern** observed in the morning hours as well. People are fresh, ready to start their day and don't want any disturbance. Hence, they are less likely to spend their prime time of the day on youtube.

2) How are the **various trending videos** related to their **captions and video title lengths?**

Answer:

- We can observe the **trend** that the title-length distribution resembles that of a normal distribution, where videos with the most number of views have title lengths between 30 and 60 characters approximately.
- This **pattern** is observed because people don't have the time to read really lengthy titles. Similarly, people will not feel like watching videos with really short titles which gives almost zilch information about the video content.
- Thus, titles which have around 30 to 50 words and give sufficient information about the underlying video, attracts viewers!

Figure 4:



3) Which **category and channels** have the largest number of trending videos? Also, what are the **common words** found in their title?

Answer:

Relationship between Views and Categories:

Figure 5:

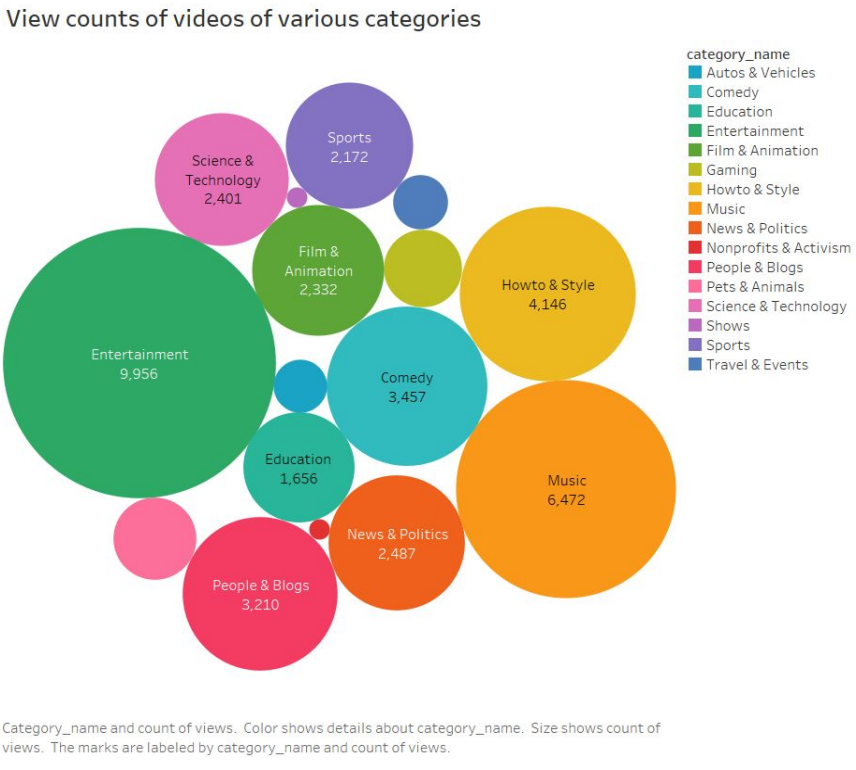
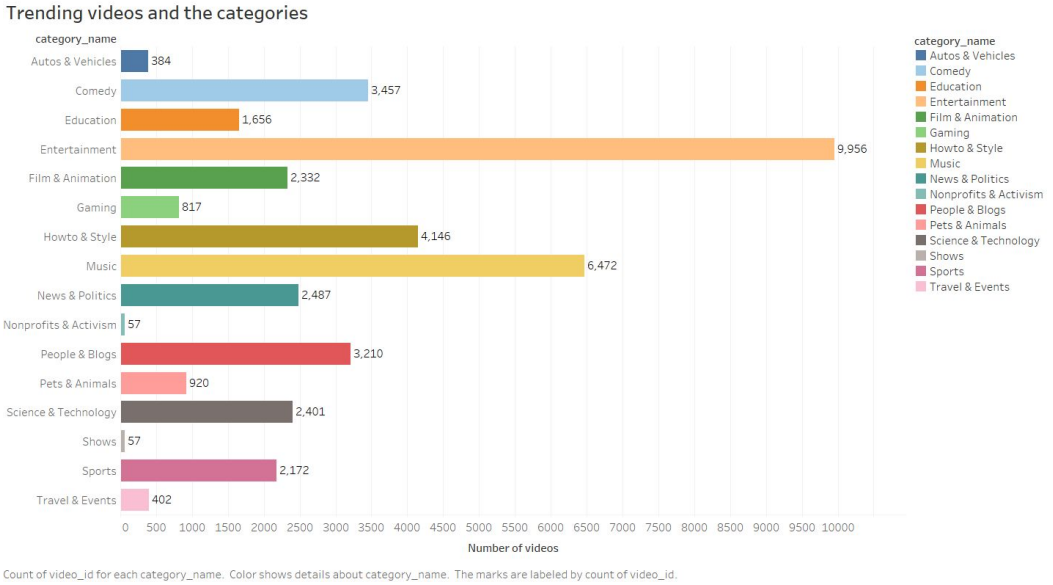


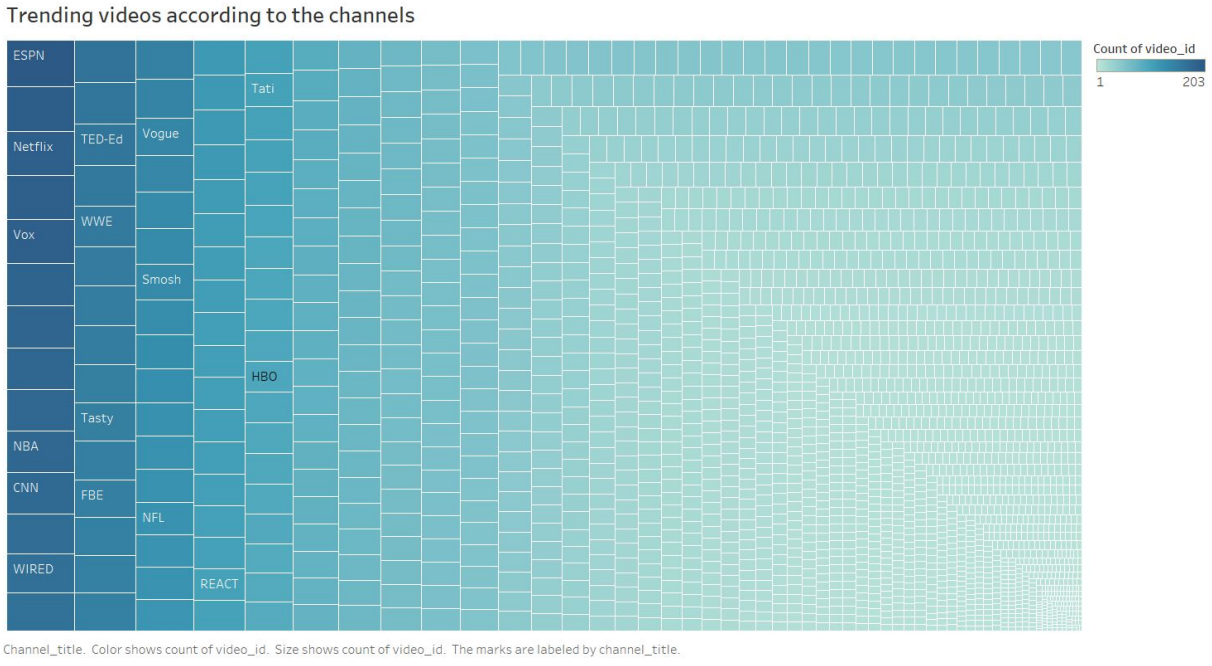
Figure 6:



- Clearly, from the two plots, we observe the **trend** that the **Entertainment category** contains the largest number of trending videos, around $10,000$ videos, followed by Music category with around $6,200$ videos. Similarly, Nonprofit and Activism related videos have the least trends - around 57 of these.
- People want to watch something light hearted, something to divert their mind and relax them. People prefer gossip over matter. People turn to youtube while eating, or while taking a break from their routine. In these periods of time, they are not in a mood to soak in more information. All that they want to do is have a laugh. As a result, **videos** related to **Entertainment and Music,** clearly have outnumbered and outshines the other categories.

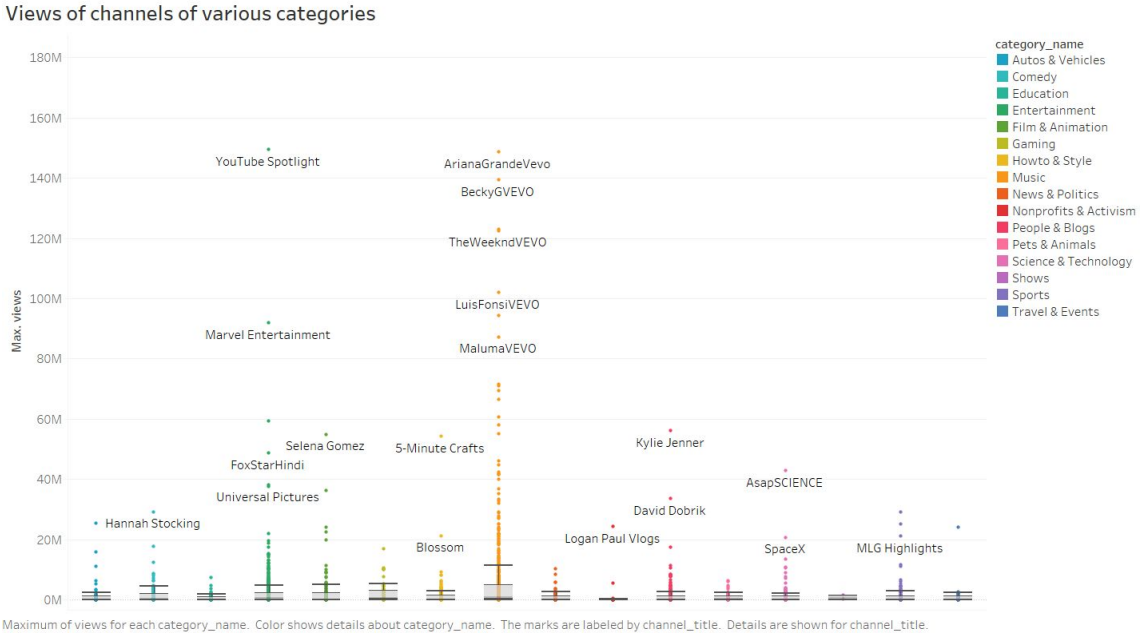
Relationship between trending videos and channels:

Figure 7:



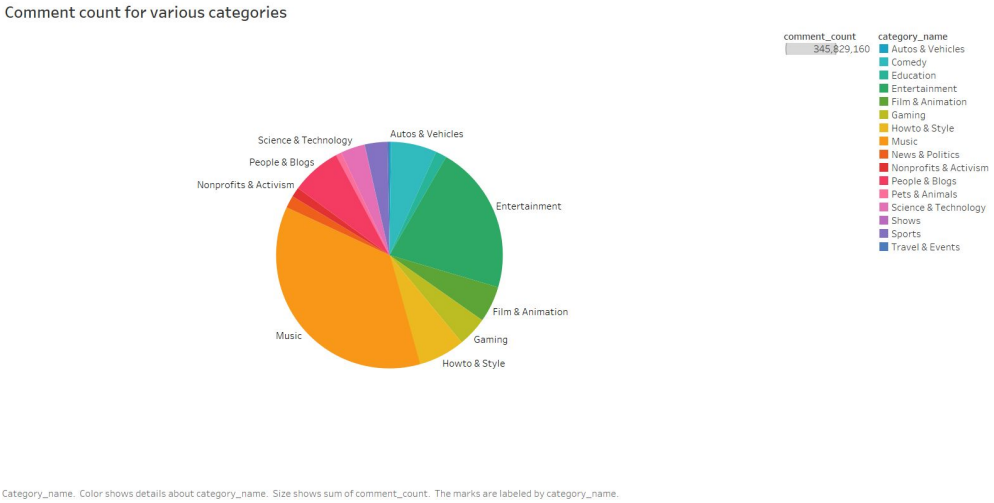
- Furthermore, this **pattern** can be from the above visualisation as well. We can see that **Entertainment** channels like ESPN, Netflix and CNN have the highest number of viewership.

Figure 8:

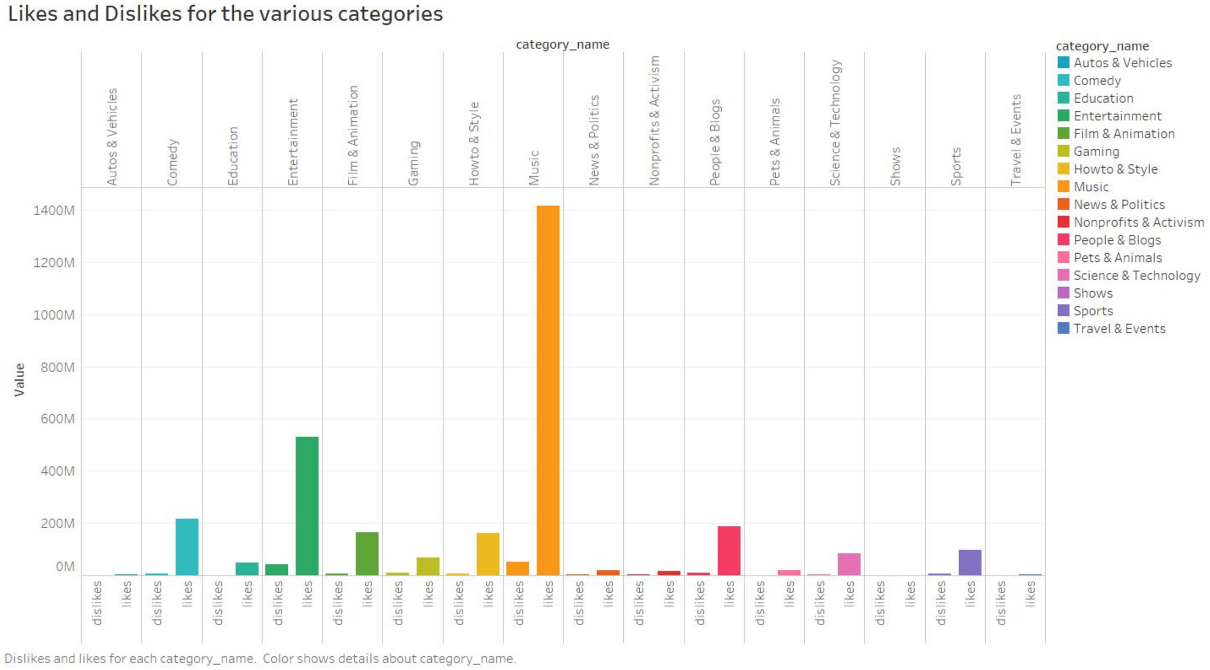


- This plots the views for various categories. Further, the outliers or highest viewed channels are mentioned on the plot and the various categories are represented with the various colours.
- This is because, most people now use social media and youtube after a tiring day at work and prefer something light. People want to listen to music to relax their minds or watch something to take their mind off work. As a result, these categories of **Music** and **Entertainment** have the highest number of viewership.

Further visualisations to show the popularity of the various categories:
Comments for the various categories:
Figure 9:



Likes and Dislikes according to the categories:
Figure 10:



Word Cloud of video titles:
Figure 11:

- Further, I analysed the most common words found in the **trending video captions** by generating a word cloud. It was observed that most of the **trending videos** had words like -"Official", "Trailer", "Audio", "Video" etc in their captions. These captions were found in more than **2000** videos.
- This **pattern** is observed because people are drawn towards latest music videos, show trailers and catchy captions!

5. Conclusion

This **process** of data exploration of youtube's trending videos was pretty **successful.** I was able to **correctly visualise my dataset** using tableau and R, point out faults and further clean it using python to remove duplicate values, unwanted and unreliable data and NaN values. As a result of data cleaning, I was able to successfully concise the huge dataset. Furthermore, I made some **transformations** in the dataset and changed the format of some columns, creating new columns for better visualisation of my data. Hence, my data exploration process was not only successful in answering my initial questions, but also answered some more extra questions that came while answering the ones that I started with. I was able to draw **relationships** between the various trending videos with their publishing day and times. Further, I was able to identify the link between the caption lengths and views. Finally, I was able to observe **patterns and trends** in the dataset which gave me the relationship between the trending videos and their channel categories.

6. Reflection

This project gave me a **thorough understanding** of data wrangling, data checking, data cleaning, formatting and visualising. Firstly, I realised that the final visualisation becomes very straightforward if our dataset is formatted correctly. However, finding the right data sources and data wrangling takes around 70% to 80% of the time. The questions helped me to further analyse and explore my data based on my requirements. I also learnt how to **extract relevant data** from the dataset for plotting the graphs. Finally, this project just helped me improve my python, tableau and R knowledge and gave me an insight into how it would be like to work as data scientists in real-time projects.

7. Bibliography

Mitchell, J(2017, October 26). Trending YouTube Video Statistics and Comments. Retrieved from https://www.kaggle.com/datasnaek/youtube

Mitchell, J(2019, June 3). Trending YouTube Video Statistics. Retrieved from https://www.kaggle.com/datasnaek/youtube-new