**Business Understanding**

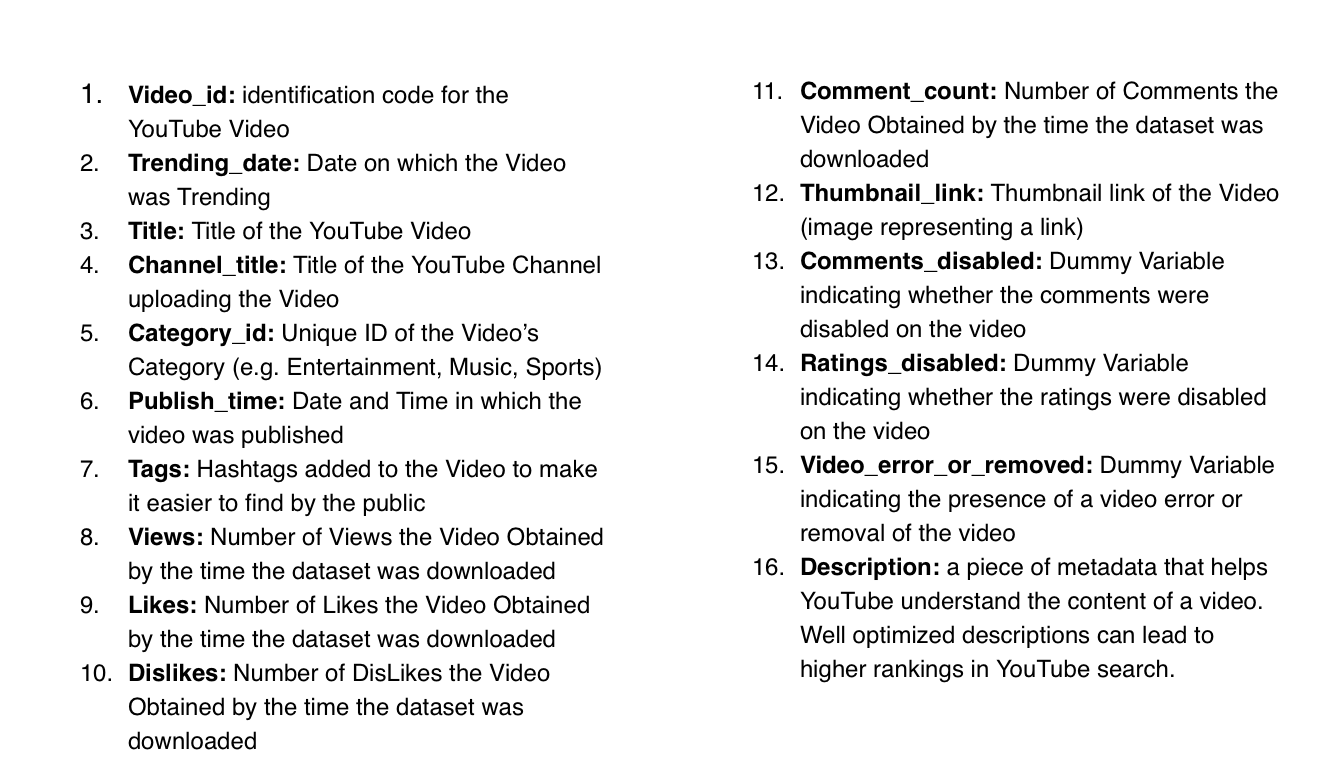
By the end of 2019, YouTube has reached 1.9 billion users worldwide, and 79% of all Internet users have their own YouTube account. Everyday, over 720 thousand hours of video are uploaded to the platform… It is not a surprise that YouTube is one if not the most popular websites on the internet, and these metrics aforementioned confirm this fact. Singers use YouTube to drop their latest music videos; Businesses use the platform to promote their new upcoming products; and with the increasing popularity of vlogs, more and more people are uploading videos of any content representing their daily lives. One thing is certain in the world of YouTube: the more views, likes or followers you get, the more popular your channel will be and the higher the opportunity for you to become Trending in the YouTube community.

This last point was our motivation for our winter competition at Duke Univerisity’s Fuqua School of Business. Our goal was to analyze what were the main characteristics of the most popular videos on YouTube, and build a model which could predict the number of views a new uploaded video could get.

**Data Understanding**

For this project, we decided to use the Extracted YouTube Trending Videos dataset from Kaggle (https://www.kaggle.com/datasnaek/youtube-new). In order to make our analysis more accurate and relevant, we only focused on US videos, reducing the dataset’s size to 40,949 rows and 16 columns.

Following is a detailed description of the 16 columns:



As we can see the dataset contains both numerical and categorical data of different YouTube Video categories including entertainment, sports, news, politics, education etc. Numerical data are how many views, likes and comments each video has received; Dummy variables are present to indicate whether a video has comments and rating enabled or not; additional Text data representing titles, tags and descriptions.

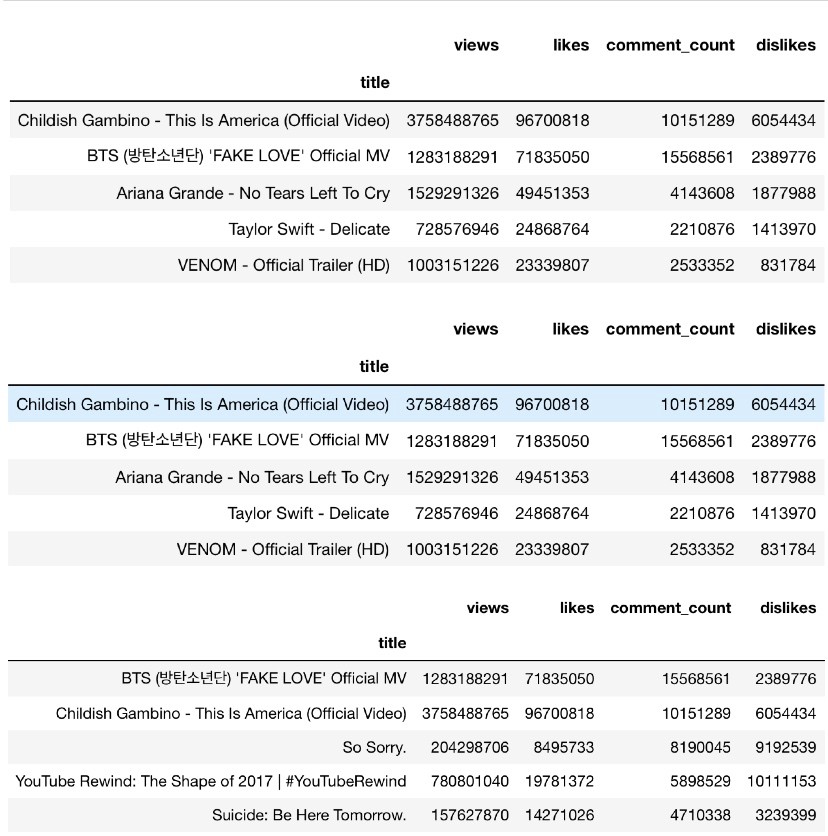
As previously mentioned our goal is to first implement an exploration of the dataset, understanding the main features of the popular videos. In addition, we will implement modeling algorithms in order to predict the number of views a certain video will obtain.

**Data Cleaning**

1. We started by getting rid of the unuseful columns in the dataset: ‘thumbnail\_link’
2. We modified ‘category ID’ by mapping the ID number we have in the dataset to the actual category name (e.g. category\_id 23 == ‘Entertainment Category’)
3. We normalized ‘views’, ‘likes’ and ‘comment count’, by taking their log values. This was used to reduce the wide ranges of these variables.
4. We converted True/False columns (e.g. ‘comments\_disabled’) into binary variables
5. We converted ‘trending\_date’ & ‘publish\_time’ columns to obtain date time formats
6. We implemented some feature engineering by creating new variables such as: ‘title\_length’, representing the count of words in the each videos’ title; ‘elapsed\_time\_to\_trending\_days’, representing the number of days it took for a video to become trending; ‘hour\_of\_publication’, representing the hour in which a video became trending

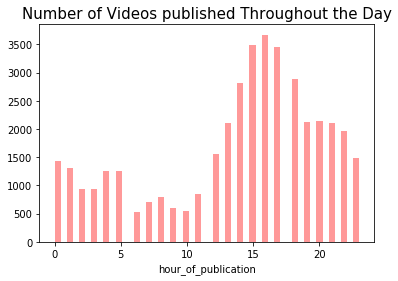
**Data Exploration**

In order to grasp a general understanding of trendy videos, we first selected top 5 videos based on views, comments, and likes.



**When are Most Videos Published?**

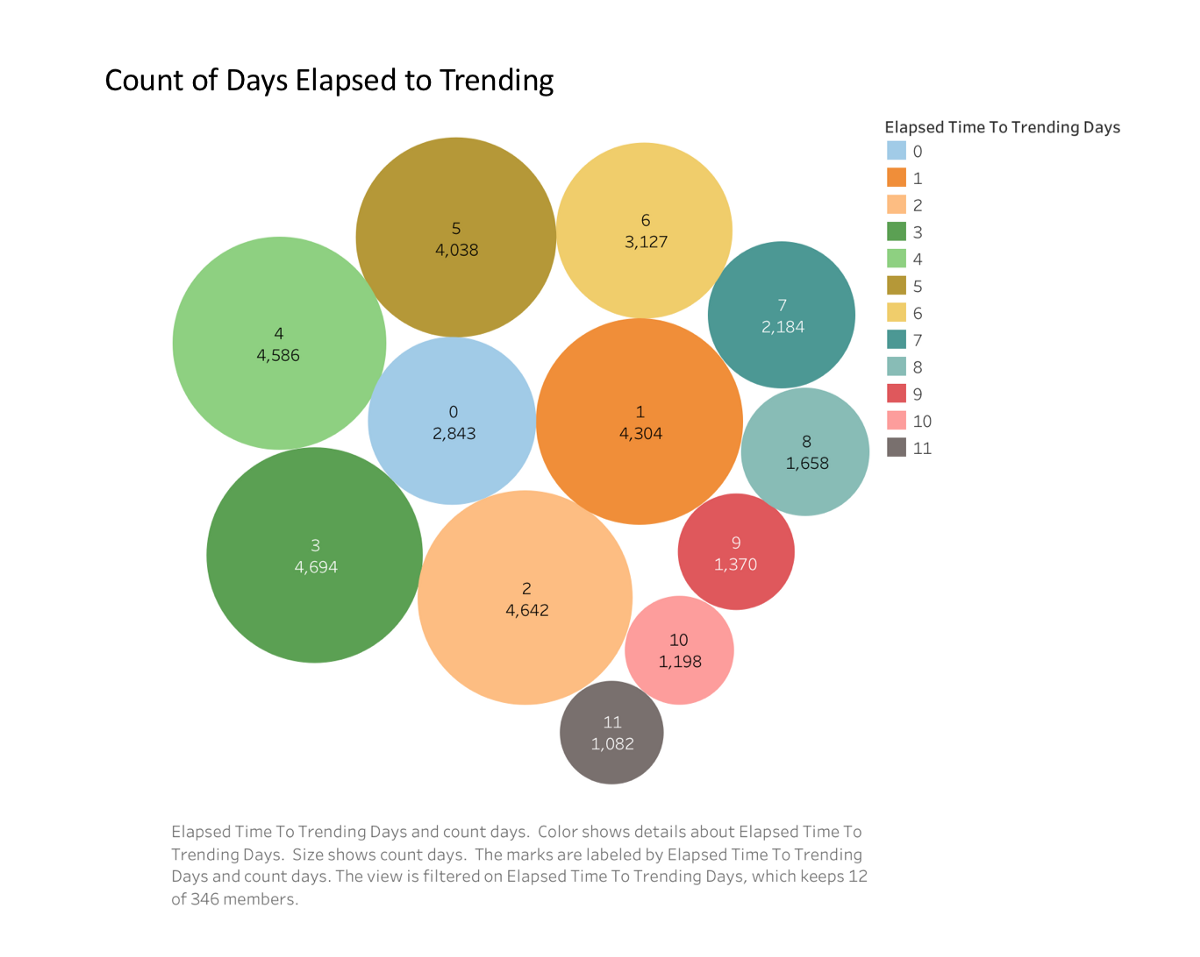
We then decided to explore the hour of publication for each video. From the distribution plot, we see that most videos are uploaded in the mid afternoon around 15:00 and 16:00.



**How long does it for a Video to Become Trending?**

Next we computed the elapsed time (in days) between publishing date and trending date, and used a BubbleChart from Tableau (below) to see how long it takes on average for a video to become trending.

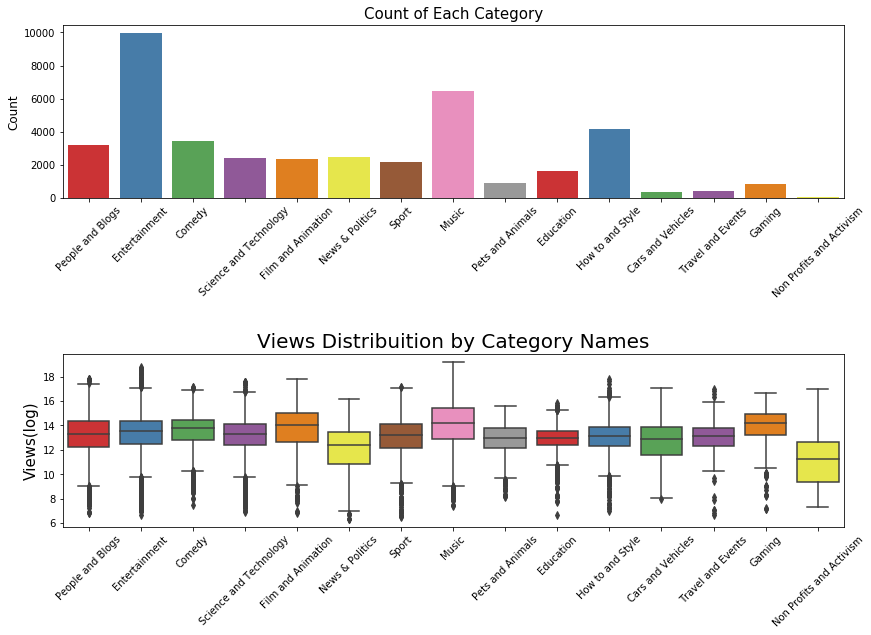
It is easy to observe that the majority of videos become trending within a 0 to 5 days span, after they are published; however, there are some instances in which some videos took a relatively longer time to become trending in the YouTube community.



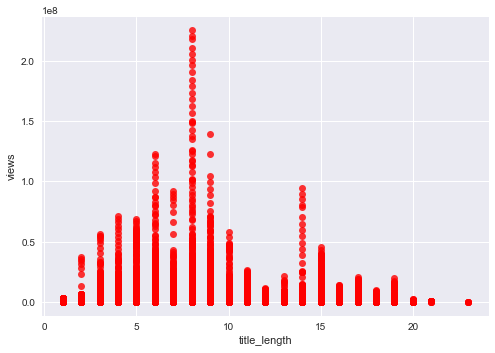
**Which are the Most Represented Categories?**

In order to better understand the differences and characteristics among categories, we plotted the following graphs. It is not hard to conclude that in US, the majority of YouTube videos belong to the “Entertainment” category, followed by “Music” and “How to and Style”.

Not surprisingly, “Entertainment” and “Music” also have the largest range for number of views as the boxplots below shows.



**Do longer titles bring in more views?**

Here we wanted to explore if there is a relationship between the length of title (number of words in title) and the number of views a video can receive, since titles are the most important factor to make a video eye-catching. Does the length of title have an effect on the views? How many words should YouTubers insert in their titles? We used the following scatterplot to verify these questions.We found that videos with around 5 to 10 words attract the most views, which implies that YouTubers should consider utilizing short and concised titles when naming their videos. A long title could actually play a negative factor on their number of views.

**Word Cloud**

To gain a better understanding of the string variables in our dataset, we developed WordClouds to analyze the main words and topics within titles and tags, to see whether there exists some interesting type of trend.



WordCloud of titles show that videos that most trending videos contain the word “official” , “VS” and “SNL”. From our personal experience we know that challenges are a never ending topic in the YouTube community (possibly justifying the presence of the word “VS”); also SNL has a very popular page on YouTube regarding Entertainment Videos. Words like ‘Spaghetti’ and ‘Burrito’ represent the popularity that Food Vlogs have gained in the last few years. Looking at the WordCloud of Tags we find that the words “TED”, “SNL” and “fortnite” are very popular and can be used by YouTubers to attract more viewers. You most likely heard of the latest videogame fortnite’s popularity and it’s presence in the WordCloud above confirms this fact.

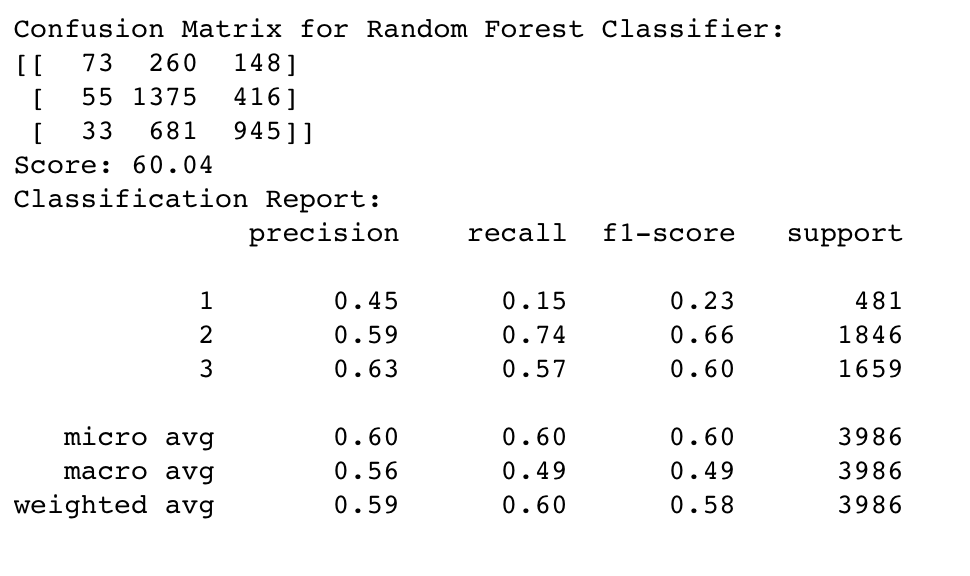
WordClouds were useful to get an idea of what people are interested in and looking for, and to inspire content creators to include catchier words in their videos.

**Machine Learning — Predicting Views**

At this point we shifted are attention to the bigger problem of predicting the number of views a video could get.

Looking at the data at hand, the only variables a YouTuber can control a priori are “hour\_of\_publication”, “title\_length”, “comm\_disabled\_true” and “ratings\_disabled\_true”. We started by running a linear regression and testing the statistical significance of the variables when predicting the number of views. All 4 variables aforementioned were statistically significant, however the initial outcome resulted in a very low R². To solve this problem we decided to divided the number of views into different range groups based on its quarter distribution.

In the picture below we show the outcome obtained on only one category — Entertainment — since different categories might share different strategies to achieve more views. By comparing different algorithms including Linear Regression, random forest, Classification Tree, and XG Boost, Random Forest classifier returned the best outcome with the Accuracy Score of 60.04%.



Outcome Table for Random Forest

The three ranges we created, are videos with less than 100,000 views; videows between 100,000 and 1,000,000 views; and videos with more than 1,000,000 views. With the random forest classifier, we predict the category of videos with 100,000 views or less correctly with a precision of 0.45. Similarly, the precision for Range 2 is 0.59 and 0.63 for Range 3.

**Business Application**

Let’s now see how our model can be applied in practice. Let’s say Duke University’s Fuqua School of Business has its own YouTube Channel to promote their programs and increase popularity among interested students on the web. We used the model we created on the Education Category and obtained a relevant Accuracy Score of 69%.

What Fuqua can do now is say they want to upload a new video called ‘MQM Students’ Everyday Life’ and upload it at 1 PM with comments enabled and rating enabled. Our model predicts that the video will obtain views in the range of 100,000 to 500,000 views.

The image to the left shows anyone can tweak or modify the parameters of the model as they wish in order to obtain the range of views it could potentially get.