**Big Data Final Report**

1. **PROJECT INTRODUCTIONC**

YouTube is the most popular video platform with the largest user base. One of the most important reasons for its popularity is that YouTube allows everyone to be a blogger. It enables content creators to share their content with a large audience. As a result, there’s a wide range of content to choose from. No matter what you want, such as makeup tutorials, cooking guides, product reviews or travel vlogs, you can always find them on YouTube.

However, with millions of contents published every day, it’s super hard for a video to stand out. Our project aims at figuring out what factors make a YouTube video a hit. Here, we use Python and Hadoop to analyze the Trending YouTube Video dataset, which includes the data on daily trending YouTube videos from 10 regions, with up to 200 listed trending videos per day.

The main goal of this project is to get insights into trending videos, to find out what’s the common features. Moreover, YouTube has a wide range of users who are from all over the world. People from different areas have different tastes on videos, so we analyzed user preference based on regions. Knowing that which category is their favorite, we could deliver exactly what they want. Hopefully, these insights will be helpful to the content creators who want to increase the popularity of their videos.

1. **DATASET DESCRIPTION**

We use the dataset of Trending YouTube Video Statistics, which is obtained from Kaggle. It contains data about trending videos for 10 countries. Here, we will use the datasets from USA, Great Britain and Canada.

The dataset is big data for the following reasons.

* VOLUME

First, let’s talk about volume. The size of our dataset is large, we can hardly deal with the data in Excel. Canada dataset is 61.1 MB, the Great Britain dataset is 50.75 MB and the US dataset is 59.85 MB. Moreover, the dataset is collected from the data generated through YouTube, which has over 1 billion users. Every day, these users contribute to billions of videos, comments, clicks etc. A great volume of data is generated every minute and even every second.

* VARIETY

Our dataset also has a wide range of data forms. There are text data, such titles, tags and stream data, such as like and comment. Also, the data is generated both by humans or by machines. For example, the comments are generated by humans while the link is generated by machine. The dataset classifies all these types of incoming data into various categories.

* VELOCITY

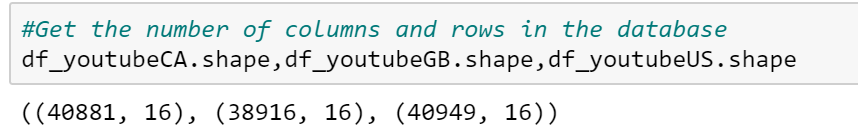
With Velocity we refer to the speed with which data are being generated. The speed of data flow on social media is quite fast, especially on YouTube, one of the most popular social media. Millions of actions happen in every second on the website. In our dataset, in every second, there are thousands of users watch the video or make comments at the same time. The trending video may also change every minute. Big data allows YouTube to get the real-time incoming flow of data and at the same time process it fast.

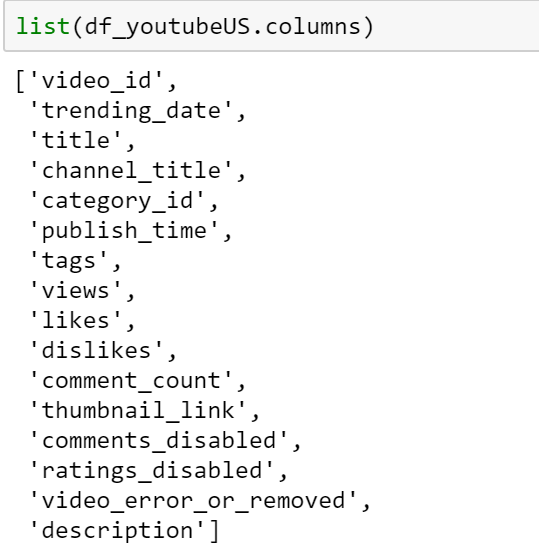
For the respective of 3Vs, our YouTube dataset is absolutely big data.

Here are the general descriptions of our datasets. Since the overview of these three datasets are similar, this part we will take US dataset as an example.

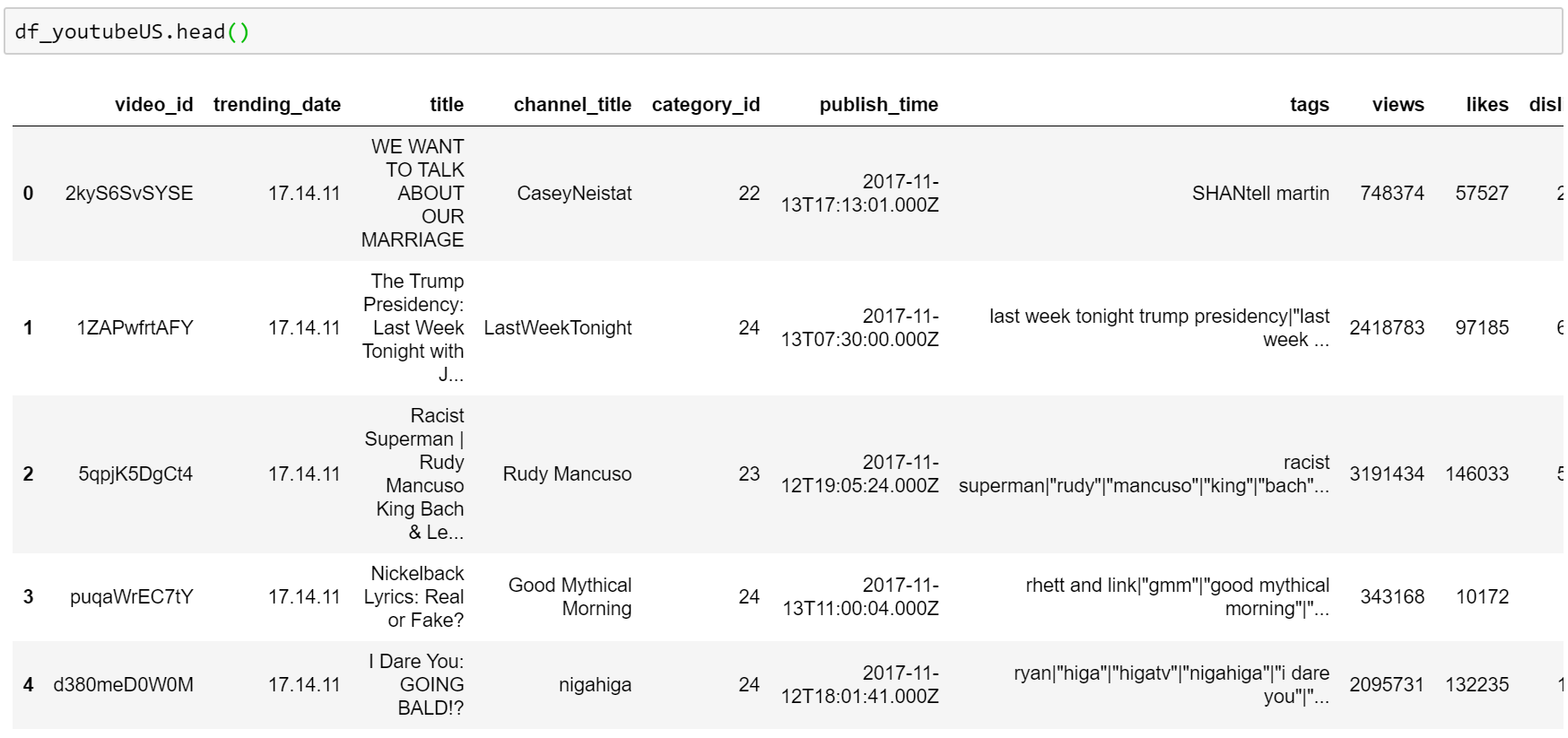
* **Dataset Overview**

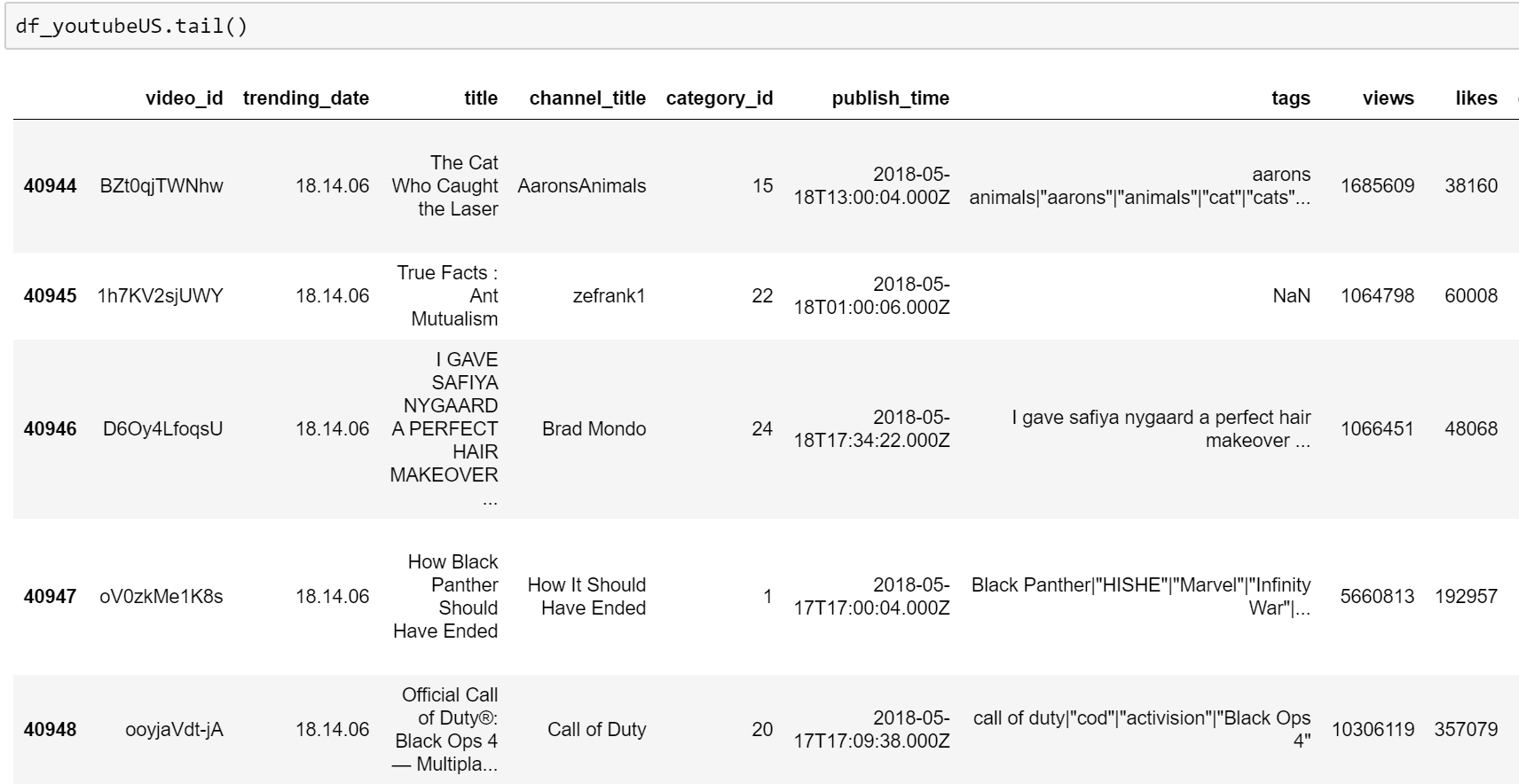
First, let’s cast an overall view over each dataset. We can see that each dataset contains 40,000 trending videos observations with 16 variables, including video id, trending data, title, channel, category, etc.





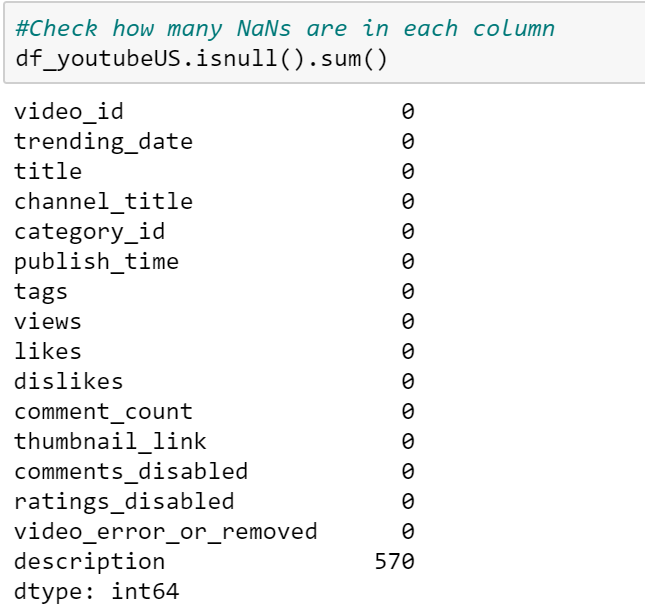
Then, we start from getting the first and last few rows dataset. From the result, we can tell that each dataset collected the top 200 trending videos over 205 days, from 2017/11/14 to 2018/06/14.





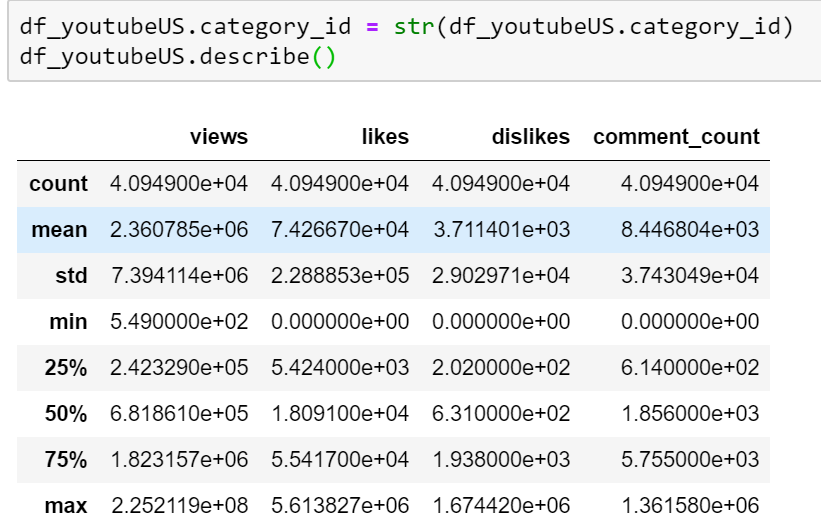
* **Data Cleaning**

Before getting down to analyze the data, we first clean the data. There are some missing data from the description column. We replace the NaN values with an empty string.



* **Statistical Description (Need more analysis here)**

After cleaning the data, let's see some statistical information about the numerical columns of our dataset.



The average number of views of a trending video is 2,360,784. The median value for the number of views is 681,861. The average number of likes of a trending video is 74,266, while the average number of dislikes is 3,711. The average comment count is 8,446 while the median is 1,856.