**ETL YouTube Video Data Analysis**

Dataset ~ Kaggle ["Trending YouTube Video Statistics"](https://www.kaggle.com/datasnaek/youtube-new)

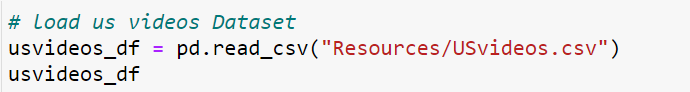
CSV: USVideos

JSON: Category\_id

Converted JSON to CSV

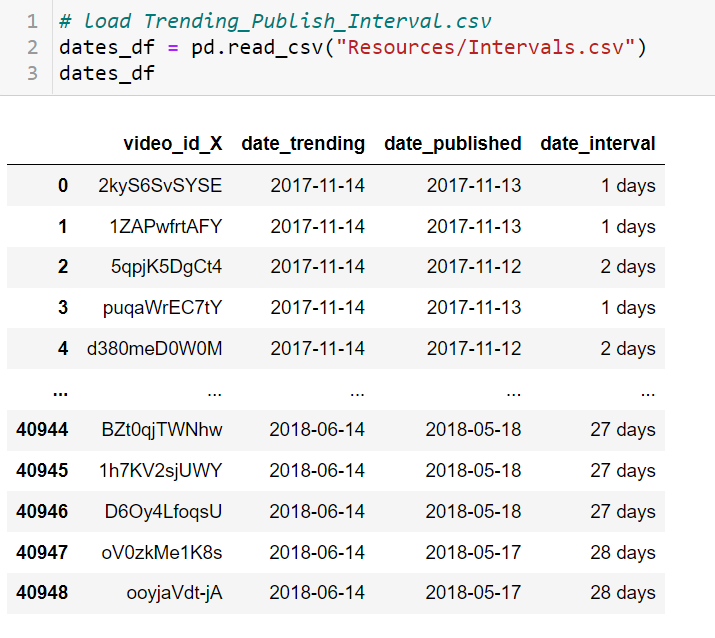
Filter CSV two columns: category\_Id and Snippet/title (genre)

Read and Examine the Dataset



Part A ~ Initial ETL of dataframe columns

* View columns
  + # of Columns 16
* Check data types
* Check Null Values
* Summary of values
  + # of Rows 40,949
  + # of unique video\_id 6,282
* Drop columns
  + Description
* Conversion of Boolean to integer datatype
  + Ratings\_disabled
  + Comments\_disabled
  + Usvideos\_error\_or\_removed
* Trending Published Interval
  + Show # of days from the date it was published to the date it trended



* Concatonate dataset with the interval date
  + Show # of days from the date it was published to the date it trended
* Exported df to CSV ~ ETL1\_columnscleaned.csv

Part B ~ Create dataframe according to number of views in descending order

* Sort views column in descending order
* Count duplicate data
* Exported df to CSV ~ ETL2\_desc\_order.csv

Part C ~ first trending date

* group data over columns 'video\_id' and 'date\_published' a
* dropped duplicate data keep first entry date only
* Exported df to CSV ~ ETL3\_first\_tdate.csv

Part D ~last trending date

* group data over columns 'video\_id' and 'date\_published'
* dropped duplicate data keep last entry date only
* Exported df to CSV ~ ETL4\_last\_tdate.csv