Analysis of Trending Youtube Videos

Group 3 Members: Vidya Giri
Rouzbeh Salehi Pour
Michael Taylor
Daniel Shackelford

Group 3 Github: https://github.com/MRDVGroupProjectCSPB4502/MRDVGroupProjectCSPB4502

Description:

This project intends to identify what determines the performance of a trending YouTube video through the analysis of string and numeric data. Questions related to the data include, but are not necessarily limited to:

String Data:

- Punctuation,
- Capitalization in video titles
- Tags
- Title letter count
- Amount of characters in a video description

Numeric Data:

- Video views
- Video like/dislike ratio
- Time of day
- Video duration

Questions

1. Perform sentiment analysis on video performance as it pertains to punctuation (exclamation points, ellipses, maybe even emojis)? Do videos titles with capital letters get more views than video titles in sentence case?
2. How does categories/tags affect trending duration? Do videos of similar category follow each other, most notably after a spike in viewership. (if someone gets an abnormal amount of views on a video, do videos in the same category see an aftershock).
3. Do videos that reach the monetization time length coincide with a higher view count? What other factors correlate with higher viewership?
    - Video like/dislike ratio
    - optimal amount of letters or words in a video title
    - length of a video description
    - Specific day/time to maximize views

Prior Work:

Existing work on Kaggle includes:

- Sentiment Analysis USA (Step by step| Emojis|Tags)
  - Analyzing if a video is receiving positive or negative sentiment based on natural language processing.
- Youtube US Trending Videos EDA
  - Exploratory data analysis with US trending videos.
- Word Cloud of South Korea YouTube
  - Analyzed the data generate a word cloud of string data.

Dataset:

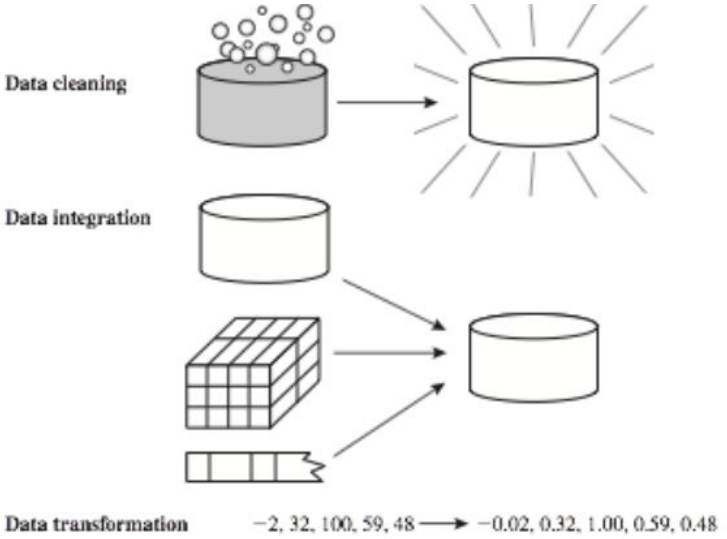
- Our dataset is from Kaggle and is the YouTube Trending Video Dataset (updated daily)
- The raw .csv dataset is uploaded on our team github repo
- The dataset includes the daily trending YouTube videos from various regions with up to 200 listed videos from each day
- Some attributes that are included are the video title, channel title, date published, video tags, views, likes, dislikes, description, thumbnail, and comment count

	video_id	title	publishedAt	channelId	channelTitle	categoryId	trending_date
0	3C66w5Z0ixs	I ASKED HER TO BE MY GIRLFRIEND...	2020-08-11T19:20:14Z	UCvtRTOMP2TqYqu51xNrqAzg	Brawadis	22	2020-08-12T00:00:00
1	M9Pmf9AB4Mo	Apex Legends | Stories from the Outlands – "Th...	2020-08-11T17:00:10Z	UC0ZV6M2THA81QT9hrVWJG3A	Apex Legends	20	2020-08-12T00:00:00
2	J78aPJ3VyNs	I left youtube for a month and THIS is what ha...	2020-08-11T16:34:06Z	UCYzPXprvl5Y-Sf0g4vX-m6g	jacksepticeye	24	2020-08-12T00:00:00
3	kXLn3HkpjaA	XXL 2020 Freshman Class Revealed - Official An...	2020-08-11T16:38:55Z	UCbg_UMjlHJg_19SZckaKajg	XXL	10	2020-08-12T00:00:00
4	VlUo6yapDbc	Ultimate DIY Home Movie Theater for The LaBran...	2020-08-11T15:10:05Z	UCDVPcEbVLQgLZX0Rt6jo34A	Mr. Kate	26	2020-08-12T00:00:00

Slice of dataset

Proposed Work:

1. **Data Cleaning:**
   - Ensure Timeliness of data values across each subset of data for each country
   - Remove redundant data and data duplicates
   - All rows containing null values will be dropped or filled
     - Measure central tendency of the data set and use to replace missing values when possible
   - Non-English Speaking country data will be removed
   - Outliers will be removed from data analysis
   - Data reduction based on less-used attributes
2. **Data Transformation:**
   - Normalization of data set in order to implement clustering techniques
   - Will need to scrub data for title attributes and associate with grammar rules (for analysis of Title trends)



✦ **Min-max normalization** to [new_min$_A$, new_max$_A$]

$$v' = \frac{v - \min_A}{\max_A - \min_A}(new\_\max_A - new\_\min_A) + new\_\min_A$$

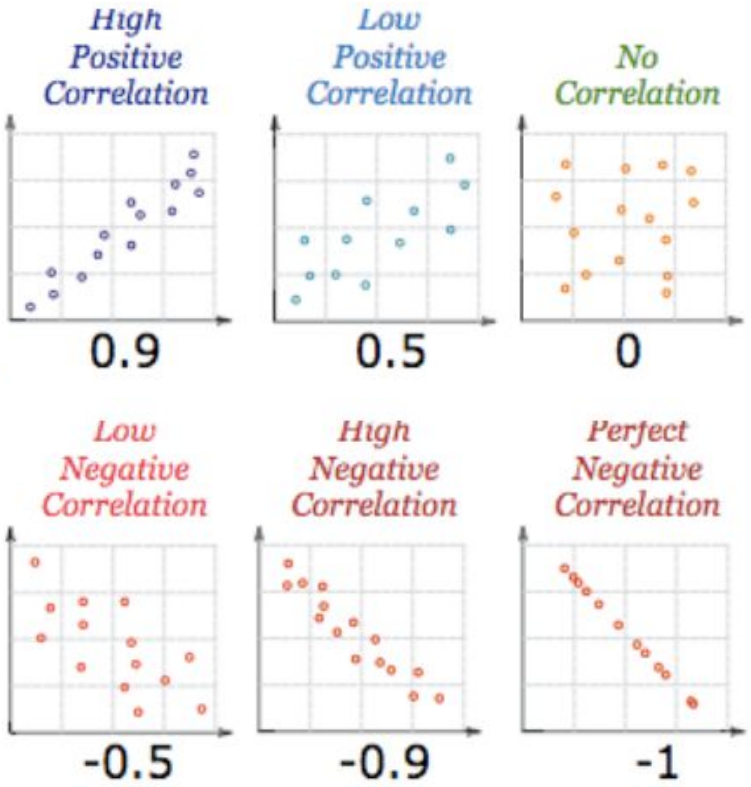
Proposed Work: (Cont)

3. Data Integration:

- Current data set is scraping data daily, will establish baseline to work off of through evaluation phase
- There are multiple similar data sets from the same source that can be used to create a more robust solution.

4. Create a correlation analysis between data attributes

- Can be ascertained through use of chi-square test for categorical data and correlation coefficient
  - Needed for analysis of tag correlation and overall viewership number analysis



List of Tools Needed:

1. **Sklearn**: Python tool with various classification, regression and clustering algorithms
2. **Seaborn**: Python data library to create information visualizations
3. **Matplotlib**: Python plotting library to visualize statistical analysis
4. **Pandas**: Library for performing data analysis easily, particularly useful for cleaning rows of data
5. **Numpy**: Library for performing computations on data

Evaluation:

- Dependant on models
- Visualizations
  - **Confusion matrix**: table of actual vs predicted classifications
  - **ROC curve**: compare classification by models
  - **Validation curve**: identify potential for overfitting
  - **Visualizing clusters**: scatter plot identifying different classes
  - **Other**: histograms, correlation heatmaps, pair plots, etc.
- Performance metrics
  - **Recall**: how many identified instances are actually true?
  - **Precision**: how many actual true instances have been identified?
  - **F1 score**: harmonic mean of recall and precision
  - **Silhouette coefficient**: scoring separation of clusters