Youtube trending video analysis

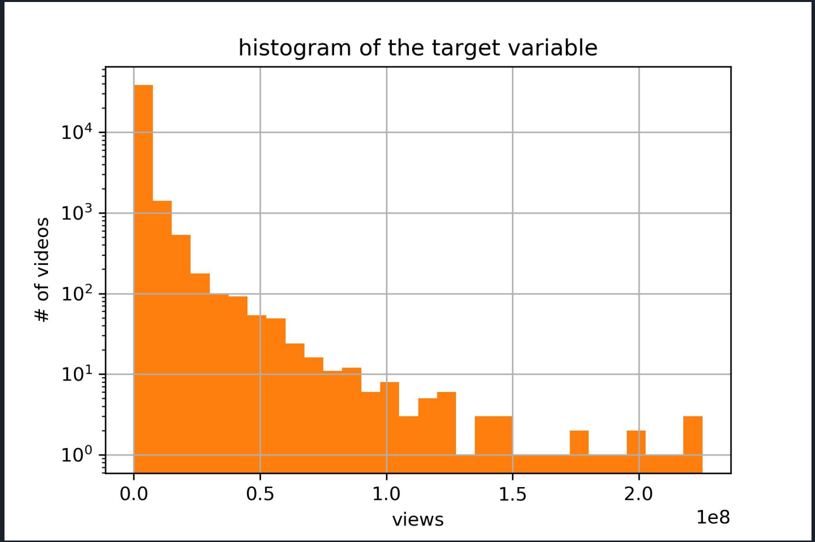
Presented by William Ward
Brown University
October 21st 2019
https://github.com/wward97/data1030project

Project Summary

- I want to analyze youtube's trending videos to see if i can determine some of the features that lead to a YouTube video being added to trending and gaining millions of views.
- To do this I analyzed 41000 videos that were marked as trending over a 1 year period.
- I analyzed the categories they belonged to and looked at likes, dislikes and comment counts to see if there was as garnable relationship.
- My target variable was views and as a result this is a regression problem.
- I think this is very interesting since most of the media I consume is from YouTube. YouTube is also becoming a space for budding talent to gain a lot of exposure, so seeing what factors lead people to gain popularity is a relevant conversation. Especially as more and more people can start using YouTube as full time careers due to revenue from advertisements, sponsorships and patreons. (a money donating app that lets you pledge to creators giving them a monthly salary of around 1 to 2 dollars per person up to 100s).
- My data set comes from kaggle. https://www.kaggle.com/datasnaek/youtube-new#USvideos.csv
- All of my data comes from the US however. This project could easily be expanded to analyze other countries and it could be pretty interesting to compare.

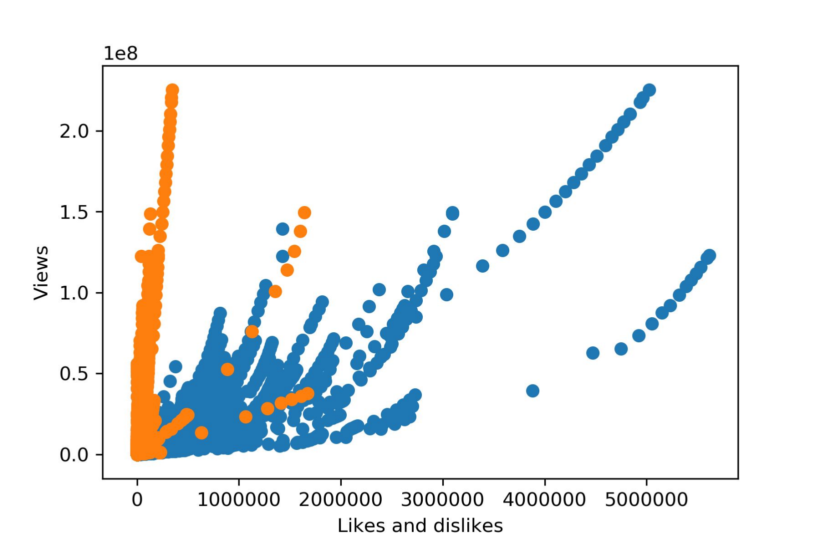
DATA CLEANING PREPROCESSING

- General trend of target variable shown on the right
- No clearly missing data
- Things that I removed from my original data set include things like:
  - Title
  - Channel names
  - Tags
  - Description (large strings of text)
- I worked with tags for a long time, separating them and oraganizing them into they're own df. In the end too large to process however.
- OHE on categorical features and scaling on continuous features.



EDA THOUGHT PROCESS

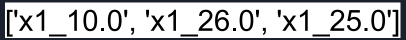
- Felt likes would be too strongly correlated with views. Not in the direction I would want. I checked likes and dislikes vs views:
- As you can see there is a strong correlation between likes and views.
- We also see that dislikes has a weird relationship with views. On one hand few dislikes does not tell us much about the views. However a lot of high view videos have no dislikes and some low view videos has no dislikes either.
- To see how much of an effect this is i use f_regression to see p score or f score to see which features are the most influential:



EDA

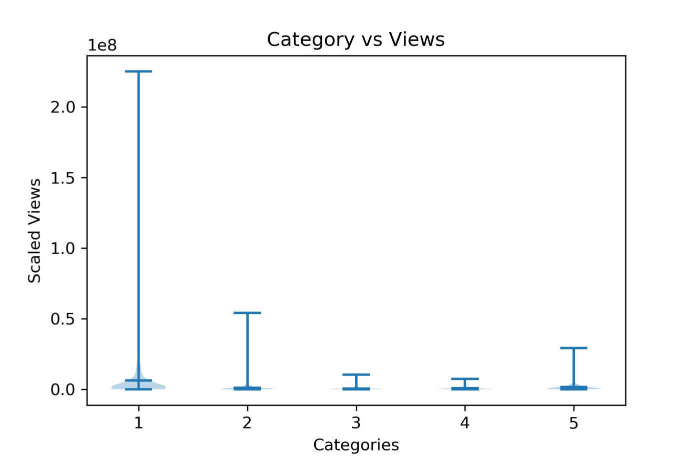
- To see how much of an effect this is i use f_regression to see p score or f score to see which features are the most influential.
- The top 3 features by f score are as expected:
    - Index(['likes', 'comment_count', 'dislikes'], dtype='object')
- This overwhelming one sidedness of the results is what prompted me to change my approach with how to decide important factors. The idea now was to use, instead of likes dislike and comment count, to use the ratio of likes to views dislikes to views and comment count to views.
- This hopefully reduced the snowball effect of hugely popular videos having incredibly high likes but the likes not contributing to the high views.

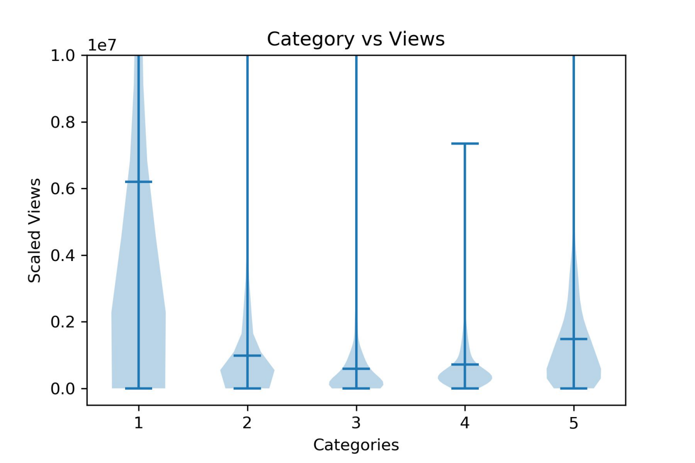
EDA

- Now I repeat the above but having removed likes dislikes and comment count and replace it with the ratio of that to views and repeat f_regression and get:
    - 
      ```
      ['x1_10.0', 'x1_26.0', 'x1_25.0']
      ```
    - These are keys tied to categories and those top three categories are in order:
        - ["Music","News & Politics","Howto & Style"]
- The fact music is the most popular is not much of a surprise. Music is often the most viewed set of videos on youtube eg 14 of the 15 most watched videos on youtube are music videos.
- The other two being the next most influential however is very interesting.

EDA

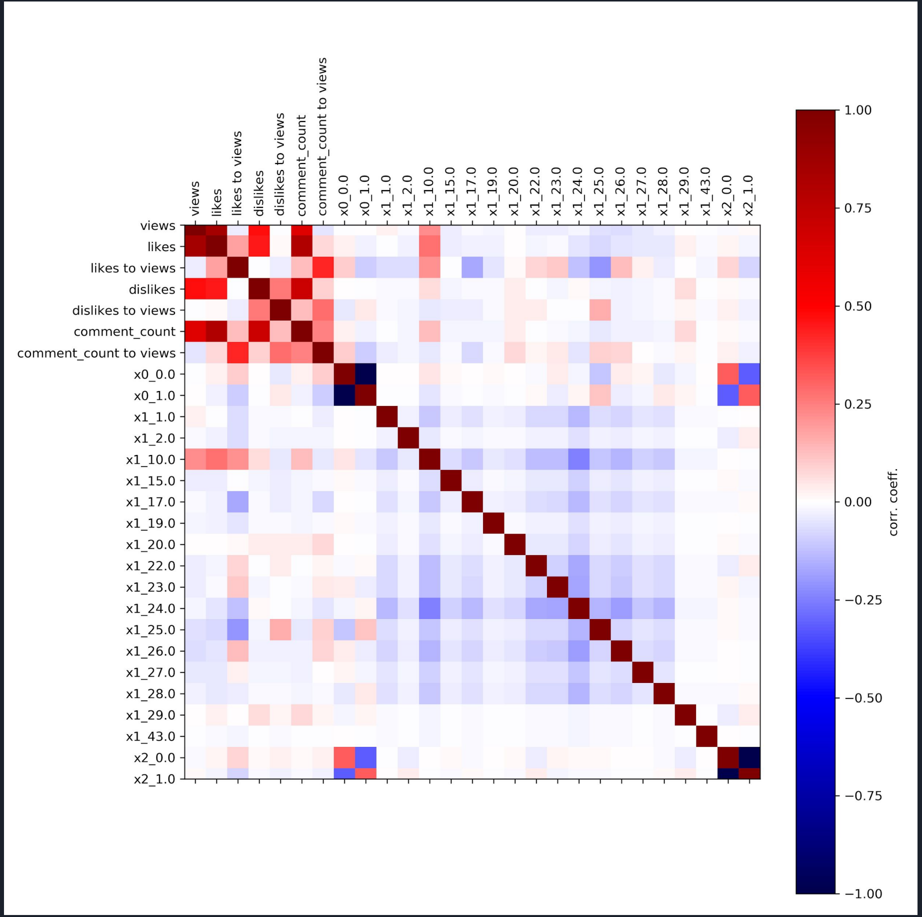
- To illustrate the relationships between categories and views i created this violin plot of the top 5 categories:
- As you can see the extremes of music are astronomically high compared to the rest. This is why these data points are problematic. Upon zooming in we get a slightly better idea of how things look
- As you can see even among youtube's trending, an area meant to show some of the most interesting and best videos to watch, there is a huge discrepancy on the number of views a video has.
- 1 to 5 Equivalent to:
  - “Music”, “News and politics”, “Howto and style” ,“Education”, “Comedy”





EDA SCATTER MATRIX

- This scatter matrix gives us a more in depth look at the relationships between each variable.
- Things of note:
  - Correlation between dislikes and comment count
  - Likes dislikes and comment count all have a high correlation coeff
  - Most of the categories have a negative corr. coeff.
  - Looking at likes to views you can see some more interaction with the categories.



Thank you for listening