# Capstone Project 1: In Depth Analysis of Trending Videos on Youtube

By Vikram Singh

# Table of Contents

# Introduction

**Data used in this project: https://www.kaggle.com/datasnaek/youtube-new**

The purpose of this project is to analyze trending videos on Youtube with the purpose of providing knowledge to companies regarding where to focus advertising and promotions on Youtube. In the current era, Youtube has become more popular than television, and has become a massive platform for advertising, and promotions of every kind. By looking at characteristics of videos that go trending and then subsequently analyzing the relevant metrics, insight can be gained into the way videos are received by viewers, the content that trends frequently, and the types of videos that receive the most views. The goal of this project is to reveal hidden knowledge regarding videos as a whole on Youtube that wouldn't be apparent otherwise.

# Cleaning and Observing the Data

The data used in this project contains trending video statistics from November 14th 2017 to June 14th 2018, and is restricted to videos that were trending in the United States. The features of this dataset include:

- Video_id : Unique identifier given to videos uploaded on youtube
- Trending_date : Date during which video was trending
- Title : Title of the video
- Channel_title : Title of the Youtube Channel/Content Creator which uploaded the Trending Video
- Category_id : Identifier to denote the category to which a video belongs
- Publish_time : Time at which video was published on Youtube
- Tags : A collection of words, phrases, and characters affiliated with a particular video, as specified by user uploading the video
- Views : The number of views a video received
- Likes : The number of likes a video received
- Dislikes: The number of dislikes a video received

- Comment_count : The number of comments a video received
- Thumbnail_link : link to thumbnail for video
- Comments_disabled : specifies whether comments for the video are disabled
- Ratings_disabled : specifies whether ratings for the video are disabled
- Video_error_or_removed : specifies if video has been removed from youtube
- Description : description provided by user uploading the video. Provides descriptive content regarding a video.
- Category Title: Custom field which I inserted into the data, mapped from the category_id field. Contains the title of the category to which a video belongs.

For the scope of this project, I decided to focus my analysis on the following metrics of interest: Views, Likes, Dislikes, Category and Comments. I am more interested in a quantitative analysis regarding how well a video performs in terms of these metrics and how this can vary across a wide variety of content. For this reason I will not be performing a detailed text analysis regarding text based columns, nor will I be heavily exploring the date/time variables. The features involving whether certain features are “disabled” are seldom True across the entire dataset, so I’ve chosen to leave them out of my analysis.

Upon finding this dataset, it was clear that the data was pretty clean and didn’t require much modification. There was no need for cleaning due to outliers, NAN’s, or reorganization due to the highly diverse nature of youtube’s trending videos. The only “missing” data found in the data set in a column of interest is the “description” column of 570 videos that simply do not have a description. I didn’t replace or alter these observations, as the description of a video is a field that can intentionally be left blank when a video is uploaded.

# Exploratory Data Analysis

Initially, I thought that since the videos in the dataset are trending, they would all be widely popular and have received a lot of attention. I was curious if there were any exceptions to this, and began my EDA with the intent of finding what I thought would be rare observations of

videos with very low quantitative metrics. As my exploration progressed, it became clear that my initial thought process was an almost flipped version of what was found in the data. While there are many videos which go trending that garner massive amounts of views, there appear to be many trending videos that are almost hidden due to how little attention they get! To my surprise, I found a much higher amount of data with values of 0 than I expected, and it was completely accurate. In this dataset of 40949 observations, I was curious exactly how frequently values of 0 were found.

**'0' Value Count for Important Features**

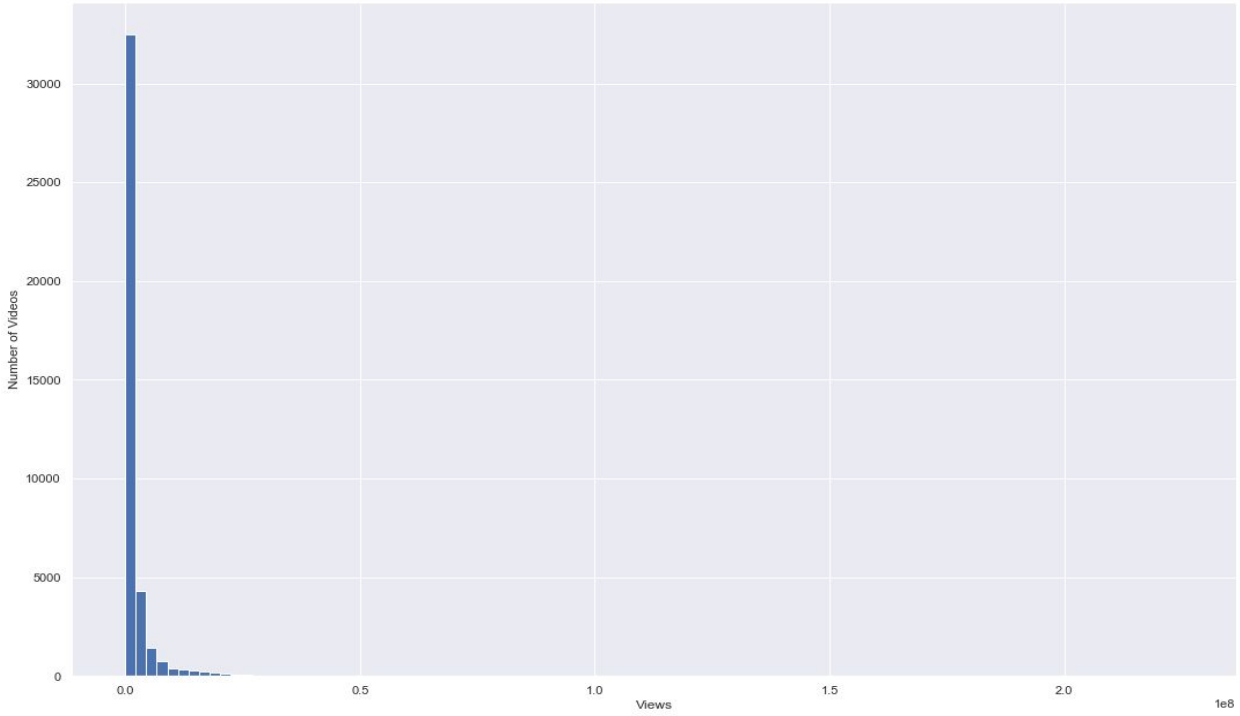
| Variable | # of observations with value of 0 | % of observations in dataset with value of 0 |
|---|---|---|
| Views | 0 | 0 |
| Likes | 172 | 0.42% |
| Dislikes | 383 | 0.935% |
| Comments | 760 | 1.855% |

Relatively speaking, none of these metrics exceed 2% in terms of values of 0 and could be considered low. However, this frequency still seems very high considering these videos are trending and would be expected to have gain significant attention, yet statistics of 0 would indicate otherwise. This analysis also only accounted for strictly 0 values and gave me an indication that many values in the dataset will be very small, so I decided to analyze the entire distribution with some summary statistics and visualizations. By looking at a set of the lowest and highest values for each variable, I found that there were videos that had extremely low, and extremely high values, and this was consistent across all variables.

**Summary Statistics**

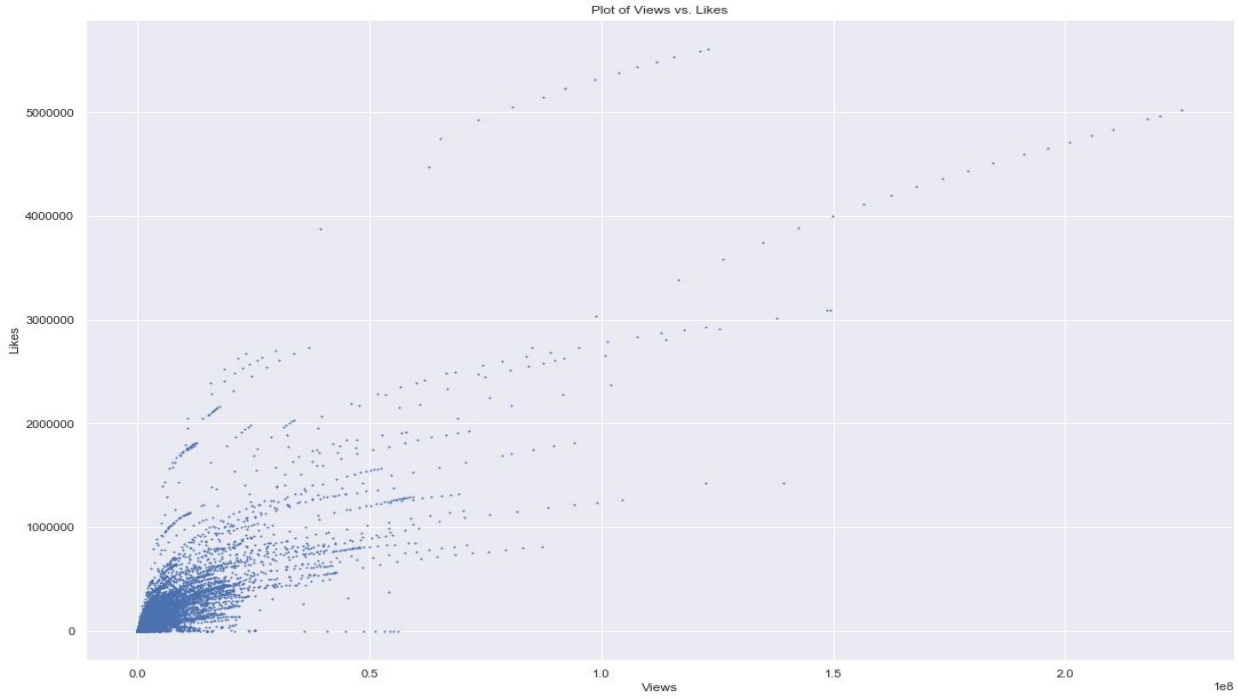
| Variable | Lowest Value | Highest Value | Range | Mean | Median | (Mean - Median) |
|---|---|---|---|---|---|---|
| Views | 549 | 225,221,923 | 225,221,374 | 2,360,784 | 681,861 | 1,678,923 |
| Likes | 0 | 5,613,827 | 5,613,827 | 74,266 | 18,091 | 56,175 |
| Dislikes | 0 | 1,674,420 | 1,674,420 | 3,711 | 631 | 3,080 |
| Comments | 0 | 1,361,580 | 1,361,580 | 8,446 | 1,856 | 6,590 |

It is apparent that there is an extremely high range across these variables, but surprisingly it seems that there is a large difference between the means and medians as well. When the mean is higher than the median, this indicates a right skewed distribution. By plotting the distribution found for views, we find that there is indeed a highly right skewed distribution.



The distribution for likes, dislikes, and comments is very similar to the above graph. Initially, I thought that videos with low values for these variables would be considered anomalies, because if a video is trending, one would think it received a lot of attention. However, these statistics reveal that most trending videos don't receive much attention, and that the videos that have massive amounts of views are almost considered to be the anomalies considering how rare they are. Such a statement can be considered subjective, depending on what one considers the threshold for a video to have received "sufficient attention". Ultimately, this data does definitively show that during this time period, a trending video is more likely to have statistics closer to the lower end of values for quantitative variables.

I then proceeded to determine the level of correlation between variables, which revealed many insights that are somewhat intuitive. Given that the amount of views a video received is the target metric when discussing the attention a video received, I began by checking the correlation of views with other variables. Between "Views" and "Likes", a high R of 0.849 was found, and ended up being the strongest correlation present between two variables.

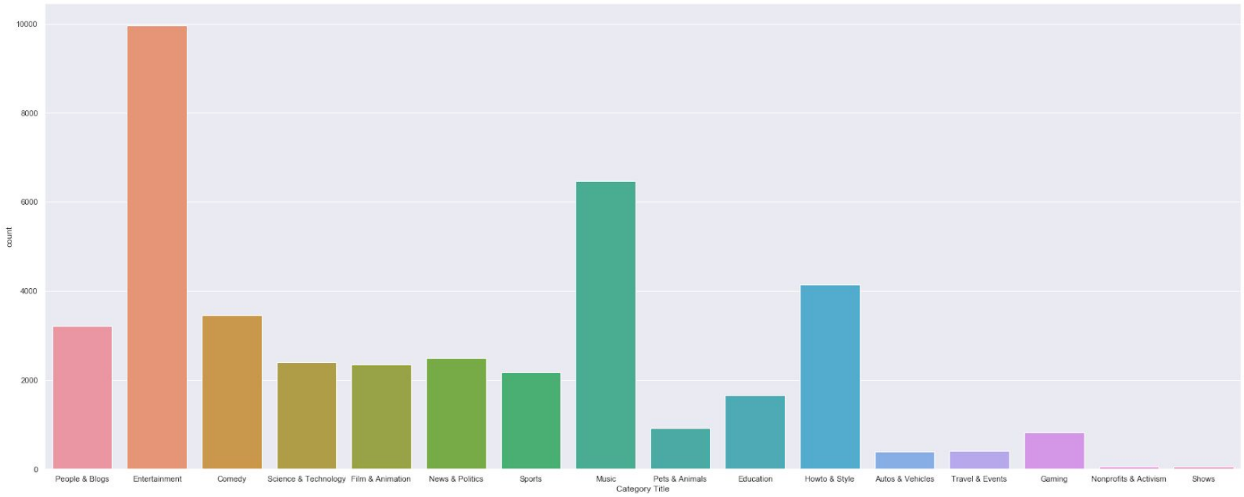


An interesting correlation was present between likes and comments, which is furthered by the correlation between dislikes and comments. Both of these combinations showed rather high correlations compared to other variables. This could indicate that videos which are received by the public very positively or negatively prompt the public to express written sentiment. However, between the two the correlation is higher between likes and comments, indicating people are more likely to express sentiment on videos that are well received. Other combinations of variables did not exhibit particularly strong correlations.

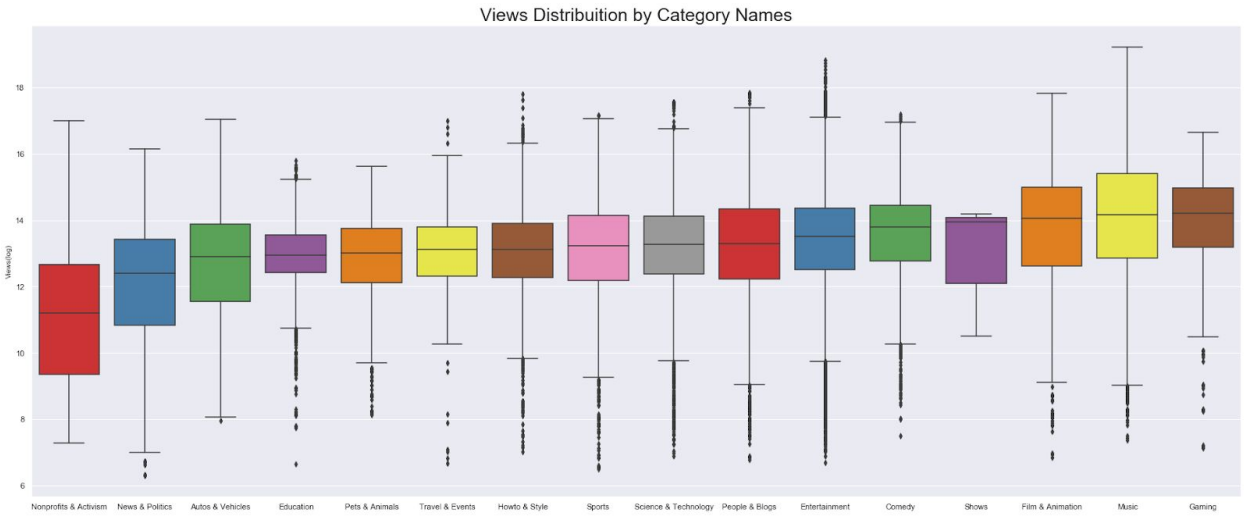
**Variable Correlations**

| Variable 1 | Variable 2 | R Value |
|---|---|---|
| Views | Likes | 0.849 |
| Views | Dislikes | 0.4722 |
| Views | Comments | 0.617 |
| Likes | Dislikes | 0.447 |
| Likes | Comments | 0.803 |
| Dislikes | Comments | 0.7 |

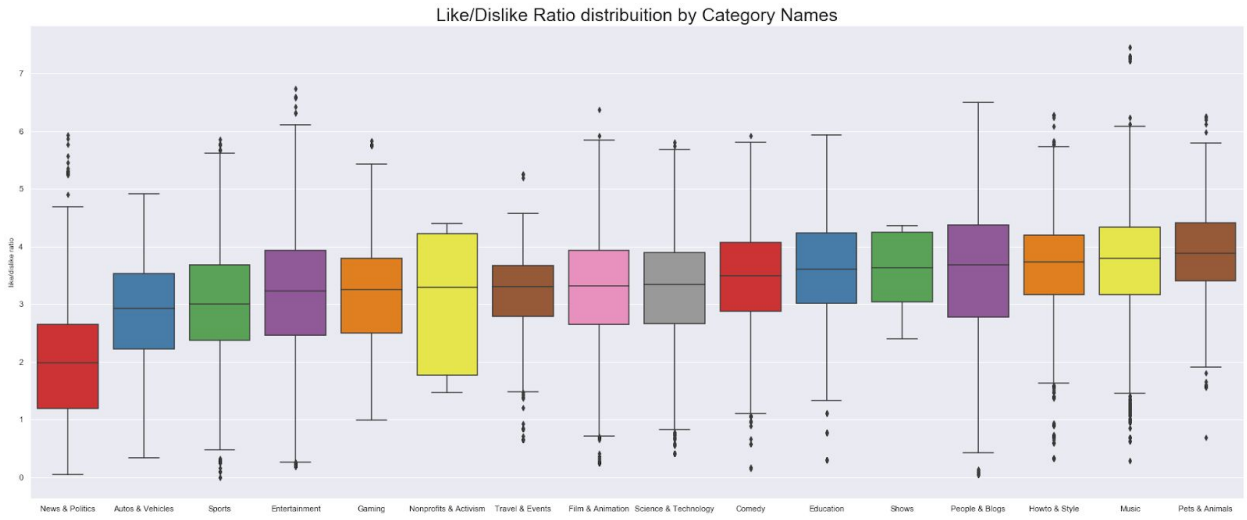
The EDA performed thus far has been on features of the entire dataset, to observe summary statistics and their relationships as a whole. It is apparent that the data is quite diverse, and when videos are categorized based on their content, it would be remiss to assume that all categories on youtube have the same summary statistics and frequency. Although this dataset contains only trending videos, the amount of videos and number of views per category can provide valuable insights into how the category of a video is impactful, and by extension, how the public receives the video.




A count plot reveals that videos of category Entertainment are most frequently trending on youtube, with Music coming in second. Youtube is a platform for content of every type, yet it is no surprise that Entertainment is the most commonly trending, because youtube is known to be commonly used for entertainment and many kinds of videos would fall into that category. The placement of Music is also expected, as youtube is known to be a musical platform for mainstream, and aspiring musicians to post audio and video recordings of their work for a respective audience. The Shows and Nonprofits and Activism categories are the least frequent, having almost no entries in the dataset. Although the frequency of videos by category provides some insights, it doesn't answer the most important question, how does the performance of a video vary by category?



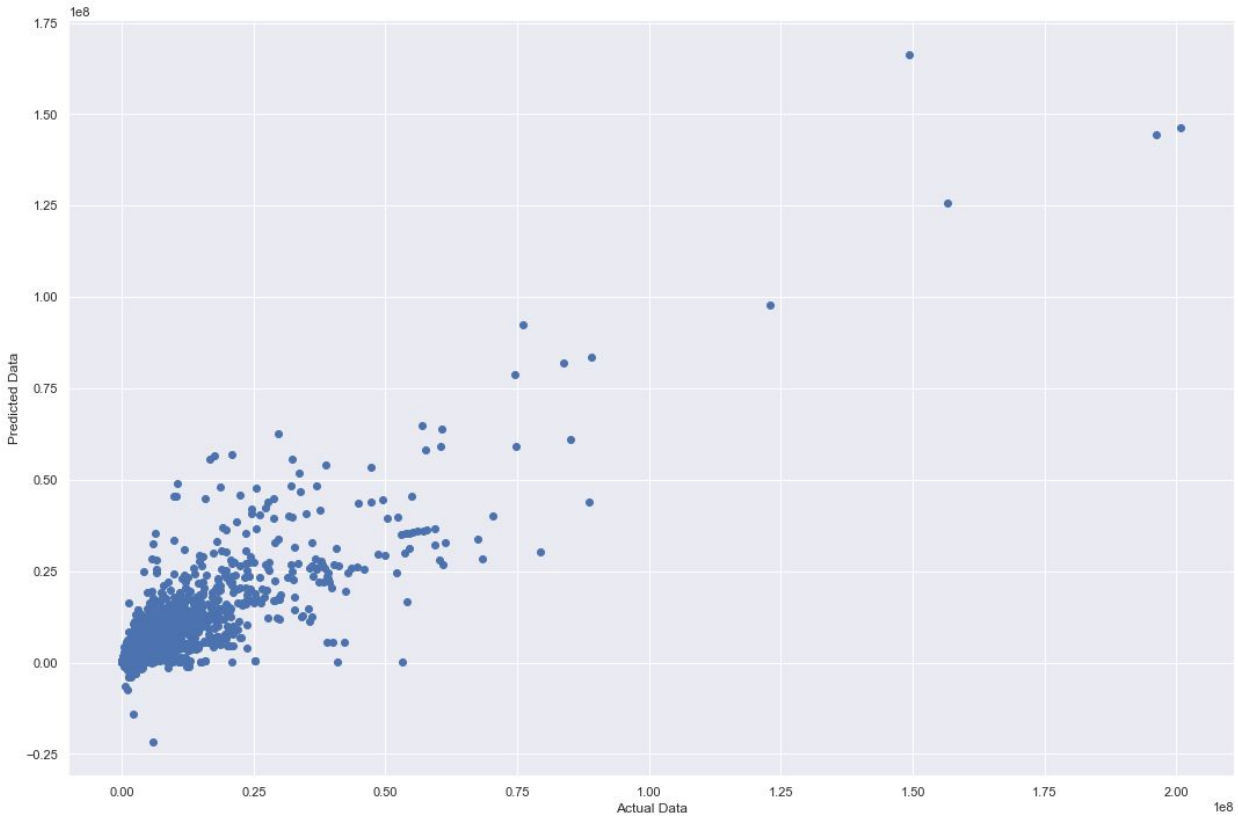

Using the median to rank categories in this logarithmic boxplot of views per category, it can be seen that gaming videos are the most popular. This is quite interesting, considering that the number of gaming videos that go trending are very low. Videos in the music category come in second place, which furthers the notion of how popular youtube is as a music platform. Even though videos of category Entertainment are by far the most frequent, they're right in the middle of the rankings, which is likely due to several videos which received few views being present in this category. Although Nonprofits & Activism come in last place, it can be seen from the countplot that there are very few videos belonging to this category, In such cases, there is insufficient data to accurately compare the views in this category to others.




Even though we know more about how frequently categories are viewed, it doesn't reveal any information on how well the public receives different kinds of content. To measure this, I've chosen to use the median of the "Like/Dislike Ratio", which is simply the amount of likes divided by the amount of dislikes for a video. Intuitively, this makes sense, as videos with more likes are met more positively by the public, while videos with more dislikes are not met well. Videos of the category Pets and Animals have the best ratio, while News and Politics has the worst. People generally enjoy animals, so this result is somewhat expected, and due to the controversial nature of various news networks and politics as a whole, it comes as no surprise that the category comes in last. The Music category also performs exceptionally well using this comparison, and demonstrates a great combination of popularity in views, and reception on Youtube. Although the Entertainment category has the most trending videos, it seems that they have one of the worst like/dislike ratios. This could be due to the fact that many of the videos which trend are not of a very high quality, or that the entertainment category as a whole is split due to the many different kinds of entertainment.

# Machine Learning

The target variable when building a model, is the amount of views that a video received, as this is the key variable in determining how much attention and reach a video has. Being the case, this is a clear regression problem where the number of views was to be predicted. From my EDA, it was clear that higher values of likes, dislikes, and comments were associated with videos which received more views, so I started with a linear regression model to form my predictions.

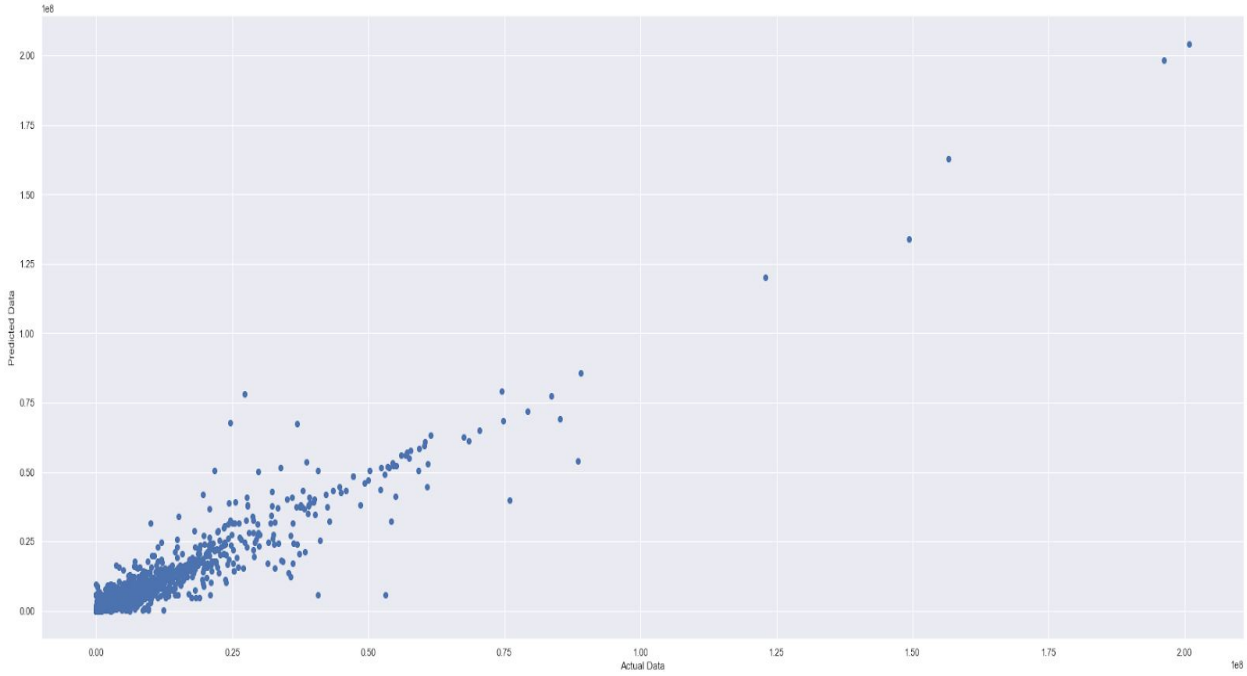



The model performed better than I expected, with actual and predicted values having R ~ 0.85 - 0.89 consistently. There are a few issues, the first being that the y intercept coefficient is 229,220, implying that a video with no likes, dislikes, or comments, would have 2292,20

views. The range of views is very large on this data set, with the least viewed video having only 549 views, but the most viewed video having over 225,000,000. Most of the entries in the dataset fall between 0 and 25,000,000 views, as can be seen from the graph, but the y intercept being much higher than the number of views many videos actually received does cause a high MSE for those entries which have lower views as a consequence. With all of this considered, it's safe to say that this model generally performs better on videos which have more views.

The next issue of using Linear Regression, comes from the coefficients of the model. For some reason the "comments" variable has a negative coefficient. This could potentially be a consequence of many videos with less views disabling comments, or is simply a limitation of the model when considering the amount of likes and dislikes typically provided on a video are much higher than the number of comments. This leads to a few negative predictions from the model, which appear to be on videos that had a low view/comment ratio. These negative predictions are inaccurate and also contribute highly to the MSE of the model.

To circumvent the limitations of the linear model, I created a model using random forest regression, and tuned the needed hyperparameters accordingly to achieve the highest score (based on R^2 value) that I could. This isn't an extremely complex model, so the main hyperparameter that required tuning was the number of estimators. I determined that n_estimators = 9 was the number that produced the best results, and the predicted vs actual values are displayed below:




This model produced excellent results, with R = ~ 0.96 - 0.97, indicating a very strong correlation between the actual and predicted values. By comparing the linear model with the Random Forest model, it's easy to tell that videos with lower numbers of views have a more linear relationship with their predicted values. There are also no negative predicted values, and the predictions across the board appear to be significantly stronger.

By cross validating the data from both of these models, the scores (based on R^2 value) appear to consistently fit with the R values for the graphs shown above. These models can also be filtered to act on data pertaining to videos of a certain category, although the graphs above demonstrate the capacity of the model to act on the dataset as a whole.

## Final Summary/Recommendations

A trending video is a video that has been selected to be featured by Youtube, which is why it is expected that such videos would be quite popular. Initially I would have thought the

same, but my analysis proves otherwise. Videos of every category over an extremely large range of several variables have gone trending, and because of this comprehensive coverage of values, I believe it is acceptable to infer findings from this project to be generally applicable to the content on youtube as a whole, within reason. Any entity which uses youtube as a promotion platform, and any current or aspiring content creator, can benefit from these recommendations.

Attention should be focused on videos in the music category for the best results. Music is the consistently top performing category on youtube in terms of public reception, and breadth. Videos of this category performed extremely well in both of these metrics, and goes to show the power of youtube as a musical platform. This might not be obvious initially when considering that youtube is platform for uploading video content rather than audio, but this is clearly not the case and music videos are amongst the most popular videos on the site. Despite the features of this category being strongly impacted by such music videos, analysis of the entire dataset confirms that this positive performance exists in general.

Popularity does not necessarily reflect positivity. With the exception of Music, the most popular categories on Youtube in terms of views, appear to be some of the least popular categories in terms of public rating. Entertainment videos are the most commonly trending by a significant margin, and rank very well when looking at view count, yet are generally received more poorly than most categories. This margin is even more apparent with gaming videos, which has the highest median number of views of any category, yet performs on the same level as Entertainment in terms of public reception. Although these negative trends clearly don't hold true with respect to every single video in these categories, as a whole they do reflect the general trends which are present. Niche categories such as Pets and Animals, Howto & Style, and People & Blogs to promote content serve as a solid balance between receiving ample

views, and public reception. Similarly, videos of a category such as Sports will attract attention, but typically aren't received well by the public and can reasonably be avoided to an extent.

## Conclusion

The process of data wrangling, exploratory data analysis, and application of machine learning has revealed insights which essentially debunk the idea that trending has significant impact on a video, and exposed the areas where youtube videos have the most success. I have truly enjoyed analyzing the features of this data, and observing how their behavior changes across categories. By observing identifiable trends in the data, in conjunction with intuitive thinking, new insights regarding youtube have been provided which provide valuable knowledge, which would not otherwise be apparent. I look forward to continuing my data science journey and taking on more complex and focused projects in the near future.