Data Analysis of U.S Youtube Dataset

Authors: Reina Yu, Michael Woo, Ary Sharifian

Summary:

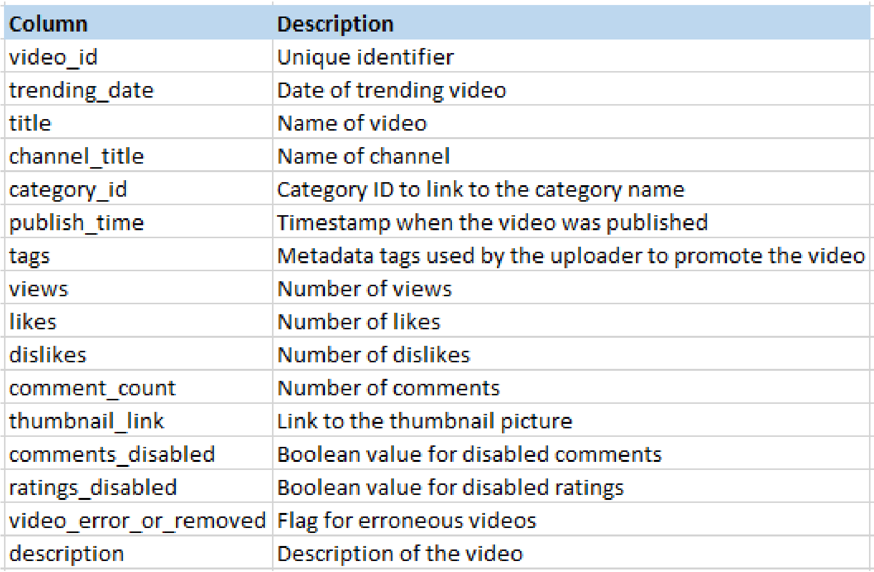
YouTube is a video platform for users to upload and share content across the globe. This mainstream form of social media has more than 2 billion users daily. To better understand the statistics and trends of this social media platform, we were determined to analyze the dimensions and metrics content owners should use in order to maximize their view counts. Here, we explored and determined common keywords, popular tags, optimal title length, viewer engagement, and most viewed categories. As a result, we concluded that the following key insight obtained from our analysis are the following: 1) common keywords amongst titles were "official, video, trailer, ft, and new", 2) title length with around 50 characters have the highest view counts, 3) popular tags are funny, comedy, how to, and pop, 4) viewer engagement for disliked videos vs. liked videos, and 5) the most popular category with the most view counts are entertainment, music, health and style, comedy and people & blogs. Refer to the *Results* section of this report for an in-depth analysis of these metrics. Given these results, we recommend users to use this strategy to optimize their view counts: upload their content to the top five most viewed categories, use common keywords in the title, ensure title length is around 55 characters and include popular tags. Additionally, we tested the correlation strength between each variable (refer to Table 1). In summary, the correlation between likes and views showed the strongest correlation and may be a good predictor for either views or likes.

To further understand the factors that affect view counts, a linear regression model was implemented to predict the number of views using likes and category ID. By using a linear regression model, we identified potential metrics that forecasts view counts. Based on our model, we determined the number of likes has the strongest predictive nature of forecasting view counts. However, the number of comments is not a strong predictive measurement of view counts. Meanwhile, the number of comments is mediocre predictor of view counts. We used an R-squared and means squared error score to determine the accuracy and robustness of our model. Refer to the *Results* section of this report for a more in-depth analysis of each metric.

Introduction:

The dataset, used in this project, is called "Trending YouTube Video Statistics" downloaded from Kaggle. It is a record of the top trending YouTube videos and consists of seven months of aggregated data on videos from November 14, 2017 to June 14, 2018 in the United States. There are 40,949 observations (videos) in this dataset. Each row represents a video that was uploaded on a specific day. There are 16 columns representing a mix of objects, integers and boolean values (Figure 1). Data cleansing activities included parsing the title and tags into lists, removing stop words such as "a" and "in", and removing null values were performed.

Figure 1. U.S YouTube Dataset Keys and Values



Column	Description
video_id	Unique identifier
trending_date	Date of trending video
title	Name of video
channel_title	Name of channel
category_id	Category ID to link to the category name
publish_time	Timestamp when the video was published
tags	Metadata tags used by the uploader to promote the video
views	Number of views
likes	Number of likes
dislikes	Number of dislikes
comment_count	Number of comments
thumbnail_link	Link to the thumbnail picture
comments_disabled	Boolean value for disabled comments
ratings_disabled	Boolean value for disabled ratings
video_error_or_removed	Flag for erroneous videos
description	Description of the video

Objective and Scope:

The scope of this project is centered around the idea of an individual aspiring to become a youtube influencer. YouTube influencers are people who built a substantial following on the video platform. In order for a novice youtuber to jump start their career, he or she has to know their target audience and the content they would be drawn to. In this project, we analyzed a U.S Youtube dataset and evaluated various levers that may be used to optimize view counts. For instance, we examined metrics like, popular video categories, commonly used clickbait title keywords, title length, viewer engageability, category ID and how the number of likes may affect view counts.

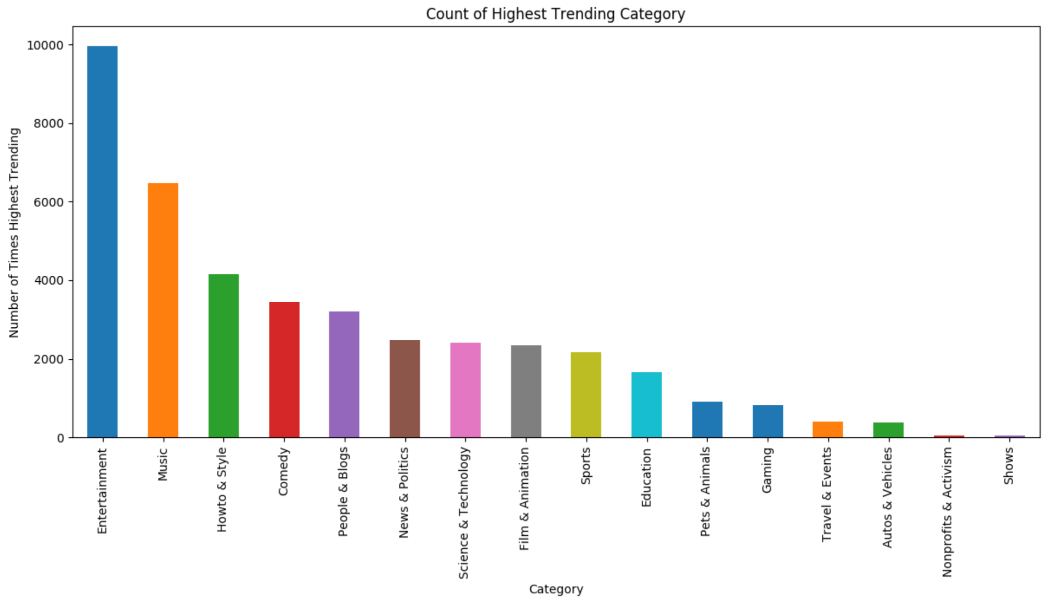
Results and Main Insights

Categories

Youtube has become a content phenomenon across the globe. According to influencer marketing hub, 5 billion videos are viewed daily. With so many videos being watched across the globe, we were determined to find the top watched categories on the social media platform. In order to evaluate this, we extracted the category ID’s for each video on the U.S Youtube dataset and created a list of counts.

Based on our analysis, Figure 2, we concluded the top five categories to be the following: 1) entertainment, 2) music, 3) how to and style, 4) comedy, and 5) people & blogs. This indicates that there is a high demand for these marketable categories and it is prudent for any user to upload in the aforementioned categories. Thus, given these insights, it is highly recommended that users upload on the most viewed categories to optimize their view counts.

Figure 2. Count of Highest Trending Categories

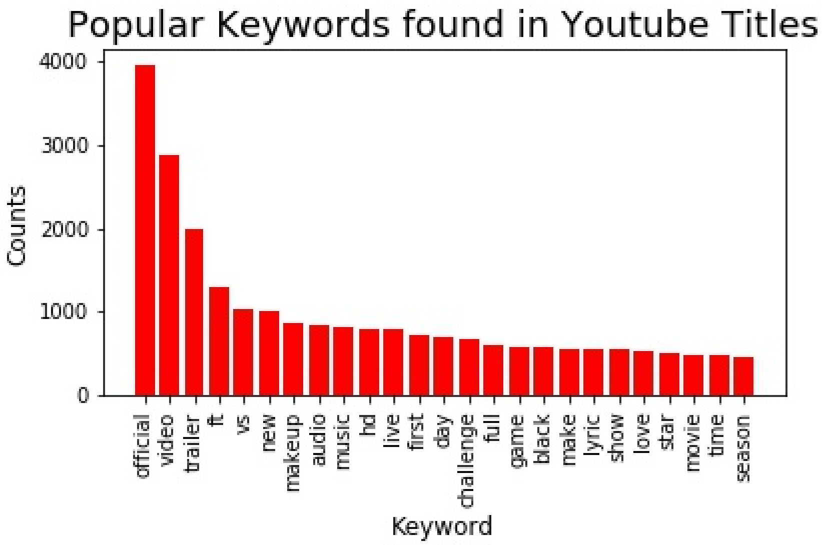


Common keywords

Clickbait is a text or thumbnail link that is designed to entice users to follow that link, listen or view to the linked piece of online content. In this case, we will refer clickbait as text that is used to attract viewers to watch the YouTuber’s videos. Using the U.S YouTube dataset, commonly used title keywords were identified by separating video titles to single string elements in python. Then, the dataset was cleaned up to remove any stop words, digits and case sensitivity. The top five commonly used keywords in video titles are the following: official, video, trailer, ft, vs and new. It is not certain that these keywords are clickbait titles but video titles containing the aforementioned keywords generally have higher view counts.

Given these results, it is recommended that users compose a title with commonly used keywords (refer to Figure 3 for top 25 common keywords). An in-depth analysis of the top 25 keywords is necessary to evaluate if there is a trend between use of common keywords to increase in view counts. However, with the current insights, users may still employ these keywords since they may be considered clickbait terms.

Figure 3. Popular Keywords found in Youtube Titles

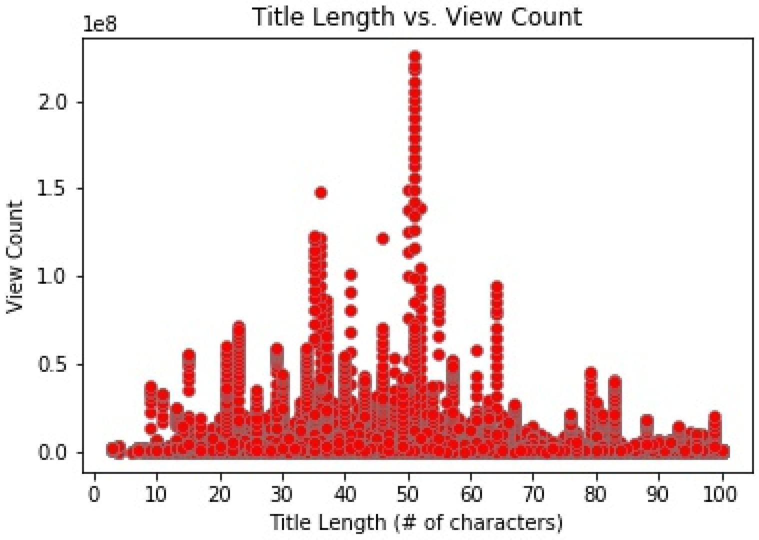


Video Title Length

YouTube has a 100 character limit for titles, anything longer than 70 characters will be truncated in the search results. Given that information, as well as the U.S Youtube dataset, we investigated if there was any trends among title length (# of characters) and view counts. We used python and a python library, called SKLearn and Matplotlib, to count the number of characters in each title and plotted it against the views for that given video. In Figure 4, users with 40 - 65 characters in their title have the highest view counts. Anything outside of that range is less common. This may be due to the fact that shorter titles do not contain enough clickbait words/keywords, while longer titles may be truncated in YouTube's search engine.

Overall, the results suggests that users should maintain a title length of 40 - 60 characters in order to increase the view counts. To be more precise, a title length of 50 characters is the optimal title length.

Figure 4: Title Length vs. View Count

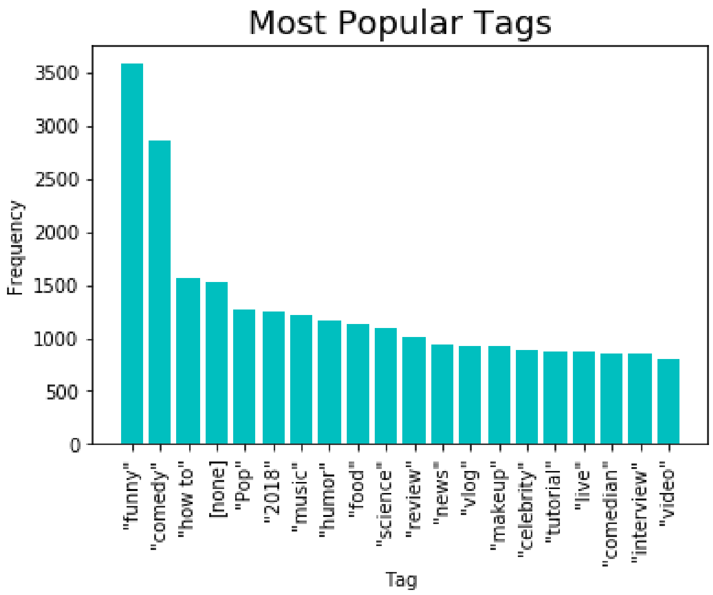


Common tags

Tags are critical in the social media world as it exists today. Creating content as a YouTube influencer is just half of the effort. The other is make it as easy as possible for the viewers to find your content online. This requires the strategic use of tags. Tags, like leaving a trail of breadcrumbs, allows viewers to indirectly navigate to your video based on related videos or searches.To accomplish finding the most popular tags we first had to prep the data by iterating through each record and parsing a pipe delimited string containing the tags. The tags were then appended to a new python list which would contain all tags in the dataset. We then used a Counter from the Collections library to identify the top 20 most frequently occuring tags and plotted this in a bar chart. Popular tags include “funny”, “comedy”, “how to”, “pop”, “2018”.

From the popular tags analysis we learn a bit about what the US audience is watching. In general we look for videos that allow us to laugh and learn. Funny videos and tutorials would be great content for an aspiring YouTube influencer to consider. This analysis could further be improved if we had data that showed which tags were most effective in bringing the viewer to your video. This could include hits on your tags or the tags of the prior referral video.

Figure 5. Most Popular Tags



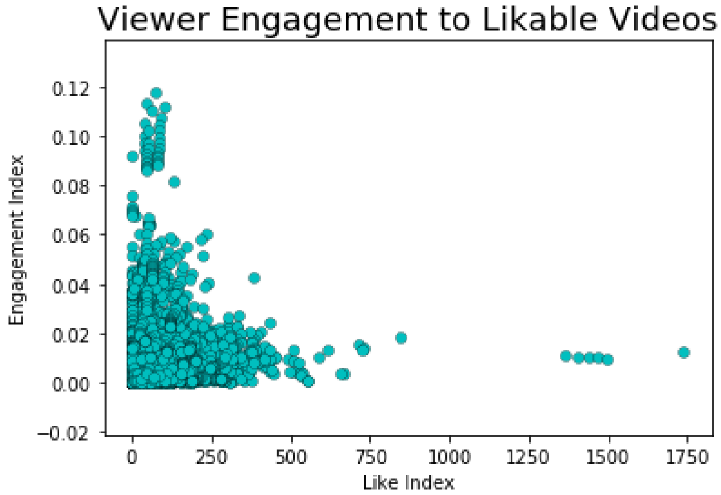
Viewer Engagement Index

View count, while important, is just one of the factors that determine whether a video is trending. This question explores measuring the degree of likable videos, viewer engagement, and their relationship to each other. We have defined two indexes for each video:

Likability Index = number of likes over number of dislikes
Engagement Index = number of comments over number of views

Using these indexes helps us better analyze engagement while removing the bias of comparing counts from someone with 100K subscribers to someone with 1M subscribers. To explore the relationship we plotted these using a scatter plot. Visually speaking, it appears that viewers are typically more engaged with less likable videos.

Figure 6. Viewer Engagement to Likable Videos

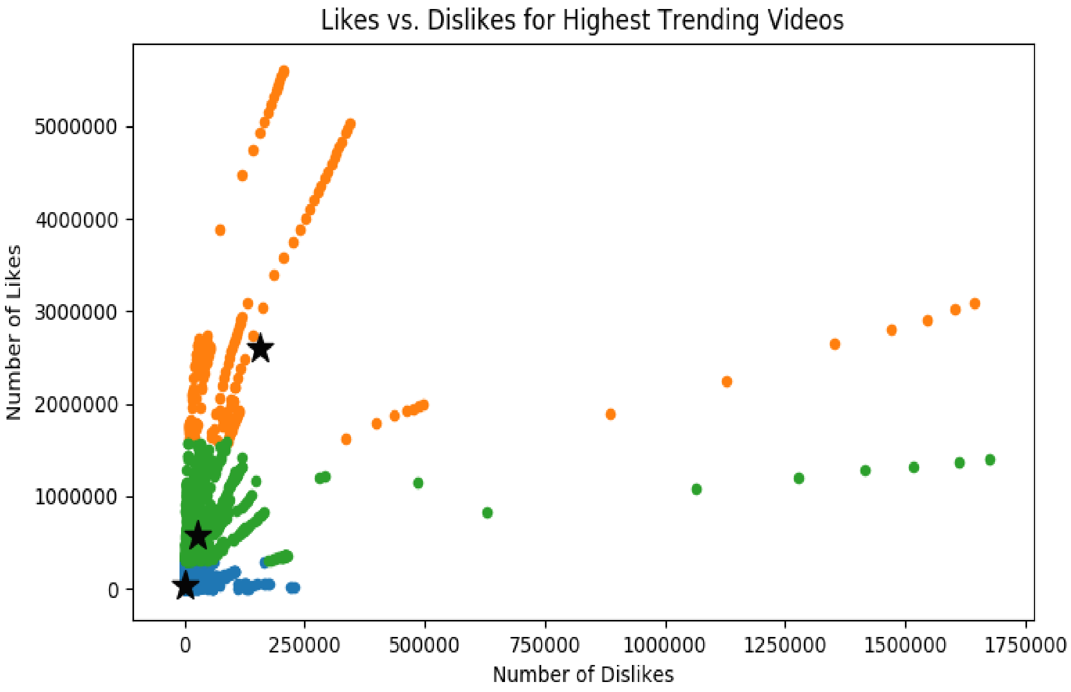


In the likability vs. engagement analysis we learn that the overly well-liked videos don't necessarily garner the most viewer engagement. As a content creator you want your videos to trigger a reaction with a strong presence of dislikes. Once you get a conversation going in your comments section, now viewers are not just watching your video, but returning to respond to comments from others. This could be improved if we had the actual comments to perform sentiment analysis to know what exactly is triggering the comments.

K-means clustering of Likes vs. Dislikes

To further understand likability of a video, a cluster analysis was required to partition the likes/dislikes clusters in which each data point belongs to a cluster with the nearest mean. In this case, we separated the likes/dislikes dataset to three clusters: least liked, liked and most liked (refer to Figure 7). In a greater sense, these insights enable users to visualize the three cluster groups and understand the correlation between the number of likes and dislikes for tending videos.

Figure 7. Likes vs. Dislikes for Highest Trending Videos



Correlation Strength between Metrics

In statistics, the correlation coefficient, r, measures the strength and direction between two linear variables. Here, we analyzed the correlation between several metrics (views, likes, dislikes, comment_count) in order to demonstrate the correlation. If the correlation strength (r) between two variables is greater than 0.70, then this indicates a strong relationship between the variables. According to the results, shown in Table 1, the correlation between likes and views, and likes and comment_counts have a correlation strength of 0.85. Therefore, likes and views have a strong linear correlation and it may be used to forecast the number of views or likes.

Table 1: Correlation (r) Strength between Metrics

	Views	*Likes*	*Dislikes*	*Comment_Count*
Views	*1*	*0.849177*	*0.472213*	*0.617621*
Likes	*0.849177*	*1*	*0.447186*	*0.803057*
Dislikes	*0.472213*	*0.447186*	*1*	*0.700184*
Comment_count	*0.617621*	*0.803057*	*0.700184*	*1*

Linear regression model

Predictive models are considerably useful for forecasting future outcomes and estimating metrics that are practical/impractical to measure. Here, we measured and predicted view counts by using the number of likes in the U.S Youtube dataset. This linear regression model allowed us to determine which metrics were indicative of estimating view counts. SKLearn, a python data science library, was used to implement a linear regression model for the three aforementioned variables to predict view counts.

Based on the results of the linear regression model (refer to Figure 8), the number of likes is a critical metric that forecasts the number of views. Additionally, the R^2 value was determined to be 0.74; this value is significant to our model since it gives us greater confidence in our predictions. In Table 2, the MAE and RMSE may be on the higher end due to the outliers present. If the outliers were removed from the dataset, the MAE and RMSE values will be normalized. Comment counts (refer to Figure 8) were somewhat predictive of views but did not show an extremely strong correlation in comparison to likes. However, the number of dislikes (refer to Figure 9) were not measurable indicators of view counts (also shown in Table 1). Given these insights, there is a strong correlation between the number of likes and view counts.

To further our understanding of our approach, we investigated how category ID and number of likes can be predictive of views. A one-hot encoding method was used to create a binary matrix for the category IDs. The arrays were then reshaped for the linear regression model. Afterwards, predictions were made based on the test data set and the R^2 value was .90. In the future, we plan to create a 3D visualization for this specific model for predicted vs. actual views.

Also, for future linear regression models it would be beneficial to evaluate other metrics like, viewership by country, number of subscribers, age of viewers, and the weather for that particular day. It is important to define such metrics since the factors that affect view counts are limitless. Therefore it is critical to identify more metrics involved in predicting view counts.

Figure 8. Linear Regression of Actual Views (based on likes)

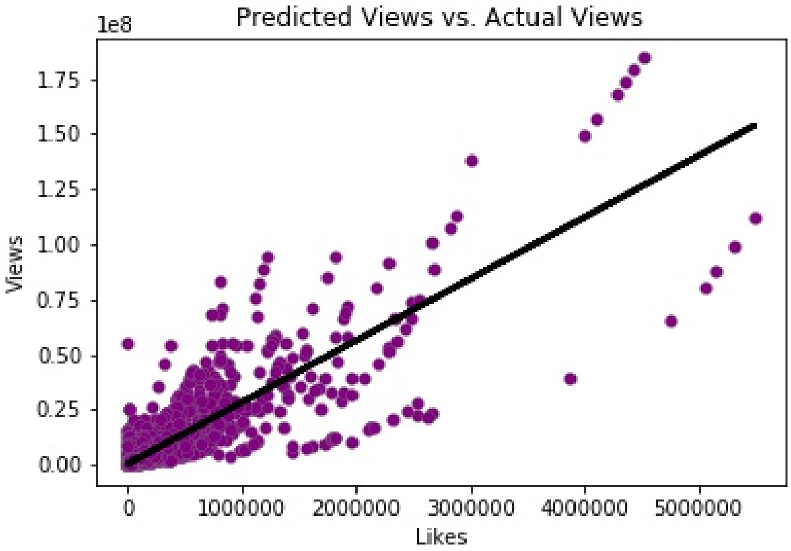
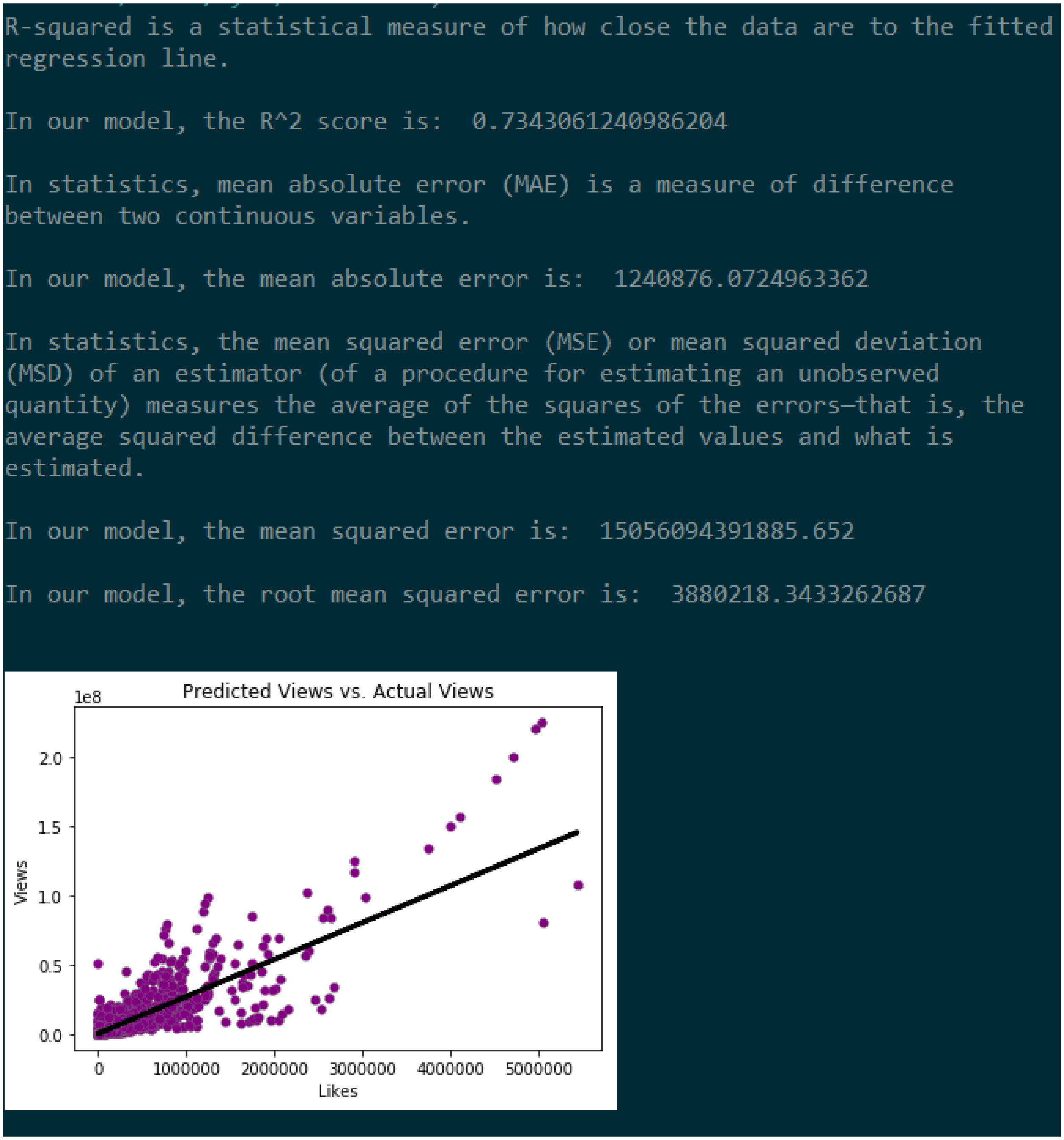


Table 2. Linear Regression Model Accuracy

Linear Regression Model Accuracy		
R^2= 0.75	MAE= 1.3e6	RMSE= 3.96e6

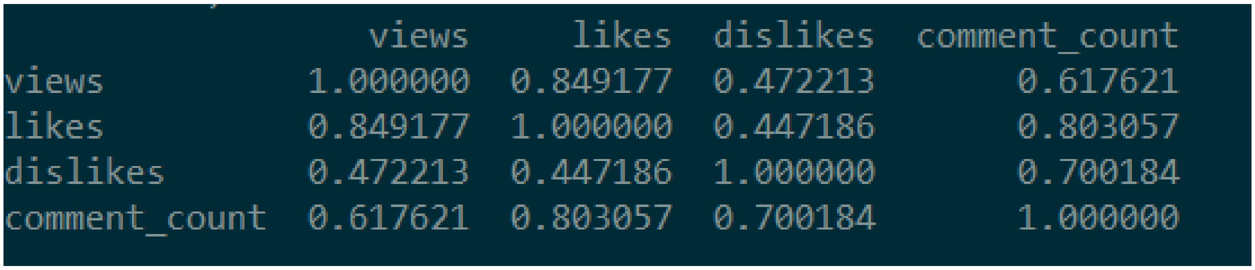
Screenshots

Linear regression model output



```
R-squared is a statistical measure of how close the data are to the fitted
regression line.

In our model, the R^2 score is:  0.7343061240986204

In statistics, mean absolute error (MAE) is a measure of difference
between two continuous variables.

In our model, the mean absolute error is:  1240876.0724963362

In statistics, the mean squared error (MSE) or mean squared deviation
(MSD) of an estimator (of a procedure for estimating an unobserved
quantity) measures the average of the squares of the errors—that is, the
average squared difference between the estimated values and what is
estimated.

In our model, the mean squared error is:  15056094391885.652

In our model, the root mean squared error is:  3880218.3433262687
```

Correlation relationship between variables



```
                 views     likes  dislikes  comment_count
views         1.000000  0.849177  0.472213       0.617621
likes         0.849177  1.000000  0.447186       0.803057
dislikes      0.472213  0.447186  1.000000       0.700184
comment_count 0.617621  0.803057  0.700184       1.000000
```