What Makes YouTube Videos Trend: An Analysis of U.S. Trending Videos

Arina Mokeeva

December 9, 2025

Executive Summary

This project examines the factors that make YouTube videos trend in the United States. Using the USvideos.csv dataset from Kaggle, I will analyze engagement metrics such as views, likes, dislikes, and comments, along with content features like category, publishing time, and tags. I will use Python with Pandas, NumPy, Matplotlib, and Seaborn for data cleaning, analysis, and visualization for interactive exploration. The goal is to identify patterns that influence audience engagement and video visibility.

Table of Contents

Project Scope

This focus is on the main question: What factors influence a YouTube video in the United States to become trending? To answer this, I will use the USvideos.csv dataset from Kaggle, which includes information such as views, likes, dislikes, comment counts, video categories, publishing times, and tags (Oh et al., 2025). Understanding this is important because YouTube is widely used among teens and young adults in the U.S., making trends on the platform relevant for content creators, marketers, and researchers (Faverio & Sidoti, 2024).

I will begin by examining the relationships between engagement variables. For example, I want to see whether videos with higher views tend to receive more likes, and how likes relate to comments as a measure of audience interaction (Yang et al., 2022). Additionally, I will analyze whether specific video categories or upload times are more likely to perform better. Another part of my analysis will focus on how tags affect engagement, specifically whether using trending or popular keywords increases a video's chances of appearing on the trending list. Finally, I plan to calculate the time difference between when a video was published and when it first appeared as trending, which will show how quickly successful videos gain traction.

The goal of this analysis is to identify meaningful patterns between engagement indicators and content features that may explain why certain videos become more visible than others. The results may help creators, marketers, and digital media professionals understand how to improve content strategies and boost audience engagement.

To complete this project, I will go through several stages of analysis using Python in Jupyter

Notebook for visual exploration. The first step will be data profiling and preparation, where I will examine the dataset's structure, check for missing or inconsistent values, and clean the data to make it accurate and ready for analysis. After preparing the data, I will create visualizations in Python to better understand trends in engagement. These visualizations will help display the relationships between variables such as views, likes, comments, categories, and tags.

Next, I will apply data modeling techniques to test which factors best predict a video's likelihood of trending. This step will involve exploring correlations and possibly running regression models to see which variables have the strongest influence. I will then summarize my findings and discuss any challenges or lessons learned along the way.

By using Python for analysis for visualization, I will gain a clearer understanding of what makes certain videos successful on YouTube. Each stage of the project will be built on the last, starting with cleaning the data, then exploring and visualizing it, and finally interpreting the results. This process will help me understand which factors truly play a role in making a video trend.

Project Importance

Understanding what drives a video to trend on YouTube is important for a variety of stakeholders. For content creators, knowing the key factors that influence a video's success can guide decisions about upload timing, video categories, and engagement strategies. This knowledge helps maximize visibility and audience reach by aligning content with what tends to capture attention early on.

For digital marketers, identifying patterns in trending content offers valuable intelligence for campaign planning and audience targeting. By understanding which video features or engagement metrics drive virality, marketers can create more effective strategies to increase user engagement and optimize advertising efforts.

Additionally, researchers studying social media dynamics and digital media behavior can use the findings from this analysis to deepen their understanding of how engagement changes over time and which factors play the most critical role in driving online attention. With YouTube being one of the most influential platforms globally, the outcomes from this study will benefit not only content creators and marketers but also in understanding broader trends in digital media success.

The findings from this project can be useful for both research and real-world content strategy. Since YouTube's algorithm focuses on things like audience retention, engagement rate, and watch time, understanding how these factors connect to a video's features can help improve how content is created and shared. With so many videos uploaded every day, even a small improvement in understanding what makes a video trend can have a big impact on how creators, brands, and viewers engage with content online.

Background of the Problem

YouTube's trending list shows which videos are gaining attention the fastest, often within a short amount of time. These trends constantly change based on audience behavior, platform algorithms, and current events. The trending list reflects both what people are watching the most and what YouTube's system decides to promote to a wider audience. Views, likes, comments, and shares are usually seen as the main signs of how popular a video is. They show how people respond and can also influence which videos the platform promotes more. The time a video is uploaded and the type of content it belongs to can also make a difference. For example, videos posted during busier hours or in popular categories may gain more traction and reach the trending list faster.

Even though YouTube's algorithm also looks at things like watch time, audience retention, and click-through rates, the full process of how video trends is still not completely clear. Some studies

suggest that small details, like the number of tags or whether comments are turned on, can affect how visible or engaging a video becomes. My goal is to determine which factors have the greatest impact on whether a video trends in the United States and to explain how audience behavior and YouTube's algorithm work together to influence what viewers watch the most.

Dataset Description

The dataset used for this project is USvideos.csv, which is part of the "Trending YouTube Video Statistics" collection on Kaggle (J, 2017). It includes over 40,000 records of videos that trended in the United States from late 2017 onward. Each record contains details about the video, including its unique ID, the date posted and trending date, the title, the channel that uploaded it, and the category ID. The dataset also provides the publish date and time, engagement metrics such as views, likes, dislikes, and comment count, as well as text-based metadata like tags and video descriptions. Additional fields include the URL for the video's thumbnail, indicators for whether comments or ratings were disabled, and whether the video was later removed or became unavailable.

Before starting the analysis, I will clean the dataset to address missing values, inconsistencies in timestamps, and text formatting issues. Once the data is cleaned, I will explore it and conduct statistical analyses to examine the relationships between engagement metrics and content features. This will allow me to identify which factors have the strongest influence on a video's likelihood of trending on YouTube in the United States.

Data Profiling

The dataset used for this project is the YouTube Trending Videos Dataset from Kaggle. It includes daily records of the most popular videos on YouTube across multiple countries, but this project

focuses only on the U.S. dataset. The file, USvideos.csv, contains 40,949 rows and 16 columns, where each row represents a trending video and each column provides details such as the video title, channel name, publish date, views, likes, dislikes, comment count, and engagement settings. The only column with missing data is the description column, which has 570 empty entries. I will be using a random 15,000 rows with 193 missing video descriptions.

This dataset is useful for exploring how people engage with trending videos and what factors influence their popularity. It captures important engagement metrics such as views, likes, comments, and shares, which can help identify what draws attention and encourages audience interaction. These patterns are helpful for content creators, marketers, and researchers who want to understand what makes a video more likely to trend on YouTube.

The data was collected using the YouTube Data API, which records daily trending videos directly from YouTube's Trending Feed. It was organized and published on Kaggle by the user Datasnaek, who combined and cleaned the data from different regions. YouTube determines trending status based on engagement factors like views, shares, likes, and comments rather than overall lifetime views, making this dataset valuable for analyzing short-term popularity.

The dataset mainly covers the years 2017 to 2018, based on the published and trending dates. It includes a mix of eight string columns, five integer columns, and three boolean columns. Because this project focuses on exploration and pattern identification, the dataset is not divided into training or validation sets.

Column Name	Description	Data Type	Outliers	Nulls	Potential Quality Issues
video_id	Unique ID for each YouTube video	string	None	0	None
trending_date	Date video appeared on trending list (YY.DD.MM)	string	None	0	Trending date and publishing time are in different formats, and need conversion for analysis
title	Video Title	string	None	0	May have long titles or special characters
channel_title	Name of the channel	string	None	0	None
category_id	Category identifier for the video	integer	None	0	The same video_id may appear in multiple categories, and some videos may have no category

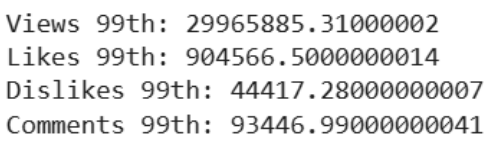
publish_time	Date and time video was published in ISO 8601 format (YYYY-MM-DDT00:00:00.000Z)	string	None	0	Needs conversion for analysis
tags	Keywords used to describe the video	string	None	0	Missing or “no tags” values, tags separated by |
views	Number of views	integer	Possible high numbers	0	High numbers, possible outliers
likes	Number of likes	integer	Possible high numbers	0	Possible outliers, 0 likes possible
dislikes	Number of dislikes	integer	Possible high numbers	0	Possible outliers, 0 dislikes possible
comment_count	Number of	integer	Possible	0	0 comments

	comments		high numbers		possible, may be affected if comments are disabled
thumbnail_link	URL to video thumbnail	string	None	0	Missing or broken URL, may have duplicates
comments_disabled	True/False if comments are disabled	boolean	None	0	None
ratings_disabled	True/False if likes/dislikes are disabled	boolean	None	0	None
video_error_or_removed	True/False if video is removed or error occurred	boolean	None	0	Affects trend analysis if True
description	Video description	string	None	193	Missing description for some videos, or may have long text

Most of the columns are complete, but the description column has 193 missing entries. Date columns such as trending_date and publish_time need to be converted to proper date formats before analysis. The tags column can have special characters or use the "|" symbol to separate multiple tags, which would need cleaning. Numeric columns, including views, likes, dislikes, and comment_count can have very high values that could affect analysis. The boolean columns show if comments or ratings are disabled, or if the video has errors, which could also impact results. This helps identify which parts of the data may require extra care before analysis.

When checking for outliers, I looked at the 99th percentile for the main numeric columns. The 99th percentile shows the point above which only the top 1% of data falls, which helps identify unusually high-performing or viral videos that could skew the results. Videos above this point are considered extreme compared to the rest. For views, the 99th percentile is around 29.9 million, likes to reach about 904,567, dislikes around 44,417, and comments reach about 93,447. These numbers show that only a small number of videos reach very high engagement levels, while most fall far below. This confirms that the dataset is strongly right-skewed, with a few viral videos heavily influencing the averages.

Figure 1: The 99th Percentile



Data Processing

To prepare the dataset for analysis, I will use Python and the pandas library to load and manipulate the U.S. YouTube Trending Videos dataset. To reduce processing time while keeping the data

representative, I will use pandas to randomly select a subset of 15,000 rows and set a fixed random seed for reproducibility.

I will use pandas datetime functions to convert the date columns into proper datetime formats. The trending_date column, currently in YY.DD.MM format, will be converted to a standard datetime, and the publish_time column, which contains timestamps in ISO 8601 format, will also be converted. I will verify the conversions by checking data types with df.info() and previewing the first few rows with df.head().

For data cleansing, I will use pandas string methods to clean the tags column by replacing the pipe character | with commas and removing placeholder values like [none]. Any missing values in the description column will be replaced with empty strings using pandas fillna().

I will also examine numeric columns such as views, likes, dislikes, and comment_count using pandas descriptive statistics (df.describe()) to identify unusually high values that could affect analysis. These steps, performed with Python, pandas, and NumPy, will ensure the dataset is structured, clean, and ready for exploration and analysis.

Descriptive Statistics

The main numeric variables analyzed in this project include views, likes, dislikes, and comment_count. These variables represent audience engagement and are the main factors in understanding why certain videos trend on YouTube. The dataset includes 15,000 randomly selected rows from the U.S. trending videos dataset.

The average number of views is approximately 2.36 million, though the standard deviation of about 7.4 million indicates large variation across videos. The median is 686,649 views, showing

that only a few videos reach high numbers, while most get fewer views. The lowest number of views is 549, and the highest is over 210 million.

The average number of likes is approximately 73,521, with a median of 18,450, showing that most videos receive moderate engagement, while a few gain millions of likes. Dislikes average 3,773, with most videos having fewer than 2,000. The number of comments also varies greatly, averaging 8,323 but with a standard deviation of over 36,000, meaning some videos attract much more audience interaction than others.

Overall, the data shows that engagement on trending videos is highly uneven. A small number of viral videos make the averages high, while most videos fall in a more typical range. Most videos fall within a normal range of engagement, but a small group of viral videos reach very high numbers, causing the data to be heavily right-skewed.

Figure 2: Descriptive Statistics

index	views	likes	dislikes	comment_count
count	15000.0	15000.0	15000.0	15000.0
mean	2355333.6132	73520.95026666667	3772.8305333333333	8323.080466666666
std	7404701.315142884	225079.37892813314	31240.856958985354	36030.98490943357
min	549.0	0.0	0.0	0.0
25%	248204.0	5500.75	204.0	626.0
50%	686649.0	18450.5	634.5	1870.5
75%	1823599.75	55445.75	1915.5	5787.5
max	210338856.0	5613827.0	1643059.0	1321281.0

Data Visualization

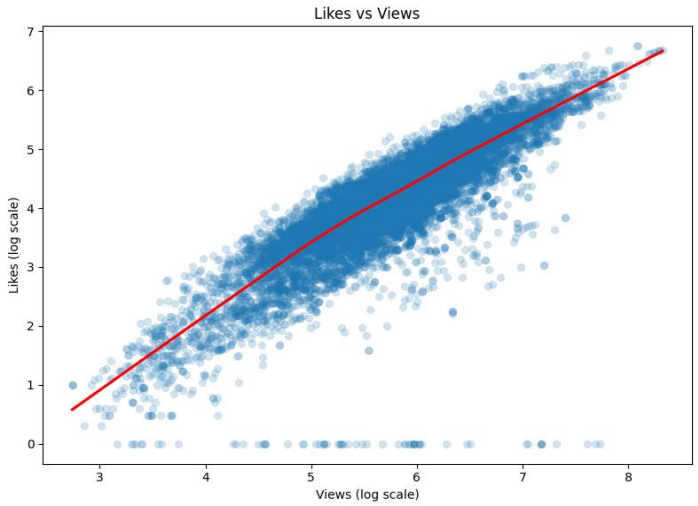
I plan on using a scatter plot to examine the relationship between likes and views for my videos. A scatter plot represents each video as a point on the graph, letting me observe whether videos with higher views also tend to receive more likes and to identify any unusual patterns. I am using a logarithmic scale because the numbers can be extremely large, and without it, most points would be clustered too tightly to distinguish clearly. I will also include a trend line to highlight the overall

relationship while keeping the individual points visible so that outliers can be identified. This method helps the detection of patterns in audience engagement and focuses on videos that perform differently than expected. Research demonstrates that scatter plots are effective for exploring correlations and understanding trends in complex datasets (Rensink, 2017).

Additionally, I plan on using a box plot to compare the division of views across different video categories. A box plot displays the median, interquartile range, and outliers, which is particularly useful because some videos have extremely high view counts. I am applying a logarithmic scale for the views to see that the wide range of values does not change the visualization. This visualization allows me to observe which categories generally receive higher views and which categories show greater irregularity. Box plots are valuable for demonstrating the differences between groups while showing extreme values, providing patterns that may influence the engagement. The use of box plots is an effective technique for comparing distributions and analyzing trends in grouped data (Hu, 2020).

Visualization 1:

Figure 3: Scatter Plot



I used a scatter plot to look at the relationship between likes and views for the videos in my dataset. The x-axis showed the number of views, and the y-axis showed the number of likes each video received. Each point on the graph represented a single video, so I could see if videos with more views also got more likes. I used a logarithmic scale on both axes because the numbers were very different, and without it, most of the points would have been too close together to see any pattern. I also added a trend line to show the general relationship while keeping all the points visible, so I could notice any unusual cases. The visualization made it easier to see how engagement varied across videos and which videos performed differently from most.

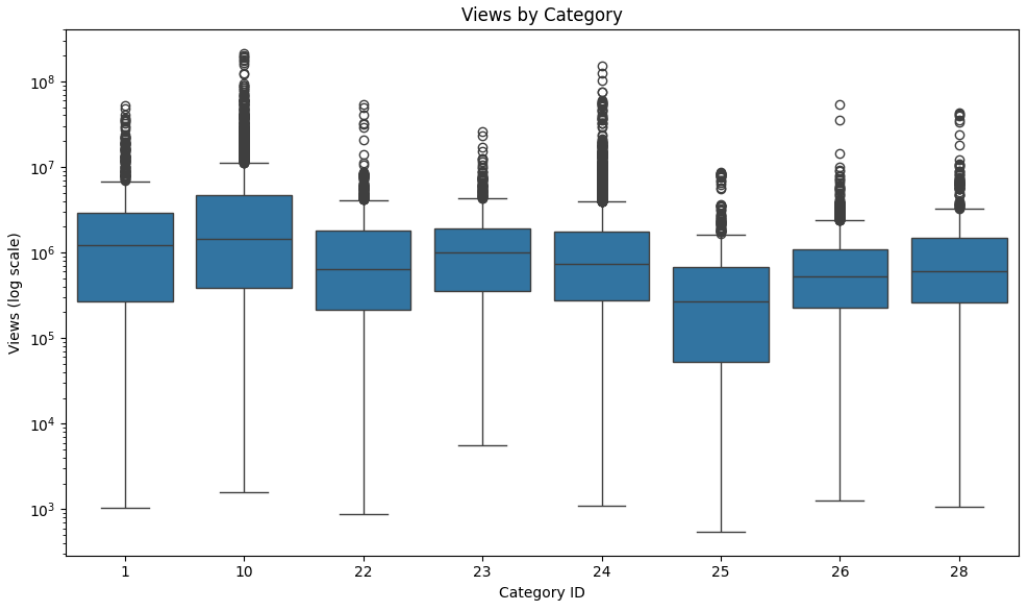
The scatter plot showed that videos with higher views usually had more likes, which makes sense because more people were watching them. Some videos had a lot of likes even though they did not have as many views, suggesting that certain types of content got people to engage more. Most videos were in the middle range of views and likes, while only a few went viral and reached very high numbers. The graph also showed groups of videos with similar patterns of views and likes, which could be related to the type of content or the time they were posted. Some videos with many views had fewer likes than expected, which might mean people watched but didn't react as much. Overall, it showed how likes and views were connected, but also that not all videos followed the same trend.

The scatter plot also showed some unusual cases. A few videos with low views still had a high number of likes, meaning they had smaller but very active audiences. On the other hand, some videos with many views had relatively low likes, suggesting people watched without interacting. The trend line confirmed that, in general, videos with more views got more likes, but the spread of points showed that engagement could vary a lot. Looking at these differences helped me understand how audience interaction worked, and which types of videos might do better on

YouTube. Overall, the graph gave a clear picture of the relationship between views and likes across all the videos in the dataset.

Visualization 2:

Figure 4: Box Plot



I used a box plot to compare the number of views across different video categories. The x-axis showed the categories, and the y-axis showed the number of views for each video. I used a logarithmic scale on the y-axis because some videos had really high view counts, and without it, most points would have been too close together to see. The box plot showed the median views, the range of most videos, and any outliers that stood out. Each category had its own box, which made it easier to see how videos performed compared to other types of content. This visualization let me see both general trends and videos that performed very differently from the rest.

The box plot showed that some categories usually had higher views than others, meaning videos in those groups tended to reach more people. Most videos in other categories had moderate views,

with fewer extreme cases. There were also a few videos in every category that got high views, showing that some videos went viral even if most stayed around average. Looking at the median views made it clear which types of content usually got more attention. The boxes and ranges also showed that performance could vary a lot within a category. Overall, it helped me see patterns in how different kinds of content attracted viewers.

The box plot also showed that even categories with lower median views could have videos that did well. A few videos in these groups got much higher views than most others, meaning things like topic, timing, or audience interest mattered. Some categories had videos that were consistent in views, while others had a wider spread, so performance wasn't always predictable. By looking at the boxes and outliers, I could tell which categories usually reached more viewers and which sometimes produced viral hits. This showed that just because a category usually had lower views, it did not mean a video couldn't get very popular. Overall, the box plot gave a clear picture of how the video category related to views while showing both trends and exceptions.

Data Modeling

To understand what makes a YouTube video trend, three different modeling techniques were used, each giving a different perspective on the data. The models applied are Linear Regression, Random Forest Classification, and K-Means Clustering. Linear Regression is useful for exploring how one variable changes in relation to others and is often used first because it is simple and easy to understand (IBM, 2021). The model fits a straight line through the data to show the overall trend, which makes it clear how each feature affects the outcome. The coefficients give a clear sense of the strength and direction of relationships, making it easier to see which factors matter most. This method is widely used because it allows basic predictions in a straightforward way and helps

explain patterns in the data (Roustaei, 2024). Even though it works best when relationships are mostly linear, it still provides valuable information about how features interact in a dataset.

Linear Regression works best when the data meets certain assumptions, such as having a consistent spread of errors and a mostly linear relationship between the predictors and the outcome (Stojiljković, 2019). When these conditions are met, the model can produce predictions that are fairly reliable and help explain why certain results occur. It is easy to run in Python, and outputs like coefficients, residuals, and performance scores can be quickly checked (IBM, 2021). Many projects start with this method because it runs fast and provides a clear baseline to compare with more advanced models. The results are also straightforward to explain, which is useful when sharing findings with people who are not familiar with technical details. If the data does not follow a linear pattern or breaks these assumptions, the predictions may be less accurate, and other techniques may give better results (Roustaei, 2024).

Classification models are used to predict a category or group for each data item rather than a continuous value. For example, a classification model can determine whether a video is likely to become viral based on its features. Random Forest is a strong classification method because it builds many decision trees and combines their results to determine the final prediction (Belcic, 2024). This reduces the chance that errors from one tree will dominate the outcome. The model works well when there are many features or the dataset is complex, because it chooses different subsets of data and features for each tree, which reduces overfitting (Pan et al., 2025). By combining many small decision trees into a larger model, Random Forest produces predictions that are reliable even when the data is inconsistent or noisy.

Classification methods like Random Forest are flexible and can handle different types of data

including numeric, categorical, or mixed types (Abdullah and Eid, 2023). When applied to a dataset of trending videos, features like views, likes, comments, category, and tags can be used to predict whether a video will become highly viral. The model provides information about which features have the strongest effect on the outcome, making the results easier to understand (Belcic, 2024). These models do not require strict assumptions such as linearity or normally distributed data, so they perform well even when the dataset is uneven or messy. This makes them especially useful for social media and engagement data where patterns can be irregular. Since Random Forest is made of multiple decision trees, its predictions are more consistent than those from a single tree (Pan et al., 2025).

K-Means clustering is an unsupervised method that groups data points into clusters based on similarity rather than predicting a particular outcome. The algorithm begins by selecting some clusters, k, and assigning each data point to the closest centroid. The centroids are then recalculated based on the points in each cluster, and this process repeats until the groups stabilize (Kavlakoglu & Winland, 2024). This method helps find patterns in data, such as grouping videos with similar engagement characteristics. One advantage is that it is simple and fast, allowing large datasets to be processed efficiently (Arvai, 2020). By keeping points within each group similar and making groups different from each other, K-Means can show patterns that might not be obvious from just looking at the data (Zubair et al., 2022).

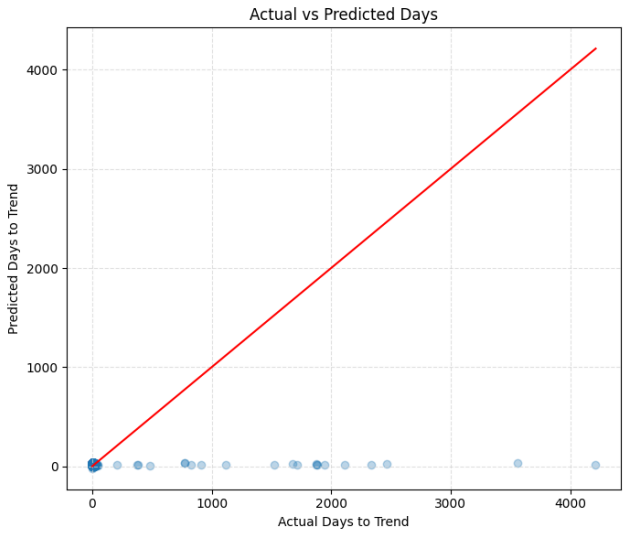
K-Means makes it possible to see clear groupings in the data, for example, which videos are highly viral, and which are less successful. Because it does not require labeled outcomes, natural patterns in the dataset can be examined without assumptions about success. Using Python and libraries like scikit-learn, different numbers of clusters can be tested to determine which grouping makes the most sense (Arvai, 2020). The method does have some limitations, such as assuming clusters are

roughly the same size and shape, which can make it less accurate if the data has unusual distributions or extreme values (Kavlakoglu & Winland, 2024). Choosing the right number of clusters is important because too few or too many groups can hide real patterns or create confusion (Zubair et al., 2022). Overall, K-Means is a useful method for exploring data and understanding different engagement behaviors.

Model 1: Linear Regression

The Linear Regression model was used to predict how many days pass between when a video is published and when it appears on the trending list.

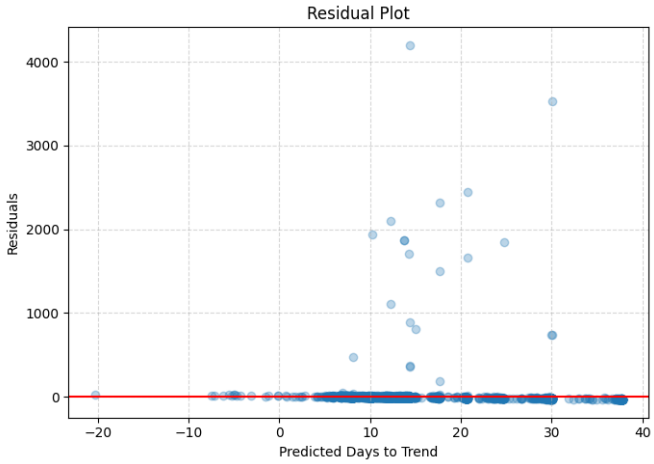
Figure 5: Actual vs Predicted Days Plot



The actual versus predicted graph shows that most videos cluster very close to zero, meaning they trend very quickly. This tells us the model does a good job predicting the speed of the videos that become trending the fastest, which are the most common and important cases. There are some extreme outliers where videos take much longer to trend, and the model does not handle these well. The residual plot shown in *Figure 6* confirms this, showing small errors for videos trending quickly

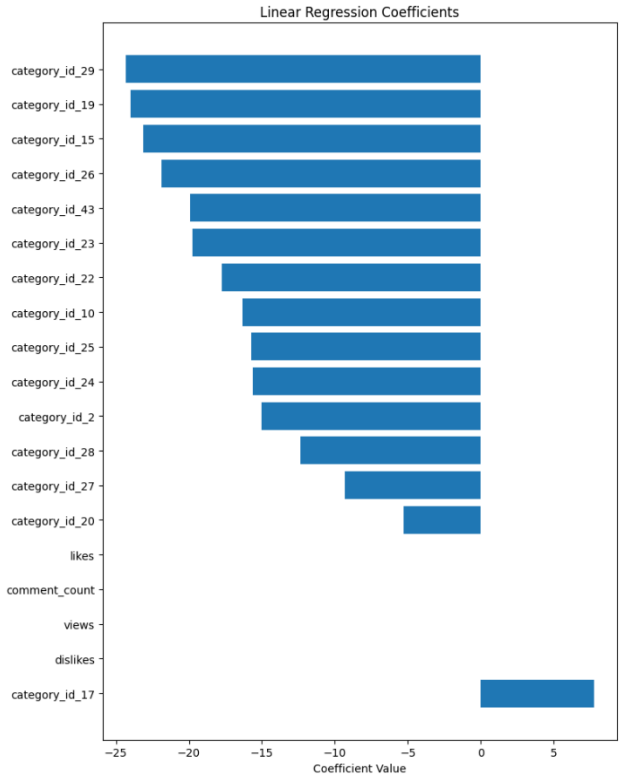
and large errors for long-term outliers. Overall, the model works well for short-term predictions but is not suitable for videos that take an unusually long time to appear on the trending list.

Figure 6: Residual Plot



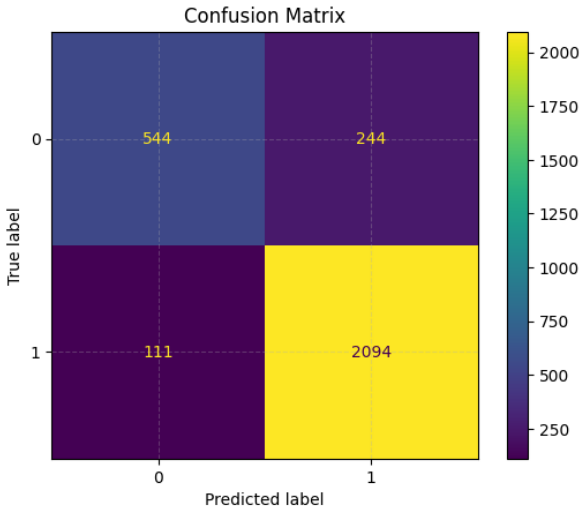
Looking at the model coefficients in *Figure 7*, it gives a clear picture of what affects how fast videos trend. Videos in categories like Nonprofits and Activism, Travel and Events, and News and Politics trend the fastest, while Sports videos tend to take longer. Engagement metrics like views, likes, dislikes, and comment counts have very little effect on the speed of trending. This shows that the type of video matters much more than its early engagement in determining how quickly it reaches the trending list. Even though the model struggles with extreme cases, it does a good job explaining the speed for most videos. This makes it useful for understanding which factors make videos trend quickly.

Figure 7: Linear Regression Coefficients



Model 2: Random Forest Classification

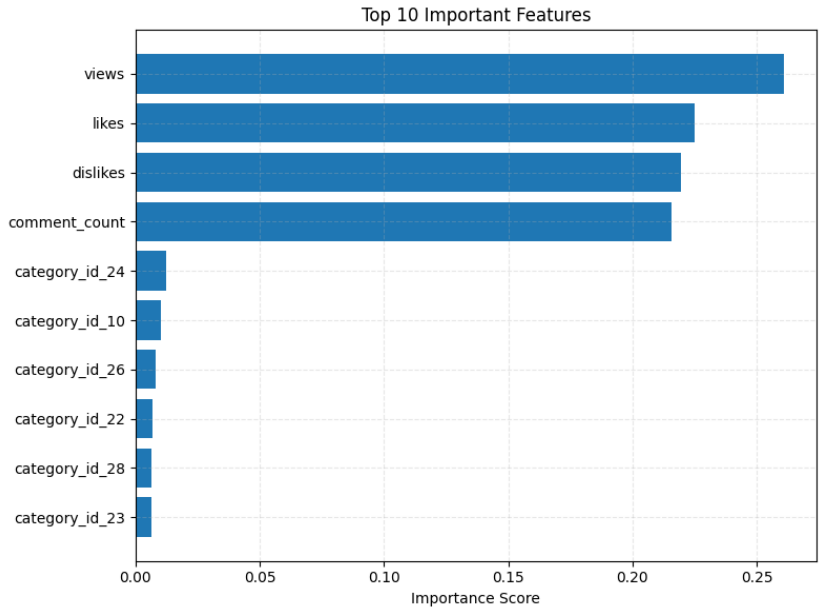
Figure 8: Confusion Matrix



The second model used was a Random Forest Classifier to predict whether a video would become

highly viral. The model's accuracy was 88.1 percent, which means that nearly nine out of ten videos were correctly classified as viral or not viral. Its F1 score is 0.9219, showing that the model does well in finding viral videos while avoiding many false alarms. The confusion matrix seen in *Figure 8* shows 2094 true positives, meaning the model correctly identified most of the viral videos, and only 111 false negatives, so very few viral videos were missed. There were 544 true negatives and 244 false positives, which means the model also did well with non-viral videos, though a few were mistakenly labeled as viral. Overall, these results show that the Random Forest Classifier is very reliable for predicting which videos are likely to become viral. It correctly identifies most viral videos while keeping the number of missed or misclassified cases low, making it a useful tool for understanding patterns of video success.

Figure 9: Top 10 Important Features

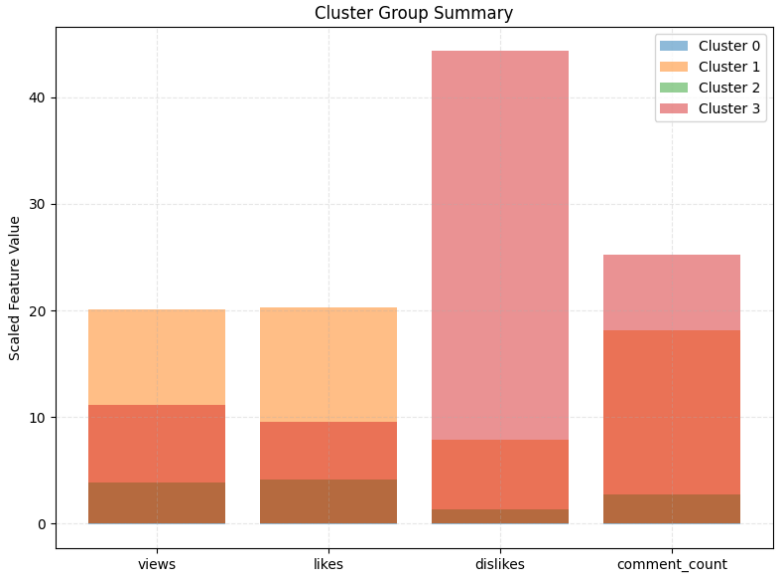


The feature importance chart shows that the most important factors for predicting virality are the engagement metrics, including views, likes, dislikes, and comment count, which are grouped at the top of the scale. The video category features, which mattered more in the Linear Regression

model for predicting days to trend, have very little impact on whether a video becomes highly viral. Category affects how quickly a video trends, while engagement metrics drive how popular it becomes. The Random Forest model shows that audience interaction is the key factor for highly viral content. The chart makes it easy to see which variables carry the most weight and helps explain what drives a video's success. Paying attention to audience response is much more effective than focusing on the video type when predicting virality.

Model 3: K-Means Clustering

Figure 10: Cluster Group Sumamry

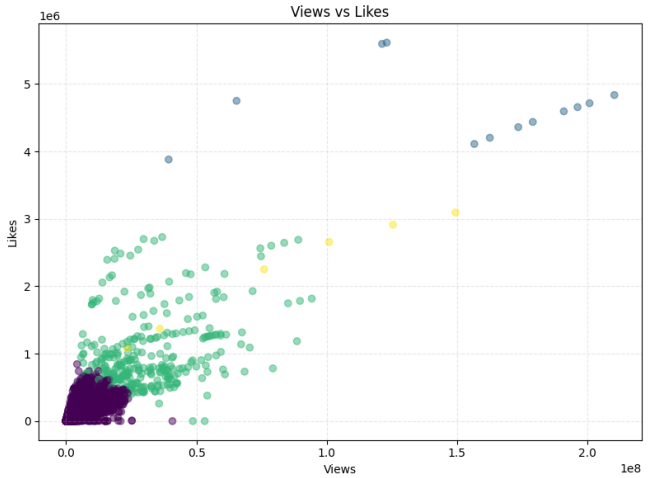


In final part of the analysis, I used K-Means clustering with four clusters on the standardized engagement features to group trending videos based on audience response. The centroid analysis separated four clear types of videos according to their average views, likes, dislikes, and comments. Most videos fell into Cluster 0, which represented typical trending content, and Cluster 2, which included niche or low-engagement videos. Cluster 1 stood out as the successful and positively received content type, with high views and likes but low dislikes and comments. This

shows that not all trending videos are the same, and audience reactions differ widely depending on the type of content. The clustering provides a clear way to describe how engagement is structured across the dataset.

The most notable finding came from Cluster 3, which represents controversial or opposing content. This group had the highest average dislikes and a very high number of comments, showing that videos with strong negative reactions also generate intense discussion. The scatterplot of views versus likes supported this, placing these videos as unusual outliers with very high activity compared to most others. This pattern recommends that videos can become successful in two ways, through consistently positive engagement like Cluster 1, or through highly active, emotionally charged engagement like Cluster 3. The K-Means model successfully revealed these underlying behavioral patterns and offers a useful framework for understanding different forms of audience interaction. This analysis can help guide content strategies by showing how engagement varies across different types of trending videos.

Figure 11: Views vs Likes



Champion Model:

For this project, the Random Forest Classifier is the Champion Model. It does the best predicting which videos will become highly viral, with an accuracy of 88.1 percent and an F1 score over 0.92. These numbers show that the model rarely misses videos that will succeed, making it very reliable for understanding which content will perform well. The results also make it clear that engagement metrics like views, likes, dislikes, and comment count are the main factors that drive a video's success. The K-Means Clustering model also did well, but it is more useful for describing patterns in the data rather than predicting outcomes. Together, the two models work well, with Random Forest showing which videos will go viral and K-Means explaining the type of engagement they get.

K-Means remains a strong secondary Champion Model because it gives important context to the classification results. By splitting the videos into four clusters, K-Means shows the different ways content can get attention. Some videos succeed through steady positive engagement, while others draw a lot of comments and reactions because they are polarizing. This explains why some videos quickly become popular, and others generate discussion in a different way. K-Means does not predict which videos will go viral, but it clearly shows how audiences respond to different types of content. When combined with the Random Forest model, it gives a complete picture of what drives success on YouTube. Using both models together makes it easier to understand both the likelihood of a video going viral and the type of engagement it will receive.

Final Results

Findings:

The Random Forest Classifier performed very well at predicting which videos would trend quickly. The model achieved an accuracy of 88.1 percent and an F1 score of over 0.92, showing it

rarely misses videos that are likely to succeed. These results confirm that the model is a reliable tool for identifying high-potential content early. The analysis focused on engagement metrics, including views, likes, dislikes, and comment count, which all had a strong influence on predictions. Among these, views had the largest effect, followed closely by likes and comment count, while dislikes had slightly less effect. The model shows that early audience behavior is very useful for understanding which videos will trend within the first week.

Looking deeper into the Random Forest, the importance of each feature shows which metrics affect the likelihood of trending. Views are the most influential, followed by likes and comment count, while dislikes provide additional context about audience reaction. This table organizes each metric's contribution, making it easy to see the differences between them. These results make it clear that paying attention to engagement measures helps explain why some videos perform better than others. Feature importance shows the relative weight of each metric in predicting early trending videos. The information in the table confirms that both views and interactions are necessary to understand which content succeeds.

Figure 12: Table 1

	Feature	Importance
0	views	0.267708
1	likes	0.256889
3	comment_count	0.240738
2	dislikes	0.234666

Looking at the dataset by category gives context on how videos perform across different types of content. Videos in categories like 10 and 1 have the highest average views, likes, and comments, which means they attract the most attention. Categories with lower averages, like 25 or 27, show smaller audience interaction and reach. Examining these numbers explains why some types of

content naturally gain more traction. The category table shows each category's mean values for key metrics, making it easy to compare them. This information works well alongside the Random Forest results to show which categories tend to trend more.

Figure 13: Table 2

	Category	Mean Views	Mean Likes	Mean Dislikes	Mean Comments
2	10	6.201003e+06	218918.199011	7907.757726	19359.764524
0	1	3.106250e+06	70787.836247	2590.681450	7627.744136
14	29	2.963884e+06	259923.614035	58076.859649	84364.859649
6	20	2.620831e+06	84502.183599	11241.696450	18042.488372
9	24	2.067883e+06	53243.325070	4314.297772	7383.229426
4	17	2.025969e+06	45363.942502	2361.339006	5148.185373
7	22	1.531835e+06	58135.825234	3173.800935	7719.013084
8	23	1.480308e+06	62582.223315	2091.521840	6521.718831
13	28	1.452627e+06	34374.276551	1894.378176	4993.721783
1	2	1.355965e+06	11056.395833	632.838542	2042.830729
11	26	9.837301e+05	39286.076942	1320.284370	5583.586589
15	43	9.035273e+05	18993.666667	429.964912	1668.719298
5	19	8.546196e+05	12030.462687	846.833333	2267.440299
3	15	8.311435e+05	21055.110870	573.238043	2892.070652
12	27	7.129408e+05	29745.031401	816.408213	3286.378019
10	25	5.925877e+05	7298.364696	1680.759550	2428.400885

K-Means Clustering reveals how different types of audiences respond to videos. The clusters group videos by the way people engage with them, ranging from steady positive attention to content that sparks lots of comments and mixed reactions. One cluster stands out for having very high comment counts and dislikes, showing that polarizing videos can generate a lot of interaction even if the responses aren't all positive. Other clusters include mainstream hits that get consistent engagement and videos with low engagement that attract little attention. Looking at these patterns helps explain why some videos quickly gain popularity while others have smaller but active audiences. This analysis gives a clearer picture of audience behavior that goes beyond just predicting which videos will trend.

When the Random Forest results are combined with K-Means clustering, it gives a fuller view of

video performance. Random Forest predicts which videos are likely to trend early, while K-Means shows how audiences actually interact with the content. Both the likelihood of trending and the type of engagement matter for understanding success. Controversial videos can draw a lot of attention, even if reactions are mixed. Mainstream videos succeed through steady engagement and broad appeal. Together, these two approaches make it easier to see what contributes to early success and the ways viewers respond to different types of content.

Overall, the analysis shows that video success on YouTube comes from a combination of factors. The Random Forest model identifies which videos are likely to trend early based on engagement metrics, while K-Means explains how audiences interact with content in different ways. Fast-trending videos often belong to categories like Music or Entertainment, which benefit from built-in audience attention. Some videos gain traction through steady positive engagement, and others draw strong reactions because they are polarizing. Looking at both prediction and behavior together provides a fuller picture of success than either model could alone. This approach makes it clear that understanding both the likelihood of trending and the type of engagement a video receives is key to analyzing performance.

Review of Completion:

Looking back at the project, it went well and met the goals I set. The Random Forest model worked as expected, with high accuracy and F1 scores, showing which videos are likely to trend. K-Means clustering helped show the different ways audiences react to content, separating videos into groups based on engagement. Cleaning the data and calculating the time-to-trend made sure the analysis was reliable and consistent. The tables make it easy to see how the models perform and how audiences interact with videos. Each step, from preparing the data to building the models, helped

me understand what makes a video successful on YouTube.

There were a few challenges along the way that needed attention. Creating the target variable for trending videos took some trial and error, and fixing timestamp formats was important to avoid errors. The feature importance from Random Forest and the cluster averages from K-Means made it clear which metrics matter most without overcomplicating the results. The models aren't perfect, but they show patterns that match what I expected from the data. Keeping the code organized helped me follow each step and make adjustments as needed. This project shows a clear process and gives a solid understanding of what affects YouTube video performance.

Potential Data Privacy and Data Security Issues:

The dataset for this project includes publicly available YouTube video information such as views, likes, dislikes, comment counts, categories, and publish times. Since the data comes from public sources, individual users are not directly identifiable, but patterns in engagement and content could still reveal information about creators or audience behavior. Care should be taken when using this data to avoid connecting it with other datasets that might compromise privacy. The models built in this project predict which videos are likely to trend and describe audience interaction, so their results could potentially influence decisions about content promotion or highlight trends in viewer engagement. Using these results responsibly means considering how the predictions are applied and avoiding misuse of sensitive patterns. Proper handling and anonymization of data ensure that insights are drawn safely without exposing personal information.

At the same time, even publicly available data carries some privacy and security considerations. Sharing model results or datasets without precautions could allow unintended identification of channels or reveal strategies that content creators use to gain engagement. The analysis focuses on

trends and patterns rather than individuals, but anyone applying these models should be mindful of ethical implications. Limiting access to the raw data and results, and using aggregated metrics, when possible, helps reduce potential risks. Being aware of these concerns helps maintain responsible use of the dataset and ensures that insights benefit understanding of video performance without compromising privacy.

Recommendations for Future Analysis:

For future analysis, it could be helpful to include more information about the videos themselves, like their length, description keywords, thumbnails, or how often the channel posts content. Adding these features might make it easier to predict which videos trend quickly and which maintain long-term attention. Looking more closely at timing, such as weekly or seasonal trends, could also show patterns in audience behavior. Trying different models, like gradient boosting or neural networks, could show whether more complex approaches improve predictions compared to Random Forest. It might also be useful to break the data into smaller groups, like by category or region, to see if some types of videos follow different trends. Updating the data over time would give a clearer picture of how viewer habits change.

Another important consideration is that the current models are solving a bit different problems, with Random Forest predicting trending videos and K-Means describing engagement patterns. If the goal is to compare champion models, focusing on similar tasks would make comparisons fairer. Examining not just how many likes, dislikes, or comments a video gets, but what the comments say, could provide extra context for engagement. Clustering could be refined with additional features to show more detailed audience behavior. Comparing YouTube trends to activity on other social media platforms might reveal how wider attention affects success. Including these

adjustments would make future analysis more consistent and give a deeper understanding of why some videos succeed and how audiences interact with content.

Project Status

Project Milestones:

Milestone	**Status**	**Notes**
Data Collection & Project Scope	Completed	Downloaded US dataset from Kaggle. Identified the importance of the project and the dataset.
Data Profiling & Preparation	Completed	Converted dates, cleaned tags, handled missing descriptions, checked missing values, and basic stats.
Project Summary Presentation	Completed	Prepared slides to summarize the dataset, project goals, and first findings.
Data Analysis & Exploration	In Progress	Exploring relationships between views, likes, comments, and categories using Python.
Data Visualization	Completed	Created the first 2 visualizations: scatter plot and box plot.
Data Modeling	Completed	Created 3 models and visualizations to

		compare and find a champion model.
Presentation 2: Final Project	Completed	Presented the findings and recommendations in a presentation form.
Final Results	Completed	Reviewed the project and finalized the main findings and possible recommendations for future analysis.

Completion History:

Date	**Task**	**Status**
10/29/2025	Downloaded dataset	Loaded a dataset of 40,000 rows successfully, identified the goal of the project, and described the dataset.
11/4/2025	Used 15,000 rows, identified possible issues in the dataset.	Chosen 15,000 rows for faster processing.
11/4/2025	Converted the date columns, cleaned tags, and the description.	Trending_date and publish_date are in datetime format now, tags reformatted, and missing descriptions filled.

11/6/2025	Project Summary Presentation	Completed slides summarizing the dataset, project goals, and original findings.
11/11/2025	Data Visualization	Created scatter plots and box plots to explore relationships between views, likes, comments, and categories.
11/27/2025	Data Modeling	Used 3 models and created 7 visualizations. Chose the champion model.
12/9/2025	Final Results	Analyzed and finalized the project and its findings.

Lessons Learned:

- I learned that converting date columns to proper datetime formats is important for comparing trending dates and publication dates.
- Writing down information about each column, including data type, missing values, and potential issues, helped me plan the analysis better.
- I saw that numeric columns like views, likes, dislikes, and comment counts can have very high values that affect analysis and need to be checked.
- Cleaning text columns like tags and descriptions takes time because of multiple values

and placeholder text.

- Choosing random 15,000 rows made it faster to work with the data while still keeping it representative. Setting a random seed makes the sample reproducible so I can get the same subset each time.
- Doing data profiling and cleaning first helped me see possible challenges and guided the rest of the project.
- Creating scatter plots and box plots helped to quickly see trends and unusual cases that were not obvious at the beginning.
- Outliers revealed interesting viral videos, showing that the most viewed or liked content does not always follow typical patterns.
- Using a logarithmic scale for highly skewed data allowed me to see patterns that would have been hidden otherwise.
- Each model has a specific role, with Random Forest performing well for predicting which videos will go viral and K-Means helping to understand patterns in audience engagement.
- Combining prediction and description provides a clearer understanding of what drives success on YouTube than relying on just one type of analysis.
- Selecting appropriate features for each model is important. Including irrelevant or overlapping variables can make interpretation harder and affect accuracy.
- I realized that exploring data visually before modeling helps spot trends or irregularities that might otherwise be missed, guiding better decisions for analysis and model design.

References

Abdullah, S., & Eid, H. F. (2023). Utilizing random forest algorithm for early detection of academic underperformance in open learning environments. *PeerJ*, *9*, e1708–e1708. https://doi.org/10.7717/peerj-cs.1708

Arvai, K. (2020). *K-Means Clustering in Python: A Practical Guide*. Realpython.com. https://realpython.com/k-means-clustering-python/

Belcic, I. (2024, October 15). *Classification in Machine Learning*. Ibm.com. https://www.ibm.com/think/topics/classification-machine-learning

Faverio, M., & Sidoti, O. (2024, December 12). *Teens, Social Media and Technology 2024*. Pew Research Center; Pew Research Center. https://www.pewresearch.org/internet/2024/12/12/teens-social-media-and-technology-2024/

Hu, K. (2020). Become Competent within One Day in Generating Boxplots and Violin Plots for a Novice without Prior R Experience. *Methods and Protocols*, *3*(4). https://doi.org/10.3390/mps3040064

IBM. (2021, August 18). *Linear Regression*. Ibm.com; IBM. https://www.ibm.com/think/topics/linear-regression

J, M. (2017). *Trending YouTube Video Statistics*. Kaggle.com. https://www.kaggle.com/datasets/datasnaek/youtube-new?resource=download&select=USvideos.csv

Kavlakoglu, E., & Winland, V. (2024, June 26). *What is k-means clustering?* IBM.

https://www.ibm.com/think/topics/k-means-clustering

Oh, M., Maeng, K., & Shin, J. (2025). Which Factors Affect Online Video Views and Subscriptions? Reference-Dependent Consumer Preferences in the Social Media Market. *Journal of Theoretical and Applied Electronic Commerce Research*, *20*(3), 197–197. https://doi.org/10.3390/jtaer20030197

Pan, Q., Agarwal, S., Nissim, N., & Sabut, S. K. (2025). *Random Forest Classifier - an overview | ScienceDirect Topics*. Www.sciencedirect.com. https://www.sciencedirect.com/topics/computer-science/random-forest-classifier

Rensink, R. A. (2017). The nature of correlation perception in scatterplots. *Psychonomic Bulletin & Review*, *24*(3), 776–797. https://doi.org/10.3758/s13423-016-1174-7

Roustaei, N. (2024). Application and interpretation of linear-regression analysis. *Medical Hypothesis Discovery & Innovation in Ophthalmology*, *13*(3), 151–159. https://pmc.ncbi.nlm.nih.gov/articles/PMC11537238/

Stojiljković, M. (2019, April 15). *Linear Regression in Python*. Realpython.com; Real Python. https://realpython.com/linear-regression-in-python/

Yang, S., Brossard, D., Scheufele, D. A., & Xenos, M. A. (2022). The science of YouTube: What factors influence user engagement with online science videos? *Plos One*, *17*(5). https://doi.org/10.1371/journal.pone.0267697

Zubair, Md., Iqbal, MD. A., Shil, A., Chowdhury, M. J. M., Moni, M. A., & Sarker, I. H. (2022). An Improved K-means Clustering Algorithm Towards an Efficient Data-Driven Modeling. *Annals of Data Science*. https://doi.org/10.1007/s40745-022-00428-2