Bachelor of Science – Management & Computer Science

Department of Business & Management

Chair of Data Analysis for Business

What It Takes to Be Trending:
An Analysis on YouTube Trending Videos

Prof. Francesco Iafrate	Giuliano Oscar Stefanelli 243331
Supervisor	Candidate

Academic Year 2021/2022

Table of Contents

Abstract

YouTube is the second most popular social media platform and the leading website for video sharing. As such, over one billion videos are watched by its users on a daily basis. YouTube, however, out of all the content uploaded every day, selects a limited number of videos to feature on its Trending List. These so-called “trending videos” represent content that is getting attention, might interest a wide range of viewers, and has the potential to become popular.

Despite the popularity of YouTube and the importance of the Trending List, trending videos have not been analyzed comprehensively. This study first briefly introduces the platform and some of its most important purposes. This thesis, then, aims to gain a deeper understanding of the nature of such videos. The research is based on a dataset of basics statistics of 40,000 trending videos collected over a period of circa 200 days. First, the research focused on investigating the features in the dataset to summarize some common characteristics among trending videos. Next, both K-means and hierarchical clustering were performed on a subset of the original dataset. Finally, the study conducted a regression analysis with the aim to predict the view count of videos. Some multivariate linear regression models were built, as well as a model based on the random forest algorithm. To evaluate the models and choose the best one, the test mean squared error was used.

CHAPTER 1

1.1 YouTube

1.1.1 The platform

YouTube was founded on February 14, 2005, by Steve Chen, Chad Hurley, and Jawed Karim. The inspiration behind the website is unclear, as even the founders have told different stories. Karim has said the idea first came from the need to find a clip of the 2004 Super Bowl halftime show [1]. Hurley and Chen have stated that YouTube was originally meant to be used as a video-version of an online dating website [2]. However, according to a story most often found in the media, the idea for this user-generated video platform came as the solution to a problem the three founders encountered when taking pictures and videos at a dinner party: how were they going to share such memories with each other [3]? Indeed, websites like Flickr[1] made sharing digital photographs easy and fast, but videos, on the other hand, could not have been shared with the same ease. The founders then decided to overcome this discrepancy by creating a video-sharing website. Their mission was simple: to build a user-friendly interface that even people lacking computer skills could use to share their memories. A little more than three months after the ideation of YouTube, on April 24, 2005, founder Jawed Karim uploaded the very first video on the website showing his visit to the San Diego Zoo in a 19-second clip [3]. His video, entitled *Me at the zoo*, now has over 200 million views and 11 million likes.[2] Seven months later, the three founders officially launched the video-sharing platform in November 2005, receiving US$11.5 billion to assist their project [3]. In 2006, YouTube achieved outstanding results: over 65,000 videos uploaded every day and over 100 million views reached daily [3]. Given the astonishing achievements, the American search-engine company Google bought YouTube for US$1.65 billion in November 2006, after its failed attempt to launch its own video-sharing platform known as Google Videos [3].

YouTube, however, is not the only video-sharing platform on the market. Indeed, it is not even the first website of this kind to be launched. Since user-generated video content is becoming the main source of entertainment among young generations, YouTube has a large pool of competitors. According to whatcompetitors.com, YouTube's biggest threats come from platforms like Vimeo, Dailymotion, Twitch, IGTV, and TikTok [4]. Vimeo, created by filmmakers Jake Lodwick and Zach Klein in 2004, counts 230

[1] Flickr: photo-sharing platform used to upload photos launched on February 10, 2004.

[2] https://www.youtube.com/watch?v=jNQXAC9IVRw&ab_channel=jawed

million monthly users, 715 million monthly views, and 350,000 new videos uploaded daily, and generated US$160 million in annual revenues in 2018 [5]. Dailymotion, created by Benjamin Bejbaum and Oliver Poitrey in 2005, counts 250 million monthly users and 3.5 billion monthly views, and generates circa US$110 million in annual revenues [6]. The live-streaming platform Twitch, launched by Justin.tv in 2011, has 140 million monthly users, 71 million hours of content viewed daily, and generates US$230 million in revenues [7]. The Meta-owned social network, Instagram, launched its video service in 2018 known as IGTV, which generates revenues of US$20 million and counts over 7 billion downloads worldwide of its smartphone application [8]. Last but definitely not least, the Chinese platform TikTok, created in 2016, counts 1.2 billion monthly users and generates US$4.6 billion in revenues [9].

Despite the huge numbers of its competitors, YouTube remains the most popular and most used video-sharing platform. According to globalmediainsight.com, YouTube counts 2.6 billion monthly users, 1 billion hours of video watched daily, and generates US$19.7 billion in revenues. The website ranks second as the most popular social media after Facebook and is considered the second-largest search engine [10].

1.1.2 The importance of YouTube today

As the second most popular social media, YouTube has revolutionized everything from entertainment to education, created new possibilities for creative individuals who want to make their online presence their career, and provided an innovative platform for advertisements and influencer marketing.

According to a survey conducted by Insider in 2019, YouTube stars make their way up in the list of the most influential people among American teenagers: “now, they [teens] are looking up to their favorite people to follow on YouTube” [11]. In this ranking, we can find some of the most famous YouTube content creators, also known as YouTubers, like David Dobrik (18.3M subscribers) and Emma Chamberlain (11.3M subscribers) competing with former U.S. President Donald Trump and reality-show star Kylie Jenner. These results do not come as a surprise, as today’s adolescents were born and are growing up in a world dominated by social media and online personalities. The reason why YouTubers now have an impressive grip and influence on the young audience is because they share content that viewers can relate to, thus creating a community. Many creators are also considered to be more engaging and appealing than traditional celebrities for the way they interact with their followers. For example, creators, often being the same age of their audience and sharing the same niche interests, can establish an almost one-to-one relationship with their subscribers through the comments section of the video.

When YouTube launched the Partner Program in 2007, it made possible for channels meeting certain requirements to put ads on their videos and earn money from doing so. Through the monetization of their content, YouTubers make sharing videos online their full-time job: 23-year-old Jimmy Donaldson (better known by his online alias ‘MrBeast’), for example, earned US$54 million in 2021 [12]. However, for the majority of these online creators, their career is not based solely on their YouTube channel. Indeed, after dominating the YouTube platform, YouTubers mark their presence in all kinds of markets. Among the most successful instances, we can find the make-up brand *Jeffree Star Cosmetics* founded by Jeffree Star, smartphone application *Dispo* founded by David Dobrik, coffee company *Chamberlain Coffee* founded by Emma Chamberlain, and novel *Girl Online* written by Zoe Sugg.

Considering the above-mentioned examples, the results from a survey conducted by toy production firm *Lego* are long anticipated: most American children between the ages of eight and twelve would rather be YouTubers than astronauts [13].

YouTube, nonetheless, is not used for mere entertainment purposes. According to research performed for Pew Research Center, the Google-owned video-sharing website has become a source of news for many Americans. 26% of U.S. adults state that, even though it is not their primary news source, YouTube is an important means they use to stay informed. Not matter the actual origin, let it be an official YouTube channel associated with news organizations (like CNN or Fox News) or an independent channel (common user sharing facts), most YouTube news consumers do not see any issues with getting news from the site [14].

Another valuable purpose of this video-sharing platform is for education. YouTube, in fact, has partnered with over 300 Universities, including top schools like Massachusetts Institute of Technology and Stanford University, to bring free lectures and content available to all [15]. Notable example is the Harvard University’s CS50 Introduction to Computer Science class streamed on its channel, that now has over 72 million views combined.

Over the years, the YouTube platform has become a perfect place to promote and boost businesses and products through advertising and influencer marketing. As a matter of fact, 90% of shoppers have discovered a new service or product through YouTube (Google), almost 80% of marketers say that YouTube is the most effective video marketing platform (Business2Community), and 90% of the ads on YouTube increase brand recall (Statista) [16].

The acquisition of YouTube by Google brought many changes to the platform, one these being the Partner Program seen previously. Since then, Google updated the program, making YouTube advertising an extremely profitable investment. Placing an ad on the video-sharing website is obviously worthwhile given its billions of monthly users, however the investment on ads would be even more valuable if users were targeted based on their interests. Among the changes brought, possibly the biggest announcement was that advertisers would be able to target users based on their Google search history, in addition to their YouTube viewing behaviors. If the ad content is closely related to a search users have made, marketers are now able to target ads at them [17].

Celebrity endorsement has been a highly common practice among marketers for years: it is a form of advertising which uses a celebrity to promote a product, business, or raise awareness.[3] Social media influencers, however, are slowly replacing celebrities and, consequently, marketers are adapting to this major shift. The marketing strategy consisting of a collaboration between a brand and an influencer, the so-called influencer marketing, has grown to US$13.8 billion in 2021. Moreover, businesses are making US$5.78 return on investment for every dollar spent on influencer marketing [18]. Among all kinds of influencers, YouTubers are remarkably influential because they come across as authentic and relatable, as previously discussed. As influencer marketing is based on the "economy of trust", YouTubers' ability to sell or recommend a product or service to their followers heavily relies on their story-telling skills. As subscribers watch more and more videos of their favorite video creators, they become invested in their lives and start to trust them. There are several examples of YouTube influencers becoming a crucial part of a brand's marketing strategy, an excellent one is the Gymshark case [16]. The workout clothing brand partnered with fitness creators, like Steve Cook (1.31M subscribers) and Whitney Simmons (2.08M subscribers), to give credibility to the name and increase its awareness, reaching US$447.9 million in global net sales in 2021.[4]

1.1.3 Trending videos

On December 9, 2015, YouTube unveiled a new element of the platform: the YouTube Trending Tab. "It [trending page] is the best way to catch the videos, creators, and trends that people watch, share, and talk about each and every day. See 'em as they take off" [19]. As it can be inferred from what Kevin Allocca (Global Head of Cultures & Trends at YouTube) said, the Trending Tab seeks to feature those videos that a wide range of viewers would watch.

The Trending List is not personalized and lists the same videos in each country to all users. The Tab is updated every 15 minutes and displays roughly 200 videos. With each update, a video might move or stay in the same position or disappear from the list. There are no requirements for a video to be featured on the YouTube Trending Tab, as long as it is not misleading and/or contains profanity, mature content, or violence. Moreover, to create the ranking, the platform considers different signals, including, but not limited to, view

[3] Definition provided by Wikipedia.com

[4] Data provided by ecommercedb.com

count, how quickly the video is making views, and when the video was published [20]. By combining these signals, the Trending Page will showcase a diverse group of videos, relevant to the viewers and reflective of the content on YouTube. This implies that the video with the most views on a certain day may not be #1 on the list and may rank below videos with fewer views [20].

On March 12, 2020, YouTube updated the Trending Tab changing its name to Explore Tab. The newly revised Tab now features the so-called Destination Pages for specific video categories, in addition to the Trending List [21].

1.2 The study

1.2.1 Literature review

To write this dissertation, it was valuable to refer to an extensive literature, prevalently on the subject of predicting popularity of online videos.

Ouyang et al. in their research "A Peek Into the Future: Predicting the Popularity of Online Videos" [22] studied the popularity of videos uploaded on the Chinese video-sharing platform YouKu. They worked with a dataset of 200,714 observations, which included basic video property features (e.g., views and category) and user statistics (e.g., number of subscribers and of videos uploaded). Their analysis was structured into a two-stage prediction model. First, they wanted to predict the future popularity level of a video with the help of classification. Ouyang et al. formed four levels of popularity based on view count. They compared the performance of several classification techniques, reaching the highest accuracy with random forests. Second, they built some regression models to predict the view count of videos. For the regression task, they proposed a log-linear and a multi-linear model.

"Characterizing and Predicting the Popularity of Online Videos" by Li et al. [23] also studied data from YouKu. First, the researchers provided a characterization of the popularity dynamics, discovering popularity evolution patters. They analyzed six evolution patterns, ranging from 'burst-slow' to 'steady'. Second, Li et al. tackled the popularity prediction problem by proposing a multivariate linear regression problem that captured the evolution patterns identified.

Moreover, there have been assorted studies conducted on YouTube, due to its popularity and prevalence among video sharing platforms.

Barjasteh et al. in "Trending Videos: Measurement and Analysis" [24] studied YouTube trending videos for the first time. The data used in their analysis was collected via the YouTube API. The paper can be divided into four sections. In the first one, the researchers conducted the exploratory data analysis on basic video statistics. In the second one, trending and non-trending videos were compared to spot the differences between the two types. In the third section, the authors ran a background check on trending YouTubers to gain a deeper understanding behind their popularity. Finally, in the fourth section, they conducted a directional-relationship analysis, revealing insight onto the viewership patter among different video categories. Similar to [24] is the paper "Analysis on YouTube Trending Videos" [25] by Gayakwad et al. who, however, focused more on the exploratory analysis of variables in the dataset.

Andry et al. in [26] took a different approach in studying the trending videos on YouTube by trying to find out how the platform determines which videos can be featured on the Trending List. The authors used several data mining methods to achieve their goal, including classification, association, and clustering. Their research found out that the YouTube algorithm for selecting trending videos is based on two factors: engagement (how users reacted to videos) and metadata (whether videos' metadata matches what users were looking for).

Feroz Khan et al. in "Virality over YouTube: an empirical analysis" [27] also tried to understand what factors help some videos go viral on the platform. The authors found out that 'networks dynamics', like in-links and hits counts, and 'offline social capital', like fan base and fame, play a fundamental role in determining a video's popularity.

1.2.2 Objectives

The primary focus of this thesis is on YouTube trending videos in the United States of America. The research concentrates on the US Trending Videos as 16.4% of YouTube traffic comes from the US (240 million viewers) [28] and because half of the thirty most popular YouTube stars in the world are from the United States [29]. The platform selects only a couple hundreds of videos as trending on a daily basis. This number is highly selective when compared to the 30,000 hours of newly uploaded content every hour.[5] Thus, it would be interesting to discover what differentiate these "trending videos" from normal ones.

First, the study aims to analyze basic statistics of trending videos. A comprehensive dataset of nearly 40,000 observations is studied to discover the general features that characterize trending videos, with the intention to derive some common rules as to what makes these videos appear on the Trending List. This dataset represents traditional information regarding trending videos, including number of views, number of comments, etc. To achieve the first objective of the dissertation, the Exploratory Data Analysis will be performed.

[5] Data provided by Statista.com

Next, the research implemented some Machine Learning algorithms to group trending videos according to feature similarities and predict the number of views for trending videos. To assemble groups, two clustering methods will be applied and commented. To predict the view count of videos, different regression models will be built on the training data and then evaluated on the test data to select the best one.

The programming language implemented to complete this thesis is R, which will be used for both the exploratory and predictive analysis. The Python programming language will be used for a small section of the EDA.

1.2.3 Contributions

The main contributions of this thesis can be summarized as follows:

- To understand the characteristics that make a video appear on the Trending List.
- To extract key knowledge that could be used to improve growth of YouTube channels.
- To provide content creators (YouTubers) with information required to upload a video that could be featured on the Trending List.
- To help businesses select videos for their advertising and marketing strategies.
- To cluster trending videos to gain a deeper understanding of their nature.
- To predict the number of views based on the basic statistics of videos and select the best model.

1.3 Thesis outline

This dissertation is structured into four chapters as follows:

- CHAPTER 1:

The first chapter is intended to give a general background on the subject of the study: YouTube's trending videos. After a brief history of the video-sharing platform, there is a summary of its current situation and a comparison with its biggest competitors. Next, we can find some of the most important applications of YouTube that certainly helped its huge success, namely: the role of YouTubers, YouTube's multiple purposes, and YouTube for advertising and marketing. Central focus of this work are videos found in the so-called "trending list", thus a short introduction to such is presented. Furthermore, the objectives and structure of the analysis were explained, providing some practical contributions to real-word situations. Finally, all academic articles which were relevant to this thesis were summarized.

- CHAPTER 2:

In the second chapter, the main theoretical concepts behind clustering and regression will be introduced, along with the specific models used in this dissertation.

- CHAPTER 3:

The third chapter will deal with the preliminary steps needed to be performed before the predictive analysis. After a brief description of the original dataset, the exploratory data analysis (EDA) is the first step in the study. For EDA, we will deal with one variable at a time, trying to extrapolate the most important features that could help better understand the data.

- CHAPTER 4:

The final chapter of the thesis will deal with the clusterization and predictive analysis performed on the dataset. We will focus on one method at a time, providing a clear explanation and commenting on the results.

CHAPTER 2

2.1 Clustering

We refer to clustering as those techniques that aim to find subsets, or *clusters*, in a given dataset. When we cluster some observations, we attempt to separate them into non-overlapping groups so that observations in the same group are similar to each other and different from observations in other sets [30]. Clustering is an unsupervised machine learning method and as such seeks to "[…] discover hidden patterns or data groupings without the need for human intervention" [31].

This chapter focuses on the two perhaps best-known approaches when it comes to clustering: K-means and hierarchical clustering. With the former, we partition the observations into a pre-defined number of clusters; in the latter, the number of clusters is not known beforehand [30].

2.1.1 K-Means clustering

To perform K-means clustering, we first need to define the number of clusters K so that the algorithm will assign each observation to only one of these K clusters. Define $C_1, \dots, C_K$ as the sets containing the indices of the n observations in each cluster [30]. These sets follow two properties:

1. Each observation is at least in one of the K clusters: $C_1 \cup \ldots \cup \ C_K \ = \ \{1, \dots, n\}$.
2. Each observation belongs to one cluster only: $C_k \ \cap \ C_{k'} \ = \ \emptyset$, for all k ≠ k'.

K-means clustering is executed accurately when the within-cluster variation is minimized. The within-cluster variation for cluster C_k measures the amount $W(C_k)$ by which the observations in the same cluster differ from each other [30]. The best partition into K clusters is achieved when the total within-cluster variation, summed over all K clusters, is as small as possible.

$$\min_{C_1,\dots C_K} \left\{ \sum_{k\,=\,1}^{K} W(C_k) \right\} \qquad (1)$$

To solve (1), we first need to define the within-cluster variation $W(C_k)$. Commonly, $W(C_k)$ is defined by the means of the squared Euclidean distance. By using the squared Euclidean distance, the within-cluster variation for the k-the cluster is interpreted as the sum of all the pairwise squared Euclidean distances between the data points in the given cluster, divided by its cardinality [30]. Thus, we define it as:

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i\prime \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i\prime j})^2 \qquad (2)$$

Note: $|C_k|$ is the cardinality of the k-the cluster

Finally, by putting (1) and (2) together the optimization problem that characterizes K-means clustering is defined.

$$\min_{C_1,\ldots C_K} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i\prime \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i\prime j})^2 \right\} \qquad (3)$$

Solving the problem in (3) can be an intricate task considering that there are almost K^n ways to divide n data points into K clusters. However, we can define the following algorithm that will help us reach a local optimum [30].

Algorithm 1 – K-Means Clustering

```
1. Initial cluster assignments: randomly assign a number from 1 to K to each data point in the
   dataset, forming K clusters.
2. Find the cluster centroid for the K clusters. The k-th centroid is the vector containing the mean of
   each variable for the observations in the k-th cluster.
3. Assign each data point to the cluster which centroid is closest in terms of Euclidean distance.
Repeat until the cluster assignments remain the same.
```

To perform K-means clustering, the number of clusters K must be pre-defined. The task of selecting K can be complicated. This problem can be solved using the Elbow Method. In [32], Nainggolan et al. explain the elbow method as a graphical technique to determine the optimal number K of clusters by looking at the percentage of the comparison among the different number of clusters, selecting the value that will form an elbow in the plot. The comparison is expressed in terms of the sum of squared error (SSE), which measures, in this case, the distance between the data point and the cluster centroid.

2.1.2 Hierarchical clustering

The main disadvantage of using K-means clustering is that it requires a pre-defined number of clusters K. Hierarchical clustering is an alternative technique that solves such a problem. As it can be implied from its name, this clustering method produces a hierarchical representation in which the clusters at one level are formed by merging clusters from the lower level [33]. At the lowest level, each data point in the dataset forms a singleton cluster (leaf). Moving along, some leaves are merged into branches based on similarities to other observations. Finally, at the highest level, all the data is grouped in only one cluster. This technique results in a tree-based representation, called "dendrogram" [30]. To identify the number clusters, a horizontal line must be drawn across the dendrogram.

There are two approaches when it comes to hierarchical clustering, perhaps the most common is the "agglomerative" (or bottom-up) one [33]. This approach refers to the fact that the dendrogram is built starting from the bottom (n clusters for n observations) and, at each level, the two closest – least dissimilar – clusters are merged. Henceforth, a so-called dissimilarity measure should be defined.

Let A and B represent two clusters. As Hastie et al. explain, the dissimilarity $d(A, B)$ between the two clusters is computed from the set of pairwise observation dissimilarities d_{xy}, where $x \in A$ and $y \in B$. Now, three measures (or linkage methods) can be defined [33]:

1. Single linkage considers the smallest of these dissimilarities and, therefore, merges A and B if a single dissimilarity is small.

$$d_{Single}(A, B) = \min_{x \in A, y \in B} d_{xy} \qquad (4)$$

2. Complete linkage takes the largest of these dissimilarities and, therefore, merges A and B if all observations in their union are similar.

$$d_{Complete}(A, B) = \max_{x \in A, y \in B} d_{xy} \qquad (5)$$

3. Average linkage considers the average dissimilarity between groups and represents a compromise between the previous methods.

$$d_{Average}(A, B) = \frac{1}{|A||B|} \sum_{x \in A} \sum_{y \in B} d_{xy} \qquad (6)$$

Note: $|A||B|$ are the cardinalities of cluster A and cluster B

2.2 Regression

The regression analysis is one of the two types of supervised learning problems. Regression is used to discover the relationship between independent (often called "predictors" or "exploratory variable") and dependent (often referred to as “outcome” or “response”) variables [34]. While there are different approaches to the regression analysis, this chapter focuses on linear regression and random forests.

2.2.1 Linear regression

Linear regression aims to model the relationship between the dependent variable Y and independent variable(s) X by fitting a linear equation to the observed data. The case of one exploratory variable is known as “simple linear regression”; the case of two or more predictors is known as “multiple linear regression”.

Simple linear regression supposes there is approximately a linear relationship between X and Y , which can be represented by the relation below.

$$Y \approx \beta_0 + \beta_1 X \quad (7)$$

In (7), β_0 and β_1 are two constants that represent the intercept and the slope terms of the linear model and are called “model coefficients”. Using the training data to estimate the coefficients, it is possible to predict the outcome variable on the assumption that $X = x$ [30].

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x \quad (8)$$

The most common approach to estimate the values of the coefficients is known as the least squares’ criterion. Consider n observation pairs $(x_1 y_1), \dots, (x_n y_n)$ consisting of a measurement of X and one of Y. The goal is to find estimates $\widehat{\beta_0}$ and $\widehat{\beta_1}$ such that $y_i \approx \widehat{\beta_0} + \widehat{\beta_1} x_i$ for $i = 1, \dots n$. This means that the resulting regression line should be close as possible to the n data points. If $\hat{y}_i = \widehat{\beta_0} + \widehat{\beta_1} x_i$ is the prediction for Y based on the i-th value of X, we can define the i-th residual as $e_i = y_i - \hat{y}_i$ [30]. Then, the residual sum of squares (RSS) is $RSS = {e_1}^2 + \dots {e_n}^2$. Finally, the least squares method estimates the values of the coefficients that minimize the RSS. Moreover, because the residuals are squared and then summed, there are no cancellations between positive and negative values [35].

Moreover, since the real relationship between X and Y may not be linear, we should consider an error term ϵ. Hence, (7) becomes:

$$Y = \beta_0 + \beta_1 X + \epsilon. \quad (9)$$

In practice, however, there are more than one predictor in a dataset. One could fit a simple linear regression model for each exploratory variable, but the results would be satisfactory. A better approach would be to extend the simple linear regression model so that it can make use of all the predictors. This is possible by associating to each variable X its own slope coefficient β [30]. If a dataset has p predictor, this means that (9) would be transformed to the equation below.

$$Y = \beta_0 + \beta_1 X_1 + \ldots \beta_p X_p + \epsilon \quad (10)$$

Finally, any linear regression model is based on four assumptions [36]:

1. Linearity: the relationship between X and Y is linear.
2. Homoscedasticity: the variance of residuals is the same for any value of X.
3. Multicollinearity: observations are independent of each other.
4. Normality: the error terms must be normally distributed.

2.2.2 Random forests

The random forest algorithm is a supervised learning algorithm that represents a type of ensemble learning. It uses the decision tree as individual model and bagging (or bootstrap aggregating) as the ensemble method. Hastie et al. in [37], describe random forests as a method to average multiple decision trees, trained on different samples of the dataset, with the aim to reduce the variance. In a regression problem, the result provided by the algorithm is the average of the results of the different decision trees.

Random forests can also be seen as an improvement of bagged trees. As in bootstrap aggregating, given the training set $X = x_1, \ldots, x_n$ with response $Y = y_1, \ldots, y_n$, it selects a random sample of the training data and fits a decision tree to these samples. In addition to this, random forests perform feature bagging: when building the trees, every time a split is considered, a random sample of *m* features is selected [30]. This process reduces the correlation among the "bagged" trees. Indeed, in the case of the presence of a strong predictor p_k in the dataset, most bagged trees will use this strong feature in the top split, making all the tree similar. Feature bagging, instead, forces the algorithm to consider a subset of the *p* predictors, so that $\frac{p-m}{p}$ splits will not even select p_k [30].

Finally, random forests can be used to rank the importance of variables via a variable importance plot.

CHAPTER 3

3.1 The dataset

The dataset used for this thesis was downloaded from Kaggle.com and it was created by user Mitchell J, who publicly made it available. The original dataset, named “Trending YouTube Video Statistics” [38], consists of ten different CSV files. Each of these files contains information about trending videos in different countries, namely: USA, UK, Germany, Canada, France, Russia, Mexico, South Korea, Japan, and India. The data was collected using the YouTube Application Programming Interface (API).

Each single dataset is structured into 16 columns reporting basic statistics regarding videos featured on the Trending List (see Table 1).

To perform the analysis in this thesis, only the dataset corresponding to trending videos in the United States of America was used. Moreover, out of the 16 columns in the original set, only three were discarded, namely: *comments_disabled, ratings_disable,* and *video_error_or_removed*.

Before starting to work on the dataset, it is fundamental to do some data cleaning. According to [39], data cleaning is the process of detecting and dealing with errors and inconsistencies in the data, present due to problems during data entry, with the intention of improving data quality. While there are some general requirements, data cleansing is a process that heavily depends on the data at hand.

After importing the original CSV file into RStudio and removing the unwanted variables, the dataset consists of a collection of 40,949 observations with 13 columns. Let’s name the dataset *us_trendingvideos*. Among the first steps to do when dealing with a new collection of data is to check for missing values. Missing data, represented by NA in R, appear when there is no value reported in one or more entries. For this reason, it is impossible to perform statistical analysis on a collection with one or more missing values [40]. There are no missing values in the columns of *us_trendingvideos*.

Next, it is also important check if there are any duplicate values. In this case, it needs to be checked whether the unique variable *video_id* appears in more than one row with same values in all columns. Given the nature of the data, there are duplicate values of some video IDs in *us_trendingvideos*. Indeed, a video can appear on the Trending List on multiple days. However, these observations cannot be considered duplicates as variables such as *trending_date, views, likes,* and *dislikes* change from row to row. Therefore, unless explicitly stated, the analysis will use the entire dataset. If a video goes trending multiple times, it is fair to think it has more "trending power", hence it would not make sense to discard those videos which appear more than once. Out of curiosity, the dataset has only 6,351 unique videos out of the 40,949 reported.

Finally, let's check the structure of the dataset to visualize the datatype of the variables and make changes accordingly during the exploratory data analysis in the next section.

Variable:	**Datatype:**	**Example:**
video_id	Factor with 6,351 levels	2kyS6SvSYSE
trending_date	Factor with 205 levels	17.14.11
title	Factor with 6,455 levels	WE WANT TO TALK ABOUT OUR MARRIAGE
channel_title	Factor with 2,207 levels	CaseyNeistat
category_id	Int	22
publish_time	Factor with 6,269 levels	2017-11-13T17:13:01.000Z
tags	Factor with 6,055 levels	SHANtell martin
views	Int	748,374
likes	Int	57,527
dislikes	Int	2,966
comment_count	Int	15,954
description	Factor with 6,902 levels	SHANTELL'S CHANNEL – https://www.youtube.com/shantellmartin [...]
thumbnail_link	Factor with 6,352 levels	https://i.ytimg.com/vi/2kyS6SvSYSE/default.jpg

Table 1

3.2 Exploratory data analysis

According to IBM in [41], the exploratory data analysis (EDA) consists in analyzing and investigating the dataset in order to summarize its aspects, often with the help of graphs. EDA helps to grasp a better understanding of the data used and to determine what are the appropriate statistical techniques to consider for the analysis.

To keep an overall organized structure, the thesis conducts the exploratory data analysis studying one variable at a time following the order from Table 1, with the only exception for *trending_date*, *publish_time*, *likes*, and *dislikes*, which will be analyzed in couples. The study will try to extrapolate some important features regarding the *us_trendingvideos* dataset, while considering the findings in Table 1.

3.2.1 Trending and publish date

The variable *trending_date* corresponds to the day the video appeared on the Trending List on YouTube. Thanks to the findings in Table 1, it can be noticed that the variable is a factor, so it should be converted to a date value.

By converting *trending_date* to the date datatype using functions from the package "lubridate", it is possible to figure out when the data were collected. The original dataset contains information of trending videos from 11th November 2017 to 14th June 2018, for a total of 213 days of observation. Moreover, considering that YouTube usually selects 200 videos to be featured on the Trending List, the dataset should contain a total of 42,600 rows. However, there are only 40,949 observations because for some days the featured trending videos were fewer than 200.

The variable *publish_time* contains both upload time and day. For future ease, let's create a new variable to differentiate between publish time and publish day. The original variable *publish_time* now contains only the upload time (for simplicity, only the hour was considered, discarding minutes and seconds) and the new variable *publish_date* contains only the upload date.

In addition, from the difference of the newly created variable *publish_date* and the original *trending_date*, a new variable corresponding to how many days it took a video to appear on the Trending List is produced: *days_to_trending*.

From this new variable, it can be observed that it takes 16 days on average from its upload day for a video to be featured on the Trending List.

Videos can be uploaded on YouTube at any time on any day of the week. It can be valuable for both YouTubers and companies to understand when they should publish a video to maximize its performance and have a higher possibility to get it featured on the Trending List.

From Figure 1, it can be noticed that the majority of videos were uploaded at 4PM (3,669 videos), followed by 3PM (3,483 videos) and 5PM (3,447 videos).

Then, we can find on what day of the week trending videos were uploaded. For further insights, each day is divided into three “upload intervals”: 8AM-3PM, 4PM-11PM, 12AM-7AM. As it can be ascertained from Figure 2, most videos that ended up trending were uploaded on a Friday in the 4PM-11PM time period.

Note: only rows with a unique video ID were considered to achieve the bar plots in Figure 2 and 3, as the variables *publish_time* and *publish_day* remain the same.

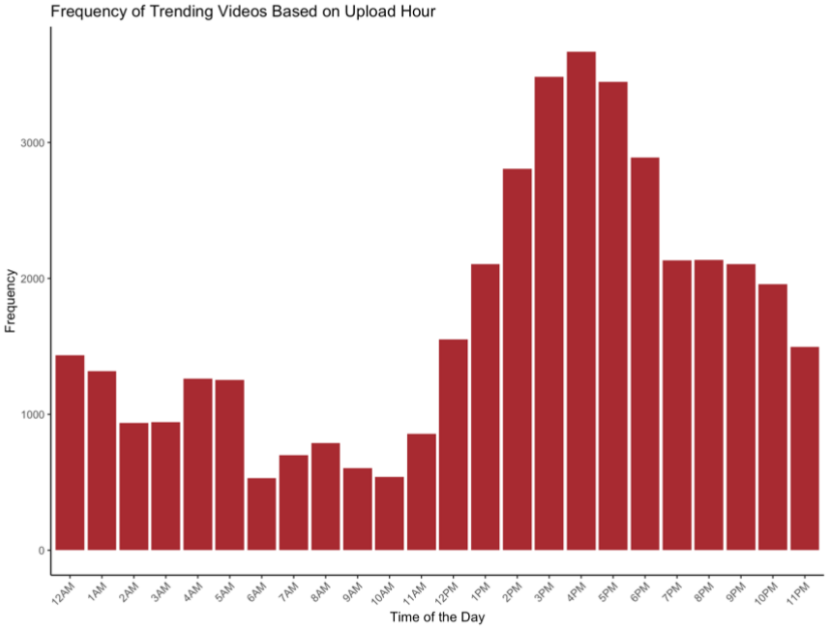


Figure 1

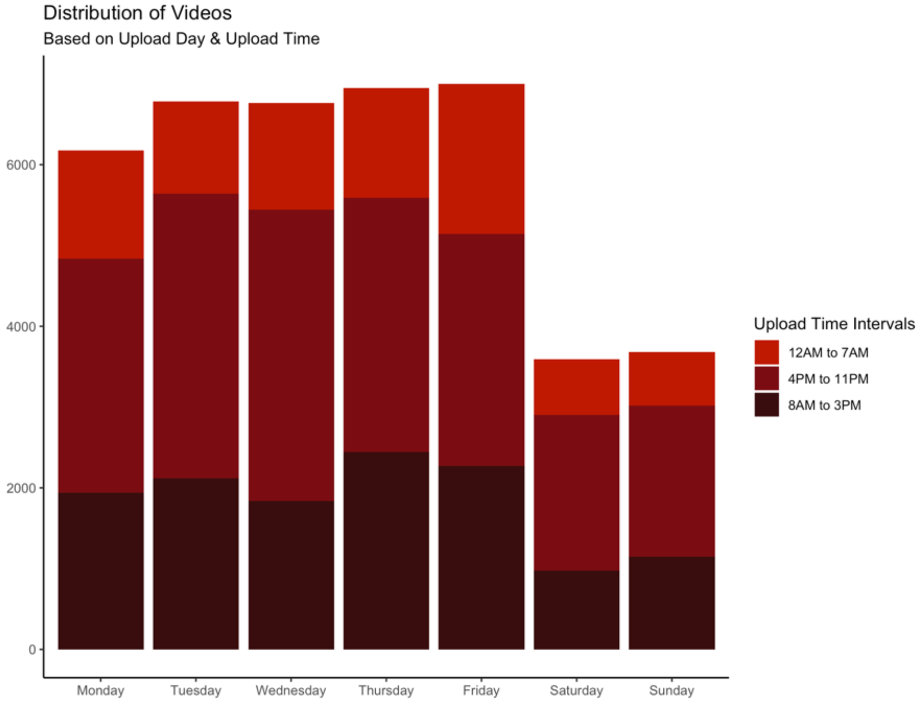


Figure 2

As discussed in section 3.1, a video can trend for multiple days. It can be interesting to find those videos that trended the longest. To achieve this, the function *group_by()* is applied on the unique variable *video_id* to group the same videos and count its occurrences. By doing this, it is possible to identify a video that trended for 30 days and seven videos that trended for 29 days (see Table 2 for an instance). However, it can be deducted from Figure 3 that these can be considered exceptions as most videos trended for less than 10 days in total. Finally, it is also insightful to calculate both the mean and the standard deviation of the number of trending days for videos. Indeed, a video on average appears on the YouTube Trending List for six days and, having a standard deviation lower than 5, the data fall somewhat close around the mean (hence more reliable). By looking at these statistics, it is possible to conclude that the videos reported in Table 2 are exceptions to the general trend.

Title:	WE MADE OUR MOM CRY… HER DREAM CAME TRUE!	Sam Smith – Pray (Official Video) ft. Logic
Channel:	Lucas and Marcus	SamSmithWorldVEVO
Trended for:	30 days	29 days
Views:	10,381,263	20,934,803

Table 2

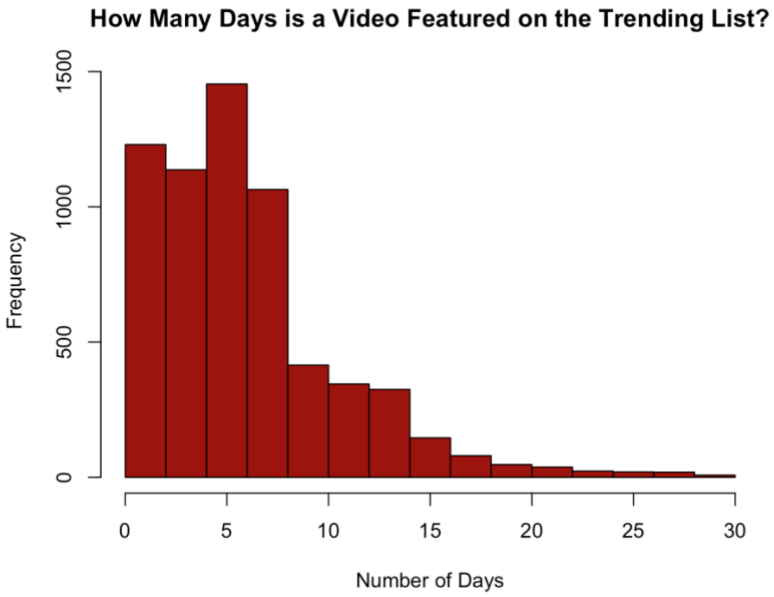


Figure 3

3.2.2 Title

The title of a video is of paramount importance as it affects the video's performance. Before playing a video users can only see the title and the thumbnail (see section 3.2.10), hence the former needs to be able to attract people and convince them to watch the video [42]. YouTube itself states that "well-written titles can be the difference between someone watching and sharing your videos, or scrolling past it" [43].

From Table 1, it can be noticed that *title* is a factor, hence it should be converted to a character. Next, let's create a new variable representing the number of characters in videos' titles: *title_length*. From Figure 4, it can be observed that most trending videos have between 30 and 50 characters in their titles.

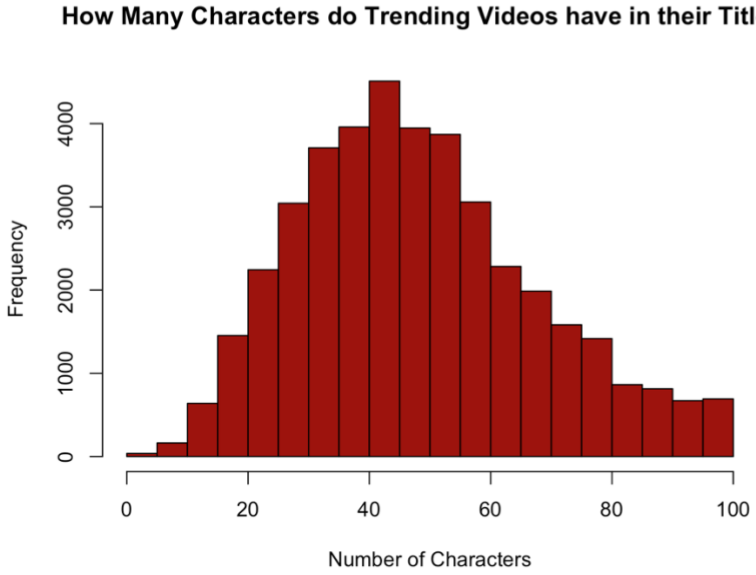


Figure 4

Next, it would be useful to find the most-used words in trending videos' titles, so that YouTubers and/or marketers can have an idea of what to include in the title. For better results, some common stop words were removed, including, but limited to: "and", "the", "an". Here are the 20 most used words.

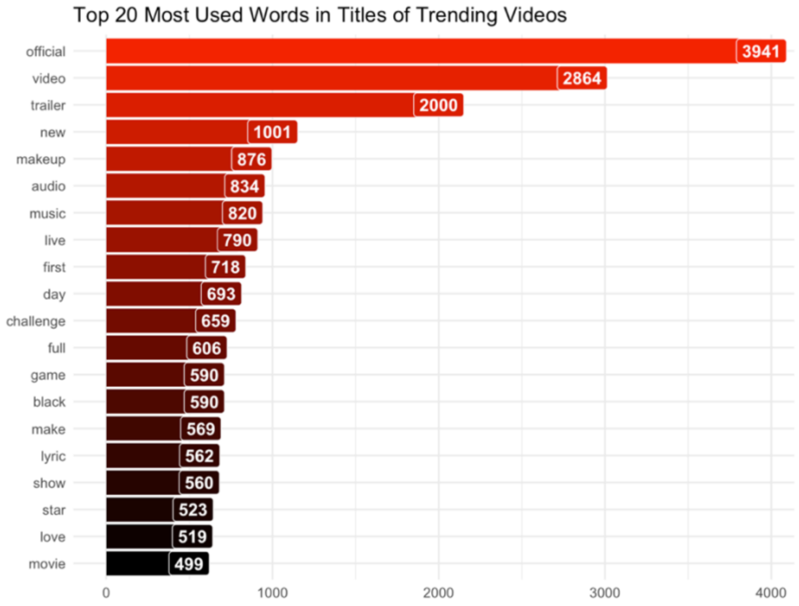


Figure 5

3.2.3 Channels

From Table 1, the variable *channel* is a factor, hence it should be converted to character. From this variable is possible to find what are the channels that have the most videos featured on the YouTube Trending List. As it can be acknowledged from Figure 6, among the first twenty channels with the most trending videos featured, seven are the official channels of famous American talk shows, like Jimmy Fallon and Ellen DeGeneres.

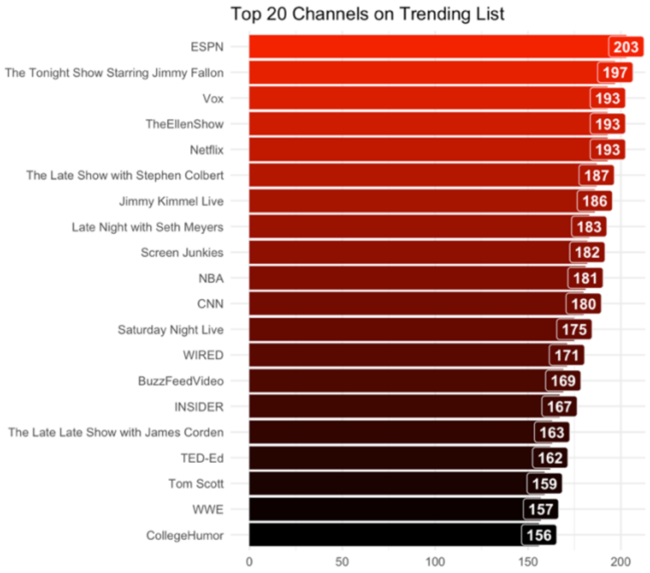


Figure 6

3.2.4 Categories

YouTube videos are organized in different categories, which help classify the videos. There are fifteen categories in total, ranging from Music to Sports and everything in between.

In *us_trendingvideos*, the variable *category_id* is an int and each number corresponds to one category. Hence, using [44] to discover what category each number represents, it is possible to map a number to its corresponding category name and store the results in a new variable: *category_name*.

Similar to the study performed on the variable *channels*, let's find those categories that have the highest number of videos on the Trending List. As it can be seen in Figure 7, the Entertainment category has the most videos, which explains why seven of the channels in Figure 6 are talk shows' official YouTube channels.

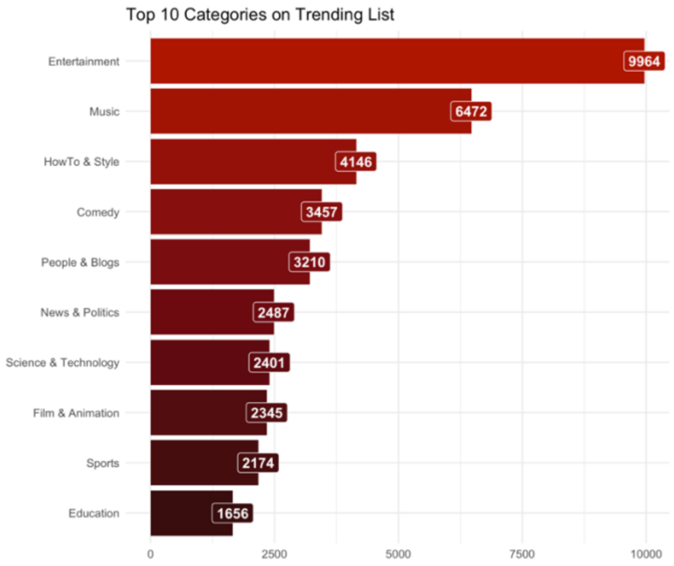


Figure 7

3.2.5 Tags

Tags are keywords a YouTuber can add to a video when uploading it to the platform that will make it easier for users to find relevant content. As YouTube itself affirms, metadata like the video's title (see 3.2.2), description (see 3.2.9), and thumbnail (see 3.2.10) are what truly help a viewer decide which videos to watch. "Tags can be useful if the content of the video is commonly misspelled. Otherwise, tags play a minimal role in the video's discovery" [45].

Similar to the variable *title*, the variable *tags* needs to be transformed from factor to character. Then, it may be useful to find out the most-used tags in US trending videos. In Figure 8 are the ten most used tags.

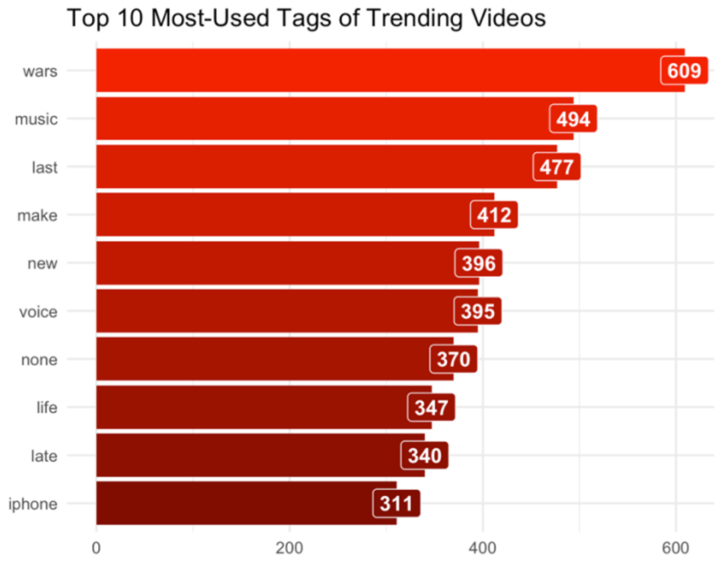


Figure 8

3.2.6 Views

Next in the exploratory data analysis is the variable *views*. The view count of a video is one of the YouTube engagement metrics and it corresponds to the number of legitimate views for a video [46]. Kayla Carmichael in [47] explains what YouTube considers “legitimate”. The view count on the platform increases when the video is played by a human being. To make sure of this, YouTube stops the view count of a video once it attains 300 views and starts a system check to detect the presence of previous or incoming deceitful (computer generated) views. Videos with fewer than 300 views do not have the power to be featured on the homepage and elude the YouTube’s algorithm, hence they are not checked.

Now that a general overview of the variable under consideration is given, let’s start the analysis by looking for the three most-watched and three least-watched videos in the dataset. If one of these videos appeared on the Trending List multiple times, the study will report the view count of the video’s most recent appearance. First, from Table 1, *views* is an int so it should be converted to numeric.

From Table 3, it can be noticed that two videos are from the second most popular category (Music) and one from the first popular category (Entertainment).

From Table 4, these videos appeared on the Trending List even with less than a thousand views. This confirms that YouTube does not look only at the view count when selecting trending videos.

Title:	Childish Gambino – This is America (Official Video)	YouTube Rewind: The Shape of 2017 | #YouTubeRewind	Ariana Grande – No Tears Left To Cry

Channel:	ChildishGambinoVEVO	YouTube Spotlight	ArianaGrandeVEVO
Views:	225,211,923	149,376,127	148,689,896
Trending date (D/M/Y):	02/06/2018	14/12/2017	14/05/2018
Category:	Music	Entertainment	Music
Upload date (D/M/Y):	06/05/2018	06/12/2017	20/04/2018

Table 3 – Most-watched videos

Title:	1 dead, others injured after Ky. School shooting	Coach Taggart Monday Presser Ahead of Arizona	Artwork Forge
Channel:	Newsy	GoDucksdotcom	Palo Alto Online
Views:	559	704	745
Trending date (D/M/Y):	28/01/2018	17/11/2017	29/01/2018
Category:	News & Politics	Sports	Sports
Upload date (D/M/Y):	23/01/2018	13/11/2017	10/01/2018

Table 4 – Least-watched videos

Next, let's look at the distribution of the view count in trending videos to understand how many views videos get on average. From Figure 9 it can be concluded that most videos' view count is between 200,000 and 2,000,000. (Note: the variable *views* was log-scaled to plot the histogram below for a better readability).

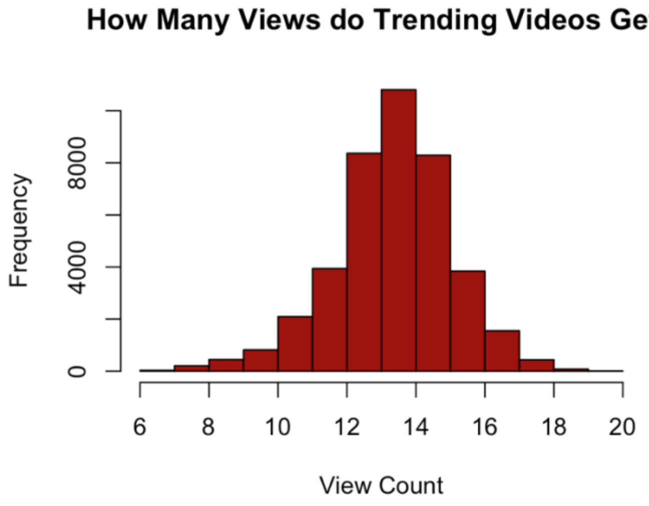


Figure 9

3.2.7 Likes and dislikes

Likes and dislikes are the other two engagement metrics that reflect how many times a video on YouTube has been interacted with. Just like the view count, the platform makes sure that these metrics are legitimate and come from human beings only [46]. In November 2021, the YouTube Team announced that viewers would not be able to see the dislike count anymore but would be able to use the dislike button; creators, however, still have access to the count. The Team opted for this solution to protect YouTubers from so-called dislike attacks [48].

Similar to the analysis for the variable *views*, let's find the three most-liked and most-disliked videos on the Trending List over the relevant period.

From Table 4, it can be noticed that two of the most-liked videos are also those which are the most-watched, this hints to a correlation between view count and like count.

From Table 5, two interesting facts can be drawn:

1) Two of the most-disliked videos are from the same creator, Logan Paul (23.4M subscribers). The YouTuber is known for making views from his controversial videos. A notable example is the video about his visit to Aokigahara (also known as suicide forest) in Japan, during which the creator filmed a suicide victim and uploaded the footage on the YouTube platform [49]. After receiving negative feedback from his viewers, he uploaded two videos: one in which he apologized ("So Sorry") and one in which he tried to raise awareness about suicide ("Suicide: Be Here Tomorrow.") [50].
2) The second most-disliked video is one of the most-watched videos as well (see Table 3). Hence, having a lot of views does not imply that a video will be necessarily popular in a positive light.

Title:	BTS 'FAKE LOVE' Official MV	Childish Gambino – This is America (Official Video)	Ariana Grande – No Tears Left To Cry
Channel:	HYBE LABELS	ChildishGambinoVEVO	ArianaGrandeVEVO
Like count:	5,613,827	5,023,450	3,094,021
View count:	123,010,920	225,211,923	148,689,896
Category:	Music	Music	Music
Trending date (D/M/Y):	01/06/2018	02/06/2018	14/05/2018

Table 4

Title:	So Sorry.	YouTube Rewind: The Shape of 2017 | #YouTubeRewind	Suicide: Be Here Tomorrow.
Channel:	Logan Paul Vlogs	YouTube Spotlight	Logan Paul Vlogs
Dislike count:	1,674,420	1,643,059	497,847
View count:	37,539,570	149,376,127	24,286,474
Category:	Entertainment	Entertainment	Nonprofit & Activism
Trending date (D/M/Y):	09/01/2018	14/12/2017	01/02/2018

Table 5

After converting both variables *likes* and *dislikes* from int to numeric, let's look at their distribution to find how many likes and dislikes trending videos get on average. Most videos get 5,000 to 50,000 likes (Figure 10) and 200 to 2,000 dislikes (Figure 11). (Note: the variables *likes* and *dislikes* were log-scaled to plot the histograms below for a better readability).

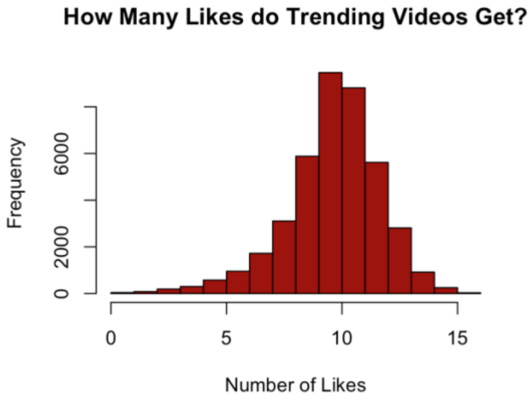


Figure 10

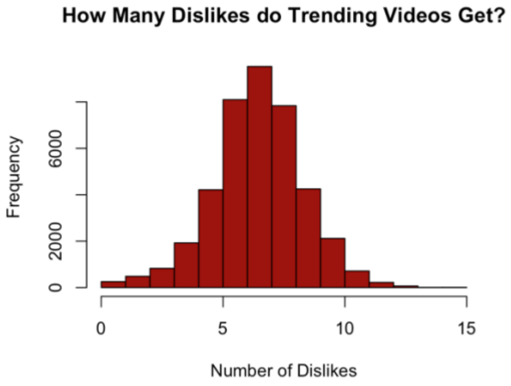


Figure 11

3.2.8 Comments

The comment section under a YouTube video is a place where viewers can share their thoughts on the content watched and YouTubers; creators, then, can reply to these comments. In this way, YouTubers acknowledge their followers and create a real online community.

The variable *comment_count* in the dataset refers the total number of comments left under a video. As for other variables, let's convert it to numeric and find the three videos with the highest number of comments and then look at the distribution of *comment_count*.

From Table 6, all three videos with the highest comment count sound familiar: "So Sorry" is one of the most disliked videos in the dataset, "BTS 'FAKE LOVE' Official MV" is one of the most liked videos in the dataset, and "YouTube Rewind: The Shape of 2017" is both one of the most watched and disliked videos in *us_trendingvideos*.

Next, by looking at the distribution of the number of comments in Figure 12, it can be concluded that most trending videos have between 700 and 5,000 comments. (Note: the variable *comment_count* was log-scaled to plot the histogram below for a better readability).

Title:	So Sorry	BTS 'FAKE LOVE' Official MV	YouTube Rewind: The Shape of 2017 | #YouTubeRewind
Channel:	Logan Paul Vlogs	HYBE LABELS	YouTube Spotlight
Comment count:	1,361,580	1,228,655	827,755

Table 6

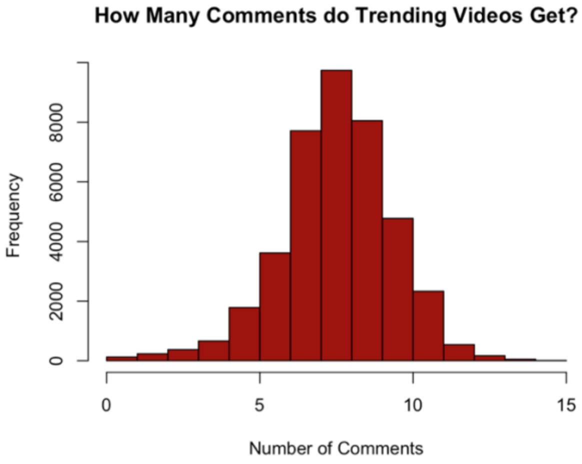


Figure 12

3.2.9 Description

The description box below a video allows the YouTuber to share some information with the viewer. Most creators use it to briefly summarize the content of the video, to mention other creators that are part of the video [51], and to include hashtags that allow the YouTuber to connect the video with similar content on the platform [52].

After converting the variable *description* to character, let's create a new variable representing the number of characters in the video's description: *description_length*. In Figure 13, it can be seen that most trending videos have fewer than 1,000 characters in their descriptions.

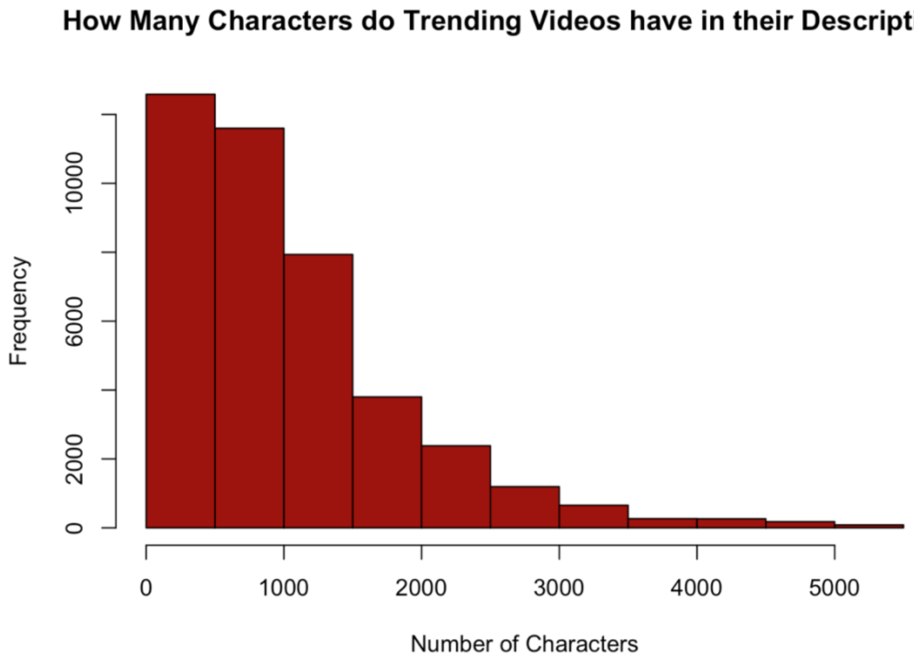


Figure 13

Next, it would be insightful to find the most-used words in trending videos' description, to see what trending creators use it for. For better results, some common stop words were removed including, but limited to: “that”, “the”, “an”. From Figure 14, it can be hypothesized that most YouTubers use the description box to invite their viewers to follow them on other social media (like Twitter, Instagram, and Facebook) and to subscribe to their channel and to like the video.

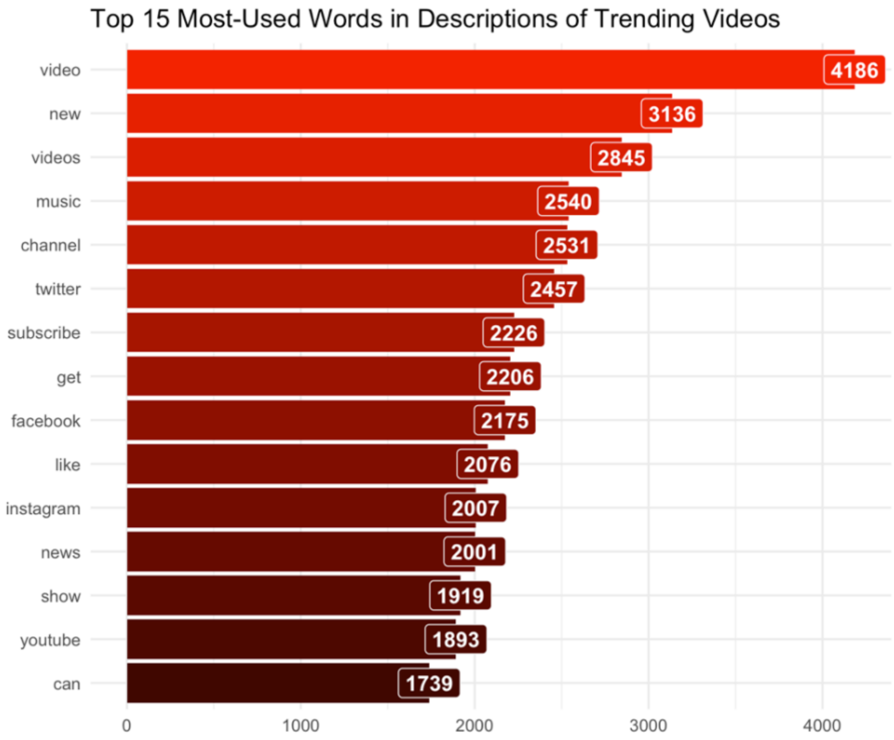


Figure 14

3.2.10 Thumbnails

Thumbnails act like a preview for videos. “Video thumbnails let viewers see a quick snapshot of your video as they are browsing YouTube” [53]. Just as titles, thumbnails play a crucial role in a video’s popularity. According to [54], thumbnails have three purposes:

- Attraction: the most critical function of a thumbnail is to capture the attention of potential viewers.
- Description: a good thumbnail should illustrate the content of the video to not give false expectations to viewers.
- Branding: thumbnails can also help viewers recognize a creator’s videos if the YouTuber keeps a consistent style and design.

It would be alluring, as well as pragmatic, to find those items that appear the most in trending videos’ thumbnails. In this way, content creators and marketers can have a general idea of what they should include in a thumbnail to have a higher possibility to get the video listed on the Trending Tab. To perform object detection, the Python programming language was used through the means of the Spyder environment.

After importing the dataset, *us_trendingvideos*, and deleting all variables but *video_id* and *thumbnail_link*, the thesis considered only unique videos. Hence, 6,351 video thumbnails were downloaded from the corresponding links.

Object detection was performed on the data using the deep learning and computer vision Python library "ImageAI", which has both prediction and detection classes. To perform the task, the *ObjectDetection* class was called. This library provides straightforward functions to detect items on any set of images using pre-trained models [55]. The model adopted in this analysis is called "YOLO v3-Tiny", proposed by Joseph Redmon in [56]. All the YOLO (You Only Look Once) models are one stage detectors, which decrease the depth of the convolutional layer[6] and increase speed [57]. Ardash et al. in [57] concluded that the YOLO v3-Tiny variant is 442% faster than older variants, but it still ensures accuracy in its results. Thesis then selected the first ten most detected items that have a probability of being correctly detected higher than 60%.

As it can be perceived from Figure 15, the majority of thumbnails report a person in them. Most likely the person is the YouTuber itself so that its viewers can recognize the creator's video, recalling the branding purpose of thumbnails. Other objects detected are day-to-day items, such as cars, clocks, and books. Among the most found objects in videos' thumbnails there are also sports balls, these are, in all likelihood, videos in the Sports category, which is one of the categories with the highest number of videos on the Trending List (see section 3.2.4).

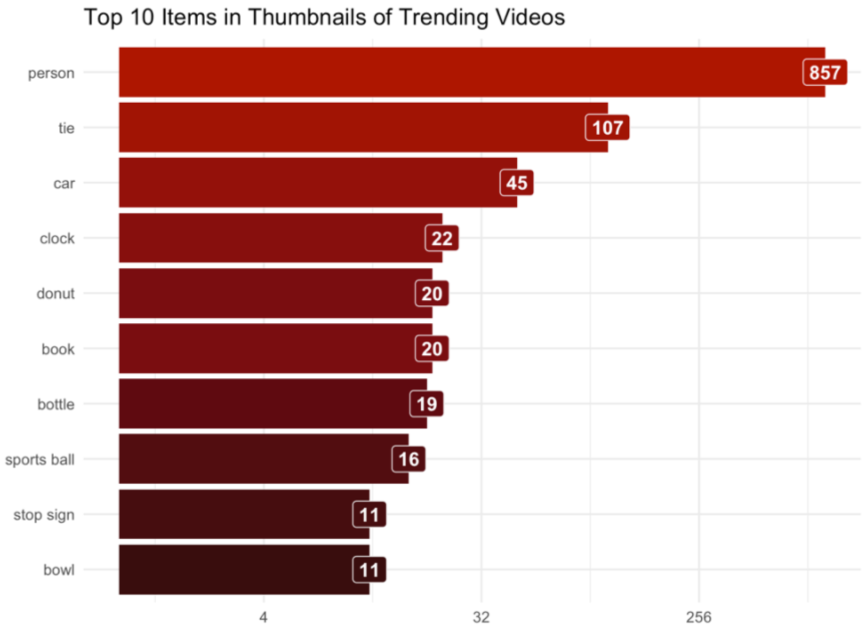


Figure 15

[6] The first layer of a CNN which converts all the pixels of an image into a single value, outputting a vector.

3.3 Summary

To conclude the exploratory data analysis in this dissertation, a brief summary of the findings is reported.

- The data from *us_trendingvideos* were collected between November 11th, 2017, and June 14th, 2018.
- It takes 16 days on average from its upload day to get a video featured on the Trending List.
- Most videos in *us_trendingvideos* trend for fewer than 10 days on average.
- Most videos in *us_trendingvideos* were uploaded on a Friday at 4PM.
- Most trending videos' titles are between 30 and 50 characters in length and include words like 'official', 'new', 'video'.
- Most trending videos get between 200,000 and 2,000,000 views, 5,000 and 50,000 likes, 200 and 2,000 dislikes, and 700 and 5,000 comments.
- Most trending videos' descriptions are shorter than 1,000 characters in length and include words like 'video, 'channel, 'Instagram', 'Twitter'.
- Most trending videos' thumbnails report a person.

CHAPTER 4

4.1 Preparation

Due to computational limitations, clustering and regression will be executed on a subset of the original dataset. The new dataset, *us_videos*, reports only the observations that coincide with the last 30 days of data collecting (from May 15th, 2018, to June 14th, 2018). After applying this filter, *us_videos* consists of 6,199 observations. Next, only the variables *views*, *likes*, *dislikes*, *comment_count*, *days_to_trending*, and *publish_time* were selected.

An extremely important preliminary step to check for outliers. Outliers are observations that are distant from all other data points [58]. In particular, it is crucial to deal and remove outliers in order to achieve the best possible results for the clustering analysis. Since K-means and hierarchical clustering force every observation to be assigned to a cluster, the clusters created may be compromised by the presence of outliers that do not fit into any cluster [30].

There are several methods to identify and, consequently, remove outliers. The thesis opted to use the interquartile range (IQR) method. The IQR is defined as the difference between the 25th and 75th percentiles [59]. Following the IQR approach, an outlier is any observation that lies outside the lower or upper bound [58]. The lower bound is found by subtracting the amount 1.5*IQR from the 1st quartile, whereas the upper bound is found by adding that same amount to the 3rd quartile.

After removing the outliers from *us_videos*, the number of observations is reduced to 2,129. In Figure 16, some boxplots are reported to show the difference between before and after the removal of outliers.

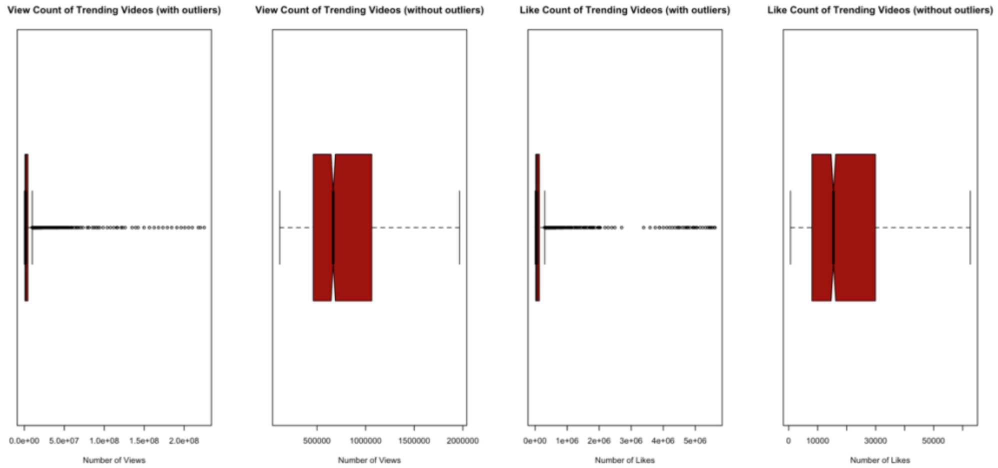


Figure 16

It is also insightful to plot the correlation matrix with the variables selected for the clustering and regression tasks. From Figure 17, it is possible to extract the following information:

- There is a relatively high positive correlation between *views* and *likes*/*dislikes*/*comment_count*.
- There is a relatively high positive correlation between *likes* and *dislikes*/*comment_count*.
- There is a relatively high positive correlation between *dislikes* and *comment_count*.

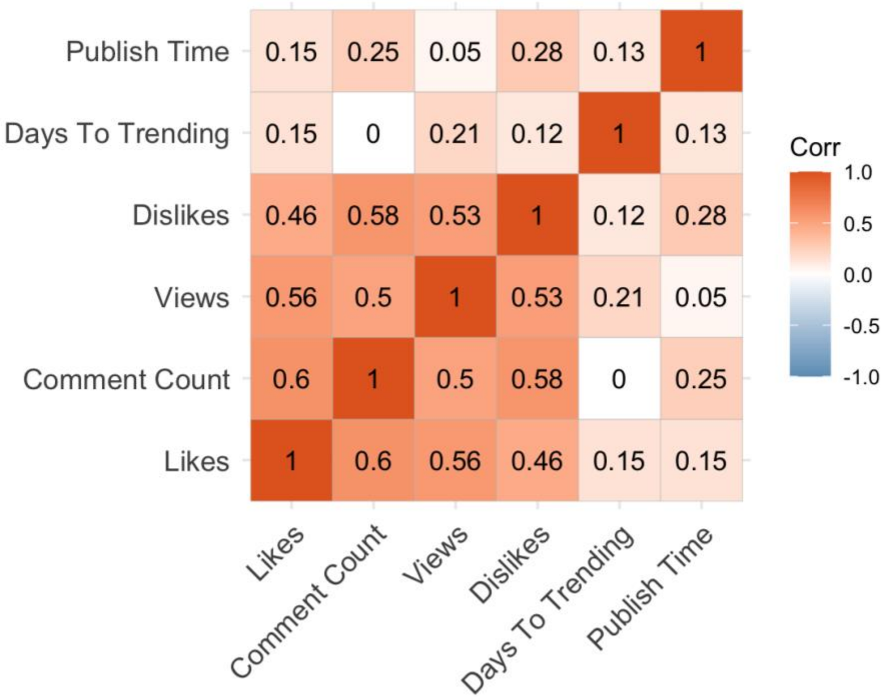


Figure 17

Let's investigate these correlations by the means of the graphs in Figure 18:

- Graph 1: the number of comments increases with the number of likes. Indeed, in the EDA we found out that some of the most-liked videos are also those with the greatest number of comments.
- Graph 2: the number of comments increases with the number dislikes. In fact, in the EDA one of the videos with the highest number of comments is also one with the greatest number of dislikes.
- Graph 3: the number of dislikes increases with the number of likes. This is an interesting insight that was not possible to extract from the EDA.

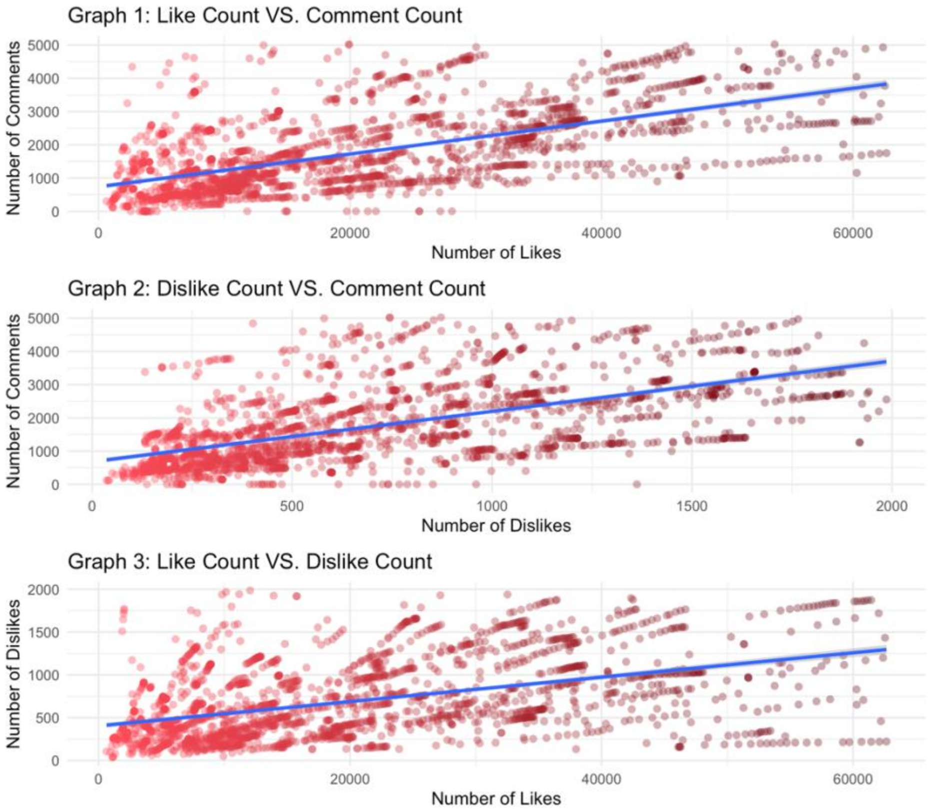


Figure 18

4.2 Clustering

Because clusters are defined by the Euclidean distance between the data points in the collection, it is important to scale the dataset used. The motivation behind standardizing the data lies in the fact that variables have different units and ranges of values [60]. Without scaling, one feature could strongly – and wrongly – influence the clustering results. This can be prevented with the standardization of the variables.

4.2.1 K-means clustering

As discussed in Chapter 2, the major drawback of K-means clustering is that it requires to pre-define the number of clusters K wanted. This problem can be overcome by applying the elbow method. This technique can be applied in R by using the function *fviz_nbclust()* from the package "factoextra". Thesis set the argument "method" of the function to total within sum of square ("wss") to achieve the plot in Figure 19. As it can be deduced from the graph below, the optimal number of clusters could be three or four.

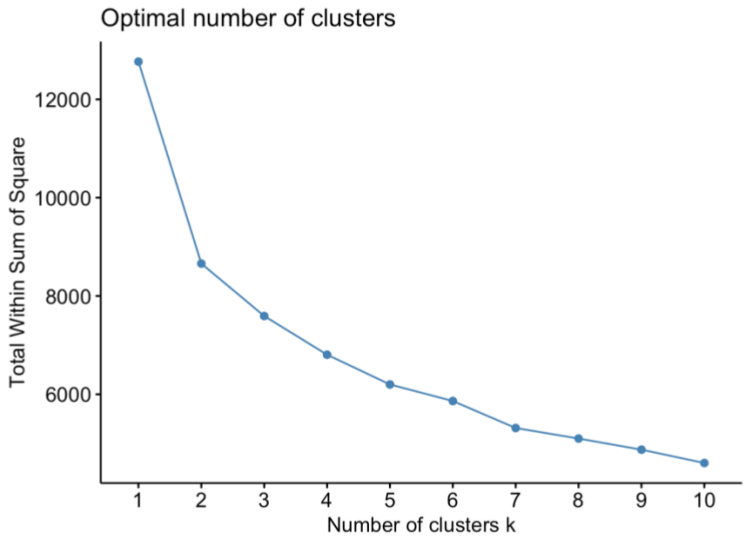


Figure 19

Thesis applied the function *kmeans()* to the standardized data, setting the number of clusters K to both three and four. Even though the WCSS is lower when K is equal to four, by looking at the cluster representation, it can be concluded that the best K-means clustering is achieved by setting K equal to three. Indeed, from Figure 20, it is possible to see how the data points are divided into three clusters and how these clusters are almost completely non-overlapping. By setting K equal to four, instead, the clusters overlap and the observations in one cluster are actually closer to the centroid of another cluster.

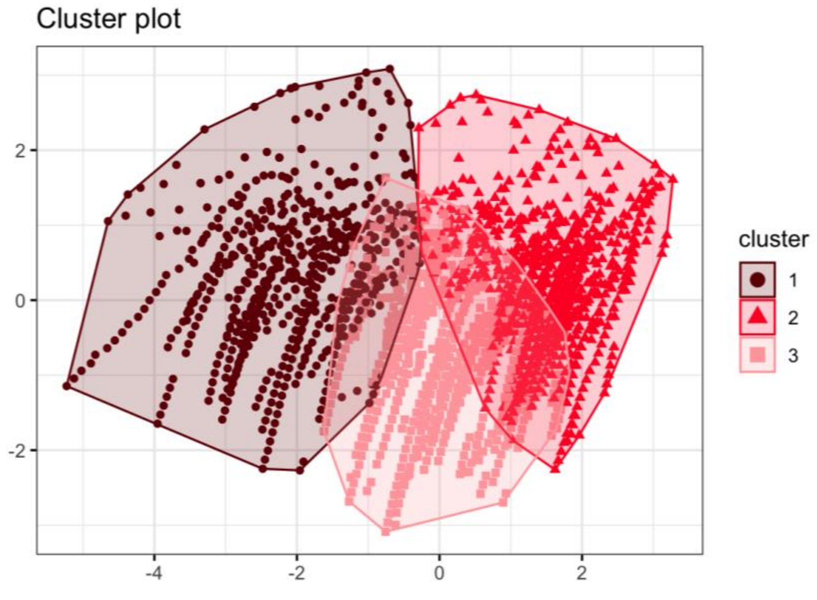


Figure 20

Moreover, other important observations coming from K-mean clustering can be summarized in the table below. From Table 8, Cluster 2 is the cluster with the most videos. It can be deduced that it is very easy for videos to be part of Cluster 2, as on average these videos have the lowest values for view, like, dislike, and comment count.

Cluster:	Number of Observations:	Views:	Likes:	Dislikes:	Comment Count:	Days to Trending:	Publish Time:
1	625	124,7578.0	37,307.87	1,093.2880	2,913.2384	10.597633	10.675200
2	859	542,950.3	10,759.73	364.9616	880.1129	10.601613	8.058207
3	645	695,986.3	16,105.93	700.8062	1,710.4992	9.014299	11.482171

Table 7

4.2.2 Hierarchical clustering

As discussed in Chapter 2, a dissimilarity measure needs to be indicated in order to perform hierarchical clustering; thesis decided to use the Euclidean distance. Moreover, thesis created a function that computes hierarchical clustering calling the function *agnes()*, using the three linkage methods explained in the second chapter. The advantage of calling *agnes()* is that it calculates the so-called agglomerative coefficients for each linkage method, measuring the strength of the clusters. The closer the coefficient is to one, the better the clustering is [61]. The highest agglomerative coefficient is achieved with linkage method set to "complete".

Next, the function *hclust()* is applied to get a graphical representation of the dendrogram. Figures 21 and 22 respectively show the original dendrogram and how the thesis thought would be best to cut the tree, forming four clusters. Finally, we can visualize how the algorithm clustered the data in Figure 23.

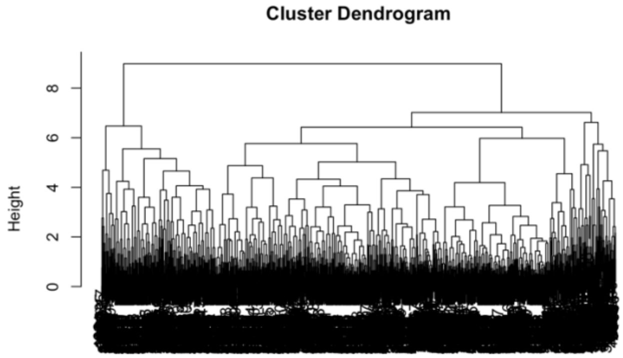


Figure 21

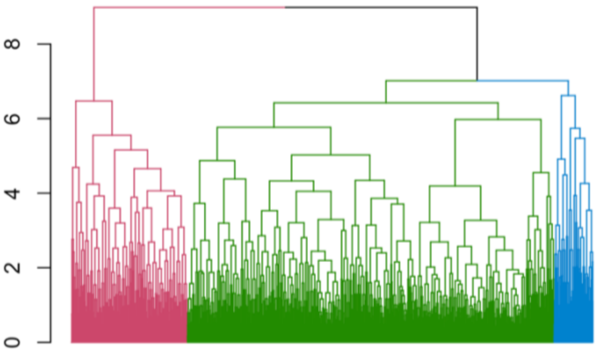


Figure 22

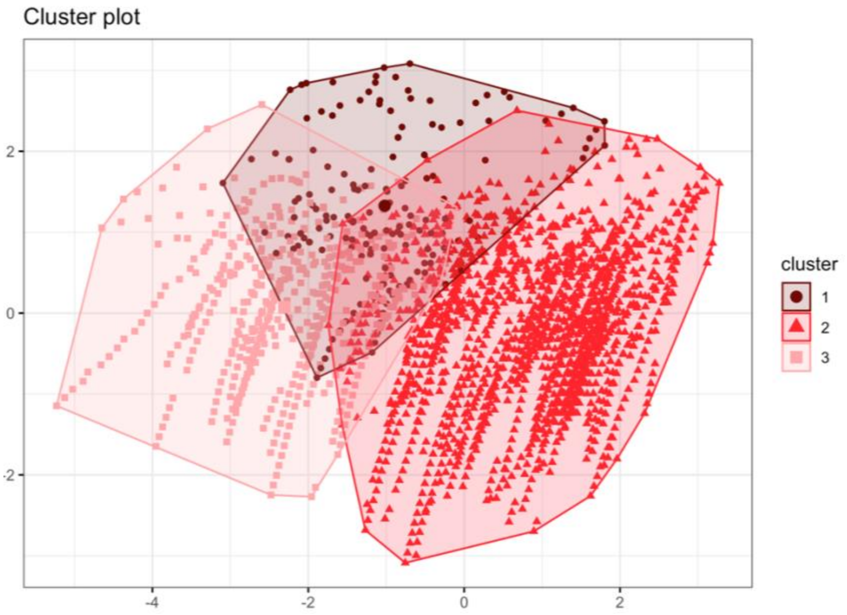


Figure 23

Further insights from hierarchical clustering can be summarized in the table below. Contrary to the results from K-means clustering, the cluster with the greatest number of videos in it, Cluster 2, is not the one with the lowest average values for views, likes, dislikes, and comments. Moreover, the cluster with the second lowest number of videos, Cluster 3, is also the one with the highest average view count, meaning that just a few videos get millions of views and thousands of likes/dislikes/comments. The average upload time for videos in Cluster 3, finally, is 4PM which is in line with the results from the exploratory data analysis (see 3.2.1).

Cluster:	**Number of Observations:**	**Views:**	**Likes:**	**Dislikes:**	**Comment Count:**	**Days to Trending:**	**Publish Time:**
1	159	85,0987.7	29,559.18	839.2453	3,352.484	3.553459	13.86164
2	1,497	625,699.4	13,651.19	505.5458	1,187.894	9.794255	14.63527
3	473	1,317,257.2	37,658.81	1,180.9408	2,893.742	12.205074	16.46723

Table 8

4.3 Regression

The aim of the regression analysis in this thesis is to find the best model to predict the number of views a video could reach. The techniques explained in Chapter 2 will be applied.

Before finding the best model, it is important to check the model assumptions. A linear regression model using all predictors, *fullmodel*, was built for this purpose.

1. Linearity: the fact that there is no patter in the ‘residual vs fitted’ plot (Figure 25) suggests that a linear relationship between the predictors and the outcome can be assumed.
2. Homoscedasticity: the scale-location plot (Figure 26) shows whether residuals are spread equally along the ranges of predictors. In this case, we do not have a linear horizontal line and neither equally spread points. We have heteroscedasticity.
3. Normality of residuals: we use the QQ plot of residuals (Figure 27) to visually check the normality assumption. The normal probability plot of residuals should approximately follow a straight line. In this case, all the points fall relatively along the reference line, hence we can assume normality.
4. Multicollinearity: even though no pair of variables has a particular high correlation, we check if collinearity exists. Any variable with a VIF value above 5 or 10 should be removed. In this case, no variables are removed (Table 11).

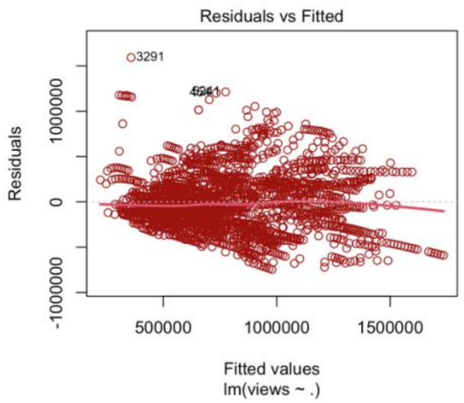


Figure 24

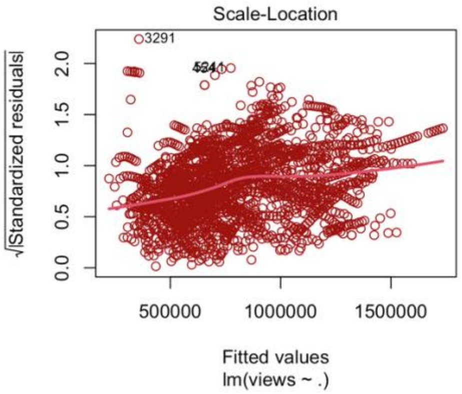


Figure 25

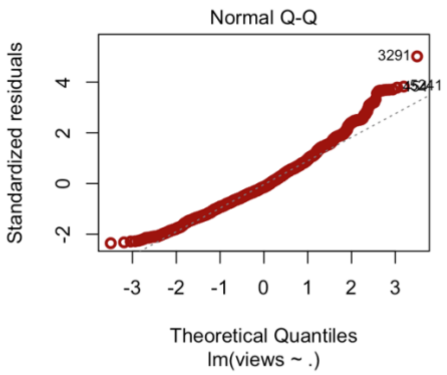
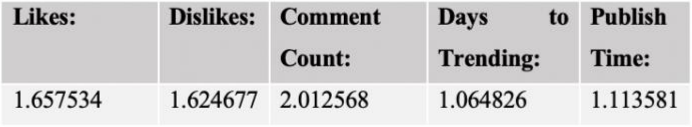


Figure 26



Likes:	Dislikes:	Comment Count:	Days to Trending:	Publish Time:
1.657534	1.624677	2.012568	1.064826	1.113581

Table 9

In order to evaluate the models built and choose the best among all, the analysis will use the mean squared error (MSE) criterion. The MSE tells how close the regression line is to the data points. The smaller the mean square error, the closer the predicted responses are to the true responses [30].

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \quad (11)$$

Finally, the dataset is split into training and test data: *trainset* has 1,490 observations and *testset* has the remaining 639 observations. Both training and test sets are then scaled.

4.3.1 Linear regression

To build a linear regression model in R, there is the function *lm()*. Let's consider again the model built to check the assumptions in the previous section. The multivariate model named *fullmodel* is built with all the features present in the training set and, if tested on *testset*, yields a MSE of 0.5260112. Moreover, all the variables considered seem to be significant, so we should not remove them from the model (see Table 12). An interesting insight we can derive from looking at the regression coefficients is that the one for variable *days_to_trending* is negative, which means that the longer it takes a video to appear on the Trending List, the lower the number of views it will get.

	Estimate:	**Std.Error:**	**t value:**	**Pr(>|t|):**	
(Intercept)	4.031e+05	2.497e+04	16.147	< 2e-16	***
Likes	9.509e+00	6.173e-01	15.405	< 2e-16	***
Dislikes	2.984e+02	1.954e+01	15.269	< 2e-16	***
Comment count	5.393e+01	8.178e+00	6.594	5.39e-11	***
Days to trending	9.519e+03	1.121e+03	8.494	< 2e-16	***
Publish time	-1.261e+04	1.541e+03	-8.186	4.60e-16	***

Table 10 - Model summary lm1

Thanks to the exploratory data analysis, it is possible to notice an interaction between the variables *comment_count* and *likes/dislikes* (respectively with interaction coefficients of 0.89 and 0.70). The presence of a significant interaction indicates that the effect of one attribute on the view count varies at different value of the other attribute. Starting from *lm1,* let's see if some interactions between these variables are statistically significant. After predicting the view count using *lm2* on the test data, the MSE is the lowest achieved for now: 0.4919779. The introduction of these interactions to the model also leads to higher regression coefficients. We can definitely see how the higher the number of likes/dislikes/comments, the higher the number of views a video will get.

	Estimate:	Std.Error:	t value:	Pr(>|t|):	
(Intercept)	0.03198	0.02383	1.342	0.179726	
Likes	0.38534	0.02570	14.997	< 2e-16	***
Dislikes	0.26270	0.02525	10.404	< 2e-16	***
Comment count	5.393e+01	0.02808	5.892	4.73e-09	***
Days to trending	0.12107	0.02020	5.994	2.57e-09	***
Publish time	-0.14853	0.02065	-7.193	9.99e-13	***
Likes*comment count	-0.13158	0.02312	-5.692	1.51e-08	***
Dislikes*comment count	0.07872	0.02379	3.309	0.000958	***

Table 11 - Model summary lm2

4.3.2 Random forests

The advantage of the random forest algorithm is that, at each splitting step, a random sample of predictors is chosen as split candidates from the full set of features. The number of random features selected can be set by the user or the choice can be left to the algorithm itself. Using the function *tuneRF()* is possible to visualize the optimal number of predictors. From Figure 30, the best number of variables randomly sampled is three, as the Out-Of-Bag (OOB) error is the lowest.

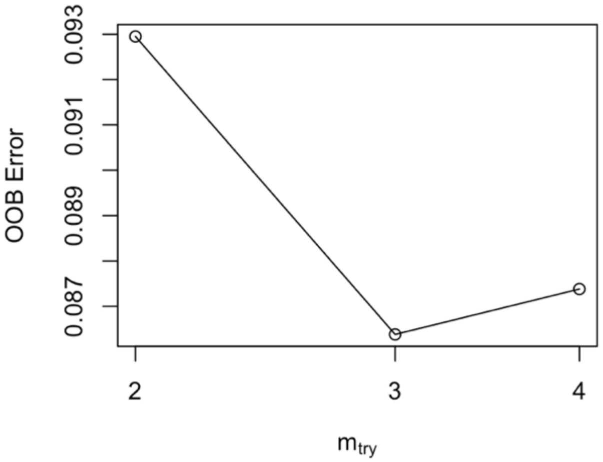


Figure 27

The random forest algorithm can be implemented via the *randomForest()* function from the homonymous library. After predicting the view count on the test data, the MSE is the lowest achieved in the entire analysis: 0.1703747.

As anticipated in Chapter 2, random forests can be used to rank the important of variables via a variable importance plot. As it can be seen in Figure 31, the variable *likes* is the most important one out of all the predictors

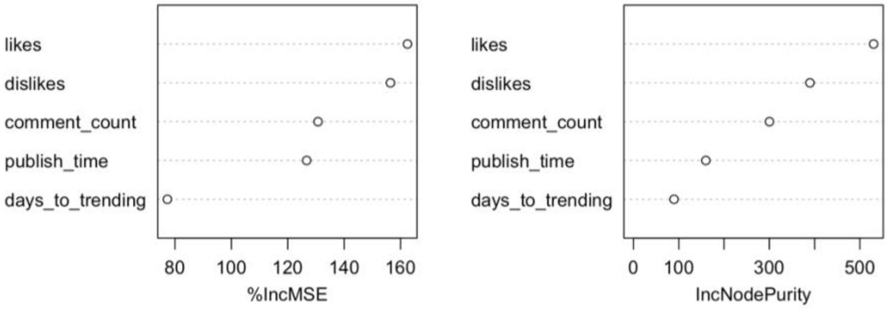


Figure 28

4.4 Summary

In Chapter 4, both clustering and regression analysis were analyzed. Here it a summary of the findings:

- K-mean clustering divided the data into three clusters.
- The best distance method to apply hierarchical clustering is complete linkage.
- Two multivariate linear regression models were built, the best of these was *lm2*, which achieved a test MSE of 0.4919779.
- With the implementation of random forests, it was possible to achieve an extremely low test MSE: 0. 1703747.
- From the variable importance plot, it can be concluded that the most important variable to predict the view count of a video is *likes*.

Finally, the main limitation in the research is that both clustering and regression analyses were performed on a subset of the original dataset, hence losing observations.

Conclusions

The thesis firstly carried out a detailed analysis and characterization of YouTube trending videos. By investigating the 40,949 observations in the dataset, it was possible to figure out some common patterns and characteristics among videos that have been featured on the Trending List. Some of the key findings were the best day and time to upload a new video, the recommendable length for titles and descriptions, and the best words to include in titles, descriptions, and tags.

Given the nature of some variables in the dataset, a clustering and regression analysis were conducted.

Clustering can be seen as an extension of the exploratory data analysis, as it allows to understand the data even further. The best clustering was obtained via the hierarchical technique with complete linkage, which resulted in having four clusters. By accessing information about these clusters, it was possible to understand that the data points were divided based on the views they got.

The regression analysis was helpful to predict the number of views for videos. Among the multivariate linear regression models built, the one with the interactions between *comment_count* and *likes/dislikes* achieved the lowest test MSE. Finally, the random forest algorithm was applied for making predictions. The random forest model achieved the lowest test MSE in the entire analysis.

The findings in this thesis can be of help to both creators and marketers. Indeed, the results achieved gave valuable information that they could apply to get videos featured on the Trending List, thus reaching a wider audience.

Reference list

[1] "Jawed Karim Success Story - How did he Founded YouTube," *StartupTalky*, Mar. 22, 2022. https://startuptalky.com/jawed-karim-youtube/ (accessed Mar. 29, 2022).

[2] R. Nieva, "YouTube started as an online dating site," *CNET*. https://www.cnet.com/tech/services-and-software/youtube-started-as-an-online-dating-site/ (accessed Mar. 28, 2022).

[3] "YouTube - Complete Guide: History, Products, Founding, and More," *History Computer*, Jan. 04, 2021. https://history-computer.com/youtube-history/ (accessed Mar. 28, 2022).

[4] "Top 10 youtube Competitors in 2022." https://whatcompetitors.com/youtube/ (accessed Mar. 31, 2022).

[5] "20 Interesting Vimeo Statistics," *DMR*, Mar. 13, 2016. https://expandedramblings.com/index.php/vimeo-statistics/ (accessed Apr. 01, 2022).

[6] "11 Interesting Dailymotion Facts and Statistics," *DMR*, Mar. 08, 2018. https://expandedramblings.com/index.php/dailymotion-facts-statistics/ (accessed Apr. 01, 2022).

[7] "Twitch Statistics 2022: How Many People Use Twitch?," Mar. 17, 2022. //earthweb.com/twitch-statistics/ (accessed Apr. 01, 2022).

[8] R. Team, "Instagram IGTV: 50 Mind-Blowing Instagram IGTV Stats For 2022," May 26, 2020. https://www.reelnreel.com/instagram-igtv-stats/ (accessed Apr. 01, 2022).

[9] "TikTok Revenue and Usage Statistics (2022)," *Business of Apps*, Jan. 10, 2019. https://www.businessofapps.com/data/tik-tok-statistics/ (accessed Apr. 01, 2022).

[10] "YouTube User Statistics 2022 | Global Media Insight," *Official GMI Blog*. https://www.globalmediainsight.com/blog/youtube-users-statistics/ (accessed Apr. 01, 2022).

[11] P. Leskin, "From David Dobrik to The Rock, these are teens' favorite people to follow on YouTube, Instagram, and Twitter," *Business Insider*. https://www.businessinsider.com/teens-favorite-influencers-celebrities-david-dobrik-kylie-jenner-2019-10 (accessed Apr. 04, 2022).

[12] A. Brown, "The Highest-Paid YouTube Stars: MrBeast, Jake Paul And Markiplier Score Massive Paydays," *Forbes*. https://www.forbes.com/sites/abrambrown/2022/01/14/the-highest-paid-youtube-stars-mrbeast-jake-paul-and-markiplier-score-massive-paydays/ (accessed Apr. 04, 2022).

[13] C. Taylor, "Kids now dream of being professional YouTubers rather than astronauts, study finds," *CNBC*, Jul. 19, 2019. https://www.cnbc.com/2019/07/19/more-children-dream-of-being-youtubers-than-astronauts-lego-says.html (accessed Apr. 04, 2022).

[14] G. Stocking, P. V. Kessel, M. Barthel, K. E. Matsa, and M. Khuzam, "Many Americans Get News on YouTube, Where News Organizations and Independent Producers Thrive Side by Side," *Pew Research Center's Journalism Project*, Sep. 28, 2020. https://www.pewresearch.org/journalism/2020/09/28/many-americans-get-news-on-youtube-where-news-organizations-and-independent-producers-thrive-side-by-side/ (accessed Apr. 05, 2022).

[15] A. Antonio and D. Tuffley, "YouTube a valuable educational tool, not just cat videos," *The Conversation*. http://theconversation.com/youtube-a-valuable-educational-tool-not-just-cat-videos-34863 (accessed Apr. 06, 2022).

[16] "The New Era of Marketing: The Power of YouTube influencers," *Digital Glue*, Mar. 08, 2020. https://digitalglue.agency/youtube-influencers/ (accessed Apr. 03, 2022).

[17] "Making YouTube Better in a Mobile, Cross-Screen World," *Google*, Jan. 20, 2017. https://blog.google/products/ads/making-youtube-better-in-mobile-cross/ (accessed Apr. 08, 2022).

[18] W. Geyser, "What is Influencer Marketing? - The Ultimate Guide for 2022," *Influencer Marketing Hub*, Nov. 01, 2016. https://influencermarketinghub.com/influencer-marketing/ (accessed Apr. 07, 2022).

[19] "Celebrating what you created, watched and shared in 2015 with #YouTubeRewind and YouTube's new Trending tab," *blog.youtube*. https://blog.youtube/culture-and-trends/youtube-rewind-2015/ (accessed Apr. 08, 2022).

[20] "Trending on YouTube - YouTube Help." https://support.google.com/youtube/answer/7239739?hl=en (accessed Apr. 06, 2022).

[21] TeamYouTube [@TeamYouTube], "The Trending tab in the YouTube app is becoming the Explore tab. You'll now find all of this👇 in □place: 📍Destination pages for Gaming, Learning & more 📍A more prominent spot for Creators & Artists on the Rise 📍What's trending on YouTube right now https://yt.be/help/QTr6 https://t.co/Rv496UOOtz," *Twitter*, Mar. 12, 2020. https://twitter.com/TeamYouTube/status/1238162785072066562 (accessed Apr. 08, 2022).

[22] S. Ouyang, C. Li, and X. Li, "A Peek Into the Future: Predicting the Popularity of Online Videos," *IEEE Access*, vol. 4, pp. 3026–3033, 2016, doi: 10.1109/ACCESS.2016.2580911.

[23] C. Li, J. Liu, and S. Ouyang, "Characterizing and Predicting the Popularity of Online Videos," *IEEE Access*, vol. 4, pp. 1630–1641, 2016, doi: 10.1109/ACCESS.2016.2552218.

[24] I. Barjasteh, Y. Liu, and H. Radha, "Trending Videos: Measurement and Analysis," p. 16.

[25] S. Gayakwad, R. Patankar, and D. Mane, "Analysis on YouTube Trending Videos," vol. 07, no. 08, p. 7, 2020.

[26] J. F. Andry, S. A. Reynaldo, K. Christianto, F. S. Lee, J. Loisa, and A. B. Manduro, "Algorithm of Trending Videos on YouTube Analysis using Classification, Association and Clustering," in *2021 International Conference on Data and Software Engineering (ICoDSE)*, Bandung, Indonesia, Nov. 2021, pp. 1–6. doi: 10.1109/ICoDSE53690.2021.9648486.

[27] G. Feroz Khan and S. Vong, "Virality over YouTube: an empirical analysis," *Internet Res.*, vol. 24, no. 5, pp. 629–647, Sep. 2014, doi: 10.1108/IntR-05-2013-0085.

[28] "YouTube users by country 2022," *Statista*. https://www.statista.com/statistics/280685/number-of-monthly-unique-youtube-users/ (accessed Apr. 19, 2022).

[29] P. L. Haasch Palmer, "These are the 30 most popular YouTube stars in the world, from PewDiePie to Ryan Kaji," *Business Insider*. https://www.businessinsider.com/most-popular-youtubers-with-most-subscribers-2018-2 (accessed Apr. 19, 2022).

[30] G. James, D. Witten, T. Hastie, and R. Tibshirani, Eds., *An introduction to statistical learning: with applications in R*. New York: Springer, 2013.

[31] "What is Unsupervised Learning?," Mar. 25, 2022. https://www.ibm.com/cloud/learn/unsupervised-learning (accessed Apr. 11, 2022).

[32] R. Nainggolan, R. Perangin-angin, E. Simarmata, and A. F. Tarigan, "Improved the Performance of the K-Means Cluster Using the Sum of Squared Error (SSE) optimized by using the Elbow Method," *J. Phys. Conf. Ser.*, vol. 1361, no. 1, p. 012015, Nov. 2019, doi: 10.1088/1742-6596/1361/1/012015.

[33] T. Hastie, R. Tibshirani, and J. Friedman, "Unsupervised Learning," in *The Elements of Statistical Learning*, New York, NY: Springer New York, 2009, pp. 485–585. doi: 10.1007/978-0-387-84858-7_14.

[34] "What is Supervised Learning?," Jun. 30, 2021. https://www.ibm.com/cloud/learn/supervised-learning (accessed Apr. 28, 2022).

[35] "Linear Regression." http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm (accessed Apr. 28, 2022).

[36] M. A. Poole and P. N. O'Farrell, "The Assumptions of the Linear Regression Model," *Trans. Inst. Br. Geogr.*, no. 52, p. 145, Mar. 1971, doi: 10.2307/621706.

[37] T. Hastie, R. Tibshirani, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction: with 200 full-color illustrations*. New York: Springer, 2001.

[38] "Trending YouTube Video Statistics | Kaggle." https://www.kaggle.com/datasets/datasnaek/youtube-new?select=USvideos.csv (accessed Apr. 11, 2022).

[39] E. Rahm and H. H. Do, "Data Cleaning: Problems and Current Approaches," p. 12.

[40] E. de Jonge, "An introduction to data cleaning with R," p. 53.

[41] "What is Exploratory Data Analysis?," Jul. 06, 2021. https://www.ibm.com/cloud/learn/exploratory-data-analysis (accessed Apr. 12, 2022).

[42] "YouTube Video Title," *Backlinko*. https://backlinko.com/hub/youtube/title (accessed Apr. 12, 2022).

[43] "Discovery and performance FAQs - YouTube Help." https://support.google.com/youtube/answer/141805?hl=en (accessed Apr. 12, 2022).

[44] TechPostPlus, “YouTube Video Categories List (Complete Guide),” *TechPostPlus*, Apr. 26, 2019. https://techpostplus.com/youtube-video-categories-list-faqs-and-solutions/ (accessed Apr. 12, 2022).

[45] “Add tags to videos - YouTube Help.” https://support.google.com/youtube/answer/146402?hl=en (accessed Apr. 13, 2022).

[46] “Understand audience engagement - Computer - YouTube Help.” https://support.google.com/youtube/answer/9313698?hl=en (accessed Apr. 12, 2022).

[47] K. Carmicheal, “How Does YouTube Count Views? We Break It Down.” https://blog.hubspot.com/marketing/how-does-youtube-count-views (accessed Apr. 12, 2022).

[48] “An update to dislikes on YouTube,” *blog.youtube*. https://blog.youtube/news-and-events/update-to-youtube/ (accessed Apr. 13, 2022).

[49] “Logan Paul,” *Wikipedia*. Apr. 08, 2022. Accessed: Apr. 13, 2022. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Logan_Paul&oldid=1081541864

[50] N. Braca, “Logan Paul Returns to YouTube With a Video About Suicide Prevention & Education,” *Billboard*, Jan. 24, 2018. https://www.billboard.com/music/music-news/logan-paul-returns-to-youtube-suicide-video-8096027/ (accessed Apr. 13, 2022).

[51] “Mention channels in video titles & descriptions - Android - YouTube Help.” https://support.google.com/youtube/answer/9637404?hl=en&co=GENIE.Platform%3DAndroid (accessed Apr. 13, 2022).

[52] “Use hashtags for video search - YouTube Help.” https://support.google.com/youtube/answer/6390658?hl=en (accessed Apr. 13, 2022).

[53] “Add video thumbnails on YouTube - YouTube Help.” https://support.google.com/youtube/answer/72431#strategies-zippy-link-3&zippy=%2Cnew-videos (accessed Apr. 12, 2022).

[54] D. James, “Why Is a YouTube Thumbnail Important?,” *tubefluence*, Sep. 27, 2020. https://tubefluence.com/why-is-a-youtube-thumbnail-important/ (accessed Apr. 21, 2022).

[55] “Detection Classes — ImageAI 2.1.6 documentation.” https://imageai.readthedocs.io/en/latest/detection/index.html (accessed Apr. 22, 2022).

[56] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 779–788. doi: 10.1109/CVPR.2016.91.

[57] P. Adarsh, P. Rathi, and M. Kumar, “YOLO v3-Tiny: Object Detection and Recognition using one stage improved model,” in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Coimbatore, India, Mar. 2020, pp. 687–694. doi: 10.1109/ICACCS48705.2020.9074315.

[58] R. singh, “It’s all about Outliers,” *Analytics Vidhya*, Aug. 31, 2020. https://medium.com/analytics-vidhya/its-all-about-outliers-cbe172aa1309 (accessed Apr. 22, 2022).

[59] G. J. G. Upton and I. Cook, *Understanding statistics*, 1. publ. Oxford: Oxford University Press, 1996.

[60] K. Rink, “Four mistakes in Clustering you should avoid,” *Medium*, Apr. 29, 2021. https://towardsdatascience.com/common-mistakes-in-cluster-analysis-and-how-to-avoid-them-eb960116d773 (accessed Apr. 26, 2022).

[61] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. 2008.