**Assessment 1b**

Predicting Video posting success on YouTube

## **Data description:**

I have sourced a few YouTube channels of a similar audience to increase the size of the dataset. This introduces more noise and limits the capacity for removal of noise specific to Chloe (eg, viewing spikes during times of controversy, advertisement of video on other social media channels, etc.). However, in doing so the model is expected to have better accuracy and generalizability (Zhou et al. 2022).

To ensure we have balanced target variables and to simplify, the question has been refined as follows:

*Using features available at the time of upload, is it possible to predict whether a video will have a like per view ratio above or below the 50th percentile with moderate accuracy (70-80%)?*

I will continue to seek similar channels to increase the dataset as well, and in doing so I may reach a size that allows me to be more selective with the percentile split.

The YouTube algorithm is a black box algorithm that many have attempted to tease clues from (Bishop 2019). However, optimizing YouTube uploads by mimicking previous success may result in improved like to view ratios.

The features of most relevance are (see part 1a for more details):

Table 1: Features obtained in assignment part 1a to be used included in future models.

|  |  |
| --- | --- |
| **Numerical Features** | **Explanation** |
| Title length | The number of words in title. |
| Description length | The number of words in description |
| Number of tags | The number of tags provided |
| Tags in title | The number of tags also present in the video title |
| Emoji count – title | The number of emojis present in the title. |
| Emoji count description | The number of emojis used in the description. |
| Description sentiment | The average VADER sentiment of the description. |
| Title sentiment | The average VADER sentiment of the title. |
| Duration in seconds | The duration of the video in seconds, video shorts have been removed (videos less than 61 seconds). |
| **Categorical Features** | **Explanation** |
| Publish day | The specific weekday a video was uploaded on (one-hot encoded). |
| Publish month | The specific month a video was uploaded (one-hot encoded). |
| **Target Variable** | **Explanation** |
| Likes per view | The like to view ratio a given video has at a given point in time. |

These features have been selected because they are all accessible/manipulatable prior to video upload and are therefore useful in making a meaningful prediction. Outliers have been removed were appropriate, and skewness and shape have been visualized for each variable (*Appendix*).

## **Clustering/Pattern:**

### *Numerical features*

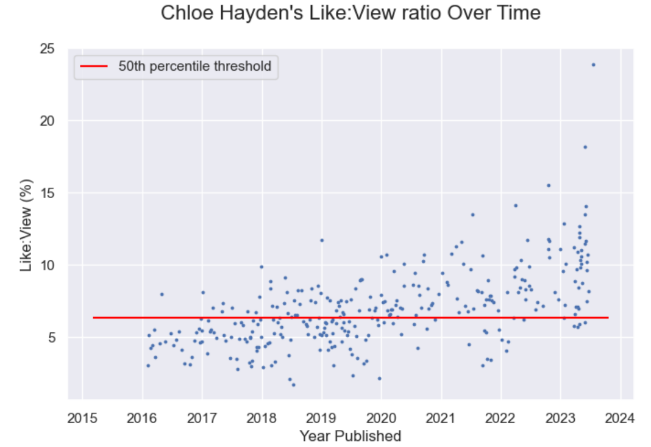
There were no obvious linear association observed between the target variable and any predictor variable (*figure 4*). However, it must be said that over time an individual youtuber tends to see growth in their like:view ratio (*figure 1*) (Page & Lopatka 1999). An ‘*Upload Count’ feature* may capture some of the information without directly introducing time. With few creators, it is unlikely to hold much weight and has been excluded to avoid overfitting for now. To demonstrate feature relevance, spearman’s rank coefficient was used and in all cases the p-value was less than 0.05 with weakly negative association existing between it and the target variable (Appendix).

Figure 1: Scatter plot demonstrating the moderate linear trend between time and like:view ratio observed for Chloe Hayden.

PCA was also conducted and identified that 95% of the variance could be explained by the first principal component. The loadings for this component put an emphasis on the importance of ‘Duration in seconds (*Figure 5*). However, PCA operates best on more linear data (Sumithra & Surendran 2015).

Two component and three component t-SNE were conducted to determine if unsupervised techniques were able to determine the target variable (*Figure 6 & 7*). Groupings were not uniformly separated by any categorical feature including the target variable. This was mirrored in the UMAP (*Figure 8*). The project’s success is therefore dependent on the detection of interactions and non-linear patterns which is consistent with similar research (Batta, Murthy & Savitri 2022; Halim, Hussain & Ali 2022).

### *Categorical features*

An association exists with upload times and the target variable as the p-values are less than 0.05 (this was not the case when we included YouTube shorts). This association is likely behavioural, as YouTubers have a preferred upload schedule, or perhaps holiday periods and relevant events in certain months have an influence.

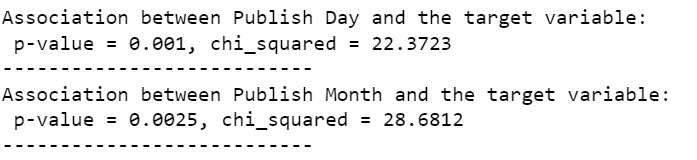


Figure 2: Snippet of the chi-squared test results for the association between time categories and like:view ratio.

It was also found that there was a strong association between the most common tags and the day uploaded:



Figure 3: Snippet of the chi-squared test results taken for the association between the most common tags (non-username tags in the top 25) and like:view ratios

However, it was decided that we are uninterested in the specifics of the video on this occasion as we do not have enough representation of each tag. By removing the ’tags’ feature, we may avoid overfitting. If creators were interested in what content specifically has optimal like to view ratios, a more extensive project could be undertaken.

Some thought has been put into the boundaries of the Publish Day. Currently, they are in Australian time. This is informative to Chloe so it shall remain as is, but there is the argument that the time zone could be place in the region of most YouTube users because it may better showcase the general behaviour of the YouTube algorithm in relation to time and would give translatable information that could be applied in other time zones. Alternatively, if user location was known, we could ‘flatten’ time and bring all instances into a shared pseudo time zone.

## **Visualizations**

### **Correlation Matrix (Pearson correlation )**

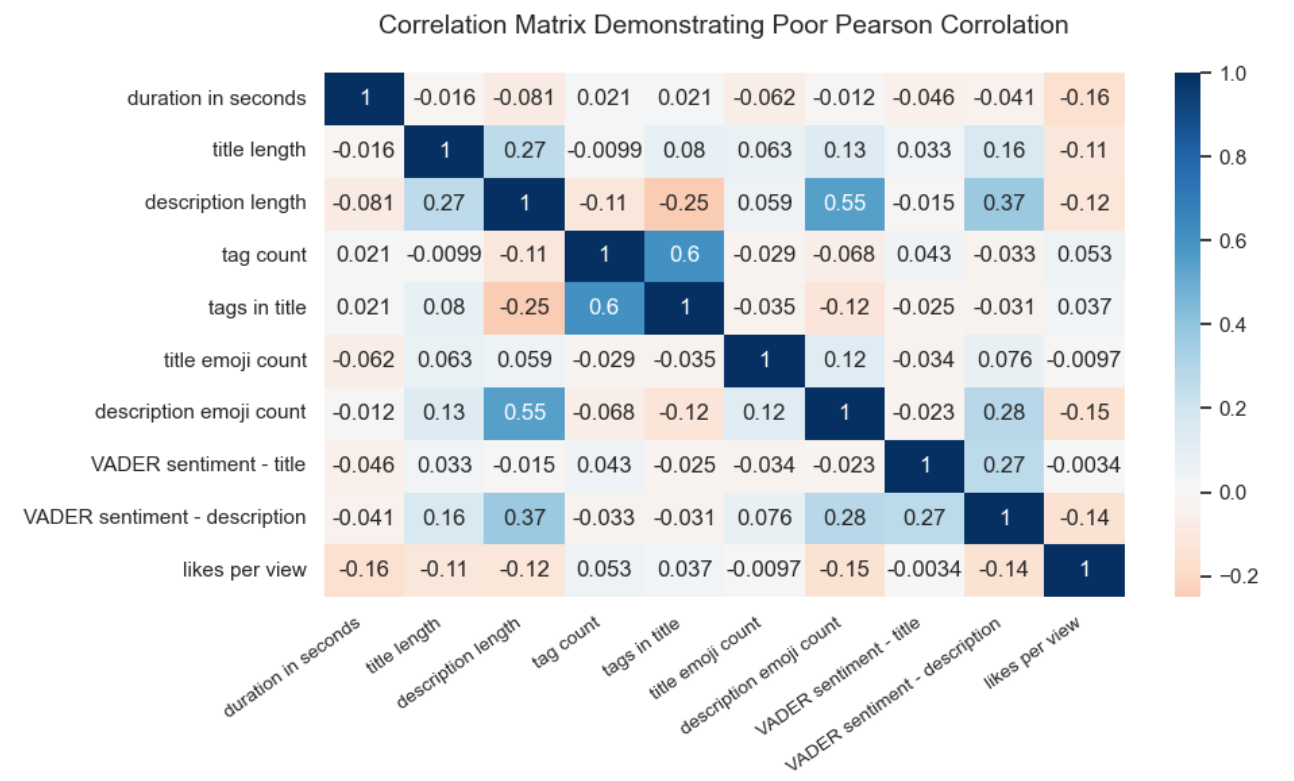


Figure 4: The correlation matrix demonstrates in the last column that no feature can be said to correlate well linearly with the like:view ratio.

### **PCA**

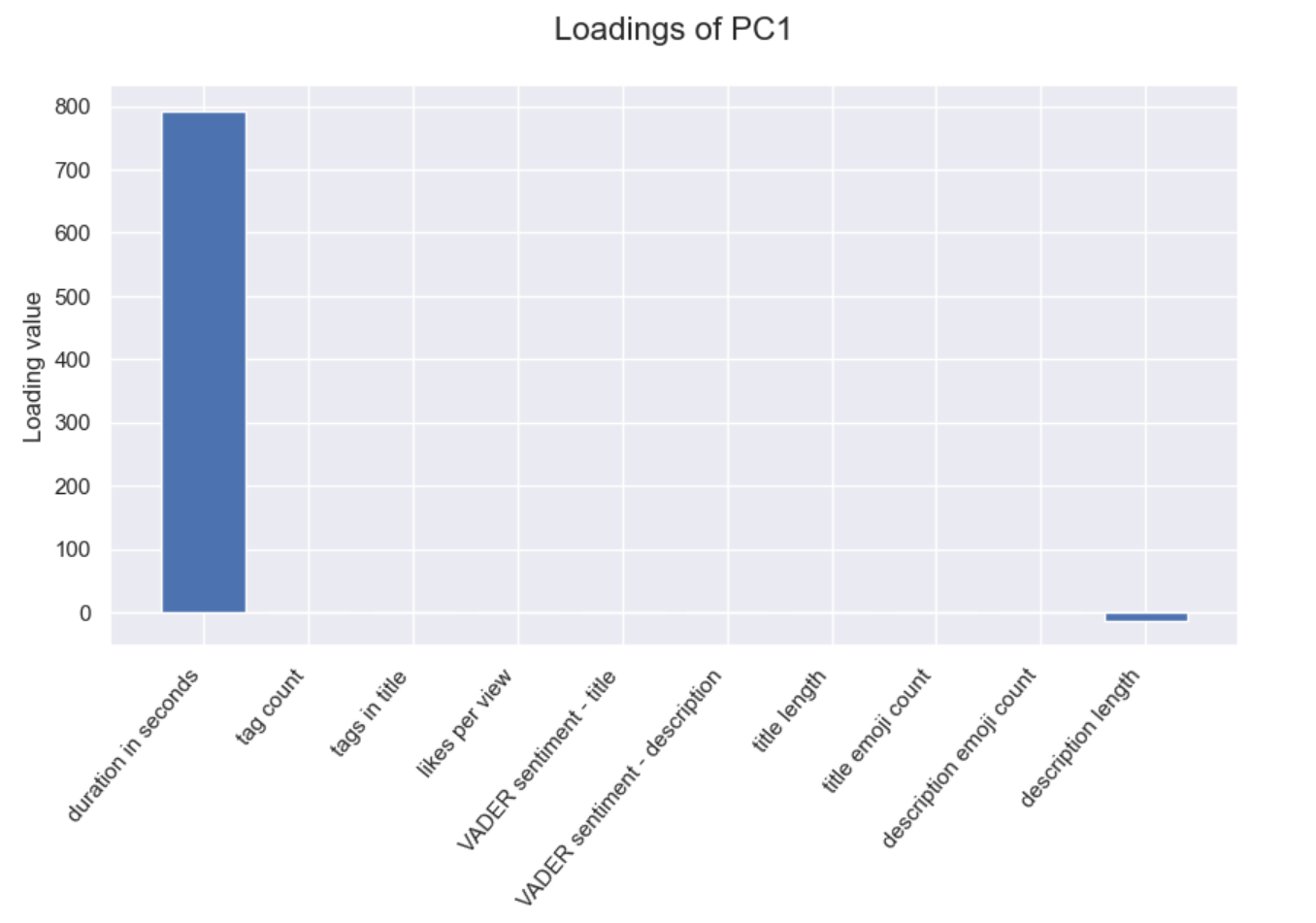
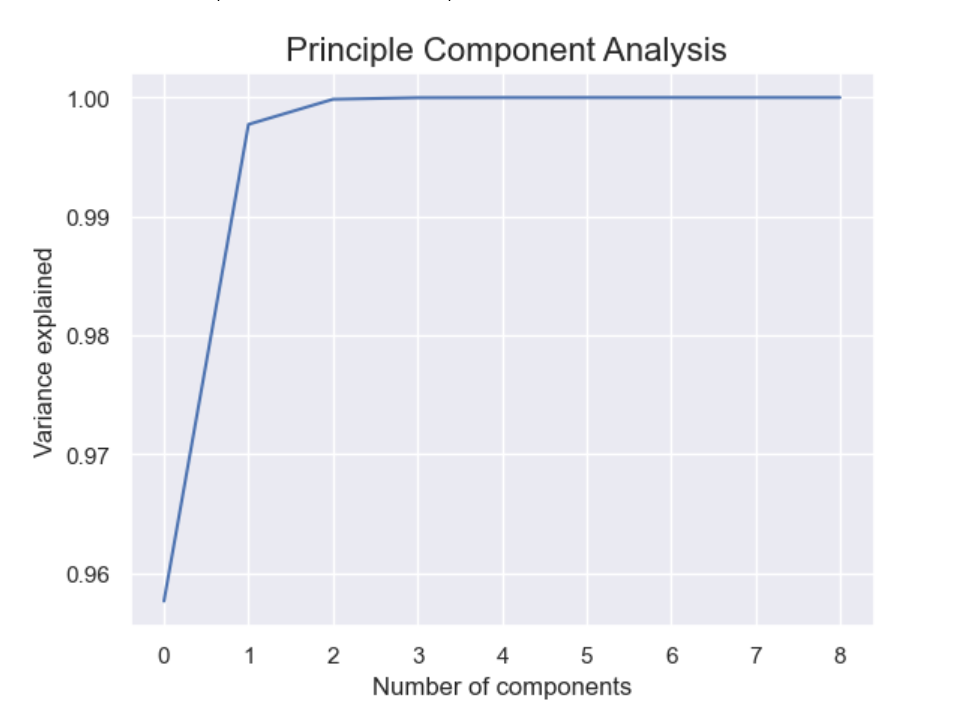


Figure 5: The variance explained by each principal component in the PCA (left) and the loadings observed for the first Principal component which explained at least 95% of the variation (right).

### **t-SNE (2 component)**

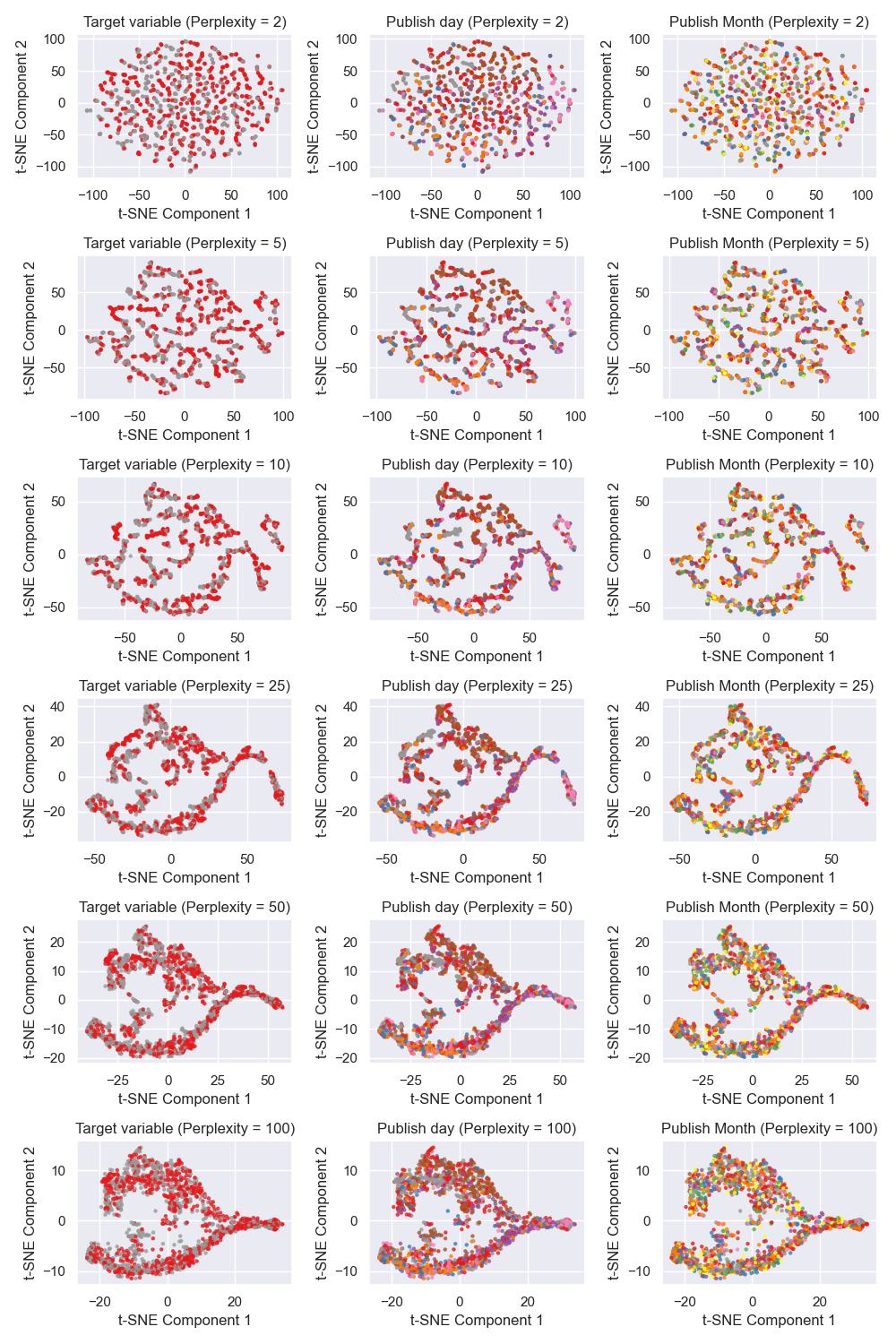


Figure 6: Dimension reduction with t-SNE was used to observe non-linear patterns in the data. This is a snippet of the informative section of the figure (2 components over a range of perplexity values, 1000 iterations was just as good as 10,000).

### **t-SNE (3 component)**

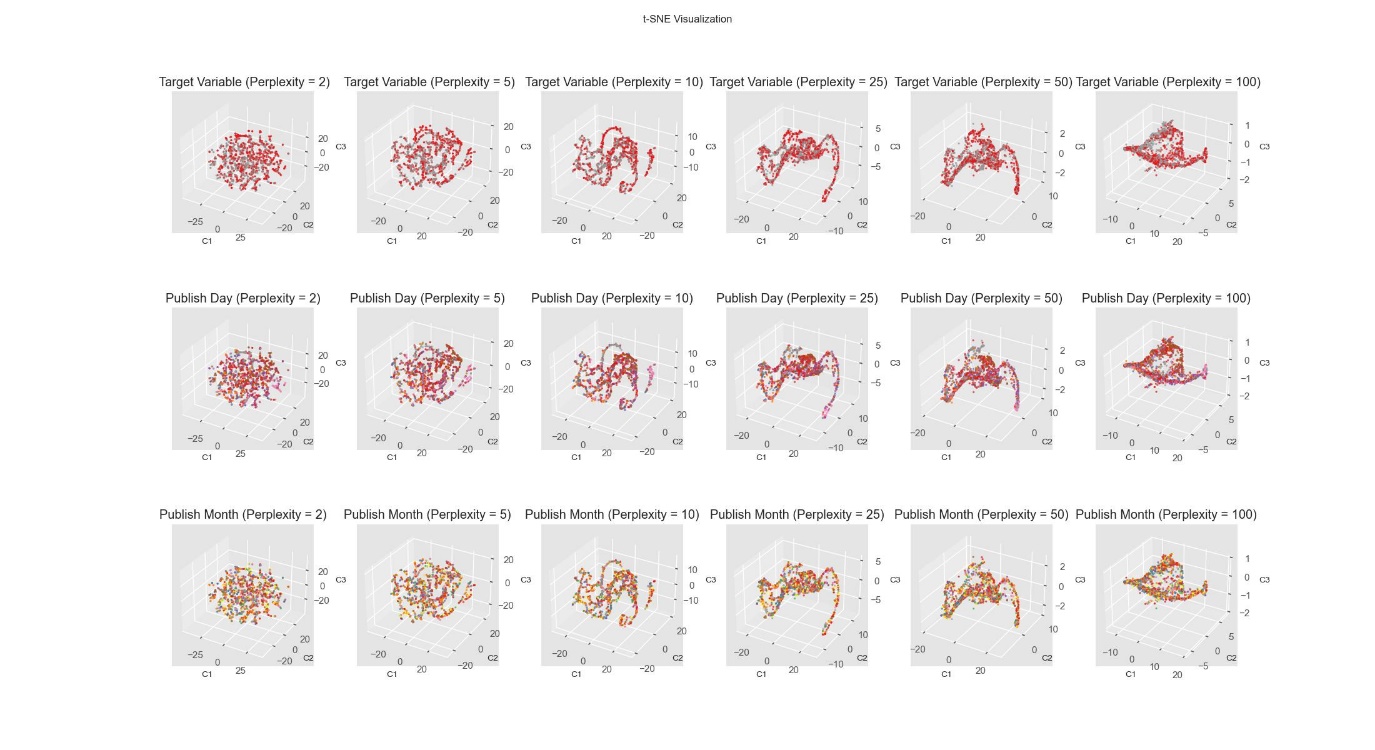


Figure 7: Three component t-SNE was completed to better understand patterns in the data.

### **UMAP (2 components)**

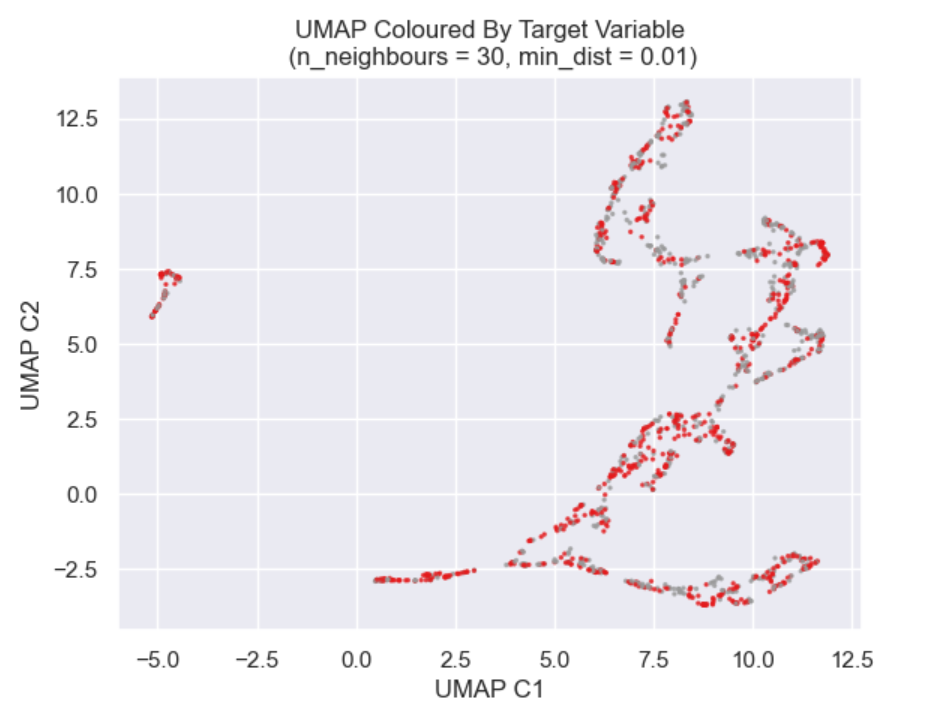


Figure 8: UMAP is similar to t-SNE, but it is known to better represent the shape of the data. A range of hyperparameters were tested but no improvement was observed.

## **Outlook**

Decision trees are efficient at finding patterns in non-linear data (Charbuty & Abdulazeez 2021). Although unsupervised t-SNE and UMAP techniques could not split the data by the target label, a supervised model may still do so with good accuracy. The nature of classification trees is such that skewness, data type, and outliers do not necessarily pose a significant problem (Greenwell 2022). They operate by iteratively splitting features to find features and values that result in an improvement of some metric (e.g., the Gini impurity, entropy, or information gain). The leaves that remain at the end democratically elect an outcome/label. The unsupervised models were ineffective, and perhaps this is an indicator that the problem is not a simple one to model with the features available. Some extra variables are still under consideration (upload count, thumbnail CNN, pseudo time zone amendment, etc.), however early modelling attempts show promise with the resources available.

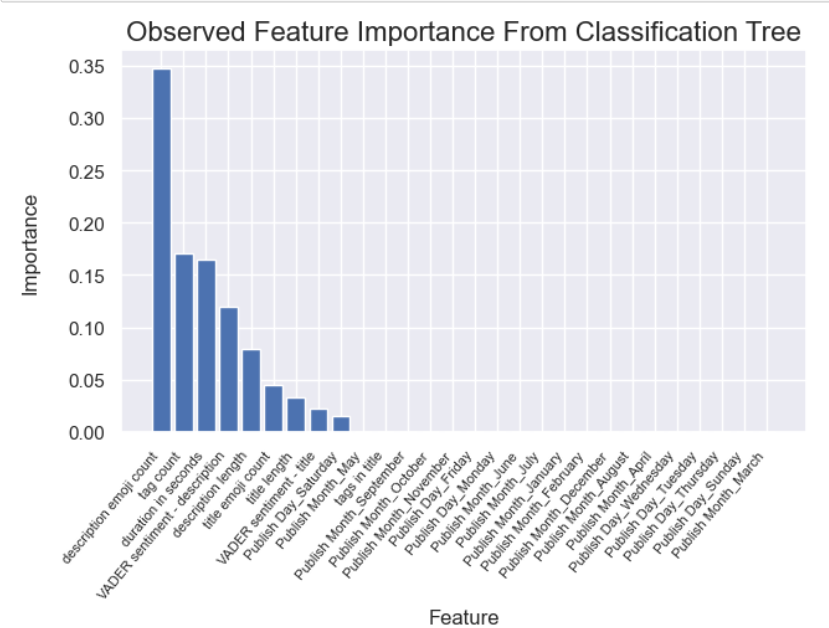


Figure 9: Variable importance from basic classification model with accuracy of 0.674.

## **References**

Batta, H, Murthy, AV & Savitri, S 2022, 'Predicting Popularity of YouTube videos using Viewer Engagement Features', in *12th International Conference on Cloud Computing, Data Science and Engineering* IEEE, India, pp. 77-81.

Bishop, S 2019, 'Managing visibility on YouTube through algorithmic gossip', *New Media &amp; Society*, vol. 21, no. 11-12, pp. 2589-2606.

Charbuty, B & Abdulazeez, A 2021, 'Classification Based on Decision Tree Algorithm for Machine Learning', *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20-28.

Greenwell, BM 2022, *Tree-Based Methods for Statistical Learning in R*, CRC Press.

Halim, Z, Hussain, S & Ali, RH 2022, 'Identifying content unaware features influencing popularity of videos on youtube: A study based on seven regions', *Expert Systems with Applications*, vol. 206, p. 117836.

Page, WH & Lopatka, JE 1999, 'Network externalities', *Encyclopedia of law and economics*, vol. 760, pp. 952-980.

Sumithra, V & Surendran, S 2015, 'A review of various linear and non linear dimensionality reduction techniques', *Int. J. Comput. Sci. Inf. Technol*, vol. 6, no. 3, pp. 2354-2360.

Zhou, K, Liu, Z, Qiao, Y, Xiang, T & Loy, CC 2022, 'Domain Generalization: A Survey', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 1-20.

# Appendix

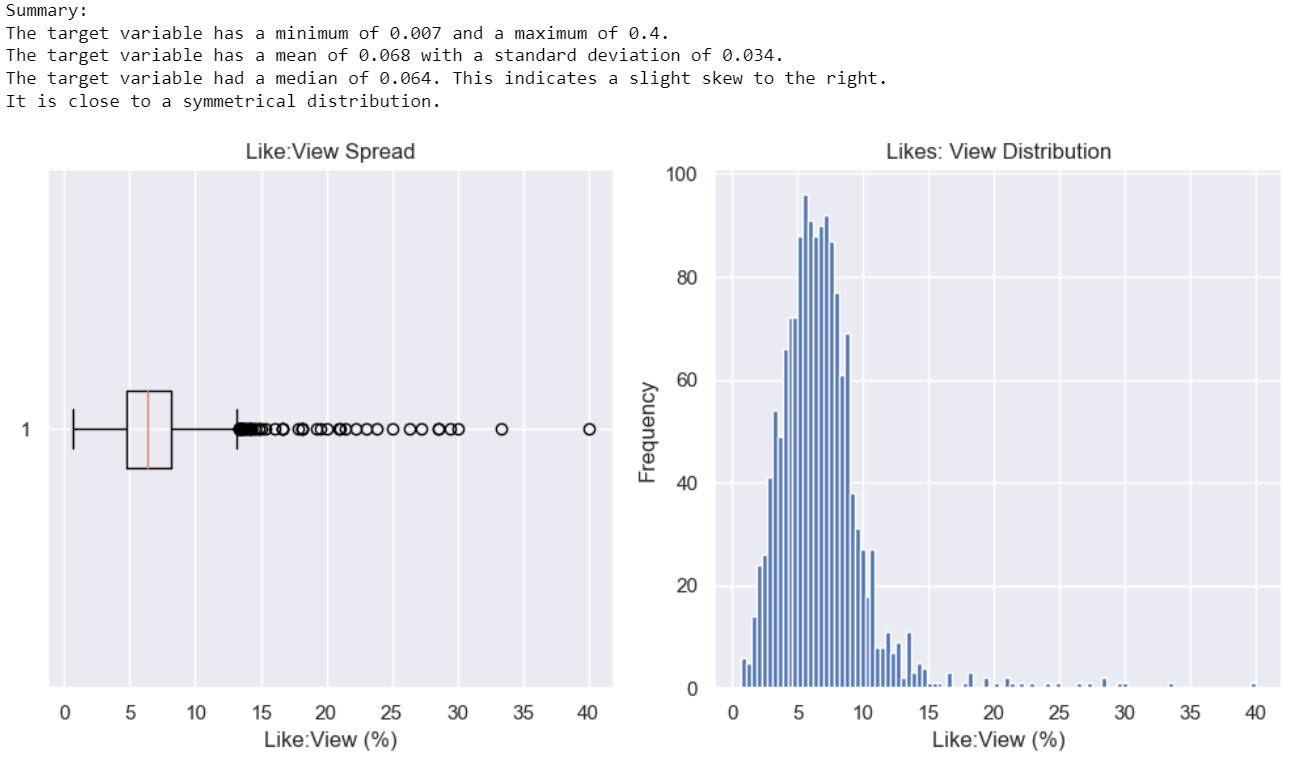


Figure 10: Snippet of target variable summary.

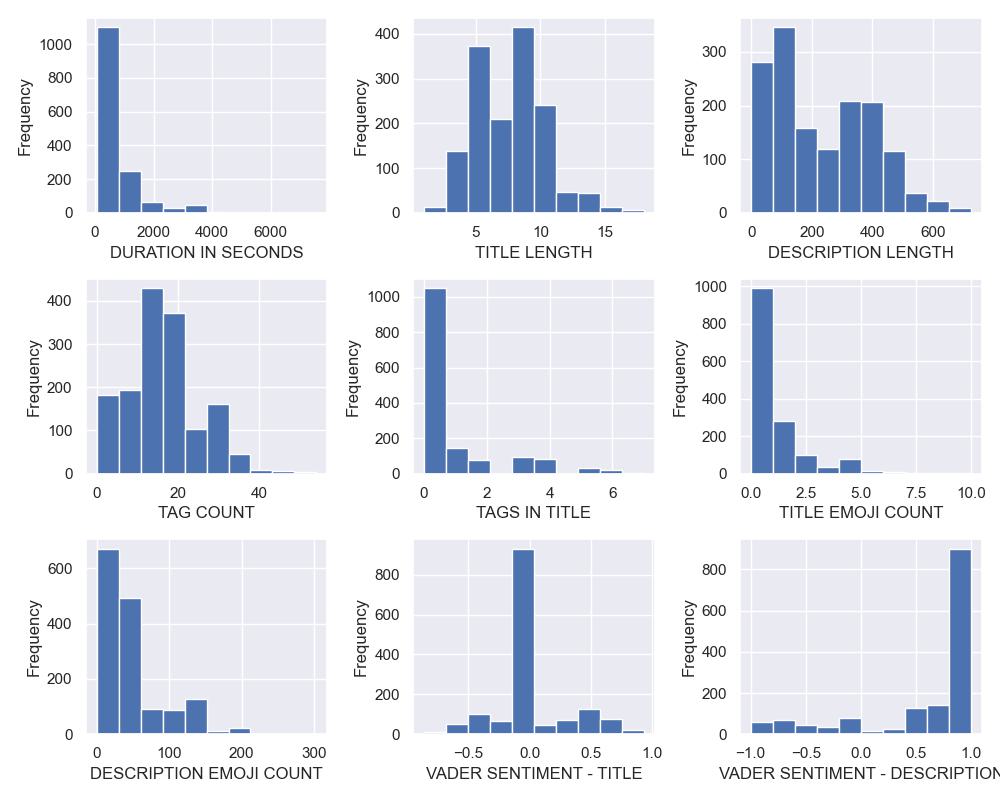


Figure 11: Histograms of numerical features.

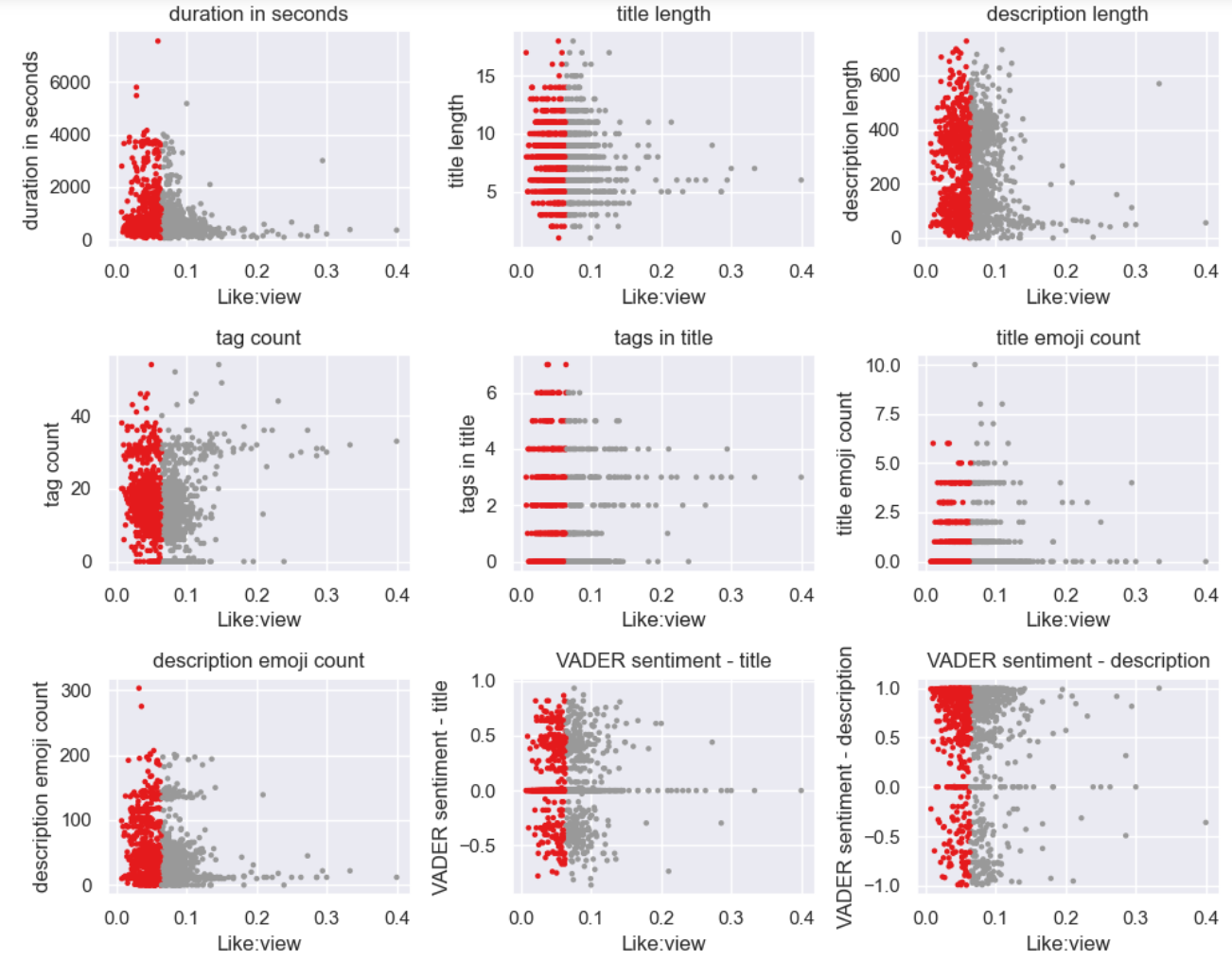


Figure 12: Observing relationships between target variable and numerical variables. Red indicates videos in the 50th percentile or lower.

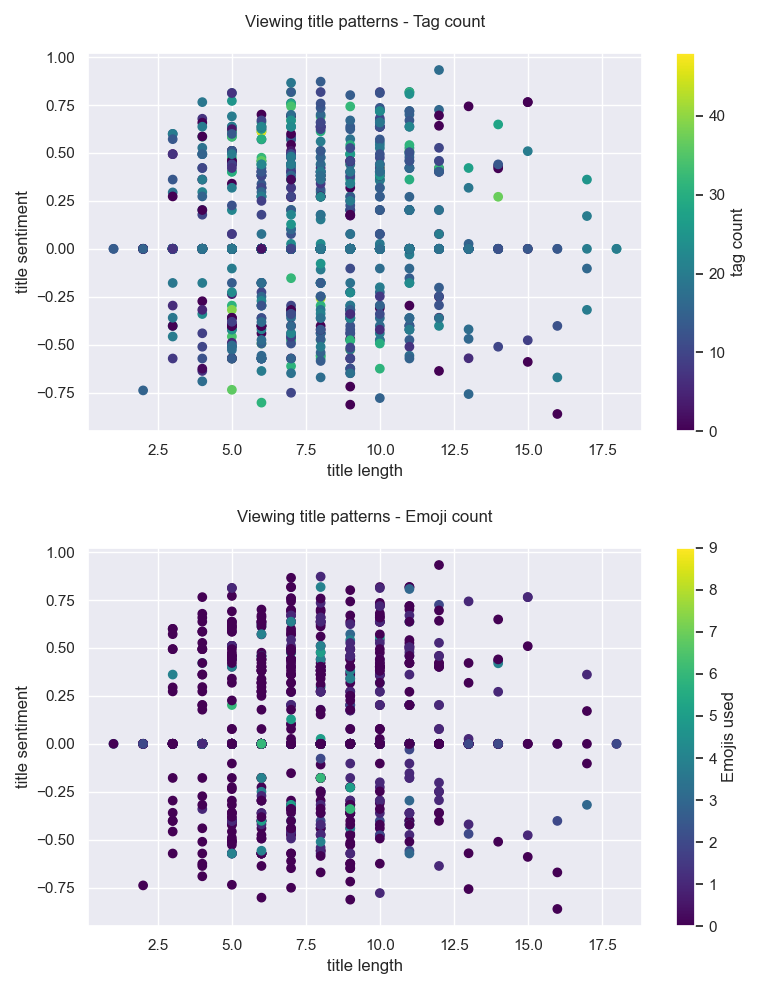
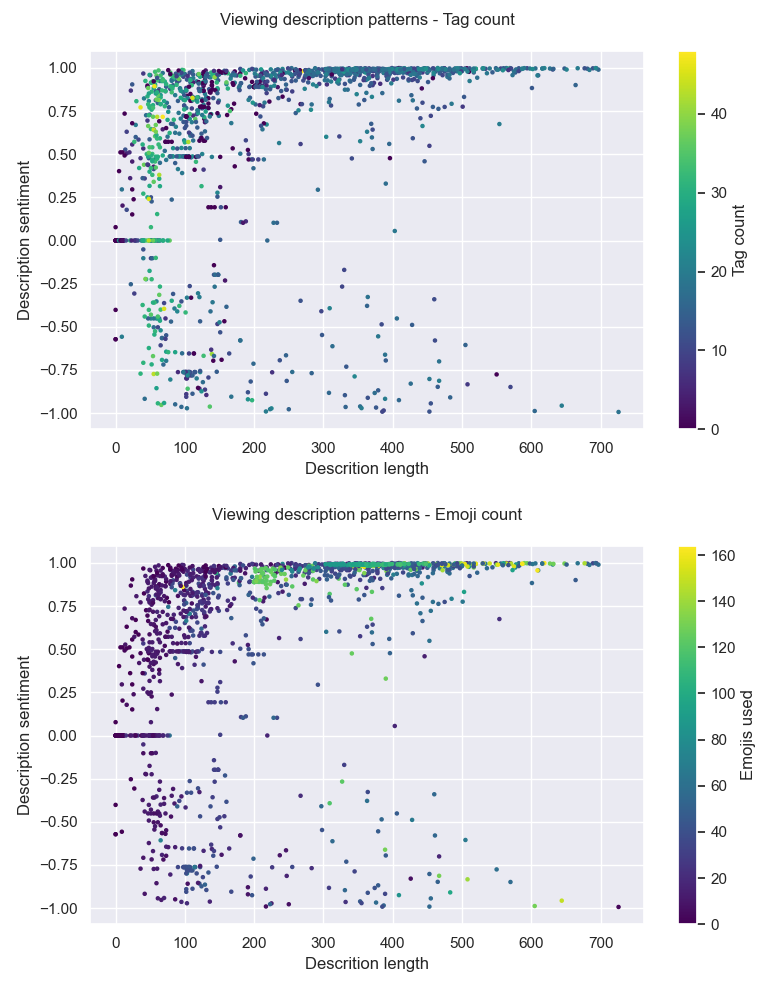


Figure 13: Multivariate analysis of description and title patterns.

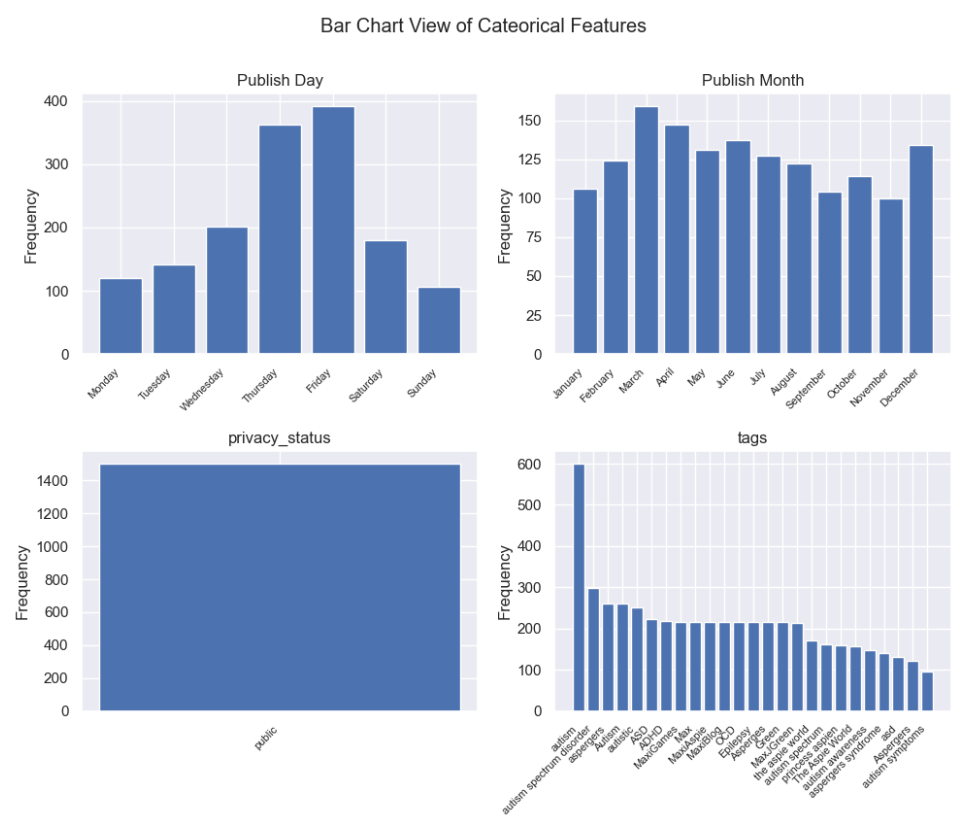


Figure 14: Univariate analysis of categorical variables. Privacy status was removed as it only had one category, and only the top 25 most common tags were viewed.

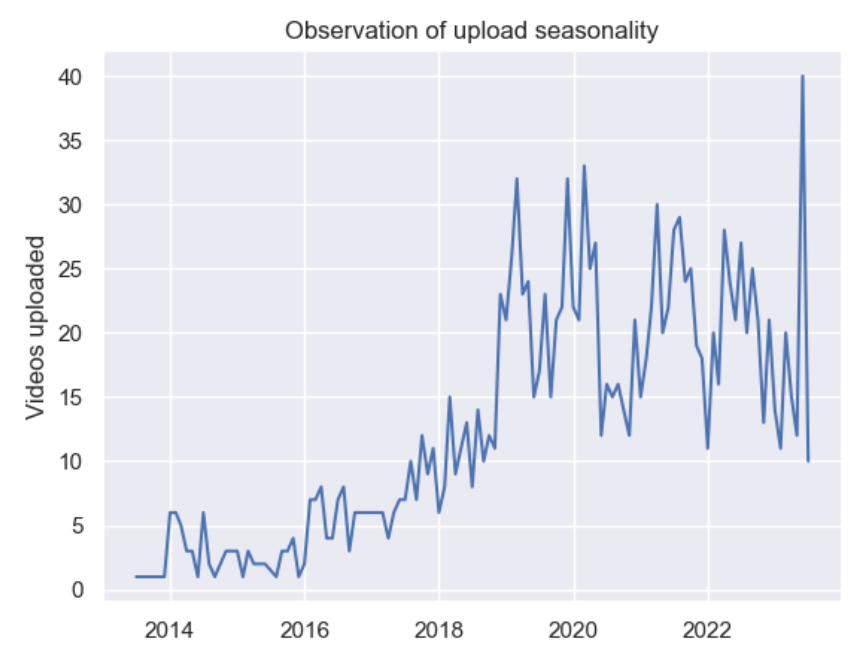


Figure 15: If any seasonality exists, the patterns are complex and difficult to unpack.