YouTube ASMR Popularity Analysis

August 14th, 2023

Emma Crawford
emcr8954@colorado.edu

Abstract

My project will attempt answer three key questions. Firstly, what factors have the most significant impact on the monetization potential of ASMR YouTube videos? Secondly, how do engagement metrics, such as likes, dislikes, comments, and subscriber count, affect view and subscribers? Finally, are there specific content genres or categories that consistently perform well in terms of views, audience, and revenue generation? By addressing these questions, I can make data-driven decisions in content creation and marketing.

Key findings:

1. Likes, comments, and subscribers are the best predictors of views
2. Lengthy descriptions, titles, tag counts, and video durations are good predictors of views
3. Time of day is an adequate predictor of views
4. Old videos with low view counts may negatively affect new videos

1 Introduction

With this project, I hope to discover the key factors that determine the view count and overall popularity of YouTube videos in the ASMR niche. ASMR (Autonomous Sensory Meridian Response) is a form of content that uses high quality sound and pleasant objects to provide relaxing or sleep-inducing content to viewers. By analyzing a dataset of ASMR YouTube videos, I can learn about the patterns that attract users and optimize content creation strategies.

2 Related Work

In recent years, several studies have delved into the world of YouTube. One study by Gupta et al.[1] focused on using machine learning to develop models for YouTube's ranking mechanism and predicting views. They considered features like likes, dislikes, comments, and subscriber count to understand what drives video popularity. Another article by Yang et al.[2] investigated user engagement specifically with online science videos, uncovering patterns in video duration, social endorsement cues like likes, and their influence on viewer engagement. Additionally, Bärtl[3] conducted a statistical analysis examining YouTube channels, uploads, and views over the past decade. Their findings revealed the dominance of a small percentage of channels in obtaining the majority of views and emphasized the importance of genre selection for a lucrative channel.

While the existing literature gives great insights, my project aims to extend previous research by using a large dataset of niche YouTube videos. By considering additional features and applying advanced machine learning algorithms, I will compile recommendations for content creators such as myself to enhance their YouTube skills in a specific area of interest.

4 Dataset

The dataset used for this project is a custom scraped set of YouTube video data. The data was collected in accordance with YouTube's ethical and legal policies. I obtained an API Key and used a Jupyter notebook to extract data related to videos that I am interested in making: ASMR relaxation videos with anonymous creators (no personally identifying information). Code for using the Youtube API is included in the GitHub repository for review.

The dataset has 22551 rows, each representing a unique video. The features included are:

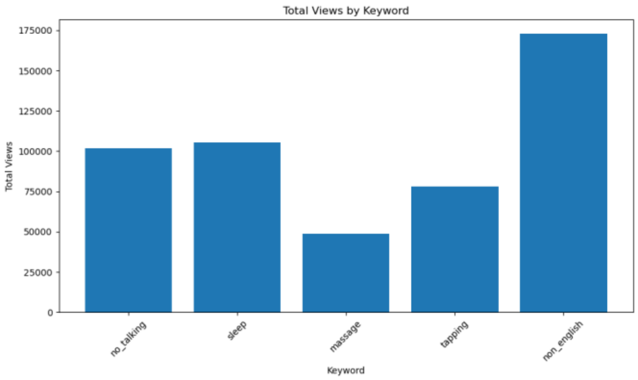
- Video information: title, description, tags, captions, definition (hd), views, like, comments, category, and published date
- Channel information: description, tags, published date, and subscriber count

5 Main Techniques Applied

Cleaning & Preprocessing: nulls, duplicates, and incorrectly formatted text were fixed prior to analysis. One tactic some youtubers use is to edit a field and re-publish their video several times. To mitigate this, I collected all videos for a channel with the same title and aggregated the data so that a duplicate video wasn't present 50-100 times in the data set when it had the same content.

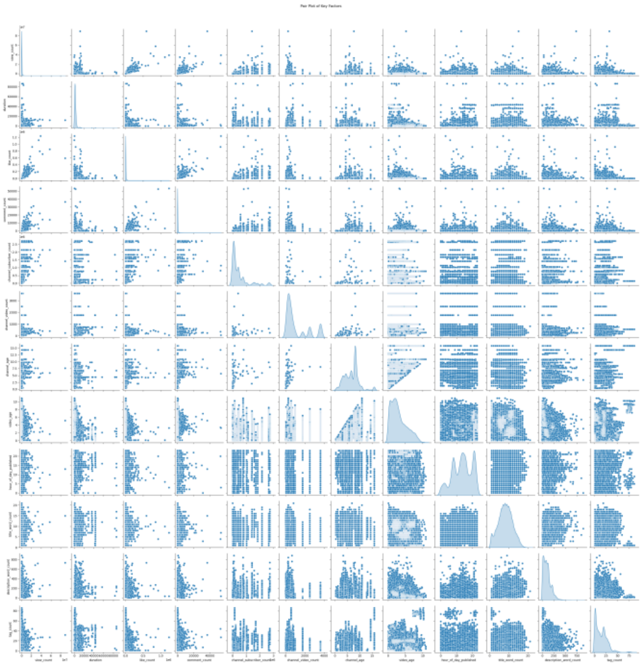
Feature Selection: channel keywords, irrelevant categories (e.g., pets & animals), and unnecessary columns (e.g., video_id) were removed during cleaning.

Feature Engineering: text fields (e.g. title, description, etc.) were summarized and tokenized to extract common keywords that were to be One-Hot Encoded for analysis. The main keywords extracted and encoded were: no_talking, sleep, massage, tapping, non_english. The non_english category indicates that at least one non_english character was found in one of the text fields. Day and hour fields were created from the published dates. Duration fields were converted to total seconds and date fields were all converted to central time to make them applicable to my current time zone.

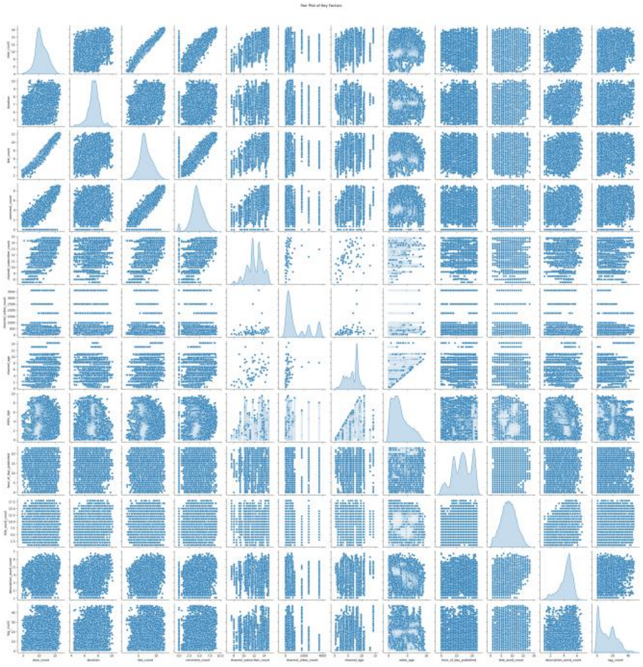


Log Transform: views, likes, comments, duration, subscribers, and description word count were all transformed due to extreme right skew. After this transformation, most variables had moderate to no skewness. Tag count, hour of day, subscriber count, and video count were multimodal which indicated potential benefit of cluster analysis.

Before log transform:



After log transform:



Outlier Handling: Outliers were removed using (z_score >= 3) after the log transform to preserve important information in the tails of the skewed variables. In the practical context, high view counts and high subscriber counts hold valuable information because they are more or less the ideal data point in our analysis. Removing them as an outlier before doing the transform may have resulted in lost information.

Collinearity Analysis: There appeared to be a few variables with high correlation which led to a Variance Inflation Factor analysis to determine if collinearity needed to be addressed. The only variable with a VIF > 5 was likes (VIF = 5.03). Because likes had such a high correlation with views (0.97) and because in a real-world context, likes are just as allusive as views, I decided to remove it from the analysis and focus on more nuanced variables. For this same reason, I chose to exclude comments (correlation with likes = 0.83) as well.

Metric Evaluation: I used mean squared error, mean absolute error, and R-squared to measure the performance of my predictive models. These metrics quantify how well my models can predict the outcomes I have chosen in the analysis process.

Linear Regression: I used linear regression as my baseline model to establish how well views are predicted assuming a linear relationship.

Random Forest Model: I used this supervised tree-based model to account for non-linear and complex relationships within the data as well as to highlight important features in the prediction of view count. Random Search (*hyperparameter tuning and cross-validation*) and Gradient Boosting (*ensemble technique*) were used to attempt to improve this model.

K-Means Clustering: This method was used as an unsupervised method to discover hidden clusters with the data that could indicate the need for segmentation or group consideration in the Random Forest Model. The elbow method was used to select the ideal number of clusters.

6 Tools & Libraries Used

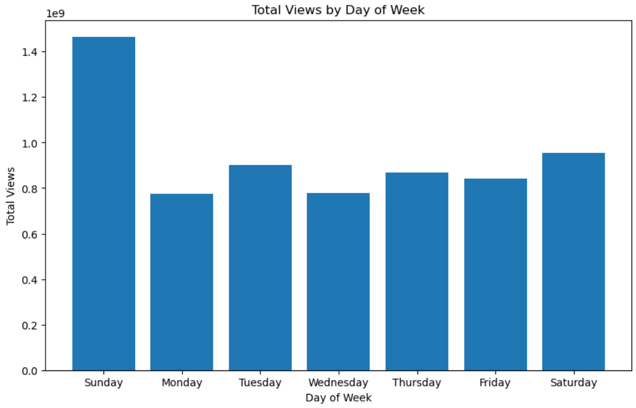
- Pandas
- Matplotlib
- Seaborn
- Numpy
- Sklearn
- Scipy
- Statsmodels
- NLTK
- Youtube API V3

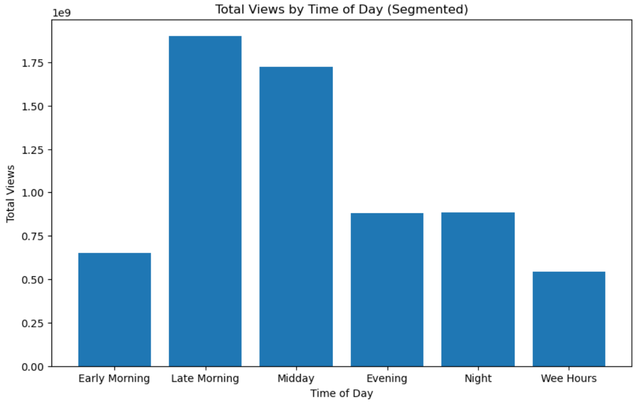
7 Results

Here are the summary statistics for view count:

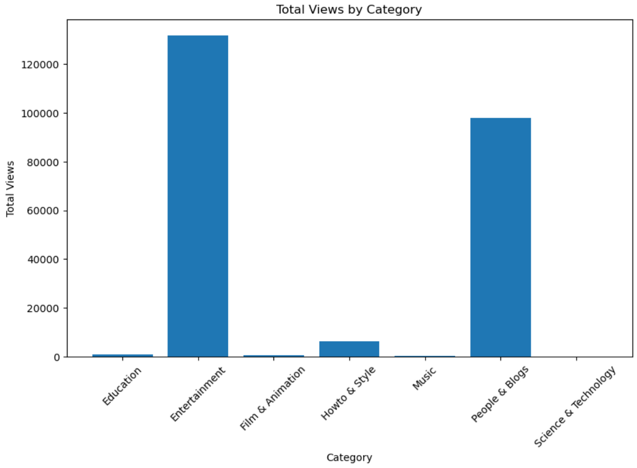
mean	2.920543e+05
std	1.363785e+06
min	2.100000e+01
25%	1.153450e+04
50%	3.284900e+04
75%	1.448890e+05
max	8.934656e+07

The average view count for this dataset is approximately 300,000 views with a large standard deviation of 1.4 million views. The highest total view counts come from videos published on a Sunday and in the late morning (8-12pm CDT) or Midday (12-4pm CDT).

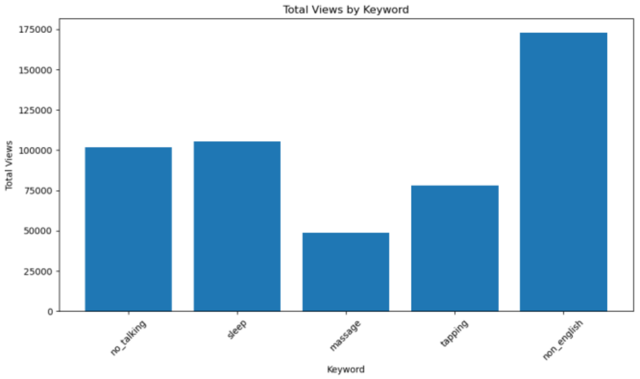




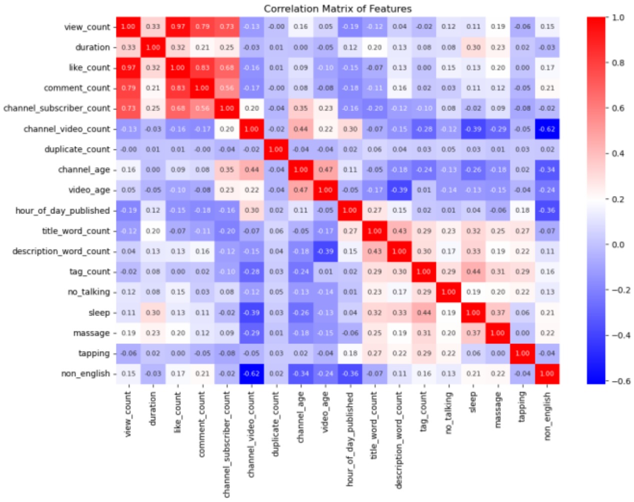
The most viewed category by far is ‘Entertainment’.



The top keywords found in either the title, description, or tags fields were no_talking, sleep, massage, tapping, and non_engligh (at least 1 non English character). It should be noted that non_english characters were not translated, so additional analysis may be needed to create a cross-language model.



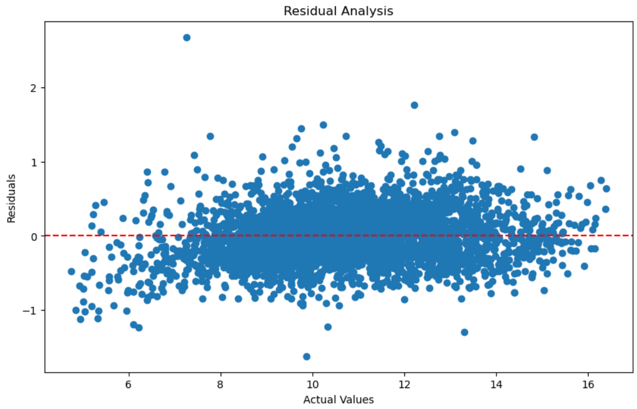
Collinearity was quite high between likes, comments, and subscribers which is expected in the real-world context and based on research of the YouTube recommendation algorithm. Other moderately positive correlations of note were: channel age and channel video count (0.44), duration and sleep tag (0.30), channel age and subscriber count (0.35). Moderately negatively correlated variables were: non_english characters and channel video count (-0.62), sleep and channel video count (-0.39), and channel age and non_english characters (-0.34).



7.1 Linear Regression

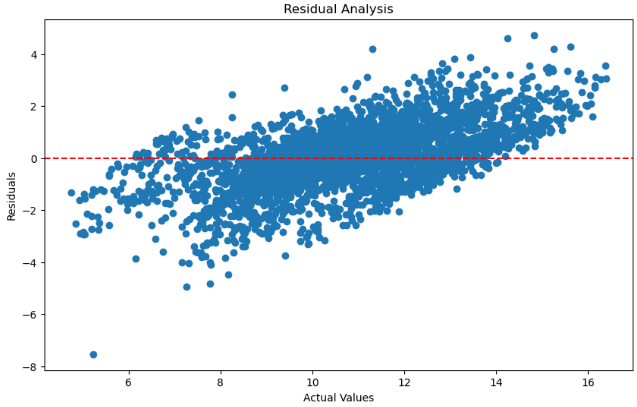
When including likes and comment counts, the linear model resulted in:

Mean Squared Error: 0.117
R-squared: 0.968



After removing likes and comment counts, the linear model outputs were:

Mean Squared Error: 1.344
R-squared: 0.636



The first model has an excellent R-squared and a random residuals plot, indicating good predication and a relatively linear relationship. However, this result is deceptive because likes and comments are highly correlated with view count. It seems very likely that these are the primary factors in the recommendation algorithm which are reflected in research about YouTube recommendations as well as intuition about how YouTube would ideally measure engagement and enjoyment.

However, in the data's context, likes and comments are just as difficult to obtain as views so there is not as nuanced a view into the algorithm as I would like. Because of this, I removed likes and comments to see what MSE and residuals come up without the highly linear relationships. This result is more what I would expect to see in a model attempting to predict views based on something other than likes and comments. The low MSE and patterned residual plot indicates that I needed a more complicated and nuanced model to capture the relationships in the data.

7.2 Random Forest Model (Tree-based Model)

The tree-based models do a better job at handling complex and non-linear relationships between features and it has a good track record with one-hot encoded features. I performed the same comparison as linear regression with including and excluding likes and comment counts. Here is the comparison:

Likes and comments included:

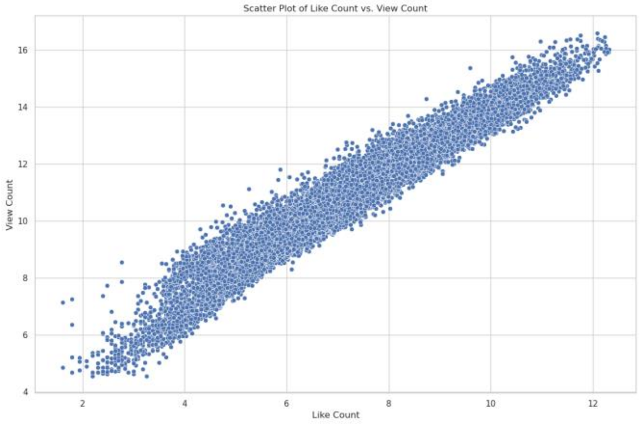
Mean Absolute Error: 0.172
R-squared: 0.984

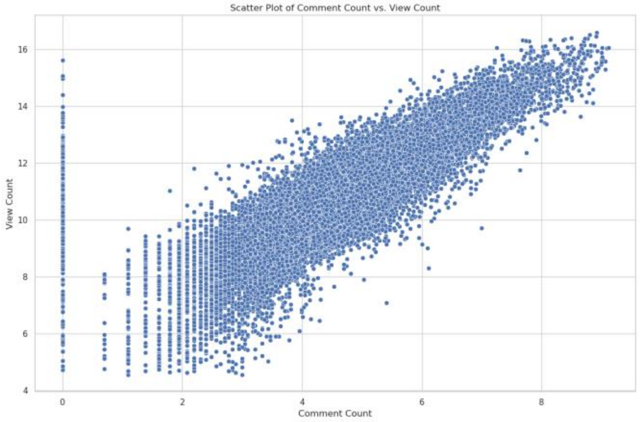
Likes and comments excluded:

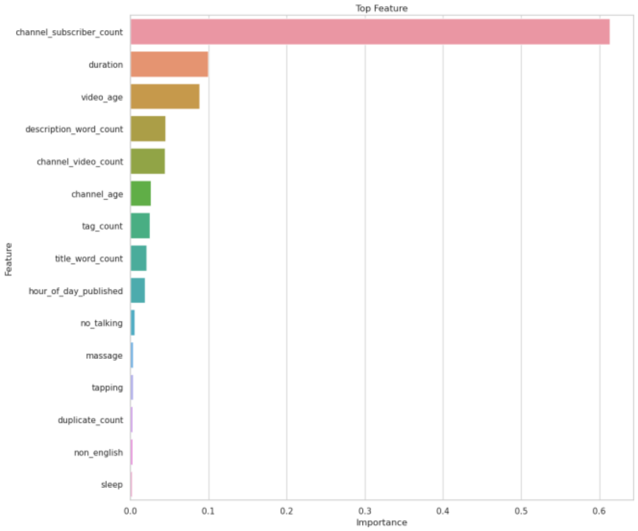
Mean Absolute Error: 0.627
R-squared: 0.807

As expected, this model outperformed the baseline linear regression in both contexts: with/without likes and comments. It performed adequately without the likes and comments with a good R-squared.

I evaluated feature importance for this model to understand what specific features were contributing to prediction. For visual demonstration, I will also include scatterplots of like and comment counts to demonstrate their importance in the models that included them.







As shown in the feature importance graph, the most important features next to likes and comments are:

1. Subscriber count
2. Duration
3. Video age
4. Description length
5. Video count
6. Tag count
7. Title length
8. Time of Day Published

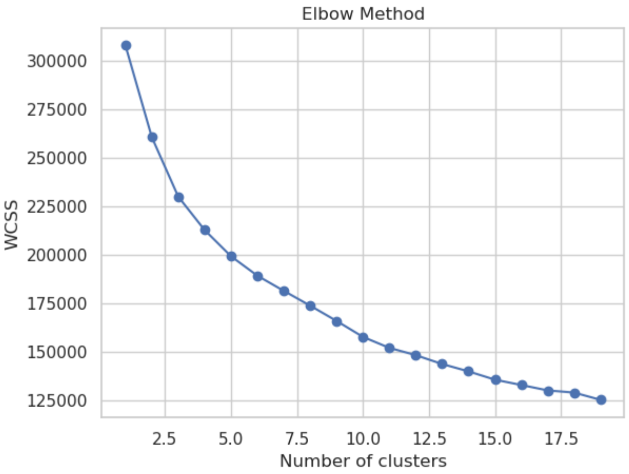
Hyperparameter tuning resulted in a minimal increase in R-squared (0.809). The Gradient Booster had no effect (0.807). However, I was only able to use 10 iterations because of computational limitations. These methods may be more effective with more available resources.

In summary, the Random Forest Model does a good job at accounting for variation within the data and predicting views without taking likes and comments into account. The main features that are important in that predication are subscribers, duration, video age, and description length.

7.3 K-Means Clustering Analysis

The keywords chosen in the feature selection and engineering were selected based on this k-means analysis because they were emerging as key clustering elements. This reflects what I know of the ASMR niche as well.

The elbow method was used for the ideal number of clusters. There was not a clear elbow as shown in the graph. However, I tested three cluster sizes within the elbow range of the graph within the context of domain knowledge and practicality to determine the best number of clusters for this analysis. Based on the graph, the curve accelerates at around 4 clusters and decelerates at around 10 clusters with a midway point at 6 clusters. Each were evaluated for practicality and cluster features.



Four clusters did not seem to capture the complexities of the data that I observed more intuitively and what I would assume based on topic niche. Ten clusters seemed to over-categorize the data and render it difficult to understand; there also seemed to be clusters with very similar data which were not different enough to be practically meaningful. I landed on six clusters because it provided the most granularity without sacrificing readability and practicality.

The clusters were divided relatively evenly by number of videos per cluster:

Cluster 4: 4055 videos
Cluster 5: 3960 videos
Cluster 1: 3541 videos
Cluster 2: 3299 videos
Cluster 3: 2932 videos
Cluster 0: 2723 videos

Here are the average values for each column based on the clusters created by the k-means analysis. The view_count, duration, channel_subscriber_count, and description_word_count were kept in log transform format for simplicity. All cluster values were statistically significant based on the Kruskal-Wallis Test with $p < 0.05$.

Cluster 0:
Average view_count: 9.616
Average duration: 7.078
Average channel_subscriber_count: 11.567
Average channel_video_count: 2086.855
Average channel_age: 8.249
Average video_age: 3.494
Average hour_of_day_published: 17.493
Average title_word_count: 10.474
Average description_word_count: 4.298
Average tag_count: 16.252
Average tapping: 0.967

Cluster 1:
Average view_count: 11.870
Average duration: 6.988
Average channel_subscriber_count: 13.488
Average channel_video_count: 974.672
Average channel_age: 7.672
Average video_age: 4.448
Average hour_of_day_published: 6.929
Average title_word_count: 4.799
Average description_word_count: 3.435
Average tag_count: 10.122
Average non_english: 0.957

Cluster 2:
Average view_count: 11.456
Average duration: 7.797
Average channel_subscriber_count: 12.426
Average channel_video_count: 367.333
Average channel_age: 5.290
Average video_age: 2.073
Average hour_of_day_published: 13.402
Average title_word_count: 10.771
Average description_word_count: 4.881
Average tag_count: 22.221
Average sleep: 0.893
Average massage: 1.000
Average non_english: 0.914

Cluster 3:

Average view_count: 10.351
Average duration: 7.565
Average channel_subscriber_count: 12.904
Average channel_video_count: 3507.631
Average channel_age: 8.154
Average video_age: 3.867
Average hour_of_day_published: 17.901
Average title_word_count: 7.427
Average description_word_count: 3.997
Average tag_count: 4.179

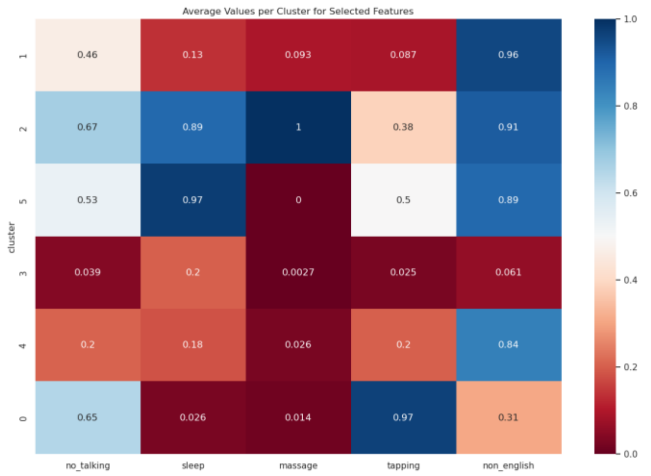
Cluster 4:

Average view_count: 9.739
Average duration: 6.982
Average channel_subscriber_count: 10.893
Average channel_video_count: 366.281
Average channel_age: 4.931
Average video_age: 2.050
Average hour_of_day_published: 13.305
Average title_word_count: 6.923
Average description_word_count: 4.134
Average tag_count: 8.714
Average non_english: 0.843

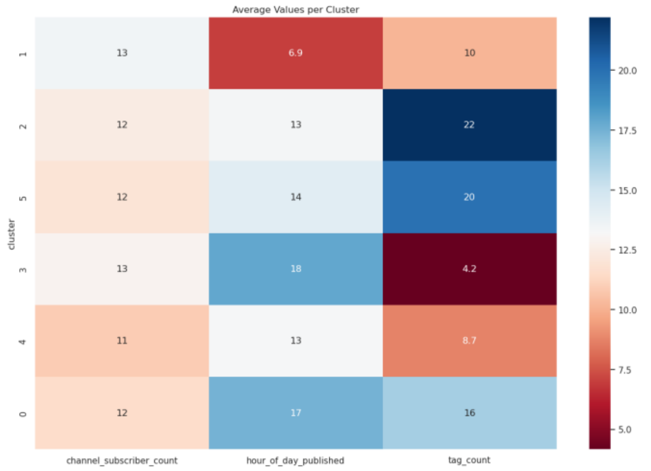
Cluster 5:

Average view_count: 10.737
Average duration: 7.623
Average channel_subscriber_count: 11.960
Average channel_video_count: 437.771
Average channel_age: 5.733
Average video_age: 2.756
Average hour_of_day_published: 14.229
Average title_word_count: 10.107
Average description_word_count: 4.933
Average tag_count: 19.935
Average sleep: 0.973
Average non_english: 0.891

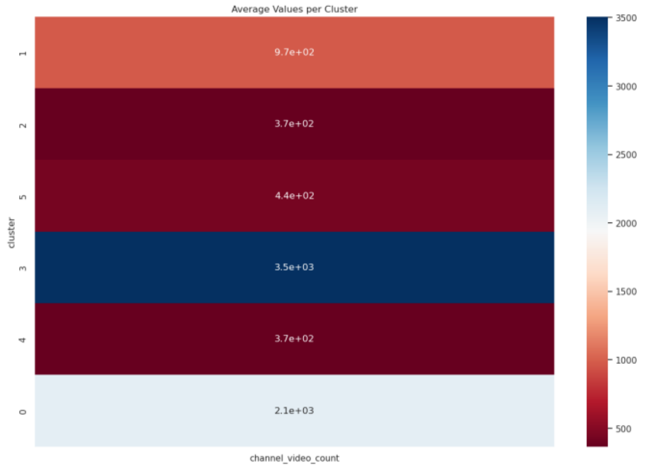
Differences in one-hot encoded keyword columns are illustrated here and ordered by average view count.



Subscribers, tag count, and time of day published are summarized here, also ordered by average view.



Lastly, video count is summarized here.



7.4 Re-evaluate Random Forest Model with Clusters

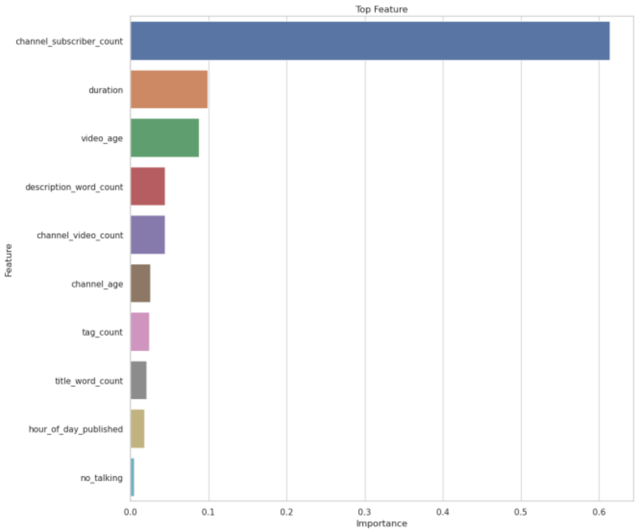
Based on these clusters, which reflect domain knowledge of ASMR, I went back to evaluate the Random Forest Model based on the new cluster column to determine if any improvement could be made to prediction based on cluster number or characteristics. The new metrics including the cluster number resulted in a slight increase to R-squared from the original model without the clusters:

Mean Absolute Error: 0.625
R-squared: 0.808

Hyperparameter tuning resulted in an additional slight increase to R-squared:

Best Parameters:
{30 fits, n_estimators: 300, min_samples_split: 5, 'min_samples_leaf': 4, 'max_depth': 40, 'bootstrap': True}
Mean Absolute Error: 0.622
R-squared: 0.809

The feature importance list remained approximately the same, with channel age becoming more important when the clusters were included in the analysis.



8 Applications

Here are the characteristics of the clusters in the ASMR niche. Applications are to apply techniques in the top clusters and avoiding the pitfalls of the lower clusters:

Cluster 0 (2723 videos)

- **Viewership**: Approximately 15,000 views.
- **Duration**: Approximately 117.38 minutes.
- **Content and Engagement**: Older channels, videos published around 5:49 PM, longer titles, and more tags. Includes tapping content.

Cluster 1 (3541 videos)

- **Viewership**: Approximately 142,000 views.
- **Duration**: Approximately 107.84 minutes.
- **Content and Engagement**: More subscribers, fewer channel videos, videos published around 6:56 AM, shorter titles, fewer tags, mostly non-English.

Cluster 2 (3299 videos)

- **Viewership**: Approximately 94,000 views.
- **Duration**: Approximately 242.04 minutes.
- **Content and Engagement**: Younger channels, videos published around 1:40 PM, longer titles, more tags, sleep and massage content, often non-English.

Cluster 3 (2932 videos)

- **Viewership**: Approximately 31,000 views.
- **Duration**: Approximately 192.36 minutes.
- **Content and Engagement**: Videos published around 5:54 PM, average title word counts, description word counts, fewer tags.

Cluster 4 (4055 videos)

- **Viewership**: Approximately 17,000 views.
- **Duration**: Approximately 107.21 minutes.
- **Content and Engagement**: Youngest channels, videos published around 1:18 PM, shorter titles, fewer tags, often non-English.

Cluster 5 (3960 videos)

- **Viewership**: Approximately 46,000 views.
- **Duration**: Approximately 205.49 minutes.
- **Content and Engagement**: Videos published around 2:14 PM, longer titles, more tags, sleep content, often non-English.

9 Key Recommendations

For a YouTube content creator in the niche of ASMR without personal identifying information and in the category of relaxation (not food, animals, or role-play), this analysis indicates the following recommendations:

1. Likes and comments are the best predictors of view count
2. Subscriber count is the next best predictor of view count
3. Post your video on a Sunday between 6am and 3pm CDT; If you are busy on Sundays, try to still post your video in this time frame during the week
4. Focus on sleep and massage content
5. Use approximately 20 tags per video
6. Create a title with approximately 10 words
7. Write long descriptions to describe your channel, content, and social media links
8. Aim for longer videos, if possible
9. Remove old videos with low view counts from your channel

REFERENCES

[1] Vandit Gupta, Akshit Diwan, Chaitanya Chadha, Ashish Khanna, Deepak Gupta. Machine Learning enabled models for YouTube Ranking Mechanism and Views Prediction. arXiv:2211.11528 [cs.IR]. Available at: https://arxiv.org/abs/2211.11528.

[2] Shiyu Yang, Dominique Brossard, Dietram A. Scheufele, Michael A. Xenos. The science of YouTube: What factors influence user engagement with online science videos? PLOS ONE, 17(5), e0267697. DOI: https://doi.org/10.1371/journal.pone.0267697.

[3] Mathias Bärtl. YouTube channels, uploads and views: A statistical analysis of the past 10 years. Convergence: The International Journal of Research into New Media Technologies, 24(1), 1354856517736979. DOI: https://doi.org/10.1177/1354856517736979.