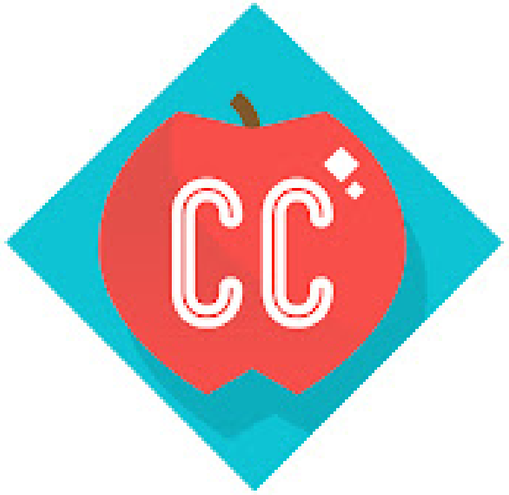
CRASHCOURSE



CHANNEL ANALYSIS

Supervised by:
Dr. Khulood alyaahya
Dr.Reem Alqifari

Table of Content

Our problem

CrashCourse is dealing with this thing where views on their videos keep going up and down. YouTube's always changing its algorithms and recommendations, which makes it tough to predict. We really need to figure out why some videos do better than others and spot any trends. That way, we can come up with solid plans to boost views and reach more people, making CrashCourse even bigger on YouTube

Data Collection

The data for our project is sourced from the **YouTube API,** data related to the CrashCourse channel with **1514** observations

One challenge we've encountered is that the YouTube API only fetches up to 50 rows of data for videos in the channel per request. To overcome this limitation, we've implemented multiple requests and integrated the results to gather comprehensive data.

About The Data

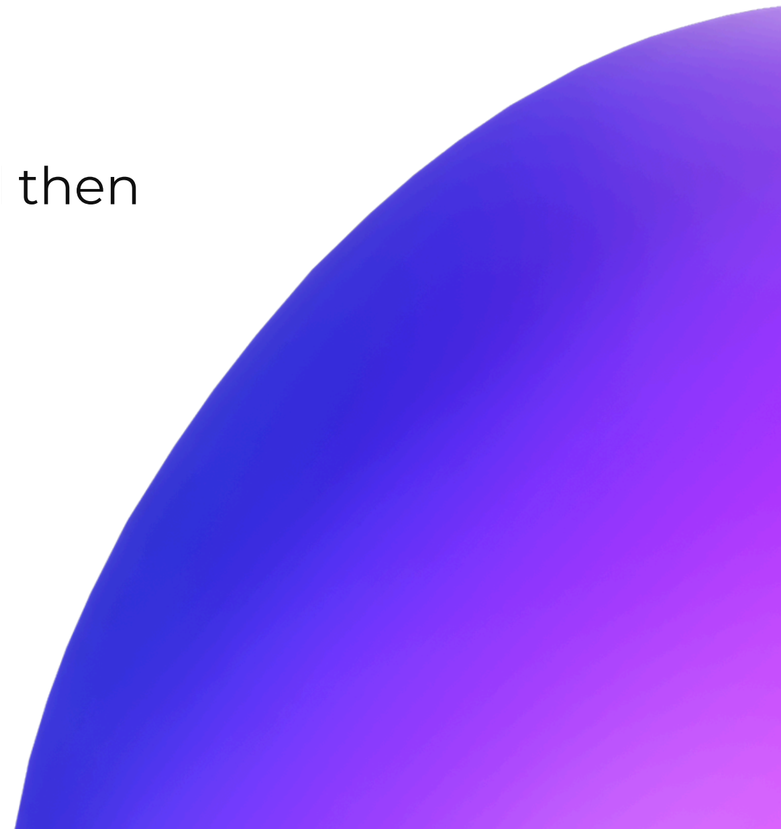
We obtained the following colomns :

1. **Title:** The title of the video.
2. **ID**: A unique identifier assigned to each video for identification purposes.
3. **Published_date**: The date and time when the video was published on the channel.
4. **Tags**: Any tags associated with the video, providing insight into its content.
5. **Views**: The number of views the video has garnered.
6. **Likes**: The count of likes received by the video.
7. **Comments**: The number of comments posted on the video.
8. **Duration**: The duration of the video.
9. **Captions**: A binary indicator representing whether captions are available for the video.

Data Processing and Cleaning

In our data preprocessing phase, we execute the following operations:

- Eliminate duplicate values and null entries.
- Convert boolean captions to binary values, assigning 1 for True and 0 for False.
- Normalize views, comments, and likes.
- Format the 'Published_date' column into datetime objects and then convert it into a specific string format.
- Transform duration into seconds.

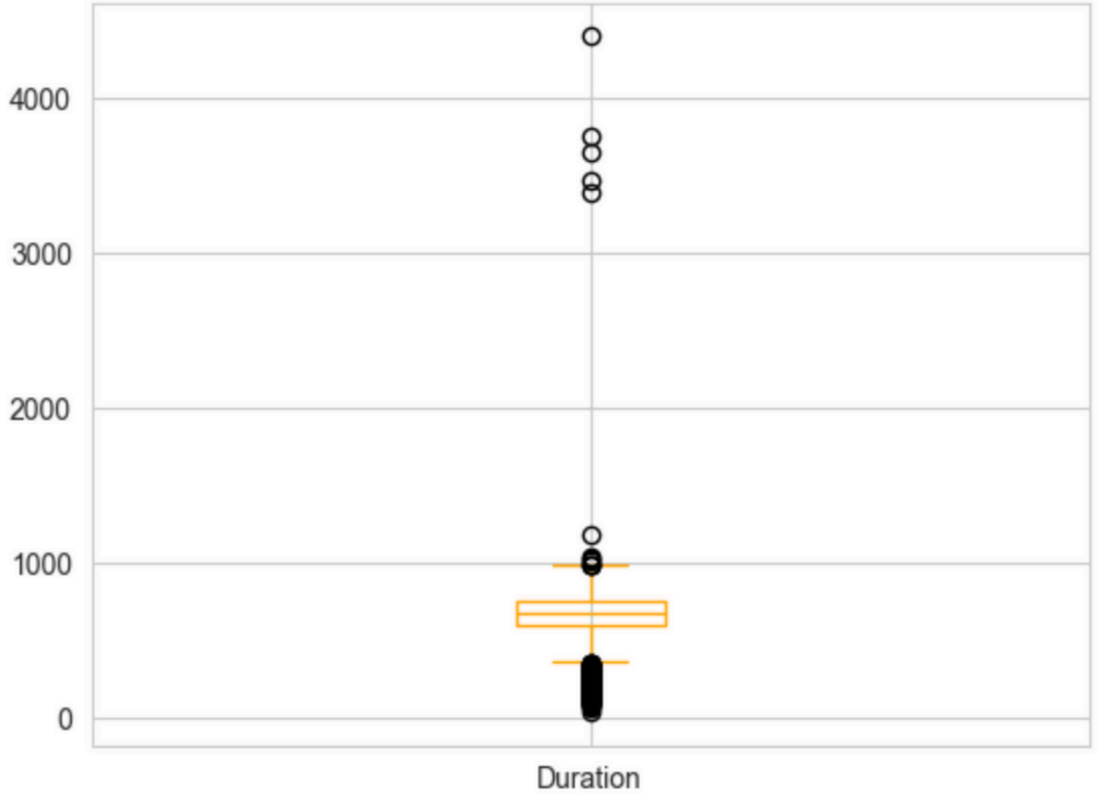


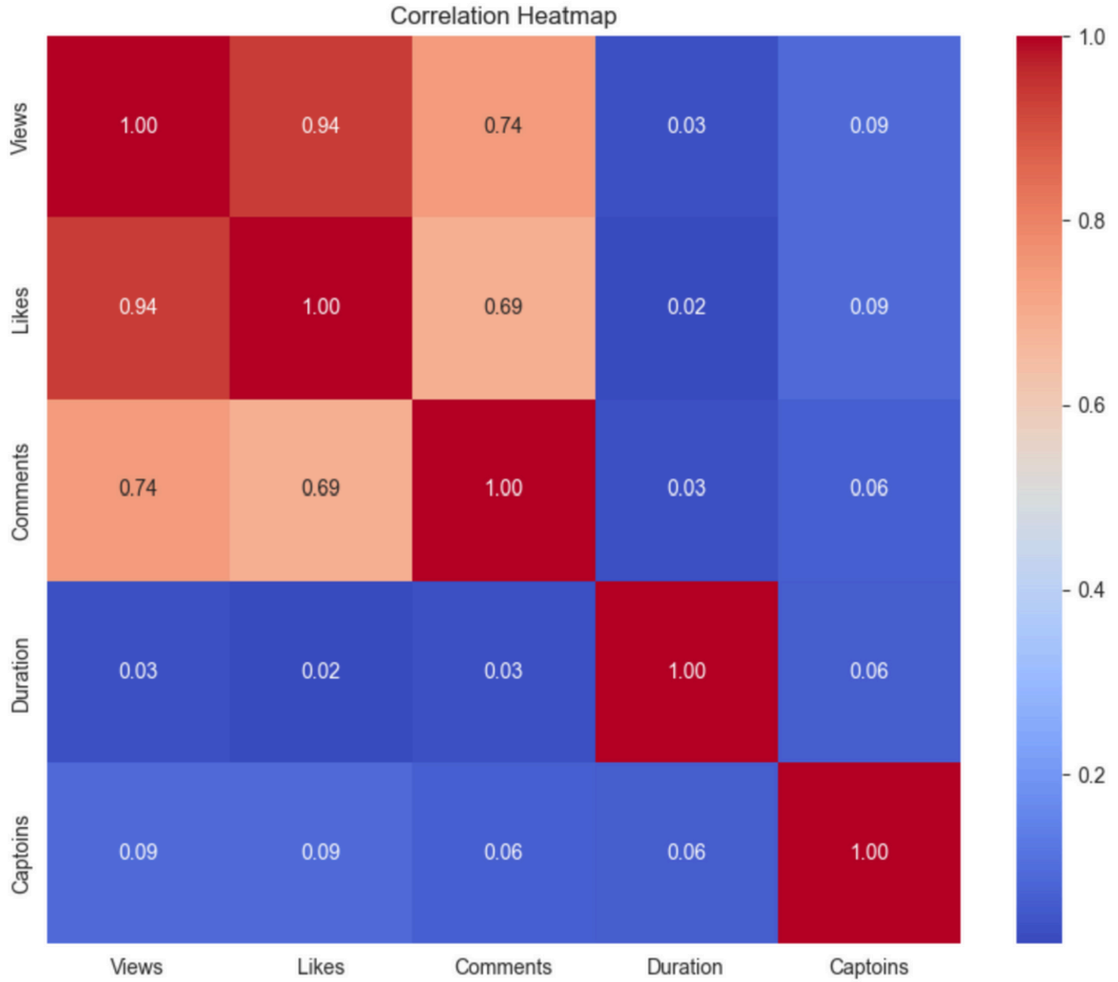
OUR DATA AFTER PREPROCESSING

	Title	ID	Published_date	Tags	Views	Likes	Comments	Duration	Captoins
0	Why Your Cat Looks Like That: Genetics: Crash ...	YnJPbphsoMY	2024-02-20 17:00:21	['vlogbrothers', 'Crash Course', 'crashcourse'...	0.002072	0.005684	0.000810	708.0	1
1	Black American History Arts & Culture | Compil...	bfH3fkIsc5U	2024-02-16 16:30:06	['vlogbrothers', 'Crash Course', 'crashcourse'...	0.001043	0.002673	0.000911	4393.0	1
2	Why Are All Humans Unique? Meiosis: Crash Cour...	pj1oFx42d48	2024-02-13 17:00:39	['vlogbrothers', 'Crash Course', 'crashcourse'...	0.002887	0.005596	0.001215	770.0	1
3	Mitosis and the Cell Cycle: Crash Course Biolo...	skPOXcVvS5c	2024-02-06 17:00:44	['vlogbrothers', 'Crash Course', 'crashcourse'...	0.003004	0.005371	0.000473	671.0	1
4	Photosynthesis: The Original Solar Power: Cras...	-ZRsLhaukn8	2024-01-30 17:00:00	['vlogbrothers', 'Crash Course', 'crashcourse'...	0.003297	0.004757	0.000878	784.0	1

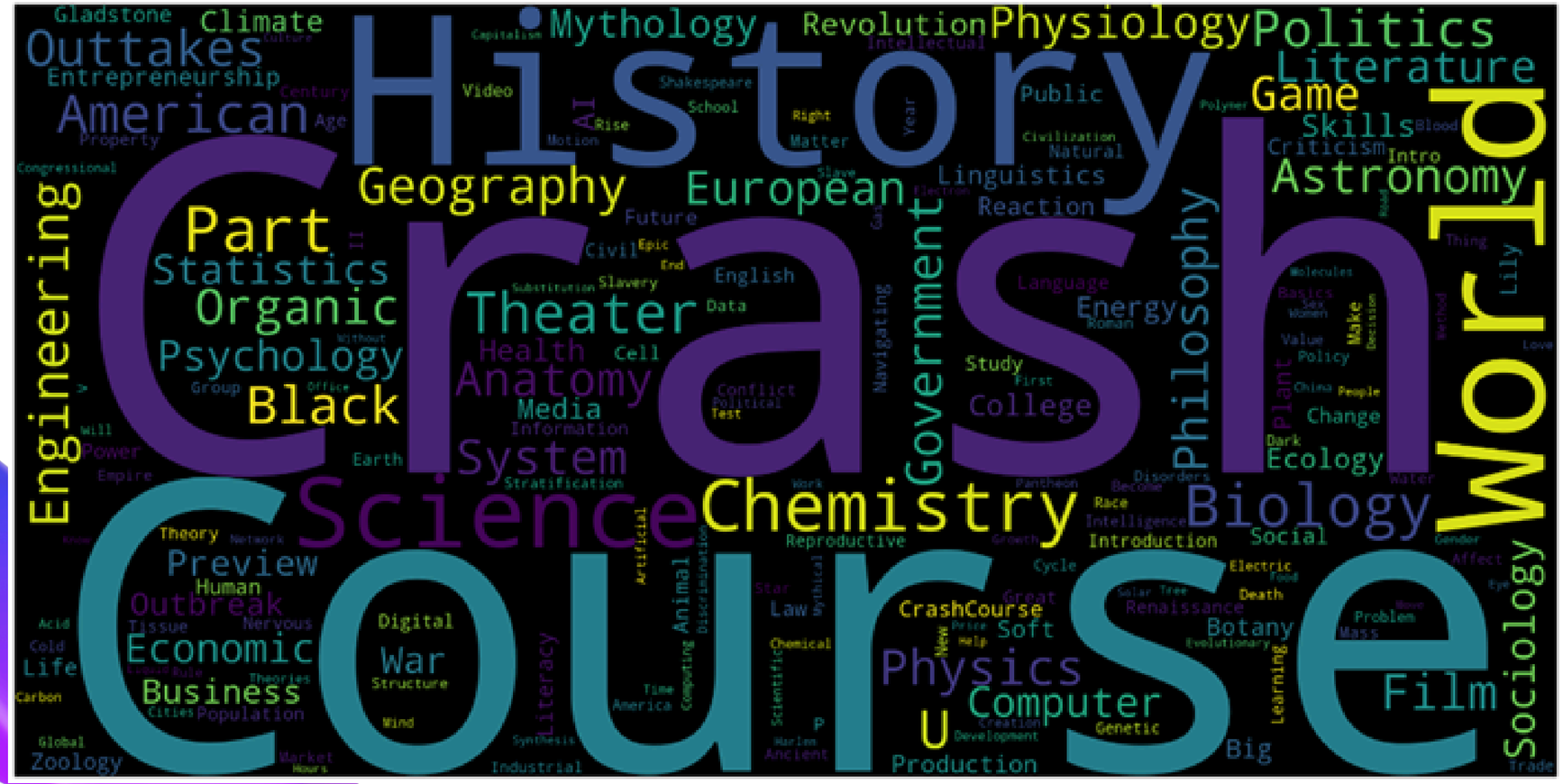
Exploratory Data Analysis

```
The average duration of the videos is 667.2052805280528 seconds
The average duration of the videos is 11.120088008800879 minutes
```





Exploratory Data Analysis
(word cloud visualization)

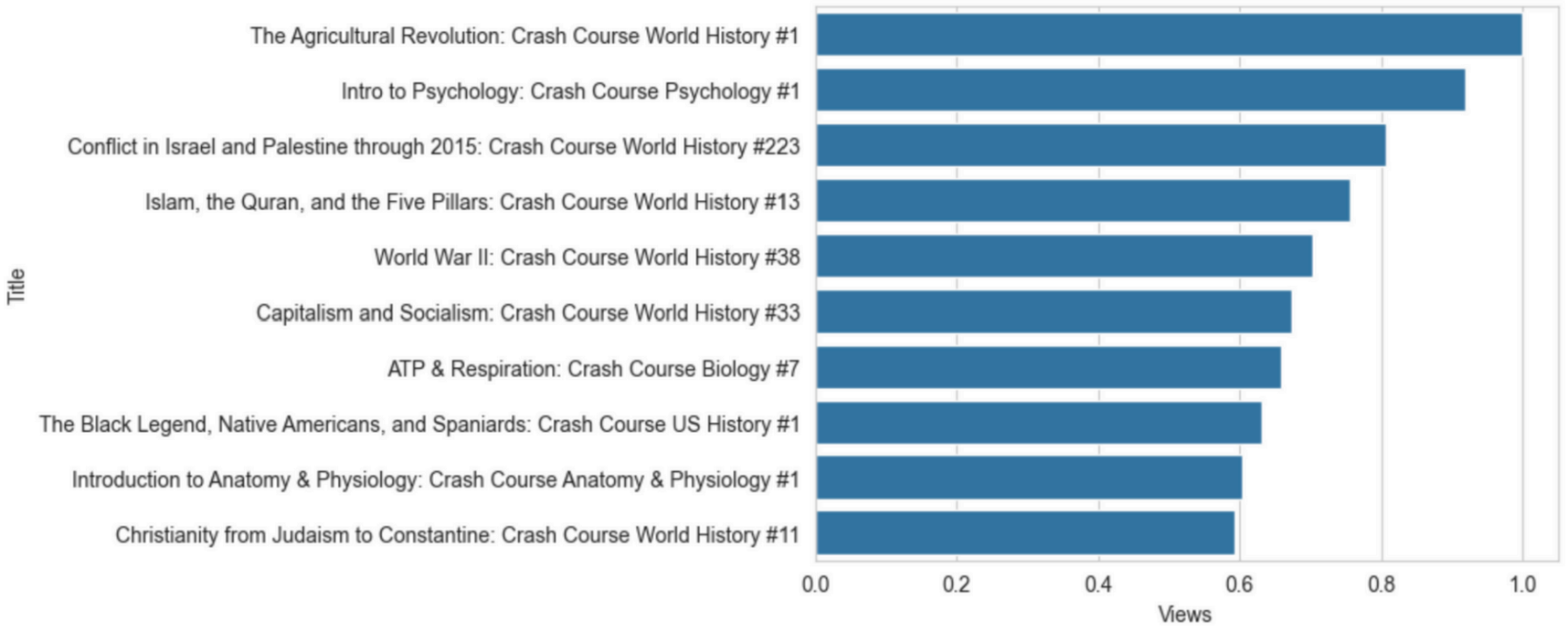


The most frequent word occurrences in the titles.

The word "Crash" appears prominently, due to its being the channel's name.

Other common words may represent recurring themes or subjects covered in the videos, providing insights into the **channel's content focus.**

Exploratory Data Analysis



top 10 most-viewed videos, revealing a strong interest in **historical topics**, then psychology and biology.

The diversity of subjects reflects CrashCourse's broad appeal, guiding future content decisions.

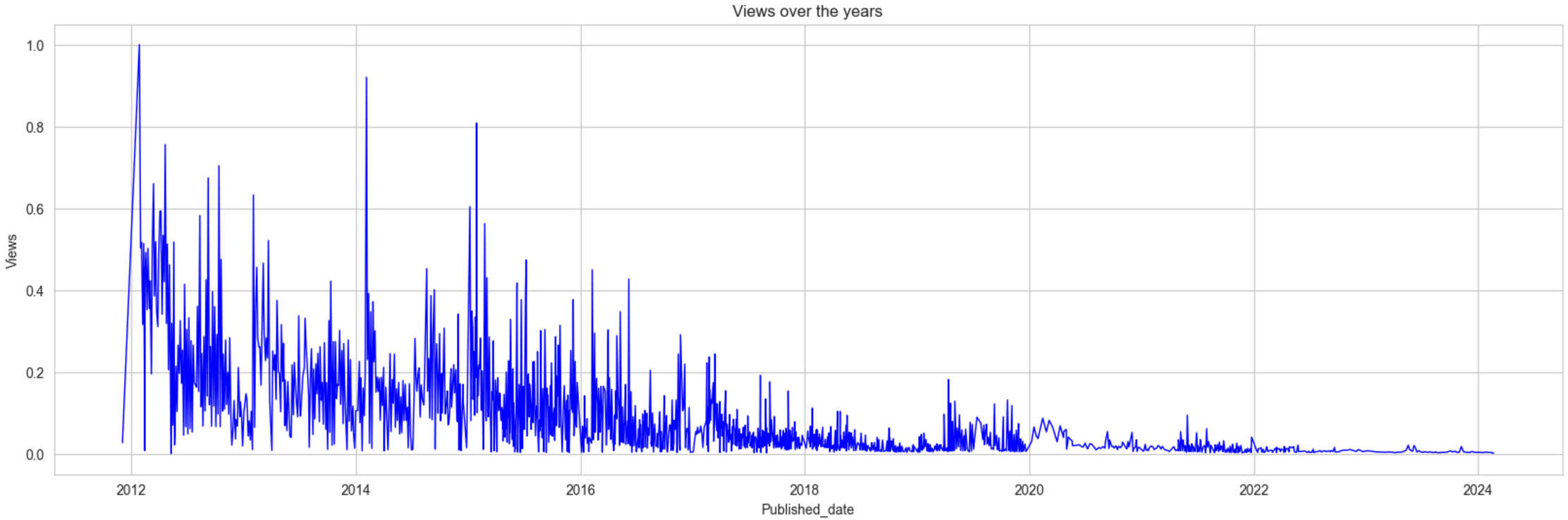
FEATURES ANALYSIS
(Tags Treemap)

The topics of these tags **"Nisa," "Neolithic Revolution," "pre-history," and "herders,"** are of significant **interest to the audience.**

These topics are relevant to **historical or cultural subjects.**

FEATURES ANALYSIS

(Published_date Line plot)



The **decrease in video views over time** could stem from several factors, including changes in platform algorithms, or shifts in viewer behavior and preferences.

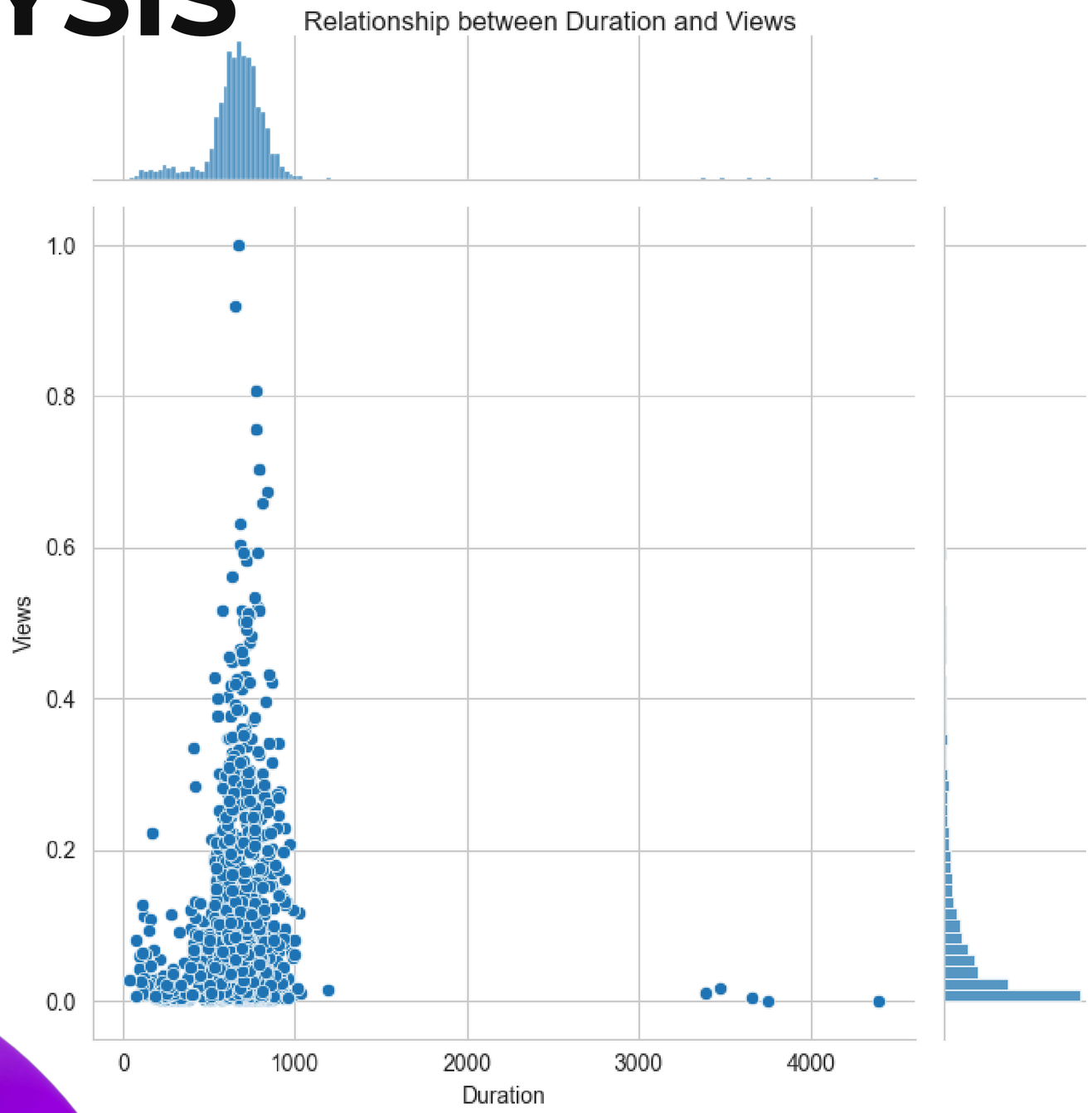
The **decline may not be unique to this YouTube channel** but could be a broader trend affecting many channels. The trend might be increasing due to the rise of alternative platforms like TikTok, drawing a significant portion of the audience away from YouTube.

FEATURES ANALYSIS

(Duration Joint plot)

1- **Short videos are more appealing** to viewers, as they require less time commitment.

2- The **ideal time** frame for a video to receive more views is **between eleven and fifteen minutes**.

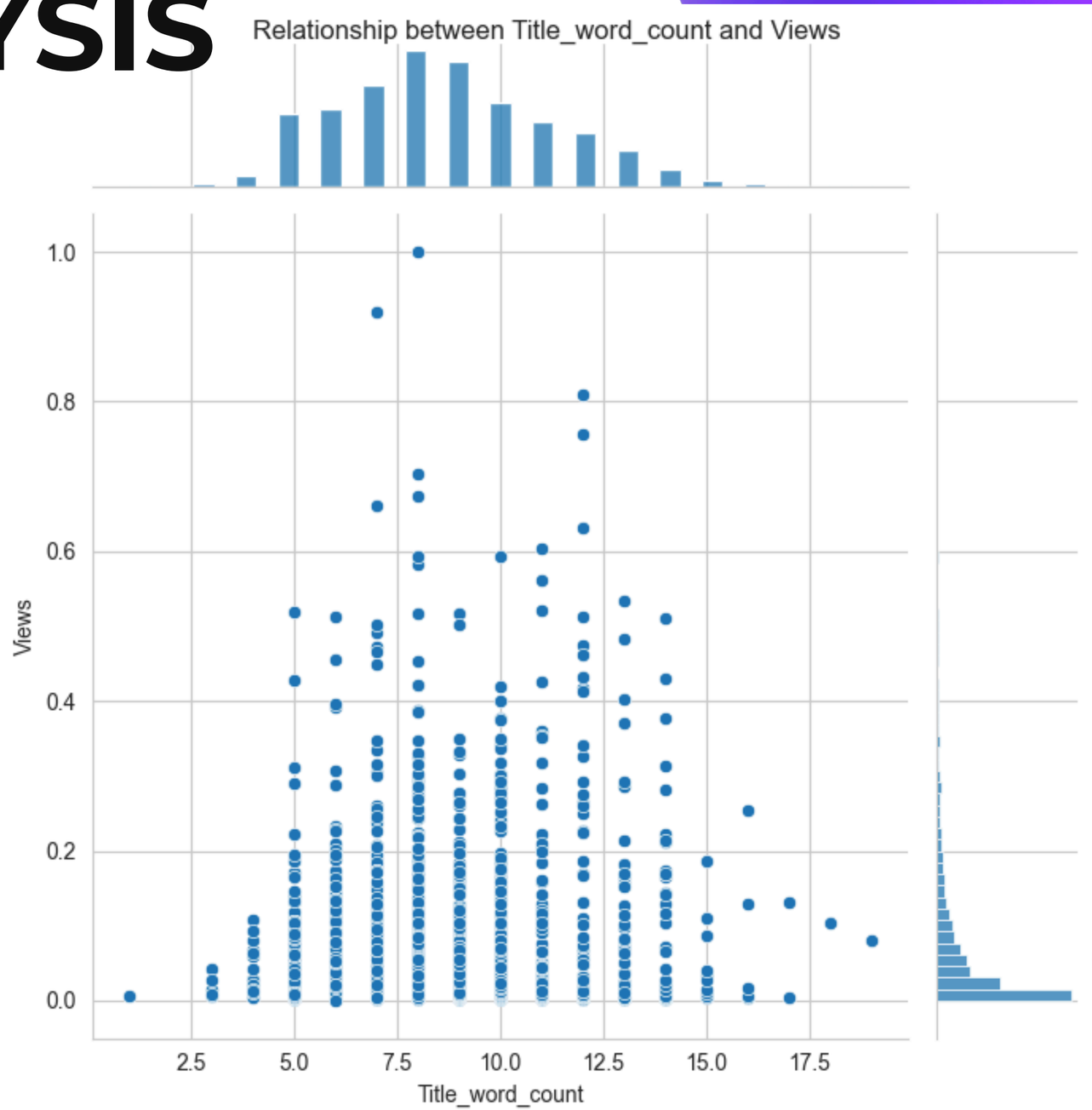


FEATURES ANALYSIS

(Title word count Joint plot)

1- Although **there isn't a clear correlation** between 'Title_word_count' and 'Views', the plot suggests that videos with **titles containing 7-11 words tend to receive more views**.

2- The **lack of a clear pattern** suggests that **factors other than title word count** might significantly **influence the number of views**.

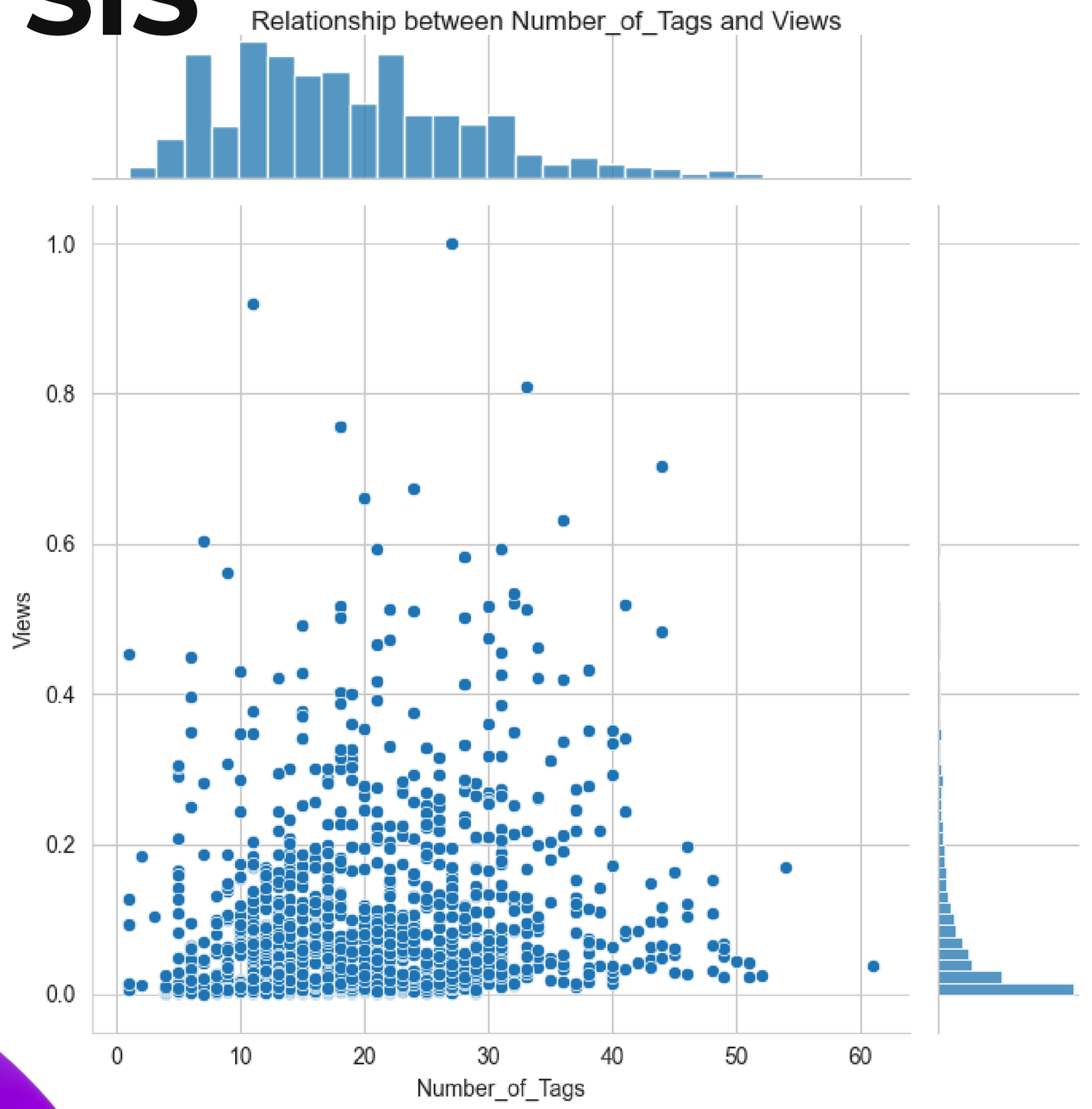


FEATURES ANALYSIS

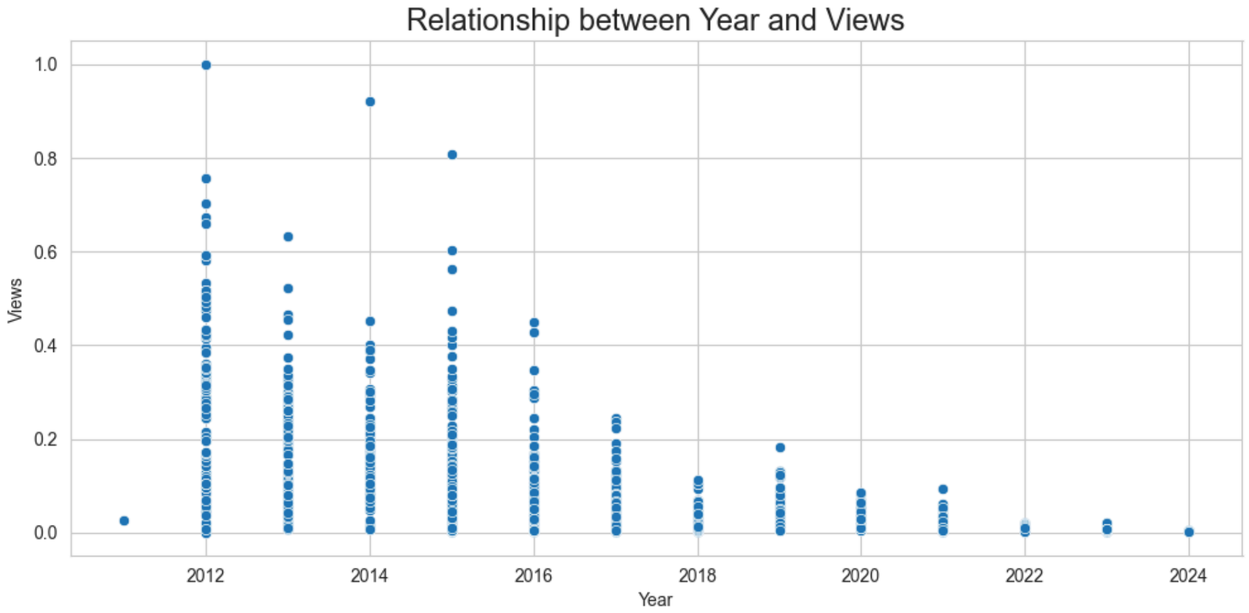
(Number of tags Joint plot)

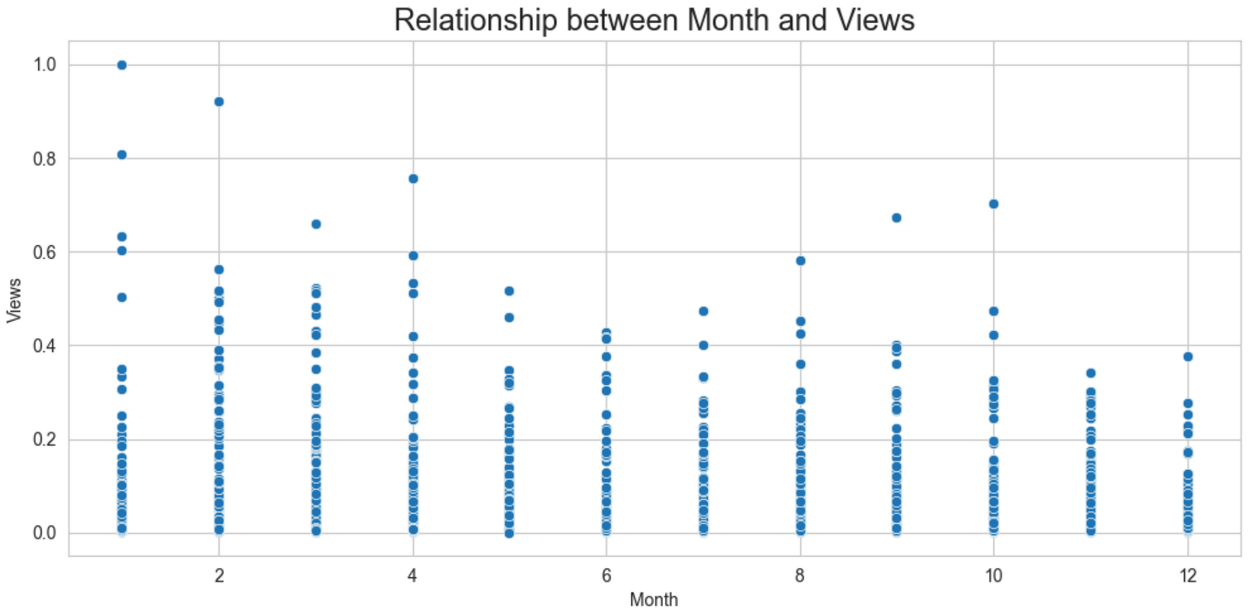
1- it's difficult to identify a specific number of tags that consistently lead to higher views.

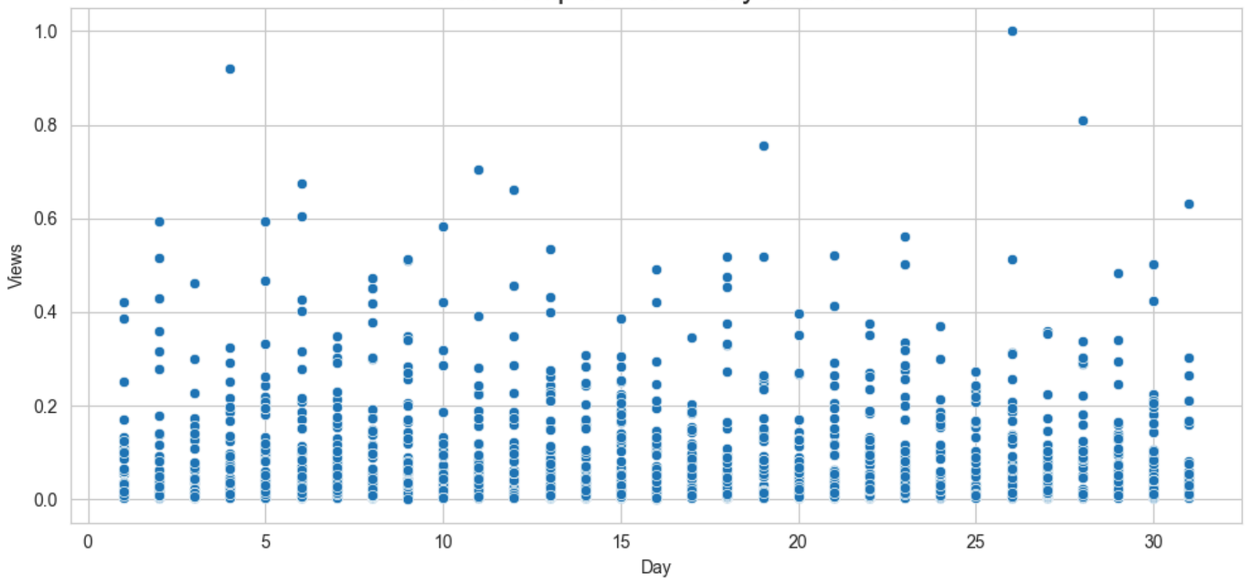
2- Factors other than the number of tags likely play a more significant role in attracting viewers.

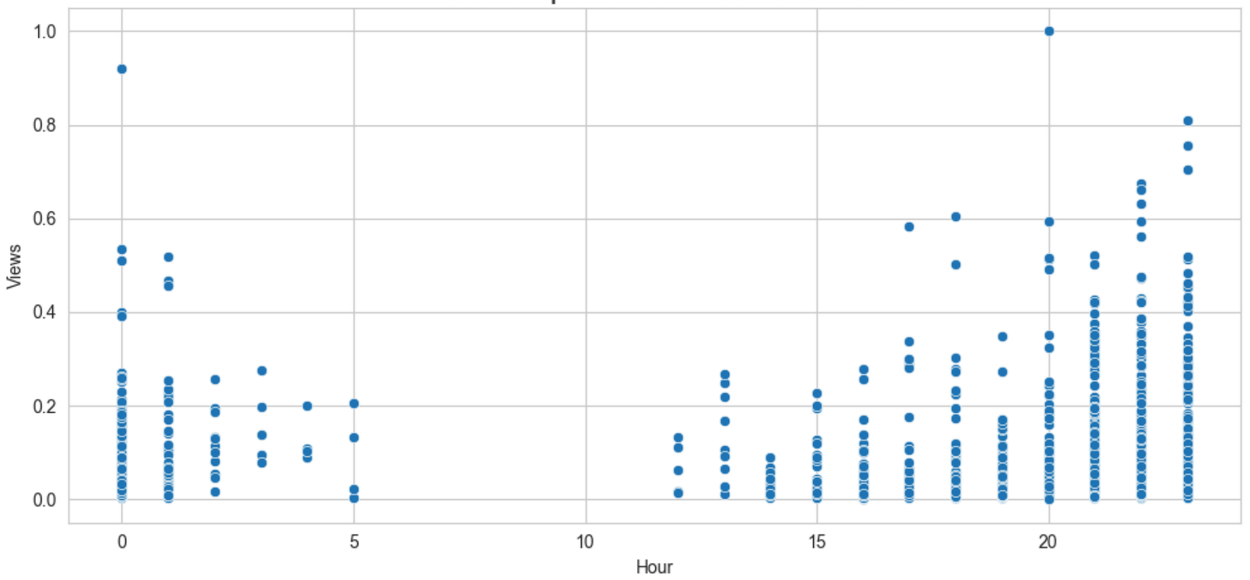


(Date / Time Scatter plots)









(Date / Time Scatter plots)

Month Scatter plot:

- The number of views doesn't vary much between different months. This could indicate that **seasonality has a minimal impact on views**. However, there's a **slight increase in views noticeable in February, March, and April.**
- Since there’s no clear monthly trend, the content creator should focus on consistent quality and engagement rather than timing releases for a particular month.

Day Scatter plot:

- The uniform distribution across days suggests that the **day of the month has little to no impact on views.**

Hour Scatter plot:

- Specific hours with peak views like **hours 21, 22, and 23 could reflect optimal times for posting** or audience online availability and patterns.
- After hour 15, there is a noticeable increase in video views.

MODELING

We built a comprehensive set of regression models to predict the number of views a video can have.

Regression is chosen in this case because we want to predict the views (a continuous outcome variable) a video can have.

Predicting views helps evaluate video performance, optimize content, allocate resources efficiently, estimate revenue potential, and gain insights into audience preferences.

These predictions will be based on factors that content creator can manipulate before uploading their videos to YouTube, including duration, title length, Year, Month, Number of Tags, Day, and Hour.

Algorithms

We've chosen a baseline model, **Linear Regression**, for its simplicity and effectiveness in capturing basic patterns

For alternative models, we opted for **Random Forest Regression and XGBoost Regression** to handle complex relationships.

Bayesian Linear Regression estimates uncertainty, while **Lasso ElasticNet** tackles correlated predictors and overfitting.

The selection of these algorithms covers a wide range of modeling approaches, from simple to more complex methods.

Methodology

01

Establishing a Baseline Model: Initially, we create a simple model (Linear Regression) to set a standard for comparison.

02

Developing and Evaluating **Four Alternative Models**: We created four extra models, these models include Random Forest Regression, XGBoost Regression, Bayesian Linear Regression, and Lasso ElasticNet.

03

We experimented with **two types of data splits** for all five models: cross-validation and an 80-20 split. Therefore, we will have a **total of ten models.**

04

Select the Best-Performing Model for Prediction: based on **having lower errors** and minimizing the average difference between predicted and actual values. choose the model with the **lower values of MAE or MSE.**

05

Improving the Chosen Model's Performance: Once we've selected the top-performing model, our next step is to improve its effectiveness. We'll achieve this by **increasing the quality of the data** to enhance the model's MAE or MSE.

Select the Best-Performing Model for Prediction

Model	K-fold				20% 80%			
	MAE	MSE	RMSE	R-squared	MAE	MSE	RMSE	R-squared
Baseline Linear Regression	0.0586	0.0078	0.0883	0.3983	0.0604	0.0077	0.0880	0.3590
Random Forest	0.0401	0.0056	0.0749	0.5668	0.0442	0.0061	0.0781	0.4951
XGBoost	0.0439	0.0066	0.0810	0.4937	0.0468	0.0070	0.0839	0.4174
Bayesian Linear Regression	0.0587	0.0078	0.0884	0.3971	0.0606	0.0078	0.0883	0.3552
Lasso ElasticNet	0.0772	0.0130	0.1139	-0.0011	0.0772	0.0121	0.1101	-0.0040

- The **Random Forest model is the best-performing model** across both splits.
- It consistently demonstrates relatively **low values for MAE, MSE**, and RMSE, coupled with higher R-squared values compared to other models.
- The Random Forest model performs slightly **better with the cross-validation** split compared to the 20% testing-80% training split.
- Thus, we used the Random Forest Regression model with cross-validation to perform the remaining data science processes.

Improving the Chosen Model's Performance

We'll achieve this by increasing the quality of the data to enhance the model's MAE or MSE. The details of the process include:

1- Removing outliers

Regression models are known to be sensitive to outliers, therefore we will attempt to eliminate them and observe any potential effects on the model.

2- Assessing and Removing Insignificant Features

We've decided to remove the 'Captions' column, since the majority of videos have a caption value of 1, and examination through feature importance plots will confirm that captions are not significant and do not impact the model.

Improving the Chosen Model's Performance

3- Building a new model after improving the data

We built a new Random Forest Regression model with cross-validation without outliers and Insignificant Features to check if it will enhance its performance.

Model	MAE	MSE	RMSE	R-squared
Random Forest Model (Before)	0.0401	0.0056	0.0749	0.5668
Random Forest Model (After)	0.0344	0.0038	0.0620	0.5683

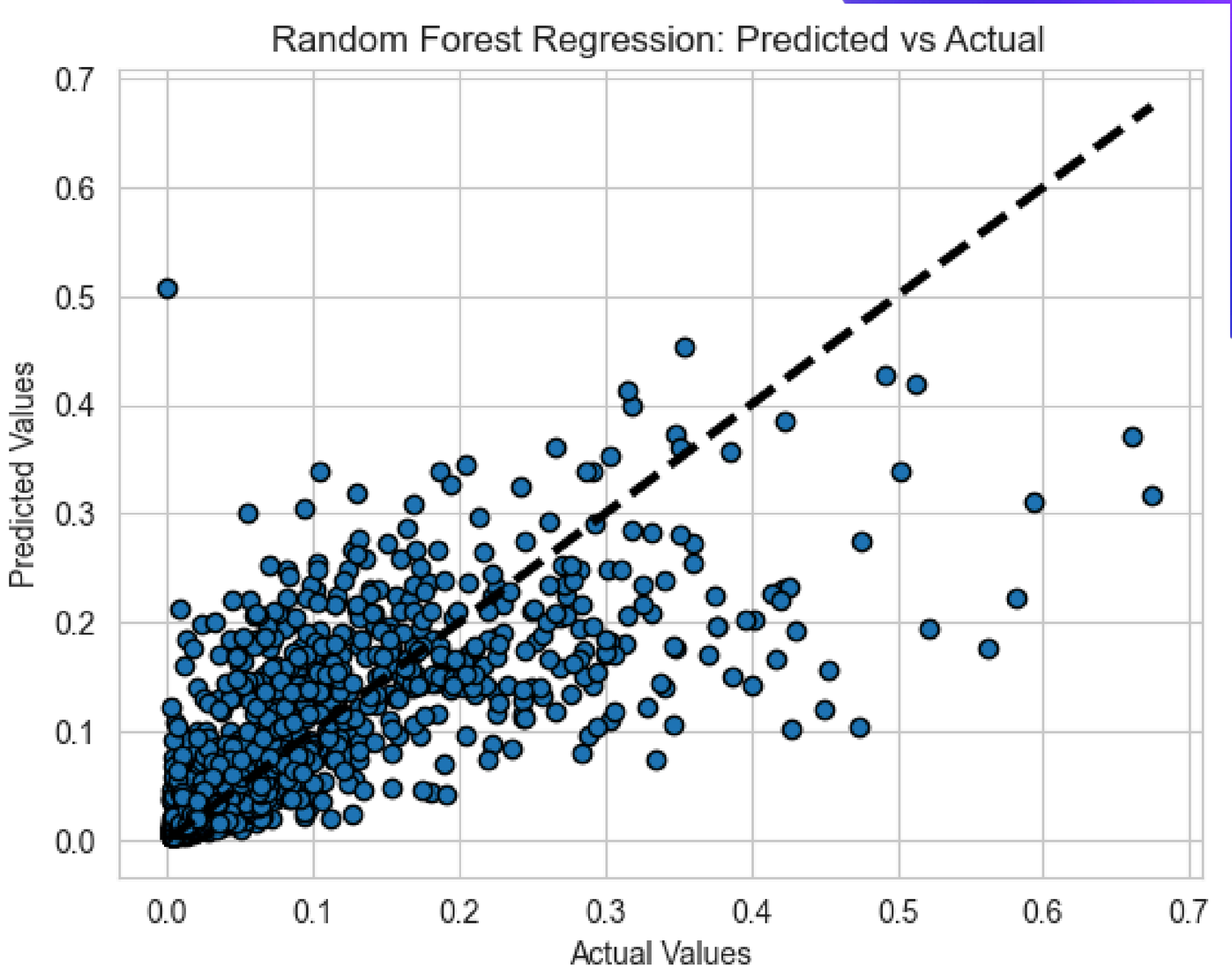
the model's performance improved as shown by an increase in MAE and MSE.

Scatter Plot of Predicted vs. Actual Values:

1- Prediction Accuracy: The concentration of blue dots around the dashed line suggests that **the model has a good level of accuracy**.

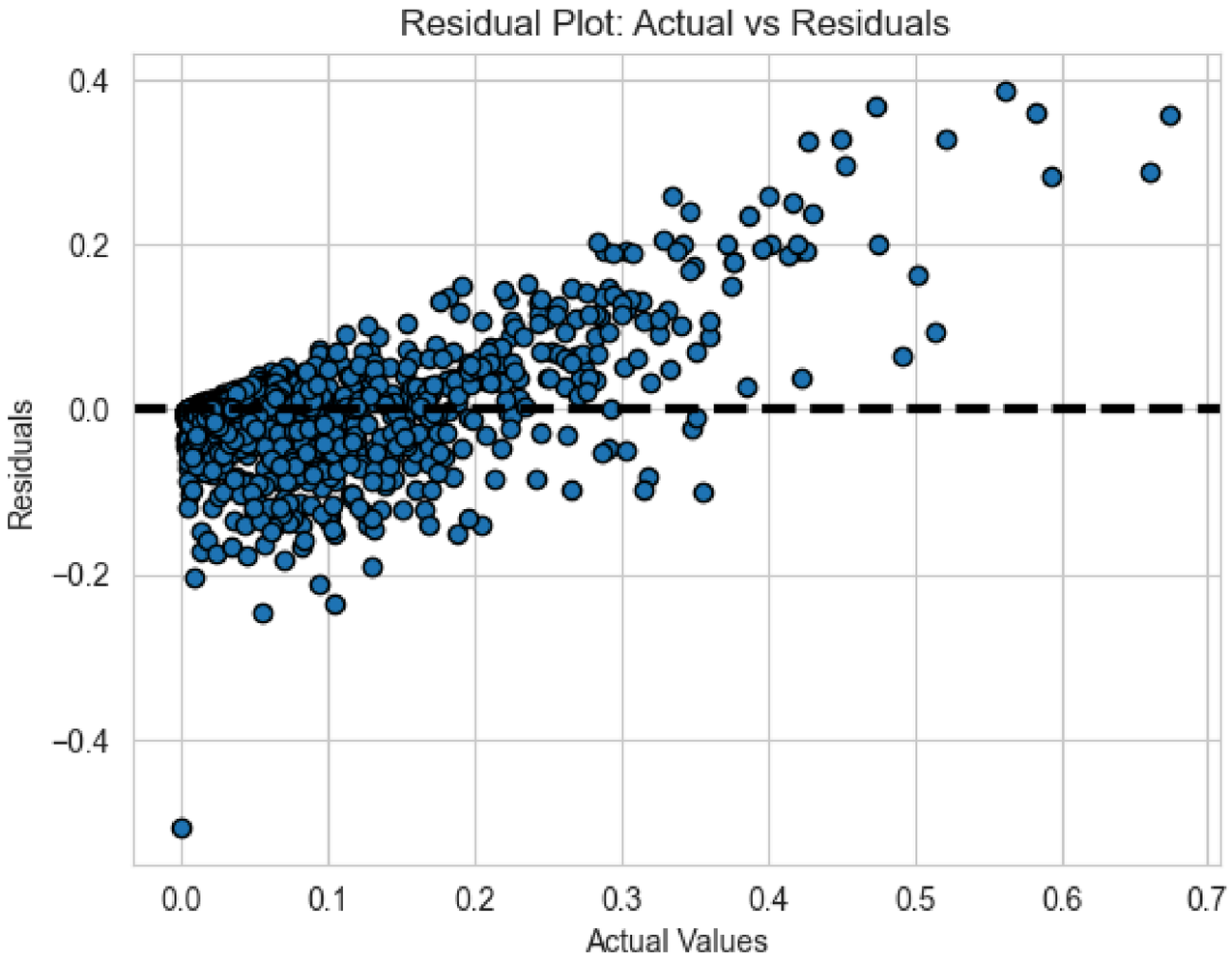
2- Prediction Errors: The majority of the data points lie above the diagonal line. This means **the model tends to overestimate the actual values.**

3- Model Fit: The overall distribution of predictions close to the prediction line indicates **a reasonably good fit** of the model.



Residual Plot:

Deviation from Zero:
The fact that the residuals curve upwards as the actual values increase indicates that **the model underestimates the higher actual values and overestimates the lower actual values.**



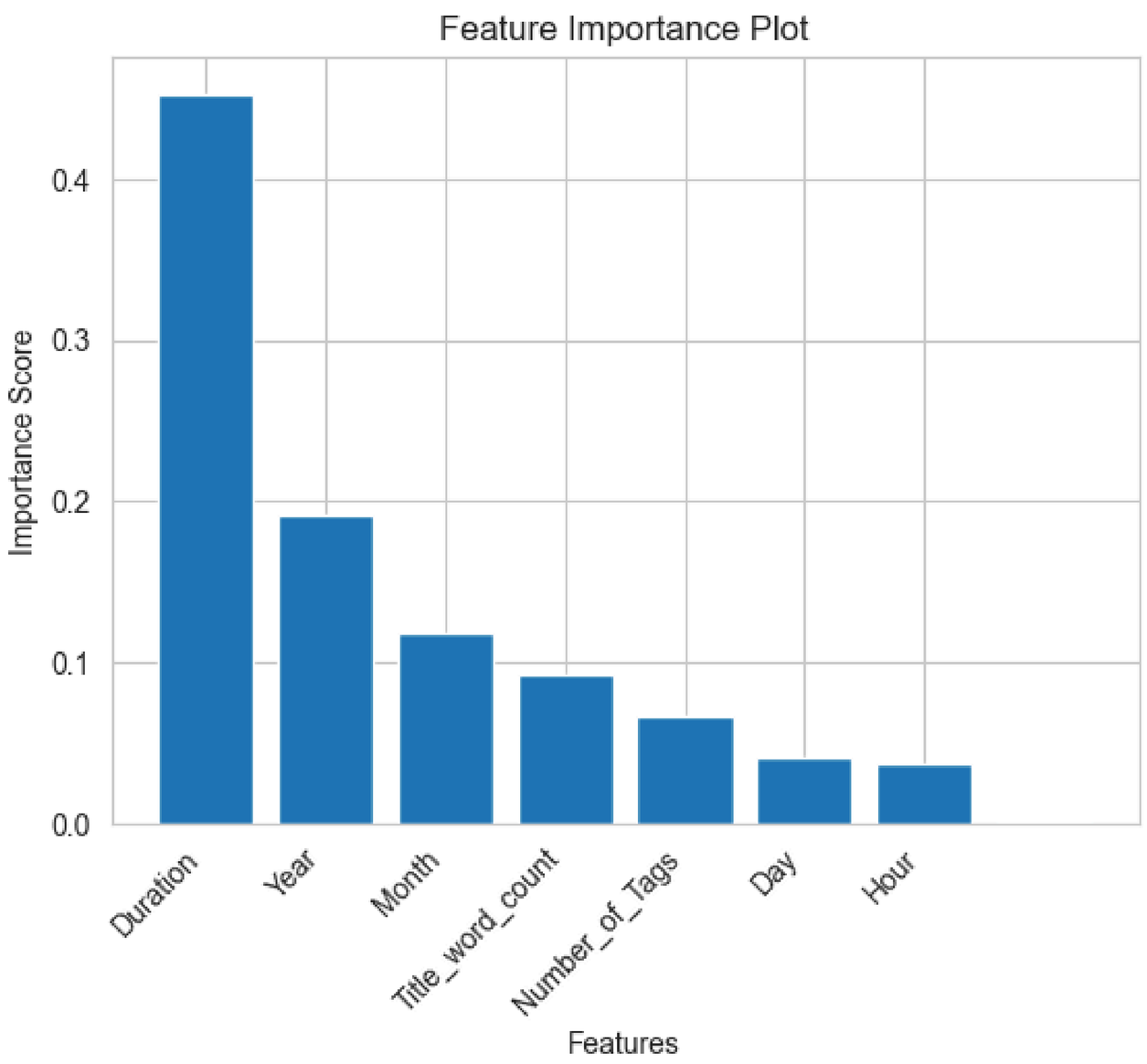
Feature Importance Plot:

1- Duration is the most important feature.

2- Year is also relatively important.

3- Month, Title word count, Number of tags all have a moderate level of importance.

4- Day and Hour have the least importance in predicting views according to this model.



Technical difficulty:

Technical difficulty :
We faced a challenge where the best model used a cross-validation split, resulting in no direct access to the tested and trained data. Consequently, visualizing the model became difficult.

The solution:
After researching visualizing models using cross-validation splits, and after several attempts, we successfully visualized the model.

How the models could be improved?

Problem

From our experience, YouTube video views are primarily influenced by the thumbnail, the title's meaning, and the video duration. These elements typically encourage viewers to click on the video.

However, our model only incorporated the duration factor. The other input fields showed weak correlations with views, rendering them unhelpful for accurate model predictions.

Solution

we should consider incorporating additional features such as thumbnail characteristics and title sentiment analysis.
By including these factors alongside video duration, we can capture more of the influential elements that prompt viewers to click on videos.

Assessment findings on LLM in data science projects

- Using LLM on our dataset showed some difficulty, as the model kept crashing in the modeling task
- The LLM’s model performed worse, this could be due to mistakes in the preprocessing
- However, the LLM provided some different insights and interpretations, and utilized new models

- LLMs are still not ready to be used officially for data science. However, they provide a great starting point, save time, and aid in exploring new data science techniques

future work

01

Building a pipeline that collects video's data over time, then analyzes the engagement and viewership, will provide additional features for analyzing the video and improving the work.

02

Analyzing the thumbnails to determine if certain features, such as colors or shapes, can attract more views.

03

Incorporating external data sources to find more features that impact views. Such as demographics, trends, and global events.

Thank you

Prepared by:
Bashair Alsadhan
Rama Alshebel
Waref Alyousef
Rana Alsayyari
Nora alwohaibi

Supervised by:
Dr. Khulood Alyahya
Dr.Reem Alqifari