**Facebook Marketplace Dataset**

The Facebook Live Sellers in Thailand dataset contains information about the Facebook pages of 10 Thai fashion and cosmetics retail sellers. Below is a description of the dataset:

1. Title: Facebook Live Sellers in Thailand Dataset

2. Source: The dataset is sourced from the UCI Machine Learning Repository.

3. Data Type: The dataset is in a tabular format, typically stored in a CSV (Comma Separated Values) file.

4. Number of Instances: There are a total of 7050 instances (rows) in the dataset.

5. Number of Attributes: The dataset initially consists of 16 attributes (columns). After removing redundant columns, there are 14 attributes remaining.

6. Attribute Information:

- status\_id: Unique identifier for each status post.

- status\_published: Date and time when the status post was published.

- status\_type: Nature of the status post (e.g., video, photo, status, link).

- num\_reactions: Number of reactions (e.g., likes, loves, wow, haha, sad, angry) received on the status post.

- num\_comments: Number of comments received on the status post.

- num\_shares: Number of shares received on the status post.

- Additional numerical and categorical attributes related to engagement metrics and status post features.

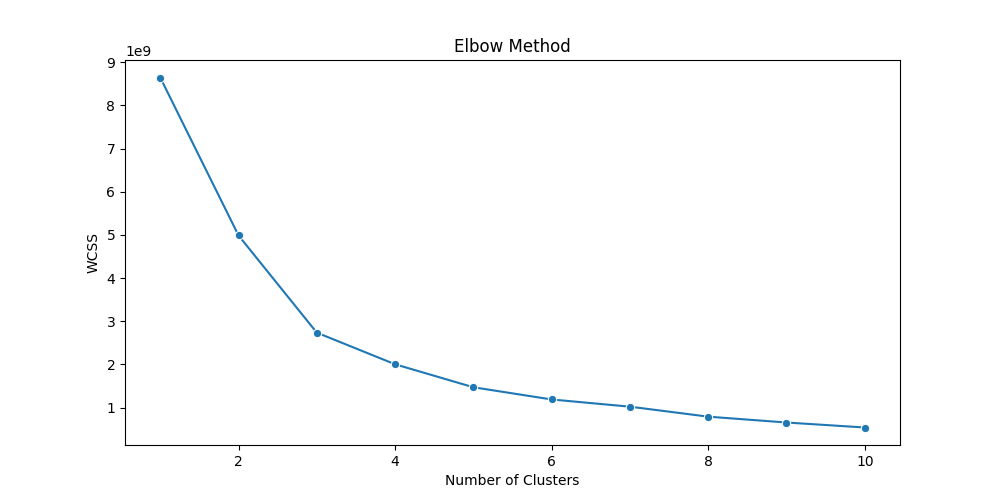
7. Missing Values: The dataset may contain missing values, which need to be handled during data preprocessing.

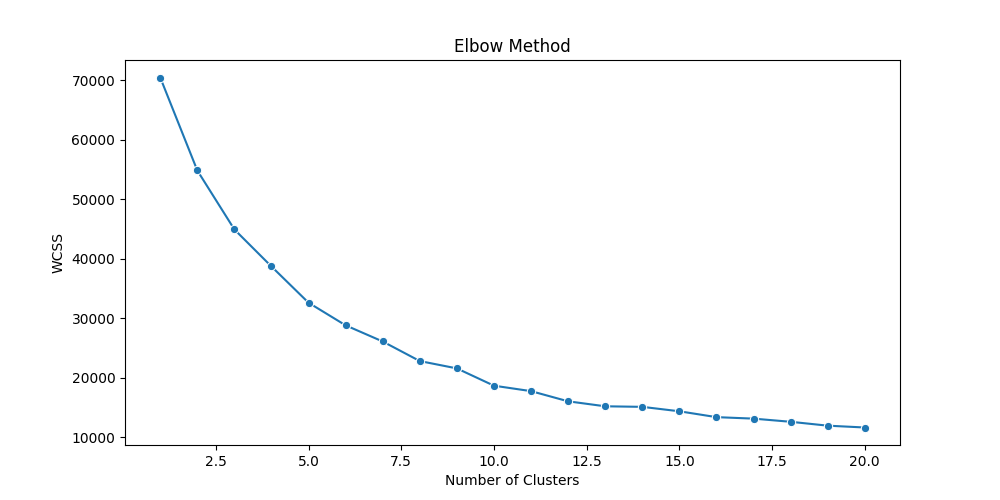
**Questions**

1. How does the time of upload (`status\_published`) affects the `num\_reaction`?
2. Is there a correlation between the number of reactions (num\_reactions) and other engagement metrics such as comments (num\_comments) and shares (num\_shares)? If so, what is the strength and direction of this correlation?
3. Use the columns status\_type, num\_reactions, num\_comments, num\_shares, num\_likes, num\_loves, num\_wows, num\_hahas, num\_sads, and num\_angrys to train a K-Means clustering model on the Facebook Live Sellers dataset.
4. Use the elbow method to find the optimum number of clusters.
5. What is the count of different types of posts in the dataset?
6. What is the average value of num\_reaction, num\_comments, num\_shares for each post type?

MY SOLUTION / WORKING

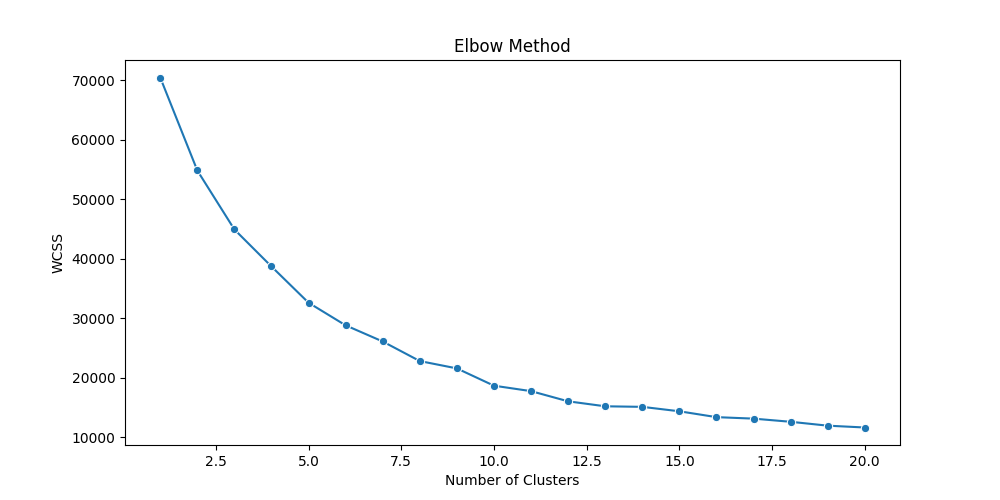
1. Looking at the Dataset
   1. Status\_published has two different data types that might be a problem later. I will deal with that. (Didn’t need to. Pandas did it itself)
2. Answering the first question.
   1. I will find the correlation between the status\_published and num\_reaction and plot a graph using matplotlib.  
      But… status\_published is not numeric type, how will I find relation between them.
   2. I split the status\_published into date and time and then found a relation between the hour of the day and the number of reactions in that hour.
   3. Done. Made 6 Graph will present the findings towards the end of the report.
3. Answering the second question.
   1. I made a tempory df with only the engagement metrics and created a correlation matrix that I plotted. It has the strength and the direction noted.
4. Answering the third question
   1. I will first convert the categorical data of status\_type into numerical
   2. Scale the features because num\_reactions is ranging from 0 1000 and num\_hugs or num\_wow ranges from 0 to 100 only.
   3. I will check the performance without scaling as well.
   4. Then using elbow method I will find the optimal value of k and then start training the cluster.

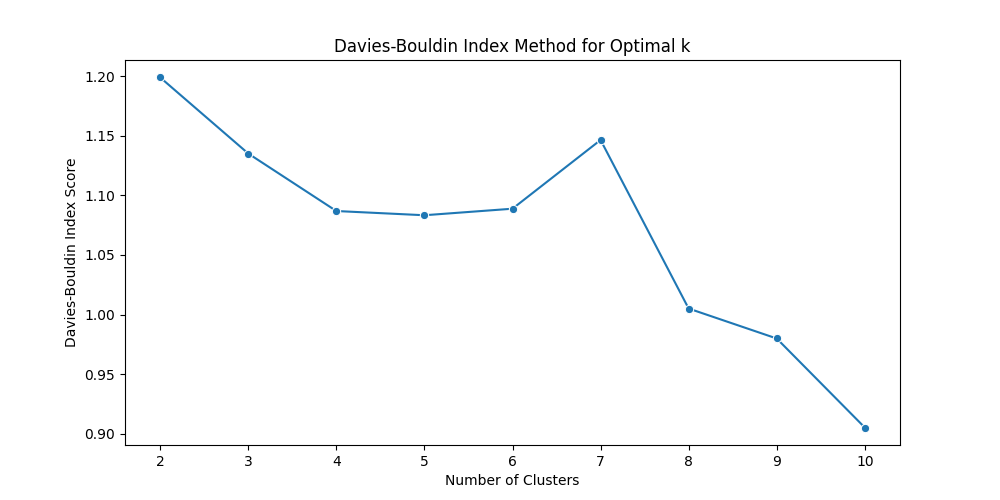
Without Scaling, the graph looks good and has a clear value for k, but the difference in the range of the feature column. Therefore, without scaling is not an option. 

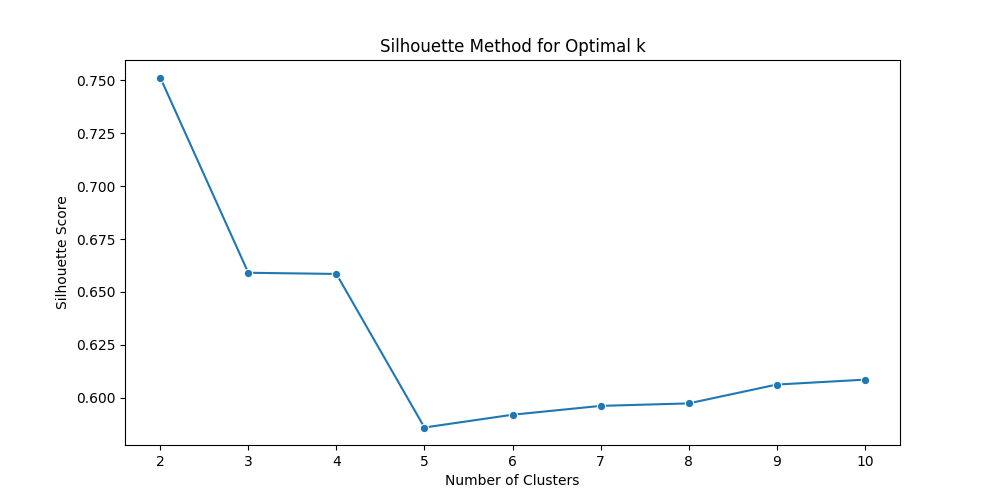
With Scaling, the value is not clear and coming out to be 10 which is again nit advisable. 

Therefore, I will employ some other method to find out the value of k.

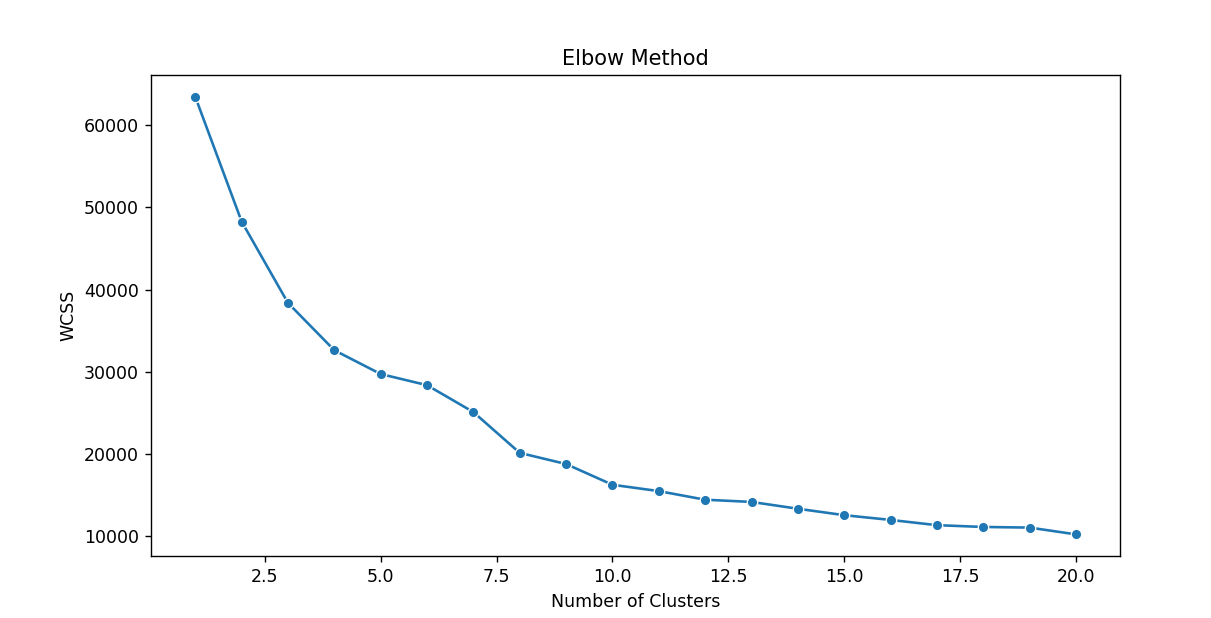
I employed Silhouette and David Bo.. Index method. At first I was not able to find an optimal value for k since elbow method was not good, silhouette gave the k value as 2, and dbi gave it as 10.

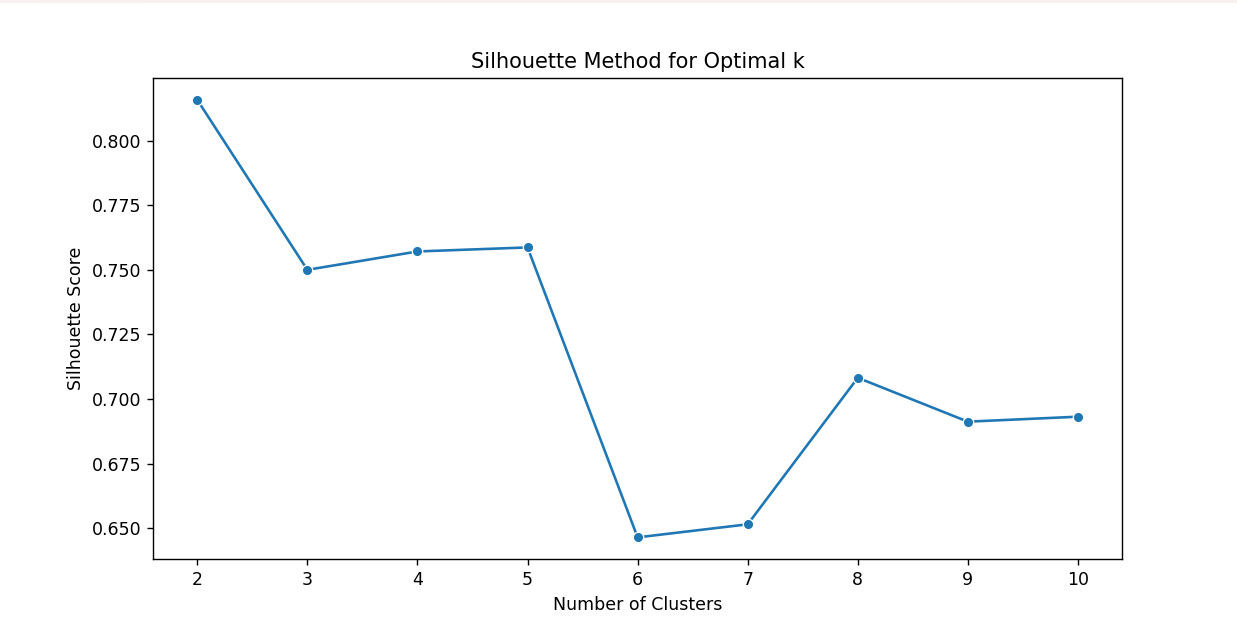


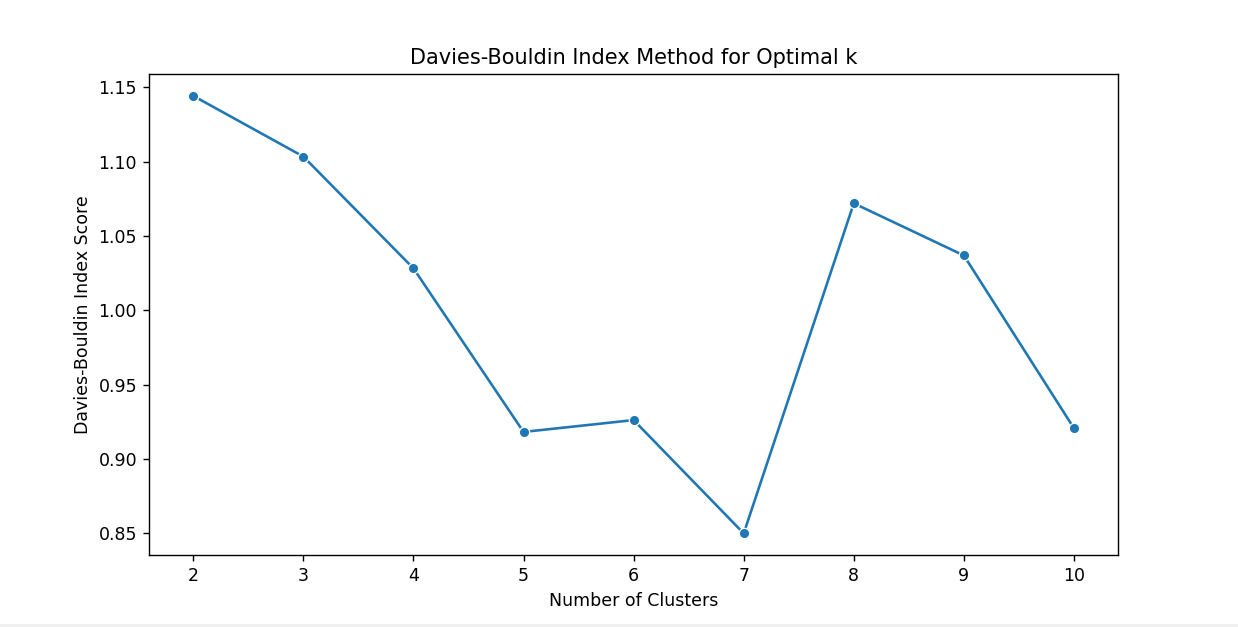




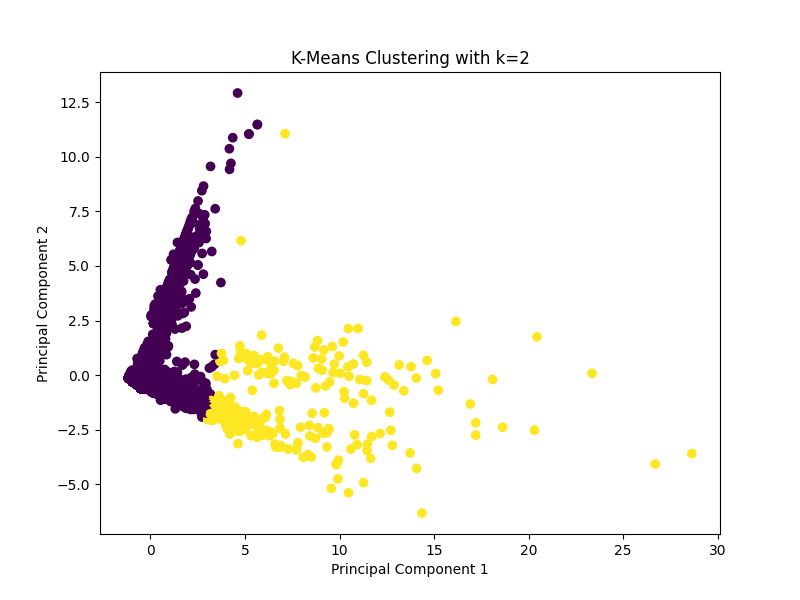
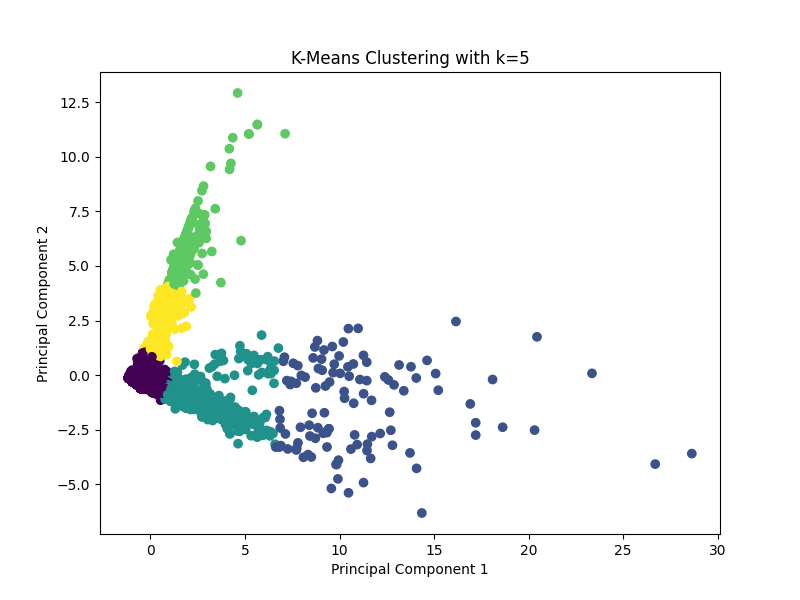
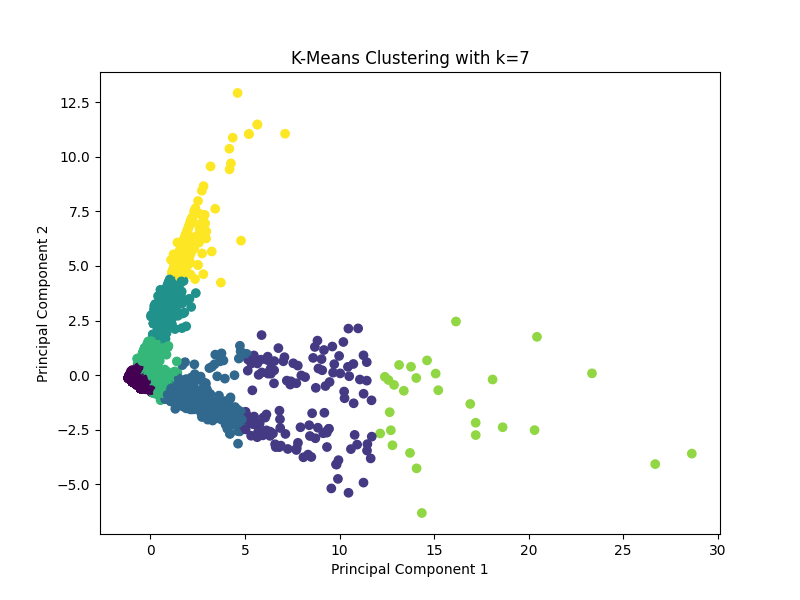
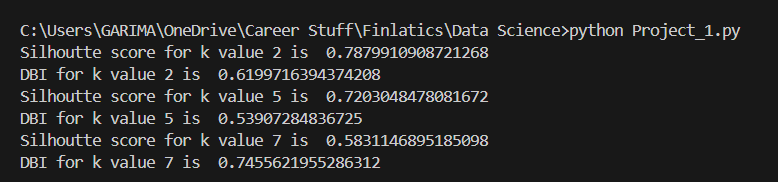
Later, I realized that it was because I was considering the categorical value in training the model. This was throwing off the model’s accuracy and misleading it since categorical value column when converted to numeric does not depict a continuous relationship. Scaling them introduces an artificial relationship and throws off the model.

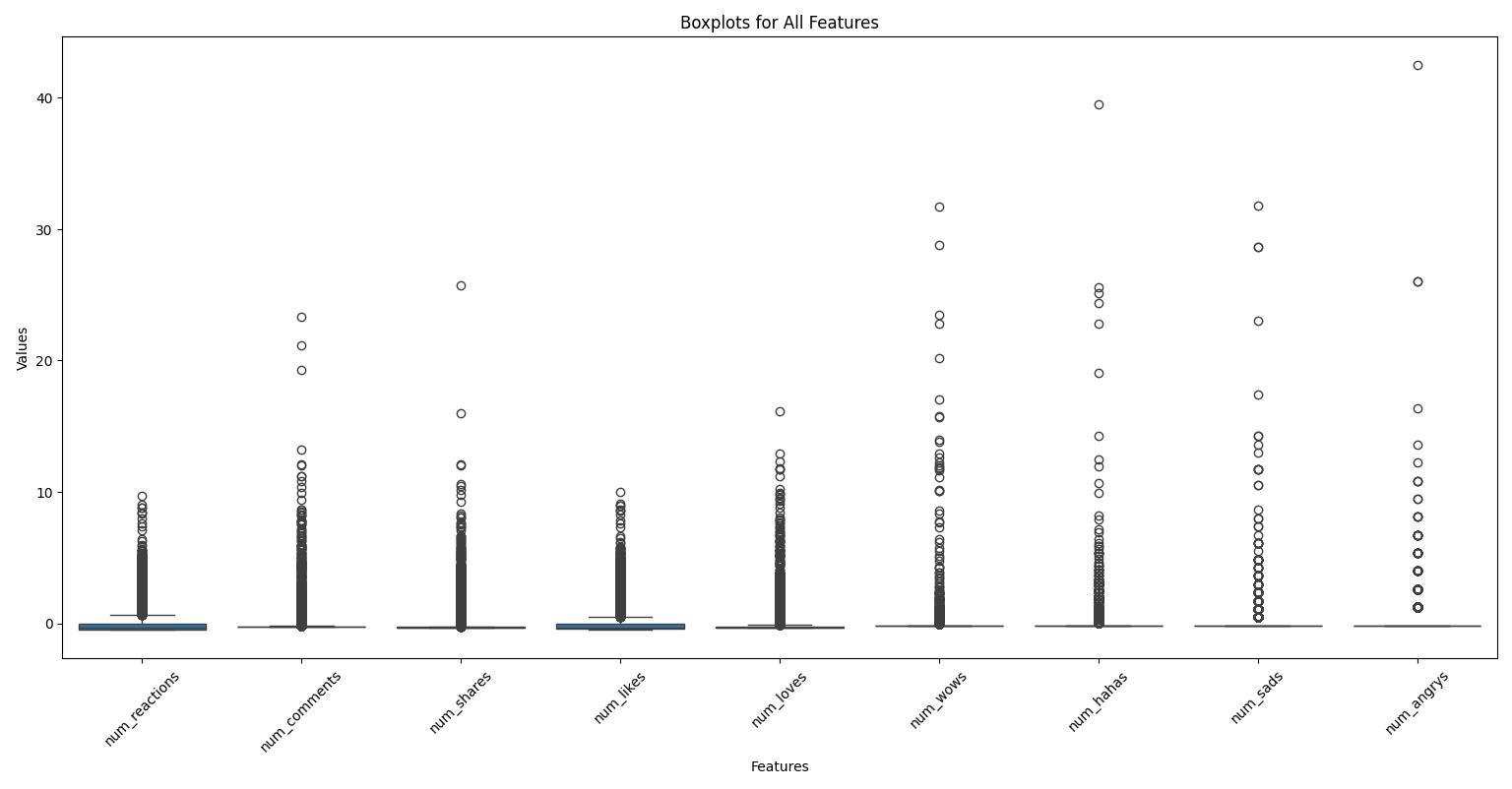






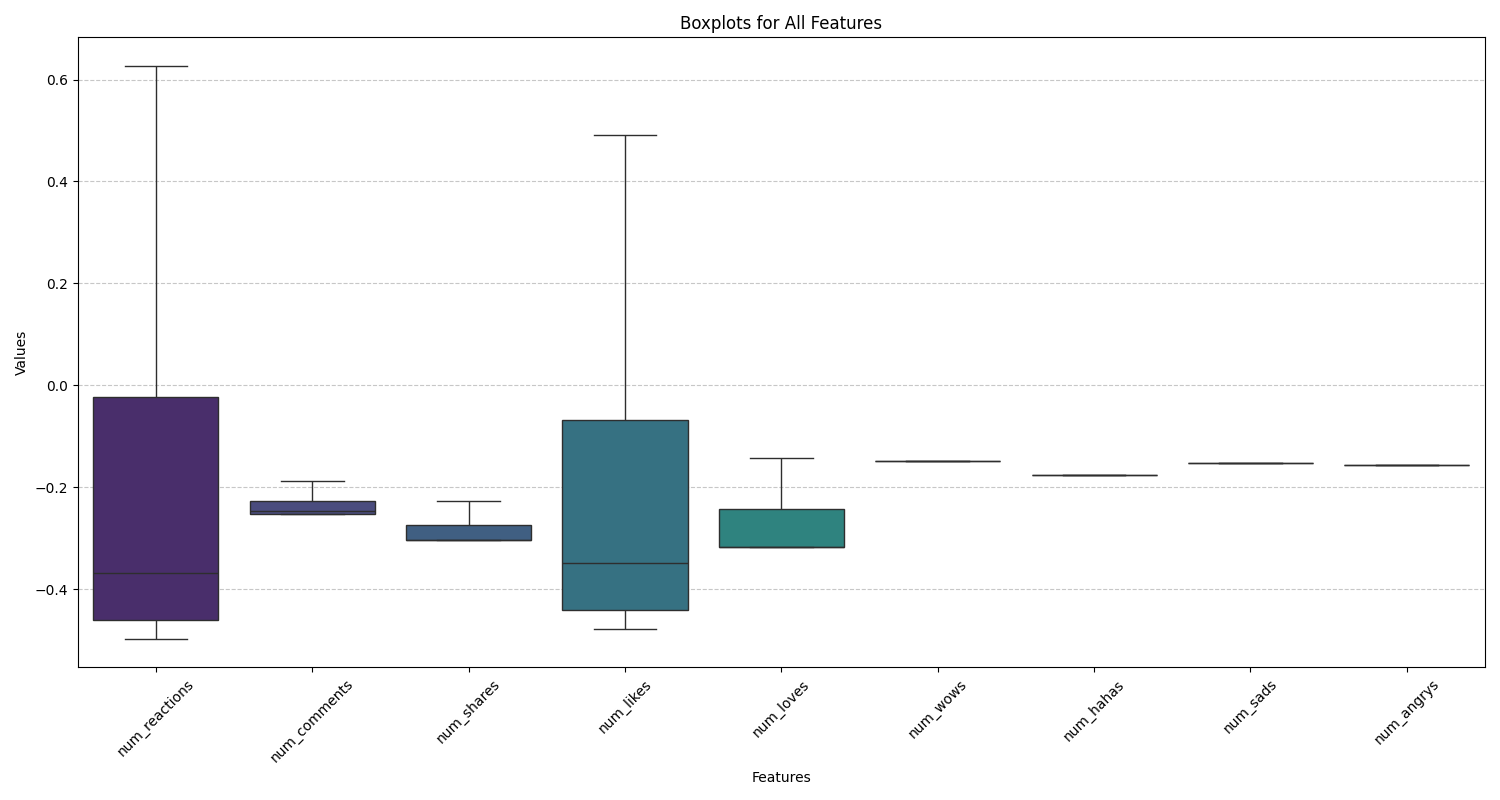
* The Silhouette Score measures how well each data point fits into its assigned cluster (higher is better).
* This graph shows the Davies-Bouldin Index against the number of clusters (lower is better, indicating better-defined clusters).
* **It is best to consider k=2, 5, and 7 as the only optimal value for K**
* Now, it is time to train the model using the k values.
* The next step is to plot the scatter plot. Since, our model is trained on multiple features to plot it on 2D graph, I will have to apply PCA.
* This is the scenario right now



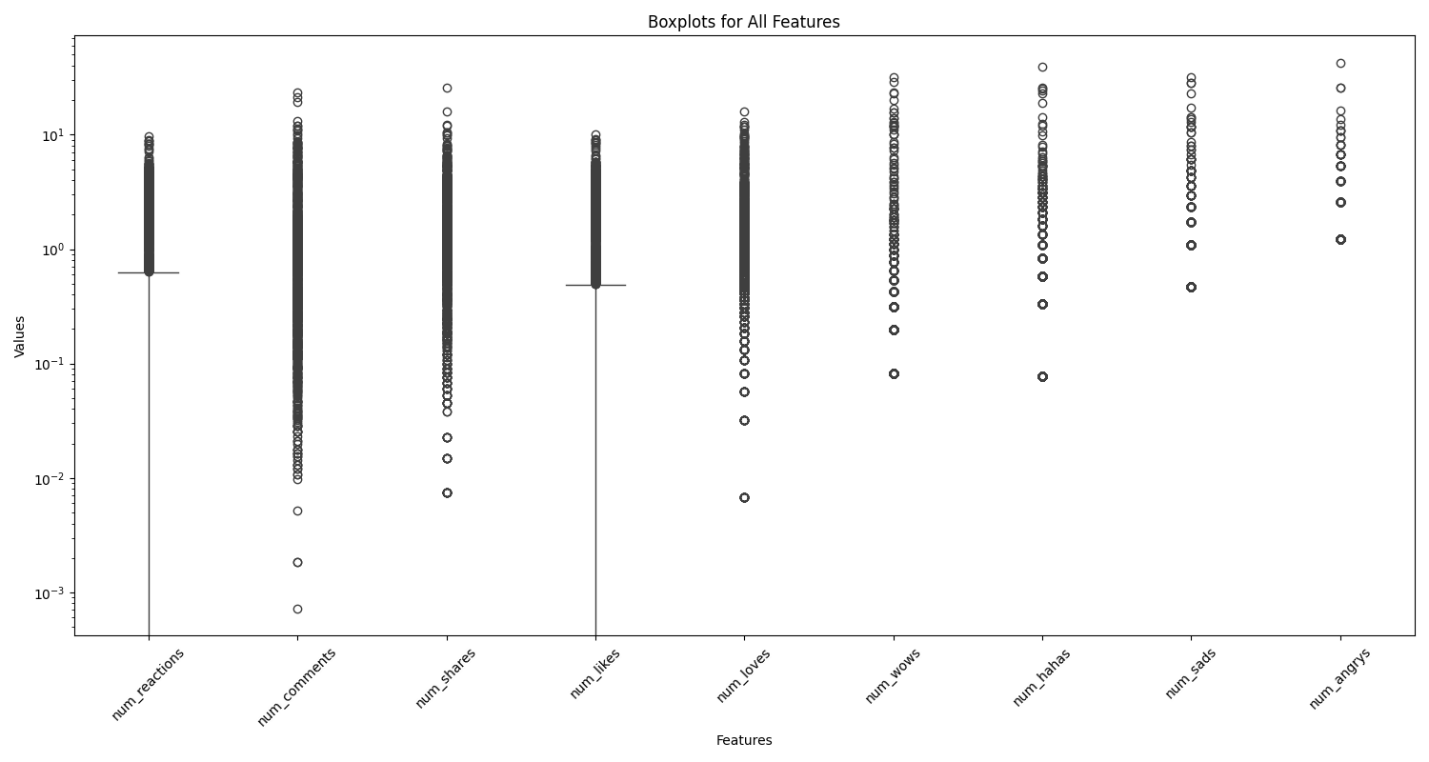


Improving the Graph

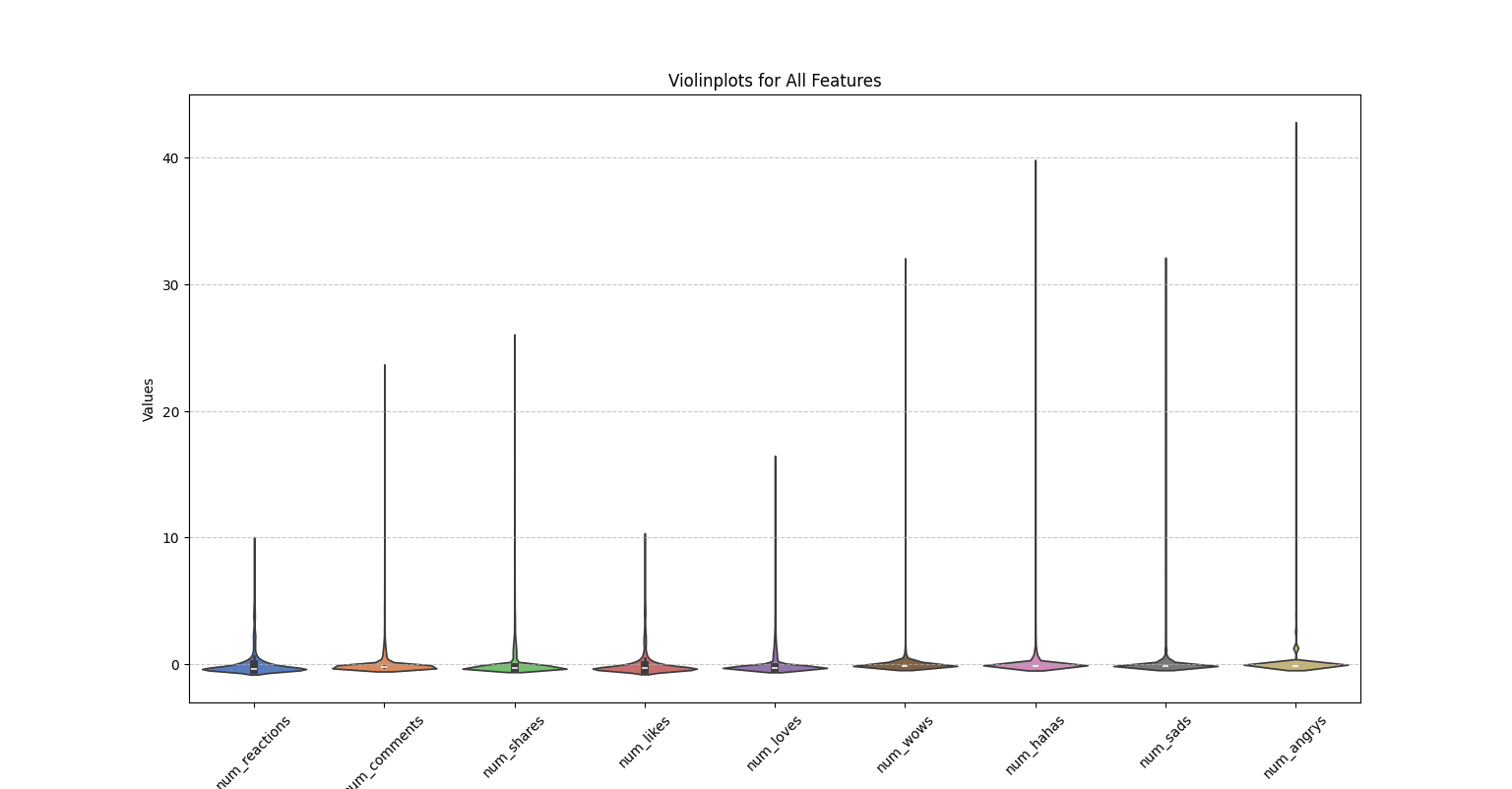
1. Hide Outliers



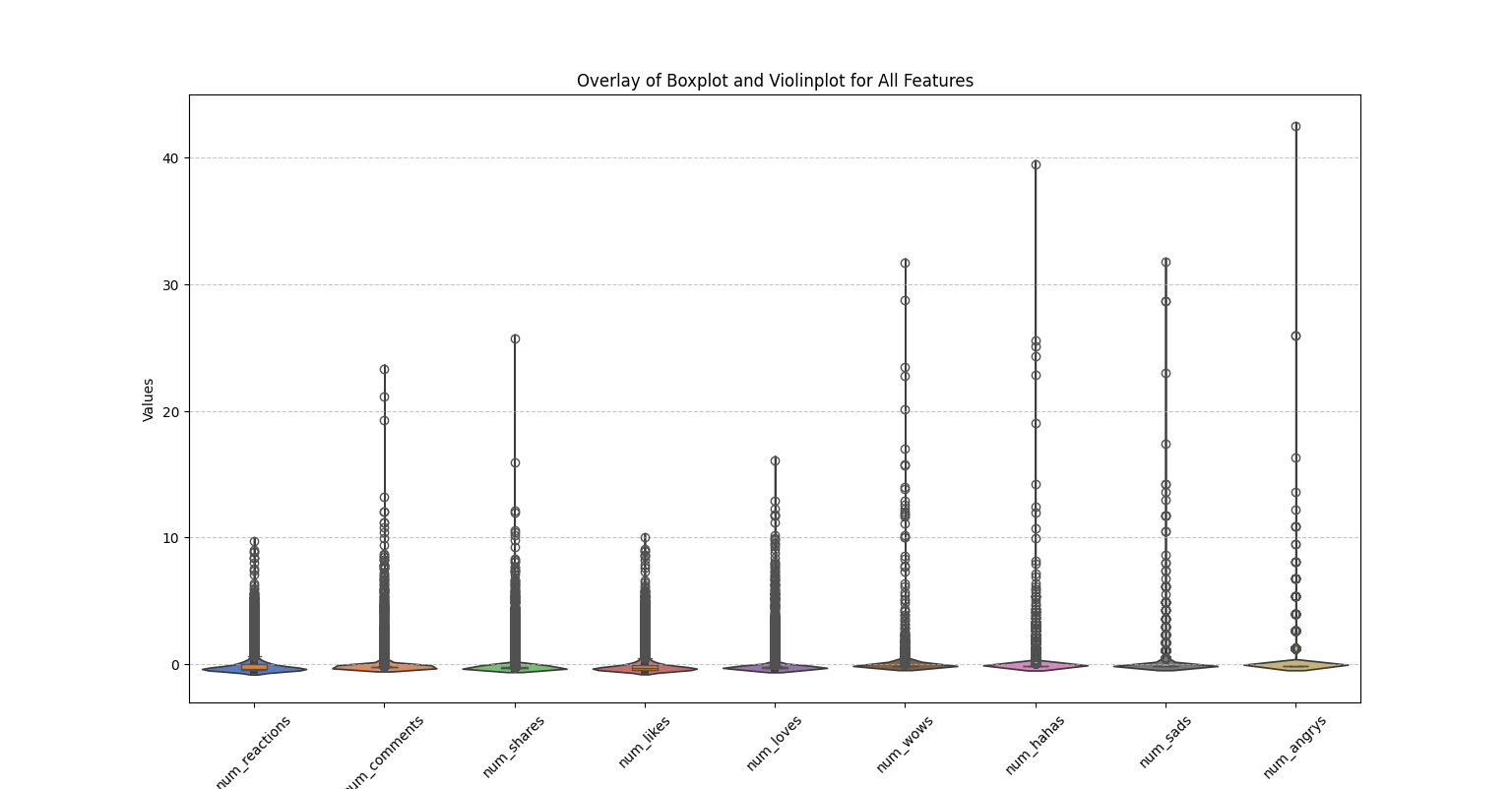
1. Using logarithmic scale on y axis



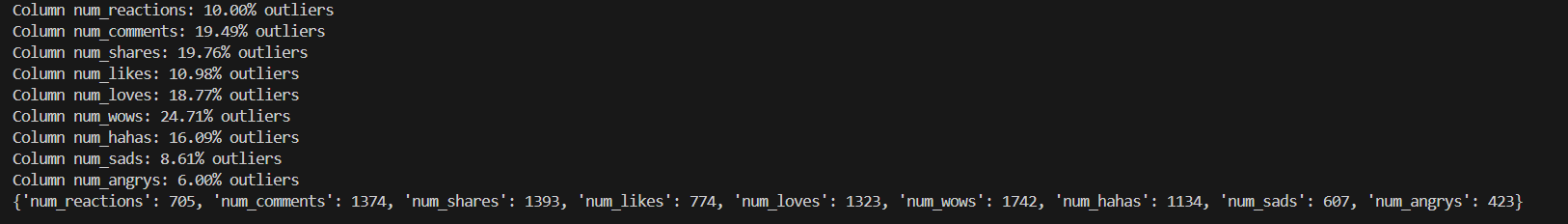
1. Trying Violin Plot



1. Violin and Boxplot overlayed



* Since K-means is a distance-based clustering, outliers must be affecting the models’ performance. Therefore, we will deal with that first.



* Strategy for Handling Outliers

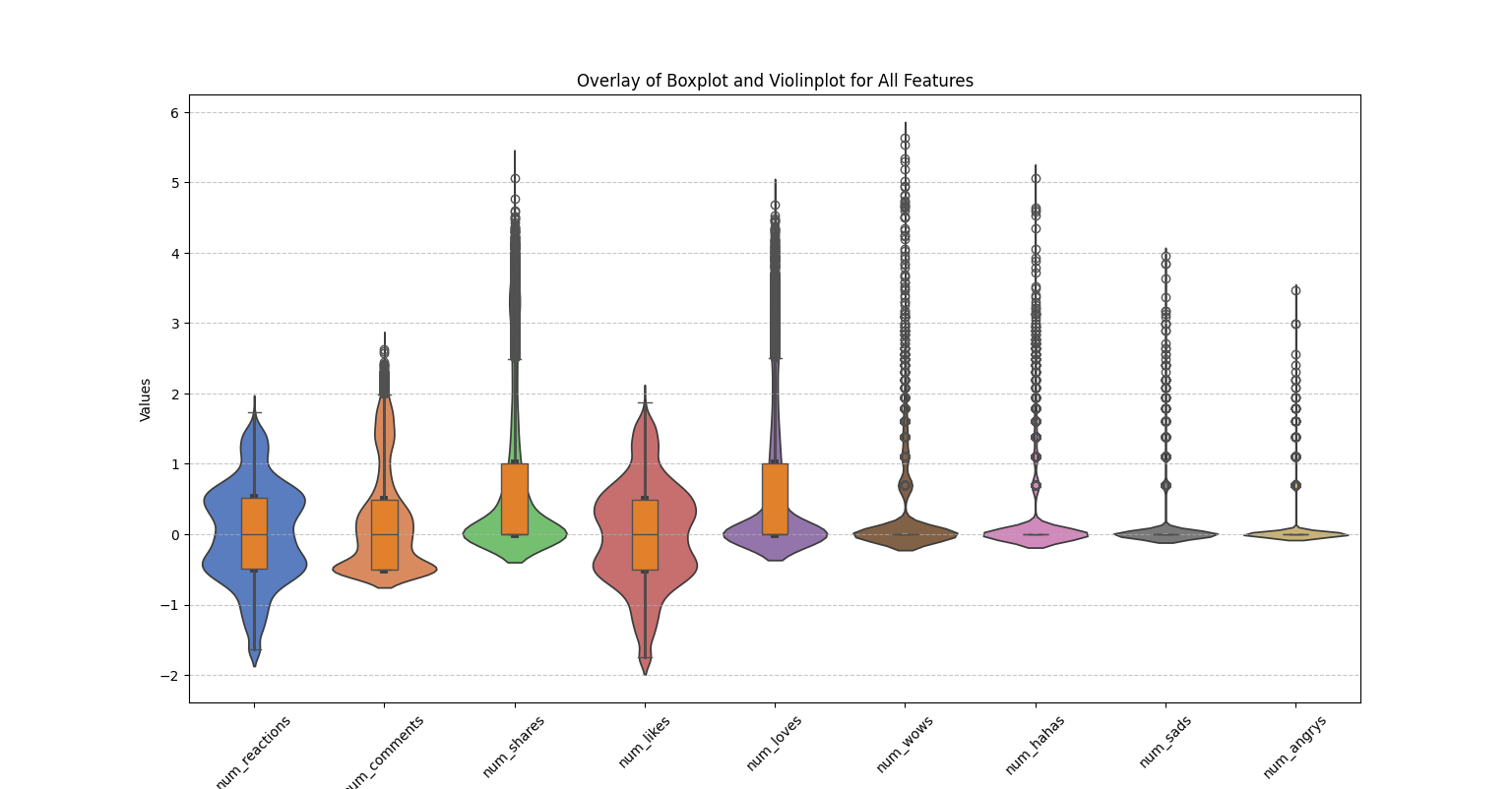
Given you have 7050 data points, here's a balanced approach to consider:

**Quantify Outliers:** The percentage of quantifiers is more than 5-10%. Therefore, removing outliers will not improve the performance of the model.

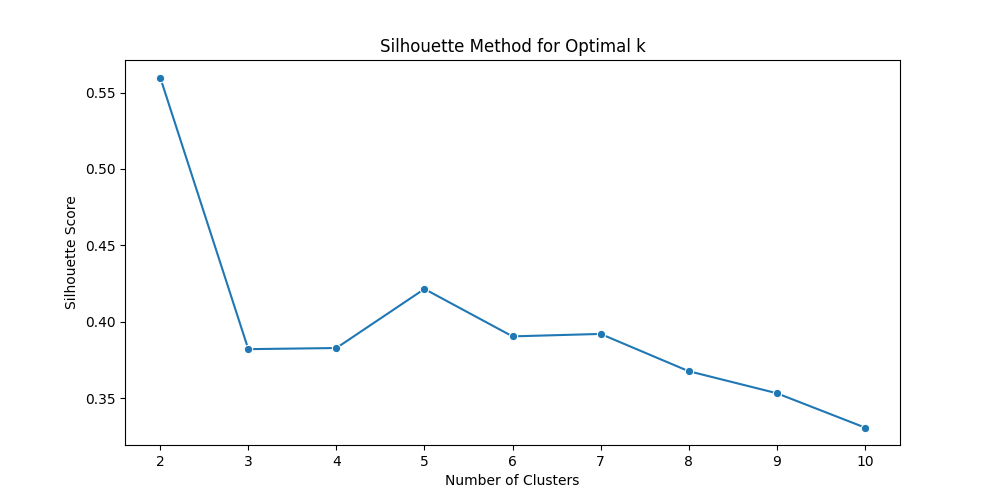
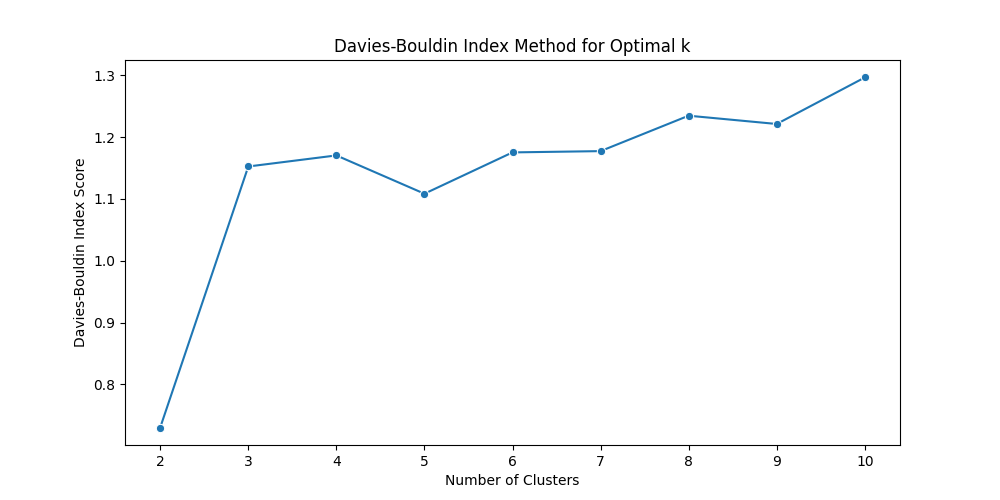
**Decide on a Course of Action:**  Also, through domain knowledge we know that outliers here just mean viral posts and keeping them would a rather good decision.

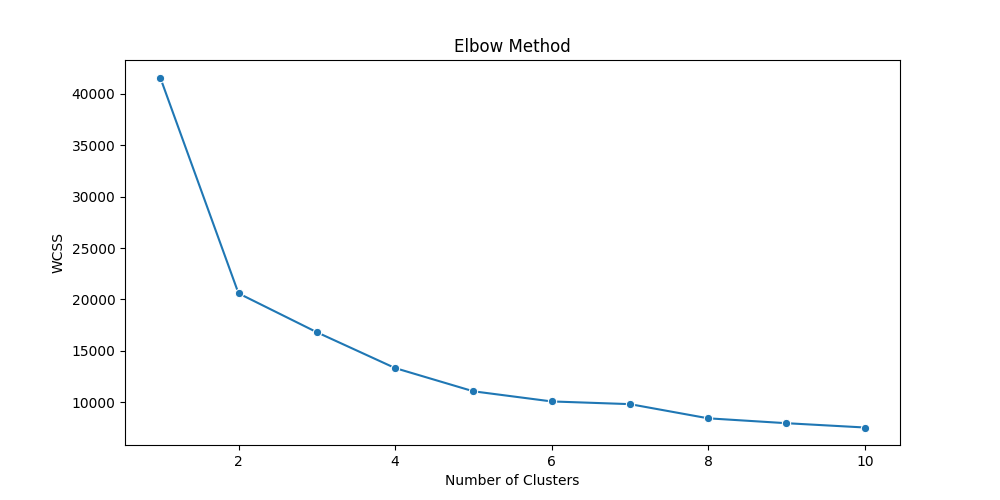
**Therefore, the best course of action would be transforming the data, since I have already performed scaling.**

And after that, check the plot.

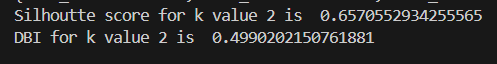


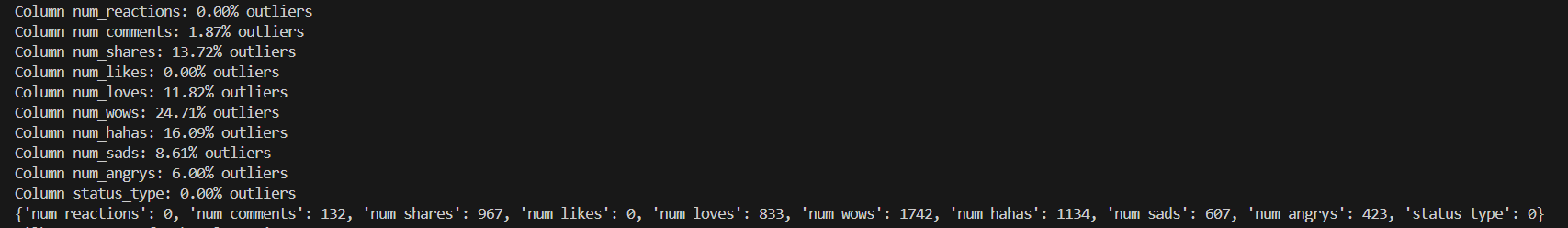
Now, I will try finding the value of k again and train the k-means clustering model.

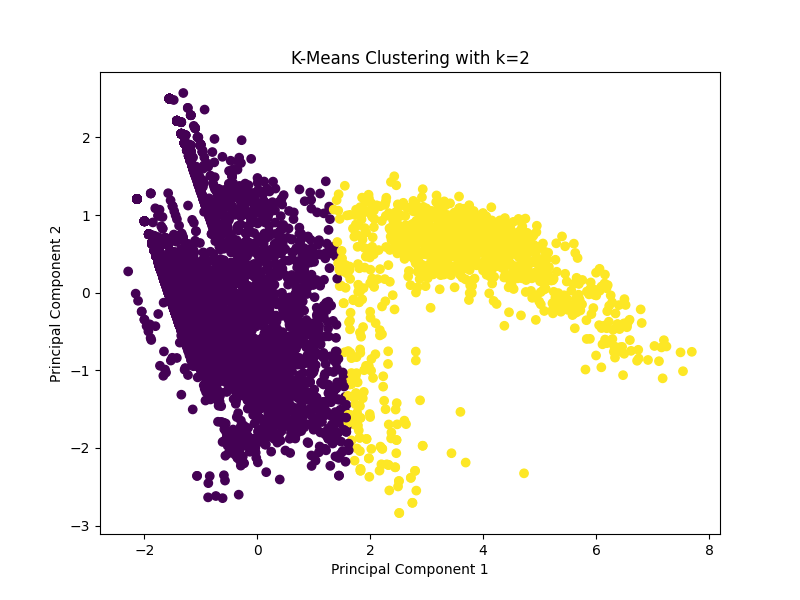
**BINGOOOOO!!! I GOT A CONSISTENT K VALUE.** 



**AS YOU CAN SEE k=2 is the clear choice.**

****





REPORT

Problem Statement

Why k means clustering?

Steps

1. Preprocessing, standard scaling
2. Finding the value of K, with and without scaling
3. Accidently took categorical value in consideration, removed that
4. K = 2,5,7
5. The score was not that good, so this time I dropped the outliers and then trained again.

Final conclusion and output

Scatter Plot that shows clusters and can be used later.

ON medium post

Feedback loop.