ADVANCED MACHINE LEARNING – YOUTUBE DATA MINING

Lecturer:
Dr. Chen Hajaj

Team members:
Ido Kapel
Aviram Klaiman

The problem: Understanding YouTube's Dynamics

- With millions of videos uploaded, understanding what makes a video popular or identifying successful content strategies is complex.

- Viewer preferences and content trends on YouTube are constantly evolving, requiring up-to-date analysis.

Original Goals

To grasp what drives a video to go viral, we must first discern the unique characteristics that distinguish one video from another.

Our goals are:

- Classifying YouTube videos into meaningful clusters.
- Identifying outliers or anomalies that deviate from common patterns.

Other methods for this problem combines deep learning for dimensionality reduction and feature representation learning, facilitating more effective clustering.

Data set

- To collect our data we used YouTube Data API, and a short glance of the data:

video_id	published_at	title	channel_id	channel_title	view_count	like_count	dislike_count	favorite_count	comment_count
4NISmeBNXy8	2024-03-16T09:30:01Z	আজ বাড়িতে সকালে রান্না খাওয়া হলো||বাড়িতে হঠ...	UCHxgoBn9vWA_gBaOI3phY6w	Sundarban Cooking	3253	460.0	NaN	0	30.0
PGHeFbpaH4Q	2024-03-15T18:30:15Z	Quick and Easy Za'atar and Labneh Spaghetti | ...	UC1rlOwTqDuWkFj87HZYRFOg	NYT Cooking	39986	2126.0	NaN	0	155.0
fl2afHLEjyM	2024-03-15T17:37:01Z	Picking A Date Based On Our Cooking! (Mexican ...	UCDP7DZOgj8VTyhyVNup83QQ	Amp World	157804	4872.0	NaN	0	368.0
6HgEZBxTsKQ	2024-03-15T17:03:52Z	After Sehri To Iftar Very Busy Routine " 1st J...	UCzR_69NUGuTW-AM_y8PCh5A	Cooking With Shabana	70457	2687.0	NaN	0	311.0
z3e2JfjHMP0	2024-03-15T17:00:48Z	Coolest Miniature: Watermelon Jelly Egg Ideas ...	UC5nA6O_7aYs3zWEX83m3jnw	Miniature Cooking	9778	120.0	NaN	0	16.0

- With a list of the columns:

```
Index(['published_at', 'title', 'channel_id', 'channel_title', 'view_count',
       'like_count', 'dislike_count', 'favorite_count', 'comment_count',
       'duration', 'definition', 'topic', 'subscriber_count',
       'total_channel_views', 'channel_description', 'channel_published_at',
       'videoCount', 'tags', 'category_id', 'default_language',
       'default_audio_language', 'license', 'content_rating', 'share_count'],
      dtype='object')
```

Methodology

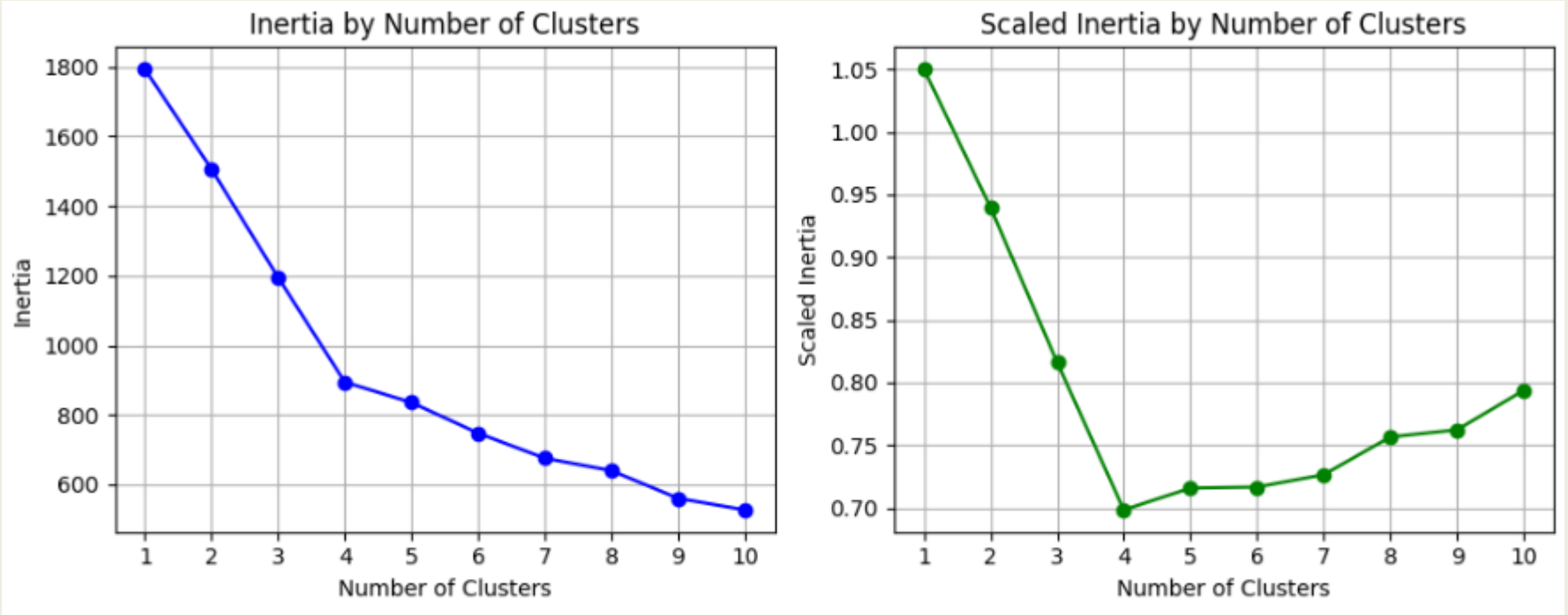
- We used t-SNE and PCA for visualization and dimentionality reduction, because of the high amount of features.
- For clustering, we used K-Means mainly because the spherical shape of the clusters in the data. After detecting some samples outside any cluster, we performed DBSCAN because it’s ability to identify outliers far from dense clusters.
- For anomaly detection, we used traditional techniques like BoxPlots, Modified Z-score, and other machine learning algorithms like DBSCAN, Local Outlier Factor, Elliptic Envelope, Isolation Forest and One Class SVM

Experiments

- We first performed the “Elbow Method” for selecting the best K – the number of clusters. We provided 2 plots, one is Inertia - within cluster sum of squares, and the scaled inertia for better results.
- Later, we performed another technique for evaluating the ideal K. We used the silhouette score, because the elbow method is not satisfying:
  - *We first applied the K-Means algorithm to the entire data set.*
  - *We then applied the K-Means algorithm to the dimension-reduced matrix.*
- After detecting some samples outside any cluster, we used DBSCAN for better clustering with respect to anomalies. Selected the right hyperparameters with a grid search function.
- For another perspective to detect anomalies, we also applied machine learning algorithms like DBSCAN, Local Outlier Factor, Elliptic Envelope, Isolation Forest and One Class SVM.

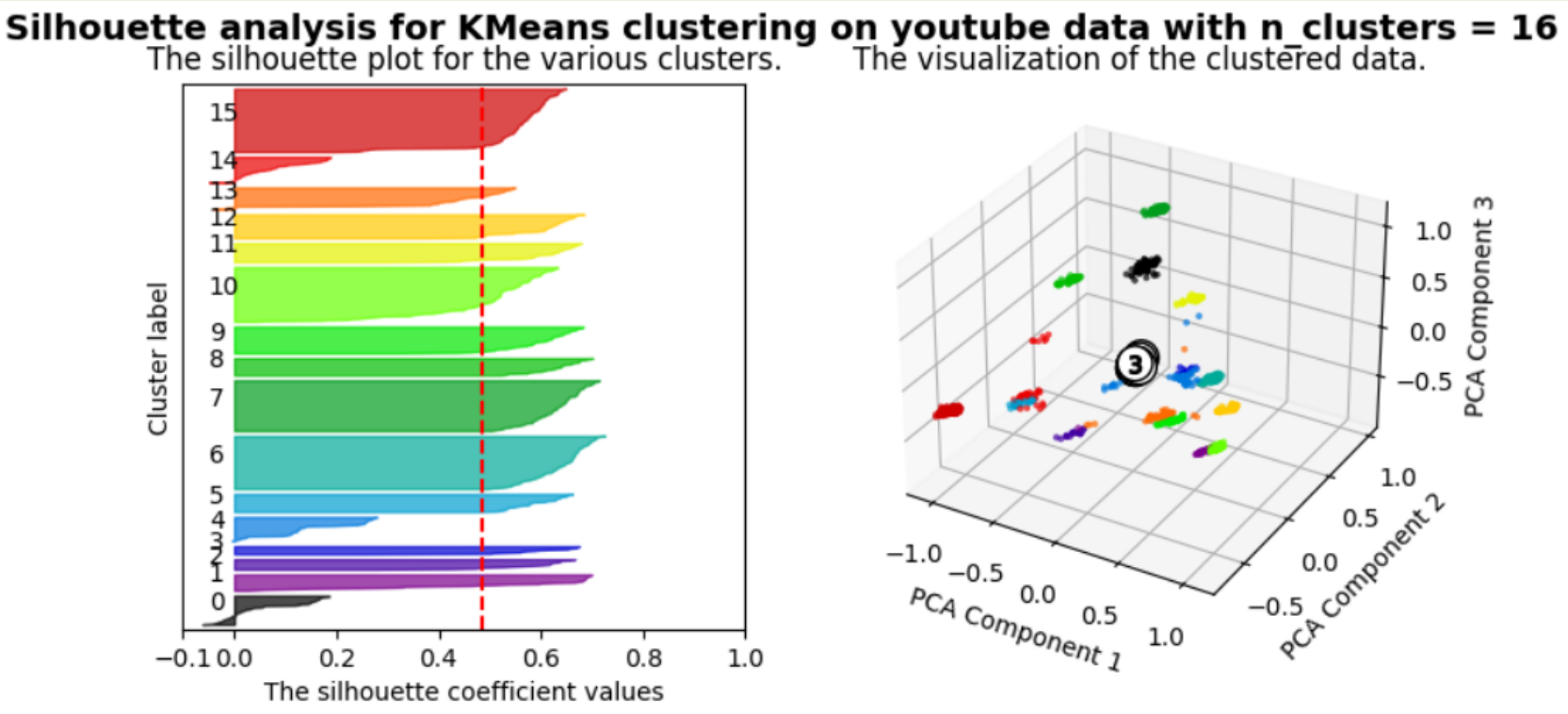
Results

- By applying the elbow method, with the inertia and scaled inertia plots, we found the right number of clusters is 4.



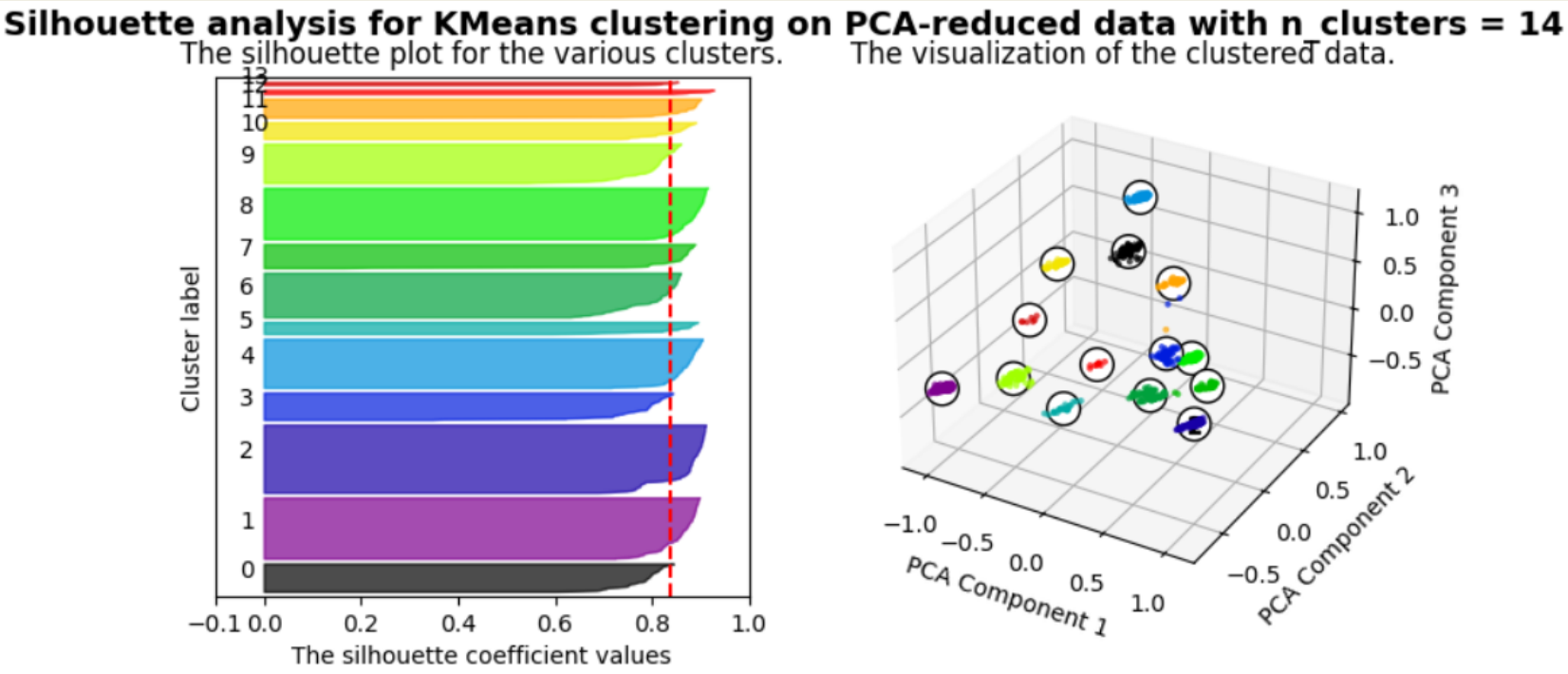
Results

- By applying K-Means to the entire dataset, we achieved a silhouette score of 0.486, with K = 16.



Results

- By applying K-Means to the dimension-reduced matrix, we achieved a much better silhouette score of 0.837, with K = 14.

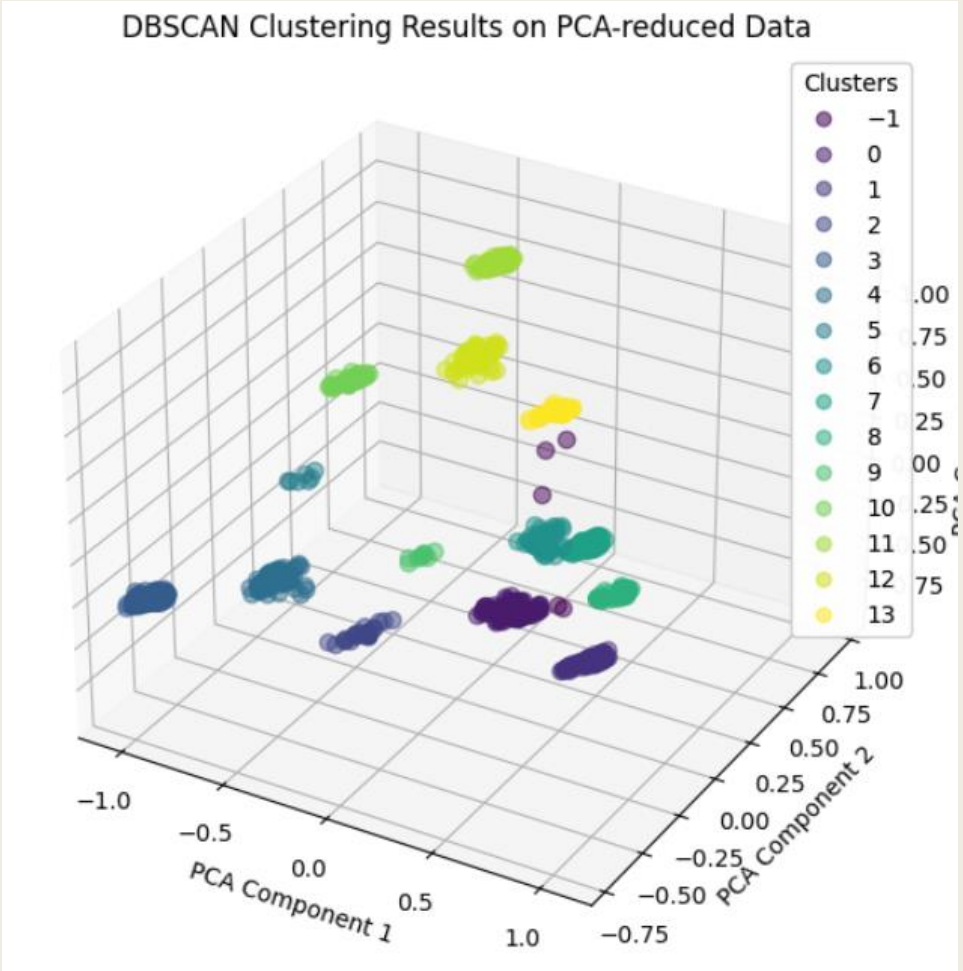


Results

- The different results of the elbow method and the silhouette were mainly because the silhouette measures both the cohesion and separation of every sample.
- Our initial clustering faced challenges due to the curse of dimensionality, where high-dimensional data makes distances between points less distinguishable, affecting the clarity of clusters.

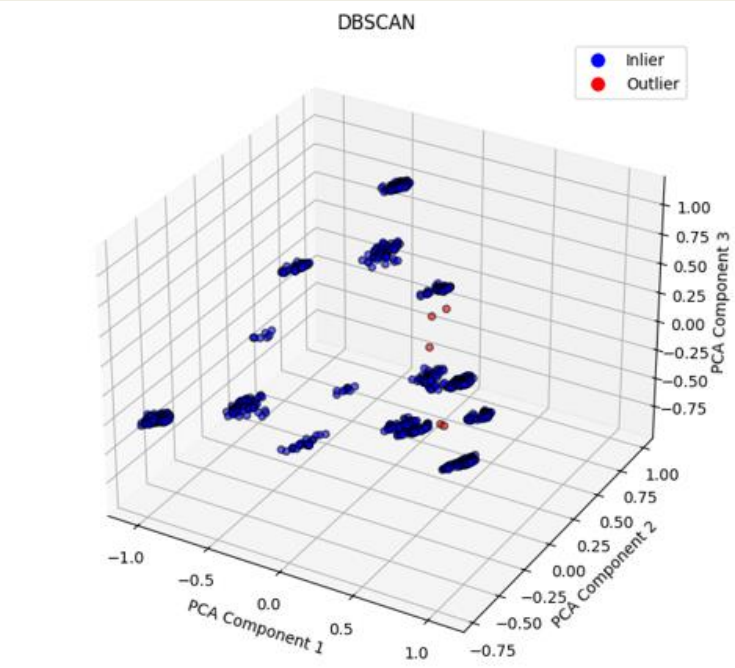
Results

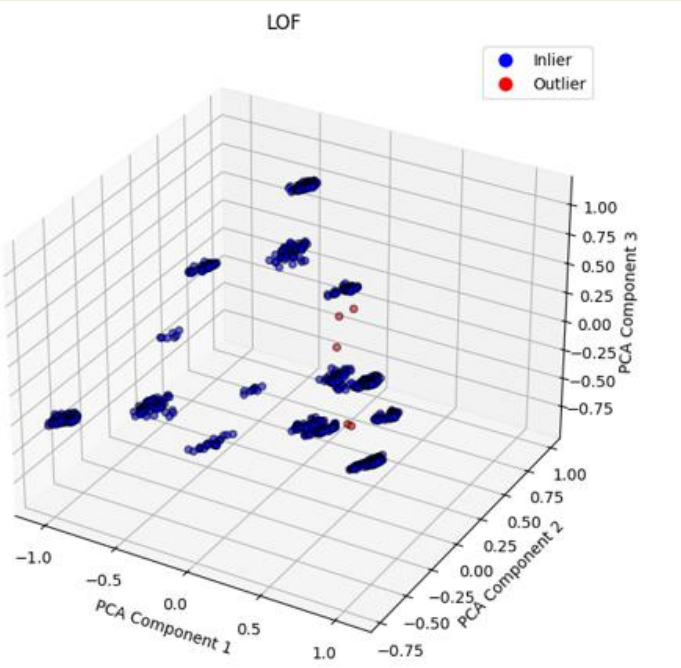
- By applying the DBSCAN algorithm, we have detected some potentially outliers.



Results

- Among the five algorithms deployed, only DBSCAN and Local Outlier Factor (LOF) identified certain samples as outliers.
- When employing a majority voting strategy to consolidate findings across methods, these samples were not classified as anomalies





Conclusion:

- This project applied machine learning to uncover patterns in YouTube video data, revealing the diverse landscape of digital content.
- Through careful preprocessing, PCA, and clustering with K-Means and DBSCAN, we identified distinct video clusters.
- Our approach highlights the importance of dimensionality reduction and strategic data analysis in extracting meaningful insights from complex datasets.
- The findings provide a basis for deeper exploration into content strategies, potentially guiding creators and marketers in optimizing digital content for better engagement.