## Abstract

This project embarked on an exploratory journey into the vast realm of YouTube videos, with the goal of uncovering patterns and anomalies within video metadata. Motivated by the growing importance of digital content and its impact on viewers' preferences and content creators' strategies, we utilized data mining techniques to cluster videos based on their attributes and detect outliers. By utilizing the YouTube API, we gathered a comprehensive dataset, which underwent preprocessing to ensure quality and relevance. We employed a combination of principal component analysis (PCA) for dimensionality reduction, K-means and DBSCAN to group similar videos, while anomaly detection was performed using multiple algorithms like DBSCAN, Local Outlier Factor, Elliptic Envelope, Isolation Forest and One Class SVM. Our methodology revealed distinct clusters that characterize the multifaceted landscape of YouTube content but did not detect any outliers in the data.

## Introduction

In the rapidly expanding universe of YouTube, where content is as diverse as its global audience, the challenge of making sense of vast amounts of video data presents a unique opportunity for exploration. This project is rooted in the application of clustering algorithms to categorize YouTube videos, aiming to distill the myriad of content into understandable and insightful groups based on shared characteristics. The motivation for this endeavor is driven by the need to understand what makes certain videos successful compared to others, within each category.

Our study focuses on leveraging advanced data mining techniques and machine learning algorithms to sift through YouTube video metadata. By doing so, we seek to cluster similar videos, thus providing a clearer picture of the content landscape on YouTube. This clustering not only helps in identifying common themes and preferences among viewers but also assists in predicting future trends in video content. The significance of our work lies in its potential to inform content strategy, enabling creators and marketers to tailor their offerings to match viewer demand more closely.

The objectives of this project are clear: to apply clustering algorithms effectively to YouTube data and to interpret the resulting clusters in a way that provides actionable insights for content creators, marketers, and platform developers. Through this analysis, we aim to contribute to the understanding of digital content consumption patterns, offering a data-driven glimpse into the types of video content that captivate and engage viewers.

This project's impact is envisaged to extend beyond academic interest, offering tangible benefits to those involved in content creation and digital strategy on YouTube. By identifying the characteristics of popular and engaging video content, our work can help stakeholders optimize their content strategies, leading to improved viewer satisfaction and engagement. Furthermore, this study enriches the field of data mining by tackling a real-world problem with significant implications for the digital content industry.

## Dataset and Features

The dataset for this project was sourced from the YouTube Data API, providing a comprehensive overview of video and channel metrics. Initially, the raw dataset included a wide array of features such as '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', and 'share_count'.

Upon thorough examination, it became apparent that certain columns, like 'default_language', 'default_audio_language', and others, contained static values or were not relevant to our analysis objectives. Moreover, 'favorite_count' and 'dislike_count' fields were found to be consistently missing due to YouTube's policy changes, rendering them unusable. Consequently, these columns were excluded from further analysis.

In the preprocessing phase, we engineered new features and transformed existing ones to enhance our dataset's usability. For instance, video 'duration' was converted into seconds, and significant insights were derived from textual columns like 'title' and 'tags' through feature engineering.

To effectively handle the refined feature set, we developed a preprocessing pipeline catering to both numerical and categorical data. For numerical features, we imputed missing values with zeros, grounded in the understanding that our data collection script returned None for any absent information in the API response. Following the imputation, we applied MinMaxScaler for normalization, a critical step for the successful application of PCA due to its sensitivity to variable scales.

For categorical features, missing values were replaced with a placeholder 'missing', and OneHotEncoder was employed to convert categorical variables into a format suitable for machine learning models. This comprehensive preprocessing approach ensured that the data was optimally prepared for clustering and analysis, allowing us to uncover meaningful patterns and insights from the YouTube dataset.

## Methodology

To address the high dimensionality of our dataset and improve the interpretability of the clustering results, Principal Component Analysis (PCA) was employed as a dimensionality reduction technique. PCA enabled us to condense the information contained in multiple variables into a smaller set of principal components that captured the most variance in the data. This not only facilitated a more efficient clustering process but also allowed for a clearer visualization of the clustered data in a reduced dimensional space.

The primary clustering algorithm employed in this project was K-Means clustering, chosen for its simplicity and effectiveness in identifying k distinct clusters within a dataset. Visual examination of the data revealed that the clusters formed spherical shapes and were distinctly separated from each other, demonstrating an ideal scenario for the application of the K-Means algorithm.

The selection of the right number of clusters, k, was informed by both the Elbow Method and the Silhouette Score, ensuring that the chosen k value offered a meaningful balance between cohesion and separation. The Silhouette Score was similar to K-Means and DBSCAN results, emphasizing the good clustering.

## Experiments

In the exploratory phase of our project, recognizing the challenge of analyzing high-dimensional data was pivotal. To address this, we employed dimensionality reduction techniques such as t-SNE and PCA. These methods were instrumental in visualizing the data

distribution within a two-dimensional space, offering initial insights into the inherent data structure.

An unexpected outcome from our initial PCA analysis revealed that a single component accounted for 99.99% of the dataset's total variance. This result prompted a reevaluation of our data collection and preprocessing steps. After consulting with Dr. Chen and revisiting our methodology, we enriched our dataset to ensure a more diverse and informative set for analysis.

To determine the most effective clustering strategy, we embarked on a two-pronged experimental approach:

1. K-Means Clustering on the Entire Dataset: We applied the K-Means clustering algorithm directly to the preprocessed dataset without dimensionality reduction. The optimal number of clusters was determined based on the silhouette score, a measure of how similar an object is to its own cluster compared to other clusters.

2. K-Means Clustering on PCA-Reduced Data: In this experiment, we first reduced the dataset's dimensionality using PCA to capture essential features while minimizing information loss. We then applied K-Means clustering to this reduced dataset. Similar to the first experiment, the optimal number of clusters was identified through the silhouette score analysis.

These experiments were designed to compare the effectiveness of clustering on the original versus dimension-reduced data, aiming to uncover the most meaningful and computationally efficient approach for segmenting YouTube video data.

## Results and Discussion

In our investigation, the utilization of three principal components enabled us to encapsulate approximately 54% of the variance inherent in the original dataset. This reduction highlighted the multifaceted nature of the data without oversimplifying its complexity.

By Applying the K-Means algorithm directly to the entire dataset (Fig. 1), the best silhouette score was relatively low - 0.486, and the clusters didn't make sense.

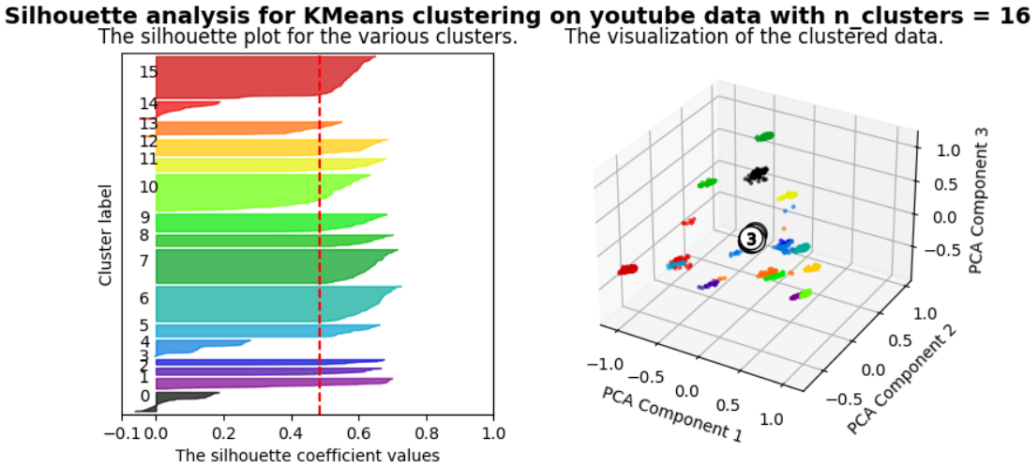


Fig. 1. K-Means algorithm applied to entire dataset.

Upon further exploration of the clustering challenges encountered, we postulated that the diminishing effectiveness of our initial clustering approach could be attributed to the curse of dimensionality. This phenomenon, inherent in high-dimensional data spaces, tends to obscure the distinctions between data points, as distances between them converge towards a uniformity that muddles meaningful groupings.

Guided by this understanding, we shifted our focus towards leveraging the dimensionality-reduced dataset for clustering. This decision was underpinned by a hypothesis that, despite encapsulating only 54% of the original data's variance through principal component analysis, the reduced dataset would still preserve the essential disparities between distinct video categories. Remarkably, this strategic pivot bore fruit, as evidenced by a significant uplift in the silhouette score to 0.837, as seen in Fig. 2. This improvement not only underscored the enhanced separation between clusters but also affirmed the effectiveness of dimensionality reduction as a means to circumvent the curse of dimensionality, thereby yielding more discernible and interpretable groupings within the YouTube video dataset.

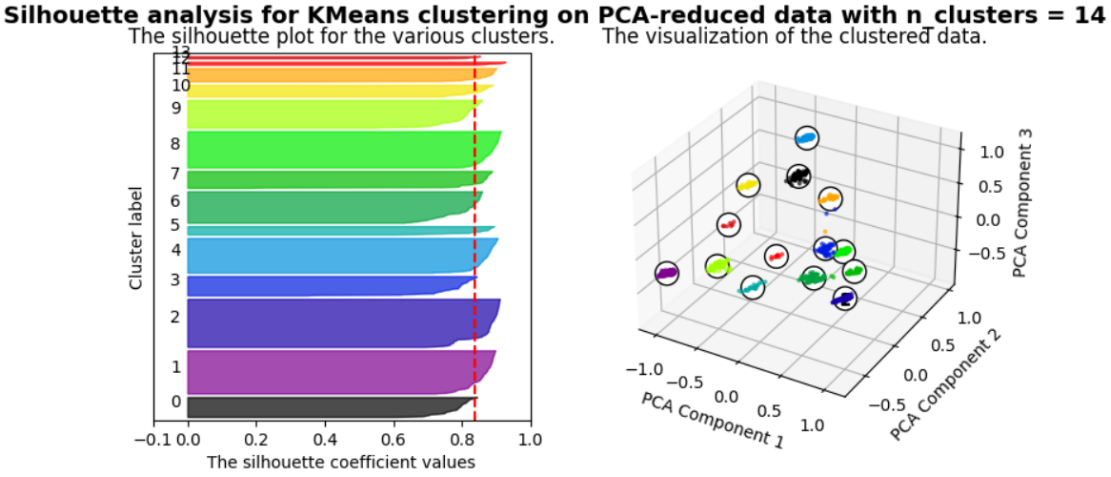


Fig. 2. K-Means algorithm applied to the dimension-reduced matrix.

Upon examining the clustering outcomes, we identified several instances that did not neatly align with the identified clusters. This observation highlighted a notable limitation of the K-Means algorithm: its tendency to force every data point into a cluster, even those that are distinctly outliers.

To address this shortcoming, we turned our attention to DBSCAN, an algorithm renowned for its adeptness at identifying outliers while forming clusters based on density. To find the best hyperparameters for DBSCAN, we defined a grid search function. This approach involved iterating through a range of values for epsilon (the radius within which to search for neighboring points) and min_samples (the minimum number of points required to form a dense region). For each parameter combination, we evaluated the clustering efficacy using the silhouette score — a measure of how similar an object is to its own cluster compared to other clusters.

The objective was to pinpoint the epsilon and min_samples values that maximized the silhouette score, thereby ensuring an optimal balance between the cohesiveness of the clusters and the correct identification of outliers. This parameters tuning was instrumental in enhancing our model's ability to discern between the core structures of our data and those data points that stood apart as anomalies.

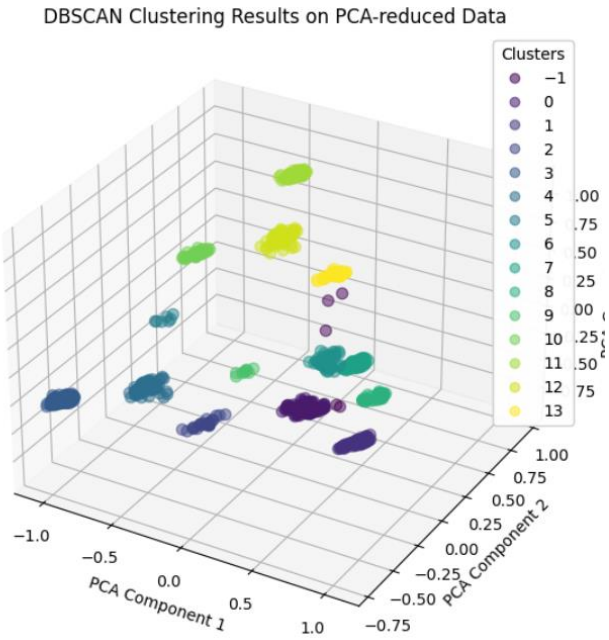


Fig. 3. DBSCAN algorithm applied to dimension-reduced matrix.

The outcomes, represented in Fig. 3, aligned closely with our expectations. By deploying DBSCAN with the chosen parameters, we discerned 14 distinct clusters that mirrored the structure revealed by K-Means, achieving a comparable silhouette score of 0.836. However, DBSCAN offered an additional advantage: it successfully identified several outliers.

Subsequently, we delved into anomaly detection, initially employing traditional methods such as box plots and modified Z-scores. However, these techniques proved insufficient for our needs, prompting us to explore more sophisticated machine learning algorithms. We tested five different models specifically designed for anomaly detection：DBSCAN, Local Outlier Factor (LOF), Elliptic Envelope, Isolation Forest, and One-Class SVM. This approach allowed us to comprehensively assess the dataset from multiple perspectives, leveraging the unique strengths of each algorithm to identify outliers effectively.

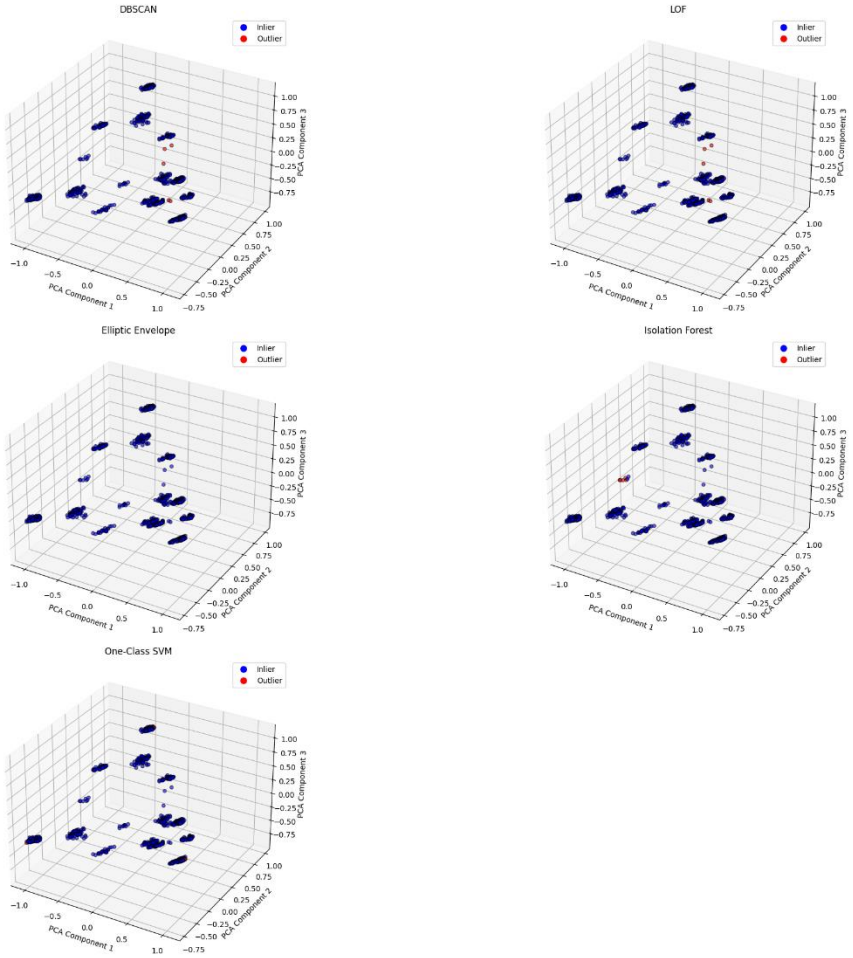


Fig. 4. Multiple anomaly detection algorithms.

Among the five algorithms deployed, only DBSCAN and Local Outlier Factor (LOF) identified certain samples as outliers. However, when employing a majority voting strategy to consolidate findings across methods, these samples were not classified as anomalies. This highlights the nuanced nature of anomaly detection, where consensus across diverse methodologies can refine our understanding of what constitutes an outlier in complex datasets.

## Conclusion and Future Work

The exploration of YouTube video data through clustering has unveiled intriguing patterns and relationships within the digital content landscape. This project has successfully demonstrated the utility of machine learning algorithms, specifically K-Means and DBSCAN, in grouping videos into clusters based on their attributes. The application of PCA for dimensionality reduction before clustering not only optimized the computational efficiency but also provided clear visual insights into the data structure and higher score.

Our methodology, centered around data preprocessing, PCA, and K-Means clustering, underscored the significance of a well-thought-out analytical approach to handle high-dimensional data. The resulting clusters revealed the diverse nature of video content on YouTube, offering a foundation for understanding viewer preferences and content strategy.

Future work could extend this analysis in several directions. Further analysis of the clusters could unveil underlying trends and patterns of popularity, offering valuable insights for content creators and YouTube alike. By understanding these dynamics, creators can tailor their content strategies to better align with viewer preferences, while YouTube can enhance its recommendation algorithms to match viewer interests more accurately. Additionally, integrating natural language processing to analyze video titles and descriptions could offer deeper understanding of content impact and viewer engagement.

## Contributions

Aviram – performed data preprocessing, Visualizations and DBSCAN

Ido – YouTube API, PCA, K-Means and streamlit app

For anomaly detection, final write-up and presentation both were included.

## GitHub

https://github.com/idokapel/youtube-data-mining.git