# YouTube Channel Analytics: A CRISP-DM Analysis to Enhance Creator Engagement and Monetization

Vijitha Gunta
MS Software Eng, San Jose State University

vijitha.gunta@sjsu.edu

## Abstract

In the realm of digital video content, YouTube stands as a behemoth. This paper delves deep into the intricate patterns and dynamics of YouTube channels, leveraging the global statistics from 2023. By employing the CRISP-DM methodology, we unearth insights and correlations that can transform audience engagement, monetization, and content strategies. Advanced clustering techniques further enable us to segment YouTube channels, culminating in strategic recommendations that promise to redefine content creation paradigms.

## 1 Introduction

### 1.1 Background

With billions of views every day, YouTube has revolutionized the way we consume video content. However, behind these massive numbers lies a complex ecosystem of content creators, each vying for the attention of viewers. Understanding the nuances of this ecosystem is paramount for creators to stand out and for businesses to effectively monetize their presence.

### 1.2 Objective of the Study

This research seeks to provide a data-driven lens to understand the dynamics of YouTube channels. Through meticulous analysis, we aim to uncover patterns in audience engagement, the competitive landscape, and effective content strategies.

## 2 Methodology

### 2.1 Data Collection and Preprocessing

#### 2.1.1 Data Source

The dataset utilized in this study encompasses global YouTube statistics for the year 2023. It provides a holistic view of channels, capturing metrics such as subscriber count, video views, category affiliations, and earnings.

#### 2.1.2 Data Cleaning

The raw dataset, though comprehensive, had its share of inconsistencies and missing values. We adopted a meticulous approach to data cleaning:

- Missing values were imputed using appropriate statistical measures.
- Outliers, identified using the IQR method, were analyzed and treated to ensure they didn't skew the analysis.

#### 2.1.3 Feature Scaling

To prepare the data for machine learning algorithms, especially clustering, features were scaled to have a mean of zero and a standard deviation of one. This step ensures that all features contribute equally to the algorithm's process.

### 2.2 Exploratory Data Analysis (EDA)

#### 2.2.1 Category Distribution

A deep dive into the category distribution provided insights into the most competitive and popular categories on YouTube. Such insights are crucial for new entrants to gauge the competition and for existing channels to pivot their content strategy if needed.

In: M. Meder, A. Rapp, T. Plumbaum, and F. Hopfgartner (eds.): Proceedings of the Data-Driven Gamification Design Workshop, Tampere, Finland, 20-September-2017, published at http://ceur-ws.org

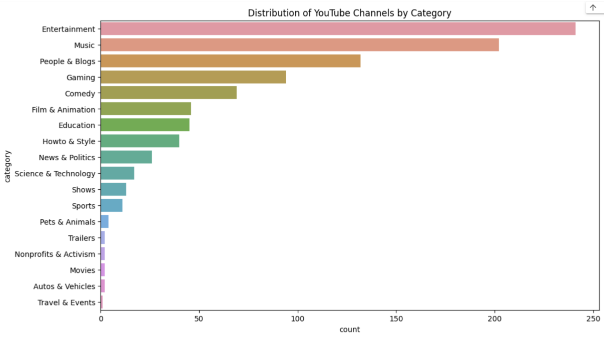


Figure 1: Distribution of YouTube Channels by Category

#### 2.2.2 Geographical Analysis

By analyzing the geographical distribution of channels, we could identify potential high-growth markets and regions where YouTube has a significant presence.

#### 2.2.3 Correlation Analysis

Understanding how different metrics correlate with each other provides a lens into potential strategies. For instance, does having more subscribers guarantee more views? Or is the content type more critical?

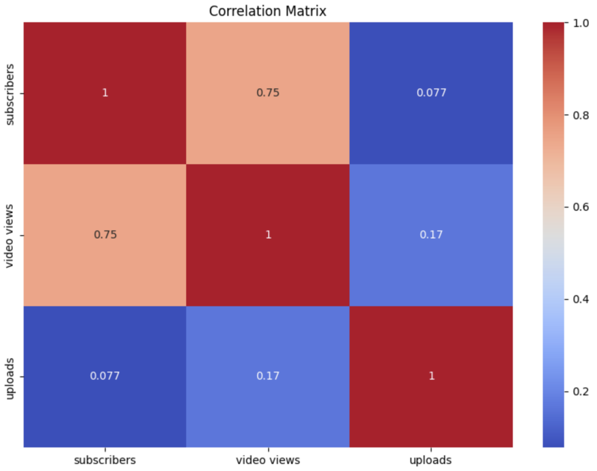


Figure 2: Correlation Matrix

### 2.3 Clustering Techniques

#### 2.3.1 K-means Clustering

K-means, a centroid-based clustering algorithm, was employed to segment YouTube channels. The algorithm iteratively assigns data points to clusters based on their proximity to cluster centroids.

#### 2.3.2 Agglomerative Hierarchical Clustering

This technique provides a hierarchical view of data clusters. Starting with each data point as an individual cluster, the algorithm successively merges clusters based on distance metrics.

#### 2.3.3 DBSCAN

DBSCAN, a density-based algorithm, identifies clusters as regions with a high density of data points. It's particularly adept at distinguishing outliers, providing a more nuanced view of the data's structure.

## 3 Results

### 3.1 Exploratory Data Analysis (EDA) Findings

#### 3.1.1 Category Insights

Our analysis revealed a high concentration of YouTube channels in the Entertainment, Music, and Gaming categories. This suggests a competitive landscape in these domains but also underscores their popularity among viewers.

#### 3.1.2 Geographical Distribution

The geographical breakdown of channels placed the United States, India, and the United Kingdom at the forefront in terms of channel count. These regions represent significant market potential and active user bases.

#### 3.1.3 Correlation Dynamics

A pivotal finding was the strong positive correlation between the number of subscribers and video views. This relationship emphasizes that channels with a higher subscriber count generally enjoy greater visibility and reach.

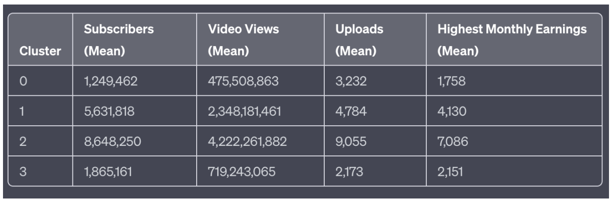
### 3.2 Cluster Analysis Outcomes

#### 3.2.1 K-means Clustering

Through the K-means clustering algorithm, we discerned four distinct channel segments, each exhibiting unique characteristics and viewer dynamics. These clusters ranged from nascent channels trying to establish their foothold, to superstar channels with millions of subscribers and views.

#### 3.2.2 Agglomerative Hierarchical Clustering

The Agglomerative Hierarchical Clustering technique furnished a nuanced, tree-like structure of channel relationships. While offering a detailed view, its complexity made direct business interpretations more challenging.



| Cluster | Subscribers (Mean) | Video Views (Mean) | Uploads (Mean) | Highest Monthly Earnings (Mean) |
|---|---|---|---|---|
| 0 | 1,249,462 | 475,508,863 | 3,232 | 1,758 |
| 1 | 5,631,818 | 2,348,181,461 | 4,784 | 4,130 |
| 2 | 8,648,250 | 4,222,261,882 | 9,055 | 7,086 |
| 3 | 1,865,161 | 719,243,065 | 2,173 | 2,151 |

Figure 3: Modeling - Channel Segmentation Using K-means Clustering

#### 3.2.3 DBSCAN Insights

The DBSCAN algorithm, recognized for its proficiency in detecting outliers, marked certain channels as anomalies based on their metrics. This outlier detection capability provides a lens to identify channels that deviate from typical patterns, warranting a closer examination.

## 4 Discussion

### 4.1 Interpreting Business Insights

#### 4.1.1 Audience Engagement Metrics

The strong correlation between subscribers and video views underscores the importance of community building. Channels with robust subscriber bases not only have a consistent viewer pool but also enjoy better organic reach, as their content is more likely to be recommended to non-subscribers.

#### 4.1.2 Competitive Landscape Insights

The data elucidated the fiercely competitive nature of the Entertainment and Music categories. Channels in these segments must continually innovate and deliver high-quality content to maintain and grow their viewer base.

#### 4.1.3 Optimal Content Strategies

An intriguing revelation was the lack of a strong correlation between upload frequency and channel popularity, be it in terms of subscribers or views. This finding advocates for a focus on content quality and relevance over sheer quantity.

### 4.2 Strategic Recommendations Based on Clustering

#### 4.2.1 Emerging Creators

Channels in the early stages of growth should prioritize content consistency, quality, and viewer engagement. Collaborations with similar-sized channels and community-building can amplify their growth.

#### 4.2.2 Mid-Tier Creators

Established channels with a sizeable viewer base should diversify their content and revenue streams. Engaging with their community through interactive content, merchandise, or sponsored content can further enhance their growth.

#### 4.2.3 Superstar Channels

For the creme de la creme of YouTube, maintaining content quality is paramount. While they enjoy massive reach, staying connected with their core audience and innovating is crucial to fend off competition.

#### 4.2.4 Outliers and Anomalies

Channels identified as outliers, especially by the DBSCAN algorithm, warrant a closer examination. These channels defy typical patterns and can offer insights into unconventional growth strategies or potential data discrepancies.

## 5 Conclusion

Our comprehensive data-driven analysis sheds light on the intricate dynamics of YouTube channels. By leveraging advanced clustering techniques, we segmented channels into distinct categories, each with its unique challenges and growth strategies. The findings and recommendations of this study promise to serve as a beacon for content creators, guiding them in their journey to achieve unparalleled growth and engagement on YouTube.

## References

[1] Rüdiger Wirth, Jochen Hipp *CRISP-DM: Towards a Standard Process Model for Data Mining*