YouTube Viewing History Analysis

Bora Sacır

January 9, 2025

Abstract

This report presents an extensive analysis of YouTube viewing history, focusing on temporal trends, channel preferences, keyword extraction, and content patterns. Various visualizations, including daily viewing trends, day-of-week trends, hourly trends, and keyword-based analysis, are used to derive insights into viewing behavior. The findings provide a comprehensive understanding of YouTube consumption habits over time.

Contents

1 Introduction

YouTube has become one of the most influential platforms for content consumption. Analyzing viewing history can provide insights into individual preferences, temporal trends, and content consumption patterns. This project aims to perform an exploratory data analysis of YouTube viewing history using advanced data visualization techniques and feature engineering.

The objectives of this project include:

- Understanding daily viewing patterns over time.
- Identifying preferences based on the day of the week and time of day.
- Extracting and visualizing keywords from video titles.
- Highlighting the most-watched channels and videos.

2 Data Description

The dataset consists of YouTube watch history exported in HTML format. Key columns extracted include:

- **Title**: The title of the watched video.
- **Channel**: The channel name.
- **Date**: The date when the video was watched.
- **Time**: The time when the video was watched.
- **Day of Week**: The weekday corresponding to the date.

The dataset was preprocessed to handle missing values, convert timestamps to readable formats, and engineer new features like rolling averages and lag values.

3 Methodology

3.1 Feature Engineering

- **Daily Counts**: Aggregated the total number of videos watched per day.
- **Lag Features**: Added features like video counts from previous days to capture temporal dependencies.
- **Rolling Averages**: Computed rolling averages over a 7-day window to identify trends.
- **Keyword Extraction**: Used TF-IDF scores to identify significant keywords in video titles.

3.2 Visualization Techniques

The following visualizations were used to explore the data:

- Line charts for daily viewing trends.
- Bar charts for day-of-week and hourly viewing trends.
- Heatmaps to analyze hourly viewing trends by the day of the week.
- Word clouds and bar charts for keyword and channel analysis.

4 Results and Analysis

4.1 Daily Viewing Trends

Figure 1 shows the number of videos watched daily over the observed time period. The data indicates significant spikes in viewing activity on certain days, potentially corresponding to holidays or specific events.

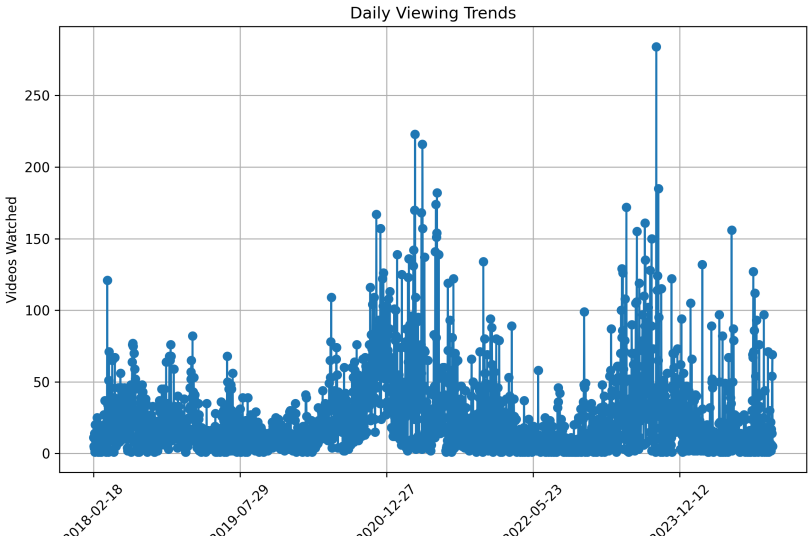

Figure 1: Daily Viewing Trends

4.2 Viewing Trends by Day of the Week

As shown in Figure 2, viewing activity peaks on certain days of the week, with Friday and Sunday showing the highest activity.

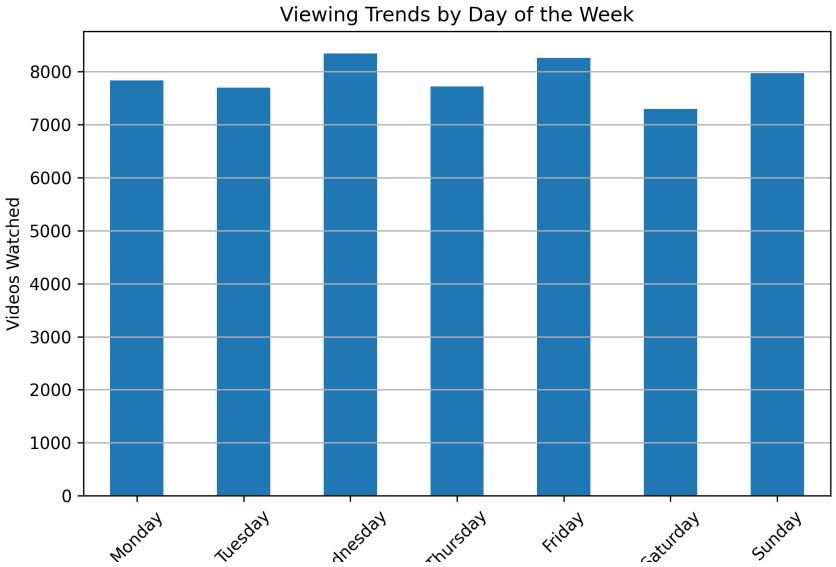

Figure 2: Viewing Trends by Day of the Week

4.3 Hourly Viewing Trends

Figure 3 illustrates the distribution of viewing activity across different hours of the day. Peak activity is observed during late evenings.

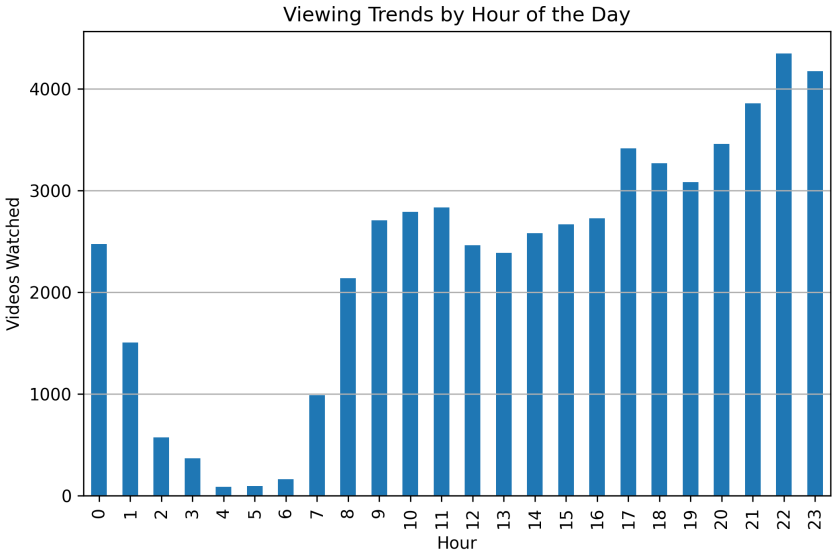

Figure 3: Viewing Trends by Hour of the Day

4.4 Hourly Viewing Trends by Day of the Week

The heatmap in Figure 4 provides a comprehensive view of viewing activity across different days and hours. It highlights specific periods of peak activity.

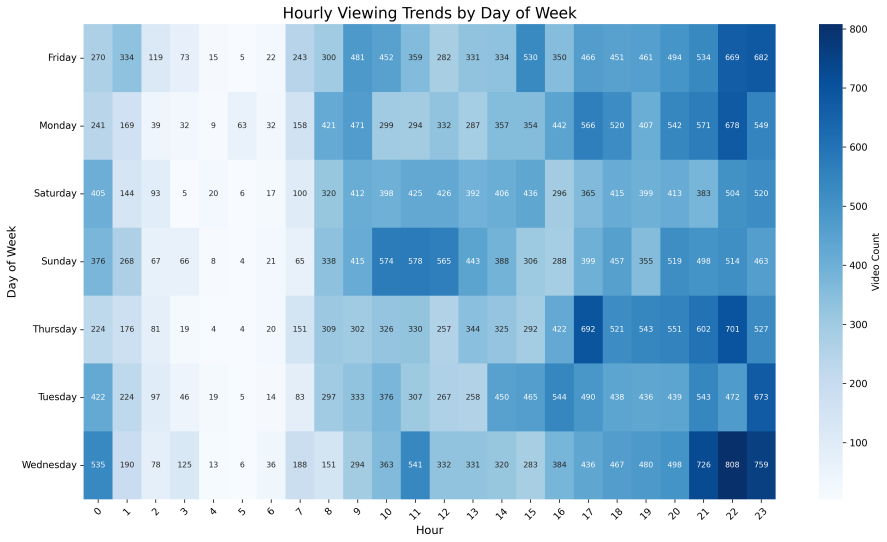

Figure 4: Hourly Viewing Trends by Day of the Week

4.5 Top 10 Most-Watched Videos and Channels

Figures 5 and 6 showcase the top 10 most-watched videos and channels. This highlights the user's specific content preferences.

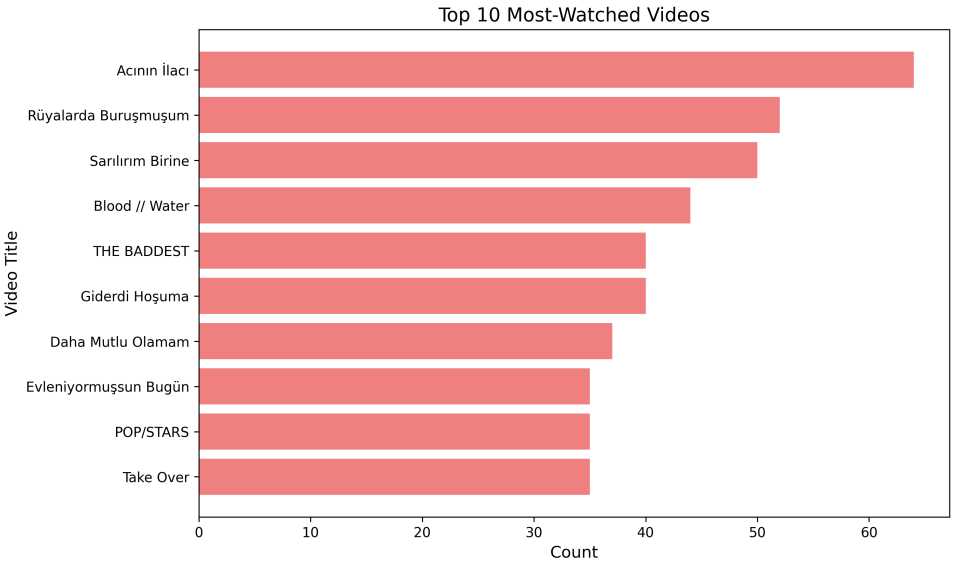

Figure 5: Top 10 Most-Watched Videos

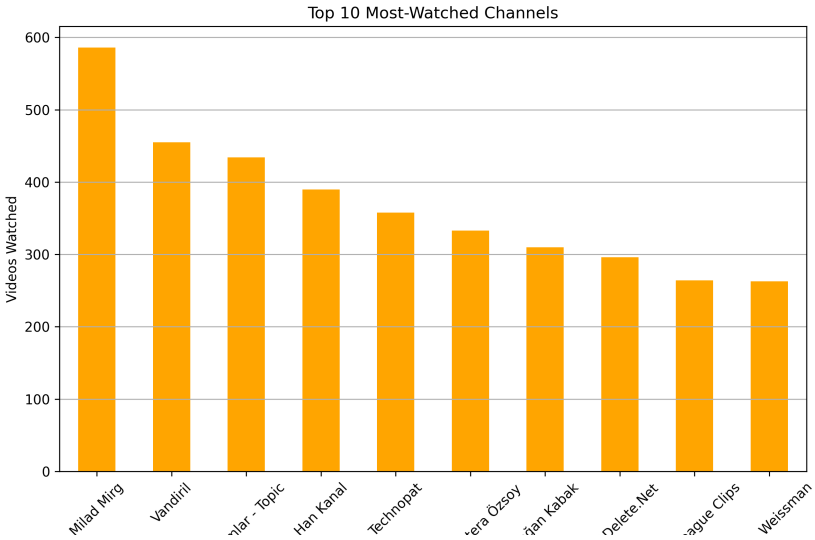

Figure 6: Top 10 Most-Watched Channels

4.6 Keyword Analysis

Significant keywords extracted using TF-IDF are shown in Figure 7, and the word cloud in Figure 8 visually represents the most frequent keywords in video titles.

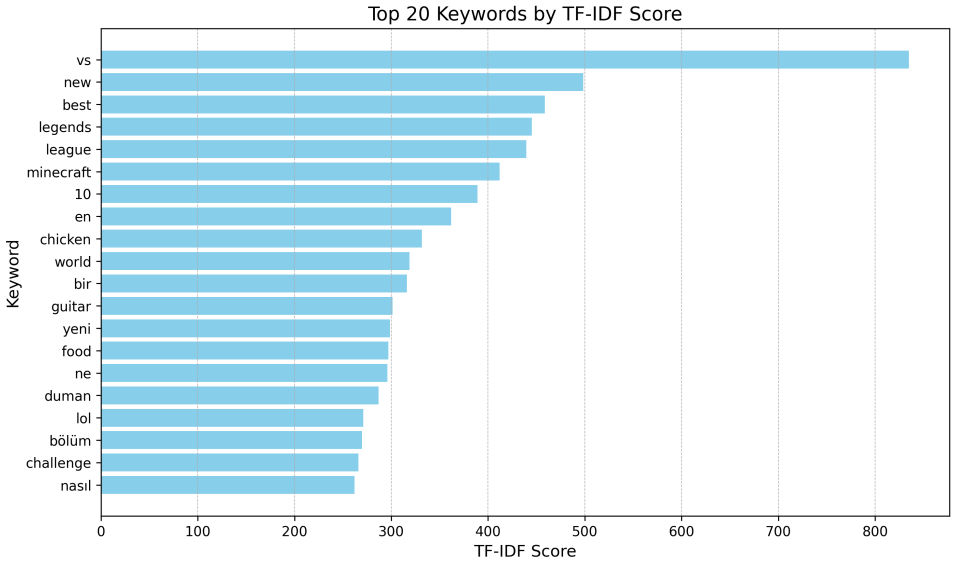

Figure 7: Top 20 Keywords by TF-IDF Score

Figure 8: Word Cloud of Keywords in Video Titles

5 Conclusion

This project provides a detailed exploration of YouTube viewing habits through data visualization and analysis. Key findings include:

- Significant temporal patterns in daily and hourly viewing trends.
- Specific days of the week and hours of the day exhibit peak activity.
- User preferences are evident in the most-watched videos, channels, and keywords.

Future work can explore predictive modeling to forecast video counts based on these patterns and trends.

6 References

- McKinney, W. (2010). *Data Structures for Statistical Computing in Python.* Proceedings of the 9th Python in Science Conference. `https://pandas.pydata.org/`
- Hunter, J. D. (2007). *Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering*, 9(3), 90-95. `https://matplotlib.org/`
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., et al. (2011). *Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research*, 12, 2825–2830. `https://scikit-learn.org/`