UAT YouTube Metrics Analysis:

Identifying Relevant Patterns in YouTube Performance Metrics

Avery Willets

University of Advancing Technology

March 27, 2025

Project Overview

- **Objective**: The goal of this project is to analyze YouTube video performance data to better understand patterns within video engagement based on video metadata.
- **Key Question:** Based on the features present in the data, can the performance of a video (in views) be predicted? Which features influence that outcome the most?
- **Dataset**: The dataset consists of publicly available data collected using YouTube's API, including: video titles, view counts, like counts, comment counts, tags used, and video duration. It was collected from the University of Advancing Technology's YouTube channel.
- **Tools & Technologies**: Python, Pandas, Seaborn, Matplotlib, Scikit-learn.
- **Project Docs:**
    - https://colab.research.google.com/drive/1eTimZAOUgv--AWLKIDZ2naVof7QHH3jx
    - https://colab.research.google.com/drive/1_9cJkO4DkoQtLnVZXWRa1PrxZtUHig9j
    - https://colab.research.google.com/drive/1xlfnptmX2S6EErh4RNIvqO37i6kbJ0FP

Introduction

- **Background**:
    - With the amount of content online, it is imperative for creators to understand their audiences in order to both find and grow their communities. The purpose of this project is to become familiar with collecting, preparing, and using real-time data to answer business questions surrounding performance metrics.
- **Problem Statement**:
    - Understanding which features are most strongly correlated with video performance can help creators optimize their content based on their needs.

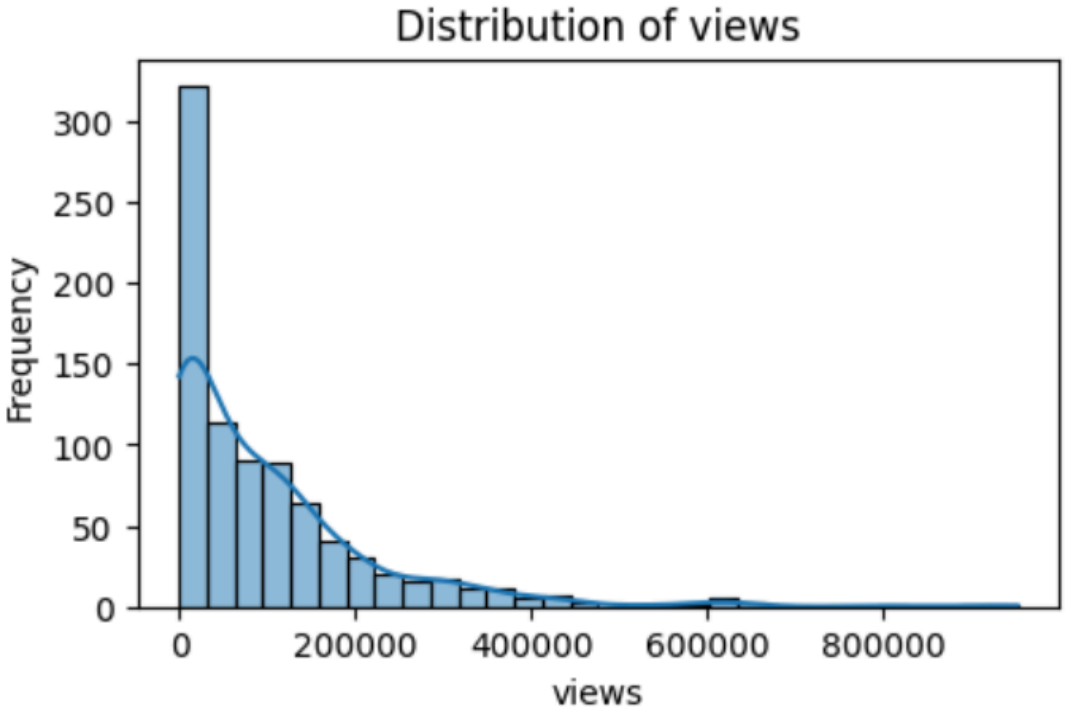
Methods Used

- **Data Collection**:
    - Data was retrieved using the YouTube Data API. This provides metadata about videos, including tags used, video length, view counts, and more. This was interacted with via the Python programming language in the Google Colab environment.
- **Data Preprocessing**:

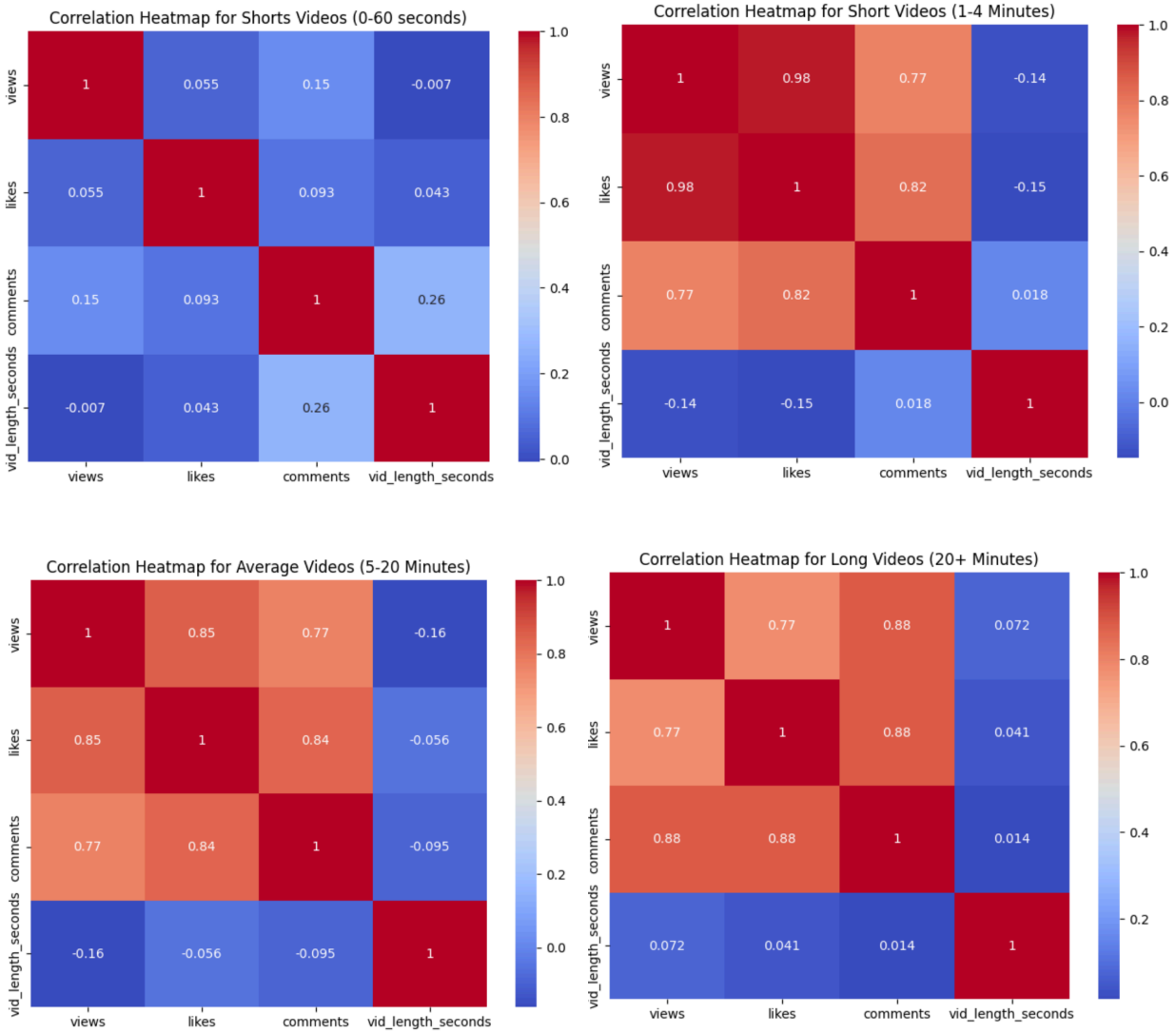
    - Null values in the data were handled by imputing missing data with the mean for numeric features and ‘Unknown’ for categorical features. Outliers were detected using a z-score method and removed. Data with unique formatting was transformed using regular expressions.
- **Techniques/Models**:
    - I primarily used regression analysis to predict video views based on features in the data. I also used grouped data based on similar features to find trends in the data.

Analysis Performed

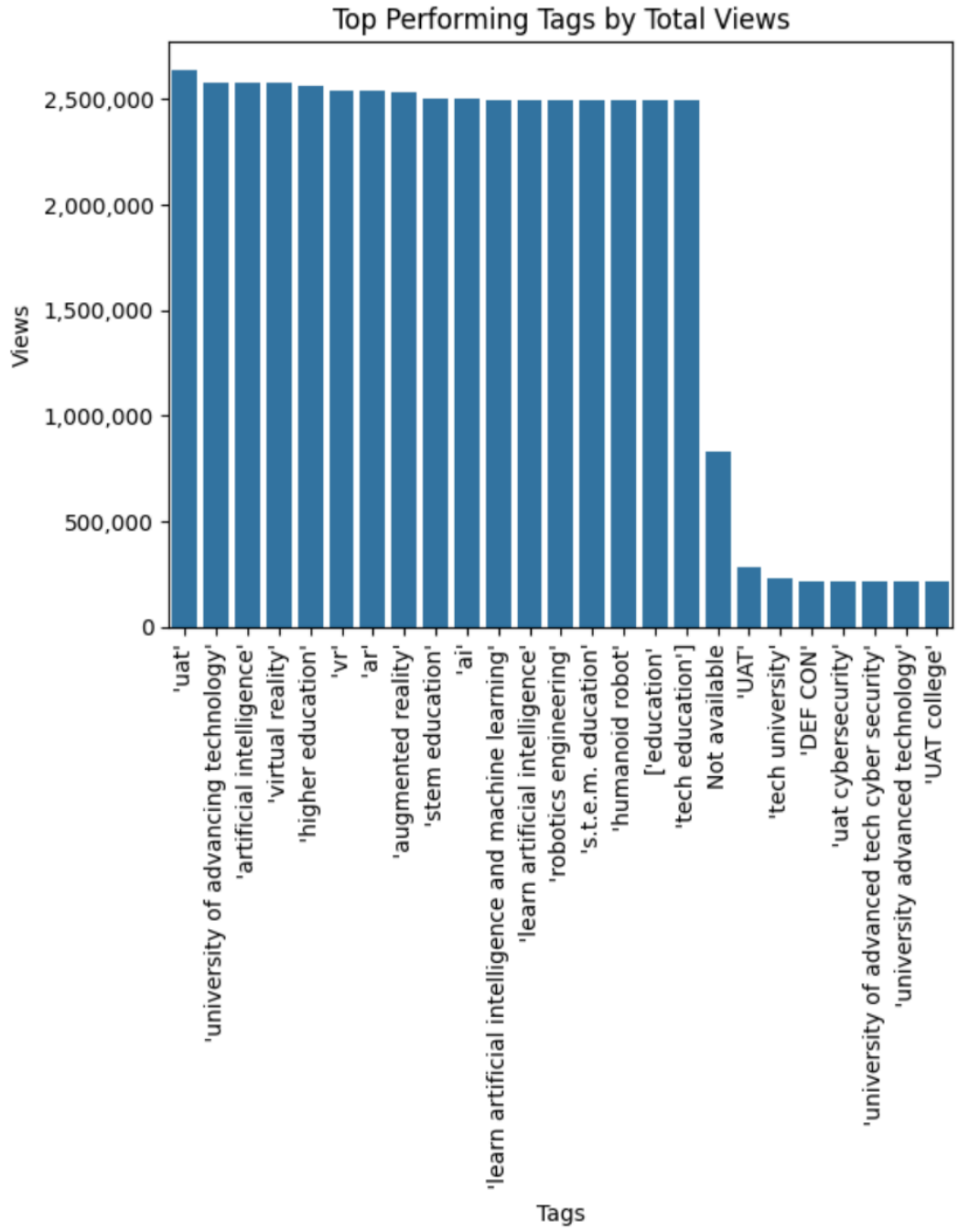
- **Exploratory Data Analysis (EDA)**:
    - I performed EDA to examine the distribution of video views (as well as likes, comments, and durations) and explore correlations between video features and performance.
- **Key Findings**:
    - There was not a significant difference found between videos of different lengths and total views.
    - It was found that videos longer than 60 seconds tended to have more interactions (likes and comments) per view, while YouTube Shorts tended to have fewer interactions per view.
    - A significant correlation was not found between tags used and video views, but some tags had much higher total views across all videos than others.
- **Visualizations**:



- A histogram of view counts indicates that while most videos have a performance of under 100,00 views, other videos have as many as 600,000+ views.



- Heatmaps of video lengths and performance metrics indicate that views correlate strongly with both likes and comments for short, medium, and long video lengths. Thus, viewers are more likely to like or comment after watching videos that are longer than 60 seconds.



- A barplot of tags vs. total views shows that tags such as ‘uat’, ‘university of advancing technology’, ‘artificial intelligence’, ‘virtual reality’, and ‘higher education’ are popular topics among viewers.

	views	video_id	title	release_date	likes	comments	tags	vid_length	z_score
0	2496244	OtRrVrHx7Do	AR+VR+AI+Robotics Degrees | University of Adva...	2018-10-29 23:26:09+00:00	1817	40	[['education', 'virtual reality', 'higher ed...	00:01:31	15.223435

- The z-score calculator ousted one video as being a rather extreme outlier in the views department. Upon inquiry, I was informed this video is at the front of many landing pages and campaigns, which makes sense. I unfortunately had to remove it from the dataset due to being such an outlier.

Insights Gained

- **Key Insights**:
  - YouTube Shorts tend to get less engagement (likes and comments), than other kinds of videos.
  - Short videos (1-4 mins) got the most views on average compared to other video lengths.
  - Tags such as ‘education’, ‘learn artificial intelligence’, and ‘humanoid robot’ were identified as being influential on total views, but the correlation between the features was weak.
- **Implications**:
  - Content creators at UAT could optimize their video length and tag usage in order to connect with a wider audience and increase engagement.

Suggested Next Steps

- **Further Research**:
  - To refine these insights, we could begin to analyze the importance of video titles, including length and keywords included.
  - We could also analyze video performance based on time of day, week, month, and year to find a more optimal publishing schedule.
  - We might also use a KNN-Clustering machine learning algorithm to group videos based on similar patterns in the data, rather than user-determined metrics. This could be used to find potential topics or audience clusters, particularly ones with high engagement.
  - If UAT would like to examine their own private data further, identifying target audience groups and understanding how best to appeal to them according to past engagement may prove beneficial.

    - Further research into trending video lengths, tags, and release schedules may also help UAT further understand how to optimize video performance.
- **Model Improvements**:
    - Most of the linear regression models implemented struggled to predict results accurately. This could be rectified by feeding them more data from similar creators, or perhaps by using a gradient boosting model instead to handle errors and make more accurate predictions.
- **Additional Features**: Identify additional features that could be explored.
    - Additional features such as video description, topic details, relevant topic IDs may provide more complex insights into engagement.
    - Comments on videos may prove useful in sentiment analysis in order to find how audiences feel about videos.
    - Video thumbnails may also be used in some machine learning models to identify patterns in engagement based on features present in the thumbnail.

Conclusion

- **Summary**:
    - This project analyzed YouTube video performances using Python to identify key factors influencing engagement with UAT's audience. The analysis provides actionable insights for UAT to reach a wider audience and increase engagement.
- **Final Thoughts**:
    - While the analysis demonstrates several insights, there is still much to be learned. Many factors still need to be explored to more fully understand UAT's YouTube video engagement.

References

- YouTube API Documentation: https://developers.google.com/youtube/v3/docs/videos
- Pandas Documentation: https://pandas.pydata.org/docs/reference/frame.html
- Scikit-learn Documentation: https://scikit-learn.org/stable/supervised_learning.html
- Regular Expressions Documentation: https://docs.python.org/3/library/re.html
- Seaborn Documentation: https://seaborn.pydata.org/
- University of Advancing Technology (UAT) YouTube Channel: https://www.youtube.com/user/UATProductions