UNIVERSITY OF MARYLAND 1856

ROBERT H. SMITH
SCHOOL OF BUSINESS

BUDT704 - 0507: DATA PROCESSING AND ANALYSIS IN PYTHON

“Exploratory Data Analysis of YouTube for Recommending Growth Strategy to Marketers and Content Creators”

Report By

Vaishnavi Mahukar
Jagadeesh Harinarayanan
Saurabh Badkas
Bhoomi Chavan
Joo Park
Justin Depinto

1 INTRODUCTION

In the ever-evolving landscape of digital content consumption, YouTube stands as a colossal platform, shaping global trends and influencing how individuals engage with diverse forms of multimedia. As we navigate this expansive digital universe, a compelling endeavor unfolds — exploring YouTube data through the lens of comprehensive data analysis.

Embarking on a journey through the vast sea of YouTube content, our objective is to unravel insights that transcend mere statistical figures. Beyond view counts and subscriber numbers, we delve into the intricacies of user preferences, content dynamics, and the elusive patterns that define success in online video.

This data analysis venture seeks to address fundamental questions and extract actionable knowledge, including:

1. Which category has the highest percentage of viewers and subscribers?
2. Can we identify the patterns during which year what categorical channels were created?
3. How long does it take for a video to go into trendy?
4. Which categories have the most comments and ratings disabled?
5. Can we identify the common words used in the titles and descriptions?
6. At what time were most of the videos uploaded?
7. Is there a way to predict the view count based on the subscriber count?

2 BACKGROUND

In the ever-expanding realm of digital content consumption, YouTube is an unparalleled behemoth, shaping global trends and revolutionizing how individuals engage with a vast array of multimedia. As we embark on a profound exploration of the YouTube data landscape, our mission extends beyond the confines of conventional statistical analysis.

Venturing into the intricate tapestry of user preferences, content dynamics, and the nuanced patterns that underscore triumph in the online video sphere, our data analysis initiative seeks to unveil a deeper understanding of this colossal platform.

The comprehensive set of inquiries guiding our analytical journey reflects a commitment to extracting actionable knowledge and uncovering the intricacies of YouTube's multifaceted ecosystem. Questions such as the categorization of videos with the highest percentage of viewers and subscribers, the temporal dynamics of channel creation, the trajectory of video virality, and the prevalence of comments and disabled ratings in specific categories all contribute to our holistic examination. Furthermore, we aim to unravel linguistic patterns within titles and descriptions, pinpoint peak upload times, and explore the potential correlation between subscriber count and view predictions.

In navigating the expansive digital universe of YouTube content, our analytical voyage is poised to transcend mere statistical figures, providing stakeholders with a roadmap to decode the underlying currents that propel videos to success.

This endeavor not only seeks to answer fundamental questions but aims to empower content creators, marketers, and enthusiasts with insights that can revolutionize content strategy, user engagement, and predictive analytics in the dynamic landscape of online video.

3 DELIVERABLE

Our deliverables include a thorough exploration of YouTube's data landscape, addressing key questions defining digital video success. We aim to unveil content categories with the highest viewership and subscribers. Our study explores temporal dynamics of channel creation, understanding patterns across years. Focusing on the velocity of video virality, we analyze how swiftly content becomes 'trendy.' Uncovering categories with high engagement and disabled comments, we shed light on engagement dynamics. We examine linguistic patterns in titles and descriptions to identify common words contributing to success. Lastly, we investigate the correlation between subscriber count and view predictions, offering insights for navigating YouTube's evolving content landscape.

4 Research Methodology

The research methodology for YouTube data analysis is a rigorous process aimed at uncovering meaningful insights from the dataset and informing strategic decision-making. The comprehensive approach encompasses key steps:

- Addressing discrepancies, removing duplications, handling missing values, and creating a clean and dependable dataset lay the foundation for robust analysis.

- Delving into the heart of the YouTube data to unveil patterns, correlations, and trends. Strategic statistical techniques are employed to decipher underlying narratives, offering a nuanced understanding of viewer preferences, content dynamics, and potential growth areas.

- Utilizing diverse visualizations such as charts, graphs, and word clouds to present complex insights in an accessible manner. Visual representations enhance the interpretation of findings and aid stakeholders in grasping key trends and patterns.

In conclusion, this meticulous methodology ensures that our recommendations are grounded in empirical evidence, empowering stakeholders to make informed decisions for strategic advancements within the YouTube platform. The combination of statistical rigor, visual insights, and targeted exploration of key research questions provides a comprehensive understanding of YouTube data dynamics.

5 Dataset

Derived from Kaggle, our dataset incorporates crucial details including genre, category classification, language and regional availability, runtime, cast specifics, rating metrics, box office performance, and accolades. Prioritizing meticulous data cleaning becomes essential to uphold the integrity of our findings, thereby augmenting the dependability of the insights drawn.

Data source:

- Channel level data
- Trending video data

5.1 Data Cleaning

Column Cleaning:

- Replaced underscores in column names with spaces and title-cased them.
- Selected specific columns of interest for analysis.

Handled Null Values:

- Removed rows with null values in the 'Created Year' column.
- Filled missing values in categorical columns with 'Other'.

Handled Specific Values:

- Replaced the 'Created Year' value for the 'YouTube' channel from 1970 to 2005.
- Removed rows with blank 'Youtuber' names (channels with only special characters).

Handled Date Columns:

- Converted the columns “publishedAt” and “trending_date” to proper date format to perform calculations.

Handled Duplicates and Zero Views:

- Checked for and removed duplicate rows.
- Replacing the values with 0 in the view count column of the channel-level data with the mean of the view count, since the dataset holds data about the top channels.

Data Type Conversion:

- Converted data types of columns such as.” Created year”, and “View count” to int64 to perform calculations on the data.

6 EXPLORATORY DATA ANALYSIS

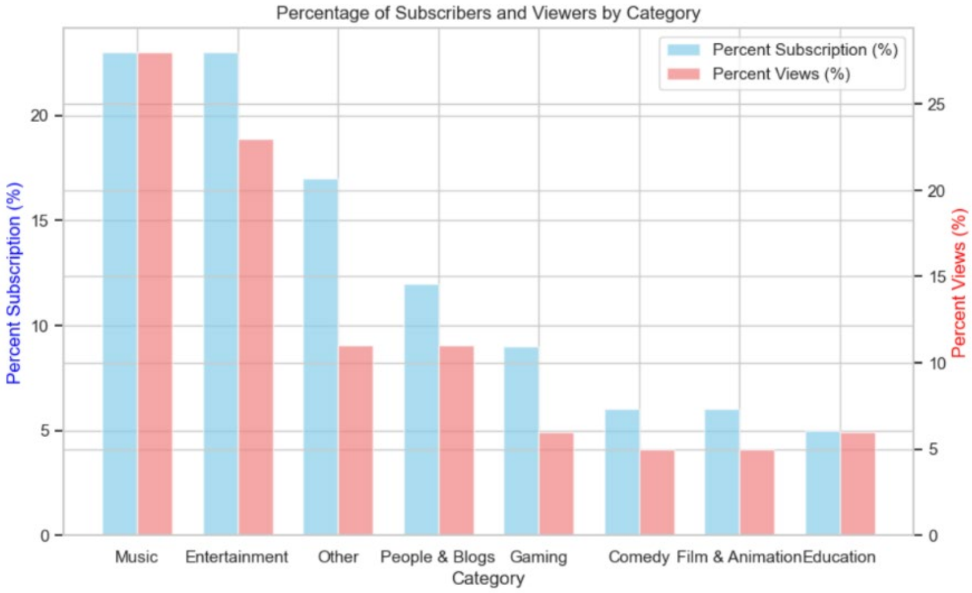
The data analysis uncovers various facets, Top categories, average time to trend, video upload patterns, common words used in trending videos, trends in channel creation over the years, and categories with comments and ratings disabled followed by a regression model.

6.1 Top categories based on subscriber and viewer %

The 3 top categories in terms of subscriber % and viewer % are:

1. Music
2. Entertainment
3. People and Blogs

Music is the top category, which has the highest subscriber percentage as well as the viewer percentage, followed by Entertainment and People and blogs.

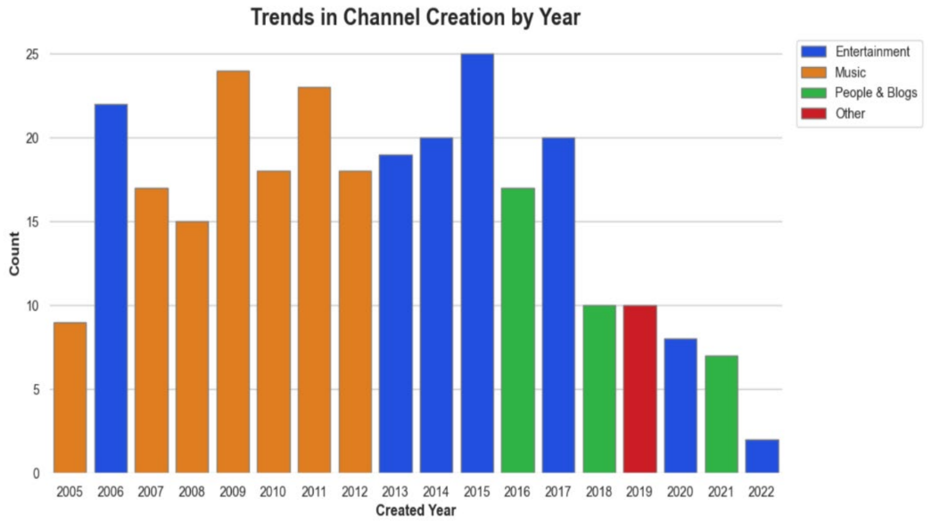


6.2 Top categories based on subscriber and viewer %

Analyzing the trends in channel creation reveals insightful patterns over the years. The data graph prominently illustrates that between 2007 and 2012, the peak era for channel inception was observed in the Music category.

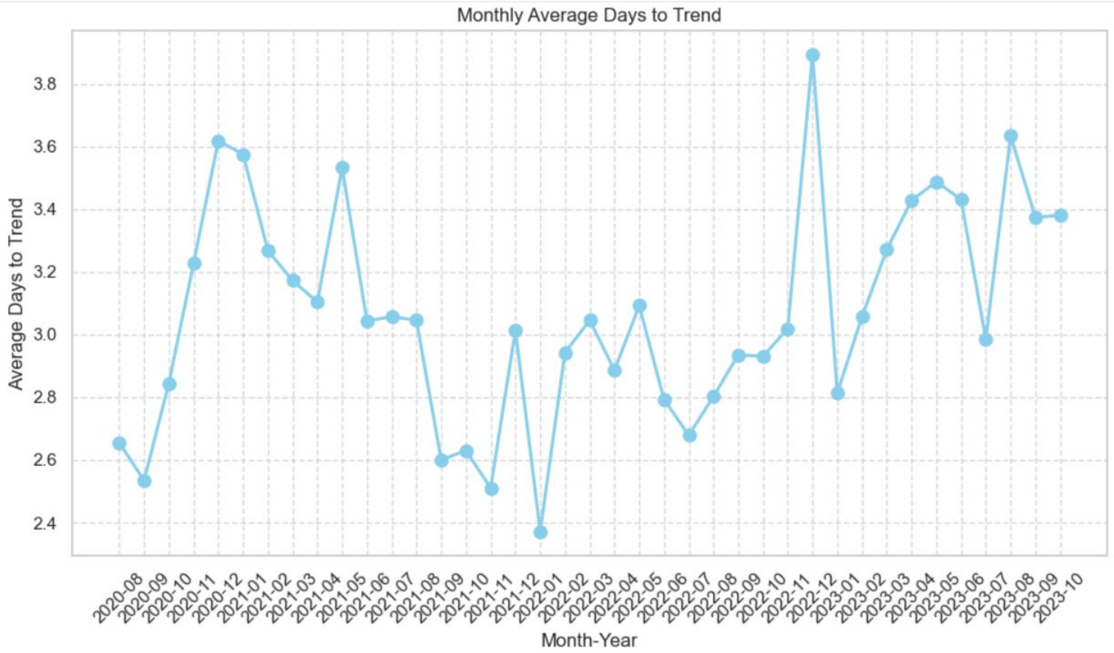
A shift in the landscape occurred during the subsequent years, particularly from 2013 to 2015 when Entertainment channels took precedence as the most actively created. However, an intriguing observation emerges as we move forward in time – a discernible decline in channel creation counts.

This notable trend suggests a nuanced evolution in the digital content space, potentially indicative of increased competition from emerging platforms. The reduced frequency of new channel creations could signify a diversification of content distribution channels or a strategic shift among content creators. This dynamic landscape invites a deeper exploration into the factors influencing channel creation trends and the broader implications for the competitive dynamics in the online content creation domain.



6.3 Average time taken for a video to trend.

It is helpful for marketers to assess how long it takes for a video to trend to optimize their sponsor the content creator so that their advertisements get viewed by more people. By looking at the graph we can say that on average it takes 3 days for the video to trend.

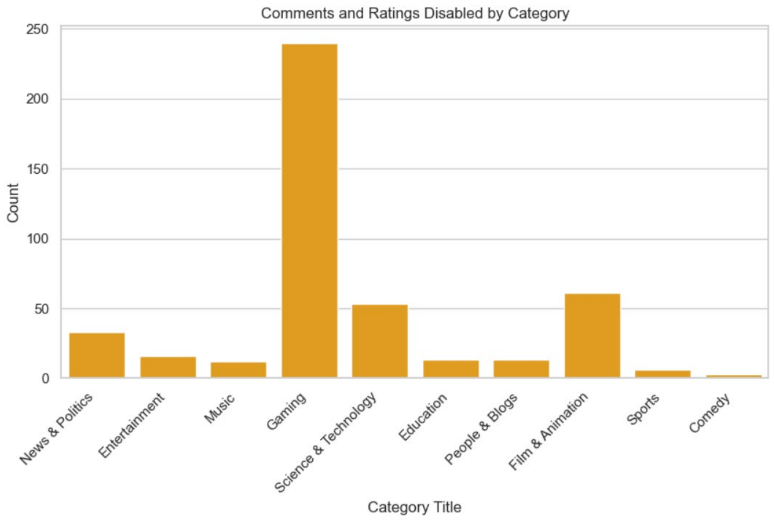


6.4 Categories by which the comments and ratings were disabled.

The graph illustrates how many videos the comments and ratings were disabled by category, this offers crucial insights for both content creators and advertisers.

On the advertiser front, the graph serves as a valuable tool for brand safety considerations. Advertisers often assess the context in which their ads will be displayed, and the knowledge of disabled comments and ratings in specific categories aids in evaluating the suitability of these environments. This insight is crucial for maintaining brand safety and ensuring alignment with the tone and nature of the content surrounding their advertisements.

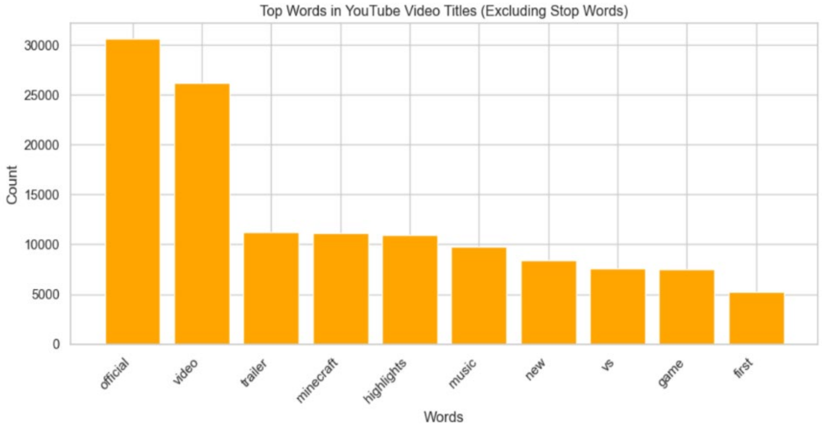
Moreover, the graph contributes to a broader understanding of the YouTube platform and its diverse communities, benefiting both content creators and advertisers.



6.5 Top words used in the titles and descriptions.

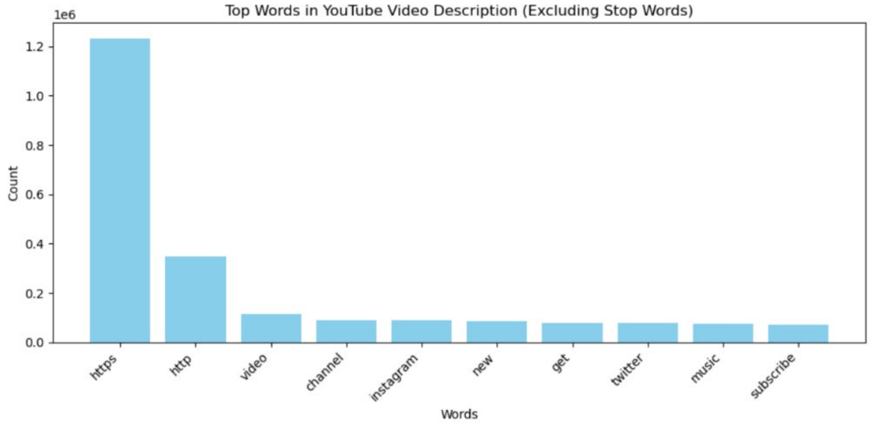
An exploration into the linguistic dynamics of trending videos exposes prevalent themes in both titles and descriptions. The frequently occurring terms in titles, such as "official," "video," "trailer," and "Minecraft," provide a snapshot of the content genres that prominently capture audience attention. In the descriptions, a distinctive set of words, including "HTTPS," "HTTP," "channel," "Instagram," and "Twitter," indicates a significant trend – the inclusion of external links to various websites. This revelation offers content creators a valuable key to understanding the connection between common words and the practice of sharing links in trending videos.

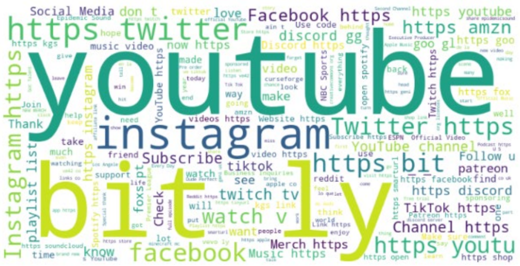
Top Words used in titles (excluding stop words):



Official Audio
GAME HIGHLIGHTS
Minecraft
Among Us
Day
New
Music Video
short
FULL GAME
Video
Oficial
Built
Made
Official Trailer
Official Video
Hour
Lyric Video
Official
Music

Top Words used in Description (excluding stop words):

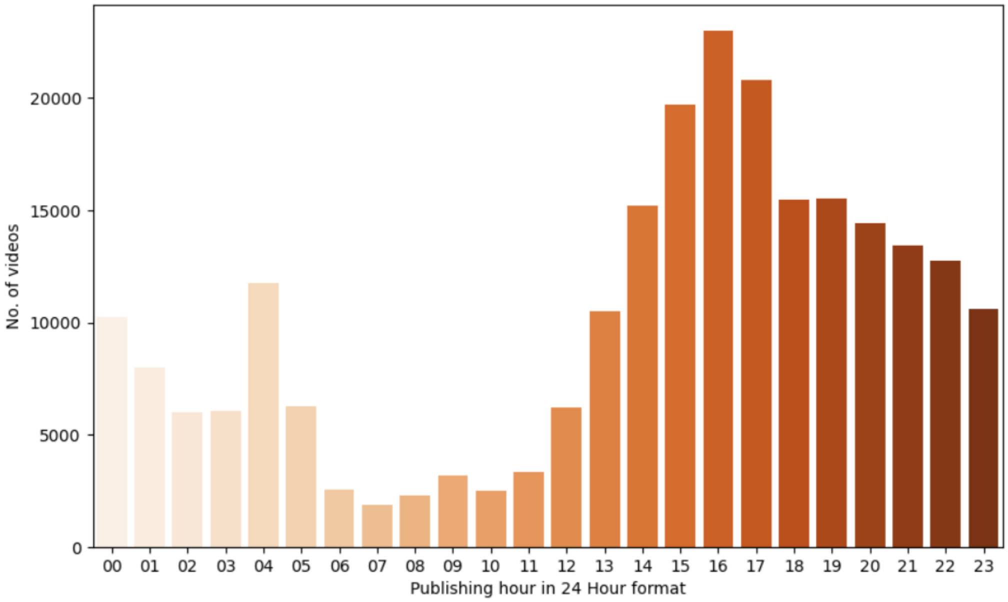




This insight serves as a practical guide for content creators seeking to optimize their video titles and descriptions effectively. By strategically incorporating these prevalent words, creators can enhance the visibility and engagement of their content, potentially positioning it for trending status. This nuanced approach empowers content creators to align their language with the prevailing terms observed in successful videos, providing them with a valuable tool to navigate the competitive digital landscape and cater to the evolving preferences of their audience.

6.6 Time frame in which the trending videos were uploaded.

The graphical representation illustrates the distribution of the number of trending videos across a 24-hour format, offering valuable insights into the temporal dynamics of video uploads. Notably, the data reveals a discernible trend, indicating that a significant majority of trending videos were uploaded during the time window spanning from 1 pm to 6 pm. This concentration of uploads during the afternoon and early evening hours suggests a strategic pattern among content creators. The choice of this time frame may be indicative of creators aiming to capitalize on peak viewer activity, aligning their video releases with periods of heightened audience engagement. This temporal analysis sheds light on optimal upload times, providing content creators with valuable information to refine their scheduling strategies and potentially maximize the visibility and impact of their videos in the competitive landscape of trending content.



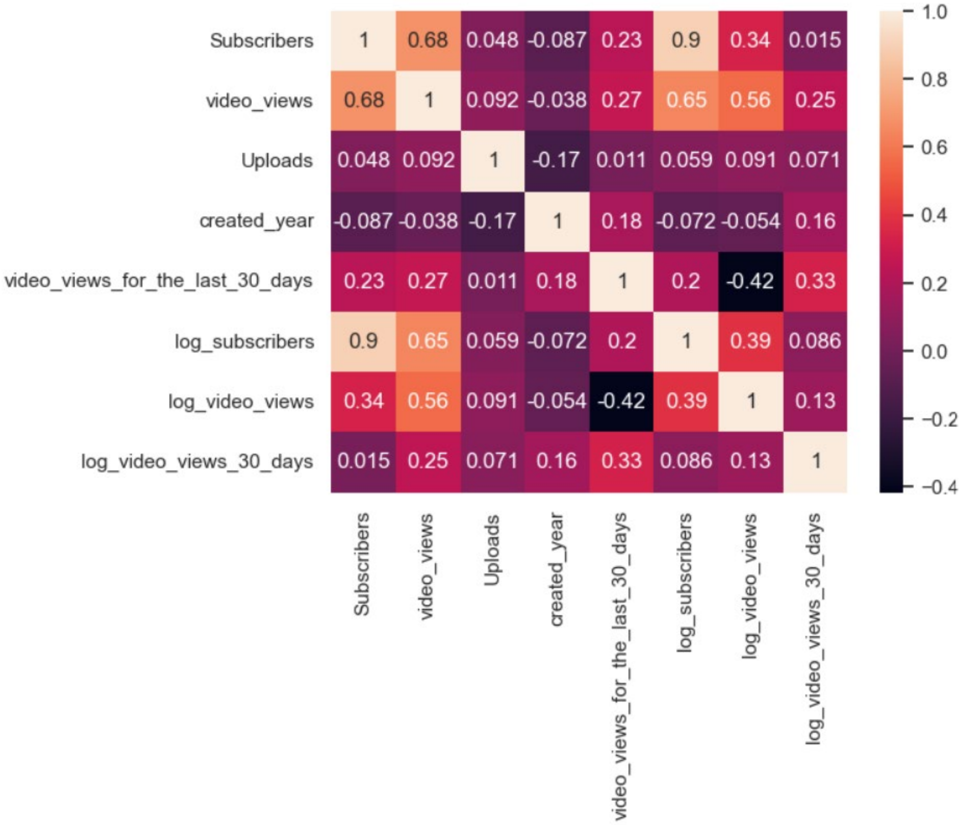
7 Regression analysis for Subscribers

The analytical insights derived from the data include a perceptive observation regarding the relationship between video views and subscriber growth on YouTube channels. The established coefficient of 0.0009 signifies that for every single view on a YouTube channel, an approximate addition of 0.0009 subscribers occurs. This quantitative metric serves as a valuable benchmark, offering content creators and channel managers a clear understanding of the subscriber acquisition rate corresponding to video views.

Another noteworthy observation pertains to the Last 30-day view coefficient surpassing that of video views. This distinction implies that videos go viral, garnering substantial attention within a specific time frame, resulting in a more pronounced surge in subscribers compared to videos with consistently high view counts. This nuanced insight underscores the significance of virality in not only amplifying immediate viewership but also fostering sustained channel growth through increased subscriber numbers.

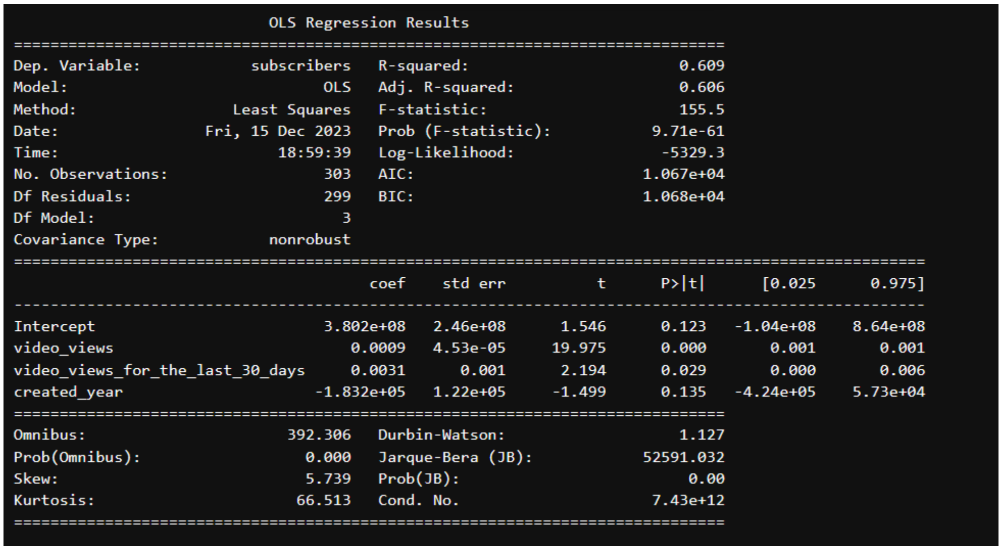
Additionally, the inverse coefficient associated with the creation year of a channel introduces a temporal dimension to subscriber accumulation. The discerned trend suggests that older channels, having been created for an extended period, exhibit a positive correlation with higher subscriber counts. This finding implies that channels with a longer history tend to accrue a more substantial subscriber base over time, emphasizing the enduring impact and longevity of content creators on the platform.

Correlation plot:



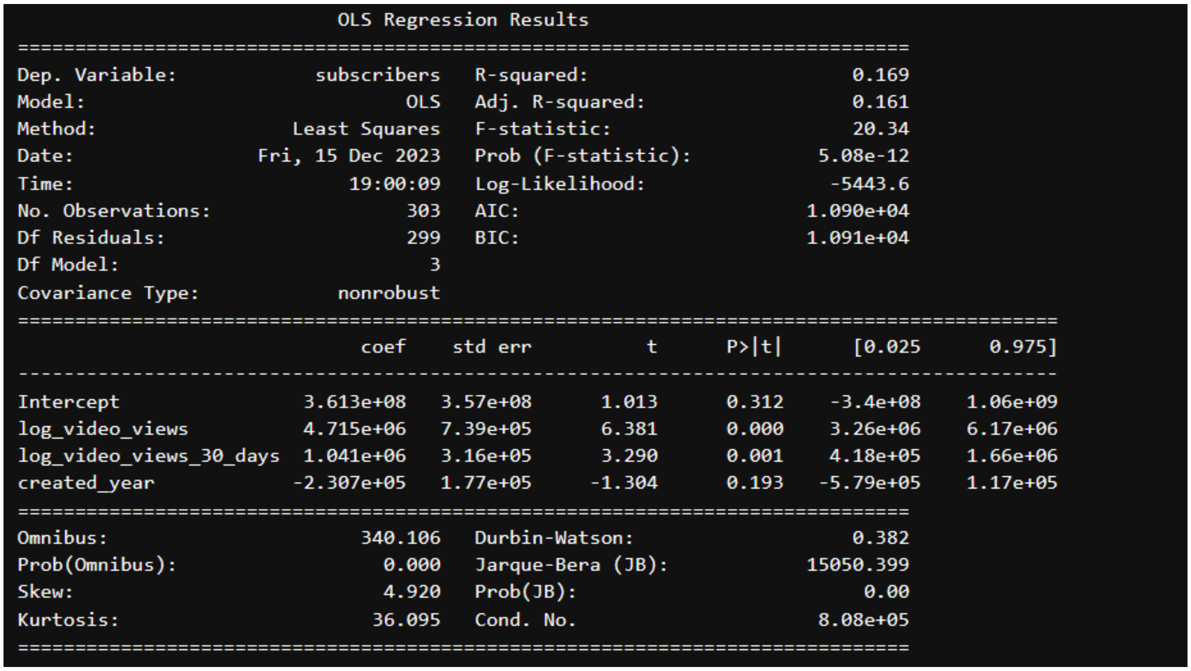
	Subscribers	video_views	Uploads	created_year	video_views_for_the_last_30_days	log_subscribers	log_video_views	log_video_views_30_days
Subscribers	1	0.68	0.048	-0.087	0.23	0.9	0.34	0.015
video_views	0.68	1	0.092	-0.038	0.27	0.65	0.56	0.25
Uploads	0.048	0.092	1	-0.17	0.011	0.059	0.091	0.071
created_year	-0.087	-0.038	-0.17	1	0.18	-0.072	-0.054	0.16
video_views_for_the_last_30_days	0.23	0.27	0.011	0.18	1	0.2	-0.42	0.33
log_subscribers	0.9	0.65	0.059	-0.072	0.2	1	0.39	0.086
log_video_views	0.34	0.56	0.091	-0.054	-0.42	0.39	1	0.13
log_video_views_30_days	0.015	0.25	0.071	0.16	0.33	0.086	0.13	1

Linear Model:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:            subscribers   R-squared:                       0.609
Model:                            OLS   Adj. R-squared:                  0.606
Method:                 Least Squares   F-statistic:                     155.5
Date:                Fri, 15 Dec 2023   Prob (F-statistic):           9.71e-61
Time:                        18:59:39   Log-Likelihood:                -5329.3
No. Observations:                 303   AIC:                         1.067e+04
Df Residuals:                     299   BIC:                         1.068e+04
Df Model:                           3
Covariance Type:            nonrobust
====================================================================================================
                                       coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------------
Intercept                         3.802e+08   2.46e+08      1.546      0.123   -1.04e+08    8.64e+08
video_views                          0.0009   4.53e-05     19.975      0.000       0.001       0.001
video_views_for_the_last_30_days     0.0031      0.001      2.194      0.029       0.000       0.006
created_year                     -1.832e+05   1.22e+05     -1.499      0.135   -4.24e+05    5.73e+04
==============================================================================
Omnibus:                      392.306   Durbin-Watson:                   1.127
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            52591.032
Skew:                           5.739   Prob(JB):                         0.00
Kurtosis:                      66.513   Cond. No.                     7.43e+12
==============================================================================
```

Log Model:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:            subscribers   R-squared:                       0.169
Model:                            OLS   Adj. R-squared:                  0.161
Method:                 Least Squares   F-statistic:                     20.34
Date:                Fri, 15 Dec 2023   Prob (F-statistic):           5.08e-12
Time:                        19:00:09   Log-Likelihood:                -5443.6
No. Observations:                 303   AIC:                         1.090e+04
Df Residuals:                     299   BIC:                         1.091e+04
Df Model:                           3
Covariance Type:            nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
Intercept                3.613e+08   3.57e+08      1.013      0.312    -3.4e+08    1.06e+09
log_video_views          4.715e+06   7.39e+05      6.381      0.000    3.26e+06    6.17e+06
log_video_views_30_days  1.041e+06   3.16e+05      3.290      0.001    4.18e+05    1.66e+06
created_year            -2.307e+05   1.77e+05     -1.304      0.193   -5.79e+05    1.17e+05
==============================================================================
Omnibus:                      340.106   Durbin-Watson:                   0.382
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            15050.399
Skew:                           4.920   Prob(JB):                         0.00
Kurtosis:                      36.095   Cond. No.                     8.08e+05
==============================================================================
```

Although the log function standardizes the large numbers of subscribers and video views, the model that represents the data more accurately is the original one, assuming by the R-squared value. Therefore, the model without any log function applied will be the one that is most adequate to use for explanation and prediction.

8 RECOMMENDATIONS

The comprehensive analysis of YouTube data undertaken by our team has provided valuable insights for content creators and marketing analysts. The mission to unravel the dynamics of YouTube was successful in achieving actionable insights for enhancing impact and optimizing marketing strategies.

- Creators can optimize content strategies based on insights into upload patterns and audience engagement.
- Understanding common words used in titles and descriptions aids in content optimization.
- Relationships between views, subscribers, and channel age offer creators a nuanced understanding of factors influencing subscriber growth.
- Creators can tailor content based on the impact of video virality, sustained views, and channel longevity.

9 CONCLUSION

In conclusion, our analysis serves as a guide for navigating the intricate landscape of YouTube, offering creators and marketing professionals actionable insights for informed decisions and strategic planning. The fusion of exploratory data analysis, regression analysis, and predictive modeling has unlocked crucial patterns and relationships within YouTube data, empowering stakeholders to optimize their strategies for enhanced success and impact.

10 FUTURE SCOPE

- Building upon the current analysis, future endeavors could delve into advanced predictive modeling techniques. Developing models that forecast not only subscriber counts based on views but also predict future trends in video virality and viewer engagement would provide an even more nuanced perspective.

- The project could evolve to offer dynamic content strategy recommendations for creators, considering fluctuating trends and audience preferences. Machine learning algorithms could be employed to identify evolving patterns and suggest adaptive content strategies for maximizing viewership and subscriber growth.

- Expanding the analysis to include social media metrics and their impact on YouTube performance could offer a comprehensive view. Integrating data from platforms like Instagram and Twitter could unveil synergies between social media presence and YouTube success.

- Investigating regional and cultural nuances influencing content popularity could be an intriguing avenue. Analyzing the impact of cultural trends on video virality and subscriber growth could provide creators with insights into global audience dynamics.

- Exploring the impact of collaborations on subscriber growth and video virality could be a valuable addition. Analyzing successful collaboration patterns and identifying influential partnerships could guide creators in forming strategic alliances.

By steering the project towards these future avenues, we can contribute to a more comprehensive and adaptive understanding of YouTube dynamics, providing stakeholders with actionable insights to navigate the ever-evolving landscape of digital content creation.

11 REFERENCES

- Python documentation: 3.12.1 Documentation (python.org)
- Seaborn documentation: seaborn: statistical data visualization — seaborn 0.13.0 documentation (pydata.org)
- Matplotlib documentation: Using Matplotlib — Matplotlib 3.8.2 documentation
- Kaggle website: Find Open Datasets and Machine Learning Projects | Kaggle