# Global Youtube Statistics*

1st Kanyasorn Phongphaitoonsin
*School of Engineering and Technology*
*Asian Institute of Technology*
Pathum Thani, Thailand
st122983@ait.asia

2nd Nitesh Ghimire
*School of Engineering and Technology*
*Asian Institute of Technology*
Pathum Thani, Thailand
st124453@ait.asia

## I. Introduction

YouTube, the world's leading video-sharing platform, has fundamentally transformed the way we consume content. In 2023, the "Global YouTube Statistics 2023" dataset offers a window into this digital powerhouse, encapsulating an extensive array of data on user demographics, content categories, engagement patterns, and more. Whether you're a data scientist exploring trends, a marketer refining strategies, or simply curious about the digital culture, this dataset provides insights into the dynamic landscape of YouTube in the current year.

YouTube's influence spans continents and cultures, and this dataset reflects that global reach. It offers a comprehensive view of how users across the world engage with the platform, highlighting regional nuances and emerging trends. Dive into this treasure trove of information to uncover the unique stories and statistics that define YouTube's place in the global digital ecosystem. With the "Global YouTube Statistics 2023" dataset, you can gain a deeper understanding of the platform's impact, providing a valuable resource for research, analysis, and exploration of the magic of online video in the modern age.

## II. Problem statement

- **Emerging Trends and Predictive Analytics**: Can the dataset be used to identify emerging trends and make predictions about the future of YouTube in 2023 and beyond? How can machine learning and data analysis techniques be applied to forecast changes in user behavior and content preferences?
- **Predictive Analytics for User Engagement**: How can predictive models be developed to forecast user engagement and content performance on YouTube in 2023? Can we identify patterns and factors that correlate with video virality, audience retention, and viewer interaction, thereby helping content creators optimize their content release schedules and video production?
- **Prescriptive Content Strategies**: Based on predictive insights, how can we develop prescriptive content strategies that advise content creators on the types of content, formats, and timing that are most likely to attract and retain viewers? What actionable recommendations can be generated to improve content quality and reach a broader audience?

## III. Related works

### *A. Predictive analysis of YouTube trending videos using Machine Learning*

This research [1] focuses on the analysis of interactive features that contribute to the trendiness of videos on YouTube. Trending videos on YouTube represent content that is gaining viewership and has the potential to become popular. Despite their significance, there has been limited research in this area. The study aims to determine the correlation and importance of variables in making a video trend on YouTube.

The research is based on viewership statistics from over 40,000 YouTube trending videos over a specific time period. It uses machine learning techniques, including linear regression, to predict the number of views for trending videos. Additionally, the study compares various classification models to predict how long it takes for a video to become trending from its upload time and how long it stays on the trending list. The research achieved a maximum accuracy of 62.53% in predicting the lifecycle of YouTube trending videos. It follows the CRISP DM methodology and uses a correlational quantitative research approach to bring objectivity to the understanding of what makes a video trend on YouTube.

### *B. Youtube Trending Videos: Boosting Machine Learning Results Using Exploratory Data Analysis*

This paper [2] delves into the analysis of YouTube trending video data to understand what captures users' attention in a short time frame. It highlights the significance of utilizing Big Data Technologies, statistics, data mining, and machine learning algorithms to transform massive data into valuable insights. The analysis involves examining time-series YouTube data from 40,000 trending videos collected over 205 days.

Through exploratory data analysis (EDA), the research explores various aspects of the data to gain insights and utilizes statistics to identify patterns and similarities between different video categories. The study also investigates how user activity varies over time, considering the nature of events that impact data quality. By directly analyzing viewership patterns in each category, the research aims to predict the next trending video categories, potentially assisting content creators in optimizing their video uploads for increased views and performance based on these predictions and analyses.

### C. Popularity Prediction of Videos in YouTube as Case Study: A Regression Analysis Study

The paper [3] focuses on predicting the popularity of videos on YouTube using regression analysis. It serves as a case study to understand the factors influencing the number of views a video receives on the platform.

The research aims to identify and analyze the variables that play a crucial role in determining a video's popularity on YouTube. It employs regression analysis, a statistical technique, to model the relationships between various factors and the number of views a video receives. By doing so, the study attempts to provide insights into the key drivers of video popularity, which can be valuable for content creators and marketers seeking to optimize their video content and strategies on YouTube.

Overall, this paper offers a quantitative analysis of YouTube video popularity, providing a data-driven approach to understanding the dynamics of video views on the platform.

### D. The Statistical Analysis based on YouTube Channel Dataset

The paper [4] conducts a statistical analysis using a dataset related to YouTube channels. While the specifics of the analysis may vary based on the paper's content, the general approach involves employing statistical methods to gain insights from the dataset.

The research likely delves into the exploration of data patterns, relationships, and trends within the YouTube channel dataset. This can involve examining variables such as subscriber counts, video views, engagement metrics, and other channel-related data to understand how they are distributed and whether there are significant correlations or dependencies among these variables.

Overall, the paper uses statistical analysis to provide a data-driven understanding of YouTube channel-related data, shedding light on key factors that contribute to a channel's performance and popularity on the platform.

## IV. Datasets

The "Global YouTube Statistics 2023 Dataset" is an expansive data collection that offers a comprehensive view of YouTube in the year 2023. Comprising 28 distinct features, it paints a detailed picture of YouTube's user base, content trends, and engagement metrics. From user demographics to content creator earnings, channel origins to engagement dynamics, and even socio-economic indicators, this dataset encapsulates the diverse facets of the YouTube ecosystem. With 995 records at its core, it serves as an invaluable resource for those seeking to understand the ever-evolving landscape of YouTube in 2023. The Table below show the key feauture in this dataset. [5]

TABLE I
KEY FEATURES

| | |
|---|---|
| rank | Position of the YouTube channel based on the number of subscribers |
| Youtuber | Name of the YouTube channel |
| subscribers | Number of subscribers to the channel |
| video views | Total views across all videos on the channel |
| category | Category or niche of the channel |
| Title | Title of the YouTube channel |
| uploads | Total number of videos uploaded on the channel |
| Country | Country where the YouTube channel originates |
| Abbreviation | Abbreviation of the country |
| channel_type | Type of the YouTube channel (e.g., individual, brand) |
| video_views_rank | Ranking of the channel based on total video views |
| country_rank | Ranking of the channel based on the number of subscribers within its country |
| channel_type_rank | Ranking of the channel based on its type (individual or brand) |
| video_views_for_the_last_30_days | Total video views in the last 30 days |
| lowest_monthly_earnings | Lowest estimated monthly earnings from the channel |
| highest_monthly_earnings | Highest estimated monthly earnings from the channel |
| lowest_yearly_earnings | Lowest estimated yearly earnings from the channel |
| highest_yearly_earnings | Highest estimated yearly earnings from the channel |
| subscribers_for_last_30_days | Number of new subscribers gained in the last 30 days |
| created_year | Year when the YouTube channel was created |
| created_month | Month when the YouTube channel was created |
| created_date | Exact date of the YouTube channel's creation |
| Gross tertiary education enrollment (%) | Percentage of the population enrolled in tertiary education in the country |
| Population | Total population of the country |
| Unemployment rate | Unemployment rate in the country |
| Urban$_{population}$ | Percentage of the population living in urban areas |
| Latitude | Latitude coordinate of the country's location |
| Longitude | Longitude coordinate of the country's location |

In this dataset, one can explore the influence and reach of content creators, the popularity of content categories, and the dynamics of viewer engagement. It goes beyond the numbers to examine the financial aspects of YouTube, channel creation details, and socio-economic factors, providing a holistic understanding of the platform's impact on society and the digital world.

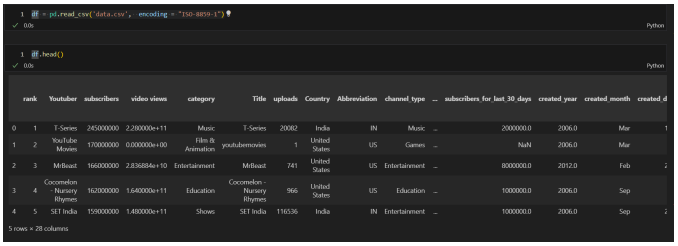
## V. Methodology

- **Data Preprocessing:**
  1) **Import Library:** Ensure necessary libraries are imported for data analysis and machine learning.

```python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np

from sklearn import metrics
from datetime import datetime
import datetime


from sklearn.preprocessing import scale
from sklearn.model_selection import train_test_split
from sklearn.metrics import euclidean_distances
from sklearn.neighbors import KNeighborsClassifier

from sklearn.model_selection import GridSearchCV
from sklearn.metrics import mean_squared_error
from sklearn.model_selection import KFold, cross_val_score
from sklearn.linear_model import LinearRegression,Ridge,Lasso,ElasticNet
from sklearn.ensemble import RandomForestRegressor
from sklearn.preprocessing import OneHotEncoder
from sklearn.preprocessing import StandardScaler

from sklearn.compose import ColumnTransformer # transform specific columns
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler, MinMaxScaler
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import OrdinalEncoder, OneHotEncoder
import scipy.stats as stats
```

Fig. 1. Enter Caption

2) **Load Data:** Load the dataset for analysis.



```
df = pd.read_csv('data.csv', encoding = "ISO-8859-1")
0.0s
Python

df.head()
0.0s
Python
```

| | rank | Youtuber | subscribers | video views | category | Title | uploads | Country | Abbreviation | channel type | ... | subscribers for last 30 days | created year | created month | created d |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | T-Series | 245000000 | 2.280000e+11 | Music | T-Series | 20082 | India | IN | Music | ... | 2000000.0 | 2006.0 | Mar | 1 |
| 1 | 2 | YouTube Movies | 170000000 | 0.000000e+00 | Film & Animation | youtubemovies | 1 | United States | US | Games | ... | NaN | 2006.0 | Mar | |
| 2 | 3 | MrBeast | 166000000 | 2.836884e+10 | Entertainment | MrBeast | 741 | United States | US | Entertainment | ... | 8000000.0 | 2012.0 | Feb | 2 |
| 3 | 4 | Cocomelon - Nursery Rhymes | 162000000 | 1.640000e+11 | Education | Cocomelon - Nursery Rhymes | 966 | United States | US | Education | ... | 1000000.0 | 2006.0 | Sep | |
| 4 | 5 | SET India | 159000000 | 1.480000e+11 | Shows | SET India | 116536 | India | IN | Entertainment | ... | 1000000.0 | 2006.0 | Sep | 2 |

5 rows × 28 columns

Fig. 2. Enter Caption

3) **See Data Information:** Explore basic information about the dataset.
4) **Check Missing Values:** Identify and handle missing values through filling or dropping.

- **Exploratory Data Analysis (EDA):**
  1) **Top 10 Channels by Subscribers:** Identify and display the channels with the highest subscriber counts.
  2) **Subscriber vs Video Views:** Analyze the relationship between subscriber count and video views.
  3) **Video Views vs Highest Monthly Earnings:** Explore the correlation between video views and monthly earnings.
  4) **Count of YouTube Channels by Categories:** Understand the distribution of channels across different categories.
  5) **Categories with the Highest Number of Subscribers:** Identify categories with the most subscribers.
  6) **YouTube Channels with Highest Subscribers Group by Country:** Examine the distribution of top subscriber channels across countries.
  7) **Top 10 Countries with the Highest Number of Channels:** Identify countries with the most YouTube channels.
  8) **Top 10 Channels by Views:** List the channels with the highest views.
  9) **Highest Viewed Channels by Categories:** Explore the highest viewed channels within each category.
  10) **Top 10 YouTube Channels with Highest Uploads to Subscriber Ratio:** Identify channels with the highest uploads to subscriber ratio.
  11) **Highest Earning Categories:** Explore which categories generate the highest earnings.
  12) **Distribution of Channel Creation Year:** Analyze the distribution of channel creation over the years.
  13) **Correlation of the Dataset Within Columns:** Explore correlations between different columns.
- **Machine Learning:**
  1) **Select Features and Target:** Discern the features and target variable for the machine learning model by examining the correlation matrix. Within the scope of this project, we have devised four distinct models. **Model 1: Predicting Video Views**
  Target Variable: Video Views
  Features:
     a) Subscribers
     b) Uploads
     c) Country
     d) Channel types

  **Model 2: Predicting Video Views Version 2**
  Target Variable: Video Views
  Features:
     a) Subscribers
     b) Uploads
     c) Average monthly earning
     d) Average yearly earning

  **Model 3: Predicting Video Views Version 3**
  Target Variable: Video Views
  Features:
     a) Subscribers
     b) Uploads
     c) Average monthly earning
     d) Average yearly earning
     e) Country
     f) Channel types

  **Model 3: Predicting Subscribers**
  Target Variable: Subscribers
  Features:
     a) Video views
     b) Uploads
     c) Average monthly earning
     d) Average yearly earning
     e) Country
     f) Channel types

  The initial models yielded suboptimal results, prompting us to reassess and modify the features. Consequently, we iterated through models 2, 3, and 4. Model 4 emerged as the most fitting solution to address the project's core challenge, delivering notably superior results when compared to its counterparts.
  2) **Cross Validation for Model and GridSerchCV for Parameter Selection:** In our project, we employed Cross Validation to systematically identify the most effective model and its optimal parameters. The three types of models explored were Linear Regression, RandomForest, and Decision Tree Regressor. In the process of conducting GridSearchCV, we explored various hyperparameter combinations, specifically selecting values for max_depth from 5, 10, and 15, and n_estimators from 5 to 15
  3) **Train the Model:** We proceeded to train the chosen machine learning model using an 80/20 train-test split, allocating 80% of the data for training and reserving the remaining 20% for testing.

4) **Pickle Model:** Save the trained model for future use.

- **Deployment:**
  1) **Create HTML:** Develop an HTML representation of the analysis.
  2) **Deploy on AWS:** Deploy the analysis, potentially including the machine learning model, on AWS.

## VI. Results and Discussions

1) Exploratory Data Analysis (EDA) Result

- **Top 10 channels by subscribers**
  The leading YouTube channel by views is T-Series, a distinguished Indian music label and film production company celebrated for its extensive catalog of Bollywood songs and captivating music videos. Following, YouTube Music secures the second position, offering a rich library of songs, playlists, and music videos. In the third spot is MrBeast, known for his inventive challenges and philanthropic endeavors. The graphical representation of these rankings is depicted in Figure 3 below.

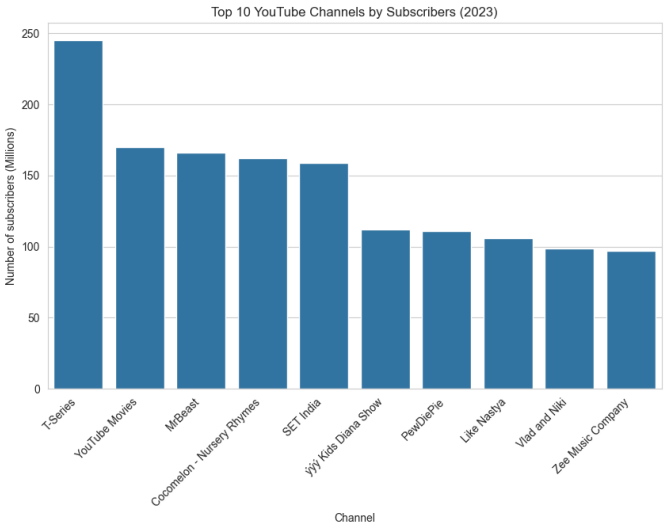



Fig. 3. Top 10 YouTube Channels by Subscribers (2023)

- **Subscriber vs Video Views**
  The captivating scatter plot, gracefully portrayed in Figure 4, unfolds a narrative of harmony, revealing a harmonious and positive correlation between the number of subscribers and the enchanting world of video views.

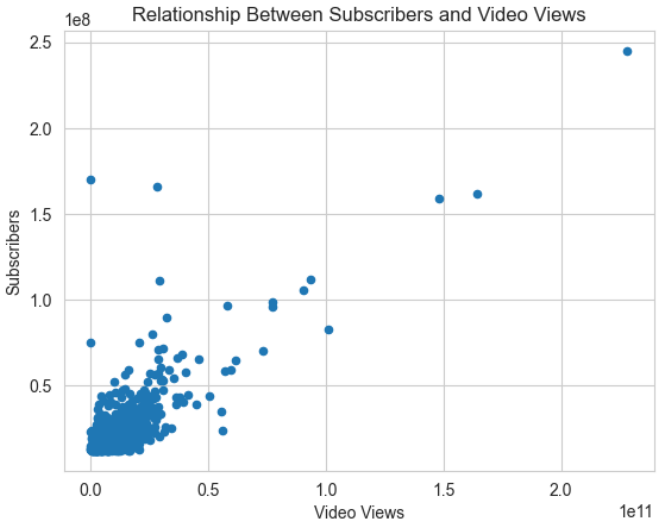



Fig. 4. Relationship Between Subscribers and Video Views

- **Video Views vs Highest Monthly Earnings**
  The evocative scatter plot, elegantly showcased in Figure 5, unveils a nuanced story. While a relationship exists, it leans towards a gentle ebb and flow rather than a robust connection, indicating that the correlation between video views and the highest monthly earnings is not overwhelmingly strong.

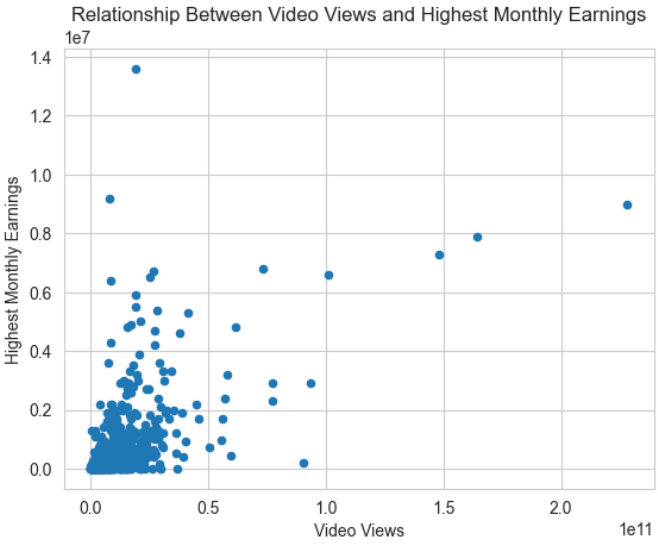



Fig. 5. Relationship Between Video Views and Highest Monthly Earnings

- **Count of Youtube channels by categories**
  At the pinnacle of YouTube's channel categories stands "Entertainment," reigning supreme with its diverse and captivating content. Following closely behind is the enchanting realm of "Music," enchanting viewers with a symphony of melodies and artistic expression. These two categories stand out prominently, outstripping others by a significant margin. The graphical representation of this categorical distribution is artfully depicted in Figure 6 below.

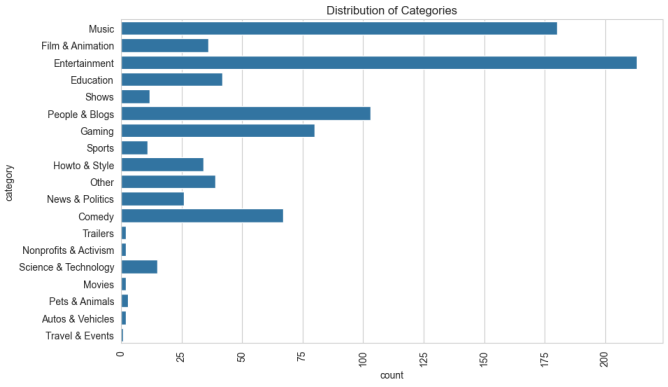



Fig. 6. Distribution of Categories

- **Categories with the highest number of subscribers**
  Among YouTube's categories, "Music" claims the top spot with the highest number of subscribers, and closely following in second place is the captivating realm of "Entertainment." These two categories stand out prominently, significantly outpacing others in terms of subscriber counts. The graphical representation of this noteworthy distribution is vividly illustrated in Figure 7 below.

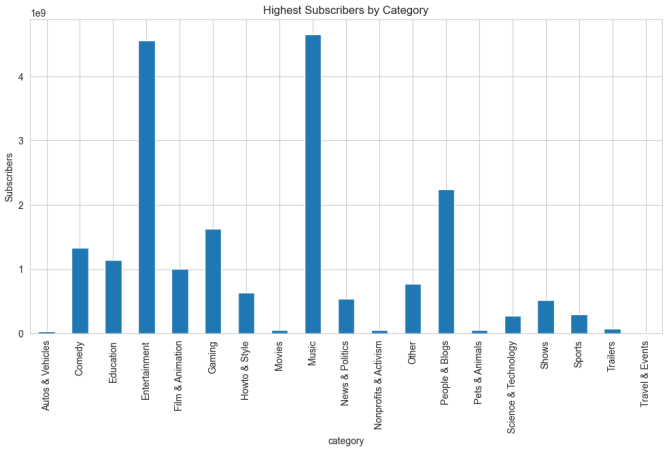



Fig. 7. Highest Subscribers by Category

- **Youtube Channels with highest subscriber group by country**

  India takes the lead as the country with the highest number of YouTube channels with subscribers, showcasing a robust and diverse digital presence. Following closely in second place, with a subscriber count hovering around 100 million, is the United States. The graphical representation of this global distribution is vividly presented in Figure 8 below.

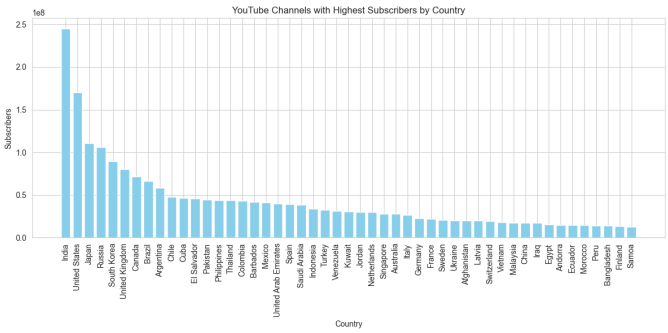



Fig. 8. YouTube Channels with Highest Subscribers by Country

- **Top 10 countries with highest number of channel**
  The United States claims the top position as the country with the highest number of YouTube channels, showcasing a vibrant and expansive digital landscape. In second place, India follows closely with a channel count that is approximately half that of the United States. The graphical representation of this global distribution is eloquently illustrated in Figure 10 below.

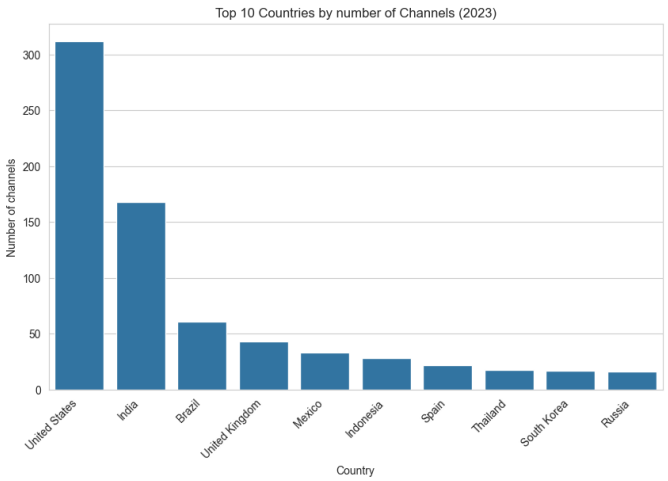



Fig. 9. Top 10 Countries by number of Channels (2023)

- **Top 10 channels by views**
  The reigning YouTube channel by views is T-Series, standing at the forefront with its extensive collection of music and film content. Following in second place is Cocomelon - Nursery Rhymes, capturing audiences with animated renditions of beloved nursery rhymes. In a strong third position, SET India trails closely, offering a diverse array of TV shows and dramas. The graphical representation of these view counts is visually depicted in Figure 10 below.

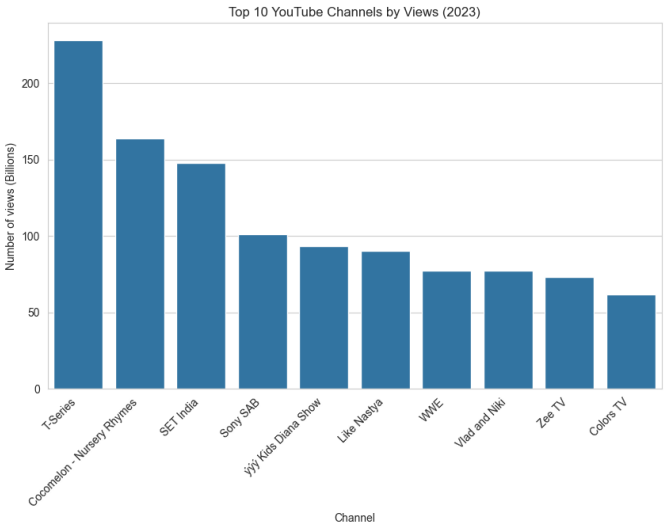



Fig. 10. Top 10 YouTube Channels by Views (2023)

- **Highest viewed channels by categories**
  The pinnacle of YouTube's video views is dominated by the "Music" category, claiming the top spot with a substantial viewership. Following closely in second place is the captivating realm of "Entertainment." These two categories stand out prominently, showcasing a significant and engaging presence on the platform. The graphical representation of this view distribution is vividly illustrated in Figure 11 below.

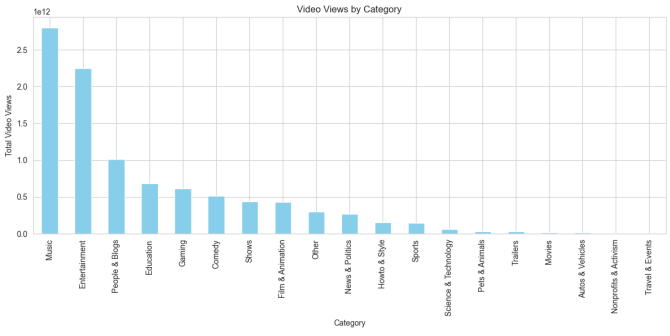



Fig. 11. Video Views by Category

- **Top 10 YouTube Channels with Highest Uploads to Subscriber Ratio**
  The GMA Integrated News secures the top position among YouTube channels with the highest uploads to subscriber ratio, indicating a dynamic and active engagement with its audience. Notably, the majority of the top 10 channels in this category belong to the news genre, emphasizing their commitment to delivering timely and frequent content. The graphical representation of these ratios is visually presented in Figure 12 below.

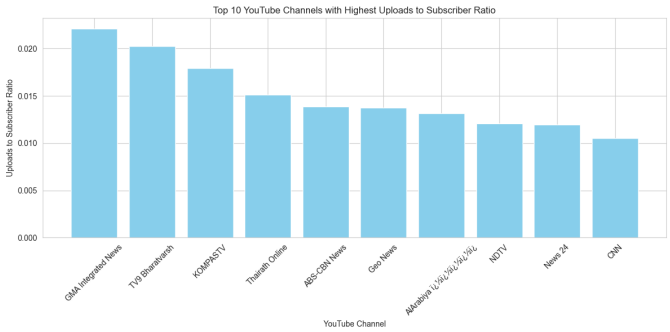



Fig. 12. Top 10 YouTube Channels with Highest Uploads to Subscriber Ratio

- **Top 10 channels by monthly earnings**
  In the bar plot depicted in Figure **??**, KIMPRO claims the top spot, generating an impressive monthly earning of approximately 7 million. Following closely, dafuq!?Boom! secures the second position, and T-Series takes the third spot, both with earnings that closely align. This graphical representation highlights the significant monthly earnings of these top-ranking YouTube channels.

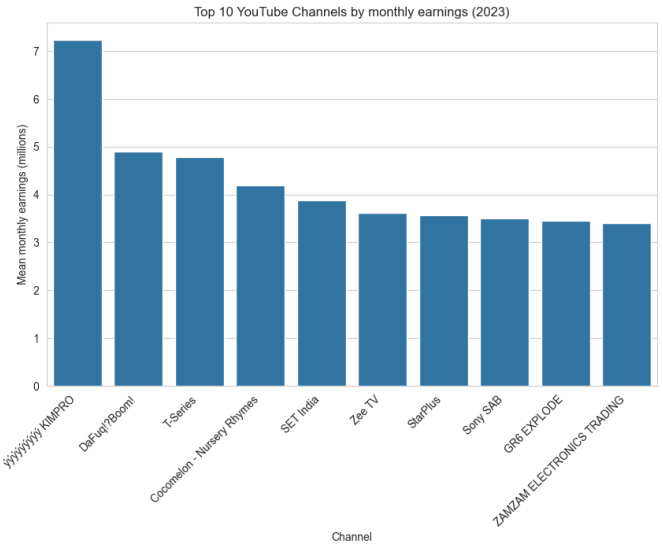



Fig. 13. Top 10 YouTube Channels by monthly earnings (2023)

- **Highest Earning Categories**
  The graph in Figure 14 unmistakably reveals that the ”Entertainment” category leads with the highest monthly earnings, underlining its financial prowess on the platform. Following closely in second place is the ”Music” category, emphasizing the substantial economic impact of these content genres on YouTube. These two categories distinctly stand out, showcasing a significant and engaging financial presence on the platform.

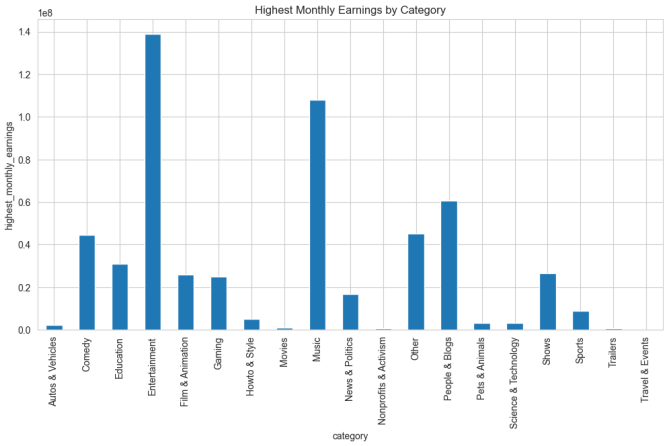



Fig. 14. Highest Monthly Earnings by Category

- **Distribution of the channel creation year**
  The distribution analysis indicates that the majority of channels were created between 2010 and 2020, with an anomalous entry suggesting a creation date in 1970, likely stemming from a data error. The dataset, sourced from the official YouTube channel itself, has been rectified by replacing the inaccurate creation date with the original launch date of YouTube, which is December 15, 2005. The updated results are visually presented in Figure **??** and Figure **??** below, reflecting a more accurate representation of the creation dates for YouTube channels.

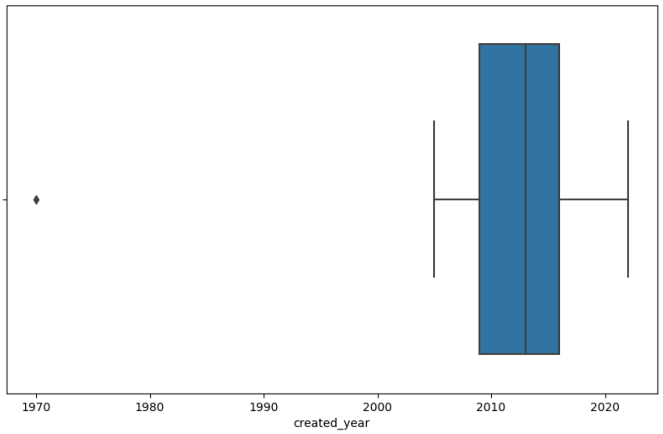



Fig. 15. Boxplot of created year before modify

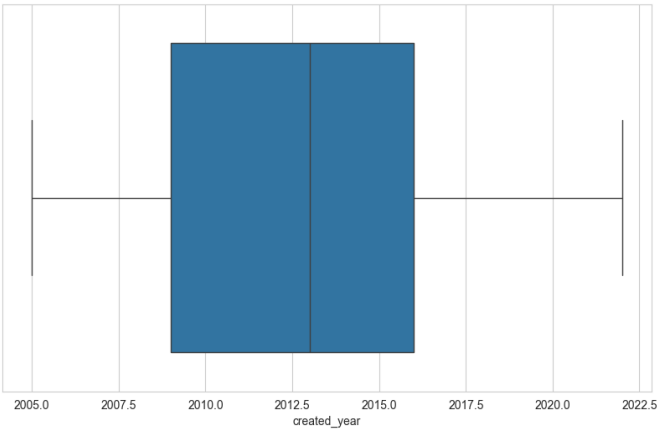



Fig. 16. Boxplot of created year

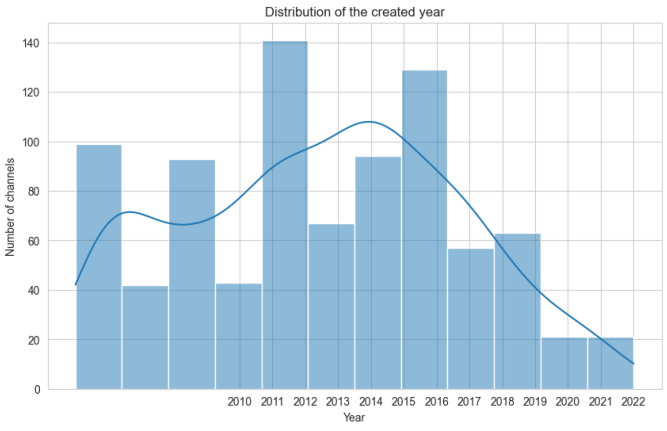



Fig. 17. Distribution of the created year

- **Correlation of the dataset within the columns** The outcome is illustrated in Figure **??** below, revealing noticeable correlations among certain features. The correlation analysis provides insights into the relationships between different variables, allowing for a better understanding of how they influence each other within the given dataset. Further exploration and interpretation of these correlations can provide valuable information for analysis and decision-making.

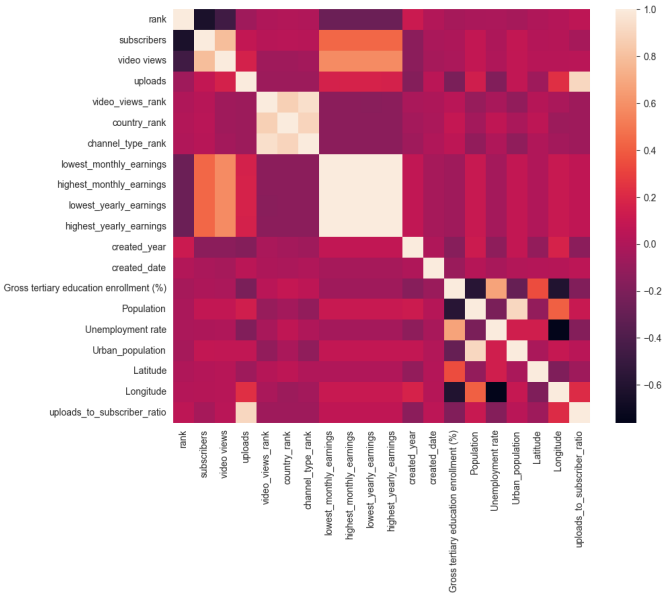



Fig. 18. Correlation of the dataset within the columns

2) Cross Validation Result
   a) Model 1 Based on the mean scores obtained, we

TABLE II
MEAN SCORES FOR DIFFERENT REGRESSION MODELS

| **Regression Model** | **Mean Squared Error Score** ($\times 10^{19}$) |
|---|---|
| Linear Regression | $-4.97 \times 10^{24}$ |
| Random Forest | $-8.14$ |
| Decision Tree Regressor | $-10.74$ |

selected the Random Forest model, which achieved the highest score among the three evaluated models.

   b) Model 2

TABLE III
MEAN SCORES FOR DIFFERENT REGRESSION MODELS

| **Regression Model** | **Mean Squared Error Score** ($\times 10^{19}$) |
|---|---|
| Linear Regression | $-6.09$ |
| Random Forest | $-6.48$ |
| Decision Tree Regressor | $-12.48$ |

Based on the mean scores obtained, we selected the Random Forest model, which achieved the highest score among the three evaluated models.

   c) Model 3

TABLE IV
MEAN SCORES FOR DIFFERENT REGRESSION MODELS

| **Regression Model** | **Mean Squared Error Score** ($\times 10^{19}$) |
|---|---|
| Linear Regression | $-4.99 \times 10^{24}$ |
| Random Forest | $-7.15$ |
| Decision Tree Regressor | $-10.72$ |

Based on the mean scores obtained, we selected the Random Forest model, which achieved the highest score among the three evaluated models.

   d) Model 4

TABLE V
MEAN SCORES FOR DIFFERENT REGRESSION MODELS

| **Regression Model** | **Mean Squared Error Score** ($\times 10^{14}$) |
|---|---|
| Linear Regression | $-1.41$ |
| Random Forest | $-1.67$ |
| Decision Tree Regressor | $-2.43$ |

Based on the mean scores obtained, we selected the Random Forest model, which achieved the highest score among the three evaluated models.

3) GridSearchCV Result

TABLE VI
HYPERPARAMETERS FOR RANDOM FOREST MODELS

| **Model** | **Regressor** | **max_depth** | **n_estimators** |
|---|---|---|---|
| **Model 1** | RandomForestRegressor | 5 | 6 |
| **Model 2** | RandomForestRegressor | 5 | 12 |
| **Model 3** | RandomForestRegressor | 10 | 11 |
| **Model 4** | RandomForestRegressor | 5 | 11 |

Clear documentation of these initial hyperparameters is essential for transparency and reproducibility in future model iterations. In summary, the consistent and thoughtful choice of hyperparameters reflects a balanced consideration of model complexity and performance, with room for potential fine-tuning to meet specific project requirements.

4) Train Result

TABLE VII
MEAN SQUARE ERROR FOR DIFFERENT MODELS

| **Model** | **Target Variable** | $R^2$ | **Mean Square Error** ($\times 10^{14}$) |
|---|---|---|---|
| **Model 1** | Video Views | 0.385 | $3.22 \times 10^{6}$ |
| **Model 2** | Video Views | 0.74 | $9.32 \times 10^{5}$ |
| **Model 3** | Video Views | 0.69 | $1.13 \times 10^{6}$ |
| **Model 4** | Subscribers | 0.71 | 1.80 |

Table VII displays mean square error (MSE) values for two regression models—Model 1,2 and 3 predicting Video Views and Model 4 predicting Subscribers. Although the R2 score may not be optimal, we have chosen to deploy Model 4 due to its minimal Mean Squared Error (MSE) values, making it the preferred selection for its overall predictive accuracy.

5) Deployment Result

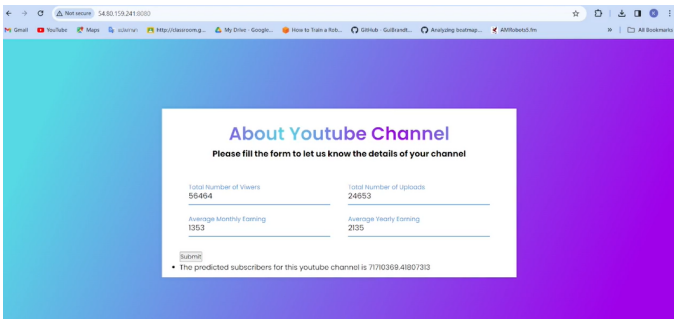



Fig. 19. Enter Caption

The deployed model on AWS functions as a visually appealing website featuring input fields for number of views, uploads, average monthly earnings and average yearly earnings. Upon user input, the website generates predictions of subscribers, displaying the results elegantly at the bottom of the page

## VII. Conclusions and Future Works

In conclusion, our comprehensive analysis of the ”Global YouTube Statistics 2023 Dataset” has provided valuable insights into the dynamics of YouTube in 2023. Through exploratory data analysis (EDA), we uncovered trends such as the dominance of the ”Entertainment” and ”Music” categories, the positive correlation between subscribers and video views, and the financial prowess of top channels in terms of monthly earnings.

We employed machine learning models to predict video views and subscribers based on various features, and after rigorous evaluation, selected the Random Forest model (Model 4) as the most accurate and reliable for our predictions. The model demonstrated notable performance with low mean squared error values, emphasizing its efficacy in forecasting YouTube metrics.

The deployment of the model on AWS as a user-friendly web application further enhances its accessibility and usability. Users can input relevant data, such as views, uploads, and earnings, to obtain predictions for the number of subscribers.

Our project not only provides valuable insights for content creators, marketers, and YouTube enthusiasts but also establishes a foundation for future research and applications in the realm of predictive analytics for online platforms.

There are several avenues for future work based on the insights and methodologies developed in this project: First, further refinement and optimization of machine learning models could enhance predictive accuracy. Fine-tuning hyperparameters and exploring advanced model architectures may lead to even more precise predictions. Second, integrating additional datasets, such as social media trends, cultural events, or technological advancements, could enrich the analysis and provide a more holistic view of the factors influencing YouTube metrics.

## References

[1] Aakash Ashok Niture, Predictive analysis of YouTube trending videos using Machine Learning,MSc Data Analytics, Dublin Business School,2020 - 21

[2] Sana Khanam, Safdar Tanweer, Syed Sibtain Khalid, Youtube Trending Videos: Boosting Machine Learning Results Using Exploratory Data Analysis, The Computer Journal, Volume 66, Issue 1, January 2023, Pages 35–46, https://doi.org/10.1093/comjnl/bxab142

[3] Soufana Mekouar, Nabila Zrira, and El-Houssine Bouyakhf. 2017. Popularity Prediction of Videos in YouTube as Case Study: A Regression Analysis Study. In Proceedings of the 2nd international Conference on Big Data, Cloud and Applications (BDCA'17). Association for Computing Machinery, New York, NY, USA, Article 51, 1–6. https://doi.org/10.1145/3090354.3090406

[4] Xintong Zhang and Zhefeng Sun. 2021. The Statistical Analysis based on YouTube Channel Dataset. In Proceedings of the 2021 1st International Conference on Control and Intelligent Robotics (ICCIR '21). Association for Computing Machinery, New York, NY, USA, 376–382. https://doi.org/10.1145/3473714.3473780

[5] https://www.kaggle.com/datasets/nelgiriyewithana/global-youtube-statistics-2023/data