CAPSTONE PROJECT

INTO THE DATA: A DEEP DIVE INTO THE TOP 1000 YOUTUBE CHANNELS

Presented By:
Yashashree Ravindra Mahajan
PCET's Nutan Maharashtra Institute of Engineering & Technology, Pune
Dept. Computer (BE CSE)
yashshreem2003@gmail.com

edunet foundation

OUTLINE

- Problem Statement
- Proposed System/Solution
- System Development Approach
- Algorithm & Deployment
- Result
- Conclusion
- Future Scope
- References

PROBLEM STATEMENT

- In today's digital landscape, YouTube plays a central role in shaping trends, influencing consumer behavior, and providing a platform for creativity, entertainment, and communication on a global scale.
- This project addresses the critical need for a systematic analysis framework to understand the landscape of the top YouTube channels, crucial for marketers, content creators, and platform developers. By identifying key factors driving channel growth, content preferences, and audience engagement strategies, it provides actionable insights for staying competitive and maximizing opportunities in YouTube content creation and distribution.
- Marketers, content creators, and platform developers stand to benefit significantly from the project's findings, enabling them to optimize strategies and capitalize on emerging opportunities in this dynamic environment.

PROPOSED SOLUTION

- **Description:** The proposed solution aims to provide comprehensive insights into YouTube's top subscribed channels through data analysis and visualization. By leveraging a dataset comprising the top 1,000 subscribed channels, we seek to uncover patterns, trends, and key metrics that can help stakeholders understand the YouTube landscape better.
- **Approach:**
    1) *Data Collection:* Utilize a dataset sourced from Kaggle, containing information on the top 1,000 subscribed YouTube channels.
    2) *Data Cleaning & Wrangling:* Address data quality issues, including missing values, incorrect data types, and inconsistencies.
    3) *Data Analysis & Visualization:* Conduct exploratory data analysis to uncover insights and visualize key metrics using charts and graphs.
    4) *Relationship Analysis*: Explore relationships between variables such as subscribers, video views, and video counts to gain deeper insights.
    5) *Classification Analysis*: Classify channels based on categories, channel age, and other relevant factors to identify trends and patterns.
- ***Expected Outcome***: A detailed analysis report highlighting the findings, trends, and actionable insights derived from the top subscribed channels dataset.

SYSTEM APPROACH

- **Software Requirements**:
  - Python: Programming language for data analysis and visualization.
  - Jupyter Notebook: Interactive development environment for Python.
  - Pandas: Python library for data manipulation and analysis.
  - Matplotlib & Seaborn: Python libraries for data visualization.
  - NumPy: Python library for numerical computing.
- **Hardware Requirements**:
  - Operating System: Windows, macOS, or Linux.
  - Processor: Intel Core i5 or equivalent.
  - RAM: Minimum 8 GB.
  - Storage: Sufficient disk space for dataset storage and analysis.

ALGORITHM & DEPLOYMENT

Algorithm & Deployment: As EDA primarily involves data exploration and analysis rather than predictive modelling, there is no specific algorithm or deployment process involved. Instead, we will focus on using descriptive statistics, visualizations, and exploratory techniques to gain insights into the dataset.

1) **Loading Necessary Libraries and Dataset**: Prepare the environment and data for analysis.
    a. **Steps**:
        i. Imported essential libraries: NumPy, Pandas, Matplotlib, Seaborn.
        ii. Loaded the dataset containing information about the top 1,000 subscribed YouTube channels.
        iii. Checked basic dataset information, such as column names, data types, and missing values.

2) **Data Cleaning & Wrangling**
    a. **Data Problems Identification**: Identify and address data quality issues.
        i. **Issues Identified**:
            1. Zero values in video views and video count.
            2. Missing values in the category column.
            3. Incorrect data types.
            4. Erroneous year values.
    b. **Cleaning & Format Data**: Resolve identified issues to ensure data quality.
        i. **Steps**:
            1. Dropped observations with zero values and missing data.
            2. Corrected erroneous year values.
            3. Addressed duplicated data.

ALGORITHM & DEPLOYMENT

3) **Wrangling the Data:** Generate additional metrics and insights from the cleaned data.
   a. **Steps**:
      i. Calculated channel age.
      ii. Computed average subscribers, video views, and video count per year.
      iii. Determined average subscribers and views per video.
   b. **Insights from Wrangling**:
      i. **Channel Age**: Median age is 10 years, suggesting channels need time to establish.
      ii. **Subscribers Growth Rate**: Median growth rate is 1,687,500 subscribers/year.
      iii. **Views Growth Rate**: Median growth rate is 644,577,700 views/year.
      iv. **Efficiency Metrics**: Average subscribers per video is 18,674.

4) **Exploratory Data Analysis:**

   Gain deeper insights into the dataset through visualization and statistical analysis. Addressed relationships between variables and category-wise statistics.

   a. **Variable Relationships**: Explored relationships between views, subscribers, video counts, and channel age.
   b. **Category Analysis**: Analyzed category-wise statistics, including subscriber counts, video views, and video counts.
   c. **Categories (2013 to 2018)** *:* We analyze how the distribution of categories has evolved over time by focusing on channels started between 2013 and 2018.
   d. **Linear Correlation**: Determined correlations between variables to understand their impact on each other.

ALGORITHM & DEPLOYMENT

d. Insights:

i. Views and subscribers exhibit a strong positive correlation.

ii. Video counts show no significant correlation with subscribers.

iii. Age of channels has a low positive correlation with subscribers.

5) Classification Analysis

a. Objective: Classify channels based on their category and rank. Investigate the top categories, newcomer categories, and channel maturity.

b. Classification Insights:

i. Identified top categories among the top 100 channels.

ii. Explored categories of the last 100 ranked channels.

iii. Analysed statistics based on channel age.

c. Findings:

i. Music, Entertainment, and People & Blogs are dominant categories among the top 100 channels.

ii. Gaming emerges as a popular category among the last 100 ranked channels.

iii. Channel maturity suggests a positive correlation with subscriber count.

RESULT


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 7 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   rank         1000 non-null   int64
 1   Youtuber     1000 non-null   object
 2   subscribers  1000 non-null   object
 3   video views  1000 non-null   object
 4   video count  1000 non-null   object
 5   category     973 non-null    object
 6   started      1000 non-null   int64
dtypes: int64(2), object(5)
memory usage: 54.8+ KB
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 989 entries, 0 to 988
Data columns (total 8 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   rank         989 non-null    int64
 1   Youtuber     989 non-null    string
 2   subscribers  989 non-null    object
 3   video views  989 non-null    object
 4   video count  989 non-null    object
 5   category     989 non-null    string
 6   started      989 non-null    int64
 7   age_count    989 non-null    int64
dtypes: int64(3), object(3), string(2)
memory usage: 61.9+ KB
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 989 entries, 0 to 988
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   rank            989 non-null    int64
 1   Youtuber        989 non-null    string
 2   subscribers     989 non-null    int64
 3   video views     989 non-null    int64
 4   video count     989 non-null    int64
 5   category        989 non-null    string
 6   started         989 non-null    int64
 7   age_count       989 non-null    int64
 8   avg_subperyrs   989 non-null    int64
 9   avg_viewperyrs  989 non-null    int64
 10  avg_vidperyrs   989 non-null    int64
 11  avg_subpervid   989 non-null    int64
 12  avg_viewpervid  989 non-null    int64
dtypes: int64(11), string(2)
memory usage: 100.6 KB
```

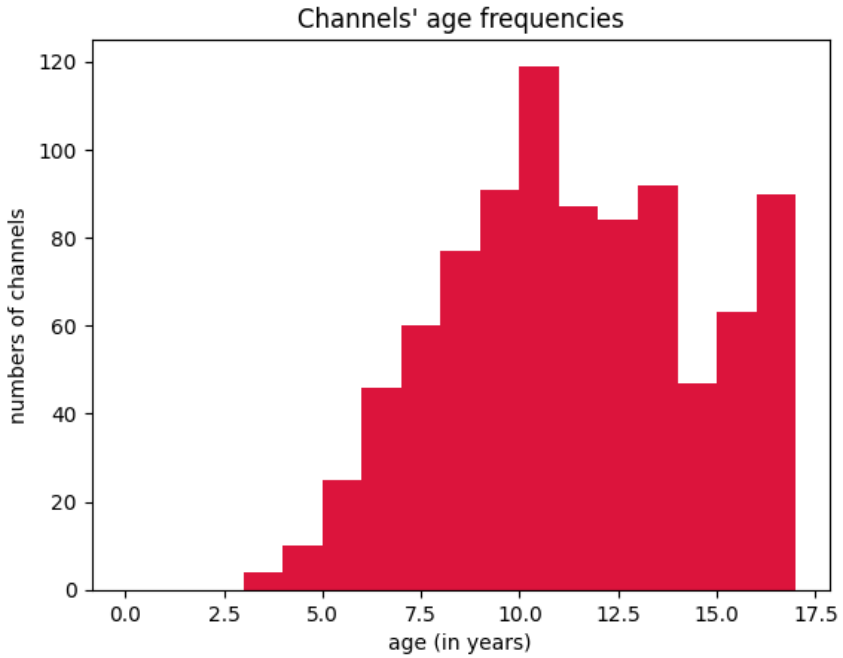
RESULT

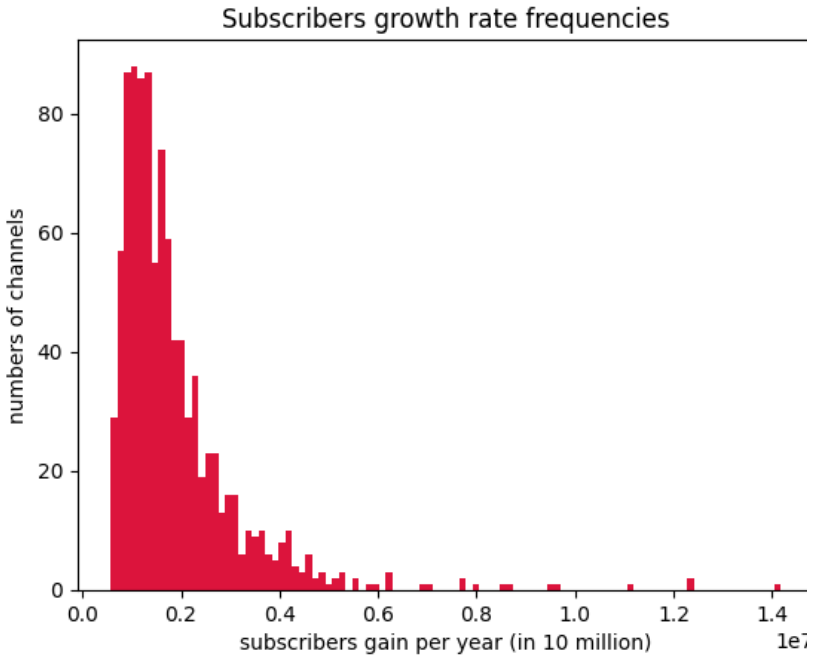
Statistical analysis

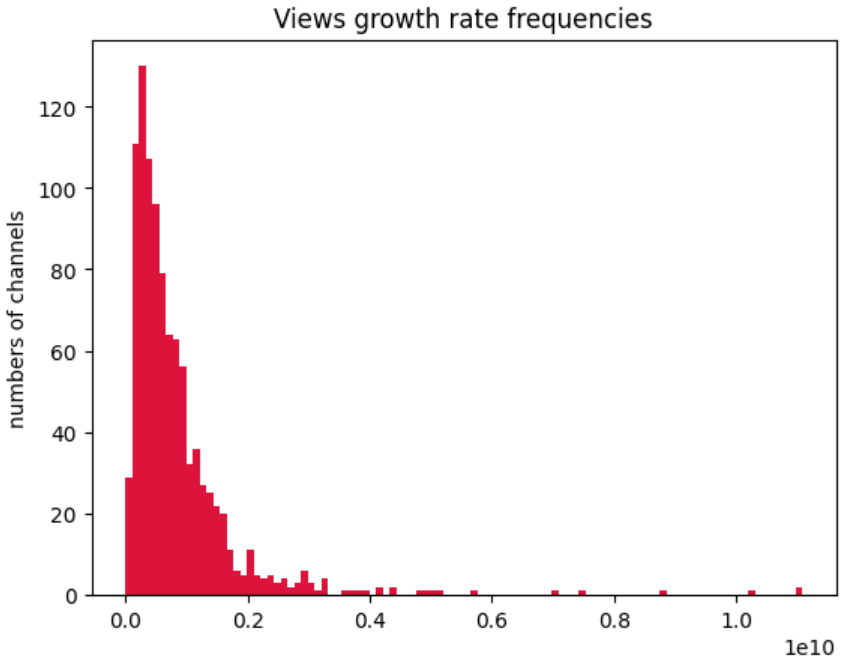
	subscribers	video views	video count
count	989.0	989.0	989.0
mean	20,166,228.5	9,293,213,552.1	8,645.0
std	14,683,134.0	12,102,173,890.6	29,509.3
min	10,900,000.0	439,098.0	1.0
25%	12,600,000.0	3,657,526,529.0	350.0
50%	15,500,000.0	6,165,983,944.0	872.0
75%	21,900,000.0	11,458,130,113.0	3,095.0
max	222,000,000.0	198,459,090,822.0	329,711.0

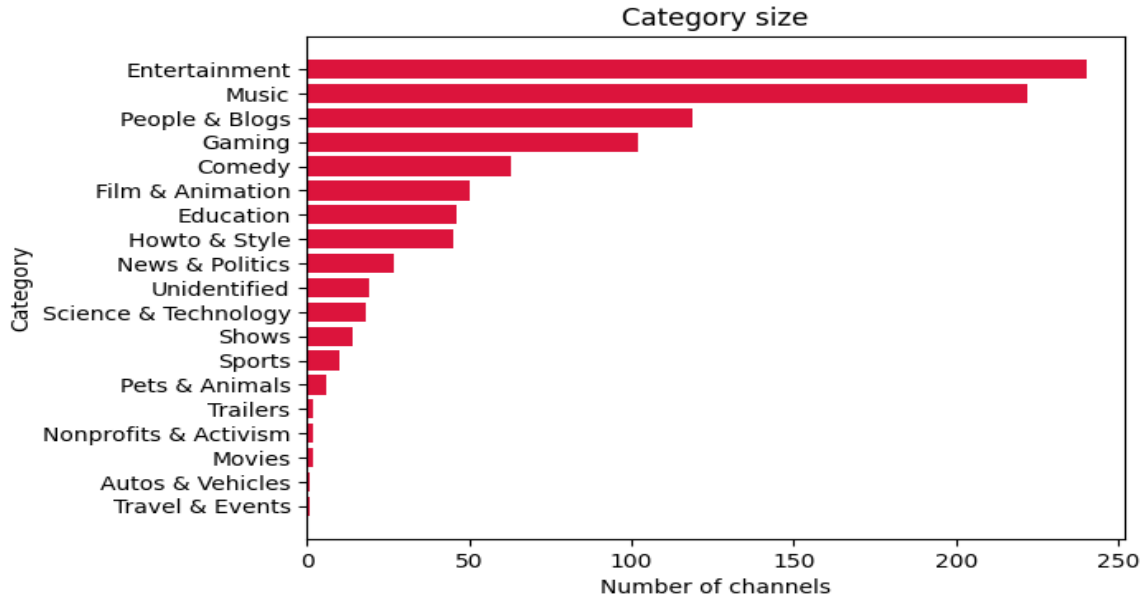
	age_count	avg_subperyrs	avg_viewperyrs	avg_vidperyrs
count	989.0	989.0	989.0	989.0
mean	11.6	1,899,850.6	846,700,447.7	710.9
std	3.8	1,343,423.1	996,074,197.6	2,467.5
min	3.0	578,947.0	27,444.0	0.0
25%	9.0	1,100,000.0	312,339,820.0	31.0
50%	11.0	1,538,462.0	583,578,992.0	89.0
75%	14.0	2,233,333.0	1,027,596,084.0	261.0
max	19.0	14,183,333.0	11,083,127,110.0	32,323.0

RESULT



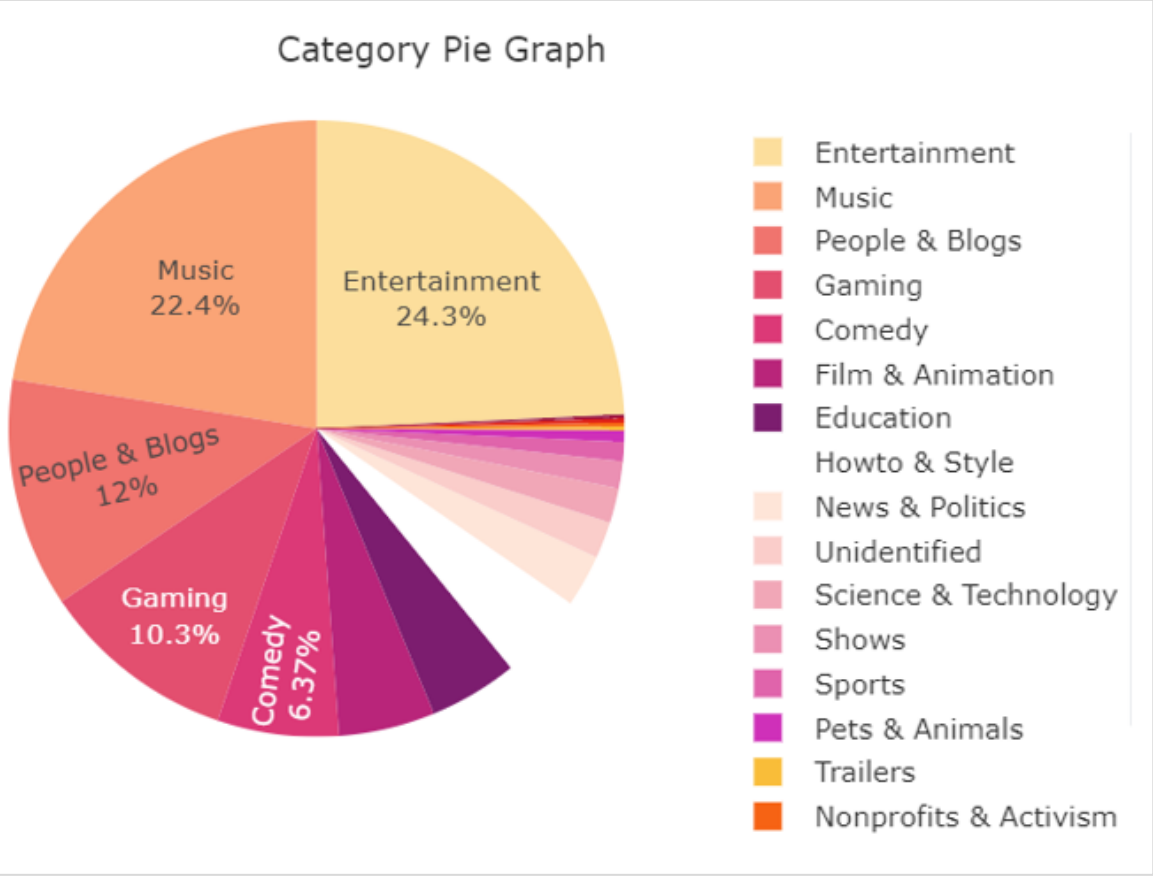


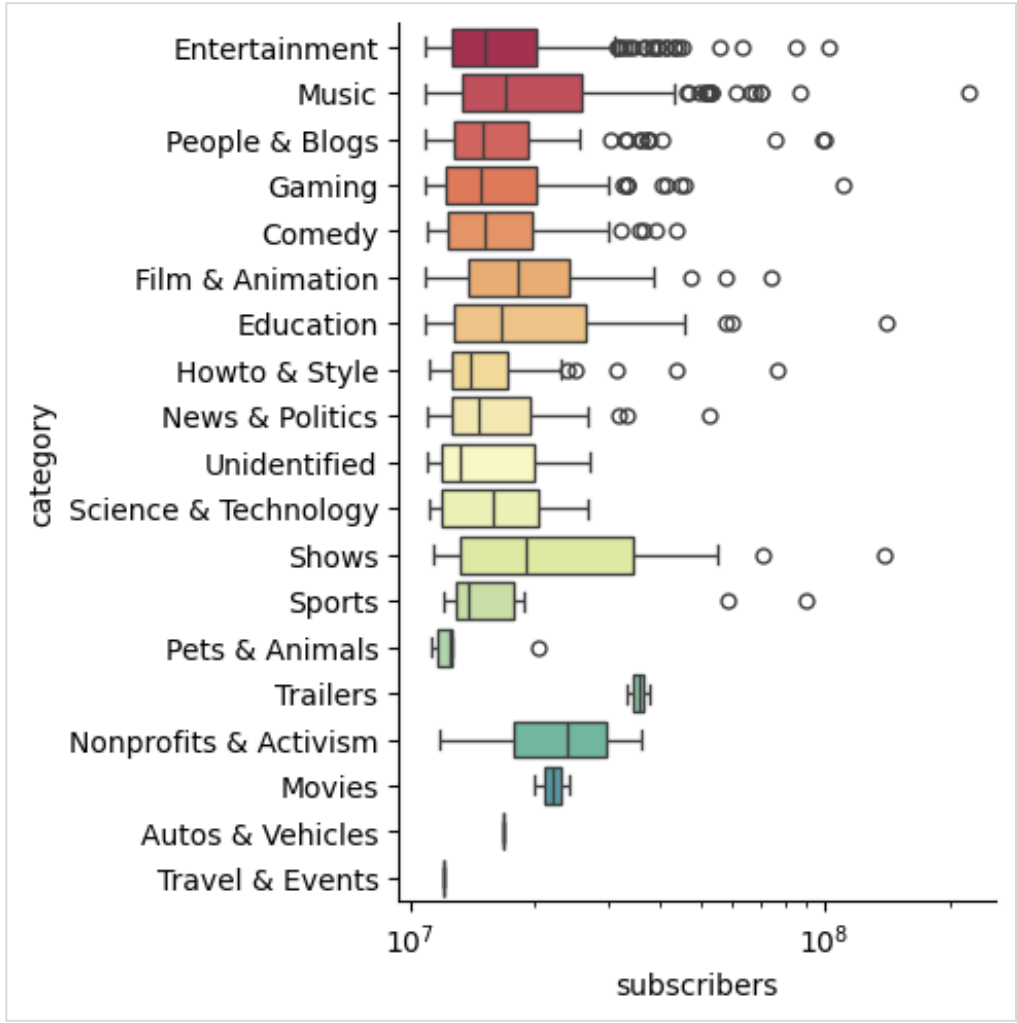




edunet foundation

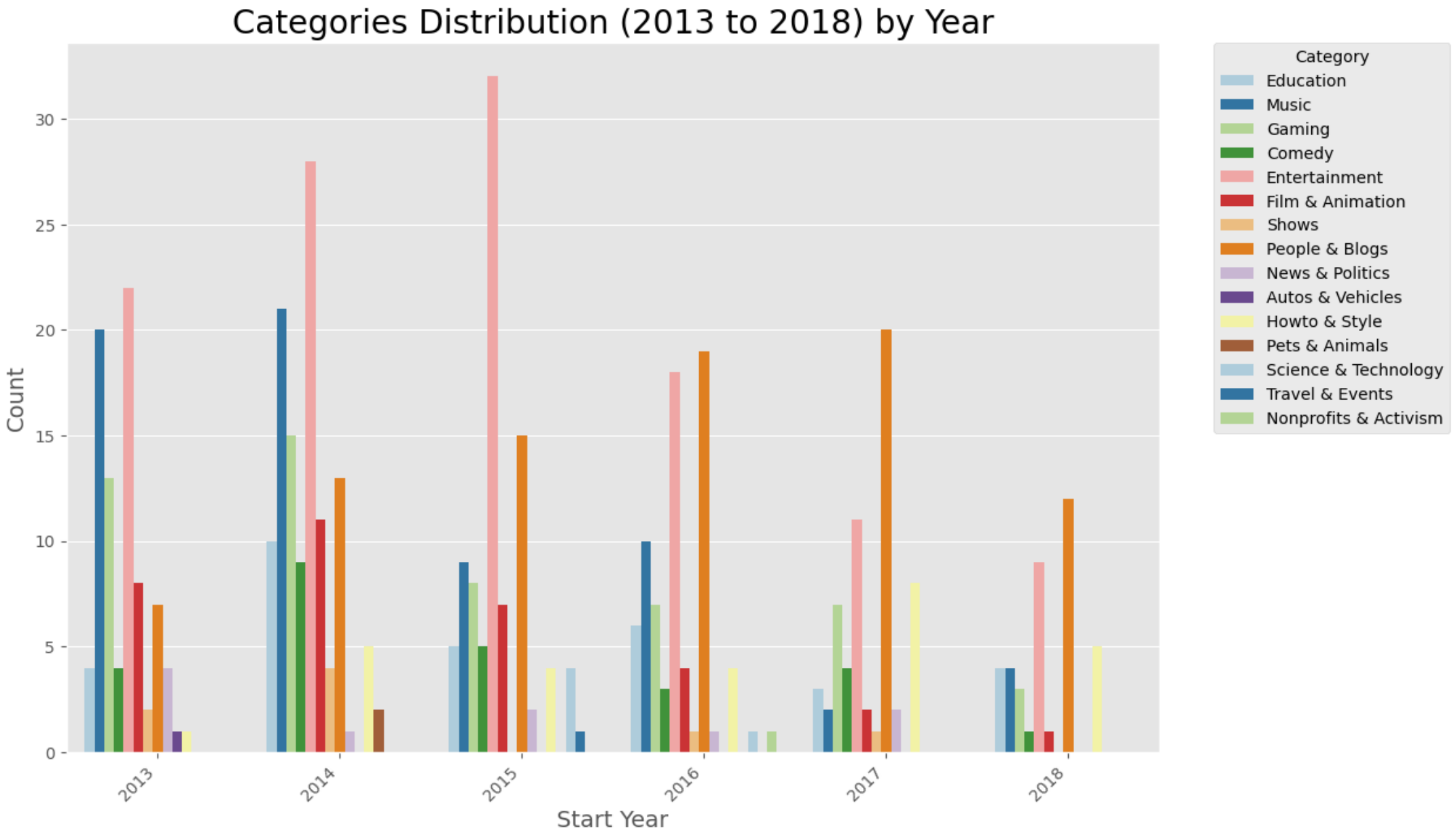
RESULT





edunet foundation

RESULT



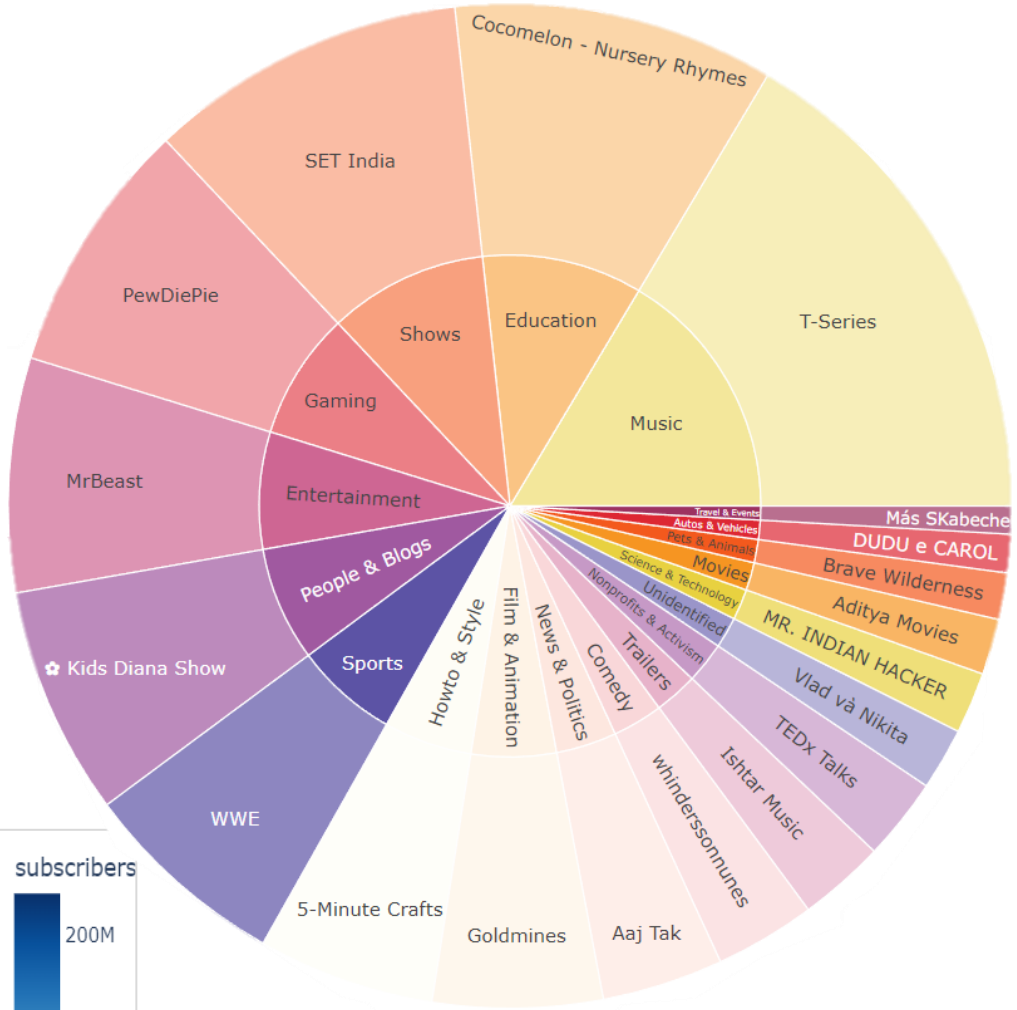
edunet foundation

RESULT

Youtuber with Most Subscribers in Each Category

Represented by Sunburst and Treemap

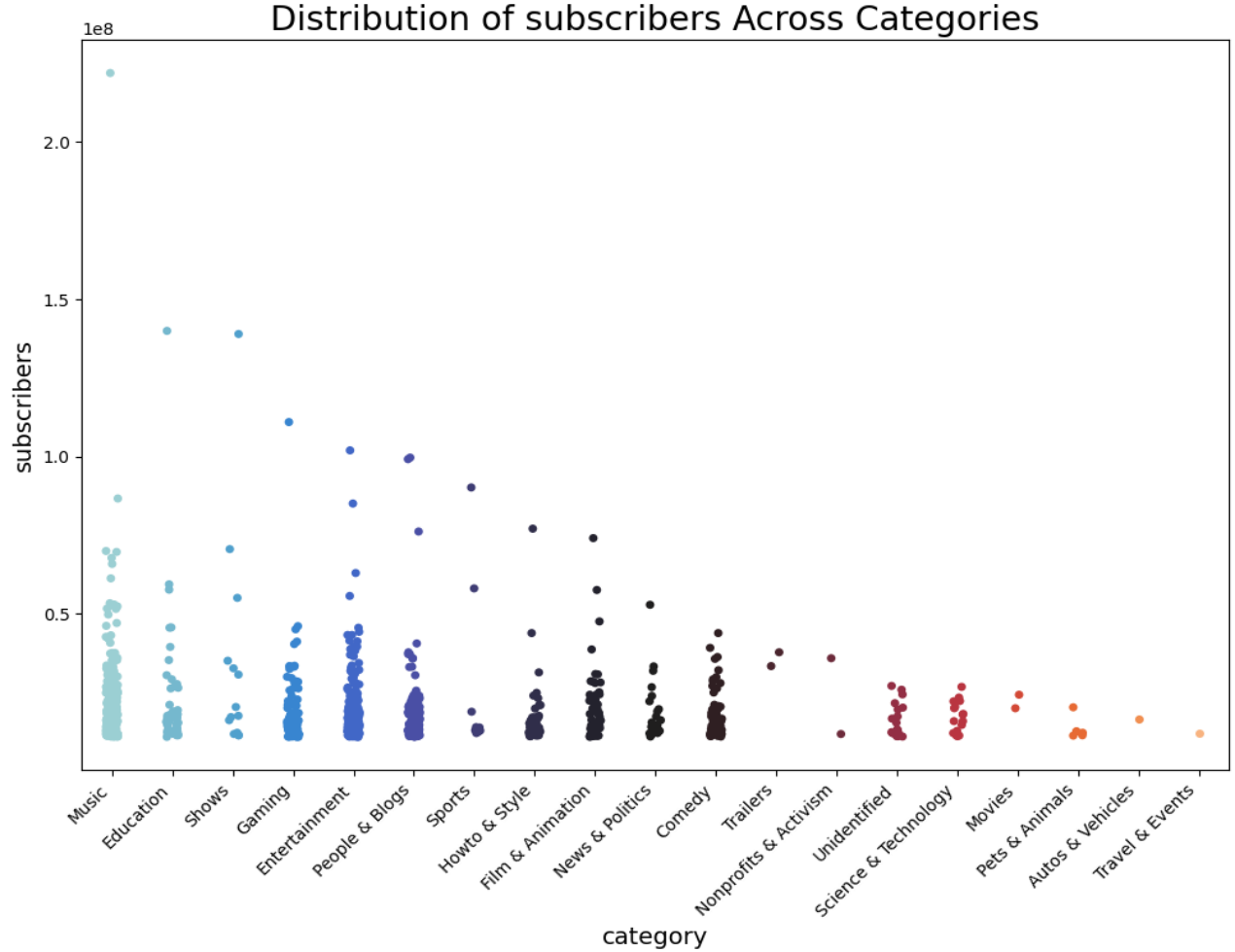
Sunburst

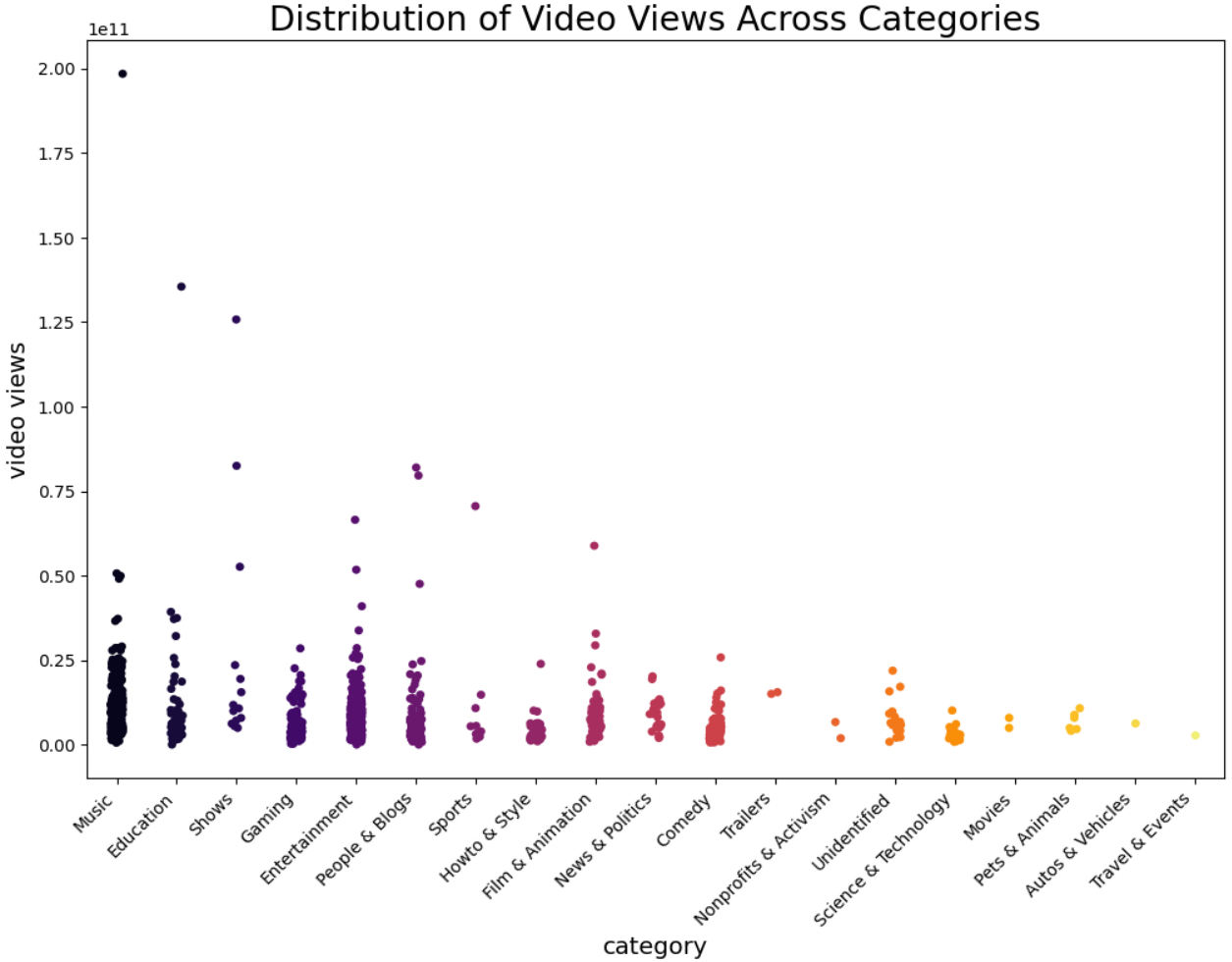


Treemap

edunet foundation

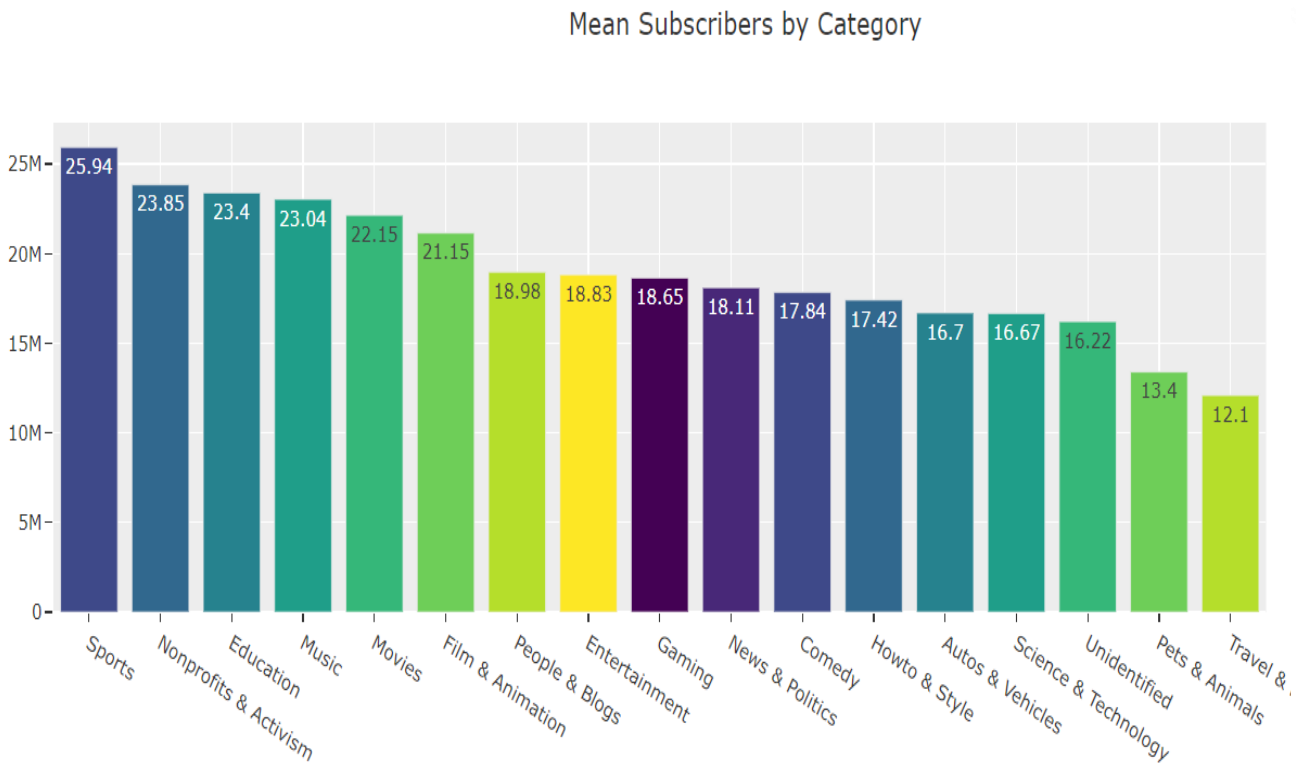
RESULT

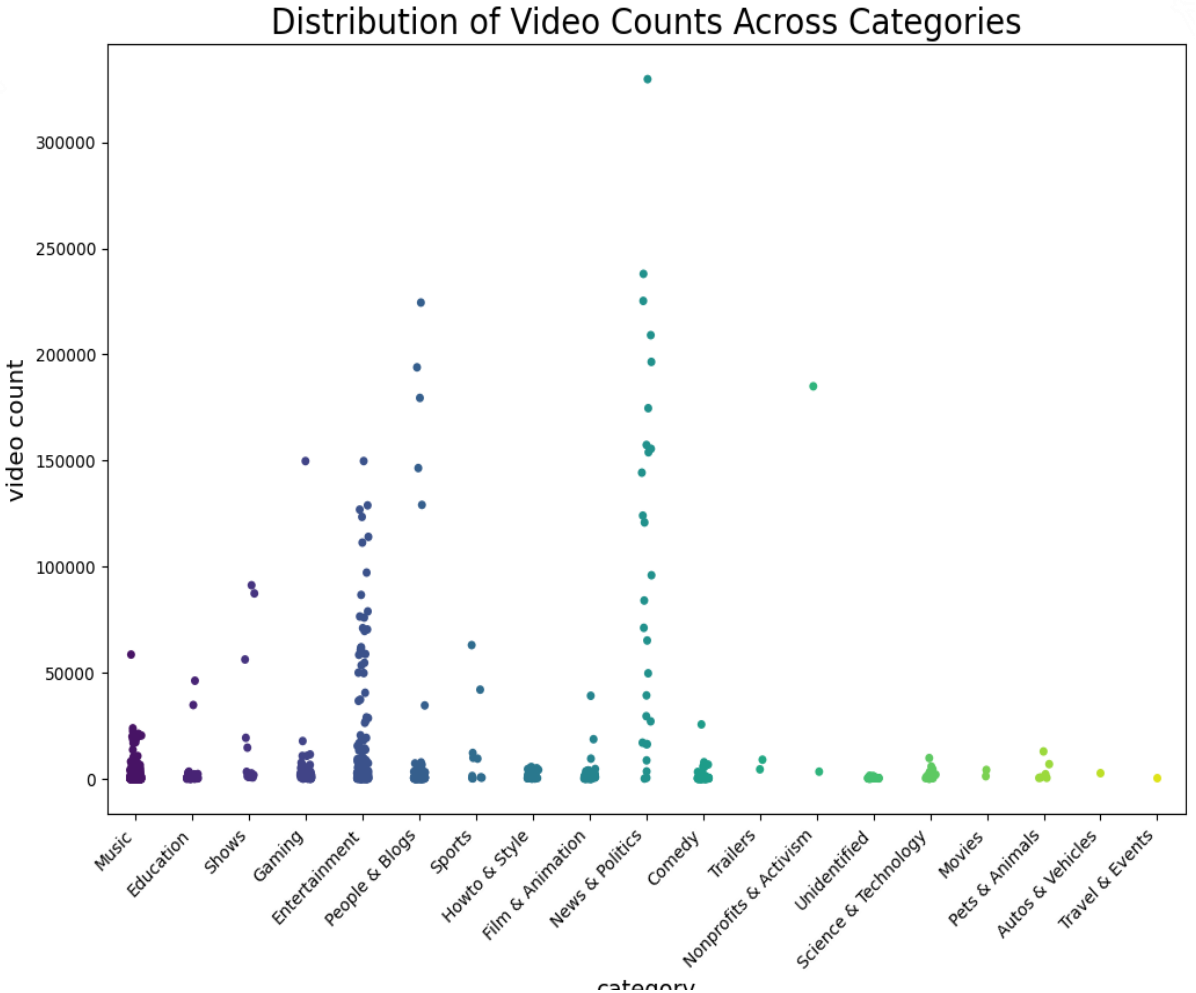




edunet foundation

RESULT

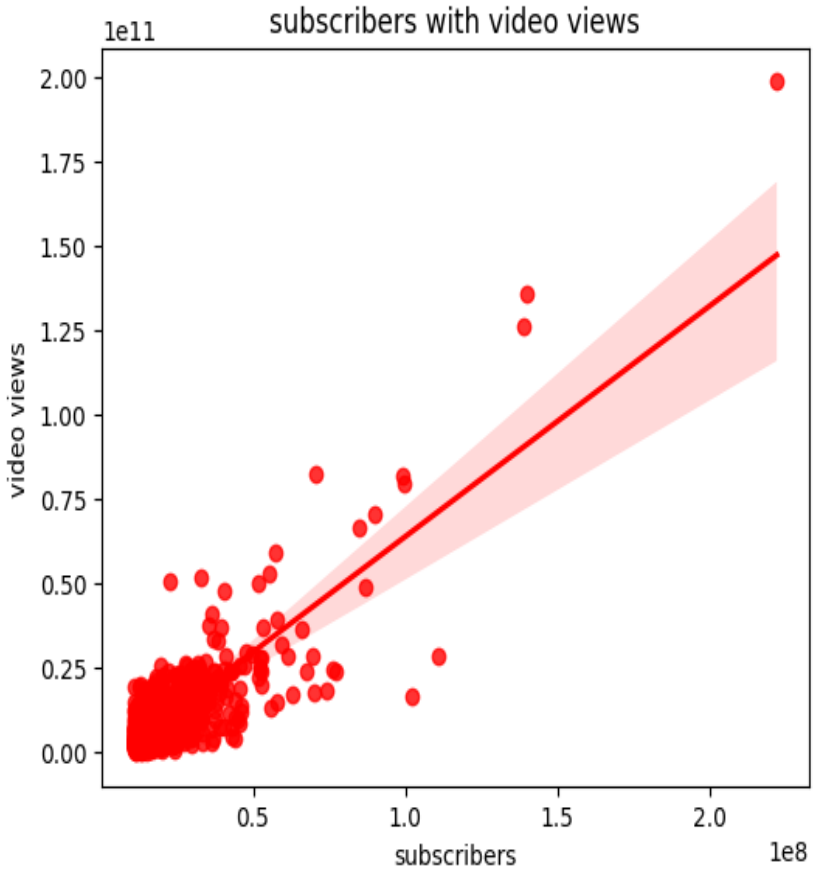


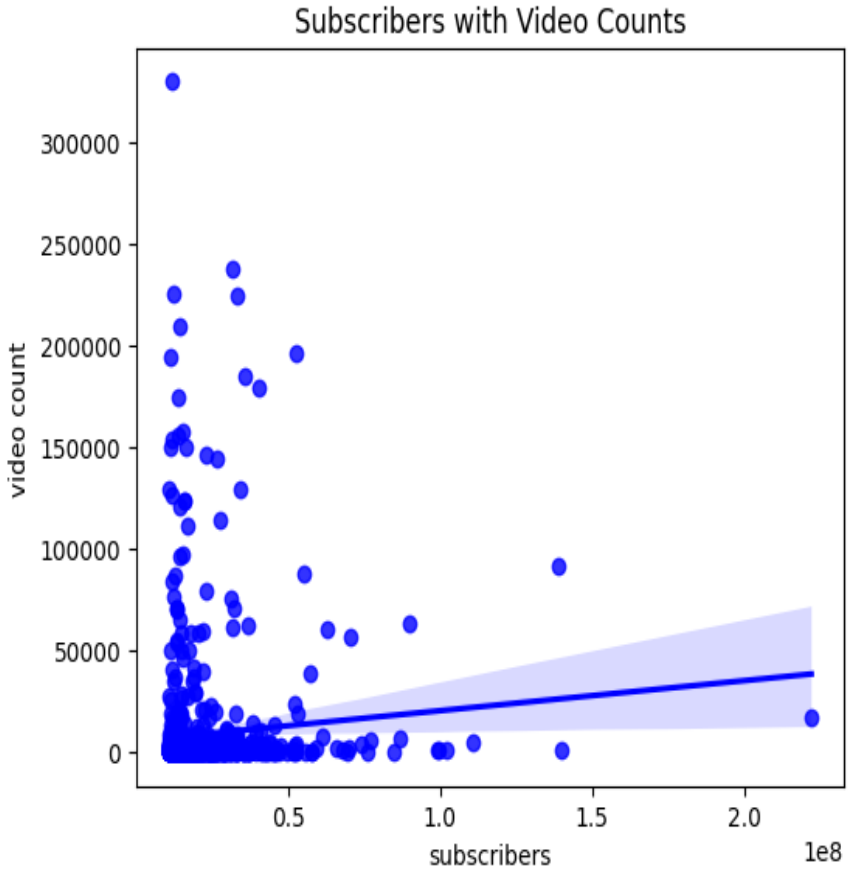


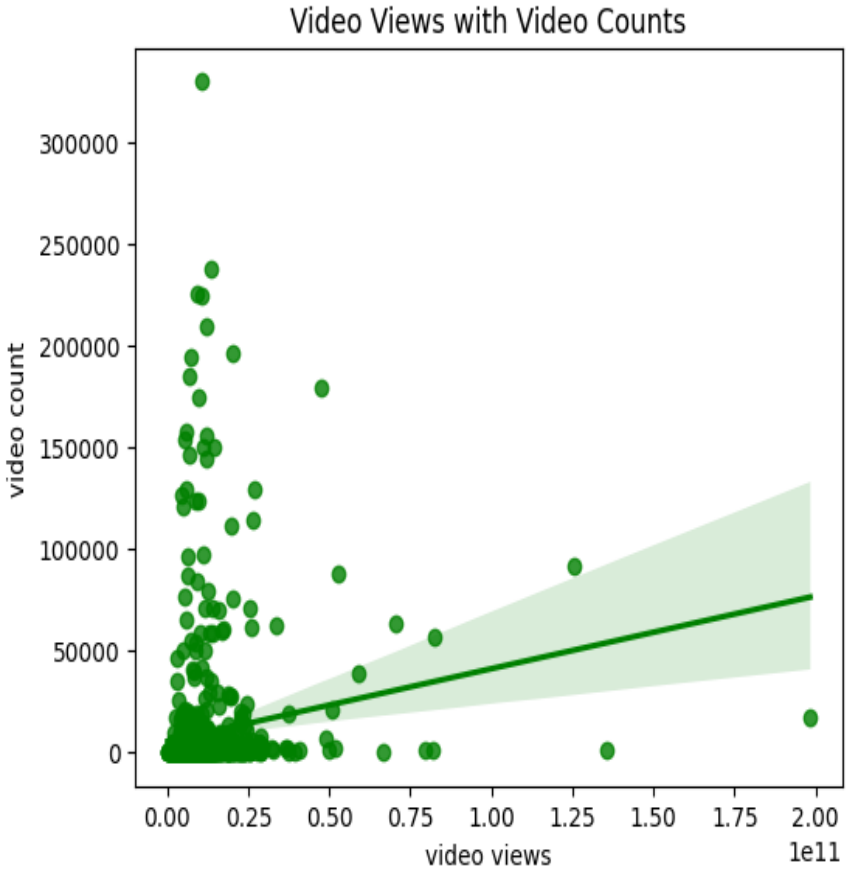
edunet foundation

RESULT

Regression plot on columns 'Subscribers', 'Video Views' and 'Video Counts'.





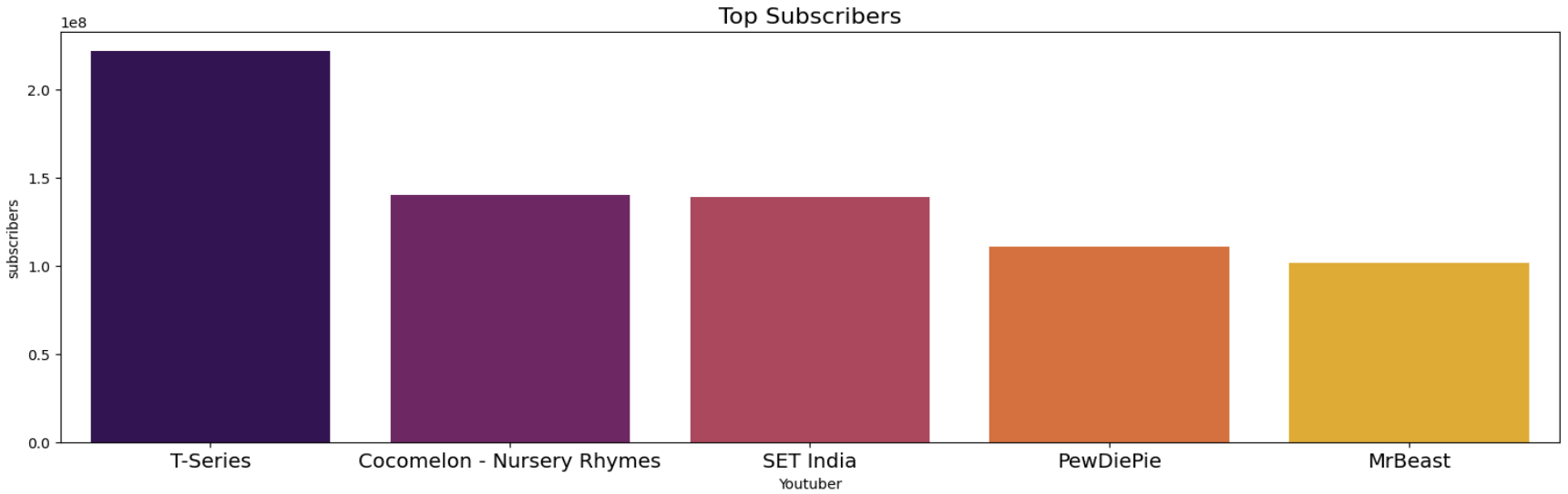


edunet foundation

RESULT

Top 5 YouTubers Analysis

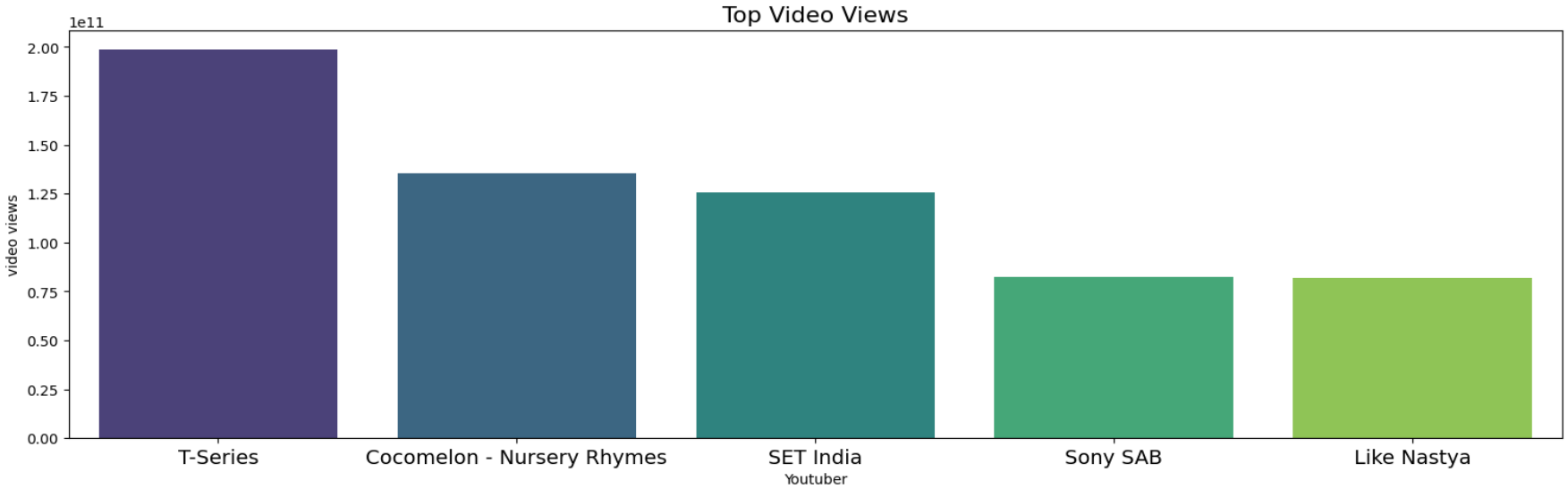
Based on Subscribers		
Rank	**Youtuber**	**subscribers**
0	T-Series	222000000
1	YouTube Movies	154000000
2	Cocomelon - Nursery Rhymes	140000000
3	SET India	139000000
4	Music	116000000



RESULT

Top 5 YouTubers Analysis

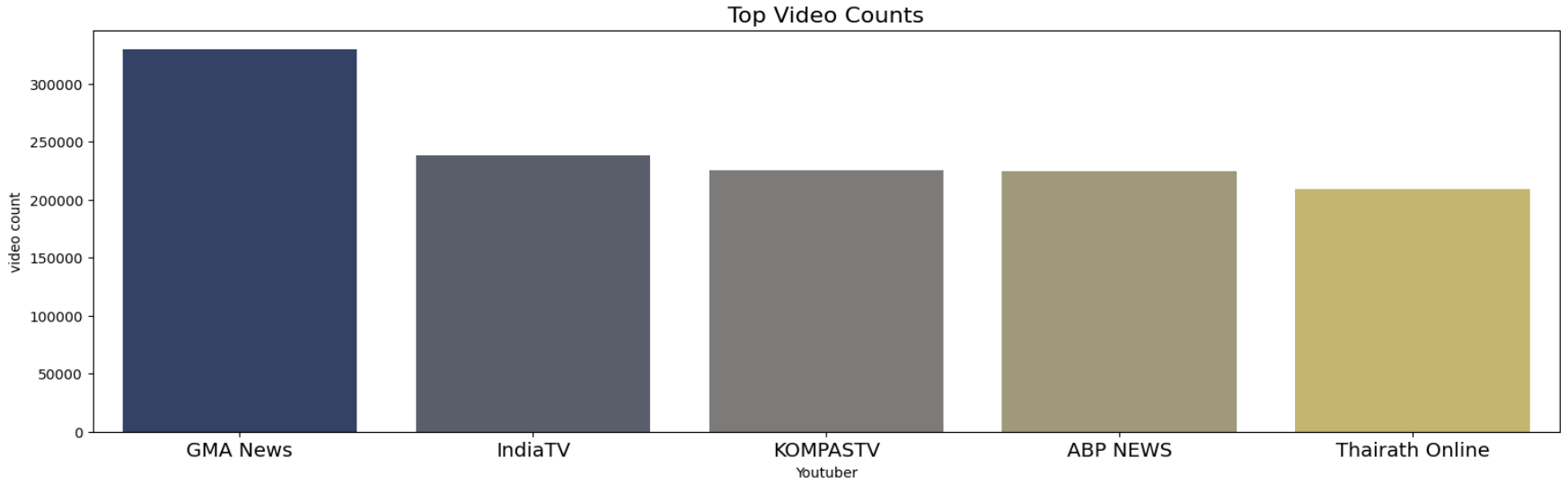
Based on Video Views		
Rank	**Youtuber**	**video views**
0	T-Series	19845909082
2	Cocomelon - Nursery Rhymes	135481339848
3	SET India	125764252686
17	Sony SAB	82473581441
8	Like Nastya	81963845811



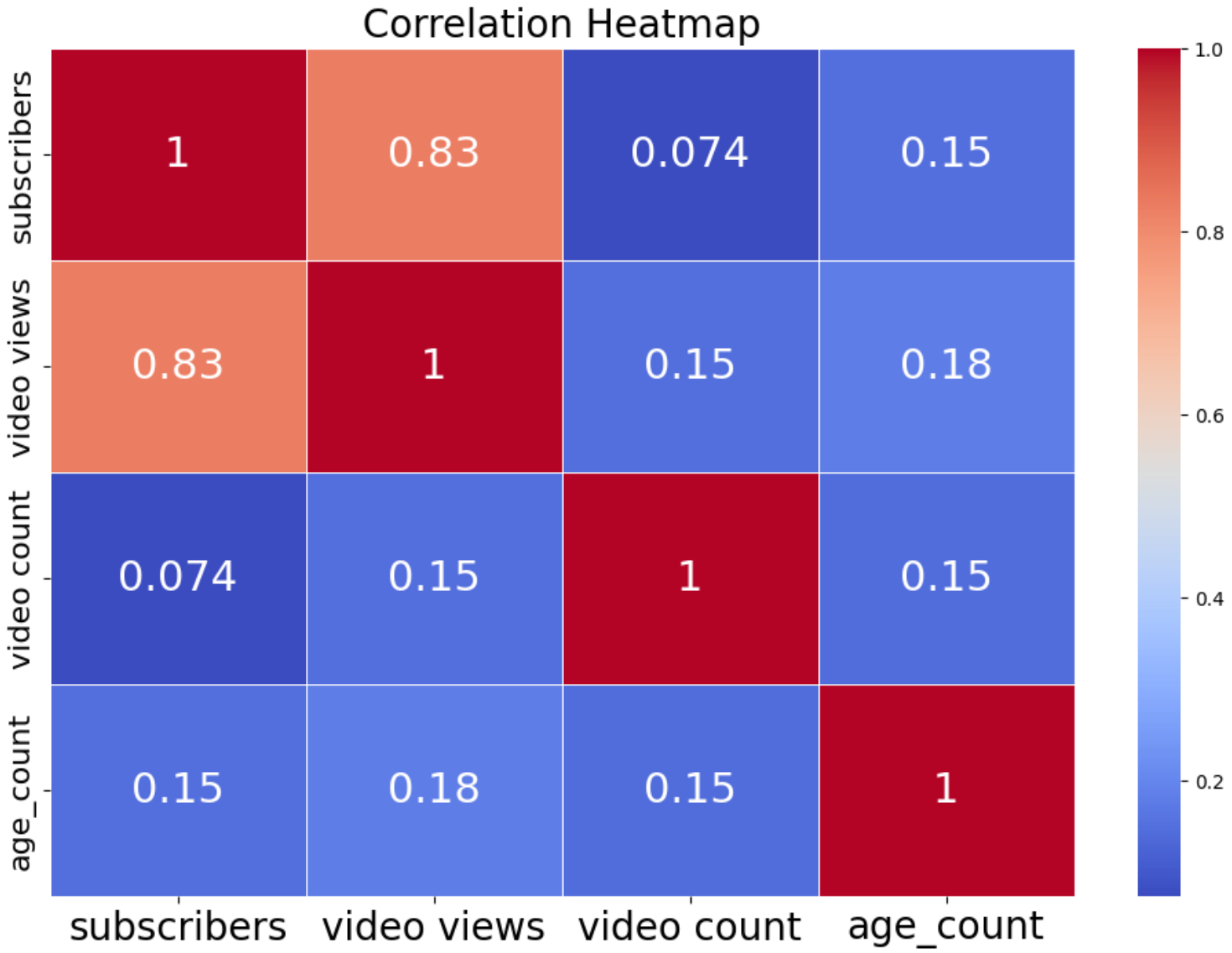
RESULT

Top 5 YouTubers Analysis

Based on Video Counts		
Rank	**Youtuber**	**video counts**
815	GMA News	329711
112	IndiaTV	237971
777	KOMPASTV	225232
100	ABP NEWS	224455
576	Thairath Online	209097



RESULT



CONCLUSION

In conclusion, this analysis provided valuable insights into the YouTube top subscribed channels ecosystem. Through thorough data exploration, cleaning, and analysis, we gained a deeper understanding of channel dynamics, category trends, and subscriber behaviors. The project successfully addressed key questions, such as the relationship between views and subscribers, the impact of video counts on channel success, and the significance of channel maturity. By examining classification insights, we identified dominant categories among the top channels, observed trends among newcomer channels, and analyzed the growth patterns of channels based on their age.

Overall, this project serves as a comprehensive guide for understanding the YouTube content landscape, offering valuable information for brands, content creators, and aspiring YouTubers to inform their strategies and decision-making processes.

CONCLUSION

Notable Findings from EDA:

- **Skewed Distribution:** Subscribers, views, and video counts exhibit skewed distributions.
- **Category Dominance:** Music, Entertainment, People & Blogs, and Gaming are dominant categories.
- **Subscriber-View Relationship:** Strong positive correlation between subscribers and views.
- **Channel Maturity Impact:** Established channels dominate top ranks; newcomer channels face challenges.
- **Content Diversity:** Wide range of content categories indicate a diverse content landscape.
- **Variable Relationships:** Views and subscribers positively correlated; video counts show no significant correlation.
- **Platform Evolution:** Shifts in category trends and channel maturity over time.
- **Content Popularity:** Top channels command substantial viewership and subscriber base.

FUTURE SCOPE

- Future enhancements may include:

a) Perform **sentiment analysis** on comments and engagement metrics to understand audience feedback and sentiment towards content.

b) Implement **machine learning** models for **predictive analytics**, such as forecasting subscriber growth or predicting video virality.

c) Conduct **comparative analysis** with competitor channels to identify strengths, weaknesses, and opportunities for improvement.

d) Market **Trend Analysis**: Analyze market trends and emerging content categories for strategic content creation.

e) Social Network Analysis: Conduct social network analysis to understand subscriber interactions and influencer networks.

REFERENCES

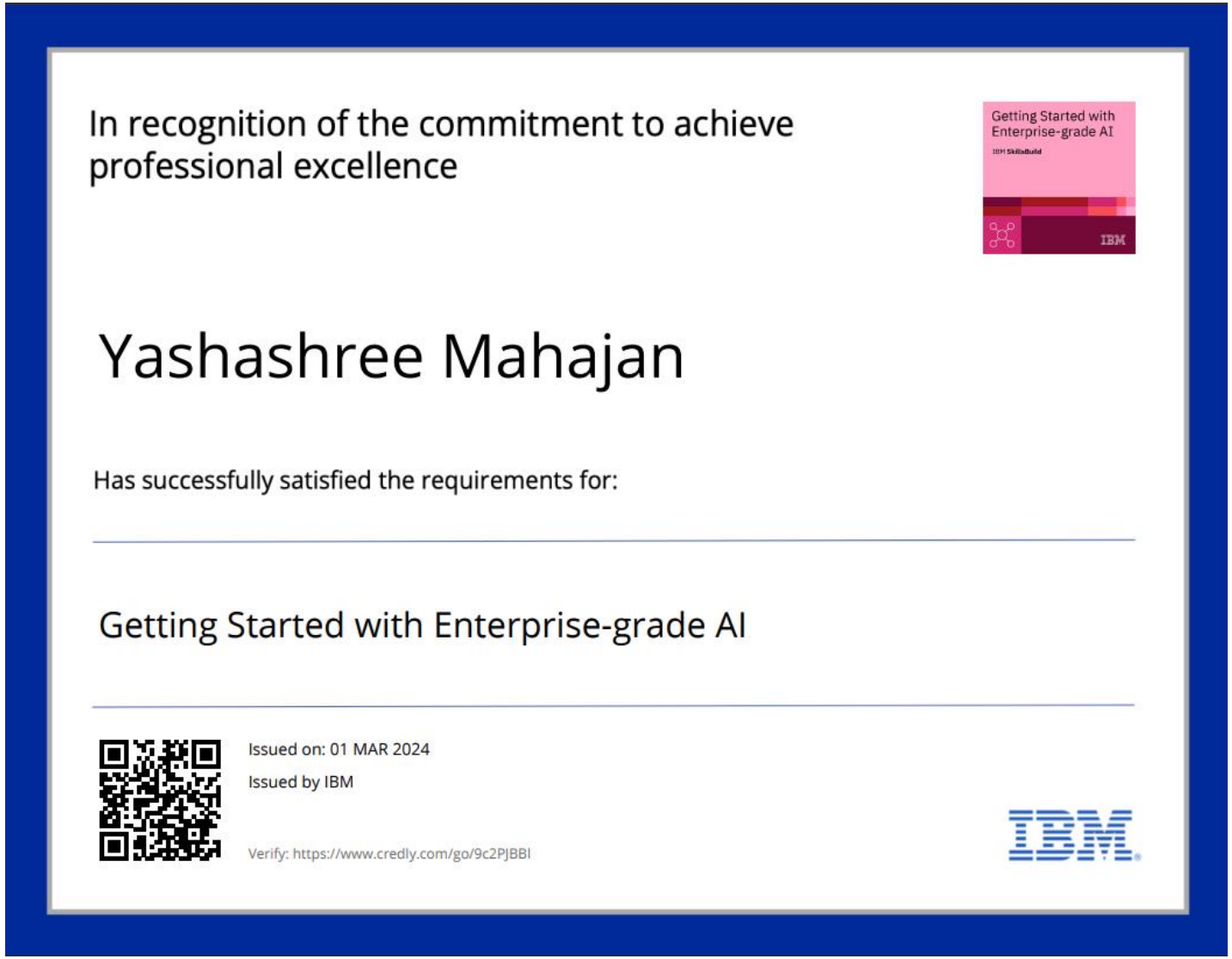
- https://pandas.pydata.org/docs/
- https://matplotlib.org/stable/users/explain/quick_start.html
- https://seaborn.pydata.org/tutorial/introduction.html
- https://www.kaggle.com/datasets/surajjha101/top-youtube-channels-data

COURSE CERTIFICATE 1

https://www.credly.com/badges/a0b9105f-7260-4cae-bc38-e0b53c202218/public_url

edunet foundation

COURSE CERTIFICATE 2



https://www.credly.com/badges/3dccedb4-41e2-460a-822f-05fd539fbad9/public_url

edunet foundation

COURSE CERTIFICATE 3

In recognition of the commitment to achieve professional excellence

Journey to Cloud: Envisioning Your Solution

IBM SkillsBuild

IBM

Yashashree Mahajan

Has successfully satisfied the requirements for:

Journey to Cloud: Envisioning Your Solution

Issued on: 01 MAR 2024

Issued by IBM

Verify: https://www.credly.com/go/6ncqIPzP

IBM

https://www.credly.com/badges/bbb12629-e6d6-446d-9879-a6b2865694a9/public_url

edunet foundation

THANK YOU

Yashashree Ravindra Mahajan
PCET's Nutan Maharashtra Institute of Engineering & Technology,Pune
IBM Skillsbuild Edunet foundation and AICTE (AI and Cloud)
SB4C-AICTE_Batch 4 G3
yashshreem2003@gmail.com

edunet foundation