Youtube data analysis

A Mini-Project Report Submitted
For
Partial Fulfillment of the Requirements of the
Degree of Bachelor of Engineering

In

COMPUTER ENGINEERING

(Semester VII)
By

Aditya Vyas 9238

Hitesh Sharma 9233

Atharva Pawar 9427

Under the guidance of

Prof. Ankita Amburle

DEPARTMENT OF COMPUTER ENGINEERING
Fr. Conceicao Rodrigues College of Engineering
Bandra (W), Mumbai - 400050
University of Mumbai

2023-2024

This work is dedicated to my family.

I am very thankful for their motivation and support.

CERTIFICATE

This is to certify that the mini-project entitled **"Youtube Data Analysis"** is a bonafide work of " Aditya Vyas (9238)**,** Hitesh Sharma (9233), Atharva Pawar (9427) **"** submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of **Bachelor of Engineering** in **Computer Engineering** (Semester-VII).

Prof. Ankita Amburle
(Name and Sign)

Guide/ Supervisor

Dr. Sujata Deshmukh

Dr. S. S. Rathod

Approval Sheet

Mini Project Report Approval for B.E. (Semester-VII)

This mini-project report entitled Youtube Data Analysis submitted by Aditya Vyas (9238)**,** Hitesh Sharma (9233), Atharva Pawar (9427) is approved for the degree of Bachelor of Engineering in **Computer Engineering** (Semester-VII).

Examiner 1._____

Examiner 2._____

Date:

Place:

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Date:

Aditya Vyas (9238)

Hitesh Sharma (9233)

Atharva Pawar (9427)

Abstract

With the exponential growth of online video content, the need for efficient video summarization techniques has become increasingly evident. This paper presents a web application that leverages state-of-the-art deep learning methods to automatically generate textual summaries of videos. The application offers a user-friendly interface for uploading videos, and within seconds, it delivers concise, coherent, and contextually relevant textual summaries that encapsulate the essential content of the video.

We employ cutting-edge pre-trained models and transfer learning to enhance the system's performance, making it adaptable to a wide range of video genres and languages..

The system's capability to capture key events, sentiments, and themes in the video content makes it valuable for a broad spectrum of applications, including content curation, video indexing, and accessibility for individuals with visual impairments.The application's user-friendly design and seamless integration with web-based platforms make it accessible to a wide audience.

This paper showcases the development and evaluation of our web application for automatic video summarization, highlighting its potential to simplify video content consumption and promote a more efficient and accessible digital media landscape.

Keywords:

Video Summarization, accessibility., web application, automatic summarization.

Acknowledgments

We have great pleasure in presenting the report on “Youtube data analysis”. I take this opportunity to express my sincere thanks towards the guide Prof. Ankita Amburle, C.R.C.E, Bandra (W), Mumbai, for providing the technical guidelines, and the suggestions regarding the line of this work. We enjoyed discussing the work progress with him/her during our visits to the department.

We thank Dr. Sujata Deshmukh, Head of Computer Engineering department, Principal and the management of C.R.C.E., Mumbai for encouragement and providing necessary infrastructure for pursuing the project.

We also thank all non-teaching staff for their valuable support, to complete our project.

Date:

Aditya Vyas (9238)

Hitesh Sharma (9233)

Atharva Pawar (9427)

Table of Content

Chapter 1

INTRODUCTION

YouTube has become an integral part of our digital lives, where a myriad of videos entertain, educate, and engage an ever-expanding audience. In this report, we present the results of our individual analysis of a YouTube dataset obtained from Kaggle, exploring fascinating facets of the platform's content and performance. We dive deep into the data to uncover what makes certain videos rise to the top, which categories dominate the platform, and other noteworthy insights.

This report serves as a record of our exploration, revealing valuable patterns, trends, and findings derived from the data. Each analysis, conducted separately, offers a unique perspective on the dataset, ranging from the most-liked videos to the top categories, and much more.

Our intended readers encompass fellow data enthusiasts, data analysts, and anyone intrigued by the data-driven exploration of YouTube content. As we progress through the report, we hope our analyses provide a deeper understanding of the intriguing world of YouTube videos and their popularity.

Thank you for embarking on this analytical journey with us, as we unravel the stories hidden within this captivating dataset.

Chapter 2

OBJECTIVES OF THE PROJECT

Our project has following objectives:

1. **Identify the Most Liked Videos**: Determine which YouTube videos have received the highest number of likes and explore the characteristics or themes that contribute to their popularity.

2. **Discover the Most Viewed Videos**: Find and analyze the videos with the highest view counts, investigating what sets them apart from others in terms of content, promotion, or other factors.

3. **Analyze the Top Categories**: Examine the distribution of videos across different categories or genres and identify the most prevalent and successful categories on the platform.

4. **Explore Video Length vs. Engagement**: Investigate whether there is a correlation between the duration of a video and the level of audience engagement, such as likes, comments, and views.

Chapter 3

SCOPE OF THE PROJECT

Following points summarize the scope of our project :

1. **Dataset Coverage**: Your analysis will focus on the YouTube dataset obtained from Kaggle. This dataset includes information about video titles, descriptions, categories, view counts, likes, dislikes, comments, and more.

2. **Temporal Scope**: The analysis covers data within a specific time frame, if applicable. This helps provide context, especially when analyzing trending videos.

3. **Data Preprocessing:** The report will include details on data preprocessing steps, such as handling missing data, data cleaning, and data transformation.

4. **Video Popularity Metrics:** The report will analyze video popularity metrics, such as the most liked videos, most viewed videos, and the relationship between likes, comments, and views.

5. **Category Analysis:** A specific focus will be given to analyzing the distribution of videos across categories and identifying the top-performing categories.

Chapter 4

REVIEW OF LITERATURE

Sr. No.	Paper Title	Citation	Database	Summary
1	YouTube Dataset on Mobile Streaming for Internet Traffic Modeling and Streaming Analysis	6	Nature	This dataset provides 1,081 hours of mobile video streaming measurements using YouTube's native client on mobile devices, with a focus on network, transport, and application layers. It covers 80 network scenarios, 171 bandwidth settings, and 332 GB of video data, making it a valuable resource for researching mobile streaming quality and performance. This dataset is particularly relevant given the dominant role of online video, with three out of five views originating from mobile devices worldwide.
2.	YouTube as a source of patient information for Coronavirus Disease (COVID-19): A content-quality and audience engagement analysis	40	Wiley	This study evaluated YouTube videos related to the SARS-CoV-2 virus during the early phase of the pandemic, finding that the quality of these videos was generally poor, with a mean DISCERN score of 31.33 out of 75. However, some top-quality videos were identified and highlighted as potential resources for patient education during the COVID-19 pandemic..
3.	Topics and destinations in comments on YouTube tourism videos during the Covid-19 pandemic	2	plos	This study examined comments on tourism-related YouTube videos during the Covid-19 pandemic, revealing discussions focused on topics like people, countries, tourists, and Covid-19, reflecting concerns about the pandemic's impact on tourism. The destinations frequently mentioned included India, Nepal, China, Kerala, France, Thailand, and Europe. The research provides theoretical insights into changing tourist perceptions and practical implications for developing safety and sustainable measures for the tourism industry during a pandemic.

Chapter 5

PROPOSED SYSTEM

5.1 PROBLEM STATEMENT

In the digital age, YouTube has grown to be a vast and dynamic platform, hosting a diverse array of video content from creators around the world. With over 2 billion logged-in monthly users, YouTube has become a powerhouse of content consumption and creation. However, within this expansive landscape lie several challenges and unanswered questions:

1. Accessibility: Existing YouTube analytics tools can be complex for non-technical users. There is a need for a user-friendly interface for accessing valuable insights.

2. Localization: Users often require country-specific data to make informed decisions regarding content creation, marketing, and audience engagement.

3. Content Creator Assistance: Content creators lack an easy way to identify top-performing videos and channels to refine their content strategies.

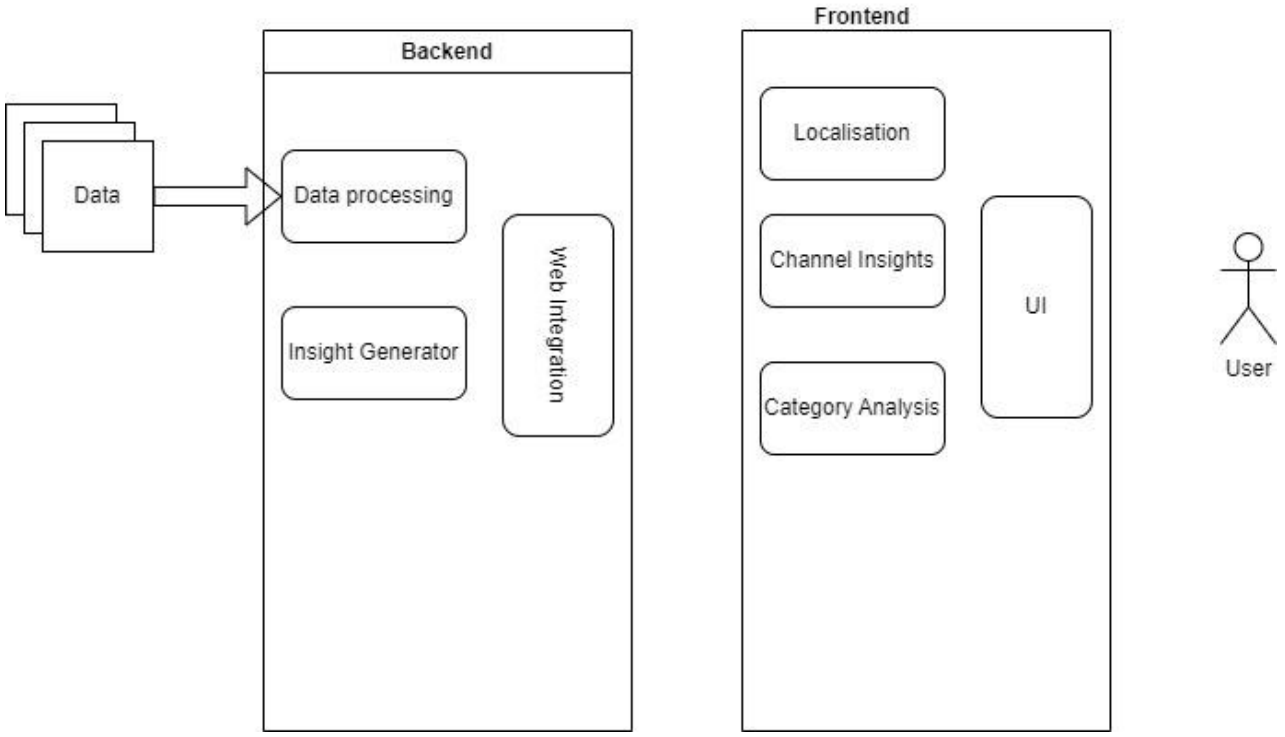
4. Marketer's Needs: Marketers need localized data to refine advertising strategies on YouTube.

5. Research and R Practice: In addition to practical insights, the project aimed to serve as a platform for honing research and data analysis skills, using R as a primary tool.

Chapter 6

SYSTEM DESIGN

6.1 MODULE DESIGN



In this chapter, we delve into the architectural design of the YouTube Data Analysis system, outlining the key components and modules that make up the system. The system design is essential for understanding how the platform functions and how it provides users with valuable insights into YouTube data.

6.2 MODULE DESCRIPTION

6.1 Frontend

The frontend of the YouTube Data Analysis system is the user-facing component responsible for providing an interactive and accessible interface for users to explore YouTube data.

Module 6.1.1: User Interface

Description: The User Interface module focuses on creating a user-friendly and visually appealing web interface. It encompasses web page design, layout, navigation, and elements for user interaction.

Module 6.1.2: Localization

Description: The Localization module enables users to select a specific country for analysis. It ensures that the displayed data and insights are tailored to the chosen region, providing relevant and localized information.

Module 6.1.3: Category Analysis

Description: The Category Analysis module allows users to explore top-performing video categories. It provides insights into the most popular content genres on the YouTube platform.

Module 6.1.4: Channel Insights

Description: The Channel Insights module empowers users to explore individual channels in depth. It offers detailed information about channel performance, including the most liked video, most viewed video, impressions data, and more.

6.2 Backend

The backend of the YouTube Data Analysis system handles data processing, analysis, and serves insights to the frontend. It forms the core of the platform.6

Module 6.2.1: Data Processing Module: This module focuses on data cleaning, transformation, and preparation, including handling missing data, data normalization, and ensuring data consistency.

Module 6.2.2: Insights Generater Module: This module derives meaningful insights from the data, identifying top-performing videos, channels, and trends contributing to video virality.

This chapter provides a detailed overview of the architectural design of the YouTube Data Analysis system, highlighting the role and functionality of each module within the frontend and backend components.

6.3 SOFTWARE AND HARDWARE USED

Software:

Operating System	Microsoft Windows 7 or higher/ Similar Linux System
Tool	Appropriate IDE like VS Code, Android Studio etc
Coding language	JavaScript , HTML, CSS, Django,R,Python
Storage	8 GB RAM and minimum 10 GB free space
Processor	Intel core I5 or higher
Other Requirements	Good Internet connectivity

Table 6.3 Software Requirements

Chapter 7

IMPLEMENTATION

We have taken our YouTube data analysis project a step further by developing a user-friendly website. This website allows users to select a specific country for analysis, providing a localized perspective on YouTube content.

Users can access a range of features, including:

1. **Top Performing Categories**: The website displays the most popular video categories for the chosen country.
2. **Top Channels**: Users can explore the top YouTube channels in the selected region.
3. **Channel Insights**: For specific channels, users gain access to valuable insights, such as the most liked video, most viewed video, impressions data, and the top-performing images.

This implementation enhances the practicality of our analysis, offering a hands-on experience for users to explore YouTube data with ease. By allowing customization and localization, our website empowers users to extract insights tailored to their specific interests or marketing strategies.

This user interface simplifies access to the wealth of data analyzed in our project, providing a user-friendly platform for data-driven decision-making in the dynamic world of online video content.

Chapter 8

RESULTS

In our analysis of YouTube video data from Kaggle, we've uncovered valuable insights, including identifying the most liked videos, top-performing video categories, and trends contributing to video virality. These findings offer practical value, empowering content creators with strategies to improve their videos and marketers with opportunities for more effective audience targeting on the platform.

As YouTube's landscape is dynamic and ever-evolving, we recognize the need for continued exploration. Future research can delve into specific aspects of content creation and audience engagement.

Despite limitations related to data constraints, we hope our exploration of YouTube data serves as an informative starting point and encourages further investigations in this digital realm.

8.1 WebApp

BDA - Youtube Video Analysis

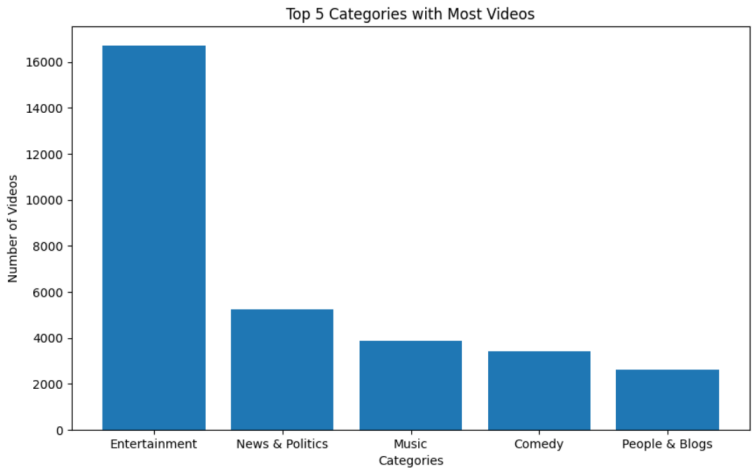
Dataset View | Channel View | Github | aggle | Code

Dataset View

Choose a dataset: USvideos

Submit

Top 5 Categories with Most Videos



Category	Counts
Entertainment	16712
News & Politics	5241
Music	3858
Comedy	3429
People & Blogs	2624

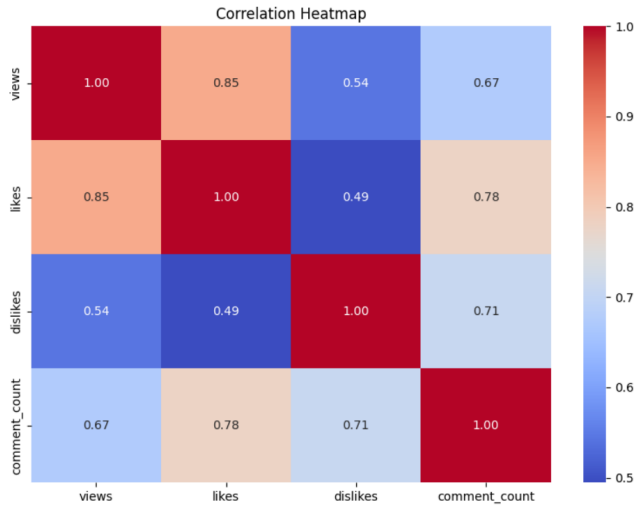
Top 10 Liked Videos

Title	Likes
YouTube Rewind: The Shape of 2017 | #YouTubeRewind	2912710
YouTube Rewind: The Shape of 2017 | #YouTubeRewind	2811216
YouTube Rewind: The Shape of 2017 | #YouTubeRewind	2656672
Marvel Studios' Avengers: Infinity War Official Trailer	2606663
Marvel Studios' Avengers: Infinity War Official Trailer	2584674
Marvel Studios' Avengers: Infinity War Official Trailer	2555411
Marvel Studios' Avengers: Infinity War Official Trailer	2513102
Marvel Studios' Avengers: Infinity War Official Trailer	2444956
Marvel Studios' Avengers: Infinity War Official Trailer	2331352
YouTube Rewind: The Shape of 2017 | #YouTubeRewind	2251815

Top 10 Most Viewed Videos

Title	Views
YouTube Rewind: The Shape of 2017 | #YouTubeRewind	125432237
YouTube Rewind: The Shape of 2017 | #YouTubeRewind	113876217
YouTube Rewind: The Shape of 2017 | #YouTubeRewind	100911567
Marvel Studios' Avengers: Infinity War Official Trailer	89930713
Marvel Studios' Avengers: Infinity War Official Trailer	87449453
Marvel Studios' Avengers: Infinity War Official Trailer	84281319
Marvel Studios' Avengers: Infinity War Official Trailer	80360459
YouTube Rewind: The Shape of 2017 | #YouTubeRewind	75969469
Marvel Studios' Avengers: Infinity War Official Trailer	74789251
Marvel Studios' Avengers: Infinity War Official Trailer	66637636

Correlation Between



BDA - Youtube Video Analysis

Dataset View | Channel View | Github | kaggle | Code

Channel View

Choose a dataset: USvideos

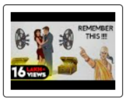
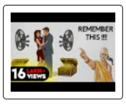
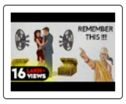
Enter a channel name:

example channel names ...
(CA:EminemVEVO) (DE:NFL) (IN:SeeKen)
(RU:Druzhko Show) (US:CaseyNeistat)

Submit

Channel : SeeKen (Videos : 57)

SeeKen : Top 10 Liked Videos

	TOP 5 HABITS OF HIGHLY SUCCESSFUL PEOPLE - YOU CAN WIN BY SHIV | SeeKen	23171	554092
	TOP 5 HABITS OF HIGHLY SUCCESSFUL PEOPLE - YOU CAN WIN BY SHIV | SeeKen	20543	479473
	TOP 5 HABITS OF HIGHLY SUCCESSFUL PEOPLE - YOU CAN WIN BY SHIV | SeeKen	20543	479473
	CHANAKYA NITI IN HINDI - à¤¬à¥□à¤¦à¥□à¤§à¤¿à¤®à¤¾à¤¨ à¤□à¤®à¥□à¤° à¤²à¥□à¤□à¥□ à¤□à¥□ à¤°à¤¾à¤□à¤¼ ANIMATED SUMMARY	18834	390119
	CHANAKYA NITI IN HINDI - à¤¬à¥□à¤¦à¥□à¤§à¤¿à¤®à¤¾à¤¨ à¤□à¤®à¥□à¤° à¤²à¥□à¤□à¥□ à¤□à¥□ à¤°à¤¾à¤□à¤¼ ANIMATED SUMMARY	17494	352205
	CHANAKYA NITI IN HINDI - à¤¬à¥□à¤¦à¥□à¤§à¤¿à¤®à¤¾à¤¨ à¤□à¤®à¥□à¤° à¤²à¥□à¤□à¥□ à¤□à¥□ à¤°à¤¾à¤□à¤¼ ANIMATED SUMMARY	14926	265125
	TOP 5 HABITS OF HIGHLY SUCCESSFUL PEOPLE - YOU CAN WIN BY SHIV | SeeKen	12444	206770
	TOP 5 HABITS OF HIGHLY SUCCESSFUL PEOPLE - YOU CAN WIN BY SHIV | SeeKen	12444	206770
	TOP 5 HABITS OF HIGHLY SUCCESSFUL PEOPLE - YOU CAN WIN BY SHIV | SeeKen	12444	206770
	TOP 5 HABITS OF HIGHLY SUCCESSFUL PEOPLE - YOU CAN WIN BY SHIV | SeeKen	12444	206770

SeeKen : Top Viewed Video

	TOP 5 HABITS OF HIGHLY SUCCESSFUL PEOPLE - YOU CAN WIN BY SHIV | SeeKen	554092

SeeKen : Top Impression Video

	WHY TOPPERS WORK FOR FAILURES (HINDI) - EDUCATION SYSTEM !!	Impression : 10.144573832601168

Code

```
# R Language Code for Youtube Video Analysis

# Team : Atharva, Aditya, Hitesh

csv_file_path <- 'static/datasets/USvideos.csv'

# Load the CSV file into a data frame
df <- read.csv(csv_file_path, encoding = 'latin1')

# Print the first few rows of the data frame to verify its contents
# head(df)
library(dplyr)
library(jsonlite)
```

```
# Define a category mapping as a named character vector
category_mapping <- c(
  "1" = 'Film & Animation',
  "2" = 'Autos & Vehicles',
  "10" = 'Music',
  "15" = 'Pets & Animals',
  "17" = 'Sports',
  "18" = 'Short Movies',
  "19" = 'Travel & Events',
  "20" = 'Gaming',
  "21" = 'Videoblogging',
  "22" = 'People & Blogs',
  "23" = 'Comedy',
  "24" = 'Entertainment',
  "25" = 'News & Politics',
  "26" = 'Howto & Style',
  "27" = 'Education',
  "28" = 'Science & Technology',
  "29" = 'Nonprofits & Activism',
  "30" = 'Movies',
  "31" = 'Anime/Animation',
  "32" = 'Action/Adventure',
  "33" = 'Classics',
  "34" = 'Comedy',
  "35" = 'Documentary',
  "36" = 'Drama',
```

```
  "36" = 'Drama',
  "37" = 'Family',
  "38" = 'Foreign',
  "39" = 'Horror',
  "40" = 'Sci-Fi/Fantasy',
  "41" = 'Thriller',
  "42" = 'Shorts',
  "43" = 'Shows',
  "44" = 'Trailers'
)

# Function to get the top 5 categories with the most videos
top_5_categories <- function(df) {
  # Ensure the "category_id" column exists in your data
  if ("category_id" %in% colnames(df)) {
    # Get the top 5 categories and their counts
    top_categories <- as.data.frame(table(df$category_id))

    # Map category IDs to names
    top_categories$category_id <- sapply(top_categories$category_id, function(cat_id) category_mapping[as.character(cat_id)])

    # Rename columns
    colnames(top_categories) <- c("Category", "Count")

    # Convert to a list
    jsondata <- as.list(top_categories)
    return(jsondata)
  } else {
    return("The 'category_id' column is not found in the dataset.")
```

```
  }
}

Top_10_Liked_Videos <- function(df) {
  # Sort the DataFrame by 'likes' in descending order and select the top 10 rows
  top_rated_videos <- head(arrange(df, desc(likes)), 10)

  # Extract the video titles and likes
  video_titles <- top_rated_videos$title
  video_likes <- top_rated_videos$likes

  # Create a list of video information
  video_info_list <- lapply(1:10, function(i) {
    list(title = video_titles[i], likes = video_likes[i])
  })

  # Convert the list to a JSON-like structure
  jsondata <- lapply(video_info_list, toJSON, pretty = TRUE)

  # Print the resulting JSON-like data
  for (item in jsondata) {
    cat(item, "\n")
  }

  return(jsondata)
}

Top_10_Most_Viewed_Videos <- function(df) {
  # Sort the DataFrame by 'views' in descending order and select the top 10 rows
```

```
  top_viewed_videos <- head(arrange(df, desc(views)), 10)

  # Extract the video titles and views
  video_titles <- top_viewed_videos$title
  video_views <- top_viewed_videos$views

  # Create a list of video information
  video_info_list <- lapply(1:10, function(i) {
    list(title = video_titles[i], views = video_views[i])
  })

  # Convert the list to a JSON-like structure
  jsondata <- lapply(video_info_list, toJSON, pretty = TRUE)

  # Print the resulting JSON-like data
  for (item in jsondata) {
    cat(item, "\n")
  }

  return(jsondata)
}


correlation_between <- function(df) {
  # Select the columns for which you want to calculate correlations
  selected_columns <- df[c('views', 'likes', 'dislikes', 'comment_count')]

  # Calculate the correlation matrix
  correlation_matrix <- cor(selected_columns, use = 'pairwise.complete.obs')
```

```
  # Convert the correlation matrix to a list
  correlation_data <- as.list(correlation_matrix)

  return(correlation_data)
}

# ----------------------------------------------------------------------

channel_Top_10_Liked_Videos <- function(df, channel_title) {
  # Filter the DataFrame for videos by the specified channel_title
  channel_df <- df[df$channel_title == channel_title, ]

  # Get the top 10 most liked videos from the filtered DataFrame
  top_liked_videos <- head(arrange(channel_df, desc(likes)), 10)

  # Extract the desired columns
  top_liked_videos <- top_liked_videos[c('video_id', 'title', 'likes', 'views', 'thumbnail_link')]

  # Convert the top 10 liked videos to a list of dictionaries (data frames)
  top_liked_videos_list <- as.data.frame(top_liked_videos)

  return(top_liked_videos_list)
}
```

```r
channel_Top_Viewed_Video <- function(df, channel_title) {
  # Filter the DataFrame for videos by the specified channel_title
  channel_df <- df[df$channel_title == channel_title, ]

  # Find the video with the highest views from the filtered DataFrame
  top_viewed_video <- channel_df[which.max(channel_df$views), ]

  # Extract the desired columns
  top_viewed_video <- top_viewed_video[c('video_id', 'title', 'views', 'thumbnail_link')]

  # Convert the top viewed video to a list of dictionaries (data frames)
  top_viewed_video_list <- as.data.frame(top_viewed_video)

  return(top_viewed_video_list[1, ])  # Return the first (and only) row
}
```

```r
channel_Top_Impression_Video <- function(df, channel_title) {
  # Filter the DataFrame for videos by the specified channel_title
  channel_df <- df[df$channel_title == channel_title, ]

  # Calculate an "impression" score for each video based on likes, dislikes, and views
  channel_df$impression <- (channel_df$likes - channel_df$dislikes) / channel_df$views

  # Find the video with the highest impression score from the filtered DataFrame
  top_impression_video <- channel_df[which.max(channel_df$impression), ]

  # Extract the desired columns
  top_impression_video <- top_impression_video[c('video_id', 'title', 'impression', 'thumbnail_link')]

  # Convert the top impression video to a list of dictionaries (data frames)
  top_impression_video_list <- as.data.frame(top_impression_video)

  return(top_impression_video_list[1, ])  # Return the first (and only) row
}
```

```r
# Example usage of the functions

# result <- top_5_categories(df)
# result <- Top_10_Liked_Videos(df)
# result <- Top_10_Most_Viewed_Videos(df)
# result <- correlation_between(df)



result <- channel_Top_10_Liked_Videos(df, "CaseyNeistat")
# result <- channel_Top_Viewed_Video(df, "CaseyNeistat")
# result <- channel_Top_Impression_Video(df, "CaseyNeistat")
print(result)

-------------------------------------------------------------------------------------------
                                    OUTPUT : channel_Top_10_Liked_Videos
-------------------------------------------------------------------------------------------
```

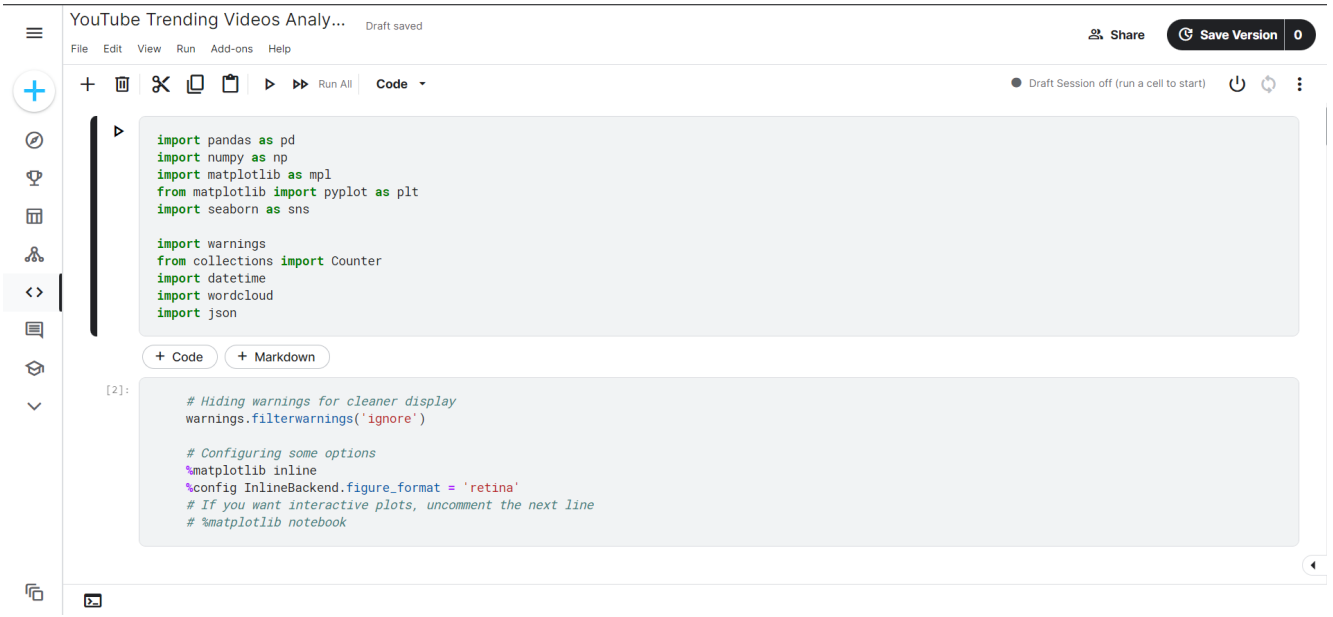
```
video_id                                                    title
1  a7NJ6Gek9v4 ALL TIME GREATEST AIRPLANE SEAT - Emirates First Class Suite
2  a7NJ6Gek9v4 ALL TIME GREATEST AIRPLANE SEAT - Emirates First Class Suite
3  a7NJ6Gek9v4 ALL TIME GREATEST AIRPLANE SEAT - Emirates First Class Suite
4  0-7wMD5ISIs                 ABANDONED MALL TURNED INTO WINTER WONDERLAND
5  0-7wMD5ISIs                 ABANDONED MALL TURNED INTO WINTER WONDERLAND
6  0-7wMD5ISIs                 ABANDONED MALL TURNED INTO WINTER WONDERLAND
7  7kLO2AB5SPM                                          about Logan Paul
8  0-7wMD5ISIs                 ABANDONED MALL TURNED INTO WINTER WONDERLAND
9  0-7wMD5ISIs                 ABANDONED MALL TURNED INTO WINTER WONDERLAND
10 0-7wMD5ISIs                 ABANDONED MALL TURNED INTO WINTER WONDERLAND


    likes   views                                thumbnail_link
1  263740 5875258 https://i.ytimg.com/vi/a7NJ6Gek9v4/default.jpg
2  255090 5507332 https://i.ytimg.com/vi/a7NJ6Gek9v4/default.jpg
3  240179 4945996 https://i.ytimg.com/vi/a7NJ6Gek9v4/default.jpg
4  232199 6647081 https://i.ytimg.com/vi/0-7wMD5ISIs/default.jpg
5  231177 6620905 https://i.ytimg.com/vi/0-7wMD5ISIs/default.jpg
6  229931 6588744 https://i.ytimg.com/vi/0-7wMD5ISIs/default.jpg
7  228488 5069042 https://i.ytimg.com/vi/7kLO2AB5SPM/default.jpg
8  228386 6550370 https://i.ytimg.com/vi/0-7wMD5ISIs/default.jpg
9  226757 6502307 https://i.ytimg.com/vi/0-7wMD5ISIs/default.jpg
10 224426 6428410 https://i.ytimg.com/vi/0-7wMD5ISIs/default.jpg
```


 Hitesh Sharma

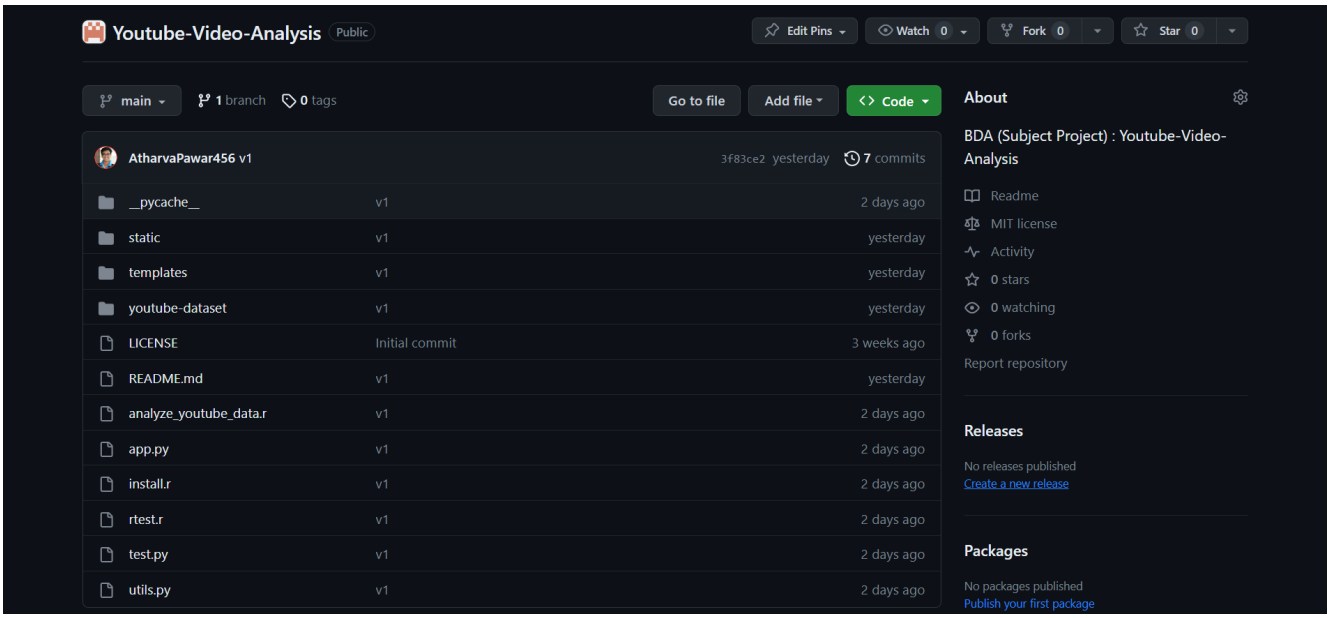
 Aditya Vyas

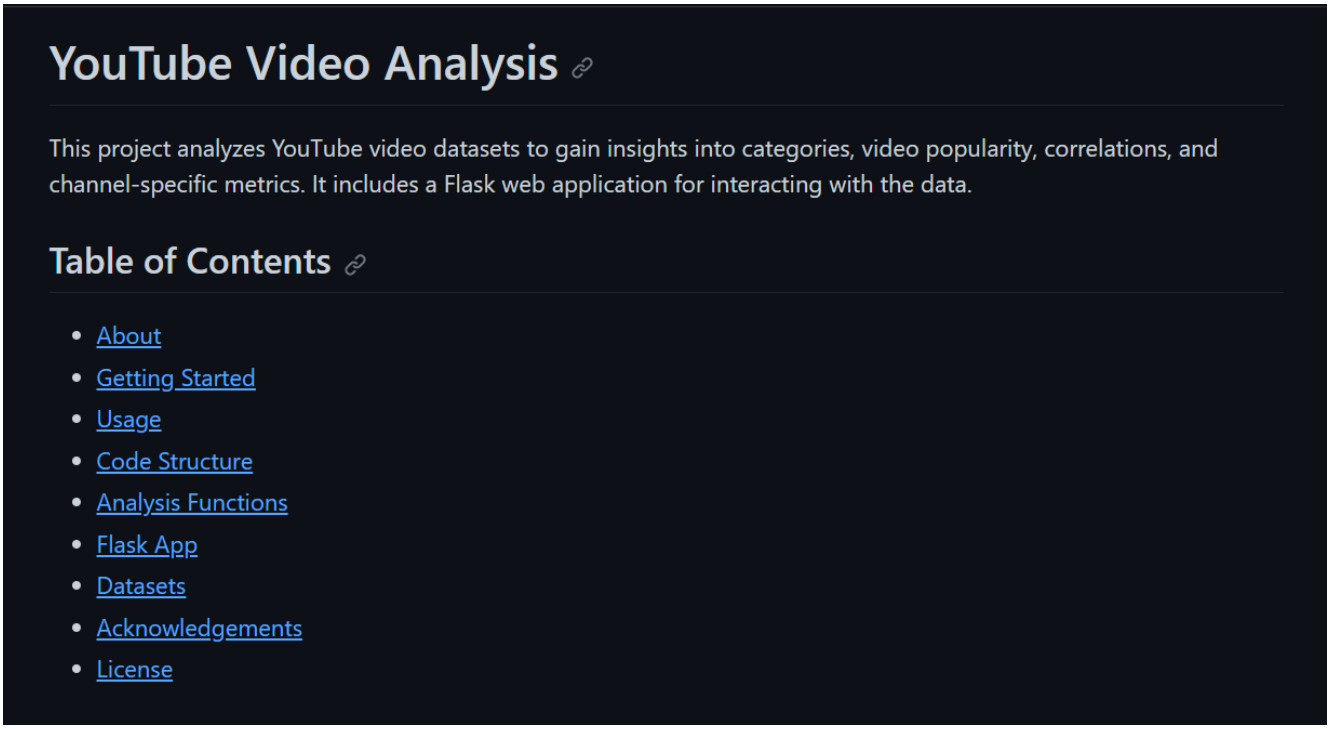
 Atharva Pawar

8.2 Kaggle Notebook: (link)



8.3 GitHub Repo: (Link)





About

This project utilizes Python, Flask, and R to analyze YouTube trending video datasets from multiple countries. The goal is to explore categories, popularity metrics, correlations, and channel-specific analytics. An interactive Flask app allows users to select datasets and view analysis results.

Key features:

- Top categories by video count
- Top 10 liked/viewed videos
- Correlations between views, likes, dislikes, comments
- Channel-specific analytics
- R scripts for data analysis
- Flask app for interacting with data

Getting Started

Prerequisites

- Python 3.x
- Flask
- Pandas
- NumPy
- Matplotlib
- Seaborn
- WordCloud

```
pip install flask pandas numpy matplotlib seaborn wordcloud
```

Installation

```
git clone https://github.com/capstone-project-SECURIX/Youtube-Video-Analysis.git
cd youtube-video-analysis
pip install -r requirements.txt
```

Usage

```
python app.py
```

The Flask app will be available at http://localhost:5000. Select a dataset and explore the video analytics.

Code Structure

```
.
├── static/
│   ├── datasets/        # CSV datasets
│   ├── images/          # Generated images
│   └── styles/          # CSS files
│
├── templates/           # HTML templates
│
├── youtube-dataset/     # Raw dataset files
│
├── analyze_youtube_data.r    # R analysis scripts
├── app.py               # Flask app
├── utils.py             # Python analysis functions
├── requirements.txt
└── README.md
```

Analysis Functions

Located in `analyze_youtube_data.r` and `utils.py`:

- `top_categories()`: Get top categories by video count
- `top_liked_videos()`: Get top 10 liked videos
- `top_viewed_videos()`: Get top 10 viewed videos
- `correlation()`: View correlations between metrics
- `channel_top_liked()`: Get top liked for channel
- `channel_top_viewed()`: Get most viewed video for channel
- ...more

Flask App

`app.py` handles routes and serves analysis results to the template files.

- `/` - Homepage, select dataset
- `/dataset` - Display analysis for selected dataset
- `/channel` - Display channel-specific analytics
- `/rlang` - Display R code

Datasets

The `youtube-dataset/` folder contains CSV files for different countries:

- USvideos.csv
- RUvideos.csv
- CAvideos.csv
- DEvideos.csv
- INvideos.csv

Acknowledgements

- Kaggle Dataset
- Kaggle Notebook

License

REFERENCES

[1] *Loh, F., Wamser, F., Poignée, F. et al. YouTube Dataset on Mobile Streaming for Internet Traffic Modeling and Streaming Analysis. Sci Data 9, 293 (2022).*

https://doi.org/10.1038/s41597-022-01418-y

[2]*Szmuda, Tomasz, et al. "YouTube as a source of patient information for Coronavirus Disease (COVID-19): A content-quality and audience engagement analysis." Risk Management and Healthcare Policy, vol. 13, 2020, pp. 2395-2400, doi: 10.1002/rmv.2132*

.

[3] *Carvache-Franco, Orly, et al. "Topics and destinations in comments on YouTube tourism videos during the Covid-19 pandemic." PLOS ONE, vol. 18, no. 3, 2023, e0281100,*

https://doi.org/10.1371/journal.pone.0281100

APPENDIX-II

<<Plagiarism Report>>