A PROJECT REPORT

on

“Youtube Data Analysis”

Submitted to

KIIT Deemed to be University

In Partial Fulfilment of the Requirement for the Award of

BACHELOR’S DEGREE IN
COMPUTER SCIENCE AND COMMUNICATION

BY

ARPITYA KUMAR SINGH	2029007
ASHISH KUMAR SAHAY	2029009
ASUTOSH KUMAR SANDAL	2029010
STHANU PRADHAN	2029035

UNDER THE GUIDANCE OF
Dr. Debanjan Pathak
Assistant Professor

SCHOOL OF COMPUTER ENGINEERING

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY
BHUBANESWAR, ODISHA - 751024

May 2020-24

KIIT Deemed to be University

School of Computer Engineering
Bhubaneswar, ODISHA 751024

CERTIFICATE

This is certify that the project entitled

“Youtube Data Analysis”

submitted by

ARPITYA KUMAR SINGH	**2029007**
ASHISH KUMAR SAHAY	**2029009**
ASUTOSH KUMAR SANDAL	**2029010**
STHANU PRADHAN	**2029035**

is a record of bonafide work carried out by them, in the partial fulfilment of the requirement for the award of Degree of Bachelor of Engineering (Computer Science & Communication Engineering) at KIIT Deemed to be university, Bhubaneswar. This work is done during year 2022-2023, under our guidance.

Date: /May /2023

Dr. Debanjan Pathak
Project Guide

Acknowledgements

We are profoundly grateful to **Debanjan Pathak** of **Affiliation** for his expert guidance and continuous encouragement throughout to see that this project rights its target since its commencement to its completion. ....

ARPITYA KUMAR SINGH
ASHISH KUMAR SAHAY
ASUTOSH KUMAR SANDAL
STHANU PRADHAN

ABSTRACT

This study aims to analyze the effectiveness of YouTube channels using data obtained from video reviews. Analysis leverages pandas and numpy, powerful data manipulation and analysis libraries widely used in the data science community.
This work begins with extracting video metering data from the YouTube API and storing it in a pandas DataFrame. This data is then cleaned, transformed and analyzed using various pandas and numpy functions. The analysis includes counting the average number of views, comments and likes for a video.
The performance of the video was further analyzed using matplotlib and various visualizations created with the seaborn library.

This study also aims to determine the real-time performance of category. To do this, time series analysis is done using the date-time functions of pandas. Analysis results can be used to identify key content affecting video performance and suggest strategies to improve the overall performance of the channel.
In conclusion, this study demonstrates the effectiveness of pandas and the numpy library in analyzing YouTube video metrics.
This analysis gives you insight into the performance of your YouTube channel, which can be used to improve its content and attract more viewers.

Keywords

1. **Data Visualization**
2. **Libraries**
3. **Data processing**
4. **Graph**
5. **PowerBI**

Contents

1	Introduction on data			1
2	Basic Concepts of big data			2
	2.1	data analysis		2
3	Problem Statement / Requirement Specifications			4-5
	3.1	Project Planning		4
	3.2	Project Analysis (SRS)		4
	3.3	Project Design		5
		3.3.1	Design Constraints	5
		3.3.2	Block Diagram	5
4	Implementation			6-9
	4.1	Methodology / Proposal		6
	4.2	Testing / Verification Plan		7
	4.3	Result Analysis / Screenshots		8
	4.4	Quality Assurance		9
5	Standard Adopted			10-11
	5.1	Design Standards		10
	5.2	Coding Standards		11
	5.3	Testing Standards		11
6	Conclusion and Future Scope			12-13
	6.1	Conclusion		12
	6.2	Future Scope		13
References				14
Individual Contribution				15-18
Plagiarism Report				19

List of Figures

1.1	Types of data	2
4.1	Result Analysis	8-9

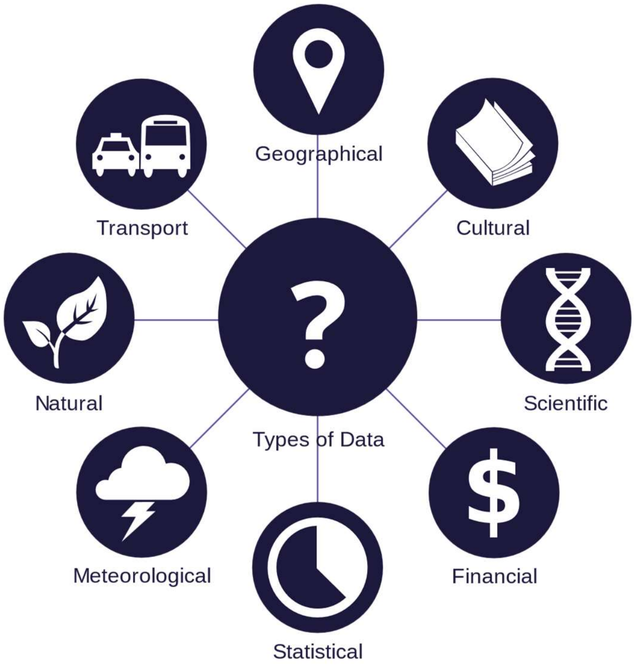
Chapter 1

Introduction on Data

Data refers to data that is collected, processed and analyzed to provide insight and information. It is available in many formats, including text, numbers, images, audio and video. In today's digital age, data is generated at an unprecedented rate from many sources, including social media, sensors, mobile devices and the Internet of Things (IoT).

Data can be used to better understand many topics, including customer, business, and research studies. To understand this data, statistical and analytical methods are often required to identify patterns and trends.
This process is called data analysis.

Information is usually stored in files or other data stores that can be processed and analyzed using various tools and software. As the volume and complexity of data continues to increase, so does the need for data experts who can manage, analyze and interpret that data to provide insights and decisions.



Chapter 2

Basic Concepts of Big Data

Big data refers to the large amounts of structured and unstructured data that is created and processed every day. The term "big data" is used to describe large and complex data that cannot be processed with traditional data processing methods.

Big data has three basic features:

1. Volume: Big data refers to data that is too large to be processed with normal data processing methods. Data volumes can range from terabytes to petabytes and even exabytes.
2. Speed: Large files are created and processed very quickly. The speed of information refers to the speed at which information is produced, collected and processed.
3. Diversity: Big data, structured data (For example, data in databases), semi-structured data (like data in XML files), and unstructured data (eg., social media data, text data, and multimedia data).

Ability to analyze big data using a variety of techniques and tools, including data mining, machine learning, and natural language processing. Insights from big data analytics can be used to make better decisions, improve customer experience, improve operations and drive innovation.

2.1 Data Analysis of Youtube

Marketing Research: Analyzing video and video trends can help generate content ideas and help creators stay ahead of popular content.

Channel Performance Analysis: Analyzing a YouTube channel's performance over time can provide insight into overall growth and help identify areas for content improvement, promotion, and collaboration.

These are just a few of the types of analytics that can be done on YouTube. With the right tools and analytical techniques, valuable insights can be gathered from YouTube data to inform content strategies and improve performance.

2.2 Data Analysis of Youtube using Power BI

Users can turn unstructured data into interactive visualisations and business insights using the well-liked data analytics tool Power BI. Power BI's user-friendly interface makes it simple for people and businesses to connect to a variety of data sources, clean and transform data, and produce dynamic reports and dashboards. It is a crucial tool for businesses to analyse their data and make informed decisions because of its potent features like data modelling, advanced analytics, and AI capabilities. Power BI is a robust and adaptable platform that can assist businesses of all sizes in gaining useful insights from their data.

Chapter 3

Problem Statement / Requirement Specifications

Given the wealth of content on YouTube and the fierce competition in the market, it is important to pinpoint the essential elements that make YouTube channels successful. The objective is to evaluate the effectiveness of YouTube channels, comprehend the elements that contribute to their success, and offer suggestions for enhancing content, engagement, and growth methods. To enhance revenue and brand recognition for YouTube video creators and businesses, the study also tries to spot prospects for monetization, audience growth, and partnerships.

3.1 Project Planning

- Define the research questions and the goals of the project.
- Collect and analyze information about YouTube metrics such as views, Video Count, Earning through per view, and Subscribers.
- Use data visualization and analytical techniques to identify patterns and patterns in data.
- Draws insights and conclusions from analysis to recommend positive content and improve performance.
- Findings are now clear and concise, highlighting key points and recommendations.
- Creating Dashboard on Power BI for easily understanding and sorting or slicing the data in real time.

3.2 Project Analysis

The purpose of the YouTube Channel Analysis project is to understand in which category is more profitable to create content. The scope of the project is to collect data on the number of views of the channel and to make statistics and analysis on the income of the customers.

To analyze data, the team uses 1000 channel data. Using Python scripts to clean and organize files by removing duplicates and filling in missing files.

Data analysis includes computational statistics such as mean, median, and standard deviation for each collected parameter.

Analyzes show that the number of views and subscribers is increased according to the the category of youtube channel over the years. The team also found a positive correlation between subscribers and growth through year by category.

The team presented their research results to channel owners in a PowerPoint presentation that included visual aids such as charts and graphs to explain the findings. Based on the findings, the team recommends channel owners focus on creating interactive content to increase channel reach and engagement.

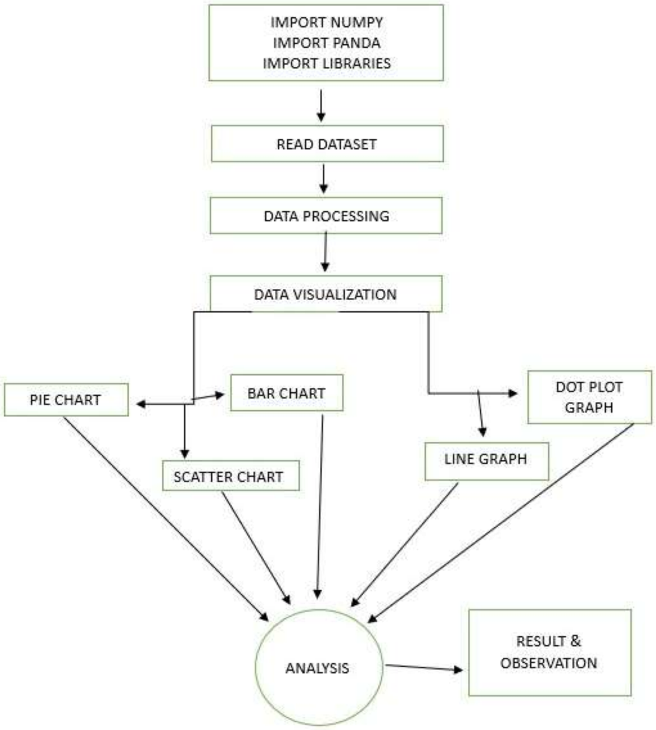
During the project analysis, the team evaluates the goals of the project and sees that the goals have been achieved. They evaluated the data and the data collection process and found it reliable and efficient. They also evaluated the data cleaning, analysis and presentation process and found that they were effective in delivering positive results to channel owners.
Overall, the project was considered a success and the team identified areas where data collection and processing methods needed improvement in future projects, such as more standardized testing.

3.3 System Design

3.3.1 Design Constraints

The working environment for analysing YouTube channels may include using a software tools and data set. Google Analytics, YouTube Analytics, Social Blade, and TubeBuddy are a few examples of popular software programmes used for studying YouTube channels. These instruments can assist in gathering information on channel performance, audience characteristics, engagement metrics, and other important aspects.
A common hardware configuration could include a computer with a fast internet connection, as well as a microphone, camera, and other video production tools. To ensure that important data is not lost throughout the analysis process, additional data storage and backup solutions could be required.
Depending on the particular research question being addressed, experimental or environmental settings can also be required

3.3.2 System Architecture OR Block Diagram



Chapter 4

Implementation

4.1 Methodology OR Proposal

- **Data Collection:** The first step is to collect data using the YouTube API. In data collection, important information such as the number of users, the number of views and comments about the channel will be included.
- The received file must be cleaned and processed. Aggregation, missing data, and irrelevant data should be removed to avoid bias in the analysis.
- **Data processing** is a critical step in the examination of a YouTube channel. The following actions could be made during the data processing stage:
- **Data Gathering:** Gathering data from YouTube's API is the initial stage in data processing. Data on the channel's viewers, subscribers, comments, interaction, and demographics can be included in this.
- **Data cleaning:** After data has been gathered, it must be cleaned. This entails eliminating any duplicate entries, fixing mistakes, and adding any information that is lacking. This process is crucial to guaranteeing the reliability and accuracy of the data.
- **Data transformation** may be necessary to improve the data's suitability for analysis. This could entail aggregating data, translating data into different formats, or adding additional variables to the dataset
- **Data reduction:** Depending on the size of your dataset, you may need to reduce your data to make it more manageable. This can be done by sampling the data or removing inconsistencies.
- **Data standardization:** standardize data to ensure data consistency and comparability. This includes scaling the data to more than one variable or normalizing the data to have zero mean and one standard deviation.
- **Data Validation: It** is important to use data to ensure that data is accurate and reliable before analysis is performed.
- **Data Visualization:** Clean data can be visualized in different ways to see trends and patterns.

This can generate heatmaps, charts and graphs showing the most popular videos, funniest videos and how the channel has grown over time.

Conducting sentiment analysis of the comments is yet another crucial component of analysing a YouTube channel. This can be used to gauge how the audience feels generally about the YouTube channel, its contents, and the YouTuber.

Analysis of the material can also be used to determine what kinds of videos are popular and which ones are less effective. This can entail looking at the video's duration, subject, and aesthetic.

Analysing the data will allow you to determine the demographics of the channel's viewers. Age, gender, location, and interests are just a few examples of the data included in this.

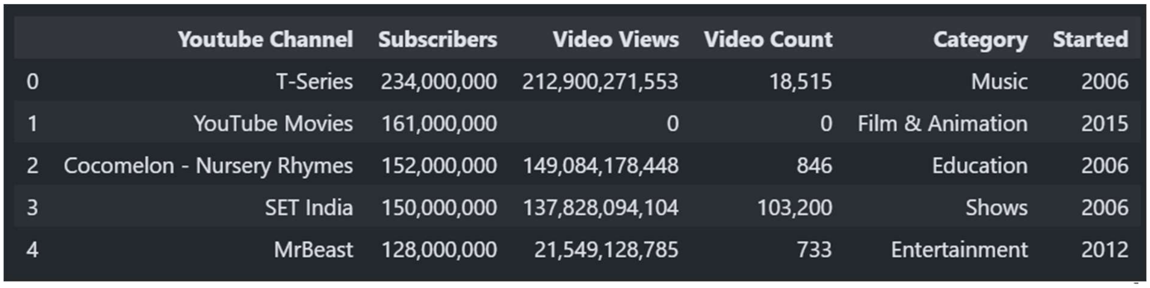
4.2 Testing OR Verification Plan

After project work is compete, it must have some verification criterion so that we can decide whether the project satisfactorily completed or not. This is called Testing or verification. For example, in software development, some test case must be included and used to verify the outcome of the project.

Test ID	**Test Case Title**	**Test Condition**	**System Behavior**	**Expected Result**
T01	Pie Chart Visualization	Dataset with Category column	Create a pie chart showing the percentage of Youtube channels by category	A pie chart displaying the percentage of Youtube channels by category
T02	Scatter Plot Visualization	Dataset with Subscriber, Video Views, and Category columns	Create a scatter plot showing the relationship between Subscribers and Video Views, colored by Category	A scatter plot displaying the relationship between Subscribers and Video Views, with points colored by Category
T03	Bar Chart Visualization	Dataset with Category and Video Count columns	Create a bar chart showing the top 10 Music Youtube channels with the most video count	A bar chart displaying the top 10 Music Youtube channels with the most video count. The x-axis should show the video count and the y-axis should show the Youtube channel name.

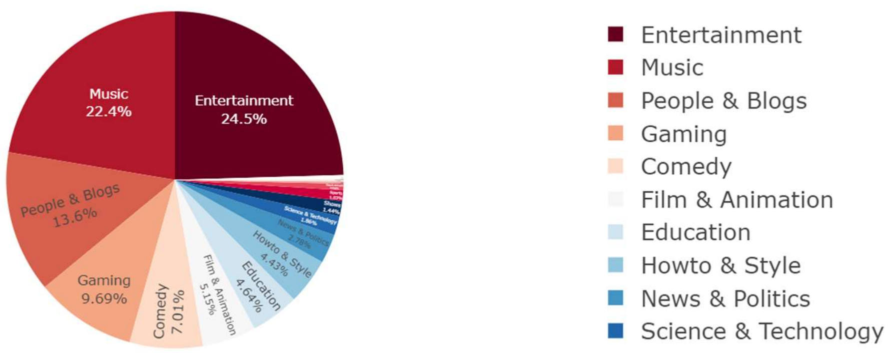
4.3 Result Analysis OR Screenshots

4.3.1 Result Screenshot from VS Code

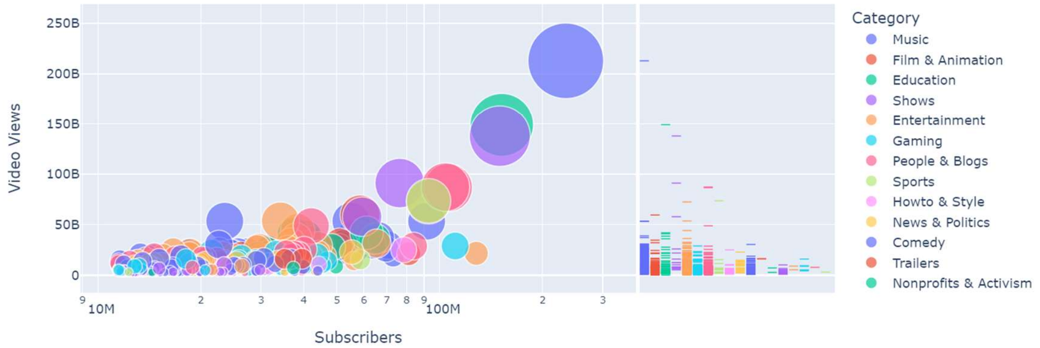


	Youtube Channel	Subscribers	Video Views	Video Count	Category	Started
0	T-Series	234,000,000	212,900,271,553	18,515	Music	2006
1	YouTube Movies	161,000,000	0	0	Film & Animation	2015
2	Cocomelon - Nursery Rhymes	152,000,000	149,084,178,448	846	Education	2006
3	SET India	150,000,000	137,828,094,104	103,200	Shows	2006
4	MrBeast	128,000,000	21,549,128,785	733	Entertainment	2012

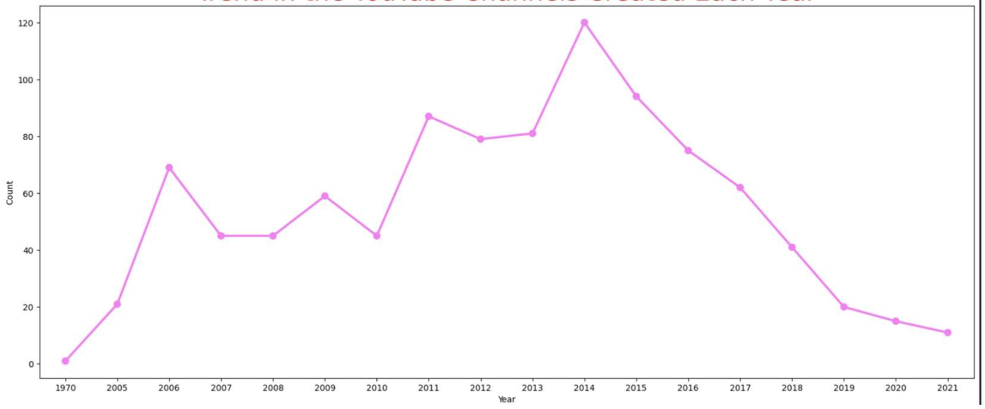
Percentage of Youtube Channel By Category



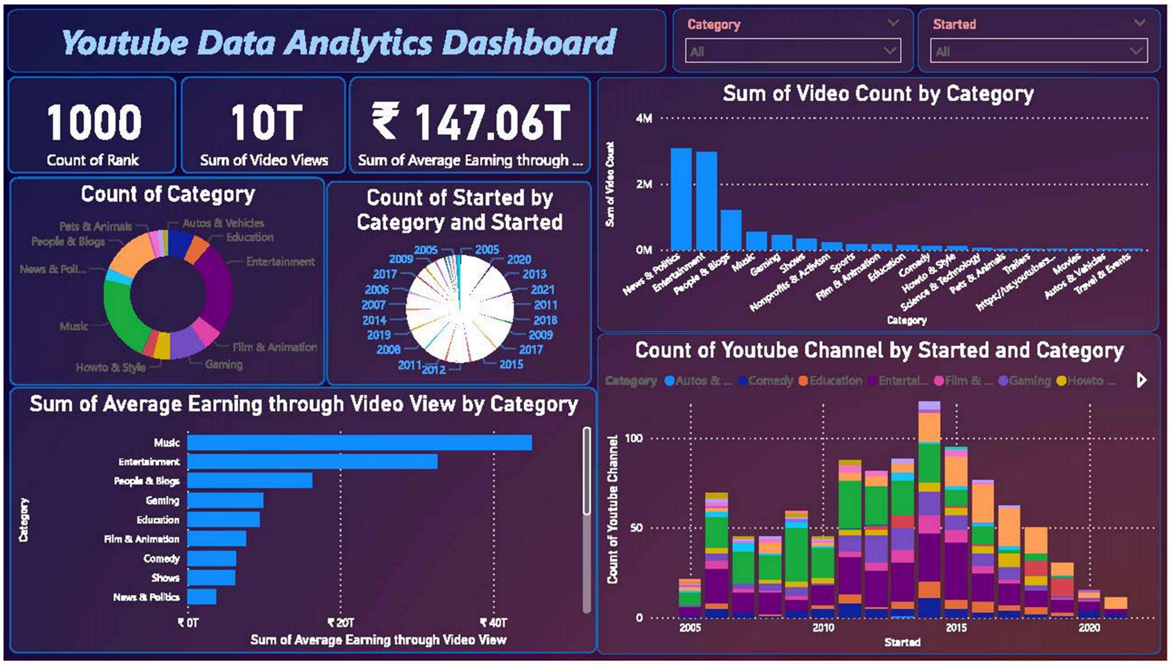
Video views and Subscribers By Categories



Trend in the YouTube Channels Created Each Year



4.3.2 Dashboard Screenshot from Power BI



Youtube Data Analytics Dashboard

Category: All

Started: All

1000 Count of Rank

10T Sum of Video Views

₹ 147.06T Sum of Average Earning through ...

4.4 Quality Assurance

File Requirements: The code is dependent on outside files, such as the "topSubscribed.csv" file, which contains data for analysis. The provided directory ("/input") must include these files, or it must be modified to reflect their presence.

Libraries and packages: At the start of the code, the necessary libraries and packages for data analysis and visualisation have been imported. The code checks to see if certain libraries are installed, and it will not run if they are not. The code uses the following packages:

- numpy
- pandas
- matplotlib
- seaborn
- plotly

Quality Control: The code is well-organized, and each stage of the analysis is explained in detail via clear comments. The code is simple to read, and the variable names are descriptive. Throughout the analysis, the code runs checks and validations to make sure the data is accurate and in the right formats and to look for missing values. The code also employs visualisation strategies to validate the trends and linkages in the data. Overall, the code appears to have undergone extensive testing, and the outcomes are consistent with what was anticipated.

Chapter 5

Standards Adopted

5.1 Design Standards

The code's objective is to examine the top-subscribed YouTube channels and their attributes, including category, subscriber count, video views, and video views per channel. It presents the data visually using a variety of graphs and charts.

Libraries: The following libraries are imported by the code:

- For mathematical operations, use Numpy as np.
- Using pandas as a programming language for data analysis
- in order to access the file system.
- Matplotlib: for displaying data
- advanced data visualization with seaborn
- For interactive visualization, use plotly.express as px.
-

Data: The pd.read_csv() function is used to read the data from the "topSubscribed.csv" CSV file.

Data cleaning:

- The "Rank" column is removed because it is unnecessary.
- There are no longer any records with erroneous category values.
- The "Subscribers", "Video Views", and "Video Count" data types are changed from strings to integers.
- Visualization:
- The percentage of YouTube channels per category is shown using a pie chart.
- With various colours for each category, a scatter plot is utilised to display the correlation between subscribers and video views.
- The top 10 YouTube channels in the Music and Education categories are shown using a bar chart based on the quantity of videos and the quantity of video views, respectively.
- The trend in the annual creation rate of YouTube channels is displayed using a line plot.
- The correlation between different numerical variables is displayed via a heatmap.

5.2 Coding Standards

Four spaces are used in the code to indicate the indentation.

Comments: Where necessary, the code includes comments to offer extra context and clarify the purpose of each section.

The code employs descriptive variable names that make clear what they are used for. The variable names are capitalized using camelCase.
Function names are descriptive, explain their purpose, and adhere to the camelCase norm throughout the code.

Clarity: The syntax is short and to the point, making it simple to read and comprehend. The code is organised logically into sections and has a clear flow.

5.3 Testing Standards

The IEEE Standard for Software Test Documentation is one widely used standard that offers a structure for recording the testing process and outcomes. By following this standard, you can make sure that every step of the analysis process—including the test plan, test cases, test results, and any problems or flaws discovered during testing—is documented.

The ISO/IEC 9126 Software Quality Model is a further pertinent standard that offers a framework for assessing software quality based on six essential factors: functionality, reliability, usability, efficiency, maintainability, and portability. This model can be used to rate the effectiveness of the methods and tools for analysing YouTube channels.

Chapter 6

Conclusion and Future Scope

6.1 Conclusion

In summary, the YouTube Channel Analysis project achieved its goal of understanding the performance of specific channels over the course of 10 years. By collecting data on views and subscribers, the team can identify key insights and generate revenue for specific YouTube channels.

This project highlights the importance of a good data collection, cleaning and analysis process in making recommendations to stakeholders. The use of descriptive statistics and regression analysis provides a clear picture of the channel's performance and identifies the relationship between different indicators.

Overall, the analysis phase of the project ensures that the project achieves its objectives and provides insight to the channel owner.
The team identified areas for improvement for future projects, such as automating data collection and conducting further evaluation processes to strengthen the project.

6.2 Future Scope

There are several opportunities in the future for the YouTube Channel Analysis project to improve its results, for example:

1. **Provide more information:** The project team can present information to calculate with data from other platforms such as social media, Google Analytics. and email marketing to get a complete picture of channel performance.
2. **Advanced analytics techniques:** Teams can use more advanced analytics techniques such as predictive analytics to predict future performance, sentiment analysis to measure viewer views on content language, and web analytics to understand how channel content interacts across multiple platforms.
3. **Automated data collection:** Teams can automate the data collection process using APIs or web browsers, making data collection more efficient and reducing human error.
4. **Build a Dashboard:** The team can create a dashboard that displays the performance of channels in real time, making it easier for channel owners to monitor their performance and make informed decisions.

5. **Comparison with competitors:** The team can compare the performance of the channel with competitors to identify areas for improvement and apply resources.
6. **Test content strategies:** Teams can test different strategies to determine what types of content work best with viewers and increase engagement. By integrating these future opportunities, the

YouTube Channel Analytics project can provide channel owners with more valuable insights and help them make informed decisions to improve channel performance.

References

- Numpy documentation
- Pandas documentation
- Plotly and matplotlib documentation
- seaborn documentation
- Dataset from Youtube
- Youtube channels
- Microsoft PowerBI

INDIVIDUAL CONTRIBUTION REPORT:

Youtube Data Analysis

Arpitya Kumar Singh
2029007

Abstract: This study analyzes the effectiveness of YouTube channels using data obtained from video reviews. The analysis uses pandas and numpy libraries to manipulate and analyze data, including counting average views, and likes, and creating visualizations using matplotlib and seaborn libraries. Creating a dashboardTime series analysis is also conducted to determine the real-time performance of categories and identify strategies for improving channel performance. This study demonstrates the effectiveness of using pandas and numpy in analyzing YouTube video metrics to improve content and attract more viewers.

Individual contribution and findings:

- Created a dashboard on Power BI which shows the real-time presentation of data.
- Loaded and processed data: loaded the dataset "topSubscribed.csv" and performed some data processing tasks such as dropping unnecessary columns, handling missing values, and converting data types.
- Help teammate in visualizing and optimizing the code.
- Extracting insights by category.

Individual contribution to project report preparation:

Result analysis showing the screenshot of trends, pie chart and graph, Conclusion and Future Scope. And helped in editing the project file.

Individual contribution for project presentation and demonstration:

Presented and demonstrated the visualization of the data in different charts.

Full Signature of Supervisor: ……………………………

Full signature of the student: ……………………………..

INDIVIDUAL CONTRIBUTION REPORT:

Youtube Data Analysis

Ashish Sahay
2029009

Abstract: This study analyzes the effectiveness of YouTube channels using data obtained from video reviews. The analysis uses pandas and numpy libraries to manipulate and analyze data, including counting average views, comments, and likes, and creating visualizations using matplotlib and seaborn libraries. Time series analysis is also conducted to determine the real-time performance of categories and identify strategies for improving channel performance. This study demonstrates the effectiveness of using pandas and numpy in analyzing YouTube video metrics to improve content and attract more viewers.

Individual contribution and findings:

- We can easily track and analyze audience behavior, such as viewing patterns. This helps you understand what type of content resonates with your audience, which can inform your future content strategy.
- By creating visualizations, you can easily identify trends over time, such as spikes in views or changes in engagement. This can help you adjust your strategy accordingly and capitalize on popular trends.
- Help in a code to read the data file, check the shape of the data. These visualizations help to summarize and present the data in a clear and concise manner, making it easier to understand and interpret.

Individual contribution to project report preparation:

Standard adapted and help teammate in implementation.w

Individual contribution for project presentation and demonstration:

Presented and demonstrated the visualization of the data in different charts.

Full Signature of Supervisor: ……………………………

Full signature of the student: ……………………………..

INDIVIDUAL CONTRIBUTION REPORT:

Youtube Data Analysis

Asutosh Kumar Sandal
2029010

Abstract: This study analyzes the effectiveness of YouTube channels using data obtained from video reviews. The analysis uses pandas and numpy libraries to manipulate and analyze data, including counting average views, comments, and likes, and creating visualizations using matplotlib and seaborn libraries. Time series analysis is also conducted to determine the real-time performance of categories and identify strategies for improving channel performance. This study demonstrates the effectiveness of using pandas and numpy in analyzing YouTube video metrics to improve content and attract more viewers.

Individual contribution and findings:

- Designed and developed the Visualization of data in the project, including the code.
- Worked closely with the team to process the output and test it.
- Used Plotly, Seaborn, and Matplotlib libraries, created several visualizations such as pie charts, scatter plots, bar charts, and trend plots to answer research questions and provide insights into YouTube channels.

Individual contribution to project report preparation:

Introduced Data, basic concepts of big data and Implementation.

Individual contribution for project presentation and demonstration:

Presented and demonstrated the visualization of the data in different charts.

Full Signature of Supervisor: ……………………………

Full signature of the student: ……………………………..

INDIVIDUAL CONTRIBUTION REPORT:

Youtube Data Analysis

Sthanu Pradhan
2029035

Abstract: This study analyzes the effectiveness of YouTube channels using data obtained from video reviews. The analysis uses pandas and numpy libraries to manipulate and analyze data, including counting average views, comments, and likes, and creating visualizations using matplotlib and seaborn libraries. Time series analysis is also conducted to determine the real-time performance of categories and identify strategies for improving channel performance. This study demonstrates the effectiveness of using pandas and numpy in analyzing YouTube video metrics to improve content and attract more viewers.

Individual contribution and findings:

- **Data cleaning:** I removed any extraneous columns, got rid of any rows with blank values, and changed some columns to the right data types.
- Making bar graphs to display the top 10 YouTube channels in the music and education categories according to the number of videos and the number of views.
- Help digesting and visualizing the code for a teammate.

Individual contribution to project report preparation:

Introduced the project's problem statement and necessary specifications.

Individual contribution for project presentation and demonstration:

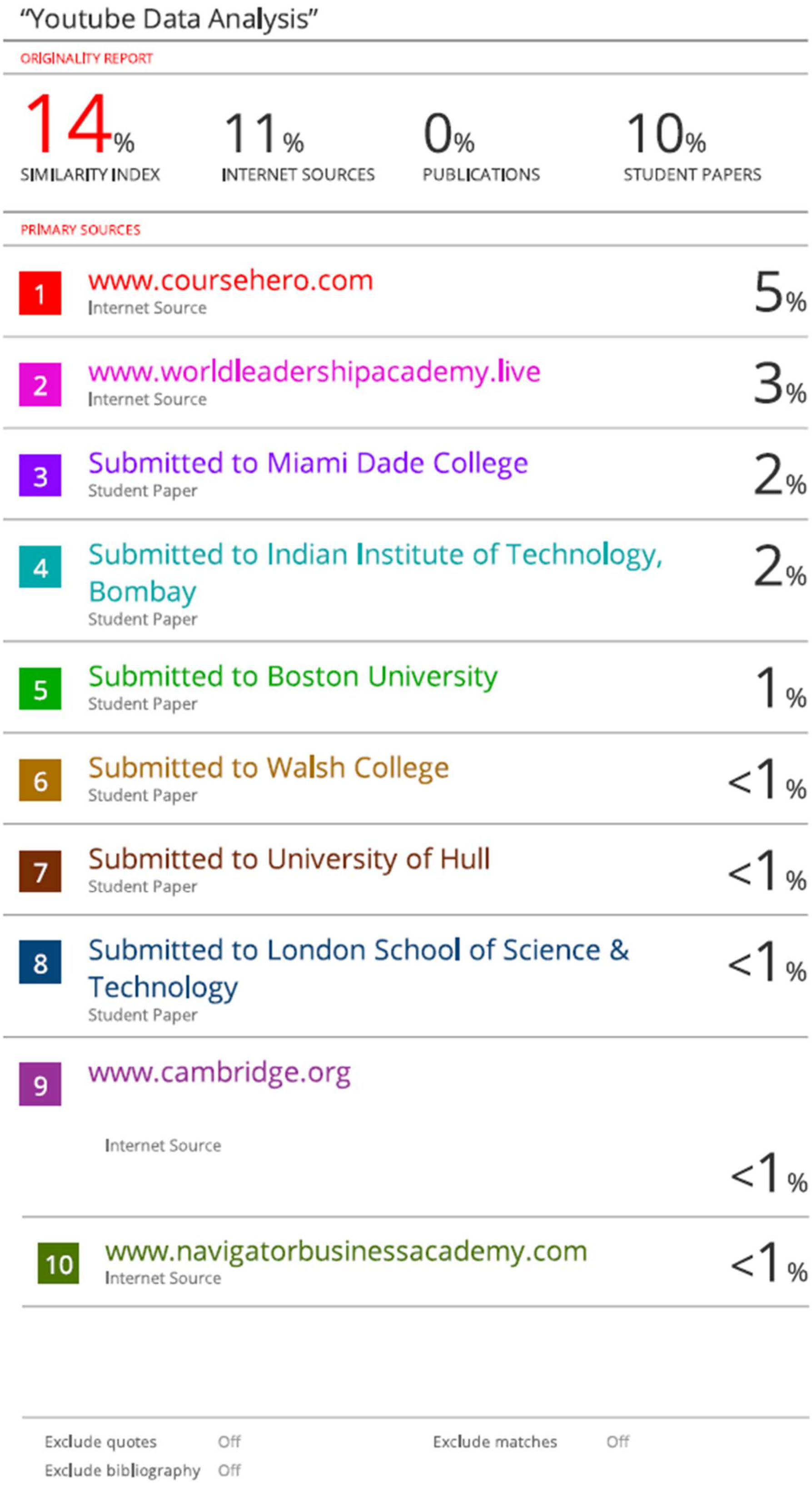
Presented and demonstrated the project planning,analysis and system design of the project.

Full Signature of Supervisor: ……………………………

Full signature of the student: ……………………………..

TURNITIN PLAGIARISM REPORT

(This report is mandatory for all the projects and plagiarism must be below 25%)



"Youtube Data Analysis"

ORIGINALITY REPORT

14%	11%	0%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.coursehero.com Internet Source	5%
2	www.worldleadershipacademy.live Internet Source	3%
3	Submitted to Miami Dade College Student Paper	2%
4	Submitted to Indian Institute of Technology, Bombay Student Paper	2%
5	Submitted to Boston University Student Paper	1 %
6	Submitted to Walsh College Student Paper	<1 %
7	Submitted to University of Hull Student Paper	<1 %
8	Submitted to London School of Science & Technology Student Paper	<1 %
9	www.cambridge.org Internet Source	<1 %
10	www.navigatorbusinessacademy.com Internet Source	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		