YILDIZ TECHNICAL UNIVERSITY

FACULTY OF CHEMISTRY-METALLURGICAL

DEPARTMENT OF MATHEMATICAL ENGINEERING

DESIGN APPLICATIONS PROJECT

Correlations Between Labels and Titles of YouTube Videos

Advisor: Prof. Dr. Fatma Aydın Akgün

Rotinda Öner

20068922

Istanbul, 2023

Table of Contents

FIGURE LIST 6
TABLE LIST 7
PREFACE 8
ABSTRACT 9
1. INTRODUCTION: 11
2. DATA 13
2.1. Data unit 13
2.2. Data item 13
2.3. Observation 13
2.4. Dataset 13
2.5. Quantitative and Qualitative Data 14
2.5.1 Importance of quantitative and qualitative data 14
2.6. Data Sources 15
2.6.1. Sourcing direct data 15
2.6.1.2. Sourcing indirect data 15
3. FREQUENCIES 15
3.1. Measuring frequency 15
3.2 Frequency Distribution 16
3.2.1. How to showcase frequency distributions 16
3.2.2. Frequency tables 16
4. STANDARDS AND CLASSIFICATIONS 16
4.1. Classifications 16
4.2. Importance of standards in statistics 17
5. METADATA 17
5.1. Metadata presentation 17
5.2. Importance of metadata 17
6. DATA VISUALISATION 17
6.1. Ways data can be visualized. 18
6.2. Best use of data visualization 18
7. DATA ANALYSIS 18
7.1. Data Analysis Process 18
7.2. Data Analysis Techniques 19
7.3. Top Data Analysis Tools 19

7.4. Correlation and Causation 19
7.4.1. Importance of correlation and causation 20
7.4.2. Measuring correlation 20
8.PYTHON 20
8.1. Advantages of using Python in Data Analytics 20
8.2. Python Libraries for Data Analytics 21
8.2.1 NumPy: 21
8.2.2 Pandas: 21
8.2.3 Matplotlib: 21
8.2.4 Seaborn: 21
8.2.5 RegEx (Regular Expression): 21
8.2.6 NLTK (Natural Language Toolkit): 21
8.2.7 JSON (JavaScript Object Notation): 21
8.2.8 Collections: 21
8.3 Jupyter Notebook 21
9. CASE STUDY INTRODUCTION 22
9.1. The Dataset 22
9.1.1 Metadata of the Dataset 23
9.2. Data Manipulation Stage 24
9.3 Correlation Between Likes, Dislike, Comment Count and Views 24
9.4. Finding the Percentage of Tag Usage on Videos 25
9.5 Correlation Between Tag Usage and Views, Likes, and Dislikes 30
9.6. Correlation Between Title Length and View Count 31
9.7. Finding the Percentage of Title Words on Videos 31
9.8. Correlation Between Title Words and View Count, Likes and Dislikes. 34
9.9. Performance of Videos with Comments Enabled/Disabled and Ratings Enabled/Disabled 35
9.10 Videos with Comments Enabled/Disabled Comparison with the Most Popular 10 Percent 35
9.11 Videos with Comments Enabled/Disabled Comparison by Most Popularity 37
9.12. Correlation Between Videos with Comments Enabled and Titles and Tags 38
9.13. Success Of a Video with Ratings Turned on Or Off 39
9.14. Determining The Most Popular Genre 40
10. CONCLUSION 43
REFERENCES 44

RESUME..... 46

FIGURE LIST

Figure 1.1: Bar chart showcasing frequency of hours worked [8]

Figure 1.2: Bar chart showcasing frequency of jobs occupied [8]

Figure 2: Screenshots of the Trending's Videos Dataset

Figure 3: Distribution of most popular tags compared to the other 90 percent in the People & Blogs genre.

Figure 4: Distribution of most popular tags compared to the other 90 percent in the Games genre.

Figure 5: Distribution of most popular tags compared to the other 90 percent in the Sports genre.

Figure 6: Distribution of most popular title words compared to the other 90 percent in the Sports genre.

Figure 7: Total views comparison of the 10 percent best performing videos between videos with comments enabled and disabled.

Figure 8: Total dislikes comparison of the 10 percent best performing videos between videos with comments enabled and disabled.

Figure 9: Total likes comparison of the 10 percent best performing videos between videos with comments enabled and disabled.

Figure 10: Total views comparison between videos with comments enabled and disabled.

Figure 11: Total likes comparison between videos with comments enabled and disabled.

Figure 12: Total dislikes comparison between videos with comments enabled and disabled.

Figure 13: Comparisons of total views between videos with comment and ratings enabled, comments disabled and ratings enabled, comments enabled and ratings disabled, and comments and ratings disabled.

Figure 14: Distribution of views comparison between videos genres.

Figure 15: Distribution of likes comparison between videos genres.

Figure 16: Distribution of dislikes comparison between videos genres.

TABLE LIST

Table 1: Dataset about people's age, gender, and income [5]

Table 2: Data about the height of children [15]

Table 3: Dataset highlighting the correlation between likes, dislikes, comments, and views.

Table 4: Dataset showcasing the frequencies of tags used on the sports genre.

Table 5: Dataset showcasing the frequencies of tags used on the sports genre cleaned up.

Table 6: Dataset showcasing the frequencies of words used on title from the sports genre cleaned up.

Table 7: Dataset showcasing the frequencies of words used on title from the sports genre.

Table 8: Dataset showcasing the frequencies of words used on title from the sports genre cleaned up.

Table 9: Dataset showcasing the correlations between views, likes and dislikes of every video with enabled comments.

Table 10: Dataset showcasing the accumulated views, likes and dislikes of every video genre.

PREFACE

I would like to thank Prof. Dr. Fatma Aydın Akgün for her continuous support throughout the course of this thesis by guiding me and giving ideas to improve certain aspects of this project.

ABSTRACT

This research delves deep into the world of YouTube and focuses on videos that have achieved trending status. The goal is to understand the factors that are leading up a video towards the visibility on the platform. The dataset that has been continuously updated since 2018 and is available on Kaggle, serves as the backbone of this exploration.

ÖZET

Bu araştırma, YouTube dünyasına derinlemesine bir bakış sunarak trend durumuna ulaşan videolara odaklanmaktadır. Amaç, bir videonun platformdaki görünürlüğüne yönlendiren faktörleri anlamaktır. Bu keşfin temelini oluşturan veri seti, 2018'den bu yana sürekli olarak güncellenmekte olup Kaggle üzerinde mevcuttur.

1. INTRODUCTION:

In the evolving landscape of online content, YouTube stands as the main platform where videos crave for attention and visibility. Videos that go viral are rewarded and end up on YouTube's trending page, where viral videos become more successful, visible for everyone to see. This research aims to the inner workings of what it takes to be successful on YouTube, with a specific focus on decoding the factors that lands videos to the trending page.

With the platform Kaggle, datasets have become accessible for everyone to gain access to use. Our dataset is a chronicle of YouTube videos gracing the trending page since 2018 and serve as the canvas for our exploration. To determine what it takes for a video to be successful, we use Excel for the dataset retrieval and Python for data analysis, manipulation, and visualization.

This research is not a mere dissection of numbers; it's an exploration of data that was unnoticed by the eye that guides a video to the trending's page. From the strategic use of keywords to the importance of likes and dislikes, each value holds a key to understanding the reason for success. The motivation behind this is not just academic; it's an attempt to teach content creators with insights for a deeper understanding of the YouTube algorithm.

The focus is not solely on statistical outcomes but on the stories behind the trends to get a perspective on the correlation between content, audience, and algorithm that demonstrates the rise of a video to the trending status.

This thesis is divided into 10 different parts, where the first part is the introduction. Chapters 2 to 6 serve as a means to tell the importance of data, how it is gathered, and how it can be used to do research.

Chapter 2 introduces what data is, how it can be sourced, and what is deemed as qualitative and quantitative data.

Chapter 3 is about frequencies and how they can be measured, since this thesis involves a lot of counting frequencies.

Chapter 4 showcases the importance of standards and classifications and how they should be utilized.

Chapter 5 explains the importance of metadata and how vital it is during research.

Chapter 6 explains how data can be visualized and all the ways it can be visualized.

Chapter 7 of this project explains the importance of Data Analysis, since it is through data analysis that this project is possible. It explains how the process is, what techniques should be applied, what tools are used in data analysis and the importance of correlations and causations in data analysis.

Chapter 8 introduces the programming language Python and how powerful of a tool it is for data analysis. In chapter 8 it explains the advantages of using python to use data analysis, the most important libraries doing data analysis and the web application Jupyter Notebook, where python is used on.

Chapter 9 is where the case study begins, using data analysis to dissect what it takes for a video to go viral. This thesis explores which correlations are the most important in a video, the significance of using tags, and the power it holds for virality. It delves into the use of title words, emphasizing the importance of specific words and the optimal length for a video title. Additionally, it discusses the consequences of enabling or disabling comments or ratings and explores which genres are definitively popular, contrasting them with less worthwhile genres to pursue.

2. DATA

Data is a collection of distinct or continuous values that try to give information, tell stories, describing the quantity, quality, fact, statistics, other basic units of meaning, or simply sequences of symbols that may be further interpreted formally. [1]

2.1. Data unit

A data unit is something that is studied in a group and is about gathering information about it. It's also called a unit record or just a record. [2]

2.2. Data item

A data item is like a detail about something that is studied about, measured, or counted. It could be things like height, where someone was born, or how much money they make. It is also called a variable because these details can change over time. [2]

2.3. Observation

An observation is when someone records a specific detail about a thing being studied. Observations can be numbers, like 173 for the height of a person in centimeters, or non-numbers, like "Australia" for the country where someone was born. [2]

2.4. Dataset

A dataset is a complete collection of all observations.

Example for a dataset [2]

	age (years)	sex	income ($)
Person 1 (John Smith)	18	m	50000
Person 2 (Joe Bloggs)	18	m	40000
Person 3 (Sally Jones)	20	f	55000
Person 4 (Linda Lee)	22	f	50000
Person 5 (Harry James)	19	m	35000

Table 1: Dataset about people's age, gender, and income [2]

2.5. Quantitative and Qualitative Data

Quantitative involves measures expressed as numbers. It deals with things like values or counts. This type of data answer questions like "how many," "how much," or "how often."

On the other hand, qualitative data involves in measures of 'types,' and it is represented with names, symbols, or number codes. This kind of data deals with categorical variables, answering questions like "what type."[3]

2.5.1 Importance of quantitative and qualitative data

Quantitative and qualitative data offer distinct insights that are frequently used in combination to obtain a comprehensive understanding of a population. For instance, when collecting data on annual income (quantitative), additional information on occupations (qualitative) can be gathered to provide more context and detail about the average annual income within each occupation category.[3]

Quantitative data

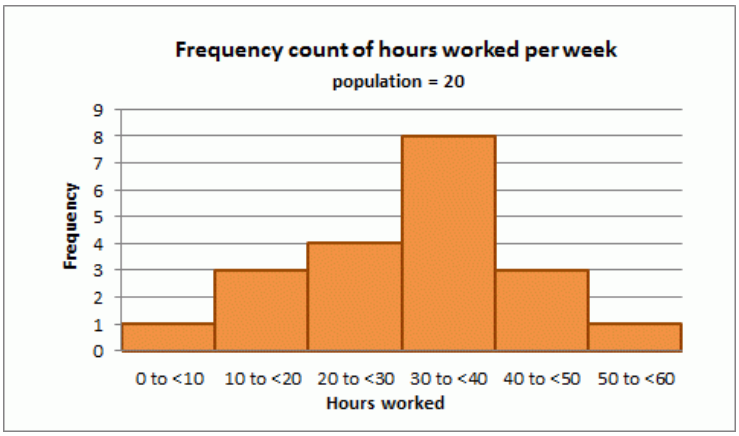


Figure 2.1: Bar chart showcasing frequency of hours worked [3]

Qualitative data

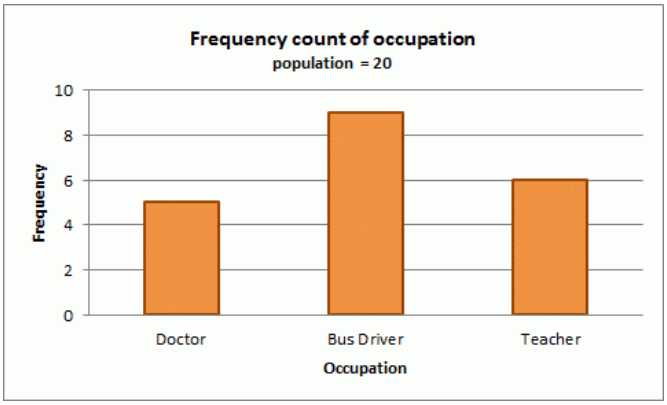


Figure 2.2: Bar chart showcasing frequency of jobs occupied [3]

2.6. Data Sources

Data can be sourced directly or indirectly.

Direct methods of data collection involve collecting new data for a specific study.

Indirect methods of data collection involve sourcing and accessing existing data that were not originally collected for the purpose of the study. [4]

2.6.1. Sourcing direct data

A survey is a process of gathering information from every individual in a population or from a selected group of individuals (a sample) within the population. In a survey, a candidate gives data either about themselves or as a representative of another unit in the population.

Direct methods of data collection include various approaches:

- Surveys administered by an interviewer.
- Self-enumerated surveys.
- In-depth interviews or focus groups.
- Observational studies.
- Experiments and clinical trials. [4]

2.6.1.2. Sourcing indirect data

Indirect data include information like school enrolments, hospital admissions, and vital records. They can be organized for statistical purposes even though they are not initially gathered for statistics. With comprehensive coverage and continuous collection, indirect data allow for long-term comparisons. They can eliminate the need for additional surveys.[4]

3. FREQUENCIES

Frequency is the number of times a particular value for a variable (data item) has been observed to occur.[5]

3.1. Measuring frequency

The frequency of a value can be presented in these ways:

- **Absolute Frequency**: It describes the number of times a specific value for a variable has been observed. It's the simplest way to express frequency.
- **Relative Frequency**: It describes the number of times a particular value for a variable has been observed in relation to the total number of values for that variable. It's calculated by dividing the absolute frequency by the total number of values for the variable.[6]

3.2 Frequency Distribution

Frequency distributions are visual representations that organize and present frequency counts, to make them more readable. They can be showcased either by absolute frequencies or relative frequencies. [6]

3.2.1. How to showcase frequency distributions

Data frequency distributions are presented through tables or graphs. Common methods for displaying frequency distributions include frequency tables, histograms, or bar charts. [6]

3.2.2. Frequency tables

A frequency table is a simple way to display the number of occurrences of a particular value or characteristic.

For example, if we have collected data about height from a sample of 50 children, we could present our findings like the table below.

Height of children		
Height (cm) of children	**Absolute frequency**	**Relative frequency**
120 – less than 130	9	18%
130 – less than 140	10	20%
140 – less than 150	13	26%
150 – less than 160	11	22%
160 – less than 170	7	14%
Total	50	100%

Table 2: Data about the height of children [6]

From this frequency table we can quickly identify information such as 7 children (14% of all children) are in the 160 to less than 170 cm height range, and that there are more children with heights in the 140 to less than 150 cm range (26% of all children) than any other height range.

4. STANDARDS AND CLASSIFICATIONS

Standards are a set of rules used to understand the collection of data and production of statistics. These standards are here to help to understand specific topics and to interpret the gathered data.

They include defining the concept, specifying variables, data collection methods, coding structures, statistical units, recommended question modules, data processing, classification, standard editing, data presentation, output categories, and data interpretation. [7]

4.1. Classifications

Classifications play a crucial role in organizing information by grouping similar pieces together. They are integral to various stages of the statistical cycle, including data collection, processing, presentation, and analysis. It's important for classifications to be exhaustive, covering all

possibilities, and mutually exclusive, ensuring each piece of information falls into only one category. [7]

Classifications should be exhaustive, and mutually exclusive.

For example, Australia, France, Japan, and Vanuatu are some of the categories from the ‘Standard Australian Classification of Countries’ (SACC). This classification is used in the ‘Country of Birth Standard’ where each response for a person's country of birth will belong in one and only one category within the classification.

4.2. Importance of standards in statistics

Standards are used to maintain consistency in the collection and communication of data of the same characteristic. This enables the comparison of data from various sources consistently and strengthens the meaningful comparisons over time. The use of statistical standards offers four key advantages:[7]

- Ensure the Quality of Statistical Outputs
- Create a Meaningful Statistical Picture of Society and Economy
- Reduce Costs
- Improve Transparency

5. METADATA

Metadata are details that give general info and describe data, often known as "data about data". It offers information about the purpose, processes, and methods employed in the data collection.[8]

5.1. Metadata presentation

Metadata gives information about all aspects of data collection from the design phase all the way to communication.

Metadata can encompass details such as the definition and description of the population, the data source, and the methodology employed.[8]

5.2. Importance of metadata

The processes used to collect data and produce statistics may influence the compatibility of the information for different statistical purposes.

Metadata provides information to enable the user to make a decision about whether the data could fit for the required purpose.[8]

6. DATA VISUALISATION

Data visualization is the art of presenting data visually to convey the story of the dataset. It transforms complex information into images and stories and makes it easier to understand.[9]

6.1. Ways data can be visualized.

Data can be communicated visually through various methods:[9]

- **Static Visualization**: This involves the use of graphs and charts, that give a glimpse of the data at a specific point in time.
- **Dynamic Visualization**: Animated presentations that emphasize key information and illustrate movements in the data that provide a more dynamic visual.
- **Interactive Visualization**: This approach enables to customize the visual data, delves deeper into specific data areas, and uses motion to track patterns over time.[9]

6.2. Best use of data visualization

Data visualization is best used when users seek an overview of a dataset rather than delving into specific values. It highlights the narratives within the data or concentrates on selected data for a specific purpose.[9]

7. DATA ANALYSIS

Data analysis is a universal process that encompasses cleaning, transforming, and processing raw data into information. Its primary goal is to assist businesses in making well-informed decisions, avoid risks through insights and statistics presented in charts, images, tables, and graphs.

In everyday life, a simple example of data analysis occurs when we make decisions by assessing past events or predicting future outcomes. It involves analyzing historical or prospective data to inform decision-making.[10]

7.1. Data Analysis Process

Data analysis involves a series of steps from gathering, processing, exploring, and finding insights from information. Here are the key steps in the data analysis process:[10]

1. Data Requirement Gathering:
    - Defines the purpose of the analysis.
    - Identifies the type of data needed.
    - Specifies the data to be analyzed.
2. Data Collection:
    - Gathers data from various sources like case studies, surveys, interviews, questionnaires, observation, and focus groups.
    - Organizes the collected data for analysis.
3. Data Cleaning:
    - Removes irrelevant data and errors.
    - Addresses white spaces, duplicate records, and basic mistakes.
    - Ensures the data is in a clean and usable format.
4. Data Analysis:
    - Utilizes data analysis tools.
    - Applies statistical methods to interpret and understand the data.
5. Data Interpretation:

- Analyzes the results obtained from the data analysis.
- Formulates conclusions and identify patterns or trends.

6. Data Visualization:
    - Presents the information graphically for easy understanding.
    - Uses charts, graphs, maps, bullet points, or other visualizations to compare datasets and observe relationships.

7.2. Data Analysis Techniques

Defining the Objectives:

It is crucial to have an end goal in mind during the data analysis faze. Understanding the questions to answer provides a clear roadmap for the analysis process.[10]

Machine Learning Algorithms:

Machine learning algorithms train models using historical data and assess their performance on new data using the right algorithms, knowledge, and functions.[10]

Text Mining and NLP:

When dealing with textual data, it is best to apply text mining and natural language processing (NLP) techniques. Analyzing sentiment, extracting topics, classifying text, or performing entity recognition to derive insights from unstructured text data is how NLP is used.[10]

7.3. Top Data Analysis Tools

- Tableau Public
- R Programming
- Python
- SAS
- Excel

7.4. Correlation and Causation

In a statistical context, two or more variables are considered related if their values change together, meaning an increase or decrease in one variable corresponds to a change in the other variable (though not necessarily in the same direction). For instance, in the case of "hours worked" and "income earned," there is a relationship if more hours worked are associated with increased income.

Correlation describes the size and direction of this relationship between variables. However, correlation doesn't imply causation. Just because variables are correlated doesn't mean one causes the other. Causation suggests a direct cause-and-effect relationship between two events, while correlation merely indicates that they vary together.

For example, smoking may be correlated with alcoholism, but it doesn't cause alcoholism. On the other hand, smoking causing an increase in the risk of developing lung cancer is an example of a cause-and-effect relationship.[11]

7.4.1. Importance of correlation and causation

The goal of much research or scientific analysis is to assess the extent of the relationship between variables. For instance:

Is there a relationship between a person's education level and their health?

Does pet ownership correlate with longer life?

Did a company's marketing campaign lead to an increase in product sales?

These questions explore potential correlations between variables. If a correlation is found, it may prompt further research to investigate whether one variable causes the other. [11]

7.4.2. Measuring correlation

The correlation between two variables is measured using a Correlation Coefficient, shown by the symbol (r). It describes the degree of relationship between the variables. The coefficient ranges from +1.0 to –1.0.[11]

- A negative correlation coefficient (below 0) proves a negative relationship between variables — when one increases, the other decreases, and vice versa.
- A positive correlation coefficient (above 0) denotes a positive relationship, indicating both variables are moving in the same direction. As one variable decreases, the other also decreases, and when one increases, the other also increases.
- A correlation coefficient of 0 shows no relationship between the variables. One variable can remain constant while the other increases or decreases.

8.PYTHON

A programming language serves as a system for writing computer programs. It is characterized by its syntax (form) and semantics (meaning) and is defined by formal language. A programming language typically has implementations, such as compilers or interpreters, that enable the execution of programs for their desired language.

Python is an example of an interpreted, interactive, object-oriented programming language. It has features like modules, exceptions, dynamic typing, high-level dynamic data types, and classes.[12]

8.1. Advantages of using Python in Data Analytics

Python is widely preferred for data analytics due to its simplicity, easy learning curve, and scalable, flexible nature. It offers a rich collection of libraries for numerical computation and

data manipulation, including NumPy and Pandas, the most popular data analysis tools. Python excels in graphics and data visualization with tools like Matplotlib.[13]

8.2. Python Libraries for Data Analytics

8.2.1 NumPy:

Supports n-dimensional arrays for numerical computing, essential for tasks like linear algebra and Fourier transforms.[13]

8.2.2 Pandas:

Handles missing data, performs mathematical operations, and enables data manipulation.[13]

8.2.3 Matplotlib:

Commonly used for plotting data points and creating interactive visualizations. .[13]

8.2.4 Seaborn:

A data visualization library, ideal for exploratory analysis and creating interactive plots.[14]

8.2.5 RegEx (Regular Expression):

Used for defining search patterns in strings, helpful for string manipulation and validation.[15]

8.2.6 NLTK (Natural Language Toolkit):

A Python library for natural language processing and computational linguistics.[16]

8.2.7 JSON (JavaScript Object Notation):

Python's built-in JSON package facilitates working with JSON data through methods like json.loads().[17]

8.2.8 Collections:

Python's collections module offers specialized container data types for specific programming problems.[18]

8.3 Jupyter Notebook

Jupyter Notebook is an open-source web application and is made to be an interactive computational environment. It contains documents (notebooks) combining code inputs and outputs into a single file. This singular document showcases:[19]

- Visualizations

- Mathematical Equations
- Statistical Modeling
- Narrative Text

Jupyter notebooks contains compatibility with over 40 programming languages, but it has a special emphasis on Python for Data Analysis.

9. CASE STUDY INTRODUCTION

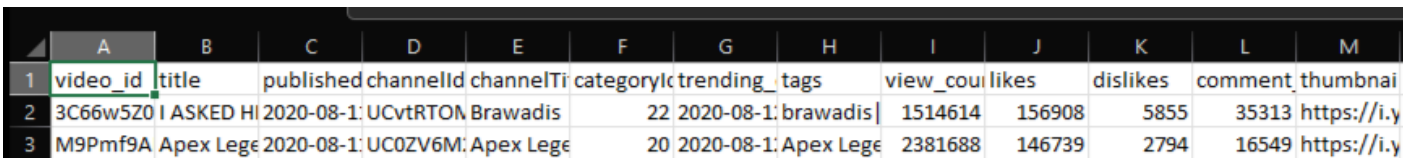
This study explores the factors influencing the virality of YouTube videos and aims to provide deeper insights into what contents push the algorithm. The main goal of this study is to explore and answer questions about the important metrics, strategies, and interactions with viewers that play a role in making a video go viral.

- **Metrics that Matter:**
  - Explore the key metrics that drive views and engagement.
  - Investigate the correlation between likes, comments, and dislikes with the overall popularity.
- **Tag Tactics:**
  - Analyze the effectiveness of tags in boosting video visibility.
  - Understanding the power of strategic tagging and its role in attracting a broader audience.
  - Evaluate whether certain types of content benefit more from specific tags.
- **Title Word Analysis:**
  - Investigate the influence of title words on a video's popularity.
  - Explore the optimal length of a video title for maximum visibility.
- **Settings for Success:**
  - Examine the importance of enabling or disabling comments and ratings.
  - Evaluate the consequences of different settings on viewer interaction and video reach.
- **Genre Dynamics:**
  - Explore the popularity dynamics across different video genres.
  - Provide insights on genre selection based on popularity.

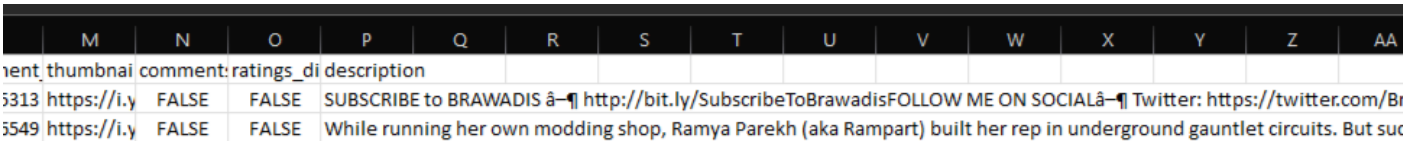
Through data analysis, the aim is to shed light on the important aspects that underly YouTube virality status, providing insights for both academic and content creation.

9.1. The Dataset

The dataset was downloaded on 06.10.2023 and consists of 15 columns and 292’589 rows [20]. In the small snapshots below, it showcases a sample of 2 rows of the dataset.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	video_id	title	published	channelId	channelTi	categoryIc	trending_	tags	view_cou	likes	dislikes	comment_	thumbnai
2	3C66w5Z0	I ASKED HI	2020-08-1	UCvtRTON	Brawadis	22	2020-08-1	brawadis|	1514614	156908	5855	35313	https://i.y
3	M9Pmf9A	Apex Lege	2020-08-1	UC0ZV6M	Apex Lege	20	2020-08-1	Apex Lege	2381688	146739	2794	16549	https://i.y



	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
ıent	thumbnai	comment	ratings_di	description											
5313	https://i.y	FALSE	FALSE	SUBSCRIBE to BRAWADIS â–¶ http://bit.ly/SubscribeToBrawadisFOLLOW ME ON SOCIALâ–¶ Twitter: https://twitter.com/Br											
5549	https://i.y	FALSE	FALSE	While running her own modding shop, Ramya Parekh (aka Rampart) built her rep in underground gauntlet circuits. But suc											

Figure 2: Screenshots of the Trending's Videos Dataset

Although certain column names are self-explanatory, it is essential to provide clarification for the specific column names that will be the focus of this project.

9.1.1 Metadata of the Dataset

Title:

- Definition: The introductory representation of the video that showcases a preview of its content.
- Purpose: To encapsulate the message of the video and entice potential viewers.

Category:

- Definition: The classification of the video.
- Purpose: For easy discovery for users interested in a particular theme or genre.

Views:

- Definition: The count of how many times the video has been watched.
- Purpose: Indicates the level of audience engagement and overall interest in the video.

Likes:

- Definition: The positive feedback metric.
- Purpose: Contributes to the overall positive reception and credibility of the video.

Dislikes:

- Definition: The negative feedback metric.
- Purpose: Gives insights into areas of improvement of the video.

Comments:

- Definition: Viewer-generated responses and discussions related to the video.
- Purpose: Provides a platform for audience interaction, feedback, and community engagement.

All the other columns were either not necessary for this project or served no purpose for further explanation. During the coding stage, they were removed or cut short and were needed further down the project to save time and to make it easier to read the data.

9.2. Data Manipulation Stage

Python was chosen for its accessibility and prompt data analysis capabilities as the programming language for this study. The codes are accessible for review on my GitHub page.

Firstly, the dataset was loaded into Jupyter Notebook by using Python's Pandas library.

```
# Load the dataset from a CSV file into the 'df' DataFrame.
df = pd.read_csv(r'c:\YoutubeProject\US_youtube_trending_data.csv')
```

The columns that won't be needed during this project will be removed, alongside NULL values and data that will not be necessary.

```
# Remove the 'thumbnail_link' column
df.drop('thumbnail_link', axis=1, inplace=True)

# Remove the 'video_id' column
df.drop('video_id', axis=1, inplace=True)

# Remove the 'channelId' column
df.drop('channelId', axis=1, inplace=True)

df = df[df['dislikes']!=0]

df= df[df['tags']!= '[none]']
```

9.3 Correlation Between Likes, Dislike, Comment Count and Views

To ensure the success of a video, it must accumulate a certain number of views within a specific timeframe to be featured on YouTube's trending page. Demonstrating a correlation between the views column and factors such as likes, dislikes, or comments would prove that a video need more than just views to be successful.

In the graph below, the correlation ratios for all variables are displayed.

```
corr=df.corr()
corr
```

	categoryId	view_count	likes	dislikes	comment_count	comments_disabled	ratings_disabled
view_count	-0.055030	1.000000	0.845533	0.696190	0.531732	0.003933	0.017287
likes	-0.083263	0.845533	1.000000	0.659794	0.709162	-0.030042	-0.039864
dislikes	-0.037977	0.696190	0.659794	1.000000	0.521833	0.008759	-0.026937
comment_count	-0.063501	0.531732	0.709162	0.521833	1.000000	-0.019139	-0.008376

Table 3: Dataset highlighting the correlation between likes, dislikes, comments, and views.

Since there is confirmation that views have a strong correlation between likes and dislikes and a light correlation between the accumulated amount of comments(comment_counts), it is safe to say that if particular video has high likes, dislikes or comments, then the video has definitely high amounts of views.

9.4. Finding the Percentage of Tag Usage on Videos

The project began with the belief that specific tags used in popular videos could boost their chances of becoming viral. If a video is already popular and has predetermined its tag usage, considering with the factors like views, likes, and dislikes, it might suggest that pre-determined tags play a role in ascending a video's popularity. To test this hypothesis, the approach was to analyze the most frequently used 10 percent of tags in all videos within certain video categories from the larger population. These tags were then categorized into smaller groups to assess whether the initial hypothesis holds true.

Throughout this hypothesis, 3 genres were categorized: Sports, People & Blogs, and Games. The reason for categorizing them is to see if there are potential similarities or huge differences between different types of categories, according to videos' popularity and the all-around population of all the videos.

Initially, the formatting involved transforming all tags into a list format for each row, removing duplicate entries, and eliminating NULL values.

```
# Defining a function for clean text
def clean_tag(tags):
    return [tag.lower().replace('"', '') for tag in tags.split('|')]

# Applying clean textf function for df
df['tags'] = df['tags'].apply(clean_tag)
```

```
df = df[df['tags'].apply(lambda x: '[none]' not in x)]
```

Then, all stop words were removed, as they served no necessary purpose, and all were taken out.

```
# Function to clean and preprocess a title
def clean_title(title):
    # Convert to lowercase
    title = title.lower()

    # Remove special characters, numbers, and punctuation
    title = re.sub(r'[^a-zA-Z\s]', '', title)

    # Tokenize the title
    tokens = title.split()

    # Remove stopwords
    stop_words = set(stopwords.words('english'))
    tokens = [word for word in tokens if word not in stop_words]

    # Rejoin the cleaned tokens
    cleaned_title = ' '.join(tokens)

    return cleaned_title

# Apply the clean_title function to the 'title' column
df['cleaned_title'] = df['title'].apply(clean_title)
df.head(2)
```

In the following code, a sepereate dataset that groups the tags frequencies in a descending order is created.

```
# Flatten the list of lists in the 'tags' column into a single list of tags
all_tags = [tag for tags in dfSport['tags'] for tag in tags]

# Count the frequency of each tag
tag_frequencies = Counter(all_tags)

# Create a DataFrame to store the tag names and their frequencies
tag_frequencies_df = pd.DataFrame(tag_frequencies.items(), columns=['Tag',
'Frequency'])

# Sort the DataFrame by frequency in descending order
tag_frequencies_df = tag_frequencies_df.sort_values(by='Frequency',
ascending=False)

# Reset the index of the DataFrame
```

```
tag_frequencies_df = tag_frequencies_df.reset_index(drop=True)
tag_frequencies_df
```

	Tag	Frequency
0	highlights	455
1	football	451
2	nba	400
3	basketball	295
4	sports	268
...	...	...
14006	tiger woods responsive	1
14007	us-sport	1
14008	woods accident	1
14009	woods crash	1
14010	greensboro swarm	1

Table 4: Dataset showcasing the frequencies of tags used on the sports genre.

For this dataset, tags with a frequency of only 1 were removed, and the top 10 percent most frequent tags were selected.

```
# Calculate the total number of unique tags
total_unique_tags = len(tag_frequencies_df)

# Calculate the top 10% threshold
top_10_percent_threshold = int(total_unique_tags * 0.10)

# Filter the tags to get the top 10% most frequent tags
top_10_percent_tags = tag_frequencies_df.head(top_10_percent_threshold)

# Remove tags with a frequency of 1
top_10_percent_tags = top_10_percent_tags[top_10_percent_tags['Frequency'] !=
1]

# Reset the index of the DataFrame
top_10_percent_tags = top_10_percent_tags.reset_index(drop=True)
top_10_percent_tags
```

	Tag	Frequency
0	highlights	455
1	football	451
2	nba	400
3	basketball	295
4	sports	268
...	...	...
1396	media	5
1397	jj redick podcast	5
1398	#badminton	5
1399	#bwf	5
1400	technique	5

Table 5: Dataset showcasing the frequencies of tags used on the sports genre cleaned up.

By taking the top 10 percent dataset into consideration, a comparison was made with the original dataset from other genres to identify if a video contains a tag featured in the top 10 percent dataset. If a tag is present in a row containing the 10 percent most frequent tags, it will be counted and accumulated; otherwise, it will not be counted.

```
# Extract the tags from the 'Tag' column of the top_10_percent_tags DataFrame
top_10_percent_words = top_10_percent_tags['Tag'].tolist()

# Create a set of the top 10% most used words
top_10_percent_words_set = set(top_10_percent_words)

# Function to check if any word from the top 10% is present in the tags list
def contains_top_words(tags):
    for tag in tags:
        if tag in top_10_percent_words_set:
            return 'Used'
    return 'Not Used'

# Apply the function to the 'tags' column of the dfSport DataFrame
dfSport['TagUsage'] = dfSport['tags'].apply(contains_top_words)
```

I also then count how many of the top 10 percent are actually featured in a video.

```
def count_top_tags(tags):
    count = 0
    for tag in tags:
        if tag in top_10_percent_tags['Tag'].values:
            count += 1
    return count
dfSport['top_10_percent_count'] = dfSport['tags'].apply(count_top_tags)
dfSport.head(2)
```

	title	channelTitle	tags	view_count	likes	dislikes	categoryName	LikesDislikeSum	TagUsage	top_10_percent_count
0	Shannon reacts to Kyle Kuzma's game-winning sh...	Skip and Shannon: UNDISPUTED	[fox, fox sports, fs1, fox sports 1, undispute...	540613	7155	308	Sports	7463	Used	31

Table 6: Dataset of the sports genre that now includes the frequency of the number of times a top 10 percent most frequent tag is used.

In the following graphs below, you will see the the distirbution between the top 10 precent most frequent used in all videos, compared to the rest of the 90 precent.

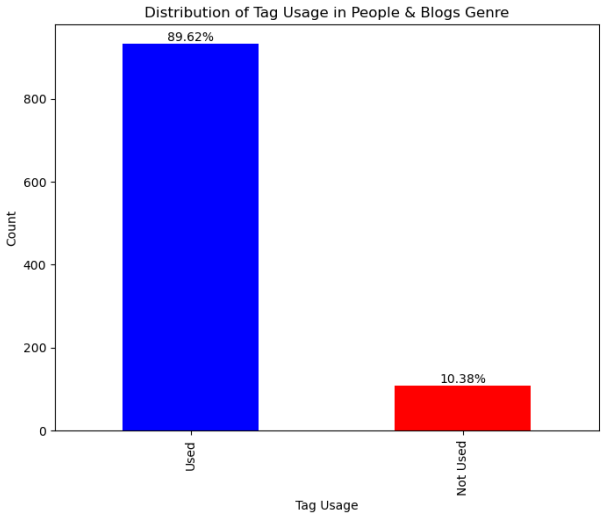


Figure 3: Distribution of most popular tags compared to the other 90 percent in the People & Blogs genre.

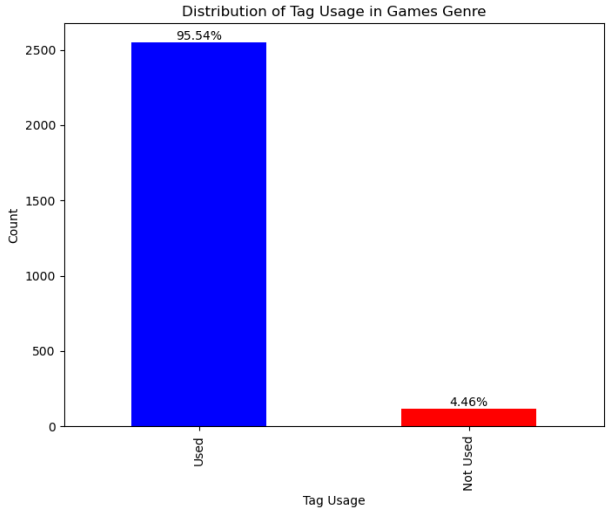


Figure 4: Distribution of most popular tags compared to the other 90 percent in the Games genre.

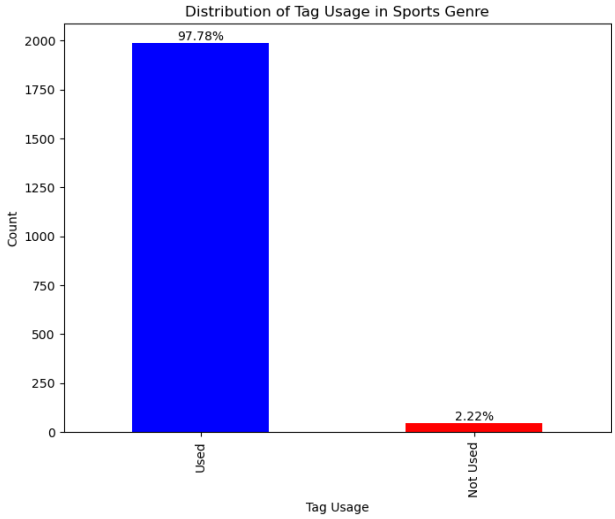


Figure 5: Distribution of most popular tags compared to the other 90 percent in the Sports genre.

According to the graphs above, for the Sports, People and Blogs, and Games genre the top 10 percent most frequently used tags are present on almost all videos from these genres.

9.5 Correlation Between Tag Usage and Views, Likes, and Dislikes

The main goal is now to see if there is a correlation between the most frequent tag usages and with likes, dislikes views or any reaction. To see this, I counted the frequencies of the top 10 precent tag usages on each video in the Games, Sports, and People & Blogs genre. On the image below you can see the correlation between the frequencies and the likes, dislikes, views and reactions.

Reactions from Games, Sports, and People & Blogs

```
Correlation between 'top_10_percent_count' and 'LikesDislikeSum     ': 0.10
Correlation between 'top_10_percent_count' and 'LikesDislikeSum     ': 0.21
Correlation between 'top_10_percent_count' and 'LikesDislikeSum     ': 0.12
```

Sports

```
Correlation between 'top_10_percent_count' and 'view_count': 0.19
Correlation between 'top_10_percent_count' and 'likes': 0.21
Correlation between 'top_10_percent_count' and 'dislikes': 0.12
```

Games

```
Correlation between 'top_10_percent_count' and 'view_count': 0.09
Correlation between 'top_10_percent_count' and 'likes': 0.10
Correlation between 'top_10_percent_count' and 'dislikes': 0.05
```

People & Blogs

```
Correlation between 'top_10_percent_count' and 'view_count': 0.12
Correlation between 'top_10_percent_count' and 'likes': 0.13
Correlation between 'top_10_percent_count' and 'dislikes': -0.01
```

As you can see, the percentage of the correlation is too low to confirm that tags play a role in determining a videos' success.

Even though the percentage of the most popular tags are high and are used, they do not affect any reactions according to this dataset.

9.6. Correlation Between Title Length and View Count

The next goal of this research was to see whether a video's title length plays a role in its popularity. In this dataset, I countedthe number of characters in each individual video's title and compared it with the views to identify any correlation. The results are presented in the image below.

```
# Add a new column for title length
df['title_length'] = df['title'].apply(len)

# Calculate the correlation between title length and view count
correlation = df['title_length'].corr(df['view_count'])

# Print the correlation value
print(f"Correlation between Title Length and View Count: {correlation:.2f}")
```

✓ 0.0s

```
Correlation between Title Length and View Count: -0.05
```

According to the dataset used, the length of the title plays no role on having a high view count.

9.7. Finding the Percentage of Title Words on Videos

Exploring the significance of particular keywords in video titles on views and engagement was a main goal in this thesis. This method involved identifying the top 10 percent most frequently used words in video titles within the sports genre and determining their presence compared to the remaining 90 percent.

To perform this analysis, the dfSports dataframe was reset, and manipulations were applied. An empty dataframe was created to accumulate and categorize the words found in video titles.

```
column_list = dfSport['cleaned_title'].tolist()

column_list = [string.split() for string in column_list]

flat_list = [word for sublist in column_list for word in sublist]

word_counts = Counter(flat_list)
```

Subsequently, from this initially empty list, another list was generated to compile the titles along with the frequency of each word's usage in those titles..

```
# Create a DataFrame to store the tag names and their frequencies
title_frequencies_df = pd.DataFrame(word_counts.items(), columns=['Title',
'Frequency'])

# Sort the DataFrame by frequency in descending order
```

```
title_frequencies_df     =     title_frequencies_df.sort_values(by='Frequency',
ascending=False)

# Reset the index of the DataFrame
title_frequencies_df = title_frequencies_df.reset_index(drop=True)
title_frequencies_df
```

	Title	Frequency
0	highlights	770
1	vs	565
2	game	261
3	full	218
4	sports	190
...	...	...
3871	supercopa	1
3872	bruno	1
3873	fernandes	1
3874	punching	1
3875	overtake	1

Table 7: Dataset showcasing the frequencies of words used on title from the sports genre.

Following that, the top 10 percent most frequent words were selected, and any words with a frequency of 1 were excluded.

```
# Calculate the total number of unique tags
total_unique_title = len(title_frequencies_df)

# Calculate the top 10% threshold
top_10_percent = int(total_unique_title * 0.10)

# Filter the tags to get the top 10% most frequent tags
top_10_percent_titles = title_frequencies_df.head(top_10_percent)

# Remove tags with a frequency of 1
top_10_percent_titles                                                    =
top_10_percent_titles[top_10_percent_titles['Frequency'] != 1]

# Reset the index of the DataFrame
top_10_percent_titles = top_10_percent_titles.reset_index(drop=True)
top_10_percent_titles
```

	Title	Frequency
0	highlights	770
1	vs	565
2	game	261
3	full	218
4	sports	190
...	...	...
382	ariel	8
383	opening	8
384	hotspur	8
385	thunder	8
386	see	8

Table 8: Dataset showcasing the frequencies of words used on title from the sports genre cleaned up.

Now, the presence of the top 10 percent words is determined in each video using the code below.

```
# Create a set of the top 10% most frequent words
top_10_percent_words_set = set(top_10_percent_titles['Title'])

# Function to check if any word from the top 10% is present in the cleaned titles
def contains_top_words(title):
    title = title.lower()
    words = title.split()
    for word in words:
        if word in top_10_percent_words_set:
            return 'Used'
    return 'Not Used'

# Apply the function to the 'cleaned_title' column of the DataFrame
dfSport['top_words_usage'] = dfSport['cleaned_title'].apply(contains_top_words)
```

And similar to the tags process, frequencies of the top 10 percent used words from each video were calculated.

```
def count_top_words(title):

    count = 0
    words = title.split()
    for word in words:
        if word in top_10_percent_titles['Title'].values:
            count += 1
    return count
```

```
dfSport['top_10_percent_count'] =
dfSport['cleaned_title'].apply(count_top_words)
dfSport.head()
```

	title	channelTitle	view_count	likes	dislikes	categoryName	cleaned_title	title_length	LikesDislikeSum	top_words_usage	top_10_percent_count
40	Shannon reacts to Kyle Kuzma's game-winning sh...	Skip and Shannon: UNDISPUTED	540613	7155	308	Sports	shannon reacts kyle kuzmas gamewinning shot im...	100	7463	Used	7
43	Giannis Gets Ejected After Headbutting Mo Wagner	Bleacher Report	756814	8278	331	Sports	giannis gets ejected headbutting mo wagner	48	8609	Used	2
47	TRAIL BLAZERS at MAVERICKS | FULL GAME HIGHLIG...	NBA	937569	10160	405	Sports	trail blazers mavericks full game highlights a...	67	10565	Used	7
65	'Don't mess with Dame Dolla' - Stephen A. reac...	ESPN	791489	13242	325	Sports	dont mess dame dolla stephen reacts lillards b...	96	13567	Used	6

Table 7: Dataset of the sports genre that now includes the frequency of the number of times a top 10 percent most frequent title word is used.

In the graph below, you will see how videos that contain the most frequent 10 percent words compare to videos that use the rest of the 90 percent.

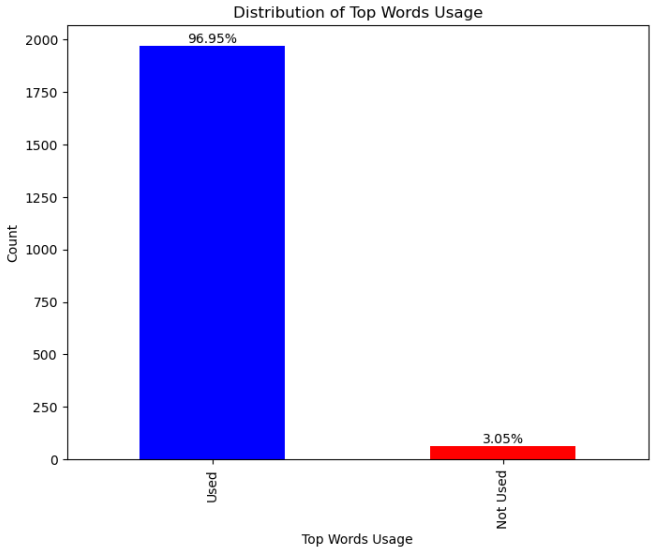


Figure 6: Distribution of most popular title words compared to the other 90 percent in the Sports genre.

It is safe to say that the majority of the most frequently used words are displayed on every video title.

9.8. Correlation Between Title Words and View Count, Likes and Dislikes.

To analyze the correlation between words and their impact on views, likes, or dislikes, the frequencies of the top 10 percent words for each video were examined and compared to the corresponding likes, dislikes, and views to identify potential correlations.

In the image below is the result for the percentage of the correlation.

```
Correlation between 'top_10_percent_count' and 'view_count': -0.01
Correlation between 'top_10_percent_count' and 'likes': -0.20
Correlation between 'top_10_percent_count' and 'dislikes': -0.01
```

As evident from the results, there is minimal correlation. Thus, based on this dataset, it can be said that the usage of specific words does not play a role in determining a video's popularity.

9.9. Performance of Videos with Comments Enabled/Disabled and Ratings Enabled/Disabled

To illustrate the impact of enabling or disabling comments and ratings on a video's popularity, YouTube provides content creators with the option to manually choose whether their videos should have comments or ratings enabled or disabled. If a content creator decides to disable comments, viewers will be unable to comment and therefore potentially reducing the video's popularity, as indicated by correlations explored in chapter 9.3. The same goes gpr disabling ratings; if a creator chooses to turn off the ability to like or dislike a video, its popularity may significantly decrease, given the earlier correlations that highlight the significance of likes and dislikes in determining a video's success.

To illustrate the impact of turning off the ability to comment or react under a video, the data frame was reverted to its original format and categorized into four groups: videos with comments enabled, videos with comments disabled, videos with ratings enabled, and videos with ratings disabled.

```
df_comments_disabled= df[df['comments_disabled'] == True]

df_comments_enabled= df[df['comments_disabled'] == False]
```

```
df_ratings_enabled= df[df['ratings_disabled'] == False]

df_ratings_disabled= df[df['ratings_disabled'] == True]
```

9.10 Videos with Comments Enabled/Disabled Comparison with the Most Popular 10 Percent

From the entire dataset, only 317 videos out of 15 thousand videos have comments disabled. So, the comparison focused on the top 10 percent best-performing videos with comments enabled and disabled. Logical comparisons were made based on their accumulated views, likes, and dislikes. All likes, dislikes, and views were accumulated separately for videos with comments enabled and disabled, respectively.

```
df_comments_disabled = df_comments_disabled.sort_values(by='view_count',
ascending=False)

df_comments_disabled31 = df_comments_disabled.head(31)
```

```
df_comments_enabled = df_comments_enabled.sort_values(by='view_count',
ascending=False)

df_comments_enabled31 = df_comments_enabled.head(31)
```

```
viewsTotalDisabled = df_comments_disabled31['view_count'].sum()
viewsTotalEnabled = df_comments_enabled31['view_count'].sum()

likesTotalDisabled = df_comments_disabled31['likes'].sum()
likesTotalEnabled = df_comments_enabled31['likes'].sum()

dislikesTotalDisabled = df_comments_disabled31['dislikes'].sum()
dilikesTotalEnabled = df_comments_enabled31['dislikes'].sum()
```

Then, the correlation was calculated between likes, dislike, and views for both parts.

```
For Comments Enabled
Correlation between 'likes' and 'view_count': 0.51
Correlation between 'dislikes' and 'view_count': 0.50
Correlation between 'likes' and 'dislikes': 0.28

For Comments Disabled
Correlation between 'likes' and 'view_count': 0.32
Correlation between 'dislikes' and 'view_count': 0.21
Correlation between 'likes' and 'dislikes': 0.67
```

The correlation values are inconsistent, and their understanding may not be meaningful due to the limited dataset containing only 31 instances. To gain a clearer understanding of which option is more beneficial, visualizations were created to compare the views, likes, and dislikes of the top 10 percent best-performing videos with comments enabled or disabled.

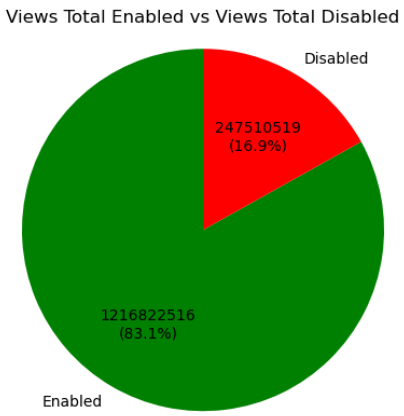


Figure 7: Total views comparison of the 10 percent best performing videos between videos with comments enabled and disabled.

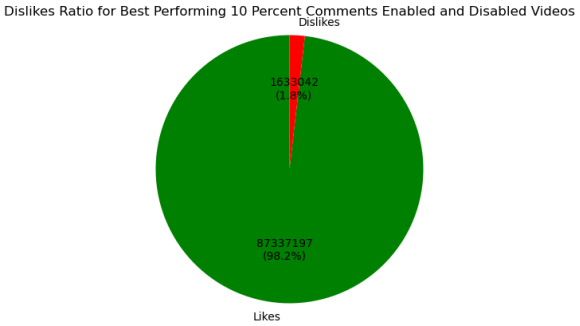


Figure 8: Total dislikes comparison of the 10 percent best performin

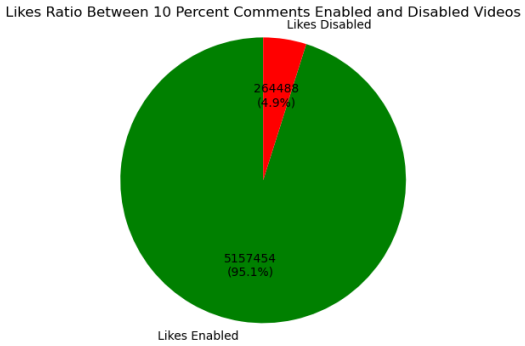


Figure 9: Total likes comparison of the 10 percent best performing videos between videos with comments enabled and disabled.

The graph implies that videos with comments enabled generally are more likely to be popular.

9.11 Videos with Comments Enabled/Disabled Comparison by Most Popularity

Instead of comparing the top 10 percent most popular videos, the analysis directly focused on the most popular videos within the 317 instances with comments disabled. The graphs below illustrate the results obtained through this approach.

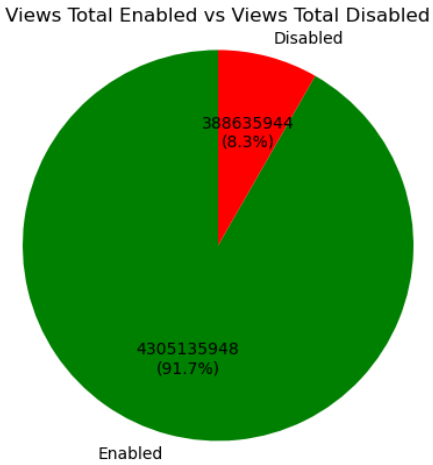


Figure 10: Total views comparison between videos with comments enabled and disabled.

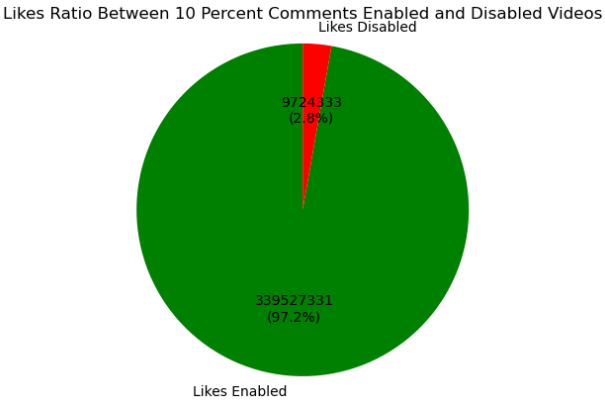


Figure 11: Total likes comparison between videos with comments enabled and disabled.

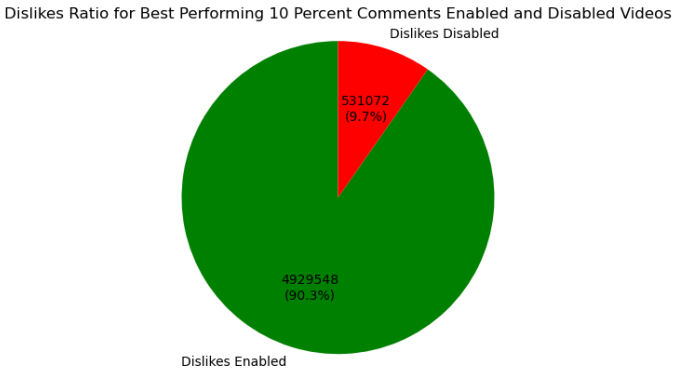


Figure 12: Total dislikes comparison between videos with comments enabled and disabled.

In these graphs, it is shown again that videos that have comments enabled tend to be more popular and more successful in the end.

9.12. Correlation Between Videos with Comments Enabled and Titles and Tags

At last, the aim was to determine if there exists a correlation between specific title words and tags for videos that have comments enabled. The following codes illustrate the percentage of correlation between them.

	view_count	likes	dislikes	comment_count	comments_disabled	top_10_percent_count	top_10_percent_count_title
view_count	1.000000	0.779049	0.581371	0.600531	NaN	0.079995	0.034690
likes	0.779049	1.000000	0.505020	0.730848	NaN	0.060737	-0.007781
dislikes	0.581371	0.505020	1.000000	0.409149	NaN	0.035414	0.027702
comment_count	0.600531	0.730848	0.409149	1.000000	NaN	0.017379	-0.006674
comments_disabled	NaN	NaN	NaN	NaN	NaN	NaN	NaN
top_10_percent_count	0.079995	0.060737	0.035414	0.017379	NaN	1.000000	0.142042
top_10_percent_count_title	0.034690	-0.007781	0.027702	-0.006674	NaN	0.142042	1.000000

Table 9: Dataset showcasing the correlations between views, likes and dislikes of every video with enabled comments.

This confirms that there exists no correlation between tags or title words between videos with comments enabled.

9.13. Success Of a Video with Ratings Turned on Or Off

To determine whether a video should enable or disable its ratings, four distinct data frames were created: one with both comments and ratings enabled, one with only comments enabled and ratings disabled, another with comments disabled and ratings enabled, and the last one with both comments and ratings disabled. To compare them, the total views were accumulated for each data frame. Given that ratings are disabled in three out of the four data frames, comparing their likes and dislikes would result in misleading conclusions. Therefore, only the views were compared, and the results are presented in the graph below.

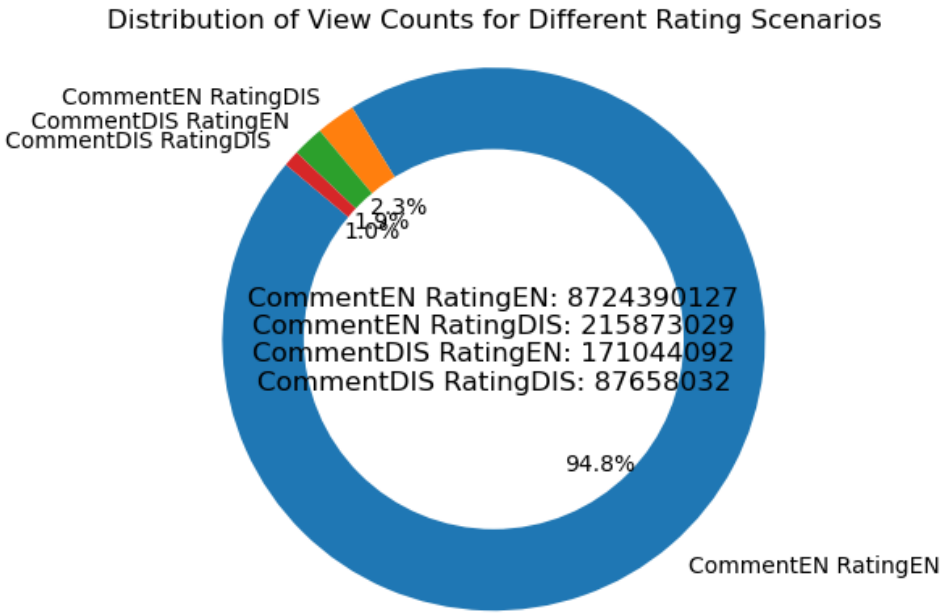


Figure 13: Comparisons of total views between videos with comment and ratings enabled, comments disabled and ratings enabled, comments enabled and ratings disabled, and comments and ratings disabled.

The graph shows that videos with comments and ratings enabled are watched way more than the rest. This proves that reaction, giving a like or dislike and the ability to comment under a video, is crucial for a video's success and should be always the key priority from content creators to let viewers know give a reaction to a video.

9.14. Determining The Most Popular Genre

Understanding the dominant genres in the market is crucial for identifying potential areas for increased popularity. To determine the most popular genre, each genre was divided into different data frames, and the total views, likes, and dislikes were calculated for each.

```
['Music' 'Comedy' 'News & Politics' 'Entertainment' 'Science & Technology'
 'People & Blogs' 'Sports' 'Autos & Vehicles' 'Howto & Style'
 'Film & Animation' 'Gaming' 'Travel & Events' 'Education'
 'Nonprofits & Activism' 'Pets & Animals']
```

```
df_grouped = df.groupby('categoryName')


category_dataframes = {}

for category, group in df_grouped:
    category_dataframes[category] = group

for category, df_category in category_dataframes.items():
    globals()[f"{category.lower()}_df"] = df_category
```

```
genres = []
total_views = []
total_likes = []
total_dislikes = []


for category, df_category in category_dataframes.items():
    genres.append(category)
    total_views.append(df_category['view_count'].sum())
    total_likes.append(df_category['likes'].sum())
    total_dislikes.append(df_category['dislikes'].sum())

combined_df = pd.DataFrame({'Genre': genres, 'TotalViews': total_views,
'TotalLikes': total_likes, 'TotalDislikes': total_dislikes})


combined_df
```

	Genre	TotalViews	TotalLikes	TotalDislikes
0	Autos & Vehicles	196987192	10020465	152911
1	Comedy	726880579	64206192	748466

	Genre	TotalViews	TotalLikes	TotalDislikes
2	Education	256350941	18182389	387310
3	Entertainment	3530764251	229808483	3454624
4	Film & Animation	629212488	36140930	563045
5	Gaming	3072071851	204862429	3354687
6	Howto & Style	333581577	20711285	573927
7	Music	5685826532	602017058	7745656
8	News & Politics	657502900	9888095	1048837
9	Nonprofits & Activism	17995032	2257812	14121
10	People & Blogs	1172882605	85962360	1676042
11	Pets & Animals	39804002	2316825	30547
12	Science & Technology	826373812	37124066	862624
13	Sports	2464493236	58002264	1572814
14	Travel & Events	41094738	3169085	37407

Table 10: Dataset showcasing the accumulated views, likes and dislikes of every video genre.

When compared visually, we get these graphs:

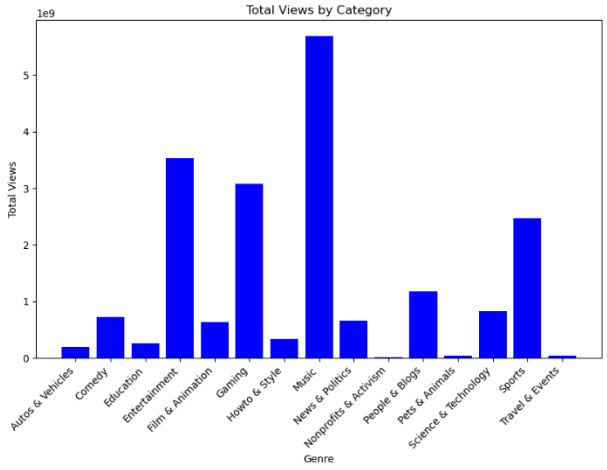


Figure 14: Distribution of views comparison between videos genres.

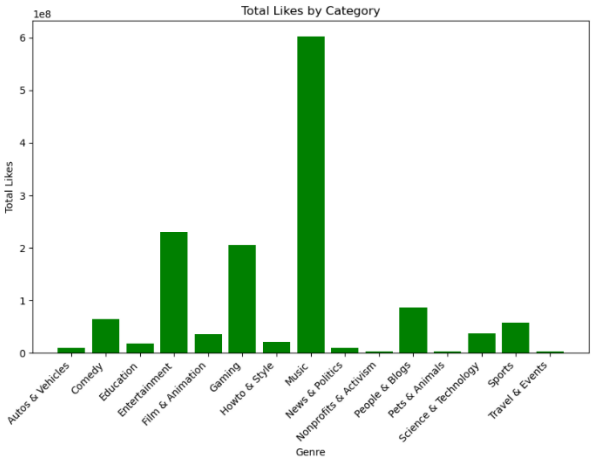


Figure 15: Distribution of likes comparison between videos genres.

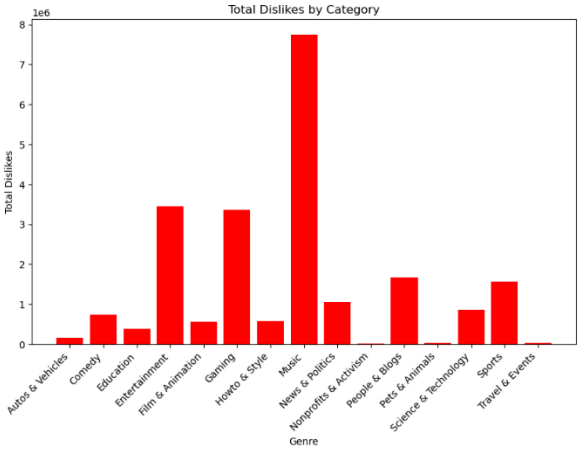


Figure 16: Distribution of dislikes comparison between videos genres.

According to the data frames, the Music genre dominates the trends page while genres like Non-practical & Activism, Pets & Animals and Travel & Events categories are rarely going viral.

10. CONCLUSION

This exploration into YouTube’s video popularity revealed several key insights. The correlation analysis shed light into the relationships between various factors and a video's success. Tags, titles, and enabling comments and ratings all played crucial roles in determining a video's visibility and engagement.

Despite the dominance of the most frequent tags in popular videos, this study showcases that they do not significantly contribute to ascend a video into viral status. Similarly, these findings reveal that title lengths and words, even within the top 10 percent frequency, do not hugely impact a video's popularity.

This research explained the importance of keeping comments and ratings enabled. Disabling these features can lead to a reduction in audience interaction, an important element for a video's popularity. Audience reactions, such as comments, likes, and dislikes, were identified as influential factors in a video's success.

Finally, the genre analysis indicates that, while the Music genre dominates the market, genres like Non-practical & Activism, Pets & Animals, and Travel & Events are less likely to achieve viral status and may be best avoided to create viral content.

REFERENCES

[1] OECD Glossary of Statistical Terms. OECD. 2008. p. 119. *ISBN 978-92-64-025561*.

(Accessed Date: 20.12.2023)

[2] Australian Bureau of Statistics (April 2019) *Statistical Terms and Concepts – Data Definition,* ABS Website, (Accessed Date:25.12.2023)

[3] Australian Bureau of Statistics (April 2019) *Statistical Terms and Concepts – Quantitative and Qualitative Data,* ABS Website, (Accessed Date:25.12.2023)

[4] Australian Bureau of Statistics (April 2019) *Statistical Terms and Concepts –Methods of sourcing data ,* ABS Website, (Accessed Date:25.12.2023)

[5] Australian Bureau of Statistics (April 2019) *Statistical Terms and Concepts – Describing Frequencies Definition ,* ABS Website, (Accessed Date: 25.12.2023)

[6] Australian Bureau of Statistics (April 2019) *Statistical Terms and Concepts - How to show a frequency distribution,* ABS Website, (Accessed Date: 25.12.2023)

[7] Australian Bureau of Statistics (April 2019) *Statistical Terms and Concepts - Standards and classifications - Statistical standards,* ABS Website, (Accessed Date: 25.12.2023)

[8] Australian Bureau of Statistics (April 2019) *Statistical Terms and Concepts – Metadata Definition* , ABS Website, (Accessed Date:25.12.2023)

[9] Australian Bureau of Statistics (April 2019*) Statistical Terms and Concepts – Data Visualization Definition*, ABS Website, (Accessed Date:25.12.2023)

[10] Simplelearn (December 2023) - *Data Analysis Methods Process Types – What is Data Analysis*, (Accessed Date:25.12.2023)

[11] Australian Bureau of Statistics (April 2019) *Statistical Terms and Concepts –Correlation and causation- The difference between correlation and causation ,* ABS Website, (Accessed Date:25.12.2023)

[12] Python FAQ - *https://github.com/python/cpython/blob/main/Doc/faq/general.rst#what-is-python-good-for* , (Accessed Date:19.12.2023)

[13] Simplelearn (July 2023) – *Data Analytics Tutorial – Data Analytics With Python – Why Data Analytics with Python*, (Accessed Date:10.12.2023)

[14] Simplelearn (January 2023) – *An Interesting Guide to Visualize Data Using Python Seaborn–* , (Accessed Date: 10.12.2023)

[15] w3schools – *Python RegEx*, (Accessed Date:18.12.2023)

[16] My Great Learning (August 2023) – *Natural Language Toolkit (NLTK) Tutorial with Python* , (Accessed Date:18.12.2023)

[17] w3schools - *Python JSON*, (Accessed Date:18.12.2023)

[18] Python FAQ - *https://github.com/python/cpython/blob/3.12/Lib/collections/__init__.py*, (Accessed Date:25.12.2023)

[19] Towards Data Science (December 2020) - *Everything You Need To Know About Jupyter Notebooks*, (Accessed Date:25.12.2023)

[20] Kaggle - *YouTube Trending Video Dataset (updated daily)*, (Accessed Date:25.12.2023)

RESUME

Name: Rotinda Öner

Surname: Öner

Birth Date: 15.11.2001

Birth Place: Schaffhausen(CH)

University: Yildiz Technical University

GitHub: *https://github.com/cRotinda*