Social Data Science Project

The Impact of Video Length and Interactivity on YouTube Channel Size and Video Popularity:

Analysis of Greece's Most-Viewed Channels

Objective:

To explore how video length, content type, and interactivity affect channel size and video popularity among the most-viewed YouTube channels in Greece using data from the YouTube API.

Byun et al. (2023). The effect of YouTube comment interaction on video engagement: focusing on interactivity centralization and creators' interactivity. Available at: https://www.emerald.com/insight/content/doi/10.1108/oir-04-2022-0217/full/html

TABLE OF CONTENTS

3 | Dataset Descriptions

3.1. | Channels Dataset

This dataset contains information about 37 YouTube channels, selected after excluding big brands, record companies, and artists from the most-viewed YouTube channels. The data was retrieved using the YouTube API and includes key metrics for each channel, such as subscriber counts, total views, and the number of uploaded videos. Additionally, the dataset provides descriptive metadata, including channel descriptions and a unique playlist identifier for videos uploaded by each channel.

Size: 37 entries, 7 columns

Purpose: To analyze channel-level information, including performance metrics and metadata, for YouTube creators

Column Name	Description	Data Type	Variable Type
Channel_name	The name of the YouTube channel	object(string)	categorical, nominal
Description	A brief description of the channel provided by the creator	object(string)	textual
Subscribers	The total number of subscribers to the channel	int64	numerical, continuous

Column Name	Description	Data Type	Variable Type
ViewCount	The total number of views across all videos uploaded by the channel	int64	numerical, continuous
Total_Videos	The total number of videos uploaded by the channel	int64	numerical, continuous
Playlist_id	A unique identifier for the playlist containing all videos from the channel	object(string)	categorical
Description_c	A cleaned version of the Description column with special characters removed and all text converted to lowercase	object(string)	textual

In [10]:
```
channel_data = pd.read_csv('channelData.csv')
channel_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 37 entries, 0 to 36
Data columns (total 7 columns):
 #   Column         Non-Null Count  Dtype 
---  ------         --------------  ----- 
 0   Channel_name   37 non-null     object
 1   Description    34 non-null     object
 2   Subscribers    37 non-null     int64 
 3   ViewCount      37 non-null     int64 
 4   Total_Videos   37 non-null     int64 
 5   Playlist_id    37 non-null     object
 6   Description_c  33 non-null     object
dtypes: int64(3), object(4)
memory usage: 2.2+ KB
```

In [11]:
```
channel_data.head()
```

Out[11]:

	Channel_name	Description	Subscribers	ViewCount	Total_Videos	P
0	Greekonomics	Πως η Οικονομία επηρεάζει την Κοινωνία!\n\nΤα ...	231000	13154190	60	UU1KjWRBCUGvxDkr
1	Dat Lilly	Νέο βίντεο κάθε Κυριακή ❤️ \nThank's for being h...	521000	100038173	190	UU9WYita8NlpXTcm
2	Pavlos Makris	Εδώ για να σε διασκεδάσω!😊 \nΠάτα το Like & Sub...	12500	1611596	33	UUhWPS3NiUzeRmh8
3	Eponimos	ναι.	392000	95878785	513	UUFOasUEk9Pkr8Ye
4	Unboxholics	TIME WELL WASTED.\nGaming | Tech | Cinema | En...	1070000	442841695	1546	UUjBCvQBVTh4XjPwt

In [12]:
```
channel_data.describe().applymap(lambda x: f"{x:,.2f}")
```

```
/var/folders/qt/nkv93n510wlcddjjxc58klyr0000gn/T/ipykernel_4780/2082133113.py:1: Futur
eWarning: DataFrame.applymap has been deprecated. Use DataFrame.map instead.
  channel_data.describe().applymap(lambda x: f"{x:,.2f}")
```

Out[12]:

	Subscribers	ViewCount	Total_Videos
count	37.00	37.00	37.00
mean	286,638.51	130,538,841.70	377.57
std	351,845.68	257,231,098.58	355.63
min	25.00	3,182.00	5.00
25%	89,800.00	20,633,094.00	128.00
50%	156,000.00	66,007,002.00	236.00
75%	301,000.00	122,569,978.00	547.00
max	1,760,000.00	1,536,553,865.00	1,546.00

3.2. | Videos Dataset

This dataset contains detailed information about 13,780 YouTube videos, retrieved using the YouTube API. It provides video-specific metadata, including titles, publication dates, descriptions, tags, and performance metrics such as views, likes, dislikes, and comment counts. Additionally, the dataset includes derived columns such as video duration in seconds, categorical video length, and a cleaned version of the video description.

Size: 13,780 entries and 20 columns
Purpose: To analyze video-level metadata and performance metrics across the 37 YouTube channels

Column Name	Description	Data Type	Variable Type	
Id	A unique identifier for each video	object(string)	categorical	
Title	The title of the video	object(string)	textual	
Published_Date	The date and time the video was published	object(string)	temporal	
Description	A brief description of the video provided by the creator	object(string)	textual	
Tags	The total number of views the video has received	object(string)	A list of tags assigned to the video by the creator	textual
Views	The total number of views the video has received	int64	numerical, continuous	
Likes	The total number of likes the video has received	int64	numerical, continuous	
Dislikes	The total number of dislikes the video has received.	int64	numerical, continuous	
Comments	The total number of comments on the video	int64	numerical, continuous	
Channel_Id	A unique identifier for the channel that uploaded the video	object(string)	categorical	

Column Name	Description	Data Type	Variable Type
Playlist_Id	A unique identifier for the playlist containing the video	object(string)	categorical
Video_Length	The duration of the video in ISO 8601 format	object(string)	
Published_Year	The year the video was published	int64	numerical, discrete
Published_Month	The month the video was published	int64	numerical, discrete
Description_c	A cleaned version of the Description column with special characters removed and all text converted to lowercase	object(string)	textual
Video_Length_Seconds	The duration of the video in seconds	int64	numerical, continuous
Video_Length_HH_MM_SS	The duration of the video formatted as HH:MM:SS	int64	temporal
Comments_Presence	Indicates whether comments are present (1 for yes, 0 for no)	int64	binary
Video_Length_Category	A categorical label for the video length (Short, Medium, Long, Super Long)	object(string)	categorical, ordinal
Channel_Username	The username of the channel that uploaded the video	object(string)	categorical
Most_Popular_Word_Count	A variable showing the count of the most popular word in descriptions per observation	int64	numerical, discrete
Popular	A new binary variable was created to classify whether a video is popular or not, based on the median of all views	int64	binary

```
In [13]: video_data = pd.read_csv('VideoData.csv')
         video_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 13780 entries, 0 to 13779
Data columns (total 24 columns):
 #   Column                  Non-Null Count  Dtype  
---  ------                  --------------  -----  
 0   Id                      13780 non-null  object 
 1   Title                   13780 non-null  object 
 2   Published_Date          13780 non-null  object 
 3   Description             10446 non-null  object 
 4   Tags                    13780 non-null  object 
 5   Views                   13780 non-null  int64  
 6   Likes                   13780 non-null  int64  
 7   Dislikes                13780 non-null  int64  
 8   Comments                13780 non-null  int64  
 9   Channel_Id              13780 non-null  object 
 10  Playlist_Id             13780 non-null  object 
 11  Video_Length            13780 non-null  object 
 12  Published_Year          13780 non-null  int64  
 13  Published_Month         13780 non-null  int64  
 14  Description_c           10445 non-null  object 
 15  Video_Length_Seconds    13780 non-null  int64  
 16  Video_Length_HH_MM_SS   13780 non-null  object 
 17  Comments_Presence       13780 non-null  int64  
 18  Video_Length_Category   13771 non-null  object 
 19  Channel_Username        13780 non-null  object 
 20  Most_Popular_Word_Count 13780 non-null  int64  
 21  Popular                 13780 non-null  int64  
 22  Predicted_Probability   13780 non-null  float64
 23  Log_Comments            13780 non-null  float64
dtypes: float64(2), int64(10), object(12)
memory usage: 2.5+ MB
```

In [14]:
```
video_data.head(2)
```

Out[14]:

	Id	Title	Published_Date	Description	Tags	Views	Likes	Dislik
0	qlifbbutkl0	Το Παγκόσμιο Μέλλον του Χρήματος | Greekonomic...	2024-11-22	Ένα ταξίδι στο μέλλον του χρηματοπιστωτικού συ...	[]	172515	14860	
1	JLQNJPg9lH4	Η "Κολομβία" της Ευρώπης | Greekonomics #45	2024-09-22	Ευχαριστώ την Freedom24 που στηρίζει το κανάλι...	[]	549229	41163	

2 rows × 24 columns

In [15]:
```
video_data.describe().applymap(lambda x: f"{x:,.2f}")
```

```
/var/folders/qt/nkv93n510wlcddjjxc58klyr0000gn/T/ipykernel_4780/2593052769.py:1: Futur
eWarning: DataFrame.applymap has been deprecated. Use DataFrame.map instead.
  video_data.describe().applymap(lambda x: f"{x:,.2f}")
```

Out[15]:

	Views	Likes	Dislikes	Comments	Published_Year	Published_Month	Vide
count	13,780.00	13,780.00	13,780.00	13,780.00	13,780.00	13,780.00	
mean	350,445.39	11,166.23	0.00	507.21	2,021.12	6.82	
std	2,175,514.05	26,985.09	0.00	2,641.94	2.94	3.50	
min	0.00	0.00	0.00	0.00	2,011.00	1.00	
25%	59,636.00	2,224.75	0.00	50.00	2,019.00	4.00	
50%	156,875.00	7,013.00	0.00	162.00	2,022.00	7.00	
75%	341,164.00	14,364.25	0.00	425.25	2,023.00	10.00	
max	125,085,728.00	1,411,134.00	0.00	161,427.00	2,024.00	12.00	

3.3. | Comments Dataset

This dataset contains comments and threads data collected from a random sample of 100 YouTube videos belonging to 37 channels. This sample offers valuable insights into audience interactions and creator presence in comment threads, while is anonymized to ensure user privacy for users.

Size: 31,115 entries and 12 columns
Purpose: To analyze user engagement and creator-audience interactions in the comments section

Column Name	Description	Data Type	Variable Type
Video_ID	A unique identifier for the video to which the comment belongs	object(string)	categorical
Channel_ID	A unique identifier for the channel that uploaded the video	object(string)	categorical
User_name	An anonymized identifier for the individual who posted the comment	object(string)	categorical
Comment	The text of the comment posted by a user	object(string)	textual
Comment_likes	The total number of likes the comment received	int64	numerical, continuous
Published_Date	The date and time the comment was published	object(string)	temporal
Total_Replies	The total number of replies to the comment	int64	numerical, continuous
Creator_Replies	The total number of replies made by the channel creator to the comment	int64	numerical, continuous
Published_Year	The year the comment was published	int64	numerical, discrete
Published_Month	The month the comment was published	int64	numerical, discrete
Comment_p	A cleaned version of the Comment column with special characters removed and all text converted to lowercase	object(string)	textual
Replies_Presence	Indicates whether replies are present (1 for yes, 0 for no)	int64	binary

Since **comments are classified as personal data under Article 4(1) of Regulation 2016/679 (GDPR)**, their **processing in this project is conducted under the legal basis of Article 6(1)(f)**, which allows processing for legitimate interests. In this case, the legitimate interest pertains to conducting academic research as part of a specific exam project. A **random sample of 62,037 comments, including usernames (personal data)**, from 100 videos was collected in adherence to the **principle of data minimization**, as outlined in **Article 5(1)(c)**, ensuring that only data necessary for the research purpose was processed.

Recognizing that usernames constitute personal data that could potentially identify individuals, **pseudonymization technique was implemented to safeguard data security and ensure user anonymity**. This aligns with the requirements of **Article 32(1)(a)**, which emphasizes the importance of technical measures to protect personal data, and the guidance provided in **Recital 26**, which underscores the value of pseudonymization in mitigating risks associated with personal data processing.

source:https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32016R0679

In [16]:

```
comments_data = pd.read_csv('CommentsDataP.csv')
comments_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 31115 entries, 0 to 31114
Data columns (total 12 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   Video_ID          31115 non-null  object
 1   Channel_ID        31115 non-null  object
 2   User_name         31112 non-null  object
 3   Comment           31099 non-null  object
 4   Comment_likes     31115 non-null  int64 
 5   Published_Date    31115 non-null  object
 6   Total_Replies     31115 non-null  int64 
 7   Creator_Replies   31115 non-null  int64 
 8   Published_Year    31115 non-null  int64 
 9   Published_Month   31115 non-null  int64 
 10  Comment_p         30288 non-null  object
 11  Replies_Presence  31115 non-null  int64 
dtypes: int64(6), object(6)
memory usage: 2.8+ MB
```

In [17]:

```
comments_data.head(2)
```

Out[17]:

	Video_ID	Channel_ID	User_name	Comment	Comment_lik
0	pxv2GXvEFqY	UCFOasUEk9Pkr8YeJxGc88Lw	@andrychristoforou5522	Cfv 0:15	
1	pxv2GXvEFqY	UCFOasUEk9Pkr8YeJxGc88Lw	@georgeanas10	Φίλε δε τραγούδησες της Ελλάδας	

In [18]:

```
comments_data.describe().applymap(lambda x: f"{x:,.2f}")
```

```
/var/folders/qt/nkv93n510wlcddjjxc58klyr0000gn/T/ipykernel_4780/2618206656.py:1: Futur
eWarning: DataFrame.applymap has been deprecated. Use DataFrame.map instead.
  comments_data.describe().applymap(lambda x: f"{x:,.2f}")
```

Out[18]:

	Comment_likes	Total_Replies	Creator_Replies	Published_Year	Published_Month	Replies
count	31,115.00	31,115.00	31,115.00	31,115.00	31,115.00	
mean	4.65	0.14	0.00	2,021.56	7.52	
std	60.76	0.59	0.00	1.97	3.97	
min	0.00	0.00	0.00	2,013.00	1.00	
25%	0.00	0.00	0.00	2,021.00	4.00	
50%	0.00	0.00	0.00	2,022.00	8.00	
75%	1.00	0.00	0.00	2,023.00	12.00	
max	3,427.00	5.00	0.00	2,024.00	12.00	

4. | Ethics Reflections

To reflect on the **ethical aspects of my project**, I followed the four principles as a guide to identify and address any ethical uncertainties (Salganik, 2019).

YouTube API & Ethics

Respect for Persons
Specifically, in this project, since obtaining consent regarding personal data (usernames) was not feasible for the collection of publicly available YouTube comments, the following measures were taken to respect individual autonomy:

- Only publicly available data was collected, ensuring no breach of privacy through unauthorized access
- Usernames were pseudonymized to protect individual identities and minimize the risk of re-identification
- No manipulation or interaction occurred with the users whose data was collected, ensuring no disruption to their online activity

Beneficence
Futhermore, to align with beneficence, the project aimed to minimize potential harms and maximize the benefits by:

- The data collected was strictly limited to what was necessary to achieve the research objectives, following the principle of data minimization under GDPR Article 5(1)(c)
- Pseudonymization was applied to further reduce the risk of re-identification and protect user privacy
- The findings of the study are intended to contribute to academic knowledg

Justice
Additionally, regarding principle of justice was upheld by ensuring fairness in the collection and processing of data:

- The collection and processing of comments were conducted under the legal basis of GDPR Article 6(1)(f), which allows processing for legitimate interests
- The random sampling of comments ensured that no specific group was over-represented or disproportionately impacted

Respect for Law and Public Interest

Lastly,the project complied with GDPR and maintained transparency to ensure accountability:

- The project adhered to GDPR Articles 4(1), 5(1)(c), and 6(1)(f), ensuring lawful and ethical processing of personal data
- Publicly available YouTube data was used, and care was taken to respect the platform's terms of service
- The research process was documented thoroughly, ensuring that methods and ethical considerations could be reviewed and scrutinized

References:

Salganik, (2019). Bit By Bit: Social Research in the Digital Age, available at: https://www.bitbybitbook.com/en/1st-ed/ethics/principles/

European Parliament and Council of the European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). Official Journal of the European Union, L119, 1–88. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679