1. Defined Business Problem/Question

Business Problem:
Understanding the factors that drive video performance on YouTube to optimise content creation.

Key Questions to Address:

- What characteristics of YouTube videos (e.g. category, country per category, views per category, yearly earnings per category).
- How does audience engagement vary across different content categories?
- What trends exist in viewer retention and how can they inform content strategy?
- What video category has the highest yearly earnings?
- What countries have the highest subscribers?
- Does the number of population per country affect the video views?
- Employment rate(how does it influence the content creation).
- Demographic Information**:** Incorporate viewer demographic data to personalise content strategies.

Objective:
Provide actionable insights to content creators, and platform managers to enhance video performance, increase audience engagement, and drive channel growth.

---

2. Data Overview

Data Source:
The dataset utilised for this project is sourced from Kaggle and specifically pertains to YouTube video statistics.

Dataset Description:

Number of Records: Approximately 500,000 videos.

A**nd we have 28 columns stated below**

Key Attributes:

1. Rank: rank of the youtuber
2. Youtuber: name of the youtube channel
3. Subscribers: numbers of subscribers per youtube channel
4. video views: Total numbers of views per video
5. Category: Numerical identifier representing the video category (e.g., Music, Education).

6. Title: Title of the video
7. Uploads: number of videos uploaded by the youtuber.
8. Country: Country where subscribers live in.
9. Abbreviation:
10. Channel_type: category of videos being uploaded on a channel.(e.g., Music, Education).
11. Video_views_rank: top viewed videos in the world.
12. Country_rank: top viewed videos per country.
13. Channel_type_rank: Top channels according to the category of videos being uploaded on it.
14. Video_views_for_the_last_30_days: numbers of video clicks per month
15. Lowest_monthly_earnings: least paid youtuber per month.
16. highest_monthly_earnings: highest paid youtuber per month.
17. Lowest_yearly_earnings: least paid youtuber per year.
18. highest_yearly_earnings: Highest paid youtuber per year.
19. Subscribers_for_last_30_days: numbers of subscribers earned by a youtube channel per month.
20. Created_year: year creation of the youtube channel
21. Created_month: month creation of the youtube channel
22. Created_date: day creation of the youtube channel
23. Gross tertiary education enrollment (%):
24. Population: population of countries.
25. Unemployment rate: rate of unemployment per country.
26. Urban_population: number of people living in an urban area per country.
27. Latitude: measurement of a location north or south of the Equator for each country.
28. Longitude: measurement of location east or west of the prime meridian at Greenwich for each country.

Data Quality Considerations:

- **Missing Values:** Some records may have missing values in fields like dislikes or comment counts.

1.	category	46 nulls.
2.	Country	122 nulls.
3.	channel_type	30 nulls.
4.	video_views_rank	1 null.
5.	channel_type_rank	33 nulls.
6.	video_views_for_the_last_30_days	56 nulls.
7.	subscribers_for_last_30_days	337 nulls.
8.	created_year	5 nulls.
9.	Population	123 nulls.
10.	Latitude	123 nulls.
11.	Longitude	123 nulls.

- **Type of data in the dataset:**
  1. Object
  2. Float
  3. Integer

---

3. SQL Database Design

To effectively manage and query the YouTube statistics data, a relational database design is proposed. The design includes the following tables and their relationships:

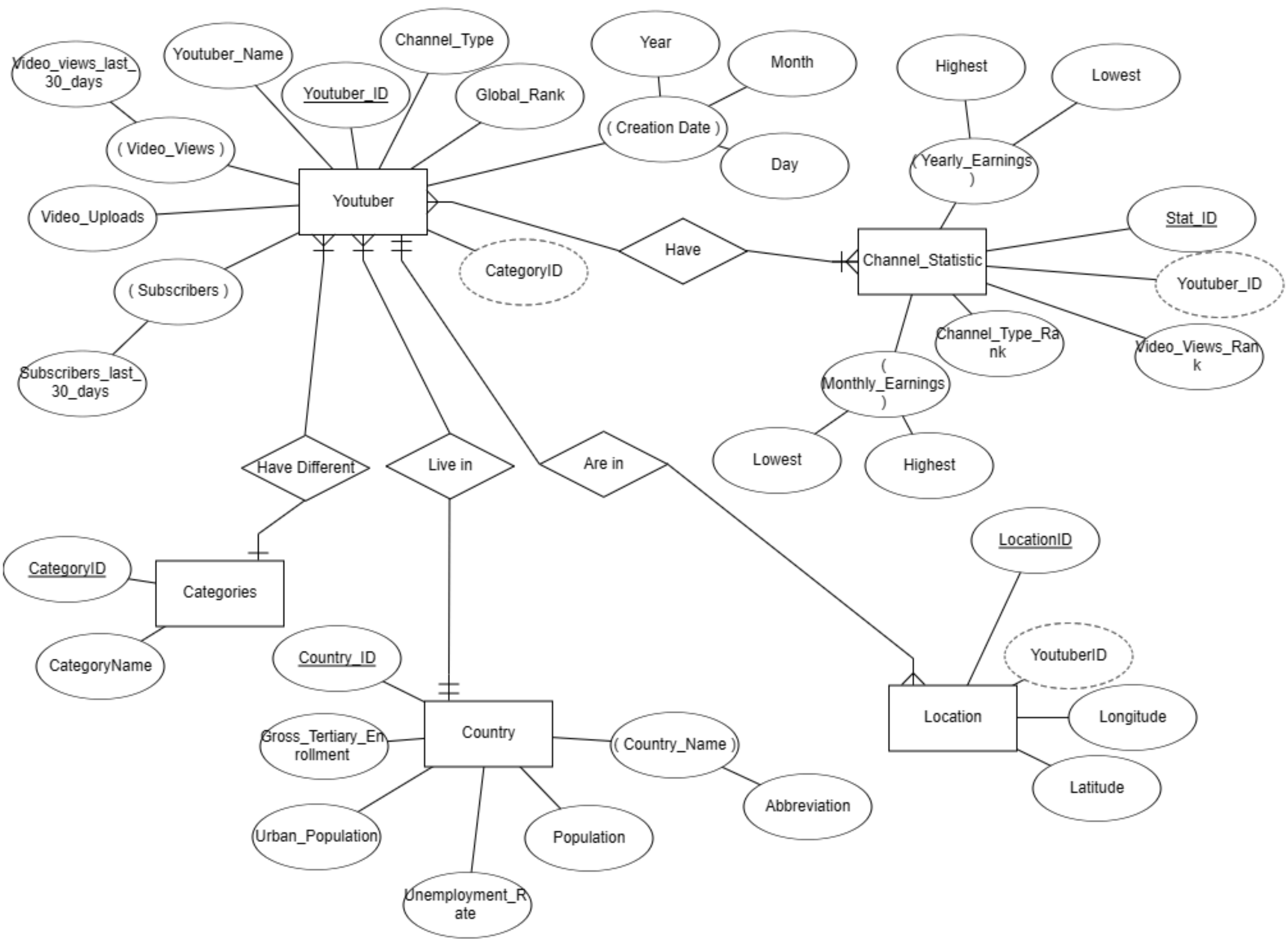
Tables and Relationships:

1. **Categories**
   - **CategoryID** (Primary Key)
   - **CategoryName**
2. **Channel Statistics**
   - **StatID** (Primary Key)
   - **YoutuberID**(Foreign Key referencing Youtuber(YoutuberID))
   - **Videos View Rank**
   - **Channel Type Rank**
   - **Lowest Monthly Earnings**
   - **Highest Monthly Earnings**
   - **Lowest Yearly Earnings**
   - **Highest Yearly Earnings**
3. **Youtuber**
   - **YoutuberID** (Primary Key)
   - **YoutuberName**
   - **Channel Type**
   - **Global Rank**
   - **Subscribers**
   - **Video Uploads**
   - **Video Views**
   - **Video Views last 30 days**
   - **Subscriber Last 30 days**
   - **Created Year**
   - **Created Month**
   - **Created Day**
   - **CategoryID** (Foreign Key referencing Categories.CategoryID)
4. **Country**
   - **CountryID (**Primary Key**)**
   - **Country Name**
   - **Abbreviation**
   - **Population**
   - **Unemployment Rate**
   - **Urban Population**

- **Gross Tertiary Enrollment**

5. **Location**
    - **LocationID** (Primary Key)
    - **YoutuberID**
    - **Latitude**
    - **Longitude**

Entity-Relationship Diagram (ERD):



Explanation:

- **YouTuber - Categories**: A YouTuber can belong to different categories (e.g., gaming, lifestyle, education).
- **YouTuber - Channel_Statistic**: Each YouTuber has their respective statistics like earnings, ranks, and other metrics.
- **YouTuber - Country**: YouTubers live in a specific country, and countries can host multiple YouTubers.

- **Country - Location**: Countries have locations identified by latitude and longitude.

Normalization Considerations:

- Ensuring data redundancy is minimized by separating categories, channels, and engagement metrics into distinct tables.
- Facilitating efficient queries and scalability for large datasets.

---

4. Visualizations and Insights

Visualizations Generated:

1. **Video views by Category:**
    - **Type:** Donut Chart
    - **Insight:** Identifies which categories have the highest views.
2. **Number of Subscribers by Category:**
    - **Type:** Clustered Bar Chart
    - **Insight:** Identifies the number of subscribers by Category.
3. **Yearly Earnings by Category:**
    - **Type:** Matrix Table
    - **Insight:** Highlights the most highest earning videos by category,allowing analysis of common categories among top performers.
4. **Countries with You-Tube**
    - **Type:** Slicer
    - **Insight:** Highlights different variables of countries in terms of category, video views,subscriptions and yearly earnings.
5. **Subscribers By Country:**
    - **Type:** Map
    - **Insight:** Highlights distribution of You-tube Subscribers globally.
6. **Geographical Distribution of Yearly earnings:**
    - **Type:** Map
    - **Insight:** Visualizes where highest yearly earnings are distributed.
7. **Unemployment rate By Country:**
    - **Type:** Tree Map
    - **Insight:** Displays the rate of unemployment rate by country and its influence on content creation,views and subscription.

Insights Derived:

1. **High-Performing Categories:**
    - Categories like **Music** and **Entertainment** consistently show higher view counts and engagement rates compared to others.
2. **Engagement Metrics Correlation:**

    - A strong positive correlation exists between views and subcriptions indicating that higher subscription counts generally lead to increased engagement and views.
3. **Subscriber Influence:**
    - Channels with larger subscriber bases experience faster growth in views and engagement per video, highlighting the importance of building a loyal audience.
4. **Geographical Viewer Trends:**
    - A significant portion of views originates from **North America** and **India**, informing targeted marketing strategies.

---

5. Next Steps for the Project

Addressing Data Gaps:

1. **Competitive Analysis:**
    - **Opportunity:** Compare channel performance against competitors to identify market positioning.
    - **Action:** Collect and analyse data from similar channels within the same categories.
2. **Content Quality Metrics:**
    - **Opportunity:** Assess qualitative aspects such as video production quality, thumbnail effectiveness, and title optimization.
    - **Action:** Utilise image and text analysis tools to evaluate thumbnails and titles for attractiveness and relevance.
3. **Monetization Data:**
    - **Opportunity:** Explore the relationship between video performance and revenue generation.
    - **Action:** If accessible, include data on ad revenue, sponsorship deals, and merchandise sales linked to video performance.

Final Recommendations:

- **Enhance Data Collection:** Expand the dataset to include additional relevant metrics that can provide deeper insights.
- **Leverage Advanced Analytics:** Utilise machine learning and AI to uncover patterns and predictions that traditional analysis might miss.
- **Focus on Content Optimization:** Apply insights to refine content strategies, aiming for higher engagement and viewer satisfaction.
- **Continuous Monitoring:** Establish ongoing data analysis practices to adapt to changing viewer behaviours and platform algorithms.