CNIT 372 Milestone 1

Group Members: Hannah Mathew, Carissa Bauerband, Ben Fan, Sophia Mahedy
Option: C

Data we plan to use: Trending YouTube Video Statistics
https://www.kaggle.com/datasets/datasnaek/youtube-new?select=CAvideos.csv
Global YouTube Statistics 2023
https://www.kaggle.com/datasets/nelgiriyewithana/global-youtube-statistics-2023

Insights derived from data: The data on YouTube creators, including their subscriber count, total views, country of residence, and video upload count, can be analyzed to gauge content engagement, geographical trends, and category preferences. These insights can inform content strategies, target demographics, and identify popular content genres on the platform.

Why listed insights are important: These insights are important because they can help us determine what videos are watched when, which is also impacted by the YouTube algorithm. We want to be able to track what types of videos are popular at what points throughout the year, such as advertising videos being popular near Black Friday and Christmas.

How we plan to design/store the data: We are going to set the data up in 3 tables: Creator, Video, and Category. We have listed the tables below with the columns, data types, constraints, and description. We are going to be flexible and if we need more tables we will add them at a later date.

Explain your rationales: We went to Kaggle and looked at the example data sources and what columns were used. Additionally, we visited YouTube to observe the various sections and categories that were displayed on the screen.

Outline tables we plan to use:

Creator:

Column Name	Data Type	Constraints	Description
Username	VARCHAR2(255)	Primary Key	Unique identifier for the user (YouTube username)
NumOfSubscribers	INTEGER	Default 0	Number of subscribers the user has
JoinDate	DATE	NOT NULL	When the creator joined youtube
TotalViews	INTEGER	Default 0	Total views across all videos by the creator
Country	VARCHAR2(50)	Not Null	Country of residence
TotalVideos	INTEGER	Default 0	Total number of videos uploaded by the creator

Video Table:

Column Name	Data Type	Constraints	Description

VideoID	VARCHAR2(150)	Primary Key	Unique identifier for each video
CategoryID	INTEGER	Foreign Key	References the Category table
Username (FK)	NVARCHAR2(50)	Not Null	Username of the video uploader
PublishedOn	DATE	Not Null	Date the video was published
NumOfComments	INTEGER	Default 0	Number of comments on the video
NumOfLikes	INTEGER	Default 0	Number of likes on the video

Category Table:

Column Name	**Data Type**	**Constraints**	**Description**
CategoryID	INTEGER	Primary Key	Unique identifier for each category
Genre	VARCHAR2(50)	Not Null	Genre associated with the category
CategoryDescription	VARCHAR2(255)	NOT NULL	Brief Describe the Category

Describe your teamwork: We unanimously decided to do group C due to our lack of interest in wanting to do research on content creators and content consumers.
Each member helped with coming up with columns and attributes for the three tables. We divided the other questions and wrote responses which were then checked by the other members. Additional information was added for questions if members thought the answers were too vague.

CNIT 372 Milestone 2

Question 1: Which country (out of USA and India) has the higher average number of subscribers per creator?

This question looks at averages, which we studied earlier in the semester. We would group by country. It is important because we can determine which countries produce the most followed Youtubers. This is good data and it's useful because then we can track popularity based on country. Youtube creators would also like this information because it relates to their metrics.

Question 2: Which creators have the most videos uploaded, and how many videos do they have?

For this question you would count the number of videos uploaded by each creator and then identify the creator with the highest count. This is important because it could help marketers, collaborators or analytics teams to identify creators' work ethic, production process, the reasoning why they are posting videos, content scheduling, identify trends on their platform, etc.

Question 3: How many videos, on average, do creators upload in a week in each category?

For this question you would find the number of videos uploaded by each creator and then calculate the average number of uploads per week for each category. This question is important as it informs content creators / social media managers about their upload frequency compared to other creators and helps them plan for better content scheduling and engagement strategies.

Question 4: Which categories have the highest ratio of comments to views, indicating strong audience engagement?

For this question we would calculate the ratio of comments to views for each category and then find the categories with the highest ratios which indicates the strongest engagement. This is important because it showcases categories where the audience engagement is high which allows creators to make strategies for similar engagement. This could be useful for content strategists, content creators and possibly marketer.s

Question 5: What is the average number of comments per video for creators in each country?

For this question you will be calculating the mean comment count for creators in each county. The creator, the video, and number of comments on each video will be the information that will be pulled. The avg function will be used to calculate the average number of comments per video in each country. This is important because it can be used to see the potential growth opportunities along with the audience's engagement.

Question 6: Which creators have the highest engagement rate (likes and comments) per video?

This question calculates the average engagement rate which in this case are likes and comments for each video for each creator(s). This is important because it can show which creators have highly engaged audiences and could help marketers, potential collaborators or influencer marketing teams to pair up with the creator.

Question 7: Which categories have the highest total views and the lowest total views globally?

We could aggregate the total number of views by category, so that we are then able to identify which categories have the highest and lowest total views globally. By using SQL's aggregate functions such as SUM to calculate the total views for each category, and using ORDER BY to sort the results, it is easy to see top and bottom figures on a list of results. This information is important because knowing what genres of content matter most with audiences from around the world can dictate how media practitioners plan their output. Knowing which genres attract the most views will help planners better align what they do with those already proven winning forms in order to expand reach or maintain engagement if desired.

Question 8: Which creators have the highest like-to-subscribers ratio for their videos?

To figure out which creators have the highest like-to-subscribers ratio, we'd take the total likes for each video (or perhaps an average across all of a creator's videos) and divide that number by the creator's total number of subscribers. This ratio gives us an understanding of how engaged a creator's subscribers are. A high like-to-subscribers ratio can indicate that the creator's content resonates strongly with their subscriber base. This metric is beneficial for both creators and potential advertisers. For creators, a high ratio suggests that their audience is highly engaged and appreciates their content. For advertisers, partnering with a creator with a high like-to-subscriber ratio can provide a better return on investment as the audience is more likely to engage with promotional content.

Question 9: What are the trends in video engagement (likes, comments) during the first half of 2017 November versus the second half of 2017 November in the United States?

We will use the SUM function to get the number of likes and comments in the first half of November and compare it to the values of the same function but in the second half of the month. We will use a JOIN statement to get data out of the Video table and the Creator table. We can also sort by country, which in this case will be the United States. It is important because creators might want to know if certain time periods/weeks are good for engagement. It is useful especially because many Youtubers depend on their video engagement for money, and bad days matter in their income. If they know a week or portion of a month will be a slow day, they might want to post another video before that to try to compensate for low views.

Question 10: How does audience engagement (likes and comments) differ between YouTube creators in the United States and India?

Compare the United States and India in terms of the average number of likes and comments per video for creators in each country. This can help us understand the level of engagement in these two regions.

Describe your teamwork: how did you come up with the 10 questions, list the contribution of each team member.

We discussed as a group what types of questions we wanted for our project. We brainstormed questions and wrote them in the document. Afterwards we discussed what needed to be changed afterwards. We then picked random questions and wrote how they related to the class and the importance of them. After completing all the questions we each reviewed them and added any additional information that we deemed necessary.

Carissa: Did questions 1, 9, and 10.
Ben: Did questions 7 and 8.
Sophia: Did question 5
Hannah: Did questions 2,3,4,6

CNIT 372 Milestone 3

GitHub repository: https://github.com/BenFan1002/CNIT372

Question 1:

```
SELECT
   Country,
   AVG(NumOfSubscribers) AS AverageSubscribers
FROM
   Creator
WHERE
   Country IN ('India', 'United States')
GROUP BY
   Country
ORDER BY
   AverageSubscribers;
```

Question 2:

```
SELECT
        Username,
        Count(TotalVideos) as TotalNumVideos
FROM
        Creator
GROUP BY
        Username
ORDER BY
        Count(TotalVideos) DESC
```

Question 3:

```
SELECT
   c.Genre,
   v.Username,
   COUNT(*) / (TRUNC(SYSDATE) - TRUNC(MIN(v.PublishedOn))) AS AvgVideosPerWeek
FROM
   Categorys c
JOIN
   Video v ON c.CategoryID = v.CategoryID
GROUP BY
   c.Genre, v.Username;
```

Question 4:

```
SELECT
   SUM(v.NumOfComments) AS TotalComments,
   SUM(cr.TotalViews) AS TotalViews,
```

```
    SUM(v.NumOfComments) / SUM(cr.TotalViews) AS CommentToViewRatio
FROM
    Categorys c
LEFT JOIN
    Video v ON c.CategoryID = v.CategoryID
LEFT JOIN
    Creator cr ON cr.username = v.username
ORDER BY
    CommentToViewRatio DESC;
```

Question 5:

```
Select c.Country, AVG(v.NumOfComments)  as AvgComments
From creator c
Inner Join video v
ON c.Username = v.Username
Group by c.Country
Order by AVG(v.NumOfComments);
```

Question 6:

```
Select c.Username, AVG((v.NumOfComments +  v.NumOfLikes) / (c.TotalVideos)) AS EngagementRate
From creator c
Inner Join video v
ON c.Username = v.Username
Group by c.Username
Order by AVG(v.NumOfComments);
```

Question 7:

```
WITH RankedCategories AS (
    SELECT
        Categorys.CategoryID,
        Categorys.Genre,
        SUM(Creator.TotalViews) AS TotalCategoryViews,
        ROW_NUMBER() OVER (ORDER BY SUM(Creator.TotalViews) DESC) AS RankDesc,
        ROW_NUMBER() OVER (ORDER BY SUM(Creator.TotalViews)) AS RankAsc
    FROM
        Video
    INNER JOIN
        Creator ON Video.Username = Creator.Username
    INNER JOIN
        Categorys ON Video.CategoryID = Categorys.CategoryID
    GROUP BY
        Categorys.CategoryID,
```

```
        Categorys.genre
)
SELECT CATEGORYid, genre, TotalCategoryviews FROM RankedCategories
WHERE RankDesc = 1 OR RankAsc = 1;
```

Question 8:

```
SELECT *
FROM RankedCategories
WHERE RankDesc = 1 OR RankAsc = 1;
Question 8:
SELECT
  cr.Username,
  SUM(v.NumOfLikes) / NULLIF(cr.NumOfSubscribers, 0) AS LikeToSubscriberRatio
FROM
  Creator cr
JOIN
  Video v ON cr.Username = v.Username
GROUP BY
  cr.Username,
  cr.NumOfSubscribers
ORDER BY
  LikeToSubscriberRatio DESC;
```

Question 9:

```
SELECT
    'First Half' AS Period,
    SUM(NumOfLikes) AS TotalLikes,
    SUM(NumOfComments) AS TotalComments
FROM
    Video v
    JOIN Creator c ON v.Username = c.Username
WHERE
    v.PublishedOn BETWEEN TO_DATE('2017-11-01', 'YYYY-MM-DD') AND TO_DATE('2017-11-15',
'YYYY-MM-DD')
    AND c.Country = 'United States'

UNION ALL

SELECT
    'Second Half' AS Period,
    SUM(NumOfLikes) AS TotalLikes,
    SUM(NumOfComments) AS TotalComments
FROM
    Video v
```

```
    JOIN Creator c ON v.Username = c.Username
WHERE
    v.PublishedOn BETWEEN TO_DATE('2017-11-16', 'YYYY-MM-DD') AND TO_DATE('2017-11-30',
'YYYY-MM-DD')
    AND c.Country = 'United States';
```

Question 10:

```
SELECT c.Country, AVG(v.NumOfLikes) AS AvgLikes
FROM Creator c
JOIN Video v
ON c.Username = v.Username
WHERE c.Country
IN ('United States', 'India')
GROUP BY c.Country;
```

Describe your teamwork: how did you come up with the 10 questions, list the contribution of each team member

Each team member chose 2 questions to write the code for. We then tested each code to ensure that it worked properly and matched the questions we had written previously. If the code did not work or if the code didn't work with our set database we changed the code or the question to make sure everything worked properly. Below is the breakdown of who did the base code for each of the questions.

Carissa - 1,2
Hannah - 4,3,10
Sophia -5,6
Ben - 7,8
Together - 9