Final Report

Alden Mo, Ao Shen, and Yunzhe Yu

Project Background

Our project is based on parsing and going through YouTube data in order to find trends of a general person. Our main goal is, given the YouTube data, to convert it into a database and then use analyze results through a set of 10 questions. The questions, in addition to focusing on general topics like the most watched channel, also targeted to show how viewing habits changed during the 2020 COVID Pandemic.

Our main use for this data is to find the average trends of the user and compare it to post-COVID watch trends. This can be used to find out how COVID changed viewership of a user and how it affected daily lives.

Database Description

The Subscriptions table could include the channelURL (STRING), the channelId (STRING), and channelName (STRING). This data can be used to correlate with the table that contains watch history.

Subscription channelURL —channelId—channelName

The Watch_History table should detail when you watched the video watchDate (TIMESTAMP), the videoURL (STRING), the channelName (STRING), and the channelURL (STRING). This can be related to subscribed channels to see what videos are correlated to which channel.

Watch_History watchDate—videoURL—channelName–channelURL

The Search_History table should detail the searchQuery (STRING) and the time of searchDate (TIMESTAMP). We can use this to check the amount of searches in a date compared to the amount of videos watched in that date to determine how long a single search kept you clicking recommended videos.

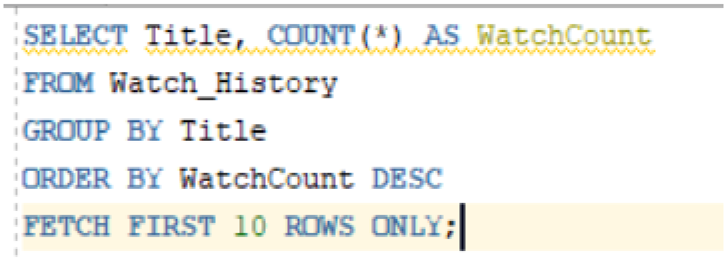
Search_History searchQuery—time of searchDate

Finally, we noticed that Youtube exports comments, this is in HTML format, but if we're still able to parse it into SQL Developer then we can get the video id of the comment videoURL (STRING), the timestamp of the comment commentDate (TIMESTAMP), and the contents of the comment comment (STRING). The table name is Comments This can be used in relation with watch history to determine what videos or channels you commented on the most.

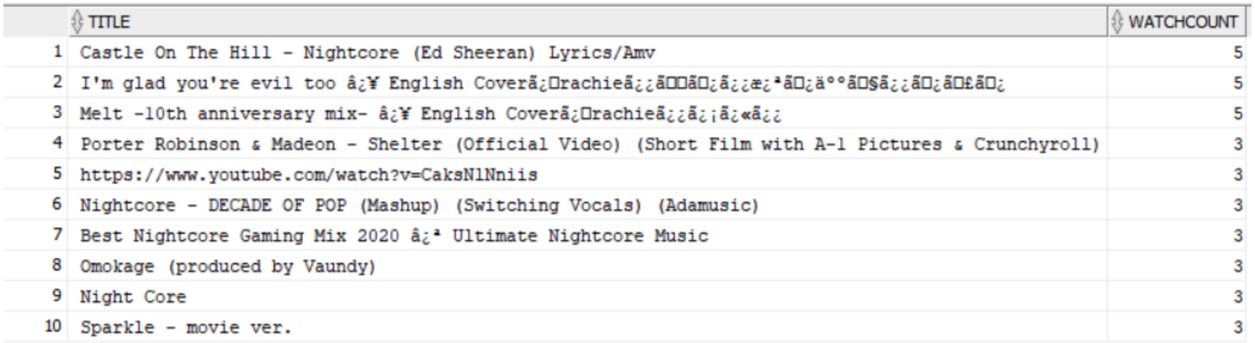
Questions and Solutions

1. Determine the most viewed videos by users

The data obtained can help YouTube determine users' preferred content and increase user engagement.



```
SELECT Title, COUNT(*) AS WatchCount
FROM Watch_History
GROUP BY Title
ORDER BY WatchCount DESC
FETCH FIRST 10 ROWS ONLY;
```

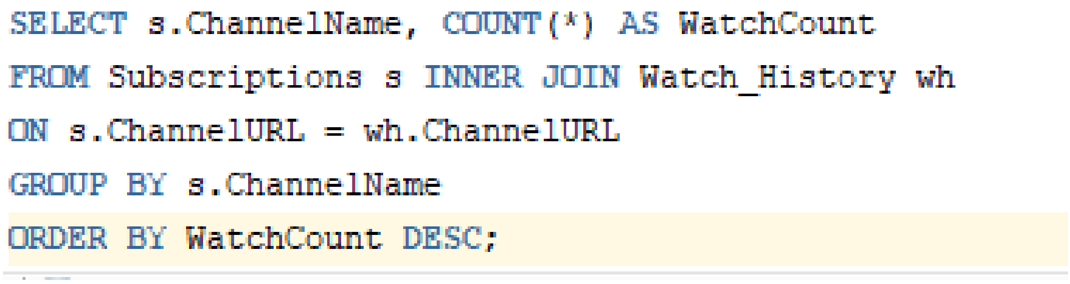


	TITLE	WATCHCOUNT
1	Castle On The Hill - Nightcore (Ed Sheeran) Lyrics/Amv	5
2	I'm glad you're evil too â¿¥ English Coverã¿Orachieã¿¿ãOOãO¿ã¿¿æ¿ªãO¿ä°°ãO§ã¿¿ãO¿ãO£ãO¿	5
3	Melt -10th anniversary mix- â¿¥ English Coverã¿Orachieã¿¿ã¿¡ã¿«ã¿¿	5
4	Porter Robinson & Madeon - Shelter (Official Video) (Short Film with A-1 Pictures & Crunchyroll)	3
5	https://www.youtube.com/watch?v=CaksN1Nniis	3
6	Nightcore - DECADE OF POP (Mashup) (Switching Vocals) (Adamusic)	3
7	Best Nightcore Gaming Mix 2020 â¿ª Ultimate Nightcore Music	3
8	Omokage (produced by Vaundy)	3
9	Night Core	3
10	Sparkle - movie ver.	3

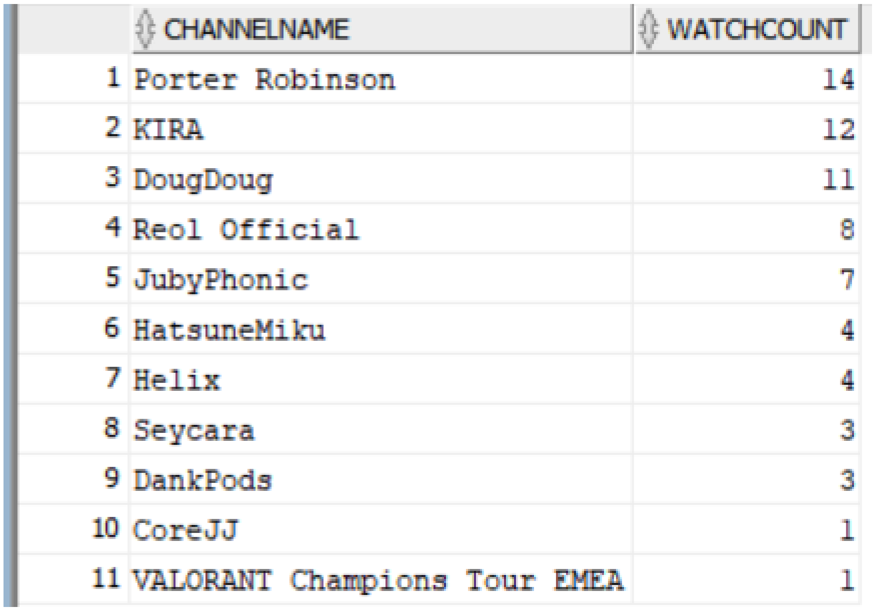
Based on the result, we can got this user likes music, and the platform can try to increase music-related push content

2. Determine which subscription channels users watch most often (excluding channels to which users are not subscribed)

Data helps YouTube identify recommended content for users, and it helps motivate video creators



```
SELECT s.ChannelName, COUNT(*) AS WatchCount
FROM Subscriptions s INNER JOIN Watch_History wh
ON s.ChannelURL = wh.ChannelURL
GROUP BY s.ChannelName
ORDER BY WatchCount DESC;
```

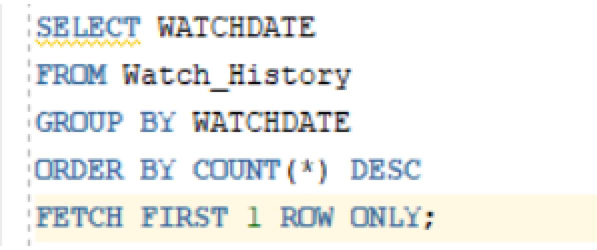


	CHANNELNAME	WATCHCOUNT
1	Porter Robinson	14
2	KIRA	12
3	DougDoug	11
4	Reol Official	8
5	JubyPhonic	7
6	HatsuneMiku	4
7	Helix	4
8	Seycara	3
9	DankPods	3
10	CoreJJ	1
11	VALORANT Champions Tour EMEA	1

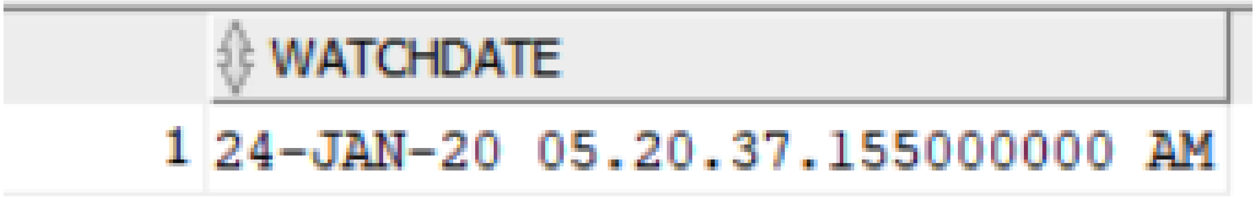
According to the results we can know that the user's favorite channel is the music channel, the platform can target the user to push, and the platform can consolidate a large number of results and inform the creators, thus motivating them

3. Get the day with the most videos watched

This lets us look at the day with the most videos watched, as it’s most videos and not minutes it gives us a look at (most likely) their favorite types of short form content



```sql
SELECT WATCHDATE
FROM Watch_History
GROUP BY WATCHDATE
ORDER BY COUNT(*) DESC
FETCH FIRST 1 ROW ONLY;
```

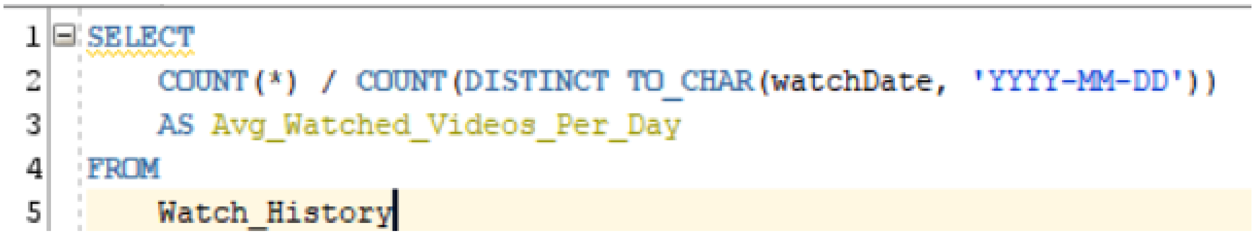


	WATCHDATE
1	24-JAN-20 05.20.37.155000000 AM

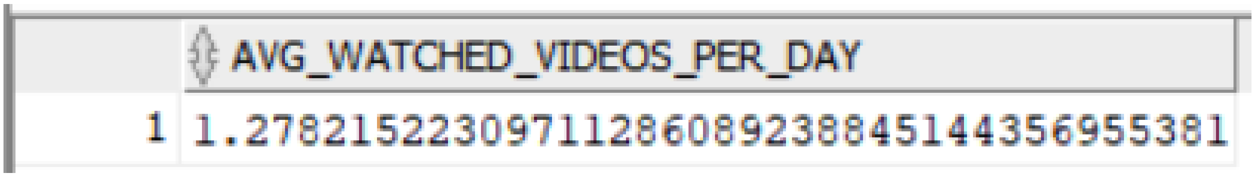
Finding this date is very important for the platform, they may be able to use this date to find the most played date per week, per month and take some action on that day.

4. Determine the average number of videos users watch per day

The data is used to understand how often users use YouTube and can be used to measure user engagement and satisfaction.



```sql
SELECT
    COUNT(*) / COUNT(DISTINCT TO_CHAR(watchDate, 'YYYY-MM-DD'))
    AS Avg_Watched_Videos_Per_Day
FROM
    Watch_History
```

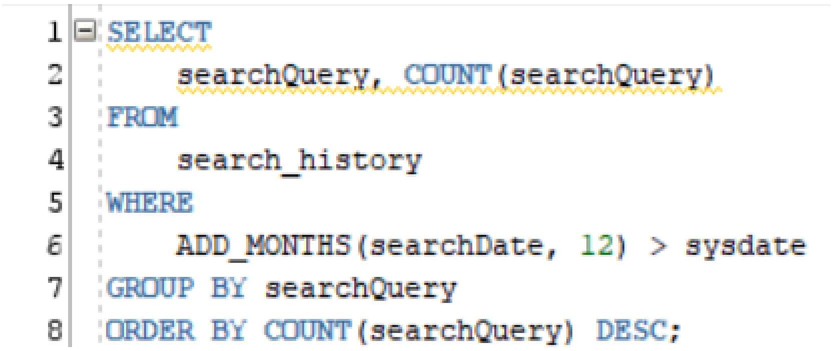


	AVG_WATCHED_VIDEOS_PER_DAY
1	1.27821522309711286089238845144356955381

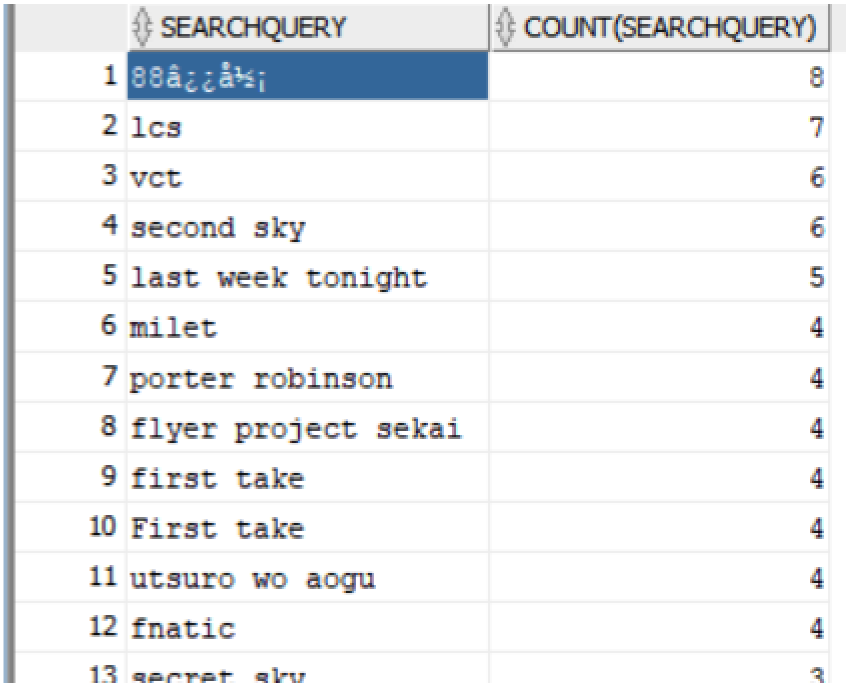
The results are used to determine the retention rate of users, and when this data changes, the platform needs to take some measures to retain users

5. What are the user's most searched queries in the past year?

This query provides insight into the user's interests and preferences. This information can be used to tailor content and marketing campaigns that align with the user's interests.



```sql
SELECT
    searchQuery, COUNT(searchQuery)
FROM
    search_history
WHERE
    ADD_MONTHS(searchDate, 12) > sysdate
GROUP BY searchQuery
ORDER BY COUNT(searchQuery) DESC;
```

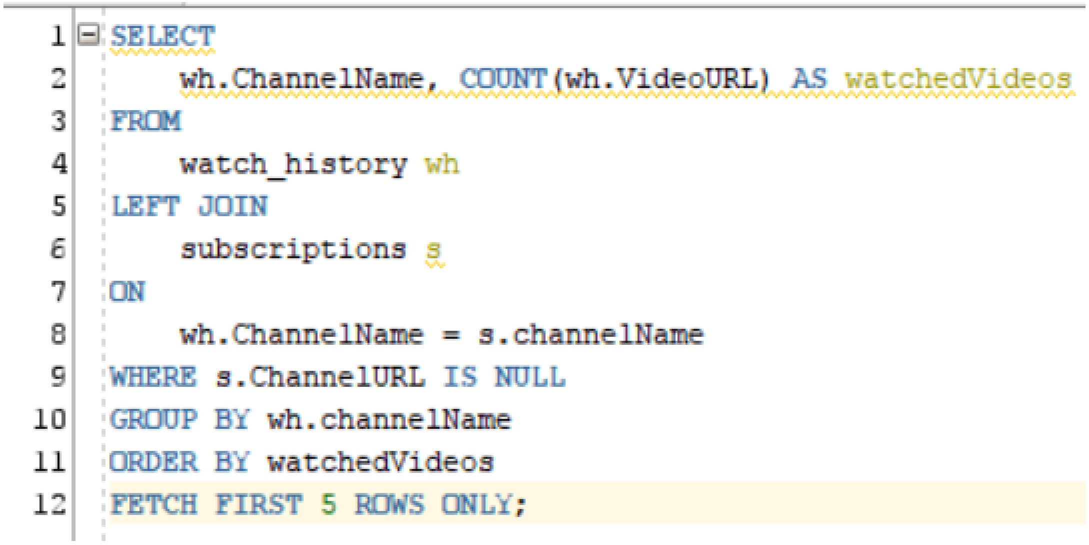


	SEARCHQUERY	COUNT(SEARCHQUERY)
1	88â¿¿å½¡	8
2	lcs	7
3	vct	6
4	second sky	6
5	last week tonight	5
6	milet	4
7	porter robinson	4
8	flyer project sekai	4
9	first take	4
10	First take	4
11	utsuro wo aogu	4
12	fnatic	4
13	secret sky	3

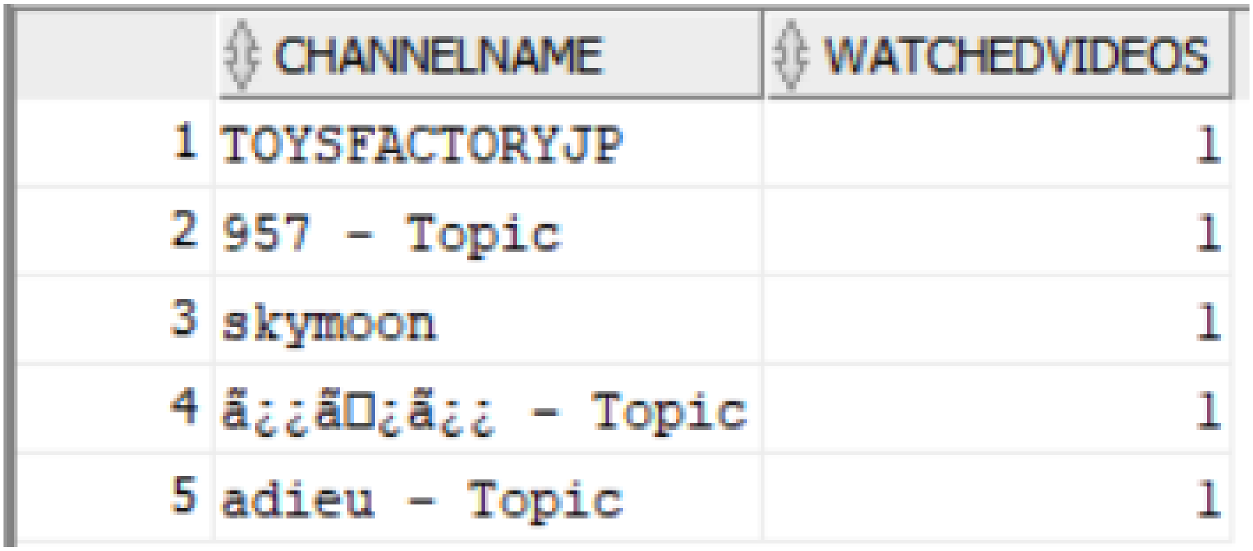
The data is used to analyze user preferences, change previously pushed content, and provide feedback for video creation.

6. Determine what the top 5 popular channels are that users are not subscribed to

Data can help platforms identify potentially popular channels that users may not have discovered yet, and use that information to make targeted marketing or content recommendations.



```
SELECT
    wh.ChannelName, COUNT(wh.VideoURL) AS watchedVideos
FROM
    watch_history wh
LEFT JOIN
    subscriptions s
ON
    wh.ChannelName = s.channelName
WHERE s.ChannelURL IS NULL
GROUP BY wh.channelName
ORDER BY watchedVideos
FETCH FIRST 5 ROWS ONLY;
```

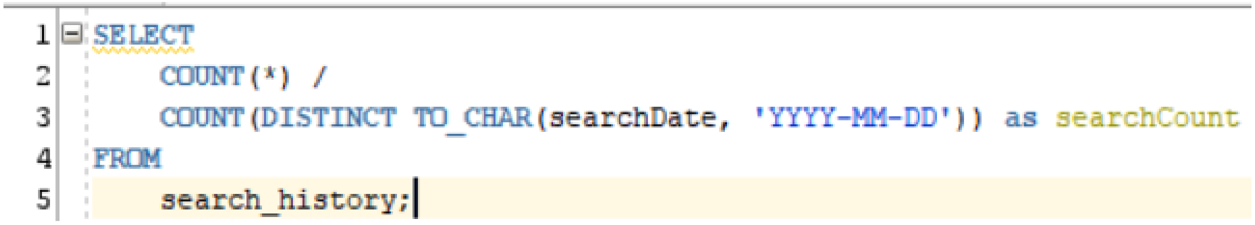


	CHANNELNAME	WATCHEDVIDEOS
1	TOYSFACTORYJP	1
2	957 - Topic	1
3	skymoon	1
4	ã¿¿ã□¿ã¿¿ - Topic	1
5	adieu - Topic	1

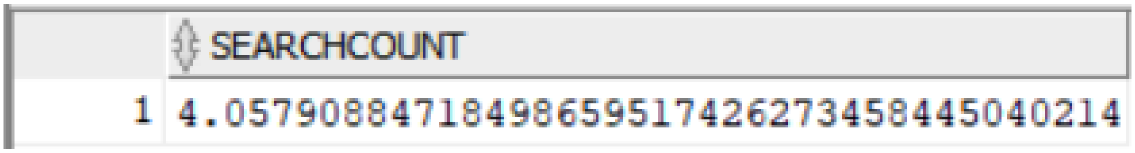
The data provides insight into potential partnerships or sponsorship opportunities for popular channels.

7. Determine the average number of searches performed by users per day

Data can help YouTube gain insight into user behavior and preferences, and use that information to improve search features and results.



```
SELECT
    COUNT(*) /
    COUNT(DISTINCT TO_CHAR(searchDate, 'YYYY-MM-DD')) as searchCount
FROM
    search_history;
```

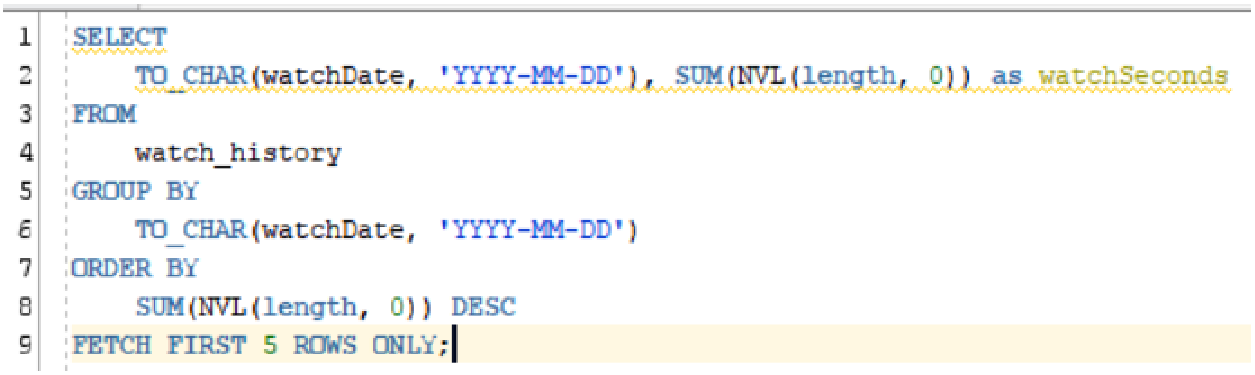


	SEARCHCOUNT
1	4.05790884718498659517426273458445040214

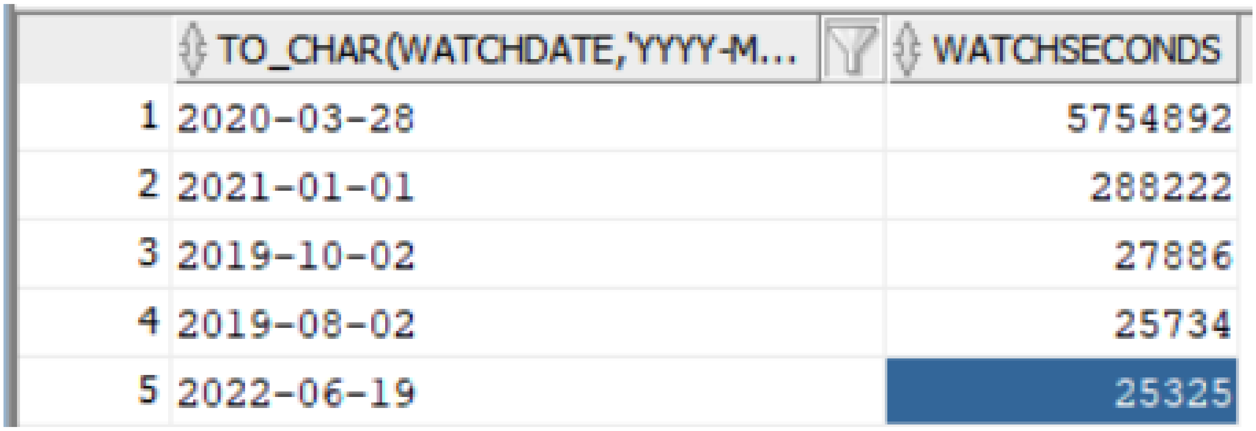
The platform can improve the search function based on the results, allowing users to get more direct access to the ethical results they want to search for.

8. Determine what are the most common search queries users use before watching a video

The data can be used to determine users' interests and preferences through their viewing history as a way to improve YouTube's search capabilities and give users a better experience when using the platform.



```
SELECT
    TO_CHAR(watchDate, 'YYYY-MM-DD'), SUM(NVL(length, 0)) as watchSeconds
FROM
    watch_history
GROUP BY
    TO_CHAR(watchDate, 'YYYY-MM-DD')
ORDER BY
    SUM(NVL(length, 0)) DESC
FETCH FIRST 5 ROWS ONLY;
```

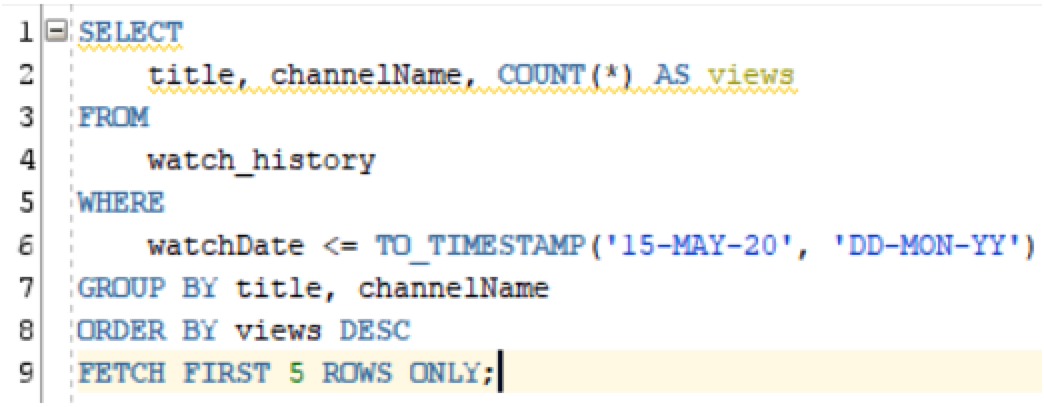


	TO_CHAR(WATCHDATE,'YYYY-M...	WATCHSECONDS
1	2020-03-28	5754892
2	2021-01-01	288222
3	2019-10-02	27886
4	2019-08-02	25734
5	2022-06-19	25325

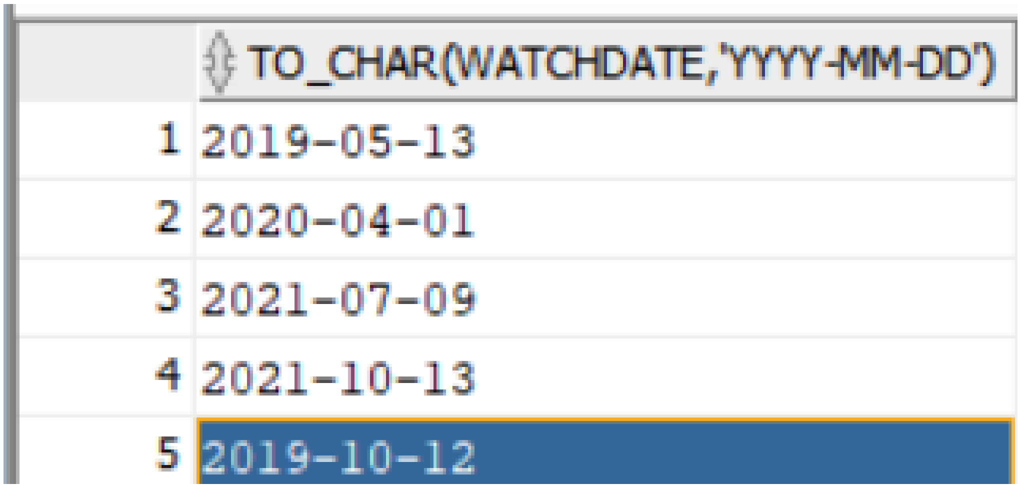
The platform can step up publicity on these dates

9. Which videos did the user watch the most in the recent month and how many times were they watched?

The data provides information on the most 3 popular videos in a recent month watched by the user. This information can be used to create targeted content and promotions that align with the user's interests and increase engagement.



```sql
SELECT
    title, channelName, COUNT(*) AS views
FROM
    watch_history
WHERE
    watchDate <= TO_TIMESTAMP('15-MAY-20', 'DD-MON-YY')
GROUP BY title, channelName
ORDER BY views DESC
FETCH FIRST 5 ROWS ONLY;
```

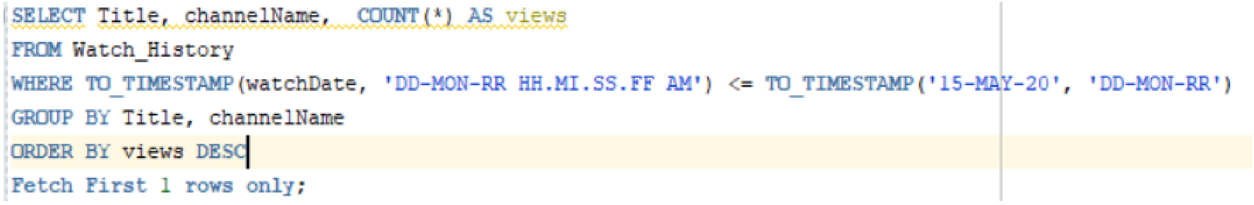


	TO_CHAR(WATCHDATE,'YYYY-MM-DD')
1	2019-05-13
2	2020-04-01
3	2021-07-09
4	2021-10-13
5	2019-10-12

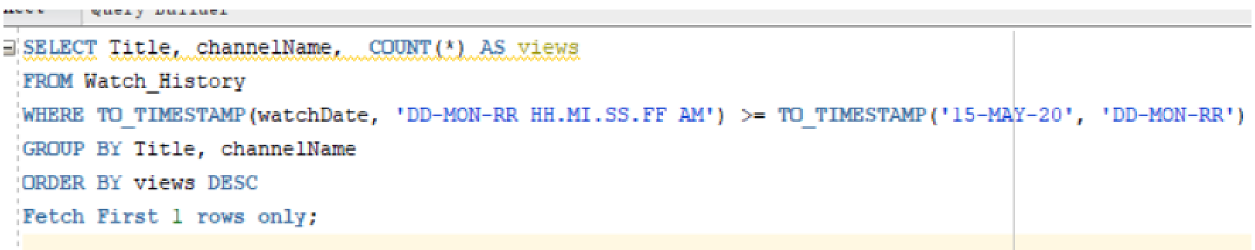
The user wanted to know some knowledge about non-ferrous metal recently

10. Using May 15, 2020 as the node, determine the most viewed videos before and after this point in time

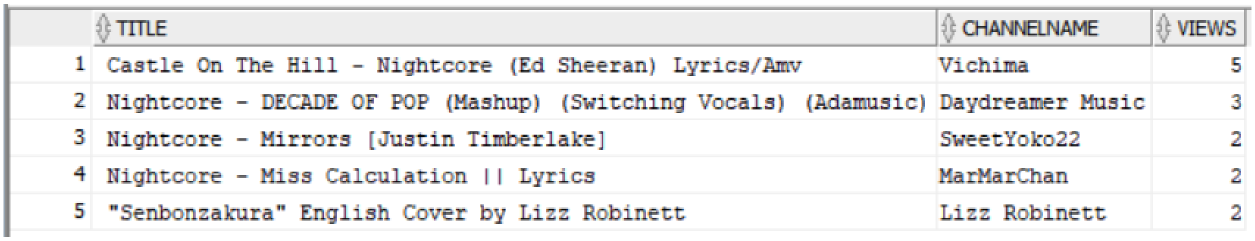
    Data to understand any changes in user sentiment during a pandemic



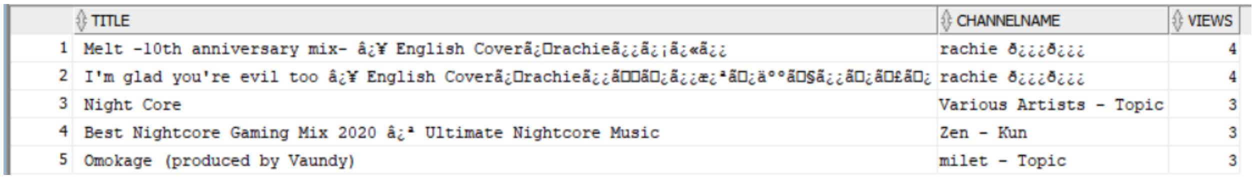
```
SELECT Title, channelName,  COUNT(*) AS views
FROM Watch_History
WHERE TO_TIMESTAMP(watchDate, 'DD-MON-RR HH.MI.SS.FF AM') <= TO_TIMESTAMP('15-MAY-20', 'DD-MON-RR')
GROUP BY Title, channelName
ORDER BY views DESC
Fetch First 1 rows only;
```



```
SELECT Title, channelName,  COUNT(*) AS views
FROM Watch_History
WHERE TO_TIMESTAMP(watchDate, 'DD-MON-RR HH.MI.SS.FF AM') >= TO_TIMESTAMP('15-MAY-20', 'DD-MON-RR')
GROUP BY Title, channelName
ORDER BY views DESC
Fetch First 1 rows only;
```



	TITLE	CHANNELNAME	VIEWS
1	Castle On The Hill - Nightcore (Ed Sheeran) Lyrics/Amv	Vichima	5
2	Nightcore - DECADE OF POP (Mashup) (Switching Vocals) (Adamusic)	Daydreamer Music	3
3	Nightcore - Mirrors [Justin Timberlake]	SweetYoko22	2
4	Nightcore - Miss Calculation || Lyrics	MarMarChan	2
5	"Senbonzakura" English Cover by Lizz Robinett	Lizz Robinett	2



	TITLE	CHANNELNAME	VIEWS
1	Melt -10th anniversary mix- â¿¥ English Coverã¿□rachieã¿¿ã¿¡ã¿«ã¿¿	rachie ð¿¿¿ð¿¿¿	4
2	I'm glad you're evil too â¿¥ English Coverã¿□rachieã¿¿ã□□ã□¿ã¿¿æ¿ªã□¿ä°°ã□§ã¿¿ã□¿ã□£ã□¿	rachie ð¿¿¿ð¿¿¿	4
3	Night Core	Various Artists - Topic	3
4	Best Nightcore Gaming Mix 2020 â¿ª Ultimate Nightcore Music	Zen - Kun	3
5	Omokage (produced by Vaundy)	milet - Topic	3

This user has been concerned about new energy vehicles since 2020

Teamwork:

We collaborated as a team to brainstorm and come up with a list of more than 10 questions. We then analyzed each question's relevance and narrowed down the list to ten questions that would provide meaningful insights and demonstrate our learning from the course CNIT 372. Each team member contributed a lot to the process, bringing in their expertise in data analysis, visualization, and text analysis to shape the questions.