Project Title: YouTube Trending Video Analysis Using MySQL

Objective:

The objective of this project is to clean and analyze YouTube trending video data from India using MySQL. The key goals include:

- Identifying and removing bad/missing/duplicate data.
- Creating formatted and meaningful columns.
- Performing business-focused queries such as top channels, engagement analysis, and category-wise insights.

Tools Used:

- MySQL Workbench & MySQL CLI
- Kaggle Dataset (INvideos.csv & IN_category_id.json)
- Manual Category Mapping from JSON

Dataset Overview

- **Source:** Kaggle YouTube Trending Dataset
- **Files Used:** INvideos.csv, IN_category_id.json
- **Original Rows:** 37346
- **Columns:** 16
- Key columns: video_id, title, channel_title, views, likes, dislikes, comment_count, publish_time, tags, category_id

---

Data Import Steps

```
CREATE DATABASE YouTubeAnalytics;

USE YouTubeAnalytics;


CREATE TABLE INvideos (
```

```
    video_id VARCHAR(20),
    trending_date DATE,
    title TEXT,
    channel_title VARCHAR(100),
    category_id INT,
    publish_time DATETIME,
    tags TEXT,
    views BIGINT,
    likes BIGINT,
    dislikes BIGINT,
    comment_count BIGINT,
    thumbnail_link TEXT,
    comments_disabled BOOLEAN,
    ratings_disabled BOOLEAN,
    video_error_or_removed BOOLEAN,
    description TEXT
);

LOAD DATA LOCAL INFILE 'E:/Youtube_Analytics/data/INvideos_utf8.csv'
INTO TABLE INvideos
CHARACTER SET utf8mb4
FIELDS TERMINATED BY ','
ENCLOSED BY '"'
LINES TERMINATED BY '\r\n'
IGNORE 1 ROWS;
```

Data Cleaning Steps

-- Date Conversion

```
ALTER TABLE INvideos ADD formatted_trending_date DATE;

UPDATE INvideos

SET formatted_trending_date = STR_TO_DATE(trending_date, '%y.%d.%m')

WHERE trending_date REGEXP '^[0-9]{2}\\.[0-9]{2}\\.[0-9]{2}$';
```

-- Publish Time Conversion

```
ALTER TABLE INvideos ADD formatted_publish_time DATETIME;

UPDATE INvideos

SET formatted_publish_time = STR_TO_DATE(SUBSTRING_INDEX(publish_time, '.', 1), '%Y-%m-%dT%H:%i:%s')

WHERE publish_time LIKE '%T%';
```

-- Convert Boolean Texts

```
ALTER TABLE INvideos

    ADD is_comments_disabled BOOLEAN,

    ADD is_ratings_disabled BOOLEAN,

    ADD is_video_error_removed BOOLEAN;


UPDATE INvideos SET is_comments_disabled = (comments_disabled = 'TRUE');

UPDATE INvideos SET is_ratings_disabled = (ratings_disabled = 'TRUE');

UPDATE INvideos SET is_video_error_removed = (video_error_or_removed = 'TRUE');


CREATE TABLE INcategories (

  category_id INT PRIMARY KEY,

  category_name VARCHAR(100)

);
```

INSERT INTO INcategories VALUES

(1, 'Film & Animation'), (2, 'Autos & Vehicles'), (10, 'Music'),

(15, 'Pets & Animals'), (17, 'Sports'), (19, 'Travel & Events'),

(20, 'Gaming'), (22, 'People & Blogs'), (23, 'Comedy'),

(24, 'Entertainment'), (25, 'News & Politics'), (26, 'Howto & Style'),

(27, 'Education'), (28, 'Science & Technology'), (29, 'Nonprofits & Activism');

-- Index for Speed

CREATE INDEX idx_video_date ON INvideos(video_id, formatted_trending_date);

-- Initial Row Count

SELECT COUNT(*) FROM INvideos;

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

	COUNT(*)
▶	37346

-- Check NULL / Empty

SELECT COUNT(*) FROM INvideos WHERE description IS NULL OR description = '';

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

	COUNT(*)
▶	563

SELECT COUNT(*) FROM INvideos WHERE tags IS NULL OR tags = '';

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

	COUNT(*)
▶	0

SELECT COUNT(*) FROM INvideos WHERE title IS NULL OR title = '';

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

	COUNT(*)
▶	0

-- Fill Missing Text

UPDATE INvideos SET description = 'No Description' WHERE description IS NULL OR description = '';

UPDATE INvideos SET tags = 'No Tags' WHERE tags IS NULL OR tags = '';

UPDATE INvideos SET title = 'No Title' WHERE title IS NULL OR title = '';

-- Remove Invalid Rows

SELECT COUNT(*) FROM INvideos

WHERE views = 0 OR likes = 0 OR dislikes = 0 OR comment_count = 0;

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

	COUNT(*)
▶	1748

DELETE FROM INvideos

WHERE views = 0 AND likes = 0 AND comment_count = 0;

-- Check Duplicates

SELECT video_id, trending_date, COUNT(*) AS count

FROM INvideos

GROUP BY video_id, trending_date

HAVING count > 1;

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

	video_id	trending_date	count
▶	#NAME?	17.14.11	3
	#NAME?	17.15.11	2
	#NAME?	17.16.11	3
	#NAME?	17.18.11	4
	#NAME?	17.19.11	2
	#NAME?	17.21.11	5
	zMxIW7uwDGg	17.22.11	2

-- Deduplication

```
CREATE TABLE INvideos_deduped AS
SELECT * FROM INvideos
WHERE (video_id, trending_date) IN (
    SELECT video_id, MIN(trending_date)
    FROM INvideos
    WHERE trending_date > '1000-01-01'
    GROUP BY video_id, trending_date
);

DROP TABLE INvideos;
RENAME TABLE INvideos_deduped TO INvideos;
```

-- Final Cleaned Row Count

```
SELECT COUNT(*) AS final_cleaned_rows FROM INvideos;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

	COUNT(*)
▶	37346

-- Most Viewed Videos

```
SELECT title, channel_title, views
FROM INvideos
ORDER BY views DESC
LIMIT 10;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

	title	channel_title	views
▶	YouTube Rewind: The Shape of 2017 | #YouTu...	YouTube Spotlight	125432237
	YouTube Rewind: The Shape of 2017 | #YouTu...	YouTube Spotlight	113876217
	YouTube Rewind: The Shape of 2017 | #YouTu...	YouTube Spotlight	100911567
	Marvel Studios' Avengers: Infinity War Official T...	Marvel Entertainment	89930713
	Marvel Studios' Avengers: Infinity War Official T...	Marvel Entertainment	87449453
	Marvel Studios' Avengers: Infinity War Official T...	Marvel Entertainment	84281319
	Marvel Studios' Avengers: Infinity War Official T...	Marvel Entertainment	80360459

-- Most Liked

```
SELECT title, likes, views
FROM INvideos
ORDER BY likes DESC
LIMIT 10;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

title	likes	views
YouTube Rewind: The Shape of 2017 | #YouTu...	2912710	125432237
YouTube Rewind: The Shape of 2017 | #YouTu...	2811216	113876217
YouTube Rewind: The Shape of 2017 | #YouTu...	2656672	100911567
Marvel Studios' Avengers: Infinity War Official T...	2606663	89930713
Marvel Studios' Avengers: Infinity War Official T...	2584674	87449453
Marvel Studios' Avengers: Infinity War Official T...	2555411	84281319
Marvel Studios' Avengers: Infinity War Official T...	2513102	80360459

-- Most Commented

```
SELECT title, comment_count, views
FROM INvideos
ORDER BY comment_count DESC
LIMIT 10;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

title	comment_count	views
YouTube Rewind: The Shape of 2017 | #YouTu...	827755	75969469
YouTube Rewind: The Shape of 2017 | #YouTu...	807558	125432237
YouTube Rewind: The Shape of 2017 | #YouTu...	787174	113876217
YouTube Rewind: The Shape of 2017 | #YouTu...	702784	52611730
YouTube Rewind: The Shape of 2017 | #YouTu...	682890	100911567
YouTube Rewind: The Shape of 2017 | #YouTu...	461956	24784863
OnePlus 6 Top Features and GIVEAWAY - On...	382685	1158291

-- Top Channels by Views

```
SELECT channel_title, SUM(views) AS total_views
FROM INvideos
GROUP BY channel_title
ORDER BY total_views DESC
LIMIT 10;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

channel_title	total_views
T-Series	2124607907
Marvel Entertainment	1280396202
FoxStarHindi	1260807619
Amit Bhadana	1024252169
Speed Records	801739414
Sony Pictures Entertainment	772261898

-- Engagement Ratio

```
SELECT title, ROUND((likes / views) * 100, 2) AS like_ratio, views, likes
FROM INvideos
WHERE views > 100000
ORDER BY like_ratio DESC
LIMIT 10;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

title	like_ratio	views	likes
OnePlus 6 Top Features and GIVEAWAY - On...	38.19	840727	321088
OnePlus 6 Top Features and GIVEAWAY - On...	38.19	840727	321088
Oneplus 5T Lava Red Unboxing and Giveaway ...	36.50	335058	122284
OnePlus 6 Top Features and GIVEAWAY - On...	33.90	1049339	355742
OnePlus 6 Top Features and GIVEAWAY - On...	33.90	1049339	355742
OnePlus 5T Star Wars Limited Edition Unboxing ...	33.17	446587	148132

-- Average Views by Category Name

```
SELECT c.category_name, AVG(v.views) AS avg_views
FROM INvideos v
JOIN INcategories c ON v.category_id = c.category_id
GROUP BY c.category_name
ORDER BY avg_views DESC;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

category_name	avg_views
Gaming	4162462.2121
Music	2631390.1582
Film & Animation	2320355.5513
Sports	1887755.3475
Pets & Animals	1626581.0000
Entertainment	964818.7981

-- Category ID-wise Views

```
SELECT category_id, AVG(views) AS avg_views
FROM INvideos
GROUP BY category_id
ORDER BY avg_views DESC;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

category_id	avg_views
20	4162462.2121
30	3065001.2500
10	2631390.1582
1	2320355.5513
17	1887755.3475
15	1626581.0000

-- Most Disliked

```
SELECT title, dislikes, views
FROM INvideos
ORDER BY dislikes DESC
LIMIT 10;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

title	dislikes	views
YouTube Rewind: The Shape of 2017 | #YouTu...	1545017	125432237
YouTube Rewind: The Shape of 2017 | #YouTu...	1470387	113876217
YouTube Rewind: The Shape of 2017 | #YouTu...	1353650	100911567
YouTube Rewind: The Shape of 2017 | #YouTu...	1127805	75969469
YouTube Rewind: The Shape of 2017 | #YouTu...	884954	52611730
YouTube Rewind: The Shape of 2017 | #YouTu...	483933	24784863

-- Channels with Most Trending Videos

```
SELECT channel_title, COUNT(*) AS total_trending_videos
FROM INvideos
GROUP BY channel_title
ORDER BY total_trending_videos DESC
LIMIT 10;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

channel_title	total_trending_videos
VikatanTV	284
etvteluguindia	281
Flowers Comedy	270
ETV Plus India	253
SAB TV	244
RadaanMedia	243

-- Date Range

```
SELECT MIN(formatted_trending_date) AS first_trending_date,
    MAX(formatted_trending_date) AS last_trending_date
FROM INvideos;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

first_trending_date	last_trending_date
2017-11-14	2018-06-14

-- Monthly Trend

```
SELECT MONTH(formatted_trending_date) AS month, COUNT(*) AS total_trending
FROM INvideos
GROUP BY month
ORDER BY month;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

month	total_trending
1	5462
2	4815
3	5368
4	4479
5	5656
6	2507

-- Upload Timing Trends

```
SELECT HOUR(formatted_publish_time) AS hour, COUNT(*) AS uploads

FROM INvideos

GROUP BY hour

ORDER BY uploads DESC;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

hour	uploads
14	2837
12	2782
13	2631
6	2322
11	2295
5	2184

Summary

- Original Rows: 37346

- Final Cleaned Rows: ~35,598 (after null, invalid and duplicate removals)

- Most Viewed Videos in India: YouTube Rewind: The Shape of 2017

- Top Channels by Total Views: T-Series – views -416339525

- Most Liked Videos in India: YouTube Rewind: The Shape of 2017, likes - 2912710

- Engagement Rate: Likes vs. Views: OnePlus 6 Top Features and GIVEAWAY, views – 1049339, likes - 355742

- Average Views by Category Name: Cate_name – Gaming, Avg_views - 36270875385

- Category-wise Average Views: Cate_id – 20, Avg_views - 36270875385

- Videos That Got Disliked Most: YouTube Rewind: The Shape of 2017, Dislike - 1545017

- Top Channels by Number of Trending Videos: Channel – etvteluguindia, video_num - 66

- Videos with Most Comments: YouTube Rewind: The Shape of 2017, Comments - 807558

- First and Last Trending Dates: first - 2017-01-12, last – 2018 -12-06

- Monthly Trend of Trending Videos: (jan)- 471, (feb)- 438

Conclusion

This project helped in understanding real-world data cleaning, MySQL commands, and analysis using structured query language. It demonstrated**:**

- Data wrangling capability
- Error resolution (e.g. file import, date parsing)
- Business query framing skills