DATA PORTFOLIO: EXCEL TO POWER BI

Identifying top YouTube Content Creators Using Excel, SQL, and Power BI

November 2024

By Sarah Park

Data Portfolio: Excel to Power BI
Identifying Top YouTube Content Creators

By Sarah Park

Contents

Introduction

Identify the top-performing YouTube content creator of 2023 and collaborate with them to keep New Zealand on the global travel radar. Our aim is to enhance NZ's tourism economy, increase opportunities to benefit local businesses, boost support for sports teams, attract potential investors, encourage an influx of foreign money, and inspire more influential figures to visit.

User Story

"As the head of marketing, I want to identify the top content creator who can effectively promote the country as a destination for its stunning natural landscapes, peaceful and laid-back lifestyle, humble and friendly people, and world-class outdoor activities. To maximise reach, we want to look into mega influencers rather than niche content creators."

Solution

Use a dashboard that identifies the top-performing channels based on metrics like subscriber count, views, and engagement rates. Select the top 3 creators, calculate the sponsorship cost for each channel, and compare them to determine which one is best to advance with.

Success Criteria (what success looks like to the user)

- Users can easily identify the top-performing YouTube channels based on subscriber count, views, and engagement rates.
- User can assess the potential for successful campaigns with top content creators based on conversion, engagement, and budget.
- User can make informed decisions on which content creator would be suitable to advance with.

Tools

SQL (MySQL): Data Exploration, Cleaning, Testing
Power BI: Visualisation, Dashboard
Excel: Calculation, Generating Findings

Steps

Step 1. Get the data

- Download as csv file from Kaggle. **[Excel]**
https://www.kaggle.com/datasets/nelgiriyewithana/global-youtube-statistics-2023?resource=download

Step 2. Data Exploration, Cleaning / Transforming **[SQL]**

- Explore data and note findings.
- Clean data based on the findings from the data exploration notes.
- Check data quality after cleaning.

Step 3. Build a Dashboard **[Power BI]**

- Import the virtual data into Power BI.
- DAX measures.
- Build a dashboard.

Step 4. Generating Findings

- Generate findings based on the insights.
- Identify top 3 creators

Step 5. Calculate YouTube Sponsorship Rate **[Excel & SQL]**

- Use both Excel and SQL to calculate sponsorship rate to avoid discrepancies.

Step 6. Recommendations and Action Plan

SQL

Data Exploration Notes

- Data shape: 847 rows 28 cols
- Data Type:

Column Name	Data Type	Description
rank	INT	Overall rank of the YouTuber
Youtuber	TEXT	Name of the YouTube channel
subscribers	INT	Number of subscribers
video_views	DOUBLE	Total number of video views
Category	TEXT	Category of the channel (e.g., Entertainment, Gaming)
Title	TEXT	Title of the YouTube channel
uploads	INT	Number of videos uploaded
Country	TEXT	Country where the channel is based
Abbreviation	TEXT	Country abbreviation (e.g., US, UK)
channel_type	TEXT	Type of channel (e.g., Individual, Company)
video_views_rank	INT	Rank based on total video views
country_rank	INT	Rank within the country
channel_type_rank	INT	Rank based on channel type
video_views_for_the_last_30_days	BIGINT	Total video views in the last 30 days
lowest_monthly_earnings	INT	Estimated lowest monthly earnings
highest_monthly_earnings	DOUBLE	Estimated highest monthly earnings
lowest_yearly_earnings	DOUBLE	Estimated lowest yearly earnings
highest_yearly_earnings	DOUBLE	Estimated highest yearly earnings
subscribers_for_last_30_days	TEXT	Number of subscribers gained in the last 30 days
created_year	INT	Year the channel was created
created_month	TEXT	Month the channel was created
created_date	INT	Day of the month the channel was created
Gross tertiary education enrollment (%)	DOUBLE	Percentage of population enrolled in tertiary education
Population	INT	Population of the country
Unemployment rate	DOUBLE	Country's unemployment rate
Urban_population	INT	Urban population of the country
Latitude	DOUBLE	Latitude coordinates
Longitude	DOUBLE	Longitude coordinates

- Some characters in the Youtubers column are corrupted.
  Solution:
  Filter out special characters to improve readability:
  `DaniRep | +6 Vï¿½ï¿` → `DaniRep`
  `AlArabiya ï¿½ï¿½ï` → `AlArabiya`
  `!!###@@@` → (removed)
  `ýýýýýýýýýýý` → (removed)

- There are unnecessary columns that aren't relevant to this project.
  Solution:
  Drop `abbreviation`, `video views rank`, `channel_type_rank`, `video_views_for_the_last_30_days`, `lowest_monthly_earnings`, `highest_monthly_earnings`, `subscribers_for_last_30_days`, `created_year`, `created_month`, `created_date`, `Gross tertiary education enrollment (%)`, `Population`, `Unemployment rate`, `Urban_population`, and `Latitude`

- Some content creator categories are not relevant to the goal of this project.
  Solution:
  Drop values `Trailers`, `Nonprofits & Activism`, `Autos & Vehicles`, `nan`, `shows`, `Music`, `film & Animation`, `News & Politics`, and `Movies`.

- There are inconsistencies in terms of case and spacing
  Solution:
  Convert to lowercase, remove spaces, and replace with underscores.
  video views → video_views
  Country → country

- The data will require a rough estimate of the engagement rate based on the number of views per subscriber.
  Solution:
  Add a new col `engagement_rate` using the following formula:
  engagement_rate = (total_views / subscribers) * 100
  (In practice, more data will need to be analysed to improve accuracy)

- Kid's channels are not properly classified and are scattered across various categories such as People & Blogs, Entertainment, and Education.
  Solution:
  Option 1: If there is not much data remaining after cleaning, it is possible to manually review the list to confirm whether they are indeed kids' content. This can involve a quick glance at the channel's content or description.
  Option 2*:* re-categorise these channels accordingly and add a new column to indicate if the channel is a `Kids`.

Data Cleaning

We cleaned the data in accordance with the data exploration notes and saved as a virtual table:

```
CREATE VIEW virtual_table AS
SELECT
    `rank` AS overall_rank,
    REGEXP_REPLACE(`Youtuber`, '[^a-zA-Z0-9 ]', '') AS `channel_name`,
    subscribers,
    `video views` AS total_views,
    ROUND((`video views`/subscribers) * 100, 2) AS engagement_rate,
    category,
    uploads,
    Country AS country,
    channel_type,
    country_rank,
    channel_type_rank
FROM `global youtube statistics`
WHERE category NOT IN (
    'Trailers',
    'Nonprofits & Activism',
    'Autos & Vehicles',
    'nan',
    'shows',
    'Music',
    'Film & Animation',
    'News & Politics',
    'Movies'
)
AND TRIM(REGEXP_REPLACE(`Youtuber`, '[^a-zA-Z0-9 ]', '')) != ''
AND REGEXP_REPLACE(`Youtuber`, '[^a-zA-Z0-9 ]', '') IS NOT NULL
```

Data shape after cleaning: 541 rows 11 cols.

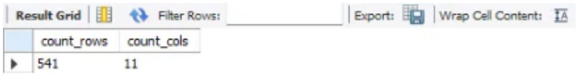
Data Quality Check After Cleaning

Before importing the cleaned data into Power BI, we want to ensure that it meets the following criteria:

- Criterion 1. Cleaned data should have 541 rows 11 cols.

```
SELECT
    (SELECT COUNT(*) FROM `virtual_table`) AS count_rows,
    (SELECT COUNT(*) FROM INFORMATION_SCHEMA.COLUMNS WHERE table_name = 'virtual_table' AND table_schema = DATABASE()) AS count_cols;
```

Output

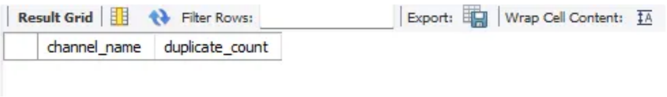


```
count_rows  count_cols
541         11
```

- Criterion 2. There should be no duplicates.

```sql
SELECT channel_name,
       COUNT(*) AS duplicate_count
FROM virtual_table
GROUP BY channel_name
HAVING COUNT(*) > 1;
```

Output

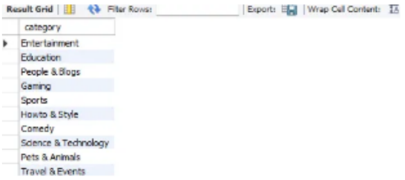


```
channel_name  duplicate_count
```

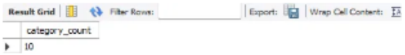
- Criterion 3. Cleaned data should have 10 unique values in the `categories` column.

```sql
-- list unique values
SELECT DISTINCT category
FROM `virtual_table`;

-- count unique values
SELECT COUNT(DISTINCT category) AS category_count
FROM `virtual_table`;
```

Output



```
category
Entertainment
Education
People & Blogs
Gaming
Sports
Howto & Style
Comedy
Science & Technology
Pets & Animals
Travel & Events
```



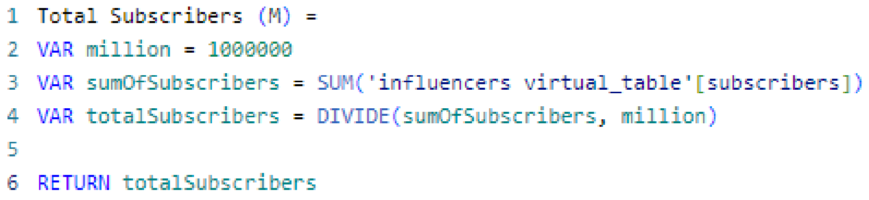
```
category_count
10
```

Power BI

Build a Dashboard with Power BI

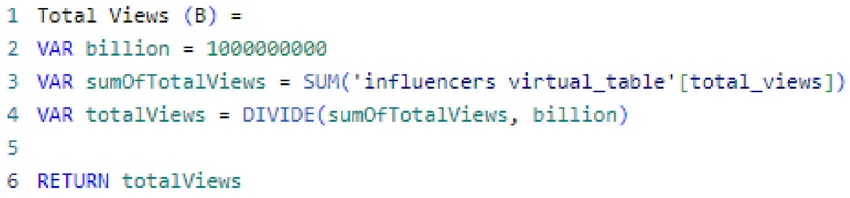
We have pushed the data into Power BI and calculated the DAX measures used to create the dashboard, including converting large numbers (e.g., billions) into a readable format e.g. 22.88 M.

Dax Measure: Total Subscribers



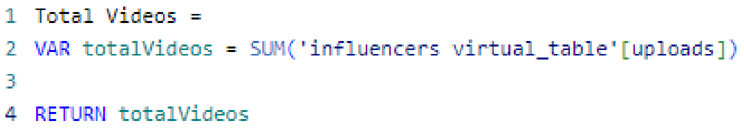
```
Total Subscribers (M) =
VAR million = 1000000
VAR sumOfSubscribers = SUM('influencers virtual_table'[subscribers])
VAR totalSubscribers = DIVIDE(sumOfSubscribers, million)

RETURN totalSubscribers
```

Dax Measure: Total Views



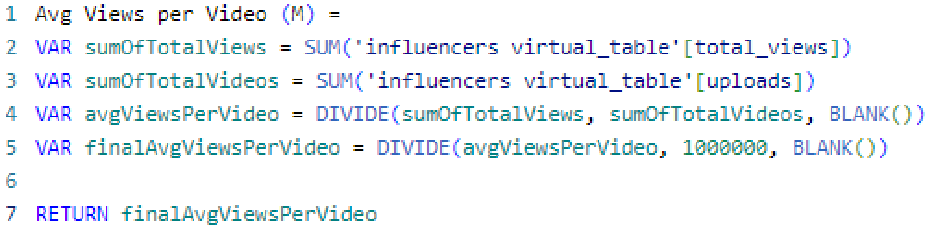
```
Total Views (B) =
VAR billion = 1000000000
VAR sumOfTotalViews = SUM('influencers virtual_table'[total_views])
VAR totalViews = DIVIDE(sumOfTotalViews, billion)

RETURN totalViews
```

Dax Measure: Total Videos



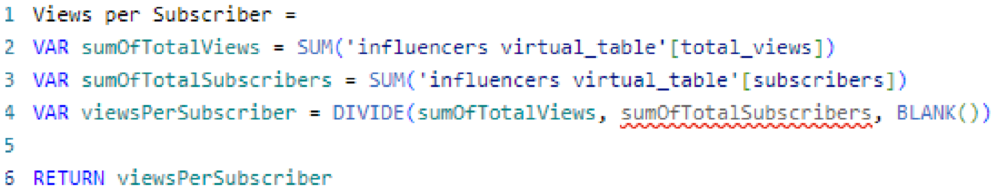
```
Total Videos =
VAR totalVideos = SUM('influencers virtual_table'[uploads])

RETURN totalVideos
```

Dax Measure: Average Views Per Video



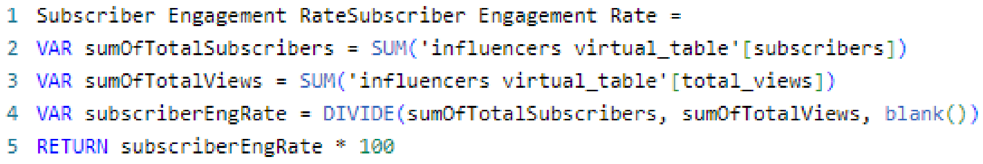
```
Avg Views per Video (M) =
VAR sumOfTotalViews = SUM('influencers virtual_table'[total_views])
VAR sumOfTotalVideos = SUM('influencers virtual_table'[uploads])
VAR avgViewsPerVideo = DIVIDE(sumOfTotalViews, sumOfTotalVideos, BLANK())
VAR finalAvgViewsPerVideo = DIVIDE(avgViewsPerVideo, 1000000, BLANK())

RETURN finalAvgViewsPerVideo
```

Dax Measure: View Per Subscriber



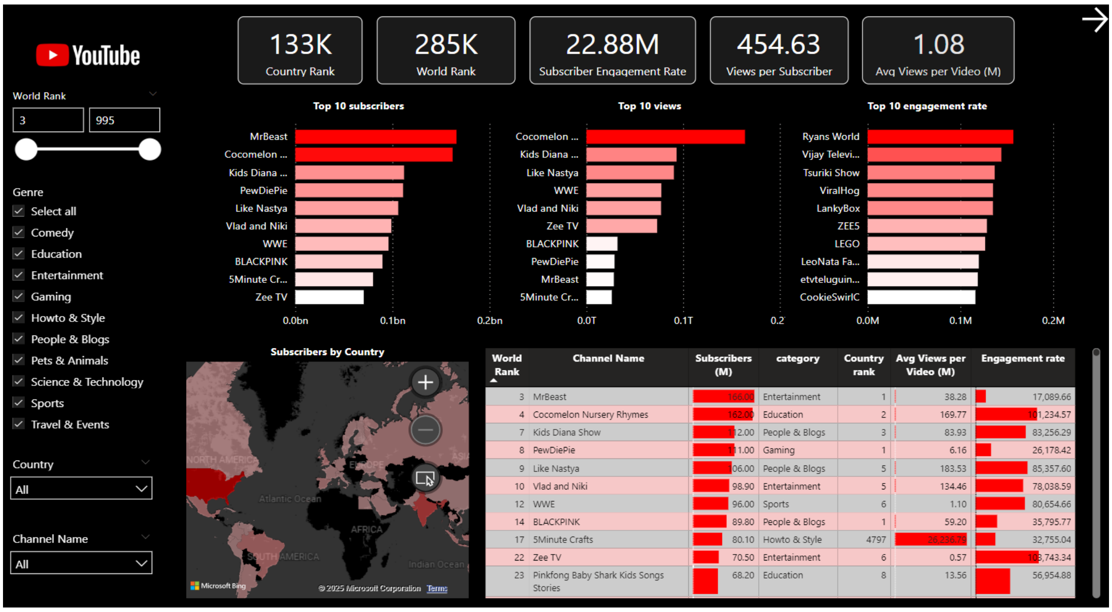
```
Views per Subscriber =
VAR sumOfTotalViews = SUM('influencers virtual_table'[total_views])
VAR sumOfTotalSubscribers = SUM('influencers virtual_table'[subscribers])
VAR viewsPerSubscriber = DIVIDE(sumOfTotalViews, sumOfTotalSubscribers, BLANK())

RETURN viewsPerSubscriber
```

Dax Measure: Engagement Rate



```
Subscriber Engagement RateSubscriber Engagement Rate =
VAR sumOfTotalSubscribers = SUM('influencers virtual_table'[subscribers])
VAR sumOfTotalViews = SUM('influencers virtual_table'[total_views])
VAR subscriberEngRate = DIVIDE(sumOfTotalSubscribers, sumOfTotalViews, blank())
RETURN subscriberEngRate * 100
```

Final Look of Our Dashboard



133K	285K	22.88M	454.63	1.08
Country Rank	World Rank	Subscriber Engagement Rate	Views per Subscriber	Avg Views per Video (M)

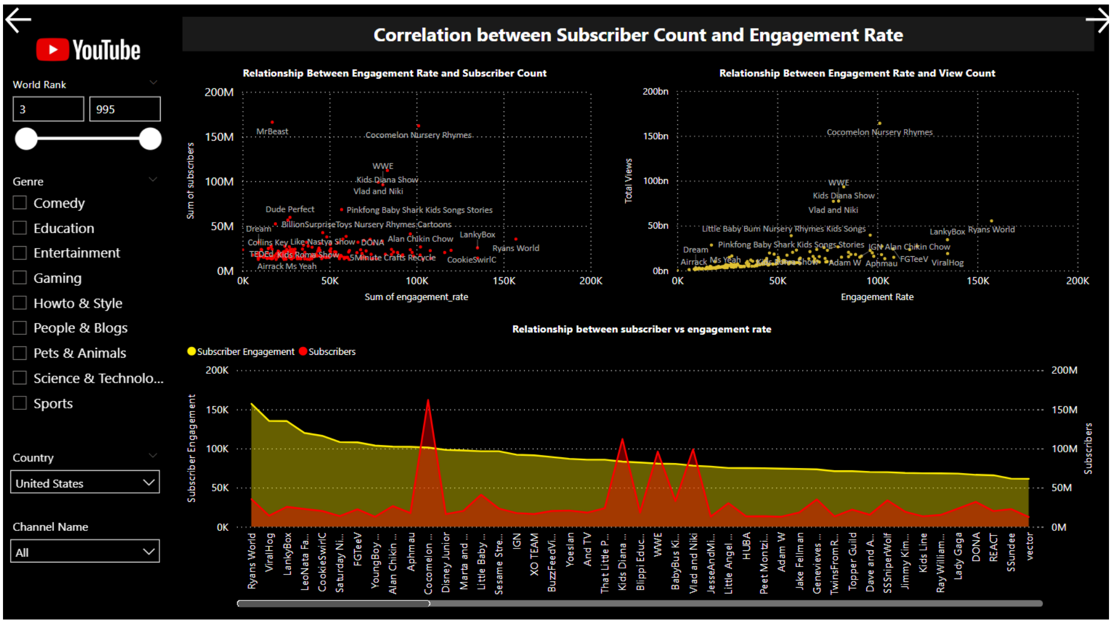
World Rank
3 995

Genre
☑ Select all
☑ Comedy
☑ Education
☑ Entertainment
☑ Gaming
☑ Howto & Style
☑ People & Blogs
☑ Pets & Animals
☑ Science & Technology
☑ Sports
☑ Travel & Events

Country
All

Channel Name
All

World Rank	Channel Name	Subscribers (M)	category	Country rank	Avg Views per Video (M)	Engagement rate
3	MrBeast	166.00	Entertainment	1	38.28	17,089.66
4	Cocomelon Nursery Rhymes	162.00	Education	2	169.77	101,234.57
7	Kids Diana Show	112.00	People & Blogs	3	83.93	83,256.29
8	PewDiePie	111.00	Gaming	1	6.16	26,178.42
9	Like Nastya	106.00	People & Blogs	5	183.53	85,357.60
10	Vlad and Niki	98.90	Entertainment	5	134.46	78,038.59
12	WWE	96.00	Sports	6	1.10	80,654.66
14	BLACKPINK	89.80	People & Blogs	1	59.20	35,795.77
17	5Minute Crafts	80.10	Howto & Style	4797	26,236.79	32,755.04
22	Zee TV	70.50	Entertainment	6	0.57	103,743.34
23	Pinkfong Baby Shark Kids Songs Stories	68.20	Education	8	13.56	56,954.88



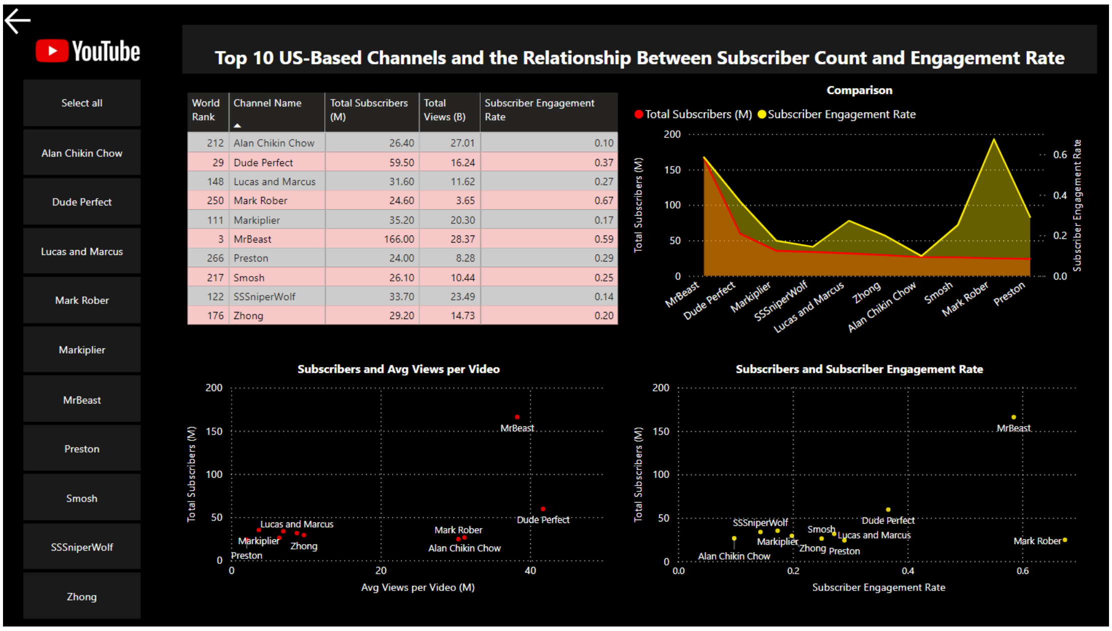
Correlation between Subscriber Count and Engagement Rate

World Rank
3 995

Genre
☐ Comedy
☐ Education
☐ Entertainment
☐ Gaming
☐ Howto & Style
☐ People & Blogs
☐ Pets & Animals
☐ Science & Technolo...
☐ Sports

Country
United States

Channel Name
All



Top 10 US-Based Channels and the Relationship Between Subscriber Count and Engagement Rate

World Rank	Channel Name	Total Subscribers (M)	Total Views (B)	Subscriber Engagement Rate
212	Alan Chikin Chow	26.40	27.01	0.10
29	Dude Perfect	59.50	16.24	0.37
148	Lucas and Marcus	31.60	11.62	0.27
250	Mark Rober	24.60	3.65	0.67
111	Markiplier	35.20	20.30	0.17
3	MrBeast	166.00	28.37	0.59
266	Preston	24.00	8.28	0.29
217	Smosh	26.10	10.44	0.25
122	SSSniperWolf	33.70	23.49	0.14
176	Zhong	29.20	14.73	0.20

Analysis

Findings

- United States has the most subscribers worldwide.
- Channels focused on children's content, entertainment, and celebrities tend to have the highest views and engagement rates, but this data is considered noise and has been removed.
- Here are top 10 channels after removing noise. We manually reviewed each channel by examining the type of content they produce and here are what we found:

Channel Name	Comment
Mr Beast	Exceeds the campaign budget.
Dude perfect	Potential replacement for Mr. Beast.
Markiplier	Focuses primarily on film-related content.
SSSniperWolf	Known for reaction videos, which cater to a somewhat niche audience.
Lucas and Marcus	Only certain videos achieve significant viewership.
Zhong	Content is vague and lacks clear focus.
Alan Chikin Chow	Content is not highly relevant to our project aim.
Smosh	Each video typically garners fewer than 1M views.
Mark Rober	Offers great reach but focuses on a niche audience with science-based content.
Preston	Known for experimentation-style content.

Calculating YouTube Sponsorship Rate Using Excel and SQL

CPM (Cost Per Mille) based formula:
Sponsorship Rate = (Average Views Per Video / 1000) × CPM

Mr Beast
CPM Range: \$30 to \$100+
Estimated Sponsorship Rate = (38,280,000 /1000) * 30
= From \$**1,148,400** to \$**3,828,000+** per video

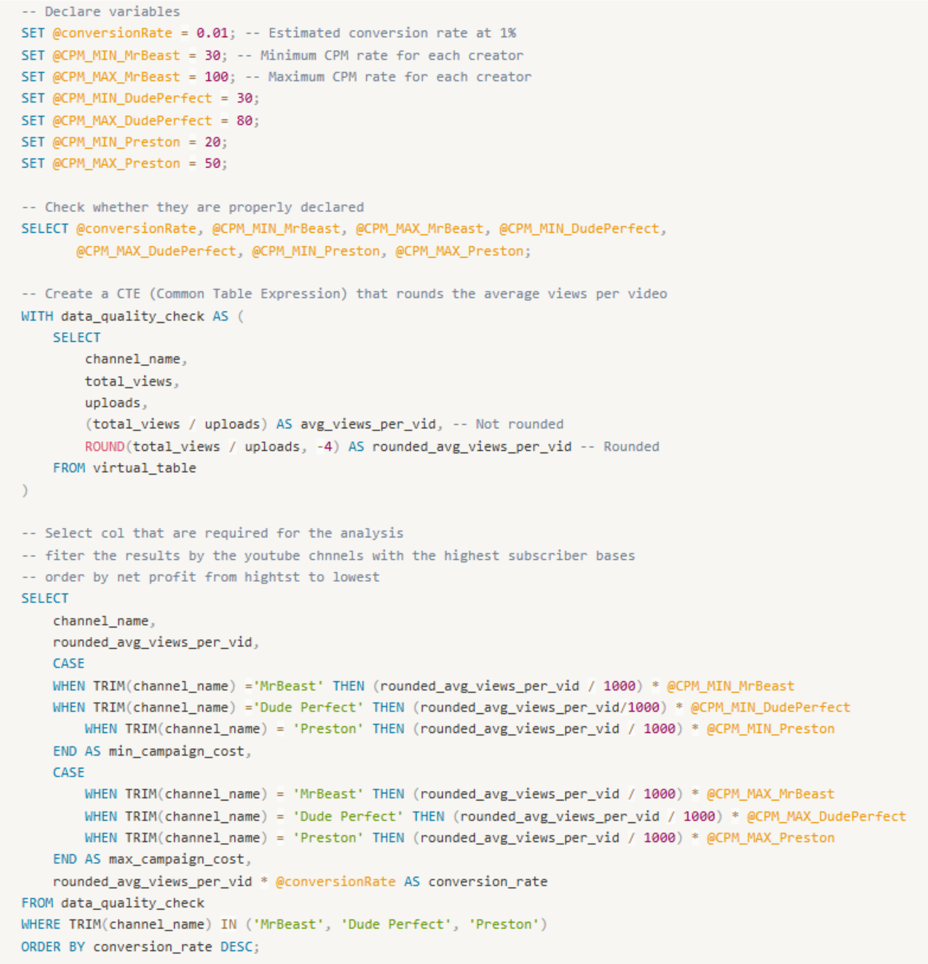
Dude Perfect
CPM Range: \$30 to \$80+
Estimated Sponsorship Rate = (41,750,000 / 1000) * 30
= From \$1,252,500 to \$**3,340,000**+ per video

Preston
CPM Range: \$20 to \$50+
Estimated Sponsorship Rate = (2,070,000 / 1000) * 20
= From **\$41,400** to **\$103,500+** per video

Given Conversion rate is 0.01:

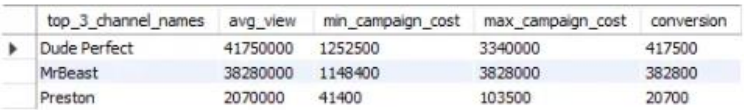
Top 3 Channel Names	Avg Views	Campaign Cost (Min-Max)	Estimated Conversion
Mr Beast	38,280,000	1,148,400 - 3,828,000	382,800
Dude Perfect	41,750,000	1,252,500 - 3,340,000	417,500
Preston	2,070,000	1,252,500 - 3,340,000	20,700

Perform calculations within SQL to confirm the Excel calculations are accurate.



```
-- Declare variables
SET @conversionRate = 0.01; -- Estimated conversion rate at 1%
SET @CPM_MIN_MrBeast = 30; -- Minimum CPM rate for each creator
SET @CPM_MAX_MrBeast = 100; -- Maximum CPM rate for each creator
SET @CPM_MIN_DudePerfect = 30;
SET @CPM_MAX_DudePerfect = 80;
SET @CPM_MIN_Preston = 20;
SET @CPM_MAX_Preston = 50;

-- Check whether they are properly declared
SELECT @conversionRate, @CPM_MIN_MrBeast, @CPM_MAX_MrBeast, @CPM_MIN_DudePerfect,
       @CPM_MAX_DudePerfect, @CPM_MIN_Preston, @CPM_MAX_Preston;

-- Create a CTE (Common Table Expression) that rounds the average views per video
WITH data_quality_check AS (
    SELECT
        channel_name,
        total_views,
        uploads,
        (total_views / uploads) AS avg_views_per_vid, -- Not rounded
        ROUND(total_views / uploads, -4) AS rounded_avg_views_per_vid -- Rounded
    FROM virtual_table
)

-- Select col that are required for the analysis
-- fiter the results by the youtube chnnels with the highest subscriber bases
-- order by net profit from hightst to lowest
SELECT
    channel_name,
    rounded_avg_views_per_vid,
    CASE
    WHEN TRIM(channel_name) ='MrBeast' THEN (rounded_avg_views_per_vid / 1000) * @CPM_MIN_MrBeast
    WHEN TRIM(channel_name) ='Dude Perfect' THEN (rounded_avg_views_per_vid/1000) * @CPM_MIN_DudePerfect
        WHEN TRIM(channel_name) = 'Preston' THEN (rounded_avg_views_per_vid / 1000) * @CPM_MIN_Preston
    END AS min_campaign_cost,
    CASE
        WHEN TRIM(channel_name) = 'MrBeast' THEN (rounded_avg_views_per_vid / 1000) * @CPM_MAX_MrBeast
        WHEN TRIM(channel_name) = 'Dude Perfect' THEN (rounded_avg_views_per_vid / 1000) * @CPM_MAX_DudePerfect
        WHEN TRIM(channel_name) = 'Preston' THEN (rounded_avg_views_per_vid / 1000) * @CPM_MAX_Preston
    END AS max_campaign_cost,
    rounded_avg_views_per_vid * @conversionRate AS conversion_rate
FROM data_quality_check
WHERE TRIM(channel_name) IN ('MrBeast', 'Dude Perfect', 'Preston')
ORDER BY conversion_rate DESC;
```

Output



	top_3_channel_names	avg_view	min_campaign_cost	max_campaign_cost	conversion
▶	Dude Perfect	41750000	1252500	3340000	417500
	MrBeast	38280000	1148400	3828000	382800
	Preston	2070000	41400	103500	20700

Both results (Excel and SQL) are identical.

- **Mr. Beast (166M)** would be the best option to maximise reach and ROI due to his large subscriber base, but the campaign cost is extremely high.
- **Dude Perfect (59.50M)** could deliver a similar outcome to Mr. Beast with a slightly lower budget. Despite having a lower subscriber base, their audience is more engaged with the content than Mr Beast's.
- **Preston (24M)** has the least conversion and engagement, but the channel is still widely known and could be a viable option within a budget-conscious strategy.

Recommendation and Action Plan

If the goal is solely to maximise reach and conversions, Mr Beast would be the best option to pursue with. While other channels, like Dude Perfect, have similar or even higher engagement rates than Mr. Beast, his brand awareness makes him the most reliable choice for ROI. For cost-effectiveness, Preston would be the ideal option, although his impact may be less certain compared to Mr Beast.

We will follow up with our client (Head of Marketing) to understand their expectations for this collaboration. Once we predict that we're on track to hit the KPIs, we will move forward with a potential partnership with one of the creators.

After reaching out and negotiating contracts, we will track each creator's performance against the KPIs. We will review how the campaigns have performed, gather insights, and optimise based on feedback from converted customers and each channel's audience.