Full End-to-End Data Project

Link: <https://www.youtube.com/watch?v=mm_sN-Elplg&list=LL&index=44>

1. Setup

Stages:

* Get requirements from the user
* Design the dashboard
* Collect data from the source
* Explore the data and its content
* Clean the data
* Test the data
* Visualize the data
* Analyze the data
* Justify and record your findings

Objective:

To run successful marketing campaigns with the top UK Youtubers

Challenges:

Struggled to find the right info on the internet from 3rd party providers from the BI reporting team.

Solution:

To create a simple dashboard that displays the top UK Youtubers by subscribers, views and video uploads.

Use Cases:

1 – Identify the top Youtubers to run campaign with

User Story:

As the Head of Marketing, I want to identify the to youtubers in the UK based on their subscription numbers, uploaded videos and views accumulated, so that I can decided which channels would be best to run a marketing campaign with to generate good ROI

Acceptance Criteria:

The dashboard should –

* List the top Youtube channels by subscribers, videos and views
* Display key metrics (channel name, subscribers, videos, views, engagement ratios)
* Be user-friendly and easy to filter/sort
* Use the most recent data possible

2 – Analyze the potential for marketing campaigns with Youtubers

User Story:

As the Head of Marketing, I want to analyze the potential for successful campaigns with the top Youtubers so that I can maximize ROI.

Acceptance criteria:

The solution should –

* Recommend Youtube channels best suited for different campaign types (e.g. product placement, sponsored video series, influencer marketing)
* Consider reach, engagement and potential revenue based on estimated conversion rates
* Clearly explain the recommendations with data-driven justifications

Success criteria

Sharon can —

* Easily identify the top performing Youtube channels based on the key metrics mentioned above
* Assess the potential for successful campaigns with top youtubers based on reach, engagement and potential revenue
* Make informed decisions on the ideal collaborations to advance with based on recommendations

This allows Sharon to achieve a good ROI and build relationships with Youtubers for future collaborations, which leads to recognition within the company.

Information needed

Sharon needs the top Youtubers in the UK, and the key metrics needed include:

* Subscriber count
* Videos uploaded
* Views
* Average views
* Subscriber engagement ratio
* Views per subscriber

Data needed

The dataset to produce the information we need should include the following fields

* Channel name(string)
* Total subscribers(integer)
* Total videos uploaded(integer)
* Total views(integer)

We’ll focus on the top 100 Youtubers for simplicity sake.

Data quality checks

We need to add measures in place to confirm the dataset contains the data required without any issues – here are some of the data quality checks we need to conduct.

* Row count check
* Column count check
* Data type check
* Duplicate check

Additional requirements

* Document the solution and include the data sources, transformation processes and walk through on analysis conclusions.
* Make source code and docs available on Github
* Ensure the solution is reproducible and maintainable so that it can support future updates

Design the Dashboard

A screenshot of a computer

Description automatically generated

Data Source

Link: <https://www.kaggle.com/datasets/bhavyadhingra00020/top-100-social-media-influencers-2024-countrywise>

Then look for the United kingdom folder and inside select the youtube dataset and download it.

A screenshot of a computer

Description automatically generated

2. Explore the Data

This stage is where we explore the data, see what we got and if theres anything wrong with it.

A screenshot of a computer

Description automatically generated

As you can see, there will be a lot of cleanup required. So these are some initial observations.

* The columns are in a different language
* The channel names also have their channel ids inside the same column
* Followers Count is using string instead of integer
* TP column is ambiguous at first and doesn’t tell us what it means and that has a couple percentage values with majority of it in blank
* Country column has United Kingdom has data , which is the correct data to have, until a certain point it has a different type of data that makes no sense.
* Influence Theme is missing a lot of data
* Reach Potential same issue as followers count
* The Save column has no data and no purpose
* And invite to campaign has odd data.

3. Using SQL to clean up data

CHARINDEX: Is to find the position of a character inside a string. We are going to use this to locate the @UC so that we can separate it to get the channel name

SQL:

SELECT CHARINDEX(‘@’, NOMBRE), NOMBRE from top\_uk\_youtubers\_2024

SUBSTRING: Then we are going to insert this inside the SUBSTRING function that will use the position to cutoff the channel name from NOMBRE column

SQL:

SELECT CAST(SUBSTRING(NOMBRE, 1, CHARINDEX(‘@’, NOMBRE)-1) AS VARCHAR(100)) AS channel\_name from top\_uk\_youtubers\_2024

We can add the other column names to get the table we want

SQL:

SELECT CAST(SUBSTRING(NOMBRE, 1, CHARINDEX(‘@’, NOMBRE)-1) AS VARCHAR(100)) AS channel\_name, total\_views, total\_subscribers, total\_videos FROM top\_uk\_youtubers\_2024

VIEWS: We can now turn this into a view for our Power BI so simply add this line

SQL:

**CREATE VIEW view\_uk\_youtubers\_2024 AS**

**SELECT CAST(SUBSTRING(NOMBRE, 1, CHARINDEX(‘@’, NOMBRE)-1) AS VARCHAR(100)) AS channel\_name, total\_views, total\_subscribers, total\_videos**

**FROM top\_uk\_youtubers\_2024**

4. Test Data

This is an important step where we run Data Quality tests to see if the data is accurate or has little errors as possible. So lets first ask ourselves what we need to from this dataset to accurately tell if theres a problem or not:

Data Quality Tests

1. The data needs to be 100 records of Youtube channels (row count test)
2. The data needs 4 fields (column count test)
3. The channel name column must be string format, while the others need to be integers
4. Each record must be unique in the dataset (duplicates count test)

Lets first check the row count. We need 100 as result

SQL:

**SELECT COUNT(\*) AS no\_of\_rows**

**FROM view\_uk\_youtubers\_2024**

If it returns 100, it passes.

Now for the column count

SQL:

**SELECT**

**COUNT(\*) as column\_count**

**FROM**

**INFORMATION\_SCHEMA.COLUMNS**

**WHERE**

**TABLE\_NAME = ‘view\_uk\_youtubers\_2024’**

If we get 4, then we passed

Now we are going to check the data type of the columns and see if it corresponds

SQL:

**SELECT**

**COLUMN\_NAME, DATA\_TYPE**

**FROM**

**INFORMATION\_SCHEMA.COLUMNS**

**WHERE**

**TABLE\_NAME = ‘view\_uk\_youtubers\_2024’**

If name is varchar and the other three columns are integer or bigint, they pass.

Now finally to check for duplicate records.

SQL:

**SELECT**

**channel\_name, COUNT(\*) as duplicate\_count**

**FROM**

**view\_uk\_youtubers\_2024**

**GROUP BY**

**Channel\_name**

**HAVING**

**COUNT(\*) > 1**

If it returns an empty table, it means no duplicates, therefore we pass.

If our dataset passed all tests, it means tis ready for use.

5. Power BI

Lets import our dataset to PowerBI. Now we are going to use DAX measures to build the components for our dashboard

A screenshot of a computer

Description automatically generated

First measure we are going to do is Total Subscribers

DAX:

Total Subscribers (M) =

VAR million = 1000000

VAR sumOfSubscribers = SUM(Sheet1[total\_subscribers])

VAR totalSubscribers = DIVIDE(sumOfSubscribers, million)

RETURN totalSubscribers

Next measure is going to be for Total views

DAX:

Total Views (B) =

VAR billion = 1000000000

VAR sumOfViews = SUM(Sheet1[total\_views])

VAR totalViews = DIVIDE(sumOfViews, billion)

RETURN totalViews

Now the measure for total videos

DAX:

Total Videos =

VAR totalVideos = SUM(Sheet1[total\_videos])

RETURN totalVideos

Next measure is for Average Views per video

DAX:

Avg Views per Video (M) =

VAR sumOfTotalViews = SUM(Sheet1[total\_views])

VAR sumOfTotalVideos = SUM(Sheet1[total\_videos])

VAR avgOfViewsPerVideo = DIVIDE(sumOfTotalViews, sumOfTotalVideos, BLANK())

VAR finalAvgViewsPerVideo = DIVIDE(avgOfViewsPerVideo, 1000000, BLANK())

RETURN finalAvgViewsPerVideo

Next is Subscriber Engagement Rate

DAX:

Subscriber Engagement Rate =

VAR sumOfTotalSubscribers = SUM(Sheet1[total\_subscribers])

VAR sumOfTotalVideos = SUM(Sheet1[total\_videos])

VAR subscriberEngRate = DIVIDE(sumOfTotalSubscribers, sumOfTotalVideos, BLANK())

RETURN subscriberEngRate

And finally the last measure we are going to create is Views per Subscriber

DAX:

Views per Subscriber =

VAR sumTotalViews = SUM(Sheet1[total\_views])

VAR sumTotalSubscribers = SUM(Sheet1[total\_subscribers])

VAR viewsPerSubscriber = DIVIDE(sumTotalViews, sumTotalSubscribers, BLANK())

RETURN viewsPerSubscriber

We have our measures completed, now time to use them to build the components of our dashboard

Lets first add our title by clicking and dragging a text box into our panel. Adjust the font size and position

A screenshot of a computer

Description automatically generated

Lets add our first table in the dashboard and pick name, total views, total subscribers and total videos.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Now lets change its visuals

A screenshot of a computer

Description automatically generated

Now lets add our tree map and is where we are going to visualize the top youtubers with highest views

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

We could lower the amount of data here just to show the top 10 channels instead. For that we need to go to the filter and make some changes

A screenshot of a computer

Description automatically generated

And we can add some data to the tooltips to make the information richer when hovering over the data

A screenshot of a video

Description automatically generated

Now lets add another component. The score cards, we will add 3 of them side by side and each one respectively will use the Avg Views per Video, Subscriber Engagement and then Views per Subscriber

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Now finally the last component will be the Horizontal Bar chart that will have channel name by total susbcribers, Will need to filter to only show the top 10

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Now we are done in adding our components, so we will need to format the dashboard visually by changing colors and formatting some indexes

A screenshot of a computer

Description automatically generated

f