Task #3: Data profiling

The task is to perform data profiling for the dataset. Expected output of the task is:

### 1. brief description of how you checked the data. If needed screenshot(s), code/query

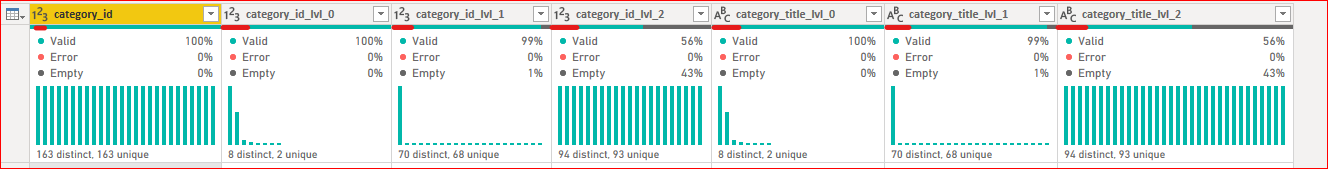
### snippets, excel formulas can be added and explained.

Answer: perform data profiling task, I have used power BI’s Power Query Editor (Used Get data 🡪 Text/CSV) and PostgreSQL database (Used Pentaho ETL tool to pull data).

Csv file categories:

Power BI tool:

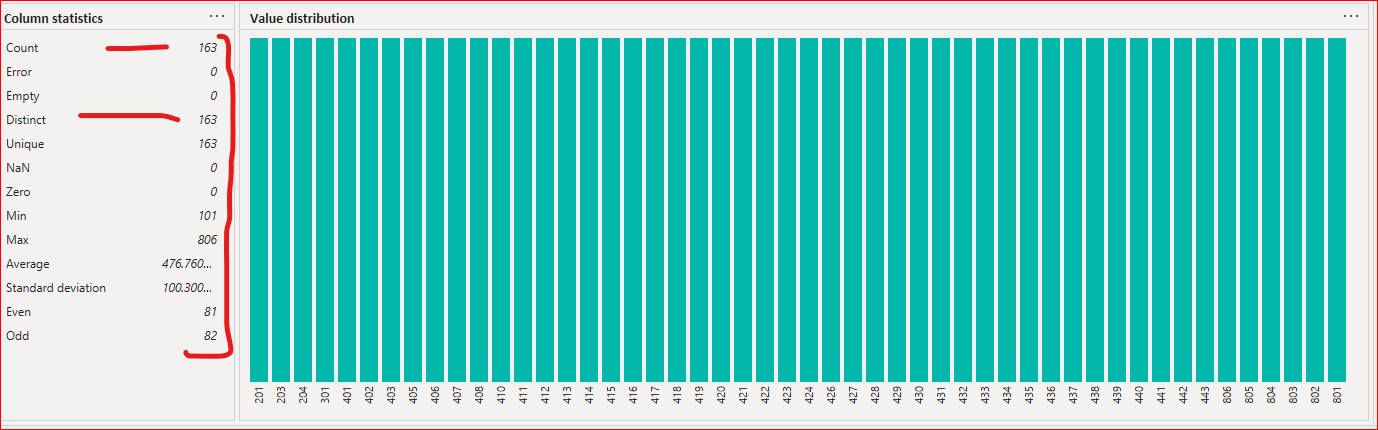
Data type of all the columns look correct (marked red) and data validation results are also looking ok.



Check data quality on column level:

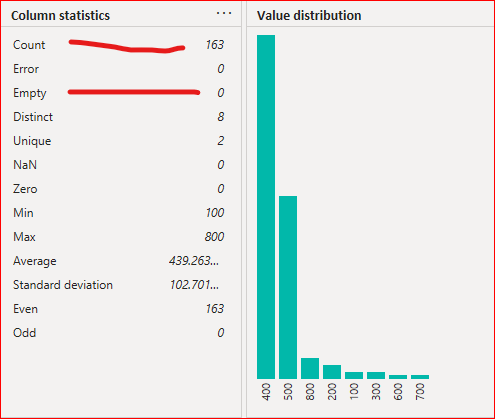
category\_id: data type is integer.

Total number of column count is 163 and all are distinct. Zero empty rows, min value is 101 and 806. Please see in below picture:



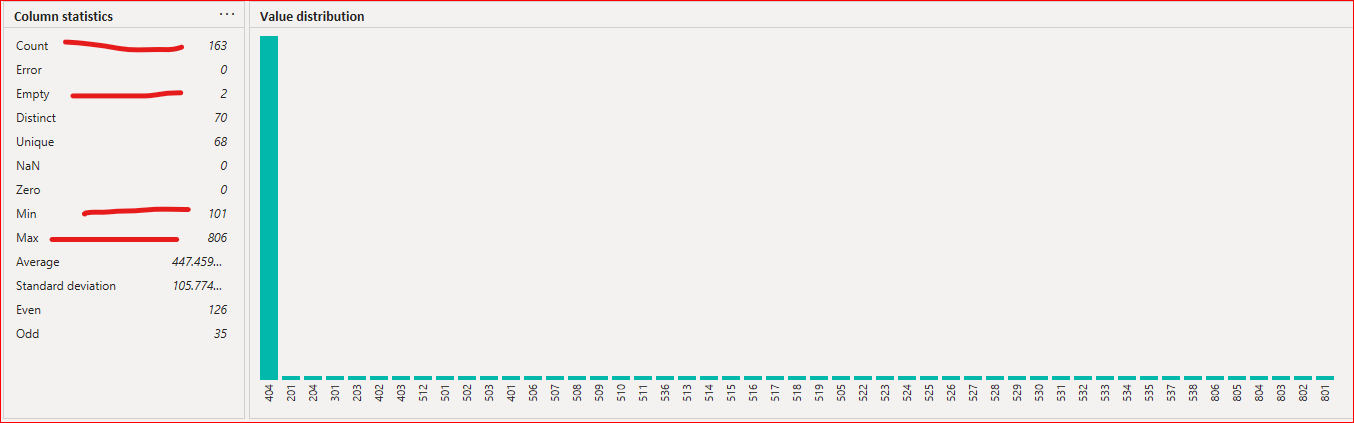
category\_id\_lvl\_0: data type is integer.

Total number of column count is 163 and Zero empty rows, min value is 100 and 800. Please see in below picture:



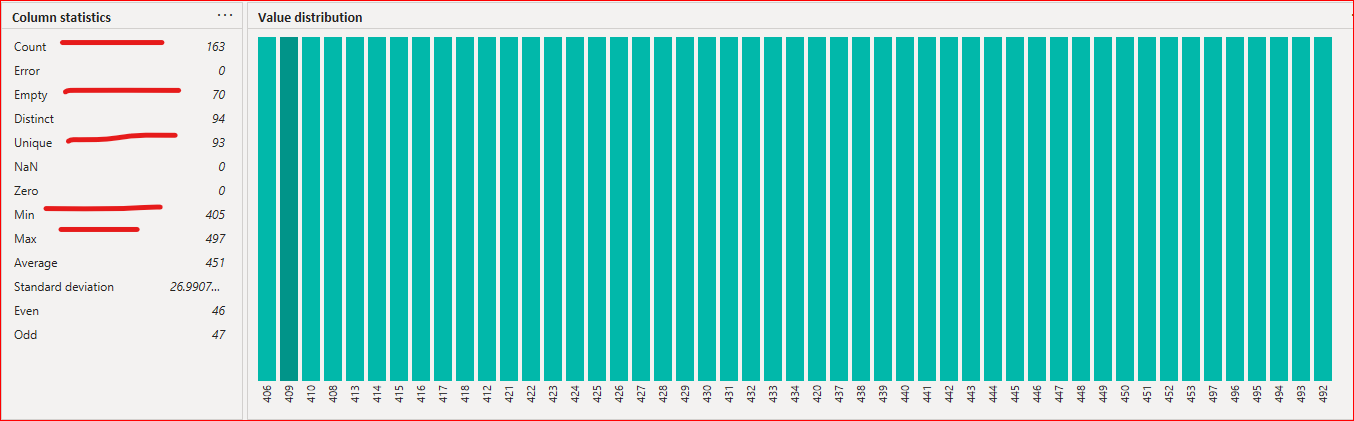
category\_id\_lvl\_1 : data type is integer.

Total number of column count is 163 and 2 empty rows, min value is 101 and 806. Please see in below picture:



category\_id\_lvl\_2 : data type is integer.

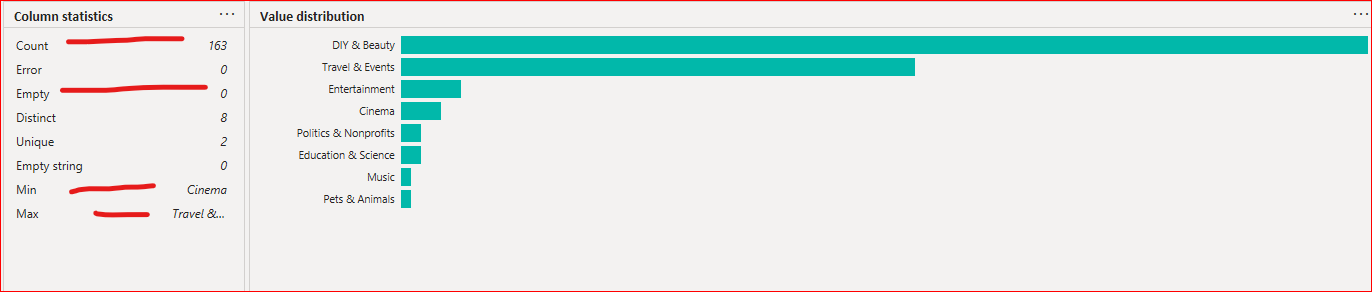
Total number of column count is 163 and 70 empty rows, min value is 405 and 497. Please see in below picture:



category\_title\_lvl\_0: data type is string.

This column data statistics are matching with column category\_id\_lvl\_0.

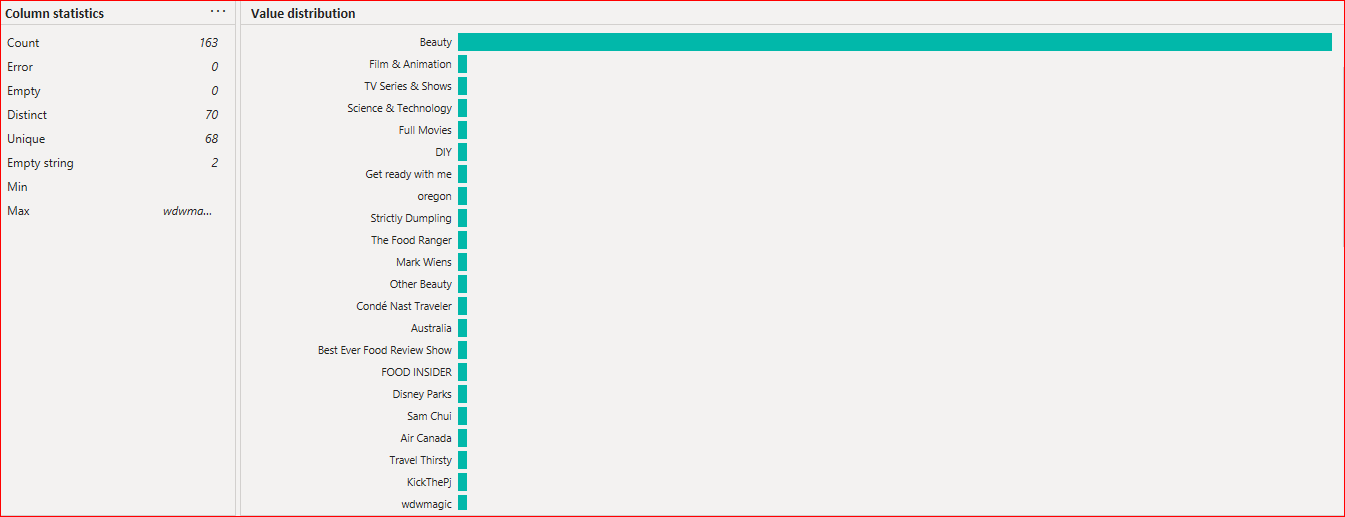
Please see in data below picture:



category\_title\_lvl\_1 : data type is string.

This column data statistics are matching with column category\_id\_lvl\_1.

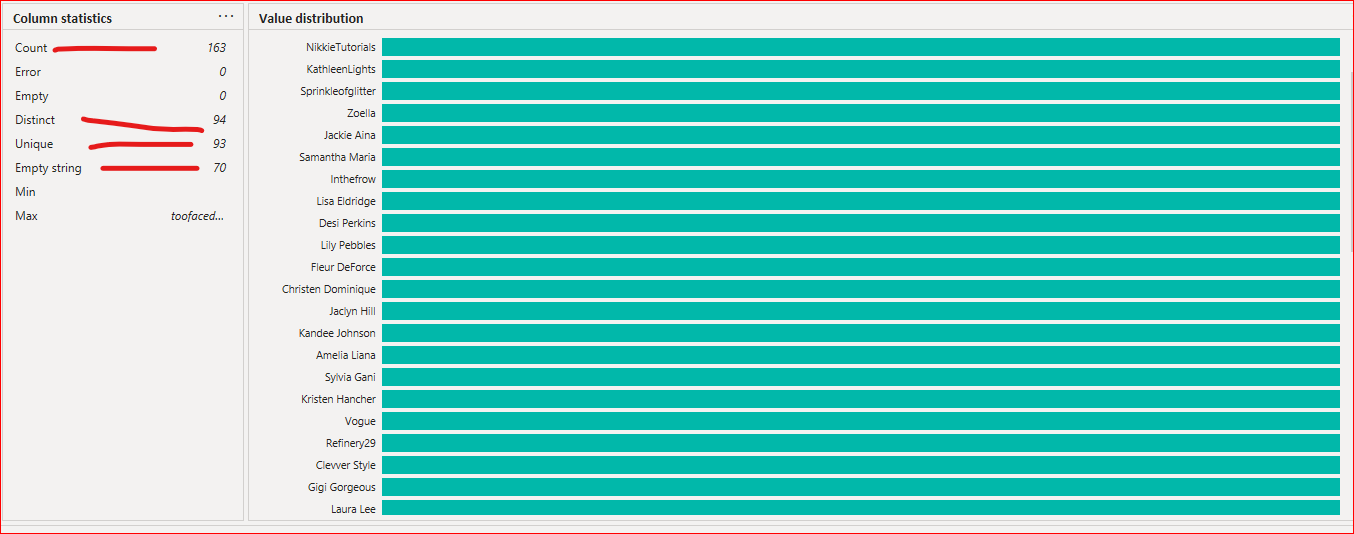
Please see in data below picture:



category\_title\_lvl\_2 : data type is string.

This column data statistics are matching with column category\_id\_lvl\_2.

Please see in data below picture:

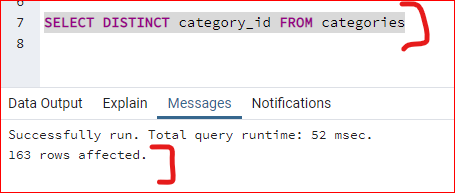


Check data via SQL queries

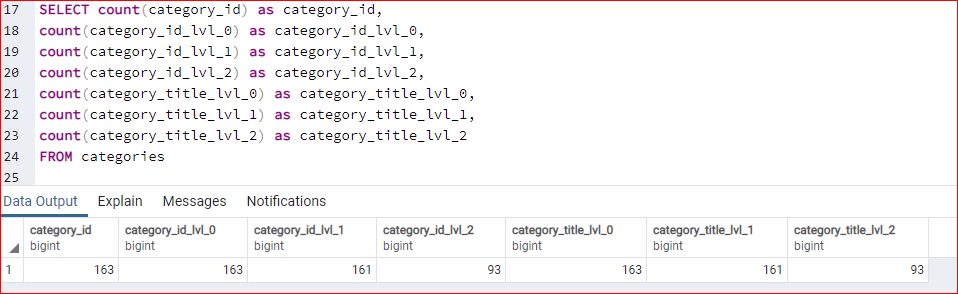
I have created test table in PostgreSQL database.

Create table query:

1. This distinct count based on category\_id is 163 rows and our result is matching with Power BI result. Therefore, we can use column category\_id as a primary key in table.



1. Number of rows count are matching one to one. For example , column category\_title\_lvl\_2 has 93 rows same as category\_id\_lvl\_2. Therefore, data is not missing in any columns.



1. Categories are assigned as per the description.

‘Categories are assigned based on a tree-based structure. It has three-level depth, where the

highest level - named lvl\_0 - can have lvl\_1 sub-category, and lvl\_1 level can have lvl\_2

sub-category. Each transaction is described with at least one level and, at most, three level granularity; and the final category for each transaction is always assigned to the lowest possible

level.

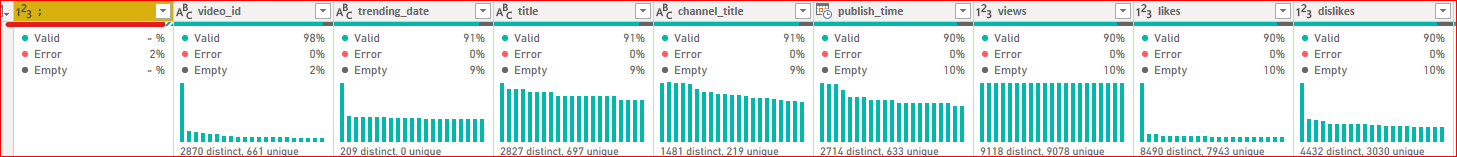
’

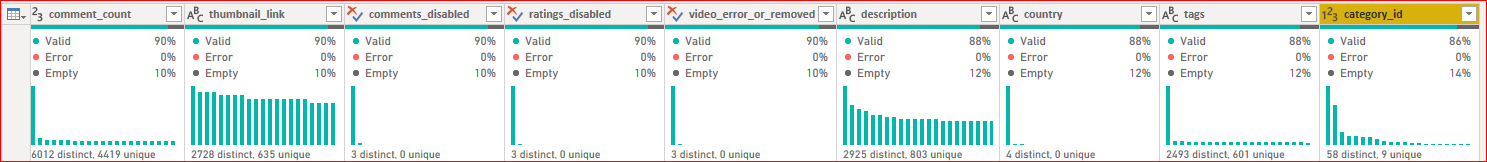


Csv file GB\_videos:

Power BI tool:

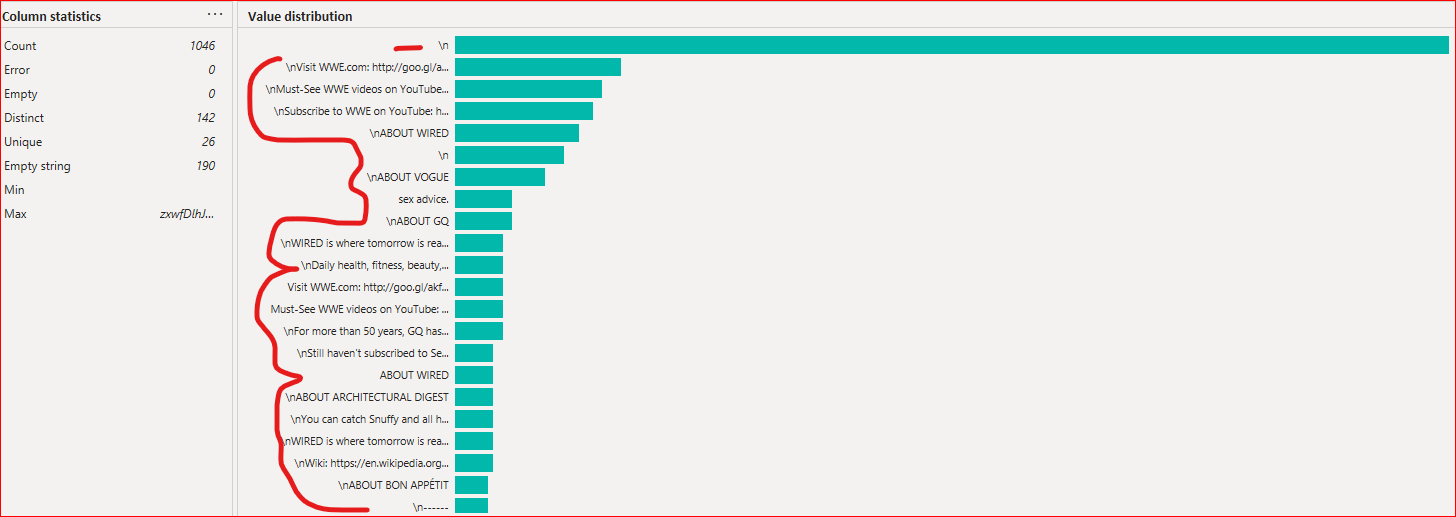
First column has error correct (marked red) and column trending\_date has a string data type but it should be date. However, other columns have correct data types and no Error found.





Individual column profile:

video\_id: data type is string and there are issues with data quality. Video\_id is word or special characters.

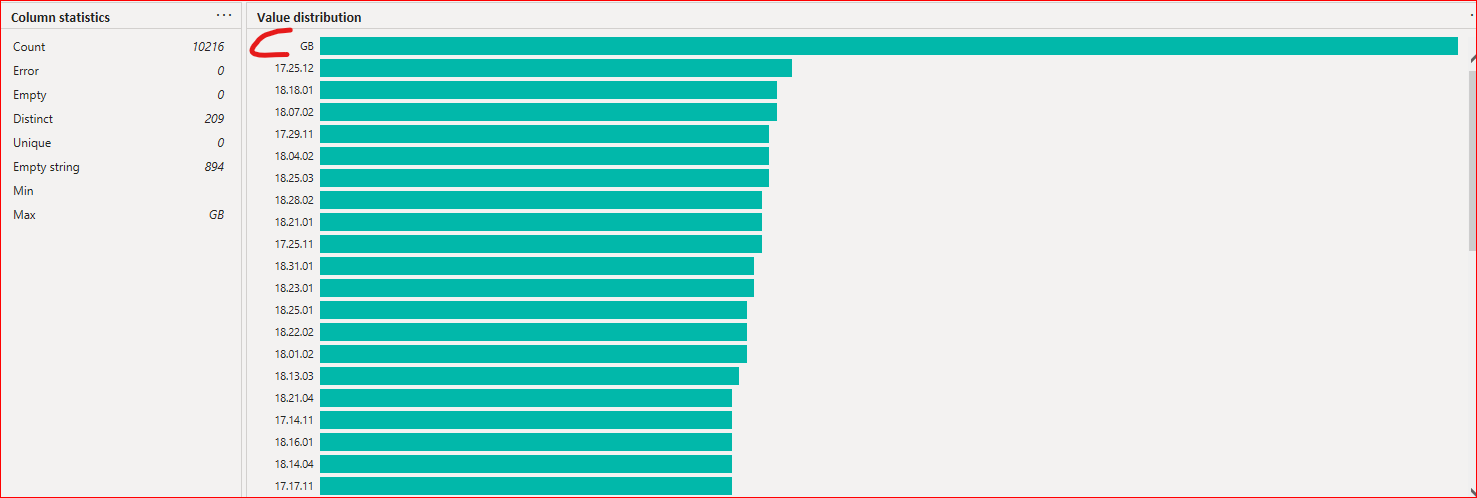


String length in video\_id. Please see data in below picture:



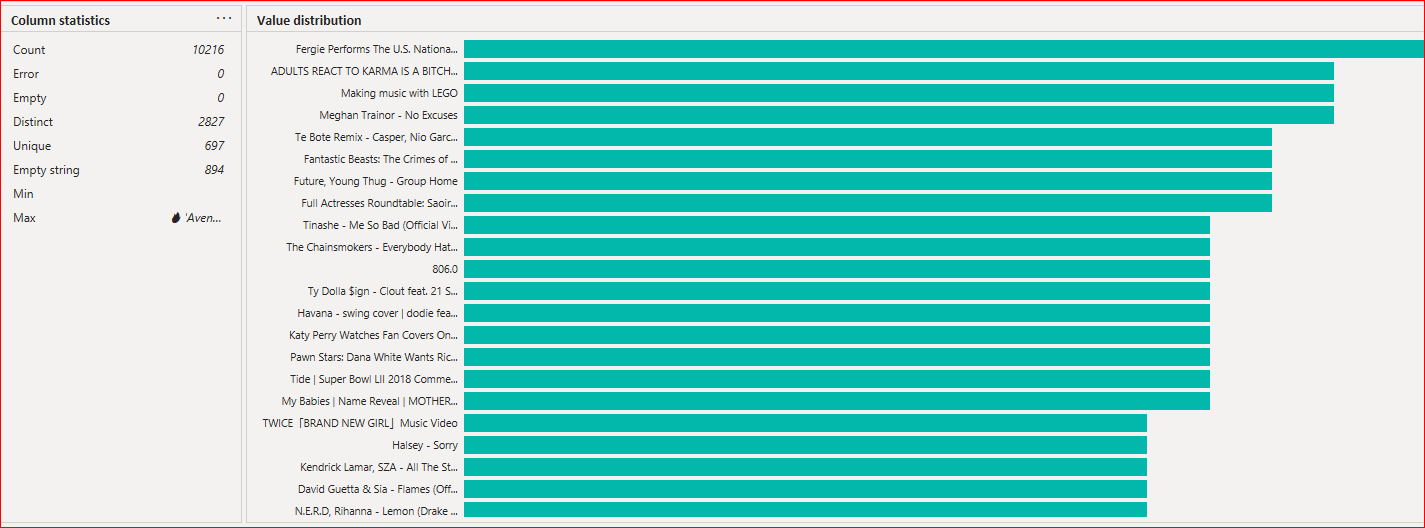
trending\_date:

Date type is a string but it should be date. Word and date are in one column. Therefore, data are not correct in this column.



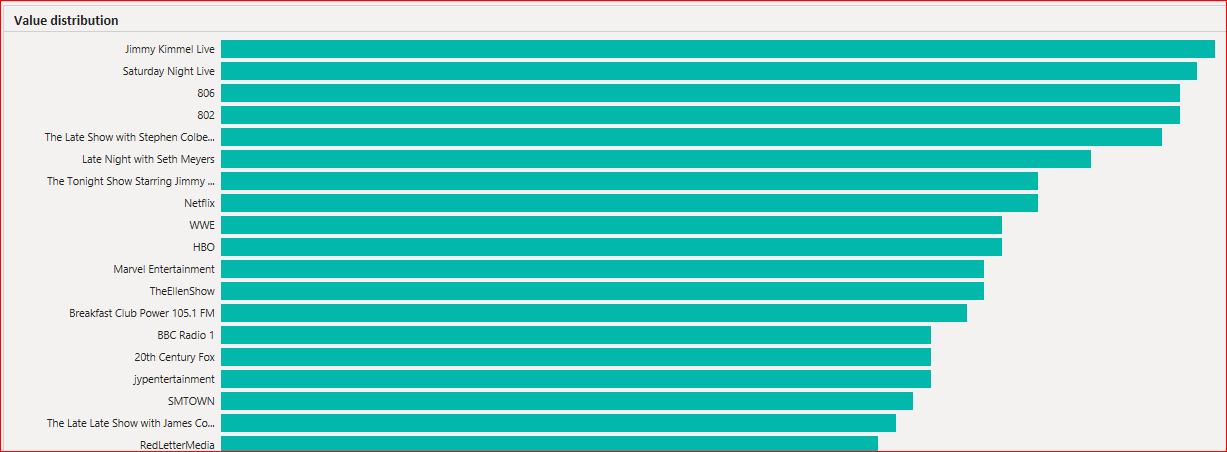
Title:

Data type is string and there are issues with data quality.



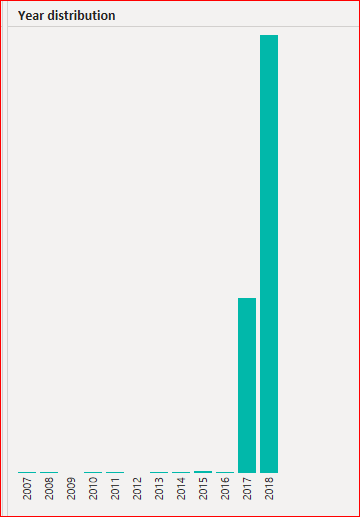
channel\_title:

Data type is string and there are issues with data quality.



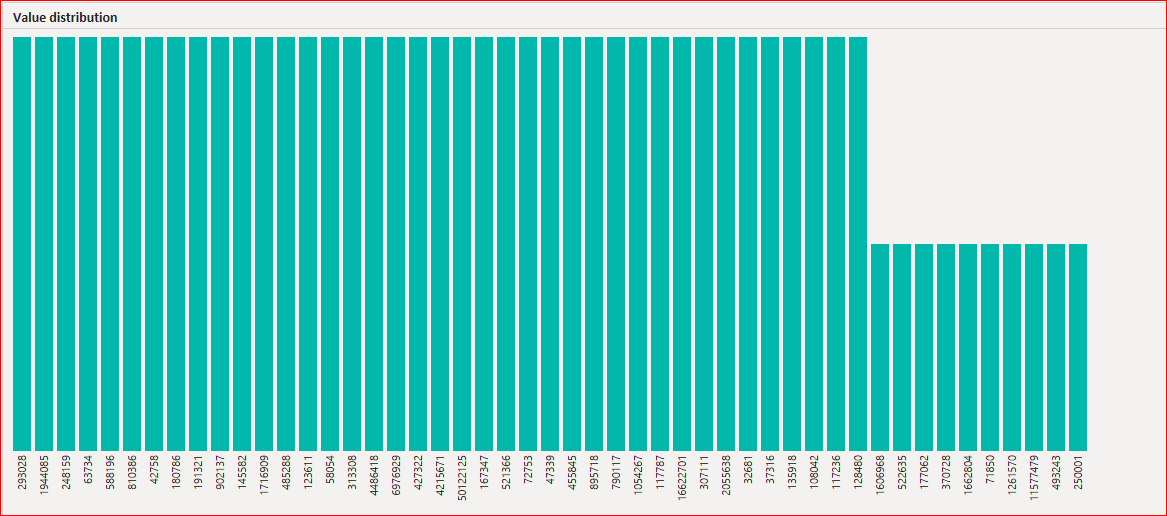
publish\_time:

Data type is datetime and we have data from 2007 till 2018. It will help us to create calendar table in Power BI.



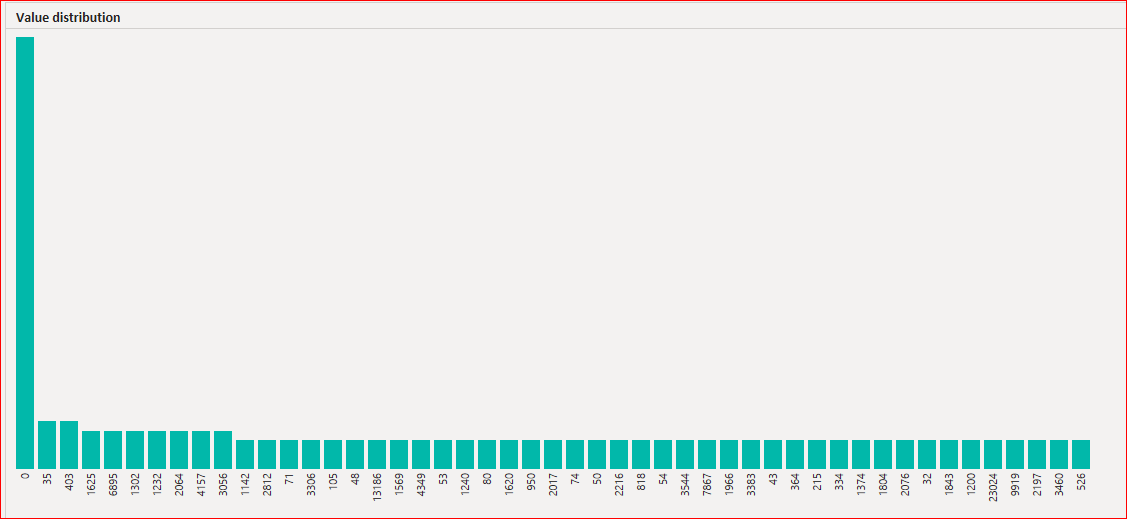
Views:

Data type is integer and have null values



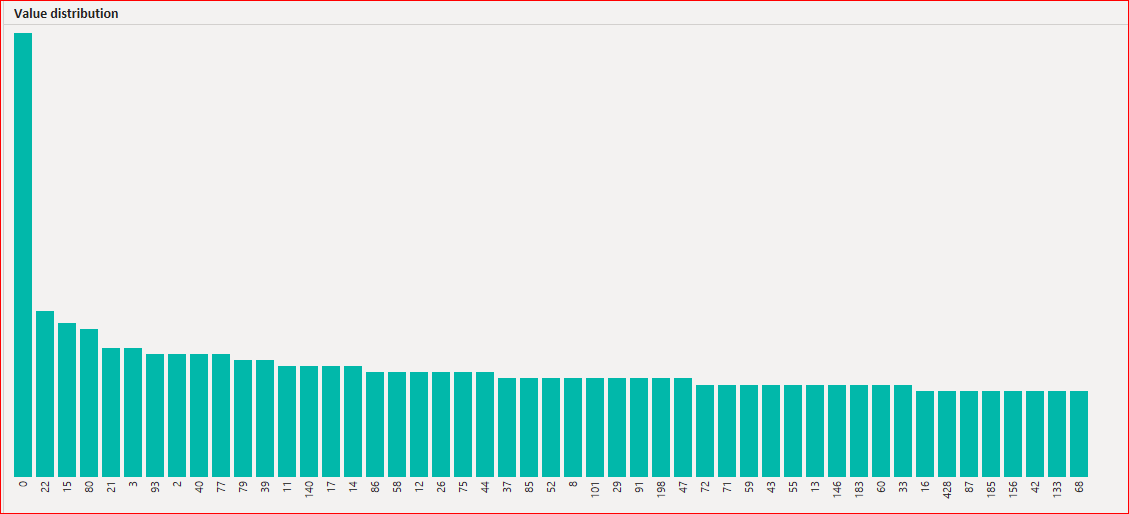
Likes:

Data type is integer and have null values



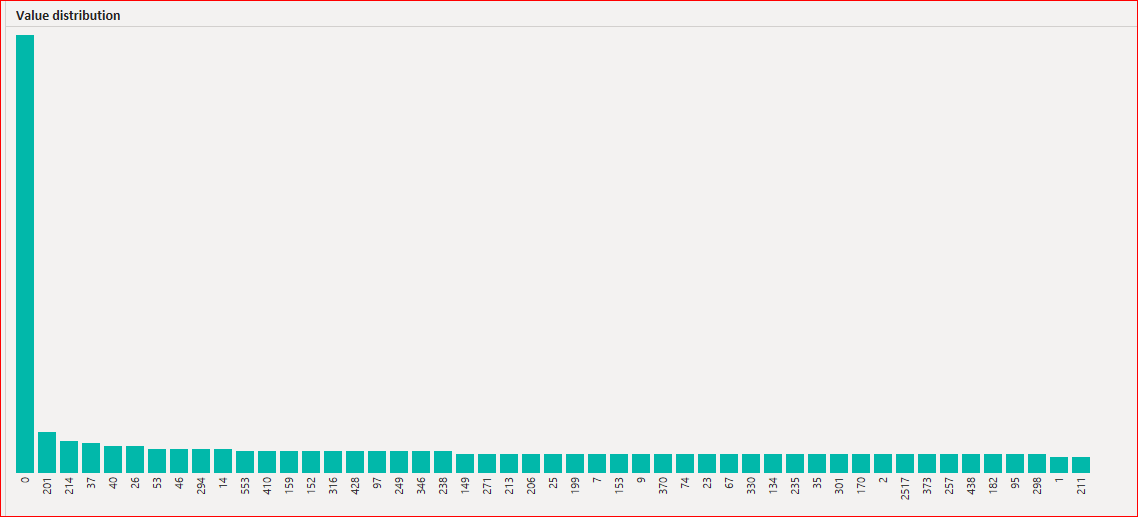
Dislikes:

Data type is integer and have null values



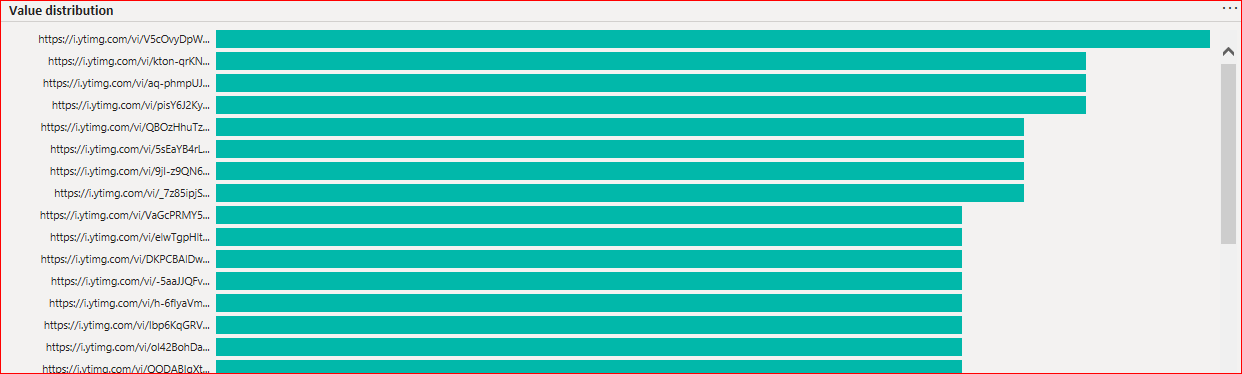
comment\_count:

Data type is integer and have null values



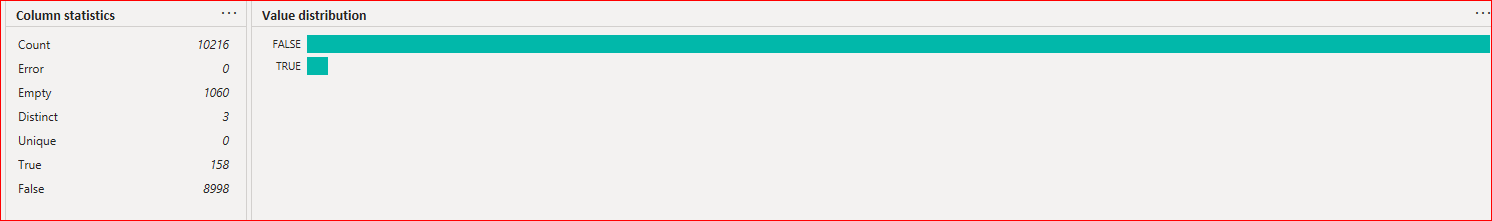
thumbnail\_link:

Data type is string and have null values



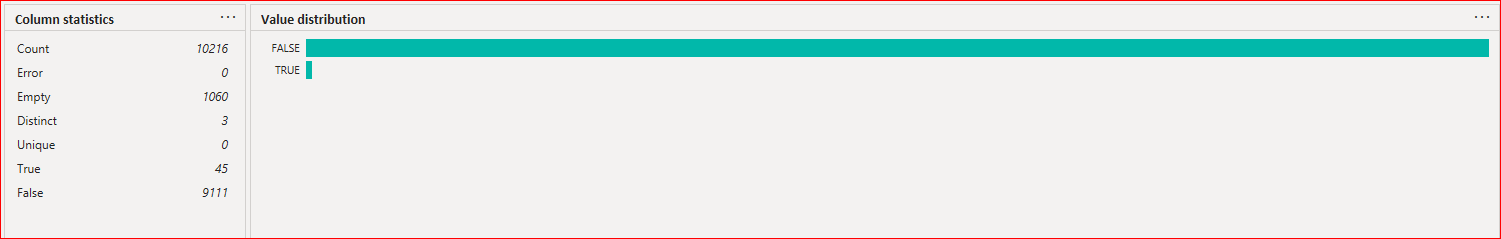
comments\_disabled:

Data type is boolean and have null values



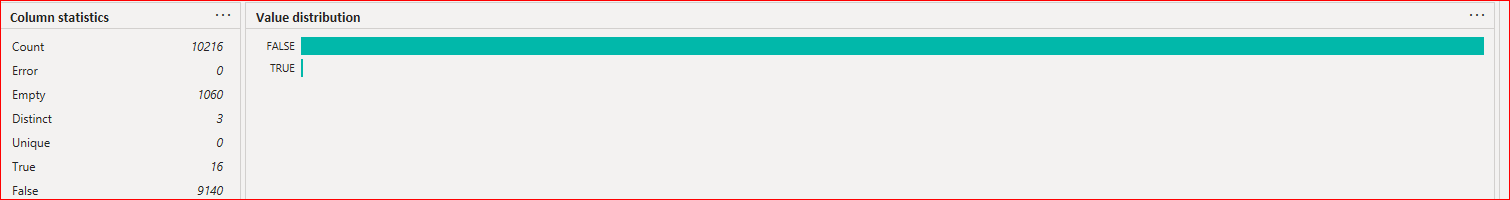
ratings\_disabled:

Data type is boolean and have null values



video\_error\_or\_removed:

Data type is boolean and have null values



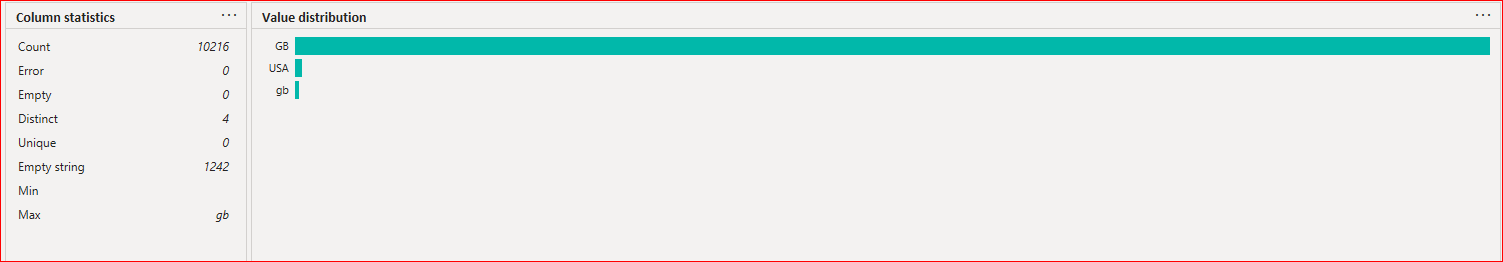
Description:

Data type is string and have null values



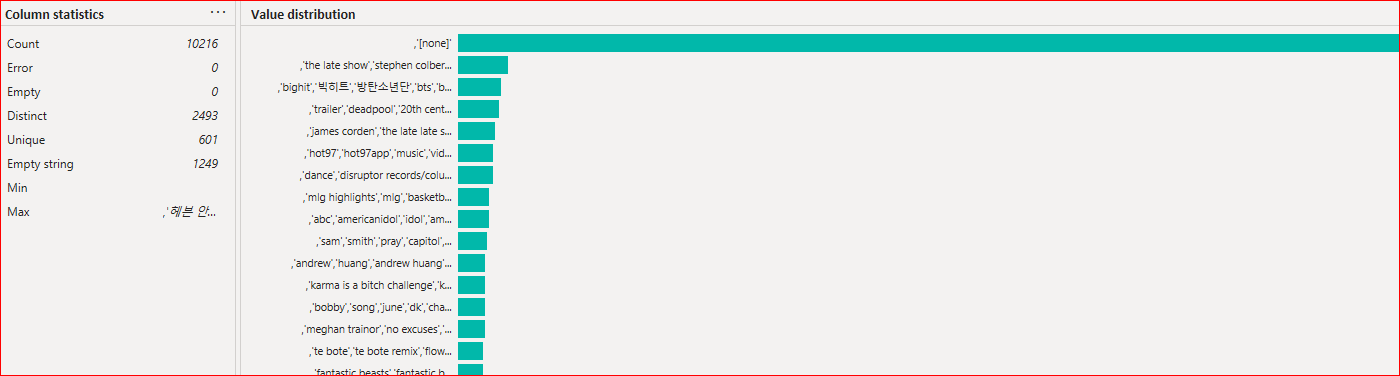
Country:

Data type is string and have null values



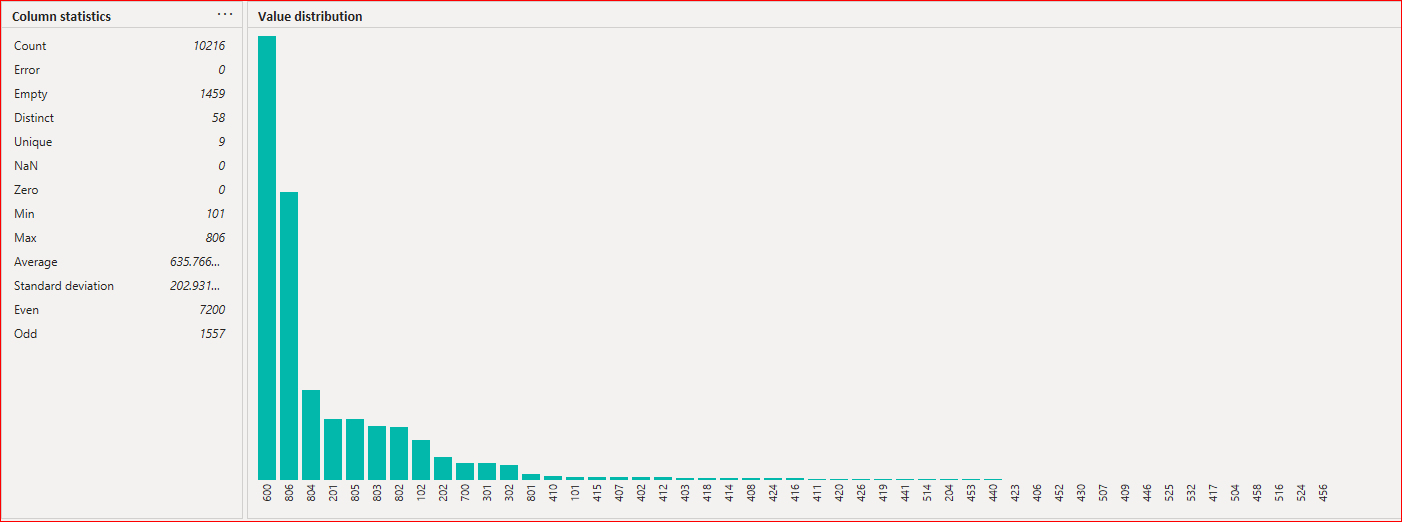
Tags:

Data type is string and have null values



category\_id:

Data type is integer and have null values



### 2. brief description of non-trivial data quality issues you noticed if any;

Table categories data quality issue:

I have not seen any major data quality issue in Table categories. There are Letter Case issues like capital letters, small letters in column category\_title\_lvl\_1 and column category\_title\_lvl\_2. Please see in below screenshots:



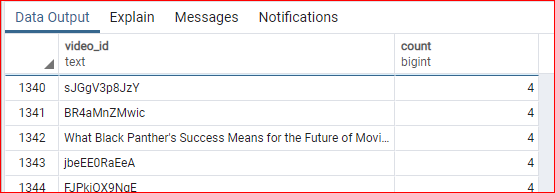


Table GB\_videos data quality issue:

video\_id:

Here are the following data issues:

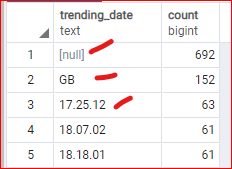
Full length sentences, website urls, null values and duplicate ids.



trending\_date:

Here are the following data issues:

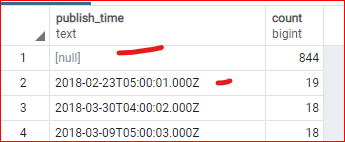
Date format, null values and word.



publish\_time:

Here are the following data issues:

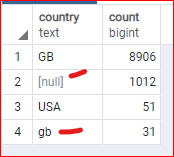
Date format and null values.



Country:

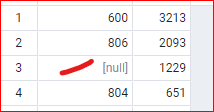
Here are the following data issues:

GB is in small and capital letter and null values.



category\_id

null values.



### 3. brief description of non-trivial improvements in data processing you would suggest if

### any;

Table categories:

INITCAP() function to convert a string to a new string that capitalizes the first letter of every word. All other letters will be lowercase.

select

category\_id,

category\_id\_lvl\_0,

category\_id\_lvl\_1,

category\_id\_lvl\_2,

category\_title\_lvl\_0,

INITCAP(category\_title\_lvl\_1) as category\_title\_lvl\_1,

INITCAP(category\_title\_lvl\_2) as category\_title\_lvl\_2

from categories

Result:



OR

In ETL process, you can fix this issue on the fly.

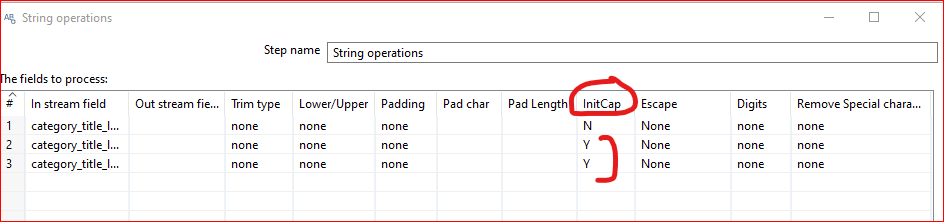


Table GB\_videos:

There are the following data qualities issues will fix if we just gather data where category\_id is not null.

We need to correct date format for two columns (trending\_date and publish\_time) and change GB value from small and capital letters in column Country. All this issues, we can fix in ETL process very easily.

Step 1:

Pull raw data in temp table and put assign data type as text.

Step 2:

* Gather data from temp table
* Perform data manipulation and cleaning task
* Insert clean data with the help of below SQL query in database .

select

video\_id,

TO\_DATE(CONCAT(substring(trending\_date,1,2), '-', substring(trending\_date,4,2), '-', substring(trending\_date,7,2)),'yy-DD-MM') as trending\_date,

title,

channel\_title,

TO\_DATE(CONCAT(substring(publish\_time,1,4), '-', substring(publish\_time,6,2), '-', substring(publish\_time,9,2), ' ', substring(publish\_time,12,8)),'yyyy-MM-dd hh24:mi:SS') as publish\_time,

views,

likes,

dislikes,

comment\_count,

thumbnail\_link,

comments\_disabled,

ratings\_disabled,

video\_error\_or\_removed,

description,

UPPER(country) as country,

tags,

category\_id

from GB\_videos

where category\_id is not null

### 4.your final conclusion whether this data set is good enough for further analysis.

We can join table GB\_videos and Categories with column category\_id. This is data is good for further analysis task.

select

---columns from table GB\_videos

gbv.video\_id,

gbv.trending\_date,

gbv.title,

gbv.channel\_title,

gbv.publish\_time,

gbv.views,

gbv.likes,

gbv.dislikes,

gbv.comment\_count,

gbv.thumbnail\_link,

gbv.comments\_disabled,

gbv.ratings\_disabled,

gbv.video\_error\_or\_removed,

gbv.description,

UPPER(gbv.country) as country,

gbv.tags,

---columns from table categories

gbv.category\_id,

ct.category\_id\_lvl\_0,

ct.category\_id\_lvl\_1,

ct.category\_id\_lvl\_2,

ct.category\_title\_lvl\_0,

ct.category\_title\_lvl\_1,

ct.category\_title\_lvl\_2

from GB\_videos gbv

left join categories ct on gbv.category\_id = ct.category\_id

where gbv.category\_id is not null

Note: I have attached file create\_table\_Task3.sql. In this file, you can find create table scripts for categories, GB\_videos\_temp and GB\_videos.