Task #3: Data profiling

1. brief description of how you checked the data. If needed screenshot(s), code/query snippets, excel formulas can be added and explained

Reviewed CSV file in Excel and used Textjoin to compile tags, then adding RIGHT to find possible tag with category\_id. Then used VLOOKUP to associate category\_id with category titles and combine the tables (GB\_videos and categories).

Afterwards, used Python and Pandas library to review data by looking at data type, null value and correlation between numeric columns. Benefits from using Pandas is that it will create a dataframe from the CSV file and we can use it to perform data analysis and data manipulation – more in this in the following points. Used the following python commands in Jupyter Notebook for the analysis



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

Before concluding on critical data issues, it needs to be understood what is important to the company in terms of data-points and let that guide the analysis. If data is not important and will be excluded from the analysis, then it might not be required to clean or remove these fields.

Looking at the data it is noticed that columns in (publish\_time, views, likes, dislikes, comment\_count, thumbnail\_link, comment\_disabled, ratings\_disabled and video\_error\_or\_removed) have exactly 2080 blank values. These are obvious candidates for deletion or manipulation unless they contain other data points that is vital for further analysis.

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

Empty cells can potentially cause incorrect results hence disrupt the analysis. Rows can be removed if the dataset is large enough to continue the analysis. In Pandas combining the pd.read\_csv and dropna will remove the rows and create a new dataframe to analyse further. Another approach is to use the fillna() method which will keep the row and add a defined value, either chosen by us or we can replace using Mean, Median or Mode (value that appears most often).

If a value is in a wrong format, then we have two options; either delete or attempt to correct the format so we can analyse the dataset. To achieve this we can combine the pd.read\_csv with pd.to\_*dataformat*(). A good approach could be to use a combination of converting the data format (example date; trending\_date) and if not possible, we can use the remove option - dropna. If the dataset was smaller, we could consider using replace wrong values, where we would target specific columns. This is meaningful with obvious entries. For larger datasets, we can use a replace if condition is meet, looping through the rows. An example we want to delete rows with video\_id with value \n (new line in programming). Here we can use Pandas to remove these entries as we view them as invalid.

for x in df.index:

if df.loc[x, "video\_id"] == 'NaN':

df.drop(x, inplace =True)

print(df.to\_string())

We can also identify rows with duplicate values by combining pd.read\_csv and duplicated() can print the occurrence in boolean expression.

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

The data set contains useable data, but will need some cleaning based on the suggested approaches described previously. We will need to identify what is the used columns so we can review the quality more in depth.

For 2080 blank values in columns publish\_time, views, likes, dislikes, comment\_count, thumbnail\_link, comment\_disabled, ratings\_disabled and video\_error\_or\_removed, I believe that these rows can be dropped. On the other hand, if we have great category data and this is of importance, then we can consider using Mean as a value if this otherwise will cause issues with calculations.

Country and Category are more often than not blank and would be candidates for review based on the value they provide for further analysis and conclusions.