**Data Wrangling Report**

Introduction: The objective of this report is to do data wrangling of the youtube dataset and see from Kaggle and find out the cleaning that needs to be done on it.

Data download: The Kaggle dataset can either be downloaded manually and read using pandas read\_csv or directly downloaded using Kaggle api.

Cleaning of data: After looking at some of the statistics which pandas give by default, it looks like some columns don’t have the proper data type. These column types are transformed.

The column description have some null values and since there are very less number of rows with this issue, I decided to drop these rows to make sure the data has no null values.

Pandas profiling library gives many important statistics and as per it, we see that there are some duplicate rows and the data is very skewed.

I decided to eliminate the duplicate rows. It might be useful to do analysis based on day of the week the video was published so a new column publish\_day is added. Some statistics are displayed using this new column.

The pandas profiling also showed that there is a lot of skewed values. Lets take one column views and do some analysis based on it. There are more than 3700 outliers (views more than 1.5 times the inner quartile from the 75th percentile) based on views. These outliers don’t look like they are bad data. Some more analysis on these outliers shows that they mostly belong to specific category id and channels. These analysis might be handy during the next steps in the project. We might need to deal with this outliers accordingly based on what kind of analysis we are performing.

There are also many texts which are not in English and might need to be dropped if we are doing any analysis based on these columns.

Conclusion: Some more changes might be needed on the dataset but for now we can proceed with this cleaned data. The jupyter notebook file shows all steps taken for data wrangling. We will also load and join data from other countries in future to perform more broader analysis at which point of time we will have to review the column differences and handle them.