1. Abstraction - I will do
2. Introduction - 1.5 page

* What is youtube
* How it’s important in data today life
* Youtube all statistics
* Youtube impact on society
* How people make a business out of it
* How it going to be in future
* Elaborate our choice why we have selected this youtube data analysis

Start from here

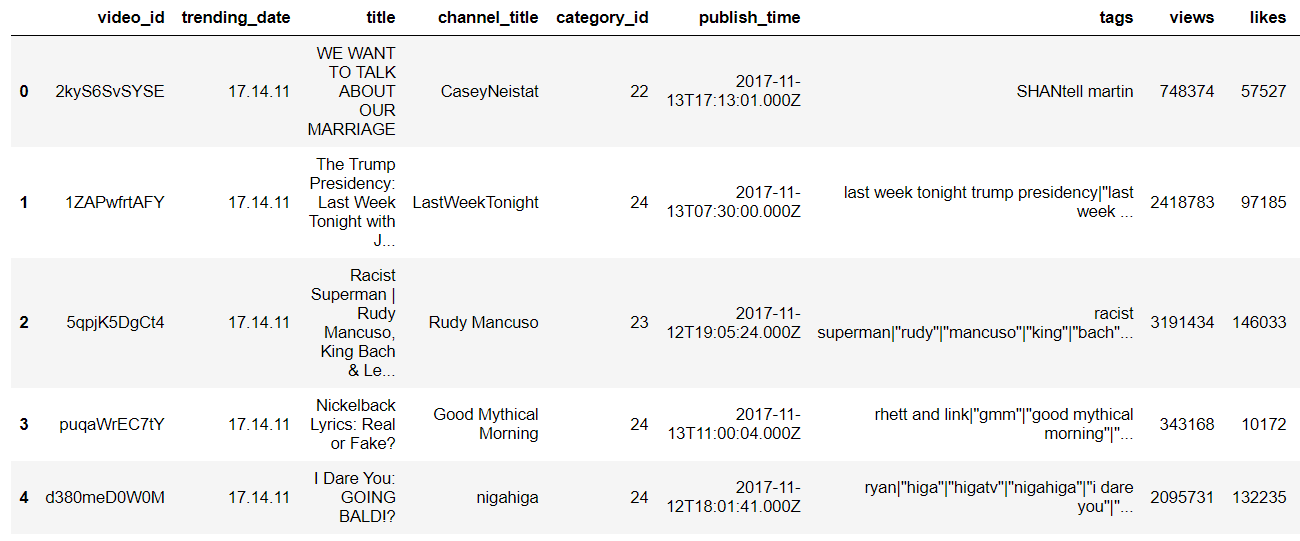
1. Motivation

* Why we are conduction this analysis
* What are the benefits we are getting out of it
* How useful this research
* What can we do from this output
* Some points from section 2(introduction)

1. Data processing analysing - I am going to do this

* About the dataset we have - overview of the data set used
* How we cleaning the data
* How handling missing values
* What are columns
* From - To date dataset
* Pic of the data set from jupyter notebook

We selected a well-structured dataset for this research from keggle, Our dataset has been titled as Trending Youtube Videos there are hundreds of videos which was trending during that particular day and also they have separated these datasets with different regions for analysis purpose we have selected US youtube trending videos dataset so that we can be focused on a single region. When we are looking at the dataset we only have data from 2017 November to 2018 June unfortunately. Apart from this, the dataset contains 40950 rows and 16 columns. Columns names are video\_id, trendin\_date, title, channel\_title, category\_id, publish\_time, tags, views, likes, dislikes, comments\_count, thumbnail\_link, comments\_disabled, ratings\_disabled, video\_error\_or\_removed, descriptions. Each column contains different types of data, some of the columns’ data we are using for our analysis purpose some of them are not.



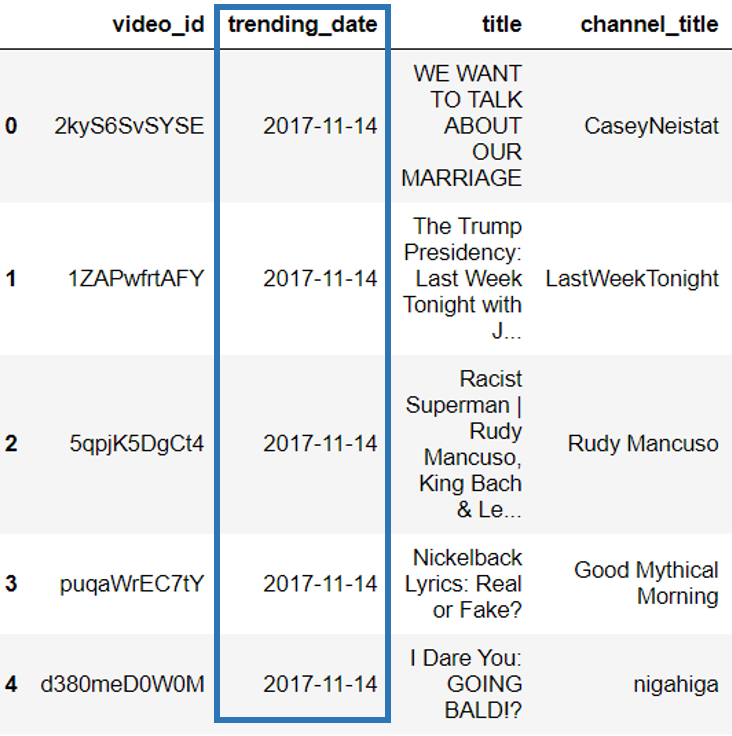


This is our dataset in fig- showing. In this dataset, we have identified some issues here we listing those issues

1. Trending\_date column contains a date which not well-formatted
2. Category\_id column contains different category id but not the category name
3. We don’t need some of these columns for our analysis
4. We need some extra columns of data for further analysis.

We cleaned our dataset in order to overcome these issues we have mentioned above

1. We have to format our trending date column as a pandas DateTime formate. For this, we have used *pandas.to\_datetime()*, here our output shows in fig-



1. Category\_id column only contains videos’ category ids only for analysis purpose we need category name and we have to map that name with this id all over the dataset. For this purpose, we have created a new CSV file which contains all the category then after we have mapped that file with our dataset.

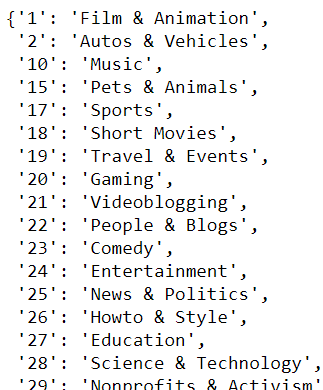
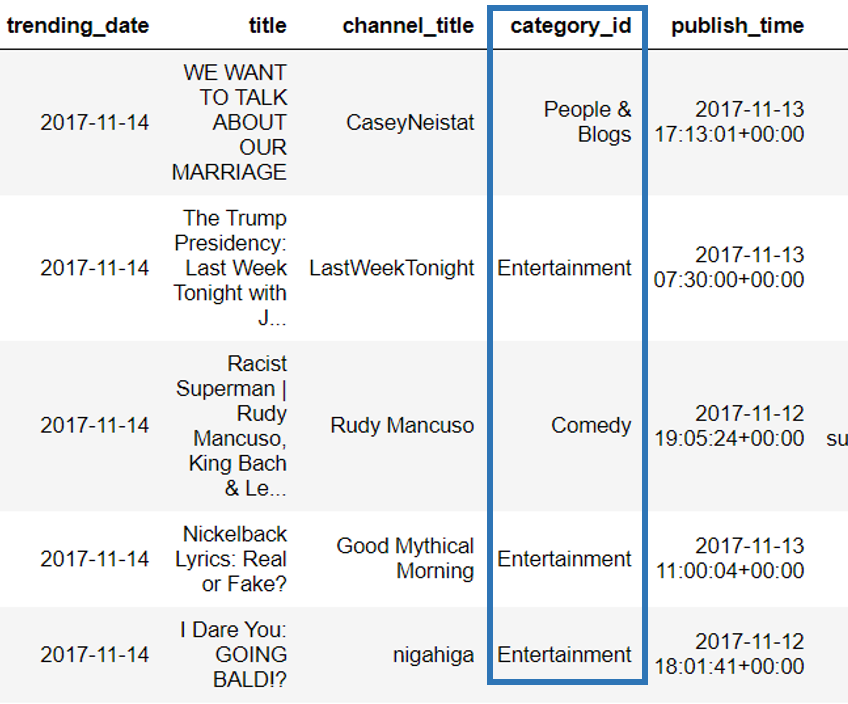
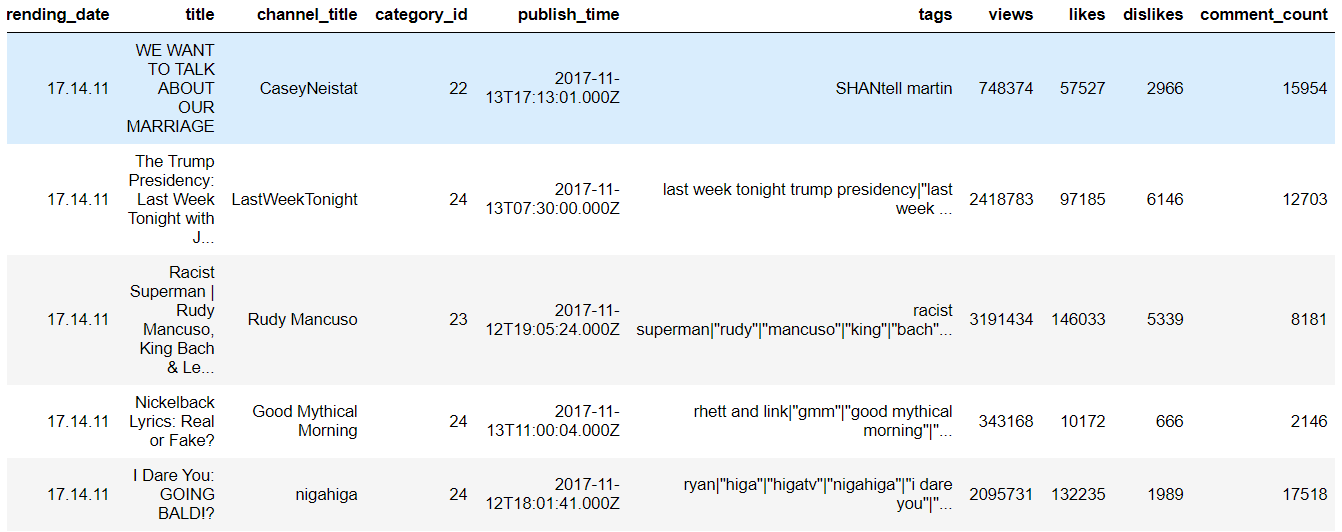


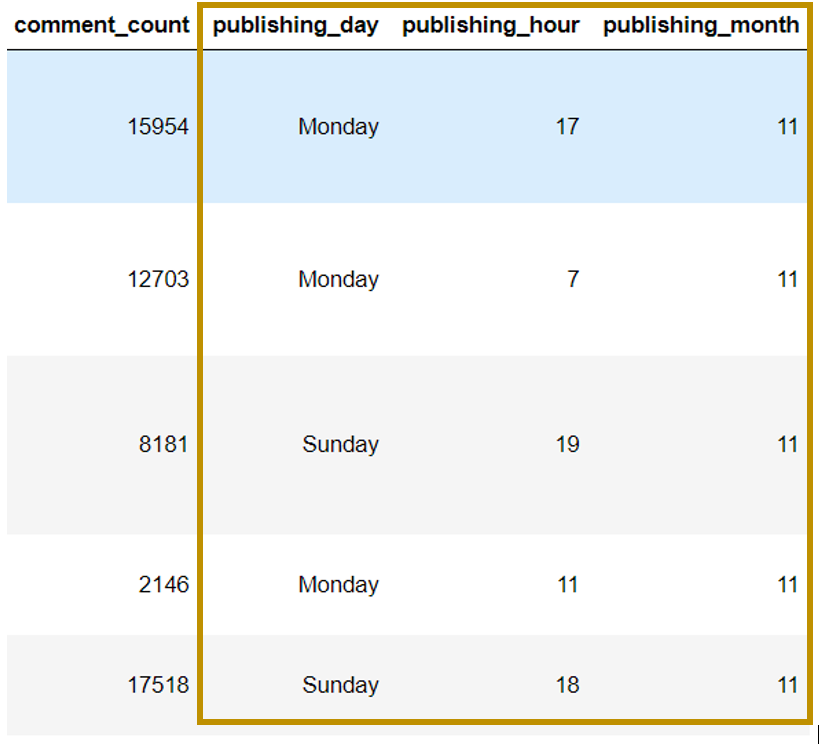
fig - shows our category id and name dataset after we mapped with our dataset it show like this in fig-



1. Finally, we have removed unwanted columns which are not needed for our analysis purpose, these the list of columns we have removed thumbnail\_link, comments\_disabled, ratings\_disabled, video\_error\_or\_removed, description. But we didn’t remove video\_id, title, channel\_title, tags columns for our future analysis purpose. After removing all these the dataset showing like this in fig-



1. Apart from these columns which already came up with dataset we have created some new columns with other existing data for our further analysis purpose, these columns have been created whenever they needed for the particular analysis purpose. These are columns we have created publishing\_day, publishing\_hour and pulishing\_month. Here in this fig- shows those created columns.



1. Methodology - I am doing

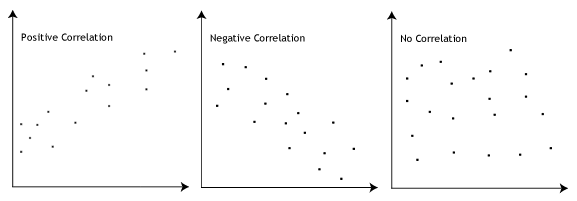
* correlations / let’s take any 2 variables
* Descriptive analysis
* Predictive analysis

In this paper, we are conducting a study on trending youtube videos’ properties in order to find some meaningful insights.

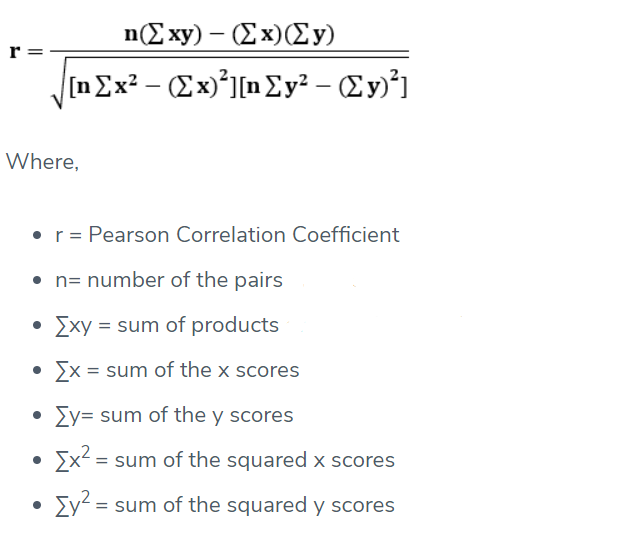
We have used the Pearson Correlation Coefficient for finding the correlations between videos’ views, like, dislike and comments\_cout.

What is the Pearson Correlation Coefficient

Pearson Correlation coefficient used to find the linear correlations or association between two variables, it can take a range of values from +1 to -1 if it’s 0 then there is no relationship between them. A value greater than 0 indicates a positive association and a value less than 0 indicates a negative association. It indicates the strength of the relationship.



Here we are finding Pearson correlations coefficient between these views, like, dislike and comments\_cout property of training videos so that we can identify an impact on one variable in another using this below formula.



Apart from this correlations coefficient analysis, we can have some different analysis as well for finding more insights so this purpose.

* trending videos’ category with its views and likes
* Which day of the week most of the trending videos had been uploaded
* Which time in a day most of the trending videos had been uploaded

1. Results and Conclusion
2. Acknowledgement
3. Future work
4. References