**Final Project Report – Youtube Trending Videos Analysis**

**ID2221 HT20-1 Data-Intensive Computing**

*Mattia Zoffoli (*[*zoffoli@kth.se*](mailto:zoffoli@kth.se)*) & Varrun Varatharajan* *(*[*varrun@kth.se*](mailto:varrun@kth.se)*)*

Table of Contents

[1. Introduction 2](#_Toc54447573)

[2. Dataset 2](#_Toc54447574)

[3. Method 2](#_Toc54447575)

[3.1. Analysis and Visualization of Data 2](#_Toc54447576)

[3.2. Analysis with MLlib 2](#_Toc54447577)

[3.2.1. Regression 2](#_Toc54447578)

[3.2.2. Clustering 2](#_Toc54447579)

[4. Results 2](#_Toc54447580)

[4.1. Analysis and Visualization 2](#_Toc54447581)

[4.2. Analysis with Mllib 3](#_Toc54447582)

[4.2.1. Linear Regression 3](#_Toc54447583)

[4.2.2. Generalized Linear Regression 3](#_Toc54447584)

[4.2.3. K-Means Clustering 3](#_Toc54447585)

[5. Code Instructions 3](#_Toc54447586)

[Resources 3](#_Toc54447587)

# Introduction

This project presents a pipeline to analyse Youtube tranding videos data for some of the most important countries in the world, trying to extract some insights from the data.

# Dataset

Kaggle dataset of trending Youtube statistics (updated daily) [1] will be used in our project. The data is made of 10 files, each for a different country, with a set of 16 columns for each trending video, reporting most importantly the views, likes, category, publication date and channel. This dataset is perfectly suited for studying relationship analyses and trends and comparing different countries.

The countries available are: US, Russia, Mexico, Korea, Japan, India, France, Germany and Canada.

# Method

## Analysis and Visualization of Data

The tools used for the analysis and visualization of data are: Apache Spark (DataFrames), Scala, Jupyter Notebook and Microsoft Office Excel (for the visualization).

In this initial phase, the following information has been extracted from the dataset for each country:

* Number of channels, videos etc
* Mean and variance values of likes, views, dislikes and comments number
* Top categories (extracting category\_id from json file)

The data obtained has been saved to file, organized in one row for each country, and then visualized, using excel to plot the data for different countries.

## Analysis with MLlib

Two different Machine Learning techniques have been used to gain some additional insights into the data: Regression (Supervised Learning) and Clustering (Unsupervised Learning).

### Regression

Initially a linear Regression algorithm [2] has been used on the most significant numerical columns of the data available: “category\_id”, “views”, “comment\_count”, “likes” and “dislikes”.

The scope of our regression has been to try to predict one of these columns at a time, using the data contained in other columns, and repeating the same process for each column.

Additionally, also Generalized Linear Regression [3] has been tested, comparing its results to simple Linear Regression.

### Clustering

In this unsupervised learning method, K-Means clustering was performed to determine the optimum number of clusters (k) required for the clustering of variables of this dataset. [5] The elbow method was used, where it runs k-means clustering on the dataset for a range of k values (from 1 to 50) and computes average score of all clusters for each k value. This method provides a good understanding on what the best value of k would be based on SSE (squared distance sum) of the data points and the respective cluster centers.

The optimal number of clusters required for this dataset is then picked at the elbow of the plot between k vs. distance. [6] The scope of this analysis is to predict how many clusters are needed for optimal learning of the parameters of this dataset.

# Results

## Analysis and Visualization

The first insight from the visualization of country data, in figure 1, is that where the number of views is highest (US and India), the number of channels is significantly lower, meaning that the few channels active in those countries get an extremely high number of views and revenue. This is particularly interesting, because it means that while Youtube and similar platforms are growing (increasing views), the attention is focusing on a handful of channels, which have probably high-quality content, drifting to resemble more traditional media.

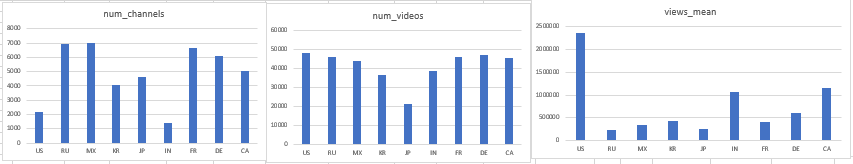


Figure 1 - Number of Channels, of videos and views per country

Addionally, there are some general patterns that are constant in

From figures 2 and 3, it is clear that the situation on Youtube in the US is extremely different from other countries

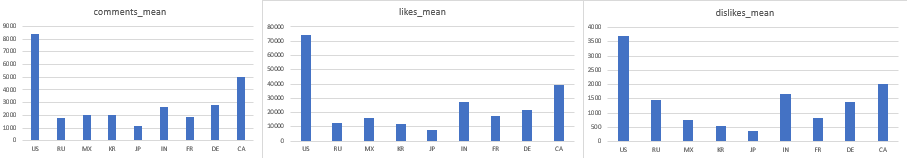


Figure 2 - Mean number of Comments, Likes and Dislikes per video, by Country



Figure 3 - Top ten categories by Country

## Analysis with Mllib

### Linear Regression

The results of Linear Regression with MLlib are shown in figure 4.  
e results of Linear Regression on the numerical values of this dataset, are not significant, because the model tends to predict a value that is close to the mean of the predicted variable, which is confirmed by the low standard deviation of predicted values.  
This means that the data is too complex for such a simple model to learn the real patterns that characterize this dataset.

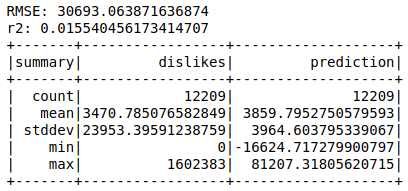
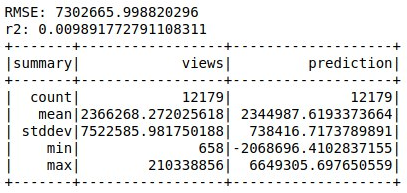
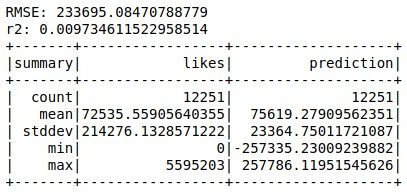
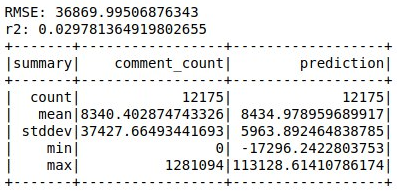
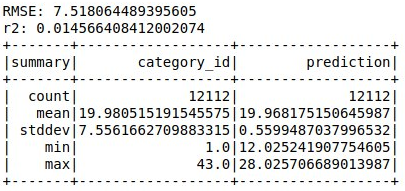
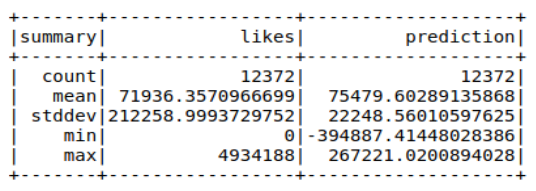
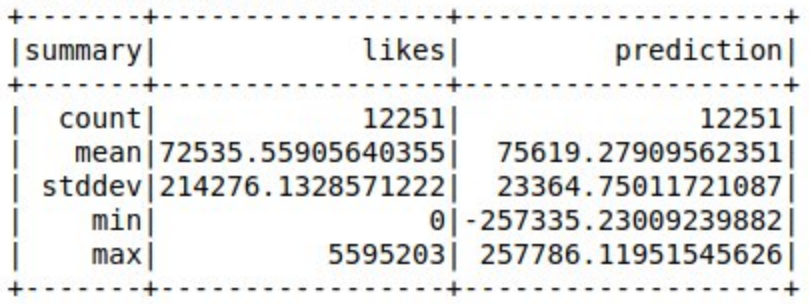


Figure 4 - Linear Regression Results separated by predicted variable

### Generalized Linear Regression





### K-Means Clustering

The results of K-Means clustering are shown in figure 5. The best value of k to cluster the dataset is determined by the elbow. The strong inflection in this plot is at k= 21. This re-affirms the conclusion that the data is a complex one, and simple models such as linear regression are insufficient for learning.

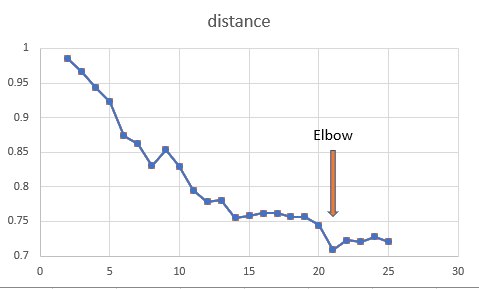


Fig. 5. Elbow method for k-means

# Code Instructions

Here are the steps that need to be followed in order to run the code and reproduce the results:

* Download Youtube data from Kaggle [1] and unzip it inside a folder named /data in the project directory.
* Create an empty /results folder in the project directory.
* Using Jupyter Notebook, run the files with country names in order, starting from 00\_US.ipynb and ending with 09\_CA.ipynb. These files will extract basic information from data, and save it in partial results inside /results folder.
* The file generated into the folder /results/partial\_09 contain data extracted for all countries, and can be visualized on excel as shown in the file Visualization.xlsx.
* To perform Regression, use Jupyter Notebook to run the files LinearRegression.ipynb and GeneralizedLinearRegression.ipynb.
* To perform Clustering with K-means, use Jupyter Notebook to run the file KMeansCLustering.ipynb.

# Resources

[1] <https://www.kaggle.com/datasnaek/youtube-new>

[2] <https://spark.apache.org/docs/latest/ml-classification-regression.html#linear-regression>

[3] <https://spark.apache.org/docs/latest/ml-classification-regression.html#generalized-linear-regression>

[4] <https://spark.apache.org/docs/latest/ml-clustering.html#k-means>

[5] <https://spark.apache.org/docs/2.2.0/ml-clustering.html#k-means>

[6] <https://www.scikit-yb.org/en/latest/api/cluster/elbow.html#:~:text=The%20elbow%20method%20runs%20k,point%20to%20its%20assigned%20center>.