

# **YOUTUBE TRENDING VIDEO ANALYSIS**

**A PROJECt REPORT**

***Submitted by***

**Nandhakumar T(17BCS210)**

**Ragu PY(17BCS212)**

**Karthi.V(17BCS208)**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**KUMARAGURU COLLEGE OF TECHNOLOGY**

**COIMBATORE 641049**

**(An Autonomous Institution Affiliated to Anna University, Chennai)**

**October 2019**

**KUMARAGURU COLLEGE OF TECHNOLOGY**

**COIMBATORE 641049**

**(An Autonomous Institution Affiliated to Anna University, Chennai)**

**BONAFIDE CERTIFICATE**

Certified that this Mini project report “**Youtube trending video analysis”** is the bonafide work of **“Nandhakumar T(17BCS210), Ragu PY(17BCS212) and Karthi.V(17BCS208)”** who carried out the Project work under my supervision.

**SIGNATURE** **SIGNATURE**

Dr.J. CYNTHIA, Ph.D., Ms. V.P. SUMATHI

**HEAD OF THE DEPARTMENT** **SUPERVISOR**

Professor and Head Assistant Professor

Department of Computer Science Department of Computer Science

and Engineering, and Engineering,

Kumaraguru College of Technology, Kumaraguru College of Technology,

Coimbatore- 641049. Coimbatore- 641049.

**ABSTRACT**

YouTube (the world-famous video sharing website) maintains a list of the top trending videos on the platform. According to Variety magazine, “To determine the year’s top-trending videos, YouTube uses a combination of factors including measuring users interactions (number of views, shares, comments and likes). Note that they’re not the most-viewed videos overall for the calendar year”. Top performers on the YouTube trending list are music videos (such as the famously virile “Gangam Style”), celebrity and/or reality TV performances, and the random dude-with-a-camera viral videos that YouTube is well-known for. This dataset is a daily record of the top trending YouTube videos. This dataset includes several months (and counting) of data on daily trending YouTube videos. Data is included for the US, GB, DE, CA, and FR regions (USA, Great Britain, Germany, Canada, and France, respectively), with up to 200 listed trending videos per day.

Each region’s data is in a separate file. Data includes the video title, channel title, publish time, tags, views, likes and dislikes, description, and comment count. The data also includes a category\_id field, which varies between regions. To retrieve the categories for a specific video, find it in the associated JSON. One such file is included for each of the five regions in the dataset.

**TABLE OF CONTENTS**

**CHAPTER NO. TITLE PAGE NO.**

**Abstract** ……………………………………………………………………… **3**

1 **Introduction** ………………………………………. 5

1.1 Domain overview …………………………... 5

1.2 Literature Survey ………………………….... 6

1.3 Motivation ………………………….............. 7

1.4 Objective ……………………………………. 7

1.5 Methodology …………………………………7

2 2.1 Dataset Description …………………………. 8

2.2 Hardware Specification ………………………9

2.3 Software Specification ……………………….9

2.4 Benefits of python in data analytics ………….10

3 **Data analysis and visualization** ……….………..….11

3.1 Largest number of trending videos in channels 11

3.2 Largest number of trending videos in category 12

3.3 Trending videos and their publishing time 13

3.4 Ratio of Likes-Dislikes in different countries 14

3.5 Correlation of trending video in countries 15

4 4.1 Conclusion ………………………….............. 16

4.2 future work …………………………............. 16

References ………………………….......................17

**CHAPTER 1**

**INTRODUCTION**

**1.1 DOMAIN OVERVIEW**

YouTube (the world-famous video sharing website) maintains a list of the top trending videos on the platform. According to Variety magazine, “To determine the year’s top-trending videos, YouTube uses a combination of factors including measuring users interactions (number of views, shares, comments and likes). Note that they’re not the most-viewed videos overall for the calendar year”.

Top performers on the YouTube trending list are music videos (such as the famously virile “Gangam Style”), celebrity and/or reality TV performances, and the random dude-with-a-camera viral videos that YouTube is well-known for. This dataset is a daily record of the top trending YouTube videos.

This dataset includes several months (and counting) of data on daily trending YouTube videos. Data is included for the US, GB, DE, CA, and FR regions (USA, Great Britain, Germany, Canada, and France, respectively), with up to 200 listed trending videos per day.

Each region’s data is in a separate file. Data includes the video title, channel title, publish time, tags, views, likes and dislikes, description, and comment count. The data also includes a category\_id field, which varies between regions.

To retrieve the categories for a specific video, find it in the associated JSON. One such file is included for each of the five regions in the dataset.

**1.2 LITERATURE SURVEY**

1. In this paper”trending videos Measurement and Analysis” we present our findings for measuring, analyzing, and comparing key aspects of YouTube trending videos. Our study is based on collecting and monitoring high-resolution time-series of the viewership and related statistics of more than 8,000 YouTube videos over an aggregate period of nine month . To retrive related data to the user who uploaded each video. For each uploader, we retrieved some feeds, such as gender, the view count of user’s channel, the subscriber count of user’s channel, and the total upload views. Further details about our datasets are provided in the different analysis sections.
2. In his paper “Deep analysis of trending youtube video statistics”includes several months (and counting) of data on daily trending YouTube videos. Data is included for the US, GB, DE, CA, and FR regions (USA, Great Britain, Germany, Canada, and France, respectively), with up to 200 listed trending videos per day. This ML algorithms like RNNs to generate their own YouTube comments. Sentiment analysis in a variety of forms and Categorising YouTube videos based on their comments and statistics.
3. In another paper,”MapReduce Use Case YouTube Data Analysis”   
   This YouTube data is publicly available and the YouTube data set is described below under the heading Data Set Description. Using that dataset we will perform some Analysis and will draw out some insights like what are the top 10 rated videos on YouTube, who uploaded the most number of videos. The training accuracy was 100% while the cross-validation accuracy was only around 69%. Distribution of the final aggregated viewership of all 3922 videos in our dataset
4. In this paper ”Classification of youtube data based on sentimental analysis” deals with analysis of YouTube Data. The analysis is done using users Sentiments features such as Views, Comments, Likes, and Dislikes. We used the Linear Regression classification approach to classify the YouTube Data. The experimental results are given accurate results which illustrated that it is influential practice and a key enabler for the social business. The insights gained from the user generated online contents and collaboration with customers is critical for success in the age of social media.

**1.3 MOTIVATION**

This dataset includes several months (and counting) of data on daily trending YouTube videos. Data is included for the US, GB, DE, CA, and FR regions (USA, Great Britain, Germany, Canada, and France, respectively), with up to 200 listed trending videos per day.

Each region’s data is in a separate file. Data includes the video title, channel title, publish time, tags, views, likes and dislikes, description, and comment count.

**1.4 OBJECTIVE**

The aim of this report is:

* To find out which channel and categories has the largest number of liked videos .

**1.5 METHODOLOGY**

The dataset undergoes normalisation, missing value treatment, choosing essential columns using filtering, deriving new columns, identifying the target variables and visualising the data as graphs.

Python is used for easy and efficient processing of data. The pandas library in Python is used to process and extract information from the given dataset. The processed data is converted into appropriate graphs for better visualisation of the results and for better understanding. For obtaining the graph Marplot, SciPy library is used.

**CHAPTER 2**

**2.1 DATASET DESCRIPTION**

This dataset includes several months (and counting) of data on daily trending YouTube videos. Data is included for the US, GB, DE, CA, and FR regions (USA, Great Britain, Germany, Canada, and France, respectively), with up to 200 listed trending videos per day.The following are its attributes as shown

Uploader Text   
trending\_date Date  
title Text  
channel\_title Text   
category\_id NumericalDiscrete  
publish\_time Date  
tags CategoricalOrdinal  
views NumericalDiscrete  
likes NumericalDiscrete  
dislikes NumericalDiscrete  
comment\_count NumericalDiscrete  
thumbnail\_link Text  
comments\_disabled CategoricalOrdinal  
ratings\_disabled CategoricalOrdinal  
description Text

**2.2 HARDWARE SPECIFICATION**

* Processors: Intel i5 8th Gen Processor
* Disk space: 1 GB
* Operating systems: Windows\* 8
* Python\* versions: 3.7.X
* Included development tools: conda\*, conda-env, Jupyter Notebook\* (IPython)
* Included Python packages: NumPy, SciPy, scikit-learn\*, pandas, Matplotlib, Jupyter, PIP\*, and others.

**2.3 SOFTWARE SPECIFICATION**

Since Exploratory data analysis is done on the dataset, the software used for the analysis is Jupyter notebook of Anaconda navigator and the Python language is used for data analysis.

Anaconda Navigator is a desktop graphical user interface included in Anaconda that allows you to launch applications and easily manage conda packages, environments and channels without the need to use command line commands.

#### The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modelling, data visualization, machine learning, and much more.

* Windows: Python 3.6.2, PIP and NumPy 1.13.1, scikit-learn 0.18.2

The framework used for this analysis is anaconda and the language used is python. Python has numerous packages that facilitates to do various analysis. NumPy, pandas and Matplot are the packages included to derive the graphical structure.

**2.4 BENEFITS OF PYTHON IN DATA ANALYTICS**

Python is an excellent tool for data analysis for four reasons:

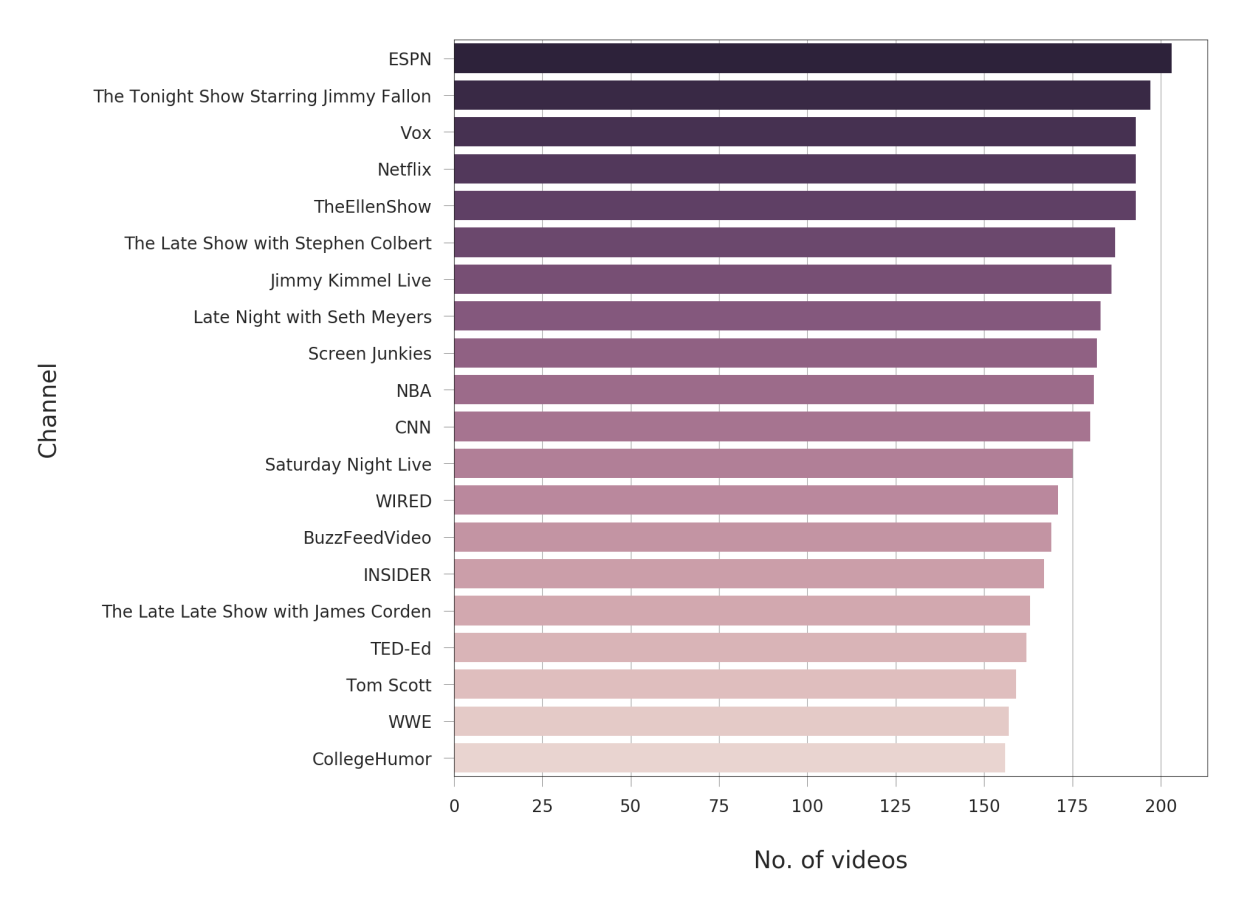
* Open source
* Speed
* Support for many devices
* Scope
* Built-in packages
* Easy visualization

**CHAPTER 3**

**DATA ANALYSIS AND VISUALIZATION**

Various cleaning process and feature selection have been done before the analysis.

**3.1 Largest number of trending videos in channels**

**** Figure 3.1 **Largest number of trending videos in channels**

In this Figure 3.1, the X axis represents No od videos . Y axis represents the channel

Inference from the Figure 3.1 is as follows:

* This graph shows that which channel has the most and least no of trending videos in that country and they are displayed in the no of videos.

**3.2 Largest number of trending videos in category**

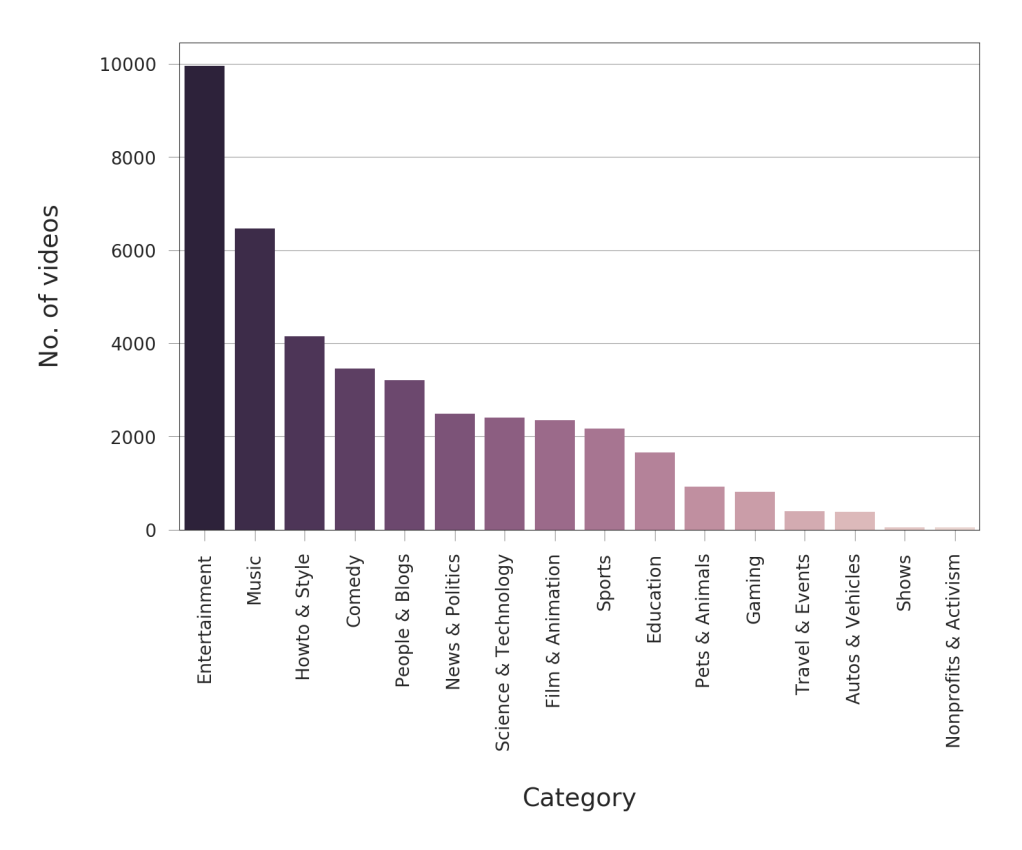
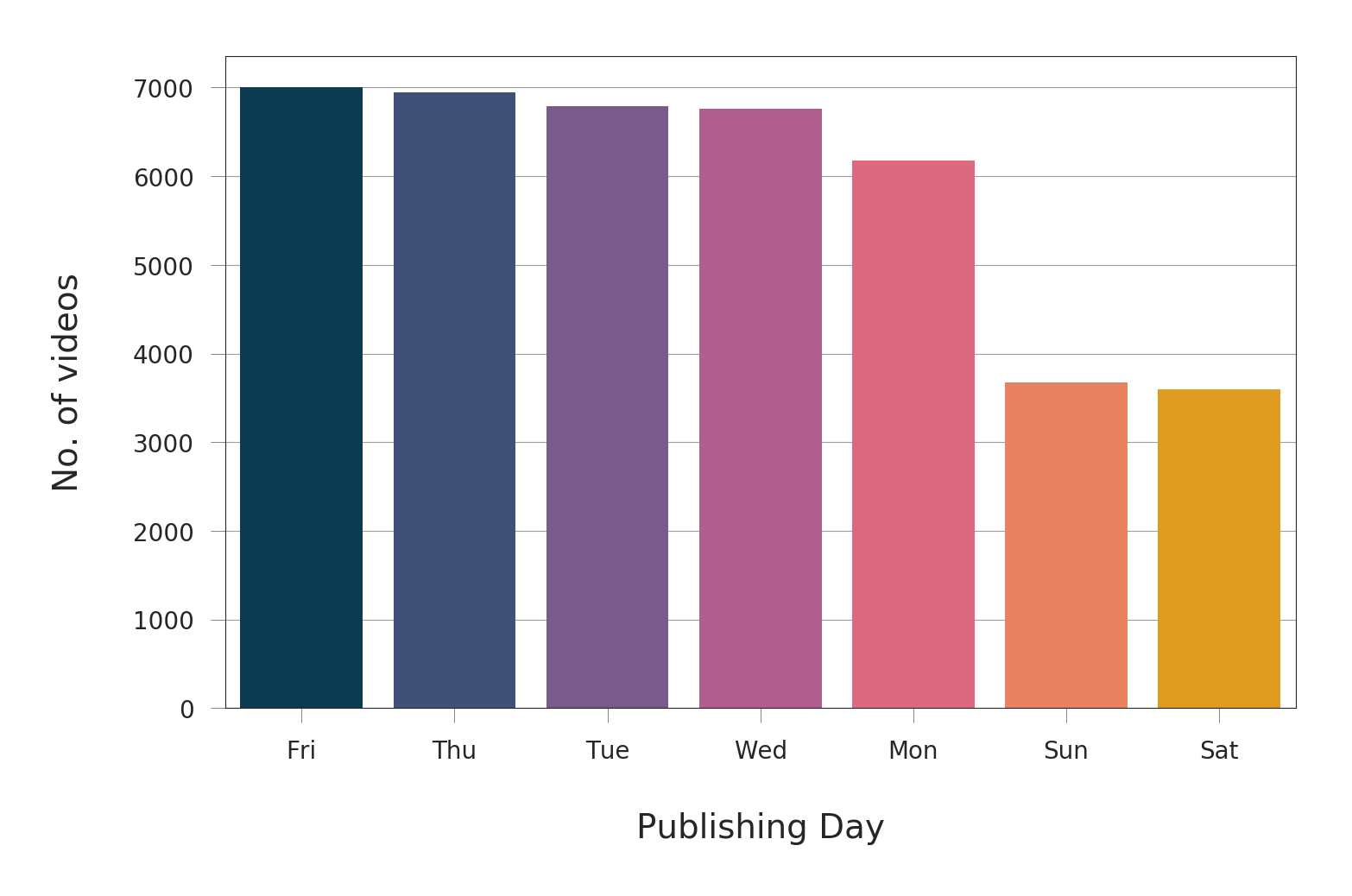


Figure 3.2 Largest number of trending videos in category

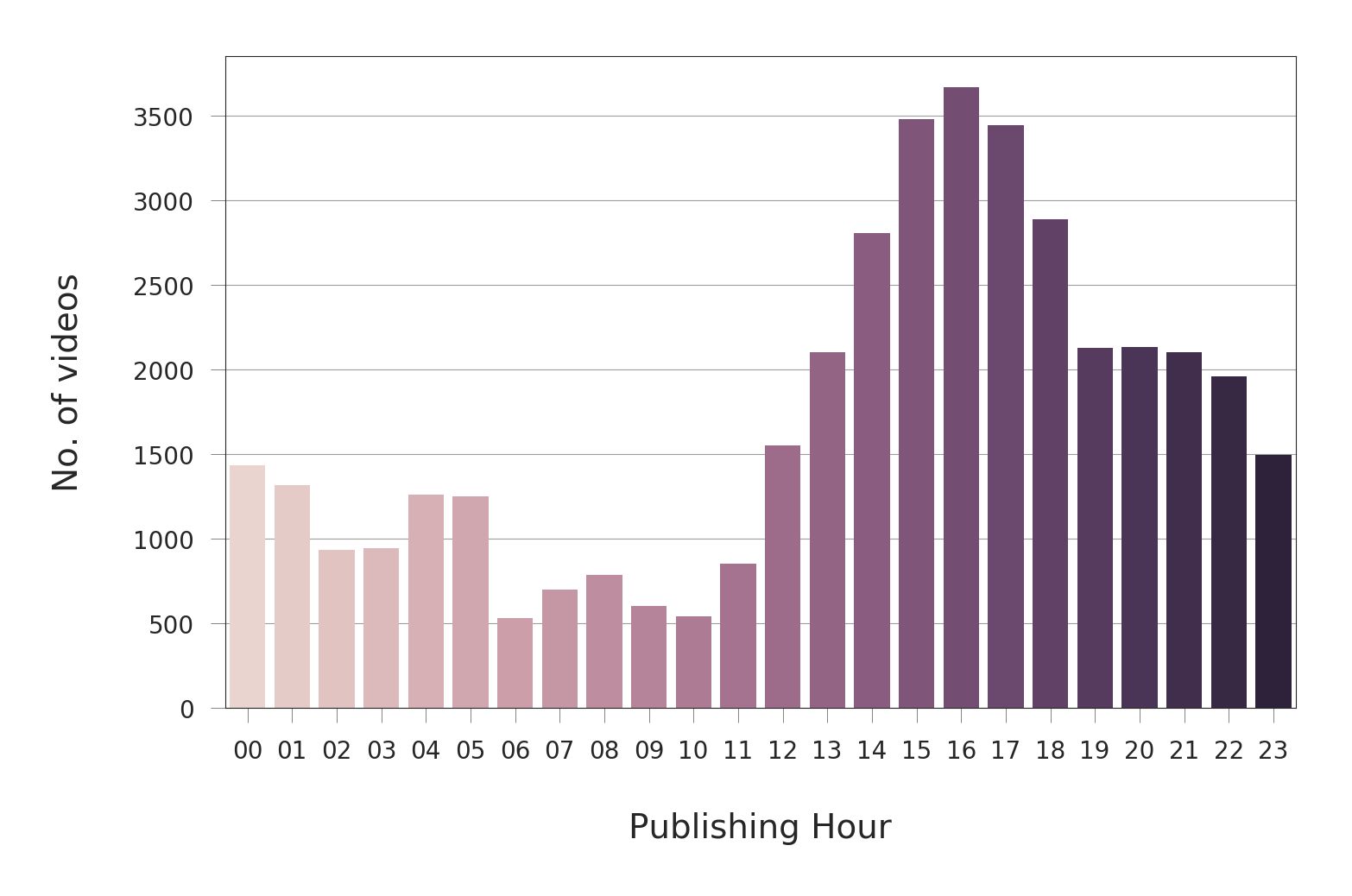
From the Figure 3.2 it is inferenced as follows:

We see that the Entertainment category contains the largest number of trending videos among other categories: around 10,000 videos, followed by Music category with around 6,200 videos, followed by Howto & Style category with around 4,100 videos, and so on.

**3.3 Trending videos and their publishing time**

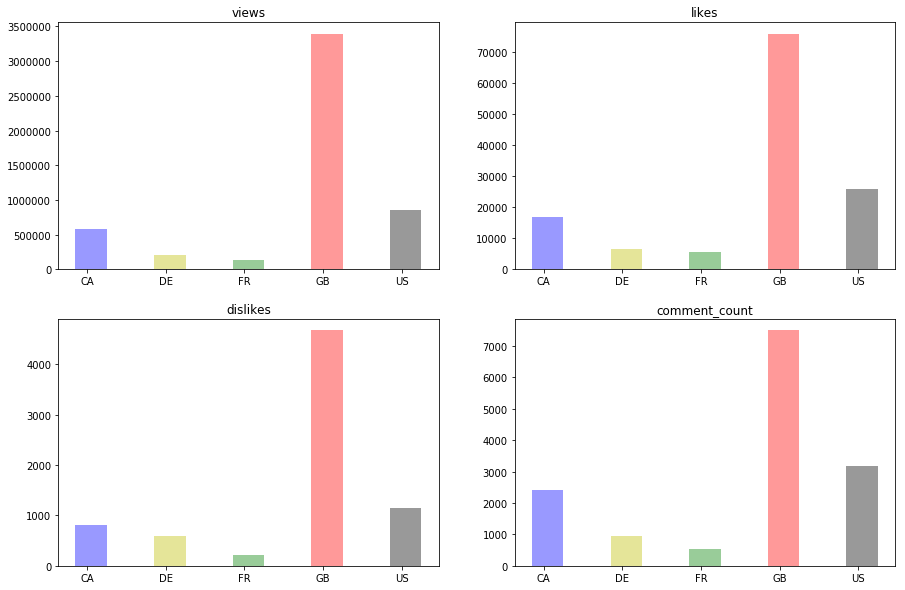
****

We can see that the number of trending videos published on Sunday and Saturday are noticeably less than the number of trending videos published on other days of the week

****

We can see that the period between 2PM and 7PM, peaking between 4PM and 5PM, had the largest number of trending videos. We notice also that the period between 12AM and 1PM has the smallest number of trending videos.

**3.4 Ratio of Likes, Dislikes, views and Comments in different countries**

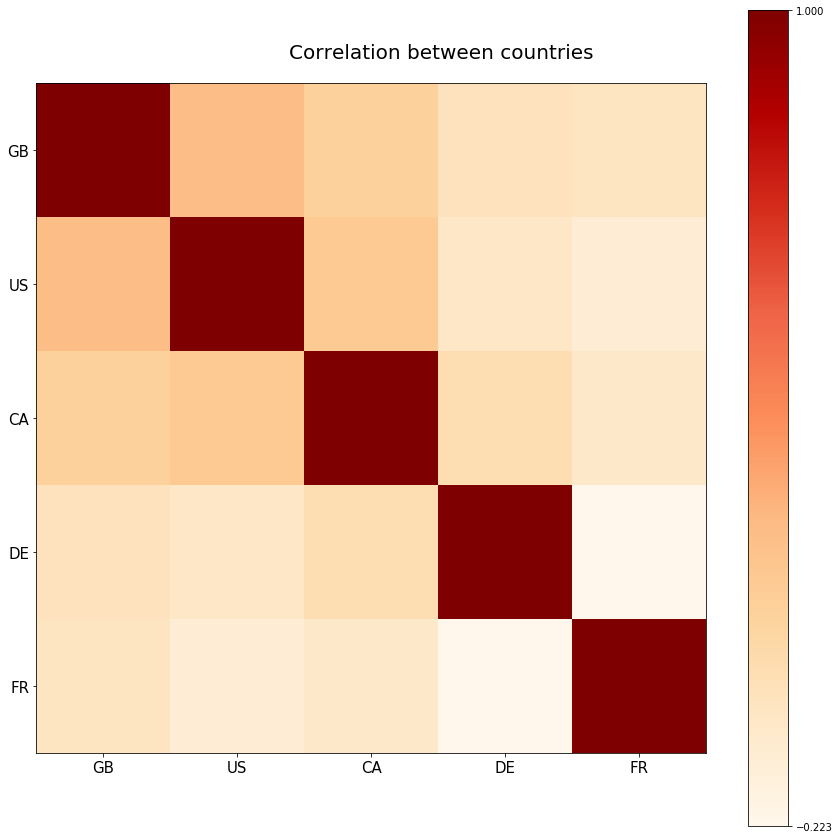


From the Figure 4.4 can conclude that:

* Obviously, four of the graphs share the similar trend in numbers. One possible reason to this is due to the video's trending duration. Enduring trending videos have the advantages in getting more views, likes, dislikes and comments)

**3.5 Correlation of trending video in between countries**

Here we got the heatmap of correlation of Youtube Trending Videos between countries. Not suprisingly, the video from United Kingdom, US and Canada is highly correlate to each other compare to Germany and France. This might due to the sharing of common language in these contries. As compare, Germany and France seems like more isolate where they did not follow the trend from english speaking countries (Just my personal opinion by observing the heatmap). This can also explain why United Kingdom has the highest number in long-trend videos, as it is contributes by multiple contries at the same time



The above heat map shows the corelation between the different countries

**CHAPTER 5**

**5.1 CONCLUSION**

The main purpose of the paper is to classify and analyse the major factors that which channel and categories has the most number of trending videos . From a proper analysis of data set fight different graphs were generated and visualized. From the graphs many conclusions have been made and information were inferred such as most number of videos liked in less publishing time are analysed in graph and most number of trending videos are analysed by their likes and views.

**5.2 FUTURE WORK**

This paper work can be extended to the higher level in future. The analysis can be referred for

* A more complex predictive model which can take newer video in cities and provide a result.
* Hadoop map reduce concept with the provided data from this paper

**REFERENCES**

1. <https://en.wikipedia.org/wiki/Exploratory_data_analysis>
2. <https://pandas.pydata.org/pandas-docs/stable/>
3. <https://acadgild.com/blog/mapreduce-use-case-youtube-data-analysis>
4. <https://www.kaggle.com/hoonkeng/deep-analysis-on-youtube-trending-videos-eda>
5. <https://www.kaggle.com/ammar111/youtube-trending-videos-analysis>
6. <https://towardsdatascience.com/youtube-data-analysis-using-pyspark-85b7cd07216f>
7. <https://www.researchgate.net/publication/266262149_Trending_Videos_Measurement_and_Analysis>
8. <https://www.reddit.com/r/statistics/comments/8li2mn/kaggle_youtube_trending_videos_detailed_analysis/>