readme

November 8, 2018

1 YOUTUBE TRENDING VIDEOS ANALYSIS

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.

The dataset is taken from Kaggle.

The given data set has data for 5 different regions. This report is a work on USA dataset which has a daily record of the top trending YouTube videos from November 2017 to June 2018.

2 Getting Started

2.1 Prerequisites

Here we will be using python3 and jupyter notebook.

we can install python using anaconda. For Jupyter please follow the link https://anaconda.org/anaconda/jupyter

After installing Python and Jupyter notebook please make sure to install the given libraries and frameworks

Pandas

```
conda install pandas
```

NumPy

Matplotlib

we get NumPy and Matplotlib when we install Pandas as dependencies

Json

```
conda install -c anaconda simplejson
```

Seaborn

```
conda install -c anaconda seaborn
```

2.2 Exploratory Finding and Explanatory Analysis

The project focuses on the analysis of US Youtube trending videos. In Exploratory analysis section, I focused on understanding each and every variable individually(Univariate) as well as understanding two or more variables and how they depend on each other(Bivariate and Multivariate).

While exploring the variables I found that the data has overplotting issues with large numbers on board(axis) and also outliers. To overcome this issue I have used log scales and log transformed data of the quantitative variables which helped me get a better insight into the data. I then started diving into how the variables of data acted on top categories and top channels which published trending videos. Here I thought that providing a story based on how the variables of data depended and acted according to top categories would be a good story to present. I prepared Explanatory analysis and slides based on what categories make most top trending videos and how are they received by the users(based on likes, dislikes, and comments).

2.3 CONCLUSIONS

Here are the key insights from the analysis and Visualizations from the project.

- Views and likes have highest corelation where as likes and dislikes have the lowest corelation between them.
- We found that 'Entertainment' category has most published trending videos with a count of 1621 followed by 'Music' with a count of 801, 'How to Style' with a count of 594 and 'shows' category has the least published trending videos.
- Based on the data of top three categories December month has most published videos with a count of 562 where as July has the least published videos with a count of 3.
- Based on the data from the top three categories Friday has the most published trending videos with a count of 593 and saturday has the least published trending videos with a count of 211
- Based on the data from the top three categories 16 Military hour has the most published trending videos and 6, 10 Military hour has the least published trending videos.
- Based on the data from the top three categories the variables likes, views, comments count show steady increase in the trend line in summer and decrease in June, shows fluctuations from July to October and tend to increase from october. But the variable dislikes shows more fluctuations.
- From the main data we can see though Entertainment category has most published videos, Music category has recorded most number of views with a count of 2.8 Billion, likes with a count of 67 Million, and comments with a count of 9.8 Million. But when it comes to dislikes, Entertainment category has most number of dislikes with a count of 6.7 Million.
- Based on the data from the top three categories the variables likes, dislikes and comments count have positive relation with variable views.

2.4 Sources

- https://www.kaggle.com/datasnaek/youtube-new
- UDACITY data visualization lessons
- https://www.geeksforgeeks.org/python-map-function/
- https://stackoverflow.com/questions/48614158/read-json-file-as-pandas-dataframe

- https://pandas.pydata.org/pandas-docs/version/0.22/api.html#datetimelike-properties
- https://stackoverflow.com/questions/29096381/num-day-to-name-day-with-pandas
- https://stackoverflow.com/questions/36410075/select-rows-from-a-dataframe-based-on-multiple-values-in-a-column-in-pandas