May 2, 2020

# Project 5 – YouTube data for US

## Udacity Business Analytics Nanodegree

*Javad Ebadi*

*This project uses Tableau to create visualizations to reveal insights from datasets. This was the final project of Udacity's Business Analytics Nanodegree to practice data visualization, including visual encodings, design principles and effective communication.*

## Contents

# YouTube US data from the US

Link for YouTube dataset: https://www.kaggle.com/datasnaek/youtube-new/data

The upload location data was added for the map visualization

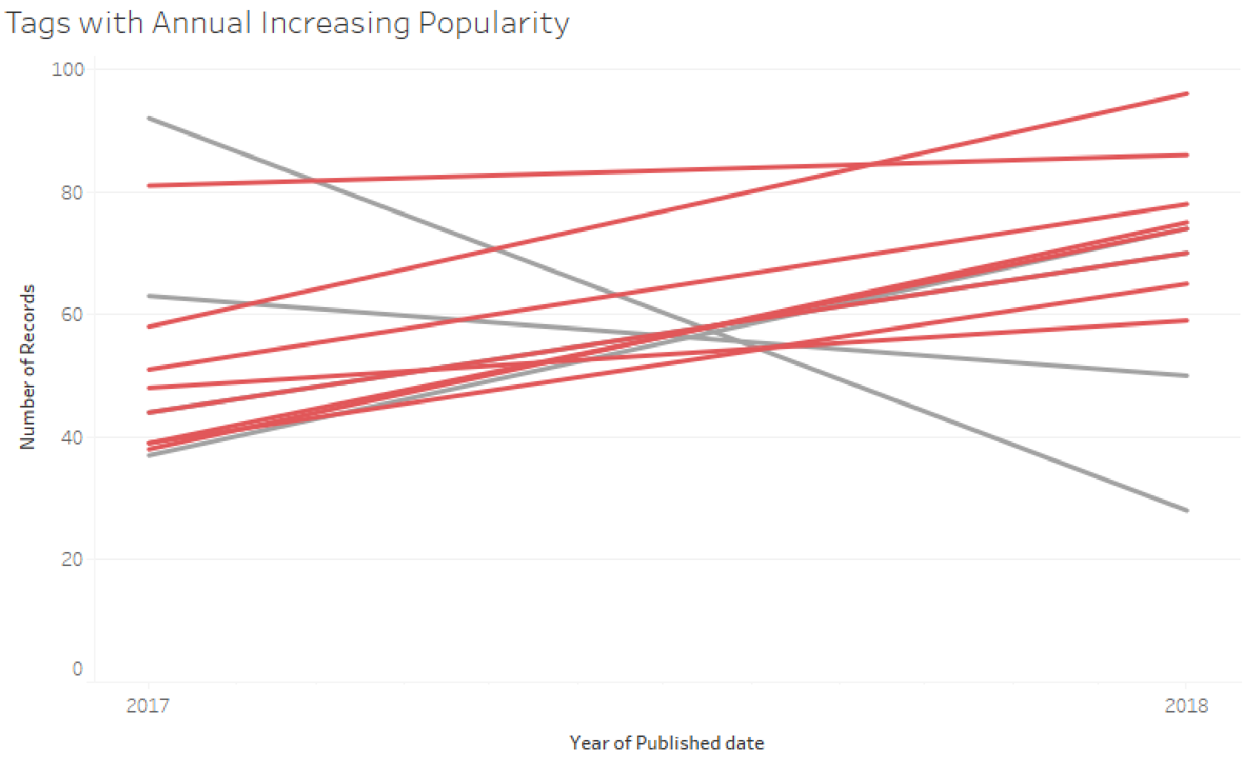
We split the tags using tableau split in Data Source page of the tableau.

# Links to dashboards or stories

## Visualization 1

Worksheet: Tags with Increasing Popularity from 2017 to 2018

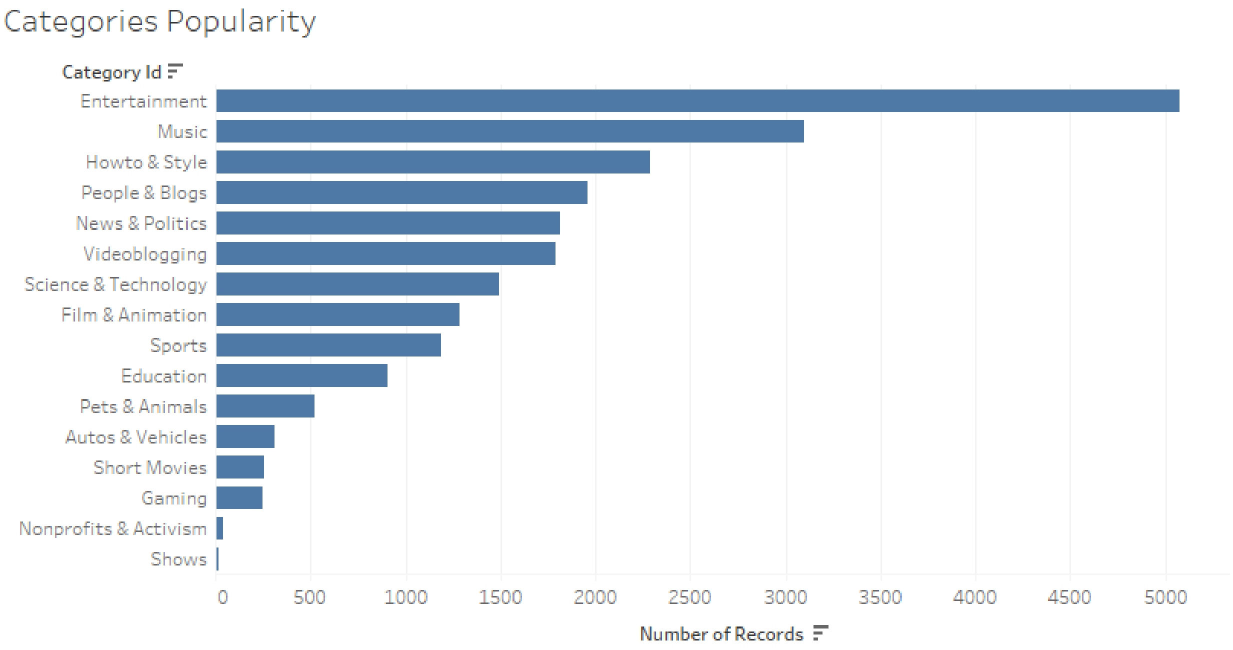
**https://public.tableau.com/profile/javadebadi#!/vizhome/TagswithIncreasingPopularityfrom2017to2018/TagswithIncreasingPopularity?publish=yes**



## Visualization 2

Worksheet: Categories Popularity

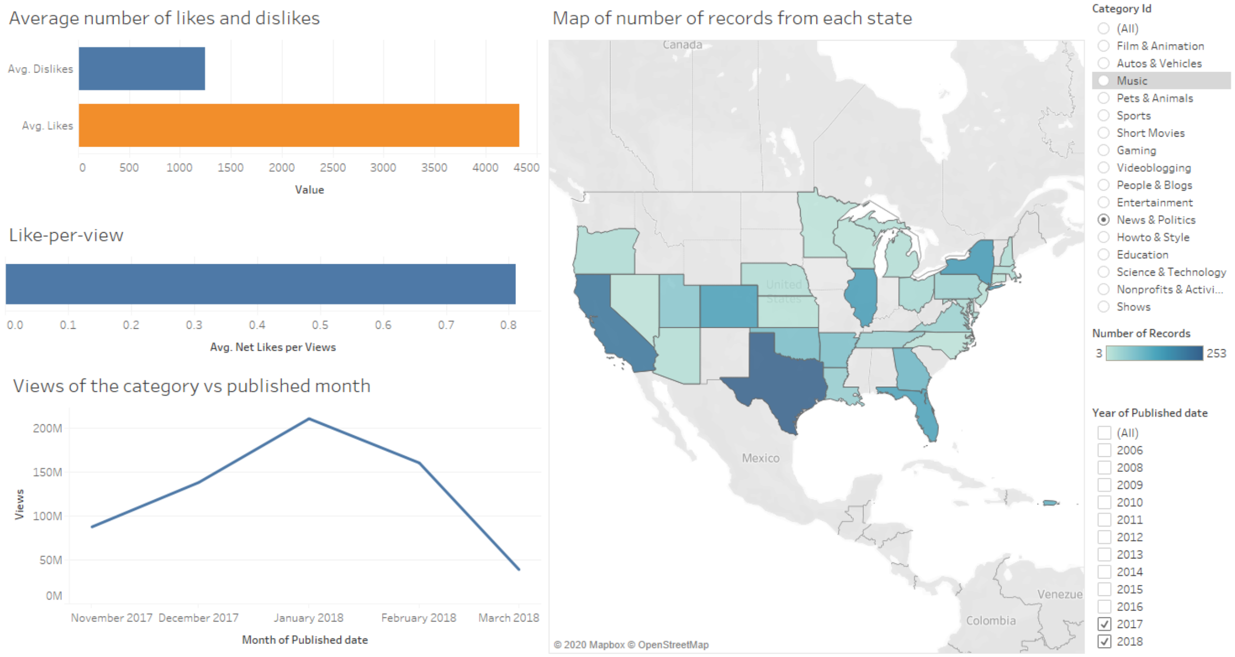
**https://public.tableau.com/profile/javadebadi#!/vizhome/CategoriesPopularity/PopularCategory?publish=yes**



# Visualization 3

Dashboard: Map, Views, and Likes of Categories

https://public.tableau.com/profile/javadebadi#!/vizhome/Mapsviewsandlikeofcategories-2/Dashboard1?publish=yes



Category Id
○ (All)
○ Film & Animation
○ Autos & Vehicles
○ Music
○ Pets & Animals
○ Sports
○ Short Movies
○ Gaming
○ Videoblogging
○ People & Blogs
○ Entertainment
● News & Politics
○ Howto & Style
○ Education
○ Science & Technology
○ Nonprofits & Activi...
○ Shows

Year of Published date
□ (All)
□ 2006
□ 2008
□ 2009
□ 2010
□ 2011
□ 2012
□ 2013
□ 2014
□ 2015
□ 2016
☑ 2017
☑ 2018

# Summary

## Main story

YouTube data set from US contains information about videos which have been trending for at least one day. The data set contains title, category, channel, number of likes, number of dislikes, number of views, geographical location, published date and trending date of videos published on YouTube.

Some tags increase and some tags decrease over time. For example, videos with “star wars” tags decreases from 2017 to 2018 while videos with “latest news” tags increases.

The Entertainment category has the highest number of records among all categories. The Music is the second.

Most of the videos are uploaded from California. However in each category, the state with highest number of uploads differs. For example, in News and Politics category most of the videos are uploaded from Texas. The average number of likes for videos of Music category published in 2017 and 2018 is highest. The number of views for Music category is the highest and then for the Entertainment. In December 2017, the number of views for both Music and Entertainment category reaches above 2Billions views and any other category is less than 1 Billion times is viewed.

## Visualization 1

In this visualization we investigate popularity increase or decrease of some tags over annual time period. The monthly behavior of each tag is stochastics and does not give us clues about answer of the question, therefore we use yearly behavior.

As we see from the visualization, tags such as “latest Bews”, “The Tonight Show”, “nba”, “Netflix”, “wwe”, “NFL” and “BuzzFeed” increase whereas tags such as “bbc” and “star wars” decrease from year 2017 to 2018.

## Visualization 2

In visualization 2, we count the number of each category. Entertainment and Music categories with 5079 and 3079 total number of records are top 2 most used categories. “Shows” and “Nonprofits & Activism” are bottom 2 less used categories with 16 and 39 number of records.

## Visualization 3

This visualization is a dashboard which contains 4 visualizations and 2 Filters. The filter shown in top right of the dashboard is a single value list of categories. The filter in bottom right of the dashboard is a multiple values list of the published date based on years from 2006 to 2018.

Let’s set category filter to “News and Politics” and years filter to “2017, 2018” and describe what each visualization in the dashboard tells.

As we see from the dashboard, the left top visualization is a bar chart which shows average likes and dislikes of the “News and Politics” category which are 4333 and 1243, respectively.

The left middle visualization shows a single bar chart which is a calculated field called Like-per-view. This field is obtained using formula `([Likes] - [Dislikes]) / [Views]`. For the “News and Politics” category the values of this field is 0.8110 which mean that from 10 views of this category about 8 views results to like.

The left bottom visualization is a line plot which is the number of views vs published data on monthly periods. The number of views of the “News and Politics” category increase from 87M on November 2017 to 210M on January 2018 and decreases to 39M on March 2018. The views of this category picks on end of the year and reaches to its minimum on the Spring season.

The right visualization is a map plot which illustrates contribution of each state of US to number of records with “News and Politics” category. As we see,  “Texas” with 253 number of records is the state which has the highest number of records in “News and Politics” category. “California” with 215 has the second rank in this category.

# Design

## Visualization 1

In visualization 1, we illustrate both increasing and decreasing tags in popularity. We use the red color for increasing tags and gray color for decreasing tags. Since, in this question, we concentrate on just one aspect of tags, which is its popularity increase or decrease, using just **two colors** was enough.

In addition, since we are investigating the popularity over time (over year), we create **line plot** with continuous data in x-axis and number of records of each tag in y-axis.

## Visualization 2

In visualization 2, we count the number of each category in the data set. We create **bar chart** because it is an appropriate visualization for counting values of categorical variable.

We use just **one color** for all bars of the bar chart because there isn’t any additional information except number of videos for each category.

## Visualization 3

In the let top visualization, we use a **bar chart** to compare to numeric values. Since bars represent to contradictory quantities we used **two non-similar color** to represent them.

In the left middle visualization, we use a **bar chart** to show a quantity which has a single values.

In the left bottom visualization, we use a **line plot** to show behavior of Views over time period.

In the right visualization, we use **map** of USA with states to show the contribution of each state to a category. We use **one color** and the **brightness of the color** in each state determines the number of records from that state. The brighter color means low number of records and darker color represents higher number of records.

# Resources

We have used the following GitHub gist to determine aliases for each category number in order to use in Tableau

https://gist.github.com/dgp/1b24bf2961521bd75d6c