Noboru Hayashi

EN 605.662 Data Visualization

Project #3 Interactive Visualization using Tableau

I. Introduction

In this paper, the process of exploring the dataset "Trending YouTube Video Statistics" and visualizations designs are demonstrated. From the dataset, the relationships between audience's comments/ratings and the view counts of the trending videos will be explored, and the over time trend of user's activities are analyzed as well.

II. Source Data

This dataset is a daily record of the top trending YouTube videos, and available at Kaggle.com[1]. It includes statistics of trending YouTube videos for the US, GB, DE, CA, and FR regions, with up to 200 listed trending videos per day. In this paper, the US data is used for exploration. There are statistics of 40.9k trending videos stored in the source file, USvideos.csv, and the schema has 16 columns (variables) as below:

Variable	Data Type	Min	Max	Mean	Std. Deviation	Description
video_id	Nominal	-	-	-	-	Unique ID of video
trending_date	Quantitative / Interval	-	-	-	-	Date of the trend video logged into the data
title	Nominal	-	-	-	-	Title of Video
channel_ title	Nominal	-	-	-	-	Title of YouTube Channel
category_ id	Ordinal / Numeric	1	43	-	-	ID of video category
publish_ time	Quantitative / Interval	-	-	-	-	Video Publish Time
tags	Nominal	-	-	-	-	Tags added to the video
views	Quantitative / Ratio	549	225211923	2360784.64	7394113.76	Number of video views
likes	Quantitative / Ratio	0	5613827	74266.70	228885.34	Number of likes

dislikes	Quantitative / Ratio	0	1674420	3711.40	29029.71	Number of dislikes
comment_count	Quantitative / Ratio	0	1361580	8446.80	37340.49	Number of comments on the video
thumbnail_link	Nominal	-	-	-	-	URL link of video thumbnail image
comments_disabled	Nominal	-	-	-	-	Whether comment feature is disabled for the video or not
ratings_disabled	Nominal	-	-	-	-	Whether rating feature is disabled for the video or not
video_error_or_removed	Nominal	-	-	-	-	Whether the video is error or removed
description	Nominal	-	-	-	-	Description text of the video

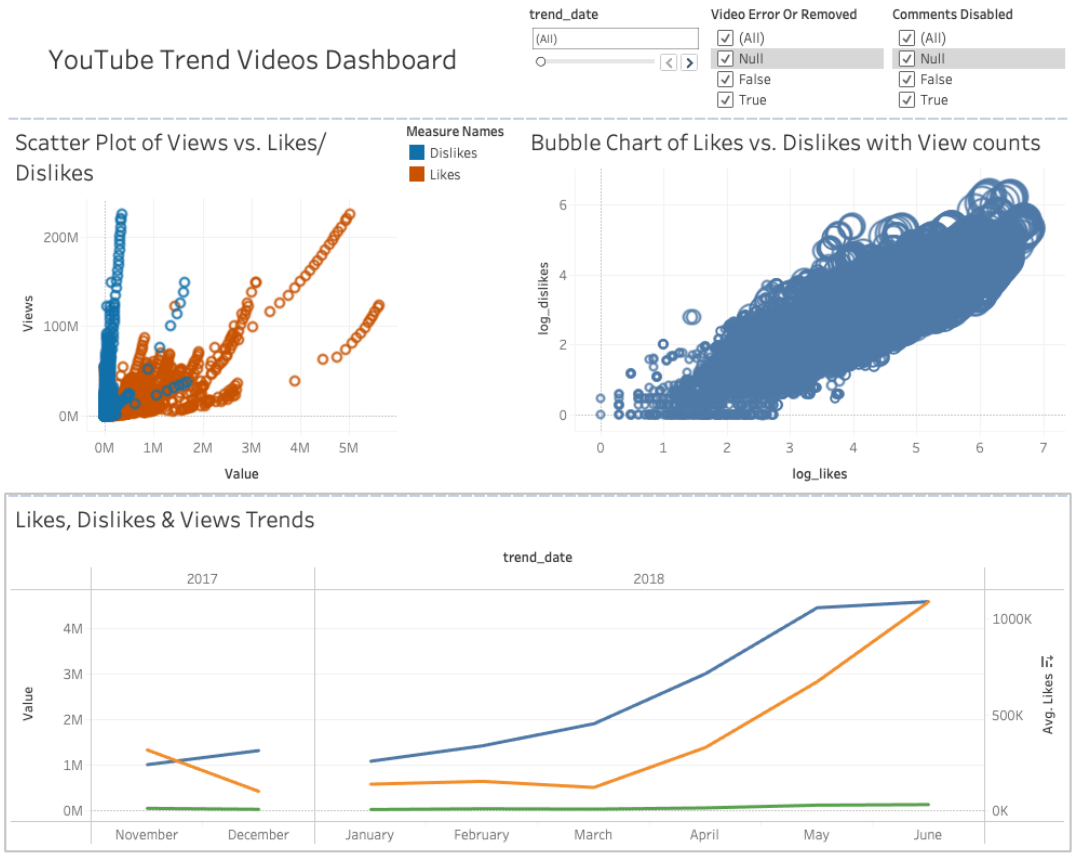
III. Analytical Questions on data

While exploring the variables of the data, I found it will be interesting to dive deep into the data and answer some analytical questions:

1. Is there any correlation between the counts of views and likes (or dislikes) ?
2. Is there any specific video category that is more likely to be trending?
3. Is there over time increase/decrease of # of users or user actions (views, likes, dislikes) for trending videos?
4. Which YouTube channel is the most trending in year 201X/202X?
5. Is there lag between published data and trending date, or most trending videos become viral right after the publish?

IV. Visualizations Design

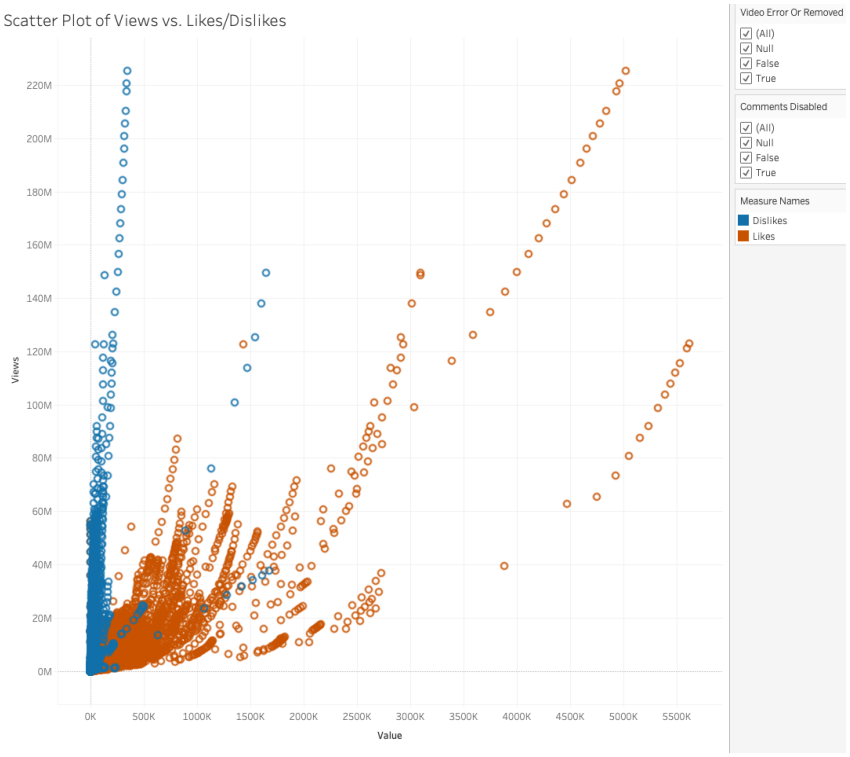
With the use of Tableau, I have developed 3 visualizations to analyze the data pattern and trend to answer some of the analytical questions above[2].



YouTube Trend Videos Dashboard

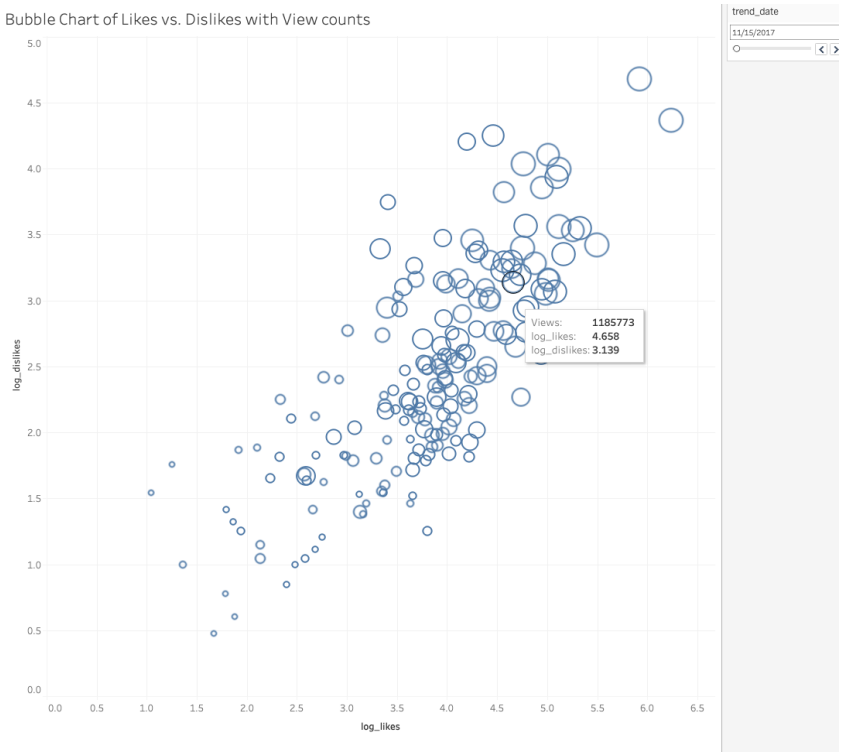
Visuals for Q1:

For the question regarding the correlation between the view counts and the user's ratings, Two types of scatter plots are adopted: One with 2 markers to visual different attributes, and the other with the bubbles indicating the third variables. Since the source data contains some null values for the fields: views, likes & dislikes, which are likely due to some data input error, filters to exclude records with null values are added. And there are also global filters for "Trend dates", "Video Errors/Removed" and "Comments Disabled" fields, so that multiple visuals are interactive with these filters.



Scatter Plot

1. In order to visualize relationships between View counts vs the counts of likes/dislikes, the scatter plot with 2 markers is used for the dashboard. From the graph, we can observe that the slopes of Views/Likes or Views/Dislikes are positive so that there are possibly positive correlations between these attributes.

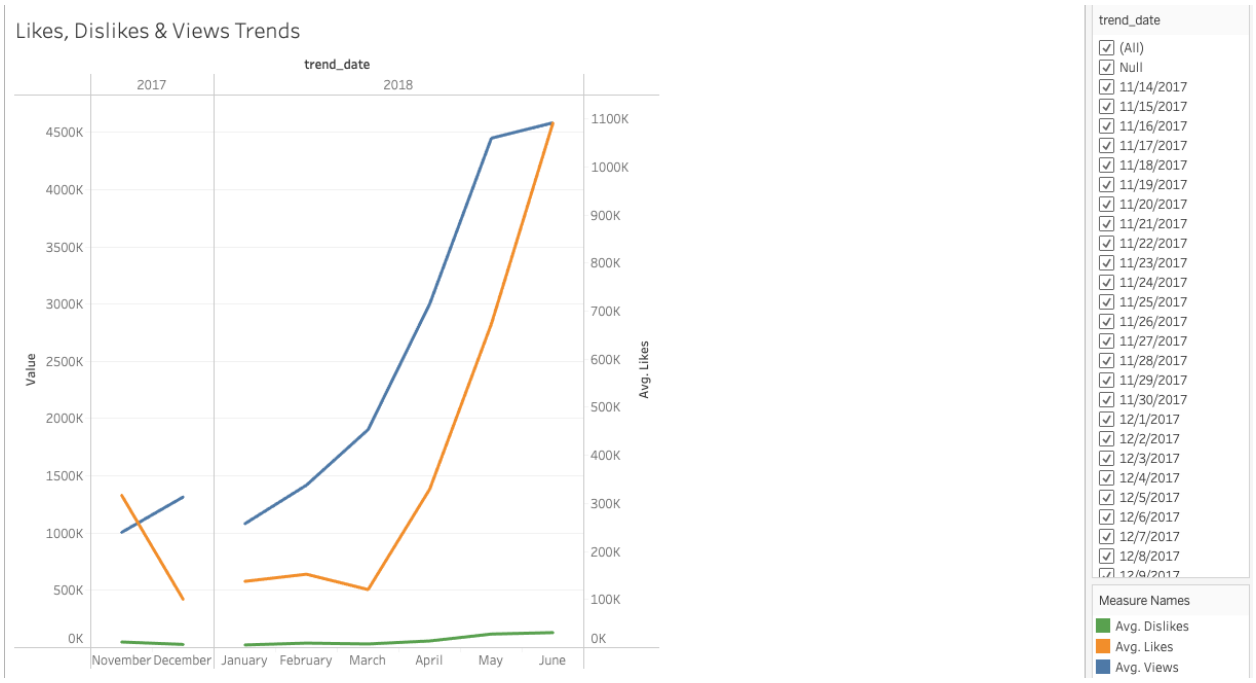


Bubble chart

2. Bubble Chart also is an effective way to show distributions of 3 variables. In the Tableau sheet, the size of the circle is proportional to the view counts of each video. By comparing the sizes of circles located at bottom left and top right, we can see as the counts of ratings increases, the video tends to have more watch counts.

Visuals for Q3:

In order to represent the trend line of User's activity (Likes, Dislikes and Views), the over time line chart is used for the Q3.



Likes, Dislikes & Views Trend line

With the multiple lines indicating Averages of likes, dislikes & views, we can see the overtime trends of those activities are growing, even though the source dataset only contains trending videos from November, 2017 to June, 2018. Besides that, we can observe that in Q1 and Q2 of 2018, the averages of action counts had doubled, which possibly indicates the users became more active or there are more new users watching and rating the videos at YouTube.

Reference:

[1] Trending YouTube Video Statistics - Kaggle.com
https://www.kaggle.com/datasnaek/youtube-new?select=USvideos.csv
[2] YouTube Trend Videos Dashboard - Tableau Public
https://public.tableau.com/profile/noboru.hayashi#!/vizhome/TrendingYouTubeVideoStatistics_16152664662650/Dashboard1?publish=yes