Noboru Hayashi

EN 605.662 Data Visualization

Project #2 Data Exploration and Design with Trending YouTube Video Statistics

I. Source Data

In this paper, the process of exploring and visually analyzing the dataset "Trending YouTube Video Statistics" will be demonstrated.

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_co unt	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

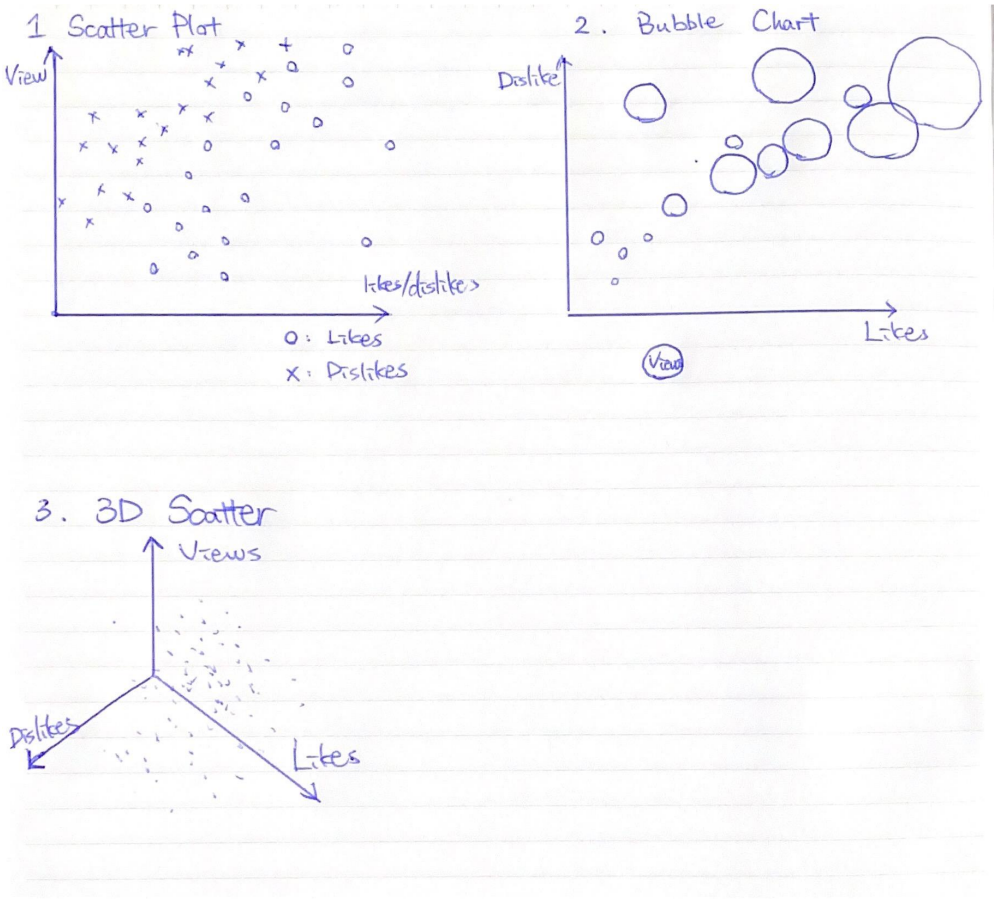
II. Analytical Questions on data

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?

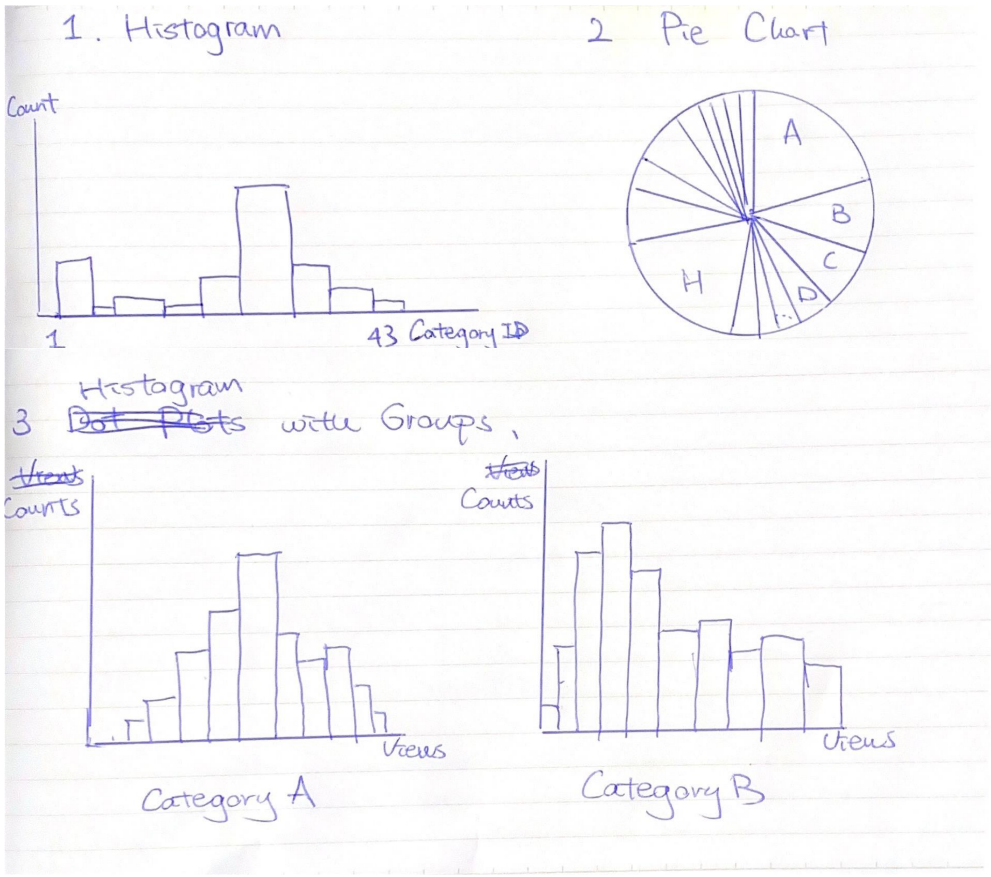
III. Visualizations Design

Visuals for Q1:



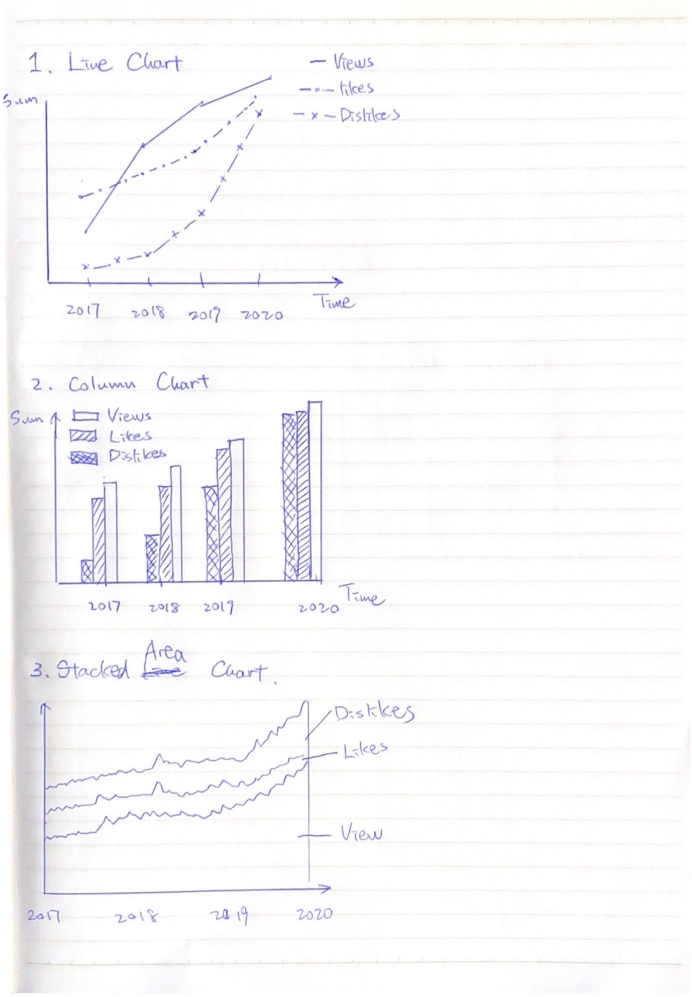
1. Scatter Plot is a simple way to visualize the distributions of the video's view counts vs # of likes(or dislikes by using multiple markers) and explore correlations between variables. With the data of YouTube stats, Views and Likes/Dislikes look positively correlated.
2. Bubble Chart also is an effective way to show distributions of 3 variables. In the sketch, the size of the bubble indicates the view counts: it seems that videos having a lot of likes and dislikes have more views, which makes sense that the more YouTube videos get exposed, the more likely audiences like/dislike the videos.
3. 3D Scatter plots distributions in a multivariate (3D) space, and it's a straightforward way to visualize 3 variables. The graph has 3 axis, and each data point represents one YouTube video with coordination of (# likes, # dislikes, # views). However 3D plots are quite hard to sketch by hand, and in order not to confuse readers, the graph has to be visually well designed.

Visuals for Q2:



1. The histogram shows the distribution of records counts vs. categories, so from the graph, we can see which category has the most videos in data by comparing the heights of the bars, and the most popular category can be inferred.
2. Pie chart shows the proportion of the frequency among categories, and the category having the largest slice is the most frequent one in data. So from the graph, the most popular category can be observed.
3. Side by side histograms can effectively visualize the difference in distributions of multiple categories. Since this type of graph contains multiple histograms, it will be very powerful and clear when the size of the category is small like Gender, Movie Rating (G, PG13, PG, R) and so on. Among YouTube statistics data, there are 43 unique category IDs, so this way of visualization is not the best for this use case.

Visuals for Q3:



1. Line chart is a basic way to visualize over time data such as the numbers of views, likes and dislikes over the trending date. In my sketch, the multiple markers for lines are used to visualize the trend of the sum of these variables.
2. Column chart is a basic chart as well and becomes useful while the few time periods of data are being used. The graph above shows the trends of sums of views, likes and dislikes with three bars for each year within 2017~2020, and we can observe the numbers of three user actions increase over time.
3. The Stacked area chart above shows the trend of user activity counts (views+likes+dislikes) over time, and since the height at each time point for each component indicates the number for that specific action, the graph is useful for comparing variables changing over an interval.

Reference:

[1] Trending YouTube Video Statistics - Kaggle.com

https://www.kaggle.com/datasnaek/youtube-new?select=USvideos.csv