Trending YouTube Video (US Data) Statistics

EDA & Data Visualization - Tableau Project

Story by Amir Awawdi /amiraw - contains 2 dashboards, 6 visuals.

Project Overview

Dataset: Trending YouTube Video Statistics

Source: https://www.kaggle.com/datasets/datasnaek/youtube-new

Background:

YouTube (the world-famous video sharing website) maintains a list of the top trending videos on the platform. To determine the year's top-trending videos, YouTube uses a combination of metrics measuring users interactions, including: number of views, shares, comments and likes.
This data comes from a Kaggle dataset, it includes a bunch of information for videos that were trending for several months (November,2017 – March,2018) mainly in USA.

Problem description:

AS a data analyst, conduct an exploratory data analysis **(EDA)** providing variety of insights to answer the following questions:

1. **What categories are the most liked and disliked ? Were there outliers channels or titles for these categories ?**
2. **Perhaps there is a time of year where one category is preferred ?**
3. **Channel Distribution: Which location exhibits the greatest channel presence ?**
4. **What tags have grown in popularity over time ?**
5. **Analyzing Seasonal Trends: Is there a peak time for a specific tag ?**
6. **Popular tags and location correlation: Are there distinct regions associated with specific tags ?**

Data Overview and Preparation:

This section elaborates on the methods used to clean and structure the data using Excel and prepare it for analysis and data visualization utilizing Tableau according to the concepts explained previously. This process typically involves the following tasks**:**

- **Cleaning** the original table in Excel included splitting the "tags" column and transposing it in a new column within a new sheet. In the original dataset, for each title (video) per trending date (with the one publish date) all the relevant tags were contained in one

string with the pipe sign " | " as a separator. The new dataset called "TagsTransposed".

- **Preparing** the "Trending Date" was required in order to match the Date structure of the " Publish Date". Using Tableau, a new column called " Updated Trending Date" had been created with the required format.

- **Joining** the original dataset with an additional table containing the names of the categories matched to the category ids was performed using a **LEFT JOIN** on the "category id" column. Regarding the newly created tag's dataset, "TagsTransposed", since it contains several rows for the each title where each row is a different tag, I chose to create a **RELATIONSHIP**, as a special feature provided by Tableau, with the original dataset sharing the "title" and "channel title" columns.

- **Creating** new objects, using Tableau, in order to perform the EDA as follows:

    - **Sets**: as an initial overview, see the first sheet within the attached tableau workbook called "Titles# & Channels# per Category per Publish - Month/Year", we observe that the distribution of titles and channels as a function of publish dates indicates clearly on a specific period of time where the majority of the data is concentrated. Therefore a set, called "PublishDates by number of titles" had been created including the period (November,2017 – March,2018) under the title " Significant number of Titles " while excluding the rest of the data considering it as negligible.
      This set was applied as a filter for all the sheets using the discussed dataset.

    - **Hierarchies**: In addition to the existing hierarchy "Location" including, in this order,: Country, State and City, that was used in the Mapping visuals, a new hierarch had been created, called "Category-Channel-Title" which was utilized to calculate the metrics: "Likes", "Dislikes", "Views" and "Comments" for each title then channel then category.

    - **Calculated Fields**: Three new fields were calculated: "Channels#" and "Titles# (Videos)", each counts distinctively the values of the columns "Channel title" and "Title" accordingly in the joined dataset. "Titles# per Tag" which counts distinctively the values of the column "title" in the "TagsTransposed" dataset.

    - **Parameters**: Two parameters had been created, the first one is "Top N Categories" which was utilized to define the number of the top categories that was inspected while filtering by the calculated field "Titles# (Videos)", this filter was applied to all the sheets using the discussed dataset.
      The second parameter is "Top N Tags" which was utilized to define the number of the most popular tags filterd by the calculated field "Titles# (Videos)", this filter was applied selected sheets related to the Tag's analysis.

Application and Analysis:

Insight #1 – What categories are the most liked and disliked ? Were there outliers channels or titles for these categories ?

- Summary – In the first dashboard "Categories & Interactions", choosing N = 6 to inspect the top 6 categories by number of titles. The bubble visual shows that the **Entertainment** category have the **highest number of titles** in each month with **Music** in the second place during November and December 2017 and third during January and February 2018 when **News & Politics** came in the second place.
  However, the table of the metrics below provide a deeper insight regarding the most liked category, without applying any filter or choosing any category in all the visuals of this dashboard, that table shows the total value for each metric where the **Music** category is clearly the **most liked** with a total of 69,357,953 likes.
  Similarly, that table shows that the most disliked category, is **Entertainment** with a total of 3,197,729 dislikes.
  In addition to the discussed table, two boxplots which contains the 4 metrics provide similar insights while drilling down on channels and titles to check for outliers.
  The upper boxplot shows clearly that in terms of "Likes" the **Music** category have a significant number of outliers, however, the in terms of "Dislikes" the **Entertainment** category had the highest outlier, the title "YouTube Rewind: The Shape of 2017 | #YouTubeRewind" from the channel " YouTube Spotlight " had a total of 1,643,059 dislikes.

- Design – The bubbles chart modified with colors according monthly number of titles per category provides a clear visual and an easy indicator to the observer's eye on which categories are the most popular in terms of number of titles monthly.
  The metrics table is a straight forward tool to measure the interactions, since it is an interactive table, it will provide fast feedback in case a drill down is required, for example you can drill down on monthly metrics for each category or filter for a specific channel or title.

- Resources – N/A

Insight #2 – Perhaps there is a time of year where one category is preferred ?

- Summary – The bubble visual provides, for each month, a clear indication that the **Entertainment** category have the **highest number of titles** and the **Music** and **News & Politics** categories switching one with other competing on the second place. **Music** in the second place during November - December 2017 and **News & Politics** came in the second during January - February 2018. However, in terms of "Likes", the **Music** category dominates with the highest monthly number of likes with a **peak**

of 16,686,095 likes during **January 2018**

- Design – See the design section of Insight #1.

- Resources - N/A

Insight #3 – Channel Distribution: Which region exhibits the greatest channel presence ?

- Summary – The distribution of the number of channels per location, while inspecting the top 7 categories , provides a list of 5 states with the highest channel presence: state of California (**CA**) comes in the first place with 153 channels, then Texas (**TX**) with 137, followed by Georgia (**GA**) with 95 then Florida (**FL**) with 92 and finally the state of Illinois (**IL**) with 81 channels.

- Design – The Bubbles inside a map provides an efficient visual where the size of each bubble relates to the number of channels located in the specific city, state.

- Resources – Google and Wikipedia to correctly locate missing location data and the following website to identify the states abbreviations https://www.faa.gov/air_traffic/publications/atpubs/cnt_html/appendix_a.html .

Insight #4 – What tags have grow in popularity over time ?

- Summary - In the dashboard "Popularity of Tags", typing N = 10 to inspect the top 10 tags by number of titles. The tag **"funny"** dominates, as seen in the treemap visual, with **279** titles, meaning it was tagged with 279 different videos which resemble nearly **47%** of the total number of titles associated with this list of top tags. Nevertheless, the line chart indicates that **"funny"** suffered from a decreasing popularity in the first three months of 2018 after a **significant 22.4%** increase in **December 2017** compared to the previous month.
  An overview of the line chart shows that all of the tags enjoyed a popularity increase in December 2017, but only the tags **"NBC"**, **"science"** and **"2018"** continued with this **positive trend** during the next month, **January 2018**, just to suffer from a decreasing popularity later on.

- Design - The treemap plot shows the most popular tags in terms of the highest number of titles associated with each, furthermore, the size of each area indicates on the relevant tag's percentage from total number of titles.
  The line chart shows the monthly popularity grow for each tag, in the tooltip one may find the monthly number of titles where the specific tag was mentioned as well as the percentage of difference compared to the previous month, which can will

show up along the line once a specific tag is chosen.

- Resources - N/A

Insight #5 – Analyzing Seasonal Trends: Is there a peak time for a specific tag ?

- Summary – The peaks would appear in the line chart as a significant positive slope
  The **tag "2018"** shows a significant, easily observed peak, within the period **December 2107 - January 2018** when its popularity increased by **333.3%** jumping from being tagged in **8** different titles to **36**; This peak is expected and understandable during the New Year holidays.
  Another peak is observed in **December 2107** where the **tag "2017"** increased by 100% , while being **tagged** in 24 titles, twice the amount within the previous month.

- Design - See the design section of Insight #4.

- Resources - N/A

Insight #6 – Popular tags and location correlation: Are there distinct regions associated with specific tags?

- Summary – The distribution of the number of titles per location for each tag, show a strong presence of the **tags "NBC"** and **"humor"** in the state of California (**CA**) where they were tagged in **83** and **94** different titles respectively. The **tags "funny"** and **"jokes"** are strongly associated with the states North Carolina (**NC**) and California (**CA**), where the first tag have **46** titles in both states and **44**, **53** titles for the second tag accordingly.

- Design - The Pie charts inside a map provides an efficient visual where the size of each bubble relates to the total number of titles located in the specific state.
  While the size of each angle within the pie related to the number of titles associated with each "tag", which mean that the total number of titles shown in the bubble plot is the sum of all the angles within all the pies for a specific tag. The map will be filtered once a specific tag is chosen from the treemap.

- Resources - See the resources section of Insight #3.