Trending on Social Media (Twitter and YouTube)

Galen Hancock
Darshan Kasat
Monica Sharma
Kyi Win
Phoebe Wu
Ruyue Zhang

OBJECTIVES

- Identify and compare trending topics on various social media platforms over time:
  - Twitter
  - Youtube
- Determine if there is any similarity between the trending subjects on both platforms.

Data Collection

1

Trending videos on YouTube: YouTube API

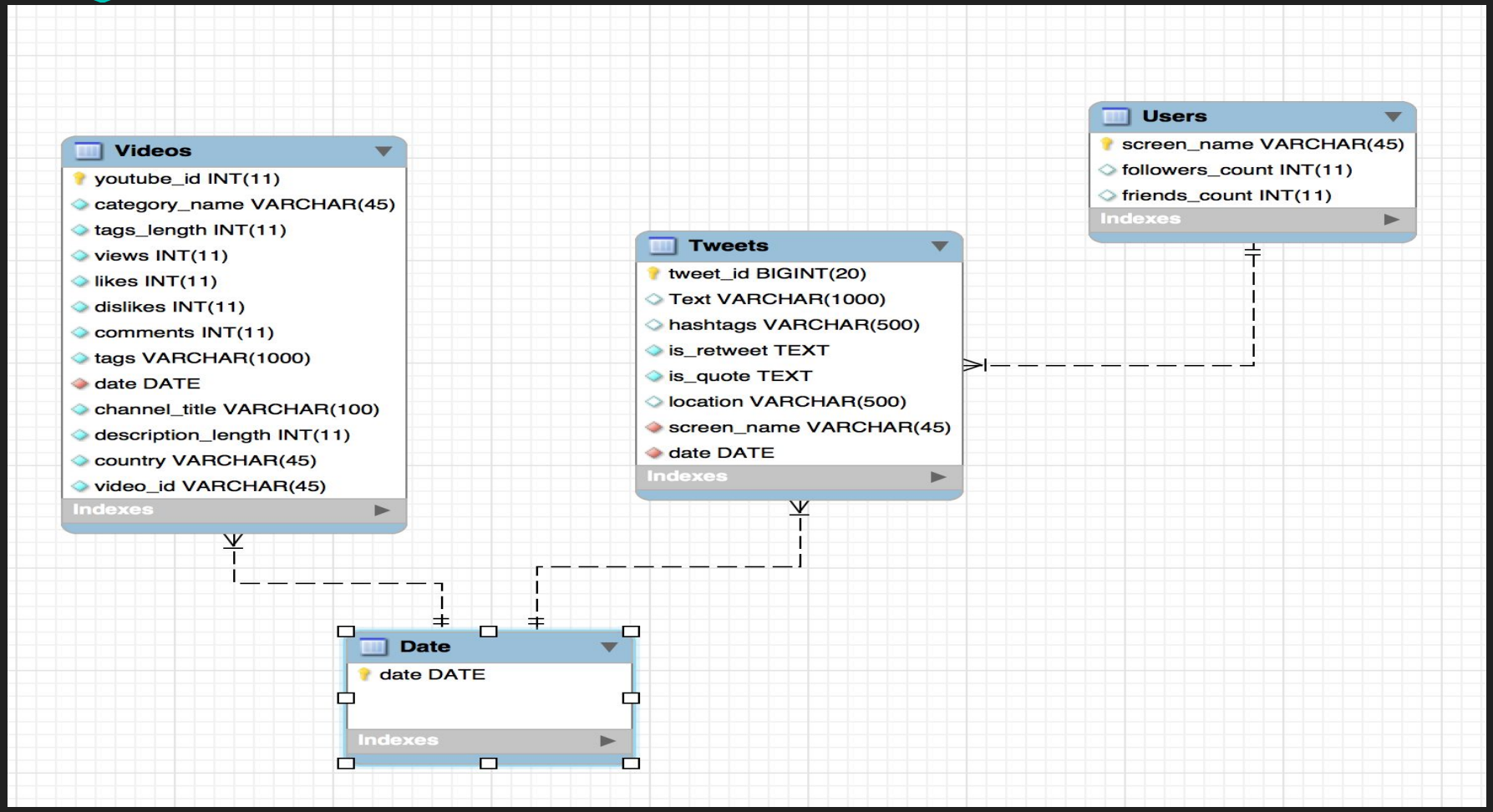
2

Trending tweets on Twitter:
Trendogate
Social Feed Manager

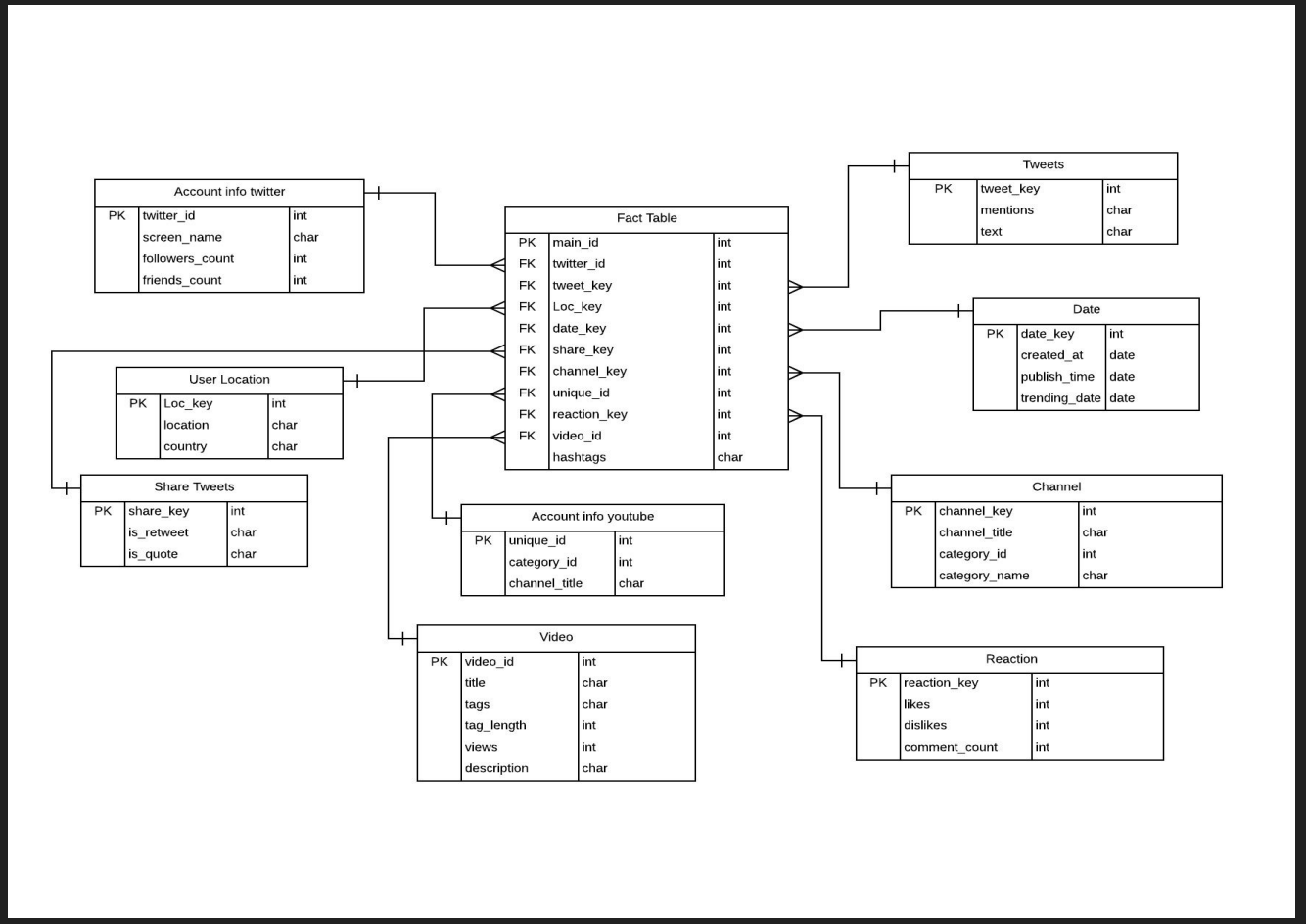
ABOUT THE DATASET

- YouTube:
  i. 200 top trending videos per day.
  ii. Countries: United States, Canada, France, Germany, Great Britain
  iii. 21 columns and around 8,000 rows.
- Twitter:
  i. Around 60,000 tweets per day.
- Time Frame:
  i. 14th November 2017 – 21st November 2017

ER Diagram


Videos
youtube_id INT(11)
category_name VARCHAR(45)
tags_length INT(11)
views INT(11)
likes INT(11)
dislikes INT(11)
comments INT(11)
tags VARCHAR(1000)
date DATE
channel_title VARCHAR(100)
description_length INT(11)
country VARCHAR(45)
video_id VARCHAR(45)
Indexes
Users
screen_name VARCHAR(45)
followers_count INT(11)
friends_count INT(11)
Indexes
Tweets
tweet_id BIGINT(20)
Text VARCHAR(1000)
hashtags VARCHAR(500)
is_retweet TEXT
is_quote TEXT
location VARCHAR(500)
screen_name VARCHAR(45)
date DATE
Indexes
Date
date DATE
Indexes

Dimensional Model



- Star Schema

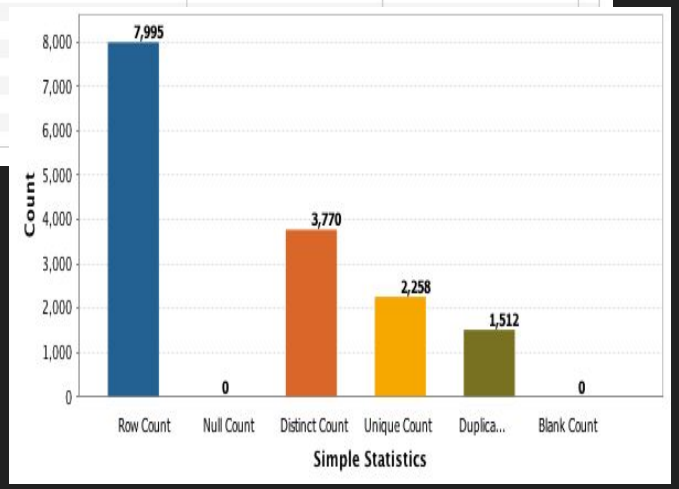
Data Quality Check

- Used Talend tool to check the data quality (Column Analysis)

Column: metadata.video_id

Simple Statistics

Label	Count	%
Row Count	7995	100.00%
Null Count	0	0.00%
Distinct Count	3714	46.45%
Unique Count	2209	27.63%
Duplicate Count	1505	18.82%
Blank Count	0	0.00%

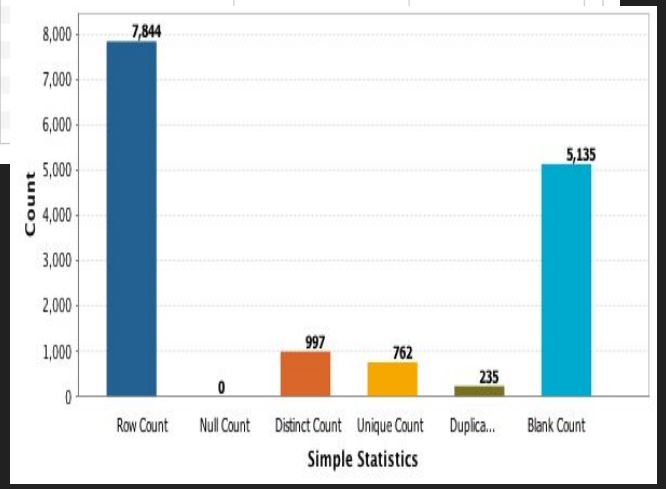


Youtube

Column: metadata.hashtags

Simple Statistics

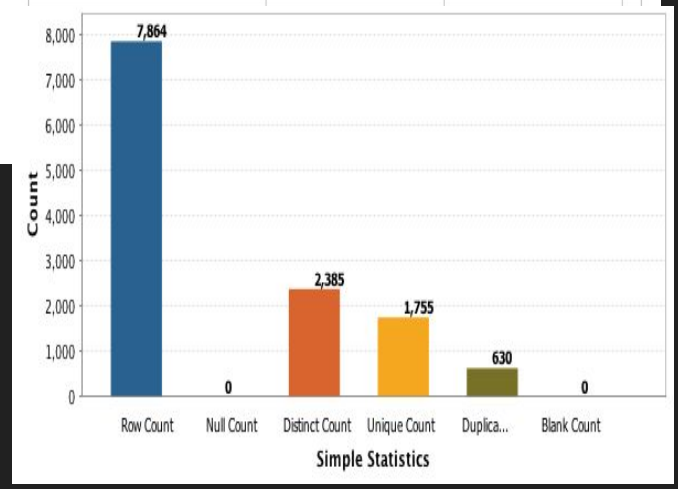
Label	Count	%
Row Count	7844	100.00%
Null Count	0	0.00%
Distinct Count	997	12.71%
Unique Count	762	9.71%
Duplicate Count	235	3.00%
Blank Count	5135	65.46%



Column: metadata.hashtags

Simple Statistics

Label	Count	%
Row Count	7864	100.00%
Null Count	0	0.00%
Distinct Count	2385	30.33%
Unique Count	1755	22.32%
Duplicate Count	630	8.01%
Blank Count	0	0.00%



Twitter

Data Cleaning

- Twitter:
  - Check for missing values:

- YouTube:
  - No missing values were found in the data

```
In [3]: df.isnull().sum()

Out[3]: created_at                    0
        twitter_id                    0
        screen_name                   0
        location                  22859
        followers_count               0
        friends_count                 0
        favorite_count/like_count     0
        retweet_count                 0
        hashtags                  36820
        mentions                  10980
        in_reply_to_screen_name   75042
        twitter_url                   0
        text                          0
        is_retweet                    0
        is_quote                      0
        coordinates               76618
        url1                      64684
        url1_expanded             64684
        url2                      76419
        url2_expanded             76419
        media_url                 65852
        dtype: int64
```

Twitter Data Cleaning

- Selected one thousand observations per day
- Dropped columns :
  - favorite_count, retweet_count, mentions, in_reply_to_screen_name, is_retweet, is_quote, twitter_url, url_1, url1_expanded, url_2, url2_expanded, media_url
- Delete foreign language
- Changed the format of Date column:
  - yyyy-mm-dd

YouTube Data Cleaning

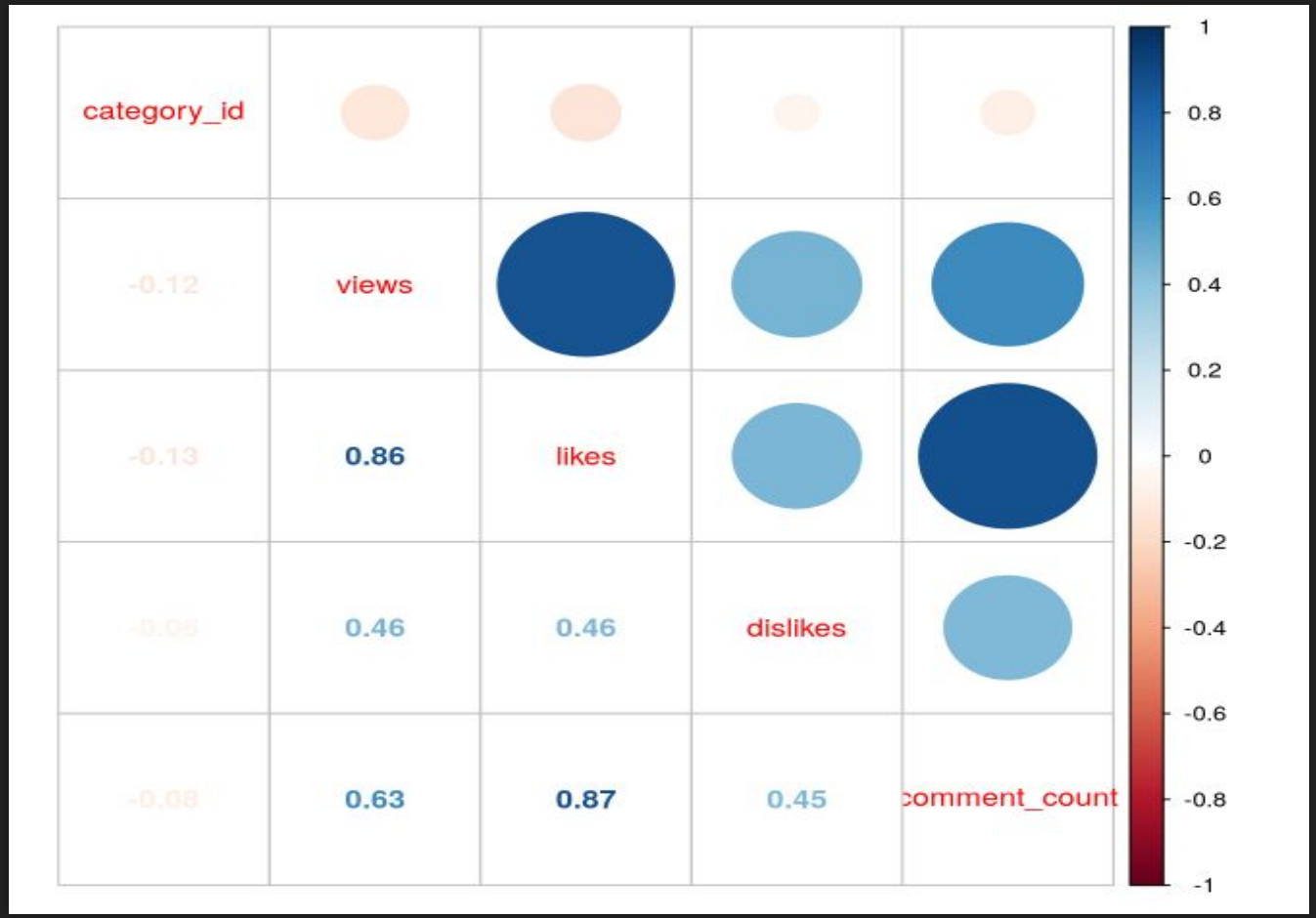
- Dropped Columns :
  - Video_title, category_id, published_time, ratings_disabled, video_error_or_removed, description
- Changed the date format:
  - yyyy-mm-dd
- Created a unique id column to identify each observation.

EDA and Visualizations

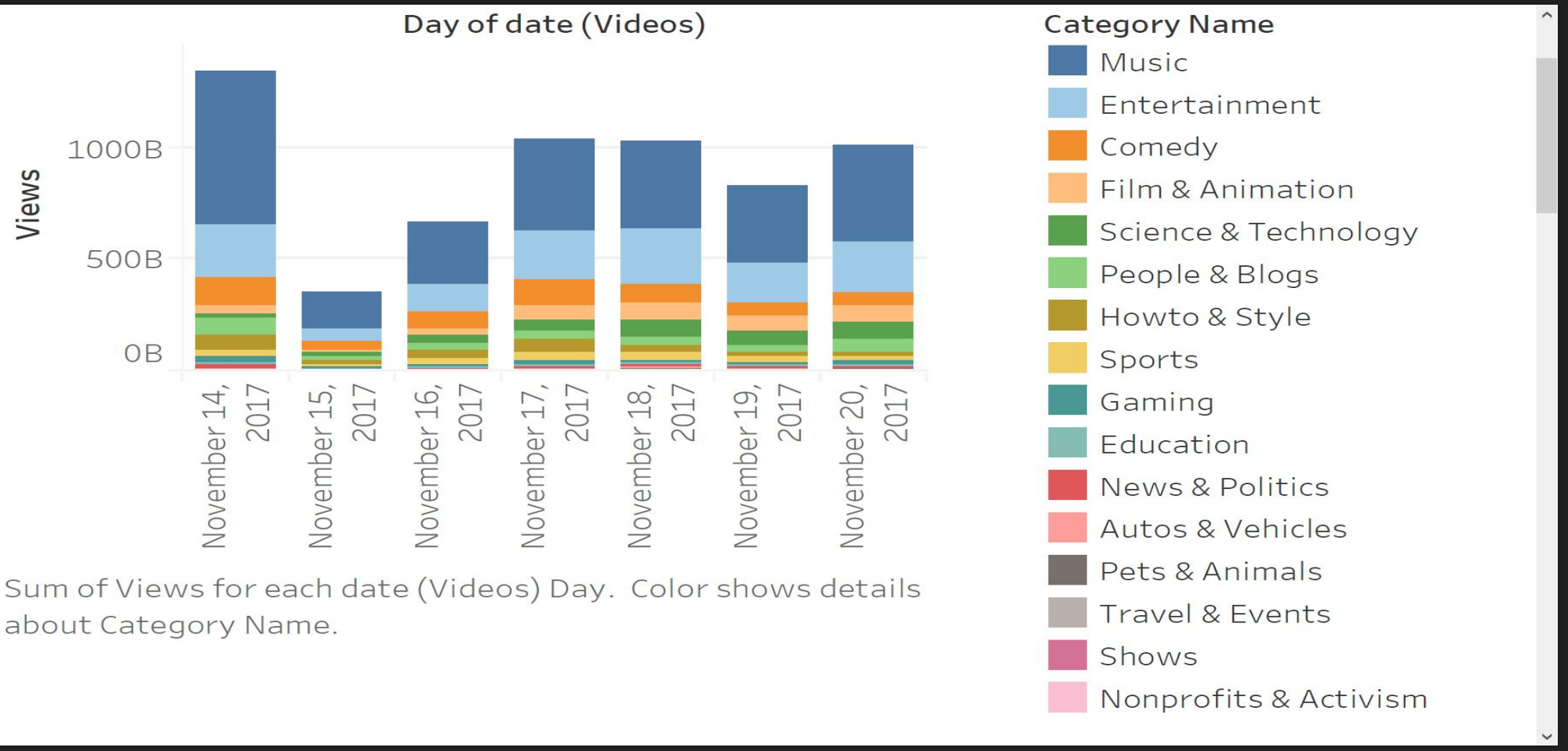
c

Correlation plot:

YouTube



Views by category by date (YouTube)



Demonstration with Tableau

1. https://public.tableau.com/profile/ruyue2989#!/vizhome/twitter_30/Story1?publish=yes
2. https://public.tableau.com/profile/galen.hancock#!/vizhome/YoutubeTags/YoutubeTags
3. https://public.tableau.com/profile/galen.hancock#!/vizhome/TwitterHashtags/TwitterHashtags
4. https://public.tableau.com/profile/galen.hancock#!/vizhome/YoutubeSummary/YoutubeSummary

Challenges

- Difficulty finding data from various social media platforms (Facebook, Tumbler)
- Collecting Twitter data
- Gathering data within the same time frame
- Cleaning data and importing into MySQL WorkBench without error

Summary of Findings

- Based on findings, subjects that "trend" on Youtube during a given timeframe do not necessarily trend on Twitter
- Trending videos and tweets generally correspond to current events
- Different countries may use Youtube for different purposes, according to Youtube category analysis

Implications and Future Research

- Determine if trends are consistent throughout the week, month, year, etc.
  - Is there a best time to post a video/tweet to optimize engagement or viewership?
- Trends on Youtube differed from trends on Twitter, so we cannot treat all social media sites the same
  - Should law enforcement/intelligence community dedicate more resources to monitoring certain websites on certain days of the week?
- Do trends consistently differ by location?
  - Perform analysis by city, state, zip, etc. if possible
  - Free text fields make this difficult (messy data)
  - Do certain sentiments stem from particular regions at particular times?
  - Law enforcement/IC could use this data to distinguish between locations with chronically negative sentiment and locations with new negative sentiment (outliers)

Thank you