5291 Project Report YouTube Video Popularity Analysis

Group 5

PART 1

Project Overview

PROJECT GOAL

1. Find what features are relevant to the video popularity

1. Predict the number of days of a YouTube Video appearing in Trending

1

Data Description

- 6 months of data on daily trending Youtube videos in U.S.
- 14 main factors
  - Video title, Channel title, Category, Tags, Description
  - Trending time, Publish time
  - Views, likes, Dislikes, Comment count
  - Comments_disabled, Ratings_disabled, Video_error_or_removed
- Dataset size: 40,950 (with 6282 unique videos)
- Data source: This dataset was collected by Kaggle using the YouTube API.

1

Define Popularity

- Define Popularity
  - Views
  - Feedback: (dislikes+likes+comments) / views
  - LikeRatio: likes / (dislike+like+1)

PART 2

Exploratory Data Analysis

- Raw Data Analysis
- Popularity Analysis

2

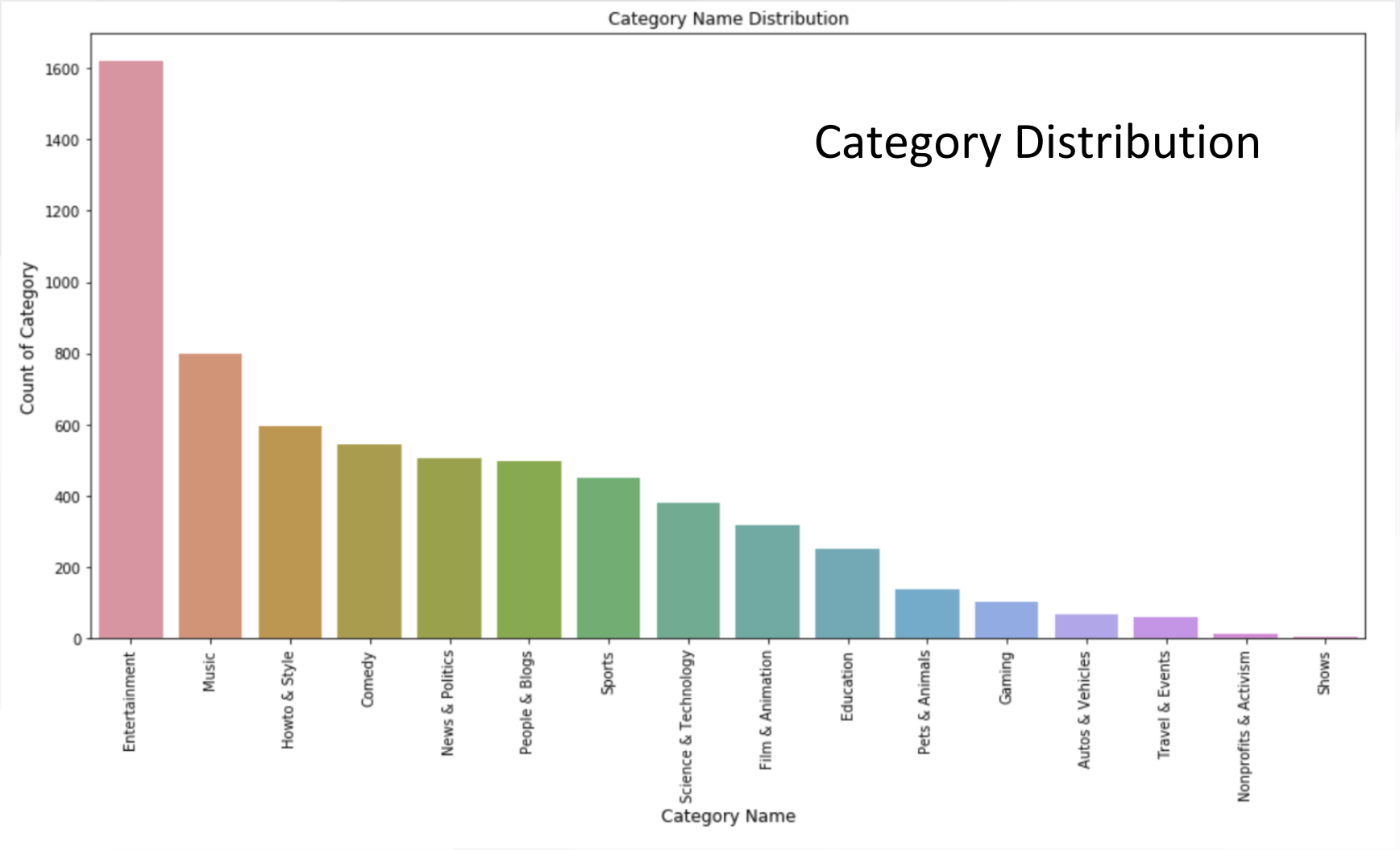
Background

“Each channel on YouTube can be associated with a content category on the platform. These categories help organize the millions of channels and billions of videos on YouTube, and enable viewers, advertisers, and creators to have a common vocabulary and understanding of each audience’s needs. The categorization helps viewers find content, allows advertisers to refine their targeting, and provides creators with common best practices.”

—— YouTube Creators
https://creatoracademy.youtube.com/page/lesson/overview-categories#strategies-zippy-link-2

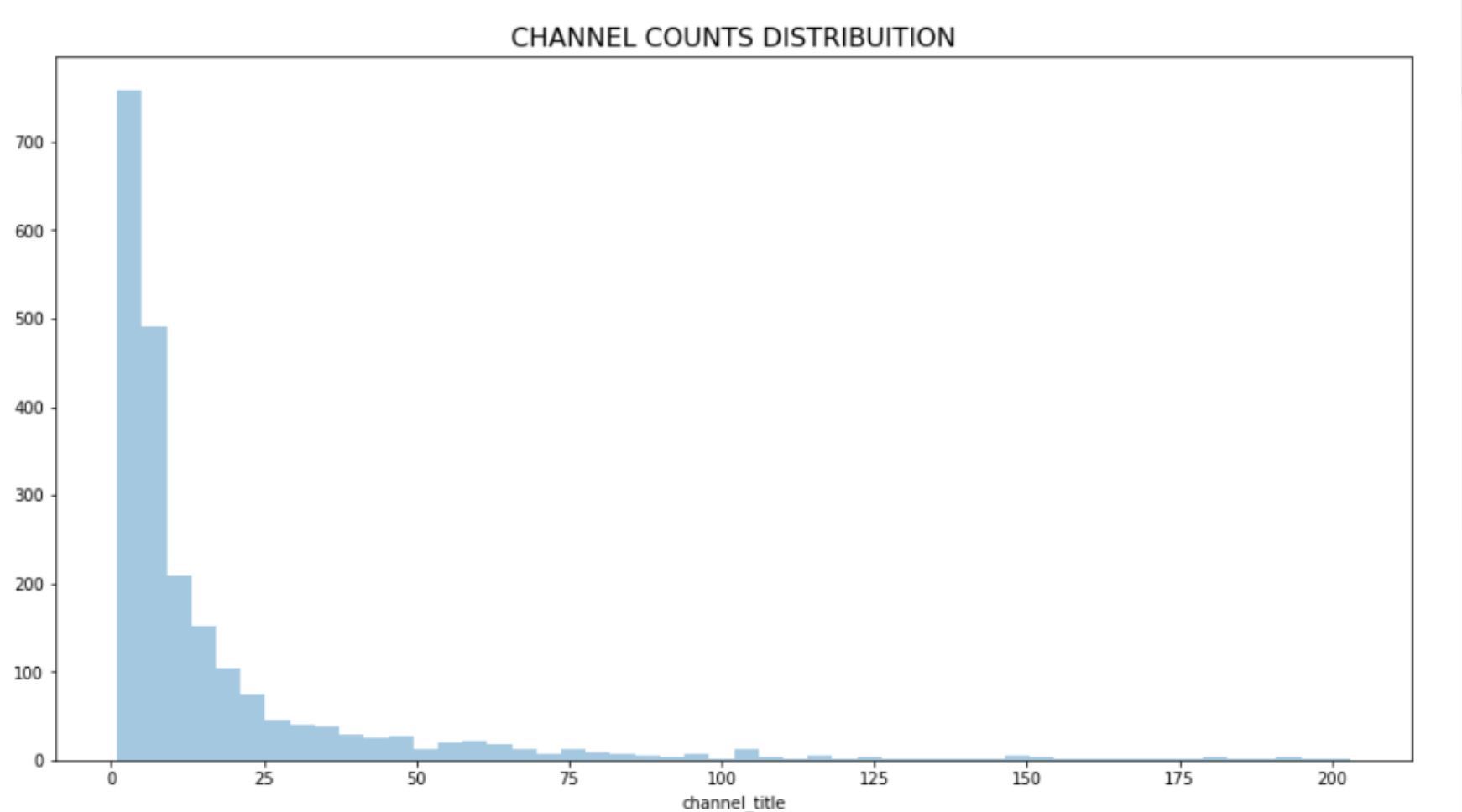
2

Exploratory Data Analysis - Category



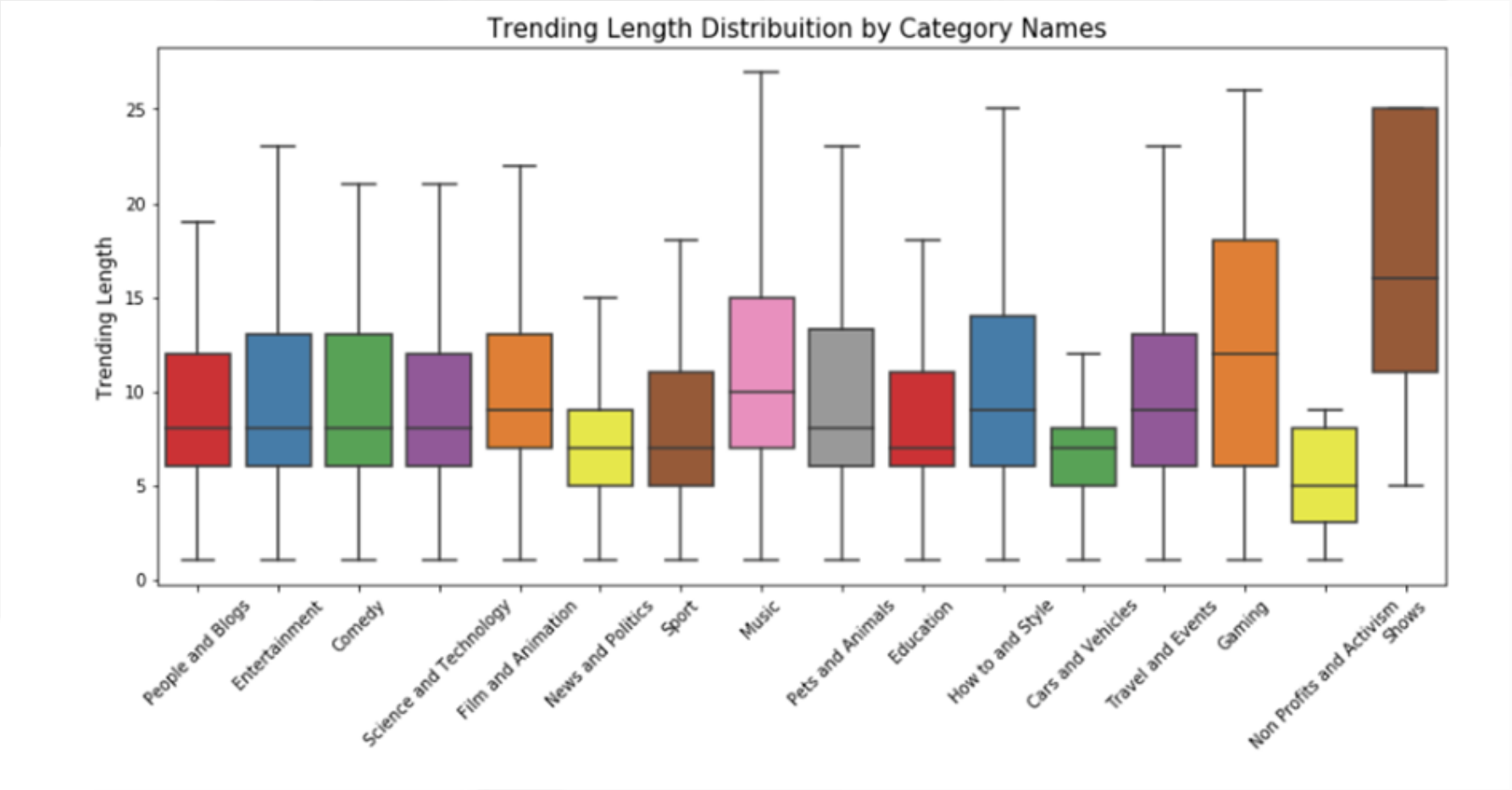
2

Exploratory Data Analysis - Channel



2 Exploratory Data Analysis

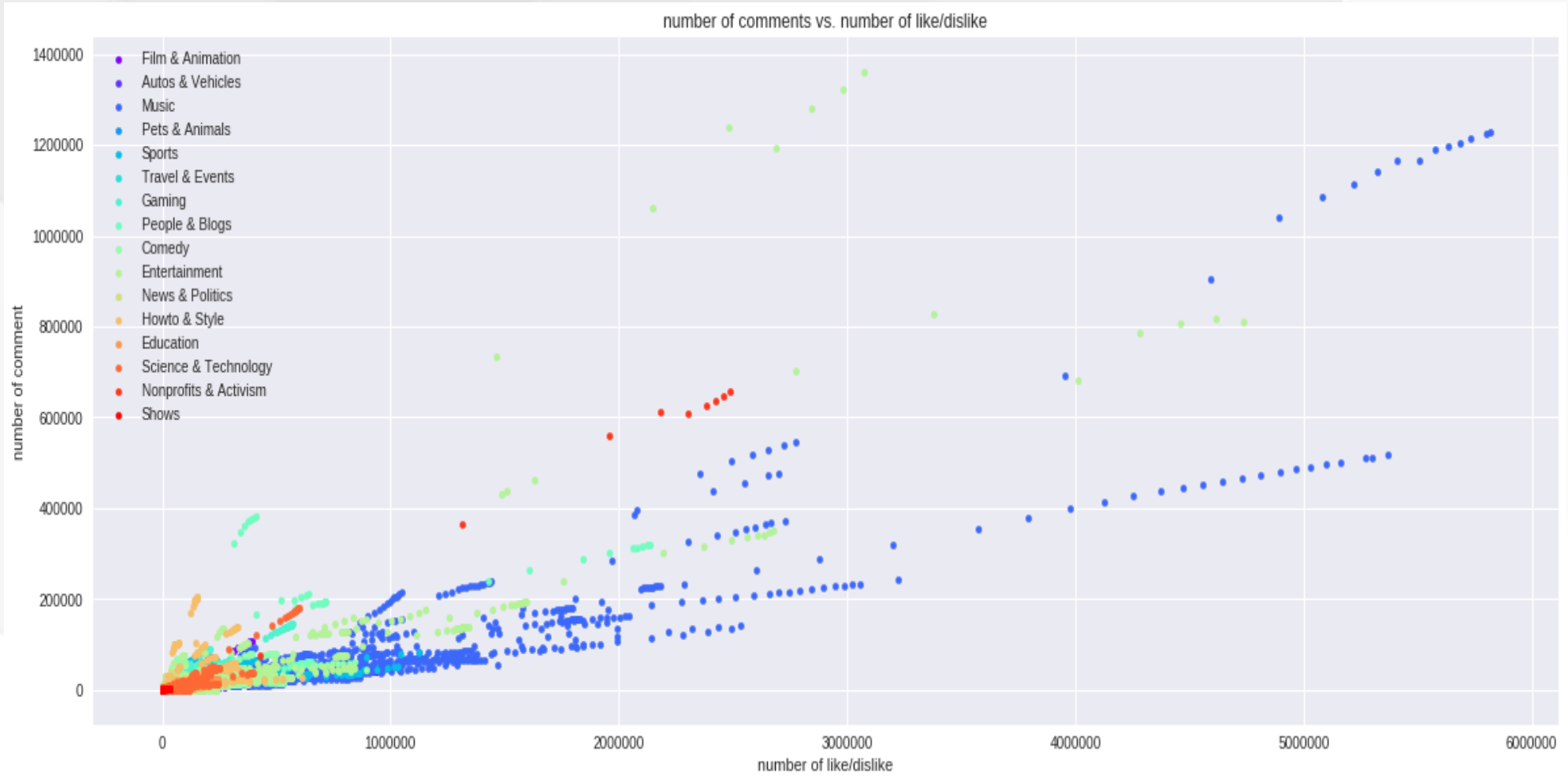
Number of Days of a video appearing in the trending dataset



2

Exploratory Data Analysis - Popularity

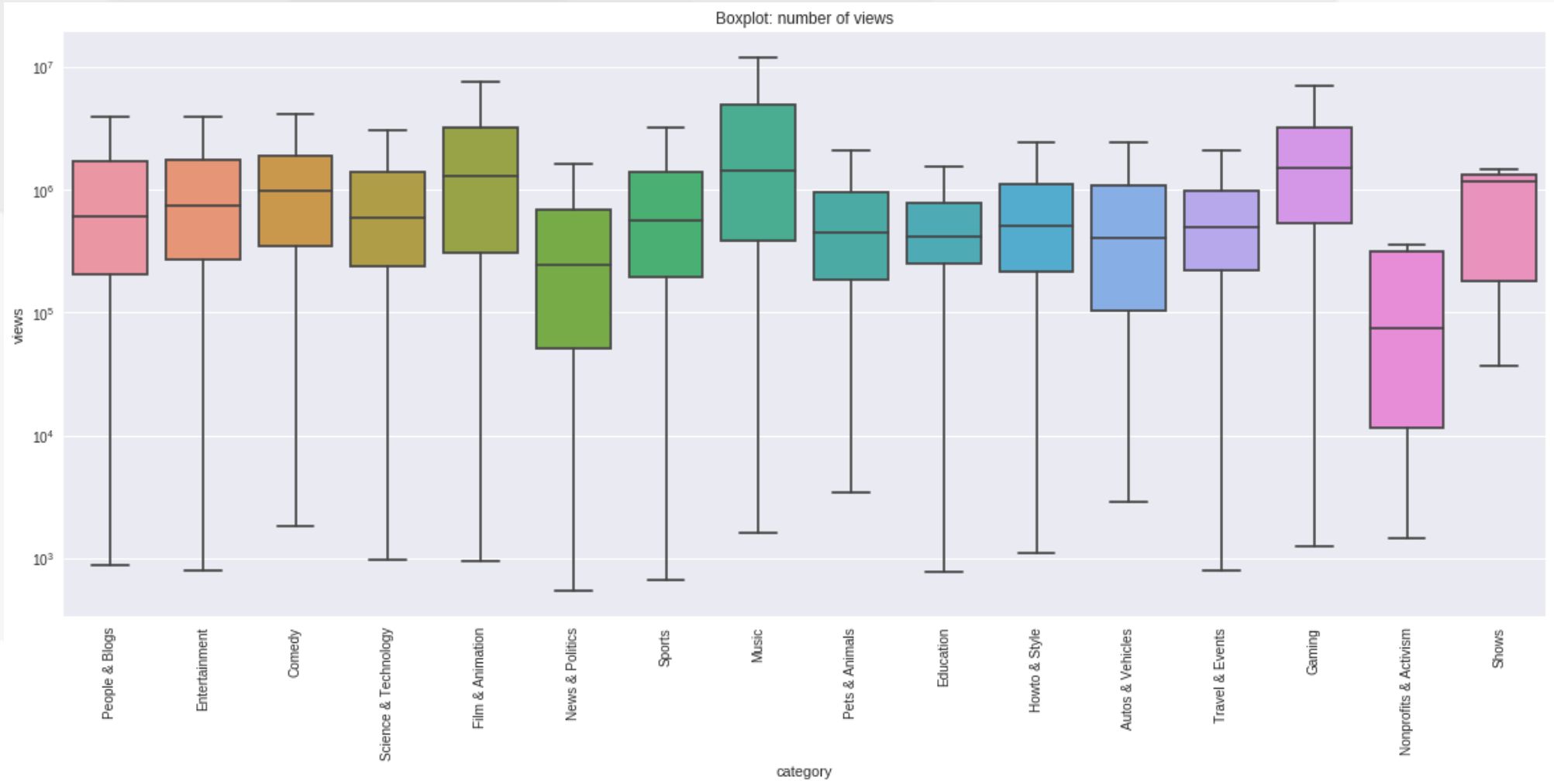
Number of Comments Against Number of (Likes+Dislikes)



2

Exploratory Data Analysis - Views

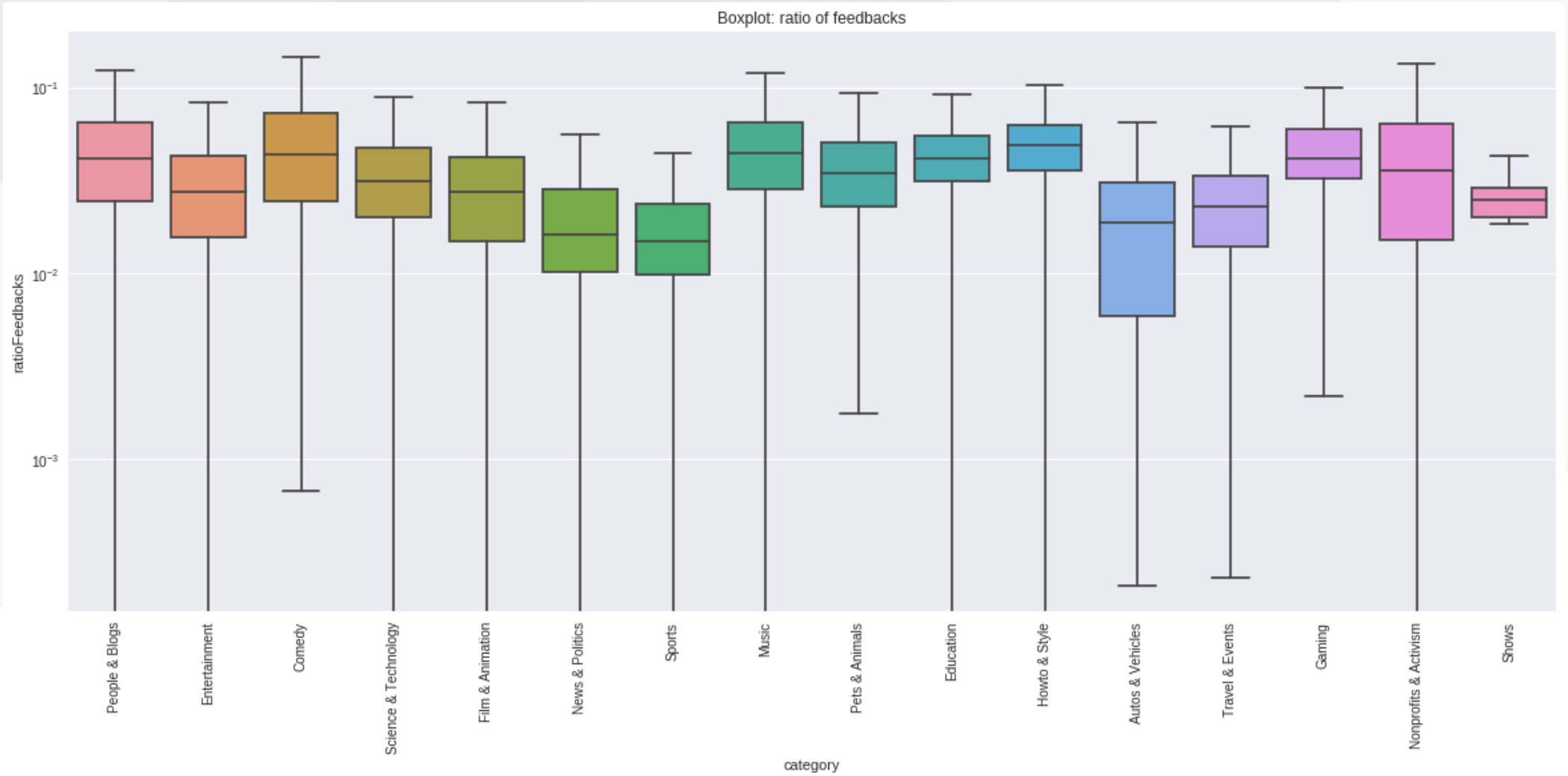
Log of number of views against each category



2

Exploratory Data Analysis - Feedback

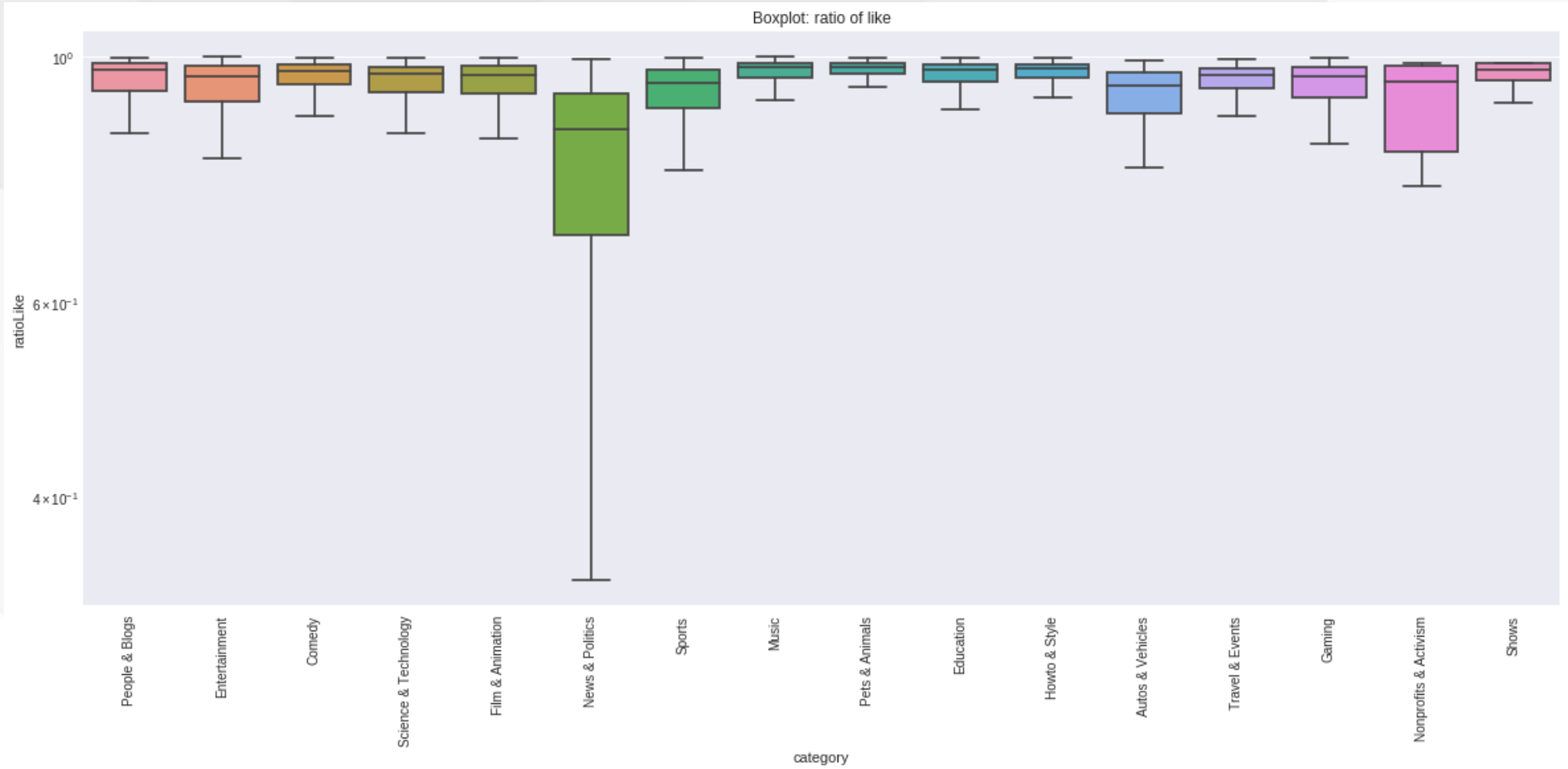
Log of Feedback Ratio (like + dislike + comments)/views against each category



2

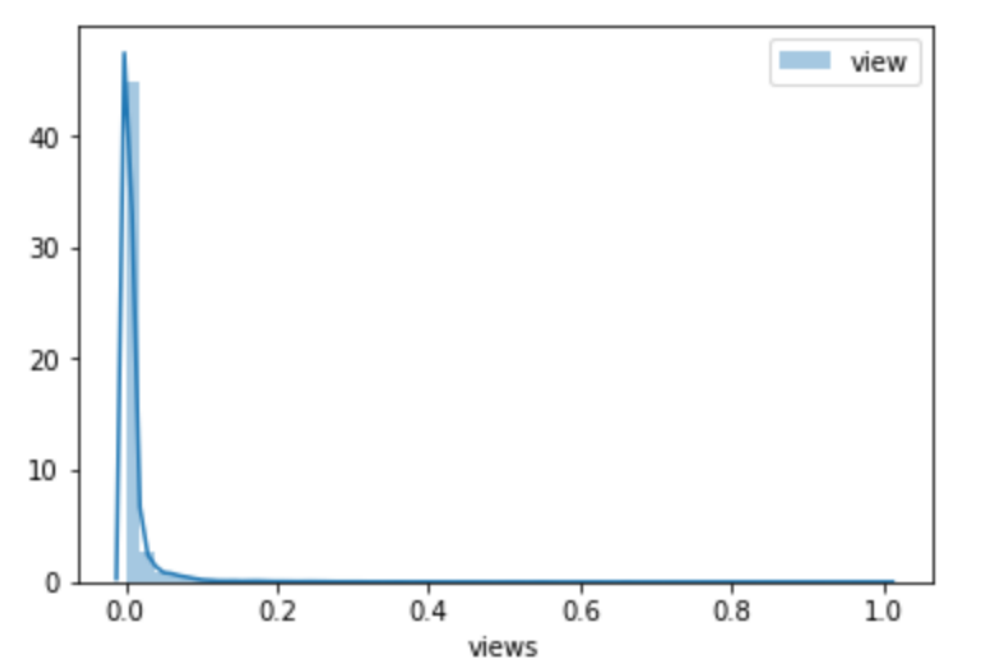
Exploratory Data Analysis - Like Ratio

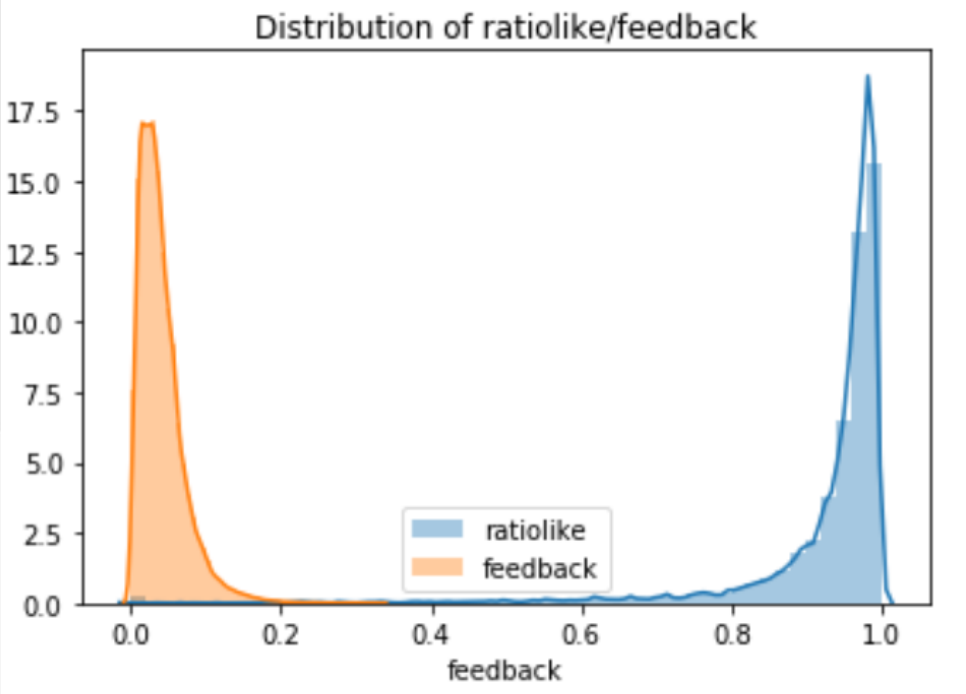
Like Ratio (like / (like + dislike +1)) against each category



2

Exploratory Data Analysis - Popularity





	feedback	ratiolike	views
feedback	1.000000	0.258196	-0.040874
ratiolike	0.258196	1.000000	0.018144
views	-0.040874	0.018144	1.000000

2

Exploratory Data Analysis - Remark

- Analysis category by category can be very helpful.
- People are more likely to express their likes towards vedio than dislike.
- Dislike and comment show similar distributions.
- Views, likeratio and feedback may indicate different aspect of popularity.
    - views defines the interest for the first glance
    - feedback defines how the video impress the viewer
    - ratiolike defines how people like the content

PART 3

Models:

- Feature Engineering
- Predictive Model

3

Model Goal:

How many times the video will be on the trending board within 30 days after publishing?

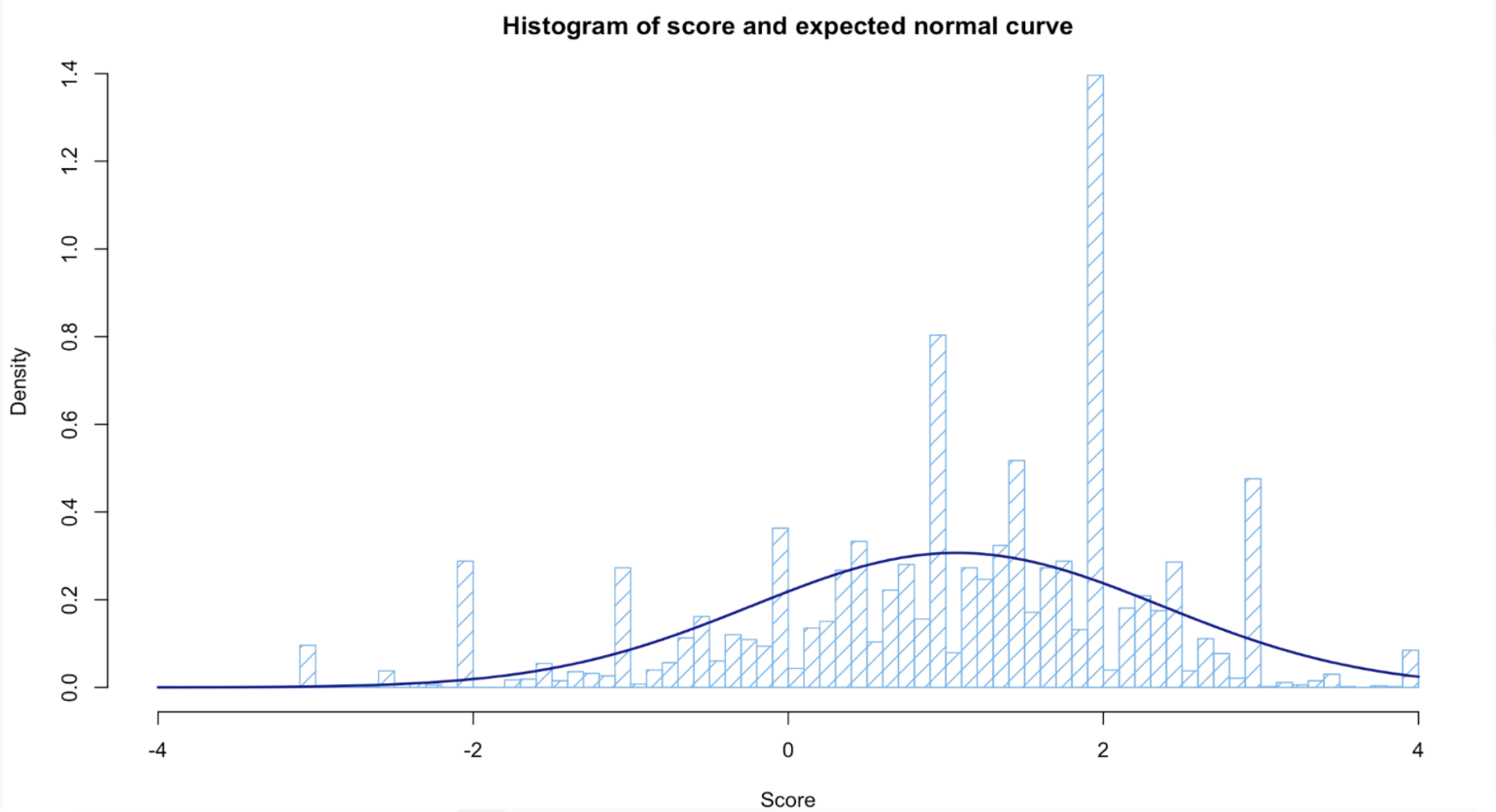
Dataset:

We select distinct videos and use the first appearance row as our dataset and count how many times it will appear within 30 days after publishing as predicted variable.
Size: 6282 rows

3

Feature Engineering - Sentiment Analysis

Description: sentiment analysis generate score



3

Feature Engineering - Popularity Metrics

- Feedback: (dislikes+likes+comments) / views
- LikeRatio: likes / (dislike+like+1)

	feedback	ratiolike	views
feedback	1.000000	0.258196	-0.040874
ratiolike	0.258196	1.000000	0.018144
views	-0.040874	0.018144	1.000000

PART 4

Future Work

- **Feature Engineering**
- **Model**

4

Future Work

Feature Engineering

1. Transform text data (description and tags) into vector using **word2vec**

1. Split by category, and conduct feature engineering (sentiment analysis) within each category

Assumption Testing

3. After featuring engineering, we will conduct assumption testing for each model we are going to use.

4

Future Work - Model

Linear Regression Model:

a. OLS Linear Regression
b. Ridge Regression
c. LASSO Regression

Non-Linear Regression Model:

a. Tree-Based Method
b. Gradient Boosting Machine

3

Appendix - Some Approaches

Here are some approaches we attempted but with poor effects

1. Perform LDA (Latent Dirichlet Allocation) on title and description to create new features. It's meaningless after knowing the category name.

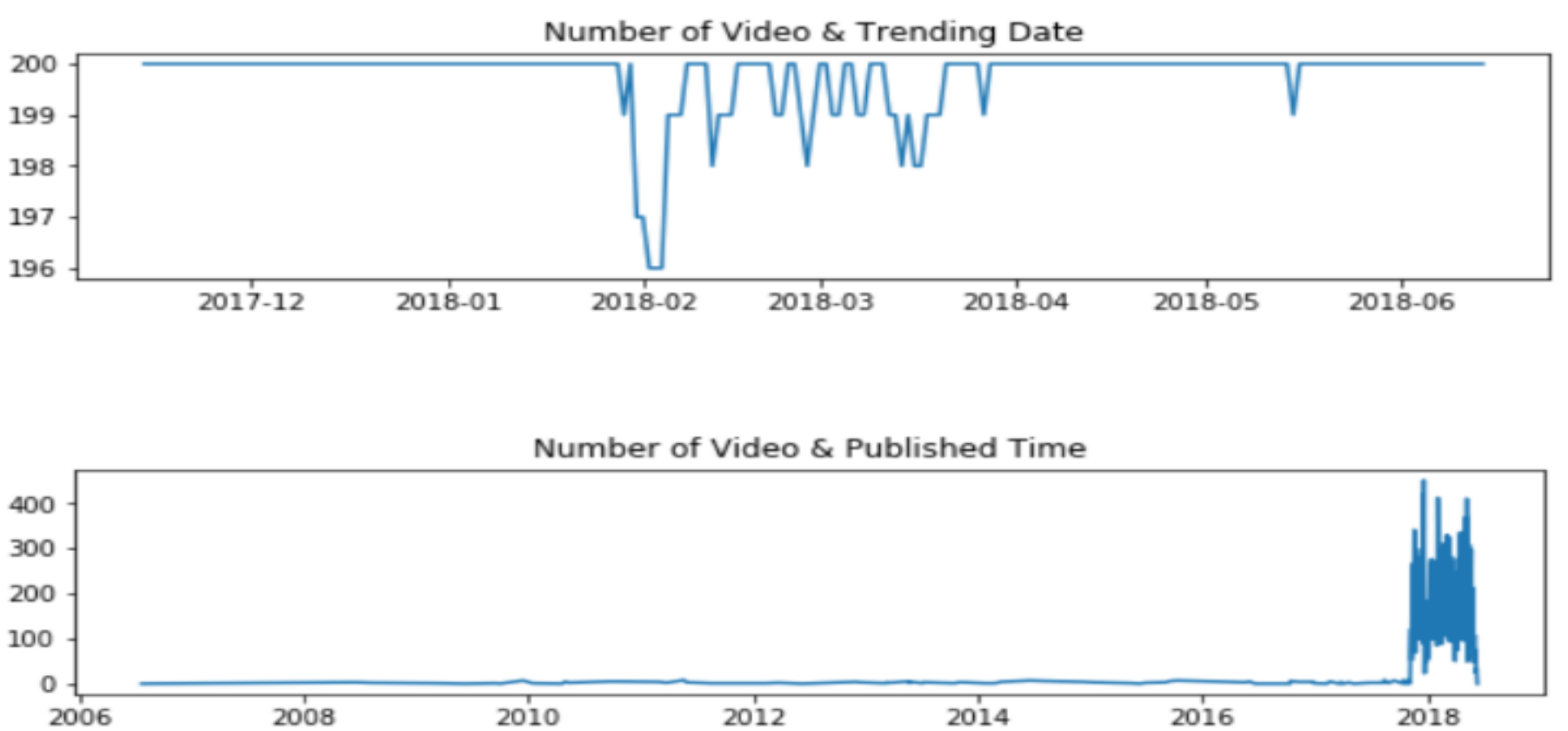
Topic #0: official, video, trailer, music, live, game, world, teaser, super, ready, theory, film, awards, highlight, lovato
Topic #1: things, true, ball, makeover, tour, hart, home, getting, cast, hair, fair, meet, secret, fashion, vanity
Topic #2: best, giant, cake, look, everything, chocolate, blind, short, wrong, slow, light, grace, spider, good, babish
Topic #3: time, year, meghan, school, netflix, song, prince, reveal, taylor, spot, league, doctor, swift, actually, found
Topic #4: official, video, trail, lyric, feat, beauty, final, room, honest, trailers, march, hawaii, gets, wild, much
Topic #5: season, like, american, never, kardashian, idol, story, king, charlie, special, coming, puth, view, performs, greatest
Topic #6: show, james, battle, voice, people, michael, lost, left, lebron, volcano, daily, using, fall, facts, times
Topic #7: first, full, royal, know, espn, high, stephen, behind, highlights, history, without, david, childish, gambino, fire
Topic #8: make, perfect, face, inside, made, shawn, miss, vogu, million, google, jennifer, dude, deep, challeng, products
Topic #9: makeup, wedding, talk, trump, makes, goes, apple, artist, scott, tutori, national, jenner, sneaker, complex, kylie
Topic #10: movie, review, tried, infinity, black, panther, america, avengers, marvel, studios, scene, dead, justin, eating, anthem
Topic #11: love, react, life, real, every, making, kevin, fake, conan, call, space, mystery, champions, asmr, following
Topic #12: challenge, black, food, christmas, take, simon, earth, amazon, japanese, breaking, scen, line, dress, white, kiss
Topic #13: audio, john, last, night, cardi, part, ever, girl, week, heart, chris, fortnite, jedi, graham, plays
Topic #14: star, bowl, watch, wars, kids, episode, adam, back, mendes, cooking, breaks, iphone, jordan, interview, open
Topic #15: test, house, youtube, golden, questions, disney, shopping, bought, taste, camila, college, solo, fish, cabello, gadgets

3

Appendix - Some Approaches

Here are some approaches we attempted but with poor effects

2. Attempt to extract patterns on Time Series and predict popularity. not started yet

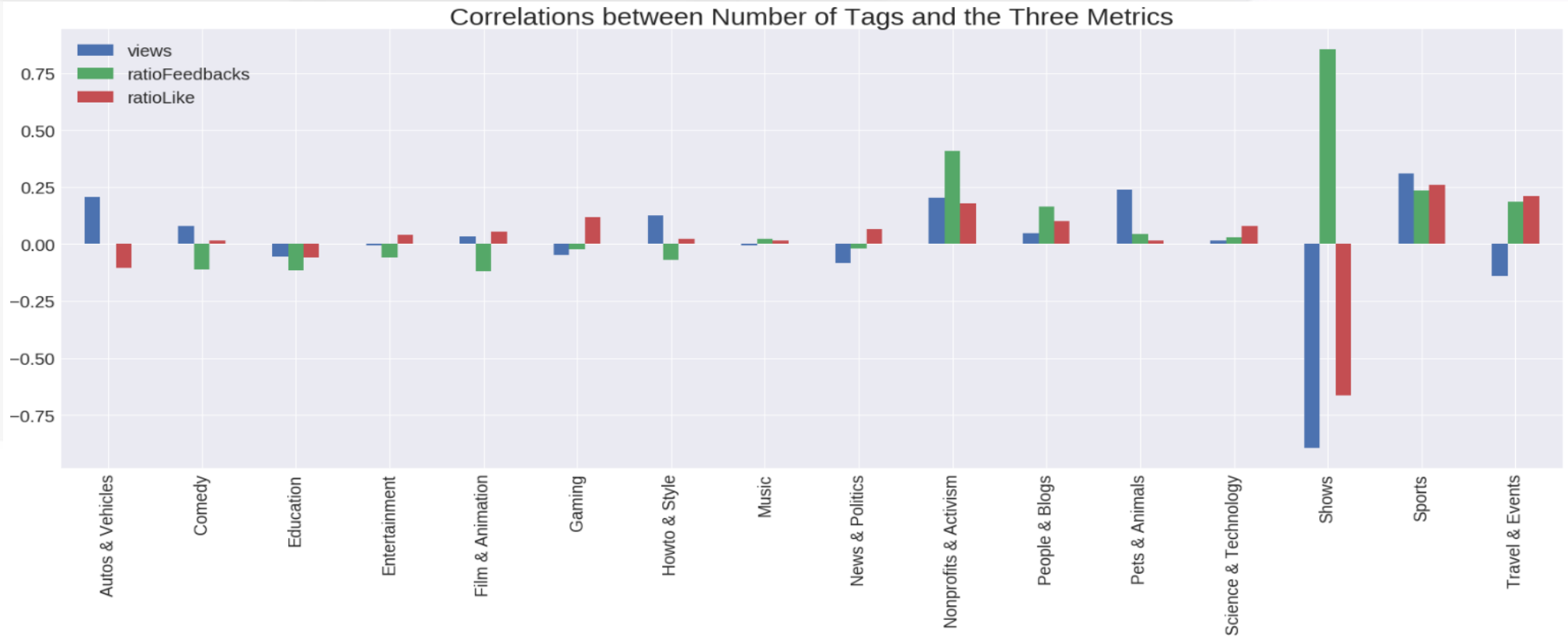


3

Appendix - Some Approaches

Here are some approaches we attempted but with poor effects

3. Analyze on the contribution of the attribute ‘tags’ e.g. analyze the correlation between number of tags and our popularity metrics no correlation



THANKS!