BT4221 Big Data Techniques and Technologies

Going Big (Data) on YouTube

Prepared by: Group 1
Chua Kai Bing, Goh Jia Yi, Ngoc Linh Chi Nguyen, Tan Zen Wei

MEET THE TEAM

NGOC **LINH CHI** NGUYEN

Y4 Computer Engineering

CHUA **KAI BING**

Y3 Business Analytics

TAN **ZEN WEI**

Y3 Business Analytics

GOH **JIA YI**

Y3 Business Analytics

AGENDA

1. Introduction
2. Data Collection

LINH CHI

3. Data Cleaning & Preprocessing
4. Data Exploration

KAI BING

5. Feature Extraction
6. Feature Selection

ZEN WEI

7. Model Building & Evaluation
8. Conclusion

JIA YI

INTRODUCTION

Hypothesis

A. “A **positive sentiment title** in a YouTube video will affect views”

B. “Having **humans in the thumbnail images** of a YouTube video will affect views”

DATA COLLECTION

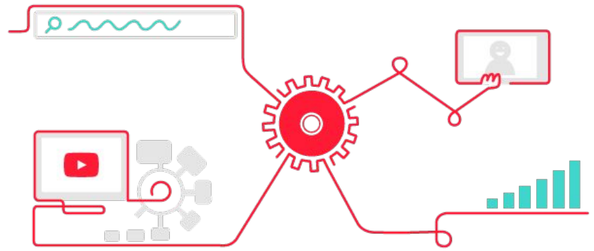
2

Data Collection

SOCIALBLADE

Extract **channel ID**

YouTube Data API v3



3 Datasets: SocialBlade, Channel, Video

Collecting Top 10 Most Subscribed Channels

17 Categories from Social Blade

- Auto & Vehicles
- Comedy
- Education
- Entertainment
- Film
- Gaming
- How to & Style
- Made for Kids
- Music
- News & Politics
- Nonprofit & Activism
- People & Blogs
- Pets & Animals
- Science & Technology
- Shows
- Sports
- Travel

“Shows” Category

Sorted by: Subscribers

Rank	Grade	Username	Uploads	Subs	Video Views
1st	D-	Sony Pictures Home Entertainment	1,341	762K	47,659,441
2nd	B	CartoonNetworkEps	1,479	367K	210,160,578
3rd	C-	ABCTVONDEMAND	20	140K	6,501
4th	D-	FHEDreamworksAnimate	26	87.8K	2,587,810
5th	B-	truTVEps	1,321	81.6K	26,121,802
6th	D-	tlcfullepisodes	70	77.4K	37,582
7th	C-	FHEFoxNetworkTV	1	74.5K	16,032

- ✗ Low subscriber counts
- ✗ Low video counts

Data Description

16 Categories from Social Blade

- Auto & Vehicles
- Comedy
- Education
- Entertainment
- Film
- Gaming
- How to & Style
- Made for Kids
- Music
- News & Politics
- Nonprofit & Activism
- People & Blogs
- Pets & Animals
- Science & Technology
- ~~Shows~~
- Sports
- Travel

Social Blade

(Example) Top 5 subscribed channels in the Gaming Category:

Sorted by: Subscribers

Rank	Grade	Username	Uploads	Subs	Video Views
1st	A	PewDiePie	4,231	107M	26,373,582,863
2nd	A-	Fernanfloo	537	38.4M	8,615,419,881
3rd	A	VEGETTA777	6,208	31.5M	13,329,834,031
4th	A	Markiplier	4,802	27.3M	14,289,042,024
5th	A	jacksepticeye	4,677	25.4M	13,417,626,042

Channel Name

https://socialblade.com/youtube/top/category/games/mostsubscribed

Dataset 1: Social Blade Data

Channel Name, Channel ID, Category

Channel Name	Channel ID	Category
Top Gear	UCjOl2AUblVmg2rA_cRgZkFg	AUTO & VEHICLES
ChrisFix	UCes1EvRjcKU4sY_UEavndBw	AUTO & VEHICLES
MotorTrend Channel	UCsAegdhiYLEoaFGuJFVrqFQ	AUTO & VEHICLES
carwow	UCUhFaUpnq31m6TNX2VKVSVA	AUTO & VEHICLES
Supercar Blondie	UCKSVUHI9rbbkXhvAXK-2uxA	AUTO & VEHICLES
CGP Grey	UC2C_jShtL725hvbm1arSV9w	AUTO & VEHICLES
Donut Media	UCL6JmiMXKoXS6bpP1D3bk8g	AUTO & VEHICLES
Doug DeMuro	UCsqjHFMB_JYTaEnf_vmTNqg	AUTO & VEHICLES
Scotty Kilmer	UCuxpxCCevIlF-k-K5YU8XPA	AUTO & VEHICLES
Hoonigan	UCxHOwUUMJZQBgXttO3OjAIA	AUTO & VEHICLES
The Tonight Show Starring Jimmy Fallon	UC8-Th83bH_thdKZDJCrn88g	COMEDY
Smosh	UCY30JRSgfhYXA6i6xX1erWg	COMEDY
Mr Bean	UCkAGrHCLFmlK3H2kd6isipg	COMEDY
shane	UCV9_KinVpV-snHe3C3n1hvA	COMEDY
JennaMarbles	UC9gFih9rw0zNCK3ZtoKQQyA	COMEDY

YouTube Data API

- ✓ Used YouTube Data API v3
- ✓ Limited to 10,000 queries per day

YouTube Data API v3

Google

The YouTube Data API v3 is an API that provides access to YouTube data, such as videos, playlists,...

MANAGE TRY THIS API API Enabled

Quota exceeded errors count (3 hr) - Queries per day

Quota Name	Limit
Queries per day	10,000
Queries per minute per user	180,000
Queries per minute	1,800,000

Dataset 2: Channel Data

Channel ID, Description, Published At, Region, Subscriber Count, Video Count, View Count

channelId	description	publishedAt	region	subscriberCount	videoCount	viewCount
UC07-dOwgza1IguKA86jqxNA	The official public health informat	2005-10-25T13:14:08Z	US	626000	1664	57909283
UCh4pyZUB0mNzieaKv831flA	We imagine a world where everyc	2008-11-26T07:24:00Z	US	696000	1060	156537119
UCc4yillQaNo6a-iG2PYbbrA	Hey everyone! Welcome to my ch	2007-01-03T22:52:37Z	US	709000	191	79229538
UCB7BryuXaMe1pUMznYAq4Jg	My channel has one simple goal, t	2013-10-24T14:56:37Z	US	770000	57	99831679
UCBP4B896svWOcWdRp8UjH2Q	Jace Norman / Xander Norman //	2012-01-02T04:40:17Z	US	880000	28	30837600
UCg3_C7BwcV0kBlJbBFHTPJQ	Global Citizen is a community of p	2008-09-22T23:18:45Z	US	975000	1607	302840751
UC3z9rRoiie3RXzhUxv1Bdqw	The OFFICIAL Planetshakers YouT	2007-09-04T05:49:37Z	US	1100000	588	291676204
UCn4sPeUomNGIr26bElVdDYg	NowThis is your premier news ou	2012-09-05T12:09:54Z	US	1160000	7620	679627487
UCYv-siSKd3Gn9IsliO95gIw	Representing God to the Lost and	2012-09-24T22:03:23Z	US	1190000	262	84967110
UCVSNOxehfALut52NbkfRBaA	Voice of America (VOA) is the larg	2008-03-14T17:13:45Z	US	1380000	34967	635231893
UCdNjexbIS_NKJC4ZRwKf9ag	This is the media channel of The C	2008-08-07T16:20:47Z	US	1470000	2642	458841789
UC0Ize0RLIbGdH5x4wI45G-A	Hey guys! My name is Drew Binsk	2012-07-26T03:24:35Z	US	1940000	781	379113325
UChLtXXpo4Ge1ReTEboVvTDg	Welcome to the official Global Ne	2009-11-04T20:46:32Z	US	2060000	20086	1016907424
UC9avFXTdbSo5ATvzTRnAVFg	NEW VIDEOS MONDAYS AND THU	2011-05-22T00:19:55Z	US	2120000	873	422298713
UCxDZs_ltFFvn0FDHT6kmoXA	Going off the map in former Sovie	2018-06-12T11:56:25Z	US	2250000	212	293687890

Dataset 3: Video Data

Channel ID, Description, Published At, Video ID, Thumbnail, Video Title, Comment Count, Dislike Count, Like Count, View Count

channelId	description	publishedAt	videoId	thumbnails	videoTitle	commentCount	dislikeCount	likeCount	viewCount
UCjOl2AU	In this	10/9/2020 9:35	NXX338W	https://i.ytim	PREVIEW: Att	528	257	5819	184447
UCjOl2AU	From the	10/9/2020 11:21	dtHcdU2c	https://i.ytim	Which car wil	568	273	7136	217619
UCjOl2AU	Here's Chris	10/7/2020 7:40	vnrtWe-R	https://i.ytim	Chris Harris o	1091	408	10189	437777
UCjOl2AU	16	10/6/2020 13:59	Ra1F0TsO	https://i.ytim	Chris Harris v	579	202	7126	191070
UCjOl2AU	The 986bhp	10/6/2020 7:36	fXysipmT	https://i.ytim	FASTEST TOP	888	168	9697	572569
UCjOl2AU	Top Gear's	10/2/2020 15:24	Wxz6wd8	https://i.ytim	Chris Harris D	1137	455	15641	848767
UCjOl2AU	Watch our	9/28/2020 20:09	4iooMysV	https://i.ytim	FIRST DRIVE:	961	673	8076	291800
UCjOl2AU	As Series 29	9/7/2020 11:40	Rdb8k4nD	https://i.ytim	Ferrari SF90,	350	276	4918	218842
UCjOl2AU	The new	9/23/2020 16:18	ImL_QbUv	https://i.ytim	FIRST DRIVE:	1699	846	14218	786868
UCjOl2AU	The BMW	9/22/2020 17:38	qBStRheB	https://i.ytim	FIRST LOOK: M	1772	750	5303	249576
UCjOl2AU	Top Gear is	9/11/2020 11:20	IMk0xEJ1	https://i.ytim	FIRST LOOK T	778	1053	3854	250697
UCjOl2AU	As the	9/10/2020 14:22	bUfeG-Vn	https://i.ytim	MCLAREN SH	695	336	9108	355497
UCjOl2AU	Ever	8/31/2020 11:54	7oveAHy8	https://i.ytim	Driving the Â	1578	944	21528	830987
UCjOl2AU	From the	8/28/2020 9:42	EASgr2Lrw	https://i.ytim	Best of Porsc	266	139	3252	129057
UCjOl2AU	In the UK?	8/10/2020 8:50	FkCIlzcYH-	https://i.ytim	EXTENDED: D	811	1019	3727	219546
UCjOl2AU	Series 28	8/10/2020 8:10	kKhkX10N	https://i.ytim	Fastest Cars c	691	602	8931	555464
UCjOl2AU	How on	8/3/2020 23:09	K4EIYQ6fk	https://i.ytim	The secrets b	3285	916	25234	1132180

DATA CLEANING & PREPROCESSING

3

Data Cleaning

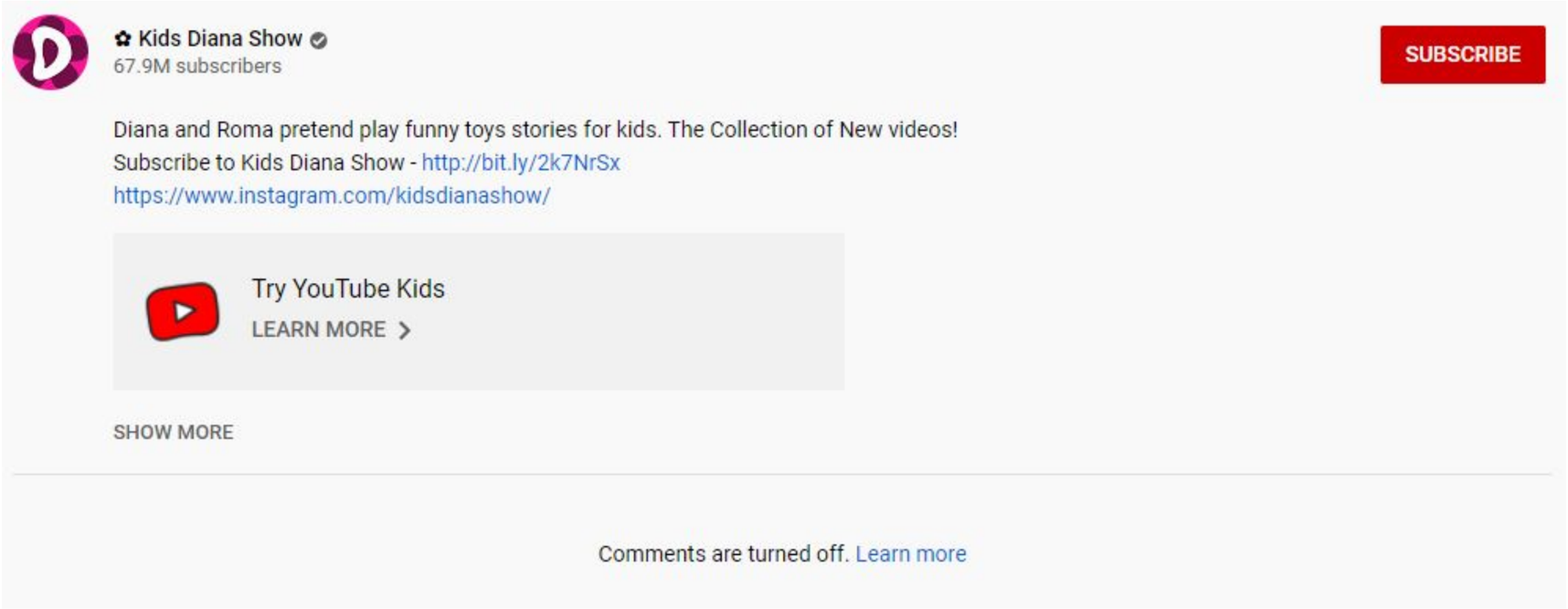
INCORRECT DATA

There were **invalid** dates and numerical fields with **strings**

MISSING DATA

Null values in numerical fields such as like count, dislike count and comment count

Missing Numerical Data



Hidden statistics

Data Preprocessing

DROP ROWS WITH:

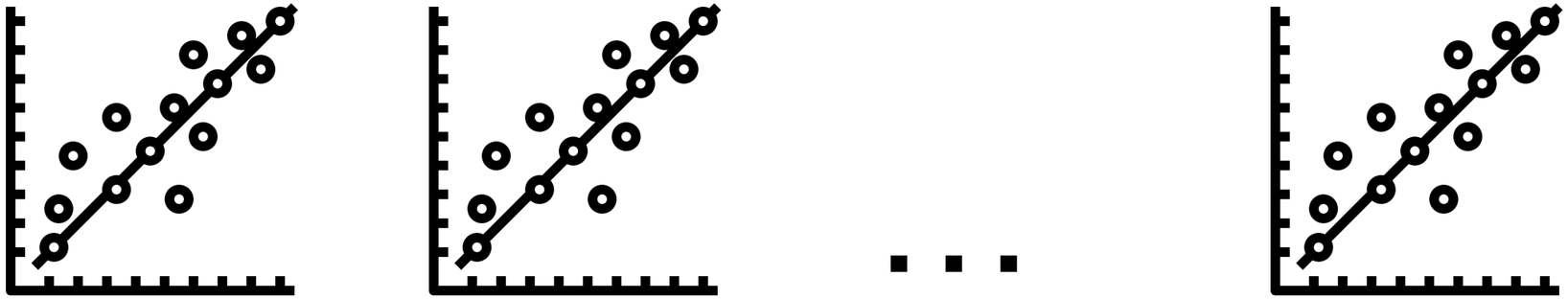
- Incorrect data
- Missing like counts/dislike count/view count

IMPUTE MISSING VALUES WITH REGRESSION:

- Comment Count

Rationale of Using Linear Regression

Using mean or 0 to replace the missing values would **misrepresent** the data

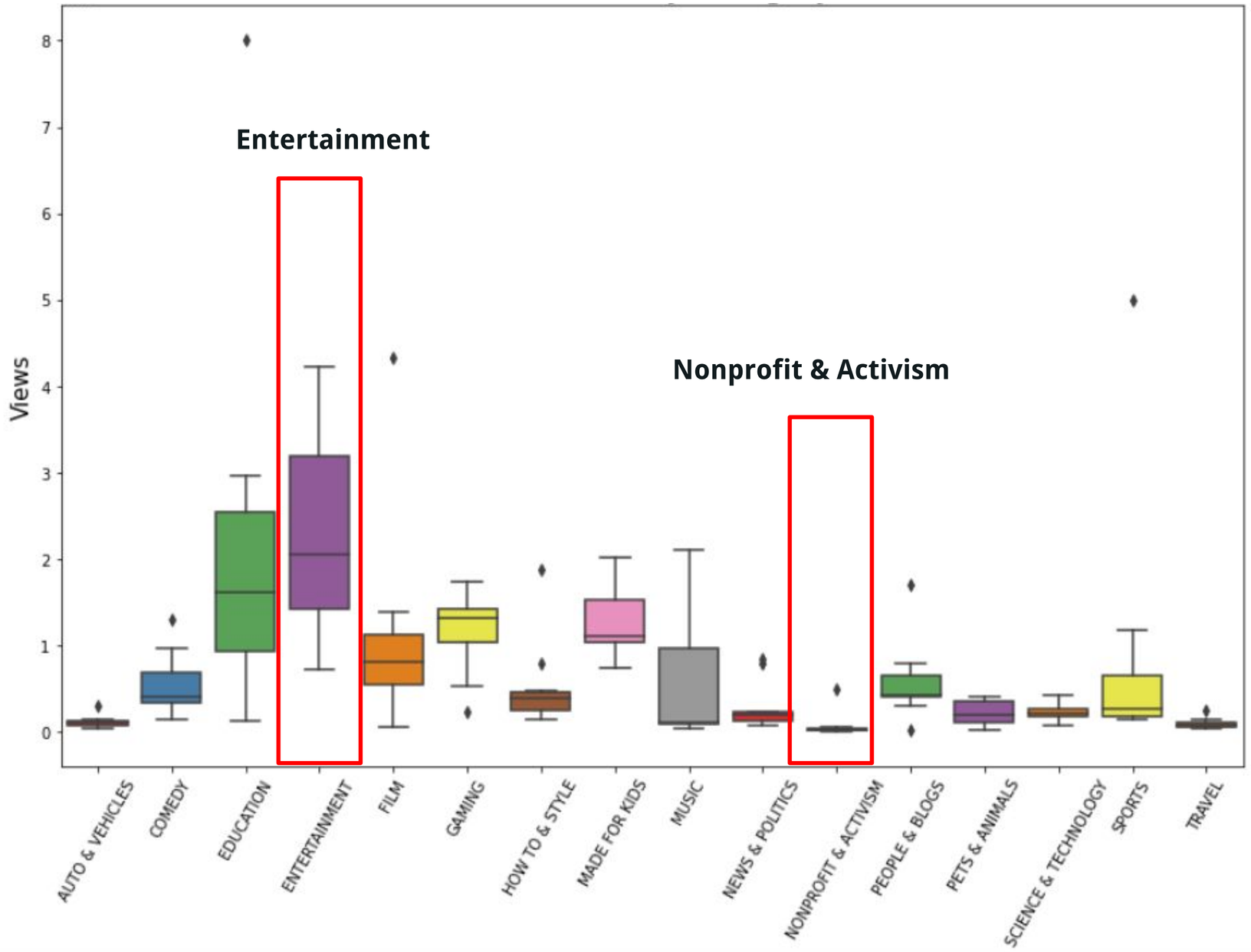


Education Music

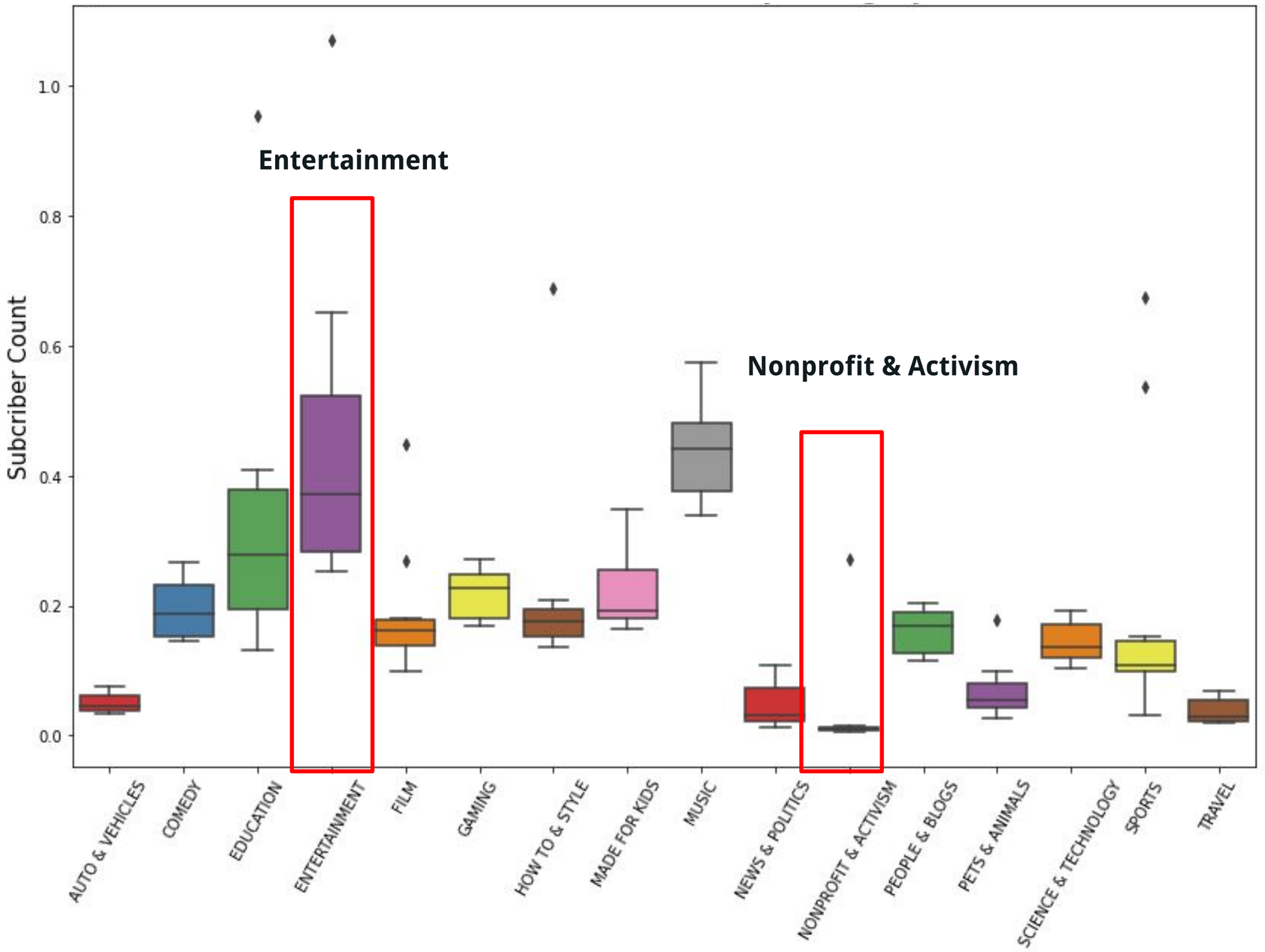
DATA EXPLORATION

4

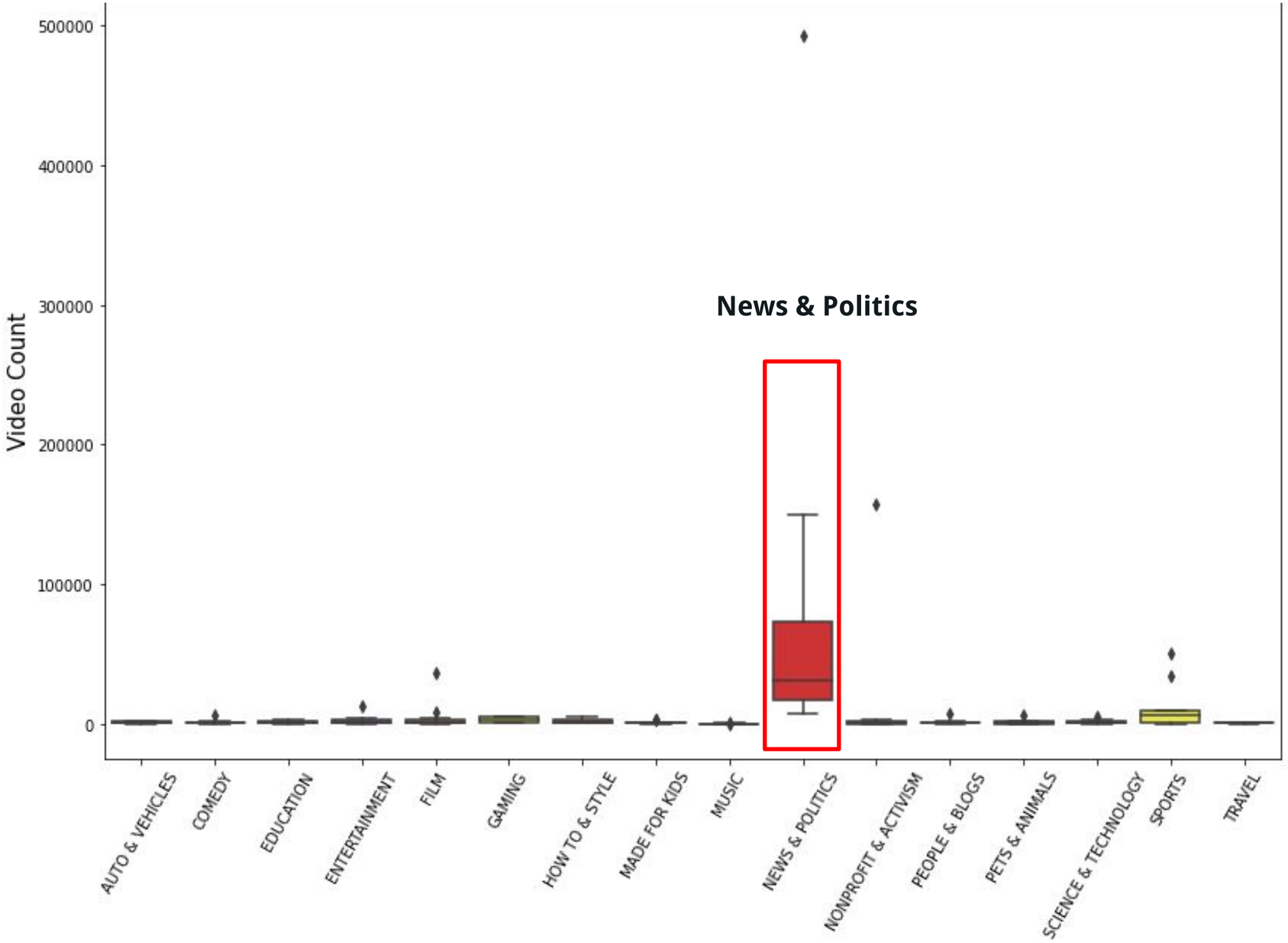
Box Plot of View Count by Category



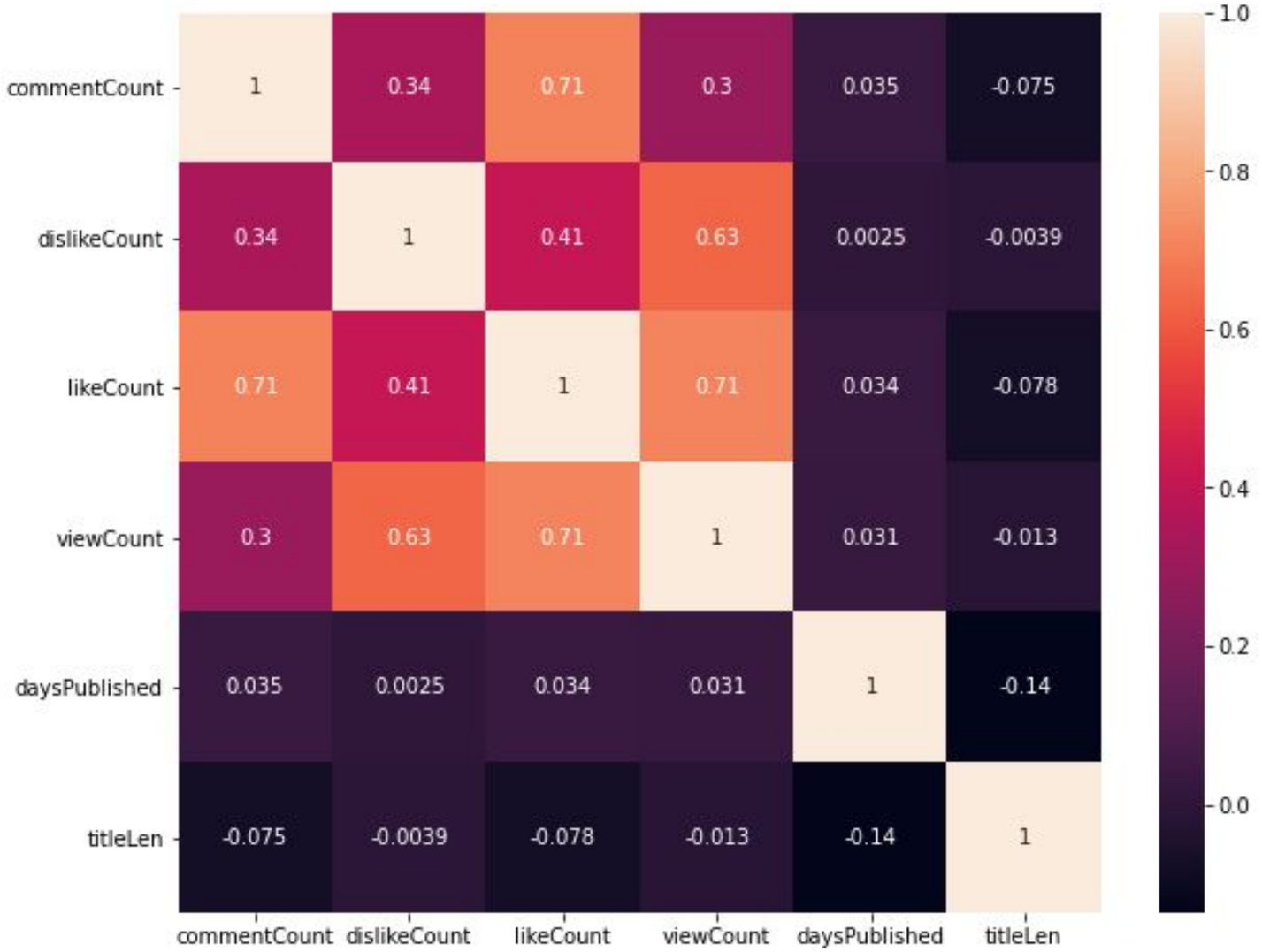
Box Plot of Subscriber Count by Category



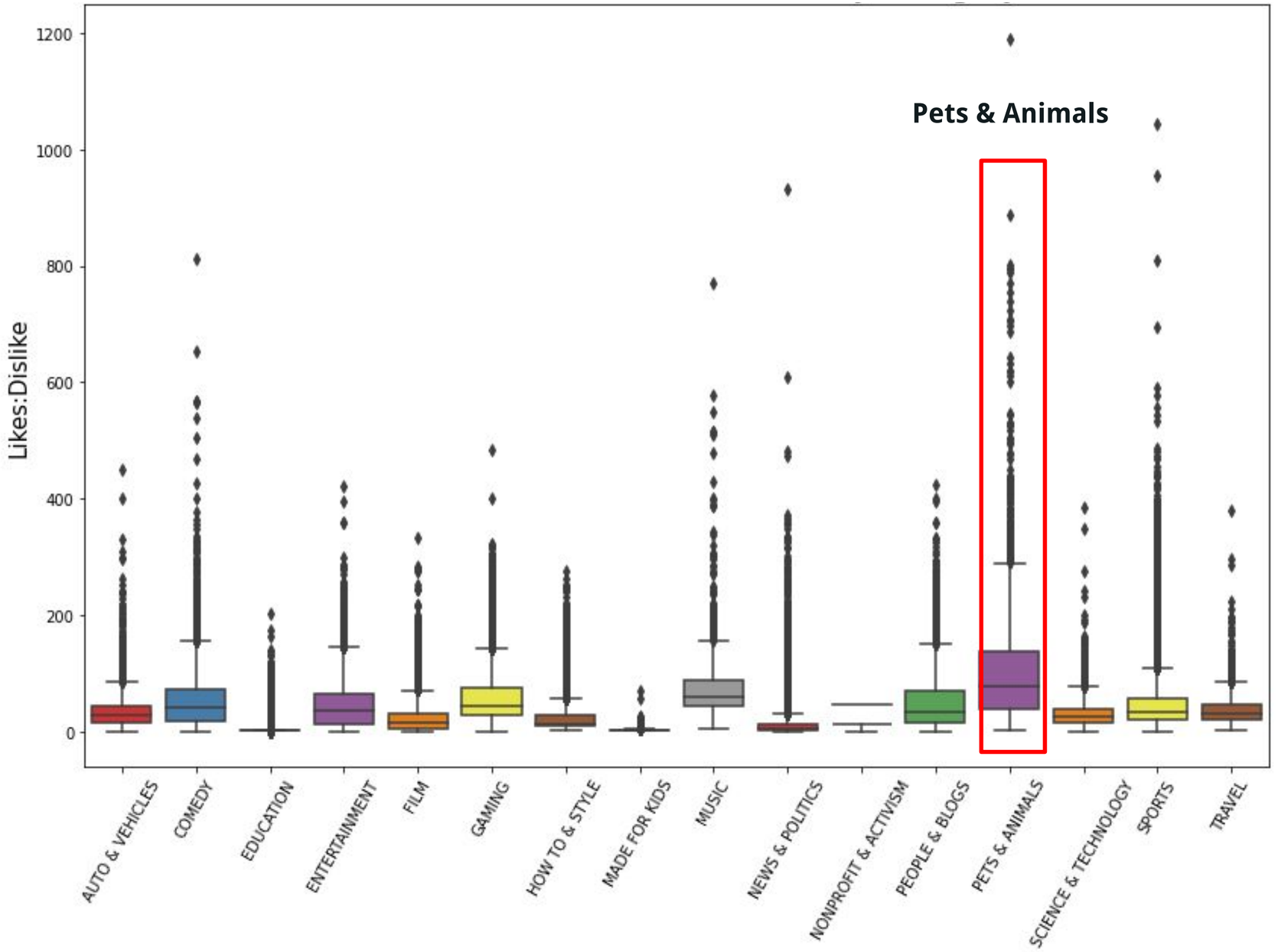
Box Plot of Video Count by Category



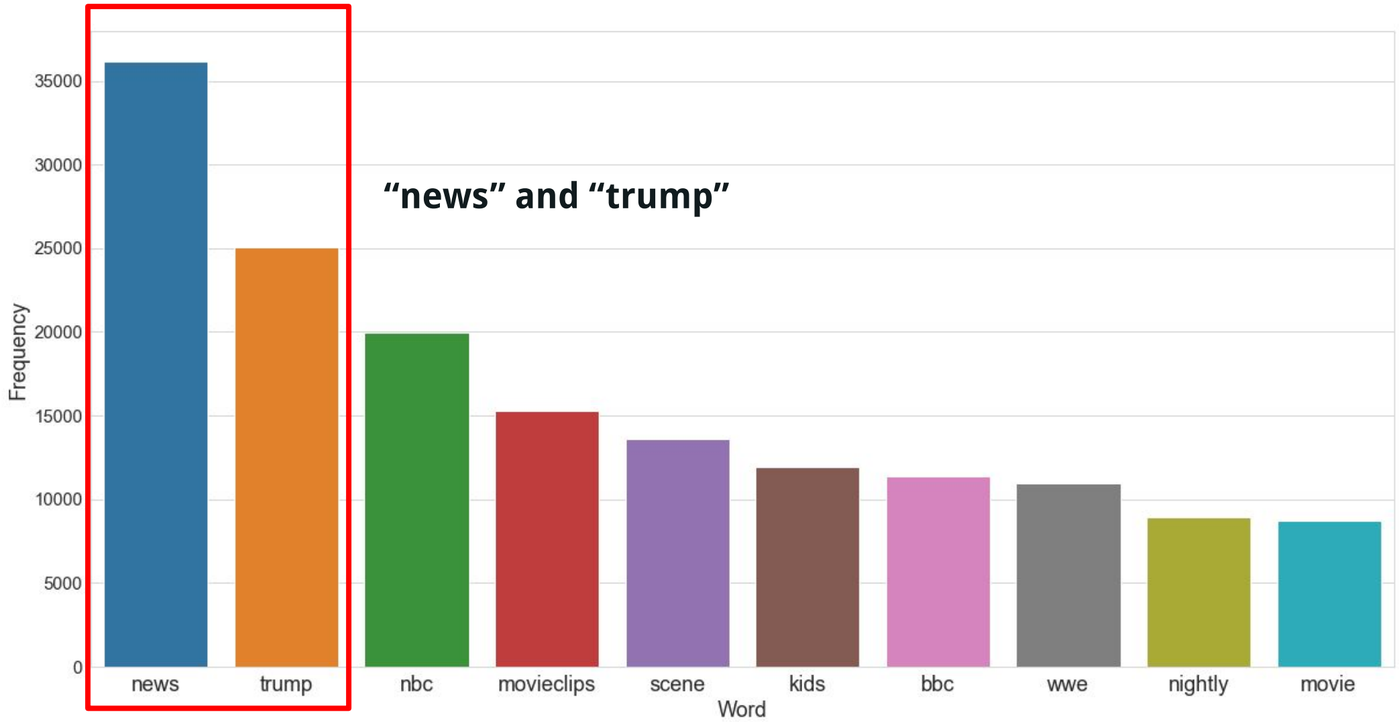
Correlation Matrix



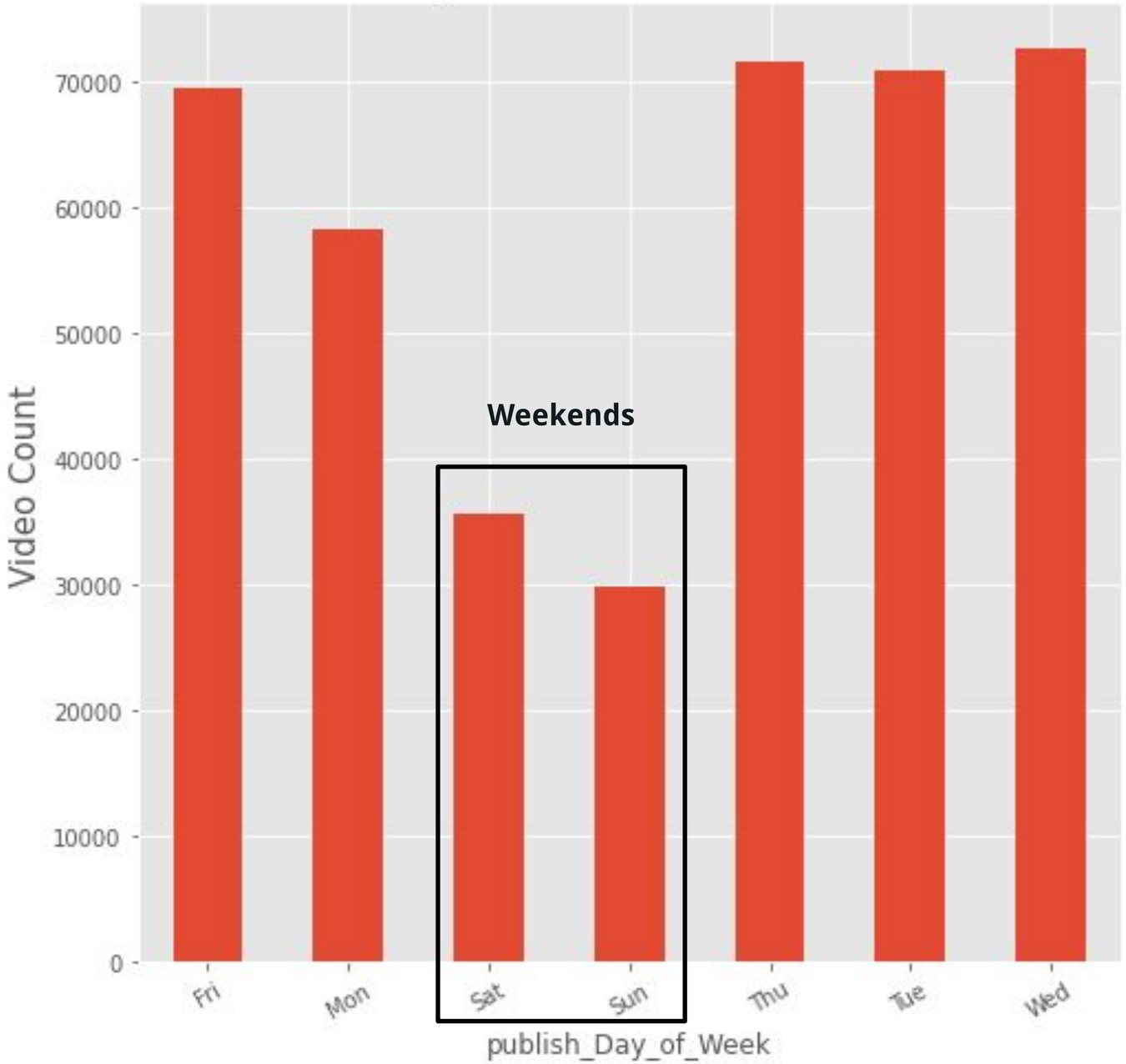
Box Plot of Like:Dislike Ratio by Category



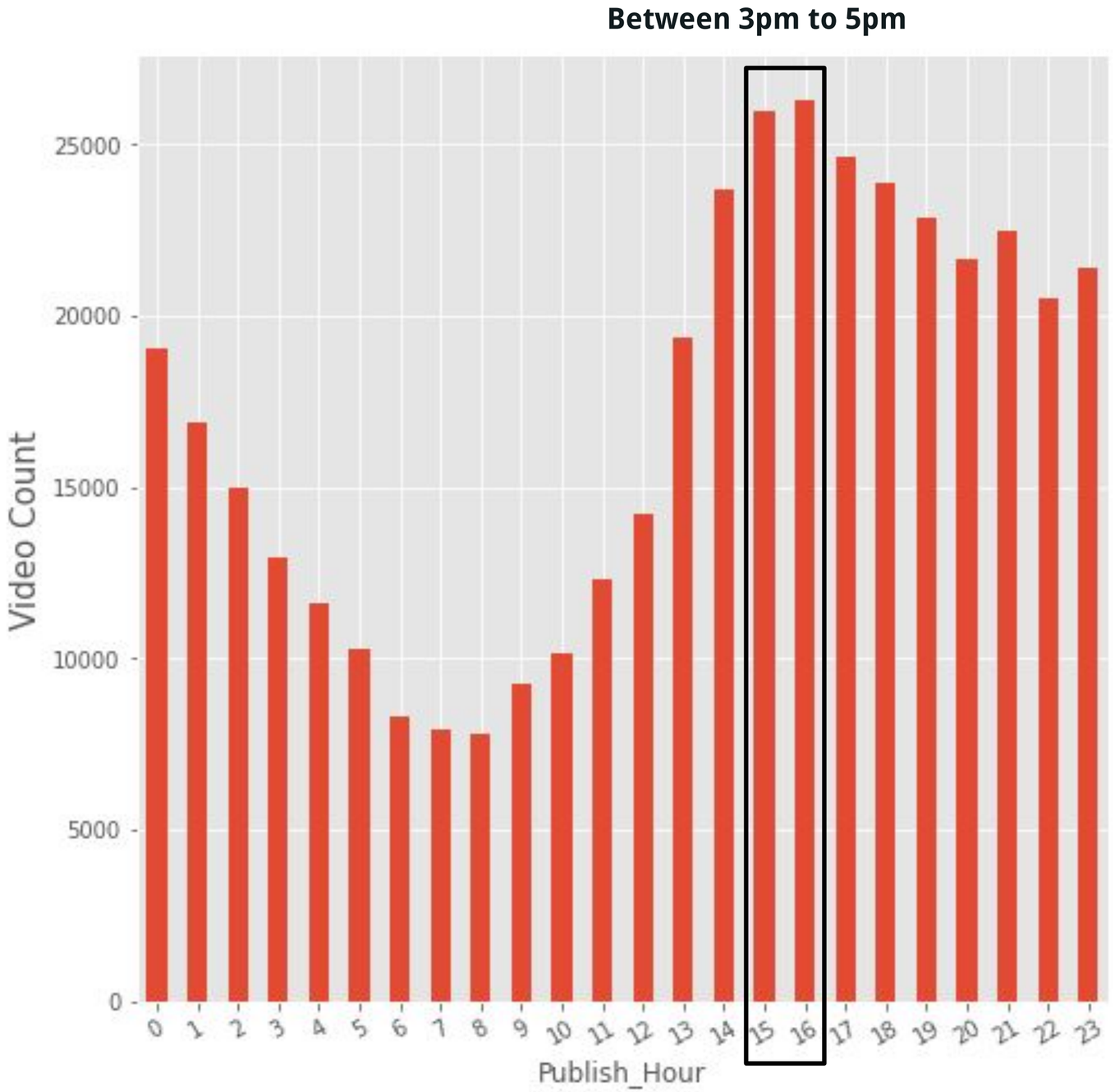
Word Cloud & Bar Graph of Word Frequencies



Bar Chart of Day of Week Published



Bar Chart of Published Time



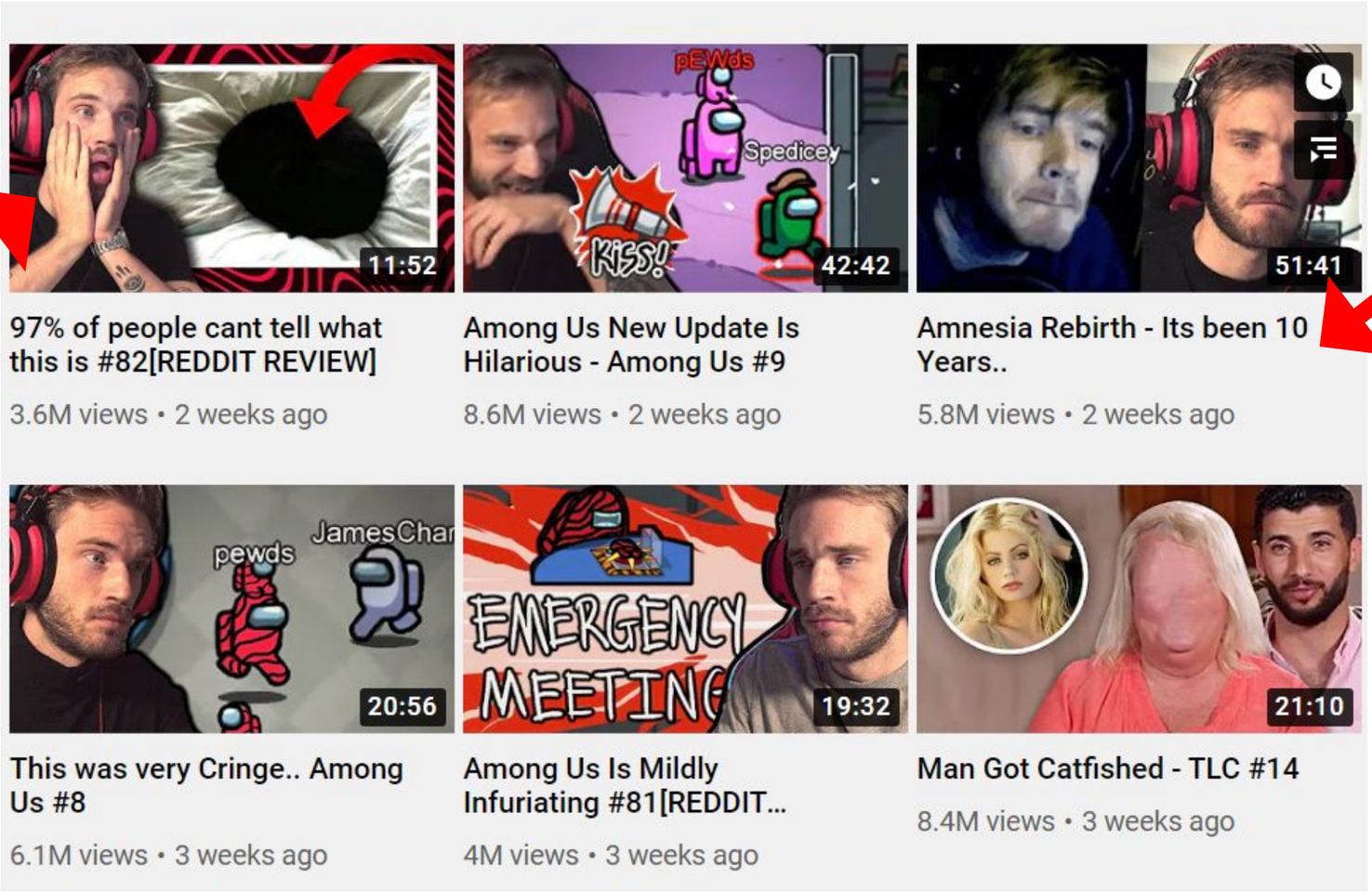
FEATURE EXTRACTION

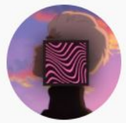
5

Components of a YouTube video

Thumbnail

Title





PewDiePie

107M subscribers

Video Title

Pranks Gone Too Far!
21M views • 4 years ago

nigahiga

There are so many great prank videos these days that prankers are being forced to top one another and some of

CC

- Long or Short in length?
- Descriptive or Simple words?

VS

Trump Supporters Upset About Fox News Calling Arizona for Biden | The Tonight Show
1M views • 1 day ago

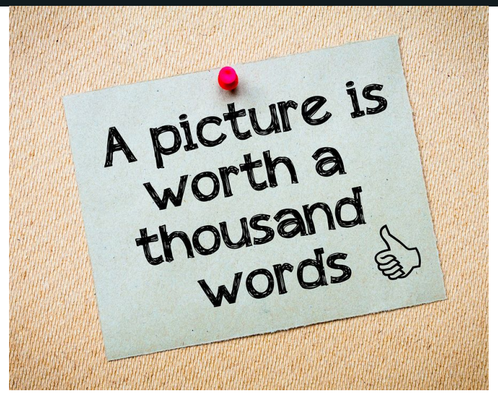
The Tonight Show Starring Jimmy Fallon

Jimmy addresses the protests Trump supporters held against Fox News after the network called Arizona as a win for Joe Biden.

New CC

Video Thumbnail

- Custom or still images taken from video?
- What types of objects to include?



vs

Feature Extraction

1

Sentiment Analysis

- On video titles
- Positive, Neutral or Negative
- Type: **categorical**

2

Object Detection

- On thumbnail images
- To find out the presence of humans
- Type: **numerical + categorical**

3

Part-of-Speech (PoS) Tagging

- On video titles
- To find out which token the title begins with
- 17 different labels
- Type: **categorical**

Feature Extraction

4

Length of Video Title

- Number of characters
- Type: **numerical**

5

Average View Count of Channel

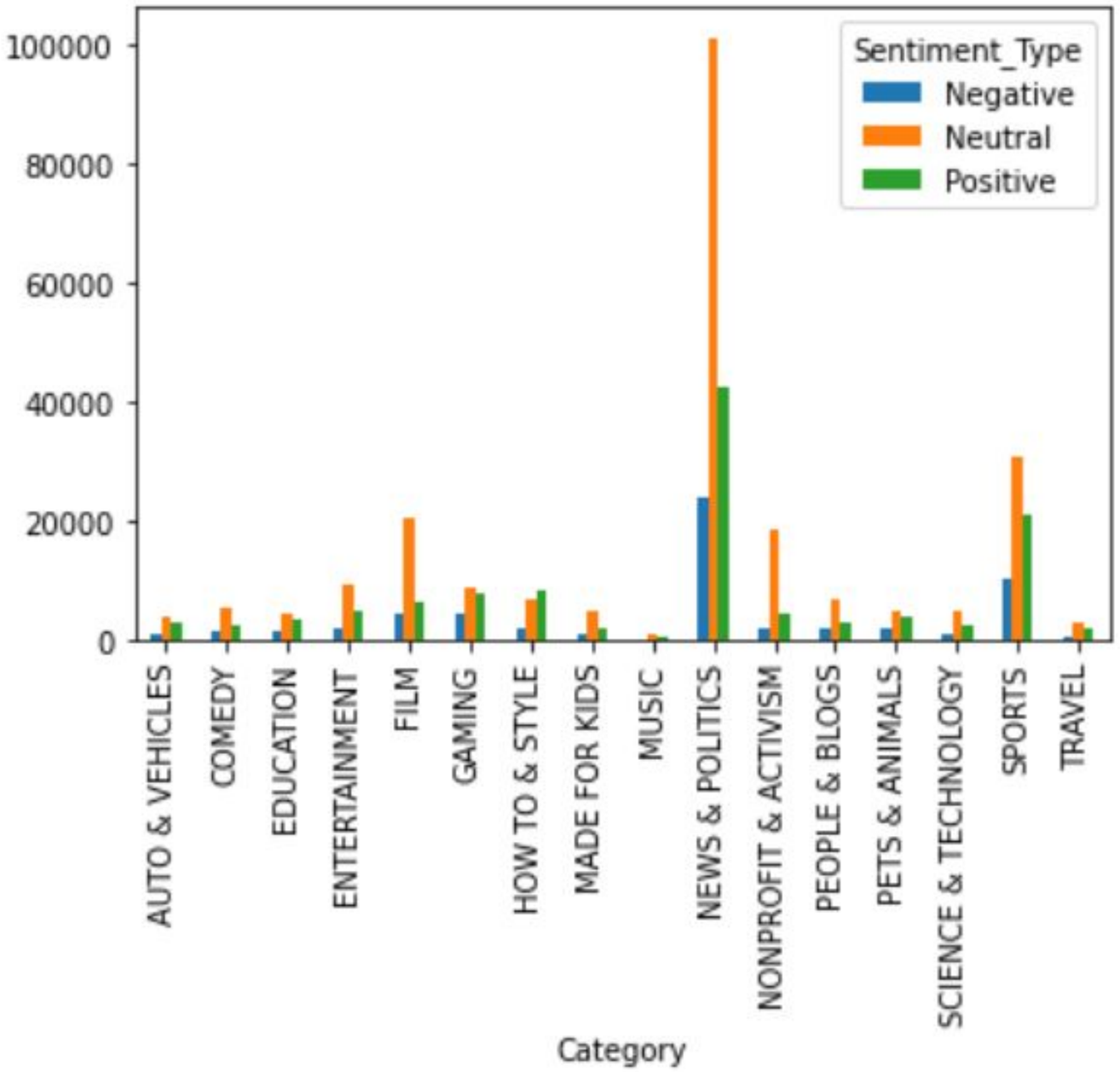
- Total view count / Number of videos
- Type: **numerical**

6

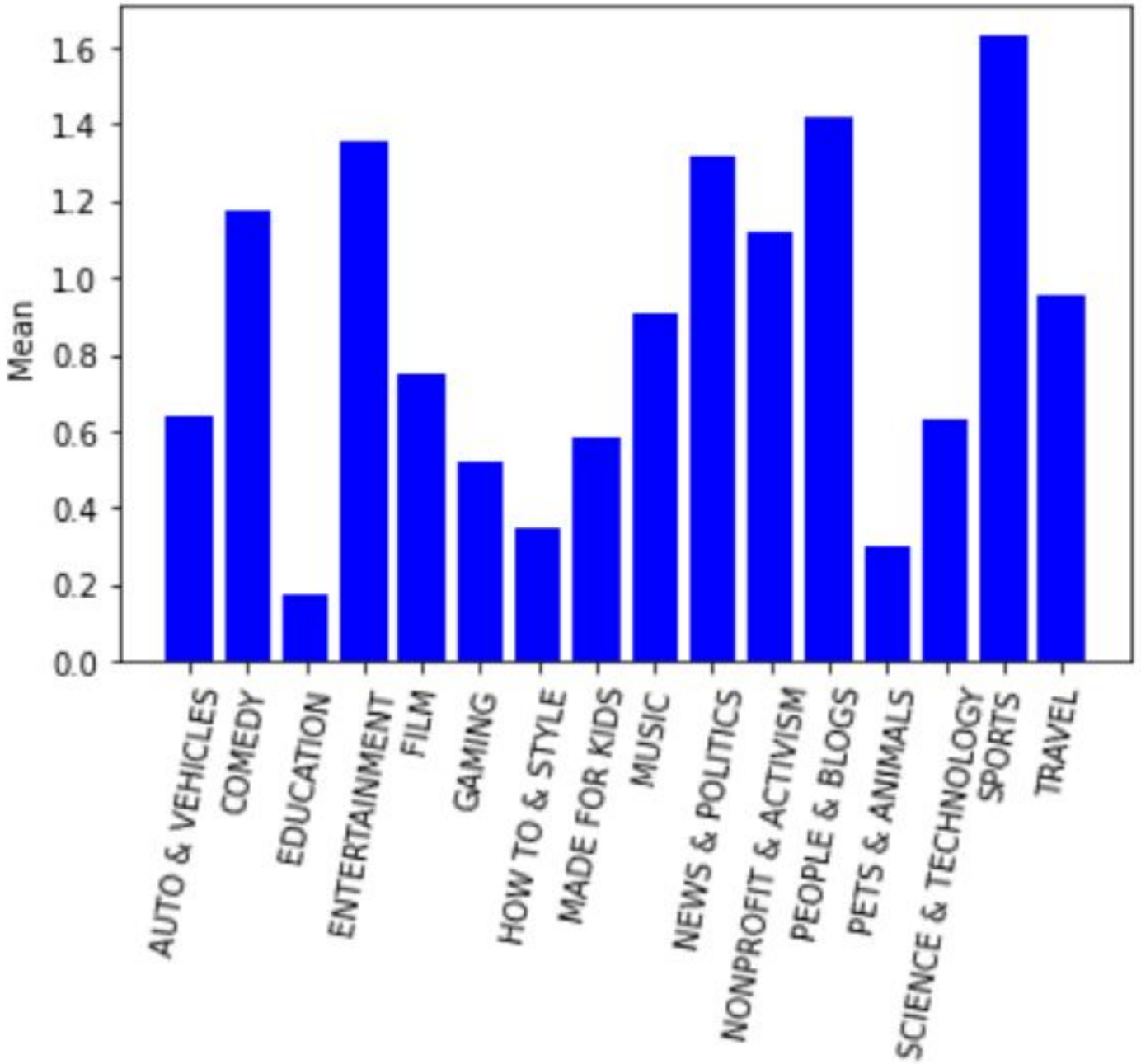
Published Timing of Videos

- Time of the day
- Day of Week
- Season
- Type: **categorical**

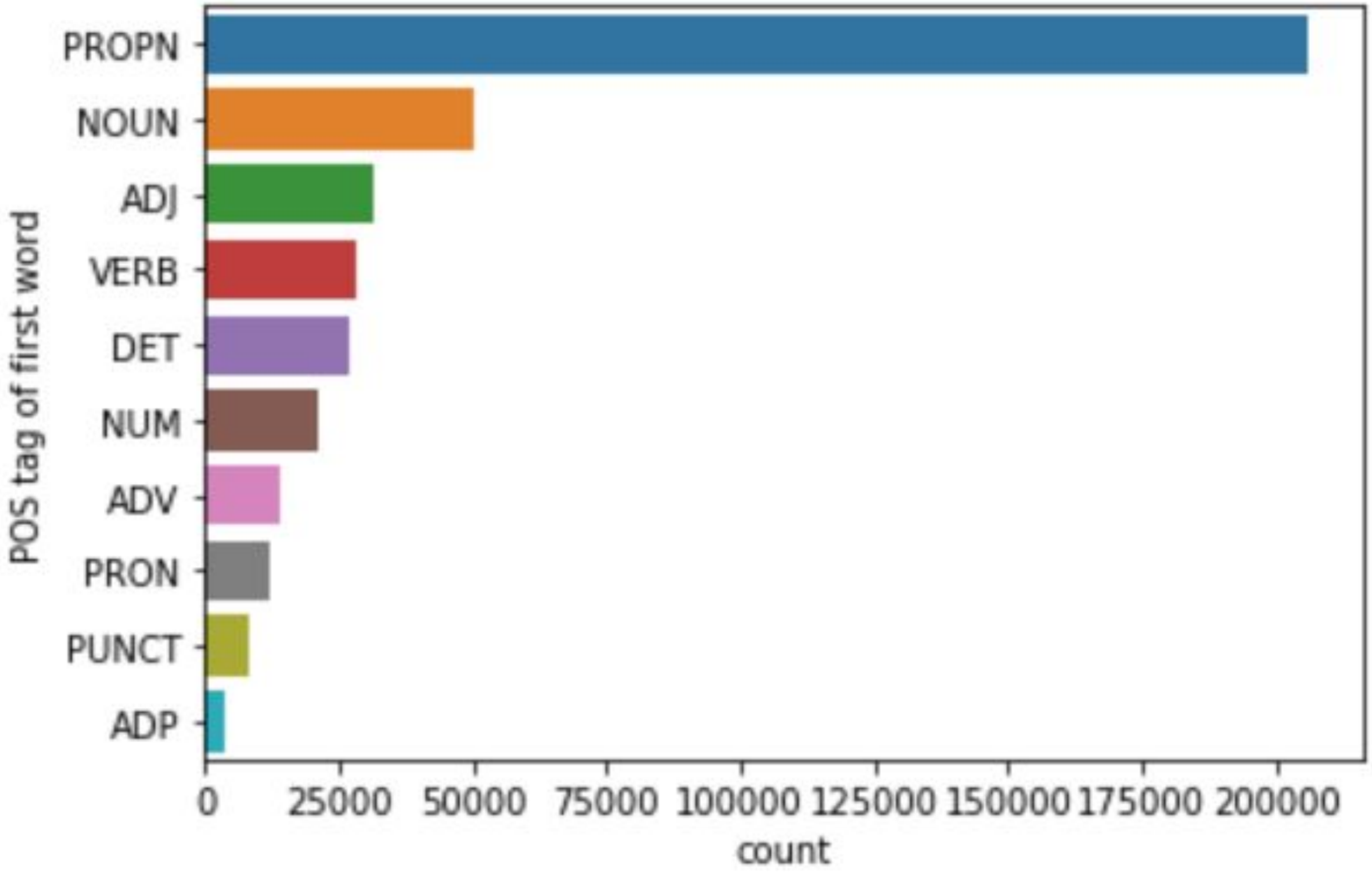
Count of Sentiment Types by Category



Average Human Count in Video Thumbnail Images by Category



Count Plot of Top 10 PoS Tags Used in the First Word of Video Title



Features

8 features

```
['videoTitle', 'sentimentType', 'first_sent_tag',
'publishedAt', 'thumbnails', 'subscriberCount',
'videoCount', 'viewCount']
```

Feature Engineering

58 features

```
['Friday', 'Monday', 'Saturday', 'Sunday', 'Thursday',
'Tuesday', 'Wednesday', 'Afternoon', 'Early Morning',
'Evening', 'Late Night', 'Morning', 'Night', 'Autumn',
'Spring', 'Summer', 'Winter', 'AUTO & VEHICLES', 'COMED
Y', 'EDUCATION', 'ENTERTAINMENT', 'FILM', 'GAMING', 'HO
W TO & STYLE', 'MADE FOR KIDS', 'MUSIC', 'NEWS & POLITI
CS', 'NONPROFIT & ACTIVISM', 'PEOPLE & BLOGS', 'PETS &
ANIMALS', 'SCIENCE & TECHNOLOGY', 'SPORTS', 'TRAVEL',
'Negative', 'Neutral', 'Positive', 'titlePROPN', 'title
DET', 'titleADJ', 'titleVERB', 'titleNOUN', 'titleNUM',
'titleAUX', 'titlePUNCT', 'titleADV', 'titleADP', 'titl
eX', 'titlePRON', 'titleSYM', 'titleSCONJ', 'titleINT
J', 'titlePART', 'titleCCONJ', 'humanPresence', 'titleL
en', 'subscriberCount', 'avgViewCount', 'humanCount']
```

4 Numerical

54 Categorical

FEATURE SELECTION

6

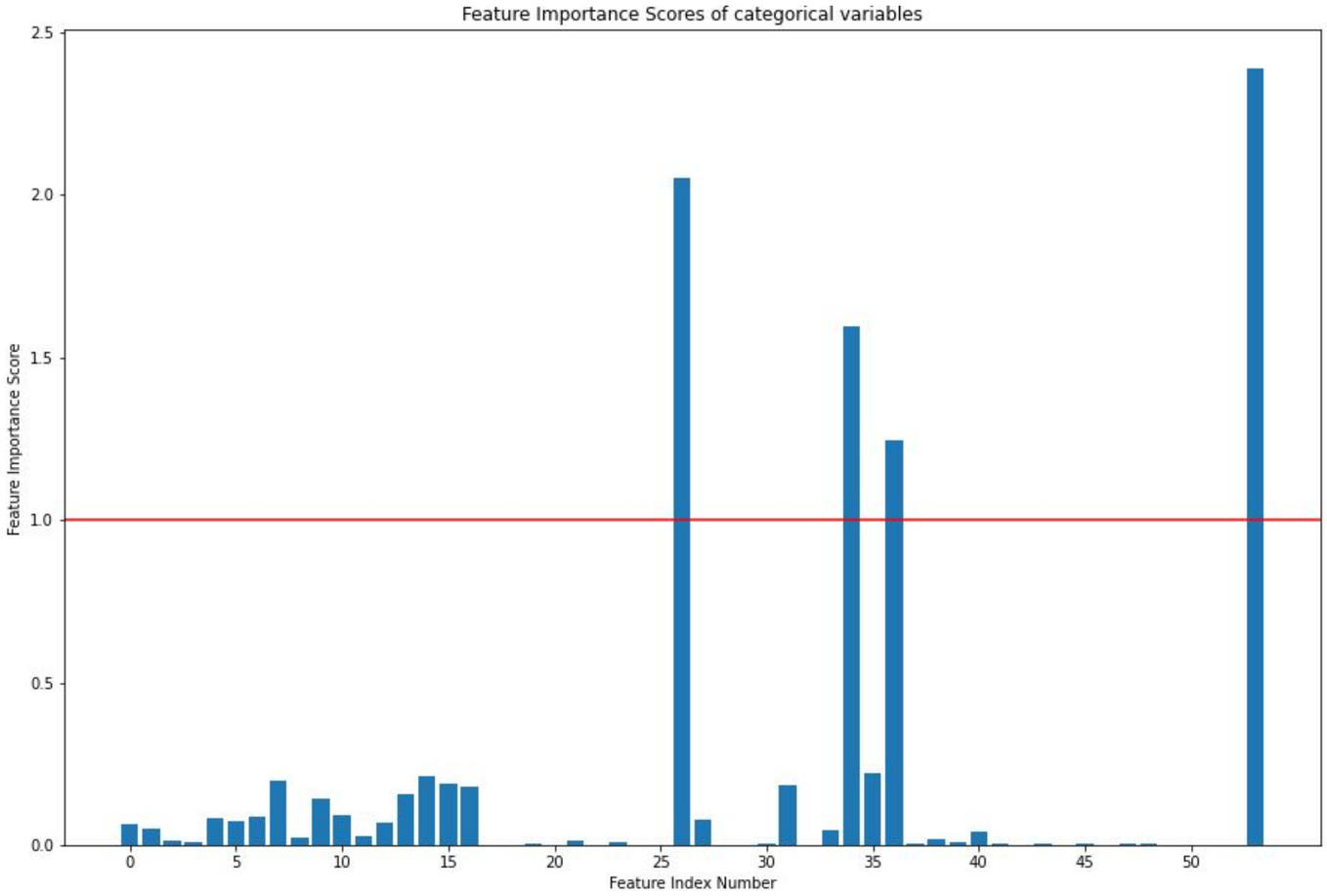
Feature Selection

- **Output Variable**: Video View Count
- **Input Variables**:

Categorical Input Variables	**Numerical** Input Variables	**ALL** Input Variables
ANOVA Test	Pearson's Correlation Test	ANOVA Test

1) Categorical Features - ANOVA Test

Results



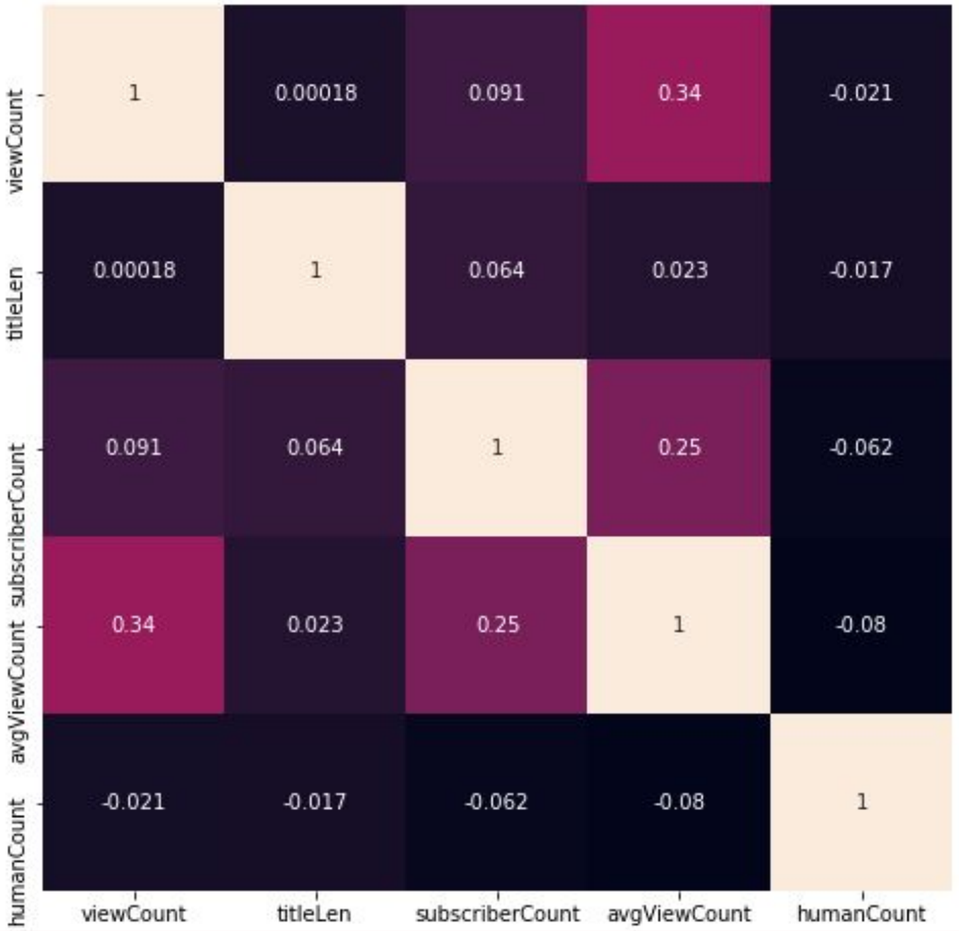
Selected Features

Index	Feature	Score
26	HOW TO & STYLE	2.051160
34	SPORTS	1.596341
36	Negative	1.246622
53	titleINTJ	2.386547

** Features with a score below 1 were dropped

2) Numerical Features - Pearson's Correlation

Results



** None of the numerical input variables are strongly correlated to one another, or to the output variable 'viewCount'

Selected Features

(ALL)

➔ subscriberCount

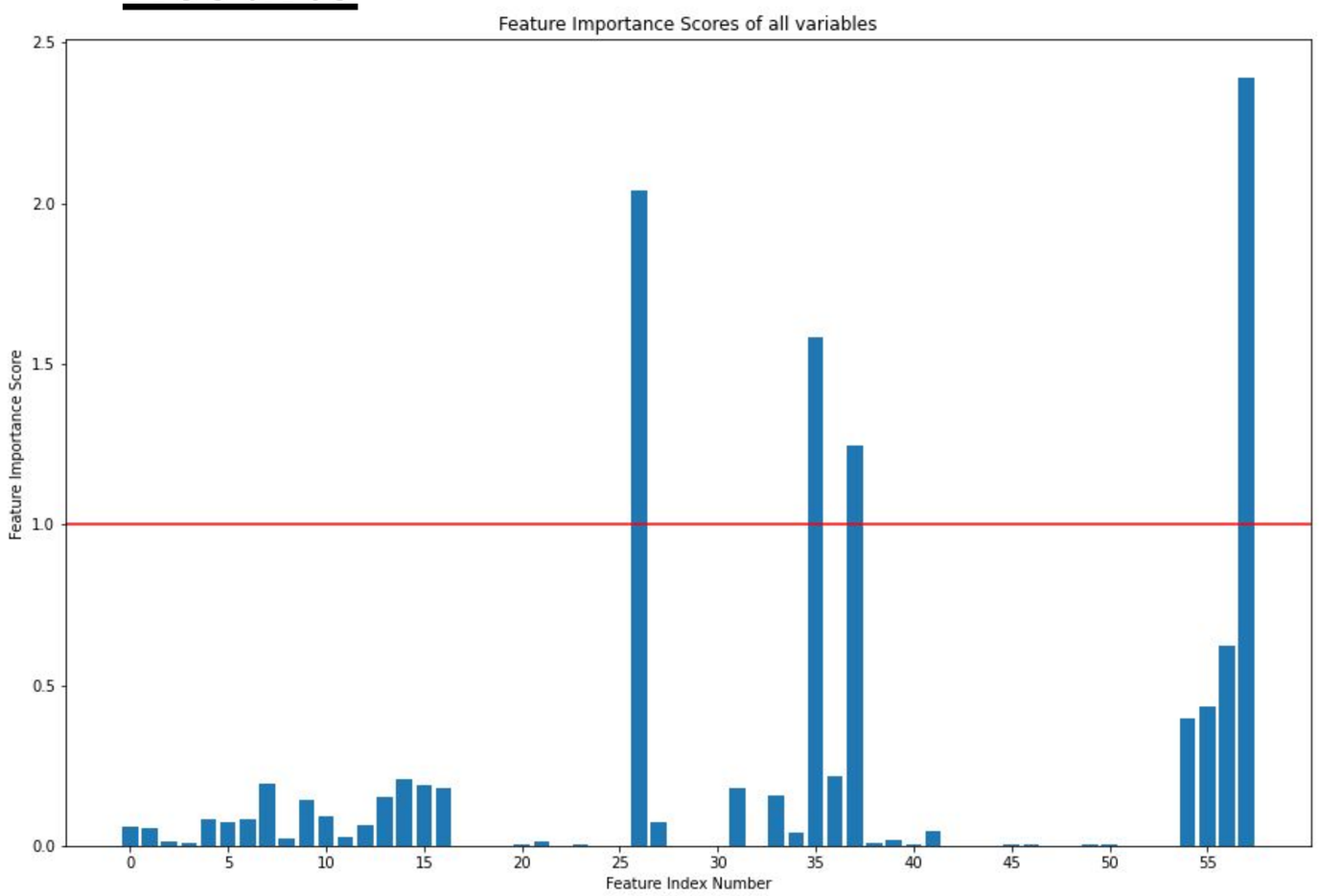
➔ avgViewCount

➔ humanCount

➔ titleLen

3) All Features - ANOVA Test

Results

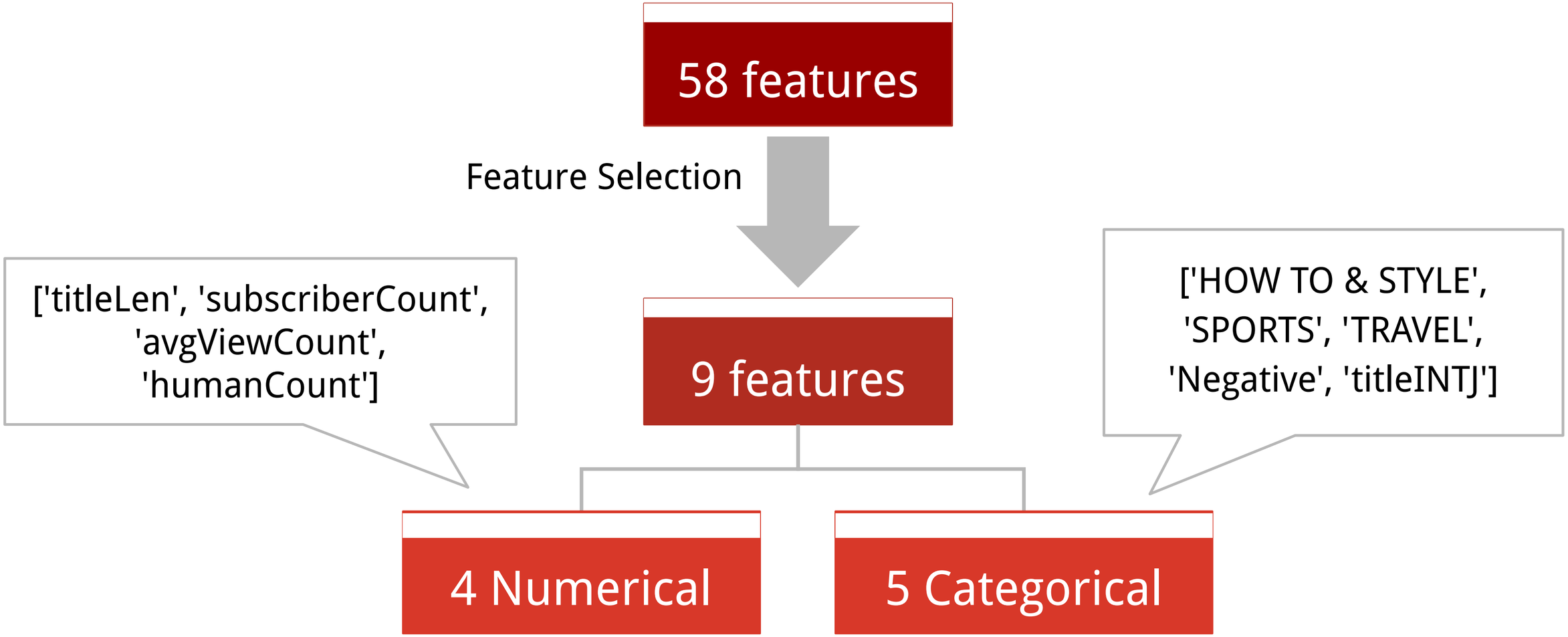


Selected Features

Index	Feature	Score
26	HOW TO & STYLE	2.041418
35	TRAVEL	1.582291
37	Negative	1.244789
57	subscriberCount	2.389420

** Features with a score below 1 were dropped

Selected Features



MODEL BUILDING & EVALUATION

7

Models Used

1 Polynomial Regression

- Fits a wide range of curvature

2 Random Forest

- Reduces overfitting of a model
- Able to handle high-dimensional data

3 Gradient Boosting

- Supports different loss functions

4 Artificial Neural Network (ANN)

- Ability to detect complex non linear relationships
- Suitable for tabular data

Tree-based model

Metrics Used for Evaluation

- Root Mean Squared Error (RMSE)
- Mean Squared Error (MSE)
- Mean Absolute Error (MAE)

The **lower** the metrics are, the **better performance** of the model.

Model Results

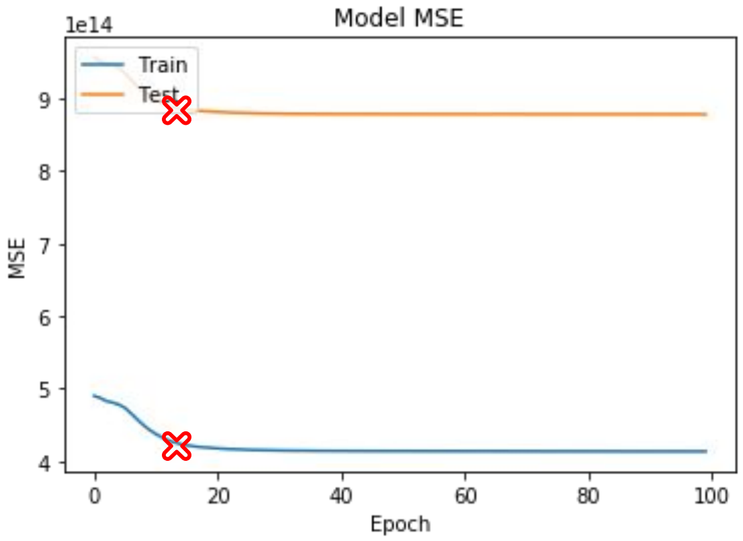
Model	RMSE	MSE	MAE
Polynomial Regression	24,486,700	599,598,500,178,391	2,410,707
Random Forest	29,491,373	869,741,069,564,456	2,369,201
Gradient Boosting	29,839,713	890,408,446,899,621	2,363,009
Artificial Neural Network	**20,330,611**	**413,333,730,000,000**	**2,338,276**

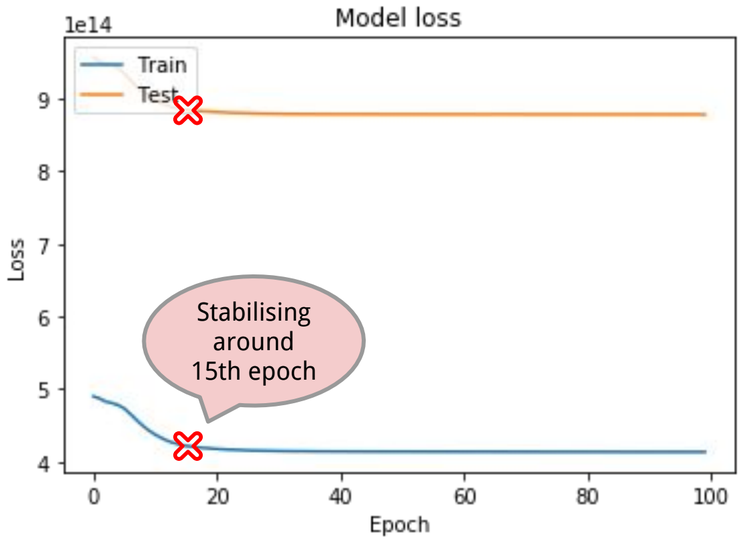
Best Model – Artificial Neural Network (ANN)

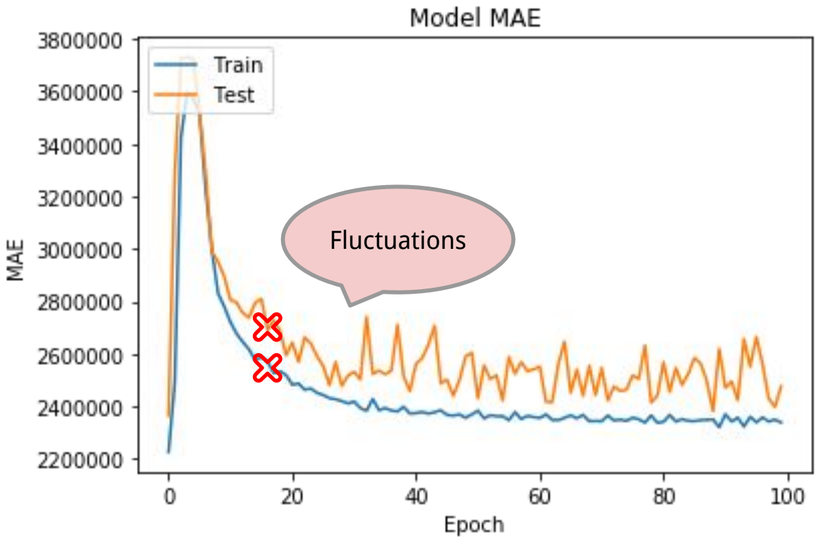
- Compiled with **Mean Squared Error (MSE)** and used **ADAM optimizer**
- Run with **100 epochs** and **batch size of 1,000**
  - Input layer: 1 layer with 128 neurons and relu activation
  - Hidden layer: 3 layers with 32 neurons each and relu activation
  - Output layer: 1 layer with 1 neuron and linear activation

Best Model – Artificial Neural Network (ANN)

Training History



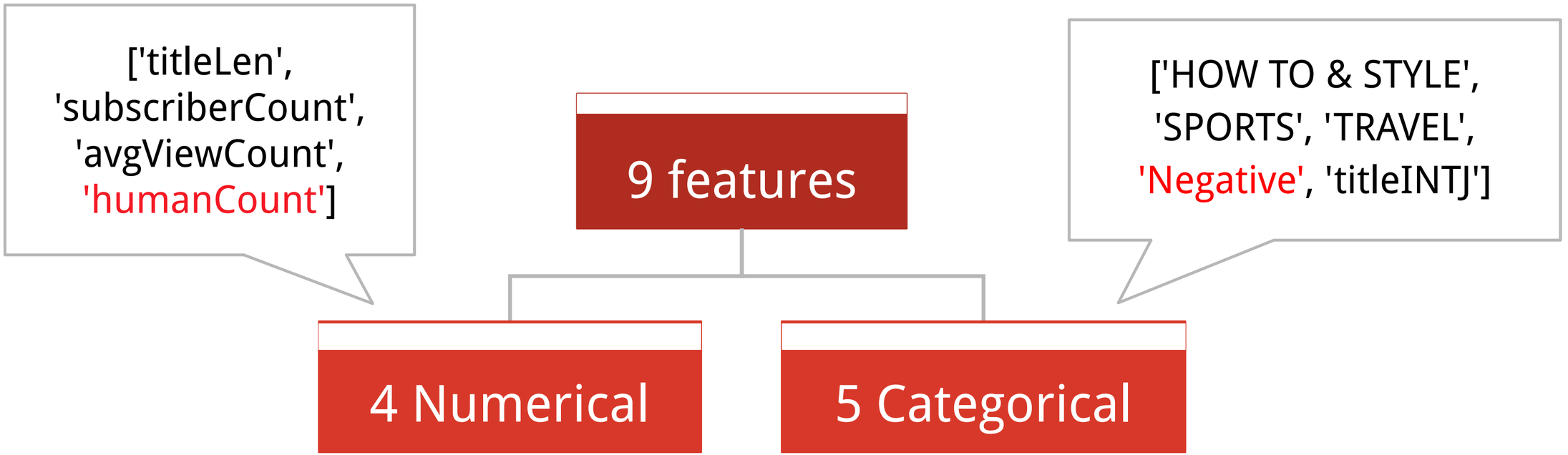




CONCLUSION

8

Selected Features (Recap)



Hypothesis

A. “A **positive sentiment title** in a YouTube video will affect views”

B. “Having **humans in the thumbnail images** of a YouTube video will affect views”

Reject

Fail to Reject

THANK YOU!

Stay Connected

NGOC **LINH CHI** NGUYEN
nguyenngoclinhchi@u.nus.edu

CHUA **KAI BING**
kai_bing.chua@u.nus.edu

TAN **ZEN WEI**
zenwei.tan@u.nus.edu

GOH **JIA YI**
gohjiayi@u.nus.edu