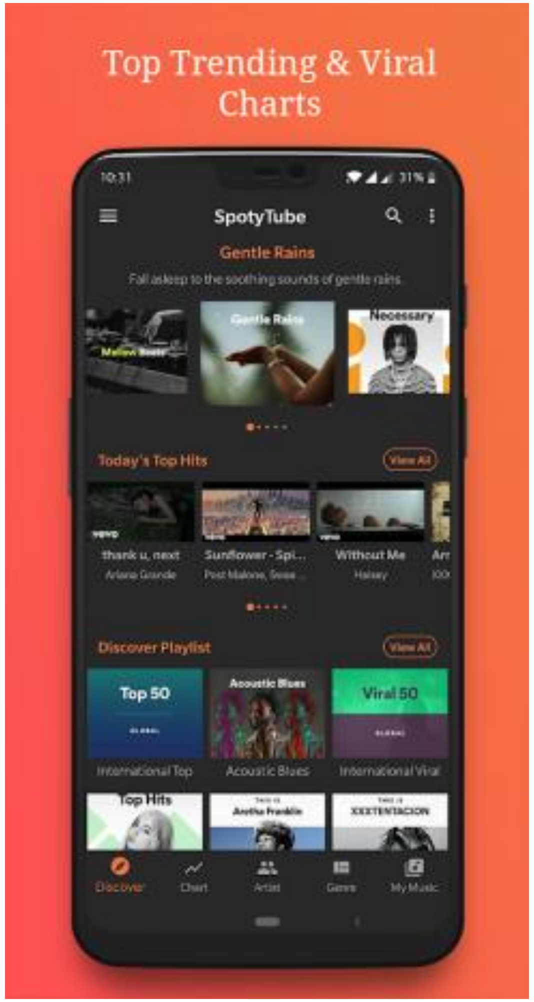
Trending YouTube Video Exploration

Yuzhe Huang
Dec. 2019

Outline

- Introduction
- Data Description
  - Dataset
  - Data processing
- Exploratory Data Analysis
  - View patterns : e.g. Top 5 most viewed categories…
  - User participation
  - Trending Lifecycle
  - Category selection: Evaluation matrix
- Modelling
  - Deep dive to the algorithm of trending charts: Multilevel Model
- Future Directions



Key Takeaways

- **View Pattern**
  Entertainment, music, gaming and auto gain most popularity in 10 countries.

- **User Participation**
  The **average voting rate, 3%** of ten countries is **10 times** of the **average comment rate,0.3%.**

- **Trending Lifecycle**
  The time interval of both going-viral and keeping trending ranges from 1 day to 2 weeks.

- **Category Selection**
  Entertainment, music, film and animation have the most business potential based on overall performance.

- **Deep Dive to the Algorithm of Trending Videos**
  Views, voting rate and comment rate will affect the going-viral days, and have different influence mechanism within different countries, especially U.S. and UK.

Introduction

YouTube is one of the largest video hosting websites in the world, which has profound influences on the society in all aspects. Therefore, analyzing of YouTube's dataset become significant for **advertisers and investors** to analyze the social trends and make a prediction for future strategic business planning.

Therefore, the key objectives of the report are as follows:

- Find the view patterns of YouTube videos by region;
- Select the category that have relatively higher business potential by constructing evaluation matrix;
- Explore the algorithm of the YouTube trending charts.

Data Description

Datasets

- The dataset includes daily trending YouTube videos of **10** countries: **United States, United Kingdom, Germany, Canada, France, Japan, Korea, Mexico, Russia and India.**
- Trending time interval: **2017.11-2018.6.**
- The variables used contain: **video id, trending date, category id, category name, publish time, views, likes, dislikes, comment count.**
- Source: https://www.kaggle.com/datasnaek/youtube-new#header

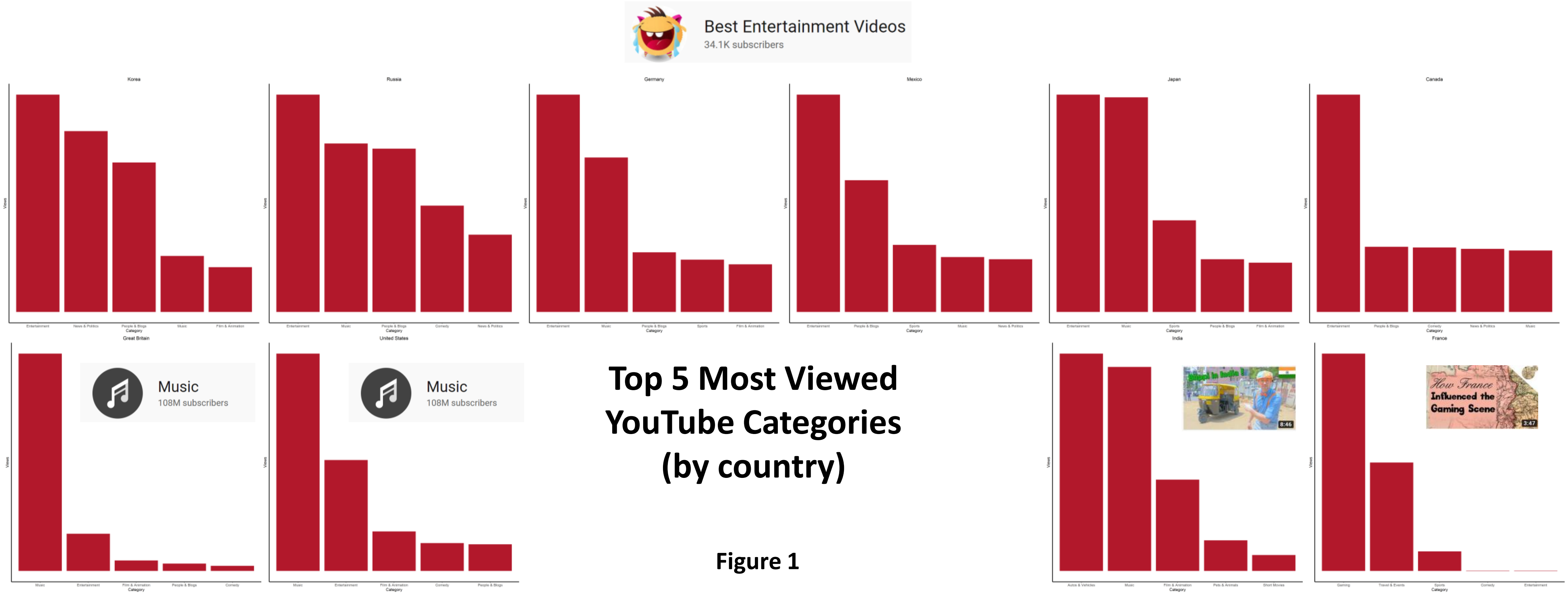
Data processing

- Drop null values and corrected time format;
- Pair the category ids with their names for further analysis.

Exploratory Data Analysis

View Pattern

- Entertainment videos are most viewed in six out of ten countries (in the first row of Figure 1), followed by music videos dominated both in U.S. and UK.
- Indian viewers like auto and vehicle videos best, while French prefer game streaming.



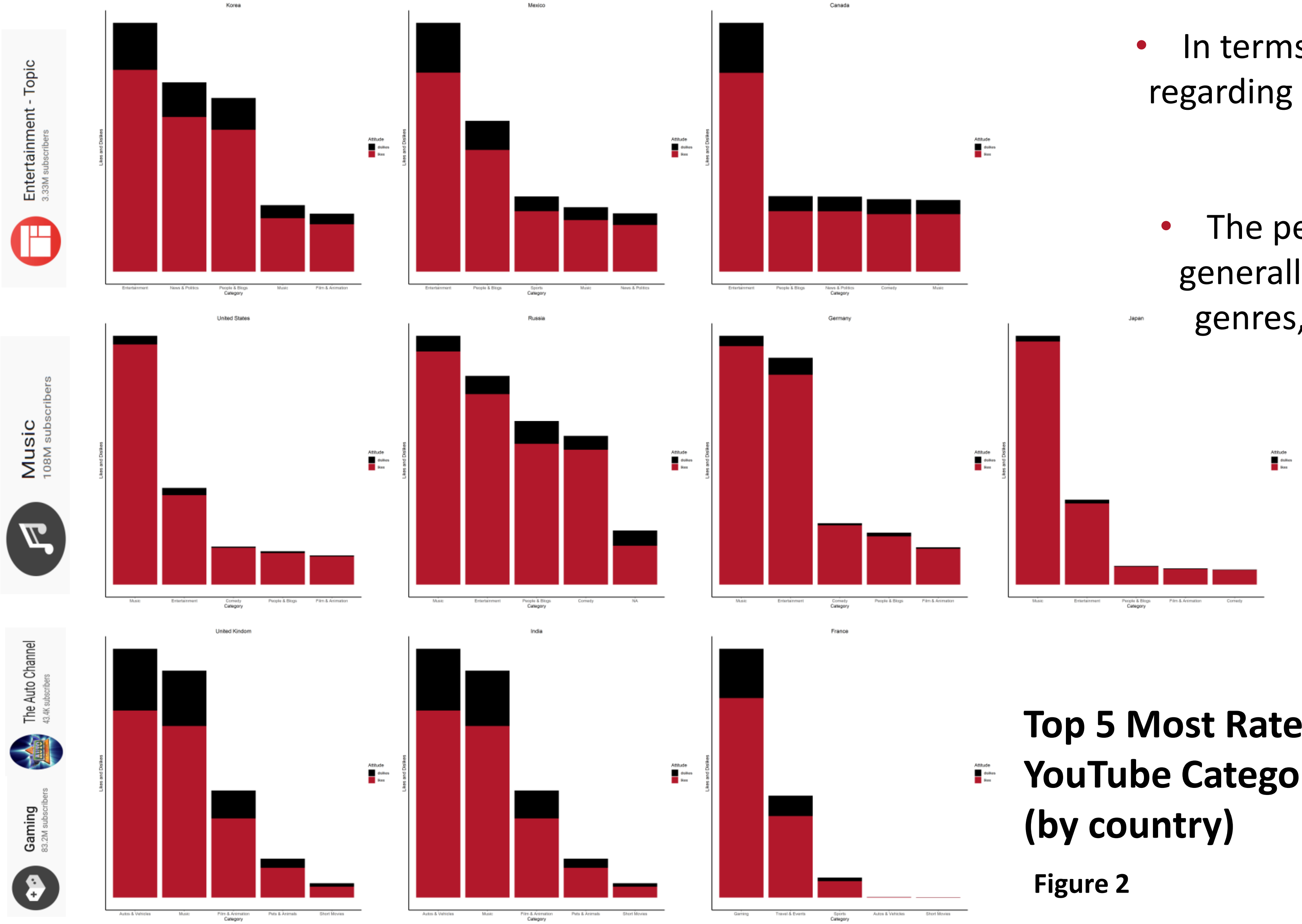
Top 5 Most Viewed YouTube Categories (by country)

Figure 1

Exploratory Data Analysis

View Pattern

Top 1 Category



- In terms of likes and dislikes, similarly, videos regarding to **entertainment, music, gaming and auto** gain most popularity.
- The percentage of dislikes in music videos is generally lower than that in other three Top 1 genres, which indicates that music videos are more acceptable than others and thus **a "safe" choice for advertiser.**

Nearly 200 years ago,
Henry Wadsworth Longfellow asserted
"Music is the universal Language of mankind."

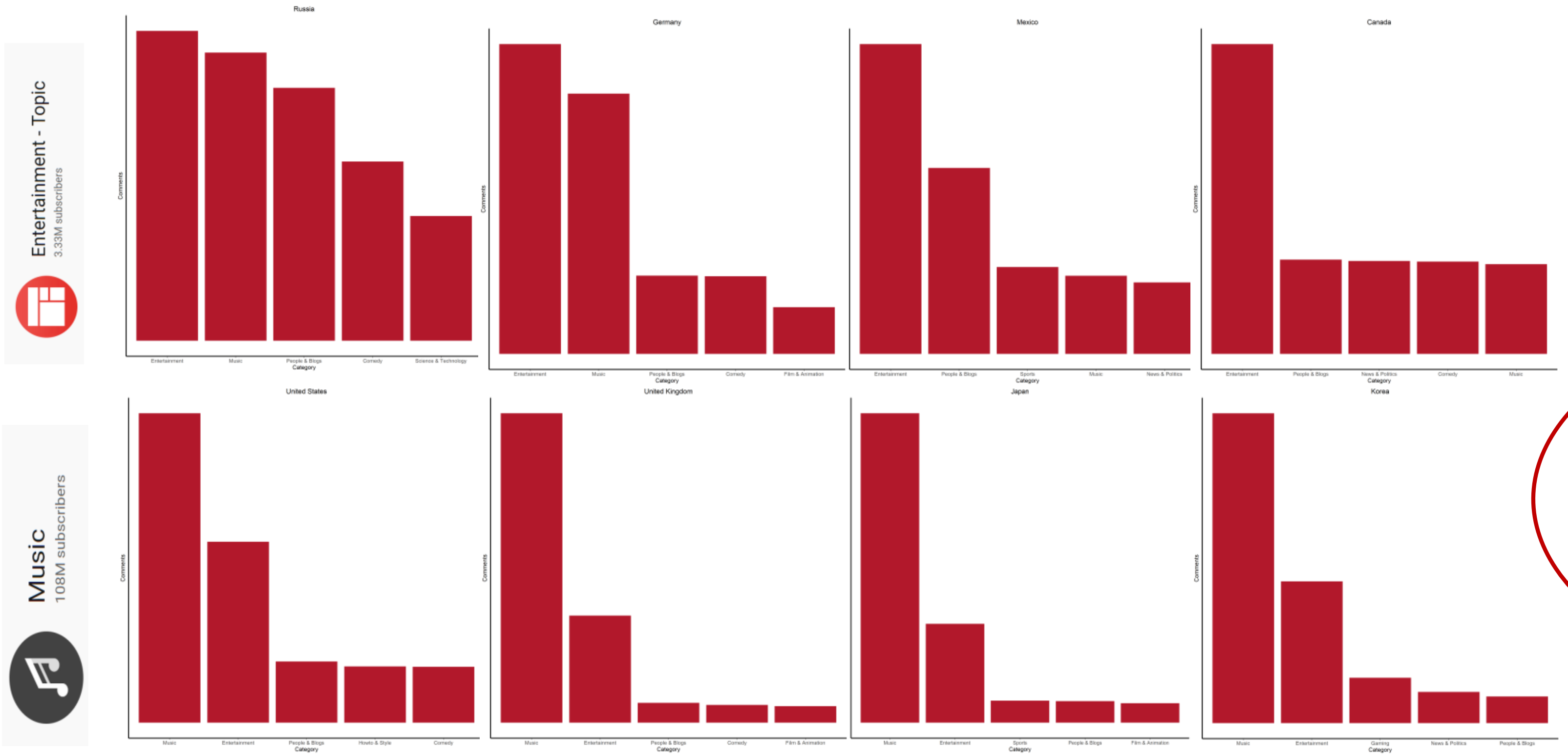
Top 5 Most Rated YouTube Categories (by country)

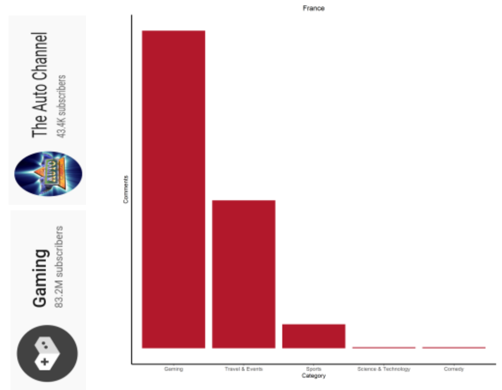
Figure 2

Exploratory Data Analysis

View Pattern

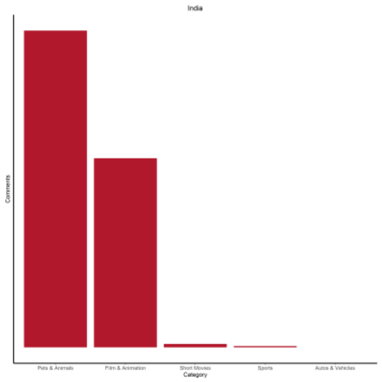
Top 1 Category





Top 5
Most Commented
YouTube Categories
(by country)

Figure 3

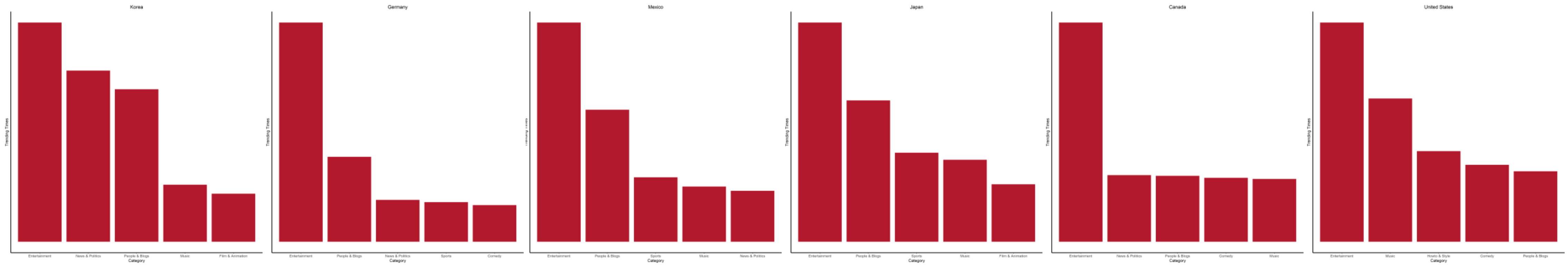


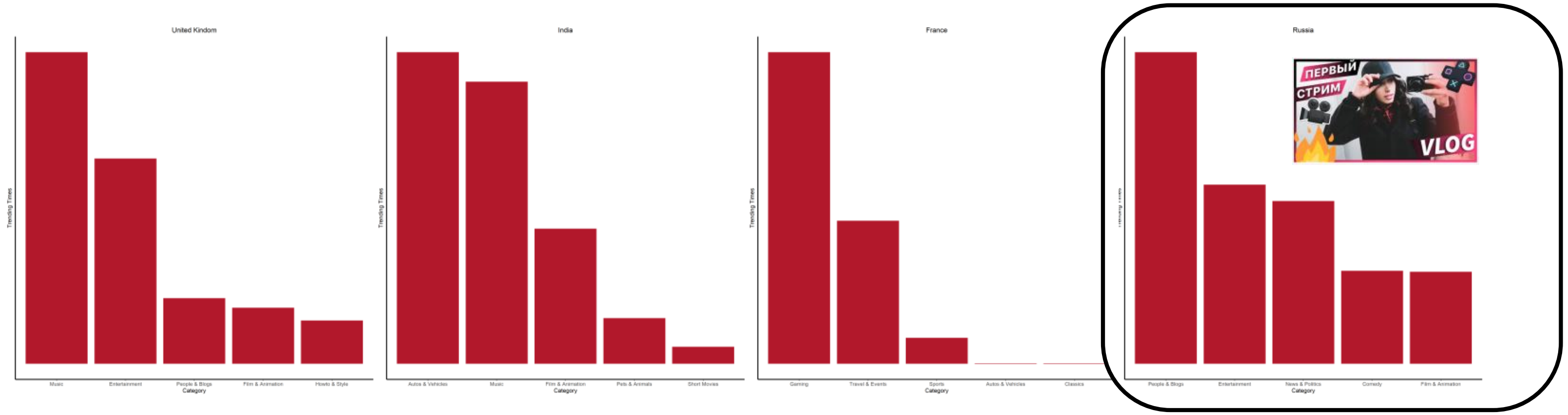
Entertainment, music, gaming and auto videos' influence continues…

Exploratory Data Analysis

View Pattern

- Finally, according to the times that videos went on charts, the four categories still dominate, which indicates that the views, likes, dislikes and number of comments are correlated with whether a video goes on chart.
- Besides, people and blog videos gain ground in Russia.





Top 5 Most Popular YouTube Categories (by country)

Figure 4

Exploratory Data Analysis

User Participation

- User participation is broke down into two parts:
  - showing their attitude to the videos, i.e. clicking “likes and dislikes”;
  - leaving comments on videos.
- The **average voting rate, 3%** of ten countries is **10 times** of the **average comment rate,0.3%** (red lines in Figure 5) :
  - Voting rate of a video=(likes+dislikes)/views X 100%;
  - Comment rate of a video=(# of comments)/views X 100%.
- The British show the least willingness to rate and comment on videos, while Russians are enthusiastic about doing so, which could be partly explained by culture difference.

Gentle, Mild

Enthusiastic!

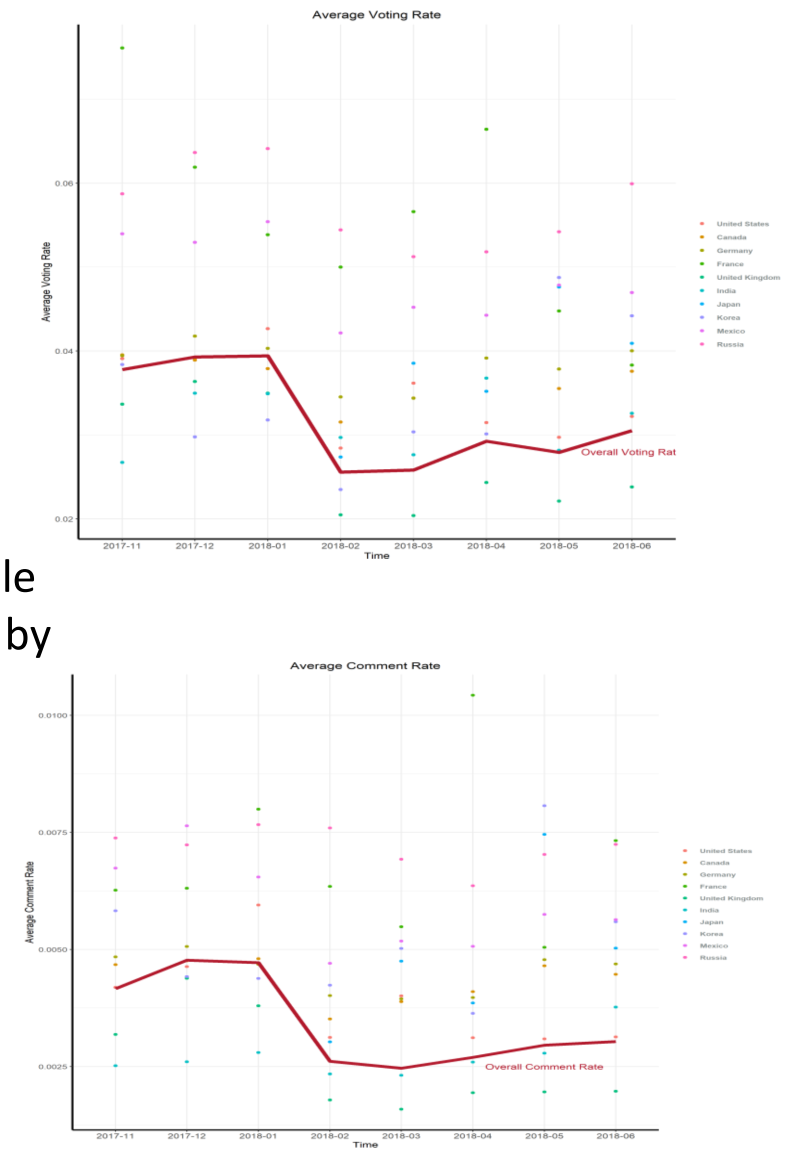


Figure 5

Exploratory Data Analysis

Trending Lifecycle

Metrics explanation

- I focus on two key time intervals:
  - Average going-viral time=Initial trending date-publish date
  - Average trending days=Final trending date-Initial trending date
- The average going-viral time for each category describes on average how fast a video can show up on the trending charts. The longer time interval is, the larger the time cost will be.
- The average trending days for each category explain on average how long a video can keep attracting audience. The longer time interval is, the larger number of target customers will be.

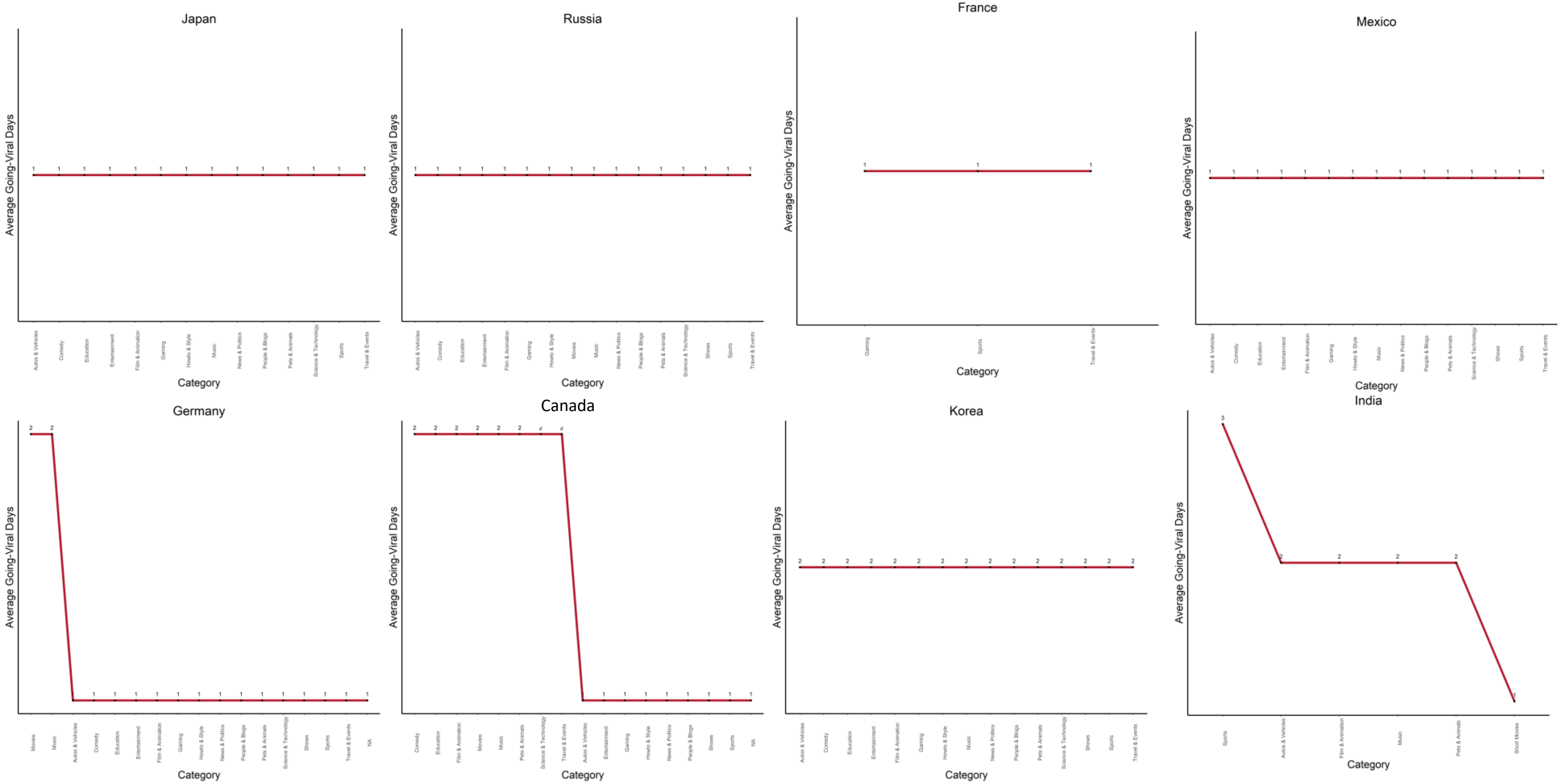
Insights from visuals (see Figure 6, Figure 7)

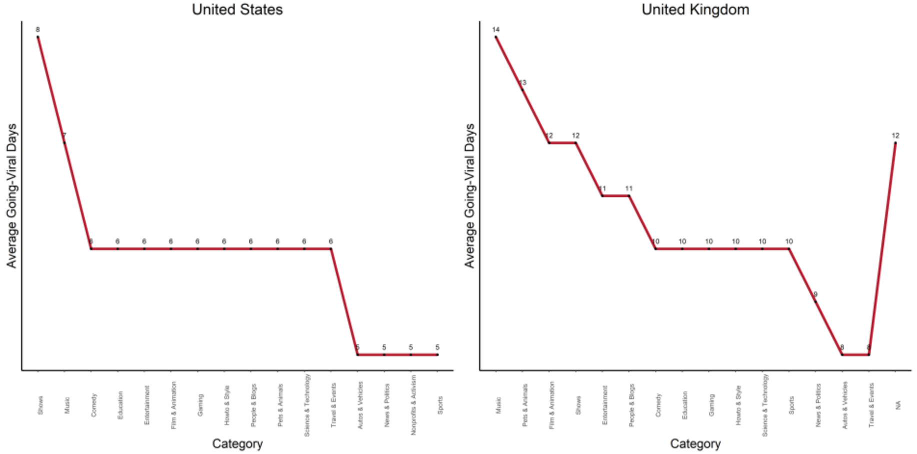
- By country
  - Average going-viral time: In most countries, it took **1-3 days** for a video to go viral. However, in U.S and UK, it took **1-2 weeks.**
  - Average trending days: It showed similar trend as the average going-viral time.
- By category
  - Average going viral time: Music videos take the longest time to go viral on YouTube. This may explained by the fact that:
    - Music videos normally lack breath-taking storylines, so they take time to attract audience;
    - There exist other more focused music video platforms, like Vimeo and TikTok, making audience switching to them.
  - Average trending days: Shows and movies are outstanding among all categories.

1 Day

2-3 Days

1-2 Weeks



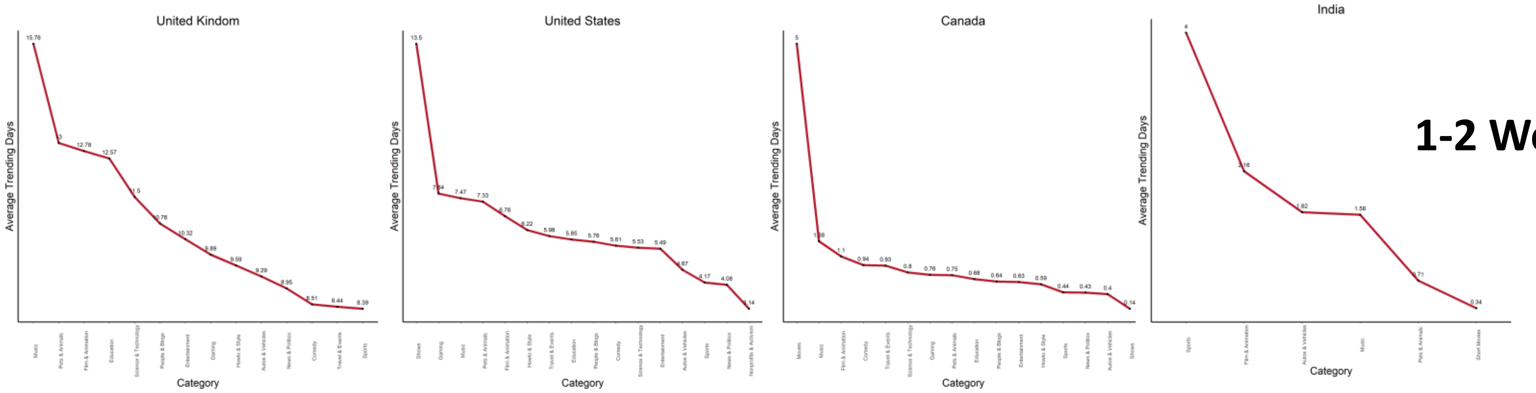


Average Going-Viral Days of Trending YouTube Videos by category, by country

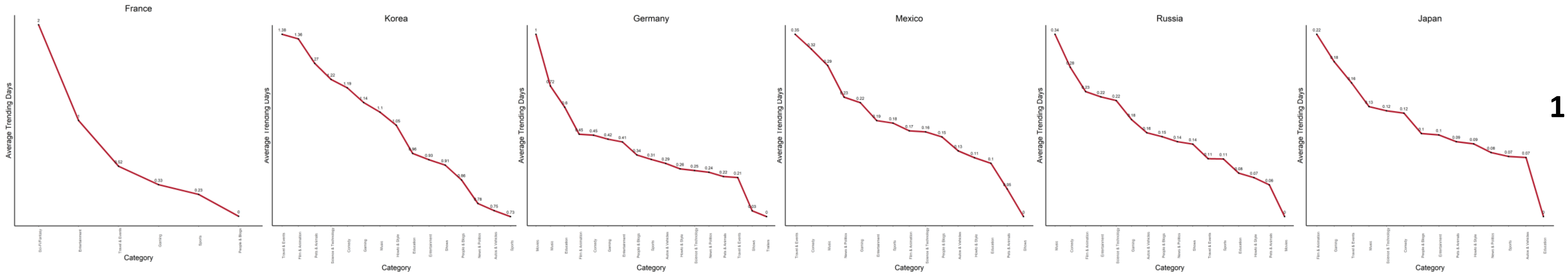
Figure 6

Average Trending Days of YouTube Videos by category, by country

Figure 7



1-2 Weeks



1 Day

Exploratory Data Analysis

Category Selection

Method

- In the evaluation matrix, I took view pattern, user participation and trending lifecycle into account and pick the variables below.
- Used the reverse ranking of each attribute as its score for categories, and then summed up all the scores to get a final score of each category to select most valuable category to do further analysis.

Findings

- **Music and entertainment** are among Top 3 popular categories based on evaluation matrix, which basically align to the previous EDA, i.e. music and entertainment videos' influence continues.
- **Film & Animation** becomes a new hitmaker and thus may be the next potential category for advertisers.

Category	Views	Voting Rate	Comment Rate	Going-Viral Time	Trending Days	Total
Music	16	16	16	2	14	64
Film & Animation	14	12	11	12	12	61
Entertainment	15	15	15	6	5	56
People & Blogs	12	13	14	8	8	55
Comedy	13	14	12	7	7	53
How to & Style	10	11	13	5	11	50
Gaming	8	8	10	9	15	50
Sports	11	10	8	15	3	47
Pets & Animals	5	5	4	11	13	38
Science & Technology	9	9	9	3	6	36
News & Politics	7	6	7	14	2	36
Education	6	7	6	4	9	32
Travel & Events	3	3	3	10	10	29
Autos & Vehicles	4	2	2	16	4	28
Nonprofits & Activism	2	4	5	13	1	25
Shows	1	1	1	1	16	20

Table 1. Evaluation Matrix-- U.S.

Modelling

Deep dive to the algorithm of trending chart

Questions of interest

- How the number of views, likes, dislikes and comments influence the going-viral time?
- Is the influence vary across different countries?

Method

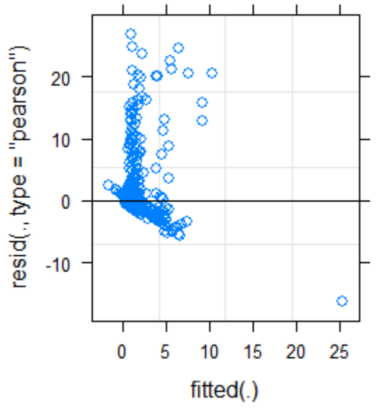
- Multilevel linear model with varying intercepts and slopes
- Predictors: Views (normalized because of scaling issues), Comment Rate (cv), Voting Rate (av)
- Response: Going-viral Days

Result and Discussion

- The influence of voting rate and comment rate and views on going-viral days is generally similar in most countries.
- In U.S. and UK, comment rate plays a more important role compared with other countries.
- In UK, since the intercept is largest, chances are are that there exist other more significant factors affecting the going-viral days.

```
Linear mixed model fit by REML ['lmerMod']
Formula: gov ~ av + cv + view_ct + (1 + av + cv + view_ct | country)
```

	(Intercept)	av	cv	view_ct
us	2.0039945	-0.8991614	-37.146220	0.04668118
ca	1.1791636	-0.9254056	-6.194495	0.10336491
de	1.1485251	-1.3955488	-2.199902	0.24190623
fr	1.2759276	-2.2908681	-1.648572	0.53410528
gb	5.2298689	-3.5678679	-142.131148	0.87074009
ind	1.0738804	-1.3684672	0.451075	0.22920014
jp	0.9205083	-0.7842695	2.712783	0.04297965
kr	1.1876241	-1.5738069	-2.495624	0.26446611
mx	1.1300223	-2.4755641	4.833146	0.62413860
ru	1.6921877	-4.9025184	-1.707702	1.38424223



Future Directions

- Dive deep into the comments to improve the analyses;
- Incorporate all video data into analyses, including trending and non-trending ones;
- Set different weight for variables used in evaluation matrix;
- Compare the characteristics of YouTube platform with others using the same methods to see different ecology of each online video platforms. If possible, I would write a report on the online video industry and become one of my portfolios for my future employers.

YouTube