YOUTUBE DATA ANALYSIS

TEAM - THE FANTASTIC FIVE

11.1.2022

MEET THE TEAM

- **Yuteng Zhang**

  Statistical Analysis
- **Dipti Sontakke**

  Presentation, Data Analysis
- **Alejandro Gutierrez**

  Data Intake, Data Cleaning
- **Patrick Brennan**

  Data Intake, Data Cleaning
- **Daniel King-Alan**

  Project Manager

AGENDA

INTRODUCTION

PROJECT DESCRIPTION

ANALYSIS

CONCLUSION

REFERENCES

INTRODUCTION : What we have achieved

YouTube is common thread for many social media users around the globe. Our projects focuses on analysis of views, likes and comments based on Country and Category.

Our analysis focuses on answering below questions:

- Is there a significant difference in the average like count of the top 50 videos within each designated region within YouTube?
  - H_a: Average like count will be significantly different between the regions
  - H_0: There is no difference between the average like count by region
- Is there a significant difference in the average like count of the top 50 videos of each designated category within YouTube?
  - H_a: Average like count will be significantly different between the categories
  - H_0: There is no difference between the average like count by category
- Additional Correlations:

  What is the average percent of videos liked by the number of views they have received?

```
ion:absolute;z-index:9
x 5px #ccc}.gbrtl .gbm
display:block;positio
pacity:1;*top:-2px;*l
/;top:-4px\0/;left:-6
ne-box;display:inlin
display:block;list-s
ne-block;line-height
pointer;display:bl
ative;z-index:1000
padding-right:9px
```

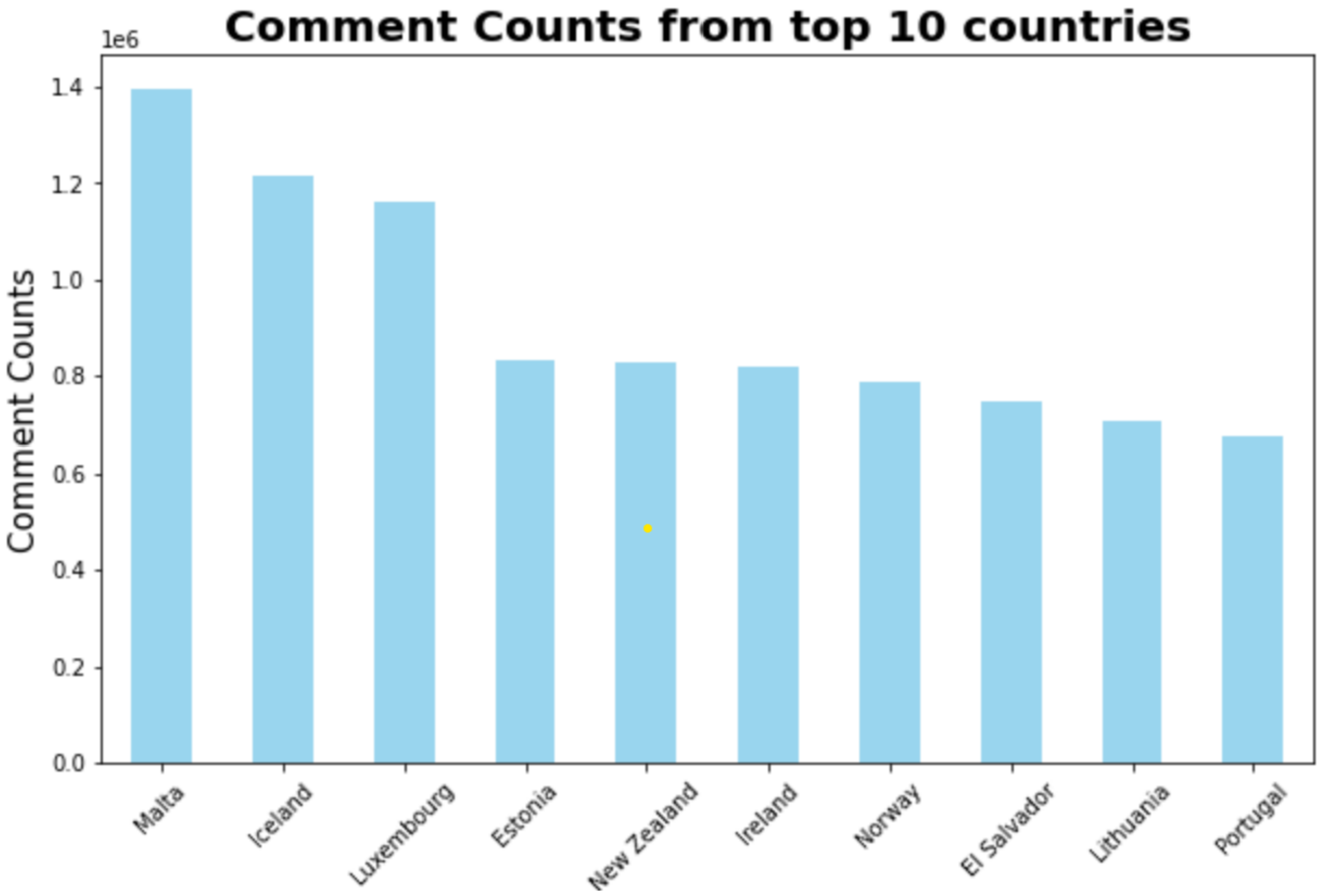
WHY YOUTUBE DATASET? LET'S DIVE IN

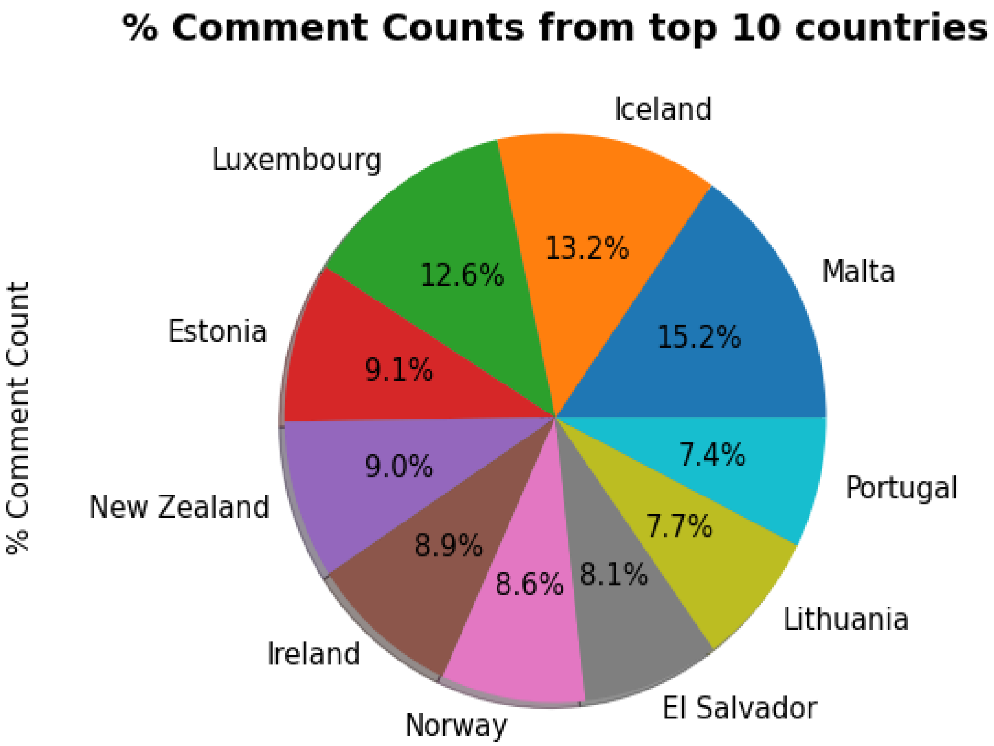
- YouTube's API falls under the umbrella of the suite of APIs offered by Google, allowing us to make use of the Google API library for easier access

- We retrieved data from two of the 'Resources' the YouTube API offers, Regions and Videos
  - From Regions, we pulled 109 country names and their IDs
  - From Videos, we pulled counts for likes, views, and comments for the top 50 videos by each country

- The API only allows us to call a maximum of the 50 most popular videos by country for the day the call is made, essentially a screenshot in time
  - As we will see, this can create for interesting results in the countries with the most views on their most popular videos

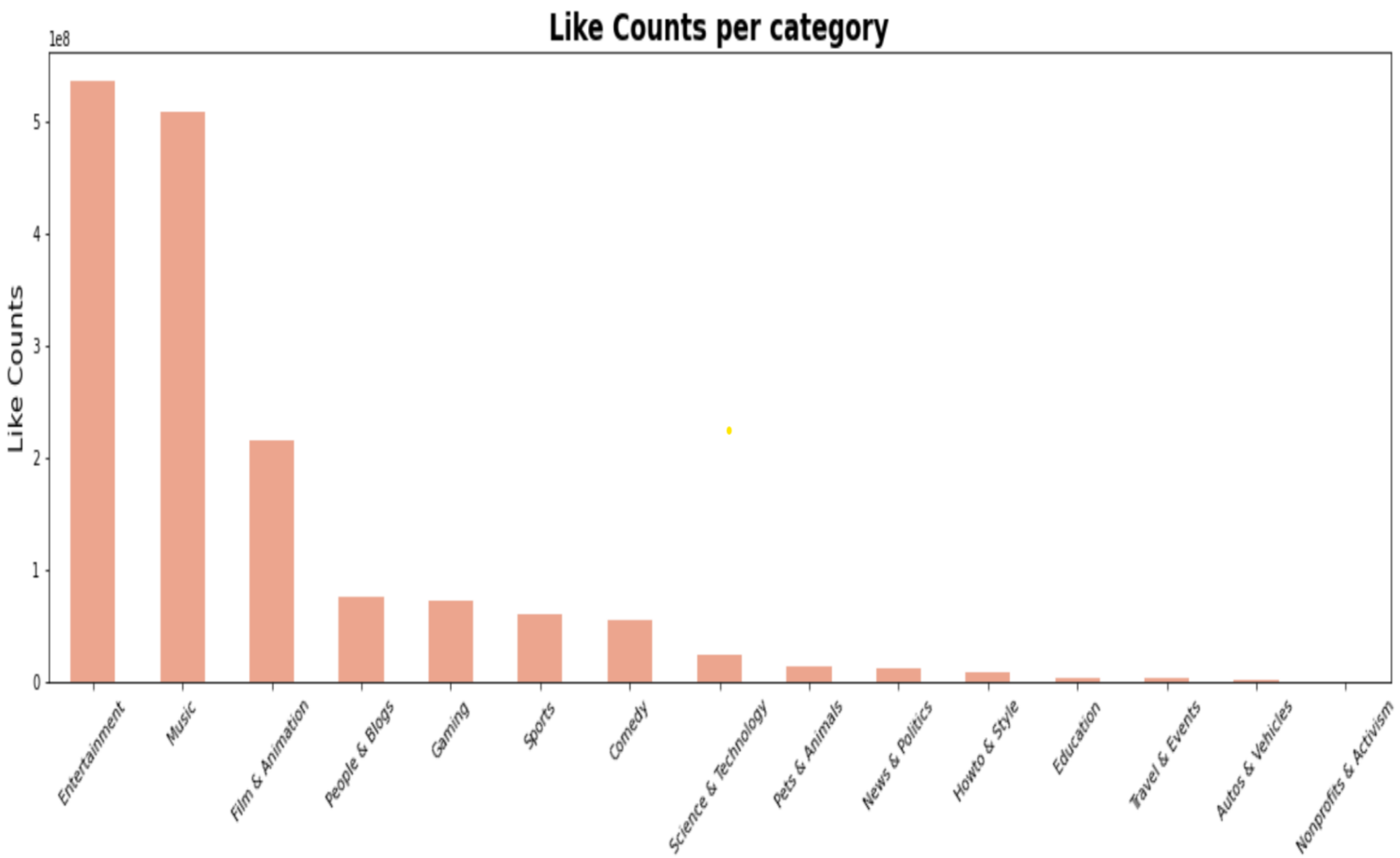
Comment counts from top 10 Countries

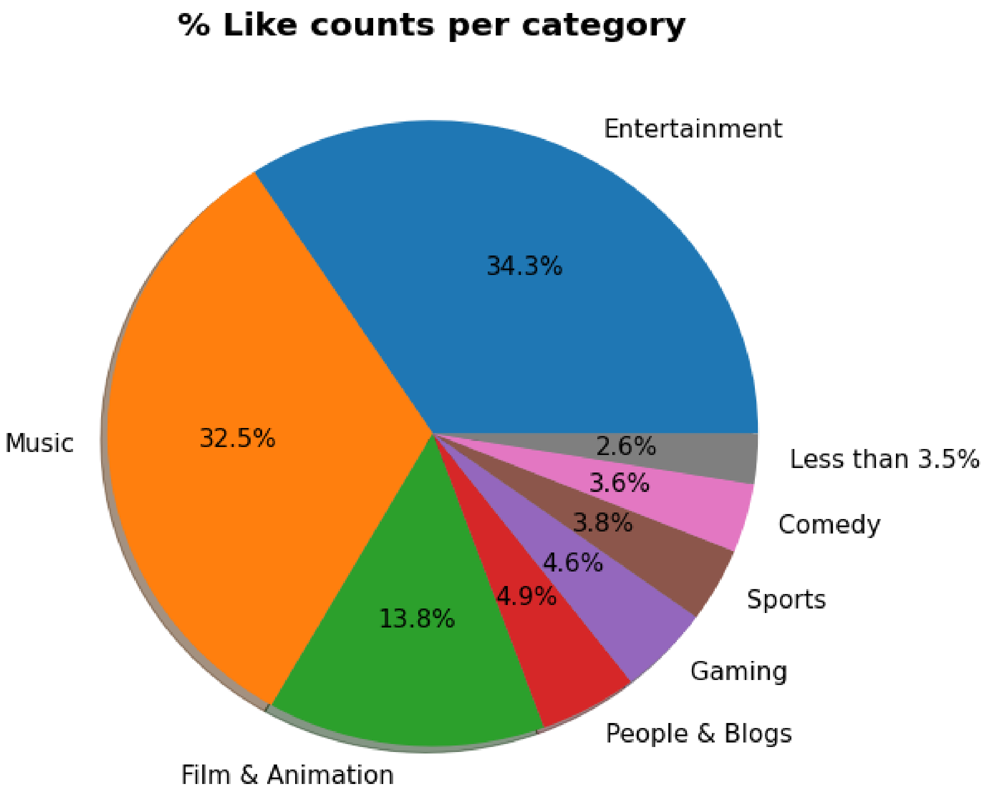




- Distribution of comment counts per million by countries.
- Distribution of percent of comment counts from top 10 countries.
- The top 10 countries listed here and in the following slides are not the ones we may expect to see based off population size, but this is a results of the limitations of the API data calls.
- One could theorize that these countries have more concentrated viewing focused on their top 50 videos than other countries, or the categories of video they watch are more broadly popular.

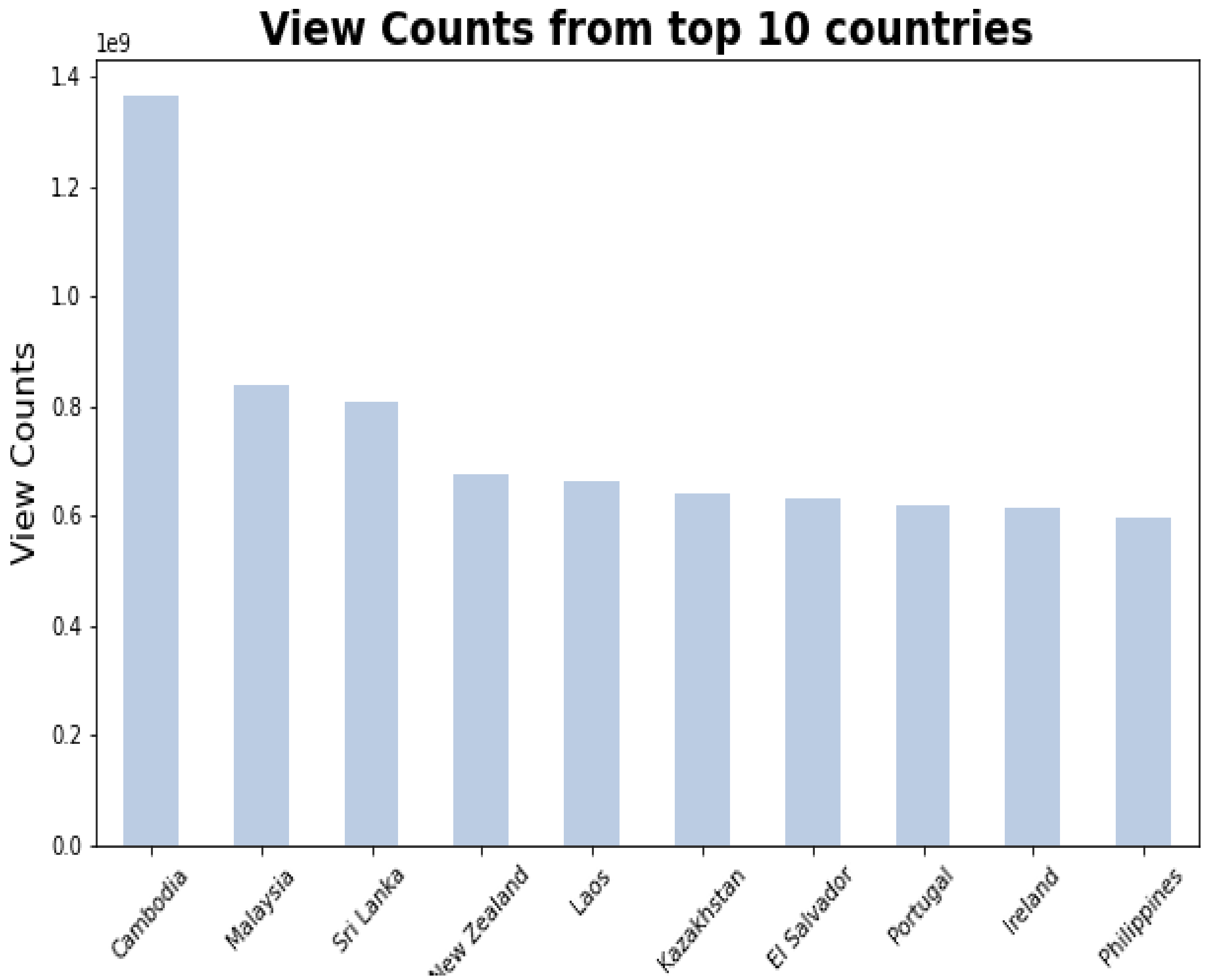
Likes counts from all categories

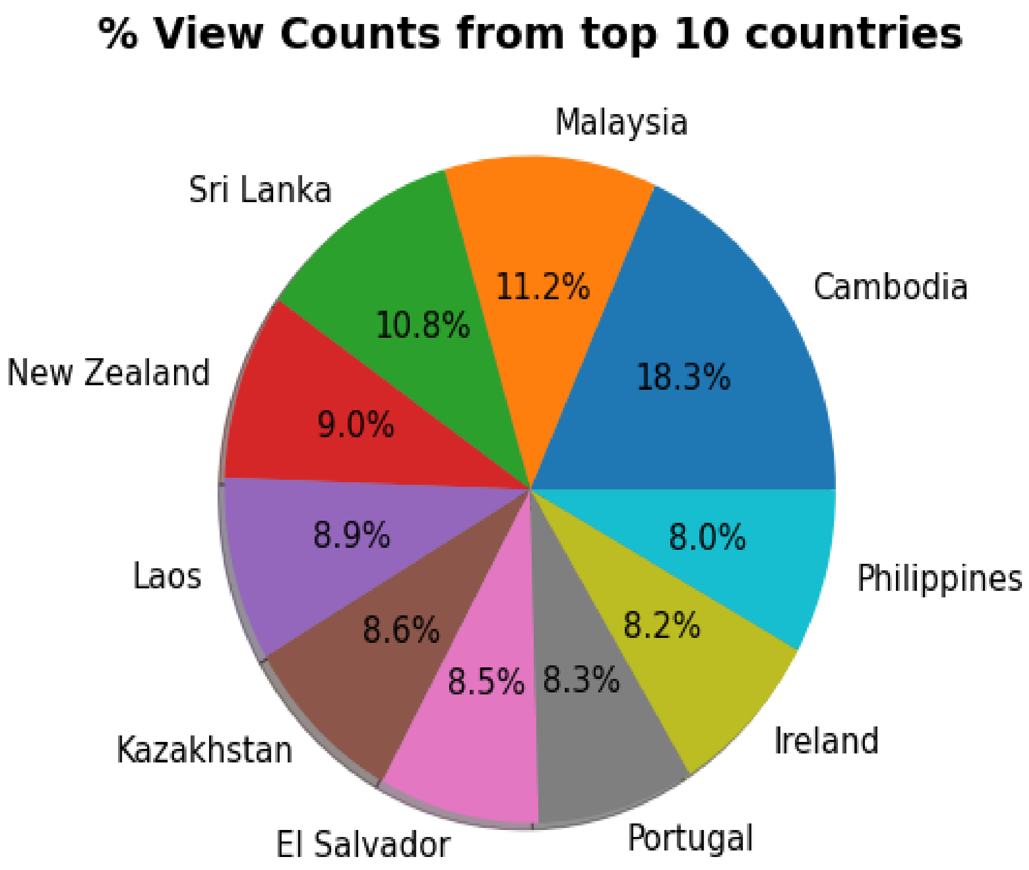




- Distribution of likes counts per hundred million by category.
- Distribution of percent of likes counts from all category.

View counts from top 10 Countries

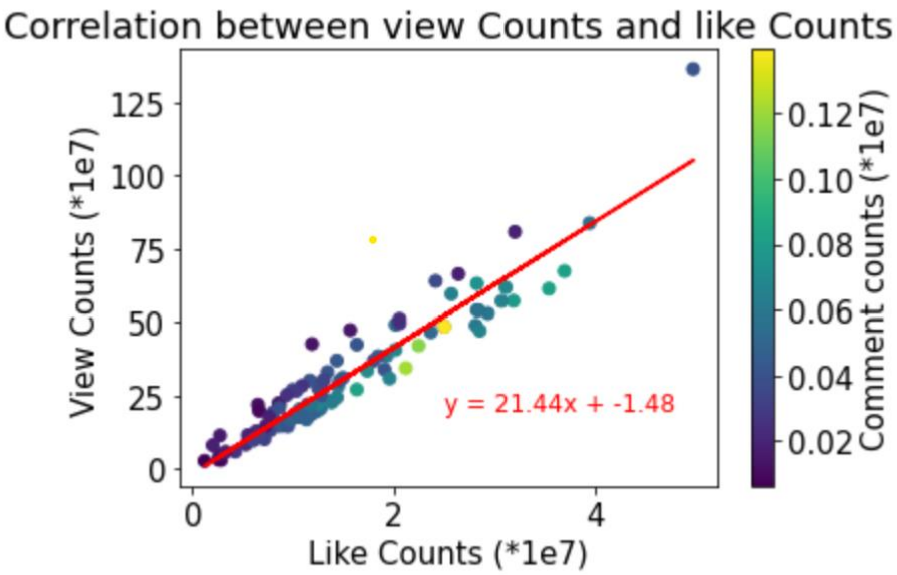




- Distribution of view counts per billion by country.
- Distribution of percent of view counts for top 10 countries.

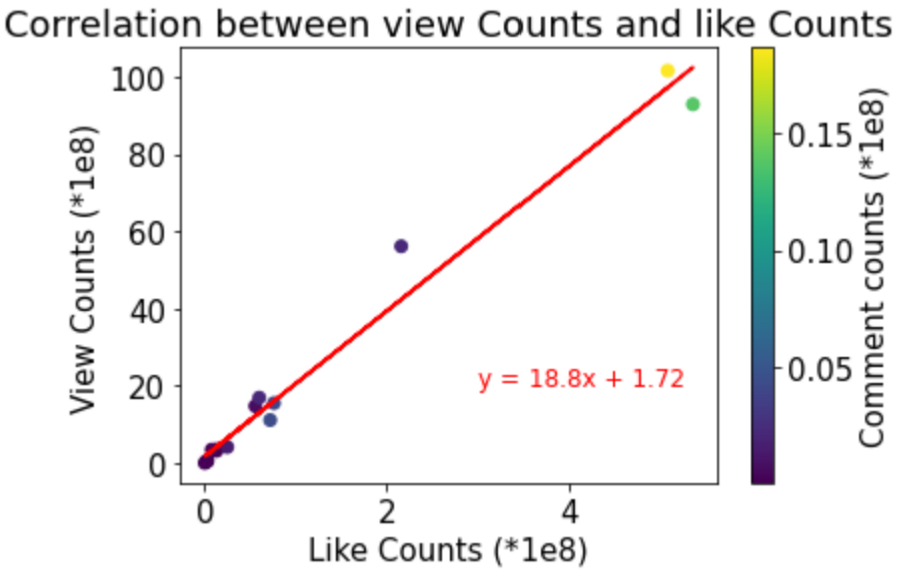
Correlation Analysis by country and category

Based on Country



```
The correlation coefficient between view counts and like counts is 0.95
```

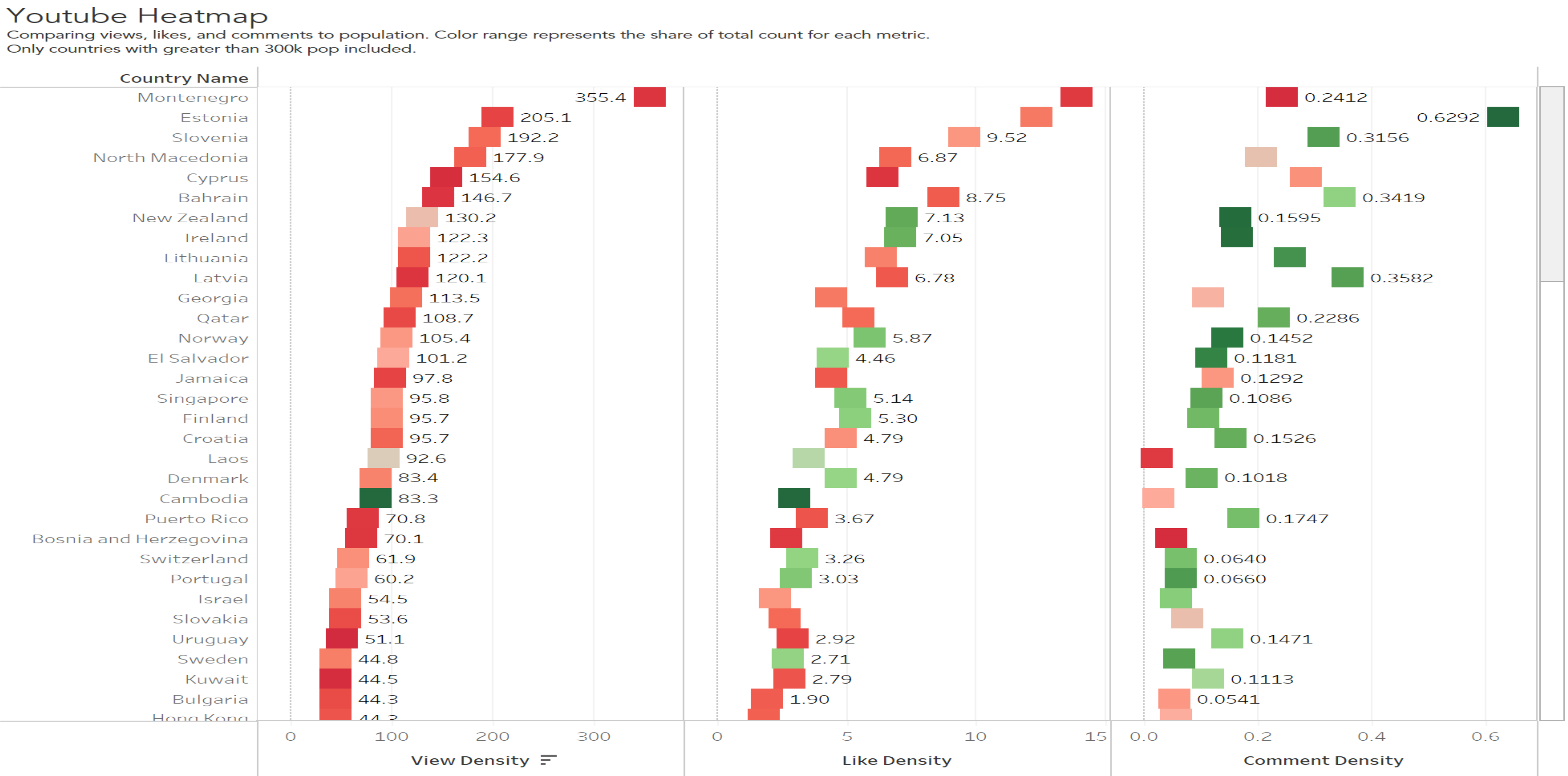
Based on Category



```
The correlation coefficient between view counts and like counts is 0.99
```

- View counts and like counts show us a strong positive correlation.
- This correlation output is applicable for both country and category.

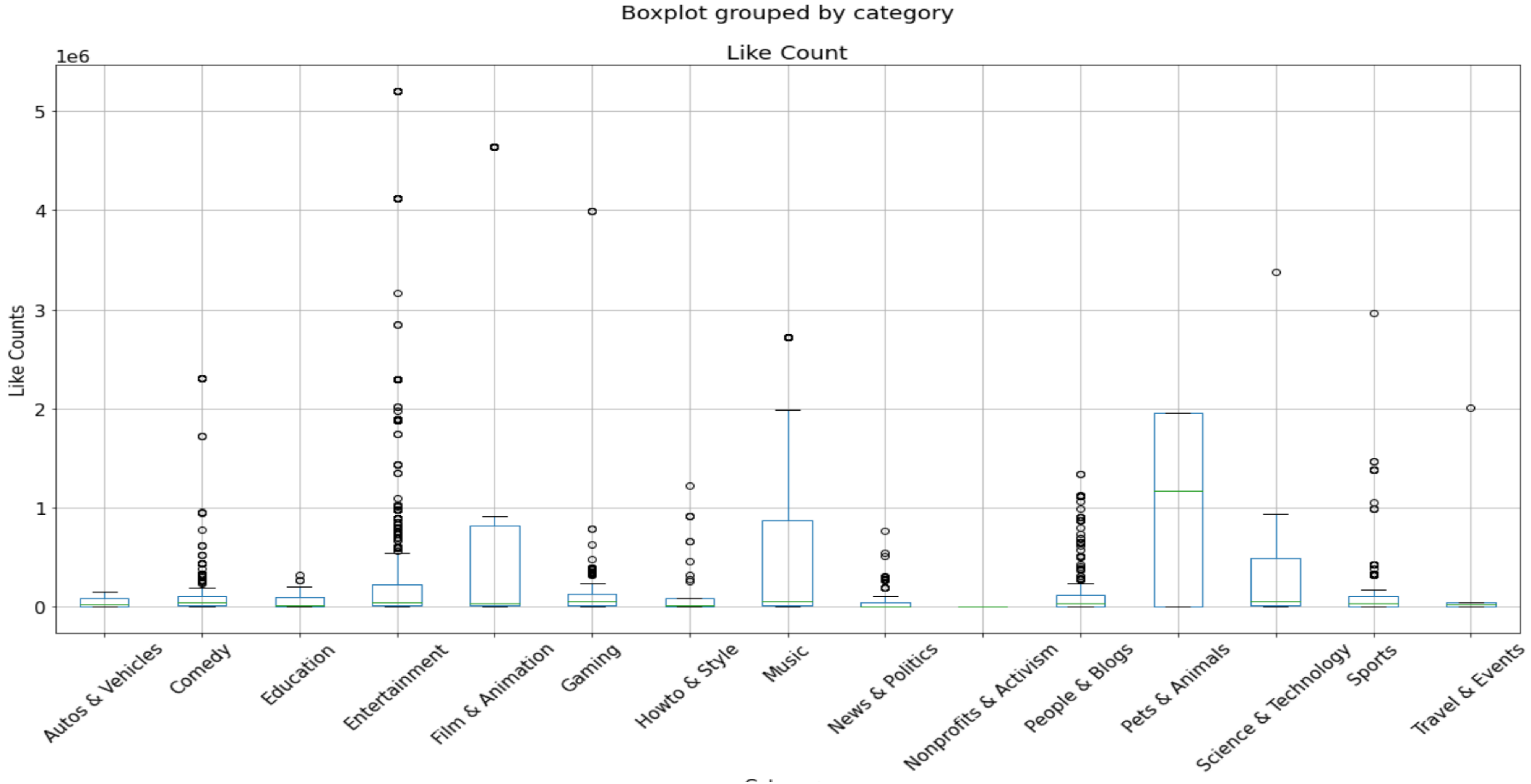
Heatmap – Comparing, Views Likes and Comments to population



Youtube Heatmap

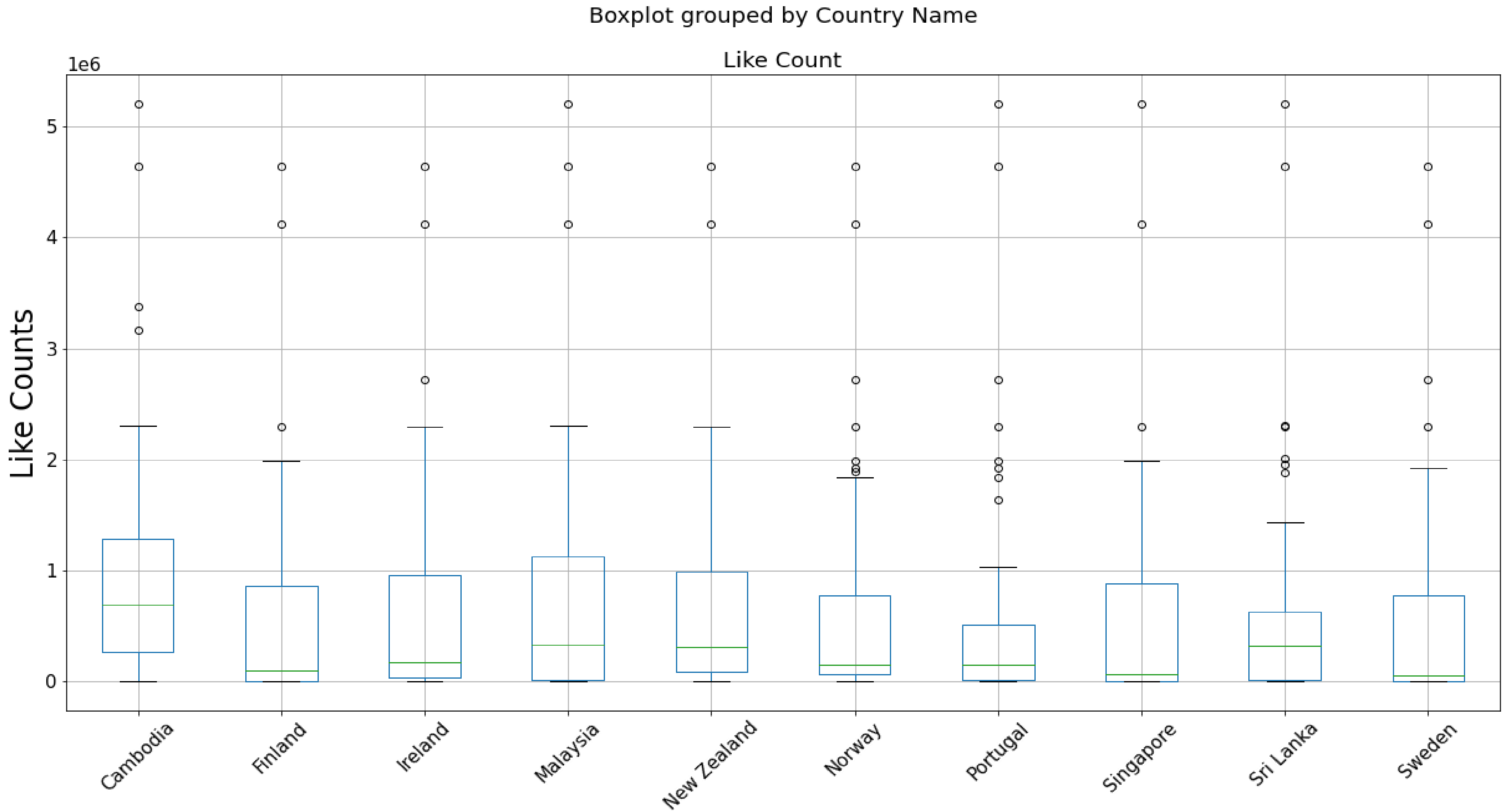
Comparing views, likes, and comments to population. Color range represents the share of total count for each metric.
Only countries with greater than 300k pop included.

ANOVA Testing by Category



- This is one way ANOVA test done by Category.
- Also used **t-test** to determine if there is a significant difference between the mean of two groups. Also calculated p-value(2.9209956568816213e-94) is smaller than 0.05 indicates that the compared two groups do have significant difference.
- The two groups compared: views vs comment; views vs likes and likes vs comments

ANOVA Testing by top 10 country



- This is one way ANOVA test doe by Category.
- Also used t-test to determine if there is a significant difference between the mean of two groups. Also calculated p-value(7.403698275134119e-32) is smaller than 0.05 indicates that the compared two groups do have significant difference.
- The two groups compared: views vs comment; views vs likes and likes vs comments

CONCLUSION

1. Accepting the findings as we have done ANOVA, t-test, and correlation coefficient that this Hypothesis testing is acceptable.
2. P-Value is less than 0.05 so we reject the null hypothesis and conclude that there are differences between regions and in their like counts and view counts.
3. P-Value is less than 0.05 so we reject the null hypothesis and conclude that there are differences between category and in their like counts and view counts.
4. Chances of getting more views, likes and comment counts from a video created for Music and Entertainment category is more compared to other categories such as DIY, Sports, News and Politics.
5. Music and Entertainment industry is in total has 65% more views compared to other categories.
6. Surprisingly, Cambodia has highest view count (approx. 14 billion) compared to other top country views. This may be due to Cambodian viewership being more highly concentrated in the top 50 most popular videos for that country.
7. As view count increases the like and comment counts also increases accordingly.
8. Top 5 countries that have the highest density of viewership are all located in Eastern or Southern Europe. It could indicate either cultural or socio-economic trend of those regions, or combination of both.
9. View count and view density showed different behaviors. From the top 10 countries for view count the majority were southeast Asian nations. Comparing this with view density a different pattern emerges where 9 of the top 10 countries were smaller European nations primarily in the Baltics and Balkans.

REFERENCES

YouTube API - https://developers.google.com/youtube/v3

Python - https://pandas.pydata.org/

GitHub - https://github.com/PatttyCakess/Youtube-API-project

THANK YOU

Fantastic Five

https://github.com/PatttyCakess/Youtube-API-project

Georgia Tech – Boot Camp