Data Analysis for Trending Content and Top Influencer on Social Media Platforms

Team: Liang Li , Mamesa El

Summary

Social media plays an important role in our daily lives. Whether it is for entertainment purposes, social interaction, or other reasons, it keeps us connected with the world. With more and more social media platforms and data available, we can begin to study the user behaviors and find the underlying connections with different social media platforms. The goal of this data analysis is to identify the features correlated with popularity for TikTok trending tracks/videos, and compare the social interaction data across social media platforms (Instagram, YouTube, TikTok). We want to understand the relationship between followers or subscribers, social engagement, and influencer category. In addition, we shared some findings and insights about the datasets, such as the top track name, top hashtag, and top video categories for influencers.

Questions

1. What features of the TikTok song (such as song duration, release date, danceability, energy, loudness, mode, speechiness, acoustiness, instrumentalness, liveness, valence, tempo, and genre) are correlated with popularity rating?
    - Statistics for all the song features
    - Which genre is most popular
    - How does speechiness compare with acousticness in influencing the popularity
2. What features of the TikTok trending video (such as caption text, author metadata, music metadata) are correlated with play count?
    - Statistics for the video features
    - Statistics for the user engagement features (e.g., like, share, comment)
    - What are the most popular hashtags in these trending videos
    - What day in a month has the most trending videos getting created
3. How do the top influencers category and social interaction differ across social media platforms (Instagram, Youtube, and Tiktok)?
    - Statistics for engagement for each social media platform
    - What social interaction (such as likes, views, and comments) and genre have the highest correlation with total subscribers?
    - What is the most popular social media platform by followers and subscribers?
    - Which influencer category dominates the trending list for YouTube and Instagram
    - Do influencers with more followers tend to have more engagement e.g. likes, comments

Datasets

- Dataset 1 - [TikTok Trending Tracks | Kaggle]
  - This dataset is stored in a CSV file with **6747** records (rows) and **24** features (columns).
  - The features are: ***index*** *(starting from 0),* ***track_id*** *(random string identifier),* ***track_name*** *(name of the track),* ***artist_id*** *(random string identifier),* ***artist_name*** *(name of the artist),* ***album_id*** *(random string identifier),* ***duration*** *(duration of the track in millisecond),* ***release_date*** *(date the track was released),* ***popularity*** *(0-100 score indicating popularity),* ***danceability*** *(0-1 score indicating danceability),* ***energy*** *(0-1 score indicating energy),* ***key*** *(key of the track),* ***loudness*** *(volume of the track),* ***mode*** *(0/1 flag),* ***speechiness*** *(0-1 score indicating the ratio of speech sound),* ***acousticness*** *(0-1 score indicating the ratio of acoustic sound),* ***instrumentalness*** *(0-1 score indicating the ratio of instrument sound),* ***liveness*** *(0-1 score indicating if the track is live),* ***valence*** *(0-1 score indicating the musical positiveness),* ***tempo*** *(number indicating how fast/slow the track plays),* ***playlist_id*** *(random string identifier),* ***playlist_name*** *(random string identifier),* ***duration_mins*** *(duration in minutes),* ***genre*** *(genre of the music).*
  - Out of these features, we think artist_name,track_id, track_name, duration, release_date, danceability, energy, loudness, mode, speechiness, acoustiness, instrumentalness, liveness, valence, tempo, and genre features are helpful for asking our questions
- Dataset 2 - [Tiktok Trending Video | Kaggle]
  - This dataset is stored in a JSON file with **1000** key-value pairs and **17** features.
  - The features are: **id** (unique integer identifier), **text** (caption text of the video), **createTime** (video creation time), **authorMeta** (metadata about the author like name and signature), **musicMeta** (metadata about the music like name and author), **covers** (URL to the video covers), **webVideoUrl** (video URL for web client), **videoUrl** (video URL for mobile client), **videoUrlNoWaterMark** (URL to video without watermark), **videoMeta** (metadata about the video), **diggCount** (maybe "like" counts), **shareCount** (count of video gets shared), **playCount** (count of video played), **commentCount** (number of comments), **downloaded** (number of downloads), **mentions** (users mentioned), **hashtags** (topics).
  - Out of these features, we think id, createTime, aurhorMeta, musicMeta, videoMeta, diggCount, shareCount, playCount, commentCount, downloaded, mentions, and hashtags features are helpful for asking our questions
- Dataset 3 - [Social Media Influencers | Kaggle]
  - Instagram
    - This dataset is stored in a CSV files with 1000 records and 8 features
    - The features are:
      - **Influencer insta name**(Influencer instagram name)**, instagram name**(instagram name), **category_1**(Types of video), **category_2**(Genres of the video), **Followers**(Total amount of followers), **Audience country(mostly)** (Country of most viewed audience), **Authentic engagement\r\n**(Total amount of audience that shares, like, or comment on relevant topic to the video), **Engagement avg\r\n**(Total amount of audience that shares, like, or comment on the video)

        - Out of these features, we think instagram name, category_1, category_2, Followers, Audience country(mostly), Authentic engagement\r\n, and Engagement avg\r\n features are helpful for asking our questions
- YouTube
    - This dataset is stored in a CSV files with 1000 records and 8 features
    - The features are:
        - **Youtuber name**(Youtuber username), **channel**(The channel name), **Category**(Type of video), **Subscribers**(Total subscribers), **Audience Country**(Country of most viewed audience)), **avg views**(Average audience views), **avg likes**(Average audience likes), **avg comments**(Average comments made on the video)
        - Out of these features, we think Youtuber name, channel, Category, Subscribers, Audience Country, avg views, avg likes, and avg comments features are helpful for asking our questions
- TikTok
    - This dataset is stored in a CSV files with 1000 records, and 7 features
    - The features are:
        - **Tiktoker name**(TikTok influencer name), **Tiktok name(**TikTok name), **Subscribers count(**Total subscribers), **Views avg.**(Total viewers), **Likes avg.**(Total likes), **Comments avg.(**Total comments made), **Shares avg(**Total time the video is shares)
        - Out of these features, we think Tiktok name, subscribers count, views ave, likes avg, comments avg, shares avg are helpful of asking our questions

Exploratory Analysis:

Dataset 1 & 2: TikTok trending tracks & videos

In the trending tracks dataset, there are 6746 rows and 24 columns. Each column has 0 null value. The first column name is broken (shown as "Unamed: 0"), we verified it is index starting from 0, and renamed the column as "index". The data is mostly clean and well formatted.
In the trending videos dataset, there are 1000 rows and 17 columns. Each column has 0 null value. The data in columns with "Meta" as suffixes are in JSON format. We normalized them so that they spread out as separate columns, e.g., previously all video related information is stored in "video" column as "{'videoWidth': 100, 'videoHeight': 200}", after JSON normalization, it expands into a few separate columns including "videoMeta.width", "videoMeta.height", and so on. For JSON columns with list of string such as "hashtags" column, we exploded them into multiple rows so that we could find the most popular hashtags later.

```
RangeIndex: 6746 entries, 0 to 6745
Data columns (total 24 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Unnamed: 0        6746 non-null   int64
 1   track_id          6746 non-null   object
 2   track_name        6746 non-null   object
 3   artist_id         6746 non-null   object
 4   artist_name       6746 non-null   object
 5   album_id          6746 non-null   object
 6   duration          6746 non-null   int64
 7   release_date      6746 non-null   object
 8   popularity        6746 non-null   int64
 9   danceability      6746 non-null   float64
 10  energy            6746 non-null   float64
 11  key               6746 non-null   int64
 12  loudness          6746 non-null   float64
 13  mode              6746 non-null   int64
 14  speechiness       6746 non-null   float64
 15  acousticness      6746 non-null   float64
 16  instrumentalness  6746 non-null   float64
 17  liveness          6746 non-null   float64
 18  valence           6746 non-null   float64
 19  tempo             6746 non-null   float64
 20  playlist_id       6746 non-null   object
 21  playlist_name     6746 non-null   object
 22  duration_mins     6746 non-null   float64
 23  genre             6746 non-null   object
dtypes: float64(10), int64(5), object(9)
memory usage: 1.2+ MB
```

```
RangeIndex: 1000 entries, 0 to 999
Data columns (total 17 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   id                   1000 non-null   object
 1   text                 1000 non-null   object
 2   createTime           1000 non-null   int64
 3   authorMeta           1000 non-null   object
 4   musicMeta            1000 non-null   object
 5   covers               1000 non-null   object
 6   webVideoUrl          1000 non-null   object
 7   videoUrl             1000 non-null   object
 8   videoUrlNoWaterMark  1000 non-null   object
 9   videoMeta            1000 non-null   object
 10  diggCount            1000 non-null   int64
 11  shareCount           1000 non-null   int64
 12  playCount            1000 non-null   int64
 13  commentCount         1000 non-null   int64
 14  downloaded           1000 non-null   bool
 15  mentions             1000 non-null   object
 16  hashtags             1000 non-null   object
dtypes: bool(1), int64(5), object(11)
memory usage: 126.1+ KB
```

Figure 1 & 2: left - overview for trending tracks; right - overview for trending videos

Dataset 3: Social Media Influencer

Social Media Influencers consist of three dataset. Each data set has 1000 rows of data. Figures 3,4, and 5 below contain the information regarding the three data frames. The tables display the total number of non-null values and the datatype of each variable. For each data frame, the numerical features containing null values are removed so that mathematical functions can be applied later. We checked and removed duplicates in 'youtuber name', 'Influencer insta name', 'and 'Tiktoker name' . Numerical features such as views, likes, comments, and shares contain strings such as 'M' or 'K'. These strings were converted to 1e6 and 1e3, respectively, while the column is set to 'int' data type.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 786 entries, 0 to 999
Data columns (total 8 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   youtuber name     786 non-null    object
 1   channel name      786 non-null    object
 2   Category          571 non-null    object
 3   Subscribers       786 non-null    int64
 4   Audience Country  786 non-null    object
 5   avg views         786 non-null    int64
 6   avg likes         786 non-null    int64
 7   avg comments      786 non-null    int64
dtypes: int64(4), object(4)
memory usage: 55.3+ KB
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 8 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Influencer insta name     1000 non-null   object
 1   instagram name            979 non-null    object
 2   category_1                892 non-null    object
 3   category_2                287 non-null    object
 4   Followers                 1000 non-null   int64
 5   Audience country(mostly)  986 non-null    object
 6   Authentic engagement
  1000 non-null   int64
 7   Engagement avg
        1000 non-null   int64
dtypes: int64(3), object(5)
memory usage: 62.6+ KB
```

Figure 3 & 4: The left table is the information regarding Youtube dataframe. The right table is the information regarding Instagram dataframe.

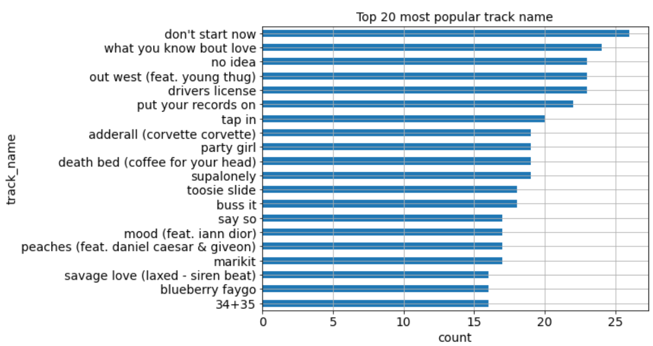
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 987 entries, 0 to 999
Data columns (total 7 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Tiktoker name      987 non-null    object
 1   Tiktok name        985 non-null    object
 2   Subscribers count  987 non-null    int64
 3   Views avg.         987 non-null    int64
 4   Likes avg          987 non-null    int64
 5   Comments avg.      987 non-null    int64
 6   Shares avg         987 non-null    int64
dtypes: int64(5), object(2)
memory usage: 61.7+ KB
```

Figure 5: This table contains the information regarding the TikTok dataframe.

Data Analysis

Dataset 1 - TikTok trending tracks

The two figures shown below show the top 20 most popular track and artist names respectively. The most popular track name is "Don't Start Now" which occurs 26 times. The most popular artist name is "Doja Cat" which occurs about 95 times.



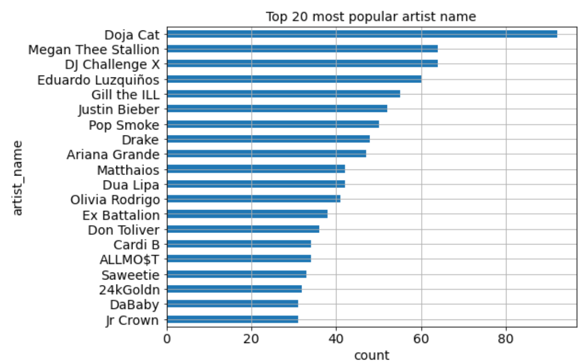
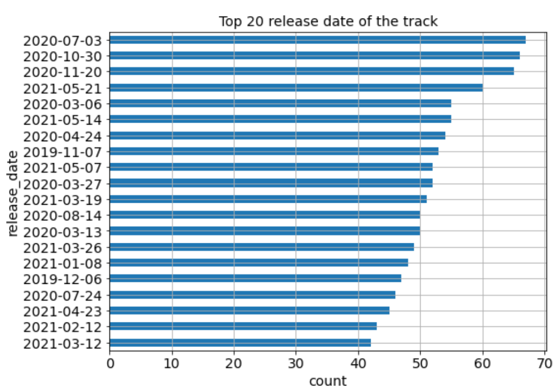


Figure 6 & 7: left - top 20 most popular track name; right - top 20 most popular artist name

The top 20 release date of the trending tracks shows the most popular release date is 2020-07-03, which is one day before Independence Day. And the most popular release day is the first day in a month.



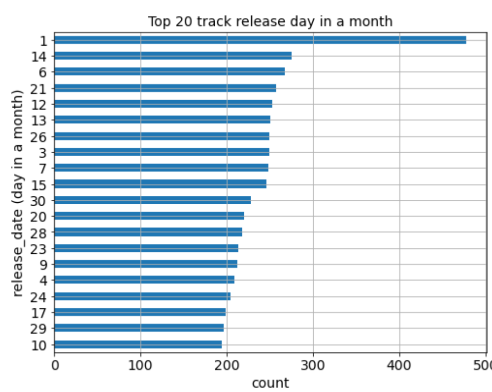


Figure 8 & 9: left - top 20 most popular release date; right - top 20 most popular release day

As we can see from the genre and music key distribution below, “TIKTOK DANCE” is the most popular genre followed by “_TIKTOK”. Key 1 is the most popular music key out of all 12 music keys. It has more than 1000 records in the datasets.

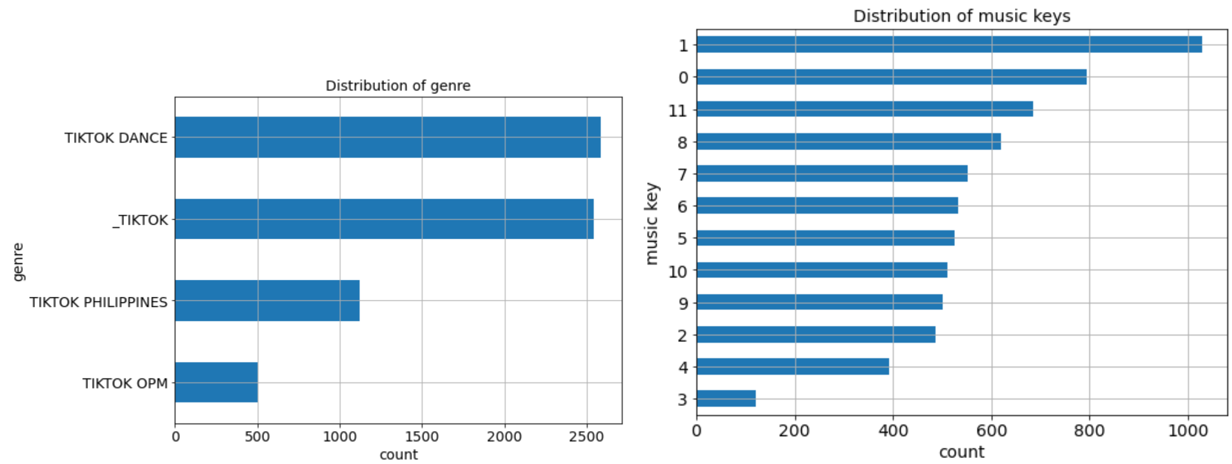


Figure 10 & 11: left - distribution of genre; right - distribution of music keys

Below is the correlation analysis for numerical features (e.g., loudness, danceability). Overall there is no direct correlation between the numerical feature and popularity rating. Energy has a relatively higher correlation coefficient (0.71) with loudness. Duration is correlated (1.00) with “duration_mins” which is expected because they can be mutually converted to each other.

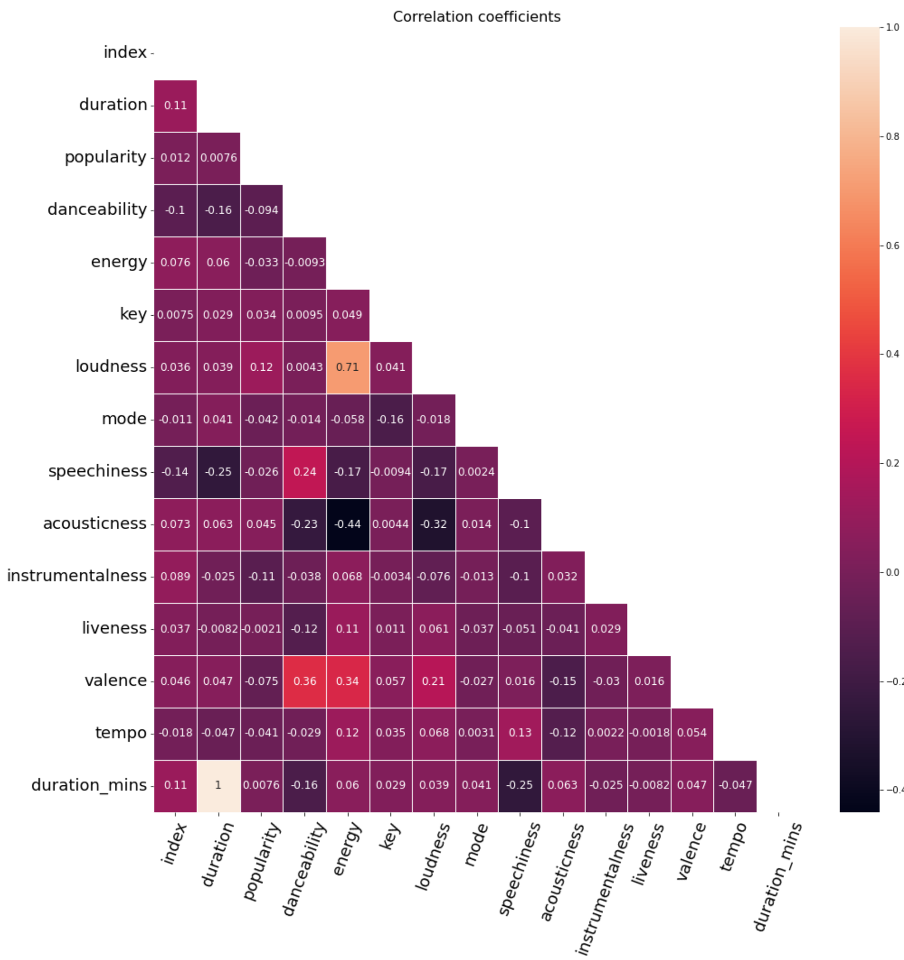


Figure 12: correlation coefficient matrix for trending track's numerical features

The correlation analysis for categorical features (e.g., track name, playlist name) shows that the correlation between track name, playlist name, and artist name with popularity is very small.

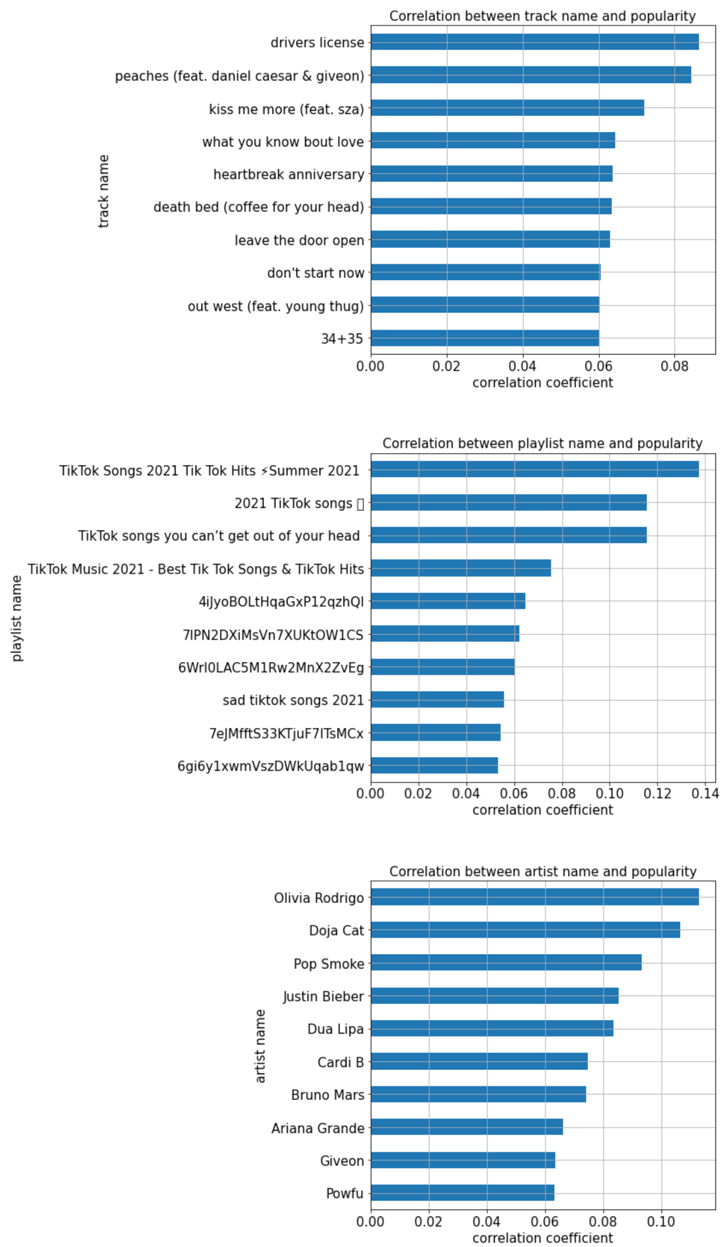


Figure 13: Correlation coefficient between trending track's categorical features and popularity

Dataset 2 - TikTok trending videos

The two figures below show the top 20 most popular video author name and music name respectively. The most popular video audio author is "timmytimmadome", and the most popular video music name is "original sound", which denotes the original video sound recorded from microphone instead of background music. The values in both columns are converted to lowercase in case we have duplicate values.

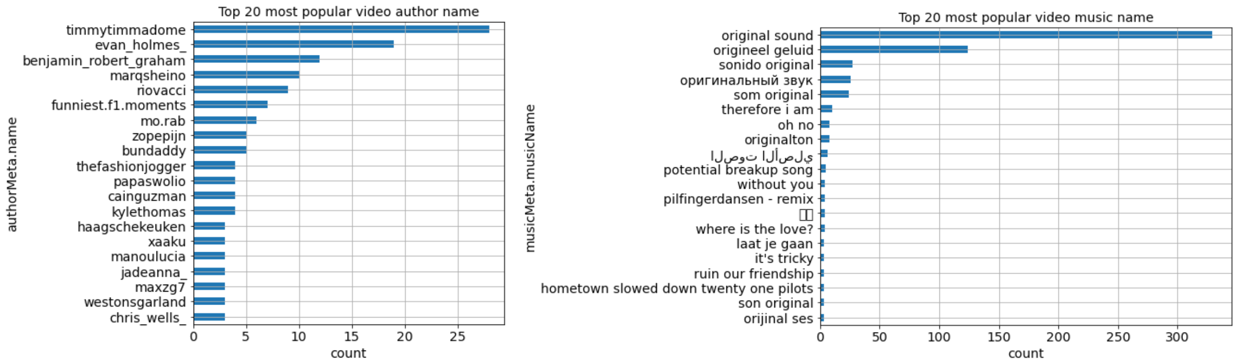


Figure 14 & 15: left - top 20 most popular video author name; right - top 20 most popular video music name

We also found the top 20 most popular hashtags, and video creation day in a month as below. The most popular hashtag is “fyp”, which occurs in more than 400 videos. The most popular video creation day is the first day in a month.

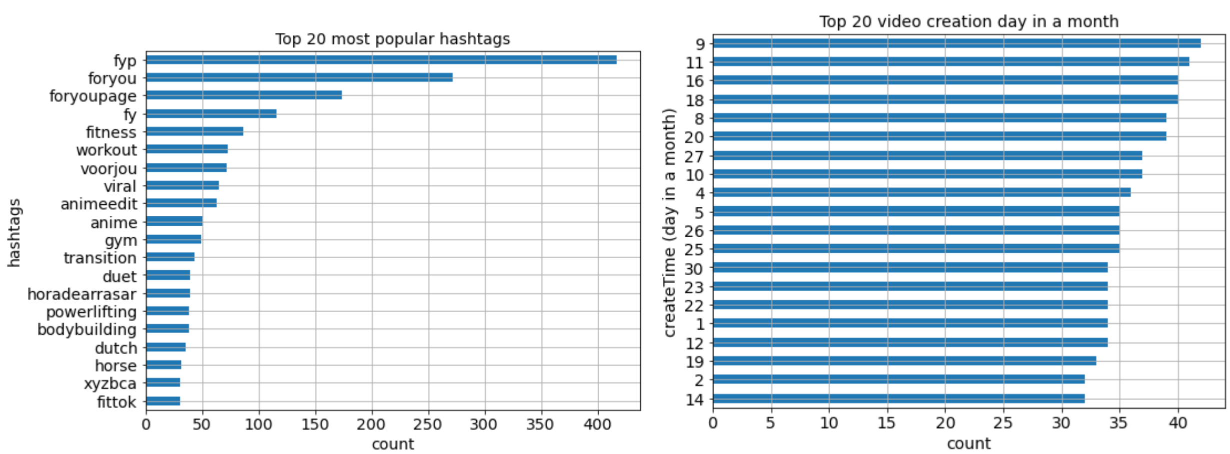


Figure 16 & 17: left - top 20 most popular hashtags; right - top 20 video creation day

The original user engagement data is skewed and it does not follow normal distribution. We tried normalizing it through box-cox transformation with scipy library. As we could see from the comparison, after applying box-cox transformation, the distribution of like/share/play/comment count becomes a normal distribution with the same scale. Most videos have like count below 2500 (after transformation, 15K before transformation), most videos are shared below 1000 times (both before and after transformation), most videos are played less than 5000 times (after transformation, 20K before transformation), most videos have less than 2000 comments (both before and after transformation).

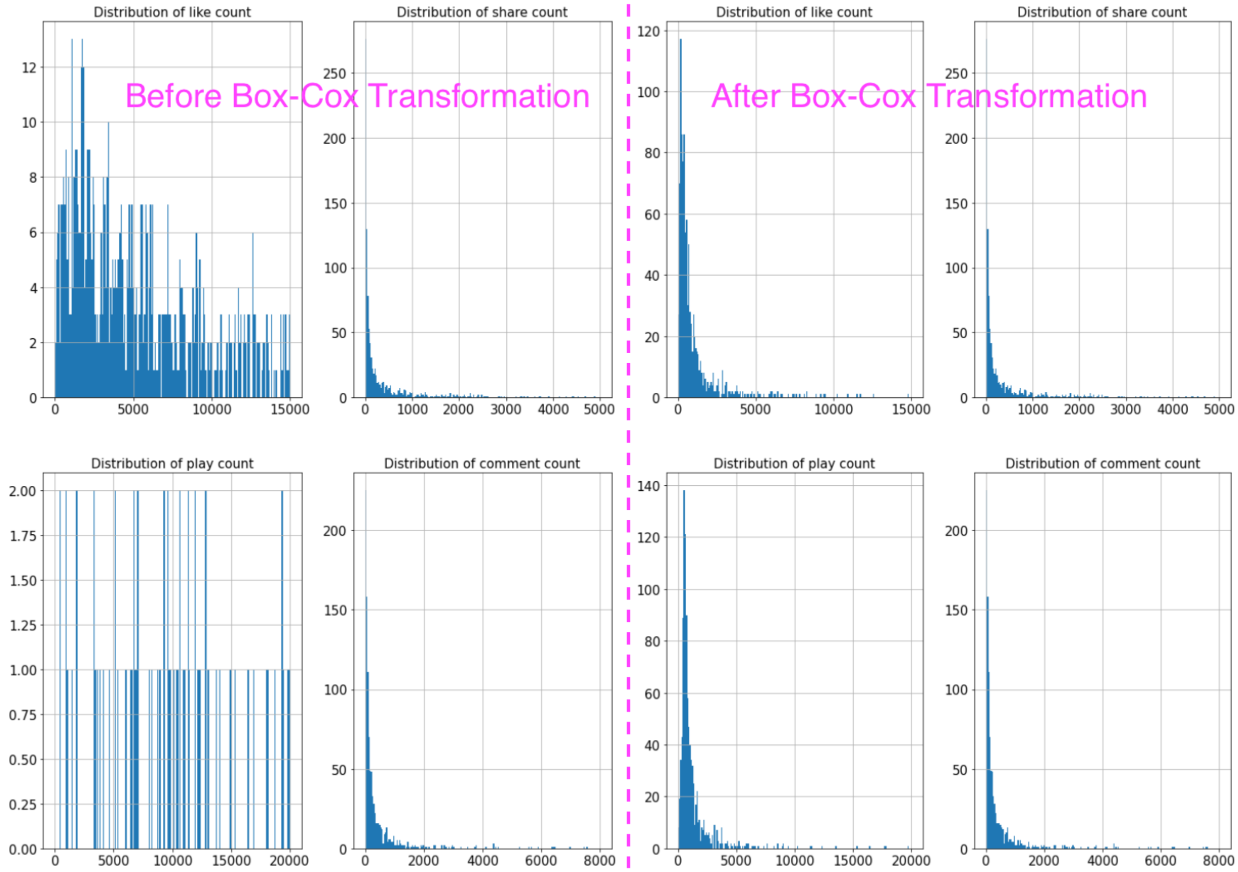


Figure 18 & 19: left - user engagement before Box-Cox transformation; right - user engagement after Box-Cox transformation

The correlation between user engagement and play count is shown below. It is found that play count ("playCount") is highly correlated with like count ("diggCount"), share count ("shareCount"), and comment count ("commentCount"). The correlation coefficient between play count and share count is 0.77. The correlation coefficient between play count and like count is 0.98. The correlation between play count and comment count is 0.91. Meanwhile, we could also see significant correlation between share count and like count (0.75), share count and comment count (0.69), like count and comment count (0.97).

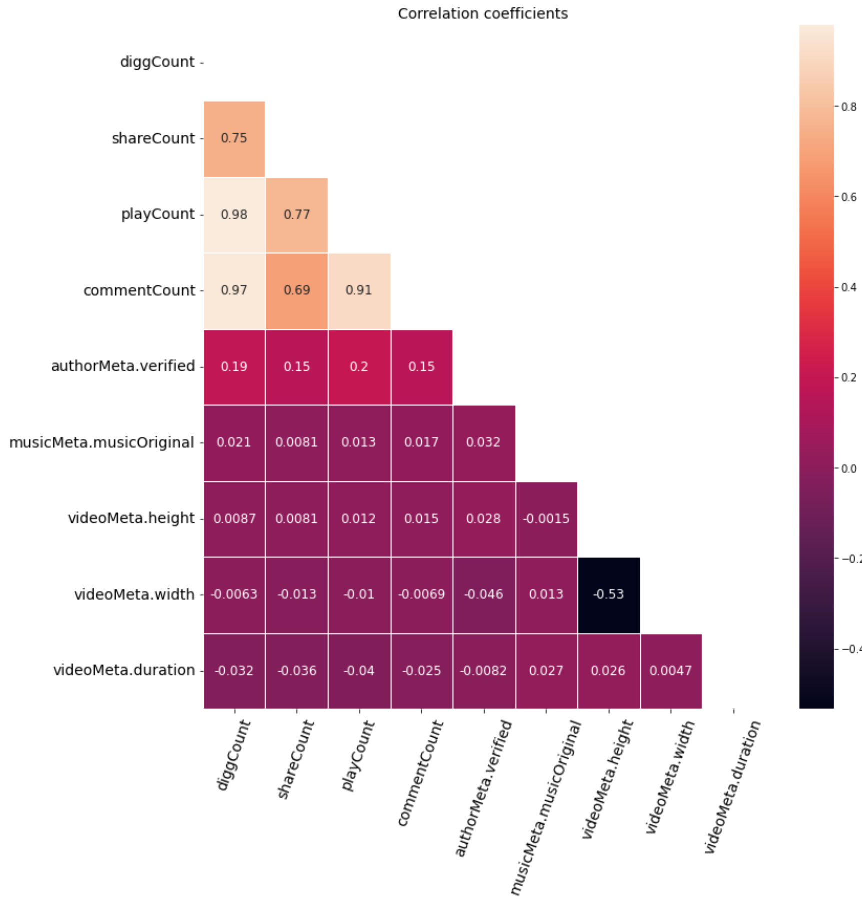
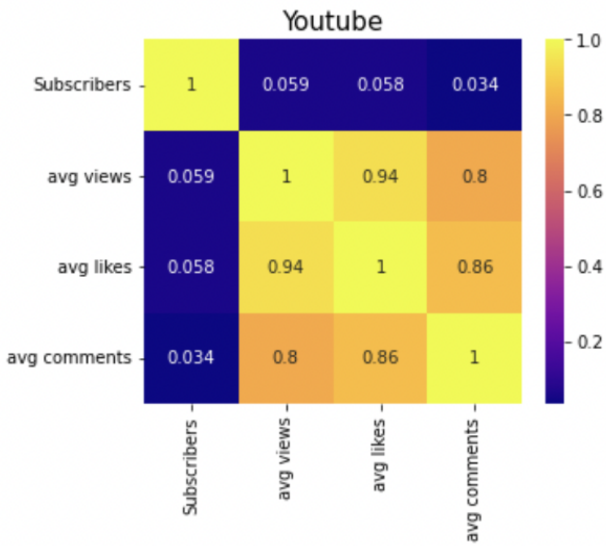


Figure 20: correlation coefficient matrix for trending video's numerical features

Dataset 3 - Social Media Influencer

The three data sets (youtube, instagram, and tiktok) were skewed such that they don't have a normal distribution. To fix that, we applied the Boxcox transformation to the data set. The transformation transforms the data such that the regression residual stays the same [4]. A normal distribution is essential as it allows us to assume that the error is normally distributed. Given this assumption, we can later construct confidence intervals and conduct hypothesis testing. However, for this project, we use the boxcox function in the scipy library to transform the data. But no further step is necessary since we are only interested in having a normal distribution.



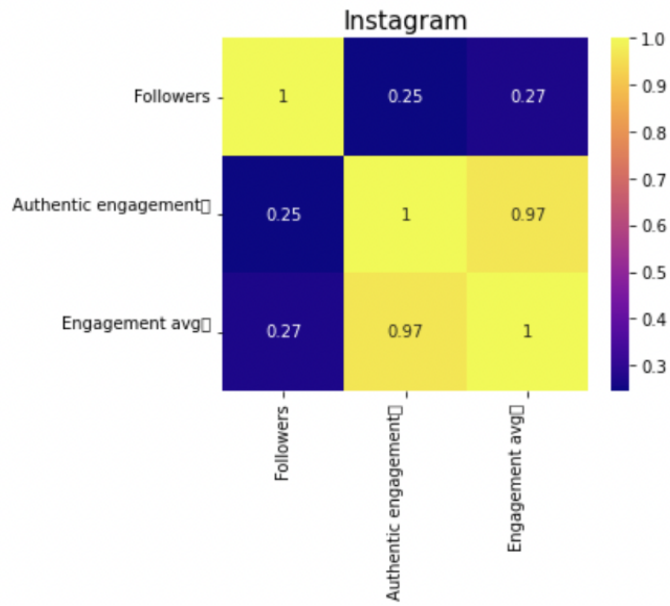


Figure 21 & 22: These are the Pearson correlation matrix for youtube and Instagram.

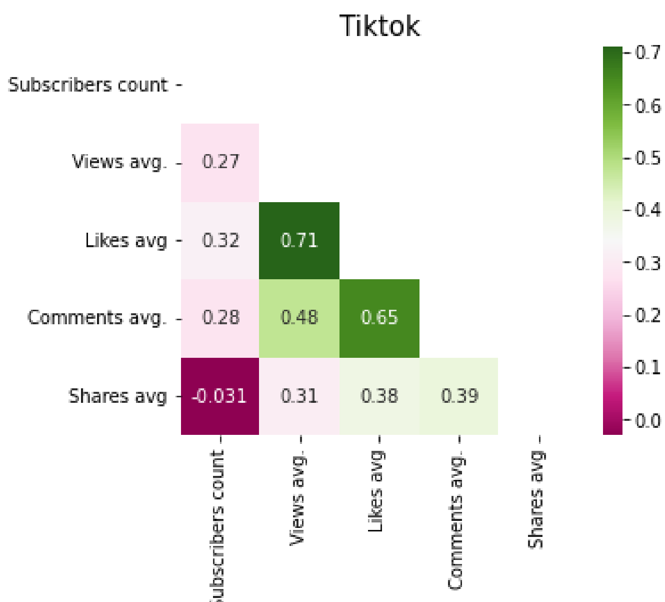


Figure 23: This figure is the Pearson correlation matrix of the Tiktok dataset.

Figures 21,22 and 23 above are the correlation matrix of Youtube, Instagram, and Tiktok. These matrices display the correlation between each feature without grouping by genre. The average views, likes, and comments for youtube data are strongly correlated, given that their correlation coefficients are higher than 0.8. For Instagram data, the authentic engagement and engagement avg features are strongly correlated, given their correlation coefficient is 0.97. Across the three platforms, social engagements (such as likes, views, comments, and shares) and subscribers/followers are weakly correlated. This indicates that we cannot use social engagement to predict the influencer's subscribers/followers.

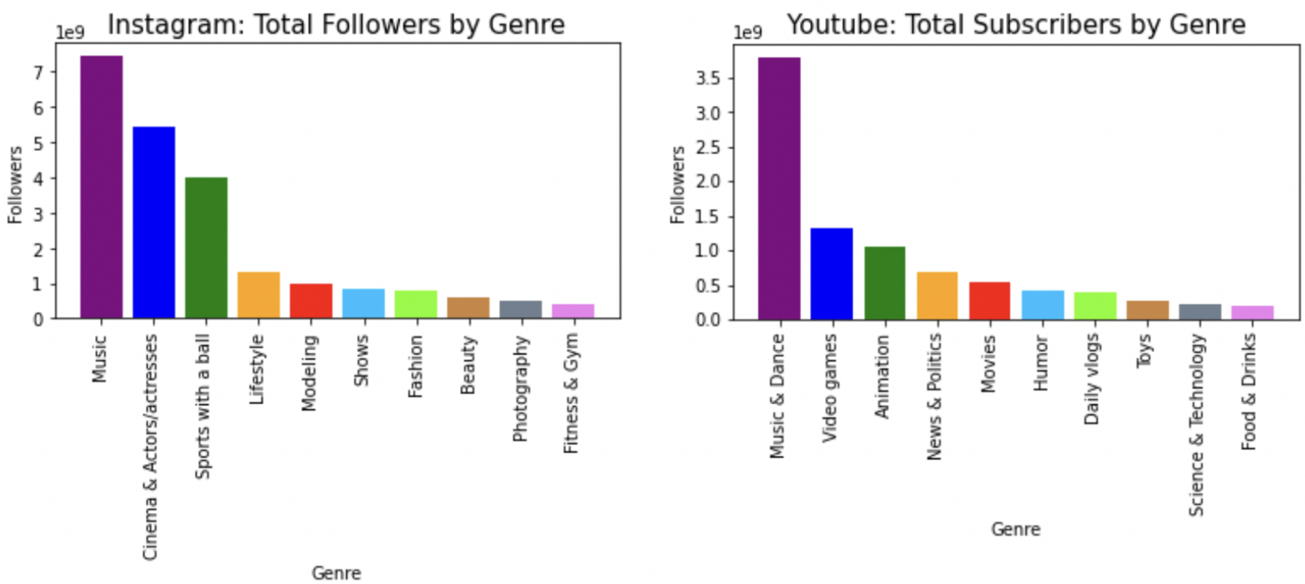


Figure 24: The total number of followers and subscribers by genre.

The Music and Music & Dance category appears to be the most popular on Instagram and Youtube. Influencers in the Music genre have a total of 7,438,299,997 followers, and influencers in the Music & Dance genre have a total of 3,797,100,000 subscribers. The total number of followers was calculated by taking the sum of all followers for each genre.

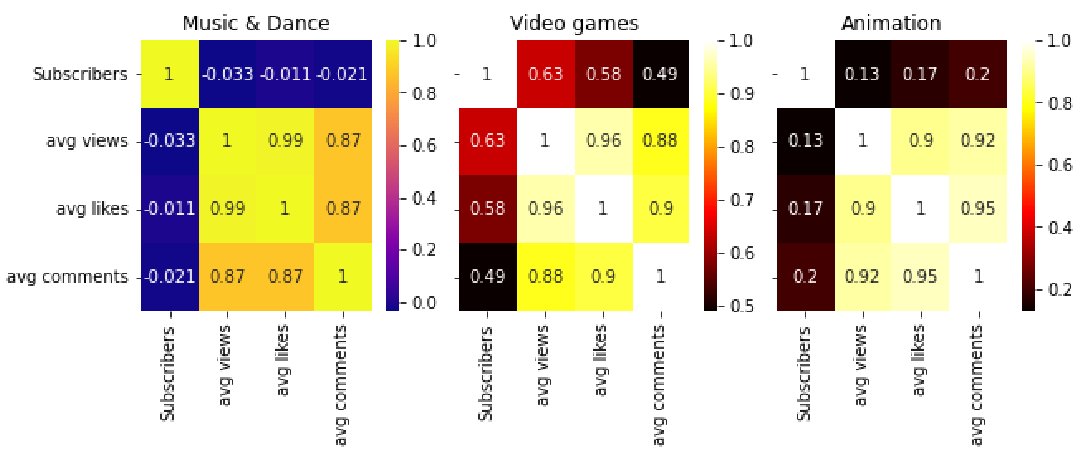


Figure 25: The correlation matrix of Youtube data frame for Music & Dance, Video games, and Animation.

The correlation matrix of the Youtube data frame for the top 3 genres is shown in figure 25. According to the figure above, the followers and social engagement of the Video games genre have a somewhat high correlation coefficient compared to the Music& Dance and Animation genre.

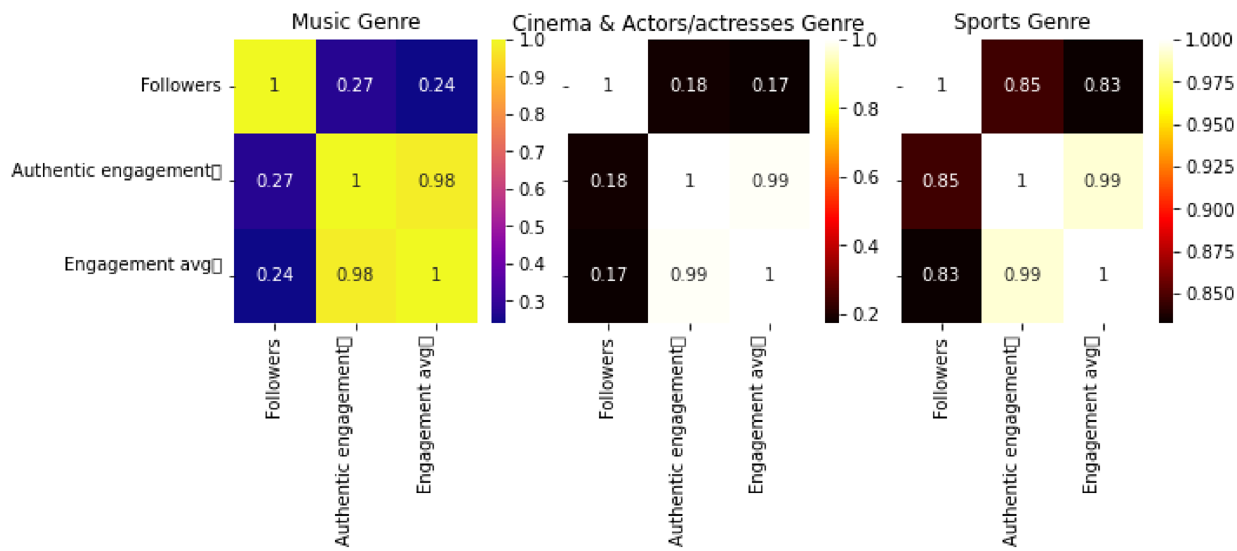


Figure 26: The correlation matrix of Instagram data frame for Music, Cinema & Actors/Actresses, and sport genre.

The correlation matrix of the Instagram data for Music, Cinema & Actors/actresses, and Sports genres is shown in figure 26. Similar to the Youtube genre in figure 25, some genres have a higher overall correlation between social engagement and followers. In this case, Authentic engagement and Followers features have a correlation coefficient of 0.83 in the sports genre. Similarly, Engagement avg and Followers have a correlation coefficient of 0.85. This indicates that people are inclined to follow/subscribe, like, share, and comment on specific genres more than others.

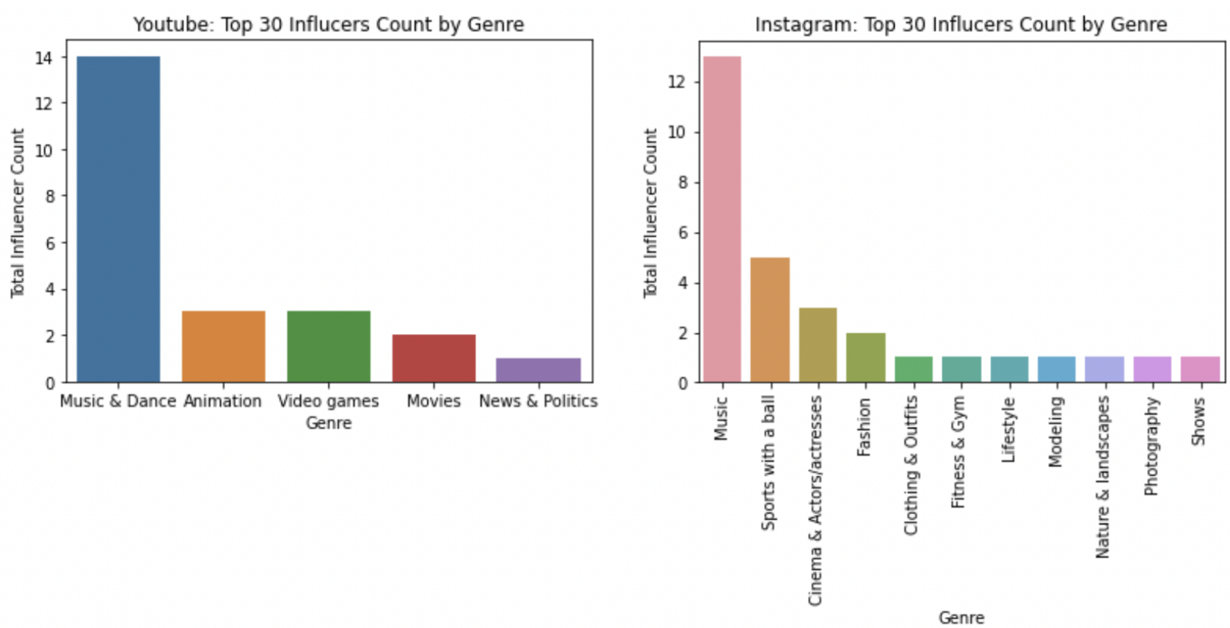


Figure 27: The distribution of the 30 most popular influencers (by followers/subscribers) by genre.

Among the 30 most popular influencers on the Youtube platform, 14 of the influencers are in the Music & Dance genre. Of the 30 most popular influencers, 13 influencers on the Instagram platform are in the music category. Furthermore, figure 27 shows that the top 30 Youtubers are classified within a small range of genres (a total of 5). In contrast, the top 30 influencers on Instagram consist of a broader range of genres(a total of 11).

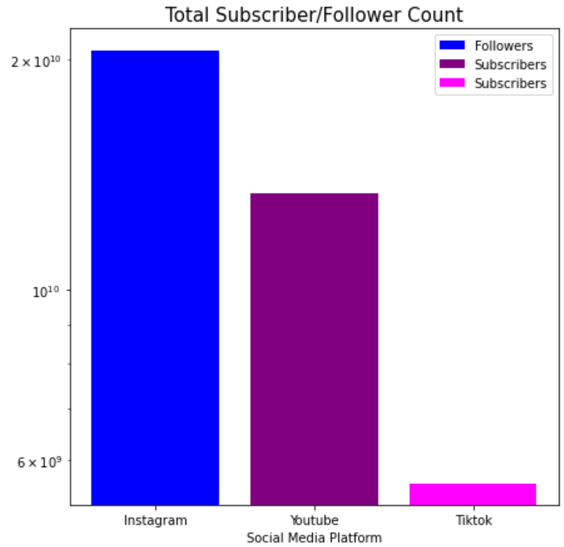


Figure 28: The total count of the followers/subscribers for the top 786 influencers.

The removal of duplicates and null-values of numerical columns reduces the data size of youtube from 1000 to 786. For consistency, the total of followers and subscribers for the top 786 influencers across Youtube, Instagram, and Tiktok were calculated. According to figure 28, Instagram is the most popular, with roughly 20B followers.

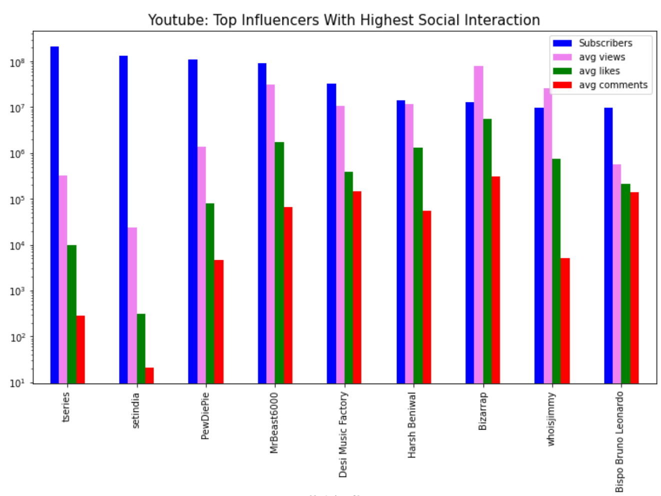


Figure 29: Influencers with highest subscribers and social engagement.

Influencers with the highest subscribers will not always have the highest engagement count. Figure 29 shows that tseries may have the highest followers, but his social engagement is lower than Bizarrap. This same trend appears for the other two social media platforms as well. Influencers with high followers will not always have the highest total social engagement count.

Conclusions

For trending tracks, there is no direct correlation between the track's features (either numerical or categorical) and popularity. We found "TIKTOK DANCE" is the most popular genre. Most trending tracks are released on the 1st day in a month. And there are 68 trending tracks released one day before Independence Day in 2020. For trending videos, we observed significant correlation between like/comment/share and popularity. The most popular hashtag is "fyp". Most trending videos are created on the 9th day in a month.

Instagram is the most popular platform based on the total number of followers and subscribers. The second most popular is youtube, followed by TikTok in last place. Music and Music & Dance is the most dominant category on Instagram and Youtube, respectively. There is no direct correlation between the overall social engagement (like, view, share, comments, authentic engagement, and engagement avg) and followers or subscribers. This indicates that other external factors influence this outcome. One possible reason is that people will socially engage with a particular genre more than others. Additionally, people are inclined to follow or subscribe to a specific genre more than others without having to engage socially in that said genre. The correlation matrix in Figures 25 and 26 further reinforces this analysis, given that the video game and sports genres tend to have higher correlation coefficients between different features of social engagement and its followers/subscribers.

Reference

TikTok Trending Track:
[1] .Edward. "Top Tiktok Tracks." *Kaggle*, Kaggle, 20 Apr. 2022, https://www.kaggle.com/code/eharian1/top-tiktok-tracks/data

TikTok Trending Video:
[2] Ven, Erik van de. "TikTok Trending Videos." *Kaggle*, Kaggle, 27 Mar. 2021, https://www.kaggle.com/datasets/erikvdven/tiktok-trending-december-2020

Social Media Influencers:
[3] Maurya, Ram Jas. "Social Media Influencers." *Kaggle*, 25 June 2022, https://www.kaggle.com/datasets/ramjasmaurya/top-1000-social-media-channels

Box-Cox Transformation:
[4] Plummer, Andrew. "Box-Cox Transformation: Explained." *Medium*, Towards Data Science, 3 Oct. 2021, https://towardsdatascience.com/box-cox-transformation-explained-51d745e34203