Factors Affecting Number of Views on YouTube Videos[*]

An Exploration of Official Music Videos uploaded by Genierock in 2015

Samita Prabhasavat

01 May 2022

Abstract

Both domestically and internationally, Thailand's entertainment and media sector is expanding. This report is based on information obtained from Genie Records' marketing department. The data is cleaned, altered, and investigated, yielding solid proof that both the time and the song itself have an impact on the number of views on a music video. The impact of the time and the song on the number of views on the music video will be discussed in this paper.

keywords: music, YouTube, views, date, Thai entertainment and media industry

1 Introduction

GMM Grammy Public Company Limited is Thailand's largest media and entertainment conglomerate. It controls 70 percent of the Thai entertainment market. The organisation has 15 music subsidiaries that specialise in various kinds of music. Genie Records, founded in 1998, is a Thai record company and a subsidiary of GMM Grammy that specialises in rock music. Klear, Cocktail, and Bodyslam are among the 23 Thai artists in the company (Grammy 2017).

The data for this report was gathered by Genie Records' marketing department with the initial goal of increasing engagement on official music videos posted on YouTube. From 2015 through 2021, the dataset includes the date the video was uploaded, the name of the band, the name of the song, and the number of views tallied in each month. The goal of this analysis is to look into the aspects that may influence how many people watch each music video. The major criteria in this report will be the time since the song was released and the song and the artist.

The data was received from Genie Records' marketing department and was altered to ensure that it was in the correct format. The data was then analysed, leading to the conclusion that the time since the song was released as well as the song have a significant impact on the number of views. In this work, a prediction model is proposed to further investigate factors affecting the number of views and provide insights for the marketing team. Understanding these elements can aid us in comprehending the effects of events that occurred during that time period, as well as the direction of the Thai entertainment industry, which contributes significantly to the Thai economy.

This report will first, look at the data collected by the marketing department of Genie Records and the methodology used to collect the data in Section 2 to provide readers with the context of the data used in this report. Second, it will discuss the insights found from exploring the data in Section 3. Last, it will conclude the insights found from exploring this data set and provide plans for further exploration of this data set in Section 4.

[*] Code and data are available at: https://github.com/PSamita/music_video.git

2 Data

The R statistical programming language (R Core Team 2020) is used to analyse the data in this report. The here package (Müller 2020) is used to access files from another location in the same R Project. The janitor package (Firke 2021) is used to tidy up the column names. Data manipulation is done with the tidyverse package (Wickham et al. 2019). To make a table from a data frame, use the knitr package (Xie 2014). Graphs are made with the ggplot2 package (Wickham 2016). The RColorBrewer (Neuwirth 2014) is used to assign colours to graphs and figures. (Xie 2016) The bookdown package is used to cross-reference graphs and figures.

2.1 Overview of the data

The marketing department of GMM Grammy Public Company Limited provided the data for this study. The data was collected in order to increase the engagement of official music videos released by artists signed to the GMM Grammy label and its sublabels. The dataset includes the number of monthly views of four official music videos, the title of the song, and the name of the band collected from 2015 to 2020.

2.1.1 Data Sources

GMM Grammy Public Company Limited is the largest media conglomerate and entertainment company in Thailand and is home to many of Thailand's top artists, such as Bird Thongchai, Bie Sukrit, Tata Young, and Bodyslam. The company is also active in concert production, artist management, film and television production, and publishing, in addition to music. GMM Grammy has 15 music subsidiaries focusing on different styles of music, including Genie Records, Grammy Gold, White Music, Musiccream, and Genelab.

Genie Records is a Thai record label that concentrates on rock music and is a subsidiary of GMM Grammy. Genie Records was founded on January 1, 1998, by GMM Grammy, and Wichian Rerkpaisan was appointed the company's first managing director. Sumet and the Pang, a musical duo who became popular with their first album, was one of the first artists released by the record label, which covered a variety of music genres. It also introduced new performers to the entertainment industry, like Tanawut Chawathanavorakul, Palaphol Pholkongseng, Noppasin Sangsuwan, Budokan, Venus, and Dajim, who went on to make hit songs. It went up against More Music, the country's premier rock music record label at the time, which featured experienced rock performers like Asanee–Wasan, Loso, Nuvo, and others.

The record label was rebranded "Genie Rock" in 2012 in order to concentrate on producing rock songs. Labanoon, The Yers, Palmy, 25 Hours, Paper Planes, and Potato, who returned to Genie Records after a brief stay on We Records and WerkGang, were among the new and returning rock musicians. As of 2022, there are 23 artists under Genie Records, including 7 soloists and 16 bands (Grammy 2017).

Four songs by four bands under Genie Records were chosen to be used in this report. The first song is 'Leave' (Thai: Ting Wai Glang Tang) by Potato. The second song is 'Soulmate' (Thai: Koo Chee Wit) by Cocktail. The third song is 'Object' (Thai: Sing Khong) by Klear. The last song is 'Never' (Thai: Mai Keay) by 25 Hours. The music videos for these songs were released on Genierock's official YouTube channel.

Potato is a Thai rock band that was established in 2001. Patchai Pukdesusook, guitarist Teekatas Taviarayakul, bassist Piyawat Anukul, drummer Kan Uamsupan, and pianist Kiattiyos Malathong make up the band's current lineup. It has eight studio albums, one extended play, and a number of compilation albums to its credit. It is regarded as one of Thailand's most successful rock bands. The chosen song, "Leave," was released in 2015 but was only included in a Chudteejed album in 2019. The song is about a couple's split, and how it's intended to be this way, and that it's best for both of them to break up.

Cocktail is a Thai rock band, founded in 2002. The band currently has four members, including lead vocalist Panthaphon Prasarnrakit, guitarist Chawarat Hadsunthathai, bassist Krirkkiat Sawangwong, and drummer Philips Prem Sirikorn. It has released seven studio albums, one extended play, and several compilation albums. It is considered one of the most successful Thai rock bands, with lots of hit songs. The chosen song, “Soulmate,” was released in 2015 but was first heard by the public at the lead vocalist Panthaphon’s wedding in 2012, as it was the song he wrote for his wife. The song expresses the feeling of how much the song writer loves his lover and that he is willing to do anything for her.

Klear is a Thai rock band, founded in 2007. The band currently has four members, including lead vocalist Rannapun Yungyeunpoonchai, guitarist Nathawat Sangvichit, bassist Keyapatr Sangwonpailert, and drummer Nut Nilvichian. It has released six studio albums and one extended play. It is considered one of the most successful Thai rock bands, with lots of hit songs. The chosen song, “Object,” was released in 2015 on the Silver Lining album. The song expresses the feelings of a person who has been unloved by everyone in his or her life until that person gets to meet someone who really loves him or her, and the person feels very thankful for that love.

25hours is a Thai rock band, founded in 2008. The band currently has five members, including lead vocalist Somphol Rungpanich, guitarists Prateep Siri-issarnunt and Piyawat Mekreu, bassist Aeksiri Kumbungpai, and drummer Krittapong Sakulnarmanake. It has released four studio albums. The chosen song, “Never,” was released in 2015 on the Mom & Popshop album. The song expresses the feelings of missing someone who is not in your life anymore.

2.1.2 Data Collection Methodology

The marketing department of the GMM Grammy, including the data scientists and data analysts, collected the view count provided on YouTube Analytics. The number of views was recorded at the end of each month after the song was released. Normally, the marketing team will analyze the number of views along with data collected from the comments under each music video and data collected from other social media platforms.

On YouTube, a view is defined as: 1) a viewer-initiated intended play; and 2) a video that has been despammed by YouTube’s algorithm. To put it another way, the video was viewed by a human on a single device, and YouTube acknowledged it.

2.1.3 Error Detection

The data was already reviewed and cleaned by the data team because it was acquired by the Marketing Department for business purposes. As a result, there is no error detected in this dataset as expected. There are some null values in the dataset, which is because the music video has not yet been uploaded on YouTube, so it does not have a number of views.

2.1.4 Imputation

There are null values in the dataset which is not because the data was not collected. At the time when the data was collected, some music videos are not yet uploaded so the number of views cannot be counted and results in null value in the dataset. All null values are replaced by 0 indicating that there are zero view for that music video at a certain time.

2.1.5 Disclosure Control

The Marketing Department of GMM Grammy contributed this dataset only for the purpose of this study. For business purposes, no additional information will be shared or released to the public.

2.2 Variables

There are five variables in the original dataset:

- Date
  - The date the number of views was recorded.
- Song
  - The name of the song in Thai.
- English Name
  - The name of the song in English.
- Band
  - The name of the band.
- Views
  - The number of views

2.3 Strengths and Weaknesses

The strength of this dataset is the quality of the data. The data in this dataset is very accurate and clean because it was collected and cleaned by data analysts and data scientists in the Marketing Department. The dataset is extracted from the real dataset the Marketing Department uses to make business decisions and increase engagement.

The fact that the data cannot be traced back is a weakness in this dataset. Because the information is not available to the public, it is impossible to verify that the numbers of views are correct. However, if applicable, the data's validity, outliers, and null values can still be recognised using statistical methods.

2.4 Focused Aspects

This paper will focus on the numbers of views on four music videos uploaded on the Genierock channel from 2015 to 2020 by month.

2.4.1 Data Exploration

Table 1 shows the example of cleaned data generated from the raw data provided by the Marketing Department of GMM Grammy. The table includes 5 variables which are year, month, song, band, and views.

Table 1: Number of Views on the Official Music Videos from 2015 to 2020

Year	Month	Song	Band	Views
2015	03	Leave	Potato	NA
2015	03	Soulmate	Cocktail	22085749
2015	03	Object	Klear	NA
2015	03	Never	25hours	NA
2015	04	Leave	Potato	NA
2015	04	Soulmate	Cocktail	51963702

The line graphs in Figure 1 represent the data provided by the Marketing Department of GMM Grammy which is used in this report to further explore the factors affecting the number of views on the official music videos. The y-axis shows the number of views counted on official music videos uploaded on YouTube while the x-axis shows the year between 2015 and 2020. Each line graph shows the trend of the number of views for each song.

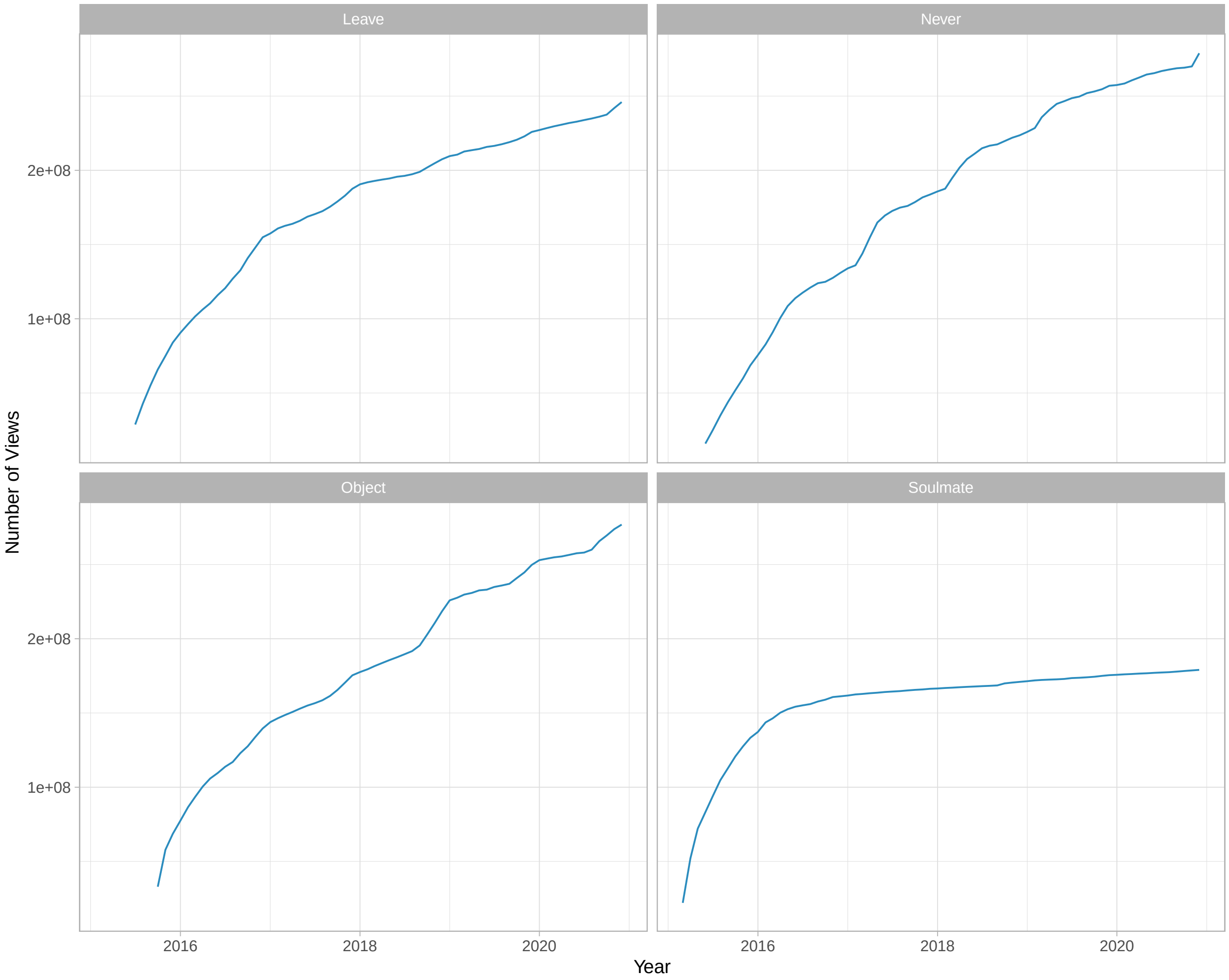

Figure 1: Number of Views from 2015 to 2020 by Song

A box plot is used to assess the data's normality before selecting the suitable statistical summary for this dataset. The data points in this dataset lie between the upper thin horizontal line and the bottom thin horizontal line, according to a box plot in Figure 2. The median number of views from this dataset is represented by a thick black horizontal line across the blue box, with 50% of the data above the median and 50% of the data below the median. Within the blue box, 25% of the data is above the median, while the remaining 25% is below the median. The thick black horizontal line, or the median line, in this box plot is a little low, indicating a possible skew in the data. Furthermore, six outliers were found in this sample. However, it is confirmed that they are all correct, so all of them are not excluded from the dataset.

In order to further confirm the non-normality of the data from looking at the box plot above, a histogram is plotted as in Figure 3. It can be seen that the histogram is left skewed and does not look like a typical bell-shaped diagram, which indicates that the data is not normally distributed.

The median has been chosen as a measure of central tendency for exploring this dataset as the data is not normally distributed.

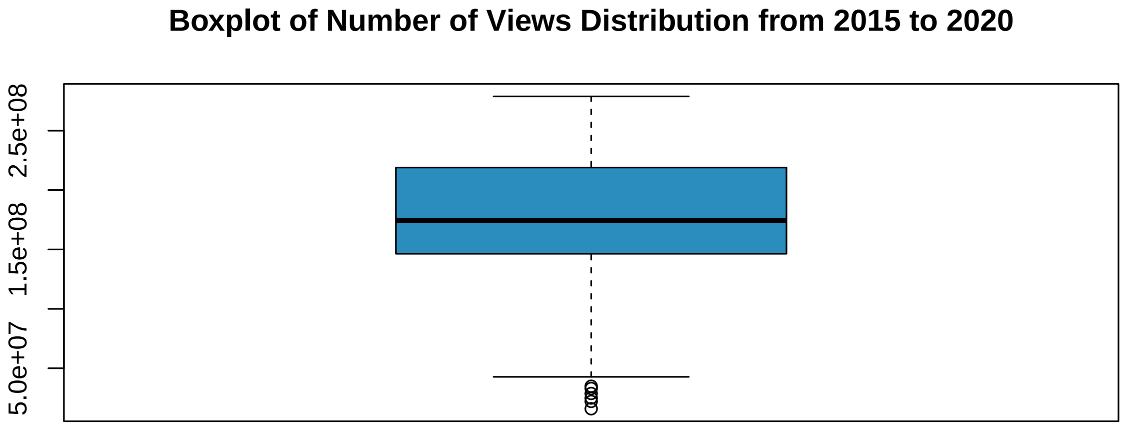

Figure 2: Boxplot of Number of Views Distribution from 2015 to 2020

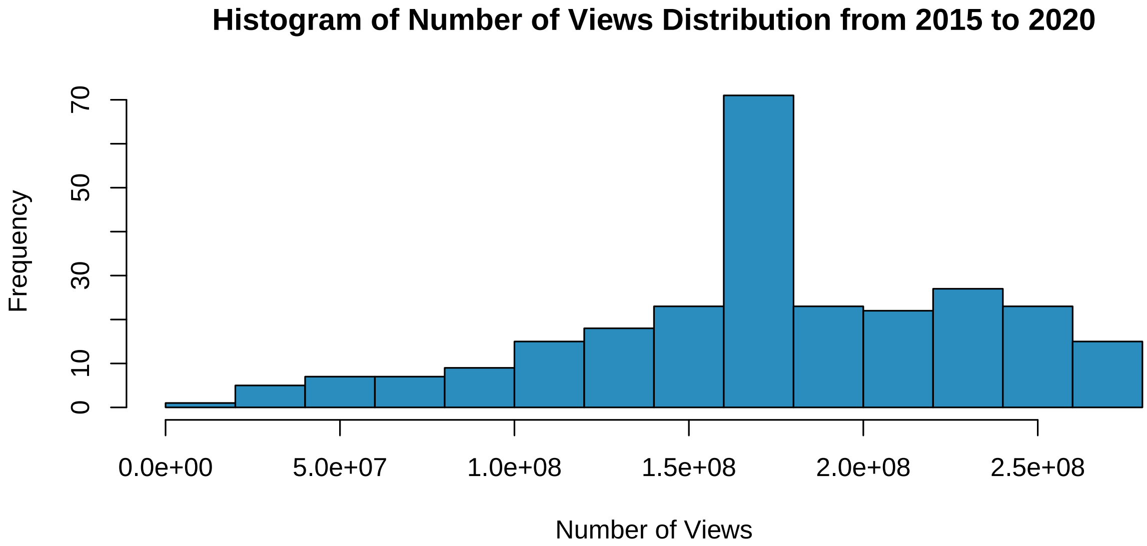

Figure 3: Histogram of Number of Views Distribution from 2015 to 2020

3 Results

3.1 Number of Views by Year

A bar graph in Figure 4 shows the median number of views on four chosen music videos uploaded on Genierock's official YouTube channel by year. It can be seen that the median number of views increased every year from 2015 to 2020. In 2020, the median number of views is the highest among the six years with around 250 million views, followed by 2019, 2018, 2017, 2016, and 2015 at around 50 million views, in that order.

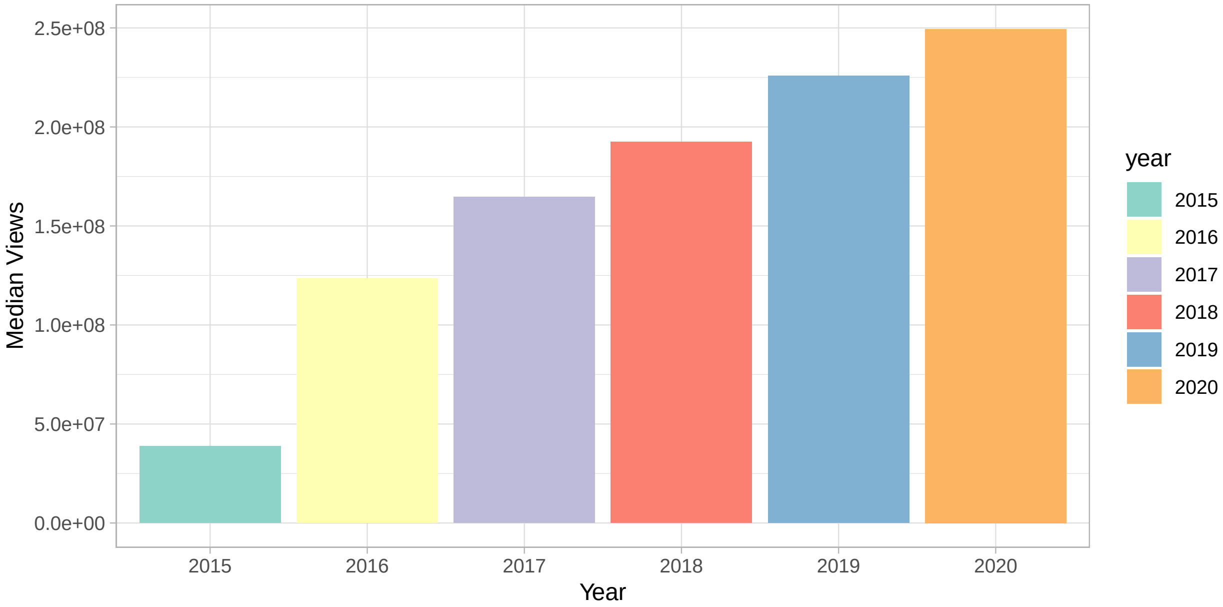

Figure 4: Number of Views by Year from 2015 to 2020

3.2 Number of Views by Month

A bar graph in Figure 5 shows the median number of views on four chosen music videos uploaded on Genierock's official YouTube channel by month. It can be seen that the median number of views hit the peak around the end of the year to the beginning of the year. The median number of views in March is the lowest. The median number of views increases from March until reaches its peak at the end of the year and remain high until February. However, the median number of views for every month are still within the range of 150 million to 200 million views.

3.3 Number of Views by Song

A bar graph in Figure 6 shows the median number of views on four chosen music videos uploaded on Genierock's official YouTube channel by song. It can be seen that, between 2015 and 2020, 'Leave' has the highest median number of views compared to the other three songs at almost 200 million views. 'Never' has the second highest median number of views at about 180 million views. The third highest median view belongs to 'Object' at around 175 million views. 'Soulmate' has the lowest median number of views at approximately 160 million views.

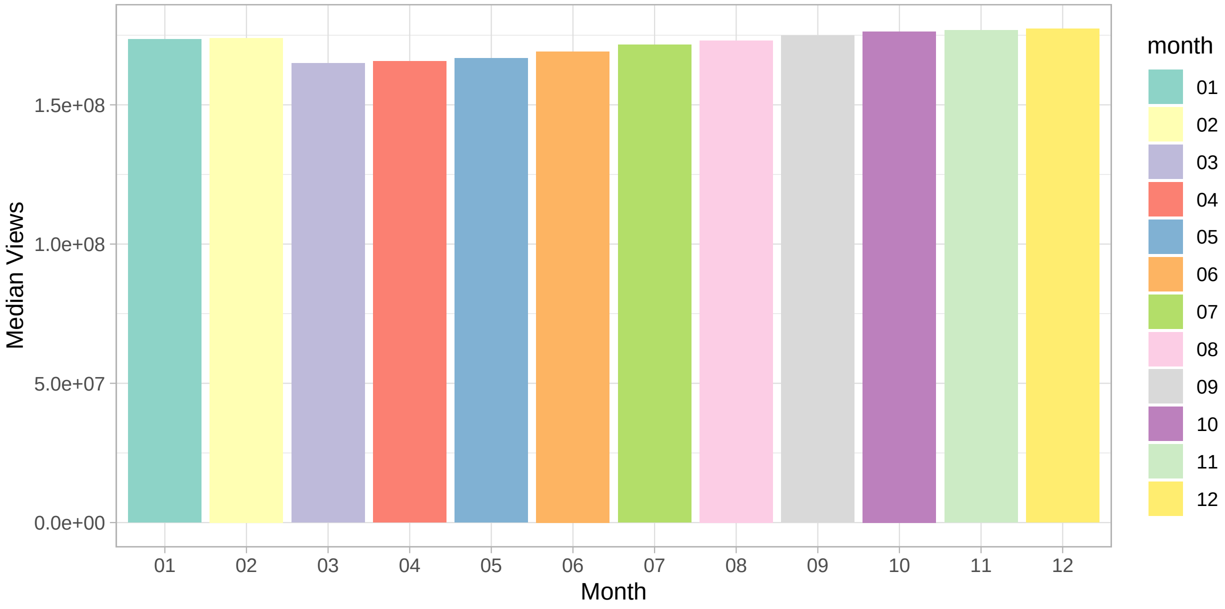

Figure 5: Number of Views by Month from 2015 to 2020

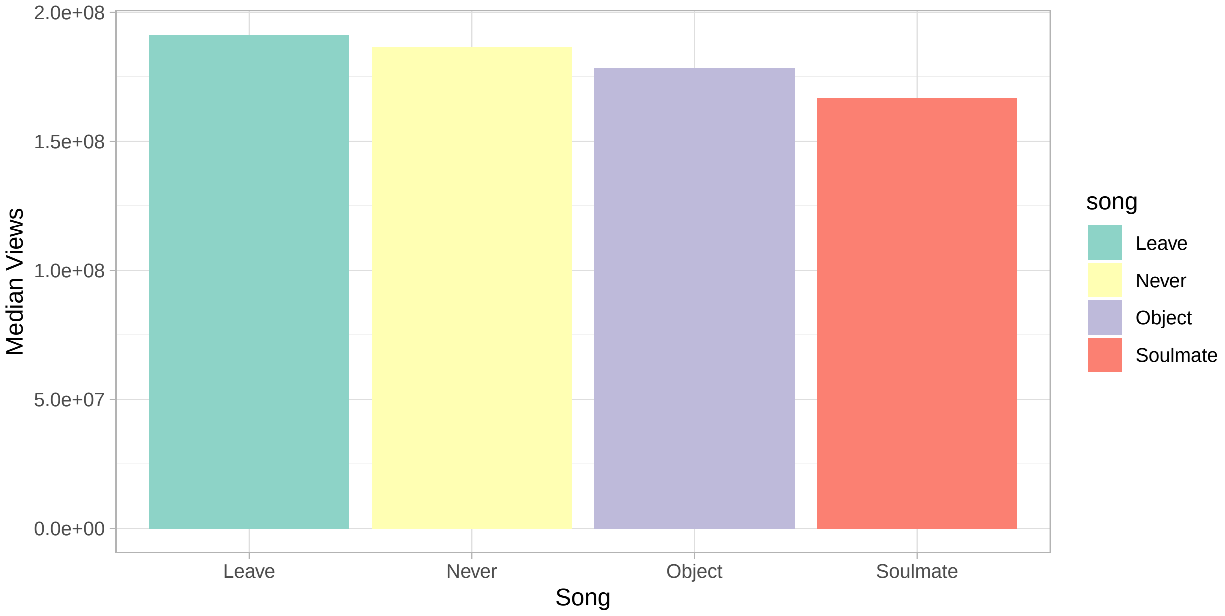

Figure 6: Number of Views by Song from 2015 to 2020

3.4 Number of Views by Year and Song

A bar graph in Figure 7 shows the median number of views on four chosen music videos uploaded on Genierock's official YouTube channel by year and song. It can be seen that the overall trend of th median number of views is positive. In 2015, the median number of views for 'Object' is zero because the music video was uploaded in the later half of the year. Most of the view count for 'Object' in 2015 is zero which resulted the median equal to zero. The medians number of views for 'Soulmate' are the highest compared to other songs in 2015 and 2016. However, from 2017 to 2020, 'Never' has the highest medians number of views.

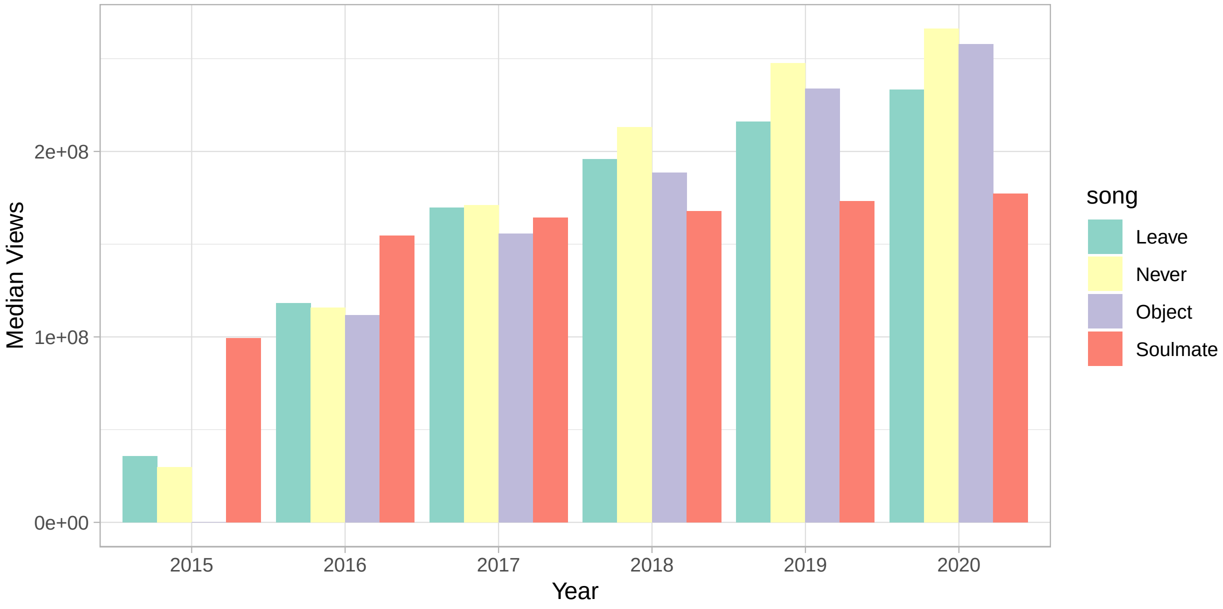

Figure 7: Number of Views by Year and Song from 2015 to 2020

3.5 Number of Views by Month and Song

A bar graph in Figure 8 shows the median number of views on four chosen music videos uploaded on Genierock's official YouTube channel by month and song. It can be seen that the medians number of views of 'Leave' are the highest compared to other songs from January to April. On the other hand, 'Never' has the highest medians number of views in the rest of the year. The medians number of views for 'Object' are high from September to February, while the medians number of views for' Soulmate' are highier than usual in February, April, June, and September to December.

4 Discussion

4.1 Methodology

This paper was created by downloading raw data provided by the Marketing Department of GMM Grammy as a csv file into RStudio software. The data was then filtered to keep only variables of interest and ensured that all data was consistent and in the right format. For example, the number of views should be in numeric format. If there is data containing characters, that data must be examined. Moreover, outliers were detected if applicable. In this case, six outliers were detected. However, after consulting with the data analyst in charge of the dataset, the data is all correct and the outliers should not be removed in this case.

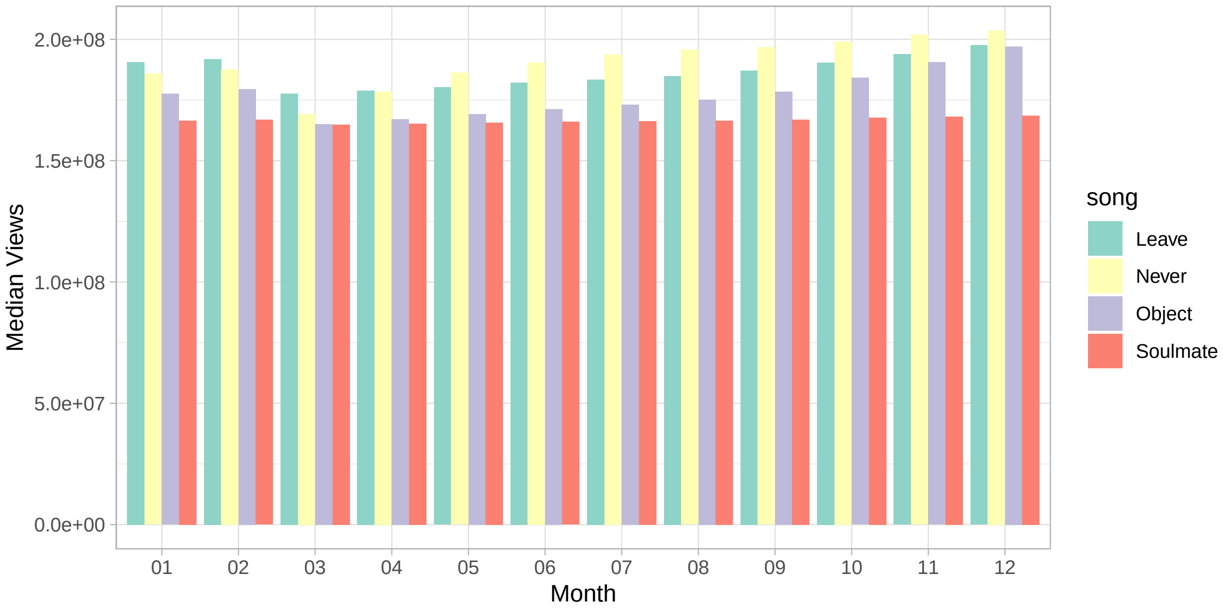

Figure 8: Number of Views by Month and Song from 2015 to 2020

After the data is ready to be used, the next step is data exploration. The normality of the data was first tested using a box plot and a histogram, which led to the decision of which statistical summary would be utilised in this study. Second, to investigate the data, a statistical summary was used. Third, graphs are created to aid in the finding of data insights. Finally, the plan for the future study is provided.

4.2 The Impact of Time on Number of Views

Figure 4 shows that the number of views increases as the year progresses. This is definite because the more people that click on the music video and watch it, the larger the amount of views will be. With each passing year, more people get access to the internet, personal computers, or mobile phones and tablets, making it easier to watch YouTube videos from anywhere. In addition, since 2017, there have been music shows in Thailand. Artists will perform the song they just released at the music show, and the artist with the most music video views combined with album sales for that week will win an award. For this reason, fans are encouraged to increase the number of views on the official music video.

It can be seen from Figure 1 that the number of views increased faster when the songs were first released. Normally the company will announce the date and time when the official music video will be uploaded on Genierock's YouTube channel. Fans will be waiting for the video to be uploaded and will watch the video as fast as possible which makes the number of views increased faster at the beginning.

From 2015 to 2016, the slope of the graph in Figure 1 for 'Object' and 'Soulmate' are very steep which means the number of views goes up very fast compared to other songs. Klear and Cocktail are loved by people in many age groups. An analysis from GMM Grammy surprisingly shows that one of the reasons why people support Klear and Cocktail, apart of their musics, is because of their educational background. Klear started their band when the members were still in university. The members of Klear attended Chulalongkorn University which has long been considered the best university in Thailand (Chulalongkorn 2021). On the other hand, Cocktail started their band back in high school years. Members of Cocktail attended Triam Udom Suksa School which has long been regarded as the best secondary school in Thailand in terms of academics. The acceptance rate of the school is around 10% which makes it the most competitive school in Thailand to get into. After graduated from high school, members of Cocktail attended either Chulalongkorn University or Thammasat University which is also a famous university in Thailand. Moreover, the lead vocalist of Cocktail attended Law School at Chulalongkorn University which is the most competitive law school in Thaialnd to get into. Since artists in Thailand have been having the image of people who are not good at studying, Klear and Cocktail, which have very good educational backgrounds, gain more popularity over other bands especially from parents who want their kids to be good at study (Grammy 2017).

Furthermore, as the year progresses and as a result of globalization, businesses will have more capacity and resources to market the music both domestically and worldwide. As a result, more individuals are visiting Genierock's YouTube account, following the company's musicians, and watching official music videos to show their support. Since 2016, GMM Grammy has been attempting to break into the worldwide market. In 2017, the good outcomes began to emerge. In China, Hong Kong, Taiwan, Indonesia, Malaysia, and the Philippines, many GMM Grammy artists have a sizable fanbase audience. Especially in Southeast Asia, where many prefer YouTube over other platforms like Apple Music or Spotify for music streaming (Sophia J. 2018).

Looking at Figure 5, the number of views on music videos is not that different between months. However, according to Figure 1, "Leave", "Object," and "Never" have fluctuations in the trend, which means there are months that people watch the music videos more than the rest of the months. According to Figure 8, there are fluctuations in the trends for every song. In Thailand, music festivals usually occur in the winter, when the weather is nice and not too hot. Winter in Thailand starts in October and ends in February. When people hear a song at a music festival, they usually go to YouTube and search for it to listen to it again because of post-concert depression, or PCD. PCD occurs when a person has a happy time, like going to a concert, which makes the body release serotonin, dopamine, endorphins, and oxytoxin. However, after the concert, the sudden stop of those chemicals makes the person feel depressed and want to go to the concert more. Because the concert just ended, people usually listen to the song they just listened to during the concert to keep the memories and feelings (Michelle S. 2017).

4.3 The Impact of Song on Number of Views

According to Figure 1 and Figure 8, there are fluctuations in the trend for every song, which are affected by the song itself. The lead vocalist first performed "Soulmate" at his wedding to express his feelings to his wife. As a result, the song has been played or performed very often at weddings. In Thailand, according to Thai norms, people normally hold their weddings in February, April, June, September, and October to December. It can be seen in Figure 8 that the number of views increases in these months. In 2019, two singers under GMM Grammy made a cover of 'Soulmate' which made the song go viral again. After listening to the cover version, people still go back to listen to the original version. Also, another singer under GMM Grammy chose 'Soulmate' to sing in "The Mask Singer." The Mask Singer is a television show where artists are invited to sing while wearing masks so that the audience does not know who is singing and will only judge the artist by his or her singing skills (Grammy 2017). The performance and the rearrangement of 'Soulmate' went viral, which made people revisit the original music video.

In 2020, the lead vocalist of Potato got married, which made people revisit the music video of 'Leave.' Even though the song is a sad song about breaking up, people still revisit the music video to congratulate the singer on his wedding as it is the most recent song the band released. Also, the bride who got married to the lead vocalist starred in one of the band's music video before. Even though it was not this song, people still see Potato's music videos as the thing that makes the two met each other and finally got married. Moreover, in 2020. the actress who stared as a female lead in the music video for 'Leave' debuted as a soloist under MBO, another subsidiary of GMM Grammy, so people came to watch the music video to see her pre-debut acting (Grammy 2017).

The song 'Object' is often performed at music festivals where people enjoy listening to both love songs and breakup songs. The lead vocalist of Klear has invented a new performing trick in 2018 where she will call a partner of one of the audience members and sing the song for him or her. The audience really love this trick as they want to tell their lovers how thankful they are to be together. This trick created a trend that people send a link to the official music video on YouTube to their partner which helps increase in the number of views. Because the song talks about a person who was never loved by anyone, but later is loved by someone, some people listen to this song when they break up with their ex-lovers too.

The song ‘Never’ is often used around the graduation time of the students because the song is about missing someone who is not in your life anymore. In Thailand, the graduation ceremonies are held in either March, April, or May. The song is used to express the feelings of the students that will have to continue their paths separately from friends in high school, but they will still miss each other even though they are not together. The song is also widely used as a background music when putting together a video when someone passed away which can happen at anytime of the year (Grammy 2017).

4.4 Weaknesses

This report was created by the data that were already cleaned and provided by the Marketing Department of GMM Grammy. As a result, it is difficult to trace back to the original raw data which gives limitations to some statistical methods that cannot be used without knowing the original data. Also, there are only 5 variables in this dataset which limits the number of aspects that can be explored in this dataset.

4.5 For the future

In order to explore more about this data, it is better to have access to the original data. In that case, more variables can be used to explore about the factors affecting the number of views on official music videos.

Also, it would be better to compute the change rate in the number of views to see how fast the number of views increases by year and month to further explore the factors.

Moreover, after exploring all the possible factors, a prediction model will be provided to predict the number of views to help making marketing decisions or creating the plan to promote the songs

Appendix

A Datasheet

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
   - The dataset was created to help GMM Grammy make more accurate and precise business decisions which will help the company to gain more profits both domestically and worldwide, and will help contributing to Thailand's economy. The dataset is not available publicly, but was provided by the Marketing Department specifically for this study. However, there are limitations on exploring the dataset because of the limited amount of data.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
   - The Marketing Department of GMM Grammy, including the data analysts and the data scientsts, created and provided the dataset for this study.
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*
   - No direct funding was received for this study.
4. *Any other comments?*
   - No.

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*
   - There are only one table in this dataset containing 281 rows of data which are all possible instances.
2. *How many instances are there in total (of each type, if appropriate)?*
   - There are 281 instaces.
3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*
   - The dataset contains all possible instances.
4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*
   - There are five variables for each instance; date, the name of the song in Thai, the name of the song in English, the name of the band, and the number of views.
5. *Is there a label or target associated with each instance? If so, please provide a description.*
   - No.

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*

   - The number of views are not available in some cases because the music video was not yet uploaded at that time.

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

   - No.

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

   - No.

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

   - No.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

    - The dataset is self-contained.

11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*

    - The data is somewhat confidential as it is used in real business but is allowed to be used in this project.

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

    - No.

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

    - No.

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

    - No.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

    - No.

16. *Any other comments?*

    - No.

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

   - The data was gathered from YouTube Analytics, and was augmented with the data gatherd from the comments section under each music video.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

   - The dataset was shared via Google Drive.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

   - The dataset is not a sample.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

   - The members of the Marketing Department of GMM Grammy.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

   - Six year, from 2015 to 2020.

6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

   - No.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

   - No.

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

   - No.

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

   - No.

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

    - No.

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*
    - No.
12. *Any other comments?*
    - No.

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*
    - Yes, cleaning the data was done.
2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*
    - I do not have access to the raw data.
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
    - R was used.
4. *Any other comments?*
    - No.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - Yes, this dataset is used by the Marketing Department of GMM Grammy to plan the promotion of the songs.
2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*
    - No.
3. *What (other) tasks could the dataset be used for?*
    - Combining with the comments on social media would be interesting
4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*
    - No.
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
    - No.
6. *Any other comments?*

- No.

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*
   - The extracted part of this dataset that is used in this paper will be available on GitHub.
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
   - GitHub
3. *When will the dataset be distributed?*
   - The dataset is available now.
4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*
   - No. MIT license.
5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*
   - None that are known.
6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*
   - None that are known.
7. *Any other comments?*
   - No.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
   - Samita Prabhasavat
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
   - samita.prabhasavat@mail.utoronto.ca
3. *Is there an erratum? If so, please provide a link or other access point.*
   - No.
4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*
   - No.
5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*

- No.

6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*

   - No.

7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*

   - Pull request on Github.

8. *Any other comments?*

   - No.

References

Chulalongkorn. 2021. "Chulalongkorn." https://www.chula.ac.th/en/.

Firke, Sam. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data.* https://CRAN.R-project.org/package=janitor.

Grammy, GMM. 2017. "GMM Grammy." http://www.gmmgrammy.com/en/index.html.

Michelle S., Christopher O., Belinda C. 2017. "Beyond Happiness: Building a Science of Discrete Positive Emotions." WASHINGTON: American Psychological Association. https://librarysearch.library.utoronto.ca/permalink/01UTORONTO_INST/fedca1/cdi_apa_psycarticlescurrent_amp_72_7_617.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files.* https://CRAN.R-project.org/package=here.

Neuwirth, Erich. 2014. *RColorBrewer: ColorBrewer Palettes.* https://CRAN.R-project.org/package=RColorBrewer.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Sophia J., Aker P., Ann W. 2018. "Streaming Music: Practices, Media, Cultures." Milton: Routledge. https://librarysearch.library.utoronto.ca/permalink/01UTORONTO_INST/fedca1/cdi_swepub_primary_oai_DiVA_org_sh_32169.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.

———. 2016. *Bookdown: Authoring Books and Technical Documents with R Markdown.* Boca Raton, Florida: Chapman; Hall/CRC. https://bookdown.org/yihui/bookdown.