Spotify on Pandemic:
Change in Trends of Music during COVID-19 Situation

[Team 7]
Peerapon Akkapusit, Wasachon Chaisirirat, Zidong Yang, Mohamed Almaazmi

1 Introduction

Life in the year 2020 has been largely influenced by the emergence of the new Coronavirus widely known as *COVID-19*. Lockdown policies are enforced in a number of countries and territories as an attempt to flatten the infection curve, and along the way, have left lingering effects on society, extensively on the social and economical sphere. The music industry, in particular, under the current circumstances in which crowds are not allowed to gather has had to adaptively change; parties could not be thrown, festivals are postponed, and live concerts are canceled—all of which direct the attention of both the providers and the consumers to streaming services even more than before [1]. Music, as we know, is often associated with emotional states of the people; in some parts of the world it is even also regarded as a kind of healing, or in a more formal form used as music therapy [2]. Now that their physical health is at stake, people are exposed to more psychosocial influence which could affect their mental states and ultimately their daily routines [3], and from there we raise the question: would this be reflected in how they are listening to music as well?

In this project, we reported our exploration into the set of data from Spotify, a music streaming service which holds quite a reputation among its peers. We provided some visualization for an overview of COVID-19 case data, reduced the dimension of our dataset, and tested our hypothesis as follows.

HYPOTHESIS

H0: There is no change in *music trends* before and after the rise of COVID-19 pandemic
H1: There is a significant change in *music trends* before and after the rise of COVID-19 pandemic

Note: We defined *music trends* as means of the audio features of the Spotify data.

2 Dataset

2.1 SPOTIFY DATASET

Audio features of more than 170,000 tracks on spotify released during 1921 - 2020 obtained from Spotify Web API. Each row includes general information of each track (e.g. spotify-generated id, name, released date) and features such as danceability, acousticness, or energy.

Courtesy of Yamaç Eren Ay on Kaggle
(https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks)

2.2 SPOTIFY WORLDWIDE DAILY SONG RANKING

Spotify's daily song rankings are obtained from Spotify Charts website (https://spotifycharts.com/regional/) using a custom crawler written in python. The data contains position on charts and number of streams for each track in the rank.

Ranking included in the analysis ranged from 8th November 2019 - 13th November 2020, consolidated into weekly ranking.

2.3 COVID-19 CASE STATISTICS

Coronavirus statistics of 218 countries and territories around the world, obtained from World-o-meter website (https://www.worldometers.info/coronavirus/). Each row contains columns such as total cases, total deaths, and more; but we are working mostly with new cases, again consolidated into weekly data.

More specifications on each dataset could be found in Appendix A.

3 Research Methodology

Our analysis is divided into 4 parts as follows.

1. **Covid-19 Data Visualization:** In order to prove our hypothesis, we first need to define a *cutting point* or where COVID-19 pandemic has started. And here we expected to find such a point from visualizing the number of new cases over time.

2. **Principal Component Analysis:** As the spotify dataset contains over 10 audio features for each track, we used PCA to reduce the dimension of the data we need to work on.

3. **Hypothesis Testing:** Here, we used t-test analysis with confidence level 0.95

4. **Linear Regression Analysis:** This step is conducted to see whether there is any relationship between number of covid cases and audio features.

Primarily, each step of the analysis is done separately for 4 countries, namely 1) Australia, 2) Brazil, 3) India and 4) the United States.

4 Results

4.1 COVID-19 DATA VISUALIZATION

From the plots in Figure 1, we tried to determine when COVID roughly started getting serious in each country, a point we will refer to as the “cutting point”. This is the point in time where the graph sharply rises, and the cluster at that point is roughly the median of the data. From this rough definition, we took the cutting points for the 4 countries to be: 31-Jul (Australia), 18-Sep (India), 7-Aug (Brazil), and 17-Jul (United States)

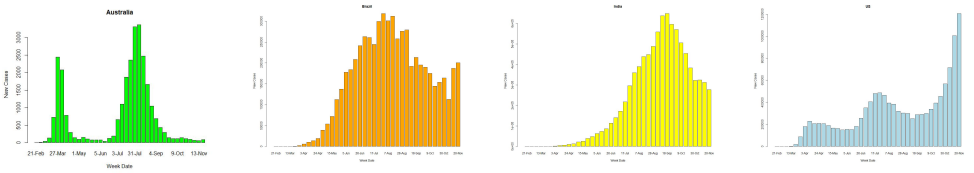

Figure 1: Barplots of COVID-19 cases over time (weekly period) of (from left to right): Australia, Brazil, India, and United States

We also looked at a boxplot showing the COVID data of the 4 countries beside each other as a percentage of their total populations in Figure 2. Here we can see that Australia & India handle COVID well as evidenced by the lower mean, while Brazil and the US seem to have badly mishandled the pandemic

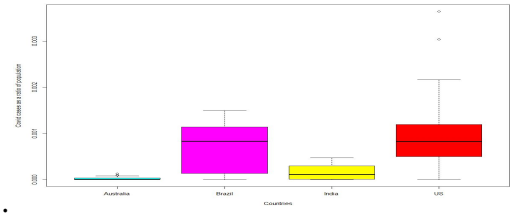

Figure 2: Boxplots of COVID-19 cases as a percentage of total population (from left to right): Australia, Brazil, India, and United States

4.2 PRINCIPAL COMPONENT ANALYSIS

11 numerical features were considered in PCA. According to the scree plot, we conducted the analysis with 4 components and retrieved the following equations in order of the proportion each explain:

$$RC1 = 0.87 \times energy + 0.85 \times loudness \quad (explain\ 0.39)$$
$$RC2 = 0.83 \times danceability \quad (explain\ 0.24)$$
$$RC3 = 0.74 \times speechiness + 0.76 \times tempo \quad (explain\ 0.21)$$
$$RC4 = 0.65 \times duration\_ms \quad (explain\ 0.17)$$

Namely, 6 features including *energy*, *loudness*, *danceability*, *speechiness*, *tempo*, and *duration* are considered as significant features. RC1 with *energy* and *loudness* explained the most proportion of the data, and thus are the most representative values. Next in the hypothesis testing, we would be using 4 columns as our features of interest: *energy, danceability, speechiness, and duration*. Noted that *loudness* and *tempo* are intentionally dropped as they are highly correlated to *energy* and *speechiness* and thus testing on them would be unnecessary.

4.3 HYPOTHESIS TESTING

To test our proposed hypothesis, we use a two sample t-test. Specifically, after determining our cutting point, we divided the music data for each country to before and after data. Thereafter, we performed a two sample t-test for each of the 4 countries and 6 features separately to test whether the mean of that feature changed after COVID for each country. The results are tabulated below:

Australia	Before	After
Danceability	0.680	0.676
Energy	0.600	0.622
Valence	0.500	0.498
Liveness	0.174	0.173
Acousticness	0.280	0.267
Speechiness	0.106	0.104

Brazil	Before	After
Danceability	0.675	0.676
Energy	0.724	0.734
Valence	0.648	0.665
Liveness	0.406	0.336
Acousticness	0.343	0.327
Speechiness	0.113	0.116

India	Before	After
Danceability	0.649	0.647
Energy	0.620	0.620
Valence	0.501	0.495
Liveness	0.167	0.158
Acousticness	0.347	0.348
Speechiness	0.085	0.091

US	Before	After
Danceability	0.722	0.695
Energy	0.588	0.608
Valence	0.485	0.464
Liveness	0.182	0.187
Acousticness	0.243	0.232
Speechiness	0.150	0.128

Figure 3: Feature means before and after COVID-19 cutting point (from left to right): Australia, Brazil, India, and United States. Highlighted rows indicate significant changes based on the t-test (p-value < 0.05)

From these tables we can see that both Australia and India don't exhibit much change in the music features (with only 1 feature significantly changing) while both Brazil and US have significant changes in more than half of the studied features (4/6 features changed significantly). We see that for the countries which handled COVID-19 badly (Brazil and the US) a lot of features changed significantly. This possibly indicates that the worse a country handles COVID, the more its music taste changes. Furthermore, upon close examination, we see that the features don't generally show the same trend direction i.e. there is no clear correlation between feature and change (increase or decrease) before and after. This possibly indicates that while COVID cases change music trends as indicated above, how this trend plays out (i.e. what features and in what direction) are dependent on other factors that vary from country to country.

4.4 LINEAR REGRESSION ANALYSIS

To check whether the music features are affected by covid new cases or not, we performed the univariate linear regressions between covid new cases and each music feature for each of 4 different countries. And we found that the feature danceability has a linear relationship with weekly new cases. Specifically, we noticed that the countries that handled COVID well, Australia and India, exhibited less correlation with danceability. In contrast, the countries that handled the virus badly had more correlation with music danceability. We found similar results for the other features although to lesser extents than for danceability.

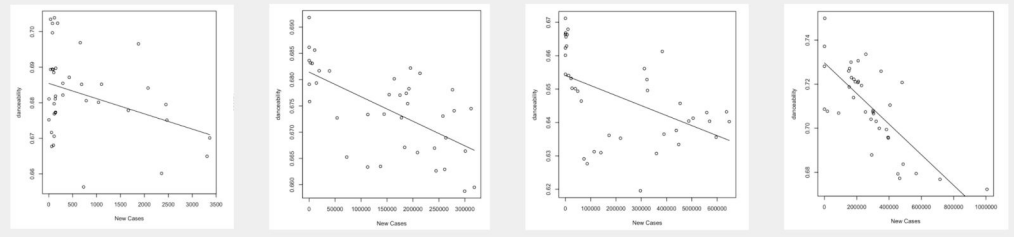

Figure 4: Linear relationship between weekly new cases and danceability for different countries (from left to right): Australia, Brazil, India, USA

5 Discussion

From the t-test results, we see that the worse a country handles COVID-19, the more change there is in the music features of the top songs in that country before and after. One possible interpretation for this is that in countries which handle COVID-19 well, there isn't much change in the daily lives of people and/or life quickly returns to normal due to the efficient handling of the virus. This small change in daily life thus leads to an equally small, mostly insignificant, change in music in that country. In contrast, countries which have not handled COVID-19 have had to deal with a lot of new COVID-19 related changes, whether it be increased family/friends sickness/deaths due to COVID-19 and/or COVID-19 related policies such as long-term quarantines and mandatory masks. This large change in people's daily lives and thoughts could have very well extended into the music taste of the people in that country. In addition, a similar discussion could also explain why countries with worse handling of COVID-19 exhibited more correlation with the music features.

Another result we notice from the t-test is that certain features may be more susceptible to COVID-19 than others. In particular, energy changed significantly for ¾ countries tested and danceability showed the most correlation with COVID-19 cases for all countries. From the table of means before and after, we can see that energy increased over the duration for all countries significantly changed (in India where the change was not significant, it remained roughly the same). A similar analysis of danceability showed a downward trend for that feature. This implies that, in general, after COVID-19, people prefer to listen to songs that are more intense and noisy but are less suitable for dancing. While reaching any specific conclusions would require more careful study that our analysis doesn't claim to do, we note that high energy low danceability songs typically refer to metal and/or rock songs which are loud and noisy but whose mood and tone are too dark to dance to. Thus, one possible interpretation is that, considering the dark atmosphere surrounding COVID-19, people's music tastes have changed to match this by increased listening to more intense yet more depressing songs such as the genre of rock/metal.

6 Limitation

6.1 DATA DIMENSION

One of the main dataset (audio features) used for analysis in this project are obtained through Spotify API, and in that we have no knowledge on the basis in which Spotify assigned values to these data. Here we are assuming that the origin of the data is reliable, though if that is not the case, our hypothesis testing could end up quite differently. Some potential representative characteristics such as track genres are also not included in the open API.

6.2 UNCONTROLLED FACTORS

Although our analysis showed some correlations between COVID-19 situation and change in trends of music, it is not guaranteed that this is a direct, exclusive cause-effect relationship. There may also be other factors—unknown and uncontrollable to us—that influence the music ranking as well, such as artists rising in popularity or contents that randomly go viral over the social network.

7 Conclusion

In this paper, we collected music-related data from Spotify API and Covid19 data with a crawler. We then visualized and consequently analyzed the data using a number of methodologies and hypothesis testing. In particular, we reached three main conclusions:

1. Trends of music as shown some change before and after COVID-19
2. The worse a country handles covid, the more (number of) features changes there
3. The better a country handles covid, the less correlated are music features and covid numbers

Lastly, based on the results of this analysis, we tried to draw a number of interpretations to try to explain the observations in the results and ended by drawing attention towards some limitations in our analysis, which largely falls on the limited access and knowledge of data and uncontrolled factors.

8 References

[1] Femke Vandenberg, Michaël Berghman & Julian Schaap. (2020). *The 'lonely raver': music livestreams during COVID-19 as a hotline to collective consciousness?*,European Societies, DOI: 10.1080/14616696.2020.1818271

[2] DeNora, T. (2013). *Music asylums : Wellbeing through music in everyday life*. ProQuest Ebook Central http://lps3.ebookcentral.proquest.com.libra.kaist.ac.kr

[3] Pfefferbaum, B. M.D., J.D., and Carol S. North. (2020, April 13). *Mental Health and the Covid-19 Pandemic.* Retrieved December 06, 2020, from https://www.nejm.org/doi/full/10.1056/NEJMp2008017

9 Appendix

APPENDIX A: DATA SPECIFICATION

Spotify Dataset

Audio features of 160k+ songs released in between 1921 and 2020. Courtesy of Yamaç Eren Ay on Kaggle (https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks)

Field	Data Type	Specification
id	Primary	
key	Categorical	All keys on octave encoded as values ranging from 0 to 11, starting on C as 0, C# as 1 and so on
artists	Categorical	List of artists mentioned
release_date	Categorical	Date of release mostly in yyyy-mm-dd format, however precision of date may vary
name	Categorical	Name of the song
acousticness	Numerical	Ranges from 0 to 1
danceability	Numerical	Ranges from 0 to 1
energy	Numerical	Ranges from 0 to 1
duration_ms	Numerical	Integer typically ranging from 200k to 300k
instrumentalness	Numerical	Ranges from 0 to 1
valence	Numerical	Ranges from 0 to 1
popularity	Numerical	Ranges from 0 to 100
tempo	Numerical	Float typically ranging from 50 to 150
liveness	Numerical	Ranges from 0 to 1
loudness	Numerical	Float typically ranging from -60 to 0
speechiness	Numerical	Ranges from 0 to 1
year	Numerical	Ranges from 1921 to 2020

mode	Dummy	0 = Minor, 1 = Major
explicit	Dummy	0 = No explicit content, 1 = Explicit content

Table A-1: Data specification of Spotify dataset

Spotify Worldwide Daily Song Ranking

Obtained from https://spotifycharts.com/regional/. Consolidated into weekly rankings

Field	Data Type
track	Categorical
position_on_chart	Numerical
streams	

Table A-2: Data specification of Spotify ranking data

COVID-19 Case Statistics

obtained from World-o-meter website (https://www.worldometers.info/coronavirus/). Consolidated into Weekly data.

Field	Data Type
country, other	Categorical
total cases	Numerical
new cases	
total deaths	
new deaths	
total recovered	
active cases	
serios, critical	
total cases / 1m population	
deaths / 1m population	
total tests	

tests / 1m population	
population	

Table A-3: Data specification of COVID-19 cases statistics.

APPENDIX B: ADDITIONAL RESULTS

Principal Component Analysis

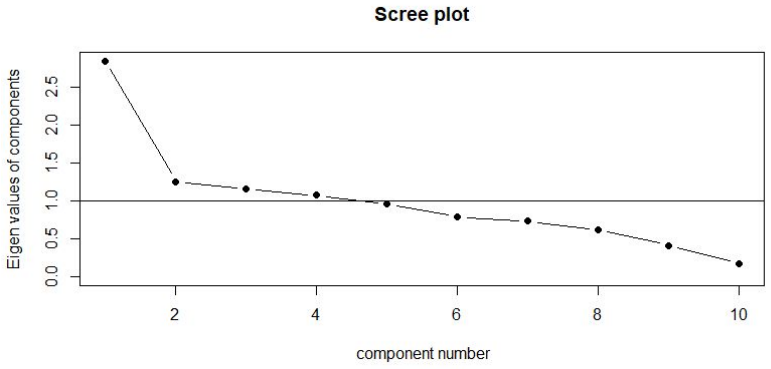

Figure B-1: Scree plot for 11 audio features of Spotify dataset

PCA Analysis	RC1	RC2	RC3	RC4
SS loadings	2.44	1.51	1.30	1.08
Proportion var	0.24	0.15	0.13	0.11
Cumulative Var	0.24	0.40	0.53	0.63
Proportion explained	0.39	0.24	0.21	0.17
Cumulative Proportion	0.39	0.62	0.83	1.00

Table B-1: PCA Analysis result