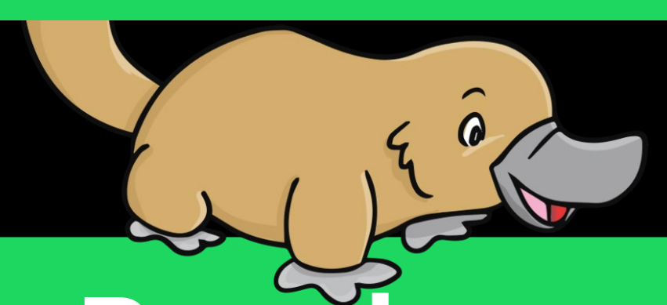
# PLATYPUS: exPLoratory dATa analYsis on sPotify's popUlar Songs



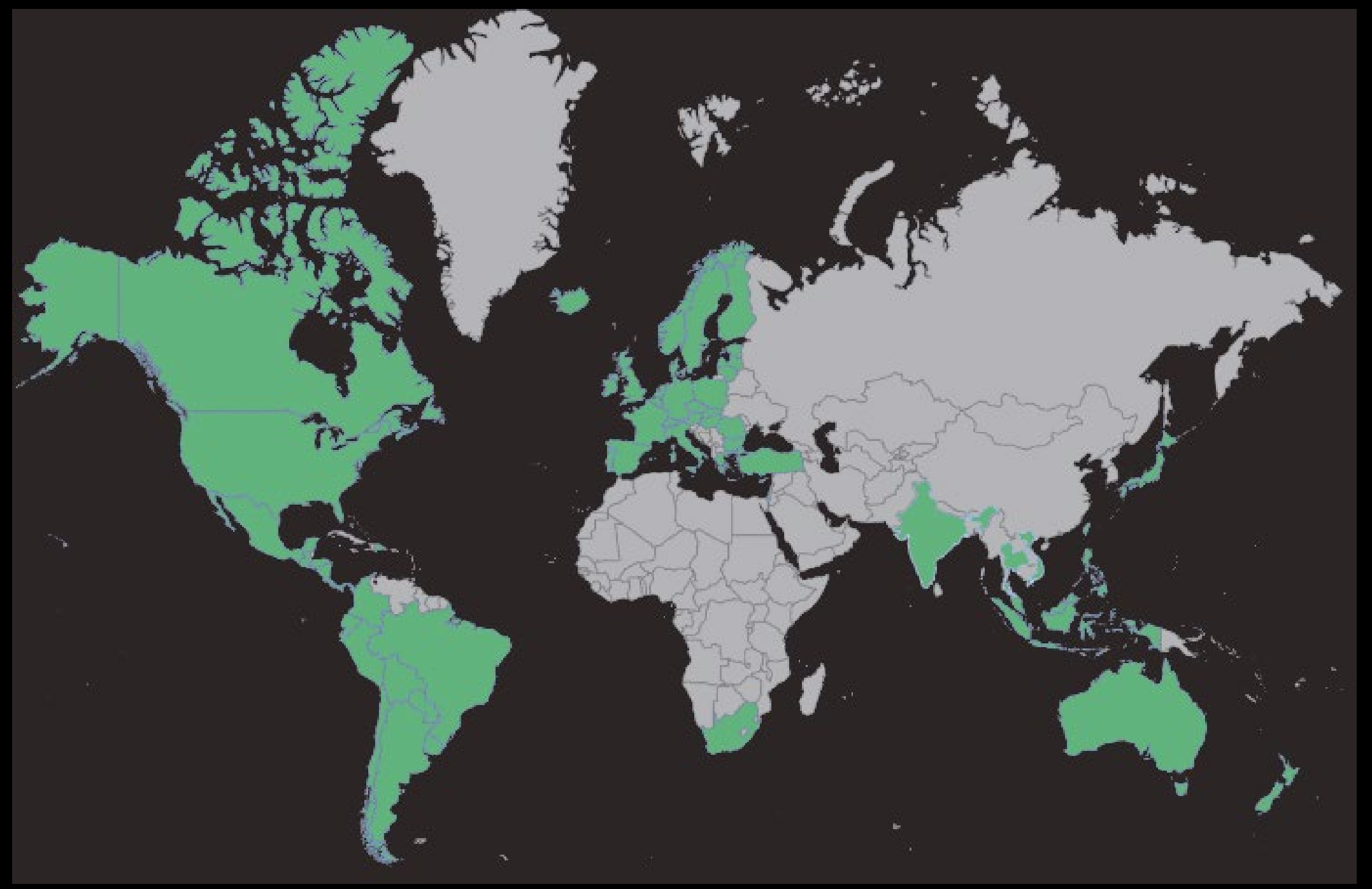
Heavy_Rotation: skaragou - berdogdu - jkennan - ccataldo

## Premise

### Introduction

Spotify is a popular music distribution service. What if we could determine which characteristics of a song cause it to appear on Spotify's Viral Charts? Could we use those characteristics to *predict* what songs may make it big in the future?

### The Data



We collected Spotify's Top 50 Viral Charts from https://www.spotifycharts.com using PowerShell. Charts were recorded daily from January 1, 2017 to March 1, 2020, and in total we gathered **3,411,816 [song, date] pairs** containing **109,658 unique songs** from **66 countries**.

We then obtained 'features' for each song through Spotify's API and stored our dataset in a SQLite database. Specifically, we gathered **valence** (how 'positive' a song sounds), **acousticness** (how acoustic or electronic a song is), **danceability** (how suitable a song is for dancing), **energy** (a song's intensity and activity), **instrumentalness** (whether a song contains vocals), **speechiness** (whether a song contains spoken words), and **tempo** (the estimated speed of the song in beats per minute).

### The Flow

We initially hypothesized that prominent 'features' of popular songs in the US were deterministic of song virality across the globe. From there, we…

- Applied Cox Proportional-Hazards and Kaplan Meier models to our dataset
- Uncovered a much more nuanced scenario regarding song features and viral longevity
- Shifted our approach to an exploratory analysis that delved deeper into the correlation of song features and a song's 'life' on the charts, specifically between the United States and Japan

## Survival Analysis

### Motivation and Results

We first tested whether US viral song features matched those of other countries using a Cox Proportional-Hazards survival analysis.

| | Japan | | | | USA | | | | Greece | | | | India | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | coef | z | Pr(>\|z\|) | Sig. Lvl. | coef | z | Pr(>\|z\|) | Sig. Lvl. | coef | z | Pr(>\|z\|) | Sig. Lvl. | coef | z | Pr(>\|z\|) | Sig. Lvl. |
| **Valence** | -0.188 | -3.014 | 0.003 | ** | -0.009 | -0.127 | 0.899 | | 0.094 | 1.212 | 0.226 | | 0.048 | 0.347 | 0.729 | |
| **Acousticness** | -0.147 | -2.362 | 0.018 | * | -0.119 | -1.832 | 0.067 | . | -0.106 | -1.421 | 0.155 | | -0.128 | -1.082 | 0.279 | |
| **Danceability** | 0.088 | 0.972 | 0.331 | | -0.857 | -9.013 | < 2e-16 | *** | -0.837 | -7.784 | 0.000 | *** | -0.162 | -0.799 | 0.424 | |
| **Energy** | -0.167 | -2.002 | 0.045 | * | 0.156 | 1.655 | 0.098 | . | -0.001 | -0.009 | 0.993 | | 0.049 | 0.271 | 0.786 | |
| **Instrumentalness** | 0.333 | 5.703 | 0.000 | *** | 0.515 | 5.915 | 0.000 | *** | 0.574 | 6.179 | 0.000 | *** | 0.596 | 4.332 | 0.000 | *** |

Significance Level codes: 0 '***' 0.001 '**' 0.05 '.' 0.1 ' '

These selected results reveal that **viral song features across countries are vastly different.** In the USA, instrumental songs and dance beats survive well. Japan liked happier songs, Greece also likes to dance, and India only preferred instrumental songs.

### Conclusions

The survival analysis model identified statistically significant features for each country but **could not conclusively confirm or deny our hypothesis**. Our model assumed that virality is only dependent on the provided song features and ignored external factors such as advertising and celebrity news. Additionally, the p-values are susceptible to outliers which could indicate undue significance between that feature and song survival.

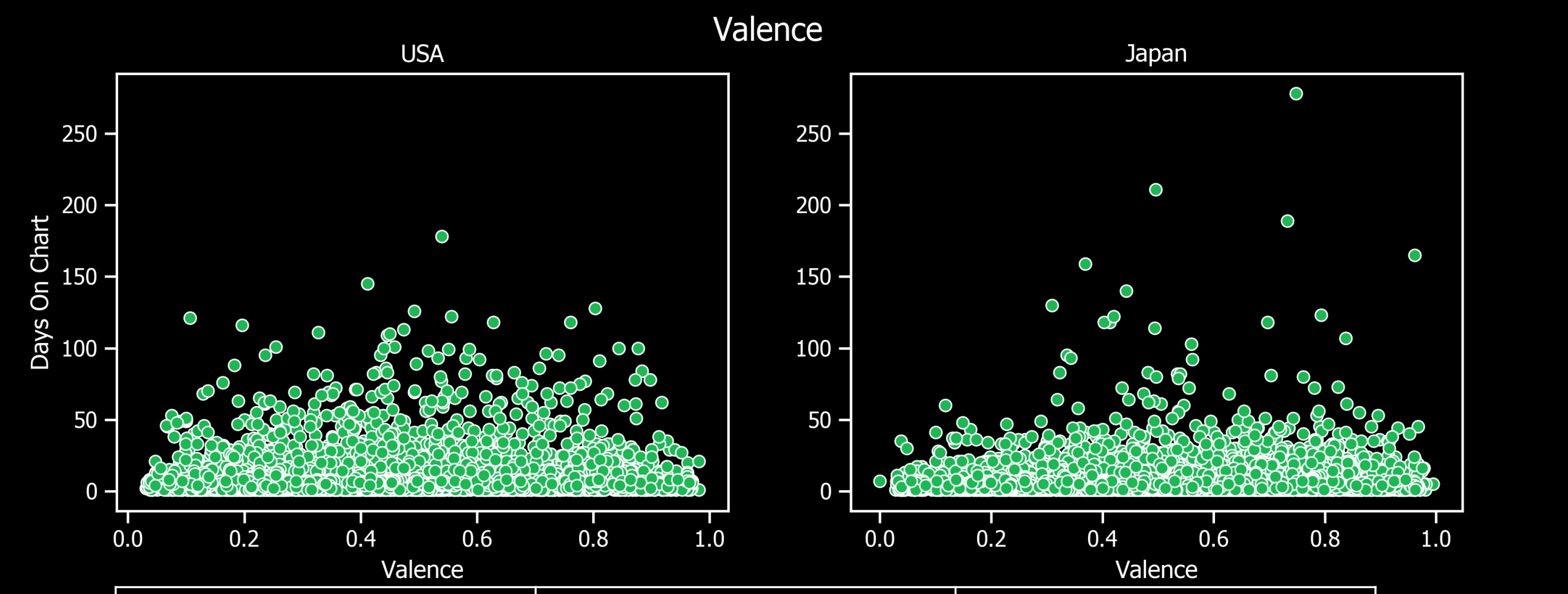
## USA & Japan

### Exploratory Analysis

To avoid suggesting causation in place of correlation, **we abandoned our hypothesis** and chose instead to explore other interesting trends within the Spotify song data. **We chose to compare the US and Japan**, as they have different significant song features and many unique tracks with which to obtain accurate correlations. Importantly, there was also very little overlap between the two charts. **Only 26.4% of the US chart, and 21.9% of the Japan chart, appeared on the other.** We also transitioned to a scatterplot + Pearson test to more accurately correlate single feature values with time on charts instead of using Cox P-H's multiple regression model.

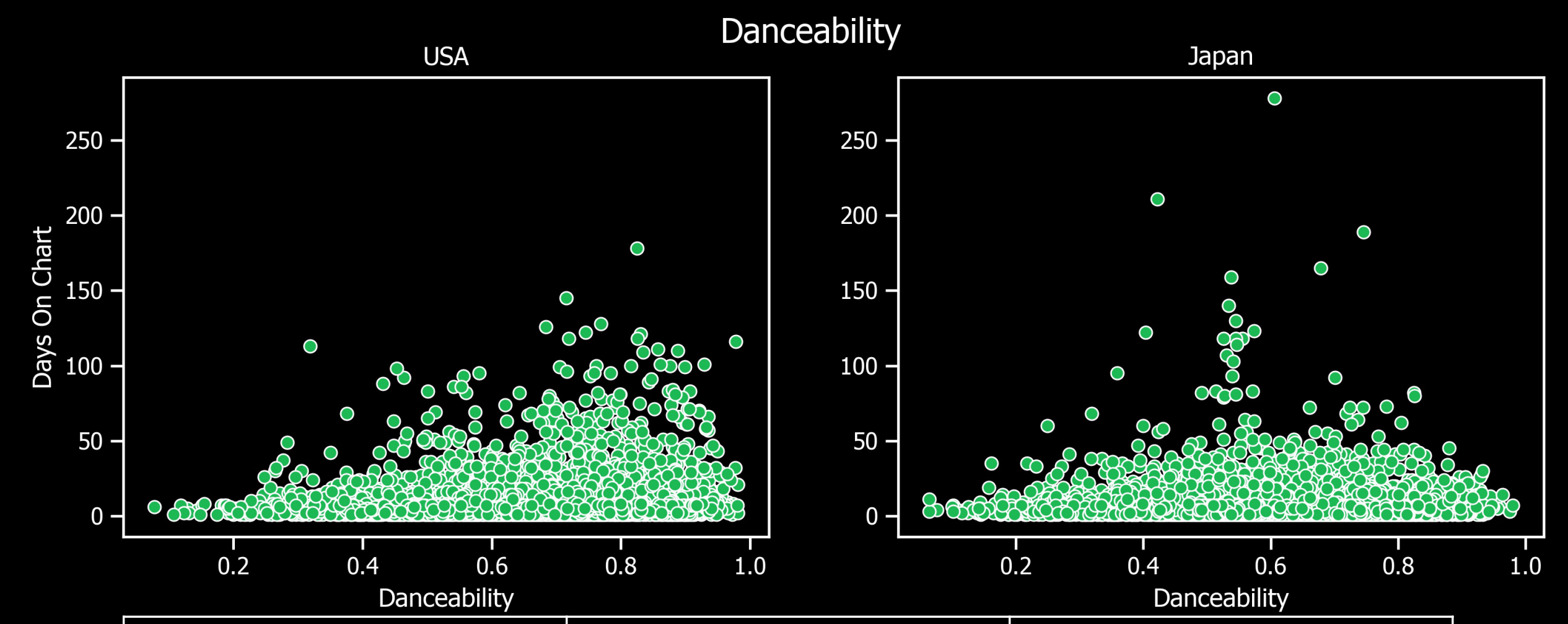
| | Days On Chart | Valence | Acousticness | Danceability | Energy | Instrumentalness | Speechiness | Tempo |
|---|---|---|---|---|---|---|---|---|
| **Mean (US)** | 9.427914 | 0.456383 | 0.231724 | 0.652117 | 0.623229 | 0.037300 | 0.121121 | 120.642003 |
| **Mean 95% CI (US)** | (9.101, 9.755) | (0.451, 0.462) | (0.225, 0.238) | (0.648, 0.656) | (0.619, 0.628) | (0.034, 0.041) | (0.118, 0.124) | (119.916, 121.368) |
| **Std Dev (US)** | 13.000791 | 0.219050 | 0.255045 | 0.153134 | 0.182677 | 0.146241 | 0.119276 | 28.883970 |
| | | | | | | | | |
| **Mean (Japan)** | 7.814416 | 0.520991 | 0.206167 | 0.608906 | 0.709561 | 0.065295 | 0.086789 | 121.873689 |
| **Mean 95% CI (Japan)** | (7.575, 8.055) | (0.516, 0.526) | (0.2, 0.212) | (0.606, 0.612) | (0.705, 0.714) | (0.061, 0.07) | (0.085, 0.089) | (121.231, 122.516) |
| **Std Dev (Japan)** | 10.492461 | 0.224410 | 0.257101 | 0.147929 | 0.203674 | 0.203300 | 0.088219 | 28.088414 |

A comparison of statistics for the US and Japan shows the **mean number of days spent on the charts is higher in the US than in Japan.** Given that there is no overlap in the confidence intervals of each country's mean, **the difference between chart longevity is statistically significant.**



| Country | ρ (Pearson Correlation Coef.) | P-value |
|---|---|---|
| US | 0.0449 | 0.00045 |
| Japan | 0.0467 | $6.23410 \times 10^{-5}$ |

We previously observed valence's significance for chart survival in Japan, but not the US. However, **we find a positive correlation for both countries** between valence and days spent on the chart using a Pearson coefficient. We also observe that Japan's viral songs have higher average valences than the US.



| Country | ρ (Pearson Correlation Coef.) | P-value |
|---|---|---|
| US | 0.1414 | 1.4309e-28 |
| Japan | -0.0083 | 0.4721 |

Similarly, we previously observed danceability's significance for survival in the US, but not Japan. We confirm this with a positive Pearson coefficient for the US. However, for Japan we do not have enough evidence to show a correlation. On average the US' viral songs have higher danceability scores than Japan's.

## Challenges & Conclusions

### Challenges and Assumptions

- Some countries may have gaps in chart data. We attribute this to errors within the source website.
- Seven countries have chart data that does not begin in January 2017 – two start in the latter half of 2017, four in 2018, and one (India, see left) in 2019. We speculate that this was due to Spotify not being available in those countries until the later dates. However, our shift in focus to Japan and the United States nullifies these data issues.
- We considered the number of days a song spent on the chart a cumulative total. It is thus possible that a song which disappeared and reappeared on the charts many times was counted as having a disproportionate number of days on the charts.

### Conclusions

- From our initial analysis, we conclude that we are unable to determine the cause of song virality with just feature set alone and require additional data and context
- After a song reaches the viral charts, there is a correlation between the time it remains there and both valence and danceability in the United States, but only valence in Japan.

## Future Work

While the origins of song virality may be too nuanced to determine from our data, we believe there is still merit in exploring the utilization of song features to predict longevity of virality (dependent of which country the song is popular in).

### Fun Facts

Though we can't make any definitive claims regarding the following, we observed a few additional interesting trends that may inspire further exploration.

- Songs that measured at 120 bpm (beats per minute) appeared to be more popular. 120 bpm is the average walking pace of an adult human!
- Songs with a high danceability value seemed less popular in Japan. A 67 year ban on dancing was in place until 2015, so this extended moratorium may have influenced Japan's music tastes.