**Spotify: Country and Genre Analysis**

BYGB 7978 Web Analytics

Instructor: Professor Yilu Zhou

December 10, 2020

Team 5

Owakhela Kankhwende

Gunita Rahman

Matthew Rogers

Jeanine Woody

**EXECUTIVE SUMMARY**

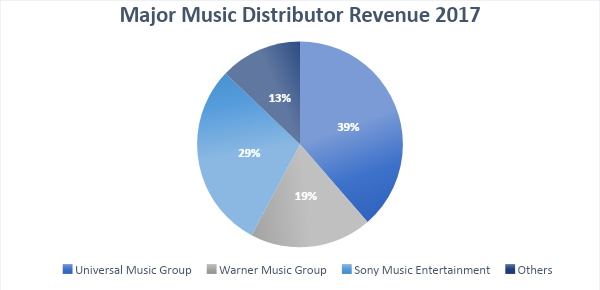
Spotify is a popular digital streaming platform that has continuously increased its services and user base. One of its most well-known features is Spotify Wrapped, which showcases an individual user’s listening history from January to November of a given year. This mixed with other key draws like Canvas, which allows an artist to add a 15 second clip to a released song, have helped Spotify become the dominant music streaming service. As of September 2020, the Swedish based company has approximately 320 million years across 92 markets. Currently, the streaming service provides access to 60 million tracks from various artists.[[1]](#footnote-1) It also allows for different insight layers by hosting communities like Artists, Brands, Developers, Investors, and Vendors. For the sake of this project, we will be utilizing the Developer community through the Spotify Web Developer API.

In this paper, we will investigate the relationship between the audio features of a given track and its relationship with the top charts for a specific country. To do so, we used data from the Spotify “This Is…” playlists, as well as the data from the nine country Top 50 Charts. The audio features were provided through scraping Spotify using the Python package, Spotipy. Our choice of artists reflected different genres that are currently prominent in the industry. These artists are also signed to major record labels that have different Music Distributors. The countries we selected to test in our model were diverse in culture because we believe this will have an impact on the success of an artist. The genres we selected are as follows: Country, K-Pop, Latin, Pop, Rap and Rock. After gaining access to the data, we compiled it to then create our prediction model. Our model of choice was the Nearest Neighbor Model, also known as kNN. This would test the similarities between the tracks already in the Top Charts and the artist tracks.

For certain artists it became clear that their particular style of music based on the audio features of their tracks would translate well into different markets. From the results of our research, we recommend that the Major Distribution Services further look into marketing their artists to markets that align with the music the artist releases. This could later be used to organize tours and campaigns to promote album releases. Based on the data we observed, it became apparent that there are less boundaries between genres since they cross different cultural backgrounds. This would lead to increased growth, not only in revenue, but also in potential markets to cater to. Record Labels could then use this information to schedule tours, promote albums, and discover new artists.

1. **Business Goal Analysis**

Music Distribution Services supply the means for an artist to market their music to a digital streaming platform (DSP) like Spotify, Apple Music, Tidal, Deezer, etc. These distribution services have been around since the inception of the music industry and have had to adapt to the current climate and shift towards the promotion of less physical record sales and more online streams. There are currently five different types of music distributors: Major Distributors, Independent Distribution Partners, White-label Distribution Solutions, Open Distribution Platforms/Aggregators, and Semi-label Distribution Services. For the purpose of this project, we will only look into Major Distributors. These distributors own an artist’s catalog, as well as smaller independent labels, which makes it easier for promotion purposes. As of 2017, approximately 85% of digital revenue goes through a Major Distributor, specifically Warner Music Group (Warner), Universal Music Group (Universal), and Sony Music Entertainment (Sony). Currently Major Distributors have a larger team and music catalog of artist repertoire that gives them a better stance to negotiate deals with a DSP.[[2]](#footnote-2) Currently Major Distributors have a larger team and music catalog of artist repertoire that gives them a better stance to negotiate deals with a DSP. To address the goal of optimizing an artist’s repertoire to different markets, these distributors normally sign deals with record labels to sell the label’s products to then make the recorded music available to the public in physical and digital formats.



**Figure 1.1**

**Pie Chart of the percentage of digital revenue accounted for through a Major Distributor**

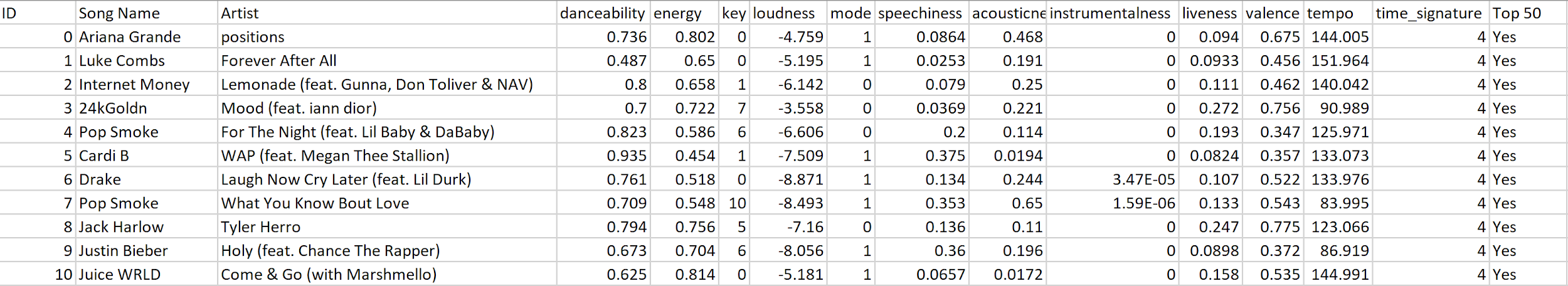
Our goal is to examine the DSP, Spotify’s, track data from the “This Is…” playlists of predetermined artists of different genres and compare it to the Top 50 Playlists of set countries. This can potentially be used to achieve a business model that would help to maximize profit for the promotion of songs and ultimately, touring. Major Distributors have the advantage to look at the qualities of the tracks an artist releases and regulate which market the song should be promoted in. Web Analytics is necessary to understand the relationship between the Major Distributor and the DSP because it would be beneficial to figure out how an artist could promote their own music to a country their musical style best reflects. For this project we looked at six different artists that are signed to different labels that utilize one of the three Major Distributors listed above. The countries we chose were the US, UK, Germany, Brazil, Australia, Japan, Philippines, India, and South Africa. By comparing the “This Is…” Playlist data to the Top 50 Country data we can then determine if a specific track will appear in the Top 50—if so, the artist will then be deemed successful in said country.

|  |  |  |  |
| --- | --- | --- | --- |
| Artist Name | Genre | Record Label | Music Distributor |
| All Time Low | Rock | Fueled By Ramen | Warner Music Group |
| Bad Bunny | Latin | Universal Music Latin Entertainment | Universal Music Group |
| BTS | K-Pop | Big Hit Entertainment | Sony Music Entertainment |
| Drake | Rap | Republic Records | Universal Music Group |
| Kacey Musgraves | Country | Columbia Records | Universal Music Group |
| The Weeknd | Pop | XO | Universal Music Group |

**Table 1: Artist Information**

Based on our knowledge of the artists and their respective genres we believe that certain songs by these artists will be more likely to appear in the Top 50. An example of this would be that The Weeknd is an extremely popular artist who Spotify currently ranks as the fifth artist in the world. We would expect more of the songs he has released to share attributes with those that are currently in the Top 50. Since his style of music fits the pop category, it makes sense as to why his songs receive more streaming in the genre. Our expectation is that the more popular an artist is and the less niche a genre is, the song attributes will be shared with the attributes present in the songs of the Top 50.

1. **Dataset Description**



**Table 2: Sample of Top 50 US Country Data**

Song data was crawled from the nine separate countries to represent diverse cultures and musical preferences globally via their Top 50 playlists. The data crawled for the artists’ is composed of the same attributes crawled for the countries, but instead from each artists’ “This Is…” playlist. The attributes crawled for each song consist of Song Name, Artist, danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo, and time signature. Song Name and Artist are simply the given name of a song in the playlist and the artist is the performer of that song.

*Danceability*

Danceability is a measure of how fitting a song is to dance to, based on specific musical elements such as “tempo, rhythm stability, beat strength, and overall regularity.”[[3]](#footnote-3) This measure ranges from 0, meaning least danceable, to 1, meaning most danceable. This is a numerical measure.

*Energy*

Energy is a measure describing the strength of a song, measured by “dynamic range, perceived loudness, timbre, onset rate, and general entropy.”[[4]](#footnote-4) Songs with a strong display of this attribute are often quick and noisy. This attribute is measured in intensity ranging from 0 to 1. This is a numerical attribute.

*Key*

The key is the attribute of the song that displays the detected key of the song. If not found, this will be displayed by a “-1.” This is categorical.

*Loudness*

Loudness is the measure of the average volume of the song, measured in decibels (db). The range typically falls between -60 and 0 db. This measure is numerical.

*Mode*

Mode is the measure of modality, meaning if the song is major or minor, indicated by 1 if major, or 0 if minor. This measure is thus categorical.

*Speechiness*

Speechinesss measures the presence of spoken word in a track and ranges from 0 to 1, with tracks over .66 typically being fully spoken, such as skits, poetry, or podcasts. This is another numerical variable.

*Acousticness*

Acousticness represents the likelihood of the song being acoustic, measured from 0 to 1, 0 being least likely to be acoustic, 1 being most likely. This is a numerical variable.

*Instrumentalness*

Instrumentalness measures the likelihood that the track contains no vocals. The measure ranges from 0 to 1, 0 meaning the track definitely contains vocals. As the measure increases the presence of vocals is more profound. This is a numerical measure.

*Liveness*

Liveness is a measure predicting the presence of an audience, meaning if the track was performed live. This measure ranges from 0 to 1, 0 being that there is no chance the track was performed live, while 1 meaning the track has features indicative of a live performance. This is a numerical variable.

*Valence*

Valence is a measure that describes the positivity or negativity of a song depending on the sound of the song, positivity being associated with a joyful sound, with negativity being associated with a more somber or enraged sound. This measure ranges from 0 to 1, with the increase from 0 to 1 indicating the level of positivity, making this a numerical variable.

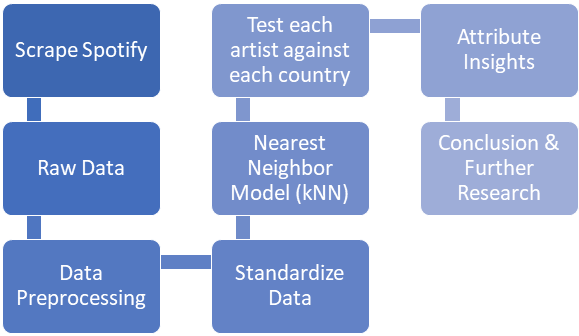
*Tempo*

Tempo is the measure of speed in music regulated by average count of beats per minute, or bpm. This is another numerical variable.

*Time Signature*

Time signature is another measure of speed in music, specific to measuring the average number of beats per measure. This last variable is also categorical.[[5]](#footnote-5)

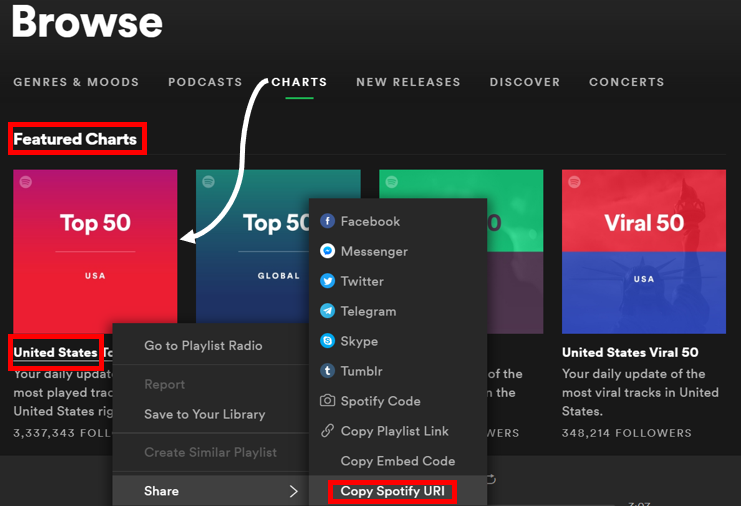
1. **System Design (Flow Chart)**

****

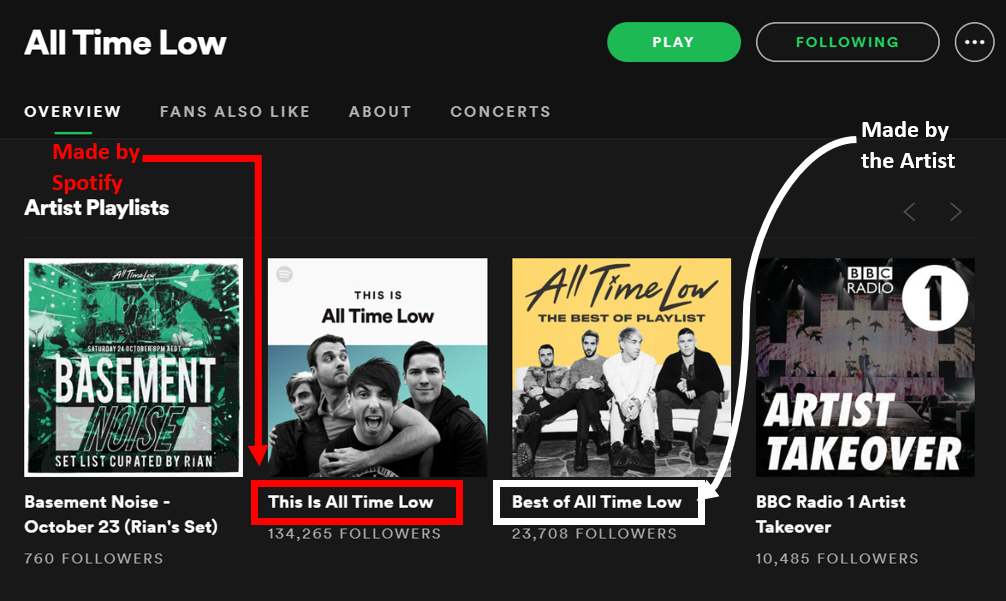
**Figure 3.1 Flow Chart of the System Design**

1. **System Implementation**

The first tool utilized was the Spotify Web Developer API for Python known as Spotipy. The Spotify Web Developer API uses Spotify URIs and IDs to determine artist, track, playlist, or user data. The Spotify URI is a specific identifier that is apparent in the Spotify Desktop platform (see Figure 3.1). According to Spotipy, the Spotify URI is “a base-62 number that you can find at the end of the Spotify URL for an artist, track, album, etc.”[[6]](#footnote-6) These components are used in the Python code to extract the audio features. We used the playlist feature to get access to the different songs and then we used the track data to get the audio analysis. This was used to scrape the playlist information for the artists and countries. To ensure we were using the proper playlist information we found the Top Charts through browsing the charts and copying the Spotify URI. We found the "This Is…" Playlist under the artist— we used the playlists that were curated by Spotify to make sure everything was uniform throughout.



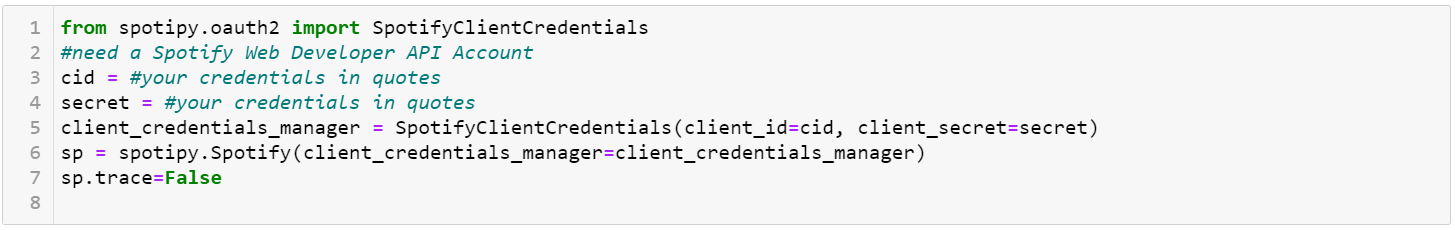
**Figure 4.1 Spotify Chart Playlist Data**



**Figure 4.2 Spotify Artist Playlist Data**

The first thing we added to our code was the authorization of our SpotifyClientCredentials to gain access to the developer through our own Spotify account. This is because the Spotify Web Developer API must be attached to an existing Spotify account. The account information was removed from the photos below since the key is unique to the user. The function relating to show\_tracks scrapes the Artist Name and the Track Name.[[7]](#footnote-7) The second for loop scrapes the individual tracks for their song features (these were discussed in the dataset description).[[8]](#footnote-8)





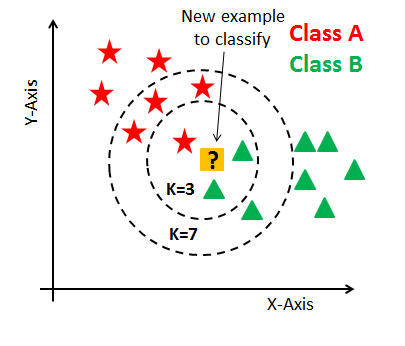
****

**Figure 4.3 Python Code to extract song data per artist & country**

After extracting Spotify's data, the country information was compiled into one dataset as a CSV file. Different versions of the dataset were implemented based on the origin country playlist. The Class variable indicated the songs were originally from the country whose model is being implemented. A song tested by a model is assigned a "1" if it's predicted as likely to appear in the country's top 50. We tried to determine which variables were necessary to keep in order to run the model. After finding not enough difference in correlation or importance to drop variables, we went with the full suite of song relevant variables decided upon by our knowledge and via Spotify's variable descriptions. These would be danceability, energy, tempo, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, time signature, and valence.

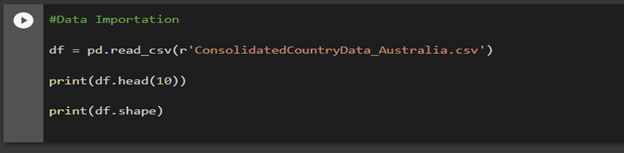
Time signature and key were later dropped due to dimensionality issues when new data was tested.

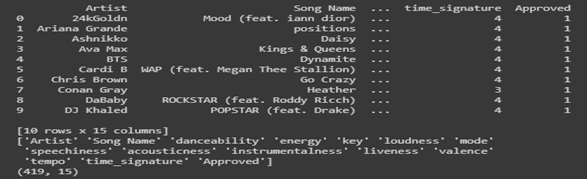
Following this, we implemented the Nearest Neighbor Model or kNN as our predictive technique. This choice was because our dataset is relatively small, with 50 songs per artist and country and 419 songs in the country dataset overall. The Nearest Neighbor or kNN model is a supervised data mining technique in which an algorithm compares the same features/measures for new data points with those of an optimized trained data model. In the end, the model will let us know the predicted success of the new data points when compared with the trained model obtained from the original data via closeness by Euclidean distance. For our project, we had to build nine trained kNN models for each country, Australia, Brazil, Germany, India, Japan, Philippines, South Africa, United Kingdom, and the United States. After we made these optimized trained models, we ran each of our six artists' datasets through as new data points to see how many songs would make it through each country's model.



**Figure 4.4 DataCamp Nearest Neighbor Example**[[9]](#footnote-9)

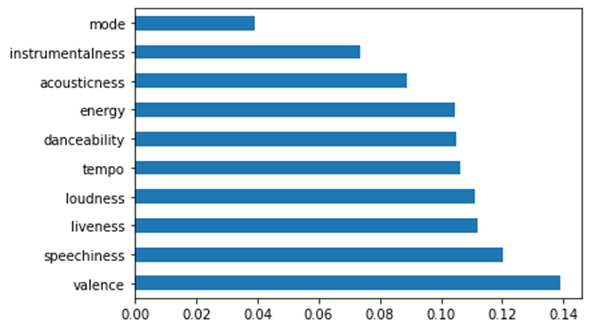
The first step of building our kNN model was to import the “Top 50” dataset we configured for each of our nine countries.



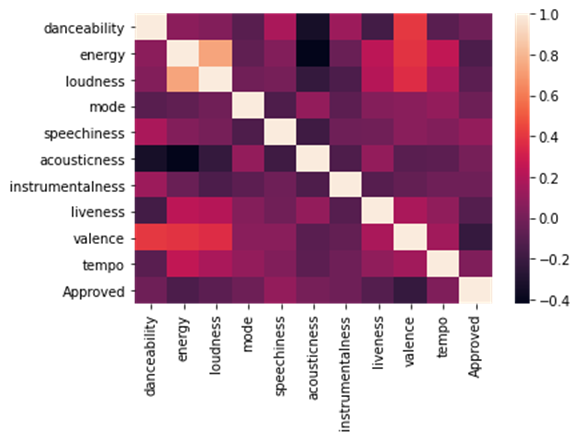


**Figure 4.5 Python Code to Import Country Datasets**

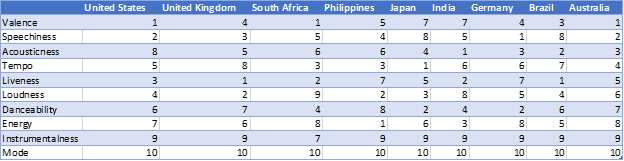
After we imported the data for each country, we looked into feature importance and tried to find clear variables to drop due to lack of significance with class. A pandas correlation method and ExtraTreesClassifier were used to find correlation and feature importance respectively for each model. Below are two charts representing feature importance visualizations for the United States. We ran these visualizations for each of our 9 countries in order to get a better understanding of how each feature ranked in their respective “Top 50” playlist. As you can see the most important features in the United States by ranking of importance were valence, speechiness, liveness, loudness, tempo, danceability, energy, acousticness, instrumentalness and lastly mode. Each countries’ “Top 50” playlist feature importance was different, laying down the framework for each country model’s uniqueness.



**Figure 4.6 United States Feature Importance via ExtraTrees Classifier**



**Figure 4.7 United States Feature Importance via Correlation Matrix**

****

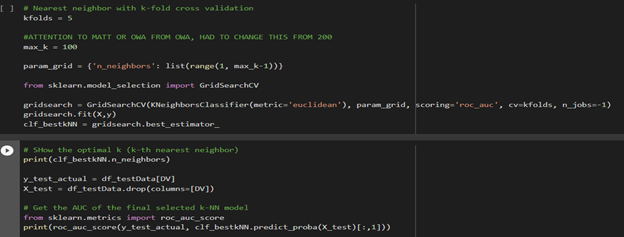
**Figure 4.8 Country Feature Rankings Chart**

We then manipulated and put the different features into categorical and numerical variables based on their values. The numerical variables in this case were danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence, tempo, and the lone categorical variable was mode. Once these features were broken down into their proper variable lists, we had to standardize the data so it would be consistent throughout our models.



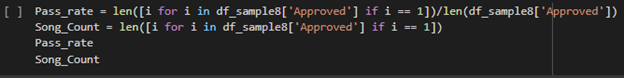
**Figure 4.9 Python Code to Set Up Variables**

The next step in the model that had to be addressed was to figure out the optimal number of nearest neighbors or K to include in each country’s kNN model. Each country has a different optimal K due to the different data points.



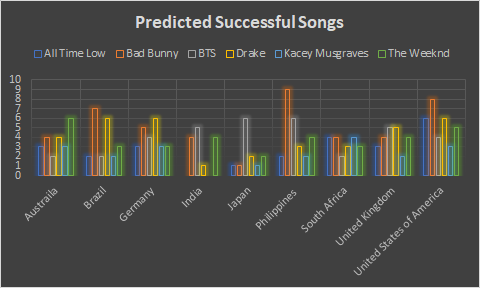
**Figure 4.10 Python Code to Discover Optimal Nearest Neighbors**

Once we obtained the optimal number of nearest neighbors to include for each country’s trained model, we then moved on to the artist side of the analysis. We took each of our six artists’ “This is” datasets and formatted them using the same techniques as the country data so they would have consistent dimensions in order to push them through the optimized country models. Finally, we ran each artist’s formatted dataset through each country’s trained model to see how many of the artist’s songs would be predicted to make it into the “Top 50” of that country.

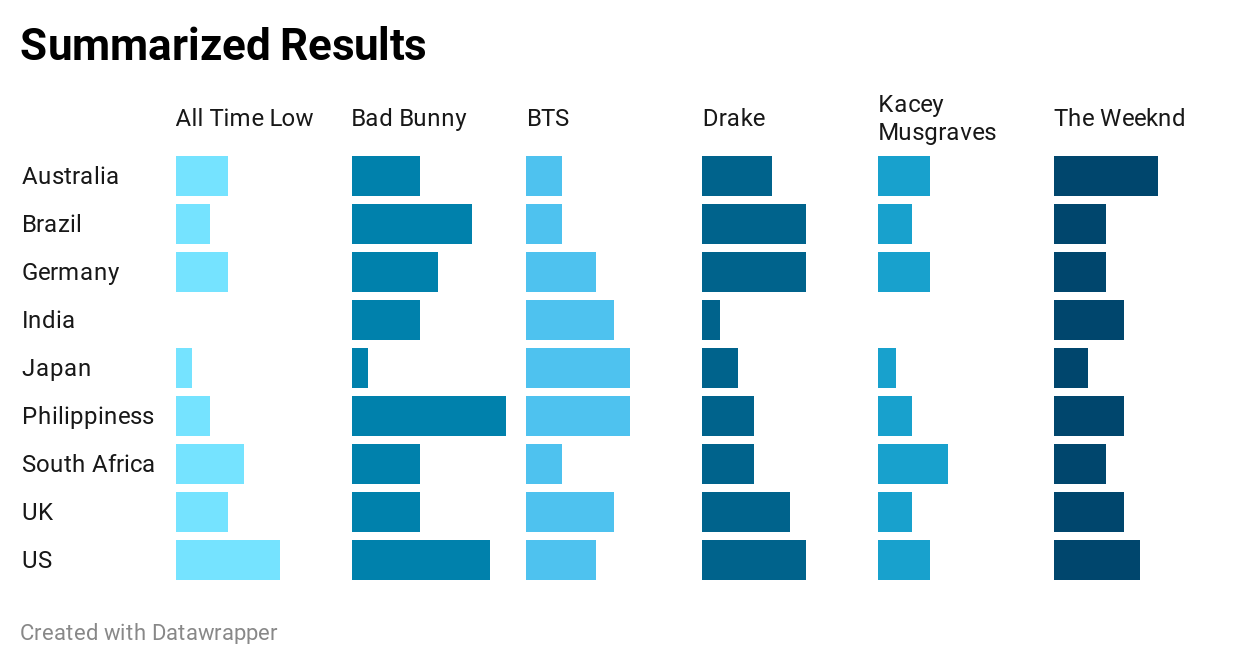


**Figure 4.11 Python Code to Display Passing Song Count**

1. **Evaluation**

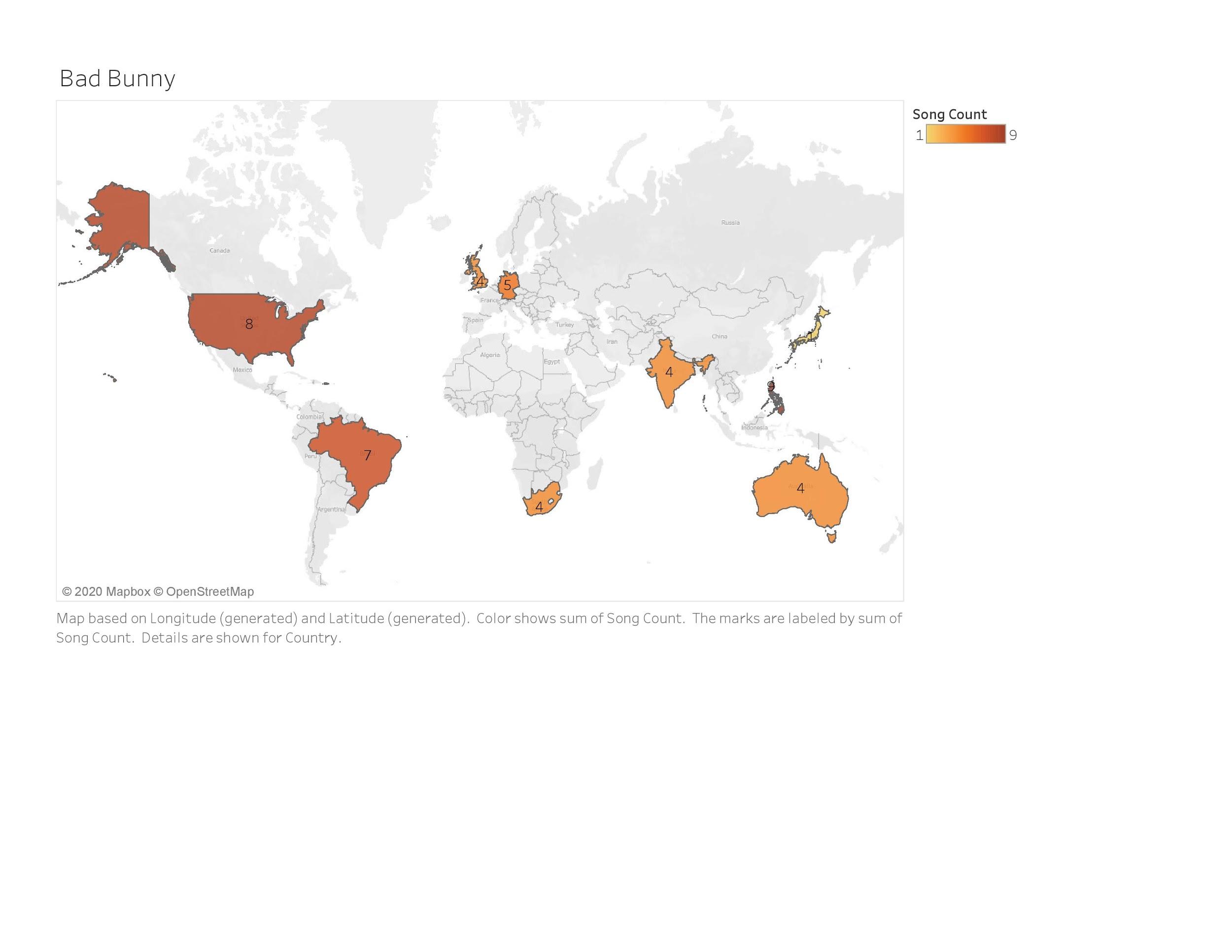
****

**Figure 5.1 Song Success by Country and Artist**

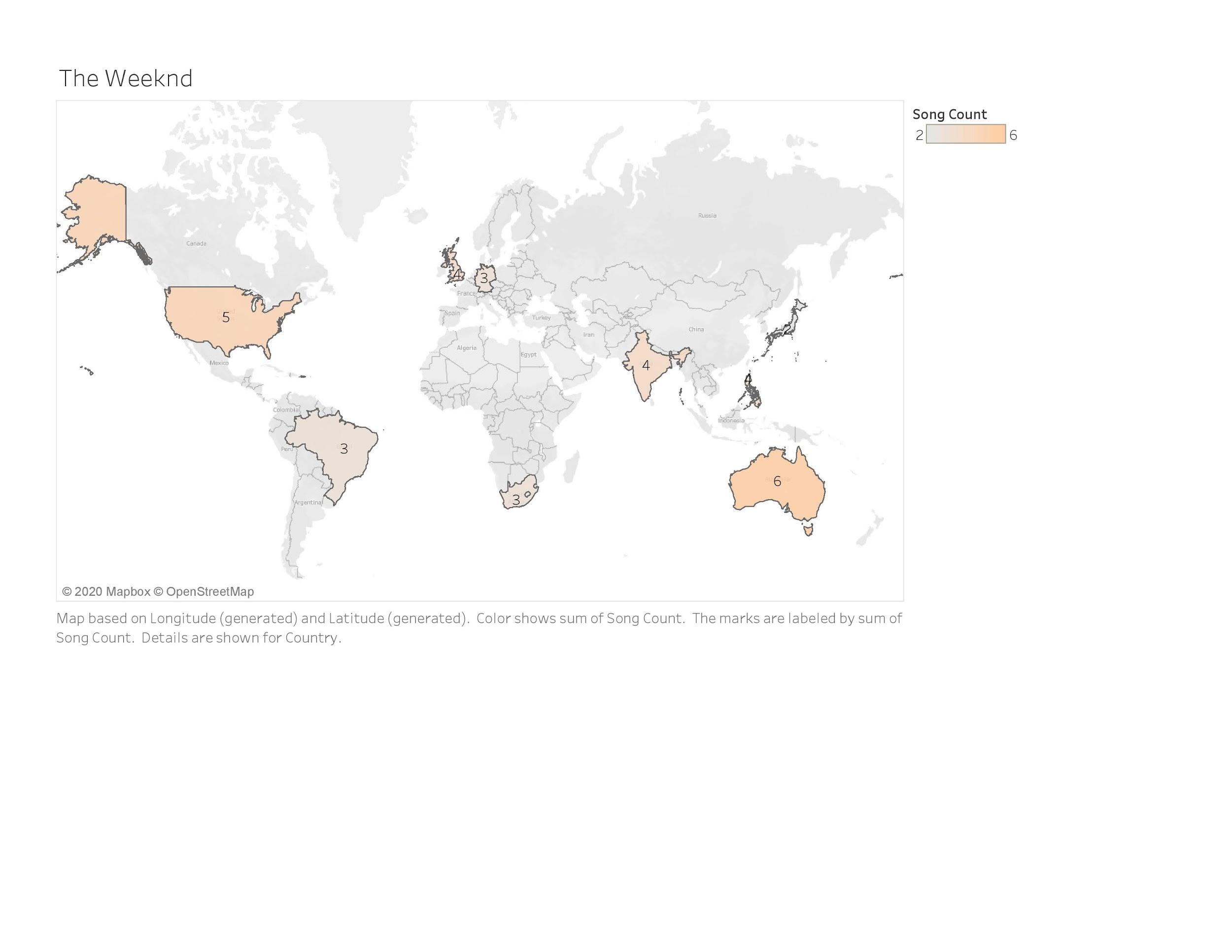
****

**Figure 5.2 Artist Success by Country**

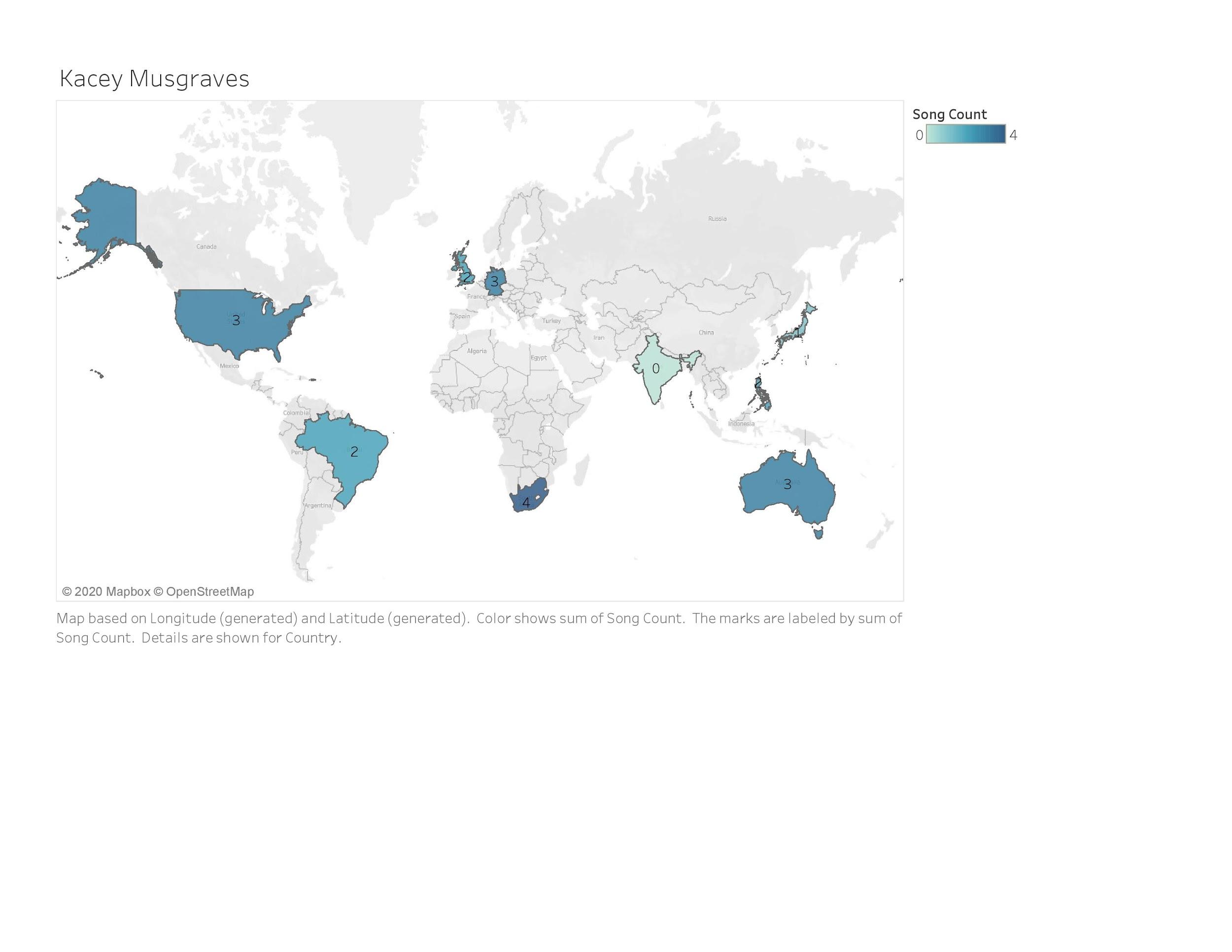
In the chart above we can see the comprehensive results for each artist’s predicted success after running their “This Is...” playlists in our kNN models for each country. For certain artists it is very clear that their particular style of music based on their features would translate well into different markets. For example, if we look at Bad Bunny in the Philippines, we can see that 9 out of his 50 “This Is...” playlist songs would be deemed a success there. This figure seems to be very high and if presented to Bad Bunny’s Major Music Distributor they might seek out putting more resources into promoting Bad Bunny in that untapped market. Similarly, The Weeknd had strong performance across all of the markets. The Weeknd had a range of 2 to 6 songs from his “This Is...” playlist making it into all of the countries’ “Top 50” playlists. This consistent success would steer The Weeknd’s Major Music Distributor to possibly put together a vast reaching business plan since The Weeknd’s music translated well into all markets. Comparably Drake also had solid success across all countries. Drake had an average of 4 songs making it from his “This is…” playlist into the “Top 50” country playlists. The fourth positive leaning artist from our analysis was BTS, they consistently had 4 or more songs from their “This is…” playlist making it to the “Top 50” country playlists. Notably BTS had 6 songs make it through in the Philippines and in Japan. This healthy performance in Asia would likely make BTS’s Major Music Distributor invest more resources for expansion in that corner of the world. On the other hand, these results could help a Major Distributor avoid wasting resources in markets. For example, if we look at All Time Low’s successful song rate in India, it is zero, telling us that their music’s features translate miserably when compared to the popular music in India. This could be advantageous information for All Time Low’s Major Distributor by preventing them from possibly streamlining resources into a market that most likely will not prove successful. The last artist to take a glance at is Kacey Musgraves, who performed the worst out of our group of artists. She had an average of 2 songs making it through our kNN models for each country, notably 0 and 1 in India and Japan respectively. Kacey Musgraves’ Music Distributor would likely decide to veer away from sending any resources to those low-performing markets. Overall, these results could be interpreted in various ways to help out Major Music Distributors. Three ways that stick out are to encourage artist expansion into a new market, avoid wasting resources in a sub-optimal market and to reaffirm and hold steady in a current market.



**Figure 5.3 Distribution of Songs for Bad Bunny**

****

**Figure 5.4 Distribution of Songs for The Weeknd**

****

**Figure 5.5 Distribution of Songs for Kacey Musgraves**

1. **Conclusion and Future Direction**

Through the analysis we conducted we concluded that the top genre worldwide was Latin music, as seen with Bad Bunny placing on all nine of the Top 50 Playlists for the nine countries we chose to crawl. Out of the songs in Bad Bunny’s “This Is…” playlist, the maximum number of songs to chart in a single country’s Top 50 playlist was nine. Following closely was The Weeknd, our representative of Pop, charting in all nine countries as well; but with the maximum amount of six songs on a singular Top 50 chart. Kacey Musgraves, our representative for Country, placed last among all the artists worldwide, only placing on eight out of nine countries, however her maximum song count on a single Top 50 chart was only four.

If seeking to pursue this research on a deeper global scale in the future, we would recommend applying the same code towards all 92 countries serviced by Spotify, using the correct URI code per a country’s Top 50 playlist. The same could be said for expanding the research to include more artists to represent a more diverse array of genres. This better represents the multitude of musical tastes across the global music market. It would be beneficial for a Music Distributor or Record Label to examine budding artists that they would consider adding to their repertoire. Given the fact that digital streaming platforms are heavily data driven, this allows new artists to gain more exposure to users through recommendations.

Music is an ever-evolving field that only continues to grow and spread cultures across the world. We see this with the expansion of new languages and genres in Western cultures as seen with artists such as BTS and Bad Bunny decimating chart records in today’s modern era with songs that are not primarily English, a departure from typical popular Western music. In our analysis we saw the beginnings of these diversifications of music tastes throughout regions of the world, and we look forward to observing these trends develop further in the future.

1. **References**

[Navlani](https://www.datacamp.com/profile/avinashnvln8), [Avinash](https://www.datacamp.com/profile/avinashnvln8) . “KNN Classification using Scikit-learn”, [August 2nd, 2018](https://www.datacamp.com/profile/avinashnvln8) https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn

Pastukhov, Dmitry. “Market Intelligence for the Music Industry.” *Soundcharts*, 23 Sept. 2019, soundcharts.com/blog/music-distribution.

Spotify, “Company Info”, *Spotify Newsroom*, September 30, 2020. https://newsroom.spotify.com/company-info/

Spotify Web Developer API. “Get Audio Analysis for a Track.” *Spotify for Developers*, 2018, developer.spotify.com/documentation/web-api/reference/tracks/get-audio-analysis/.

Spotify Web Developer API, “Get a Playlist's Items.” *Spotify for Developers*, 2018, developer.spotify.com/documentation/web-api/reference/playlists/get-playlists-tracks/.

“Welcome to Spotipy!”, *Read The Docs*, https://spotipy.readthedocs.io/en/2.16.1/

1. Spotify, “Company Info”, *Spotify Newsroom*, September 30, 2020.

   https://newsroom.spotify.com/company-info/ [↑](#footnote-ref-1)
2. Dmitry Pastukhov, “Market Intelligence for the Music Industry.” *Soundcharts*, 23 Sept. 2019, soundcharts.com/blog/music-distribution. [↑](#footnote-ref-2)
3. Spotify Web Developer API, “Get a Playlist's Items.” *Spotify for Developers*, 2018, developer.spotify.com/documentation/web-api/reference/playlists/get-playlists-tracks/. [↑](#footnote-ref-3)
4. Spotify Web Developer API, “Get a Playlist's Items.” *Spotify for Developers*, 2018, developer.spotify.com/documentation/web-api/reference/playlists/get-playlists-tracks/. [↑](#footnote-ref-4)
5. Spotify Web Developer API, “Get a Playlist's Items.” *Spotify for Developers*. [↑](#footnote-ref-5)
6. “Welcome to Spotipy!”, *Read The Docs*, https://spotipy.readthedocs.io/en/2.16.1/ [↑](#footnote-ref-6)
7. Spotify Web Developer API, “Get a Playlist's Items.” *Spotify for Developers*. [↑](#footnote-ref-7)
8. Spotify Web Developer API, “Get Audio Analysis for a Track.” *Spotify for Developers*, 2018, developer.spotify.com/documentation/web-api/reference/tracks/get-audio-analysis/. [↑](#footnote-ref-8)
9. [Avinash Navlani](https://www.datacamp.com/profile/avinashnvln8), “KNN Classification using Scikit-learn”, [August 2nd, 2018](https://www.datacamp.com/profile/avinashnvln8)

   https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn [↑](#footnote-ref-9)