Country Music

Examining the relationship between highly-played Spotify songs and overall happiness in different countries

Joey Hotz

April 21, 2022

Introduction

Music is often thought of as one of the greatest cultural achievements which the human race has accomplished. From Stradivarius and Bach to Kanye West and Taylor Swift, music has continued to change and develop over time. With the advent of technology in the 21st Century, and the popularity of music streaming platforms, it is not a surprise that music is so prevalent in our daily lives.

Beyond the fact that music is so easily accessible, it is also incredibly popular. Many people listen to music throughout the day, as they may find that it helps them to focus, or even improves their mood. Many people find that their mood changes alongside the mood of the music which they are listening to, which is why people may create playlists for specific emotions, such as a workout playlist or a "sad boi hours" playlist. Music, and in particular, the ability to play music easily at the touch of a button, presents an incredibly accessible way for people to change and regulate their emotions, as it can help you to be calm, or angry, or happy.

In this report, we seek to answer the following question: "*Is there a correlation between countries' happiness levels and how happy the most popular songs are in that country?*" To answer this question, we will examine the most popularly-streamed songs on Spotify in different countries during the first three months of 2022, and compare these songs to the country happiness rankings based on the World Happiness Report.

Spotify is one of the largest and most popular music streaming apps in the world. The platform allows subscribers to easily play songs from a seemingly endless collection of music.

On top of their vast collection of music and other forms of entertainment (such as podcasts and shows), Spotify also has a robust API which allows users to access 'advanced features' of audio files which are available to stream through Spotify. These advanced features which are available for songs on Spotify include measurements of whether a song is more instrumental or more vocal, whether a song was recorded live or in a recording studio, how energetic the track is, and how happy a song is.

The World Happiness Report is an annually-published report which is written by members of the United Nations Sustainable Development Solutions Network. The report aims to classify the countries of the world based on how happy the citizens of the country are with their lives in the country. The SDSN collects their data by polling the citizens of the respective countries, and bases their happiness rankings on six primary explanatory factors; GDP per capita, social support, life expectancy, freedom to make life choices, generosity, and corruption.

The World Happiness Report is a well-regarded statistical survey which can be used as a numerical quantifiable metric for how happy citizens of a specific country are, with regard to the lives which they are able to lead in that country.

To do this, we will collect the audio features of the most played tracks on Spotify in different countries around the world, and we will compare this information to the happiness index values in the World Happiness Report, to determine if there is a statistically significant relationship.

Table 1: Happiness Metrics for four different songs by Kanye West

Song	Danceability	Energy	Valence
Champion	0.693	0.504	0.723
Good Life (feat. T-Pain)	0.439	0.808	0.487
Love Lockdown	0.760	0.524	0.112
Only One (feat. Paul McCartney)	0.732	0.243	0.106

Methodology

The fundamental research question which we seek to answer is "*Is there a correlation between countries' happiness levels and how happy the most popular songs are in that country?*"

In order to answer this question, we will need to quantify the happiness of the people in a country, as well as the happiness of a song. As mentioned above in the Introduction section, we can quantify the happiness of a song using the three metrics (danceability, energy, and valence) which are provided through Spotify's API.

To determine the happiness level of a country, we will use the happiness index values which were estimated in the 2021 edition of the World Happiness Report, and we will compare the countries' happiness levels to the happiness metrics for the most popular songs on Spotify in these respective countries.

Song Happiness Metrics

In order for us to create a robust metric to determine how happy the most popular songs on Spotify are in a particular country, we will use three track features from the Spotify API. The three features which we will use as a metric for how happy a song is are the "danceability", "energy", and "valence" metrics. These three metrics are all computed by Spotify themselves, and although the company has not provided specific information on the algorithm they use to compute these metrics, their API documentation does provide information about what these three values mean.

- **Danceability** describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
- **Energy** is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.
- **Valence** is a measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).

The table below shows the values of these three song happiness metrics for four different songs by Kanye West; Champion, Good Life, Love Lockdown, and Only One. Although all four songs are by the same artist, these four songs cover a different range of sounds. Of these songs, some of them (Champion and Good Life) are upbeat and have positive lyrics, whereas Love Lockdown and Only One are sad and sentimental songs.

Although Spotify's algorithms to determine the danceability, energy, and valence of individual songs are essentially "black box algorithms", these four examples can help to illustrate the differences between these three metrics, and how these are evaluated for different songs.

Data Collection

To collect the data regarding the happiness levels of individual countries, we had to download the data as a Microsoft Excel spreadsheet from the World Happiness Report's website. The SDSN publishes a new edition of the World Happiness Report annually, and their organization takes care to ensure that the data which they use for their report is easily accessible for the general public.

For us to collect the data which the SDSN used in creating the 2021 edition of the World Happiness Report, we simply need to navigate to the 2021 report on their website, and download the Excel spreadsheet labelled "Data for Figure 2.1", which is a neatly-formatted table which summarized the organization's findings.

To collect the data relating to the most popular songs on Spotify in different countries across the world during the first three months of 2022, we used weekly Top 200 Tracks charts for each country, which were published on the Spotify Charts website.

The Spotify Charts website has information available for 72 countries, a majority of which are located in North America or Europe, and are primarily richer countries. Of the 72 countries with data available on the Spotify Charts website, only two of these countries (Nigeria and South Africa) are located in Sub-Saharan Africa, despite the fact that this is the geographic region with the largest number of countries, according to the World Happiness Report's list of regions.

Within Spotify's API, as well as the Spotify Charts website, each of these 72 countries has a corresponding two-letter country code. For example, Canada's code is "ca", and New Zealand's code is "nz". For any of these 72 countries, the link to the Weekly Top 200 Tracks in that country in a given week (from Friday to the following Friday) is `https://spotifycharts.com/regional/[COUNTRY CODE]/weekly/[START OF WEEK]--[END OF WEEK]`, where the country code is the country's unique two-letter identifier, and the start and end of week are the first and last days of the week, respectively, both of which are written in `YYYY-MM-DD` format.

To collect the data from these charts, we created a web scraper in Python which utilized the `requests` and `BeautifulSoup` modules in order to read in the raw HTML code for these Top 200 Tracks charts and extract the information pertaining to these tables, before using the `pandas` library to coerce these tables into CSV files.

This process allowed us to collect up to 13 weeks' worth of Top 200 Tracks charts for 72 countries across the globe, for a grand total of 918 individual Top 200 Tracks charts during the first three months of the year. However, the charts which were collected from this website only displayed the names of the song, the country, the dates, and the number of streams which the song had during that week in that country. In order to collect the information about how happy the songs on these charts were, we had to import the data into R, and use the `spotifyR` wrapper for Spotify's AP to collect the necessary information about each of these individual tracks. In total, there were 11479 individual tracks which we needed to collect information about with the Spotify API.

Data Cleaning

With respect to the data which we collected from the World Happiness Report's website, there was nearly no cleaning required for us to utilize the data. The "Data for Figure 2.1" spreadsheet on the World Happiness Report website is already cleanly formatted for user readability, which allows us to easily load in the data from the Excel spreadsheet into R, with virtually no cleaning required, other than renaming the columns in the dataset to match our use cases, and removing extraneous columns.

The happiness data which we read directly from the World Happiness Report's published Excel spreadsheet was very clean and well formatted. The table below shows a small subsection of the data which was read directly from the Excel spreadsheet which we downloaded from the World Happiness Report's website.

As we can see in the table above, this table is already formatted in an easily readable and easily usable way, which allows us to essentially just load in the data from this Excel document and get straight to work.

Table 2: Sample of the data read directly from the 2021 World Happiness Report's spreadsheet

Country name	Regional indicator	Ladder score	Standard error of ladder score	upperwhisker	lowerwhisker
Finland	Western Europe	7.8421	0.0316457	7.904126	7.780075
Denmark	Western Europe	7.6195	0.0346571	7.687428	7.551572
Switzerland	Western Europe	7.5715	0.0362430	7.642536	7.500464
Iceland	Western Europe	7.5539	0.0593734	7.670271	7.437528
Netherlands	Western Europe	7.4640	0.0273260	7.517559	7.410441
Norway	Western Europe	7.3925	0.0354486	7.461979	7.323021
Sweden	Western Europe	7.3627	0.0356996	7.432671	7.292729

Table 3: Sample of the raw data collected from the Spotify Charts website via Python

Source
<td class="chart-table-image"> <a href="https://open.spotify.com/track/7rglLriMNBPAyuJOMGwi39" target="_blank"
<td class="chart-table-image"> <a href="https://open.spotify.com/track/02MWAaffLxlfxAUY7c5dvx" target="_blank">
<td class="chart-table-image"> <a href="https://open.spotify.com/track/5PjdY0CKGZdEuoNab3yDmX" target="_blan
<td class="chart-table-image"> <a href="https://open.spotify.com/track/4fouWK6XVHhzl78KzQ1UjL" target="_blank"
<td class="chart-table-image"> <a href="https://open.spotify.com/track/5Z9KJZvQzH6PFmb8SNkxuk" target="_blank

On the other hand, the data which we collected from the Spotify Charts website using our automated Python script was very poorly formatted, as we essentially scraped raw HTML code from the Spotify Charts website. To rectify this issue, we had to perform a significant amount of data cleaning in order for us to utilize the data which we collected with our Python web scraper.

The table below shows the data which we read from the Spotify Charts website prior to any cleaning. As we can see in the output below, the data is nearly unreadable in its current format, as there are a ton of extraneous characters in many of the columns which need to be removed in order for us to access the information which we actually care about.

To fix the multitude of issues with the raw Spotify Charts data which we collected, we needed to find and extract the particular parts of each cell which we cared about using in our report.

To do this, we used the `stringr` library to edit the contents of these cells using a series of regular expressions, which would help us chip away at the excess HTML code in these cells until only the necessary information which we required was left.

For the 'Streams' and 'Position' columns, this editing process was rather straightforward, as these columns each needed to represent numeric quantities, and the HTML code (aside from the number we wanted) did not contain any numeric characters. In order to extract the numeric quantities which we wanted from these two columns, we were able to directly extract the numeric characters from the HTML code outputs by using a regular expression to remove all non-numeric characters from the cells in these columns.

Since the 'Trend' column was not relevant towards answering our guiding research question, we simply removed this column from the dataset entirely, as there was no purpose in cleaning the values in this column when the cleaned values were not going to be used.

Similarly, we removed the 'Track' column from the raw data which we collected using our Python web scraper. Although the information in the Track column is relevant, the information is recoverable based on the unique Spotify ID of the track which we were able to extract from the information in the 'Source' column. The information in the 'Track' column was added back into our dataset later when we combined our cleaned dataset with the information retrieved via the Spotify API.

Lastly, the most important column which we needed to clean was the 'Source' column. The HTML code

Table 4: Sample of the raw Spotify data, after the data cleaning process

track_id	country	country_code	begin	end	plays	song_rank
7rglLriMNBPAyuJOMGwi39	Canada	ca	2021-12-31	2022-01-07	1110076	1
02MWAaffLxlfxAUY7c5dvx	Canada	ca	2021-12-31	2022-01-07	1066897	2
5PjdY0CKGZdEuoNab3yDmX	Canada	ca	2021-12-31	2022-01-07	1064281	3
4fouWK6XVHhzl78KzQ1UjL	Canada	ca	2021-12-31	2022-01-07	980807	4
5Z9KJZvQzH6PFmb8SNkxuk	Canada	ca	2021-12-31	2022-01-07	970135	5

inside of this column corresponds to a working embed link for the Spotify track which was usable on the Spotify Charts website, similar to the contents of the "Spotify Link" columns in the tables hosted on the GitHub.io website for this project.

Within the framework of Spotify, and their API in particular, each track is given a unique track ID, which corresponds to a URL for the track, as well as an identifier for that track. The `spotifyR` API relies on these track IDs in order to retrieve information about the particular track, such as the track length, artists, album, and the three metrics which are relevant for measuring the song's happiness; it's danceability, energy, and valence. To extract the track IDs from the HTML code, we first used a regular expression to extract the "https://open.spotify.com/track/7rglLriMNBPAyuJOMGwi39" URL from the 'Source' column, and we then extracted the unique identifier from this URL, by removing everything before the final slash character.

The table below depicts a sample of the Spotify data after these cleaning procedures. As we can see, it is clearly much simpler than the initial data, and the data in this table is actually readable by humans, as opposed to being strictly machine-readable.

After we completed cleaning the raw data which we collected, the only remaining step was to collect information about these tracks, and append this data to the table, in order to create a larger table which contained all of the necessary information which we need.

This data collection process was done using the `spotifyR` package, which allowed us to access information about these tracks using the Spotify API, based on the individual track IDs which we had collected from the Spotify Charts website.

Table 5: Counts of Variables in our Spotify Dataset

	Variable Type	Present Values	Missing Values
Track_ID	character	183600	0
Track_URL	character	183600	0
Country	character	183600	0
Country_Code	character	183600	0
Weekly_Streams	double	183600	0
Weekly_Rank	double	183600	0
Week_Start	double	183600	0
Week_End	double	183600	0
Song_Name	character	183600	0
Artists	character	183600	0
Song_Length	double	183600	0
Danceability	double	183557	43
Energy	double	183557	43
Valence	double	183557	43
Occurrences	double	183600	0

Lastly, we will check if there are any missing values in our dataset, so that we can mitigate any potential issues which could be caused by empty cells within our dataset.

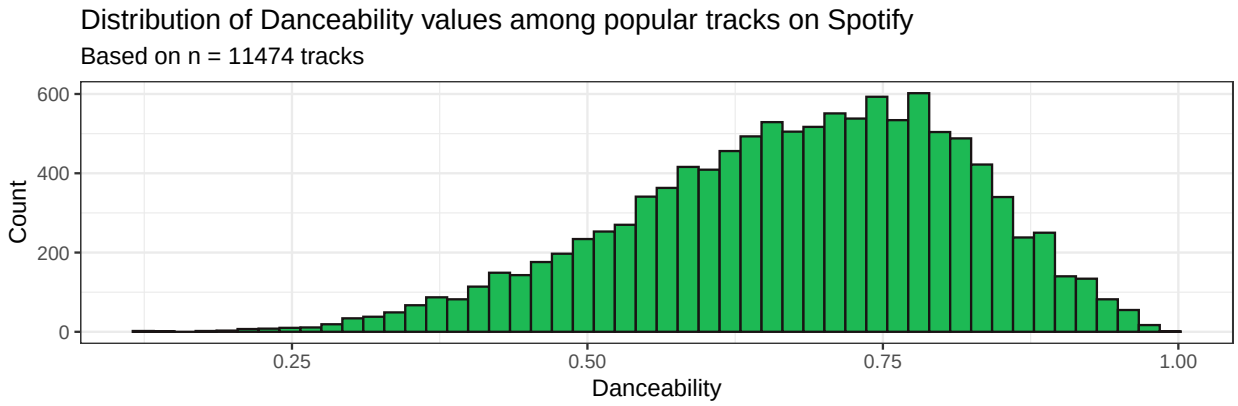
As we see above, there are only three columns which have any missing data, which are unfortunately the three most important columns for answering our research question.

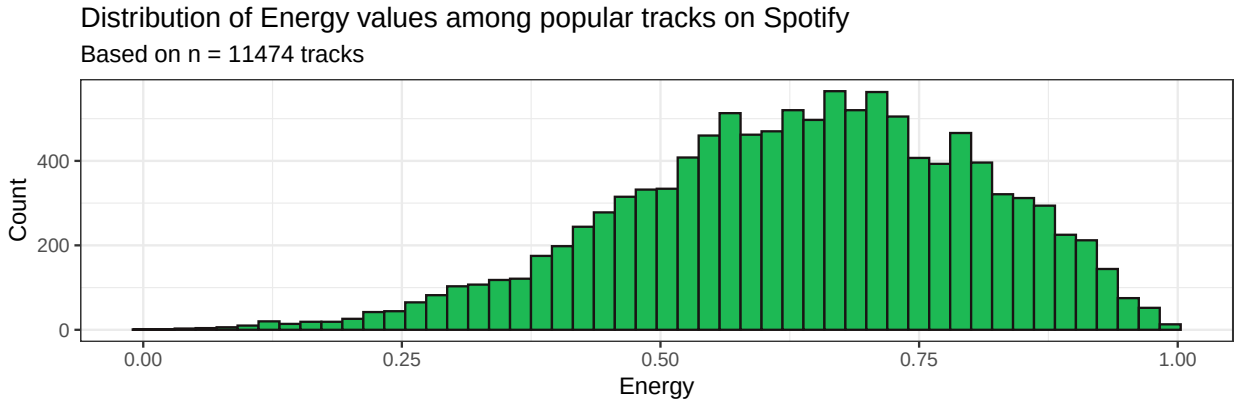
The existence of these 43 missing values in the Danceability, Energy, and Valence columns of our dataset has likely occurred because there are a handful of songs which do not have available values for any of these three metrics. As the number of missing values in our dataset is very small, and these rows all still contain valuable information, we will not be directly removing the rows which contain missing observations from our dataset altogether, and we will instead only remove them when the missing observations may cause errors, such as in the calculation of summary statistics for our variables of interest.

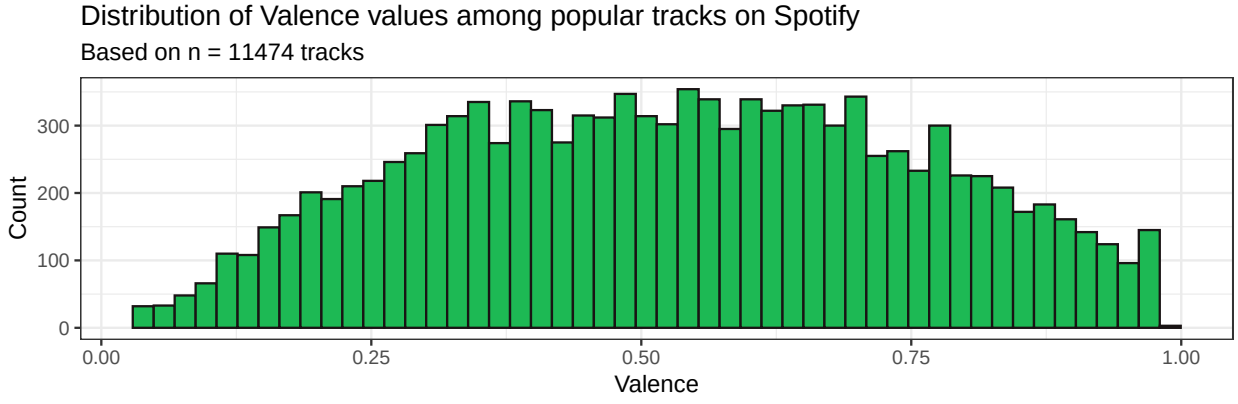
Data Exploration

Now that all of our data has been collected and cleaned, we can begin to explore the actual data which we are examining for the purposes of this report.

First, we will create histograms of the distribution of our three chosen metrics to measure the happiness of different songs. These histograms are displayed in the output below.



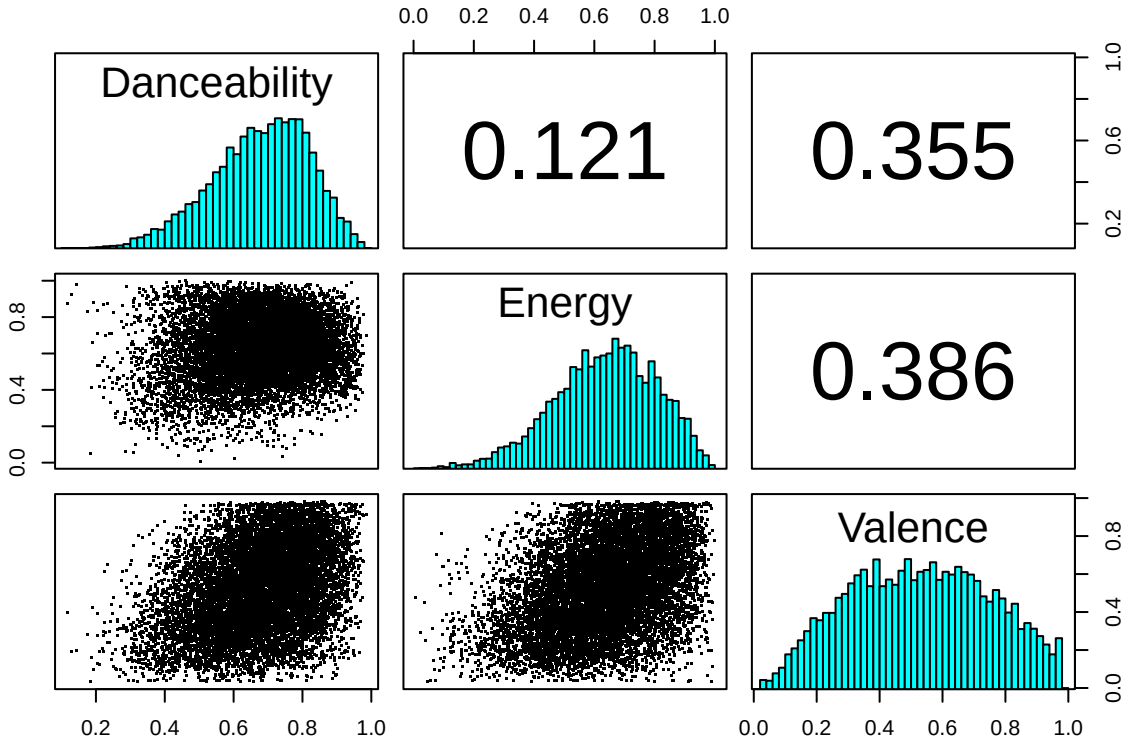




Next, we will examine the correlation between the three metrics which we have chosen to measure the overall happiness level of a song.

As we can see in the histograms above, as well as the example data based on a select few tracks by Kanye West, these metrics are not always perfectly in line with one another. However, one might expect that overall, these metrics are correlated with one another, as the concepts of danceability, energy, and valence described in the Spotify API documentation seem to overlap with one another.

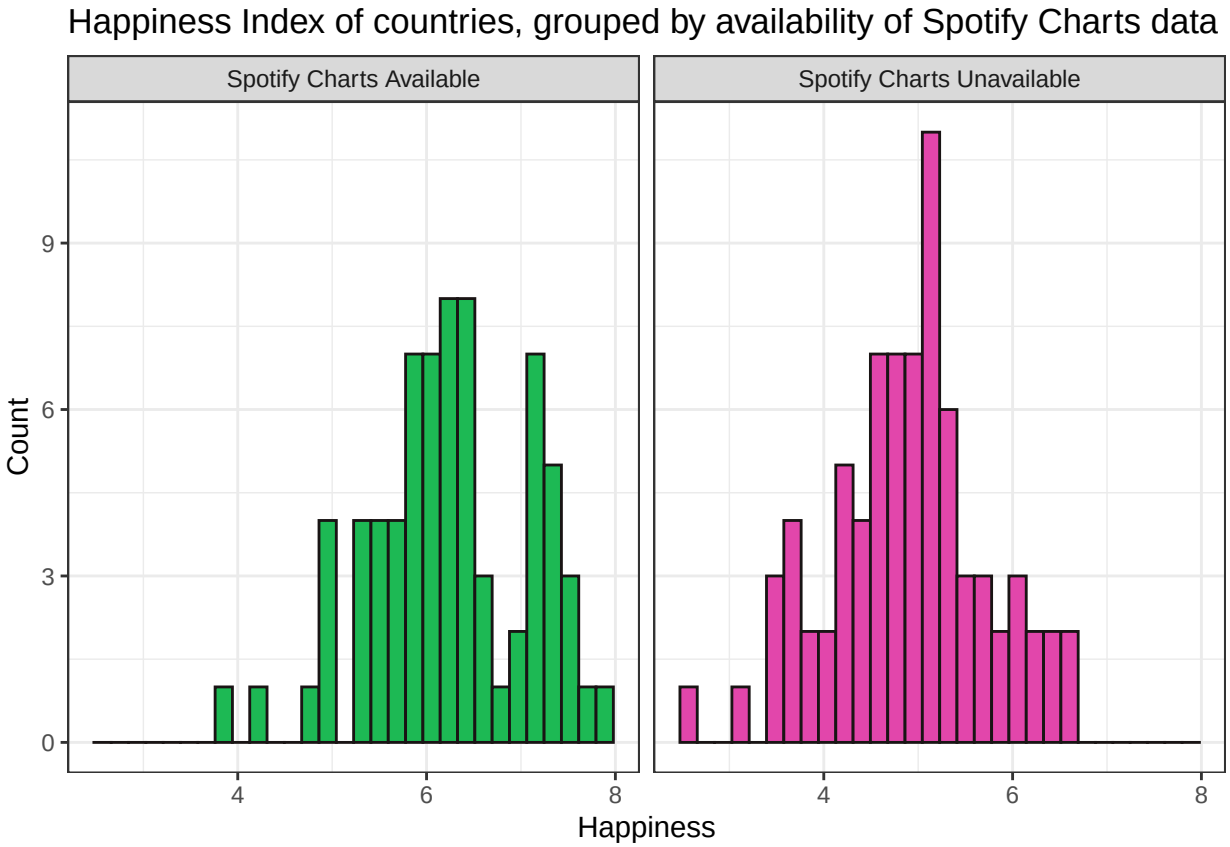
The plot below depicts histograms for each of these three values, alongside scatter plots which compare two of these metrics at a time, and the overall correlation coefficients between each pair of these three metrics.



As we see in the output above, these three metrics all appear to have some level of positive correlation with one another. However, these three variables are not particularly highly correlated with one another, as the largest correlation between these song happiness metrics is between the Energy and Valence of a song, with a correlation coefficient of 0.3862557.

Next, we will examine the distribution of the happiness index values among the different countries which had data available on the Spotify Charts website. The 2021 World Happiness Report dataset contained estimates of the average happiness levels for 149 different countries across the world. However, there were only 72 countries which had data available on the Spotify Charts website that we were able to collect.

The histogram below displays the distribution of happiness levels among the countries of the world, grouped based upon whether or not there was available data for that particular country on the Spotify Charts website.



As we can clearly see in the plot above, there is a massive disparity in the happiness levels of countries which had data available on the Spotify Charts website and those which did not have data on the website.

One likely explanation for this massive disparity between these two groups is that countries which were marked as 'happier' tended to have higher levels of economic and technological development than the countries which have lower happiness index values. Countries which tend to have large amounts of people with access to Spotify may tend to have higher levels of economic and technological access, as Spotify requires its users to have stable access to the internet in order to stream tracks with their platform.

Results

The guiding question for this research project was "*Is there a correlation between countries' happiness levels and how happy the most popular songs are in that country?*"

To examine this question in greater detail, we created scatter plots which compared the average happiness levels in countries to the average danceability, energy, and valence of songs on the Spotify Top 200 Tracks charts in these countries.

In total, we created 6 plots to visually examine the association between these variables, as we created plots which simply examined the mean values of these three metrics per country, and we created an additional three plots which instead took a weighted mean of these metrics based on the number of streams which the track had on Spotify during the given timeframe. These 6 plots are all available on the Interactive Visualizations page of the GitHub.io website for this report.

As we saw in the Data Exploration section above, there is not a particularly high amount of correlation between the three metrics which we chose to quantify how happy a song on Spotify is, as the highest correlation value among the three pairs of metrics was only 0.3862557.

Since these metrics do not have particularly high levels of correlation, we can safely create models which incorporate multiple metrics at once as predictors without a fear of multicollinearity skewing our results.

Model Creation

In addition to these six visualizations, we will also create four new models to compare the happiness levels in different countries to the values of these three Spotify song metrics. As we did when creating the interactive visualizations for these metrics which are displayed on the GitHub.io website, we will create two models which are based solely on the means of these metrics per country, and another two models based on the means per country, weighted by the number of plays which the track had in that particular country.

Using the Spotify data which we have collected, we will create a strictly linear model to predict the happiness of a country based on the three Spotify song metrics, as well as a model which takes into account second and third-order interaction terms among these three variables.

The 'strictly linear' model is a model which follows an equation of the form

$$\text{Happiness} = \beta_0 + \beta_1 \cdot \text{Danceability} + \beta_2 \cdot \text{Energy} + \beta_3 \cdot \text{Valence}$$

, where Happiness represents the findings of the 2021 World Happiness Report, and Danceability, Energy, and Valence represent the means of these respective metrics per country.

The linear model based on the unweighted means will have the same overall formula as the model based on the weighted means, but the values of the β variables will be different among these two models.

Unlike the strictly linear model, the model which considers second and third-order interaction terms will take into account the products of two or three of these terms at once, on top of the values of these individual terms. This 'third-order model' can be represented by the following equation:

$$\begin{aligned}\text{Happiness} = \beta_0 + \beta_1 \cdot \text{Danceability} + \beta_2 \cdot \text{Energy} + \beta_3 \cdot \text{Valence} + \beta_{1,2} \cdot \text{Danceability} \cdot \text{Energy} \\ + \beta_{1,3} \cdot \text{Danceability} \cdot \text{Valence} + \beta_{2,3} \cdot \text{Energy} \cdot \text{Valence} + \beta_{1,2,3} \cdot \text{Danceability} \cdot \text{Energy} \cdot \text{Valence}\end{aligned}$$

To assess the quality of the four models which we have created that incorporate all three of our Spotify-based happiness metrics, we will examine the adjusted R-squared values of the fitted models.

The adjusted R-squared value of a model is a metric which compares how well a particular model fits an outcome compared to a model which does not include any predictors, and essentially just predicts the mean of the outcome variable each time. Adjusted R-squared models are used instead of "simple" R-squared values when a model contains multiple predictors, as the adjustment of the R-squared value is meant to account

Table 6: Adjusted R-Squared values for our four predictive models

Model Type	Adjusted R-Squared
Linear Model - Unweighted Means	0.007347
Linear Model - Weighted Means	-0.002087
Interaction Model - Unweighted Means	-0.002151
Interaction Model - Weighted Means	0.013546

for the fact that a model with a higher number of predictors should have better predictive accuracy, due to the increased complexity of the model.

The table above shows the adjusted R-squared values for each of the four models which we have created, based on the description of the models given above. Among these four models, the model which performed the best (in terms of adjusted R-squared) is the interaction model which used weighted means, which had an adjusted R-squared value of 0.013546, which is incredibly low.

In each of these four models, the effect of these the three predictors (Danceability, Energy, and Valence) on the overall happiness of a country appears to be negligible, if there even is any effect at all. The adjusted R-squared values for all four of these models are incredibly close to zero, which renders them as practically useless predictors for the variable which we aimed to predict using these values.

Conclusions

In this report, our primary guiding question was to determine if there was a relationship between the happiness of popular music in a particular country and how happy the citizens are in that country.

Based on the visualizations on the GitHub.io page, and the adjusted R-squared values for our models which we computed in the previous section, it is safe to say that there is at best a spurious correlation between how happy popular tracks on Spotify are in a particular country and how happy the overall populace is within the country, among the countries which had accessible data on the Spotify Charts website.

Because the countries which had data available on the Spotify Charts website were significantly skewed towards wealthier and more populated countries, it is difficult to accurately say whether this complete lack of a relationship would generalize to the other countries which were unable to be included in this study, but it is likely that this effect would be negligible at best, if it did exist at all.

Limitations

Overall, the most noticeable limitation of this research project is the total lack of available data on the Spotify Charts website for more than half of the countries of the world, and the massive bias in country's happiness which was correlated with the presence (or lack thereof) of a country on the Spotify Charts website. Unfortunately, this limitation was essentially a factor of the design of the question itself, and this was not easily rectifiable within the scope of our research question.

Another potential issue with this research question is that collecting data about popular music in a country based on the Spotify Top 200 Tracks in that country requires an implicit assumption that Spotify users in a given country are representative of the country's population as a whole.

From a statistical and logistical perspective, there are multiple reasons why this assumption may not be a well-founded statistical assumption to make. Two particularly large issues with this assumption are:

- Due to socioeconomic factors which could limit access to technology and Wi-Fi, people in a given country who can access Spotify may not be representative of the whole country's population

- Spotify demographics in a given country may not be representative of the whole population, as Spotify tends to attract a younger demographic

Future Steps

Although the data which we collected from the Spotify Charts website and the World Happiness Report did not provide us with a satisfactory answer to our guiding question, it does not mean that these variables are entirely unrelated, and that the data which we have collected is worthless.

There was an unmistakable difference between the happiness index values of the countries which had data available on the Spotify Charts website and the countries which did not have Spotify data. A possible explanation for this difference is that the countries with higher Spotify usage tend to be countries where a large proportion of their citizens have easy access to the Internet, and are able to afford Spotify (if they are Spotify Premium users). The countries which fit this profile may also tend to be the countries which are rated more favourably on the World Happiness Report, likely in part due to their citizens' access to technology and financial freedom.

A future step for this research could be to incorporate data from the United Nations' Human Development Index to determine if there are other external factors which can affect Spotify usage in a country as well as the happiness values estimated by the World Happiness Report.

This additional information could potentially also help us control for the demographics relating to which people have access to Spotify and which people regularly use Spotify in a given country, as these limitations were unable to be addressed based on the data which we collected in the creation of this report.