8/14/2020

INSIGHTS ON MUSIC DIVERSITY – RECENT HITS

Group Project

*How top listened music genres evolve over time*

*A story about recent music hits*

**AUTHORS**

**L. B.**

**C. G.**

**C. N.**

**TERESA VENTAJA**

hIGHER DIploma in science in data analytics

INDEX

[1. Introduction 2](#_Toc48299416)

[2. Datasets definition & description 2](#_Toc48299417)

[3. Research case & Data preparation 4](#_Toc48299418)

[4. Analysis & Results 6](#_Toc48299419)

[a. Exploring music geographical bias – is there any artist location whose popularity is substantially more remarkable than others? 6](#_Toc48299420)

[b. Artists geographic location *time-lapse* 8](#_Toc48299421)

[c. Music diversity in a nutshell – how genre popularity evolves over time 9](#_Toc48299422)

[d. Music vibes - How consumers value the energy and positivity in a song. Is music tempo still a thing? 10](#_Toc48299423)

[e. Acoustic vs. Lyrics – How much consumers care about them 11](#_Toc48299424)

[f. Danceability & Popularity 12](#_Toc48299425)

[5. Infographic 14](#_Toc48299426)

[Sectioning 16](#_Toc48299427)

[BIBLIOGRAPHY 17](#_Toc48299428)

# Introduction

Both the type of music people is interested in and the technology used for listening to it evolve over time. In the 4th industrial revolution, we went digital, and analogic devices, such as vinyl, have become more of a vintage nostalgic item for collectors rather than a real competition for digital platforms when it comes to the consumption of music. Remarkable changes are more visible over decades, but the evolution of music consumption habits is fast enough to present differences in smaller periods of time.

The purpose of this project is to narrate how recent music hits trends are evolving over the period from 2010 to 2019. We divided this assignment into 5 sections. The next section gives details on what kind of data are we using and where have we got it from. Section 3 defines the research case, dividing it into manageable objectives aligned with the final purpose of our narrative. The following section analyses the data by means of visualizations and findings that emerged. Finally, Section 5 holds the commentary on an infographic that has been produced with the goal to resume the main findings of the project.

# Datasets definition & description

This project has been created based primarily on 2 datasets: “*Top Spotify songs from 2010-2019*” (Kaggle, 2020a) and “*Music artists popularity*” (Kaggle, 2020b). The main advantage of using those datasets is that it allows the extraction of interesting insights about both songs and artists, and the combination of those enriches the narrative.

The dataset “*Top Spotify songs from 2010-2019*” (hereafter “Spotify dataset”) contains insightful information on top 10 hits by year, but also some technical measures of the characteristics of the songs, which we will contrast and compare. This is our core data, the basis for building our story about recent music hits. The .csv file contains 14 columns and 604 rows, only one row containing missing data for some observations. This dataset is illustrated on the table 1.

*Table 1: Dataset Description: “Top Spotify songs from 2010-2019”*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Column name** | **Column description** | **Data type** | **Range** | **% missing data** |
| **(blank)** | column number | numeric - ordinal - discrete | 0 - 603 | 0 |
| **title** | Song title | character - nominal - categorical | 584 different categories | 0 |
| **artist** | band/artist name | character - nominal - categorical | 184 different categories | 0 |
| **top genre** | the genre of the track | character - nominal - categorical | 50 different categories | 0 |
| **year** | the release year of the recording (for this version of the song) | date | 2010 - 2019 | 0 |
| **bpm** | Beats Per Minute - The tempo of the song | numeric - ordinal - continuous | 43 - 206 | 0.17 |
| **nrgy** | Energy - The energy of a song: the higher the value, the more energtic | numeric - ordinal - continuous | 4 - 98 | 0.17 |
| **dnce** | Danceability - The higher the value, the easier it is to dance to this song | numeric - ordinal - continuous | 23 - 97 | 0.17 |
| **dB** | Loudness (measured in dB) - The higher the value, the louder the song | numeric - ordinal - continuous | (-15) to (-2) | 0.17 |
| **live** | Liveness - The higher the value, the more likely the song is a live recording | numeric - ordinal - continuous | 2 - 74 | 0.17 |
| **val** | Valence - The higher the value, the more positive mood for the song | numeric - ordinal - continuous | 4 - 98 | 0.17 |
| **dur** | Duration - The length of the song | numeric - ordinal - continuous | 134 - 424 | 0 |
| **acous** | Acousticness - The higher the value, the more acoustic the song is | numeric - ordinal - continuous | 0 - 99 | 0.17 |
| **spch** | Speechiness - The higher the value, the more spoken word the song contains | numeric - ordinal - continuous | 3 - 48 | 0.17 |
| **pop** | Popularity - The higher the value, the more popular the song is | numeric - ordinal - continuous | 0 - 99 | 0.17 |

*Source: Playlist machinery, 2020.*

The dataset “*Music artists popularity”* (hereafter “artists dataset”) focuses on the artists and contains meaningful information about the number of listeners each artist has. It is based on *MusicBrainz* and *Last.fm* data, including also tags and scrobbles. The .csv file contains 10 columns and 1466083 rows, some columns presenting a significant amount of missing data (over 50%). This dataset is illustrated on the table 2.

*Table 2 - Dataset Description: “Music artists popularity”*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Column name** | **Column description** | **Data type** | **Range** | **% missing data** |
| **mbid** | *MusicBrainz* ID | character - nominal - arbitrary | 1466083 unique values | 0 |
| **artist\_mb** | Artist name according to M*usicBrainz* | character - nominal - categorical | 1353002 categories | 0 |
| **artist\_lastfm** | Artist name according to *last.fm* | character - nominal - categorical | 986732 categories | 33 |
| **country\_mb** | Artist country according to *MusicBrainz* | character - nominal - categorical | 496746 categories | 55 |
| **country\_lastfm** | Artist country, based on *last.fm* tags | character - nominal - categorical | 185500 categories | 86 |
| **tags\_mb** | Artist tags on M*usicBrainz,* separated by semicolon (;) | character - nominal - categorical | 107858 categories | 92 |
| **tags\_lastfm** | Artist tags on *last.fm*, separated by semicolon (;), sorted by frequency decreasing | character - nominal - categorical | 379615 categories | 74 |
| **listeners\_lastfm** | Number of listeners on *last.fm* | numeric - ordinal - continuous | 91283 to 5.38m | 33 |
| **scrobbles\_lastfm** | Number of scrobbles on *last.fm* | numeric - ordinal - continuous | 268193 to 517m | 33 |
| **ambiguous\_artist** | TRUE if more than one artist shares the same *last.fm* page | logical | TRUE, FALSE | 0 |

# Research case & Data preparation

The purpose of this project is to narrate how recent music hits trends are evolving over the period from 2010 to 2019. To get an insightful understanding of it, we are going to look at the evolution of multiple song features. We will first explore geographical bias and music diversity, and then consumer habits. To get the appropriate level of detail on this analysis, we have defined the following **operational objectives**:

1. Exploring music geographical bias – is there any artist location whose popularity is substantially more remarkable than others?
2. Artists geographic location time-lapse
3. Music diversity in a nutshell – how genre popularity evolves over time
4. Music vibes - How consumers value the energy and positivity in a song. Is music tempo still a thing?
5. Acoustic vs. Lyrics – How much consumers care about them
6. Danceability & Popularity

In terms of **data preparation**, to accomplish the objective of geographically visualize the top 20 genres by Top10 single count (objective 1), some pre-processing of the two datasets has been performed in Excel. To add the associated country to the “*Top Spotify songs from 2010-2019*” dataset, the V-lookup function in excel has been utilized in order to retrieve the country of the artist.

Few missing values have been identified due to either difference in spelling and/or actual missing data; to manage missing values, decision has been taken to manually check Country of origin for the ones that are most common/with the greatest number of listeners. Three were not found hence the corresponding observations have been disregarded and removed from final total count. The resulting dataset thus contains the same columns as the original plus Country column.

Data pre-processing has been performed in the “Artists dataset” in Excel to make feasible the visual representation of cumulative number of listeners analysed by country. To achieve this result, a pivot table has been inserted and the variable Country and sum of listeners selected, such that each row could show a country and the relative cumulative number of listeners.

For Objective 2, we decided to merge datasets, as nationality of the artists held in one file has been linked to the Spotify dataset using the powerful VLOOKUP function available in Microsoft Excel to link artist name to their country of origin. Four rows in the Spotify dataset displayed the value “0” as a result of the VLOOKUP. They have been deleted.

25 rows returned N/As due to special characters contained in the name of the artists (i.e. *Beyoncé*). In this case, nationality of the singers has been searched on the Internet and manually added. Data have then been manipulated to fit the time series format required to produce a dynamic bar chart race on the free online software Flourish (2020) with the purpose of counting the top 10 nationalities of artists in the Spotify dataset by year. The total number of records left after pre-processing and filtering has been 583 for that visualisation with total count adding up for each year starting from 2010 to 2019. This has been achieved by applying SUM function in Excel.

# Analysis & Results

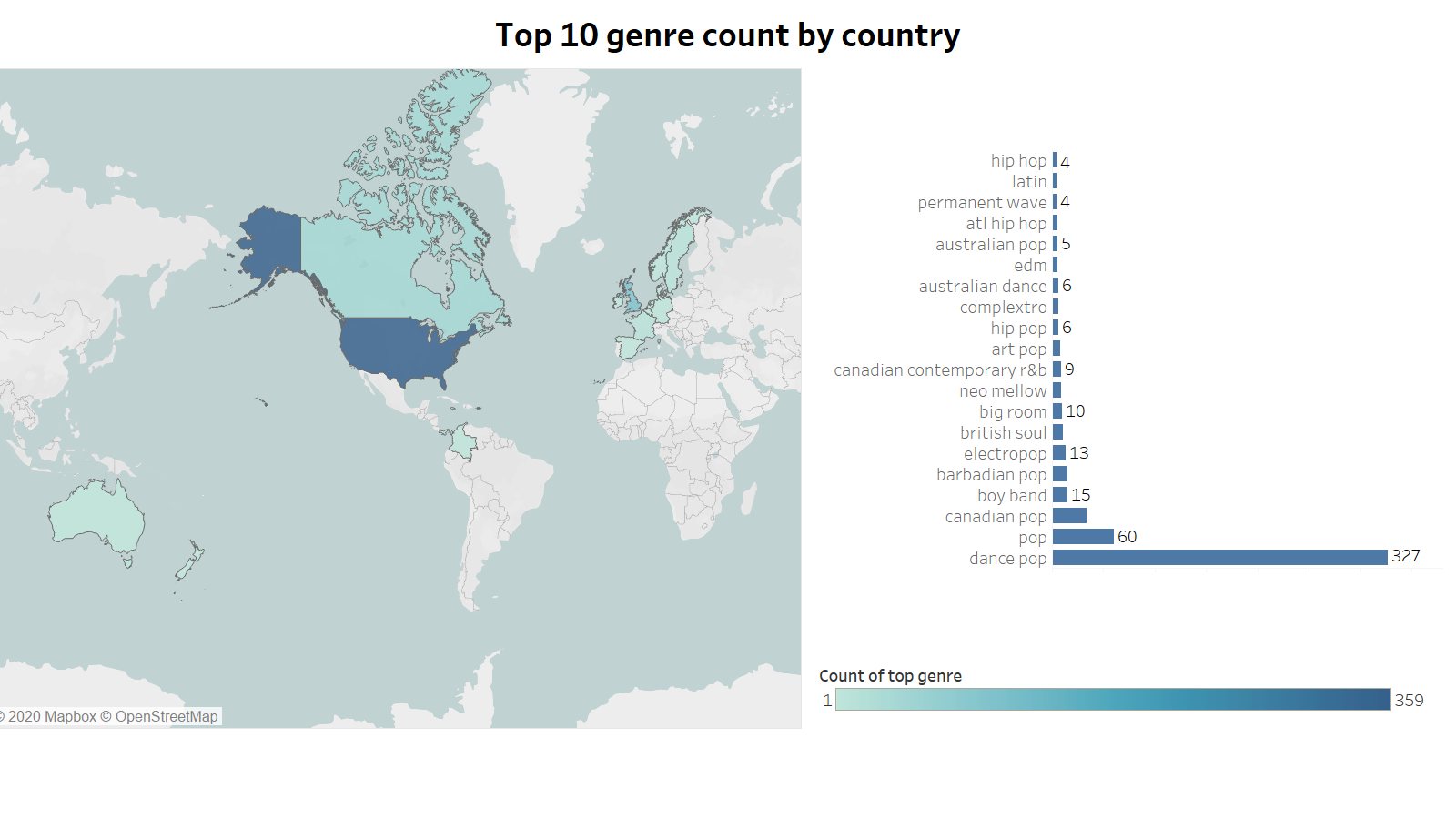
## Exploring music geographical bias – is there any artist location whose popularity is substantially more remarkable than others?

Top 10 genre count by country

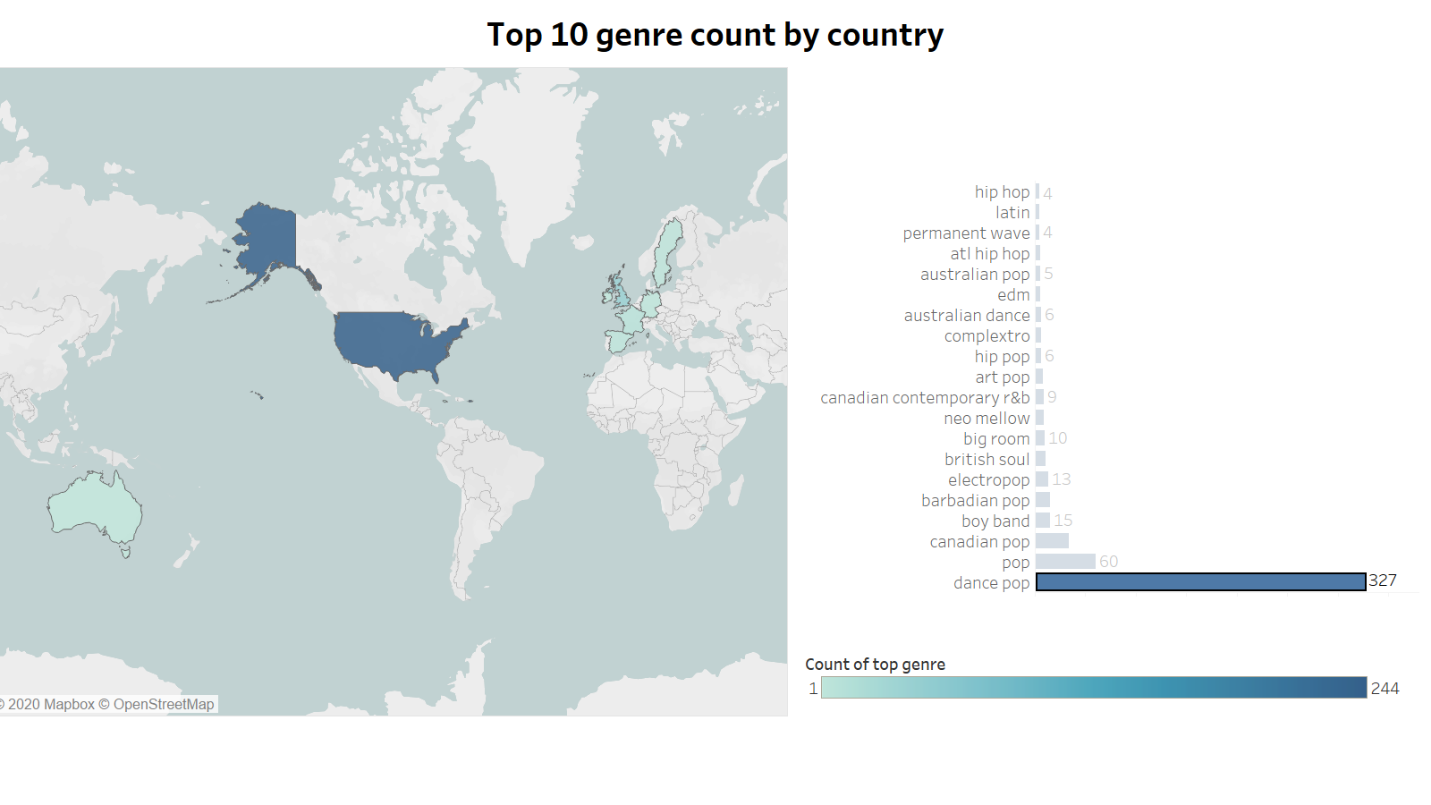
The variable Country has been selected, set as geographical variable in Tableau, such that Longitude and Latitude variables could be generated; also Top genre variable has been selected and set as dimension and Count, such that the number of times each genre would be present in the dataset could be captured. The resulting dashboard has been created including also a bar chart with the Top 20 genres per number of singles, to be used as filter to see Top genre by country.

As clearly shown in the dashboard, dance pop is the most common genre, with a count of 327 songs in the Spotify dataset, out of which 244 belongs to artists based in U.S, followed by 45 artists in UK. With these results it is possible to states that the genre dance pop is clearly dominated by the Anglo-American culture.

*Figure 1. Top 10 genre by country.*

**

*Figure 2. Top 10 genre count by country*

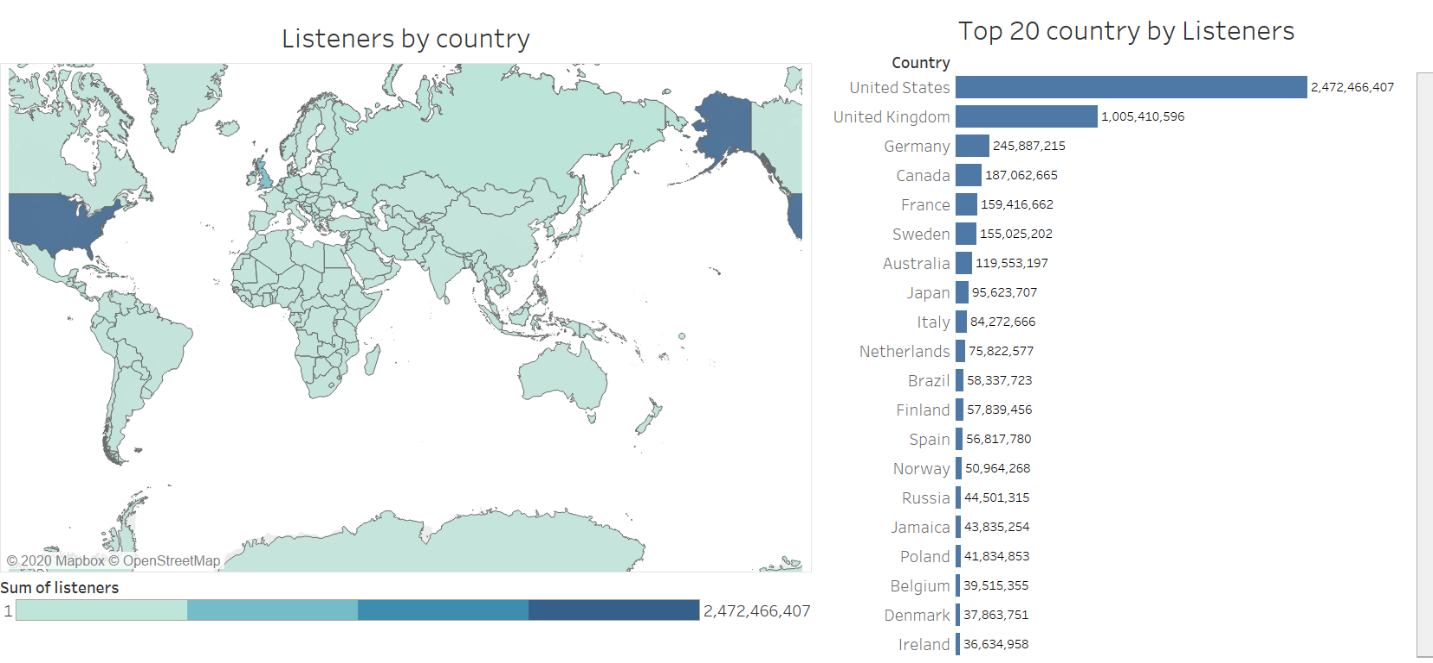


Listeners by country

In order to limit the visual representation just to the most relevant countries in terms of number of listeners, the top 20 countries have been showed in the bar graph. As we can see from the Map, US is clearly dominant in terms of number of listeners (2,472,466,407) for Top10 songs, followed by UK with 1,005,410,596 and Germany with 245,887,215.

Following the results from both dashboards, it can be stated that the music industry is clearly dominated by the United States with a number of listeners for Top10 songs which is more than double the one reported by the U.K at the second position and almost 10 times higher than the Germany in 3rd position.

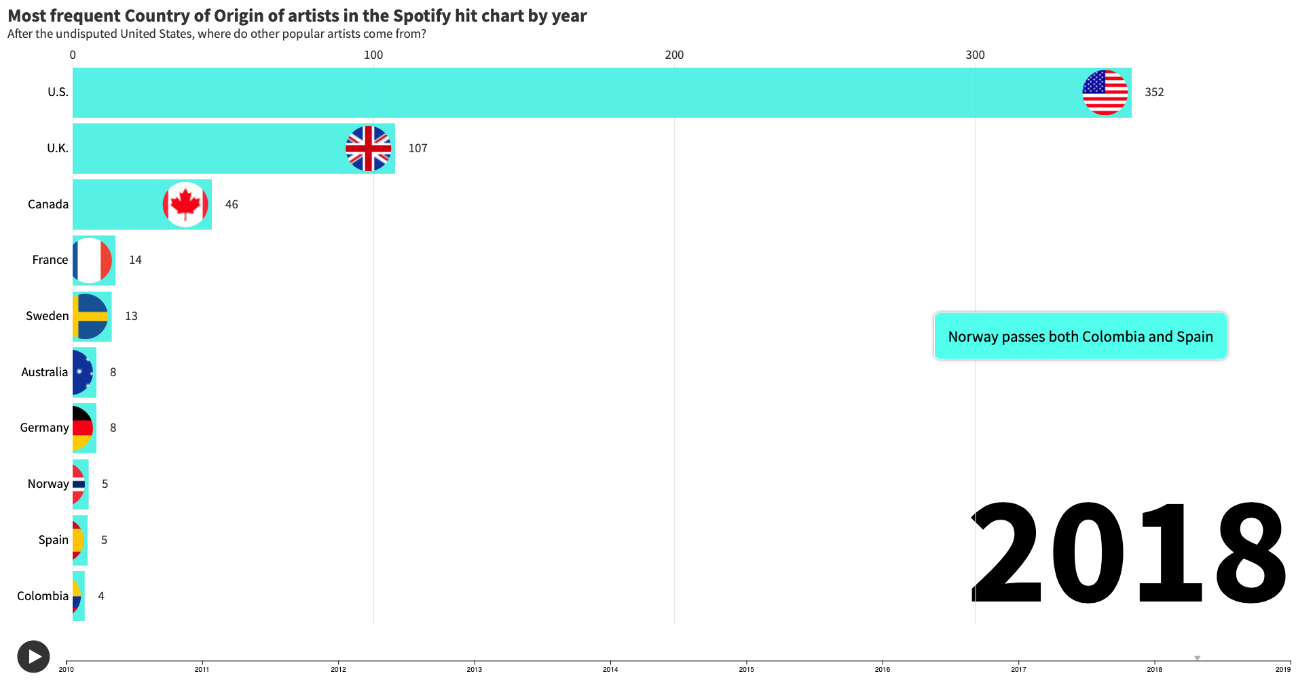
*Figure 3. Dashboard on number of top 10 Spotify songs listeners by country.*



## Artists geographic location *time-lapse*

The bar chart race created using Flourish to visualise the top 10 most represented nationalities of the artists in the top chart by year allows, thanks to its dynamic nature, to spot trends that otherwise could not be easily detectable. Apart from the United States being by far the most represented country throughout the whole timeframe analysed, France, the only non-English speaking country able to make the top 3 between 2010 and 2011, was undertaken by Canada in 2012, event that consolidates the primacy in the world “music” of the English language, with United Kingdom constantly occupying the second spot. It is from 2013 onwards that also Nordic artists started to attract the attention of the fans over Latin singers with Sweden surpassing both Spain and Colombia. Norway also overtook the pair between 2017 and 2019, to certify a growth in likeability for Scandinavian musicians.

*Figure 4: Flourish Bar Chart Race of most frequent country of origin of artists in the Spotify hit chart by year.*



Link source: <https://public.flourish.studio/visualisation/3384395/>; Flourish, 2020.

## Music diversity in a nutshell – how genre popularity evolves over time

This visualisation represents each genre with a colour coded bubble, whose size represents the popularity of the genre with each bubble labelled with the genre name. The years were passed into the Pages function to control the year represented, with a slider on the right-hand side of the screen which allows for both manual year selection and a slideshow effect. Hovering the mouse over each bubble brings up a textbox containing the genre, year and popularity (pop) of bubble. An Index() function was built to filter only the top 5 results for each year in each page.

*Figure 5. Top genres by year (in this case 2016 is displayed).*

A picture containing screenshot

Description automatically generated

## Music vibes - How consumers value the energy and positivity in a song. Is music tempo still a thing?

Energy and positivity in a song can both be measured, but they are more subjective concepts than tempo (beats per minute). For instance, maybe the rhythm of a song is accelerated, so it is calculated to be a positive song, but the notes and keys used inspire sadness instead. Even though Spotify measurements are unlikely to spot such those subtle changes, we expect the overall perception of energy and positivity to be a relevant figure, because the population size is large enough to compensate small inaccuracies.

We decided to perform this analysis for 2 main reasons: firstly, because now genres such as pop, rap, techno, R&B, punk, etc. are trendy, and overall, they represent more upbeat music than in earlier decades. We can find such kind of upbeat music in a variety of everyday scenarios, such as in clothing stores, gyms, malls, etc. Secondly, because now there is a new trend on pop songs where the chorus are voice (no instrumental) with low energy (YouTube, 2017), whereas in the 2000 the chorus were the climax of the song, full of energy. We are going to check what how is this trend progressing:

*Figure 6. Songs’ energy, positivity and tempo evolution over time.*

A close up of a fence

Description automatically generated

We can see that the average tempo does not fully matches the changes on energy and positivity. For instance, in 2014 we have a very high value, but the energy is quite low, and positivity is among the lowest. However, the overall trend is a reduction on average of the tempo, energy and positivity in the time series. We can conclude than all variable (energy, positivity and tempo) are decreasing on average as time goes by.

## Acoustic vs. Lyrics – How much consumers care about them

This visualisation shows a scatterplot of average Acoustic vs average Speech for each genre of music per year, each scatterplot point was a square whose size represented the popularity of the acoustic/speech combination. Each scatterplot square is colour coded to represent the genre of music that it is calculated for. A column along the right-hand side of the screen can be clicked to filter for each year and the colour coded column can filter for reach genre.

*Figure 7. Scatterplot of average Acoustic vs average Speech for each genre of music per year.*

A screenshot of a computer

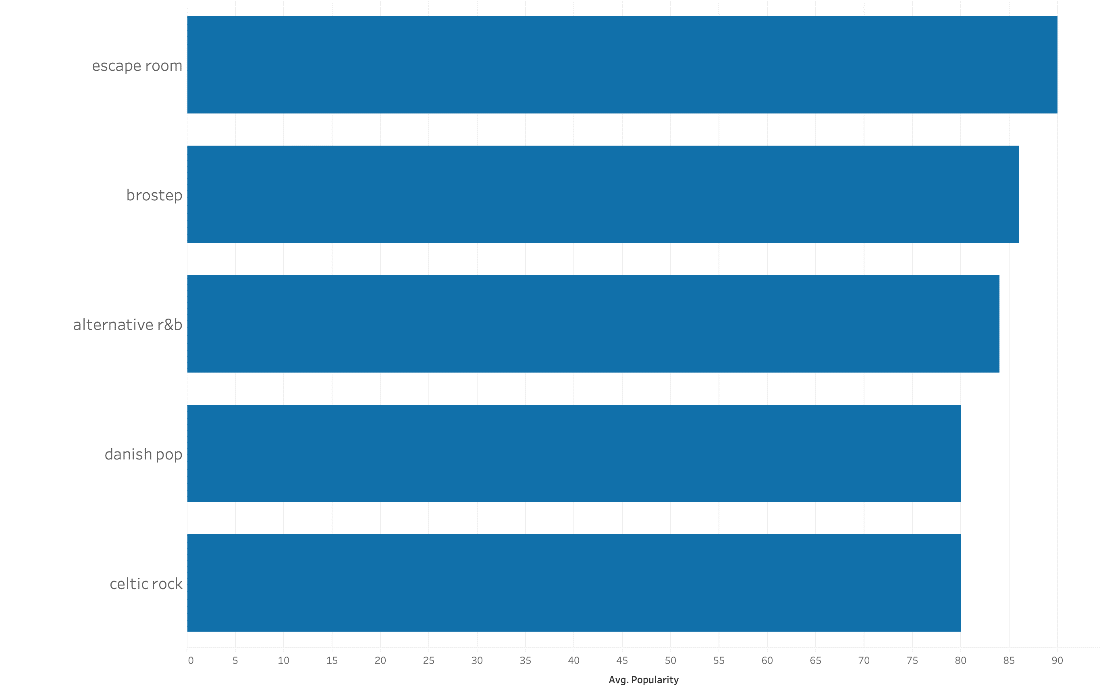
Description automatically generated

## Danceability & Popularity

Civilizations from different areas in the world have very different habits but a constant is that at social gatherings everyone has their own traditional dances. For this reason, it has been decided to investigate whether a song that stimulates individuals to come together, move and follow the rhythm can positively influence the popularity of that track.

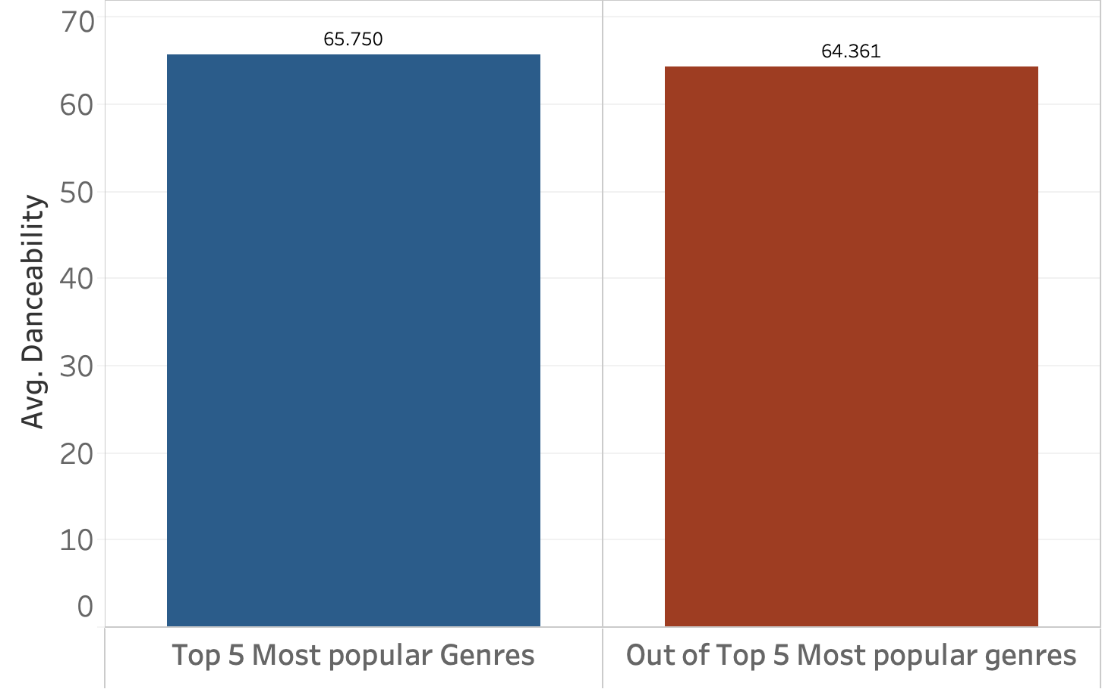
To attempt the above, genres of the songs in the Spotify dataset have been grouped and the average popularity calculated utilizing the functionalities of Tableau. The resulting top 5 genres, that have all an average popularity of above 80, are displayed below.

*Figure 8. Top 5 most popular genres measured using variable popularity.*



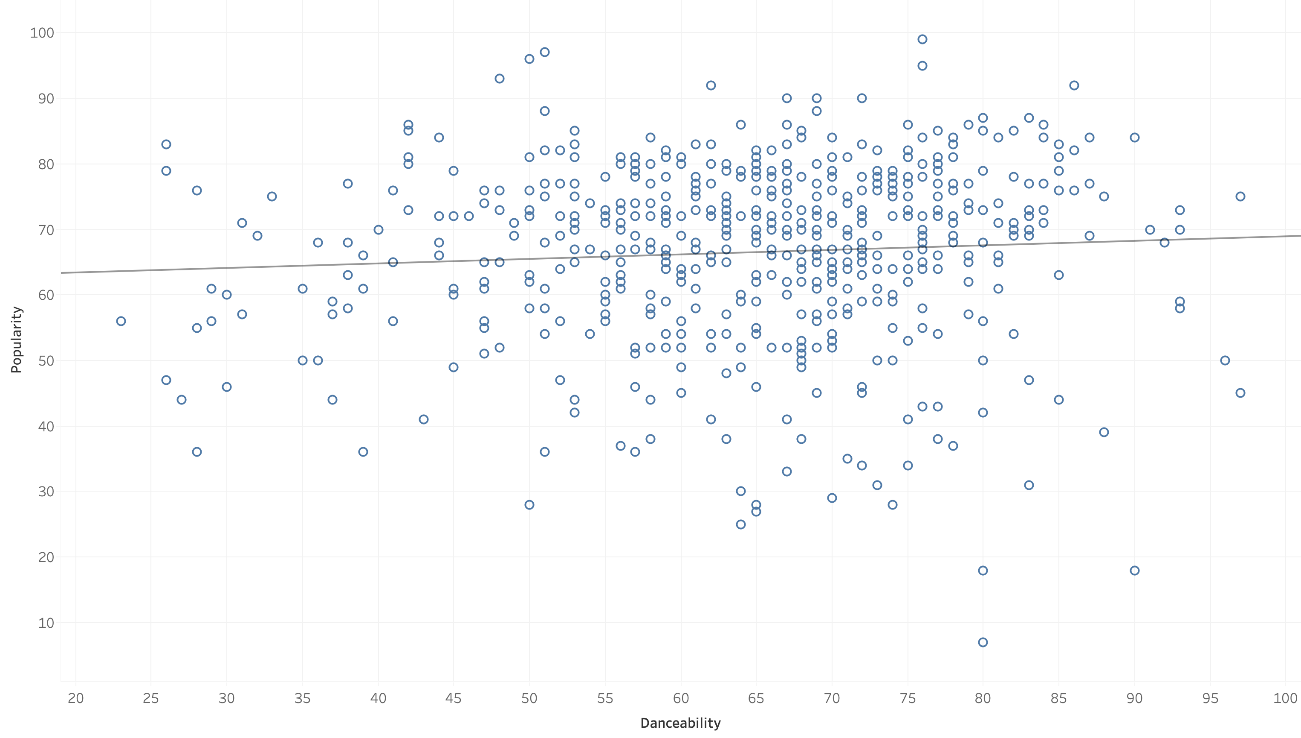
The top 5 most popular genres have been subsequently grouped together. All the other genres have instead been assigned to another group. The danceability measure has been introduced and the average for it has been calculated across the two congregations. Popular genres average slightly higher in danceability, but the overall difference is small, which leads to assume that the popularity of a song is not strongly influenced by whether individuals can actively move at its rhythm.

*Figure 9. Average danceability of top 5 most popular genres vs non top 5 most popular genres.*



Further, a regression model has been conceived selecting as independent variable popularity, and as dependent variable danceability. Five outliers with value 0 in either popularity or danceability were filtered out.

*Figure 10. Regression model with danceability as independent variable and popularity as dependent variable.*



The model reveals a p-value superior to α=0.05, which does not allow for rejection of the Null hypothesis that there is not statistically significant relationship between the two variables selected. Indeed, this conclusion is supported by the small R-Squared value that suggest that danceability only explains 0.46% of the variability in levels of popularity (McClave et al., 2014). The regression equation implies that for an increase of one unit in danceability, popularity will increment by 0.069 units. Due to the model not being significant, interpreting the regression equation adds little if no value.

*Table 3. Regression model results.*

|  |  |
| --- | --- |
| **P-value:** | 0.129155 |
| **Equation:** | Popularity = 0.0690006\*Danceability + 62.045 |
| **R-Squared:** | 0.004627 |

Following the results obtained after applying inferential and descriptive statistics to the data, it can be concluded that for a song to become popular, it does not necessarily need to have a danceable beat.

# Infographic

The infographic has been designed aiming to summarize the main insights from the report; the software utilized is Piktochart (2020).

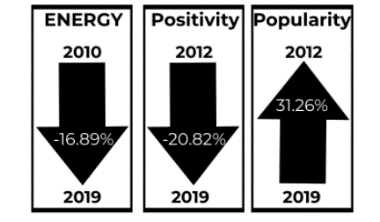
The chosen **typography style** has been selected because of the clarity and consistency with the overall report designs, the goal was to achieve a simple yet modern style. The selected font named “Chivo” in black, in its bold version for the main title and for the headings in each section.

As defined by Google fonts (2020): “*The strength of Chivo Black makes it ideal for highlights and headlines. Chivo Regular's elegance makes it ideal for combining with the strength of Chivo Black for continuous reading. Its design details make it an indispensable ally for any designer*.”

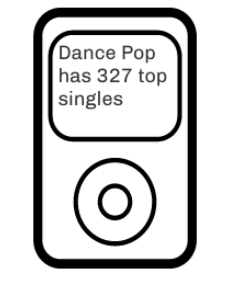
The **colour** background has been left white, to further highlight the black icons and the text. We decided to combine black and white because it is how music notes are presented in the score. We also have blue and red, inspired by the American flag, the choice has been driven by the fact that U.S turned out to be the main country leading the Top10 charts and artists’ country of origin. Also, these 2 colours match very well with the scale of greys and black/white.

Besides the two graphs already commented, the **icons** inserted are:

* Arrows to visually highlight the increase or decrease over time of Energy, Positivity and Popularity



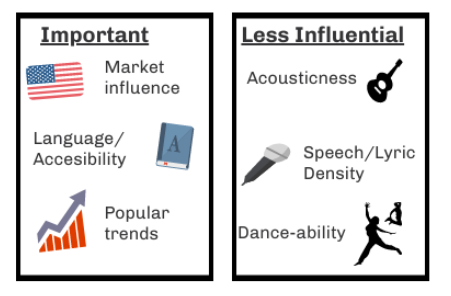
* Stylized iPod with number of top genre, to visually support the idea of popularity of Dance pop among Top10 songs. Genre mostly representing commercial music industry and wide-spread and popular, I.e. a genre you can find in every iPod



* American Flag, to highlight the supremacy of United States in the Music Industry, not only from the Offer side, in terms of artists and songs, but also from the Demand side, in terms of cumulative listeners



* Icons representing main elements influencing the popularity, U.S flag to highlight market influence, a dictionary representing the language accessibility, trend graph to show popularity, a guitar representing the acousticness aspect, while a microphone to represent the lyrics aspect and a dancing figure to represent the danceability.



## 

## Sectioning

The infographic is separated in the middle with a white band. The reason why we decided to do so is because it is meant to fit in a flyer. Front page of the flyer would be the half top part of the infographic, and back page would be the half bottom part of it.

**Section 1: Front of the flyer**

We went into detail on consumer habits, such as the number of unique genres per year, which gives an idea on how diverse the most popular hits are, which has not changed much since 2012.

A screenshot of a cell phone

Description automatically generatedThen we shared some insights on the changes on song Energy, Positivity and Popularity over time and shared a graph showing the acousticness vs. lyrics, where we can clearly see that there is no obvious association.

**Section 2: Back of the flyer**

A screenshot of a cell phone

Description automatically generatedThis section contains general insights about the entire dataset, starting with the most popular genre, the influence of the US in nowadays’ music, and information about what determine popularity in songs.

The last two pieces of icons summarize the most remarkable information in 6 visual snippets, acting as a closure section. Important to notice that these 2 sections complement each other. We understand the infographic as a whole, rather than a sum of unlinked pieces of information. Even though they talk about different stats, they both combined provide a comprehensive overview of our data.

# BIBLIOGRAPHY

1. Flourish (2020) “*Most frequent country of origin of artists in the Spotify hit chart by year*”. Lorenzo Battistin. Available at: https://public.flourish.studio/visualisation/3384395/ (created 11th August 2020).
2. Google Fonts (2020) *Google Fonts*. Available at: https://fonts.google.com/ (Accessed: 13 August 2020).
3. Kaggle (2020a) “*Top Spotify songs from 2010-2019*”, datafile.csv. Available at: https://www.kaggle.com/leonardopena/top-spotify-songs-from-20102019-by-year?select=top10s.csv (viewed 05th August 2020).
4. Kaggle (2020b) “*Music artists popularity*”, datafile.csv. Available at: https://www.kaggle.com/pieca111/music-artists-popularity (accesed 05th August 2020).
5. McClave, J. T., Benson, P. G., & Sincich, T. (2014). *Statistics for business and economics*. Upper Saddle River, NJ, Pearson/Prentice Hall.
6. Piktochart (2020), Available at: <https://piktochart.com/>.
7. Playlist Machinery (2020), “*Organise your music*”, Playlist Machinery. Available at: http://organizeyourmusic.playlistmachinery.com/ (accesed 05th August 2020).
8. YouTube (2003) *“Crazy in Love”,* YouTube video, added by Beyoncé [Online]. Available at: https://www.youtube.com/watch?v=ViwtNLUqkMY (accessed 05th August 2020).
9. YouTube (2017) “*Taylor Swift – Look what you made me do chorus loop*”, YouTube video, added by SkyLINX [Online]. Available at: https://www.youtube.com/watch?v=A-FQh3V05Ko (accesed 05th August 2020).