|  |  |
| --- | --- |
| A picture containing screen, clock, building, window  Description automatically generated (Source:, n.d.)  **visual music analysis**  *Higher Diploma in Science in Data Analytics* | ***Million Songs Dataset***  A brief visual music analysis  **Teresa Ventaja** |

INDEX

[1. Dataset selection 1](#_Toc45479410)

[2. Data description 1](#_Toc45479411)

[3. Essential literature review – research case 4](#_Toc45479412)

[4. Objectives 4](#_Toc45479413)

[5. Analysis 4](#_Toc45479414)

[Selection of visualizations 4](#_Toc45479415)

[1. Scatter Plot 5](#_Toc45479416)

[2. World map 5](#_Toc45479417)

[3. Dashboard: bar plots & box-and-whisker plots 6](#_Toc45479418)

[Report of results / findings 9](#_Toc45479419)

[1. Song loudness and duration over time 9](#_Toc45479420)

[2. *(Discarded – only for contrasting purposes)* Popularity: artists and songs. How do they differ? 9](#_Toc45479421)

[3. Where the majority of artists covered in the dataset live 9](#_Toc45479422)

[4. Technical music measures and their level of accuracy 10](#_Toc45479423)

[6. Bibliography 11](#_Toc45479424)

# 1. Dataset selection

In this project I decided to use a sample of the Million Song Dataset, the data collection of one million popular contemporary songs. This machine learning project for feature analysis and metadata was produced by The Echo Nest in collaboration with LabROSA, also partially funded by the National Science Foundation of America (NSF). They aim to promote research on algorithms with the ability to scale to commercial sizes related to Music Information Retrieval. Audio is not included in the data, only derived features[[1]](#footnote-2).

Since the dataset size is 280 GB, unreasonably big for the purpose of this project, I downloaded a random subset of 10.000 songs (1% of the original dataset). Even though the analysis I am going to perform is not going to be as thorough as if I had worked with the original dataset, the randomization of the sample allows to get a sense of what the entire dataset looks like, because it is big enough.

The reason why I selected this topic is because I like music, and because it represents a challenge due to the technical features measured that I am not familiar with. For me to succeed on the elaboration of this project I need to research and get out of my comfort zone, and that will have an impact on my self-confidence when it comes to learning data analytics.

# 2. Data description

First, I have collected information on the description of the different columns. I noticed immediately that 5 columns did not contain any relevant information, or did not contain any information whatsoever, so I eliminated them[[2]](#footnote-3). As a schematic summary of the data description, I have created a simple table (Table 1). Important to note that **the data was collected on December 2010, so the findings I will check will refer to year 2010 and not 2020**. To complete the description, I just need to add that "*Terms*" are tags provided by The Echo Nest, and they come mostly from blogs, whereas "*Mbtags*" are musicbrainz tags, selected by humans to a particular artist[[3]](#footnote-4).

[[4]](#footnote-5)

In the following table (Table 2) I will present a more comprehensive table with some information on the data type, the range, and the percentage of empty rows. Also, the column names and the total rows and columns on the original dataset:

[[5]](#footnote-6)

# 3. Essential literature review – research case

In music, the **time signature** informs about the meter of the piece to play. It is the number of beats per measure. The most used time signatures are 4/4, 3/4, 2/4 and 6/6[[6]](#footnote-7). In our dataset there are 8 numerical values assigned to different time signatures, but we have no information about the equivalence with real values. **Bar or measure** refers to a segment of time holding a certain number of beats[[7]](#footnote-8). Beats can be defined as "a regular, repeating pulse that underlies a musical pattern”, whereas **tempo** is the speed at which music patters play back and it’s measured in beats per minute (BPM)[[8]](#footnote-9). A **tatum** is the "lowest regular pulse train that a listener intuitively infers from the timing of perceived musical events: a time quantum. It is roughly equivalent to the time division that most highly coincides with note onsets"[[9]](#footnote-10). This concept helps to quantify tempo mathematically.

Another concept we have in the dataset is the minor and the major **scales** (mode in this dataset). The greatest difference them is the 3rd note. More specifically, the distance between the root note and the 3rd note: In major scale, the distance is 2 steps, also called a major 3rd interval and in minor scale, the distance is one and a half step, also called a minor 3rd interval[[10]](#footnote-11). In practice, generally speaking songs played in major scale tend to sound “happy”, whereas in minor case might sound “sadder”. However, most songs contain notes in both scales, as much as they have high and bass notes. Finally, the **key** is the major or minor scale around which a piece of music revolves. For instance, a song played in the ‘key of C major’ revolves around the seven notes of the C major scale – C, D, E, F, G, A, and B. That means the fundamental notes making up the song’s melody, chords, and bassline are all derived from that group of notes[[11]](#footnote-12).

# 4. Objectives

1. To explore how songs loudness and duration have changed over the years
2. *(Discarded – only for contrasting purposes)* Popularity: artists and songs. How do they differ?
3. To find out where the majority of artists covered in the dataset live
4. To explore technical music measures and their level of accuracy: beats start time, tatum start time, mode (scale), key and time signature.

# 5. Analysis

## Selection of visualizations

### 1. Scatter Plot

I selected this visual (Table 3) to get a sense of the changes on the duration and loudness of songs over time at a glimpse. Roughly half of the observations have unknown song year, and that is going to define the results in this plot and the others below. However, we can see that the triangle showing “Unknown” median value is located approximately in the middle of the rest of filter groups, which suggests that our data containing song year is a good representation of the overall observations.

I have selected shape marks because I am going to use colours very often in other plots, and this functionality works fine here, the comparison is visually easy to spot.

A close up of a map

Description automatically generated[[12]](#footnote-13)

While I was trying different options for scatter plots, I used the data on artists and songs popularity by artist familiarity to listeners. However, I discarded the use of this plot because it is way too crowded. It is not as useful as the previous one, because changes between groups of years cannot be easily spotted. However, I collected them into a Dashboard and kept them in my Tableau file for comparison purposes.

### 2. World map

I decided to plot a world map to uncover the geographical limitations of the dataset. I explained in the dataset selection section that this data has been collected by organizations from the U.S. I want to check if their goal was to analyse music representative of most of the regions on the earth, or only specific regions. 34% of the data does not have geographical coordinates but filtering and scrolling on the dataset for those I can still see that most artists are English speaking.

A close up of a map

Description automatically generated[[13]](#footnote-14)

### 3. Dashboard: bar plots & box-and-whisker plots

For technical features, I considered very important to check confidence level: how certain are we that the measure is correct. I used box plots for that purpose and bar charts, because of the level of detail they provide and its relevance. Since beats time starts and tatum time start are both time related, I compared them to spot differences over the years. I added another chart below so we can see for each bar very easily how sure we can be about the measure. Since time signature is also a time themed feature, I introduced it below those charts that I’ve just mentioned.

On the right side, I added 2 charts[[14]](#footnote-15) about the mode and the key (Table 5 and Table 6). They are first to catch eyes’ attention because they measure what I consider to be the most revealing technical music features. We can see a bar and associate it with a specific note, so it is not such an abstract term as time is. **This dashboard is adapted to colour blindness** to promote diversity and inclusion, hence the colour selection.

A screenshot of a cell phone

Description automatically generated[[15]](#footnote-16)

A picture containing pencil

Description automatically generated[[16]](#footnote-17)

A screenshot of a cell phone

Description automatically generated[[17]](#footnote-18)

A screenshot of a cell phone

Description automatically generated[[18]](#footnote-19)

A close up of a device

Description automatically generated[[19]](#footnote-20)

## Report of results / findings

### 1. Song loudness and duration over time

The first thing to notice is that the most recent the song release date is, the loudest its median value and median duration are. Note that loudness values are negative, so that should be taken into consideration.

### 2. *(Discarded – only for contrasting purposes)* Popularity: artists and songs. How do they differ?

As per the 2 scatter plots that I discarded, even though subtle variances are not easily perceived, overall, we can affirm that songs popularity score slightly higher in middle and high artist familiarity than artists popularity.

### 3. Where the majority of artists covered in the dataset live

That map is quite revealing. We can affirm that for most of the music analysed on the dataset, artist’s home country is the U.S.A. There is also a significant amount or artists whose home country is in Europe. Even taking into consideration that an artist location is not always related with his country of origin, or the kind of music it produces, we can affirm that the data collection is biased. It mostly analyses music produced in the Western countries. Perhaps a more ethical title for the dataset would have been “*Million Western Songs Dataset*”.

### 4. Technical music measures and their level of accuracy

The majority of songs are in C, D, G and A minor key, Eb and F# in minor mode being the least used. For those in major mode, Db, F#, G and Ab are abundant, whereas there is scarcity of C and F. Songs in minor mode have a greater level of accuracy when predicting the key than those in major mode. Important to note that most songs use a wide variety of notes, and very often also multiple keys, so this is just referring to the most prevalent key.

Both the tatum and the beats time start are longer the earlier the song’s release date is, generally speaking. However, the more recent the song’s release date is, the greater is the confident level of the measurement, hence the accuracy of the data. Lastly, we can claim that time signatures corresponding to number 3 and 5 have been very accurately predicted by the algorithm, but we can’t claim the same about number 1, whose level of confidence is too low. Unfortunately, no information has been found about the time signature numbers and their equivalence in real music language (4/4, 4/3, etc.), so no interpretation can be made on that regard.

# 6. Bibliography

1. Bertin-Mahieux, T.; Ellis, D. P.W.; Whitman, B. & Lamere, P. (2011): “*The Million Song Dataset*”. In Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011) [Online] Available at: https://corgis-edu.github.io/corgis/csv/music/ (Accessed 11/07/2020)
2. Million Song Dataset (2012): “*Frequently Asked Questions*”. [Online] Available at: <http://millionsongdataset.com/faq/> (Accessed 11/07/2020)
3. Dummies, A Willey Brand (2016): “*Common Music Time Signatures”* [Online] Available at: <https://www.dummies.com/art-center/music/piano/common-music-time-signatures/> (Accessed 11/07/2020)
4. MightyExpert (2020): “*What is a Bar in Music?”* [Online] Available at: <https://www.mightyexpert.com/what-is-a-bar-music/> (Accessed 11/07/2020)
5. Ableton (no date): “*Beat and tempo*” [Online] Available at: <https://learningmusic.ableton.com/make-beats/beat-and-tempo.html> (Accessed 11/07/2020)
6. Jehan, T. (2005): “*Creating Music by Listening*”, Massachusetts Institute of Technology [Online]. Available at: <http://web.media.mit.edu/~tristan/phd/dissertation/chapter3.html#x1-390003.4.3> (Accessed 11/07/2020)
7. Gil, D. (2017): “DIFFERENCES BETWEEN MAJOR AND MINOR SCALES (#2 IS THE KEY)” [Online] Available at: <https://shredaddict.com/differences-between-major-and-minor-scales/> (Accessed 11/07/2020)
8. Pouska, A. (2020): “Keys in Music” [Online]. Available at: <https://www.studybass.com/lessons/harmony/keys-in-music/> (Accessed 11/07/2020)
9. Cover image source: <https://towardsdatascience.com/step-by-step-to-visualize-music-genres-with-spotify-api-ce6c273fb827> (Accessed 11/07/2020)

1. Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere.

   The Million Song Dataset. In Proceedings of the 12th International Society

   for Music Information Retrieval Conference (ISMIR 2011), 2011. Available at:

   <https://corgis-edu.github.io/corgis/csv/music/> (Accessed 11/07/2020) [↑](#footnote-ref-2)
2. Columns eliminated: artist.location, artist.similar, release.name, song.artist\_mbtags and song.title [↑](#footnote-ref-3)
3. http://millionsongdataset.com/faq/ [↑](#footnote-ref-4)
4. Table 1 [↑](#footnote-ref-5)
5. Table 2 [↑](#footnote-ref-6)
6. <https://www.dummies.com/art-center/music/piano/common-music-time-signatures/> [↑](#footnote-ref-7)
7. <https://www.mightyexpert.com/what-is-a-bar-music/> [↑](#footnote-ref-8)
8. <https://learningmusic.ableton.com/make-beats/beat-and-tempo.html> [↑](#footnote-ref-9)
9. <http://web.media.mit.edu/~tristan/phd/dissertation/chapter3.html#x1-390003.4.3> [↑](#footnote-ref-10)
10. <https://shredaddict.com/differences-between-major-and-minor-scales/> [↑](#footnote-ref-11)
11. <https://www.studybass.com/lessons/harmony/keys-in-music/> [↑](#footnote-ref-12)
12. Table 3 [↑](#footnote-ref-13)
13. Table 4 [↑](#footnote-ref-14)
14. One of the charts was returning me error “6ea18a9e” after I saved the worksheet, so I added to the zip folder the image in .png format [↑](#footnote-ref-15)
15. Table 5 [↑](#footnote-ref-16)
16. Table 6 [↑](#footnote-ref-17)
17. Table 7 [↑](#footnote-ref-18)
18. Table 8 [↑](#footnote-ref-19)
19. Table 11 [↑](#footnote-ref-20)