Tempo and duration analysis

In these exercises, we will use a sample of the huge dataset called “million songs” (http://millionsongdataset.com/). This dataset has been used for a computational challenge about recommendation. It contains all information about the songs.

The files are stored in the hdf5 format. A special format for storing big databases.

The database is huge (300 Gigabytes) so we will not work on the whole database but just a sample of songs written by The Beatles. And we will study if the **tempo** and **duration** of their songs (1) have an impact on their popularity and (2) change over time.

Original files are in hdf5, you need to install the package rhdf5 from Bioconductor. Hence you load the library. The command to read files is h5read and you specify the group name of data to read. In our case 3 group names are storing information about music data: ‘analysis’, ‘musibrainz’, and ‘metadata’. Each group name contains different labels and descriptions. We will use the following labels to build a data.frame :
'tempo'
'duration'
'loudness'
'song_hotttnesss'
'year'

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager") BiocManager::install()
BiocManager::install( "rhdf5" )

library(rhdf5) G1
=
h5read("D:/Utilisateurs/turenne/Dropbox/UIC/LectureDataProcessing/NicolasM
usicAnalysis/ex.h5","analysis") G2 =
h5read("D:/Utilisateurs/turenne/Dropbox/UIC/LectureDataProcessing/NicolasM
usicAnalysis/ex.h5","musicbrainz") G3 =
h5read("D:/Utilisateurs/turenne/Dropbox/UIC/LectureDataProcessing/NicolasM
usicAnalysis/ex.h5","metadata")
```

We will use library ggplot2 and dplyr. Here are the weblink describing functions:
ggplot2: https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf dplyr:
https://cran.r-project.org/web/packages/dplyr/dplyr.pdf

1. **convert data** into a single data frame by selecting tempo, duration, loudness, song_hotttnesss, year

(here we will use *list.files* and *h5read functions ;* hence *do.call rbind lapply* to make a loop over each file)

2. **Clean the data frame** by selecting songs having no NaN , year different from 0, and year before 1970, because the Beatles stopped producing songs after 1970. And export the data.frame

3. **read the dataset** and display the first values

4. **Checking the Count**. Display the number of songs, all the years when the Beatles produced songs, and the histogram of production.

5. Duration analysis

Select songs from the year 1963 (using command filter from dplyr package). Display summary Select songs from the year 1970. Display summary What do you observe?

6. hotttnesss analysis by ranking

Make a ranking by hotness (using command arrange from dplyr). Display the first 10, and the last 10.

We want to see if hotness is like a normal distribution.
Display hotttnesss using command ggplot from package ggplot2 using the histogram.

What do you observe?

Display summary of hotness. What are the mean and the median?

This is a function to compute the mode :

```
#creating the mode function
getmode <- function(v)
{    uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}
```

What is the mode ? In summary are the mode, median, and mean the same? Is the distribution of hotness normal ?

Display hotness by a boxplot .

7. hotttnesss analysis by plotting duration and tempo

Display hotness in function of duration using the function plot.
What do you observe?

```
with(df_beatles, plot(duration,  hotttnesss, xlab="Duration (sec)", ylab="Song Hotness")) #we see that duration impact hotness
```

Now we make a 3-dimensional plotting by adding **tempo** dimension being a function of point size. (We will use command ggplot with geom_point)

What do you observe?

Now we make a 3-dimensional plotting by adding the **loudness** dimension being a function of point size. (We will use command ggplot with geom_point)

What do you observe?