2023

Georgiy Farina, Enkh-Oyu Nomin

SUPSI-DTI: 2nd year DS&AI Bachelor

16/05/2023



**Data Challenge: Spotify Dataset**

Sommario

[Introduction and problem setting 1](#_Toc135324851)

[Data structure and source 1](#_Toc135324852)

[Research questions 1](#_Toc135324853)

[Data Cleaning and exploration 2](#_Toc135324854)

[Exploratory data analysis 2](#_Toc135324855)

[Modelling 2](#_Toc135324856)

[Summary of results and conclusion 2](#_Toc135324857)

# Introduction and problem setting

## Data structure and source

The dataset involved in this project is the Spotify collection of the top 2000 (around 100 per year) songs and their audio characteristics, throughout the years 1999-2019. We went for this choice since we both are fans of music and users of Spotify, therefore, we were familiar with the domain. We also picked this dataset because it has many features to be used (18 total columns out of which 16 to be used for the models), a decent number of rows (roughly 2000) and two valid candidates for the project requested tasks: regression and classification.

## Research questions

Firstly, we wanted to understand which artists were the most consistent throughout the years, meaning which ones had most hit songs in the charts. Then we also wanted to know the average popularity of these latest artists, and by doing that we could see which artists out of the most present are also the most popular.

After having analysed a bit the singers, we wanted to analyse more the songs themselves. As first research questions we were curious about which are the most popular genres and what is the duration of the most popular songs. There were some expectations based on the last decade that we witnessed and also on the most popular songs we knew from 2000 and on, but still we wanted it either confirmed, or to be surprised.

Since it is songs we’re analysing, we wanted to see which ones were the top 10 most popular songs of the current millennium, and also how the songs changed over the time in terms of duration, noticing ourselves that, with respect to the past, they tend to be shorter.

Then, since Spotify has been released first in 2008, we figured that we also wanted to know if there was any correlation between the release year and some other features such as popularity, genre or others, to see how the charts and music evolved starting from the foundation year of Spotify, service that became huge in the last decade.

At last, we wanted to figure out the most impactful and prevalent correlations between the feature, for example if there is a correlation between duration and popularity, genre and popularity, duration and genre, and so on. That is because we wanted to see

“In this section plots could be used to showcase the data and anticipate the answers.”

# Data Cleaning and exploration

present how the data were loaded, cleaned and put in a format suitable for the rest of the study. This part will show pre-eminently in the report with the code chunks needed to clean the data. Some plots could be useful here to understand the structure of the data.

# Exploratory data analysis

present the data in details before modelling with plots (boxplots, scatter plots, etc.) and summary statistics (means, median, sd). This part serves as a first justification of why some models are chosen to answer the questions. For example if we are interesting in questions related to a binary variable, then we will use classification models. In this part we can show the summary statistics for the binary variable.

# Modelling

for each question asked in the introduction, present one (or more) methods to answer the question. The output of each model should be explained. When more than one method is compared, advantages and limitations of those methods related to the particular case should be highlighted through plots and tables. Moreover plots should be used to showcase the power of the best method chosen.

# Summary of results and conclusion

This section should summarize the main findings with reference to the questions asked in the introduction.



