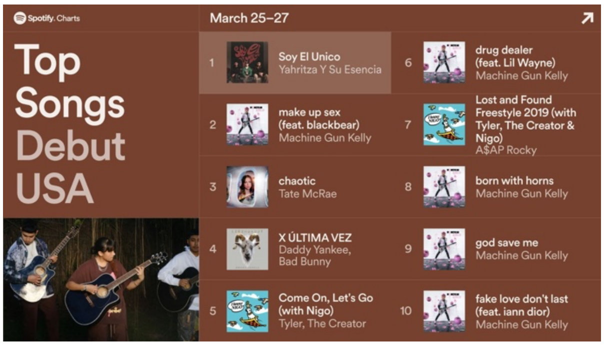
Part 2) Statistical Analysis & Machine Learning (10%)

In this exam, we aim to classify song into 11 music genres (the target variable is "Class"). Each record refers to each song along with 17 variables (including target).



Spotify. Charts

March 25–27

Top Songs Debut USA

1	Soy El Unico	Yahritza Y Su Esencia
2	make up sex (feat. blackbear)	Machine Gun Kelly
3	chaotic	Tate McRae
4	X ÚLTIMA VEZ	Daddy Yankee, Bad Bunny
5	Come On, Let's Go (with Nigo)	Tyler, The Creator
6	drug dealer (feat. Lil Wayne)	Machine Gun Kelly
7	Lost and Found Freestyle 2019 (with Tyler, The Creator & Nigo)	A$AP Rocky
8	born with horns	Machine Gun Kelly
9	god save me	Machine Gun Kelly
10	fake love don't last (feat. iann dior)	Machine Gun Kelly

You have to submit the following items to MyCourseVille:

- File.ipynb – a source code.
  - It cannot be graded if we cannot map your answer to the question, so please add a comment to identify the questions, e.g., Q1.1, Q1.2
  - All answers must be shown or written in this file (the notebook file).
  - For the question with "Write your answer", please write your answer in text or comment in your notebook.
- All files must be renamed as "{student_id}_{firstname}_Part2", e.g., 6030133421_Chaiyatad_Part2.ipynb

Tasks

Load data from the link below into data frame in Pandas:

https://github.com/kaopanboonyuen/2110446_DataScience_2021s2/raw/main/datasets/msff_rv22v1_1.csv

1. Pre-process the data [1 point]

   1.1. Remove 3 variables "Artist Name", "Tract Name", and "Popularity"

   1.2. Change the variable "duration_in min/ms" to "duration"

   1.3. The output data must be in the data frame "selected_df" and show info()

2. Provide exploratory data analysis (EDA) and statistical analysis [1 point]

   2.1. The variable "valence" refers to an emotion, where 0 and 1 refer to negative and positive emotions, respectively. Perform a suitable statistical analysis to answer whether different music genres have different music emotion (valence).

   - "Write your answer" What should be a suitable statistical test? Is it significantly different?

3. Data cleansing [3 points]
    - 3.1. Show the number of missing values for each variable
        - "Write your answer" How many missing records for the variables "Key" and "Instrumentalness"?
    - 3.2. For variable "Key", impute missing with 1
    - 3.3. For the variable "Instrumentalness", impute missing with 0 and if it is less than (<) 0.01, replace the value with 0.
    - 3.4. Create a dummy code for the variable "Key" with dropping the first level. Please ensure that the variable "Key" should not be in the data frame anymore.
    - 3.5. The output data must be in the data frame "cleaned_df" and show info()
    - 3.6. Show basic statistics of variables using describe()
4. Train/Test split [1 point]
    - 4.1. Split train/test **with stratification** by setting testing size to be 30% and random seed to be 1234
    - 4.2. Show the amount of whole data, train, and test.
5. Create Model1: Decision Tree [2 points]
    - 5.1. Create and train a decision tree model with the following settings (the variable name "tree"):
        - The splitting criteria is 'gini'
        - The maximum tree depth is 10.
        - The minimum number of examples at leaf node should be 10.
    - 5.2. Evaluate on test and show a classification report with 4 decimal points
        - "Write your answer" What is the accuracy?
    - 5.3. Sort (descending) and print the features based on feature_importances from the decision tree
        - "Write your answer" What is the rank of the variable "valence"?
6. Create Model2: Regression (You have to choose the right regression) [1 point]
    - 6.1. Select only 4 features including ['acousticness','duration','instrumentalness','speechiness']
    - 6.2. Create and train a regression model with the following settings (the variable name "reg")
        - random_state=1992
        - solver='newton-cg'
    - 6.3. Evaluate on test and show a classification report with 4 decimal points
        - "Write your answer" What is the accuracy?
7. Compare the models [1 point]
    - 7.1. Since there is a single test data set, using McNemar's test to answer as to whether or not both methods are significantly different.
        - "Write your answer" Is it significant different?