CSEN911: Data Mining
Winter 2025
Author: Hazem Nabil
Course Instructor: Dr. Ayman Alserafi

German University in Cairo

Clustering Mini-Project

Due: Friday 12th of December, 2025 11:59 PM

Introduction

With the advent of music streaming platforms, music is more accessible and easier to listen to than ever before. At the same time, finding new music (sub)genres to explore can be intimidating due to the huge variety available. One possible solution to this is to recommend similar music genres to the ones one already likes. Clustering could allow us to group similar music sub-genres together for that purpose, as well as find out what makes each subgenre–or each group of music genres–similar or different to another.

Description

You are required to apply the **K-Means clustering algorithm** to group similar music genres together based on simplified music features. In addition, you are required to answer the following queries using appropriate markdown text by using an appropriate visualization method for each question to find the answer, while justifying and explaining each method using appropriate markdown text:

1. What factors affect the Popularity of the song? (Mention the two biggest factors, with **interpretation**)
2. Which genre is the most popular and which is the least?
3. What are the most correlated pairs of attributes? (Mention 2 pairs, with **interpretation**)
4. Plot the frequency of words in genres. Which main genre categories have the most sub-genres? (mention at least 3)
5. Create an additional visualization that **differs** from earlier ones. Your visualization must highlight a new data relationship or finding. Provide a short explanation of the insight and its relevance.

As part of the model-building & interpretation phases, you are required to:

1. Select and justify the appropriate parameter values for the clustering algorithm, while clearly explaining the method used to find these values.
2. Display a random sample of **at least 5 genres** from each cluster group.
3. Visualize and interpret/describe each cluster with respect to the features present in the data.
4. Evaluate the clustering quality using the appropriate method mentioned in the lecture, comment on its score and give a final recommendation.

Info: You will be given a notebook with the dataset description and link. No starter cells will be provided this time. You will be **assessed** on the **organization, clarity and structure** of your notebook. Use markdown cells to split your code into discrete steps, and group logically related code into code cells.

Evaluation Criteria

Warning: Using AI to solve the project will result in a **ZERO**. You can use external sources to search for methods or syntax; However, all the explanations and comments must be your own.

The maximum grade for this submission is 50 and your Jupyter notebook submission will be graded on the following:

- Justifying and commenting on every step taken or decision made (submitting the code only will result in a **ZERO**).
- Visualizing and answering each question correctly in detail [20]
- Applying appropriate and necessary data pre-processing methods with justification. [5]
- Model Building, interpretation and Evaluation [20]
- Notebook organization and clean structure [5]

Notice: If you make changes to the dataset, you are required to display the effect of your changes using an appropriate method. You are also required to comment on your code, clearly explaining and justifying each step you are doing.

Submission

To accept the assignment and gain access to the private repository for your team, register using the following GitHub classroom link `https://classroom.github.com/a/wDKG8-rU` as a team of two. **Make sure you are submitting your work in the correct repository; incorrectly placed submissions will NOT be graded.**

Warning: Your Jupyter notebook submission will be re-run after submission. Submitting a notebook that errors mid-way through execution will result in losing the grade of all subsequent cells! The order of the notebook's code cells and the order in which you execute them matters. Please make sure your notebook runs and outputs the expected results by restarting the kernel and running all the cells before finalizing your submission.

Bonus (2.5 Points)

Use your clustering model in a function that takes in a music genre as an input and recommends a few similar genres to explore.

```
>> # Function Call Example
>> genre_recommender('jazz')
['hard bop', 'electric bass', 'jazz clarinet', 'cool jazz', 'bebop']
```

Note: Code related to this bonus task should be performed at the end of your mini-project notebook, not in between the mini-project's requirements. Indicate the start of your bonus code, either with a comment at the start of the code cell, or add a new section/heading titled 'Bonus' for clarity (similar to the existing sections provided).