Milestone 1 Project Proposal

CS 4412 - Data Mining

1st Cannon Blocker
Department of Computer Science
Kennesaw State University
Kennesaw, United States
cblocke8@students.kennesaw.edu

***Abstract*—This document is a project proposal for the CS 4412 semester long data mining project. This proposal includes the intended dataset that I will be using, the discovery questions I intended to find a conclusion for, the planned techniques I intend to use for analysis of the dataset, and a rough plan for the additional project milestones.**

I. Dataset Description

The dataset I intend to use is "Osu Beatmap stats from 2007 to 2019" which can be found at [1] The dataset has 83,641 rows, 37 columns and a file size of 32 MB. The dataset is a collection of all beatmaps for the standard game mode that have the status of ranked/loved/qualified up to August 1st of 2019. Out of the 37 columns I intend to use primarily attributes that characterize the beatmap. The most important of which are Title, Genre, BPM, Playcount, Passcount, etc.

title	bpm	genre_id	language_	favourite_	rating	playcount
DISCO PRI	119.999	2	3	515	8.11412	369455
1,2,3,4, 00	172	2	2	117	7.72277	87273
1,2,3,4, 00	172	2	2	117	7.72277	117868
1,2,3,4, 00	172	2	2	117	7.72277	91073
1,2,3,4, 00	172	2	2	117	7.72277	118536
Love Fight	125	2	6	47	7.67679	94878
Scatman	136.016	5	2	533	8.70632	1741352
Pop Star	139.867	5	3	62	7.63661	185093
Re:Re:	153.001	4	3	263	8.55465	238059

Fig. 1. A sample of the data set

II. Discovery Questions

A. What map and song characteristics are most commonly shared among the most popular beatmaps?

1) Discovery focus: Identify the patterns in play count and Likes by exploring how popularity relates to factors like Genre, BPM, Length, Star Rating, and object composition (Circle/Slider Count):

2) Why it's valuable: This question helps uncover what drives player attention and replay behavior in rhythm games. Popularity isn't random — it may reflect patterns in: song genre preferences, BPM ranges that players enjoy most, difficulty levels that attract the largest audience, chart styles that feel satisfying or iconic.

3) Why it's interesting: Instead of assuming "harder = better" or "faster = more popular," this discovery approach lets the data reveal: what types of beatmaps and genres dominate the top charts, whether popularity is linked more to music features or game play design, and what common structures appear in widely played beatmaps. This provides insight into the design formula behind successful rhythm game content.

B. Are there distinct groups of songs that achieve popularity in different ways (high replay vs high appreciation)?

1) Discovery focus: Explore whether beatmaps naturally cluster into engagement types, such as:

- high play count but low likes (frequently played)
- low play count but high likes (underrated favorites)
- high in both (iconic beatmaps)

2) Why it's valuable: Play count and likes measure different kinds of engagement:

- Play count = what people repeatedly play
- Likes = what people emotionally or qualitatively appreciate

3) Why it's interesting: This question can reveal hidden categories like:

- "grind maps" (played a lot but not loved)
- "community favorites" (loved even if less played)
- "iconic masterpieces" (high on both)
- "overlooked gems" (high likes, low exposure)

This could be a powerful discovery because it can show that engagement is multi-dimensional, and different beatmaps succeed for different reasons.

III. Planned Techniques

The two techniques that I primarily want to use are clustering and association rules. Clustering appears to be a good way to group beatmaps into natural categories based on features like: BPM, Genre, and Rating. This should be primarily used for answering the first discovery question. Association rules should be useful for discovering "successful combinations". For example,

- If Genre = J-Pop and BPM ¿ 180 → often high likes
- If Star Rating between 5–6 and AR high → often high play count

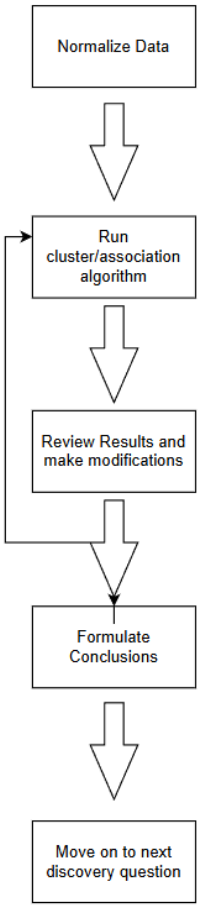

Fig. 2. Hypothetical analysis pipeline

IV. Preliminary Timeline

I anticipate there will be some challenges in actually understanding a lot of the data. There is more than 80,000 individual beatmap rows with dozens of attributes that are all loosely connected. I also believe there will be some challenge with classification of certain maps / outliers due to unpredictable factors (EX: Map does not fall in to expected classification group because it's a notable "meme" map and has community popularity despite being short in other metrics).

A. Milestone 2

I want to have at least an operational notebook with a normalized dataset and a functional pipeline for putting the dataset through the aforementioned algorithms.

B. Milestone 3

I want to have a fully operational notebook with a completely normalized dataset and a fully functional pipeline for putting the dataset through the aforementioned algorithms and generating visuals to assist in my conclusions.

C. Milestone 4

By milestone 4 I want to have already completed all functional requirements and focus solely on the interpretation and evaluation of the discovered knowledge.

References

[1] https://github.com/jxu/osu-stats/releases/tag/2019-08-01