CAP5771 Spring 2025 Project

Haylee Zuba

Milestone 1

Objective

The purpose of this project is to build a tool that demonstrates knowledge of data science principles. For my project, I have chosen to combine media that interests me by comparing metadata from songs and video games to explore trends, patterns, and correlations between video game choice and music taste. There are many datasets available on Kaggle that offer data that will assist in the creation of this project. In order to achieve this, I will be using a million songs dataset, an emotional analysis of music dataset, and a video game trend dataset. By comparing the data from these datasets, I can construct a training and testing set of data to train a machine learning model to identify patterns and output results.

Type of Tool

For my Spring project, I have chosen to create a project that is exemplary of option 3: a recommendation engine. One of the most intriguing aspects of data science, to myself as a student, is training a machine learning model to recognize patterns and watch it learn to make its own decisions.

To create this model, I will need to utilize different python libraries. For the data analysis portion, pandas and numpy will be crucial to analyze the data. Once we ascertain the statistical summaries of the data and account for outliers and null values, we can use libraries such as matplotlib and seaborn for visualization of the data and store it using SQLite After the data preprocessing and exploratory data analysis are performed, we can move on to constructing the machine learning model using Pytorch and Scikit libraries. These are the intended libraries for now, but there may crop up other libraries that will assist with functionality or function better for the intended use rather than the ones listed.

The total tech stack for this project is as follows:

- Backend: Python in Pycharm IDE, utilizing libraries such as Pandas, Numpy, Seaborn, Matplotlib, Pytorch, Scikit, and SQLite.
- UI: Javascript and Streamlit to create a functional and aesthetically pleasing UI.

Data

The Million Song Dataset is a collection of metadata and audio features from a million songs. Some of the attributes that each song in this dataset has are:
“Song ID: Unique identifier for each song (String)(contains alphanumeric characters)
Title: Name of the song (String)
Artist ID: Unique identifier for the artist (String)
Artist Name: Name of the performing artist (String)
Release Year: Year the song was released (Integer)
Tempo: Speed of the track in beats per minute (BPM) (float)
Loudness: Overall perceived loudness of the track (float)
Key: Musical key of the song (Integer)
Time Signature: Rhythmic structure of the track (Integer)
Duration: Length of the track in seconds (Integer)” (According to the authors of this dataset: Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere.)

This is a very large dataset, 300gb, so for the EDA, I used a subset. This will still offer information about trends within the data, but is feasible to run on my laptop. During the implementation of the data modeling and when creating the predictive algorithm, I will use methods such as chunk processing to analyze the full set.

The DEAM Dataset - Emotional Analysis in Music

The DEAM dataset consists of 1802 excerpts and full songs annotated with valence values over the whole song and per each second measurements. For this project, we don't need to examine every second of every song. This dataset contains key columns:
“Song ID: Unique identifier for each track (integer)
Valence Mean: Average emotional positivity score (float)
Arousal Mean: Average energy level of the song (float)
Tempo: Speed of the track (BPM) (float)
Spectral Features: Various frequency-based audio characteristics(list[float])” (According to the files posted on Kaggle).

For the EDA, I will not be utilizing the second-by-second annotations, as the mean valence and arousal values will be more important to the overall trends, and will be more efficient to perform the analysis on.

Discovering Hidden Trends in Global Video Games

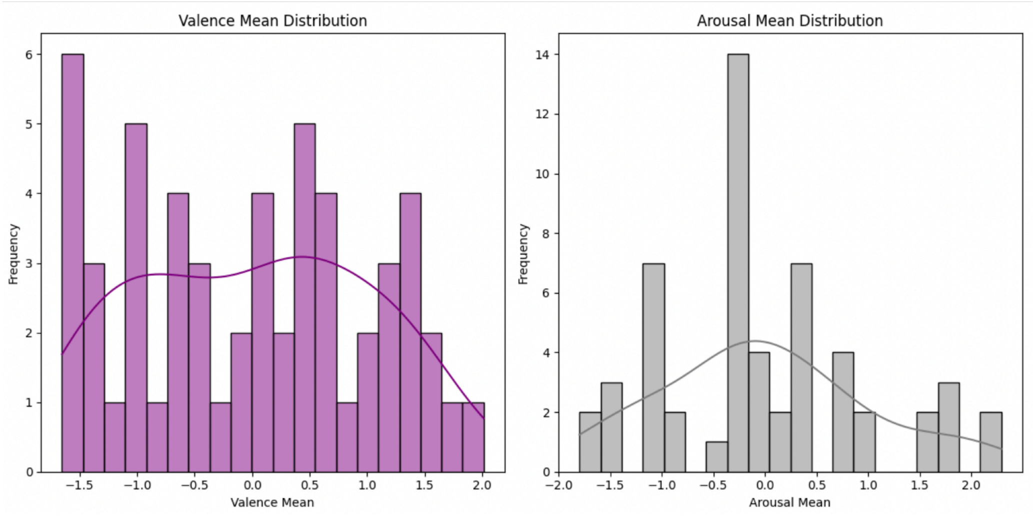
This dataset holds data for nearly 2,000 video games and their sales across different platforms and countries. There are 13 columns within the data set that represent (and their respective data types):

"Rank: The ranking of the game in terms of global sales. (Integer)
Game Title: The title of the game. (String)
Platform: The platform the game was released on. (String)
Year: The year the game was released. (Integer)
Genre: The genre of the game. (String)
Publisher: The publisher of the game. (String)
North America: The sales of the game in North America. (Integer)
Europe: The sales of the game in Europe. (Integer)
Japan: The sales of the game in Japan. (Integer)
Rest of World: The sales of the game in the rest of the world. (Integer)
Global: The total global sales of the game. (Integer)
Review: The review score of the game. (Float)" (Author of the dataset, unnamed on Kaggle)

While this dataset mainly explores the fiscal success of global video games, the dataset also holds metadata crucial to identifying any possible correlations between song and video game metadata.

Exploratory Data Analysis

The DEAM dataset contains 55 entries and 13 columns. After running EDA on this dataset, we can conclude that the means are very close to 0 for valence_mean and arousal_mean, and a slight left skew - meaning more positive valence and arousal values for some songs within the dataset. The visualization of the distributions for valence mean and arousal mean and the fact that they are close to 0 show that a majority of the songs fall within a similar neutral range.



A correlation plot, visualized on the Kaggle website, shows a moderately strong correlation between valence mean and arousal mean. This means that songs that are high valence, are also high arousal, and vice versa.

The video game sales dataset shows a large spread. Some games have extremely high sales, and others have a few. This is a sign of a very thorough dataset that contains statistics from many different games from many different countries. The sales data.csv shows a high standard deviation, indicating few popular games, but mostly mediocre/decently performing games are in the dataset. The visualizations for this dataset show that consoles like Playstation and Xbox are some of the most popular companies for games. The genre plot shows that action and shooter games have the highest sales, and adventure games tend to have lower sales. The scatter plot shows that global sales and user score have a moderate correlation. Games with higher user counts tend to have higher global sales.

The million song dataset subset contains 10,000 songs that shows an even distribution between older and newer songs, with a large standard deviation showing that the dataset includes songs from many different decades. Visualizations of this dataset show that most of the songs are from the 2000s, and that there is a correlation between different characteristics of the songs such as tempo, loudness, and danceability.

Project Timeline

For this next Milestone, we have until March 21st, so 5 weeks from now, to complete the features and data modeling for the project. Listed below is an estimated timeline of the tasks I intend to complete by that date - at the latest. The estimated time is a very rough estimate, everything will probably require different times than anticipated.

Task	Task Description	Date to be completed by	Time Estimated to complete
Feature Engineering	Create new features from existing ones to improve model performance, encode categorical variables	3/10	~8 hours
Feature Selection	Evaluate feature importance, and reduce dimensionality	3/10	~8 hours
Create the training and testing datasets	Create datasets to train and test the predictive algorithm	3/10	~2 hours
Data Modeling	Train multiple machine learning models, and optimize hyperparameters	3/17	~10 hours
Evaluate model performance	Evaluate model performance using appropriate metrics (e.g., accuracy)	3/20	~6 hours
UI initial development	For this project, I intend to create a creative UI to	3/21	~5 hours

	showcase the findings of this project.		
Milestone2.pdf	Write the report for Milestone 2	3/20	~3 hours

Once this has been completed, I will then move on to the tasklist for milestone 3, which is tentatively as follows:

Task	Task Description	Date to be completed by	Time Estimated to complete
Model Evaluation/Improvement	Evaluate model performance on the test set, find any weak spots/limitations	4/01	~6 hours
Model Interpretation	Interpret the outputs and explain interpretations	4/05	~4 hours
Bias identification and analysis	Identify any biases the model might show, and figure out why that may be.	4/08	~4 hours
UI Finalization	Finish creating an aesthetically pleasing (or at the very least, functional) UI to display the project	4/20	~6 hours
Presentation and Final Report	Prepare a 4 minute presentation, 4 minute demo video, and write the final report detailing the workflow and findings of the project.	4/22	~6 hours

Sources/Links:

Preprocessed data:

Million Song Dataset: Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The Million Song Dataset. In Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011), 2011. Available: http://millionsongdataset.com/ (Originally found on Kaggle)

Discovering Hidden Trends in Global Video games, Andy Bramwell, Available: https://www.kaggle.com/datasets/thedevastator/discovering-hidden-trends-in-global-video-games

DEAM Dataset, available: https://www.kaggle.com/datasets/imsparsh/deam-mediaeval-dataset-emotional-analysis-in-music

Other data that may be considered later in the process, but was not used for the EDA:

https://www.kaggle.com/datasets/arnabchaki/popular-video-games-1980-2023
https://www.kaggle.com/datasets/gregorut/videogamesales
https://www.kaggle.com/datasets/uciml/msd-audio-features