CS210 Project

Ayça Pektekin

January 2024

1 Introduction

In this project, I used my Spotify data in order to explore more about my listening behavior. I have used different visualization strategies in order to explore about the data. These include (but not limited to) artists who I listened the most, the month I have listened the most etc. We will discuss more these in detail in the upcoming chapters.

2 Data and Its Source

The data used in this project is provided by Spotify. Spotify provides various data for an account. For this project, two of them are used: YourLibrary, StreamingHistory

2.1 YourLibrary

This dataset includes all tracks a user liked since the creation of the user's account. The data format was JSON and looked like the following:

```
[{
  "artist": "Ezhel",
  "album": "Müptezhel",
  "track": "İyi Bil",
  "uri": "spotify:track:6Sg9kVAVm4Xn6MVh2mYNMd"
},
{
  "artist": "Soner Avcu",
  "album": "Yeniden",
  "track": "Yeniden",
  "uri": "spotify:track:4lfe318AeoHaKbxEfPl9SV"
}]
```

However, this data did not include much about the songs' features. For this purpose, I have aggregated the library data using the Spotify API. The API provides audio features of the tracks using the track uri. Because of API limitations, it was not possible to get all tracks' audio features; yet, I was able to get most of them. Having this process done, the data had the following extra features.

```
danceability, energy, key, loudness, mode, speechiness,
instrumentalness, liveness, valence, tempo, duration_ms
```

2.2 StreamingHistory

This dataset includes all the streaming history for the year 2023. Data format was again JSON and looked like the following. However, it was not possible to aggregate this dataset using the Spotify API since it included too many entries (24000).

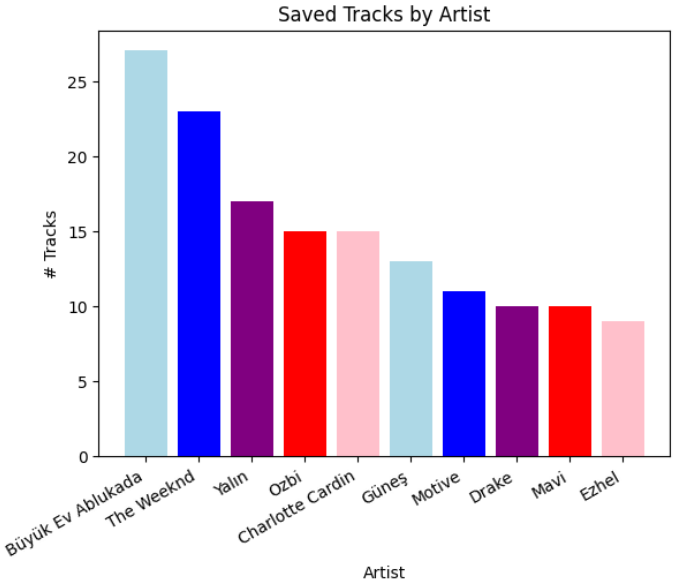
```
[{
  "endTime" : "2023-06-18 13:03",
  "artistName" : "Büyük Ev Ablukada",
  "trackName" : "ne deve ne kush",
  "msPlayed" : 16021
},
{
  "endTime" : "2023-06-18 13:03",
  "artistName" : "Büyük Ev Ablukada",
  "trackName" : "HOŞÇAKAL KADAR",
  "msPlayed" : 14272
}],
```

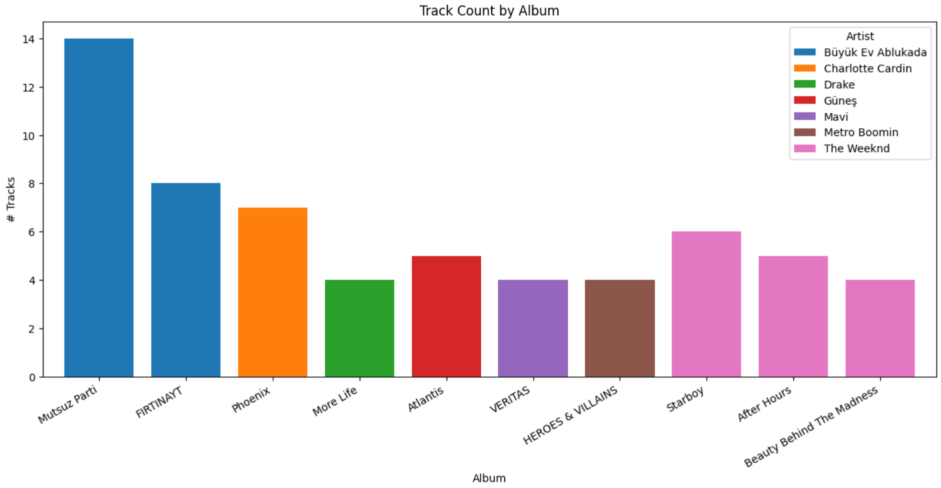
3 Data Analysis and Findings

In this section, I have analysed the mentioned datasets using various visualization and EDA techniques.

3.1 YourLibrary

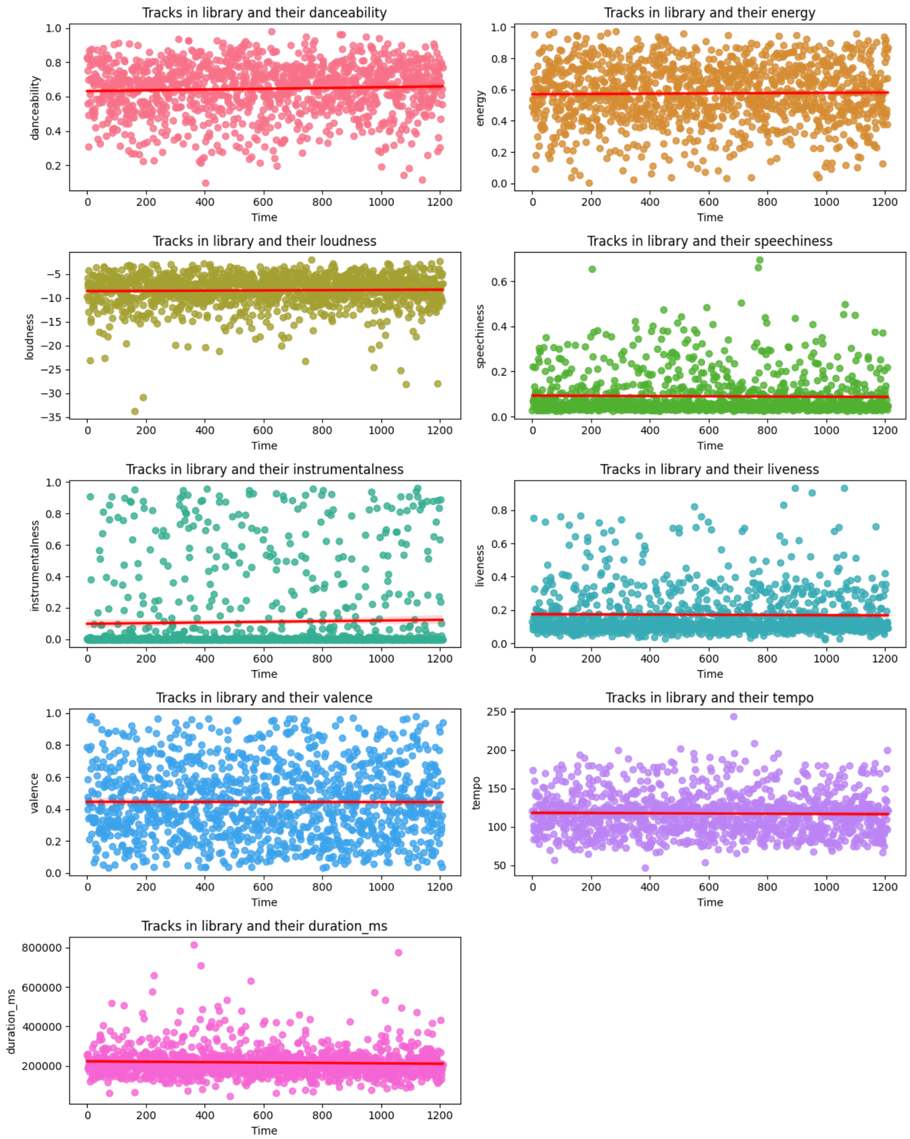
The most basic analysis idea was to see the artists and albums which have the most tracks that I liked. For this purpose, the following bar charts are plotted.



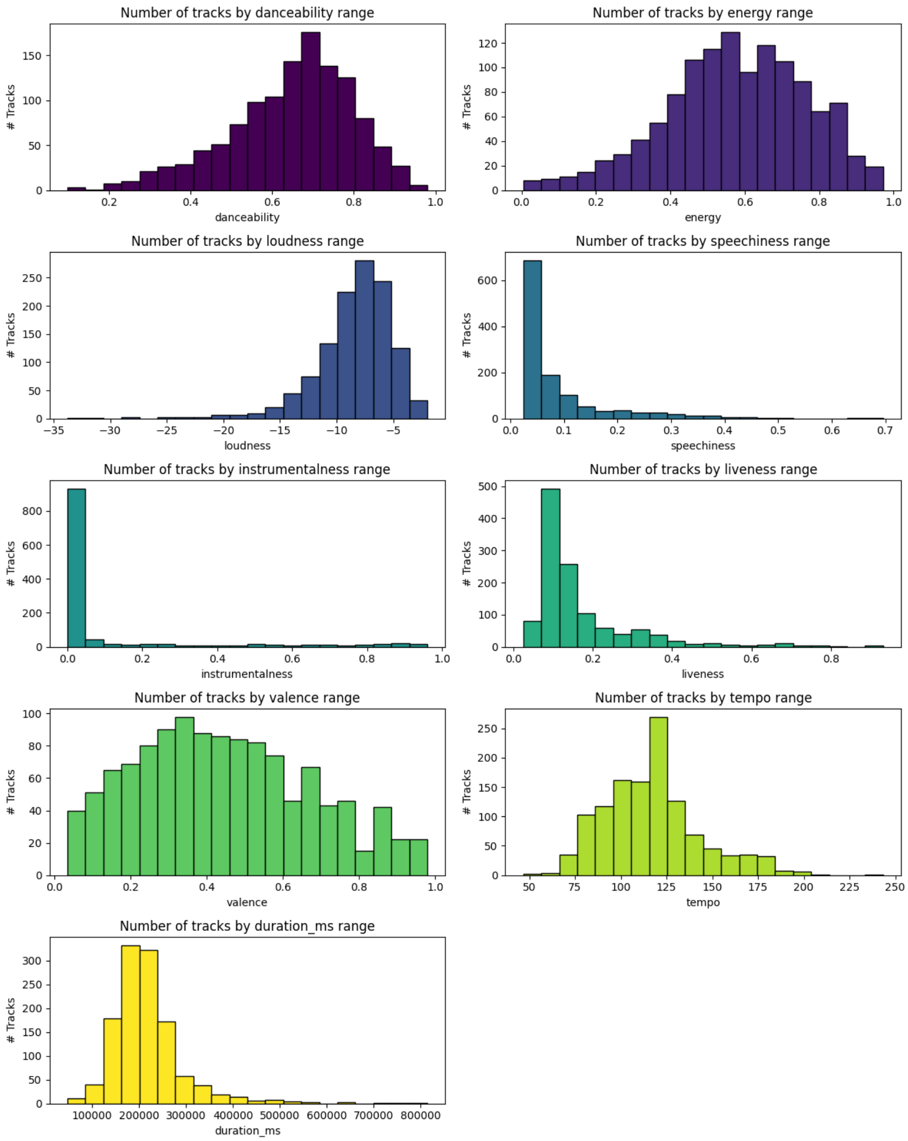


As it can be seen in the figures, the artist that I have the most liked tracks is ”Büyük ev Ablukada” and the their album ”Mutsuz Parti”.

Secondly, I wanted to explore whether there is a correlation between the audio features and my liking behavior change over time. This led me to plot audio features of my liked tracks with respect to the time. The time axis is from 2021 to 2024, and each dot corresponds to a track.



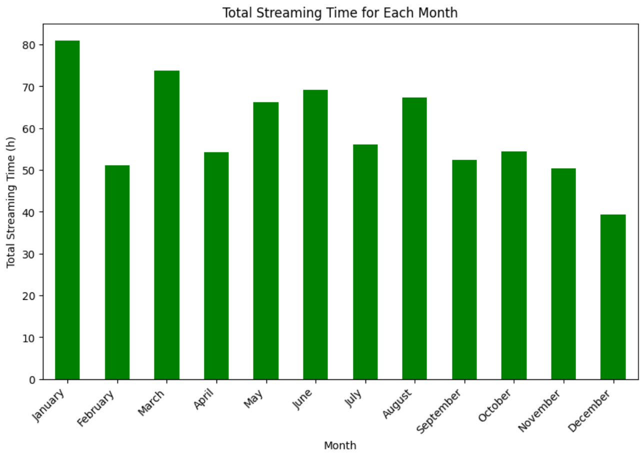
As it can be seen, audio features of my liked tracks does not change significantly over time since regression lines' slope are near to zero. Therefore I have decided to plot number of tracks I liked with respect to audio features as a histogram.



These plots clearly indicate that most of the tracks I liked share common audio features. For instance, I tend to like tracks whose loudness is in range (-10, -5)

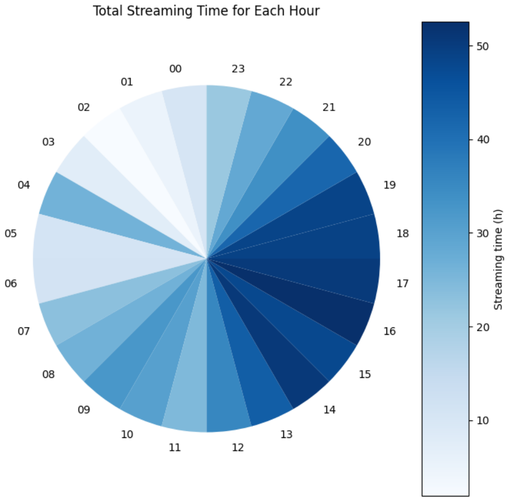
3.2 StreamingHistory

As we mentioned earlier, this dataset included my streaming history in 2023. First of all, I grouped the data entries with respect to the month and calculated total streaming time for each month. Results appear to be the following.



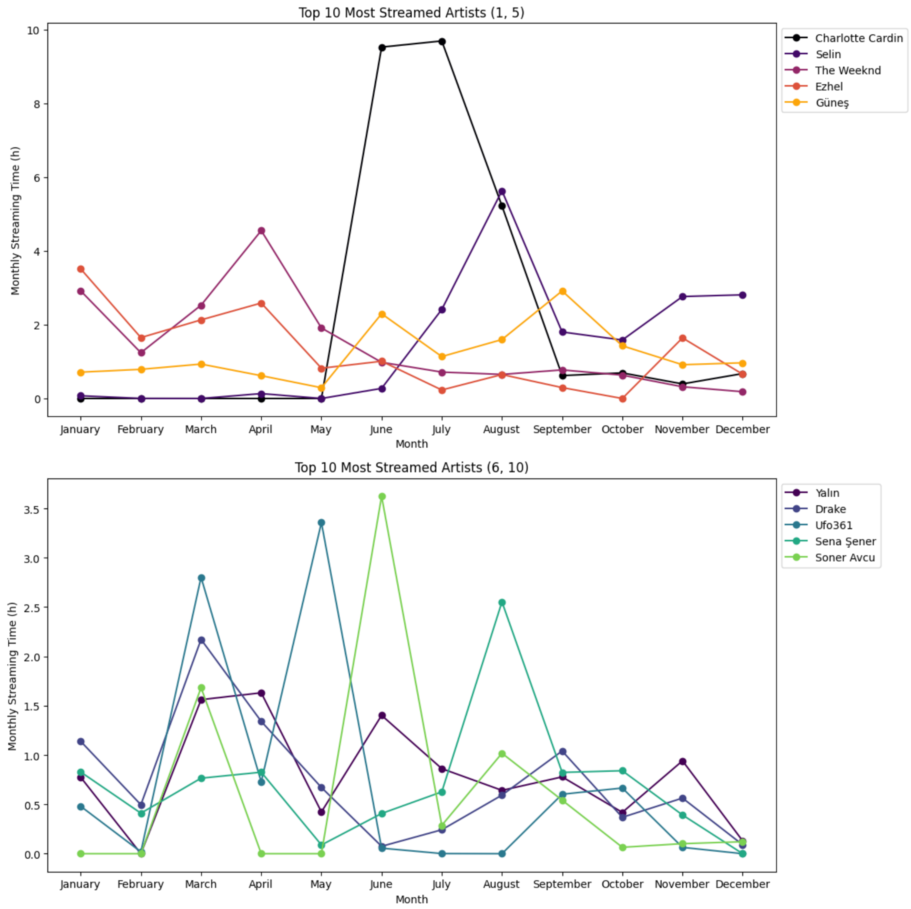
This figure displays that January is the month that I have listened music the most (≈ 80 hours), followed by March (≈ 75 hours) and June (≈ 70 hours).

Secondly I grouped the data by the hours of a day so that I can see my listening distrubition within a day.



As it can be seen, I have listened to music the most between 14.00 and 20.00. Moreover, the leading time interval is 16.00 - 17.00.

Finally, I wanted to see my most streamed artists and their change over months. This led me to plot a line chart displaying the monthly stream of my top 10 artists (top 10 is calculated with respect to the total streaming time of me in 2023).



As you can see, Charlotte Cardin is the artist that I listened the most in 2023, whom I listened the most in June and July. It can also be seen that I started listening to Charlotte Cardin in June. On the other hand, I have listened to Güneş less, who is the fifth; however, it is distributed accros the year. This implies that I have listened to Güneş in every month of the year.

4 Predicting Song Likes

In this chapter, I have tried to train a model that will predict whether I will add a song to my library or not, using the audio features of the song.

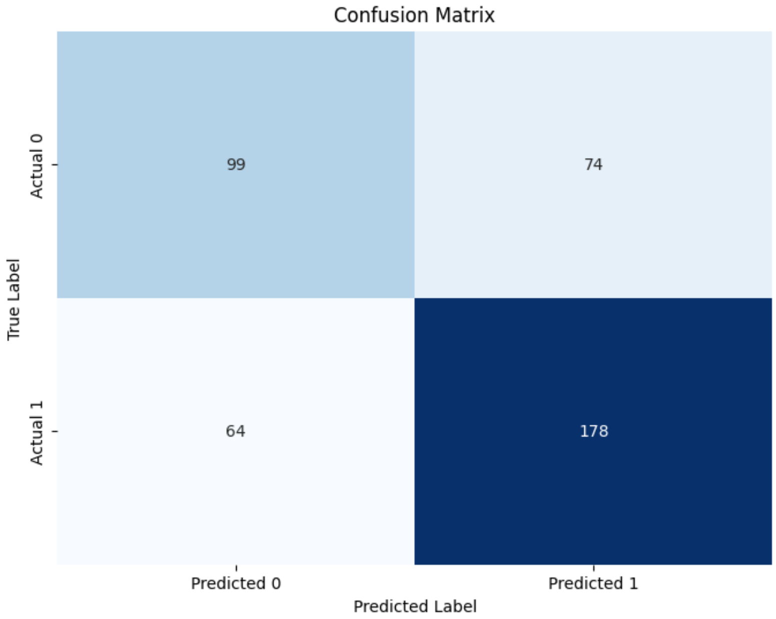
4.1 Dataset Preperation

I used the tracks in YourLibrary dataset as the positive class since they are in my library. However, I also needed a negative class representing songs whom I will not add to my library. To accomodate this, I have downloaded this dataset from Kaggle. This dataset includes the mostly streamed songs (953 entries) in 2023. I removed the songs that also exist in my library, from this dataset and marked these songs as the negaative class. Finally, I selected the audio features to include and merged two datasets to create the final dataset that looked like the following.

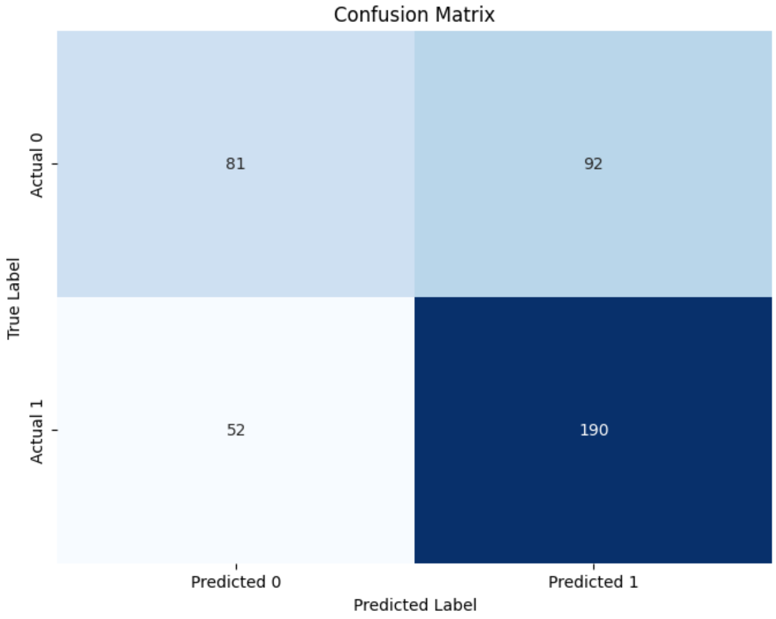
```
track_name,danceability,energy,speechiness,instrumentalness,liveness,valence,is_liked
Inner Light,65,70,3,0,7,12,1
Entre Nosotros (Remix) [con Nicki Nicole],70,44,4,0,37,61,0
```

4.2 Training the Model

Before training the model, I have split the dataset into train and test. Then I have trained a Random Forest with the target variable: *is_liked*. It had a test accuracy score of 66.75%, and the followig confusion matrix.



This model was not successful, which led me to tune the hyper-parameters of the random forest to create a tuned model. Bayesian search is used for this purpose. After tuning the model, the accuracy declined more: 65.30%.



One possible reason that the model was not successful is the fact that the negative class is generated from the most streamed songs in 2023 but not the songs I actually do not like. I have not listened most of the songs in that list, so there may exist songs that I would like. Another reason may be there are no correlation between the audio features and my song likings.

5 Limitations and Future Work

As we discussed earlier, the negative class for the machine learning may have been prepared better. A true dataset would be the one prepared by me with songs that I actually do not like. Moreover, the model may also need more features like the genre of the song. My dataset did not include genre information because Spotify do not provide genre information through their API.

Future work over this project may be solving these limitations by preparing a better dataset and generating more features through other external services. For instance, one may analyze the ”sentiment” of the song using the song lyrics and add it as a feature. If these fixes lead to a more successful model, a personal song recommendation system may be created.