Report on Spotify Consumption Analysis for DSA201 Course

Ahmet Can Toksoy
30885

Motivation

The motivation of this project is to understand my habits and gain insights into my behaviour, preferences and trends over time. The project is a product of my curiosity to explore my Spotify listening patterns to identify how my musical preferences and engagement evolved over yeras as music became a pivotal part of my life. Additionally this project is an excellent opportunity to enrich my portfolio. The scope of this project is downgraded to only

The project initially aimed to analyze my general social media consumption, however the scope of this project decreased because of time constraints and as I couldn't gather useful data.

Data Source

The data for this project comes directly from Spotify's streaming history, which I requested and downloaded from Spotify's data download future. Spotify provides users with their detailed listening history in JSON format. Spotify API was used less frequently as useful endpoints deprecated in November 2024.

I combined the given data into a dataset spanning from 2016 to 2024. The dataset includes key attributes like timestamp, track titles, track artists, listening durations, the platform, etc.

Data Analysis

The techniques used in different stages of the analysis are as the following:

- **Data Preprocessing**

The JSON files were loaded into Python using pandas. After that, through preprocessing missing/ irrelevant data was handled and data is processed for further analysis. Some specific details are listed below:

- The "ts" column is converted to a datetime object for easier manipulation. This is then used to extract meaningful time based features such as date, time, day_of_week and hour_of_day.
- Missing values in critical columns like master_metadata_track_name and master_metadata_album_artist_name were removed to safeguard the data integrity.

- The platform column was standardized. Extracted only the main part of the platform name (e.g., "osx" to "os"), converted them to lowercase for consistency. Finally unified categories by mapping variations like "osx" to "os".
- Categorical encoding is done to prepare the data for analysis. Platform and day_of_week features were labeled using the “Label Encoder” to convert the textual categories into numeric values.
- In some part of the analysis the already existing data was processed to create sessions. Sessions were defined based on a 30-minute gap between consecutive listening timestamps. A unique session ID was assigned to each group of consecutive listening events. Sessions were categorized into, short- less than 30 minutes, medium- 30 to 90 minutes, long**-** more than 90 minutes.

- **Exploratory Data Analysis**

Aggregate statistics were computed, including total listening time, most played tracks/artists, and listening trends over time. Yearly breakdowns were analyzed to detect patterns and changes in listening behavior.

Found the overview of my listening behaviour by identifying top tracks and artists by counting occurrences. Top 10 artists and tracks were calculated and displayed for a comprehensive view of listening preferences. Furthermore the data was grouped by date to calculate the number of tracks played each day. Summary statistics provided. Finally calculated the cumulative listening time in hours by summing the ms_played column and converting it into hours.

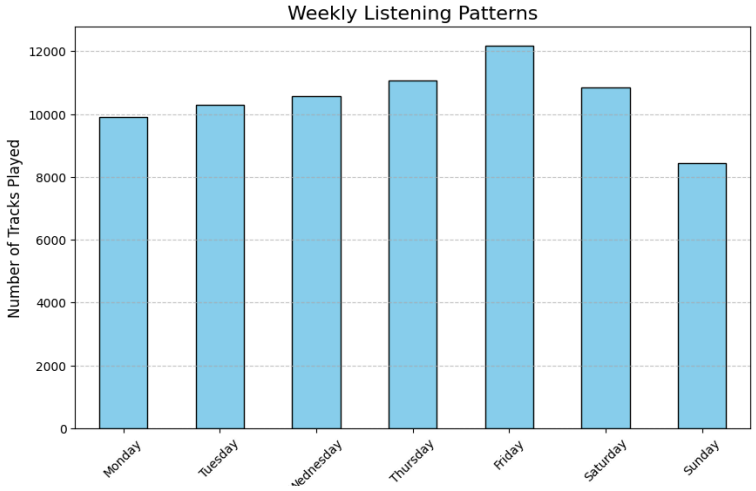
After that, I continued with analysis of the temporal patterns. Weakly, hourly listening patterns were drawn. Grouped data by the day of the week to examine the distribution of listening behavior. Extracted the hour of the day from timestamps and grouped data to understand the distribution of listening behavior across hours. Then these were visualized to present the reader. Finally the complete dataset was divided into yearly subsets and then top artists, top tracks, and daily listening statistics for each year was calculated.

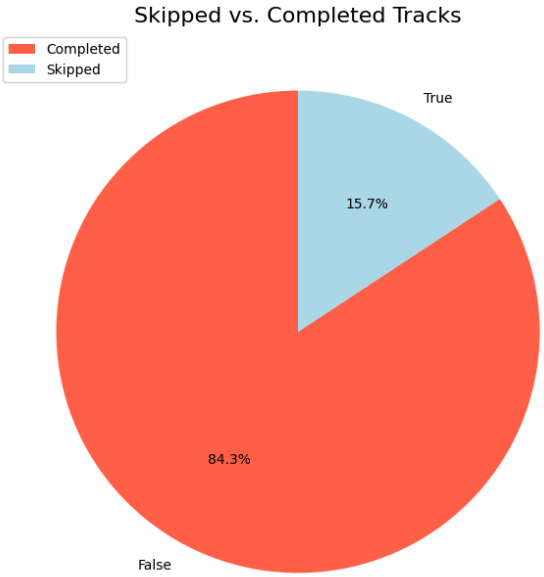
To analyze the behavioural insights I analyzed the distribution of skipped versus completed tracks and presented the results in a pie chart to illustrate the proportion of skipped and completed tracks. For each year, identify the day with the maximum listening time in minutes.

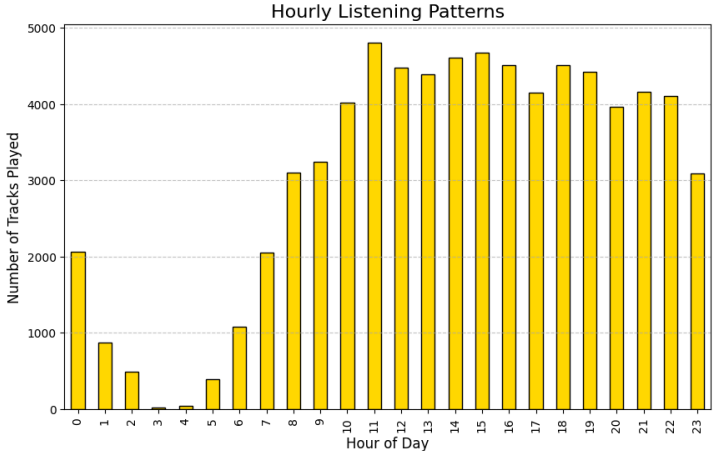
Also I conducted EDA to provide detailed understanding of session durations and platform preferences.

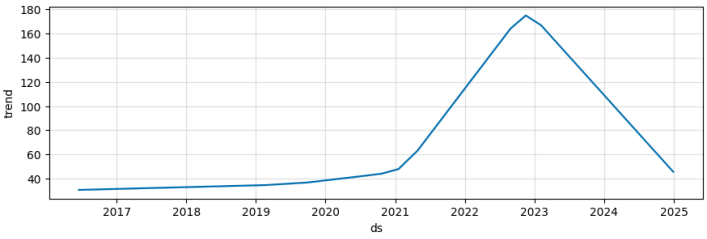
- **Data Visualization**

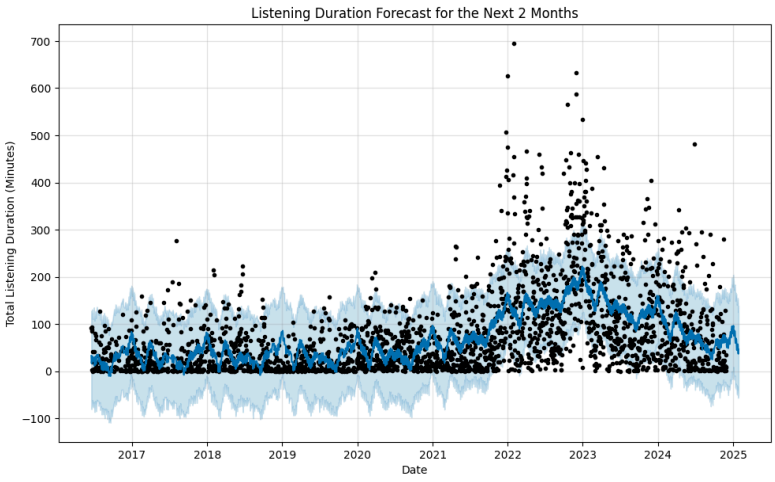
The analyzed data is visualized through line plots, bar charts, scatter plots, stack plots etc. Some of the graphs are listed below:

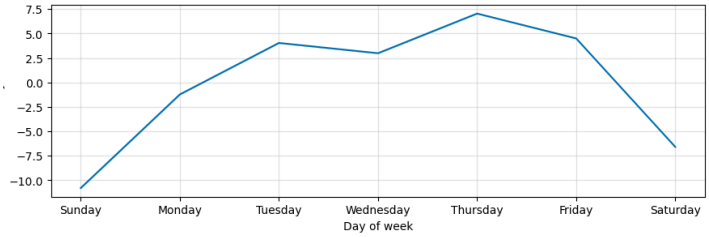

Weekly Listening Patterns
Number of Tracks Played
0
2000
4000
6000
8000
10000
12000
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday

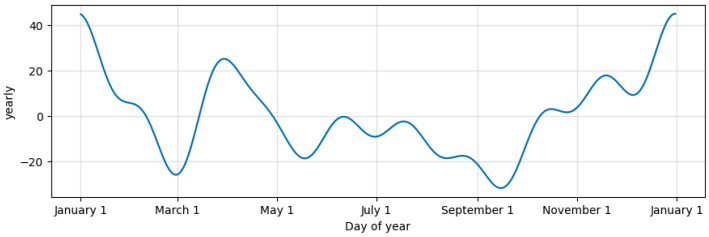

Skipped vs. Completed Tracks
Completed
Skipped
True
15.7%
84.3%
False


Hourly Listening Patterns
Number of Tracks Played
0
1000
2000
3000
4000
5000
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Hour of Day


trend
40
60
80
100
120
140
160
180
2017
2018
2019
2020
2021
2022
2023
2024
2025
ds


Listening Duration Forecast for the Next 2 Months
Total Listening Duration (Minutes)
−100
0
100
200
300
400
500
600
700
2017
2018
2019
2020
2021
2022
2023
2024
2025
Date


−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Day of week


yearly
−20
0
20
40
January 1
March 1
May 1
July 1
September 1
November 1
January 1
Day of year

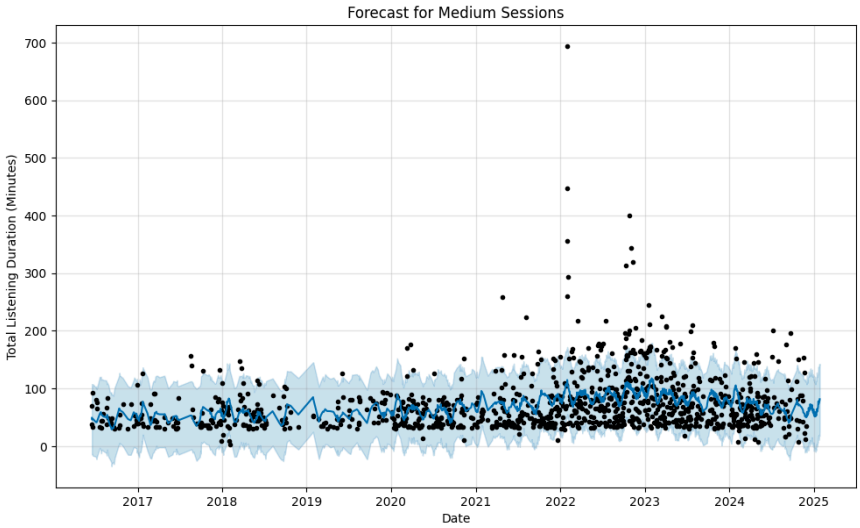
- **Hypothesis Testing**

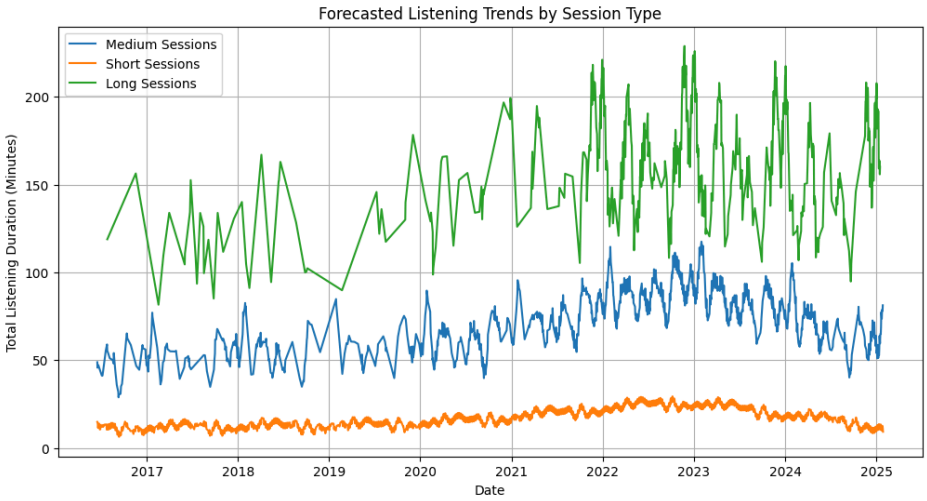
Tested several simple hypothesis, particularly:

- Yearly Listening Time Trends: Conducted a one-way ANOVA test to determine if there were significant differences in total listening time across years.
- Shuffle vs. Non-Shuffle Modes: Performed a normality test to decide between parametric or non-parametric methods (e.g., t-test or Mann-Whitney U test) to compare listening behaviors.
- Repeat vs. Unique Listening: Evaluated whether repeated listening behavior significantly differed from unique listening using statistical methods.

- **Time Series Analysis and Forecasting**

Used “Prophet” library to model and forecast the total listening duration for the next 60 days. Also did a session based forecasting where I modeled forecasts separately for “Short", " Medium", and “Long” sessions to understand session-specific growth or decline. Furthermore ı highlighted how different session types contribute to daily totals using stack plots.





- **Classification Techniques**

Built a platform prediction model using an XGBoost Classifier. The classifier tries to predict for the platform based on input features such as listening hour, day of the week, listening duration and time block. The class imbalances are addressed using “SMOTE”.

Achieved a really low accuracy of %59. Still open to improvements

Findings

- I spent a total of 3173 hours listening to music.
- On average, I played around 29 tracks per day, with peak listening days exceeding 500 tracks. My listening habits show variability, with notable spikes on significant days like 2021-12-30 (626.55 minutes) and 2022-01-30 (694.95 minutes).
- My most frequently played artists include; Hayko Cepkin(4088 times), Nightwish(2183 times), and Rammstein(1975 times) dominate the list, reflecting my preference for rock, metal, and alternative genres.
- Besides the metal genre, over the years some artists like Eminem and Howard Shore dominated the list.
- My top tracks include Sweet Child O' Mine, Bertaraf Et, and Dans Et
- Friday is my most active listening day, with 12,178 tracks played over the years. Sundays show reduced listening activity,
- My listening peaks occur in the evenings, especially between 18.00 to 22.00.
- I mostly listen on iOS devices, followed by Windows, suggesting that most of my music consumption happens on personal devices like phones and laptops. Occasionally, I use partner platforms like smart TVs.
- In 2016-2018, I explored diverse genres, including metal, pop, EDM, and hip-hop, with artists like Twenty One Pilots, Imagine Dragons, and 2Pac. Starting in 2019, my focus completely shifted to heavier rock and metal genres, with Rammstein and Sabaton dominating my playlists. Recent years (2022-2024) saw an increase in Turkish rock artists like Hayko Cepkin, Şebnem Ferah, and Pentagram.
- Most of my listening sessions fall into the short to medium categories (less than 90 minutes), though longer sessions are not uncommon.

Limitations and Future Work

I would like to increase the scope of the project, and analyze different social media. However, for that I need to increase my usage of these platforms.

Spotify previously allowed users to get audio features of a particular track like danceability, etc. This is no longer supported so it constraints the analysis. For improvement I am thinking of obtaining these features from an external source such as a dataset and analyzing the content of the song I listen to. An analysis for more comprehensive data like genres and mood labels.

Machine learning model was unsuccessful and had low accuracy and other metrics. One improvement could be getting more suitable data and training a better model. Due to the time constraints and knowledge the implementation in the project is the best I could.