Spotify User Behavior Analysis

Panashe Dione

CPT_S 475

12/5/25

1. Abstract

This project analyzes a dataset of 520 Spotify users to explore how listening behavior varies across genres, discovery methods, moods and time of day. The dataset was found on Kaggle, and was cleaned by standardizing categorical labels, organizing multiple responses, and filling missing entries to ensure consistency. Using Python and visualization techniques, including bar charts, a pie chart and heatmaps, this analysis identifies key patterns like the dominance of Melody and Classical music as preferred genres, the importance of recommendation and playlists in music discovery, and the influence of mood and time on music/content choices. The results highlight the behavioral trends that provide insight into how users interact with music and podcasts on the platform

2. Introduction

Music platforms like Spotify and apple music play a major role in how people consume content, whether its music or podcasts. As someone who uses Spotify regularly, I have always been curious about how my listening habits compare to other users, and what factors influence people's choices when interacting with the platform. This interest made Spotify an appealing topic for exploring user behavior through data.

The goal of this project is to analyze a survey dataset containing responses from 520 Spotify users to better understand how listening habits vary across genres, moods, discovery methods and time of day. The dataset, sourced from Kaggle, provides a wide range of self-reported information about user preferences and motivations, making it perfect for exploratory data analysis Because the data reflects user experience, and there's always something you can learn from them, it offers an insight into how people think about their own music behavior and taste.

To investigate these patterns, the project uses Python for data cleaning and visualization. Several important steps were taken to make sure the dataset is prepared for analysis, starting with standardizing categorial labels, organizing multiple responses, and filling in missing values to ensure consistency. After cleaning, Different visualizations charts such as bar charts, a pie chart, and heatmaps were created to highlight key relationships and trends within the data.

This report presents the findings from those visualizations and talks about user behavior in terms of genre preferences, discovery habits, mood-related choices, and time-based listening patterns. By looking deep into these trends, the project aims to provide a clear understanding of how individuals engage with Spotify and what factors most strongly influence their listening decisions

3. Problem Definition

The purpose of this project is to examine how Spotify users engage with the platform and which factors influence their listening behavior. Specifically, the analysis focuses on identifying trends and relationships within self-reported survey data. The project focuses on the following questions:

1. Which music genres are most preferred among Spotify users?
2. How do users typically discover new music on Spotify?
3. Does mood influence whether a user chooses music or podcasts?
4. How do genre preferences vary across different times of day

By answering these questions, the project aims to highlight meaningful patterns in user behavior and provide insight into the motivations behind listening choices.

4. Models and Measures

This project utilizes exploratory data analysis methods to examine user behavior withing the Spotify dataset. The dataset consists of categorical and user response, so the analysis focuses on descriptive techniques rather than predictive modeling. The goal was to identify patterns, and relationships across key variables such as genre preferences, discovery methods, mood influences, and listening times,

4.1 Data Cleaning Methods

Before analysis, there were many steps taken to ensure consistency and interpretability”

- **Standardizing categories labels**: Multiple variations of similar responses (different spellings or combined categories) were merged into consistent labels.
- **Organizing multiple responses**: Some entries containing multiple genres or moods were pared and grouped into broader or meaningful categories.
- **Handling missing values**: blank or incomplete responses were replaced with “unknown” to make sure the dataset maintained its completeness and avoid disruptions in visualization.

These measures ensured the dataset was cleaned enough to perform exploratory analysis.

4.2 Visualization Choices

The following visualization methods were selected based on the type and purpose of each variable:

- **Bar chart**: used to show the distribution of user’s favorite music genres. This helps identify dominant genre preferences within the sample

- **Pie chart**: Used to display how users discover new music Spotify. This visually highlights the proportion of recommendations, playlist, radio, and other methods
- **Heatmap (Mood X Content Preference)**: Used to explore how mood influences whether users choose music or podcast. This measure captures relationships between the two variables.
- **Heatmap (Time of Day X Genre)**: used to look at how listening patterns change across morning, afternoon, and night. This visualization allows for easy comparison of genre popularity across the day.

4.3 Reasoning for Visuals

These methods were chosen because they provide a clear way to interpret the categorical survey data. Since the dataset does not include numerical measures, descriptive charts and heatmaps offer the most effective approach for identifying trends and relationships. These methods directly support the project's goal of uncovering behavioral patterns among Spotify users.

5. Implementation and Analysis

The project was done in Jupyter Notebook using python. Several library were included for analysis, libraries like pandas for data manipulation, seaborn and matplotlib for visualization, and NumPy for basic operations. To sum it all up, first the dataset was loaded, then prepared for analysis through cleaning and organization, and last generating visualizations to answer the main questions

5.1 Loading the Dataset

The dataset was downloaded into Kaggle and imported into Jupyter using pandas (the python library) next was the inspection, viewing the first few rows, checking column names, and identifying general patterns or inconsistencies. This step provided an overview of the dataset's structure to help ease into the cleaning process

5.2 Data Cleaning and Preparation

Because the dataset was based on survey responses, many of the fields were categorical and contained variations in formatting. To create consistency and reliability, the following were done:

- Standardizing category names: Some responses included slight variations or combined categories (for example "Classical & melody, dance"). These were standardized into consistent labels and improved interpretability

- Handling multi-select responses: Certain fields, such as genre or mood, contained multiple values separated by commas. These were grouped into broader categories where appropriate to improve visualization clarity
- Filling missing values: Any blank entries were replaced with the label "Unknown" to avoid issues during plotting and maintain dataset completeness.

These steps ensured the dataset was properly structured for exploratory analysis

5.3 Visualization Process

Four visualizations were created to address the project's main questions

1. Bar Chart for Favorite Genres:
   This chart was made by counting the frequency of each genre and plotting the distribution. It provided a clear view of the preferred genres among the respondents
2. Pie Chart for Music Discovery Methods:
   This visualization shows the proportion of users who rely on recommendations, playlists, radio, or other methods to discover new music. It highlighted the dominance of algorithm-driven and playlist-based discovery
3. Heatmap for Mood x preferred content:
   A pivot table was created to summarize how often each mood category corresponded to choosing music or podcasts. A heatmap was then applied to visualize the relationship between emotion and content type
4. Another pivot table was made to examine genre preferences at different listening times (morning, afternoon, night)

6. Results and Discussion

6.1 Favorite Music Genres

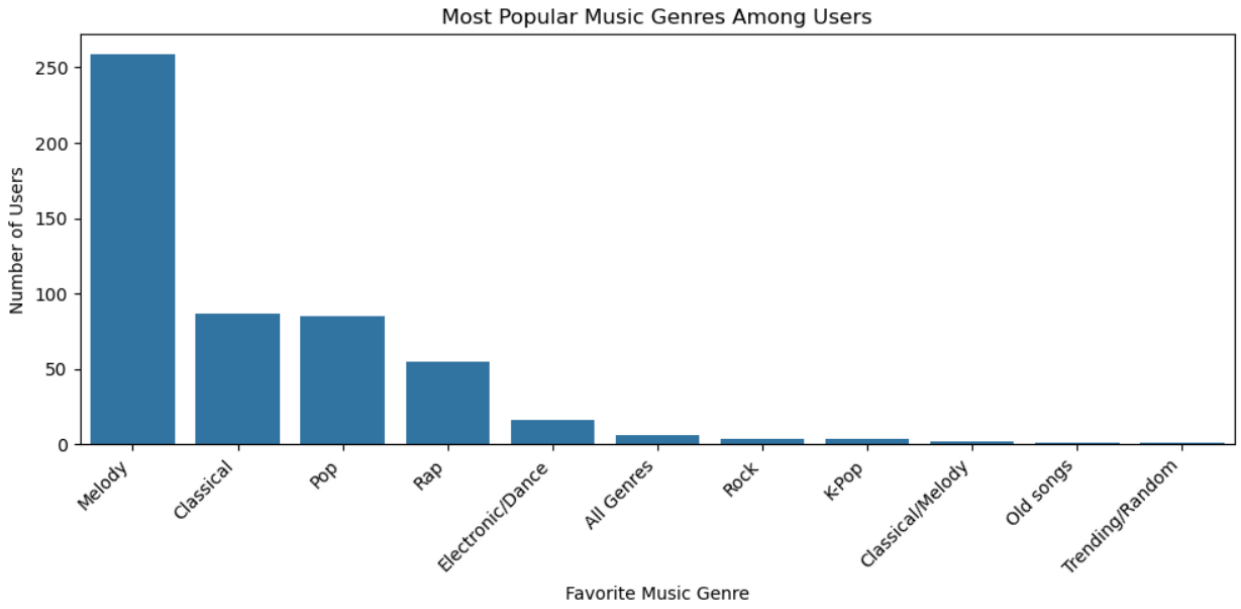

Figure 1: Bar Chart of Favorite Genres

The bar chart shows clear differences in genre preference among respondents. “Melody” appears as the most frequently chosen category, followed by Classical, Pop, and Rap. The strong presence of Melody and Classical suggests that many users prefer softer or more relaxing styles of music. Pop and Rap remain popular but appear less frequently compared to those top categories. This distribution provides a general overview of the musical landscape within the dataset and highlights the diversity of user tastes.

6.2 Music Discovery Methods

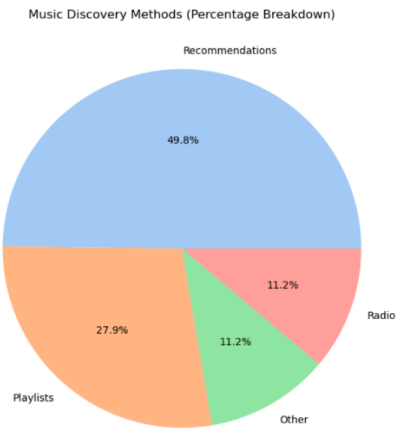

Figure 2: pie Chart of Music Discovery Methods

The pie chart reveals that most users discover new music through recommendations and playlists, showing the importance of Spotify's algorithmic systems and playlists. Radio and "Other" methods make up a much smaller portion of user discovery. This indicates that users rely heavily on Spotify to guide them toward new content rather than searching manually or using external sources. The finding supports the broader trend of personalized recommendation engines shaping listening behavior

6.3 Mood X Preferred Content

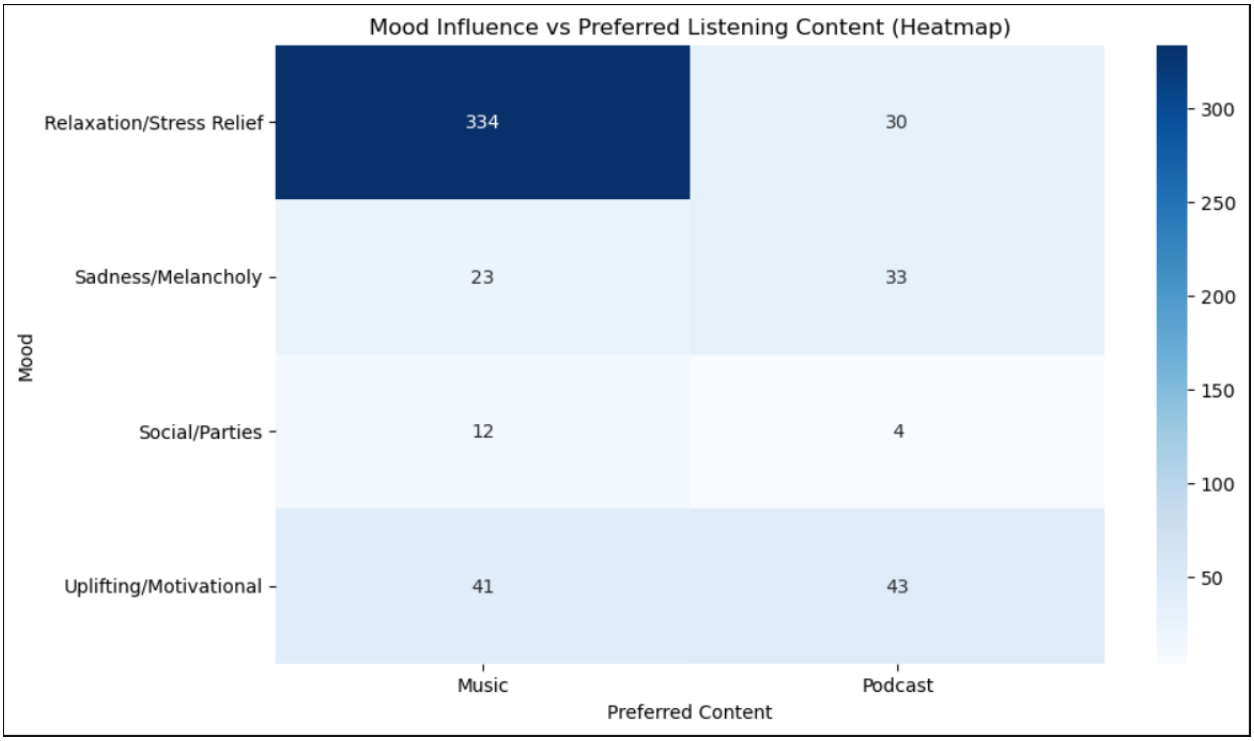

Figure 3: Heatmap of Mood X Content Preferences

This heatmap illustrates how emotional state influences whether users choose music or podcasts. Relaxation and stress-relief moods overwhelmingly correspond to choosing music. Sadness and motivational moods show a slightly higher interest in podcasts, although the overall difference remains small. Social or party moods also primarily lead to music selections, but this category includes fewer responses, making the trend less strong than the others.

Overall, the visualization suggests that while podcasts are used across moods, music remains the dominant choice in nearly all emotional contexts. Mood does influence content selection, but the shifts are relatively subtle rather than dramatic.

6.4 Time of Day X Genre

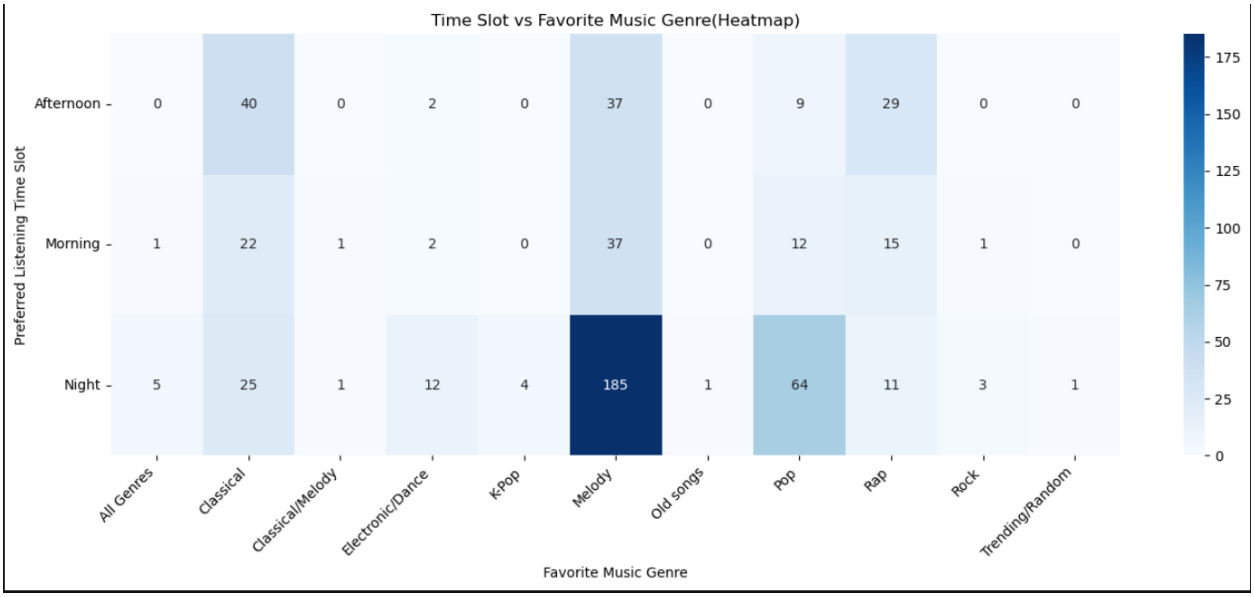

Figure 4: Heatmap of Time Slot x Genre

The final heatmap shows how listening patterns vary throughout the day. Nighttime is the most active listening period, showing higher activity across all genres. Melody is particularly popular at night. Afternoon and morning listening activity is lower overall, and although genre distributions during these times are similar, they show fewer total listeners.

These results suggest that listening habits align with typical daily routines: users listen most frequently at night, likely during downtime or relaxation periods, while daytime listening is more sporadic.

6.5 Summary of findings

- Melody and Classical are the most popular genres among survey respondents
- Recommendations and playlists are the most dominant methods for discovering new music.
- Mood influences content choice, with slight increases in podcast use during sad or motivational moods.
- Nighttime is the peak listening period, with consistently higher engagement across genres

Together, these results provide meaningful insight into how Spotify users choose and interact with audio content

7. Conclusion

This project explored listening behavior among Spotify users by analyzing survey responses from 520 participants. Through data cleaning, exploratory analysis, and visualization, several meaningful patterns emerged. Melody and Classical were the most popular genres, indicating a preference for softer or more relaxing styles of music among many users. Discovery habits showed a strong reliance on Spotify's recommendation system and curated playlists, highlighting the platform's influence on shaping user experience.

Mood played a role in content selection, with relaxation-related moods strongly associated with music, and sad or motivational moods showing a slight increase in podcast usage. Additionally, time-of-day patterns revealed that nighttime is the most active listening period across all genres, suggesting that users engage with audio content most frequently during evening or winding-down hours.

Overall, the project demonstrates how survey-based data can provide meaningful insight into user preferences and behaviors on digital platforms. While the dataset is limited to self-reported responses, the analysis still offers a useful perspective on the factors that influence listening choices. Future work could incorporate larger or more detailed datasets, such as real streaming history, to further enhance the understanding of user behavior on Spotify.

8. Appendix A – Code Repository

The complete Python code and dataset used for data cleaning, analysis, and visualization in this project is available on GitHub:

GitHub Repository:
https://github.com/Panashedione211/Spotify-User-Behavior-Analysis.git

The repository includes:

- Jupyter Notebook containing all analysis steps
- Generated visualizations
- Supporting documentation