DATA SCIENCE MINOR PROJECT  
  
PROJECT REPORT  
(Project Semester January-April 2025)

**EDA on Spotify History**

Submitted by

**Ayush Sahu**

Registration Number **12321402**

BTech CSE K23GR

Course Code INT-375

Under the Guidance of

**(Gargi Sharma (UID: 29439))**

Discipline of CSE/IT  
Lovely School of Computer Science and Engineering  
Lovely Professional University, Phagwara

**CERTIFICATE**  
This is to certify that **Ayush Sahu** bearing Registration no. **12321402** has completed **INT 375** project titled, **“EDA on Spotify History”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his original development, effort and study.  
  
  
**Signature and Name of the Supervisor**  
**Designation of the Supervisor**  
School of Computer Science and Engineering  
Lovely Professional University  
Phagwara, Punjab.  
  
**Date:**

**DECLARATION**  
I**, Ayush Sahu**, student of **K23GR (BTech CSE)** under CSE/IT Discipline at Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.  
  
Date:  
  
  
Signature  
  
Registration No. **12321402**  
  
Name of the student: **Ayush Sahu**

**ACKNOWLEDGEMENT**

I would like to express my heartfelt gratitude to my faculty coordinator, Ms. Gargi Sharma, for her continuous guidance, valuable feedback, and encouragement throughout the course of this project. Her insights were instrumental in completing this analysis successfully.  
  
I also thank Lovely Professional University for providing the necessary infrastructure and environment to work on this project.

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| S.no. | Content | Page no. |
| 1. | DECLARATION | 3 |
| 2. | ACKNOWLEDGEMENT | 4 |
| 3. | Introduction | 6 |
| 4. | Source of Dataset. | 6 |
| 5. | EDA Process | 6 |
| 6. | Analysis on Dataset  5.1. Introduction  5.2. General Description  5.3. Specific Requirements, functions and formulas  5.4. Analysis results  5.5. Visualization | 7-9 |
| 7. | Conclusion | 10 |
| 8. | Future Scope | 10 |

**1. Introduction**

The project titled “EDA on Spotify History” focuses on understanding user listening behavior through Exploratory Data Analysis (EDA). By analyzing personal streaming history from Spotify, the project uncovers insights such as listening frequency, preferred time slots, top artists, and more. These insights help to better understand music consumption patterns over time.

**2. Source of Dataset**

[**https://mavenanalytics.io/data-playground**](https://mavenanalytics.io/data-playground)

**3. EDA Process**

The Exploratory Data Analysis involved several steps:  
- Loading and viewing the dataset  
- Checking for null values and missing data  
- Changing the datatype of endTime to timestamps and creating new time-based features (day, hour, weekday)  
- Renaming and converting duration units  
- Dropping unnecessary columns  
- Filling missing values with appropriate statistical methods  
  
This preprocessing ensured the dataset was clean and ready for meaningful analysis and visualization.

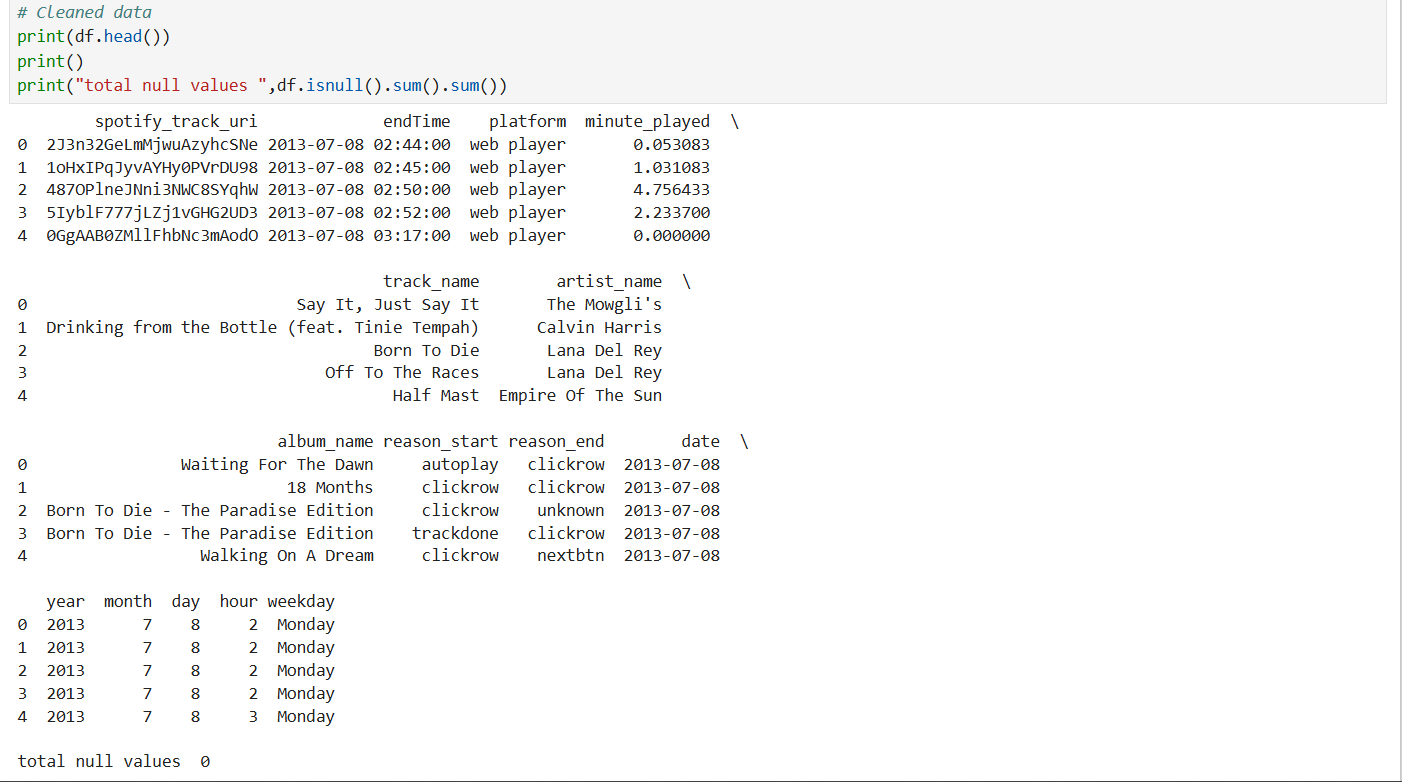
A screenshot of a computer

AI-generated content may be incorrect.

**A screenshot of a computer screen

AI-generated content may be incorrect.**

**Clean data**

****

**4. Analysis on Dataset**

**4.1 Daily Stream Count**

i. Introduction  
This analysis identifies how many tracks the user streamed each day over the period.

ii. General Description  
Streaming data was grouped by 'date' to compute the total number of songs listened to per day.

iii. Specific Requirements, Functions and Formulas  
Used Pandas groupby: **df.groupby('date').size().reset\_index(name='listen\_count')**

iv. Analysis Results  
The result showed some clear peak and low streaming days.

v. Visualization  
A line plot was generated using Matplotlib to visualize streaming volume over time.

**4.2 Peak Listening Hours**

i. Introduction  
This analysis determines the hours of the day when the user streams most.

ii. General Description  
The 'hour' extracted from the timestamp was grouped and counted to identify listening trends.

iii. Specific Requirements, Functions and Formulas  
Used **df.groupby('hour').size()** and visualized using Seaborn barplot.

iv. Analysis Results  
The results revealed specific hours where listening was highest.

v. Visualization  
Displayed using a bar plot for each hour from 0 to 23.

**4.3 Top Tracks by Play Count**

i. Introduction  
This analysis lists the most frequently played tracks by the user.

ii. General Description  
The 'trackName' column was counted and sorted to identify popular songs.

iii. Specific Requirements, Functions and Formulas  
Used **df['trackName'].value\_counts().head(10)** for top 10 tracks.

iv. Analysis Results  
Showed user's favorite songs over time.

v. Visualization  
Visualized as a horizontal bar plot of top 10 tracks.

**4.4 Top Artists Listened**

i. Introduction  
Finds the most listened-to artists in the dataset.

ii. General Description  
Grouped and counted entries in 'artistName'.

iii. Specific Requirements, Functions and Formulas  
Used **df['artistName'].value\_counts().head(10).**

iv. Analysis Results  
Revealed user’s favorite artists.

v. Visualization  
Displayed using a bar chart of top 10 artists.

**4.5 Distribution of Play Duration per Track**

i. Introduction  
Analyzes how long tracks were played on average.

ii. General Description  
The 'minute\_played' column was visualized.

iii. Specific Requirements, Functions and Formulas  
Used **sns.histplot(df['minute\_played'])**

iv. Analysis Results  
Most tracks had short to medium duration. Few had long durations.

v. Visualization  
A histogram showed a right-skewed distribution.

**4.6 Weekday vs Weekend Listening Trends**

i. Introduction  
Compares listening patterns between weekdays and weekends.

ii. General Description  
Dates were classified as weekday or weekend based on **'day\_name()'.**

iii. Specific Requirements, Functions and Formulas  
Used a custom label map and counted frequency.

iv. Analysis Results  
Weekends generally showed higher or different listening behavior.

v. Visualization  
A bar plot compared listening counts between each weekday.

**4.7 Correlation Analysis and Heatmap**

i. Introduction  
Explores correlations between time-based features and listening duration.

ii. General Description  
Computed Pearson correlation matrix of numerical fields.

iii. Specific Requirements, Functions and Formulas  
Used **df.corr() and plotted with sns.heatmap().**

iv. Analysis Results  
Some relationships were revealed.

v. Visualization  
Heatmap visualizes correlation strengths.

**4.8 Z-Test: Statistical Comparison Between Two Artists**

i. Introduction  
Tests if two artists have significantly different average play durations.

ii. General Description  
Samples of 'minute\_played' were taken for two popular artists.

iii. Specific Requirements, Functions and Formulas  
Used **z-test** based comparison.

iv. Analysis Results  
The result showed whether the difference in user behavior toward the two artists is statistically significant.

v. Visualization  
Output displayed as printed test result with p-value.

**5. Conclusion**

This exploratory analysis of Spotify streaming history provides valuable insights into the user's listening behavior. From understanding peak hours and favorite artists to identifying usage patterns over days and durations. Techniques like correlation mapping and hypothesis testing further enhanced the depth of the study, making it both statistically insightful and visually informative.

**6. Future Scope**

Future work on this project can expand in several directions:  
- Analyze listening trends across genres.  
- Build a recommendation system based on usage patterns.  
- Integrate with real-time Spotify API for live behavior tracking.

