Tuning into Success:
Song Attributes and Popularity Based on Spotify Data

By: Nakita Hancock, Narmeen Mohammed, Tang Ng, Anu Panday, Maya Welford and Ildiko Wurth

INTRODUCTION

The aim of the project was to investigate a mutually interesting field within the team and formulate relevant questions aimed at addressing issues within the music industry.

The team collectively agreed to focus on 'Media and Entertainment,' with a united decision to delve into Spotify datasets sourced from Kaggle and the Last.fm API. This data not only aligned with the team's shared interests in the entertainment industry but also presented an opportunity to explore various questions relevant to challenges within the music industry.

Objectives

Collaboratively, the team identified specific questions to address, based on research and the data available:

1. What is the correlation between song attributes and song popularity?
    a. What is the relationship between these song attributes and song genre and popularity?
2. How have different song attributes and song genre popularity changed over the past decade?
3. Is there any correlation between song attributes and the song being 'Shazamed'?

This report outlines the team's approach to the analysis. It will first delve into the background of the project to contextualise the aim and objectives of the project, followed by the team's project specification and design in planning and the structure of analysis. It will then outline the team's implementation process and challenges. Finally, a conclusive section to summarise key findings and insights to end the report on in depth analysis.

At the end of the report, we hope to illuminate the current state of the music industry, shed light on the challenges faced by industry stakeholders, and elucidate why the chosen questions are pivotal for our target audience.

BACKGROUND

The music industry is an entertaining and highly profitable industry. It is heavily reliant on streaming services such as Spotify, accounting for 67% of sales. Stakeholders, including Spotify leaders, developers, record labels, current and upcoming artists benefit from understanding trends and attributes shaping song popularity. Using data from Spotify and the Last.fm API, our project is to study the likelihood of song popularity correlating with specific attributes, such as liveness and

energy. We will then analyse the difference in song attributes based on genre. Attributes will also be analysed for trends over the past decade to visualise the change in music with time. This helps to give a better understanding as to how music is evolving and how dependent this is on individual musical features.

SPECIFICATIONS AND DESIGN

Requirements

To conduct an in depth analysis, the technical and non-technical requirements listed in Table 1 need to be met. These requirements are a combination of those set by our team members and Code First Girls.

Table 1. Project Requirements

Technical Requirements	Non-Technical Requirements
● Code to be written in Python script in Jupyter Notebook with clearly defined sections: loading data, cleaning data, transforming data, analysis, visualisation, reporting. ● Use at least one API to fetch data. ● Use the following scientific packages to analyse and visualise the dataset: Pandas, NumPy, MatplotLib ● Data should be explored and analysed before visualisation. ● The code shouldn't time out by taking too long to run, or take up too much storage	● Clear, easy to read graphs for readers ● Evidenced conclusions and explanations for readers ● The results should be reliable and reproducible ● The code should be readable, well-explained and easy to follow ● Project should aim to answer the questions we set out in our group homework ● PDF document with clear project specification ● PowerPoint slide deck with key points for presentation. ● Ready to be submitted by midnight on 17th December!

Design and Architecture

The team used Colab to collectively write, review and contribute to the data analysis for the project. The project design is shaped ultimately by the questions set out, for which the team conducted a series of explorative, descriptive and statistical analyses for.

Data visualisation is a key aspect of the project design, and the team extensively discussed which graphs would be most appropriate to provide useful insights for the intended audience of the report.

The datasets from Kaggle included a data dictionary which summarised what the datasets included, and described the columns. This was helpful to understand the possibilities for analysis, and to shape the research questions.

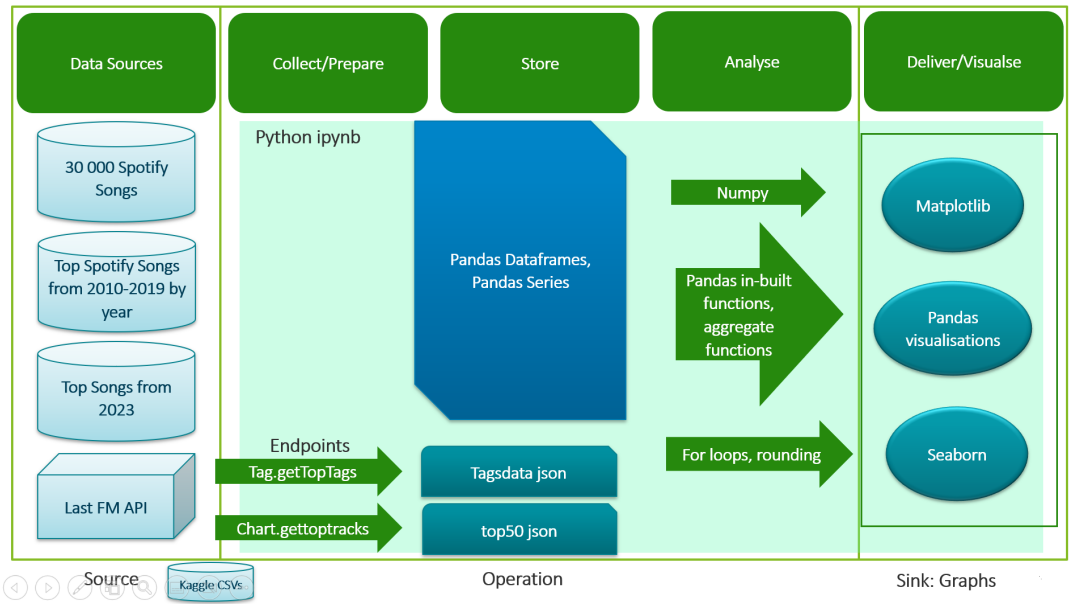


Fig 1. Project design

IMPLEMENTATION AND EXECUTION

The initial phase involved data exploration in our Jupyter Notebook, where we extracted key statistics such as mean values to familiarise ourselves with the data-sets. Subsequently, we encountered data cleanliness issues, including null values and non-numeric data in certain data-sets which we encoded to numerical data in order to use them in data analysis. To enhance readability, we consolidated subsets of pop genres into a singular genre.

Development Approach and Roles

Regular collaborative sessions were conducted to review the notebook, make decisions on graph selection, and establish a structured code layout. Opting to compartmentalise the code into sections for ease of navigation and adding clear comments explaining each graph's significance. This organisational approach streamlines access to specific data-sets or analyses with minimal scrolling.

Communication was maintained through various channels, including Slack and comments on shared documents, ensuring continuous collaboration even between meetings. Meetings involved screen sharing to collectively review data and reports. A collaborative walk-through allowed for discussions and expansion on any added or modified elements. Each meeting concluded with a to-do list, outlining individual responsibilities and tasks to be completed before the subsequent meeting or deadline.

The team incorporated multiple data-sets and utilised an API to gain a comprehensive understanding of music and its popularity. During collaborative brainstorming sessions, the team identified focal questions and determined the optimal graphs and libraries for conveying analyses

effectively. Workload was then distributed based on a preliminary SWOT analysis of each member's strengths and weaknesses and availability during an initial meeting.

Following the SWOT analysis and based on people's strengths, we created a project timeline (Table 2) and distributed tasks to members into the following main sections: data cleaning, data analysis, data visualisation, and report writing. Whilst each section was allocated to different members to incorporate everyone's ideas and inputs, we all decided that we would help where needed.

Table 2. Project Timeline and Responsibility

Date	**Name**	**Milestone**
By 1st Dec	Maya, Anu	Download data; use API to fetch data
By 1st Dec	Maya, Anu	Load data, clean, transform
By 7th Dec	Ildiko, Tang, Narmeen	Analysis
By 10th Dec	Nakita, Narmeen	Visualisation
By 14th Dec	All	Report/insights
By 15th Dec	All	Proofreading/checking
17 Dec		Submit

The project timeline helped us keep on track of time. However, when it came to analysing and visualising the data, it became clear that more of the team members were needed in order to meet the project deadline. We then had a meeting to discuss what questions each of us were going to tackle and what graphs/plot were required for each question.

Tools and Libraries

As shown in Table 3, we used a variety of tools and libraries to carry out the analysis. They were all essential for our project.

Table 3. Tools and Libraries used

Tools and Libraries	**What were they used for?**
Pandas Library	Loading, cleaning and analysing the data
Matplotlib and Seaborn Library	Data visualisation
NumPy and SciPy Library	Basic statistics
Requests Library	To make an API request for the Last.fm API
Google Colab	Sharing code and where we analysed and visualised our data in
Google Docs	Making notes and project documentation report
Slack/Google Meet	Communication

Implementation Process and Challenges

All team members worked collaboratively and independently, allowing the development of skill sets. The team openly shared their strengths and development areas, and agreed on what to focus on.

The initial phase involved data exploration in our Jupyter Notebook, where we extracted key statistics such as mean values to familiarise ourselves with the data-sets. Subsequently, we encountered data cleanliness issues, including null values and non-numeric data in certain data-sets. To counter this, depending on the number of null values, type of value, and depth of analysis, we either imputed the null values with means or the nearest valid observation, or dropped the null values. This resulted in data that did not error based on null values, leading to easier analysis.

One challenge was data types, whereby the data type was not appropriate for analysis, for example numbers were formatted as strings instead of floats or integers as they were in quotation marks. We deleted the quotation marks , which allowed us to use numerical methods to analyse the data.

Another challenge was dealing with duplicates. One dataset contained duplicates of songs, but also each song could belong to multiple genres, which is important for analysis by genre. Rows that were completely duplicated were dropped. To keep the integrity of the data, instead of removing songs from all but the first appearing genre, two dataFrames were made: one where all duplicates, regardless of genre, were dropped, so this dataset had no duplicates and could be analysed across genres, and a second dataframe to be grouped by genre. This allowed for analysis both as a complete set of songs, and by genre.

During initial analysis, it was noted that a lot of songs were not popular, with a track popularity of less than 10. It seemed rather anomalous, with large histogram peaks at 0-10, but on inspection of the data, these values are not anomalous. Instead it indicates that with 30000 songs, not all songs are going to be popular. It also shows how data extraction affects the data and subsequent analysis.

On producing the correlation heatmap for attributes versus track popularity, it was noted there was only weak correlation. As such, instead of analysing all the attributes against popularity, we chose 3 attributes that correlated the most with popularity, and further analysed these by genre. This led to a greater understanding of the relationship (or lack of) between energy, duration and instrumentalness.

Flexibility in meeting times and effective communication on Slack allowed ongoing collaboration despite individual scheduling constraints and other responsibilities.

Initiating our exploration with line graphs to depict genre-popularity relationships, we realised their limited clarity and informativeness. We experimented with stacked bar charts, which provided a clearer visualisation of genres and their respective popularity levels.

One of the team's main challenges was script readability, which was identified early in the project. This was resolved by team members including commentary throughout the code, for enhanced readability.

Agile Development

The team used an agile approach with timed sprints, for example we worked iteratively and adjusted analysis of the questions after further exploring the datasets. Furthermore, code review was done throughout the project by various team members, which involved testing the code and discussing comments on the Colab file. Data cleaning and data analysis was edited and improved on each review. The team had weekly standup meetings, where each member shared what they had been working on, what they would work on next, and any blockers.

DATA COLLECTION

There is a vast number of datasets available on Kaggle. After the team browsed Kaggle and each suggested 3 potential datasets to use, we agreed on Spotify / song data. Our decision was further supported given that there are multiple Spotify-related datasets on Kaggle, which we therefore used.

To answer our question, we required data that spanned many songs and their associated attributes over several genres. We also needed track data that covered the last decade. Finally, we also needed data that included how popular Shazam was.

Use of API

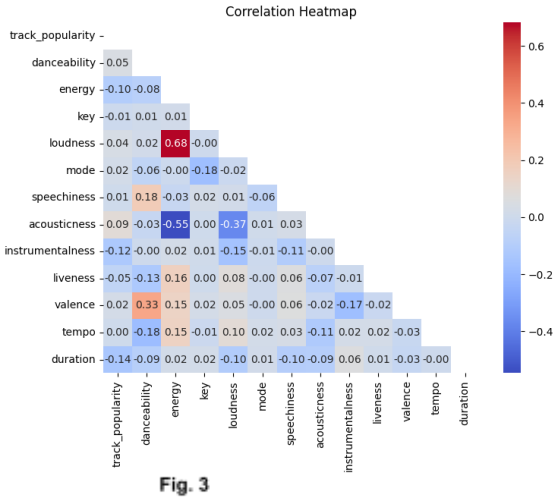
To include current information, we used the Last.fm API. Last.fm is the world's largest online music service, and its API allows data including album, artist, chart, location to be obtained. This was used to provide the current top genres and top 50 chart artists. Get requests are coded into the .ipynb analysis so that the data is live. For API keys, please refer to https://www.last.fm/api/authentication.

RESULT REPORTING - Summary of Results in .ipynb file:

Song Attributes and Popularity

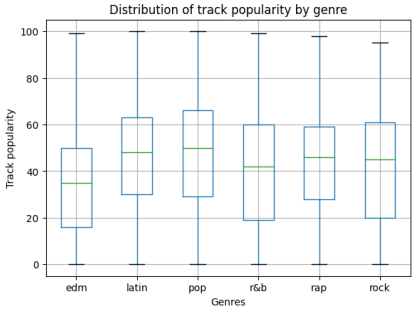
In terms of track popularity, the factors with the most effect are duration, instrumentalness, and energy. However, the correlation is still weak, with decreased duration, decreased instrumentalness, and decreased energy. (Fig. 2)

Fig. 2



When analysing the popularity distribution by genre, we found rap and R&B have the widest distributions, whereas rap and latin have the least distributed popularities. (Fig. 3)

Fig. 3



The most popular genre is pop, unsurprisingly (Fig. 4). We then looked at duration, instrumentalness and energy as these correlated with popularity.

For song duration, most genres appear positively skewed, with songs leaning towards shorter durations. The shortest songs are most likely to be EDM. Whereas for energy, EDM and rock playlists have the highest energy, followed by latin and pop. R&B playlists are the least energetic. (Fig. 5)

Fig. 6 shows there is little variation in instrumentalness, but on analysing the light blue area with high popularity and high instrumentalness, many of these songs were in rap playlists (contradicting the high instrumentalness and low vocal). Otherwise, there is a range of genres with varying instrumentalness. This shows that a variety of songs were present per genre.

Song Attributes Across the Decade

Of all the attributes, song duration has the most noticeable change across the decade, showing a declining trend, possibly due to reducing attention span due to the increase in social media, and cost-savings associated with producing shorter tracks (Fig. 7). We also noted pop remained the most popular genre across the decade. (Fig. 7) We did some further analysis to compare with current trends using the Last FM API. (Fig. 9)

Correlation of song attributes with Shazam

For correlation with Shazam, there was very weak correlation to attributes. As such, it appeared no individual attribute correlates with the song being Shazamed. (Fig. 8)

Fig. 4

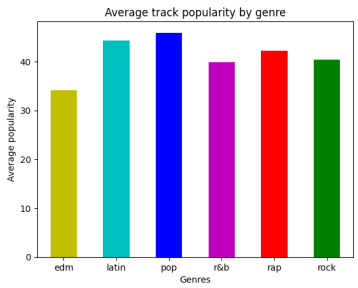


Fig. 5

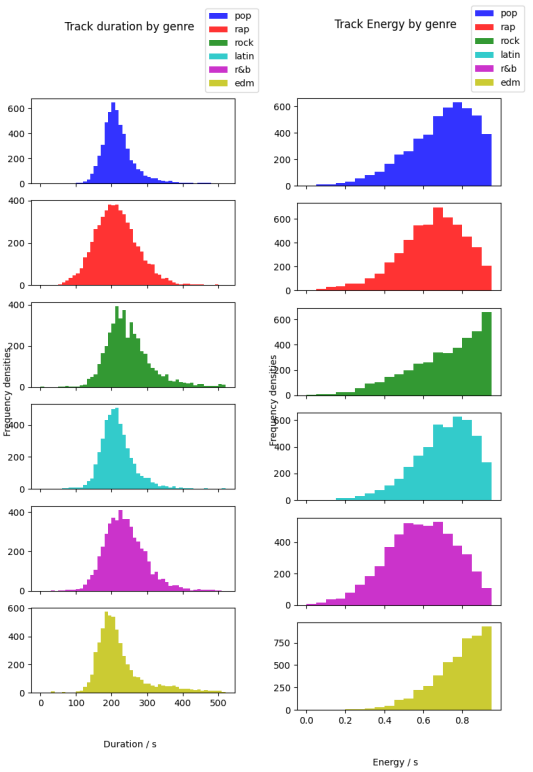
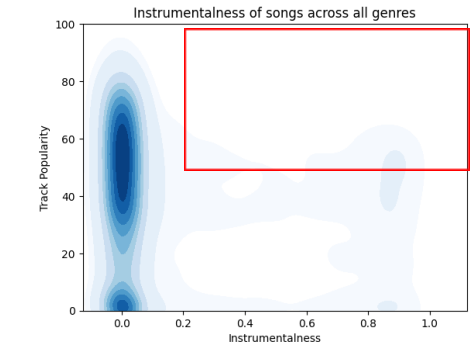


Fig. 6



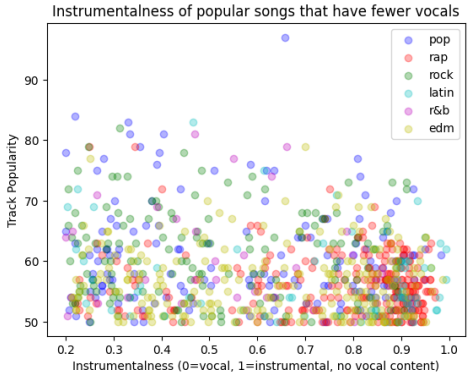
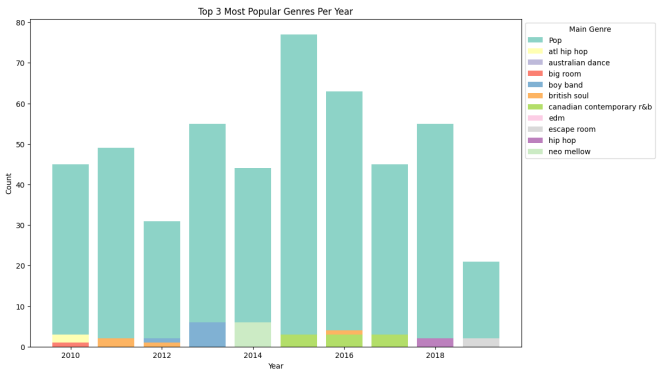


Fig. 7



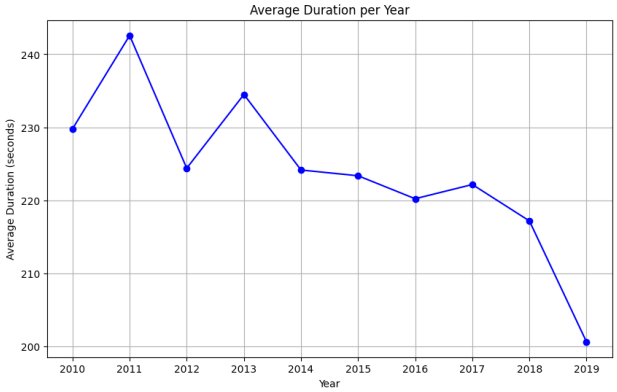


Fig. 8

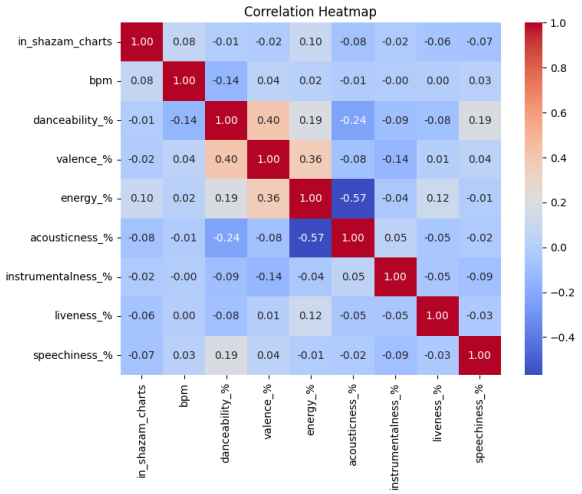
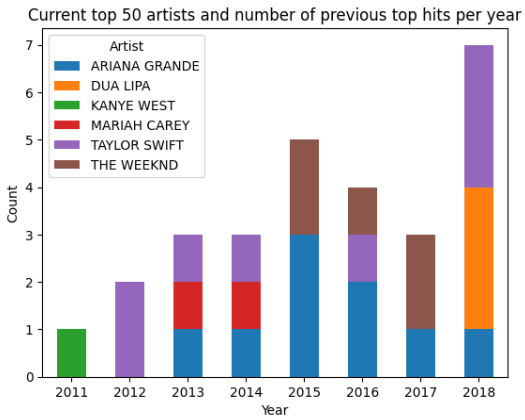


Fig. 9



CONCLUSION

In conclusion, our project involved a comprehensive exploration of Spotify song attributes and their correlation with popularity, aligning with the broader context of the entertainment industry. Our objectives were designed to address pertinent questions surrounding the relationships between song features and their popularity, including over time.

By researching the music industry more widely, we uncovered the substantial role of streaming platforms like Spotify, hence confirming our decision to conduct our analysis on Spotify datasets. Our project aimed to contribute valuable insights, shedding light on the nuanced dynamics that influence song popularity, thereby benefiting stakeholders ranging from streaming services to record labels and aspiring artists.

Regarding specifications and design, our project adhered to a set of technical and non-technical requirements, ensuring a robust and effective analytical process. The design, architecture, and implementation phases were characterised by a thoughtful approach, leveraging tools like Colab, and using agile methodologies for flexibility.

The implementation and execution phase involved a collaborative effort to tackle challenges such as data cleanliness and script readability. The team dynamically adapted to changes, optimising our approach through iterative development and agile practices. The SWOT analysis and project timeline facilitated efficient task distribution, ensuring that each team member's strengths were maximised.

The data analysis and insights gained from this project have highlighted the correlation between song attributes and popularity, links with genre, and changes over a 10-year period, therefore offering a further understanding of trends in the music industry to our audience.

For more comprehensive research and future improvements, it is crucial to consider incorporating additional data sources. Exploring year-by-year trends and delving deeper into the relationship between product features and popularity could uncover nuanced patterns. Enriching our dataset with information on customer demographics and economic indicators would also provide a more holistic view.