Analyzing Spotify Dataset

Lucas Gaspar
Bachelors in Cyber Security
Wentworth Institute of Technology
Boston, Massachusetts
gasparl@wit.edu

***Abstract*— Spotify is the world's largest music streaming platform, hosting countless artists ranging from international icons to independent artists trying to make a name for themselves. Each song uploaded to Spotify has its listening activity tracked, along with various characteristics of the song. By analyzing these traits, we learn about what makes a song reach the top of the charts as well as what might keep some from getting there. Additionally, genre analysis and prediction can be done, leading to more accurate Spotify recommendation and song grouping.**

Keywords*— Spotify, Music, Machine Learning, Genre, Marketing

I. Introduction (*Heading 1*)

Begin your introduction by clearly presenting your topic and explaining its significance - why it is important or interesting. Instead of listing questions separately, weave them together into a cohesive narrative that naturally connects the topic, its relevance, and its context. Provide an overview of existing research and key findings in this area, incorporating necessary citations to support your discussion. Your goal is to create a compelling introduction that sets the stage for your report.

Spotify is the world's largest music streaming platform, hosting countless artists ranging from international icons to independent artists trying to make a name for themselves. Each song uploaded to Spotify has its listening activity tracked, along with various characteristics of the song. By analyzing these traits, we learn about what makes a song reach the top of the charts as well as what might keep some from getting there. Additionally, genre analysis and prediction can be done, leading to more accurate Spotify recommendation and song grouping.

There are some important questions that can be answered.

………

II. Datasets

A. Source of dataset (Heading 2)

In this section, introduce your dataset by explaining its source—where you obtained it and whether it is from a credible provider. Include details such as when the dataset was generated and how it was created by its original author. If you generated the dataset yourself, describe the methods and processes you used.

The data that I will analyzing is titled, "Spotify Music Dataset". It was uploaded on Kaggle in December 2024 by Solomon Ameh. Although it is not officially published by Spotify, we can confirm its validity by the fact that all information contained within the Spotify API can be accessed and used for free. The dataset's creator collected this information using python scripts.

B. Character of the datasets

Describe the dataset's format and size. Additionally, provide an overview of the dataset's characteristics, including its features, size, structure, and any relevant attributes that are important for your analysis. Describe the dataset's format and size, as well as its key features, including the parameters, columns, rows, and character attributes along with their respective units. Using a table to present this information is recommended for clarity. Explain whether you cleaned the data or converted any units, specifying the formulas or rules applied. If multiple datasets were combined, describe how they were merged. Additionally, mention if you created any new categories for analysis, detailing what they are and how they were generated. Providing this background ensures transparency and helps readers understand the reliability and relevance of your data.

The dataset contains exactly 1685 entries, each representing a specific song hosted on Spotify. Each song has stats about the specifications of the song such as its key and tempo, as well as the song's creators and release date. Some of the tracks were taken specifically for their popularity and others for their unpopularity. With this combined dataset, many questions can be answered regarding what traits songs in similar genres have as well as what causes songs to score high on popularity rankings.

III. Methodology

In this part, you should give an introduction of the methods/model. First, what's the method/model. What's the assumption of this method/model. What's the advantage/disadvantage of this method/model. Why did you choose it. What Python module or function do you apply to apply this method/model. Any optional input/extra work did you adjust to make the results better. If you have multiple methods, feel free to use subsection A., B. to separate them.

To answer each question as accurately as possible, I implemented several different models and visualizations, each specializing in looking at the data in a new way. Each of these methods, as well as the question that they were used to answer, can be seen below in parts A though ___.

(notes:

- Talk ab machine learning queston
- Maybe don't group by question?
- Iono check flow)

979-8-3315-0993-4/25/$31.00 ©2025 IEEE

Questions:

1. What parts of a song have the highest impact on its popularity?
2. What is the most interesting outlier in the data? (What song is popular despite not being like the rest and why?)
3. How do the statistics and rankings compare on Spotify vs YouTube?
4. Can songs be automatically categorized to certain genres based on their features such as tempo and energy?

A. *Method A*

Example: The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

$$a + b = \gamma \tag{1}$$

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation. Use “(1)”, not “Eq. (1)” or “equation (1)”, except at the beginning of a sentence: “Equation (1) is . . .”

B. *Method B*

- Bulletin 1
- Bulletin 2.
- Bulletin 3

C. *Method C*

Example: The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

$$a + b = \gamma \tag{1}$$

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation. Use “(1)”, not “Eq. (1)” or “equation (1)”, except at the beginning of a sentence: “Equation (1) is . . .”

An excellent style manual for science writers is [7].

IV. Results

In this section, present your findings using an appropriate method, such as equations, numerical summaries, or visualizations like charts and graphs. Clearly explain all results and provide guidance on how to interpret them. If any unexpected results arise, discuss possible reasons or contributing factors. To improve clarity and organization, consider using subsections (e.g., A, B) to separate different aspects of your results.

Example: After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

A. *Result A*

Example: XXX

1) For papers with more than six authors: Add author names horizontally, moving to a third row if needed for more than 8 authors.

2) For papers with less than six authors: To change the default, adjust the template as follows.

a) Selection: Highlight all author and affiliation lines.

b) Change number of columns: Select the Columns icon from the MS Word Standard toolbar and then select the correct number of columns from the selection palette.

c) Deletion: Delete the author and affiliation lines for the extra authors.

B. *Results B*

Example: Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

C. *Results C*

a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation “Fig. 1”, even at the beginning of a sentence.

TABLE I. TABLE TYPE STYLES

Table Head	**Table Column Head**		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

[a] Sample of a Table footnote. (*Table footnote*)

Fig. 1. Example of a figure caption. (*figure caption*)

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

V. Discussion

Every method/project has its shortage or weakness. Please discuss the unsatisfied results in your project. And discuss the feasible suggestions of future work to revise/improve your result.

Example: xxx

Identify applicable funding agency here. If none, delete this text box.

(notes:

- Look at songs following trends vs old popular songs
- Maybe change dataset to be more random
- Problem is many of the things that make a song are not measurable with numbers. Release time, artist sentiment, etc. are important
-)

VI. Conclusion

In this part, you should summarize your project. What important results did you find for your topic and what's the effect of this result on the real-world?

Example: xxx

Acknowledgment *(Heading 5)*

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

(look up what to put here)

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an MSW document, this method is somewhat more stable than directly inserting a picture.

To have non-visible rules on your frame, use the MSWord "Format" pull-down menu, select Text Box > Colors and Lines to choose No Fill and No Line.

References

Use the IEEE format for the citation. The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..." Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

[1] S. Ameh, "Spotify Music Dataset." Kaggle, 2024. [Online]. Available: https://www.kaggle.com/datasets/solomonameh/spotify-music-dataset/data

[2]

IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.