SPOTIFY

BUSA8090 DATA AND VISUALISATION*FOR BUSINESS*

MINH ANH (JANE) TRAN - 46720898

MACQUARIE UNIVERSITY

Table of Contents

Introduction

Data analytics has become essential in the modern music industry for making informed decisions and developing strategic plans. This paper examines Spotify data to extract key insights that can inform marketing strategies and optimise content placement for Sony Music's Chief Marketing Officer (CMO). The study is divided into two sections. The first section consists of visualisations that display different aspects of Spotify data. The second section discusses data governance mechanisms and ethical implications.

PART A

Visualisation 1: Influence of Playlist Inclusions on Streams

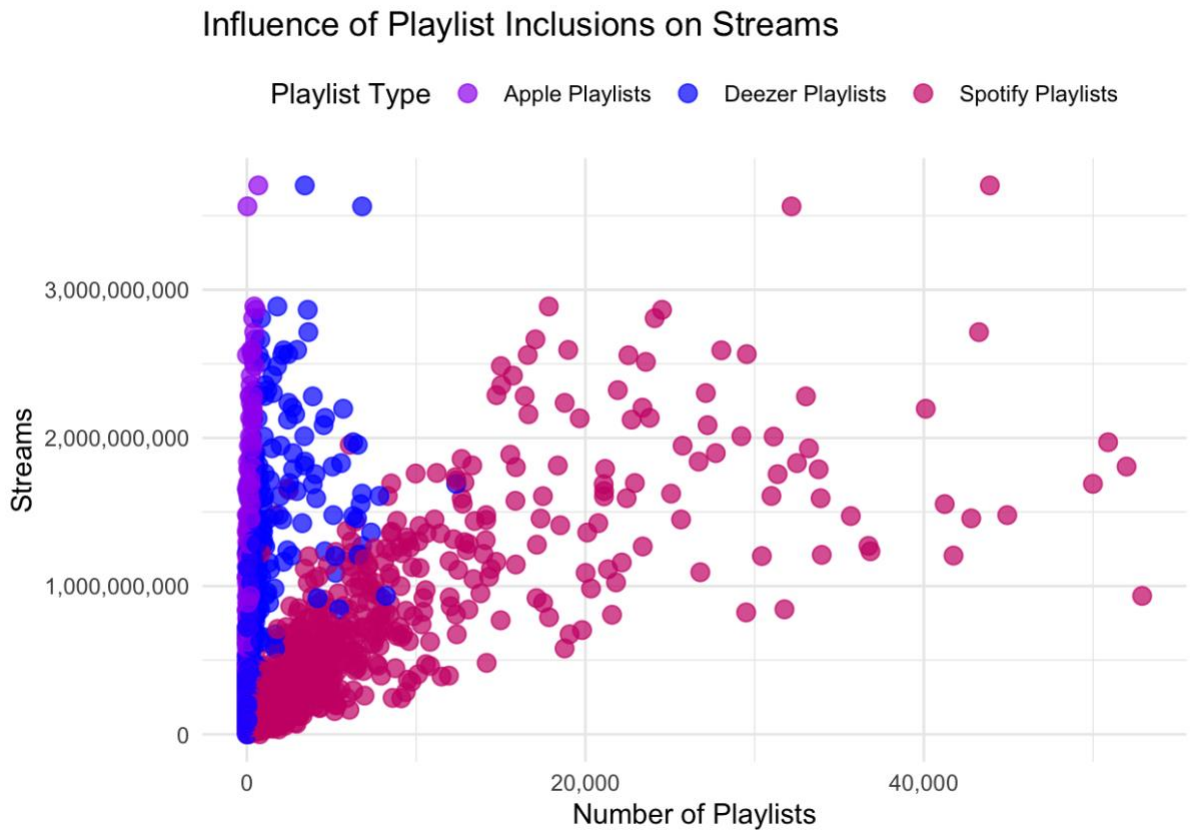

Figure 1. Influence of Playlist Inclusions on Streams

Using the "pivot_longer" function, the code transformed the data from a wide format to a long format. Different types of playlists (in_spotify_playlists, in_apple_playlists, and in_deezer_playlists) were merged into a single column called "playlist_type," and the related numbers were placed in a separate column called "playlist_count." Then, the data is filtered to only contain relevant playlists, and it was renamed to improve readability. After defining the specific colours for each type of playlist, we plot the scatter plot using the ggplot function. The plot matched the number of playlists to the relevant stream numbers, with points colored by playlist type for ease of readability.

The scatter plot portrays the correlation between the number of playlists a song is included in and the number of streams it receives, distinguished by playlist types (Apple, Deezer, and Spotify). Based on the graph, songs included in more playlists are more likely to receive higher streams, indicating a positive connection. Notably, playlists by Spotify show a larger

distribution among the axes. This suggests an interconnection between the larger number of Spotify playlists and the higher streaming counts. On the other hand, Deezer playlists show high streaming records regardless of fewer playlist inclusions, while Apple playlists tend to cluster more closely together.

The more songs included in a Spotify playlist, the larger the streams they get. Therefore, if a song is added to Spotify playlists, it is anticipated to witness a significant increase in streaming numbers. The finding predicts that users tend to create their preferred playlists and replay their favourite tracks on Spotify more frequently than other platforms such as Apple and Deezer. Marketing and promotion strategies can feature fan-made playlists on Sony Music's social channels and official Spotify profile to encourage playlist creation and receive incentives for exclusive content, shout-outs, or prizes for the most creative playlists. Artists and music marketers should concentrate on having their songs added to as many Spotify playlists as they can in order to boost streams. Spotify seems to have the biggest influence on streaming statistics because of its greater distribution and the possibility of high streams with more playlist entries. This could be because of its superior branding and increased appeal in the entertainment sector.

Visualisation 2: Streams by Released Month

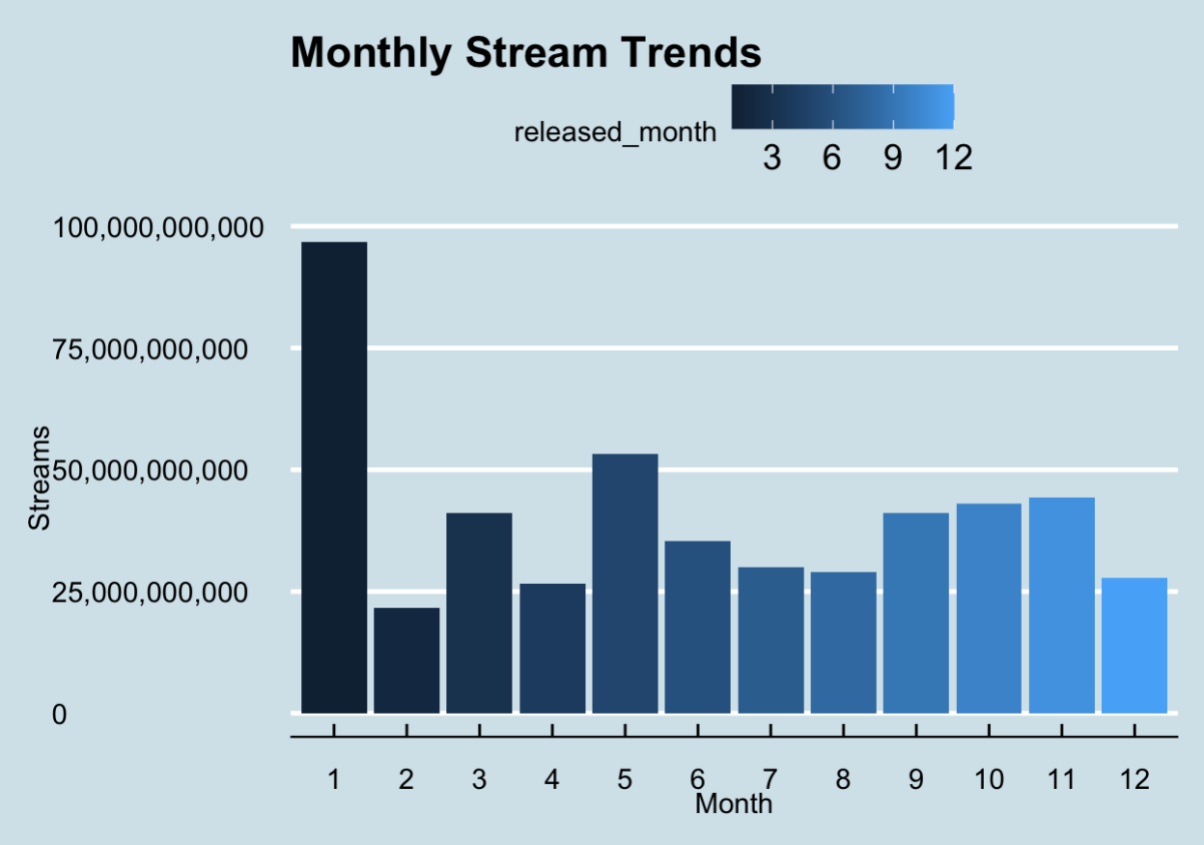

Figure 2. Streams by Released Month

The "as.numeric()" function is used to first convert the “streams" data to numerical form. By doing this, it is ensured that the values from the “streams" is handled effectively as numerical values. The analysis then produces a bar plot using the ggplot2 package. The “ggplot()” function specifies the data frame “df” and constructs the aesthetics "aes," with the x-axis representing the “released_month” (as a factor to treat it as categorical data) and the y-axis representing the streams. The analysis then employed the "fill = released_month" function to color-code the bars according to their month. The “geom_bar(stat = "identity")” code creates the bars representing the actual values of streams. The function “labs()” is used to add titles and labels to the plot. Then, the y-axis is formatted with “scale_y_continuous(labels = scales::comma)” to display the streams values with commas for better readability. The highlight of the graph is the application of the economist theme to portray the bar plot aesthetically.

Using Spotify's historical streaming data, the visualisation offers insights into the streaming activity throughout the twelve months. According to the plot, January has a exceptionally

higher number of streams than other months, suggesting a notable surge in streaming activity. This might result from a variety of factors, including popular content releases, resolutions made for the new year, and listening trends post holidays. The streaming activity levels over the remaining months remain steady, although there is a slight fluctuation. Finally, there is a discernible uptick in May and an ongoing rise from September to November. This implies that streaming behaviour may exhibit periodicity or occasional surges.

The plot's insights indicate that streaming activity ascends from September to November, with a peak in January. To take advantage of increased streaming activity, record labels and artists should carefully schedule their biggest releases and marketing initiatives during these times. During these busiest months, enhanced marketing initiatives like targeted advertising and unique content should be concentrated. Furthermore, ongoing trend analysis can assist in fine-tuning these tactics for maximum effect, guaranteeing that promotions and releases correspond with listener behaviour patterns.

Visualisation 3: Streams by Artist Count

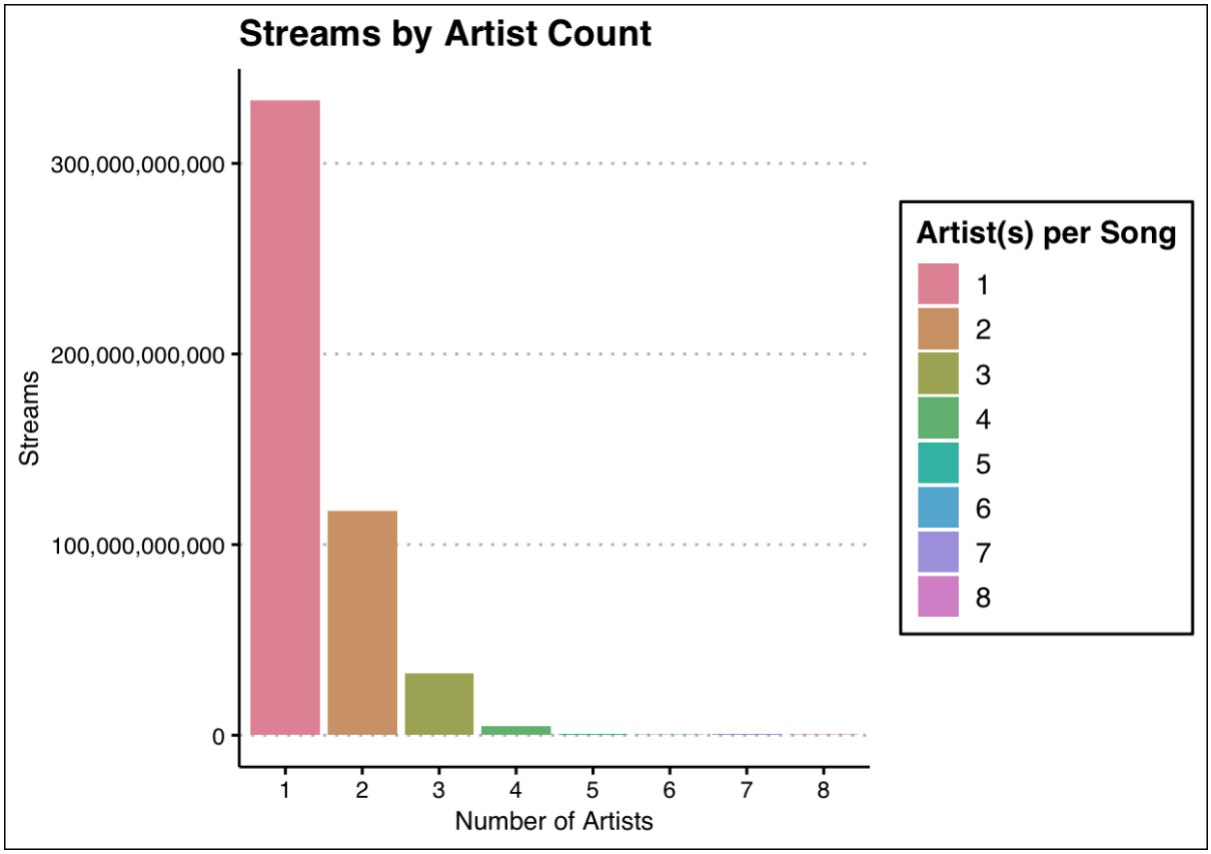

Figure 3. Streams by Artist Count

The analysis starts by generating a color palette with the “rainbow_hcl()” function to create distinc colour for each number of artist that collaborate in a song and better readability. To achieve this, the library “colorspace” was installed. The ggplot2 package is used to plot the data. In this plot, the x-axis represents the artist_count (the number of artists per song) and the y-axis represents the number of streams. The analysis uses the function “fill = factor(artist_count)” again to color the bars based on the number of artists per song. The “geom_bar(stat = "identity")” function creates the bars representing the actual values of streams. Titles and labels uses the “labs()” function and y-axis is scaled with “scales::comma” for better readability. The plot was styled using theme_clean() for a minimalist look, scale the colour of the plot manually and customise the legend for aesthetical purpose.

The visualisation displays the distribution of streams based on the number of artists contributing to a song. The figure suggests that single artists typically garner more streams, as songs featuring only one artist have the largest number of streams by a significant margin. While still rather high in terms of streaming, songs with two artists have far fewer than those

featuring only one artist. As the number of artists rises, the number of streams falls even further, with extremely few streams for songs with more than three performers. This tendency suggests that audiences prefer songs by solo or duet musicians over tracks with several collaborators.

This occurs due to larger number of solos was included in the data included to the dataset, with 587 songs by single artists, while collaboration of 2 only accounts for half of the solo's. Collaboration of 6, 7, and 8 artists only account for 2-3 songs included to the set list, which heighten the disparity. This can be due to solos is more preferred among artists according to various factors, including cost, time, and their personal preferences. The quality of the dataset can be improved to reduce this difference. Adding more tracks with larger collaboration can shows whether collaboration gains higher training, or will it be any improvement to the disparity.

The plot suggests that songs by solo artists typically receive more streams than tracks with several collaborators, potentially as a result of marketability or listener preference. As a result, artists and record labels may concentrate on solo releases while making sure these songs are well-marketed. Future datasets should have more collaborative tracks, with customised marketing strategy, in order to enhance the analysis and give a more comprehensive picture of their effects on streaming. Furthermore, it is imperative to highlight superior partnerships and tactical marketing initiatives in order to amplify the appealing qualities of multi-artist tracks.

Visualisation 4: Trends in Audio Features Over Time

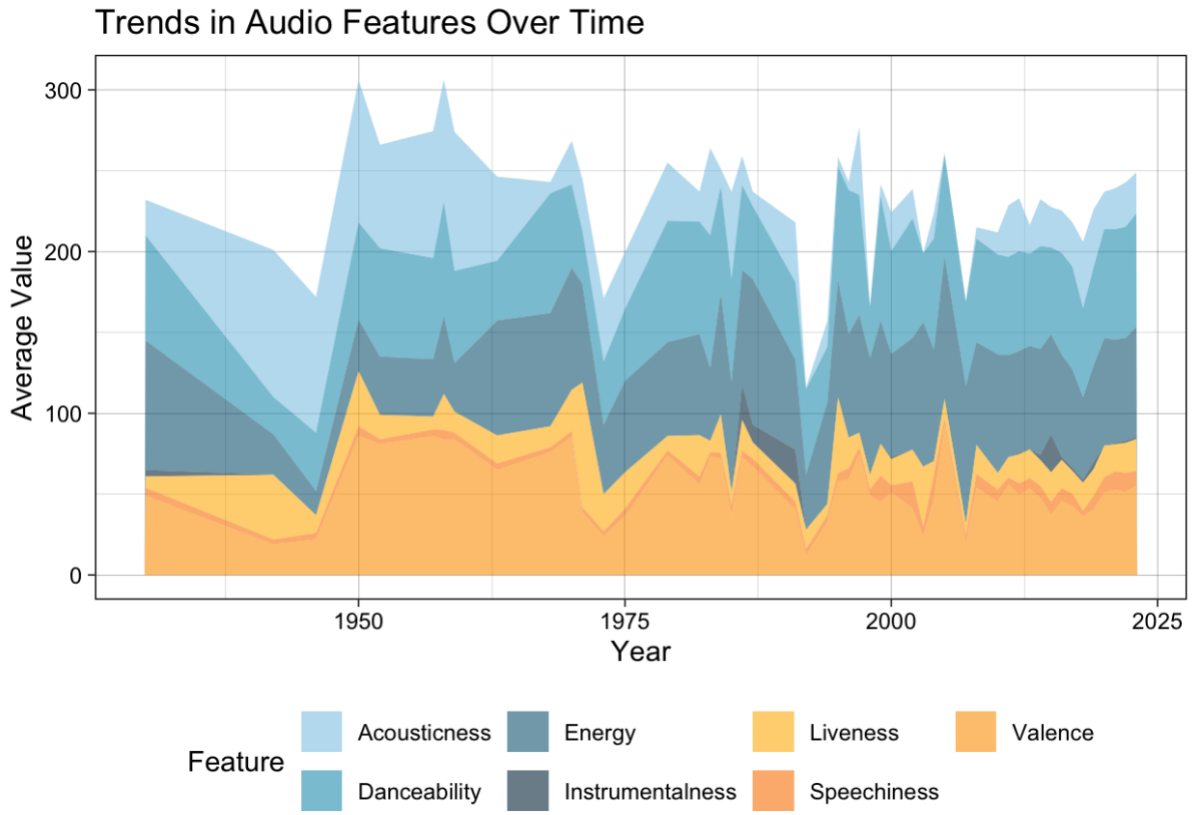

Figure 4. Trends in Audio Features Over Time

The data was first combined year-by-year to get the average values of the different auditory characteristics (liveness, instrumentality, acousticness, danceability, valence, energy, and speechiness) for every year. This was done using the "group_by" and summarise functions. Once the annual averages were acquired, the data was once more converted using the "pivot_longer" function from wide format to long format. Plotting became simpler as a result of the adjustment, which yielded separate rows for the average value of each feature. Then, the code recoded the feature names for clarity and created an area chart using ggplot2. Custom colours were defined for every feature using the "scale_fill_manual" function, and the plot aesthetics were adjusted for improved presentation.

The visualisation offers insight into the evolution of musical audio elements from the mid-19th century to the present. It demonstrates how attributes including Danceability, Valence, Energy, Acousticness, Instrumentality, Liveness, and Speechiness are changing over time. Overall, there is a strong correlation between the area along the period. Valence appeared in the fewest

tracks, while Acousticness was found to be the most favourable rhythm in the recordings. Over the years, average Danceability and Energy level of music have increased, suggesting a trend towards more and upbeat and rhythmically engaging compositions. Although Acousticness has varied greatly throughout time, it has generally decreased, indicating a move away from acoustic music and towards more electronically preference. Liveness and Speechiness remained relatively constant albeit with some slight fluctuation, which possibly due to changes in recording and production techniques.

Since these qualities are in line with the rising popularity of genres like pop, EDM, and hip-hop, it is likely that music will continue to favour stronger danceability and energy, given historical trends. It is observable that electronic music production gets progressively more attention and accessible. In order to satisfy modern listener preferences, artists and producers might consider adding aspects to their compositions that increase danceability and energy. This could entail employing dynamic production techniques, powerful rhythms, and lively tempos. To appeal to current listeners, embrace modern production techniques that lessen acousticness, such as digital effects and electronic instrumentation. Given the context, the Visualisation 5 will evaluate whether a higher Danceability level actually affects Streams.

Visualisation 5: Streams and Danceability

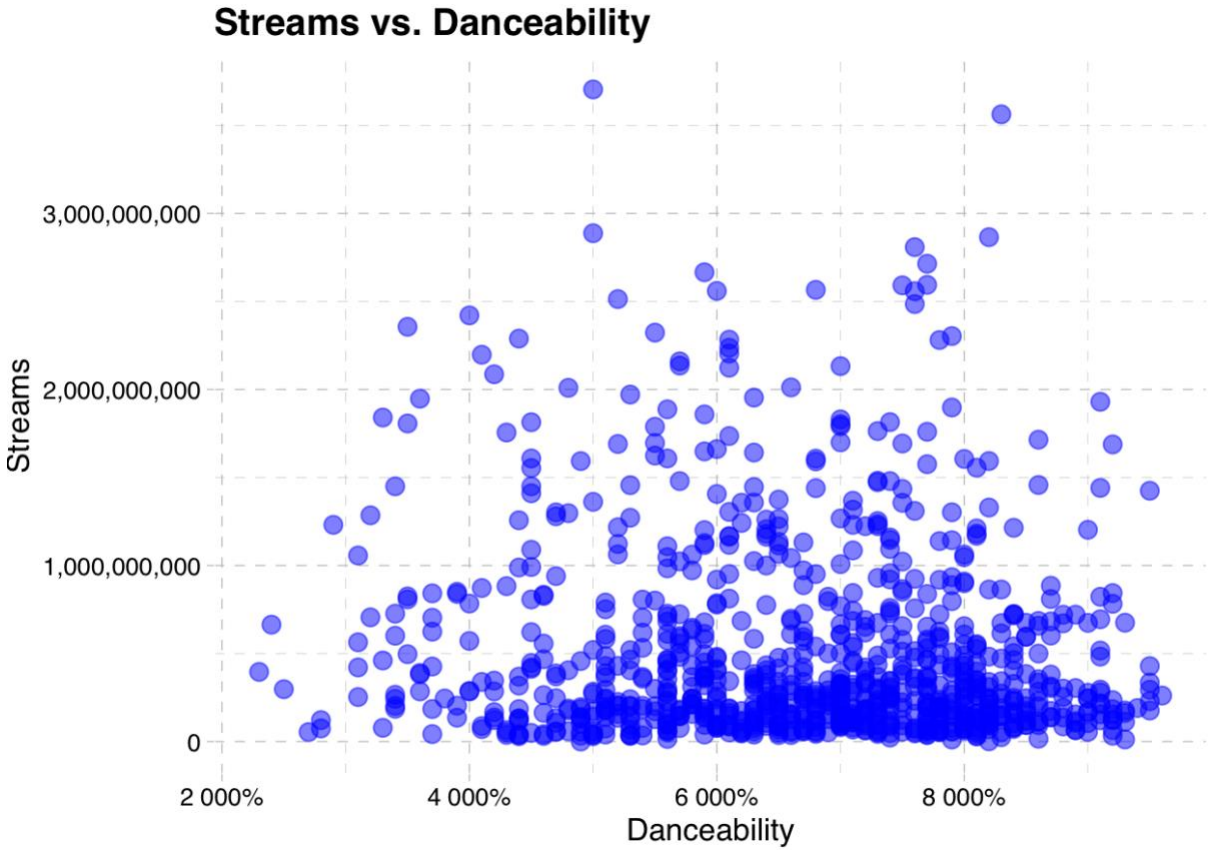

Figure 5. Streams and Danceability

The scatter plot is ideal for visualising the relationship between danceability and streams because it effectively displays the distribution and variability of individual data points. It allows for easy identification of patterns, trends, and outliers. The ggplot package is employed again, with danceability on the x-axis and streams on the y-axis. The aes() function mapped the columns to the respective axes. The points were added to the plot with geom_point() in blue, as well as readable size and transparency level, provides better visual clarity and distinction between overlapping points. display percentages using scale_x_continuous(labels = scales::percent) and the y-axis to display In this scatter plot, the analysis uses theme “pander” to minimalise the visualisation’s appearance.

This scatter plot investigates the correlation between the danceability of tracks and their streams. It assists the CMO in determining whether more danceable tracks tend to perform better, which can affect the type of music Sony Music prioritises in production and promotion to align with audience preferences. Each point represents a song, plotted according to its danceability score and total streams. The scatter plot shows that there is a wide range of streams

across all levels of danceability. Songs with low to moderate danceability (roughly 2% to 6%) tend to have a slightly higher variability in streams, including some very high stream counts. Higher danceability scores (above 6%) do not appear to correlate strongly with higher streams, as points are more scattered with fewer songs achieving extremely high stream counts in this range.

The visualisation's intention is to analyse whether danceability and higher user streaming are correlated, further complementing the analysis in Visualisation 4. That graph suggests there is a rare relationship between danceability and stream count. Danceability by itself is not an adequate indicator of a song's performance on streaming services, as evidenced by the large variations in streams across all danceability levels. Increasing danceability simply will unlikely to result in greater streams, given that higher streams are observed at different danceability levels. Artists and producers should take into account other features in combination with danceability to enhance their products' quality, rather than solely focusing on such factors. Promotion should consider aspects such as marketing, artist popularity, and genre trends in order to optimise streaming success.

PART B

1. Data Governance

Data governance is the process of guaranteeing that data is effectively administered in order to generate value for an organisation (Brous et al. 2020, DAMA International 2017). This piece of paper will discover four data analytics processes for data governance and the ways in which R features can facilitate these processes.

Data Collection and Effective Processing: In the past, managers have encountered challenges in establishing trust in data science products due to the frequent discovery of data that is not of the necessary quality (Symons & Alvarado 2016, Huang 2018). This issue highlights the importance of data collection from credible sources (Wallis et al. 2007). Guaranteeing that data is accurately ingested into the system and obtained from reputable sources enhances the intrinsic quality of the data and the integrity of the data systems, which facilitate future analysis, prediction, and organisational decision-making (Brous et al. 2020).

R function: Provides functions like read.csv(), import() %>% as_tibble()) and packages like DBI for database connections to ensure accurate data ingestion.

Data Quality and Cleansing: Poor quality data may have a detrimental impact on the efficacy and efficiency of an organisation, as it is increasingly involved in supporting organisational activities and driving business decisions (Ridzuan & Wan Zainon 2019). It is crucial for these businesses to have high-quality data in order to accomplish more precise and useful results, as they depend on data such as customer relationship management and supply chain management to run day-by-day operations (Ali & Mubeen Ahmed Warraich 2010).

R function: packages like “dplyr” for data manipulation, “tidyr” for data tidying, and “assertthat” for validation. Functions such as “mutate()” and “rename” are also commonly used for better supporting later visualisation.

Data Integration and Transformation: It is essential to integrate data from a variety of sources and convert it into a format that is usable for the purpose of conducting effective data analysis. This procedure guarantees that data from various sources can be combined, cleansed, and reorganised to generate a precise and cohesive dataset for analysis (Calabrese 2019).

R function: Packages and functions such as the "tidyverse", "reshape2", and "dplyr" for data wrangling allow for effective data manipulation and transformation, preparing the data for further analysis.

Data Storing: A fundamental component of data governance is the effective storage of data, which guarantees the integrity, availability, and security of the data over time. Effective data storage allows efficient retrieval and utilisation of data for future analytics (Dhudasia et al., 2021), which also guarantees compliance with regulatory requirements.

R function: functions such as “store()”, “save()”, or “saveRDS()” allow us to save one or more R objects to a file for later use in analysis, and modify the original data for future visualisation. Other functions, such as “write.csv()” and "write.table(), are also widely used in writing data frames to CSV files, and are useful for data storage and transfer.

2. Ethical Values

Accuracy of Data, Information, and Knowledge

Inaccurate interpretations and decisions can result from misrepresentation, whether through scale manipulation, selective data omission, or misleading graphical effects. Ensuring accurate presentation of data enables stakeholders to make well-informed decisions and avoid deceptive insights (Chen et al. 2009). This entailed guaranteeing that the data utilised was accurately sourced, cleansed, and represented in the visualisations without any distortions or manipulations (Bisoux 2023).

In Visualisation 2, the “streams” column was converted to numeric type and accurately displayed on the y-axis. This conversion enabled the bar heights to accurately reflect the true streaming number each month, while producing informative insights. Accuracy was necessary in identifying the peak streaming number in January and predicting future trends.

Accessibility of Data and Information

Accessibility was considered by selecting visual formats and concepts that are easily understood by a wide range of audiences. This entailed utilising legible labels, legends, and colour schemes that were understandable to people with colour vision problems (Quadri et al.

2024). Accessible visualisations democratise data insights, allowing them to reach a broader population.

In Visualisation 3 (Streams by Artist Count), the “rainbow_hcl()” function was used to set a colour palette with distinct colours for readability and a clean theme for minimal visual clutter. This strategy renders the visualisation approachable and clear to an extensive spectrum of audiences.

Invasion of Individual's Privacy

Preserving privacy entails the process of anonymizing data to prevent the discernment of persons, ensuring the security of sensitive data, and complying with privacy laws and regulations (Andrew et al. 2023). Adhering to ethical visualisation practices is essential for upholding trust and complying with legal regulations by respecting individual privacy (Brous et al. 2020).

All processes of visualisation in the report were ethically managed to not include any privacy invasion behaviours. For instance, the data in Visualisation 1 (Influence of Playlist Inclusions on Streams) was reviewed and analysed based on the type of playlist and the number of streams. It was ensured that no personal information on individual users or their listening habits was disclosed. This method ensured the protection of personal privacy while also offering valuable insights for prediction.

3. Justification on the prediction

The primary objective of the Spotify data analysis report and its visualisation is to have an advantageous effect on Sony Music's Chief Marketing Officer's decision-making process and accurately forecast the future conduct of consumers (Kappen et al. 2018). The five visualisations have been justified in the work in Part A. Here, the report discusses how prediction/prescription can be affected positively/negatively.

The utilisation of historical Spotify data for predictive analytics and decision-making can provide both valuable insights and potential pitfalls. In the positive sense, the examination of historical streaming trends facilitates the development of targeted marketing strategies and informed decision-making. Artists and publishers can optimise their platform impact by strategically planning releases and promotions based on historical patterns in user behaviour,

such as maximum streaming sessions and favourable genres (Soares Araujo et al., 2020). Furthermore, the data obtained from Spotify can be used to generate insights that can be used to inform playlist placement strategies. This enables artists to prioritise postings on high-performing playlists in order to extend their reach.

However, it is imperative to recognise the constraints and possible disadvantages associated with depending entirely on historical Spotify data. The presence of biases within the dataset, such as an overabundance of specific genres or demographics, might cause the analysis to be distorted and result in misleading strategies (Xiong, 2023). Furthermore, the existence of partial or faulty data (e.g. the absence of certain streams or incorrectly ascribed playlist places) can compromise the dependability of forecasts and recommendations that are derived from past trends (Zimran, 2020). Relying excessively on past information without taking into account evolving trends and external influences can lead to missed opportunities or ineffective tactics. This underscores the significance of complementing historical analysis with up-to-date data and industry expertise.

Finally, historical Spotify data serves as a great basis for predictive analytics and decision-making in the entertainment industry. However, it is crucial to approach this data with a critical mindset and supplement it with additional sources of information. By recognising the advantages and constraints of the existing data and employing a sophisticated method of examination, artists and labels can utilise Spotify data in a productive manner to amplify their streaming achievements while minimising any hazards and biases.

Reference

Ali, K. and Mubeen Ahmed Warraich (2010) 'A framework to implement data cleaning in enterprise data warehouse for robust data quality', *2010 International Conference on Information and Emerging Technologies* [Preprint]. doi:10.1109/iciet.2010.5625701.

Andrew, J., Eunice, R.J. and Karthikeyan, J. (2023) 'An anonymization-based privacy-preserving data collection protocol for Digital Health Data', *Frontiers in Public Health,* 11. doi:10.3389/fpubh.2023.1125011.

Bisoux, T. (2023) *The ethics of Data Visualization, AACSB.* Available at: https://www.aacsb.edu/insights/articles/2019/12/the-ethics-of-data-visualization (Accessed: 02 June 2024).

Brous, P., Janssen, M. and Krans, R. (2020) 'Data Governance as success factor for Data Science', *Lecture Notes in Computer Science,* pp. 431–442. doi:10.1007/978-3-030-44999-5_36.

Calabrese, B. (2019) 'Data Integration and transformation', *Encyclopedia of Bioinformatics and Computational Biology,* pp. 477–479. doi:10.1016/b978-0-12-809633-8.20459-7.

Chen, M. *et al.* (2009) 'Data, information, and knowledge in visualization', *IEEE Computer Graphics and Applications*, 29(1), pp. 12–19. doi:10.1109/mcg.2009.6.

DAMA International (2017) *Data Management Body of Knowledge*. Basking Ridge, NJ: Technics Publications.

Dhudasia, M.B., Grundmeier, R.W. and Mukhopadhyay, S. (2021) 'Essentials of Data Management: An overview', *Pediatric Research,* 93(1), pp. 2–3. doi:10.1038/s41390-021-01389-7.

Huang, J. (2018) 'From Big Data to knowledge: Issues of provenance, trust, and Scientific Computing Integrity', *2018 IEEE International Conference on Big Data (Big Data)* [Preprint]. doi:10.1109/bigdata.2018.8622561.

Kappen, T.H. *et al.* (2018) ‘Evaluating the impact of prediction models: Lessons learned, challenges, and recommendations’, *Diagnostic and Prognostic Research*, 2(1). doi:10.1186/s41512-018-0033-6.

Quadri, G.J. *et al.* (2024) ‘Do you see what I see? A qualitative study eliciting high-level visualization comprehension’, *Proceedings of the CHI Conference on Human Factors in Computing Systems* [Preprint]. doi:10.1145/3613904.3642813.

Ridzuan, F. and Wan Zainon, W.M. (2019) ‘A review on data cleansing methods for Big Data’, *Procedia Computer Science*, 161, pp. 731–738. doi:10.1016/j.procs.2019.11.177.

Soares Araujo, C.V., Pinheiro de Cristo, M.A. and Giusti, R. (2020) ‘A model for predicting music popularity on streaming platforms’, *Revista de Informática Teórica e Aplicada*, 27(4), pp. 108–117. doi:10.22456/2175-2745.107021.

Xiong, X. (2023) ‘Evaluating random sampling bias in sentiment analysis of social media data’, *Theoretical and Natural Science*, 25(1), pp. 36–42. doi:10.54254/2753-8818/25/20240895.

Zimran, A. (2020) ‘Recognizing sample-selection bias in historical data’, *Social Science History*, 44(3), pp. 525–554. doi:10.1017/ssh.2020.11.