**UNIVERSITY OF BRADFORD**

Text

Description automatically generated with low confidence

**University of Bradford**

**Faculty of Management,**

**Law and Social Sciences**

**BUSINESS DATA ANALYTICS**

**TITLE: SPOTIFY GROUP REPORT**

**GROUP 108**

**NAME: ZAINAB OLUWADAMILOLA AJAYI- 21037118**

**NWABUEZE OKOJI- 22013602**

**ONYEMAECHI CHUKWUKA 21035599**

**OUSMAN H. MBALLOW – 16015929**

**KENECHUKWU OFFOR – 22018958**

**VICTOR ODOGWU- 22029644**

# Executive Summary.

The music streaming industry has seen significant growth in recent years, with a significant increase in profit and royalties for both artists and companies (Terroso-Saenz, F., Soto, J. and Muñoz, A. (2023). This has led to a decline in the popularity of traditional music distribution platforms and a shift towards digital streaming services such as Spotify and Apple Music. Spotify is a widely used music streaming platform that possesses a large quantity of data on the listening habits and preferences of its users. This shift has greatly impacted the music industry, helping it recover from the negative effects of piracy and becoming the main form of music distribution for artists and labels.

This research aims to use Spotify's data on users' listening habits and preferences to gain insights into how song attributes such as valence, danceability, tempo, energy, genre, acousticness, and intensity impact subscribers' listening behaviour, to improve Spotify’s algorithms, user experience, and overall business strategy. Furthermore, it aims to analyse data from Spotify's music streaming platform to identify ways to increase the number of premium subscribers, which would lead to a higher revenue for the platform. The research focuses on identifying popular music genres and using regression analysis to predict listeners' preferred music based on their previous listening habits and contextual variables.

Secondary data was sourced from Kaggle from 2017-2020 with a total of 3,974 observations. The trends indicate that subscribers tend to stream music with higher danceability and valence attributes, rather than music with higher tempo and lower danceability attributes. These findings could be used by music and marketing companies to better understand streaming music consumption. However, the study has some limitations, such as a limited time frame and lack of consideration for certain contextual factors.

# Literature Review

Business analytics is a comprehensive and multi-disciplinary approach that employs advanced mathematical, statistical, machine learning and network science techniques to transform data into useful insights that assist in making informed and timely decisions. It is a process that encompasses different analytical methodologies to address business challenges, assess actions and inform decisions (Duan and Xiong, 2015; Kunc and O’Brien, 2019).

Business analytics can be considered as a connection between decision support systems (DSS) or business intelligence and data management systems (Frazzetto et al. 2019), acting as a bridge between data and decisions. DSS, which are computer-based solutions that are crucial tools for assisting problem-solving and decision-making, have undergone evolution over time, with the focus shifting from mathematical models to data-driven DSSs, which view data as the primary DSS tool.

Figure 1 illustrates the key areas that contribute to business analytics.

**Diagram

Description automatically generated**

**Source: Frazzetto et al. (2019)**

**Figure 1: Key areas that contribute to business analytics**

CRISP-DM, a well-known method for extracting knowledge from data, provides a thorough understanding of the data mining process. It is a goal-oriented approach that focuses on the essential steps required to achieve objectives. The CRISP-DM process starts by defining clear business goals and using available data to analyse and produce results that aid in decision making. There are 6 steps of CRISP-DM which includes business understanding, data understanding, data preparation, modelling, evaluation, and deployment.

Business analytics can be divided into three phases: descriptive analytics, predictive analytics, and prescriptive analytics. Prescriptive analytics builds upon the insights provided by descriptive and predictive analytics by helping to identify and evaluate the best options to achieve business objectives after analysing past and current activities, enabling data-driven optimization for decision support and planning (Frazzetto et al. 2019). These phases are:

1. **Descriptive Analytics:** This is the most basic and well-understood phase of analytics as it describes data in its current form and is data-driven (Raghupathi and Raghupathi, 2021). It involves summarizing business activities and reporting on them using historical data analytics, typically through data aggregation and mining techniques (Frazzetto et al. 2019). Both structured and unstructured data sources are used, with the primary focus being collecting, categorizing, characterizing, aggregating, classifying, identifying, and visualizing relevant patterns in the data (Lustig et al. 2010; Raghupathi and Raghupathi, 2021).
2. **Predictive Analytics:** This attempts to make accurate predictions and forecasts about possible future outcomes with the aid of analytical techniques based on previous patterns using machine learning and data mining (Delen and Ram, 2018; Frazzetto et al. 2019; Lepenioti et al. 2020; Raghupathi and Raghupathi, 2021). It extracts and synthesizes useful information from current and historical data, enabling business managers to forecast the probability of certain events, identify repetitive relationships or patterns, and determine patterns between events and extrapolating these patterns to make informed forecasts (Gandomi and Haider, 2015).
3. **Prescriptive Analytics:** This is the use of optimization, simulation, and heuristics-based decision modelling techniques (Duan and Xiong, 2015; Delen and Ram, 2018) to select the best alternatives for a decision process (identified by descriptive and predictive analytics) (Frazzetto et al. 2019; Lepenioti et al. 2020). Prescriptive analytics attempts to prescribe the best option to achieve the predicted future after analyzing and understanding past activities (Frazzetto et al. 2019). Sivarajah et al (2017) and Bihani and Patil (2014) state that predictive analytics determines the cause-effect relationship among analytical results and based on feedback from predictive analytical models, can help organizations with business process optimization.

Figure 2 presents the 3 phases of business analytics.

Diagram

Description automatically generated

**Source: Delen and Ram (2019)**

**Figure 2: 3 Phases of Business Analytics**

* 1. **Future Business Analytics Trends**

There has been a new technology paradigm shift in the way data is collected and processed since the 2010s. Figure 3 depicts the evolution of analytics over time.Timeline

Description automatically generated

**Source: A historical view to the evolution of analytics**

**Figure 3: Delen and Ram (2018)**

Big data or data deluge (Sivarajah et al. 2017) refers to the overwhelming amount of structured, semi-structured and unstructured information that cannot be processed adequately using traditional data processing systems and technologies (Bendre and Thool, 2016; Lee, 2017). This emerging technology development is as vital as nanotechnology and quantum computing (Sivarajah et al. 2017), growing at a rapid speed and transforming businesses into data-driven organizations (Lee, 2017).

* 1. **Characteristics of Big Data**

1. **Volume:** This pertains to the substantial amount of data that an organization generates. For instance, Sivarajah et al. (2017) report that Facebook generates more than 500 terabytes of data every day, and Walmart generates over 2.5 petabytes of data every hour from customer transactions.
2. **Velocity:** This pertains to the speed at which data is generated and processed. Nowadays, data is generated in real-time through digital devices and large complex networks connected daily to collect, analyze, and share data (Sivarajah et al. 2017; Lee, 2017).
3. **Variety:** This refers to the different data formats which include structured and unstructured data. Due to the proliferation of unstructured data, there is an increase in data variety as data is captured in different forms and sources such as messages (text, email, tweets, blogs), healthcare data, social media (image, audio, video), and sensor data readings (Sivarajah et al. 2017; Lee, 2017).
4. **Veracity:** This highlights the unreliability, uncertainty, incompleteness, inconsistency, and imprecision in large datasets. This raises concerns about data quality as veracity features measures data accuracy and its potential use for analysis (Vasarhelyi et al. 2015; Sivarajah et al. 2017; Lee, 2017).
5. **Variability:** This underscores that the rate of data is constantly and rapidly changing. It is distinct from variety as it highlights data whose meaning is constantly changing as data flows with unpredictable peaks and troughs can be volatile (Lee, 2017).
6. **Value:** This deals with gaining insights from large amount of structured and unstructured data. It involves the understanding that in its original form, pieces of data may be insignificant, through data analytics, high-valued data can be extracted which can then be used by organizations to increase revenue, decrease costs, and improve customer satisfaction (Sivarajah et al. 2017; Lee, 2017).
7. **Visualization:** This refers to the representation of data in an effective readable format using visual tools and platforms (Taheri et al. 2014; Bendre and Thool, 2016). For instance, Tableau – a big data visualization tool has the capability to transform large and complex datasets into easy-to-understand graphs, pictures, and charts, displayed in dashboards

**2.3 Opportunities and Challenges of data analytics for businesses**

According to Sivarajah et al. (2017), opportunities in data analytics include value creation, providing valuable business intelligence for decision-making, and improving the visibility and flexibility of supply chain and resource allocation. However, Lee (2017) suggests that big data poses several challenges for its development and management. He identified six specific challenges in the field, which include:

i. Data Quality: Ensuring that the data being used for decision-making is accurate and reliable is a challenge for big data, as it often comes from a wide variety of sources.

ii. Data Security: Ensuring the proper security of big data is a challenge as poor security can lead to financial and intellectual loss, as well as damage to a company's reputation.

iii. Privacy: The collection and use of personal data by companies is a growing concern as more and more companies rely on this data for their operations. However, the question of how much personal data is healthy enough to be given to third parties remains a concern.

iv. Investment Justification: Companies find it difficult to justify investments into big data projects due to unclear problem definitions and the use of emerging technologies.

v. Data Management: The volume, variety, veracity, velocity, and variability of big data requires significant investment in data storage, which can be expensive to manage. This can also require advanced data warehouse technologies.

vi. Shortage of Qualified Data Scientists: As the need for data manipulation increases, so does the need for qualified data scientists. However, a shortage of qualified data scientists is also a challenge.

1. **Top of Form**
2. **Methodology**

For our analysis, we used a dataset of the top songs from 2017 to 2020, daily released by Spotify and retrieved from Kaggle. This dataset allows us to record the most popular songs and their features over almost three years, providing a more reliable and comprehensive assessment of global music trends. The dataset includes 3,974 observations and focuses on thirteen mood-related features of a song, such as valence, danceability, instrumental, duration, energy, explicit, liveness, loudness, mode, popularity, speechiness, tempo, and acoustiness. However, it should be noted that the dataset is not comprehensive enough to represent global music trends from various countries and the number of people covered is restricted.

The analytical techniques used in this research are ANOVA, chi-squared test, one sample t-test, correlation test, simple linear regression, and multiple linear regression and are explained below:

ANOVA (Analysis of Variance) is a statistical technique used to compare the means of two or more groups. In this research, ANOVA was used to compare the average danceability across four years (2017, 2018, 2019, and 2020). The null hypothesis states that there is no statistically significant difference of the average danceability across the four years.

A chi-squared test is used to determine if there is a significant association between two categorical variables. In this research, the chi-squared test was used to determine the association between the mode rate and explicit music. The null hypothesis states that there is no association between the mode rate and explicit music.

A one sample t-test is used to examine whether the sample mean of a variable is statistically different from the population mean. In this research, the one sample t-test was used to examine whether the sample mean of danceability is statistically different from the population mean of danceability. The null hypothesis states that there is no statistical difference in the population mean and sample mean of danceability across 2017, 2018, and 2019.

Correlation test is a statistical technique used to test the relationship between two quantitative variables. In this research, correlation test was used to determine the strength and direction of the association between valence, tempo and the danceability variable. The results of the correlation test showed that there is a moderate positive correlation between valence and danceability and a weak positive correlation between tempo and danceability, indicating that valence may have a stronger impact on danceability than tempo.

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. In this research, linear regression was used to model the relationship between danceability and energy. The results of the linear regression analysis showed that there is a statistically significant relationship between energy and danceability, but the relationship is relatively weak.

Multiple linear regression is used to model the relationship between a dependent variable and multiple independent variables. In this research, multiple linear regression was used to model the relationship between danceability (dependent variable) and multiple independent variables: energy, valence, duration ms, instrumentalness, liveness, loudness, mode, speechiness, accousticness, and tempo.

1. **Findings.**

The tables below present the basic statistics of our variables.

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generatedGraphical user interface, application

Description automatically generated with medium confidenceChart, histogram

Description automatically generatedChart, histogram

Description automatically generated

**Chart, histogram

Description automatically generated**

|  |
| --- |
| In our research, ANOVA test was used to compare the average danceability across four years (2017,2018,2019, and 2020) and to determine if there was a statistically significant difference between the means. The null hypothesis proposed that there would be no difference. The decision rule for the study was to reject the null hypothesis if the probability value was less than 0.05, and not reject it if the probability value was greater than 0.05. The results of the study revealed that there was a statistically significant difference in terms of average danceability across the four years, leading to the rejection of the null hypothesis at a 5% significance level.  The results further revealed that there is a statistically significant difference between the means of 2017 and 2018, 2017 and 2019, and 2019 and 2020. Additionally, there is a statistically significant difference between the means of 2018 and 2019. Therefore, the null hypothesis is rejected at the 5% significant level. However, there was no statistically significant difference found between the means of 2018 and 2020, thus the null hypothesis is not rejected in this comparison.  Our research findings indicate that a statistical analysis was performed to compare the means of a control group (2017) with the means of the other groups (2018, 2019, and 2020). The results reveal that there is a statistically significant difference between the control group and the other groups at a significance level of 5%. Based on this, we reject the null hypothesis and conclude that there is a significant difference between the means of the control group and the other groups. The tables below present the ANOVA results. |

Table

Description automatically generated with medium confidenceDiagram, engineering drawing

Description automatically generatedChart, box and whisker chart

Description automatically generatedChart

Description automatically generatedChart, line chart

Description automatically generated

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | |  | | --- | |  | | |  | | --- | |  | |  | |  | | | |

|  |
| --- |
|  |

Table

Description automatically generated with low confidence

Chart, box and whisker chart

Description automatically generatedA chi-squared test is used to determine if there is a significant association between two categorical variables. The research findings show that a chi-squared test was used to determine the association between the mode rate and explicit music on Spotify, by comparing across four years (2017, 2018, 2019, and 2020). The null hypothesis proposed that there would be no association between the two variables. The results of the study revealed that the association between the mode variable and the explicit variable was statistically significant for the year 2017,2018, and 2020, thus rejecting the null hypothesis and indicating that there is a significant association between the two variables. However, for the year 2019, the association was not significant, and the null hypothesis was not rejected. Additionally, a 2 by 2 contingency table was used for all the four years and the Phi coefficient equals the Cramer’s V respectively. The research further revealed that the odds ratio for explicit music having lower odds in mode rate than non-explicit music in year 2017 was 40.13%, in year 2018 was 29.29%, in year 2019 was 21.66% and in year 2020 was 37.32% respectively.

The tables below present results from the Chi-squared test.

A screenshot of a computer

Description automatically generated with low confidenceA close-up of a document

Description automatically generated with medium confidenceA close-up of a document

Description automatically generated with medium confidenceA screenshot of a computer

Description automatically generated with low confidence

One sample t-Test was used to in this research to examine the difference between the sample mean of danceability and the population mean of danceability across four years (2017,2018,2019, and 2020). The null hypothesis proposed that there would be no difference between the sample mean and population mean. The decision rule for the study was to reject the null hypothesis if the p-value is less than 0.05, and not reject it if the p-value is greater than 0.05. The results of the study revealed that there was no statistical difference in the population mean and sample mean of danceability across 2017, 2018 and 2020, but revealed statistical difference in 2019. The sample mean and population mean for the four years were found to be x̄ =0.6450 and μ = 0.6428 respectively with t-value = 0.00, Pr = 0.9980, which indicates that there is a small interval between the sample mean and population mean and fail to reject the null hypothesis.

In this research, a correlation test was used to examine the relationship between two quantitative variables, valence, tempo, and danceability. The correlation coefficient was used to determine the strength and direction of the association between the three variables. The decision rule for the correlation test is to reject the null hypothesis if the calculated correlation coefficient (r) is outside the acceptance region. The acceptance region is determined by the level of significance (alpha) and the sample size. The results of the analysis showed a moderate positive correlation between valence and danceability with a correlation coefficient of 0.4189 and a p-value of <0.0001 which indicates that there is a moderate positive correlation between the two variables and is highly statistically significant. Also, the results showed a weak positive correlation between tempo and danceability with a correlation coefficient of 0.10214 and a p-value of <0.0001 which indicates that there is a weak positive correlation between the two variables and it's highly statistically significant. It's important to consider other factors that may affect the danceability of the song beside tempo. In summary, the correlation between danceability and valence is stronger than the correlation between danceability and tempo, indicating that valence may have a stronger impact on danceability than tempo. The tables below present the findings.

Text, letter

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

A linear regression was used in this research to model the relationship between danceability (dependent variable) and energy (independent variable). The decision rule for simple linear regression is to reject the null hypothesis if the calculated test statistic is greater than the critical value from the appropriate statistical table, at a pre-determined significance level (such as 0.05). The results of the analysis showed that there is a statistically significant relationship between energy and danceability with a p-value of <0.0001, which suggests that the relationship observed in the sample data is highly statistically significant and unlikely to be due to chance. The F-value of 67.78 and the t-value of 67.06 both indicate that the model is a good fit for the data. The R-squared value of 0.0168 represents the proportion of variance in the dependent variable that can be explained by the independent variable. In this case, the R-squared value is small, indicating that only 1.68% of the variation in danceability can be explained by the energy of the song, other factors may play more important role in determining danceability. In summary, the results of this linear regression analysis indicate that there is a statistically significant relationship between energy and danceability. The tables below present the findings.

Table

Description automatically generatedDiagram, engineering drawing

Description automatically generatedChart, scatter chart

Description automatically generated

A multiple linear regression was used to model the relationship between danceability (dependent variable) and multiple independent variables: energy, valence, duration ms, instrumentalness, liveness, loudness, mode, speechiness, accousticness, and tempo. The results of the analysis showed that there is a statistically significant relationship between danceability and the independent variables with a p-value of 0.0001, which suggests that the relationship observed in the sample data is highly statistically significant and unlikely to be due to chance. The F-value of 262.72 indicates that the model is a good fit for the data. The t-values and p-values for each independent variable give information about the importance of each variable in the model. A high t-value and a low p-value for an independent variable indicates that the variable is a strong predictor of the dependent variable. For example, valence has a t-value of 21.71 and a p-value of <0.0001 which indicates that valence is a strong predictor of danceability. Energy has a t-value of -15.21 and a p-value of <0.0001, this indicates that energy is a strong predictor of danceability but has a negative relationship with danceability, as the energy increases, danceability decreases. In summary, the results of this multiple linear regression analysis indicate that there is a statistically significant relationship between danceability and the independent variables. The R-squared value of 0.3987 suggests that the independent variables can explain about 39.87% of the variation in danceability. The tables below highlight the results.

**Table

Description automatically generatedDiagram

Description automatically generated**

|  |
| --- |
|  |

1. **Discussion/Conclusion.**

In this research, the business objective is to utilize Spotify's data on users' listening habits and preferences to gain insights into how song attributes, such as mood, valence, danceability, tempo, acousticness and other contextual factors may impact subscribers' listening behaviour. By understanding these patterns, Spotify can improve its algorithms, user experience, and ultimately its overall business strategy.

Findings emanating from this study are as follows:

1. ANOVA test findings suggest that danceability may be an important factor in determining which songs are streamed on Spotify from 2017 to 2020.

2. The Chi-squared test findings from 2017 to 2020 suggest that explicit music may be less likely to be streamed on Spotify, which aligns with the business objective of increasing streamed songs according to customer preferences.

3. The One sample t-Test findings suggest that danceability may be an important factor in determining which songs are streamed on Spotify, and the data aligns with the business objective as the sample mean of danceability for all four years does not show a significant difference from the population mean.

4. The Correlation test findings from 2017 to 2020 suggest that valence and tempo may be important factors in determining which songs are streamed on Spotify, which aligns with the business objective.

5. The Linear regression findings suggest that energy may not be a strong predictor of which songs are streamed because of a relatively weak relationship with danceability from 2017 to 2020.

Based on the findings, there are several implications and recommendations:

* ANOVA test findings suggest that the danceability of songs on Spotify may have changed over time. As a result, it's important for Spotify to understand how the danceability of songs is affected over time and consider it when curating playlists and making recommendations to listeners.
* Chi-squared test finding indicates that songs with different mode rate and explicit music may have different chances of being streamed on Spotify. As a result, it's important for Spotify to consider the mode rate and explicitness when curating playlists and making recommendations to listeners.
* The Correlation test findings indicate that valence and tempo may be related to danceability and could be used to select songs with high danceability. As a result, it's important for Spotify to consider valence and tempo when curating playlists and making recommendations to listeners.
* The Linear regression findings suggest that other variables such as valence, duration ms, instrumentalness, liveness, loudness, mode, speechiness, accousticness, and tempo may play a more important role in determining danceability. As a result, it's important for Spotify to consider these variables when curating playlists and making recommendations to listeners.

Thus, it is expected that Spotify take a more data-driven approach by employing machine learning and artificial intelligence (AI) to identify patterns and relationships between the variables that affect the danceability of songs. This can be used to create a more personalized listening experience for users and increase the chances of them streaming more songs.

**References**

Bendre, M.R. and Thool, V.R. (2016). Analytics, challenges and applications in big data environment: a survey. *Journal of Management Analytics* 3(3), 206-239

Bihani, P. and Patil, S. T. (2014). A comparative study of data analysis techniques. *International journal of emerging trends and technology in computer science* 3(2), 95-101

Chen, H., Chiang, R.H.L and Storey, V.C. (2012). Business intelligence and analytics. *MIS quarterly* 36(4), 1165-1188

Delen, D, and Ram, S. (2018). Research challenges and opportunities in business analytics. *Journal of business analytics* 1(1), 2-12.

Duan, L. and Xiong, Y. (2015). Big data analytics and business analytics. *Journal of management analytics* 2(1),1-21.

Frazzetto, D., Nielsen, T.D., Pedersen, T.B. and Siksnys, L. (2019). Prescriptive analytics: a survey of emerging trends and technologies. *The VLDB journal* 28,575-595

Gandomi, A. and Haider, M. (2015). Beyond the hype: big data concepts, methods, and analytics. *International journal of information management*, 35(2), 137-144.

Hassan, N.R. (2019). The origins of business analytics and implications for the information systems field. *Journal of business analytics*, 2(2), 118-133.

Holsapple, C.W., Lee-Post, A. and Pakath, R. (2014). A unified foundation for business analytics. *Decision support systems*, 64, 130-141.

Kunc, M. and O’Brien, F.A. (2019). The role of business analytics in supporting strategy processes: opportunities and limitations. *Journal of the operational research society* 70(6), 974-985.

Lee, I. (2017). Big data: dimensions, evolution, impacts, and challenges. *Business horizons* 60, 293-303

Lepenioti, K., Bousdekis, A., Apostolou, D. and Mentzas, G. (2020). Prescriptive analytics: literature review and research challenges. *International journal of information management*, 50, 57-70

Lustig, I., Dietrich, B., Johnson, C. and Dziekan, C. (2010). The analytics journey. *Analytics magazine* 3(6), 11-13

Martinez-Plumed, F., Contreras-Ochando, C.S., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M.J. and Flach, P. (2021). CRISP-DM twenty years later: from data mining processes to data science trajectories. *IEEE transactions on knowledge and data engineering*, 33(8), 3048-3061.

Raghupathi, W. and Raghupathi, V. (2021). Contemporary business analytics: an overview. *Data* 6(86), 1-11

Sivarajah, U., Kamal, M.M., Irani, Z. and Weerakkody, V. (2017). Critical analysis of big data challenges and analytical methods. *Journal of business research* 70, 263-286

Taheri, J., Zomaya, A.Y., Siegel, H.L. and Tari, Z. (2014). Pareto frontier for job execution and data transfer time in hybrid clouds. *Future generation computer systems*, 37, 321-334.

Vasarhelyi, M.A., Kogan, A. and Tuttle, B.M. (2015). Big data in accounting: an overview. *Accounting horizons*, 29(2), 381-396