Spotify Chart Data: Musical Attributes and Artist Popularity as Predictors of Chart Position.

Bertie Harte

Student No: R00206922

Supervisor: Dr. Justin Mc Guinness

For the module DATA8006 – Data Science Analytics Project

as part of the Higher Diploma of Science in Data Science and Analytics, Department of Mathematics, 04th May 2022.

Declaration of Authorship

I, Bertie Harte, declare that this thesis titled 'Spotify Chart Data: Musical Attributes and Artist Popularity as Predictors of Chart Position.' and the work presented in it are my own. I confirm that,

- This work was done wholly or mainly while in candidature for the Higher Diploma at Munster Technological University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at Munster Technological University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. Except for such quotations, this project report is entirely my own work.
- I have acknowledged all main sources of help.
- I understand that my project documentation may be stored in the library at MTU and may be referenced by others in the future.

Signed: Bertie Harte

Date: 04th May 2022

Acknowledgment

Thanks to Dr. Justin Mc Guinness for clear, precise direction and advice on my approach for this thesis; to the lecturing staff for the HDip in Data Science who have facilitated my journey over the last two years in what can best be described as challenging circumstances.

As always, many thanks must go to my wife Mary for her continued support of our family while I had my head stuck in learning, researching, programming and the general business of continued education while working full time.

Table of Contents

Abstract 7

1. Introduction 8

2. Literature Review 9

2.1 Musical Attribute and Track Performance 10

2.2 Artist and Track Performance 10

2.3 Attribute Measurement 11

2.4 Prediction 12

2.5 Literature Review Conclusion 13

3. Data source 14

3.1 Basic Dataset 14

3.2 Expanded Dataset 16

3.3 Data Exploration 19

3.4 Artist Popularity and Longevity 22

3.5 Analysis of Stream Counts 25

3.6 Musical Attributes 28

4. Results 35

4.1 Model Design 35

4.2 Model 1.1 Multiple Regression without artist variable 37

4.3 Model 1.2 Neural Network without artist variable 38

4.4 Initial Model Results 40

4.5 Model 2.1 Multiple Regression with artist variable included 40

4.6 Model 2.2 Neural Network with artist variable included 41

4.7 Model Outcomes, Initial and Revised models 43

4.8 Model 3.1 Multiple Regression with artist groups and musical attributes 44

4.9 Model 3.2 Neural Network with artist groups and musical attributes 45

5. Discussion & Conclusion 46

6. Further Work 48

References 49

Appendix A. Supplementary information Spotify API 50

Appendix B. R Code utilised to expand the dataset 51

List of Figures

Figure 3.1 Mean attribute values for all chart positions. 18
Figure 3.2 Artist proliferation – full dataset. 19
Figure 3.3 Artist proliferation – Top 1 position artists. 20
Figure 3.4 Boxplot: count of days in dataset – Top 1 tracks. 23
Figure 3.5 Boxplot of Maximum Daily Stream Counts. 25
Figure 3.6 Percentage of Daily Stream Counts 2021-10-15. 26
Figure 3.7 Word Cloud 2019-12-15. 26
Figure 3.8 Percentage of Daily Stream Counts 2019-12-15. 27
Figure 3.9 Top 1 % of daily Stream Counts over time. 28
Figure 3.10 Grouped danceability over time. 28
Figure 3.11 Grouped energy over time. 29
Figure 3.12 Grouped valence over time. 29
Figure 3.13 Grouped tempo over time. 30
Figure 3.14 Grouped loudness over time. 30
Figure 3.15 Grouped key over time. 31
Figure 3.16 Grouped speechiness over time. 31
Figure 3.17 Grouped acousticness over time. 32
Figure 3.18 Grouped instrumentalness over time. 32
Figure 3.19 Grouped liveness over time. 33
Figure 3.20 Grouped liveness over time. 33
Figure 4.1 Distribution of Artist count, full dataset. 36
Figure 4.2 Model 1.1 musical attributes only, streams response on the reduced dataset. 38
Figure 4.3 Neural Network Model 1.2 musical attributes only, reduced dataset. 39
Figure 4.4 Multiple Regression Model 2.1 musical attributes and artist, reduced dataset. 40
Figure 4.5 Neural Network Model 2.2 musical attributes and artist, reduced dataset. 42
Figure 4.6 Neural Network Model 2.2 liveness and key analysis for 100 most accurate predictions and complete model dataset. 43
Figure 4.7 Multiple Regression Model 3.1 musical attributes and artist groups, reduced dataset. 44
Figure 4.8 Neural Network Model 3.2 musical attributes and artist groups, reduced dataset. 45
Figure 4.9 Neural Network Model 3.2 using with only artist groups, reduced dataset. 46

List of Tables

Table 3.1 Variable descriptions basic Spotify daily chart data. 14
Table 3.2 Summary of basic Spotify daily chart data. 14
Table 3.3 Summary of Spotify daily chart data rank 1 position. 14
Table 3.4 Artist track summary basic Spotify daily chart data. 15
Table 3.5 Variable Descriptions for the Expanded Spotify daily chart data. (See footnote*) 16
Table 3.6 Summary of musical attributes, Expanded Dataset vs Filtered Dataset. 17
Table 3.7 Correlation analysis for musical attributes. 18
Table 3.8 Artist Details for Total Stream Counts and Top 1 Positions Achieved. 20
Table 3.9 Top 10 Count of days for which artist had more than one track in Daily Top 200 Chart. 21
Table 3.10 Details of Extreme count of tracks in Daily Chart. 21
Table 3.11 Minimum and Maximum count of days in dataset of Top 1 tracks. 22
Table 3.12 Top 10 Artists – Total streams. 24

Table 3.13 Top 10 artists count and percentage of Top 1 positions.... 24
Table 4.1 Predictor screening results for rank and streams variables using the Top 5% of dataset.... 35
Table 4.2 Predictor Screening Results for rank and streams variables on the entire dataset with artist omitted. 36
Table 4.3 Predictor Screening Results for rank and streams variables on the reduced dataset with artist omitted. 37
Table 4.4 Most accurate predictions of streams count Model 1.1.... 38
Table 4.5 Most accurate predictions Model 1.2 39
Table 4.6 Most accurate predictions Model 2.1 41
Table 4.7 Correlation of prediction error to musical attributes Neural Network Model 2.1 42
Table 4.8 Artist rank recoding divisions. 43

Nomenclature

Specific names are used for the variables defined in the Spotify dataset and this thesis. The variable names have been defined by Spotify and have not been edited as part of this work. All variables are shown as un-quoted, italicised in font Courier-New. e.g. *track, artist, speechiness, acousticness, valence*. Descriptive passages in this document that are referring to the effects of named variables related to overarching concepts may use the same word in different contexts. E.g., The following sentence is concerned with the relationships between the artist (as a noun) and the named variables. “The dataset was analysed to determine if there was correlation between an artist’s success and the principle musical attributes of *danceability*, *tempo* and *valence*. “ Artists in this context is distinct from the variable *artist* in the dataset.

An exception to this text format is present in the preliminary “Basic Dataset” discussed in Chapter 3.1 created from the downloaded csv files. Five of the six variables in this basic dataset are capitalised: “Artist”, “Track Name”, “Position”, “URL”, “Streams”. Usage of the Spotify Application Programming Interface (API) to expand this simple dataset creates and expanded dataset which uses the lowercase variables outlined above. Full details of the dataset structure are given in table 3.5 in chapter 3.2.

Abstract

Spotify AB is currently the world's largest provider of music streaming services. At the time of this work the Spotify music library contains in excess of 82 million music tracks and there are 406 million users, of whom 180 million are paid subscribers. Since its launch in 2008, the service has been credited as having a disruptive influence on the music industry, allowing listeners access to a greater range of music than previously available on radio or in record stores. As a result of this user base, the service has been described as "democratising" music, allowing users to freely choose which tracks they listen to, as opposed to having curated music streamed to them. Listener data is gathered on an enormous scale allowing for near instantaneous reporting of listener preferences and real-time monitoring of the historical performance data of every track. Beginning in 2015 Spotify have promoted an open data model, publishing a daily chart at a global and regional level while allowing access to their datasets for third-party analysis. This open data model has encouraged research into the mechanisms and relationships that exist within the Spotify dataset.

Existing research largely explores the effects of musical attributes, (tone, key, tempo, genre, among others), as being primary influencers on chart position. This thesis proposes that the performing artist may be a stronger influence on chart position than the musical attributes alone. Namely that artists that achieve a high level of fame / success will in turn have more success, independently of the actual attributes of their released tracks. Data exploration determined that there were no significant relationships established between chart position and musical attributes at an individual track or artist level. It was similarly determined that changes in the listener patterns over the four-year observation period suggested that specific types of tracks tend to perform better, particularly where they are closely associated with the trending musical attributes. The results also indicate that the best performing tracks, measured by `rank` = 1, are achieved by only 42 artists from a dataset of 1621 artists. There is little evidence to define these artists, or their musical tracks as significantly different from many of the other artists in the dataset. This finding influences the models as the historical performance data of artists and the overall chart trend towards specific types of music is seen to be a major influence on chart position.

1. Introduction

This Thesis aims to explore the relationships between the musical attributes and artists for music tracks along with the performance (`rank`) of those tracks. The source data is the Spotify “Top 200” daily chart covering a four-year period from 2018-01-01 to 2021-12-31, containing 292,200 observations. This work will explore, by way of analysis, if musical attributes can be used as predictors of chart performance, or if the performing artist(s) have a stronger influence on chart performance, irrespective of the musical attributes of the tracks. Furthermore, using the historical data, a performance profile of track attributes against the peak “chart position” achieved by those tracks is developed. This thesis examines if this historical profile(s) can then be used to predict a subsequent `rank` for a track once the input parameters of track attributes and artist performance data are entered. There are two related measurements of track performance, namely the daily `rank`, measured 200 to 1, and the calculated daily `streams` count measured typically in 1,000s of `streams`. Spotify `rank` is an inverse scale, where `rank` 1 indicates the best performing `track`, by `streams` count. Both measurements are explored in this work, predictive methods are utilised to attempt a robust `rank` prediction and an estimate of `streams` count.

Analysis of the dataset has been leveraged to give an understanding of the data relationships between the music tracks and the success achieved. This thesis addresses problems with data quality, namely replication of data on a large scale, and the presence of extreme outlier events. Data processing techniques such as recoding, binning, text analysis, and data transformations are used to prepare the dataset for subsequent analysis and act as inputs to the final models, resulting in a predictive neural network model capable of predicting `rank` with an accuracy of 85%.

Thus, this thesis is both a data analysis of the Spotify dataset and an investigation of predictive chart performance.

2. Literature Review

This chapter outlines a review of existing literature and previous research concerned with the areas of musical attribute data, classification, big data analysis and predictive methods, based on Spotify data sources. A search for research papers, conference papers and scholarly articles was conducted on the primary sources: Emerald, Research Gate, Science Direct, Google Scholar and Sci-Hub in December 2021/January 2022, for papers published after 2015. The lower bound of 2015 was selected as prior to this year the Global user base of Spotify was not sufficiently large to generate large datasets, after this time Spotify growth accelerated to the current levels which increased data availability. Additionally, a search for earlier papers yielded results that were focused on the technological or commercial aspects of Spotify as a service, rather than on analysis of the musical data. An additional contributor to the absence of research was that Spotify did not openly publish their dataset prior to 2015, while some data was released in 2013, there was not the availability of data to encourage or promote significant research.

Four initial questions were addressed in this study:

1. Is there a body of research established in the area of Spotify listener data?
2. What are the main predictors identified within the Spotify dataset variables?
3. Can the contributing variables to track position or streams count be refined or categorised?
4. Is a robust method of prediction proposed or likely based on the variables in the dataset?

Suitable papers were searched for on each of the databases, papers that failed, or inadequately answered questions 2, 3 & 4 were subsequently excluded. This stage involved the reading of the full text, or as much as was required to form an opinion on its eligibility. In several cases, additional papers were cited, where specific papers were cited on more than one occasion the original paper was accessed (if available) and evaluated to determine relevance to this thesis. Each of the remaining papers was read in full, the key points were noted. These points, and the papers were categorised into the following broad divisions.

- Do the papers conclude a link between musical attribute and track performance?
- Do the papers conclude a link between the artist and track performance?
- Do the papers discuss a measurement methodology to attribute influence?
- Do the papers propose methods or techniques to predict performance?

The final selection of papers forms the basis of this review.

2.1 Musical Attribute and Track Performance

Understanding the musical attributes and their interactions within the dataset is an important step in determining any possible relationships between chart rank and stream counts. Identification of predictor influence, independently and as part of correlated variable is a key part of much of the research. Much effort has been undertaken by existing research in attempting to create a descriptive method for music that can accommodate the personalised nature of music, (Han *et al*, 2018). The usage of basic numerical systems for describing each attribute independently ignores what researchers' term "*enactive representations*", whereby the listening experience is not a stand-alone event. Listeners are affected by external influences such as mood, historical (nostalgic) events and recommendations which affect their choices.

Lee and Lee (2018) examined the feasibility of prediction based on the audio signal characteristics of a track, but not a single value as used in the Spotify dataset. Ni *et al* (2015), have noted musical trends and correlation between the specific attributes of "loudness, duration and harmonic simplicity", text analysis of track lyrics has also been proposed by Dhanraj & Logan (2005). Spotify use a proprietary "popularity" algorithm for their own recommendation engine; research efforts to emulate this have shown moderate success, such as by Berger (2017). The overall sentiment in the literature is that the unique data structures in music and external influences not reflected in the data could lead to biased predictions (Sciandra & Spera, 2019) and that the independent simplified musical attributes, as reported by Spotify, do not lend themselves to use as predictors.

2.2 Artist and Track Performance

The influence of a musical artist on chart success is critically evaluated by Aguiar & Waldfogel (2018). This research focused on promotion by record labels, as well as social media influencers of artists and their inclusion (or exclusion) from the Spotify "Discovery" and "New Music Friday" playlists. The findings would suggest that artists (or tracks) that are promoted by the playlists tend to achieve large stream counts and high daily rankings but tend to fall in both metrics once no longer promoted in the recommendation playlists. The analysis determines that the performance of the tracks is driven more by the promotion than by the track attributes. The paper further proposes an amplification effect, whereby the tracks that are promoted have already achieved "some" success, they are not completely unknown. The inclusion on the recommendation playlists simply promotes the tracks to a wider audience, the promotion is based on an existing performance metric. While this doesn't exclude the impact of the musical attributes on the initial success, it creates a supercategory for promoted artists that drives further success, the musical attributes are not considered in the exposure of these Artists and tracks.

This independence of the track attributes from the promotion playlists may mean that many millions of listeners may not like or would not ordinarily listen to the track if it were not promoted, rather the listener passively accepts the recommendation and listens to a curated playlist as opposed to actively selecting a track or tracks. (Aguiar & Waldfogel, 2018). These promotional relationships can then affect the dataset, leading in turn to misclassification of the Artist, attribute and performance measurements by including the measures of daily streams from Spotify generated and curated playlists with independently selected stream from listener selection. The concern of the research, as evidenced by Aguiar & Waldfogel is that the specific promotion of tracks skews the data in favour of those promoted tracks, at the expense of genuine success for tracks independently chosen by the listeners. This research brings into focus the independence of the Spotify data, where a significant influence on the listener data is the recommendation mechanism.

2.3 Attribute Measurement

The treatment of the musical attributes is discussed by various sources, the topics of therapeutic effects of music for mood management during the Covid19 Pandemic is researched by Kalustian & Ruth (2021). Similarly, the topic of correlations between attributes and time sensitive diurnal listening patterns is researched by Heggli *et al*. (2021). These studies focus on the emotive aspects of the attributes, specifically valence; defined as a measure of how happy or sad a song is, and energy; defined as a perceptual measure of intensity and activity, as the key variables. The studies determine that music with specific characterises may act as a healing influence. Other research focuses on the musical form of the variables; for example, tempo, rhythm and time signature are evaluated by Savelsberg (2020), while the energetic and danceable characterises are explored as part of trend analysis of the changing types of music, (Sciandra & Spera, 2019).

Much of the literature shares a similar critique of the Spotify attributes, namely that the 0 to 1 range of the attributes average flattens the music, which makes meaningful analysis less robust. Furthermore, many of the top performing tracks share the same ranges of key musical attributes, which makes segregation difficult. The difficulty as expressed in the literature is that music is dynamic, with variation existing for each of the musical attributes over the duration of the track. Reducing this complex, multi-attribute and track specific dynamic range to a set of average values loses the essence of the track. Han (2021) explores in detail the relationships between variables. While the paper focuses on the variables of “liveness”, “Danceability”, “Property” and “Popularity” the utilised method of exploring interdependence is interesting and may be in-scope for this project. The understanding that the attribute variables do not act independently is significant. In this context the research explores how tracks with high tempo tend to be more danceable which in-turn give a higher associative energy, thus the correlation between tempo and danceability and energy is established to give an overall sense of the track which is of greater significance than the sum of the individual component attributes.

2.4 Prediction

Due to the availability and scale of Spotify datasets there is considerable variety in predictive methods across the literature. Linear Regression (Bonet *et al.*, (2014), Generalised Linear Mixed Models (Sciandra & Spera, 2019), Random Forest, K-Nearest Neighbour and Linear Support Vector Classifier (Pareek *et al.*, 2022) are all documented as existing methods of prediction, with varying degrees of success. Of importance in all papers is that the core Spotify attribute data, in the 0 to 1 range, was not utilised. Data manipulation, recoding and the creation of classifications were all deemed to be required to achieve any measure of accuracy. Principal Component Analysis (PCA) is leveraged by South *et al.* (2021), to create classifiers using the eigen values as measurements of the covariance of the attributes while centrality values evaluate how related the attributes influences are to each other, to further refine the predictor inputs. Adaption of existing models in unrelated fields have achieve results of up to 83% accuracy (70 % recall) as recorded by Pareek *et al.* (2022). This research cites previous work by Poorna & Vrushsen (2020), as the foundation for a classification algorithm utilising pattern recognition, feed forward and cascade forward neural networks developed as part of research into global carbon emission.

No standout method has been identified, there are problems identified with all methods and results are varied, depending heavily on the input data, the output target and the depth of manipulation conducted, while avoiding overfitting or producing impractical models. Ingle *et al.* (2021) highlights the inherent problems caused by noise in the dataset due to the inclusion of so many repetitive entries for tracks at different positions. For tracks that gradually move up the chart to a high position the values are assessed for static musical attributes with a varying rank performance attribute, this feature of the dataset, applicable to every track that is repeated, makes any prediction more difficult. There are, it would seem, no effective predictor variables if the musical attribute values never change in any way with the dynamic chart position variable.

The conclusion of this research sets the tone for what is likely to be a significant element of the analysis and modelling herein, successful artists are likely to achieve further success and the performance of a track will not "…depend just on how happy, energic, danceable or loud your song is, but more likely it would be related to your current popularity as an artist.", Ingle *et al.* (2021).

2.5 Literature Review Conclusion

The critical outcome of this review is that the analysis of the dataset must consider the possibility that the use of simple attribute values, rather than complex audio signal analysis may reduce the musical attributes to common levels for tracks that are actually very different. Additional efforts may be required to clarify or categorise the artists or tracks into classes to aid in both the analysis and modelling. As researched by Han (2021) the combining of musical attribute variables into classifiers may be of use in modelling, similar classification of the *artist* and *rank* variables, or combinations of *artist* and musical attributes may also be necessary. It is also possible that some bias already exists in the dataset, most likely for the most successful tracks or artist.

Any modelling effort is expected to contain some element of recoding or classification to achieve reasonable accuracy results, and that because of the potential super-influence of the artist, and the unknown effects of external influences, such as the recommendation playlists identified by Aguiar & Waldfogel (2018), the results are likely to not translate into accurate models for blind data. Reasonable care must therefore be taken to not overfit models by creating classifications, or combinations of attributes into derived variables that reinforce any possible predictors present in the dataset.

3. Data source

3.1 Basic Dataset

The dataset for this work is available directly from Spotify servers using freely accessible developer accounts and an Application Programming Interface (API). The data source was created in January 2022 by direct download of a csv file which was subsequently manipulated in R-Studio using API calls. In order to reduce the download size Spotify utilises a primary key for the *URL* variable. This initial dataset is not of sufficient detail to allow for detailed musical attribute analysis or effect exploration, the usage of the API with the *URL* key is required to expand the dataset to include a variety of desirable attributes, that can be tailored to the specific areas of research interest. The core data from Spotify contains the six variables, as shown in Table 3.1**.**

Table 3.1 Variable descriptions basic Spotify daily chart data.

Variable	Type	Description
Track Name	Nominal	Track Name.
Position	Continuous	Daily Chart Position range 200 to 1 (inverse, #1 is top position).
date	Nominal	Date of daily Top 200 chart results.
Artist	Nominal	Artist / Artists Names.
URL	N/A	Hyperlink to Spotify servers, used to access detailed track information.
Streams	Continuous	Count of daily streams per track.

The basic dataset contains 200 observations per day of the six variables from Table 3.1 for each date beginning on 2018-01-01 and concluding on 2021-12-31. As 2020 was a leap year the total number of observations in the dataset is: 200 x (365 + 365 +366 + 365) = 292,200. Analysis found that there were no missing values and no anomalies in the data structures or formats, however it did indicate that there was a significant volume of repetition within the dataset, particularly for the Artist and Track Name variables. Table 3.2 contains summary information on the basic dataset. Additional analysis on foot of these findings indicated that of the available Top 1 positions in the dataset the top ranked position was achieved by only 43 Artists and 62 unique tracks, shown in Table 3.3.

Table 3.2 Summary of basic Spotify daily chart data.

Count DISTINCT Track Name	Position Range	date range		Count DISTINCT Artist	Count DISTINCT URL	Minimum Streams	Streams Range
4,622	1 - 200	2018-01-01	2021-12-31	1,621	4,622	2,947	2947 - 319678

Table 3.3 Summary of Spotify daily chart data rank 1 position.

Count DISTINCT Artist	Count DISTINCT Track Name	Minimum Streams	Maximum Streams
43	62	31045	319678

These findings would suggest that there is a low variability in the artists and tracks that achieve the top ranked position, which may extend to the overall dataset. Additional data exploration, expanded to other chart positions is discussed later in this work in chapters 3.4 to 3.6 inclusive.

The URL variable in the basic dataset is a primary key, unique for each of the 82 million tracks in the Spotify music library. The Spotify API was subsequently utilised to request detailed track musical attributes and historical track information to populate additional variables. R-Studio was used to handle the API requests to Spotify servers. For each distinct URL, a “get audio-features” request was processed, the returned data was parsed to a new data frame, which was subsequently joined with the original dataset on the common URL key. The resulting dataset contained the musical attributes for each track, while maintaining the date, artist and track positions from the original dataset. Details of the Spotify API “get audio-features” and the r code is provided in Appendix A and B.

Examination of the basic data also identifies that many artists are prolific, having multiple tracks in the dataset. Of the 1,621 artists in the dataset, 490 artists have more than one track present. Furthermore 102 of the artists have more than ten tracks in the observations and accounts for 50% of all the tracks in the dataset, i.e., 6% of all artists are responsible for 50% of all tracks. Table 3.4 contains details of the five most prolific Artists in the dataset, along with summary information for the entire dataset.

Table 3.4 Artist track summary basic Spotify daily chart data.

Most Prolific Artists			**Dataset Summary**	
Artist	**Count Tracks**		Mean Tracks	3
Taylor_Swift	139		Min Tracks	1
Drake	71		Max Tracks	139
Ed_Sheeran	65		Mode Tracks	1
Juice_WRLD	64			
Ariana_Grande	59			

The basic dataset lacks any of the details required to fully understand the track - position relationship outside of that established by the artist. In order to fully explore the data, with a view to determining and understanding more complex relationships, it is necessary to expand the dataset to include the musical attributes for each track. This data is also available from Spotify using their API.

3.2 Expanded Dataset

Table 3.5 Variable Descriptions for the Expanded Spotify daily chart data. (See footnote*)

Variable	Type	Description
title	Nominal	Track Name.
rank	Continuous	Chart Position (inverse, #1 is top position).
date	Nominal	Date of daily Top 200 chart results.
artist	Nominal	Artist / Artists Names.
streams	Continuous	Count of daily streams per track.
Trend	Categorical	Change in rank relative to previous rank (Up, Down, Same, New Entry).
danceability	Continuous	Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
energy	Continuous	Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.
key	Continuous	The key the track is in. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on. If no key was detected, the value is -1.
loudness	Continuous	The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typically range between -60 and 0 db.
speechiness	Continuous	Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.
acousticness	Continuous	A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
instrumentalness	Continuous	Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.
liveness	Continuous	Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.
valence	Continuous	A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).
tempo	Continuous	The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.
duration_ms	Continuous	The duration of the track in milliseconds.

***Footnote: Additional variables created with the Spotify API are identified as the shaded rows. All descriptions as per Spotify developer documentation detailed in Appendix A.**

Table 3.5 represents an overview of the expanded dataset. The additional variables created using the API calls have been appended to the basic data, identified by shaded rows. The dependant variables of interest for this project are `streams` (count) and `rank` (chart position). The core predictor variables are those that can be described as musical attributes; `danceability`, `energy`, `key`, `loudness`, `speechiness`, `acousticness`, `instrumentalness`, `liveness`, `valence` and `tempo`. All subsequent discussions regarding data analysis and predictive modelling will refer to this expanded dataset or defined sub-sets thereof. Summary statistics of the musical attribute variables and a brief analysis of each is found given in Table 3.6. As previously mentioned, the volume of repetition in the dataset is significant, particularly with tracks and artists that perform well. Any summary statistic on the full dataset may be impacted by tracks that have multiple entries in the dataset, which may skew the results and subsequent interpretation. This effect is illustrated in Table 3.6, where the summary information for all musical attributes is shown for both the whole dataset and a filtered sub-set containing only a single occurrence of each distinct track.

Table 3.6 Summary of musical attributes, Expanded Dataset vs Filtered Dataset.

Attribute	danceability		energy		key		loudness		speechiness	
Statistics	Full dataset	Distinct tracks	Full dataset	Distinct tracks	Full dataset	Distinct tracks	Full dataset	Distinct tracks	Full dataset	Distinct tracks
Mean	0.658	0.641	0.634	0.623	5.328	5.164	-6.48	-6.915	0.125	0.168
Std Dev	0.145	0.154	0.175	0.185	3.625	3.664	2.502	2.865	0.179	0.246
Min	0.073	0.073	0.005	0.005	0	0	-34.475	-34.475	0	0
Max	0.975	0.975	0.993	0.993	11	11	0.175	0.175	1	1
Median	0.676	0.654	0.653	0.64	6	5	-6.062	-6.382	0.059	0.062

Attribute	acousticness		instrumentalness		liveness		valence		tempo	
Statistics	Full dataset	Distinct tracks	Full dataset	Distinct tracks	Full dataset	Distinct tracks	Full dataset	Distinct tracks	Full dataset	Distinct tracks
Mean	0.236	0.232	0.021	0.043	0.161	0.164	0.478	0.455	114.497	110.22
Std Dev	0.256	0.263	0.099	0.153	0.128	0.141	0.228	0.237	35.207	43.665
Min	0	0	0	0	0	0	0.026	0.026	0.035	0.035
Max	0.994	0.994	0.989	0.989	0.989	0.989	0.982	0.982	207.975	207.975
Median	0.14	0.119	0	0	0.114	0.116	0.476	0.449	118.016	116.942

This exploration demonstrates that the full dataset may be influenced by the presence of multiple observations, of the same track attributes. The evidence, while subtle, is in the changes in the key statistics of mean and standard deviation for each of the variables, when including the replicated tracks from the full dataset. Of the musical attributes in Table 3.6 there is evidence at this early stage, indicated with shaded cells, to suggest that highly prevalent tracks tend to be of increased values in *danceability*, *energy*, *key*, *valence*, *tempo* and *acousticness*, while simultaneously having lower values in *loudness*, *speechiness*, *instrumentalness* and *liveness*. While this may seem conflicting, many of these variables are inversely related. Tracks that are danceable would not ordinarily be acoustic and would be expected have a low *speechiness* and have a high *tempo*. Correlation analysis on the dataset reinforces this understanding, the result of the which is shown in Table 3.7.

Table 3.7 Correlation analysis for musical attributes.

	danceability	energy	key	loudness	speechiness	acousticness	instrumentalness	liveness	valence	tempo
danceability	1.000									
energy	0.160	1.000								
key	0.032	0.032	1.000							
loudness	0.211	0.725	- 0.002	1.000						
speechiness	0.045	-0.015	- 0.027	-0.046	1.000					
acousticness	-0.247	-0.591	0.007	-0.498	-0.165	1.000				
instrumentalness	-0.147	-0.152	- 0.032	-0.204	0.355	-0.067	1.000			
liveness	-0.062	0.122	0.019	0.035	-0.188	0.013	-0.155	1.000		
valence	0.329	0.330	0.063	0.195	-0.223	-0.084	-0.243	0.153	1.000	
tempo	0.012	0.082	0.017	0.064	-0.528	0.078	-0.384	0.257	0.314	1.000

From the correlation table it is evident that the attributes of danceability, energy, key, loudness, valence, and to a lesser extent tempo are related, while the attributes of speechiness, acousticness, instrumentalness and liveness are similarly related but opposing. At a simple level, it is possible to describe the data as belonging to two main classes: danceable energetic uplifting tracks or slow acoustic instrumental tracks. Figure 3.1 shows the change in the mean musical attributes for danceability, energy and valence over all chart positions. While the changes are minor, there is some evidence that tracks further from the Top 1 position are less danceable, less energetic and less uplifting. However, the question remains if these differences in attributes are sufficient to be used as predictors. The indications at this point are that the ranges of the musical attributes that achieve the top positions are contained within the overall ranges for tracks that do not achieve top positions, so there may be little to define or identify those tracks that perform well from the larger body of track attribute data.

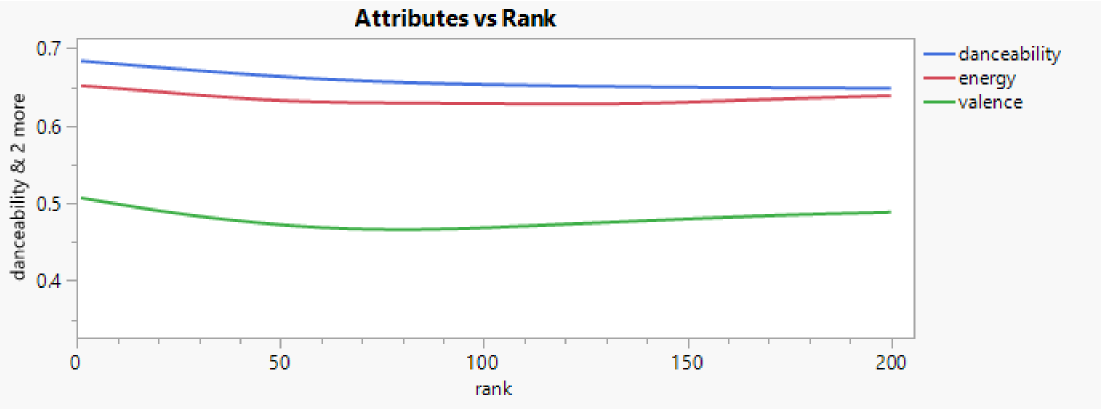


Figure 3.1 Mean attribute values for all chart positions.

3.3 Data Exploration

From the findings of section 3.2, it was determined that the `artist` variable may be a significant contributor to chart position and rank. As the dataset contains categorical data, in the artist's name, it would make sense to explore the large dataset in a manner that is visual rather than simply quantitative. While summary tables of the quantitative data and distribution plots of that data are valid, a pictorial representation of the artist popularity can be instantly informative, while aiding discovery and insight into the data relationships. Figure 3.2 represents all artists from the full dataset, with `artist` name scaled by count of appearances in the dataset, irrespective of positions achieved. Applying text analysis in this manner to the count by total `streams` of all `artist` variables clearly shows not only the most popular artist, but given their sheer magnitude informs the viewer of how prevalent these top performing artists are.

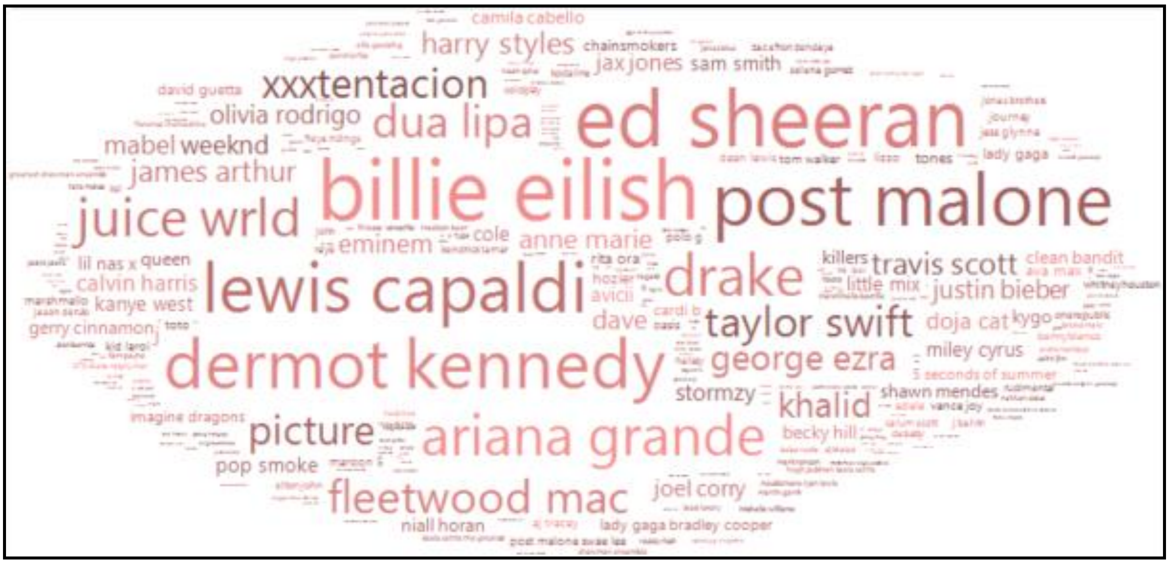


Figure 3.2 Artist proliferation – full dataset.

When the same text analysis is repeated in Figure 3.3 with the addition of a filter for artists that have achieved the top position, or `rank` = 1, it is clear to see that many of the names that were prominent in the Figure 3.2 are absent; Billie Eilis, Post Malone and Dermot Kennedy are noticeable examples of this. The increase in relative scale of Ariana Grande, Drake and Olivia Rodrigo is also evident.

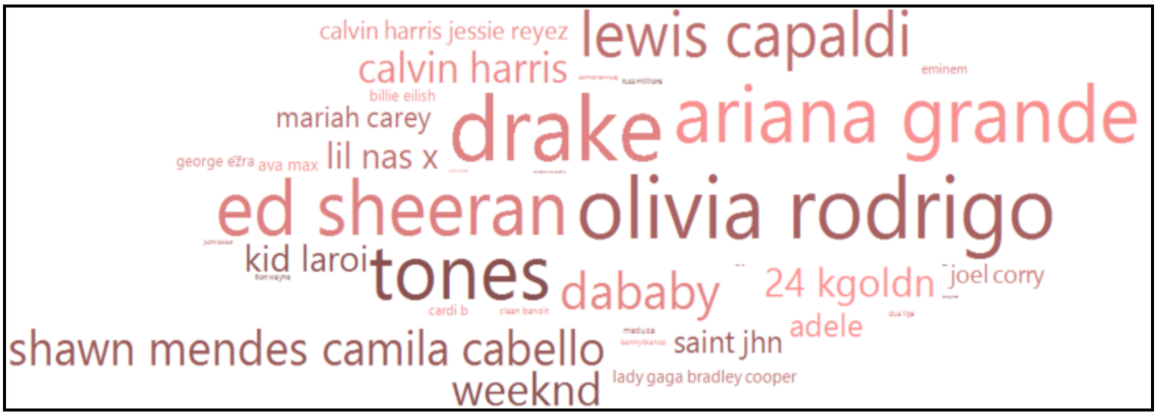


Figure 3.3 Artist proliferation – Top 1 position artists.

From this simple text based visual analysis on an otherwise cluttered dataset, it is evident that the most popular artists, as measured by total streams, may not achieve the top position often, or at all. As the Spotify chart data is so dynamic, being a daily record of the Top 200 `streams`, it may be premature to oversimplify the relationship between artist popularity on a daily increment, which one would expect to lead to a Top 1 position, and the overall actual popularity of the artists.

It is possible that artists may be prolific, having multiple tracks in the dataset and having multiple tracks in the Top 200 on any given day. The combined popularity of the tracks, as measured by total `streams` may overestimate the artists popularity, relative to the singularly most popular individual track on any given day. Extending this total popularity measure over the entire dataset leads to the types of word clouds seen in Figure 3.2 with prominence given to artists who rarely or never achieve a Top 1 position. It may be that artists with a high total `streams` count have achieved that total via multiple contributing tracks, none of which are streamed in a high enough quantity daily to achieve the Top 1 spot. A case in point is for the artist Post Malone relative to two other high performing artists. This artist is noticeably absent from Figure 3.3, having not achieved a Top1 position. Analysis shows that while ranked #6 in total **streams** and appearing 6,022 times in the dataset, this artist has never achieved the Top 1 position, despite achieving a Top 10 position 399 times and a Top 5 position on 205 occasions. Table 3.8 shows Post Malone's overall ranking, based on total `streams` along with the additional artists including Top 1 position data for comparison.

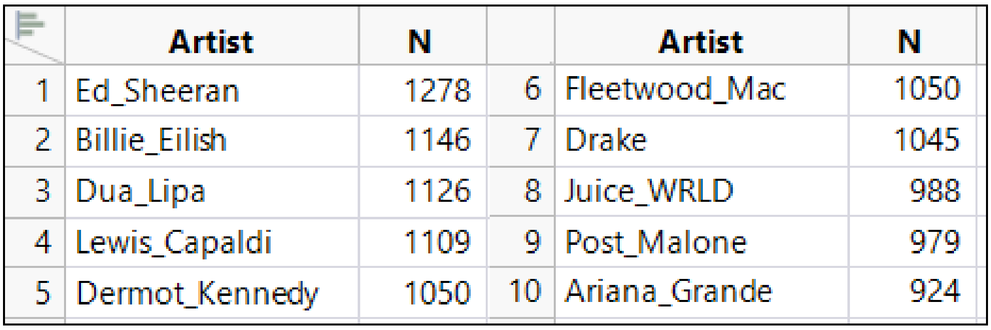
Table 3.8 Artist Details for Total Stream Counts and Top 1 Positions Achieved.

Artist	Appearances	Total Streams	Dataset Rank (Total Streams)	% of Total Streams	Count of Top 1 Positions
Lewis Capaldi	5,613	88,025,705	1	2.62%	81
Post Malone	6,022	74,720,678	6	2.22%	0
Olivia Rodrigo	2,190	52,714,896	9	1.57%	112

As can be seen in Table 3. 8 Post Malone with a value of 0 does not feature in the Top 1 result despite being the artist with the highest appearance count in the dataset, achieving the sixth highest total `streams` count and a percentage of total streams that is surpassed by only the top 0.3% of all artists in the dataset. Furthermore, it can be seen that Olivia Rodrigo has achieved a significant rate of success in terms of Top 1 tracks, despite having the lowest appearance count of the three artists and the lowest total `streams` count from Table 3.8. While Lewis Capaldi is the most streamed artist in the dataset this artist also has a lower value for the number of days where he was in the Top 1 position. A point of note is that the artists in Table 3.8 are all within the top 0.5% of artists when measured by `streams` count, yet clearly this isn't sufficient to consistently achieve a Top 1 position.

Given that the count of days in the dataset is 1461, all these artists have more than one song in the Top 200 on more than 1 day. Table 3.9 shows a sample of the data, grouped by artist for the number of days where the artist had multiple tracks in the daily chart.

Table 3.9 Top 10 Count of days for which artist had more than one track in Daily Top 200 Chart.

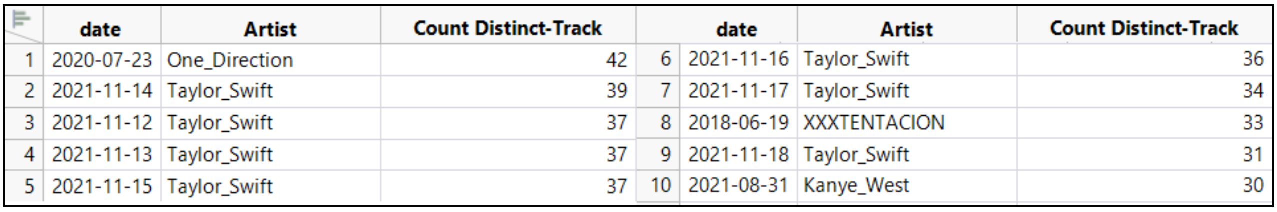


	Artist	N		Artist	N
1	Ed_Sheeran	1278	6	Fleetwood_Mac	1050
2	Billie_Eilish	1146	7	Drake	1045
3	Dua_Lipa	1126	8	Juice_WRLD	988
4	Lewis_Capaldi	1109	9	Post_Malone	979
5	Dermot_Kennedy	1050	10	Ariana_Grande	924

***Footnote: N represent the number of days from the total of 1461 where the Artist had at least 2 tracks in the daily Top 200 chart.**

While not a measure of the Top 1 performance this count of the number of days where an artist had more than one track in the daily Top200 chart is a valuable indicator of how popular an artist is over the longer period. This may prove to be a key feature of input into subsequent performance modelling. An extreme example of artists with multiple tracks in the daily Top 200 on a given day is illustrated in Table 3.10.

Table 3.10 Details of Extreme count of tracks in Daily Chart.



	date	Artist	Count Distinct-Track		date	Artist	Count Distinct-Track
1	2020-07-23	One_Direction	42	6	2021-11-16	Taylor_Swift	36
2	2021-11-14	Taylor_Swift	39	7	2021-11-17	Taylor_Swift	34
3	2021-11-12	Taylor_Swift	37	8	2018-06-19	XXXTENTACION	33
4	2021-11-13	Taylor_Swift	37	9	2021-11-18	Taylor_Swift	31
5	2021-11-15	Taylor_Swift	37	10	2021-08-31	Kanye_West	30

The largest single count of unique tracks by an artist on any given day in the dataset was 42 tracks by One Direction on 2020-07-23, it should be noted that none of these tracks were the Top 1 track on that day. Taylor Swift had between 31 and 39 tracks in the daily Top 200 for a period covering 2021-11-12 to 2021-11-18. The presence of this type of individual artist effect on the data would of course influence the overall data shape. Artists that have a specific sound or musical style, and who represent a significant percentage of the daily chart may affect the distribution of the musical attribute variables for any affected dates.

An additional driver of the Top 1 position may be external events. As noted in Table 3.10 the count of 42 distinct One Direction songs in the daily chart on 2020-07-23 was exceptional; a quick search for news headlines indicates that this date was the 10th anniversary of One Direction's formation. As Spotify utilises a recommendation engine for users it is highly probable that this stand-out performance by One Direction was strongly influenced by Spotify's recommendation algorithms, rather than actual user preferences.

It is this artist related aspect in the data relationships that present a significant challenge to prediction. The analysis has found that artist proliferation doesn't always correlate with Top 1 success, as shown by Post Malone. Similarly, One Direction are only ranked 421 of 1621 artists by total `streams` which appears at odds with their single day performance. Taylor Swift on the other hand is the twelfth most popular artist in the dataset by total `streams` counts yet has only achieved three Top 1 positions. To capture these relationships, it is necessary to combine summary analysis of the individual artists' total `streams` count, their peak chart positions and their count of number of days present in the dataset. In essence, analysis of the dataset should add to understanding of the artists' body of work as a possible contributor to position and predictions.

3.4 Artist Popularity and Longevity

As artist proliferation has been discussed in terms of the number of tracks by artists present in the dataset a second strand is a measurement of the duration artist and tracks remain in the dataset, and if artist longevity influences overall performance. To investigate further, the dataset was analysed for how many days tracks that achieve the Top 1 position remain in the chart, at any position. A sample of this analysis showing the minimum and maximum values is in Table 3.11 and Figure 3.4.

Table 3.11 Minimum and Maximum count of days in dataset of Top 1 tracks.

Track Title	Count of Days	Max Rank	Min Rank
All Too Well (Taylor's version)	19	9	1
Shotgun	1200	200	1

The data from Table 3.11 shows large variability in the longevity of the Top 1 track in the dataset. The associated boxplot in Figure 3.4 for the analysis clearly shows that titles that having achieved the top 1 position tend to remain in the dataset for a considerable time, with a mean of 451 days. The range of duration, 19 to 1200 days is also of importance, as this likely means that where analysis is conducted at a track level there will be, as has been noted, effects and impacts on the result due to the large-scale repetition in the dataset.

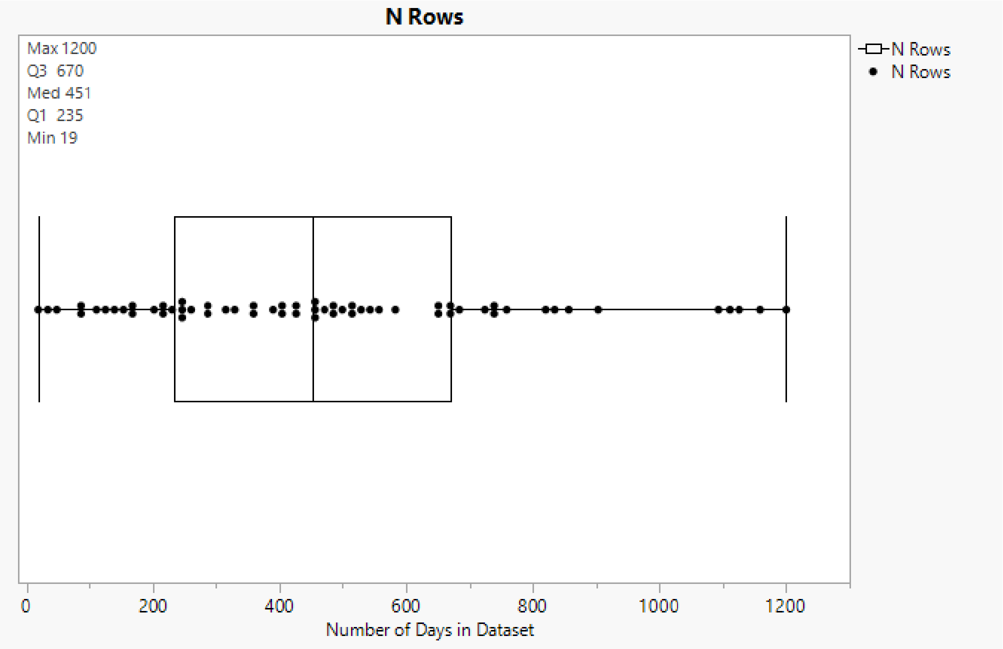
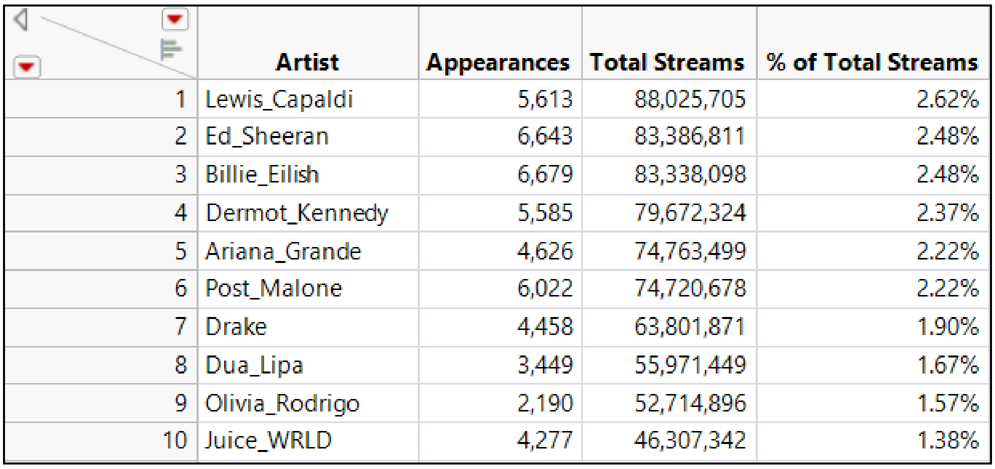


Figure 3.4 Boxplot: count of days in dataset – Top 1 tracks.

Comparative analysis of artists in the measurements of `streams` count and Top 1 positions also gives insight into the possibility that there are additional data relationships. As the daily Top 1 position is determined by the daily `streams` count, it would be expected that those artists with high total `streams` count would have similarly high counts of Top 1 positions. While this assumption is at least partially accurate, in that there is correlation between those artists that have achieved a Top 1 position and their total `streams` count, it excludes the relative relationships between artists and ignores the effect of an artist having multiple tracks in the daily chart on a given day. Similarly, there is also a risk that a successful artist, with a large daily `streams` count may be denied the Top 1 position by a similarly successful artist who happens to release a track on a given day. This interplay between artists, tracks and external effects adds a level of complex variability to the dataset that may not be easily explained, as seen with the One Direction example. Table 3.12 contains the Top 10 performing Artists as measured by total `streams` count in the full dataset.

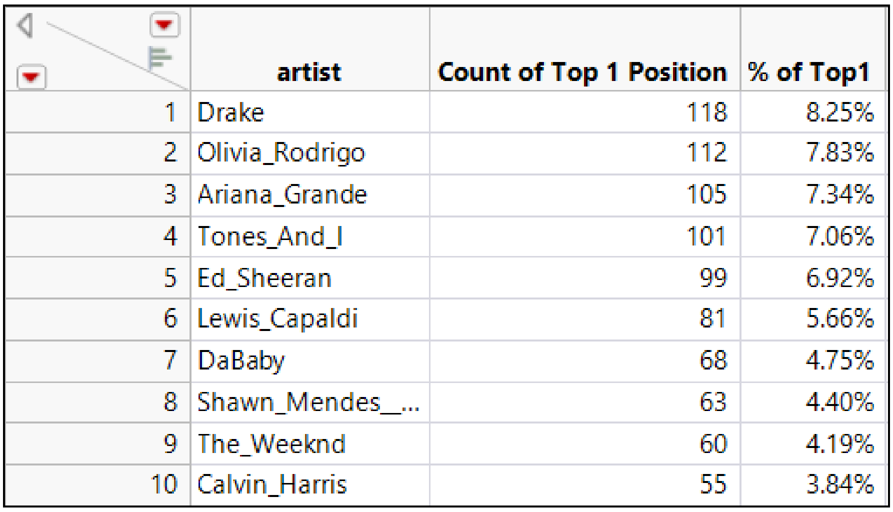
Table 3.12 Top 10 Artists – Total streams.



	Artist	Appearances	Total Streams	% of Total Streams
1	Lewis_Capaldi	5,613	88,025,705	2.62%
2	Ed_Sheeran	6,643	83,386,811	2.48%
3	Billie_Eilish	6,679	83,338,098	2.48%
4	Dermot_Kennedy	5,585	79,672,324	2.37%
5	Ariana_Grande	4,626	74,763,499	2.22%
6	Post_Malone	6,022	74,720,678	2.22%
7	Drake	4,458	63,801,871	1.90%
8	Dua_Lipa	3,449	55,971,449	1.67%
9	Olivia_Rodrigo	2,190	52,714,896	1.57%
10	Juice_WRLD	4,277	46,307,342	1.38%

It is clear from Table 3.12 that Lewis Capaldi would be considered as the “most popular” artist in the dataset based on total `streams`, followed by Ed Sheeran and Billie Eilish. Olivia Rodrigo, in position 9, is 1% “less” popular than Lewis Capaldi based on total `streams`; however, in terms of success in achieving the daily Top 1 position and maintaining those positions over time, Olivia Rodrigo outperforms Lewis Capaldi by over 2% when measured by count of days at the Top 1 position, as shown in Table 3.13.

Table 3.13 Top 10 artists count and percentage of Top 1 positions.



	artist	Count of Top 1 Position	% of Top1
1	Drake	118	8.25%
2	Olivia_Rodrigo	112	7.83%
3	Ariana_Grande	105	7.34%
4	Tones_And_I	101	7.06%
5	Ed_Sheeran	99	6.92%
6	Lewis_Capaldi	81	5.66%
7	DaBaby	68	4.75%
8	Shawn_Mendes_...	63	4.40%
9	The_Weeknd	60	4.19%
10	Calvin_Harris	55	3.84%

Similarly, Drake although appearing in position 6 for total `streams` count in Table 3.12, is the artist with the most days in Top 1 position. The complication for prediction is that measurement of total `streams` count does not equate to Top 1 positions, Drake is “more” successful when measured by Top 1 position while “less” successful in total `streams` count than Lewis Capaldi. It is clear then that while there is a clear measurement of daily popularity, it is less clear over the date-range of the dataset, which may hinder predictive efforts. At a basic level, any artist needs only to be the most successful on any given day as each daily chart is an independent event, however as it becoming evident there are additional factors that can influence this independence.

3.5 Analysis of Stream Counts

The Top 1 daily track is defined as the track having the maximum `streams` count for each day. Analysis of the full dataset shows that the maximum daily values fall within the range of 31,045 to 319,678 daily `streams`, as seen in the boxplot in Figure 3.5 which represents the distribution of the stream counts for only the Top 1 position for the dataset.

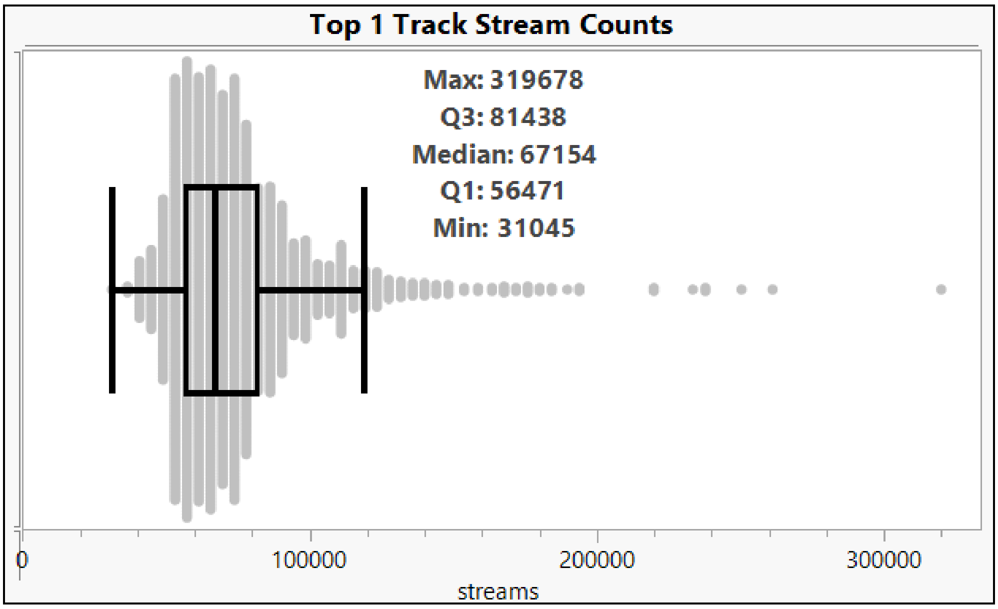


Figure 3.5 Boxplot of Maximum Daily Stream Counts.

In percentage terms, the top 1 tracks account for 1.76% to 10.99% of the total `steams` per day. These values show the incredible variability in "how popular" the Top 1 track can be relative to the remaining tracks in the daily chart. At the upper end of the scale, a track that accounts for an individual 10.99% of the Total daily `streams` can no-doubt be described as popular, whereas at the lower end, a track that achieves the Top 1 position while "only" accounting for 1.76% of the total `streams` is not significantly more popular than its nearest neighbours.

Further analysis of the data values was conducted to investigate if there was a rational explanation for why a single track represented such an extreme upper outlier. The track which accounted for 10.99% of total `streams` on a single day was identified as "Easy on Me" by Adele. This track was released on 2021-10-15, which was the same day the maximum percentage of daily `streams` was recorded. Figure 3.6 shows the distribution of the **streams** data for this particular date, which clearly shows just have much of an outlier this single track was.

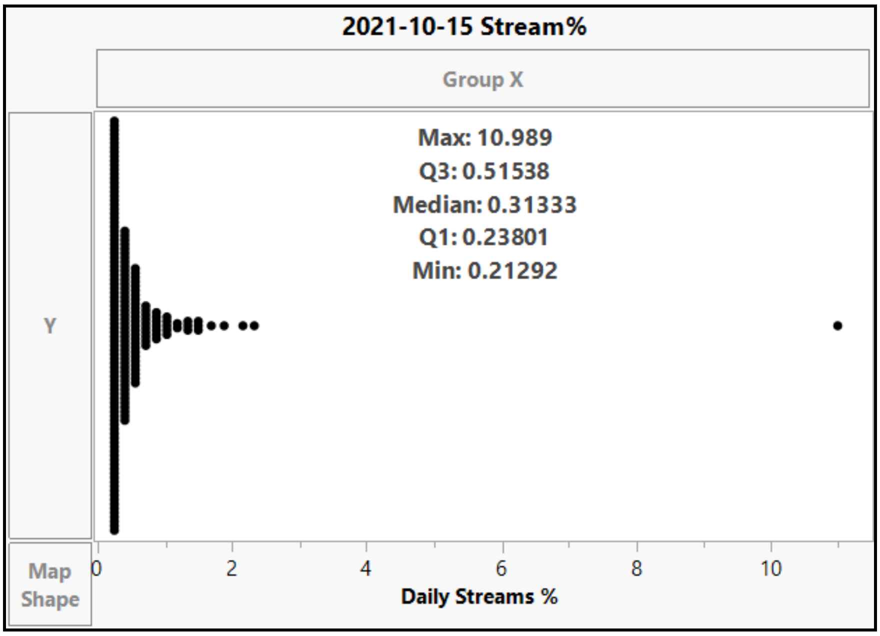


Figure 3.6 Percentage of Daily Stream Counts 2021-10-15.

Repeating the analysis for the track with the lowest value of 1.76% of daily `streams` shows that it occurred on 2019-12-15, for the track “All I Want for Christmas Is You” by Mariah Carey. Additional exploration on the subset for this date indicates that the daily chart was heavily influenced by seasonal themes, with sixty-five tracks classified as “Christmas” tracks, as shown in the word cloud in Figure 3.7. The subset is also less variable, the Top 10 positions are only separated by 0.5% The distribution of daily `streams` percentage for this date is shown in Figure 3.8.

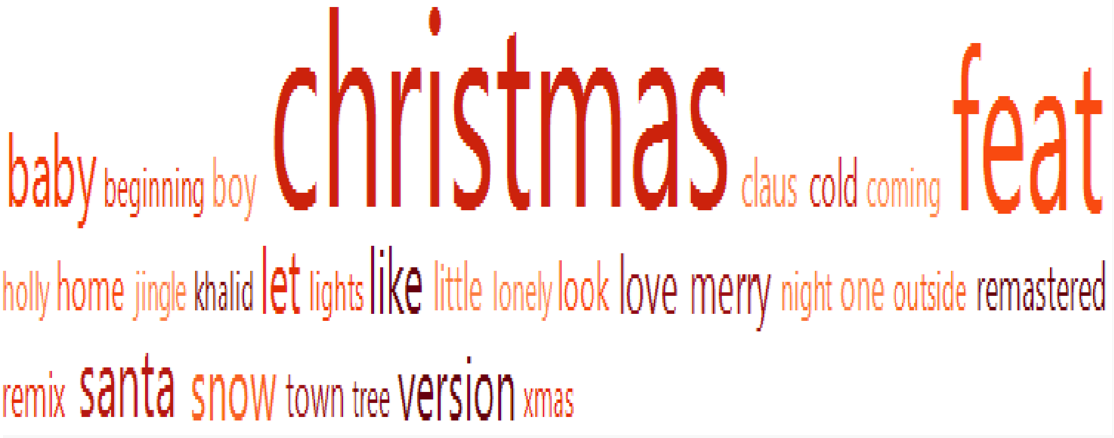


Figure 3.7 Word Cloud 2019-12-15.

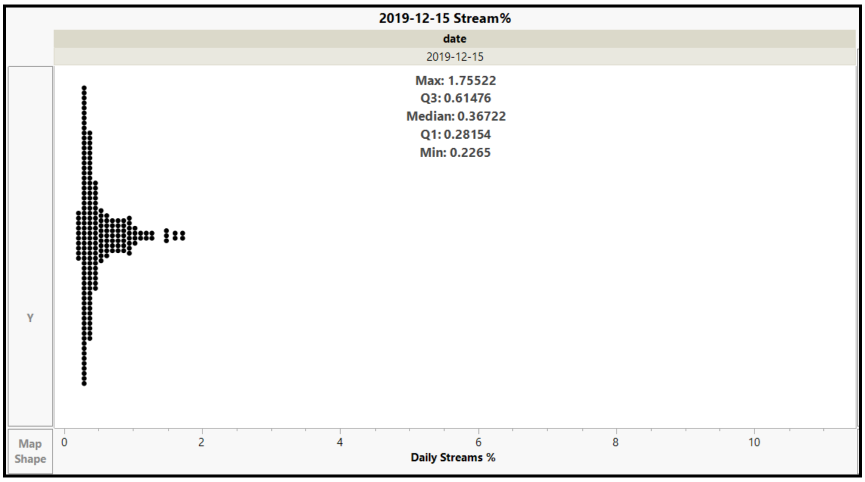


Figure 3.8 Percentage of Daily Stream Counts 2019-12-15.

Analysis of the `streams` count data for the Top 1 position has demonstrated that the range of values within the dataset will further complicate predictions, the lowest value for a Top 1 track of 31,045 recorded for Mariah Carey, would have only been the Top 14 position on the day that Adele achieved the highest value for a Top 1 track of 319,678. On Christmas Day 2021, 74 of the daily 200 tracks had `streams` count higher than 31,045, thus an additional influence in the total number of daily `streams` is identified. With such variation in the `streams` counts for successful tracks, utilising the `streams` count as a predictor would not be practical. To address this problem the percentage of total daily `streams` may be a better measurement. Figure 3.8 shows the percentage of total daily `streams` of the Top 1 position over the entire dataset, with the addition of a smoothing curve and a trend line. The smoothing applied is standardised values, calculated by subtracting the mean and dividing by the standard deviation of the daily variable. This gives a clearer indication of the ranges of values that are achieving the Top 1 position. An observation on the general upward trend in percentage of daily stream counts over time and the number of extreme outliers in 2021 is another interesting development.

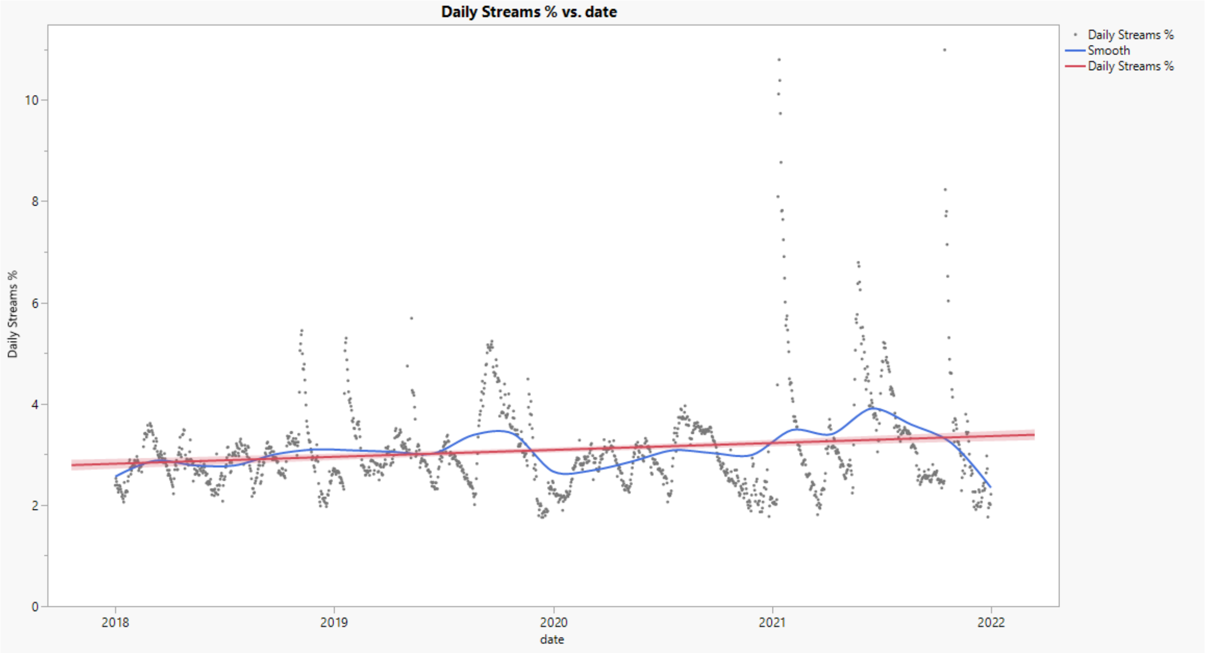


Figure 3.9 Top 1 % of daily Stream Counts over time.

3.6 Musical Attributes

The additional musical attributes variables in the extended dataset represents an opportunity to add detail and insight to the basic positional and stream counts information. An area of specific interest is if there is evidence to suggest that specific types of tracks tend toward the top of the charts. The initial analyse will concentrate on three subsets of the dataset: the Top 1 track only, the Top 10 tracks and all other tracks outside of the top 10, measured over the four-year observation. Figures 3.10 to 3.20 visualise the yearly averages for each of the musical attributes for the uneven bins of "Top 1", "Top 10" and "Outside Top 10" over the four-year observation.

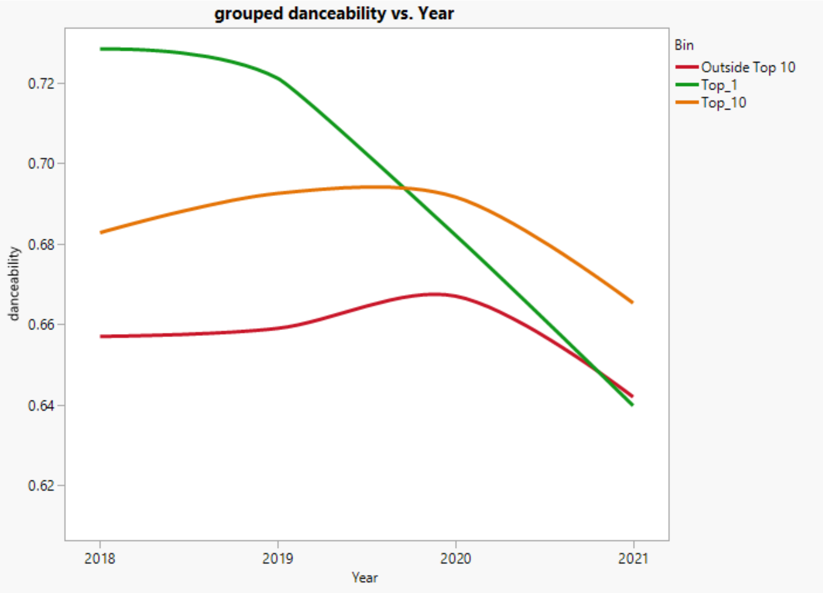


Figure 3.10 Grouped danceability over time.

The mean `danceability` of the overall dataset shown in Figure 3.10 has fallen over the four-year observation period. The rate of decline is more pronounced for the Top 1 position. The Top tracks in 2021 tend to be less danceable than in each of the preceding years, and lower than the general trend in the Top 10 grouping, and the overall dataset mean.

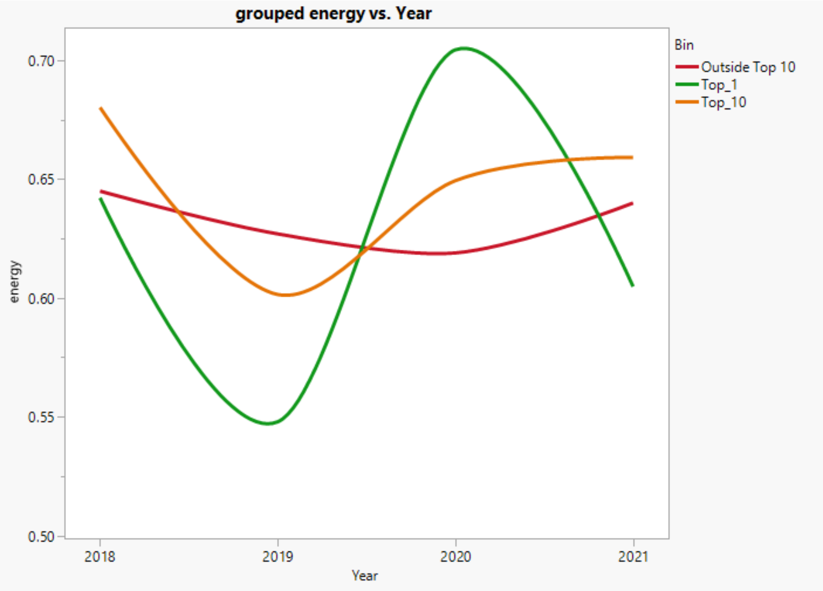


Figure 3.11 Grouped energy over time.

Figure 3.11 shows Yearly `energy` variability over the observation period, while there is a reduction from the initial 2018 value for each of the bins there is not a robust indication that this trend will continue. The Top 1 and Top 10 trend is of interest for 2020, with the top performing tracks indicating an increase in `energy`, while the remainder of the dataset showed a downward movement.

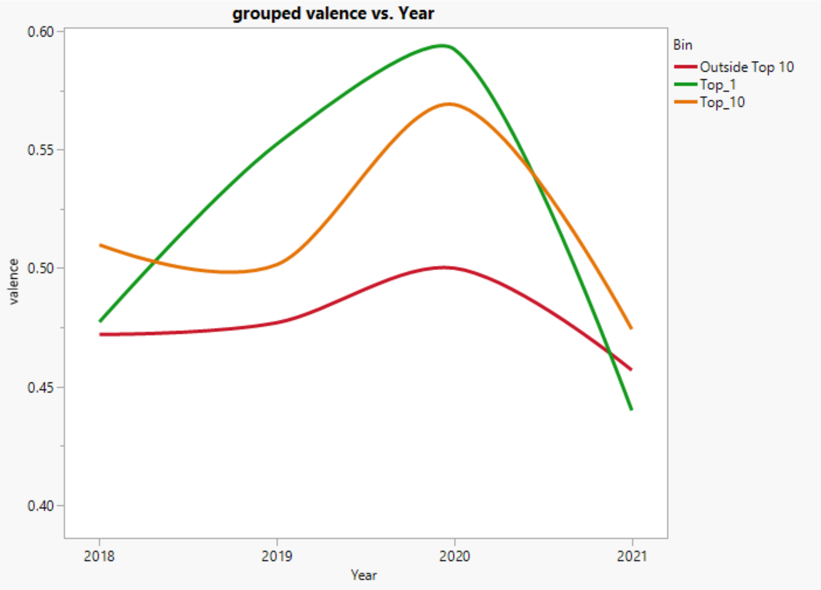


Figure 3.12 Grouped valence over time.

Overall, the mean value for `valence` has similarly fallen over the observation period, as shown in Figure 3.12. However, 2020 shows that in the dataset the value was higher, with much the same pattern as with `energy` in Figure 3.11, before dropping sharply. Of interest is a similar, although not as pronounced rise is seen in `danceability` and `energy` for tracks that are not the Top 1.

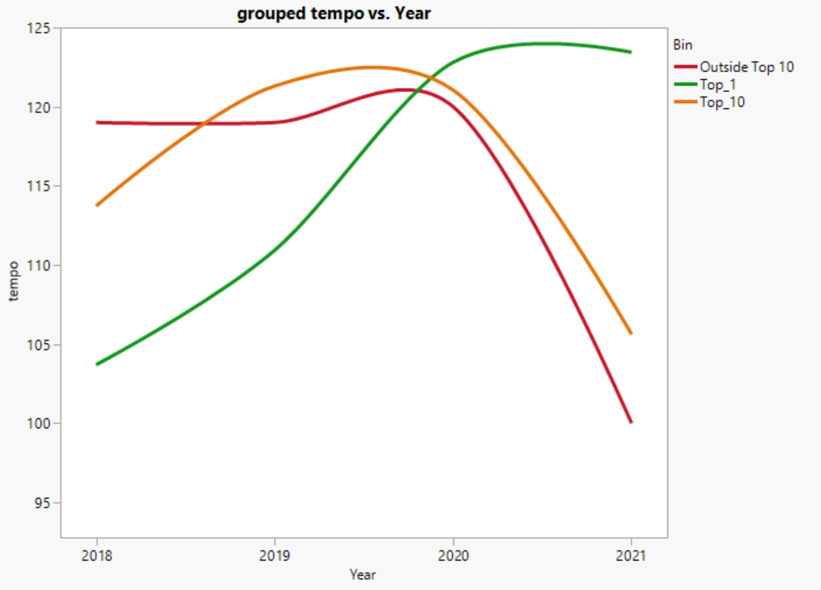


Figure 3.13 Grouped tempo over time.

Figure 3.13 shows that `tempo` for the Top 1 track has shown some evidence of opposing the general dataset trend. From 2018, where the `tempo` of the Top 1 track trailed the remainder of the dataset it has continued to rise, while the overall trend in the dataset is downward. Again, there is an uplift noted in the general sense in 2020.

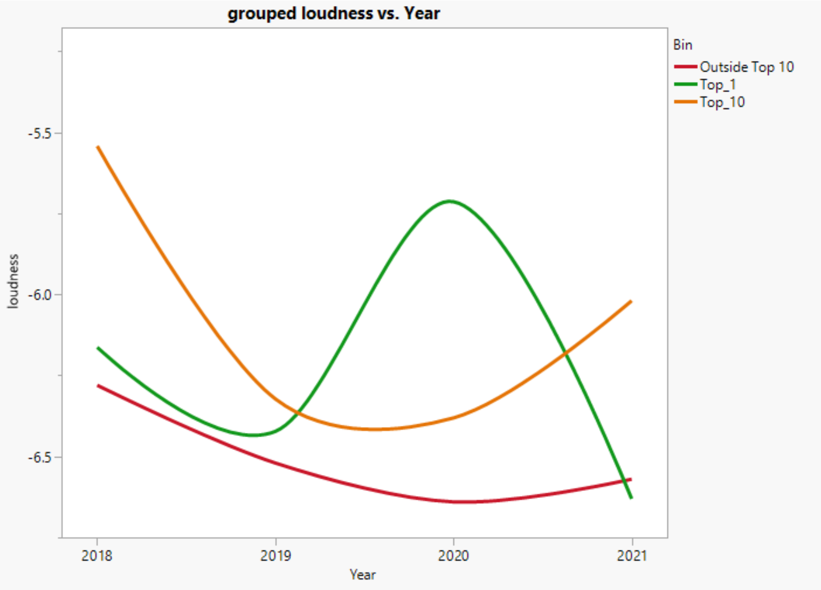


Figure 3.14 Grouped loudness over time.

The `loudness` variable can be considered as a measure of overall volume in a track. What this means is that the Top 1 tracks tend to be quieter. Figure 3.14 shows a noticeable increase in 2020 that goes against the overall general downward trend but aligns with the previous attributes of `energy`, `valence` and `danceability`, as does the Top 10 track values for 2020 and 2021.

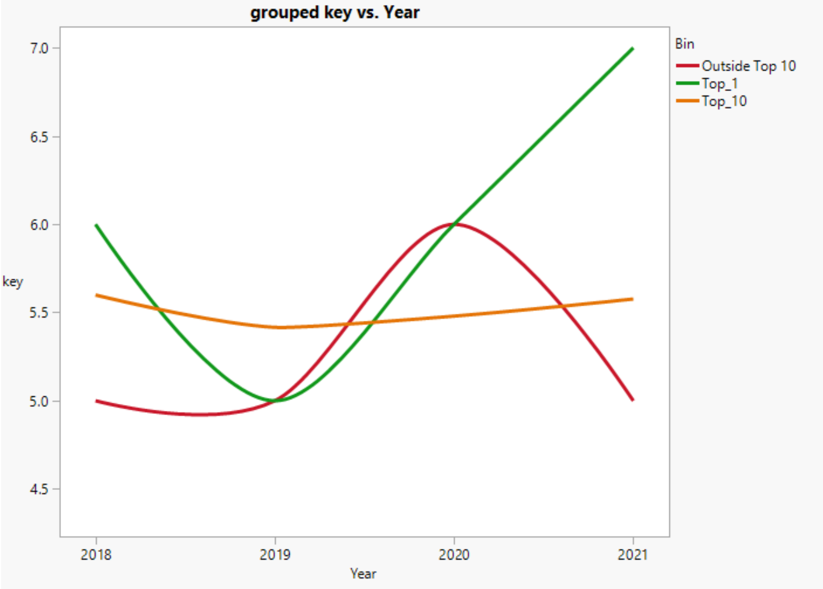


Figure 3.15 Grouped key over time.

The most striking feature of Figure 3.15 for `key` is that in 2021 the Top 1 tracks tended towards a full octave higher. There may be a correlation with the trending toward acoustic and instrumental driven tracks, particularly with the emergence of strong, female singer – songwriters such as Olivia Rodrigo.

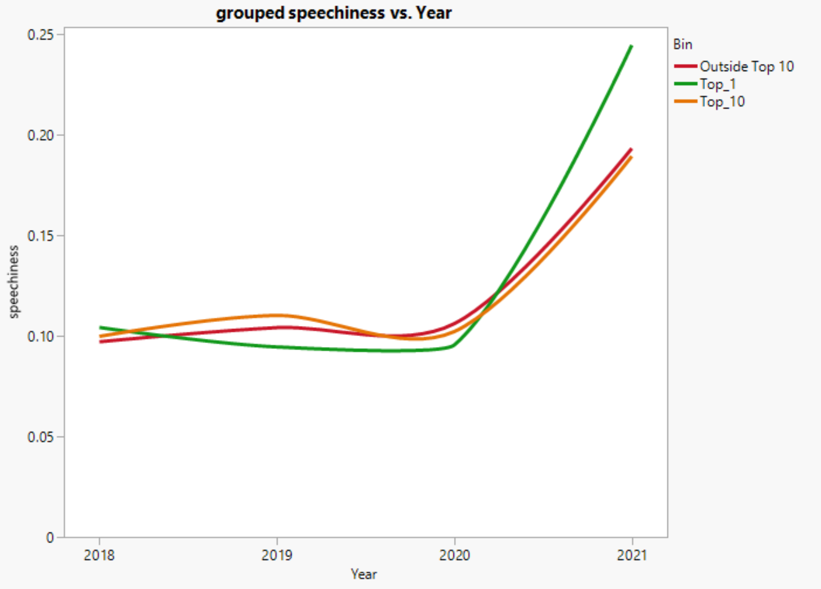


Figure 3.16 Grouped speechiness over time.

A very clear trend across the entire dataset is evident in Figure 3.16, whereby `speechiness` has increased quiet significantly since 2020. The entire dataset is trending towards tracks with a higher vocal content. While all values are less than 0.33, which Spotify indicates are mostly musical or instrumental tracks, the change in the attribute in 2021 is obvious.

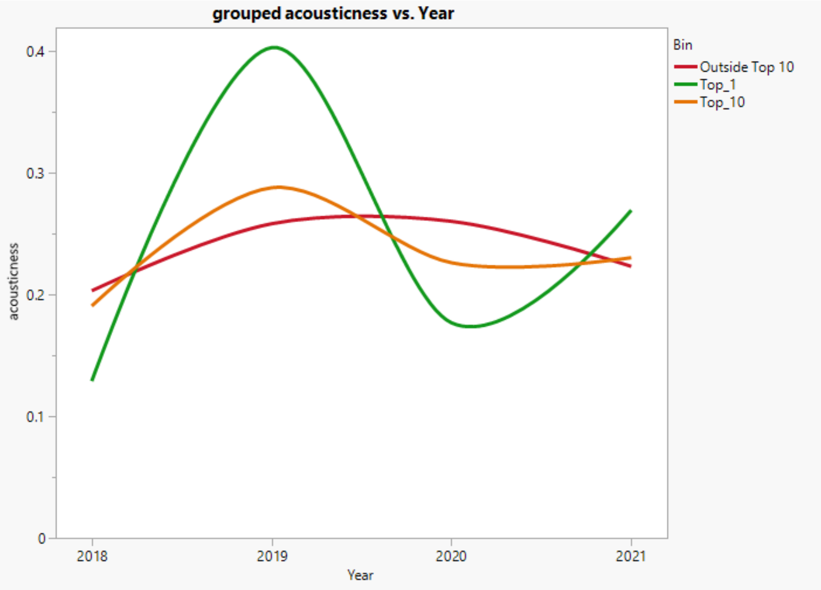


Figure 3.17 Grouped acousticness over time.

The `acousticness` variable shown in Figure 3.17 indicates that the Top1 and Top 10 tracks in 2019 tended towards a higher degree of `acousticness`. While the trend was reversed in 2020 and into 2021, the Top 1 tracks indicate a tendency towards more acoustic tracks, however the acoustic values are still low.

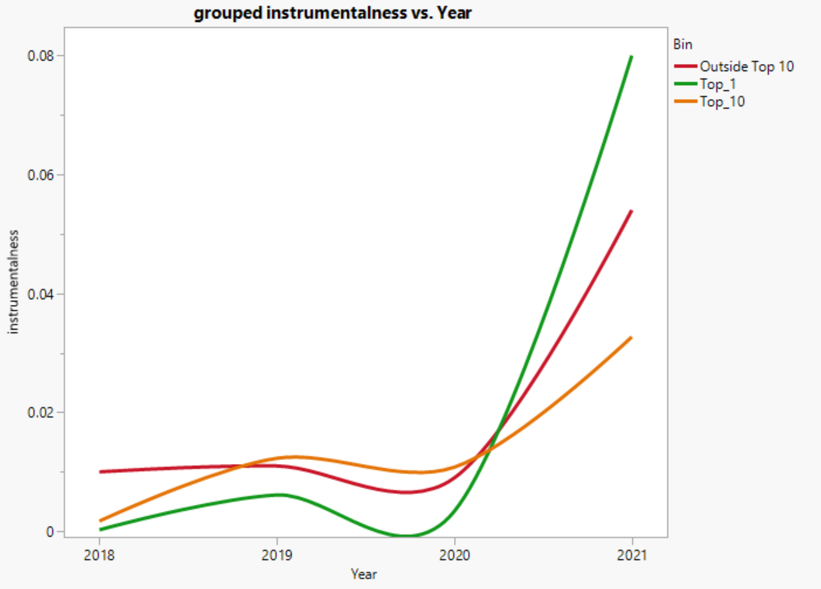


Figure 3.18 Grouped instrumentalness over time.

As with the `speechiness` variable there is a very clear upward trend from 2020 to 2021. However, as the actual values are all low the actual difference may not be significant. Figure 3.18 would suggest the dataset is trending towards tracks that are more instrumental than in previous years but coming from a low baseline.

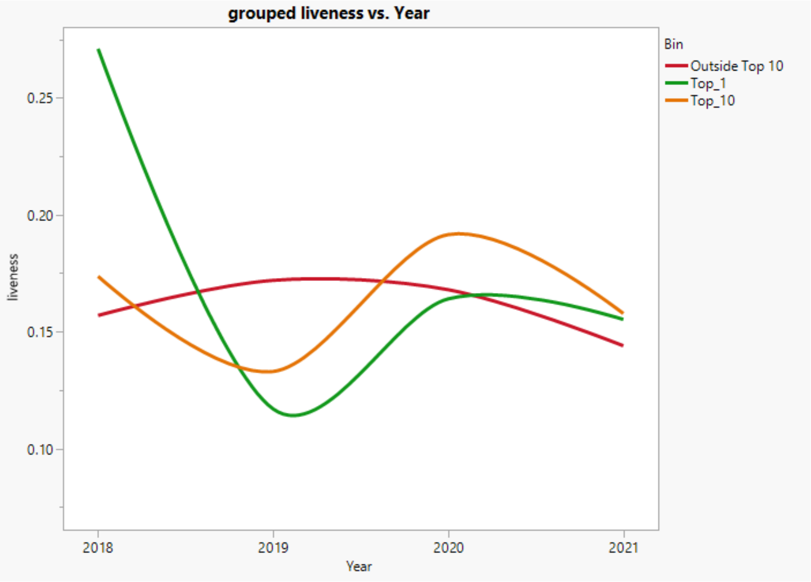


Figure 3.19 Grouped liveness over time.

Figure 3.19 shows that overall, the dataset is trending away from live tracks, yet again 2020 reverses some of the downward change. There is little difference between the three groupings, suggesting that the whole dataset is moving to tracks that are less live, this may be correlated to external event. The global Covid19 pandemic resulted in the cancellation or postponement of live performances during much of 2020 and 2021, which may have affected listener patterns.

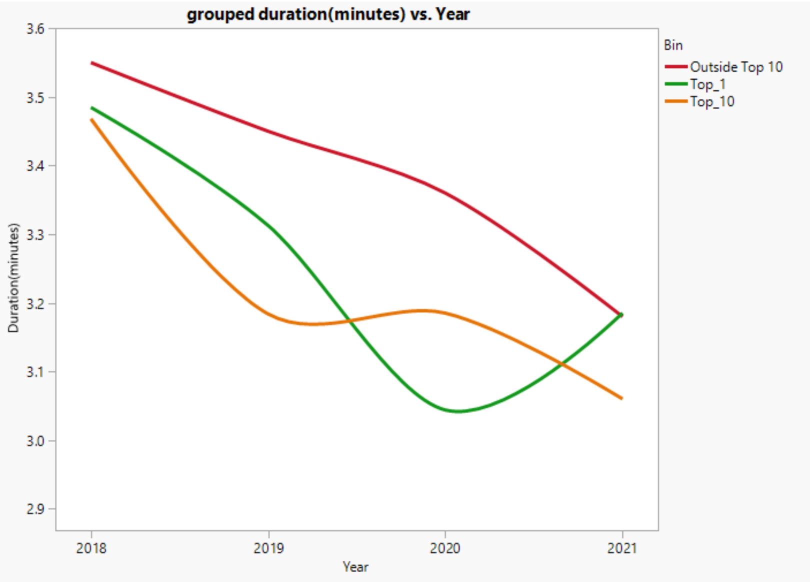


Figure 3.20 Grouped liveness over time.

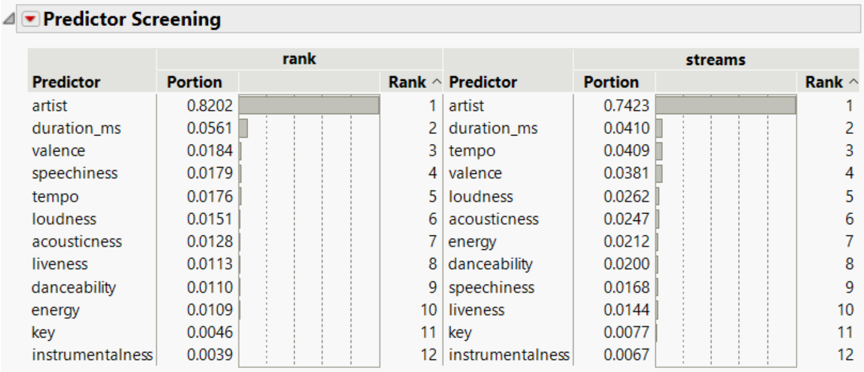
Figure 3.20 indicates that the `duration` of tracks in general has fallen over the four-year period, with the mean track length for the Top 1 tracks in 2021 being almost 18 seconds shorter than in 2018. A similar reduction in duration is also seen for the Top 10 tracks, being almost 25 seconds shorter, and the outside Top 10 tracks reducing by 22 seconds.

The independent examination of the musical attributes demonstrates an overall change in the types of music performing well in the dataset over the four years, especially for the Top 1 tracks. Figures 3.10, 3.11 and 3.12 show that the tracks that perform well tend to be less danceable, less energetic and with a lower valence. Figures 3.14 to 3.20 show that these tracks also tend to be less loud, in a higher key, with more vocals, somewhat acoustic and instrumental, but not live recordings, and shorter.

4. Results

4.1 Model Design

Table 4.1 Predictor screening results for rank and streams variables using the Top 5% of dataset.



Predictor Screening

	rank			streams	
Predictor	Portion	Rank	Predictor	Portion	Rank
artist	0.8202	1	artist	0.7423	1
duration_ms	0.0561	2	duration_ms	0.0410	2
valence	0.0184	3	tempo	0.0409	3
speechiness	0.0179	4	valence	0.0381	4
tempo	0.0176	5	loudness	0.0262	5
loudness	0.0151	6	acousticness	0.0247	6
acousticness	0.0128	7	energy	0.0212	7
liveness	0.0113	8	danceability	0.0200	8
danceability	0.0110	9	speechiness	0.0168	9
energy	0.0109	10	liveness	0.0144	10
key	0.0046	11	key	0.0077	11
instrumentalness	0.0039	12	instrumentalness	0.0067	12

Predictor screening was conducted using bootstrapped random forest. As the top position on the daily chart is determined by streams count, the predictor screening was completed for both rank and streams as independent outputs. As the primary focus is on tracks that perform well, a local filter was applied for only those tracks with rank in the range of 1 to 10, i.e., the top 5% of the dataset and the results are shown in Table 4.1.

The screening was conducted using the source data variables, excluding title, date and trend. The rationale behind excluding these variables is that trend is a function of rank changing over time. The data analysis previously showed that there was little variability in the Top 1 position, i.e., songs tended to remain in the top positions, this was discussed in section 3.1 and shown in Table 3.3. As a result, the effect of trend and date as predictors was being skewed by the longevity of tracks in the top positions. As can be seen in the Table 4.1, while the musical attributes vary in order and effect, the primary predictor is artist and is responsible for 82% of rank and 74% of streams output.

The artist was included specifically as the initial data exploration clearly showed that from the total variability of artist within the dataset the Top 1 position was limited to 2.52% of artists. The Top 10 positions achieved by 11% of artists and the expanding bins for highest chart position achieved showed that 82% of the artists never make it to the Top 20 position. Furthermore, 66% of the artists do not achieve a position higher than rank = 50. Simply put, forty-five of all artists in the dataset monopolise the top 1 chart position. Figure 4.1 shows the count of artist distribution in the dataset for different chart position bins. The total number of distinct artists in the dataset is 1,780. The inclusion of such a strong predictor in any model would lessen the general applicability of any model as a predictive tool. Effectively, the artist variable is such a strong predictor that all other (musical) attributes are potentially nullified by its inclusion.

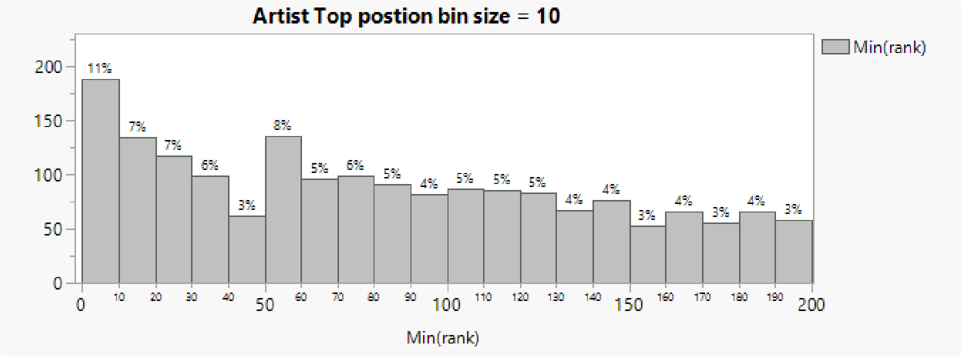
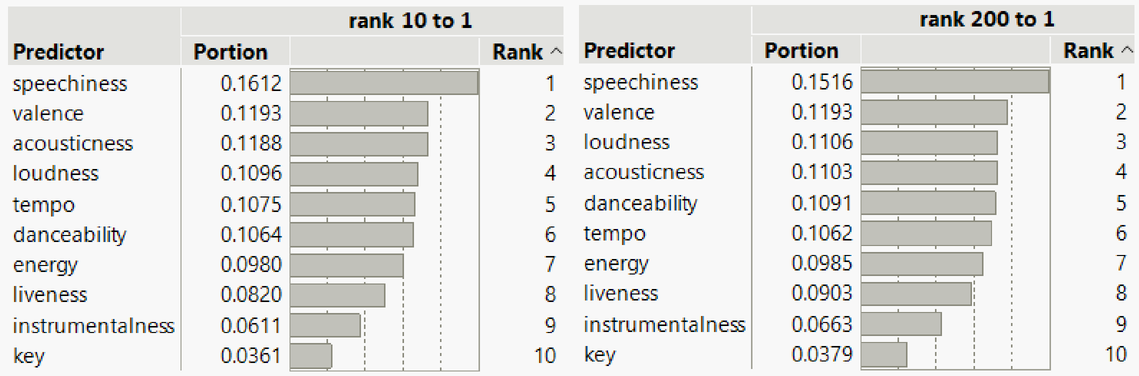


Figure 4.1 Distribution of Artist count, full dataset.

The research question of this paper is if prediction of chart `rank` or `streams` count is possible using the variables present in the dataset. The data analysis shows that the `artist` is the most significant influencer irrespective of musical attributes, this requires a two-pronged. One modelling effort will use the `artist` variable, while the second will exclude `artist` as a predictor. The outcome of the models can then be assessed in terms of accuracy, generalisation and usefulness to inform if the `artist` variable does indeed nullify the musical attributes. Re-running the predictor screening on for the top 5% positions (`rank` 10 to 1) and the full dataset (`rank` 200 to 1) using only the musical attributes, excluding `artist` and `duration`, shows that the most influencing variables are those associated with uplifting vocal tracks with an acoustic overtone. The screening results are shown in Table 4.2. These attributes: `speechiness`, `valence`, `acousticness` were also identified in the preliminary analysis in section 3.6, which indicated that the most prominent artists in the database have tracks that largely fit these musical attributes. As the most popular artists feature so prominently in the dataset, it is highly probable that the influence of these artists is driving the similarities in the full and subset predictor screening, even when that actual `artist` variable is not included.

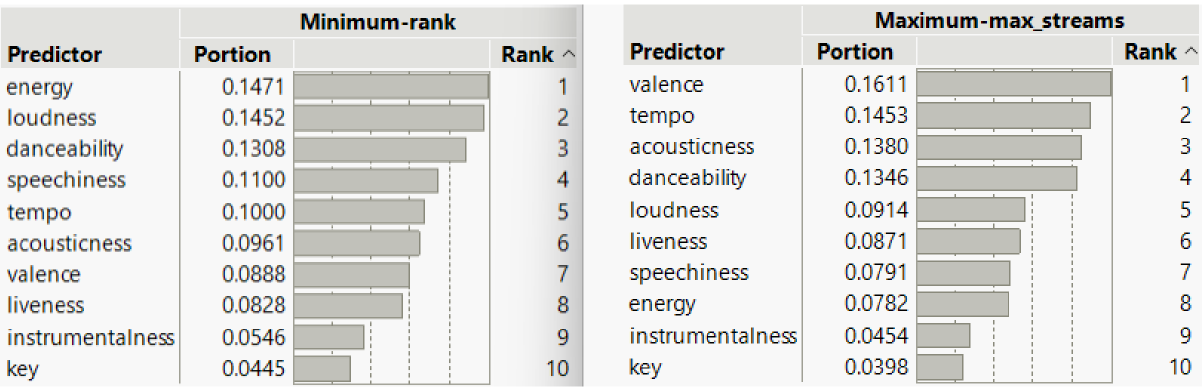
Table 4.2 Predictor Screening Results for rank and streams variables on the entire dataset with artist omitted.



	rank 10 to 1				rank 200 to 1		
Predictor	Portion		Rank ^	Predictor	Portion		Rank ^
speechiness	0.1612		1	speechiness	0.1516		1
valence	0.1193		2	valence	0.1193		2
acousticness	0.1188		3	loudness	0.1106		3
loudness	0.1096		4	acousticness	0.1103		4
tempo	0.1075		5	danceability	0.1091		5
danceability	0.1064		6	tempo	0.1062		6
energy	0.0980		7	energy	0.0985		7
liveness	0.0820		8	liveness	0.0903		8
instrumentalness	0.0611		9	instrumentalness	0.0663		9
key	0.0361		10	key	0.0379		10

As was noted during the analysis phase in section 3.1, the repetitive structure of the dataset would lead to problems with multi-collinearity as the same track attributes would be assessed multiple times. A greater influencing effect being noted for tracks that are more frequent, i.e. the top performing tracks. To address this problem of repeated identical observations, the dataset was reduced to a single occurrence of each distinct `track`. Filtering was applied for the maximum value of the `streams` count, as the maximal values typically correlated at+70% with the highest chart positions, this resulted in a dataset of 6,035 observations. The initial modelling activity will involve only those musical attributes as shown in Table 4.2 on this reduced dataset (N=6,035). Predictor screening was again completed to determine if a significant change in the response was observed on this reduced dataset, with the output shown in Table 4.3.

Table 4.3 Predictor Screening Results for rank and streams variables on the reduced dataset with artist omitted.



Minimum-rank			Maximum-max_streams		
Predictor	Portion	Rank	Predictor	Portion	Rank
energy	0.1471	1	valence	0.1611	1
loudness	0.1452	2	tempo	0.1453	2
danceability	0.1308	3	acousticness	0.1380	3
speechiness	0.1100	4	danceability	0.1346	4
tempo	0.1000	5	loudness	0.0914	5
acousticness	0.0961	6	liveness	0.0871	6
valence	0.0888	7	speechiness	0.0791	7
liveness	0.0828	8	energy	0.0782	8
instrumentalness	0.0546	9	instrumentalness	0.0454	9
key	0.0445	10	key	0.0398	10

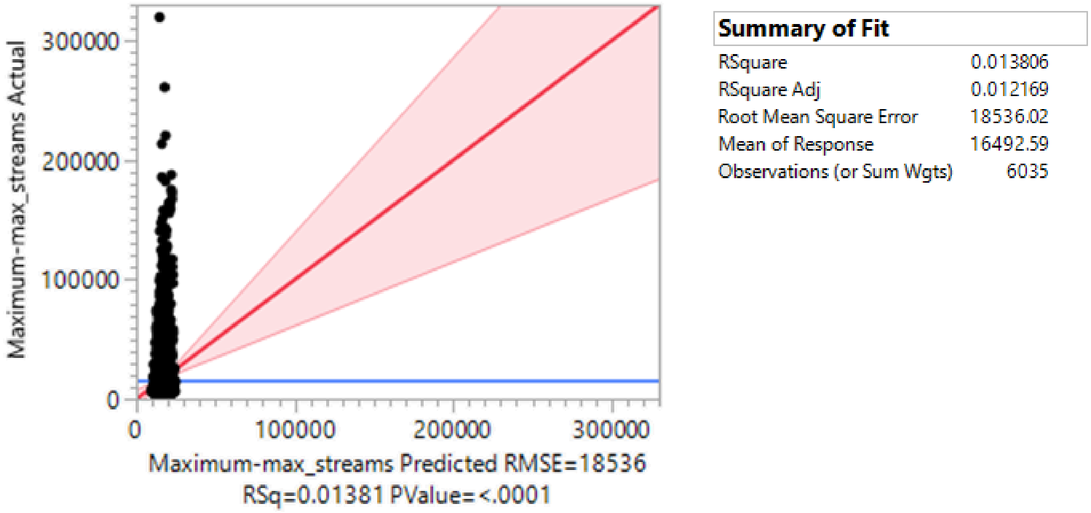
Table 4.3 shows that there is a significant re-ordering of the predictors for both dependant variables of `rank` and max `streams` when using the reduced dataset. The models will utilise this reduced dataset, primarily because of the concerns surrounding repetition. Two predictive modelling methods are utilised, multiple regression and neural network. For both methods the response variable of `streams` count is estimated.

4.2 Model 1.1 Multiple Regression without artist variable

The initial multiple regression model utilising least squares was developed with the response variable of `streams` count. The predictors were: `danceability`, `energy`, `key`, `loudness`, `speechiness`, `acousticness`, `instrumentalness`, `liveness`, `valance` and `tempo`. The performance of the model with an R^2 of 0.013, Adjusted R^2 of 0.012 and an RMSE of 18536, indicates that this initial model is not of any practical use in predicting max `streams` values using the musical attributes. Figure 4.2 shows the actual vs predicted plot of the model output, along with summary information of the results. This clearly shows the model trend to underestimate the `streams` count, with all prediction points less than 30,000 `streams`.

The complete model estimate formula is:

Estimate = 21102.89 + (-3750.35 * danceability) + (-6327.97 * energy) + (-51.28 * key) + (279.89 * loudness) + (996.63 * speechiness) + (5366.87 * acousticness) + (-2461.07 * instrumentalness) + (2392.03 * liveness) + 5140.67 * valence) + (-1.57 * tempo)



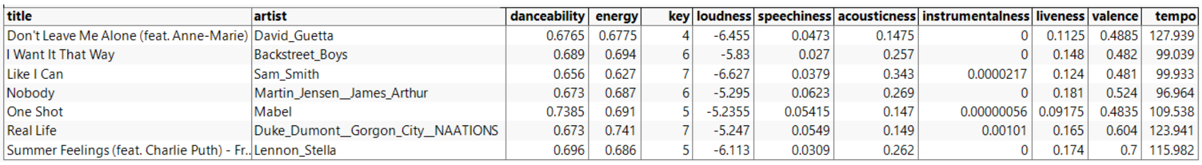
Summary of Fit

RSquare	0.013806
RSquare Adj	0.012169
Root Mean Square Error	18536.02
Mean of Response	16492.59
Observations (or Sum Wgts)	6035

Figure 4.2 Model 1.1 musical attributes only, streams response on the reduced dataset.

Analysis of the model results does lead to some insight; Table 4.4 shows the predicted values with the lowest standard error, it also shows that these tracks have remarkably similar musical attributes that fit a much narrower range band than from the whole dataset.

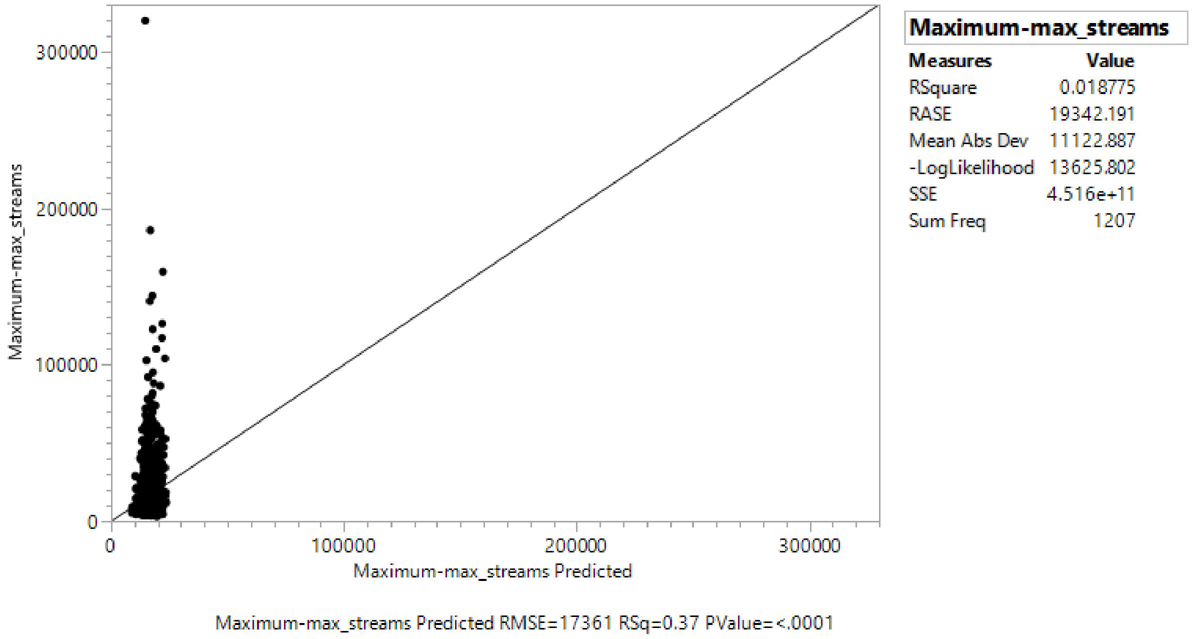
Table 4.4 Most accurate predictions of streams count Model 1.1



title	artist	danceability	energy	key	loudness	speechiness	acousticness	instrumentalness	liveness	valence	tempo
Don't Leave Me Alone (feat. Anne-Marie)	David_Guetta	0.6765	0.6775	4	-6.455	0.0473	0.1475	0	0.1125	0.4885	127.939
I Want It That Way	Backstreet_Boys	0.689	0.694	6	-5.83	0.027	0.257	0	0.148	0.482	99.039
Like I Can	Sam_Smith	0.656	0.627	7	-6.627	0.0379	0.343	0.0000217	0.124	0.481	99.933
Nobody	Martin_Jensen__James_Arthur	0.673	0.687	6	-5.295	0.0623	0.269	0	0.181	0.524	96.964
One Shot	Mabel	0.7385	0.691	5	-5.2355	0.05415	0.147	0.00000056	0.09175	0.4835	109.538
Real Life	Duke_Dumont__Gorgon_City__NAATIONS	0.673	0.741	7	-5.247	0.0549	0.149	0.00101	0.165	0.604	123.941
Summer Feelings (feat. Charlie Puth) - Fr..	Lennon_Stella	0.696	0.686	5	-6.113	0.0309	0.262	0	0.174	0.7	115.982

4.3 Model 1.2 Neural Network without artist variable

The initial neural network model was developed with the response variable of streams count. As with the multiple regression model 1.1 the predictors were: danceability, energy, key, loudness, speechiness, acousticness, instrumentalness, liveness, valance and tempo. The neural network utilised 3 neural layers and was validated by K-fold cross validation, with 5 folds. The performance of the model with an R^2 of 0.018 and an RASE of 19342, indicates that this initial model is also not of any practical use in predicting max streams values using the musical attributes. Figure 4.3 shows the actual vs predicted plot of the neural network model output, along with summary information of the results. This clearly shows the model trend to underestimate the *streams* count, with all prediction points less than 30,000 *streams*, which is similar to the values observed by the multiple regression results from Figure 4.2.



Maximum-max_streams

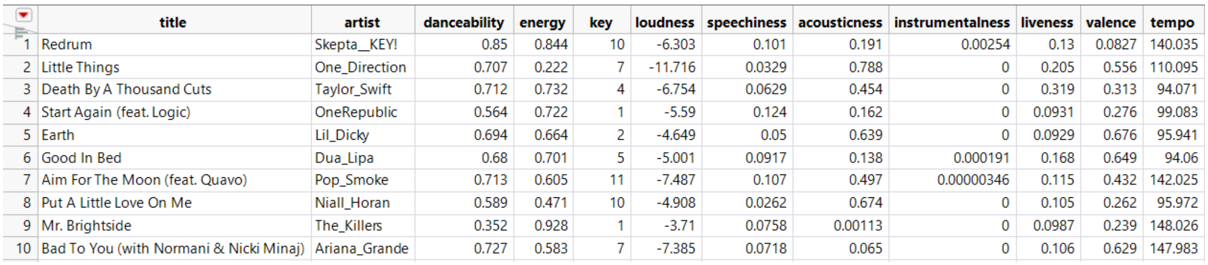
Measures	Value
RSquare	0.018775
RASE	19342.191
Mean Abs Dev	11122.887
-LogLikelihood	13625.802
SSE	4.516e+11
Sum Freq	1207

Maximum-max_streams Predicted RMSE=17361 RSq=0.37 PValue=<.0001

Figure 4.3 Neural Network Model 1.2 musical attributes only, reduced dataset.

The analysis of the neural model prediction data did not point to any significant trend, however the subset of most accurately predicted tracks was interesting, because it represented a sample of some of the lowest of "maximum stream count" for popular artists (i.e., one of their worst performing songs as measured by *streams* count). An exception to this was "Mr. Brightside" which is not only the most successful track by The Killers, but also the single most prevalent track in the entire dataset, an input that the model did not have. No real determination can be made as to why these were the most accurate predictions, yet it remained an interesting observation and one that may warrant further research. Table 4.2 shows the prediction data for the ten most accurately predicted results.

Table 4.5 Most accurate predictions Model 1.2



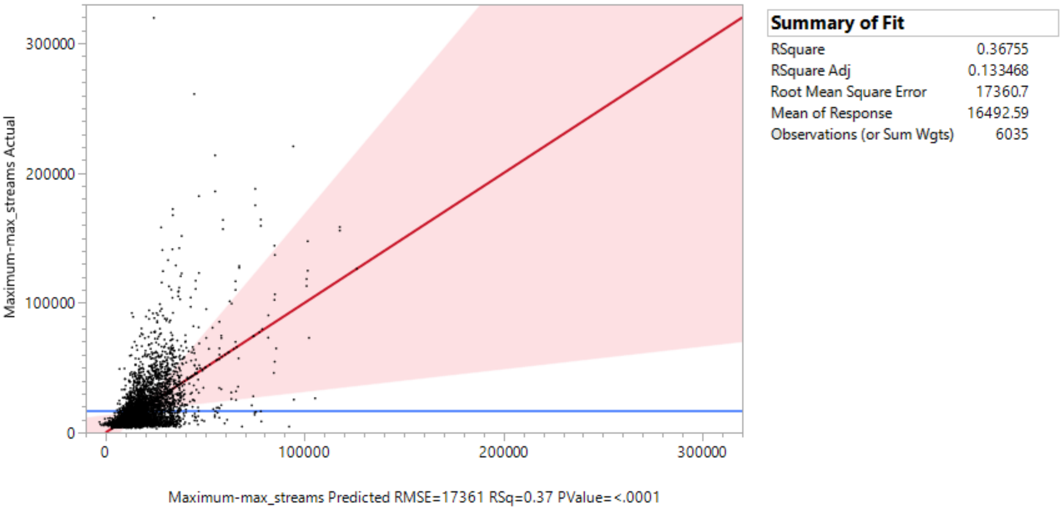
	title	artist	danceability	energy	key	loudness	speechiness	acousticness	instrumentalness	liveness	valence	tempo
1	Redrum	Skepta__KEY!	0.85	0.844	10	-6.303	0.101	0.191	0.00254	0.13	0.0827	140.035
2	Little Things	One_Direction	0.707	0.222	7	-11.716	0.0329	0.788	0	0.205	0.556	110.095
3	Death By A Thousand Cuts	Taylor_Swift	0.712	0.732	4	-6.754	0.0629	0.454	0	0.319	0.313	94.071
4	Start Again (feat. Logic)	OneRepublic	0.564	0.722	1	-5.59	0.124	0.162	0	0.0931	0.276	99.083
5	Earth	Lil_Dicky	0.694	0.664	2	-4.649	0.05	0.639	0	0.0929	0.676	95.941
6	Good In Bed	Dua_Lipa	0.68	0.701	5	-5.001	0.0917	0.138	0.000191	0.168	0.649	94.06
7	Aim For The Moon (feat. Quavo)	Pop_Smoke	0.713	0.605	11	-7.487	0.107	0.497	0.00000346	0.115	0.432	142.025
8	Put A Little Love On Me	Niall_Horan	0.589	0.471	10	-4.908	0.0262	0.674	0	0.105	0.262	95.972
9	Mr. Brightside	The_Killers	0.352	0.928	1	-3.71	0.0758	0.00113	0	0.0987	0.239	148.026
10	Bad To You (with Normani & Nicki Minaj)	Ariana_Grande	0.727	0.583	7	-7.385	0.0718	0.065	0	0.106	0.629	147.983

4.4 Initial Model Results

Following a review of the accuracy results of the initial multiple regression and neural network models, and analysis of the most accurate predictions the models were re-run with the inclusion of the artist variable. The expectation was that as indicated by the predictor screening, discussed in chapter 4.1 and shown in Table 4.1, the inclusion of the artist variable would improve the models. All additional model parameters remained unchanged.

4.5 Model 2.1 Multiple Regression with artist variable included

The multiple regression model utilising least squares was re-run with the inclusion of the artist variable as a predictor, with the same response variable of streams count. The predictors were: artist, danceability, energy, key, loudness, speechiness, acousticness, instrumentalness, liveness, valance and tempo. The performance of the model with an R^2 of 0.36, Adjusted R^2 of 0.13 and an RMSE of 17360, indicates that the revised model, inclusive of artist, is more accurate in predicting maximum streams values. However, with an adjusted R^2 of 0.13 and the explicit usage of 1620 individual artist predictors the actual usefulness of the model is questionable. Figure 4.4 shows the actual vs predicted plot of the model output, along with summary information of the results. The model trend to underestimate the streams count is still present, yet some reasonable results are achieved as indicated by the existence of prediction points along the line of fit, and within the 5% confidence banding, however, the model still fails to predict the upper ranges of the streams counts.



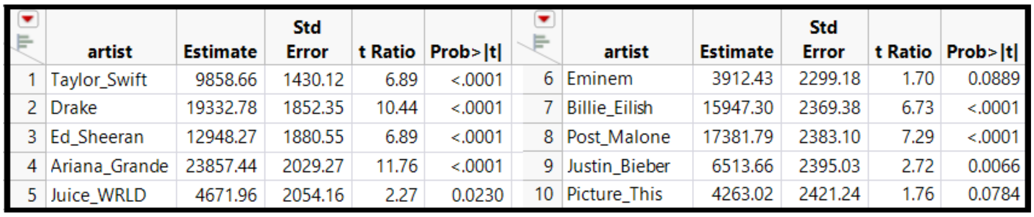
Summary of Fit	
RSquare	0.36755
RSquare Adj	0.133468
Root Mean Square Error	17360.7
Mean of Response	16492.59
Observations (or Sum Wgts)	6035

Figure 4.4 Multiple Regression Model 2.1 musical attributes and artist, reduced dataset.

Of interest, is that analysis of model results, specifically those with those with the lowest standard error showed that many of the best performing artists were most accurately predicted. Table 4.3

shows an excerpt from the overall results. In this case the musical attributes are not of specific interest as the analysis is on the artists with most accurate predictions.

Table 4.6 Most accurate predictions Model 2.1

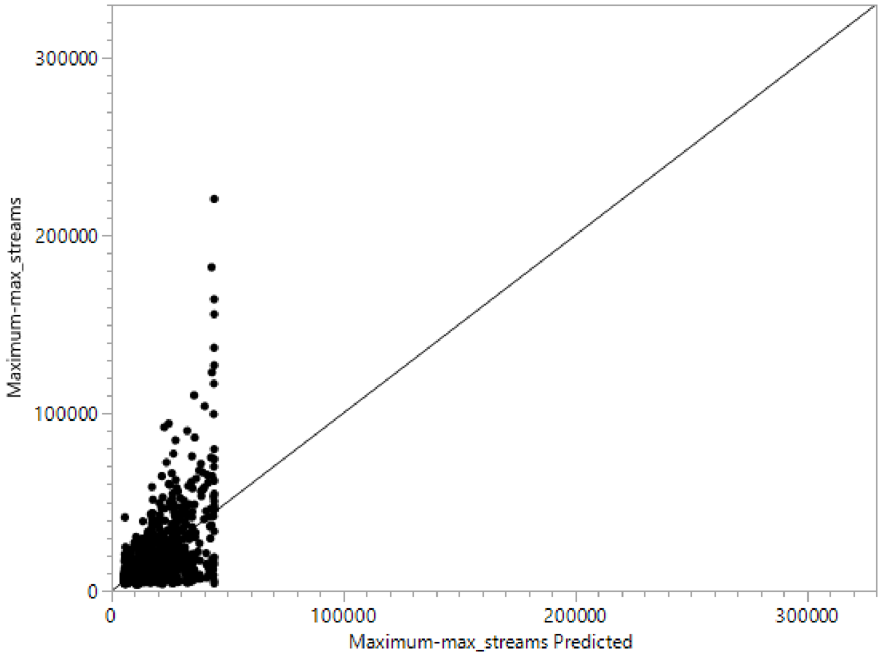


	artist	Estimate	Std Error	t Ratio	Prob>|t|
1	Taylor_Swift	9858.66	1430.12	6.89	<.0001
2	Drake	19332.78	1852.35	10.44	<.0001
3	Ed_Sheeran	12948.27	1880.55	6.89	<.0001
4	Ariana_Grande	23857.44	2029.27	11.76	<.0001
5	Juice_WRLD	4671.96	2054.16	2.27	0.0230
6	Eminem	3912.43	2299.18	1.70	0.0889
7	Billie_Eilish	15947.30	2369.38	6.73	<.0001
8	Post_Malone	17381.79	2383.10	7.29	<.0001
9	Justin_Bieber	6513.66	2395.03	2.72	0.0066
10	Picture_This	4263.02	2421.24	1.76	0.0784

Many of the `artist` names are recognisable as those artists that had the most Top 1 positions as shown in Table 3.13, highest total `streams` count as shown in Table 3.12 and most days in the dataset as explored in the initial analysis in chapter 3. An additional point of interest is that the model is more accurate at predicting those tracks with `streams` count closer to the dataset mean of 16492. Analysis of the model results indicated that of the 992 tracks that were predicted within the 5% confidence boundary, all had actual `streams` count of less than 16492. When the extended analysis was conducted on the results of the model, it became clear that the effect of the `artist` variables far outweighed the effect of the musical attributes, to the point where the highest `streams` count could not be predicted. An anomalous finding is that the 150 data points with the lowest standard error were all by Taylor Swift. Simply put, the association of a specific `artist` variable greatly improved the accuracy of prediction. This was due in no small part to how often these top performing artists were present, even in the reduced dataset. The revised model, including the `artist` variable, tended to be more accurate for those artists most prevalent in the dataset.

4.6 Model 2.2 Neural Network with artist variable included

The neural network model was re-developed with the response variable of `streams` count and with the additional `artist` variable included. No additional modification to the model were made, the predictors were: `artist`, `danceability`, `energy`, `key`, `loudness`, `speechiness`, `acousticness`, `instrumentalness`, `liveness`, `valance` and `tempo`. The neural network utilised 3 neural layers and was validated by K-fold cross validation, with 5 folds. The performance of the model with an R^2 of 0.33 and an RASE of 15052, indicates a significant improvement in predicting max `streams` values when including the `artist` variable with the musical attributes. While the revised neural model continues to underestimate the `streams` count, there are indications that higher estimates are now being generated, this is similar to those values observed by the revised multiple regression model 2.1 results from Figure 4.4. The R^2 values of both of the revised models indicates a similar effect with the inclusion of the `artist` variable on the models. Figure 4.5 shows the actual vs predicted plot of the revised neural network model output, along with summary information of the results.



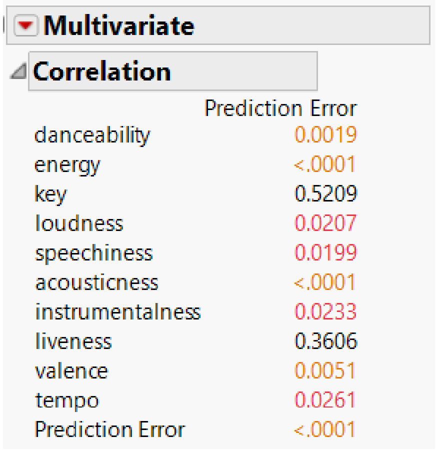
Maximum-max_streams

Measures	Value
RSquare	0.3368558
RASE	15052.046
Mean Abs Dev	7918.041
-LogLikelihood	13323.117
SSE	2.735e+11
Sum Freq	1207

Figure 4.5 Neural Network Model 2.2 musical attributes and artist, reduced dataset.

Further analysis of the predicted results data for this model did not point to any significant trend. However, the subset of most accurately predicted tracks was once again interesting. The model proved to be incredibly accurate for some tracks, with an actual prediction within an error of less than 1% for 373 tracks. The 10 most accurate predictions had an error of less than 0.02%. However, as with the previous multiple regression model 2.1 no real determination can be made as to why these were the most accurate predictions. The only possible indicator of a relationship is in the correlation between the error rate and the attributes of key and liveness, as shown in Table 4.4.

Table 4.7 Correlation of prediction error to musical attributes Neural Network Model 2.1

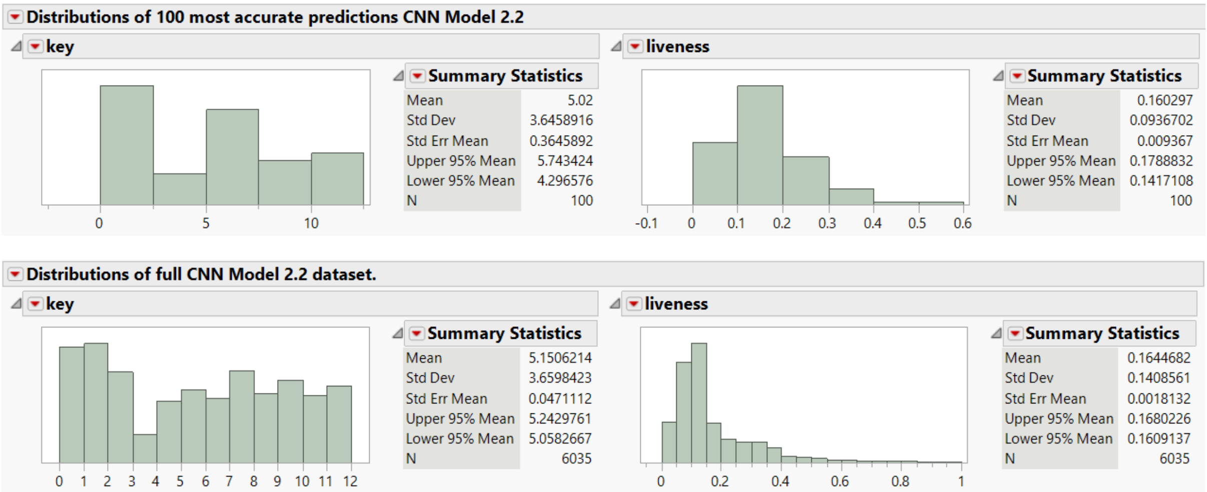


Multivariate

Correlation

	Prediction Error
danceability	0.0019
energy	<.0001
key	0.5209
loudness	0.0207
speechiness	0.0199
acousticness	<.0001
instrumentalness	0.0233
liveness	0.3606
valence	0.0051
tempo	0.0261
Prediction Error	<.0001

A subset comprising of the 100 most accurate prediction results from neural network model 2.2 was examined and the tracks do tend to be non-live tracks. This subset is compared to the dataset used in the model. Figure 4.6 illustrate this comparison, there is no real understanding or explanation of why the model was more accurate for specific tracks. There isn't a significant difference between the key-values for liveness and key.



Distributions of 100 most accurate predictions CNN Model 2.2

key

Summary Statistics

Mean	5.02
Std Dev	3.6458916
Std Err Mean	0.3645892
Upper 95% Mean	5.743424
Lower 95% Mean	4.296576
N	100

liveness

Summary Statistics

Mean	0.160297
Std Dev	0.0936702
Std Err Mean	0.009367
Upper 95% Mean	0.1788832
Lower 95% Mean	0.1417108
N	100

Distributions of full CNN Model 2.2 dataset.

key

Summary Statistics

Mean	5.1506214
Std Dev	3.6598423
Std Err Mean	0.0471112
Upper 95% Mean	5.2429761
Lower 95% Mean	5.0582667
N	6035

liveness

Summary Statistics

Mean	0.1644682
Std Dev	0.1408561
Std Err Mean	0.0018132
Upper 95% Mean	0.1680226
Lower 95% Mean	0.1609137
N	6035

Figure 4.6 Neural Network Model 2.2 liveness and key analysis for 100 most accurate predictions and complete model dataset.

4.7 Model Outcomes, Initial and Revised models

The four models did not prove to be accurate in terms of predicting `streams` count. The initial data analysis outlined in chapter 3, the predicator screening outlined in chapter 4.1 and the results of the revised models which included the `artist` variable all indicate that the `artist` variable is a valid predictor. As the predictor is certainly of merit, it would be erroneous to exclude it fully. The proposal for the next phase of the modelling is to recode the `artist` variable into derived categorical variables that may more accurately map the effect of the artist performance as a predictor. The approach taken is to define a set of variables that represent new attributes associated with the artists. These variable as shown are shown in Table 4.5, along with descriptions. For the `rank` grouping, uneven bin sizes were chosen as the data showed that outside of the top 50 there is not any likelihood that an artist will be successful. These groupings are not exclusive, an artist can exist in each group.

Table 4.8 Artist rank recoding divisions.

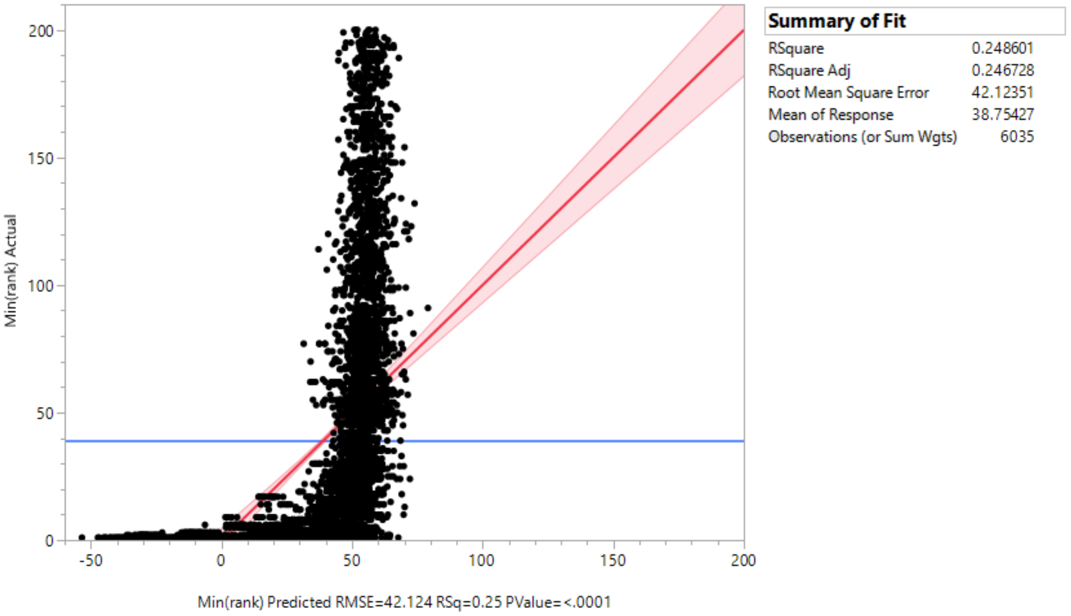
Artist Group	Description	Type
Top_1	Has the artist ever had rank #1	Binary: Yes / No
Top_10	Has the artist ever had rank >1 <=10	Binary: Yes / No
Top_20	Has the artist ever had rank >11 <=20	Binary: Yes / No
Top_50	Has the artist ever had rank >21 <=50	Binary: Yes / No
Top_200	Has had rank >51	Binary: Yes / No

To support the generation of these variables a query was created to return the minimum rank (best chart position) for each artist; the resulting data table was merged with the original dataset to append a column containing the best chart position achieved by the artist, for every row in the dataset. irrespective of the actual rank of the track per row. Conditional formatting was then applied to create the five artist group variables as shown in Table 4.5. Multiple regression and neural network models were then created, with these artist groups in place of the individual distinct artist variable.

However, at this point it was observed that the introduction of the artist group variables created a conflict with the streams count values for those artists that had multiple tracks in the dataset. The streams count values for each independent track would be evaluated against a static “best” rank value, this would make accurate prediction impossible. The decision was made to therefore modify the modelling approach to look at the minimum rank (best position) as the response variable as opposed to the maximum streams count.

4.8 Model 3.1 Multiple Regression with artist groups and musical attributes

The multiple regression model was re-developed with the response variable of minimum rank and with the additional artist group variables included. No additional modification to the model was made, the predictors were: Top_1, Top_10, Top_20, Top_50, Top_200, danceability, energy, key, loudness, speechiness, acousticness, instrumentalness, liveness, valance and tempo. Figure 4.7 shows the actual vs. predicted plot for this model.



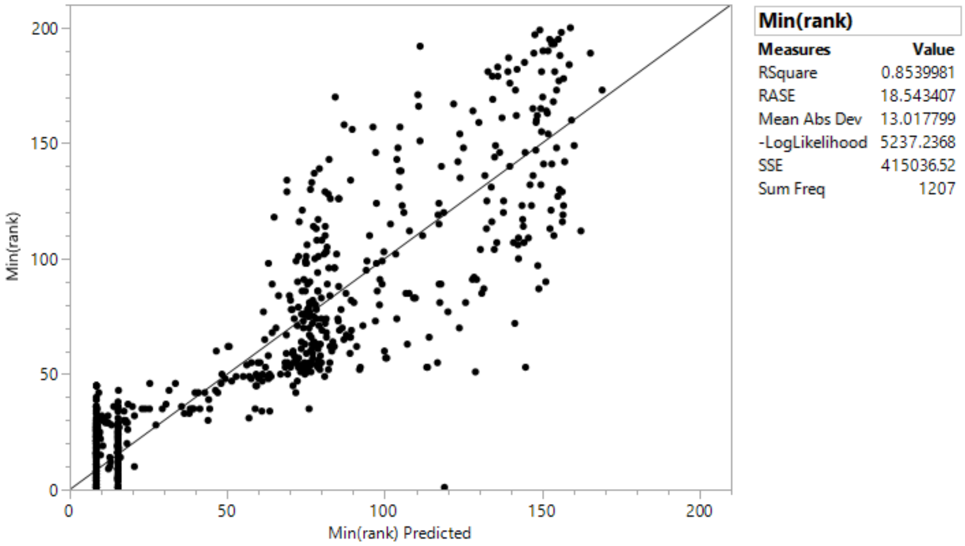
Summary of Fit	
RSquare	0.248601
RSquare Adj	0.246728
Root Mean Square Error	42.12351
Mean of Response	38.75427
Observations (or Sum Wgts)	6035

Figure 4.7 Multiple Regression Model 3.1 musical attributes and artist groups, reduced dataset.

The performance of the model with an R^2 of 0.24 is not particularly accurate in predicting the minimum `rank` values when including the `artist` group variables with the musical attributes. Of significant concern is the prediction of negative values for minimum rank. In this use case multiple regression has proven to be an unsuitable method for prediction, when using the `artist` derived groups variable. Given the negative predictions in the results set the results were excluded from further analysis.

4.9 Model 3.2 Neural Network with artist groups and musical attributes

The neural network model was re-developed with the response variable of minimum `rank` and with the additional artist group variables included. No additional modification to the model was made, the predictors were: `Top_1`, `Top_10`, `Top_20`, `Top_50`, `Top_200`, `danceability`, `energy`, `key`, `loudness`, `speechiness`, `acousticness`, `instrumentalness`, `liveness`, `valance` and `tempo`. The performance of the model with an R^2 of 0.85 and an RASE of 18.54, indicates strong accuracy in predicting the minimum `rank` values when including the artist group variables with the musical attributes. Figure 4.7 shows the actual vs. predicted plot for this model.



Min(rank)

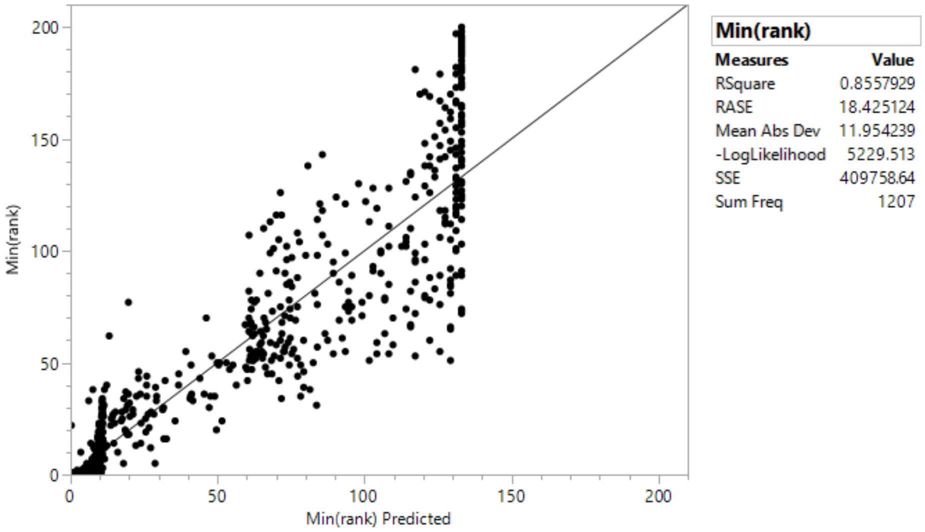
Measures	Value
RSquare	0.8539981
RASE	18.543407
Mean Abs Dev	13.017799
-LogLikelihood	5237.2368
SSE	415036.52
Sum Freq	1207

Figure 4.8 Neural Network Model 3.2 musical attributes and artist groups, reduced dataset.

Figure 4.7 shows the strong performance of the model over the range of `rank` values, indicating that the use of the artist groups in this neural network model has proven to be of value. The model results indicate problems with the fit, evidence of both over and under estimation is shown. However, the model does appear to perform well, for this thesis the final neural network model using the artist groups variables and the musical attributes has achieved better than anticipate results.

5. Discussion & Conclusion

The creation and utilisation of the artist classifiers has proven to be the most useful method to influence prediction, and these are certainly of interest and importance. While the final neural network model 3.2 did demonstrate greater accuracy when using the artists group classifiers there is little practical use to the model. The model is estimating positions based on derived `artist` variables, based on historical data, which may potential override the actual musical attributes. This has been discussed at length in this thesis in the initial data analysis and this phenomenon was also discussed in the literature, Ingle *et al.* (2021). The Initial data analysis in chapter 3, and the predictor screening in chapter 4 identifies the `artist` as the primary driver of success, any variable created from the artists historical record inherits this characteristic. For an unknown artist no such history can exist which means that the prediction would be based exclusively on the musical attributes, which has been demonstrated to be inaccurate. To assess the effect of this on the final model the neural model was re-run with only the artist group variable. i.e. no musical attributes were used as predictors. Figure 4.9 shows the actual vs predicted plot of the results.



Min(rank)	
Measures	**Value**
RSquare	0.8557929
RASE	18.425124
Mean Abs Dev	11.954239
-LogLikelihood	5229.513
SSE	409758.64
Sum Freq	1207

Figure 4.9 Neural Network Model 3.2 using with only artist groups, reduced dataset.

The modified neural network model, which uses no musical attributes, achieves a similar R^2 result of 0.85 and RASE of 18.42 and is in line with the un-modified model. This would concur with the predictor screening results that indicated the likely influence of the `artist` variable. The derived group variables simple inherited the artist influence, however it is important to note that this isn't necessarily a negative outcome as the predictions are more accurate when using the artist groups. In terms of the model useability the preference would be to continue to include the musical attributes, as in model 3.2, the fit of the model is better for the `rank` values in the upper range.

For the models and data in general, variability within the musical attributes, even for tracks that achieve similar positions, complicates the relationships between musical attributes and chart performance to the point that it is not reliably possible to determine a pattern. This is evident by the low R^2 of 0.013 for model 1.1 and a similar low R^2 of 0.018 for model 1.2. Additionally, changes in the types of music that performed well over the four-year observation period similarly implies that predictions are also influenced by the relationships to other tracks in the same time frame. However, this is not always assured, there is not a clearly defined transition as to why specific types of tracks became more successful over similar tracks as measured by musical attributes, the evidence presented in chapter 3.6 shows the general change in the dataset, with no underlying reason. It is possible that musical trends are self-contained, whereby tracks that achieve success inspire other tracks that are musically similar. While outside of the scope of this thesis, this is an area where further research may be required. In practical terms even if it were possible to align an unknown artist to a known artist by similarity of musical attributes it is not possible to reliably estimate how the new artist will perform as this is related to the independent performance of all other tracks at the same time.

A similar difficulty is present in the structure of the data itself, being a daily chart, the volume of repetition is a significant problem. As identified during the analysis in chapter 3 popular songs tend to remain popular, skewing the entire dataset. While reducing the dataset to single occurrences using the best chart positions achieved per track does address this repetition, it also introduces a bias to the dataset. By looking at the highest positions achieved by the distinct tracks there is no distinction made between tracks that move gradually up the chart to a high position and those that achieve the position more quickly. Adding a running count of days the track is present in dataset or utilising the position trend variable can address this, however this again introduces bias for the most popular tracks, artists and musical attributes, contributing to a skew in the dataset that favours the most successful tracks and artists.

Any modelling efforts using the entire dataset would suffer from overfitting, while any models using the reduced dataset and derived artist attributes would suffer from bias. The volatility and variability within the dataset required careful consideration to attempt extraction of meaningful results. The conclusion must be that the musical attributes alone cannot be used as predictors; there is simple too much overlap of the attribute values for tracks that achieve a high chart positions and tracks that do not. Equally, when the artist is included as a variable, either directly or indirectly as derived variable the result is to artificially inflate the accuracy results for known data.

The overall sentiment remains that the artist is the single largest contributor to performance, which is heavily biased by the most successful artists. There is not sufficient data to fully explain this; the actual nuances of music cannot be readily transcribed to the eleven musical attributes that Spotify have chosen to use in their data. The intrinsic relationships between attributes are lost with a single average value measurement, tone, timbre and intonation are all lost in flattening the musical depth. Ultimately it is this reduction of the tracks to a series of core attributes, with little understanding of the complex relationships, that determines the less-than-optimal modelling outcomes.

6. Further Work

Several areas of interest were identified during this thesis. These relate to the average measurements of the musical attributes that Spotify utilise in their data. A more robust method of comparing musical tracks may be in understanding the variability within the tracks relative to each other, rather than comparisons of flattened average values. At present this data isn't available in the Spotify databases.

The effect of external events and recommendation algorithms is also significant, if listener patterns are being heavily influenced by external factors, then it is probable that the predictive outcomes will be affected. An understanding of the influence on the dataset may improve understanding, the addition of a categorical variable for each track on a given data indicating if the track were on a recommendation playlist would add to the dataset. This data is available via web scraping from the historical Spotify records but is not readily accessible in the main databases. Significant programming effort is required to reliable capture this data.

The addition of a genre classification may also aid in the understanding of the data set. While the changes in musical taste are loosely captured by the flattened attribute value an additional explanatory variable would again add to the dataset. This descriptive variable is present in the Spotify dataset but was not considered as part of this thesis.

References

Aguiar, L. & Waldfogel, J. (2021). “Platforms, Power, and Promotion: Evidence from Spotify Playlists”. The Journal of Industrial Economics. 69. pp. 653-691. doi: 10.1111/joie.12263.

Araujo, C., Cristo, M. & Giusti, R. (2020). “Will I Remain Popular? A Study Case on Spotify.” Conference: XVI Encontro Nacional de Inteligência Artificial e Computacional pp. 599-610. doi: 10.5753/eniac.2019.9318.

Chen, H. (2021) “The Relationship between Music Property and Degree of its Popularity in Spotify”, *BCP Business & Management*, 13, pp. 386–390. doi: 10.54691/bcpbm.v13i.112.

Gomes, I., Pereira, I., Soares, I., Antunes, M., Au-Yong-Oliveira, M. (2021). “Keeping the Beat on: A Case Study of Spotify.” doi: 10.1007/978-3-030-72651-5_33.

Heggli, O., Stupacher, J. & Vuust, P. (2021). “Diurnal fluctuations in musical preference.” Royal Society Open Science. 8. doi: 10.1098/rsos.210885.

Ingle, S. (2021). “Big Data Analytics: A Spotify Case Study”. International Journal for Research in Applied Science and Engineering Technology. 9. pp.1823-1829. doi: 10.22214/ijraset.2021.38702.

Jacobson, K., Murali, V.,Newett, E., Whitman, B. & Yon, R. (2016). “Music Personalization at Spotify.” Pp 373-375. doi:10.1145/2959100.2959120.

Kalustian, K. & Ruth, N. (2021). “Evacuate the dancefloor: Exploring and classifying spotify music listening before and during the COVID-19 pandemic in DACH countries.” Jahrbuch Musikpsychologie. 30. doi: 10.5964/jbdgm.95.

Lacognata, A. & Poole, J. (2021). “The Melodies of Politics: Assessing a Correlation Between Music Taste and Political Views with Spotify.” Journal of Student Research. 10. doi: 10.47611/jsrhs.v10i3.1898.

Ochi, V., Estrada, R., Gaji, T., Gadea, W. & Duong, E. (2021). “Spotify Danceability and Popularity Analysis using SAP”.

Savelsberg, J. (2021). “Visualizing music structure using Spotify data.“ https://www.researchgate.net/publication/353829597_Visualizing_music_structure_using_Spotify_data#fullTextFileContent

Sciandra, M. & Spera, I. (2020). “A model-based approach to Spotify data analysis: a Beta GLMM.” Journal of Applied Statistics,.pp. 1-16. doi: 10.1080/02664763.2020.1803810.

Sergeevich, K. & Mironchuk, V. (2021). “Automated systemic cognitive analysis of the top 100 songs on Spotify for 21 century.” doi: 10.13140/RG.2.2.34336.71685/1.

Sharma, Dr & Pareek, Dr & Pathak, Mr & Sakariya, Ms. (2022). “Predicting Music Popularity Using Machine Learning Algorithm and Music Metrics Available in Spotify.” Journal Of Development Economics and Management Research Studies. 09. 10-19. doi: 10.53422/JDMS.2022.91102.

South, T., Roughan, M. & Mitchell, L. (2021). “Popularity and centrality in Spotify networks: critical transitions in eigenvector centrality.” Journal of Complex Networks. 8. doi: 10.1093/comnet/cnaa050.

Appendix A. Supplementary information Spotify API

Description: Get audio feature information for a single track identified by its unique Spotify ID.

Source: https://developer.spotify.com/documentation/web-api/reference/#/operations/get-audio-features

Sample:

GET / audio-features/{id}

Id string **(**required)

The Spotify ID for the track.

Example value:" 2takcwOaAZWiXQijPHIx7B "

Response Example:

```
{
  "acousticness": 0.00242,
  "analysis_url": "https://api.spotify.com/v1/audio-analysis/2takcwOaAZWiXQijPHIx7B\n",
  "danceability": 0.585,
  "duration_ms": 237040,
  "energy": 0.842,
  "id": "2takcwOaAZWiXQijPHIx7B",
  "instrumentalness": 0.00686,
  "key": 9,
  "liveness": 0.0866,
  "loudness": -5.883,
  "mode": 0,
  "speechiness": 0.0556,
  "tempo": 118.211,
  "time_signature": 4,
  "track_href": "https://api.spotify.com/v1/tracks/2takcwOaAZWiXQijPHIx7B\n",
  "type": "audio_features",
  "uri": "spotify:track:2takcwOaAZWiXQijPHIx7B",
  "valence": 0.428
}
```

Appendix B. R Code utilised to expand the dataset

```
library(spotifyr) # Spotify API library for R

Sys.setenv(SPOTIFY_CLIENT_ID = '#################') # client id user key

Sys.setenv(SPOTIFY_CLIENT_SECRET = '#############') # client password

# Generate a secure single use key to connect to the Spotify API services

access_token <- get_spotify_access_token()

# Function to execute the API calls

MyFunction <- function(x) return(get_track_audio_features(x,authorization =
get_spotify_access_token()))

# Function to pass the distinct track URL to the API function

track_data <-cbind.data.frame(track_uri, t(sapply(track_uri$id,MyFunction,
USE.NAMES=F)))

# Function to merge the newly fetched track data with the original data

Irl_2018_2021_track_data <- merge(Irl_2018_2021,track_data,by = 'id')
```