What Metrics Affect a Song's Popularity on Spotify:

An Analysis Using Multiple Regression

Ryan James Calabio

5 Sept. 2023

Abstract

The topic of this paper is understanding what variables commonly occur in the most popular songs on Spotify. To determine this we will use multiple regression analysis in Excel using XLStat. This paper draws conclusions using information including Spotify's API statistics about artists and their songs that are both accessible to the public. Among these statistics are ones that are common like BPM, key, and streams. Spotify also creates their own metrics to help quantify songs like danceability percentage and energy percentage. For many artists, Spotify is a platform to reach new fans, as their algorithm for music discovery allows users to discover music that may not have a large audience yet. One of the most important metrics for users and artists are streams on a song; this is used to measure or prove whether or not the song could be potentially good / popular. In this paper, I am going to use regression analysis to measure what metrics could potentially have an influence on the number of streams. Our goal is to create a regression equation that you could input into features of a new song, and then potentially predict its popularity. This regression equation helps us understand what variables one can focus on to increase the popularity of their music. From our results variables related to time were important like release year and month, which had the strongest positive correlations with streams.

Introduction

Spotify as a platform is revolutionary for many smaller artists. The reason for this is because the Spotify algorithm works to introduce smaller more unknown artists to its users, giving a lot of artists a chance that they may never have if they were to utilize other platforms.

In this paper, we will use regression analysis to measure the relationships between certain metrics and the number of streams. This can potentially predict the number of streams a song may get by inputting metrics related to its release and current state. This could also be used to understand what metrics have a strong correlation with streams in order to assist artists in understanding what metrics they should focus on if getting more streams is a goal of theirs.

Objectives

My objective with this is to define what metrics can be useful for someone to grow on Spotify using the data available to me. In order to do this best, I have looked to different related literature that has already been written to understand relevant topics more. Spotify has its own section on its website that has papers in which they share their own internal research on all sorts of technical topics like machine learning, search & recommendations, and human-computer interaction. Two papers stand out to me as relevant to the experiment we are conducting in this paper: “Automatic Music playlist Generation via Simulation-based Reinforcement Learning” and “Mostra: Balancing multiple objectives for music recommendation”. These are music recommendation focused research papers that are very relevant to the number of streams an artist will get on Spotify, as the artist first needs to be shown to users by Spotify. In the former research paper on playlist generation, they discuss the challenges related to utilizing “satisfaction metrics” in order to create optimal playlists per user (Tomasi, Cauteruccio, Kanoria, Ciosek, Rinaldi, & Dai, 2023). The latter research paper discusses MOSTRA (Multi-objective Set Transformer), which focuses on four separate metrics that it wants to optimize its search engine for. These metrics are SAT, Discovery, Exposure, and Boosting. SAT is whether a user completely listens to a song. Discovery is whether a user has never listened to a song before. Exposure is whether the song belongs to an emerging artist, in which case they would like to promote them as they believe that “this makes Spotify a sustainable platform to support a wide variety of creators” (Bugliarello & Lalmas, 2023). Boosting is whether or not the song belongs to a group that platform is interested in promoting. An example of this is whether or not an artist belongs to a cultural group that is being given special attention at that time. The purpose overall is to provide a regression equation that gives artists that use Spotify some understanding of what variables and commonalities many of the most popular songs on the platform have in common.

Methodology

In this experiment, I aim to discover a relationship between Spotify API metrics and the popularity of songs. In order to discover potential relationships, I decided to use a multiple regression analysis in order to discover correlations between certain metrics and stream numbers. This analysis was done in Excel using XLStat on a dataset that included 953 of the most popular songs on Spotify. In this experiment, the stream numbers for a song are the dependent variable in this experiment, and the other metrics provided by the data set are the independent variables that will be changed to measure the dependent variable. In this study, we use a total of 14 Spotify metrics that are provided to use in analysis. They are the following: artist count, release year, release month, number of Spotify playlists that the song is included in, number of Spotify charts that the song is included in, BPM (beats per minute), modes (major or minor), danceability, valence, energy, acousticness, instrumentalness, liveness, and speechiness. The latter 7 are Spotify generated metrics that they generate on their platform. The data was then imported from Kaggle into Excel. The columns were reduced down to ones that are only focused on Spotify and usable for the regression analysis. Within Excel, the regression analysis used will not accept non-numeric character or nulls. Any column that could not be properly coded into a quantitative measure or that had nulls were removed. Once the analysis is run in Excel, the output is shown as a table of coefficient values along with the metric that they are associated with (see Appendix A).

Empirical Results

Once we input the data into the regression model and learn the coefficients, this leaves us with the following regression equation:

$$\begin{aligned} Streams = {} & -30{,}838{,}637 \ast (\#\ of\ artists) + 516{,}6491.1 \ast (year\ released) + 9{,}408{,}304.46 \\ & \ast (month\ released) + 58{,}248.27 \ast (\#\ of\ spotify\ playlists) + 256{,}504.55 \\ & \ast (bpm) + 2{,}785{,}557.65 \ast (mode) - 340{,}588.14 \ast (danceability\ \%) \\ & + 278{,}141.31 \ast (valence\ \%) - 1{,}607{,}375 \ast (energy\ \%) + 686{,}022.99 \\ & \ast (acousticness\ \%) - 1{,}991{,}528.9 \ast (instrumentalness\ \%) + 87{,}200.22 \\ & \ast (liveness\ \%) - 2{,}374{,}900.3 \ast (speechiness\ \%) - 1.018 \ast 10\text{^}(-10) \end{aligned}$$

We can then order the correlation strength with streams in our data by organizing each factor by how large the coefficient associated with that metric is. That leaves us with this chart with metrics by coefficient size in descending order

Coefficient Strength by Metric

Metric	Coefficient
Month Released	9408304.46289843
Year Released	5166491.09814717
# of Spotify Charts Included In	3155057.01592243
Musical Mode	2785557.65211777
Acousticness %	686022.991938621
Valence %	278141.312895902
BPM	256504.552139923
Liveness	87200.2167213899
# of Spotify Playlists Included In	58248.2746062717
Danceability %	-340588.138308661
Energy %	-1607375.19567529
Instrumentalness %	-1991528.86919476
Speechiness %	-2374900.33371411
Artist Count	-30838636.5313186

Looking at this table, we can see that there are both positive and negative coefficients. The strongest correlation is release month, and streams seem to increase as the months go on in the year. More recent songs seem to be correlated with higher stream levels as well. The third highest positive coefficient is the number of Spotify charts that the song is included in. This is reasonable because the charts are supposed to track songs that are popular in certain groups of people (i.e. Top Songs in Japan). Musical mode is next in correlation strength, which is coded with 1 for Major and 2 for minor. Since there is a positive correlation, it means minor seems to occur more often in popular songs. Minor is sadder, which may suggest that there is sadder music in the popular songs of today. Acousticness tracks how acoustic a song is, so there may be more of a trend towards more acoustic genres, like country or guitar driven tracks in recent years. Valence is Spotify’s way of tracking happiness vs sadness within a song. This is interesting since the correlation with modes seem to suggest opposite correlations; however, modes in minor do not necessarily mean that a song is a sad one. A higher BPM (Beats Per Minute) means a faster song, and the positive correlation suggests faster songs are more popular. Liveness is a measure of the audience within a recording, which has a positive coefficient. The last positive coefficient is the number of Spotify playlists that a song is in. This is expected, seeing that the more playlists that a song is in, the more likely it is that a user would replay the song. Danceability, energy, instrumentalness, speechienss, and artist count were all negatively correlated.

Conclusions

As Spotify grows more and more as a platform, the understanding of what portions of music to focus on will become more and more valuable. By determining correlations between stream numbers of popular songs and various Spotify metrics, we may gain insight into what could potentially affect a song's popularity. Nine out of the fourteen coefficients are positive, implying that the larger the number related to the metric, the larger the number of streams the song will have garnered. The remaining five coefficients are negative, which implies that the higher the unit of measurement for these metrics, the song will be less popular and accrue less Spotify streams.

Discussions

There are a few implications for the conclusions that we derived from our regression analysis. The most interesting thing that we derived from this is that the two strongest positive correlations are related to release time. A lot of the positive correlations made sense, and there was a lot to derive from the negative ones too. An interesting thing to not between coefficients is that valence and mode were almost suggesting different things. With mode coefficient suggesting a move towards sadder songs and valence towards positive. However, the way these metrics measure that are done in different ways. Since happiness is a somewhat subjective measure, measure of happiness may vary. Doing more tests focused on more specific variables may provide more accurate results too, as there are a considerable number of independent variables that this multiple regression analysis takes into consideration. It would also be interesting to do some feature engineering to find the impact of variable like number of Billboard Hot 100 hits or more information related to release schedule / promotion.

Future Research

In the future, it would be interesting to do an analysis on more than just the most popular songs, but also include as much imported data from Spotify's API as it will allow us to download. Doing an analysis over historical trends to see if the metrics that affect popularity of certain songs (i.e. measured by streams) can help identify trends in what correlates with higher levels of streams per song.

References

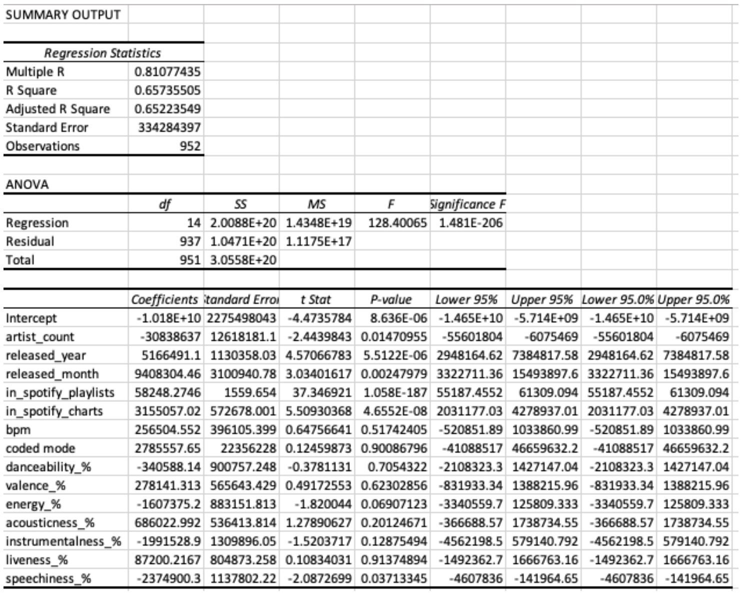
Elgiriyewithana, N. (2023, August 26). *Most streamed Spotify Songs 2023*. Kaggle. https://www.kaggle.com/datasets/nelgiriyewithana/top-spotify-songs-2023

Tomasi, F., Cauteruccio, J., Kanoria, S., Ciosek, K., Rinaldi, M., & Dai, Z. (2023, July 24). *Automatic Music Playlist Generation via simulation-based reinforcement learning*. Spotify Research. https://research.atspotify.com/2023/07/automatic-music-playlist-generation-via-simulation-based-reinforcement-learning/

Bugliarello, E., & Lalmas, M. (2023, March 29). *Mostra: Balancing multiple objectives for music recommendation*. Spotify Research. https://research.atspotify.com/2022/04/mostra-balancing-multiple-objectives-for-music-recommendation/

Appendices

Appendix A



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.81077435
R Square	0.65735505
Adjusted R Square	0.65223549
Standard Error	334284397
Observations	952

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regression	14	2.0088E+20	1.4348E+19	128.40065	1.481E-206
Residual	937	1.0471E+20	1.1175E+17		
Total	951	3.0558E+20			

	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-1.018E+10	2275498043	-4.4735784	8.636E-06	-1.465E+10	-5.714E+09	-1.465E+10	-5.714E+09
artist_count	-30838637	12618181.1	-2.4439843	0.01470955	-55601804	-6075469	-55601804	-6075469
released_year	5166491.1	1130358.03	4.57066783	5.5122E-06	2948164.62	7384817.58	2948164.62	7384817.58
released_month	9408304.46	3100940.78	3.03401617	0.00247979	3322711.36	15493897.6	3322711.36	15493897.6
in_spotify_playlists	58248.2746	1559.654	37.346921	1.058E-187	55187.4552	61309.094	55187.4552	61309.094
in_spotify_charts	3155057.02	572678.001	5.50930368	4.6552E-08	2031177.03	4278937.01	2031177.03	4278937.01
bpm	256504.552	396105.399	0.64756641	0.51742405	-520851.89	1033860.99	-520851.89	1033860.99
coded mode	2785557.65	22356228	0.12459873	0.90086796	-41088517	46659632.2	-41088517	46659632.2
danceability_%	-340588.14	900757.248	-0.3781131	0.7054322	-2108323.3	1427147.04	-2108323.3	1427147.04
valence_%	278141.313	565643.429	0.49172553	0.62302856	-831933.34	1388215.96	-831933.34	1388215.96
energy_%	-1607375.2	883151.813	-1.820044	0.06907123	-3340559.7	125809.333	-3340559.7	125809.333
acousticness_%	686022.992	536413.814	1.27890627	0.20124671	-366688.57	1738734.55	-366688.57	1738734.55
instrumentalness_%	-1991528.9	1309896.05	-1.5203717	0.12875494	-4562198.5	579140.792	-4562198.5	579140.792
liveness_%	87200.2167	804873.258	0.10834031	0.91374894	-1492362.7	1666763.16	-1492362.7	1666763.16
speechiness_%	-2374900.3	1137802.22	-2.0872699	0.03713345	-4607836	-141964.65	-4607836	-141964.65

Figure 1: This is the output provided by XLStat