What are the factors that increase streams on Spotify?

TEAM POLIWAG

Abdussamet Yaşar
201180761

Gizem Değirmencioğlu
201180763

Ender Güz
201180760

ABSTRACT

With Spotify becoming the center of music worldwide, it has also been seen as a tool for becoming famous and making money. This report, which includes analyses on how to increase the number of streams on Spotify, is a must-read for anyone who wants to get their songs heard by more people and make money doing so. The main question was answered through the creation of sub-questions and answering them through the report. The report includes analysis of what types of music are listened to more, which countries listen to more music, the relationship between the number of followers and the number of streams, analysis of the words in the names of popular songs, analysis of song duration, and analysis of dancebility, energy, tempo, and musical positiveness of popular songs, as well as visualizations of these analyses.

1. INTRODUCTION

In order for a song to be listened to by more people on Spotify, it must have certain characteristics. Those who will produce content on Spotify must understand and solve the connection between songs and people, and produce content accordingly, in order to increase the number of streams of their songs. For a song to be popular, first and foremost, the person singing the song must have a good voice. There is nothing that can be done about this inherent characteristic, but knowing what types of music are listened to the most is an advantage in increasing the number of strams of a song. Making songs in the most listened to or popular genres will increase the audience potential of the song. People who want to increase the number of streams of their songs should pay attention to the words in the names of their songs. An analysis showed that emotional words like "love" and "live" and words like "me" and "you" that can be used in relationships stand out. Putting a song in the Spotify list of the countries where it is listened to the most will also increase the potential audience. Knowing the values of tempo, positivity, and danceability of music also helps the person making the song understand the preferences of people. If a song has a high tempo and high energy, it is more likely to be listened to at a party or while exercising. On the other hand, a song with lower energy and lower tempo may

be more suitable for relaxing or calming activities. Knowing these preferences of people can help content creators tailor their songs to specific audiences and increase the chances of their songs being listened to more. Additionally, analyzing the danceability of a song can also give insights into what kind of activities it may be suitable for. By understanding these preferences and creating content that aligns with them, content creators can increase the chances of their songs being listened to more on Spotify. In addition to all these, analyzes were made on many different topics and as a result of these analyzes, some results were obtained by making use of Linear Regression and Random Forest Regression.

Keywords

"Python, data science, data analysis, pandas, matplotlib, seaborn, machine learning."

2. MOTIVATION

We have chosen this topic because Spotify is one of the most popular song streaming platforms and many people listen to music every day. Some people just listen to songs, while others make money by sharing their songs on this platform. Some songs become very popular in the first few days they are uploaded to Spotify, while others never achieve this popularity. So what are the factors that determine this? In addition to God-given voice, what are the factors that have an impact on popularity and why do popular songs receive more attention on this platform than others? Our motivation is to analyze the data of popular songs and prepare a report on what people should pay attention to in order to become famous and make money.

3. DATASETS

For this report, we used 9 data sets. Some of the data sets were not used directly because we decided that other data sets containing the data of those data sets were more comprehensive, but we still used those data sets for observation. We obtained these data sets from Kaggle. When selecting the data sets, we paid attention to whether they had the necessary data to answer the main question and sub-questions we wanted to answer, and also made sure that the data sets had a sufficient number of data points. The data sets mainly contain Spotify Top100, Top200 data. These data are different by year and country. During the analysis, we used streams, followers, genres, titles, durations, energy, danceability, musical positivity, tempo and popularity data found in the data sets. The links to the data sets are below.

1. https://www.kaggle.com/datasets/dhruvildave/spotify-charts

2. https://www.kaggle.com/datasets/bricevergnou/spotify-recommendation

3. https://www.kaggle.com/datasets/leonardopena/top50spotify2019

4. https://www.kaggle.com/datasets/geomack/spotifyclassification

5. https://www.kaggle.com/datasets/zaheenhamidani/ultimate-spotify-tracks-db

6. https://www.kaggle.com/datasets/muhmores/spotify-top-100-songs-of-20152019

7. https://www.kaggle.com/datasets/jfreyberg/spotify-artist-feature-collaboration-network?select=nodes.csv

8. https://www.kaggle.com/datasets/sashankpillai/spotify-top-200-charts-20202021

9. https://www.kaggle.com/datasets/iamsumat/spotify-top-2000s-mega-dataset

4. METHODOLOGY

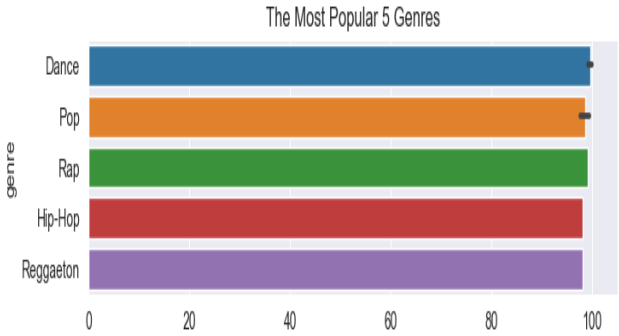
In this research, we used pandas, numpy, seaborn, matplotlib, sklearn libraries. It was used to read the Pandas library database. We used the Numpy library to convert the values in the data set into a format suitable for linear regression. The heat map feature was used in the Seaborn library. The heat map allowed us to understand

the level of relationship between the variables. With the matplotlib library, we provided a meaningful visualization of the values in the dataset. For example barchart, pie chart, line chart, bar-line chart, bubble chart, scatter. visualizations were used. A bar chart is a way of summarizing a set of categorical data. You can easily see the change of data according to time in Barchart. This possibility is not available in the pie chart. Our purpose in using the scatter is to see the distribution of values better. With the double y-axis in bar-linechar, we can visualize multiple variables on the same graph. In the Sklearn library, we used linear regression and metrics structures. Thanks to Metrics, there are structures where we can analyze how accurately the results obtained as a result of machine learning are predicted or whether they consist of a false prediction. We used it to analyze the amount of measurement errors in linear regression, which we use as machine learning. Linear regression was also used as machine learning. The reason for this is that we want to reach a numerical data rather than making a classification. Linear regression is a popular and uncomplicated algorithm used in data science and machine learning

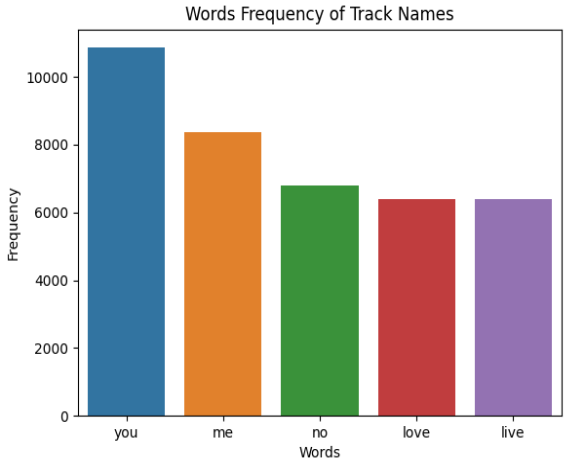
The final model we decided to use is a meta-model which contains 2 Linear regressors and a Random Forest regressor. We used the most impactful attributes on our dataset to train our model. At the base we have Linear regressor and Random Forest algorithms. At the top we have another Linear regressor using their predictions as input.

Instead of directly trying to predict the popularity value of a song we are trying to predict if the song would have a popularity over 74, which is the popularity average of top 100 songs from 2010 to 2019.

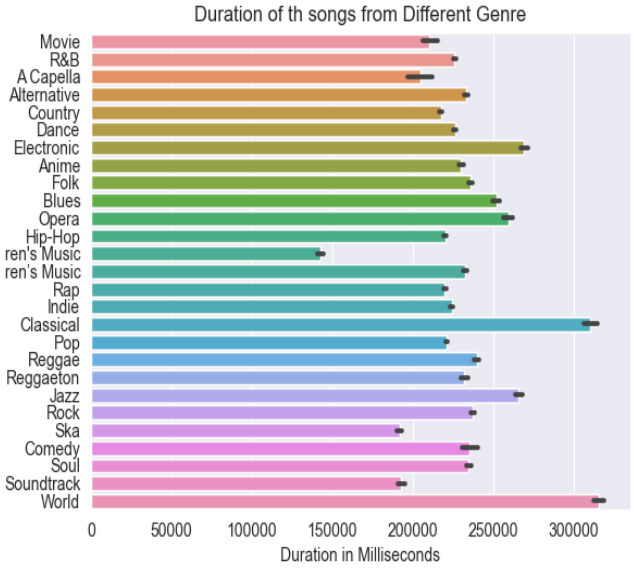
5. EXPERIMENTS & RESULTS



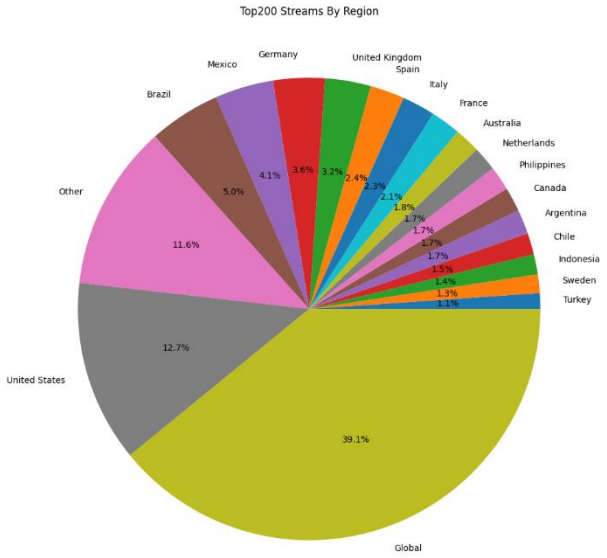
When we look at the top 5 most popular genres, we see that these are dance, pop, rap, hip-hop, and reggaeton. The first four genres are well-known to everyone, but reggaeton is not. This is because reggaeton is popular in Central America, Spain, Portugal, and especially Latin America. Although it is not a music genre accepted by the entire world, it has entered the list of the most listened to genres thanks to its popularity in certain regions. This shows us the importance of geography in addition to song genres. Looking at this figure, it can be seen that the top 5 music genres are generally more energetic music genres. The fact that the most popular genre is dance also confirms this. In addition to individual music listening, music plays an important role in the service and entertainment sectors. The music in the first two ranks is the music used in these sectors. Therefore, this factor should also be considered in the selection of music genre in addition to individual music pleasure.



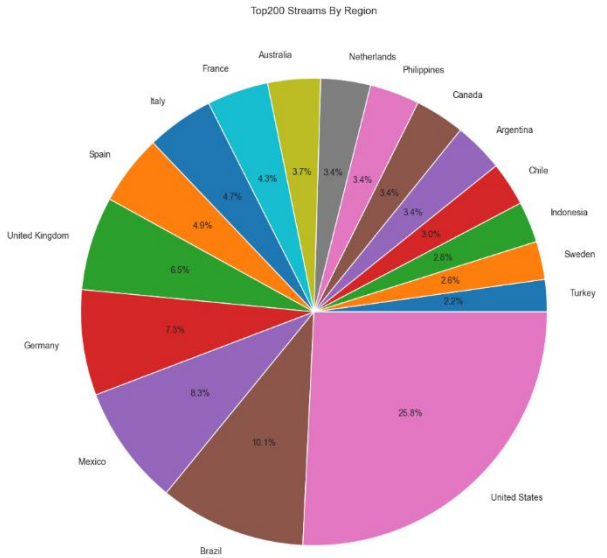
The top 5 most frequent words in popular song titles are shown in the above graph. These are you, me, no, love, and live. The words in the first and second ranks are words that express dual relationships. The other 3 words are words that appeal more to emotions. When we evaluate the 5 words together, we see that the titles of the songs are about dual relationships and experienced emotions. Using the words in the top 5 in a song contributes to more people discovering that song and thus helps increase the number of times the song is listened to.



The above figure shows the average duration of song genres. The Classical and World genres are the ones with the longest duration and the shortest is Ren's Music. These are extreme values and we see that the duration of all genres except a few is between 200-250 seconds, except for these. When we look at the duration of the most popular music genres listed in the above graph, they are all slightly over 200 seconds and have very similar durations. The average duration of song genres gives us information about what the duration of the music to be made in that genre should be after the song genre is selected.

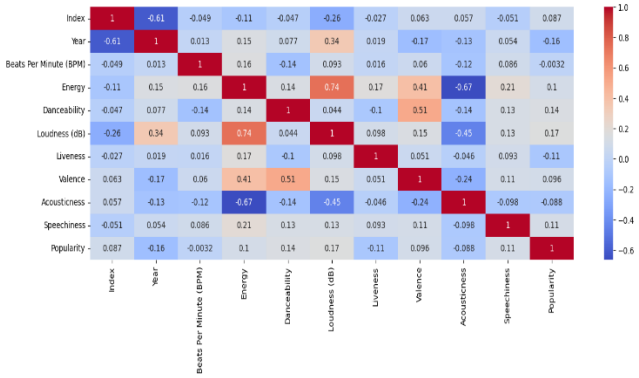


In the above graph, we see the number of times the Top200 list is listened to by country. We see that Global has a very dominant presence in the list. We did not consider countries with less than 1% share and grouped them under the heading "Other", but when all of them are added together, they have the third largest share. The main graph we will evaluate here is not this one, but we used it to make the graph below more understandable.



This pie chart has been created without including Global and Others data. The most striking aspect

of the graph is that China, which has a population of 1.4 billion, is not included in the pie chart. This is not because the number of Chinese people listening to music is low. China is not among the countries where Spotify can be registered, so we cannot see China in the graph, but the fact that India has a very low share despite its population shows that they really listen to very little music. The graph indicates that the United States is the most suitable country for publishing music on Spotify.



heatmap is a table showing the relationships of components to each other. Each square shows the correlation between the variables on each axis. Correlation ranges from -1 to +1. Values close to zero indicate that there is no linear relationship between the two variables. Values close to 1 represent a linear relationship, while values close to -1 indicate an opposite relationship.

6. CONCLUSION

In this project, we aimed to create a roadmap for people who publish songs on Spotify to increase the number of plays of their songs. As a result of this project, we obtained the following results.

- The most common words in song titles are words that appeal to people's emotions and using these words can increase the streams.

- Except for a few song genres, all song genres have an average duration of between 3 and 4 minutes.

- The United States is the country where the most music is listened to. (China and India are not included in the analysis because Spotify is not used in these two countries).

- The ratio of population to number of plays is similar.

- There is a linear relationship since the value in the intersection of Energy and Loudnessis very close to 1. Otherwise there is an opposite relationship between Energy and Acusticness because the value is very close to -1

REFERENCES

[1]https://www.ibm.com/topics/linear-regression#:~:text=Resources-,What%20is%20linear%20regression%3F,is%20called%20the%20independent%20variable.

[2]https://www.javatpoint.com/machine-learning-random-forest-algorithm

[3]https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

[4]https://matplotlib.org/stable/tutorials/introductory/quick_start.html

[5]https://numpy.org/doc/stable/user/absolute_beginners.html

[6]https://seaborn.pydata.org/tutorial/introduction

[7]https://pandas.pydata.org/docs/user_guide/10min.html

[8]https://www.kaggle.com/code/kierondrumm/spotify-charts-exploration

[9]https://www.kaggle.com/code/kevinkwan/spotify-eda

[10] Hartwig, F., & Dearing, B. E. (1979). Exploratory data analysis (No. 16). Sage.