Proposal

For this project we will be trying to figure out what factors make music artists popular (by total song listens) on Spotify. The answer to such a problem could be used by music publishers in order to figure out potential artists they should sign before artists become extremely popular. This will allow the publishers to get contracts that are more likely to make money for them. Another use of the answer to the problem is so that artists themselves know what to focus on if they want to become popular.

Since Spotify tracks and artists data are widely studied, there are several great ways of obtaining our data. We plan on directly making request calls to the Spotify API for retrieving the most up-to-date data along with searching within large dataset collections such as Kaggle. Through some research, we have found the “Spotify Track Dataset” created by Maharshi Pandya to contain most of the attributes for track data that we need. This dataset has also been used and implemented by other users, and will be helpful for benchmarking and comparing our results. For the artists, the dataset “Spotify Artist Metadata Top 10k” by Jack Harding documented related information for each of the most popular songwriters and artists on Spotify, including their age, sex, location, and ranked popularity. We aim to combine these two datasets together to address the correlation between the success of an artist and their published records. More data sources might be added later on in the project for feature engineering.

For the data source of this project, we would like to explore the correlation between the expansion of artists and their album release patterns and conditions. Then we plan to use several datasets from Kaggle and combine them to get more unique variables. We plan to explore the data from Spotify because we believe that most young adults use Spotify and the data from this platform would be more reliable.

For this project, we will be using hypothesis tests and linear regression analysis. The hypothesis is for determining if a factor is important to the singer, and if some data mean and variance agree with our assumptions. The linear regression analysis is to see if the data fit into the model we proposed and for making future predictions.

Our plans for this project are to create a comprehensive guide with valid evidence to suggest the best plan of action for anyone hoping to succeed on spotify. With enough analysis on every feature related to music success, we hope to be able to gauge the importance and consistency with which they impact an artist's success. We hope to be able to understand how certain artists succeed while others fail and simultaneously gather insight on the differences between them.

1. Why is this problem important?

For the project we will be trying to figure out what factors make music artists popular (by total song listens) on Spotify. The answer to such a problem could be used by music publishers in order to figure out potential artists they should sign before artists become extremely popular. This will allow the publishers to get contracts that are more likely to make money for them. Another use of the answer to the problem is so that artists themselves know what to focus on if they want to become popular.

1. Where do you plan on acquiring the data from? What are the data sources for this?

Since Spotify tracks and artists data are widely studied, there are several great ways of obtaining it. We could directly make request calls to the Spotify API for retrieving the most up-to-date data, or search within large dataset collections, such as Kaggle. Through some research, we found the “Spotify Track Dataset” created by Maharshi Pandya to contain most of the attributes for track data that we need. This dataset has also been used and implemented by other users, and will be helpful for benchmarking and comparing our results. For the artists, the dataset “Spotify Artist Metadata Top 10k” by Jack Harding documented related information for each of the most popular songwriters and artists on Spotify, including their age, sex, location, and ranked popularity. We aim to combine these two datasets together to address the correlation between the success of an artist and their published records. More data sources might be added later on in the project for feature engineering.

For the data source of this project, we would like to explore the correlation between the expansion of artists and their album release patterns. Then we plan to use several datasets from Kaggle and combine them to get more unique variables. We plan to explore the data from Spotify because we believe that most young adults use Spotify and the data from this platform would be more reliable.

1. What type of analyses do you propose for the dataset?

For this project, we will be using hypothesis tests and linear regression analysis. The hypothesis is for determining if a factor is important to the singer, and if some data mean and variance agree with our assumptions. The linear regression analysis is to see if the data fit into the model we proposed and for making future predictions.

1. What are your expected outcomes for the project?

Our plans for this project are to create a comprehensive guide with valid evidence to suggest the best plan of action for anyone hoping to succeed on spotify. With enough analysis on every feature related to music success, we hope to be able to gauge the importance and consistency with which they impact an artist's success. We hope to be able to understand how certain artists succeed while others fail and gather insight on the differences between them.

**References**

Harding, J. (2020). *Spotify Artist Metadata Top 10k*. https://www.kaggle.com/datasets/jackharding/spotify-artist-metadata-top-10k

Pandya, M. (2022). 🎹 Spotify Tracks Dataset [Data set]. In *Kaggle*. https://doi.org/10.34740/KAGGLE/DSV/4372070

<https://www.kaggle.com/datasets/raj713335/tbo-hotels-dataset?resource=download>

<https://www.kaggle.com/datasets/lesageethan/lego-model-information-and-return-on-investment>

<https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview>

<https://www.kaggle.com/datasets/rupindersinghrana/airbnb-price-dataset>

<https://www.kaggle.com/datasets/patricklford/water-and-air-quality>

Spotify:

<https://www.kaggle.com/datasets/adnananam/spotify-artist-stats>

<https://www.kaggle.com/datasets/jackharding/spotify-artist-metadata-top-10k>

<https://www.kaggle.com/datasets/nelgiriyewithana/top-spotify-songs-2023>

<https://www.kaggle.com/datasets/hedizekri/top-charts-artists-country>