## **FP3 – Data Analysis Plan**

**Overview**

* **What is the topic of your project?**

In our FP we will be predicting a song’s popularity based on the track features.

* **Why is it important to tackle this problem in your project?**

Knowing what music to release can be difficult for artists and producers, so this project can help users explore how popular a given song would be today. Artists would then know what kind of music is the most popular.

* **What is your dataset about? How many samples (rows) does it have?**

Our dataset is a random sample of 10,000 songs from the Spotify API.

* **What are your features? What is your target variable?**

The features are danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo, duration, days since release and artist popularity. Our target variable is popularity, which Spotify calculates as an integer between 1 and 100 based on the total number of plays and how recent those streams were.

**Questions/Hypotheses**

* **What questions have you formulated regarding your dataset? These questions may tie back to the problem statement (just present them as questions this time). You should have at least one question tapping into the comparison of multiple ML algorithms in predicting your target variable from your feature variables.**
  + What features are important in determining popularity for a given song? Which features have comparatively little impact on determining popularity?
  + Why are certain songs more popular than others?
  + What types (genre, region, artist, etc.) are the most popular? What types are least popular?
  + Which ML algorithm is most accurate in predicting song popularity?
  + How does the average popularity for a given group (genre, artist, release date) change over time?
* **What hypotheses are you going to test? You should have at least one hypothesis test in your final report.**
  + Songs that are released within the past 90 days will be categorized as popular.
  + Songs that have an artist popularity above 85 will be categorized as popular.
  + Songs that are from the pop genre will be categorized as popular.
* **What hypotheses have you formulated regarding potential causal relationships among some of the variables in your dataset?**
  + Songs that are released closer to today will be more popular than songs that were released further in the past.
  + Songs that have a medium-high value for “energy” will be more popular than other songs.
  + Songs that have a high value for “danceability” will be more popular than other songs.
  + Songs that have a higher artist popularity will be more popular than other songs.

**Data Analysis Plan**

In this section, describe some of the analyses you are planning to conduct. This is tentative and can change later, but you should have a clear idea of what to do for your ML problem. Answer the following questions:

* **Is your project tackling a classification problem or a regression problem?**

We are trying to predict the popularity of songs from the Spotify library. Each song comes with a popularity score, but we are also planning on discretizing this score by binning the data into several categories. This is so that we can implement classification as well as regression algorithms.

* **Which ML algorithms are you planning to use?**

For regression we would implement ridge, lasso, and standard regression algorithms as well as the KNN, SVM, Decision Tree versions of regression. We would likewise use KNN, SVM, Decision Tree, and Naive Bayes for classification analysis.

* **Which feature extraction/engineering techniques are you planning to use?**

For classification we would need to discretize the data. We also plan to extract the features using the Spotify API.

* **What are some variables that would be useful to visualize? What types of visualizations (e.g., scatter plots, bar graphs, etc.) are you planning to make? Name at least two different visualizations and explain why you want to produce them.**

We definitely want to use scatter plots to visualize correlation between target and feature variables. Additionally it would be helpful to visualize metrics using bar graphs to compare performance of different models. In particular, variables such as the correlation between artist popularity and song popularity, or days since release to song popularity would be useful to see. These visualizations would help determine which features have the largest impact on song popularity.