Question: What attributes are most useful in predicting the performance of a song on the Spotify platform?

What is an observation:

In the original dataset, each observation contains twenty five variables. Five are categorical: track title, artist name(s), the key of the song, the mode (major or minor), the Spotify pages' url, and our created streams category variable. The pages' URL is irrelevant to our analysis, the track name or artist are good key values for each observation however in terms of regression neither are consequential. The mode and key values contain information about the song itself and have the potential to be predictors contingent on our model-building process. The rest are numerical variables and include: our response variable, number of streams, the artist count (how many artists worked on the song), variables for how many playlists the song is included and its place on the charts in four different platforms: spotify, apple, deezer, and shazam, the BPM (Beats per minute), and percentages of danceability, valence, energy, acousticness, instrumentalness, liveliness, and speechiness some of which are measured subjectively and others objectively but regardless contain information about the song's characteristics and may serve as useful predictors. Time related release dates (the year and month) could serve as a categorical or numeric variable to track trends in the event it is used for our regression analysis portion.

During the initial cleaning process we created a streams category (which delineates between Low Medium High and Very High) and intuitively chose some variables which we found less compelling. This new variable turns the question into one of classification, and we decided that regression is a more comprehensive method to go about answering our research question. After preparing for this portion of the project we reconsidered which variables we ought to keep for the purpose of model building and especially regression. Much of our revised cleaning process will be determining which of these variables are the most important in predicting how successful a song will be, not through intuition alone but systematically using that entirety of the data we have. Some basic exploratory data analysis methods might also omit certain variables from our consideration. Another potential step for selecting these variables could be a variable importance plot produced from the random forest method.

Supervised vs. unsupervised learning:

After performing EDA, we found that our data suffers from a few cases of multicollinearity. To help combat against this, we plan on using unsupervised learning on the front end of our model building to make our regression more robust against these problems. Furthermore, unsupervised learning may help with discovering natural patterns and, in turn, reduce dimensionality through techniques such as PCA.

After using unsupervised learning to reduce dimensionality, we plan on using supervised learning in our regression model in an attempt to minimize error in predictions and train the model towards a specific goal (in this case predicting the number of streams). Furthermore, we chose to do regression over classification because our goal is to find the statistics that reflect trends or patterns associated with the popularity of songs, which is described by the number of streams a song has.

Models & Algorithms:

We plan on using a couple of different methods to perform our analysis. First, we will use LASSO to assist in reducing multicollinearity problems and prevent overfitting on the training data through regularization. Additionally, this will help keep better interpretability of the model through techniques such as ridge regression. The automated variable selection will allow us to simplify our model and enhance prediction accuracy.

We are also considering using Random Forest. Random Forest can reduce overfitting of our model by averaging predictions across multiple decision trees. Averaging our decision trees will smooth out extreme predictions and should lead to more generalizable results. Further, using the Random Forest method is beneficial for handling both numerical and categorical data well. Similarly to LASSO, Random Forest will assist us in choosing the most influential variables and will help build a simpler and more effective model.

Success:

Our research question aims to find out which aspects of a song are most useful in predicting a song's success on the spotify platform. Our approach is to create multiple models, through LASSO, random forests, and other forms of variable and regression analysis to create the best predictive model we can. A success then involves an iterative process wherein each model improves upon another version or is ruled out mainly based on its R-squared and adjusted R-squared values until we find the best model we can create from these data. R-squared and adjusted R-squared are useful starting points, but we will need to be cautious about solely improving predictiveness and holistically compare each of these models.

Issues:

We anticipate that we will face several issues in our regression analysis, mostly tied to the dataset itself. Firstly, our dataset has 95 null values in the column 'key' and 50 in the column 'in_shazam_charts'. While there are several ways that we could reconcile this, we likely will have to pull in additional data sets to fill in any null data, or provide additional data when needed. In turn, this will force us to clean an additional dataset; however, it should be far easier since we have already done so with a similar dataset. Then we will have to join the two datasets, checking to make sure that we do not interfere with the integrity of the data during the process.

Furthermore, our data is skewed significantly towards more recent release years. In turn, this will likely lead to a non-representative sample, which keeps the model from accurately capturing trends in the number of streams over time. Furthermore, this may lead to further bias towards recent trends, and make it harder to isolate the effects of release timing on overall streaming popularity. To reconcile this, bringing in an additional data set may help us analyze the effect of release year and/or potential changes in trends across time.

Lastly, the data that we originally collected, cleaned, and prepared for analysis, no longer fits our plans. As our focus has shifted since first landing on this data set, we will likely have to go back through the cleaning process for our data to ensure that we include all of the necessary variables in our final dataset, in addition to each of these variables being prepared for overall regression analysis.

If our initial approach fails, we may be forced to reconsider the structure of our dataset, and whether or not it is actually usable in answering our research question. Furthermore, to solve this problem, we would have to follow the steps listed above in order to reconcile our data with additional datasets to fill in the gaps.

Feature Engineering:

When looking at a corrplot of our numeric variables, there are only two relationships that may pose a problem for our model. Firstly, the relationship between released_year and streams is highly positively correlated. This may partially be due to a significantly higher number of observations for more recent years, compared to the past. Furthermore, accousticness_% and energy_% have a relatively high negative correlation. In order to deal with these, we would conduct PCA to determine which of these components explain the greatest percentage of the variance, in order to reduce the dimensionality of our model. Additionally, one-hot encoding may be useful for converting certain categorical variables, such as artist_name. However, it could be best used to help regress on the variables key and mode, because one-hot encoding artist_name will likely greatly expand our dataset, whereas key and mode are much more concise.

Results:

We have decided on four different outputs that will summarize the results of our model. We will first use a table of regression coefficients in order to display the final variables in our model and assist in increasing the interpretability of our model. We could also include p-values in this table if we wanted to display statistical significance. We will then use both R-squared and Adjusted R-squared to quantify the proportion of the variance in y explained by the model. We will use Adjusted R-squared because we want to be able to control for the number of variables in our model if we end up with many variables. Furthermore, we will use R-squared and Adjusted

R-squared to compare the predictability of our models with each other, to determine the best model to use in our final regression. Finally, we will include RMSE in order to have a measure of the predictive error in our model.