Statistical Analysis

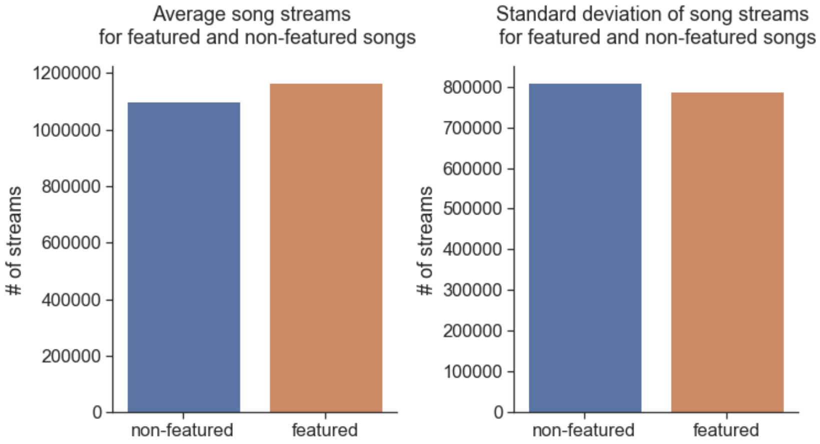
Prior to building the predictive model, the following statistical tests were performed to help

A) answer questions stemming from the EDA

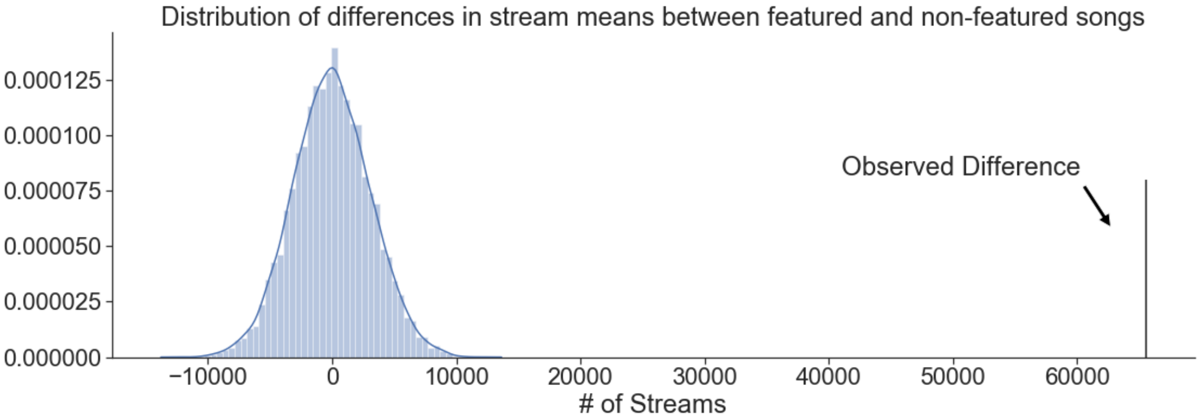
B) answer questions regarding the features

1) Featuring another artist on a song leads to an average of 5.98% or ~66k more streams. Is this observation statistically significant?

To determine if the observed difference in means is statistically different, 10,000 bootstrap samples of adjusted means were taken from the two groups (featured/non-featured).



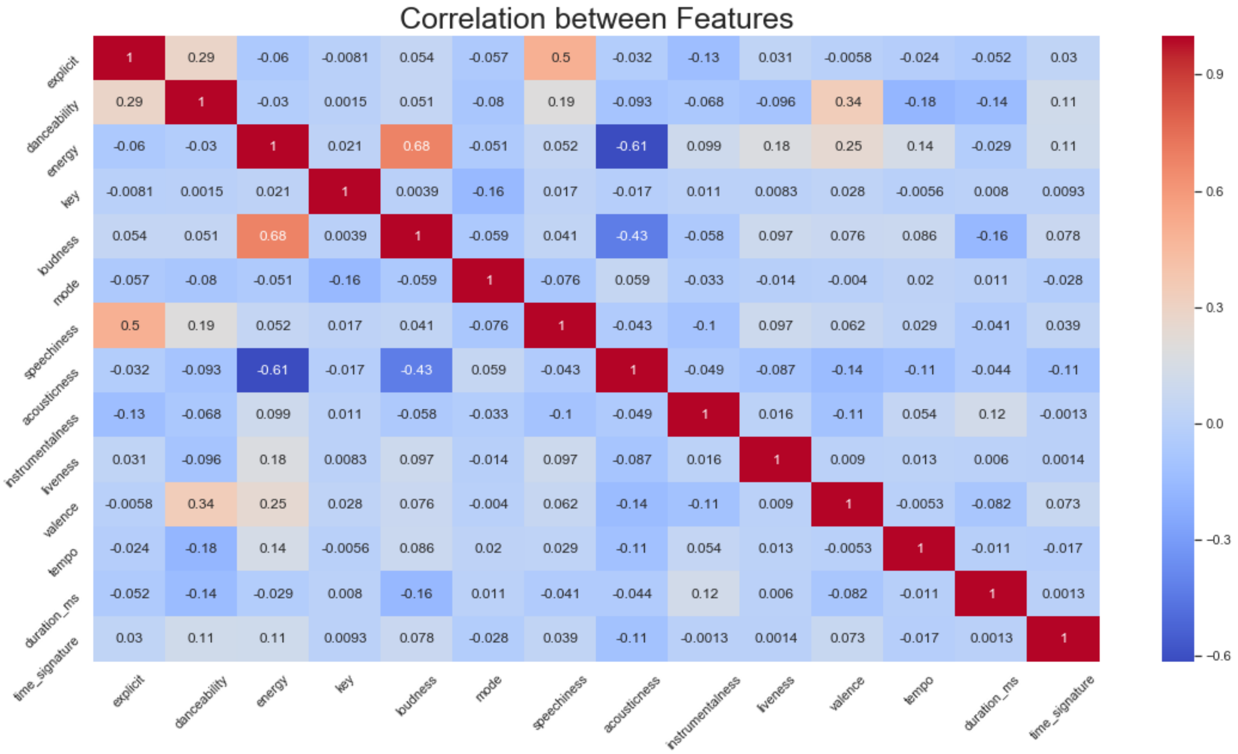
The resulting pvalue from the sampling was 0 suggesting there is a statistically meaningful difference in means between the two groups. The plot below shows the observed difference in means and the distribution of the bootstrap sampling.



2a) Is there any relationship between any of the features? If so, will the correlation between two features impede the classification model?

Prior to building the classification model, the features were evaluated for collinearity.

To determine collinearity, a correlation matrix of the features was constructed. Most notably, there appears to be a negative correlation between accousticness and energy and a positive correlation between loudness and energy.



2b) Will correlation between two features impede a ML model?

To determine if correlation between features will impede the ML model, a regression was run to see if there is a significant R^2 between the two features.

Accousticness & Energy

Running a linear regression with Accousticness and Energy as the dependant and independent variables, we observe an R^2 of 37%. This observation is not significant enough to exclude either variable from our ML model.

Loudness & Energy

Important to note: In this dataset, loudness is measured as the logarithm of decibels from a baseline 0 to encapsulate the wide range of numeric decibel values that a song can take on. For the purposes of determining linear collinearity, the exponent and normalization has not been applied to the data. This does introduce a small amount of inaccuracy in the linear model, however, it is likely not significant enough to bias the end result.

Loudness & Energy have an R^2 of 39%. This is not significant enough to exclude one of the features in the final model.

3) Is there a meaningful difference between subgenres in terms of features? Put another way, are there distinguishing features in the dataset difference between for example 'pop' music and 'dance pop' music?

Keeping each subgenre separate might induce a high amount of error in the model due to potential marginal differences in the features between similar genres (ex. Pop & Dance Pop).
A possible solution is to introduce sparsity into the model output by grouping similar genres.

To determine if similar genres share similar statistically similar distributions and values, two tests were carried out on the similar subgenres:

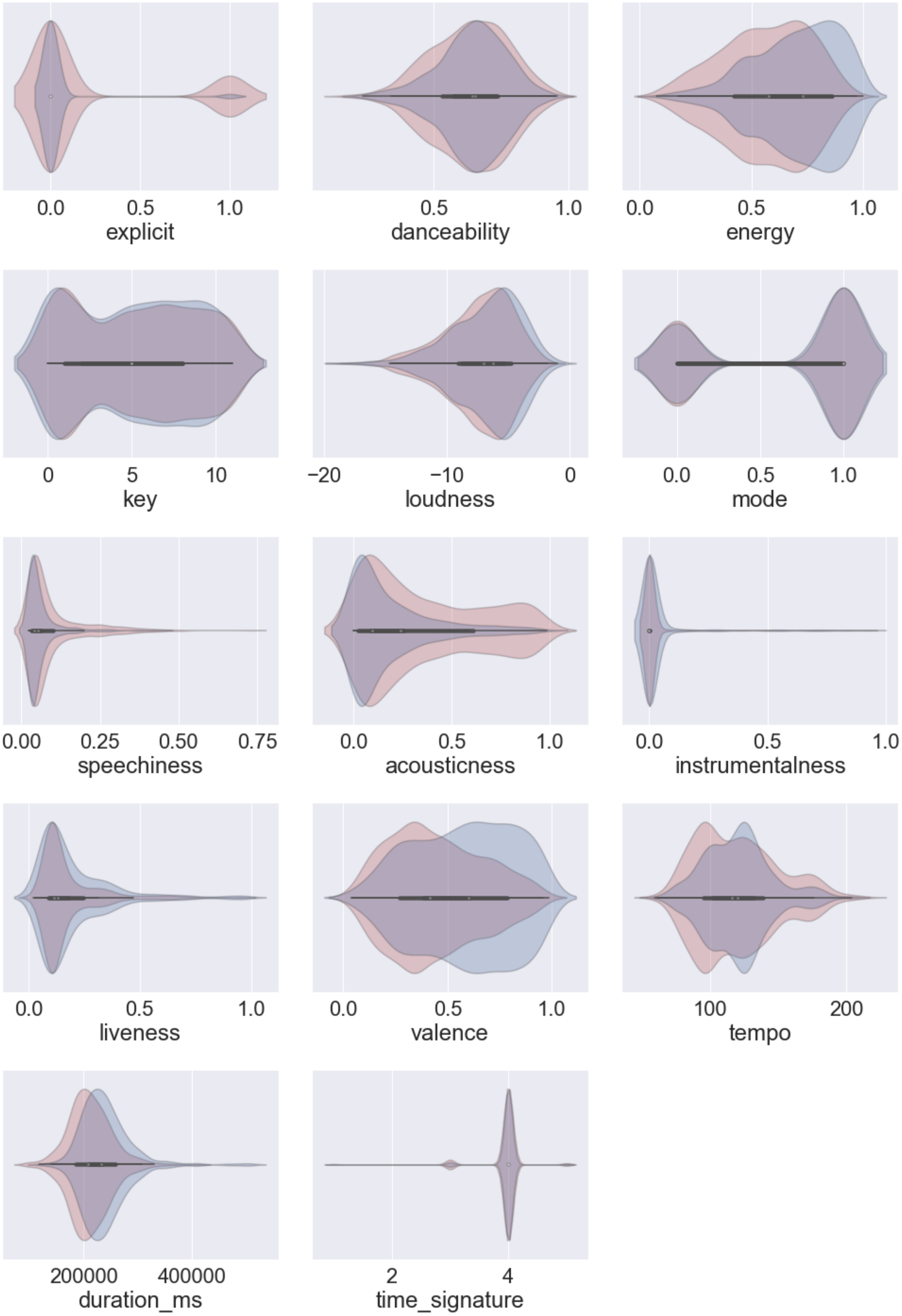
- Bootstrap sampling of the means for each feature
- Mann-Whitney u test for distribution differences for each feature

Example: Working with pop & dance pop distributions, we see that only a 3 features exhibit both a statistically meaningful difference in mean and distribution.

Pop versus Dance Pop

Feature	Bootstrap µ pvalue	Mann-Whitney u test pvalue
explicit	0.0	0.0
danceability	0.9256	0.0684
energy	1.0	0.0
key	0.3157	0.2467
loudness	1.0	0.0
mode	0.7036	0.2733
speechiness	0.0	0.0
acousticness	0.0	0.0
instrumentalness	0.9899	0.0
liveness	0.9995	0.0317
valence	1.0	0.0
tempo	0.6991	0.0196
duration_ms	1.0	0.0
time_signature	0.9463	0.0221

The below plot displays the difference in mean values and distribution for each feature:



From the statistical tests and plots, there are observed slight differences between subgenres. For the purposes of the classification model, merging subgenres will likely improve model performance while not losing meaningful granularity in terms of the models ability to appropriately classify a song.

To confirm this hypothesis, both the original dataset and a condensed version will be used to determine the optimal data for the model.