

Using Spotify Audio Features to Model Track Genre

Introduction

Humans are capable of understanding the difference between a rock song and a classical song based on audio features. Thus, there must be some relationship between the audio features of a song and its genre. However, inferring the genre of a song computationally is a much more complex problem because it requires inferring the human perceptible audio features of a song from a very high dimensional space.

Spotify has algorithms which are designed to quantify human perceptible audio features such as “danceability” and “loudness”. But how well do these algorithms capture the human perception of music? Can these computed features be used to infer the genre of a song?

This statistical analysis is designed to predict the genre of a Spotify track (response) based on Spotify’s audio features (predictors) using multinomial logistic regression. If we find evidence that a relationship exists between Spotify’s audio features and a genre, then this would be evidence to support that Spotify’s audio features capture the human perception of audio.

Data

The dataset includes 6588 tracks with variables such as audio features and track genre. The dataset also includes information about the track such as release date, artists names, and track name. However, this model is only concerned with audio features and genre.

The response is a categorical variable of the tracks genre. The dataset contains 8 genres; classical, electronic dance music (edm), country, hip-hop, metal, punk, pop, and rock.

There are 11 total predictor variables which describe different audio features of the track. See the table on the next page for the names, variable type, and a description of the audio features provided by the Spotify API that are used as predictor variables for this model.

The dataset was scraped using Spotify’s API with a python script. The script used Spotipy, a Python library for Spotify API calls. The script first retrieved a list of genres from the API. Then it retrieved a list of playlists from each genre and the tracks for each of the playlists. Duplicate tracks were removed in the script by removing tracks which had the same track name and artist ids. Spotify’s API only allows for a limited number of genres and playlists to be retrieved by the API user.

Predictor Variables

Variable Name	Type	Description
acousticness	float	A confidence measure from 0.0 to 1.0 of whether the track is acoustic.
danceability	float	A value from 0.0 to 1.0 describing how suitable a track is for dancing.
duration_ms	integer	The duration of the track in milliseconds.
energy	float	A measure from 0.0 to 1.0 and represents a measure of intensity and activity.
instrumentalness	float	A confidence measure from 0.0 to 1.0 of whether a track contains no vocals.
key	integer	The key the track is in as an integer in Pitch Class notation.
liveness	float	A confidence measure from 0.0 to 1.0 of whether the track contains live audience.
loudness	float	The overall loudness of a track in decibels (dB).
mode	integer	Indicates the modality. Major is represented by 1 and minor is 0.
speechiness	float	A measure from 0.0 to 1.0 of the presence of spoken words in a track.
tempo	float	The overall estimated tempo of a track in beats per minute (BPM).
time_signature	integer	An estimated time signature.
valence	float	A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track

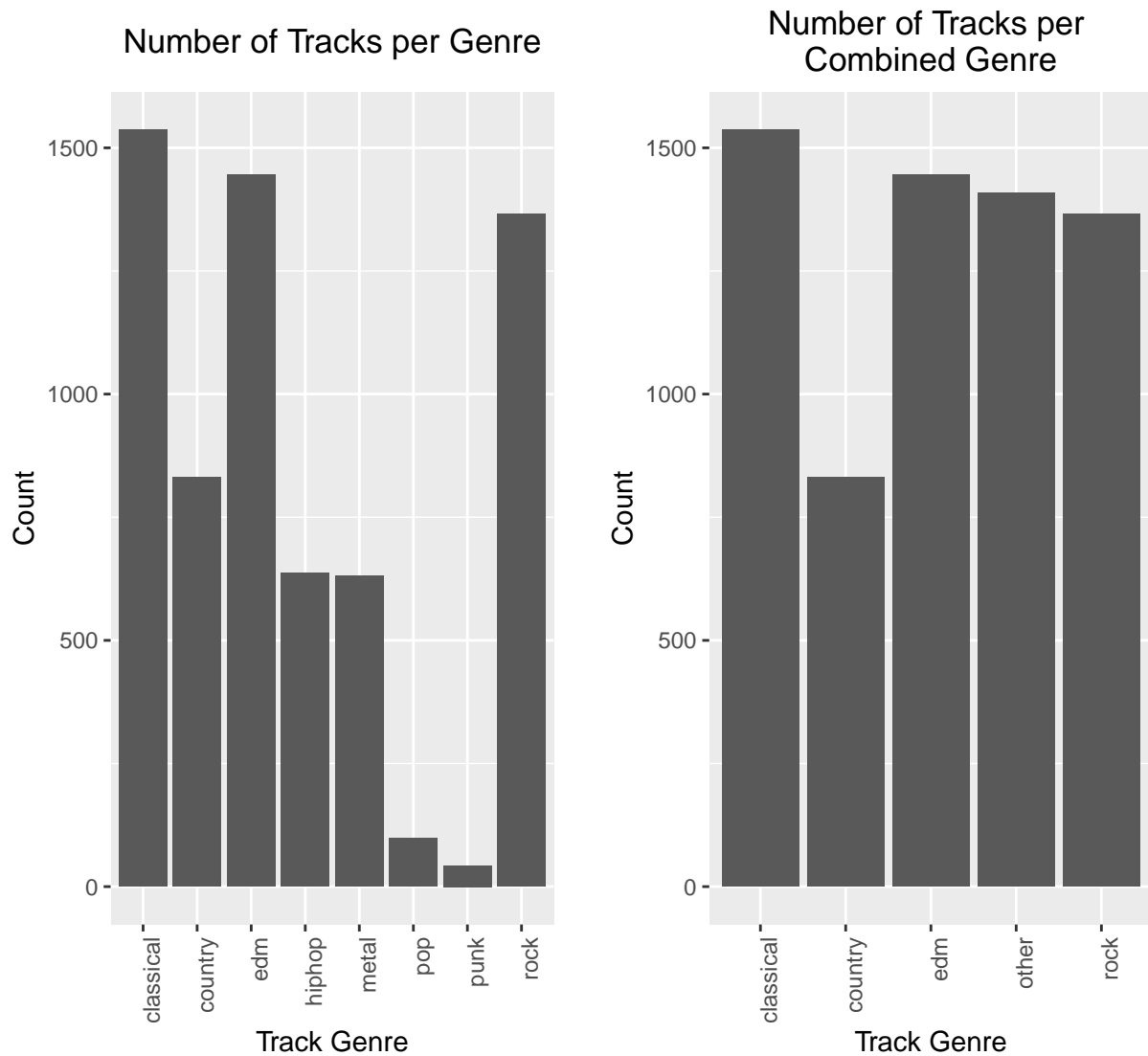
References

[Data Scraping Script](#)

[Spotify API Reference](#)

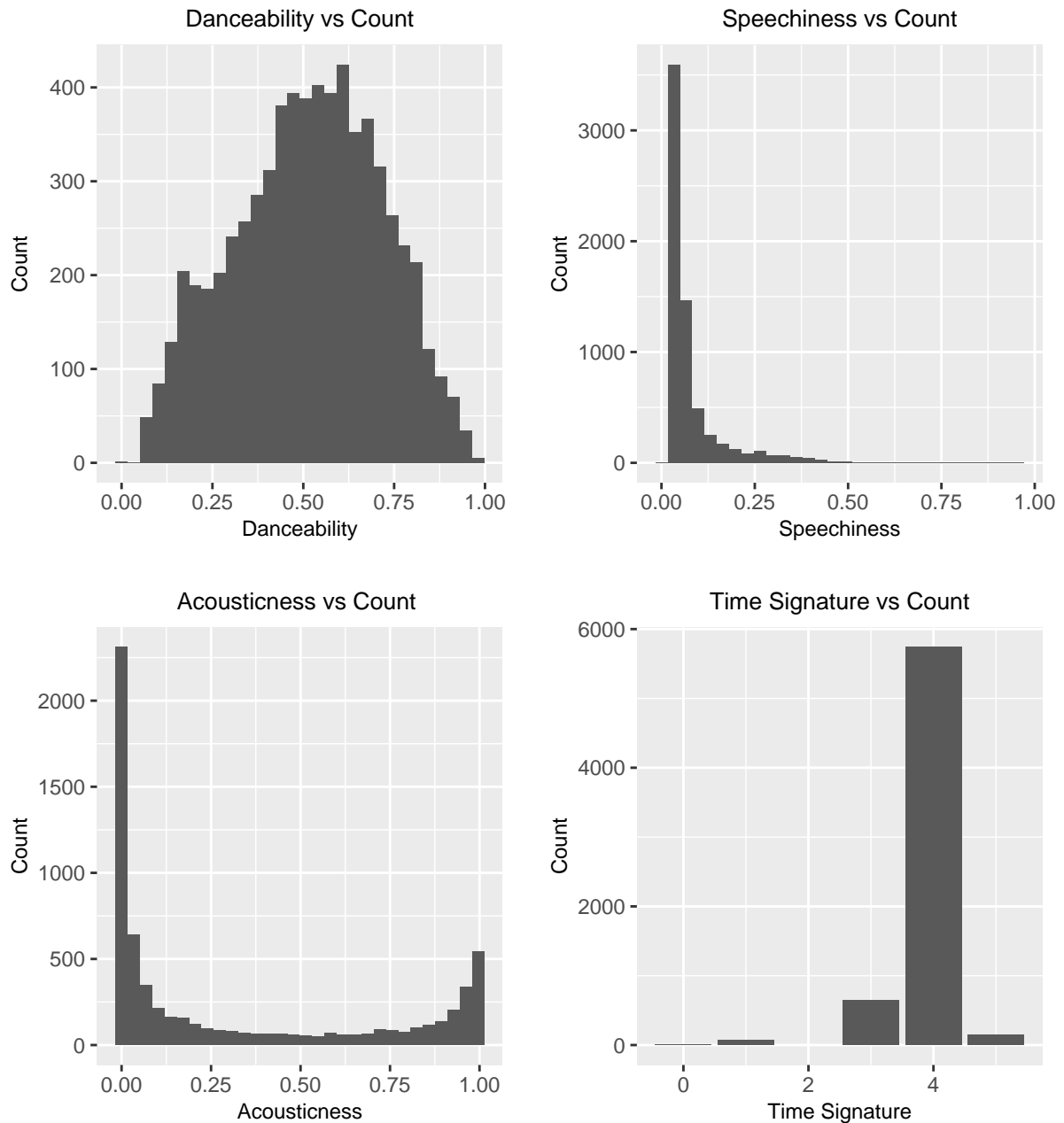
Towards Data Science published an [article](#) describing several modeling approaches to predict whether a Spotify track will be on Spotify's Top 100 billboard using similar predictor variables.

Exploratory Data Analysis



The plot on the left is the raw data containing 8 possible genres. In this plot, there is a lot of variation between the counts for each genre. The plot on the right uses a data frame which combines hip-hop, pop, punk, and metal to one genre called “other”. This plot has a much more even distribution of counts and only contains 5 possible genres. Multinomial models perform best with 5 or less response categories and an even distribution of counts for each category. Therefore, the dataframe of combined genres plotted on the right will be used for the model instead of the original raw dataset.

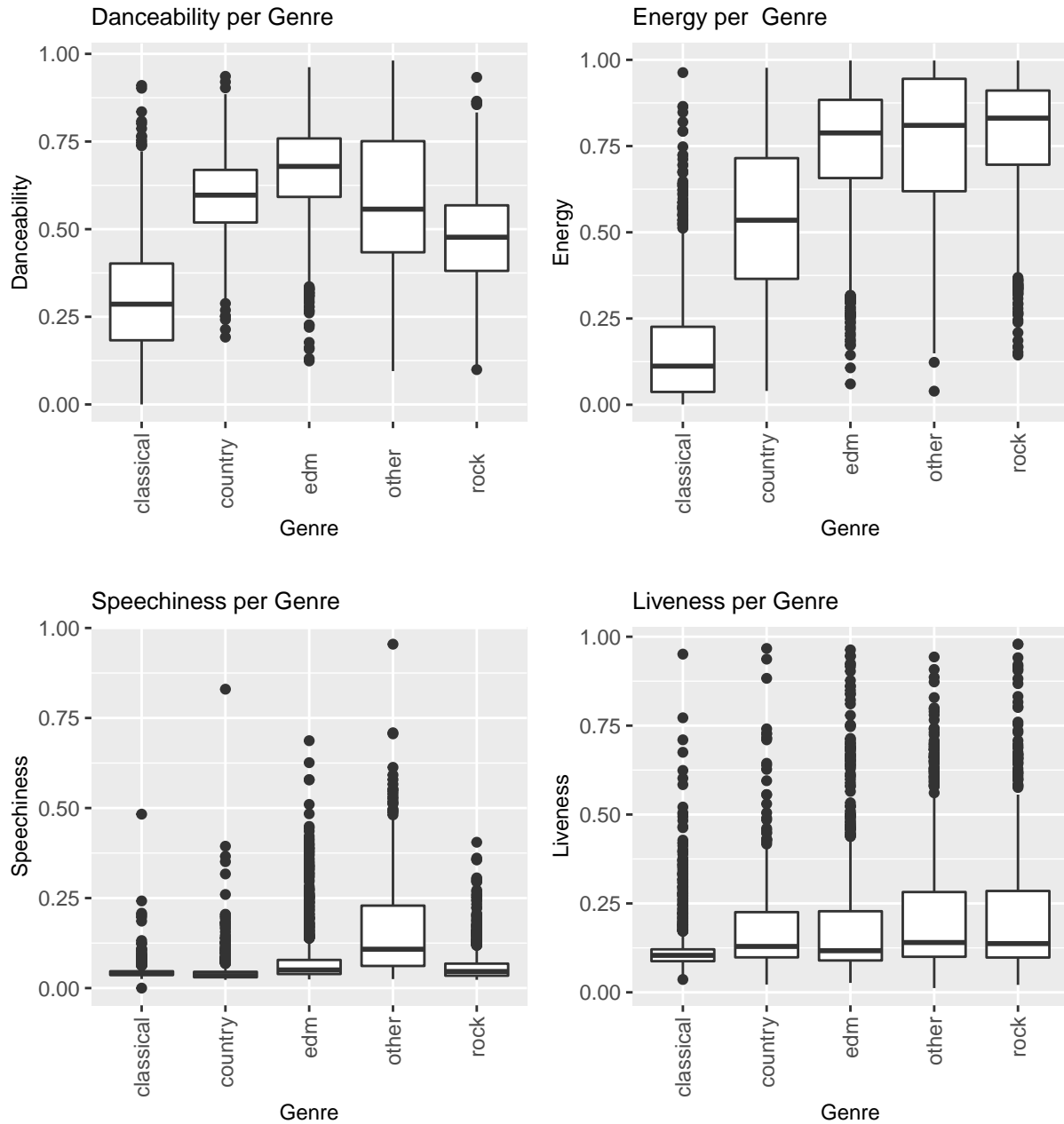
Distribution of Predictor Variables



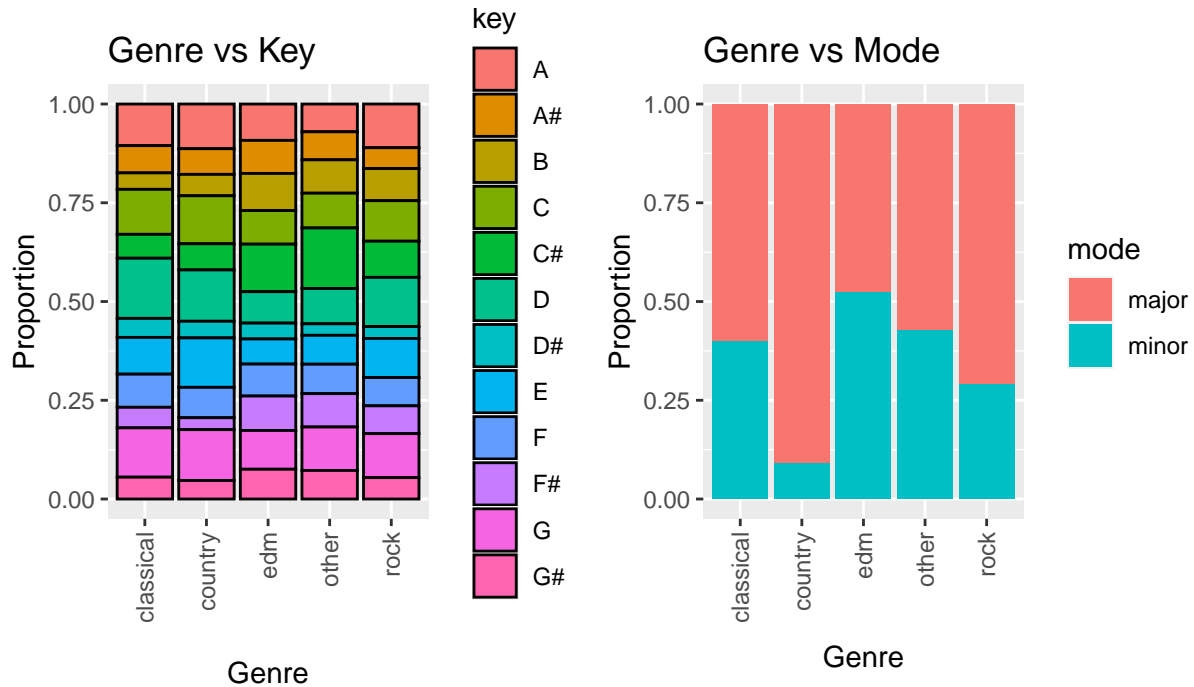
There is a wide range of distributions for the predictor variables. Some predictors like danceability, roughly follow a normal distribution. Acousticness however seems to roughly follow a bimodal distribution. Some plots, particularly the speechiness plot do not show a lot of variation. The vast majority of observations have a speechiness of around 0. This is expected as it would be rare for a musical track to contain spoken (not sung) words. The time signature plot also indicates that the majority of the tracks are in 4/4 time. This is the widely accepted convention for time signature in modern music.

Some predictor variables such as time signature and speechiness may be less effective for predicting genre due to their lack of variation.

Continuous Predictors vs Genre

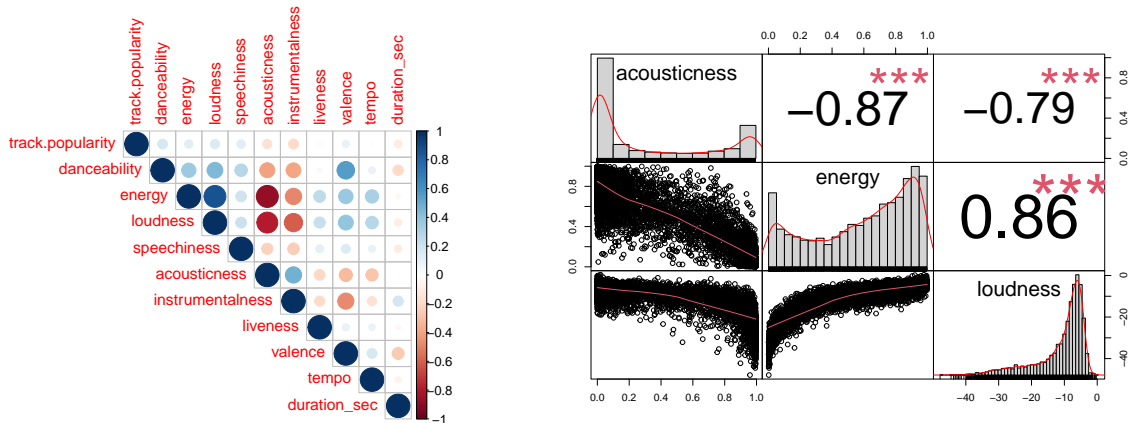


Speechiness and liveness do not show much variability between genres. Throughout the predictors, classical is shown to differ greatly from the other genres. This may suggest that classical would be a good baseline category for the multinomial model. The most promising predictors seem to be energy and danceability based on their variation between genres.



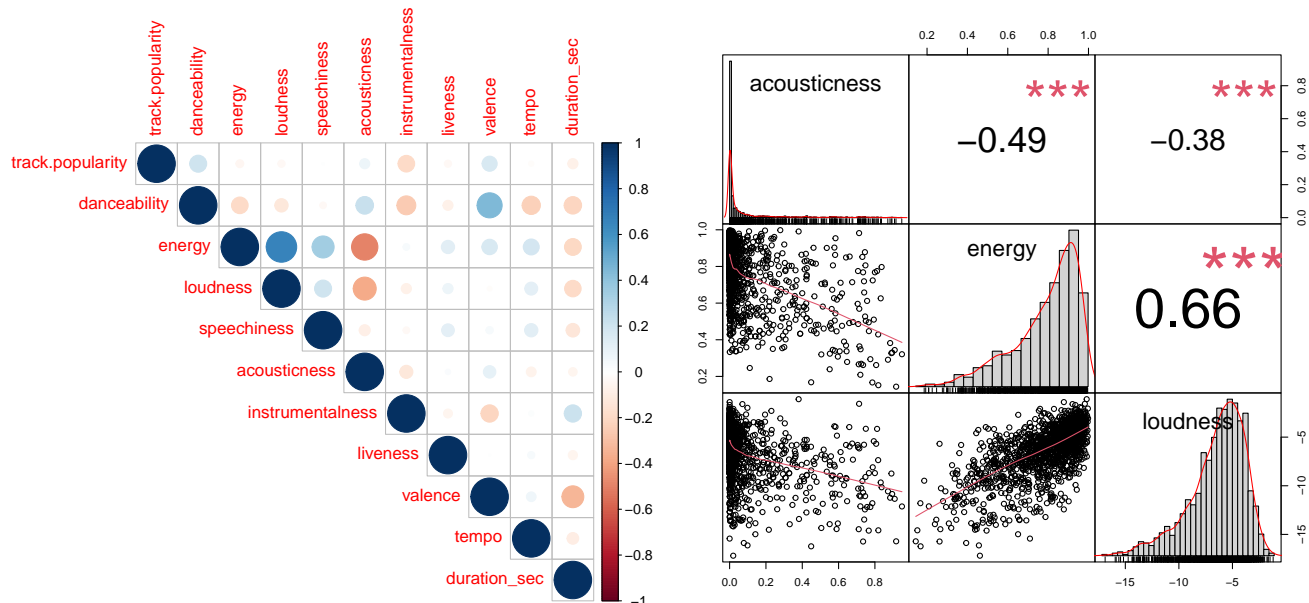
The distribution of keys seems to be consistent throughout each genre. This suggests that genre may be independent from key. The distribution modes contains more major modes throughout the dataset. However, country has a much higher proportion of major modes than others.

Predictor Correlation

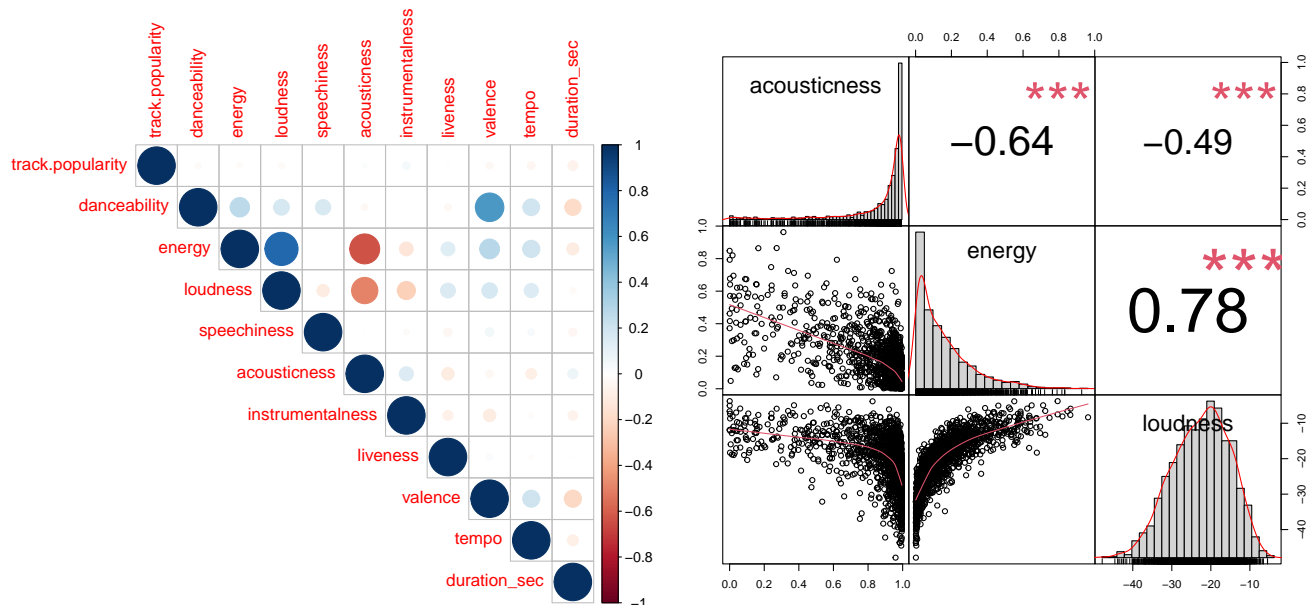


The plot on the left shows the correlation between the continuous predictor variables. It is shown that loudness and energy have a high correlation with acousticness. The plot on the right gives more detailed information about these variables which seem to be highly correlated. In this plot, it is shown that energy and loudness have a strong negative linear relationship with acousticness. This reflects an assumption that acoustic tracks are generally more “mellow”. Similarly loudness and energy have a strong positive linear relationship which reflects the human interpretation that “loud” tracks should have more “energy”.

Correlation with Rock Genre



Correlation with Classical Genre



The above plots demonstrate how the correlation between the variables is not consistent throughout all genres. While both Classical and Rock genres show energy and loudness have some relationship with acousticness, the magnitude and type of correlation is widely different between the two genres. This suggests that a multinomial model may not have sufficient complexity.

Regression Analysis

Model Selection

The model selected was a multinomial model containing all original the original predictor variables and an interaction between key and mode. The model paramaters are included in the “Additonal Work” section. Classical was selected as the baseline category as it had the most different predictor variable values. It was difficult to justify removing any predictor variables because ANOVA tests indicated that all predictor variables had at least one parameter not equal to 0. The key variable only contained the base note. However, musical keys are not defined by a single base note, but rather the base note and the mode (C major, A minor, etc.) making an interaction between the two seem very likely. The interaction was justified with a lower residual deviance and higher prediction accuracy compared to the model without an interaction.

Model Performance

Predicted vs Actual Genre

genre	classical	country	edm	other	rock	accuracy
classical	1461	25	32	2	17	0.957
country	22	545	38	29	197	0.656
edm	26	69	1037	189	125	0.717
other	10	50	251	759	339	0.539
rock	8	119	130	186	922	0.675

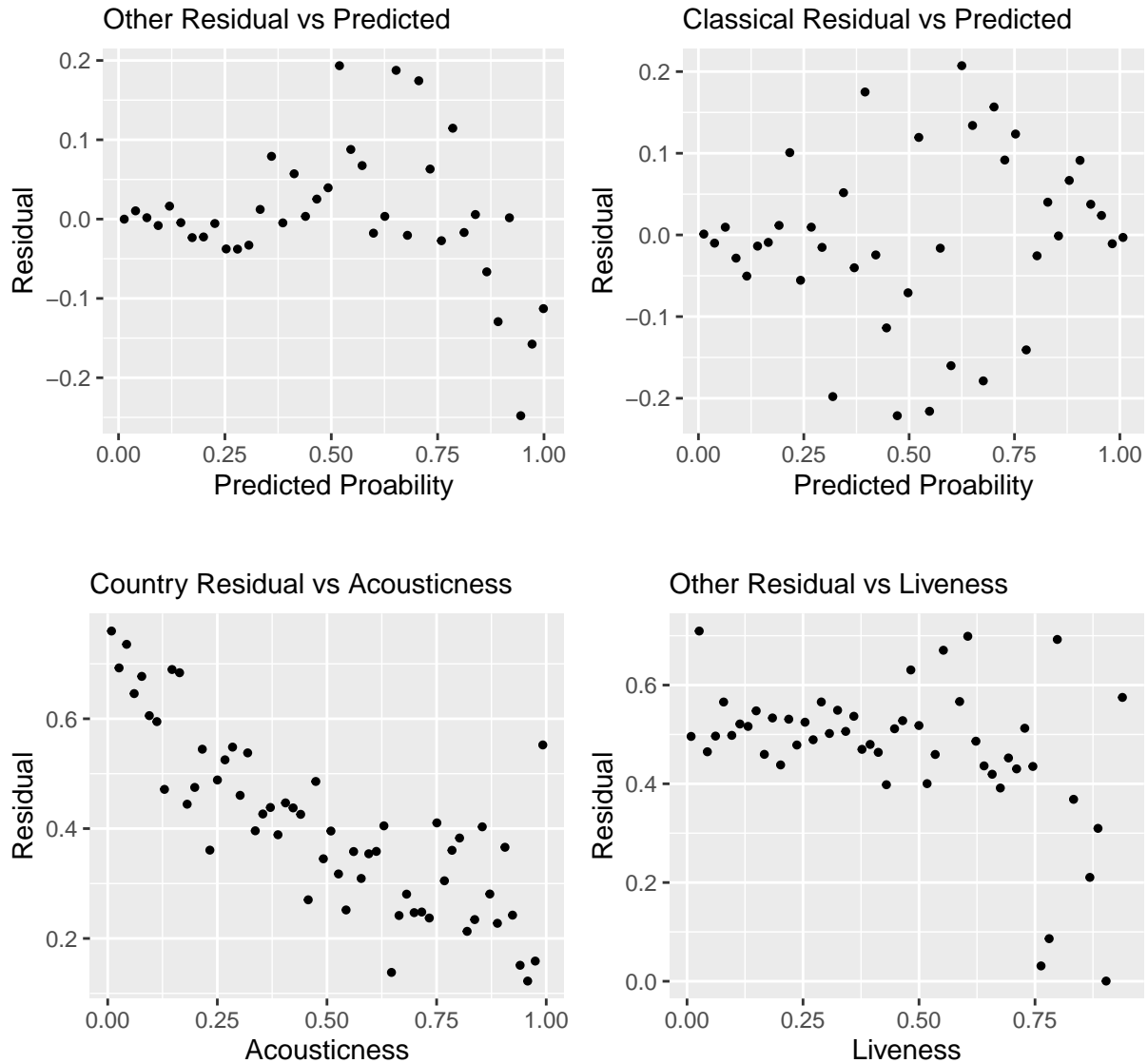
The model shows high performance with a total accuracy of around 72%. Classical has the highest prediction accuracy. This is expected as classical had the most different values for predictors. Other has the least accuracy which is also expected because it contains multiple genres making it more complex to model.

Model Assumptions

Random Sample: The sample was randomly selected from Spotify’s available genres and playlists on the API. However, it is unknown if the available genres and playlists are a random sample from the total population of genres and playlists on Spotify. It is possible that Spotify selects which genres and playlists are available to API users. Therefor, it is difficult to assume randomness.

Independence From the correlation plots, energy and loudness both have a strong linear relationship with acousticness. However, an ANOVA test suggested that acousticness should still be included in the model. This may be due to the number of categories in the response. Since it was difficult to justify removing variables based on the ANOVA test, variables that did not appear to be independent from EDA are included in the model.

Linearity



In the residual plot vs predicted other (top left) and the residual plot vs predicted classical (top right) constant variance is not shown. In the residual plot vs predicted other, a clear pattern is observed. However, in the residual plot vs predicted classical plot, a clear pattern is not observed.

The plot on the bottom left shows the country genre residual vs acousticness. A clear linear pattern is observed. However, the variance seems relatively constant. The plot on the bottom right shows the other genre residual vs danceability. A clear pattern is not observed but the variance is not constant.

The four above plots are examples of the wide range of residual patterns across the predictor variables. It is difficult to assume linearity. This makes it difficult to make accurate inferences of the model parameters.

Discussion

The overall performance of the model is quite impressive. A total prediction accuracy of 72% is very high considering there are 5 response categories. This is strong evidence to support that Spotify’s audio features accurately measure the human interpretation of audio.

Other Genre Prediction Matrix

	True Other	False Classical	False EDM	False Country	False Rock
Number of Predictions	759	10	251	50	339
% Predicted	53.9	0.07	17.8	3	24

Other genre showed the the most error. However, false predictions of other genre may reflect the human interpretation of track genre. Other genre contained metal and punk genres. Humans find metal and punk very similar to rock and their genres could easily be mistaken. This could explain the high proportion of false rock predictions for other genre and could actually be evidence that the audio features capture human interpretation of audio. A similar interpretation does not explain the high proportion of false edm predictions however.

Examples of Model Parameters Reflecting Human Interpretation

Genre	Variable	Param. Est.	P-Value	Conf. Low	Conf. High
EDM	Danceability	14.124	0.000	13.502	14.747
Country	Instrumentalness	-7.773	0.000	-8.872	-6.674
Rock	Acousticness	-5.014	0.000	-5.732	-4.266

Humans would expect eletronic dance music (EDM) to be much more “danceable” than classical music. This is reflected in the model showing a 14.124 increase of the log odds of a track belonging to the EDM genre as oppose to the classical genre based on a tracks danceability value.

Humans would also expect country music to have more singing than classical music. This is reflected in the model showing a 7.773 decrease of the log odds of a track belonging to the country genre as oppose to the classical genre based on a tracks instrumentalness value (instrumentalness is intended to decrease with more singing in a track)

Finally, humans expect rock music to not be acoustic and classical music to be acoustic. This is reflected in the model showing a 5.014 decrease of the log odds of a track belonging to the rock genre as oppose to the classical genre based on the acousticness value.

These parameters are examples of how the model may reflect some of the human interpretation of audio to predict genre. It is evidence to support that Spotify’s audio features accurately represent human interpretation.

Overall, the model does contain evidence that Spotify’s musical features reflect human interpretations of music. However, we must be cautious when interpreting the model due to the limitations discussed in the next section.

Limitations

The limitations of the multinomial model are stretched with this model due to its complexity. The high number of predictor variables also makes it difficult to assume linearity for all variables. As shown in the analysis section, many of the variables had poor residual plots. The high variation of p-values and linearity suggest that many of the parameters may not actually be useful for the model. The non-linearity makes the interpretation of the parameters questionable; even with low p-values as the linearity condition is not met. The high number of response categories also make it difficult to justify removing variables with an ANOVA test. This suggests that a different statistical model which allows for non-linear relationships and more response categories may be better suited for the problem.

Interpretation is also difficult without the details to the Spotify algorithms used to calculate the predictor variables. While Spotify does provide a good description of the variables, exactly how those variables are calculated could provide more information on how to interpret the coefficients.

The prediction accuracy should also have been compared with a human prediction accuracy. Without a baseline of human performance for genre prediction, it is difficult to make conclusions about the relationship between the audio features and human interpretation.

The other genre also may be a bit problematic. Having widely different genres combined into one category likely complicates the model. This was also reflected with a much lower prediction accuracy for the other genre. Future studies should consider having distinct genres instead of combining genres.

Finally, this model did not include cross validation. It is possible that over-fitting is present in the dataset. Future studies should include cross validation to prevent over-fitting.

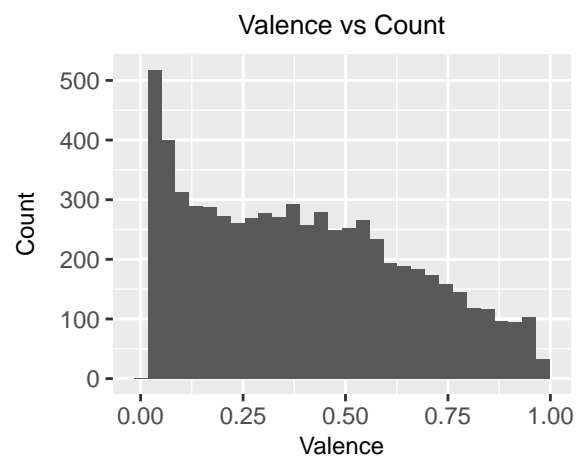
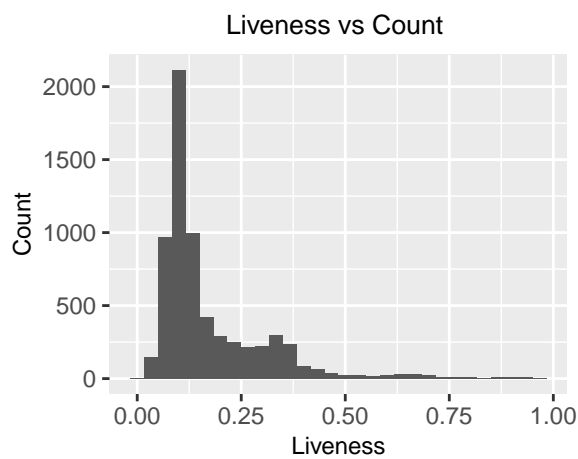
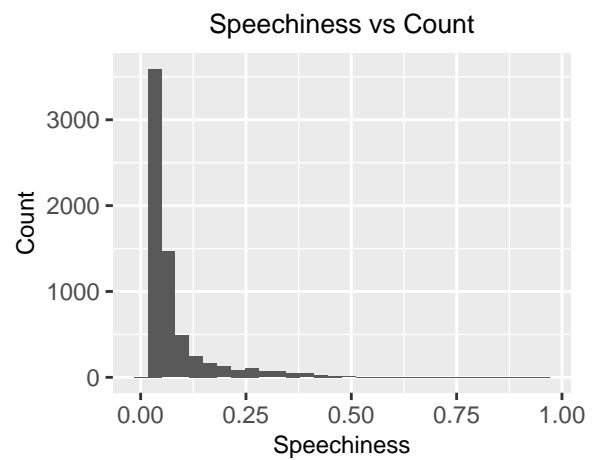
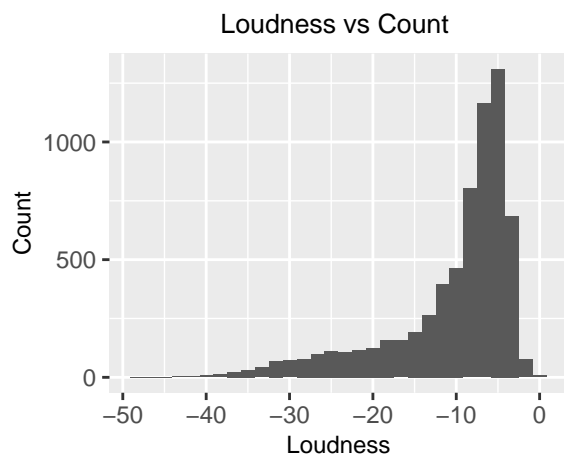
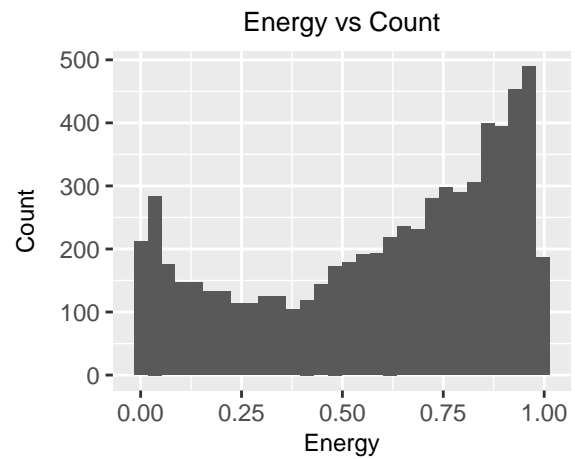
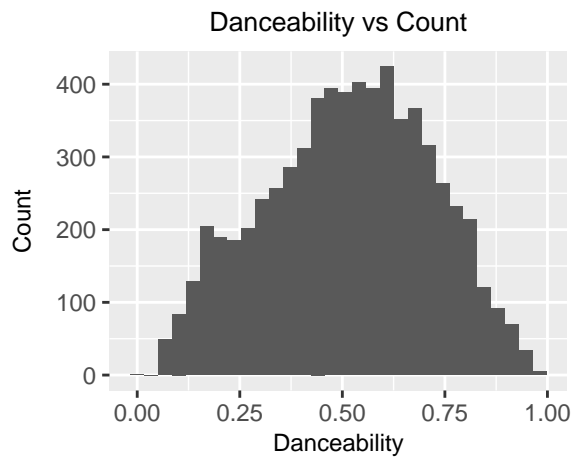
Conclusion

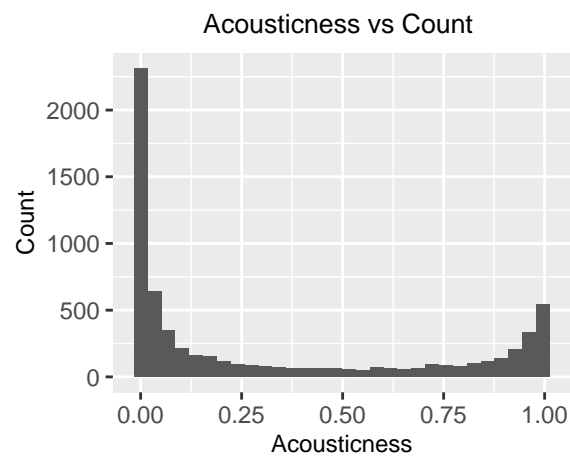
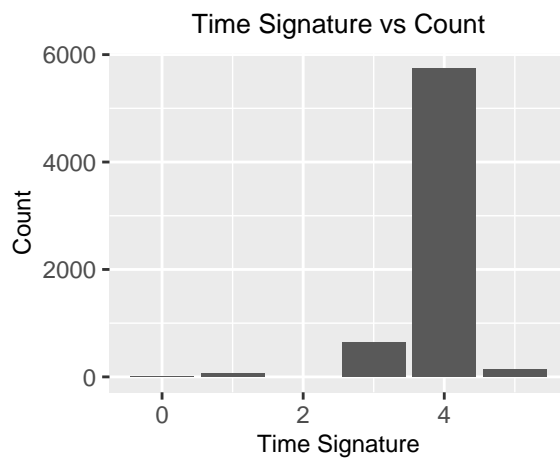
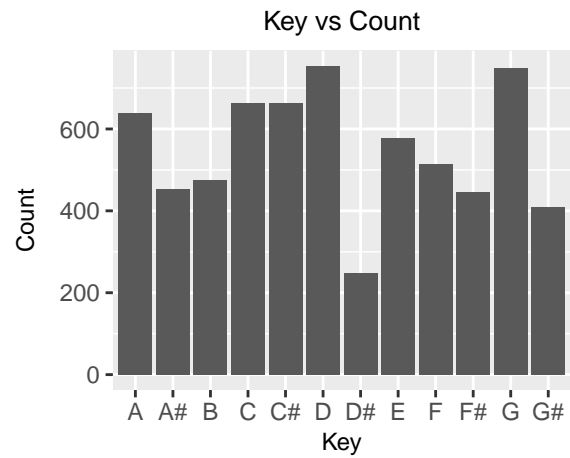
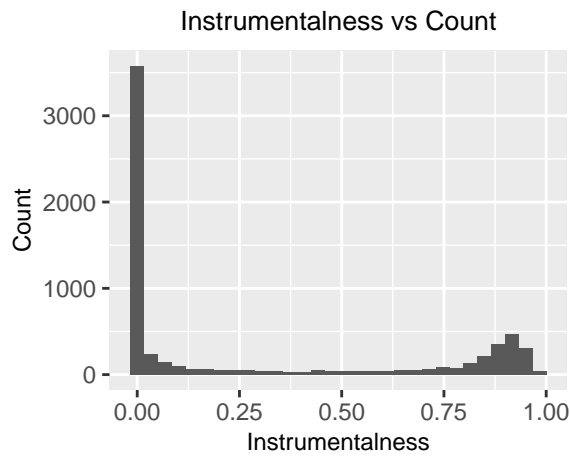
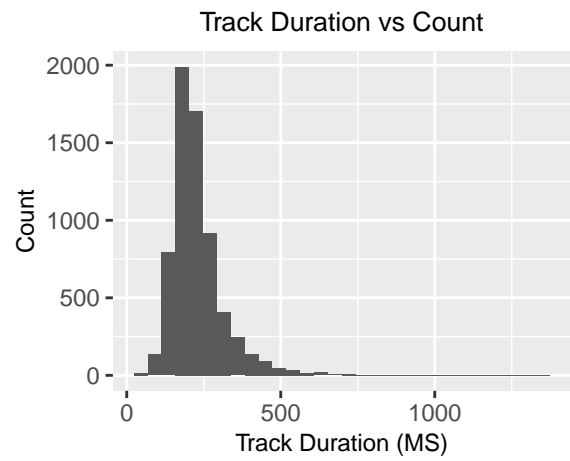
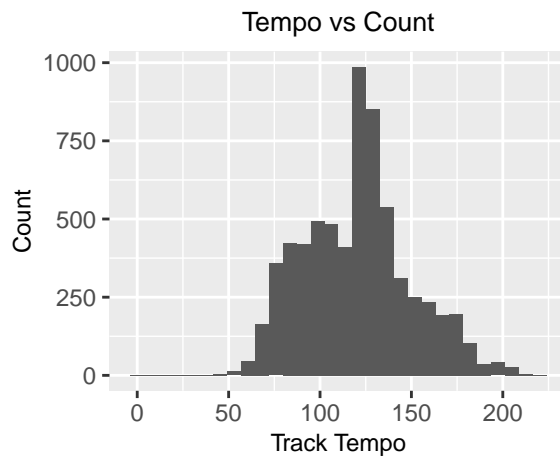
While the model has many limitations, it still shows evidence that Spotify's audio features can be used in a multinomial statistical model to predict track genre. The benefit to using a simple statistical model for prediction is that it allows us to interpret the audio features directly. This is important because we would expect a model to reflect that tracks that are more "danceable" are more likely to belong to the electronic dance music genre as opposed to the classical. However, evidence is also shown that a more complex model may be more sufficient due to the lack of linearity and number of response categories. In any case, the fact that a 72% prediction accuracy of genre was obtained suggests that Spotify's audio features reflect human interpretation of audio.

This model is also an example of how statistical models can be combined with machine learning algorithms to address complex relationships while still maintaining interpretation. In this project, algorithms handled the complexity of the problem by encoding track audio features into variables to be used in an interpretable statistical model.

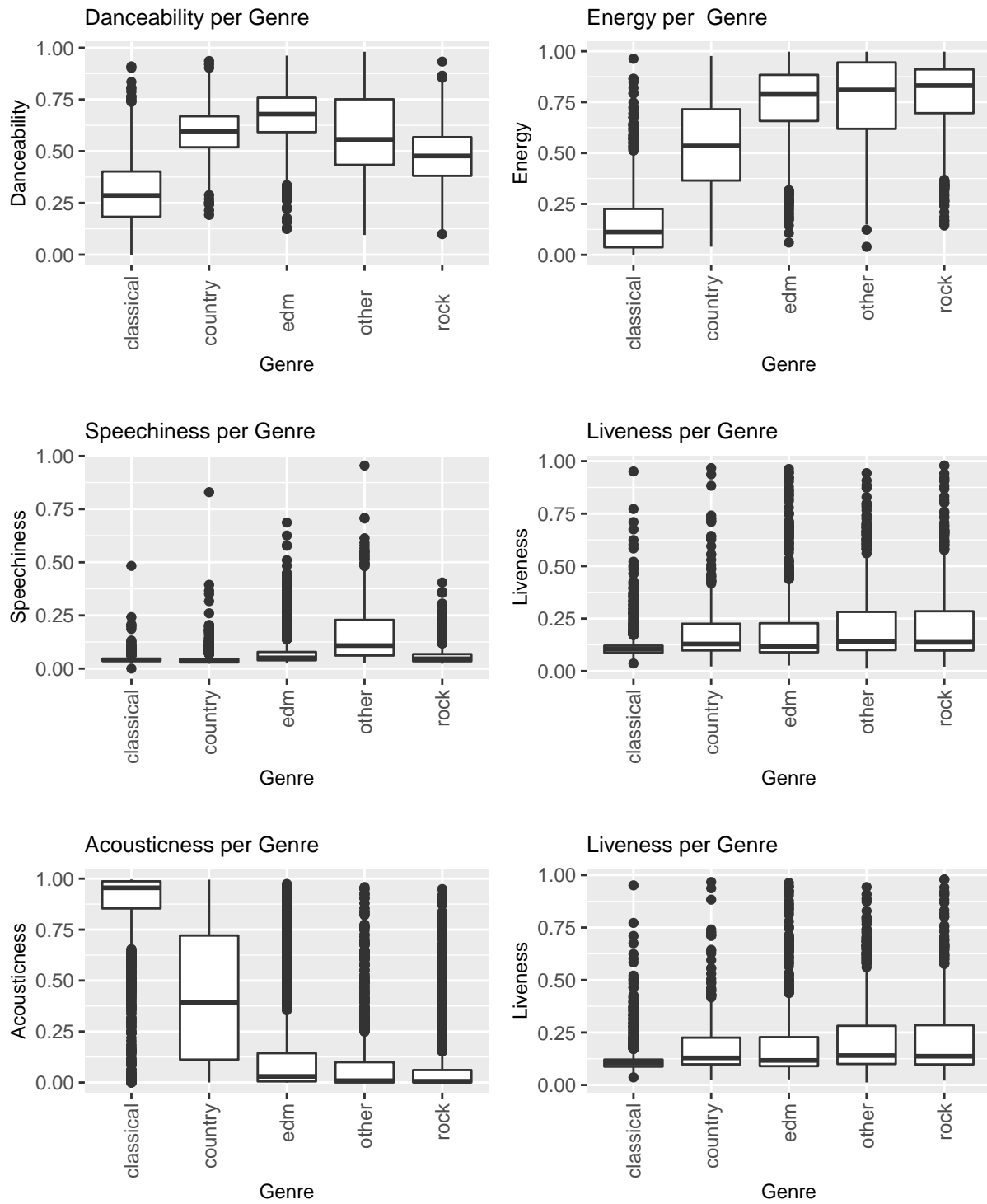
Additional Work

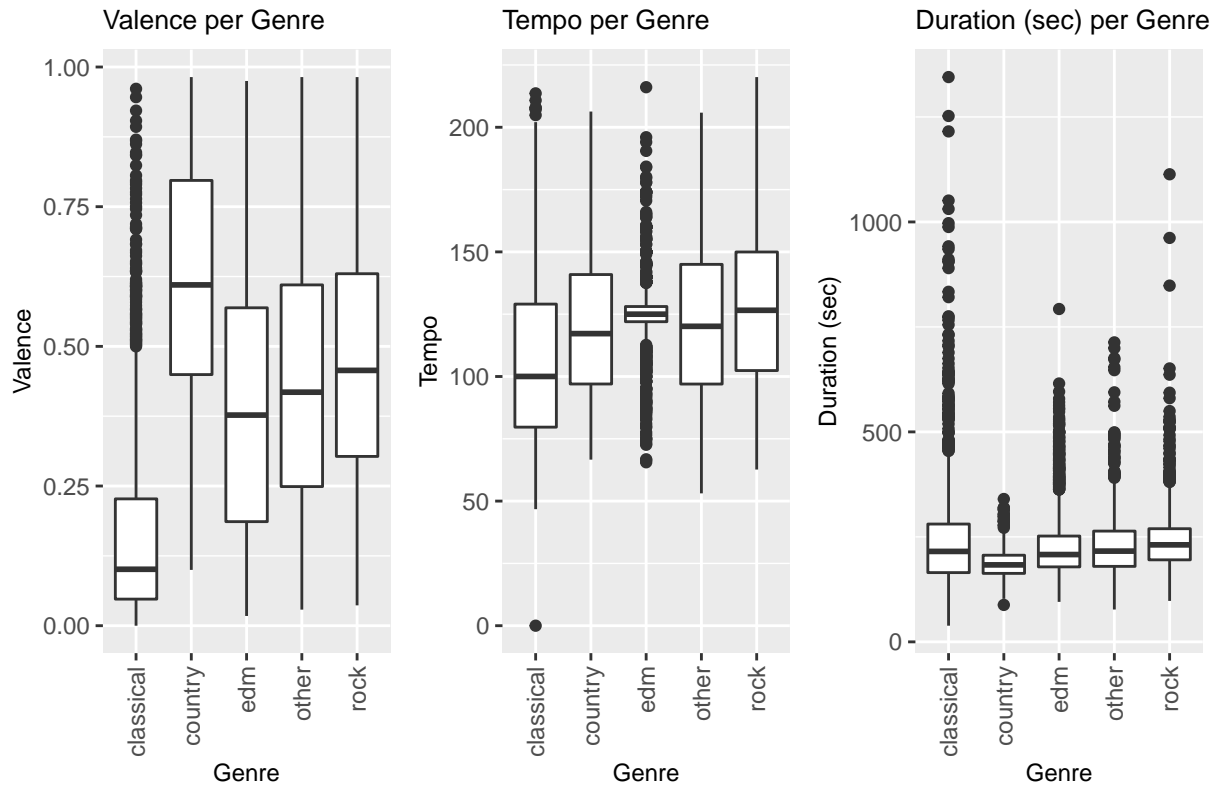
Distribution of Plot Counts vs Predictors





Genres vs Predictors





Full Model Output

y.level	term	estimate	std.error	statistic	p.value	conf.low	conf.high
country	(Intercept)	5.253	0.533	9.858	0.000	4.209	6.297
country	danceability	4.519	0.378	11.945	0.000	3.777	5.260
country	energy	-0.624	0.406	-1.538	0.124	-1.421	0.172
country	loudness	0.223	0.029	7.727	0.000	0.166	0.280
country	speechiness	-8.923	0.050	-179.032	0.000	-9.021	-8.826
country	acousticness	-3.113	0.363	-8.566	0.000	-3.826	-2.401
country	instrumentalness	-7.773	0.561	-13.862	0.000	-8.872	-6.674
country	liveness	4.126	0.303	13.622	0.000	3.532	4.719
country	valence	3.775	0.335	11.285	0.000	3.120	4.431
country	tempo	0.000	0.003	0.143	0.886	-0.006	0.007

y.level	term	estimate	std.error	statistic	p.value	conf.low	conf.high
country	duration_sec	-0.007	0.001	-5.481	0.000	-0.010	-0.005
country	keyA#	-1.344	0.356	-3.771	0.000	-2.042	-0.645
country	keyB	-0.308	0.388	-0.795	0.427	-1.068	0.452
country	keyC	-0.706	0.306	-2.305	0.021	-1.307	-0.106
country	keyC#	-0.983	0.352	-2.792	0.005	-1.674	-0.293
country	keyD	-0.912	0.279	-3.274	0.001	-1.459	-0.366
country	keyD#	-0.672	0.396	-1.698	0.089	-1.447	0.104
country	keyE	0.139	0.291	0.477	0.633	-0.431	0.708
country	keyF	-1.277	0.304	-4.201	0.000	-1.873	-0.681
country	keyF#	-0.211	0.263	-0.802	0.422	-0.726	0.304
country	keyG	-1.264	0.280	-4.508	0.000	-1.813	-0.714
country	keyG#	-1.723	0.370	-4.654	0.000	-2.449	-0.997
country	modeminor	-3.675	0.233	-15.785	0.000	-4.131	-3.218
country	keyA#:modeminor	1.229	0.452	2.717	0.007	0.343	2.116
country	keyB:modeminor	2.661	0.397	6.702	0.000	1.883	3.439
country	keyC:modeminor	2.047	0.371	5.522	0.000	1.320	2.773
country	keyC#:modeminor	2.542	0.327	7.778	0.000	1.901	3.182
country	keyD:modeminor	0.648	0.444	1.462	0.144	-0.221	1.518
country	keyD#:modeminor	1.452	0.615	2.360	0.018	0.246	2.658
country	keyE:modeminor	0.089	0.389	0.230	0.818	-0.674	0.852
country	keyF:modeminor	1.651	0.357	4.627	0.000	0.952	2.350
country	keyF#:modeminor	1.140	0.416	2.736	0.006	0.323	1.956
country	keyG:modeminor	2.072	0.405	5.121	0.000	1.279	2.865
country	keyG#:modeminor	2.743	0.485	5.657	0.000	1.793	3.694
edm	(Intercept)	-7.016	0.432	-16.254	0.000	-7.861	-6.170

y.level	term	estimate	std.error	statistic	p.value	conf.low	conf.high
edm	danceability	14.124	0.317	44.488	0.000	13.502	14.747
edm	energy	6.247	0.347	18.025	0.000	5.568	6.927
edm	loudness	0.200	0.031	6.563	0.000	0.140	0.260
edm	speechiness	3.515	0.475	7.395	0.000	2.583	4.446
edm	acousticness	-2.189	0.343	-6.373	0.000	-2.862	-1.516
edm	instrumentalness	-2.166	0.288	-7.515	0.000	-2.731	-1.601
edm	liveness	4.701	0.229	20.548	0.000	4.253	5.150
edm	valence	-2.625	0.301	-8.718	0.000	-3.215	-2.035
edm	tempo	0.008	0.003	2.257	0.024	0.001	0.014
edm	duration_sec	0.003	0.001	2.829	0.005	0.001	0.005
edm	keyA#	-1.160	0.377	-3.079	0.002	-1.898	-0.422
edm	keyB	0.078	0.380	0.204	0.838	-0.667	0.823
edm	keyC	-0.578	0.302	-1.918	0.055	-1.169	0.013
edm	keyC#	0.021	0.326	0.063	0.950	-0.618	0.659
edm	keyD	-0.881	0.273	-3.225	0.001	-1.417	-0.346
edm	keyD#	-0.417	0.423	-0.985	0.325	-1.247	0.413
edm	keyE	-0.594	0.310	-1.919	0.055	-1.201	0.013
edm	keyF	-0.556	0.298	-1.864	0.062	-1.140	0.029
edm	keyF#	1.062	0.188	5.645	0.000	0.693	1.431
edm	keyG	-1.212	0.275	-4.408	0.000	-1.751	-0.673
edm	keyG#	-0.610	0.354	-1.724	0.085	-1.304	0.083
edm	modeminor	-0.410	0.208	-1.975	0.048	-0.818	-0.003
edm	keyA#:modeminor	0.809	0.276	2.929	0.003	0.268	1.350
edm	keyB:modeminor	1.099	0.294	3.733	0.000	0.522	1.675
edm	keyC:modeminor	1.121	0.251	4.473	0.000	0.630	1.613

y.level	term	estimate	std.error	statistic	p.value	conf.low	conf.high
edm	keyC#:modeminor	-0.317	0.234	-1.354	0.176	-0.775	0.142
edm	keyD:modeminor	-0.314	0.270	-1.160	0.246	-0.844	0.216
edm	keyD#:modeminor	2.070	0.369	5.614	0.000	1.347	2.793
edm	keyE:modeminor	-0.302	0.280	-1.077	0.281	-0.851	0.247
edm	keyF:modeminor	0.236	0.260	0.910	0.363	-0.273	0.746
edm	keyF#:modeminor	-0.196	0.249	-0.786	0.432	-0.684	0.292
edm	keyG:modeminor	0.903	0.251	3.592	0.000	0.410	1.396
edm	keyG#:modeminor	1.234	0.286	4.312	0.000	0.673	1.795
other	(Intercept)	-0.155	0.409	-0.379	0.705	-0.957	0.646
other	danceability	6.616	0.280	23.600	0.000	6.067	7.165
other	energy	3.522	0.325	10.842	0.000	2.885	4.159
other	loudness	0.297	0.031	9.465	0.000	0.236	0.359
other	speechiness	13.821	0.400	34.544	0.000	13.037	14.605
other	acousticness	-4.322	0.379	-11.407	0.000	-5.065	-3.580
other	instrumentalness	-4.764	0.316	-15.060	0.000	-5.384	-4.144
other	liveness	4.302	0.208	20.646	0.000	3.893	4.710
other	valence	-0.555	0.302	-1.837	0.066	-1.148	0.037
other	tempo	-0.005	0.003	-1.378	0.168	-0.011	0.002
other	duration_sec	0.007	0.001	7.489	0.000	0.005	0.009
other	keyA#	-1.406	0.389	-3.617	0.000	-2.168	-0.644
other	keyB	0.249	0.380	0.654	0.513	-0.497	0.994
other	keyC	-0.509	0.312	-1.630	0.103	-1.121	0.103
other	keyC#	0.160	0.336	0.475	0.635	-0.499	0.818
other	keyD	-0.918	0.284	-3.231	0.001	-1.475	-0.361
other	keyD#	-0.731	0.443	-1.649	0.099	-1.599	0.138

y.level	term	estimate	std.error	statistic	p.value	conf.low	conf.high
other	keyE	-0.374	0.305	-1.228	0.220	-0.972	0.223
other	keyF	-0.720	0.310	-2.323	0.020	-1.328	-0.113
other	keyF#	1.130	0.193	5.864	0.000	0.752	1.508
other	keyG	-0.959	0.284	-3.379	0.001	-1.516	-0.403
other	keyG#	-0.795	0.368	-2.163	0.031	-1.516	-0.075
other	modeminor	-1.321	0.256	-5.166	0.000	-1.822	-0.820
other	keyA#:modeminor	1.697	0.311	5.448	0.000	1.086	2.307
other	keyB:modeminor	1.198	0.322	3.722	0.000	0.567	1.829
other	keyC:modeminor	1.815	0.288	6.312	0.000	1.252	2.379
other	keyC#:modeminor	0.545	0.261	2.089	0.037	0.034	1.057
other	keyD:modeminor	0.253	0.302	0.839	0.401	-0.338	0.845
other	keyD#:modeminor	2.622	0.400	6.562	0.000	1.839	3.405
other	keyE:modeminor	0.210	0.304	0.690	0.490	-0.386	0.805
other	keyF:modeminor	0.718	0.300	2.393	0.017	0.130	1.305
other	keyF#:modeminor	0.026	0.283	0.092	0.927	-0.528	0.580
other	keyG:modeminor	1.404	0.290	4.848	0.000	0.837	1.972
other	keyG#:modeminor	1.910	0.323	5.919	0.000	1.277	2.542
rock	(Intercept)	2.864	0.393	7.279	0.000	2.093	3.635
rock	danceability	1.372	0.290	4.733	0.000	0.804	1.940
rock	energy	3.000	0.318	9.446	0.000	2.377	3.622
rock	loudness	0.235	0.029	7.989	0.000	0.177	0.293
rock	speechiness	-0.995	0.597	-1.666	0.096	-2.165	0.176
rock	acousticness	-5.014	0.366	-13.692	0.000	-5.732	-4.296
rock	instrumentalness	-4.866	0.306	-15.899	0.000	-5.466	-4.266
rock	liveness	4.409	0.201	21.917	0.000	4.015	4.803

y.level	term	estimate	std.error	statistic	p.value	conf.low	conf.high
rock	valence	2.698	0.299	9.023	0.000	2.112	3.284
rock	tempo	-0.003	0.003	-0.930	0.352	-0.009	0.003
rock	duration_sec	0.006	0.001	7.202	0.000	0.005	0.008
rock	keyA#	-1.939	0.373	-5.196	0.000	-2.671	-1.208
rock	keyB	-0.263	0.372	-0.709	0.478	-0.992	0.465
rock	keyC	-0.787	0.297	-2.653	0.008	-1.369	-0.206
rock	keyC#	-0.691	0.333	-2.076	0.038	-1.343	-0.039
rock	keyD	-0.759	0.266	-2.858	0.004	-1.280	-0.239
rock	keyD#	-0.699	0.402	-1.740	0.082	-1.486	0.088
rock	keyE	-0.311	0.277	-1.121	0.262	-0.854	0.233
rock	keyF	-0.786	0.284	-2.764	0.006	-1.342	-0.229
rock	keyF#	0.425	0.182	2.330	0.020	0.067	0.782
rock	keyG	-1.254	0.270	-4.653	0.000	-1.783	-0.726
rock	keyG#	-1.520	0.359	-4.229	0.000	-2.225	-0.816
rock	modeminor	-1.950	0.255	-7.645	0.000	-2.450	-1.450
rock	keyA#:modeminor	2.124	0.318	6.680	0.000	1.501	2.747
rock	keyB:modeminor	2.159	0.323	6.678	0.000	1.525	2.793
rock	keyC:modeminor	1.259	0.317	3.976	0.000	0.638	1.880
rock	keyC#:modeminor	1.254	0.277	4.525	0.000	0.711	1.797
rock	keyD:modeminor	-0.010	0.310	-0.033	0.974	-0.617	0.597
rock	keyD#:modeminor	1.474	0.422	3.496	0.000	0.648	2.301
rock	keyE:modeminor	0.415	0.296	1.401	0.161	-0.165	0.995
rock	keyF:modeminor	0.208	0.309	0.674	0.500	-0.397	0.814
rock	keyF#:modeminor	0.495	0.295	1.681	0.093	-0.082	1.073
rock	keyG:modeminor	0.837	0.326	2.569	0.010	0.198	1.475

y.level	term	estimate	std.error	statistic	p.value	conf.low	conf.high
rock	keyG#:modeminor	2.204	0.345	6.397	0.000	1.529	2.879