Spotify Genre Classifier Final Report

Problem

- Music streaming services like Spotify are the go to hub for people around the world to stream their favorite music. These streaming platforms offer many services such as recommending songs and artists but a long overdue feature that was recently implemented only on mobile platforms was using tags for genre classification of tracks. But this still falls short of how robust and widespread the genre classification feature needs to be. Improvements and more exposure by increasing the number of labels of classification and spreading the feature to different computer and web applications would greatly improve user satisfaction and revenue for the streaming platform.
- Spotify has about 40-60,000 new songs added to the platform everyday which makes it very difficult for the platform to manually label each and every track that is uploaded. A solution to this is to use a machine learning genre classification model that can make predictions of the genre for each new track that is added. By adding this tag based genre classification feature across all platforms (web, computer, mobile) that use Spotify, Spotify could potentially convert free users of the platform into paying premium users as user satisfaction could increase due to the added feature. Free users might convert to premium users as they find Spotify to be the platform that offers the best and largest number of features to customize their listening experience.

Data Wrangling

- To create this classification system I used a kaggle dataset containing 232,725 tracks classified by the creator of the dataset into 18 columns and 26 genres.
- Kaggle Dataset Link: https://www.kaggle.com/zaheenhamidani/ultimate-spotify-tracks-db
- Example Row Of Dataset:
	genre	artist_name	track_name	track_id	popularity	acousticness	danceability	duration_ms	energy	instrumentalness	key
0	Movie	Henri Salvador	C'est beau de faire un Show	0BRjO6ga9RKCKjfDqeFgWV	0	0.6110	0.389	99373	0.9100	0.000000	C#
liveness	loudness	mode	speechiness	tempo	time_signature	valence
0.3460	-1.828	Major	0.0525	166.969	4/4	0.8140
- 26 genres:
  - 'Movie', 'R&B', 'A Capella', 'Alternative', 'Country', 'Dance', 'Electronic', 'Anime', 'Folk', 'Blues', 'Opera', 'Hip-Hop', "Children's Music", 'Children's Music', 'Rap', 'Indie', 'Classical', 'Pop', 'Reggae', 'Reggaeton', 'Jazz', 'Rock', 'Ska', 'Comedy', 'Soul', 'Soundtrack', 'World'
- 18 columns:
  - Genre, acousticness, artist name, danceability, duration_ms, energy, mode, track id, instrumentalness, key, liveliness, loudness, song name, popularity, speechiness, tempo, valence, and time signature.

- Pandas Profile Report Takeaways
    a. No null values that needed to be addressed in the data nor any duplicate rows.
- Data Modifications I Made
    a. I deleted the rows containing the 'movie' genre from the data set as I felt the label of 'movie' wasn't a very useful or proper genre tag for which the model could classify on.
    b. The second modification I made in this notebook was to combine the genre labels of "Children's Music" and "Children's Music" as the apostrophes in the data file for each of the labels was different, therefore the genres being treated as different when they should be the same. I was able to combine the 2 genres into one by changing the apostrophe in the genre column to be uniform for all 'Children's Music' labels.
- Genre & Track Features Overall Examination
    a. After examining the value counts of each of the genres which had approximately 9000-5000 rows of data except the genre 'A Capella' which only had 119 rows of data.
    b. The data of the column features of the tracks were 'object' for the categorical variables, 'int' for 'duration_ms' and the 'popularity' features, and the remaining numerical columns were float data types.

```
Genre:
 Comedy              9681
Soundtrack           9646
Indie                9543
Jazz                 9441
Pop                  9386
Electronic           9377
Children's Music     9353
Folk                 9299
Hip-Hop              9295
Rock                 9272
Alternative          9263
Classical            9256
Rap                  9232
World                9096
Soul                 9089
Blues                9023
R&B                  8992
Anime                8936
Reggaeton            8927
Ska                  8874
Reggae               8771
Dance                8701
Country              8664
Opera                8280
Movie                7806
Children's Music     5403
A Capella             119
Name: genre, dtype: int64
```

```
genre               object
artist_name         object
track_name          object
track_id            object
popularity           int64
acousticness       float64
danceability       float64
duration_ms          int64
energy             float64
instrumentalness   float64
key                 object
liveness           float64
loudness           float64
mode                object
speechiness        float64
tempo              float64
time_signature      object
valence            float64
dtype: object
```

- Categorical Column Features Examination

a. I discovered the following unique values for the important categorical features (after 'genre' column modification explained above):

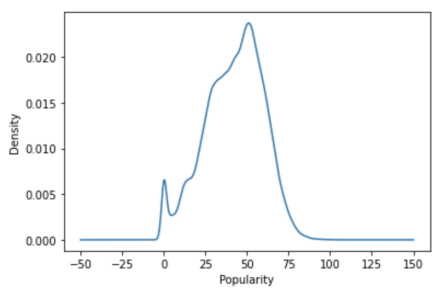
```
Genre:
 ['R&B' 'A Capella' 'Alternative' 'Country' 'Dance' 'Electronic' 'Anime'
 'Folk' 'Blues' 'Opera' 'Hip-Hop' 'Children's Music' 'Rap' 'Indie'
 'Classical' 'Pop' 'Reggae' 'Reggaeton' 'Jazz' 'Rock' 'Ska' 'Comedy'
 'Soul' 'Soundtrack' 'World']

Key:
 ['D' 'C' 'F' 'B' 'E' 'G' 'G#' 'A#' 'C#' 'A' 'F#' 'D#']

Time Signature:
 ['4/4' '3/4' '5/4' '1/4' '0/4']

Mode:
 ['Minor' 'Major']
```

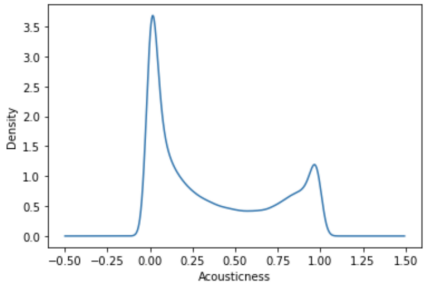
- Numerical Column Features Examination
  a. I examined the numerical columns data by getting the median, mode, and standard deviation of the track features:

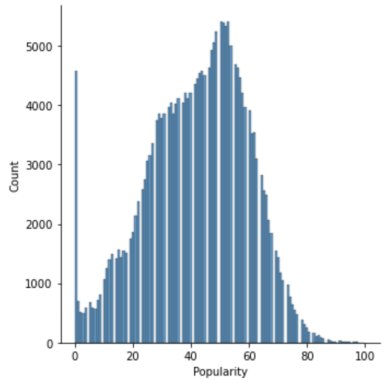
```
popularity:
 median:  44.0  mode: 0    53
dtype: int64 std: 17.487793638151285

acousticness:
 median:  0.21600000000000003  mode: 0    0.995
dtype: float64 std: 0.35167706648943226

danceability:
 median:  0.573  mode: 0    0.597
dtype: float64 std: 0.18545232652268254

duration_ms:
 median:  221387.0  mode: 0    240000
dtype: int64 std: 114234.1463242093

energy:
 median:  0.614  mode: 0    0.721
dtype: float64 std: 0.26157118554535336

instrumentalness:
 median:  4.7e-05  mode: 0    0.0
dtype: float64 std: 0.30343474081654914

liveness:
 median:  0.128  mode: 0    0.111
dtype: float64 std: 0.19829119729478722

loudness:
 median:  -7.643  mode: 0   -5.318
dtype: float64 std: 5.976155633206901

speechiness:
 median:  0.0504  mode: 0    0.0337
dtype: float64 std: 0.18572328191291343

tempo:
 median:  115.97  mode: 0    120.016
dtype: float64 std: 30.909935166135575

valence:
 median:  0.445  mode: 0    0.961
dtype: float64 std: 0.2593840383530541
```

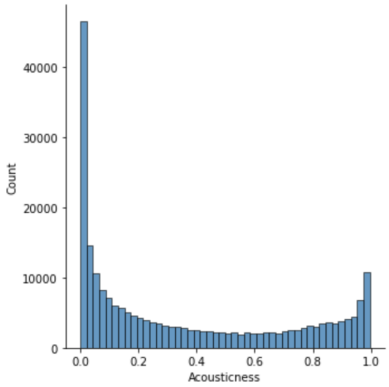
b. I also plotted the density and count distribution plots for the numerical columns which showed skewed distributions for multiple features of track data:

- Popularity & Acousticness

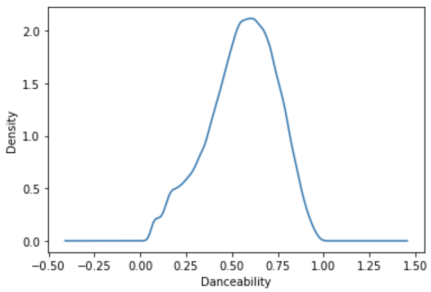


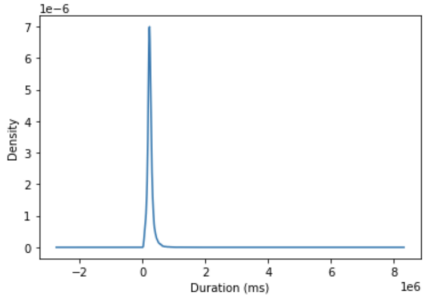


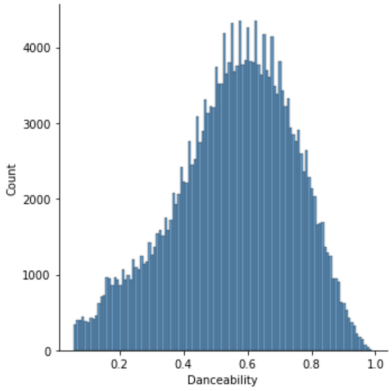


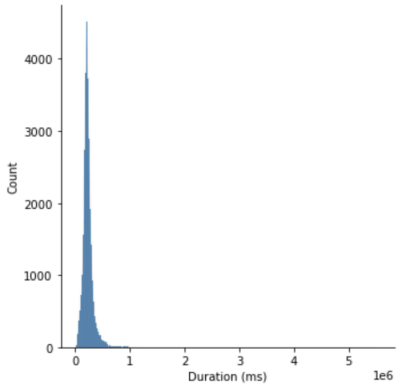


- Danceability & Duration

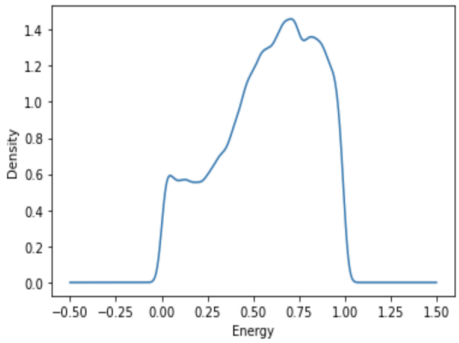


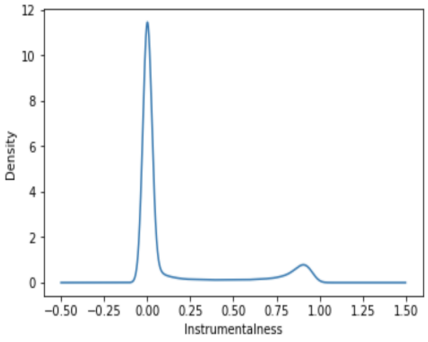


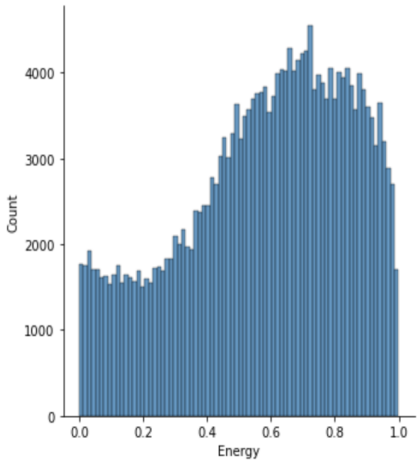


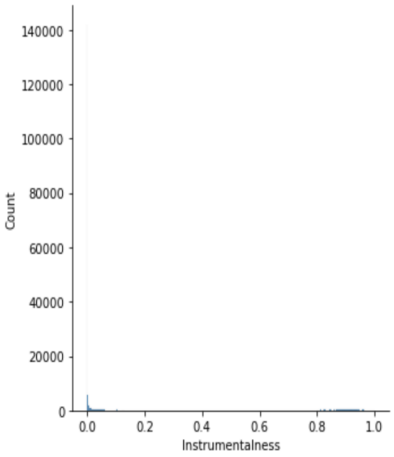


- Energy & Instrumentalness

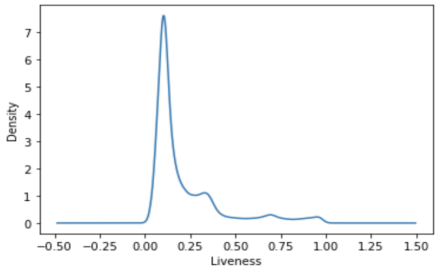


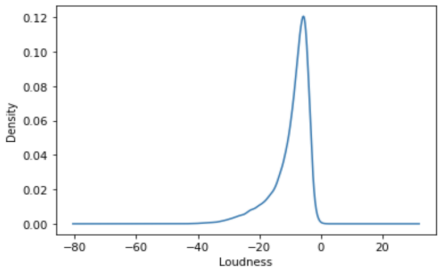


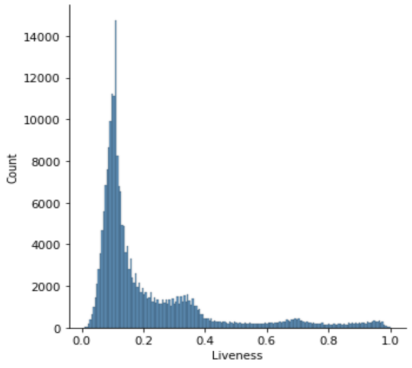


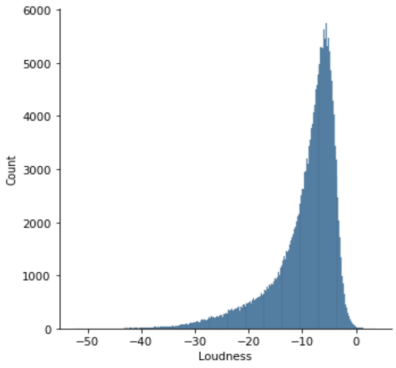


- Liveness & Loudness

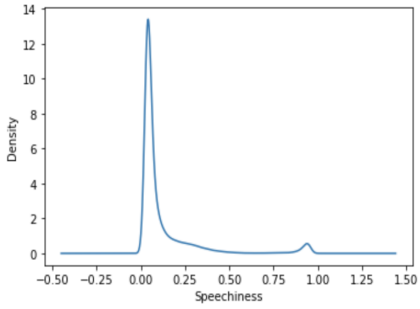


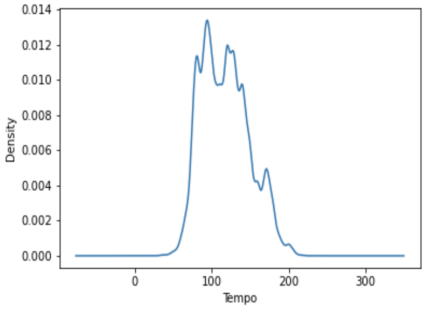


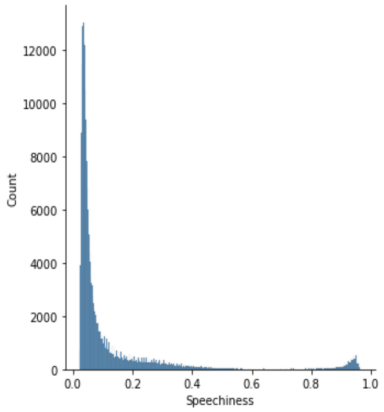


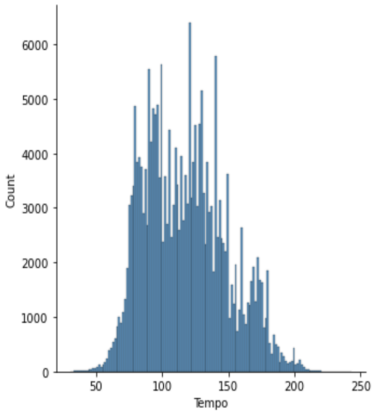


- Speechiness & Tempo

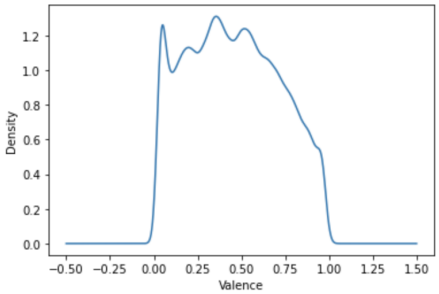


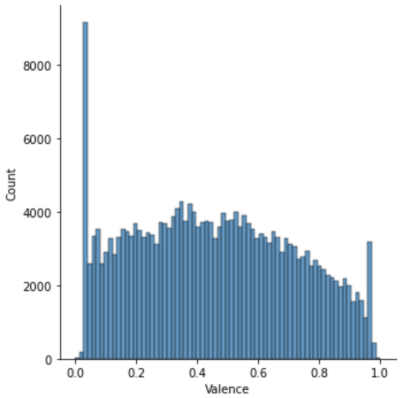






- Valence



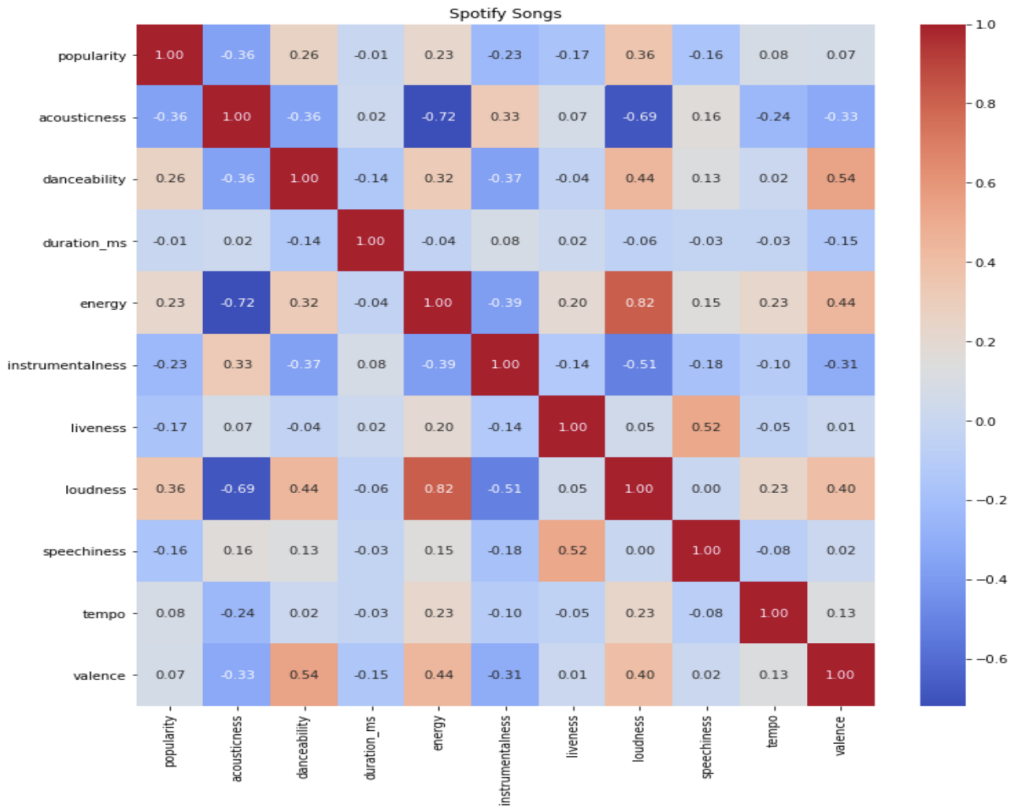


- After examining the Spotify API, I was able to learn more about the description of the features and their range of data:
    - 7 categorical: genre, artist_name, track_name, track_id, key, mode, time_signature
    - 11 numerical: popularity, acousticness, danceability, duration_ms, energy, instrumentalness, liveness, loudness, speechiness, tempo, valence
    - Genre: There are 25 genres of music after editing the data set.
    - Artist_name: Name of artist
    - Track_name: Name of track
    - Track_id: ID of track
      Key: ['D' 'C' 'F' 'B' 'E' 'G' 'G#' 'A#' 'C#' 'A' 'F#' 'D#']. The key or pitch the track is in
    - Mode: [Major, Minor]. Modality of track, the type of scale from which its melodic content is derived.
    - Time_signature: ['4/4' '3/4' '5/4' '1/4' '0/4']. An estimated overall time signature of a track. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure).
    - Popularity: Scale: [0, 100]. The popularity of the track with users on Spotify. 0 to 100 is an increase in popularity.
    - Acousticness: Scale: [0.0, 1.0]. A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
    - Danceability: Scale: [0.0, 1.0]. Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
    - Duration_ms: The duration of the track in milliseconds.
    - Energy: Scale: [0.0, 1.0]. It represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy.
    - Instrumentalness: Scale: [0.0, 1.0]. Predicts whether a track contains no vocals. “Ooh” and “aah” sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly “vocal”. The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.
    - Liveliness: [0.0, 1.0]. Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.
    - Loudness: Scale: [-60, 0]. The overall loudness of a track in decibels (dB).

- Speechiness: Scale: [0.0, 1.0]. Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.
- Tempo: The overall estimated tempo of a track in beats per minute (BPM).
- Valence: Scale: [0.0, 1.0]. Describes the musical positiveness conveyed by the track. High valence tracks sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).

Exploratory Data Analysis (EDA)

- Column Feature Correlation w./ Seaborn Heatmap & Pairplot
  a. I plotted a seaborn correlation heatmap to understand the relationship between the 11 numerical attributes of a track and found energy and loudness had a strong correlation.



  b. I also plotted a seaborn pairplot to see the correlation between features of a track.

- <u>25 Genres & Associated Attributes Summary Statistics</u>
    a. I created a table to show the summary statistics of each of the attributes of a track.

	popularity	acousticness	danceability	duration_ms	energy	instrumentalness	liveness	loudness	speechiness	tempo	valence
count	224919.000000	224919.000000	224919.000000	2.249190e+05	224919.000000	224919.000000	224919.000000	224919.000000	224919.000000	224919.000000	224919.000000
mean	42.132354	0.357150	0.556557	2.359802e+05	0.577908	0.149095	0.214534	-9.452503	0.121159	117.795684	0.455164
std	17.487794	0.351677	0.185452	1.142341e+05	0.261571	0.303435	0.198291	5.976156	0.185723	30.909935	0.259384
min	0.000000	0.000000	0.056900	1.538700e+04	0.000020	0.000000	0.009670	-52.457000	0.022200	30.379000	0.000000
25%	30.000000	0.034700	0.439000	1.844780e+05	0.399000	0.000000	0.097100	-11.530000	0.036800	92.992000	0.239000
50%	44.000000	0.216000	0.573000	2.213870e+05	0.614000	0.000047	0.128000	-7.643000	0.050400	115.970000	0.445000
75%	55.000000	0.697000	0.694000	2.665470e+05	0.791000	0.037700	0.263000	-5.450000	0.106000	139.479000	0.660000
max	100.000000	0.996000	0.989000	5.552917e+06	0.999000	0.999000	1.000000	3.744000	0.967000	242.903000	1.000000

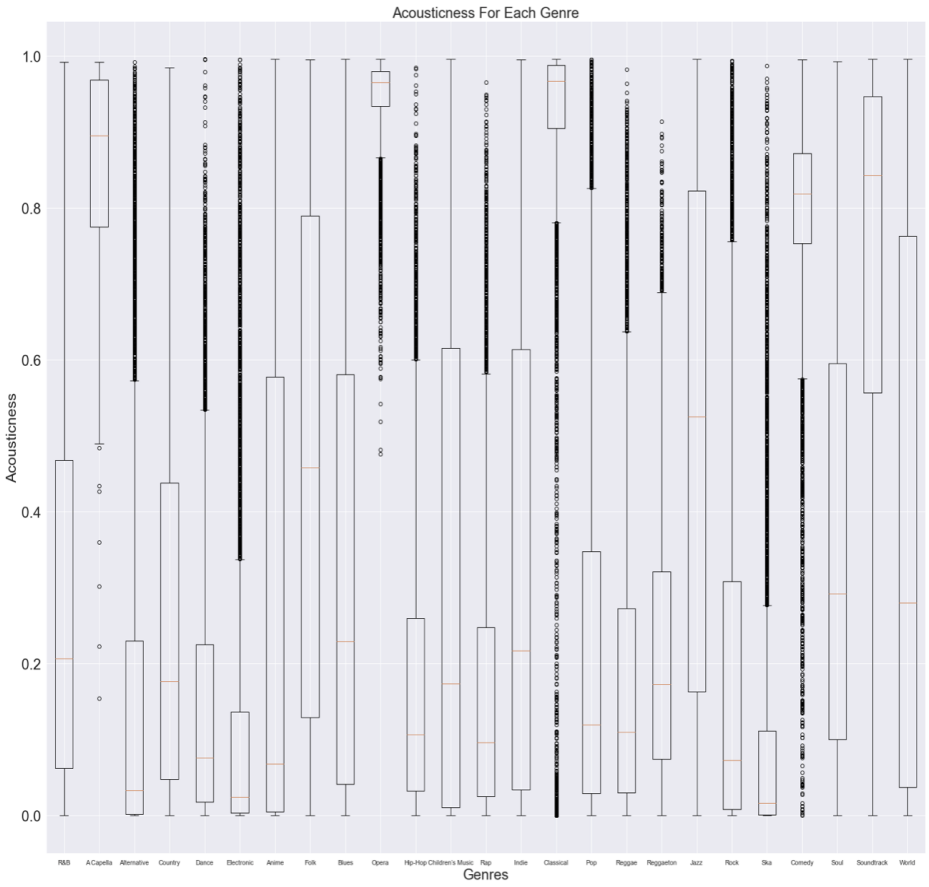
    b. I created a table to show the summary statistics grouped by genre of each attribute of a track.

	popularity								acousticness		...	tempo		valence				
	count	mean	std	min	25%	50%	75%	max	count	mean	...	75%	max	count	mean	std	min	25%
genre																		
A Capella	119.0	9.302521	7.868145	0.0	4.0	8.0	13.0	44.0	119.0	0.829941	...	129.68300	181.714	119.0	0.328724	0.255005	0.0380	0.1
Alternative	9263.0	50.213430	7.661040	0.0	45.0	49.0	55.0	83.0	9263.0	0.162313	...	143.96550	213.788	9263.0	0.449590	0.216426	0.0321	0.2
Anime	8936.0	24.258729	9.648703	0.0	17.0	23.0	30.0	65.0	8936.0	0.286843	...	149.99125	220.276	8936.0	0.441682	0.249619	0.0000	0.2
Blues	9023.0	34.742879	9.755328	0.0	28.0	33.0	40.0	80.0	9023.0	0.327840	...	140.63550	242.903	9023.0	0.579425	0.224677	0.0315	0.4
Children's Music	14756.0	36.202426	25.536529	0.0	2.0	49.0	56.0	86.0	14756.0	0.320112	...	140.09250	220.119	14756.0	0.532251	0.250929	0.0000	0.3
Classical	9256.0	29.282195	14.133541	0.0	25.0	32.0	38.0	69.0	9256.0	0.868843	...	127.18125	212.923	9256.0	0.214463	0.200275	0.0000	0.0
Comedy	9681.0	21.342630	8.428764	0.0	15.0	20.0	26.0	61.0	9681.0	0.793098	...	115.12800	207.157	9681.0	0.412764	0.207258	0.0237	0.2
Country	8664.0	46.100416	9.745975	0.0	39.0	45.0	52.0	82.0	8664.0	0.270172	...	144.48075	217.538	8664.0	0.535160	0.219819	0.0395	0.3
Dance	8701.0	57.275256	11.208370	0.0	51.0	57.0	64.0	100.0	8701.0	0.152888	...	134.03000	218.081	8701.0	0.517754	0.226822	0.0340	0.3
Electronic	9377.0	38.056095	9.741981	0.0	31.0	37.0	44.0	96.0	9377.0	0.119839	...	144.99000	220.169	9377.0	0.388129	0.236938	0.0205	0.1
Folk	9299.0	49.940209	8.218284	0.0	44.0	49.0	55.0	84.0	9299.0	0.463201	...	136.30850	236.799	9299.0	0.440237	0.241416	0.0277	0.2
Hip-Hop	9295.0	58.423131	8.269761	14.0	52.0	57.0	63.0	98.0	9295.0	0.176172	...	141.98650	214.126	9295.0	0.473381	0.222325	0.0336	0.3
Indie	9543.0	54.701561	7.355754	0.0	49.0	54.0	59.0	97.0	9543.0	0.331214	...	137.93400	219.331	9543.0	0.428665	0.221606	0.0277	0.2
Jazz	9441.0	40.824383	9.588840	0.0	35.0	40.0	46.0	79.0	9441.0	0.499606	...	127.87100	239.848	9441.0	0.508961	0.251218	0.0266	0.3
Opera	8280.0	13.335628	8.460264	0.0	7.0	11.0	17.0	63.0	8280.0	0.945202	...	120.66900	236.735	8280.0	0.189864	0.172322	0.0207	0.0
Pop	9386.0	66.590667	7.248797	3.0	62.0	66.0	71.0	100.0	9386.0	0.224819	...	140.01575	213.990	9386.0	0.481371	0.225029	0.0277	0.3
R&B	8992.0	52.308719	9.246359	0.0	46.0	51.0	58.0	92.0	8992.0	0.288216	...	134.17175	216.636	8992.0	0.450346	0.215387	0.0321	0.2
Rap	9232.0	60.533795	8.177226	14.0	55.0	59.0	65.0	99.0	9232.0	0.168080	...	142.00050	216.115	9232.0	0.455918	0.213913	0.0331	0.2
Reggae	8771.0	35.589328	10.779762	0.0	28.0	34.0	42.0	78.0	8771.0	0.185783	...	142.31400	218.184	8771.0	0.679665	0.198141	0.0331	0.5
Reggaeton	8927.0	37.742915	13.544414	0.0	28.0	35.0	46.0	98.0	8927.0	0.218923	...	146.03850	234.923	8927.0	0.659439	0.202052	0.0381	0.5
Rock	9272.0	59.619392	7.474083	0.0	54.0	59.0	64.0	95.0	9272.0	0.196429	...	142.00875	219.331	9272.0	0.517113	0.231137	0.0277	0.3
Ska	8874.0	28.612351	10.757129	5.0	21.0	27.0	35.0	74.0	8874.0	0.099728	...	155.57725	221.578	8874.0	0.653472	0.223245	0.0331	0.4
Soul	9089.0	47.027836	9.235035	0.0	41.0	46.0	53.0	85.0	9089.0	0.360679	...	132.44500	216.636	9089.0	0.480562	0.245857	0.0287	0.2
Soundtrack	9646.0	33.954800	8.639645	0.0	28.0	33.0	39.0	72.0	9646.0	0.717349	...	126.14200	216.429	9646.0	0.118483	0.139913	0.0000	0.0
World	9096.0	35.524077	9.399816	0.0	29.0	34.0	41.0	76.0	9096.0	0.393341	...	141.55675	212.923	9096.0	0.295657	0.230914	0.0000	0.1

c. I displayed the summary statistics and plotted box plots grouped genre for each of the 9 important numerical features of a track: 'acousticness', 'danceability', 'energy', 'instrumentalness', 'liveness', 'loudness', 'speechiness', 'tempo', and 'valence' for each genre. The data showed many outliers when grouped by each genre as seen in the box plots.

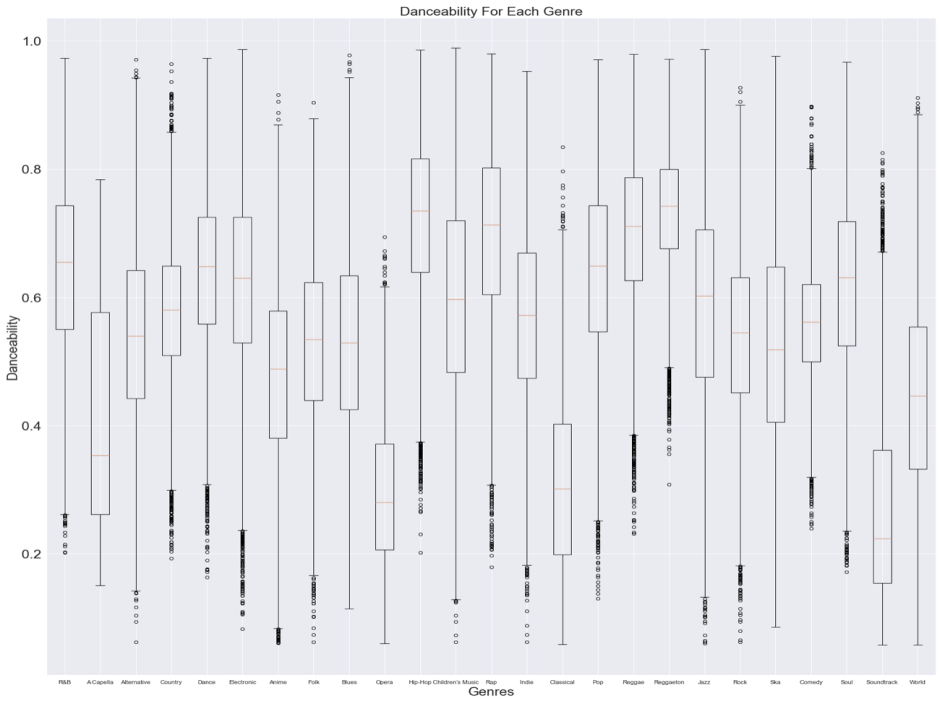
   i. <u>Acousticness by Genre</u>

genre	count	mean	std	min	25%	50%	75%	max
A Capella	119.0	0.829941	0.181866	0.154000	0.775500	0.89500	0.96900	0.992
Alternative	9263.0	0.162313	0.241155	0.000001	0.001950	0.03350	0.23050	0.992
Anime	8936.0	0.286843	0.362341	0.000000	0.005050	0.06795	0.57725	0.996
Blues	9023.0	0.327840	0.309981	0.000001	0.041100	0.22900	0.58100	0.996
Children's Music	14756.0	0.320112	0.339331	0.000001	0.011200	0.17350	0.61525	0.996
Classical	9256.0	0.868843	0.255269	0.000001	0.905000	0.96700	0.98800	0.996
Comedy	9681.0	0.793098	0.130313	0.000363	0.753000	0.81900	0.87200	0.995
Country	8664.0	0.270172	0.262801	0.000028	0.048000	0.17700	0.43800	0.985
Dance	8701.0	0.152888	0.184252	0.000004	0.018500	0.07600	0.22500	0.996
Electronic	9377.0	0.119839	0.200477	0.000002	0.003610	0.02460	0.13700	0.995
Folk	9299.0	0.463201	0.334784	0.000001	0.129000	0.45800	0.79000	0.995
Hip-Hop	9295.0	0.176172	0.188891	0.000015	0.033000	0.10700	0.26000	0.985
Indie	9543.0	0.331214	0.321618	0.000001	0.034100	0.21700	0.61400	0.995
Jazz	9441.0	0.499606	0.337637	0.000001	0.163000	0.52500	0.82300	0.996
Opera	8280.0	0.945202	0.057516	0.476000	0.934000	0.96500	0.98000	0.996
Pop	9386.0	0.224819	0.250306	0.000006	0.029200	0.12000	0.34800	0.995
R&B	8992.0	0.288216	0.262520	0.000030	0.062075	0.20700	0.46800	0.992
Rap	9232.0	0.168080	0.189780	0.000007	0.025700	0.09600	0.24825	0.965
Reggae	8771.0	0.185783	0.204736	0.000004	0.030000	0.11000	0.27300	0.982
Reggaeton	8927.0	0.218923	0.180025	0.000010	0.074850	0.17300	0.32100	0.914
Rock	9272.0	0.196429	0.252861	0.000001	0.008770	0.07310	0.30825	0.994
Ska	8874.0	0.099728	0.174256	0.000001	0.001550	0.01625	0.11200	0.987
Soul	9089.0	0.360679	0.288262	0.000014	0.100000	0.29200	0.59500	0.993
Soundtrack	9646.0	0.717349	0.292065	0.000004	0.557000	0.84250	0.94700	0.996
World	9096.0	0.393341	0.363125	0.000002	0.037300	0.28000	0.76300	0.996



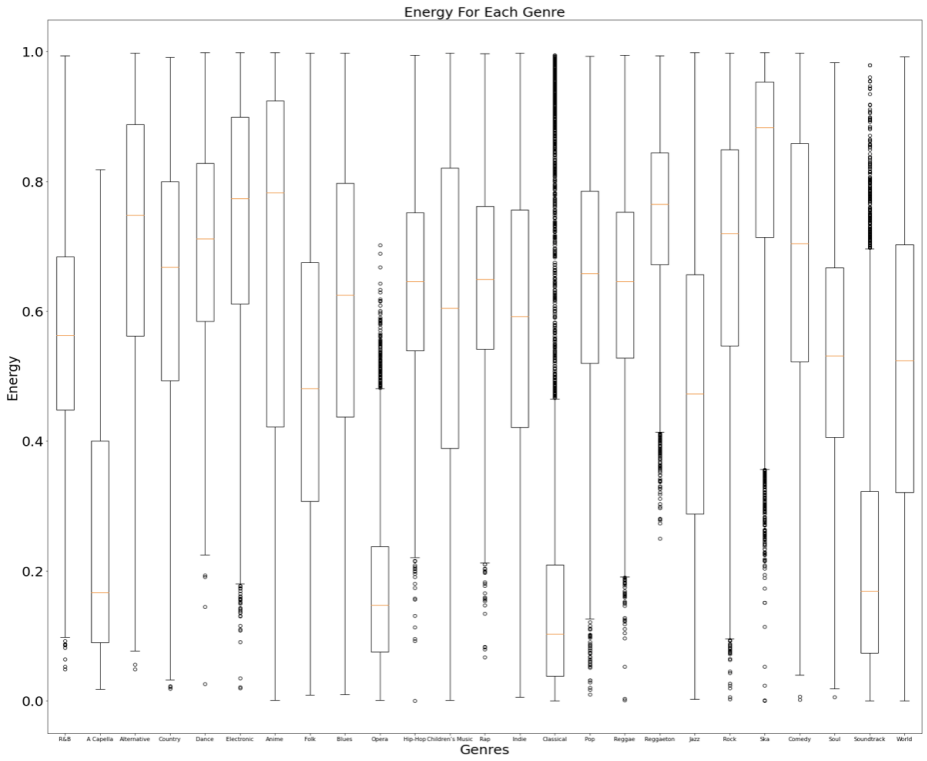
ii. Danceability by Genre

genre	count	mean	std	min	25%	50%	75%	max
A Capella	119.0	0.412252	0.179719	0.1500	0.2610	0.3530	0.57600	0.784
Alternative	9263.0	0.541898	0.150391	0.0617	0.4420	0.5390	0.64200	0.971
Anime	8936.0	0.472090	0.149229	0.0600	0.3800	0.4880	0.57825	0.916
Blues	9023.0	0.528232	0.145147	0.1140	0.4245	0.5290	0.63350	0.978
Children's Music	14756.0	0.598829	0.168058	0.0617	0.4830	0.5970	0.72000	0.989
Classical	9256.0	0.305958	0.135201	0.0582	0.1980	0.3010	0.40200	0.834
Comedy	9681.0	0.559038	0.089625	0.2390	0.4990	0.5610	0.62000	0.898
Country	8664.0	0.577038	0.109771	0.1920	0.5090	0.5800	0.64900	0.964
Dance	8701.0	0.638191	0.126575	0.1630	0.5580	0.6480	0.72500	0.973
Electronic	9377.0	0.619542	0.146418	0.0822	0.5290	0.6300	0.72500	0.987
Folk	9299.0	0.527276	0.132496	0.0617	0.4390	0.5340	0.62300	0.904
Hip-Hop	9295.0	0.718808	0.130642	0.2010	0.6390	0.7350	0.81600	0.986
Indie	9543.0	0.566821	0.140346	0.0617	0.4740	0.5720	0.66900	0.953
Jazz	9441.0	0.585638	0.158903	0.0596	0.4750	0.6020	0.70500	0.987
Opera	8280.0	0.290650	0.112660	0.0592	0.2060	0.2800	0.37100	0.694
Pop	9386.0	0.640236	0.141266	0.1300	0.5460	0.6485	0.74300	0.971
R&B	8992.0	0.642125	0.137701	0.2010	0.5500	0.6550	0.74300	0.973
Rap	9232.0	0.697244	0.141120	0.1790	0.6040	0.7130	0.80200	0.980
Reggae	8771.0	0.699271	0.121033	0.2310	0.6260	0.7110	0.78700	0.979
Reggaeton	8927.0	0.731260	0.096116	0.3080	0.6760	0.7420	0.80000	0.972
Rock	9272.0	0.538292	0.133388	0.0617	0.4510	0.5450	0.63100	0.927
Ska	8874.0	0.526799	0.164059	0.0850	0.4050	0.5180	0.64700	0.976
Soul	9089.0	0.617645	0.137804	0.1710	0.5240	0.6310	0.71800	0.967
Soundtrack	9646.0	0.265616	0.150734	0.0572	0.1540	0.2230	0.36100	0.825
World	9096.0	0.443293	0.166094	0.0569	0.3320	0.4460	0.55400	0.911



iii. Energy by Genre

genre	count	mean	std	min	25%	50%	75%	max
A Capella	119.0	0.250313	0.198740	0.018000	0.09005	0.1670	0.40050	0.818
Alternative	9263.0	0.711519	0.206063	0.048400	0.56200	0.7480	0.88800	0.998
Anime	8936.0	0.665356	0.299668	0.000943	0.42200	0.7830	0.92400	0.999
Blues	9023.0	0.606171	0.229498	0.009570	0.43750	0.6250	0.79700	0.998
Children's Music	14756.0	0.593204	0.253964	0.000499	0.38900	0.6050	0.82100	0.998
Classical	9256.0	0.177984	0.225483	0.000020	0.03840	0.1030	0.20925	0.995
Comedy	9681.0	0.676094	0.211705	0.001410	0.52200	0.7040	0.85900	0.998
Country	8664.0	0.636318	0.200344	0.018800	0.49300	0.6680	0.80000	0.991
Dance	8701.0	0.698067	0.160860	0.025900	0.58500	0.7120	0.82800	0.999
Electronic	9377.0	0.739299	0.188646	0.019300	0.61100	0.7740	0.89900	0.999
Folk	9299.0	0.491733	0.230968	0.009320	0.30700	0.4810	0.67500	0.998
Hip-Hop	9295.0	0.643275	0.150037	0.000243	0.53900	0.6460	0.75200	0.995
Indie	9543.0	0.581002	0.217834	0.005540	0.42100	0.5920	0.75600	0.998
Jazz	9441.0	0.472776	0.237807	0.002110	0.28800	0.4730	0.65700	0.999
Opera	8280.0	0.168779	0.117803	0.000909	0.07560	0.1470	0.23800	0.702
Pop	9386.0	0.642208	0.182855	0.009930	0.52000	0.6580	0.78500	0.993
R&B	8992.0	0.564248	0.167602	0.048400	0.44800	0.5630	0.68400	0.994
Rap	9232.0	0.650520	0.156686	0.066800	0.54200	0.6495	0.76200	0.997
Reggae	8771.0	0.635749	0.161206	0.000957	0.52800	0.6460	0.75300	0.995
Reggaeton	8927.0	0.748836	0.125012	0.250000	0.67200	0.7650	0.84400	0.994
Rock	9272.0	0.683670	0.203043	0.002590	0.54700	0.7200	0.84900	0.998
Ska	8874.0	0.815585	0.175297	0.000267	0.71425	0.8830	0.95300	0.999
Soul	9089.0	0.532506	0.184366	0.005900	0.40600	0.5310	0.66700	0.983
Soundtrack	9646.0	0.221110	0.186378	0.000020	0.07390	0.1690	0.32300	0.979
World	9096.0	0.506453	0.252054	0.000098	0.32100	0.5240	0.70300	0.992

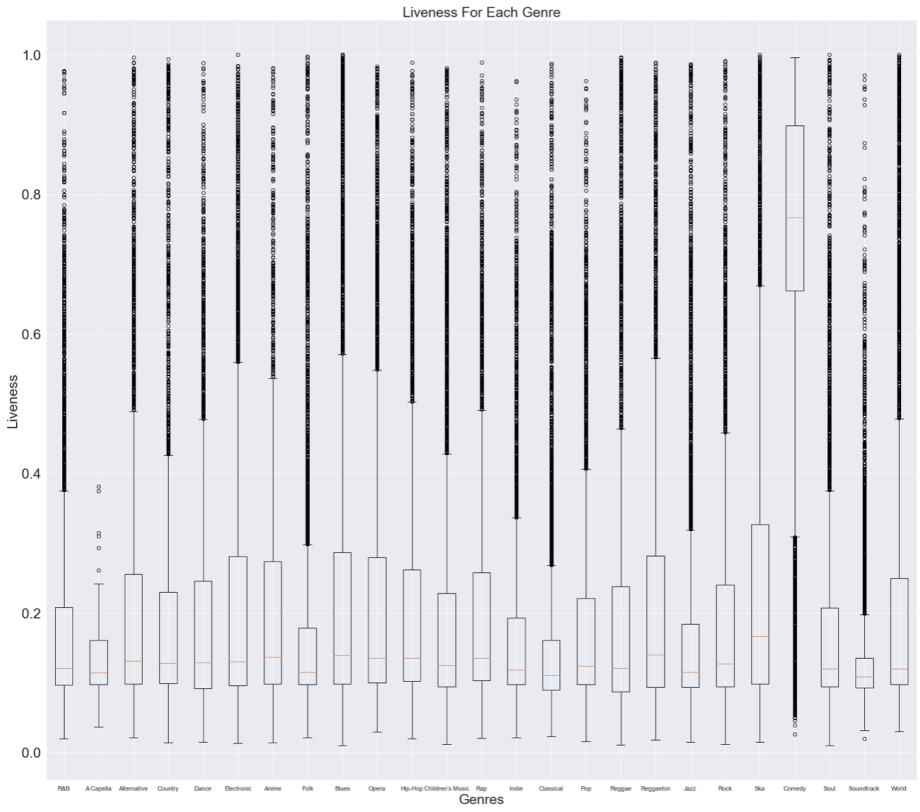


iv. Instrumentalness by Genre

genre	count	mean	std	min	25%	50%	75%	max
A Capella	119.0	0.007491	0.040491	0.0	0.000000	0.000000	0.000097	0.336
Alternative	9263.0	0.061303	0.176176	0.0	0.000000	0.000073	0.006635	0.955
Anime	8936.0	0.280592	0.391047	0.0	0.000000	0.000414	0.765000	0.997
Blues	9023.0	0.095175	0.213789	0.0	0.000010	0.001030	0.040650	0.975
Children's Music	14756.0	0.087013	0.230714	0.0	0.000000	0.000023	0.005352	0.997
Classical	9256.0	0.599425	0.377911	0.0	0.145000	0.829000	0.905000	0.994
Comedy	9681.0	0.000574	0.009987	0.0	0.000000	0.000000	0.000000	0.352
Country	8664.0	0.005610	0.041811	0.0	0.000000	0.000001	0.000044	0.880
Dance	8701.0	0.035449	0.143123	0.0	0.000000	0.000001	0.000221	0.979
Electronic	9377.0	0.350955	0.361364	0.0	0.003240	0.190000	0.751000	0.994
Folk	9299.0	0.084934	0.209306	0.0	0.000003	0.000330	0.021850	0.985
Hip-Hop	9295.0	0.011200	0.073822	0.0	0.000000	0.000000	0.000012	0.927
Indie	9543.0	0.085317	0.209438	0.0	0.000000	0.000174	0.018750	0.976
Jazz	9441.0	0.358009	0.383488	0.0	0.000204	0.133000	0.801000	0.985
Opera	8280.0	0.232013	0.347569	0.0	0.000164	0.009140	0.448250	0.994
Pop	9386.0	0.016599	0.094935	0.0	0.000000	0.000000	0.000045	0.972
R&B	8992.0	0.025558	0.115646	0.0	0.000000	0.000001	0.000221	0.969
Rap	9232.0	0.009317	0.065828	0.0	0.000000	0.000000	0.000010	0.951
Reggae	8771.0	0.039972	0.150182	0.0	0.000000	0.000006	0.000707	0.961
Reggaeton	8927.0	0.003666	0.029829	0.0	0.000000	0.000000	0.000015	0.726
Rock	9272.0	0.053288	0.165608	0.0	0.000000	0.000047	0.004293	0.974
Ska	8874.0	0.056709	0.173950	0.0	0.000000	0.000042	0.003455	0.963
Soul	9089.0	0.062295	0.182135	0.0	0.000000	0.000061	0.005800	0.970
Soundtrack	9646.0	0.783611	0.245675	0.0	0.772000	0.881000	0.927000	0.999
World	9096.0	0.233882	0.369701	0.0	0.000000	0.000131	0.542000	0.996

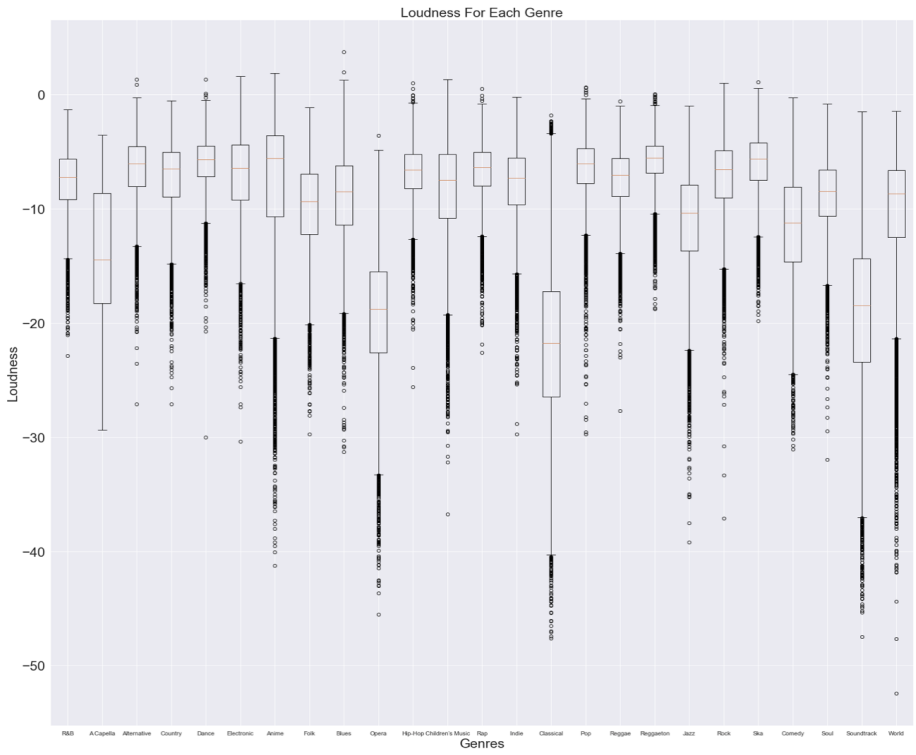
v. Liveness by Genre

genre	count	mean	std	min	25%	50%	75%	max
A Capella	119.0	0.136924	0.063144	0.03690	0.097600	0.114	0.1605	0.381
Alternative	9263.0	0.196985	0.157880	0.02110	0.098300	0.131	0.2550	0.996
Anime	8936.0	0.192391	0.139406	0.01370	0.098200	0.137	0.2740	0.981
Blues	9023.0	0.233125	0.219053	0.00967	0.098100	0.139	0.2870	1.000
Children's Music	14756.0	0.183986	0.148114	0.01190	0.094300	0.125	0.2280	0.981
Classical	9256.0	0.162810	0.145103	0.02250	0.089500	0.110	0.1610	0.987
Comedy	9681.0	0.724775	0.223437	0.02580	0.662000	0.767	0.8980	0.996
Country	8664.0	0.187216	0.153139	0.01430	0.098775	0.128	0.2300	0.994
Dance	8701.0	0.187753	0.148386	0.01490	0.092000	0.129	0.2460	0.988
Electronic	9377.0	0.210006	0.177842	0.01300	0.095500	0.130	0.2810	1.000
Folk	9299.0	0.170773	0.144907	0.02110	0.097800	0.115	0.1780	0.998
Hip-Hop	9295.0	0.201146	0.156169	0.01960	0.102000	0.135	0.2620	0.989
Indie	9543.0	0.168919	0.126682	0.02110	0.097400	0.118	0.1930	0.962
Jazz	9441.0	0.173355	0.151554	0.01460	0.093800	0.115	0.1840	0.986
Opera	8280.0	0.223264	0.193768	0.02960	0.099500	0.135	0.2790	0.983
Pop	9386.0	0.179967	0.135340	0.01570	0.097200	0.124	0.2210	0.962
R&B	8992.0	0.175350	0.135170	0.01960	0.096475	0.121	0.2080	0.977
Rap	9232.0	0.198939	0.152349	0.02080	0.102750	0.135	0.2580	0.989
Reggae	8771.0	0.193076	0.181150	0.01050	0.086850	0.121	0.2380	0.996
Reggaeton	8927.0	0.207510	0.171025	0.01770	0.093100	0.140	0.2820	0.989
Rock	9272.0	0.186981	0.149458	0.01200	0.094500	0.127	0.2400	0.991
Ska	8874.0	0.243473	0.207786	0.01520	0.098400	0.166	0.3270	1.000
Soul	9089.0	0.179252	0.150928	0.00967	0.094400	0.120	0.2070	1.000
Soundtrack	9646.0	0.137555	0.093915	0.02000	0.092800	0.109	0.1350	0.970
World	9096.0	0.227793	0.228101	0.02990	0.097700	0.120	0.2500	1.000



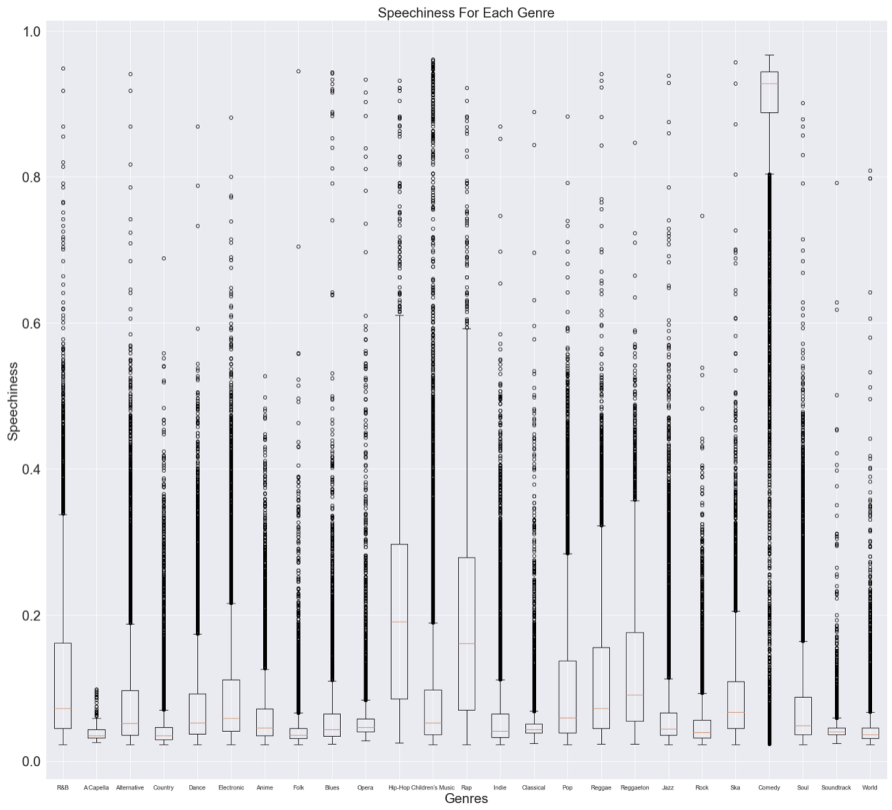
vi. Loudness by Genre

genre	count	mean	std	min	25%	50%	75%	max
A Capella	119.0	-13.660387	5.437118	-29.368	-18.27300	-14.4640	-8.61050	-3.555
Alternative	9263.0	-6.540803	2.764235	-27.119	-8.02650	-6.0280	-4.52200	1.342
Anime	8936.0	-7.917802	6.184689	-41.234	-10.69125	-5.6015	-3.58375	1.893
Blues	9023.0	-9.053807	3.855575	-31.284	-11.39800	-8.4830	-6.20350	3.744
Children's Music	14756.0	-8.399591	4.201918	-36.721	-10.82700	-7.4870	-5.20275	1.342
Classical	9256.0	-21.544477	7.682214	-47.599	-26.45925	-21.7970	-17.23300	-1.786
Comedy	9681.0	-11.689321	4.777983	-31.047	-14.64900	-11.2090	-8.08800	-0.255
Country	8664.0	-7.341693	3.250495	-27.119	-8.94300	-6.5090	-5.01900	-0.521
Dance	8701.0	-6.054241	2.334848	-30.016	-7.19100	-5.6890	-4.48700	1.342
Electronic	9377.0	-7.035868	3.705659	-30.361	-9.23900	-6.4500	-4.37500	1.585
Folk	9299.0	-9.870282	3.880259	-29.729	-12.21850	-9.3340	-6.93400	-1.101
Hip-Hop	9295.0	-6.860286	2.449047	-25.602	-8.19800	-6.5590	-5.20500	1.012
Indie	9543.0	-7.915142	3.277045	-29.729	-9.61950	-7.3030	-5.55550	-0.198
Jazz	9441.0	-11.210457	4.622718	-39.198	-13.68700	-10.3530	-7.89000	-1.002
Opera	8280.0	-19.339767	5.618621	-45.539	-22.60550	-18.7630	-15.47725	-3.581
Pop	9386.0	-6.495423	2.715295	-29.729	-7.75000	-6.0280	-4.70500	0.634
R&B	8992.0	-7.597064	2.742233	-22.881	-9.15700	-7.2125	-5.64825	-1.289
Rap	9232.0	-6.669916	2.450655	-22.589	-7.97075	-6.3470	-5.01475	0.496
Reggae	8771.0	-7.518107	2.785819	-27.689	-8.91350	-7.0220	-5.59150	-0.586
Reggaeton	8927.0	-5.875960	2.162146	-18.779	-6.86250	-5.5300	-4.47700	0.070
Rock	9272.0	-7.285875	3.333181	-37.124	-9.03525	-6.5260	-4.88175	1.023
Ska	8874.0	-6.172705	2.780476	-19.811	-7.51075	-5.6035	-4.22925	1.100
Soul	9089.0	-8.866409	3.252463	-31.981	-10.62200	-8.4310	-6.56400	-0.810
Soundtrack	9646.0	-19.282684	6.735061	-47.499	-23.43400	-18.4745	-14.37125	-1.497
World	9096.0	-10.705435	6.276421	-52.457	-12.51125	-8.6650	-6.61175	-1.462



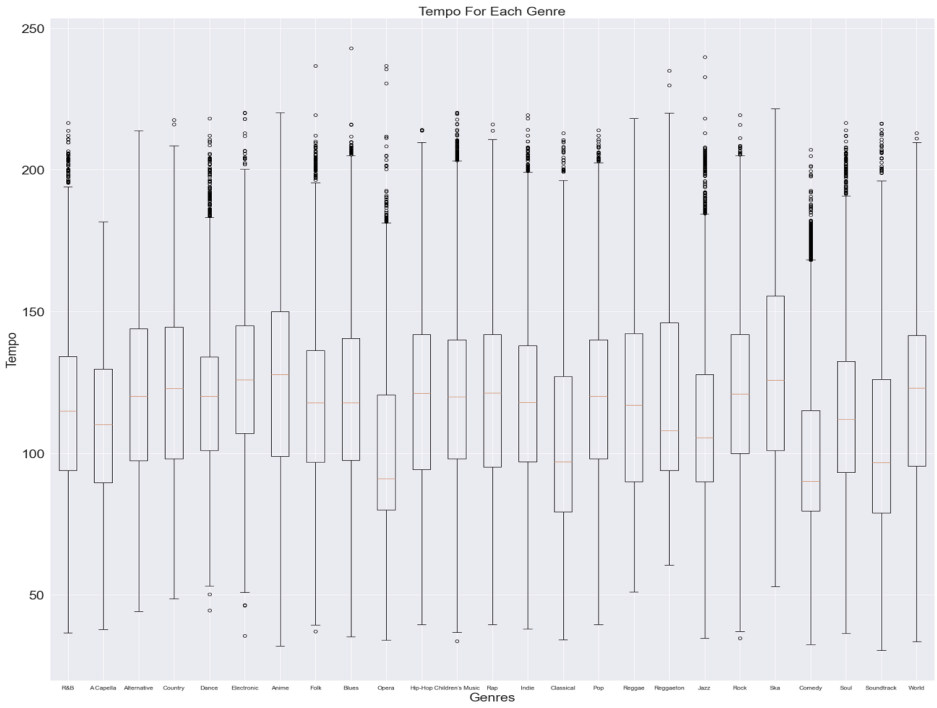
vii. Speechiness by Genre

genre	count	mean	std	min	25%	50%	75%	max
A Capella	119.0	0.042414	0.018254	0.0254	0.03175	0.03450	0.043350	0.0983
Alternative	9263.0	0.088783	0.091241	0.0229	0.03580	0.05200	0.096900	0.9410
Anime	8936.0	0.065102	0.053889	0.0227	0.03470	0.04515	0.071525	0.5270
Blues	9023.0	0.061809	0.061537	0.0231	0.03380	0.04310	0.064500	0.9430
Children's Music	14756.0	0.097763	0.123095	0.0229	0.03630	0.05230	0.097400	0.9610
Classical	9256.0	0.052001	0.040798	0.0238	0.03840	0.04330	0.050500	0.8890
Comedy	9681.0	0.853532	0.205918	0.0230	0.88800	0.92800	0.944000	0.9670
Country	8664.0	0.048989	0.048392	0.0223	0.02960	0.03480	0.045900	0.6890
Dance	8701.0	0.083608	0.077924	0.0228	0.03740	0.05270	0.092200	0.8690
Electronic	9377.0	0.098988	0.097870	0.0228	0.04060	0.05820	0.111000	0.8810
Folk	9299.0	0.045077	0.037122	0.0222	0.03060	0.03540	0.044900	0.9450
Hip-Hop	9295.0	0.205396	0.134865	0.0246	0.08505	0.19100	0.297000	0.9320
Indie	9543.0	0.066724	0.070631	0.0226	0.03270	0.04080	0.064300	0.8690
Jazz	9441.0	0.072304	0.079359	0.0226	0.03520	0.04360	0.066500	0.9390
Opera	8280.0	0.059720	0.055130	0.0276	0.04040	0.04610	0.057800	0.9330
Pop	9386.0	0.107963	0.105474	0.0225	0.03840	0.05890	0.137000	0.8830
R&B	8992.0	0.120994	0.110921	0.0228	0.04460	0.07235	0.162000	0.9490
Rap	9232.0	0.188186	0.133128	0.0229	0.07010	0.16100	0.279000	0.9220
Reggae	8771.0	0.116163	0.102424	0.0231	0.04500	0.07240	0.156000	0.9410
Reggaeton	8927.0	0.127616	0.096444	0.0230	0.05480	0.09030	0.176000	0.8470
Rock	9272.0	0.053664	0.043249	0.0224	0.03180	0.03910	0.056425	0.7470
Ska	8874.0	0.089158	0.070205	0.0225	0.04500	0.06660	0.109000	0.9570
Soul	9089.0	0.082531	0.084651	0.0225	0.03600	0.04880	0.087500	0.9010
Soundtrack	9646.0	0.043852	0.023250	0.0243	0.03610	0.03990	0.045300	0.7920
World	9096.0	0.045766	0.036822	0.0222	0.03130	0.03650	0.045700	0.8090



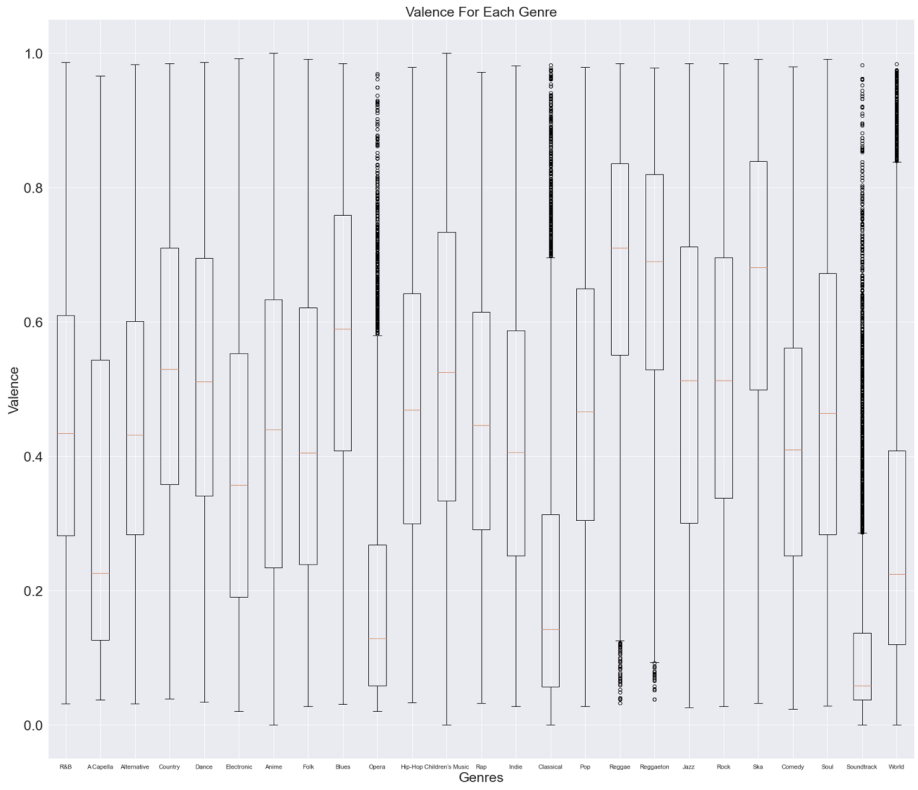
viii. Tempo by Genre

genre	count	mean	std	min	25%	50%	75%	max
A Capella	119.0	111.518950	29.523057	37.861	89.57900	110.1210	129.68300	181.714
Alternative	9263.0	122.534485	30.129437	44.194	97.33150	120.0230	143.96550	213.788
Anime	8936.0	126.629156	33.160207	32.080	98.92125	127.7995	149.99125	220.276
Blues	9023.0	121.137637	30.387693	35.204	97.56650	117.7940	140.63550	242.903
Children's Music	14756.0	121.638247	30.144981	33.792	98.03100	119.9585	140.09250	220.119
Classical	9256.0	104.341807	31.038727	34.208	79.32000	97.0965	127.18125	212.923
Comedy	9681.0	98.235488	27.839986	32.451	79.72000	90.1240	115.12800	207.157
Country	8664.0	123.414419	30.237831	48.718	98.04575	122.9045	144.48075	217.538
Dance	8701.0	120.795919	26.092474	44.573	100.97200	120.0390	134.03000	218.081
Electronic	9377.0	125.845967	26.686039	35.551	107.02000	126.0110	144.99000	220.169
Folk	9299.0	118.748882	28.558870	37.114	96.86300	117.9200	136.30850	236.799
Hip-Hop	9295.0	120.791039	29.755364	39.497	94.29800	121.1270	141.98650	214.126
Indie	9543.0	119.290814	28.598398	38.017	97.01450	117.9670	137.93400	219.331
Jazz	9441.0	111.783658	29.343372	34.765	90.05200	105.4720	127.87100	239.848
Opera	8280.0	101.802977	30.213146	34.151	80.02175	91.0420	120.66900	236.735
Pop	9386.0	121.175844	28.699376	39.497	98.00425	120.0160	140.01575	213.990
R&B	8992.0	116.373834	28.733638	36.710	93.89575	114.8835	134.17175	216.636
Rap	9232.0	121.100808	29.557365	39.497	95.06250	121.3245	142.00050	216.115
Reggae	8771.0	118.162491	31.472660	51.039	90.02100	117.0130	142.31400	218.184
Reggaeton	8927.0	120.987507	33.308933	60.623	93.99350	107.9620	146.03850	234.923
Rock	9272.0	122.629630	29.127513	34.717	99.90450	121.0090	142.00875	219.331
Ska	8874.0	129.427622	32.837509	53.065	101.01725	125.6980	155.57725	221.578
Soul	9089.0	115.322493	28.759171	36.542	93.24300	112.0050	132.44500	216.636
Soundtrack	9646.0	104.083509	32.049869	30.379	78.99125	96.6375	126.14200	216.429
World	9096.0	119.821152	29.894906	33.593	95.52500	122.9955	141.55675	212.923



ix. Valence by Genre

genre	count	mean	std	min	25%	50%	75%	max
A Capella	119.0	0.328724	0.255005	0.0380	0.1265	0.2260	0.54350	0.966
Alternative	9263.0	0.449590	0.216426	0.0321	0.2840	0.4320	0.60100	0.983
Anime	8936.0	0.441682	0.249619	0.0000	0.2340	0.4400	0.63325	1.000
Blues	9023.0	0.579425	0.224677	0.0315	0.4080	0.5900	0.75900	0.985
Children's Music	14756.0	0.532251	0.250929	0.0000	0.3340	0.5250	0.73400	1.000
Classical	9256.0	0.214463	0.200275	0.0000	0.0572	0.1430	0.31325	0.982
Comedy	9681.0	0.412764	0.207258	0.0237	0.2520	0.4100	0.56100	0.980
Country	8664.0	0.535160	0.219819	0.0395	0.3580	0.5300	0.71000	0.985
Dance	8701.0	0.517754	0.226822	0.0340	0.3410	0.5110	0.69500	0.986
Electronic	9377.0	0.388129	0.236938	0.0205	0.1910	0.3570	0.55300	0.992
Folk	9299.0	0.440237	0.241416	0.0277	0.2390	0.4050	0.62100	0.991
Hip-Hop	9295.0	0.473381	0.222325	0.0336	0.3000	0.4690	0.64250	0.979
Indie	9543.0	0.428665	0.221606	0.0277	0.2520	0.4060	0.58750	0.981
Jazz	9441.0	0.508961	0.251218	0.0266	0.3010	0.5130	0.71200	0.985
Opera	8280.0	0.189864	0.172322	0.0207	0.0589	0.1290	0.26800	0.969
Pop	9386.0	0.481371	0.225029	0.0277	0.3050	0.4670	0.64975	0.979
R&B	8992.0	0.450346	0.215387	0.0321	0.2820	0.4340	0.61000	0.986
Rap	9232.0	0.455918	0.213913	0.0331	0.2910	0.4460	0.61500	0.972
Reggae	8771.0	0.679665	0.198141	0.0331	0.5510	0.7100	0.83600	0.985
Reggaeton	8927.0	0.659439	0.202052	0.0381	0.5290	0.6900	0.81950	0.978
Rock	9272.0	0.517113	0.231137	0.0277	0.3380	0.5130	0.69600	0.985
Ska	8874.0	0.653472	0.223245	0.0331	0.4990	0.6810	0.83900	0.991
Soul	9089.0	0.480562	0.245857	0.0287	0.2840	0.4640	0.67200	0.991
Soundtrack	9646.0	0.118483	0.139913	0.0000	0.0374	0.0589	0.13700	0.982
World	9096.0	0.295657	0.230914	0.0000	0.1200	0.2250	0.40800	0.984



Pre-Processing & Training

- For this stage, I converted 3 categorical columns (mode, key, time_signature) to numerical values by assigning integer values to each of the unique categorical values and then merged the numerical conversions into the dataset dataframe as additional columns. In addition, I used one hot encoding to create dummy variables for the categorical columns although it wasn't necessary due to the previous conversion.
- Next, I eliminated any rows that classified the same song into more than one genre since I'm treating this as a multi classification problem and not as a multi-label classification problem. This elimination left only one row for each unique song that classifies it into one genre.
    - To do this, first I changed the artist name and track name columns to uppercase strings so no two same strings with different capitalizations would be treated as different tracks.
    - Then, I used the pandas 'drop_duplicates()' function with 'track_id' to find and drop rows with the same track id.
    - Lastly, I used 'drop_duplicates()' again but with the subset ['artist_name', 'track_name]. This would identify identical songs that have the different 'track_id' values for reasons such as appearing in different albums, among others. These songs could be potentially classified into different genres and thus this final filtering ensures each unique track has one row and genre associated with it.
    - The result of all the filtering created a final data set of 152,864 tracks from the original 232,725 tracks.
- I set the features that would be used to train the model to be the columns: features = ['acousticness', 'danceability', 'energy', 'instrumentalness', 'key_num', 'liveness', 'loudness', 'mode_num', 'speechiness', 'tempo', 'time_signature_num', 'valence'].
- I created the variable X = df[features] the features of each track, and y = df['genre'] the list of genres. I used test_train_split to split the data into 75% training and 25% testing data stratified on the distribution of genres in the prediction class variable y.
- The final thing I did was remove the outliers from the training and test set. To do this, I first found the standard deviation of the training set, and the cutoff (3 * std), and then found the lower and upper bounds of permissible values by taking (mean-cutoff) and (mean+cutoff) respectively. By using these bounds I filtered out rows from the training and testing sets that were not within these bounds.

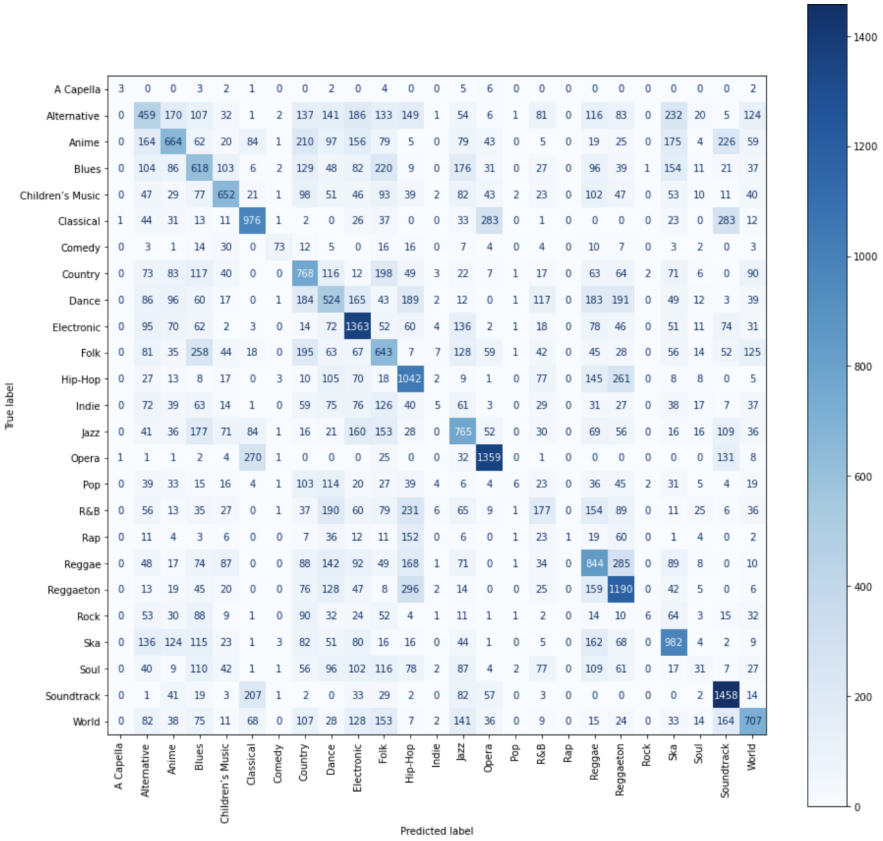
Modeling

- Before training and testing, I used the StandardScalar() method to fit and transform my training set and used the training set fit on the testing set to avoid data leakage.
- I trained and tested on 4 ML models: logistic regression (multinomial), Random Forest Classifier, Gradient Boosting Classifier, and Linear Support Vector Classifier.
- I took the following 5 steps to implement each of the 4 ML models:
    - 1. I used GridSearchCV for hyperparameter tuning to pick the best model version
        - Random Forest Classifier: n_estimators = 500, criterion = gini
    - 2. I fit the model and madee predictions for the test set.

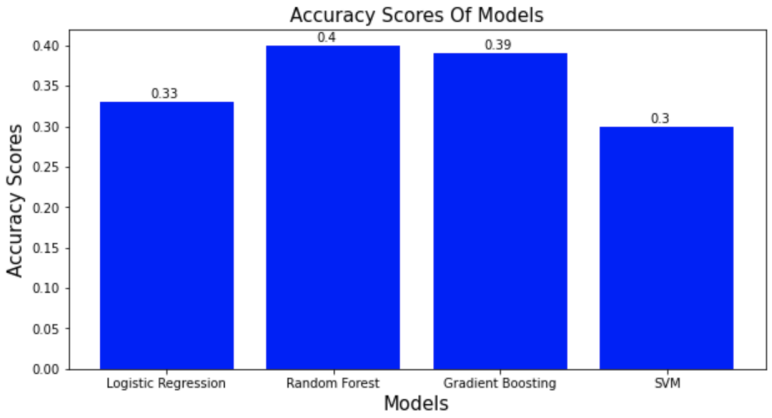
- 3. Printed the classification report displaying the accuracy, precision, recall, and F1-scores for each unique class variable (‘genre’).

	precision	recall	f1-score	support
A Capella	0.60	0.11	0.18	28
Alternative	0.26	0.20	0.23	2240
Anime	0.39	0.31	0.34	2177
Blues	0.28	0.31	0.29	2000
Children’s Music	0.50	0.42	0.45	1569
Classical	0.56	0.55	0.55	1777
Comedy	0.78	0.35	0.48	210
Country	0.31	0.43	0.36	1802
Dance	0.25	0.27	0.25	1974
Electronic	0.45	0.61	0.52	2245
Folk	0.27	0.33	0.30	1968
Hip-Hop	0.40	0.57	0.47	1829
Indie	0.11	0.01	0.01	820
Jazz	0.36	0.39	0.38	1937
Opera	0.68	0.74	0.71	1836
Pop	0.32	0.01	0.02	596
R&B	0.21	0.14	0.16	1308
Rap	1.00	0.00	0.01	359
Reggae	0.34	0.40	0.37	2108
Reggaeton	0.44	0.57	0.50	2095
Rock	0.55	0.01	0.02	543
Ska	0.45	0.51	0.48	1924
Soul	0.13	0.03	0.05	1075
Soundtrack	0.57	0.75	0.64	1954
World	0.47	0.38	0.42	1842
accuracy			0.40	38216
macro avg	0.43	0.33	0.33	38216
weighted avg	0.39	0.40	0.38	38216

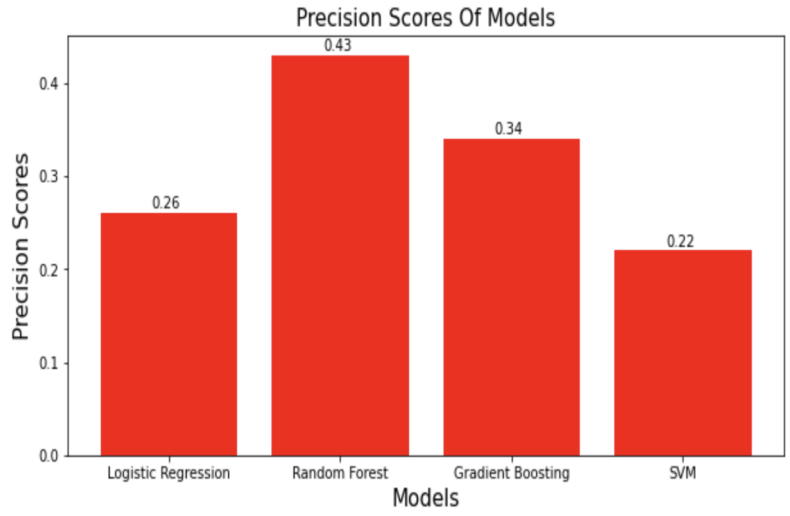
- 4. Calculated the accuracy, precision, recall, F1-score, and log loss scores, and ROC-AUC (area under the curve) for the model.
- 5. Lastly, displayed the confusion matrix to see the distribution of predictions being made across all genres.



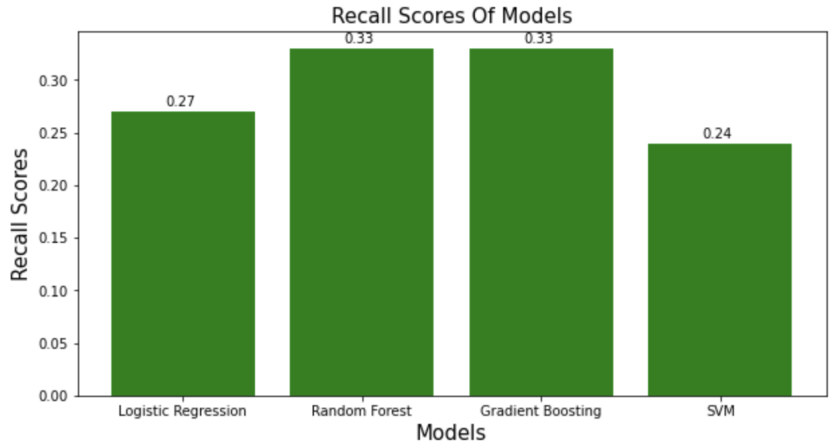
- I plotted bar graphs represent the 5 metrics for comparison of the 4 ML models:
  - <u>Accuracy Scores Of Models</u>



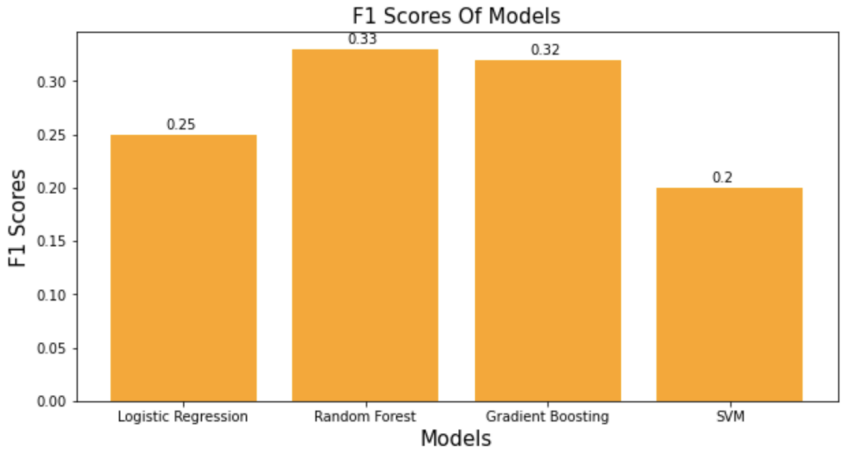
  - <u>Precision Scores Of Models</u>



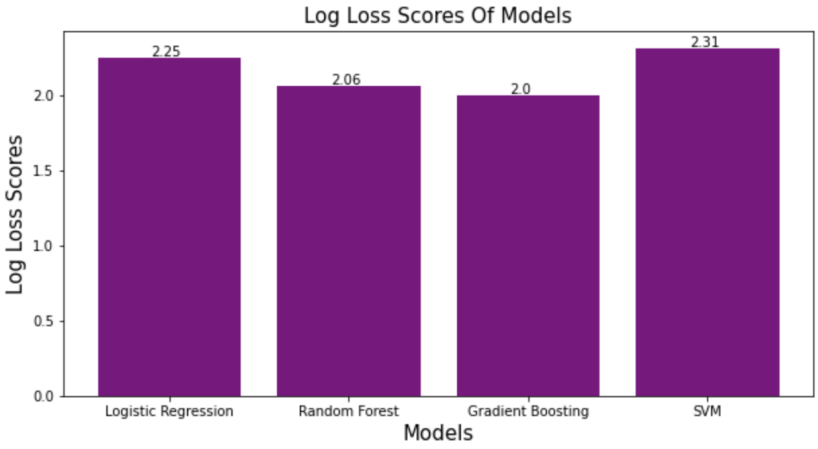
- Recall Scores Of Models



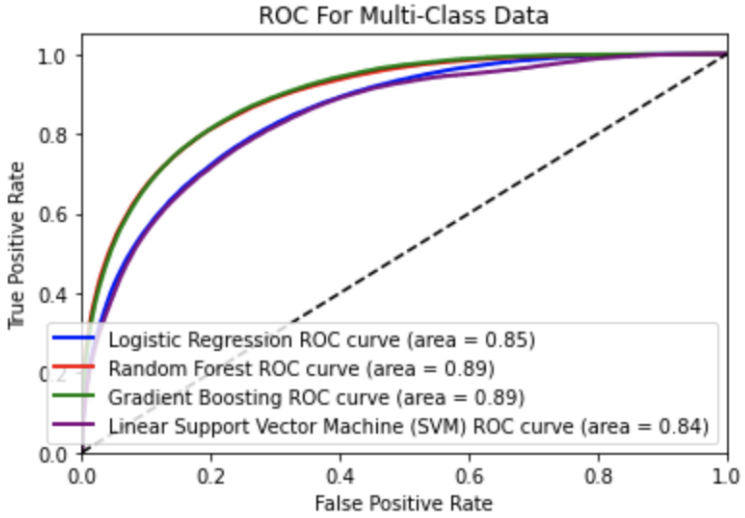
- F1 Scores Of Models



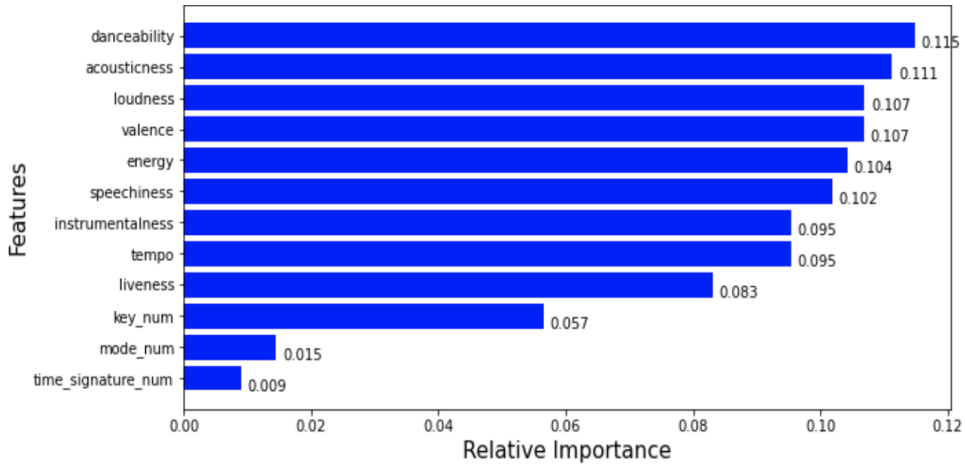
- Log Loss Scores Of Models



- ROC-AUC Of Models



- The best model that resulted in the best fit was the random forest model. It had the second lowest log loss score (2.06), the highest accuracy (40%), precision (43%), recall (33%), f1 (33%), and ROC-AUC (0.89).
- I plotted a horizontal bar graph displaying the feature importances for the model:



- Genre Classification Function:
  - I created a function that classifies song inputs from users into their predicted genre label and the next 2 most likely genres the song can be classified into in case Spotify decides to classify songs into more than one genre based on likelihood.
  - To accomplish this I built methods called 'get_token(),' 'search_for_track(),' 'get_features(),' and 'predict_genres()' to use the API and get the required audio features in cases where tracks inputted by a user don't exist in the dataset. Then, I used the random forest model to predict on the track and get the genre it predicts is the classification. Lastly, I use the 'predict_proba()' method to get the likelihood of classification into the different genres and display the top 2 likelist ones after the actual predicted genre label.
  - Below is an image of an example input and result of the function:

```
In [211]: predict_genres()

Please enter track name to find its genre:
sad
Please enter artist name of song to find its genre:
bo burnham

The genre the song can be classified into is: Blues at 13.80% probaility.
The next 2 likeliest genres the song can be classified into are: Jazz at 10.2
0% probability and Children's Music at 9.00% probability.
```

- Similar Tracks Recommender Function:
  - I created a function that recommends the top K (user inputted integer K) songs similar to the one the user inputs that the user might want to listen to.
  - I ask for the inputs of artist name and track name as well as the k number of songs they want recommended to them.
  - To accomplish this I mimicked the KNN (K-Nearest Neighbor) algorithm by finding the euclidean distance of each of the inputted track's features with every other songs features in the dataset and recommending the top k songs in order of smallest euclidean distance.
  - Below is an image of an example input and result of the function:

```
In [24]: recommend_songs()

Please enter the number of similar songs you would like to get recommended be
tween 1 and 152864:
10
Please enter track name to find its 10 closest similar songs:
bambi
Please enter artist name of song to find its 10 closest similar songs:
baekhyun

Your top 10 song recommendations are:

1. TRIO FOR HORN, VIOLIN AND PIANO IN EB, OP. 40; I by TOM'S MUSIC BOX
2. LACRIMOSA DOMINAE by IMMEDIATE
3. テルーの唄(ゲド戦記より) by YUKA
4. RIDE TO THE NAZI HIDEOUT by JOHN WILLIAMS
5. JESUS ON THE MOUNT by MATT BRAUNGER
6. INTRODUCTION - LIFE OF BRIAN / SOUNDTRACK VERSION by MONTY PYTHON
7. JACQUARD CAUSEWAY by BOARDS OF CANADA
8. EVERYTHING'LL CHANGE by MICHL
9. LEON WITH CLAIRE by CAPCOM SOUND TEAM
10. CAT RESTAURANT by BECKY DONOHUE
```

Takeaways

- The random forest classifier was the best model although the gradient boosting classifier was also very close in performance with a 0.06 lower log loss score but was 7% less in precision than the random forest classifier.
- Based on the displayed feature importances horizontal bar graph, the three categorical variables seem to not be very useful in training the model.
- There is a possibility that the tracks have not been accurately labeled as this dataset was compiled by a kaggle user who didn't explain the method by which they were able to label all the tracks. 'Garbage in, garbage out' could possibly be an accurate saying to represent the reason for the lack of an accurately created model. Further experimentation and testing would be necessary to come to this conclusion of whether the data is intrinsically problematic and needs further feature engineering or the data compilation by the creator of the dataset is the issue.

Further Research

- In the future, I would like to eliminate outliers for each genre grouped by its associated features and see the results of refitting the model with the filtered data. As seen in the EDA step there were many outliers that could potentially have significantly hurt the fit of the random forest classifier.
- Another thing to try would be to try and use different binning of features to see if that could help differentiate the genres more from one another. This binning could be done with grouping the categorical variables such as keys (ex. C and C#, etc.) or even certain ranges of numerical features could be taken together to represent certain integer values. Trying different binning techniques would allow us to see what leads to more distinct stratification of data between genres.
- The third thing to try would be to compile the dataset myself by accessing the API and getting genres based on artist and album and synthesizing them in a way that would lead to more accurate pre-labeled data to use to train the model. I already tried to look at genres based on tracks which don't exist, and I similarly couldn't find genre labels for albums as well. Maybe the genre labels

for albums are rare in the Spotify database but there seems to be some genre labels for artists. Although using artist overall genre labels might seem improper when constructing a dataset, only further testing will determine how accurate it could be.

- The aforementioned techniques could also be used by Spotify when engaging in improving their own ML model.