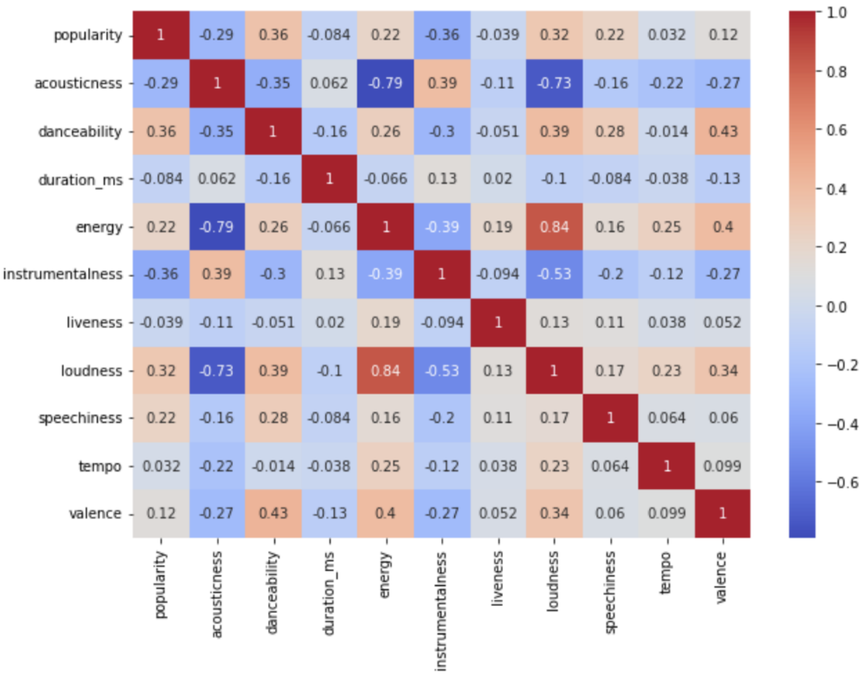
Anya Zorin FML

Capstone Classification Project Report

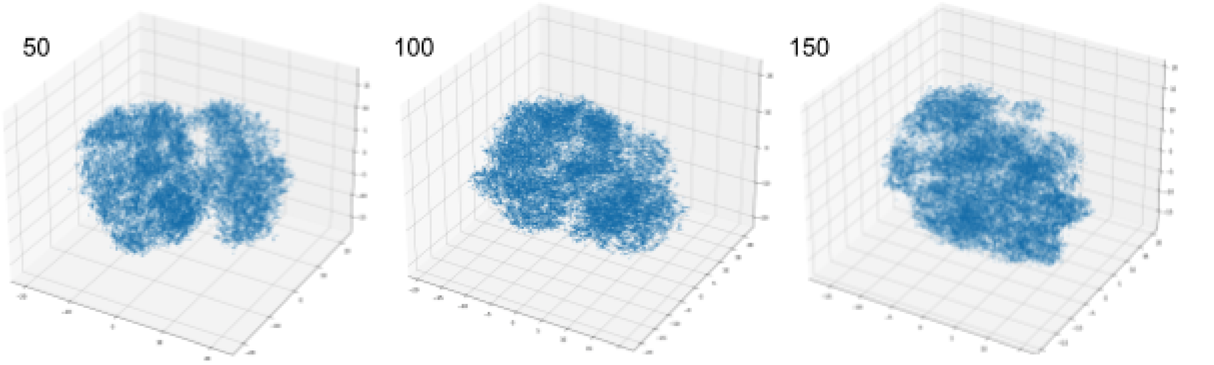
I first loaded the data and checked for where there were null values, and found that 5 rows were na, so I dropped those rows. I then took the linguistic data columns (artist name and song name) as well as the spotify id and the date obtained by spotify since these are irrelevant towards genre. I then checked the datatypes of the columns and found that tempo was a string, for which the cause is missing tempo values, with string '?' in the tempo's place. Since the number of missing entries was roughly 10% of the total data entries, I chose to estimate the tempo values by using a simple multi-linear regression using the data that we did have. After the missing tempo data had been estimated, I converted the whole column to float so that it is usable.



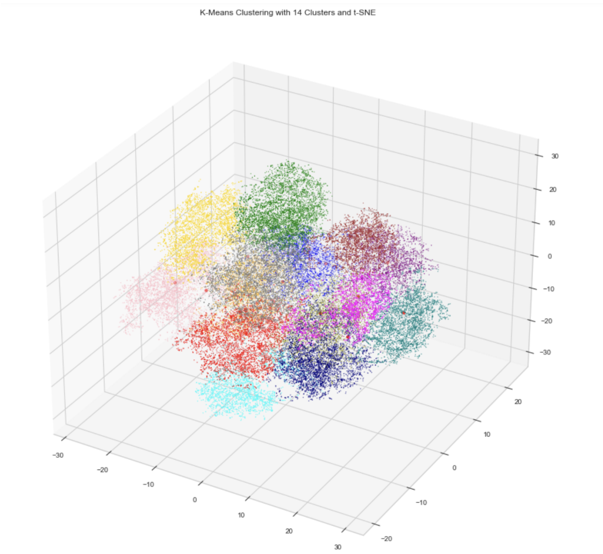
	popularity	acousticness	danceability	duration_ms	energy	instrumentalness	liveness	loudness	speechiness	tempo	valence
popularity	1	-0.29	0.36	-0.084	0.22	-0.36	-0.039	0.32	0.22	0.032	0.12
acousticness	-0.29	1	-0.35	0.062	-0.79	0.39	-0.11	-0.73	-0.16	-0.22	-0.27
danceability	0.36	-0.35	1	-0.16	0.26	-0.3	-0.051	0.39	0.28	-0.014	0.43
duration_ms	-0.084	0.062	-0.16	1	-0.066	0.13	0.02	-0.1	-0.084	-0.038	-0.13
energy	0.22	-0.79	0.26	-0.066	1	-0.39	0.19	0.84	0.16	0.25	0.4
instrumentalness	-0.36	0.39	-0.3	0.13	-0.39	1	-0.094	-0.53	-0.2	-0.12	-0.27
liveness	-0.039	-0.11	-0.051	0.02	0.19	-0.094	1	0.13	0.11	0.038	0.052
loudness	0.32	-0.73	0.39	-0.1	0.84	-0.53	0.13	1	0.17	0.23	0.34
speechiness	0.22	-0.16	0.28	-0.084	0.16	-0.2	0.11	0.17	1	0.064	0.06
tempo	0.032	-0.22	-0.014	-0.038	0.25	-0.12	0.038	0.23	0.064	1	0.099
valence	0.12	-0.27	0.43	-0.13	0.4	-0.27	0.052	0.34	0.06	0.099	1

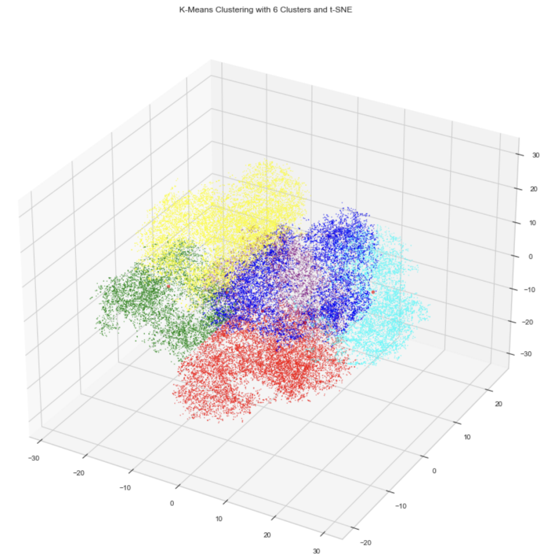
After this I made a correlation matrix so that I could get a better sense of the data before moving forward. [Extra Credit] An interesting observation to be made is that loudness and energy are extremely correlated, while loudness and energy are negatively correlated with acousticness. But outside of this correlation does not really go above 0.4.

Before dimensionality reduction and clustering I z-scored the data. I used t-sne for dimensionality reduction, and tried perplexities 50,100, and 150, and found silhouette scores

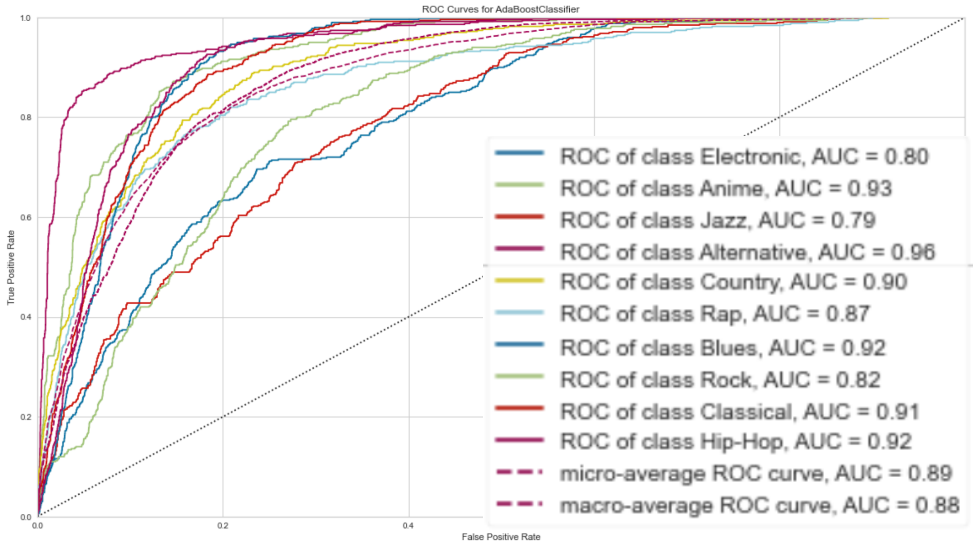


for ks from 2 to 10 on each of them. Based on these results as well as the 3D plots (3 components) for these dimension reductions I chose to do t-SNE with perplexity = 100. With perplexity 50, the clusters were not well defined, and perplexity 150 is generally very high and with not enough improvement from 100, so I chose 100. I used the silhouette score to find a suitable k for k-means clustering for the lower dimension data. I found that the silhouette score peaked in 2 places, so I decided to use both of them since they were close in value, one was k=6 and the other was k= 14.


K-Means Clustering with 14 Clusters and t-SNE


K-Means Clustering with 6 Clusters and t-SNE

After clustering, I added the cluster number as another column in the dataset so that it can be used in the classification model, and I normalized it. This way the grouping found through the dimensionality reduction and clustering can be used. I then split the data into test and train sets as instructed. I then used an AdaBoost Forest model to classify the genres, and got the AUC for each class, as well as the overall (0.89). I used this to graph the ROC for each class and the overall. Yellowbrick makes it very simple to graph by class.


ROC Curves for AdaBoostClassifier