Project 2: Write Up

By Pomaikai Canaday, Milan Champion, Brandon Herren

Summary:

Through the analysis of artist gender and genre, we continued to make insight into possible patterns and connections in popular music over the period of 1963-2018. Using more directed classification method of the initial genre 'tags' we scraped for each song, we were able to assign each tag relative association values with a variety of genres. From there, we used these genre values to cluster tags, analyze the connection between gender and chart rank, as well as that of genre and release year, achieving fairly successful classification results at a level that suggests that each pair of variables is connected in some way. This could reflect upon the continued effect of gender and race in the music industry, as female vs. male had a clear impact on the song's performance over the time period, and genre clustering the continued divide between traditionally White-associated genres like rock, metal, and alternative, versus Black-associated genres such as soul, R&B, and hip hop. We will continue to pursue these findings beyond the hypotheses detailed below.

Basic Statistical Analysis and Data Cleaning Insight:

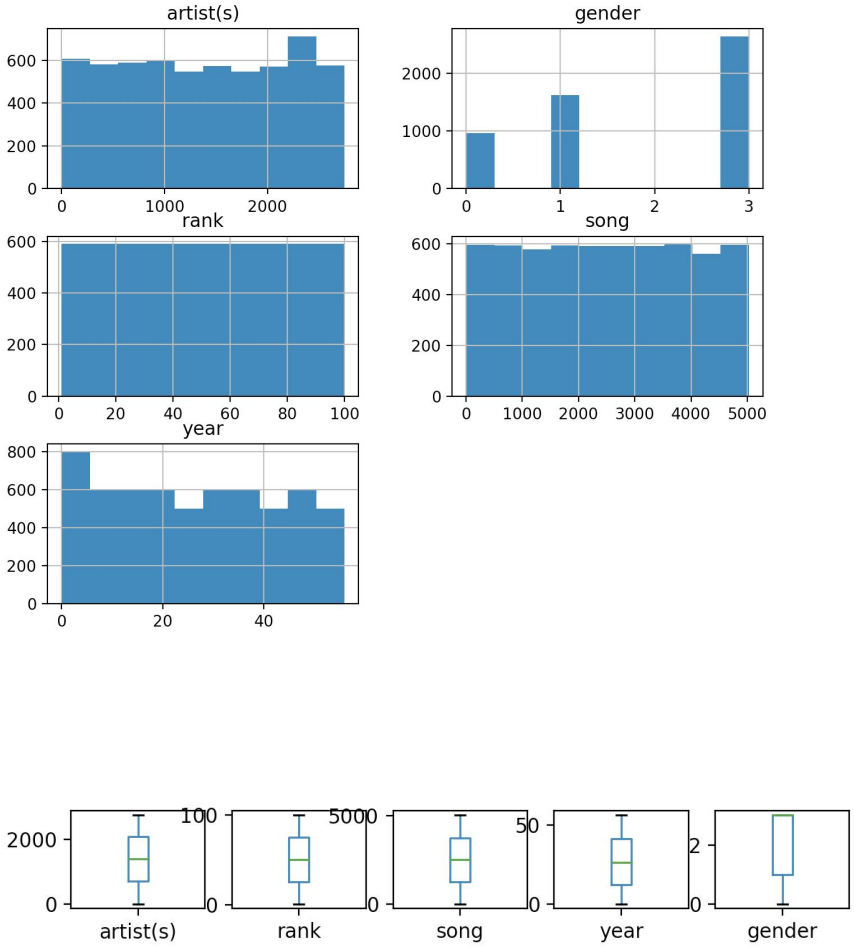
Data Cleaning Insight

While performing analysis, we realized that there was a mistake in the initial chart data - likely due to our method of Wikipedia scraping. For whatever reason, the years 1963, 1966, and 1975 were duplicated in the data set, with the songs from each year appearing twice. Upon this realization, we re-tooled the JSON file to remove these duplicates and continued on with the analysis.

Music Chart Data and Gender Data

While running the Basic Statistical analysis on the following data sets we got the following values and graphs:

artist(s)	rank	song	year	gender
Madonna	1	Angel	1963	unknown (band/multiple artist)
-	10	Heaven	1966	-
-	100	My Love	1975	-
	range(1-100)		(all the years)	



It is interesting to note the following:

- Madonna is the mode for artist despite being a single female artist (which is categorized as 0 numerically) as females rank the lowest in terms of chart distribution, being that bands/multiple had the most charted songs. This could be interesting for later analysis of Madonna's impact on social movements specifically looking at feminist movements in the late 1900's.
- There were no outliers for artist, rank, song, year and gender as shown with the box and whiskers plots where if there was an outlier, it would be shown on these graphs. We suspected this because they are all categorical.
- Seems like most of the charted songs came from bands/multiple artist, which could be interesting to look by year, how many bands/artist charted. This could be interesting in

terms of social movements, for example, we see a lot of racial movements correlated with groups such as N.W.A and the Beatles.

Genre Work (Pre-Analysis Classification):
In Project 1, we had genre tag data for each of the 5600 songs, each of which was in the format of a dictionary with keys being the tag names and the values being the relative weight of how much the song had been tagged with that key.

EX. {rock: 100, pop: 78, dance: 12}

This data had plenty of small pitfalls - as user tagged data meant that plenty of song tags were relatively nonsensical or irrelevant to genre (a shocking number of "disney" tagged songs). In order to categorize the tags into more specific genre subcategories, we created, based on the most common tags in the set, the following 14 groups of broader genre categories:
[rock, pop, indie, alternative, soul, r&b, hip hop, rap, dance, electronic, folk, jazz, country, metal]

From there, we calculated genre "overlap ratios" for each of the individual tags, for instance, for the tag "indie rock" - we looked at how much overlap there was between that tag (in each song it appeared in) and each category of genres (see *findGenreCorrelation*). Based on the overlap for the entire data set, we calculated the overall overlap value for each tag, which is then saved into a file titled: genre_correlation_11-06-2019_170429.csv.

We also performed some basic statistical analysis for each of the tags, finding the mean/standard deviation/spread of the release years for each song that included the given tag. These matched our expectations (who would've guessed that the '80s' tag has its mean release year… in the 80's!).

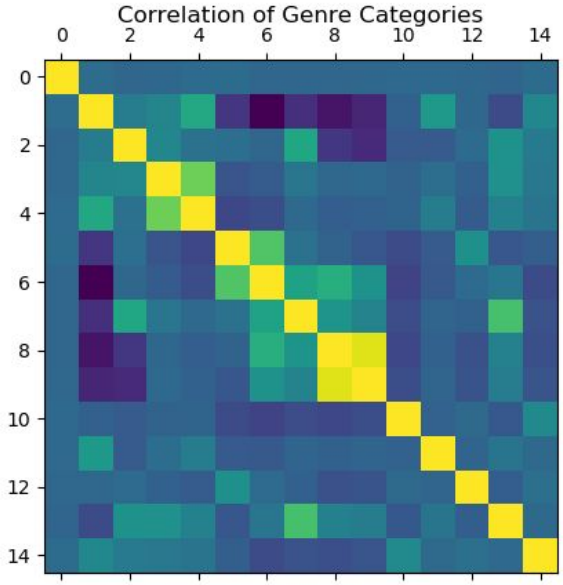
<u>Histograms and Correlations</u>

Using these new overlap values, We calculated the correlations between each of the ratios, which produced values similar to what We expected. Rock correlated with metal/alternative/indie, hip hop correlated with rap/r&b, while rock and hip hop themselves were quite strongly inversely correlated. This makes sense with our broad understanding of musical genres and those two specifically are generally not considered similarly musically. It also speaks to possible social trends, as rock is primarily associated (despite its mixed origins) with white artists, whereas hip hop is associated with African-American artists. In many ways, the separation of genres speaks to the ongoing segregation of music. Here were the overall results:

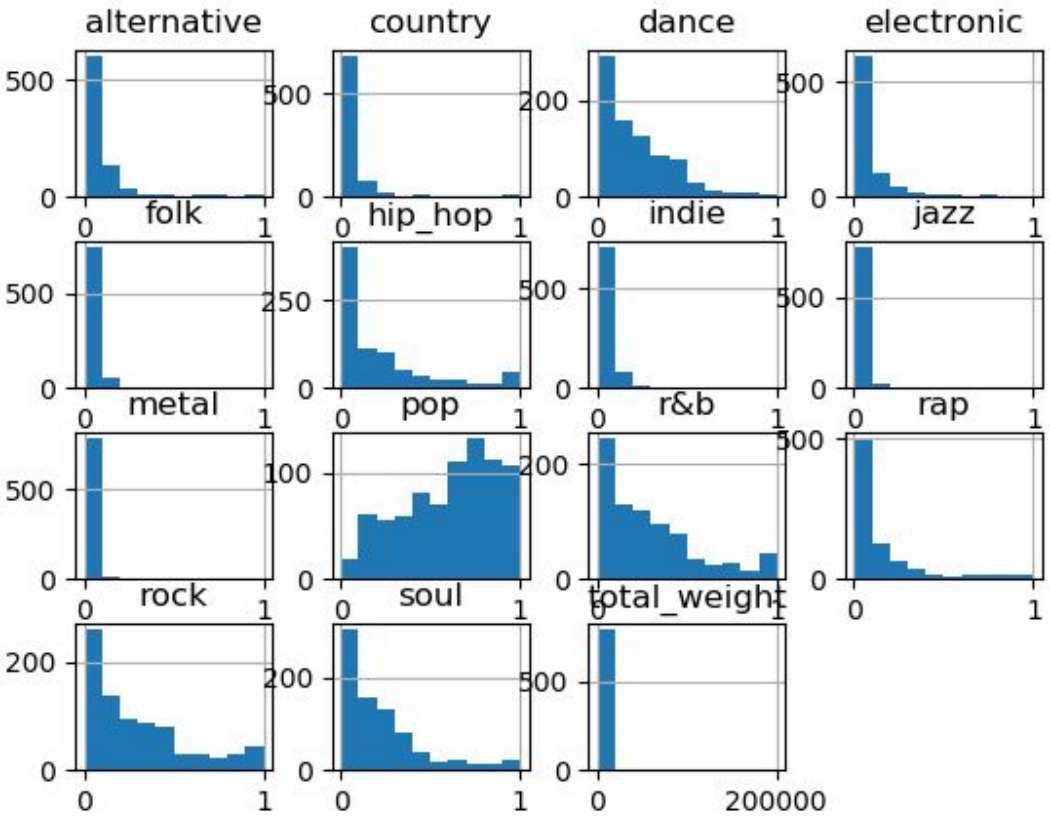
	total_weight	rock	pop	indie	alternative	soul	r&b	dance	hip_hop	rap	country	metal	jazz	electronic	folk
total_weight	1.000000	0.003022	-0.039817	-0.025266	-0.010394	-0.004163	-0.037836	-0.042902	-0.042505	-0.035276	-0.021086	-0.024670	-0.031080	-0.046883	-0.002665
rock	0.003022	1.000000	0.107546	0.152716	0.381985	-0.306289	-0.552727	-0.342623	-0.464203	-0.386728	-0.068648	0.285452	-0.036401	-0.200591	0.170690
pop	-0.039817	0.107546	1.000000	0.167297	0.029895	0.011987	-0.042401	0.374469	-0.301471	-0.362536	-0.109470	-0.109315	-0.007681	0.235914	0.081109
indie	-0.025266	0.152716	0.167297	1.000000	0.654875	-0.144323	-0.099350	0.059563	-0.032365	-0.022595	-0.055586	0.009035	-0.072568	0.230063	0.069699
alternative	-0.010394	0.381985	0.029895	0.654875	1.000000	-0.214682	-0.178726	-0.020713	-0.087720	-0.072288	-0.055054	0.105674	-0.102349	0.128466	0.046293
soul	-0.004163	-0.306289	0.011987	-0.144323	-0.214682	1.000000	0.577460	0.031994	-0.059108	-0.132752	-0.197487	-0.104956	0.226923	-0.131890	-0.084587
r&b	-0.037836	-0.552727	-0.042401	-0.099350	-0.178726	0.577460	1.000000	0.342339	0.417767	0.245138	-0.232184	-0.121290	-0.012409	0.055265	-0.190695
dance	-0.042902	-0.342623	0.374469	0.059563	-0.020713	0.031994	0.342339	1.000000	0.249933	0.140344	-0.185522	-0.043439	-0.067665	0.548078	-0.154982
hip_hop	-0.042505	-0.464203	-0.301471	-0.032365	-0.087720	-0.059108	0.417767	0.249933	1.000000	0.924929	-0.213916	-0.067985	-0.166903	0.131397	-0.173943
rap	-0.035276	-0.386728	-0.362536	-0.022595	-0.072288	-0.132752	0.245138	0.140344	0.924929	1.000000	-0.177950	-0.043430	-0.149759	0.100138	-0.143482
country	-0.021086	-0.068648	-0.109470	-0.055586	-0.055054	-0.197487	-0.232184	-0.185522	-0.213916	-0.177950	1.000000	-0.064448	-0.014972	-0.124101	0.188692
metal	-0.024670	0.285452	-0.109315	0.009035	0.105674	-0.104956	-0.121290	-0.043439	-0.067985	-0.043430	-0.064448	1.000000	-0.046873	0.047869	-0.014749
jazz	-0.031080	-0.036401	-0.007681	-0.072568	-0.102349	0.226923	-0.012409	-0.067665	-0.166903	-0.149759	-0.014972	-0.046873	1.000000	-0.088869	0.021537
electronic	-0.046883	-0.200591	0.235914	0.230063	0.128466	-0.131890	0.055265	0.548078	0.131397	0.100138	-0.124101	0.047869	-0.088869	1.000000	-0.020542
folk	-0.002665	0.170690	0.081109	0.069699	0.046293	-0.084587	-0.190695	-0.154982	-0.173943	-0.143482	0.188692	-0.014749	0.021537	-0.020542	1.000000

(clearer output is displayed in genreClustering.py)

Also in heatmap form:



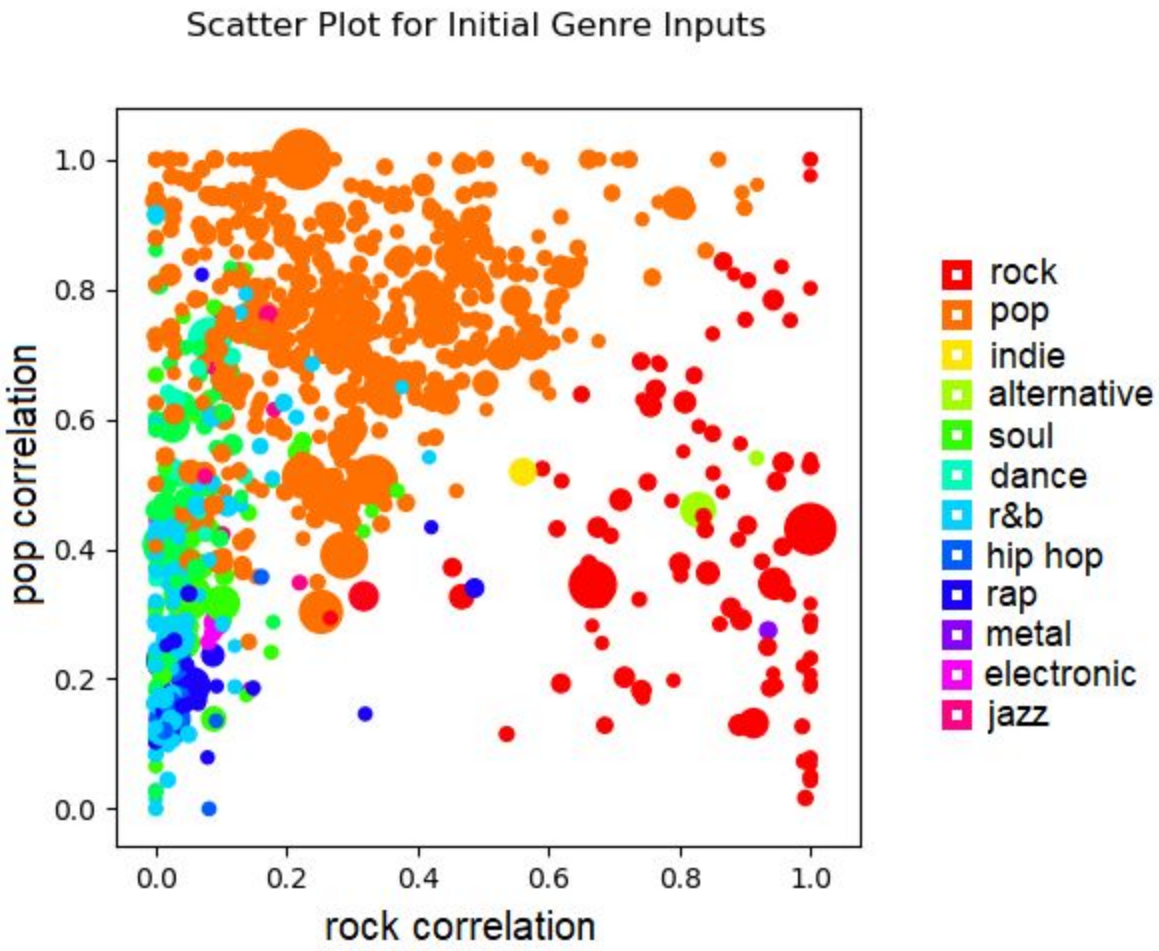
We also produced histograms for the distribution of each of the ratios, seen below:



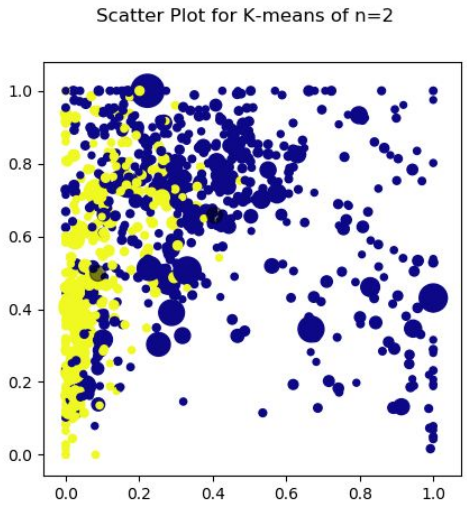
The histograms provide some insight into the distribution of each of the genre ratios. Pop occurs by far the most common, with the highest proportion above 0.5. Other popular genres, such as rock, R&B, and soul, all have smaller niches which are highly matched to them. Conversely, genres like jazz, electronic, and metal, which are even smaller musical niches, have a large proportion at less than 0.1 matching.

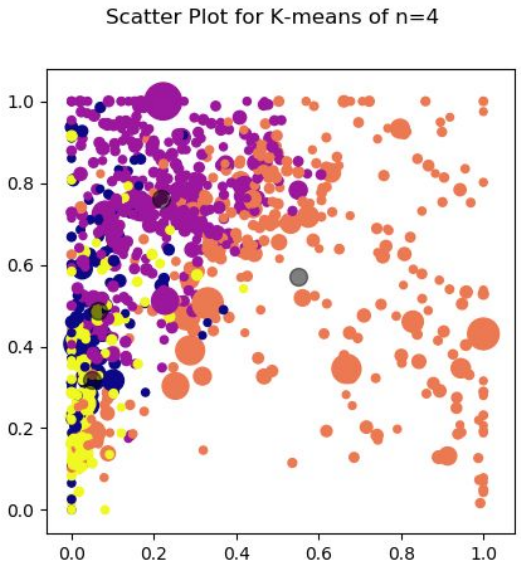
Cluster Analysis:

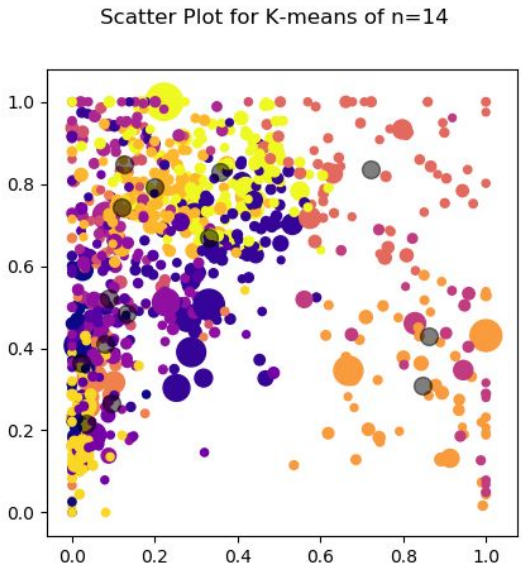
In order to further test the performance of the genre ratio values, we attempted to cluster the various tags based on the given genre ratio values. Naturally, there were some issues with this, as clustering 14-dimensional values did not lend itself to very high silhouette values. Prior to clustering, we attempted to cluster based on simply taking the single highest genre association for each tag in placing it in that genre's cluster (making it 14 clusters total). The following scatter plot was produced from those results.



While attempting k-means clustering for these tags, we were able to get a maximum silhouette score of 0.27 for two clusters. Here are a few examples of the results. The plots show that k-means appears fairly ineffective at categorizing the tags based on their associated genres.

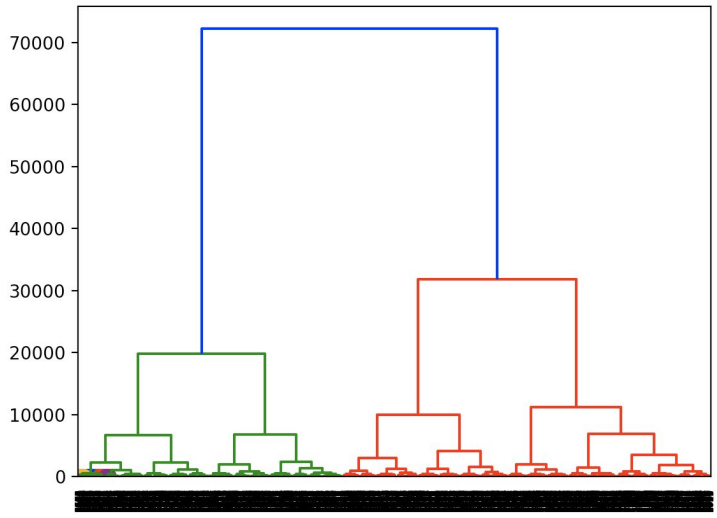






For DBSCAN, the results were consistently 1 cluster with some number of noise points, with there being 6 noise points for an EPS value of 0.65 which produced the highest silhouette value of 0.39.

We were also interested in how artists and rank cluster together hierarchically. So we ran a Wards clustering algorithm on those two attributes and got the following results:



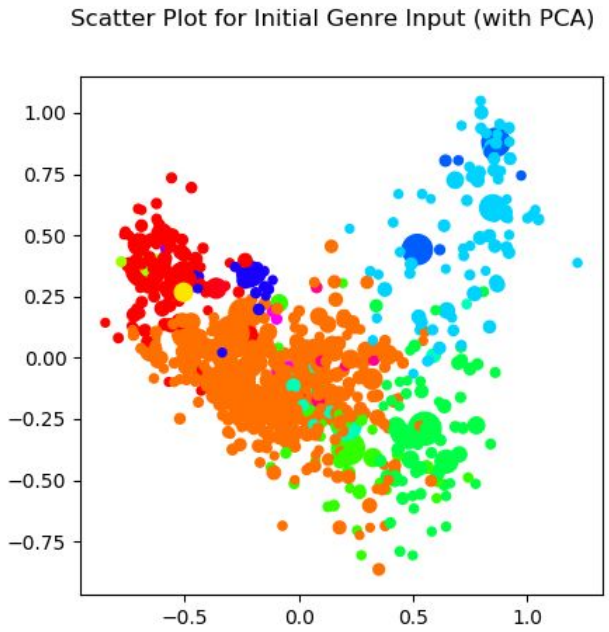
Silhouette Score for cluster: 0.5161460924088213

Based on the Silhouette Score of 0.52 approximately, we can say there is a reasonable structure found. This gives us insight into the popularity of an artist in terms of rank, which we could link to their impact on society. Broadly, we can look at artists who were tagged in Hip-Hop and the rank during a specific year and see what movements were popular at the time.

PCA Analysis

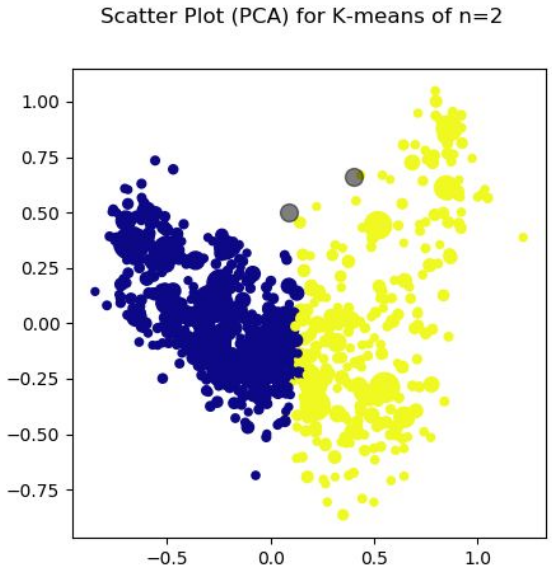
Given that the 2D visualizations weren't exactly the most enlightening for the 14 dimensions of data, we also performed a PCA transformation to better visualize the data. Here are the results:

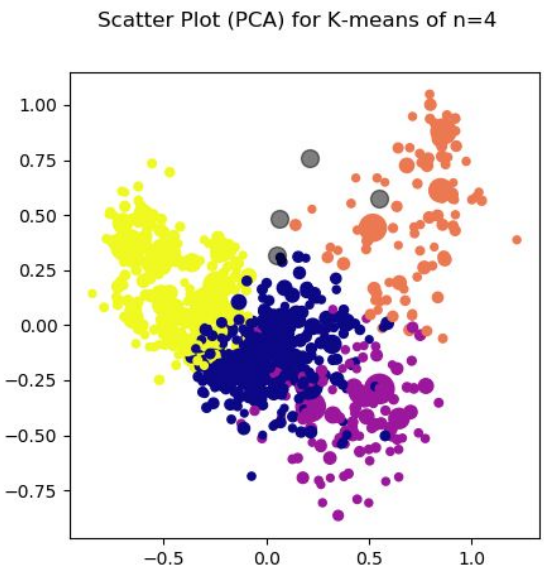
For the initial genre classifications:

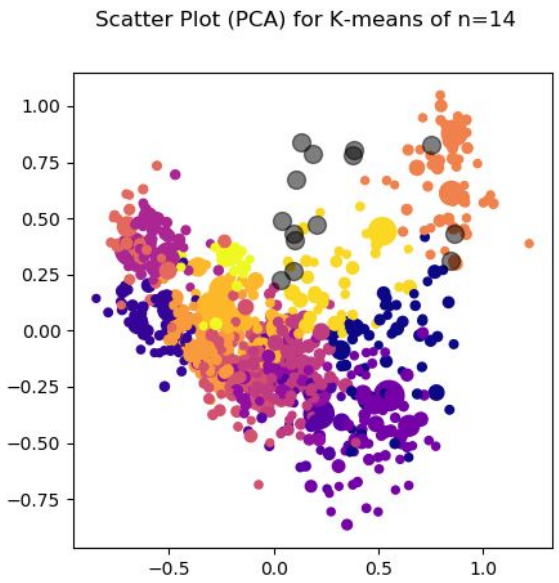


This provides a much better overview of the genres and their differences, with pop (in orange), unsurprisingly occupying the middle space of the genres, while rock, soul, and r&b all push off in separate directions clustered densely together.

Here were the PCA visualizations for the k-means clusters:







Using PCA analysis, we can see that despite n=2's higher silhouette value, n=4 seems to perform better at classifying tags by genre, especially when compared to the initial classification clusters above. N=14 appears extremely ineffective, in spite of that being the number of "genre categories" that exist in the data set, suggesting that a k-means analysis wasn't able to best parse this situation as not all genres are of equal size and density.

Association Rules / Frequent Itemset Mining Analysis

For association rules, we used the initial genre arrays of string tags for each song to determine their association.

Our results were,
13 rules at S=0.2, C=0.4
2 rules at S=0.3, C=0.5
1 rule at S=0.6, C=0.6

The full output can be found in genreClustering.py. The rules were overall somewhat weak, with the strongest ones simply being the occurrence of pop and rock as genre tags, as they occurred commonly. Another one was the association between '80s' and 'pop', which was also fairly unsurprising. Nevertheless, the tags were not very good predictors of each other, with only one rule meeting the relatively low threshold of 0.6 support and 0.6 confidence.

Predictive Analysis

Hypothesis 1: Genre data can be used to predict the relative release year of each song.

In order to pursue this hypothesis, we used Naive Bayes and Random Forest classifiers to predict if each song was released before or after 1990. The following data was used for prediction: title, artist, and each of the genre 'weight' values, calculated based on our genre overlap ratios for each tag multiplied by the weight of that tag in the initial genre dictionary for that song (see: *getEstimatedGenreOverlapBySong*). We first created a class column based on the actual true/false value of the song occurring before or after 1990. From there, we dropped the year column, and performed a train-test split of 80-20. The following results were produced:

For **NAIVE BAYES (Gaussian)**:

GaussianNB: 0.872111 (0.015731)

0.8746594005449592
[[459 119]
 [19 504]]

	precision	recall	f1-score	support
0	0.96	0.79	0.87	578
1	0.81	0.96	0.88	523
accuracy			0.87	1101
macro avg	0.88	0.88	0.87	1101
weighted avg	0.89	0.87	0.87	1101

For **RANDOM FOREST**:

Random Forest: 0.954571 (0.009340)

0.9600363306085377
[[555 23]
 [21 502]]

	precision	recall	f1-score	support
0	0.96	0.96	0.96	578
1	0.96	0.96	0.96	523
accuracy			0.96	1101

macro avg	0.96	0.96	0.96	1101
weighted avg	0.96	0.96	0.96	1101

As such, both performed quite well, with Random Forest being 96% accurate, and Naive Bayes being 87% accurate, which for the most part confirmed our hypothesis. The fact that Random Forest was more accurate than Naive Bayes is not surprising, given the large number of attributes used in our analysis which likely contributed to more complicated trees.

Hypothesis 2: Can we predict whether a song was in the top 20 of the year (rank) based on the artist, year, and song?

In order to pursue this hypothesis, we used all classifiers listed on the project description: Decision Tree (CART), KNeighbors (KNN), Naive Bayes(GNB), and Random Forest(RFC) classifiers to predict if a select song was in the top 20 of the year it was released based on the following data: artist, year, and song title. We first created a class column based on the actual true/false value of whether the song was in the top twenty of the year. From there, we dropped the rank column, and performed a train-test split of 80-20. The following results were produced:

Validation: KNN

0.7821428571428571

[[863 42]
[202 13]]

	precision	recall	f1-score	support
0	0.81	0.95	0.88	905
1	0.24	0.06	0.10	215
accuracy			0.78	1120
macro avg	0.52	0.51	0.49	1120
weighted avg	0.70	0.78	0.73	1120

Validation: CART

0.69375

[[722 183]
[160 55]]

	precision	recall	f1-score	support
0	0.82	0.80	0.81	905
1	0.23	0.26	0.24	215

accuracy			0.69	1120
macro avg	0.52	0.53	0.53	1120
weighted avg	0.71	0.69	0.70	1120

Validation: GNB

0.69375

[[722 183]
 [160 55]]

	precision	recall	f1-score	support
0	0.82	0.80	0.81	905
1	0.23	0.26	0.24	215
accuracy			0.69	1120
macro avg	0.52	0.53	0.53	1120
weighted avg	0.71	0.69	0.70	1120

Validation: RFC

0.69375

[[722 183]
 [160 55]]

	precision	recall	f1-score	support
0	0.82	0.80	0.81	905
1	0.23	0.26	0.24	215
accuracy			0.69	1120
macro avg	0.52	0.53	0.53	1120
weighted avg	0.71	0.69	0.70	1120

As such, all performed okay, with KNN being 78% accurate, GNB being 69% accurate, CART being 69% accurate, and RFC being 69% accurate. With these numbers, we do not see that we can positively support the hypothesis. We were not surprised that the averages were what they were, since we used three attributes to predict which is a relatively small set needed for a good prediction.

Hypothesis 3: Based on the artist, song, rank, and year can we predict whether the artist was a band/multiple artist?

In order to pursue this hypothesis, we used all classifiers listed on the project description: Decision Tree (CART), KNeighbors (KNN), Naive Bayes(GNB), and Random Forest(RFC) classifiers to predict if we could predict whether the song was sung by a single artist or a band/multiple artist based off the artist(this gives no indication of gender), song, rank, and year. We first created a class column based on the actual true/false value of whether the artist was described as band/multiple or a single-gender (male or female). From there, we dropped the gender column, and performed a train-test split of 80-20. The following results were produced:

Validation: KNN

0.6240474174428451
[[261 255]
 [189 476]]

	precision	recall	f1-score	support
0	0.58	0.51	0.54	516
1	0.65	0.72	0.68	665
accuracy			0.62	1181
macro avg	0.62	0.61	0.61	1181
weighted avg	0.62	0.62	0.62	1181

Validation: CART

0.79424216765453
[[415 101]
 [142 523]]

	precision	recall	f1-score	support
0	0.75	0.80	0.77	516
1	0.84	0.79	0.81	665
accuracy			0.79	1181
macro avg	0.79	0.80	0.79	1181
weighted avg	0.80	0.79	0.79	1181

Validation: GNB

0.6054191363251482
[[258 258]

[208 457]]

	precision	recall	f1-score	support
0	0.55	0.50	0.53	516
1	0.64	0.69	0.66	665
accuracy			0.61	1181
macro avg	0.60	0.59	0.59	1181
weighted avg	0.60	0.61	0.60	1181

Validation: RFC

0.7256562235393734

[[359 157]
[167 498]]

	precision	recall	f1-score	support
0	0.68	0.70	0.69	516
1	0.76	0.75	0.75	665
accuracy			0.73	1181
macro avg	0.72	0.72	0.72	1181
weighted avg	0.73	0.73	0.73	1181

As such, all performed okay, with KNN being 62% accurate, GNB being 61% accurate, CART being 79% accurate, and RFC being 69% accurate. With these numbers, we do not see that we cannot positively support the hypothesis. We were not surprised that the averages were what they were, since we used four attributes to predict which is a relatively small set needed for a good prediction.

Hypothesis 4: R&B and Rock generally have the same tags and so it can be assumed that a song under the genre of R&B will also be under the genre of Rock.

In order to test this hypothesis, we used a t-test on our genre correlation data. Unfortunately, it was difficult to find data that fell under a normal distribution, so we used rock and r&b, which we believed would be of some statistical significance. Using the column for r&b exact, as well as the column for rock_exact, we ran a t-test. The findings of the t-test did not support our hypothesis, meaning that these differences between the 2 datasets did not happen by chance. Any similarities between the 2 groups were insignificant.

Ttest_indResult(statistic=-0.3077861198415526, pvalue=0.7582847964168427)

We also used a linear regression to see the correlation between r&b and rock. While it doesn't appear absolutely random, there isn't a strong correlation between the data points. The regressor intercept was about 0.45 and the regressor coefficient was -0.57.

```
Regressor intercept: [0.45727782]
Regressor Coefficient: [[-0.57203369]]
```