Spotify Project: Application Development

Elizabeth Kerrigan
Joel Rodriguez
Wills Mckenna

March 2021

Overview

For this project we worked with the Spotify data set-

(https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks).

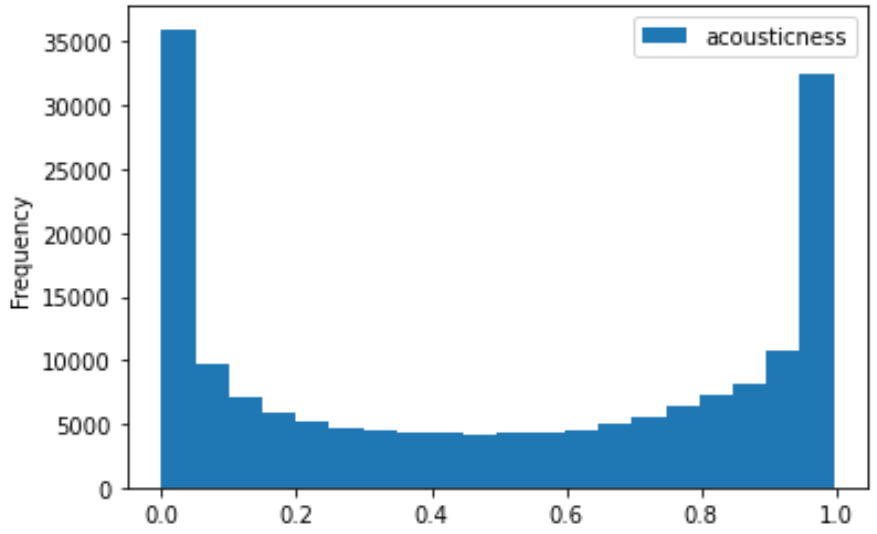
This dataset is over 170,000 songs, with features that capture many aspects of music. We used the dataset to implement several machine learning models with the ultimate goal of creating a "playlist generator"- where based upon user input, the user could be provided with a new collection of 20 songs. Steps in the project included data analysis and exploration, preprocessing of the data, principal component analysis, cluster analysis, and then finally K-nearest-neighbors classification and an app.py that implements the playlist generator task.

01 - Data Exploration

Acousticness:

A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic. Does not follow a normal distribution. Most of the acousticness appears to occur in the .00 and .95 range.¶

```
Datatype:  float64
Mean:      0.499
Median:    0.517
Mode:      0.995
Std:       0.380
Min:       0.000
25%:       0.088
50%:       0.517
75%:       0.895
Max:       0.996
Skew:     -0.035
```



Top 10 Songs by Acousticness

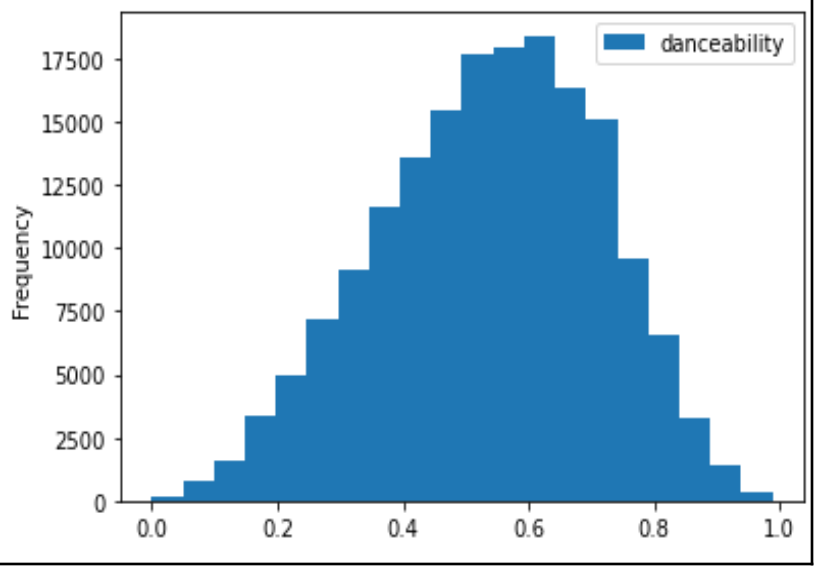
	artists	name	acousticness
5	['Mamie Smith & Her Jazz Hounds']	Crazy Blues - 78rpm Version	0.996
7	['Mamie Smith & Her Jazz Hounds']	Arkansas Blues	0.996
8	['Francisco Canaro']	La Chacarera - Remasterizado	0.996
11	['Francisco Canaro']	Desengaño - Remasterizado	0.996
17	['Francisco Canaro']	El Africano - Remasterizado	0.996
32	['Maurice Chevalier']	Oh Maurice	0.996
56	['Mamie Smith & Her Jazz Hounds']	Frankie Blues	0.996
59	['Mamie Smith & Her Jazz Hounds']	The Darktown Flappers' Ball	0.996
63	['Francisco Canaro']	El Baccarat - Remasterizado	0.996
80	['Mamie Smith & Her Jazz Hounds']	Mean Daddy Blues	0.996

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.

It's not quite a normal distribution, it appears to skew to the left.¶

```
Datatype:  float64
Mean:      0.537
Median:    0.548
Mode:      0.565
Std:       0.176
Min:       0.000
25%:       0.414
50%:       0.548
75%:       0.669
Max:       0.988
Skew:     -0.236
```



Top 10 Songs by Danceability

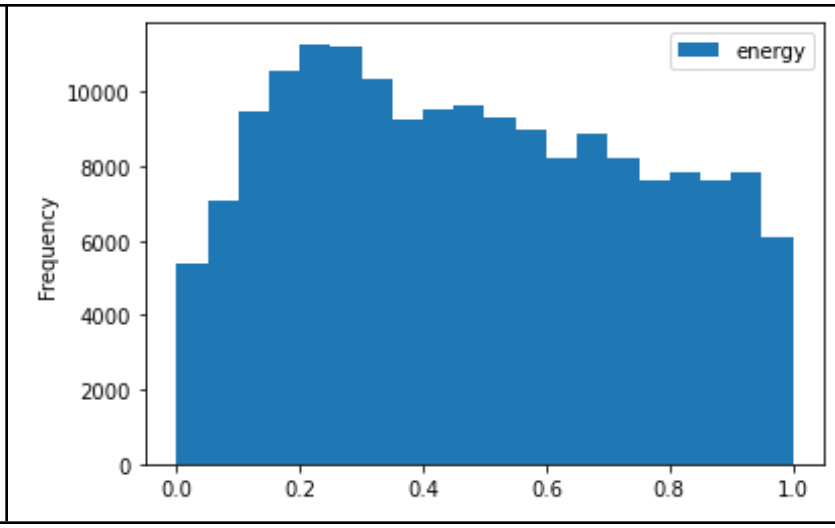
	artists	name	danceability
13734	['Tone-Loc']	Funky Cold Medina	0.988
54525	['Spooner Street', 'Rio Dela Duna', 'Leonardo ...	Cool - Leonardo La Mark Remix	0.987
141441	['Pitbull', 'Trina', 'Young Bo']	Go Girl	0.986
37455	['Tone-Loc']	Funky Cold Medina - Re-Recorded	0.985
92739	['Nilla Pizzi']	O mama mama - Remix 2014	0.985
171536	['Dan McKie', 'Zigmund Slezak']	Dddance - Zigmund Slezak Remix	0.985
173266	['Michael Beyer']	Stuck in Your Brain	0.982
13910	['Vanilla Ice']	Ice Ice Baby	0.980
39276	['347aidan']	Dancing in My Room	0.980
51218	['The Jacksons', 'Mick Jagger']	State of Shock	0.980

Energy:

Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy.

Does not follow a normal distribution. The distribution skews to the right.

```
Datatype: float64
Mean:     0.483
Median:   0.465
Mode:     0.195
Std:      0.273
Min:      0.000
25%:      0.249
50%:      0.465
75%:      0.711
Max:      1.000
Skew:     0.144
```



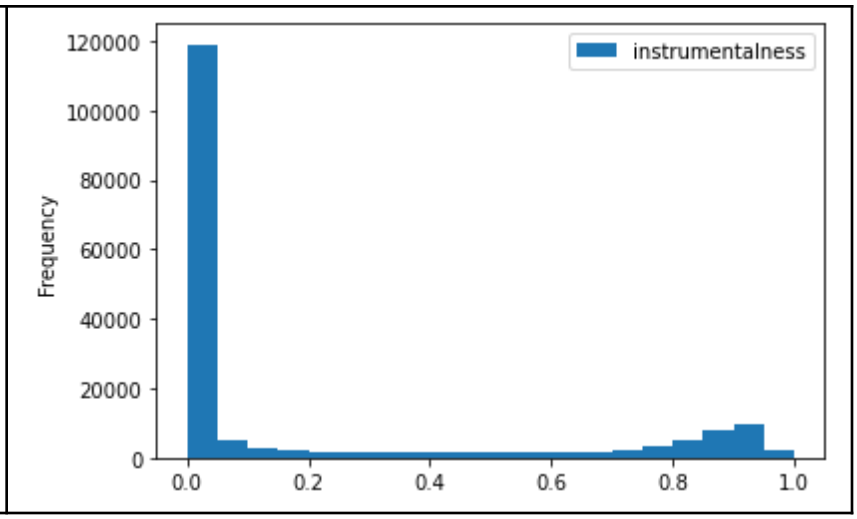
Top 10 Songs by Energy

	artists	name	energy
3543	['Benny Goodman']	Applause; Benny Goodman's 'No Encore' Announce...	1.0
3563	['Benny Goodman']	Applause; Transition Back to Goodman Orchestra...	1.0
23044	['Benny Goodman']	Applause as Lionel Hampton Enters - Live	1.0
39601	['Maurice Chevalier']	Moi J'fais Mes Coups En Dessous	1.0
41853	['Benny Goodman']	Applause; Martha Tilton Returns to Stage - Live	1.0
56761	['Komprex']	Victim	1.0
56805	['Running Man', 'SoundLift']	Amnesia (Mix Cut) - SoundLift's Emotional Take	1.0
56927	['Adam Ellis']	Napalm Poet (Mix Cut) - Original Mix	1.0
56957	['ReOrder', 'STANDERWICK', 'Sky Patrol']	Folding Your Universe (Mix Cut) - Original Mix	1.0
56961	['Sneijder', 'Bryan Kearney']	Proper Order (Mix Cut) - Original Mix	1.0

Instrumentalness:

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 and is purely instrumental, so for example a classical symphony (with no choral part) would be close to 1.0. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0. Does not follow a normal distribution. Appears to skew to the right. Most of the instrumentalness occurs at the .00 range.

```
Datatype:  float64
Mean:      0.197
Median:    0.001
Mode:      0.0
Std:       0.335
Min:       0.000
25%:       0.000
50%:       0.001
75%:       0.252
Max:       1.000
Skew:      1.364
```



Top 10 Songs by Instrumentalness

	artists	name	instrumentalness
19302	['Erik Eriksson', 'White Noise Baby Sleep', 'W...	Clean White Noise - Loopable with no fade	1.000
38892	['High Altitude Samples']	Soft Brown Noise	1.000
76018	['High Altitude Samples']	Cabin Back Noise	1.000
93318	['Erik Eriksson', 'White Noise for Babies', 'W...	Pure Brown Noise - Loopable with no fade	1.000
125791	['Erik Eriksson', 'Lullabies for Deep Meditati...	White Noise - Loopable With No Fade	1.000
142261	['Zen Sounds']	White Noise: Mindfulness Meditations (Loopable)	1.000
157709	['The White Noise Zen & Meditation Sound Lab']	Calm Rain Storm & Gentle White Noise	1.000
61338	['K Dutta']	Suno Suno Tumhen Sunaye	0.999
70239	['Nataural']	Under Shelter Rain	0.999
92672	['Sleep Baby Sleep', 'Meditation Spa', 'White ...	Pouring Rain - Loopable with No Fade	0.999

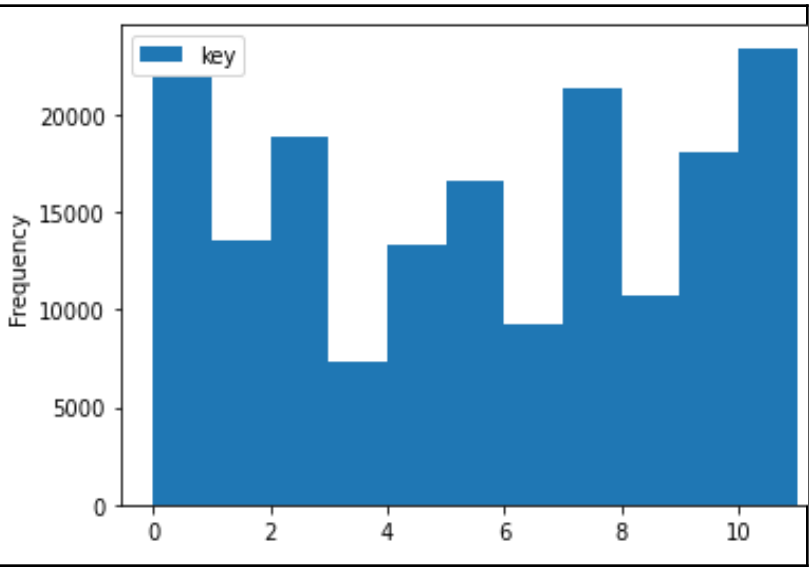
Key:

The key the track is in. Integers map to the pitches using standard Pitch Class notation.

E.g. 0 = C, 1 = C♯, D♭, 2 = D, 3 = D♯, E♭, 4 = E, 5 = F, 6 = F♯, G♭, 7 = G, 8 = G♯, A♭, 9 = A, 10 = A♯, B♭, 11 = B

https://en.wikipedia.org/wiki/Pitch_class

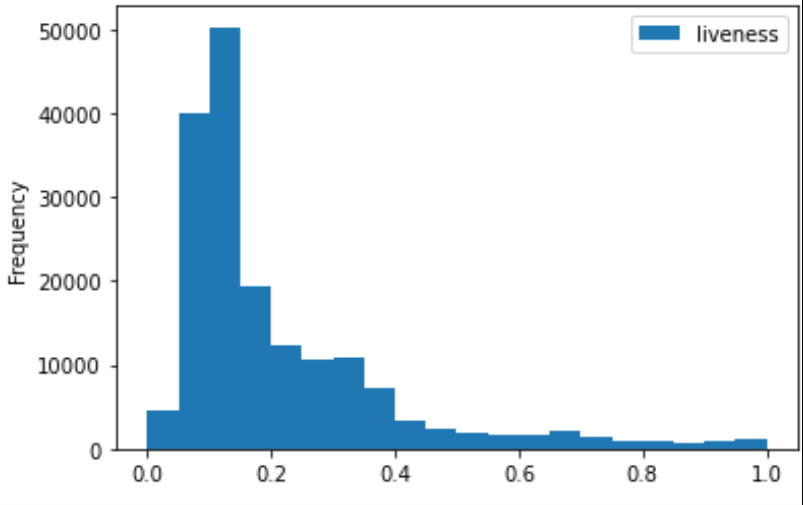
Does not follow a normal distribution. It appears to be sparsely populated, which makes sense as there is no one dominant key in all of western music.

Datatype: int64 Mean: 5.205 Median: 5.000 Mode: 0 Std: 3.518 Min: 0.000 25%: 2.000 50%: 5.000 75%: 8.000 Max: 11.000 Skew: 0.004	

Liveness:

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.

Does not follow a not follow a normal distribution. Appears to skew to the right. Most of the liveness occurs at the .1 range.

Datatype: float64 Mean: 0.211 Median: 0.138 Mode: 0.111 Std: 0.180 Min: 0.000 25%: 0.099 50%: 0.138 75%: 0.270 Max: 1.000 Skew: 2.078	

Top 10 Songs by Liveness

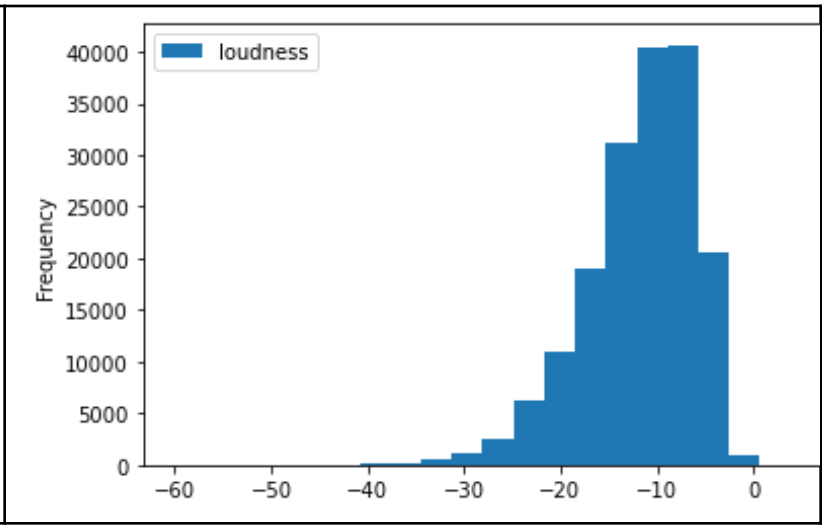
	artists	name	liveness
170448	['The Band']	Up on Cripple Creek - Concert Version	1.000
63836	['Duke Ellington']	Skin Deep - Live	0.999
105900	['Fleetwood Mac']	The Chain - Live at Warner Brothers Studios in...	0.998
31937	['Cheryl Lynn']	Encore	0.997
68656	['Eagles']	Life in the Fast Lane - Live; 1999 Remaster	0.997
85274	['Bob Marley & The Wailers']	Trenchtown Rock - Live At The Roxy Theatre	0.997
154512	['Banda Maguey']	Pero Te Amo - Live Version	0.997
15981	['Billy Joel']	Auld Lang Syne - Live at Madison Square Garden...	0.996
35027	['Bee Gees']	Islands In The Stream - Live At The MGM Grand/...	0.996
84934	['KISS']	Strutter - Live/1975	0.996

Loudness:

The overall loudness of a track in decibels (db). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db.

Does not follow a normal distribution. Appears to skew to the left. Most of the loudness occurs at -10db.¶

```
Datatype:  float64
Mean:      -11.751
Median:    -10.836
Mode:       -7.578
Std:         5.692
Min:       -60.000
25%:       -14.908
50%:       -10.836
75%:        -7.499
Max:         3.855
Skew:       -0.989
```



Top 10 Songs by Loudness

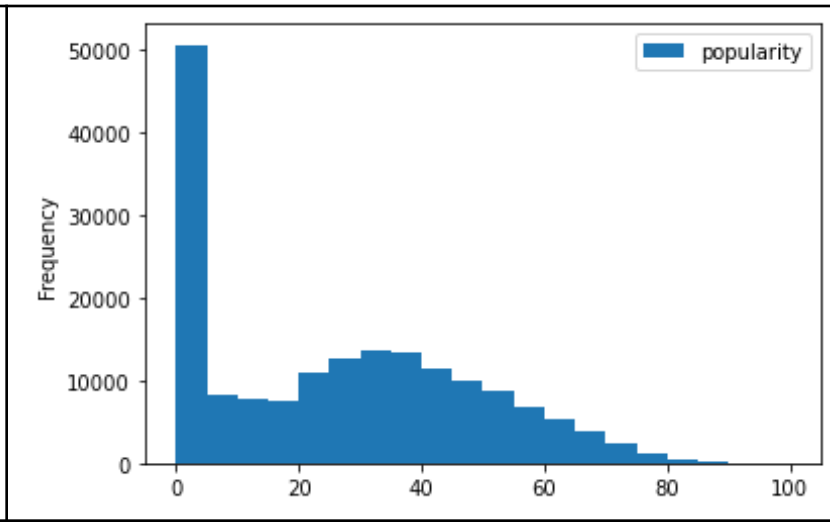
	artists	name	loudness
23004	['Benny Goodman']	Pause Track - Live	-60.000
23073	['Benny Goodman']	Pause Track - Live	-60.000
62366	['Future Rapper']	StaggerLee Has His Day at the Beach	-60.000
62458	['Sarah Vaughan']	Pause Track	-60.000
138635	['Time Bomb Symphony']	You R Heaven	-60.000
144701	['Sarah Vaughan']	Pause Track	-60.000
146842	['Connie Francis']	Hava Nagilah	-60.000
128704	['Igor Stravinsky', 'Michael Tilson Thomas']	Le sacre du printemps (The Rite of Spring): Pr...	-55.000
157919	['HI-FI CAMP']	Cabin Sound	-54.376
144512	['Igor Stravinsky', 'Leonard Bernstein', 'Lond...	The Rite of Spring (Scenes of Pagan Russia in ...	-48.587

Popularity:

The higher the value, the more popular the song is.

Does not follow a normal distribution. Appears to skew to the right. Can't explain or interpret what going on around the 0 range. Could be due to there being a high number of songs that are not listened to on Spotify.

Datatype:	int64
Mean:	25.693
Median:	25.000
Mode:	0
Std:	21.873
Min:	0.000
25%:	1.000
50%:	25.000
75%:	42.000
Max:	100.000
Skew:	0.363



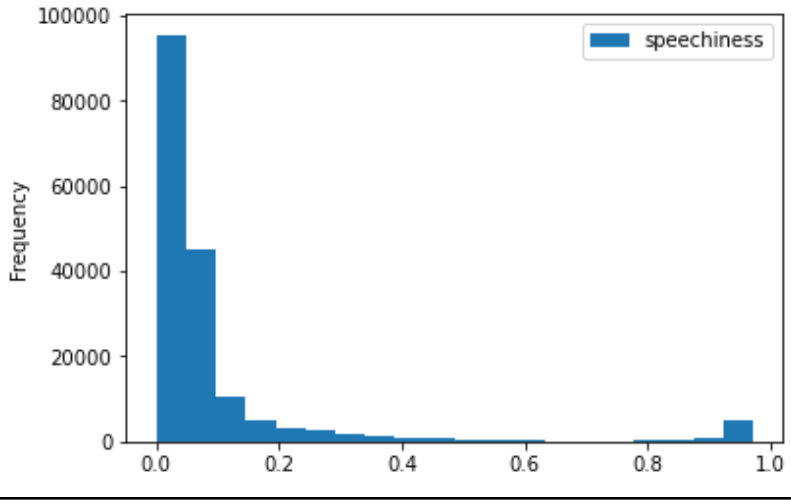
Top 10 Songs by Popularity

	artists	name	popularity
20062	['Olivia Rodrigo']	drivers license	100
19862	['24kGoldn', 'iann dior']	Mood (feat. iann dior)	96
19866	['Ariana Grande']	positions	96
19886	['Bad Bunny', 'Jhay Cortez']	DÁKITI	95
19976	['KAROL G']	BICHOTA	95
19868	['Ariana Grande']	34+35	94
19870	['CJ']	Whoopty	94
19872	['The Kid LAROI']	WITHOUT YOU	94
19876	['Billie Eilish']	Therefore I Am	94
19928	['Bad Bunny', 'ROSALÍA']	LA NOCHE DE ANOCHE	94

Speechiness:

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, such as audio-books. 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.

Does not follow a normal distribution. Appears to skew to the right. Most of the speechiness occurs at the 0.0 range.

Datatype: float64 Mean: 0.106 Median: 0.045 Mode: 0.0337 Std: 0.182 Min: 0.000 25%: 0.035 50%: 0.045 75%: 0.076 Max: 0.971 Skew: 3.751	

Top 10 Songs by Speechiness

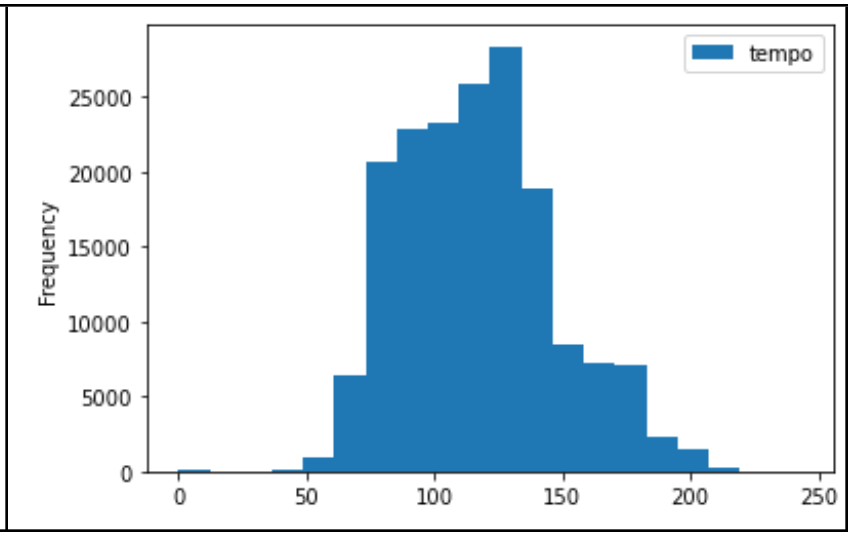
	artists	name	speechiness
162011	['Harper Lee', 'Eva Mattes']	Wer die Nachtigall stört ..., Kapitel 1	0.971
25872	['Ernest Hemingway', 'Christian Brückner']	Kapitel 15 - Der alte Mann und das Meer - Erzä...	0.970
25873	['Ernest Hemingway', 'Christian Brückner']	Kapitel 16 - Der alte Mann und das Meer - Erzä...	0.970
40598	['Эрих Мария Ремарк']	Часть 38.4 & Часть 39.1 - Обратный путь	0.970
41619	['Tadeusz Dolega Mostowicz']	Chapter 16.9 - Doktor Murek zredukowany	0.969
44288	['Georgette Heyer', 'Brigitte Carlsen']	Kapitel 220 - Die drei Ehen der Grand Sophy	0.969
44331	['Georgette Heyer', 'Brigitte Carlsen']	Kapitel 174 - Die drei Ehen der Grand Sophy	0.969
62567	['Georgette Heyer', 'Brigitte Carlsen']	Kapitel 187 - Die drei Ehen der Grand Sophy	0.969
62626	['Georgette Heyer', 'Brigitte Carlsen']	Kapitel 262 - Die drei Ehen der Grand Sophy	0.969
77991	['Tadeusz Dolega Mostowicz']	Chapter 16.19 - Doktor Murek zredukowany	0.969

Tempo:

The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.

Although the distribution almost looks normal. Skews to the right, but not by much. Can't explain why the mode is 0. Requires further investigation.

```
Datatype:  float64
Mean:      117.006
Median:    115.816
Mode:        0.0
Std:        30.254
Min:         0.000
25%:        93.931
50%:       115.816
75%:       135.011
Max:       243.507
Skew:        0.423
```



Top 10 Songs by Tempo

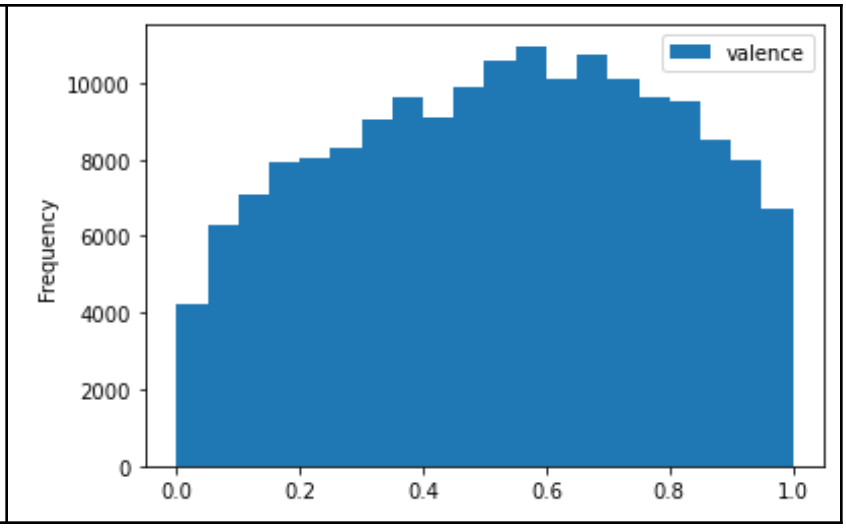
	artists	name	tempo
84474	['Bill Withers']	I Don't Want You on My Mind	243.507
29835	['J.J. Cale']	Call The Doctor	243.372
65985	['Bob Dylan']	Dear Landlord	238.895
29374	['Grateful Dead']	Candyman - 2013 Remaster	236.799
167770	['Suicide']	Surrender - 2005 Remastered Version	224.437
121924	['Portishead']	Undenied	222.605
156668	['Aviation Weather']	Sail over the Storm	221.954
2753	['Bimal Gupta']	Biyer Pare	221.741
136469	['Big Black']	L Dopa	221.112
3947	['Anestis Delias']	To xaremi sto xamam	221.058

Valence:

A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).

Does not follow a normal distribution. It appears to skew to the left.

```
Datatype:  float64
Mean:      0.525
Median:    0.536
Mode:      0.961
Std:       0.264
Min:       0.000
25%:       0.311
50%:       0.536
75%:       0.743
Max:       1.000
Skew:     -0.101
```



Top 10 Songs by Valence

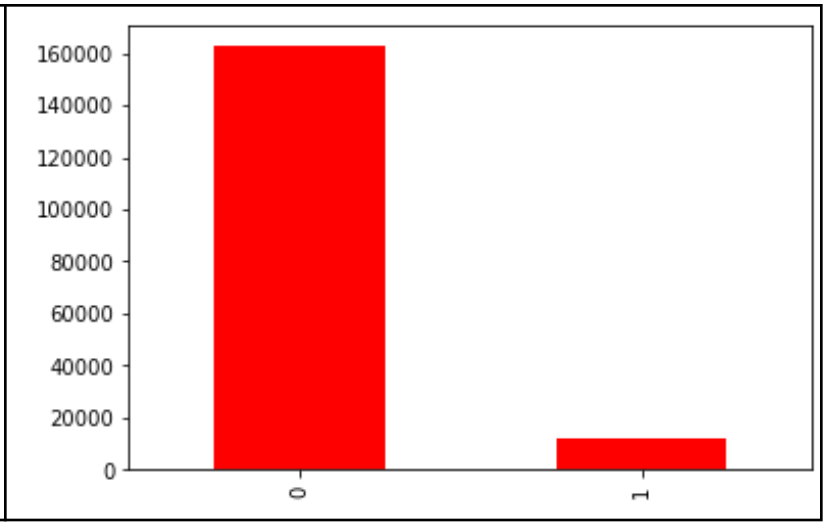
	artists	name	valence
758	['Ignacio Corsini']	La Mina del Ford - Remasterizado	0.0
845	['Francisco Canaro']	Oh Mujer Mujer - Remasterizado	0.0
860	['Ignacio Corsini']	Shangai Bay - Remasterizado	0.0
1217	['Iván Rolón']	Cuatro melodías al unísono, No. III	0.0
2820	['The Moors']	Santa Claus Is Coming To Town	0.0
3302	['Billie Holiday']	Back In Your Own Backyard - Take 1	0.0
3796	['MGM Studio Orchestra']	Munchkinland Insert - Alternate Tag	0.0
3806	['Bert Lahr', 'Judy Garland', 'Ray Bolger', 'B...	If I Were King of the Forest - Partial Take; A...	0.0
4211	['The Slobs']	The Christmas Raid	0.0
5830	['Charlie Parker']	Embraceable You - Live At Carnegie Hall, New Y...	0.0

Explicitness:

Indicates if explicit language was used in the song.

93% of the songs were not explicit and 7% of the songs were explicit.¶

```
Datatype:        int64
Not Explicit:    0.932
Explicit:        0.068
```

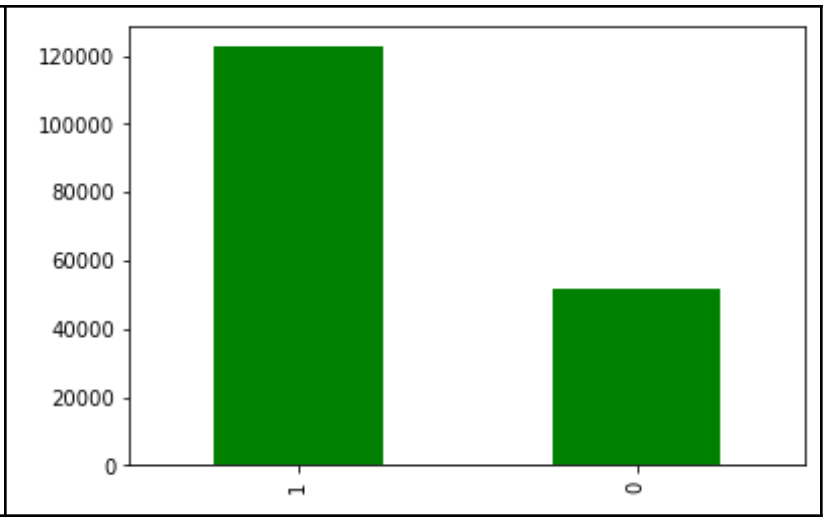


Mode:

Mode indicates the modality (major or minor) of a track, the type of scale from which melodic or harmonic content is derived. Major is represented by 1 and minor is 0.

70% of the melodic content has a major modality and 30% has a minor modality.

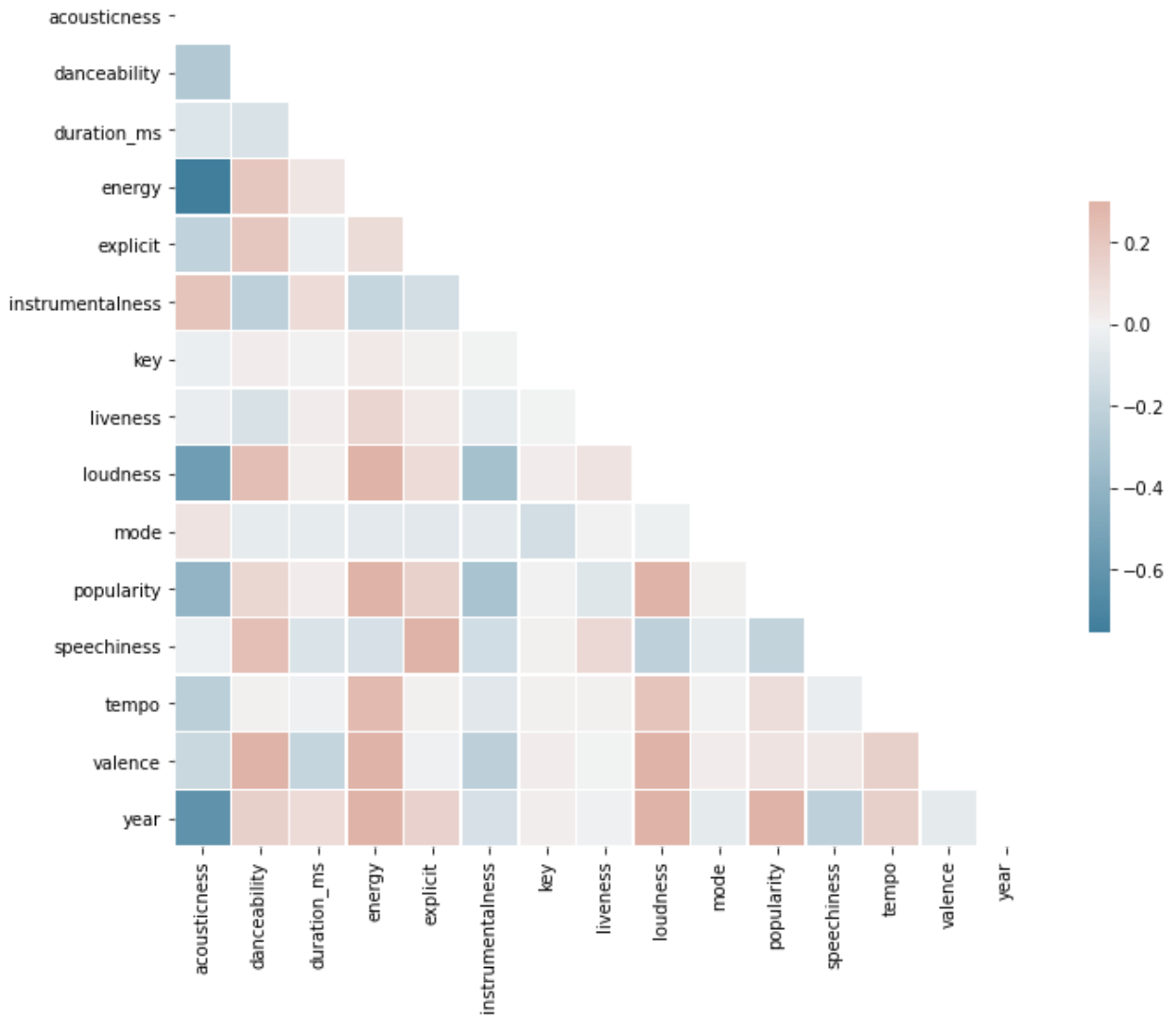
```
Datatype:        int64
Major:           0.702
Minor:           0.298
```



Correlation Matrix

After viewing the correlation matrix, there weren't any high correlation candidates for feature reduction.

Acousticness and energy appears to be the most significantly correlated, but still not enough for consideration.



Top Negative Correlations

```
acousticness      loudness           -0.546639
loudness          acousticness       -0.546639
popularity        acousticness       -0.396744
acousticness      popularity         -0.396744
instrumentalness  loudness           -0.317562
loudness          instrumentalness   -0.317562
popularity        instrumentalness   -0.300625
instrumentalness  popularity         -0.300625
danceability      acousticness       -0.263217
acousticness      danceability       -0.263217
dtype: float64
```

Top Positive Correlations

```
energy        loudness        0.779267
              year            0.540850
year          energy          0.540850
valence       danceability    0.536713
danceability  valence         0.536713
popularity    year            0.513227
year          popularity      0.513227
loudness      year            0.465189
year          loudness        0.465189
speechiness   explicit        0.353872
dtype: float64
```

02 – Pre-Processing

Performed the following tasks:

1) Checked for null values

```
acousticness         0
artists              0
danceability         0
duration_ms          0
energy               0
id                   0
instrumentalness     0
key                  0
liveness             0
loudness             0
name                 0
popularity           0
speechiness          0
tempo                0
valence              0
year                 0
explicit_0           0
explicit_1           0
mode_0               0
mode_1               0
dtype: int64
```

2) Of the 16 numeric attributes, 5 attributes ('key', 'loudness', 'popularity', 'tempo', 'duration_ms') needed additional normalization

	key	loudness	popularity	tempo	duration_ms
0	5	-12.628	12	149.976	168333
1	5	-7.261	7	86.889	150200
2	0	-12.098	4	97.600	163827
3	2	-7.311	17	127.997	422087
4	10	-6.036	2	122.076	165224
...	...	...	...	...	...
174384	6	-5.089	0	125.972	147615
174385	4	-11.665	0	94.710	144720
174386	4	-12.393	0	108.058	218147
174387	0	-12.077	69	171.319	244000
174388	7	-12.237	0	112.208	197710

174389 rows × 5 columns

	key	loudness	popularity	tempo	duration_ms
0	0.454545	0.741868	0.12	0.615900	0.030637
1	0.454545	0.825918	0.07	0.356823	0.027237
2	0.000000	0.750168	0.04	0.400810	0.029792
3	0.181818	0.825135	0.17	0.525640	0.078215
4	0.909091	0.845102	0.02	0.501324	0.030054
...	...	...	...	...	...
174384	0.545455	0.859933	0.00	0.517324	0.026752
174385	0.363636	0.756949	0.00	0.388942	0.026209
174386	0.363636	0.745549	0.00	0.443757	0.039977
174387	0.000000	0.750497	0.69	0.703549	0.044824
174388	0.636364	0.747992	0.00	0.460800	0.036145

174389 rows × 5 columns

3) 2 categorical attributes ('explicit', 'mode') required dummy attribute conversions.

explicit_0	explicit_1	mode_0	mode_1
1	0	1	0
1	0	1	0
1	0	0	1
1	0	0	1
0	1	1	0
...	...	...	...
1	0	1	0
1	0	0	1
1	0	1	0
0	1	0	1
1	0	0	1

4) Binned 'year' and 'release_date' attribute for every 10 years.

	year	year_bin
0	1920	1920s
1	1920	1920s
2	1920	1920s
3	1920	1920s
4	1920	1920s
...	...	...
174384	2020	2020s
174385	2021	2020s
174386	2020	2020s
174387	2021	2020s
174388	2020	2020s

174389 rows × 2 columns

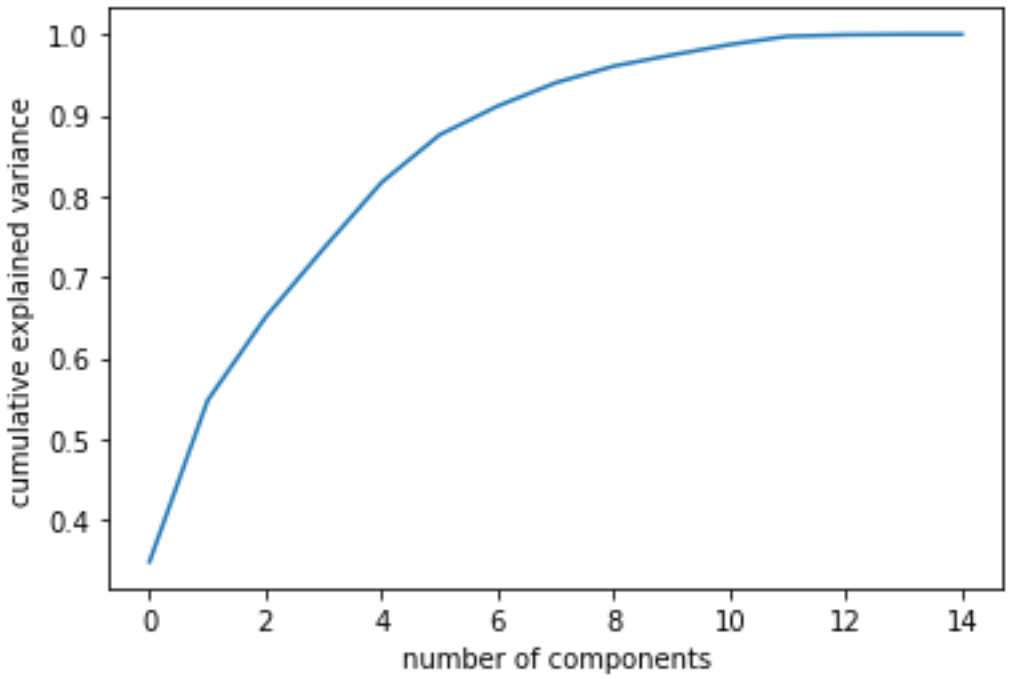
5) Binning 'release_date' would produce the same results as 'year'. Attribute is redundant, so it will be dropped from the dataframe.

03 – Principle Component Analysis

6 Attributes accounted for 0.88 of the variance.

```
mode_0:              0.35
acousticness:        0.20
explicit_0:          0.10
instrumentalness:    0.08
key:                 0.08
valence:             0.06
```

Scree Plot



04 - Cluster Analysis

Cluster model 1: k-means with our PCA attributes

The best number of K clusters was 13 after looking at our silhouette values:

```
df_silhouette_values = df_silhouette_values.sort_values('Silhouette Mean', a
df_silhouette_values
```

	K	Silhouette Values	Silhouette Mean
10	13	[0.5043133969549626, 0.12129033611852735, 0.45...	0.325729
11	14	[0.5041760651787965, 0.12082550614254375, 0.44...	0.322958
12	15	[0.5041677375375535, 0.12082837722410662, 0.45...	0.320459
8	11	[0.5027327944517269, 0.171960102762028, 0.4540...	0.316400
13	16	[0.10999961368428665, 0.06828125808006094, 0.1...	0.313898
16	19	[0.12398628113536839, 0.06323588640800475, 0.2...	0.312377
9	12	[0.5027327944517269, 0.171960102762028, 0.4549...	0.310722
14	17	[0.5041807665896382, 0.12088230678792537, 0.24...	0.310606
17	20	[0.12452963488841297, 0.06281076180394472, 0.1...	0.309180
15	18	[0.12408768859075439, 0.0635879679059271, 0.23...	0.307994

The cluster centroids:

	mode_0	acousticness	explicit_0	instrumentalness	key	valence
0	-0.00	0.09	1.00	0.77	0.43	0.45
1	-0.00	0.80	1.00	0.03	0.73	0.54
2	1.00	0.84	1.00	0.04	0.48	0.50
3	-0.00	0.17	1.00	0.03	0.14	0.62
4	-0.00	0.82	1.00	0.04	0.18	0.48
5	-0.00	0.93	1.00	0.83	0.45	0.40
6	1.00	0.07	1.00	0.79	0.61	0.44
7	-0.00	0.16	1.00	0.02	0.71	0.61
8	1.00	0.16	1.00	0.03	0.29	0.58
9	1.00	0.92	1.00	0.82	0.47	0.39
10	-0.00	0.20	0.00	0.03	0.40	0.51
11	1.00	0.19	1.00	0.03	0.87	0.60
12	1.00	0.21	0.00	0.03	0.60	0.52

Size of each cluster:

```
Size of Cluster 0 =  7290
Size of Cluster 1 =  23732
Size of Cluster 2 =  14037
Size of Cluster 3 =  20718
Size of Cluster 4 =  23099
Size of Cluster 5 =  16989
Size of Cluster 6 =  4760
Size of Cluster 7 =  23589
Size of Cluster 8 =  10617
Size of Cluster 9 =  8014
Size of Cluster 10 =  7071
Size of Cluster 11 =  9692
Size of Cluster 12 =  4781
```

We then added the cluster assignment to each data point, and created a new csv value with the new findings. The cluster assignments later became class labels during KNN.

Cluster model 2: k-means without PCA.

The best number of K clusters was 3 after looking at silhouette values:

:

	K	Silhouette Values	Silhouette Mean
0	3	[0.32491168 0.37797116 0.28125591 ... 0.306523...	0.251984
1	5	[0.30632806 0.0360266 0.24481317 ... 0.186661...	0.247469
2	6	[0.30638074 0.03588191 0.27893815 ... 0.186689...	0.239975
3	8	[0.30695186 0.03480905 0.27013015 ... 0.186707...	0.233284
4	4	[0.36286317 0.08450712 0.28091655 ... 0.252633...	0.228270
5	14	[0.35240864 0.06631845 0.24586632 ... 0.225948...	0.221531
6	7	[0.30662062 0.03536158 0.2702072 ... 0.186775...	0.218903
7	11	[0.35433616 0.14229329 0.28037828 ... 0.300021...	0.218696
8	9	[0.30691768 0.03484066 0.27597587 ... 0.186733...	0.218532

Centroids (subset):

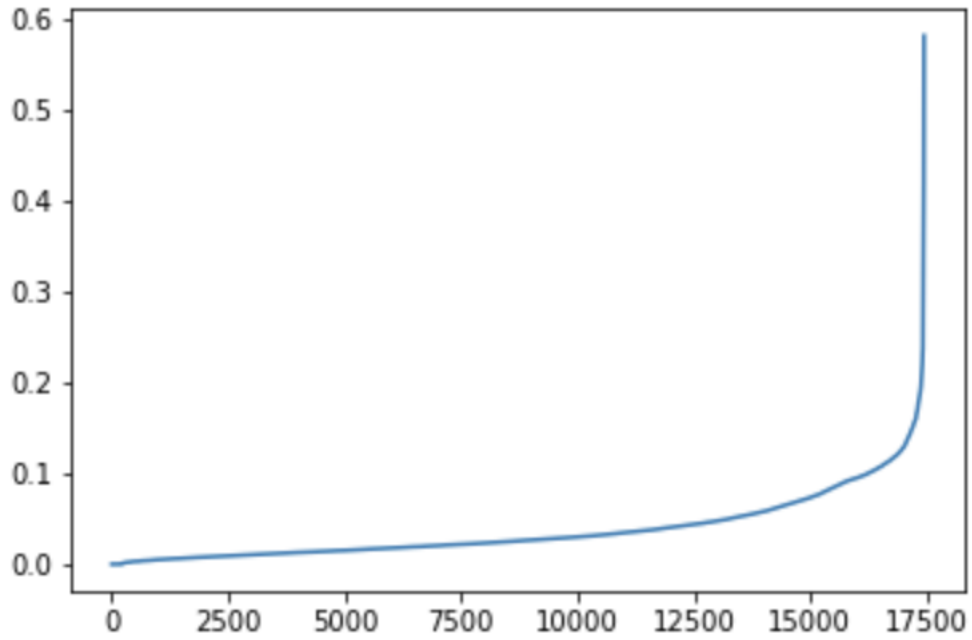
	acousticness	danceability	duration_ms	energy	instrumentalness	key	liveness	loudness	popularity	speechiness	tempo	valence	explicit_0	explicit_1
0	0.19	0.58	0.04	0.67	0.11	0.44	0.22	0.80	0.35	0.09	0.51	0.60	0.90	0.10
1	0.46	0.55	0.04	0.51	0.23	0.54	0.21	0.76	0.25	0.12	0.48	0.52	0.91	0.09
2	0.85	0.48	0.04	0.27	0.27	0.45	0.21	0.71	0.17	0.11	0.45	0.45	0.99	0.01

Size of each cluster:

```
Size of Cluster 0 =  51901
Size of Cluster 1 =  60903
Size of Cluster 2 =  61585
```

A new csv file was created with these cluster assignments.

Cluster model 3: DBSCAN with our PCA attributes
Using the 6 attributes from PCA, a random sample of the data and the knee plot below, we determined the best fit for epsilon to be roughly 0.18, seen below as the point of highest curvature in the plot.

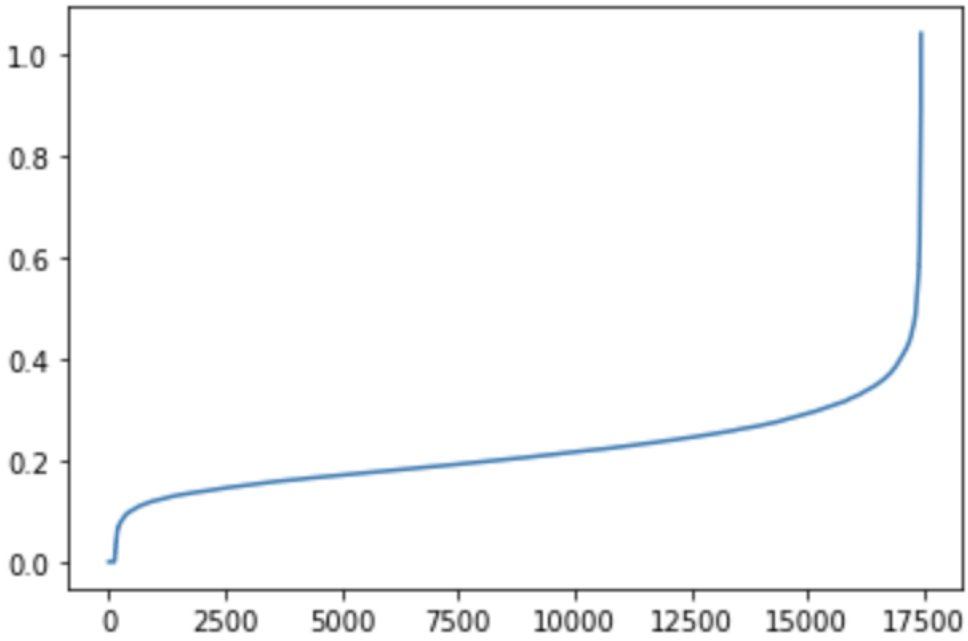


Applying DBSCAN with epsilon = 0.18 and min_sample = 20, we found an estimated 11 clusters with 7.3% of the sample as outliers. We also found the silhouette value to be 0.186.

```
Estimated number of clusters: 11
Estimated number of noise points: 1275
Percent estimated outliers: 7.3
Silhouette Coefficient: 0.186
```

We then applied the cluster assignments to the entire dataset to create labels and saved the new csv file.

Cluster model 4: DBSCAN without PCA attributes
Without first applying PCA, we used DBSCAN using each numerical attribute of the Spotify dataset. Using the knee plot on a random sample of 10% of the total data below, we can see that the greatest curvature occurs at about 0.4.



Applying DBSCAN with epsilon = 0.4 and min_sample = 20, we found an estimated 7 clusters with 8.7% of the sample as outliers and a silhouette value of 0.264. This silhouette coefficient is a 40% improvement than DBSCAN with PCA.

```
Estimated number of clusters: 7
Estimated number of noise points: 1518
Percent estimated outliers: 8.7
Silhouette Coefficient: 0.264
```

Using silhouette values to assess these 4 clustering techniques, we determined the best model to be K-Means using PCA with a mean silhouette value of 0.33 for K = 13. That is the model we will employ for our app.

05 – Classification

KNN using K-Means clusters with and without PCA

With PCA

We found K = 37 to have the highest accuracy with PCA reduction.

	K	Accuracy
7	37.0	0.992603
1	31.0	0.992574
2	32.0	0.992545
3	24.0	0.992517
0	30.0	0.992459
3	33.0	0.992431
5	26.0	0.992402
15	45.0	0.992402
5	35.0	0.992373
14	44.0	0.992345

Without PCA

We found K = 21 to have the highest accuracy for K-Means without PCA reduction.

	K	Accuracy
20	21.0	0.991427
20	21.0	0.991427
18	19.0	0.991427
18	19.0	0.991427
19	20.0	0.991399
32	33.0	0.991399
32	33.0	0.991399
19	20.0	0.991399
28	29.0	0.991313

This section of our project has shown that effective KNN classifiers can be developed using our clusters found from previous experiments. The first KNN was using our PCA-Kmeans clusters, and the most effective K was 37. The second KNN classifier was with all features and clusters using Kmeans, and the best K was found to be 21. KNN with our DBSCAN or HAC clusters will not be implemented for this project.

06 – App Implementation

The app integrated our cluster data and KNN classifier to create an example of what command line user input could look like. This app addressed the original aim of the project, which was to create a “playlist generator” that could provide the user with a list of similar songs. The app has two main features- one, a new playlist is generated given the user chooses one song out of a list of ten already existing in the dataset; and two, a new playlist is generated given the user inputs a completely new song the app does not know about. The first feature is based solely off of the cluster results from the cluster analysis, while the second feature utilizes the KNN classifier that was developed using the KMeans with PCA cluster data set.

Example screenshot of first feature:

```
**********************************************************************************************************
Welcome to the Playlist generator using song data from Spotify. First, choose your method of clustering:
**********************************************************************************************************


            What cluster model would you like to use?
            (1) kmeans with PCA
            (2) kmeans without PCA
            (3) DBSCAN with PCA
            (4) DBSCAN without PCA

Enter cluster method, or press enter with no number for default=KMeans with PCA: 1
Do you want to generate playlist based on an existing song, or enter your own? default=existing
(n) for new song, (e) for existing: e

        Choose the song by entering corresponding number from the following list which will be used
        to generate playlist of similar songs, (enter) to see new list, (q) to quit.
        (Note- using your own song will only use the default=KMeans with PCA clusters for the classification)


            (0)
            Artist: ['Mala Karim']
            Song: Aw Kcha Law Mala
            Decade: 1920s


            (1)
            Artist: ['Johann Sebastian Bach', 'Karl Erb', 'Concertgebouworkest', 'Willem Mengelberg']
            Song: St. Matthew Passion, BWV 244 - Part Two: No.31 Evangelist: "Die aber Jesum gegriffen hatten"
            Decade: 1950s


            (2)
            Artist: ['Shakira']
            Song: Octavo Día
            Decade: 1990s
```

As can be seen, the user is asked to choose one of the four clustering types that was done in the cluster analysis and the app then uses the appropriate dataset that has the cluster assignments; when the user chooses one of the listed songs, the app simply finds 20 random songs in the same cluster.

Example screenshot of second feature:

```
Enter cluster method, or press enter with no number for default=KMeans with PCA: 1
Do you want to generate playlist based on an existing song, or enter your own? default=existing
(n) for new song, (e) for existing: n
Song name: Song1
Artist: Artist1
Major (1) or minor (0): 1
Level of acousticness (from 0-1): .80
Explicit content (1) or not (0): 0
Level of instrumentalness (from 0-1): .50
Key (approximate from 0-1 where C is 0 and B is 1, i.e C#=.09): .80
Valence (musical happiness level): .11
[8]
                                              artists                                                name   year
0                                        ['DJ Shadow']              What Does Your Soul Look Like - Pt. 4  1990s
1               ['Francisco Canaro', 'Carlos Roldán']                        Mi Castigo - Remasterizado  1940s
2                                     ['Jesse Cook']                                Mario Takes A Walk  1990s
3                                  ['R. Carlos Nakai']                          Song For the Morning Star  1980s
4                                ['Kiev Chamber Choir']   Seven Spiritual Songs, Op. 3: III. Gloria (2005)  2010s
5                 ['Frédéric Chopin', 'Maryla Jonas']               Waltz in C-Sharp Minor, Op. 64 No. 2  1950s
6                                   ['Mohammed Rafi']                      Ae Dil Tujhi Ko Nind Na Aai  1940s
7                                    ['Wynton Kelly']                         On Green Dolphin Street  1950s
8                          ['Brian Eno', 'Harold Budd']           Above Chiangmai - 2004 Digital Remaster  1980s
9                                    ['Miklós Rózsa']                                    Anno Domini  1960s
10                 ['Los Yumbos', 'Los Provincianos']                                         Sicuris  1930s
11                              ['Οδυσσέας Μοσχονάς']                        Μες της Πεντέλης τα βουνά  1930s
12           ['Modest Mussorgsky', 'William Kapell']    Pictures at an Exhibition: Il vecchio castello  1940s
13                    ['Lalita Phadke', 'Balakram']                        Bachpan Mera Bachpan Tera  1940s
14                         ['Mohammed Abdel Wahab']                                  El Naby Habibak  1950s
15                                   ['Joe Hisaishi']                          Departure -To the West-  1990s
16  ['Pyotr Ilyich Tchaikovsky', 'Leningrad Philha...  Symphony No.5 In E Minor, Op.64, TH.29: 1. And...  1960s
17          ['Frédéric Chopin', 'Arthur Rubinstein']               Nocturnes, Op. 15: No. 3 in G Minor  1960s
18              ['Frédéric Chopin', 'Robert Lortat']  12 Études, Op. 10: No. 2 in A Minor "Chromatique"  1930s
19               ['Herb Alpert & The Tijuana Brass']                               A Taste Of Honey  1960s

Create new playlist? (y/n)
```

Here, the user is asked to input aspects of their new song. The attributes are those that were discovered during our PCA analysis of the dataset. The KNN classifier then assigns it a cluster number, and then 20 songs in the existing data set are chosen based on that cluster assignment.

The video demonstrating the app from our presentation is linked here.