Spotify Popularity Prediction And Recommender

DSI 19 Capstone Project
Adrian Teng

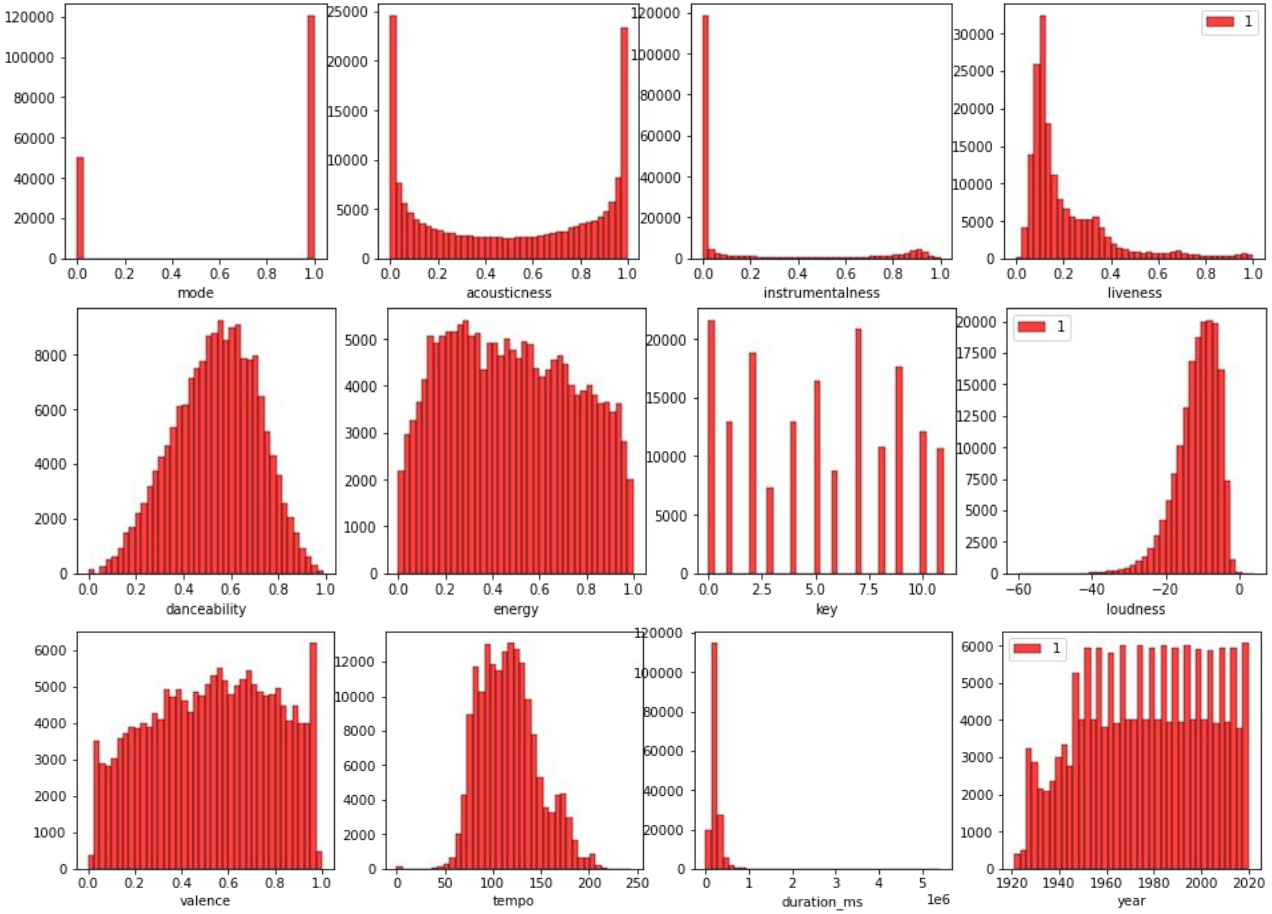
Problem Statement

An event company hopes to come out with an albums that matches their theme of event due to the rise for varieties of events and keeping the similarities of each song tightly. As they hope to reduce the workloads of music coordinator, especially during this period of time , when live performance is not allowed due to pandemic and using music playlist will be the most optimal choice.

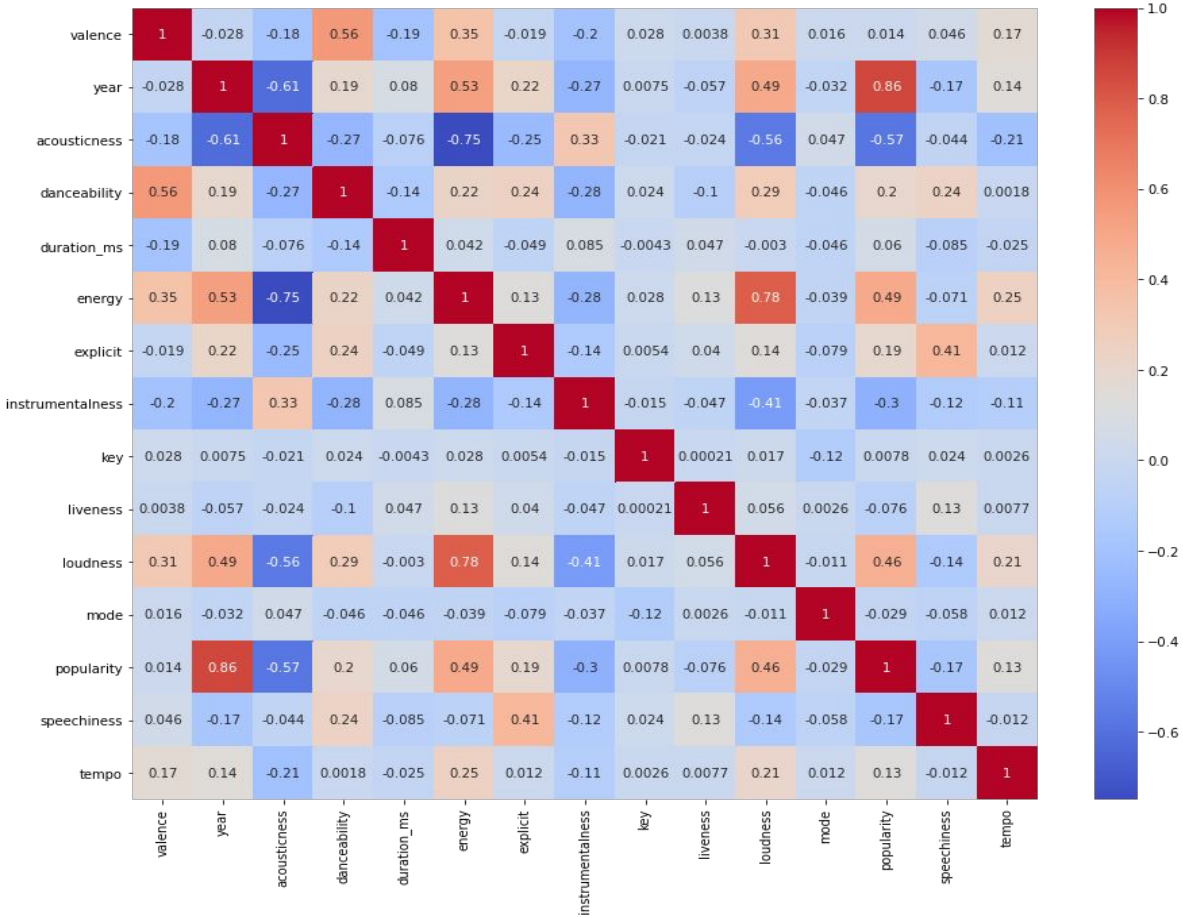
Objective

- Find the best Regression Model to predict the popularity from the song attributes taken from Spotify data.
- Predict and recommend songs according to song attributes similarities.

EDA



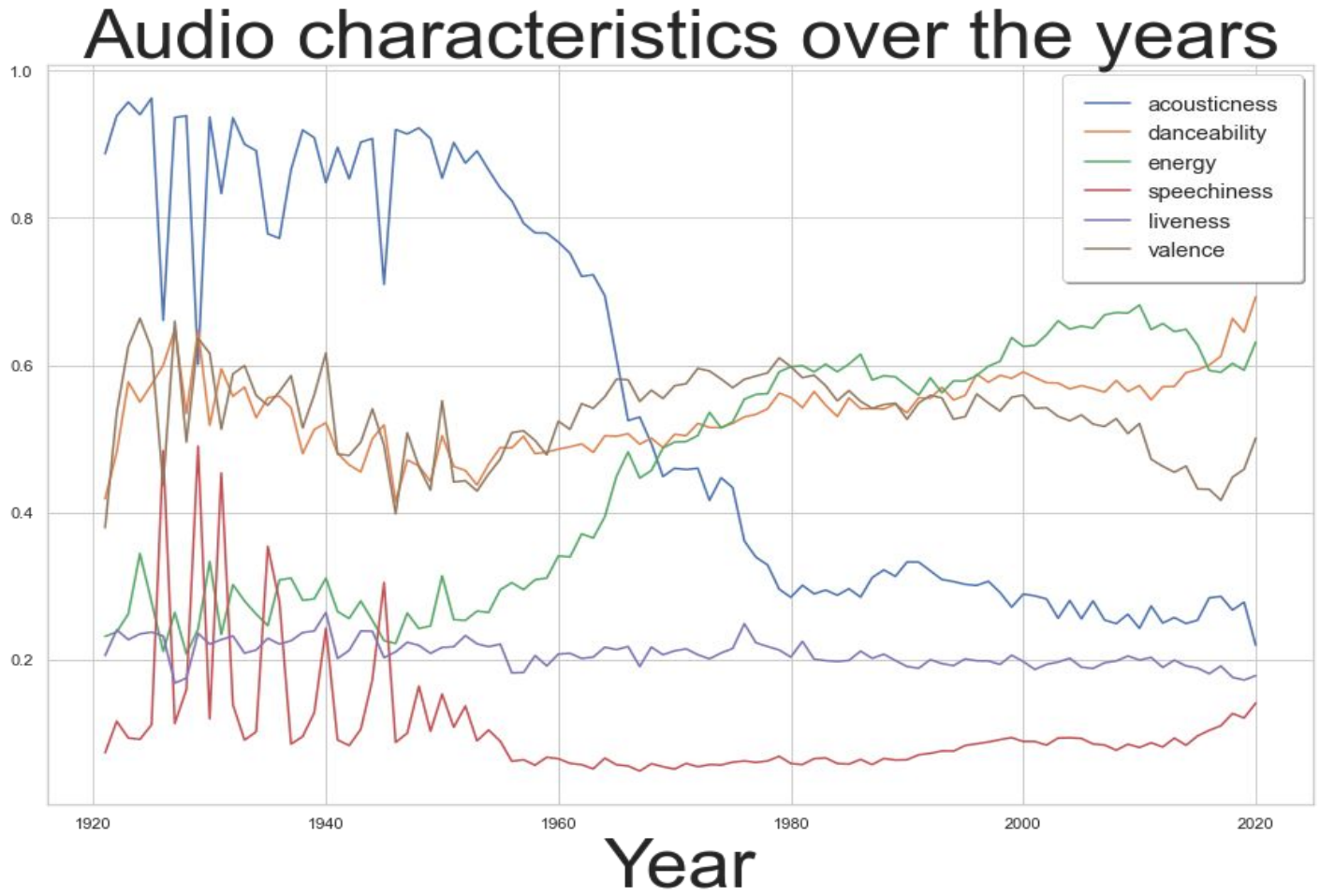
- Acousticness
- Danceability
- Energy
- Valence



The dataset has very low correlation.

The highest correlation is between year and popularity

Energy is slightly correlated to popularity

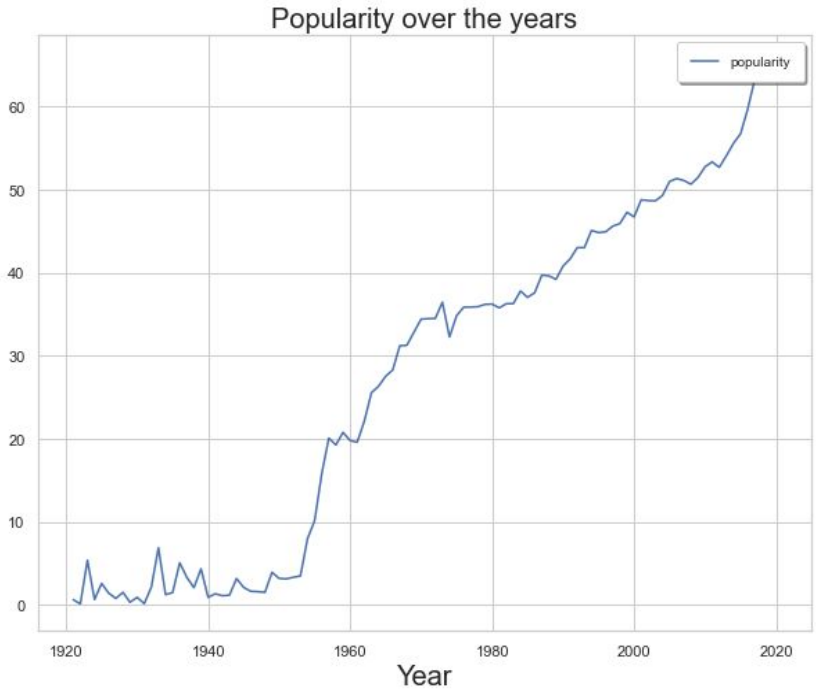


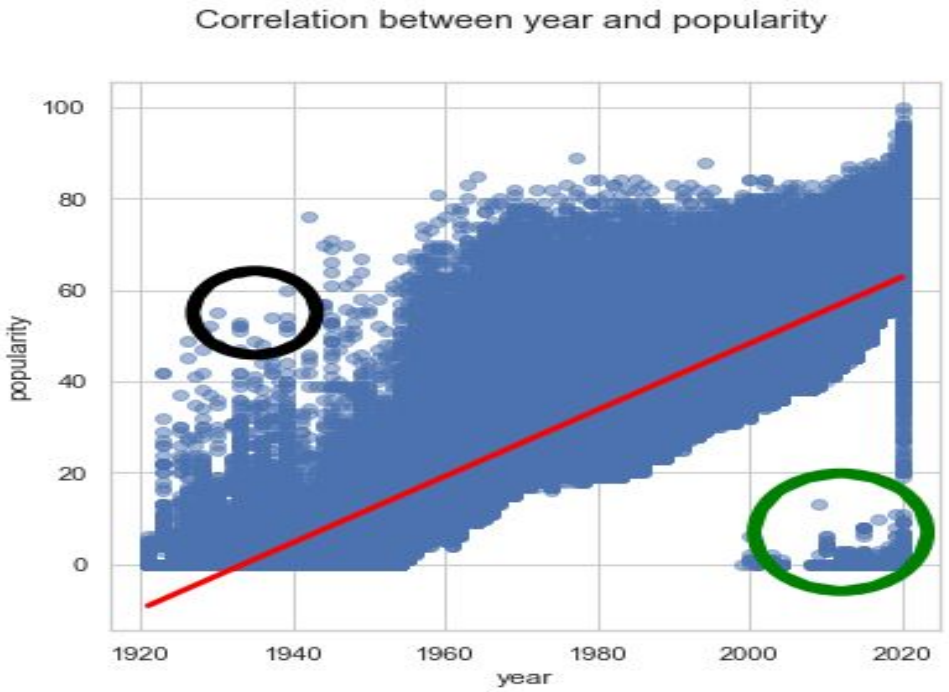
Over the years increased:
- energy
- danceability

Over the years decreased:
- acousticness
- speechiness

Over the years no change:
- liveness
- valence

Popularity over the years





Feature Selection

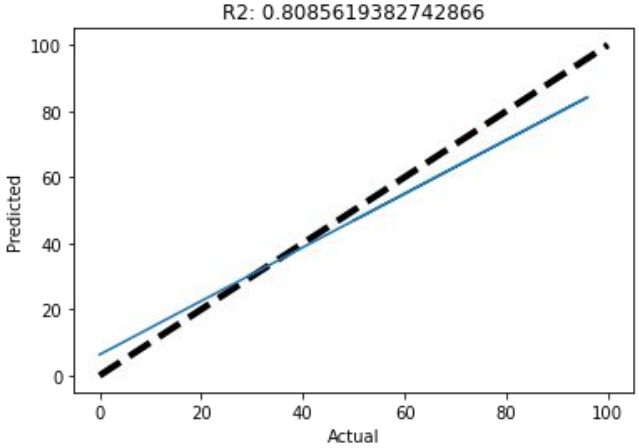
Recursive feature elimination:

- Key
- Duration

	1	2	3	4	5	6	7	8	9	10	11	12	13
valence	9	8	7	6	5	4	3	2	1	1	1	1	1
year	1	1	1	1	1	1	1	1	1	1	1	1	1
acousticness	3	2	1	1	1	1	1	1	1	1	1	1	1
danceability	5	4	3	2	1	1	1	1	1	1	1	1	1
duration_ms	13	12	11	10	9	8	7	6	5	4	3	2	1
energy	7	6	5	4	3	2	1	1	1	1	1	1	1
explicit	8	7	6	5	4	3	2	1	1	1	1	1	1
instrumentalness	2	1	1	1	1	1	1	1	1	1	1	1	1
key	14	13	12	11	10	9	8	7	6	5	4	3	2
liveness	6	5	4	3	2	1	1	1	1	1	1	1	1
loudness	11	10	9	8	7	6	5	4	3	2	1	1	1
mode	10	9	8	7	6	5	4	3	2	1	1	1	1
speechiness	4	3	2	1	1	1	1	1	1	1	1	1	1
tempo	12	11	10	9	8	7	6	5	4	3	2	1	1

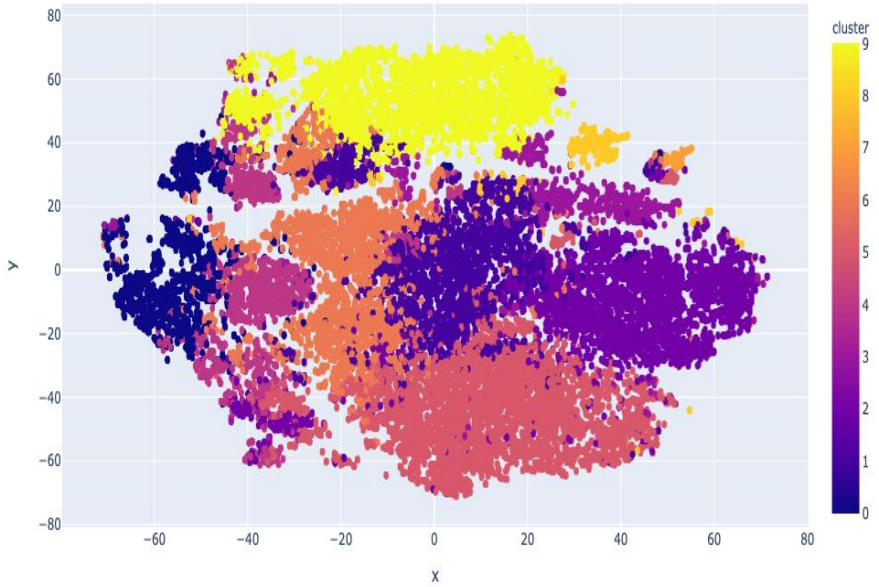
Regression Model

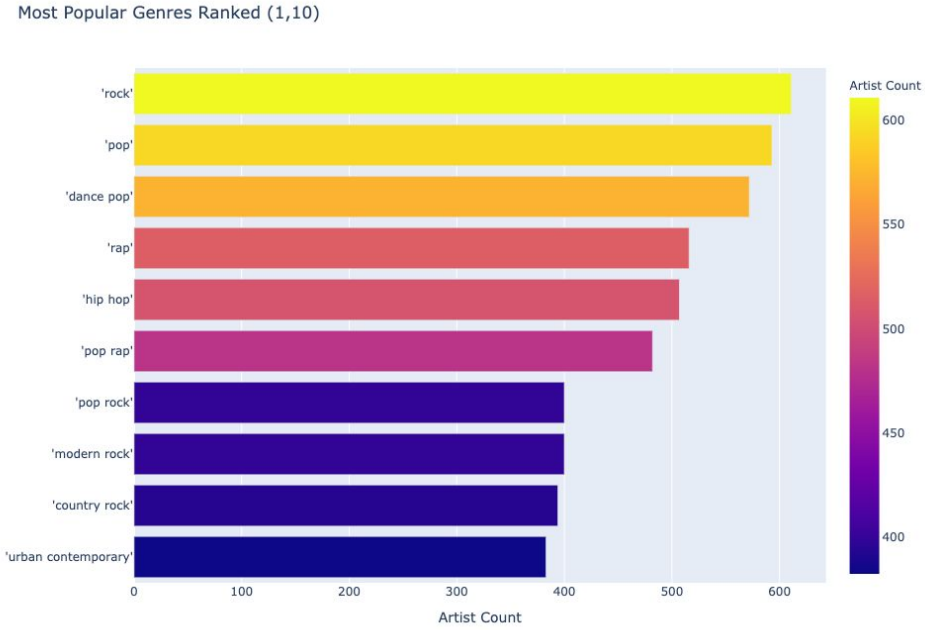
- Linear Regression
- DecisionTree Regression
- RandomForest Regression
- AdaBoost Regression



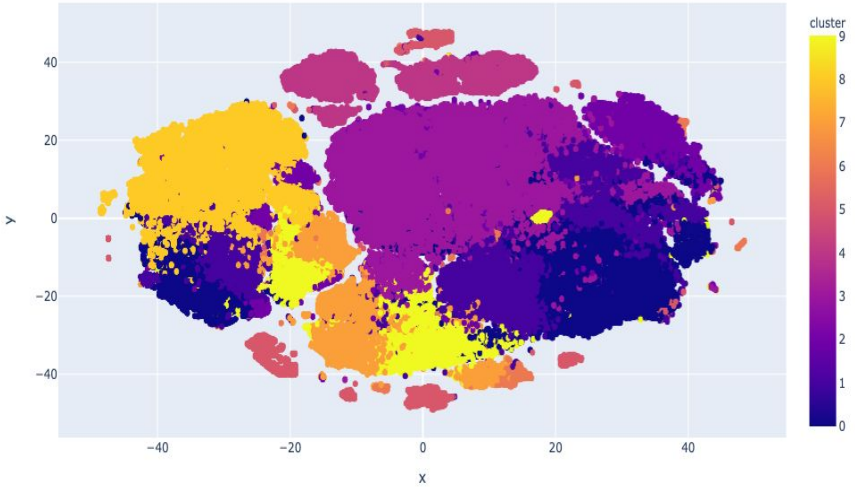
	step	train_accuracy	test_accuracy	r2_score	mean_squared_error	mean_absolute_error
0	lr	0.753055	0.758100	0.758100	115.637734	7.982060
1	rfr	0.929969	0.811810	0.811810	89.962224	6.697909
2	dtr	0.997145	0.607110	0.607110	187.816956	9.245266
3	abr	0.683610	0.683541	0.683541	151.279952	10.009932

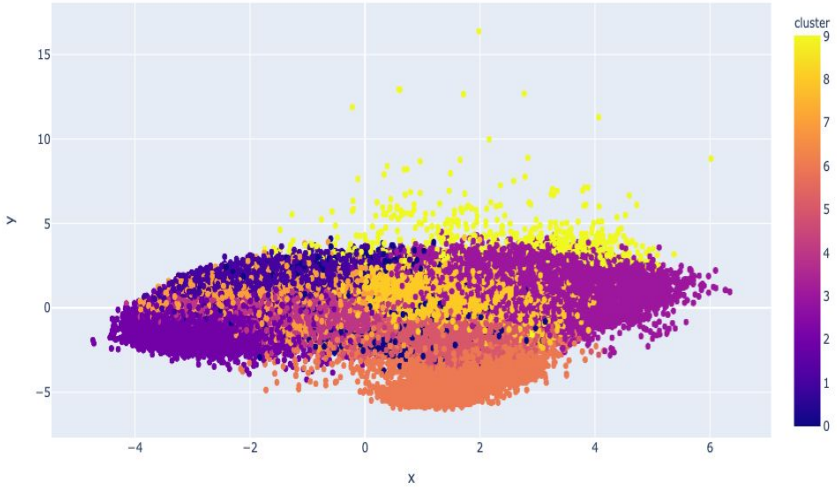
Clustering(Kmeans) for genre





TSNE & PCA Clustering(Kmean) for songs by song attributes

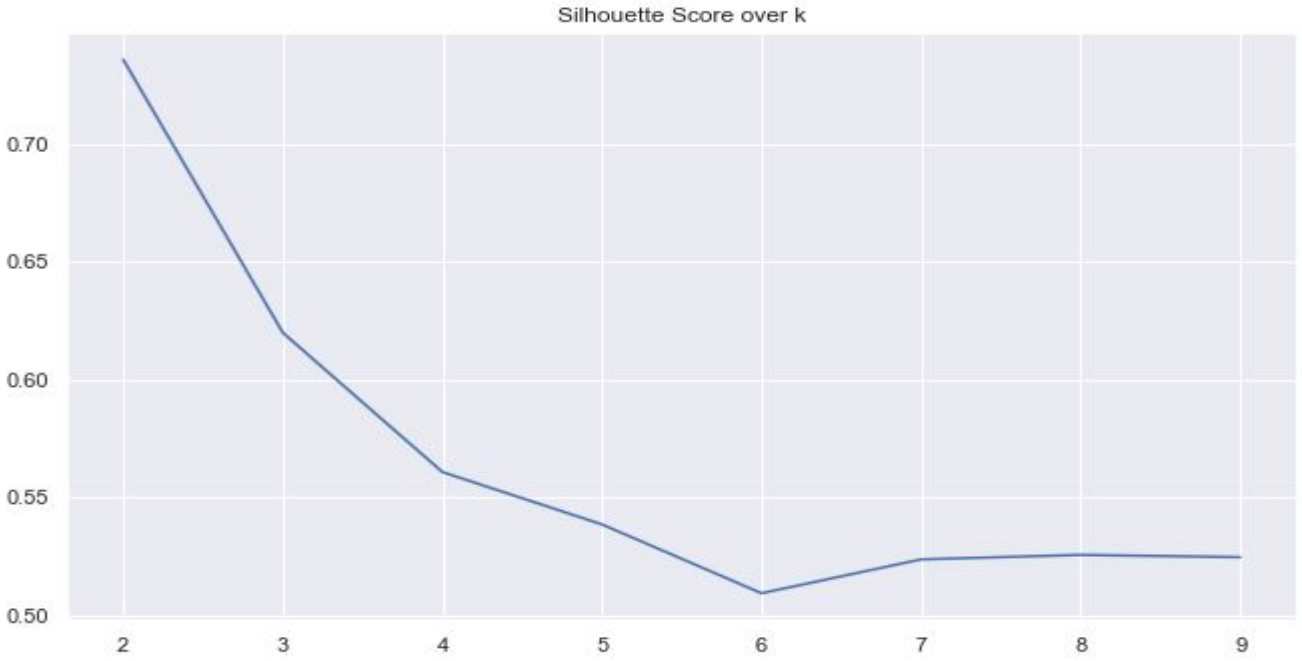




KMeans TSNE Scaled Silhouette Score: 0.35231417417526245

KMeans PCA Scaled Silhouette Score: 0.3387676713434714

Silhouette Score over no of clusters



Recommender(cosine similarity)

```
rank_song_similarity_by_measure(final_merged_df, 'Learn to Fly', 'Foo Fighters',4, '2000s')
```

	artists	similar song to Learn to Fly	popularity	cluster_label	genres	decade	similarity with song	acousticness	danceability	energ
72007	Foo Fighters	Monkey Wrench	50	1	['post-grunge', 'modern rock', 'alternative me...	2000s	0.994196	0.000021	0.399798	0.9
70090	Weezer	Don't Let Go	40	1	['alternative rock', 'modern rock', 'permanent...	2000s	0.993160	0.000038	0.380567	0.9
77296	Puddle Of Mudd	Control	63	2	['alternative rock', 'post-grunge', 'alternati...	2000s	0.992941	0.003092	0.449393	0.9
82375	Silversun Pickups	Little Lover's so Polite	40	1	['alternative rock', 'modern rock', 'alternati...	2000s	0.992735	0.001536	0.480769	0.8
80848	Arcade Fire	Neighborhood #3 (Power Out)	52	1	['alternative rock', 'modern rock', 'permanent...	2000s	0.992709	0.000545	0.528340	0.9

Conclusion

- Based on it's audio features, RandomForest outperform the rest, however looking at the outliers in EDA. We can assume that predicting popularity with audio features are not enough. Artists and probably song lyrics will play a part in a more accurate prediction.
- TSNE is better than PCA for this dataset.
- Kmeans and Cosine similarity are comparable in the calculation for similarities with audio features.

Thank You

Q &A?