

Spotify's Top 200 Charts:

Unsupervised Learning & Recommendations

Members: Nate, Lisa, Mateo, Geo, Sam



Introduction

Since we've never worked with audio data or classification of audio data we wanted to try working with data that is structured as such.



Overarching Question

We ask the question:

Is there a relationship between the some of these features and, if so, how are they correlated?

Additionally, how can we use these features to cluster songs based on these audio tracks of songs represented by numeric features?

Hypothesis

The distribution of certain genres will be different among different groups of songs. These differences in distributions will allow us to perform unsupervised learning on the data to cluster the songs into different groups/listening personas.

For example, the mean tempo of Pop artists will be higher than that of Ballad singers since Pop songs tend to be more upbeat and fast.

If numerical data is extracted from the songs then models can be trained to cluster/classify songs into different groups since there will be enough difference between certain features. This approach of comparing audio features between two groups can then be applied to other projects, such as comparisons of living beings/objects to classify the two.

Challenges



Data

~3000 rows, 15 columns



Models

K-Means? Gaussian Mixture?
Nearest Neighbors?



Interpretation

Why did one model work
'better' than another one?

01



Exploratory Data Analysis

Data *Storytelling*



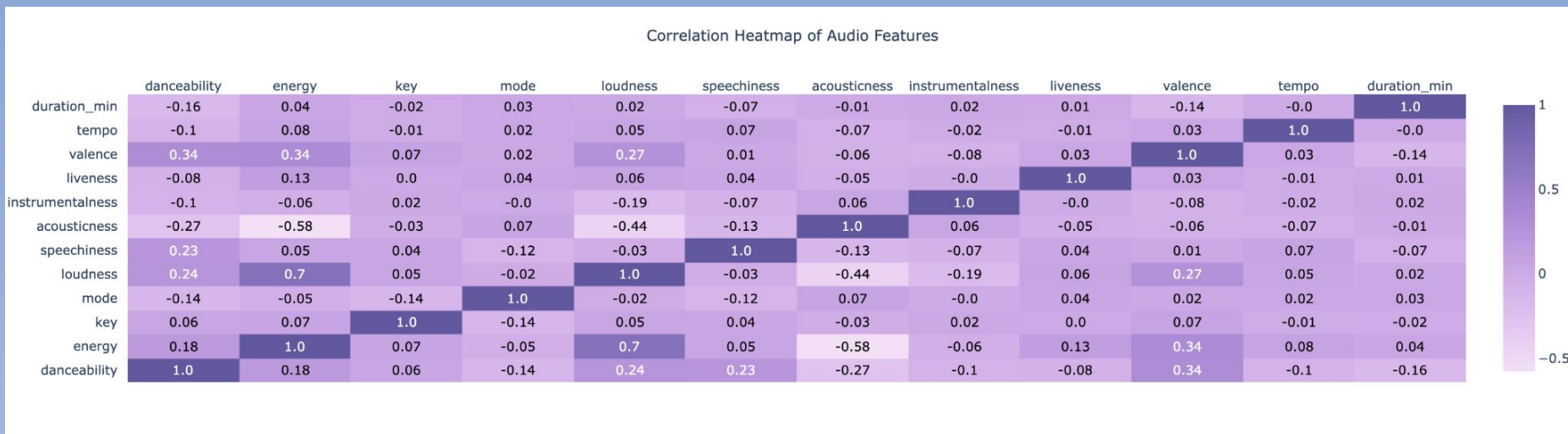
About the Data



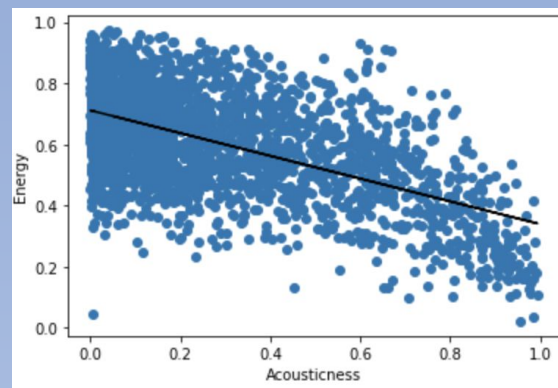
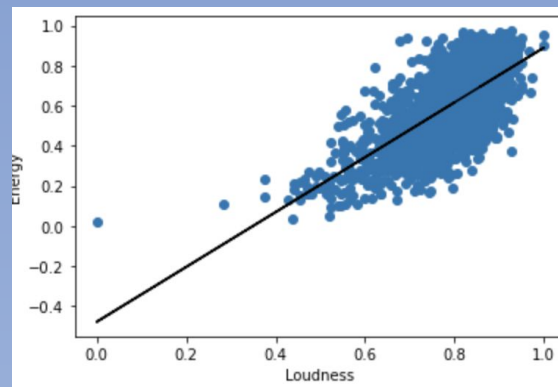
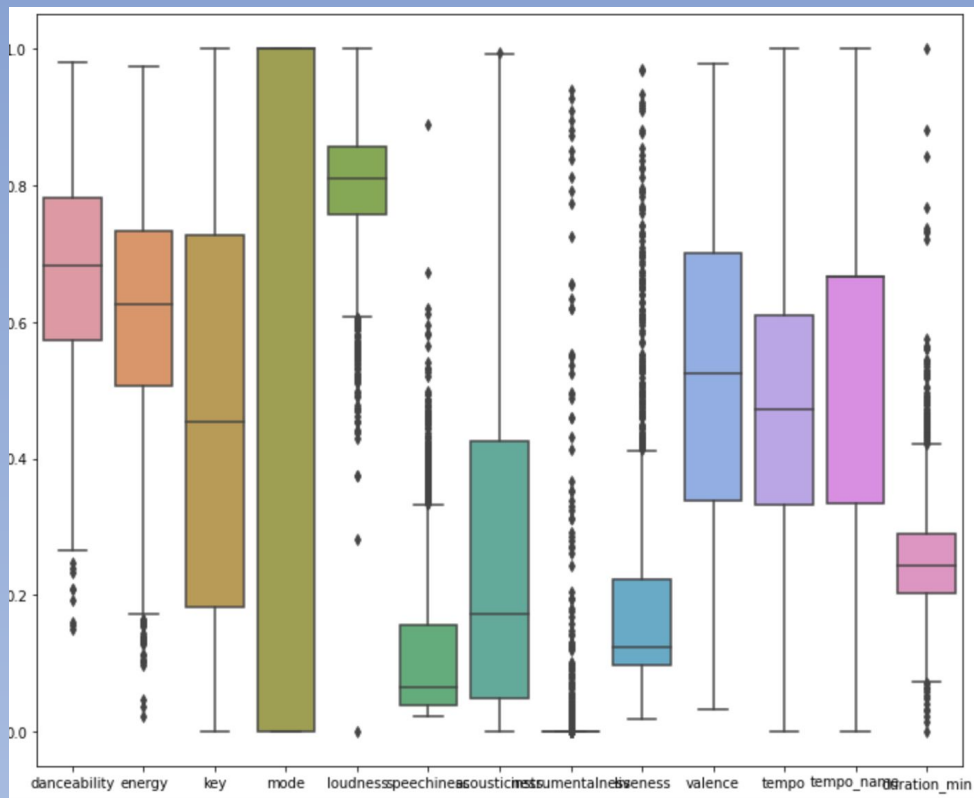
Data *Visualization*

Inspired by the story of the original creation of the current Spotify recommending system by Erik Berghardsson.

Correlation Between Features



Distributions & Correlations of Audio Features





02

Data Cleaning

[illegible]



How we chose variables





How we chose variables:

- **PCA?** Not good for categorical, did not give us significant and easily interpretable components.
- **Solution:** Correlation Threshold

Why a threshold?

- Easy to access and evaluate
 - Pipeline objects allow for easy implementation due to custom transformers if we decide to convert to supervised
- 

03

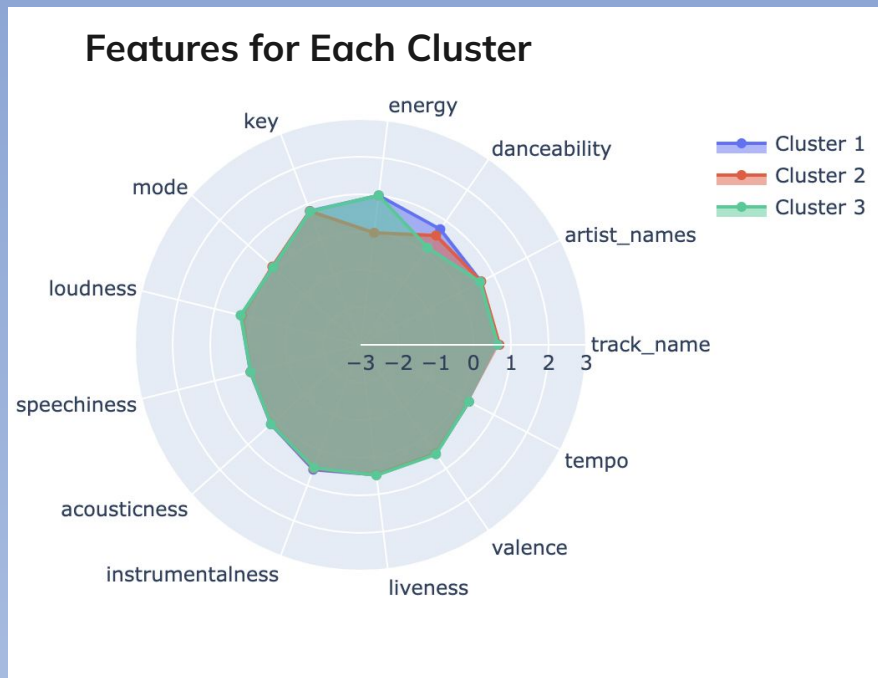
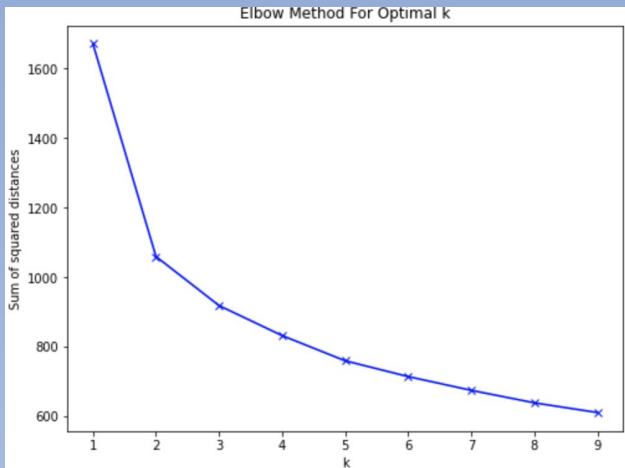
Different Models / Approaches



K-Means Clustering

High level concept: creating k clusters where the center is a circular shape and points are assigned to their closest centroid.

Through the elbow method, we found that 3 was the best hyperparameter.

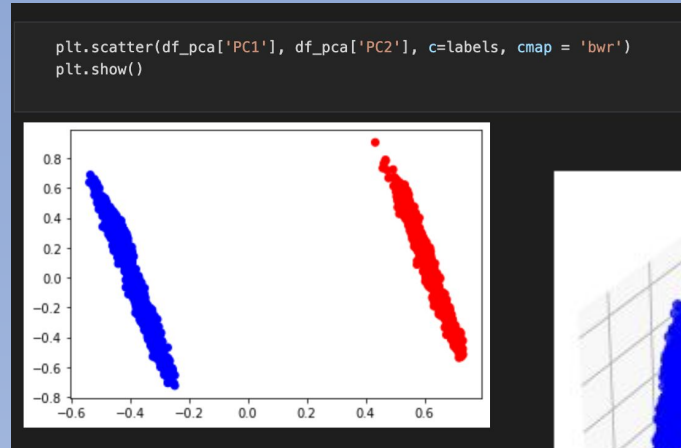


Gaussian Mixture Model

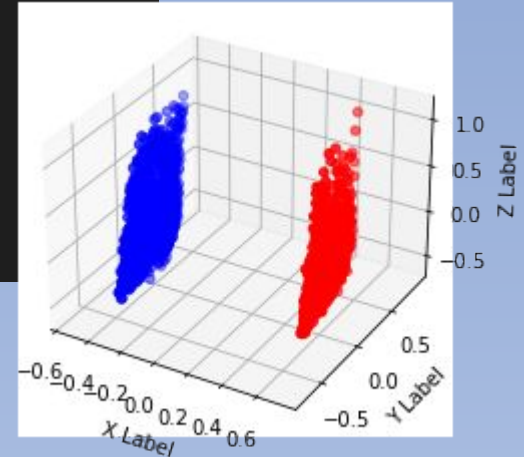
High level concept: assuming the data set represents multiple Gaussian (normal) distributions, clustering based on each point's probability of belonging to each distinct distribution. Soft clustering method

Using the optimal K, we observed 2 distinct groups, meaning 2 different distributions were discovered once the data was transformed by principal component analysis.

The lower dimensionality still portrays 72% of the variance due to the 3 principal components.



Silhouette Score = .53



K-Nearest Neighbors

High level concept:

Supervised learning by calculating the distance between a given data point and its k-nearest neighbors in the training dataset.

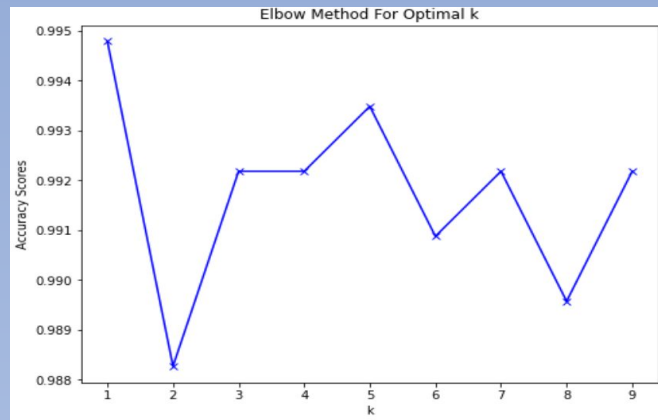
Then, the model makes predictions based on the majority class **or** the average value of the nearest neighbors.

We performed k-nearest neighbors clustering with train-test split of 0.3 for the test data.

- > 0.99 accuracy for nearly all values of nearest neighbors.

This algorithm was not particularly helpful in our project, as our dataset is imbalanced and possibly overfit the model.

- high accuracy could be explained by the fact that the clusters themselves were very similar, with the data points within each cluster being tightly packed together.



Hard vs. Soft Clustering

Hard Clustering (K-means/KNN): each point belongs to one centroid or cluster.

Soft Clustering (GMM): each point belongs to multiple clusters with weights since it is a distribution.

What does this mean?

We used hard clustering and soft clustering methods on the off chance that the points would be close enough to be closely related to multiple clusters. However, it was unnecessary. In a scenario such as belonging to playlists based on genre, it would be useful to use soft clustering for multi-labeling tasks.

04 Final Model



Evaluation:

With no ground truth label and with an unsupervised method such as clustering, an empirical evaluation is difficult to depict. For the Gaussian Mixture model, the silhouette score was used as a mathematical evaluation of its performance.

Fill in

At the end of the day we have to answer the question ourselves, did the clustered songs make sense and were they enjoyable? Could we use this to make a playlist we would listen to? This is why we tested ourselves by picking a few songs from a cluster made by each method to listen to and decide.




Thoughts: Both playlists from the GMM model sound pretty good, they don't fit into a genre or central idea but they encapsulate a good group of songs.

GMM playlists:




cluster 0

cluster 1




GMM Cluster 1

1		simple times Kacey Musgraves
2		Voy A Conquistarte Diego Verdaguer
3		WUSYANAME (feat. Young... Tyler, The Creator, Youn...

K-Means Cluster 1

1		Slow Down Summer Thomas Rhett
2		Qué de Raro Tiene Vicente Fernández
3		Laugh Now Cry Later (feat. Lil... Drake, Lil Durk

K-Means Cluster 2

1		Straightenin E Migos
2		I AM WOMAN E Emmy Meli
3		Nos Comemos (feat. Ozuna) E Tiago PZK, Ozuna

K-Means Cluster 3

1		MÍA Danna Paola
2		La Zona E Bad Bunny
3		Family Affair Mary J. Blige

Where do we go next?

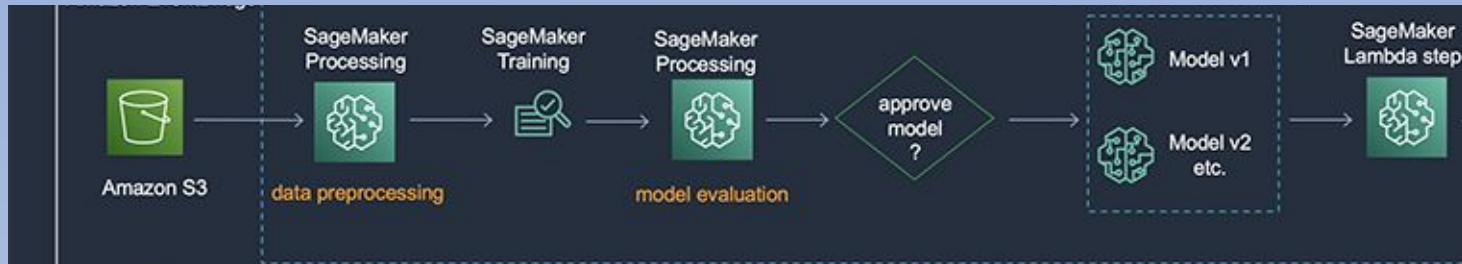
- Deploy it on the cloud
- Create a pipeline to feed new data

Repeat the cycle as time goes on

What new questions can we ask?

Mainly it begs two questions

- Would user interaction data make better clustering decisions?
- Could we make ground truth labels? Such as songs originally belonging to a playlist or based on genre.



05

Most Important Factors



Conclusion



It is hard to quantify the effectiveness of a model with no ground truth labels but it comes down to, do we like the results from the model?

Yes, the models are not horribly bad and the small playlists we created from them were enjoyable. There is much to be done to improve the cohesiveness of the song selection but for a first try it was a success.