# Spotify Music Recommendation Model


By: Yasmin Abu Helal

Aashrita Pillutla

Adeena Syed

Parto Aflatounian

# Project Overview

Problem we Chose to Study:

- Addressing the overload of choice in the current day music industry by improving music recommendation algorithms and increasing engagement and satisfaction (Schwartz, 2004)
- Increasing user retention though music recommendation by providing accurate and specific recommendations to a user's personal taste
- Providing an algorithm that can help support and recommend emerging and diverse artists to combat current recommendation systems which favor popular artists

Benefits:

- Economic benefits
  - Personalization in recommendations leads to 40% increased revenue generation (McKinsey & Co, 2020)
  - Increased engagements drives sales and streaming of new music for a variety of artists and increases effectiveness of advertising
- Social benefits
  - Streaming platform users experience increased satisfaction with the platform they subscribe to
  - By promoting a wider range of music, the algorithm encourages cultural diversity in music consumption. It breaks down barriers between different communities, fostering a more inclusive cultural landscape.

Parto

# Dataset

https://www.kaggle.com/datasets/theoverman/the-spotify-hit-predictor-dataset?resource=download

```
Number of rows: 24698
Number of cols: 20
Column names and their data types:
danceability        float64
energy              float64
key                   int64
loudness            float64
mode                  int64
speechiness         float64
acousticness        float64
instrumentalness    float64
liveness            float64
valence             float64
tempo               float64
type                 object
id                   object
uri                  object
track_href           object
analysis_url         object
duration_ms           int64
time_signature        int64
track                object
artist               object
dtype: object

Sample Rows:
   danceability  energy  key  loudness  ...  duration_ms  time_signature                    track             artist
0         0.509   0.277    6   -14.323  ...       161893               4            Walking Blues  Big Joe Williams
1         0.716   0.753    2    -5.682  ...       222000               4    Suddenly Last Summer        The Motels
2         0.360   0.542    5   -13.885  ...       444907               4                Sanctuary        Béla Fleck
3         0.656   0.512    7   -11.872  ...       157893               3           The Wild Rover        The Pogues
4         0.642   0.889    2    -5.620  ...       162293               4  In The Driver's Seat    John Schneider

[5 rows x 20 columns]
```

- Obtained from Kaggle
- Contains Spotify tracks from 1980 - 2019 & their audio features
  - Features: danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo, duration_ms, time_signature
- Features help us capture what someone might like in a song, making the dataset well-suited for training a music recommendation model.

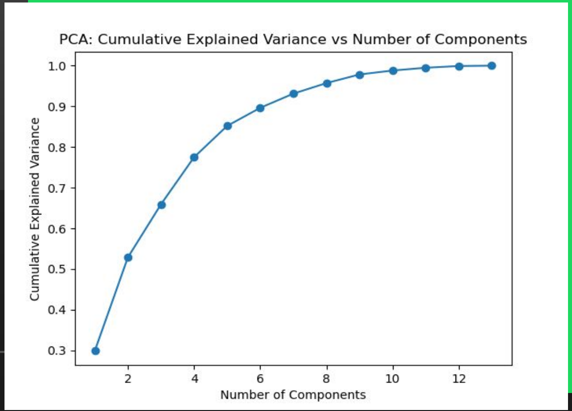
# Original Model

https://github.com/AnitaSoroush/MusicRecommendationSystem

1. The model takes the training set and cleans it by removing irrelevant columns, dropping rows with missing information, and removing duplicate rows.
2. The model transforms the training set by normalizing it/scaling it between 0 and 1.
3. The model applies k-means clustering on the training set using 10 clusters (the training set is segmented into 10 clusters based on the similarity of music features).
4. The input data (user playlist) is taken and normalized/scaled with the same scaler that was used on the training data.
5. Music recommendations are generated by matching each row of the input data to a cluster and then randomly recommending 5 tracks from that cluster.

This is an unsupervised ML model because we are not training it with inputs that are matched to outputs. Instead we are training it by assigning each track to a cluster based on its similarity to other tracks, which is determined by its features. In the context of our model, the “class” or “label” can be thought of as the cluster that a track is assigned to. These clusters do not exist in our dataset but are instead generated by the k-means clustering method, therefore making this model unsupervised.

# Refined Model

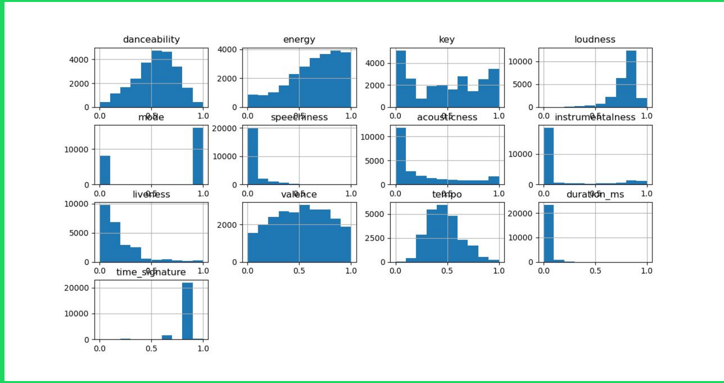



To improve the model we:

1. Cleaned the training set prior to providing it to the model so that no data cleaning needed to take place.
2. Implemented Principal Component Analysis (PCA). We fit a PCA model to the training set to determine the principal components. We generated a plot comparing the cumulative explained variance to the principal components and identified (7, 0.9) as the elbow point. We fit a new PCA model to the training set with n_components equal to 0.9 to retain 90% of the training set’s variance or 7 components. In simple terms, we modified the training set to retain the important patterns of the data with fewer variables.
3. Used the optimal value for the number of clusters. In the original model, the number of clusters was hardcoded to 10. We applied the knee locator method to find the elbow point in the plot of the sum of squared errors (SSE) against the number of clusters. This point provides a good balance between the minimization of the SSE and the number of clusters, and is therefore the optimal number of clusters for the model to use.

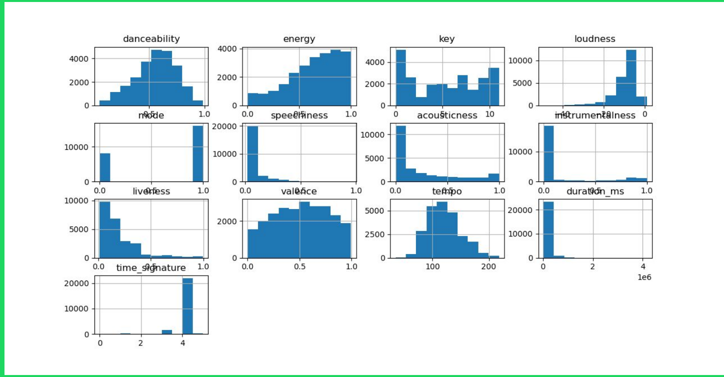
Parto

Data After Original Model’s Transformation

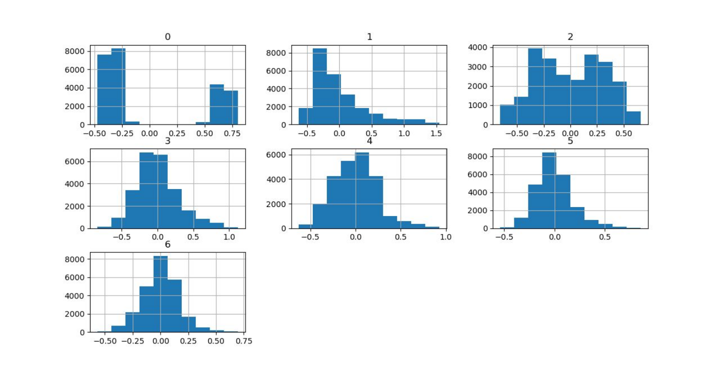



# Data Transformation

Data Before Transformation




Data After Refined Model’s Transformation




## The difference?

The original model only scales the features. The refined model creates 13 components (linear combinations of the features) from the 13 features and keeps the 7 most important ones/the ones that retain 90% of the training set’s variance, in addition to scaling the features.

Yasmin

# Testing Accuracy of Original Model vs Refined Model

- To compare the accuracy of the original model and the refined model, we created playlists containing 5 of our favourite songs. These playlists were the test sets for the models. The playlists the models generated were presented to us unlabelled to eliminate bias. We listened to both playlists and rated them on a scale of 1-10 depending on how well it aligned with our music taste.

- We also compared the SSE vs Num Clusters plot to see which model had a lower SSE value per cluster (this would indicate a more accurate model).

Home
Search
Playlists
Create Playlist
Liked Tracks
Results

Aashrita

# Results - Aash

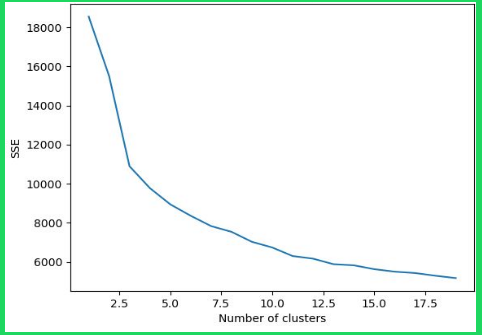
**Original model playlist rating:** 5/10

**Refined model playlist rating:** 7.5/10

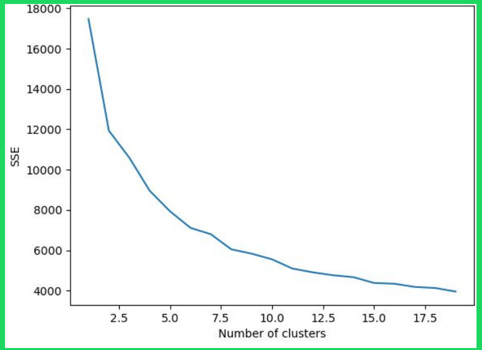
**Comments:** The original model's playlist has more songs that I know but I like the songs in the refined model's playlist better. The original model's results leaned toward matching artists/songs in popularity, while the refined model provided songs which I enjoyed the sound of more.

From the SSE vs Num Clusters curves we can see that the accuracy for the refined model is better. The refined model has a lower SSE value for every number of clusters compared to the original model.

## Original Model




## Refined Model



Home

Search

Playlists

Create Playlist

Liked Tracks

Results

Adeena

# Results - Adeena

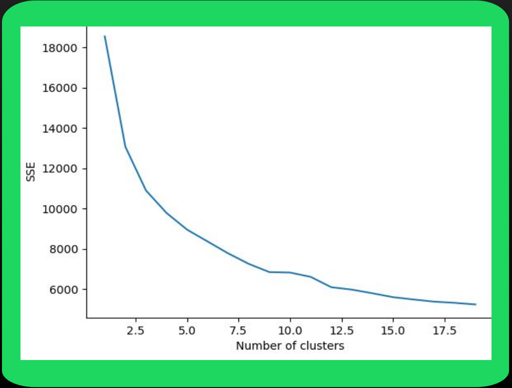
**Original model playlist rating:** 5.5/10

**Refined model playlist rating:** 8/10

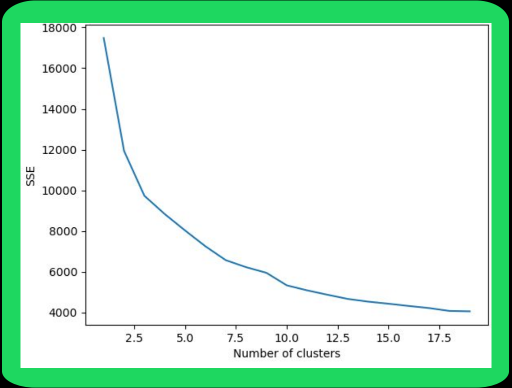
**Comments:** I noticed the same thing as Aash, the original model's playlist had more tracks and artists that I recognized, but the refined model's playlist was way more aligned with my music taste.

Once again, the SSE vs Num Clusters curves shows that the accuracy for the refined model is better (the refined model has a lower SSE value for every number of clusters compared to the original model).

## Original Model




## Refined Model



Home

Search

Playlists

Create Playlist

Liked Tracks

Results

Yasmin

# Results – Yasmin

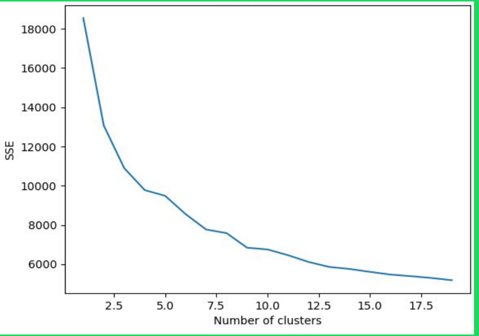
**Original model playlist rating:** 3/10

**Refined model playlist rating:** 5/10

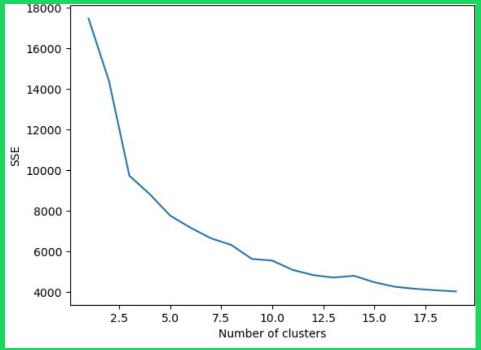
**Comments:** Unlike Aash and Adeena, I actually didn't think either of my playlists were that aligned to my music taste. With that being said, the refined model was the closest to my music taste and songs that I would listen to on a daily basis.

Once again, the SSE vs Num Clusters curves shows that the accuracy for the refined model is better (the refined model has a lower SSE value for every number of clusters compared to the original model).

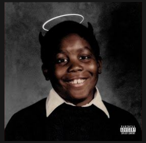
## Original Model




## Refined Model



# Thank You!



SCIENTISTS & ENGINEERS

Killer Mike, André 3000,
Future, & Eryn Allen Kane

0:23 -3:25

# Bibliography


Kaggle - The Spotify Hit Predictor Dataset (1960-2019), Farooq Ansari, https://www.kaggle.com/datasets/theoverman/the-spotify-hit-predictor-dataset?resource=download, (only 1980 - 2010 datasets were used)

GitHub - MusicRecommendationSystem, Anita Saroush, https://github.com/AnitaSoroush/MusicRecommendationSystem

The Paradox of Choice: Why More Is Less, Barry Schwartz, 2004

The Value of Getting Personalization Right—or Wrong—is Multiplying, McKinsey & Company, 2020, https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/the-value-of-getting-personalization-right-or-wrong-is-multiplying