Music Recommendation System

Presented by :

Rahul V Patil
Radhika S Joshi

Objective

- In this project, we perform clustering on Spotify dataset and develop a recommendation system for the users.
- The dataset includes songs and the features related to them. Features like danceability, energy, etc.
- We perform clustering on the dataset and recommend songs to the users based on clusters.
- This project makes use of machine learning model.

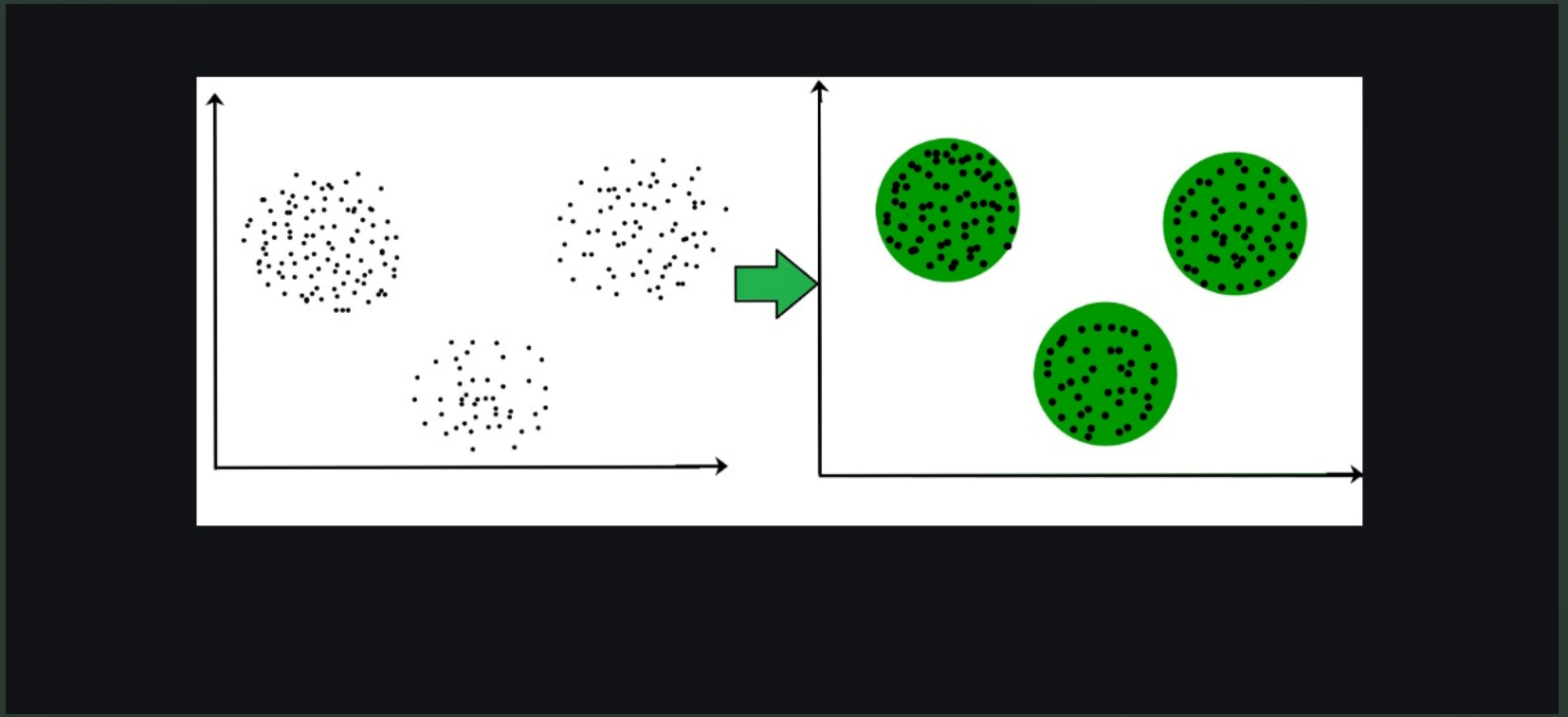
Data Description

- At first the given data contains 42305 rows and 22 columns, where some of the columns are as follows :
- Acousticness : A confidence measure from 0-1, 1 representing high acousticness.
- Danceability : Measuring from 0-1, it suggests how suitable a song is to dance to.
- Duration : Duration of the track in ms.
- Few other columns include – id, energy, key, liveness, tempo, speech

What is clustering?

- It is basically a type of *unsupervised learning method*. An unsupervised learning method is a method in which we draw references from datasets consisting of input data without labeled responses. Generally, it is used as a process to find meaningful structure, explanatory underlying processes, generative features, and groupings inherent in a set of examples.
- **Clustering** is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

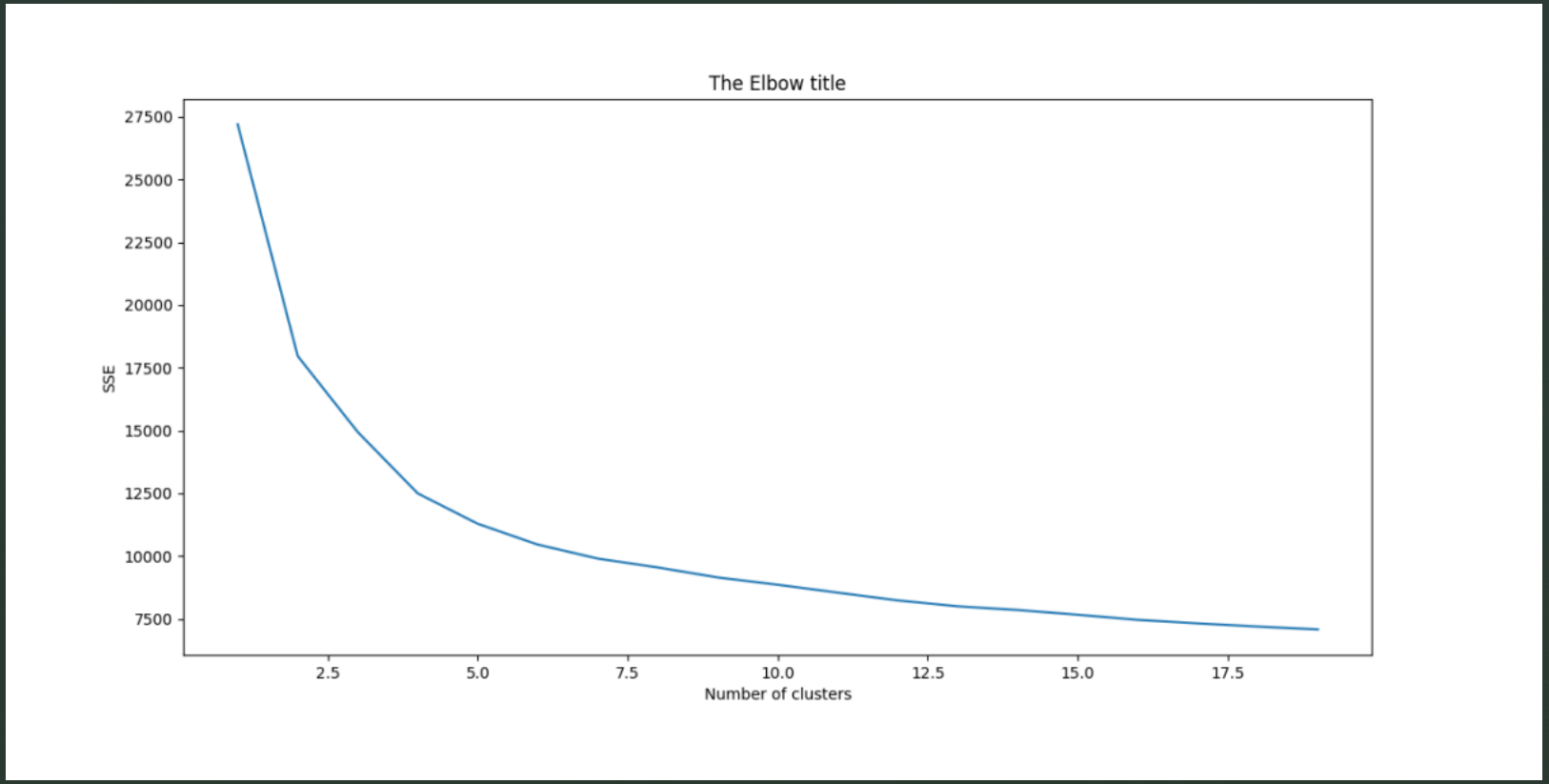
Clustering Example



K-means Algorithm

- K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science.
- K-Means Clustering is an algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.
- It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.

Elbow Method



What are the steps in K-means algorithm?

- Step-1: Select the number K to decide the number of clusters.
- Step-2: Select random K points or centroids. (It can be other from the input dataset).
- Step-3: Assign each data point to their closest centroid, which will form the predefined K clusters.
- Step-4: Calculate the variance and place a new centroid of each cluster.

What are the steps in K-means algorithm?

- Step-5: Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.
- Step-6: If any reassignment occurs, then go to step-4 else go to FINISH.
- Step-7: The model is ready.

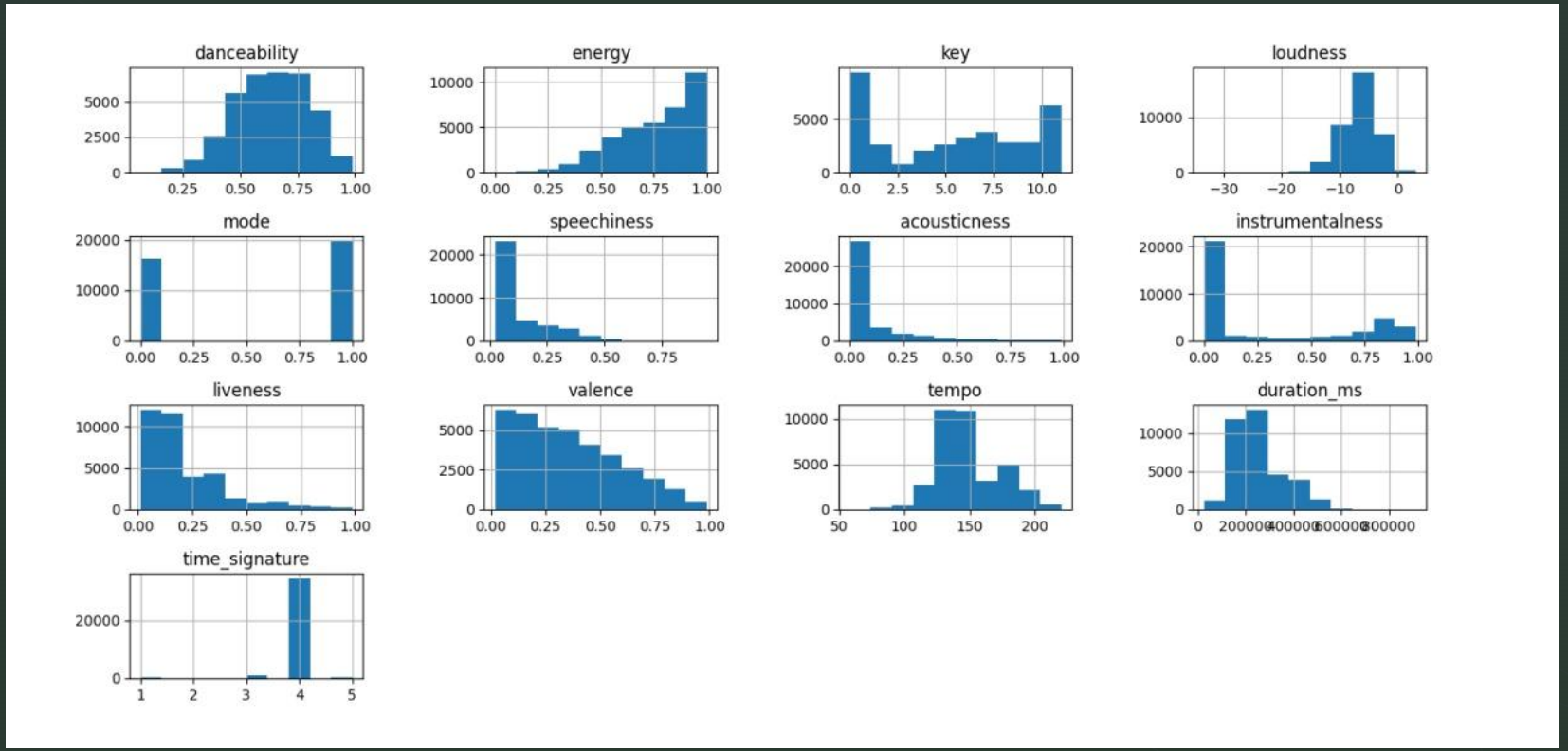
Data preprocessing – Data Cleaning

- Removing unnecessary columns – columns like ‘type’, ‘url’, ‘genre’ are not useful so we drop all such columns.
- Handling missing data – Columns consist of NaN values, so we drop these columns too.
- We also remove all the duplicates.

Data preprocessing – Data Transformation

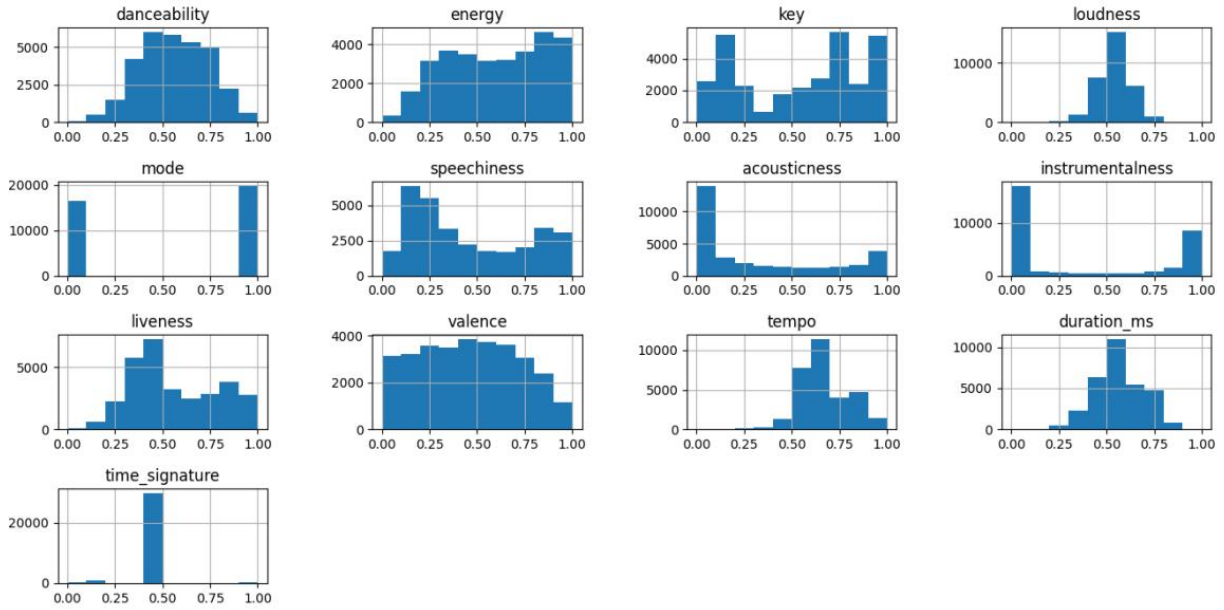
- Data transformation is the process of converting data from one format to another, typically from the format of a source system into the required format of a destination system.
- Here we convert data to visualise it better using histogram plots.
- If we want to use K-means, the distribution should be symmetric.
- Most of the features have a skewed distribution. So we should normalize them.
- Normalization is a technique often applied as part of data preparation for machine learning. The goal of normalization is to change the values of numeric columns in the dataset to use a common scale, without distorting differences in the ranges of values or losing information.

Data Transformation Example



Normalization

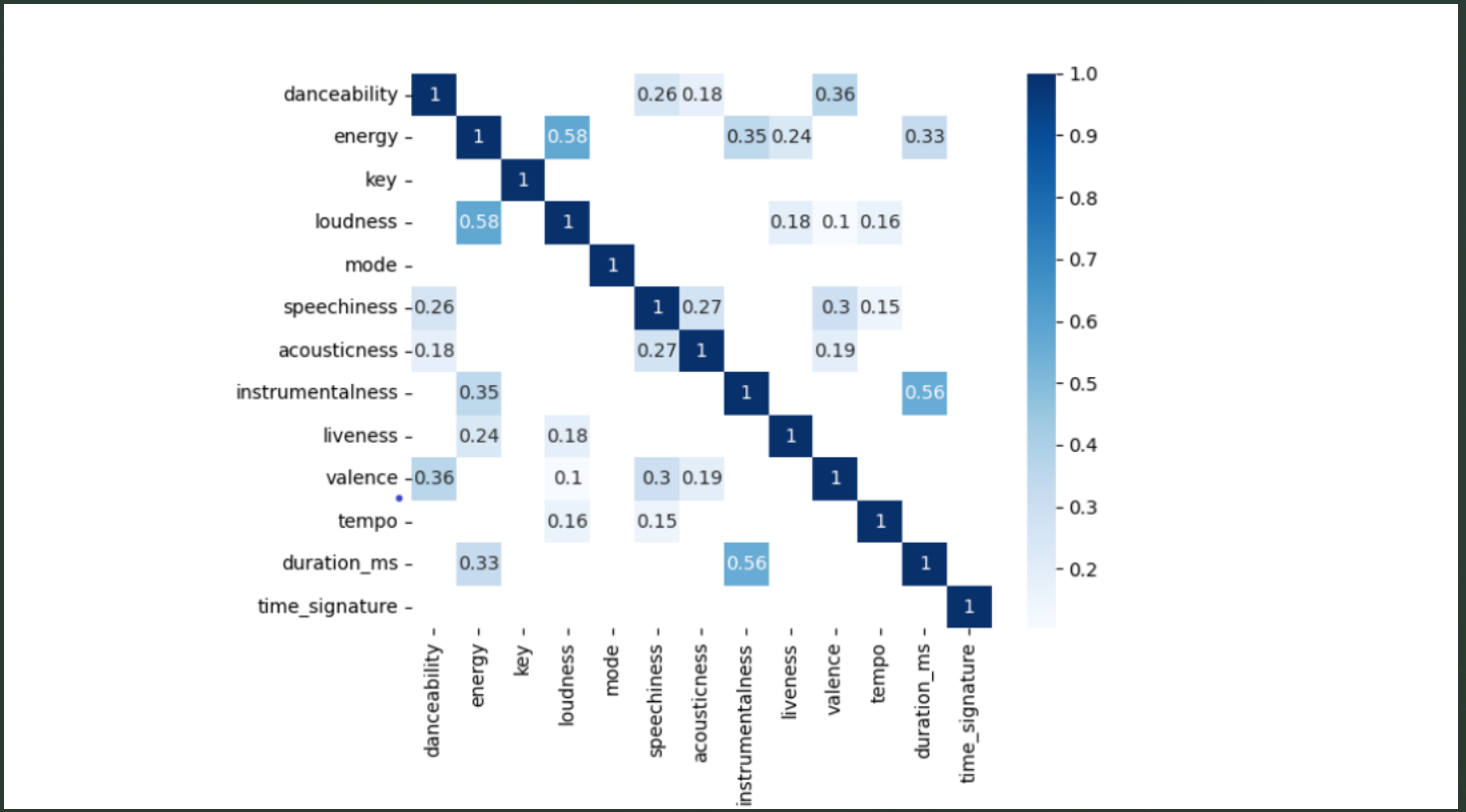
After normalizing the result is as follow:



Data Reduction

- Before moving on to clustering let’s just make sure about another thing: For a better result, we should find highly correlated variables, and if there are any, we should drop them.
- Data reduction is a process that reduces the volume of original data and represents it in a much smaller volume. Data reduction techniques are used to obtain a reduced representation of the dataset that is much smaller in volume by maintaining the integrity of the original data.

Data Reduction Example



Output

- Output contains 6 .csv files which are as follow :
- single_playlist.csv
- We matched each row of the user preference dataset to a cluster, and then randomly recommended one of the songs from the same cluster. The cluster playlists are :
- pl1.csv , pl2.csv , pl3.csv , pl4.csv , pl5.csv
- For pl[i].csv, we matched i’th row of the user preference dataset to a cluster, and then randomly recommended 5 songs from that cluster.

Functions

- `raw_data.isnull().sum()` – identifying null values.
- `raw_data.drop(['type', 'uri', 'track_href', 'analysis_url','song_name', 'Unnamed: 0', 'title', 'genre'], axis=1, inplace=False)`
- `MinMaxScaler()`
- `training_data.hist(),plt.show()`
- `corr = training_data.corr()`
- `sns.heatmap(corr[corr > 0.1], cmap="Blues", annot=True),plt.show()`
- `kmeans.fit(training_data.drop(['id'], axis=1, inplace=False))`
- `single_playlist.append((cluster_songs.sample()).values.flatten().tolist())`
- Main function to be called when the script is run from the command line. This function will recommend songs based on the user's input and save the playlist to a csv file.

Summary

- A music recommendation system was developed that can learn users' preferences. The system can classify a wide range of stored music using automatic music content analyses. Users can opt for music according to their mood, using such words as "bright", "exciting", "quiet", "sad" and "healing".
- The machine learning model, takes the data from the single_playlist csv file which is already created by the user on the Spotify platform. Later we use k-means algorithm for clustering the type of songs user might be interested in, and the machine recommends the songs based on these formed clusters at random.
- The clusters are the 5 playlists mentioned in the previous slide.

Thank You.