Lab 1: Report

Ameya Zope

Feb 15, 2024

Lab Description

In this lab, I have created a dashboard to vizualize data on ScreePlot, PCA based Biplot K Means MSE Plot and finally the Scatterplot Matrix (SPLOM). One goal of this lab is to understand and get more hands on experience with PCA, KMeans and dimensionality reduction. Another gols of this lab is to simulate a real life dashboard wherein the heavy computation (like PCA and KMeans) is done by a backend server(python in this case) and served to the frontend server which then displays the data for visualization. The frontend server in this case is a React based server which utilizes d3 for visualization and creation of graphs. Many interactivity elements and animations have been added to the frontend to enable the data scientist to draw observations like correlation in the data. This dashboard has been made in accordance with the lab-2 for CSE 564 : Visualization & Visual Analytics course at Stony Brook University.

Dataset

I have used the same dataset that was used for Lab 1. The dataset comprises numerical features from Spotify tracks, focusing on attributes that quantify different aspects of the music. These attributes include the total number of streams, danceability, valence (positivity), energy, acousticness, instrumentalness, liveness, and speechiness percentages. Each feature plays a vital role in characterizing the musical pieces, from their suitability for dancing to the presence of acoustic elements and live performance aspects.

Link to Dataset

https://www.kaggle.com/datasets/nelgiriyewithana/top-spotify-songs-2023

How to Run

There are two servers created for this project.

1. **Python based Flask Backend Server** - This server is used to do the heavy calculations like PCA , KMeans, etc. This server exposes several APIs that server to provide the frontend with the data that is needed for rendering the plots.

2. **React based Frontend Server** - This server is mainly used to serve the UI which includes the Scree Plot, the PCA Based Biplot, the K Means MSE Plot, the Scatterplot Matrix. This server fetches all the data that it needs to server from the flask server.

Run the python server

1. Run the below commands

```
cd ./backend
pip3 install -r requirements.txt
python3 server.py
```

Preferably use python3.9 for running the above server because development and testing has been done on python3.9

Run the frontend server

1. Run the below commands

```
npm install
npm start
```

Interesting Observations

The top four attributes were identified based on their squared sum of PCA loadings, which signifies their contribution to the variance explained by the principal components. These attributes were key in understanding the dataset's underlying structure and patterns.

The Biplot visualizations between PC1 and PC2 revealed significant contributions to data clustering, illustrating the distinction between the components in terms of their variance capture. However, further components (PC3, PC4, PC5) showed diminishing returns in terms of clarity in clustering, reflecting their lesser significance. Infact, if we plot PC6 and PC7 on the biplot we can see that the clustering goes haywire. On the other hand with PC1 and PC2 on the biplot, we can see clearly separated clusters.

A quick look at the biplot suggests that energy and acousticness attributes have a strong, inverse correlation and significant magnitudes, indicating that these features are crucial in the clustering process.

The k-means clustering analysis, visualized through an MSE plot, identified the optimal number of clusters at the elbow point, where the clustering quality was maximized. Deviating from this optimal number resulted in less clear clustering, affirming the selected k-value's effectiveness.