Project Check-in 5

```
%pip install --upgrade pip
%pip install scikit-lego
%pip install seaborn
%pip install nbstripout
!nbstripout --install

Requirement already satisfied: pip in
/opt/anaconda3/lib/python3.11/site-packages (24.3.1)
Note: you may need to restart the kernel to use updated packages.
Requirement already satisfied: scikit-lego in
/opt/anaconda3/lib/python3.11/site-packages (0.9.1)
Requirement already satisfied: narwhals>=1.0.0 in
/opt/anaconda3/lib/python3.11/site-packages (from scikit-lego) (1.9.4)
Requirement already satisfied: pandas>=1.1.5 in
/opt/anaconda3/lib/python3.11/site-packages (from scikit-lego) (2.1.4)
Requirement already satisfied: scikit-learn>=1.0 in
/opt/anaconda3/lib/python3.11/site-packages (from scikit-lego) (1.2.2)
Requirement already satisfied: numpy<2,>=1.23.2 in
/opt/anaconda3/lib/python3.11/site-packages (from pandas>=1.1.5-
>scikit-lego) (1.26.4)
Requirement already satisfied: python-dateutil>=2.8.2 in
/opt/anaconda3/lib/python3.11/site-packages (from pandas>=1.1.5-
>scikit-lego) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in
/opt/anaconda3/lib/python3.11/site-packages (from pandas>=1.1.5-
>scikit-lego) (2023.3.post1)
Requirement already satisfied: tzdata>=2022.1 in
/opt/anaconda3/lib/python3.11/site-packages (from pandas>=1.1.5-
>scikit-lego) (2023.3)
Requirement already satisfied: scipy>=1.3.2 in
/opt/anaconda3/lib/python3.11/site-packages (from scikit-learn>=1.0-
>scikit-lego) (1.11.4)
Requirement already satisfied: joblib>=1.1.1 in
/opt/anaconda3/lib/python3.11/site-packages (from scikit-learn>=1.0-
>scikit-lego) (1.2.0)
Requirement already satisfied: threadpoolctl>=2.0.0 in
/opt/anaconda3/lib/python3.11/site-packages (from scikit-learn>=1.0-
>scikit-lego) (2.2.0)
Requirement already satisfied: six>=1.5 in
/opt/anaconda3/lib/python3.11/site-packages (from python-
dateutil>=2.8.2->pandas>=1.1.5->scikit-lego) (1.16.0)
Note: you may need to restart the kernel to use updated packages.
Requirement already satisfied: seaborn in
/opt/anaconda3/lib/python3.11/site-packages (0.13.2)
Requirement already satisfied: numpy!=1.24.0,>=1.20 in
/opt/anaconda3/lib/python3.11/site-packages (from seaborn) (1.26.4)
Requirement already satisfied: pandas>=1.2 in
```

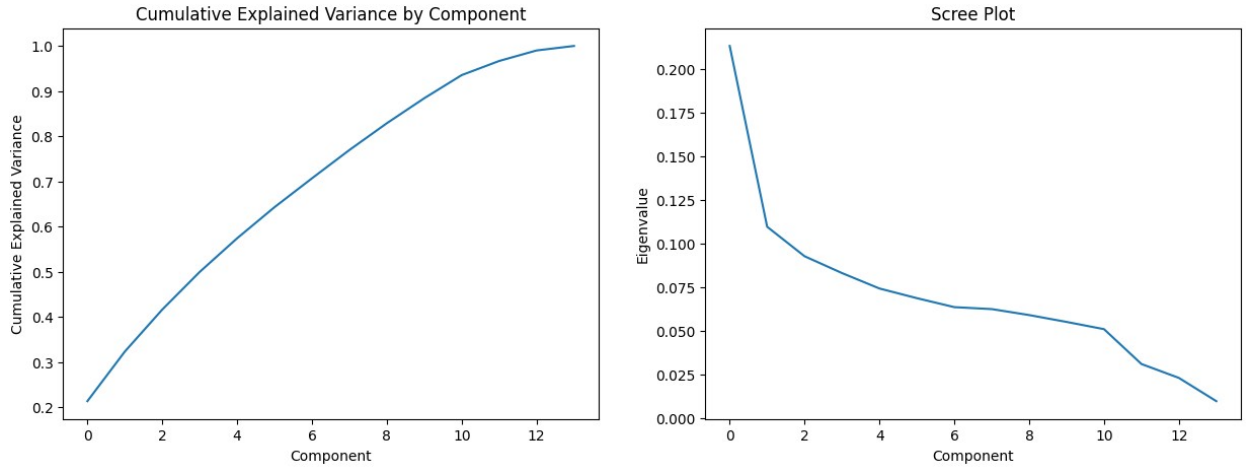
```
/opt/anaconda3/lib/python3.11/site-packages (from seaborn) (2.1.4)
Requirement already satisfied: matplotlib!=3.6.1,>=3.4 in
/opt/anaconda3/lib/python3.11/site-packages (from seaborn) (3.8.0)
Requirement already satisfied: contourpy>=1.0.1 in
/opt/anaconda3/lib/python3.11/site-packages (from matplotlib!
=3.6.1,>=3.4->seaborn) (1.2.0)
Requirement already satisfied: cycler>=0.10 in
/opt/anaconda3/lib/python3.11/site-packages (from matplotlib!
=3.6.1,>=3.4->seaborn) (0.11.0)
Requirement already satisfied: fonttools>=4.22.0 in
/opt/anaconda3/lib/python3.11/site-packages (from matplotlib!
=3.6.1,>=3.4->seaborn) (4.25.0)
Requirement already satisfied: kiwisolver>=1.0.1 in
/opt/anaconda3/lib/python3.11/site-packages (from matplotlib!
=3.6.1,>=3.4->seaborn) (1.4.4)
Requirement already satisfied: packaging>=20.0 in
/opt/anaconda3/lib/python3.11/site-packages (from matplotlib!
=3.6.1,>=3.4->seaborn) (23.1)
Requirement already satisfied: pillow>=6.2.0 in
/opt/anaconda3/lib/python3.11/site-packages (from matplotlib!
=3.6.1,>=3.4->seaborn) (10.2.0)
Requirement already satisfied: pyparsing>=2.3.1 in
/opt/anaconda3/lib/python3.11/site-packages (from matplotlib!
=3.6.1,>=3.4->seaborn) (3.0.9)
Requirement already satisfied: python-dateutil>=2.7 in
/opt/anaconda3/lib/python3.11/site-packages (from matplotlib!
=3.6.1,>=3.4->seaborn) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in
/opt/anaconda3/lib/python3.11/site-packages (from pandas>=1.2-
>seaborn) (2023.3.post1)
Requirement already satisfied: tzdata>=2022.1 in
/opt/anaconda3/lib/python3.11/site-packages (from pandas>=1.2-
>seaborn) (2023.3)
Requirement already satisfied: six>=1.5 in
/opt/anaconda3/lib/python3.11/site-packages (from python-
dateutil>=2.7->matplotlib!=3.6.1,>=3.4->seaborn) (1.16.0)
Note: you may need to restart the kernel to use updated packages.
Requirement already satisfied: nbstripout in
/opt/anaconda3/lib/python3.11/site-packages (0.7.1)
Requirement already satisfied: nbformat in
/opt/anaconda3/lib/python3.11/site-packages (from nbstripout) (5.9.2)
Requirement already satisfied: fastjsonschema in
/opt/anaconda3/lib/python3.11/site-packages (from nbformat-
>nbstripout) (2.16.2)
Requirement already satisfied: jsonschema>=2.6 in
/opt/anaconda3/lib/python3.11/site-packages (from nbformat-
>nbstripout) (4.19.2)
Requirement already satisfied: jupyter-core in
/opt/anaconda3/lib/python3.11/site-packages (from nbformat-
```

```
>nbstripout) (5.5.0)
Requirement already satisfied: traitlets>=5.1 in
/opt/anaconda3/lib/python3.11/site-packages (from nbformat-
>nbstripout) (5.7.1)
Requirement already satisfied: attrs>=22.2.0 in
/opt/anaconda3/lib/python3.11/site-packages (from jsonschema>=2.6-
>nbformat->nbstripout) (23.1.0)
Requirement already satisfied: jsonschema-specifications>=2023.03.6 in
/opt/anaconda3/lib/python3.11/site-packages (from jsonschema>=2.6-
>nbformat->nbstripout) (2023.7.1)
Requirement already satisfied: referencing>=0.28.4 in
/opt/anaconda3/lib/python3.11/site-packages (from jsonschema>=2.6-
>nbformat->nbstripout) (0.30.2)
Requirement already satisfied: rpds-py>=0.7.1 in
/opt/anaconda3/lib/python3.11/site-packages (from jsonschema>=2.6-
>nbformat->nbstripout) (0.10.6)
Requirement already satisfied: platformdirs>=2.5 in
/opt/anaconda3/lib/python3.11/site-packages (from jupyter-core-
>nbformat->nbstripout) (3.10.0)
Note: you may need to restart the kernel to use updated packages.
fatal: --local can only be used inside a git repository
Installation failed: not a git repository!

import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
from sklearn.preprocessing import StandardScaler
import pandas as pd
from sklearn.decomposition import PCA, TruncatedSVD
df = pd.read_csv("./dataset.csv")
```

1. We chose to run PCA on our data.
2. See below
3. N/A

```
# Step 1: Clean Data
# Remove duplicates
df_cleaned = df.drop(columns='Unnamed:
0').drop_duplicates(subset=['track_id','album_name','artists','track_n
ame'])

# Remove columns with every row unique. Also dropping artist and album
because it would be too much one-hot encoding
df_cleaned.drop(columns=['track_id', 'track_name',
'artists','album_name'], inplace=True)
df_cleaned.dropna(axis=0,inplace=True)
```

```
#The columns with object datatype will be categorical
columns = df_cleaned.select_dtypes(include=['int64',
'float64']).columns.tolist()
df_cleaned = df_cleaned[columns]

scaler = StandardScaler() # Scale the data so that the variances for
each feature can be similarly weighted
df_cleaned = scaler.fit_transform(df_cleaned)
df_cleaned = pd.DataFrame(df_cleaned, columns=columns)

pca = PCA(n_components=14) # 14 principal components for 14 features
(don't have to use them all)
transformed_data = pca.fit_transform(df_cleaned)
eigenvalues = pca.explained_variance_ratio_
cumulative_explained_variance = eigenvalues.cumsum()

fig, ax = plt.subplots(1,2, figsize=(15,5))

# For both graphs, the principal component number starts at 0 not 1

# Plot the cumulative explained variance
ax[0].plot(cumulative_explained_variance)
ax[0].set_xlabel("Component")
ax[0].set_ylabel("Cumulative Explained Variance")
ax[0].set_title("Cumulative Explained Variance by Component")

# Plot the eigenvalues to create the scree plot
ax[1].plot(eigenvalues)
ax[1].set_xlabel("Component")
ax[1].set_ylabel("Eigenvalue")
ax[1].set_title("Scree Plot")

Text(0.5, 1.0, 'Scree Plot')
```



1. As our dataset has 14 numerical features, we are trying to see if we can use principal component analysis as a dimensionality reduction technique, in order to reduce the complexity of our data. Judging from the cumulative explained variance graph, it seems like we would still need a decent number of components to accurately describe our data. In future iterations, we were considering clustering prior to PCA, so that we could graph our data on the first 2 principal components and see if the clusters spread out, however, it is unlikely that plotting the first two principal components will capture a lot of the variability in the data, considering that they only make up around 30% of the variance. Given this, the clusters may not be easily separated on the first two components, and will likely heavily overlap.