Project Check-in 1

1. We chose the Spotify Prediction dataset.
2. We are studying how different aspects of tracks such as loudness, instrumentalness, tempo, and genre individually predict popularity, and how they combine to predict popularity.
3. See below.
4. See below.

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

```
Requirement already satisfied: scikit-lego in c:\users\isaac\appdata\
local\programs\python\python311\lib\site-packages (0.9.1)
Requirement already satisfied: narwhals>=1.0.0 in c:\users\isaac\
appdata\local\programs\python\python311\lib\site-packages (from
scikit-lego) (1.9.3)
Requirement already satisfied: pandas>=1.1.5 in c:\users\isaac\
appdata\local\programs\python\python311\lib\site-packages (from
scikit-lego) (2.1.2)
Requirement already satisfied: scikit-learn>=1.0 in c:\users\isaac\
appdata\local\programs\python\python311\lib\site-packages (from
scikit-lego) (1.3.2)
Requirement already satisfied: numpy<2,>=1.23.2 in c:\users\isaac\
appdata\local\programs\python\python311\lib\site-packages (from
pandas>=1.1.5->scikit-lego) (1.26.1)
Requirement already satisfied: python-dateutil>=2.8.2 in c:\users\
isaac\appdata\roaming\python\python311\site-packages (from
pandas>=1.1.5->scikit-lego) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in c:\users\isaac\appdata\
local\programs\python\python311\lib\site-packages (from pandas>=1.1.5-
>scikit-lego) (2023.3.post1)
Requirement already satisfied: tzdata>=2022.1 in c:\users\isaac\
appdata\local\programs\python\python311\lib\site-packages (from
pandas>=1.1.5->scikit-lego) (2023.3)
Requirement already satisfied: scipy>=1.5.0 in c:\users\isaac\appdata\
local\programs\python\python311\lib\site-packages (from scikit-
learn>=1.0->scikit-lego) (1.11.3)
Requirement already satisfied: joblib>=1.1.1 in c:\users\isaac\
appdata\local\programs\python\python311\lib\site-packages (from
scikit-learn>=1.0->scikit-lego) (1.3.2)
Requirement already satisfied: threadpoolctl>=2.0.0 in c:\users\isaac\
appdata\local\programs\python\python311\lib\site-packages (from
scikit-learn>=1.0->scikit-lego) (3.2.0)
Requirement already satisfied: six>=1.5 in c:\users\isaac\appdata\
roaming\python\python311\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.


[notice] A new release of pip available: 22.3 -> 24.2
[notice] To update, run: python.exe -m pip install --upgrade pip

Requirement already satisfied: seaborn in c:\users\isaac\appdata\
local\programs\python\python311\lib\site-packages (0.13.2)
Requirement already satisfied: numpy!=1.24.0,>=1.20 in c:\users\isaac\
appdata\local\programs\python\python311\lib\site-packages (from
seaborn) (1.26.1)
Requirement already satisfied: pandas>=1.2 in c:\users\isaac\appdata\
local\programs\python\python311\lib\site-packages (from seaborn)
(2.1.2)
Requirement already satisfied: matplotlib!=3.6.1,>=3.4 in c:\users\
isaac\appdata\local\programs\python\python311\lib\site-packages (from
seaborn) (3.8.0)
Requirement already satisfied: contourpy>=1.0.1 in c:\users\isaac\
appdata\local\programs\python\python311\lib\site-packages (from
matplotlib!=3.6.1,>=3.4->seaborn) (1.1.1)
Requirement already satisfied: cycler>=0.10 in c:\users\isaac\appdata\
local\programs\python\python311\lib\site-packages (from matplotlib!
=3.6.1,>=3.4->seaborn) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\isaac\
appdata\local\programs\python\python311\lib\site-packages (from
matplotlib!=3.6.1,>=3.4->seaborn) (4.43.1)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\isaac\
appdata\local\programs\python\python311\lib\site-packages (from
matplotlib!=3.6.1,>=3.4->seaborn) (1.4.5)
Requirement already satisfied: packaging>=20.0 in c:\users\isaac\
appdata\roaming\python\python311\site-packages (from matplotlib!
=3.6.1,>=3.4->seaborn) (23.2)
Requirement already satisfied: pillow>=6.2.0 in c:\users\isaac\
appdata\local\programs\python\python311\lib\site-packages (from
matplotlib!=3.6.1,>=3.4->seaborn) (10.1.0)
Requirement already satisfied: pyparsing>=2.3.1 in c:\users\isaac\
appdata\local\programs\python\python311\lib\site-packages (from
matplotlib!=3.6.1,>=3.4->seaborn) (3.1.1)
Requirement already satisfied: python-dateutil>=2.7 in c:\users\isaac\
appdata\roaming\python\python311\site-packages (from matplotlib!
=3.6.1,>=3.4->seaborn) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in c:\users\isaac\appdata\
local\programs\python\python311\lib\site-packages (from pandas>=1.2-
>seaborn) (2023.3.post1)
Requirement already satisfied: tzdata>=2022.1 in c:\users\isaac\
appdata\local\programs\python\python311\lib\site-packages (from
pandas>=1.2->seaborn) (2023.3)
Requirement already satisfied: six>=1.5 in c:\users\isaac\appdata\
roaming\python\python311\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.


[notice] A new release of pip available: 22.3 -> 24.2
[notice] To update, run: python.exe -m pip install --upgrade pip

Collecting nbstripout
  Downloading nbstripout-0.7.1-py2.py3-none-any.whl (15 kB)
Collecting nbformat
  Downloading nbformat-5.10.4-py3-none-any.whl (78 kB)
     ---------------------------------------- 78.5/78.5 kB 4.3 MB/s
eta 0:00:00
Collecting fastjsonschema>=2.15
  Downloading fastjsonschema-2.20.0-py3-none-any.whl (23 kB)
Collecting jsonschema>=2.6
  Downloading jsonschema-4.23.0-py3-none-any.whl (88 kB)
     ---------------------------------------- 88.5/88.5 kB ? eta
0:00:00
Requirement already satisfied: jupyter-core!=5.0.*,>=4.12 in c:\users\
isaac\appdata\roaming\python\python311\site-packages (from nbformat-
>nbstripout) (5.4.0)
Requirement already satisfied: traitlets>=5.1 in c:\users\isaac\
appdata\roaming\python\python311\site-packages (from nbformat-
>nbstripout) (5.11.2)
Collecting attrs>=22.2.0
  Downloading attrs-24.2.0-py3-none-any.whl (63 kB)
     ---------------------------------------- 63.0/63.0 kB ? eta
0:00:00
Collecting jsonschema-specifications>=2023.03.6
  Downloading jsonschema_specifications-2024.10.1-py3-none-any.whl (18
kB)
Collecting referencing>=0.28.4
  Downloading referencing-0.35.1-py3-none-any.whl (26 kB)
Collecting rpds-py>=0.7.1
  Downloading rpds_py-0.20.0-cp311-none-win_amd64.whl (213 kB)
     -------------------------------------- 213.6/213.6 kB 6.6 MB/s
eta 0:00:00
Requirement already satisfied: platformdirs>=2.5 in c:\users\isaac\
appdata\roaming\python\python311\site-packages (from jupyter-core!
=5.0.*,>=4.12->nbformat->nbstripout) (3.11.0)
Requirement already satisfied: pywin32>=300 in c:\users\isaac\appdata\
roaming\python\python311\site-packages (from jupyter-core!
=5.0.*,>=4.12->nbformat->nbstripout) (306)
Installing collected packages: fastjsonschema, rpds-py, attrs,
referencing, jsonschema-specifications, jsonschema, nbformat,
nbstripout
Successfully installed attrs-24.2.0 fastjsonschema-2.20.0 jsonschema-
4.23.0 jsonschema-specifications-2024.10.1 nbformat-5.10.4 nbstripout-
```

```
0.7.1 referencing-0.35.1 rpds-py-0.20.0
Note: you may need to restart the kernel to use updated packages.


[notice] A new release of pip available: 22.3 -> 24.2
[notice] To update, run: python.exe -m pip install --upgrade pip
UsageError: Line magic function `%nbstripout` not found.
```

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.metrics import mean_squared_error, mean_absolute_error,
r2_score
from sklearn.linear_model import LinearRegression

df = pd.read_csv("./dataset.csv")

print(df.columns)
print("************************************")
print(df.head(10))
print("************************************")
print(df.tail(10))
print("************************************")
print(df.describe())
print("************************************")
print(df.info())
print(df.shape)
```

```
Index(['Unnamed: 0', 'track_id', 'artists', 'album_name',
'track_name',
       'popularity', 'duration_ms', 'explicit', 'danceability',
'energy',
       'key', 'loudness', 'mode', 'speechiness', 'acousticness',
       'instrumentalness', 'liveness', 'valence', 'tempo',
'time_signature',
       'track_genre'],
      dtype='object')
*************************************
   Unnamed: 0                track_id
artists  \
0           0  5SuOikwiRyPMVoIQDJUgSV                           Gen
Hoshino
1           1  4qPNDBW1i3p13qLCt0Ki3A                          Ben
Woodward
2           2  1iJBSr7s7jYXzM8EGcbK5b                Ingrid
Michaelson;ZAYN
3           3  6lfxq3CG4xtTiEg7opyCyx                          Kina
Grannis
```

```
4           4  5vjLSffimiIP26QG5WcN2K                      Chord
Overstreet
5           5  01MVOl9KtVTNfFiBU9I7dc                          Tyrone
Wells
6           6  6Vc5wAMmXdKIAM7WUoEb7N  A Great Big World;Christina
Aguilera
7           7  1EzrEOXmMH3G43AXT1y7pA                            Jason
Mraz
8           8  0IktbUcnAGrvD03AWnz3Q8             Jason Mraz;Colbie
Caillat
9           9  7k9GuJYLp2AzqokyEdwEw2                        Ross
Copperman

                                           album_name  \
0                                              Comedy
1                                    Ghost (Acoustic)
2                                      To Begin Again
3  Crazy Rich Asians (Original Motion Picture Sou...
4                                             Hold On
5                                 Days I Will Remember
6                          Is There Anybody Out There?
7                  We Sing. We Dance. We Steal Things.
8                  We Sing. We Dance. We Steal Things.
9                                              Hunger

                   track_name  popularity  duration_ms  explicit  \
0                      Comedy          73       230666     False
1            Ghost - Acoustic          55       149610     False
2              To Begin Again          57       210826     False
3  Can't Help Falling In Love          71       201933     False
4                     Hold On          82       198853     False
5        Days I Will Remember          58       214240     False
6              Say Something           74       229400     False
7                  I'm Yours           80       242946     False
8                      Lucky           74       189613     False
9                     Hunger           56       205594     False

   danceability  energy  ...  loudness  mode  speechiness
acousticness  \
0         0.676  0.4610  ...    -6.746     0       0.1430
0.0322
1         0.420  0.1660  ...   -17.235     1       0.0763
0.9240
2         0.438  0.3590  ...    -9.734     1       0.0557
0.2100
3         0.266  0.0596  ...   -18.515     1       0.0363
0.9050
4         0.618  0.4430  ...    -9.681     1       0.0526
0.4690
5         0.688  0.4810  ...    -8.807     1       0.1050
```

```
0.2890
6          0.407  0.1470  ...    -8.822     1       0.0355
0.8570
7          0.703  0.4440  ...    -9.331     1       0.0417
0.5590
8          0.625  0.4140  ...    -8.700     1       0.0369
0.2940
9          0.442  0.6320  ...    -6.770     1       0.0295
0.4260

   instrumentalness  liveness  valence    tempo  time_signature
track_genre
0          0.000001    0.3580   0.7150   87.917               4
acoustic
1          0.000006    0.1010   0.2670   77.489               4
acoustic
2          0.000000    0.1170   0.1200   76.332               4
acoustic
3          0.000071    0.1320   0.1430  181.740               3
acoustic
4          0.000000    0.0829   0.1670  119.949               4
acoustic
5          0.000000    0.1890   0.6660   98.017               4
acoustic
6          0.000003    0.0913   0.0765  141.284               3
acoustic
7          0.000000    0.0973   0.7120  150.960               4
acoustic
8          0.000000    0.1510   0.6690  130.088               4
acoustic
9          0.004190    0.0735   0.1960   78.899               4
acoustic

[10 rows x 21 columns]
*************************************
        Unnamed: 0                track_id           artists  \
113990      113990  2A4dSiJmbviL56CBupkh6C    Lucas Cervetti
113991      113991  0CE0Y6GM75cbrqao8EOAlW      Chris Tomlin
113992      113992  3FjOBB4EyIXHYUtSgrIdY9     Jesus Culture
113993      113993  4OkMK49i3NApR1KsAIsTf6      Chris Tomlin
113994      113994  4WbOUe6T0sozC7z5ZJgiAA    Lucas Cervetti
113995      113995  2C3TZjDRiAzdyViavDJ217     Rainy Lullaby
113996      113996  1hIz5L4IB9hN3WRYPOCGPw     Rainy Lullaby
113997      113997  6x8ZfSoqDjuNa5SVP5QjvX     Cesária Evora
113998      113998  2e6sXL2bYv4bSz6VTdnfLs  Michael W. Smith
113999      113999  2hETkH7cOfqmz3LqZDHZf5     Cesária Evora

                                             album_name  \
113990          Frecuencias Álmicas en 432hz (Solo Piano)
113991                             The Ultimate Playlist
```

```
113992                                      Revelation Songs
113993                     See The Morning (Special Edition)
113994                           Frecuencias Álmicas en 432hz
113995  #mindfulness - Soft Rain for Mindful Meditatio...
113996  #mindfulness - Soft Rain for Mindful Meditatio...
113997                                               Best Of
113998                                     Change Your World
113999                                        Miss Perfumado

                               track_name  popularity  duration_ms
explicit  \
113990  Frecuencia Álmica XI - Solo Piano          22       369049
False
113991        At The Cross (Love Ran Red)          32       250629
False
113992              Your Love Never Fails          38       312566
False
113993        How Can I Keep From Singing          39       256026
False
113994           Frecuencia Álmica, Pt. 4          22       305454
False
113995                Sleep My Little Boy          21       384999
False
113996                   Water Into Light          22       385000
False
113997                     Miss Perfumado          22       271466
False
113998                            Friends          41       283893
False
113999                          Barbincor          22       241826
False

        danceability  energy  ...  loudness  mode  speechiness
acousticness  \
113990         0.579   0.245  ...   -16.357     1       0.0384
0.97000
113991         0.387   0.531  ...    -4.788     1       0.0290
0.00305
113992         0.475   0.860  ...    -4.722     1       0.0421
0.00650
113993         0.505   0.687  ...    -4.375     1       0.0287
0.08410
113994         0.331   0.171  ...   -15.668     1       0.0350
0.92000
113995         0.172   0.235  ...   -16.393     1       0.0422
0.64000
113996         0.174   0.117  ...   -18.318     0       0.0401
0.99400
113997         0.629   0.329  ...   -10.895     0       0.0420
```

```
0.86700
113998         0.587   0.506  ...   -10.889     1       0.0297
0.38100
113999         0.526   0.487  ...   -10.204     0       0.0725
0.68100

        instrumentalness  liveness  valence    tempo
time_signature  \
113990          0.924000    0.1010   0.3020  112.011               3

113991          0.000000    0.2010   0.1530  146.003               4

113992          0.000002    0.2460   0.4270  113.949               4

113993          0.000000    0.1880   0.3820  104.083               3

113994          0.022900    0.0679   0.3270  132.147               3

113995          0.928000    0.0863   0.0339  125.995               5

113996          0.976000    0.1050   0.0350   85.239               4

113997          0.000000    0.0839   0.7430  132.378               4

113998          0.000000    0.2700   0.4130  135.960               4

113999          0.000000    0.0893   0.7080   79.198               4


        track_genre
113990  world-music
113991  world-music
113992  world-music
113993  world-music
113994  world-music
113995  world-music
113996  world-music
113997  world-music
113998  world-music
113999  world-music

[10 rows x 21 columns]
*************************************
          Unnamed: 0     popularity   duration_ms   danceability  \
count  114000.000000  114000.000000  1.140000e+05  114000.000000
mean    56999.500000      33.238535  2.280292e+05       0.566800
std     32909.109681      22.305078  1.072977e+05       0.173542
min         0.000000       0.000000  0.000000e+00       0.000000
25%     28499.750000      17.000000  1.740660e+05       0.456000
50%     56999.500000      35.000000  2.129060e+05       0.580000
```

```
75%     85499.250000      50.000000  2.615060e+05       0.695000
max    113999.000000     100.000000  5.237295e+06       0.985000

              energy            key       loudness           mode  \
count  114000.000000  114000.000000  114000.000000  114000.000000
mean        0.641383       5.309140      -8.258960       0.637553
std         0.251529       3.559987       5.029337       0.480709
min         0.000000       0.000000     -49.531000       0.000000
25%         0.472000       2.000000     -10.013000       0.000000
50%         0.685000       5.000000      -7.004000       1.000000
75%         0.854000       8.000000      -5.003000       1.000000
max         1.000000      11.000000       4.532000       1.000000

         speechiness   acousticness  instrumentalness
liveness  \
count  114000.000000  114000.000000     114000.000000  114000.000000

mean        0.084652       0.314910          0.156050       0.213553

std         0.105732       0.332523          0.309555       0.190378

min         0.000000       0.000000          0.000000       0.000000

25%         0.035900       0.016900          0.000000       0.098000

50%         0.048900       0.169000          0.000042       0.132000

75%         0.084500       0.598000          0.049000       0.273000

max         0.965000       0.996000          1.000000       1.000000


             valence          tempo  time_signature
count  114000.000000  114000.000000   114000.000000
mean        0.474068     122.147837        3.904035
std         0.259261      29.978197        0.432621
min         0.000000       0.000000        0.000000
25%         0.260000      99.218750        4.000000
50%         0.464000     122.017000        4.000000
75%         0.683000     140.071000        4.000000
max         0.995000     243.372000        5.000000
*************************************
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 114000 entries, 0 to 113999
Data columns (total 21 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   Unnamed: 0        114000 non-null  int64
 1   track_id          114000 non-null  object
 2   artists           113999 non-null  object
```

```
 3   album_name        113999 non-null  object
 4   track_name        113999 non-null  object
 5   popularity        114000 non-null  int64
 6   duration_ms       114000 non-null  int64
 7   explicit          114000 non-null  bool
 8   danceability      114000 non-null  float64
 9   energy            114000 non-null  float64
 10  key               114000 non-null  int64
 11  loudness          114000 non-null  float64
 12  mode              114000 non-null  int64
 13  speechiness       114000 non-null  float64
 14  acousticness      114000 non-null  float64
 15  instrumentalness  114000 non-null  float64
 16  liveness          114000 non-null  float64
 17  valence           114000 non-null  float64
 18  tempo             114000 non-null  float64
 19  time_signature    114000 non-null  int64
 20  track_genre       114000 non-null  object
dtypes: bool(1), float64(9), int64(6), object(5)
memory usage: 17.5+ MB
None
(114000, 21)
```

```
for col in df.columns:
    print(col, ": ", df[col].nunique())

# # unique track_ids should be equal to the number of tracks, but it
isn't
# Seems like some track_ids show up multiple times with different
track_genres (possibly some of the other features are different as
well, but that hasn't been confirmed yet)
# Seems like 1 track_genre disappears when the duplicates of the
track_ids are removed
```

```
Unnamed: 0 :  114000
track_id :  89741
artists :  31437
album_name :  46589
track_name :  73608
popularity :  101
duration_ms :  50697
explicit :  2
danceability :  1174
energy :  2083
key :  12
loudness :  19480
mode :  2
speechiness :  1489
acousticness :  5061
instrumentalness :  5346
```

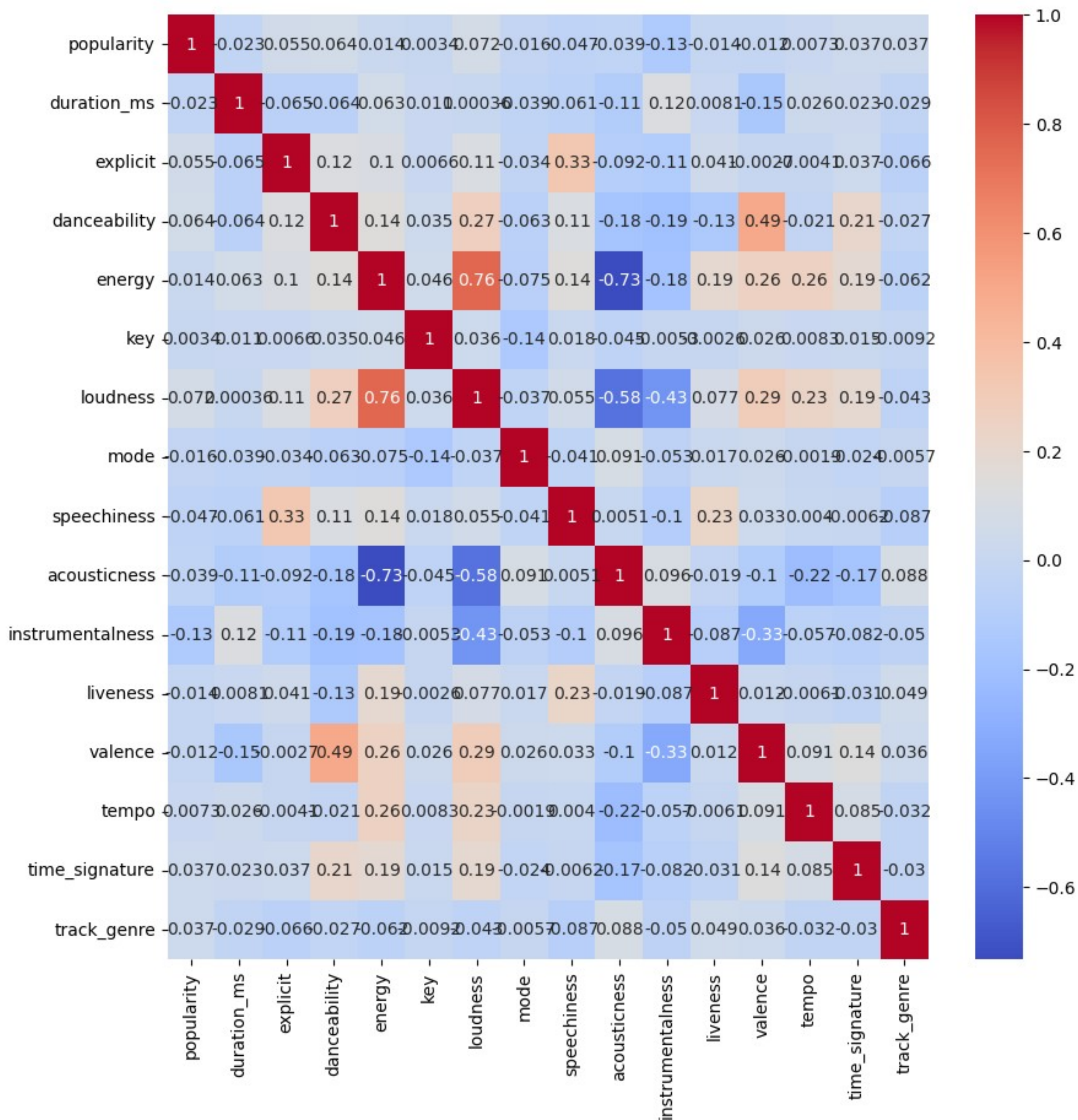
```
liveness :  1722
valence :  1790
tempo :  45653
time_signature :  5
track_genre :  114
```

```
# Drop all duplicates (need to remove the first column because these
are just indices)
revised_df = df.drop(columns='Unnamed:
0').drop_duplicates(subset=['track_id'])

# Double check that everything lines up
for col in revised_df.columns:
    print(col, ": ", revised_df[col].nunique(), "; Type: ",
revised_df[col].dtypes)
print(revised_df.shape)
```

```
track_id :  89741 ; Type:  object
artists :  31437 ; Type:  object
album_name :  46589 ; Type:  object
track_name :  73608 ; Type:  object
popularity :  101 ; Type:  int64
duration_ms :  50697 ; Type:  int64
explicit :  2 ; Type:  bool
danceability :  1174 ; Type:  float64
energy :  2083 ; Type:  float64
key :  12 ; Type:  int64
loudness :  19480 ; Type:  float64
mode :  2 ; Type:  int64
speechiness :  1489 ; Type:  float64
acousticness :  5061 ; Type:  float64
instrumentalness :  5346 ; Type:  float64
liveness :  1722 ; Type:  float64
valence :  1790 ; Type:  float64
tempo :  45653 ; Type:  float64
time_signature :  5 ; Type:  int64
track_genre :  113 ; Type:  object
(89741, 20)
```

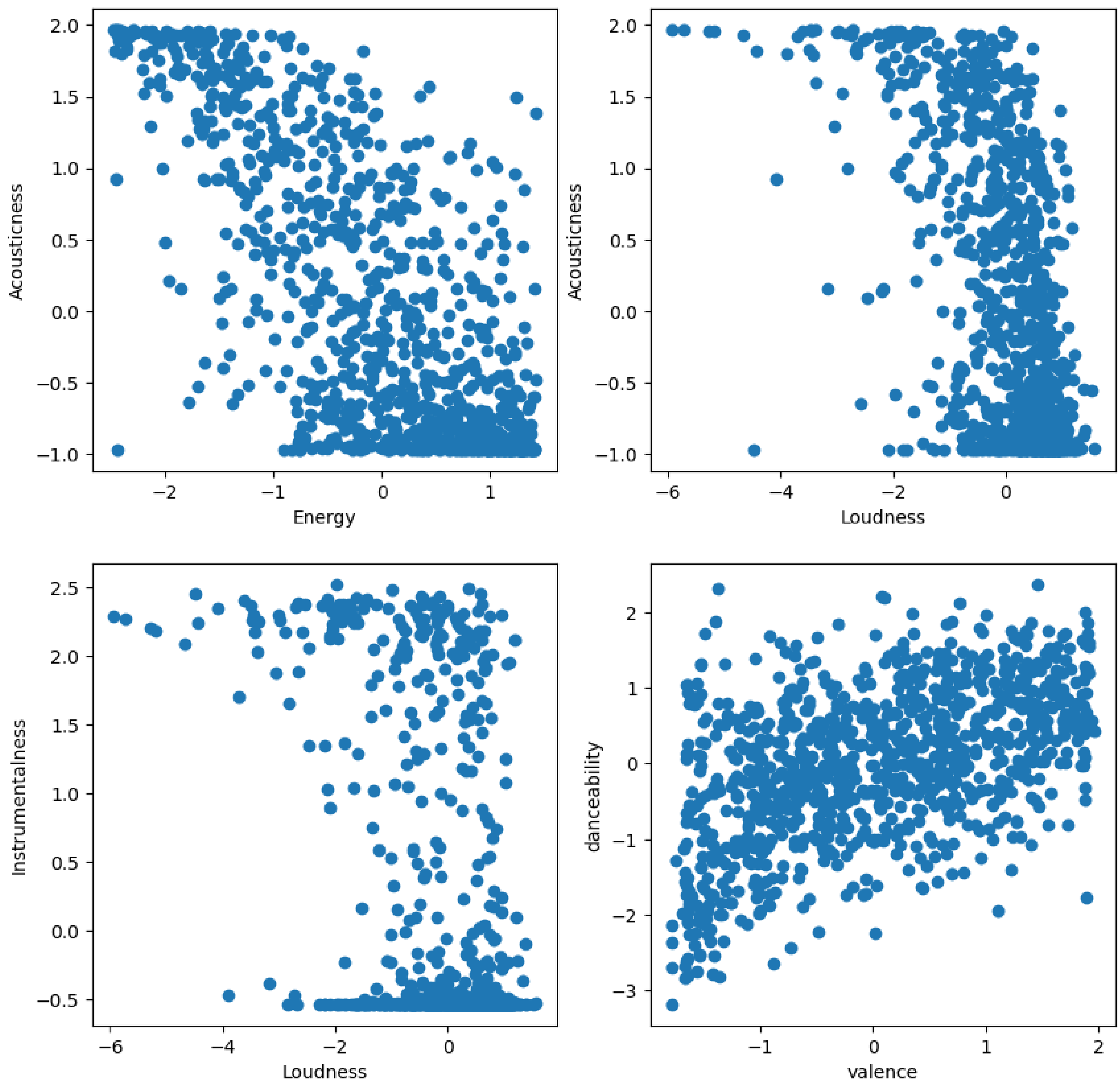
```
revised_df.drop(columns=['track_id', 'artists', 'album_name',
'track_name'], inplace=True)
columns = revised_df.columns
le = LabelEncoder()
revised_df['track_genre'] =
le.fit_transform(revised_df['track_genre'])

scaler = StandardScaler()
revised_df = scaler.fit_transform(revised_df)
revised_df = pd.DataFrame(revised_df, columns=columns)
```

```python
# Look at linear correlations between features
corr_matrix = revised_df.corr(method='pearson')
plt.figure(figsize=(10 , 10))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm')
plt.show()
```

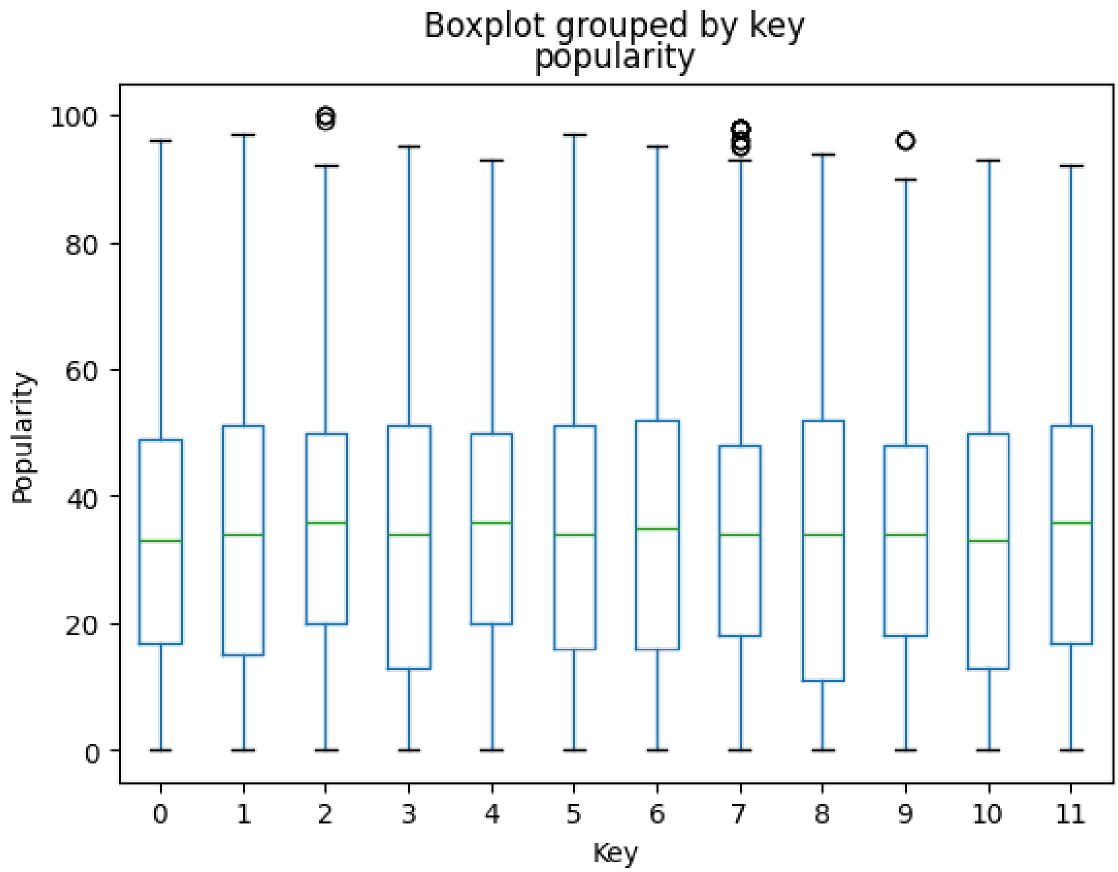



```python
sample = revised_df.sample(n=1000, random_state=42)

fig,ax = plt.subplots(2,2,figsize=(10,10))
```

```
ax[0,0].scatter(sample['energy'], sample['acousticness'])
ax[0,0].set_xlabel('Energy')
ax[0,0].set_ylabel('Acousticness')

ax[0,1].scatter(sample['loudness'], sample['acousticness'])
ax[0,1].set_xlabel('Loudness')
ax[0,1].set_ylabel('Acousticness')

ax[1,0].scatter(sample['loudness'], sample['instrumentalness'])
ax[1,0].set_xlabel('Loudness')
ax[1,0].set_ylabel('Instrumentalness')

ax[1,1].scatter(sample['valence'], sample['danceability'])
ax[1,1].set_xlabel('valence')
ax[1,1].set_ylabel('danceability')
plt.show()
```




```
df.boxplot(column='popularity', by='key', grid=False)
plt.xlabel('Key')
plt.ylabel('Popularity')

Text(0, 0.5, 'Popularity')
```


Boxplot grouped by key
popularity
Popularity
Key
100
80
60
40
20
0
0
1
2
3
4
5
6
7
8
9
10
11