CS210 INTRODUCTION TO DATA SCIENCE

FALL 2023-24

COURSE PROJECT REPORT

İlke Öncü
28930

PROJECT DESCRIPTION

The aim of this project is to analyze my different Spotify playlists to understand my music listening habits and understanding relationships between various attributes that might have an effect on that. In addition, the correlation between outside factors such as weather conditions in two different cities and the mood of the songs were also investigated. To see the analysis in a wider perspective, a more general playlist analysis using more various ways were conducted.

The project consists of two consecutive and related parts. In the first part of the project I analyzed 2 different playlists from 2 different time spans (Erasmus Times in Karlsruhe & After Erasmus is Istanbul). After analyzing some of the attributes from the playlists, I made an analysis of weather conditions during my time in those places (weather data were taken from a historical weather API website). I tried to understand whether there is a relation between music mood and the weather in different geographical locations. The analysis include extracting the data, exploratory data analysis, visualizations, machine learning, finding correlations and making comparisons as well making an hypothesis testing at the at the end (related with weather and mood of the playlist).

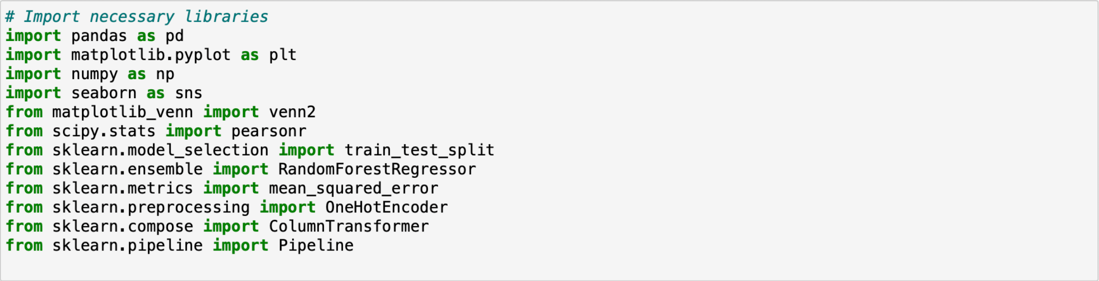
In the second part of the project I wanted to conduct more detailed analysis that might be interesting for the outcome of the project. Since I listen the songs I like repeatedly, I wanted to see my music taste covering all the months in both of the playlists that are analyzed in the first part (April-January). As an extra, at the end, I use a recommendation system using Spotify API to recommend me songs from my favourite songs' tracks. The analysis in this part includes collecting data & data cleaning, exploratory data analysis (EDA), interactive visualizations, unsupervised learning, Spotify API recommendation.

DATA ANALYSIS

FIRST PART: Two Different Playlists

Erasmus Times & After Erasmus Playlist Analysis

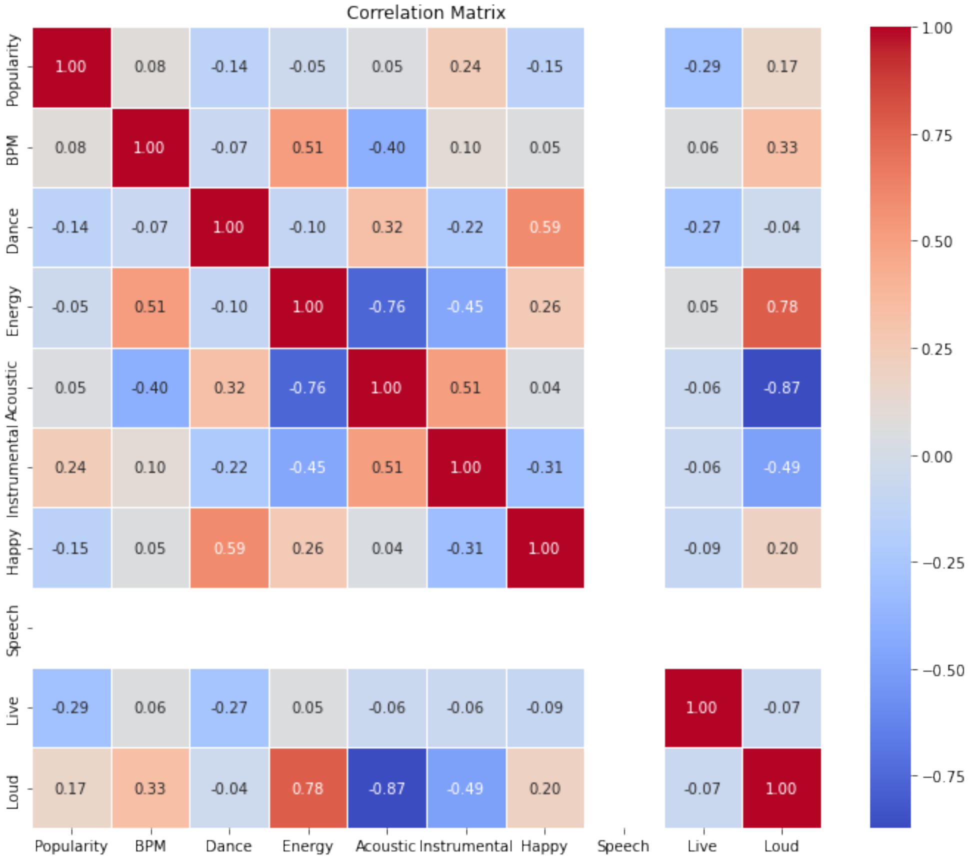
Data analysis starts with importing the necessary libraries.

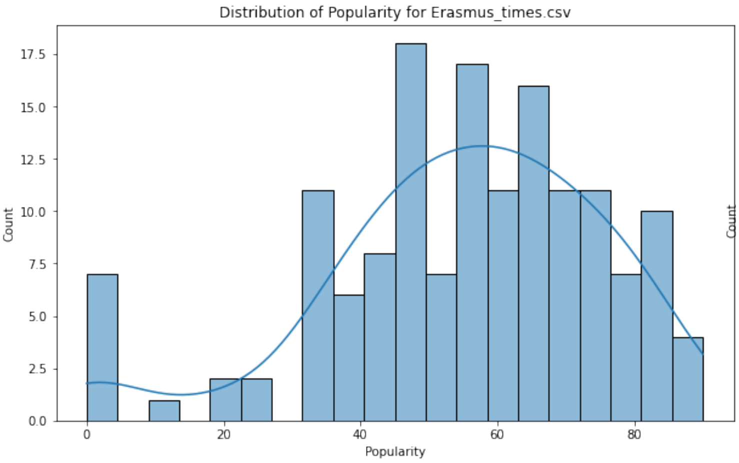


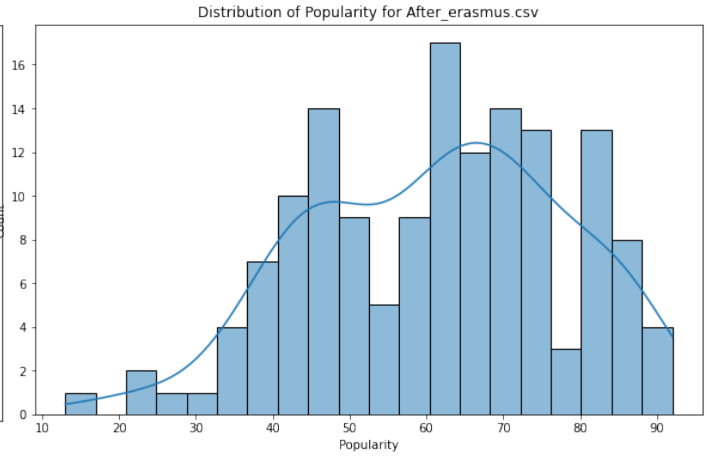
```
# Import necessary libraries
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
from matplotlib_venn import venn2
from scipy.stats import pearsonr
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error
from sklearn.preprocessing import OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
```

After importing the necessary libraries, csv files were read separately using pd.read_csv function and exploratory data analysis (EDA) was made using various ways. Performing EDA was to understand the characteristics of my playlist data. It includes exploring summary statistics, correlations, distributions, and other relevant metrics.

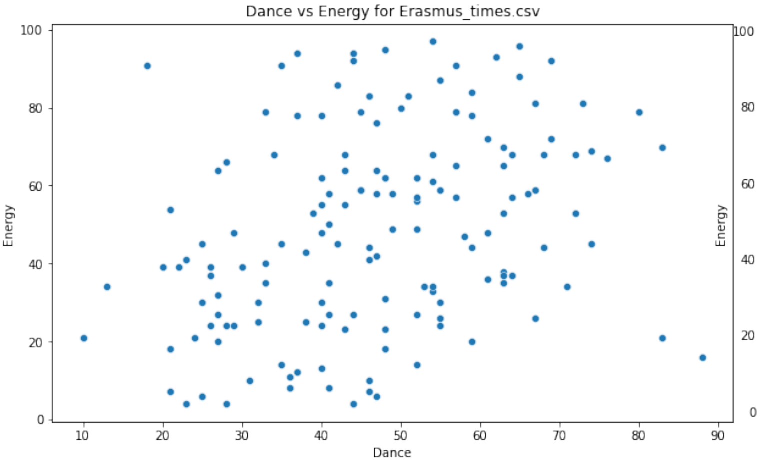
The below figure shows the correlation between different attributes in the dataset.

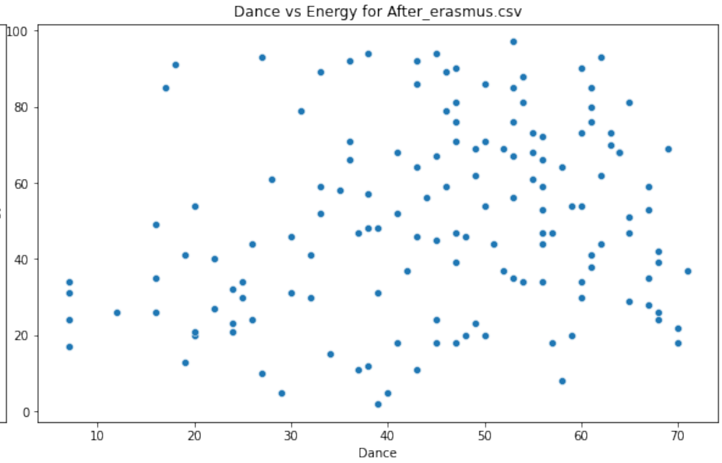






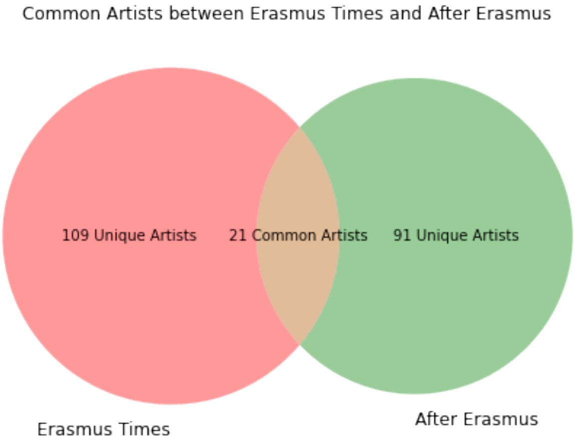
Above figures show the popularity of the songs in the playlists, sns.histplot function was used here. sns.scatterplot was used to show the relation between “dance” and “energy” as can be seen below.





Line plots, bar plots and box plots were also used to show relations between other attributes.

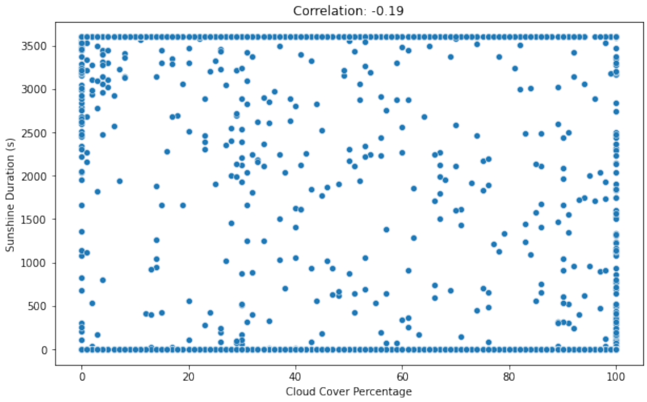
Common artists are found using the Venn diagram method as can be seen in the below figure.

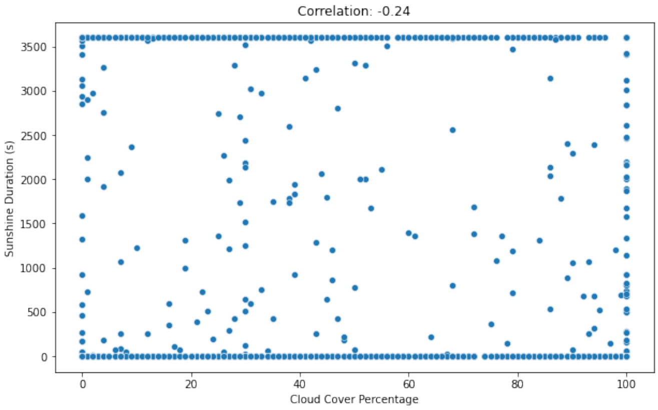


```
{'Sons Of The East', 'Tom Odell', 'Bob
Dylan', 'BØRNS', 'Michael Schulte',
'Young the Giant', 'Sleeping At Last',
'Cage The Elephant', 'The Lumineers',
'Kodaline', 'Woodkid', 'M83', 'Nick
Mulvey', 'Hozier', 'Layup', 'mor ve
ötesi', 'Sum 41', 'Coldplay', 'Thirty
Seconds To Mars', 'Linkin Park',
'Tamino'}
```

Weather Analysis

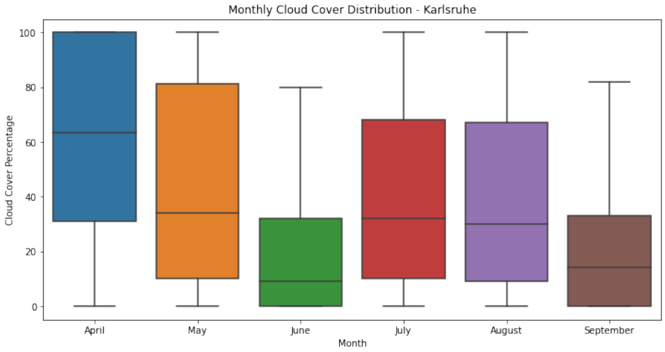
In this part, different relations were investigated using visualizations and different type of graphs.

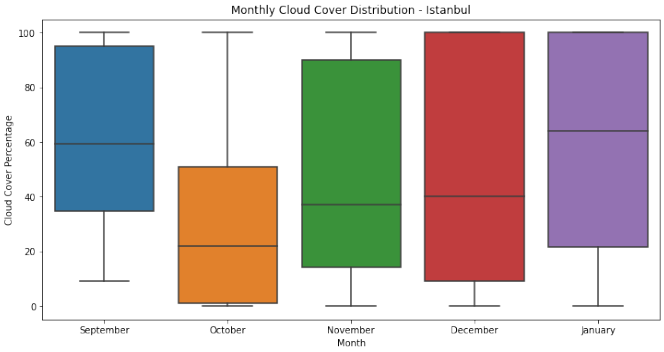


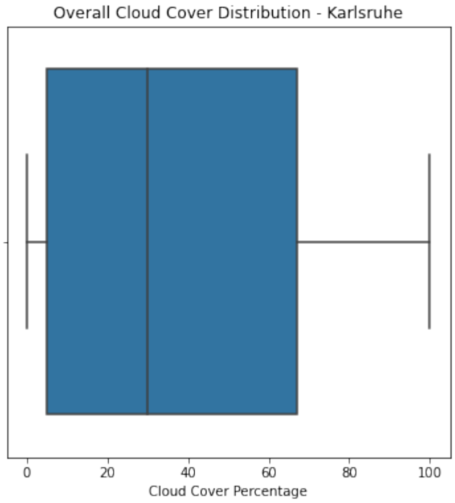


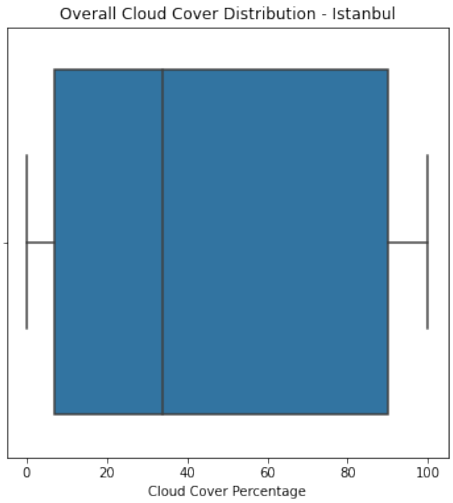
It can be seen from the above figure that there is a negative correlation between cloud cover percentage and sunshine duration in both of the datasets.

Cloud distribution is analyzed using box pilots monthly and in total shown below.









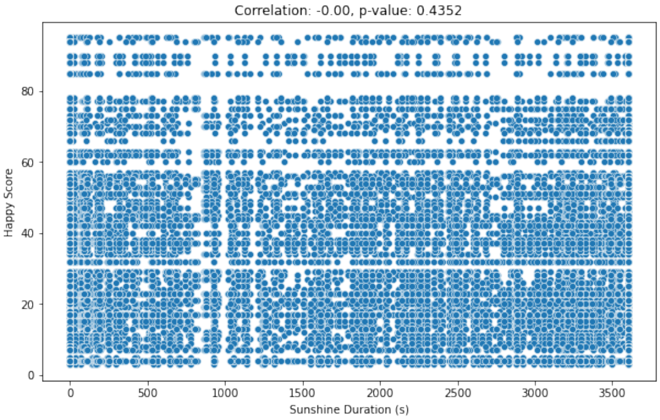
Hypothesis Testing

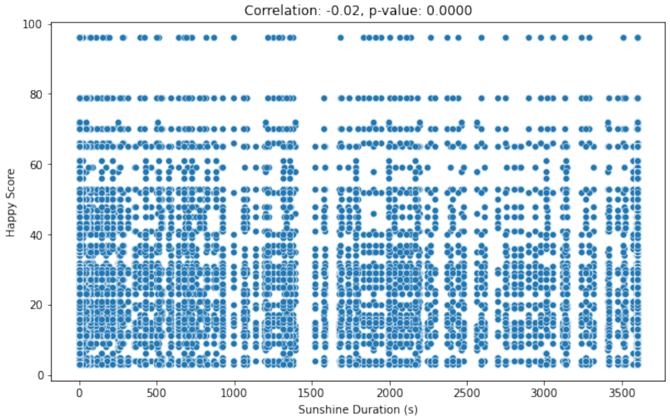
Three hypothesis tests were conducted in order to gain deeper insights about the correlation between the mood of the music and weather related factors.

Hypothesis Testing#1

For Playlist 1 (Erasmus Times): Null Hypothesis (H0): There is no significant relationship between sunshine duration and mood scores in Playlist 1. Alternative Hypothesis (H1): There is a significant relationship between sunshine duration and mood scores in Playlist 1.

For Playlist 2 (After Erasmus): Null Hypothesis (H0): There is no significant relationship between sunshine duration and mood scores in Playlist 2. Alternative Hypothesis (H1): There is a significant relationship between sunshine duration and mood scores in Playlist 2.

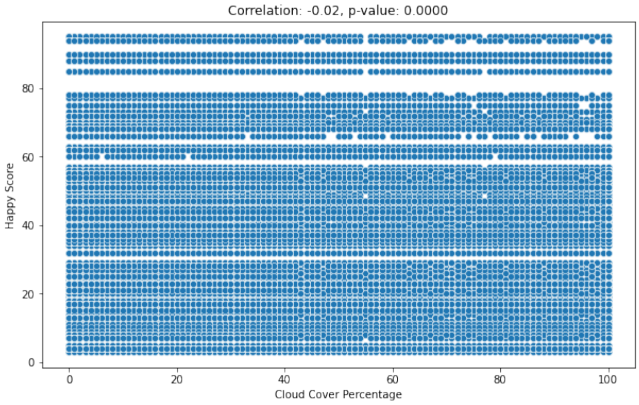


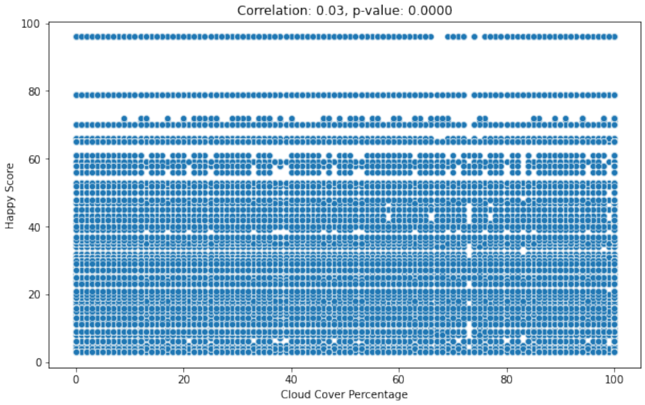


Hypothesis Testing#2

Playlist 1 (Erasmus Times): Null Hypothesis (H0): There is no significant relationship between cloud cover percentage (Weather 1) and mood scores in Playlist 1. Alternative Hypothesis (H1): There is a significant relationship between cloud cover percentage (Weather 1) and mood scores in Playlist 1.

Playlist 2 (After Erasmus): Null Hypothesis (H0): There is no significant relationship between cloud cover percentage (Weather 2) and mood scores in Playlist 2. Alternative Hypothesis (H1): There is a significant relationship between cloud cover percentage (Weather 2) and mood scores in Playlist 2.

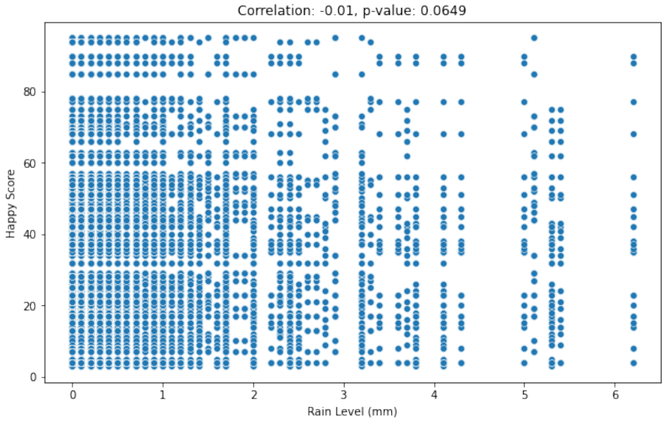


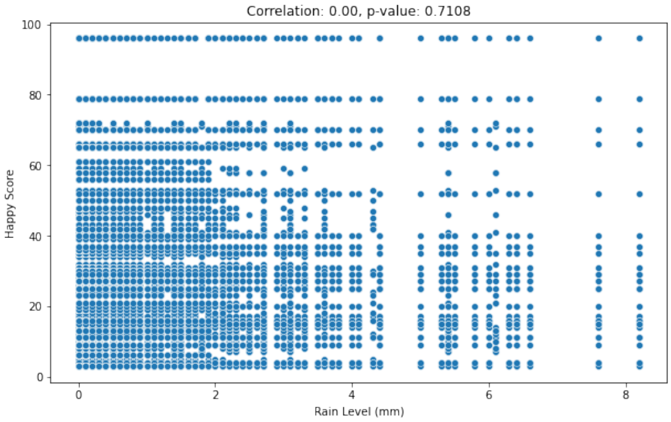


Hypothesis Testing#3

Playlist 1 (Erasmus Times): Null Hypothesis (H0): There is no significant relationship between rain levels (Weather 1) and mood scores in Playlist 1. Alternative Hypothesis (H1): There is a significant relationship between rain levels (Weather 1) and mood scores in Playlist 1.

Playlist 2 (After Erasmus): Null Hypothesis (H0): There is no significant relationship between rain levels (Weather 2) and mood scores in Playlist 2. Alternative Hypothesis (H1): There is a significant relationship between rain levels (Weather 2) and mood scores in Playlist 2.

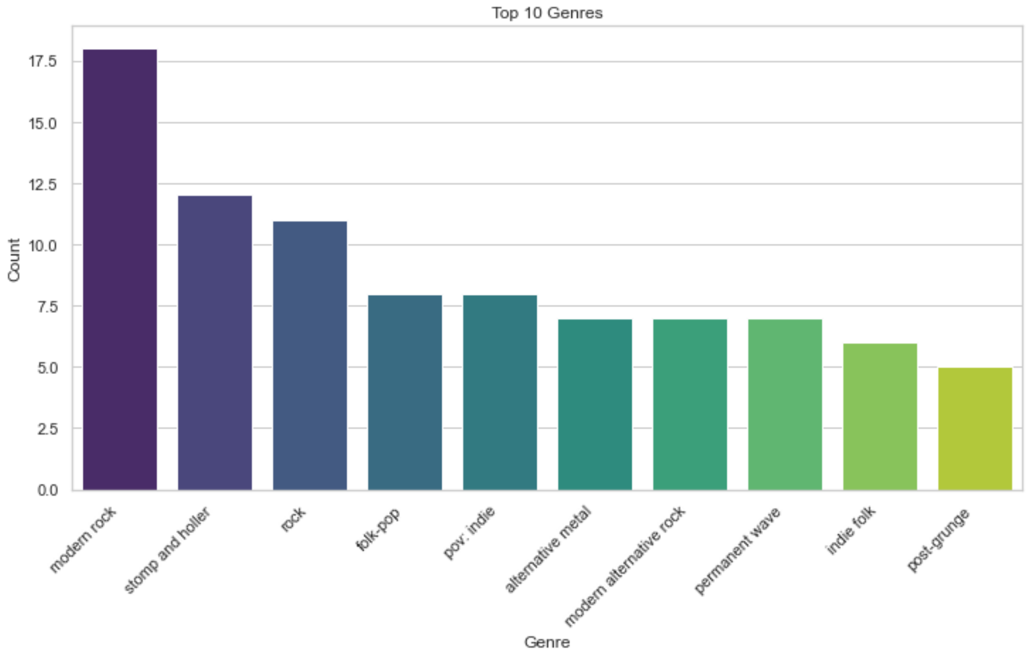




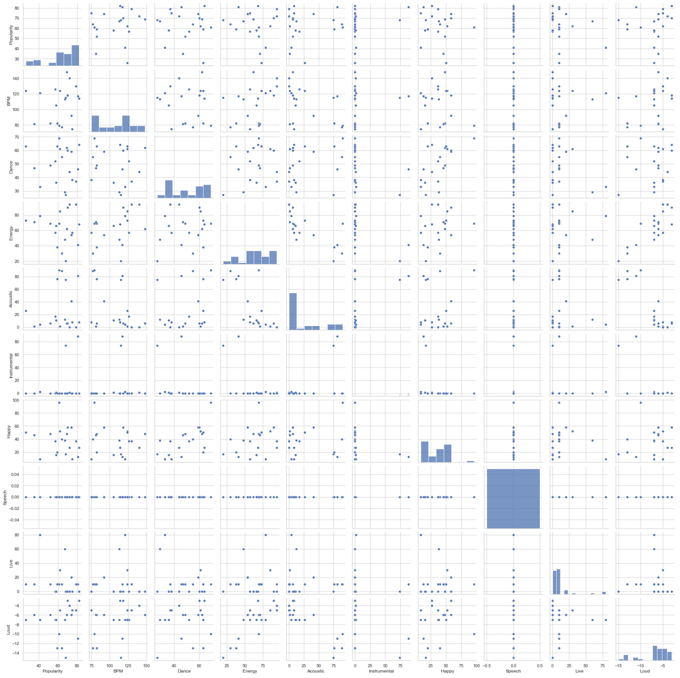
SECOND PART: Collective Playlist

Collecting Data & Data Cleaning

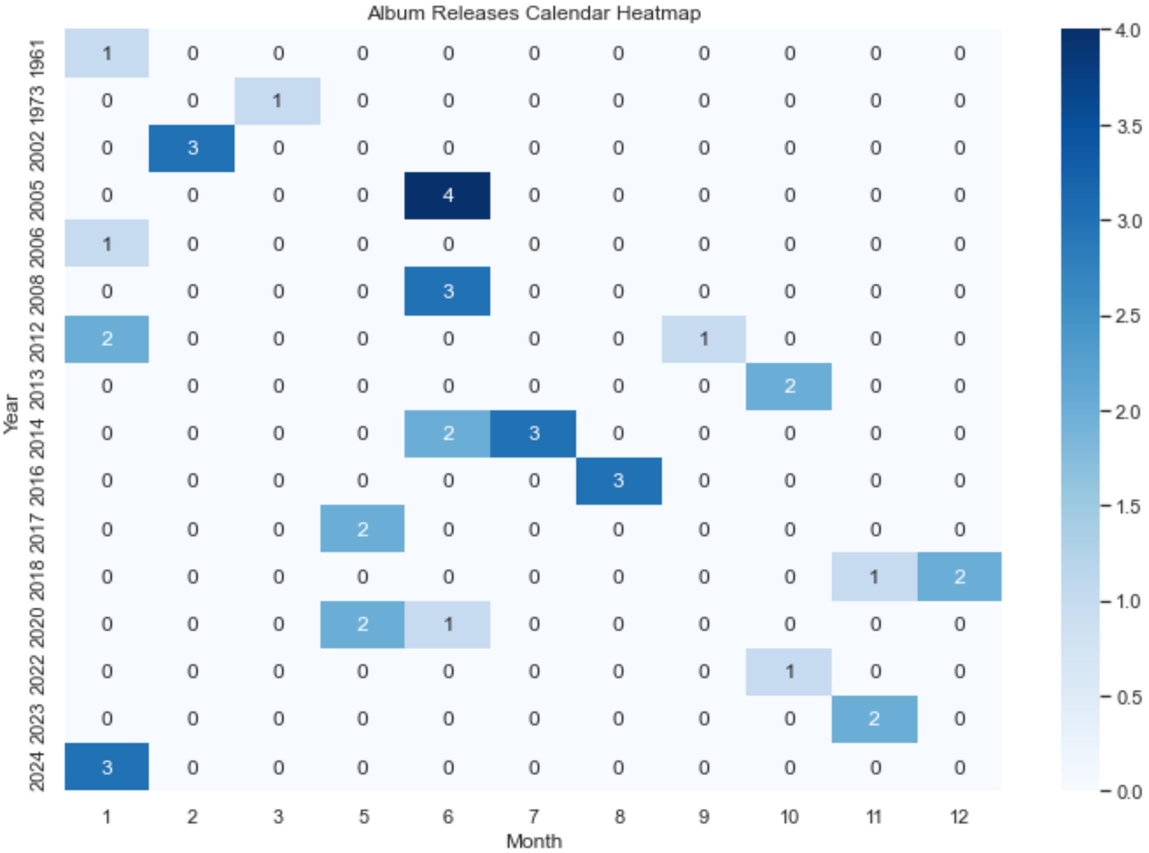
Data was clean by handling missing values, duplicates, and outliers to ensure the data is in a format suitable for analysis. EDA was also conducted in this part. Data distribution was analyzed using histograms. Top 10 genres were found for the later anaylsis as can be seen in the below figure.



Pair plots were created using sns.pairplot code as can be seen below.



Heatmap also shows different relationships between several attributes.



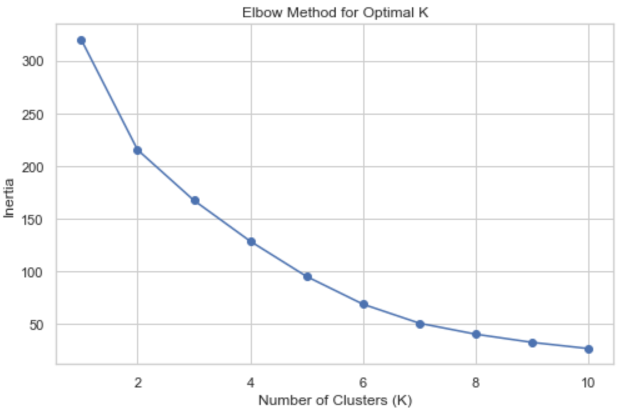
Year	1	2	3	5	6	7	8	9	10	11	12
1961	1	0	0	0	0	0	0	0	0	0	0
1973	0	0	1	0	0	0	0	0	0	0	0
2002	0	3	0	0	0	0	0	0	0	0	0
2005	0	0	0	0	4	0	0	0	0	0	0
2006	1	0	0	0	0	0	0	0	0	0	0
2008	0	0	0	0	3	0	0	0	0	0	0
2012	2	0	0	0	0	0	0	1	0	0	0
2013	0	0	0	0	0	0	0	0	2	0	0
2014	0	0	0	0	2	3	0	0	0	0	0
2016	0	0	0	0	0	0	3	0	0	0	0
2017	0	0	0	2	0	0	0	0	0	0	0
2018	0	0	0	0	0	0	0	0	0	1	2
2020	0	0	0	2	1	0	0	0	0	0	0
2022	0	0	0	0	0	0	0	0	1	0	0
2023	0	0	0	0	0	0	0	0	0	2	0
2024	3	0	0	0	0	0	0	0	0	0	0

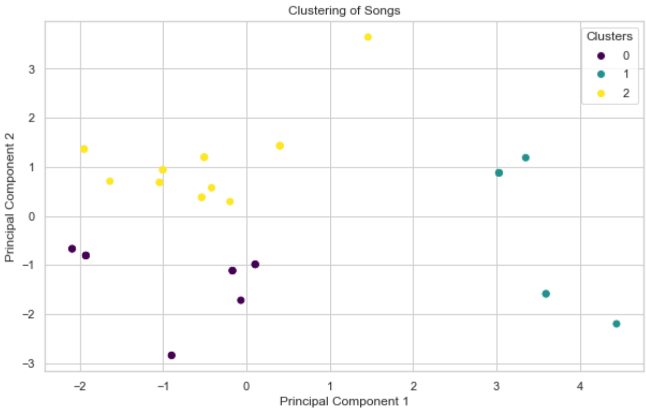
Interactive Visualizations

They were added for fun purposes.

Unsupervised Learning

Features were being standardized and based on the elbow method the optimal K was chosen and fitted the KMeans model as can be seen in the below visuals.





Spotify API Music Recommendation

By using Spotify API and Spotify function, some songs were recommended from different playlists in Spotify.

CONCLUSION

Data can show us many things and different methods for data analysis opens so many new doors and perspectives to research are.