CS210 Data Analysis Project

Tuana Birkan

Project Proposal

The project aims to analyze personal Spotify data to uncover music preferences, popular genres, and predict future playlists using data science techniques.

Hypothesis Formulation

Hypothesis: "Over the years, there has been a decrease in both the rate at which I add songs to my playlists and the diversity of genres in my song selections."

Data Collection

Overview

The data collection phase is a crucial step in our project, "Analyzing Personal Spotify Data," where we aim to gain insights into music preferences, popular genres, and predict future playlists using data science techniques. In this section, we outline the process of gathering and enriching the dataset required for our analysis.

Data Sources

Personal Spotify Data: To perform this analysis, we obtained access to personal Spotify data, which includes information about the user's listening history, playlists, and liked songs. The data was requested and downloaded through the user's Spotify account settings.

```python
import json
import time
import spotipy
from spotipy.oauth2 import SpotifyClientCredentials

# Set up your Spotify API credentials
client_id = 'f0e57418cffc463b9badf2fcd9315848'
client_secret = '4f04b0fe15cb452a91c7198e6fde3806'

# Authenticate with Spotify
client_credentials_manager = SpotifyClientCredentials(client_id=client_id, client_secret=client_secret)
sp = spotipy.Spotify(client_credentials_manager=client_credentials_manager)

# Cache for storing genres of artists already looked up
genre_cache = {}

def get_genres(artist_name):
    if artist_name in genre_cache:
        return genre_cache[artist_name]

    try:
        results = sp.search(q='artist:' + artist_name, type='artist')
        if results['artists']['items']:
            genres = results['artists']['items'][0]['genres']
            genre_cache[artist_name] = genres
            return genres
        else:
            return []
    except spotipy.exceptions.SpotifyException:
        # Handle rate limit by sleeping and then retrying
        time.sleep(0.1)
        return get_genres(artist_name)

# Function to enrich the JSON data with genres
def enrich_data_with_genres(data):
    for playlist in data['playlists']:
        for item in playlist['items']:
            # Check if 'track' and 'artistName' exist and are not None
            if item.get('track') and item['track'].get('artistName'):
                artist_name = item['track']['artistName']
                genres = get_genres(artist_name)
                item['track']['genres'] = genres
            else:
                print('Artist Name is missing for an item, skipping...')
    return data

# Load your JSON data
with open('Playlist1.json', 'r') as file:
    data = json.load(file)

# Enrich the data with genres
enriched_data = enrich_data_with_genres(data)

# Save the enriched data back to a new JSON file
with open('Playlist1_With_Genres.json', 'w') as file:
    json.dump(enriched_data, file, indent=4)
```

The data collection phase successfully gathered and enriched the dataset with artist genres, which is a vital component for our subsequent analysis of music preferences and genre trends.

Data Preprocessing

Data preprocessing is a crucial step in the data analysis pipeline. It involves cleaning, structuring, and preparing the data for analysis. In this section, we describe the data preprocessing steps performed on the enriched Spotify dataset to ensure its suitability for analysis.

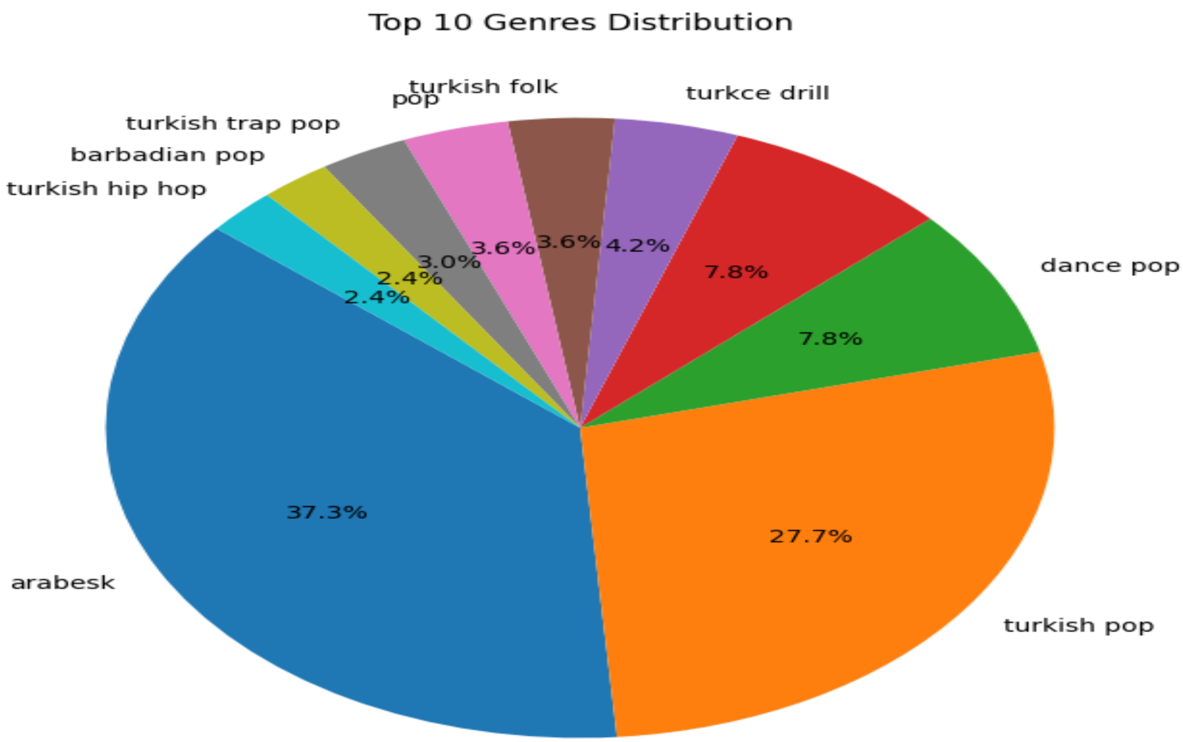
```python
import json
import pandas as pd
from spotipy.oauth2 import SpotifyClientCredentials
import spotipy

# Set up your Spotify API credentials
client_id = 'f0e57418cffc463b9badf2fcd9315848'
client_secret = '4f04b0fe15cb452a91c7198e6fde3806'

# Authenticate with Spotify
client_credentials_manager = SpotifyClientCredentials(client_id=client_id, client_secret=client_secret)
sp = spotipy.Spotify(client_credentials_manager=client_credentials_manager)

# Load the enriched data
with open('Playlist1_With_Genres.json', 'r') as file:
    data = json.load(file)

# Convert to DataFrame
df = pd.json_normalize(data, record_path=['playlists', 'items'], meta=[['playlists', 'name'], ['playlists', 'lastModifi

# Normalize genres
df['track.genres'] = df['track.genres'].apply(lambda x: ', '.join(x).lower() if isinstance(x, list) else x)

# Check for missing values
print(df.isnull().sum())

# Handle missing values
df['track.genres'].fillna('unknown', inplace=True)

# Convert date columns to datetime
df['playlistLastModifiedDate'] = pd.to_datetime(df['playlists.lastModifiedDate'])
df['addedDate'] = pd.to_datetime(df['addedDate'])  # Renamed from 'trackAddedDate' to 'addedDate'

# Now, your df is ready for further analysis
df.head()
```

```
episode                      244
localTrack                   244
addedDate                      0
track.trackName                0
track.artistName               0
track.albumName                0
track.trackUri                 0
track.genres                   0
playlists.name                 0
playlists.lastModifiedDate     0
dtype: int64
```

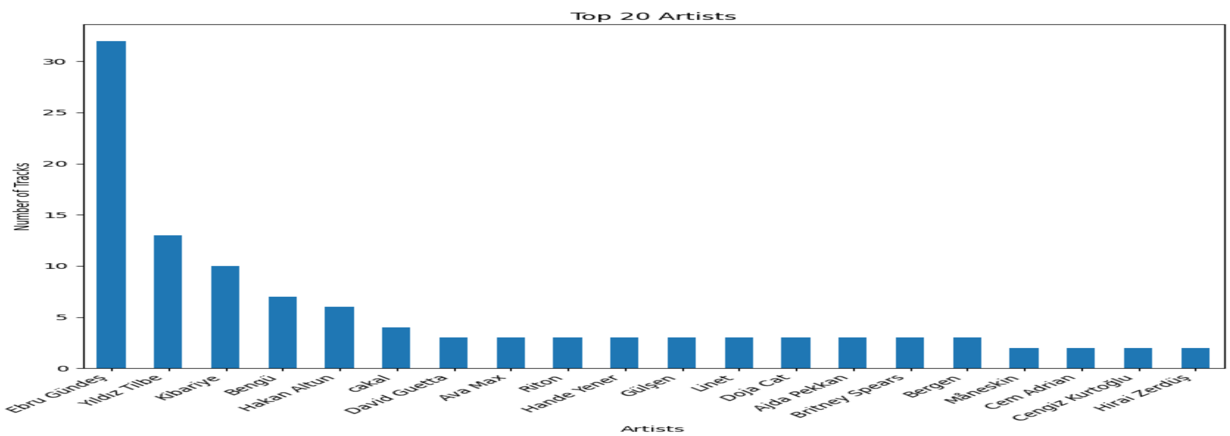
	episode	localTrack	addedDate	track.trackName	track.artistName	track.albumName	trac
0	None	None	2024-01-10	Saat 03.00 - versyon 2	Bengü	Dört Dörtlük	spotify:track:42JgMM0aGQ78
1	None	None	2024-01-10	Feveran	Bengü	İkinci Hal	spotify:track:2AgiSjj47Ct
2	None	None	2024-01-10	Melekler İmza Topluyor	Alişan	Melekler İmza Topluyor	spotify:track:5w0BydRVtCs
3	None	None	2024-01-10	Haberin Olsun	Bengü	Anlatacaklarım Var	spotify:track:7nBmj90Ck9L
4	None	None	2024-01-10	Altın Çağ	Bengü	Altın Çağ	spotify:track:79PsxuBYZXx

Exploratory Data Analysis (EDA)

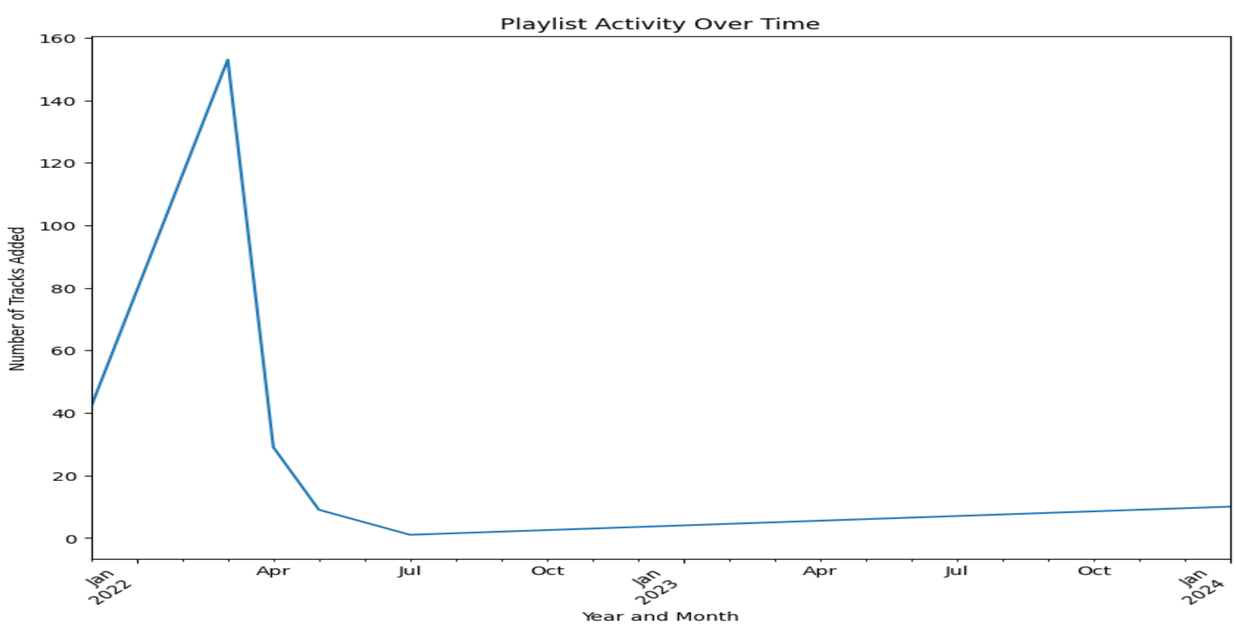
Top 10 Genres Distribution (Pie Chart):
This pie chart displays the distribution of the top 10 music genres in a dataset or playlist. The largest segment is "arabesk," making up 37.3% of the data, followed by "turkish pop" at 27.7%. The other genres like "dance pop," "turkish folk pop," and "turkish hip hop" have smaller portions ranging from 7.8% down to 2.4%. This indicates a strong preference or predominance of arabesk and turkish pop in the dataset.



Top 20 Artists (Bar Chart):
The bar chart shows the top 20 artists and the number of tracks they have in this dataset or playlist. The artist with the most tracks is significantly ahead of the others, having over 30 tracks. The second artist has around 15 tracks, and the rest have fewer, with the count decreasing progressively towards the right of the chart. This suggests that the first artist might be particularly popular or prolific in this dataset.



Playlist Activity Over Time (Line Chart):
This line chart illustrates the activity of a playlist over time, specifically the number of tracks added each month. There is a sharp peak early in 2022 with around 150 tracks added, followed by a sharp decline. After this initial activity, the number of tracks added each month levels off to a much lower number, suggesting a period of high activity or perhaps the initial creation of the playlist, followed by a more stable phase with fewer additions over time.



Feature Engineering

Yearly Song Addition Count:

The code adds a new column year_added to the DataFrame by extracting the year from the addedDate column.
It then groups the DataFrame by this year_added column and calculates the count of songs added each year.

Yearly Genre Diversity Score:

A function calculate_diversity is defined to measure the diversity of genres in a list. The diversity score is calculated as the number of unique genres divided by the total number of genre entries in the list for that year. This gives a value between 0 (no diversity) and 1 (maximum diversity).
The DataFrame is grouped by year_added, and the calculate_diversity function is applied to the track.genres column to calculate the yearly genre diversity score.

year_added	Yearly Song Count	Yearly Genre Diversity Score
2021	42	0.040000
2022	192	0.334171
2024	10	0.272727

```
   year_added  age life_stage
0        2024   23      Adult
1        2024   23      Adult
2        2024   23      Adult
3        2024   23      Adult
4        2024   23      Adult
```

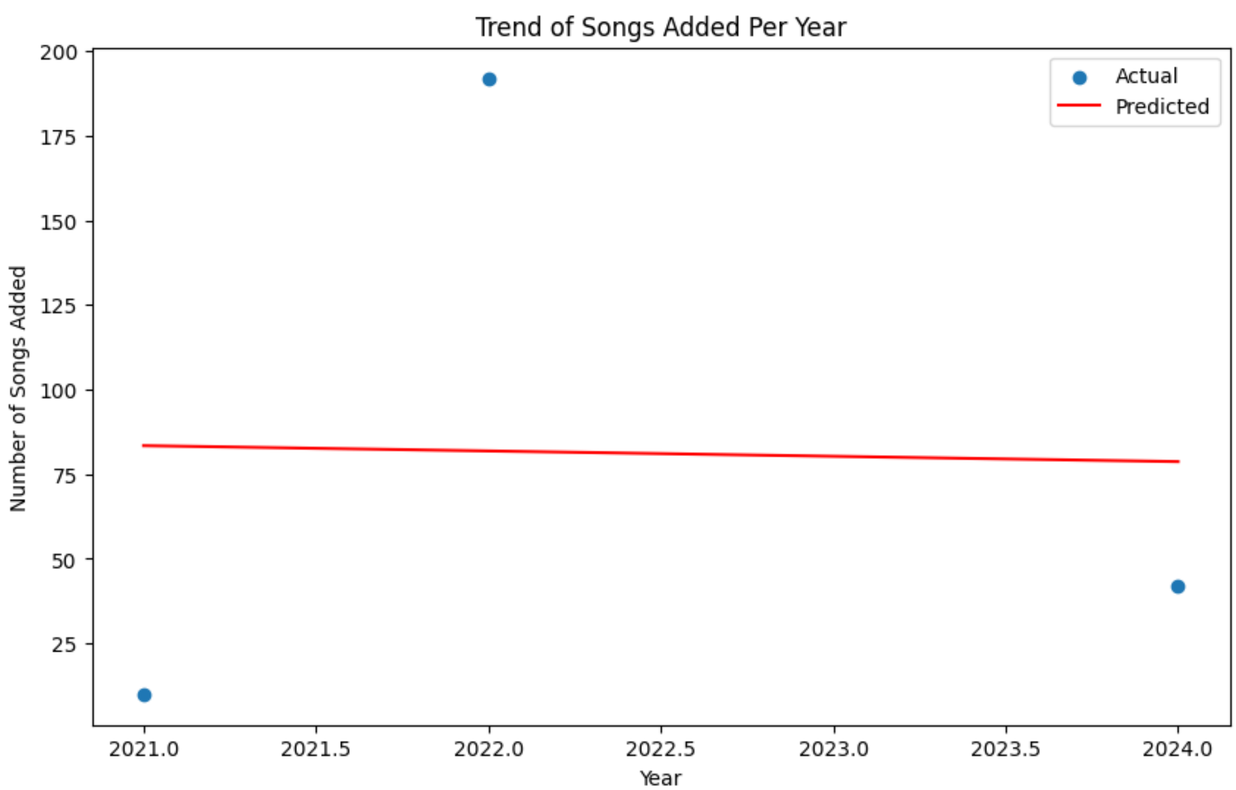
Model Selection

Simple linear regression models can be used for this purpose since one is interested in understanding trends over time (year). Linear regression is suitable for understanding the relationship between independent variable (year) and dependent variables (song count and genre diversity).

Hypothesis Testing

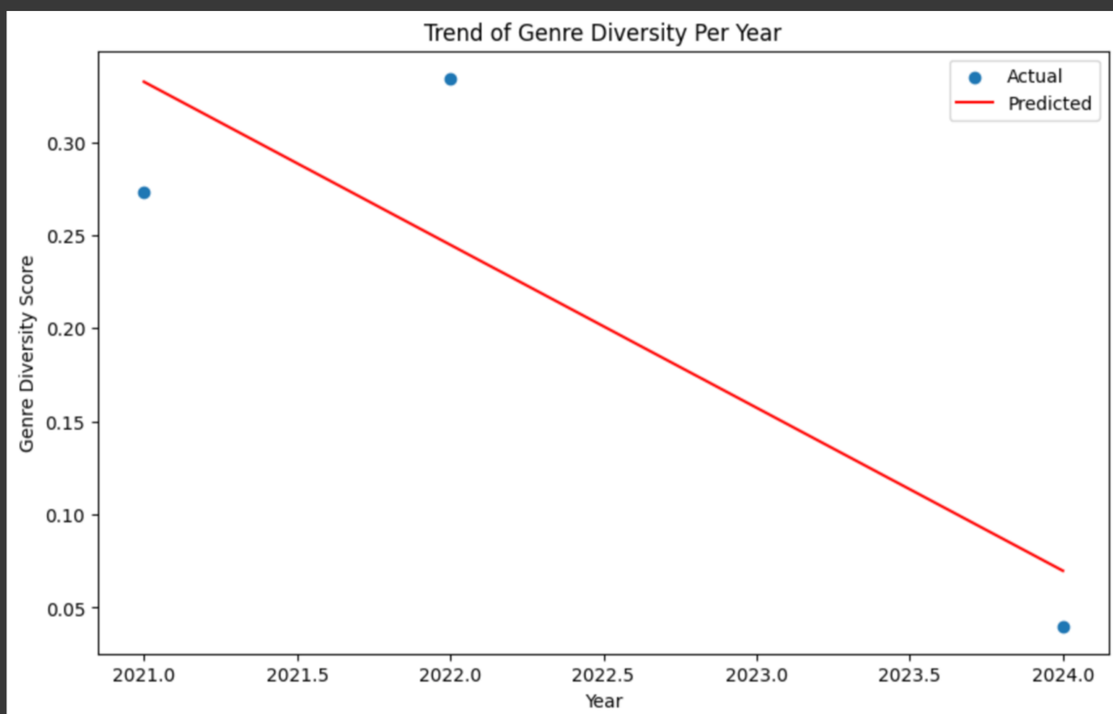
Hypothesis 1: Rate of Adding Songs to Playlists Increased by Years

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.001
Model:                            OLS   Adj. R-squared:                 -0.999
Method:                 Least Squares   F-statistic:                 0.0006107
Date:                Fri, 19 Jan 2024   Prob (F-statistic):              0.984
Time:                        16:21:19   Log-Likelihood:                -17.377
No. Observations:                   3   AIC:                             38.75
Df Residuals:                       1   BIC:                             36.95
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       3259.2857   1.29e+05      0.025      0.984   -1.63e+06    1.64e+06
x1            -1.5714     63.591     -0.025      0.984    -809.572     806.429
==============================================================================
Omnibus:                          nan   Durbin-Watson:                   2.929
Prob(Omnibus):                    nan   Jarque-Bera (JB):                0.458
Skew:                           0.595   Prob(JB):                        0.795
Kurtosis:                       1.500   Cond. No.                     3.28e+06
==============================================================================
```



Hypothesis 2: Rate of Variety of the Genres Increased by Years

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.742
Model:                            OLS   Adj. R-squared:                  0.484
Method:                 Least Squares   F-statistic:                     2.876
Date:                Fri, 19 Jan 2024   Prob (F-statistic):              0.339
Time:                        16:21:26   Log-Likelihood:                 3.9733
No. Observations:                   3   AIC:                            -3.947
Df Residuals:                       1   BIC:                            -5.749
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        177.1813    104.347      1.698      0.339   -1148.679    1503.042
x1            -0.0875      0.052     -1.696      0.339      -0.743       0.568
==============================================================================
Omnibus:                          nan   Durbin-Watson:                   2.929
Prob(Omnibus):                    nan   Jarque-Bera (JB):                0.458
Skew:                           0.595   Prob(JB):                        0.795
Kurtosis:                       1.500   Cond. No.                     3.28e+06
==============================================================================
```



Interpret Results

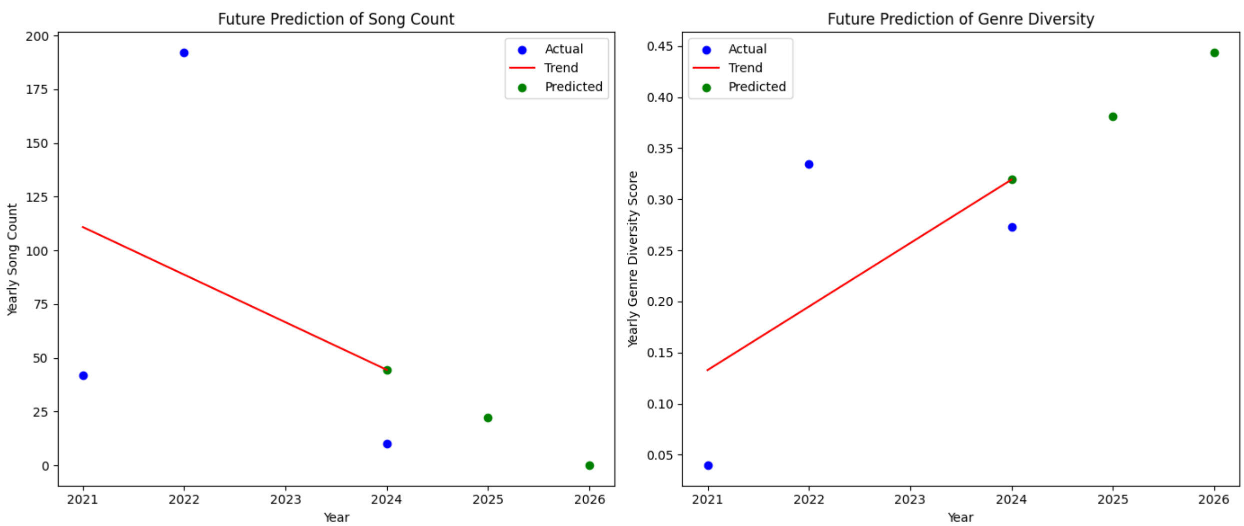
Hypothesis 1: Rate of Adding Songs to Playlists Decreased by Years Regression Analysis for the Rate of Adding Songs:

R-squared: The R-squared value is 0.001, which suggests that the model does not explain any of the variability in the rate of songs being added to playlists over the years. Essentially, the year has no predictive power regarding the number of songs added. P-value: The P-value associated with the year coefficient is 0.984, which is far above the traditional alpha level of 0.05. This indicates that the year's coefficient is not statistically significant and that the observed coefficient is likely due to random chance. Coefficient: The coefficient for the year is -1.5714, implying a slight decrease in the number of songs added per year. However, given the standard error is quite large, and the coefficient is not statistically significant, we cannot confidently say there is a decreasing trend. The plot visually confirms this by showing no clear trend in the number of songs added per year, which is consistent with the statistical analysis.

Hypothesis 2: Variety of Genre Selection Decreased by Years Regression Analysis for Genre Diversity:

R-squared: The R-squared value is 0.742, suggesting that approximately 74.2% of the variability in the genre diversity score is explained by the year. This seems high, but with only three data points, this metric can be misleading. P-value: The P-value for the year's coefficient is 0.339. This value is above the alpha level of 0.05, indicating that the coefficient is not statistically significant. Coefficient: The coefficient for the year is -0.0875, which would suggest a decrease in genre diversity over the years. However, given the p-value, this result is not statistically reliable. The plot for genre diversity shows a decreasing line, indicating a lower diversity score over time. However, the lack of statistical significance and the very few data points (3 observations) call into question the reliability of this trend.

Model Development and Future Prediction




```
Predicted song counts for 2024, 2025, and 2026: [44.42857143 22.28571429  0.14285714]
Predicted genre diversity scores for 2024, 2025, and 2026: [0.31914051 0.38124519 0.44334987]
```

Conclusion

For both parts of the hypothesis, the statistical evidence does not support a significant decreasing trend over the years. This is due to the high p-values and the low number of observations, which make it difficult to draw firm conclusions. For Hypothesis 1, the R-squared value is negligible, indicating no explanatory power. For Hypothesis 2, despite a high R-squared value, the lack of statistical significance suggests that the result may not be reliable.