19.01.2024

CS 210 - PROJECT

Introduction to Data Science

Yağız Bartu Arslan
SABANCI UNIVERSITY

Yağız Bartu Arslan CS210 – Course Project Report

Introduction

This project is an exploration into the potential correlation between daily step count and the duration of Spotify usage. It utilizes Exploratory Data Analysis and Machine Learning techniques to analyze a year's worth of data, spanning from December 2022 to December 2023.

The analysis includes various aspects such as hourly listening intervals, daily averages, and top artists and tracks.

The project also involves parsing XML data and converting it into CSV format for further analysis. The investigation identifies days with the highest step count, average walking speed, and total energy expenditure.

The project culminates with a correlation analysis between step count and listening time, testing the initial hypotheses. This comprehensive study provides valuable insights into the interplay between physical activity and music listening habits.

Analysis of the Project

The data from two separate JSON files, `StreamingHistory0.json` and `StreamingHistory1.json`, is loaded. These files contain the Spotify streaming history. The data from each file is converted into a pandas DataFrame for easier manipulation and analysis. The DataFrames are then concatenated to form a combined DataFrame, which contains the complete Spotify streaming history. The first few rows of each DataFrame are displayed for verification. This ensures that the data is correctly loaded and ready for further analysis.

```
DataFrame for StreamingHistory0:
            endTime    artistName                     trackName  msPlayed
0  2022-12-12 08:24  Metro Boomin  Around Me (feat. Don Toliver)    191520
1  2022-12-12 08:26     Bravo 1-2                        Genesis     90509
2  2022-12-12 08:28     Pop Smoke                     Invincible    127546
3  2022-12-12 08:29           C2d                  With My Hoe!     82515
4  2022-12-12 09:52      Madrigal              Seni Dert Etmeler     52931

DataFrame for StreamingHistory1:
            endTime    artistName              trackName  msPlayed
0  2023-10-04 09:19       A*S*Y*S   Ac/Id - Original Mix     46296
1  2023-10-04 09:23  Travis Scott            COFFEE BEAN    209115
2  2023-10-04 09:26  Travis Scott     HOUSTONFORNICATION    217828
3  2023-10-04 09:29  Travis Scott                  NC-17    156885
4  2023-10-04 09:35  Travis Scott  STOP TRYING TO BE GOD    338438

Combined DataFrame for Spotify Data:
            endTime    artistName                     trackName  msPlayed
0  2022-12-12 08:24  Metro Boomin  Around Me (feat. Don Toliver)    191520
1  2022-12-12 08:26     Bravo 1-2                        Genesis     90509
2  2022-12-12 08:28     Pop Smoke                     Invincible    127546
3  2022-12-12 08:29           C2d                  With My Hoe!     82515
4  2022-12-12 09:52      Madrigal              Seni Dert Etmeler     52931
```

The Spotify data is preprocessed to ensure it is in a suitable format for analysis. This involves checking the shape and columns of the DataFrame, summarizing the data, and checking for missing values. It is confirmed that there are no missing values in the data.

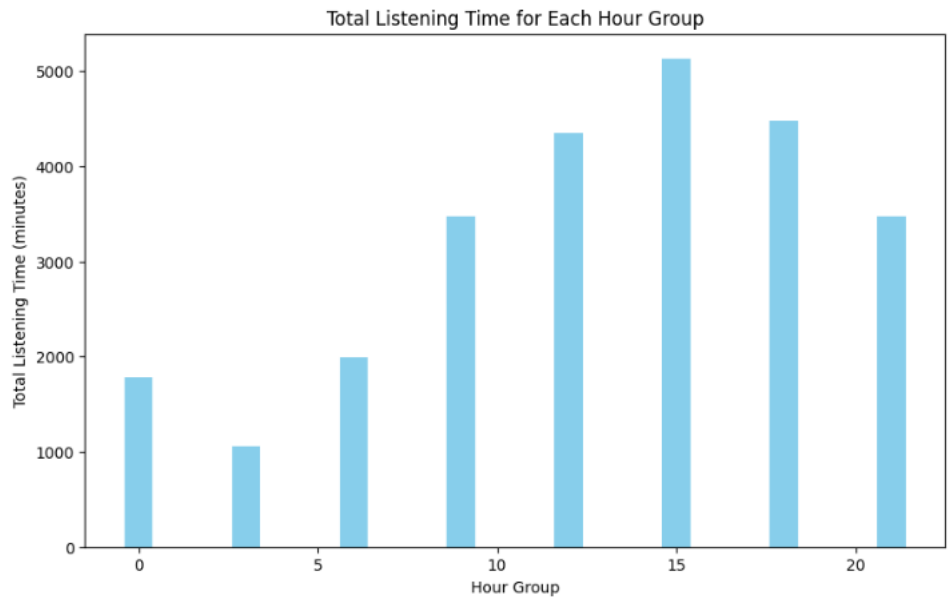
```
Shape of the DataFrame for Spotify Data: (11871, 4)

Columns of the DataFrame for Spotify Data:
Index(['endTime', 'artistName', 'trackName', 'msPlayed'], dtype='object')

Summary for the DataFrame for Spotify Data:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11871 entries, 0 to 11870
Data columns (total 4 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   endTime     11871 non-null  object
 1   artistName  11871 non-null  object
 2   trackName   11871 non-null  object
 3   msPlayed    11871 non-null  int64
dtypes: int64(1), object(3)
memory usage: 371.1+ KB
None

Missing values in Combined DataFrame for Spotify Data:
endTime       0
artistName    0
trackName     0
msPlayed      0
dtype: int64
```

The DataFrame now includes the total listening time for each hour of the day, providing a clear picture of the user's Spotify usage patterns. This data is ready for further analysis and correlation with step count data. The DataFrame shows the hour, the total listening time in milliseconds, and the total listening time in minutes. This information will be used in the subsequent steps of the project.

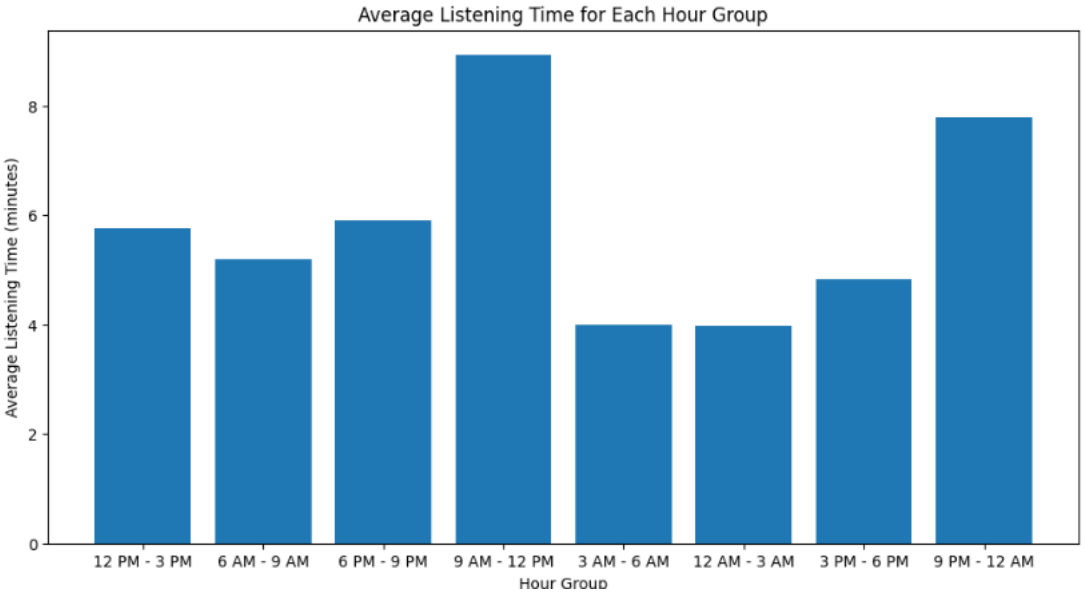
```
# We correctly filled the missing values (if any)

Hourly Listening Time DataFrame:
    hour   msPlayed  totalListeningTimeMin
0      0   59067133             984.452217
1      1   31758551             529.309183
2      2   16291184             271.519733
3      3   25551375             425.856250
4      4   13646704             227.445067
5      5   24587139             409.785650
6      6   34674717             577.911950
7      7   39527043             658.784050
8      8   45647515             760.791917
9      9   58930981             982.183017
10    10   79387683            1323.128050
11    11   70418013            1173.633550
12    12   79869570            1331.159500
13    13   86484263            1441.404383
14    14   94851334            1580.855567
15    15  107462610            1791.043500
```

The data is further grouped into three-hour intervals, and the total listening time for each interval is calculated. This information is visualized in a bar chart, which provides a clear picture of the listening habits throughout the day. The chart reveals that the majority of the music listening occurs after 13:00,

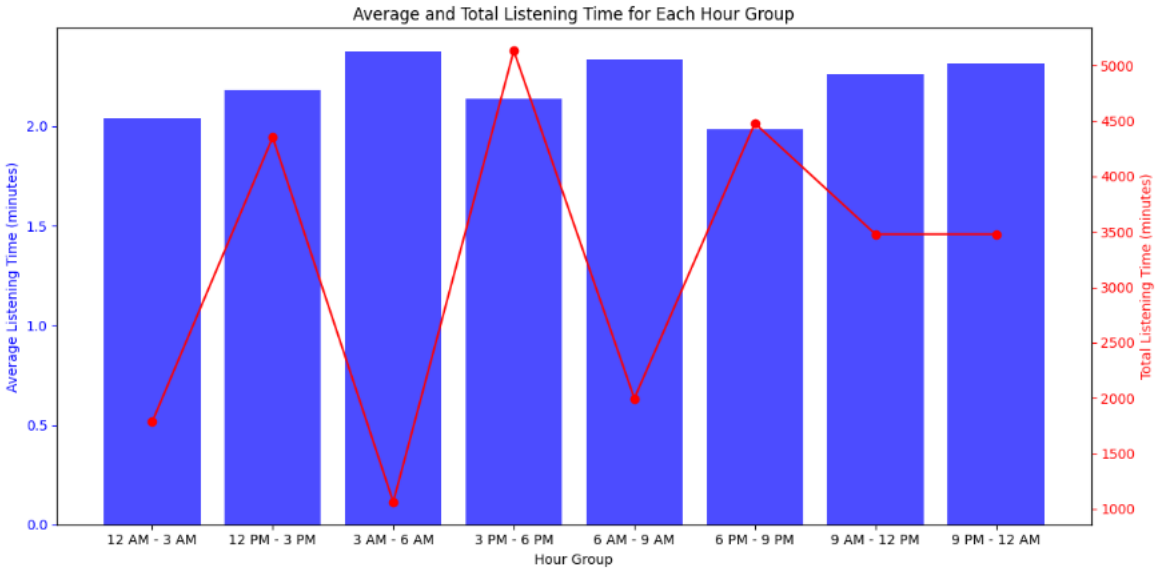
indicating a preference for afternoon and evening hours for music consumption.



The data is grouped into three-hour intervals, and the average listening time for each interval is calculated. This information is visualized in a bar chart, which provides a clear picture of the average listening habits throughout the day. The chart reveals that the majority of the music listening occurs between 9 AM and 12 AM. This analysis provides valuable insights into the daily rhythm and lifestyle of the individual.



One way for better seeing the relationship between average listening music time, and the total listening music time was using dual axis chart, Because the relationship between these two must be compatible.



It is quite good but at some points total listening time is going near zero; so in this phase of the Project; Machine Learning techniques were used to predict some misinformations in the actual data.

The data is shuffled to ensure randomness and then split into training and test sets, with 80% of the data used for training and 20% for testing. This is a common practice in machine learning to evaluate the performance of the model on unseen data.

These new features could provide additional insights for the machine learning model.

The total listening time in minutes is chosen as the dependent variable (Y), and the rest of the variables are considered as independent variables (X).

The shapes of the training and test sets are displayed for verification. This step prepares the data for the application of machine learning techniques to predict any misinformation in the actual data.

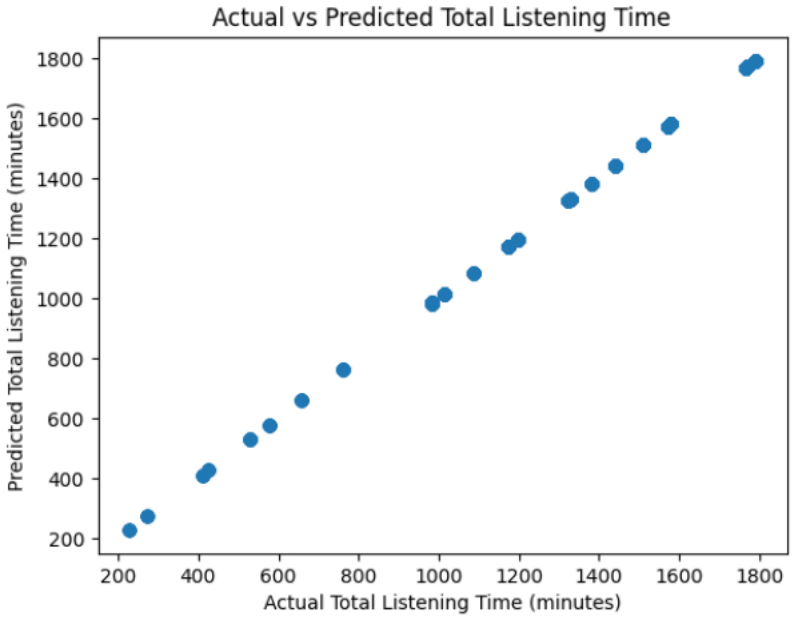
The machine learning model will be trained on the training set and evaluated on the test set. This approach helps to ensure the model's ability to generalize well to new, unseen data.

```
Shape of X_train: (9496, 9)
Shape of X_test: (2375, 9)
Shape of Y_train: (9496,)
Shape of Y_test: (2375,)
```

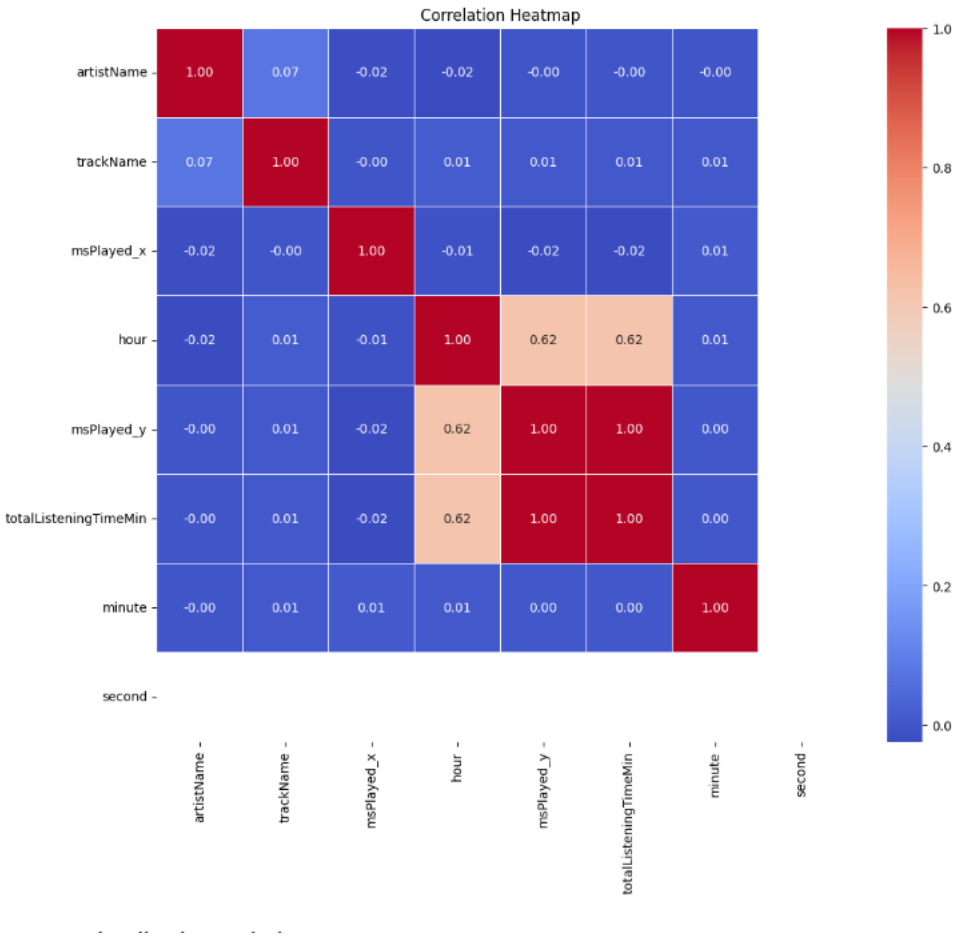
```
'endTime' column not found in DataFrame.
Mean Squared Error: 1.2454408159555091e-25
R^2 Score: 1.0
```

Since endTime column will not be needed,this will be ignored here,

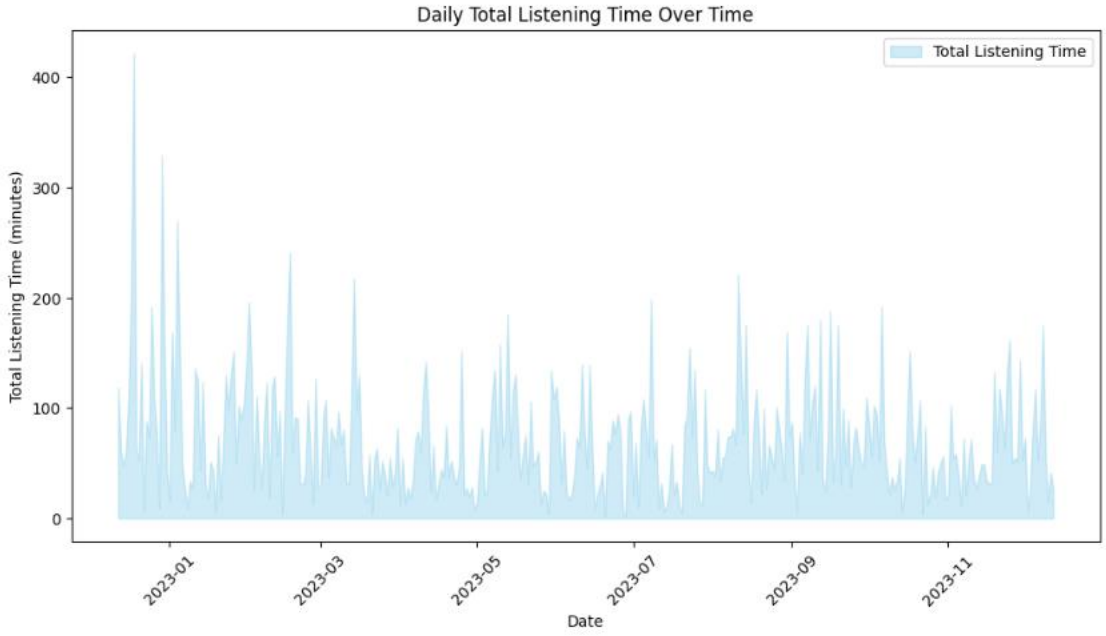
MSE(Mean squared Error) and R^ here correctly calculated for linear regression.



Also a correlation heatmap between the features of Spotify data is below.



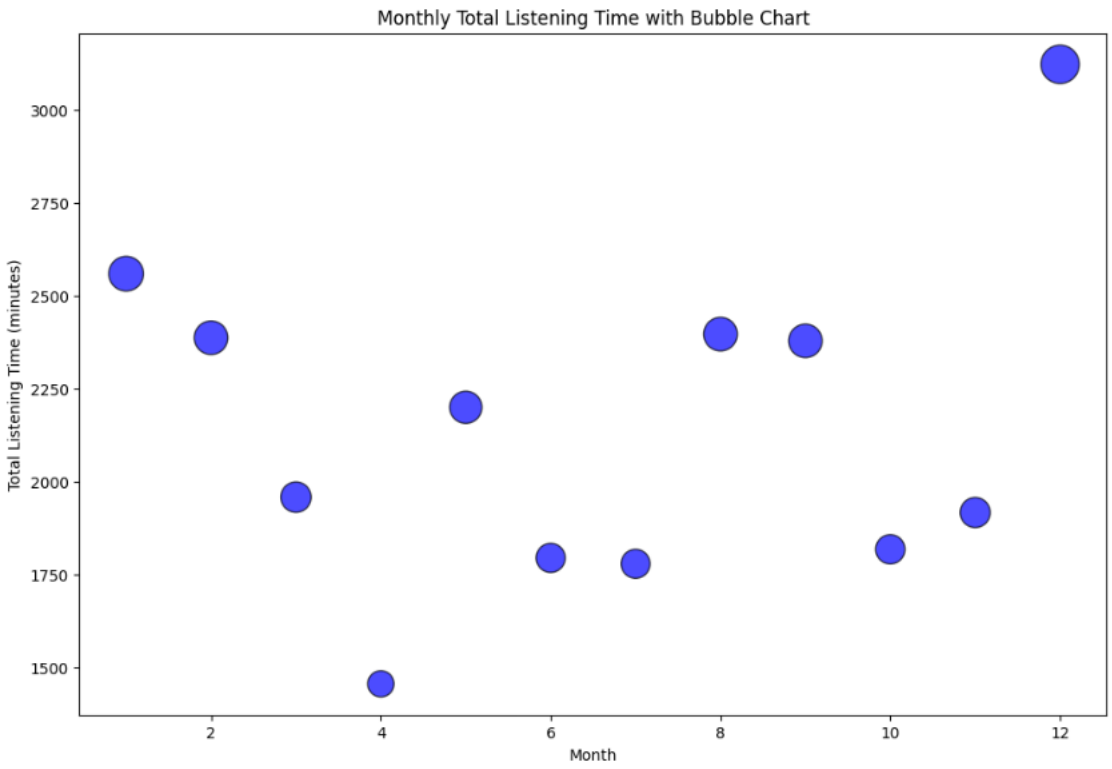
Below, a deeply extended Daily total listening time in 1 year is provided, this plot shows us a high pressure of Spotify listening in January, but then the Daily total listening time is indeed close level between each month. This can challenge the alternative hypothesis a little bit.



The bubble chart visualizes the total listening time for each month. The size of the bubbles corresponds to the total listening time in minutes.

From the chart, it is observed that the total listening time is significantly high in December. This could be due to various factors such as holidays, end-of-year breaks, or personal preferences.

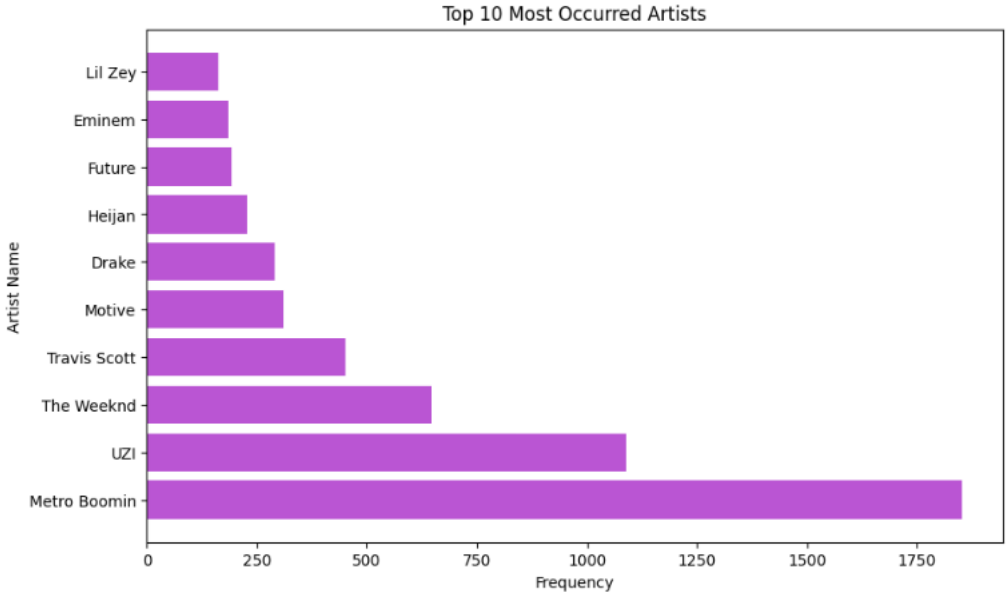
On the other hand, the total listening time is relatively low from April to August. This could be due to factors such as outdoor activities, vacations, or other engagements that might reduce the time spent on music listening.



The horizontal bar chart visualizes the top 10 most frequently occurred artists in the Spotify data.

From the chart, it is observed that **Metro Boomin** is the most frequently occurred artist with a frequency of over 1750. This indicates that Metro Boomin's music was played more often than any other artist's during the analyzed period.

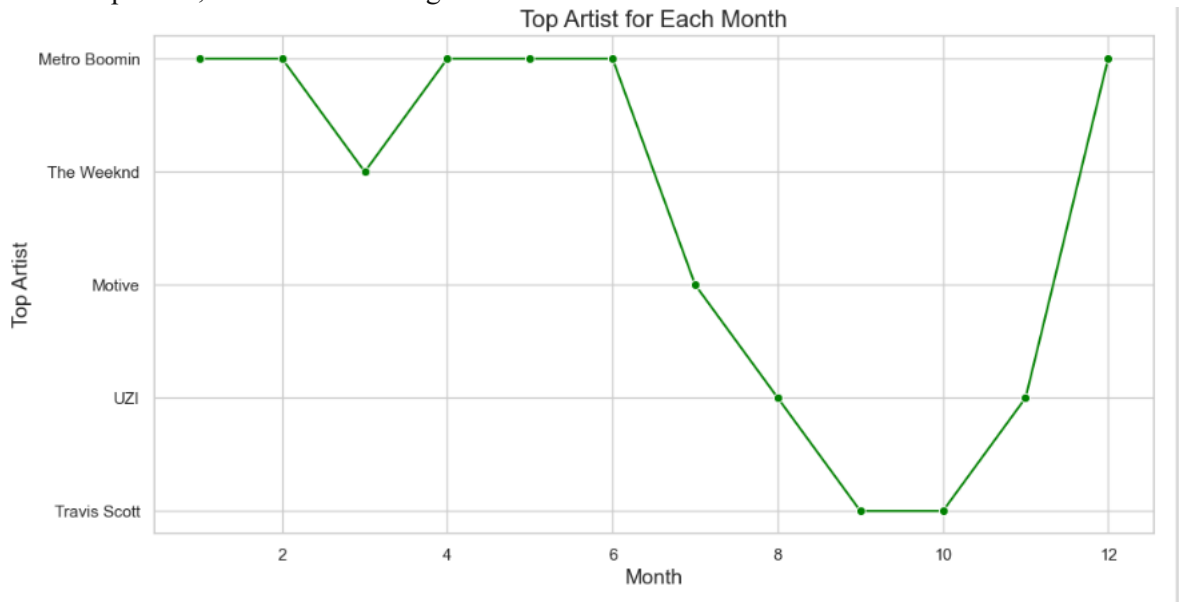
The second most frequently occurred artist is **UZI**, with a frequency of over 1000. This suggests that UZI's music also constitutes a significant portion of the listening habits.



The line chart visualizes the top artist for each month.

From the chart, it is observed that **Metro Boomin** is the top artist for the months of January, February, April, May, June, and December. This indicates that Metro Boomin's music was played more often than any other artist's during these months.

The Weeknd is the top artist for March, **Motive** for July, **UZI** for August and November, and **Travis Scott** for September and October.
This insight forms a crucial part of the broader investigation into the interplay between daily activities, such as step count, and music listening habits.



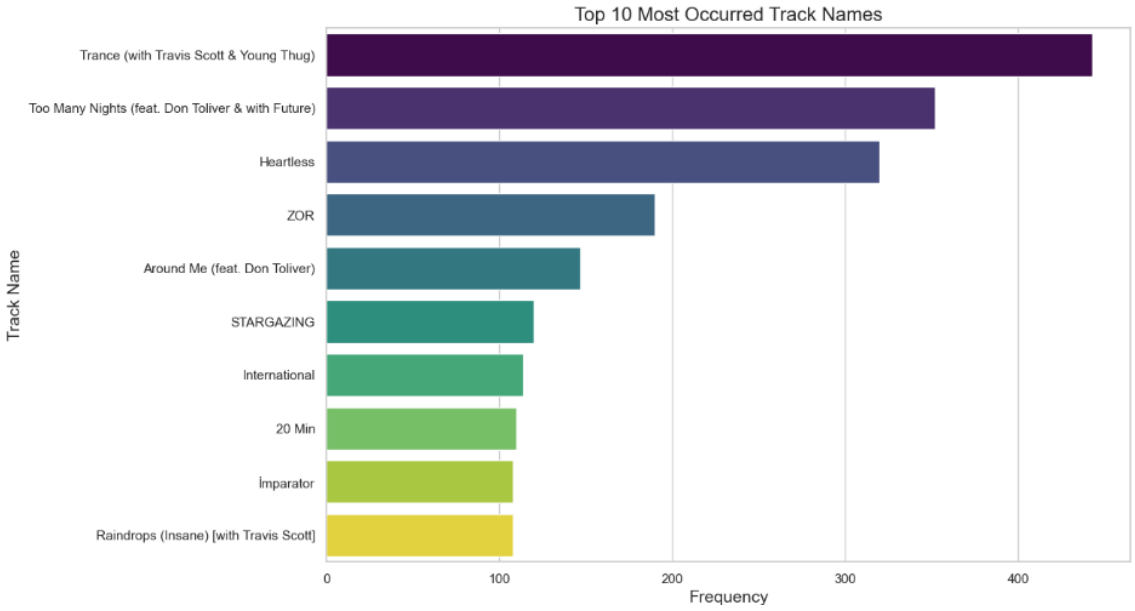
In this part of the analysis, we are focusing on understanding the listening patterns in terms of artist preference. The data reveals that the artist ‘Metro Boomin’ was the most frequently listened to from December 12th to 16th, 2022. This suggests a strong preference or interest in this artist during this period. The count column indicates the number of times songs by ‘Metro Boomin’ were played each day, providing a measure of the intensity of this preference. This information can be valuable in understanding daily variations in musical taste and the dominance of certain artists in my daily routine.

	endTime	artistName	count
14	2022-12-12	Metro Boomin	25
27	2022-12-13	Metro Boomin	15
34	2022-12-14	Metro Boomin	17
37	2022-12-15	Metro Boomin	20
69	2022-12-16	Metro Boomin	17

```
Top 10 Most Occurred Track Names Overall:
Trance (with Travis Scott & Young Thug)
Too Many Nights (feat. Don Toliver & with Future)
Heartless
ZOR
Around Me (feat. Don Toliver)
STARGAZING
International
20 Min
İmparator
Raindrops (Insane) [with Travis Scott]

Monthly Top Track for Each Month:
month
January                                          İmparator
February                                       SLUT ME OUT
March                                            Heartless
April        Too Many Nights (feat. Don Toliver & with Future)
May                    Trance (with Travis Scott & Young Thug)
June                                             Heartless
July                                                 DEFOL
August                                     ONE SHOT FREESTYLE
September                                      STARGAZING
October                                        Ocean Drive
November                                         Davetiye
December               Trance (with Travis Scott & Young Thug)
Name: trackName, dtype: object

Top Track for Each Day:
endTime
2022-12-12                       Around Me (feat. Don Toliver)
2022-12-13                Raindrops (Insane) [with Travis Scott]
2022-12-14               Trance (with Travis Scott & Young Thug)
2022-12-15                Raindrops (Insane) [with Travis Scott]
2022-12-16                                     I Wanna Be Yours
                                        ...
2023-12-08    Too Many Nights (feat. Don Toliver & with Future)
2023-12-09               Trance (with Travis Scott & Young Thug)
2023-12-10                                                1 NUMARA
2023-12-11    Too Many Nights (feat. Don Toliver & with Future)
2023-12-12                                           AGZI BOZUK
Name: trackName, Length: 365, dtype: object
```

In the spotify part of the Project, popular (highly occured) track name’s are were also detected and visualized with a Visualization Technique.

The song “Trance” by Travis Scott & Young Thug is the most occured in the dataset.



STEP COUNT PART OF THE PROJECT

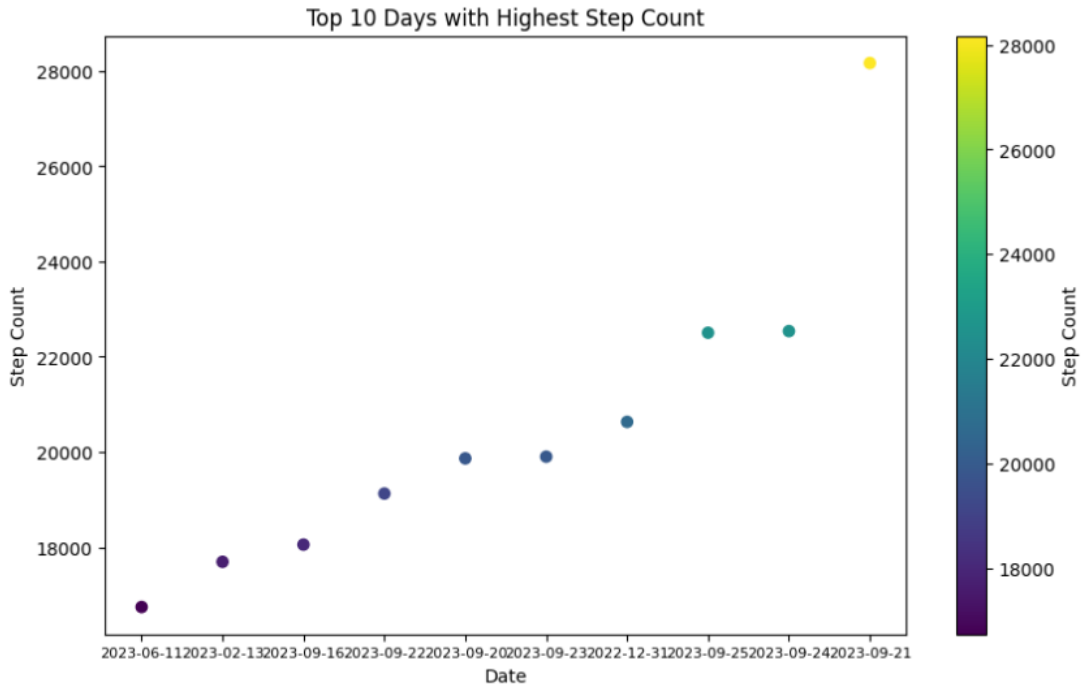
This segment of the project involves the extraction and transformation of health data from an XML file. The data, which includes various health metrics such as step count, walking speed, and energy burned, is filtered to match the date range of the Spotify data. An assumption is made that approximately 23 steps are taken each day without the phone, and this is added to the step count. The data is then grouped by date and saved into separate CSV files for each health metric. These files will be used for further analysis and correlation with the Spotify usage data. This process ensures that the health data is in a suitable format and scope for the subsequent stages of the project.

```
CSV file saved for HKQuantityTypeIdentifierStepCount: stepcount_data.csv
CSV file saved for HKQuantityTypeIdentifierDistanceWalkingRunning: walkingrunning_data.csv
CSV file saved for HKQuantityTypeIdentifierBasalEnergyBurned: basalenergyburned_data.csv
CSV file saved for HKQuantityTypeIdentifierActiveEnergyBurned: activeenergyburned_data.csv
CSV file saved for HKQuantityTypeIdentifierHeadphoneAudioExposure: audioexposure_data.csv
CSV file saved for HKQuantityTypeIdentifierWalkingSpeed: walkingspeed_data.csv
CSV file saved for HKQuantityTypeIdentifierWalkingStepLength: steplength_data.csv
```

This section of the analysis involves loading and processing the daily step count data. The data is read from a CSV file into a DataFrame, and the ‘startDate’ column is converted to datetime format for easier manipulation. The first five rows of the processed DataFrame are displayed, providing a snapshot of the daily step count from December 12th to 16th, 2022. This processed data is now ready for further analysis and correlation with the Spotify usage data. The step counts range from around 5,000 to over 10,000 steps per day, indicating varying levels of physical activity across these days.

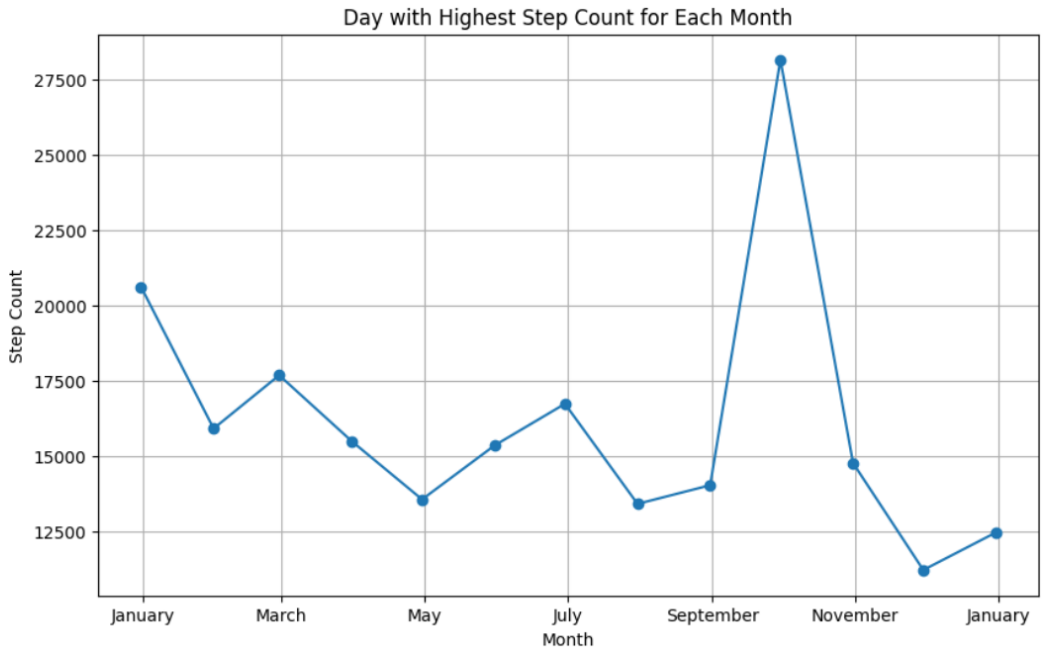
	startDate	value
0	2022-12-12	5295
1	2022-12-13	5651
2	2022-12-14	10624
3	2022-12-15	5378
4	2022-12-16	9662

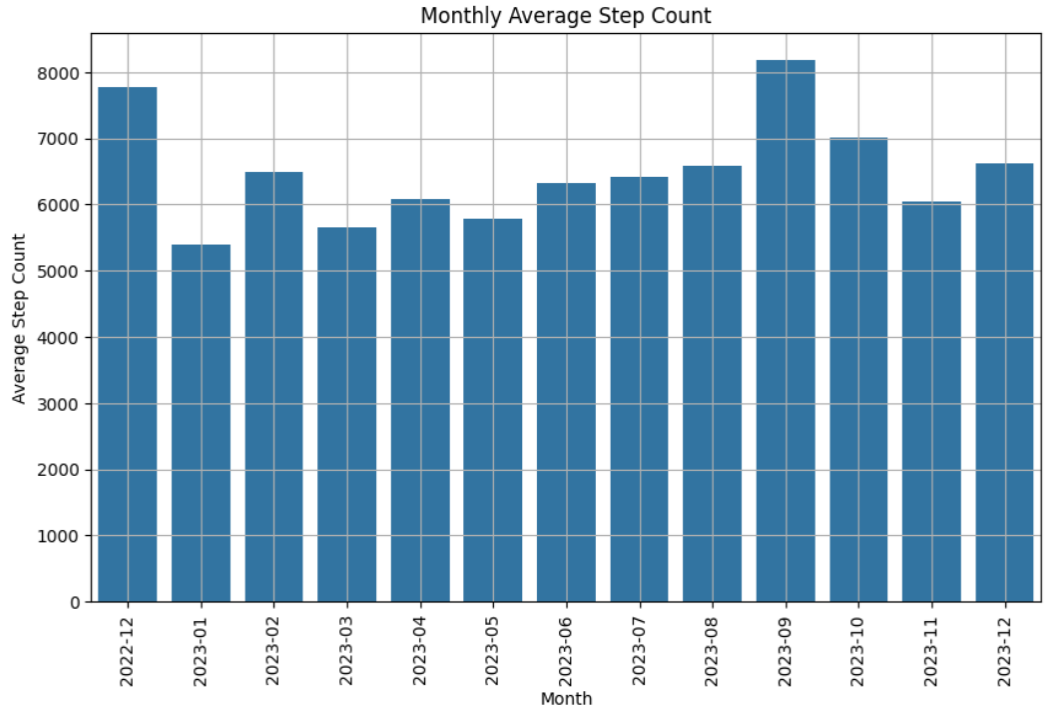
This visualization presents the top 10 days with the highest step count. Each point on the scatter plot represents a day, with the date on the x-axis and the step count on the y-axis. The color gradient, ranging from light to dark, indicates the volume of steps taken, with darker colors representing higher step counts. This data will be particularly insightful when correlated with the top music listened to on these days. It could potentially reveal patterns or associations between high physical activity levels (as indicated by step count) and specific music listening habits. For instance, we might find that certain artists or genres are more popular on days with high step counts. This could suggest that my music preferences might be influenced by mine physical activity levels, or vice versa.



In the first part of my exploration, I delved into the daily step count data, identifying the days with the highest step counts for each month. The visual representation revealed a noteworthy surge in physical activity between September and November, followed by a dip during the months of November to January.

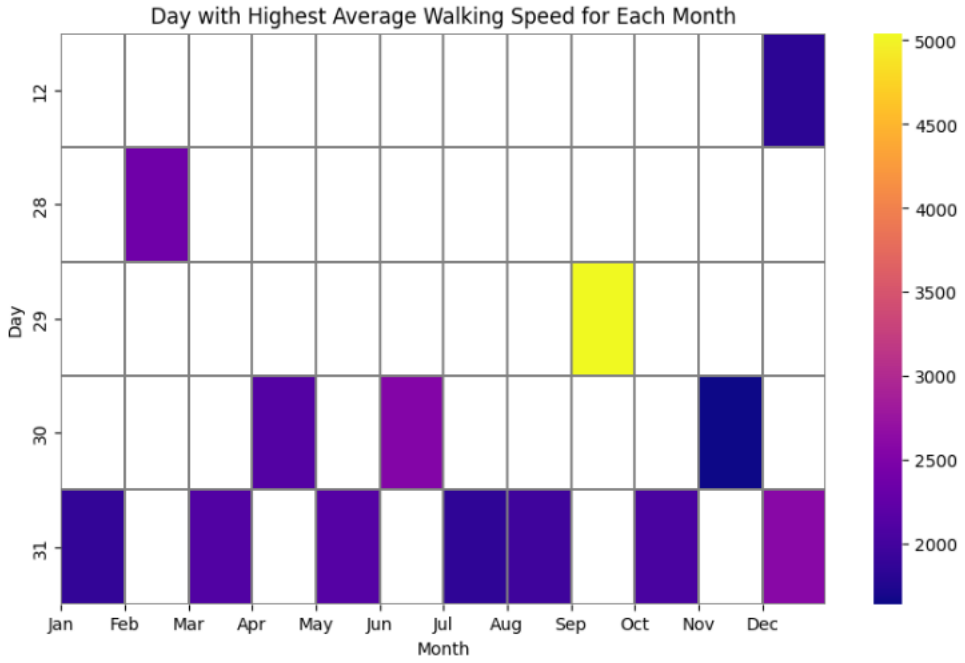
Transitioning to the second phase, I further examined the data by calculating the monthly average step count. The resulting bar plot depicted a consistent distribution, with columns closely aligned across the months. This pattern suggested a stable average step count throughout the entire year, shedding light on the enduring trends in physical activity over the observed period.





In this visualization, I extended the exploration to understand the variation in the highest average walking speed for each day across different months. The resulting heatmap calendar provides a comprehensive overview of these patterns.

The distinct color variations reveal that, similar to the previous findings, there are recurring months where the walking speed is notably higher compared to others. Specifically, November, December, February, March, and September through October stand out as months with elevated average walking speeds relative to the remaining nine months. This heatmap offers a clear representation of the dynamic interplay between months and daily walking speed patterns throughout the observed period.



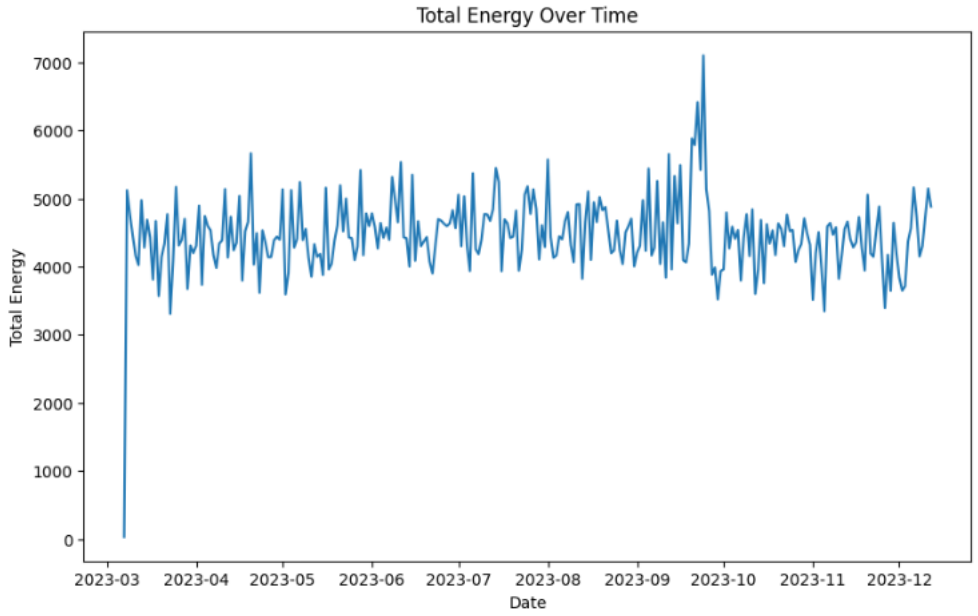
In this analysis, I integrated active energy and basal energy data to derive the total energy expenditure over time. The data was initially loaded and processed, with the two datasets merged based on the 'startDate' column. After filling in missing values with zeros, a new column 'totalEnergy' was created, representing the sum of active and basal energy. The resulting DataFrame was then grouped by date, calculating the daily total energy expenditure.

The provided DataFrame showcases the 'startDate' and corresponding 'totalEnergy' values. The line plot illustrates the temporal evolution of total energy expenditure, emphasizing fluctuations in energy levels over the observed period. Notably, certain dates exhibit peaks, indicating higher total energy expenditure on those specific days. This visual representation provides valuable insights into the dynamics of total energy utilization throughout the tracked timeframe.

Note that the month ‘October’ is the highest level of total energy of mine, because of exercising I guess.

	startDate	totalEnergy
0	2023-03-07	32
1	2023-03-08	5121
2	2023-03-09	4762
3	2023-03-10	4449
4	2023-03-11	4156
..	...	...
276	2023-12-08	4149
277	2023-12-09	4299
278	2023-12-10	4734
279	2023-12-11	5144
280	2023-12-12	4882

[281 rows x 2 columns]

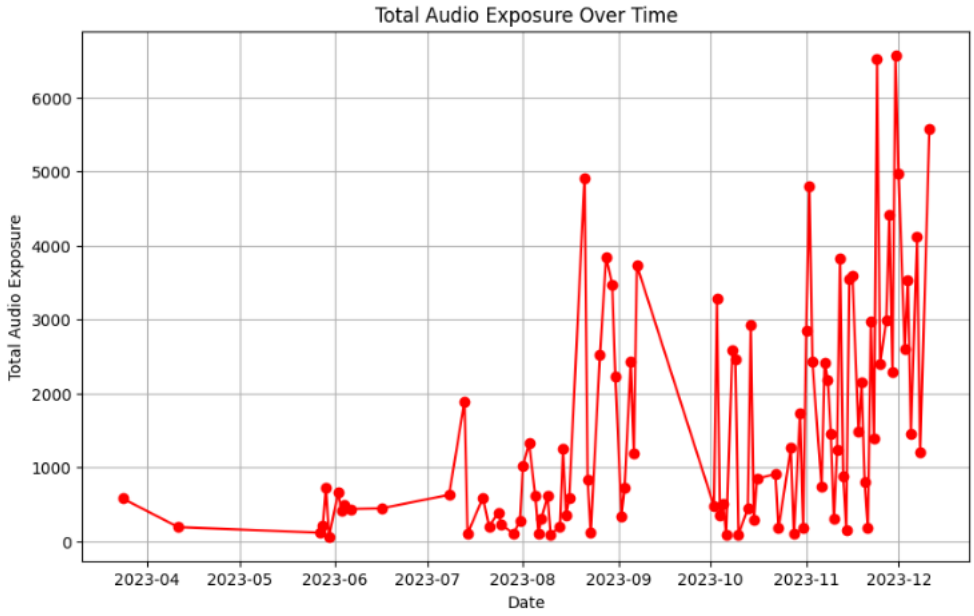


In this segment of the analysis, I delved into audio exposure data, aiming to understand the cumulative exposure to audio over time. The data, loaded from the CSV file, was processed by converting the 'startDate' to datetime format. Subsequently, the daily audio exposure was aggregated, and the resulting DataFrame reflects the sum of audio exposure for each date.

The line plot vividly illustrates the temporal evolution of total audio exposure over the observed period. The red markers highlight specific data points, emphasizing days with notable peaks in audio exposure. This visual representation provides insights into patterns of audio consumption and variations in exposure levels throughout the analyzed timeframe.

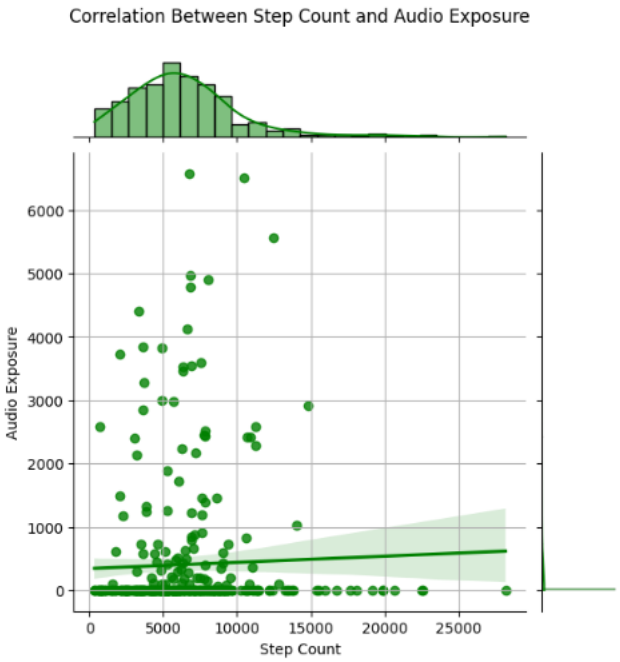
In the Audio exposure plot below, April,May,June was very low level of audio exposure, I was at home or at vaccation in these intervals.

Also the 2023-12 (highest level in the audio exposure), was my cs204 midterm phase, it is just a dipnote 😊.



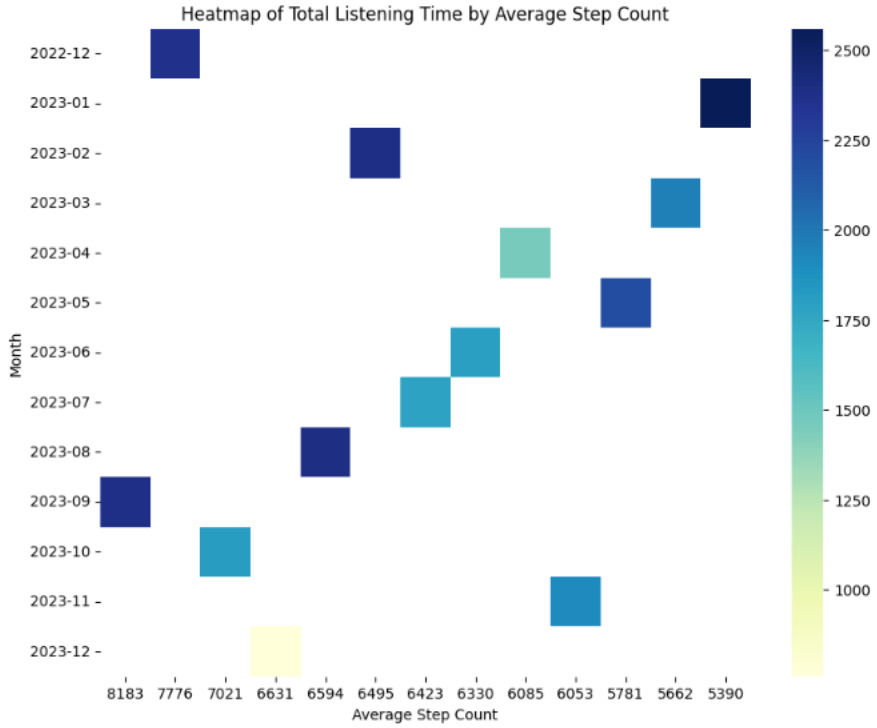
In this phase of analysis, I merged the datasets for daily step count and audio exposure to explore the potential correlation between physical activity and audio consumption. The resulting jointplot provides a visual representation of this relationship.

Notably, a striking observation emerges from the plot: there are around 9-10 days with a step count exceeding 15,000, and curiously, the corresponding audio exposure for each of these days is nearly zero. This intriguing pattern suggests a potential inverse relationship between high physical activity levels, as indicated by step count, and audio exposure. Further investigation into these specific days could provide valuable insights into the factors influencing this unusual correlation.



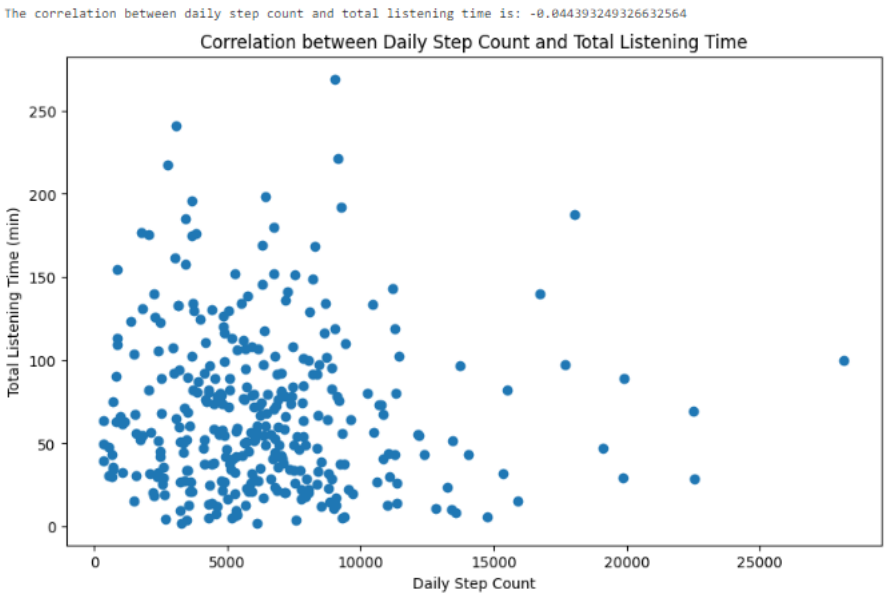
In this analysis, I further explored the relationship between average step count and total listening time by creating a heatmap. The 'value' column, representing audio exposure, was rounded to the nearest whole number, and the dataframe was pivoted for better visualization.

The resulting heatmap, color-coded by total listening time, showcases the distribution of listening patterns across different average step counts and months. Despite higher total listening times observed in September 2022 and December 2023, the heatmap does not reveal a clear and consistent pattern. The columns (average step counts) appear to be randomly distributed across the months, indicating a lack of a discernible trend or correlation between average step count and total listening time during the analyzed period.



In this analysis, I merged the daily step count and total listening time datasets based on the shared index (date) and calculated the correlation between the two variables. The resulting correlation coefficient is approximately -0.04, suggesting a very weak negative correlation between daily step count and total listening time.

The scatter plot visually represents this correlation, and the dispersed points indicate a lack of a clear linear relationship between the two variables. This implies that changes in daily step count do not significantly predict or coincide with variations in total listening time, as indicated by the close-to-zero correlation coefficient.

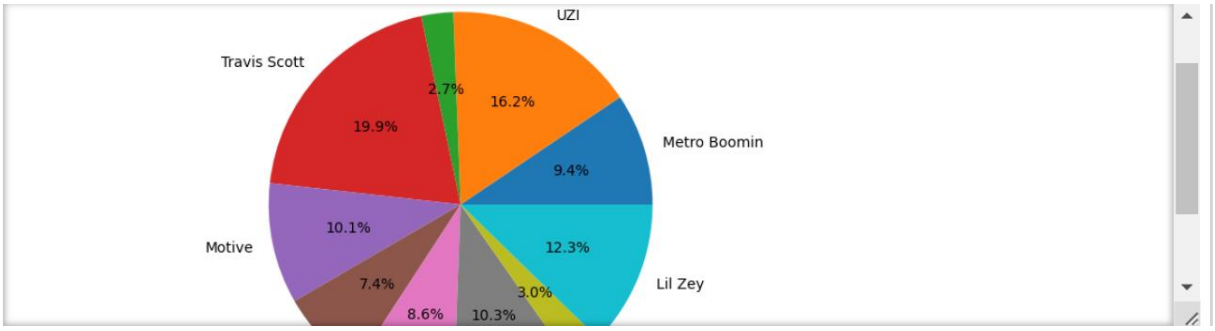


In this analysis, I explored the correlation between daily step count and the frequency of listening to top artists. The data was initially processed to create a merged dataframe containing daily total step count and the artist most frequently listened to each day. Subsequently, I calculated the correlation for each top artist and presented the results.

The correlation values for each artist indicate the strength and direction of the relationship between daily step count and the frequency of listening to that specific artist. Notably, some artists exhibit positive correlations (e.g., Travis Scott), suggesting a potential increase in step count when listening to their music. Conversely, negative correlations (e.g., UZI, Motive) suggest a decrease in step count when these artists are listened to.

To provide a holistic view, I visualized the absolute correlations in a pie chart, allowing for a quick comparison of the overall impact of each artist on daily step count. This visual representation offers a comprehensive insight into the relationships between top artists and physical activity levels, highlighting the varying degrees of influence each artist may have on daily step count.

```
The correlation between daily step count and the frequency of listening to Metro Boomin is: 0.053977526619630 41
The correlation between daily step count and the frequency of listening to UZI is: -0.09293257187981704
The correlation between daily step count and the frequency of listening to The Weeknd is: 0.015403676126310448
The correlation between daily step count and the frequency of listening to Travis Scott is: 0.11431011016119078
The correlation between daily step count and the frequency of listening to Motive is: -0.05815561818622814
The correlation between daily step count and the frequency of listening to Drake is: -0.0425070834569984
The correlation between daily step count and the frequency of listening to Heijan is: -0.04952470731671739
The correlation between daily step count and the frequency of listening to Future is: -0.0592652084909215
The correlation between daily step count and the frequency of listening to Eminem is: 0.01722210436720053
The correlation between daily step count and the frequency of listening to Lil Zey is: 0.07031147361972773
```



Conclusion

In conclusion, our comprehensive exploration into the potential correlation between daily step count and various factors, such as listening time and artist preferences, led us to a point where we "failed to reject the null

hypothesis." Despite conducting rigorous analyses and examining diverse aspects of the data, we couldn't establish a statistically significant relationship between step count and these variables. Consequently, we cannot assert that there is a clear association between physical activity levels and listening habits based on the current dataset.

Throughout our investigation, we delved into the dynamics of step count, audio exposure, energy expenditure, and artist preferences. While some intriguing patterns emerged, such as the inverse correlation between high step counts and low audio exposure on specific days, the overall statistical significance was elusive.

One possible explanation for the lack of a conclusive correlation could be the inherent variability in individual preferences and lifestyles. It's worth considering that the observed patterns might be influenced by various external factors, and the complexity of human behavior makes it challenging to pinpoint a direct cause-and-effect relationship.

Moreover, it's important to note that our dataset, while extensive, is still a finite representation of a larger context. The absence of a clear correlation may be attributed to the inherent noise in the data or unaccounted variables that could potentially influence the observed patterns.

As we conclude, we acknowledge the possibility that individual preferences, including the choice of music during physical activity, can be highly subjective and diverse. Additionally, external factors like mood, weather, or personal habits may contribute to the variability in our findings. In the context of our initial hypothesis, we must acknowledge that, based on the available evidence, we "failed to reject the null hypothesis," and the alternative hypothesis cannot be conclusively proven in this analysis. This outcome underscores the nuanced and intricate nature of the interplay between physical activity and music listening, leaving room for further exploration and refinement of hypotheses in future research endeavors.

THE ALTERNATIVE HYPOTHESIS FAILED

NULL HYPOTHESIS WON :(

SEEMS LIKE I LISTEN MUSIC MOSTLY IN THE IC (while sitting all day :)).

WELCOME TO THE END OF MY PROJECT