Ece Ayda Kırmızıgül - DSA210 Project Report

Introduction

This project, conducted as part of the Sabancı University CS210 Introduction to Data Science course, examines the relationship between music listening habits and physical activity by analyzing two personalized datasets: Spotify Listening Data and Apple Health Step Count Data, each spanning one year. The objective is to explore potential patterns and correlations between daily step counts and music listening behavior, offering deeper insights into how these aspects of daily life intersect.

Project Motivation

As someone who values both music and maintaining an active lifestyle, I am intrigued by the potential connection between my activity levels (specifically step counts) and my music preferences. This project seeks to identify whether variations in daily step counts influence the type or duration of music I listen to, or if my listening habits affect my activity levels. By delving into these datasets, the goal is to uncover patterns that reflect how physical activity and listening behavior interact, providing meaningful insights into personal habits.

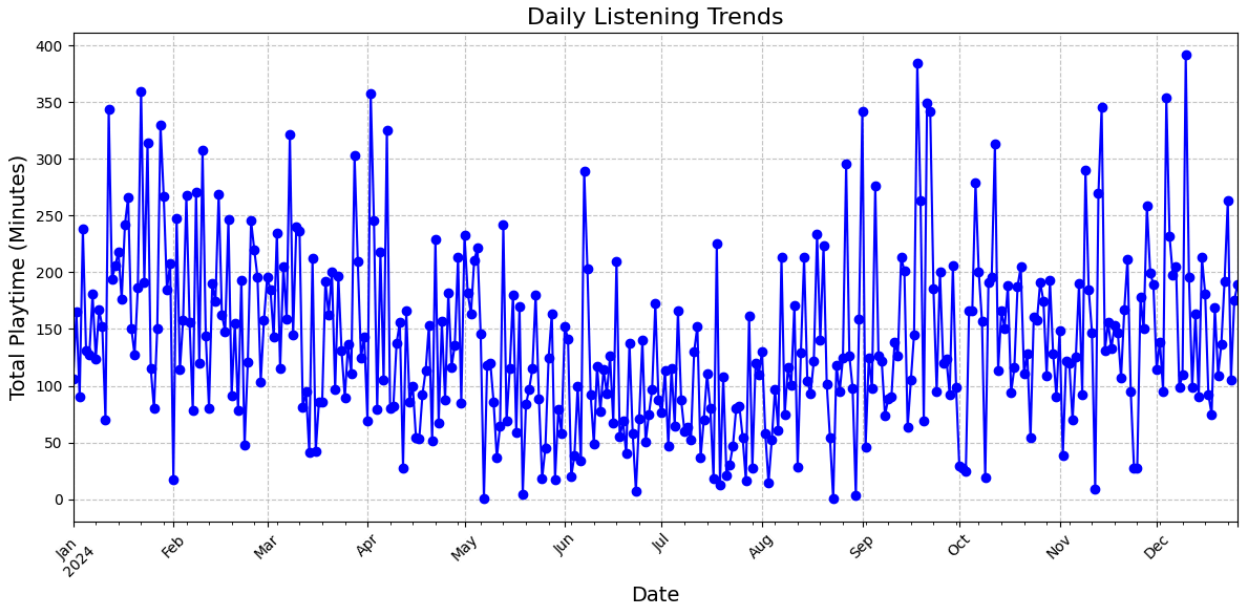
Data Sources

For this project, I used two datasets: the first is my Spotify listening data taken from my Spotify account, and the second is my Apple Health data taken from my Apple Health account.

Spotify Data

A full year of personal music listening history, downloaded via Spotify's data export feature. The data was provided in JSON format, which I opened and explored to make it usable for this project. This dataset includes detailed information about my listening habits on Spotify, such as track names, artists, and listening durations. Relevant portions of the data were analysed and visualized to align with the project's aim. Here are some important visuals that show my dataset.

Daily Listening Trends

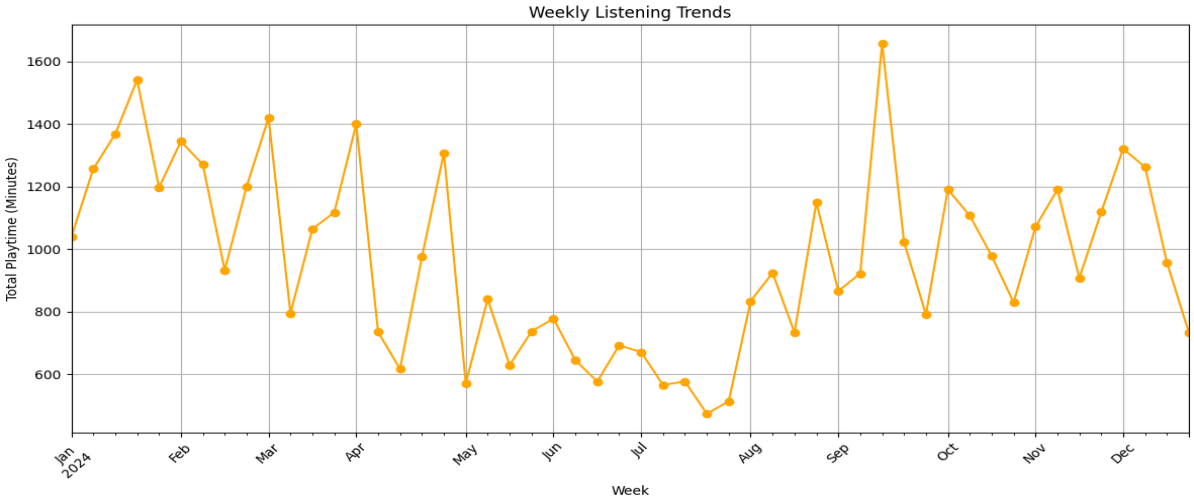


This line chart provides a detailed view of daily Spotify listening duration in minutes over the course of one year. The x-axis represents individual dates from January to December, and the y-axis shows the total playtime in minutes. Each point on the chart corresponds to a specific day, capturing the variations in listening habits at a granular level.

The chart reveals significant fluctuations in daily listening patterns, with occasional spikes where listening time exceeds 300 minutes, suggesting periods of heightened engagement with music. These spikes could correspond to specific events, such as long commutes, dedicated relaxation time, or specific emotional states that prompted extended listening sessions. Conversely, troughs, where listening time drops below 50 minutes, might reflect busier days or shifts in focus away from music.

Notable trends include periods of consistently high or low listening durations, which could align with certain times of the year, such as vacations or exam periods. Additionally, there appears to be greater variability in listening times during the earlier months of the year, which gradually stabilizes as the year progresses. This visualization provides a rich foundation for exploring correlations between daily activity levels and listening behavior.

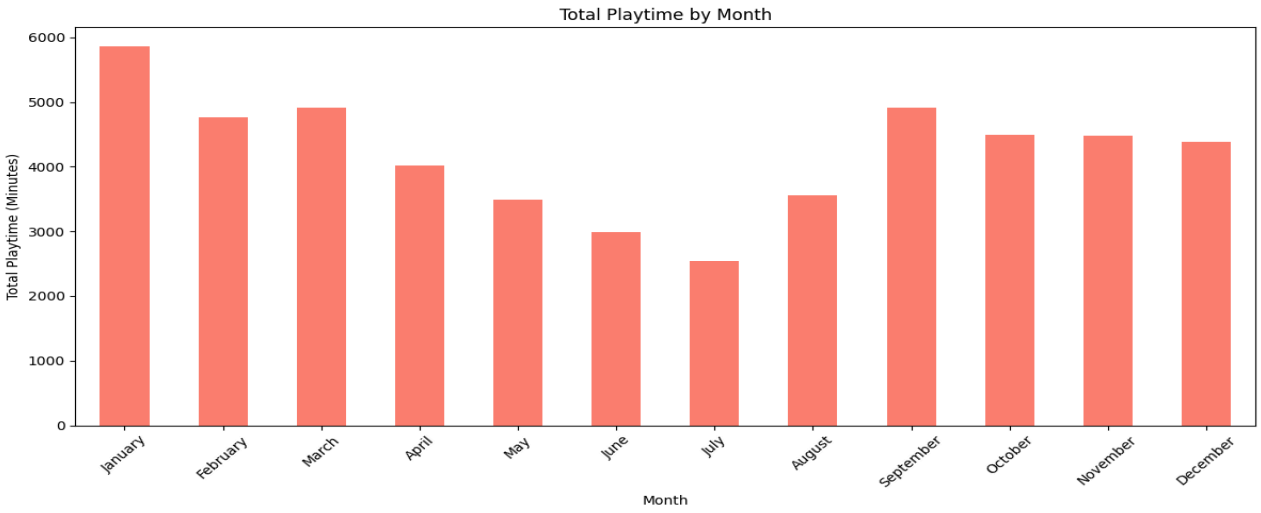
Weekly Listening Trends



This visualization represents the total Spotify listening duration in minutes for each week over the span of one year. The x-axis displays the weeks chronologically from January to December, while the y-axis shows the total playtime in minutes. The graph highlights fluctuations in weekly listening habits, revealing potential patterns or significant changes in music consumption behavior over time.

Notably, certain peaks and troughs indicate periods of increased or decreased activity, which could be associated with external factors such as holidays, workload variations, or personal routines. This visual serves as a foundation for analyzing trends in music listening habits and their possible connection to physical activity levels.

Total Playtime by Month

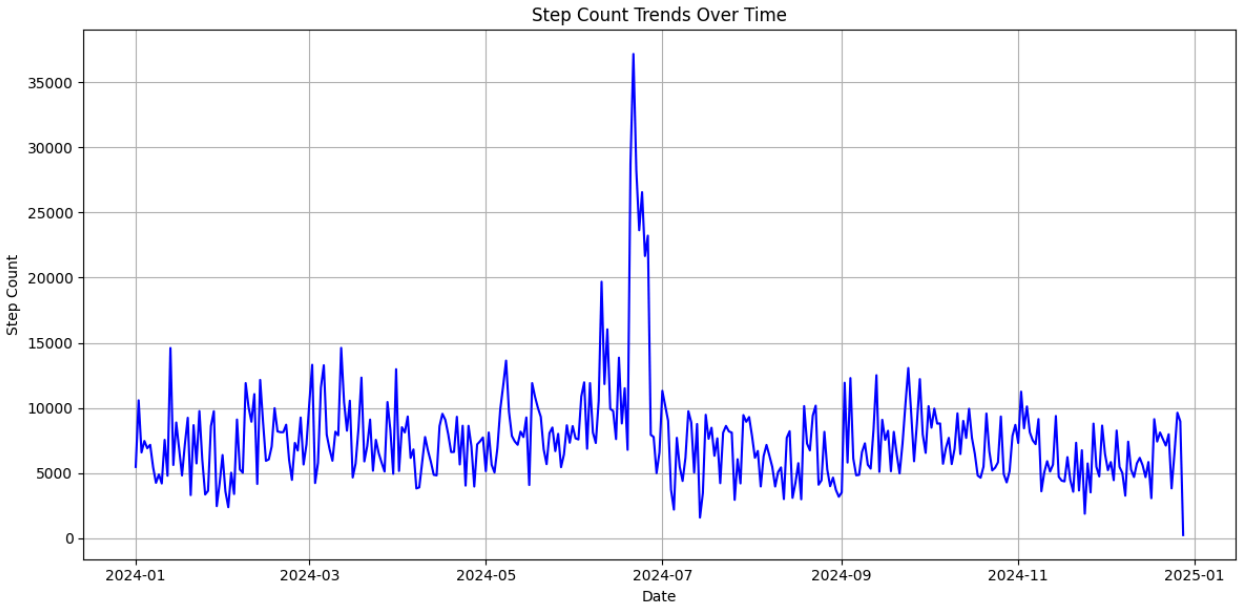


This bar chart aggregates Spotify listening duration by month, providing a high-level view of listening habits throughout the year. The x-axis lists the months from January to December, while the y-axis represents total playtime in minutes. This visualization highlights seasonal trends, such as higher listening times during colder months like January and February and lower durations during summer months like July, possibly due to changes in routines or vacation periods.

Apple Health Data

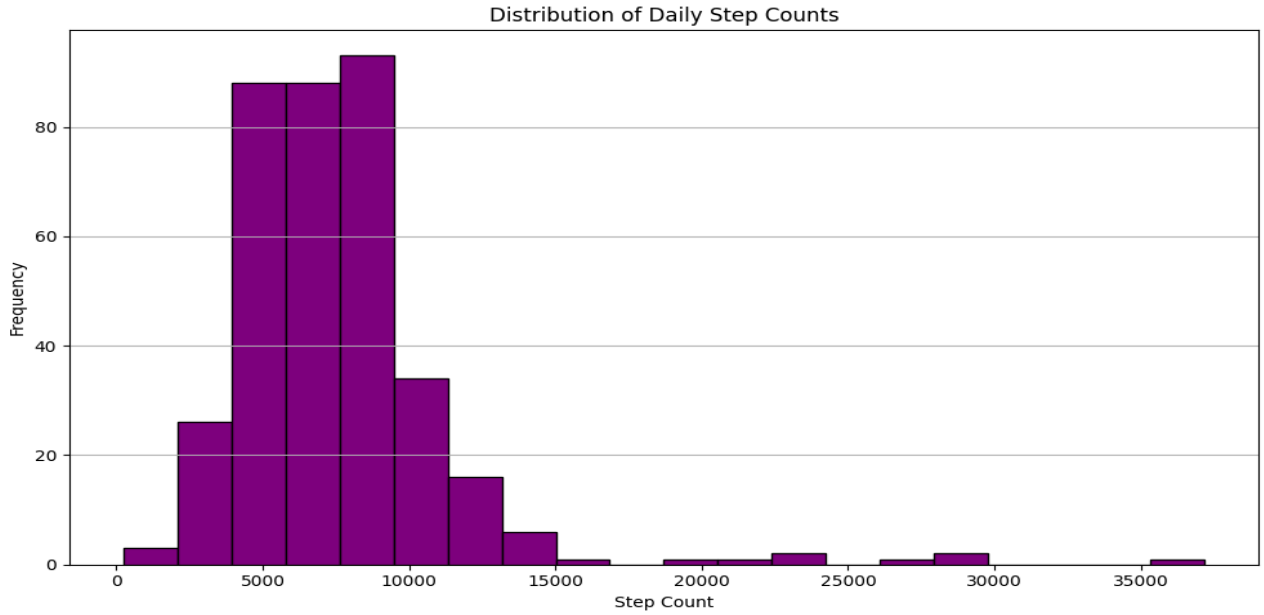
The second dataset consists of a full year of daily step count records, retrieved from my Apple Health account using the iPhone's data export tools. This dataset provides a detailed record of physical activity measured in steps. The data was initially exported in a structured format and then processed to align with the goals of this project. Key details, such as daily step counts and corresponding dates, were extracted to enable comparison with the Spotify listening data.

Step Count Trends Over Time



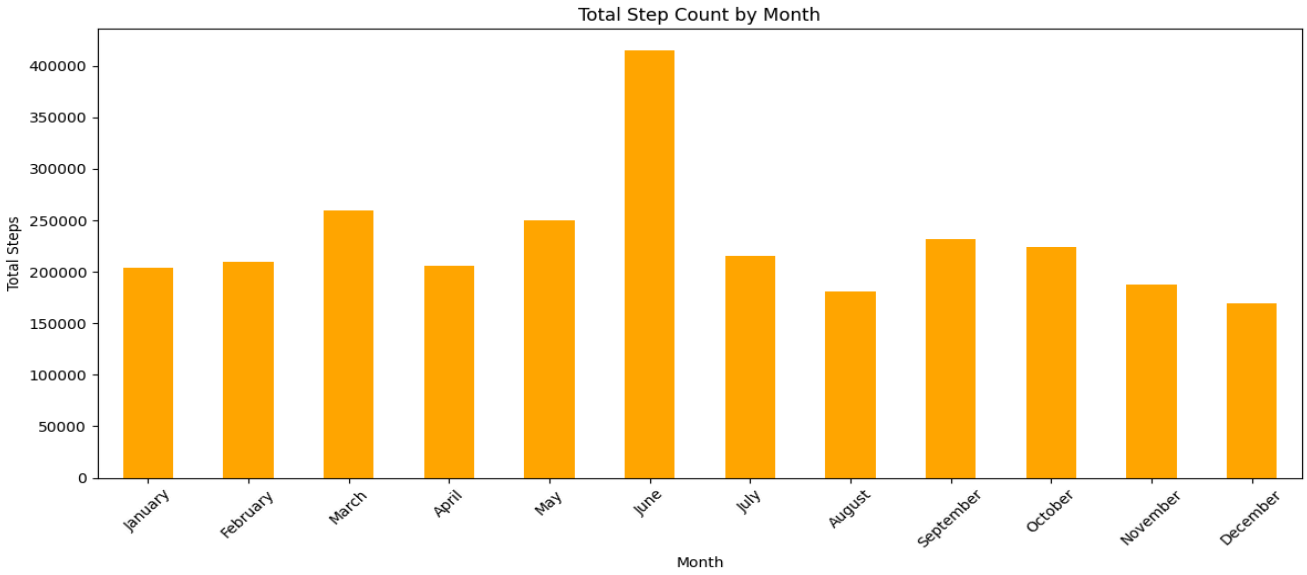
This line chart displays the daily step counts over the entire year. The x-axis represents the dates, while the y-axis shows the step count for each day. The chart captures day-to-day fluctuations in physical activity, with noticeable peaks indicating exceptionally active days and troughs representing periods of lower activity. The significant spike around mid-year suggests a specific event or period of heightened activity, such as participation in an event or a fitness challenge. Meanwhile, consistent patterns of moderate activity may reflect established daily routines.

Distribution of Daily Step Counts



This histogram illustrates the distribution of daily step counts recorded over the year. The x-axis represents step count ranges, while the y-axis shows the frequency of days within each range. The data reveals that most daily step counts fall between 5,000 and 10,000 steps, aligning with common activity goals for maintaining a healthy lifestyle. The skewed distribution, with fewer days exceeding 15,000 steps, highlights occasional spikes in physical activity. Conversely, the presence of lower step counts indicates days with reduced mobility, which might align with rest days or specific circumstances such as bad weather or illness.

Total Step Count by Month



This bar chart aggregates the total number of steps taken each month throughout the year. The x-axis lists the months, and the y-axis represents the cumulative step count for each month. The data shows that June had the highest total step count, potentially corresponding to outdoor activities or vacation periods. In contrast, months like December and July exhibit relatively lower totals, possibly due to seasonal changes, holidays, or other personal factors. This visualization provides a clear overview of step count trends at a higher temporal resolution, helping to identify potential seasonal patterns in physical activity.

Correlation Analysis and Hypothesis Testing

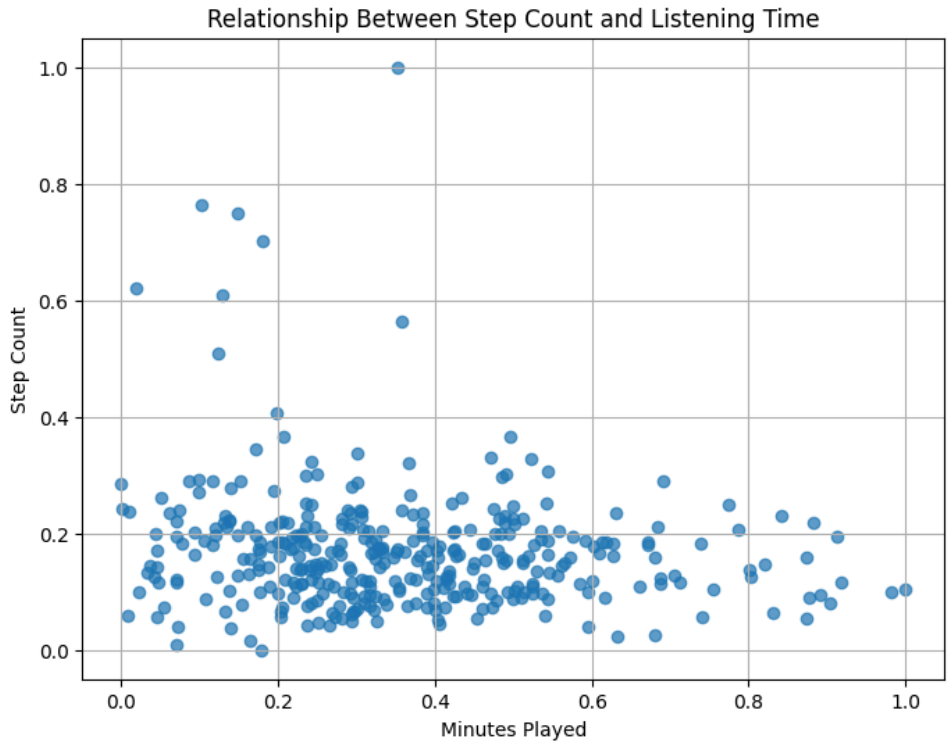
The next phase of this project involves exploring the relationship between daily step counts and music listening habits through a detailed correlation analysis. This section focuses on merging the datasets, scaling them appropriately, and investigating the potential connection between the two variables. By utilizing statistical methods such as correlation analysis, scatter plots, and linear regression, we aim to quantify and visualize the degree of association between step counts and listening minutes. Additionally, hypothesis testing is employed to validate the statistical significance of the observed correlations.

Key steps in this process include:

- **Merging and Scaling the Data**: Combining the datasets to align step counts with corresponding listening minutes and normalizing them for analysis.
- **Investigating the Connection**: Analyzing scatter plots and trends over time to uncover visual patterns in the relationship.
- **Hypothesis Testing and Linear Regression**: Using statistical tests to establish the strength and reliability of the correlation, followed by regression models to predict one variable based on the other.

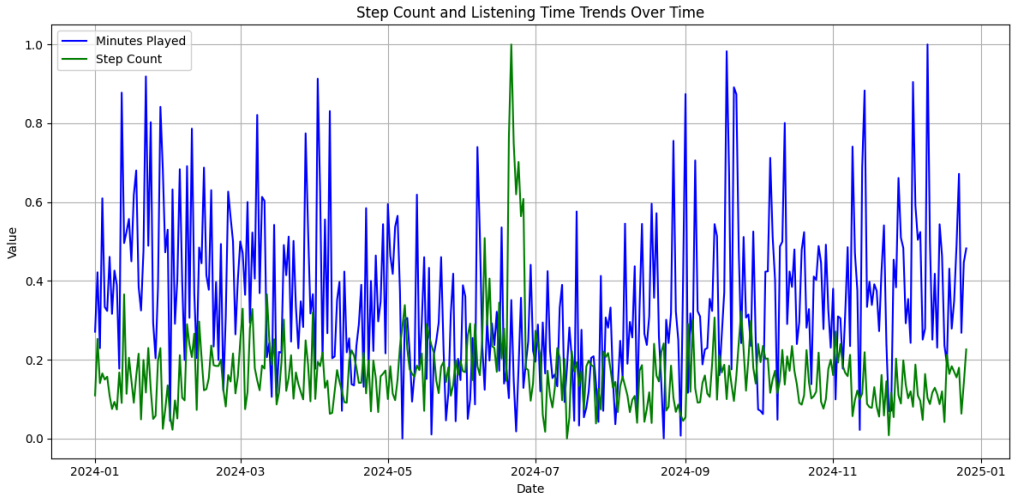
This section lays the groundwork for understanding how physical activity and music listening habits might influence each other, supported by quantitative evidence and visualization.

Scatter Plot: Relationship Between Step Count and Listening Time



This scatter plot illustrates the relationship between daily step counts and minutes of music listening. Each point represents a single day's data, with the x-axis showing normalized minutes played and the y-axis showing normalized step counts. The distribution of points suggests a weak or negligible direct correlation, as no clear linear trend is observed. Some clustering in specific areas may indicate shared tendencies or constraints in daily behavior.

Trend Over Time: Step Count and Listening Time



This line chart shows the normalized trends of step counts and music listening time over the year. The x-axis represents the dates, while the y-axis represents the normalized values of step counts and listening minutes. The two lines provide a comparative view, allowing us to observe how these variables fluctuate over time. While some overlapping peaks and troughs hint at a potential connection, the overall trends seem independent, suggesting that step counts and listening time are largely influenced by different factors.

Correlation and Statistical Analysis: Step Count vs. Listening Minutes

To evaluate the relationship between step count and listening minutes, multiple statistical tests were conducted, starting with hypothesis formulation and proceeding to correlation, t-tests, and chi-square tests.

Hypotheses

- Null Hypothesis (H_0): There is no correlation between step count and listening minutes.
- Alternative Hypothesis (H_1): There is a significant correlation between step count and listening minutes.

Results and Interpretations

1. **Pearson Correlation**
    - **Correlation Coefficient (r):** -0.1536
    - **P-value:** 0.0034
    - **Interpretation:** The Pearson correlation indicates a weak negative but statistically significant relationship between step count and listening minutes. Although the correlation is weak, the result suggests that as step count increases, listening minutes slightly decrease, or vice versa.
2. **T-Test**
    - **T-statistic:** -1.6434
    - **P-value:** 0.1012
    - **Interpretation:** The t-test results show no statistically significant difference in listening minutes between groups with high and low step counts. This suggests that grouping data into high and low step count categories does not reveal a substantial effect on listening behavior.
3. **Chi-Square Test**
    - **Chi²-statistic:** 3.5912
    - **P-value:** 0.4641
    - **Interpretation:** The chi-square test result indicates no significant association between categorized step counts and listening minutes. This reinforces the

notion that there is minimal interaction between these variables when grouped categorically.

Linear Regression Analysis

To further investigate the relationship, an Ordinary Least Squares (OLS) regression model was applied, with step count as the dependent variable and listening minutes as the independent variable.

1. **Regression Summary**
    - **R-squared:** 0.024
    - **Adjusted R-squared:** 0.021
    - **F-statistic:** 8.670
    - **P(F-statistic):** 0.003
    - **Coefficient for Minutes Played:** -0.0819 (P-value = 0.003)
2. **Interpretation**
    - The regression model explains only 2.4% of the variance in step count, indicating a very weak predictive power. The negative coefficient for minutes played (-0.0819) aligns with the Pearson correlation, confirming a weak negative relationship between step count and listening minutes. The P-value of 0.003 for this coefficient suggests that the relationship is statistically significant, albeit weak.

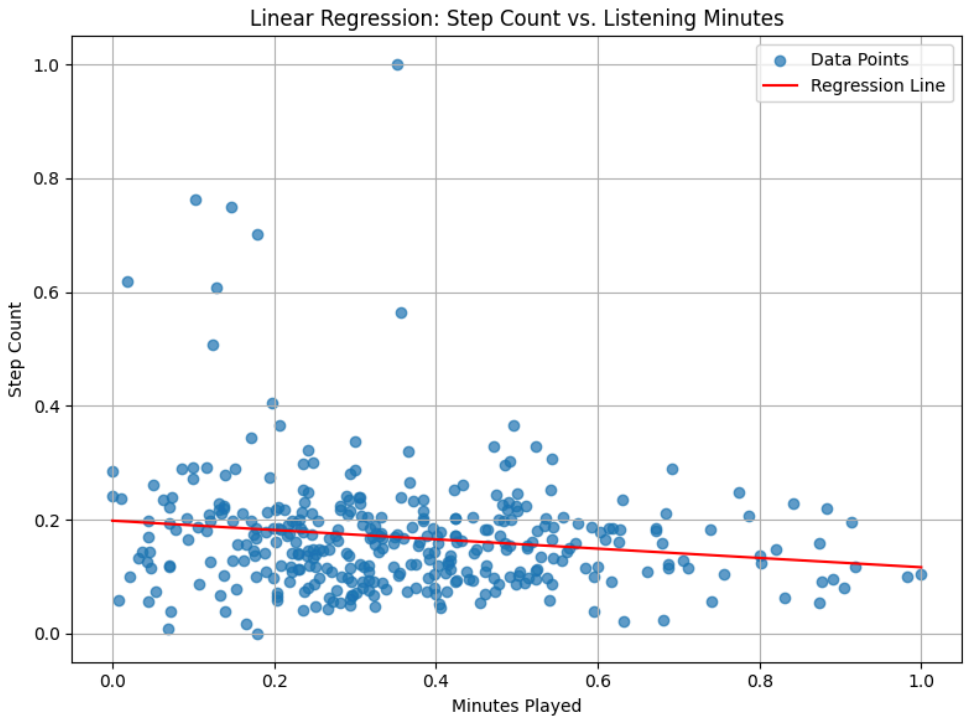
Random Forest Analysis

To complement the regression analysis, a Random Forest model was applied to predict step count based on listening minutes. The performance of the model was evaluated using the Mean Squared Error (MSE):

- **Mean Squared Error (MSE):** 0.1613

Interpretation

- The Random Forest model demonstrates limited predictive accuracy with an MSE of 0.1613. This result further supports the conclusion that listening minutes have minimal predictive power for step count, reinforcing the weak relationship observed in previous analyses.



- **Regression Line**
    - The scatter plot with the regression line visualizes the weak negative trend between step count and listening minutes. Most data points are scattered widely, reinforcing the weak correlation observed in the statistical tests.

Conclusion

The analyses reveal the following:

- A weak but statistically significant negative correlation between step count and listening minutes ((r = -0.1536), (p = 0.0034)).
- No significant differences in listening minutes between high and low step count groups based on the t-test ((p = 0.1012)).
- No significant association between categorized step count and listening minutes from the chi-square test ((p = 0.4641)).
- The OLS regression confirms the weak negative correlation but suggests minimal predictive power ((R^2 = 0.024), (p = 0.003)).
- The Random Forest model yields a relatively high Mean Squared Error (0.1613), further demonstrating the weak predictive capability of listening minutes for step count.

Overall, while a statistically significant relationship exists, the weak correlation, low predictive power, and high MSE indicate that step count and listening minutes are only minimally related. This suggests that other factors may have a more substantial influence on these variables.

Conclusion of the Project

As an Industrial Engineering student undertaking the Introduction to Data Science course, this project provided a unique opportunity to apply data science methodologies to a real-world problem. The primary aim was to explore the potential relationship between physical activity, represented by step counts, and music listening habits, represented by Spotify listening minutes, using personalized datasets collected over a year.

This project aimed to bridge theoretical knowledge with practical applications, utilizing tools such as data visualization, hypothesis testing, correlation analysis, and machine learning models. By doing so, the project demonstrated the power of data-driven decision-making in understanding complex behavioral patterns.

Beyond just analyzing data, the project emphasized key skills such as data preprocessing, merging datasets, and applying statistical tests to draw meaningful conclusions. The integration of various methods, including Pearson correlation, t-tests, chi-square tests, linear regression, and Random Forest modeling, provided a comprehensive approach to understanding the data.

The findings revealed a weak but statistically significant negative correlation between step counts and listening minutes, showing that while the relationship exists, it is minimal. This outcome reflects the complexity of human behavior and highlights the need for a multidisciplinary approach to understanding such interactions. As an Industrial Engineering student, this project aligns well with the principles of systems thinking and optimization by exploring how various factors interconnect to influence personal habits.

In conclusion, this project served as a valuable exercise in applying data science techniques to a personal and relatable dataset. It emphasized the importance of critical thinking, statistical analysis, and the practical implementation of data-driven approaches in uncovering insights. Moving forward, this experience not only enhances my understanding of data science principles but also equips me with tools to analyze and optimize systems, a fundamental aspect of Industrial Engineering.