



DSA210



# MUSIC ACTIVITY AND CALORIE CORRELATION

SIMAY REEL  
31190

Start



# Introduction

This project explores the correlation between physical activity (daily steps and calories burned) and music listening habits (Spotify listening time). Using data analysis and machine learning techniques, the project examines trends and relationships across datasets collected over a specified time frame.

The project incorporates exploratory data analysis, data visualization, and statistical modeling to investigate the hypothesis that increased music listening time correlates positively with higher physical activity levels. It concludes with a discussion of the results and their implications.





# MY HYPOTHESIS

HO: The more music I listen to in a day,  
the more I walk and burn calories.

HA: The amount of music listened to in  
a day does not have a significant  
positive relationship with the number  
of steps taken and calories burned.





# Data sources and preparation

## 1. **Spotify Listening Data:**

- Extracted from JSON files containing detailed listening history, including timestamps, track names, and listening durations.
- Preprocessed into a daily aggregated format to calculate total listening time in minutes.

## 2. **Health Data:**

- Extracted from XML files, including metrics for daily steps and calories burned.
- Converted to CSV format and preprocessed to ensure consistency in date formatting and values.

## 3. **Merged Dataset:**

- Combined Spotify and health datasets using the date as a common key.
- Ensured data alignment and consistency across all features.



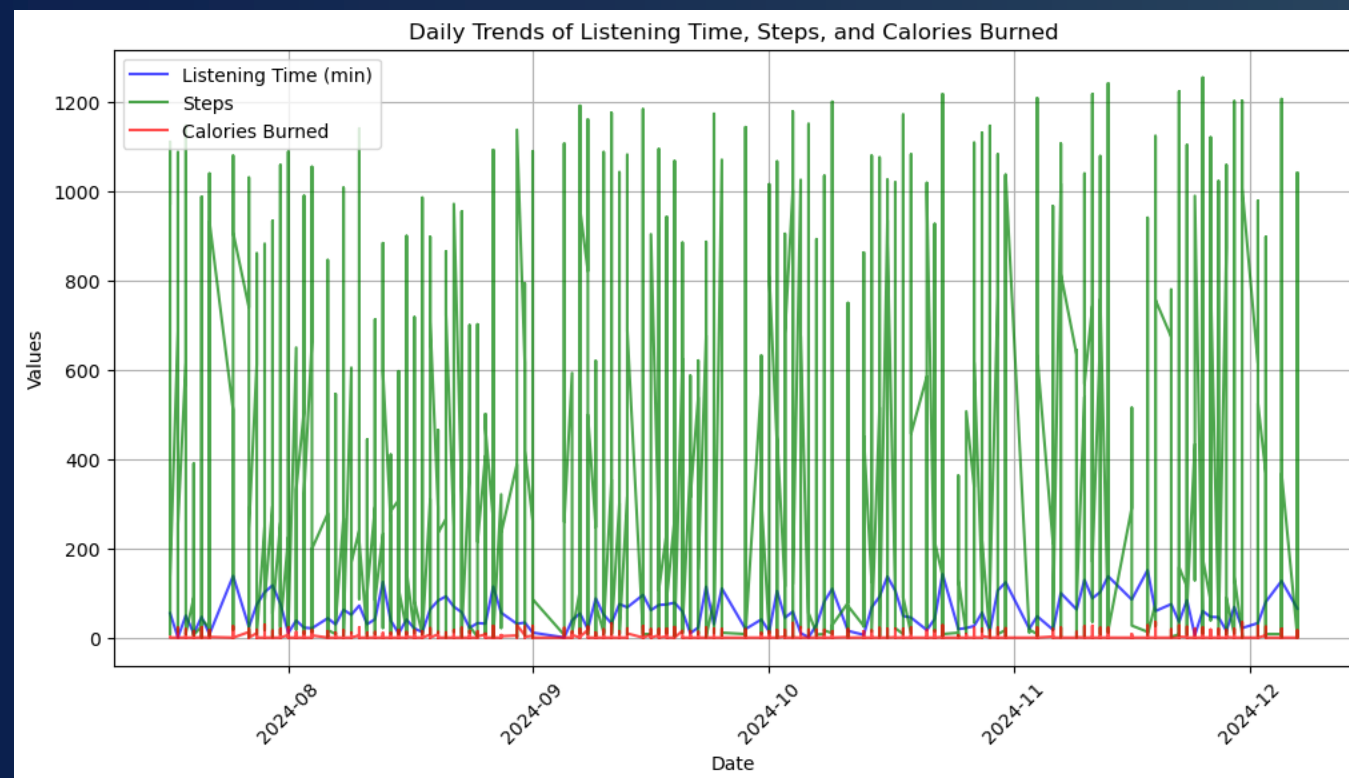



# Exploratory Data Analysis (EDA)



## 1. Daily Trends:

- Visualized daily listening time, steps, and calories burned.
- Observed patterns such as peaks in music listening during evenings and fluctuations in physical activity on weekends.



- 
- There doesn't seem to be a clear pattern linking spikes in music listening time to steps or calories burned on a daily level.
  - Steps show significant variability, while music listening and calories burned are relatively stable.



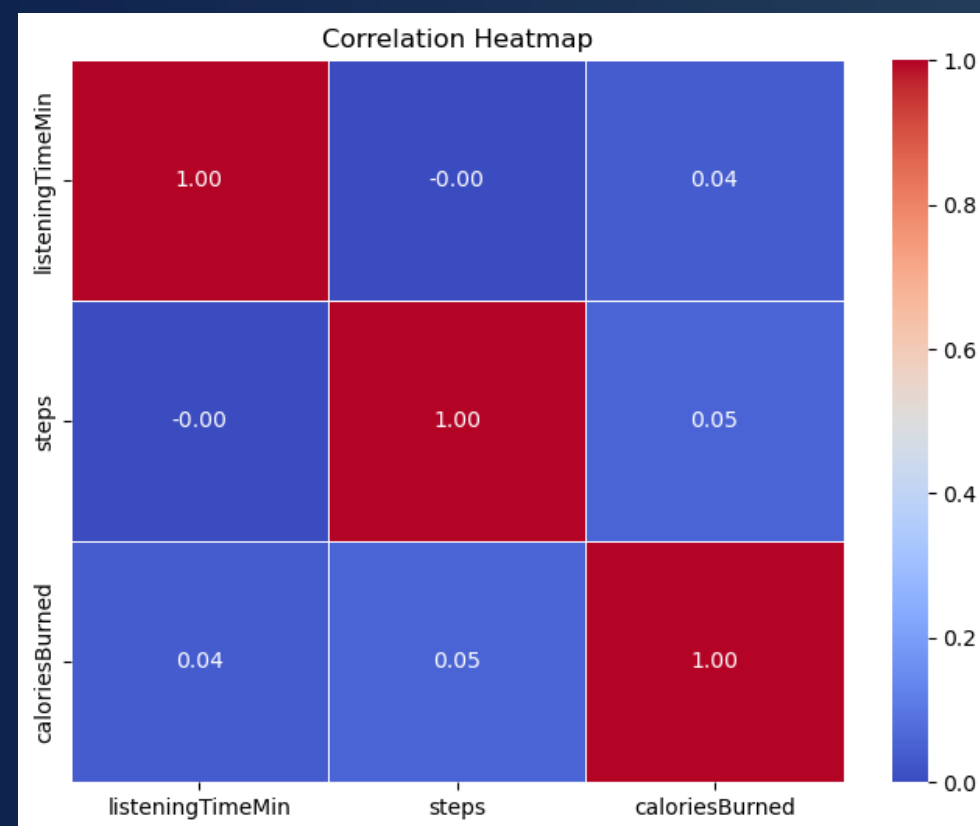


# Exploratory Data Analysis (EDA)



## 2. Correlation Analysis:

- Computed correlation coefficients between listening time, steps, and calories burned.
- Generated heatmaps and scatter plots to illustrate relationships.



- ❑ **Listening Time vs. Steps:**  $\sim 0.00$ , indicating no linear correlation.
- ❑ **Listening Time vs. Calories Burned:**  $\sim 0.04$ , showing a very weak positive correlation.
- ❑ **Steps vs. Calories Burned:**  $\sim 0.05$ , indicating weak correlation.
- ❑ The hypothesis that music listening positively correlates with physical activity levels (steps or calories burned) isn't strongly supported.



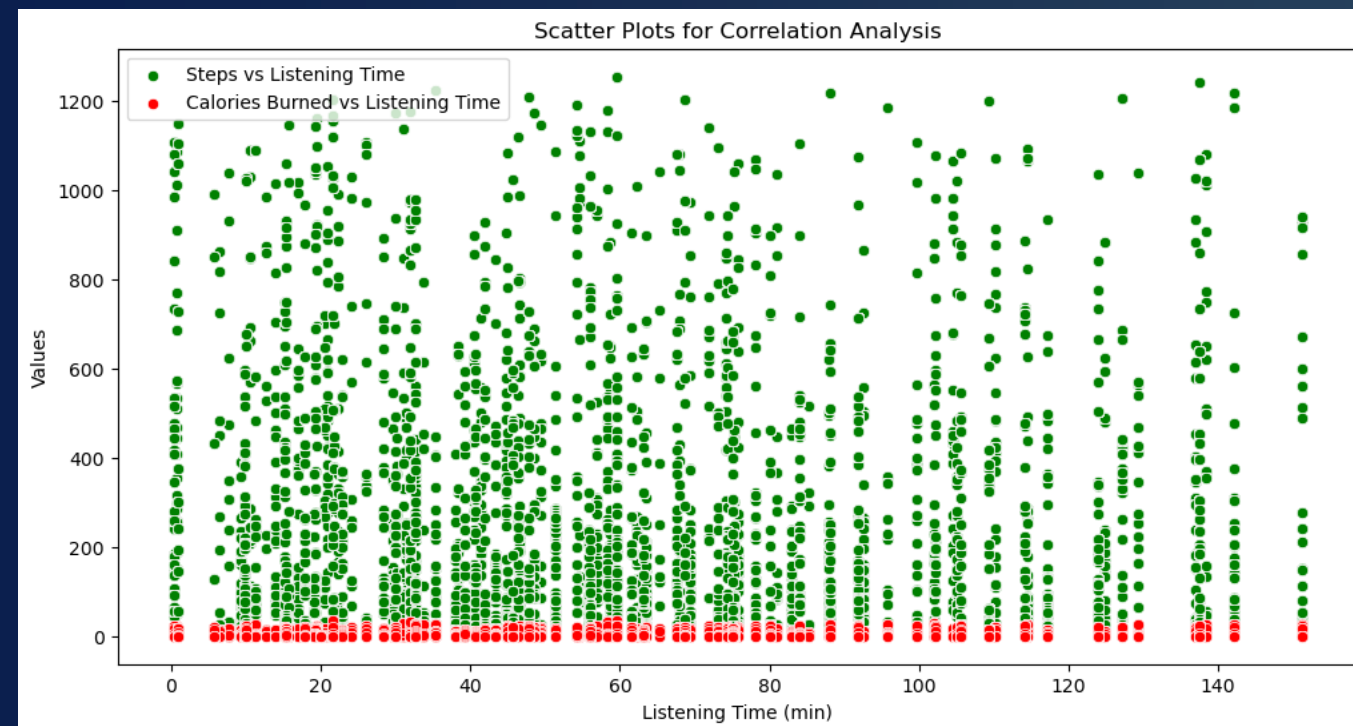


# Exploratory Data Analysis (EDA)



## 2. Correlation Analysis:

- Computed correlation coefficients between listening time, steps, and calories burned.
- Generated heatmaps and scatter plots to illustrate relationships.



- ❑ **Steps vs. Listening Time:** Points are widely scattered with no apparent trend, confirming the lack of a strong correlation.
- ❑ **Calories Burned vs. Listening Time:** Most data points are clustered near the origin, indicating minimal variability in calories burned relative to listening time.
- ❑ Scatter plots reinforce the heatmap findings—there's no evident relationship between music listening time and physical activity metrics.





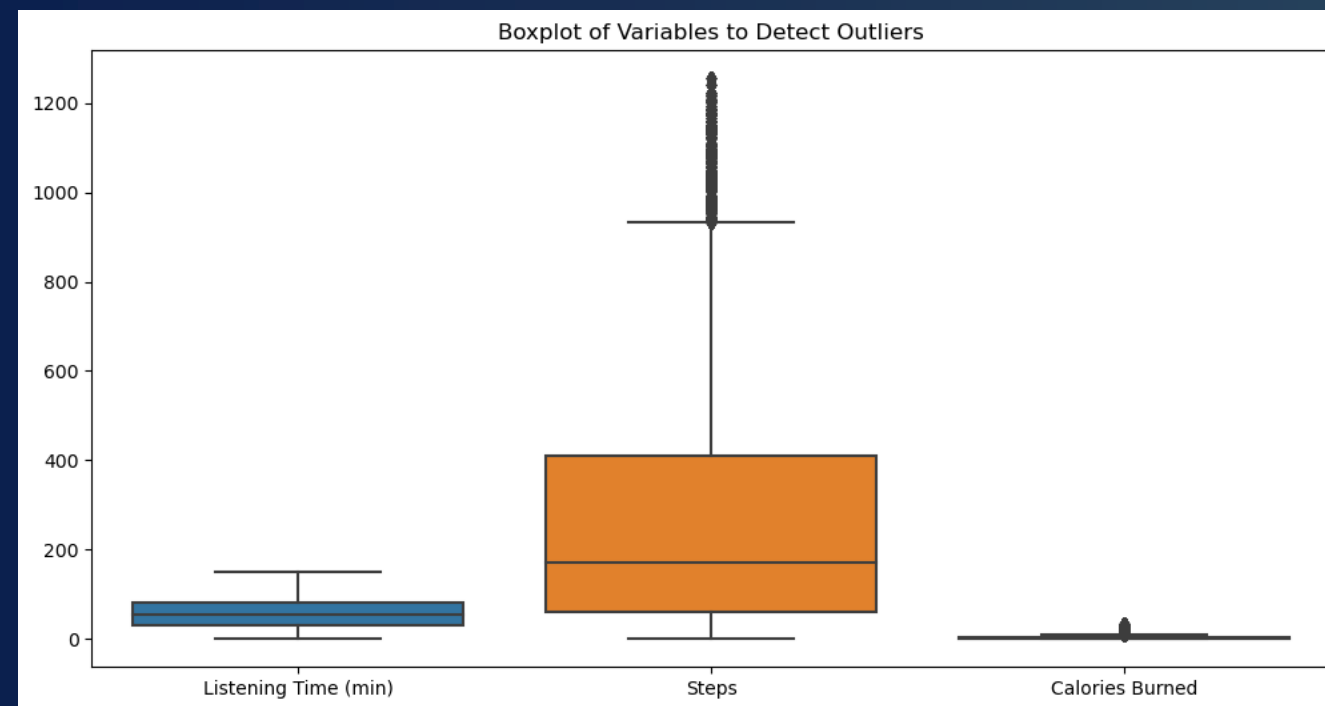


# Exploratory Data Analysis (EDA)



## 3. Outlier Detection:

- Identified and removed outliers using boxplots and percentile thresholds to improve data quality.



- ☐ **Listening Time (min):** No visible outliers. It appears that listening time is relatively consistent day-to-day.
- ☐ **Steps:** A large variability is evident, with many outliers above 800 steps. This indicates that on some days, the steps are significantly higher than average.
- ☐ **Calories Burned:** There are minimal outliers. Indicates that calorie burn due to walking is relatively stable.





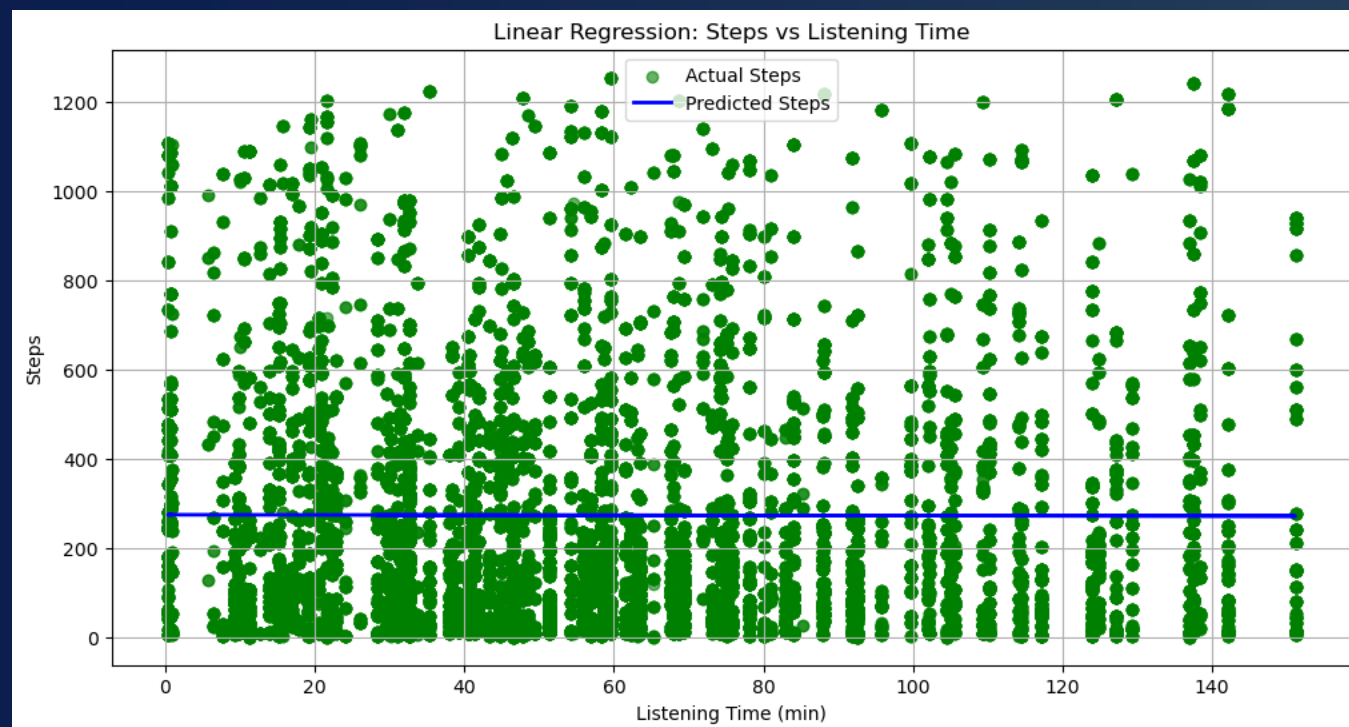


# Machine learning analysis

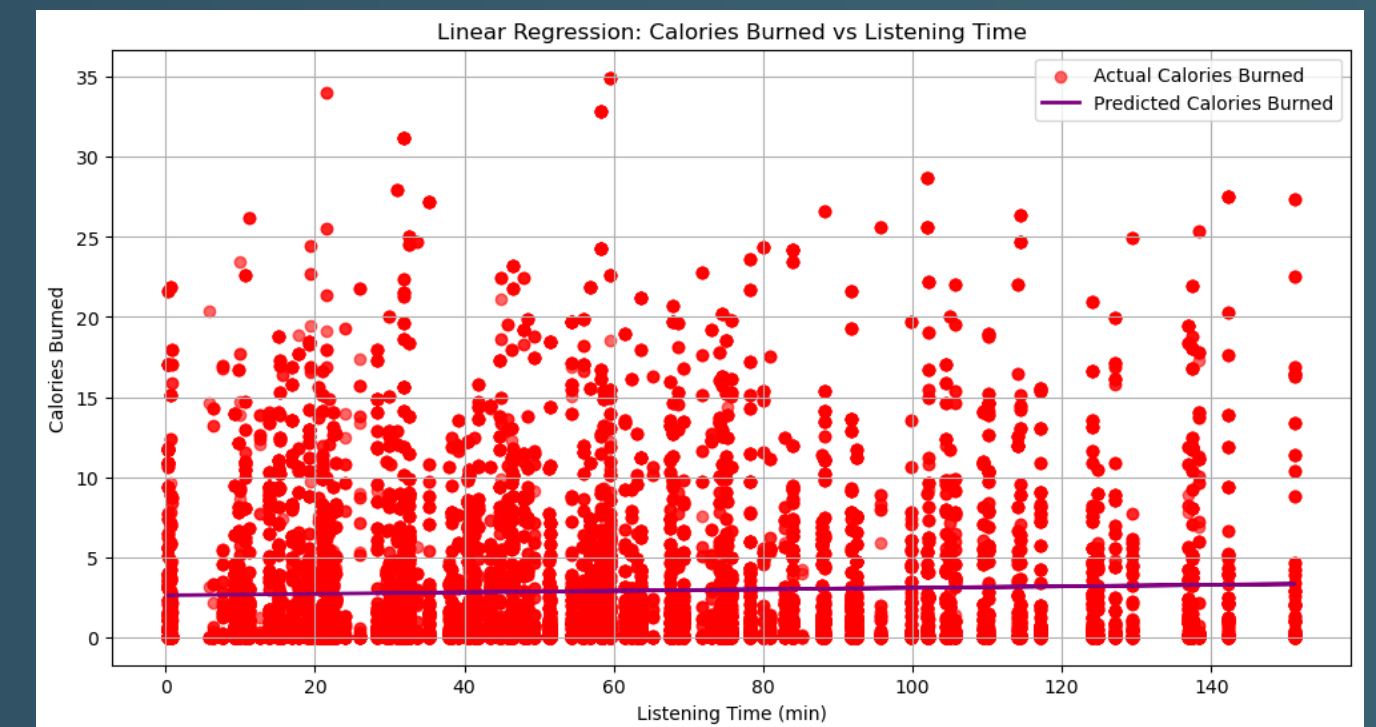


## 1. Linear Regression Models:

- Trained separate models to predict steps and calories burned based on listening time.
- Evaluated models using Mean Squared Error (MSE) and  $R^2$  scores, which indicated weak predictive relationships.



The lack of upward trends in both scatter plots suggests that increased music listening time does not directly influence physical activity levels in my data.

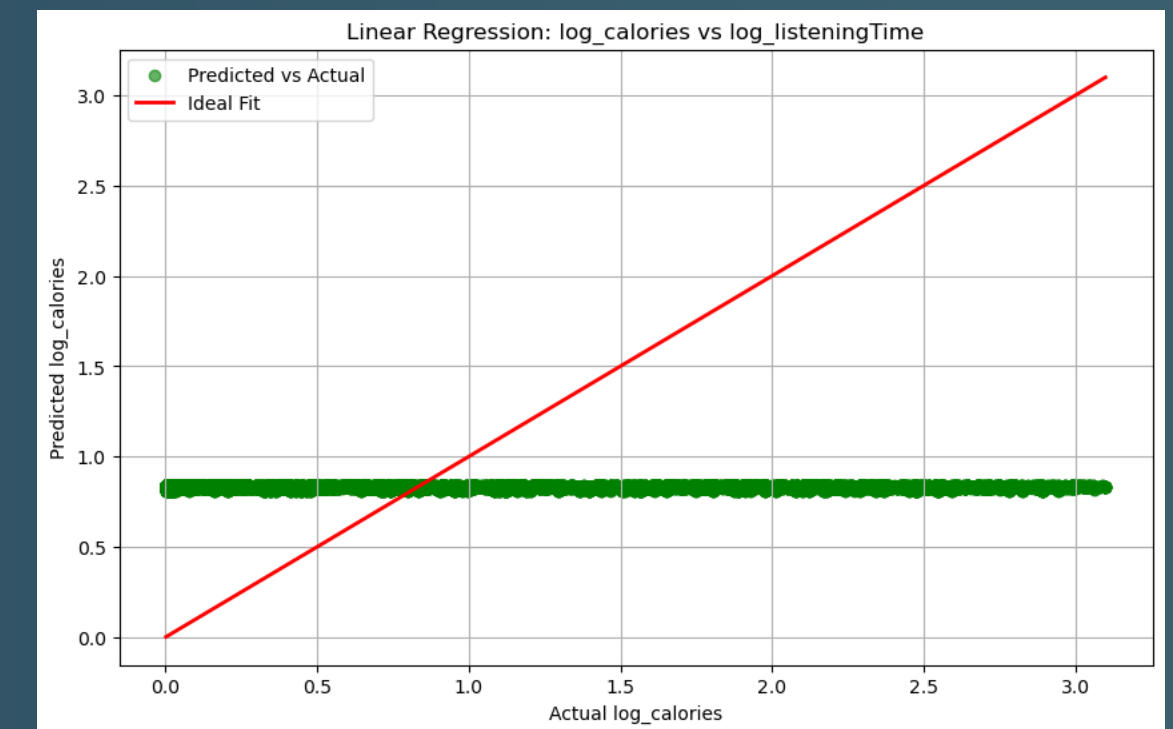
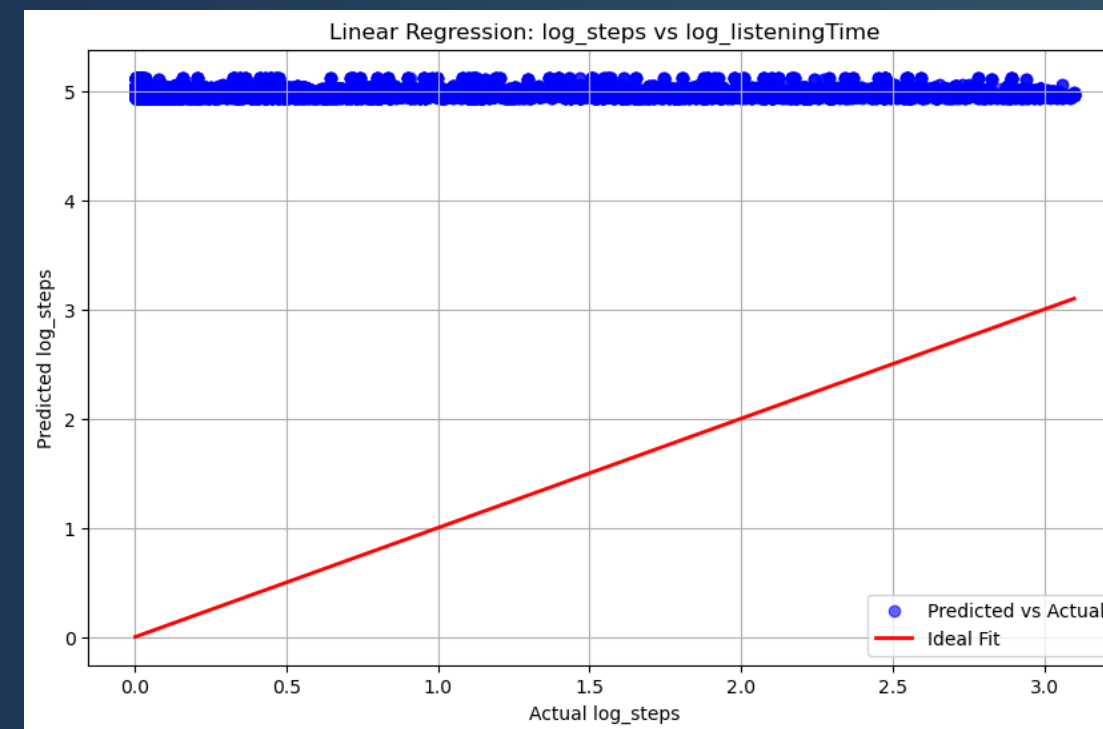
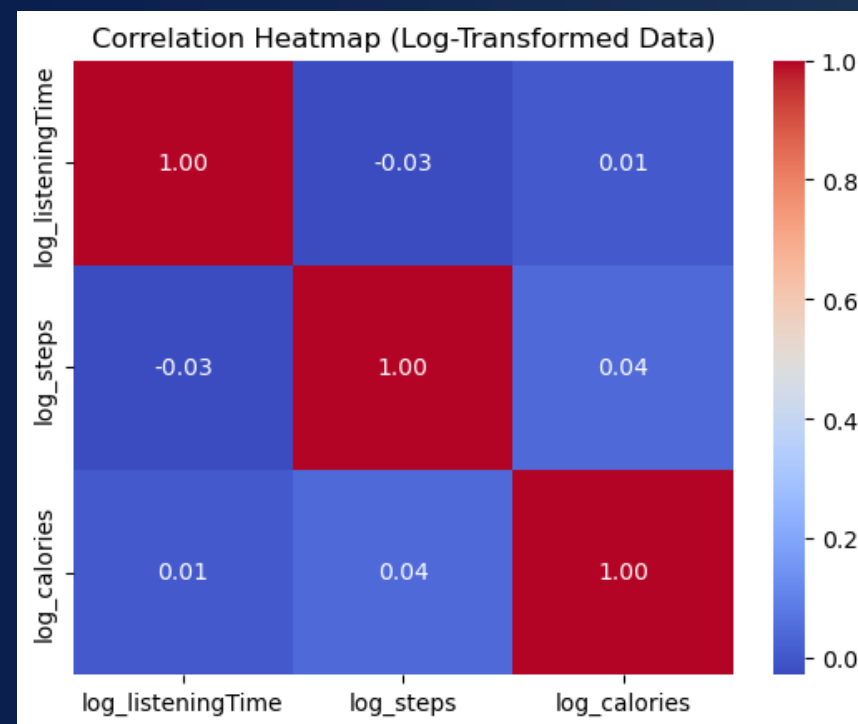




# Machine learning analysis

## 2. Log Transformation:

- Applied log transformation to normalize data distributions.
- Recomputed correlations and regression models, yielding similar results.





# Machine learning analysis



## 3. Decision Tree Regression:

- Implemented a decision tree model for steps prediction.
- Achieved slightly improved performance but still weak predictive power.

```
from sklearn.tree import DecisionTreeRegressor

# Decision tree for steps prediction
X = merged_df[['listeningTimeMin']]
Y_steps = merged_df['steps']

X_train, X_test, Y_train, Y_test = train_test_split(X, Y_steps, test_size=0.2, random_state=42)

tree_model = DecisionTreeRegressor(max_depth=5, random_state=42)
tree_model.fit(X_train, Y_train)
Y_pred = tree_model.predict(X_test)

mse = mean_squared_error(Y_test, Y_pred)
r2 = r2_score(Y_test, Y_pred)
print(f"Decision Tree Model - MSE: {mse:.2f}, R^2: {r2:.2f}")
```

Decision Tree Model - MSE: 66434.01, R^2: 0.05



$R^2$  measures how well the variance in the target variable is explained by the input features. A value of 0.05 means that only 5% of the variability in my target is explained by the listening time or other features. Hence, this is a very low value, indicating a weak relationship or poor predictive performance.





# Findings

**1.**

## **Weak Correlation:**

Correlation coefficients for listening time vs. steps and calories burned were close to zero, suggesting no strong relationship.

**2.**

## **Outliers and Variability:**

High variability in daily physical activity and music listening habits likely diluted any potential correlation.

**3.**

## **Subjective Factors:**

Factors such as mood, weather, and personal routines may influence both metrics independently, complicating the analysis.





# Visualizations Used



## 1. **Daily Trends Plot:**

- Overlaid line plots of daily listening time, steps, and calories burned.

## 2. **Heatmaps:**

- Correlation heatmap for raw and log-transformed data.

## 3. **Scatter Plots:**

- Visualized relationships between listening time and health metrics.

## 4. **Weekly Averages:**

- Aggregated data by week to identify broader trends.







# Limitations and future improvements



- Future research could incorporate additional variables (e.g., music genre, mood) and explore larger, more diverse datasets to better understand these relationships.
- Data might not represent all factors affecting physical activity.
- Linear regression assumes a specific type of relationship, which may not exist.
- Collect more data over a longer period.
- Investigate non-linear relationships or seasonal patterns.





# Conclusion

- The hypothesis that music listening time positively correlates with physical activity was not supported by the data. Hence, we reject the null hypothesis and fail to reject the alternative hypothesis.
- The analysis failed to establish a statistically significant correlation between Spotify listening time and physical activity levels. While interesting patterns were observed, such as increased music listening during specific periods, the overall relationships were weak.







## Hypothesis Testing Results

Steps  $R^2$ : 0.00  
Calories  $R^2$ : -0.00

Conclusion: Correlation Strength is Weak



## Final Conclusion

There is no significant correlation between music listening time and physical activity.

Hypothesis Rejected.

