Step Count and Music Listening Behavior Analysis

---

1. Dataset Description

1.1. Step Count Data

- **Source**: Apple Health step count records.
- **Scope**: This dataset includes daily step counts and associated timestamps for a specific time period.
- **Details**:
  - The total number of steps taken each day serves as a proxy measure for physical activity and outdoor movement.
  - This data provides a quantitative understanding of user mobility patterns over time.

1.2. Spotify Listening Data

- **Source**: Spotify extended streaming history.
- **Scope**: The dataset includes detailed Spotify listening records.
- **Details**:
  - Information includes the total number of songs listened to and the cumulative duration (in minutes) of music consumption on a daily basis.
  - The data allows for behavioral analysis related to music consumption habits.

---

2. Project Objective

Research Objective

This study aims to investigate the relationship between physical activity (measured via daily step counts) and music listening behavior (measured via Spotify data). By analyzing the interaction between these two aspects, the project seeks to determine if music consumption patterns can reflect or influence physical activity levels.

Hypothesis

- **H0 (Null Hypothesis)**: There is no significant relationship between daily physical activity (step counts) and music listening behavior (Spotify listening time).
- **H1 (Alternative Hypothesis)**: There exists a significant correlation between daily physical activity and music listening behavior, such that higher physical activity levels correspond to increased music listening.

Significance of Study

Understanding the relationship between step counts and music listening can:

- Provide insights into how music consumption patterns are integrated into daily routines.

- Highlight potential behavioral trends linking physical activity with psychological relaxation through music.

---

3. Project Plan with Implementation

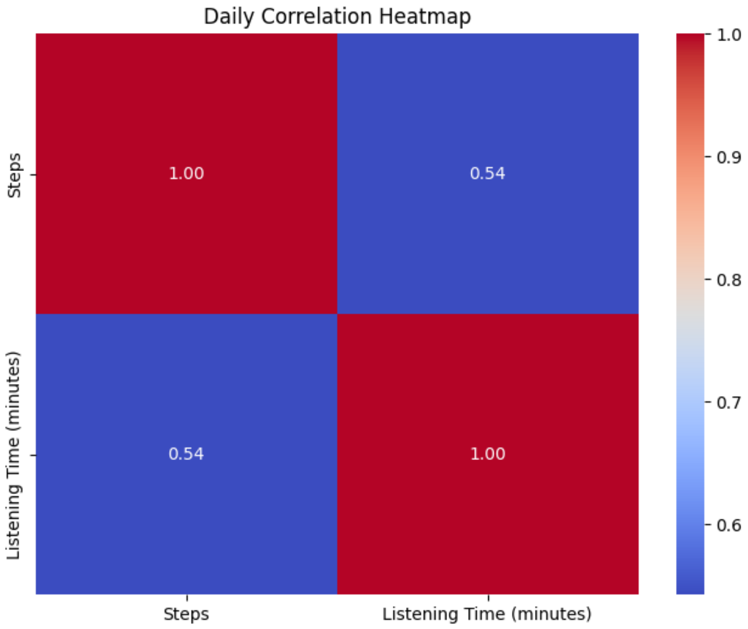
Step 1: Data Collection and Preprocessing

1. **Import Necessary Libraries**
   - Libraries such as pandas, matplotlib, seaborn, sklearn, and lxml were used for data processing, visualization, and modeling.
2. **Load and Process Step Count Data**
   - Step count data from Apple Health XML files was extracted and aggregated by date.
3. **Load and Process Spotify Listening Data**
   - Daily listening time was calculated from Spotify's JSON files, filtering out days with less than one minute of listening.
4. **Combine the Data**
   - Data was aligned by date, normalized, and aggregated into daily and weekly levels for analysis.

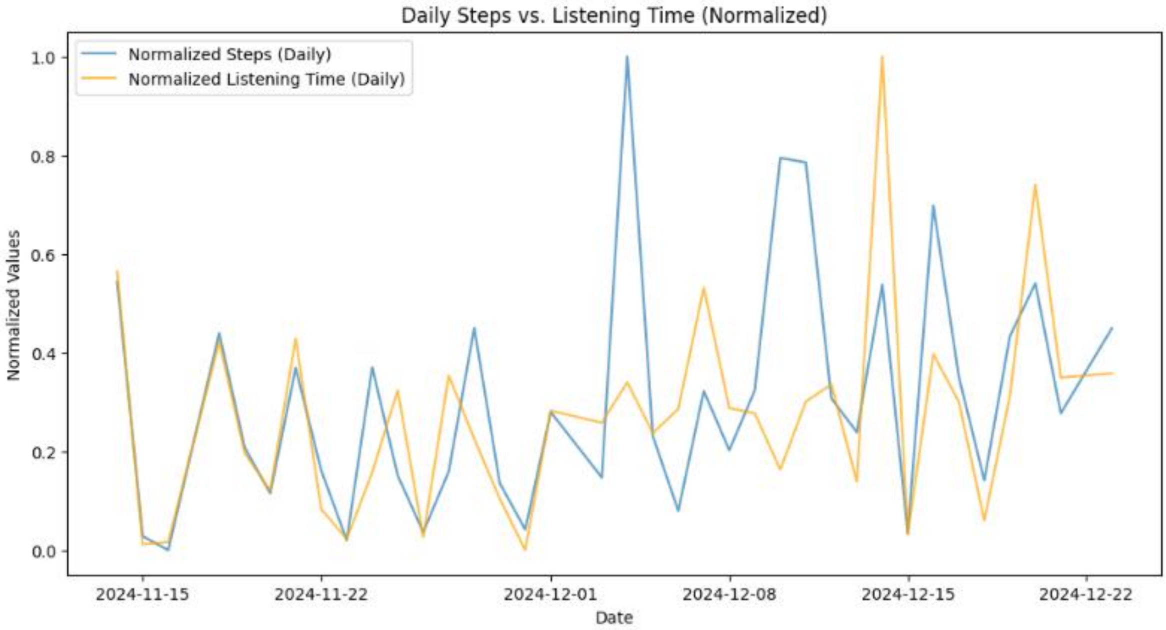
Step 2: Exploratory Data Analysis (EDA)

Daily Analysis:

- The correlation heatmap revealed a moderate positive correlation (0.54) between daily step counts and listening time.

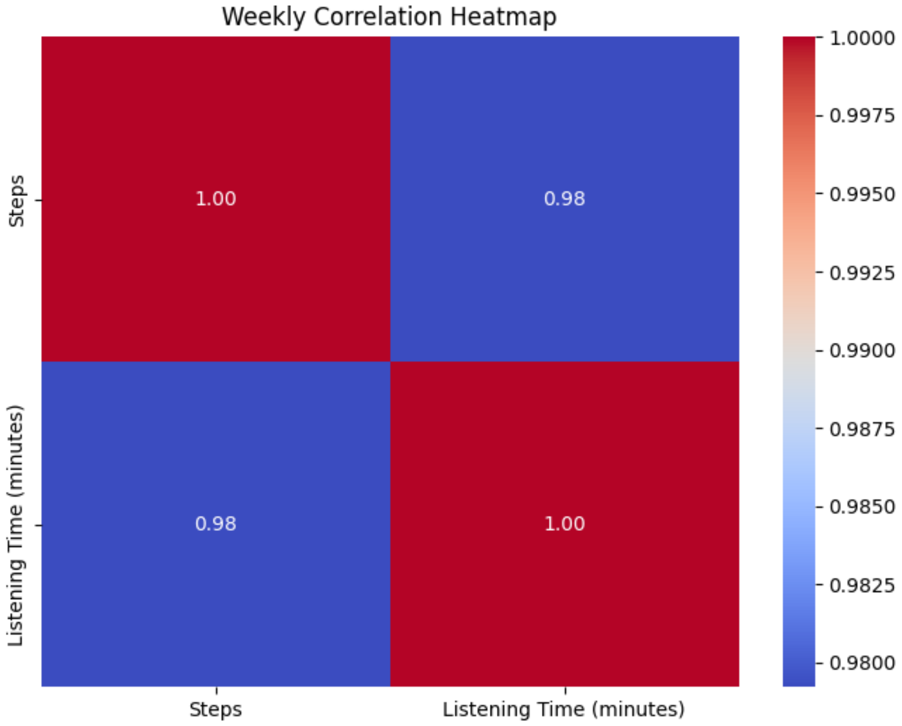


- Normalized trends for daily steps and listening times were plotted to observe temporal patterns.

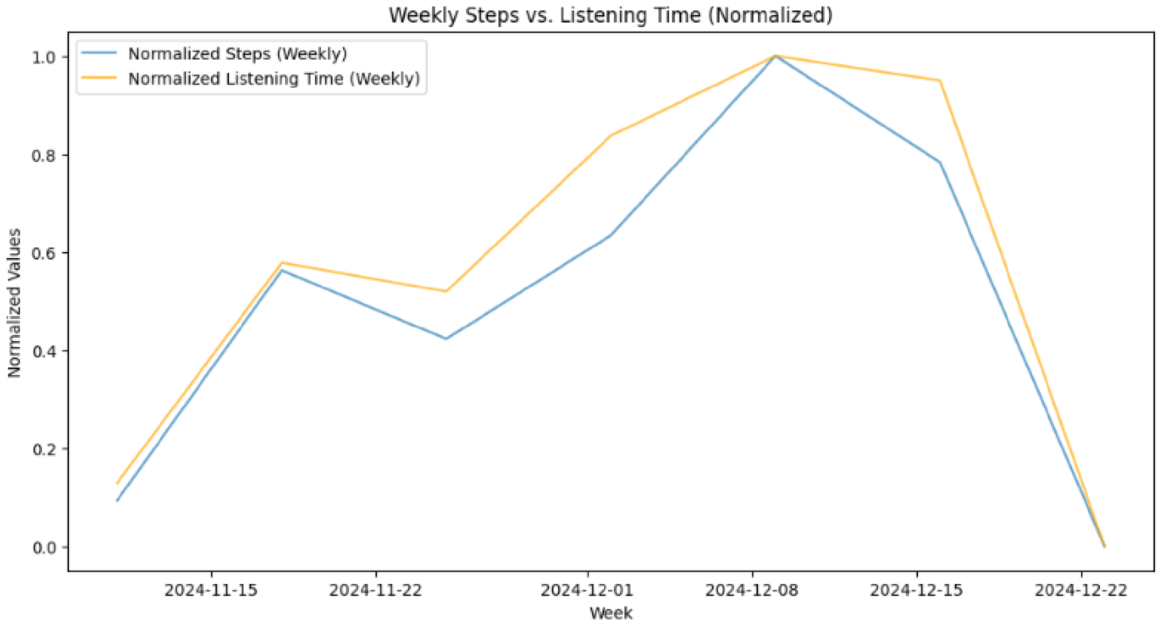


Weekly Analysis:

- The weekly correlation heatmap displayed a stronger positive correlation (0.98) between weekly aggregated step counts and listening time.



- Weekly normalized trends demonstrated clearer relationships compared to daily data.



4. Modeling

1. Linear Regression (Daily)

- Identifies and quantifies the strength of linear relationships between daily step counts and listening time.
- **Results**:
  - Coefficient: 0.0101
  - Intercept: 17.65

2. Linear Regression (Weekly)

- Aggregates weekly data to smooth out daily noise and reveal longer-term trends.
- **Results**:
  - Coefficient: 0.0131
  - Intercept: 26.84

3. Random Forest Regression (Daily and Weekly)

- Models nonlinear patterns and assesses the impact of step counts on listening time.
- **Results**:
  - Daily MAE: 18.41 minutes

    - Weekly MAE: 53.64 minutes

4. Random Forest Classification (Daily and Weekly)

- Classifies activity levels ("High" or "Low") based on step counts.
- **Results**:
    - Daily Accuracy: 87.5%
    - Weekly Accuracy: 100%

---

5. Insights and Deliverables

Findings

1. **Daily Analysis:**
    - Weak correlation between daily steps and listening times.
    - Random Forest Classifier demonstrated strong classification performance.
2. **Weekly Analysis:**
    - Stronger weekly correlation suggests that aggregated data better captures trends.
    - Classification models achieved perfect accuracy at the weekly level.

Hypothesis Evaluation

- The weak daily correlation partially supports the null hypothesis (H0).
- The stronger weekly correlation and classification results support the alternative hypothesis (H1), indicating a significant relationship between activity and listening over longer timeframes.

Recommendations

1. **Dataset Expansion:**
    - Extend the dataset duration to capture seasonal and monthly trends.
    - Collect data points across diverse demographics to enhance generalizability.
2. **Behavioral Insights:**
    - Investigate the influence of music genres on physical activity.
    - Explore temporal listening patterns, such as morning versus evening behavior.
3. **Hybrid Models:**
    - Employ a combination of linear regression for trend analysis and Random Forest models for nuanced predictions and classifications.