**Analyzing the Relationship Between Audio Exposure and Physical Activity Dynamics**

*Ahmet Alperen Öznam 28928*

**Introduction**

In recent years, personal audio devices have become ubiquitous, seamlessly blending entertainment and productivity into daily routines. Whether it is listening to music while studying or walking, the prevalence of headphone usage has increased significantly. However, prolonged exposure to headphone audio, particularly at higher volumes, raises questions about its potential influence on physical activity and walking dynamics. This study investigates the relationship between headphone audio exposure and physical activity metrics, specifically walking performance, energy expenditure, and step counts, using data derived from Apple Health.

The primary objective of this exploration is to determine whether headphone audio exposure correlates with changes in physical activity patterns. By integrating audio exposure data with walking and energy-related metrics and aligning timestamps, this project seeks to analyze trends and assess any potential impacts. Furthermore, it explores whether certain audio exposure patterns could optimize or maintain walking performance.

Key components of this analysis include exploratory trend visualization, correlation studies, and predictive modeling. An essential aspect is distinguishing between correlation and causation, ensuring a robust interpretation of results and findings.

Tools and technologies such as Python, Pandas, NumPy, and Scikit-learn were employed for data processing and modeling, while Matplotlib and Seaborn were used for data visualization. Challenges such as accurately interpreting audio exposure metrics and addressing the causation vs. correlation dilemma were carefully considered to provide actionable and meaningful insights.

**Hypothesis Testing**

To examine the relationship between headphone audio exposure and physical activity, this study formulates the following null and alternative hypotheses:

* **Null Hypothesis (H0)**: There is no relationship between headphone audio exposure and physical activity metrics such as walking performance, energy expenditure, and step counts.
* **Alternative Hypothesis (H1​)**: There is a relationship between headphone audio exposure and physical activity metrics.

**Methodology for Hypothesis Testing**

The statistical significance of correlations is assessed using p-values derived from a permutation test. This approach involves:

1. **Distance Correlation Calculation**: Compute the observed correlation (Robs) between headphone audio exposure and each physical activity metric. Why distance correlation is used for corralation matrix calculation will be discussed in further sections.
2. **Permutation Test**:
   * Generate a distribution of correlation values by randomly permuting one of the variables (e.g., step count) while keeping the other fixed (e.g., audio exposure).
   * Repeat this process 1,000 times to build a distribution under the null hypothesis.
3. **P-Value Determination**: Calculate the p-value as the proportion of permuted correlations with an absolute value equal to or greater than ∣Robs∣∣Robs​∣.

**Decision Rule**

* If p ≤ 0.05, reject the null hypothesis, indicating a statistically significant relationship.
* If p > 0.05, fail to reject the null hypothesis, suggesting insufficient evidence to claim a significant relationship.

This approach ensures robust and reliable conclusions about the potential influence of headphone audio exposure on physical activity metrics.

**Code Implementation**

**1. Data Export and Organization**

The data-exporting.ipynb notebook is dedicated to processing and organizing raw data exported from Apple Health. The primary goal was to handle the large dataset effectively by splitting it into smaller, more manageable chunks based on the type of health data. This organization ensures that specific metrics can be analyzed more easily in subsequent steps.

The process begins with parsing the raw XML file containing all health records. Unique record types are identified, and the data is categorized into meaningful groups such as "Headphone Audio," "Walking Metrics," "Physical Activity," "Sleep Metrics," and "Body Stats." This categorization reflects the key areas of interest for the analysis.

To clean the data, unnecessary attributes like source details and device information were removed, keeping only the essential information required for analysis. Finally, the cleaned records were saved as separate XML files within a structured directory, grouped by category. This approach streamlines access to specific datasets and ensures that each category can be analyzed independently, reducing complexity.

The organized data provides a clear foundation for exploratory and hypothesis-driven analysis in the following stages.

**2. Data Preparation**

The data-prep.ipynb notebook focuses on transforming the cleaned data into a format suitable for analysis. This includes calculating daily values for each dataset and addressing missing data using imputation techniques.

* **Daily Value Calculation**:
  + For metrics like step count, active energy burned, and basal energy burned, daily totals were calculated by summing the values logged throughout the day.
  + For metrics like walking speed and headphone exposure, daily averages were calculated to represent the overall trend for each day.
* **Data Imputation**:
  + Missing data for metrics such as step count or walking speed was identified and addressed using mean imputation. For each metric, the mean value was calculated from the available data and used to fill missing entries.
  + Imputed data points were flagged to distinguish them from original data, ensuring transparency in the dataset.
* **Data Merging**:
  + Data from different metrics (e.g., exposure, step count, energy burned) was aligned by date and combined into a unified dataset. This allowed for seamless analysis of relationships between variables.

The resulting dataset contained daily aggregated values for each metric, with imputation flags clearly marking filled data points. This structured and enriched dataset is essential for reliable trend analysis and hypothesis testing.

**3. Data Interpretation**

The data-interpretation.ipynb notebook is dedicated to analyzing the relationship between headphone audio exposure and physical activity metrics. The primary tool for this analysis is distance correlation, chosen for its versatility and ability to handle non-linear relationships.

* **Initial Observations**:
  + Data visualization revealed that the relationships between variables such as exposure, walking speed, and energy burned were neither linear, monotonically increasing/decreasing, nor ordinal. This made traditional correlation metrics like Pearson, Spearman, or Kendall unsuitable for capturing the nuances of these relationships.
* **Why Distance Correlation?**:
  + Distance correlation measures both linear and non-linear associations between variables, making it ideal for this dataset.
  + Unlike traditional correlation metrics, it does not require assumptions about the nature of the relationship between variables, providing a more comprehensive analysis.
* **Implementation**:
  + Distance correlation matrices were calculated to quantify the strength of associations between headphone audio exposure and physical activity metrics.
  + To ensure the statistical significance of these correlations, p-values were derived using a permutation test. This involved comparing the observed correlation against a distribution of correlations generated under the null hypothesis.
* **Visualization**:
  + Heatmaps were used to display both the distance correlation values and their corresponding p-values, making it easy to identify significant relationships and patterns.

Distance correlation provided a robust and nuanced understanding of the relationships between headphone audio exposure and physical activity metrics. This analysis highlighted potential associations that might not have been captured by traditional correlation methods, offering valuable insights for hypothesis testing.

**4. High-Low Classifier**

The high-low-classifier.ipynb notebook focuses on building a predictive model to classify audio exposure levels (high or low) using physical activity metrics as input features. The objective is to explore whether exposure levels can be effectively predicted based on walking speed, energy burned, and step count.

* **Data Preparation**:
  + Data from the merged dataset was used, with "low exposure" labeled as 0 and "high exposure" labeled as 1.
  + Non-numeric columns and imputation flags were excluded from the analysis to retain only meaningful features.
* **Feature Selection**:
  + The following features were selected for classification:
    - Step count
    - Active energy burned
    - Basal energy burned
    - Walking speed
* **Model Training**:
  + A Random Forest classifier was chosen for its ability to handle non-linear relationships and feature importance analysis.
  + The dataset was split into training and testing sets (80% training, 20% testing) to evaluate model performance.
* **Evaluation**:
  + The classifier achieved high accuracy when predicting exposure levels. However, this high accuracy may indicate potential overfitting due to the dataset's simplicity or imbalanced classes.
  + A classification report was generated to assess precision, recall, and F1 scores for both exposure levels.

The model demonstrated the feasibility of using physical activity metrics to predict exposure levels, though the results may need further validation. In the report's "Shortcomings and Improvements" section, the limitations of this approach and recommendations for more robust modeling techniques are addressed.

**User-Defined Parameters**

The analysis includes two adjustable parameters that users can tune to modify the dataset and its compatibility:

1. **Threshold**:
   * The threshold parameter defines the decision boundary between high and low exposure levels. By adjusting this value, users can redefine what qualifies as high or low exposure. This allows for flexibility in classifying exposure data based on different criteria or contexts.
2. **last\_x\_days**:
   * The last\_x\_days parameter controls the number of recent data points included in the final dataset. Specifically, it extracts the last x elements of the exposure data and dumps them into separate JSON files for high and low exposure. This ensures that the dataset contains an equal number of samples for each class, maintaining compatibility and balance for subsequent analysis or modeling.

These parameters allow for a tailored approach to dataset preparation, enabling users to adapt the analysis to their specific needs while ensuring the data remains well-structured and balanced.

**Results**

After preparing the datasets, the first step was to explore the trends of the data over time. By plotting various metrics against their respective dates, it became evident how imputation affected the dataset. These visualizations also highlighted the non-linear nature of the relationships between variables, underscoring the complexity of the collected data. Such insights were crucial in shaping the subsequent steps of the analysis. One can see the plot in the appendix section.

To ensure that these issues did not compromise the integrity of the correlation analysis, further research was conducted into suitable techniques. Traditional correlation methods, such as Pearson or Spearman, were deemed inadequate for this dataset due to their reliance on specific assumptions, such as linearity or monotonic relationships. Through this research, an article on **distance correlation** was identified, which offered a robust solution. Distance correlation is capable of measuring both linear and non-linear associations, making it well-suited for datasets with diverse and complex relationships [1].

Additionally, special consideration was given to imputed values in the dataset. For the calculation of correlation and p-values, imputed data points were excluded to avoid introducing bias or inaccuracies. The analysis proceeded exclusively with the original data to ensure the validity of the findings. This decision further reinforced the robustness of the results.

The following correlation heatmaps are from 3 datasets, high exposure, low exposure and combined. The distance correlation value is between [0,1] 0 meaning independent and 1 meaning dependent.

A screenshot of a graph

Description automatically generated

For the high exposure data, the first row of the heatmap reveals the effects of headphone audio exposure on other physical activity parameters. While the correlation values are not particularly high across the range, the data suggests that the most significant relationship is between exposure and walking speed, with a correlation value of 0.26. This indicates that higher audio exposure levels may have a moderate influence on walking speed, though the relationship is not strongly pronounced. Unrelated to the project topic but it is also possible to see a strong relationship between active energy burned and step count (0.71).

A screenshot of a graph

Description automatically generated

For the low exposure data, the first row of the heatmap illustrates the relationship between headphone audio exposure and other physical activity parameters. Similar to the high exposure scenario, the correlation values are relatively low. However, the most notable effect of exposure in this case is on basal energy burned, with a correlation value of 0.19, followed closely by walking speed at 0.18. These values suggest a weak association between lower audio exposure levels and these specific physical activity metrics. It is important to note that the relation between step count and active energy burned has increase from 0.71 to 0.89. This can be by pure chance but we will see in the p-value analysis.

A screenshot of a graph

Description automatically generated

Combined data does not provide any valueable insight on the effect of exposure since the vlues are all pretty low. From this analysis importance of class based analysis of the data can be seen from this corralation matrices analysis I have seen that I was more active physically while I was exposed to high volume but this is not enough to reject of fail to reject the null-hypothesis.

A screenshot of a graph

Description automatically generated

From the p-value matrix of high exposure dataset it is possible to se that for walking speed we are failing to reject H0 since it is below alpha value. For other parameters we cannot say so but for ease of simplicity if for one value the hypothesis fail to be rejected I will completely fail to reject the null hypothesis.

A screenshot of a graph

Description automatically generated

At low exposure dataset there seem not to be any relation with respect to p-value since for all parameters it is above alpha value so for low exposure dtaset we reject the null hypothesis.

A screenshot of a graph

Description automatically generated

Also when combined datset p-value matrix is evaluated it is possible to see that for walking speed we conclude similarly for high exposure. This means that we fail to reject the null-hypothesis.

Last step of the implementation of the project was to develop a basic random forest classifier based on physical activity data where the labels were high-low exposure. Here is the confusion matrix of the model.

A screenshot of a computer screen

Description automatically generated

It seems that it is able to predict with 100% accuracy.

Potential developments and comments on how reliable the data is provided in the next section.

**Short-Comings of the Dataset and Further Development**

This analysis faced several limitations that impacted the reliability and generalizability of the results:

1. **Dataset Sparsity and Limited Overlaps**:
   * The dataset had relatively few samples for exposure data, with many non-overlapping dates between the various physical activity metrics. This sparsity limited the ability to draw definitive conclusions about the relationships between exposure and activity metrics. A larger, more complete dataset would likely yield clearer and more robust insights.
2. **Suspicious Energy Burn Data**:
   * Upon examining the active and basal energy burned values, they appeared inconsistent with expected averages. Comparing these values with reference data for average male calorie expenditure revealed significant discrepancies. This raises concerns about the accuracy of the energy data, likely stemming from limitations in the phone's ability to correctly capture and estimate calorie burn. Such inaccuracies may have introduced noise into the analysis, affecting the overall reliability of the results.
3. **Classification Model Constraints**:
   * The classification model was trained on a dataset of approximately 600 samples (80% making it 480 samples), which is relatively small for machine learning purposes. While the model achieved high accuracy, this could be due to overfitting or the dataset's simplicity. Increasing the dataset size would provide the model with more diverse examples, leading to more realistic and reliable predictions.

**Recommendations:**

* Collect a larger dataset with better temporal alignment between metrics to reduce sparsity and improve analysis accuracy.
* Investigate alternative methods or devices for capturing energy expenditure data to ensure reliable and accurate measurements.
* Re-train the classification model on a more extensive dataset to enhance its robustness and predictive power.

These limitations highlight areas for improvement and emphasize the importance of comprehensive, high-quality data in producing reliable and actionable findings.

**Conclusion**

This analysis aimed to explore the relationship between headphone audio exposure and physical activity metrics. While the trends in the data provided insights into potential relationships, limitations in dataset size, sparsity, and data quality significantly impacted the results.

Using distance correlation, the analysis revealed weak associations between exposure and physical activity metrics such as walking speed and basal energy burned. For high exposure, walking speed showed the strongest correlation (0.26), while for low exposure, basal energy burned (0.19) and walking speed (0.18) were the most affected metrics. However, combined data offered no additional value, emphasizing the importance of class-based analysis.

The p-value analysis concluded that:

* For high exposure, the null hypothesis (H0H) could not be rejected for walking speed, resulting in a failure to reject H0​ for the dataset as a whole.
* For low exposure, H0​ was not rejected across all parameters, reinforcing the lack of significant relationships under low exposure conditions.
* For combined data, the results aligned with those for high exposure, again failing to reject H0​.

A random forest classification model was developed, achieving 100% accuracy on the limited dataset. However, with only 600 samples, this result is likely influenced by overfitting and dataset simplicity.

In conclusion, while the null hypothesis could not be rejected for high and combined exposure datasets, the analysis suggests subtle trends that warrant further exploration. A larger, higher-quality dataset with better temporal alignment and improved energy data accuracy is recommended to draw more definitive conclusions.

**References**

1. D. Edelmann, T. F. Móri, and G. J. Székely, "On relationships between the Pearson and the distance correlation coefficients," Statistics and Probability Letters, vol. 169, p. 108960, 2021.

**Appendix**

Appendix 1


A graph with blue lines and dots

Description automatically generated A graph with blue lines and dots

Description automatically generatedA graph with blue lines and dots

Description automatically generated

A graph with blue lines and numbers

Description automatically generated A graph with blue lines and numbers

Description automatically generatedA graph with blue lines and numbers

Description automatically generatedA graph with blue lines and numbers

Description automatically generated