Project Report

Onur Orman

Introduction

The personal data might be considered one of the most valuable things for a person, particularly in terms of more sensitive data such as the personal information related to the official institutions, identity, financial status, etc. Besides these types of sensitive data, though it may not be considered as sensitive as the aforementioned types of data, there is another type of at least partially sensitive data: The health records, recorded by the devices that billions of people use every single day, in every single possible context. As much as this kind of data are sensitive, it is also important to properly analyze them in order to have a better comprehension into one's health and personal life, potentially for the benefit of the person whose records are analyzed. For that purpose, this project tries to understand the relationship between the physical activity of the author, a sophomore computer science and engineering student in Sabancı University, and his educational life, whose intensity soars increasingly over time.

The Data, the Source of the Data, and the Types of Data

The data contain the health records of the author, covering the last six years (precisely, from November 23, 2017 to November 10, 2023) of his life, which includes his education life in high school, the process of the university entrance exam, and the education life in Sabancı University. The data were collected from Apple Health on iPhone using the local exporting option, and were obtained as an XML file comprising the data of the following types: the step count, the distance of walking and running, the number of climbed flights, the length of a step, the speed of walking, the percentage of walking asymmetry, the percentage of walking double support, by which the amount of time when both feet are on the ground while walking is measured (*HKQuantityTypeIdentifierWalkingDoubleSupportPercentage*, n.d.), the percentage of walking steadiness, the amount of burned active energy, which is the type of energy consumed because of physical activity (*ActiveEnergyBurned*, n.d.), the amount of burned basal energy, which is the

energy used by the body for basic metabolic activities (*BasalEnergyBurned*, n.d.) and can be increased by increasing the physical activity and exercise (Marcus, 2013), and the level of audio from headphones.

The Hypothesis for the Data

The main idea for constructing the hypothesis was the claim that the author's physical activity increased as his education life intensified further. Therefore, the following pair of null hypothesis and alternative hypothesis was developed:

- Null hypothesis: The mean values of the data from the two periods that are subject to the test are equal; physical activity of the author in terms of the analyzed data does not increase on average as my education life intensifies.
- Alternative hypothesis: The mean value of the data from the first period to be tested is greater than the mean value of the data from the second period to be tested; physical activity of the author in terms of the analyzed data increases on average as my education life intensifies.

Analysis Methods for the Data

In order to analyze the data, as a first step, the XML file obtained from Apple Health was parsed using the ElementTree module in Python (*Xml.etree.ElementTree — the ElementTree XML API*, n.d.). After the initial parsing, the data with the identifiers that are not viable to analyze in this project, namely *HKQuantityTypeIdentifierBodyMass*, *HKQuantityTypeIdentifierHeight*, and *HKCategoryTypeIdentifierHeadphoneAudioExposureEvent* were removed from the obtained list of records, the source versions, which describe the iOS version installed on the device during the recording, were corrected to major versions and one-point minor versions, that is, for instance, if an example datum was collected in iOS 16.3.1, it was counted as collected in iOS 16.3, and the values whose unit was defined as percentage were changed with their multiplications by 100, because the values were recorded in the range between 0 and 1. Later, a dataframe using Pandas was created to store all of the records, where the creation times, starting times, and ending times of the records were converted into Datetime objects, and the column denoting source name was removed, since the name of the source was "Onur'un iPhone'u" in all the records, which was not a distinguishing feature of the records at all. Later, based on the identifiers indicating the data types, separate dataframes for each data type were created. The exploratory data analysis was

first conducted with the dataframes purely based on the distinct recordings, in which the shape, the object types, and the descriptive statistics of the dataframe were computed and box plots were prepared using Matplotlib and Seaborn, and the visualizations demonstrating the distributions of the individual data types, their changes over time, and the relations between those data types were created using the same packages, i.e., Matplotlib and Seaborn, in Python.

After the processing of the initial dataframes, the dataframes were reconsidered in the following manner: the records were summed based on the creation date for the data types of step count, distance of walking and running, number of climbed flights, the amount of burned active energy, and the amount of burned basal energy, while the daily averages based on the creation date were taken for the data belonging to the types of step length, speed of walking, percentage of walking asymmetry, percentage of walking double support, percentage of walking steadiness, and the level of audio from headphones. Then, exploratory data analysis was also conducted with these dataframes, including the printing of the shapes and the object types and the computations of descriptive statistics for each dataframe, and generation of box plots so as to visually describe the statistical properties of the dataframes, which can be found in appendices. Later, histograms were created in order to depict the distributions of the data in the dataframes, line charts were generated for showing the change in the values of different data types over time, and scatter plots were drawn on the purpose of describing the relations between data types. Having decided that the analysis of the data was going to become more convenient utilizing the dataframes created using the second approach, which is the approach where the data were grouped by the date of creation, correlations between the data types were calculated using .corr() method of Pandas, and correlation matrices were generated using Matplotlib and Seaborn.

For the stage of hypothesis testing, the data were separated into four main periods, nine subperiods for the fourth period, and eleven subperiods for the first period as follows:

- Period 1 (November 23, 2017 – March 10, 2020): The author's education in high school (pre-pandemic)
    - Period 1.1 (November 23, 2017 – January 18, 2018): The first term of the author in high school

    - Period 1.b1 (January 19, 2018 – February 4, 2018): The first semester break of the author in high school
    - Period 1.2 (February 5, 2018 – June 7, 2018): The second term of the author in high school
    - Period 1.b2 (June 8, 2018 – September 16, 2018): The first summer break of the author in high school
    - Period 1.3 (September 17, 2018 – January 17, 2019): The third term of the author in high school
    - Period 1.b3 (January 18, 2019 – February 3, 2019): The second semester break of the author in high school
    - Period 1.4 (February 4, 2019 – June 13, 2019): The fourth term of the author in high school
    - Period 1.b4 (June 14, 2019 – September 8, 2019): The second summer break of the author in high school
    - Period 1.5 (September 9, 2019 – January 16, 2020): The fifth term of the author in high school
    - Period 1.b5 (January 17, 2020 – February 2, 2020): The third semester break of the author in high school
    - Period 1.6 (February 3, 2020 – March 10, 2020): The sixth term of the author in high school, until the first case of COVID-19 in Turkey was officially announced.
- Period 2 (March 11, 2020 – June 27, 2021): The processes related to YKS, the centralized college admission exam in Turkey (COVID-19 pandemic continues in the world and in Turkey during this period); this period is considered as a period without any breaks in between.
- Period 3 (June 28, 2021 – September 26, 2021): The process of placement in a university after the college admission exam (COVID-19 pandemic continues with lower intensity in Turkey)
- Period 4 (September 27, 2021 – November 10, 2023): The author's education in Sabancı University

- Period 4.1 (September 27, 2021 – January 12, 2022): FDY (Foundations Development Year) as a preparatory course for improving skills in English; education was conducted as hybrid.
- Period 4.b1 (January 13, 2022 – February 27, 2022): The first semester break of the author in the university
- Period 4.2 (February 28, 2022 – June 10, 2022): First term of the freshman year of the author; education was conducted as hybrid; however, it was *de facto* online due to the low attendance in the physical lectures; therefore, it is considered as online.
- Period 4.b2 (June 24, 2022 – October 2, 2022): The first summer break of the author in the university
- Period 4.3 (October 3, 2022 – January 6, 2023): Second term of the freshman year of the author; the first term when the author experienced entirely face-to-face education in the university
- Period 4.b3 (January 21, 2023 – February 26, 2023): The second semester break of the author in the university
- Period 4.4 (February 27, 2023 – May 31, 2023): First term of the sophomore year of the author; education was conducted as online due to the constraints imposed because of the earthquakes in Turkey
- Period 4.b4 (June 12, 2023 – October 1, 2023): The second summer break of the author in the university
- Period 4.5 (October 2, 2023 – November 10, 2023): Second term of the sophomore year of the author; education was conducted as entirely face-to-face.

In terms of educational intensity, the periods can be ordered in a descending manner as follows:

- **Main periods:** Period 4, Period 2, Period 1, Period 3
- **Subperiods of Period 4:** Period 4.5, Period 4.4, Period 4.3, Period 4.2, Period 4.1, Period 4.b4, Period 4.b3, Period 4.b2, Period 4.b1

    - Period 4.b3, Period 4.b2, and Period 4.b1 can be considered as possessing an equal intensity in terms of education, which is a remarkably low intensity given that those periods were the breaks.
    - Period 4.b4 diverges from the other break subperiods since the author was taken part in a project (explained in the FAQ part of the GitHub repository); this subperiod is considered more intensive in terms of education than other break subperiods
- **Subperiods of Period 1:** Period 1.6, Period 1.5, Period 1.4, Period 1.3, Period 1.2, Period 1.1, Period 1.b5, Period 1.b4, Period 1.b3, Period 1.b2, Period 1.b1
    - Period 1.6 can be considered more intensive in terms of education than Period 1.5, and the two subperiods can be considered more intensive than Period 1.4, Period 1.3, Period 1.2, and Period 1.1, which have virtually the same educational intensity.
    - Period 1.b5, Period 1.b4, Period 1.b3, Period 1.b2, and Period 1.b1 are considered to have equal educational intensity.

After this separation was performed, the dataframes belonging to these periods and subperiods, when the data were available, were subjected to T-test with a level of significance equal to 0.05 in pairs. In this process, main periods, subperiods of Period 4, and subperiods of Period 1 were compared among themselves, i.e., a subperiod of Period 4 was not compared to a subperiod of Period 1 or a main period. The reasons why the T-test was utilized were that the test does not require the samples to possess a normal distribution provided that the size of the sample is sufficiently large (Tushev, 2021), and that it focuses on the difference between the means of the two samples, which was compliant with the main idea of the project. For this purpose, the data associated with step count, distance of walking and running, climbed flights, amount of burned active energy, and the amount of burned basal energy, which are directly related to the level of physical activity, were selected and subjected to the tests, while the other data types were used to visualize the relations, whose charts can be found in the appendices, and to calculate the correlation matrices.

Findings and Analysis

The findings in the project can be grouped into two categories: the correlations between the different data types,[1] and the results related to the hypothesis.

1) Correlations between Different Types of Data

The first correlation matrix, given in Figure 1, demonstrates the correlations between the data types which were recorded during the entire timeframe covered by the project: step count, distance of walking and running, and the number of climbed flights.

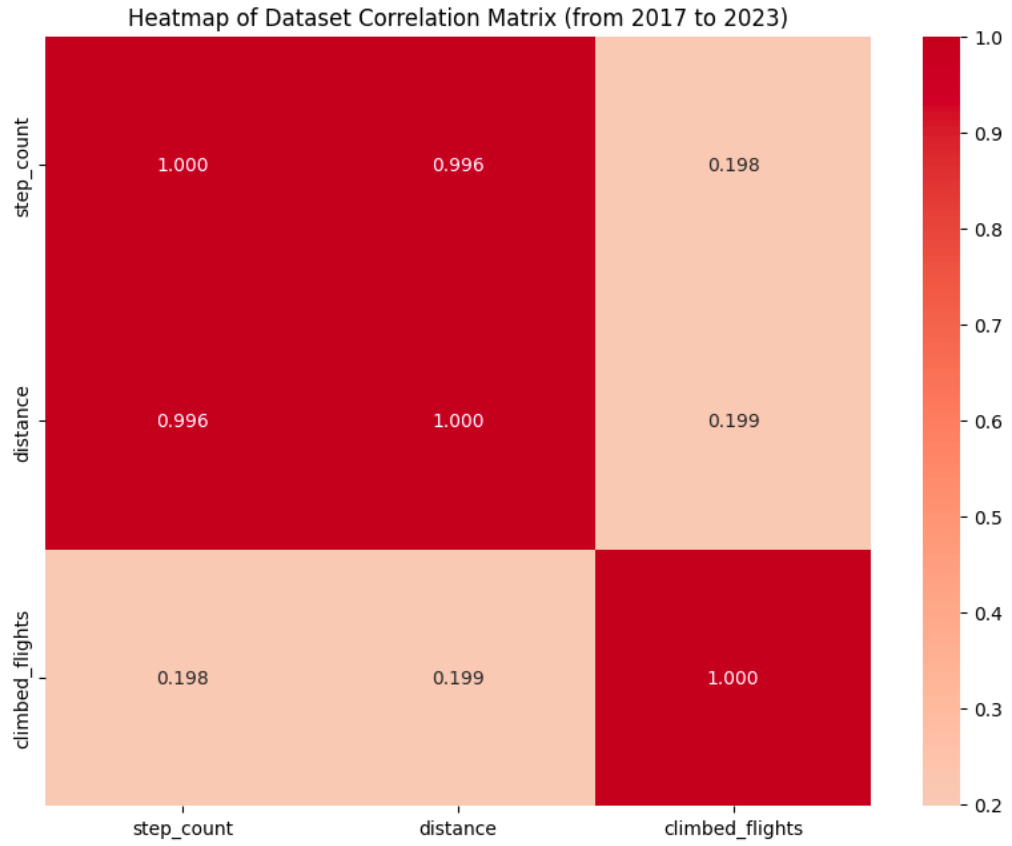


Figure 1: The correlation matrix of the data types available from 2017 to 2023.

[1] The correlations that could not be determined and would be given as NaN in the outputs were counted as 0 by using .fillna() method of Pandas while calculating the correlations.

As it can be observed from the matrix, if the lower boundary for strong correlation is to be accepted as 0.5 in terms of the absolute value of the correlation coefficient, the only data types between which there is a strong correlation were the step count and the distance of walking and running, with a correlation coefficient of 0.996, indicating a strong positive correlation, which can also be observed in the form of a scatter plot as in the Appendix JJ. On the other hand, the step count and the number of climbed flights were weakly and positively correlated with a coefficient of 0.198, which can be found in Appendix KK, while a similar situation occurred with the distance of walking and running and the number of climbed flights, where the correlation was expressed with the coefficient equal to 0.199.

The second correlation matrix, given in Figure 2, covers four years of data, starting from the installation of iOS 13.1 on the device, which is the starting of availability of the level of audio from the headphones.

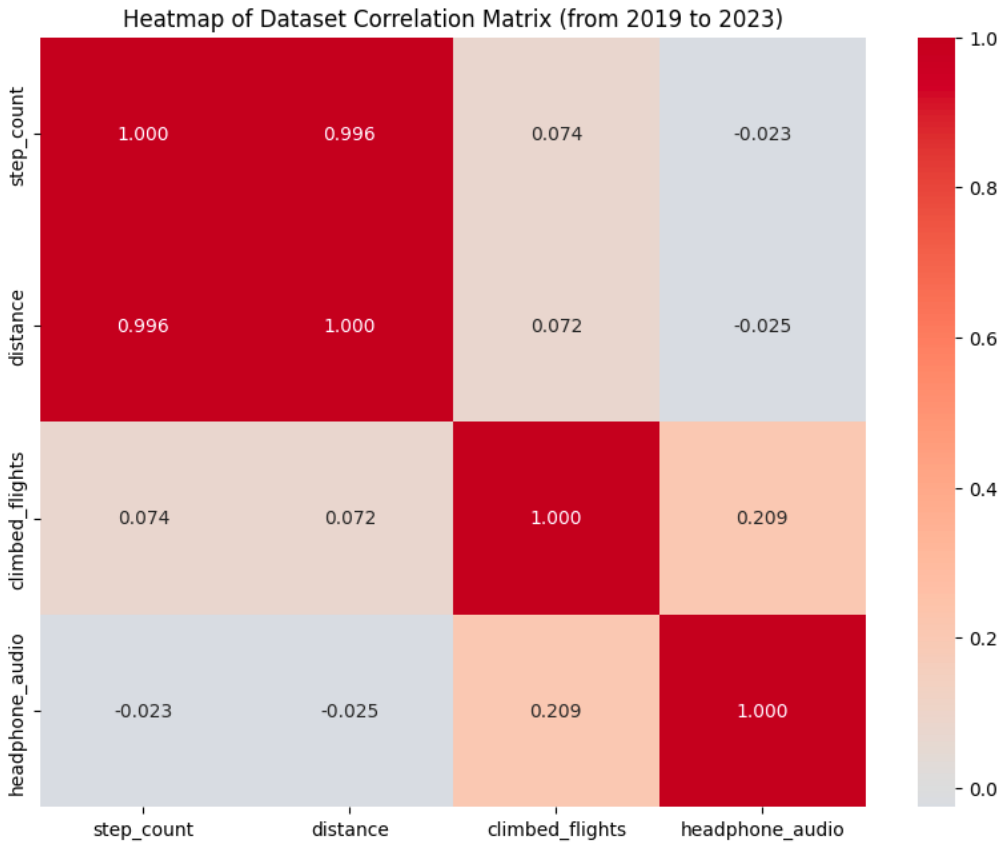


Figure 2: The correlation matrix of the data types available from 2019 to 2023.

In a similar manner to the first correlation matrix, the only data types to have a strong correlation are the step count and the distance of walking and running, where a correlation coefficient of 0.996 was observed. As it can be inferred from the coefficients in the matrix, the other correlations are weak correlations. The notable difference in this matrix is that the level of audio from the headphones is negatively correlated to the step count, as it can be observed in Appendix LL, and the walking and running distances with the respective coefficients of -0.023 and -0.025, which indicates that the values recorded within these two data types decrease as the audio level while using headphones increases.

The third correlation matrix contains the data collected within three years, from the release of iOS 14 in 2020 to the end of the timeframe.

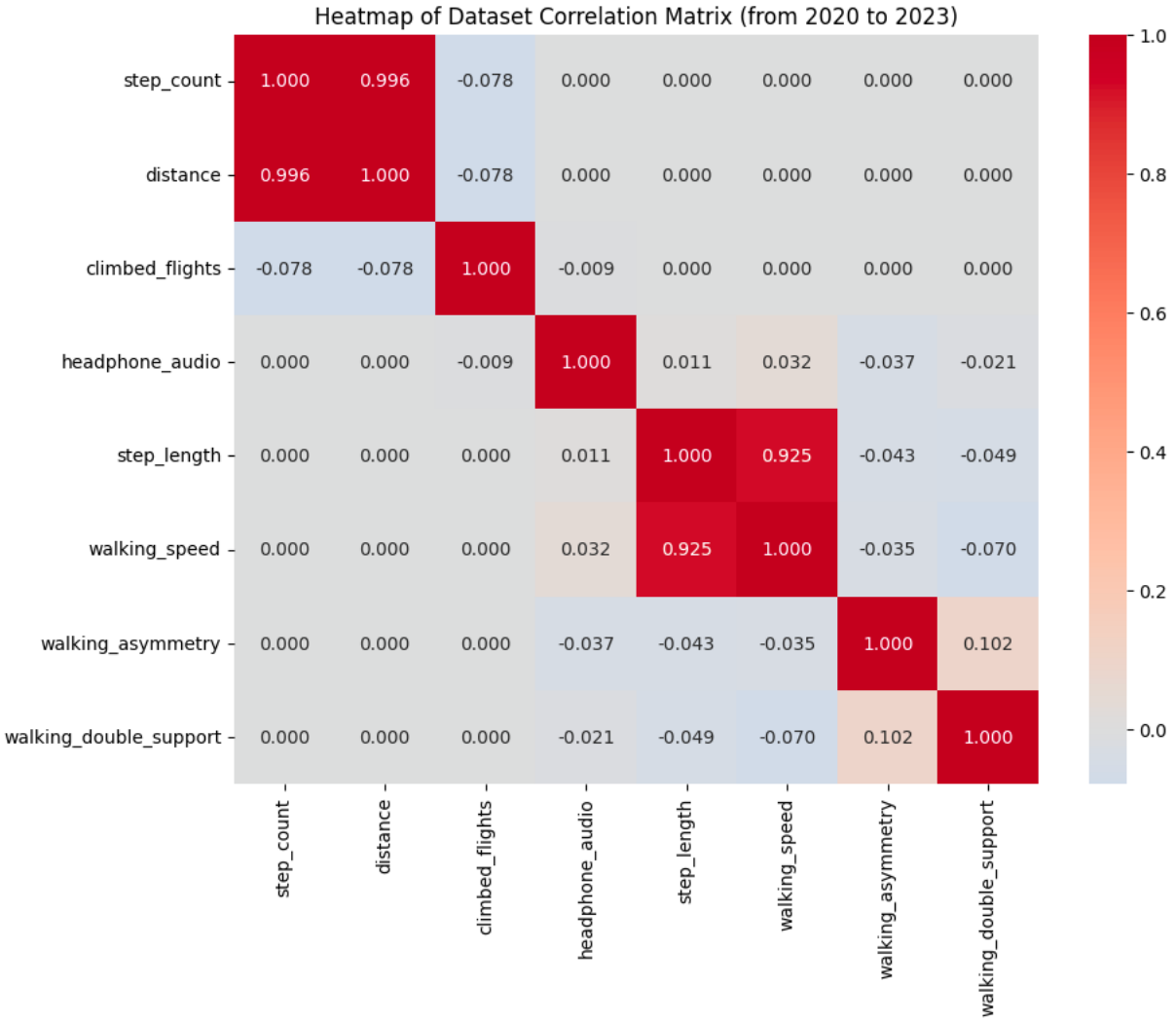


Figure 3: The correlation matrix of the data types available from 2020 to 2023.

Diverging from the previous correlation matrices, there are two strong positive correlations noticed in this matrix: between the step count and the distance of walking and running, with the same correlation coefficient as before, and between the step length and the speed of walking, with a correlation coefficient of 0.925. Other positive correlations were recorded between the percentages of walking asymmetry and walking double support, between the walking speed and the audio level from headphones, and between the step length and the audio level from headphones, meaning that as the audio level increases in the headphones, both the speed of walking and the length of a particular step also increase. Besides that, there are negative correlations between some of the data types: for instance, step length and the percentage

of walking asymmetry has a negative correlation with the coefficient equal to -0.043, indicating that as the step length increases, the overall walking of the author becomes more and more symmetric, even though it may be unnoticeable in the macroscale.

The last correlation matrix covers all of the data types that can be analyzed numerically; as a consequence, it contains only one year of data, beginning with the release of iOS 16, where the remaining data types became available.

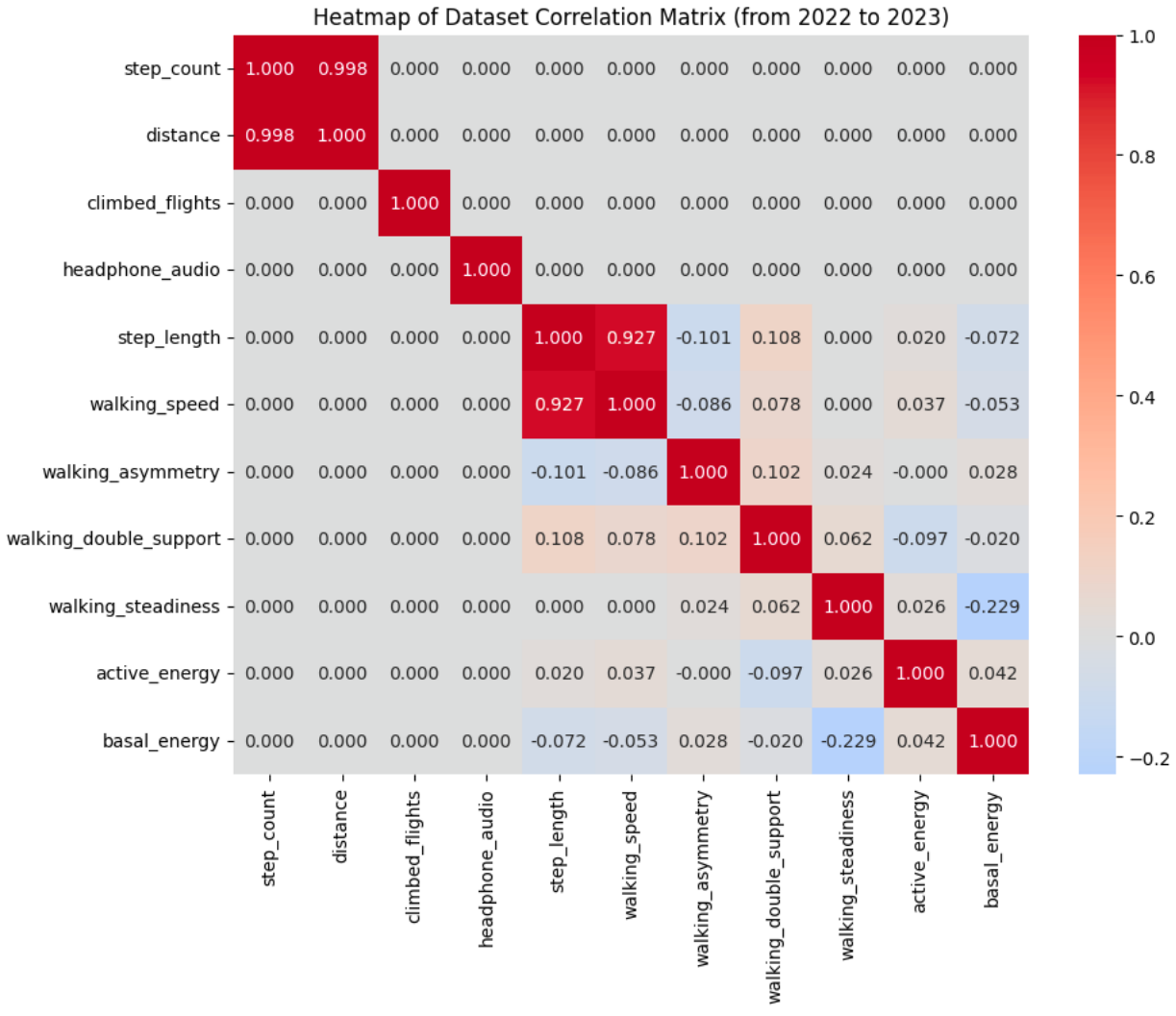


Figure 4: The correlation matrix of the data types available from 2022 to 2023.

As it was the case with the previous correlation matrices, the only strong correlations have taken place between the step count and the distance of walking and running, as a positive correlation with the coefficient of 0.998, and between the step length and the walking speed, as a positive correlation which can be mathematically described by the coefficient 0.927. A remarkable circumstance with the correlation data in the matrix is that the amount of burned active energy, which is the type of energy that was supposed to be directly related to the physical activity (*ActiveEnergyBurned*, n.d.), did not have any correlation with the step count or the distance of walking and running; instead, it showed weak positive correlations with the step length, the speed of walking, and the amount of burned basal energy, the energy type burned during the ordinary metabolic activities (*BasalEnergyBurned*, n.d.), with coefficients equal to 0.020, 0.037, and 0.042, respectively. Even though the last correlation may indicate that the amount of burned active energy increases as the amount of burned basal energy increase, which might indirectly connect it to the increase in the level of involvement in physical activity, it seems that it is too early to generate a direct interpretation between the author's active energy consumption and the physical activity solely on the basis of the data constructing the above correlation matrix. This situation may have arisen from the fact that the data considered here covers only one year, which may be inadequate at the stage of conducting the project in terms of determining a correlation, or the confinement that the data were collected using only the smartphone, not a smartwatch or any similar fitness tracker.

2) Hypothesis Tests with respect to Different Categories

As described above, in the relevant sections, the hypothesis tests were conducted by pairing main periods with main periods, subperiods of Period 4 with subperiods of Period 4, and subperiods of Period 1 with subperiods of Period 1, based on the order of educational intensity. P-values for the tests are available in the notebook on the GitHub repository as outputs of the relevant code cells; in this report, the verdicts regarding the null hypothesis are given. The test results utilizing the data from the main periods are presented in the following table, Table 1:

Table 1: The results of the hypothesis tests with the data from main periods.

		Data types				
		Step Count	W&R Distance	Climbed Flights	Burned Active Energy	Burned Basal Energy
Compared main periods	P4 vs. P2	H_0 is rejected	H_0 is rejected	H_0 is not rejected	Data were not available	Data were not available
	P4 vs. P1	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected	Data were not available	Data were not available
	P4 vs. P3	H_0 is rejected	H_0 is rejected	H_0 is rejected	Data were not available	Data were not available
	P2 vs. P1	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected	Data were not available	Data were not available
	P2 vs. P3	H_0 is not rejected	H_0 is not rejected	H_0 is rejected	Data were not available	Data were not available
	P1 vs. P3	H_0 is rejected	H_0 is rejected	H_0 is rejected	Data were not available	Data were not available

When the results from the tests with main periods are examined, the results show that in terms of step count and distance of walking and running, there is an obvious pattern verifying the main claim in terms of the physical activity in some tests; the averages in Period 4 are significantly greater than the averages in Period 2 and the averages in Period 3; on the contrary, this is not the case when compared to Period 1, in which the education life of the author was less intensive than Period 4. Another exception occurred in the tests where Period 2 was compared to Period 1 and Period 3: The averages did not differ significantly between the periods. This situation may have stemmed from the fact that Period 2 covers the timeframe when the COVID-19 pandemic was at its peaks. Other than those circumstances, there is not anything which diverges from the main claim in the tests where the data of step count and walking and running distances were utilized.

Similarly, the tests conducted using the data of climbed flights deviate from the other tests with respect to the results in only one case, where Period 4 and Period 2 were compared. The reason for such a deviation might be that even though the author is involved in physical activities significantly more in Period 4 than Period 2, he is mostly engaged in buildings that have at most three flights at the campus of the university. Another possible reason may be that in some subperiods of Period 4, the education was conducted as online because of various reasons, such subperiods might have reduced the average of the number of flights in Period 4. Regardless of both prospects, a more systematic approach and research might be required so as to gain a better comprehension of this situation.

The second category of the hypothesis tests were applied by matching subperiods of Period 4 in order to compare different terms and educational conditions within the context of university. The results of these tests are presented on Table 2.

Table 2: The results of the hypothesis tests with the data from subperiods of Period 4.

		Data types				
		Step Count	W&R Distance	Climbed Flights	Burned Active Energy	Burned Basal Energy
Compared subperiods	P4.5 vs. P4.4	H_0 is rejected	H_0 is rejected	H_0 is not rejected	H_0 is rejected	H_0 is not rejected
	P4.5 vs. P4.3	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected
	P4.5 vs. P4.2	H_0 is rejected	H_0 is rejected	H_0 is not rejected	Data were not available	Data were not available

	P4.5 vs. P4.1	H_0 is rejected	H_0 is rejected	H_0 is not rejected	Data were not available	Data were not available
	P4.5 vs. P4.b4	H_0 is rejected	H_0 is rejected	H_0 is not rejected	H_0 is rejected	H_0 is not rejected
	P4.5 vs. P4.b3	H_0 is rejected	H_0 is rejected	H_0 is not rejected	H_0 is rejected	H_0 is not rejected
	P4.5 vs. P4.b2	H_0 is rejected	H_0 is rejected	H_0 is rejected	H_0 is not rejected	H_0 is not rejected
	P4.5 vs. P4.b1	H_0 is rejected	H_0 is rejected	H_0 is rejected	Data were not available	Data were not available
	P4.4 vs. P4.3	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected
	P4.4 vs. P4.2	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected	Data were not available	Data were not available
	P4.4 vs. P4.1	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected	Data were not available	Data were not available
	P4.4 vs. P4.b4	H_0 is not rejected	H_0 is not rejected	H_0 is rejected	H_0 is not rejected	H_0 is not rejected
	P4.4 vs. P4.b3	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected

	P4.4 vs. P4.b2	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected
	P4.4 vs. P4.b1	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected	Data were not available	Data were not available
	P4.3 vs. P4.2	H_0 is rejected	H_0 is rejected	H_0 is not rejected	Data were not available	Data were not available
	P4.3 vs. P4.1	H_0 is rejected	H_0 is rejected	H_0 is not rejected	Data were not available	Data were not available
	P4.3 vs. P4.b4	H_0 is rejected	H_0 is rejected	H_0 is rejected	H_0 is rejected	H_0 is not rejected
	P4.3 vs. P4.b3	H_0 is rejected	H_0 is rejected	H_0 is not rejected	H_0 is rejected	H_0 is not rejected
	P4.3 vs. P4.b2	H_0 is rejected	H_0 is rejected	H_0 is rejected	H_0 is not rejected	H_0 is not rejected
	P4.3 vs. P4.b1	H_0 is rejected	H_0 is rejected	H_0 is rejected	Data were not available	Data were not available
	P4.2 vs. P4.1	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected	Data were not available	Data were not available
	P4.2 vs. P4.b4	H_0 is not rejected	H_0 is not rejected	H_0 is rejected	Data were not available	Data were not available

	P4.2 vs. P4.b3	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected	Data were not available	Data were not available
	P4.2 vs. P4.b2	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected	Data were not available	Data were not available
	P4.2 vs. P4.b1	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected	Data were not available	Data were not available
	P4.1 vs. P4.b4	H_0 is not rejected	H_0 is not rejected	H_0 is rejected	Data were not available	Data were not available
	P4.1 vs. P4.b3	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected	Data were not available	Data were not available
	P4.1 vs. P4.b2	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected	Data were not available	Data were not available
	P4.1 vs. P4.b1	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected	Data were not available	Data were not available
	P4.b4 vs. P4.b3	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected
	P4.b4 vs. P4.b2	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected
	P4.b4 vs. P4.b1	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected	Data were not available	Data were not available

	P4.b3 vs. P4.b2	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected
	P4.b3 vs. P4.b1	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected	Data were not available	Data were not available
	P4.b2 vs. P4.b1	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected	Data were not available	Data were not available

By comparing the results in this table, it can be observed that when the tests with the step counts and the distances of walking and running are considered, Period 4.5 and Period 4.3, the only terms with entirely face-to-face education in the timeframe of university, clearly outperform the other subperiods of Period 4: the average step count and the average walking and running distance in these two subperiods are significantly greater than any other subperiod in Period 4. However, the two subperiods, Period 4.5 and Period 4.3, do not differ on the basis of the averages of these values in a significant manner according to the results, since the null hypothesis is not rejected in that test. These conclusions are also supported by the tests utilizing the amount of burned active energy when such data were available: it is entirely compliant with the aforementioned results, not demonstrating any particular divergence in terms of the decision regarding the null hypothesis.

As it was the case with the tests on main periods, the outcomes of the tests with the data of climbed flights also differ from the tests with other data types between the subperiods of Period 4. Precisely, the null hypothesis, which is that the means of the two samples do not differ significantly from each other, fails to be rejected when all the terms that are not the subperiods of breaks, regardless of the mode of education, are subjected to the tests. The previously explained situation about the buildings in the campus, along with the fact that the break subperiods mostly pass on the home for the author, could have given rise to such a distinction in terms of the test results. In fact, it can be argued that these possible reasons are validified by the rejection of null

hypothesis in the tests where the fully face-to-face terms are compared to the break subperiods most frequently.

Another distinguishing fact on the results of the tests concerning the subperiods of Period 4 is that the null hypothesis fails to be rejected in all of the tests using the amount of burned basal energy, including the comparisons of terms with face-to-face education to each other and to the break subperiods when the data were available. This outcome shows that the author's metabolism remains unaffected of changes in his physical activity, maintaining a steady level of basal energy consumption most of the time.

The third, and the last category of hypothesis tests was formed by the ones where the data from subperiods of Period 1 were subjected to the T-tests. The conclusions concerning the null hypothesis are given as follows on Table 3:

Table 3: The results of the hypothesis tests with the data from subperiods of Period 1.

		Data types				
		Step Count	W&R Distance	Climbed Flights	Burned Active Energy	Burned Basal Energy
Compared subperiods	P1.6 vs. P1.5	H_0 is not rejected	H_0 is not rejected	H_0 is rejected	Data were not available	Data were not available
	P1.6 vs. P1.4	H_0 is not rejected	H_0 is not rejected	H_0 is rejected	Data were not available	Data were not available
	P1.6 vs. P1.3	H_0 is not rejected	H_0 is not rejected	H_0 is rejected	Data were not available	Data were not available

	P1.6 vs. P1.2	H_0 is not rejected	H_0 is not rejected	H_0 is rejected	Data were not available	Data were not available
	P1.6 vs. P1.1	H_0 is not rejected	H_0 is not rejected	H_0 is rejected	Data were not available	Data were not available
	P1.6 vs. P1.b5	H_0 is rejected	H_0 is rejected	H_0 is not rejected	Data were not available	Data were not available
	P1.6 vs. P1.b4	H_0 is rejected	H_0 is rejected	H_0 is rejected	Data were not available	Data were not available
	P1.6 vs. P1.b3	H_0 is rejected	H_0 is rejected	H_0 is rejected	Data were not available	Data were not available
	P1.6 vs. P1.b2	H_0 is rejected	H_0 is rejected	H_0 is rejected	Data were not available	Data were not available
	P1.6 vs. P1.b1	H_0 is rejected	H_0 is rejected	H_0 is rejected	Data were not available	Data were not available
	P1.5 vs. P1.4	H_0 is not rejected	H_0 is not rejected	H_0 is rejected	Data were not available	Data were not available
	P1.5 vs. P1.3	H_0 is not rejected	H_0 is not rejected	H_0 is rejected	Data were not available	Data were not available
	P1.5 vs. P1.2	H_0 is not rejected	H_0 is not rejected	H_0 is rejected	Data were not available	Data were not available

	P1.5 vs. P1.1	H_0 is not rejected	H_0 is not rejected	H_0 is rejected	Data were not available	Data were not available
	P1.5 vs. P1.b5	H_0 is rejected	H_0 is rejected	H_0 is not rejected	Data were not available	Data were not available
	P1.5 vs. P1.b4	H_0 is rejected	H_0 is rejected	H_0 is rejected	Data were not available	Data were not available
	P1.5 vs. P1.b3	H_0 is rejected	H_0 is rejected	H_0 is rejected	Data were not available	Data were not available
	P1.5 vs. P1.b2	H_0 is rejected	H_0 is rejected	H_0 is rejected	Data were not available	Data were not available
	P1.5 vs. P1.b1	H_0 is rejected	H_0 is rejected	H_0 is rejected	Data were not available	Data were not available
	P1.4 vs. P1.3	H_0 is rejected	H_0 is rejected	H_0 is rejected	Data were not available	Data were not available
	P1.4 vs. P1.2	H_0 is rejected	H_0 is rejected	H_0 is rejected	Data were not available	Data were not available
	P1.4 vs. P1.1	H_0 is rejected	H_0 is rejected	H_0 is rejected	Data were not available	Data were not available
	P1.4 vs. P1.b5	H_0 is rejected	H_0 is rejected	H_0 is not rejected	Data were not available	Data were not available

	P1.4 vs. P1.b4	H_0 is rejected	H_0 is rejected	H_0 is rejected	Data were not available	Data were not available
	P1.4 vs. P1.b3	H_0 is rejected	H_0 is rejected	H_0 is rejected	Data were not available	Data were not available
	P1.4 vs. P1.b2	H_0 is rejected	H_0 is rejected	H_0 is rejected	Data were not available	Data were not available
	P1.4 vs. P1.b1	H_0 is rejected	H_0 is rejected	H_0 is rejected	Data were not available	Data were not available
	P1.3 vs. P1.2	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected	Data were not available	Data were not available
	P1.3 vs. P1.1	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected	Data were not available	Data were not available
	P1.3 vs. P1.b5	H_0 is rejected	H_0 is rejected	H_0 is not rejected	Data were not available	Data were not available
	P1.3 vs. P1.b4	H_0 is rejected	H_0 is rejected	H_0 is rejected	Data were not available	Data were not available
	P1.3 vs. P1.b3	H_0 is rejected	H_0 is rejected	H_0 is rejected	Data were not available	Data were not available
	P1.3 vs. P1.b2	H_0 is rejected	H_0 is rejected	H_0 is rejected	Data were not available	Data were not available

	P1.3 vs. P1.b1	H_0 is rejected	H_0 is rejected	H_0 is rejected	Data were not available	Data were not available
	P1.2 vs. P1.1	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected	Data were not available	Data were not available
	P1.2 vs. P1.b5	H_0 is rejected	H_0 is rejected	H_0 is not rejected	Data were not available	Data were not available
	P1.2 vs. P1.b4	H_0 is rejected	H_0 is rejected	H_0 is rejected	Data were not available	Data were not available
	P1.2 vs. P1.b3	H_0 is rejected	H_0 is rejected	H_0 is rejected	Data were not available	Data were not available
	P1.2 vs. P1.b2	H_0 is rejected	H_0 is rejected	H_0 is rejected	Data were not available	Data were not available
	P1.2 vs. P1.b1	H_0 is rejected	H_0 is rejected	H_0 is rejected	Data were not available	Data were not available
	P1.1 vs. P1.b5	H_0 is rejected	H_0 is rejected	H_0 is not rejected	Data were not available	Data were not available
	P1.1 vs. P1.b4	H_0 is rejected	H_0 is rejected	H_0 is rejected	Data were not available	Data were not available
	P1.1 vs. P1.b3	H_0 is rejected	H_0 is rejected	H_0 is rejected	Data were not available	Data were not available

	P1.1 vs. P1.b2	H_0 is rejected	H_0 is rejected	H_0 is rejected	Data were not available	Data were not available
	P1.1 vs. P1.b1	H_0 is rejected	H_0 is rejected	H_0 is rejected	Data were not available	Data were not available
	P1.b5 vs. P1.b4	H_0 is not rejected	H_0 is not rejected	H_0 is rejected	Data were not available	Data were not available
	P1.b5 vs. P1.b3	H_0 is not rejected	H_0 is not rejected	H_0 is rejected	Data were not available	Data were not available
	P1.b5 vs. P1.b2	H_0 is not rejected	H_0 is not rejected	H_0 is rejected	Data were not available	Data were not available
	P1.b5 vs. P1.b1	H_0 is not rejected	H_0 is not rejected	H_0 is rejected	Data were not available	Data were not available
	P1.b4 vs. P1.b3	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected	Data were not available	Data were not available
	P1.b4 vs. P1.b2	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected	Data were not available	Data were not available
	P1.b4 vs. P1.b1	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected	Data were not available	Data were not available
	P1.b3 vs. P1.b2	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected	Data were not available	Data were not available

	P1.b3 vs. P1.b1	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected	Data were not available	Data were not available
	P1.b2 vs. P1.b1	H_0 is not rejected	H_0 is not rejected	H_0 is not rejected	Data were not available	Data were not available

The verdicts regarding the means of the data from the subperiods on the table demonstrate that during the time in which the author used to study in high school, the average step count and the average distance of walking and running significantly differ between the educational terms and the break subperiods, and the averages on the educational terms do not differ that much compared to each other, with one remarkable exception: when Period 1.4 is compared to the subperiods prior to it, the null hypothesis is always rejected in the tests conducted with step counts and walking and running distances. The main cause of this difference can be the hiking which occurred in the Period 1.4, where the teacher of biology took the students to Mount Ganos, a mountain in Tekirdağ, Turkey, the hometown of the author and where he had completed his education of high school, in the Tekirdağ (Ebru Nayim) High School. The failure to reject the null hypothesis also sustains its existence in the comparisons of the break subperiods; it seems that these subperiods are not different from each other in terms of physical activity as the results suggest.

On the contrary to the previous tests involving other periods and the subperiods of the university education of the author, the tests with the data of climbed flights generally do not diverge from the results with the other data types until Period 1.5; Period 1.5 is the subperiod where that difference starts compared to the other subperiods. The circumstance that as the students progress to the upper grades, their classrooms change to the classrooms at the upper flights of the building in the high school might have contributed to this distinction by increasing the frequency of climbing flights for the author with an increasing acceleration starting at the described subperiod. As a shortcoming of the data from these subperiods, due to the release of associated features in iOS 16, the data related to the burned active energy and burned basal energy were not available during Period 1, Period 2, and Period 3, which was why those data

were utilized for the hypothesis tests only in the second category, where the subperiods of Period 4 were analyzed; as a result, those tests cannot be exerted on this category of tests.

Limitations

The primary limitation of the project was related to the availability of some of the features in the dataset; while some of the data types were available for recording in the starting date of the dataset, November 23, 2017, when the most recent iOS version was iOS 11, some of the data types became available in later versions of iOS. As a result, the hypothesis tests for the amount of burned active energy and the amount of burned basal energy could not be conducted for the entire timeframe; those tests were conducted only for the subperiods of Period 4, starting at Period 4.b2. Likewise, due to the aforementioned circumstance, instead of one correlation matrix, four of them were created: one comprising the data recorded during the entire timeframe, one comprising the data starting the release and installation of iOS 13.1 in October 2019, when the level of audio from the headphones became available as a data type, one consisting of the types of data available since the release and installation of iOS 14 in September 2020, when the percentages of walking asymmetry and walking double support, alongside the measurements of step length and the walking speed were introduced, and the last one for the data from September 2022, when iOS 16 was released and the percentage of walking steadiness, the amount of burned active energy, and the amount of burned basal energy were started to be recorded.

Another limitation of the project may be the covered device types: aside from the audio level of headphones, where the EarPods, the default wired headphones by Apple, another wired headphone from Sony, and a wireless headphone, Huawei FreeBuds SE were the three different devices from which the data were recorded, there is only one device type for the other data types, which is an iPhone. Since the author does not use smartwatches or any other wearable device acting as a fitness tracker, the opportunity of collecting even more detailed data does not exist within the context of this project.

Future work

Future work on this project may include a further investigation to Period 2 in the search for a better comprehension of the pandemic and its impacts on the lives of ordinary people, a possible revisit to the analyses after further data is collected as the life of the author progresses as a whole, including the educational life in university, mandatory internship and the later ones, a probable study of master's degree, and the business life as possible milestones for the future of his life, and a possible development of a machine learning model in order to predict the future statistics of the author related to his physical activity.

Conclusion

The main conclusion which can be inferred from the hypothesis tests can be that the primary claim, stating that physical activity of the author increases as the intensity of his education life increases, is actually the case when the subperiods with education terms are compared to the subperiods of breaks. This situation is also valid when the subperiods with face-to-face education are compared to the other subperiods, particularly in the context of Period 4, further verifying the claim in this perspective. On the other hand, aside from the distinctions between the education terms in Period 4 and the hiking in Period 1.4, the levels of physical activity of the author are not distinguishable in a significant manner on average between different educational terms. Besides that, even though Period 2 was remarkably more intensive than Period 1 and Period 3, the comparison of Period 2 with the two main periods have failed to reject the null hypothesis, indicating there is not any noticeable change on average in terms of physical activity, most probably due to the COVID-19 pandemic and related precautions. To sum up, the physical activity of the author obviously increases from the first approach, that is, the author is physically more active in educational terms than he is in the break subperiods. The second approach, i.e., the one considering the changing educational intensities within different educational terms, on the contrary, needs more data collection and analysis due to extraordinary situation with the main period concerning what may be one of the most important parts in an individual's education life in Turkey, and also most of the subperiods of Period 4, namely, their overlapping with the pandemic and the later imposition of online education in the afterwards of the earthquakes.

This project presented a not-so-much-considered insight into the author's life, including for the author himself, primarily. As the data were analyzed in more and more sophisticated manners, the physical life of the author came into the life in different shapes than it used to come. It may be stated that such a project recreates the memories from one's life, just from a completely different perspective which also may create new opportunities and prospects for the future life, which may be the entire point of the past and the memories as a whole.

References

activeEnergyBurned. (n.d.). Apple Developer Documentation. https://developer.apple.com/documentation/healthkit/hkquantitytypeidentifier/1615771-activeenergyburned

basalEnergyBurned. (n.d.). Apple Developer Documentation. https://developer.apple.com/documentation/healthkit/hkquantitytypeidentifier/1615512-basalenergyburned

HKQuantityTypeIdentifierWalkingDoubleSupportPercentage. (n.d.). Apple Developer Documentation. https://developer.apple.com/documentation/healthkit/hkquantitytypeidentifierwalkingdoublesupportpercentage

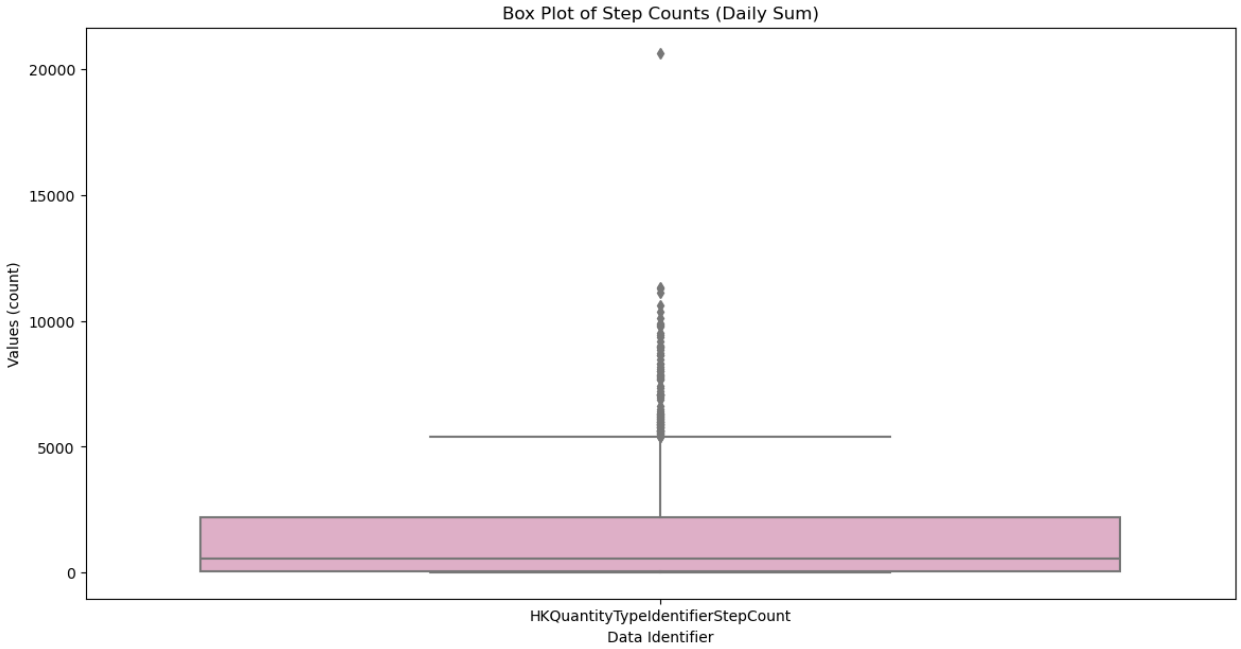
Marcus, J. B. (2013). Weight management: finding the healthy balance. *Culinary Nutrition*, 431–473. https://doi.org/10.1016/b978-0-12-391882-6.00010-8

Tushev, M. (2021, October 26). *Does my sample have to be normally distributed for a t-test?* Medium. https://miroslavtushev.medium.com/does-my-sample-have-to-be-normally-distributed-for-a-t-test-7ee91aaaca2a

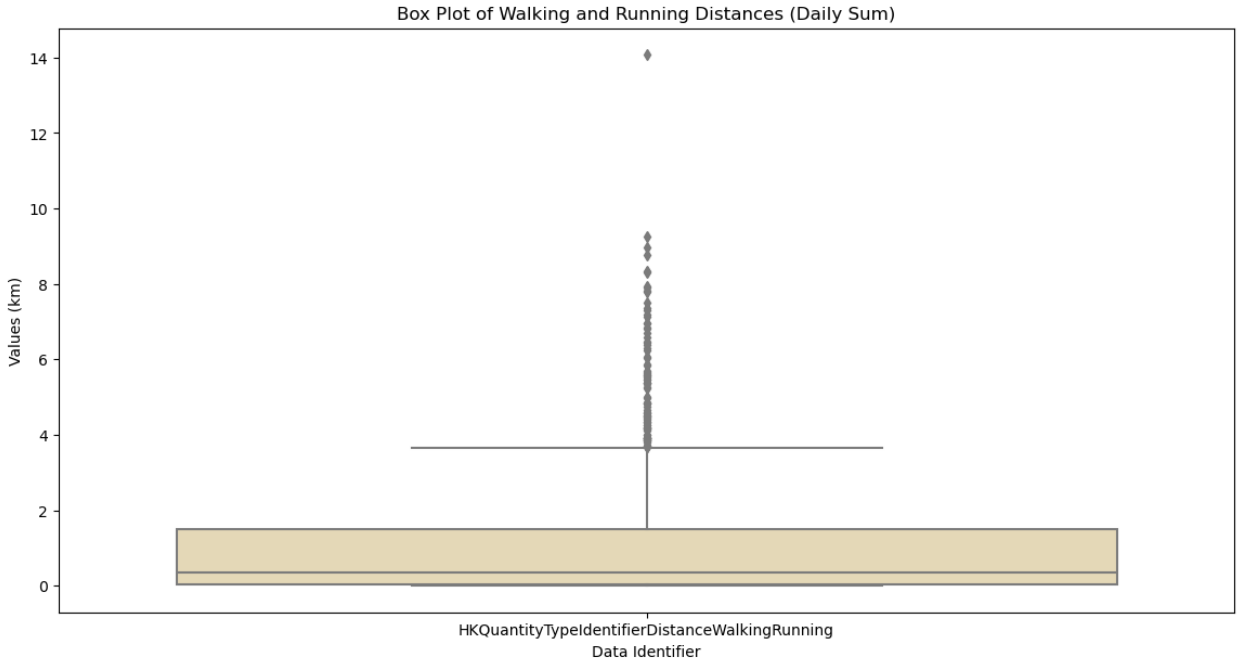
xml.etree.ElementTree — The ElementTree XML API. (n.d.). Python Documentation. https://docs.python.org/3.11/library/xml.etree.elementtree.html

Appendices

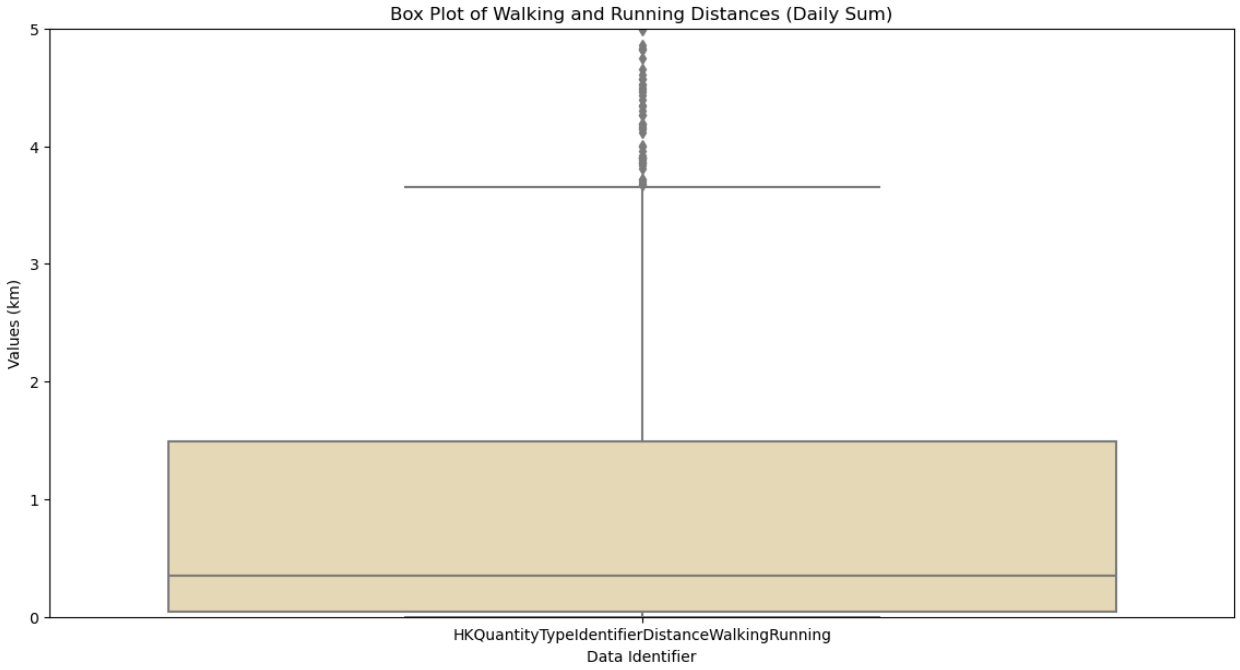
Appendix A: The box plot of the step counts.



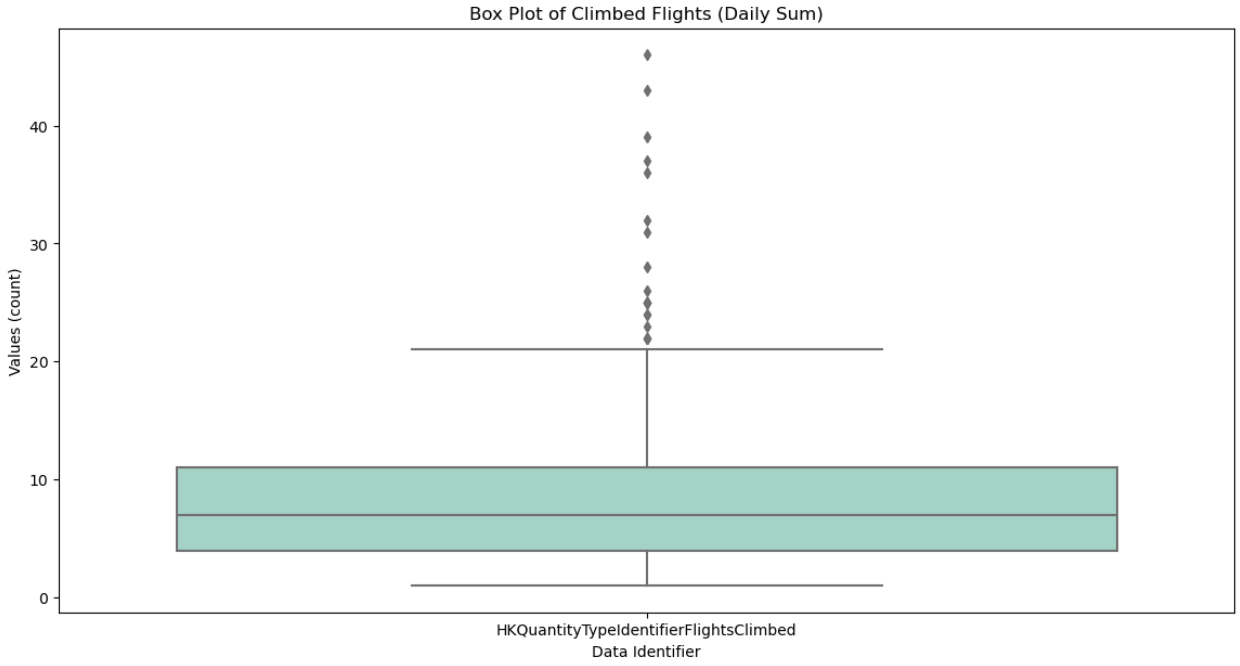
Appendix B: The box plot of walking and running distances.



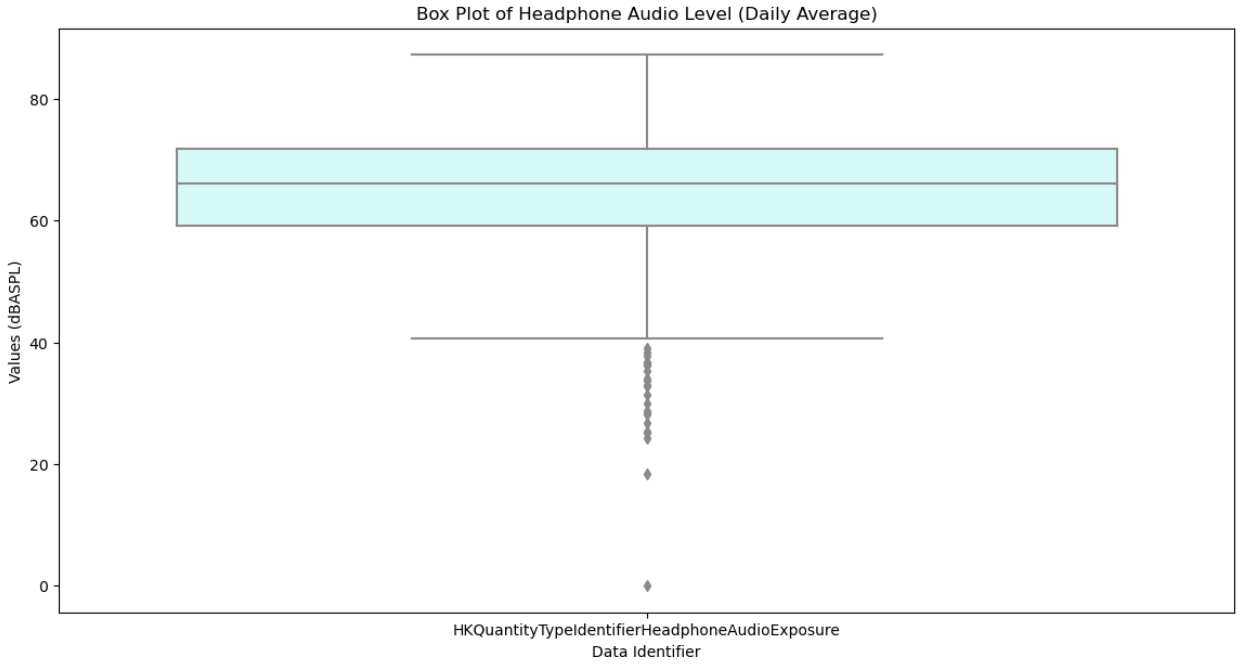
Appendix C: The box plot of the walking and running distances, from a closer perspective.



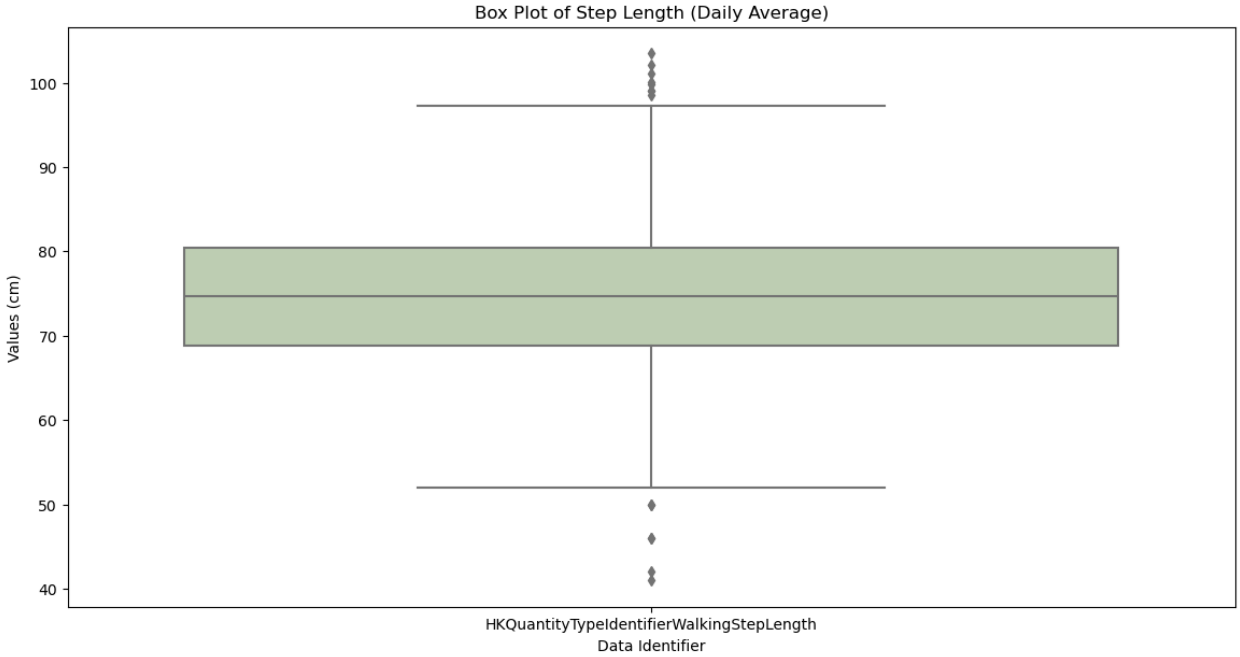
Appendix D: The box plot of climbed flights.



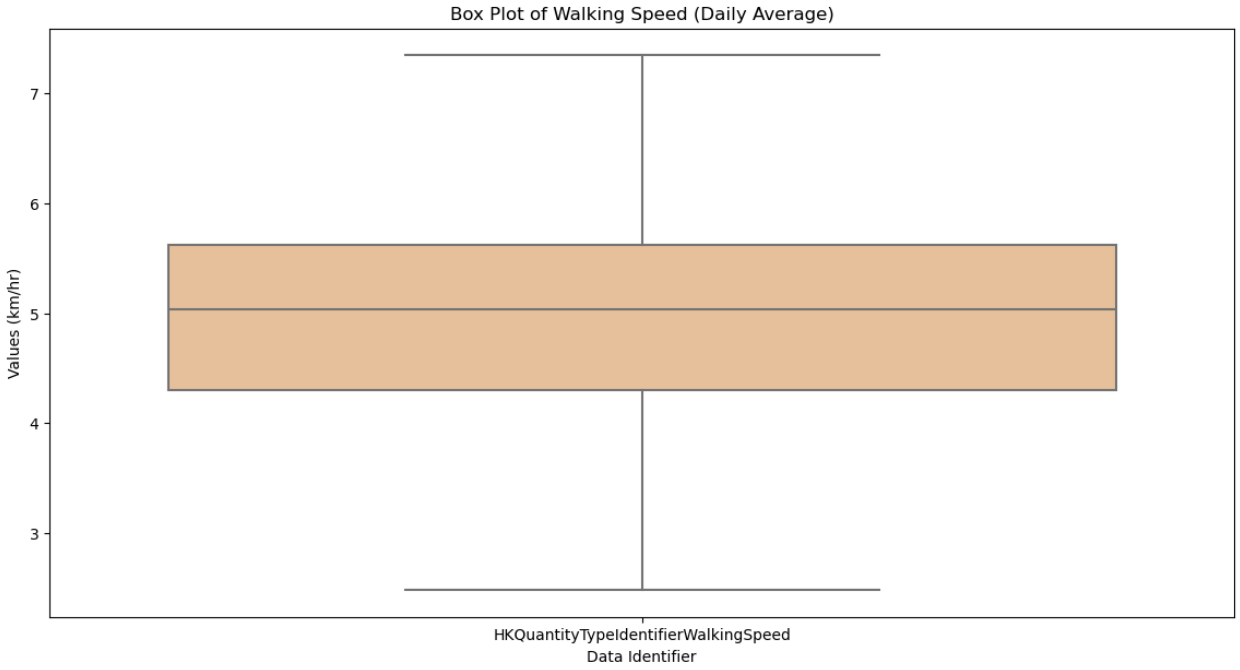
Appendix E: The box plot of headphone audio levels.



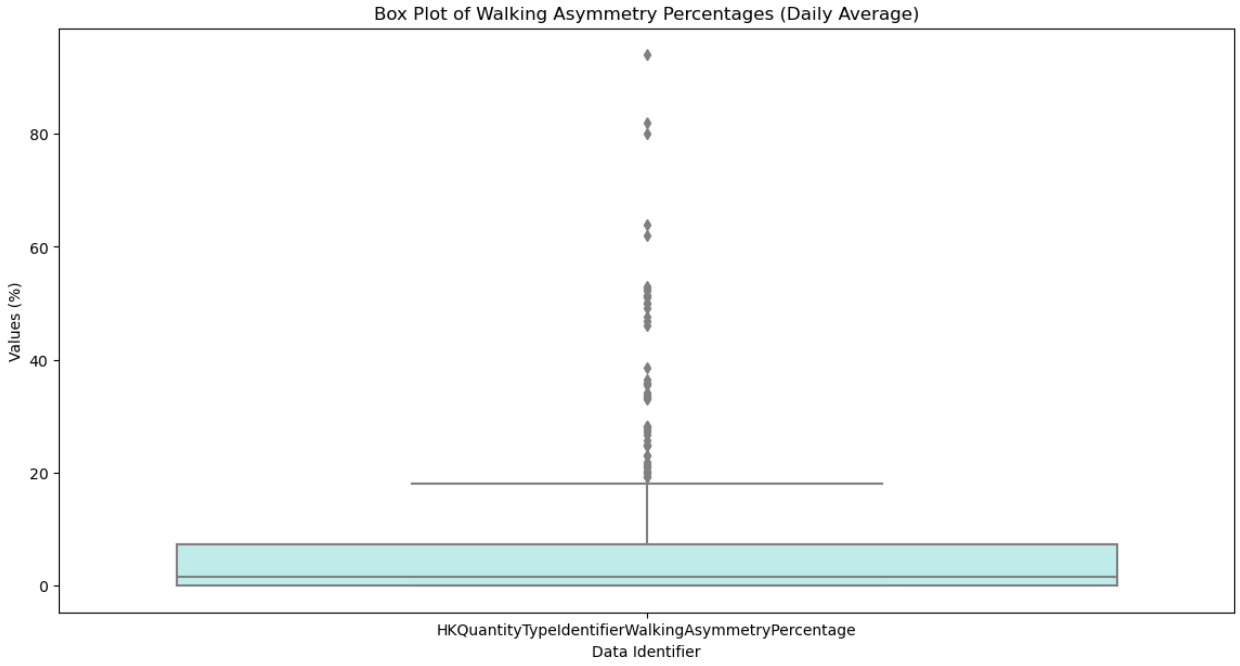
Appendix F: The box plot of step lengths.



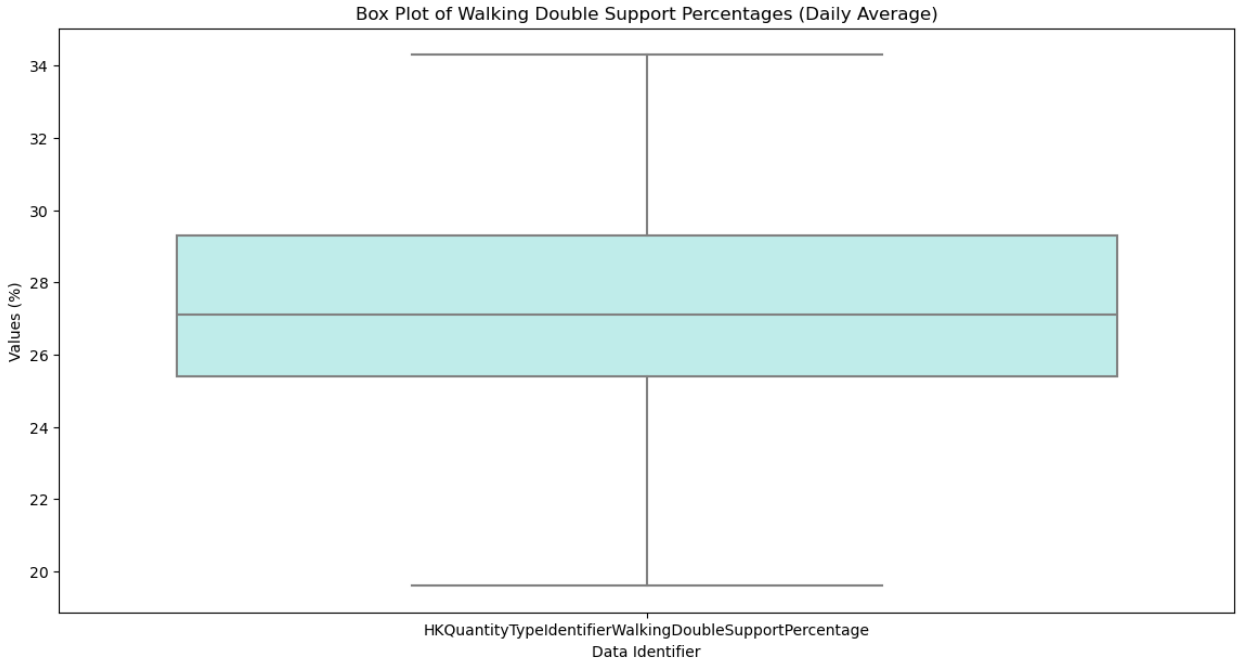
Appendix G: Box plot of walking speeds.



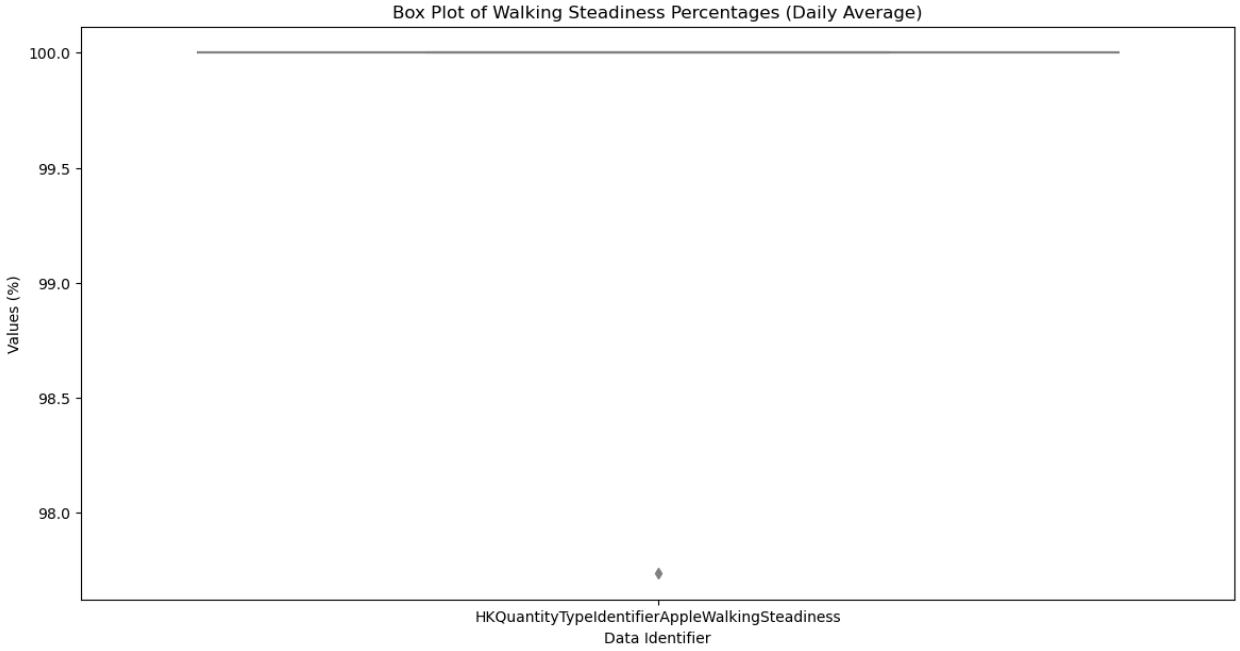
Appendix H: The box plot of walking asymmetry percentages.



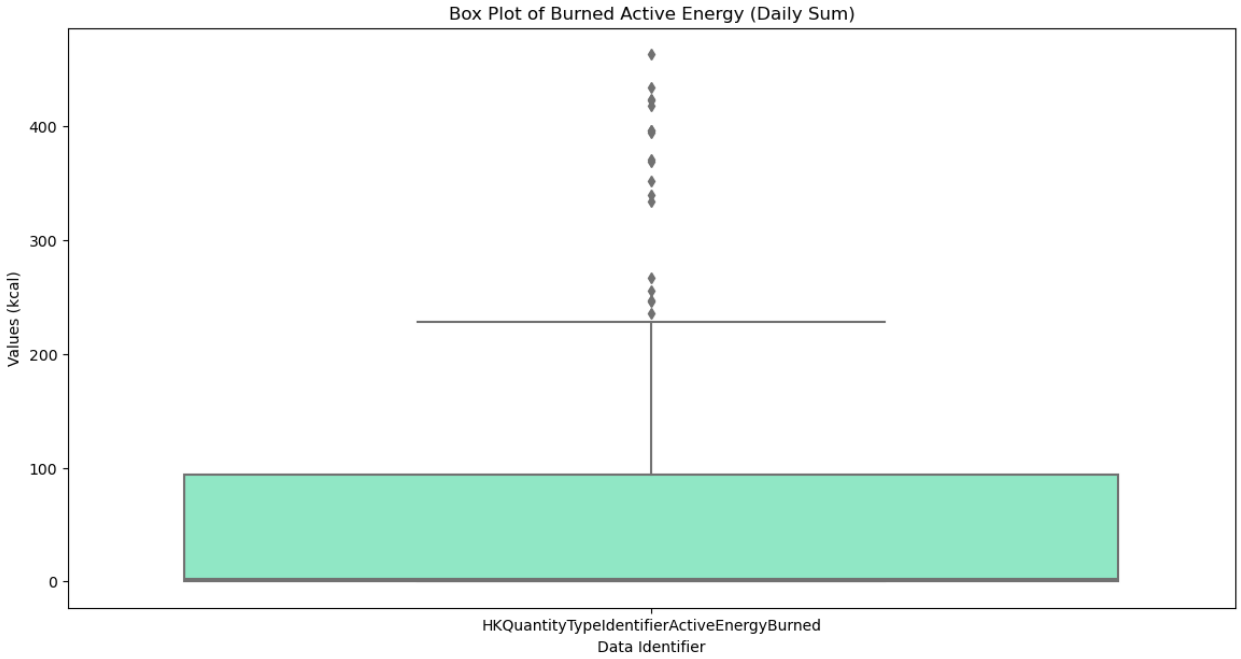
Appendix I: The box plot of walking double support percentages.



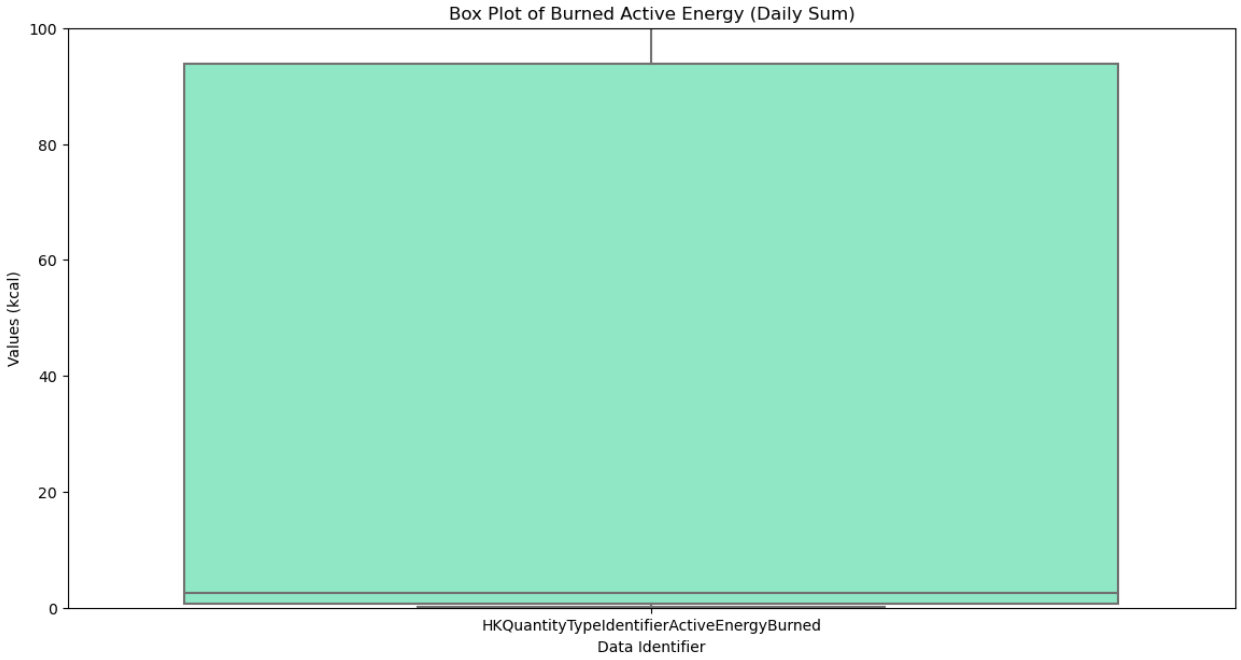
Appendix J: The box plot of walking steadiness percentages.



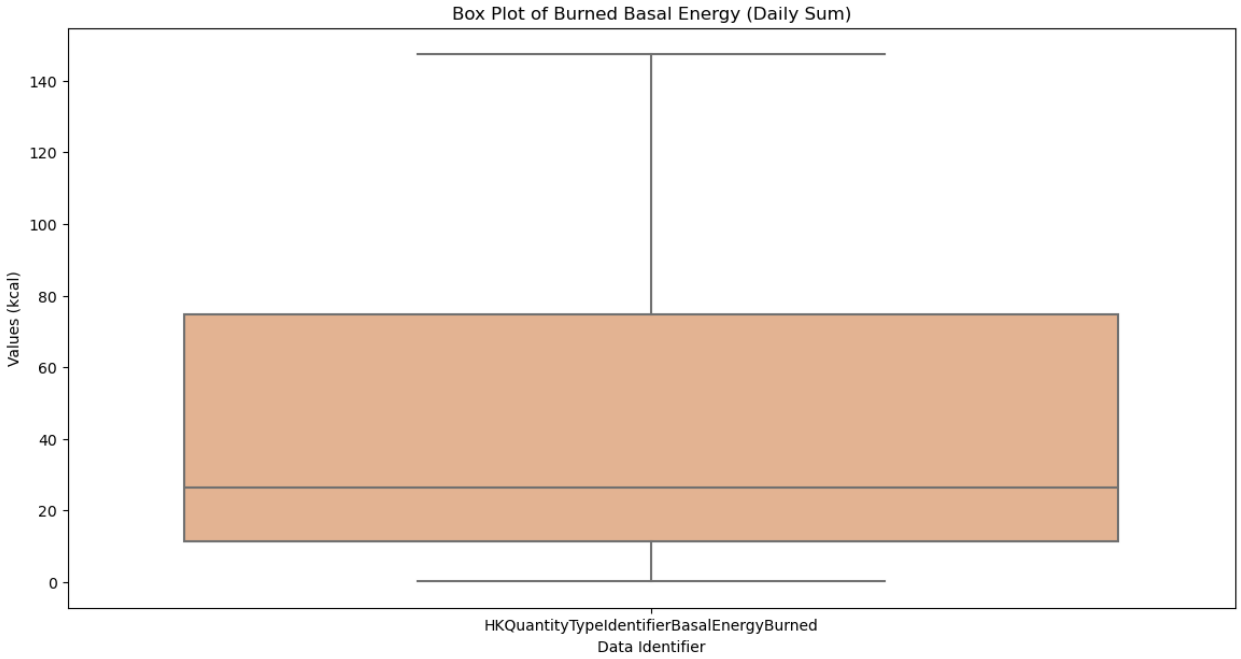
Appendix K: The box plot of the amounts of burned active energy.



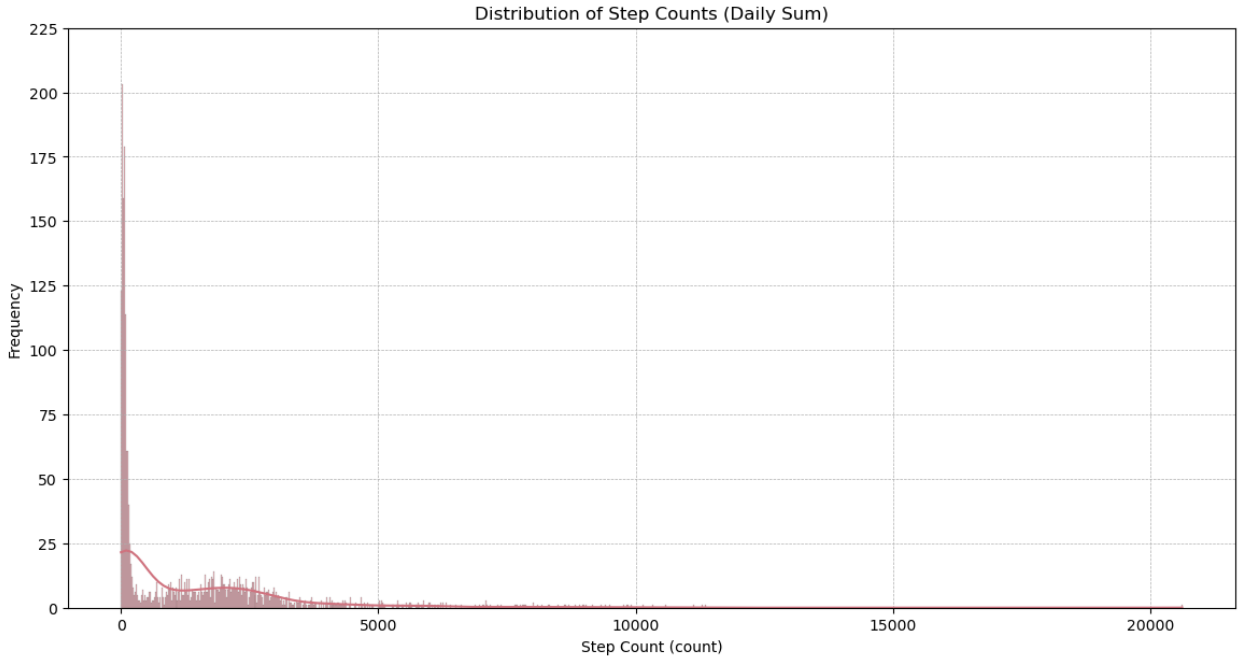
Appendix L: The box plot of the amounts of burned active energy, from a closer perspective.



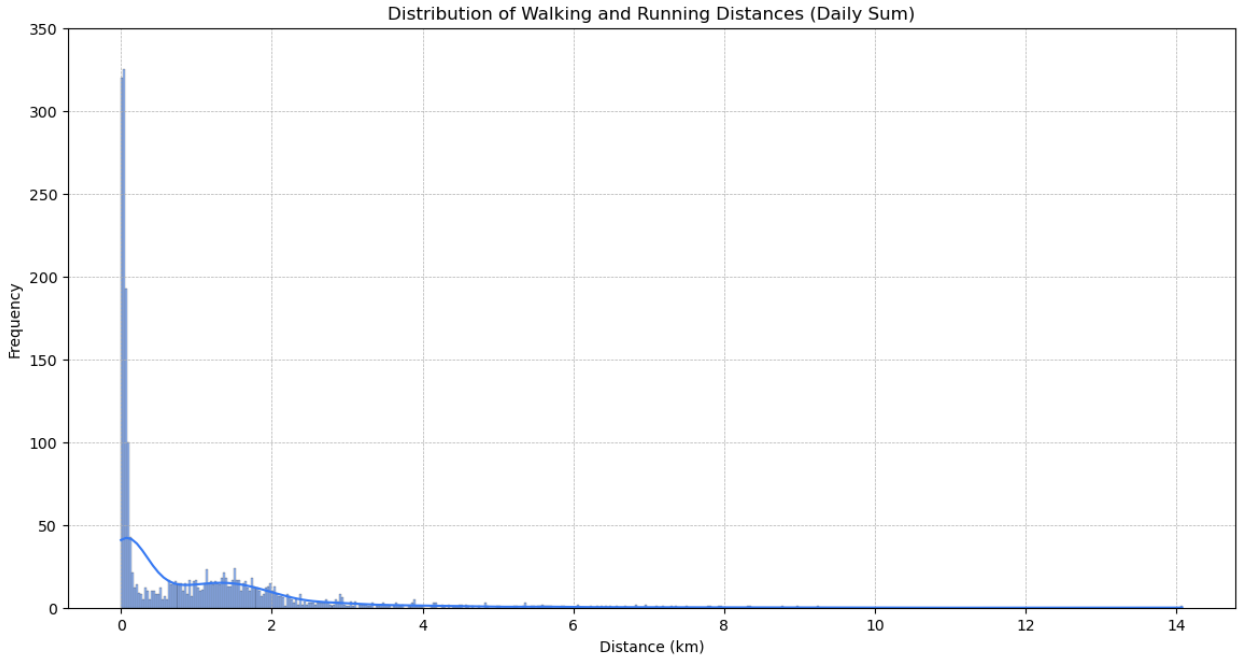
Appendix M: The box plot of the amounts of burned basal energy.



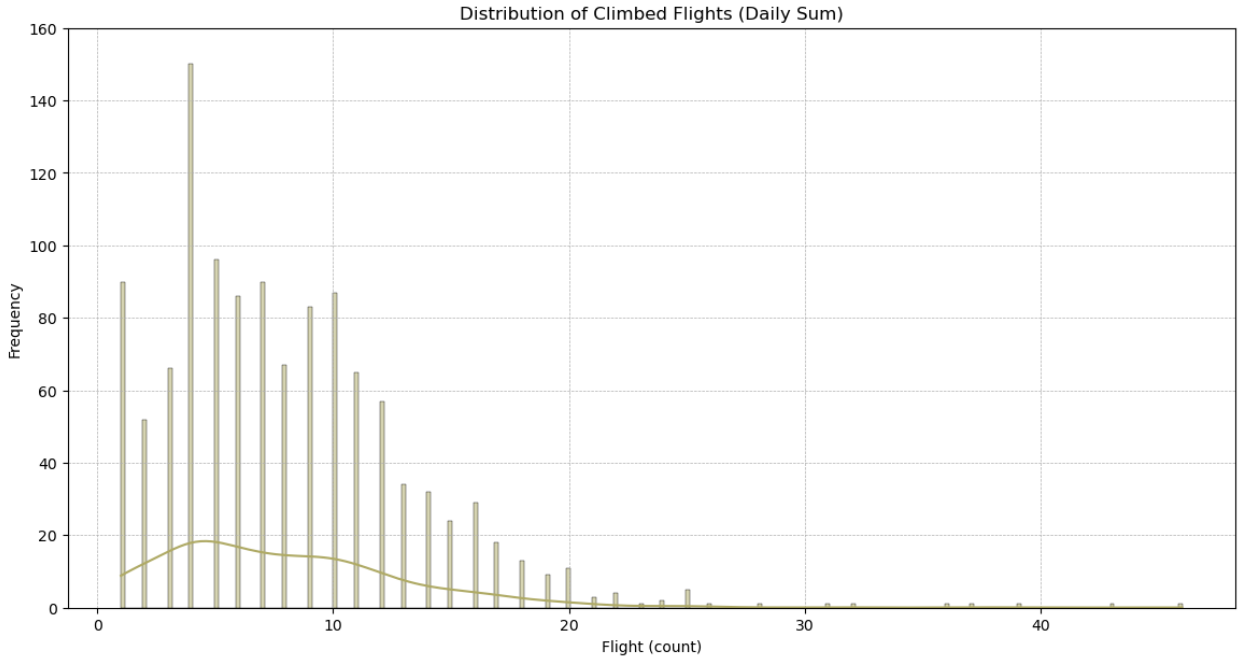
Appendix N: The histogram showing the distribution of step counts.



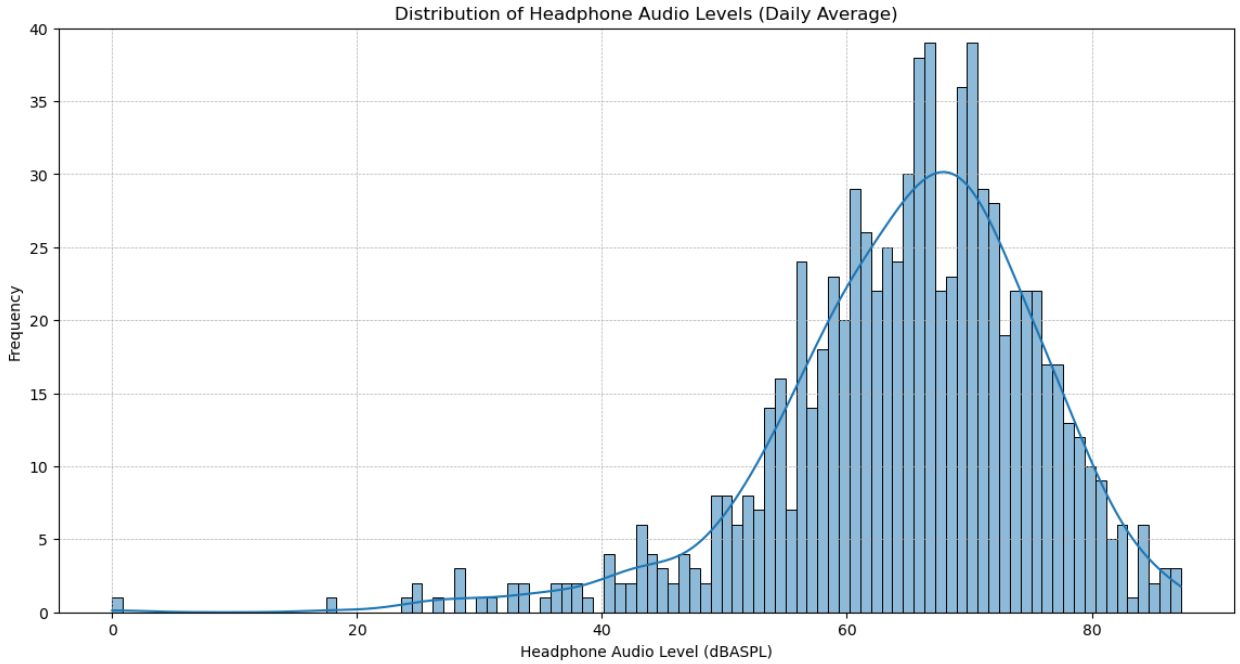
Appendix O: The histogram showing the distribution of walking and running distances.



Appendix P: The histogram showing the distribution of climbed flights.

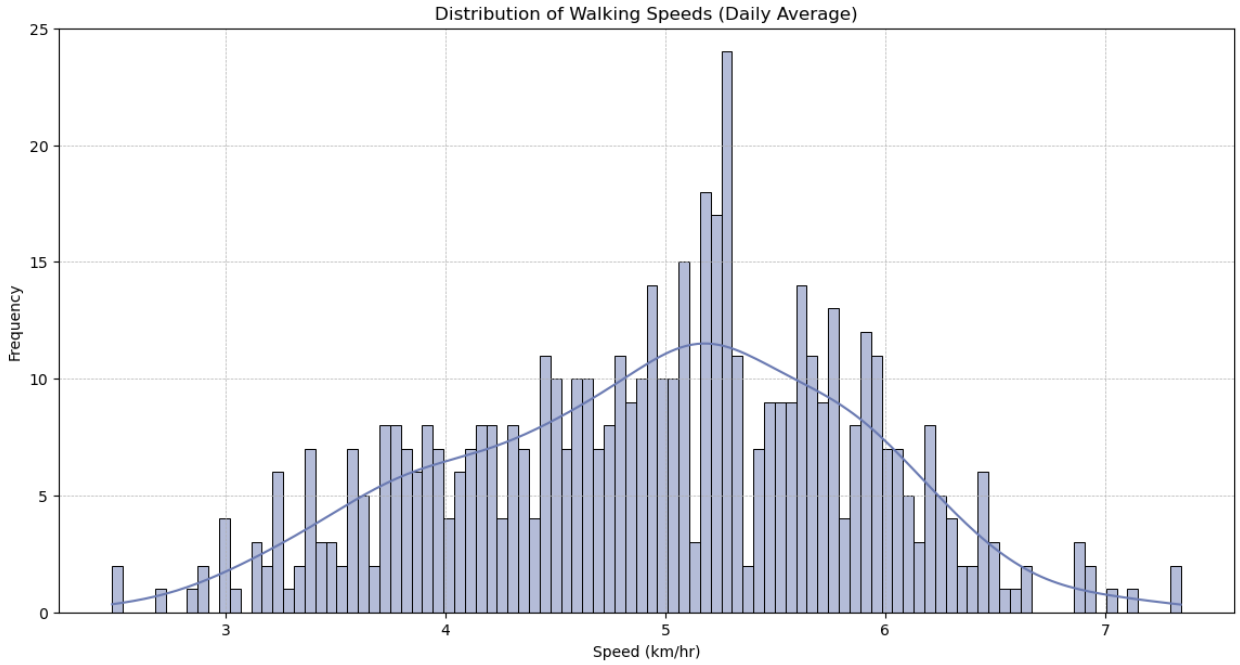


Appendix Q: The histogram showing the distribution of headphone audio levels.

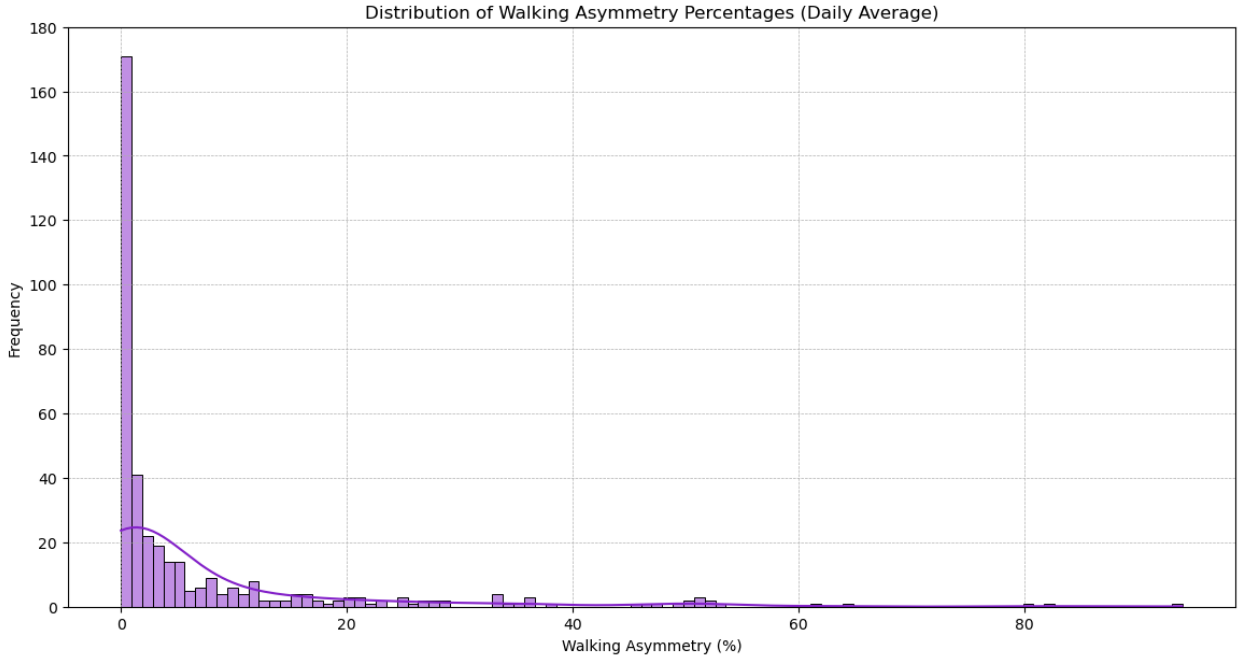


Appendix R: The histogram showing the distribution of step lengths.

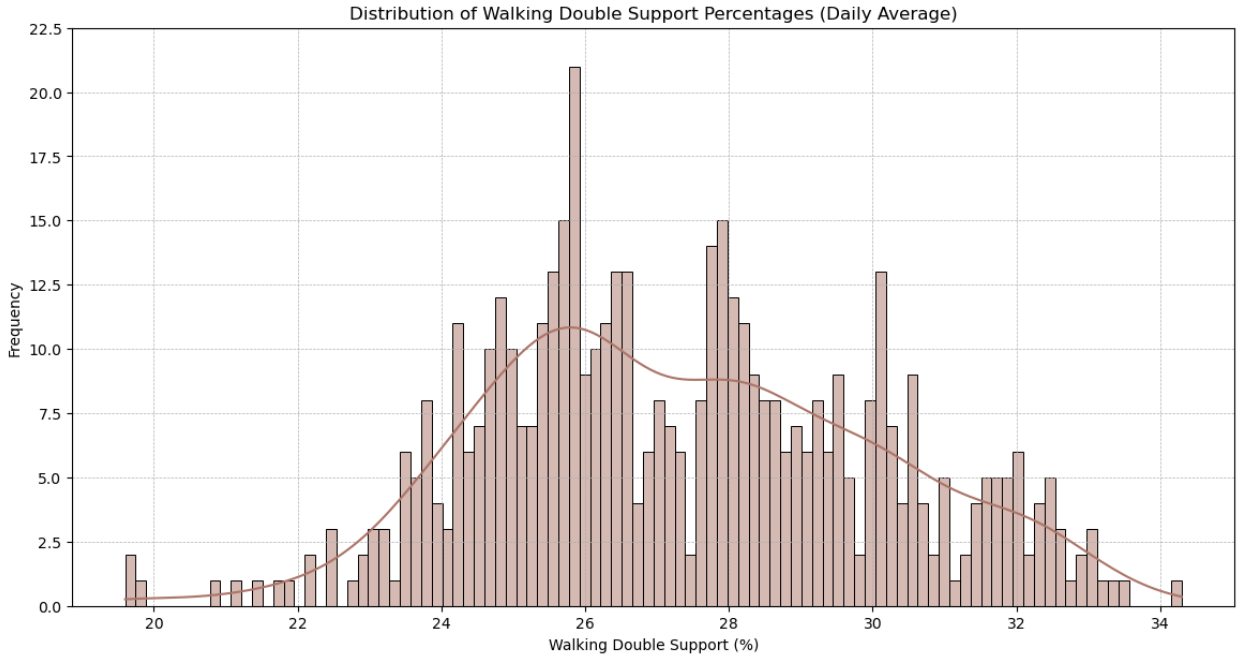
Appendix S: The histogram showing the distribution of walking speeds.



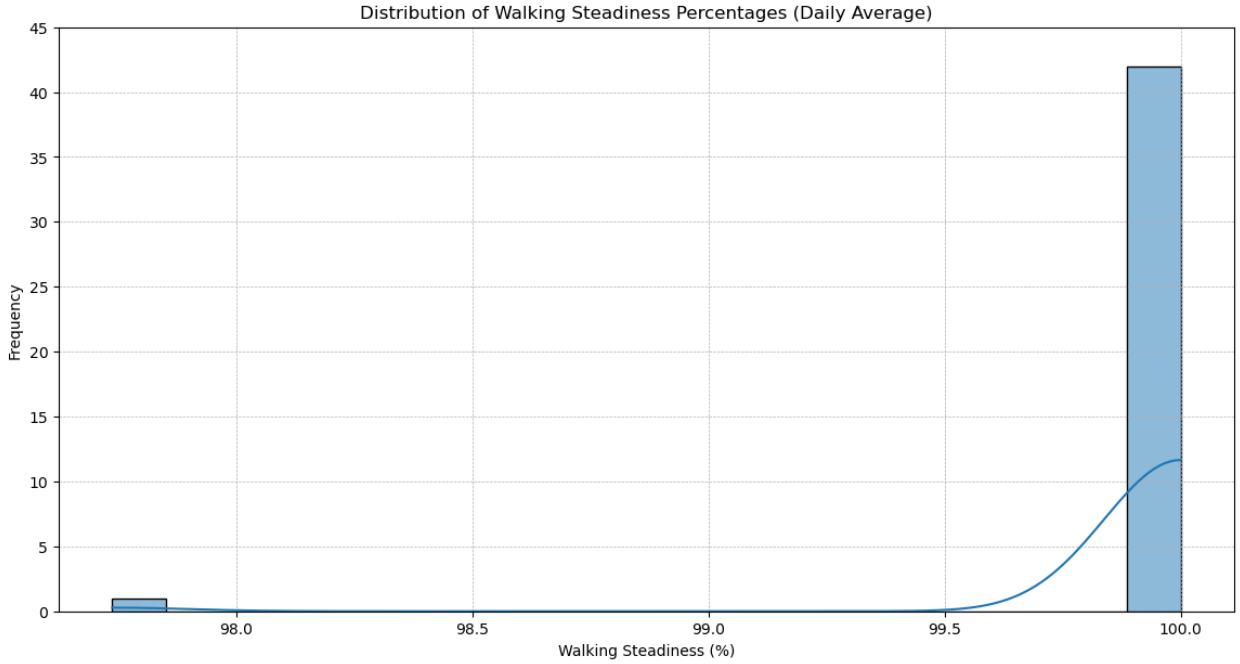
Appendix T: The histogram showing the distribution of walking asymmetry percentages.



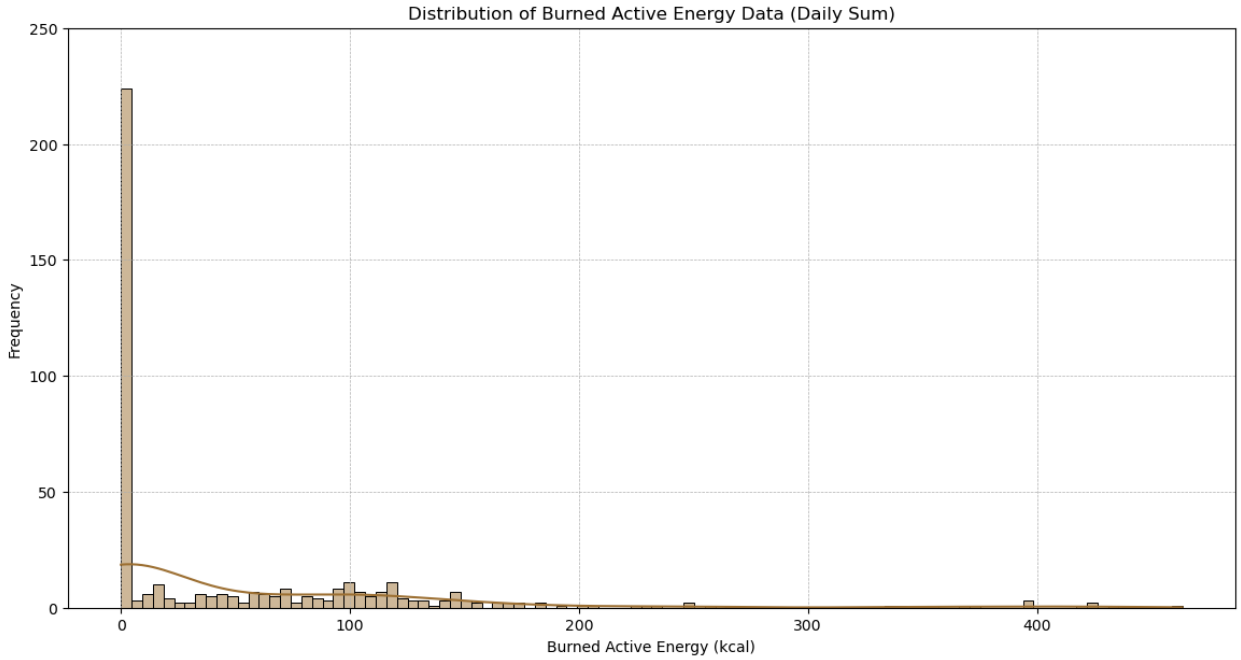
Appendix U: The histogram showing the distribution of walking double support percentages.



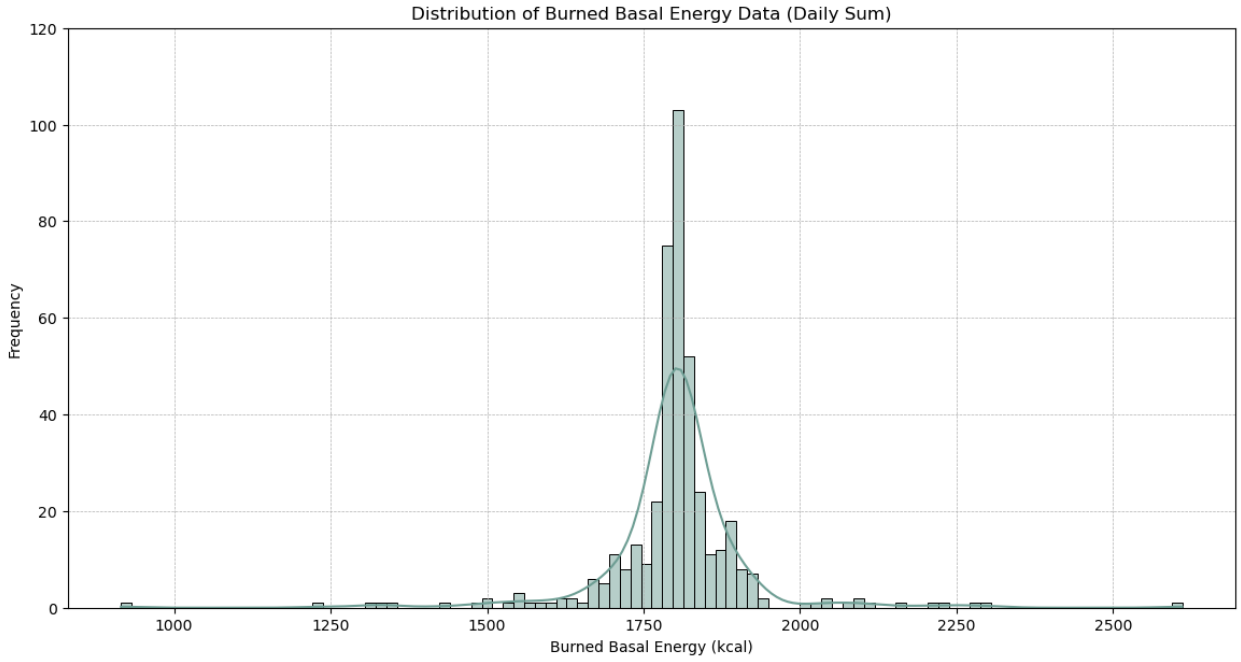
Appendix V: The histogram showing the distribution of walking steadiness percentages.



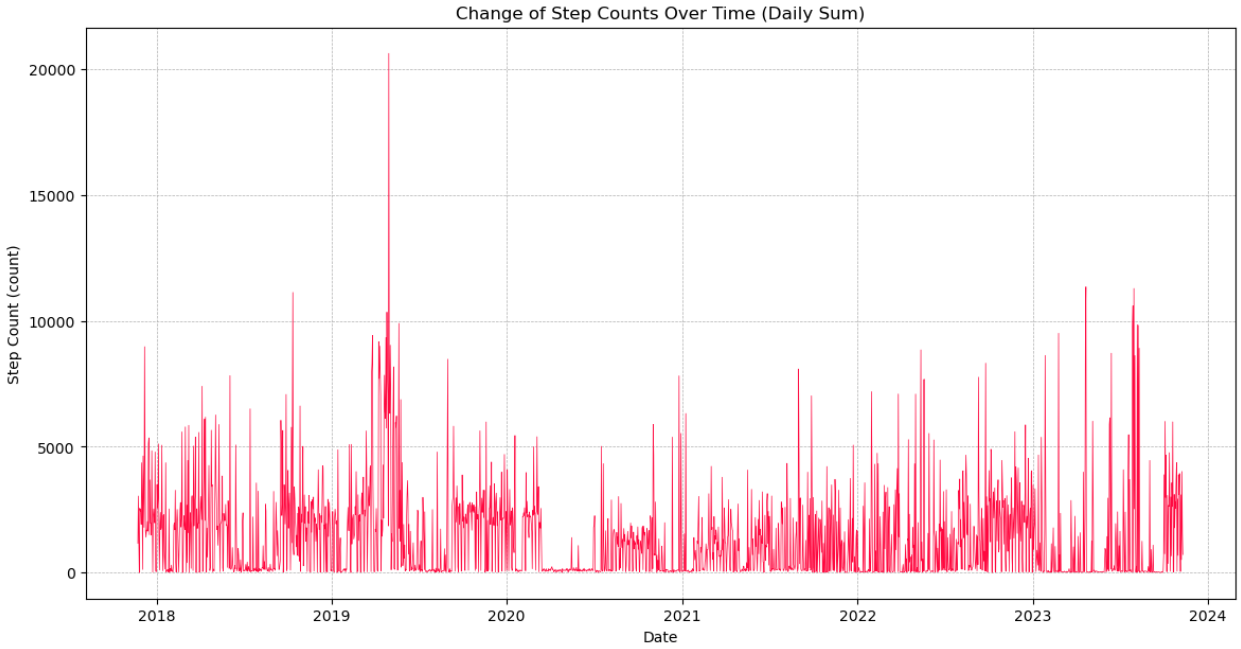
Appendix W: The histogram showing the distribution of the amounts of burned active energy.



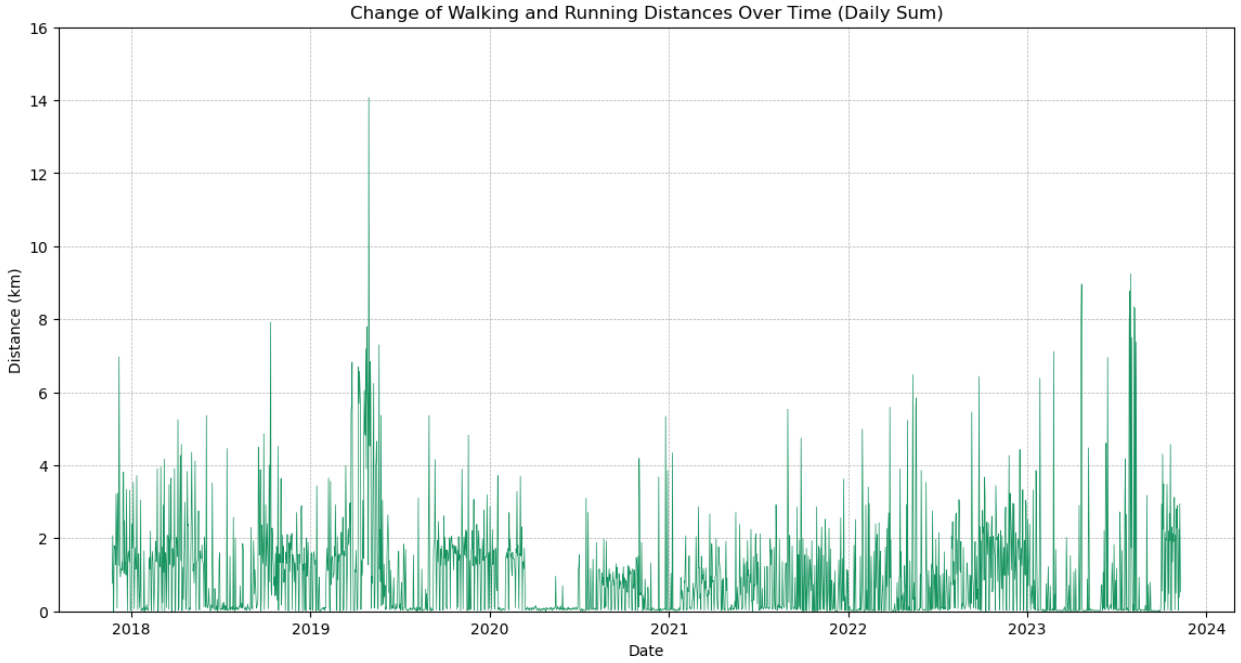
Appendix X: The histogram showing the distribution of the amounts of burned basal energy.



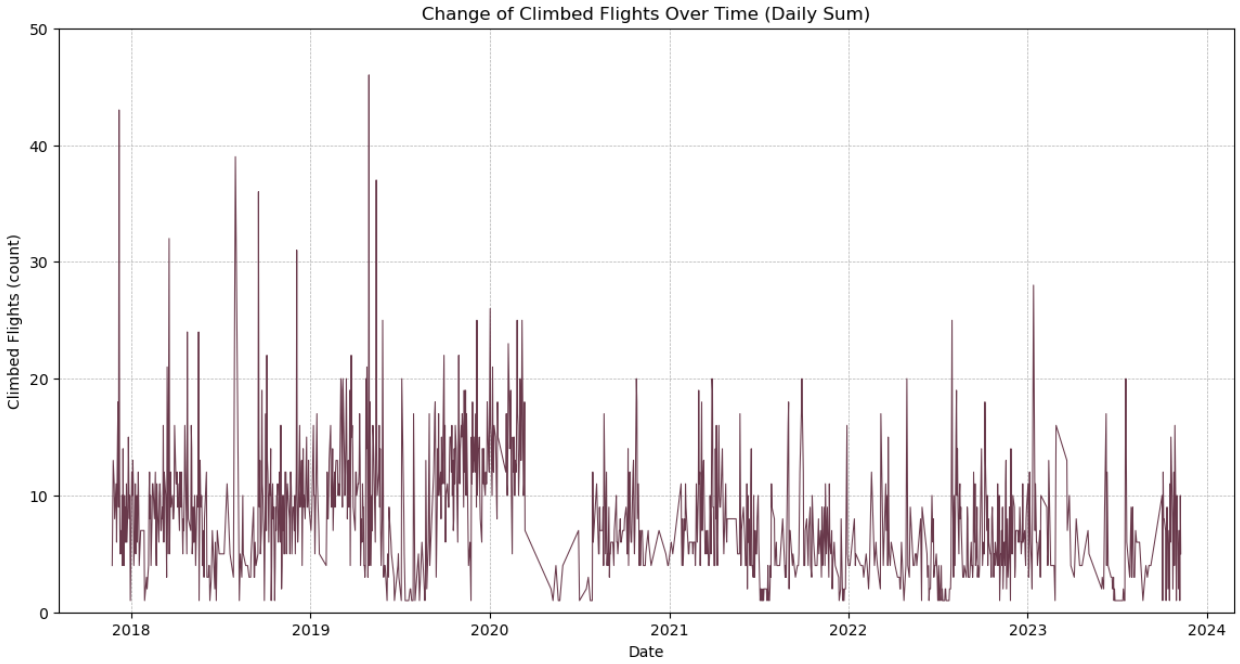
Appendix Y: The line plot showing the change of step count over time.



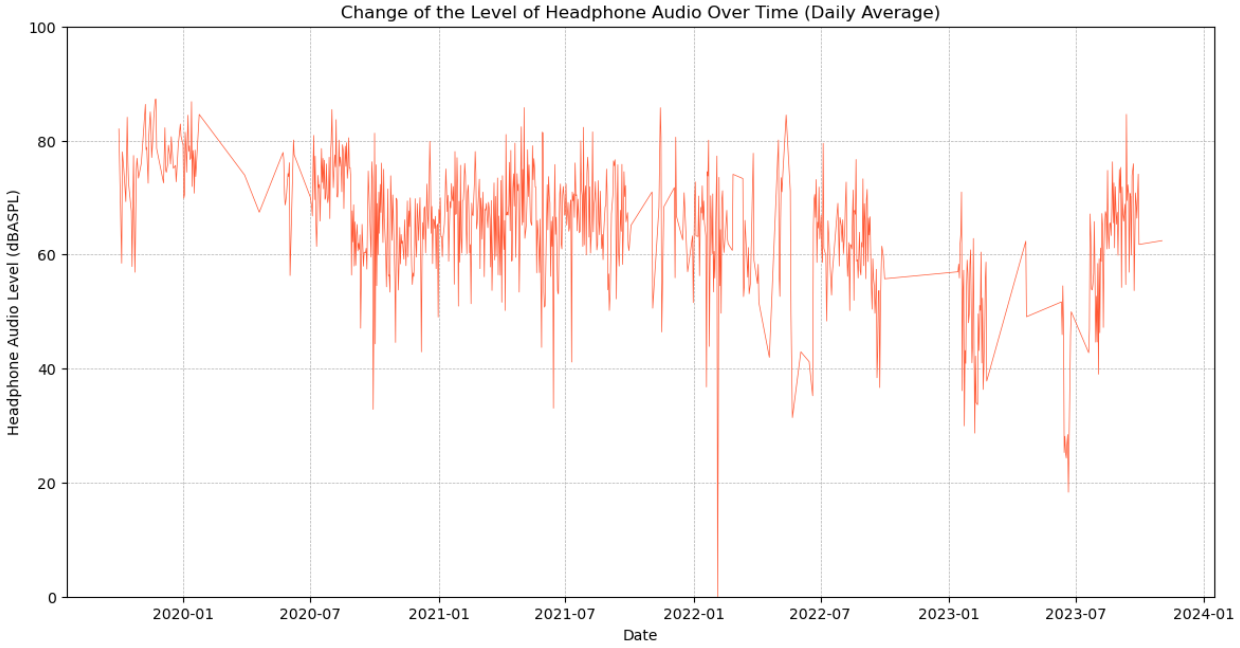
Appendix Z: The line plot showing the change of walking and running distance over time.



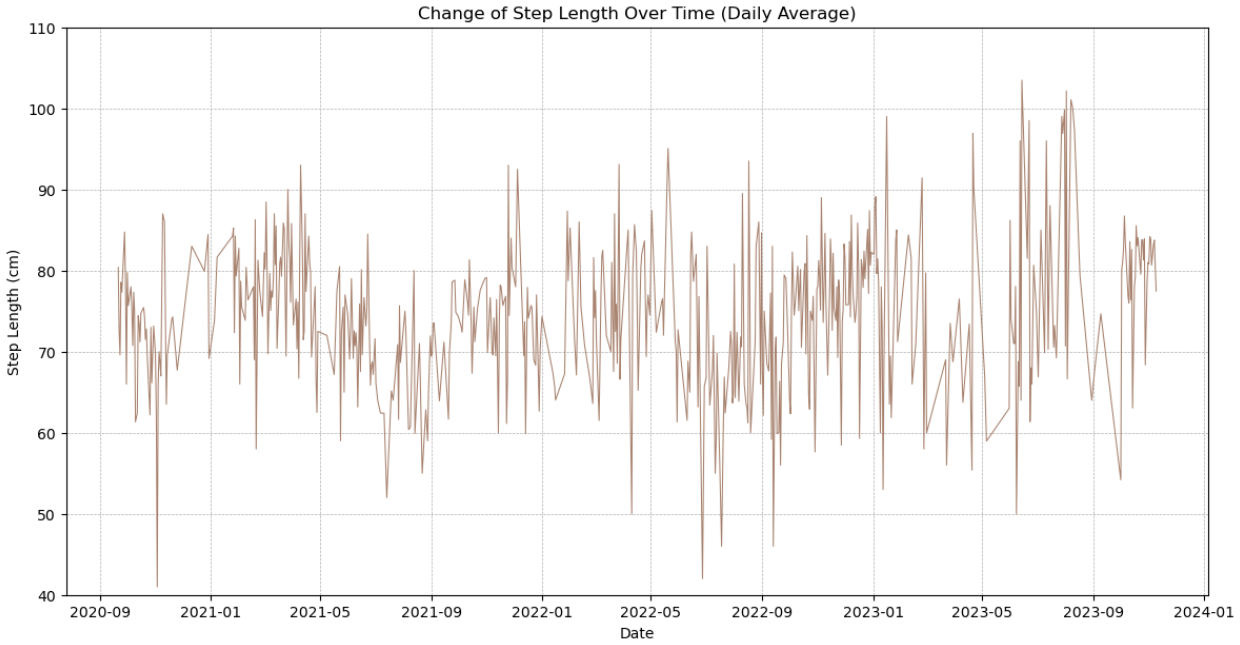
Appendix AA: The line plot showing the change of climbed flights over time.



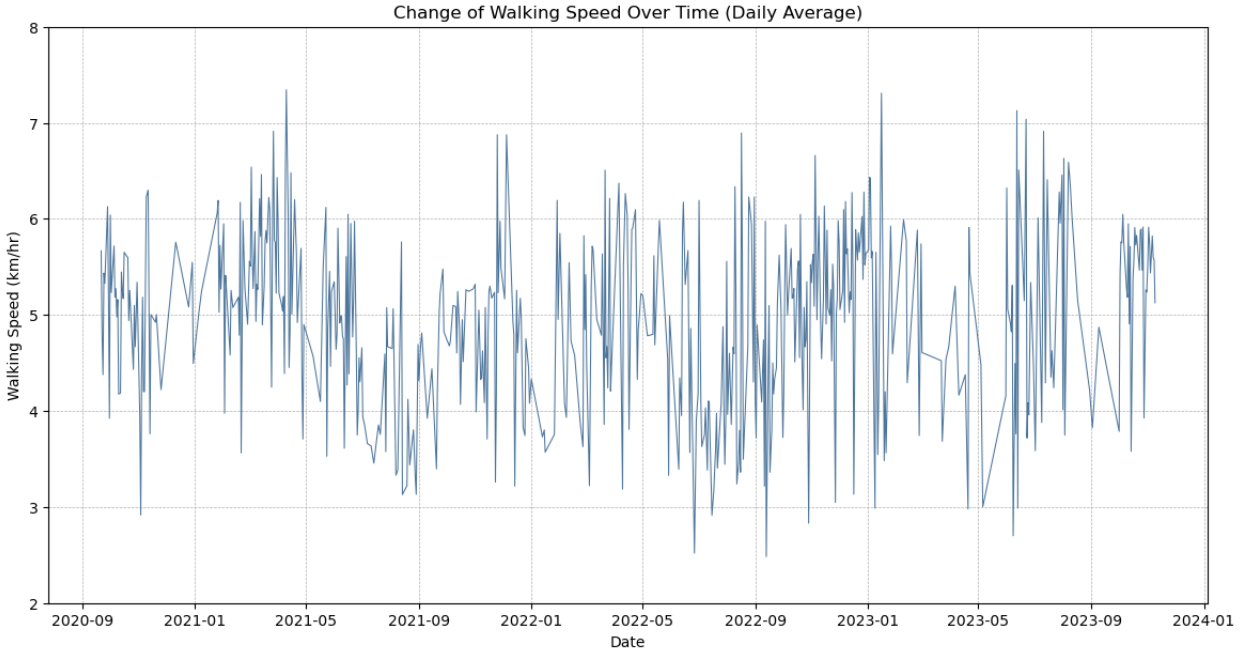
Appendix BB: The line plot showing the change of headphone audio level over time.



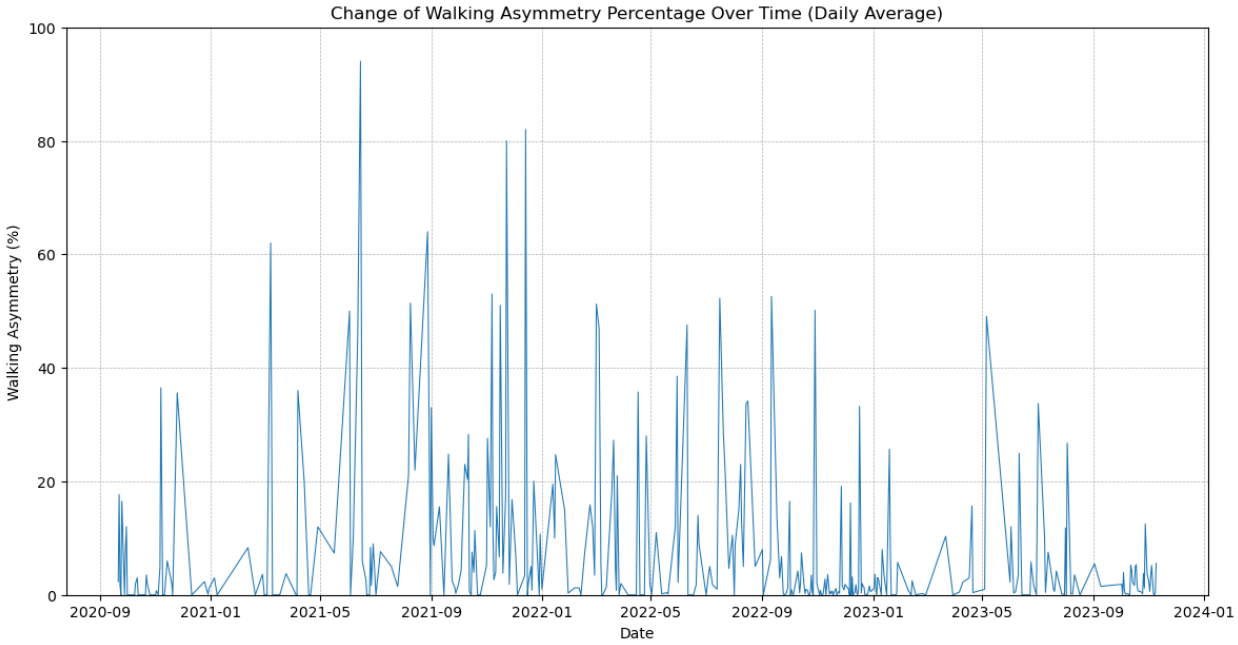
Appendix CC: The line plot showing the change of step length over time.



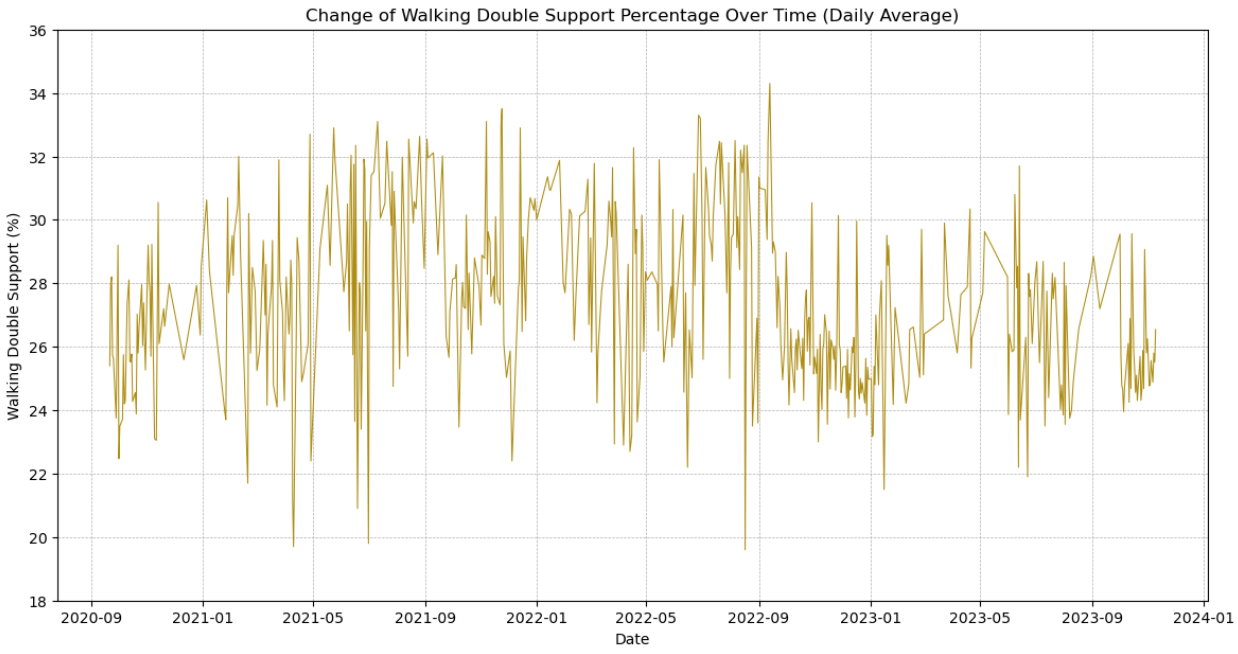
Appendix DD: The line plot showing the change of walking speed over time.



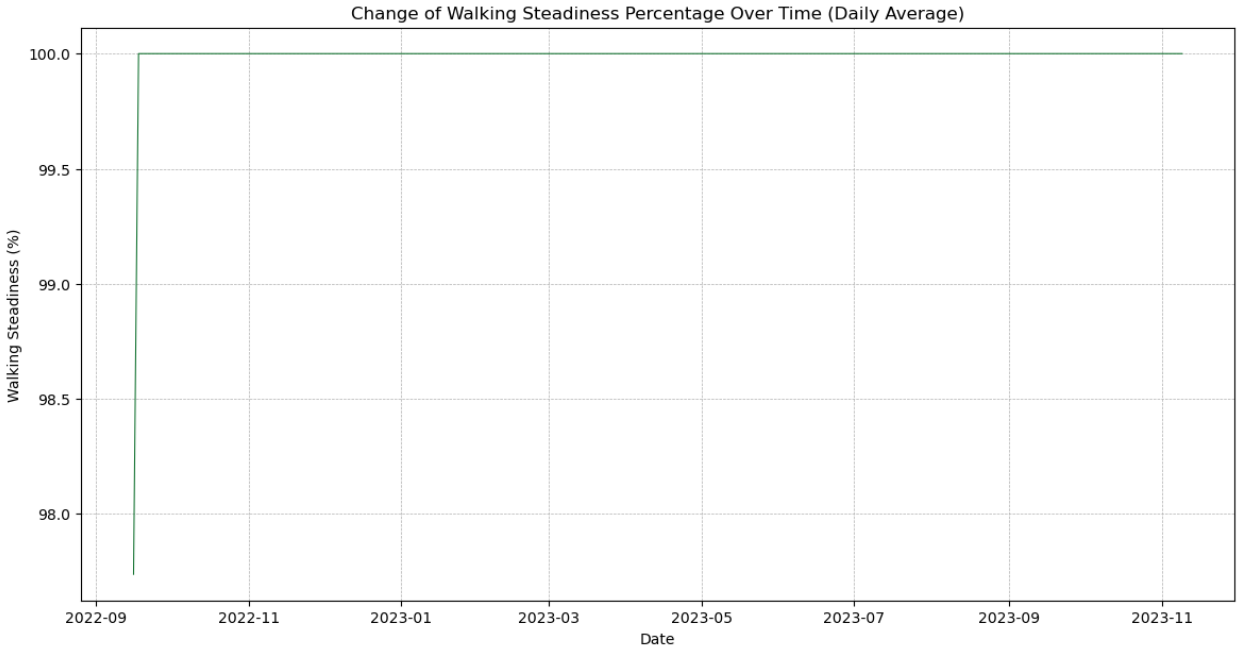
Appendix EE: The line plot showing the change of walking asymmetry percentage over time.



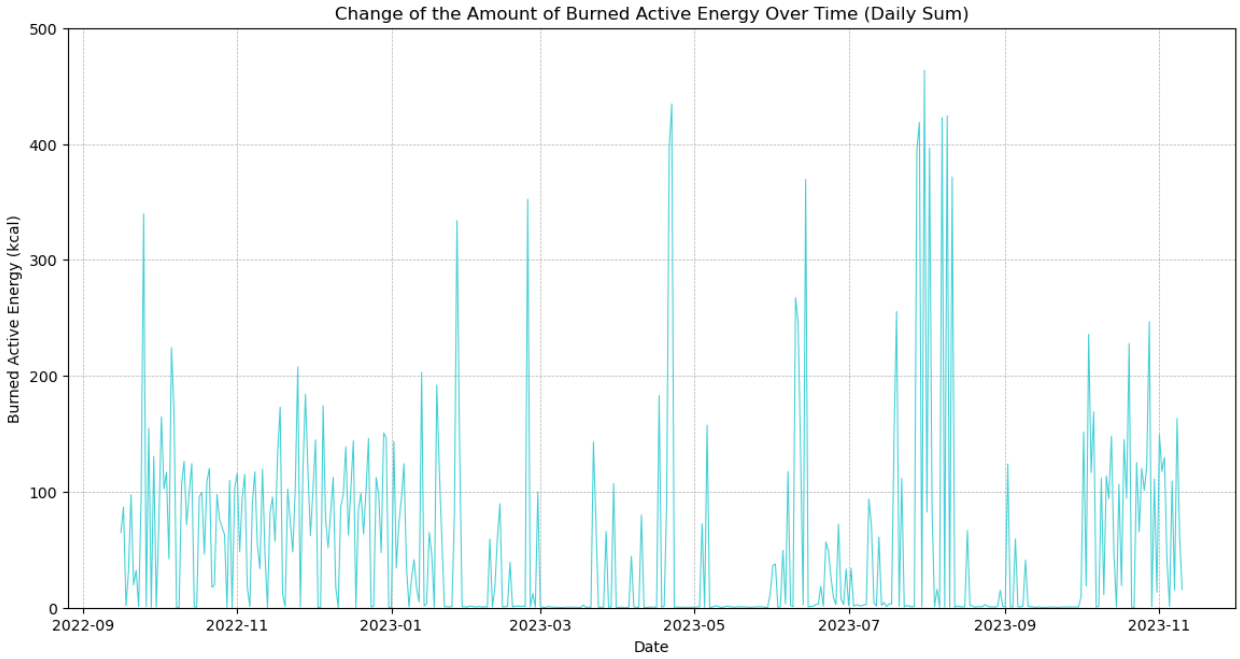
Appendix FF: The line plot showing the change of walking double support percentage over time.



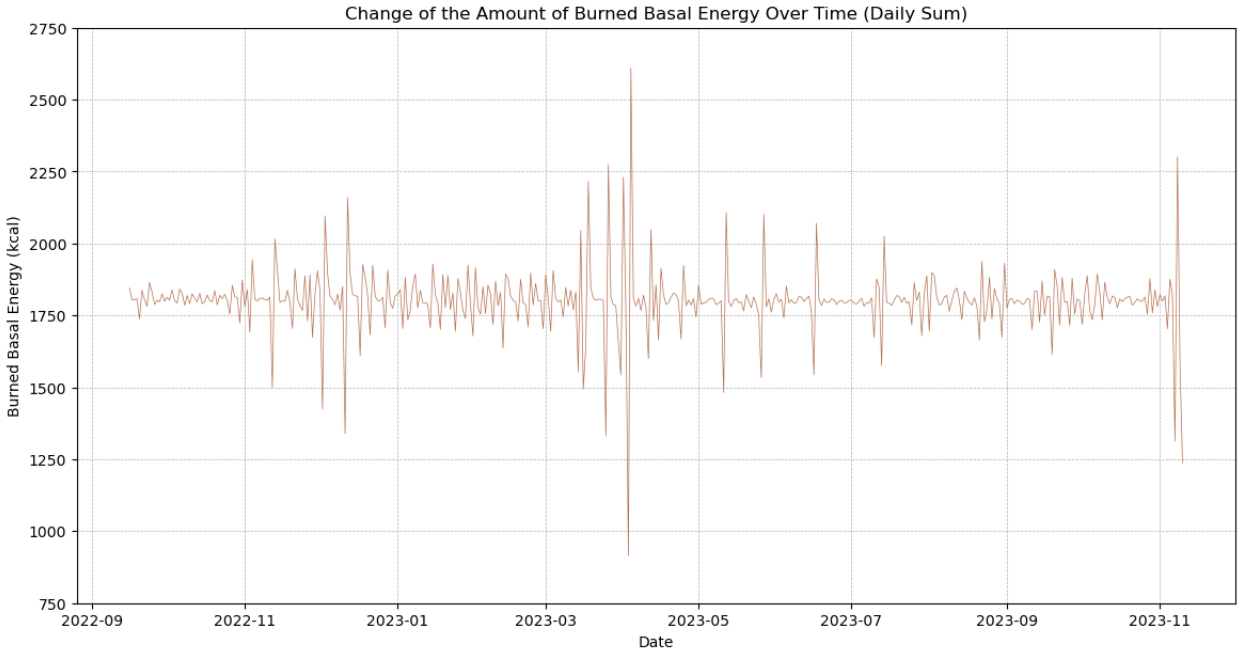
Appendix GG: The line plot showing the change of walking steadiness percentage over time.



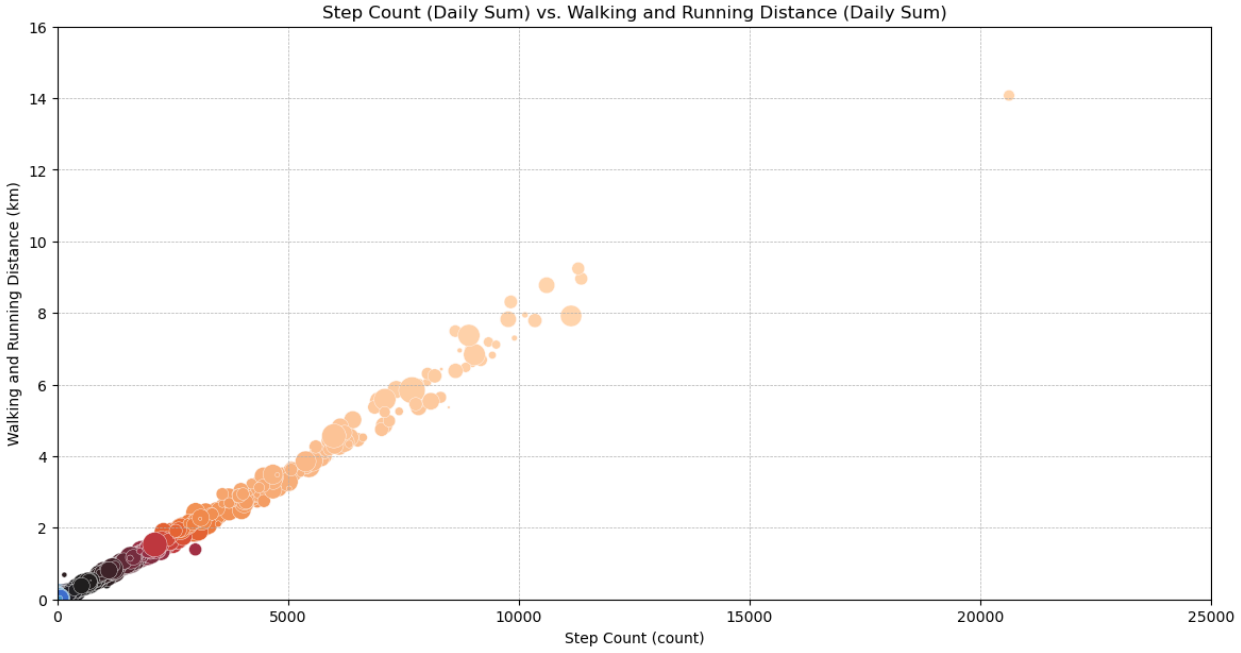
Appendix HH: The line plot showing the change of the amount of burned active energy over time.



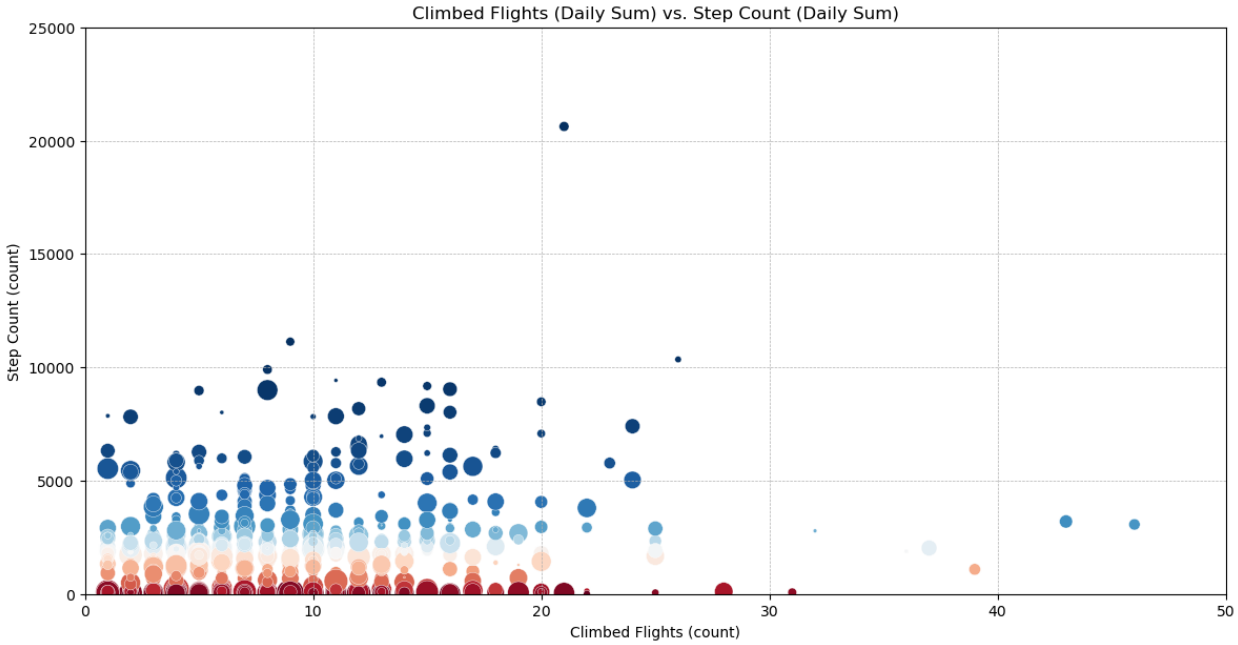
Appendix II: The line plot showing the change of the amount of burned basal energy over time.



Appendix JJ: The scatter plot showing the relationship between step count and distance of walking and running.



Appendix KK: The scatter plot showing the relationship between climbed flights and step count.



Appendix LL: The scatter plot showing the relationship between headphone audio level and step count.