**353 Group Project Report**

**Project overview**

Using the accelerometer sensor of a phone taped to our ankle we investigated what kinds of information we can derive from such data. We did a systematic analysis to determine the best settings for filtering and cleaning the data. We were able to determine the pacing that individuals have. By looking at the forces in a single axis we were able to determine speed and distance as well as how many times a person stopped during their voyage. Finally we tried to investigate if we could predict which exercise an individual is doing using Machine Learning.

**Research Questions and Approach**

Throughout the project our main objective was to answer the general question “can we determine the walking style of a person”. We hypothesized that determining the pacing, speed, and distance would be essential for this. Ultimately, we hoped to determine differences in the walking of two people as well as how much data was needed to come to a conclusion. Lastly we investigated whether we could apply machine learning to determine differences between exercises by giving it a training set.

The question of how to determine the difference in people's walking styles can be explored in many aspects and it's hard to pick any single aspect that would be representative. Pace is interesting because how often someone takes a step can be due to physical limitations and style, the hope is to use a fourier transformation on the data and have the fourier domain tell us something about the pace. Velocity and acceleration are picked because they are one of the most straightforward data to visualize, and it is also expected to have periodic behaviour that can be used to determine the walking frequency. To determine the velocity we used the integral of the acceleration and used the distance as a way to validate our findings. Using the acceleration on all three dimensions together with the frequency they appear,hopefully that would tell us a bit more about how people would walk differently.

**Materials and Methods**

***Tools used***

Data was collected using the [Physics Toolbox Sensor Suite App](https://play.google.com/store/apps/details?id=com.chrystianvieyra.physicstoolboxsuite&hl=en_CA) on phones pertaining to individual group members. Each phone has a different sensor in their circuit board with different physical limitations in both the magnitude of the forces they can measure, the threshold, and how often, the Frequency (see table 1).

|  |  |  |  |
| --- | --- | --- | --- |
| Person | Phone | Frequency (Hz) | Threshold (+/-g) |
| Sam | Samsung Galaxy Note 10 | 400 | 8 |
| Vera | iPhone X | 100 | 6 |
| Kevin | ZenFone 3 Max | 100 | 3 |

Table 1. Information on different phones used and their limitations

***Sampling Method***

Convenience sampling, a non-probability sampling method, was used to collect the data that we needed for the project. This means that samples were taken only from the subjects who were available and willing to participate in the sampling, which were the three group members. Although the samples might not represent the population and might have bias problems, convenience sampling was the best option because it is quick, simple, inexpensive, and the group members were readily available.

***Systematic Reviews***

The data were collected using the *Physics Toolbox Sensor Suite App* on both the Android smartphones and iPhones. The tri-axial accelerometer records the gravitational force (g) in three dimensions along the x, y, and z axes and the total gravitational force of our walking motion (see picture 1). When we stand still, perpendicular to the ground, the g-force values are (x=0, y= 1, z=0), meaning that we are experiencing gravity only in the vertical direction (y-axis), and the x and z axes are always orthogonal. If we started walking, the x, y, and z g-forces as well as the total g-force would increase or decrease whenever we speed up or slow down. Different phones also cap off at different g’s. For example, one group member’s phone caps off at 2g, so his phone is not able to record any g-forces higher than that (See thresholds in Table 1).

The sampling frequency varied between 100HZ and 400HZ on different phones, meaning 100 to 400 samples were obtained per second. The data is exported as a CSV file from the App with four columns, including “time” for the real user time, “gFx” for the g-force along the x-axis, “gFy” for the g-force along the y-axis, “gFy” for the g-force along the z-axis, and the total force (called “TgF” on Android and “gFTotal” on iPhone).

***Collecting Data***

Phones were either taped to the right side of our right ankle or placed on the right side in our right shoe in the same orientation. The orientation of the phone is shown in photo 1. The x-axis is the horizontal g-force in the forward and backward direction. The y-axis is the vertical g-force in the upward and downward direction. The z-axis is perpendicular to the y-axis in the outward and inward direction. Each one of the group members would walk for about ten to twenty minutes on both flat roads and slopes. The walking motion was a continuous series of stride cycles with one or two short stops in between.

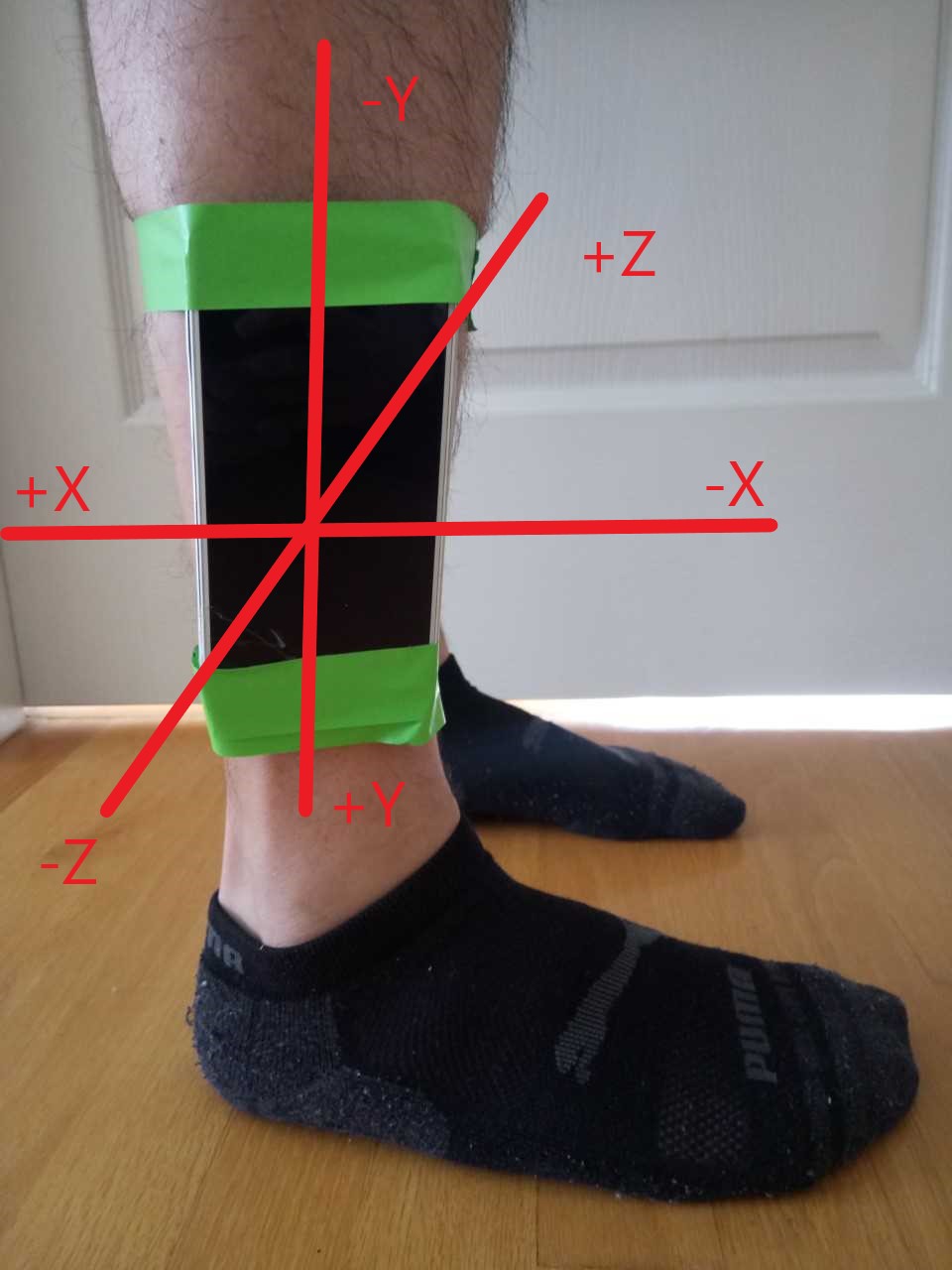


Figure 1. Photo depicting how measurements were made

***Errors***

Although we ensured that the phones were tightly taped to our ankles to minimize noise in the data, there was still an inevitable non-sampling error, which was the orientation of our phones when collecting the data. It was nearly impossible for all of us to place our devices perfectly at the exact same location above our ankles or in our shoes because we did not have access to the appropriate equipment. Therefore, there were some differences in the x, y, and z axes between the group members. One group member’s x-axis might be tilting a little bit upward while the other group member’s was tilting downward. Yet, as we progressed through the data analysis step, we realized that the differences are insignificant and negligible.

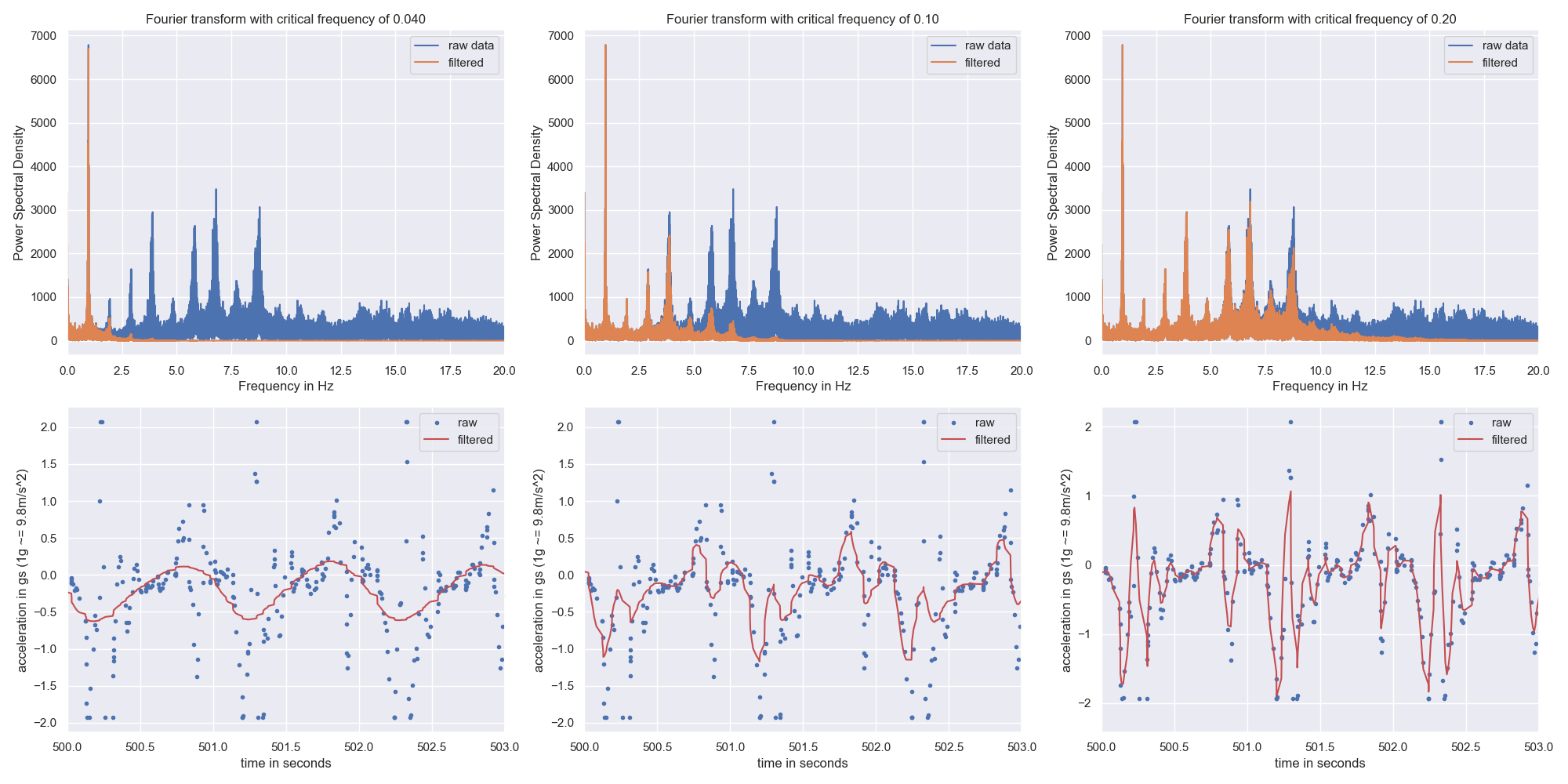
**Assumptions**

We assume that the force of gravity is negligible in the x and z axis. Even though this is not true, the probability of placing the phone perfectly perpendicular to the ground is close to zero. Moreover, the leg bone and how one walks is not going to be at a perpendicular position at all times. This means that there can be a bias due to the force of gravity now having a component in not only the y. To account for acceleration due to force of gravity we would need to know something about the angle at which the phone is relative to the ground at each step, which we do not have. Additionally, the fourier domain can have frequencies greater than 10Hz. It is reasonable to assume that no one has a walking frequency in the 50 Hz rate, therefore we assume frequencies above 10Hz to be noise.

**Cleaning the data**

Choosing the techniques to clean up the data we got and extract the truth from the raw observation is tricky and very challenging. Due to the different measuring devices and different sampling frequency. Data collected was not in a good shape, with different base noise, threshold and sampling frequency. To make our data representative and more meaningful we decide to filter the raw data with a low-pass filter first to get rid of the base noise. We found that because the data were sampled from different devices, we would need essentially apply different filtering parameters to reach the same filtering effect.

The low-pass filter is chosen because the walking frequency is happening on the scale of 1-2hz, which is relatively lower than most of our noise that we collected, see data transformed into frequency domain and analysis, figure 1. From the Fourier domain, it seems like the most significant frequency of the signal is within 5hz, however after such frequency a train of continuous pulses can be seen from the frequency > 10hz. This train of pulses is very apparent in the x axis of the g force data.

Figure 1. Comparison of Fourier Transformed Data to Raw Data, bottom 3 plots of scatter plots showing effects of filtering

Looking at the raw data we suspect that this might have to do with the fact that the majority of the data we collected were very close to the capacity of our device. This can be seen in the raw data plot where with some the upper peak of the pulses looks like to be cutted out. In the other axis where this behaviour is not observed the fourier domain seems rather clean, with the majority of signal is within 5 hz. Thus we believe that there is less truth in the higher frequency signal than the lower frequency however, this is not to say that there is nothing valuable there. The butterworth addresses the need quite clearly, it allows us to adjust critical frequency where we can keep all the signal features below that frequency and for higher frequency we can adjust the weight they contribute to the filter result.

By plotting the filtered data next to the raw data with different filtering parameters we can determine the best critical frequency (Figure 1). Here we see that as our critical frequency is reduced from .1 to .04 we get a more symmetrical sinusoidal curve due to over-filtering, resulting in unrealistic data. Additionally, there is a loss in the spectral density at ~1Hz so too much data is being cut out. In contrast, when too stringent of a filter is used, 0.2, a connect the dots effect is obtained which means we are not reducing the noise enough. With this information we concluded that a critical frequency of .1 would be the best.

**Data analysis**

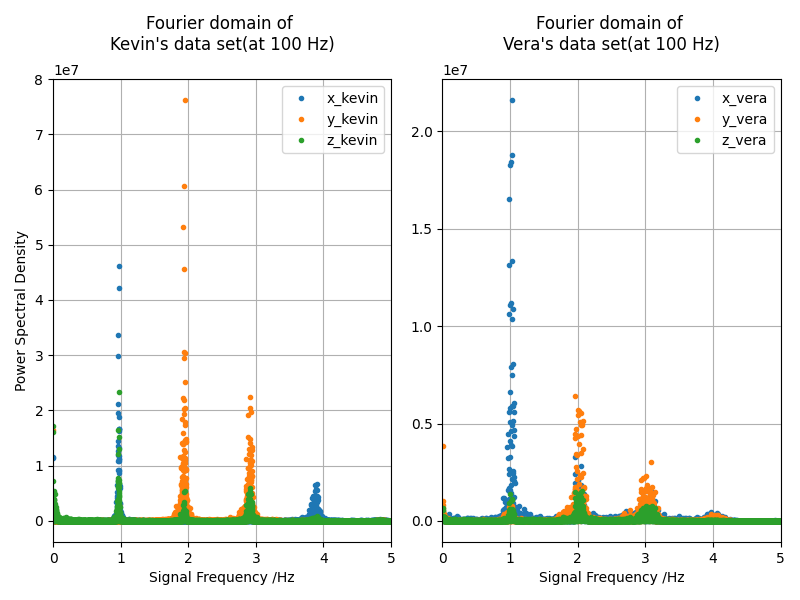
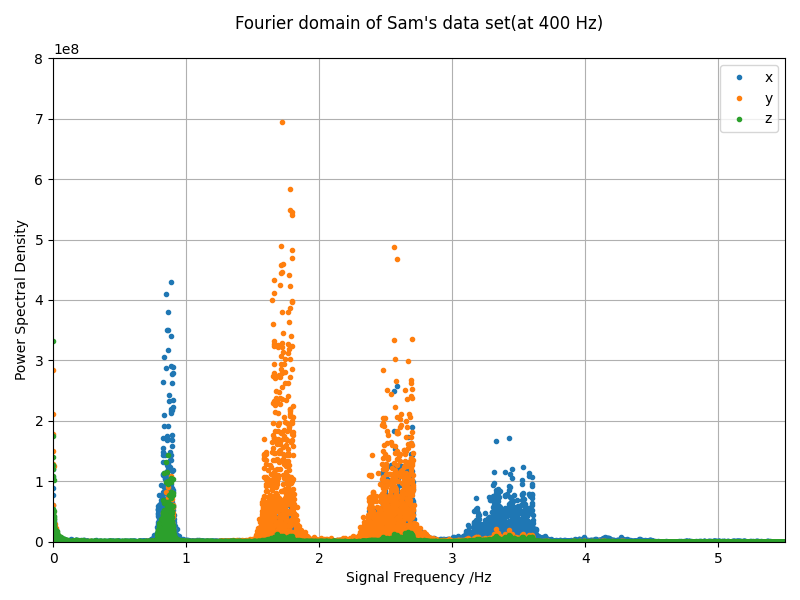
The analysis of the “walking styles” starts with some investigation with signal analysis happening in the Fourier domain. We are interested in the signal components in the x, y and z axis, as well as the frequency which the signal repeats itself. 

Figure 2. Comparison of Fourier Domain of the three axes

Figure 2 shows a plot of the Power Spectral Density (PSD) and the signal frequency of all three axes. As shown in the figure, the PSD for x-axis(blue) and z-axis(green) peaks around 1Hz, and the PSD for y-axis(orange) peaks around 2Hz. This pattern of the majority frequency of y-axis is about double than the majority frequency in x-axis and z-axis can be observed in all three dataset, shown in Figure 2.

This effect may be due to the fact that the data is only recorded from a signal leg, so the movement recorded on the x-axis and z-axis are relatively unaffected by the movement of the other leg whereas the in y-axis is more sensitive. From the same figures, it is also obvious that both x and y axis have recorded significantly higher PSD than the z axis. This might suggest the changes on the z-axis are perhaps insignificant, likely due to it being pointed outward from the ankle, where not much motion can be captured. Also, it is important to take note that on the y axis the constant gravitational pull of positive 1 g is not removed from the filtering, but since gravity is constant it would contribute to the 0Hz, therefore it is automatically ignored in the Fourier domain.

To further explore the aspect of how walking frequency differs from each group member, a semi-log plot of comparison on the most significant frequency for all three axis is shown below.

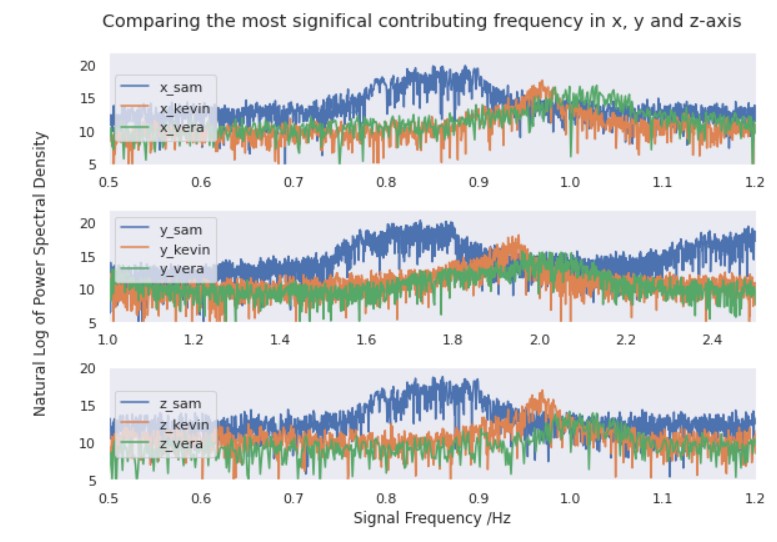


Figure 3. Zooming in and comparing the most significant frequency of all three data sets.

As shown in Figure 3, all three axes seem to agree on the fact that Sam’s data set seems to have a lesser frequency than the other 2, which suggests the walking frequency might be somewhat slower. There seems to be less difference between Kevin’s and Vera’s data but if taken a closer look in the y and z-axis Kevin's data seems to be reaching its peak before Vera's data does.

The information we obtained from the fourier domain suggested that although all three axes may have very different values, the frequency of most of the steps can be obtained by comparing the signal composition in the fourier domain. The three axis seem to agree a lot on the most significant frequency, with the exception that y would actually be 2 times x or z axis due to it also taking account of the movement from the other leg.

In conclusion, having different walking frequency is one of the sufficient conditions to say that the walking styles are different, but it is not a necessary condition to make such a claim. For example, Kevins’s data and Vera’s data are so close together they might be seen as the same frequency or inconclusive. To further determined how to determined different working style between the group members we would to need more information on other aspects of walking style

**Deriving Distance and Velocity**

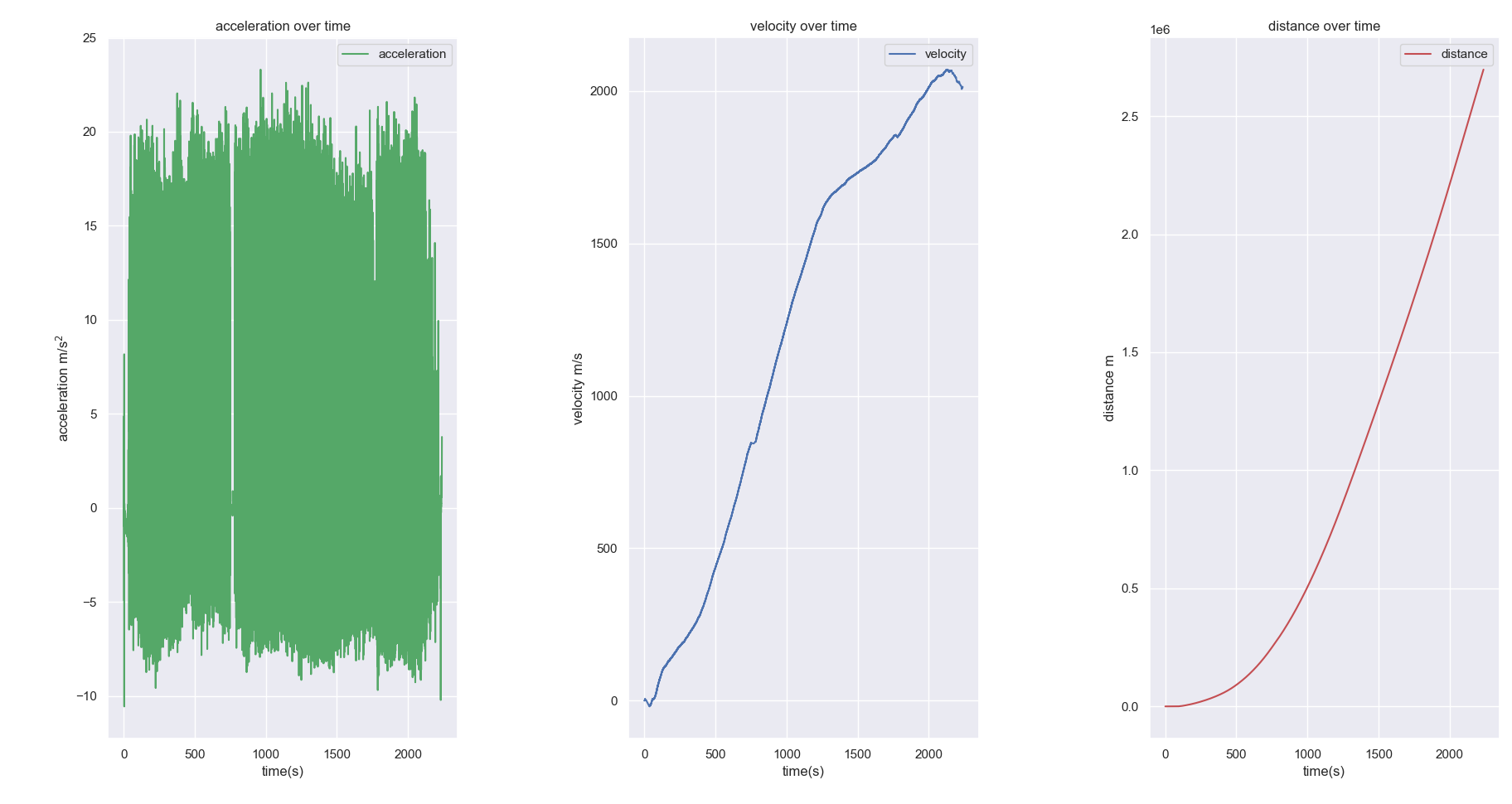
To determine velocity and distance, we assumed that the major component of moving forward is due to forces in the x direction and so the problem was initially investigated by only looking at data collected in the x-axis. After filtering the data with a butterworth we have g forces so by multiplying this series with the gravitational constant, g= 9.81 m/s2, acceleration is obtained. From here we naively attempted to integrate the acceleration over time for our entire dataset to obtain velocity over time. We then integrated the velocity to get distance. This proved to be disastrous as we can see in figure 4 below.

Figure 4. Initial attempt of integrating to get velocity and distance from forces in the x direction.

The results above are clearly incorrect as velocity should remain more or less constant at around 1.4m/s. Additionally, traveling 1e6 m in 2000 seconds is ridiculous so clearly there was a flaw in this plan. We suspect that there may be some gravitational forces due to uneven terrain so we decided to perform the integration on a tighter time frame of 10 seconds (figure 5). This still proved to be problematic as well however it gives us insight into what is causing the phenomena. The amplitude between steps varies greatly between each step, likely due to measuring error.

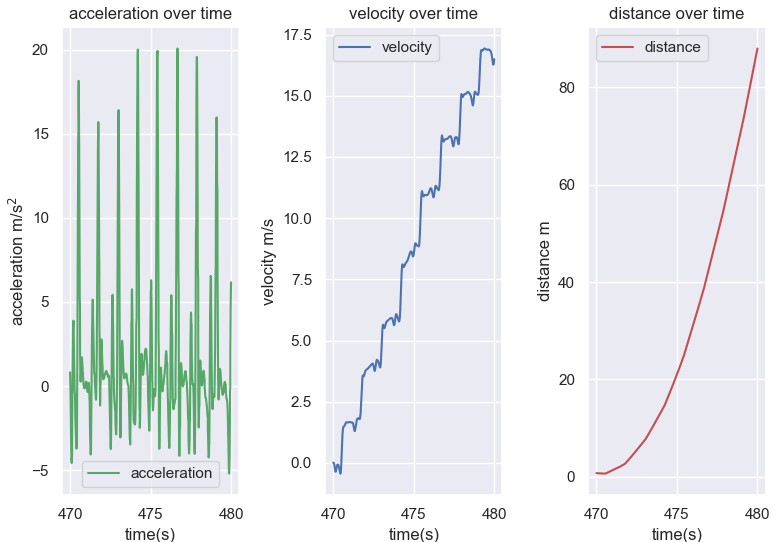


Figure 5. Looking at what is happening in the x-axis over a 10 second period of time with first method

To circumvent this issue a rolling mean was calculated to normalize the acceleration relative to the window we are integrating. Simply normalizing by the mean acceleration at given intervals helped when looking in a small window, figure 6, but proved problematic on the larger scale, figure 7. The rationale behind this was that a step should be more or less symmetrical however this is in itself problematic as the forces exerted are not symmetrical. Overall you should have a greater magnitude in the pulses of acceleration in the forward direction than the backward, if this was not the case you would be walking on the same spot. Notice that although this solution is incorrect the velocity is getting more constant. It was hypothesized that whenever the individual stops and the acceleration is 0, we are accidentally giving velocity by using this normalization method. This is an interesting find in that we can now easily visualize when and how many times a person stops during their travels by counting the hills, which appears to be 5 in this set, figure 7. Although this method is an improvement it is still incorrect and in order to get a good result we attempted to integrate on rolling windows of 1 second.

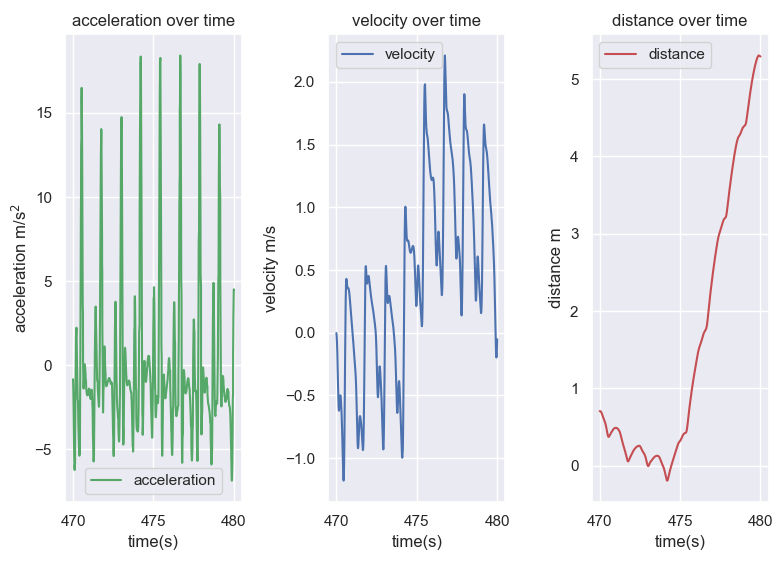
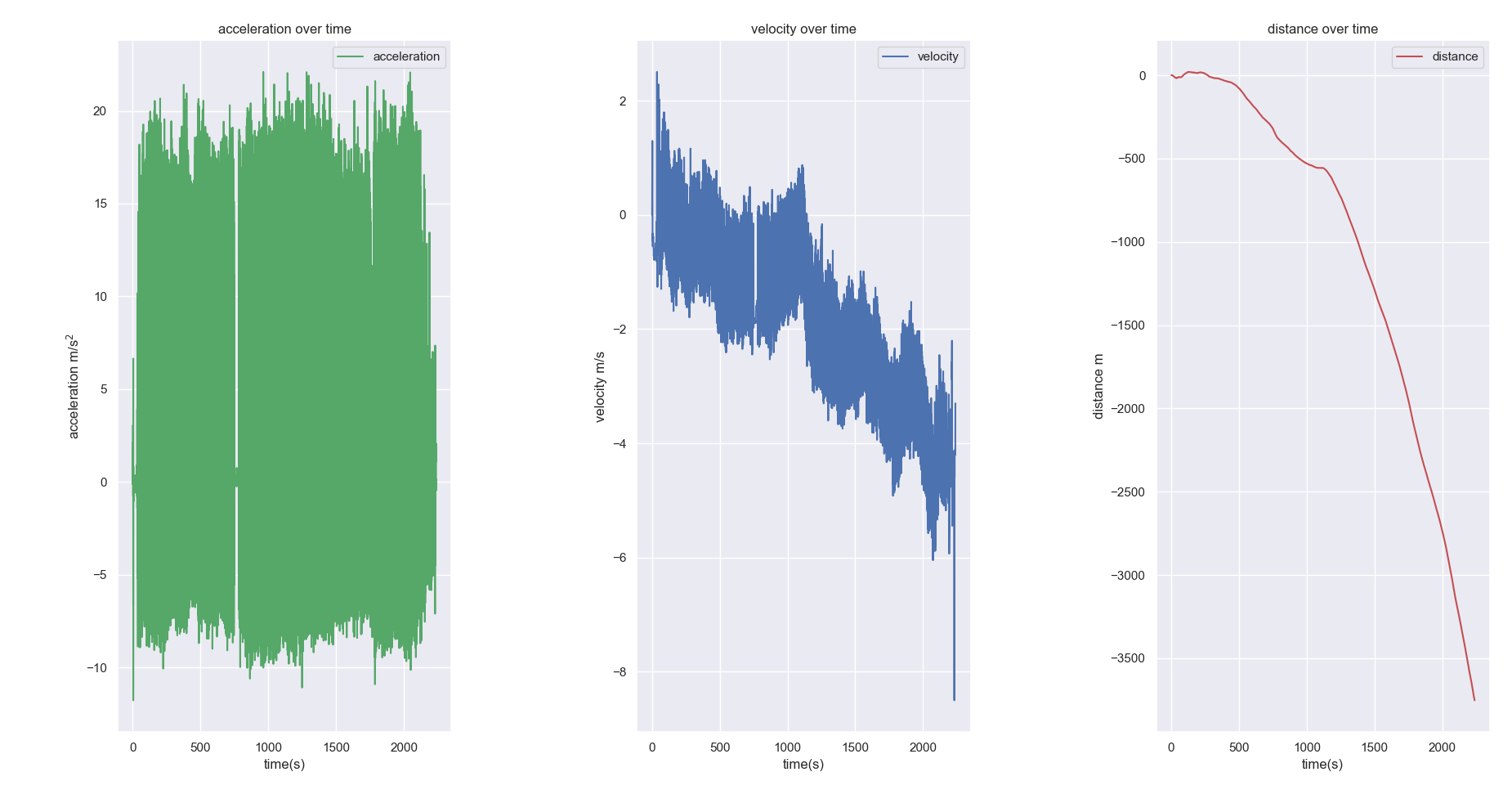
Figure 6. velocity and distance by normalizing by mean acceleration within a 10 second window

Figure 7. Overall velocity and distance by normalizing by mean acceleration

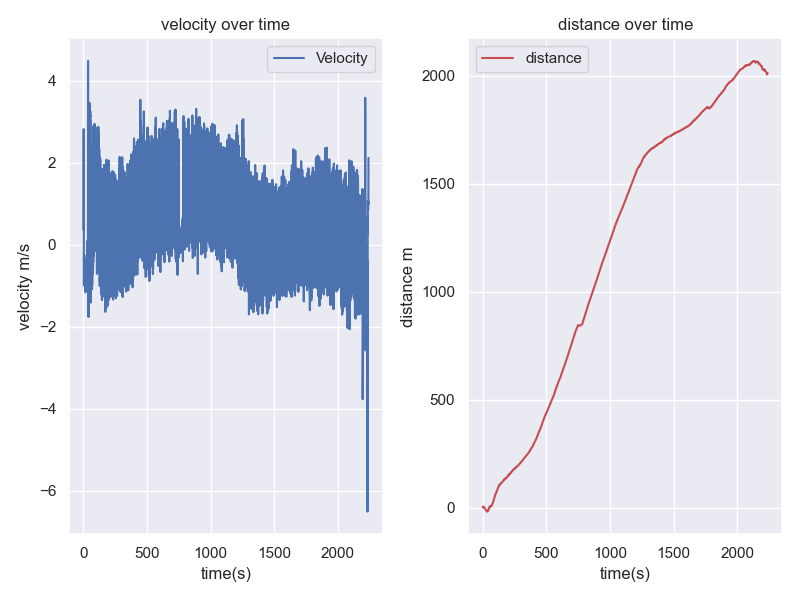
Finally, we applied integration with a rolling window of one seconds since between one second it is likely that you have only taken maybe 1 or 2 steps at most. This proved to be difficult to program because the x values would be forced to be equally spaced by a dx of 1. This meant that the integral we obtained is overestimating by giving each row the value of 1 second when in reality it is a fraction of a second. We eventually determined a way to deal with this; since we knew our window was the size of our frequency, which amounts to 1 second we could divide the velocity we obtained by our section size, which was the frequency since the frequency is how many rows occur in 1 second. This proved to work extremely well and gave the most accurate results, figure 8. By looking at the change in concavity in the distance graph we can determine the individual stopped 5 times. This same result can be determined by looking at the velocity for stretches where velocity remains at zero but is harder to see in figure 8.

Figure 8. Final method for Determining Velocity and Distance over time given by looking at x axis

Out of curiosity we applied the velocity and distance calculating program to the other two axises. Unsurprisingly the y axis proved to have issues since it has a constant acceleration downwards due to gravity, figure 9. This can be corrected by subtracting 1 from the original y values at the beginning however it over estimates the distance traveled, figure 10. Surprisingly, the forces in the z direction were rather informative and got a similar distance when calculating the distance to those calculated in the x, see figure 11. Unlike the z the velocity here is much less which suggests that perhaps the phone was slightly tilted on the leg towards the front of the body. Another explanation, assuming perfect positioning, could be that each step the individual moves their leg inward slightly. Now that we have better insight into speed we can compare and identify individuals based on average velocities.

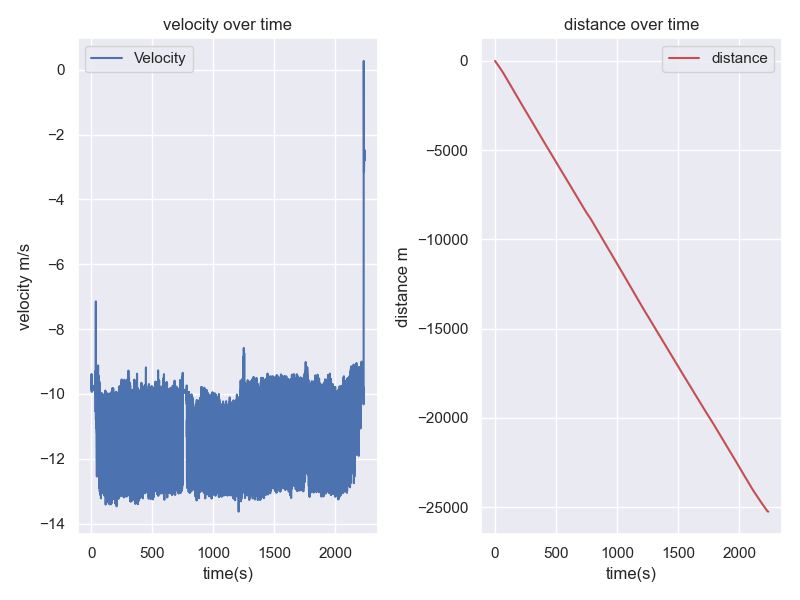
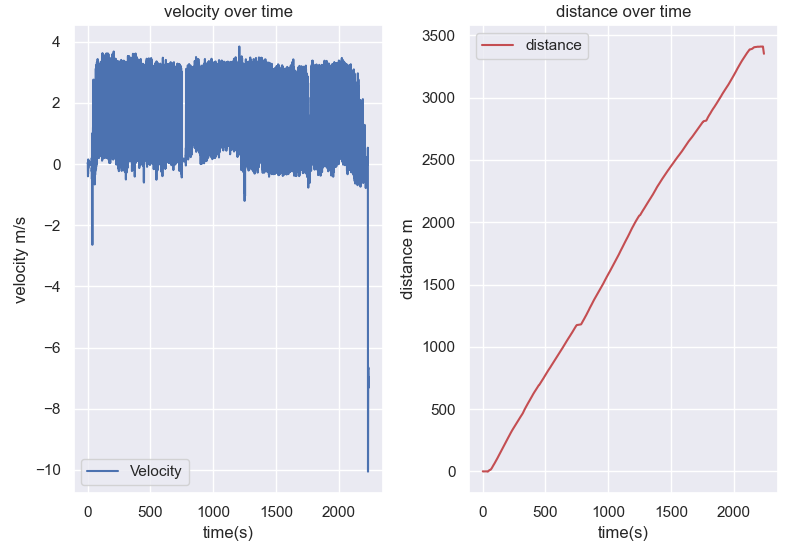
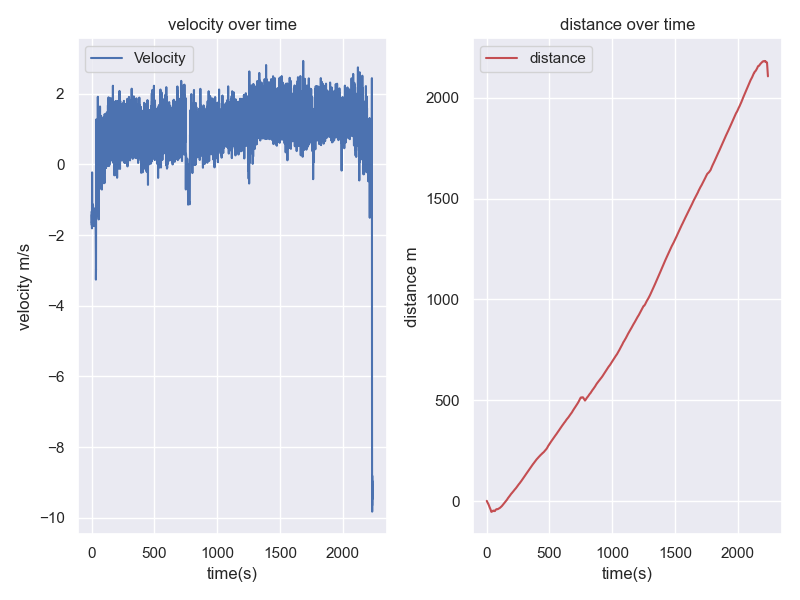
Figure 9. Velocity and Distance in the Y axis Figure 10. Velocity and Distance in the Y axis post correction

Figure 11. Velocity and Distance in the z

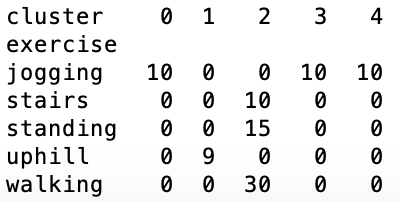
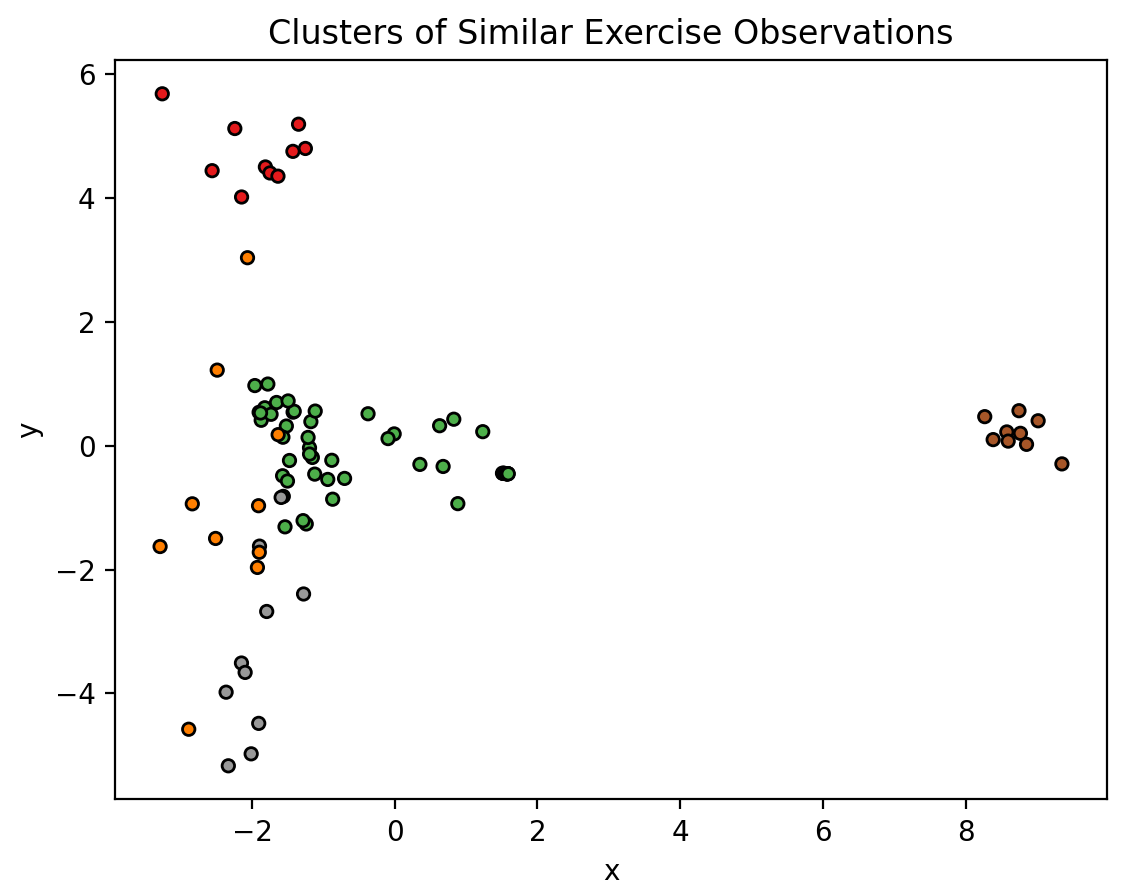
**Machine Learning**

Beside using the Fourier analysis to compare our walking pattern, can we predict individual walking style using the three vector g-forces from the accelerometer? Due to the differences in device specifications, it would be unreasonable to make predictions between the group members because it would be too obvious to tell which data set came from which phone just by looking at the sample frequency or where the data caps off . Therefore, we approached the question in another way: can we create a machine learning model to predict the type of exercise a person is doing? This way, we ensured the person could record various exercises using the same device and predictions for that person were made based on just their own data.

We created a labelled dataset that contains five exercises: walking, jogging, walking uphill, walking up and down the stairs, and standing still. Each exercise has several features: two seconds of g-force in x-axis (in g), two seconds of g-force in y-axis (in g), and two seconds of g-force in z-axis (in g). That makes up to 600 features. The noise in the forces was filtered using the Butterworth lowpass filter. We also have an unlabelled dataset with the same features, but the exercise name redacted. The truth data shows the actual exercises of the unlabelled observations. The labelled data was then used to train and validate the machine learning model to predict the exercises that the unlabelled data contains.

The model that we used is Random Forest because it gives us more options for the parameters to fit the data. We also used PCA to reduce the dimensions from 600 features to 10 most important features. The training score and validation score are 0.986928 and 0.980769 respectively, which shows that the model will predict the never-before-seen values very well and it is neither overfitting nor underfitting.

To visualize how our machine learning model was able to take the unknown exercise observations and come up with the correct exercises, we used KMeans to see how the exercises were clustered based on similar features (Figure 12). First, we transformed the 600 features into the same scale using MinMaxScalar, then we fit a PCA model to reduce the dimensions to 2 so we could plot the clusters. We noticed that walking, stairs, and standing are clustered into the same groups. This could mean that the motions for these exercises are similar (Figure 13).

Figure 12. Clusters based on KMeans = 5 Figure 13. Clusters based on KMeans = 5

**Concluding Remarks**

One of the problems that we have with our data is the g-forces measured in all three axes do not transform directly into the axes we are used to, i.e. with negative direction of y axes is parallel to gravitational pull. It was also found that if we wish to eliminate all noise and transform the data into the standard axis we would need a gyroscope. This would allow us to determine the rotation of the phone at any given time and translate the phone axis into the axis that we conventional would use and get more accurate value for speed and distance traveled.

We decide on butterworth low-pass filtering based on the Fourier domain of the major pulses are in the low frequency domain. However just like there is noise high frequency, it is completely possible that there is some low frequency noise that is in the data but did not get filtered out because they are low frequency. For example, the constant 1 g of gravitational pull down in the y axis, which is a 0hz constant that affects all g-force int the y axis and requires to be dealt with separately. If there is some other low frequency noise in other axes we would not be able to get rid of them from the filtering.

A lot of the noise and inaccuracy came from the fact that most of the data are measured in different devices, which have different specifications. Using the same device would decrease the measurement error introduced since there is no difference in threshold or other unnoticed bias that arises from using different measuring devices. Additionally there is a lack of control in the environment where the data was collected, different locations have different slopes, terrain, and obstacles. This could cause the g force which is normally in the y axis to affect the other axises.

From the inferential statistical standpoint, the data is far from making any groundbreaking conclusions or generalizations. Although we try to control as many variables as we can, a lack of participants and a lack of data is probably one of the reasons that the data is not representative enough. In order to answer the research question, we would need more participants and standardise the measuring standard to reach a conclusion that could be generalized and tell us more about how people’s walking styles. It would have been nice to use machine learning to categorize people’s walking styles but data collection was limited due to covid and it is hard to mimic a different walking style for prolonged periods of time. The ideal dataset would come from volunteers with medical conditions that affect their walking. Nonetheless, we feel we have made good progress to an interesting research project.

**Project Experience summary**

Kevin

* Developed a python script which returns seaborn plots of accurately estimated Velocity and Distance that was calculated using a rolling window that integrated a pandas DataFrame containing acceleration data measured by a phone.

Sam

* Visualize and analyze walking data in the Fourier domain, structured the limitations and methodology section of the report, make use of background knowledge on statistics, signal filtering.

Vera

* Implemented machine learning model using scikit-learn in Python to predict accelerometer data, manipulated and cleaned datasets with Numpy, Pandas, and noise filtering, and created data visualization with Fourier Transform and Matlibplot.