# Analysis of Kinematics Motion Data

Austin Chen, Ambar Mishra, Koby Lieu, Su-Ting Tan, Sai Sriya Mudigonda

## Introduction

Kinematics is a branch of mechanics focused on the motion of objects without referencing the forces which cause the motion. The dataset includes a single file with 88588 sensor data samples collected from accelerometer and gyroscope from iPhone 5c in 10 seconds interval and ~5.4/second frequency.

The attributes of this data [1] are described at the end of this section. Among the columns, the dataset contains "date", "time" and "username" columns which provide information about the exact date, time and user which collected these measurements.

Furthermore, the gyroscope accounts for the rotational orientation of the user’s iPhone. The accelerometer accounts for the linear change relative to the frame of reference of the device; it is used to measure the force of acceleration caused by movement.

The motivation behind analyzing this data is to develop patterns in the acceleration and motion in space during the different activities and investigate

We used several methods and approaches to analyzing the data. These include basic visualizations and aggregation in the exploratory data analysis section as well as time series analysis, logistic regression models, and artificial neural networks used in the classification section.



The image above shows what the x, y, and z coordinates mean in our variables.

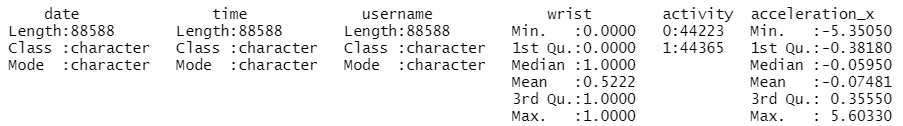
The following are the attributes that are included in our dataset:

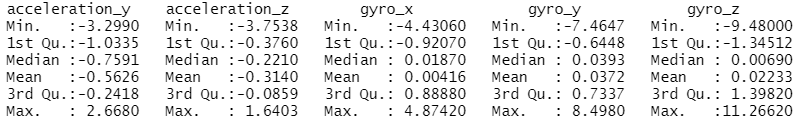
* acceleration\_x: “Acceleration in X direction” (moving leg left or right from current position)
* acceleration\_y: “Acceleration in Y direction” (moving in forward or backwards directions)
* acceleration\_z: “Acceleration in Z direction” (lifting leg up or down)
* gyro\_x: “Gyroscope reading in X direction”
* gyro\_y: “Gyroscope reading in Y direction”
* gyro\_z: “Gyroscope reading in Z direction”
* date
* time
* username: Name of contributor (which in our dataset has one value: “Viktor”).
* wrist: "0" denotes wearing the device on the left wrist; "1" denotes wearing the device on the right wrist.
* activity: "0" denotes walking; "1" denotes running.

We have no missing values in our dataset and data cleaning isn't necessary.

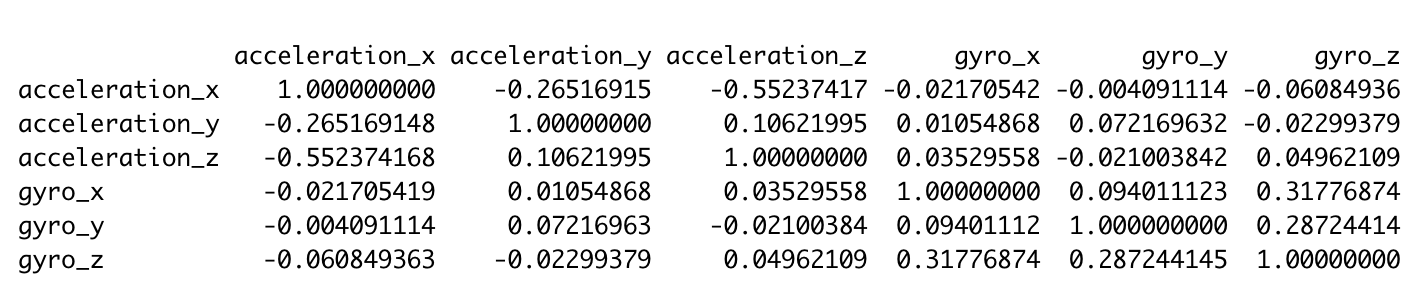
## 

## Exploratory Data Analysis



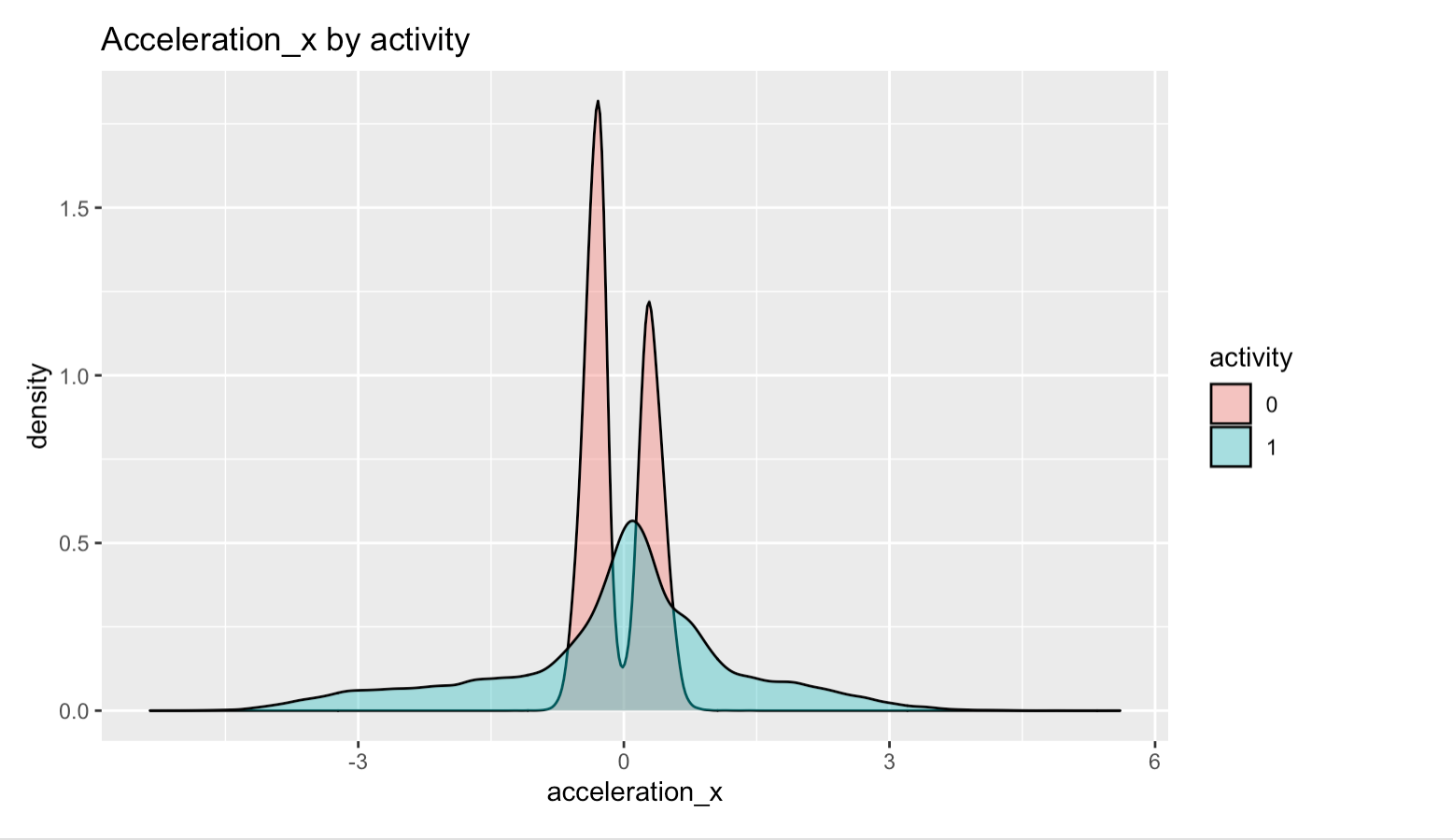


The above summary statistics do not provide any particularly interesting findings in aggregating the data across the multiple days of activity, so we will specify particular categories to investigate those individual patterns. First, we decided to see if there were any obvious correlations between the accelerometer and gyroscope variables:

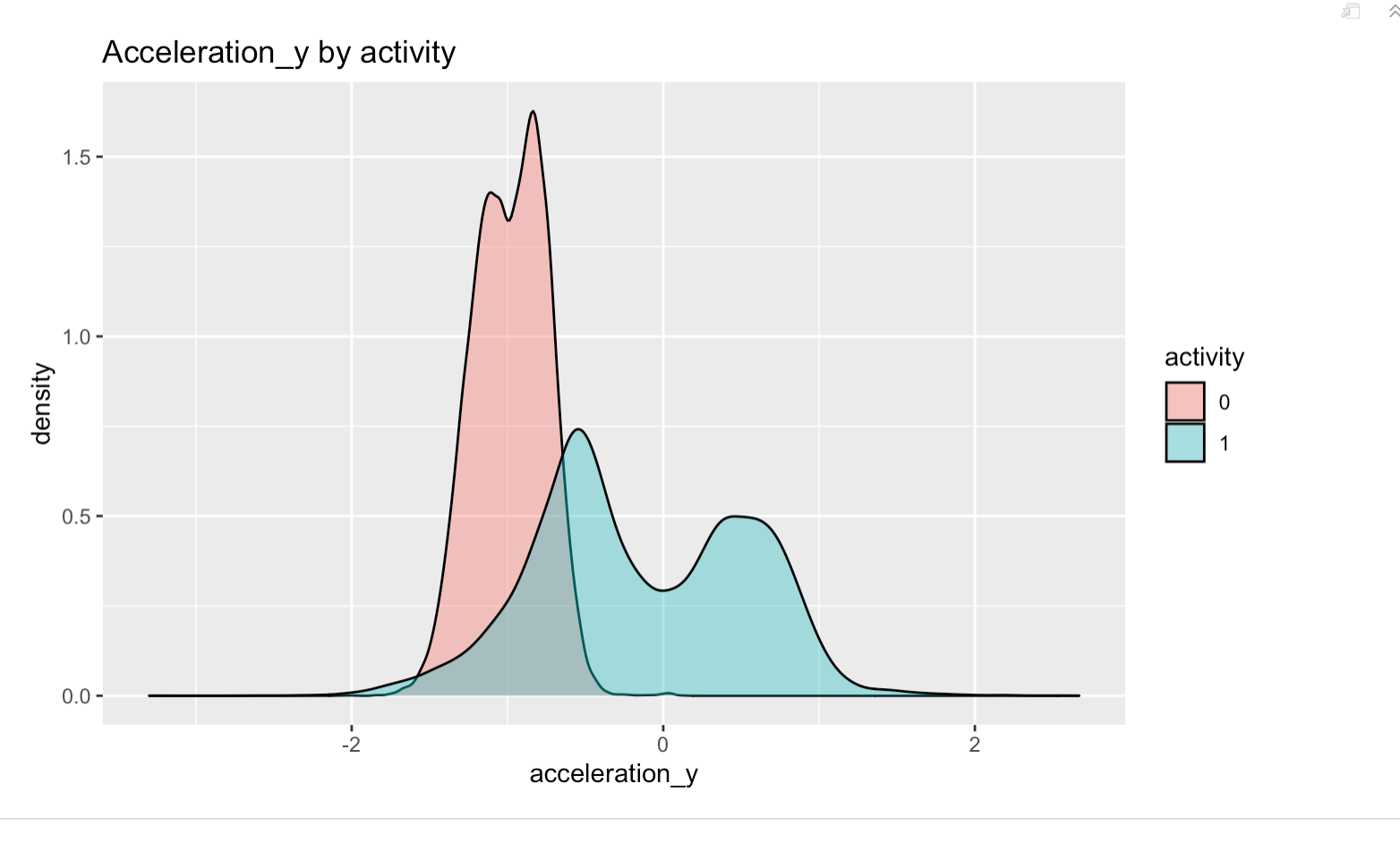


The strongest correlation by magnitude was the correlation between `acceleration\_x` and `acceleration\_z` with a -0.55237 correlation coefficient.

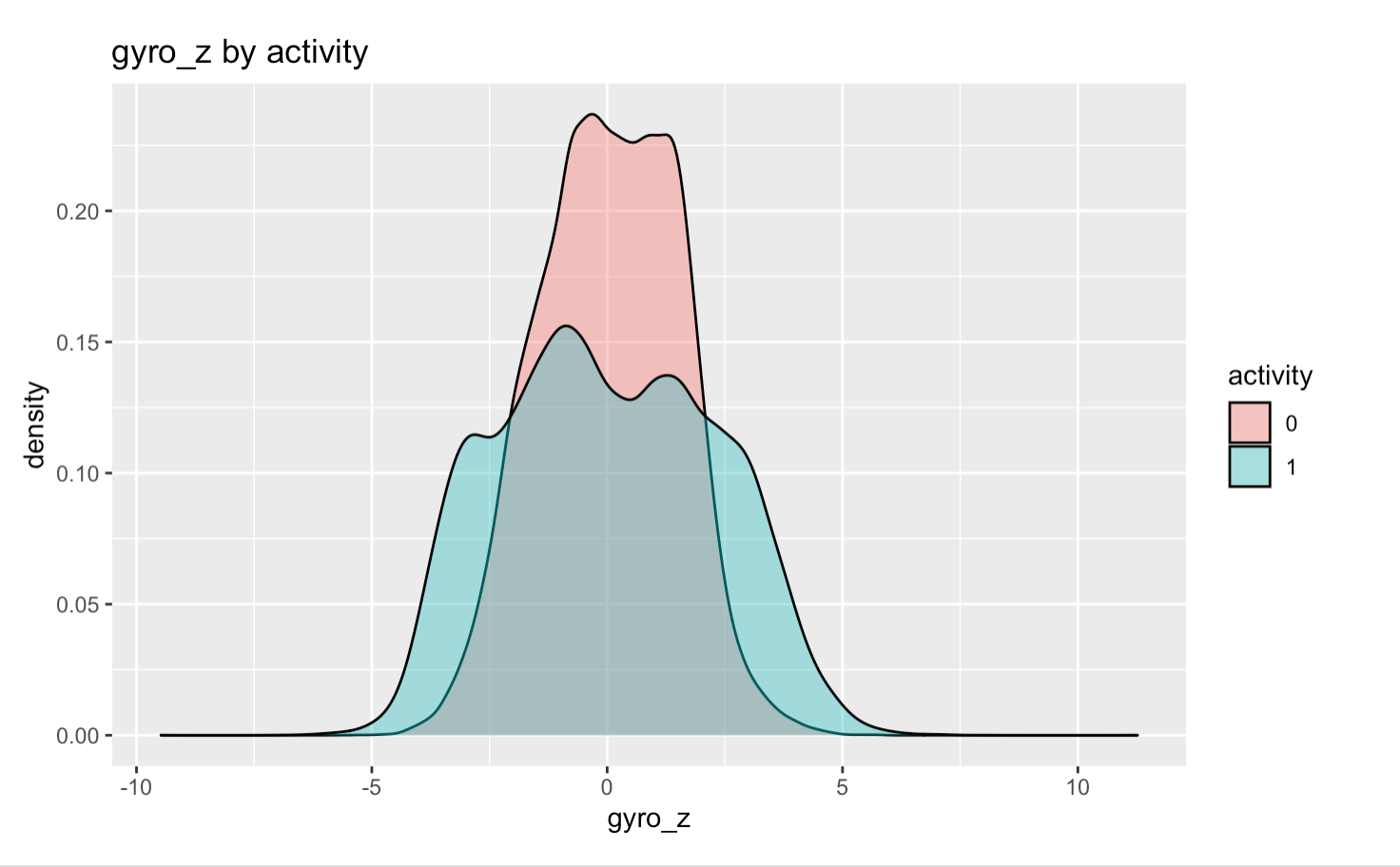
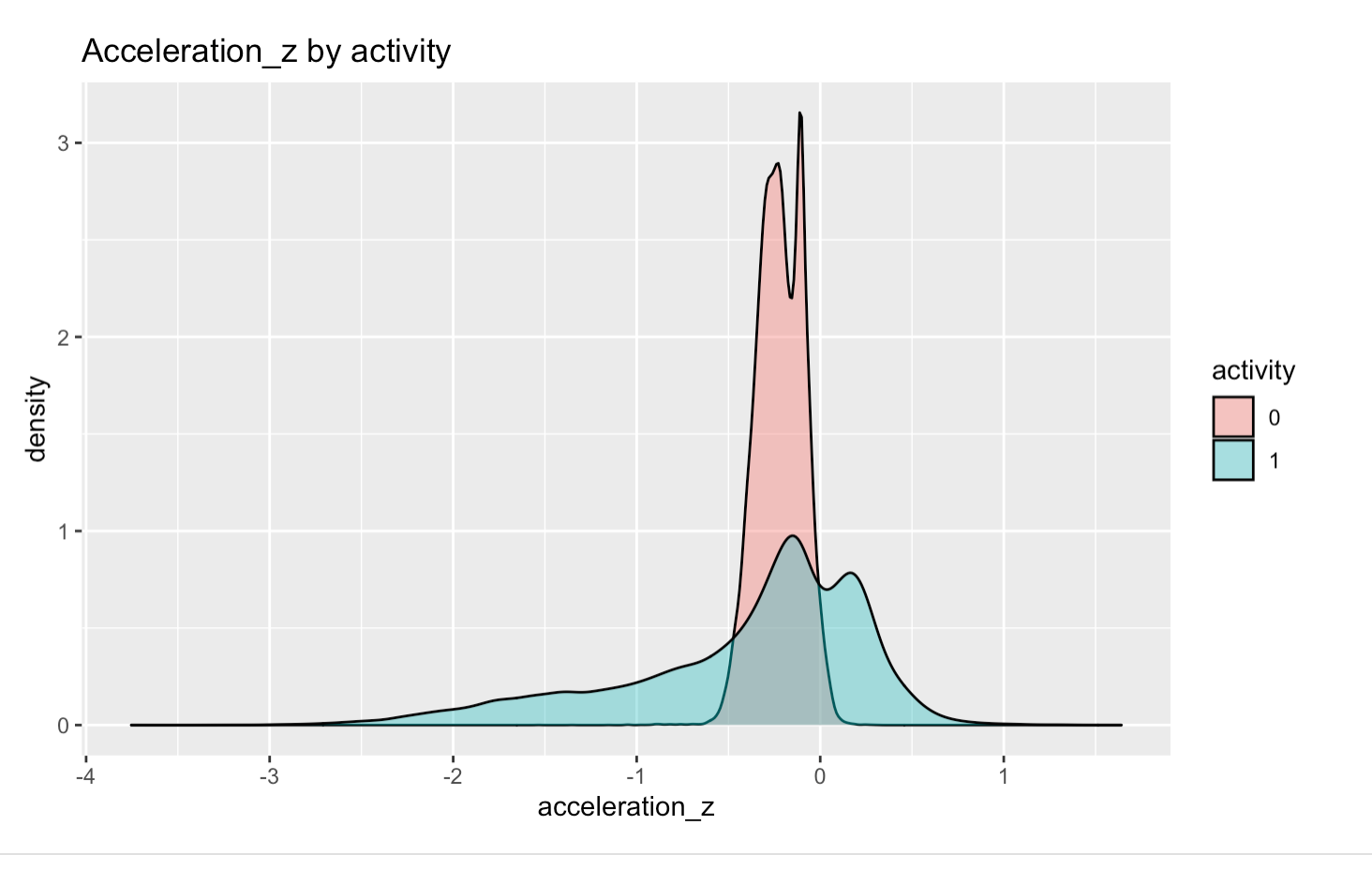
After that, we plotted the density graphs for each individual accelerometer and gyroscope variable below:



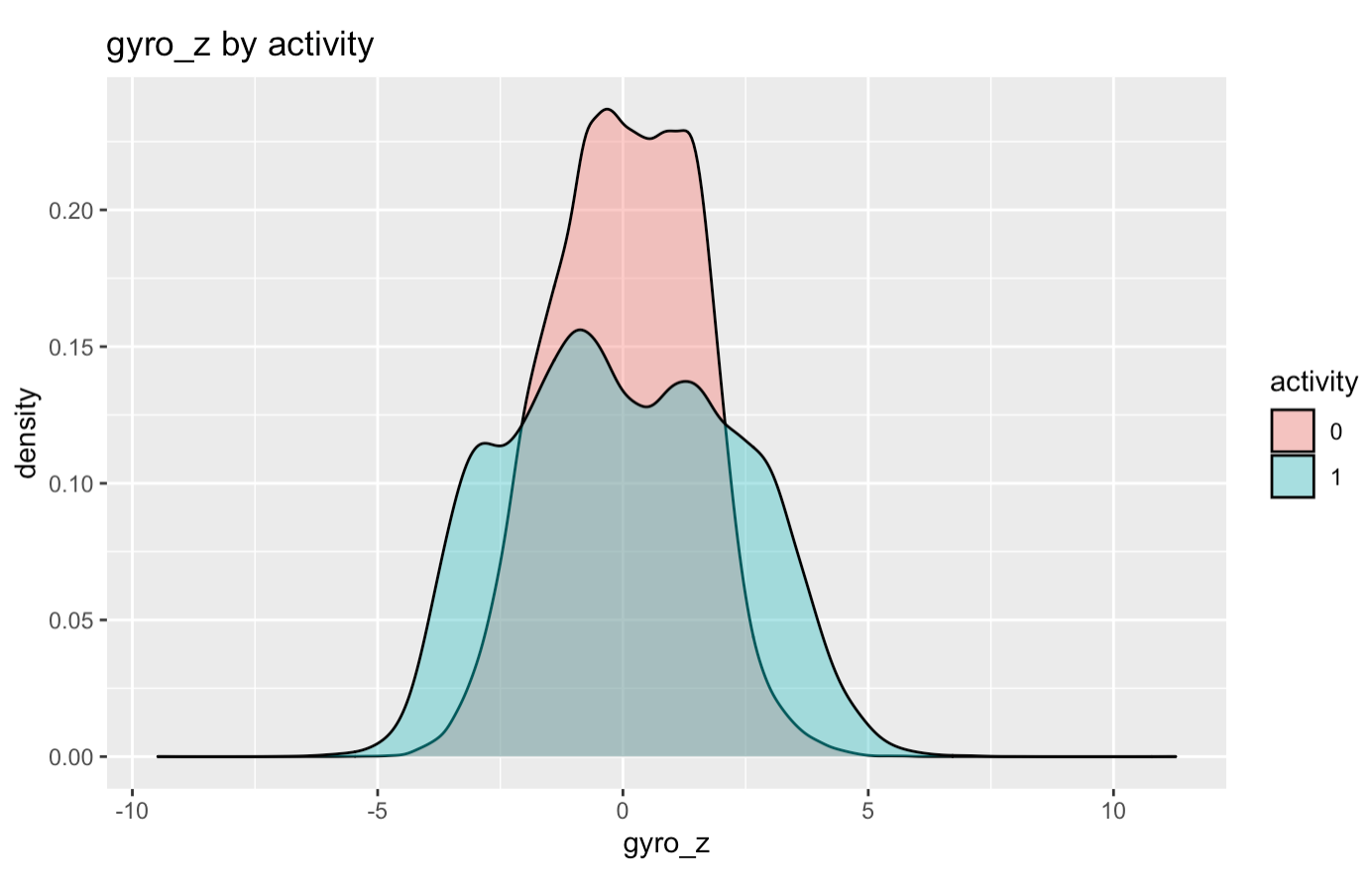
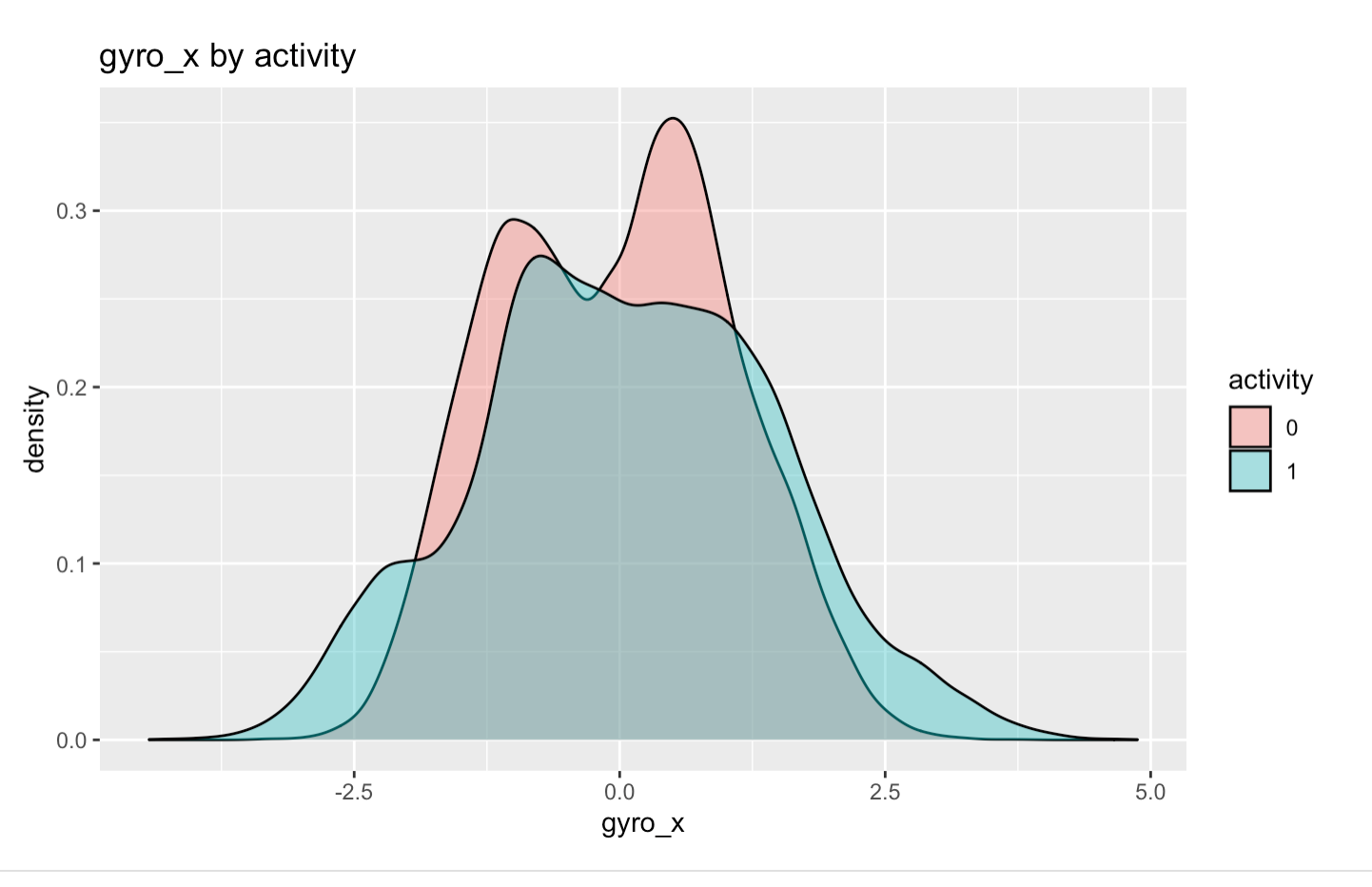
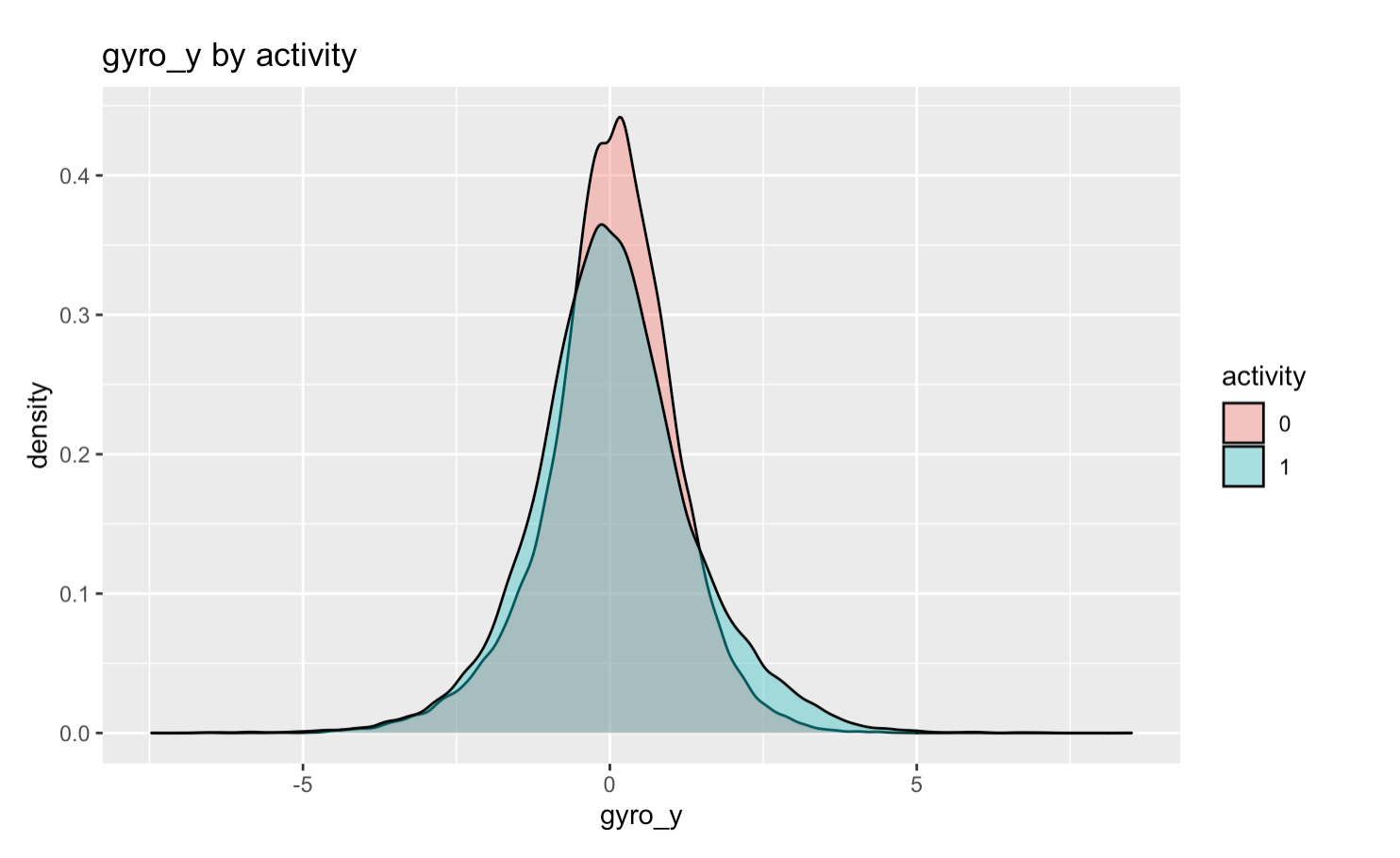
This plot showcases the acceleration in X direction. We observed that when the test subject is walking, the values in the plot have a bimodal distribution with peaks at two particular values on either side of 0. By comparison, when the test subject is running, the values are more spread out throughout the plot with a monomodal distribution peaking at zero. An observation worth noting from this is the observation of abnormally large values, likely outliers, in this data set. We may address those later on.



This plot showcases the acceleration in Y direction. We observed that when the test subject is walking, the bimodal distribution of acceleration in the Y direction is nearly entirely negative. By comparison, when the test subject is running, we encounter another bimodal distribution with peaks slightly favoring the negative acceleration. While it is much more difficult to hypothesize what has happened in the graph of acceleration in the X direction, we know that with acceleration in the Y direction, we see that the walking speed only gets slower by varying rates throughout the sessions recorded. We may interpret this finding as the subject beginning each walking session with a certain speed and then slowing that speed throughout the course of the session. With running, we see the test subject decelerates a bit less frequently than he accelerates, suggesting a fast-and-slow pattern of running with multiple speeds that alternate. We may verify this by looking at the individual sessions.



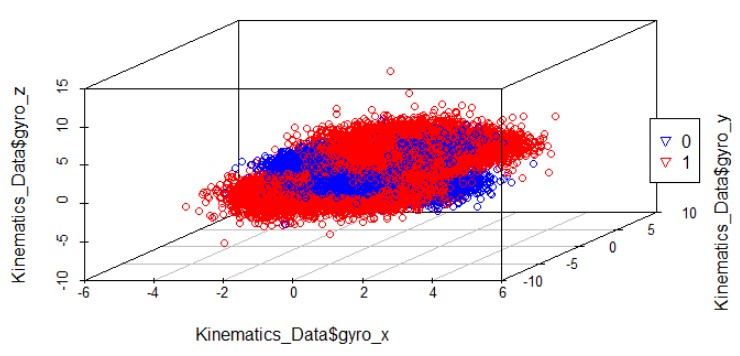
This plot showcases the acceleration in Z direction. We observed that when the test subject is walking, the values in the graph are bimodal and generally negative. On the other hand, while the test subject is running, the values have a larger distribution peaking at two values symmetric about the y-axis (a positive and negative value with roughly equal distance from 0). As acceleration in the Z direction is motion perpendicular to the ground, we interpret the readings of the test subject walking to be largely downward motion towards the ground during the step. When the test subject is running, we see still a large amount of downward motion towards the ground with a small amount of upward motion, raising the leg and foot as part of the step.

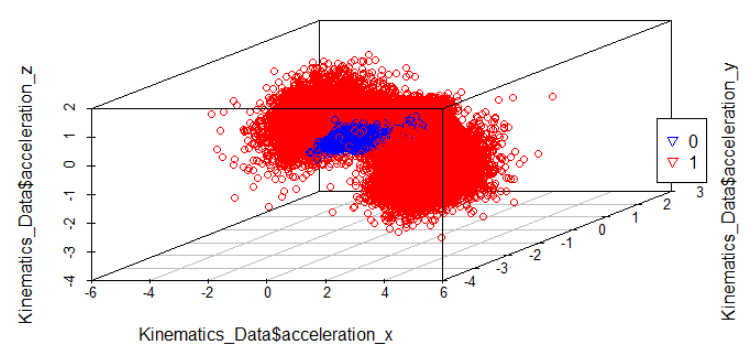


These plots showcase how the gyroscope changes in the x, y and z directions. Gyroscope is a type of angular activity and it helps with our internal body balance. For all of the gyroscope readings in all directions for both walking and running, the value stays around 0 most of the time, meaning the angular activity of the test subject is overall fairly centered.

The differences in these plots may be interpreted as follows:

* We observe a singular peak distribution in the gyro\_x plot at a negative value for the running graph, whereas we observe both a peak there and one at a positive value in that of the walking graph. This observation contrasts with the graph of acceleration\_x while walking, which had the negative value at a higher density.
* The gyro\_y plots tend to be similar between walking and running, though the walking graph may be slightly narrower leading to a higher peak on its density plot.
* The gyro\_z plots show very different shapes between walking and running; with both roughly symmetric plots, we may interpret the data to show equal upward and downward angular activity for each activity.

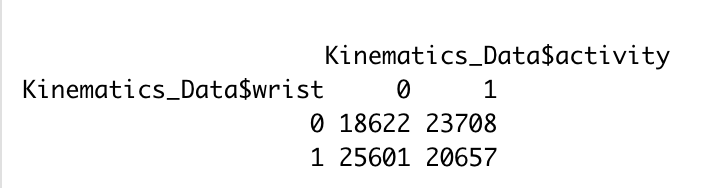




In the 3D plots above for acceleration and gyro variables, we can notice some distinctions between the observations of walking and running (0:walking, 1:running). In the 3D plot for the gyro variables, the running observations take more extreme values in the x direction and y direction compared to the walking observations. This makes sense since the angular velocity for side to side and up and down will be much greater in running than walking.

As for the 3D plot for the acceleration variables, we can see that the running observations surround the walking observations. This is interesting to see, because we observe the walking observations more clustered together and stay within a certain range. For the running observations on the other hand, we can see that the values are much greater than running in all directions. This makes sense, since acceleration increases when running.

In addition to analyzing the dataset as a whole, we wanted to divide the dataset into 4 subsections. We were interested in studying how the values differed between the right and left wrist, and if the values in walking or running are affected as well.



In the contingency table above, we can observe how the different combinations of activity and wrist are distributed. As stated in the introduction, 0 in wrist represents the left wrist, and 1 in wrist represents the right wrist. For activity, 0 represents walking and 1 represents running. The most popular combination of wrist and activity is the right wrist and walking. Additionally, for running, the left wrist is used more often than the right wrist. Each subsection has enough observations to compare directly to one another. To further analyze the different combinations of activity and wrist, we will take a look at the density graphs of each variable as done previously.

**Differences in Walking Acceleration Readings by Wrist:**

|  |  |
| --- | --- |
|  |  |
|  |  |

In the acceleration\_x graphs, we observe that the left wrist values make up the majority of the positive acceleration in the x direction previously observed, and similarly that the right wrist values make up the majority of negative acceleration in the x direction. We also observe a fairly symmetric bimodal distribution when the device is worn on the right wrist and the test subject is walking, whereas we see an asymmetric density plot when the device is worn on the left wrist and the subject is walking. There are only minor differences for the acceleration in the z direction. As with the following plots, we cannot causally attribute the differences entirely to the wrist which the device is worn, as we must then also control for the other variables such as date, time, and activity. However, given the size of the samples, we can reasonably infer that the wrist has some role in affecting the acceleration readings for walking.

**Differences in Walking Gyroscope Readings by Wrist:**

|  |  |
| --- | --- |
|  |  |
|  |  |

We observe very different density plots for the gyroscope readings in the x direction while walking once separated by wrist. In comparison, we do observe fairly similar plots for gyroscope readings in the y direction, as well as fairly symmetric density plots that create the symmetric density plot we observed previously. While the y direction plots show symmetry, the x direction and z direction plots provide further evidence that there are differences in readings for the gyroscope based on which wrist the device is on.

**Differences in Running Acceleration Readings by Wrist:**

|  |  |
| --- | --- |
|  |  |
|  |  |

We observe very different and asymmetric acceleration plots for running. The x direction plots are very nearly reflections of one another about 0. The other two plots show very little resemblance to one another as well.

**Differences in Running Gyroscope Readings by Wrist:**

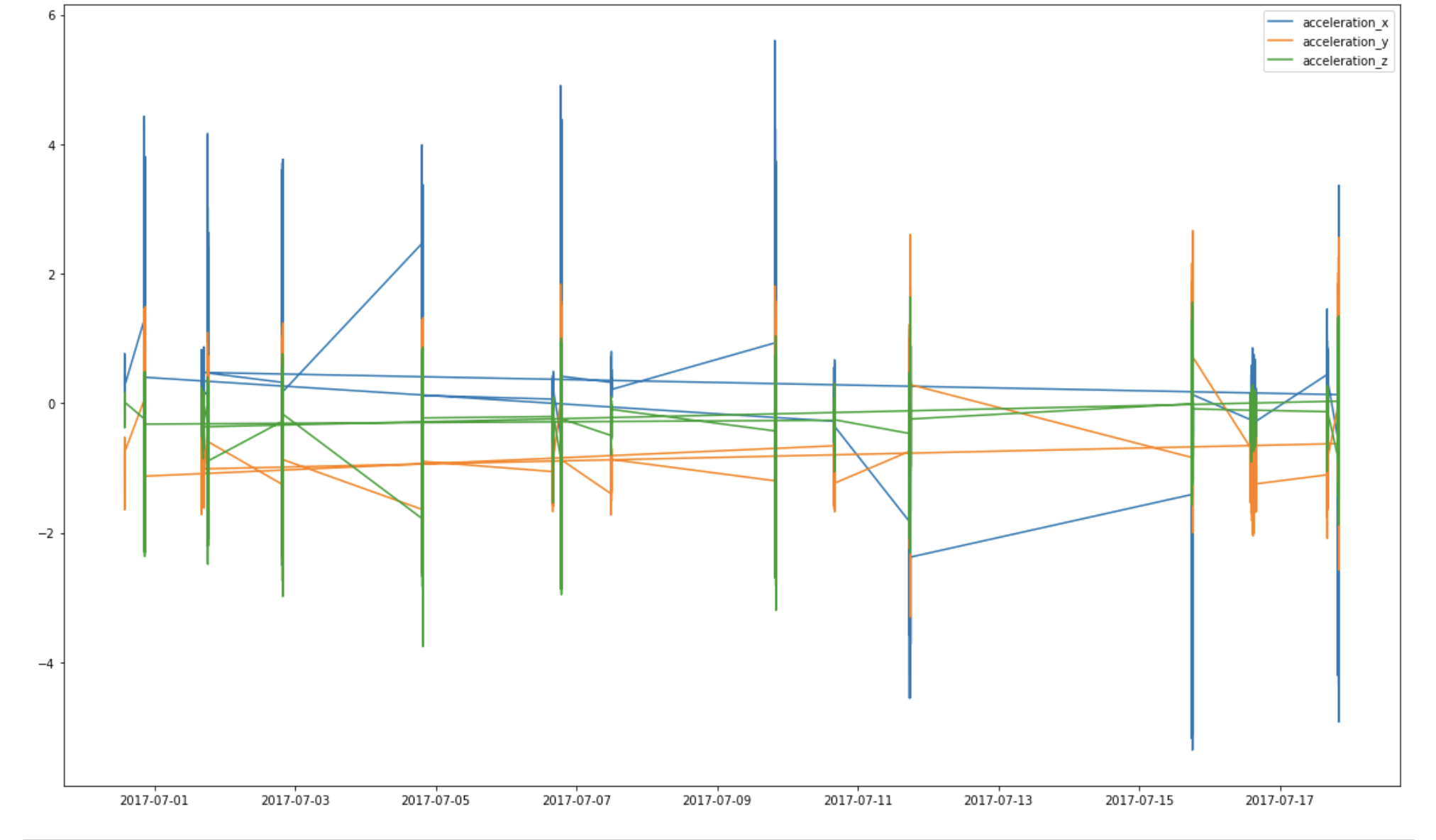
|  |  |
| --- | --- |
|  |  |
|  |  |

We see there are large differences in the graphs for gyroscope readings for running when separated by wrist. With the exception of the y direction plots, which once again show a fairly symmetric plot, we see some bimodal graphs that do not look like one another, which adds to our conclusion that the accelerometer and gyroscope readings have some significant differences based on which wrist is capturing the data.

## 

**Time Series Plot Analysis:**

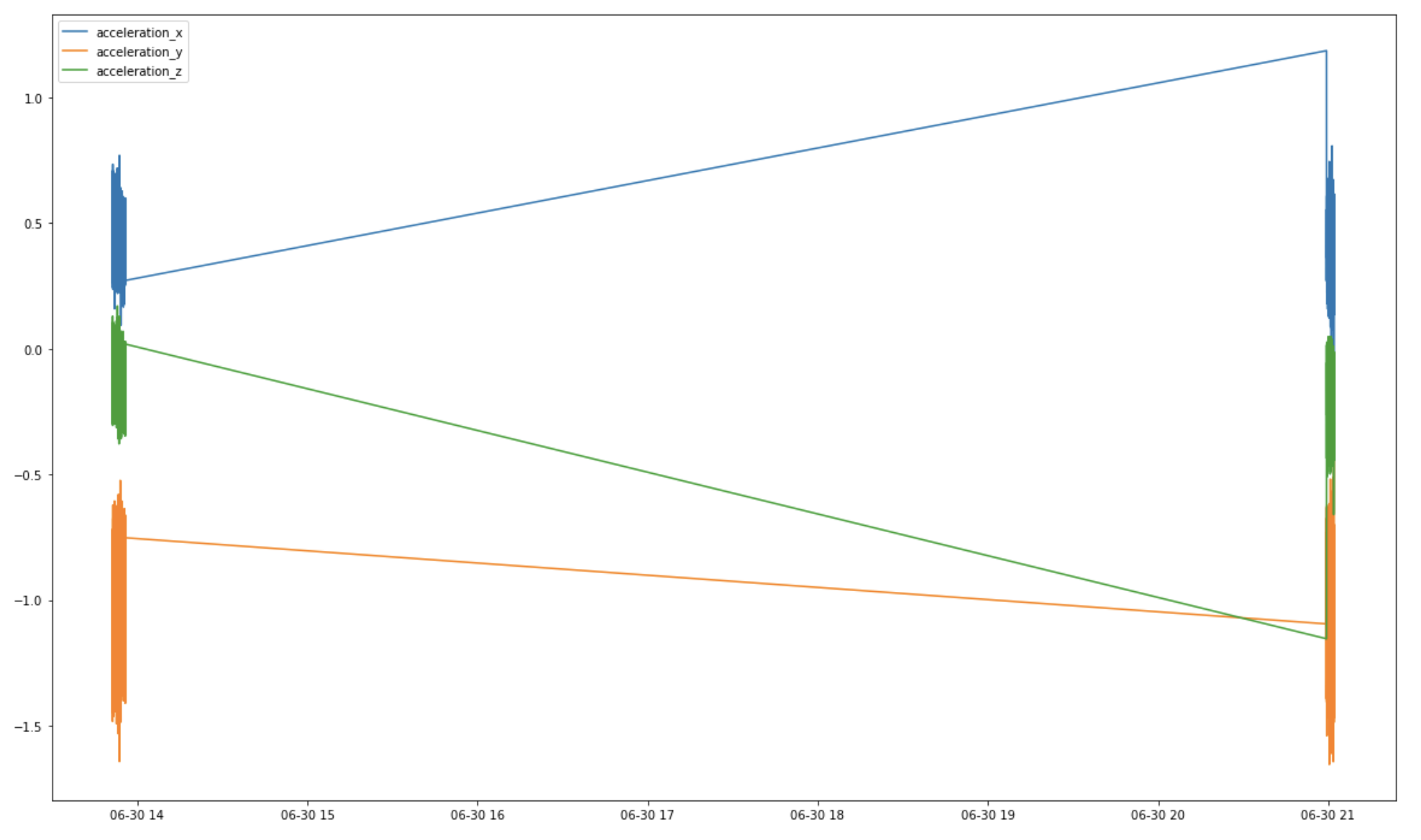
## Now that we understand how the values vary greatly among running and walking, we wanted to analyze how the acceleration values are affected over time. We experimented with gyro, but nothing meaningful could be drawn from the plots. We visualize this by creating time series graphs with time on the x axis, and the acceleration value on the y axis. The different colored lines represent the different directions as labeled in the legend.



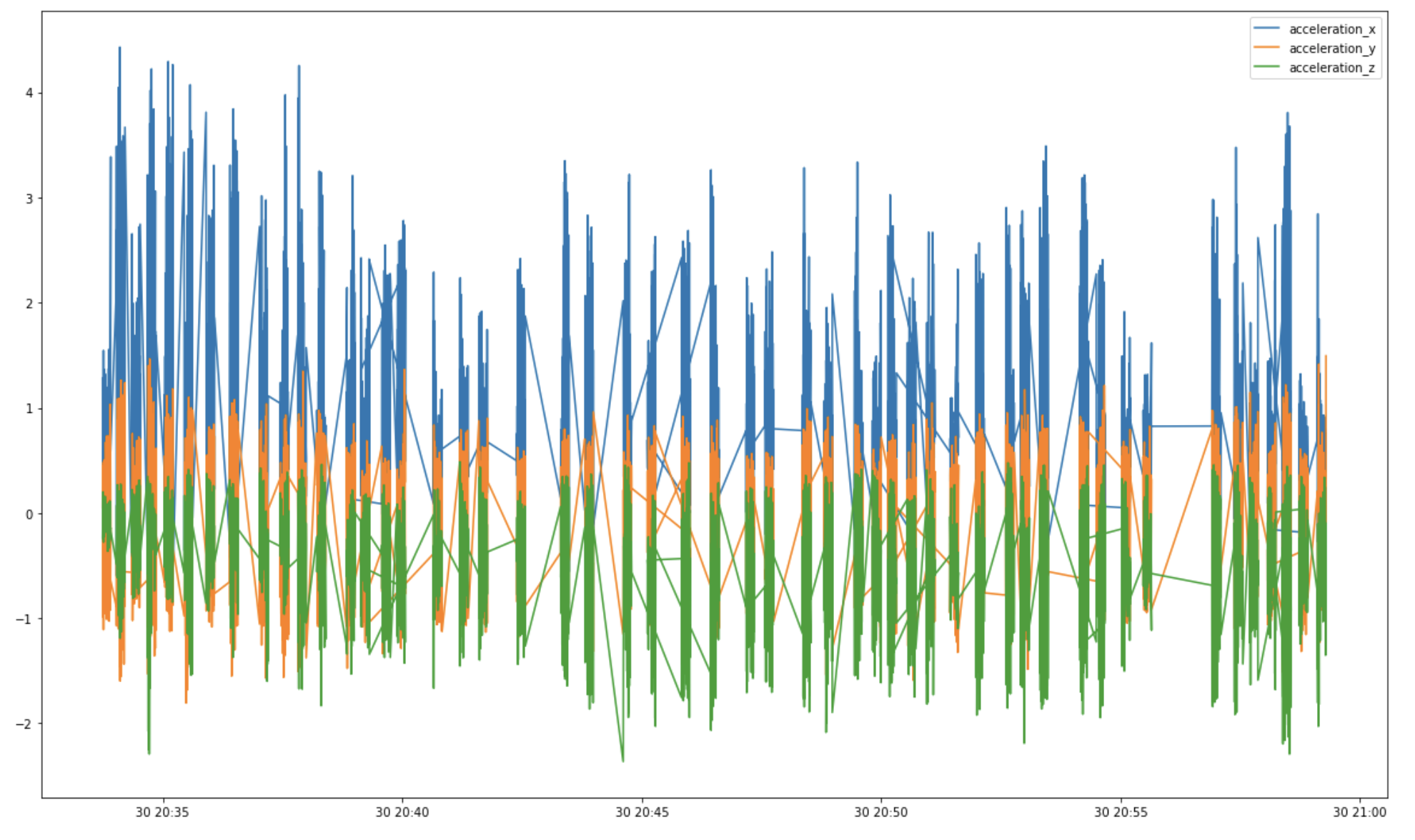
As we can observe, there are many observations each day, so we cannot clearly see the trend for each day. To analyze the trend for each day closely, we decided to focus on the first day and last day of the experiment. Analyzing each day would be time consuming and unnecessary. This way, we can see how the acceleration values changed in the first and last trial of the experiment.

We can see the time series graphs for the first day June 30, 2017 on the next page.

Walking for June 30, 2017



Running for June 30, 2017



From the visuals above, we can clearly understand the trend of how the subject started their workout by looking at the time values. They first started walking in the workout, and took a 5 hours break before continuing to run, but in the end, walked for the rest of the workout. Additionally, we can notice how the values fluctuate between running and walking.

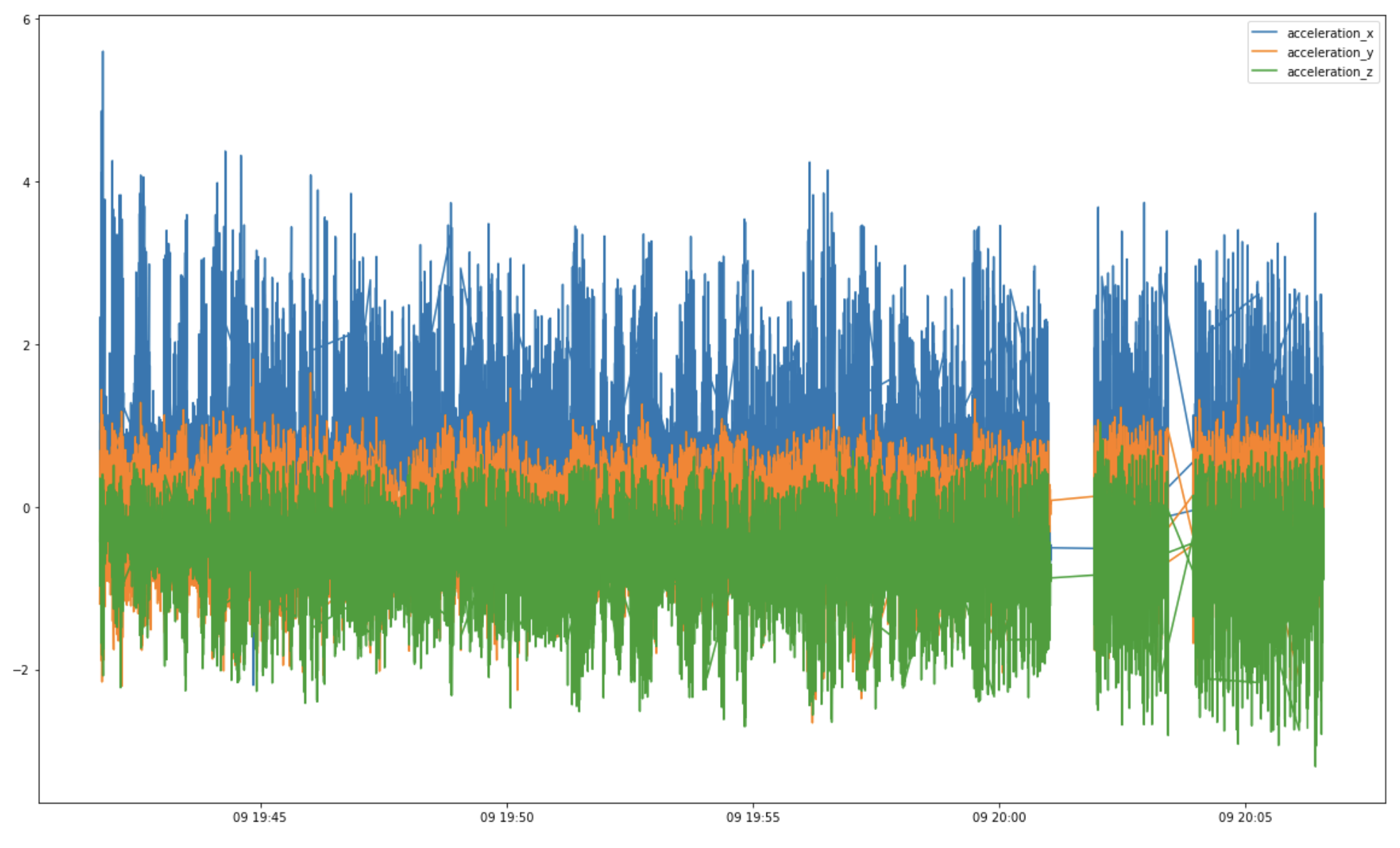
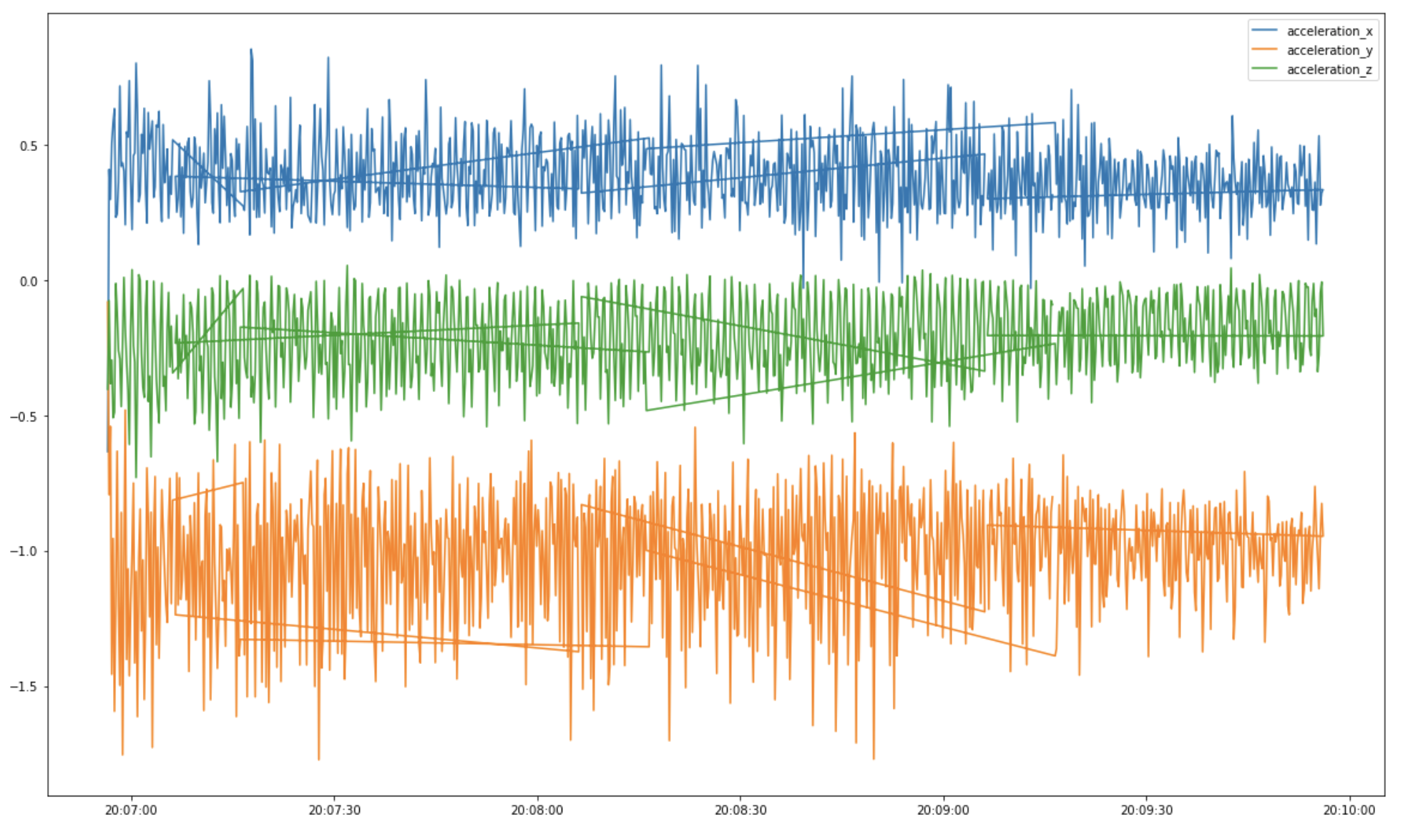
We see when walking, acceleration\_x is in a value range of about 0.2 to 0.7, acceleration\_y is in a value range of -1.5 to -0.5, and acceleration\_z is in a value range of about -0.4 to 0.1.

Since walking is a low acceleration activity, it makes sense that the size by size acceleration is low. It is positive because when walking, there is force applied to the side during the step. For the forward and backward acceleration which is acceleration\_y, it is negative since people are walking slowly and do not speed up when walking. For the up and down acceleration, the body is rarely moving up and down when walking, so it would make sense that the up and down acceleration is mostly negative.

Now when we analyze the running time series graph, we see a noticeable difference in the changes in value. The side by side acceleration now has greater values. For the forward and backward acceleration, we now see positive values since running requires some acceleration for higher speeds. We also see some positive values for up and down acceleration, since running would cause the body to move up and down at high running speeds.

After understanding how the walking and running values differ in the time series graph in a day, we are interested in how the runner changes acceleration speeds in the last day of the experiment. The time series graphs for the last day can be seen on the next page.

Walking for July 9, 2017



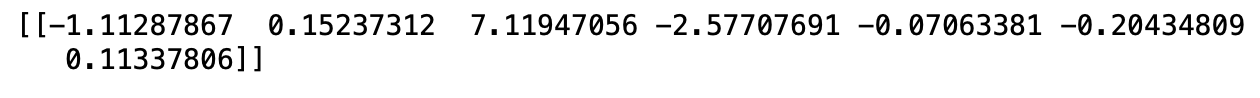
Running for July 9, 2017

## Compared to the first day, we can immediately notice that the graphs are smoother with little break in between. This could indicate that the test subject was experimenting with the equipment at first, or was not wearing the equipment properly. Additionally, we can see higher values of acceleration in the last day as the highest side by side acceleration had a value of above 5 and the highest forward and backward acceleration had a value of almost 2. Compared to the first day values, the test subject accelerated harder on the last day.

## Model Building

In addition to data analysis exploration, we wanted to build the best model to predict whether it is walking or running given our data. The variables we included in all models are 'wrist', 'acceleration\_x', 'acceleration\_y', 'acceleration\_z', 'gyro\_x', 'gyro\_y', 'gyro\_z'. We utilized logistic regression, a logistic generalized additive model, and neural networks as our options for the best model. Before using each technique, we randomly split 30% of our data to represent the test data, and the other 70% as the training data in order to train our model.

Since our response model is binary, we definitely wanted to utilize a logistic regression model to predict the activity variable. Using the logistic regression function in python from the sklearn.linear\_model package, we fitted the training data and predicted the running and walking values utilizing the test data. To obtain the error rate, we counted the number of incorrect observations and divided them by the total observations to obtain the error rate. Logistic regression obtained an error rate of 0.13966963916168115. To analyze the model even further, we printed out the coefficient values shown below which are in order of the following predictor variables: 'wrist', 'acceleration\_x', 'acceleration\_y', 'acceleration\_z', 'gyro\_x', 'gyro\_y', 'gyro\_z.' Clearly, the largest coefficient in magnitude is ‘acceleration\_y,’ and it is also positive, indicating that increasing acceleration\_y will make it more probable that the activity is running. This is no surprise to use as the graphs we analyzed in walking and running indicated that acceleration\_y was the biggest difference between the two activities. Running tends to have positive acceleration in the y direction, and walking tends to have negative acceleration in the y direction.



We also wanted to include a logistic generalized additive model to see if there is any nonlinear relationship in the data. Following the same process as logistic regression and using the LogisticGAM function from the pygam package, we obtained an error rate of 0.0205817059863792. This is significantly better than the standard logistic regression model, and clearly we see that there is a nonlinear relationship in our data. This is expected since from the graphs, we can notice a pattern visually among the walking and running observations.

In the model using neural networks, we implemented a three layer neural networks model utilizing the MLPClassifier model from the sklearn.neural\_network package. For the activation parameter, we used a logistic solver, since logistic regression produced a low error rate. We used the “adam” optimizer since we were dealing with a large data set. The minibatch size for the optimizer was 200. The maximum number of iterations was 1000. An adaptive learning rate was used for the algorithm. The multi-layer neural networks model produces an error rate of 0.009218497196824322, which is the most accurate model out of the three.

We found the neural networks model was the best fit for our dataset because we have a large dataset and we noticed the patterns among walking and running observations from our initial exploratory data analysis. Neural networks detect patterns similar to how the human brain works, so it would make sense that it can achieve an extremely low error rate for the data. Neural networks also work well for large datasets, and that turned out to be the case for the dataset we were using.

## Conclusion

As we analyzed the kinematics motion data set, we set to classify the type of activity based on the accelerometer and gyroscope readings as well as some other elements. From the exploratory data analysis sections, we found (and later confirmed during logistic regression) that there were significant changes between the readings depending on which wrist the device was on.

In our analysis of the time series plots, we looked at the notable differences in the first and last days of activity by the test subject. We saw that compared to the first day, the last day of activity had much smoother readings as well as higher readings in acceleration, suggesting a stronger effort by the subject during the later running workouts.

Lastly, we applied multiple classifier models to the data to classify the activity based on the readings of the accelerometer, the gyroscope, and the wrist that the device was on. Among the three models we applied, the three-layer artificial neural network had the lowest error rate in classifying the observations correctly.

Using machine learning and a variety of statistical methods, we can train technology to accurately classify the type of activity based on the accelerometer and gyroscope readings. As technology improves, we can incorporate additional readings and information into the models to improve the classification accuracy further. Later still, one may expand the application of these models beyond just fitness tracking into other industries and usages.

## References

[1] M. Yasser H.. Kinematics Motion Data. Retrieved April 1, 2022 from <https://www.kaggle.com/datasets/yasserh/kinematics-motion-data>.