ACTIVITY RECOGNITION

ASSIGNMENT #1

Vanessa Ulloa | CSE 572: Data Mining | 07 Apr 2019

TABLE OF CONTENTS

PHASE 1: DATA COLLECTION

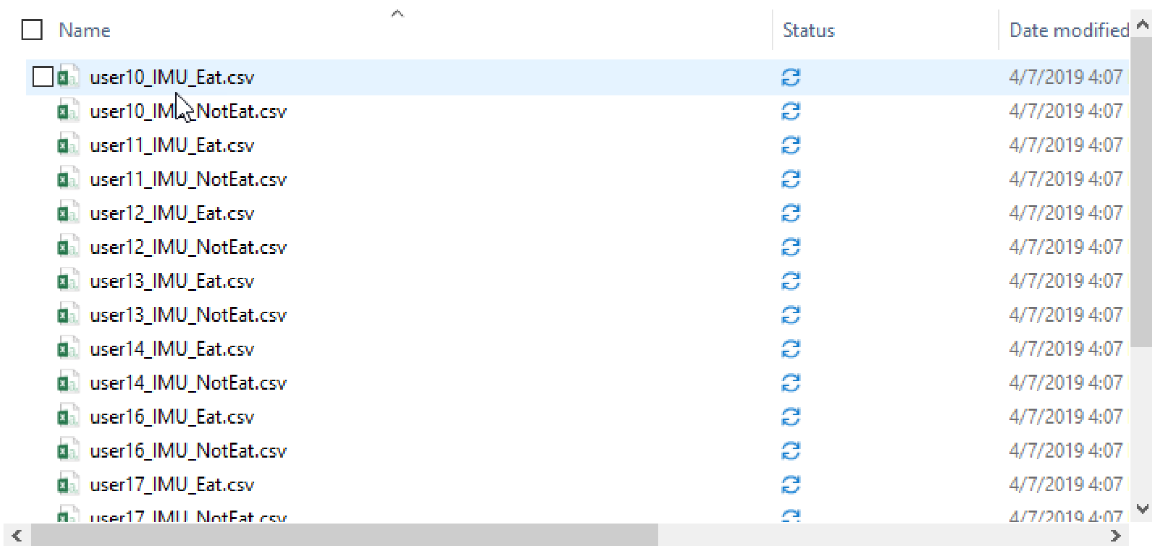
The first part of this project was to organize the raw data. The files that were given per user were one file with IMU sensor data, EMG sensor data and a file with frame numbers for either an action with a fork or spoon for a user. The file with the frame numbers (labeled {userid}.txt) contains 3 columns of data, the first column is a start frame number and the second column is the end frame number; the third column is not used. The significance of these start and end frames are that they are in relation to video taken of a user with the Myo wrist band and when they are engaging in an eating action. We will use this frame numbers in order to synchronize the IMU and EMG data and sort out which data belongs to eating actions and which data belongs to non-eating actions.

The video data will be the ground truth data used in this project. The assumed frames per second in this video is 30 fps. The sampling rate for the IMU data is 50 Hz or 50 samples per second, the sampling rate for the EMG data is 200 Hz or 200 samples per second. In order to synchronize the IMU data the start frame in the ground truth file was multiplied by (50/30) and similarly the end frame. This new start and end sampling value is used in order to retrieve the start and end row from the IMU file that correlates to the eating action frames. The same method is used for the EMG data except for the calculation is (200/30) since the sampling rate is not the same as the IMU sampling rate. Similarly, the new start and end sampling value is used to retrieve the data from the EMG file that correlates to the eating action.

The non-eating data was found using a similar method except that we use values in between a start frame and the previous end frame in order to indicate non-eating actions. For example, we iterate through the ground truth file start at $i=2$, then we apply the calculation (either for EMG or IMU) to the start frame at row i and the end frame at row $i-1$ (the previous row). This range gives us the rows that correlate in either the IMU or EMG files for a non-eating action.

The IMU data that is provided includes sensor data for orientation, accelerometer and gyroscope. This results in ten different sensor data points (OriX, OriY, OriZ, OriW, AccX, AccY, AccZ, GyroX, GyroY, GyroZ). The EMG data was not used in this project due to poor synchronization, there were many issues synchronizing the ground truth file with the EMG data.

The IMU sensor data was stored into a matrix where all the eating actions were appended and each column signifies a particular sensor (ten columns) and this data is stored per user into a CSV file.



	A	B	C	D	E	F	G	H
1	0.779	0.565	0.083	-0.259	0.895	-0.3	-0.371	-1.312
2	0.779	0.565	0.086	-0.26	0.84	-0.261	-0.354	15.25
3	0.777	0.567	0.09	-0.258	0.841	-0.375	-0.354	33.688
4	0.776	0.57	0.095	-0.254	0.87	-0.504	-0.3	34.25
5	0.773	0.574	0.098	-0.251	0.913	-0.454	-0.279	31.875
6	0.771	0.579	0.101	-0.247	0.845	-0.396	-0.299	39.812
7	0.769	0.583	0.105	-0.24	0.868	-0.458	-0.228	44.688
8	0.768	0.586	0.109	-0.236	0.89	-0.515	-0.235	33.688
9	0.766	0.588	0.111	-0.235	0.811	-0.491	-0.228	13.188
10	0.765	0.59	0.11	-0.233	0.826	-0.483	-0.209	1.75
11	0.766	0.592	0.105	-0.23	0.818	-0.479	-0.132	-11.188
12	0.768	0.591	0.098	-0.226	0.882	-0.418	-0.136	-25.688
13	0.77	0.589	0.09	-0.226	0.854	-0.307	-0.213	-39
14	0.774	0.585	0.083	-0.227	0.805	-0.243	-0.289	-37.5
15	0.778	0.582	0.076	-0.225	0.889	-0.245	-0.355	-26.438
16	0.781	0.579	0.07	-0.221	0.884	-0.385	-0.257	-23.062
17	0.785	0.576	0.065	-0.22	0.935	-0.477	-0.159	-33.125
18	0.787	0.573	0.058	-0.221	0.959	-0.34	-0.243	-45
19	0.789	0.569	0.051	-0.224	0.866	-0.236	-0.336	-43.625
20	0.792	0.566	0.045	-0.225	0.833	-0.221	-0.329	-32.938

This was done for ten users for a total output of 20 csv files. The data for the non-eating actions is significantly longer than those for eating actions which provides a variation in time between eating and non-eating actions across all sensor data points.

PHASE 2: FEATURE EXTRACTION

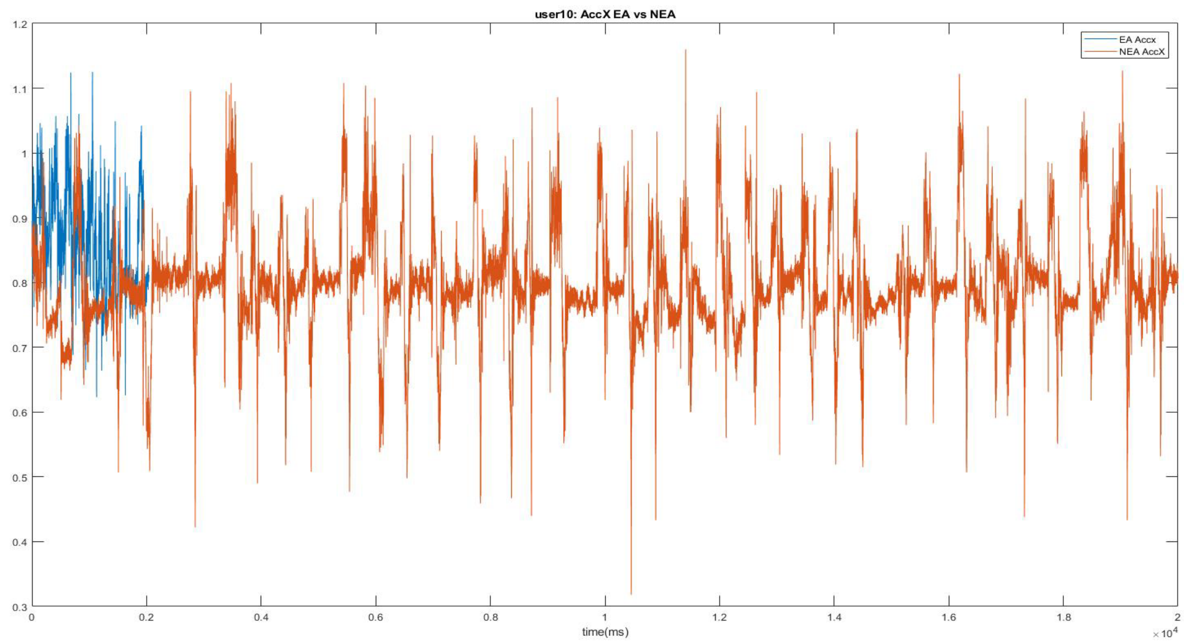
The data is organized into .csv files where all eating actions are concatenated together but each sensor data point resides in it's own column. For example:

	A	B	C	D	E	F	G	H	I	J	K
1	0.779	0.565	0.083	-0.259	0.895	-0.3	-0.371	-1.312	5.875	19.312	
2	0.779	0.565	0.086	-0.26	0.84	-0.261	-0.354	15.25	0.125	23.25	
3	0.777	0.567	0.09	-0.258	0.841	-0.375	-0.354	33.688	7.812	9.25	
4	0.776	0.57	0.095	-0.254	0.87	-0.504	-0.3	34.25	14.438	-3.062	
5	0.773	0.574	0.098	-0.251	0.913	-0.454	-0.279	31.875	16.625	-0.812	
6	0.771	0.579	0.101	-0.247	0.845	-0.396	-0.299	39.812	25.375	-5.5	
7	0.769	0.583	0.105	-0.24	0.868	-0.458	-0.228	44.688	13.062	-15.438	
8	0.768	0.586	0.109	-0.236	0.89	-0.515	-0.235	33.688	10.625	-8.625	

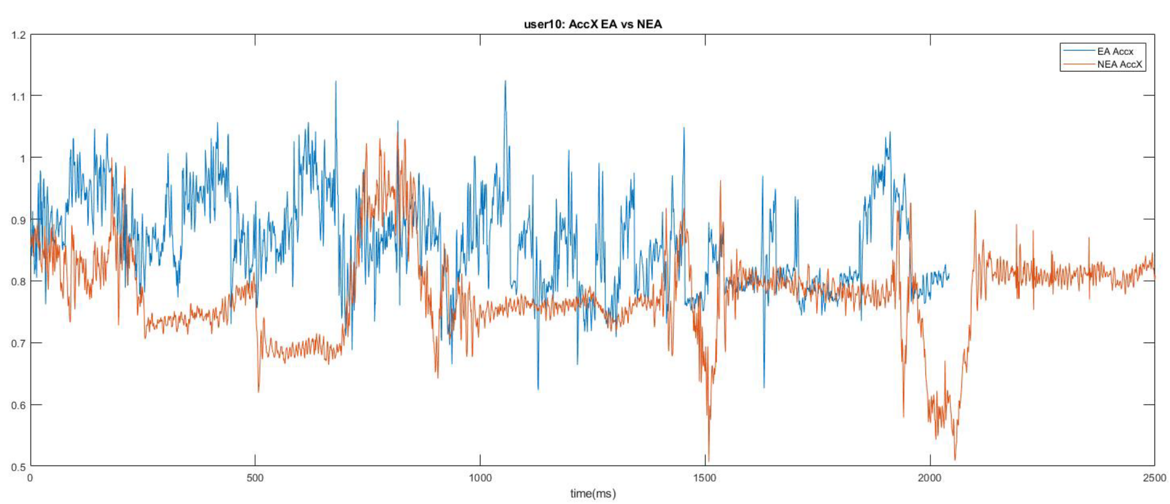
In this diagram, the columns are related to the IMU data:

- Column A: Orientation X
- Column B: Orientation Y
- Column C: Orientation Z
- Column D: Orientation W
- Column E: Acceleration X
- Column F: Acceleration Y
- Column G: Acceleration Z
- Column H: Gyroscope X
- Column I: Gyroscope Y
- Column J: Gyroscope Z

I checked various sensor data points for individual users, for example an Acceleration X for user ten comparison between Eating and Non-Eating actions.



The first thing we notice is that the length of time for the non-eating actions is significantly greater than those for eating actions. The same graph with a shortened x-axis also shows that the Acceleration X data points have a lot of variance in comparison to the non-eating data.



Variances were checked among other dimensions (sensors) to look for variation and where the following feature extraction methods could be applied to the entire sample set (ten users).

1. Mean
2. Standard Deviation
3. FFT (Fast Fourier Transformation)
4. Minimum
5. Maximum

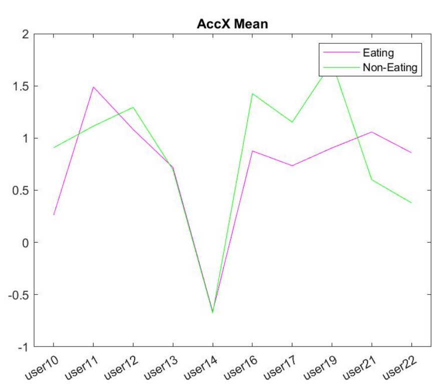
MEAN

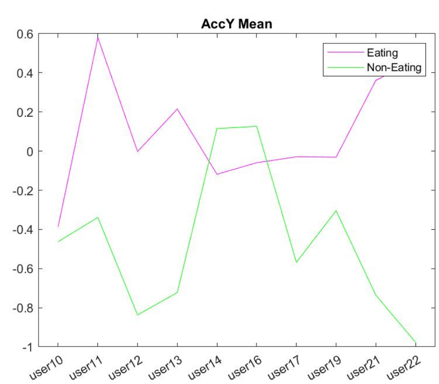
The mean is used in data mining with continuous data and is essentially an average of a given data set. All the values are added together and then divided by the total number of data points in the set. Mean was chosen because graphs such as the graph for Acceleration X for user10 displayed different values for Eating and Non-Eating actions. Using the mean the graphs will display the average across all the users (ten users) for Eating and Non-Eating actions.

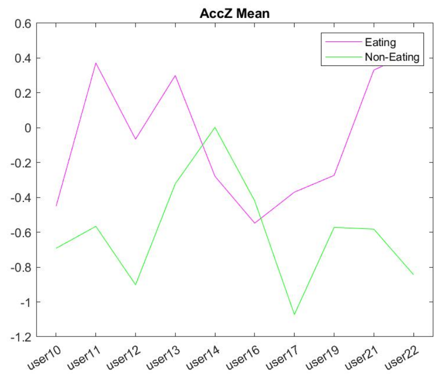
In MATLAB the mean was taken for every sensor value for every user, these values where then normalized and saved so that they could be graphed. The normalization of a mean value indicates how many standard deviations that value is the from the mean, this is also known as the z-score.

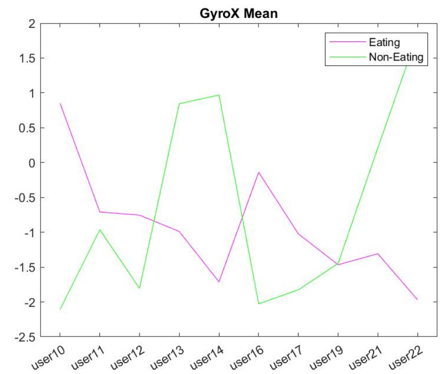
Graphs were created for ten users displaying the mean per sensor and the results varied.

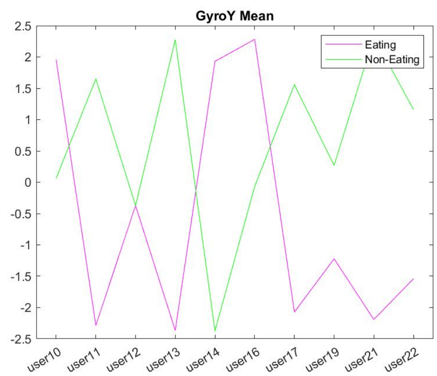
- Acceleration X showed similar data, showing a lack of variance of acceleration along the horizontal axis:
- Acceleration Y however showed a large variance not only among the users but also the vertical axis movements. Acceleration also Z showed a large variance across the Z-axis.
- The Gyroscope dimensions all showed a lot of variance from Eating and Non-Eating actions.
- The Orientation dimensions varied, where Orientation W did not show much variance and Y a little more, but X and Z were consistently different.

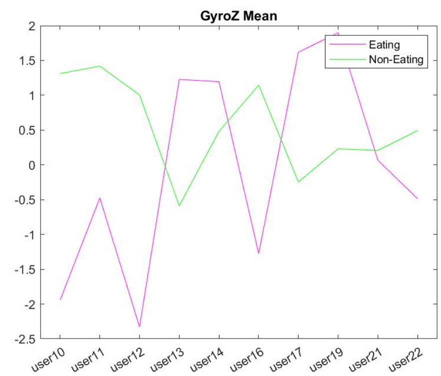


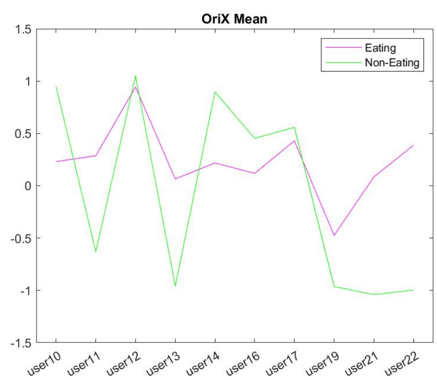


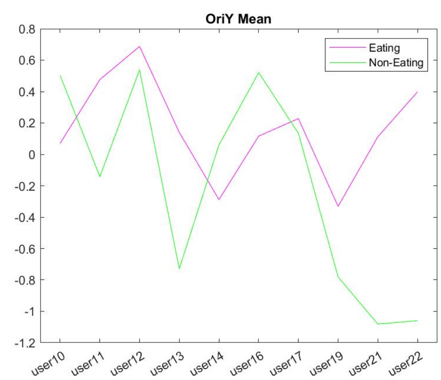


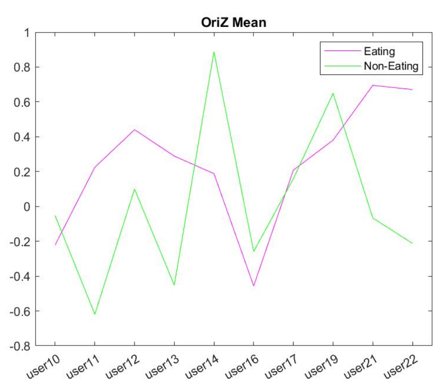


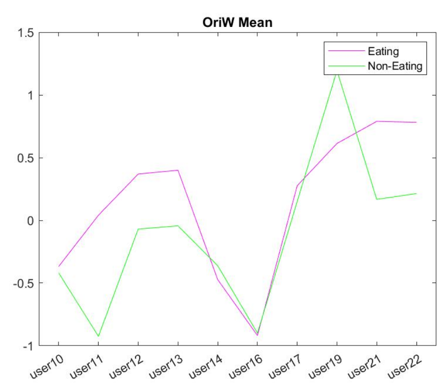










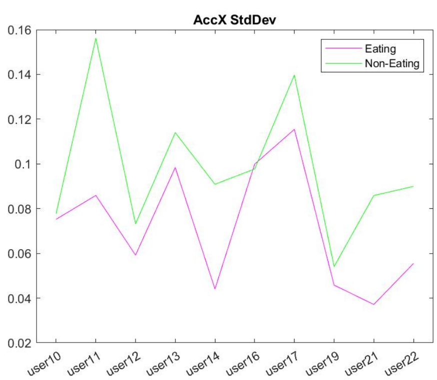


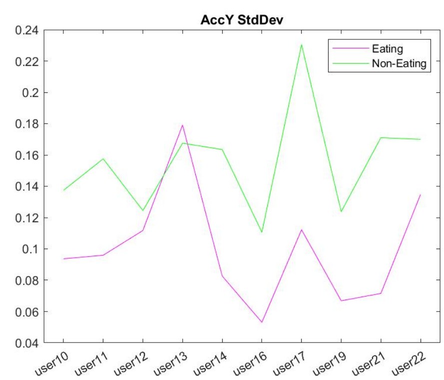
STANDARD DEVIATION

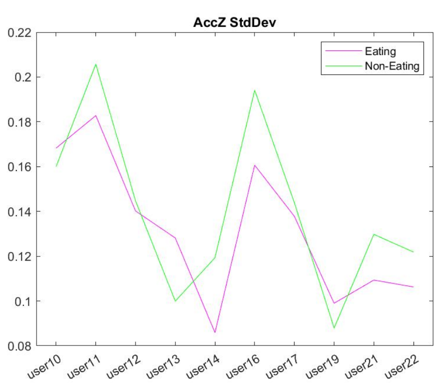
Standard deviation is a measurement that can be used to indicate how close a value or values are to the mean and therefore can provide a display of how varied a data set is. Since the standard deviation is related to the mean, it made sense to also graph and view how the standard deviation appeared for each of the sensor dimensions among the users.

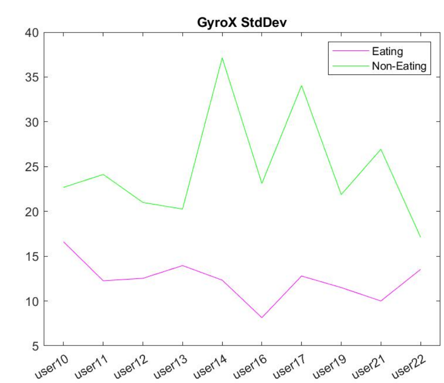
In MATLAB the standard deviations were taken for every sensor dimension across all the users and compile together. A graph was created per dimension in order to view the variance of the data.

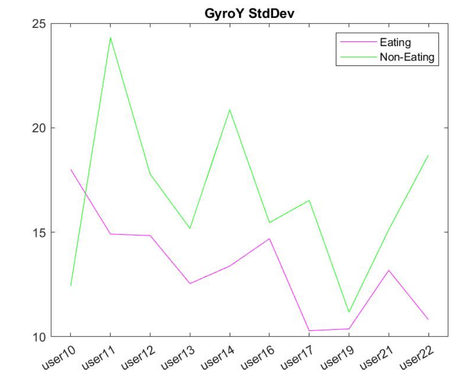
The results from the graphs showed that for seven of the 10 sensor dimensions there was a similarity from Mean. The dimensions that were different were Acceleration X which showed little variance and Orientation X and Orientation Y. Upon closer examination, however on the standard deviation graph or Orientation Z for user17 there is a consistent spike between eating and non-eating activities and in comparison, on the Mean graph for Orientation Z for user17 the points on the graph are close to one another. While visually the graphs appear different, if we take the Mean and Standard Deviation Graphs for Orientation Z and compare each user's data points we can see a consistency in that points that are close to one another on one graph, are also on the other. This shows that there is not much Orientation variance between eating and non-eating actions.

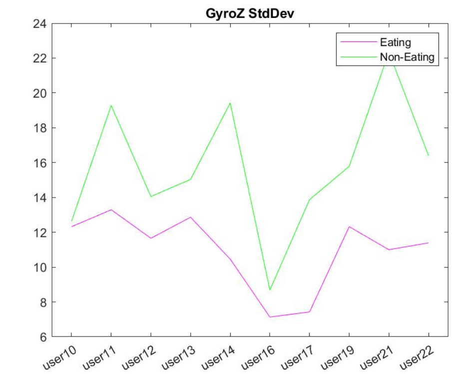


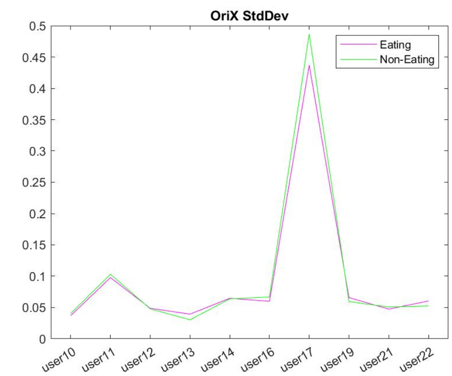


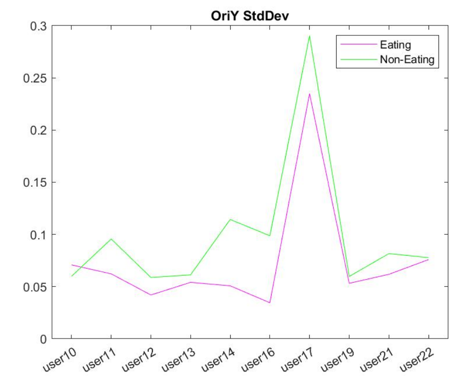


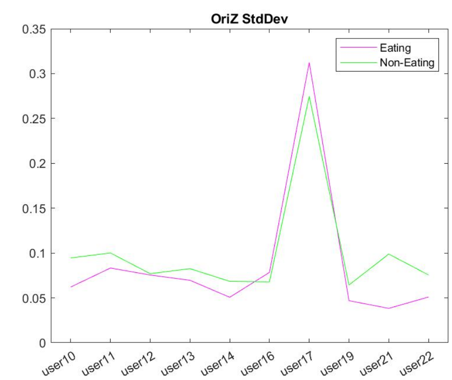


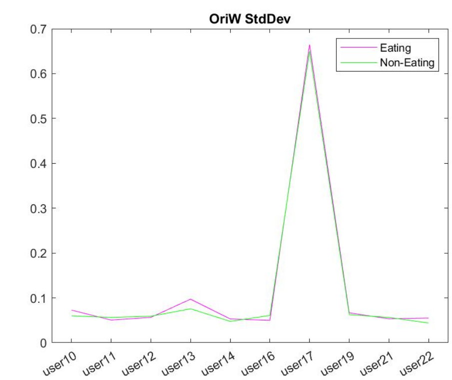










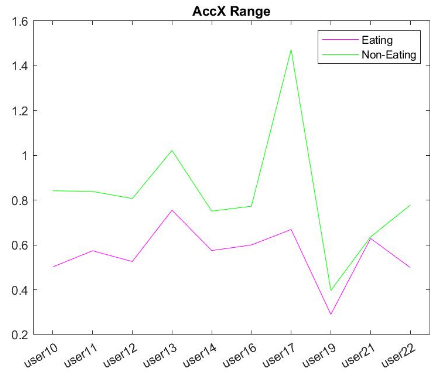


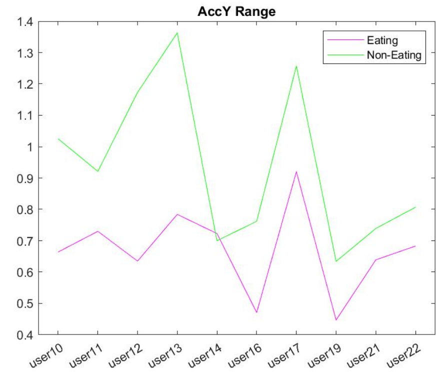
RANGE

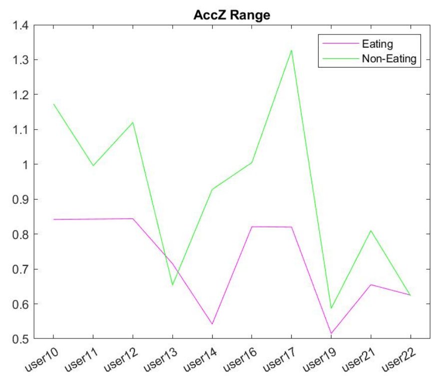
Range is the calculation of the difference between the maximum and minimum values in a data set. This can be used on this data set to distinguish between eating and non-eating actions. Certain dimensions such as acceleration and gyroscope will have a higher range since eating actions are limited in the variance in those dimensions versus non-eating actions.

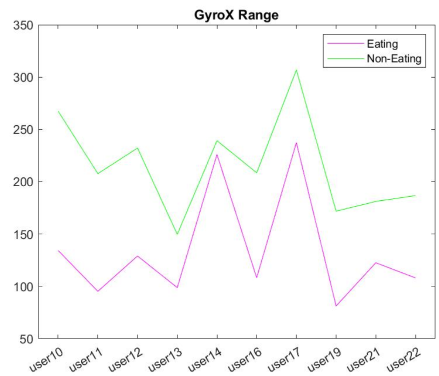
In MATLAB the range was acquired for every dimension for every use by subtracting the minimum value by the maximum value. This produced a range.

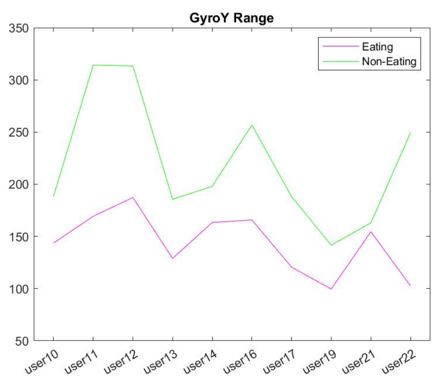
As predicted the variance between all the dimensions except for the four orientation dimensions show that there is much more movement in non-eating actions vs eating actions.


AccX Range
Eating
Non-Eating
user10 user11 user12 user13 user14 user16 user17 user19 user21 user22

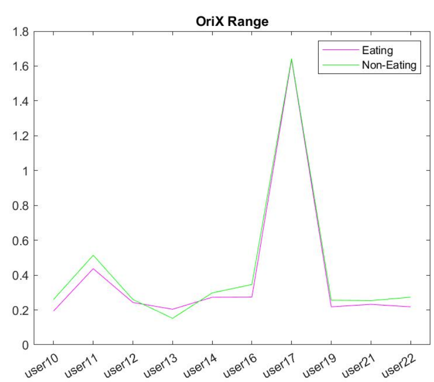

AccY Range
Eating
Non-Eating
user10 user11 user12 user13 user14 user16 user17 user19 user21 user22

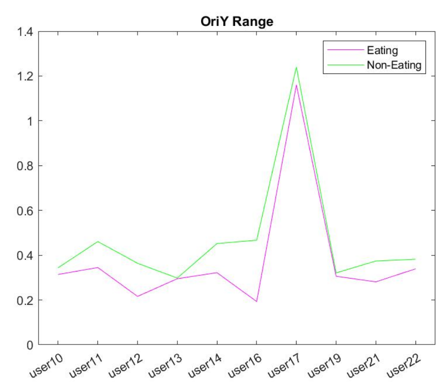

AccZ Range
Eating
Non-Eating
user10 user11 user12 user13 user14 user16 user17 user19 user21 user22

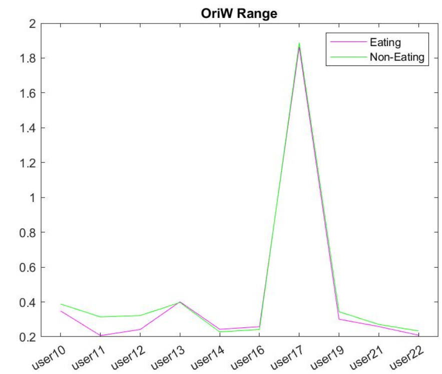

GyroX Range
Eating
Non-Eating
user10 user11 user12 user13 user14 user16 user17 user19 user21 user22


GyroY Range
Eating
Non-Eating
user10 user11 user12 user13 user14 user16 user17 user19 user21 user22


GyroZ Range
Eating
Non-Eating
user10 user11 user12 user13 user14 user16 user17 user19 user21 user22


OriX Range
Eating
Non-Eating
user10 user11 user12 user13 user14 user16 user17 user19 user21 user22


OriY Range
Eating
Non-Eating
user10 user11 user12 user13 user14 user16 user17 user19 user21 user22

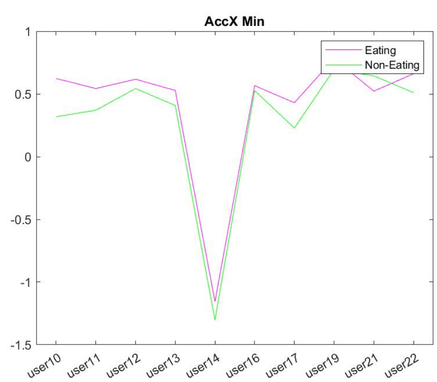


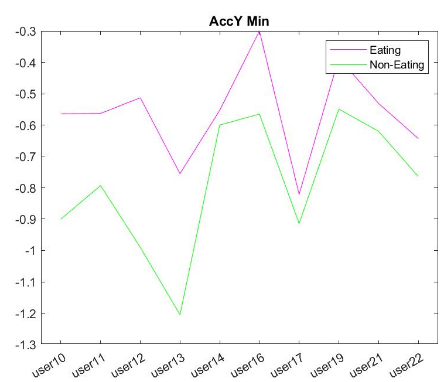
MINIMUM

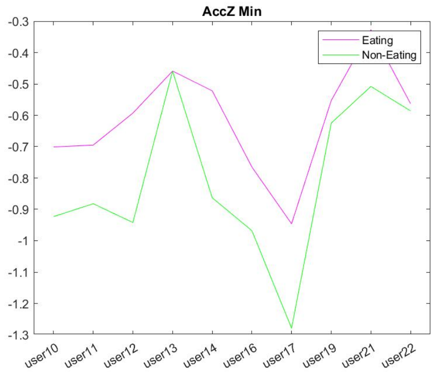
The minimum calculation is taking the minimum values per sensor dimensions per user. This distribution of values can show the differences between eating and non-eating data by exploring that variances in the minimum values in relation to each dimension. Dimensions such as Orientation may show lower variance between the activities as an eating action is a very specific action and the range of values for such an action especially in comparison to a non-eating action where there are many options of movement.

The Minimum values were extracted using MATLAB and graphed.

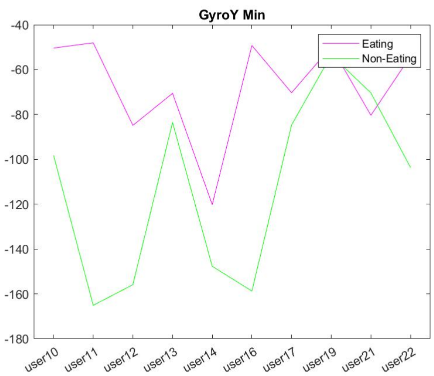
The distribution on the graphs shows that for the dimensions such as Orientation there was not much of a difference, this tells me that in terms of range and with the range and maximum graphs that the Orientation values for X, Y, Z and W did not have much variance and cannot accurately distinguish between eating and non-eating actions. However, the sensor dimensions for Acceleration and Gyroscope showed a variance in minimum values where eating actions were always a larger value than the minimums of non-eating actions. This shows, along with the maximum graphs, that the range for these sensor dimensions is much larger in non-eating actions possibly due to more movement in general versus a typical eating action.

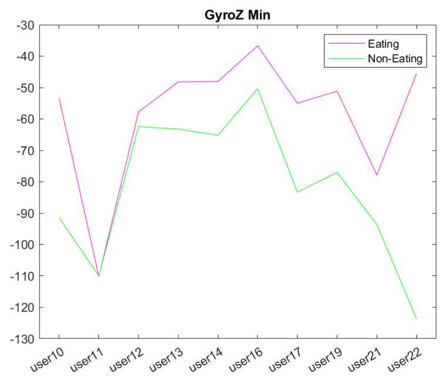


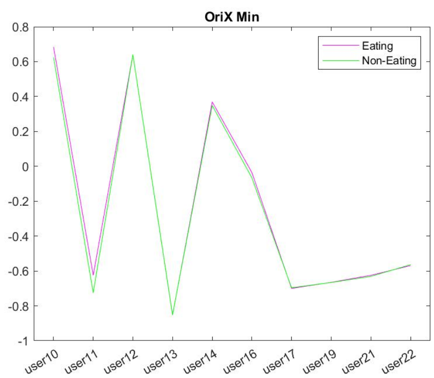


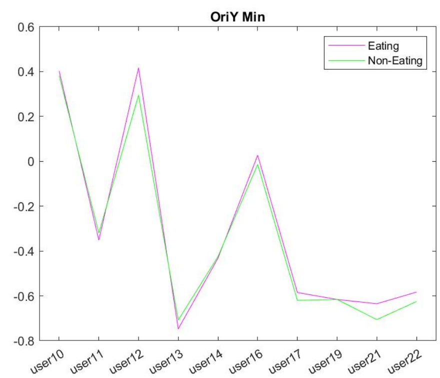

AccZ Min
Eating
Non-Eating
-0.3 -0.4 -0.5 -0.6 -0.7 -0.8 -0.9 -1 -1.1 -1.2 -1.3
user10 user11 user12 user13 user14 user16 user17 user19 user21 user22

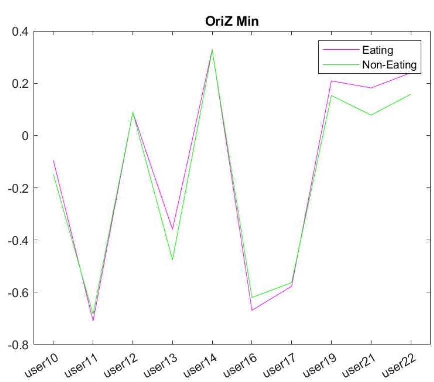

GyroX Min
Eating
Non-Eating
-20 -40 -60 -80 -100 -120 -140 -160 -180
user10 user11 user12 user13 user14 user16 user17 user19 user21 user22

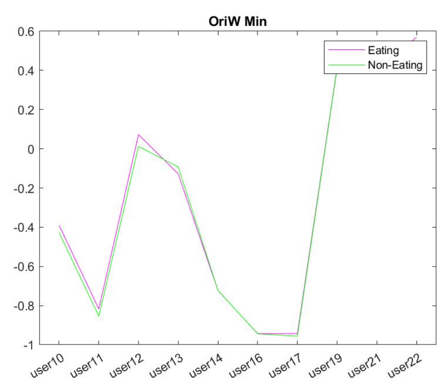

GyroY Min
Eating
Non-Eating
-40 -60 -80 -100 -120 -140 -160 -180
user10 user11 user12 user13 user14 user16 user17 user19 user21 user22


GyroZ Min
Eating
Non-Eating
-30 -40 -50 -60 -70 -80 -90 -100 -110 -120 -130
user10 user11 user12 user13 user14 user16 user17 user19 user21 user22


OriX Min
Eating
Non-Eating
0.8 0.6 0.4 0.2 0 -0.2 -0.4 -0.6 -0.8 -1
user10 user11 user12 user13 user14 user16 user17 user19 user21 user22


OriY Min
Eating
Non-Eating
0.6 0.4 0.2 0 -0.2 -0.4 -0.6 -0.8
user10 user11 user12 user13 user14 user16 user17 user19 user21 user22


OriZ Min
Eating
Non-Eating
0.4 0.2 0 -0.2 -0.4 -0.6 -0.8
user10 user11 user12 user13 user14 user16 user17 user19 user21 user22

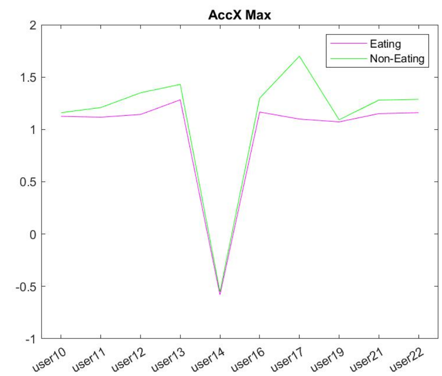

OriW Min
Eating
Non-Eating
0.6 0.4 0.2 0 -0.2 -0.4 -0.6 -0.8 -1
user10 user11 user12 user13 user14 user16 user17 user19 user21 user22

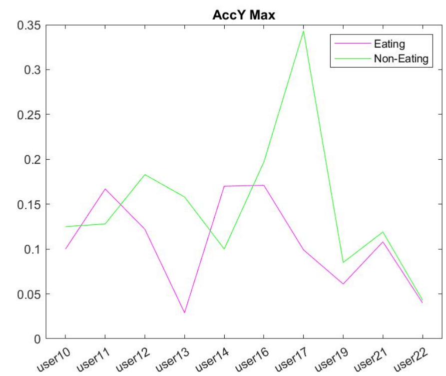
MAXIMUM

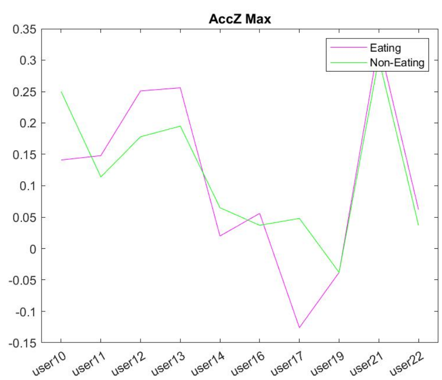
The maximum calculation is simply taking the maximum values per sensor dimension per user. This data set can show what the max value is per dimension and the distribution of that value and how it relates to eating and non-eating data. The distribution of data in terms of the Orientation values I suspect will not vary as much as the other sensor values as eating data has a smaller range overall of values than non-eating data.

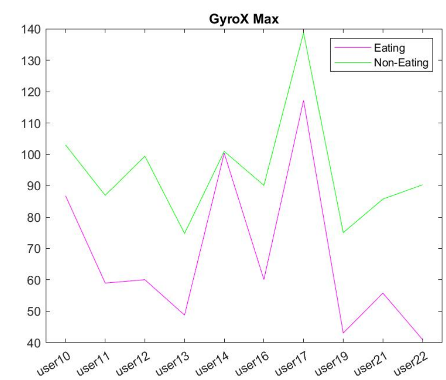
The Max values were extracted using MATLAB and graphed.

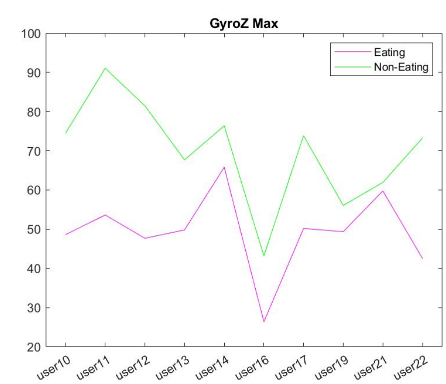
As suspected, the Orientation dimensions did not show a large distribution of maximum values as well as the Acceleration X sensor. The other dimensions showed a variance which can be used to say that typically non-eating actions have higher maximum values.

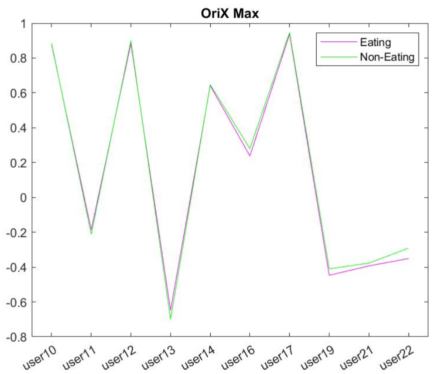


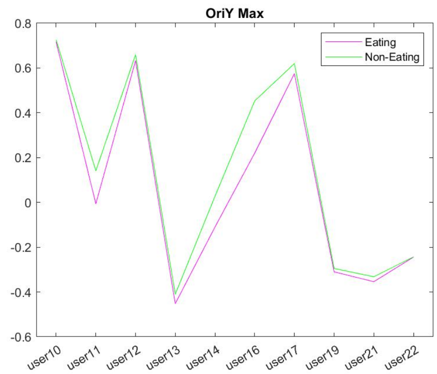


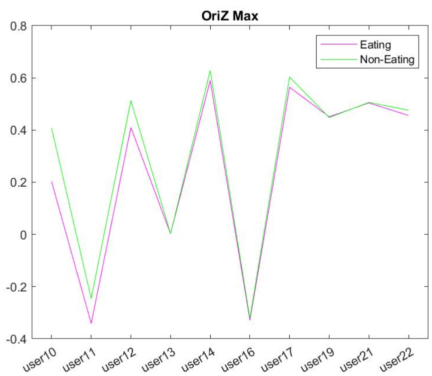


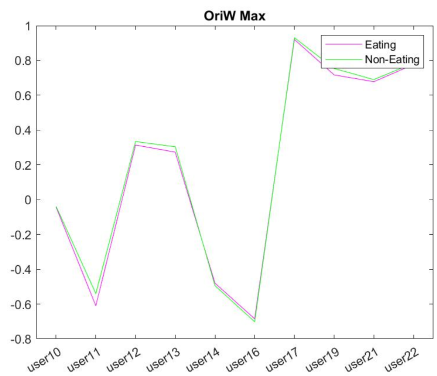










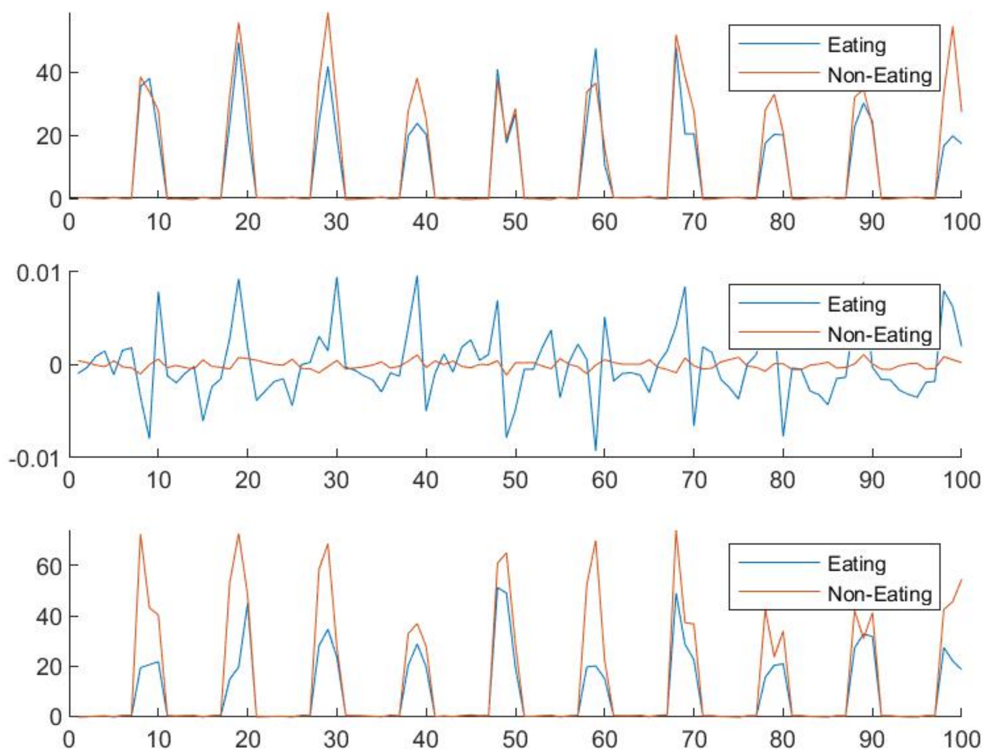


PHASE 3: FEATURE SELECTION

The feature matrix was arranged by appending each of the feature data for Eating and Non-Eating into columns. For example, the first column contained all the max feature data for Eating Actions for all users, the second column contained the mean, etc. The order of features was Max, Mean, Min, Range and Standard Deviation.

The results of PCA are eigenvalues and eigenvectors. The eigenvectors symbolize a vector in a given dimensional space. New feature matrices can be generated for each feature space by taking the Coeff matrix and multiplying the original feature matrix by this matrix.

This graph of the 3 principal components shows some variation in these features, the results might have been different if other feature selection methods would be used.



When looking at the resulting eigenvalues, there were only 5 values, so reduction was not necessary and the spider and subplots were generated using the feature matrix multiplied by the coeff matrix.

1
5.9623e+03
134.1096
3.0949
0.7855
1.6682e-29

1
1.3589e+04
126.0009
8.4942
0.6911
4.4554e-29