Emiel Steerneman - s1499262
Mina Atef Moussa Atia - s1944037

4.1

a) The training set consists of 7352 rows and 561 columns. The test set consists of 2947 rows and 561 columns. Each row represents measurements at a specific point in time, and each column represents a feature.

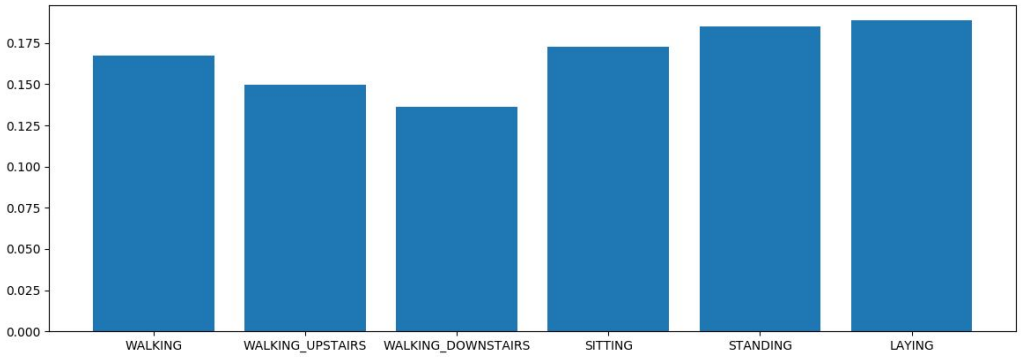
b) The following statistics are calculated by combining both the training data and test data :

Feature	**Mean**	**Median**	**Standard Deviation**
tBodyAcc-mean()-X	0.279	0.279	0.003
tBodyAcc-max()-X	-0.940	-0.941	0.002
tGravityAcc-mean()-Y	-0.147	-0.145	0.005
fBodyAcc-kurtosis()-Y	-0.442	-0.481	-0.372
fBodyGyro-bandsEnergy()-49,64	-1.000	-1.000	0.000

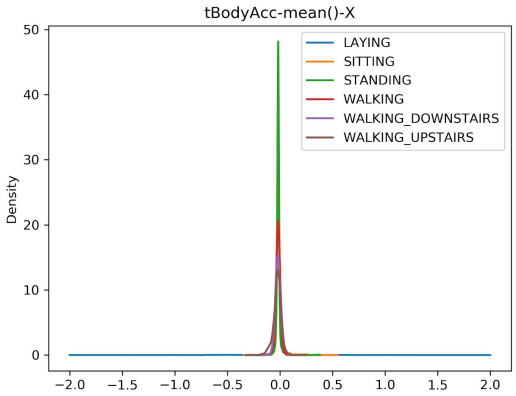
4.2

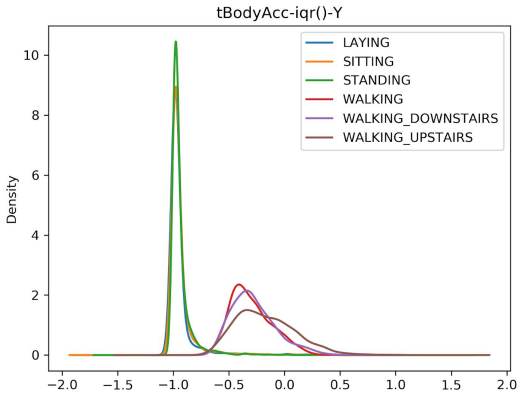
a) The training set consists of 7352 rows and 561 columns. The test set consists of 2947 rows and 561 columns. Each row represents a point in time, and the column represents the label of the activity happening at that point in time.

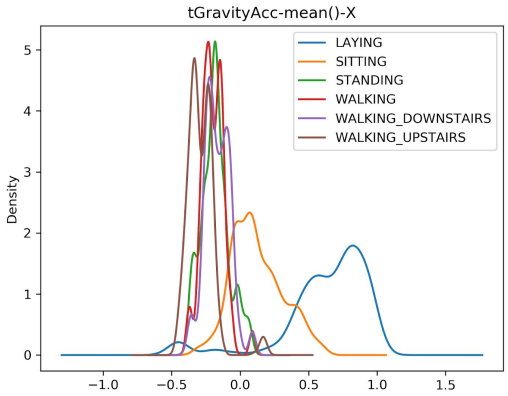
b) The distribution of the labels over te data can be called balanced. Each label represents around 17% of the data. It is important that each label has around the same amount of labels, to prevent a model from focussing on a single label with many more samples. Most models train by reducing some kind of error. When one label is overrepresented, the model can focus on that label to reduce the overall error, at the cost of bad performance on other underrepresented labels.

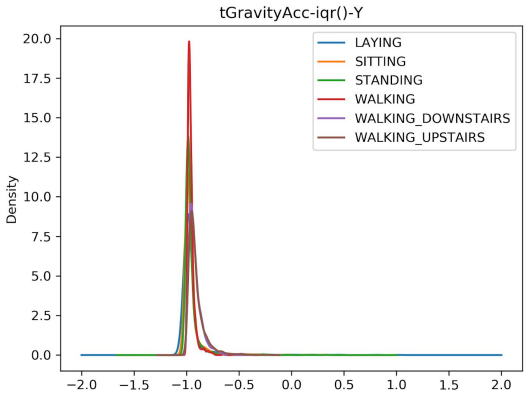


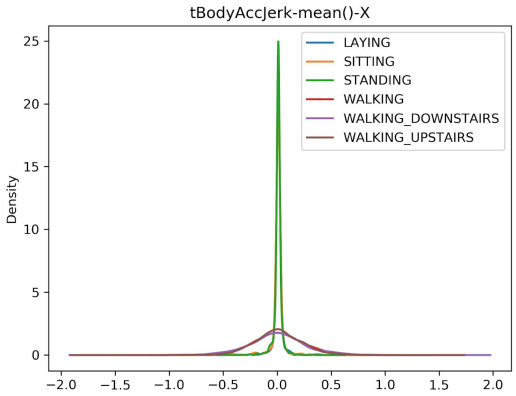
4.3

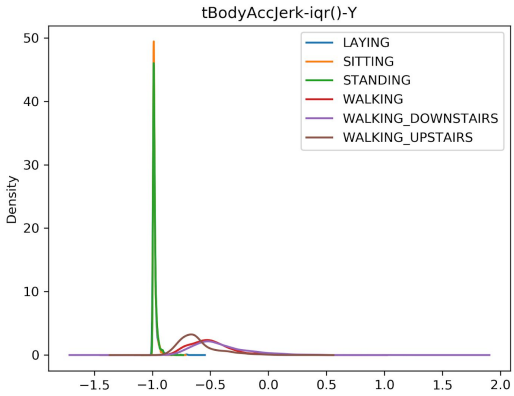

tBodyAcc-mean()-X
Density
LAYING
SITTING
STANDING
WALKING
WALKING_DOWNSTAIRS
WALKING_UPSTAIRS

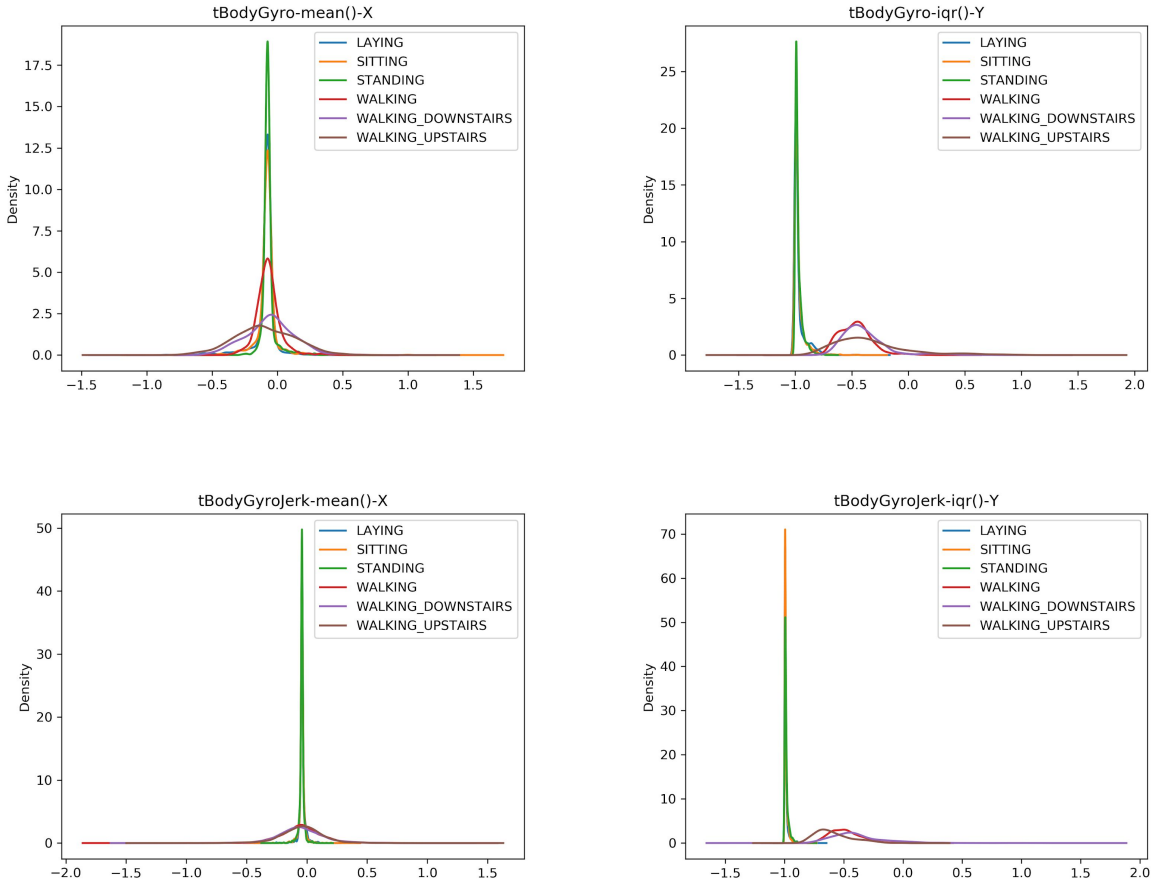

tBodyAcc-iqr()-Y
Density
LAYING
SITTING
STANDING
WALKING
WALKING_DOWNSTAIRS
WALKING_UPSTAIRS


tGravityAcc-mean()-X
Density
LAYING
SITTING
STANDING
WALKING
WALKING_DOWNSTAIRS
WALKING_UPSTAIRS


tGravityAcc-iqr()-Y
Density
LAYING
SITTING
STANDING
WALKING
WALKING_DOWNSTAIRS
WALKING_UPSTAIRS


tBodyAccJerk-mean()-X
Density
LAYING
SITTING
STANDING
WALKING
WALKING_DOWNSTAIRS
WALKING_UPSTAIRS


tBodyAccJerk-iqr()-Y
Density
LAYING
SITTING
STANDING
WALKING
WALKING_DOWNSTAIRS
WALKING_UPSTAIRS



4.4

a) Combining both the training set and the test set, the dataset consists of 10299 rows and 128 columns. Each row is described as follows:

"The sensor signals (accelerometer and gyroscope) were pre-processed by applying noise filters and then sampled in fixed-width sliding windows of 2.56 sec and 50% overlap (128 readings/window). The sensor acceleration signal, which has gravitational and body motion components, was separated using a Butterworth low-pass filter into body acceleration and gravity. The gravitational force is assumed to have only low frequency components, therefore a filter with 0.3 Hz cutoff frequency was used."

b) When windows and thus datapoints overlap, it can occur that a datapoint belongs to two different classes. The consequence of this is that the training model will be less accurate, because using these points for training diminishes the difference between two classes. The most obvious solution would be to remove these datapoints. However, this would mean that the original signal cannot be fully reconstructed. Another solution would be to assign the class only to the first half of the window. This would prevent overlap and loss of data. Only the last half of

the last window would go unlabeled. This can be easily done in Pandas, simply by dropping the last 64 columns.

4.5

a) Time domain features

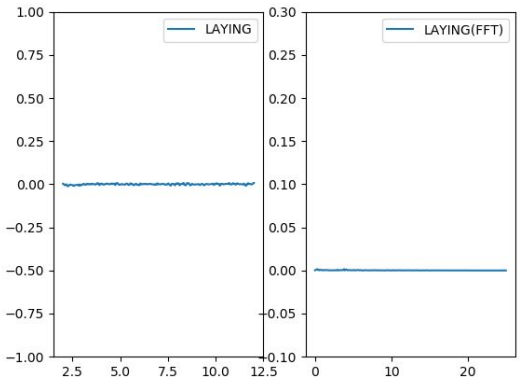
Range: [-1.1890, 1.2386]
Mean: -0.0011
stddev: 0.1922
kurtosis: 4.5849

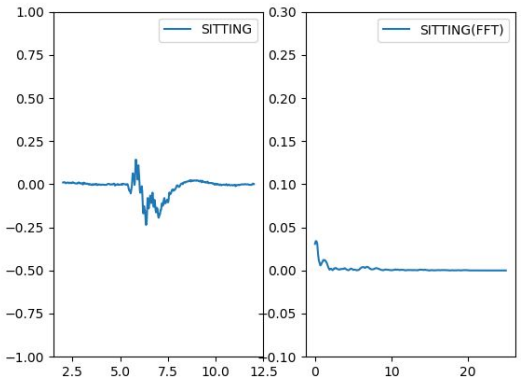
b)

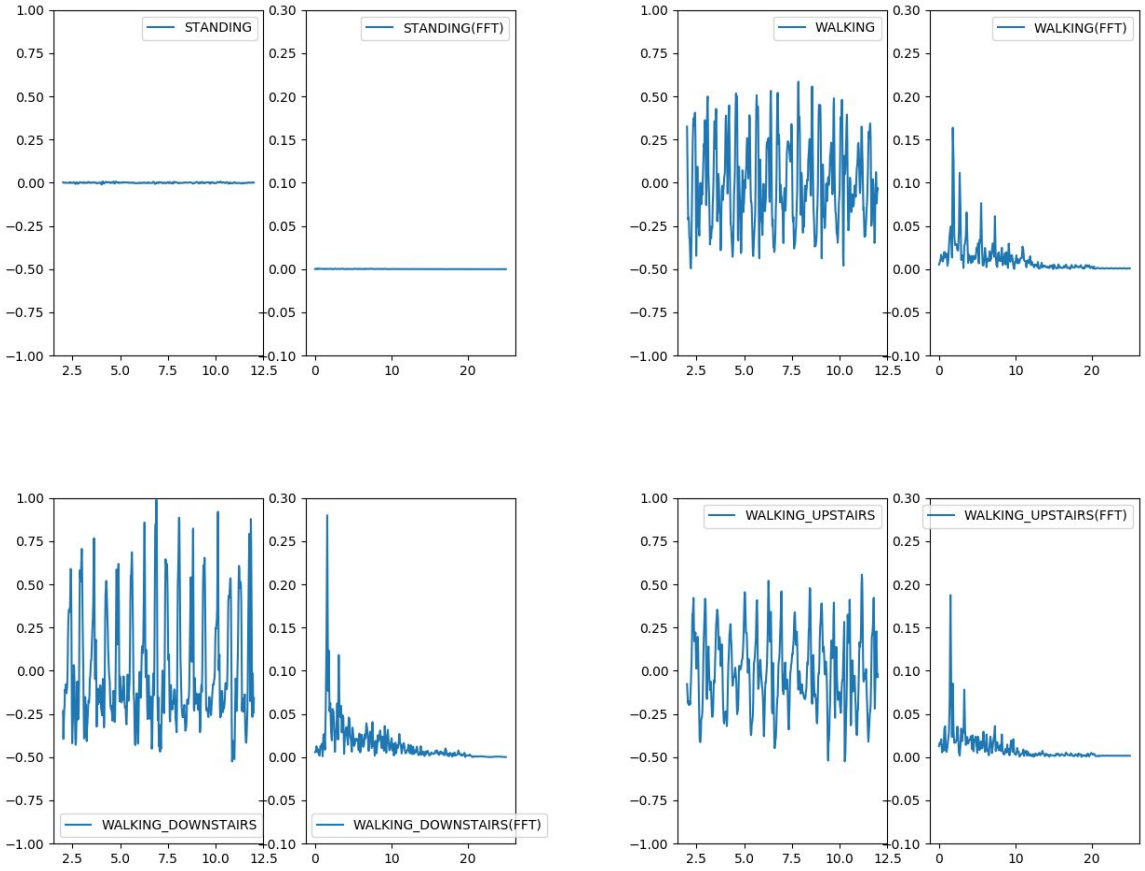
	LAYING	SITTING	STANDING	WALKING	WALKING_DOWNSTAIRS	WALKING_UPSTAIRS
range	[-1.0349, 0.3277]	[-0.3581, 0.3673]	[-0.3720, 0.2794]	[-0.8738, 1.0445]	[-1.1890, 1.2386]	[-0.9054, 1.0106]
mean	-0.0033	-0.0012	0.0007	-0.0005	0.0025	-0.0045
stddev	0.0447	0.0177	0.0128	0.2281	0.3642	0.2539
kurtosis	102.8159	84.3442	65.3205	-0.1022	-0.0733	-0.0067

Looking at the table above, there are a few features that, when combined, could help classify a signal. Kurtosis is a good discriminator between walking and not walking. The same goes for range. Kurtosis could possibly be used as the only feature to distinguish signals, but it would not be accurate, especially when deciding between WALKING and WALKING_DOWNSTAIRS. Other than this, there does not seem another useful combination of features. Overall, these features would probably make for a poor classifier.

4.6



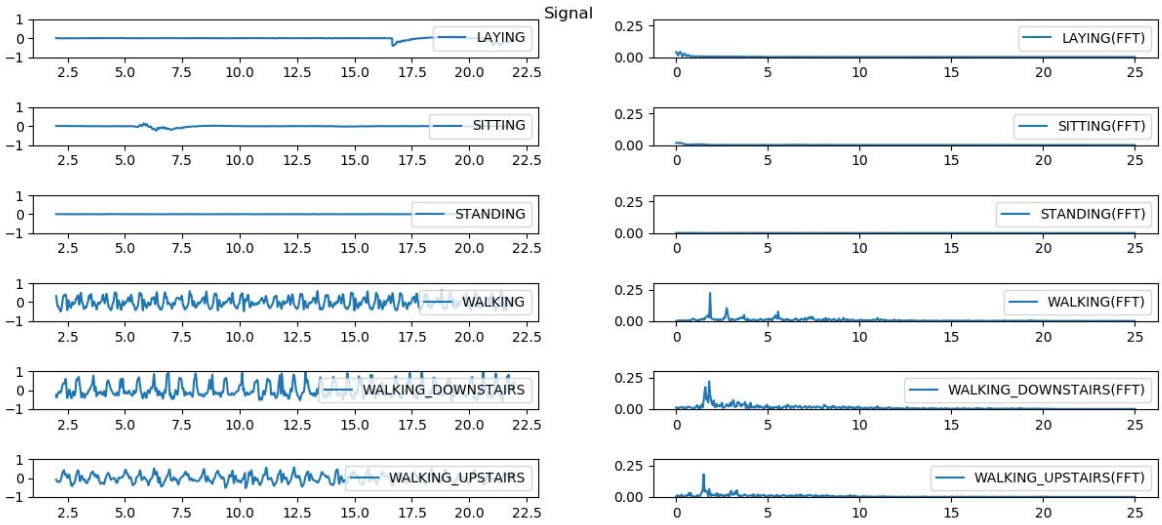


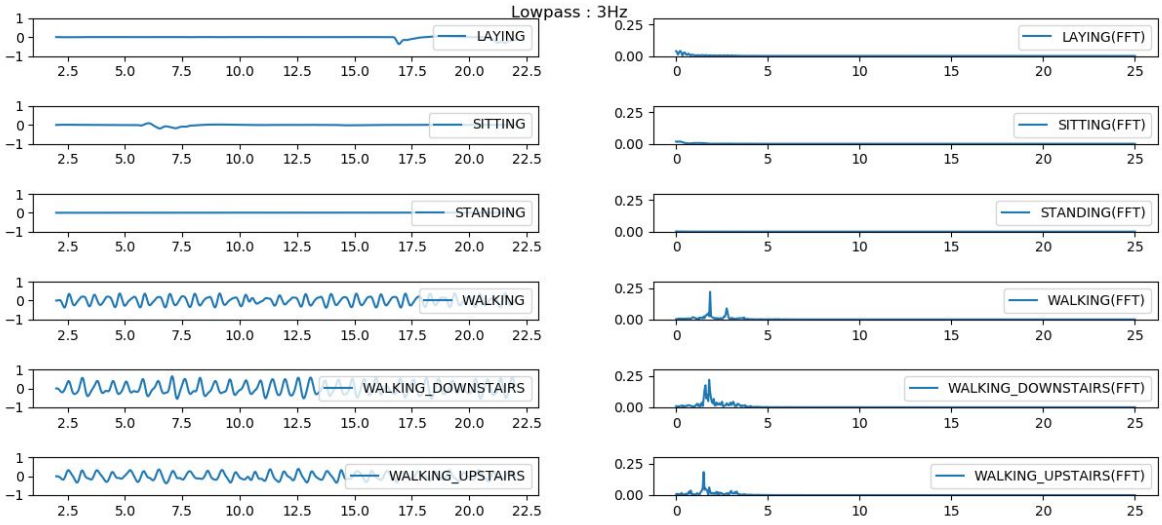


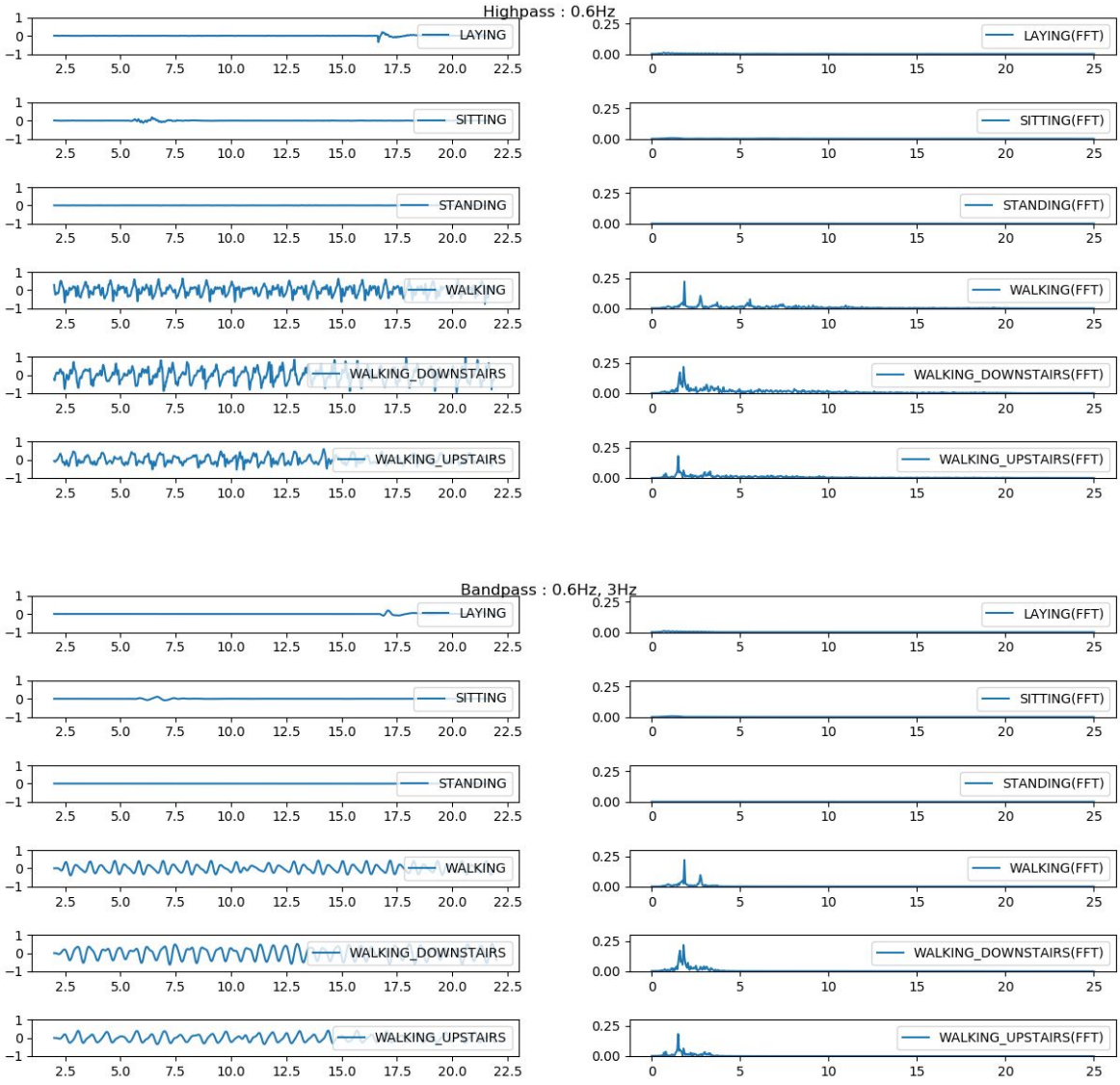
b) Looking at the spectra, it does not seem possible to differentiate between LAYING, SITTING, and STANDING, and between WALKING, WALKING_DOWNSTAIRS, and WALKING_UPSTAIRS.

4.7

a) The frequencies that are expected in all signals are the ones caused by breathing, which should be around 0.25Hz to 0.33Hz. The frequencies expected in the walking activities are around 1Hz to 1.5Hz, which is the average walking frequency. The breathing frequency is probably insignificant, because it really depends on each person's health. It might however give some distinction between moving and not moving. A good lower bound might therefore be 0.6Hz, somewhere between breathing and walking. Walking 4 miles per hour takes around 150 steps per minute, so a good higher bound might be around 3Hz.







The best filter is the bandpass filter.

4.8

Summary of confusion matrix of DTW on feature data

	Precision	Recall	f1-score	Support
WALKING	0.96	0.80	0.87	60
WALKING UPSTAIRS	0.85	0.80	0.83	51
WALKING DOWNSTAIRS	0.68	0.97	0.80	31
SITTING	0.78	0.78	0.78	51
STANDING	0.84	0.76	0.80	55
LAYING	0.90	1.00	0.95	47
avg / total	0.85	0.84	0.84	295

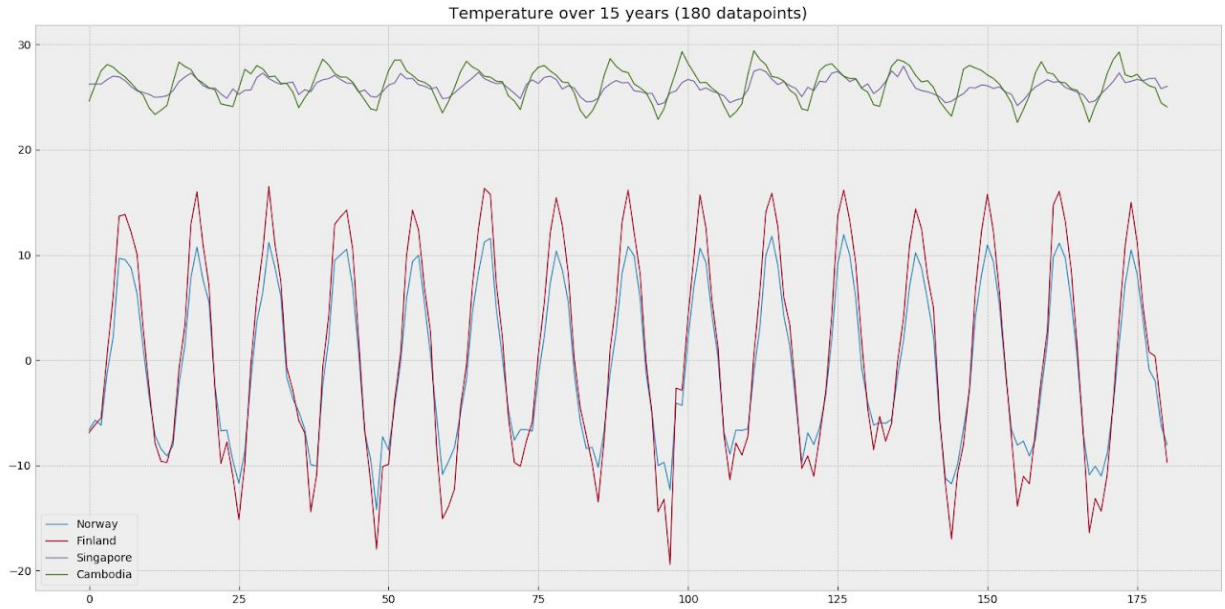
Summary of confusion matrix of DTW on raw data (867890 samples in 4.5 hours)

	Precision	Recall	f1-score	Support
WALKING	0.95	0.80	0.87	119
WALKING UPSTAIRS	0.88	0.85	0.86	99
WALKING DOWNSTAIRS	0.69	0.92	0.79	64
SITTING	0.77	0.86	0.81	88
STANDING	0.93	0.80	0.86	122
LAYING	0.93	1.00	0.97	98
avg / total	0.88	0.86	0.86	590

The results between the two confusion matrices are not that significant. All cells with a difference higher than 0.05 are marked, of which there are only four. Overall, DTW on raw datas performs slightly better than DTW on feature data.

4.9

a)



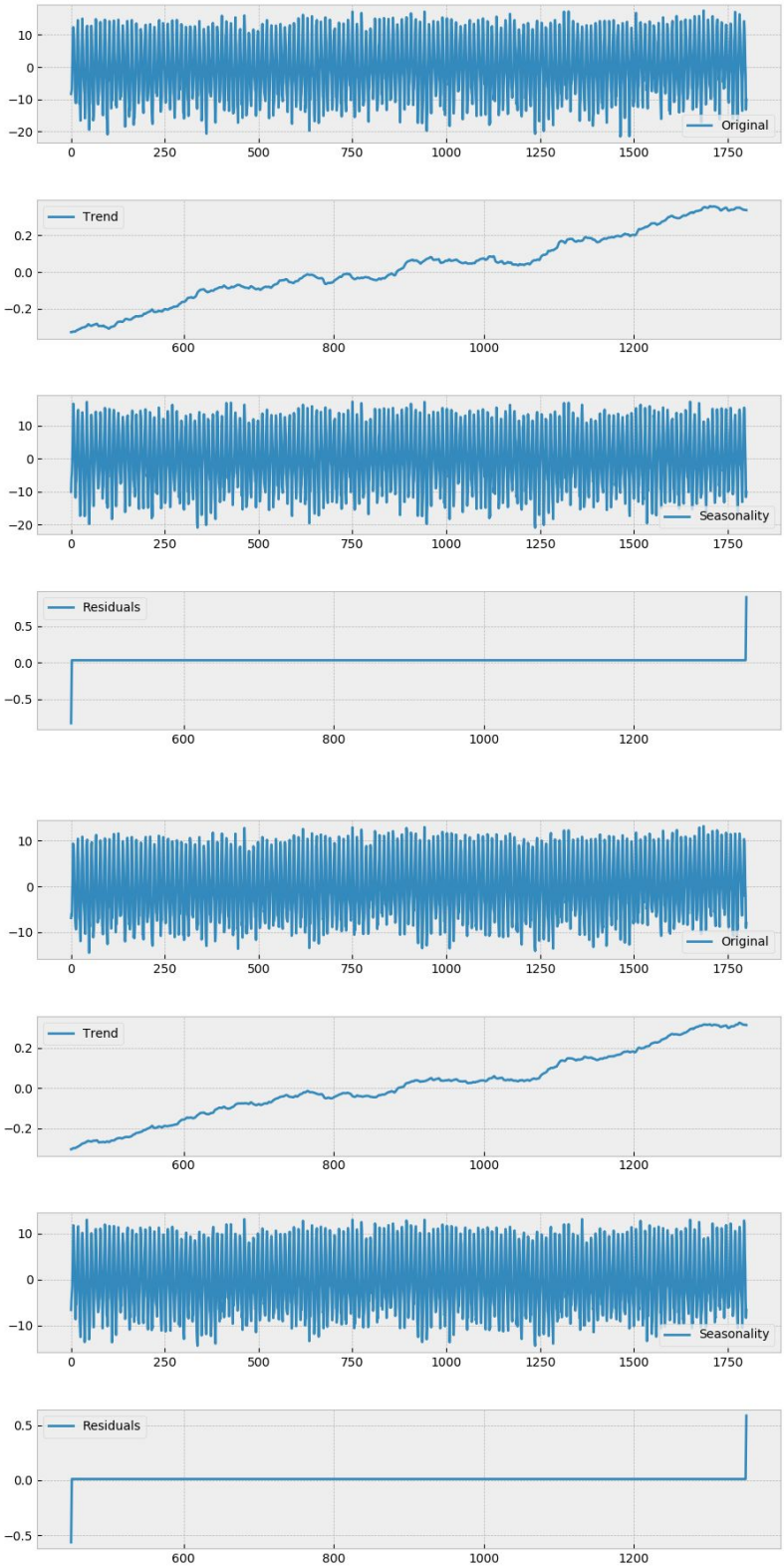
DISTANCE	Norway	Finland	Singapore	Cambodia
Norway	0	4028	47371	47617
Finland	4028	0	45262	45509
Singapore	47371	45262	0	1279
Cambodia	47617	45509	1279	0

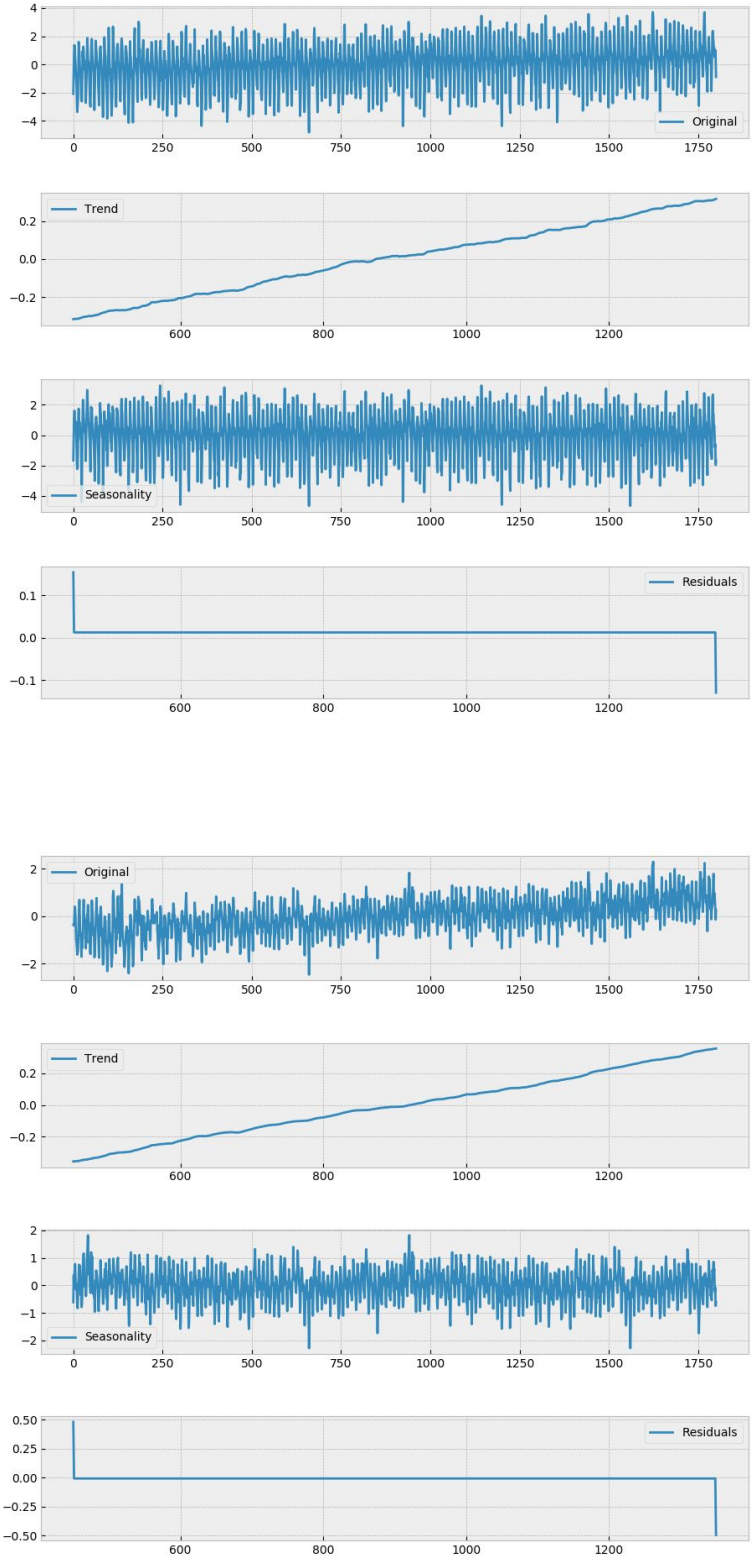
The plot shows that Norway and Finland have similar temperatures and that Singapore and Cambodia have similar temperatures. The temperature differences between the Scandinavian lands and the Southeast Asian lands are relatively large. This is supported by the DTW table. The distance between Finland-Norway and Singapore-Cambodia is relatively low, compared to the distance between e.g. Norway-Singapore, and Finland-Cambodia.

b) Dickey-Fuller Test:
P-value for Norway 0.0162%
P-value for Finland 0.0001%
P-value for Singapore 3.2006%
P-value for Cambodia 0.1719%

Taking 5% as the critical value, the P-values for all the signals are low enough to be considered stationary.

c) Figures of Norway, Finland, Singapore, and Cambodia. We are unsure what to do with the result, since there seems to be no signal left after removing the trend and seasonality.




Original
Trend
Seasonality
Residuals
Original
Trend
Seasonality
Residuals