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.

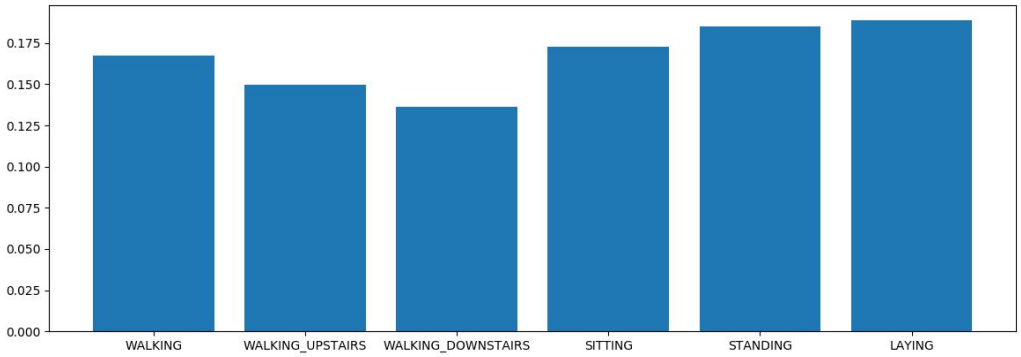
b) The following statistics are calculated by combining both the training data and test data :

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

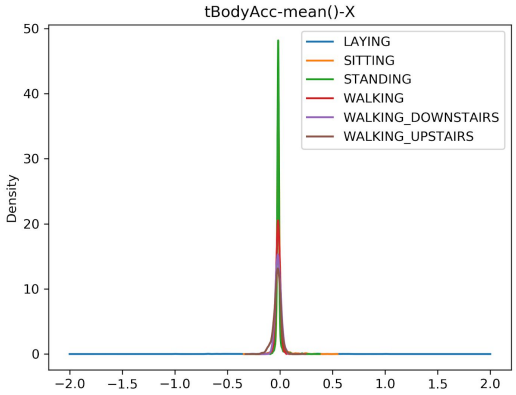
4.2

a) The training labels consist of 7352 rows and a single column. The test labels consist of 2947 rows and a single column. Each row represents a point in time, and the column represents the label of the activity happening at that point in time.

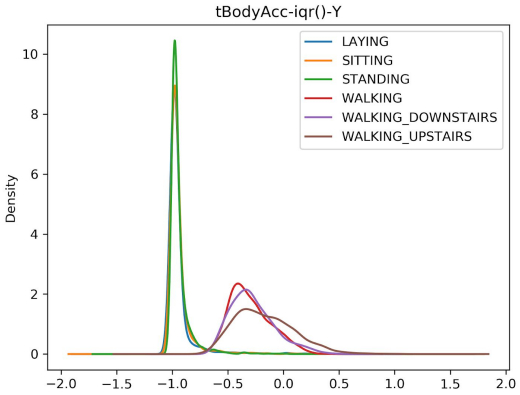
b) The distribution of the labels over the data can be called balanced. Each label represents around 17% of the data. It is important that each label has around the same amount of labels, to prevent a model from focussing on a single label with many more samples. Most models train by reducing some kind 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.



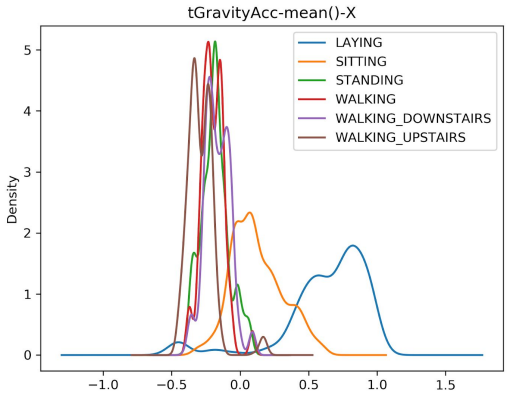
4.3



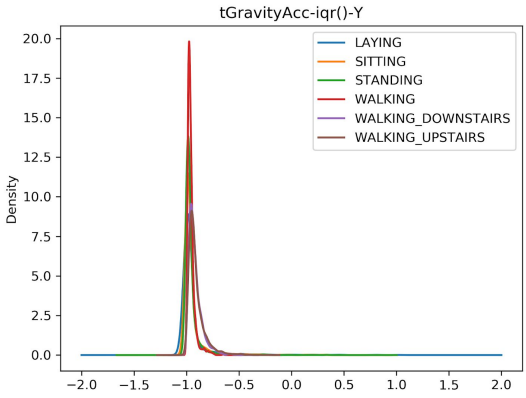
The means of the x-accelerations of all activities are placed around zero. This makes sense, because there should be almost no x-acceleration when doing any of these activities. This feature is unusable to discriminate between activities.



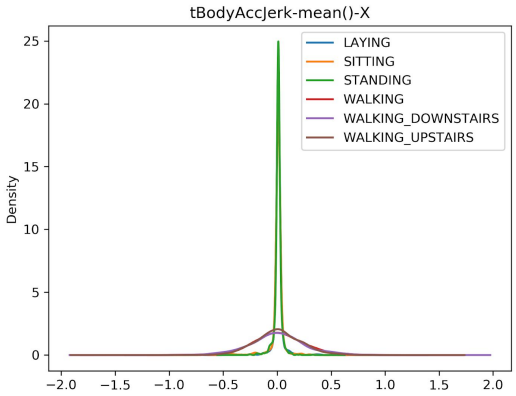
The interquartile range of the activities show a clear difference between walking and not walking. The IQR of the non-walking activities is clearly centered around -1, while the IQR of the walking activities is more spread out between -0.5 and 0. This feature could be used to differentiate between walking and not walking. We are unsure how the interquartile range can be negative.



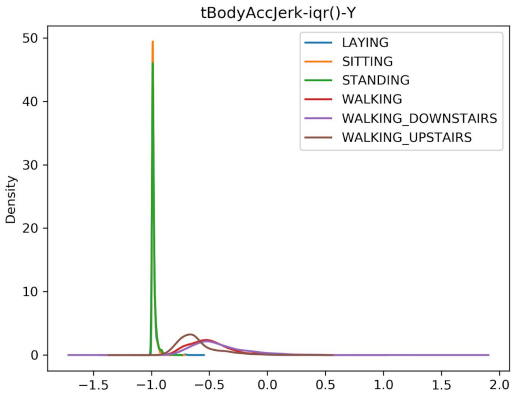
Most of these distributions are multimodal, quite steep, and centered around -0.25. The two exceptions are sitting and laying. This feature could be used to differentiate between sitting, walking, and other activities. The reason for the difference of these two activities might be because of how the angles of the sensors change when sitting or laying down. Gravity will show up as a constant acceleration, which can be seen on the x-axis when rotation the sensors.



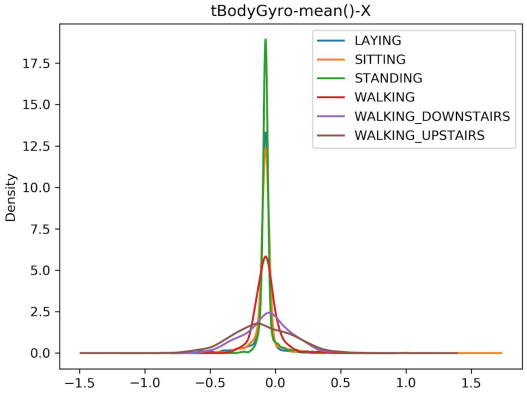
All distributions are centered around -1, and very slightly skewed to the right. This feature is of no use for distinguishing between any activity.



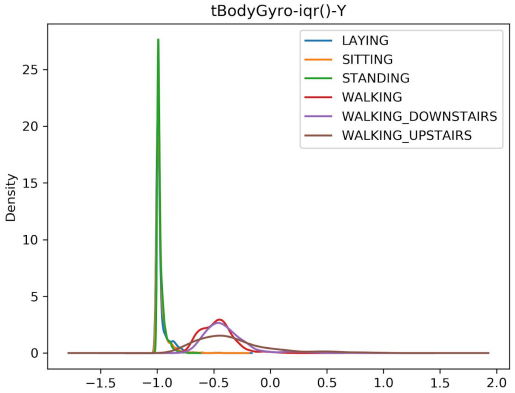
All distributions are unimodal and centered around 0. The non-walking activities are very steep. This makes sense, because the jerk is the second derivative of the velocity, and the velocity is 0 when doing any of these activities. The jerk is more spread out for the walking activities, because velocity is constantly changing a little when walking. This feature could be used to distinguish between walking and not walking.



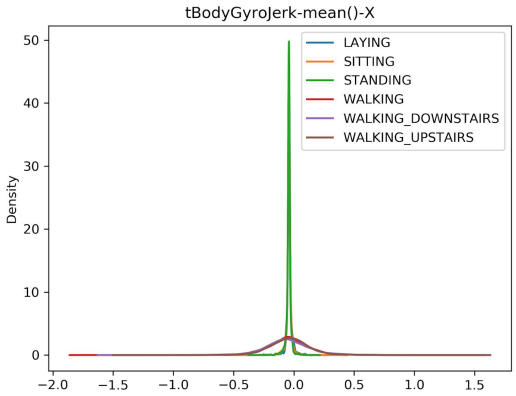
The non-walking distributions are very steeply centered around -1.0. The walking distributions are more spread out between -1.0 and 0. Walking upstairs is more skewed to the right. This feature could be used to distinguish between walking and non-walking, and possibly to distinguish walking upstairs.



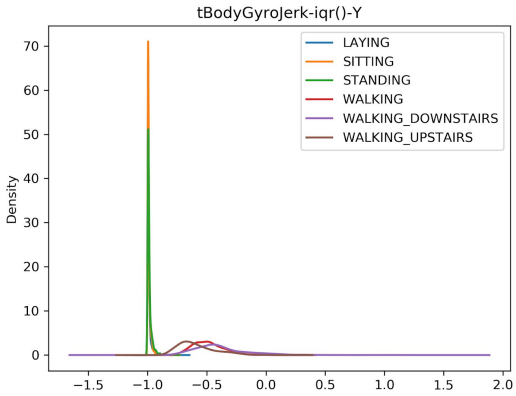
All distributions are unimodal and centered around -0.1. The distributions differ significantly in their spread. The non-walking distributions are all very steep, especially standing. The Walking activity is quite steep, but not as steep as the non-walking activities. The walking upstairs activity and walking downstairs activity are the flattest. This feature could be used to distinguish between non-walking and walking activities, and possibly to distinguish walking and standing.



These distributions are very similar to the distributions of tBodyAccJerk-iqr()-Y. The only difference is that the walking upstairs distribution is more flat, and centered around -0.5. This feature could be used to distinguish between walking and non-walking, and possibly to distinguish walking upstairs.



All distributions are unimodal and centered around 0. The non-walking activities are very steep. This makes sense, because the jerk is the second derivative of the velocity, and the velocity is 0 when doing any of these activities. The jerk is more spread out for the walking activities, because velocity is constantly changing a little when walking. This feature could be used to distinguish between walking and not walking.



These distributions are very similar to the distributions of tBodyAccJerk-iqr()-Y and tBodyGyro-iqr()-Y. The difference here is that the walking upstairs activity doesn't really stand out anymore from the other walking activities. This feature could be used to distinguish between walking and non-walking.

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:

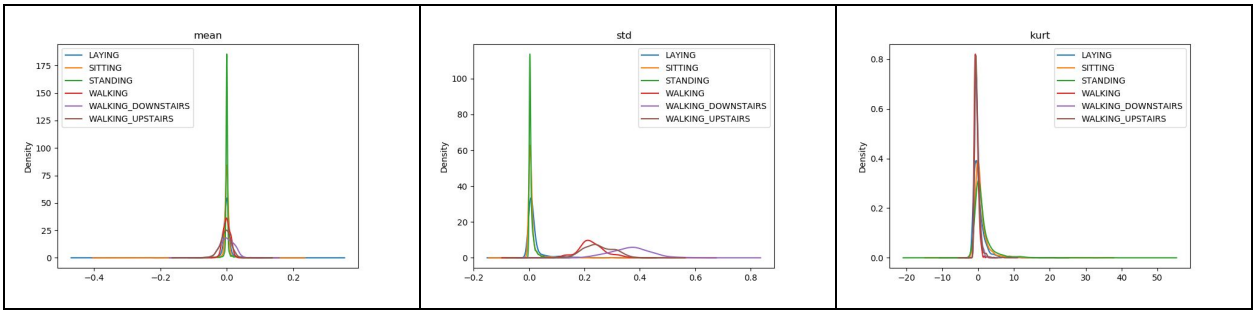
"The sensor signals (accelerometer and gyroscope) were pre-processed by applying noise filters and then sampled in fixed-width sliding windows of 2.56 sec and 50% overlap (128 readings/window). The sensor acceleration signal, which has gravitational and body motion components, was separated using a Butterworth 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

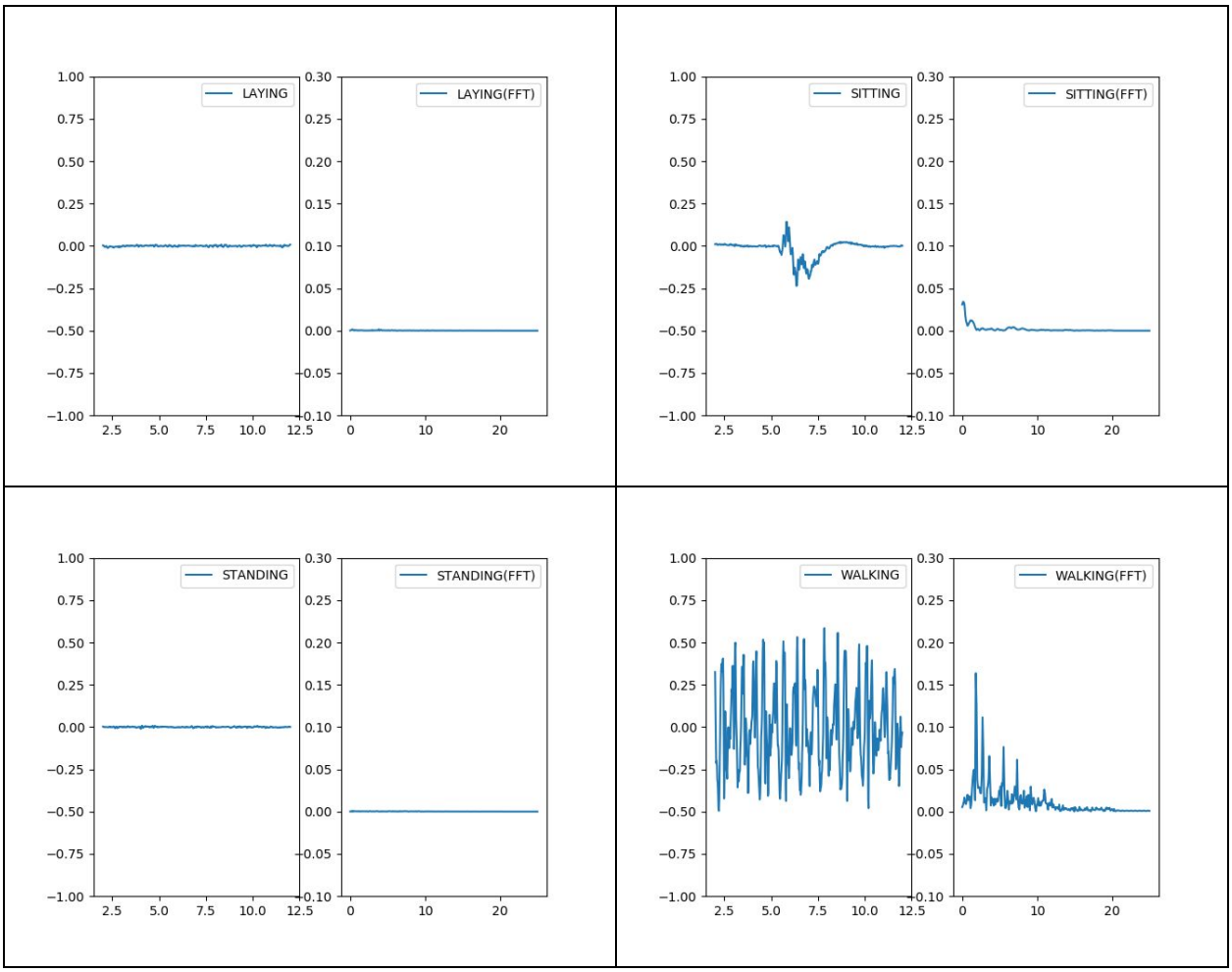
a) Features chosen : mean, standard deviation, kurtosis

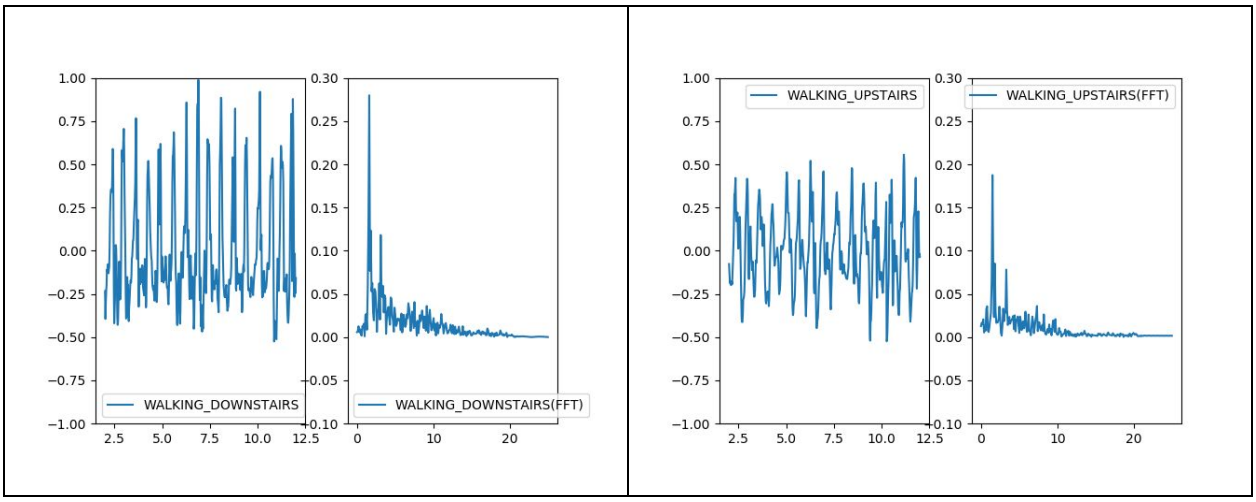
b)



Looking at the images 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 the standard deviation. The mean could be used to discriminate between sitting and standing, and the other four activities. The standard deviation might even be used to discriminate all six activities, but it might have trouble distinguishing walking and walking upstairs.

4.6

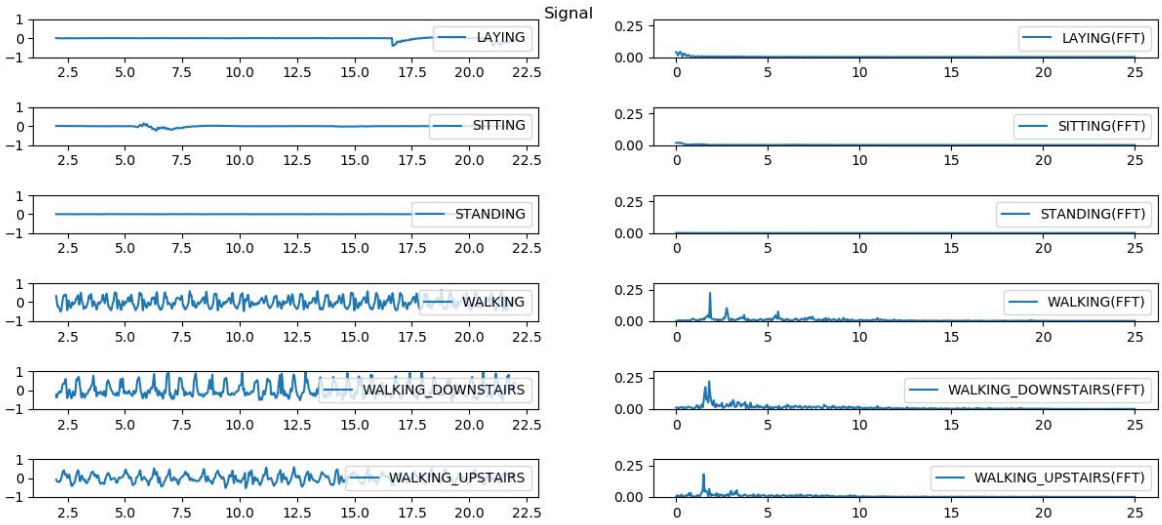


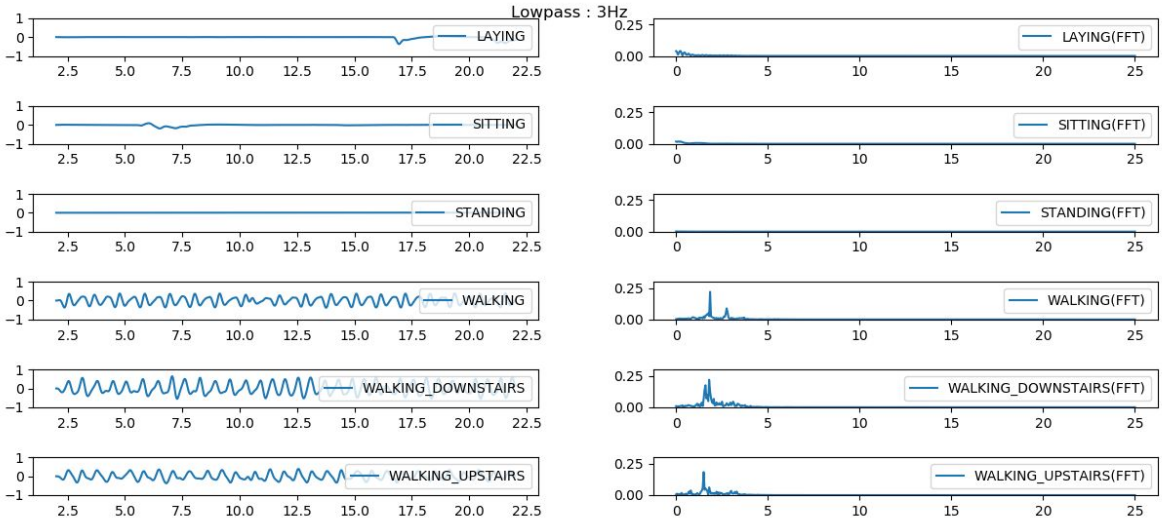


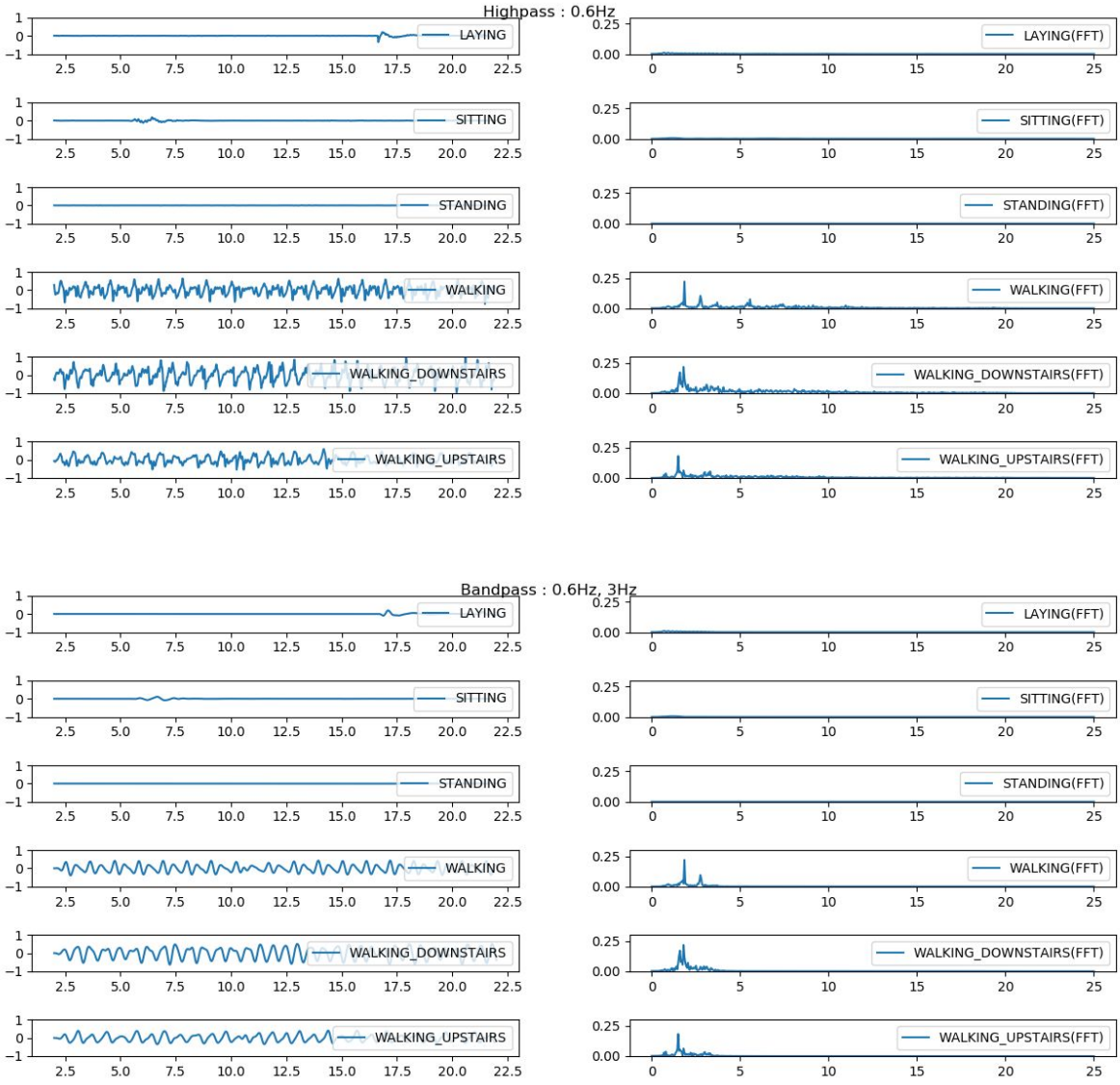
b) Looking at the spectra, it does not seem possible to differentiate between laying, sitting, standing or between walking, walking downstairs, walking upstairs. The signals can however be used to discriminate between these two groups. For laying, sitting, and standing, the signal is basically zero, except for some small bumps in sitting which can probably be chalked up to noise or random movement, looking at the signal. Looking at walking, walking downstairs and walking upstairs, all signals have a strong peak around 2Hz, another at 4Hz and a lot of noise, slowly declining. Walking as a few more peeks than walking downstairs and walking upstairs, but I doubt it will be enough to create a good classifier.

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 would be a bandpass filter because it filters both the lower breathing frequency and any noise higher than the walking frequency.

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

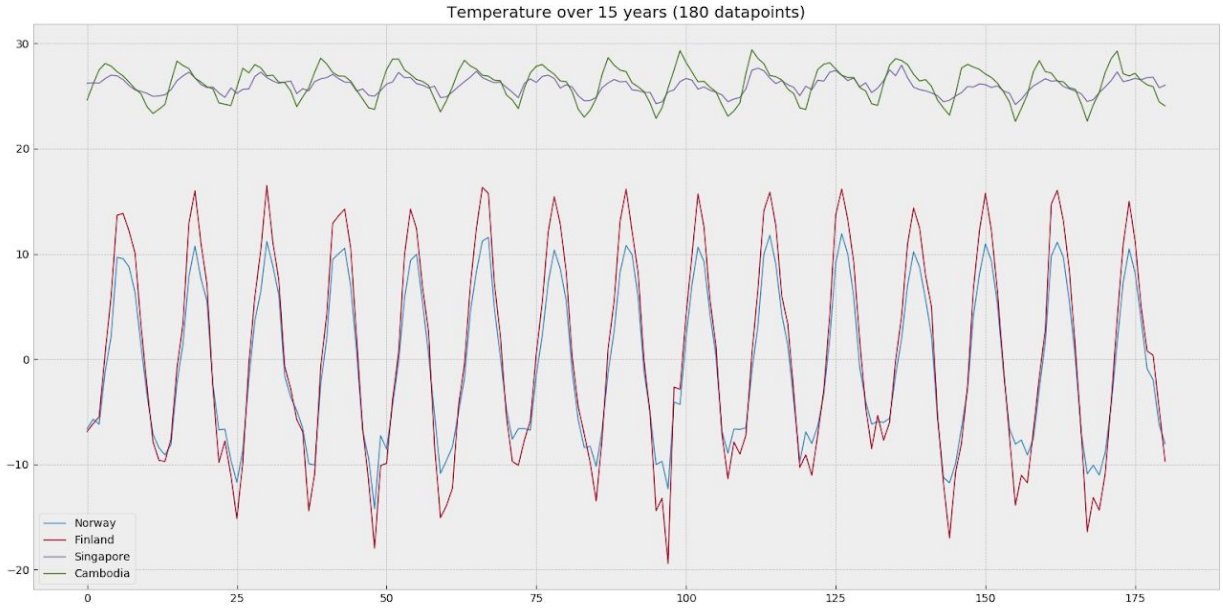
Summary of confusion matrix of DTW on raw data (867890 samples in 4.5 hours)

	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 matrices are not that significant. All cells with a difference higher than 0.05 are marked, of which there are only five. Overall, DTW on raw data performs slightly better than DTW on feature data. I am not sure why DTW on raw data performs better. Looking at the very slight differences between all cells, I would chalk it up to randomness. Another dataset might favour the DTW on feature data. I am not sure why it has the largest influence on standing, but once again I think this is just randomness. I expected the DTW on feature data to perform better because it seems to hold much more information, but in any practical application these two classifiers would perform equally well.

4.9

a)



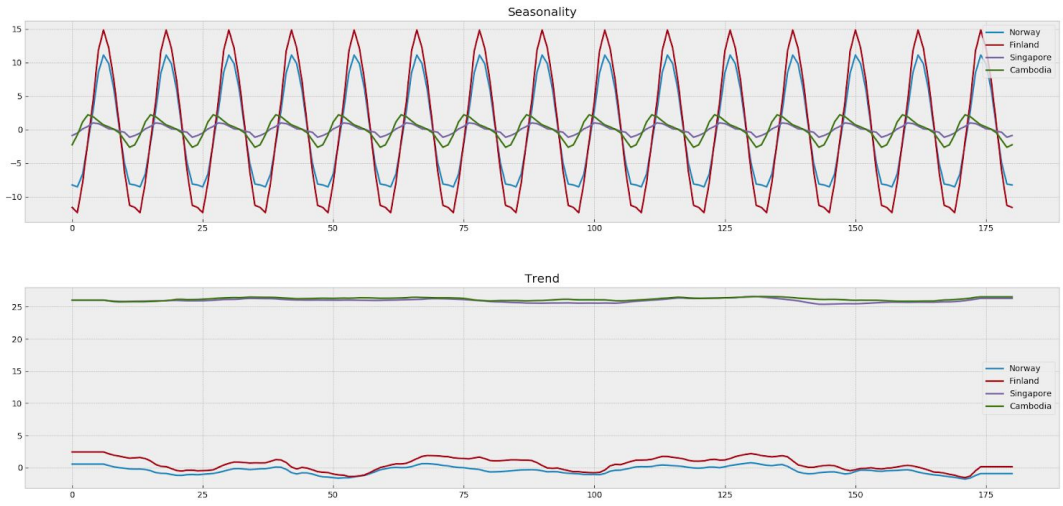
Distance Time Warp table over 150 years

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

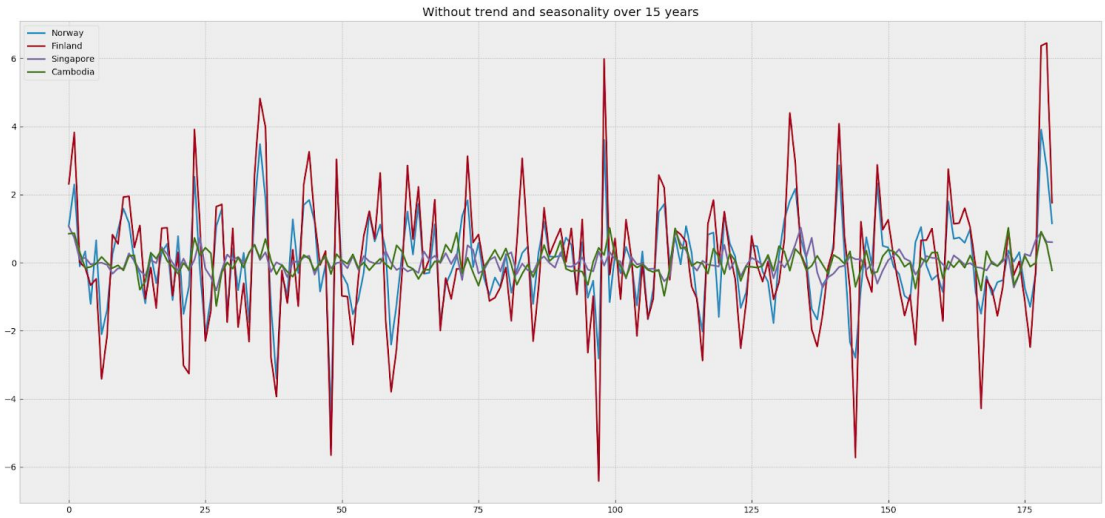
The plot shows that Norway and Finland have similar temperatures and that Singapore and Cambodia have similar temperatures. The temperature differences between the Scandinavian lands and the Southeast Asian lands are relatively large. This is supported by the DTW table. The distance between Finland-Norway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



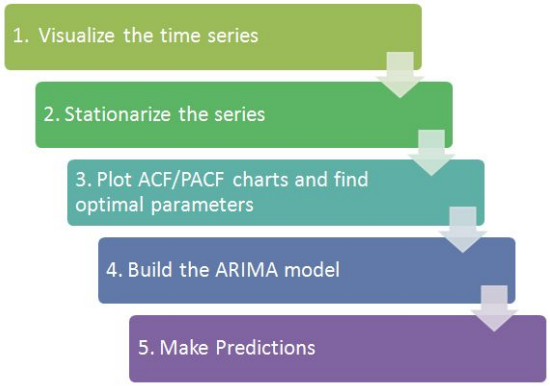
Distance Time Warp table after eliminating seasonality and trend over 150 years

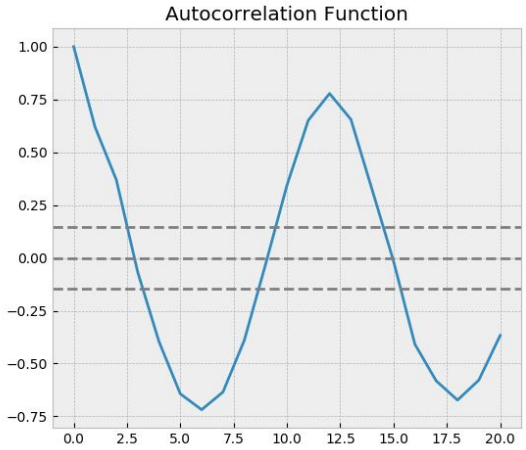
DISTANCE	**Norway**	**Finland**	**Singapore**	**Cambodia**
Norway	0	1443	1656	1460
Finland	1443	0	2570	2337
Singapore	1656	2570	0	423
Cambodia	1460	2337	423	0

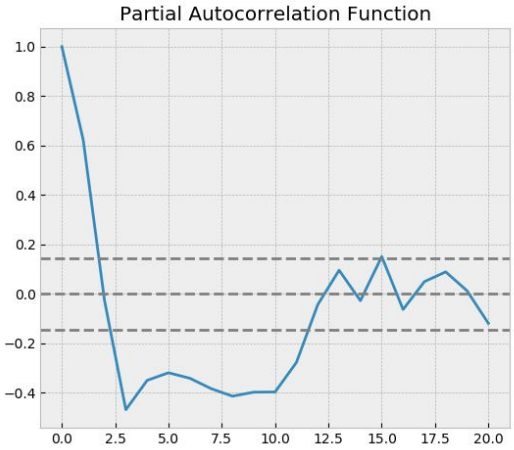
Looking at this new table, removing trend and seasonality made a huge impact on the distances. In the first table Singapore and Cambodia were close together (1279), and Norway and Finland were close together (4028). These two groups were separated from each other by around 46000. Now, Norway is equally close to all three other countries. The smallest distance is between Singapore and Cambodia (423), and the largest distance is between Finland and Singapore (2570).

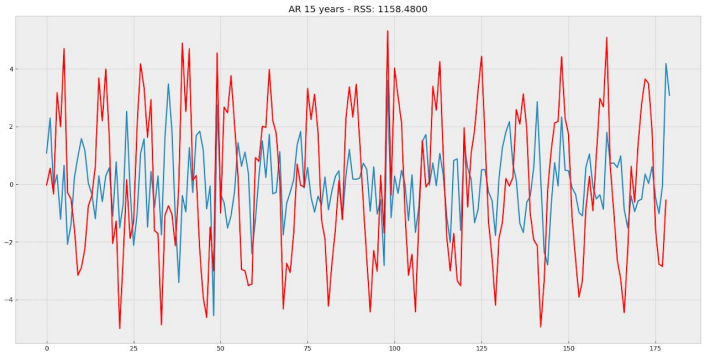
d)

According to [11], this image should describe my approach. The data has been visualized, and it has been stationarized by removing the trend and seasonality. The ACF and PACF plots can be found in the image below. According to the plots, the parameters should be p=2, q=3. d=1 according to [19]. The results of the AR and MA models can be found below

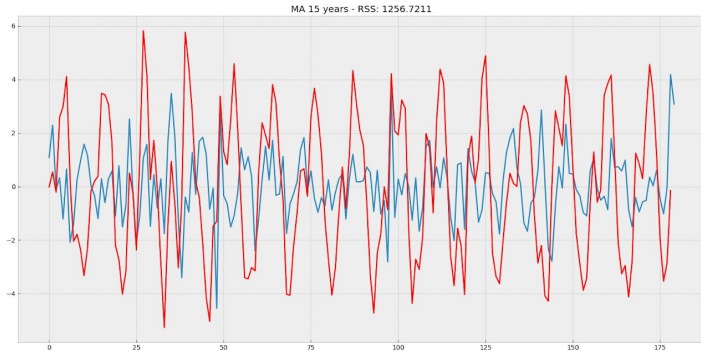






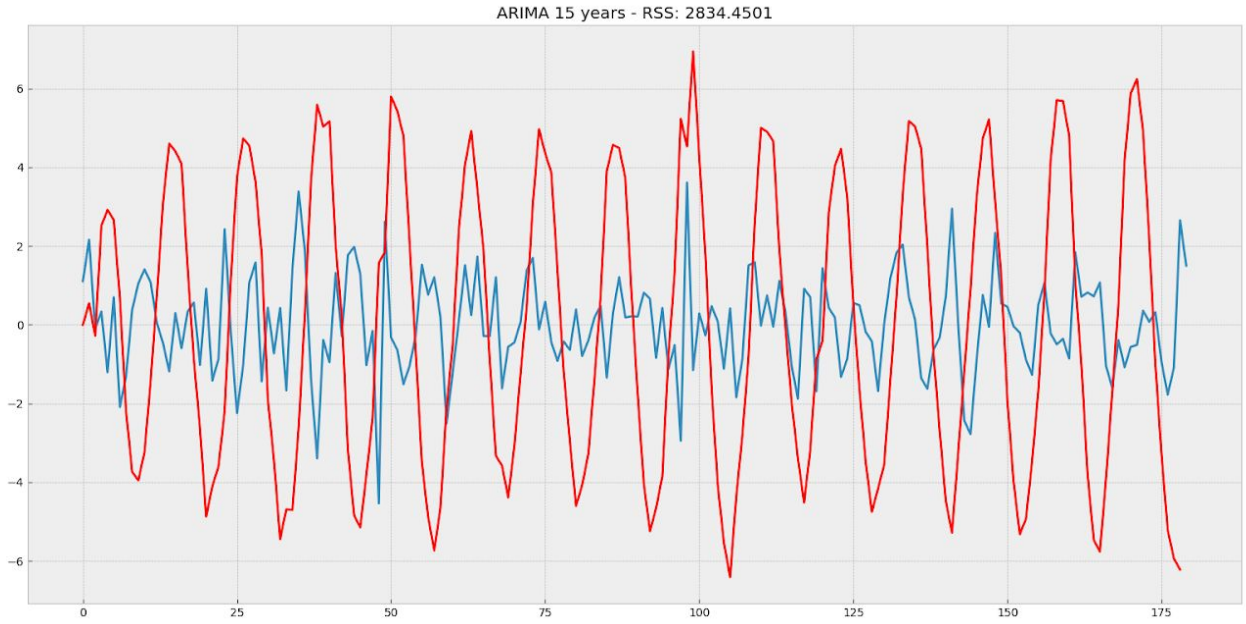


Result of the AR model

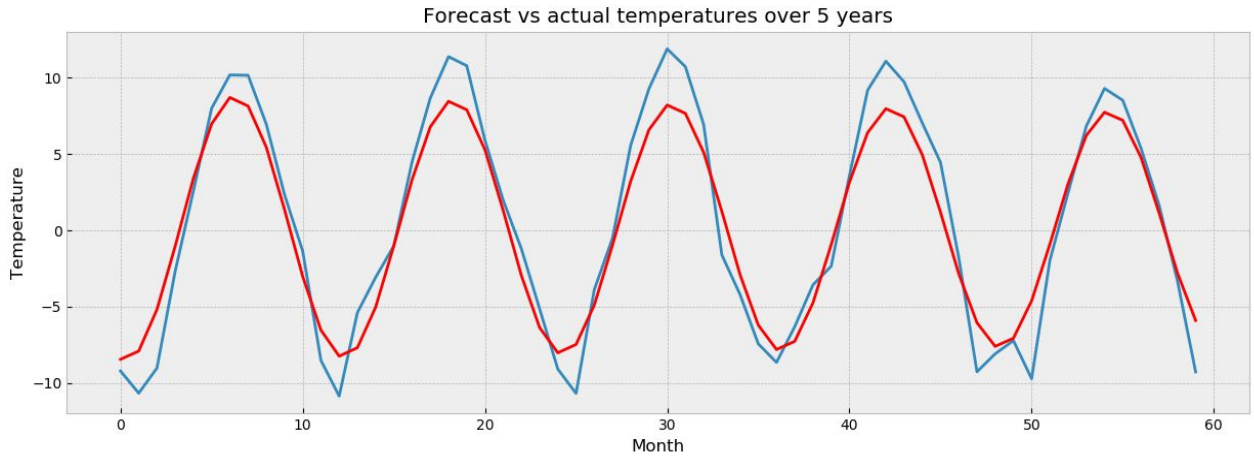


Result of the MA model

I am not really impressed by either the AR or MA model. Both seem to correctly predict 15 years of peaks but they are not as accurate as I had hoped. Unfortunately, the ARIMA model could not be tested using the parameters found because of the following error: "*The computed initial MA coefficients are not invertible You should induce invertibility, choose a different model order, or you can pass your own start_params.*" Taking p=3 and q=2 does work however, and gives the results below. These results are shown below. This model looks better than the results from AR and MA.



e)



The results above is five years of actual data in blue, and five years of predicted data in red. The ARIMA model with p=3 q=2 d=1 has been used to create this forecast. I'd say that this prediction is very accurate. The peaks should have been a little stronger however.