# Assignment 1

DATA MINING

CSE 572: Spring 2020

**Submitted by**

**Akshay Kumar(akuma216@asu.edu)**

# 1. Assignment Phase 1: Feature Extraction

As a part of assignment 1 we are given CGMData and CGMTime file to extract 4 features that are useful for the prediction process. I have started by looking into 5 different types of feature extraction techniques that can be useful for this time series data.

1. Fast Fourier Transform
2. Discrete Wavelet Transform
3. Velocity
4. Moving Average
5. Expanding Window

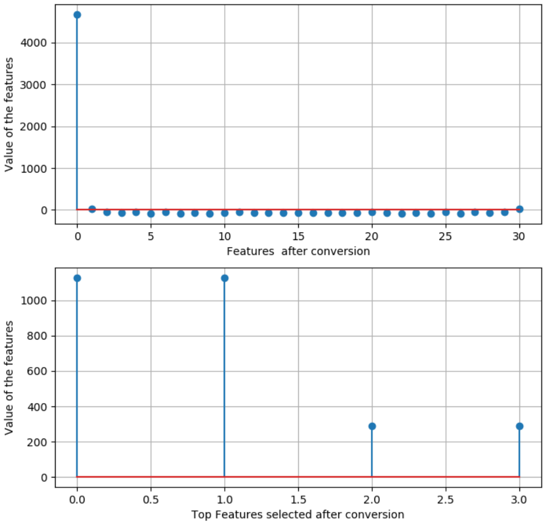
## Question 1 and 2

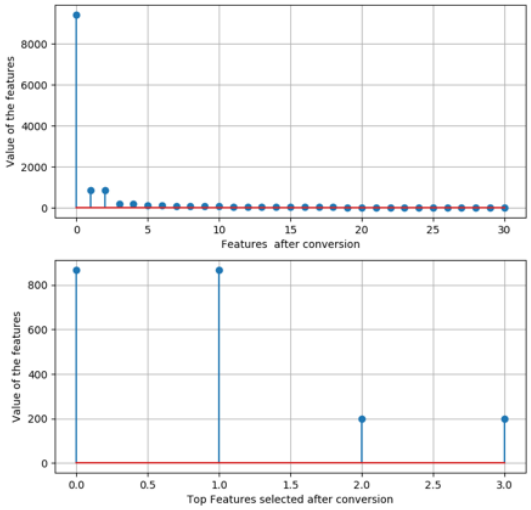
## 1.FAST FOURIER TRANSFORM

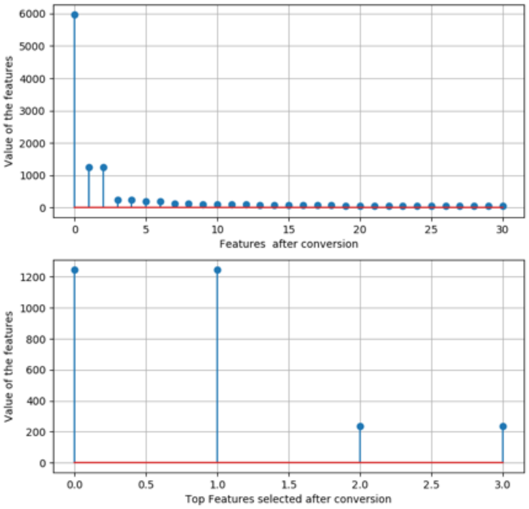
Fast Fourier Transform is used to extract frequency-based information from the time series data. We only select those frequency components which show high variance and keep it as a good feature among the whole array. For this project I have kept top 4 features for each row of data. I also have ignored the first highest feature as it comes from a constant 1 which act as a noise. I have plotted the original extracted data in **decreasing order** with the selected **top 4 data** on the bottom.

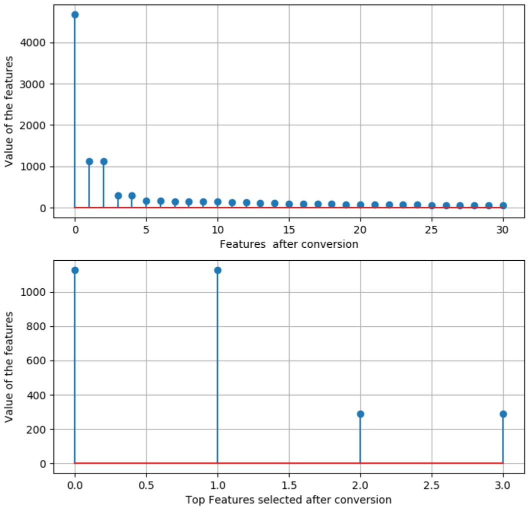
### Reason for choosing FFT

It gives the peak of the CGM data at which meal was taken showing the rise of glucose level in the body. For now, this can be considered as a good feature for detecting the meal intake in the graph









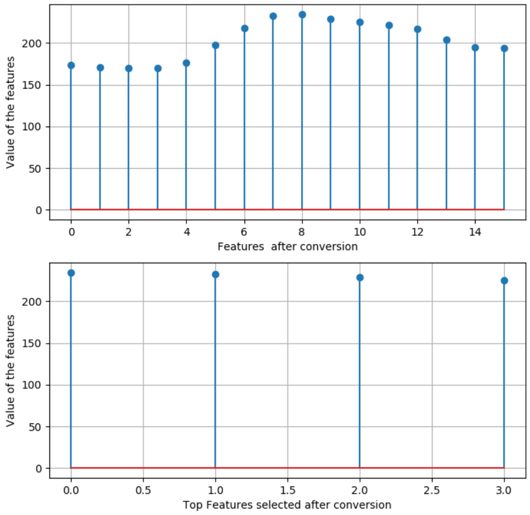
All 31 graph for each and every person can be found under Person folder and subfolder FFT.

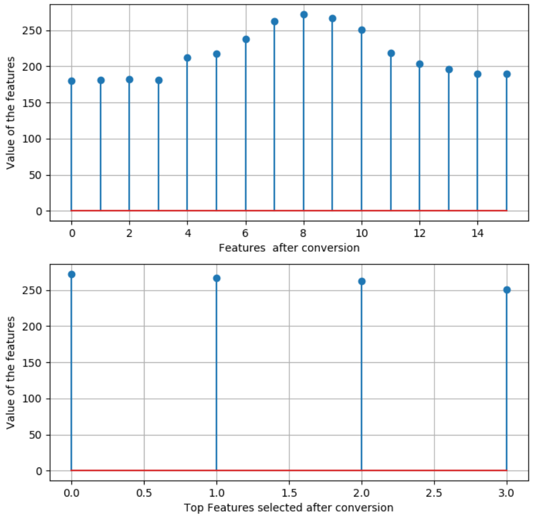
## 2.DISCRETE WAVELET TRANSFORM

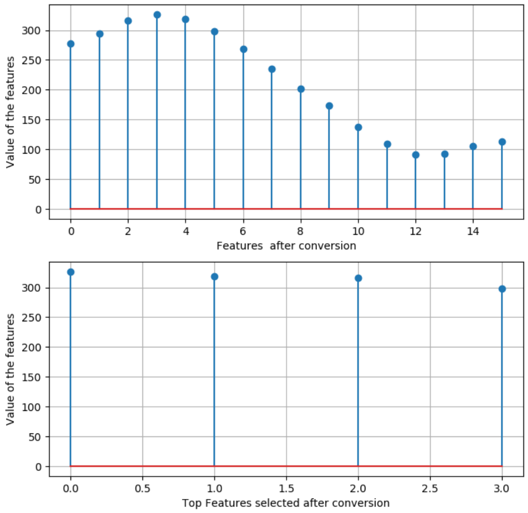
Discrete wavelet transform uses different sets of wavelet scales and transition to find the frequency domain of the given series. It is more helpful in finding the change as compared to Fourier transform because it also captures the rate and location. For this also I have taken top 4 features as my feature matrix

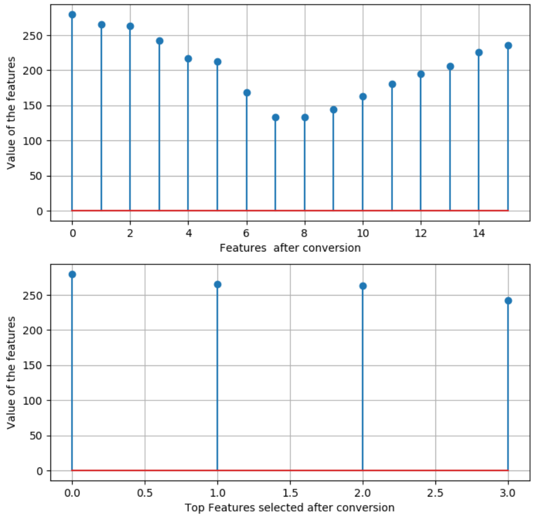
### Reason for choosing DWT

Discrete Wavelet Transform is similar to FFT but the advantage of choosing DWT is that it capture the rate of change more accurately as compared to the FFT and also the time stamp if we want to know the exact location.









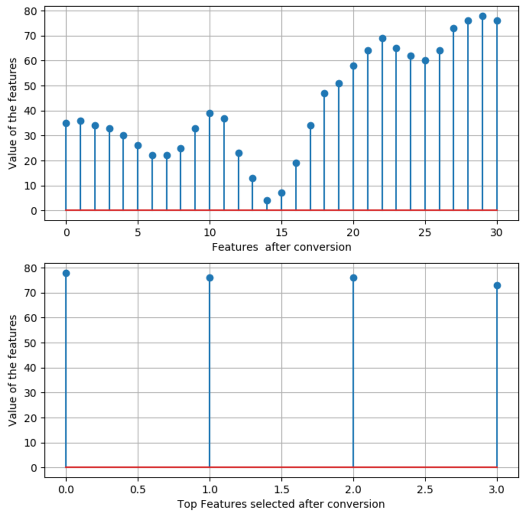
All 31 graph for each and every person can be found under Person folder and subfolder DWT.

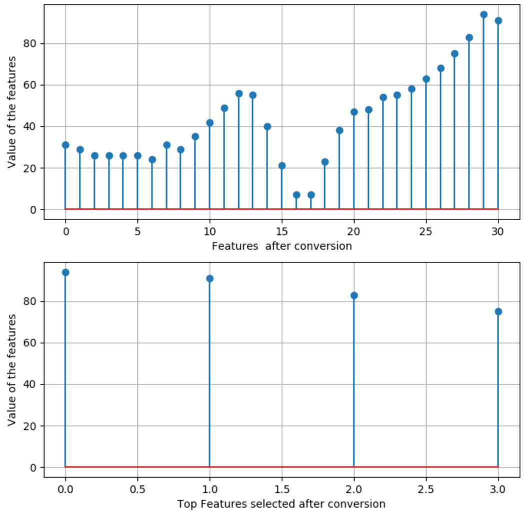
# 3.VELOCITY

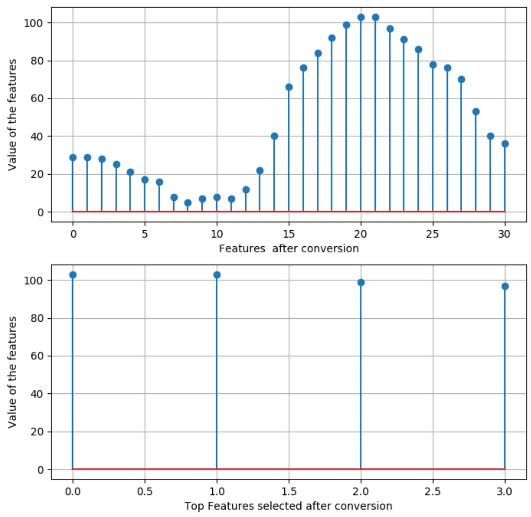
Velocity finds the rate of change between the t+1 and t timestamp of the time series data. Its find the rate of change of the value between these points.
For this also I have taken top 4 features as my feature matrix

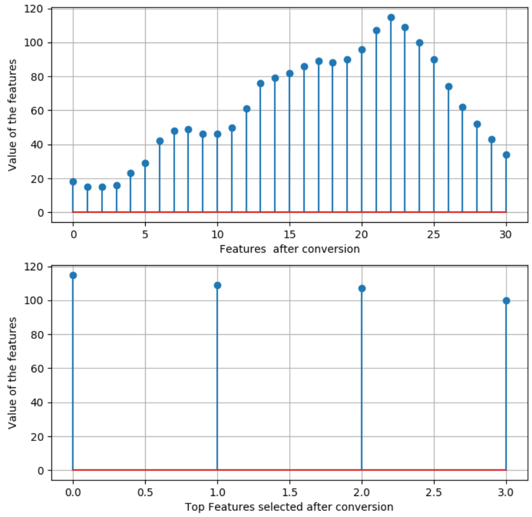
### Reason for choosing Velocity

It captures rate of change of y value which will help us to locate the maximum point and points around it and I selected the top 4 of those points as my features









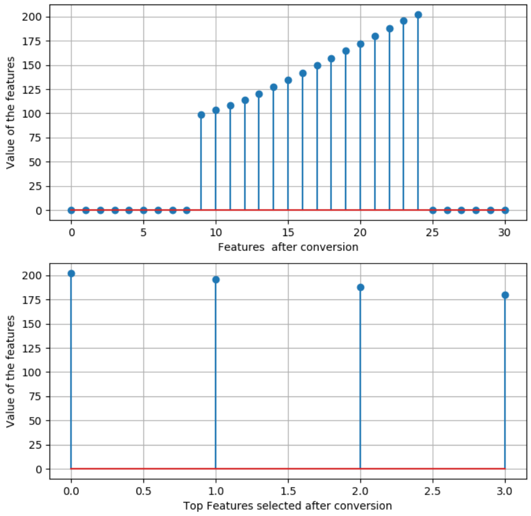
All 31 graph for each and every person can be found under Person folder and subfolder Velocity.

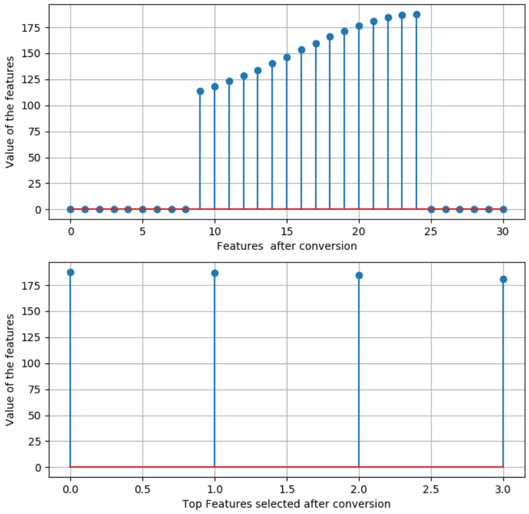
## 4. MOVING AVERAGE

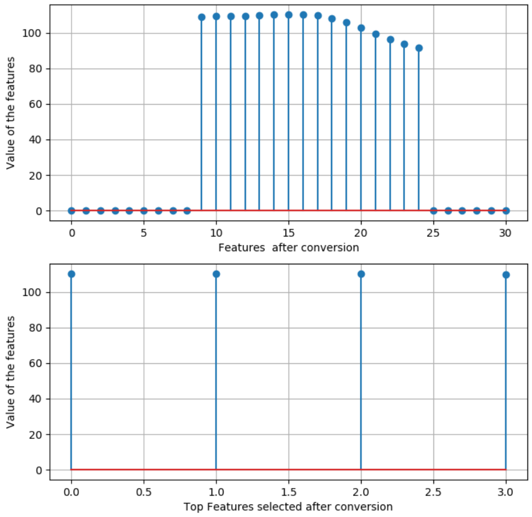
Moving average is used to create smooth version of the time series data set. It helps to remove any random noise that can be present over the series or any random fluctuations. I have used rolling function of pandas creating moving average for length 16. I have removed the NAN values in order to create plots and finding the maximum 4 values. For this also I have taken top 4 features as my feature matrix

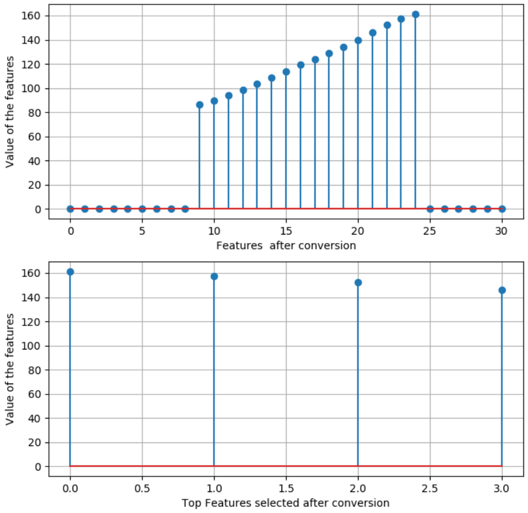
### Reason for choosing Moving Average

Moving average smoothes the data and form a good features for creating a stationary points also keeping previous information.





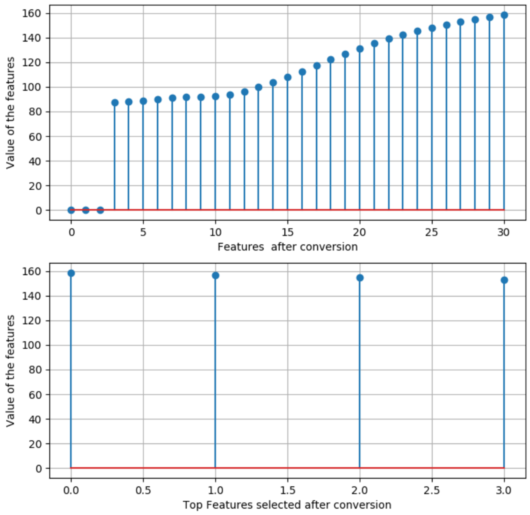


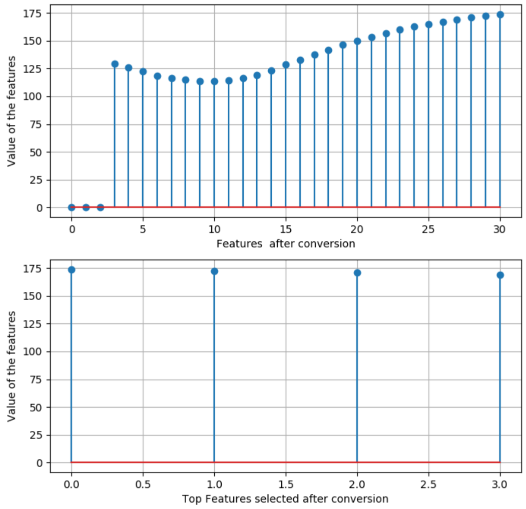


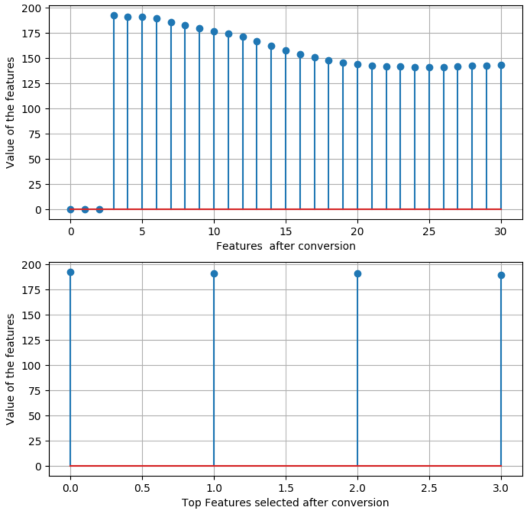
All 31 graph for each and every person can be found under Person folder and subfolder RollingWindow.

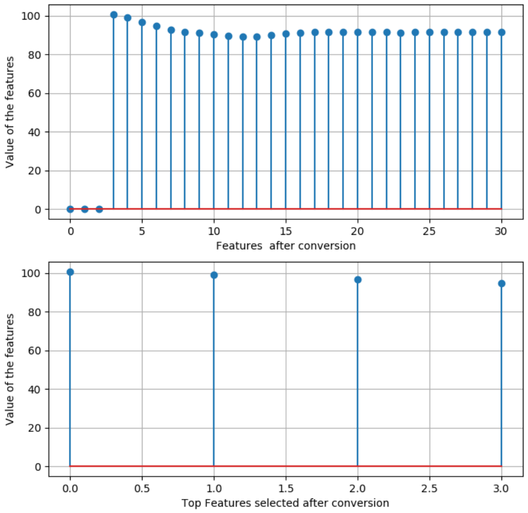
## 5. EXPANDING WINDOW

Expanding windows is same as cumulative sum which keeps the record of all the past things and increases over time. Same happens in expanding window where number of observation increases over time.









All 31 graph for each and every person can be found under Person folder and subfolder ExpandingWindow.

# QUESTION 3

Drawbacks and Advantages of choosing all the 5 features above

1. Fast Fourier Transform- We have used FFT to capture the increase in the glucose level of a person when there is food intake. According to initial hypothesis it should have selected the points with high change in value but in some of the case it selected peak from graph with multiple maxima. It

doesn’t capture the exact point where there is a spike but can also capture multiple points where there is not sudden increase. The other thing is we have to deal with real and imaginary points while dealing with FFT. Since most of the dataset has one single maxima it forms a decent feature. To avoid this I used DWT to look further

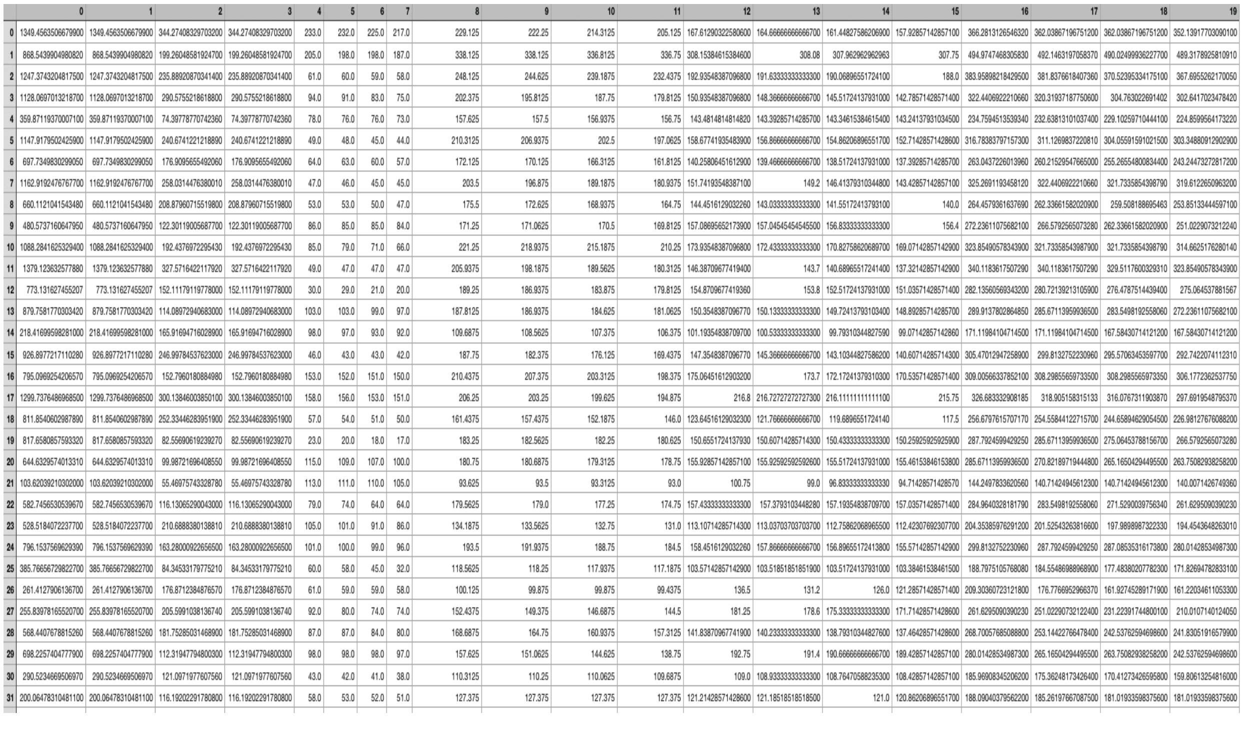
2. Discrete Wavelet Transform- In DWT both higher frequency window and lower frequency window helps to capture rate of change in glucose level more accurately as compared to the FFT.
3. Velocity- It tries to find the maximum point in the graph but sometimes lead to find the maximum point where these is least rate of change in the glucose level.
4. Moving Average- Its tries to smooth the graph and create stationary points with minimum change in the glucose level. It provides a constant value keeping the value close to the maxima if window size is selected properly.
5. Expanding Window- Expanding window is not at all helpful in this case as it keeps all the information from the past and doesn’t help in finding the exact instantaneous change in the value.

## Question 4

## FEATURE MATRIX

I have created a FeatureExtracted.csv file which contains all the features that I have extracted using the above feature extraction technique. Since I have extracted 5 features having top 4 from each method we have matrix of size 32X20 for each person

Screenshot of the feature matrix



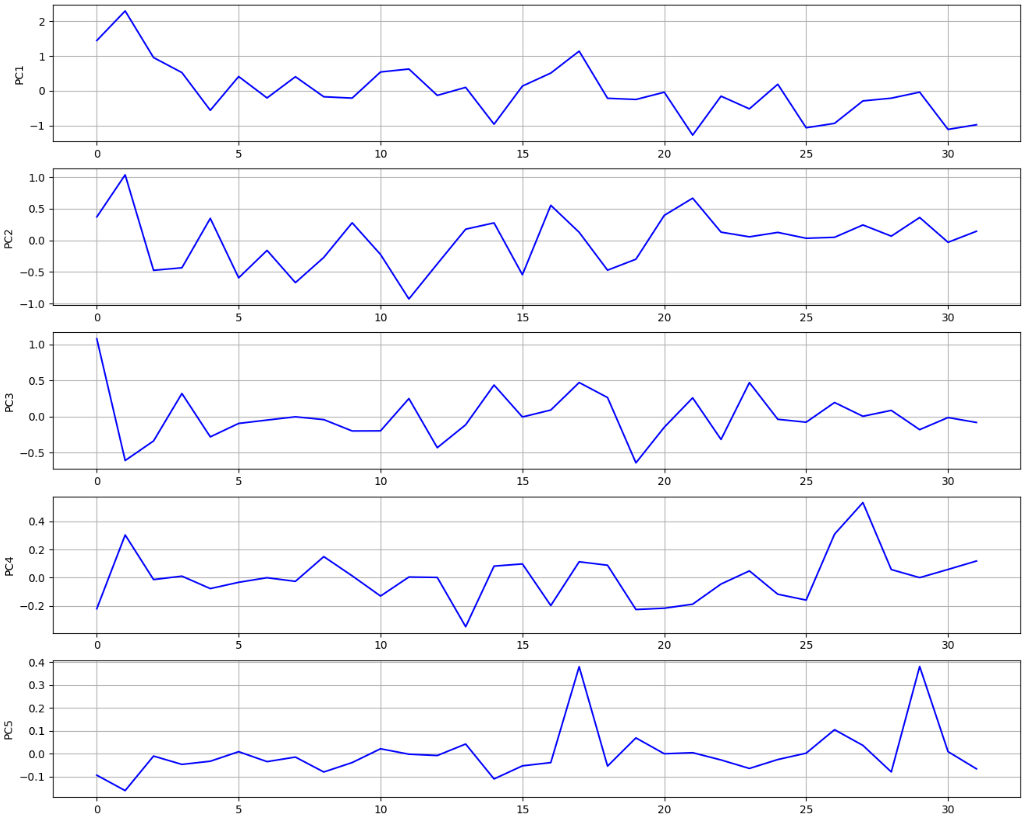
|  | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1349.4563506679900 | 1349.4563506679900 | 344.27408329703200 | 344.27408329703200 | 233.0 | 232.0 | 225.0 | 217.0 | 229.125 | 222.25 |
| 1 | 868.5439904980820 | 868.5439904980820 | 199.26048581924700 | 199.26048581924700 | 205.0 | 198.0 | 198.0 | 187.0 | 338.125 | 338.125 |
| 2 | 1247.3743204817500 | 1247.3743204817500 | 235.88920870341400 | 235.88920870341400 | 61.0 | 60.0 | 59.0 | 58.0 | 248.125 | 244.625 |
| 3 | 1128.0697013218700 | 1128.0697013218700 | 290.5755218618800 | 290.5755218618800 | 94.0 | 91.0 | 83.0 | 75.0 | 202.375 | 195.8125 |
| 4 | 359.87119370007100 | 359.87119370007100 | 74.39778770742360 | 74.39778770742360 | 78.0 | 76.0 | 76.0 | 73.0 | 157.625 | 157.5 |
| 5 | 1147.9179502425900 | 1147.9179502425900 | 240.6741221218890 | 240.6741221218890 | 49.0 | 48.0 | 45.0 | 44.0 | 210.3125 | 206.9375 |
| 6 | 697.7349830299050 | 697.7349830299050 | 176.9095655492060 | 176.9095655492060 | 64.0 | 63.0 | 60.0 | 57.0 | 172.125 | 170.125 |
| 7 | 1162.9192476767700 | 1162.9192476767700 | 258.0314476380010 | 258.0314476380010 | 47.0 | 46.0 | 45.0 | 45.0 | 203.5 | 196.875 |
| 8 | 660.1121041543480 | 660.1121041543480 | 208.87960715519800 | 208.87960715519800 | 53.0 | 53.0 | 50.0 | 47.0 | 175.5 | 172.625 |
| 9 | 480.5737160647950 | 480.5737160647950 | 122.30119005687700 | 122.30119005687700 | 86.0 | 85.0 | 85.0 | 84.0 | 171.25 | 171.0625 |
| 10 | 1088.2841625329400 | 1088.2841625329400 | 192.4376972295430 | 192.4376972295430 | 85.0 | 79.0 | 71.0 | 66.0 | 221.25 | 218.9375 |
| 11 | 1379.123632577880 | 1379.123632577880 | 327.5716422117920 | 327.5716422117920 | 49.0 | 47.0 | 47.0 | 47.0 | 205.9375 | 198.1875 |
| 12 | 773.131627455207 | 773.131627455207 | 152.11179119778000 | 152.11179119778000 | 30.0 | 29.0 | 21.0 | 20.0 | 189.25 | 186.9375 |
| 13 | 879.7581770303420 | 879.7581770303420 | 114.08972940683000 | 114.08972940683000 | 103.0 | 103.0 | 99.0 | 97.0 | 187.8125 | 186.9375 |
| 14 | 218.41699598281000 | 218.41699598281000 | 165.91694716028900 | 165.91694716028900 | 98.0 | 97.0 | 93.0 | 92.0 | 109.6875 | 108.5625 |
| 15 | 926.8977217110280 | 926.8977217110280 | 246.99784537623000 | 246.99784537623000 | 46.0 | 43.0 | 43.0 | 42.0 | 187.75 | 182.375 |
| 16 | 795.0969254206570 | 795.0969254206570 | 152.7960180884980 | 152.7960180884980 | 153.0 | 152.0 | 151.0 | 150.0 | 210.4375 | 207.375 |
| 17 | 1299.7376486968500 | 1299.7376486968500 | 300.13846003850100 | 300.13846003850100 | 158.0 | 156.0 | 153.0 | 151.0 | 206.25 | 203.25 |
| 18 | 811.8540602987890 | 811.8540602987890 | 252.33446283951900 | 252.33446283951900 | 57.0 | 54.0 | 51.0 | 50.0 | 161.4375 | 157.4375 |
| 19 | 817.6580857593320 | 817.6580857593320 | 82.55690619239270 | 82.55690619239270 | 23.0 | 20.0 | 18.0 | 17.0 | 183.25 | 182.5625 |
| 20 | 644.6329574013310 | 644.6329574013310 | 99.98721696408550 | 99.98721696408550 | 115.0 | 109.0 | 107.0 | 100.0 | 180.75 | 180.6875 |
| 21 | 103.62039210302000 | 103.62039210302000 | 55.46975743328780 | 55.46975743328780 | 113.0 | 111.0 | 110.0 | 105.0 | 93.625 | 93.5 |
| 22 | 582.7456530539670 | 582.7456530539670 | 116.13065290043000 | 116.13065290043000 | 79.0 | 74.0 | 64.0 | 64.0 | 179.5625 | 179.0 |
| 23 | 528.5184072237700 | 528.5184072237700 | 210.6888380138810 | 210.6888380138810 | 105.0 | 101.0 | 91.0 | 86.0 | 134.1875 | 133.5625 |
| 24 | 796.1537569629390 | 796.1537569629390 | 163.28000922656500 | 163.28000922656500 | 101.0 | 100.0 | 99.0 | 96.0 | 193.5 | 191.9375 |
| 25 | 385.76656729822700 | 385.76656729822700 | 84.34533179775210 | 84.34533179775210 | 60.0 | 58.0 | 45.0 | 32.0 | 118.5625 | 118.25 |
| 26 | 261.4127906136700 | 261.4127906136700 | 176.8712384876570 | 176.8712384876570 | 61.0 | 59.0 | 59.0 | 58.0 | 100.125 | 99.875 |
| 27 | 255.83978165520700 | 255.83978165520700 | 205.5991038136740 | 205.5991038136740 | 92.0 | 80.0 | 74.0 | 74.0 | 152.4375 | 149.375 |
| 28 | 568.4407678815260 | 568.4407678815260 | 181.75285031468900 | 181.75285031468900 | 87.0 | 87.0 | 84.0 | 80.0 | 168.6875 | 164.75 |
| 29 | 698.2257404777900 | 698.2257404777900 | 112.31947794800300 | 112.31947794800300 | 98.0 | 98.0 | 98.0 | 97.0 | 157.625 | 151.0625 |
| 30 | 290.5234669506970 | 290.5234669506970 | 121.0971977607560 | 121.0971977607560 | 43.0 | 42.0 | 41.0 | 38.0 | 110.3125 | 110.25 |
| 31 | 200.06478310481100 | 200.06478310481100 | 116.19202291780800 | 116.19202291780800 | 58.0 | 53.0 | 52.0 | 51.0 | 127.375 | 127.375 |

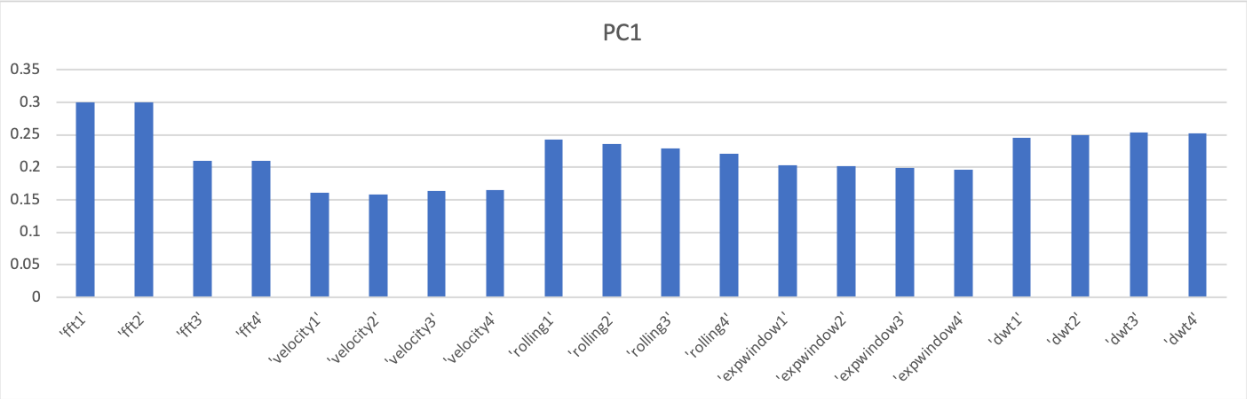
|  | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 214.3125 | 205.125 | 167.61290322580600 | 164.66666666666700 | 161.44827586206900 | 157.92857142857100 | 366.2813126546320 | 362.03867196751200 | 362.03867196751200 | 352.13917703090100 |
| 1 | 336.8125 | 336.75 | 308.15384615384600 | 308.08 | 307.962962962963 | 307.75 | 494.9747468305830 | 492.1463197058370 | 490.02499936227700 | 489.3178925810910 |
| 2 | 239.1875 | 232.4375 | 192.93548387096800 | 191.63333333333300 | 190.06896551724100 | 188.0 | 383.95898218429500 | 381.8376618407360 | 370.52395334175100 | 367.6955262170050 |
| 3 | 187.75 | 179.8125 | 150.93548387096800 | 148.36666666666700 | 145.51724137931000 | 142.78571428571400 | 322.4406922210660 | 320.31937187750600 | 304.763022691402 | 302.6417023478420 |
| 4 | 156.9375 | 156.75 | 143.4814814814820 | 143.39285714285700 | 143.34615384615400 | 143.24137931034500 | 234.7594513539340 | 232.63813101037400 | 229.10259710444100 | 224.8599564173220 |
| 5 | 202.5 | 197.0625 | 158.67741935483900 | 156.86666666666700 | 154.86206896551700 | 152.71428571428600 | 316.78383797157300 | 311.1269837220810 | 304.05591591021500 | 303.34880912902900 |
| 6 | 166.3125 | 161.8125 | 140.25806451612900 | 139.46666666666700 | 138.51724137931000 | 137.39285714285700 | 263.0437226013960 | 260.21529547665000 | 255.26554800834400 | 243.24473272817200 |
| 7 | 189.1875 | 180.9375 | 151.74193548387100 | 149.2 | 146.41379310344800 | 143.42857142857100 | 325.2691193458120 | 322.4406922210660 | 321.7335854398790 | 319.6122650963200 |
| 8 | 168.9375 | 164.75 | 144.4516129032260 | 143.03333333333300 | 141.55172413793100 | 140.0 | 264.4579361637690 | 262.33661582020900 | 259.508188695463 | 253.85133444597100 |
| 9 | 170.5 | 169.8125 | 157.08695652173900 | 157.04545454545500 | 156.83333333333300 | 156.4 | 272.23611075682100 | 266.5792565073280 | 262.33661582020900 | 251.0229073212240 |
| 10 | 215.1875 | 210.25 | 173.93548387096800 | 172.43333333333300 | 170.82758620689700 | 169.07142857142900 | 323.85490578343900 | 321.73358543987900 | 321.7335854398790 | 314.6625176280140 |
| 11 | 189.5625 | 180.3125 | 146.38709677419400 | 143.7 | 140.68965517241400 | 137.32142857142900 | 340.1183617507290 | 340.1183617507290 | 329.5117600329310 | 323.85490578343900 |
| 12 | 183.875 | 179.8125 | 154.8709677419360 | 153.8 | 152.51724137931000 | 151.03571428571400 | 282.13560569343200 | 280.72139213105900 | 276.4787514439400 | 275.064537881567 |
| 13 | 184.625 | 181.0625 | 150.3548387096770 | 150.13333333333300 | 149.72413793103400 | 148.89285714285700 | 289.9137802864850 | 285.67113959936500 | 283.5498192558060 | 272.23611075682100 |
| 14 | 107.375 | 106.375 | 101.19354838709700 | 100.53333333333300 | 99.79310344827590 | 99.07142857142860 | 171.11984104714500 | 171.11984104714500 | 167.58430714121200 | 167.58430714121200 |
| 15 | 176.125 | 169.4375 | 147.3548387096770 | 145.36666666666700 | 143.10344827586200 | 140.60714285714300 | 305.47012947258900 | 299.8132752230960 | 295.57063453597700 | 292.7422074112310 |
| 16 | 203.3125 | 198.375 | 175.06451612903200 | 173.7 | 172.17241379310300 | 170.53571428571400 | 309.00566337852100 | 308.29855659733500 | 308.2985565973350 | 306.1772362537750 |
| 17 | 199.625 | 194.875 | 216.8 | 216.72727272727300 | 216.11111111111100 | 215.75 | 326.683332908185 | 318.905158315133 | 316.0767311903870 | 297.6919548795370 |
| 18 | 152.1875 | 146.0 | 123.64516129032300 | 121.76666666666700 | 119.6896551724140 | 117.5 | 256.6797615707170 | 254.55844122715700 | 244.65894629054500 | 226.98127676088200 |
| 19 | 182.25 | 180.625 | 150.6551724137930 | 150.60714285714300 | 150.43333333333300 | 150.25925925925900 | 287.7924599429250 | 285.67113959936500 | 275.06453788156700 | 266.5792565073280 |
| 20 | 179.3125 | 178.75 | 155.92857142857100 | 155.92592592592600 | 155.51724137931000 | 155.46153846153800 | 285.67113959936500 | 270.82189719444800 | 265.16504294495500 | 263.75082938258200 |
| 21 | 93.3125 | 93.0 | 100.75 | 99.0 | 96.83333333333330 | 94.71428571428570 | 144.2497833620560 | 140.71424945612300 | 140.71424945612300 | 140.0071426749360 |
| 22 | 177.25 | 174.75 | 157.43333333333300 | 157.3793103448280 | 157.19354838709700 | 157.03571428571400 | 284.9640328181790 | 283.5498192558060 | 271.5290039756340 | 261.6295090390230 |
| 23 | 132.75 | 131.0 | 113.10714285714300 | 113.03703703703700 | 112.75862068965500 | 112.42307692307700 | 204.35385976291200 | 201.52543263816600 | 197.9898987322330 | 194.4543648263010 |
| 24 | 188.75 | 184.5 | 158.4516129032260 | 157.86666666666700 | 156.89655172413800 | 155.57142857142900 | 299.8132752230960 | 287.7924599429250 | 287.08535316173800 | 280.01428534987300 |
| 25 | 117.9375 | 117.1875 | 103.57142857142900 | 103.51851851851900 | 103.51724137931000 | 103.38461538461500 | 188.7975105768080 | 184.55486988968900 | 177.48380207782300 | 171.82694782833100 |
| 26 | 99.875 | 99.4375 | 136.5 | 131.2 | 126.0 | 121.28571428571400 | 209.30360723121800 | 176.7766952966370 | 161.92745289171900 | 161.22034611053300 |
| 27 | 146.6875 | 144.5 | 181.25 | 178.6 | 175.33333333333300 | 171.71428571428600 | 261.6295090390230 | 251.02290732122400 | 231.22391744800100 | 210.0107140124050 |
| 28 | 160.9375 | 157.3125 | 141.83870967741900 | 140.23333333333300 | 138.79310344827600 | 137.46428571428600 | 268.70057685088800 | 253.14422766478400 | 242.53762594698600 | 241.83051916579900 |
| 29 | 144.625 | 138.75 | 192.75 | 191.4 | 190.66666666666700 | 189.42857142857100 | 280.01428534987300 | 265.16504294495500 | 263.75082938258200 | 242.53762594698600 |
| 30 | 110.0625 | 109.6875 | 109.0 | 108.93333333333300 | 108.76470588235300 | 108.42857142857100 | 185.96908345206200 | 175.36248173426400 | 170.41273426595800 | 159.80613254816000 |
| 31 | 127.375 | 127.375 | 121.21428571428600 | 121.18518518518500 | 121.0 | 120.86206896551700 | 188.09040379562200 | 185.26197667087500 | 181.01933598375600 | 181.01933598375600 |

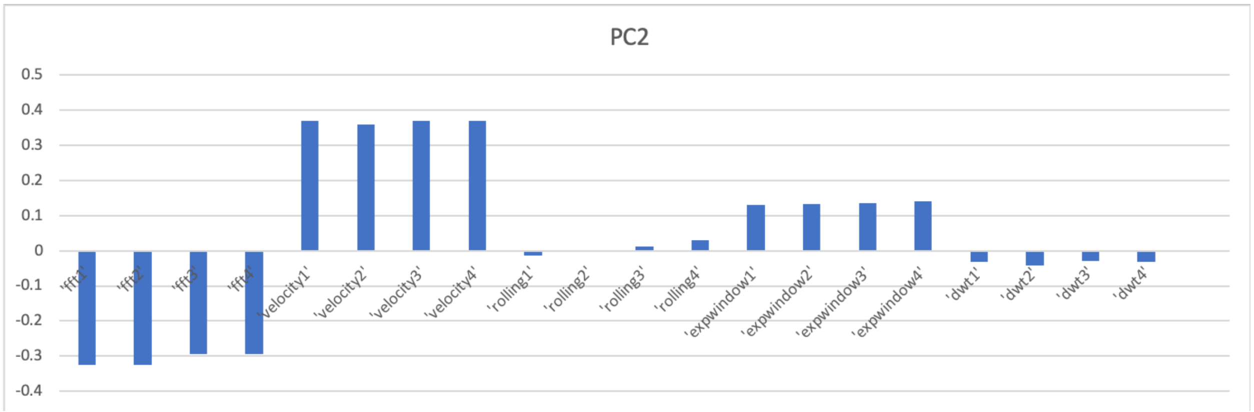
## Question 5

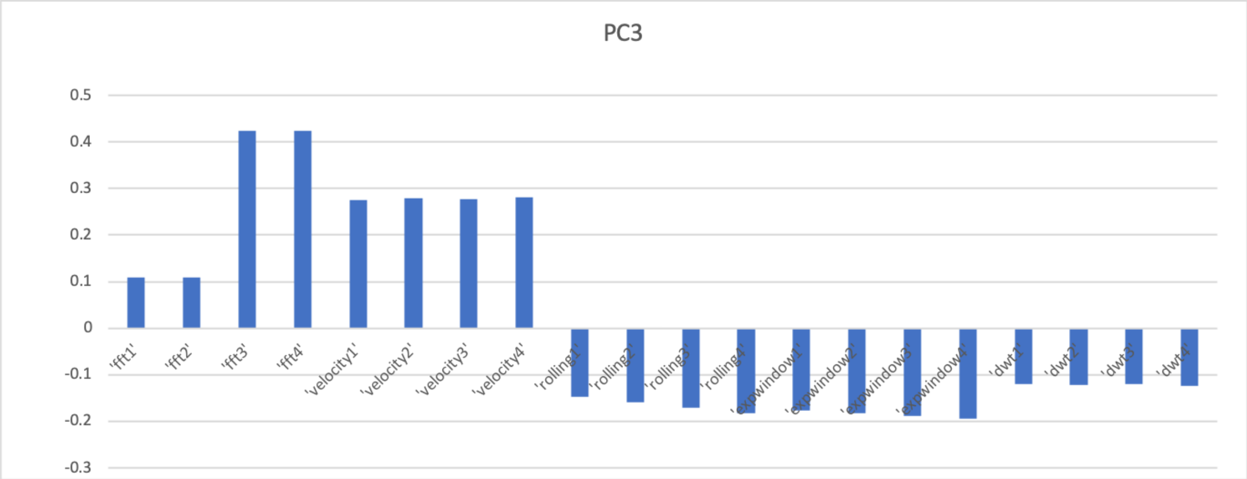
## PRINCIPAL COMPONENT ANALYSIS

For principal component analysis I am using the extracted feature matrix as input for PCA function from sklearn. It outputs an array of 31X5 matrix consisting of 5 principal components. Along with that I have created an eigen value matrix of 5X20 to see the contribution of each extracted feature on the 5 components Plots of eigen value matrix. Plotted all the five components in time series below.




PC1
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
'fft1'
'fft2'
'fft3'
'fft4'
'velocity1'
'velocity2'
'velocity3'
'velocity4'
'rolling1'
'rolling2'
'rolling3'
'rolling4'
'expwindow1'
'expwindow2'
'expwindow3'
'expwindow4'
'dwt1'
'dwt2'
'dwt3'
'dwt4'


PC2
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
'fft1'
'fft2'
'fft3'
'fft4'
'velocity1'
'velocity2'
'velocity3'
'velocity4'
'rolling1'
'rolling2'
'rolling3'
'rolling4'
'expwindow1'
'expwindow2'
'expwindow3'
'expwindow4'
'dwt1'
'dwt2'
'dwt3'
'dwt4'


PC3
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
'fft1'
'fft2'
'fft3'
'fft4'
'velocity1'
'velocity2'
'velocity3'
'velocity4'
'rolling1'
'rolling2'
'rolling3'
'rolling4'
'expwindow1'
'expwindow2'
'expwindow3'
'expwindow4'
'dwt1'
'dwt2'
'dwt3'
'dwt4'

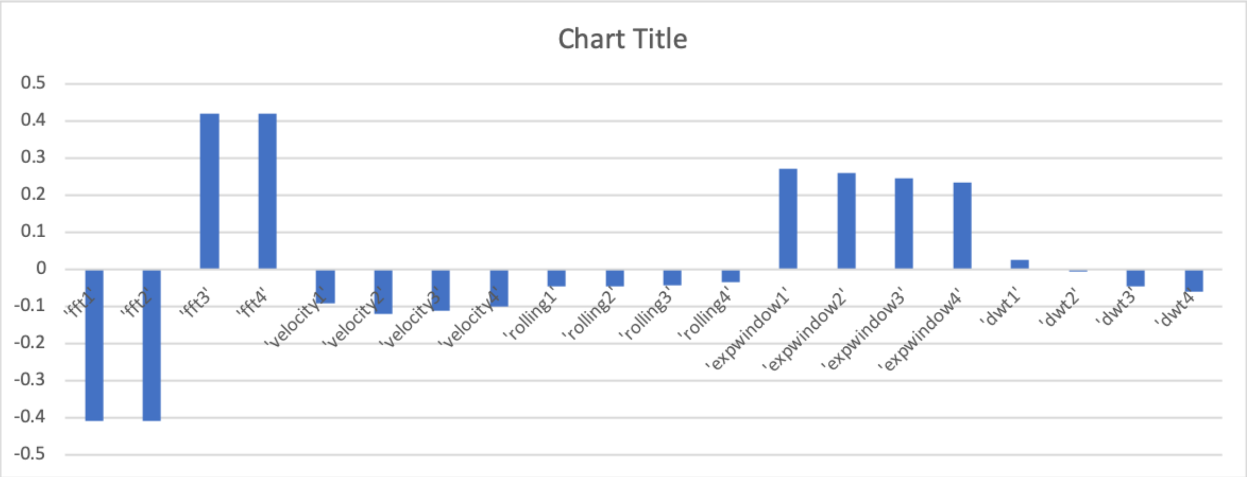
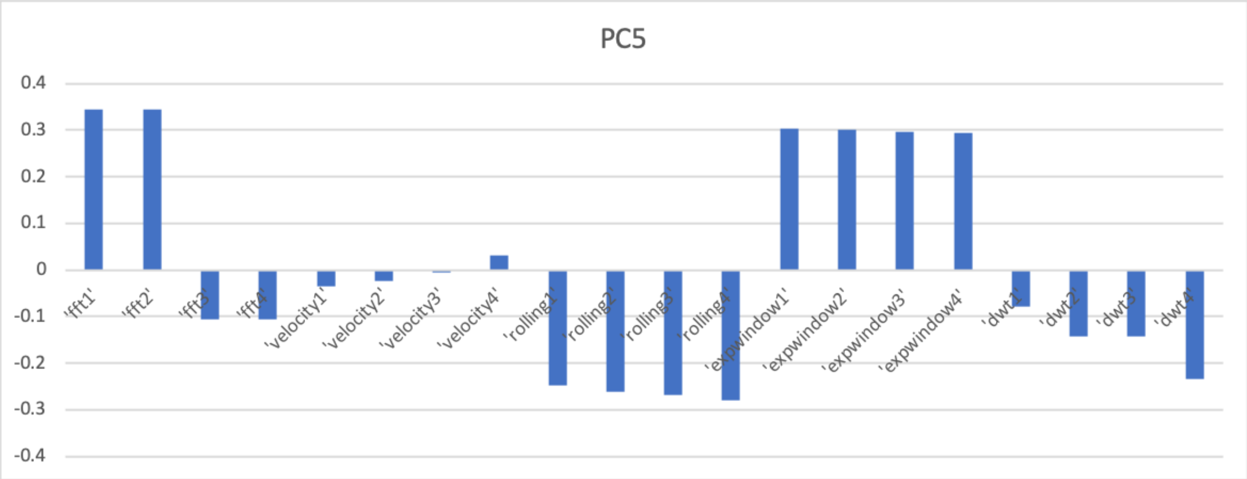
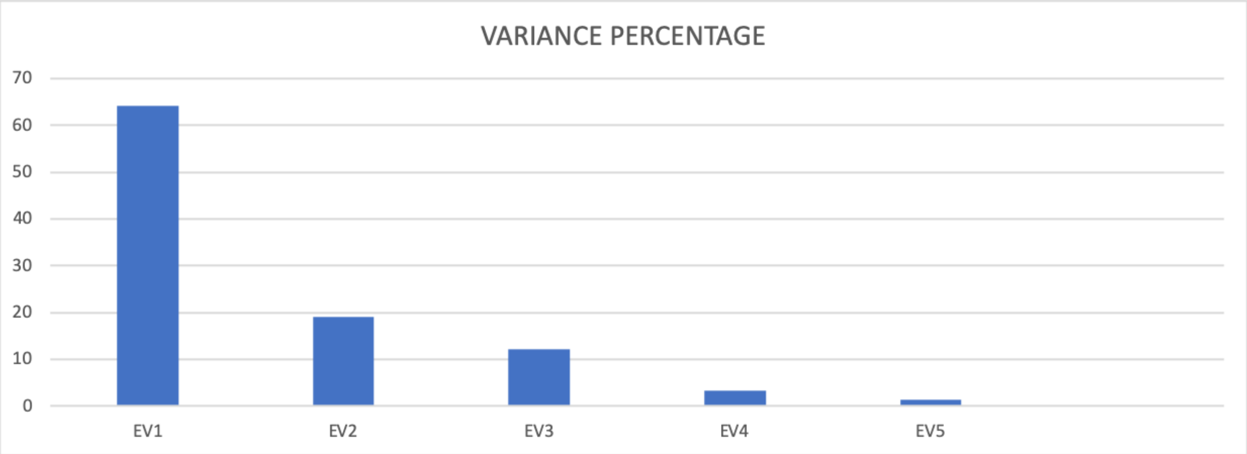


Chart Title


PC5


VARIANCE PERCENTAGE

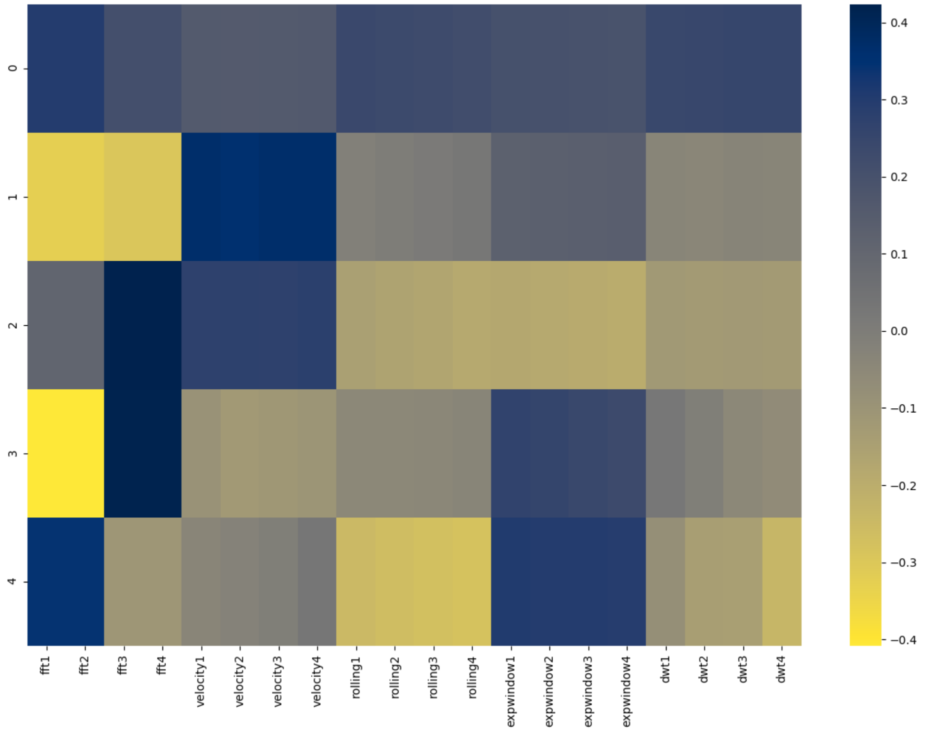
## Question 6

First principal component shows FFT, Moving Average and DFT as the most important feature. The variance for PC1 is around 63% among all the principal components
The second principal component shows Velocity and Expanding window as the most important feature. The variance for PC1 is around 19% among all the principal components

The third principal component shows Velocity and FFT as the most important feature. The variance for PC3 is around 11% among all the principal components

The fourth principal component shows Expanding window as the most important feature. The variance for PC4 is around 3% among all the principal components

Similarly it can be seen for all the Principal components in the figure plotted above. I have also made a heatmap using seaborn library for analysis.


fft1
fft2
fft3
fft4
velocity1
velocity2
velocity3
velocity4
rolling1
rolling2
rolling3
rolling4
expwindow1
expwindow2
expwindow3
expwindow4
dwt1
dwt2
dwt3
dwt4