Project 1 Report

DATA MINING

CSE 572: Spring2020

Submitted to: Professor Ayan Banerjee Ira A. Fulton School of Engineering Arizona State University

Submitted by: Rohith Varma Gaddam(rgaddam2@asu.edu**)**

January 2, 2020

Introduction

We have the time series data of the glucose levels at the time of meals the type diabetes patients. We have the data records of 5 patients. The main aim of this assignment is to extract the new features from the time series data and apply principal component analysis to the new features and get top 5 components.

Feature Extraction

In this step, we extract the new features from the given CGM time series data. To extract the new features we will use the following feature extraction methods. Before going into the feature extraction we have cleaned the data Na values.

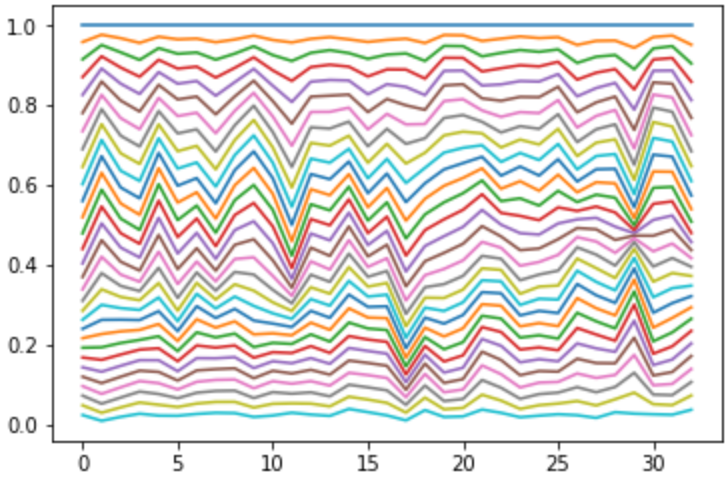
These are the following feature extraction methods we used:

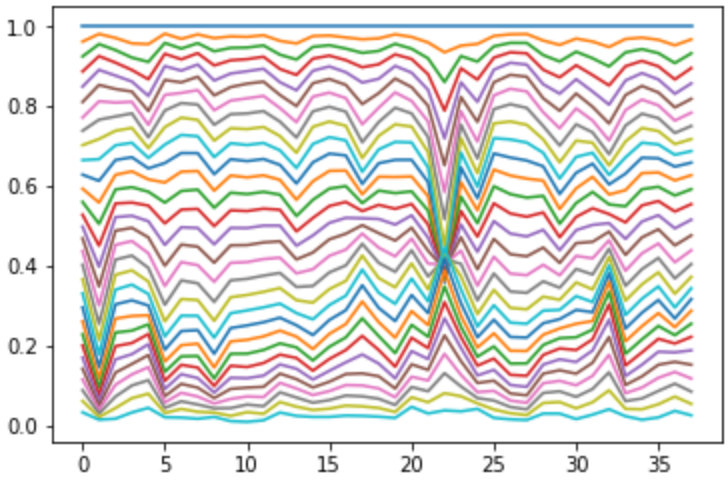
1. Auto Correlation
2. Velocity
3. Fast Fourier Transform
4. Moving average

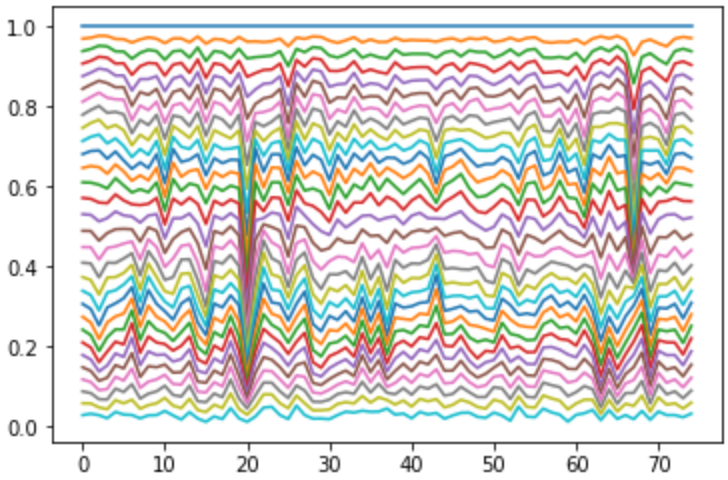
I. **AutoCorrelation:**

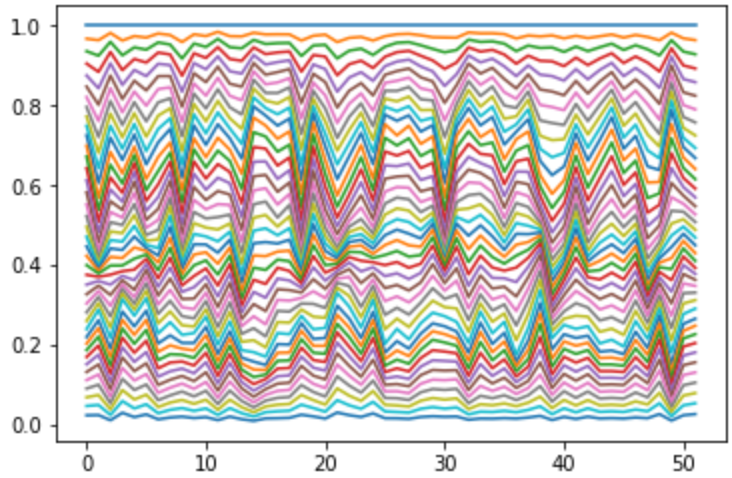
Generally correlation is used for the representing the dependence and similarity between the two data series. But in time series data we have only one series. Now, auto correlation comes in to the picture.For a time series data, we apply correlation on the series itself, with it's previous values. With the help of auto correlation we observe if there is a trend in time series.

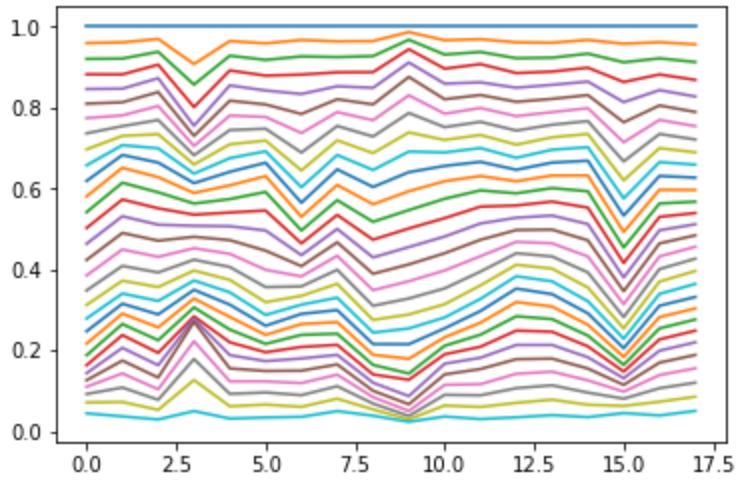
- Correlation between x,y is corr(x,y)
- AutoCorrelation of x is corr(x(t),x(t-K)), K is constant


1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30


1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30
35


1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
50
60
70

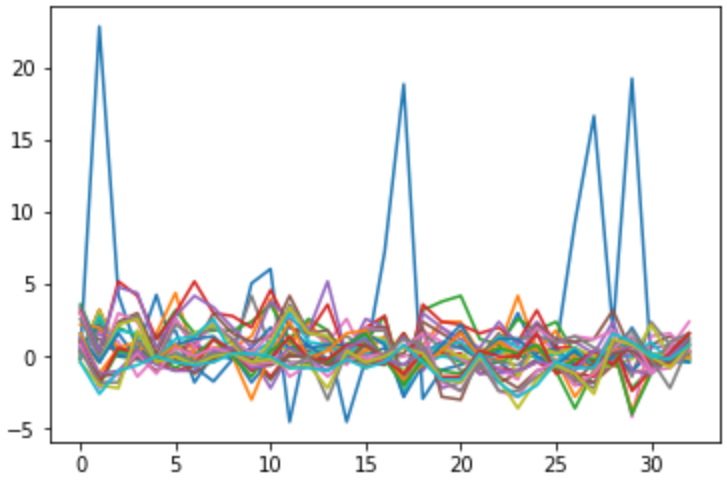


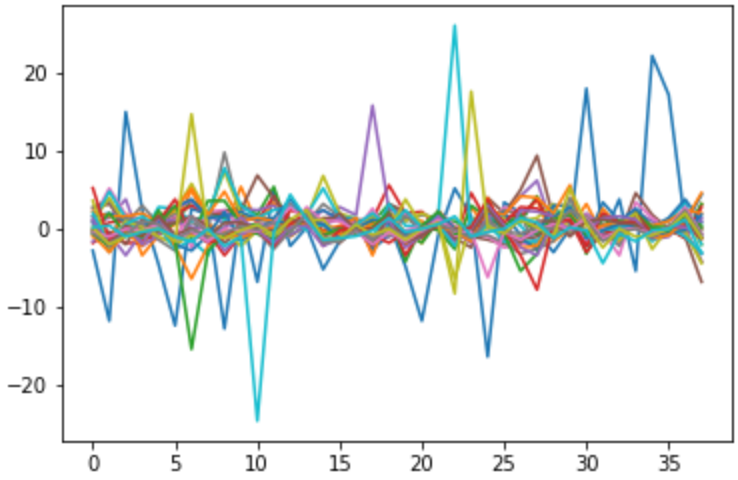


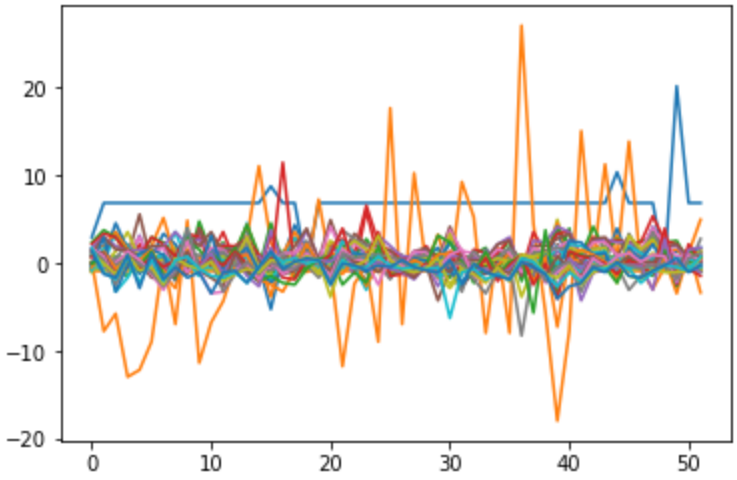
- **<u>Outcome:</u>** As the motive of auto correlation of to get the trends in time series data. Here no plot intersects with other plots. So, we can say that there is trend in the time series data.

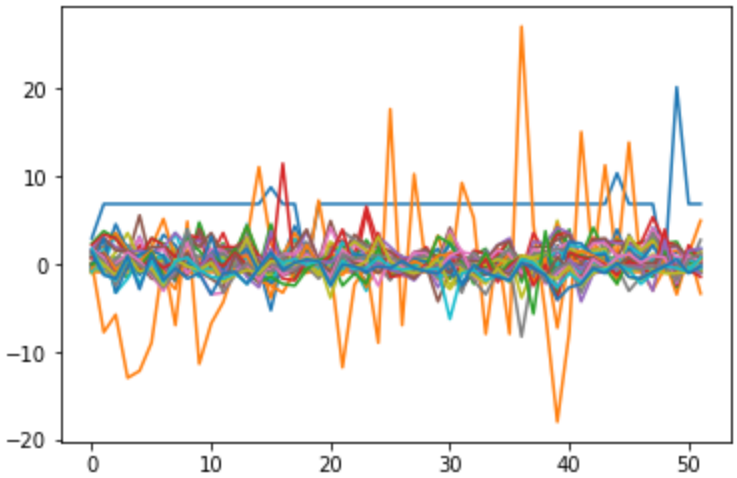
II. **Velocity:**

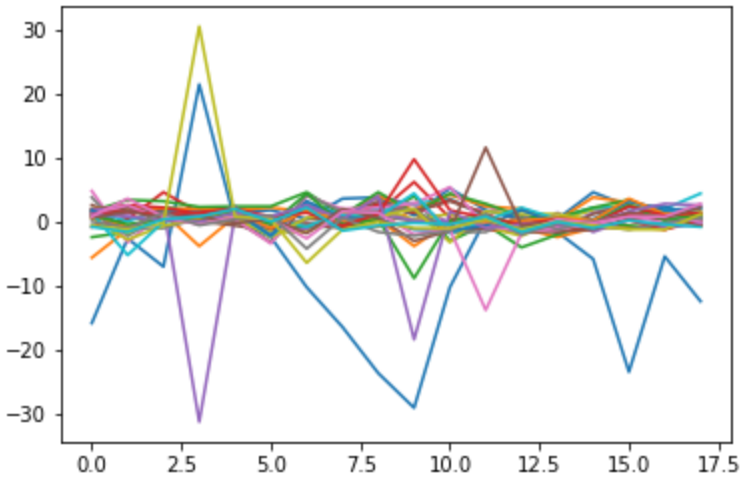
Here velocity is nothing but rate change glucose level for a time interval. For whole series calculated the velocity of glucose levels for contiguous intervals. With help of velocity, if there is a high rate of change of glucose level, we can predict that has meal intake.







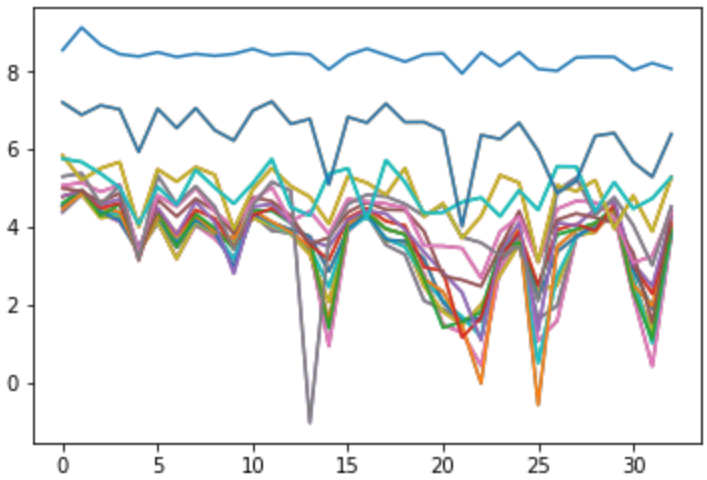


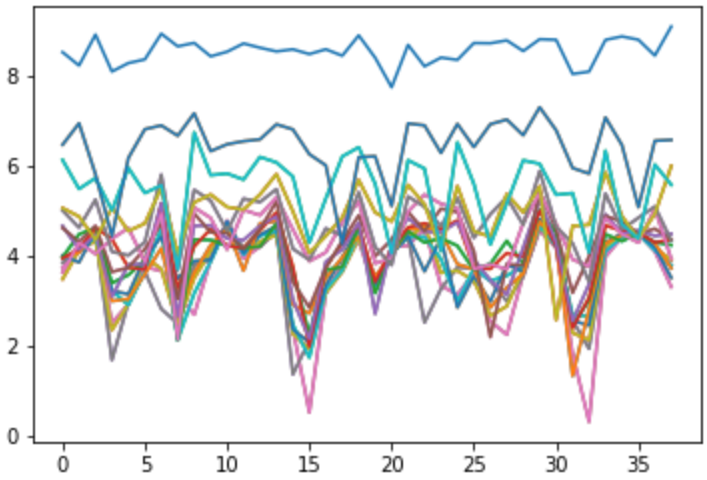


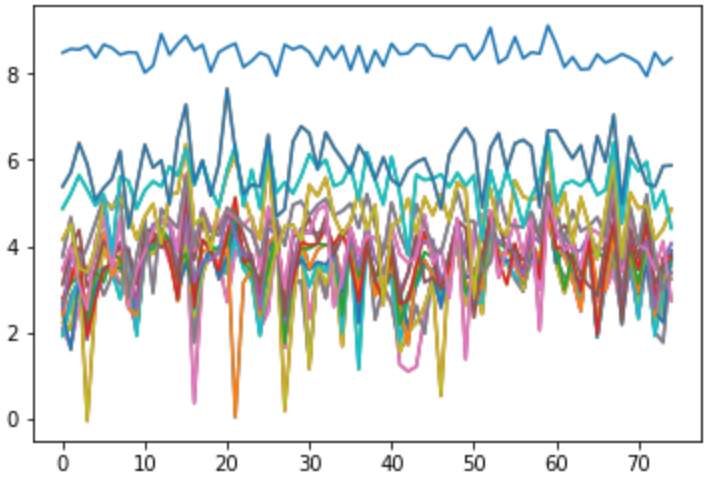
- **<u>Outcome:</u>** By observing the graphs, we can see that we have achieved the motive. The peaks in the graph represent the high rate of change of glucose leve

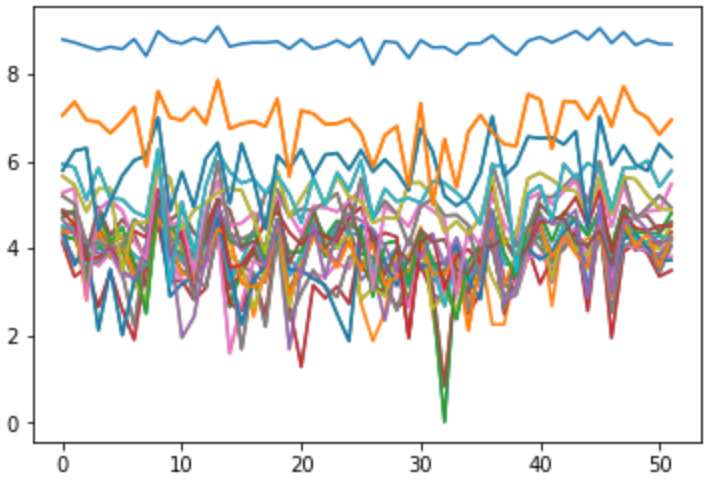
III. **Fast Fourier Transform:**

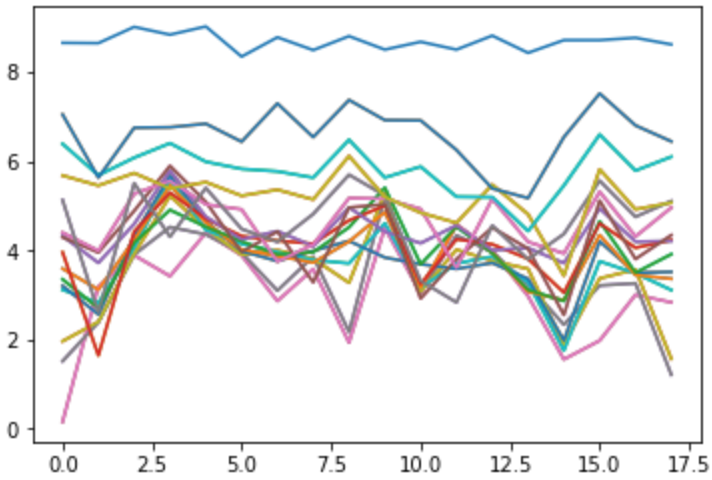
Fourier transform helps us to transform the signal from time domain to frequency domain. It gives the output in the form of the complex numbers. We have plotted the graph with the magnitude of the vectors. It helps us to get the peaks of the time series where the meal is taken







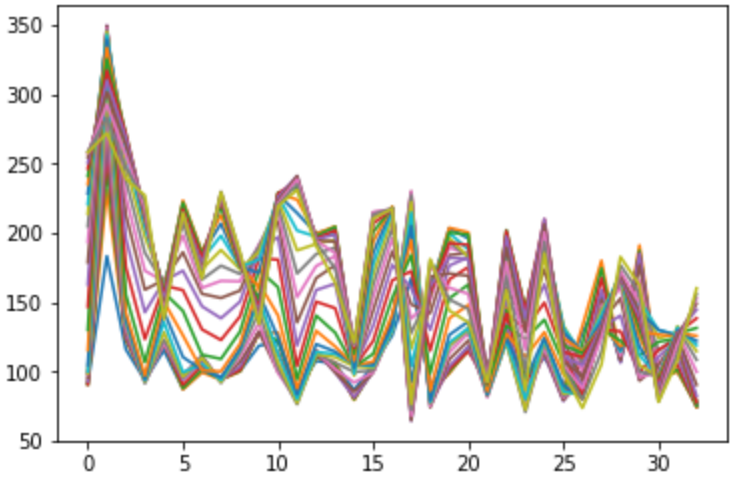


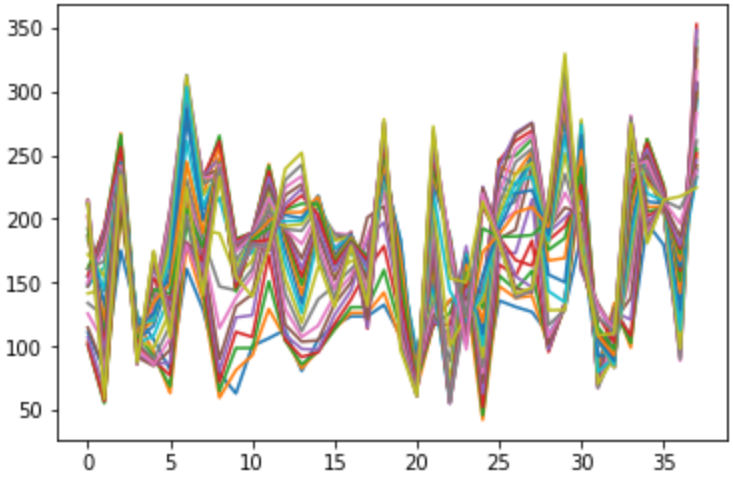


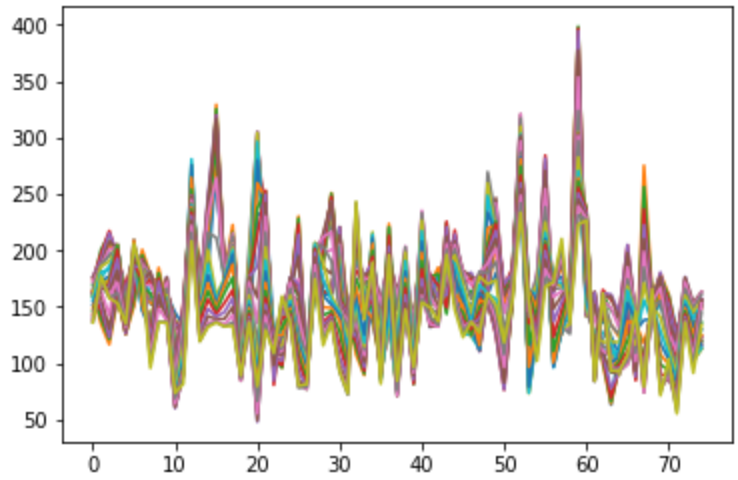
- **<u>Outcome:</u>** In the frequency domain the high peaks and bottom peaks both represent the amplitude of the signal. So, the represent high glucose level.

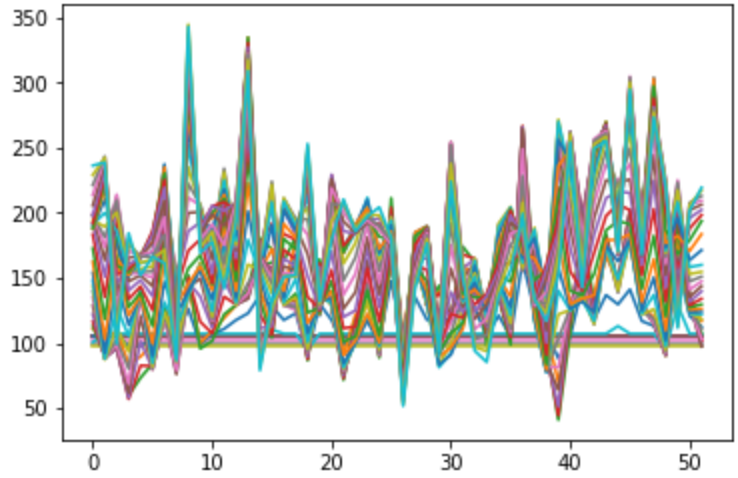
IV. **Moving Average:**

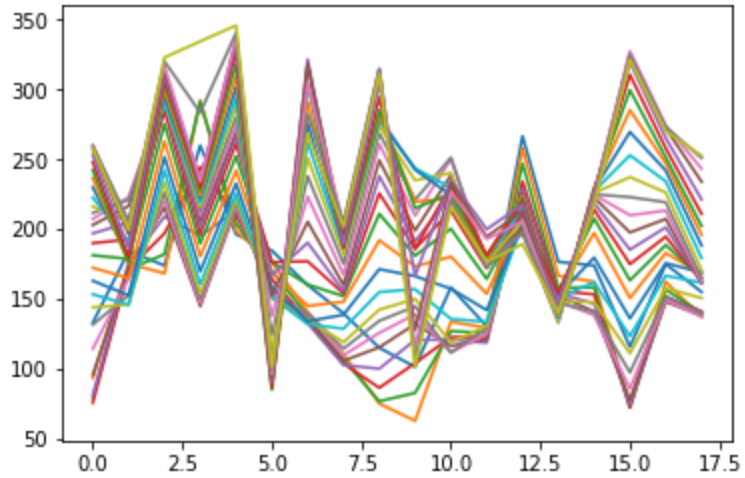
Moving average is a mean value of the contiguous subset set series of the times series data.It helps us to identify interesting trends in the data. It can be calculated with help of mean with window size. It helps us to observe where the glucose level goes high.






400
350
300
250
200
150
100
50
0
10
20
30
40
50
60
70


350
300
250
200
150
100
50
0
10
20
30
40
50


350
300
250
200
150
100
50
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5

- **Outcome:** The peaks of the moving average plot represent the highest in the time series data in the particular window area.

Feature Selection

Subtask 1: Arranging the feature matrix

In this step we need to combine features we have extracted till now. So, that we can do the further analysis on the features at once. Below screen shots are the data frame of combined feature matrix.

	0	1	2	3	4	5	6	7	8	9	...	110	111	112	113	
0	1.0	0.958655	0.915175	0.870791	0.826055	0.780690	0.735330	0.690181	0.645642	0.602391	...	220.666667	228.333333	235.000000	241.000000	246.333
1	1.0	0.976842	0.951032	0.922750	0.892110	0.859873	0.826120	0.790490	0.752720	0.713152	...	343.000000	339.666667	333.666667	325.333333	317.000
2	1.0	0.968162	0.932840	0.894653	0.854322	0.811996	0.768054	0.723892	0.679816	0.636179	...	257.666667	264.333333	268.666667	271.333333	271.333
3	1.0	0.956404	0.913716	0.871550	0.828458	0.784444	0.740195	0.696786	0.653099	0.610411	...	204.000000	208.000000	209.666667	212.000000	213.333
4	1.0	0.971888	0.943231	0.913965	0.883176	0.851184	0.818532	0.785480	0.752121	0.717225	...	159.666667	158.333333	157.333333	156.333333	154.666

113	114	115	116	117	118	119
241.000000	246.333333	250.666667	254.666667	257.333333	258.666667	258.000000
325.333333	317.000000	309.666667	301.666667	292.666667	282.666667	272.000000
271.333333	271.333333	268.666667	263.000000	256.000000	247.333333	239.000000
212.000000	213.333333	214.333333	215.333333	217.666667	223.000000	227.000000
156.333333	154.666667	151.333333	146.666667	142.333333	139.000000	137.666667

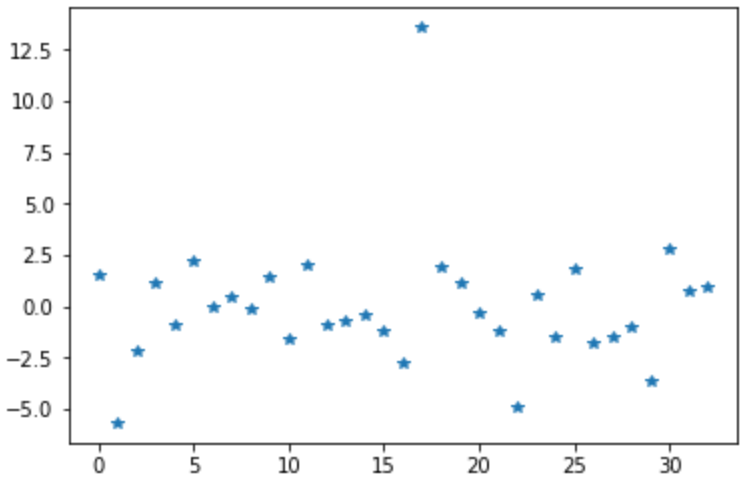
Subtask 2 : Execution of PCA

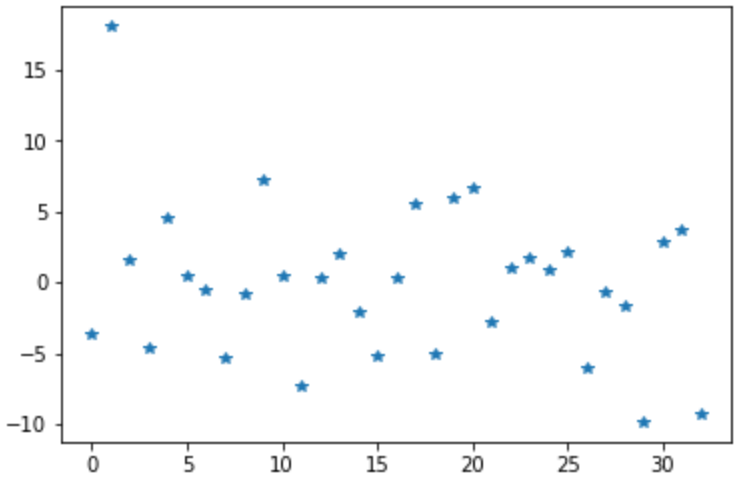
The principal component analysis done with help of scipy package. It gives the top 5 components and projects the data on to new principle components

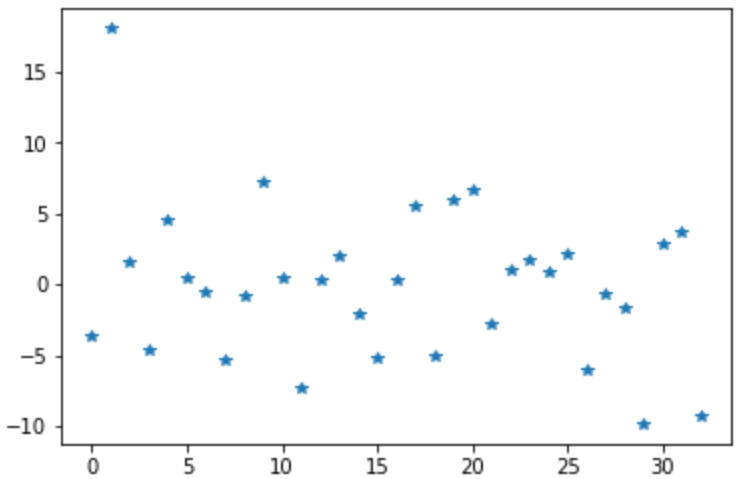
Subtask 3: Results of PCA

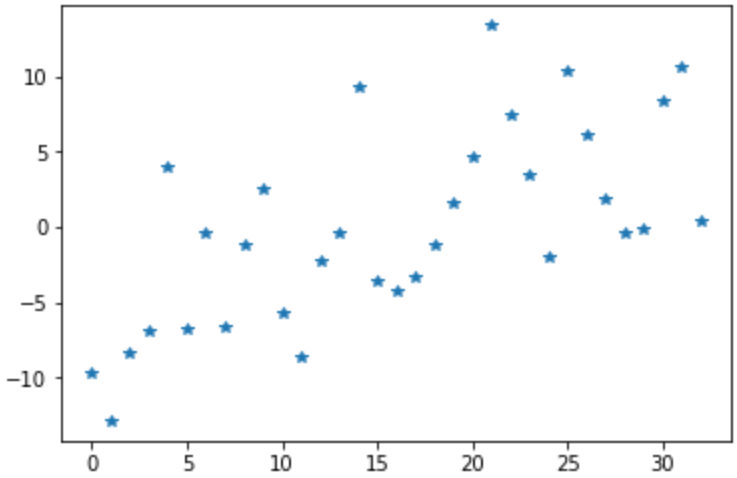
We have plotted the final feature space individually to study the variance and the level of spread

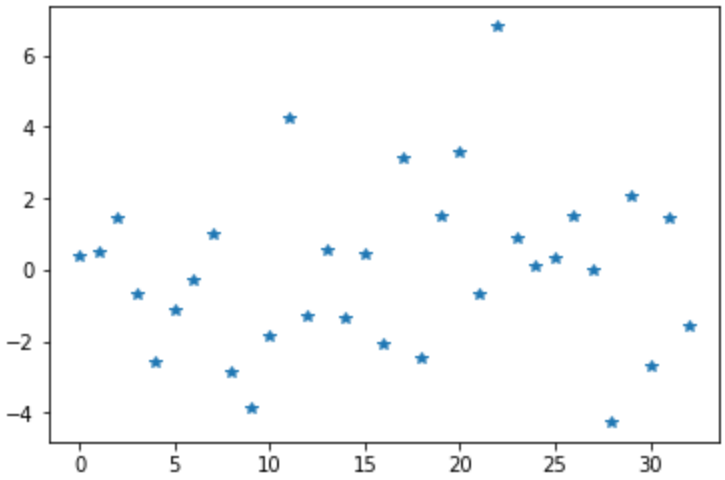
PCA results for CGM values:


12.5
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
0
5
10
15
20
25
30


15
10
5
0
−5
−10
0
5
10
15
20
25
30


15
10
5
0
−5
−10
0
5
10
15
20
25
30

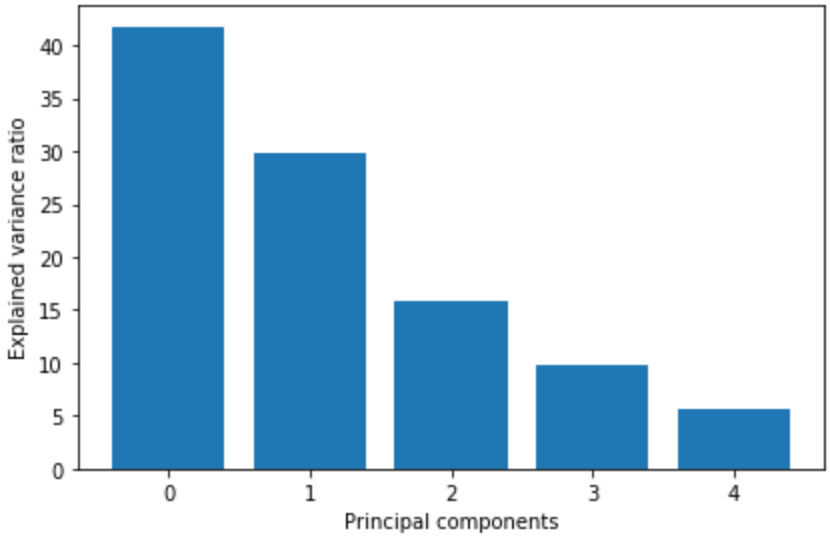


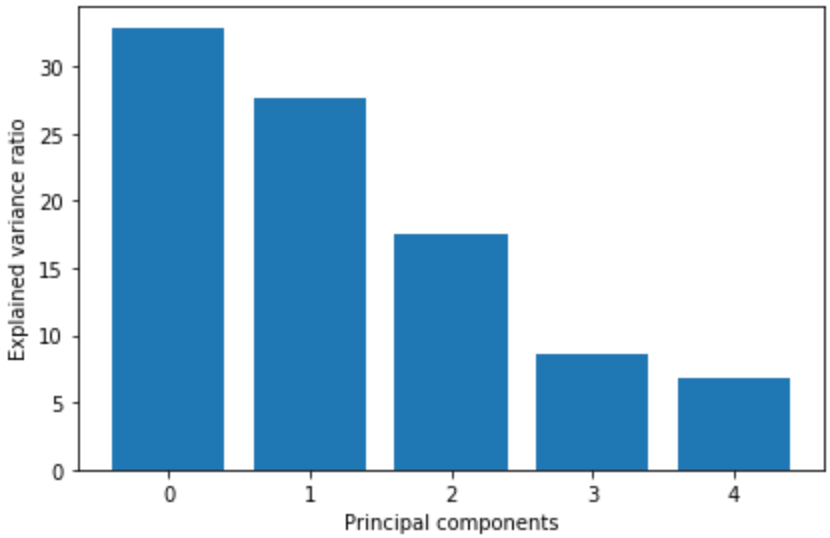


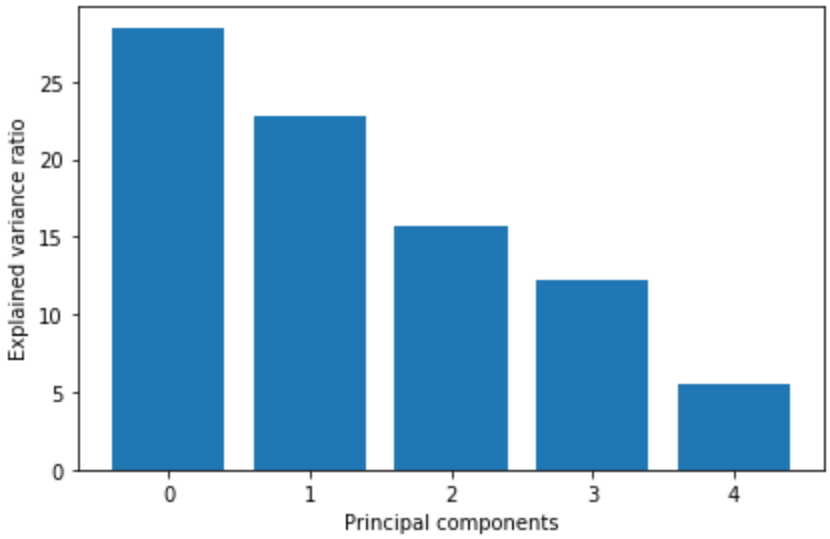
➔ As one can observe here the variance of data decreases as we come down in the order of principal components. First component preserves more variance of the data

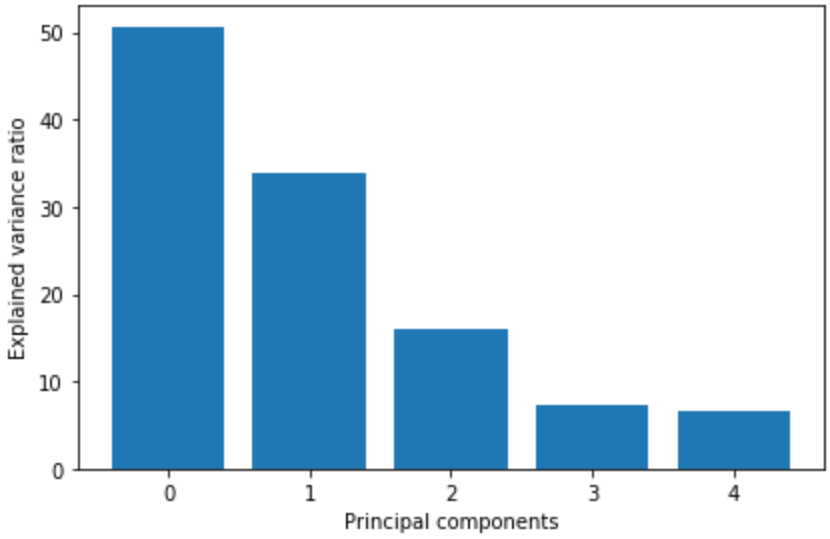
Subtask 4: Was doing PCA helpful?

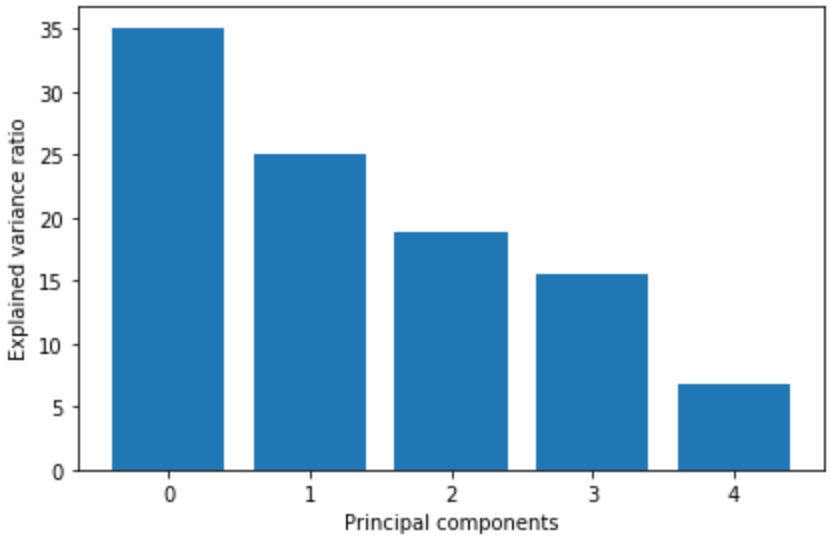
PCA was very helpful in reduction of the feature space as we were able to identify the most important features from a set of 120 features which otherwise would have caused us an overfitted model as most of the features could only express less than 2% of the variance. From the below bar plots you can observe that only 5 components are preserving nearly 80% of variance.






Explained variance ratio
0
5
10
15
20
25
0
1
2
3
4
Principal components


Explained variance ratio
0
10
20
30
40
50
0
1
2
3
4
Principal components



REFERENCES:

1)https://pandas.pydata.org/pandas-docs/stable/

2)https://docs.scipy.org/doc/

3)https://towardsdatascience.com/understanding-pca-fae3e243731d