CSE 572-Data Mining (Fall 2019)
Project Phase I Report
Group 8

Submitted to:

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

Submitted By:

Smruti Sudha Dash
Aditi Panigrahi
Diksha Thakur
Sanika Shah

Abstract

The project involves extracting four types of features from a given dataset of CGM level and the corresponding time stamp. PCA is performed on the resulting feature matrix to get the top 5 features which are helpful in deciding the meal time of the person based on CGM level.

Keywords : Feature, RMS, Standard Deviation, FFT, CGM, feature matrix, PCA

I. Introduction

This project is a part of Data Mining course as a requirement in the session of Fall 2019 at Arizona State University. The goal of this project is to detect and predict meals based on the blood glucose level of Diabetic patients who use automated Insulin injection methods depending on their blood glucose level.

Terminologies:

1. Feature: A feature is any term in the document/dataset which is specific to it and helps in distinguishing the document/dataset from others.
2. RMS: Root Mean Square (RMS) , also known as the quadratic mean, is the square root of the mean of the squares of a set of numbers.
3. Standard Deviation: Standard deviation is a measure of the amount of variation or dispersion of a set of values.
4. FFT: Fast Fourier Transform (FFT) is an algorithm that calculates the various frequency components of a recorded signal. This feature includes peak detection with automatic calculation of associated distances and length values.
5. CGM: Continuous Glucose Monitoring (CGM) tracks the glucose level in the blood in the day and night. CGM is calculated using data sensor and transmitter.
6. Feature Matrix: Feature matrix gives us the values of all the features corresponding to every data point.
7. PCA: Principal Component Analysis is a dimensionality reduction technique that projects the given feature matrix onto another dimensionality with reduced number of features/dimensions by choosing the top k features/dimensions

II. Team members (Group - 8)

1. Smruti Sudha Dash (ssdash@asu.edu)
2. Aditi Panigrahi (apanigr2@asu.edu)
3. Diksha Thakur (dthakur2@asu.edu)
4. Sanika Shah (ssshah33@asu.edu)

III. Project Phase I

The phase I of the Project involved identifying features from the raw data that could potentially serve as good features for conducting PCA in order to detect a meal from blood sugar levels of a patient

recorded at a definite interval. In this phase, we only focus on selection of features that would contribute to successfully predict a meal from the blood sugar level data of the subject.

The data is masked data and the time interval between the blood glucose level recordings are 5 mins. Feature selection can be done in many different ways. Here we focus on 4 different types of features namely:

1) Statistical features
2) CGM velocity
3) FFT
4) Power Spectral Density

These features are explained later on.

Before extracting any features and then performing PCA on them to find out the features which contribute most to our variance in the data we need to first pre-process the raw data provided to us.

A. DATA PREPROCESSING

The data used for this assignment is from CGMSeriesLunchPat and CGMDatenumLunchPat where CGMDatenumLunchPat indicates the timestamps and CGMSeriesLunchPat indicates the glucose values corresponding to those timestamps. There were some missing values in the timestamps. Since we know that the interval between timestamps is 5 minutes or 0.003472222 timestamp, we added this value to the previous value in order to get the missing values. The CGMSeriesLunchPat also had some missing values which were populated by using the same values as their previous ones. One of the rows for the person 4 includes 42 columns, we removed the values for the columns after 32 for that row since the data for all others time series consisted of 32 columns and populating missing values for all those rows would be lot of dummy data which may result in incorrect analysis.

B. FOUR TYPES OF FEATURES

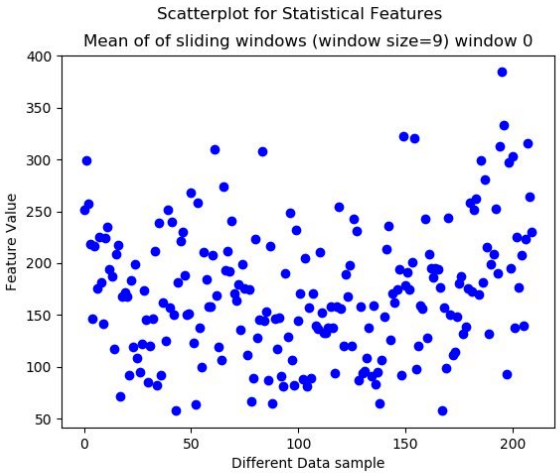
For our assignment we have combined the data points for all 5 people to generate one single file and then did feature extraction. This step was done to improve feature extraction process.

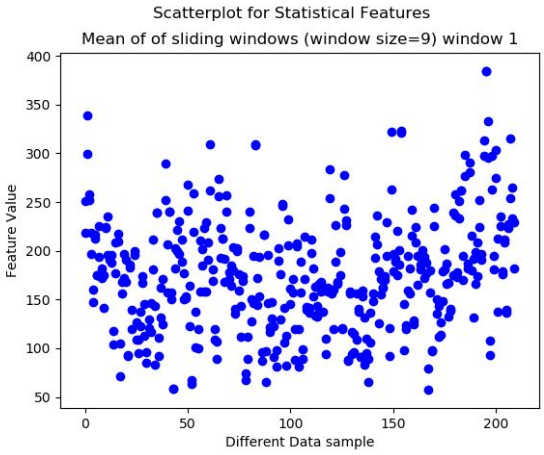
1. **Statistical Features:**

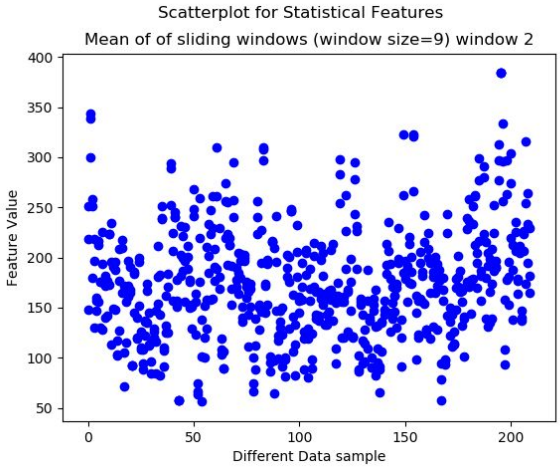
   For statistical features we have calculated very basic attributes of data like mean, standard derivation, minimum and maximum value and how they are changing over interval of 2.5hrs. We created 4 overlapping windows, each window has CGM value for 45 minutes and values for the last 15minutes overlaps with the next window. On these windows we calculated different features as explained further in this report.

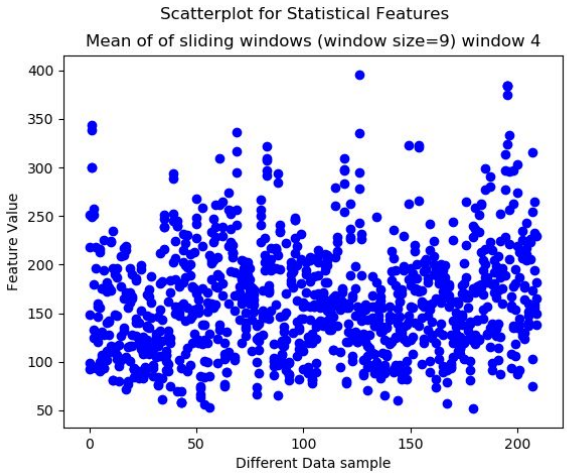
a. Windowed Mean

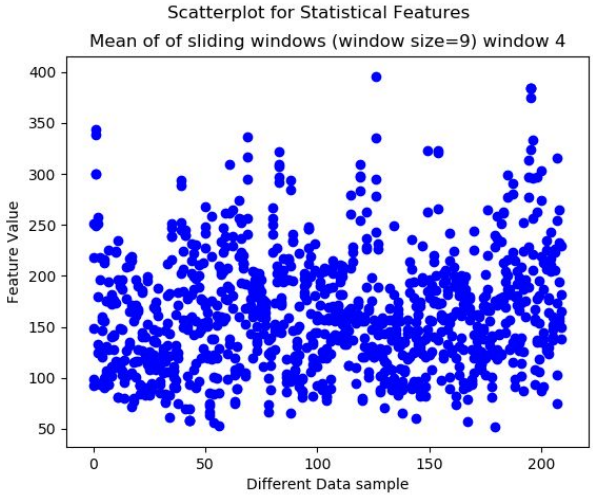
Windowed Mean is average value of CGM values. As our data is time sensitive, we want to see what is average value of CGM value during different intervals. To see this, we calculated average value in each window. We have five windows hence we got 5 features from windowed mean, where each window contributes to one feature or data point. Following are the plots for this feature.



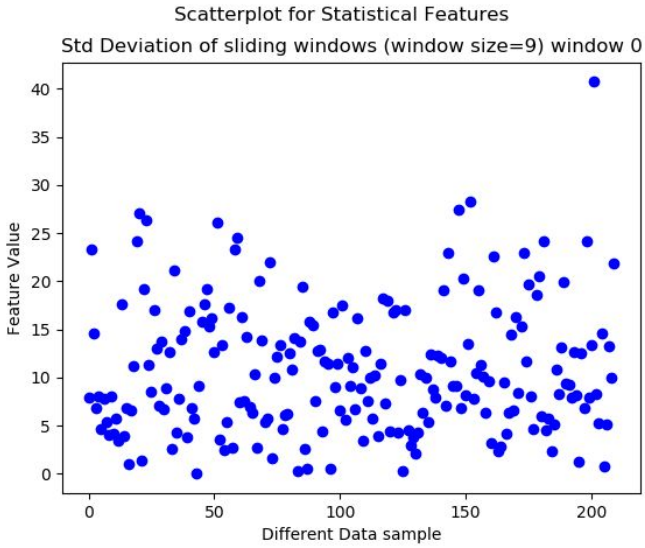


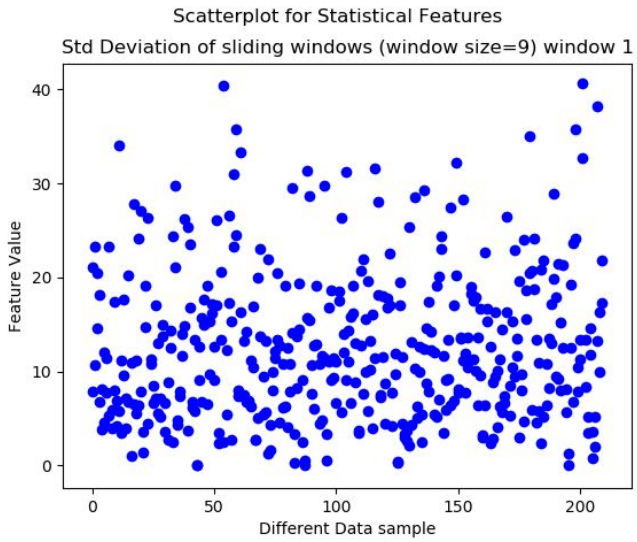


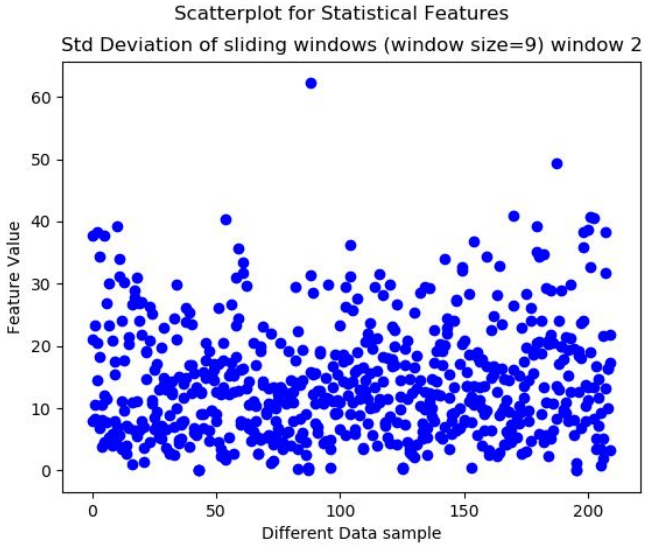


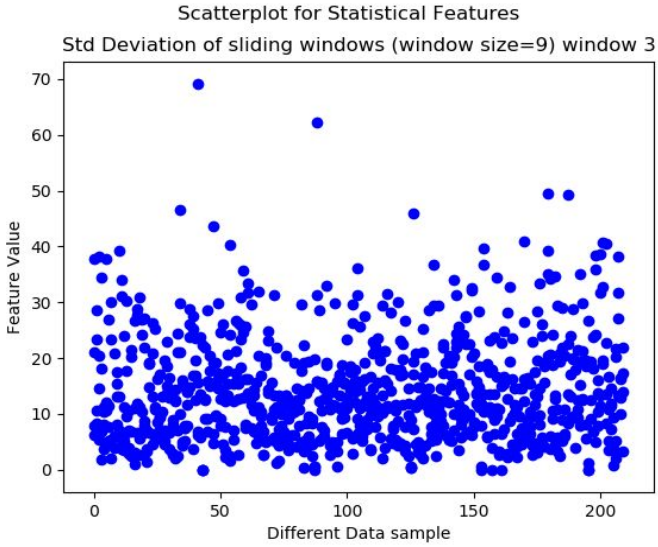


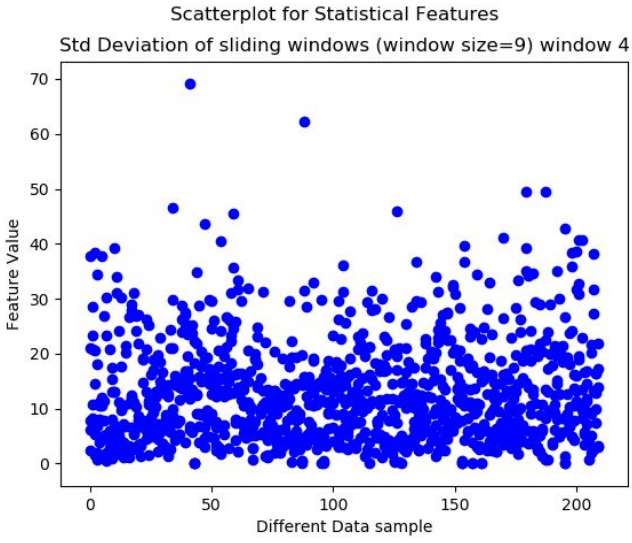
b. Windowed Standard Deviation:







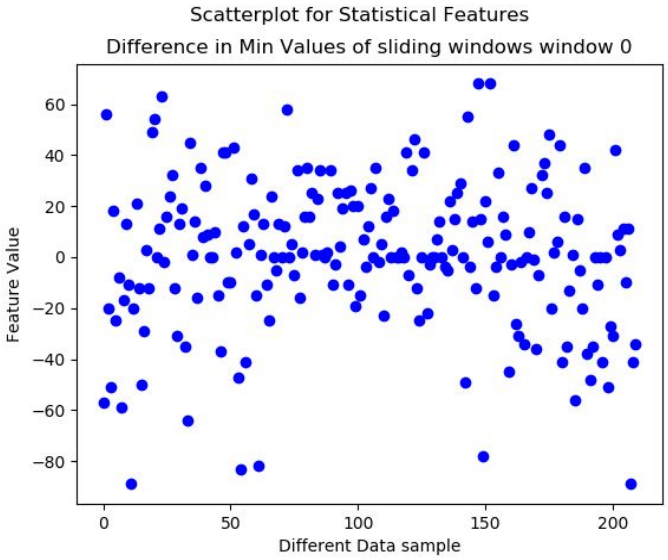


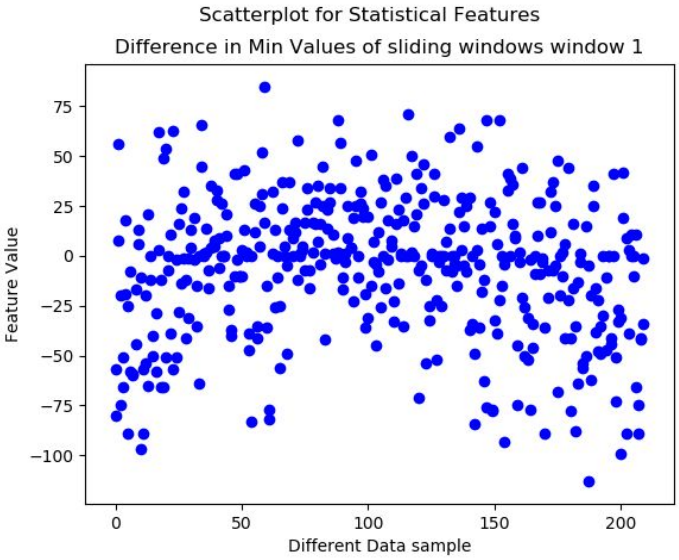


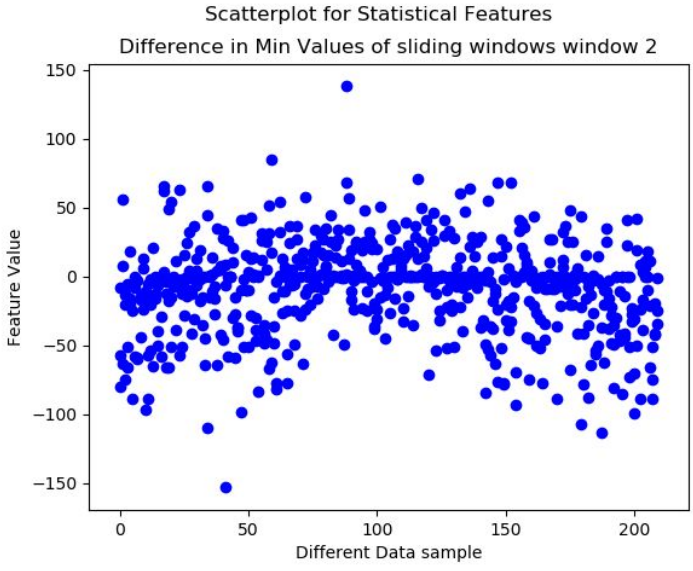
Standard deviation is measure of variation, it gives an estimate how varied data we have. To understand how CGM values are varying locally we calculated standard deviation within overlapping windows. We have five windows each one which contribute to one standard deviation for each time series. We plotted these feature values for each time series and got graphs as shown above.

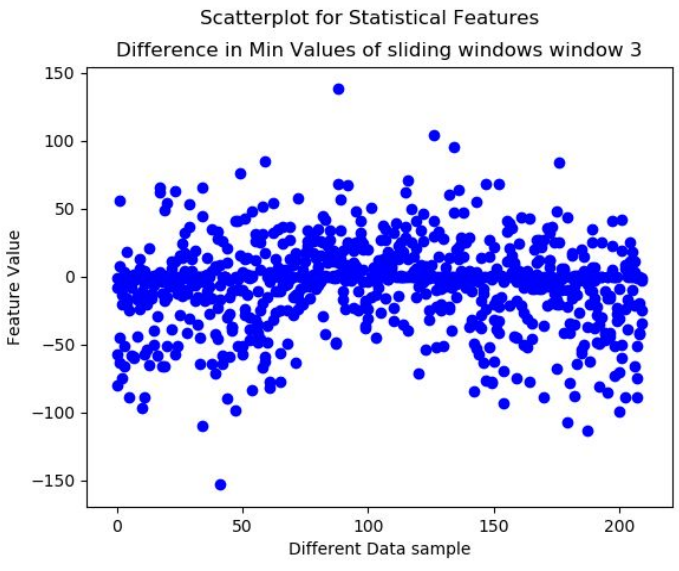
c. Change in Minimum CGM Value of each Window:

Each window has its minimum CGM values, it is important to see how these minimum values are changing while moving from one window to another. We considered both positive and negative change to see the direction of change as well. Difference between each window will correspond to one feature hence creating four features and their scatter is as follows.



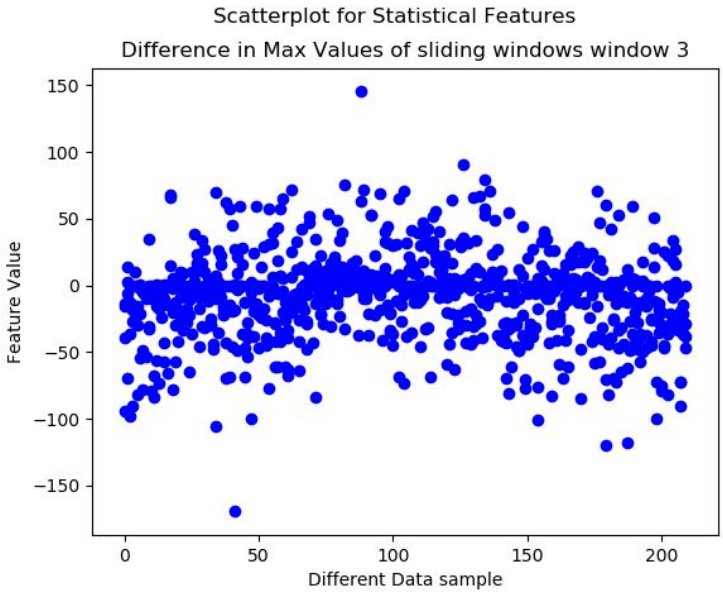


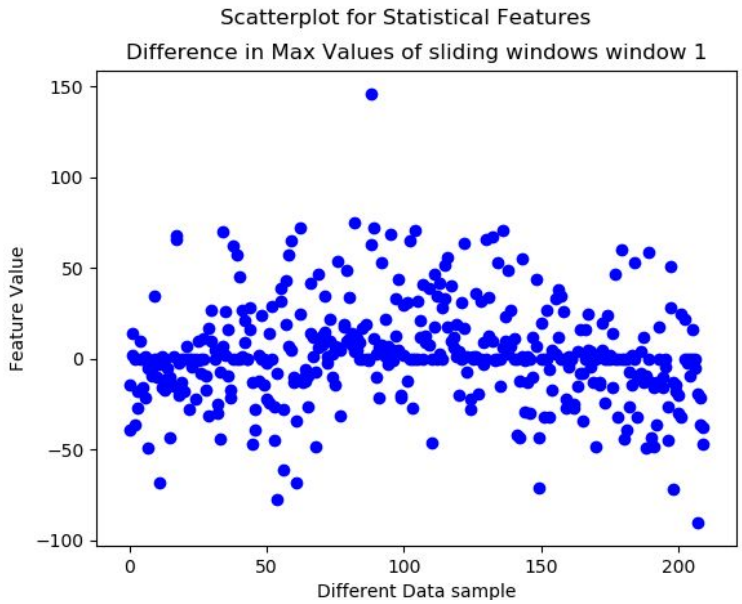


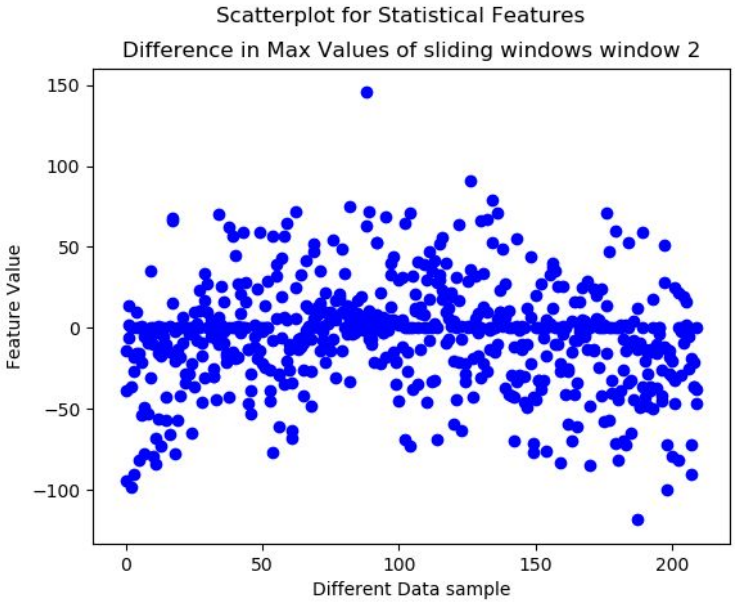


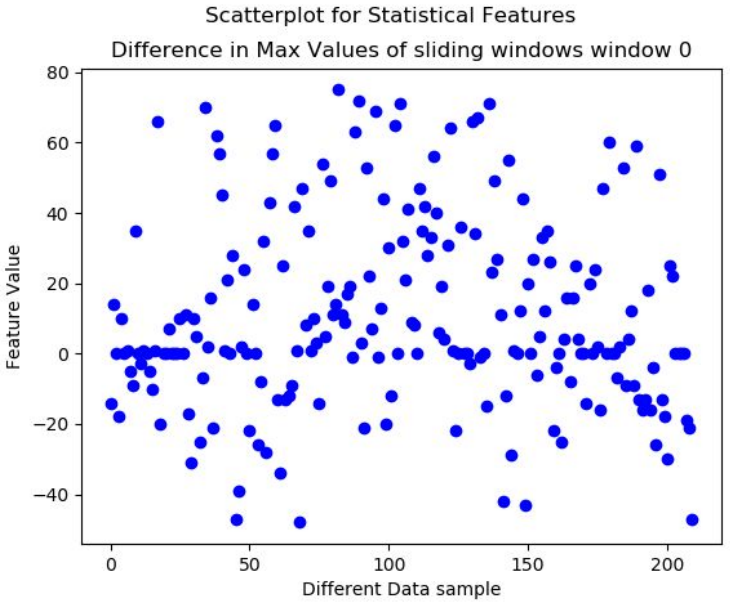
d. Change in Maximum CGM Value of each Window:

Each window has its maximum CGM values, it is important to see how these maximum values are changing while moving from one window to another. We considered both positive and negative change to see the direction of change as well. Difference between each window will correspond to one feature hence creating four features and their scatter is as follows.



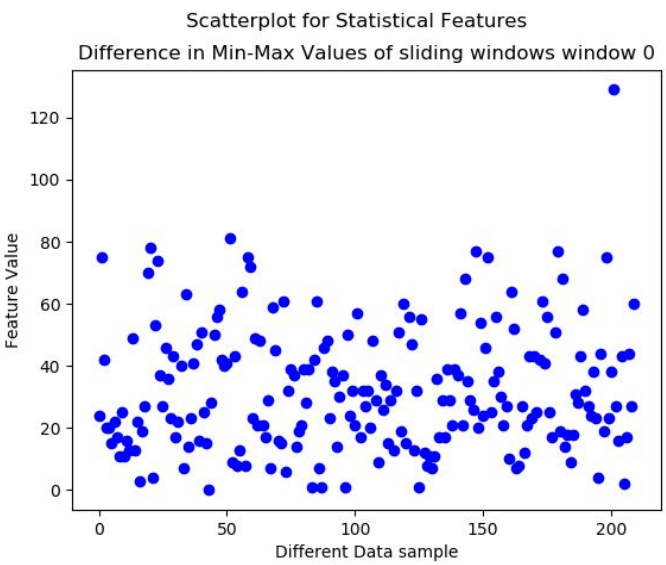


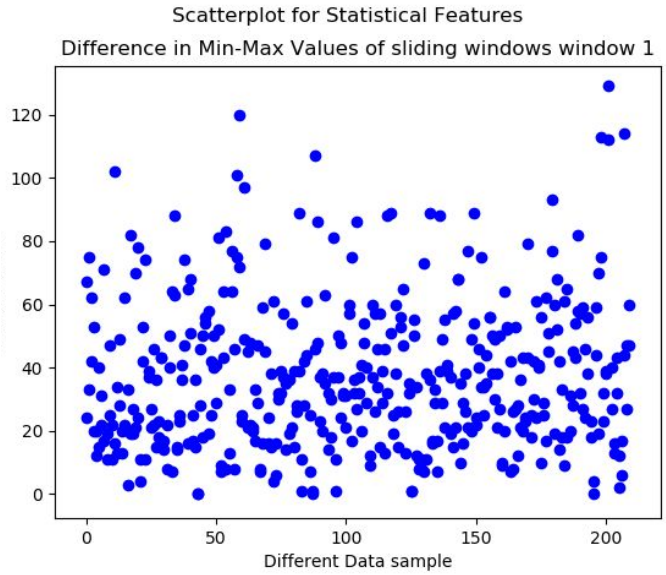


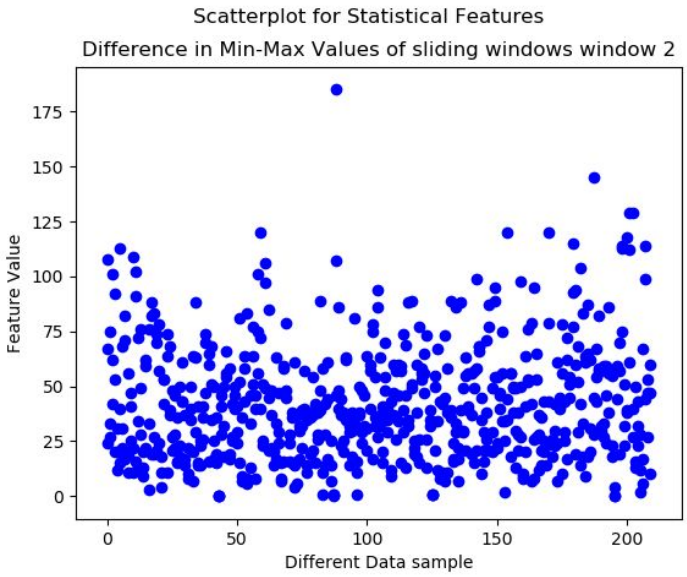


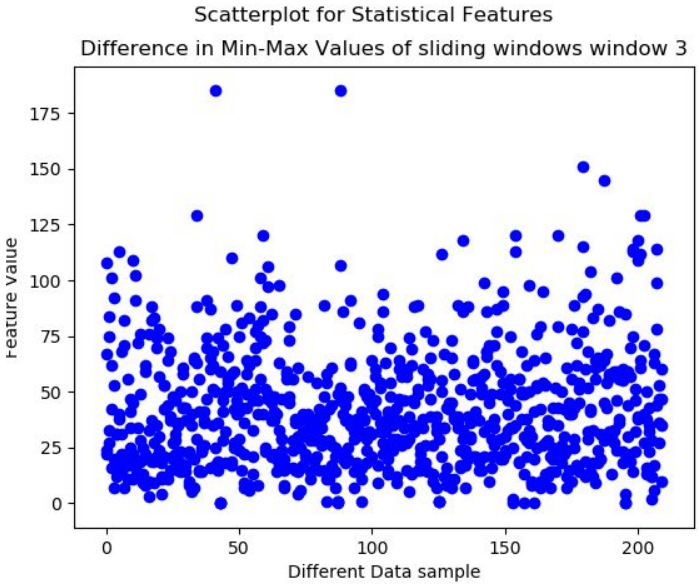
e. Difference between the Minimum and maximum CGM value within each window:

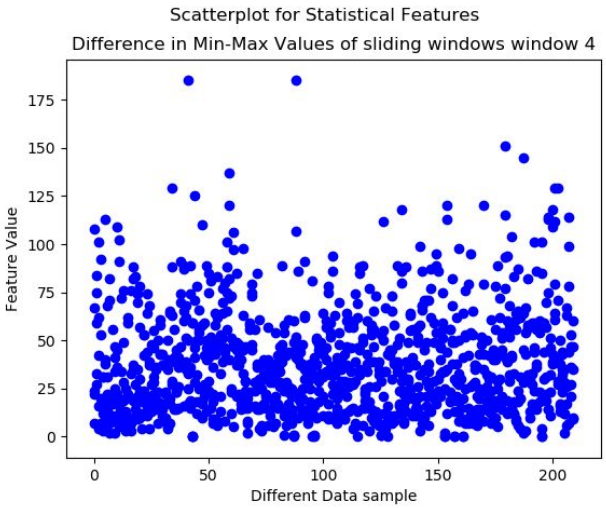
This is another statistical feature where we considered difference between minimum and maximum CGM value inside a given widow. Although standard deviation captures the overall spread of the data points, this feature would capture the absolute difference between maximum and minimum value within each value.











All above features fall into a single category of statistical features as these features include straightforward statistical values and may play an important role in recognising any pattern but they are not enough to recognise any complex relation between data points over a period of time.

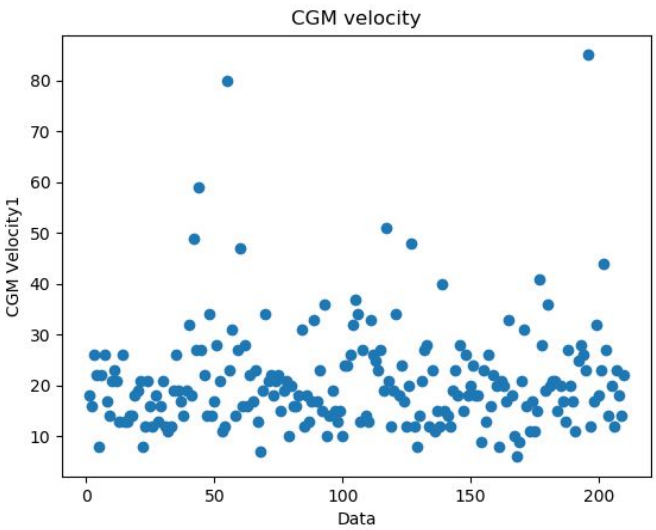
2. CGM Velocity:

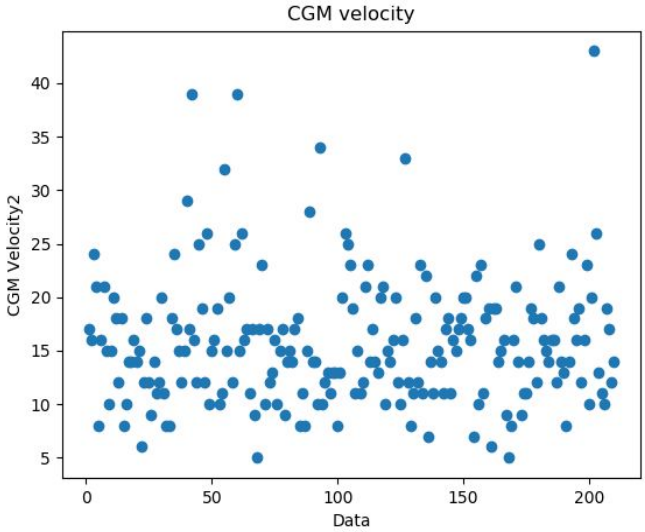
CGM velocity is the change is CGM value with change in time. The CGM velocity is calculated by using the formula :

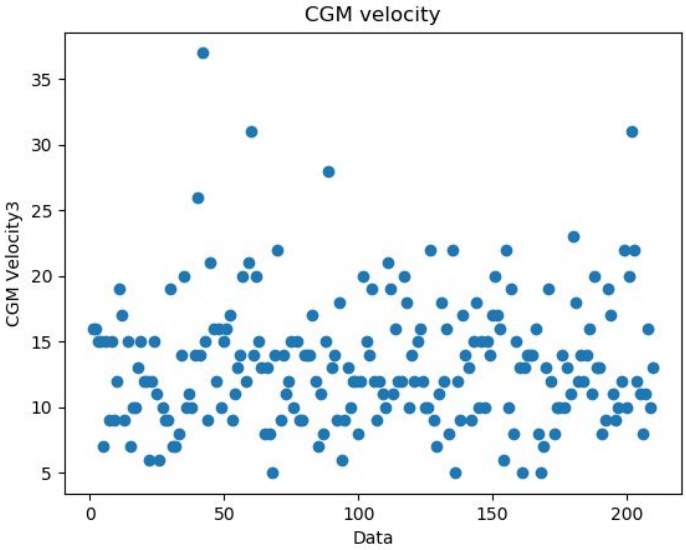
CGM velocity = (CGM[t] - CGM[t-1])/(change in time)

The CGM velocity is calculated for each day(per person) and then the top 3 CGM velocity are taken into consideration as they record the maximum change in CGM data in a particular time series.

CGM velocity could serve as an important feature in detecting the meal plan as it records when the maximum increase or decrease in glucose level. So, the time when meal is taken the velocity could rise suddenly which may be indicated by the CGM velocity.



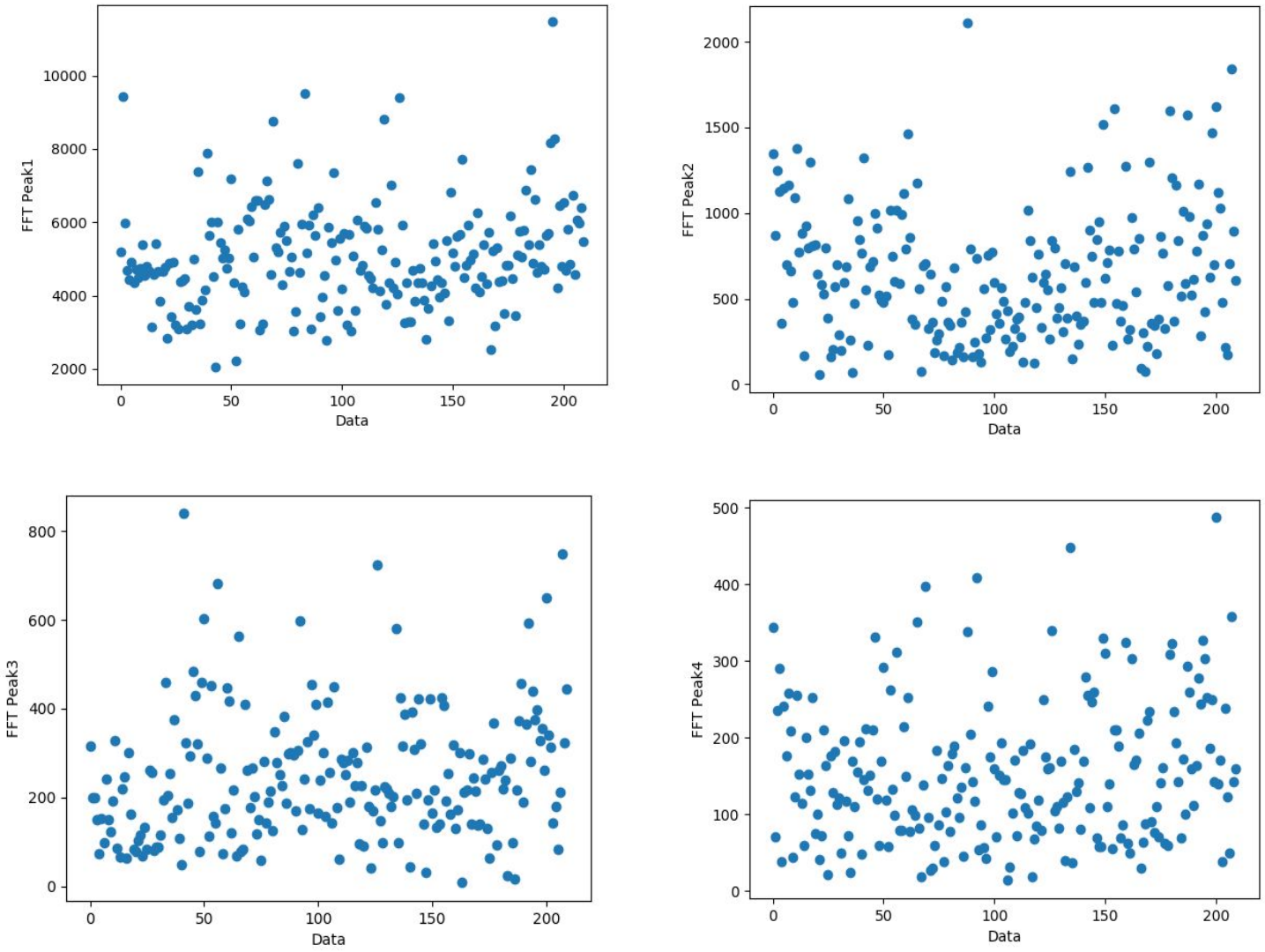




Scatter Plots for CGM velocity

3. FFT (Fourier Fast Transform):

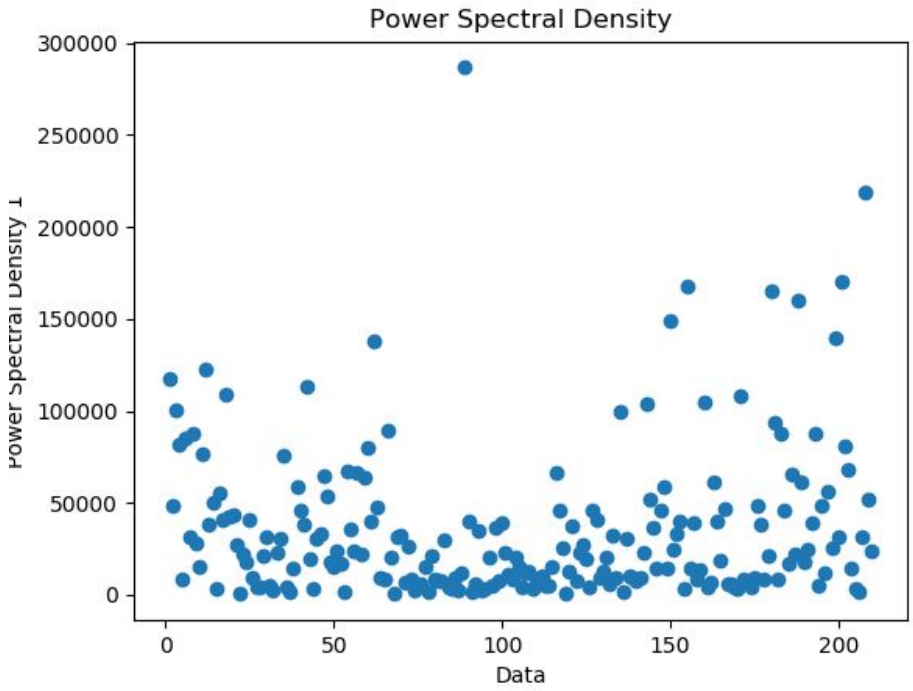
A FFT function calculates the Discrete Fourier Transform for the given data. DFT converts an original time series data from the time domain to the frequency domain. It gives us the most frequent values in the cgm time series data and we considered taking the top 4 out of that. We used the fft function from scipy library in python to calculate the DFT for the time-series data and then took the highest 4 values or the peaks as 4 features. As it can be seen in the plots below for fft peaks vs data points we see the feature scatters the data pretty well and we get a good spread across the y axis for our data. Therefore, we think it is a good feature for PCA.



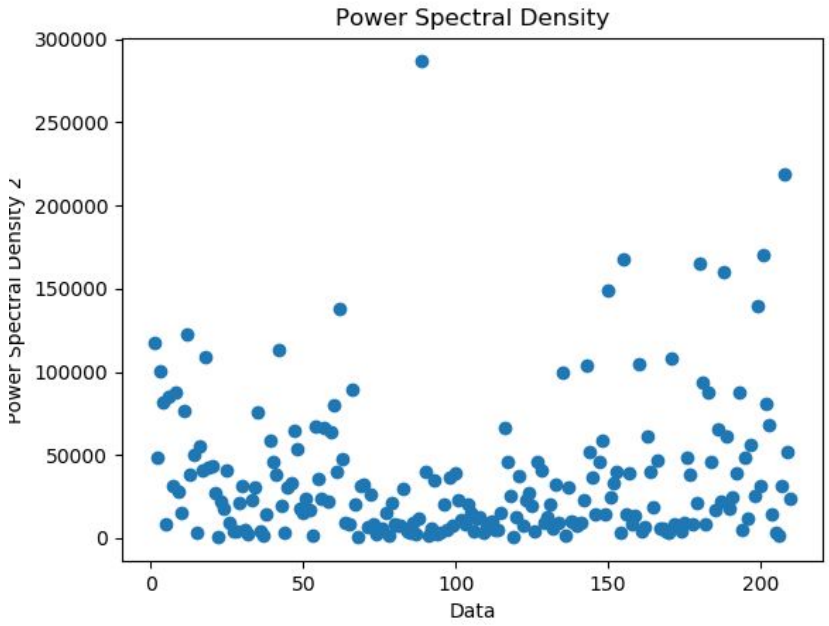
Scatter plots for four features of FFT

4. **Power Spectral Density:**

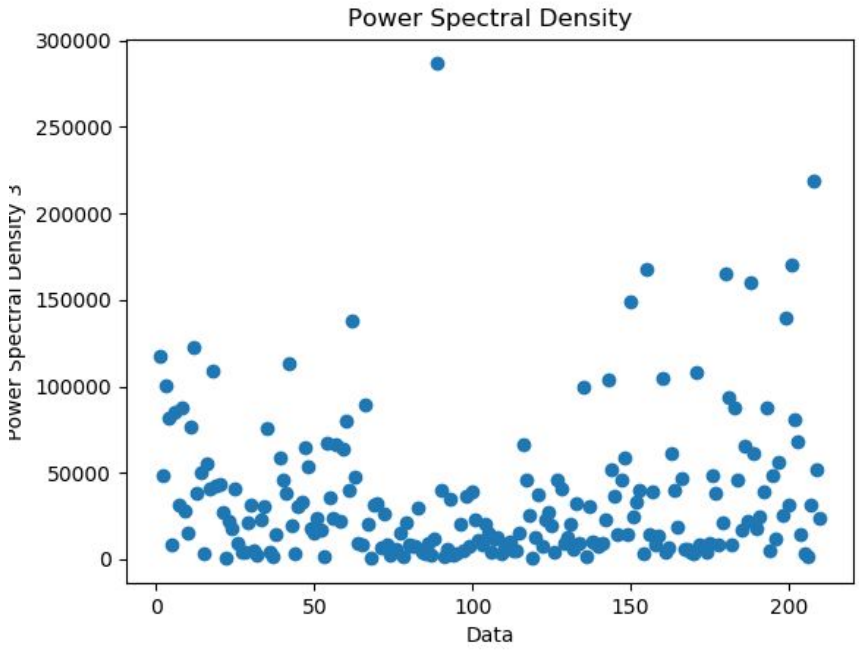
The power spectrum of a time series describes the distribution of power into frequency components composing that signal. Power spectral density function (PSD) shows the strength of the variations as a function of frequency. In other words, it shows at which frequencies, the variations are strong and at which frequencies, the variations are weak. So, it can be used to show the variations in the CGMSeries data over time. High frequency wavenumber of the Power Spectral Density for a particular time series data, show the sharp variations which can be useful to predict when meal was consumed by a person. For our calculation, we have considered the top 3 values of the power spectral density for a given time series data for better clarity of the variations in CGM value.



Scatter Plot for first feature of Power Spectral Density



Scatter Plot for second feature of Power Spectral Density



Scatter Plot for third feature of Power Spectral Density

C. FEATURE MATRIX CREATION FROM SELECTED FEATURES

We have combined the data of all the subjects resulting in 210 data points in total. The total number of features that we finally derived from above mentioned features are 30. 4 features from FFT using the peak values, top 3 CGM velocity values, 5 features each from windowed mean, standard deviation, local minima and local maxima, 3 features from power spectral density resulting in 30 features total. Thus the feature matrix has dimension of 210*30. The feature matrix now has all the above selected and calculated features from the pre-processed data.

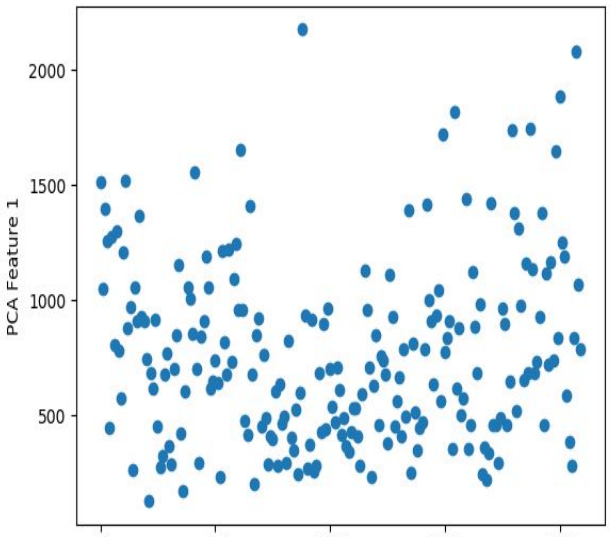
D. PCA ON FEATURE MATRIX:

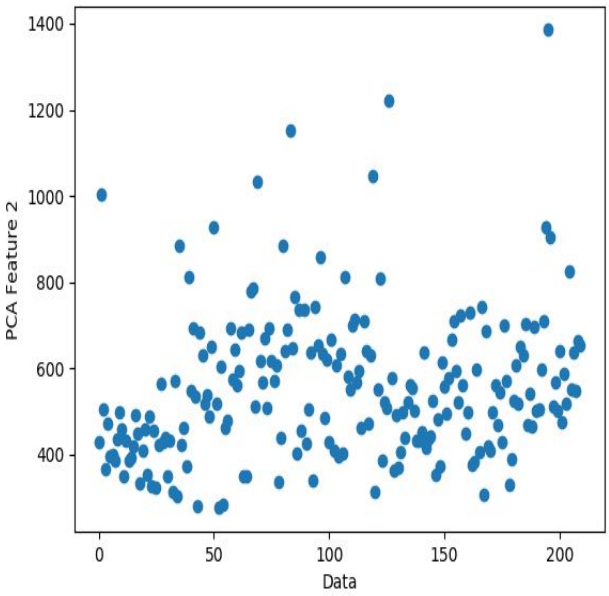
After getting the feature matrix we applied PCA on it to get the principal components. For this we used the pca function from the sklearn library in python. These components are the Eigen Vectors for the PCA performed in the features matrix. PCA projects the data onto another dimensional space with eigen vectors as the dimensions. The eigen vectors represent the variance that we get from each of the features used and tells us how much variance is contributed by the calculated Principal components as well. We took the top 5 eigen vectors/components that maintained 80% variance.
To get the feature projection onto this 5 eigen vectors we apply the following formula:

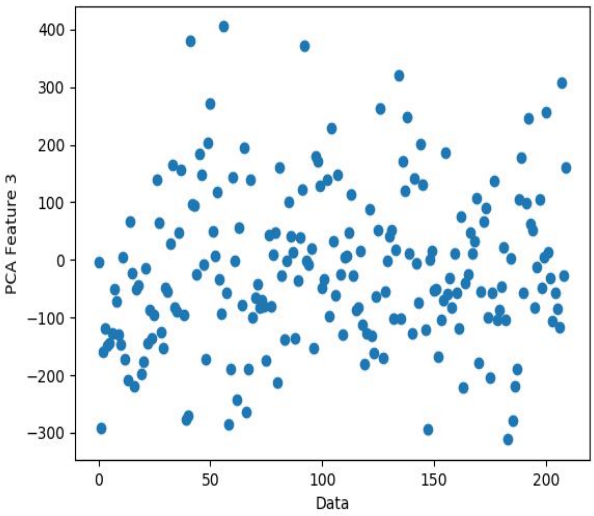
$$\mathbf{X' = X * E^T}$$

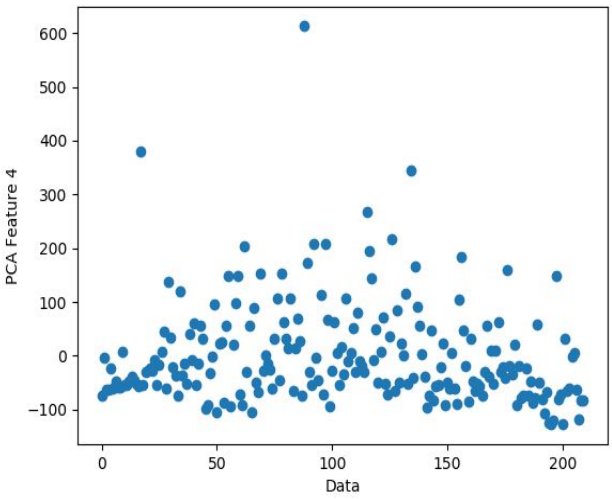
Where X is the feature matrix with **210*30** dimensions
The PCA components has dimension of 30*30 dimensions, we choose top 5 eigen vectors to get the dimension of 5*30. This is the E matrix. The resulting X' matrix has the dimension of 210*5 which is the projection of data points on the new top 5 components. Thus we get the final feature matrix with reduced features i.e 5 features which is called dimensionality reduction.

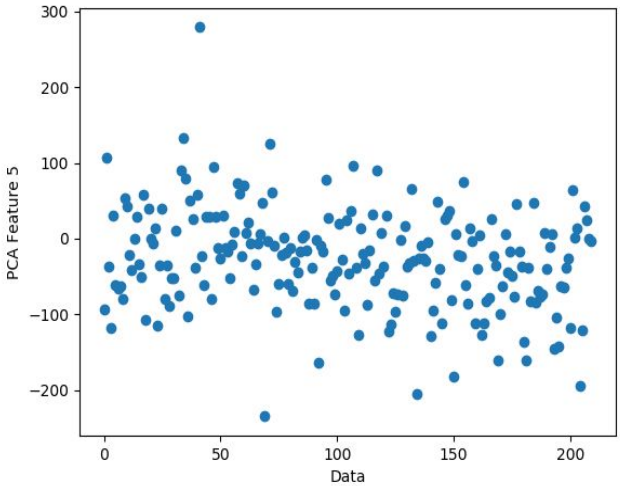
We then plotted the final feature values for our data points as shown below:


PCA Feature 1
2000
1500
1000
500


PCA Feature 2
1400
1200
1000
800
600
400
0
50
100
150
200
Data


PCA Feature 3
400
300
200
100
0
−100
−200
−300
0
50
100
150
200
Data


PCA Feature 4
600
500
400
300
200
100
0
−100
0
50
100
150
200
Data


PCA Feature 5
300
200
100
0
−100
−200
0
50
100
150
200
Data

As we can see from the plots, the final 5 features all have very high variance and also contain the spread of the data which will ultimately help us for classification if we choose to do so. As seen from the explained variance ratio, we can say that the top 5 eigen capture 80% of the variance and that is reflected in the final feature plots for 5 features. To capture 90% of variance we need to take the top 7 features after Principal Component Analysis on the data.

E. TOP 5 FEATURES FROM PCA:

So next, when we plot the top 5 Principal components and the original features contributing to the top 5 Principal components, we see that PSD(Power Spectral Density) is the feature that is contributing most to the variance to the data in the transformed space of the components, After PSD(Power Spectral Density), FFT(Fast Fourier Transform) is another feature contributing significantly to the variance. Compared to these 2 features, statistical features and cgm velocity do not contribute as significantly to the variance in the data. We see that the idea of using FFT(Fast Fourier Transform) and PSD(Power Spectral Density) to represent the features of the data can help us predict and detect meals in the given data.

F. REFERENCES:

[1] Feature Extraction of EEG Signals Using Power Spectral Entropy, Aihua Zhang ; Bin Yang ; Ling Huang, 2008 International Conference on BioMedical Engineering and Informatics
[2] https://en.wikipedia.org/wiki/Spectral_density#Power_spectral_density
[3] https://en.wikipedia.org/wiki/Fast_Fourier_transform