Data Mining project 1

Submitted by Srija Saha(ASU ID:1218443577)

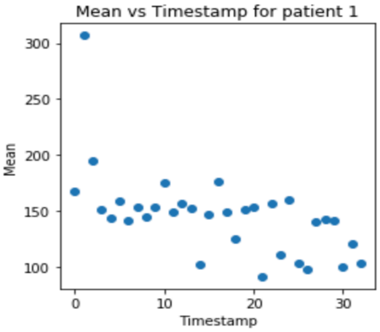
Introduction: The project is about analyzing and finding patterns in the glucose level data in human cells in every 5 mins for 2.5 hours during a lunch meal. I am given five types (glucose level data, timestamps for glucose levels, insulin basal infusion data, insulin bolus infusion data and timestamps for both types of insulin levels) of input files for 5 patients. I have analyzed the glucose level data with respect to time stamps for 5 patients and could draw significant outsights from them.

Tasks:

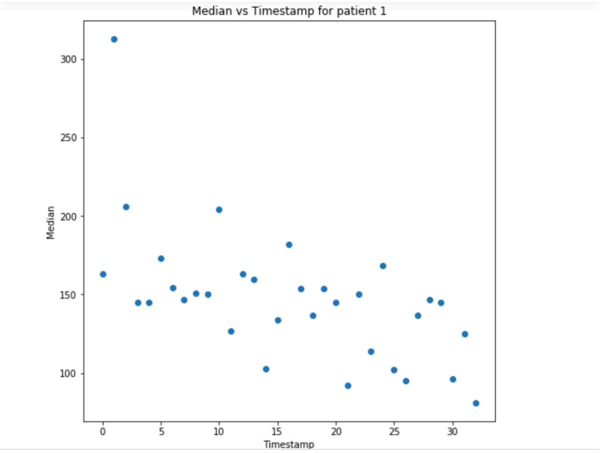
a. I have extracted few important features from CGM data files and plotted each feature with timestamp for patient 1

:

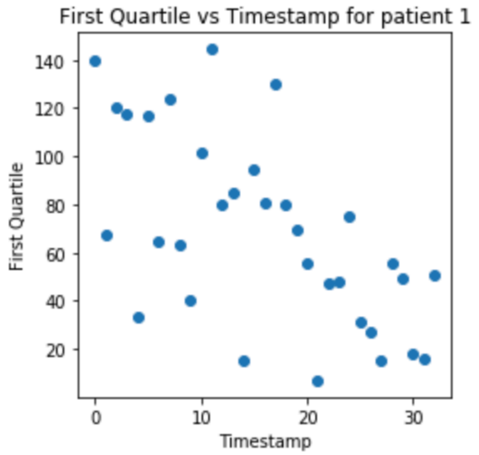
1.Mean –



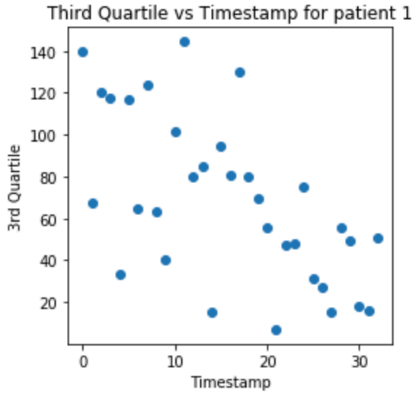
2.Median-



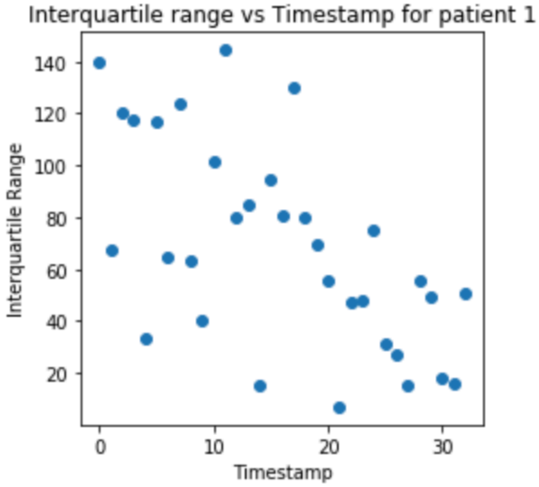
3.First Quartile-



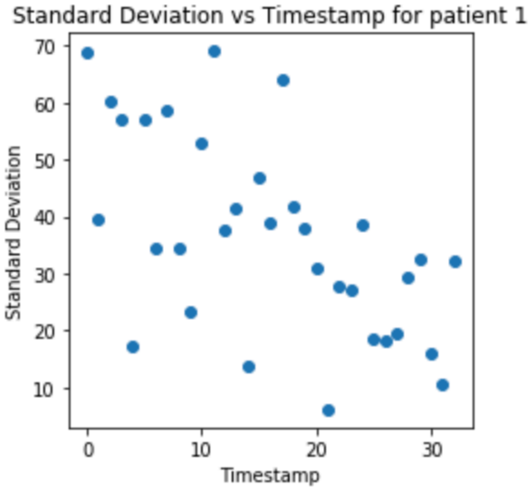
4.Third Quartile-



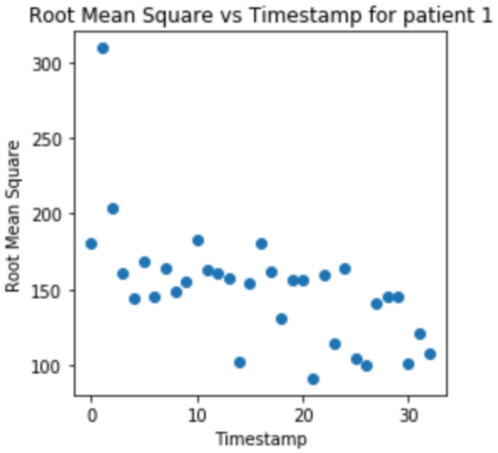
5.Interquartile range–



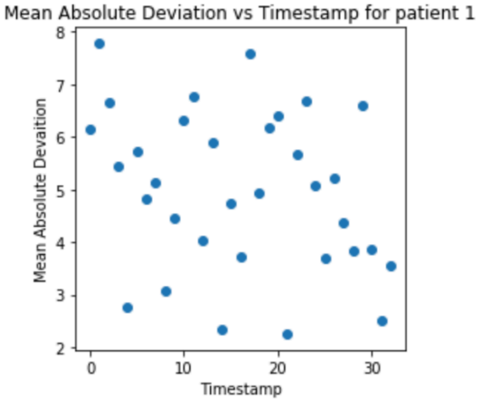
6.Standard Deviation-



7.Root mean square –



8.Mean Absolute Deviation-



b. The reason for choosing the above features explained below-

Mean- The statistical mean is used to understand the central tendency of the data. It is determined by adding all the data points in a population unit and then dividing the total by the number of points. Here I have taken the mean of glucose levels of particular patient in a particular day which have reduced the 33*31 matrix to 33*1.So the mean was used to get an idea average glucose level and whether the glucose level in particular timestamp deviated a lot from the mean value.

Median-Median gives the middle value in a sorted list of data. It splits the list in two parts-list having values less than median and list having values more than median.

Here I have calculated the center value(mean of 15th and 16th column) of glucose levels for a patient in particular day. So we can figure out if the glucose level changed abruptly or smoothly in the middle of the time stamp.

First Quartile and 3rd Quartile- The first quartile (q1) is the median of the lower half of the data set splitted by median. So 25% of the numbers in the data set lie below and about 75% lie above q1.

The third quartile (q3) is the median of the upper half of the data set splitted by median. So 75% of the numbers in the data set lie below q3 and about 25% lie above q3 .

By q1 and q3,we can get the idea of fluctuations in glucose level in small range of data.

Interquartile range(IQR)- Interquartile range is the difference between third and first quartile and it gives the idea about middle 50% data in a data set.So by using IQR, we get an insight about the range of middle 50% glucose data. If the range is higher,then the variation in glucose data around median is higher and if the range is lower, variation in glucose data around median is smaller.

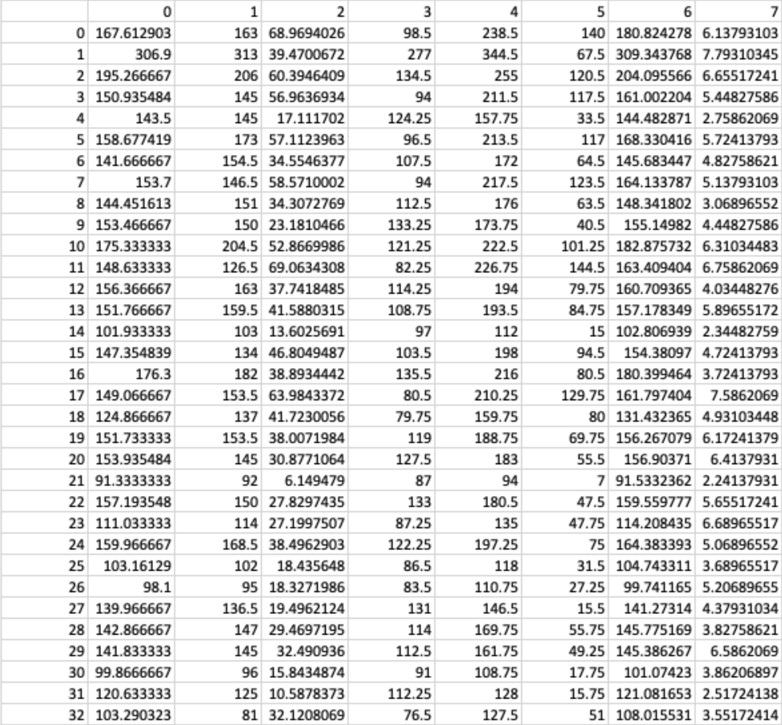
Standard Deviation- Standard deviation gives the measure of deviation of each points in the data set from the mean value. So here I calculated the standard deviation of glucose level of one patient per day to know in which timestamp the glucose level variation is higher than mean value. So we can inspect the patient in that specific time when the variation in glucose level is higher and can take necessary actions accordingly.

Root Mean Square(RMS)- RMS gives the magnitude of set of data. If there is any negative data in glucose level, the positive and negative values do not cancel out as we take the squares of all the values. So we get the actual magnitude of the glucose in time series data.

Mean Absolute Deviation(MAD)- It gives the average distance between each data point and mean. So this is a good measure to understand the fluctuations in glucose /insulin level whole day in the patient.

c. The mean,median,standard deviation,first quartile,third quartile,interquartile range,rms and mean absolute deviation values for first patient in first day is 167.6,163,68,98,238 ,140,180,6.1379 respectively .Similarly we got values these data for other days.If these values for every feature in all days are similar, they are highly correlated,otherwise there is deviation in the glucose or insulin levels in each day for the particular patient.

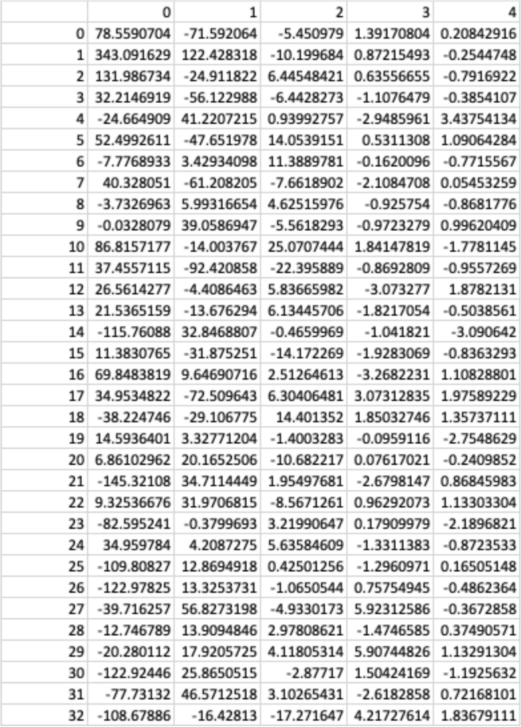
d.I have created a matrix of size 33*8 with 8 features.



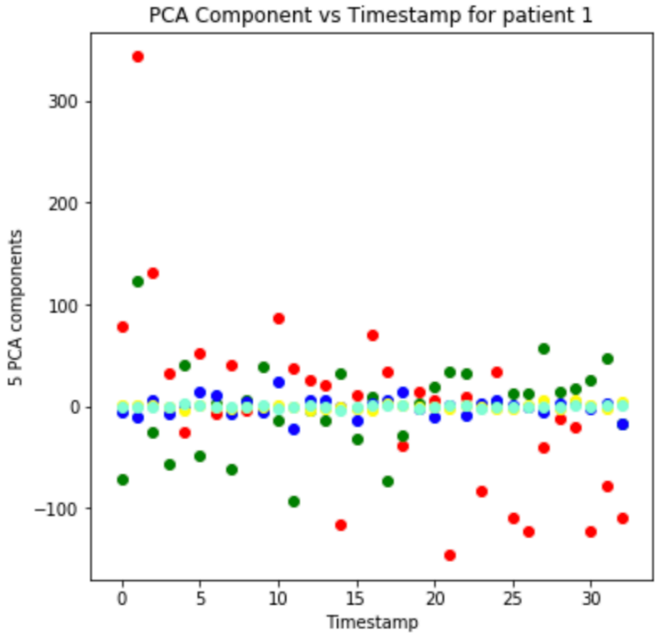
	0	1	2	3	4	5	6	7
0	167.612903	163	68.9694026	98.5	238.5	140	180.824278	6.13793103
1	306.9	313	39.4700672	277	344.5	67.5	309.343768	7.79310345
2	195.266667	206	60.3946409	134.5	255	120.5	204.095566	6.65517241
3	150.935484	145	56.9636934	94	211.5	117.5	161.002204	5.44827586
4	143.5	145	17.111702	124.25	157.75	33.5	144.482871	2.75862069
5	158.677419	173	57.1123963	96.5	213.5	117	168.330416	5.72413793
6	141.666667	154.5	34.5546377	107.5	172	64.5	145.683447	4.82758621
7	153.7	146.5	58.5710002	94	217.5	123.5	164.133787	5.13793103
8	144.451613	151	34.3072769	112.5	176	63.5	148.341802	3.06896552
9	153.466667	150	23.1810466	133.25	173.75	40.5	155.14982	4.44827586
10	175.333333	204.5	52.8669986	121.25	222.5	101.25	182.875732	6.31034483
11	148.633333	126.5	69.0634308	82.25	226.75	144.5	163.409404	6.75862069
12	156.366667	163	37.7418485	114.25	194	79.75	160.709365	4.03448276
13	151.766667	159.5	41.5880315	108.75	193.5	84.75	157.178349	5.89655172
14	101.933333	103	13.6025691	97	112	15	102.806939	2.34482759
15	147.354839	134	46.8049487	103.5	198	94.5	154.38097	4.72413793
16	176.3	182	38.8934442	135.5	216	80.5	180.399464	3.72413793
17	149.066667	153.5	63.9843372	80.5	210.25	129.75	161.797404	7.5862069
18	124.866667	137	41.7230056	79.75	159.75	80	131.432365	4.93103448
19	151.733333	153.5	38.0071984	119	188.75	69.75	156.267079	6.17241379
20	153.935484	145	30.8771064	127.5	183	55.5	156.90371	6.4137931
21	91.3333333	92	6.149479	87	94	7	91.5332362	2.24137931
22	157.193548	150	27.8297435	133	180.5	47.5	159.559777	5.65517241
23	111.033333	114	27.1997507	87.25	135	47.75	114.208435	6.68965517
24	159.966667	168.5	38.4962903	122.25	197.25	75	164.383393	5.06896552
25	103.16129	102	18.435648	86.5	118	31.5	104.743311	3.68965517
26	98.1	95	18.3271986	83.5	110.75	27.25	99.741165	5.20689655
27	139.966667	136.5	19.4962124	131	146.5	15.5	141.27314	4.37931034
28	142.866667	147	29.4697195	114	169.75	55.75	145.775169	3.82758621
29	141.833333	145	32.490936	112.5	161.75	49.25	145.386267	6.5862069
30	99.8666667	96	15.8434874	91	108.75	17.75	101.07423	3.86206897
31	120.633333	125	10.5878373	112.25	128	15.75	121.081653	2.51724138
32	103.290323	81	32.1208069	76.5	127.5	51	108.015531	3.55172414

e. As I measured the correlation value of a particular dataset,I observed there are many highly correlated data present in the data set of glucose level and we do not need data of similar characteristics for data mining.The data which have larger variation are only useful for finding important patterns in data. So I applied PCA(Prinicipal Component Analysis) method to extract only the important features.So the dimension of 33*8 matrix reduced to 33*5 which only contains the important component of data which contains valuable information about the glucose levels in patients.

The featured matrix 33*5 after applying PCA extraction mechanism



	0	1	2	3	4
0	78.5590704	-71.592064	-5.450979	1.39170804	0.20842916
1	343.091629	122.428318	-10.199684	0.87215493	-0.2544748
2	131.986734	-24.911822	6.44548421	0.63556655	-0.7916922
3	32.2146919	-56.122988	-6.4428273	-1.1076479	-0.3854107
4	-24.664909	41.2207215	0.93992757	-2.9485961	3.43754134
5	52.4992611	-47.651978	14.0539151	0.5311308	1.09064284
6	-7.7768933	3.42934098	11.3889781	-0.1620096	-0.7715567
7	40.328051	-61.208205	-7.6618902	-2.1084708	0.05453259
8	-3.7326963	5.99316654	4.62515976	-0.925754	-0.8681776
9	-0.0328079	39.0586947	-5.5618293	-0.9723279	0.99620409
10	86.8157177	-14.003767	25.0707444	1.84147819	-1.7781145
11	37.4557115	-92.420858	-22.395889	-0.8692809	-0.9557269
12	26.5614277	-4.4086463	5.83665982	-3.073277	1.8782131
13	21.5365159	-13.676294	6.13445706	-1.8217054	-0.5038561
14	-115.76088	32.8468807	-0.4659969	-1.041821	-3.090642
15	11.3830765	-31.875251	-14.172269	-1.9283069	-0.8363293
16	69.8483819	9.64690716	2.51264613	-3.2682231	1.10828801
17	34.9534822	-72.509643	6.30406481	3.07312835	1.97589229
18	-38.224746	-29.106775	14.401352	1.85032746	1.35737111
19	14.5936401	3.32771204	-1.4003283	-0.0959116	-2.7548629
20	6.86102962	20.1652506	-10.682217	0.07617021	-0.2409852
21	-145.32108	34.7114449	1.95497681	-2.6798147	0.86845983
22	9.32536676	31.9706815	-8.5671261	0.96292073	1.13303304
23	-82.595241	-0.3799693	3.21990647	0.17909979	-2.1896821
24	34.959784	4.2087275	5.63584609	-1.3311383	-0.8723533
25	-109.80827	12.8694918	0.42501256	-1.2960971	0.16505148
26	-122.97825	13.3253731	-1.0650544	0.75754945	-0.4862364
27	-39.716257	56.8273198	-4.9330173	5.92312586	-0.3672858
28	-12.746789	13.9094846	2.97808621	-1.4746585	0.37490571
29	-20.280112	17.9205725	4.11805314	5.90744826	1.13291304
30	-122.92446	25.8650515	-2.87717	1.50424169	-1.1925632
31	-77.73132	46.5712518	3.10265431	-2.6182858	0.72168101
32	-108.67886	-16.42813	-17.271647	4.21727614	1.83679111



f.PCA is a feature extraction mechanism which only extracts the variant data from a large pool of data set thereby reducing the dimension of the data set. Here I have taken top 5 features from the dataset out of 8 features which contains maximum information about the glucose levels.

References:

https://www.techopedia.com/definition/26136/statistical-mean
https://www.thoughtco.com/what-are-first-and-third-quartiles-3126235
https://www.khanacademy.org/math/cc-sixth-grade-math/cc-6th-data-statistics/cc-6-mad/v/mean-absolute-deviation
https://stackoverflow.com/questions/40963659/root-mean-square-of-a-function-in-python
https://www.geeksforgeeks.org/absolute-deviation-and-absolute-mean-deviation-using-numpy-python/
https://towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60
https://www.dezyre.com/data-science-in-python-tutorial/principal-component-analysis-tutorial
https://medium.com/@kasiarachuta/importing-and-exporting-csv-files-in-python-7fa6e4d9f408
https://www.youtube.com/watch?v=kApPBm1YsqU
https://cmdlinetips.com/2018/04/how-to-concatenate-arrays-in-numpy/
https://www.youtube.com/watch?v=a9UrKTVEeZA