Shi Yu Lin 1216287878

Assignment 1

a) Four different types of feature: Glucose level, CGM velocity, moving rms, and FFT

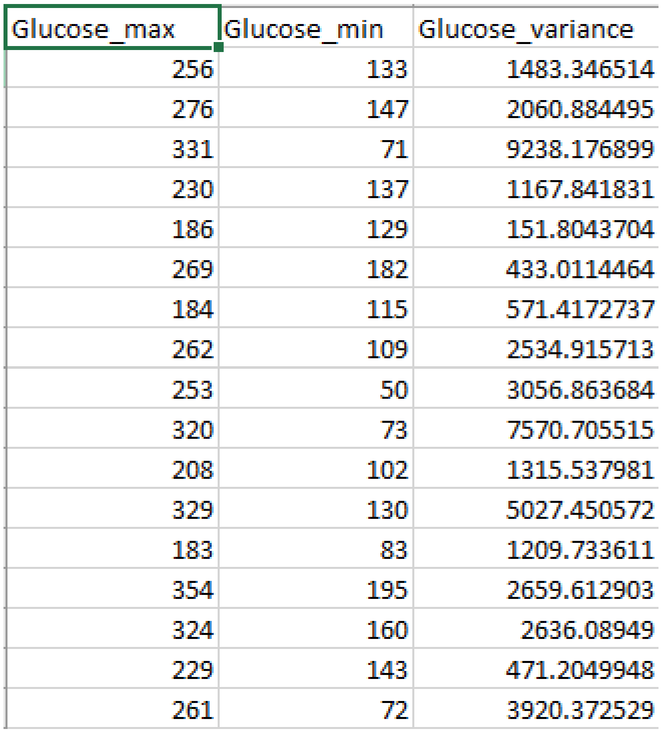
b) Glucose level can be used to check the health status of patient given their MAX, MIN, Variance values. For example, the blood glucose target range for diabetics, according to the American Diabetes Association, should be 90 – 130 mg/dL before meals, and less than 180 mg/dL two hours after meals.

CGM velocity can be used to blood level fluctuations between the current timestamp and the next timestamp, which can come in handy to check the health status of the patient.

Moving rms is defined as the rms of the current index up to current value plus an offset, in this case the offset is 5. This statistical value is useful if the model is a normal distribution.

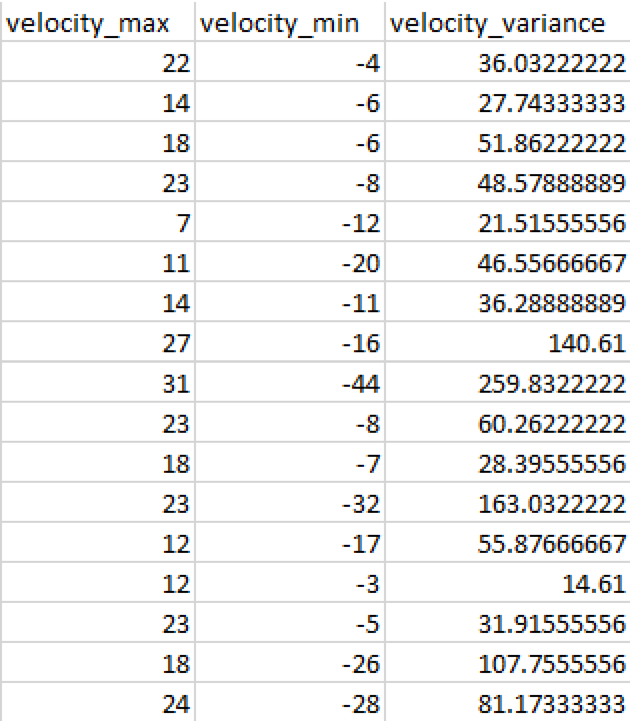
Fast fourier transform is a technique that samples a signal over a period of time and divides it into its frequency components. FFT feature was used to segregate the data based on its frequency.

c) Feature matrix is saved in feature_matrix.csv.



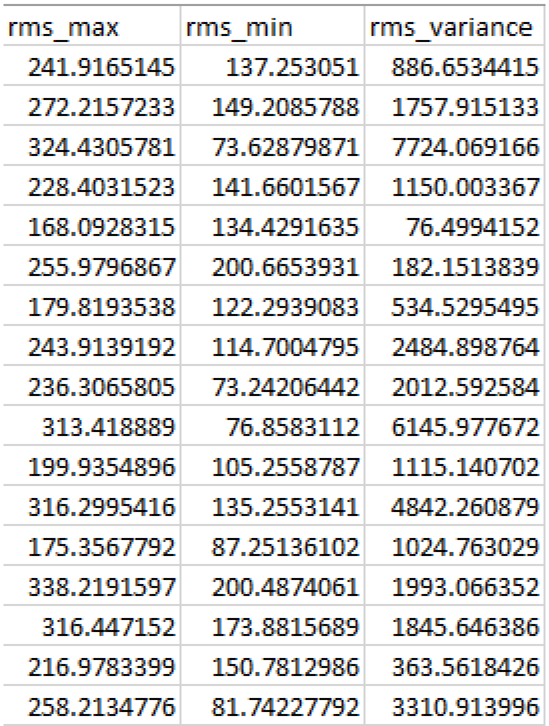
Glucose_max	Glucose_min	Glucose_variance
256	133	1483.346514
276	147	2060.884495
331	71	9238.176899
230	137	1167.841831
186	129	151.8043704
269	182	433.0114464
184	115	571.4172737
262	109	2534.915713
253	50	3056.863684
320	73	7570.705515
208	102	1315.537981
329	130	5027.450572
183	83	1209.733611
354	195	2659.612903
324	160	2636.08949
229	143	471.2049948
261	72	3920.372529

From the table, we can observe that for this subject, there are times that shows abnormal results, for example, the maximum glucose level are higher than 300. For those data, it is likely the subject did not take medicine on time. From my perspective, these are valid and useful features and can be save for future use.



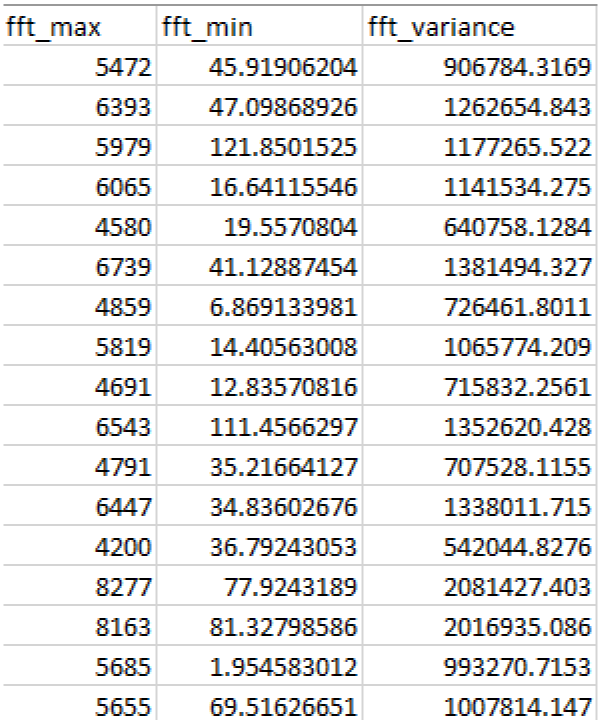
velocity_max	velocity_min	velocity_variance
22	-4	36.03222222
14	-6	27.74333333
18	-6	51.86222222
23	-8	48.57888889
7	-12	21.51555556
11	-20	46.55666667
14	-11	36.28888889
27	-16	140.61
31	-44	259.8322222
23	-8	60.26222222
18	-7	28.39555556
23	-32	163.0322222
12	-17	55.87666667
12	-3	14.61
23	-5	31.91555556
18	-26	107.7555556
24	-28	81.17333333

From the table, we can observe the glucose level variation over given time period. Variance of velocity gives an idea of the range of the fluctuation. However, minor change should be applied to maximum and minimum, I should extract these features with absolute measurement.



rms_max	rms_min	rms_variance
241.9165145	137.253051	886.6534415
272.2157233	149.2085788	1757.915133
324.4305781	73.62879871	7724.069166
228.4031523	141.6601567	1150.003367
168.0928315	134.4291635	76.4994152
255.9796867	200.6653931	182.1513839
179.8193538	122.2939083	534.5295495
243.9139192	114.7004795	2484.898764
236.3065805	73.24206442	2012.592584
313.418889	76.8583112	6145.977672
199.9354896	105.2558787	1115.140702
316.2995416	135.2553141	4842.260879
175.3567792	87.25136102	1024.763029
338.2191597	200.4874061	1993.066352
316.447152	173.8815689	1845.646386
216.9783399	150.7812986	363.5618426
258.2134776	81.74227792	3310.913996

From the table, apparently, the value does not change drastically, it is because I choose a window of 5, calculating the neighbor values. However, this feature type compensates cgm velocity in a sense that it gives an idea of the average trend of the data, so it is easier to check if the variation is within acceptable range.



fft_max	fft_min	fft_variance
5472	45.91906204	906784.3169
6393	47.09868926	1262654.843
5979	121.8501525	1177265.522
6065	16.64115546	1141534.275
4580	19.5570804	640758.1284
6739	41.12887454	1381494.327
4859	6.869133981	726461.8011
5819	14.40563008	1065774.209
4691	12.83570816	715832.2561
6543	111.4566297	1352620.428
4791	35.21664127	707528.1155
6447	34.83602676	1338011.715
4200	36.79243053	542044.8276
8277	77.9243189	2081427.403
8163	81.32798586	2016935.086
5685	1.954583012	993270.7153
5655	69.51626651	1007814.147

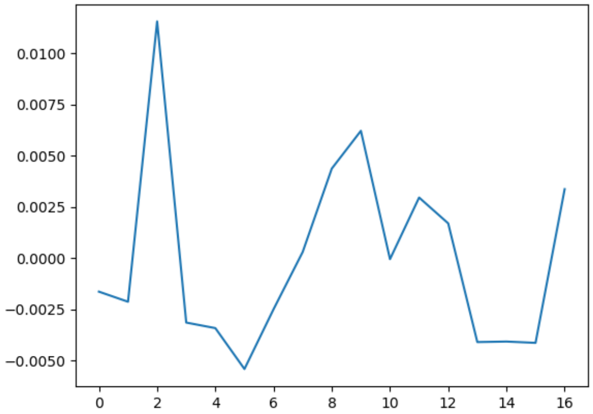
This table gives the information of the FFT of the dataset. Now to think of it, these numbers alone doesn't make any sense. To understand the frequency component of the dataset, it should be observed in a continuous perspective.

d) New derived feature matrix is saved in final_pca.csv.

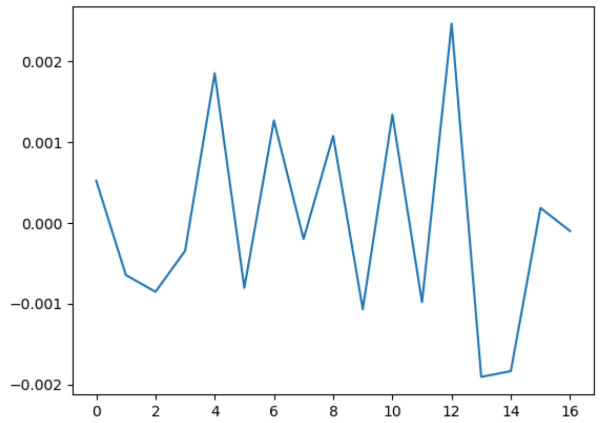
e)

For each PCA component, I plot graphs among each time series.

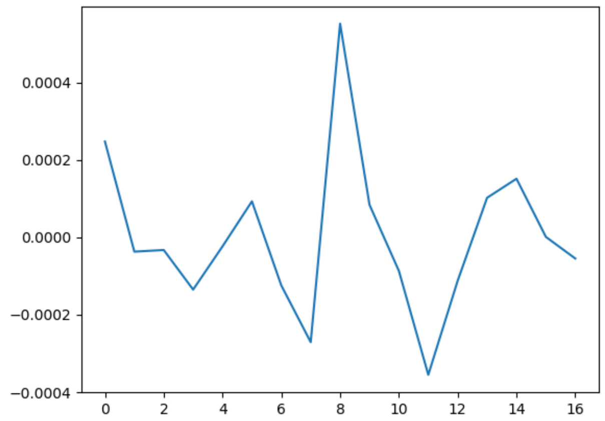
PCA component 1



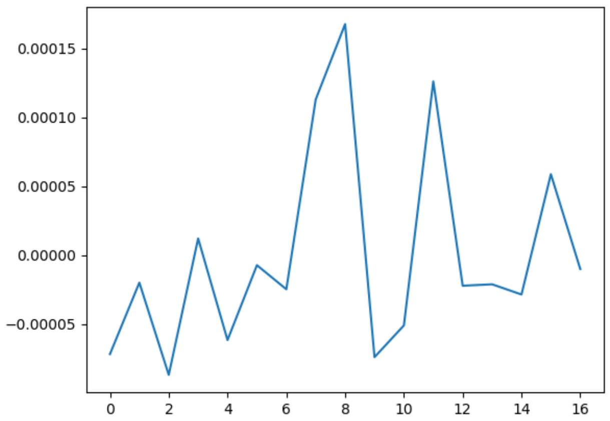
PCA component 2



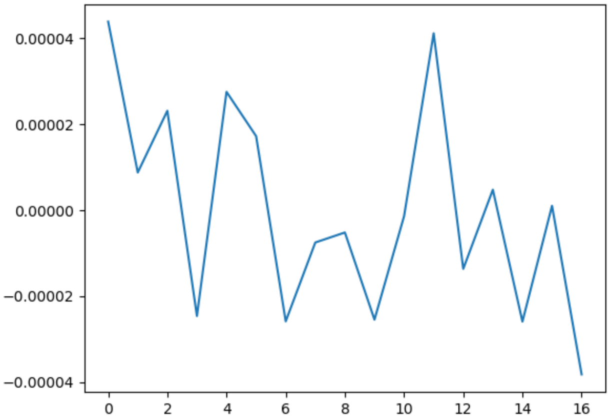
PCA component 3



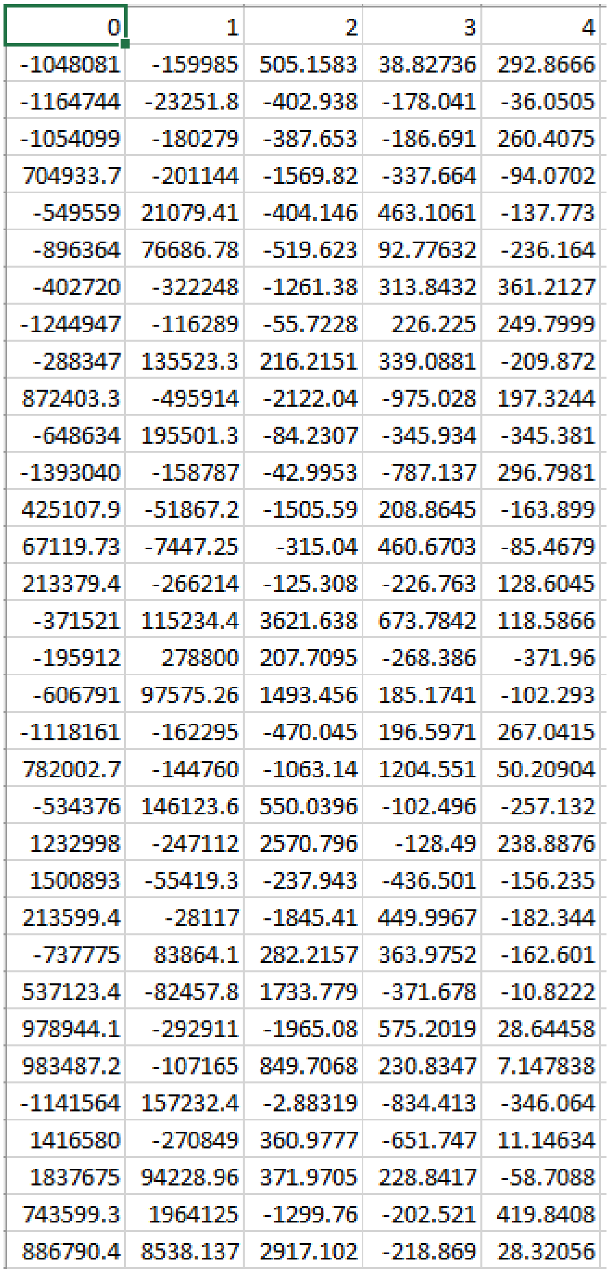
PCA component 4



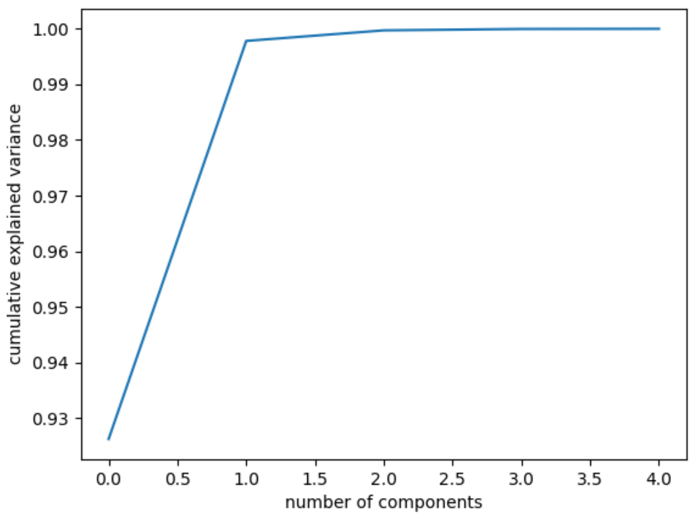
PCA component 5



f)



0	1	2	3	4
-1048081	-159985	505.1583	38.82736	292.8666
-1164744	-23251.8	-402.938	-178.041	-36.0505
-1054099	-180279	-387.653	-186.691	260.4075
704933.7	-201144	-1569.82	-337.664	-94.0702
-549559	21079.41	-404.146	463.1061	-137.773
-896364	76686.78	-519.623	92.77632	-236.164
-402720	-322248	-1261.38	313.8432	361.2127
-1244947	-116289	-55.7228	226.225	249.7999
-288347	135523.3	216.2151	339.0881	-209.872
872403.3	-495914	-2122.04	-975.028	197.3244
-648634	195501.3	-84.2307	-345.934	-345.381
-1393040	-158787	-42.9953	-787.137	296.7981
425107.9	-51867.2	-1505.59	208.8645	-163.899
67119.73	-7447.25	-315.04	460.6703	-85.4679
213379.4	-266214	-125.308	-226.763	128.6045
-371521	115234.4	3621.638	673.7842	118.5866
-195912	278800	207.7095	-268.386	-371.96
-606791	97575.26	1493.456	185.1741	-102.293
-1118161	-162295	-470.045	196.5971	267.0415
782002.7	-144760	-1063.14	1204.551	50.20904
-534376	146123.6	550.0396	-102.496	-257.132
1232998	-247112	2570.796	-128.49	238.8876
1500893	-55419.3	-237.943	-436.501	-156.235
213599.4	-28117	-1845.41	449.9967	-182.344
-737775	83864.1	282.2157	363.9752	-162.601
537123.4	-82457.8	1733.779	-371.678	-10.8222
978944.1	-292911	-1965.08	575.2019	28.64458
983487.2	-107165	849.7068	230.8347	7.147838
-1141564	157232.4	-2.88319	-834.413	-346.064
1416580	-270849	360.9777	-651.747	11.14634
1837675	94228.96	371.9705	228.8417	-58.7088
743599.3	1964125	-1299.76	-202.521	419.8408
886790.4	8538.137	2917.102	-218.869	28.32056



As you can see from the variance explained graph, first three(component 0,1,2) components account for 100% of the variance, which entirely explained the data set.