CS 584-04: Machine Learning

Fall 2019 Assignment 1 [NAME – RAHUL NAIR (A20438470)]

ANSWER 1:

a) According to Izenman (1991) method, the recommended bin-width for the histogram for x should be:

$$h = 2(IQR)\,N^{-1/3}$$

where:

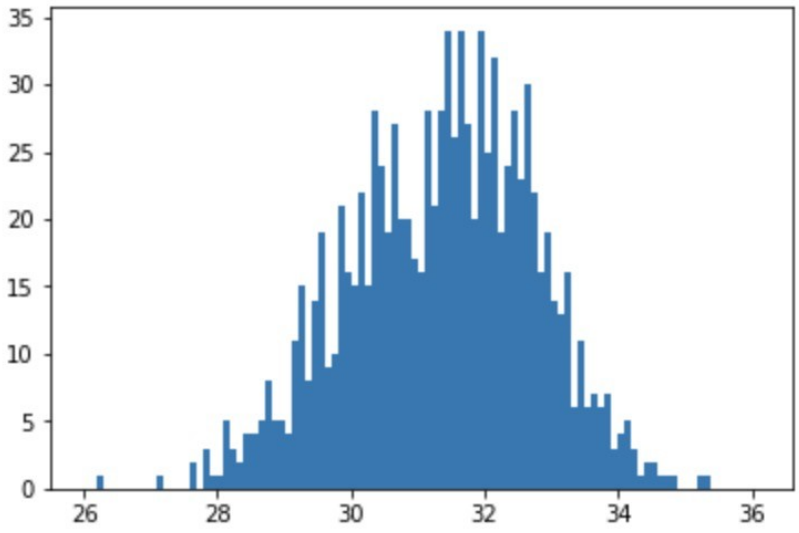
h: bin-width
IQR: Inter-Quartile range
N: Total number of Observations

Now, in this problem the recommended bin-width would be 40.01332889135637.

b) Minimum value of Field x: 26.3
Maximum value of Field x: 35.3

c) Value of a: 26
Value of b: 36

d)

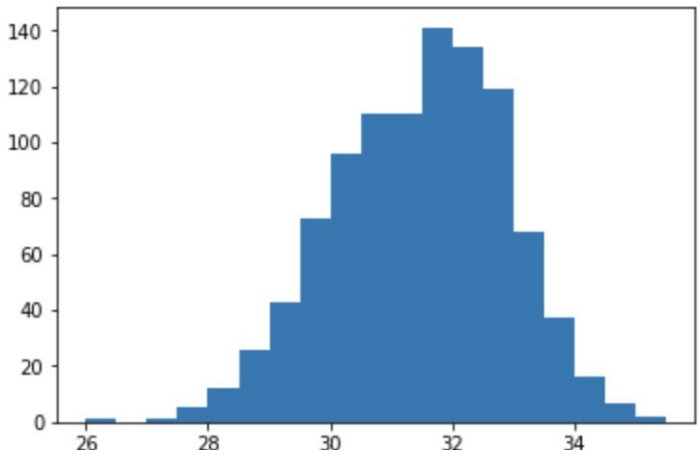


m(i)	p(i)
26.05	0.0
26.150000000000002	0.0
26.250000000000004	0.00999000999000999
26.350000000000005	0.0
26.450000000000006	0.0
26.550000000000008	0.0
26.65000000000001	0.0
26.75000000000001	0.0

26.850000000000012	0.0
26.950000000000014	0.0
27.050000000000015	0.0
27.150000000000016	0.00999000999000999
27.250000000000018	0.0
27.35000000000002	0.0
27.45000000000002	0.0
27.550000000000022	0.0
27.650000000000023	0.01998001998001998
27.750000000000025	0.0
27.850000000000026	0.02997002997002997
27.950000000000028	0.00999000999000999
28.05000000000003	0.00999000999000999
28.15000000000003	0.049950049950049945
28.250000000000032	0.02997002997002997
28.350000000000033	0.01998001998001998
28.450000000000035	0.03996003996003996
28.550000000000036	0.03996003996003996
28.650000000000038	0.049950049950049945
28.75000000000004	0.07992007992007992
28.85000000000004	0.049950049950049945
28.950000000000042	0.049950049950049945
29.050000000000043	0.03996003996003996
29.150000000000045	0.10989010989010987
29.250000000000046	0.14985014985014983
29.350000000000048	0.07992007992007992
29.45000000000005	0.13986013986013984
29.55000000000005	0.1898101898101898
29.650000000000052	0.0899100899100899
29.750000000000053	0.09990009990009989
29.850000000000055	0.20979020979020976
29.950000000000056	0.15984015984015984
30.050000000000058	0.14985014985014983
30.15000000000006	0.21978021978021975
30.25000000000006	0.14985014985014983
30.350000000000062	0.2797202797202797
30.450000000000063	0.23976023976023975
30.550000000000065	0.1898101898101898
30.650000000000066	0.2697302697302697
30.750000000000068	0.19980019980019978
30.85000000000007	0.19980019980019978
30.95000000000007	0.16983016983016982
31.05000000000007	0.15984015984015984
31.150000000000073	0.2797202797202797
31.250000000000075	0.20979020979020976
31.350000000000076	0.2797202797202797
31.450000000000077	0.33966033966033965
31.55000000000008	0.2597402597402597
31.65000000000008	0.33966033966033965
31.75000000000008	0.2697302697302697
31.850000000000083	0.19980019980019978
31.950000000000085	0.33966033966033965
32.05000000000008	0.24975024975024973
32.15000000000009	0.3196803196803197
32.250000000000085	0.1898101898101898
32.350000000000094	0.23976023976023975
32.45000000000009	0.2797202797202797
32.5500000000001	0.22977022977022976

32.65000000000009	0.29970029970029965
32.7500000000001	0.21978021978021975
32.850000000000094	0.15984015984015984
32.9500000000001	0.1898101898101898
33.0500000000001	0.13986013986013984
33.150000000000105	0.12987012987012986
33.2500000000001	0.15984015984015984
33.35000000000011	0.05994005994005994
33.4500000000001	0.10989010989010987
33.55000000000011	0.05994005994005994
33.650000000000105	0.06993006993006992
33.750000000000114	0.05994005994005994
33.85000000000011	0.06993006993006992
33.95000000000012	0.02997002997002997
34.05000000000011	0.03996003996003996
34.15000000000012	0.049950049950049945
34.250000000000114	0.02997002997002997
34.35000000000012	0.00999000999000999
34.45000000000012	0.01998001998001998
34.550000000000125	0.01998001998001998
34.65000000000012	0.00999000999000999
34.75000000000013	0.00999000999000999
34.85000000000012	0.00999000999000999
34.95000000000013	0.0
35.050000000000125	0.0
35.150000000000134	0.0
35.25000000000013	0.00999000999000999
35.350000000000136	0.00999000999000999
35.45000000000013	0.0
35.55000000000014	0.0
35.650000000000134	0.0
35.75000000000014	0.0
35.850000000000136	0.0

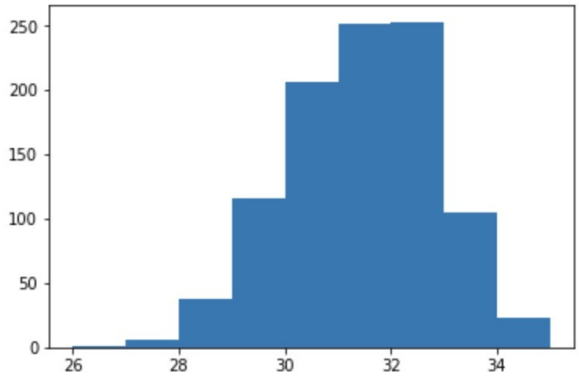
e)



m(i)	p(i)
26.25	0.001998001998001998
26.75	0.0

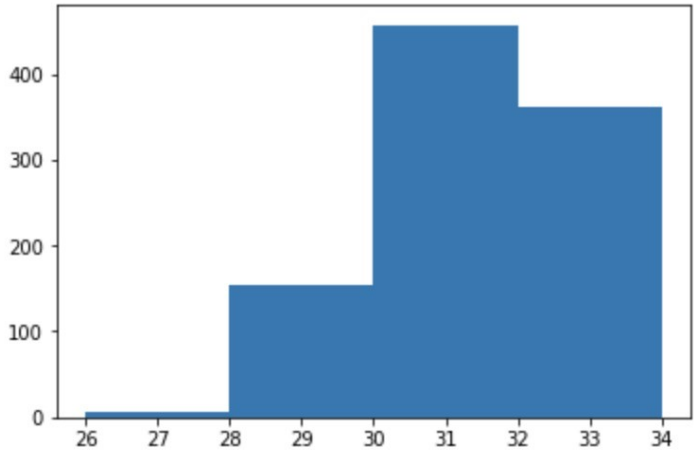
27.25	0.001998001998001998
27.75	0.011988011988011988
28.25	0.03196803196803197
28.75	0.061938061938061936
29.25	0.11388611388611389
29.75	0.17782217782217782
30.25	0.23976023976023977
30.75	0.25374625374625376
31.25	0.28771228771228774
31.75	0.34965034965034963
32.25	0.32367632367632365
32.75	0.2757242757242757
33.25	0.15784215784215785
33.75	0.07992007992007992
34.25	0.03596403596403597
34.75	0.013986013986013986
35.25	0.003996003996003996

f) _



m(i)	p(i)
26.5	0.000999000999000999
27.5	0.006993006993006993
28.5	0.04295704295704296
29.5	0.13186813186813187
30.5	0.22277722277722278
31.5	0.28471528471528473
32.5	0.27172827172827174
33.5	0.10789210789210789
34.5	0.022977022977022976

g)



m(i)	p(i)
26.5	0.000999000999000999
27.5	0.006993006993006993
28.5	0.04295704295704296
29.5	0.13186813186813187
30.5	0.22277722277722278
31.5	0.28471528471528473
32.5	0.27172827172827174
33.5	0.10789210789210789
34.5	0.022977022977022976

h) In my opinion, the histogram with bin-width 0.5 will provide us with more insights into the shape and distribution of the field x. As we can see in the histogram and density estimations, the histogram with bin-width 0.1 is erratically distributed and when it comes to histograms with bin-widths 1 and 2 don't give us much insights. So with proper balance between the marginalized distribution of data histogram with bin-width 0.1 is the best in these list.

ANSWER 2:

a) Five- point summary of x:
   Minimum value = 26.3
   First Quartile = 30.4
Median = 31.5
Third Quartile = 32.4
Maximum value = 35.4

Values of 1.5 IQR whiskers:
   Q1- 1.5*IQR = minimum value = 27.4
   Q3+1.5*IQR = maximum value = 35.4

b) Five- point summary of group 0:

Minimum value = 26.3
First Quartile = 29.4
Median = 30
Third Quartile = 30.6
Maximum value = 32.2

Values of 1.5 IQR whiskers:

Q1- 1.5*IQR = minimum value = 27.6
Q3+1.5*IQR = maximum value = 32.4

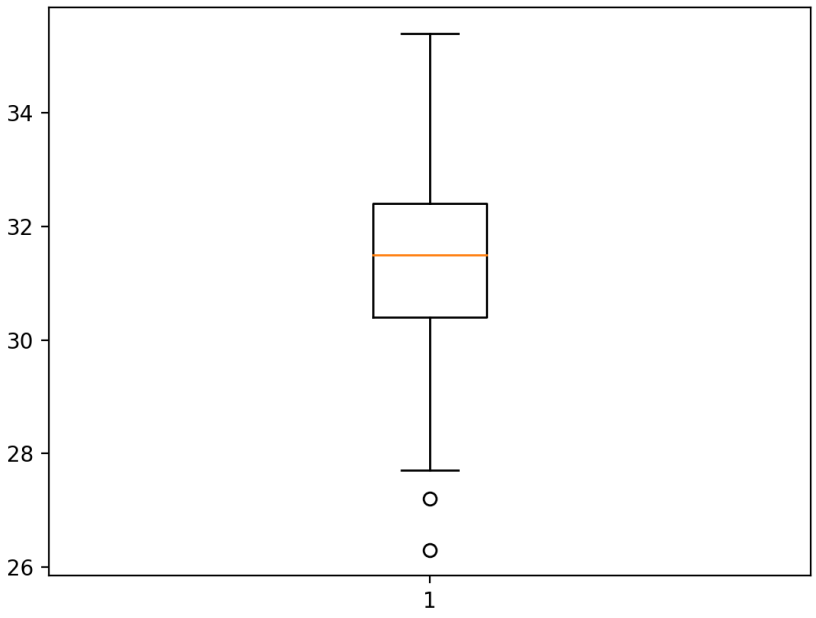
Five- point summary of group 1:

Minimum value = 29.1
First Quartile = 31.4
Median = 32.1
Third Quartile = 32.7
Maximum value = 35.4

Values of 1.5 IQR whiskers:

Q1- 1.5*IQR = minimum value = 29.45
Q3+1.5*IQR = maximum value = 34.65

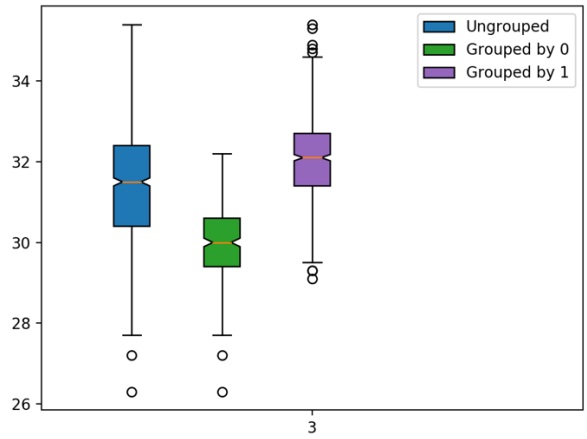
c) Boxplot:



The 25th percentile value of the data is 30.4. When we apply 1.5 IQR whiskers, the answer is 27.4 which is precisely shown by python boxplot.

The 75th percentile of the data is 32.4. When we apply 1.5 IQR whiskers, the answer is 35.4 which is the same as that calculated by python. So, it is exactly correct.

d)



For the general data (Ungrouped), the ouliers found using 1.5 whiskers are:-
27.21
26.28

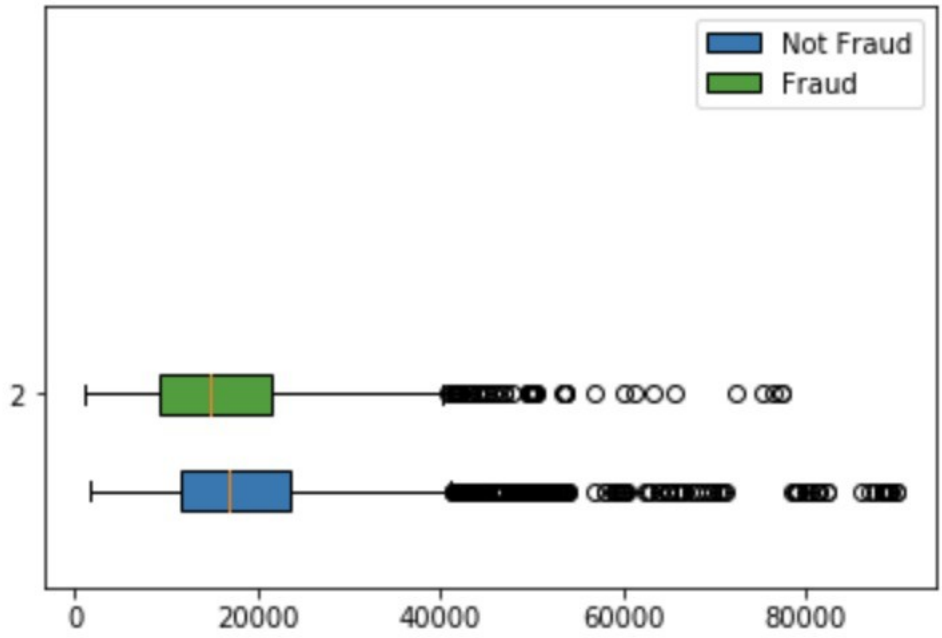
For the data which are grouped by 0, the outliers found using 1.5 whiskers are:-
27.21
26.28

For the data which are grouped by 1, the outliers found using 1.5 whiskers are:-
29.33
29.1
35.39
35.28
34.88
34.77
34.70
34.66

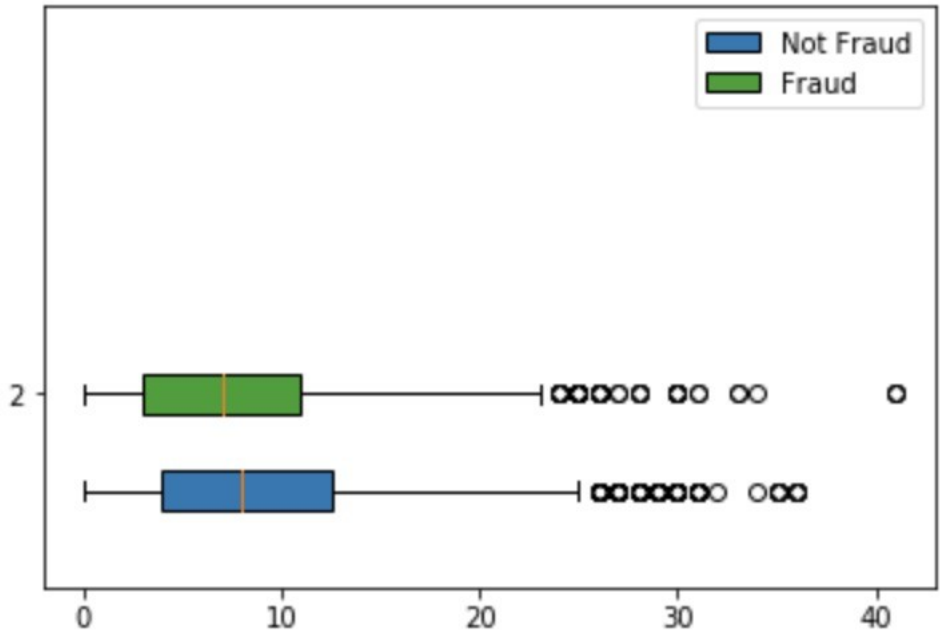
ANSWER 3:

a) Percentage of investigations found to be fraudulent are: 19.9497%

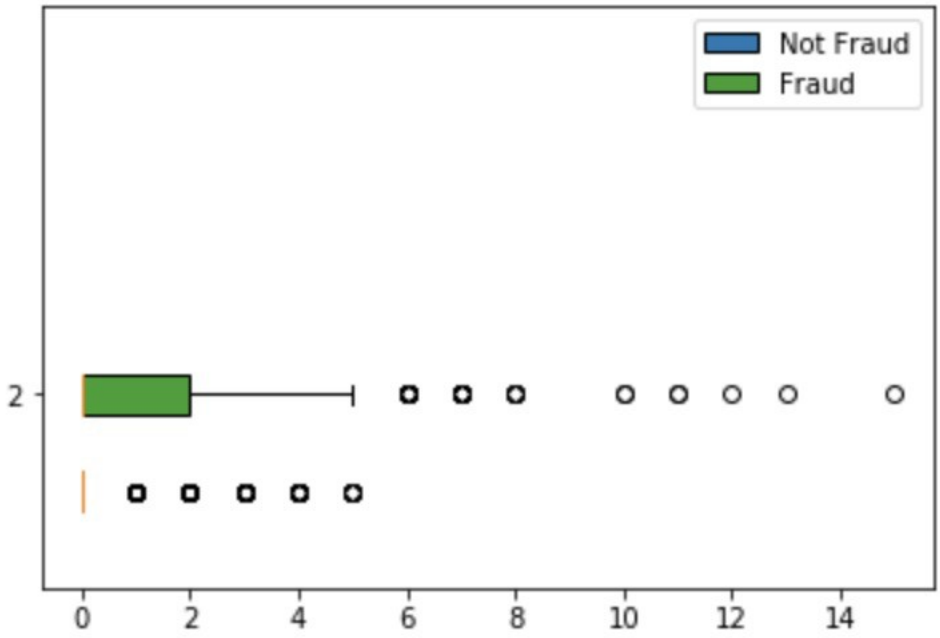
b) For ‘TOTAL_SPEND’:



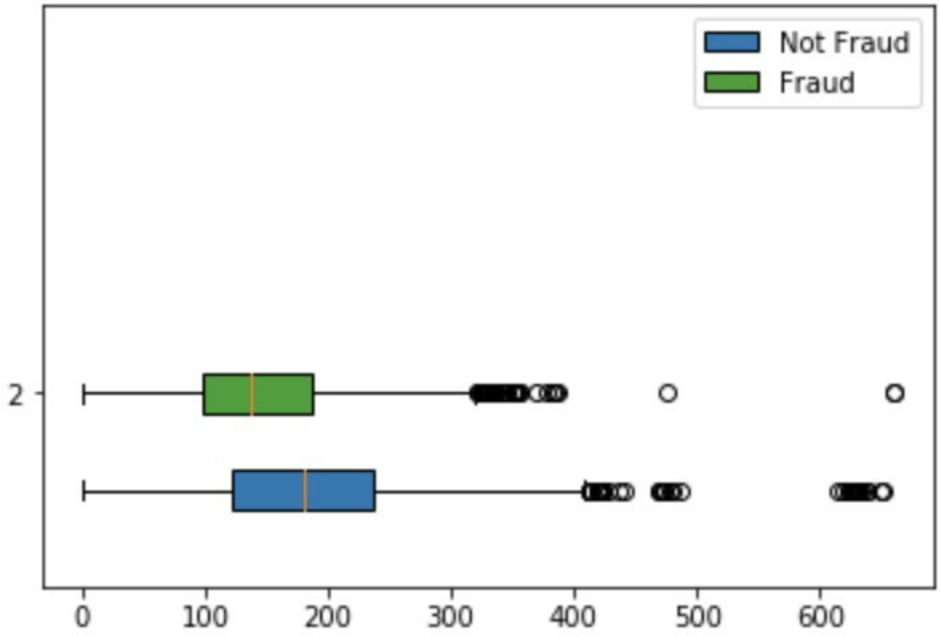
FOR ‘DOCTOR_VISITS’:



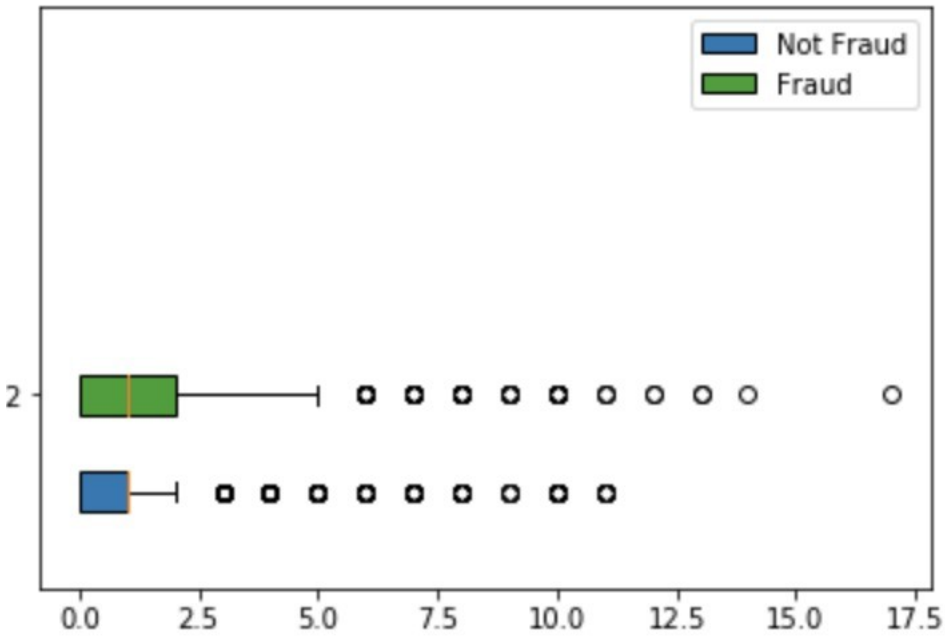
FOR ‘NUM_CLAIMS’:



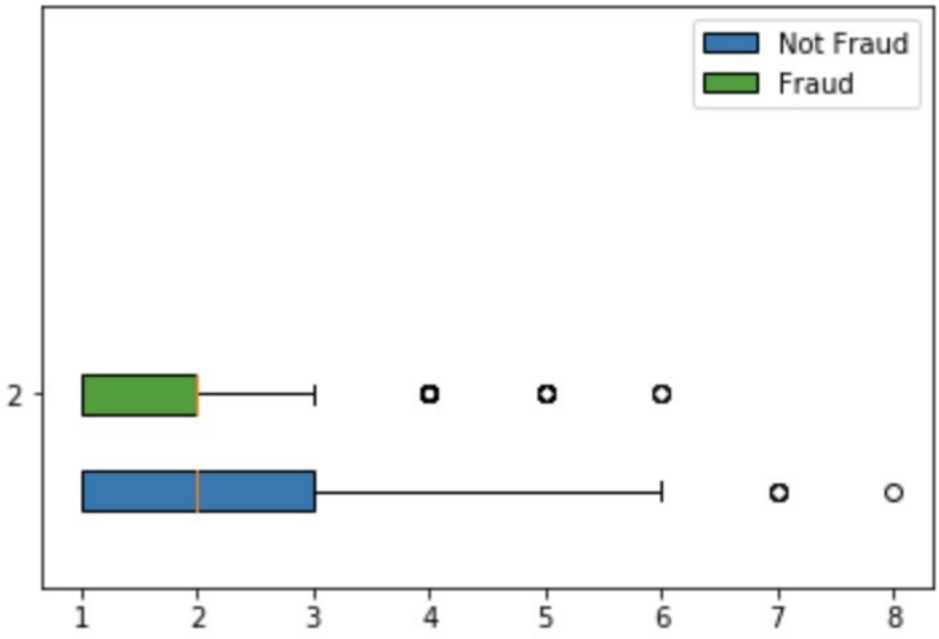
FOR ‘MEMBER_DURATION’:



FOR ‘OPTOM_PRESC’:



FOR ‘NUM_MEMBERS’:



c) (i) Number of Dimensions used: 6
(ii) Transformation matrix:
[[-6.49862374e-08 -2.41194689e-07 2.69941036e-07
-2.42525871e-07
-7.90492750e-07 5.96286732e-07]
[7.31656633e-05 -2.94741983e-04 9.48855536e-05
1.77761538e-03

3.51604254e-06 2.20559915e-10]
[-1.18697179e-02 1.70828329e-03 -7.68683456e-04
2.03673350e-05
1.76401304e-07 9.09938972e-12]
[1.92524315e-06 -5.37085514e-05 2.32038406e-05
-5.78327741e-05
1.08753133e-04 4.32672436e-09]
[8.34989734e-04 -2.29964514e-03 -7.25509934e-03
1.11508242e-05
2.39238772e-07 2.85768709e-11]
[2.10964750e-03 1.05319439e-02 -1.45669326e-03
4.85837631e-05
6.76601477e-07 4.66565230e-11]]

We can prove that the resulting variables are orthonormal if the multiplication of transpose of the transformed matrix and the transformed matrix gives us an identity
matrix which is:
[[1.00000000e+00 -2.87703888e-15 1.90299165e-15
7.06552872e-15
1.16226473e-15 -1.35308431e-16]
[-2.87703888e-15 1.00000000e+00 -1.37216627e-15
-1.98244199e-14
-6.59194921e-16 7.21644966e-16]
[1.90299165e-15 -1.37216627e-15 1.00000000e+00
4.96366728e-15
-6.24500451e-17 -1.17961196e-16]
[7.06552872e-15 -1.98244199e-14 4.96366728e-15
1.00000000e+00
1.10432496e-14 -4.20496971e-15]
[1.16226473e-15 -6.59194921e-16 -6.24500451e-17
1.10432496e-14
1.00000000e+00 -6.66133815e-16]
[-1.35308431e-16 7.21644966e-16 -1.17961196e-16
-4.20496971e-15
-6.66133815e-16 1.00000000e+00]]

Therefore, the resulting variables are orthonormal.

d) i) the score value is 0.8778523489932886.
ii) The score function is showing the accuracy of the model. It depicts that the model is correct about 87% of the time. This is a pretty good accuracy rate for a model.

e) The five neighbors are:-

	CASE_ID	FRAUD	TOTAL_SPEND	DOCTOR_VISITS	NUM_CLAIMS	MEMBER_DURATION	\
588	589	1	7500	15	3	127	
2897	2898	1	16000	18	3	146	
1199	1200	1	10000	16	3	124	
1246	1247	1	10200	13	3	119	
886	887	1	8900	22	3	166	

	OPTOM_PRESC	NUM_MEMBERS
588	2	2
2897	3	2
1199	2	1
1246	2	3
886	1	2

f) The predicted probability of the fraudulent is 100%.
The predicted probability of the fraudulent is greater than the answer we found in a) part. Based on the criterion this will be classified as fraudulent, but this observation will be misclassified. Since the accuracy of the model is 87.79%(approx), we still cannot be sure whether the model is correct in classifying it as fraudulent.