Assignment 1

Name : Sourav Yadav

AID: A20450418

CS584 Spring 2020

Question 1

Answer (a):

Recommended Bandwidth= 2*iqr*pow(N,-1/3) =0.3998667554864774

Answer (b):

Max=35.4

Min=26.3

Answer(c):

A=26

B=36

Answer (d):

Histogram is given below

Co-ordinates of histogram given in format: (mid point of histogram, p(mid point of histogram))

(26.125,0.0)

(26.375,0.003996003996003996)

(26.625,0.0)

(26.875,0.0)

(27.125,0.003996003996003996)

(27.375,0.0)

(27.625,0.007992007992007992)

(27.875,0.011988011988011988)

(28.125,0.027972027972027972)

(28.375,0.01998001998001998)

(28.625,0.05194805194805195)

(28.875,0.05194805194805195)

(29.125,0.07992007992007992)

(29.375,0.0919080919080919)

(29.625,0.16783216783216784)

(29.875,0.12387612387612387)

(30.125,0.21178821178821178)

(30.375,0.17182817182817184)

(30.625,0.27972027972027974)

(30.875,0.15984015984015984)

(31.125,0.24375624375624375)

(31.375,0.1958041958041958)

(31.625,0.3756243756243756)

(31.875,0.1878121878121878)

(32.125,0.36363636363636365)

(32.375,0.17182817182817184)

(32.625,0.32367632367632365)

(32.875,0.15184815184815184)

(33.125,0.1838161838161838)

(33.375,0.08791208791208792)

(33.625,0.0959040959040959)

(33.875,0.05194805194805195)

(34.125,0.04795204795204795)

(34.375,0.015984015984015984)

(34.625,0.01998001998001998)

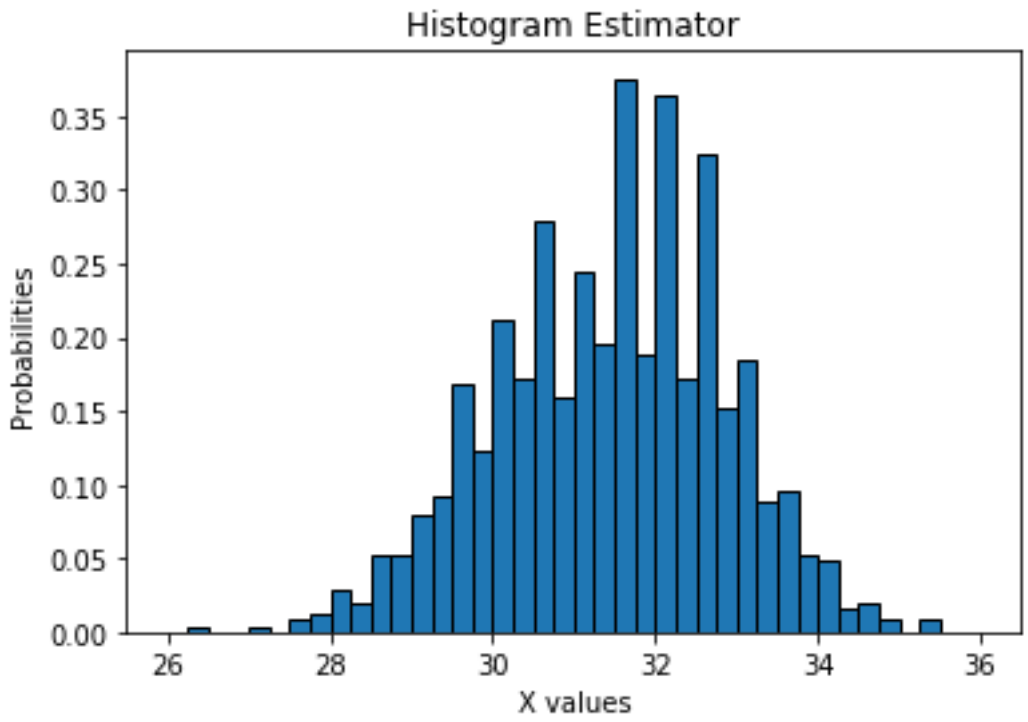
(34.875,0.007992007992007992)

(35.125,0.0)

(35.375,0.007992007992007992)

(35.625,0.0)

(35.875,0.0)



Answer (e):

Co-ordinates of histogram given in format: (mid point of histogram,p(mid point of histogram))

(26.25,0.001998001998001998)

(26.75,0.0)

(27.25,0.001998001998001998)

(27.75,0.00999000999000999)

(28.25,0.023976023976023976)

(28.75,0.05194805194805195)

(29.25,0.08591408591408592)

(29.75,0.14585414585414586)

(30.25,0.1918081918081918)

(30.75,0.21978021978021978)

(31.25,0.21978021978021978)

(31.75,0.2817182817182817)

(32.25,0.2677322677322677)

(32.75,0.23776223776223776)

(33.25,0.13586413586413587)

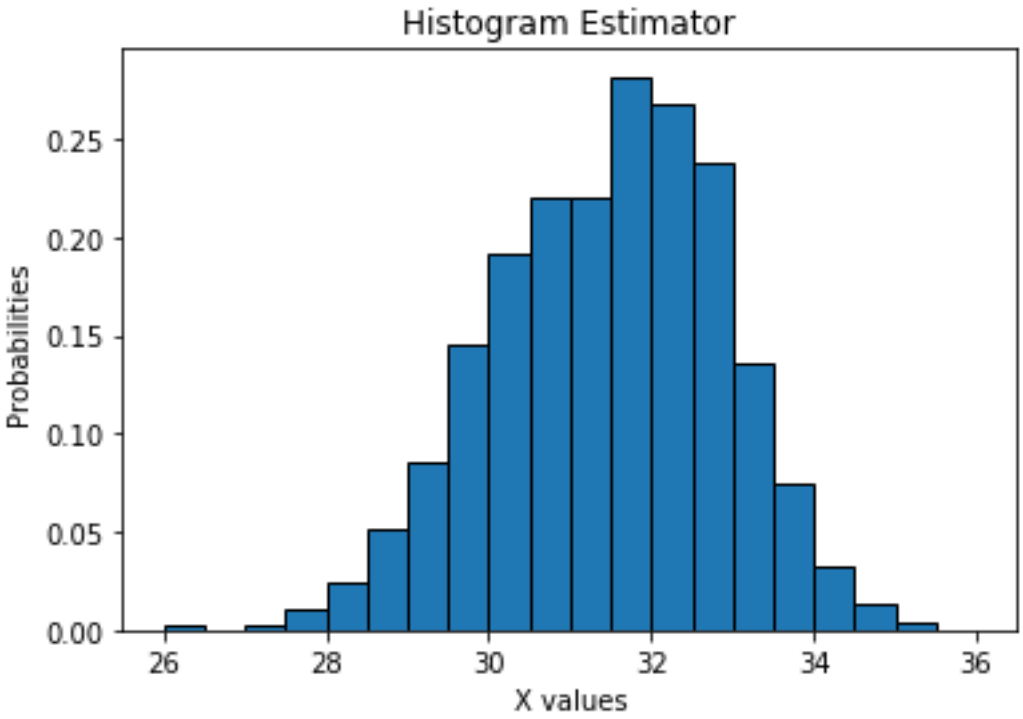
(33.75,0.07392607392607392)

(34.25,0.03196803196803197)

(34.75,0.013986013986013986)

(35.25,0.003996003996003996)

(35.75,0.0)



Answer (f):

Co-ordinates of histogram given in format: (mid point of histogram,p(mid point of histogram))

(26.5,0.000999000999000999)

(27.5,0.005994005994005994)

(28.5,0.03796203796203796)

(29.5,0.11588411588411589)

(30.5,0.2057942057942058)

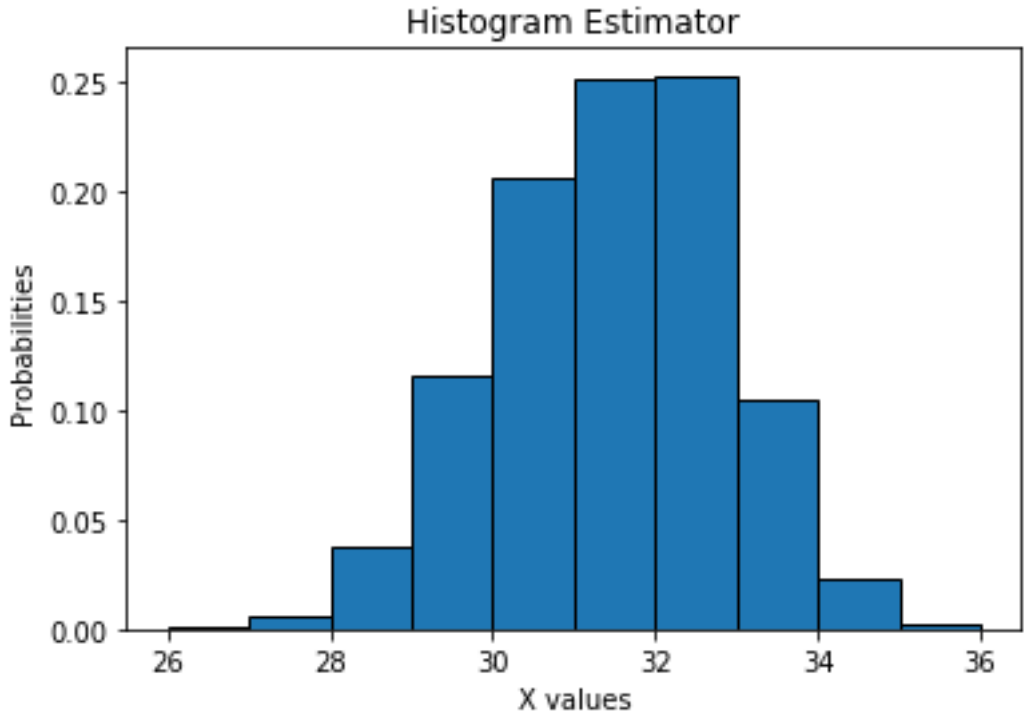
(31.5,0.25074925074925075)

(32.5,0.25274725274725274)

(33.5,0.1048951048951049)

(34.5,0.022977022977022976)

(35.5,0.001998001998001998)



Answer (g): Co-ordinates of histogram given in format: (mid point of histogram, p(mid point of histogram))

(27.0,0.0034965034965034965)

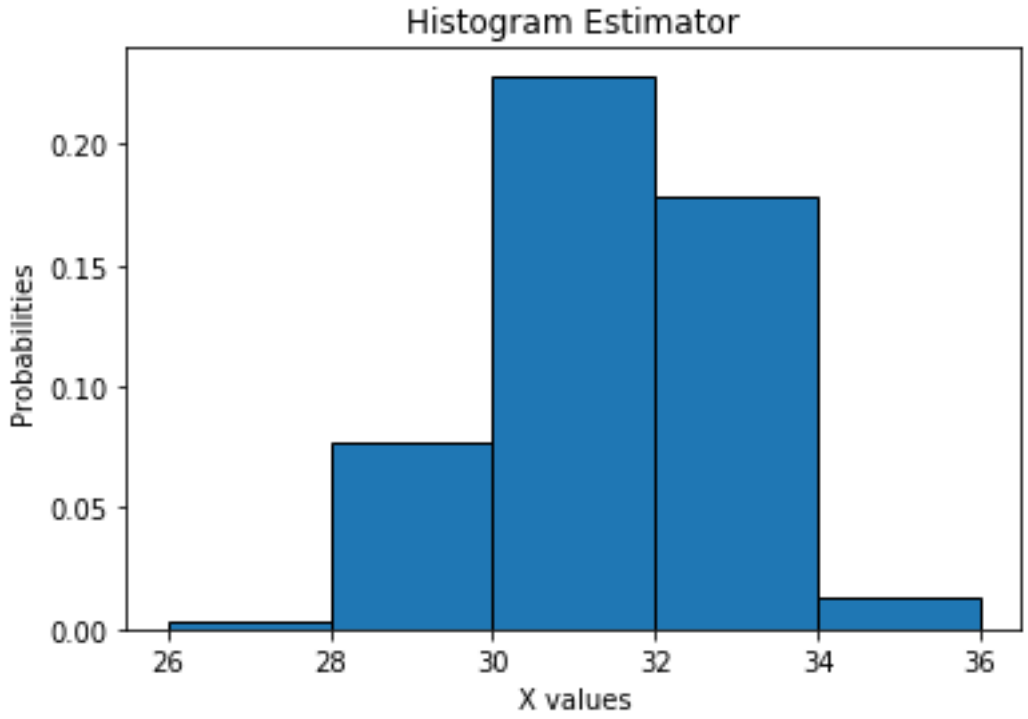
(29.0,0.07692307692307693)

(31.0,0.22827172827172826)

(33.0,0.17882117882117882)

(35.0,0.012487512487512488)

Histogram is given below



Answer (h): In my opinion, among four histogram the histogram with bin-width 0.5 provide best insight into the distribution of the field of x. h=0.25 produced the histogram with lot of uneven probability estimation whereas h=1 and h=2 histogram have wider bin-width which fails to represent true distribution.

Question 2:

Answer (a):

Five number summary of x:

Q1 :30.4

Median :31.5

Q3 :32.4

Maximum :35.4

Minimum :26.3

1.5 IQR whiskers :35.4,27.4

Answer (b):

Five number summary of x of group1:

Q1 :31.4

Median :32.1

Q3 :32.7

Maximum :35.4

Minimum :29.1

1.5 IQR whiskers :34.650000000000006,29.449999999999992

Five number summary of x of group0:

Q1 :29.4

Median :30.0

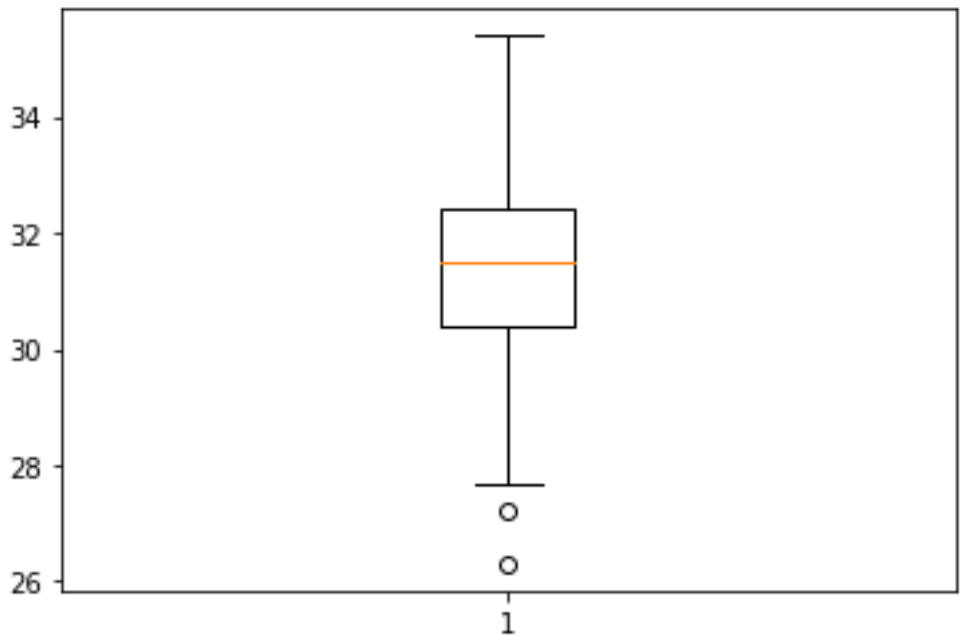
Q3 :30.6

Maximum :32.2

Minimum :26.3

1.5 IQR whiskers:32.400000000000006,27.599999999999994

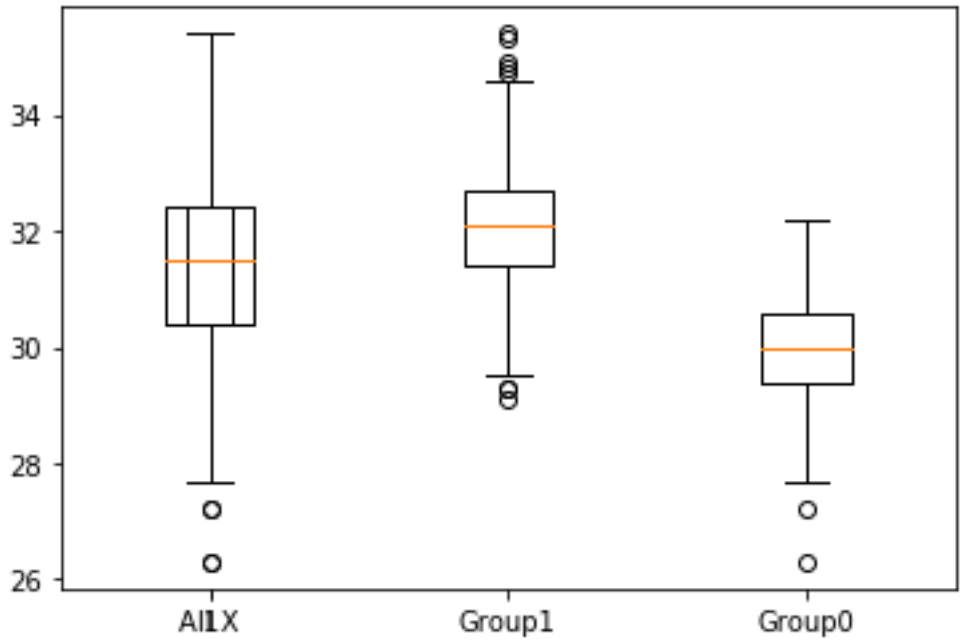
Answer (c):



1.5 IQR Whiskers are calculated above in Answer(a): 35.4,27.4

From the above boxplot I can verify that shown Boxplot whiskers are correct using the calculated 1.5 IQR values.

Answer (d):



Outliers of all X values :[26.3, 27.2]

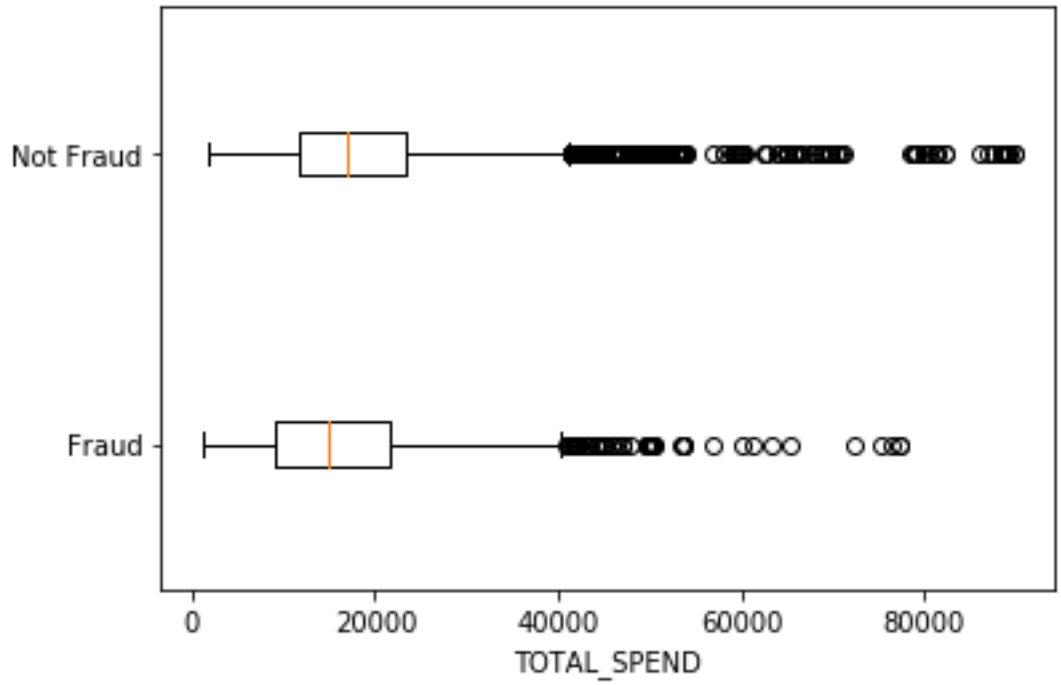
Outliers of Group 1 :[35.3, 29.3, 35.4, 34.9, 34.7, 34.8, 29.3, 29.1]

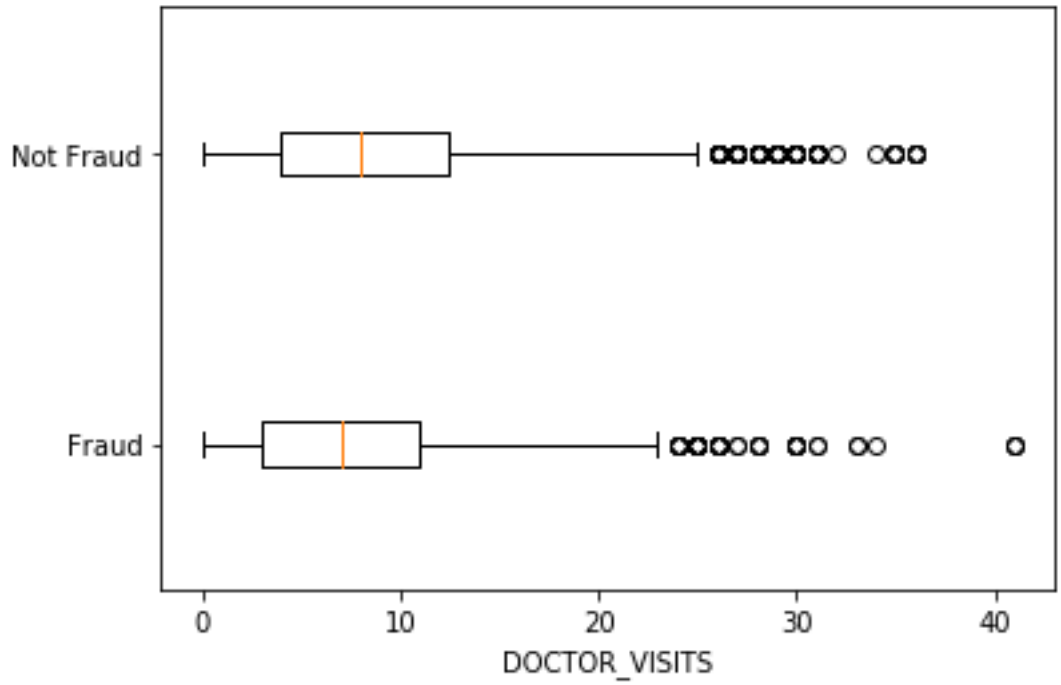
Outliers of Group 0 :[27.2, 26.3]

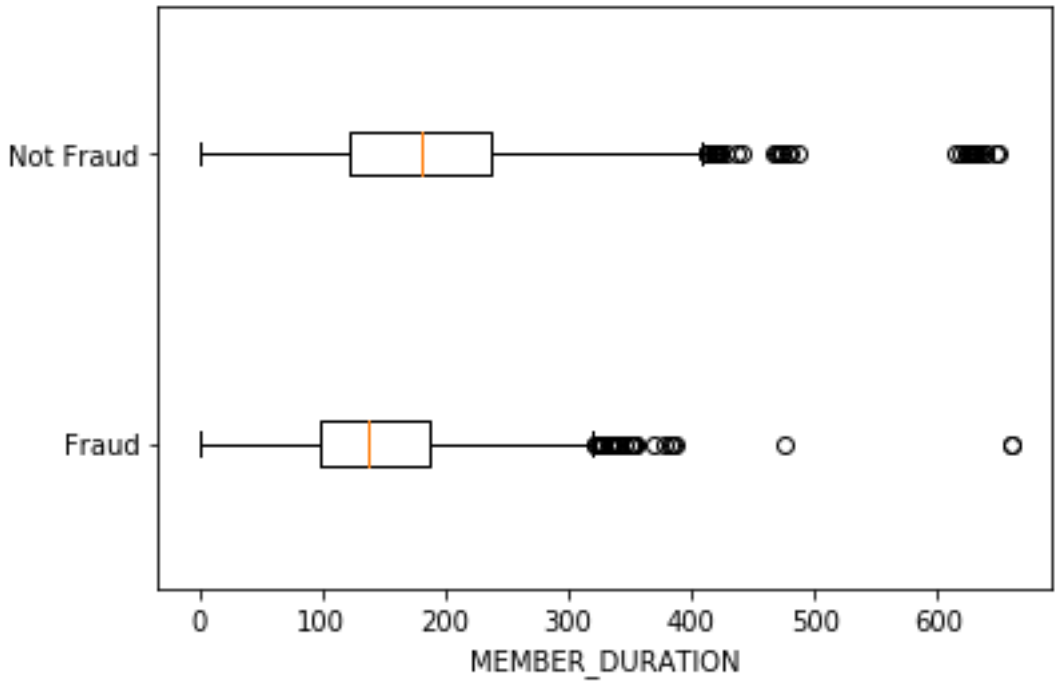
Question 3:

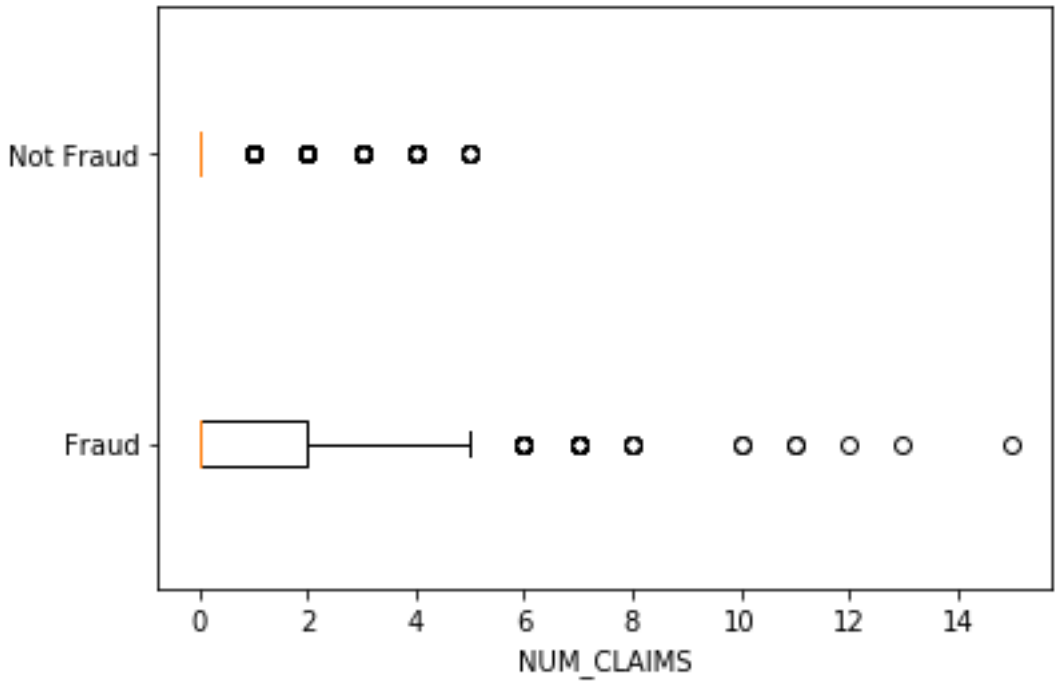
Answer (a): Percentage of investigations found fraudulent: 19.9497

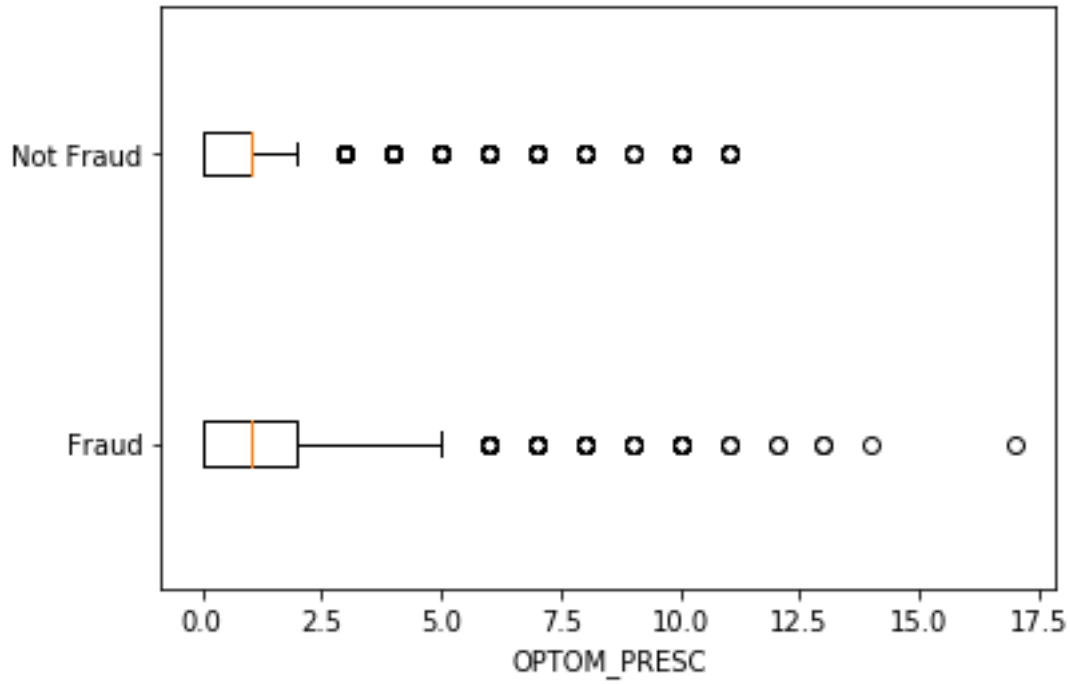
Answer (b):

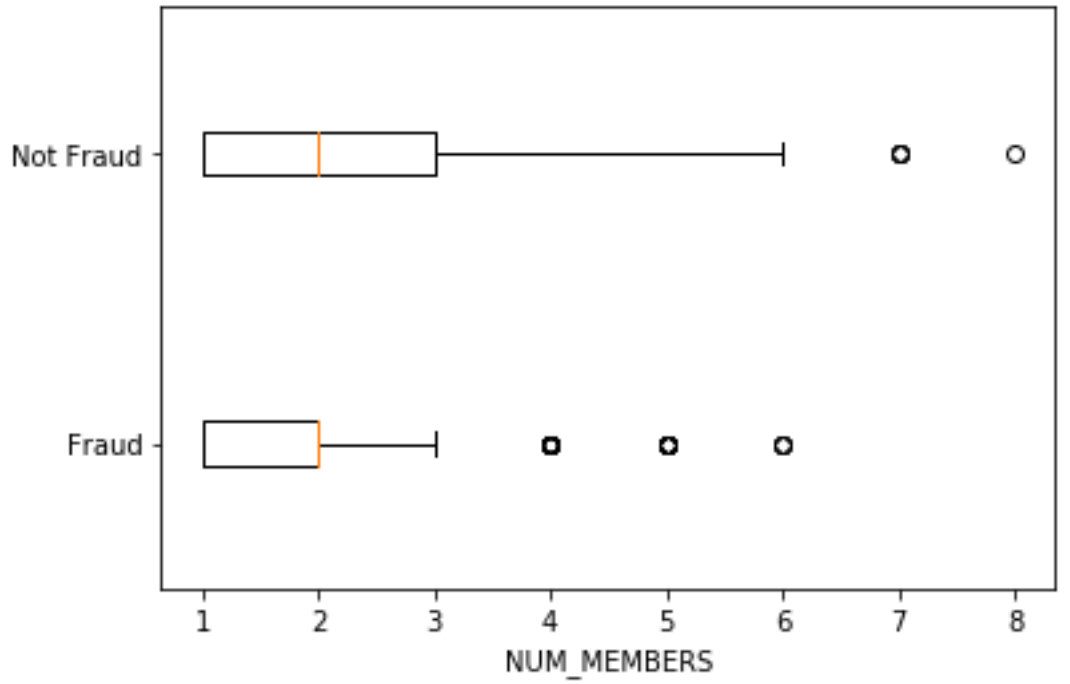

Not Fraud
Fraud
0
20000
40000
60000
80000
TOTAL_SPEND


Not Fraud
Fraud
0
10
20
30
40
DOCTOR_VISITS


Not Fraud
Fraud
0
100
200
300
400
500
600
MEMBER_DURATION


Not Fraud
Fraud
0
2
4
6
8
10
12
14
NUM_CLAIMS





Answer (c):

1.The number of dimensions used 6

2. 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]]

The resulting variable are orthogonal if we multiply resulting variable matrix with the transpose of itself ,we get identity matrix i.e.,

Expecting an Identity Matrix by multiplying tranpose of transformed X matrix with itself =

[[1.00000000e+00 -2.99781901e-16 -4.56882795e-16 5.45884952e-15

1.20129601e-15 -1.27176915e-16]

[-2.99781901e-16 1.00000000e+00 -6.56592836e-16 -2.76891140e-14

-1.22818422e-15 7.71951947e-16]

[-4.56882795e-16 -6.56592836e-16 1.00000000e+00 3.50110566e-15

1.14491749e-16 -2.32452946e-16]

[5.45884952e-15 -2.76891140e-14 3.50110566e-15 1.00000000e+00

1.14825684e-14 -3.47768689e-15]

[1.20129601e-15 -1.22818422e-15 1.14491749e-16 1.14825684e-14

1.00000000e+00 -6.27969898e-16]

[-1.27176915e-16 7.71951947e-16 -2.32452946e-16 -3.47768689e-15

-6.27969898e-16 1.00000000e+00]]

The above matrix is identity matrix and hence the resulting variable are orthogonal.

Now taking the norm of Transformed matrix X.

Norm of the Transformed matrix: 1.000000000000081

Hence the Transformed matrix X is both Orthogonal and its norm is 1. Therefore, the it is Orthonormal Matrix.

Answer (d)

(i) Score returned by Score Function :0.8778523489932886

(ii) Score value return by score function indicates the accuracy of our KNN algorithm. It is the average accuracy of our algorithm on our test data and labels. The above score value returned by KNN algorithm indicates that our algorithm return approximately 88% correct results.

Answer (e):

Row index of 5 neighbours of given point : [[588 2897 1199 1246 886]]

Input value and Target Values:

	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

Answer (f):

Predicated Probability of Fraudulent: 1.0

Since 1 is greater than 0.19, the observation will be fraudulent and hence will not be misclassified.