Machine Learning

Assignment 1

```
1.          i  group     x
0       0      1  29.8
1       1      1  33.3
2       2      0  30.9
3       3      1  32.2
4       4      0  31.1
...    ...    ...   ...
996    996      0  29.6
997    997      1  31.5
998    998      1  30.1
999    999      0  28.8
1000  1000      0  30.6

[1001 rows x 3 columns]
```

a) Recommended bin width according to Izeman:

```
Recommended bandwidth 0.3998667554864774
```

b) Minimum and Maximum values:

```
Minimum value 26.3
Maximum value 35.4
```

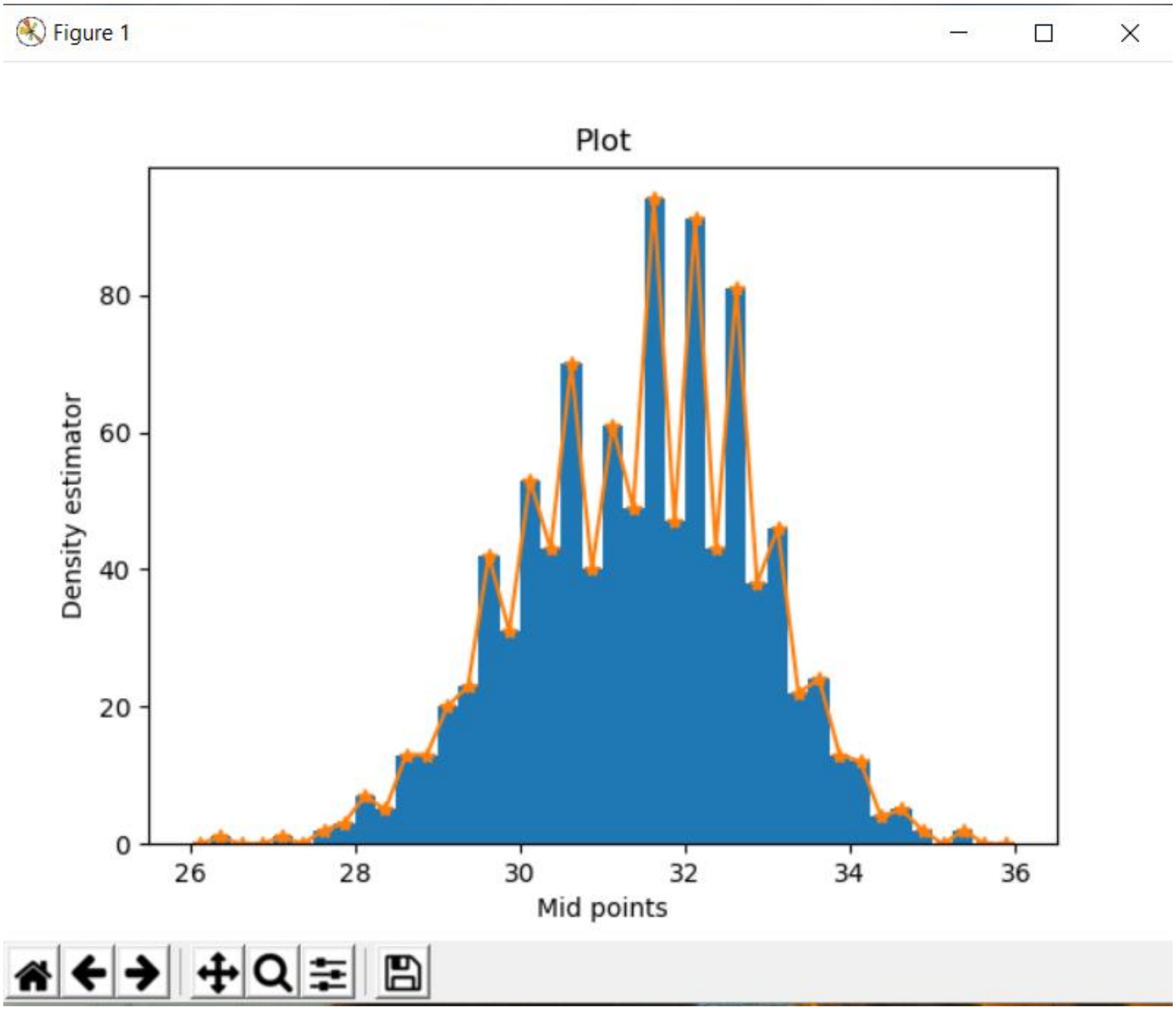
c) Largest number less than minimum and smallest number greater than maximum:

```
Largest value less than minimum 26
Smallest value greater than maximum 36
```

d) Histogram for h=0.25

```
Coordinates of Density Estimator
   mi       p(mi)
-----  ----------
26.12  0
26.37  0.003996
26.62  0
26.87  0
27.12  0.003996
27.37  0
27.62  0.00799201
```

27.87	0.011988
28.12	0.027972
28.37	0.01998
28.62	0.0519481
28.87	0.0519481
29.12	0.0799201
29.37	0.0919081
29.62	0.167832
29.87	0.123876
30.12	0.211788
30.37	0.171828
30.62	0.27972
30.87	0.15984
31.12	0.243756
31.37	0.195804
31.62	0.375624
31.87	0.187812
32.12	0.363636
32.37	0.171828
32.62	0.323676
32.87	0.151848
33.12	0.183816
33.37	0.0879121
33.62	0.0959041
33.87	0.0519481
34.12	0.047952
34.37	0.015984
34.62	0.01998
34.87	0.00799201
35.12	0
35.37	0.00799201
35.62	0
35.87	0

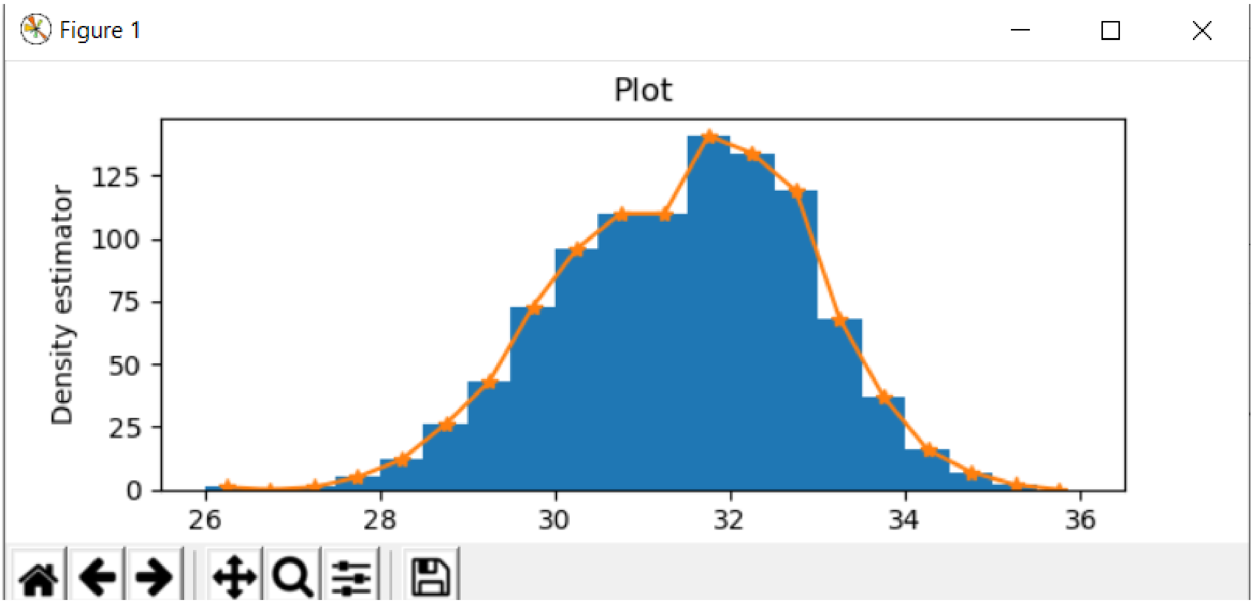


e) Histogram for h=0.5

Coordinates of Density Estimator

```
  mi      p(mi)

-----  ----------

26.25  0.001998

26.75  0

27.25  0.001998

27.75  0.011988

28.25  0.02997

28.75  0.0539461
```

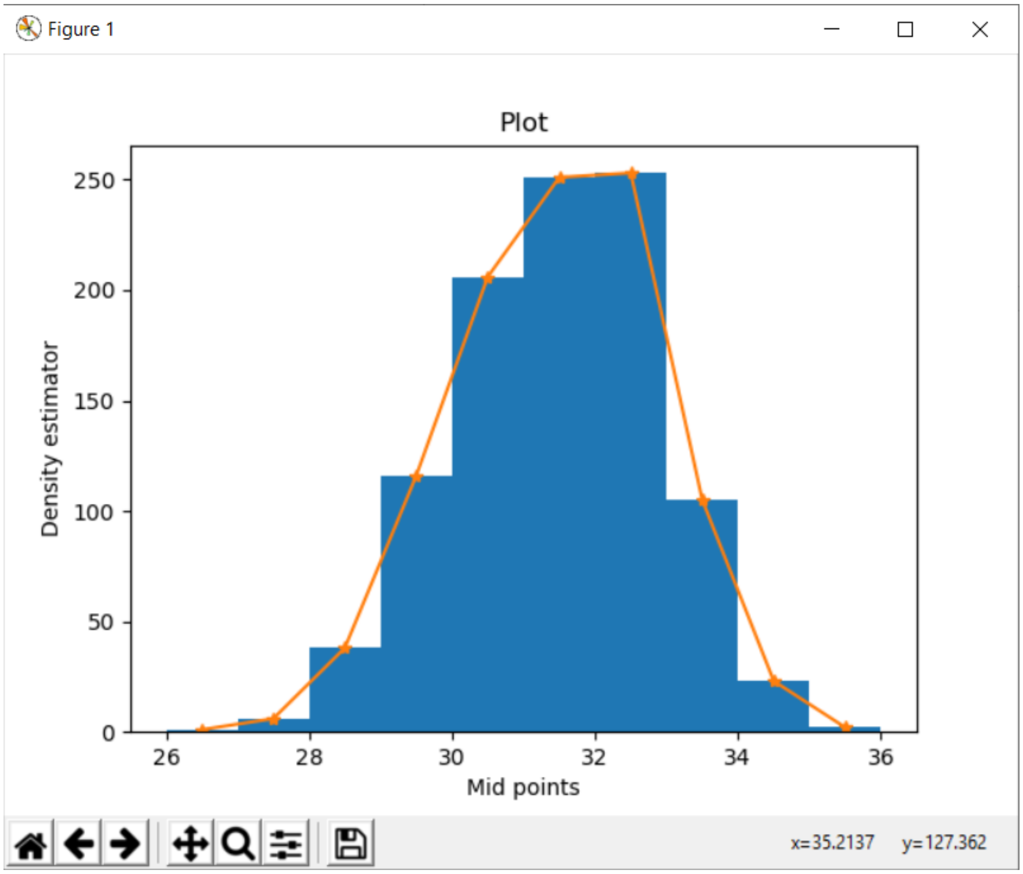
```
29.25  0.103896

29.75  0.14985

30.25  0.207792

30.75  0.205794

31.25  0.253746

31.75  0.281718

32.25  0.255744

32.75  0.21978

33.25  0.11988

33.75  0.0579421

34.25  0.02997

34.75  0.00999001

35.25  0.003996

35.75  0
```



f) Histogram for h=1

Coordinates of Density Estimator

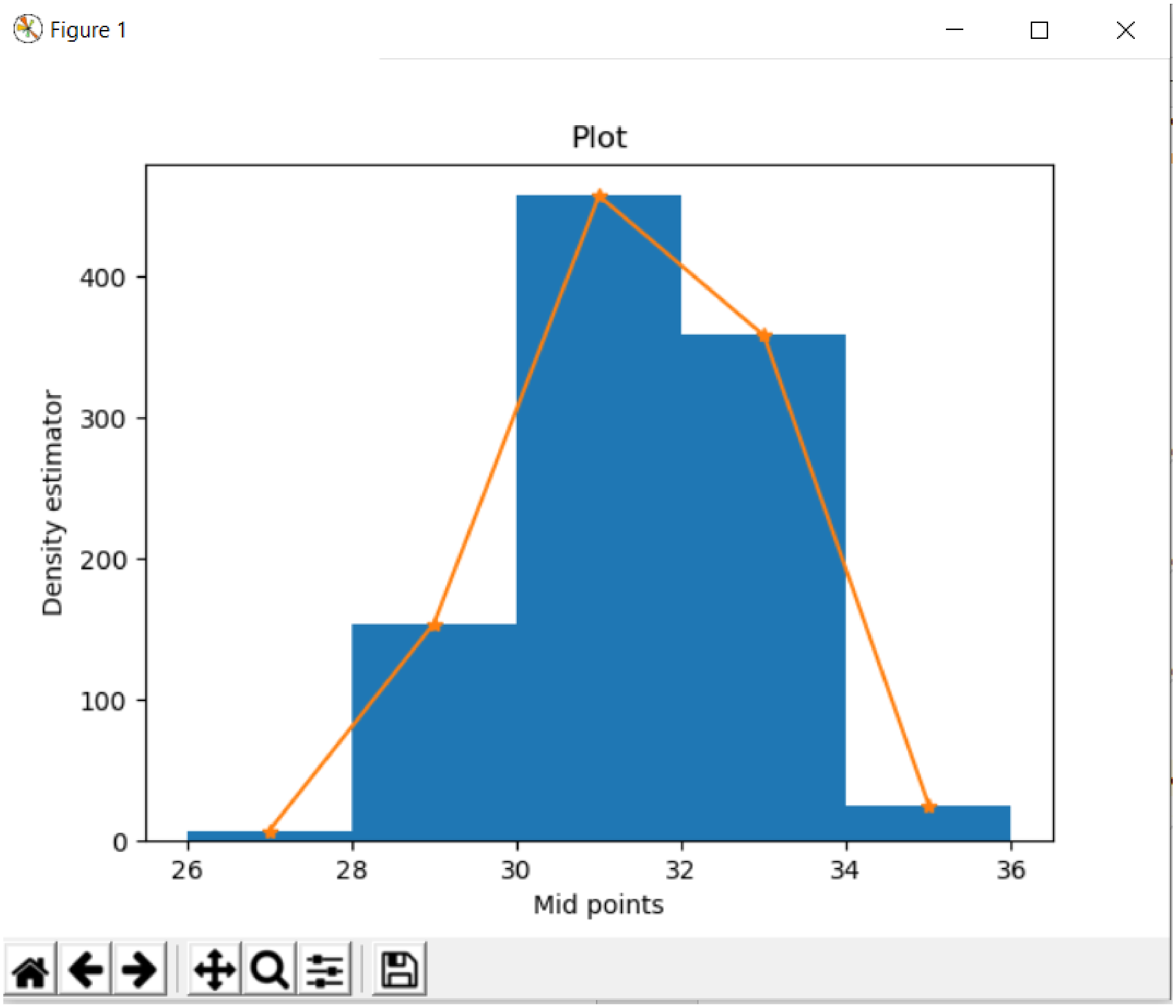
```
  mi       p(mi)
----  -----------
26.5  0.000999001
27.5  0.00699301
28.5  0.041958
29.5  0.126873
30.5  0.206793
31.5  0.267732
32.5  0.237762
33.5  0.0889111
34.5  0.01998
35.5  0.001998
```



g) Histogram for h=2

Coordinates of Density Estimator

```
  mi     p(mi)
----  ---------
  27  0.003996
  29  0.0844156
  31  0.237263
  33  0.163337
  35  0.010989
```



h) From the above calculations, the histogram with h=0.5 gives the best results. Thus histograms with minimum bandwidth gives best results.

2.a)

Five number summary of X

The minimum value is: 26.3

First quartile Q1 is 30.4

The median (Quartile 2) is 31.5

The third quartile is 32.4

Interquartile range is 2.0

b)

Five number summary of group 0

The minimum value is: 26.3

First quartile Q1 is 29.4

The median (Quartile 2) is 30.0

The third quartile is 30.6

Interquartile range is 1.200

Five number summary of group 1

The minimum value is: 29.1

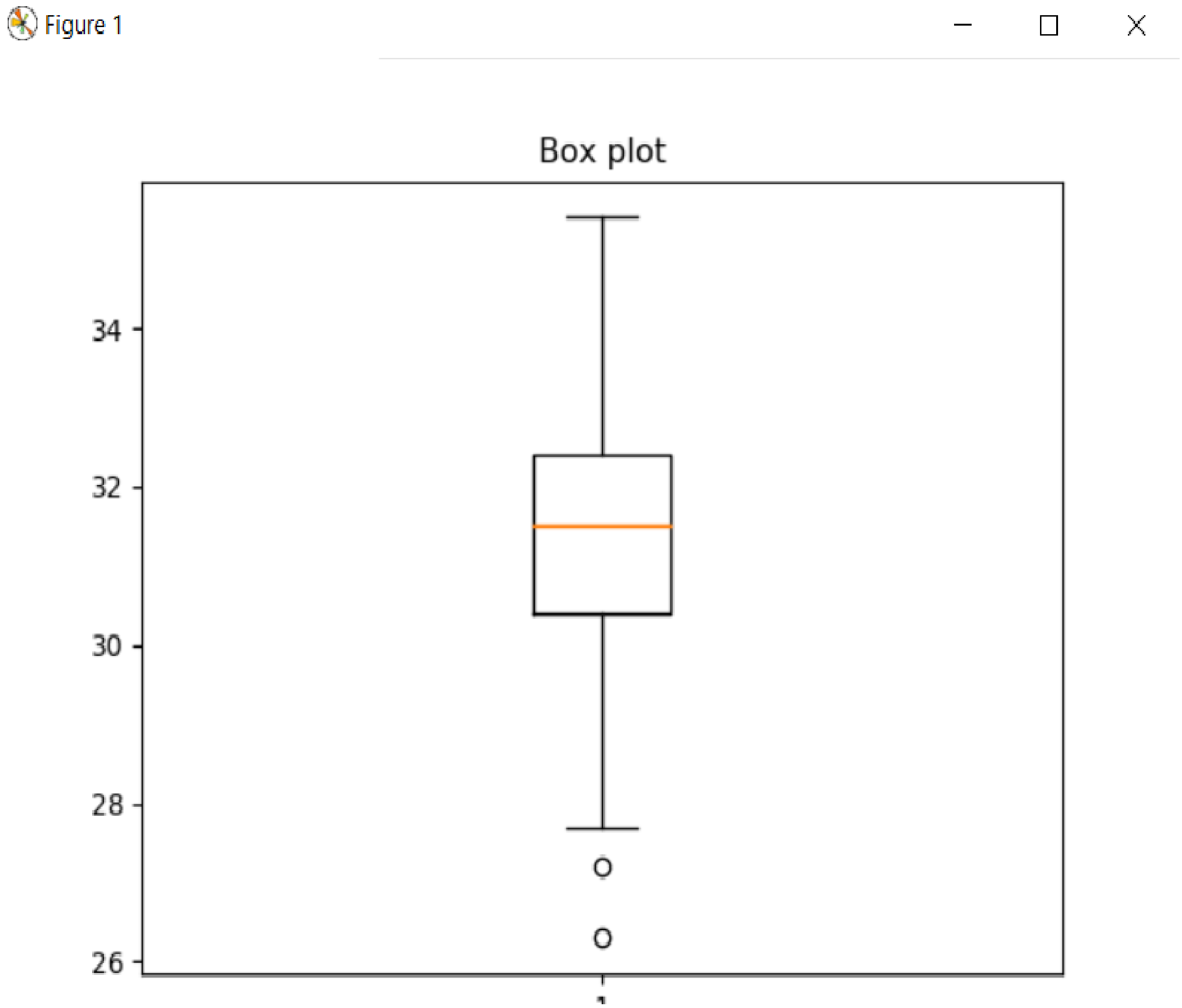
First quartile Q1 is 31.4

The median (Quartile 2) is 32.1

The third quartile is 32.7

Interquartile range is 1.300

c)



The whiskers from the box plot are

Lower whisker – 27

Upper whisker – 36

Thus the boxplot reports them incorrectly.

d) Outliers are

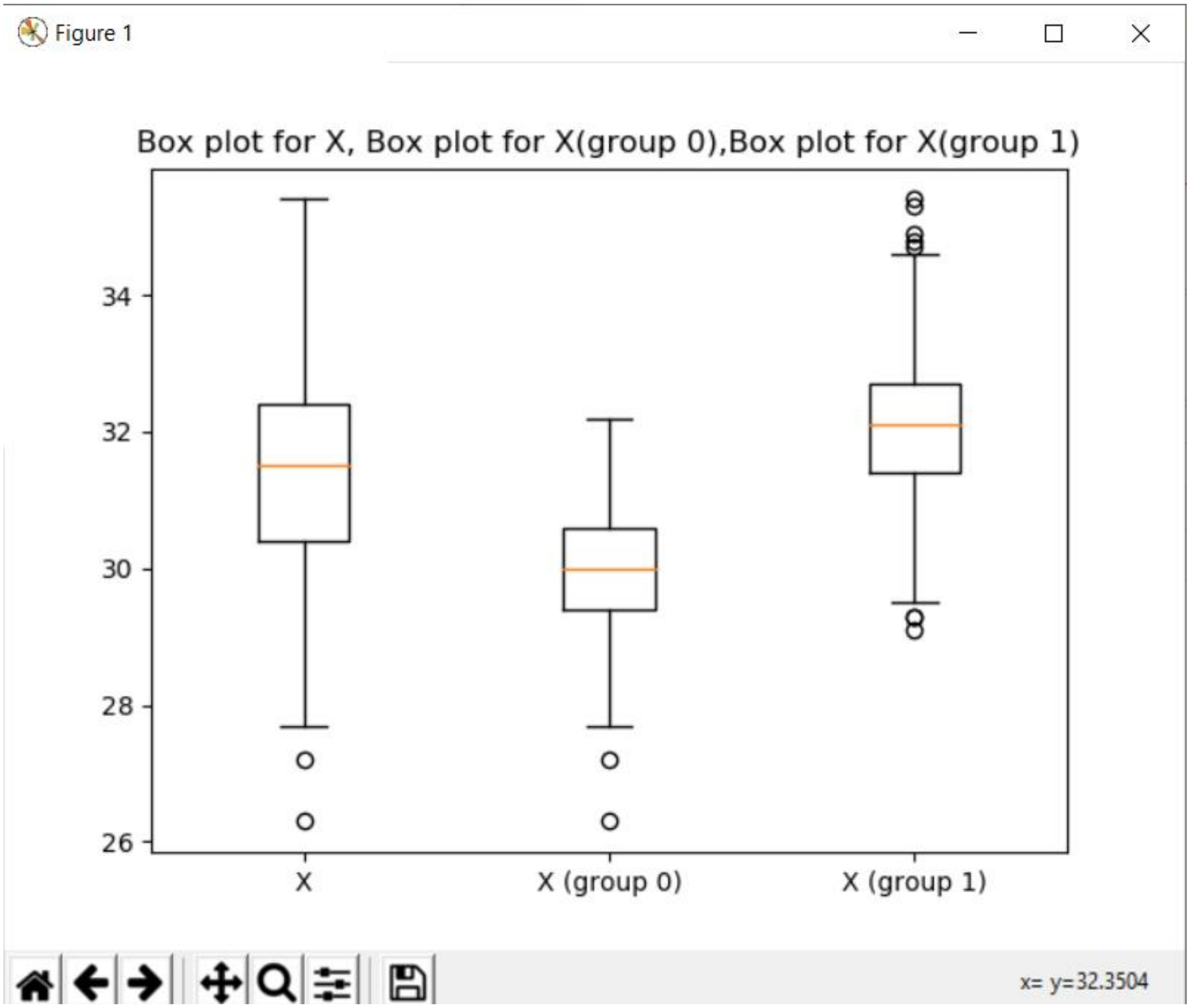


Figure 1
Box plot for X, Box plot for X(group 0),Box plot for X(group 1)
34
32
30
28
26
X
X (group 0)
X (group 1)
x= y=32.3504

outlier for X [26.3, 27.2]

outlier for group 0 [26.3, 27.2]

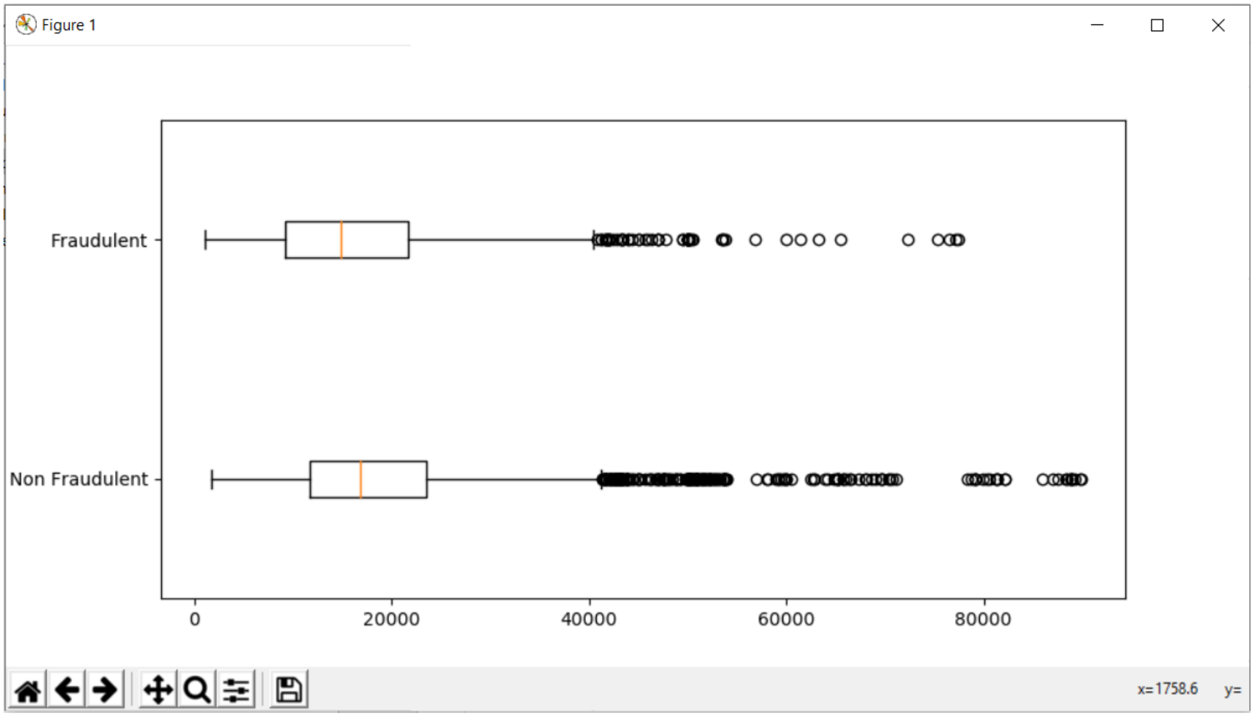
outlier for group 1 [29.1, 29.3, 29.3, 34.7, 34.8, 34.9, 35.3, 35.4]

3.

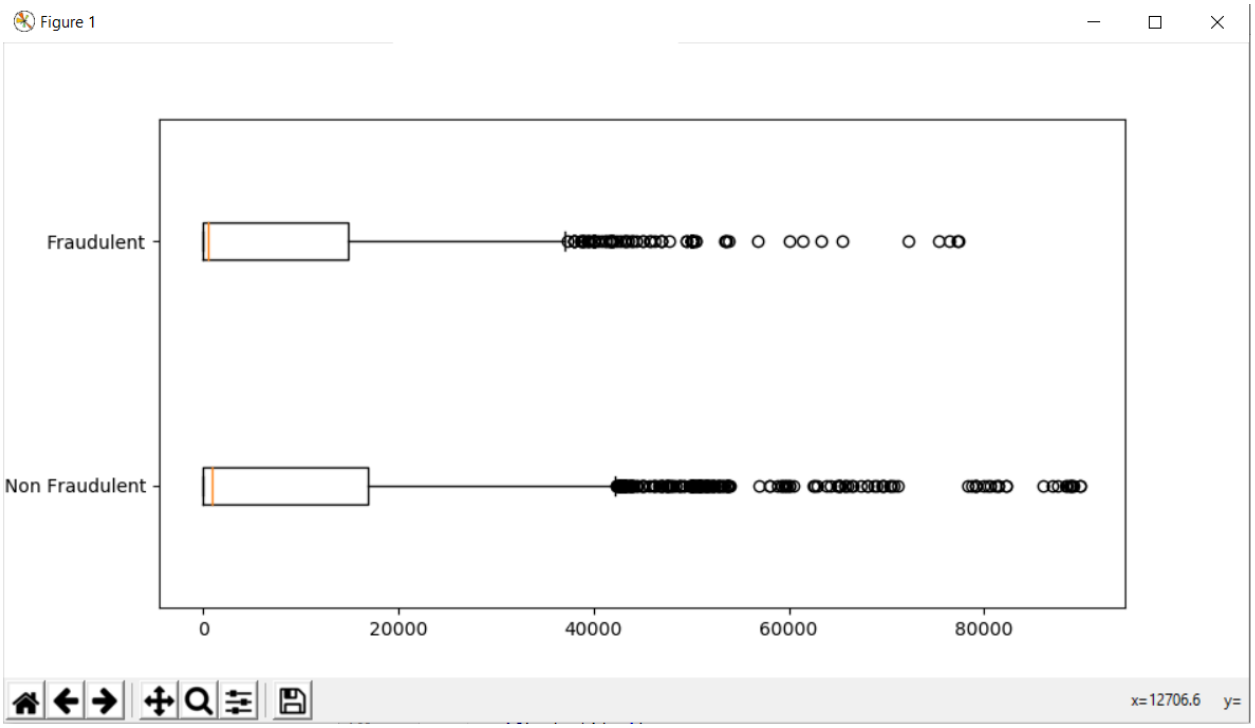
a)Fraudulent percentage = 19.95

b)

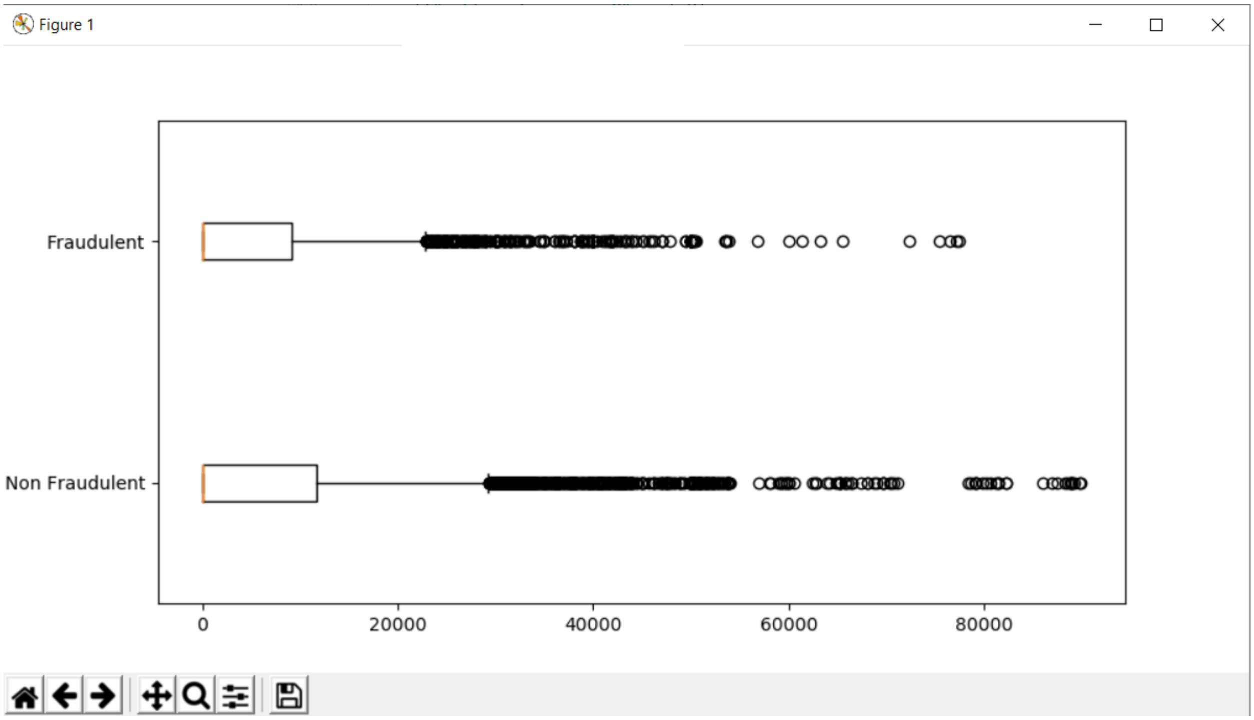
boxplot for total spent



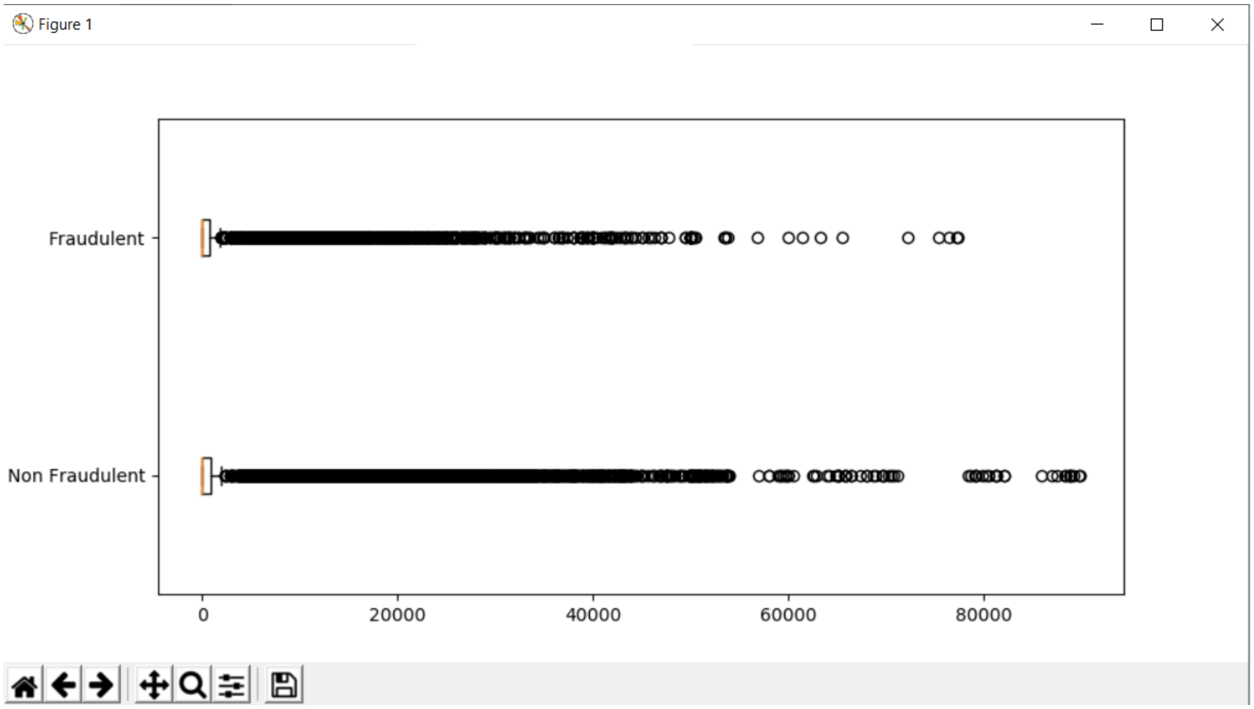
Boxplot for doctor visit



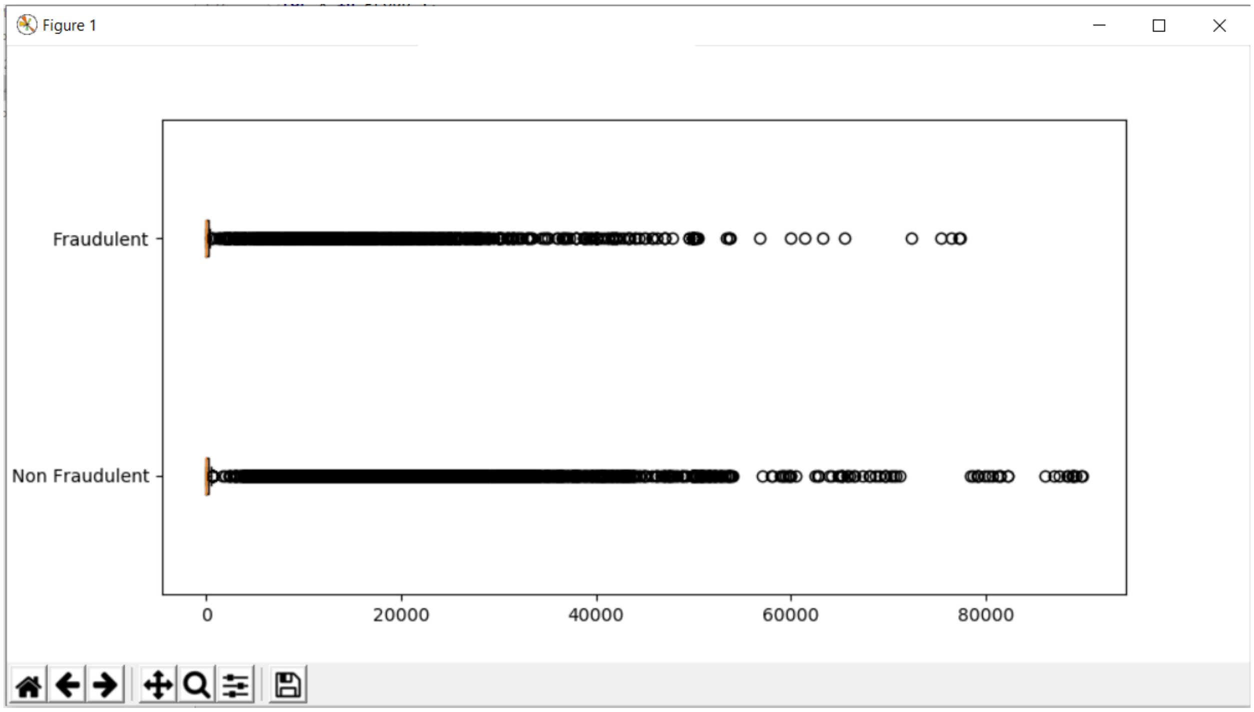
Box plot for number claims



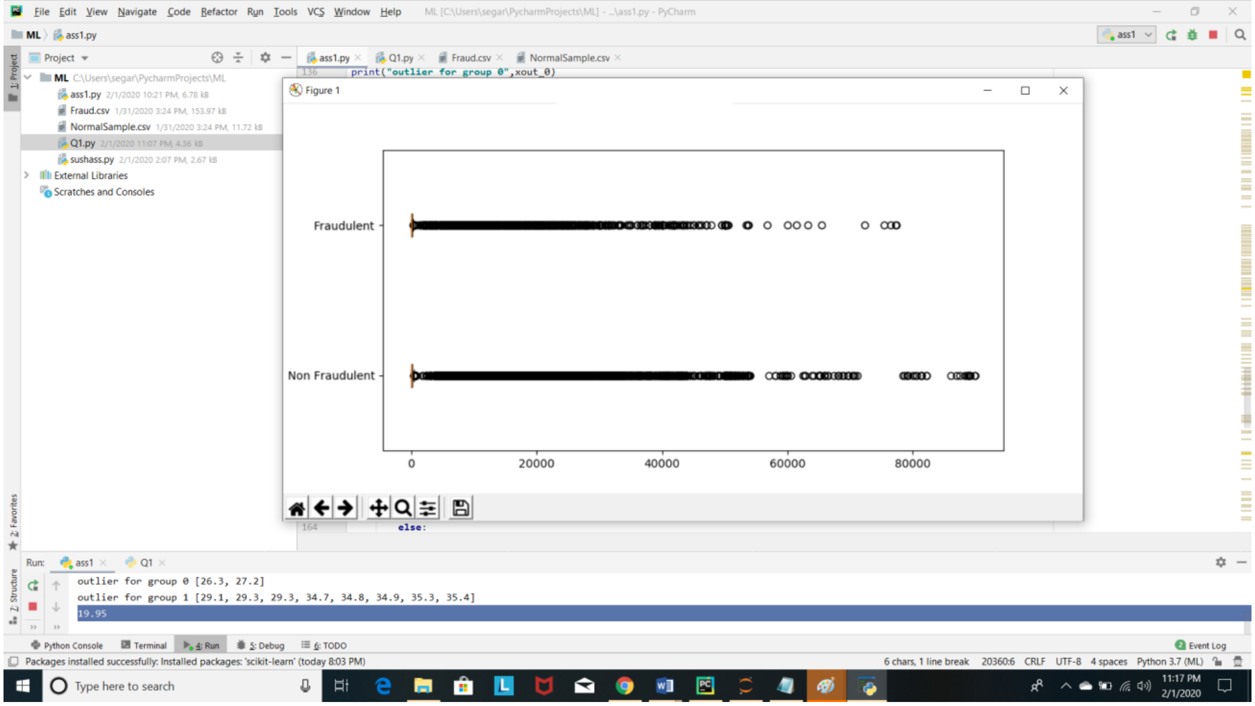
Box plot for member duration



Box plot for optom



Box plot for number of members


File Edit View Navigate Code Refactor Run Tools VCS Window Help
ML [C:\Users\segar\PycharmProjects\ML] - ...\ass1.py - PyCharm
ML
ass1.py
Project
ass1.py
Q1.py
Fraud.csv
NormalSample.csv
ML C:\Users\segar\PycharmProjects\ML
ass1.py 2/1/2020 10:21 PM, 6.78 kB
Fraud.csv 1/31/2020 3:24 PM, 153.97 kB
NormalSample.csv 1/31/2020 3:24 PM, 11.72 kB
Q1.py 2/1/2020 11:07 PM, 4.36 kB
sushass.py 2/1/2020 2:07 PM, 2.67 kB
External Libraries
Scratches and Consoles
Figure 1
Fraudulent
Non Fraudulent
0
20000
40000
60000
80000
Run:
ass1
Q1
Python Console
Terminal
4: Run
5: Debug
6: TODO
Event Log
Type here to search
11:17 PM
2/1/2020

```
print("outlier for group 0",xout_0)
```

```
else:
```

```
outlier for group 0 [26.3, 27.2]
outlier for group 1 [29.1, 29.3, 29.3, 34.7, 34.8, 34.9, 35.3, 35.4]
19.95
```

Packages installed successfully: Installed packages: 'scikit-learn' (today 8:03 PM)

6 chars, 1 line break 20360:6 CRLF UTF-8 4 spaces Python 3.7 (ML)

(The code for the following sums is referred from the code given by the professor (Nearest neighbor supervised and unsupervised algorithm, and Eigen value).

c)

i) Number of dimensions

t(x) * x =

```
 [[2812184770000    1040176400      42913200   20404919400     134771800
      220035900]
 [   1040176400       788159        23809      10264845        57654
        106717]
 [     42913200        23809         7922       448090         3459
          4765]
 [  20404919400     10264845       448090     232422585      1163391
       2121127]
 [    134771800        57654         3459      1163391        24460
         13581]
 [    220035900       106717         4765      2121127        13581
         29423]]
```

Eigenvalues of x =

```
 [6.84728061e+03 8.38798104e+03 1.80639631e+04 3.15839942e+05
 8.44539131e+07 2.81233324e+12]
```

Eigenvectors of x =

```
 [[-5.37750046e-06 -2.20900379e-05  3.62806809e-05 -1.36298664e-04
  -7.26453432e-03  9.99973603e-01]
 [ 6.05433402e-03 -2.69942162e-02  1.27528313e-02  9.99013423e-01
   3.23120126e-02  3.69879256e-04]
 [-9.82198935e-01  1.56454700e-01 -1.03312781e-01  1.14463687e-02
   1.62110700e-03  1.52596881e-05]
 [ 1.59310591e-04 -4.91894718e-03  3.11864824e-03 -3.25018102e-02
   9.99428355e-01  7.25592222e-03]
 [ 6.90939783e-02 -2.10615119e-01 -9.75101628e-01  6.26672294e-03
   2.19857585e-03  4.79234486e-05]
 [ 1.74569737e-01  9.64577791e-01 -1.95782843e-01  2.73038995e-02
   6.21788707e-03  7.82430481e-05]]
Number of Dimesions used is  6
```

ii)

Transformation matrix

```
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 Transformed x =  [[ 5.96859502e-03  1.02081629e-02 -6.64664861e-03  1.39590283e-02
   9.39352141e-03  6.56324665e-04]
 [-2.09672310e-02  5.01932025e-03  8.51930607e-04  5.16174400e-03
   1.22658834e-02  7.75702220e-04]
 [ 7.64597676e-03  1.97528525e-02 -7.38335310e-03 -1.71350853e-03
   1.50348109e-02  8.95075830e-04]
 ...
 [-7.18408819e-05 -1.62580211e-02  2.75078514e-02 -7.13245766e-03
  -4.74021952e-02  5.31896971e-02]
 [-1.80147801e-04 -1.62154130e-02  2.76213381e-02 -9.17125411e-03
  -4.76625006e-02  5.35474776e-02]
 [-2.21157680e-03 -2.73884697e-02  2.93391341e-02 -7.81347172e-03
  -4.70861917e-02  5.36071324e-02]]
```

The identity matrix is obtained as follows:

```
Expect an Identity Matrix =  [[ 1.00000000e+00 -2.16948855e-15  7.97972799e-17  7.65967151e-15
   1.04083409e-17 -2.98372438e-16]
 [-2.16948855e-15  1.00000000e+00 -2.33320308e-16 -1.92970639e-14
  -5.20417043e-16  7.49400542e-16]
 [ 7.97972799e-17 -2.33320308e-16  1.00000000e+00  4.57874840e-15
  -6.93889390e-17 -2.08166817e-16]
 [ 7.65967151e-15 -1.92970639e-14  4.57874840e-15  1.00000000e+00
   7.39339145e-15 -9.18015663e-15]
 [ 1.04083409e-17 -5.20417043e-16 -6.93889390e-17  7.39339145e-15
   1.00000000e+00 -5.82867088e-16]
 [-2.98372438e-16  7.49400542e-16 -2.08166817e-16 -9.18015663e-15
  -5.82867088e-16  1.00000000e+00]]
```

Since the product of the matrix and the transpose of the matrix gives the identity matrix, the matrix is orthonormal.

d)

i) The result of score function is 0.8779

ii) The score function gives the accuracy between the actual and the predicted value

e)

The focal observation is [7500, 15, 3, 127, 2, 2]

The Transformed focal observation is [[-0.02886529 0.00853837 -0.01333491 0.0176811   0.00793805 0.0044727]]

The indices of the five neighbors of the focal are [[588 2897 1199 1246 886]]

The input and target values of the nearest neighbors are

```
   ID  TOTAL_SPEND  DOCTOR_VISITS  ...  OPTOM_PRESC  NUM_MEMBERS  Target Value
```

```
0   588       7500         15  ...        2        2        1
1  2897      16000         18  ...        3        2        1
2  1199      10000         16  ...        2        1        1
3  1246      10200         13  ...        2        3        1
4   886       8900         22  ...        1        2        1

[5 rows x 8 columns]
```

f)

No of fraud observations / Total no of neighbors = 5/5 =1

Thus the focal is fraudulent

Also the focal is in the training data and the target value is also 1. Thus observation is not misclassified.