CS 584-04: Machine Learning

Fall 2018 Assignment 1

Suhas Sreenivas

CWID: A20423132

Question 1 (40 points)

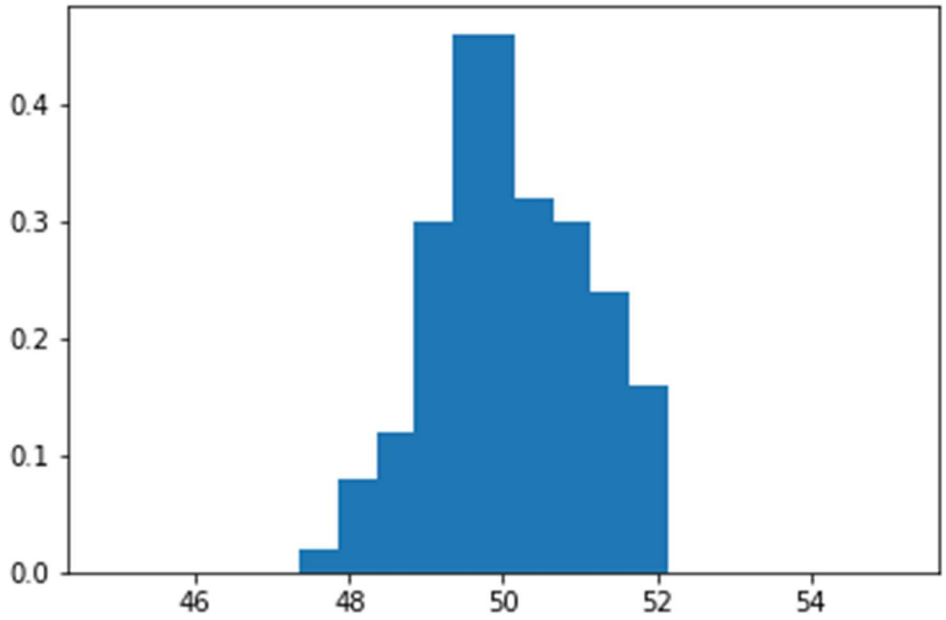
a) (4 points) According to Izenman (1991) method, what is the recommended bin-width for the histogram of x?

ANS:
h = 0.6237088427642294

b) (3 points) What is the bin-width after applying the beautification step?
ANS: 0.1

c) (10 points) Use h = 0.5, minimum = 45 and maximum = 55. List the coordinates of the density estimator. Paste the histogram drawn using Python or your favorite graphing tools.

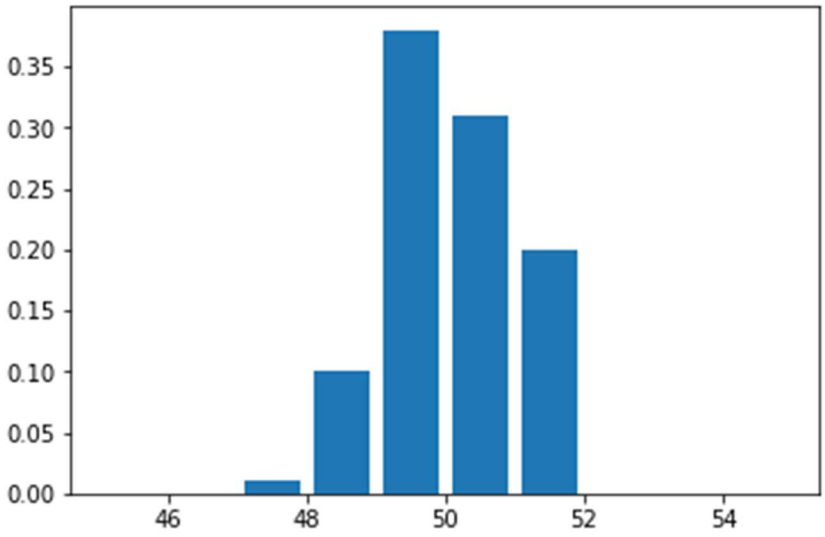
ANS: Coordinates of density estimates for h= 0.5 :
[0.0, 0.0, 0.0, 0.0, 0.0, 0.02, 0.08, 0.12, 0.3, 0.46, 0.32, 0.3, 0.24, 0.16, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0]



d) (10 points) Use h = 1, minimum = 45 and maximum = 55. List the coordinates of the density estimator. Paste the histogram drawn using Python or your favorite graphing tools.

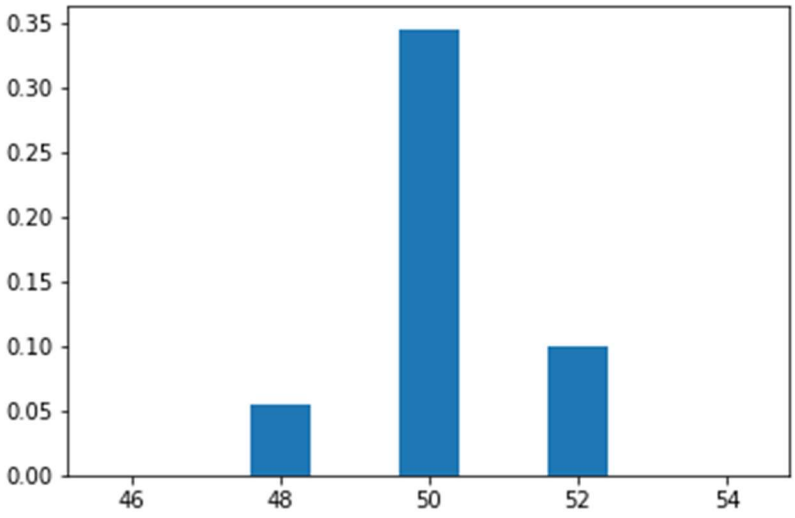
ANS: Coordinates of density estimates for h= 1 :

[0.0, 0.0, 0.01, 0.1, 0.38, 0.31, 0.2, 0.0, 0.0, 0.0]



e) (10 points) Use h = 2, minimum = 45 and maximum = 55. List the coordinates of the density estimator. Paste the histogram drawn using Python or your favorite graphing tools.

ANS: Coordinates of density estimates for h= 2:
[0.0, 0.055, 0.345, 0.1, 0.0]



f) (3 points) Among the three histograms, which one, in your opinions, can best describe the distribution of the field x?
The extent of the gradients is captured lesser by each subsequent histogram. However, the second one (h=1) captures it sufficiently and the first one (h=0.5) captures it in detail.
The first one seems to best describe the data but the second isn't too far behind.

Question 2 (20 points)

a) (2 points) What are the five-number summary of x?

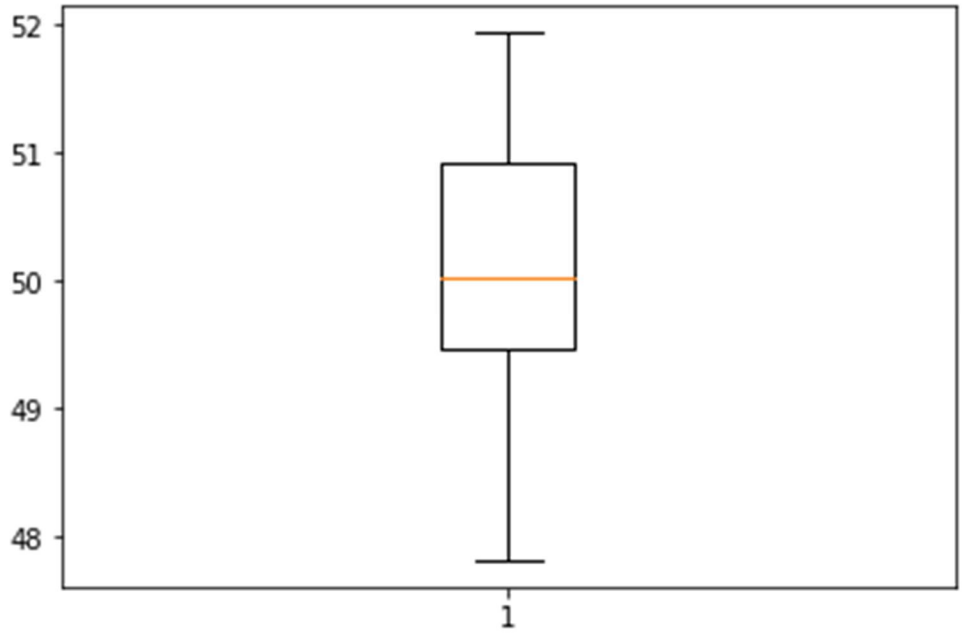
Min = 47.82
q-25 = 49.4675
q-50 = 50.03
q-75 = 50.915
max = 51.94

b) (3 points) What are the five-number summary of x for each category of Group?

Group 0:
Min = 47.82
q-25 = 49.295
q-50 = 50.22
q-75 = 50.96
max = 51.94

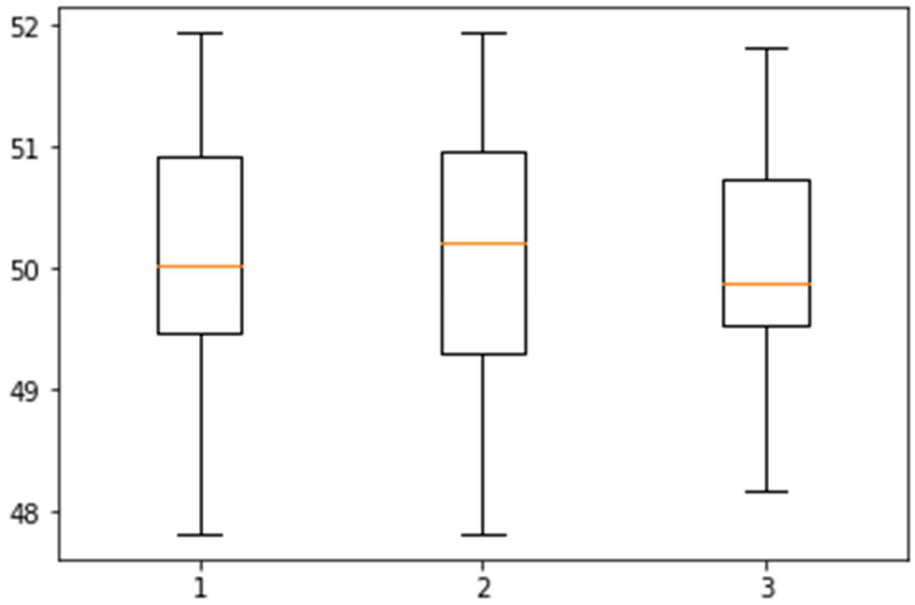
Group 1:
Min = 48.17
q-25 = 49.53
q-50 = 49.88
q-75 = 50.74
max = 51.82

c) (5 points) Draw a boxplot of x (without Group) using the Python boxplot function. Can you tell if the Python's boxplot has displayed the 1.5 IQR whiskers?



Python's (matplotlib's) boxplot displays 1.5 IQR as one of the parameter, 'whis' is by default set to 1.5 according to the official documentation.

d) (10 points) Draw a graph where it contains the boxplot of x, the boxplot of x for each category of Group (i.e., three boxplots within the same graph frame). Use the 1.5 IQR whiskers, identify the outliers of x, if any, for the entire data and for each category of Group.



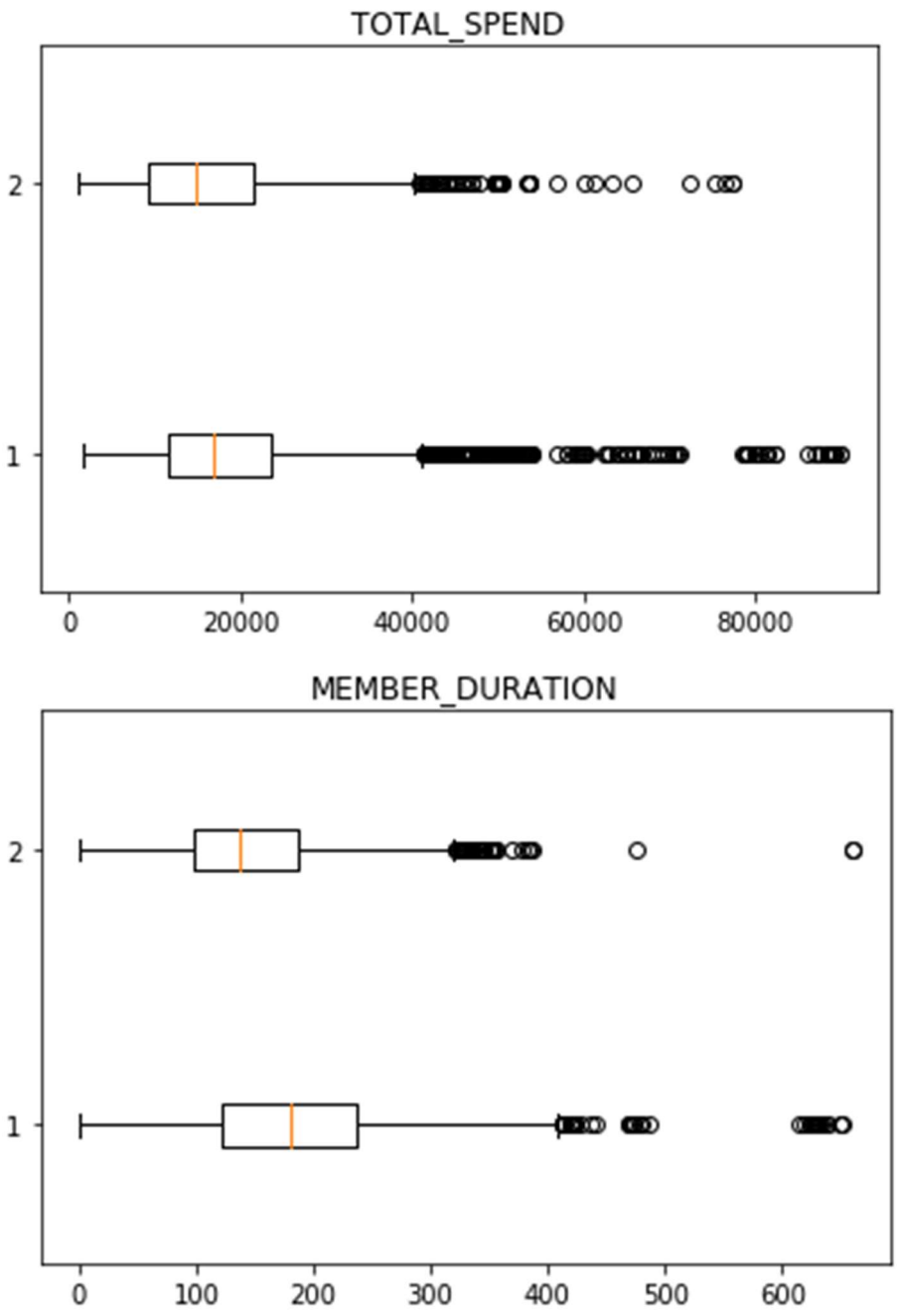
There are no outliers for the data

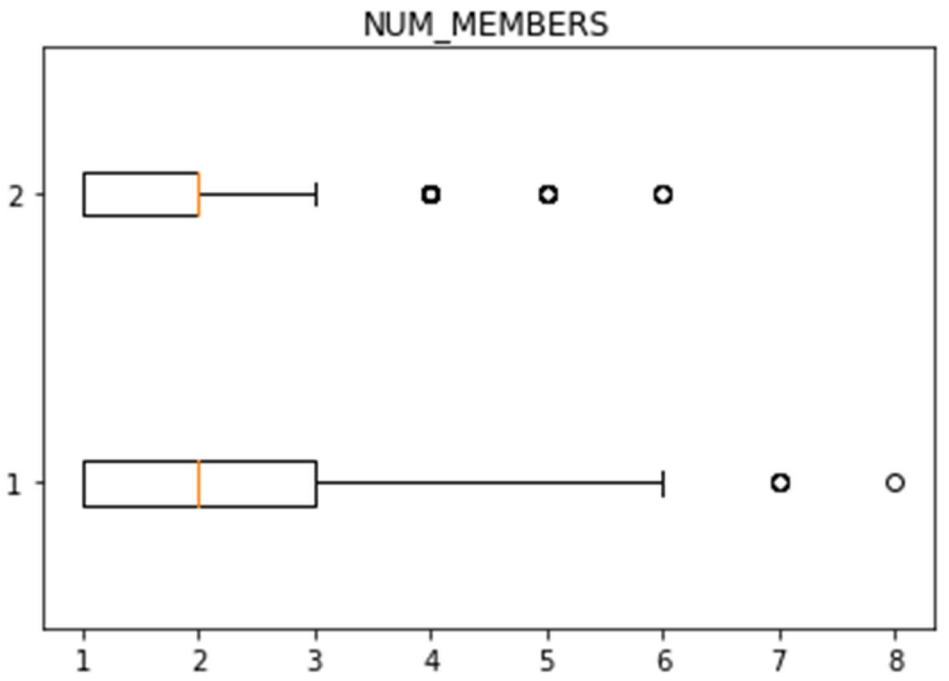
Question 3 (40 points)

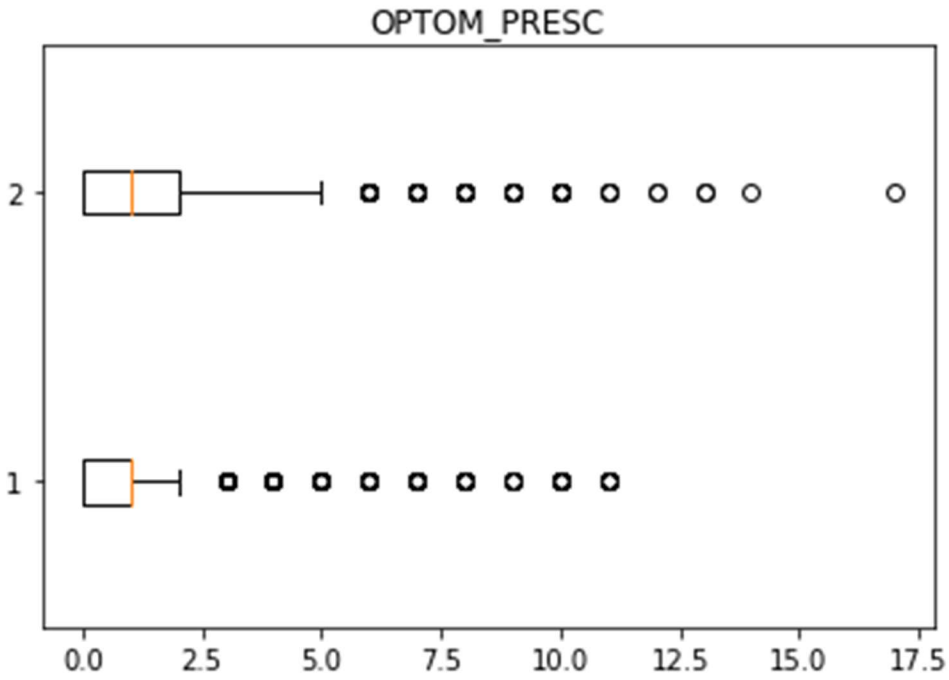
a) (5 points) What percent of investigations are found to be fraudulent? Please give your answer up to 4 decimal places.

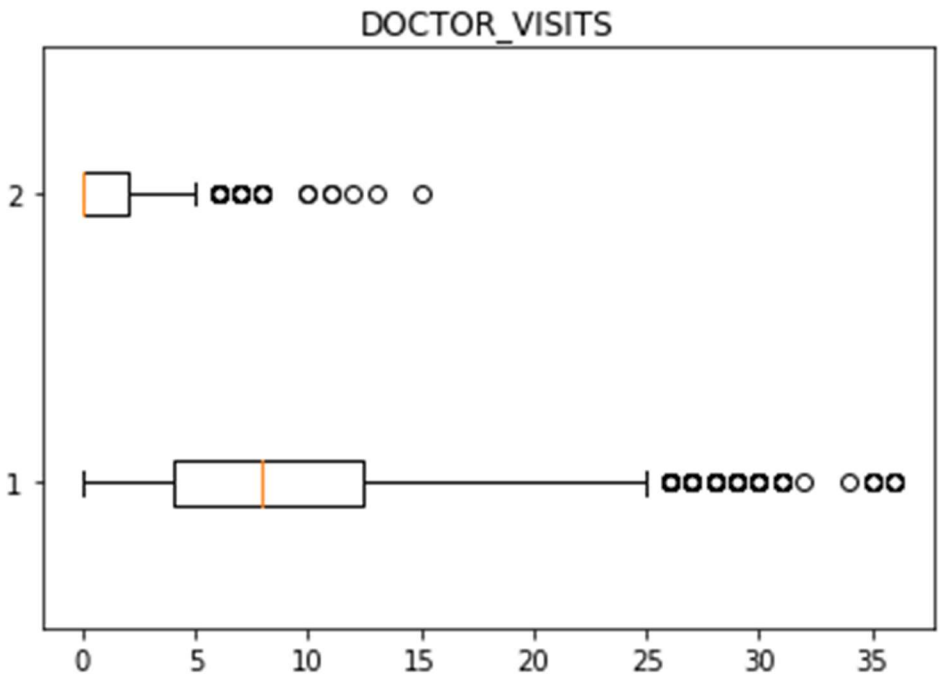
ANS: Fraud % = 19.949664429530202

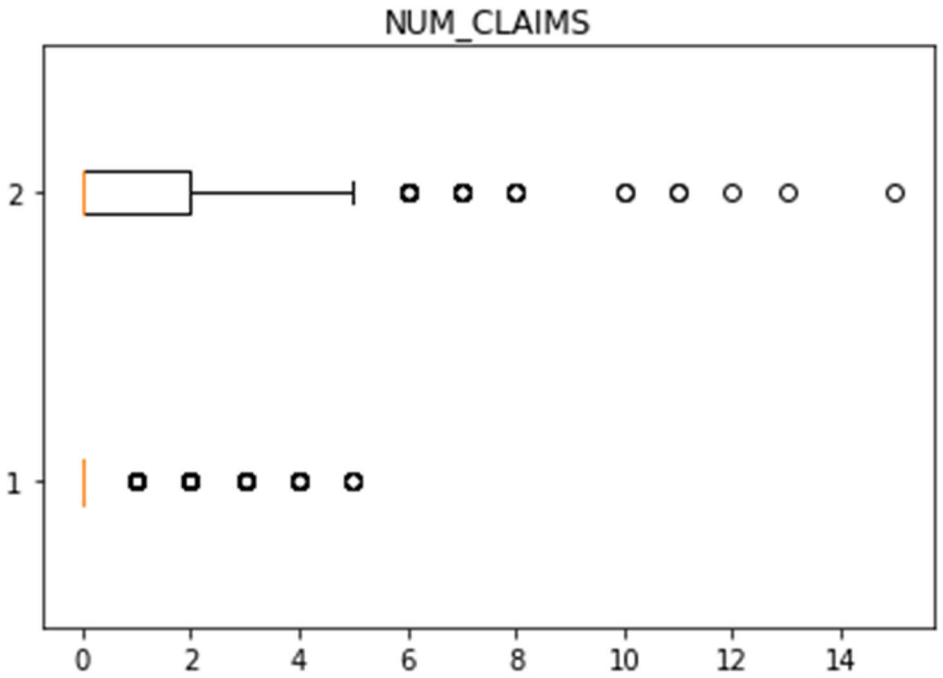
b) (5 points) Use the BOXPLOT function to produce horizontal box-plots. For each interval variable, one box-plot for the fraudulent observations, and another box-plot for the non-fraudulent observations. These two box-plots must appear in the same graph for each interval variable.


TOTAL_SPEND
2
1
0
20000
40000
60000
80000
MEMBER_DURATION
2
1
0
100
200
300
400
500
600


NUM_MEMBERS
2
1
1
2
3
4
5
6
7
8


OPTOM_PRESC
2
1
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5


DOCTOR_VISITS
2
1
0
5
10
15
20
25
30
35


NUM_CLAIMS
2
1
0
2
4
6
8
10
12
14

c) (10 points) Orthonormalize interval variables and use the resulting variables for the nearest neighbor analysis. Use only the dimensions whose corresponding eigenvalues are greater than one.

i. (5 points) How many dimensions are used?
ANS: 2

ii. (5 points) Please provide the transformation matrix? You must provide proof that the resulting variables are actually orthonormal.

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 transpose of the transformed matrix multiplied by the transformed matrix should be an identity matrix
```
xtx = np.matmul(transf_x.transpose(),transf_x);
```

In the given code xtx is an identity matrix. Exerpt from the code output:
Expect an Identity Matrix =
```
[[ 1. -0.  0.  0.  0. -0.]
 [-0.  1. -0. -0. -0.  0.]
 [ 0. -0.  1.  0. -0. -0.]
 [ 0. -0.  0.  1.  0. -0.]
 [ 0. -0. -0.  0.  1. -0.]
 [-0.  0. -0. -0. -0.  1.]]
```

d) (10 points) Use the NearestNeighbors module to execute the Nearest Neighbors algorithm using exactly <u>five</u> neighbors and the resulting variables you have chosen in c). The KNeighborsClassifier module has the score function.

i. (5 points) Run this function, provide the function return value
Score: 0.8778523489932886

ii. (5 points) Explain the meaning of the function return value.
The score represents the mean accuracy on the given test data and labels. (from scikit-learn documentation)

e) (5 points) For the observation which has these input variable values: TOTAL_SPEND = 7500, DOCTOR_VISITS = 15, NUM_CLAIMS = 3, MEMBER_DURATION = 127, OPTOM_PRESC = 2, and NUM_MEMBERS = 2, find its **five** neighbors. Please list their input variable values and the target values.

CASE_ID 589
FRAUD 1
TOTAL_SPEND 7500
DOCTOR_VISITS 15
NUM_CLAIMS 3
MEMBER_DURATION 127
OPTOM_PRESC 2
NUM_MEMBERS 2
Name: 588, dtype: int64

CASE_ID 2898
FRAUD 1
TOTAL_SPEND 16000
DOCTOR_VISITS 18
NUM_CLAIMS 3
MEMBER_DURATION 146
OPTOM_PRESC 3
NUM_MEMBERS 2
Name: 2897, dtype: int64

CASE_ID 1200
FRAUD 1
TOTAL_SPEND 10000
DOCTOR_VISITS 16
NUM_CLAIMS 3
MEMBER_DURATION 124
OPTOM_PRESC 2
NUM_MEMBERS 1
Name: 1199, dtype: int64

```
CASE_ID          1247
FRAUD               1
TOTAL_SPEND      10200
DOCTOR_VISITS       13
NUM_CLAIMS           3
MEMBER_DURATION    119
OPTOM_PRESC          2
NUM_MEMBERS          3
Name: 1246, dtype: int64
```

```
CASE_ID          887
FRAUD              1
TOTAL_SPEND     8900
DOCTOR_VISITS     22
NUM_CLAIMS         3
MEMBER_DURATION  166
OPTOM_PRESC        1
NUM_MEMBERS        2
Name: 886, dtype: int64
```

f) (5 points) Follow-up with e), what is the predicted probability of fraudulent (i.e., FRAUD = 1)? If your predicted probability is greater than or equal to your answer in a), then the observation will be classified as fraudulent. Otherwise, non-fraudulent. Based on this criterion, will this observation be misclassified?

The predicted probability of a fraud for the sample in the previous question is 1, which is greater than 19%. The above sample will not be misclassified.