CS 584-04: Machine Learning

Fall 2019 Assignment 1

Question 1 (40 points)

Write a Python program to calculate the density estimator of a histogram. Use the field *x* in the NormalSample.csv file.

a) (5 points) According to Izenman (1991) method, what is the recommended bin-width for the histogram of x?

Ans:

0.3998 => 0.4 after rounding

b) (5 points) What are the minimum and the maximum values of the field x?

Ans:

Minimum value of x = 26.3
Maximum value of x = 35.4

c) (5 points) Let a be the largest integer less than the minimum value of the field x, and b be the smallest integer greater than the maximum value of the field x. What are the values of a and b?

Ans:

Value of a = 26
Value of b = 36

d) (5 points) Use h = 0.1, minimum = a and maximum = b. List the coordinates of the density estimator. Paste the histogram drawn using Python or your favorite graphing tools.

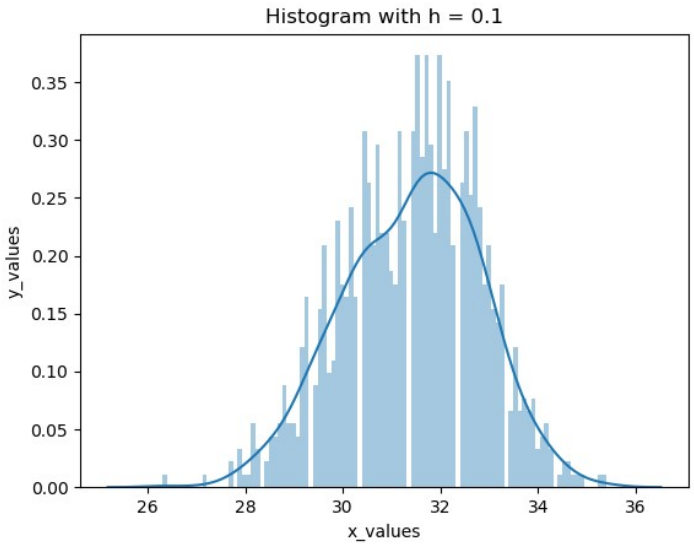
Ans:

Coordinates of the density estimator are as follows:

(26.05, 0.0), (26.150000000000002, 0.0),
(26.250000000000004, 0.009990009990009999),
(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, 0.009990009990009999),
(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),
(32.150000000000084, 0.3196803196803197),
(32.250000000000085, 0.1898101898101898),
(32.35000000000009, 0.23976023976023975),
(32.45000000000009, 0.2797202797202797),

(32.55000000000009, 0.22977022977022976),
(32.65000000000009, 0.29970029970029965),
(32.75000000000009, 0.21978021978021975),
(32.850000000000094, 0.15984015984015984),
(32.950000000000095, 0.1898101898101898),
(33.0500000000001, 0.13986013986013984),
(33.1500000000001, 0.12987012987012986),
(33.2500000000001, 0.15984015984015984),
(33.3500000000001, 0.05994005994005994),
(33.4500000000001, 0.10989010989010987),
(33.550000000000104, 0.05994005994005994),
(33.650000000000105, 0.06993006993006992),
(33.75000000000011, 0.05994005994005994),
(33.85000000000011, 0.06993006993006992),
(33.95000000000011, 0.02997002997002997),
(34.05000000000011, 0.03996003996003996),
(34.15000000000011, 0.049950049950049945),
(34.250000000000114, 0.02997002997002997),
(34.350000000000115, 0.00999000999000999),
(34.45000000000012, 0.01998001998001998),
(34.55000000000012, 0.01998001998001998),
(34.65000000000012, 0.00999000999000999),
(34.75000000000012, 0.00999000999000999),
(34.85000000000012, 0.00999000999000999),
(34.950000000000124, 0.0), (35.050000000000125, 0.0),
(35.15000000000013, 0.0),
(35.25000000000013, 0.00999000999000999),
(35.35000000000013, 0.00999000999000999),
(35.45000000000013, 0.0), (35.55000000000013, 0.0),
(35.650000000000134, 0.0), (35.750000000000135, 0.0),
(35.850000000000136, 0.0), (35.95000000000014, 0.0)

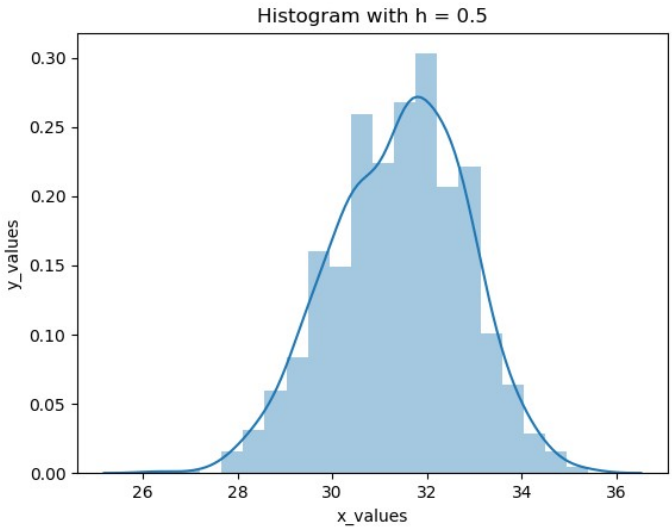


e) (5 points) Use h = 0.5, minimum = a and maximum = b. List the coordinates of the density estimator. Paste the histogram drawn using Python or your favorite graphing tools.

Ans:

Coordinates of the density estimator are as follows:

(26.25, 0.001998001998001998), (26.75, 0.0),
(27.25, 0.001998001998001998), (27.75, 0.011988011988011988),
(28.25, 0.029970029970029972), (28.75, 0.053946053946053944),
(29.25, 0.1038961038961039), (29.75, 0.14985014985014986),
(30.25, 0.2077922077922078), (30.75, 0.2057942057942058),
(31.25, 0.25374625374625376), (31.75, 0.2817182817182817),
(32.25, 0.25574425574425574), (32.75, 0.21978021978021978),
(33.25, 0.11988011988011989), (33.75, 0.057942057942057944),
(34.25, 0.029970029970029972), (34.75, 0.00999000999000999),
(35.25, 0.003996003996003996), (35.75, 0.0)



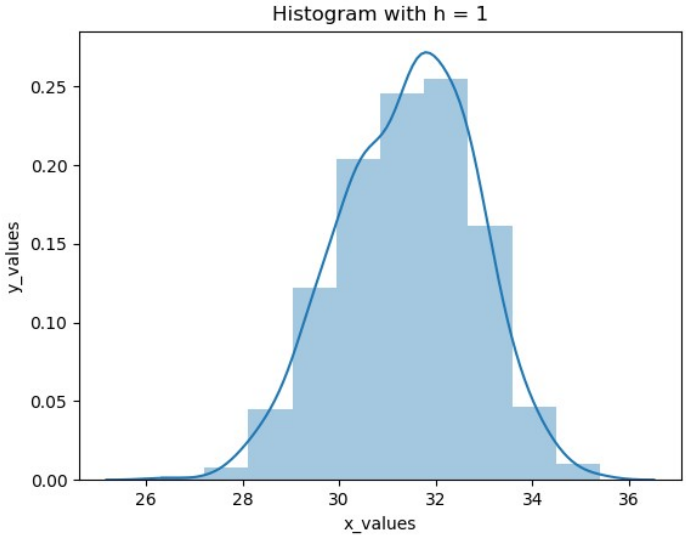
f) (5 points) Use h = 1, minimum = a and maximum = b. List the coordinates of the density estimator. Paste the histogram drawn using Python or your favorite graphing tools.

Ans:

Coordinates of the density estimator are as follows:

(26.5, 0.000999000999000999), (27.5,
0.006993006993006993), (28.5, 0.04195804195804196), (29.5,
0.12687312687312688),

(30.5, 0.20679320679320679), (31.5, 0.2677322677322677),
(32.5, 0.23776223776223776), (33.5, 0.08891108891108891),
(34.5, 0.01998001998001998), (35.5, 0.001998001998001998)

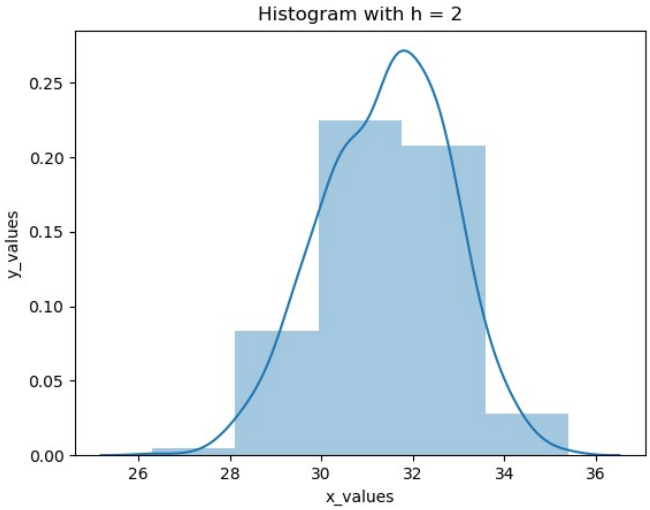


g) (5 points) Use h = 2, minimum = a and maximum = b. List the coordinates of the density estimator. Paste the histogram drawn using Python or your favorite graphing tools.

Ans:

Coordinates of the density estimator are as follows:

(27.0, 0.003996003996003996), (29.0, 0.08441558441558442),
(31.0, 0.23726273726273725), (33.0, 0.16333666333666333),
(35.0, 0.01098901098901099)



h) (5 points) Among the four histograms, which one, in your honest opinions, can best provide your insights into the shape and the spread of the distribution of the field x? Please state your arguments.

Ans:

I believe the one with h=0.5 will be able to provide more insight into the distribution of the field x. This is because of a few reasons, first is that the resulting histogram in this case manages to show just enough detail and at the same time refrains from losing the big picture. For example, in the case where h=0.1 the bin-width is so small that it practically loses the whole point of observing data through histograms. In the case of h=1, this histogram can also be considered a good representative of distribution of x but it misses out on some details due to its slightly more generalized nature; details such as the bin in which the maximum value of x may lie can be more accurately described when observing the h=0.5 histogram. In the case of h=2, it becomes extremely hard to practically deduce accurate information about x due to its over-generalization. h=0.5 hits that sweet spot by neither going too accurate nor being too generalized. Apart from this, 0.5 also comes the closest, when compared to others in this scenario, to the value of h obtained using Izenman's method which is 0.4.

Question 2 (20 points)

Use in the NormalSample.csv to generate box-plots for answering the following questions.

a) (5 points) What is the five-number summary of x? What are the values of the 1.5 IQR whiskers?

Ans:

The five-number summary of x is as follows:

Minimum of Sample = 26.3
First Quartile of Sample = 30.4
Median of Sample = 31.5
Third Quartile of Sample = 32.4
Maximum of Sample = 35.4

Value of Left Whisker = 27.4
Value of Right Whisker = 35.4

b) (5 points) What is the five-number summary of x for each category of the group? What are the values of the 1.5 IQR whiskers for each category of the group?

Ans:

For Category 0:

Five-number summary of x:

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

Value of Left Whisker = 27.599
Value of Right Whisker = 32.2

For Category 1:

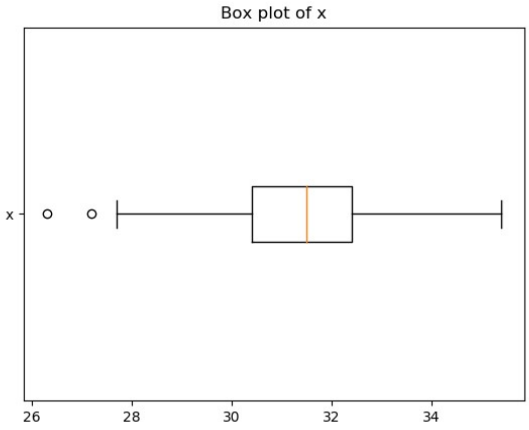
Five-number summary of x:

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

Value of Left Whisker = 29.4499
Value of Right Whisker = 34.65

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

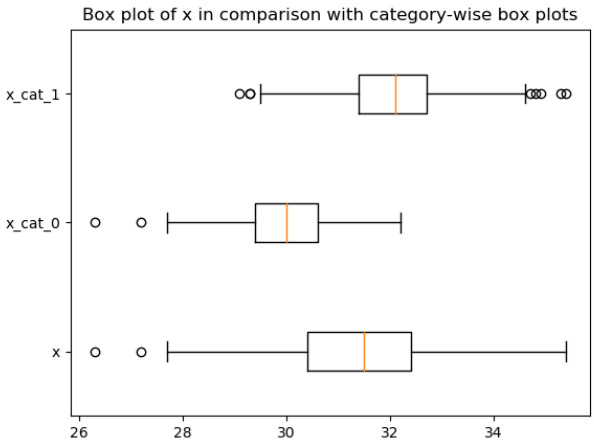
Ans:



By observing the box plot it can be seen that this function has displayed the whiskers correctly

d) (5 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 the group.
Hint: Consider using the CONCAT function in the PANDA module to append observations.

Ans:



Outliers for the entire data are:
[27.2, 26.3]

Outliers for Category 0:
[27.2, 26.3]

Outliers for Category 1:
[35.3, 29.3, 35.4, 34.9, 34.7, 34.8, 29.3, 29.1]

Question 3 (40 points)

The data, FRAUD.csv, contains results of fraud investigations of 5,960 cases. The binary variable FRAUD indicates the result of a fraud investigation: 1 = Fraudulent, 0 = Otherwise. The other interval variables contain information about the cases.

1. TOTAL_SPEND: Total amount of claims in dollars
2. DOCTOR_VISITS: Number of visits to a doctor
3. NUM_CLAIMS: Number of claims made recently
4. MEMBER_DURATION: Membership duration in number of months
5. OPTOM_PRESC: Number of optical examinations
6. NUM_MEMBERS: Number of members covered

You are asked to use the Nearest Neighbors algorithm to predict the likelihood of fraud.

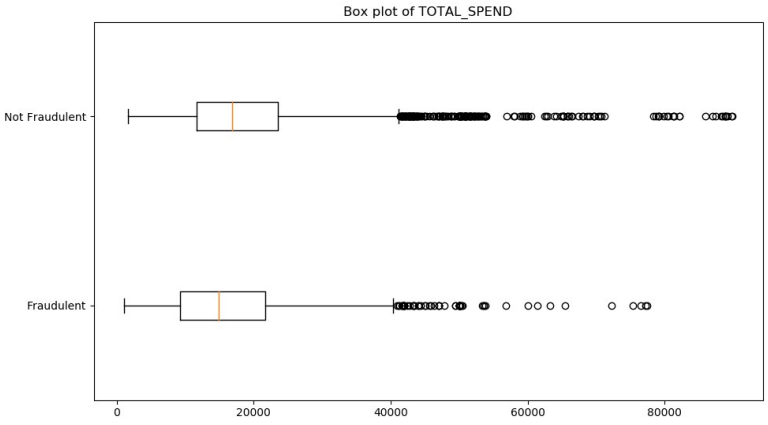
a) (5 points) What percent of investigations are found to be fraudulent? Please give your answer up to 4 decimal places.

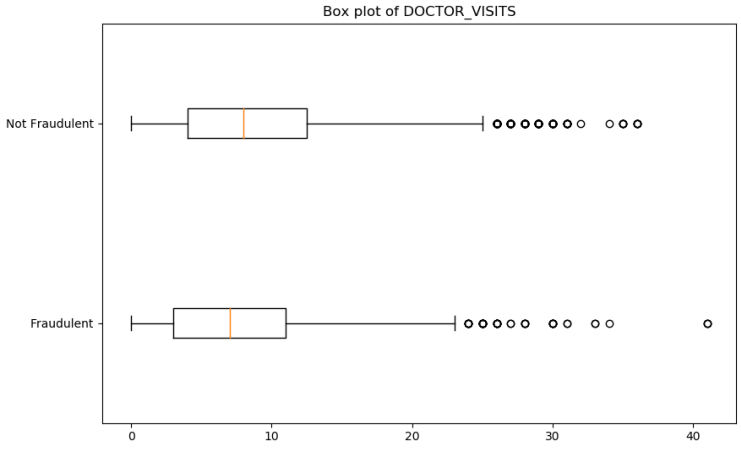
Ans:

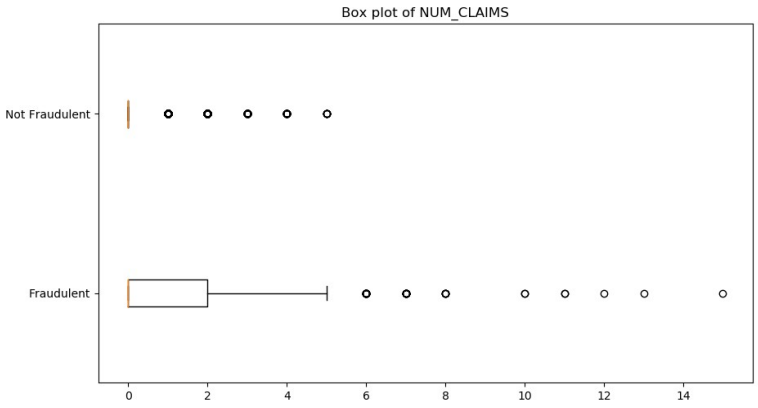
Fraudulent Investigations' percentage = 19.9496%

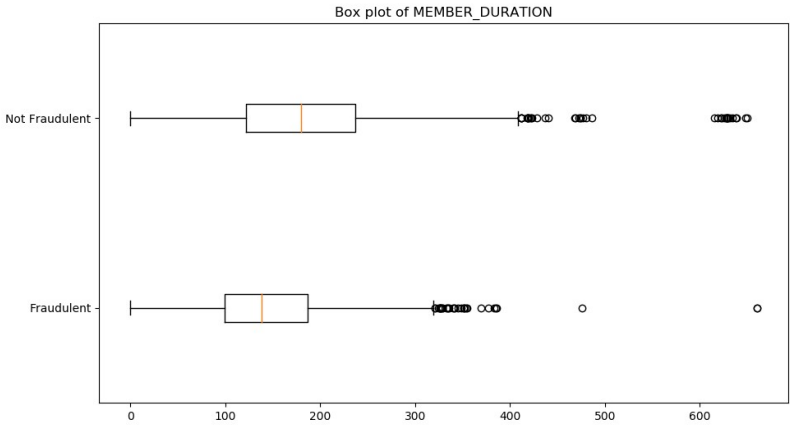
b) (5 points) Use the BOXPLOT function to produce horizontal box-plots. For each interval variable, one box-plot for the fraudulent observations, and another box-plot for the non-fraudulent observations. These two box-plots must appear in the same graph for each interval variable.

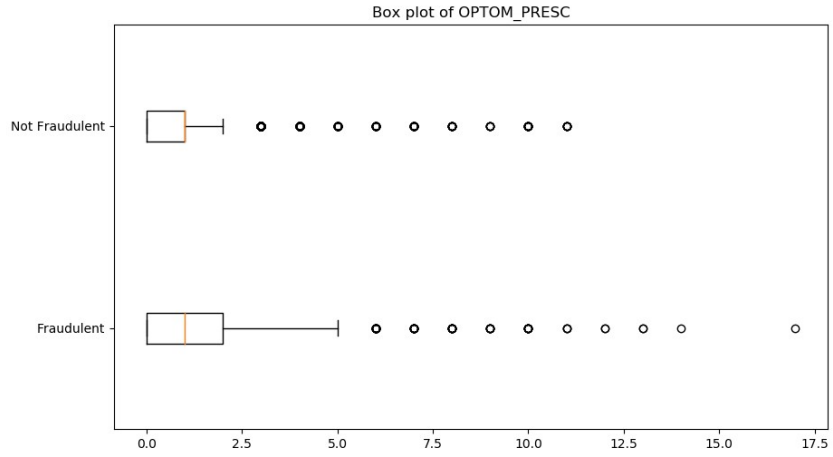
Ans:

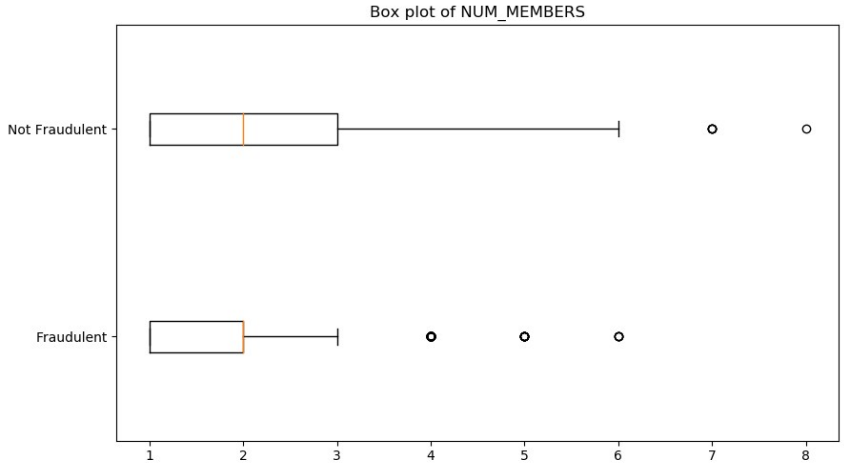



Box plot of DOCTOR_VISITS
Not Fraudulent
Fraudulent
0
10
20
30
40


Box plot of NUM_CLAIMS
Not Fraudulent
Fraudulent
0
2
4
6
8
10
12
14


Box plot of MEMBER_DURATION
Not Fraudulent
Fraudulent
0
100
200
300
400
500
600





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

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 a score function.

   i. (5 points) Run the score function, provide the function return value
   ii. (5 points) Explain the meaning of the score function return value.

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. *Reminder: transform the input observation using the results in c) before finding the neighbors*.

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?