CS 584: Machine Learning

Spring 2020 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. **Bin width: 0.4**

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

Ans**. Minimum Value: 26.3**

Maximum Value: 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**. a: 26.0**

b: 36.0

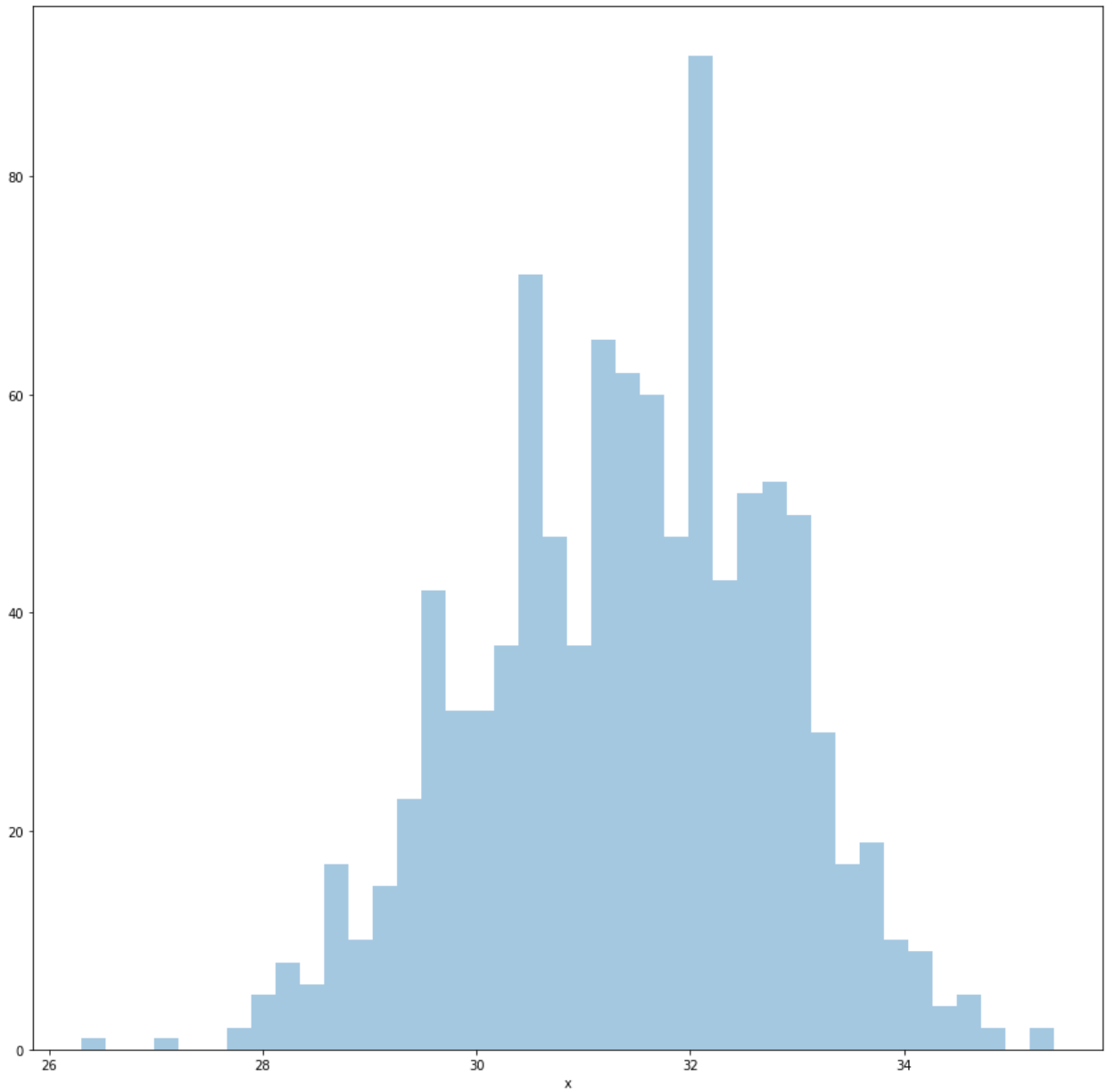
d) (5 points) Use h = 0.25, minimum = a and maximum = b. List the coordinates of the density estimator. Paste the histogram drawn using Python or your favorite graphing tools.

Ans**. -density estimates, midpoint values-**

[(0.0, 26.125), (0.003996003996003996, 26.375), (0.0, 26.625), (0.0, 26.875), (0.003996003996003996, 27.125), (0.0, 27.375), (0.007992007992007992, 27.625), (0.015984015984015984, 27.875), (0.023976023976023976, 28.125), (0.03596403596403597, 28.375), (0.03596403596403597, 28.625), (0.07192807192807193, 28.875), (0.059940059940059943, 29.125), (0.14785214785214784, 29.375), (0.11188811188811189, 29.625), (0.1878121878121878, 29.875), (0.14785214785214784, 30.125), (0.2677322677322677, 30.375), (0.1838161838161838, 30.625), (0.22777222777222778, 30.875), (0.17582417582417584, 31.125), (0.33166833166833165, 31.375), (0.23976023976023977, 31.625), (0.32367632367632365, 31.875), (0.22777222777222778, 32.125), (0.2837162837162837, 32.375), (0.21178821178821178, 32.625), (0.22777222777222778, 32.875), (0.10789210789210789, 33.125),

(0.13186813186813187, 33.375), (0.05194805194805195, 33.625), (0.06393606393606394, 33.875), (0.03596403596403597, 34.125), (0.023976023976023976, 34.375), (0.011988011988011988, 34.625), (0.007992007992007992, 34.875), (0.0, 35.125), (0.007992007992007992, 35.375), (0.0, 35.625), (0.0, 35.875)]

Histogram with h = 0.25

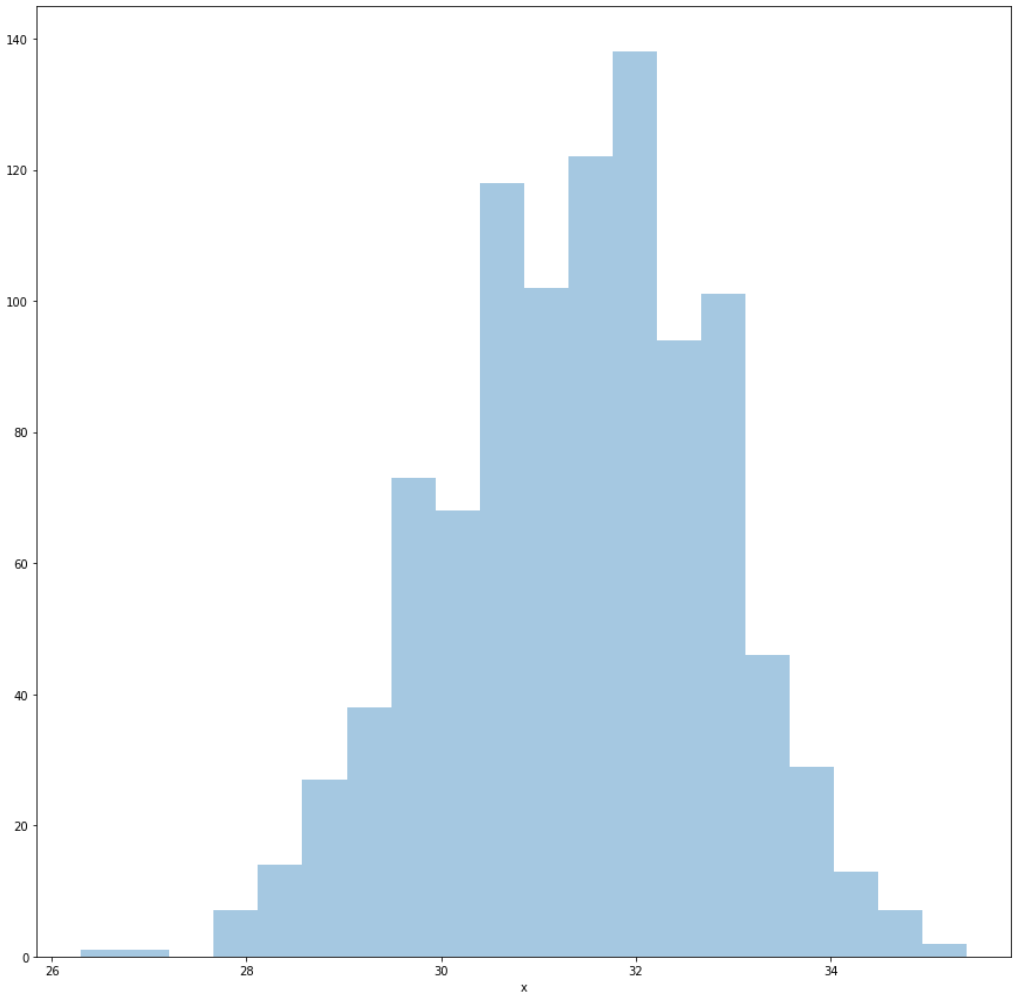


e) (5 points) Use h = 0.5, minimum = a and maximum = b. List the coordinates of the density estimator. Paste the histogram drawn using Python or your favorite graphing tools.

Ans. **--density estimates, midpoint values---**

[(0.001998001998001998, 26.25), (0.0, 26.75), (0.001998001998001998, 27.25), (0.011988011988011988, 27.75), (0.029970029970029972, 28.25), (0.053946053946053944, 28.75), (0.1038961038961039, 29.25), (0.1498501498501498, 29.75), (0.2077922077922078, 30.25), (0.2057942057942058, 30.75), (0.25374625374625376, 31.25), (0.2817182817182817, 31.75), (0.25574425574425574, 32.25), (0.21978021978021978, 32.75), (0.11988011988011989, 33.25), (0.057942057942057944, 33.75), (0.029970029970029972, 34.25), (0.00999000999000999, 34.75), (0.003996003996003996, 35.25), (0.0, 35.75)]

Histogram with h = 0.5

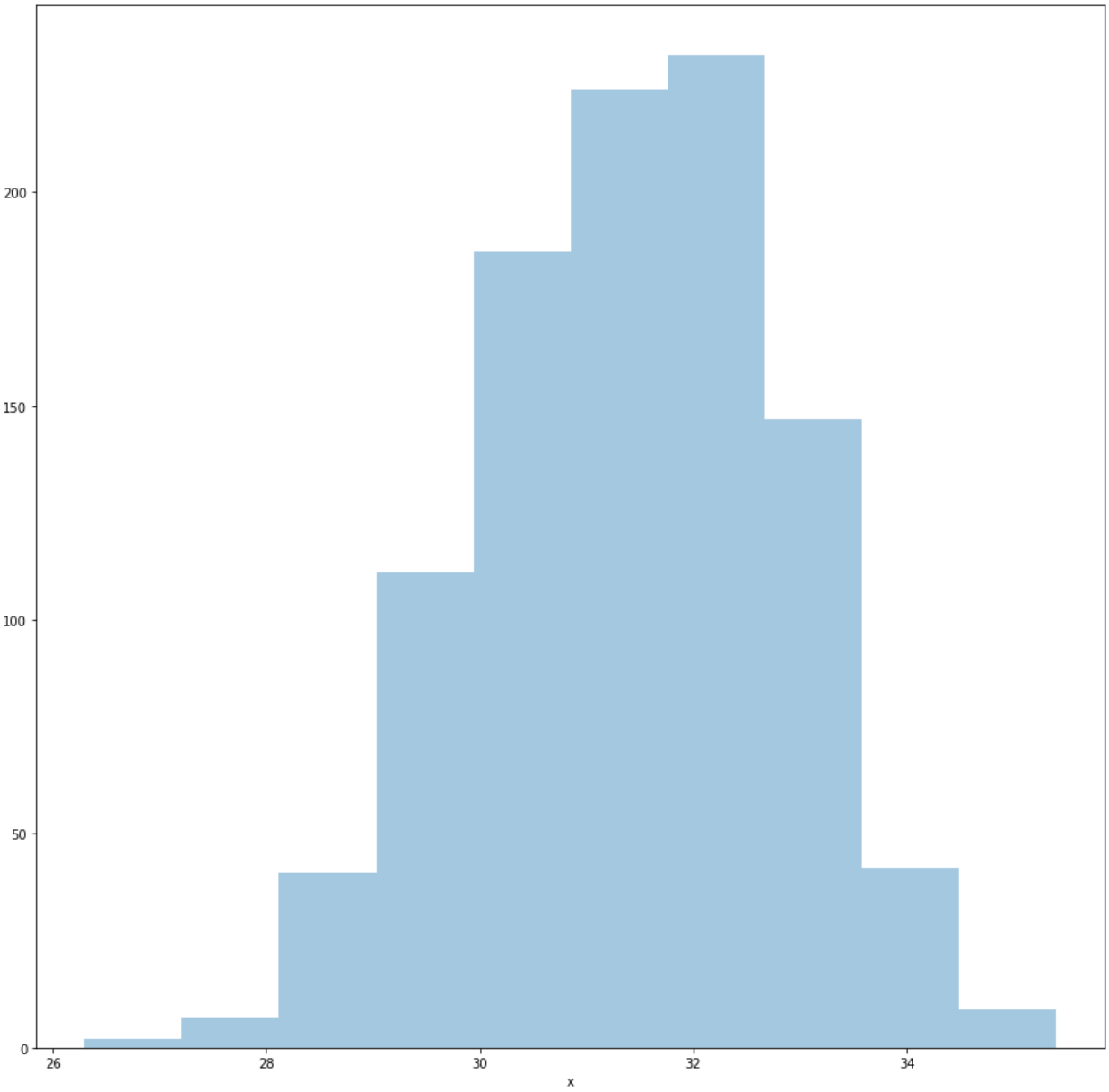


f) (5 points) Use h = 1, minimum = a and maximum = b. List the coordinates of the density estimator. Paste the histogram drawn using Python or your favorite graphing tools.

Ans. **--density estimates,midpoint values----**

[(0.000999000999000999, 26.5), (0.006993006993006993, 27.5), (0.04195804195804196, 28.5), (0.12687312687312688, 29.5), (0.20679320679320679, 30.5), (0.2677322677322677, 31.5), (0.23776223776223776, 32.5), (0.08891108891108891, 33.5), (0.01998001998001998, 34.5), (0.001998001998001998, 35.5)]

Histogram with h=1

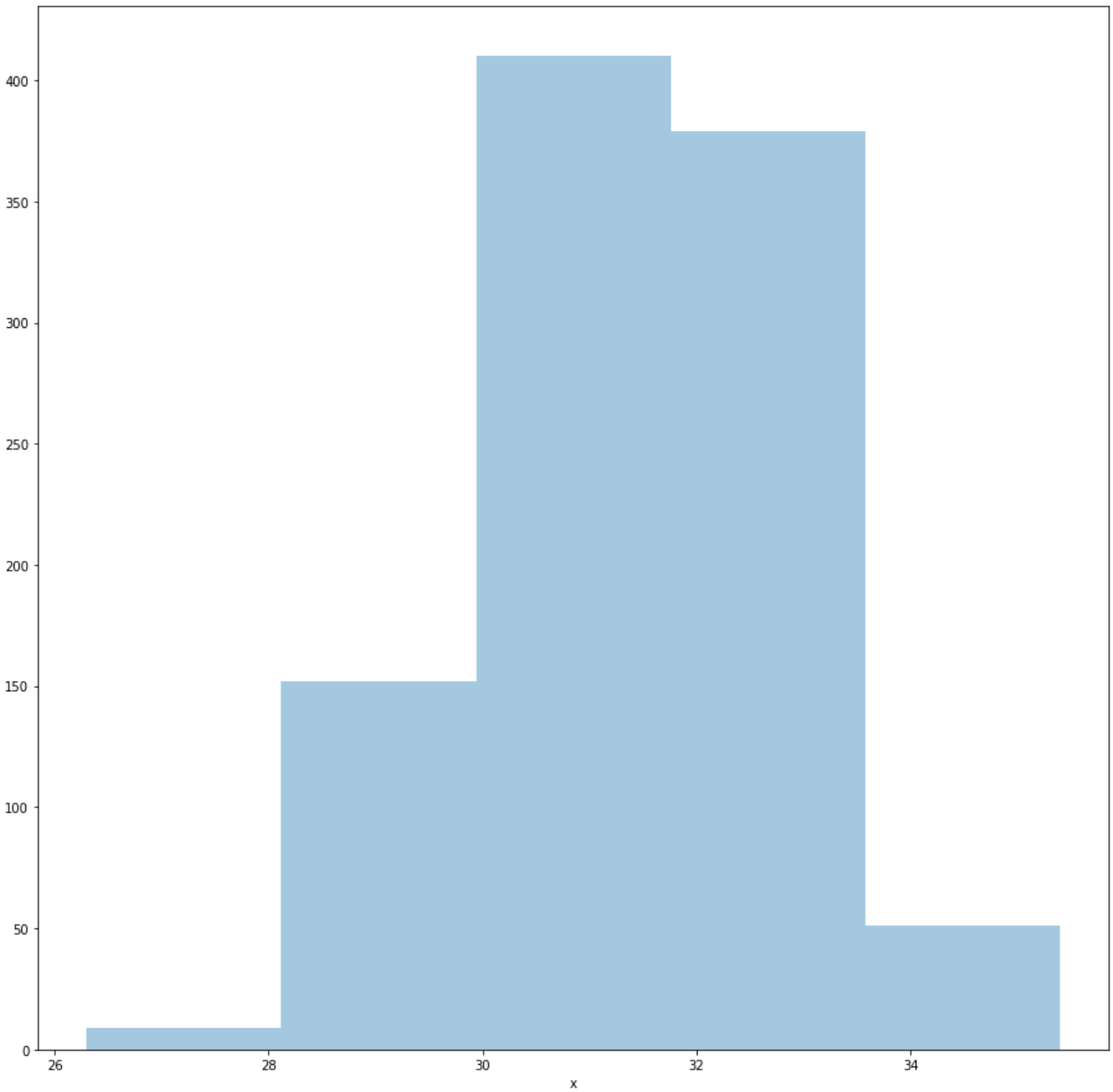


g) (5 points) Use h = 2, minimum = a and maximum = b. List the coordinates of the density estimator. Paste the histogram drawn using Python or your favorite graphing tools.

Ans**. --density estimates,midpoint values--**

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

Histogram with h=2



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. **The histogram when h = 0.25, the bin width is very small. Due to this, we are not able to find the pattern or trend in the data as it is not following the frequency distribution of data. In the same way, the histogram when h = 2, the bin width is large. Here also the same story, we cannot find the trend in the data as more data is exposed. Whereas, in the histogram with h = 1, the data follows normal distribution of the field x.**

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. **Five Number Summary**

count 1001.000000
mean 31.414585
std 1.397672
min 26.300000
25% 30.400000
50% 31.500000
75% 32.400000
max 35.400000
Name: x, dtype: float64 IQR Whiskers

{'Lower Whisker': 27.4, 'Upper 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?

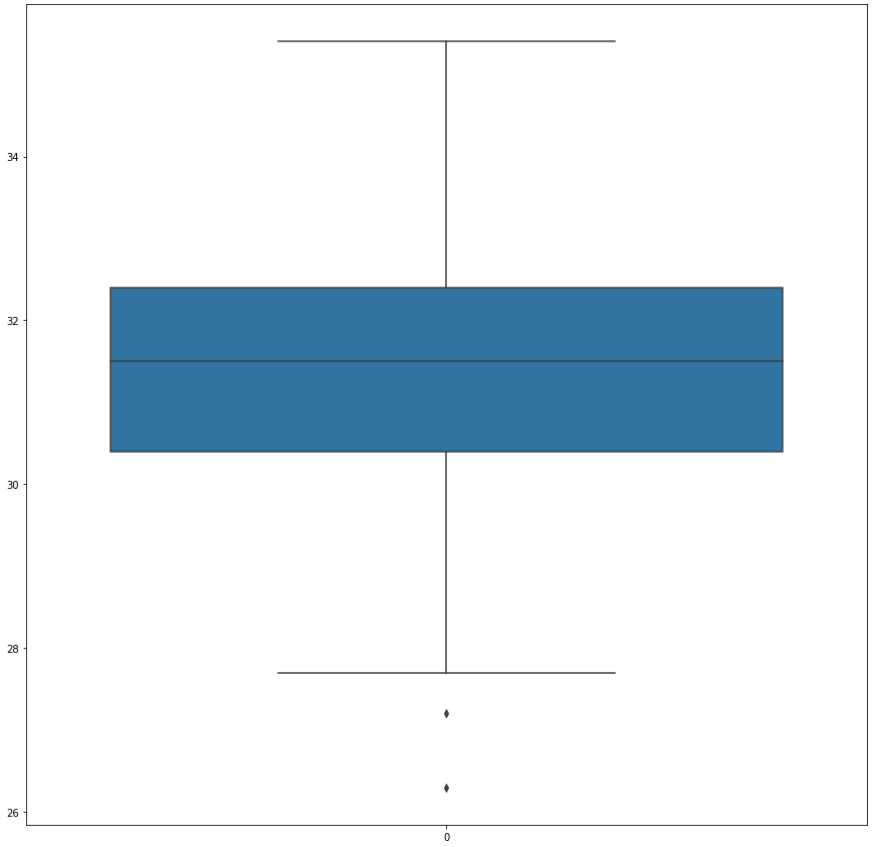
Ans. **count 315.000000**
mean 30.004127
std 0.973935
min 26.300000
25% 29.400000
50% 30.000000
75% 30.600000
max 32.200000
dtype: float64,

count 686.000000
mean 32.062245
std 1.040236
min 29.100000
25% 31.400000
50% 32.100000
75% 32.700000
max 35.400000
dtype: float64

Group 0 Whiskers
{'Lower Whisker': 26.4, 'Upper Whisker': 33.6}
group 1 Whiskers
{'Lower Whisker': 28.4, 'Upper Whisker': 35.7}

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?

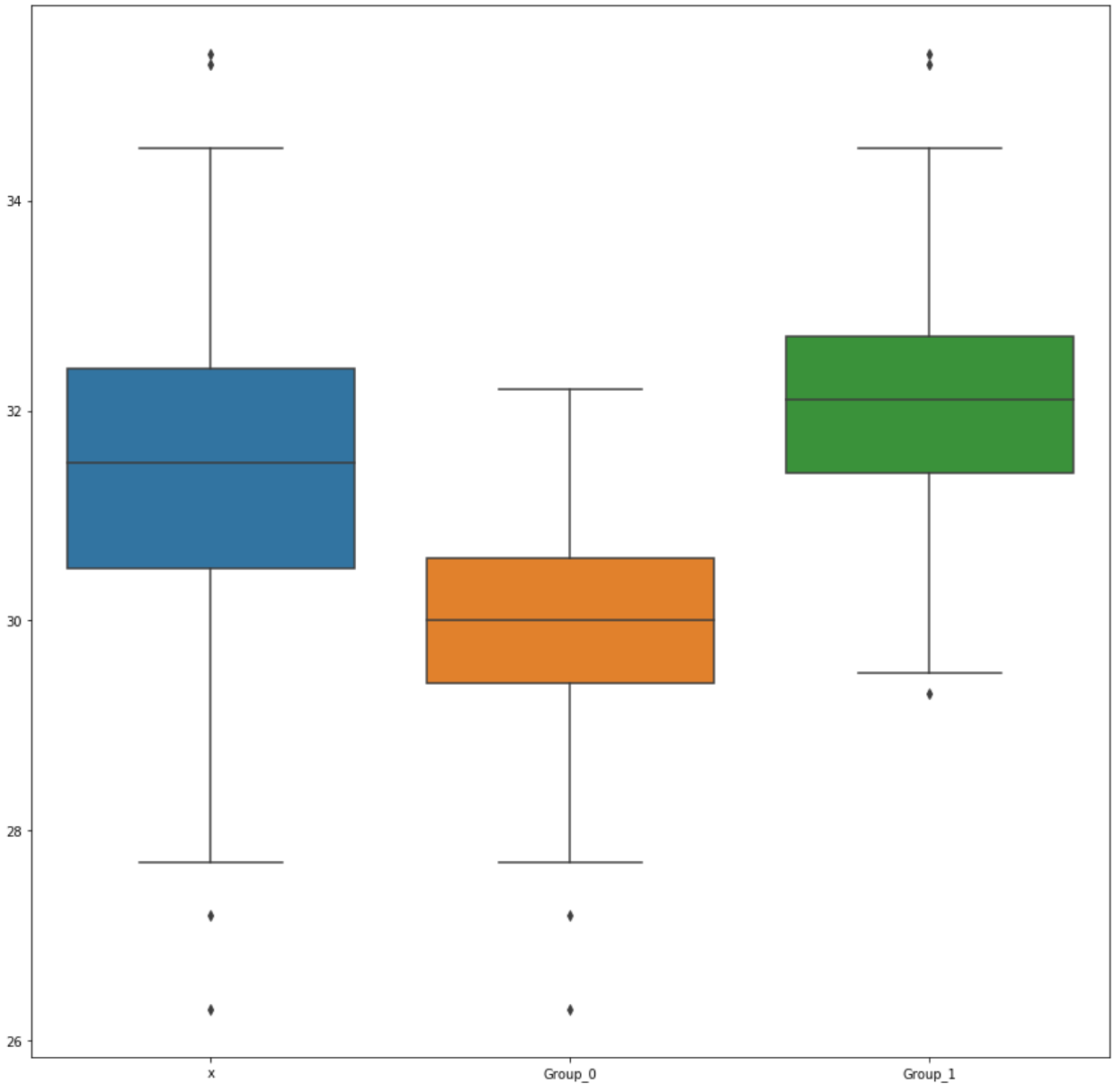
Ans.



Yes, the plot the shows the Lower Whisker 27.4 and the Upper Whisker 35.4 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 of Data x: [27.2, 26.3]
Outliers of Group 0(Zero): [27.2, 26.3]
Outliers of Group 1(One): []

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

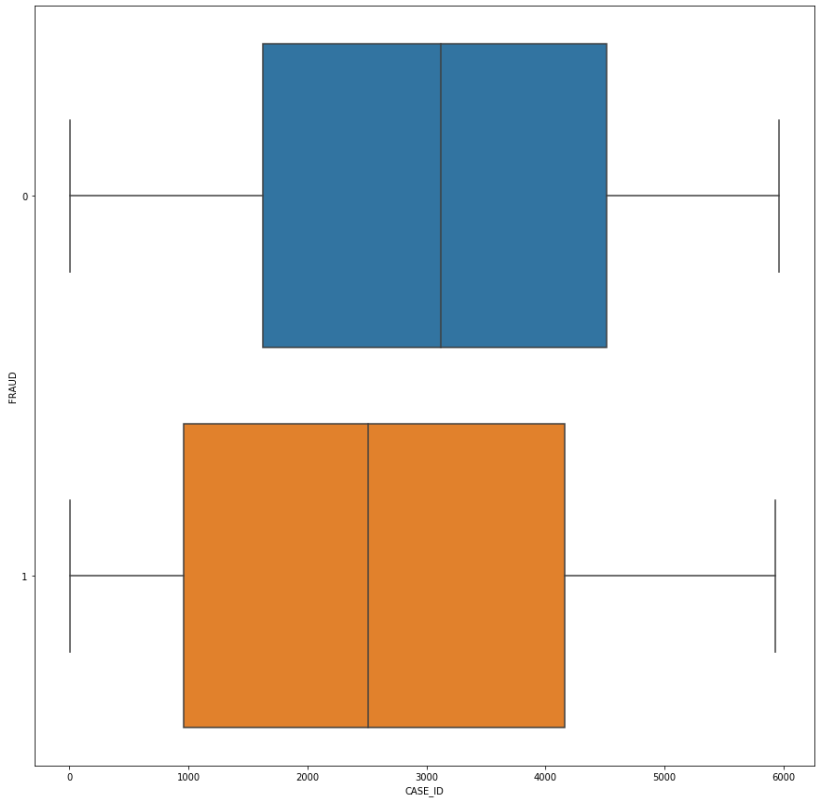
You are asked to use the Nearest Neighbors algorithm to predict the likelihood of fraud.

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

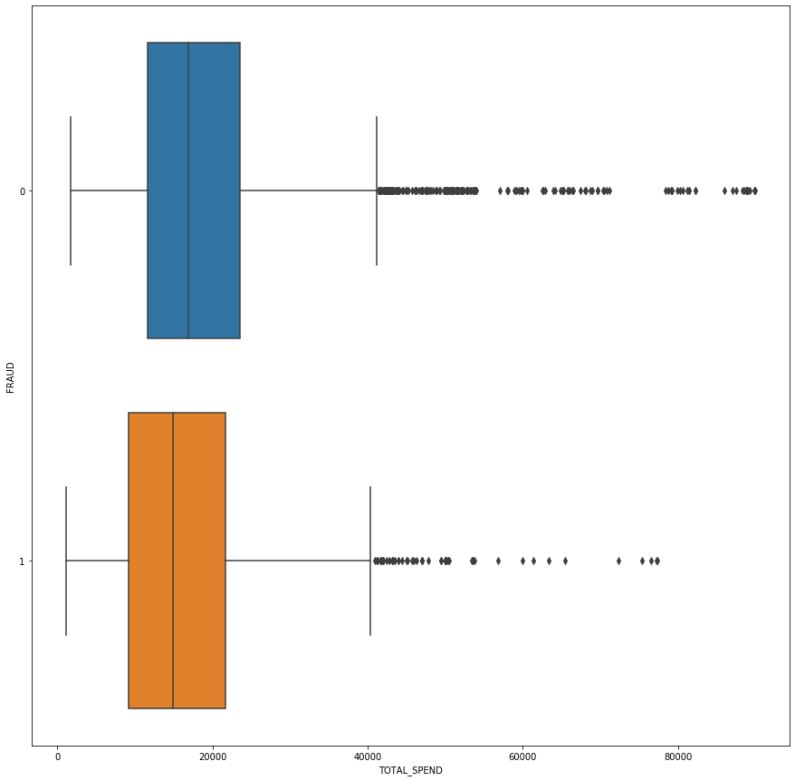
Ans. **Percentage of Fraud investigations 19.9497%**

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.

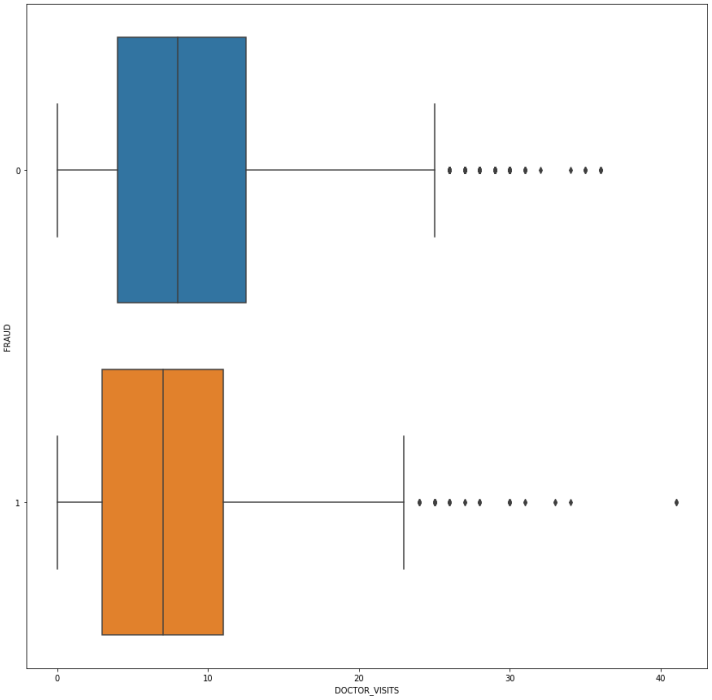
Ans. **CASE_ID**



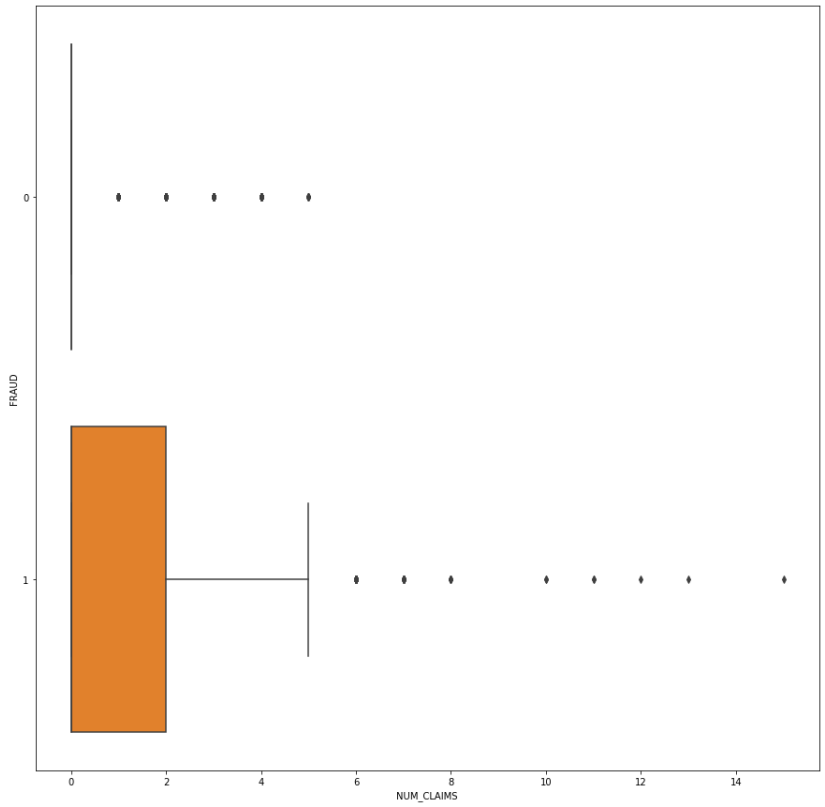
TOTAL_SPEND



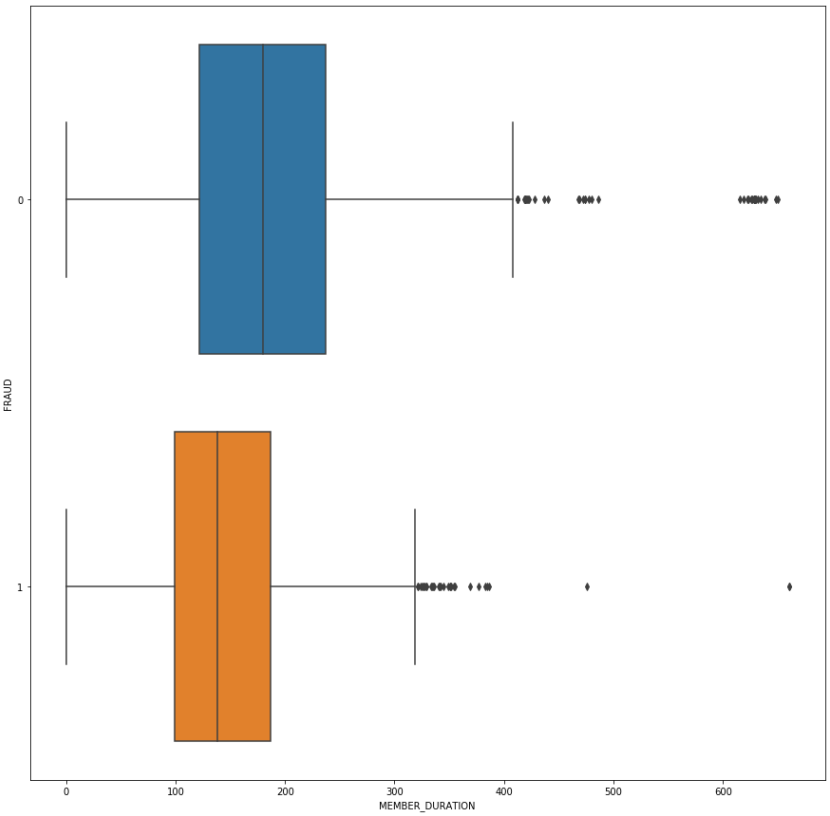
DOCTOR_VISITS



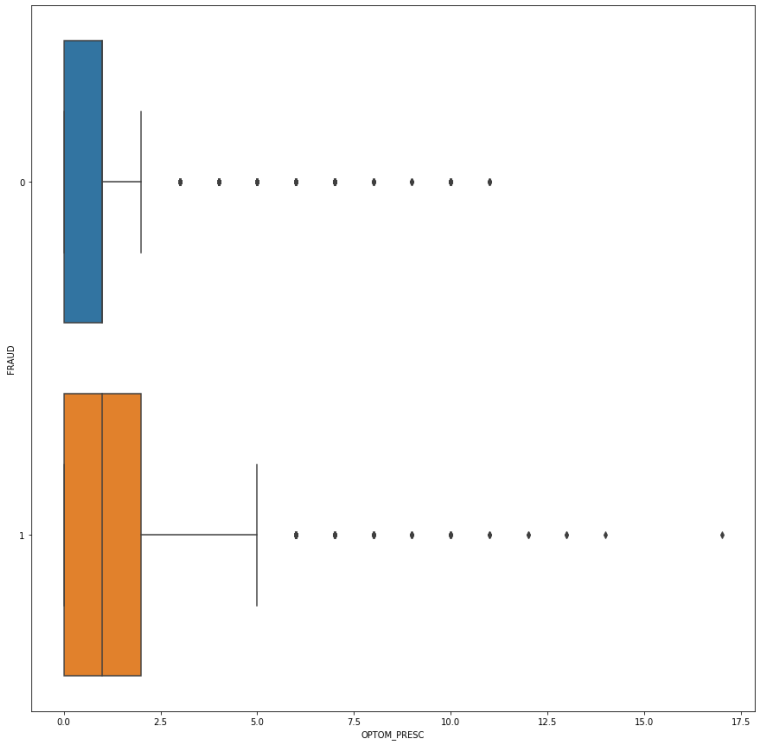
NUM_CLAIMS



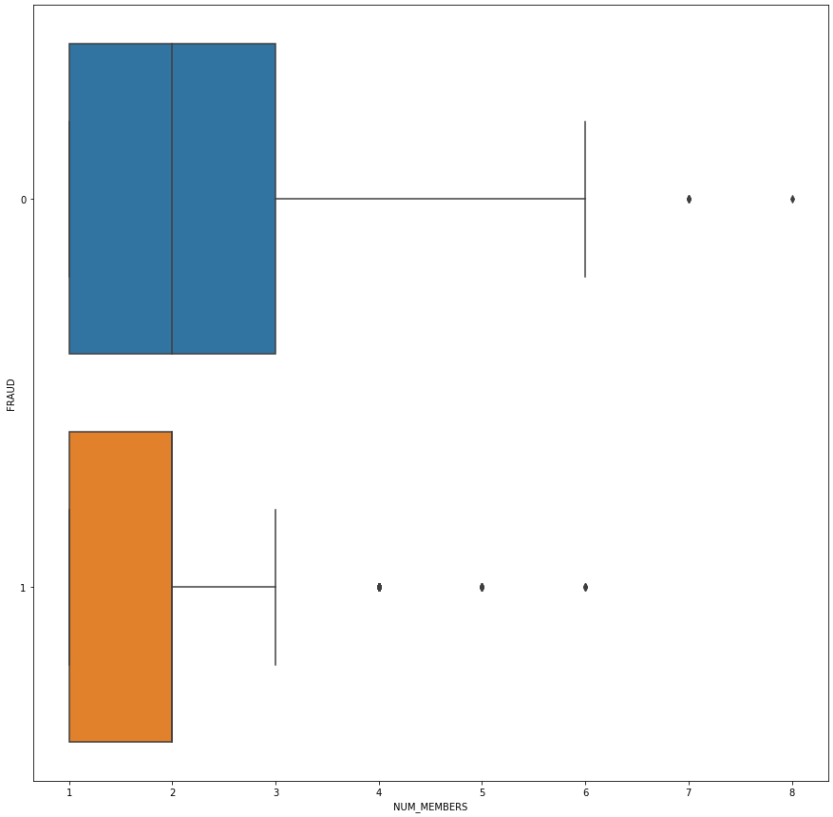
MEMBER_DURATION



OPTOM_PRESC



NUM_MEMBERS



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. **Number of Dimensions used: 6**

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

Ans. ***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]]

Variables are orthonormal:

[[1.00000000e+00 -1.11022302e-16 9.67108338e-17 -7.63278329e-17
1.99493200e-17 -7.91467586e-18]
[-1.11022302e-16 1.00000000e+00 1.83447008e-16 2.25514052e-17
-1.38777878e-17 -3.03576608e-18]
[9.67108338e-17 1.83447008e-16 1.00000000e+00 -6.67868538e-17
-7.91467586e-18 2.55465137e-17]
[-7.63278329e-17 2.25514052e-17 -6.67868538e-17 1.00000000e+00
-9.10729825e-17 1.63660318e-16]
[1.99493200e-17 -1.38777878e-17 -7.91467586e-18 -9.10729825e-17
1.00000000e+00 3.25748543e-16]
[-7.91467586e-18 -3.03576608e-18 2.55465137e-17 1.63660318e-16
3.25748543e-16 1.00000000e+00]]

d) (10 points) Use the NearestNeighbors module to execute the Nearest Neighbors algorithm using exactly five 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

Ans. **Score: 0.8414429530201343**

ii. (5 points) Explain the meaning of the score function return value.

Ans. **The Score Function gives the mean accuracy of the data. It tells how well the predictions match the original data.**

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*.

Ans. **Neighbors of the data: [[588 2897 1199 1246 886]]**

Neighbors input Variables:

	TOTAL_SPEND	DOCTOR_VISITS	...	OPTOM_PRESC	NUM_MEMBERS
588	**7500**	**15**	**...**	**2**	**2**
2897	**16000**	**18**	**...**	**3**	**2**
1199	**10000**	**16**	**...**	**2**	**1**
1246	**10200**	**13**	**...**	**2**	**3**
886	**8900**	**22**	**...**	**1**	**2**

Predicted Target Class: [1]

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?

Ans. **Predicting Probabilites on Traininig data: [[1. 0.]**

[1. 0.]

[1. 0.]

...

[0.8 0.2]

[0.8 0.2]

[0.8 0.2]]

Predicted Probabilites: [[0. 1.]]

The percentage of fraud in question a id 20%. The predicted Probability is 1. This is greater than 20%. Hence, the observation is not misclassified.