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?

The recommended bin-width for the histogram of x is 0.3998667554864774

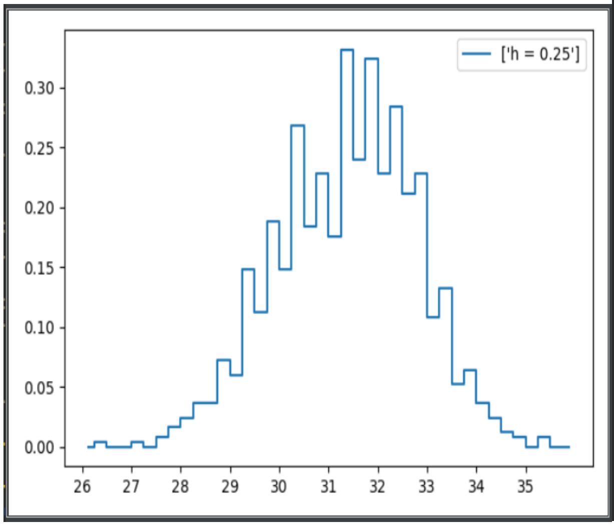
b) (5 points) What are the minimum and the maximum values of the field x?

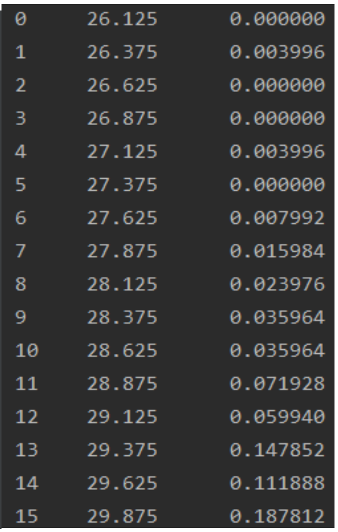
The minimum value is : 26.3
The maximum value is : 35.4

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?

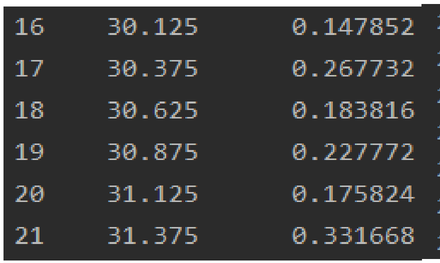
a = 26
b = 36

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.

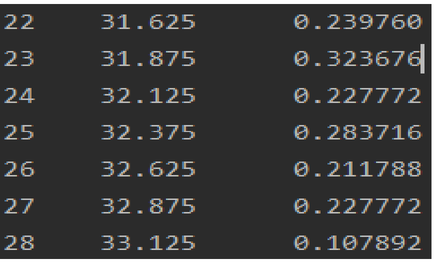




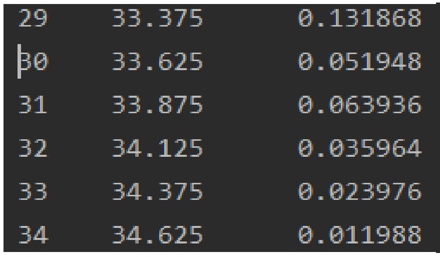
```
0     26.125     0.000000
1     26.375     0.003996
2     26.625     0.000000
3     26.875     0.000000
4     27.125     0.003996
5     27.375     0.000000
6     27.625     0.007992
7     27.875     0.015984
8     28.125     0.023976
9     28.375     0.035964
10    28.625     0.035964
11    28.875     0.071928
12    29.125     0.059940
13    29.375     0.147852
14    29.625     0.111888
15    29.875     0.187812
```



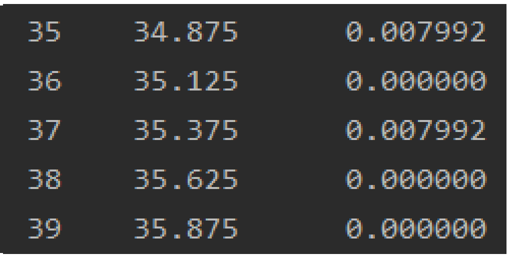
```
16      30.125          0.147852
17      30.375          0.267732
18      30.625          0.183816
19      30.875          0.227772
20      31.125          0.175824
21      31.375          0.331668
```



```
22      31.625          0.239760
23      31.875          0.323676
24      32.125          0.227772
25      32.375          0.283716
26      32.625          0.211788
27      32.875          0.227772
28      33.125          0.107892
```

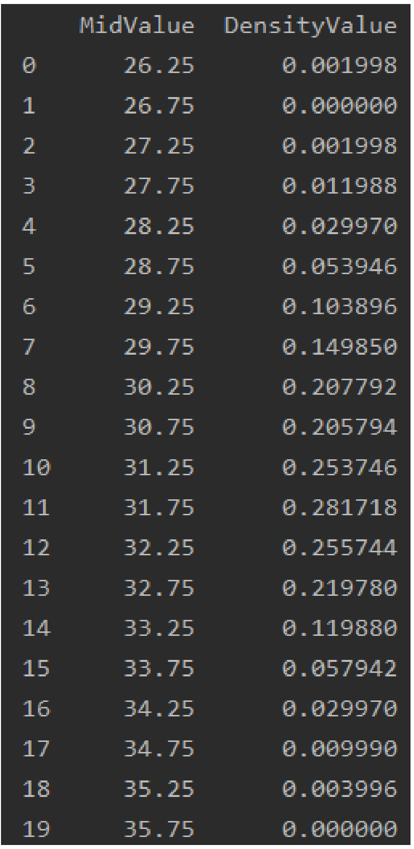


```
29      33.375          0.131868
30      33.625          0.051948
31      33.875          0.063936
32      34.125          0.035964
33      34.375          0.023976
34      34.625          0.011988
```

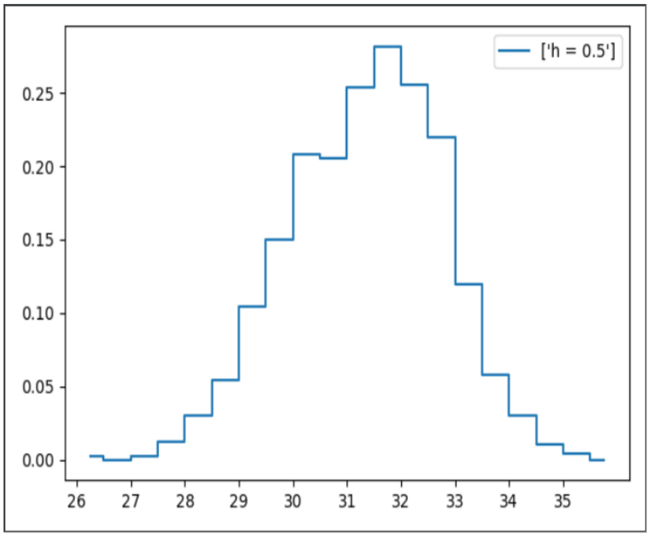


```
35      34.875          0.007992
36      35.125          0.000000
37      35.375          0.007992
38      35.625          0.000000
39      35.875          0.000000
```

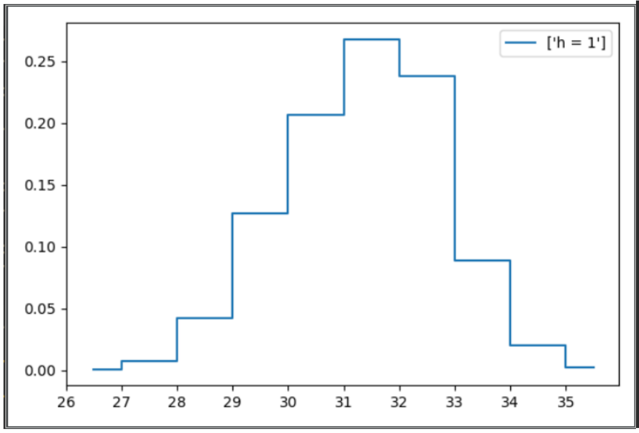
e) (5 points) Use h = 0.5, minimum = a and maximum = b. List the coordinates of the density estimator. Paste the histogram drawn using Python or your favorite graphing tools.

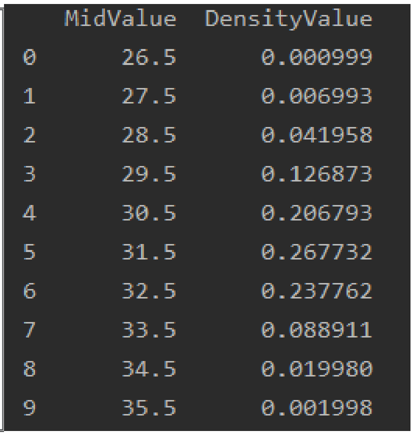


```
     MidValue  DensityValue
0       26.25      0.001998
1       26.75      0.000000
2       27.25      0.001998
3       27.75      0.011988
4       28.25      0.029970
5       28.75      0.053946
6       29.25      0.103896
7       29.75      0.149850
8       30.25      0.207792
9       30.75      0.205794
10      31.25      0.253746
11      31.75      0.281718
12      32.25      0.255744
13      32.75      0.219780
14      33.25      0.119880
15      33.75      0.057942
16      34.25      0.029970
17      34.75      0.009990
18      35.25      0.003996
19      35.75      0.000000
```



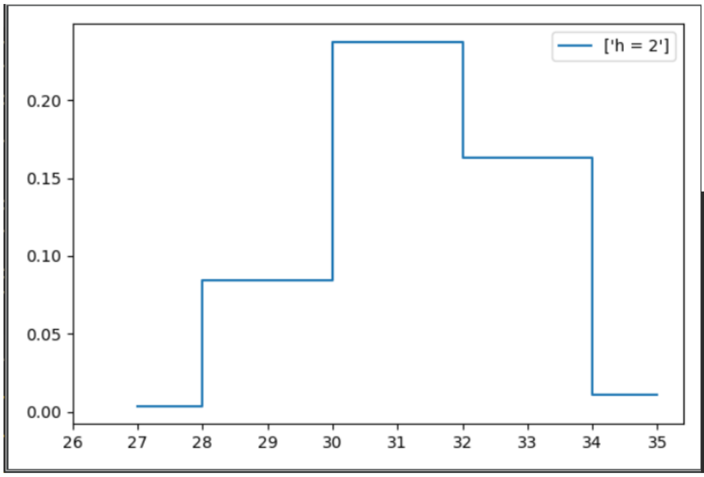
f) (5 points) Use h = 1, minimum = a and maximum = b. List the coordinates of the density estimator. Paste the histogram drawn using Python or your favorite graphing tools.

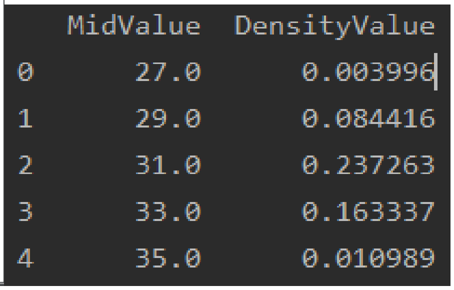




```
   MidValue  DensityValue
0      26.5      0.000999
1      27.5      0.006993
2      28.5      0.041958
3      29.5      0.126873
4      30.5      0.206793
5      31.5      0.267732
6      32.5      0.237762
7      33.5      0.088911
8      34.5      0.019980
9      35.5      0.001998
```

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.





```
   MidValue  DensityValue
0      27.0      0.003996
1      29.0      0.084416
2      31.0      0.237263
3      33.0      0.163337
4      35.0      0.010989
```

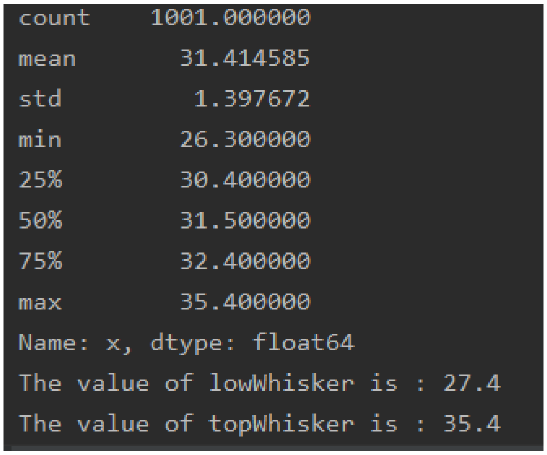
h) (5 points) Among the four histograms, which one, in your honest opinions, can best provide your insights into the shape and the spread of the distribution of the field x? Please state your arguments.

I will choose h = 0.5. Because this bin-width is closest to the Izenman's recommended value, which is 0.3999. Meanwhile, the h = 0.5 of histograms is more like the normal distribution.

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?



```
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
The value of lowWhisker is : 27.4
The value of topWhisker is : 35.4
```

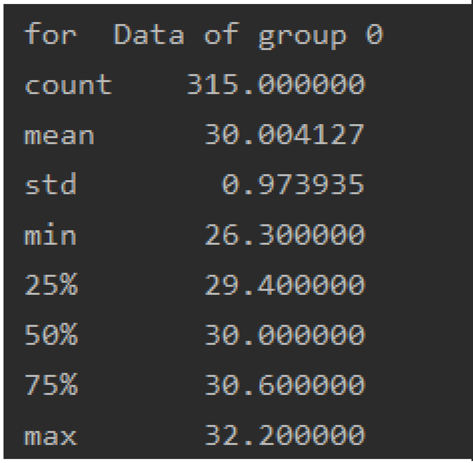
Minimum	Q1	Median	Q3	Maximum
26.3	30.4	31.5	32.4	35.4

IQR = Q3-Q1 =2, 1.5*IQR = 3

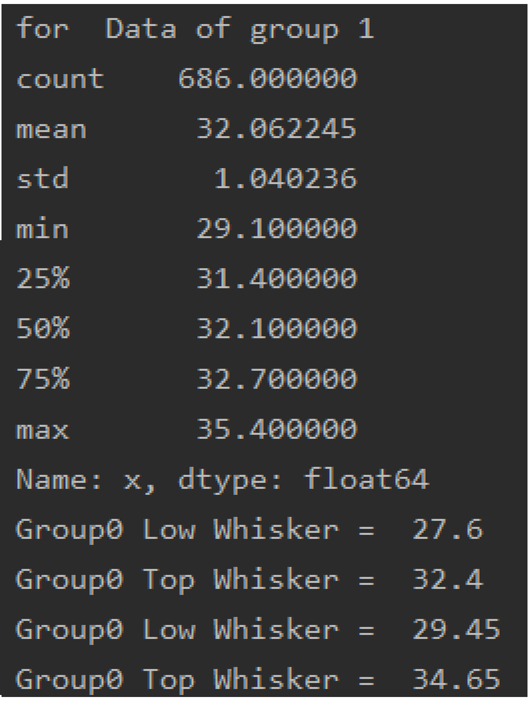
Low whisker : max {minimum, Q1-1.5*IQR} = 27.4

Upper whisker : {Q3+1.5*IQR, maximum} = 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?



```
for  Data of group 0
count    315.000000
mean      30.004127
std        0.973935
min       26.300000
25%       29.400000
50%       30.000000
75%       30.600000
max       32.200000
```



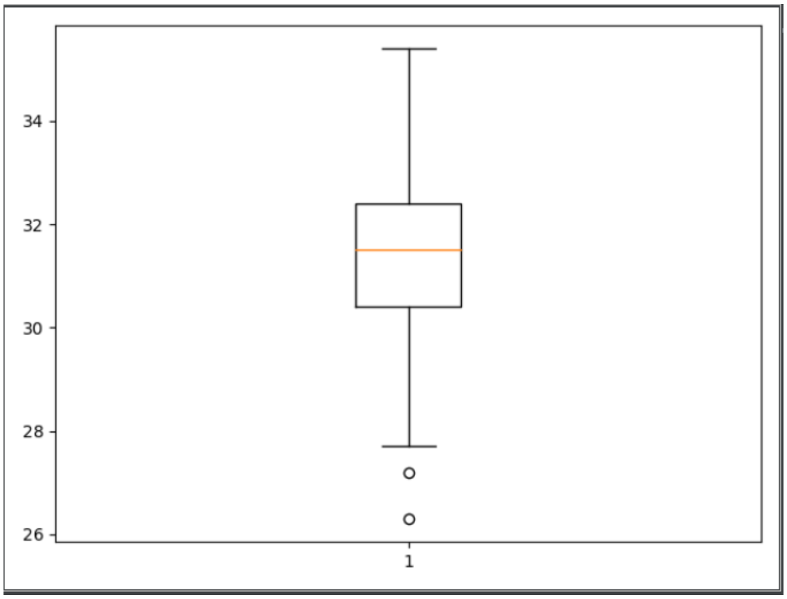
```
for  Data of group 1
count    686.000000
mean      32.062245
std        1.040236
min       29.100000
25%       31.400000
50%       32.100000
75%       32.700000
max       35.400000
Name: x, dtype: float64
Group0 Low Whisker =  27.6
Group0 Top Whisker =  32.4
Group0 Low Whisker =  29.45
Group0 Top Whisker =  34.65
```

Group	Minmum	Q1	Median	Q3	Maximum
0	26.3	29.4	30	30.6	32.2
1	29.1	31.4	32.1	32.7	35.4

For group0 : IQR = Q3-Q1 =1.2, 1.5*IQR = 1.8 Low whisker is max {minimum, Q1-1.5*IQR} = 27.6, and upper whisker is min {Q3+1.5*IQR, maximum} = 32.2

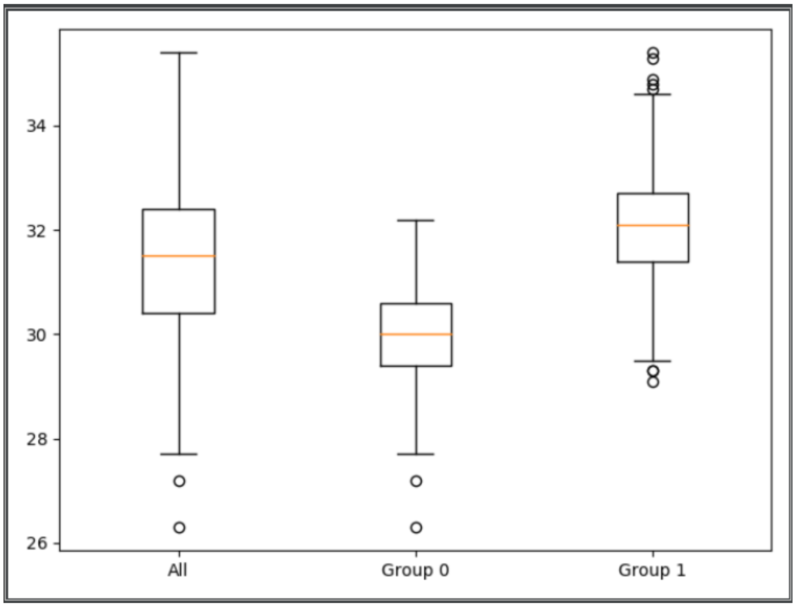
For group1 : IQR = Q3-Q1 =1.3, 1.5*IQR = 1.95 Lower whisker is max {minimum, Q1-1.5*IQR} = 29.45, and upper whisker is min {Q3+1.5*IQR, maximum} = 34.65

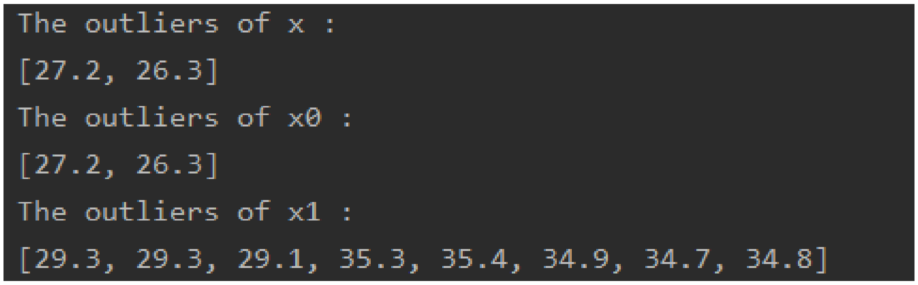
c) (5 points) Draw a boxplot of x (without the group) using the Python boxplot function. Can you tell if the Python's boxplot has displayed the 1.5 IQR whiskers correctly?



Compare with the result (Low Whisker = 27.4 and Top Whisker = 35.4) from 2(a), we can obviously see that this boxplot displayed the 1.5 IQR whiskers

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.





```
The outliers of x :
[27.2, 26.3]
The outliers of x0 :
[27.2, 26.3]
The outliers of x1 :
[29.3, 29.3, 29.1, 35.3, 35.4, 34.9, 34.7, 34.8]
```

For all x, there are several outliers outside the low whisker. These two values are come from group 0.

For group 0, there are several outliers outside the low whisker.

For group 1, there are several outliers outside both the low whisker and top whisker

Group	IQR	Left Whisker	Right Whisker
all	1.2	27.6	32.2
0	1.3	29.45	34.65
1	2	27.4	35.4

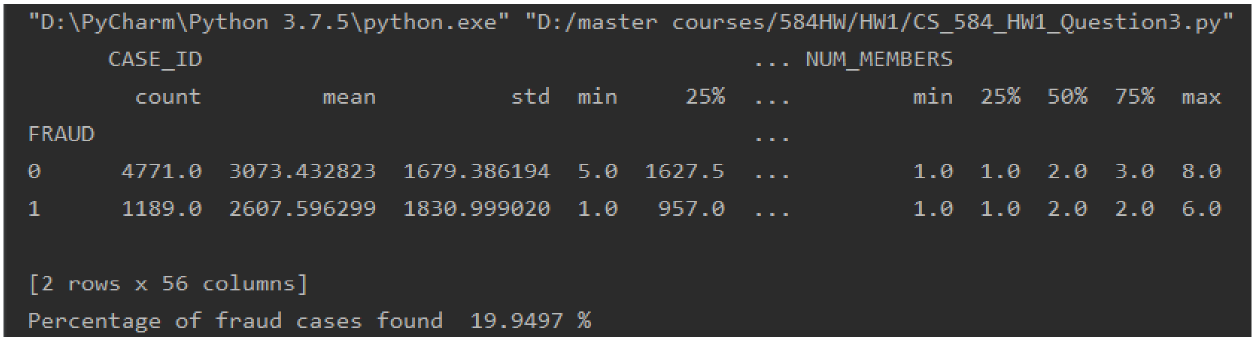
Question 3 (40 points)

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

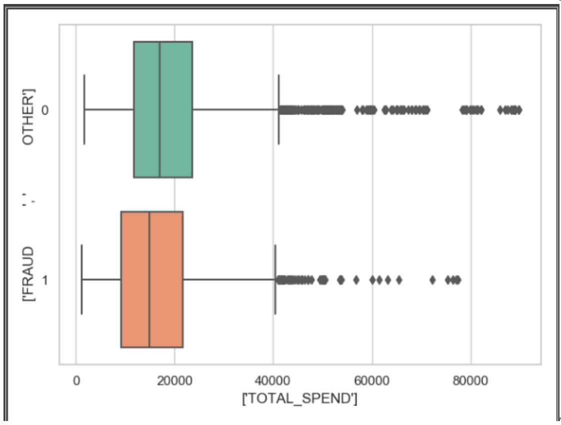
You are asked to use the Nearest Neighbors algorithm to predict the likelihood of fraud.

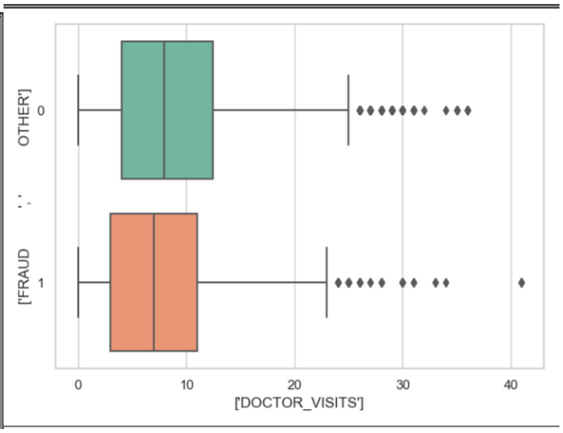
a) (5 points) What percent of investigations are found to be fraudulent? Please give your answer up to 4 decimal places.

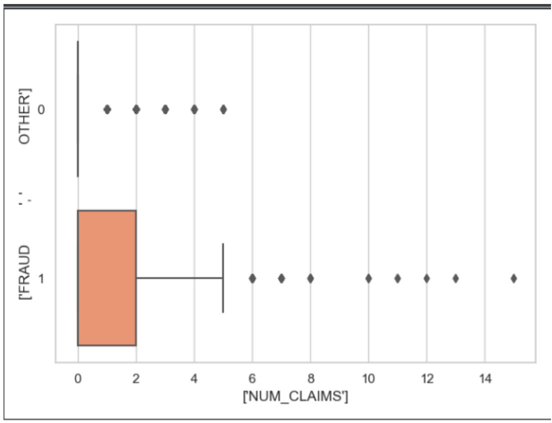


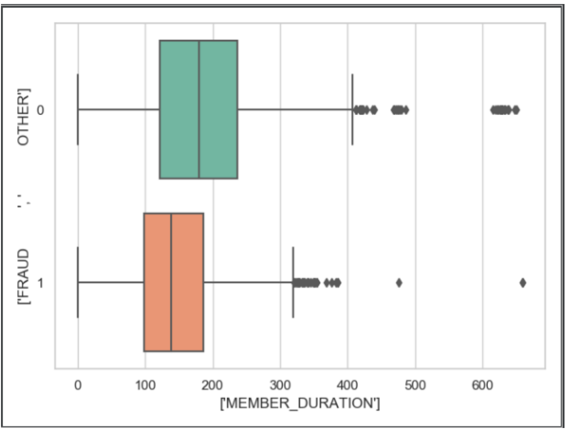
```
"D:\PyCharm\Python 3.7.5\python.exe" "D:/master courses/584HW/HW1/CS_584_HW1_Question3.py"
      CASE_ID                                          ... NUM_MEMBERS
        count          mean          std  min     25%  ...         min  25%  50%  75%  max
FRAUD                                                  ...
0      4771.0  3073.432823  1679.386194  5.0  1627.5  ...         1.0  1.0  2.0  3.0  8.0
1      1189.0  2607.596299  1830.999020  1.0   957.0  ...         1.0  1.0  2.0  2.0  6.0

[2 rows x 56 columns]
Percentage of fraud cases found  19.9497 %
```

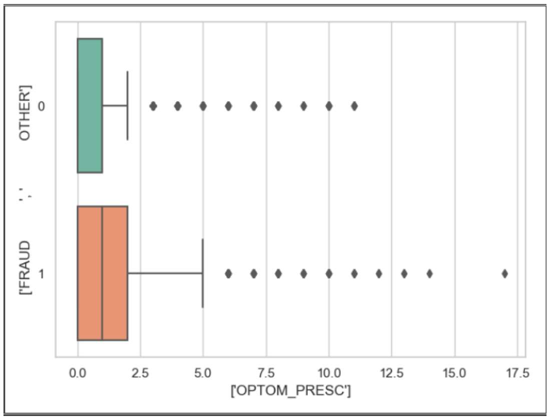
b) (5 points) Use the BOXPLOT function to produce horizontal box-plots. For each interval variable, one box-plot for the fraudulent observations, and another box-plot for the non-fraudulent observations. These two box-plots must appear in the same graph for each interval variable.

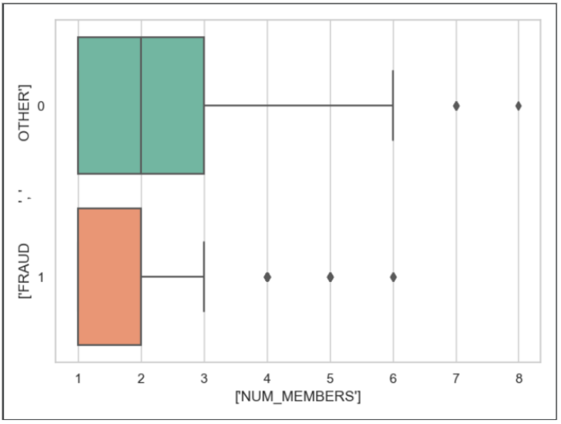






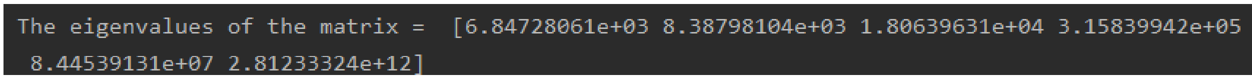






c) (10 points) Orthonormalize interval variables and use the resulting variables for the nearest neighbor analysis. Use only the dimensions whose corresponding eigenvalues are greater than one.

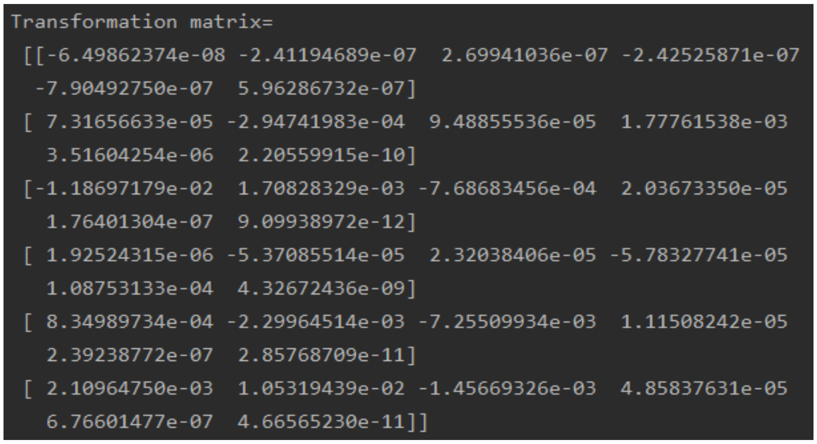
i. (5 points) How many dimensions are used?



```
The eigenvalues of the matrix =  [6.84728061e+03 8.38798104e+03 1.80639631e+04 3.15839942e+05
 8.44539131e+07 2.81233324e+12]
```

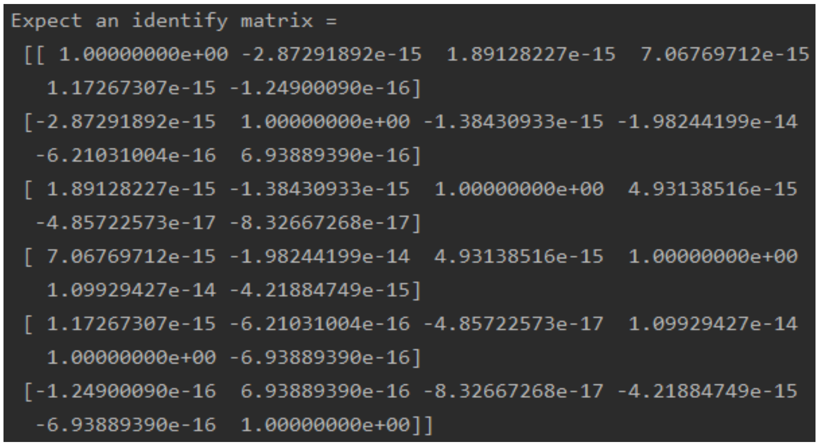
They are all bigger than 1. So the six dimensions are used

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]]
```

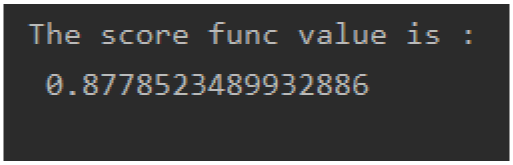
By multiply the input matrix with the transformation matrix (shown above), we can have a transformed matrix. Then we multiply the transposed transformed matrix and it self, we got a 6X6 identity matrix (shown below). Because this is an identity matrix, so the resulting variables are orthonormal.



```
Expect an identify matrix =
 [[ 1.00000000e+00 -2.87291892e-15  1.89128227e-15  7.06769712e-15
   1.17267307e-15 -1.24900090e-16]
 [-2.87291892e-15  1.00000000e+00 -1.38430933e-15 -1.98244199e-14
  -6.21031004e-16  6.93889390e-16]
 [ 1.89128227e-15 -1.38430933e-15  1.00000000e+00  4.93138516e-15
  -4.85722573e-17 -8.32667268e-17]
 [ 7.06769712e-15 -1.98244199e-14  4.93138516e-15  1.00000000e+00
   1.09929427e-14 -4.21884749e-15]
 [ 1.17267307e-15 -6.21031004e-16 -4.85722573e-17  1.09929427e-14
   1.00000000e+00 -6.93889390e-16]
 [-1.24900090e-16  6.93889390e-16 -8.32667268e-17 -4.21884749e-15
  -6.93889390e-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

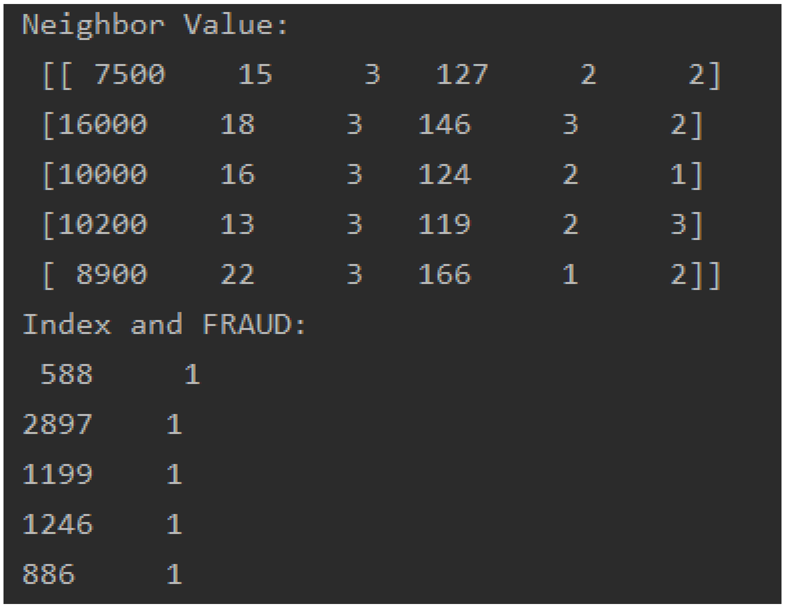


```
The score func value is :
 0.8778523489932886
```

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

This value means about 87.79% observations are correctly. In another word, the misclassification rate is around 12.21%

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*.



```
Neighbor Value:
 [[ 7500    15     3   127     2     2]
 [16000    18     3   146     3     2]
 [10000    16     3   124     2     1]
 [10200    13     3   119     2     3]
 [ 8900    22     3   166     1     2]]
Index and FRAUD:
 588      1
2897     1
1199     1
1246     1
886      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?

The predicted probability of fraudulent is 1 because the FRAUD values are all 1. Therefore, the observation will be classified as fraudulent, they are not misclassified.