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?

Aa) 0.3998667554864774

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

Ab) **Minimum(x)=26.3 and Maximum(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?

Ac) a=26 and 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.

Ad)

m	p(m)
26.05	0
26.15	0
26.25	0.009990001
26.35	0.009990001
26.45	0
26.55	0
26.65	0
26.75	0
26.85	0

26.95	0
27.05	0
27.15	0.009990001
27.25	0.009990001
27.35	0
27.45	0
27.55	0
27.65	0.019980002
27.75	0.019980002

27.85	0.029970003
27.95	0.039960004
28.05	0.019980002
28.15	0.059940006
28.25	0.079920008
28.35	0.049950005
28.45	0.059940006

28.55	0.07992008
28.65	0.08991009
28.75	0.12987013
28.85	0.12987013
28.95	0.0999001
29.05	0.08991009
29.15	0.14985015
29.25	0.25974026
29.35	0.22977023
29.45	0.21978022
29.55	0.32967033
29.65	0.27972028
29.75	0.18981019
29.85	0.30969031
29.95	0.36963037
30.05	0.30969031
30.15	0.36963037

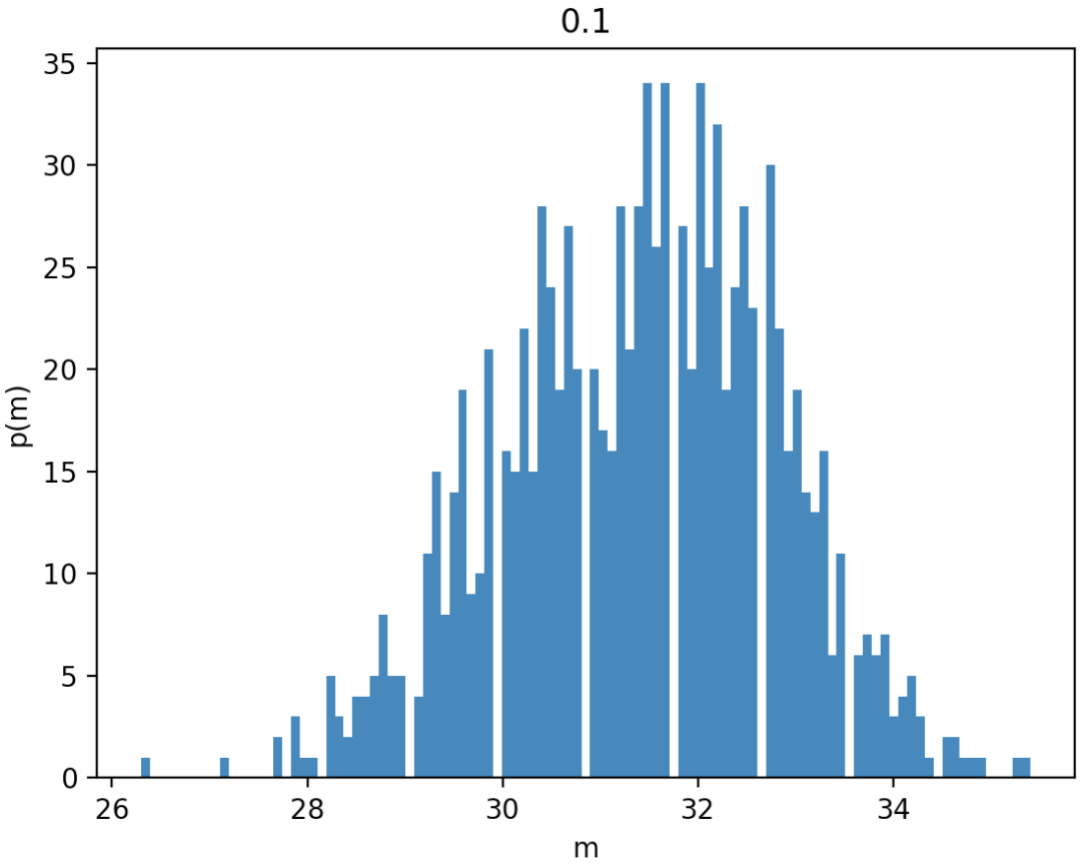
30.25	0.36963037
30.35	0.42957043
30.45	0.51948052
30.55	0.42957043
30.65	0.45954046
30.75	0.46953047
30.85	0.3996004
30.95	0.36963037
31.05	0.32967033
31.15	0.43956044
31.25	0.48951049
31.35	0.48951049
31.45	0.61938062
31.55	0.5994006
31.65	0.5994006
31.75	0.60939061
31.85	0.46953047

31.95	0.53946054
32.05	0.33966034
32.15	0.31968032
32.25	0.50949051
32.35	0.18981019
32.45	0.27972028
32.55	0.27972028
32.65	0.2997003
32.75	0.51948052
32.85	0.21978022
32.95	0.18981019
33.05	0.18981019
33.15	0.12987013
33.25	0.28971029
33.35	0.15984016
33.45	0.10989011
33.55	0.10989011

33.65	0.06993007
33.75	0.12987013
33.85	0.05994006
33.95	0.02997003
34.05	0.02997003
34.15	0.04995005
34.25	0.07992008

34.35	0.02997003
34.45	0.01998002
34.55	0.01998002
34.65	0.00999001
34.75	0.01998002
34.85	0.00999001
34.95	0
35.05	0

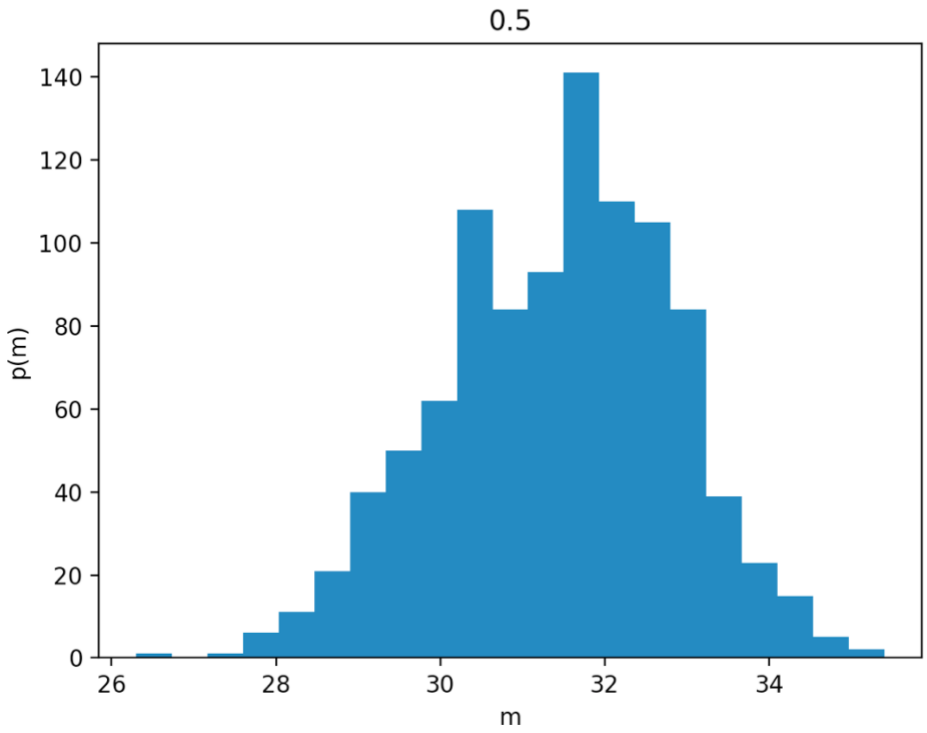
35.15	0
35.25	0.00999001
35.35	0.00999001
35.45	0
35.55	0
35.65	0
35.75	0
35.85	0
35.95	0



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.

Ae)

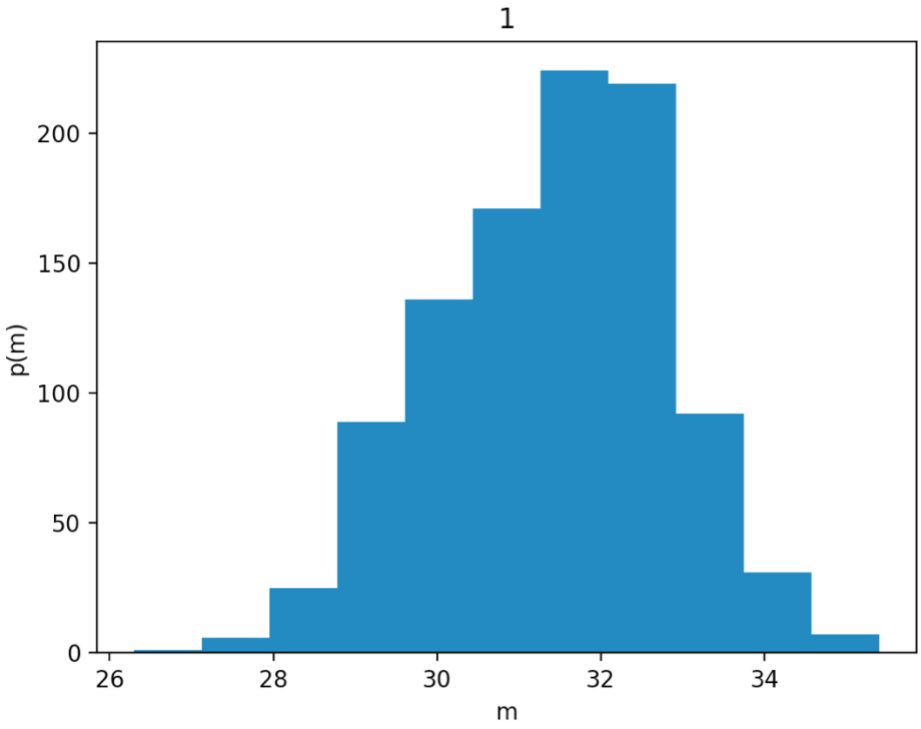
m	p(m)
26.25	0.001998
26.75	0
27.25	0.001998
27.75	0.011988
28.25	0.031968
28.75	0.061938
29.25	0.113886
29.75	0.177821
30.25	0.239760
30.75	0.253746
31.25	0.287712
31.75	0.349650
32.25	0.323676
32.75	0.275724
33.25	0.157842
33.75	0.079920
34.25	0.035964
34.75	0.013986
35.25	0.003996
35.75	0



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.

Af)

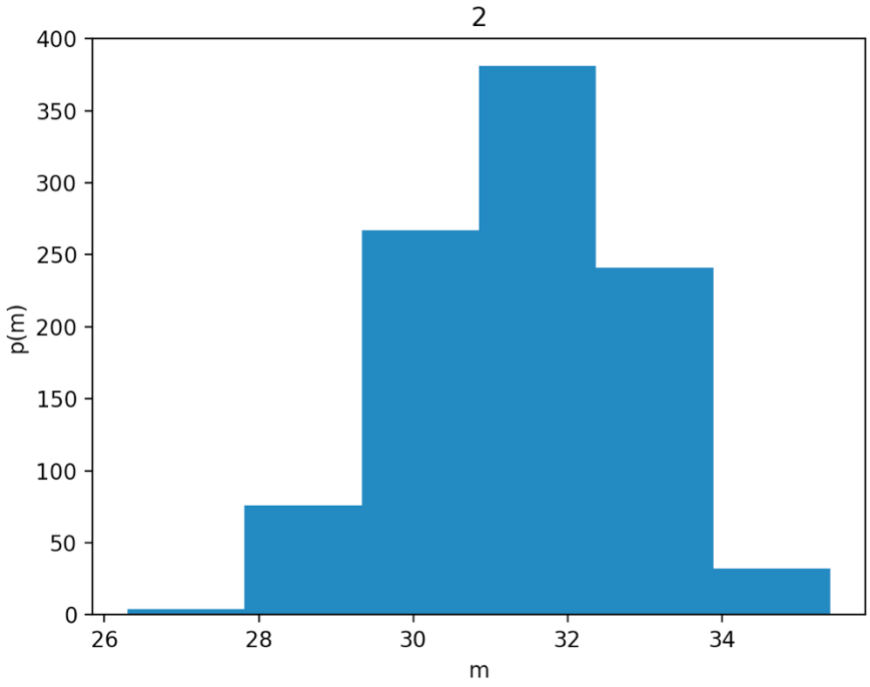
m	p(m)
26.5	**0.000999**
27.5	**0.00699301**
28.5	**0.04295704**
29.5	**0.13186813**
30.5	**0.22277722**
31.5	**0.28471528**
32.5	**0.27172827**
33.5	**0.10789211**
34.5	**0.02297702**
35.5	**0.001998**



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.

Ag)

m	p(m)
27	**0.003996**
29	**0.08491508**
31	**0.24525475**
33	**0.18031968**
35	**0.01248751**



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.

Ah) I feel that the histogram with the bin-width of 0.5 can provide better insight to the data distribution compared to the others. It doesn't have too many information unlike h=0.1, which has too many data points and is difficult to form ideas from the graph, while h=0.5 gives a simpler view, but not too simple like the h=1 and h=2, so that some sort of assessment can be done from the graph.

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?

Aa) median: 31.5

Min: 26.3

Max: 35.4

A: 26

B: 36

q1: 30.4

q2: 32.4

IQR 2.0

W1 27.4

W2 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?

Ab)

N:686

****

Median: 32.1

Min: 29.1

Max: 35.4

A: 29

B: 36

q1: 31.4

q2: 32.7

IQR 1.3000000000000043

W1 29.449999999999992

W2 34.650000000000006

N: 315

****

Median: 30.0

Min: 26.3

Max: 32.2

A: 26

B: 33

q1: 29.4

q2: 30.6

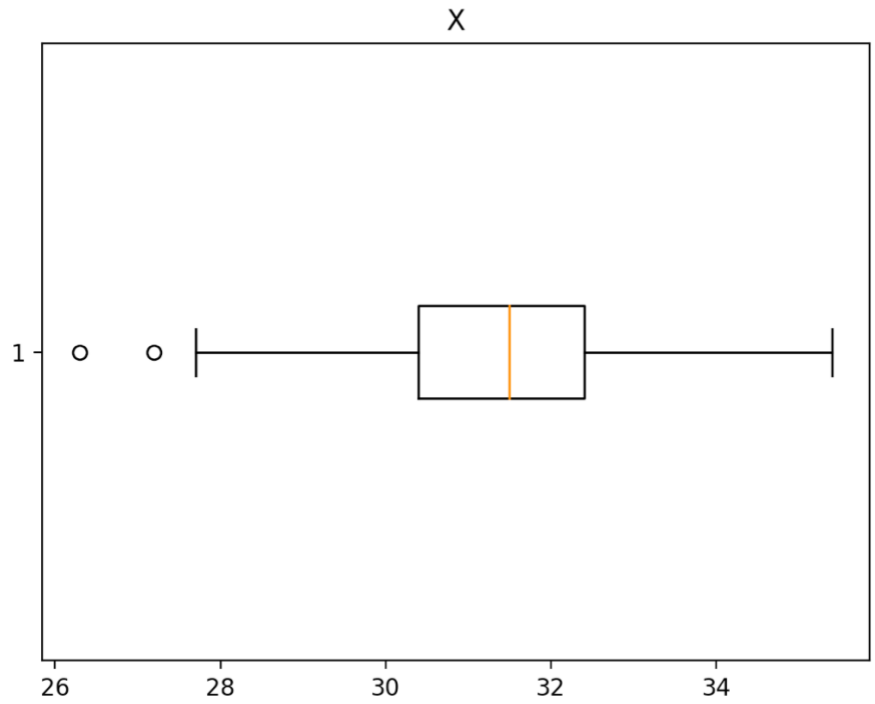
IQR: 1.2000000000000028

W1: 27.599999999999994

W2: 32.400000000000006

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?

Ac)



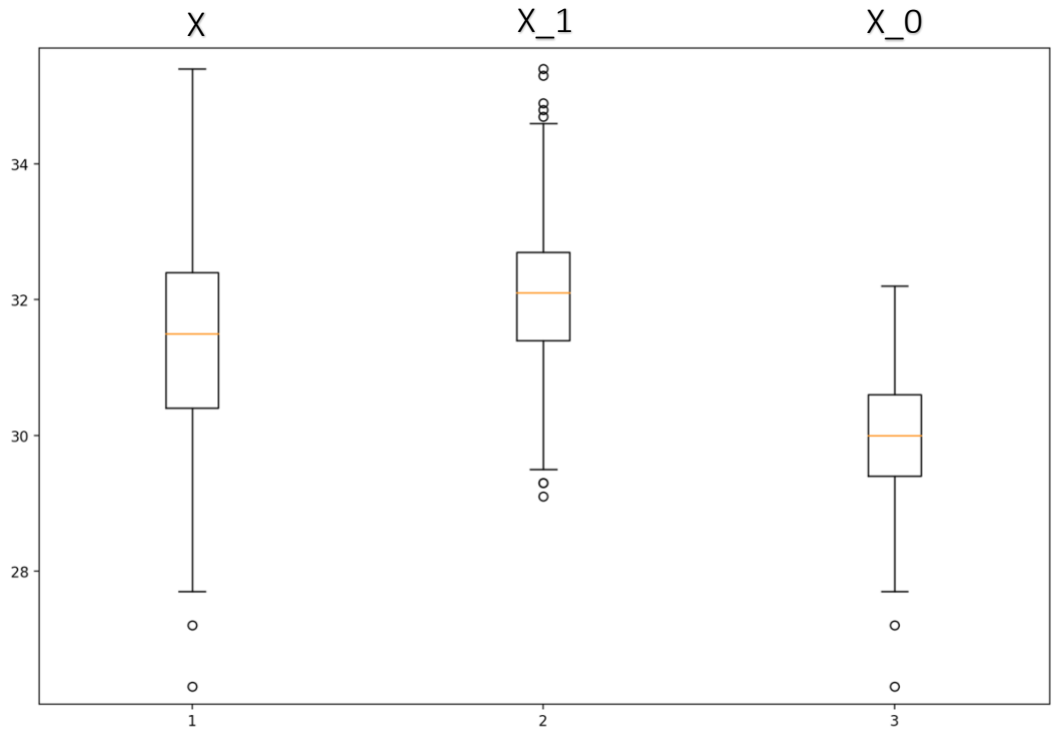
Yes, it can be clearly seen that there two outliers, i.e, 26.3, and 27.2.

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.

Ad) **The outliers of X are: (26.3, 27.2)**

The outliers for X_1 are: (29.1, 29.3, *29.3***) and (35.3, 35.4, 34.9, 34.7, 34.8)**

The outliers for X_0 are: 27.2, 26.3)


X
X_1
X_0
34
32
30
28
1
2
3

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

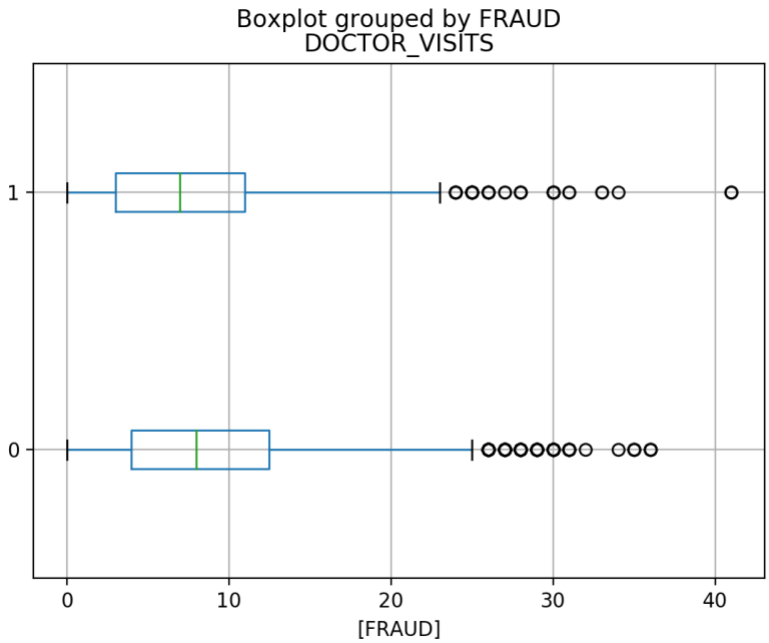
You are asked to use the Nearest Neighbors algorithm to predict the likelihood of fraud.

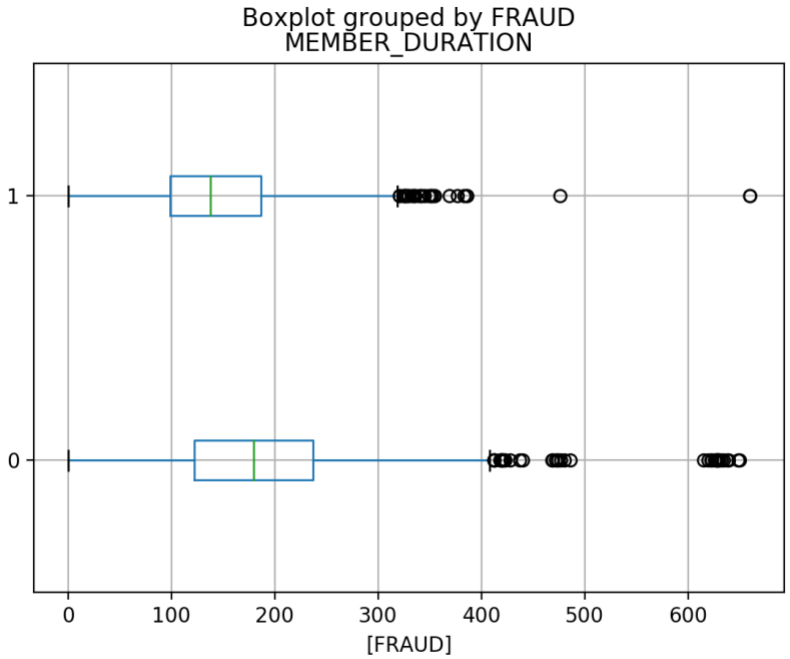
a) (5 points) What percent of investigations are found to be fraudulent? Please give your answer up to 4 decimal places.

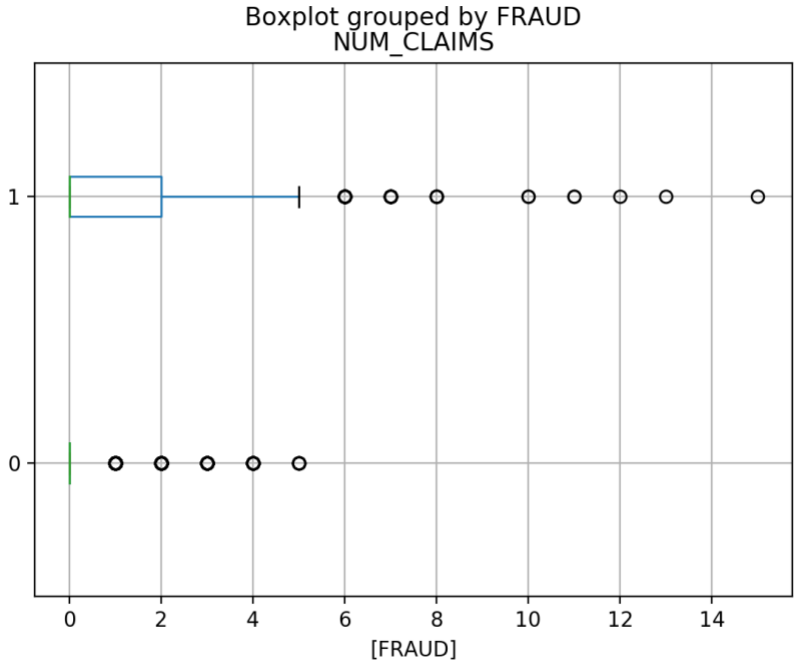
Aa) Fraudulent percentage: 19.9497%

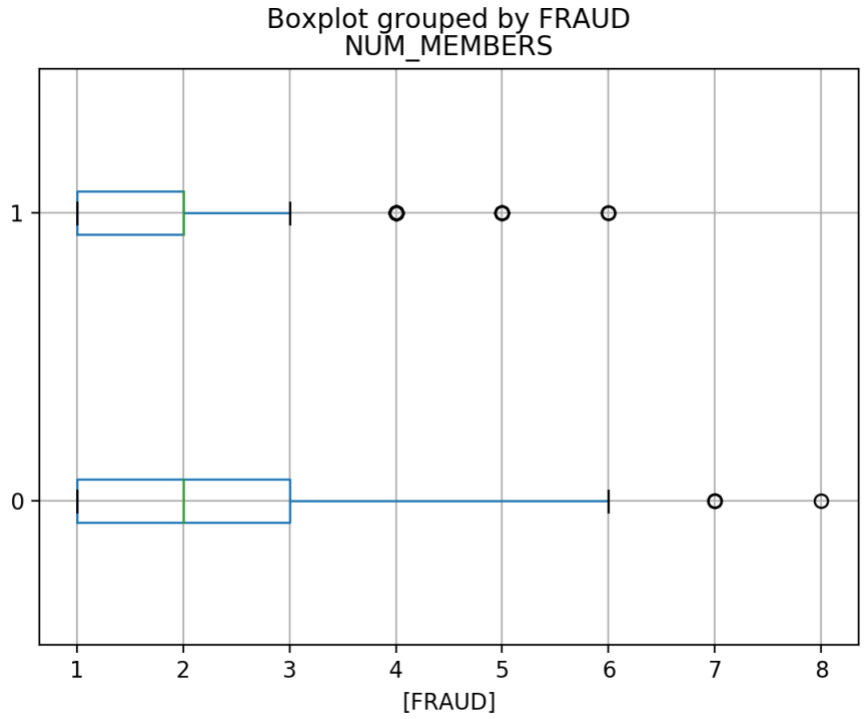
b) (5 points) Use the BOXPLOT function to produce horizontal box-plots. For each interval variable, one box-plot for the fraudulent observations, and another box-plot for the non-fraudulent observations. These two box-plots must appear in the same graph for each interval variable.

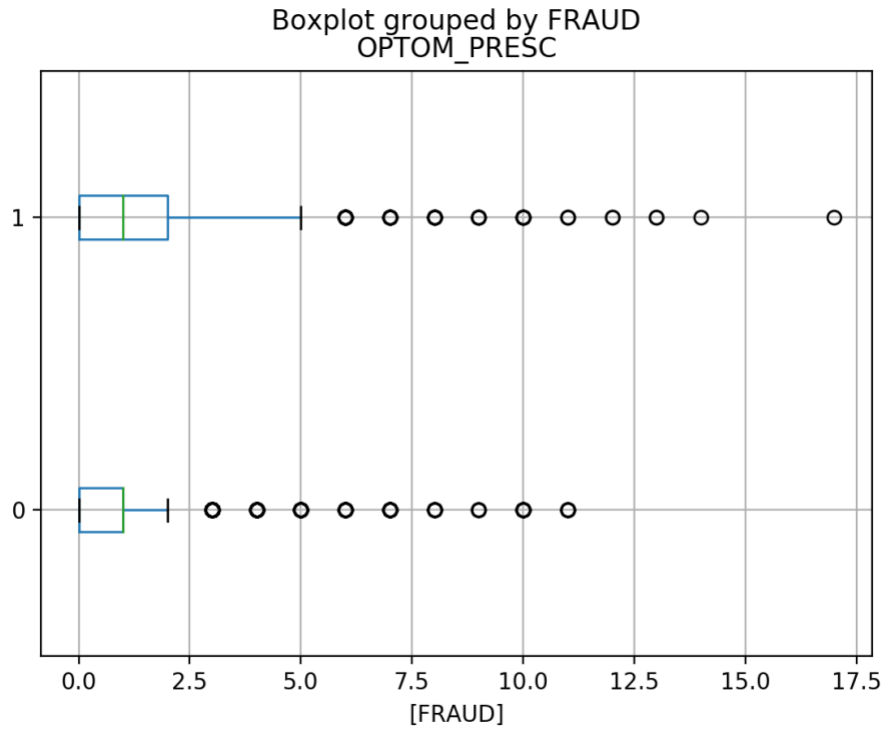
Ab)

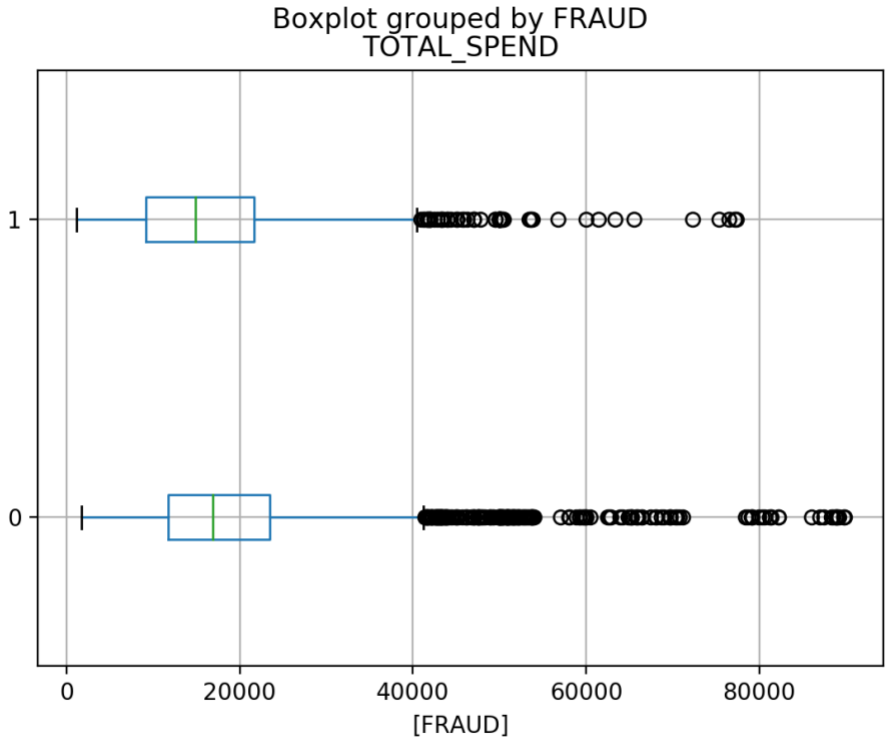



Boxplot grouped by FRAUD
MEMBER_DURATION
1
0
0
100
200
300
400
500
600
[FRAUD]


Boxplot grouped by FRAUD
NUM_CLAIMS
1
0
0
2
4
6
8
10
12
14
[FRAUD]


Boxplot grouped by FRAUD
NUM_MEMBERS
1
0
1
2
3
4
5
6
7
8
[FRAUD]


Boxplot grouped by FRAUD
OPTOM_PRESC
1
0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
[FRAUD]



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.

Ac) i. 6 dimensions

ii. Transformation Matrix

[[-6.49862374e-08 0.00000000e+00 0.00000000e+00 -0.00000000e+00 0.00000000e+00 -0.00000000e+00]

[0.00000000e+00 -2.94741983e-04 0.00000000e+00 0.00000000e+00 -0.00000000e+00 -0.00000000e+00]

[-0.00000000e+00 0.00000000e+00 -7.68683456e-04 0.00000000e+00 -0.00000000e+00 -0.00000000e+00]

[0.00000000e+00 -0.00000000e+00 0.00000000e+00 5.78327741e-05 -0.00000000e+00 -0.00000000e+00]

[0.00000000e+00 -0.00000000e+00 -0.00000000e+00 -0.00000000e+00 -2.39238772e-07 -0.00000000e+00]

[0.00000000e+00 0.00000000e+00 -0.00000000e+00 -0.00000000e+00 -0.00000000e+00 4.66565230e-11]]

If, the transpose of the Transformed_matrix multiplied by the transformed matrix gives an Identity matrix then the matrix is orthonormal.

i.e,

IMat = trans_m.transpose().dot(trans_m)

IMat = Transpose(Transformed_matrix)*Transformed_matrix:

IMat =

[[1.00000000e+00 -5.55111512e-17 1.17961196e-16 4.85722573e-17 1.38777878e-17 -3.46944695e-17]

[-5.55111512e-17 1.00000000e+00 -1.66533454e-16 -2.08166817e-17 6.93889390e-18 -3.46944695e-17]

[1.17961196e-16 -1.66533454e-16 1.00000000e+00 -1.95427442e-17 3.46944695e-18 -3.51281504e-17]

[4.85722573e-17 -2.08166817e-17 -1.95427442e-17 1.00000000e+00 -2.38524478e-16 -1.97758476e-16]

[1.38777878e-17 6.93889390e-18 3.46944695e-18 -2.38524478e-16 1.00000000e+00 -6.85215773e-17]

[-3.46944695e-17 -3.46944695e-17 -3.51281504e-17 -1.97758476e-16 -6.85215773e-17 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
ii. (5 points) Explain the meaning of the score function return value.

Ad) i: 0.8825503355704698

Ii: The score gives the accuracy of the predicted 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*.

Ae) The neighbors are: [[588 1199 2264 1246 3809]]

The distances are:
[[1.64636127e-10 3.78641748e-04 5.36680515e-04 7.69632938e-04 8.81391774e-04]]

CASE_ID 589
FRAUD 1
TOTAL_SPEND 7500
DOCTOR_VISITS 15
NUM_CLAIMS 3
MEMBER_DURATION 127
OPTOM_PRESC 2
NUM_MEMBERS 2

CASE_ID 1200
FRAUD 1
TOTAL_SPEND 10000
DOCTOR_VISITS 16
NUM_CLAIMS 3
MEMBER_DURATION 124
OPTOM_PRESC 2
NUM_MEMBERS 1

CASE_ID 2265
FRAUD 1
TOTAL_SPEND 13800
DOCTOR_VISITS 15
NUM_CLAIMS 3
MEMBER_DURATION 121
OPTOM_PRESC 1
NUM_MEMBERS 1

CASE_ID 1247
FRAUD 1
TOTAL_SPEND 10200
DOCTOR_VISITS 13
NUM_CLAIMS 3
MEMBER_DURATION 119
OPTOM_PRESC 2
NUM_MEMBERS 3

CASE_ID 3810
FRAUD 1
TOTAL_SPEND 20000
DOCTOR_VISITS 16
NUM_CLAIMS 3
MEMBER_DURATION 124
OPTOM_PRESC 0
NUM_MEMBERS 2

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?

Af) The predicted probability is 100% and since it is greater than the fraudulent percentage i.e., 19.9497%, it will be classified as fraudulent. Based on this criterion, the observation is not misclassified.