CS 584: Machine Learning

Spring 2020 Assignment 1

Question 1 (40 points)

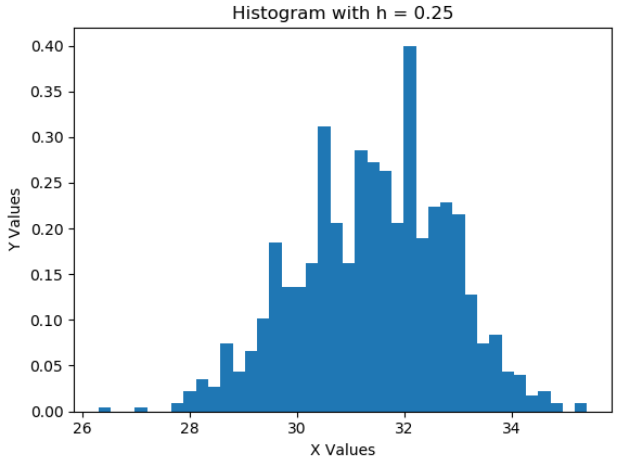
Write a Python program to calculate the density estimator of a histogram. Use the field *x* in the NormalSample.csv file.

a) (5 points) According to Izenman (1991) method, what is the recommended bin-width for the histogram of x?
Ans. Recommended bin-width according to Izenman method: 0.3998667554864774

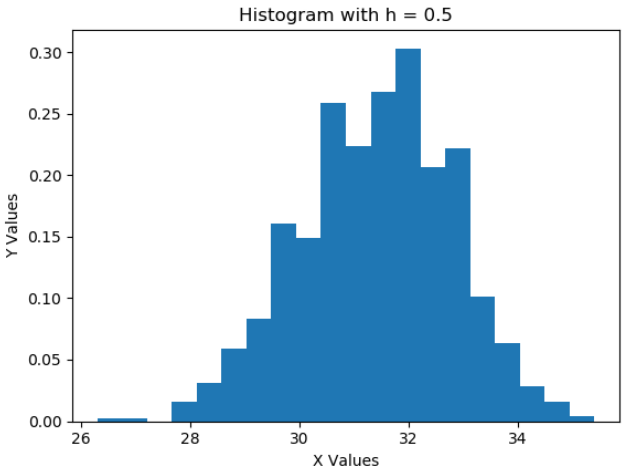
b) (5 points) What are the minimum and the maximum values of the field x?
Ans. Minimum Value: 26.3, Maximum Value: 35.4

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: Value of a: 26, Value of 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.
Ans: Coordinates: [(26.125, 0.0), (26.375, 0.003996003996003996),
(26.625, 0.0), (26.875, 0.0),
(27.125, 0.003996003996003996), (27.375, 0.0),
(27.625, 0.007992007992007992), (27.875, 0.015984015984015984),
(28.125, 0.023976023976023976), (28.375, 0.03596403596403597),
(28.625, 0.03596403596403597), (28.875, 0.07192807192807193),
(29.125, 0.059940059940059943), (29.375, 0.14785214785214784),
(29.625, 0.11188811188811189), (29.875, 0.1878121878121878),
(30.125, 0.14785214785214784), (30.375, 0.2677322677322677),
(30.625, 0.1838161838161838), (30.875, 0.22777222777222778),
(31.125, 0.17582417582417584), (31.375, 0.33166833166833165),
(31.625, 0.23976023976023977), (31.875, 0.32367632367632365),
(32.125, 0.22777222777222778), (32.375, 0.2837162837162837),
(32.625, 0.21178821178821178), (32.875, 0.22777222777222778),
(33.125, 0.10789210789210789), (33.375, 0.13186813186813187),
(33.625, 0.05194805194805195), (33.875, 0.06393606393606394),
(34.125, 0.03596403596403597), (34.375, 0.023976023976023976),
(34.625, 0.011988011988011988), (34.875, 0.007992007992007992),
(35.125, 0.0), (35.375, 0.007992007992007992),
(35.625, 0.0), (35.875, 0.0)]

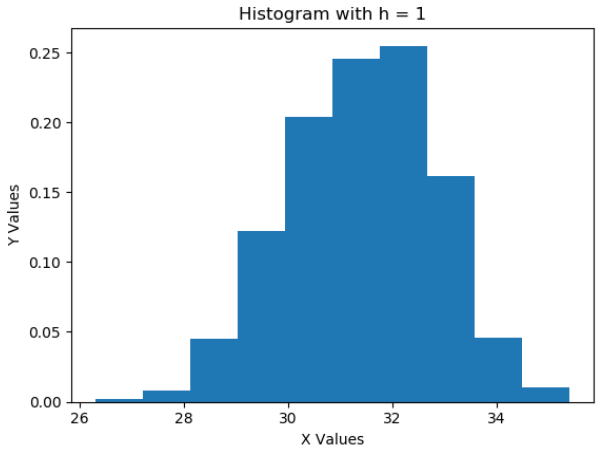


e) (5 points) Use h = 0.5, minimum = a and maximum = b. List the coordinates of the density estimator. Paste the histogram drawn using Python or your favorite graphing tools.
Ans: Coordinates: [(26.25, 0.001998001998001998), (26.75, 0.0),
(27.25, 0.001998001998001998), (27.75, 0.011988011988011988),
(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)]

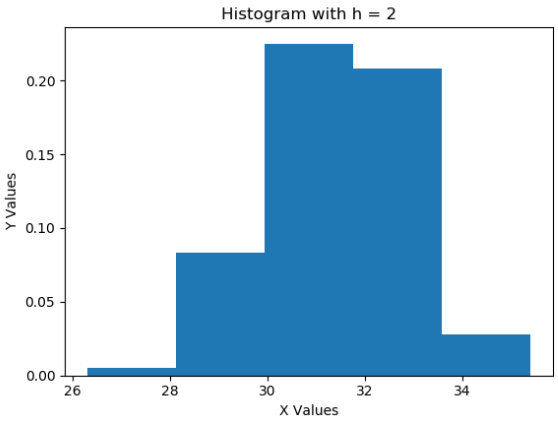


f) (5 points) Use h = 1, minimum = a and maximum = b. List the coordinates of the density estimator. Paste the histogram drawn using Python or your favorite graphing tools.
Ans: Coordinates: [(26.5, 0.000999000999000999), (27.5, 0.006993006993006993),
(28.5, 0.04195804195804196), (29.5, 0.12687312687312688),
(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: [(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: According to me histogram with h=0.5 provides best insights into the shape and the spread of the distribution of the field x, because even though histogram with h=2 provides a bigger picture but it becomes difficult to deduce accurate information about the data. For h=0.25 the bin-width is too small, so it becomes too hard to observe the histogram and deduce the information from it. Histogram with h=1 can be considered as a good representation but it misses some details. Apart from this according to Izenman's method the recommended Bin-width for histogram of x=0.4(approximately) and histogram with h=0.5 is close to this

recommendation. Thus h=0.5 provides best insights into the shape and the spread of the distribution of the field x.

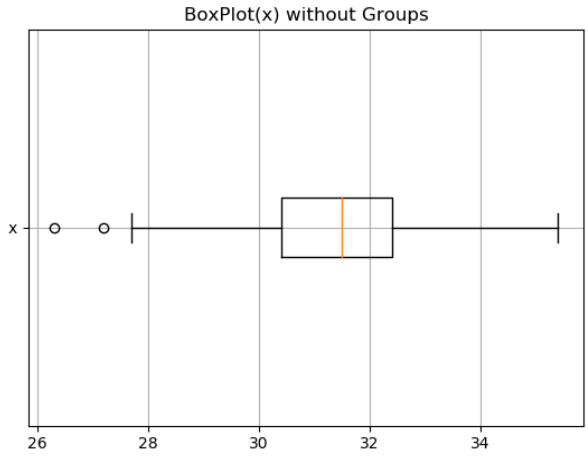
Question 2 (20 points)

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

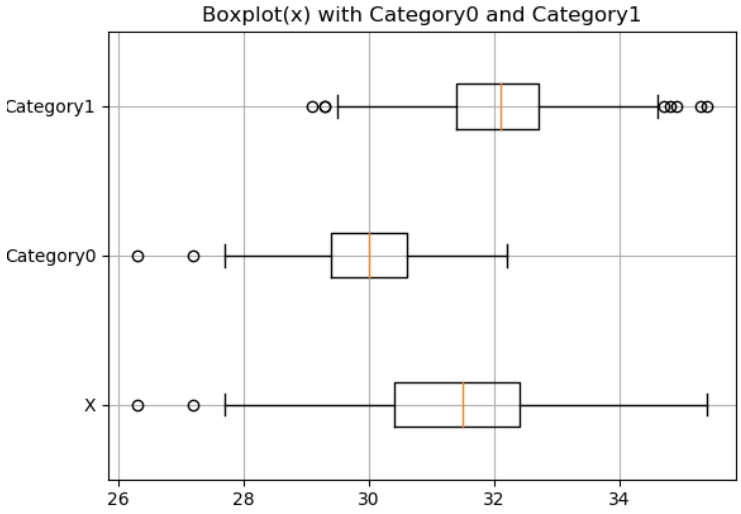
a) (5 points) What is the five-number summary of x? What are the values of the 1.5 IQR whiskers?
Ans: Minimum: 26.3
Q1: 30.4
Median: 31.5
Q3: 32.4
Maximum: 35.4
Lower whisker value: 27.4
Upper whisker value: 35.4

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: Group 0:
Minimum: 26.3
Q1: 29.4
Median: 30.0
Q3: 30.6
Maximum: 32.2
Lower whisker value for Category 0: 27.599999999999994
Upper whisker value for Category 0: 32.2
Group 1:
Minimum: 29.1
Q1: 31.4
Median: 32.1
Q3: 32.7
Maximum: 35.4
Lower whisker value for Category 1: 29.449999999999992
Upper whisker value for Category 1: 34.650000000000006

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, Python's boxplot has displayed the 1.5 IQR whiskers correctly.
Lower whisker value: 27.4
Upper whisker value: 35.4



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 Entire Data: [27.2, 26.3] (Small Circles on x Boxplot)
Outliers for Category 0: [27.2, 26.3] (Small Circles on Category0 Boxplot)
Outliers for Category 1: [35.3, 29.3, 35.4, 34.9, 34.7, 34.8, 29.3, 29.1] (Small Circles on Category1 Boxplot)

Question 3 (40 points)

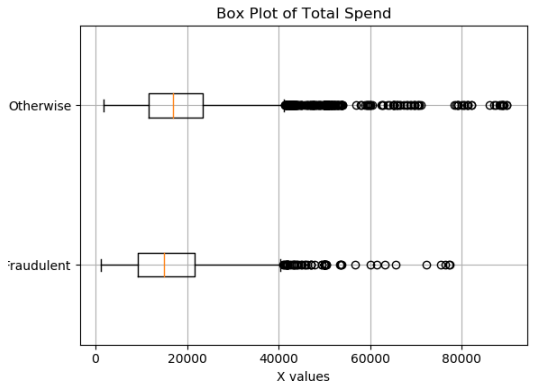
The data, FRAUD.csv, contains results of fraud investigations of 5,960 cases. The binary variable FRAUD indicates the result of a fraud investigation: 1 = Fraudulent, 0 = Otherwise. The other interval variables contain information about the cases.

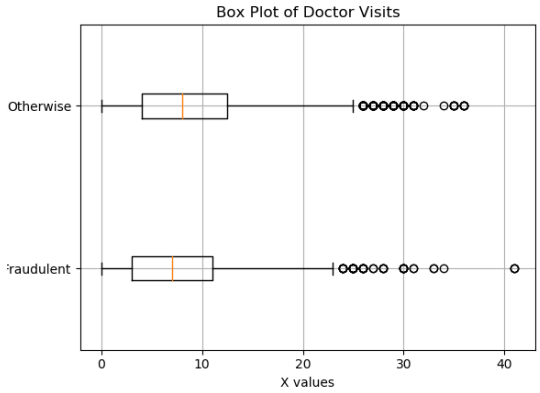
1. TOTAL_SPEND: Total amount of claims in dollars
2. DOCTOR_VISITS: Number of visits to a doctor
3. NUM_CLAIMS: Number of claims made recently
4. MEMBER_DURATION: Membership duration in number of months
5. OPTOM_PRESC: Number of optical examinations
6. NUM_MEMBERS: Number of members covered

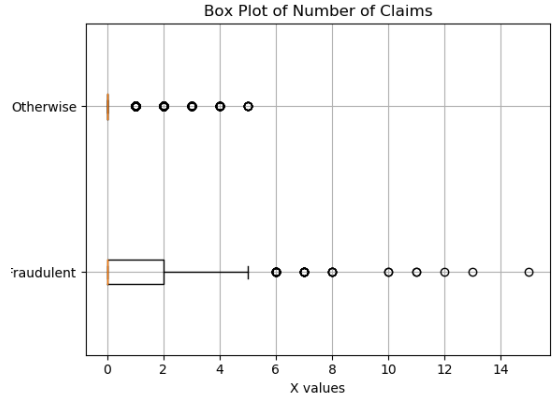
You are asked to use the Nearest Neighbors algorithm to predict the likelihood of fraud.

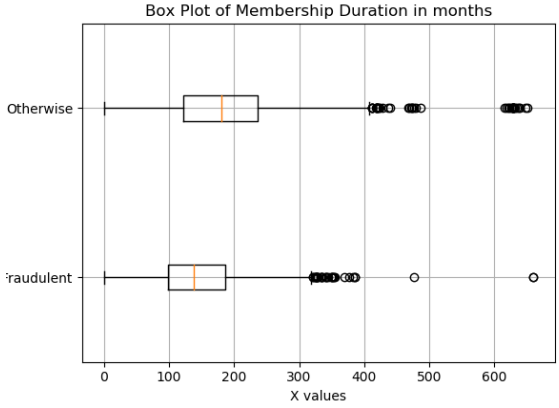
a) (5 points) What percent of investigations are found to be fraudulent? Please give your answer up to 4 decimal places.
Ans: 19.9497% of investigations are found to be fraudulent.

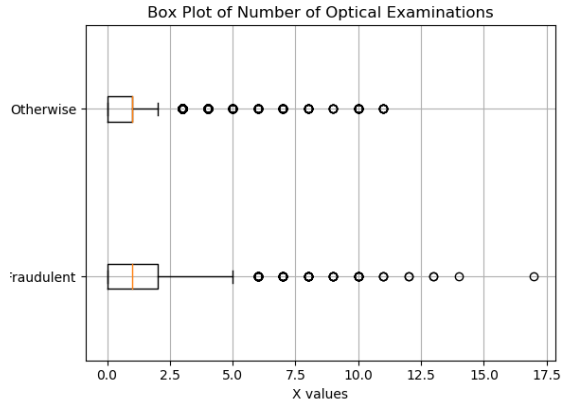
b) (5 points) Use the BOXPLOT function to produce horizontal box-plots. For each interval variable, one box-plot for the fraudulent observations, and another box-plot for the non-fraudulent observations. These two box-plots must appear in the same graph for each interval variable.
Ans:

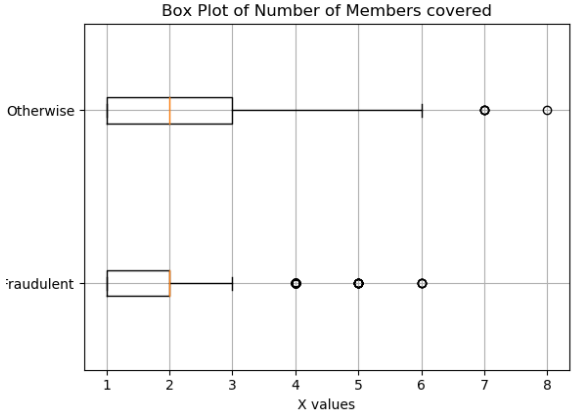



Box Plot of Doctor Visits
Otherwise
raudulent
0
10
20
30
40
X values


Box Plot of Number of Claims
Otherwise
raudulent
0
2
4
6
8
10
12
14
X values


Box Plot of Membership Duration in months
Otherwise
raudulent
0
100
200
300
400
500
600
X values





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: Total 6 dimensions are used and they are (TOTAL_SPEND, DOCTOR_VISITS, NUM_CLAIMS, MEMBER_DURATION, OPTOM_PRESC, and NUM_MEMBERS) because all the eigenvalues are greater than 1.

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]]

2. Proof that the resulting variables are actually orthonormal.
Identity Matrix = Transpose(Transformation of x) * (Transformation of x)
= [[1.00000000e+00 -2.99781901e-16 -4.56882795e-16 5.45884952e15
  1.20129601e-15 -1.27176915e-16]
[-2.99781901e-16 1.00000000e+00 -6.56592836e-16 -2.76891140e-14
 -1.22818422e-15 7.71951947e-16]
[-4.56882795e-16 -6.56592836e-16 1.00000000e+00 3.50132250e-15
  1.14491749e-16 -2.32452946e-16]
[5.45884952e-15 -2.76891140e-14 3.50132250e-15 1.00000000e+00
  1.14821347e-14 -3.47768689e-15]
[1.20129601e-15 -1.22818422e-15 1.14491749e-16 1.14821347e-14
  1.00000000e+00 -6.27969898e-16]
[-1.27176915e-16 7.71951947e-16 -2.32452946e-16 -3.47768689e-15
 -6.27969898e-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 Function result : 0.8778523489932886
   ii. (5 points) Explain the meaning of the score function return value.
   Ans: It returns the mean accuracy on the given test data and labels that means our model has 87% accuracy on training 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: Transformed Input variables : [[-0.02886529 0.00853837 -0.01333491 0.0176811
0.00793805 0.0044727]]
Target Values: All the values in the "FRAUD" column from Fraud.csv
Neighbors: [[588 2897 1199 1246 886]]

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: Predicted label for Test Data: [1]

Probability values of test set: [[0. 1.]]

The value of predict probability function for label 1 has come to 1.0, which is greater than 0.19 that we got in (3.a). Thus the observation is fraudulent and not misclassified.