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.

1. (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

1. (5 points) What are the minimum and the maximum values of the field x?

The minimum value is : 26.3

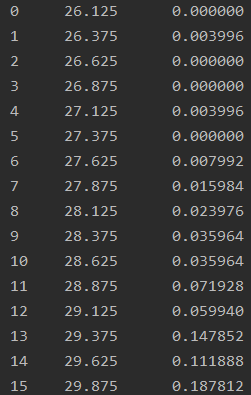
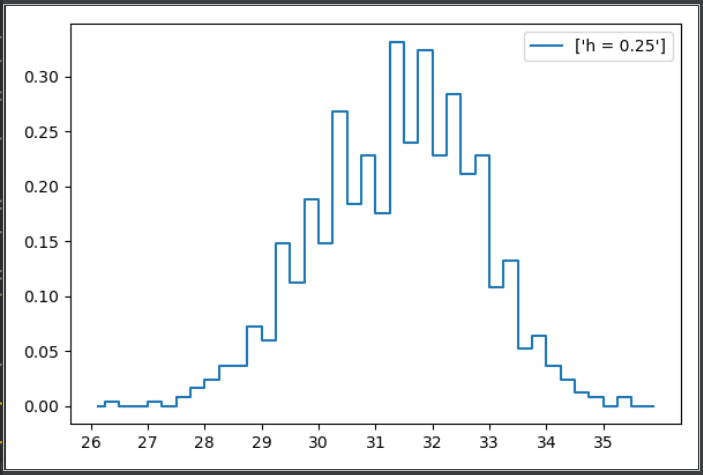
The maximum value is : 35.4

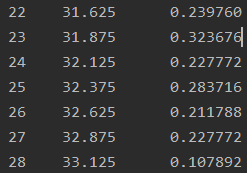
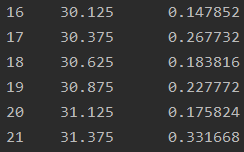
1. (5 points) Let a be the largest integer less than the minimum value of the field x, and b be the smallest integer greater than the maximum value of the field x. What are the values of a and b?

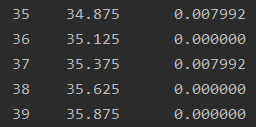
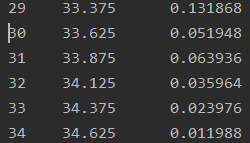
a = 26

b = 36

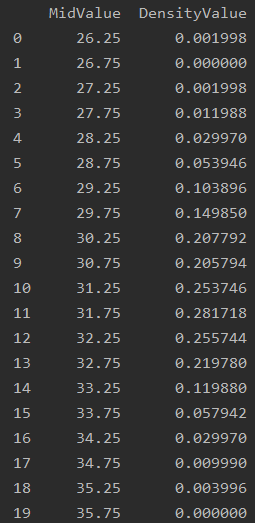
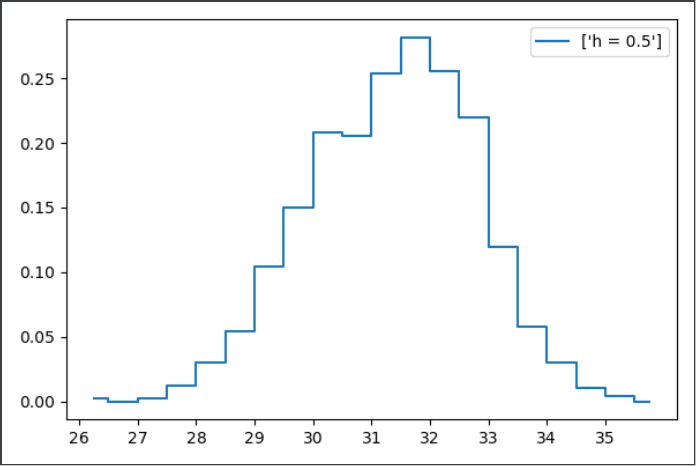
1. (5 points) Use h = 0.25, minimum = a and maximum = b. List the coordinates of the density estimator. Paste the histogram drawn using Python or your favorite graphing tools.



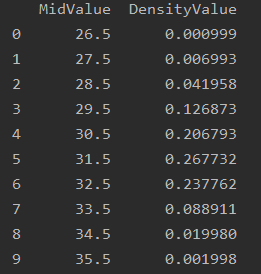
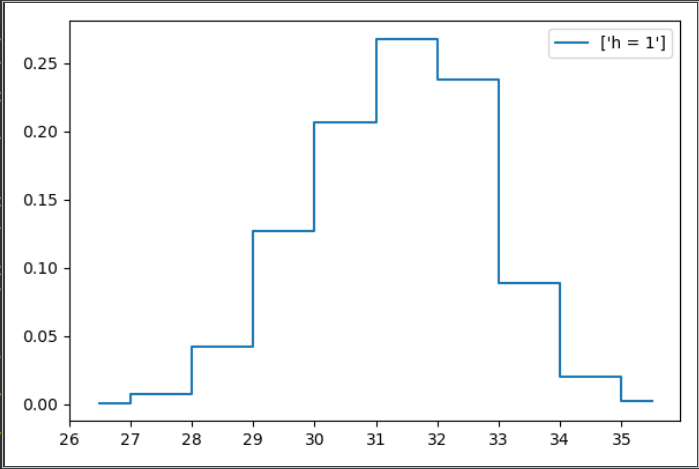




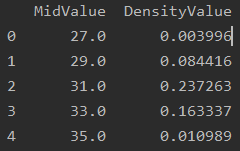
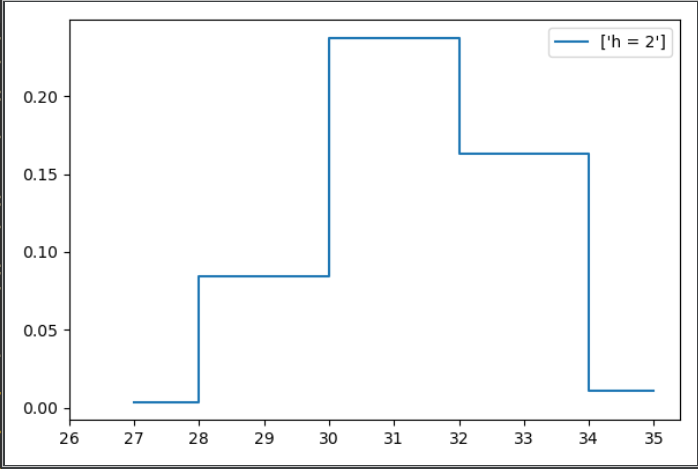
1. (5 points) Use h = 0.5, minimum = a and maximum = b. List the coordinates of the density estimator. Paste the histogram drawn using Python or your favorite graphing tools.



1. (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.



1. (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.



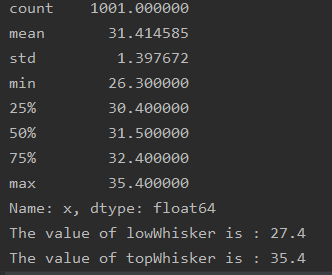
1. (5 points) Among the four histograms, which one, in your honest opinions, can best provide your insights into the shape and the spread of the distribution of the field x? Please state your arguments.

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.

1. (5 points) What is the five-number summary of x? What are the values of the 1.5 IQR whiskers?



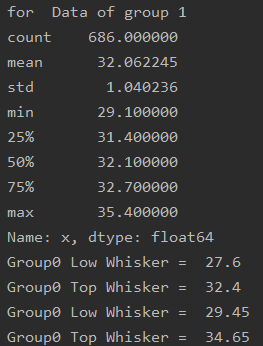
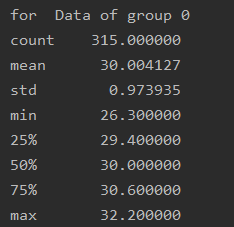
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Minimum | Q1 | Median | Q3 | Maximum |
| 26.3 | 30.4 | 31.5 | 32.4 | 35.4 |

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

1. (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?

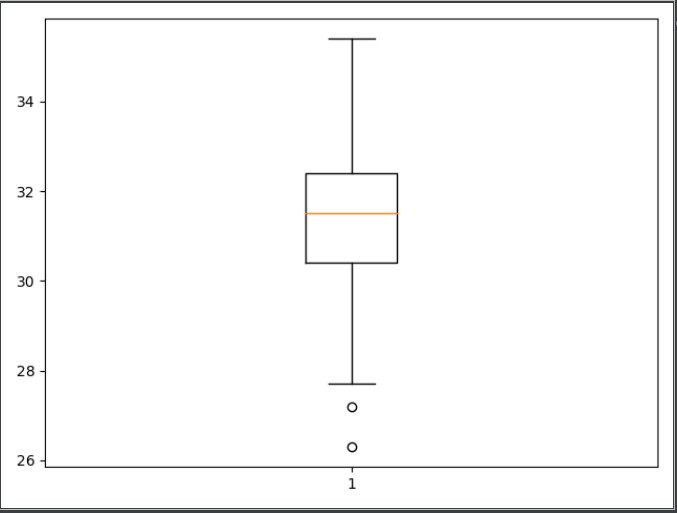


|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Group | Minmum | Q1 | Median | Q3 | Maximum |
| 0 | 26.3 | 29.4 | 30 | 30.6 | 32.2 |
| 1 | 29.1 | 31.4 | 32.1 | 32.7 | 35.4 |

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

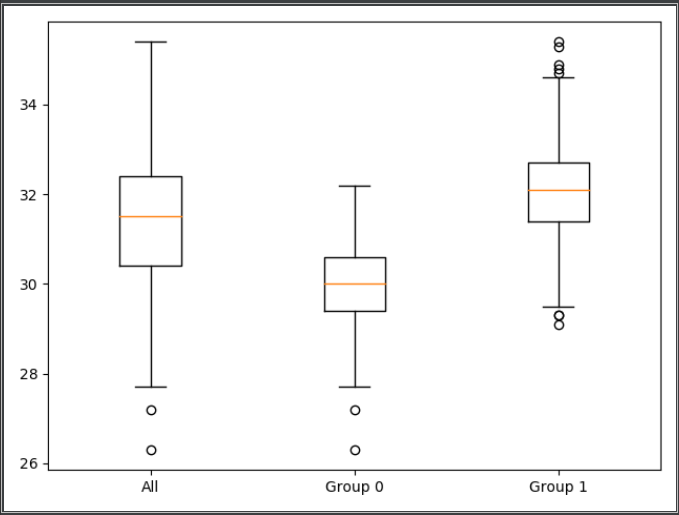
For group1 : IQR = Q3-Q1 =1.3, 1.5\*IQR = 1.95 Lower whisker is max {minimum, Q1-1.5\*IQR} = 29.45, and upper whisker is min {Q3+1.5\*IQR, maximum} = 34.65

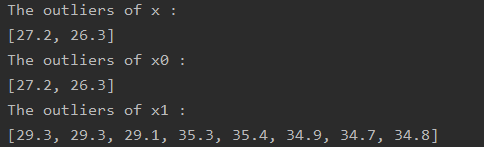
1. (5 points) Draw a boxplot of x (without the group) using the Python boxplot function. Can you tell if the Python’s boxplot has displayed the 1.5 IQR whiskers correctly?



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

1. (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*.





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 |

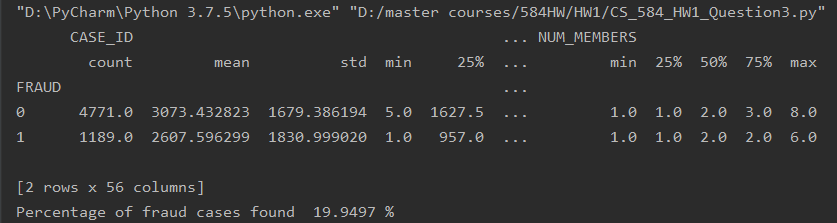
# Question 3 (40 points)

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.

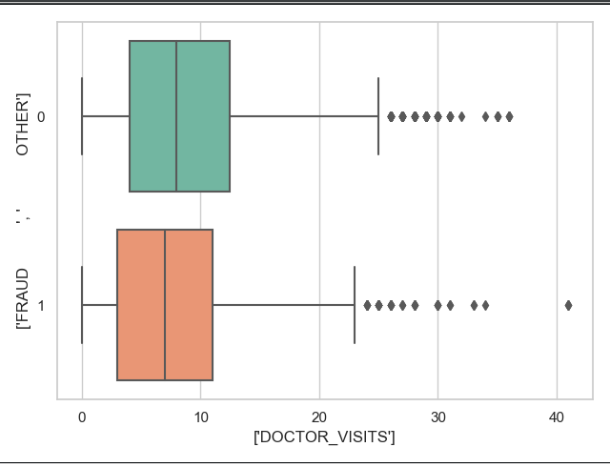
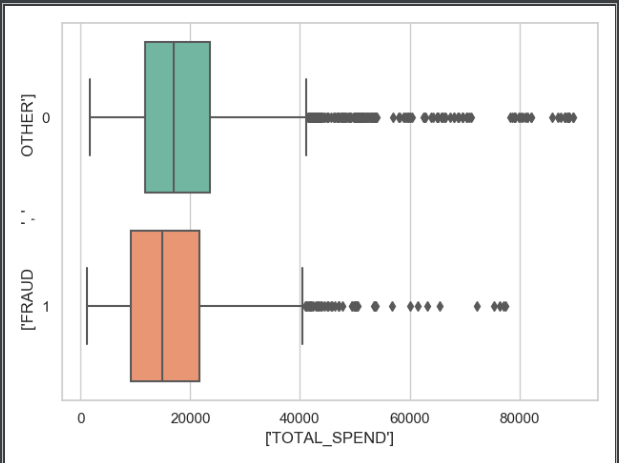
1. TOTAL\_SPEND: Total amount of claims in dollars
2. DOCTOR\_VISITS: Number of visits to a doctor
3. NUM\_CLAIMS: Number of claims made recently
4. MEMBER\_DURATION: Membership duration in number of months
5. OPTOM\_PRESC: Number of optical examinations
6. NUM\_MEMBERS: Number of members covered

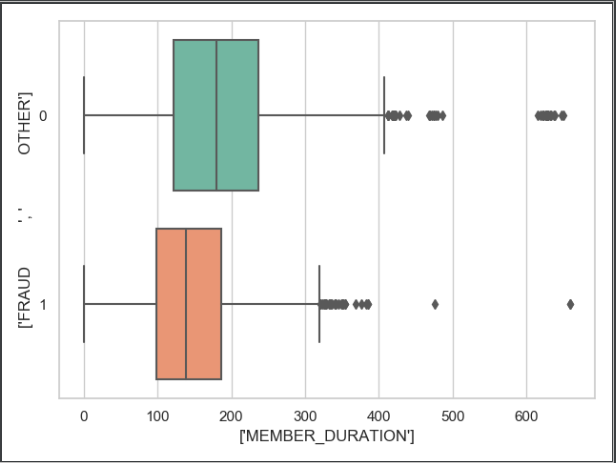
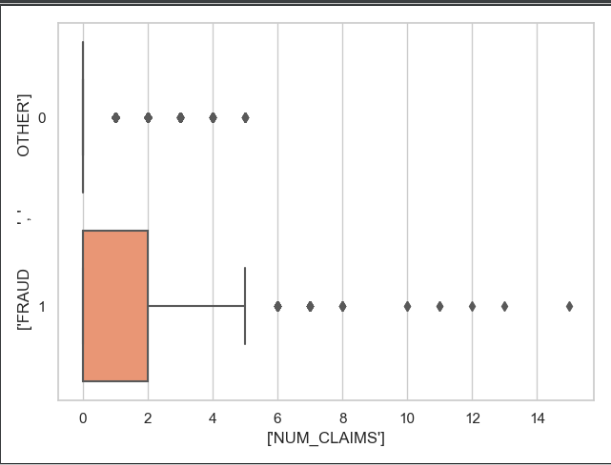
You are asked to use the Nearest Neighbors algorithm to predict the likelihood of fraud.

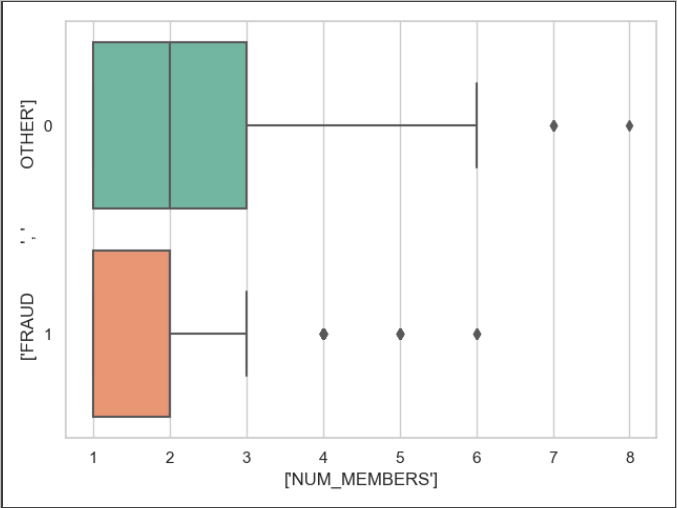
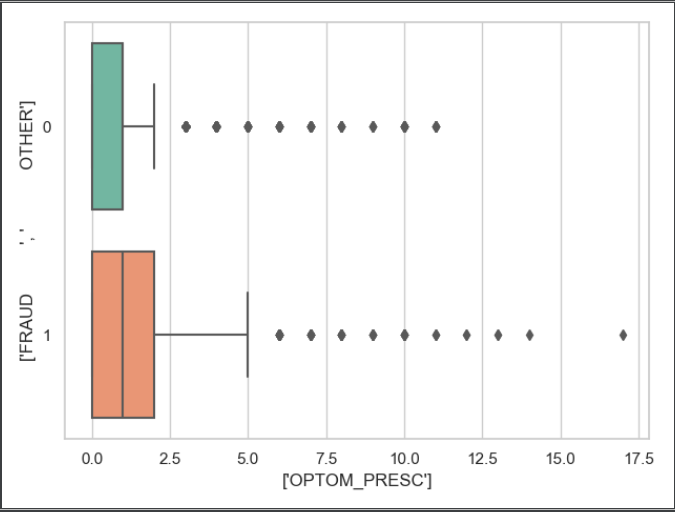
1. (5 points) What percent of investigations are found to be fraudulent? Please give your answer up to 4 decimal places.



1. (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.





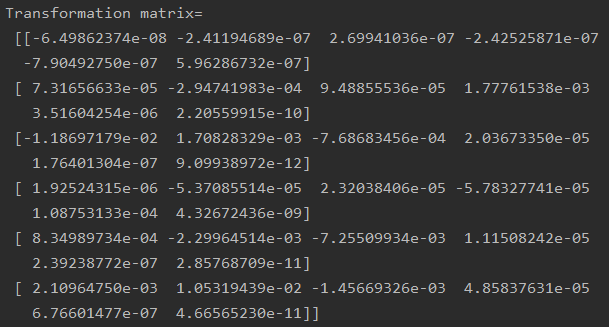


1. (10 points) Orthonormalize interval variables and use the resulting variables for the nearest neighbor analysis. Use only the dimensions whose corresponding eigenvalues are greater than one.
   1. (5 points) How many dimensions are used?

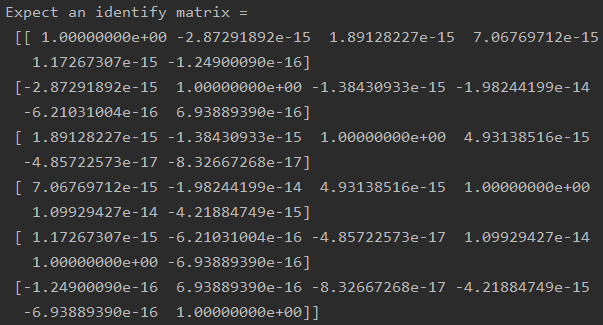


They are all bigger than 1. So the six dimensions are used

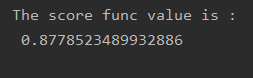
* 1. (5 points) Please provide the transformation matrix? You must provide proof that the resulting variables are actually orthonormal.



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.



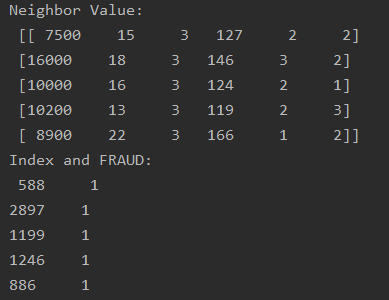
1. (10 points) Use the NearestNeighbors module to execute the Nearest Neighbors algorithm using exactly five neighbors and the resulting variables you have chosen in c). The KNeighborsClassifier module has a score function.
   1. (5 points) Run the score function, provide the function return value



* 1. (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%

1. (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*.



1. (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.