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?
Recommended bin-width for the histogram of x **= 0.3998667554864774**

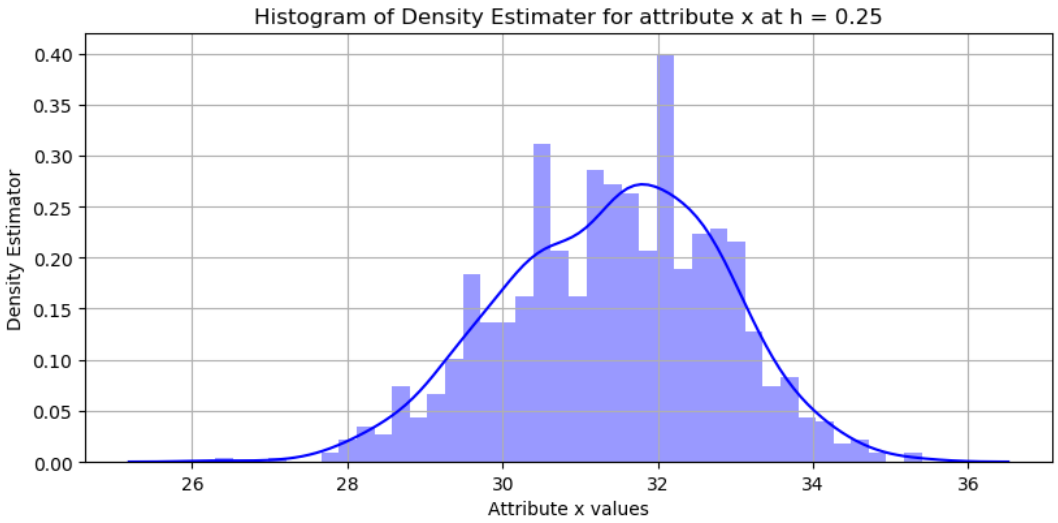
b) (5 points) What are the minimum and the maximum values of the field x?
Minimum of the field x = **26.3**
Maximum of the field x = **35.4**

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?
Value of 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.

Density Estimator p(mi) for respective mid-point mi as follow:
[mi : p(mi)] at h = 0.25

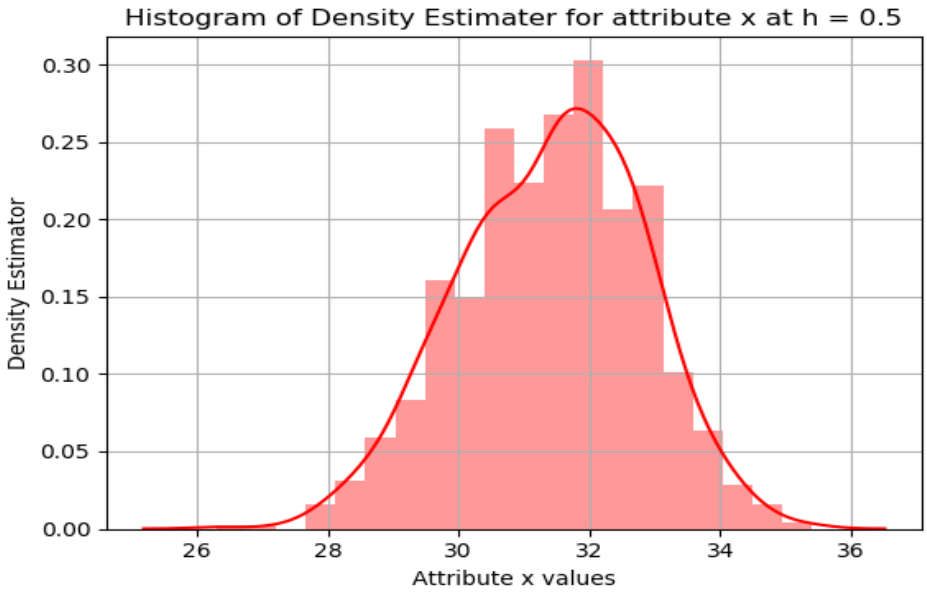
[26.125 : 0.0] [26.375 : 0.003996003996003996] [26.625 : 0.0] [26.875 : 0.0] [27.125 : 0.003996003996003996] [27.375 : 0.0] [27.625 : 0.007992007992007992] [27.875 : 0.011988011988011988] [28.125 : 0.023976023976023976] [28.375 : 0.01998001998001998] [28.625 : 0.03596403596403597] [28.875 : 0.05194805194805195] [29.125 : 0.059940059940059943] [29.375 : 0.0919080919080919] [29.625 : 0.11188811188811189] [29.875 : 0.12387612387612387] [30.125 : 0.14785214785214784] [30.375 : 0.17182817182817184] [30.625 : 0.1838161838161838] [30.875 : 0.15984015984015984] [31.125 : 0.17582417582417584] [31.375 : 0.1958041958041958] [31.625 : 0.23976023976023977] [31.875 : 0.1878121878121878] [32.125 : 0.22777222777222778] [32.375 : 0.17182817182817184] [32.625 : 0.21178821178821178] [32.875 : 0.15184815184815184] [33.125 : 0.10789210789210789] [33.375 : 0.08791208791208792] [33.625 : 0.05194805194805195] [33.875 : 0.05194805194805195] [34.125 : 0.03596403596403597] [34.375 : 0.015984015984015984] [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.

[mi : p(mi)] at h = 0.5

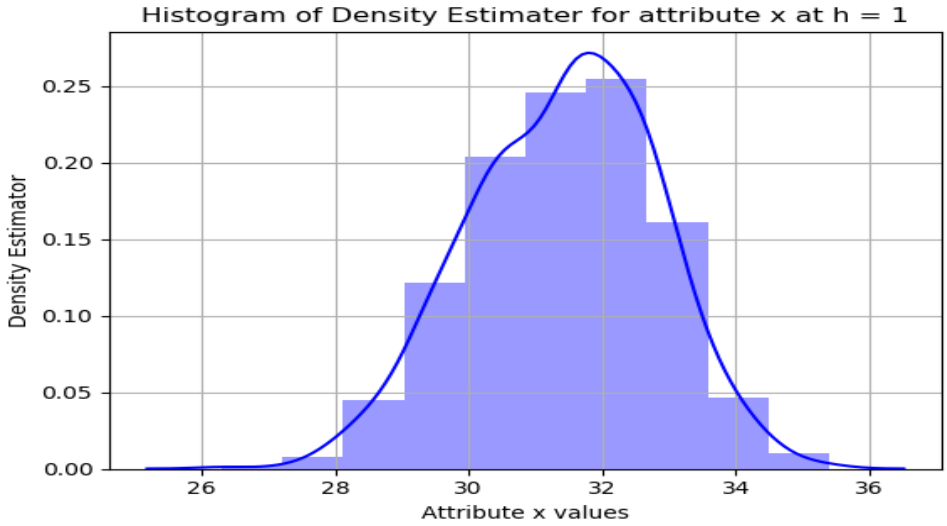
[26.25 : 0.001998001998001998] [26.75 : 0.0] [27.25 : 0.001998001998001998] [27.75 : 0.00999000999000999] [28.25 : 0.02197802197802198] [28.75 : 0.04395604395604396] [29.25 : 0.0759240759240] [29.75 : 0.11788211788211] [30.25 : 0.159840159840159] [30.75 : 0.17182817182817184] [31.25 : 0.18581418581418582] [31.75 : 0.21378621378621] [32.25 : 0.1998001998001] [32.75 : 0.18181818181818182] [33.25 : 0.0979020979020979] [33.75 : 0.05194805194805] [34.25 : 0.025974025974025976] [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.

[mi : p(mi)] at h = 1

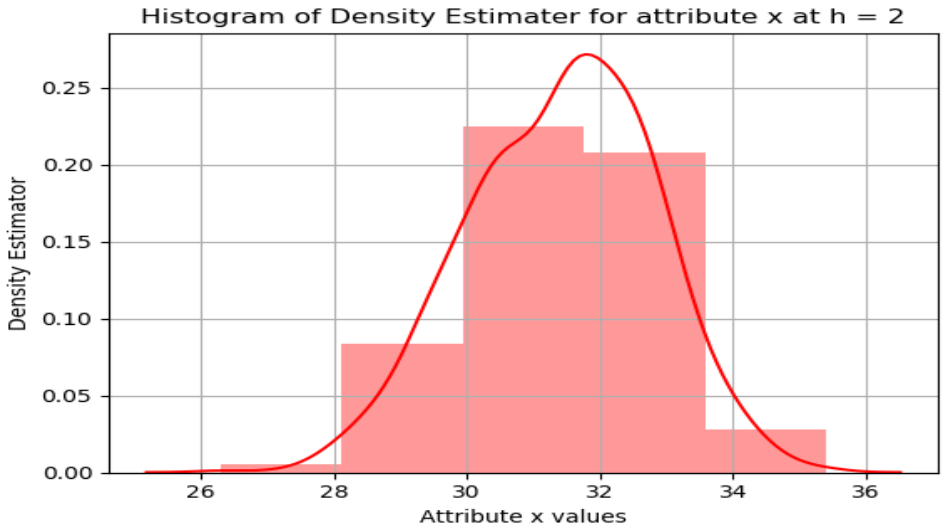
[26.5 : 0.000999000999000999] [27.5 : 0.005994005994005994] [28.5 : 0.03696303696303] [29.5 : 0.1108891108891109] [30.5 : 0.18981018981018982] [31.5 : 0.23376623376623376] [32.5 : 0.21878121878121878] [33.5 : 0.08591408591408592] [34.5 : 0.01998001998001998] [35.5 : 0.001998001998001998]



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.

[mi : p(mi)] at h = 2

[27.0 : 0.0034965034965034965] [29.0 : 0.07642357642357642] [31.0 : 0.2202797202797203] [33.0 : 0.16183816183816183] [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.

As per my opinion, h = 0.5 histogram can best provide insights into the shape and the spread of the distribution of the field x among four histograms.
As h = 0.25 gives fine grained analysis of the given input data distribution but the same time interpretation is much harder to understand, while h =1 and 2 give big picture of given input but not much informatic. But, h =0.5 satisfied both, in the form of ease of interpretation and informatic analysis.

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?

Five Number Summary of x:

Min	*Q1*	*Q2*	*Q3*	*Max*
26.3	*30.4*	*31.5*	*32.4*	*35.4*

IQR = |Q3 – Q1| = 32.4 – 30.4 = 2

Lower Wishker = Q1 – (1.5 * IQR)	Upper Wishker = Q3 + (1.5*IQR)
27.4	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?

Five Number Summary of x: (Group = 0)

Min	*Q1*	*Q2*	*Q3*	*Max*
26.3	29.4	30.0	30.6	32.2

IQR = |Q3 – Q1| = 32.4 – 30.4 = 1.2

Lower Wishker = Q1 – (1.5 * IQR)	Upper Wishker = Q3 + (1.5*IQR)
27.599999999999994	32.400000000000006

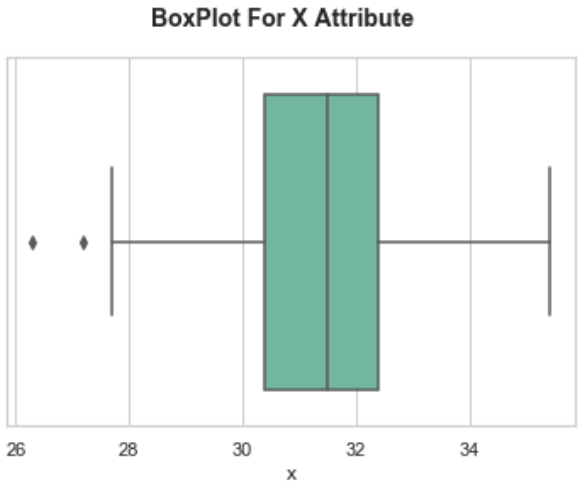
Five Number Summary of x: (Group = 1)

Min	*Q1*	*Q2*	*Q3*	*Max*
29.1	31.4	32.1	32.7	= 35.4

IQR = |Q3 – Q1| = 32.4 – 30.4 = 1.3

Lower Wishker = Q1 – (1.5 * IQR)	Upper Wishker = Q3 + (1.5*IQR)
29.449999999999992	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?

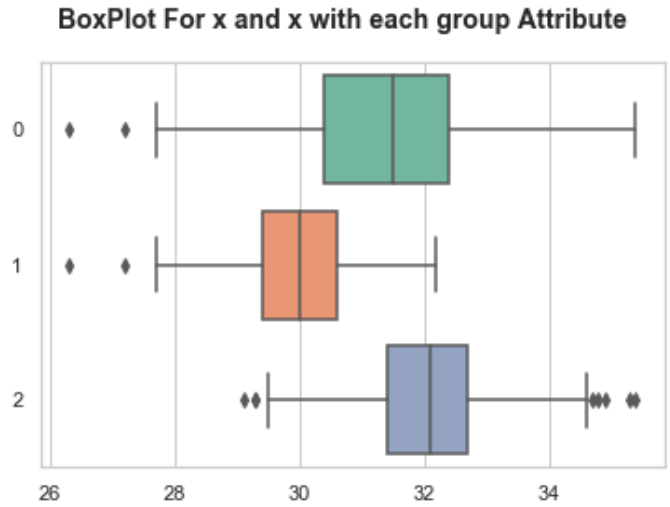


Verify the Box Plot Values as follow:
Five Number summery of x: min = 26.3, Q1 = 30.4, Q2 = 31.5 , Q3 = 32.4 and max = 35.4
IQR = |Q3-Q1| = 2.1

Lower Wishker = Q1 – (1.5 * IQR)	Upper Wishker = Q3 + (1.5*IQR)
27.4	35.4

Comparing Box-Plot values with above table values, we can say that Boxplot has displayed 1.5 IQR whiskers 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.



Group	Y Axis
x Attribute	0
x Attribute with Group = 0	1
x Attribute with Group = 1	2

Identifying Outlier in above three groups:

OutLiers for x – attribute:

Lower whisker outliers	27.2, 26.3
Upper whisker outliers	N/A

OutLiers for x – attribute with group = 0:

Lower whisker outliers	27.2, 26.3
Upper whisker outliers	N/A

OutLiers for x – attribute with group = 1:

Lower whisker outliers	29.3, 29.3, 29.1
Upper whisker outliers	35.3, 35.4, 34.9, 34.7, 34.8

Question 3 (40 points)

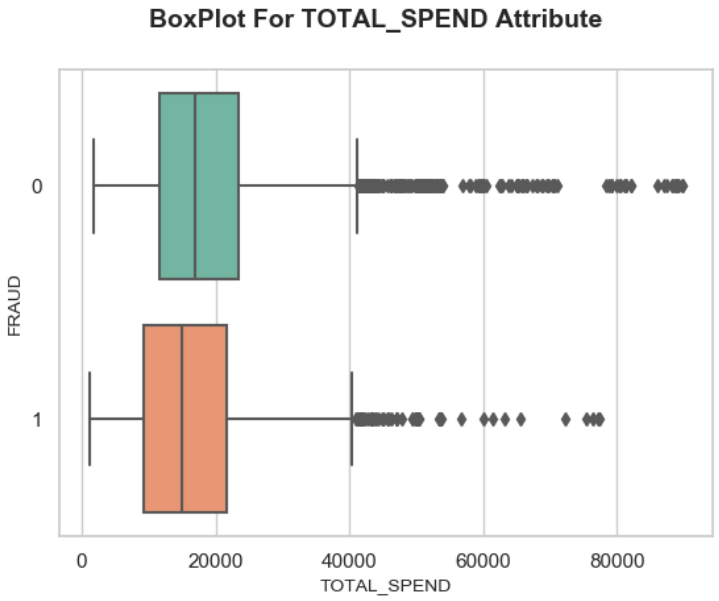
The data, FRAUD.csv, contains results of fraud investigations of 5,960 cases. The binary variable FRAUD indicates the result of a fraud investigation: 1 = Fraudulent, 0 = Otherwise. The other interval variables contain information about the cases.

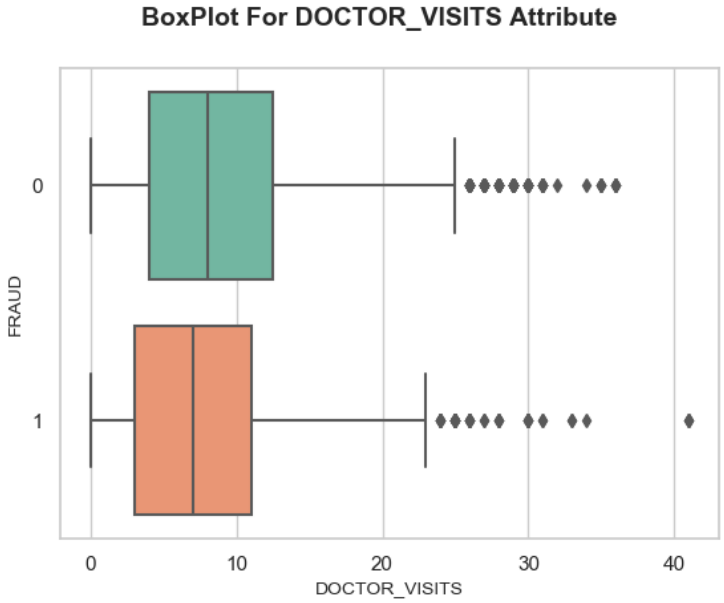
1. TOTAL_SPEND: Total amount of claims in dollars
2. DOCTOR_VISITS: Number of visits to a doctor
3. NUM_CLAIMS: Number of claims made recently
4. MEMBER_DURATION: Membership duration in number of months
5. OPTOM_PRESC: Number of optical examinations
6. NUM_MEMBERS: Number of members covered

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

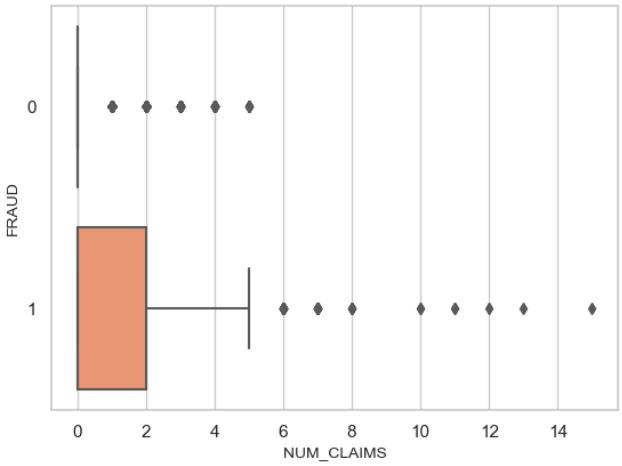
a) (5 points) What percent of investigations are found to be fraudulent? Please give your answer up to 4 decimal places.
19.9497 % of investigations are found to be fraudulent.

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.

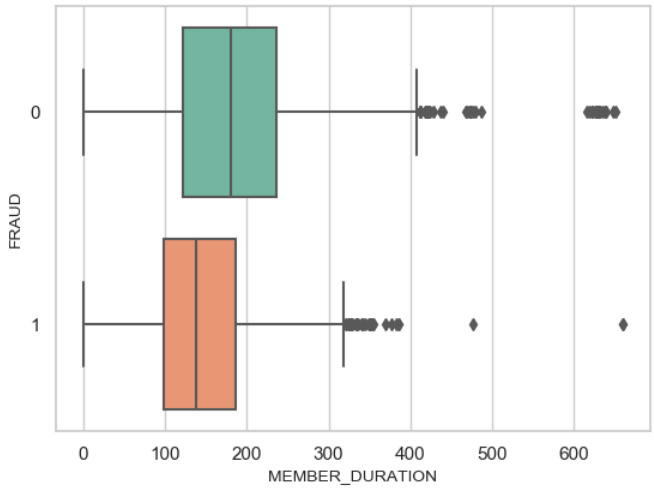




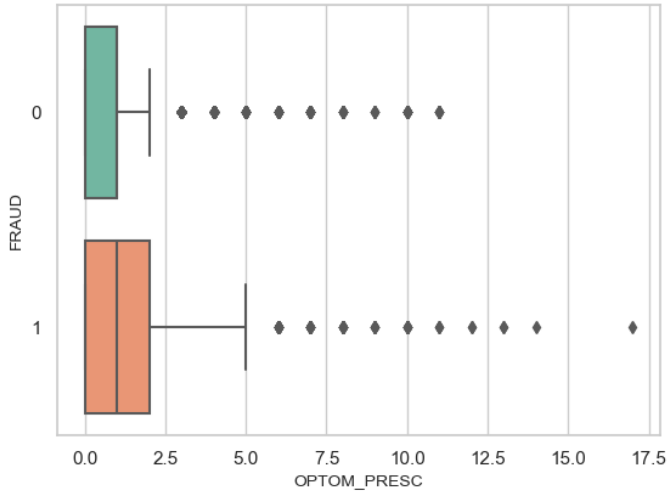
BoxPlot For NUM_CLAIMS Attribute

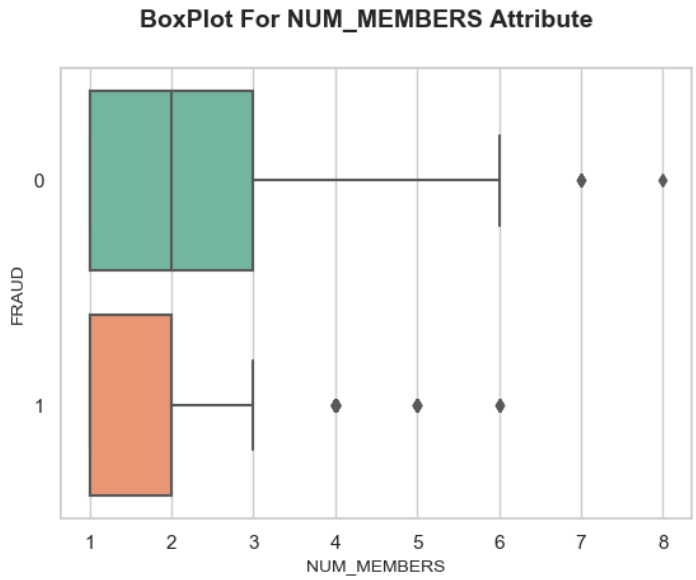


BoxPlot For MEMBER_DURATION Attribute



BoxPlot For OPTOM_PRESC Attribute





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?

Eigenvalues for all the dimensions as follows:

6.84728061 e+03	8.38798104 e+03	1.80639631 e+04	3.15839942 e+05	8.44539131 e+07	2.81233324 e+12

We can see that all the eigenvalues are greater than one. So total no of dimensions to be used = **6**

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

Transformation 6 × 6 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]]
```

Proof:

1. First the actual matrix of size 5960 × 6 is multiply with transformation of matrix of 6 × 6 size so, transformed matrix of 5960 × 6 is derived.

2. Then transpose of transformed matrix (6 × 5960)multiple by transformed matrix (5960 × 6), it gives resultant Identity matrix (6 × 6):

```
[[ 1.00000000e+00 -3.00432422e-16 -4.61219604e-16  5.45323877e-15
   1.20996962e-15 -1.28911638e-16]
 [-3.00432422e-16  1.00000000e+00 -6.44449771e-16 -2.76820667e-14
  -1.23512311e-15  7.78890841e-16]
 [-4.61219604e-16 -6.44449771e-16  1.00000000e+00  3.49546780e-15
   1.21430643e-16 -2.39391840e-16]
 [ 5.45323877e-15 -2.76820667e-14  3.49546780e-15  1.00000000e+00
   1.14968798e-14 -3.47812057e-15]
 [ 1.20996962e-15 -1.23512311e-15  1.21430643e-16  1.14968798e-14
   1.00000000e+00 -6.31439345e-16]
 [-1.28911638e-16  7.78890841e-16 -2.39391840e-16 -3.47812057e-15
  -6.31439345e-16  1.00000000e+00]]
```

So, it is proven that resulting variables are actually orthonormal.

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 function value = **0.8414429530201343**

ii. (5 points) Explain the meaning of the score function return value.
Score function defines mean accuracy for the given test data and labels. That mean it gives the how much fraction of test data are correctly classified.
So, Misclassified rate = 1 – score function
= 1 – **0.8414429530201343**
= 0.15855704697
So here misclassification rate is 15.85 %.

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*.
Input variable = [7500,15,3,127,2,2]
After Transformed **Input Variable** = [[-0.02886529, 0.00853837, -0.01333491, 0.0176811 , 0.00793805,0.0044727]]
Target =[[1,0]]
Five nearest neighbors: [[588 2897 1199 1246 886]]
Nearest neighbor with its case id and target value (Fraud Attribute) and other input variables:

	CASE_ID	FRAUD	TOTAL_SPEND	...	MEMBER_DURATION	OPTOM_PRESC	NUM_MEMBERS
588	589	1	7500	...	127	2	2
2897	2898	1	16000	...	146	3	2
1199	1200	1	10000	...	124	2	1
1246	1247	1	10200	...	119	2	3
886	887	1	8900	...	166	1	2

[5 rows x 8 columns]

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?

Based on the five nearest neighbor target value (FRAUD==1), we can see all the neighbors having target attribute = 1.
So, (5/5)*100 = 100.0 >19.9497 value derived in (ans 3.a). So, the observation is classified as fraudulent. So, Observation will not be misclassified.