CS 484: Introduction to Machine Learning

Spring 2021 Assignment 1

Question 1 (25 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) Use the Pandas describe() function to find out the count, the mean, the standard deviation, the minimum, the 25th percentile, the median, the 75th percentile, and the maximum.

"NormalSample.csv" file contains 1001 records. After using the pandas' describe() function the following information is found.

Item	Value
count	1001.000000
mean	31.414585
std	1.397672
min	26.300000
25%	30.400000
50%	31.500000
75%	32.400000
max	35.400000

b) (5 points) What is the bin width recommended by the Izenman (1991) method? Please round your answer to the nearest tenths (i.e., one decimal place).

From Q1.a, it is known that Q3(i.e., 75%) = 32.4 and Q1(i.e., 25%) = 30.4
IQR = Q3 – Q1
= 32.4 – 30.4
= 2

Bin-width, $h = 2(IQR)N^{-1/3}$
= 0.4 (rounded to one decimal place)

c) (10 points) Use the Shimazaki and Shinomoto (2007) method and try d = 0.1, 0.2, 0.5, 1.0, 2.0, and 5.0. What is the recommended bin width? You need to show your calculations to receive full credit.

To apply the Shimazaki and Shinomoto (2007) method first divide the data range into m bins of width d.

For this calcCD() function is written and from the function The following Data is found.

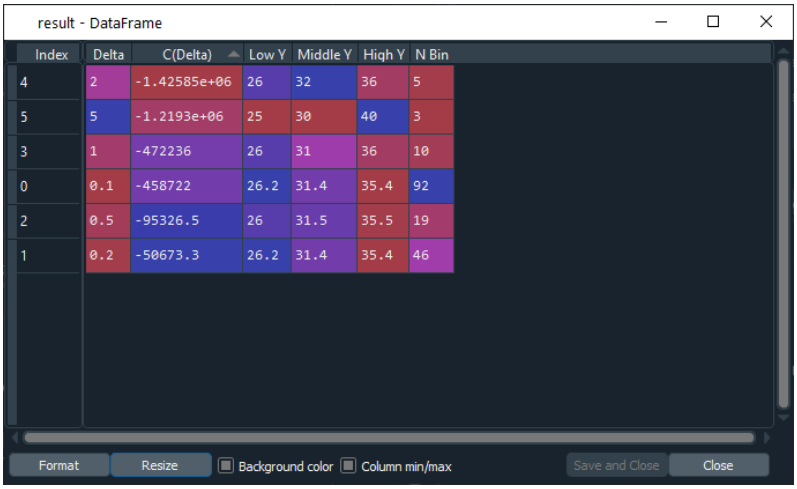

Figure 1

Data is sorted with ascending order of C(Delta). The calculation done using the following formula:

- Mean, $\bar{n} = \frac{1}{m} \sum_{i=1}^{m} n_i$
- Variance, $v = \frac{1}{m} \sum_{i=1}^{m} (n_i - \bar{n})^2$
- C(Delta), $C(d) = (2\bar{n} - v)/d^2$

Where,

- ✓ n_i → number of observations that enter the i^{th} bin.
- ✓ d → width of bin
- ✓ m → number of bins

Now, according to the Shimazaki and Shinomoto (2007) method , the optimal bin width as the d that minimizes $C(d)$. So, from the Figure 1 $C(d)$ is minimum for $d = 2$. So recommended bin width is 2.

The histogram for the data is given as follows:

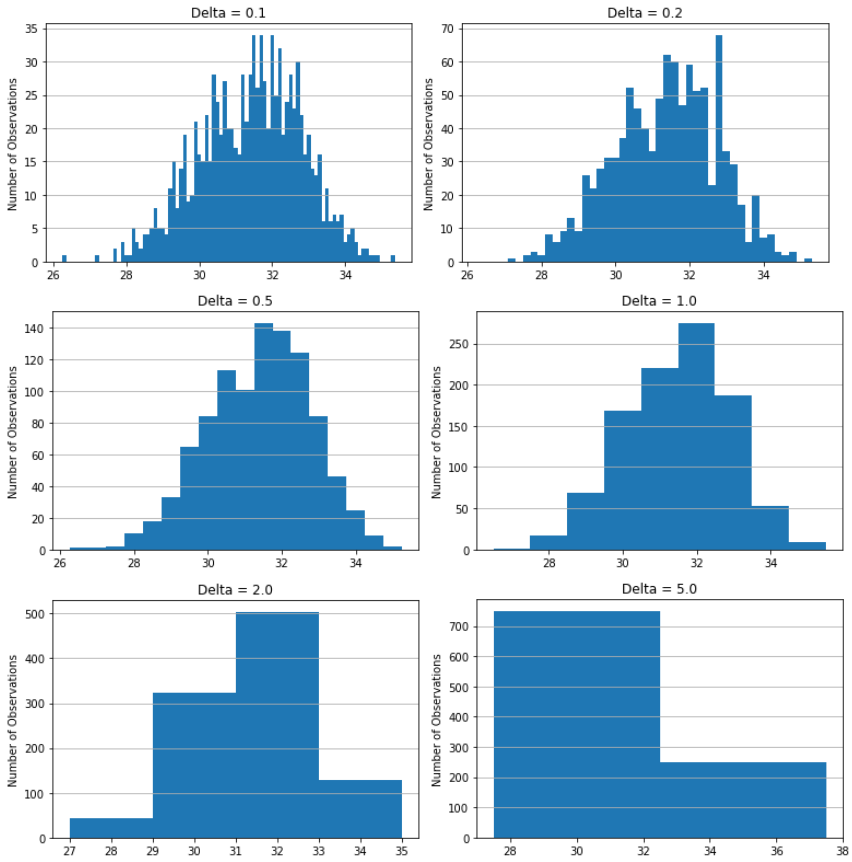

Figure 2

d) (5 points) Based on your recommended bin width answer in (c), list the mid-points and the estimated density function values. Draw the density estimator as a vertical bar chart using the matplotlib. You need to properly label the graph to receive full credit.

The recommended bin width is $d = 2$. The list of Mid-points and the Estimated Density Function Values are as below:

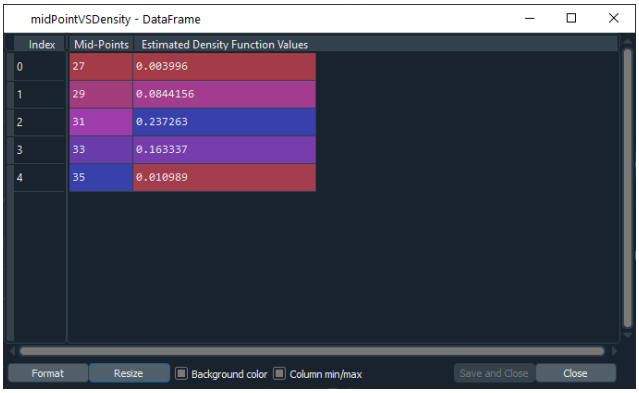

Figure 3

Vertical Bar Chart for Mid-points v/s Estimated Density Function Values is given below:

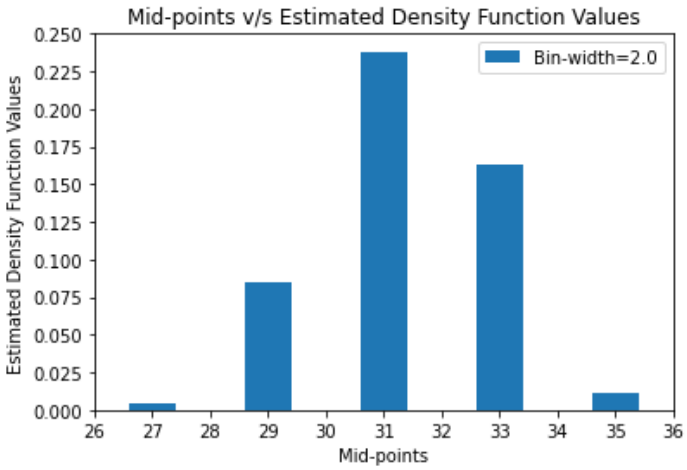

Figure 4

Question 2 (15 points)

The NormalSample.csv contains the *group* variable that has two values, namely, 0 and 1.

a) (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?

The below figure shows five-number summary of x for each category of the group:

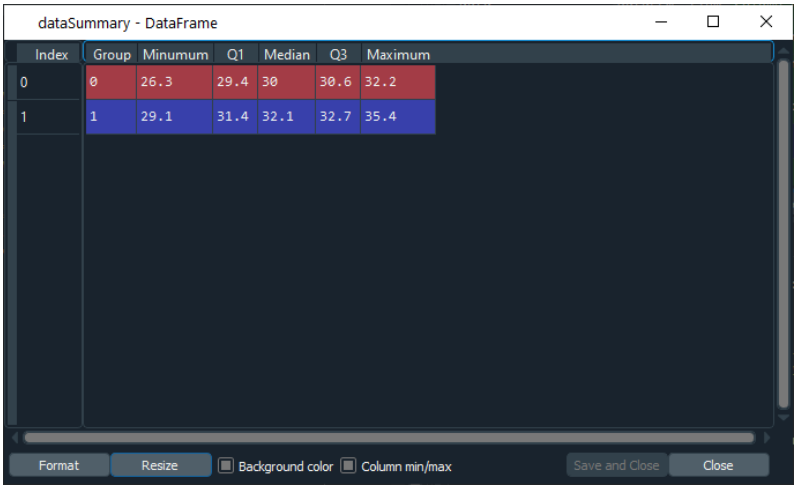

Figure 5

The values of the 1.5 IQR whiskers from each category of group are as follows:

- For group = 0:

$$IQR = Q3 - Q1 = 30.6 - 29.4 = 1.2$$
$$Lower\ Whisker = Q1 - 1.5 * IQR = 29.4 - 1.5 * 1.2 = 27.6$$
$$Upper\ Whisker = Q3 + 1.5 * IQR = 30.6 + 1.5 * 1.2 = 32.4$$

- For group = 1:
  $IQR = Q3 - Q1 = 32.7 - 31.4 = 1.3$
  $Lower\ Whisker = Q1 - 1.5 * IQR = 31.4 - 1.5 * 1.3 = 29.45$
  $Upper\ Whisker = Q3 + 1.5 * IQR = 32.7 + 1.5 * 1.3 = 34.65$

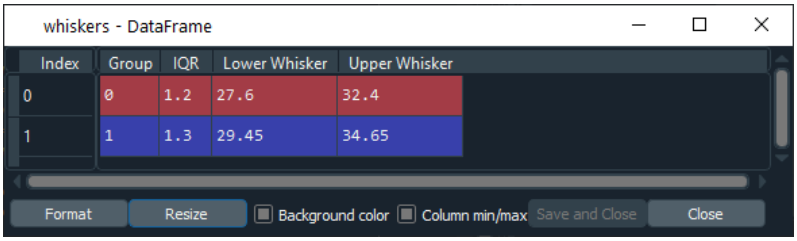

Figure 6

b) (10 points) Draw a graph where it contains the overall boxplot of x, the boxplot of x for each category of *group* (i.e., three horizontal boxplots within the same graph frame). Use the 1.5 IQR whiskers, identify any outliers of x for the entire data and each category of the group. You must properly label your boxplots to receive full credits. *Hint: Consider using the CONCAT function in the PANDA module to append observations*.

Graph for overall boxplot of X:

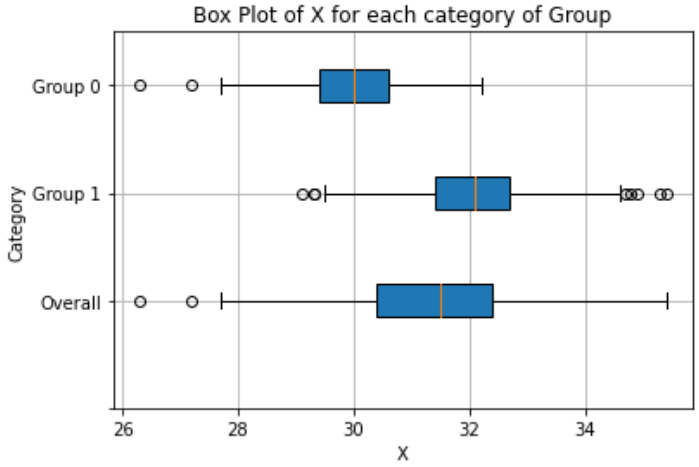

Figure 7

There are 2 outliers in the Overall. They are:

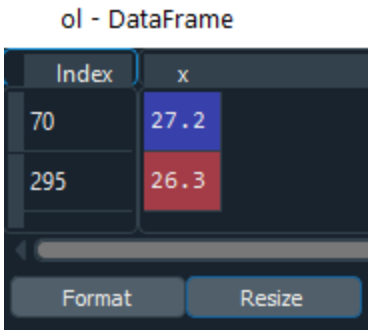

Figure 8

There are 8 outliers in the Group 1. They are:

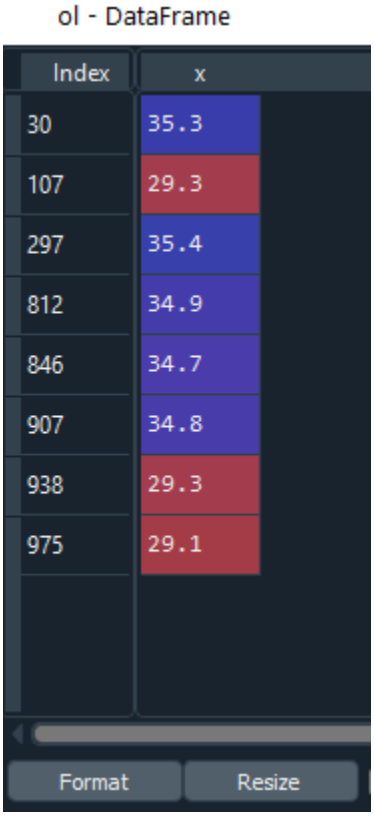

Figure 9

There are 2 outliers in the Group 0. They are:

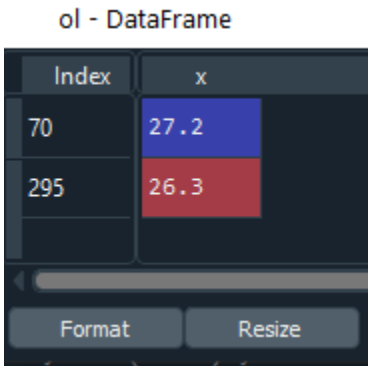

Figure 10

Question 3 (35 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 = Fraud, 0 = Not Fraud. 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 frauds? Please give your answer up to 4 decimal places.

Total Case: 5960
Fraud Case: 1189
Percentage: $1189/5960 * 100 = 19.9497\%$

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

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

As we have the eigen values greater than 1 as below:
Eigenvalues of x =
[6.84728061e+03 8.38798104e+03 1.80639631e+04 3.15839942e+05
8.44539131e+07 2.81233324e+12]

The number of dimensions for this condition is: 6 as the eigenvalues are greater than 1

ii. (5 points) Please provide the transformation matrix? Show evidence that the orthonormalized columns 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]]

Expecting an Identity Matrix, so the resulting variables are orthonormal (PROVED)=

[[1.00000000e+00 -6.30463563e-17 -1.70013743e-15 8.77897473e-15
  1.00245333e-15 -2.16948855e-16]
[-6.30463563e-17 1.00000000e+00 -5.66387215e-16 -1.91920048e-14
 -1.49880108e-15 -2.59341160e-16]
[-1.70013743e-15 -5.66387215e-16 1.00000000e+00 1.74296341e-15
  8.84708973e-17 1.21430643e-17]
[8.77897473e-15 -1.91920048e-14 1.74296341e-15 1.00000000e+00
  1.13216728e-14 -3.84154514e-15]
[1.00245333e-15 -1.49880108e-15 8.84708973e-17 1.13216728e-14
  1.00000000e+00 -6.31439345e-16]
[-2.16948855e-16 -2.59341160e-16 1.21430643e-17 -3.84154514e-15
 -6.31439345e-16 1.00000000e+00]]

c) (10 points) Use the NearestNeighbors module to execute the Nearest Neighbors algorithm using exactly five neighbors and the orthonormalized columns you have chosen in (b). The KNeighborsClassifier module has a score function.

i. (5 points) Find out from the documentation the purpose of the score function.

Return the mean accuracy on the given test data and labels. In multi-label classification, this is the subset accuracy which is a harsh metric since we require for each sample that each label set be correctly predicted. Score is basically how much cases given percentage how much percentage correctly classified w.r.t. to target. It also defines that how much accuracy can be achieve by defined model.

ii. (5 points) Run the score function, show and explain the function return value.

Score function values:0.8414429530201343

Misclassification $= (1 - scoreValue) * 100 = 15.855704697986573\%$
The score value represents mean accuracy of the given data set wrt target and how much data is correctly classified. Whereas misclassification rate is how much data is not correctly classified.

d) (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 (b) before finding the neighbors*.

The input focal point is given as [7500, 15, 3, 127, 2, 2]

Trasformed input
[[-0.02886529 0.00853837 -0.01333491 0.0176811 0.00793805 0.0044727]]

Neighbhr of target input values =
[[588 2897 1199 1246 886]]

Nearest Neighbhr of taget case id and value =

	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]

e) (5 points) Follow-up with (d), what is the predicted probability of fraud (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 a fraud. Otherwise, not a fraud. Based on this criterion, will the observation in (d) be misclassified?

For 5 nearest neighbor target is 1:Fraud and all neighbors is having the vale as 1([1 1 1 ... 1 1 1]) making it 100% which is greater than answer of 3.a. So we can say this observation cannot be misclassified, hence it is fraud.

Question 4 (25 points)

I found the following flights from Chicago O'Hare Airport (ORD) to Shanghai Pudong Airport (PVG).

Flight	Carrier 1	Carrier 2	Airport 1	Airport 2	Airport 3	Airport 4
A	American	Cathay Pacific	ORD	LAX	HKG	PVG
B	American	Cathay Pacific	ORD	SFO	HKG	PVG
C	American	China Southern	ORD	LAX	CAN	PVG
D	American	Virgin Atlantic	ORD	LHR		PVG
E	British Airways	Virgin Atlantic	ORD	LHR		PVG
F	Delta		ORD	SEA	ICN	PVG
G	United	Austrian	ORD	LHR	VIE	PVG
H	United	Virgin Atlantic	ORD	LHR		PVG
I	United		ORD	DCA	EWR	PVG
J	United		ORD	DEN	LAX	PVG
K	United		ORD	EWR		PVG
L	United		ORD	IAD	EWR	PVG
M	United		ORD	LAS	LAX	PVG
N	United		ORD	LAX		PVG
O	United		ORD	LGA	EWR	PVG

To answer the following questions, please replace empty string values in **Airport 3** with three underscore characters (i.e., '___').

a) (5 points) Generate a scatterplot of **Airport 3** (y-axis) versus **Airport 2** (x-axis). Please properly label the axes to receive full credits.

Scatterplot is given as:

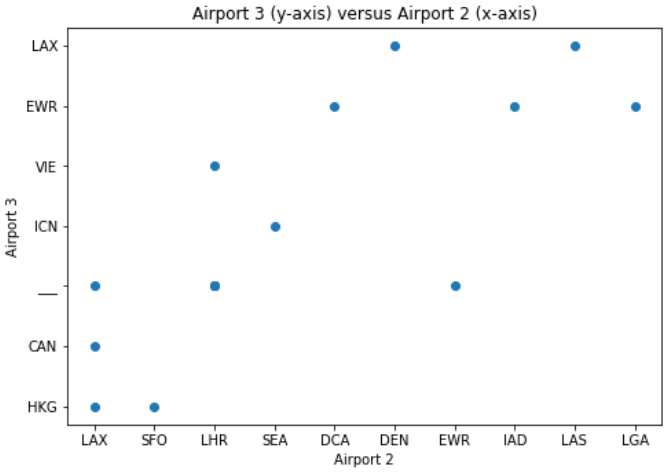

Figure 11

b) (5 points) Generate a frequency table of the airport codes in **Airport 2** and **Airport 3** combined.

The frequency table of the airport codes in Airport 2 and Airport 3 combined is given below:

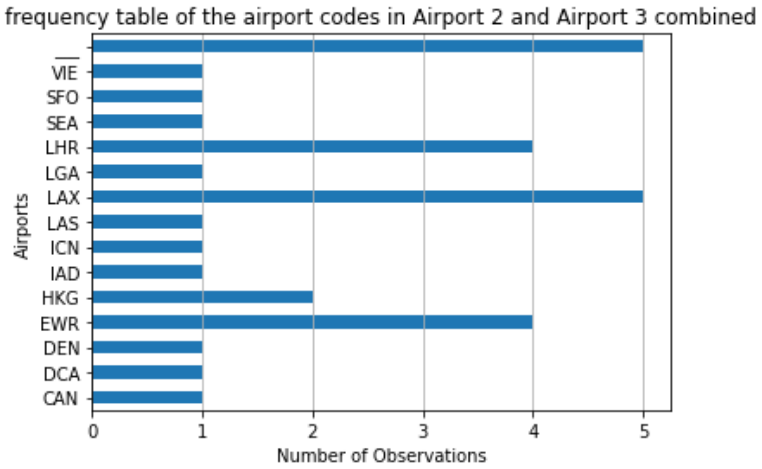

Figure 12

c) (10 points) Suppose a new airline creates a new flight from ORD to PVG that first stops at LAX and then ICN. I want to know which flight(s) most resembles this new flight. Use the Cosine Distance to measure the differences between this flight and the existing flights.

i. Create a vector of word counts for each flight. This vector has as many elements as the number of unique values found in (b).

X = Airport Frequency Data

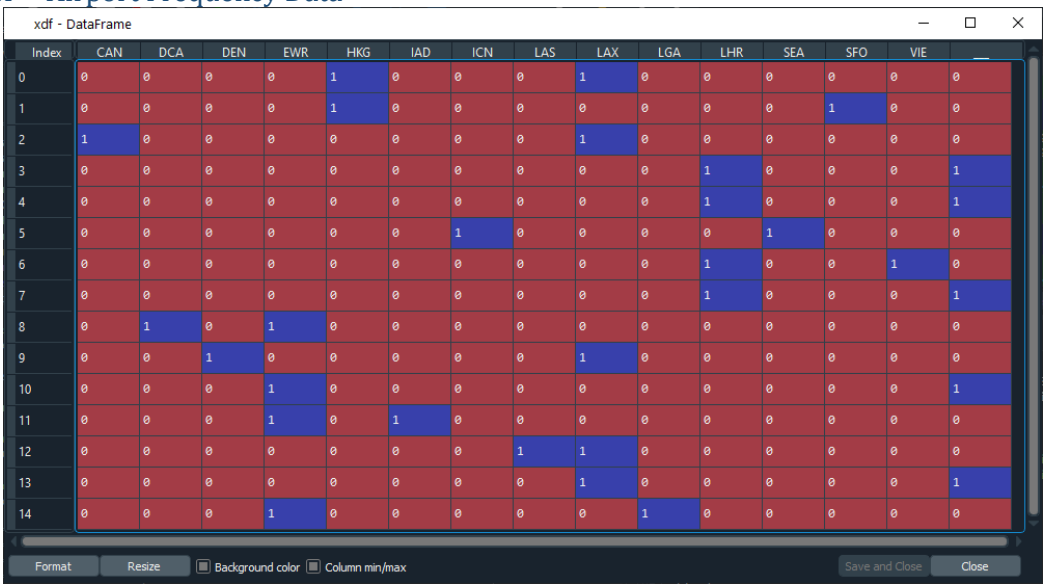

Figure 13

Airport_probeData:

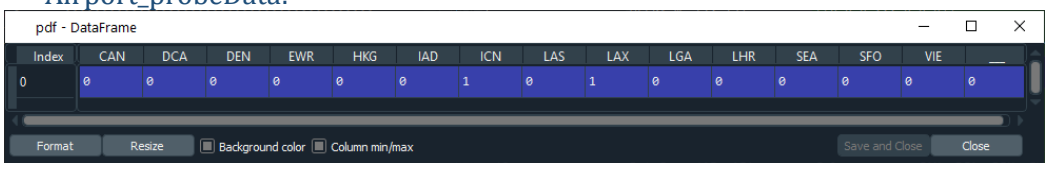

Figure 14

ii. Initialize all elements in the vector to zeros.
iii. Count the number of times the airport codes appeared in **Airport 2** and **Airport 3**.
iv. Calculate the Cosine Distance between the new flight and the Flights A to O.

You will list the Cosine Distances in a table.

Cosine Distance
[[0.5]
[1.]
[0.5]
[1.]
[1.]
[0.5]
[1.]
[1.]
[1.]
[0.5]
[1.]
[1.]
[0.5]
[0.5]
[1.]]

Figure 15

d) (5 points) Which flight(s) have the shortest Cosine Distance from the new flight?

Flights of shortest Cosine Distance from the new flight is of minimum cosine distance of 0.5 And the flights are:

Index	Flight	Cosine Distance
0	A	0.5
1	C	0.5
2	F	0.5
3	J	0.5
4	M	0.5
5	N	0.5