Aman Rangapur
A20517739

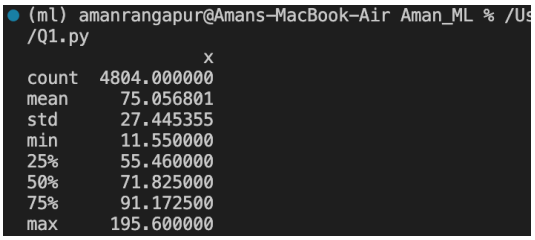
CS484: Introduction to Machine Learning Assignment-1

Question 1

a. What are the count, the mean, the standard deviation, the minimum, the 25th percentile, the median, the 75th percentile, and the maximum of the feature x? Please round your answers to the seventh decimal place.

Statistic	Value	Statistic	Value
count	4804.0000000	25%	55.4600000
mean	75.0568010	Median	71.8250000
Standard deviation	27.4453550	75%	91.1725000
Min	11.5500000	Max	195.6000000

Console Screenshot:



b. Use the Shimazaki and Shinomoto (2007) method to recommend a bin width. We will try d = 0.1, 0.2, 0.25, 0.5, 1, 2, 2.5, 5, 10, 20, 25, 50, and 100. What bin width would you recommend if we want the number of bins to be between 10 and 100 inclusively? You need to show your calculations to receive full credit.

Calculations: For delta = **0.1**
BinRight_n = Ceil value of ((Max – Middle) / delta) = math.ceil(195.6-75.1)/0.1 = 1205
BinLeft_n = Ceil value of ((Middle - Min) / delta) = math.ceil(75.1-11.55)/0.1 = 636
Total Number of Bins = BinRight_n + BinLeft_n = 1841
c_delta = (2*mean_count - ssd_count) / delta / delta = (2*4.1342 - 8.4088)/0.1/0.1 = -14.032

The following table shows the results of trying these thirteen values of delta.

Delta	C Delta	No. of bins	Solution
0.1	-14.0324	1841	No
0.2	-373.8800	921	No
0.25	-448.2728	737	No
0.5	-576.4073	369	No
1	-637.2622	185	No

Delta	C Delta	No. of bins	Solution
2	-657.8489	93	Yes
2.5	-661.0811	75	Yes
5	-667.2308	38	Yes
10	-662.0939	19	Yes
20	-630.1331	10	Yes
25	-606.6692	8	No
50	-514.4502	4	No
100	-236.3717	2	No

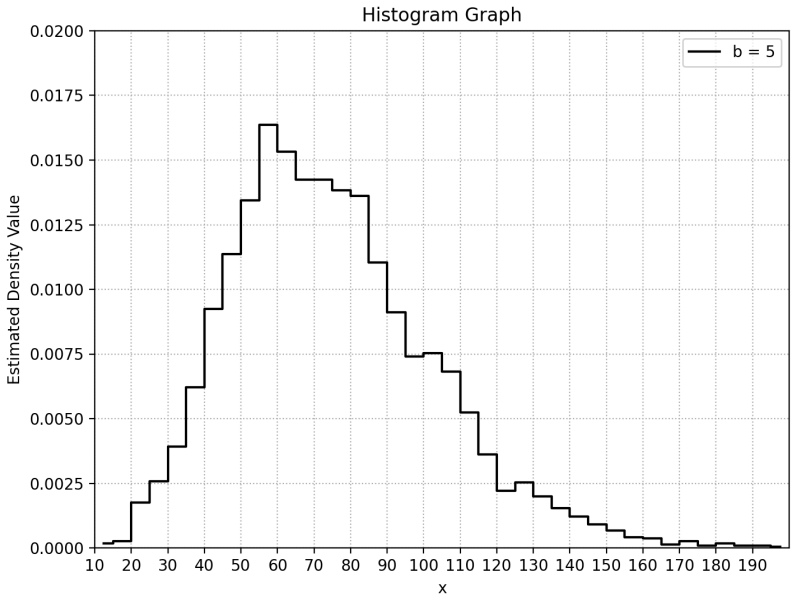
Among the thirteen possible bin-widths, only five solutions are admissible according to the requirement on the number of bins. Among the four admissible solutions, the one where d = 5 yields the lowest $C(d)$ value. Therefore, the recommended bin-width is 5.

Console Screenshot:

	Delta	Criterion Delta	No. of Bins
0	0.10	-14.032427	1841.0
1	0.20	-373.880076	921.0
2	0.25	-448.272884	737.0
3	0.50	-576.407374	369.0
4	1.00	-637.262281	185.0
5	2.00	-657.848992	93.0
6	2.50	-661.081145	75.0
7	5.00	-667.230803	38.0
8	10.00	-662.093961	19.0
9	20.00	-630.133100	10.0
10	25.00	-606.669200	8.0
11	50.00	-514.450200	4.0
12	100.00	-236.371700	2.0

c. Draw the density estimator using your recommended bin width answer in (b). You need to label the graph elements properly to receive full credit.

The right screenshot shows the mid points and density function values of bin width 5 with a total of 38 bins. The histogram is shown below.



	MidPoint	Density
0	12.5	0.000167
1	17.5	0.000250
2	22.5	0.001749
3	27.5	0.002581
4	32.5	0.003913
5	37.5	0.006203
6	42.5	0.009242
7	47.5	0.011366
8	52.5	0.013447
9	57.5	0.016361
10	62.5	0.015321
11	67.5	0.014238
12	72.5	0.014238
13	77.5	0.013822
14	82.5	0.013614
15	87.5	0.011032
16	92.5	0.009117
17	97.5	0.007410
18	102.5	0.007535
19	107.5	0.006828
20	112.5	0.005246
21	117.5	0.003622
22	122.5	0.002206
23	127.5	0.002540
24	132.5	0.001998
25	137.5	0.001540
26	142.5	0.001207
27	147.5	0.000916
28	152.5	0.000666
29	157.5	0.000416
30	162.5	0.000375
31	167.5	0.000125
32	172.5	0.000250
33	177.5	0.000083
34	182.5	0.000167
35	187.5	0.000083
36	192.5	0.000083
37	197.5	0.000042

Question 2.

We need to create the Training and Testing partitions from the observations in the hmeq.csv. We will use all observations (including those with missing values in one or more variables) for this task. The Training partition will contain 70% of the observations. The Testing partition will contain the remaining 30% of the observations. We initialize the random seed with the integer 20230101.

a. Before we partition the observations, we need a baseline for reference. How many observations are in the dataset? What are the frequency distributions of BAD (including missing)? What are the means and the standard deviations of DEBTINC, LOAN, MORTDUE, and VALUE?

Number of observations in the dataset are 5960. Frequency distribution for BAD is shown below:

Frequency Distribution	0	1
BAD	4771	1189

Mean and standard deviations are shows in below table.

	Mean	Std
DEBTINC	33.77	8.60
LOAN	18607.96	11207.48
MORTDUE	73760.81	44457.61
VALUE	101776.04	57385.76

b. We first try the simple random sampling method. How many observations (including those with missing values in at least one variable) are in each partition? What are the frequency distributions of BAD (including missing) in each partition? What are the means and the standard deviations of DEBTINC, LOAN, MORTDUE, and VALUE in each partition?

Observations in Train Partition: 4172

Frequency Distribution	0	1
BAD	3344	828

Mean and standard deviations in train partition are shows in below table.

	Mean	Std
DEBTINC	33.77	8.44
LOAN	18609.41	11300.34
MORTDUE	74067.99	44640.12
VALUE	101716.90	56671.25

Observations in Test Partition: 1788

Frequency Distribution	0	1
BAD	1427	361

Mean and standard deviations in test partition are shows in below table.

	Mean	Std
DEBTINC	33.80	8.96
LOAN	18604.58	10990.88
MORTDUE	73055.47	44040.99
VALUE	101912.31	59015.12

c. We next try the stratified random sampling method. We use BAD and REASON to jointly define the strata. Since the strata variables may contain missing values, we will replace the missing values in BAD with the integer 99 and in REASON with the string 'MISSING'. What are the frequency distributions of BAD (including missing) in each partition? What are the means and the standard deviations of DEBTINC, LOAN, MORTDUE, and VALUE in each partition?

Frequency Distribution of REASON after replacing missing values with "MISSING"

Frequency Distribution	DebtCon	HomeImp	MISSING
REASON	3928	1780	252

Observations in Train Partition: 4172.

The below table shows the frequency distribution of BAD in test partition.

Frequency Distribution	0	1
BAD	3340	832

Mean and standard deviations of train partition are shows in below table.

	Mean	Std
DEBTINC	33.82	8.37
LOAN	18493.09	11021.75
MORTDUE	73241.51	44516.04
VALUE	101274.82	56367.52

Observations in Test Partition: 1788.

Frequency Distribution	0	1
BAD	1431	357

Mean and standard deviations in test partition are shown in below table.

	Mean	Std
DEBTINC	33.66	9.12
LOAN	18876.00	11628.04
MORTDUE	74976.34	44310.37
VALUE	102949.76	59702.35

Question 3

a. What percent of investigations are found to be frauds? This is the empirical fraud rate. Please round your answers to the fourth decimal place.

FRAUD = 0 in 4,771 observations and FRAUD = 1 in 1,189 observations. Therefore, the percent of fraud is 1189 / (4771 + 1189) = 0.199497 = 19.9497%.

b. We will divide the complete observations into 80% Training and 20% Testing partitions. A complete observation does not contain missing values in any of the variables. The random seed is 20230225. The stratum variable is FRAUD. How many observations are in each partition?

Observations in train partition: 4768. Observations in test partition: 1192.

c. Use the KNeighborsClassifier module to train the Nearest Neighbors algorithm. We will try the number of neighbors from 2 to 7 inclusively. We will classify an observation as a fraud if the proportion of FRAUD = 1 among its neighbors is greater than or equal to the empirical fraud rate (rounded to the fourth decimal place). What are the misclassification rates of these numbers of neighbors in each partition?

The below table shows the observations for k = [2,7].

d. Which number of neighbors will yield the lowest misclassification rate in the Testing partition? In the case of ties, choose the smallest number of neighbors.

I observed lowest misclassification rate for k=6 in test partition. Test misclassification is 18.0369.

k	Train Accuracy	Train Misclassification	Test Accuracy	Test Misclassification
2	86.0529	13.9471	80.7886	19.2114
3	85.7802	14.2198	79.4463	20.5537
4	83.5361	16.4639	80.9564	19.0436
5	83.599	16.401	80.3691	19.6309
6	82.9279	17.0721	81.9631	18.0369
7	82.9908	17.0092	81.3758	18.6242

e. Consider this focal observation where DOCTOR_VISITS is 8, MEMBER_DURATION is 178, NUM_CLAIMS is 0, NUM_MEMBERS is 2, OPTOM_PRESC is 1, and TOTAL_SPEND is 16300. Use your selected model from Part (d) and find its neighbors. What are the neighbors' observation values? Also, calculate the predicted probability that this observation is a fraud.

The focal observation has TOTAL_SPEND = 16300, DOCTOR_VISITS = 8, NUM_CLAIMS = 0, MEMBER_DURATION = 178, OPTOM_PRESC = 1, and NUM_MEMBERS = 2. The six neighbors have indices 2679, 3046, 3201, 2426, 3282, 2416. Therefore, the six neighbors are

INDEX	TOTAL_SPEND	DOCTOR_VISITS	NUM_CLAIMS	MEMBER_DURATION	OPTOM_PRESC	NUM_MEMBERS	FRAUD
2679	15200	8	0	180	1	2	0
3046	16600	9	0	180	1	2	0
3201	17200	9	0	180	1	2	0
2426	14600	8	0	180	1	2	0
3282	17500	7	0	180	1	2	0
2416	14500	7	0	180	1	2	0

Since the FRAUD values of all six neighbors are 0, the predicted probability of fraud of the focal observation is also 0 and, therefore, the predicted FRAUD is 0. The predicted probability of fraud is 0.