Dr. Eick

COSC 6335*“Data Mining”*

ProblemSet1 Fall 2022

*Task 2: Creating Good Histograms*

*Fifth Draft*

Deadlines; **We., October 05**, 2022, 11.59 PM

Last updated: September 28, 2022, 8.30 PM

**Task2 Submission Guideline**

1. All submissions will be on teams
2. Each person need to submit: a program file, a report and a readme.
3. The program file will be a function with clear instruction how to provide inputs and what the outputs will be. We will use the instruction run your program.
4. The module should include all the dependencies needed to run the file, inputs and outputs.
5. Instructions for the report (for more details see file Task2.pptx near the project specification in the course webpage:
	1. It will be three to five pages
	2. Will describe the algorithm used to generate the results
	3. If any further algorithm used to compare or generate the results, that will be explained properly
	4. Interpretation of the results
	5. An appendix with results for all three datasets (will not count towards page limits).

*Task2 Specification*

Your task which is called GHISTO is to create a good 7-equi-bin or 10-equi-bin histogram for a Dataset D consisting of 30 numerical observations which can include duplicates. The histogram depicts the percentage of observations that fall into the 7 or 10 bins of your solution. However, you are allowed to remove high or low outliers from the dataset to obtain a better histogram. Basically, to create a histogram you need to select the following:

1. Which observations are removed from D because they are high or low outliers; we call the subset of D which is obtained after outlier removal D’
2. A bin origin—the left boundary of the left most bin
3. A bin-size
4. Number of bins (7 or 10)

Next, you can compute the percentages for each bin by counting what percentage of the values in D’ belong to a particular bin; e.g. [a, b]. Moreover, each value belonging to D must be associated with a bin and cannot be outside the bin boundary.

The output of GHISTO should have the following format:

<number of bins>

<origin>

<bin-size>

<bin\_1><fraction of values>

….

<bin\_n> <fraction of values>

Outliers: <list outliers>

Where bins are described by the bin boundaries and the percentage of observation

For example for an input dataset D2=which consisting of integers in {35,74,75,76,76,76,85,87,87,88,88,92,95,96,97,100,105,106,107,108,109,110,111,111,115,115,116,116,175,180}

The output of your program could be as follows:

|  |  |  |
| --- | --- | --- |
|  | **Histogram without Outliers[[1]](#footnote-1)** | **Histogram Using Outliers** |
| **Histogram** | C:\Users\mdmah\PycharmProjects\ProfessorEick\DataMiningFall2022\Outputs\Histograms\Histograms\Histogram of D2 with outliers _10 bins.png | C:\Users\mdmah\PycharmProjects\ProfessorEick\DataMiningFall2022\Outputs\Histograms\Histogram of D2 without outliers _7 bins.png |
| **Outputs** | 103514.51[35, 49.51) 1/30[49.5, 64.2) 0/30[64.2, 78.53) 5/30[78.53, 93.04) 6/30[93.04, 107.55) 7/30[107.55, 122.6) 9/30[122.06,136.57) 0/30[136.57,151.08) 0/30[151.08,165.59) 0/30[165.59,180.1) 2/30D2=(35, 74, 75, 76, 76, 76, 85, 87, 87, 88, 88, 92, 95, 96, 97, 100, 105, 106, 107, 108, 109, 110, 111, 111, 115, 115, 116, 116, 175, 180) | 7746.1[74, 80.1) 5/27[80.1, 86.2) 1/27[86.2, 92.3) 5/27[92.3, 98.4) 3/27[98.4, 104.5) 1/27[104.5, 110.6) 6/27[110.6, 116.7) 6/27Outliers: 35, 175, 180D2=(35, 74, 75, 76, 76, 76, 85, 87, 87, 88, 88, 92, 95, 96, 97, 100, 105, 106, 107, 108, 109, 110, 111, 111, 115, 115, 116, 116, 175, 180) |
| **Density Function** | (27.745 0) (42.255 0.033) (56.265 0) (71.275 0.17) (85.785 2) (100.295 0.23) (114.805 0.3) (129.315 0) (143.825 0) (158.335 0) (172.875 0) (187.355 0.67) | (70.95 0) (77.05 0.19) (83.15 0.04) (89.25 0.19) (95.35 0.11) (101.45 0.04) (107.55 0.22) (113.65 0.22) (119.75 0)  |

The output histogram with outliers has a bin origin of 35 and bin-width of 16.1 and the output histogram without outliers has a bin origin of 74 and bin-width of 6 following bag of outliers: {35, 175, 180}

Additionally, you program should output the following unnormalized density function with following format,

<center of bean 1><density of bean 1>, <center of bean 2><density of bean 2>,,,…,,<center of bean n><density of bean n>

This density function will be compared with a ground truth density function of the dataset to evaluate the quality of the histogram your program GHISTO produced. More details how this evaluation is done will added in the next version of the Task2 specification which should be available by September 10 the latest.

Moreover, your program will be tested for input datasets D1, D2, D3, D4, and D5 and where D1 D2, D3 look as follows:

D1=(10,15,18,51,51,52,52,54,54,54,56,58,58,60,60,61,61,62,63,63,64,65,65,67,67,68,68,69,73,130)

D2=(35,74,75,76,76,76,85,87,87,88,88,92,95,96,97,100,105,106,107,108,109,110,111,111,115,115,116,116,175,180)

D3= (20, 36,38,39,40,41,42,44,48,51,52,54,55,56,58,63,65,66,67,68,69,73, 78,81,84,85,87,88,92, 105)

Moreover, you program will be tested by two more input datasets D4 and D5 which are not know to you during program development time.

1. Actually, you are supposed to produce a 10-bin or 7-bin histogram, and not a 9-bin histogram but the main purpose of this discussion is what output your program should produce. [↑](#footnote-ref-1)