Data Visualization HW1

Hongsup Oh

February 2020

1 Part 1

1. problem 1

   (a) Create an array with 200 elements from 1 to 200 in order
   arr = [1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200]

   (b) Create a box plot to visualize this array of 200 elements

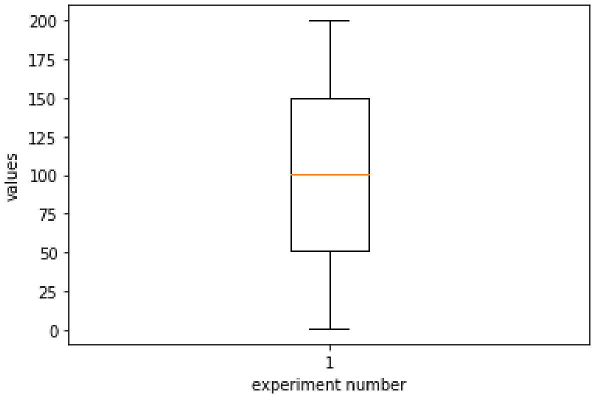

Figure 1: Box Plot

2. problem 2.

   (a) Create an array x with 10,000 floats in the range [1,10].
   x = [9.11800697 7.67144927 6.96370485 ... 8.6995738 6.22030777 1.0190432]

   (b) Plot, for the array x, a histogram showing a uniform distribution of bin sizes.

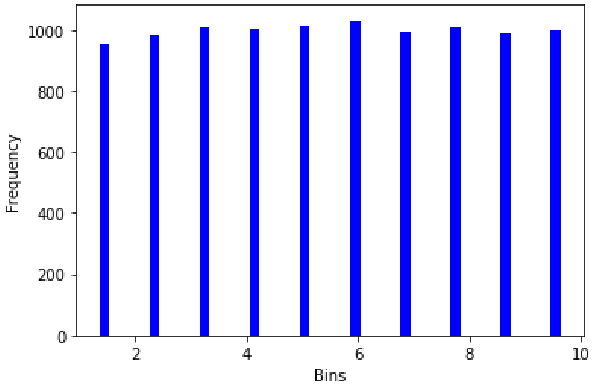

Figure 2: histogram for uniform distribution

   (c) For the same array x, plot a histogram showing a monotonically increasing distribution of bin sizes.

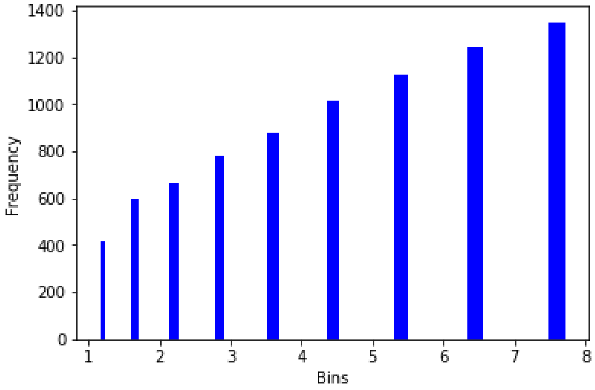

Figure 3: histogram for monotonic increase

(d) For the same array x, plot a histogram showing a monotonically decreasing distribution of bin sizes

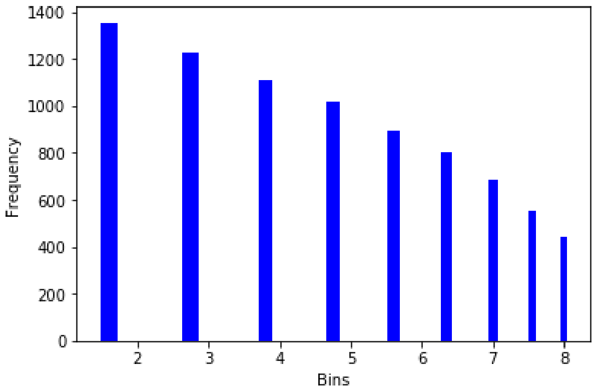

Figure 4: histogram for monotonic increase

(e) Which of the histograms you created before can be considered a misleading visualizations, and why?
The histograms of (c) and (d) could be considered a misleading visualization. Since histogram of (c) and (d) are monotonic increasing and decreasing, it looks like x array made at part (a) can be considered non uniformed distributed array. For example, in the case of (c), larger numbers' distribution in array x looks like larger than that of small numbers.

3. problem 3.

(a) Create an array by generating 100 random numbers with normal (aka Gaussian) distribution.
x = [-1.29852548 0.17421459 1.62592145 0.74494864 0.12122266 0.00551402
2.03460393 3.10755519 0.54939308 -0.42049008 0.27971693 -1.16756766
0.32026703 -0.37516533 -0.25943104 -1.85371692 -0.9699492 -0.20694932
-0.33059412 1.88606807 -0.13060656 0.75864859 -0.64850143 -0.99126438
-0.87755527 -0.24908185 -1.06358645 -1.50114347 0.96024504 1.92183763
0.61762413 -1.35069648 -1.49420104 -2.4491237 -1.72960236 0.92023467
-1.09314846 1.41175278 -0.38537055 0.10679591 -0.01934057 0.79196378
0.79342477 -0.77246093 -1.34839539 -0.17399014 -0.19984607 0.91434379
0.54896194 -2.28464439 -1.30166642 0.62815881 0.87304977 1.47110113
0.0839767 -0.87344343 0.85027773 -0.13111132 -0.42976318 -1.17406441
0.72227636 0.2680992 0.16871371 0.98795262 0.03192119 -1.38355898
0.66653011 -0.16853815 1.51342611 -0.06476895 -0.30581175 0.30119017
-1.29221054 0.49888526 1.19240664 -0.65093927 -1.44499119 -0.77878429
0.38101964 -1.19282633 -0.44132625 0.29231691 -0.80810246 -1.07896116
1.32329095 0.26961839 -0.68402199 -2.59070183 -1.60856375 0.38689839

-0.70239757 -0.31432652 -0.40469997 -0.29444783 1.4818246 -1.24182288
1.0619542 0.58157742 -0.34487632 1.18578477]

(b) Write the numbers out to a binary file (using numpy). Read the binary file back into an array (using numpy).
bnum = [-1.98869644 -1.03118327 -0.0367118 0.82740767 -1.51419995
-0.72468518 0.41074809 -0.15855966 0.4975847 -2.1554405 -0.56031343
0.01948672 -0.3021431 -0.05639669 1.21485755 -0.37976124 0.67963786
-0.49716904 0.71196702 -1.10032969 0.40092525 -0.6339815 -1.55290376
0.1697776 3.62552707 0.38658025 0.23660326 -2.45429937 -1.23000175
0.41270739 -0.59549653 -1.07400213 1.63950366 0.09803048 -0.29996035
0.95004331 2.12670825 -1.86746705 2.20952292 0.95853336 -1.60362796
1.06003771 0.3499571 0.16852299 0.98900188 -0.95612295 0.54976627
-0.72663862 0.42352858 -1.4801347 -0.03333396 0.27847037 1.3847852
0.1282461 0.98038607 -0.19863935 -0.21165918 0.68149992 1.60178753
-0.42234198 -0.25422639 -0.06937915 -1.47614908 0.67630508 0.42725092
-1.33031208 0.3919864 -1.49604401 0.13383753 0.96270201 -1.55823503
0.37725757 0.19718604 0.20609721 1.66048327 -1.34468452 0.43461508
1.20655496 -0.95186921 -1.0374365 -0.1604798 0.87691912 -0.3527396
-1.14286511 -0.25205102 0.27245213 1.29449425 1.54959846 -0.00468624
-0.83139156 -0.66691728 -0.60372889 -1.13853244 0.56419383 2.10427439
-0.2243308 -0.38522358 -0.59219562 -0.90533196 0.55760478]

(c) Create a histogram and a bar chart of the data you read back from the binary file.

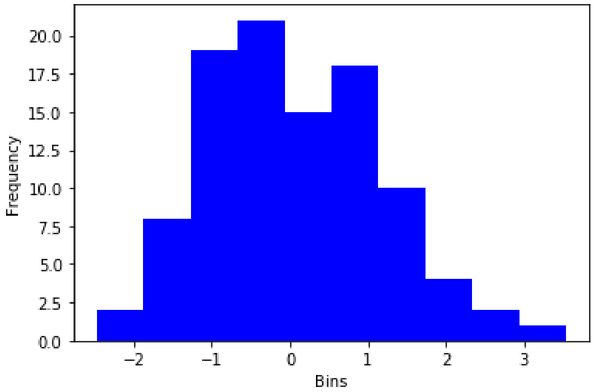

Figure 5: histogram

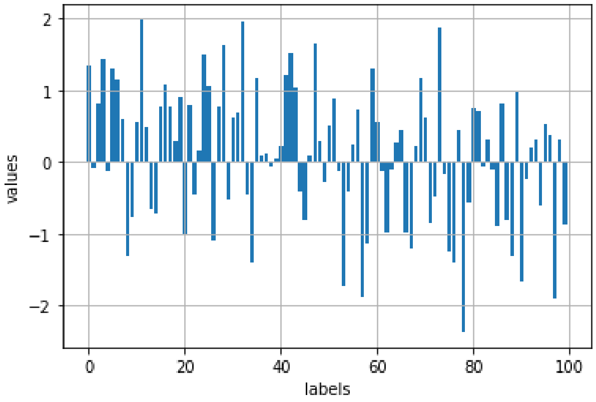

Figure 6: barchart

(d) What information does the histogram shows that the bar chart does not?
Histogram shows the distribution of numerical data. It is probability distribution of continuous variable.

(e) What information does the bar chart shows that the histogram does not?
It is used for non continuous variable. It shows frequency of each label. In the problem, indexes of array are considered as labels and elements of array are considered as frequency.

(f) Opinion
This problem has continuous data. Thus histogram would be better method in the problem.

2 Part 2

1. problem 1

   (a) Download the NOAA Land Ocean Temperature Anomalies Data Set and Load Data.
   Embedded open function and readline function are used and separate year and temperature data to yr array and value array respectively.

(b) Create a Scatter Plot and a Bar Plot. Include a label called "Year" along the x-axis and a label called "Degrees Celsius +/- From Average" along the y-axis.

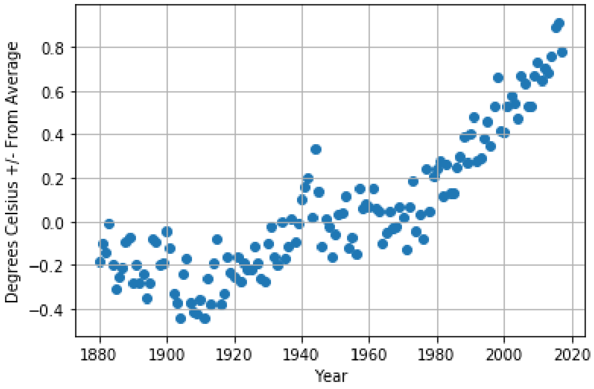

Figure 7: Scatter plot

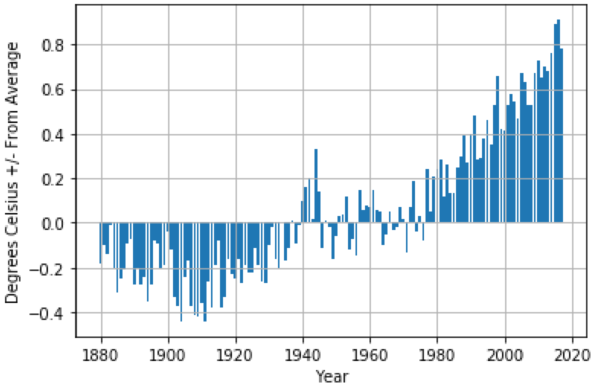

Figure 8: Bar plot

(c) Describe the trends you see in the data.
There are two trends. First, temperature increases as time passes. Second before 1950s, trend is minus degrees and after 1950s, trend is plus degrees

(d) Discuss which plot you believe shows those trends better and why.
Both plots show well for increasing temperature trend. But Bar plot is better to show where is minus degree and plus degree. Thus, I think bar plot is better to shows those trends.

(e) Provide an example of a different plot which better represents the long-term trends.
Moving average (or rolling average) would be good to show the long-term trends. If data would be separated by month or days from 1880 to 2020, line plot would be much more fluctuated. So it is difficult to find trend of graph. In order to avoid the situation, moving average method can remove the fluctuation. Made function named movingAvg is used for moving average.

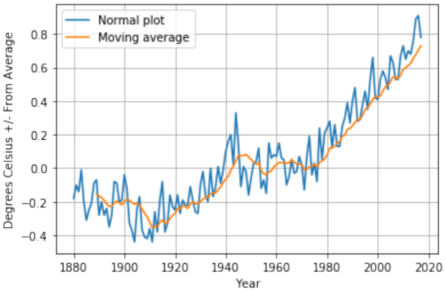

Figure 9: moving average

2. problem 2.

(a) Download the statistical data about marriage from and Read carefully the description of the dataset on the webpage. Load the data. labelsCandi array has name of labels for data. And data2 hash map has each year's data as a array form.

(b) Create a Star Plot and a Line Graph using at least three fields.

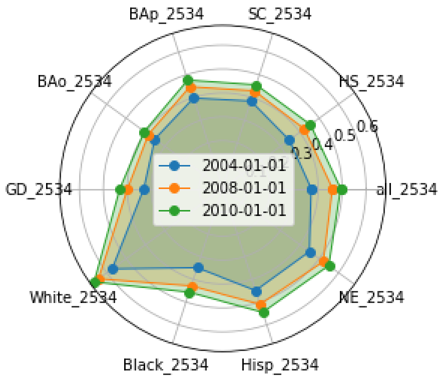

Figure 10: Star Plot

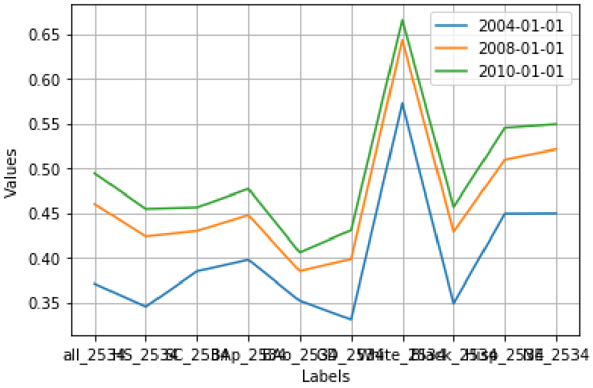

Figure 11: Line Plot

(c) Describe the trends you see in the plots for the three different fields. White2534 shows the highest value in three years data. In addition, all labels' values increase from 2004 to 2010 at graph.

(d) What are some pros and cons of using a Star Plot vs a Line Graph? For the star plot, it is clear to see which is the highest and lowest values. In addition, it is easy to check each label's value. However, it is not easy to see clear trend. For the line plot, it is clear to see the trend of graph, if the graph has trend. But it is not easy to check each label's value compared with star plot.

3. problem 3.

(a) Download the U.S. Birth data set and Load the data. And Create visualizations to support your answers for the following questions:

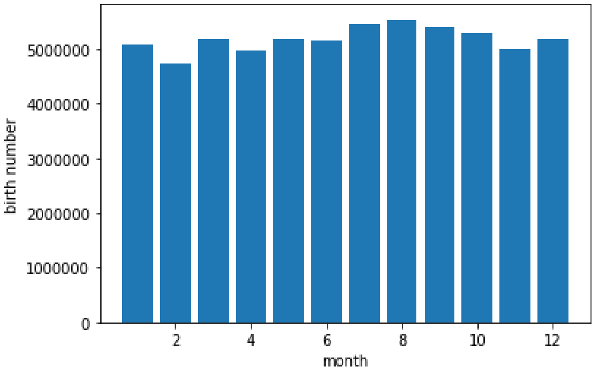

Figure 12: Bar Plot

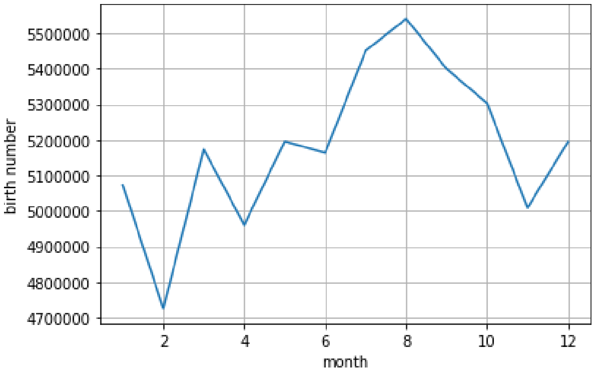

Figure 13: Line Plot

(b) What month had the highest number of births?
It is August (8)

(c) What month had the lowest number of births?
It is February (2)

(d) Are there any interesting trends in the data?
From winter to summer season(FEB to AUG), the number of birth increase and it decreases from summer to winter season(AUG to DEC).

4. problem 4.
I choose fifa countries audience csv file.

(a) Produce three good visualizations which convey a unique trend in the data. Discuss the trends you see briefly

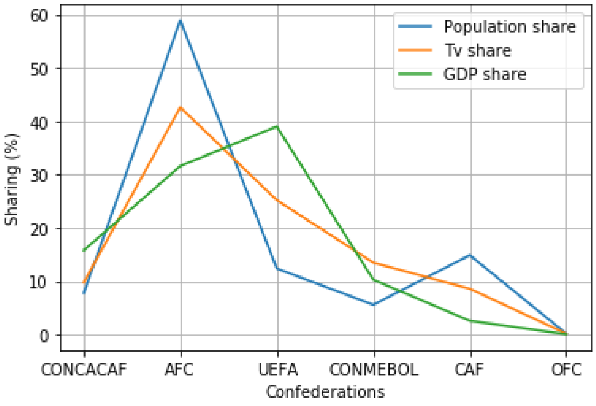

Figure 14: Line Plot

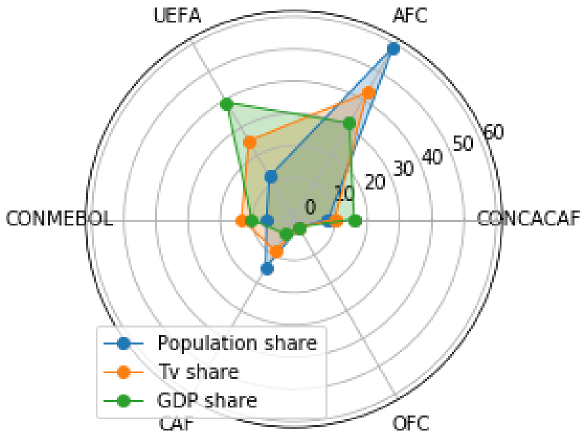

Figure 15: Star Plot

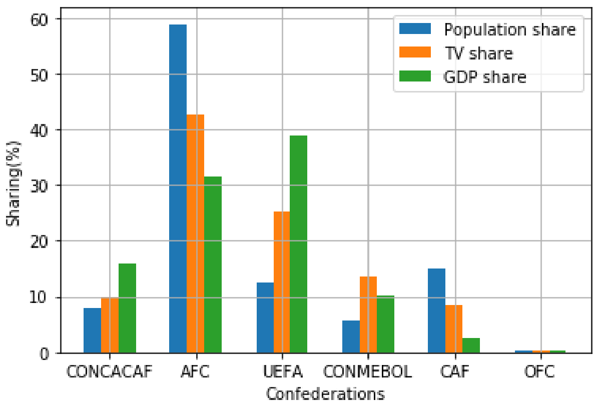

Figure 16: Bar Plot

Since data have too many nations, confederations are used instead of nations. Hash maps are used to calculated each confederation's sharing. Among three trends such as population, TV, and GDP sharing, AFC shows the highest sharing at population and TV and UEFA shows the highest sharing at GDP. OFC shows the lowest sharing at all threes.

(b) Produce three bad visualizations and explain briefly why these visualizations are not suitable for the dataset you picked.

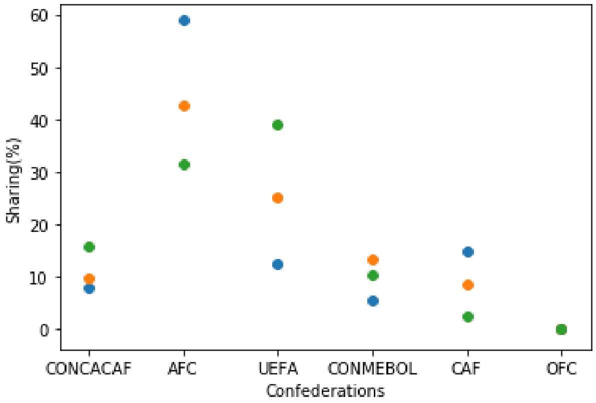

Figure 17: Scatter Plot

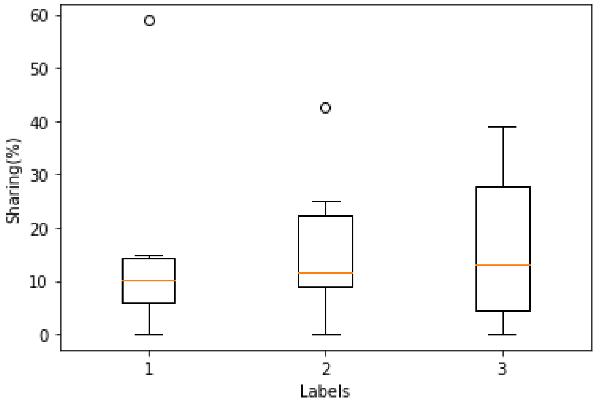

Figure 18: Box Plot

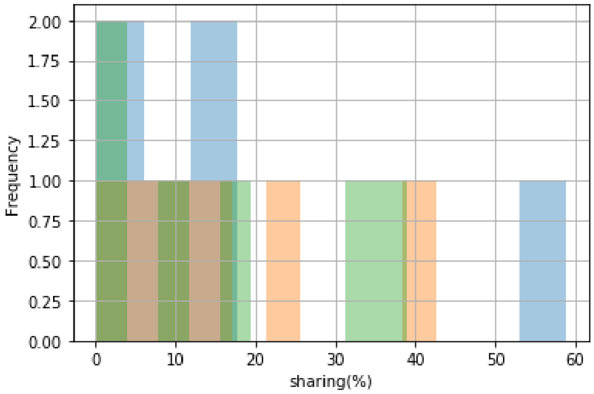

Figure 19: Histogram

First, scatter plot (Figure 17) is not fit for the data. It is not easy to see trend of graph, due to small data. In addition, data is just differentiated by color, so it is not easy to see the difference of each data. Second, box plot (Figure 18) is not fit for the data. It just show the range of each data such as the lowest and highest and average point. But it is impossible to see which confederation has more sharing. Third, histogram (Figure 19) is the worst method for the data, because the data is not continuous data. Instead, it is labelled by different confederation.

3 Part 3

1. problem 1

   (a) Why is assessing value of visualizations important?
   (ANS): Now a day, many new methods, techniques, and systems have been developed. However, many of these new methods are not used in real-world situations. In addition, judgement of the value of visualization is in varying senses. Thus, in order to make a good choice, assessing value of visualizations is important.

   (b) What are the two measures for deciding the value of visualizations?
   (ANS): effectiveness and efficiency

2. problem 2

   (a) Briefly describe the mathematical model for the visualization block shown in Fig.1.
   First, V process use Data D which is transformed according to a specification S, and make time varing image output I(t).
   Second, p (perception) process receive image I with an increase in

knowledge K.
Third, the user decide to adpat the specification of the visualization.

3. problem 3

   (a) State four parameters that describe the costs associated with any visualization technique.
   (ANS): Initial development costs, Initial costs per user, Initial costs per session, and perception and exploration costs.

4. problem 4

   (a) What are the pros and cons of interactivity of visualizations?
   Pros: Enhances the understanding of the data.
   Cons: First, allowing the user to modify S freely will lead to subjectiveness. Second, interaction is costly, and lead to high perception and exploration costs

4 Part 4

1. problem 1.

   (a) Download the brain MRI dataset "T2.raw" from canvas and load it into python. The data format is float32 with dimensions 320 x 320 x 256
   original shape: (26214400,)
   changed shape: (320, 320, 256)

   (b) Extract one slice from the volume and save it as a PNG image.
   It is saved as myslice.png in submitted folder.

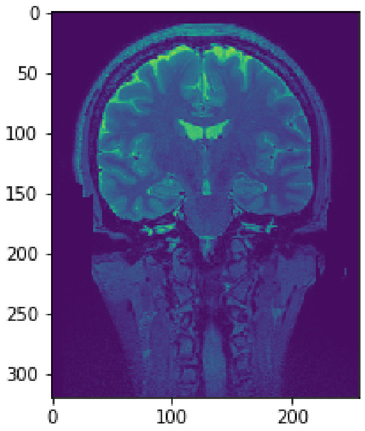

Figure 20: one slice from the volume

2. Extra Credit

   (a) Threshold problem
   Function named threshold is built to do process. Variable named tr is used to control threshold. If pixel value of image is lower than threshold , it is modified to zero. Otherwise, it is modified to one. And function named makeNew is used to prevent original data changing.

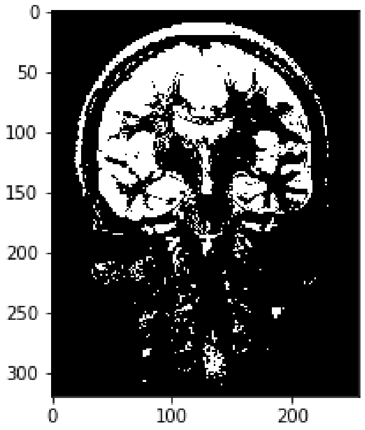

Figure 21: Threshold at 102

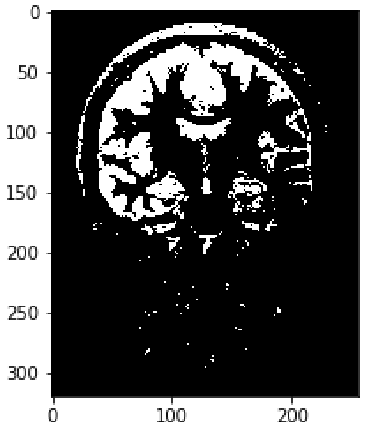

Figure 22: Threshold at 128