See Dropbox for Due Date
T. Kinsman

Homework is to be programmed in Python, R, Matlab, or Java.

When coding, assume that the grader has no knowledge of the language or API calls but can read comments. Use prolific comments before each section of code, or function call to explain what the code does, and why you are using it.

Do not use single letter variable names. That is for theory classes. This is an applied class.

Hand in your results, and the commented code, in the associated dropbox. Again, put all files in a directory with your name on it, HWNN_LastName_Firstname.
Then zip up the entire directory and submit the zip file.

Inside the directory we should find two files:

A. HWNN_<LASTNAME>_<Firstname>_results.pdf – your write up of what you did.
B. HWNN_<LASTNAME>_<Firstname>_program.ext – your program.

Substitute the homework number for NN.

Feel free to look over each other's shoulders, at each other's work, but do you own work.
Let me know whom you worked with. Do not hand in copies of each other's code.

Background Ridiculousness from a Fictional Universe:
Abominable snowfolk have been observed in the Himalayan Mountains. After decades of careful and expensive observations, the associated data was collected. What most of us do not know is that when they mature, the hair on the shoulders of most abominable snowfolk turns gray.
This is like the silver-backed great apes in other parts of the world.

It is now known that stress causes mammal's hairs to lose color and turn gray. The mechanism was recently discovered. (Not kidding: https://thenextweb.com/science/2020/01/23/scientists-figured-out-why-stress-turns-your-hair-gray/)

The data collected includes information about each individual observed. The curated data available to you now includes the age at which the individual started having gray hair on their shoulders, and the individual's approximate height in cm.

Overview:
It is believed that there are two sub-species of these snowfolks. Cluster them into two types, based on the age at which they matured.

Details follow:

(continued)

1. **Exploratory Data Analysis:** (1 pts)
   You are also provided with some mystery data, in the file MysteryData.txt.
   It consists of two underlying groups. This data is pre-quantized to the nearest unit.

   Questions:

   a. Compute the average and standard deviation of this data?
   b. Remove the last value from the data, and then re-compute the average value.
      How did average values change? Why do you think you observed these changes?
      What caused this amount of change?

2. **1D Clustering using Otsu's method on the age.** (4 pts total)

   Implement Otsu's method from scratch.
   Use Otsu's method to find the best age to form two clusters with.
   Make the decision based on "age ≤ each_threshold".

   What is the best threshold you found?

   - Quantize the snowfolks age into bins that are 2 years interval using the ranges the ranges [0-2), [2,4), ... so the first bin is age zero up-to (but not including) age 2.

     ```
     quantized_data  = floor(raw_data / BIN_SIZE_for_AGE) * BIN_SIZE_for_AGE
     ```

   - Implement Otsu's method to separate the snowfolk's population into two clusters.
     That is, we are using Otsu's method to binarize the data. There are other methods.
     We are quantizing the data into two groups.

   Questions:

   a. What age should we use to best separate the two clusters? (1 pt.)
   b. What is the minimum mixed variance that resulted? (1 pt.)
   c. Breaking Ties: How would your program handle a situation (1 pt.)
      where the same value of minimum mixed variance occurred twice?
      Does this situation happen?
   d. What other methods could we have used to quantize the data? (1 pt.)

3. **Graphing** (1 pts)
   Ignoring regularization, plot a graph of:
   the mixed variance for the snowfolk's data based on age versus the quantized age.

   Add a circular point indicating the value used to segment the data into two clusters.
   Clearly label all the axes.

(continued)

4. **Exploring Regularization:** (2 pts)

   Let: Cost_Function = Objective_Function + Regularization.
   Let the mixed variance be the objective function, as previously defined.

   We want to add a regularization term to encourage the two clusters to be the same size, however, it should not overwhelm the objective function.

   Regularization = abs((Number of Points in First Group) – (Number of Points in Second Group)) / NormFactor * alpha.

   Set the NormFactor to 100. This is about half the size of the data.

   Explore different values of alpha around:
   [100, 1, 1/5, 1/10, 1/20, to 1/25, to 1/50, and 1/100, 1/1000].
   Which value of alpha causes the "best" splitting point to change?
   Do you notice anything? There is not necessarily a change.

5. **Use quantization based on height.** (2 pts)
   Repeat the use of Otsu's method for clustering.
   Use a bin size of 5 cm.
   What do you find for the best height?
   What about the minimum mixed variance?

6. **Conclusion and Discussion:** (1 pt)

   Write a conclusion that describes what you learned in this homework.
   This is college. Write at least a five-sentence paragraph here.