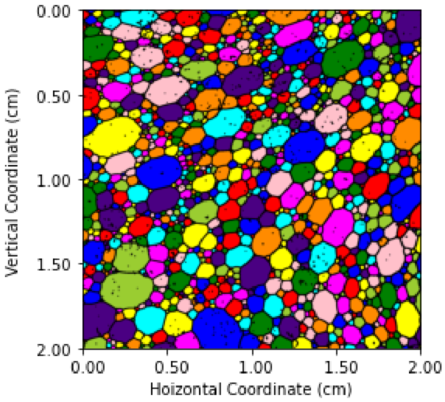
Data Science Competition

Fall 2023

Please submit your solutions to the Google spreadsheet provided at the start of the competition. You must provide both an answer and at least a quick explanation to how the solution was found to receive full credit. In the case of a tie-breaker, we will rank higher solutions which are (i) easy to explain/code and (ii) computationally efficient (run fast).

The ad wizards at Popsee Cola are interested in creating the perfect snapshot of their new soda creation: Popsee Techno. The gimmick here is that the soda bubbles are designed to be different colors.

A glass of Popsee is poured into a glass, and the ad wizards take a picture of the soda head from the side of the glass. Here's a snapshot of the foam:



You might have noticed that this form is much, much more disordered than the sample problem! Relatively small bubbles seem much more common, for instance. To make improvements to the picture aesthetics, the ad wizards at Popsee want to be able to quantify what the foam looks like. Similar to the sample dataset for the stained glass window, the Popsee dataset gives the following information regarding bubbles:

- `Area` The area of a bubble given in squared centimeters.

- `Perimeter` The perimeter of a bubble given in centimeters.
- `Centroid_y` The y-coordinate centroid (middle point) of a bubble given in centimeters, as shown in the picture (note that 0 is at the top).
- `Centroid_x` The x-coordinate centroid (middle point) of a bubble given in centimeters, as shown in the picture (note that 0 is on the left).
- `Degrees` The number of neighbors bordering a bubble.
- `cell_number` A label, or tag, for the bubble.

The following questions are of interest to Popsee. Don't sweat it if you can't answer all of these questions in the allotted time. Just focus on the easier ones to start off and see how far you can get. Better to answer a few precisely than to answer all of them poorly.

(**Easier questions**: 10 pts each)

1. How many bubbles are in the image?
2. What is the mean and variance of the bubble perimeters?
3. On average, how many neighbors does a bubble have?
4. Create a frequency plot for the bubble degrees (number of neighbors for each bubble)? Take a look at the mode of this plot. How does this differ from your answer in the previous question?
5. What's the largest area bubble in the picture? What is its coordinates and area? How much bigger is it compared to the average bubble area? How many sides does it have? Does this bubble also have the most sides?
6. What is the average bubble size on the top half of the image? What about the bottom half? Can you give a statistical statement comparing average bubble sizes?
7. What percentage of bubbles are "lonely", meaning having at most four neighbors? What's the average area of a lonely bubble?

   (**Trickier plotting questions:** 20 pts each)

8. Make a scatter plot that has for the x axis the number of neighbors, and y value of the average perimeter for cells with n numbers. If possible, make a horizontal line on the graph with a y value of the average cell perimeter. Find the number of neighbors a cell should have to have mean perimeter closest to that of the entire dataset?

9. Give a histogram of the cell areas. It should look quite terrible. Can you provide an appropriate scaling to make the histogram look more readable?

10. What's the relation between perimeter and area? In a graph of number of sides vs. average cell perimeter, is there some kind of a pattern? Can you give a function which approximates this relation?

(...and an even trickier modeling question: 30 pts)

11. Create a crude estimate for determining whether a cell is on the border of the image or not. What is the average number of neighbors for cells on the border of the image? How does this compare to average number of neighbors over all cells?