Data Science Competition Solutions

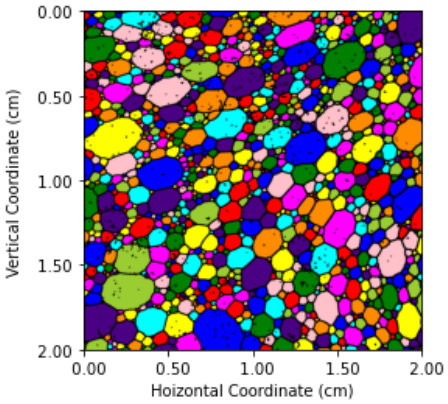
Fall 2023

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).You are encouraged to save code from this example, which

---

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**)

1. How many bubbles are in the image?

   There are $\boxed{1239}$ rows in the dataset, each corresponding to a bubble.

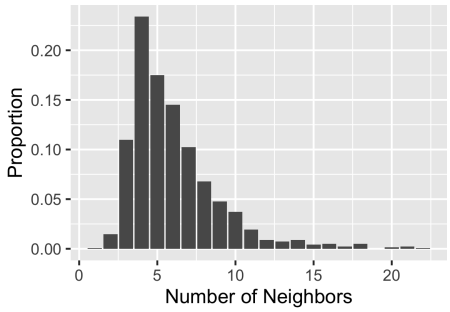
2. What is the mean and variance of the bubble perimeters?

   Calling on a mean and standard deviation function (or square root of variance), we find a mean of $\boxed{.157 \text{ cm}}$ and variance of $\boxed{.031 \text{ sq. cm}}$

3. On average, how many neighbors does a bubble have?

   The mean of the degree column gives an average of $\boxed{6.01}$ neighbors, which is what we would expect for a network having degree three at each vertex.

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?



The mode is at $\boxed{\text{4 sides}}$, which is quite different from the average number of sides near 6.

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?

   Using a built-in function which finds the index of the largest cell (in R, we can use the "which" function), we find the largest cell has label $\boxed{463}$. This has (x,y) centroid coordinates of $\boxed{(.179, .751)}$ and an area of $\boxed{.054}$ sq. cm. Note that we can use the image to check that this coordinate lands on a yellow, largish looking cell for a quick check. This cell has lots of neighbors- $\boxed{21}$ in fact, but it's not the most! Take a look at the previous plot and note that there is a cell with 22 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?

   We can separate our data by looking at cells with centroid with y-coordinates less than or greater than 1. Careful with how we set up the coordinates! Lower y-values correspond to the upper part of the image. The upper part of the image has a mean area of $\boxed{\text{.0022 sq. cm}}$, while the bottom has an average of $\boxed{\text{.0027 sq. cm}}$. Seems like a minor difference, but percentage-wise the top cells have a mean area that is (.0027-.0022)/.0027= 19% smaller than the bottom cells.

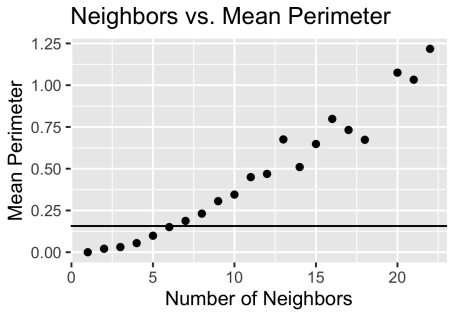
   We can run a two-sample t-test or permutation test to compare the means. A permutation test gives us a p-value of .0476 that the means are different. Decent evidence, but nothing out of the park here.

7. What percentage of bubbles are "lonely", meaning having at most four neighbors? What's the average area of a lonely bubble?

   Filtering on bubbles having at most four sides, we find that the proportion of lonley cells is $\boxed{.359}$. Perhaps unsurprisingly, the area of these cells are much smaller at $\boxed{\text{.000226 sq. cm}}$.

   (**Trickier plotting questions**)

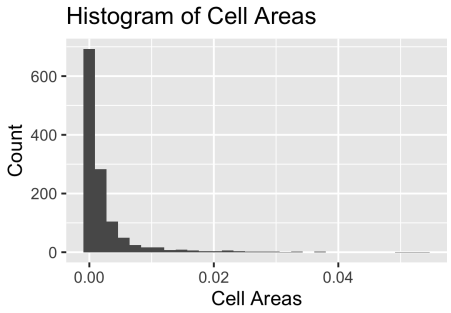
8. Make a scatter plot that has for the x axis the number of neighbors, and y value of the average perimeter for cells with n numbers. If possible, make a horizontal line on the graph with a y value of the average cell perimeter. Find the number of neighbors a cell should have to have mean perimeter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



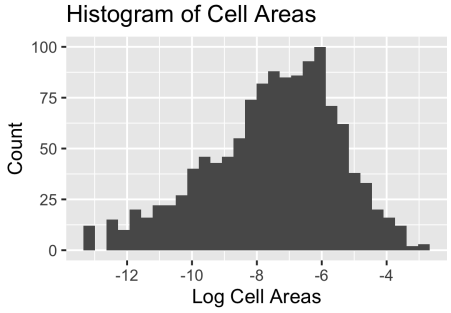
Here we group with respect to degree, and then find mean areas for each class of neighbors. In R, this is done with a combination of dplyr and ggplot tools (see R code). We see that the average perimeter is quite close to the average perimeter conditioned on 6 sided cells.

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?

   Here's a histogram of cell areas with no tinkering:



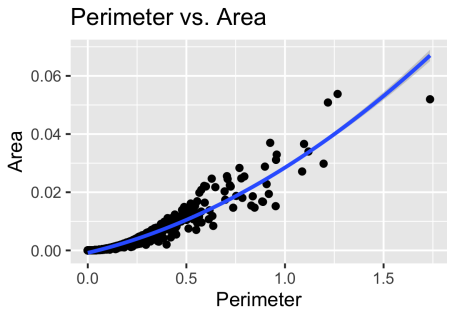
This is quite scrunched up near zero, and it's hard to see that distribution of larger cells. One quick fix (are there others?) is to take the log area of each cell. The resulting distribution is as follows:



10. What's the relation between perimeter and area? In a graph of number of perimeter vs. area, is there some kind of a pattern?

    Can you give a function which approximates this relation?

    Here's a scatterplot of cell perimeter vs. area



    The blue function you see is approximating by a quadratic function. The regression function is $A(P) = (1.33 \times 10^{-2})P^2 + (1.60 \times 10^{-2})P - (8.28 \times 10^{-4})$.

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?

    If you were given pixel data of every cell, you can simply check to see if any pixels lie on the image boundary. Since we're only given centroids, we'll have to make some estimates. Here's a crude one:

    (a) A cell has median area of .00068 sq cm. (note: we could also use mean area. Why did we use median instead?)

    (b) Assuming cells to be squares (!!), this would mean a median side length of the square root of the cell area, with a value of $d = .026$ cm. (this step is also quite crude in many ways: using a square instead of more representative shapes, that we're using diameter instead of radius).

    (c) Now that we have our typical cell length, we declare a cell to be in a boundary cell if either the x or y centroid are within a distance $d/2$ of the boundary.

    Using this (cruuuude) estimate, we filter cells near the boundary, and find that the average number of neighbors of such cells is $\boxed{4.01}$, quite smaller than those of the entire foam. This is a common occurrence-boundary cells can skew statistics because the entire cell is censored, including in this case the total number of cell neighbors.