Data Science Competition

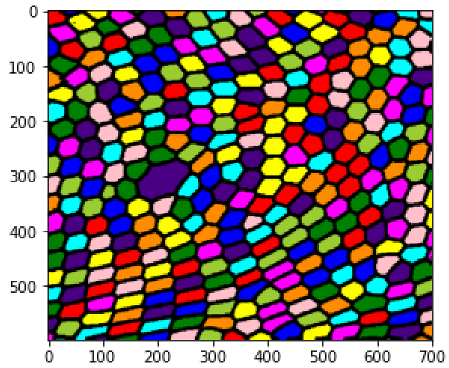
Fall 2023

Sample Questions

Note: You are encouraged to save code from this example, which can be referred to during the actual data competition. Many of the questions for the actual competition will resemble those seen here!

The mad medieval scientists at Stained Glass Creations (SGC) have devised a machine that automates their window making process. The scientists claim that the machine will create a perfectly uniform grid of stained glass cells. If their machine is successful, they can corner the market on churches, castles, and medieval-themed pizza companies (note: google image "Vince the Pizza Prince" for a glimpse of Scranton history).

At the moment their machine is far from ideal. A sample window looks like this:



Note that the numbers on the axes area given in millimeters. It ought to look like a perfect honeycomb pattern, with each cell a regular hexagon, and all cells having the same area. To make improvements, the scientists want to be able to quantify the deformities of their windows. The dataset provided gives the following information regarding these cells:

- `Area` The area of a cell given in millimeters squared.

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

To be more specific about the centroid, consider a cell C that has pixel coordinates $\{(x_1, y_1), \ldots, (x_n, y_n)\}$. Then the centroid of C has coordinates $(\bar{x}, \bar{y})$ defined by the average positions of the x and y values, or $\bar{x} = (x_1 + \ldots + x_n)/n$ and $\bar{y} = (y_1 + \ldots + y_n)/n$. For something like the cells we're considering, it's approximately where you would point to if asked to identify the "center" of the cell.

The following questions are of interest to SGC:

(**Easier questions**)

1. How many cells are there?
2. What's the average area of a cell?
3. What is the histogram of the cell areas?
4. What's the largest cell in the picture? How much bigger is it compared to the average cell size? How many sides does it have?
5. What about the second largest cell? What is its area and centroid coordinates?
6. What percentage of cells have "defective sides", meaning not having six sides?
7. Assuming all windows from SGC are produced with the same distribution of cell areas and number of neighbors. Given data from this image, what statistical evidence is there that seven sided cells have large areas than five sided cells?

(**Harder questions**)

1. What's the relation between number of neighbors and area? In a graph of number of neigbors vs. average cell area, is there some kind of a pattern? Can you give a function which approximates this relation? Can you give a statistical statement about how good your fit is?

2. Can you give another graph of cell perimeter vs. cell area? What does this function look like?

3. How many five neighbor cells have areas larger than the average cell area?

4. How many seven neighbor cells have areas less than the average cell area?

5. Estimate how many cells are on the border of the window. How many neighbors are there for border cells, on average? Compute a frequency plot of number of neighbors for both cells and interior cells.

Solutions and hints

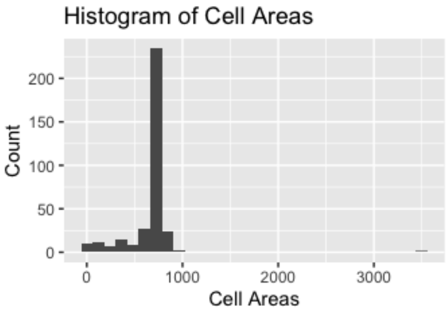
The plots and code used to generate answers were generated in R. However, the answers should be available by using basic filtering and regression tools from any popular programming language (including Excel).

(Easier)

1. Take a length of any column to get that there are **340** cells.

2. You can either use a mean function or sum up the total areas of cells and divide by 340 to obtain an average size of **661.7 mm sq**.

3. Here's a histogram



   Areas look pretty concentrated around the mean, and we can also see our very large outlier cell at about five times the size of the mean cell.

4. The megacell is number 165. It has an area of 3500 mm sq. It's an 8-gon (having 8 sides) A good way to check this is the right cell is to check the centroid of the cell, which has coordinates (210, 307). It is 3500/661.7 = 5.29 bigger than the average cell.

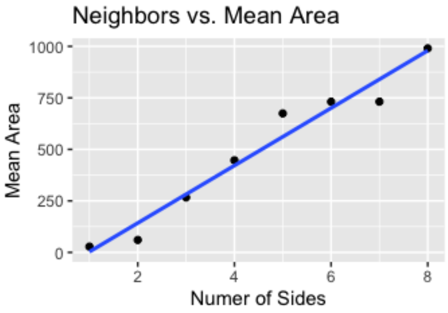
5. Sort rows according to descending size, you'll be looking at the second row, belong to cell 36. The area is 986 mm squares, with a center located at (370, 50).

6. There 129 cells not having 6 cells. Kind of surprising at first blush, but not so much when we consider boundary cells.

7. We can use a one-sided, two-sample t-test (or permutation test, take your pick) between filtered vectors containing only areas 5 and 7 sided cells to obtain a p-value of .009.

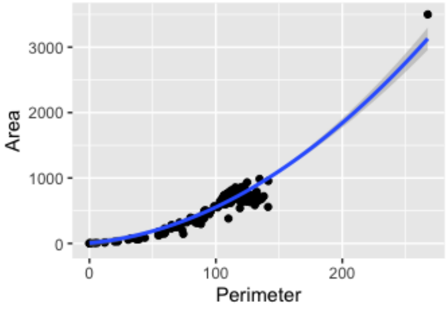
Harder questions

1. Here's a graph of a linear fit.



   The linear fit is Area = -136.38 + 139.47· Sides. The correlation coefficient is $r = .98$, which is a pretty good fit!

2. Here's a plot of perimeter vs area



   We've included a quadratic polynomial relation

$$\text{Area} = .037\ \text{Perimeter}^2 + 1.79\text{Perimeter} + 6.41$$

The coefficient in front of the square term might suggest that the quadratic term is not important, but for values over 100, the quadratic term is, in fact, dominant. The multiple R-squared value is .85 which is still a pretty good fit.

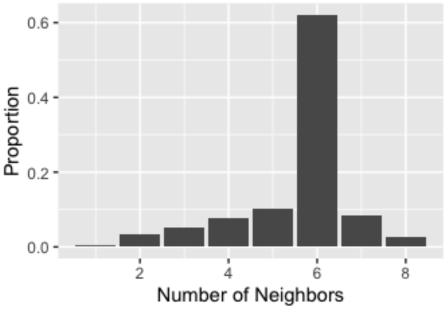
3. Here we need to filter on five sides, and then further filter on those sides being greater than average. Surprisingly 25 out of 35 cells are larger than average! This is a bit surprising, since one might reason

    - There's (about) six neighbors on average
    - Average areas increase with the number of neighbors.
    - Therefore, most cells with less than six neighbors (including 5 neighbor cells) should have areas smaller than average.

    However, we need to be careful about considering mean versus the median, which is more impervious to outliers. In this case, the median area is about 713 mm squared, and only 13 of the 35 five neighbor cells have area greater than this median. By the way, what might be causing the difference between the mean and median?

4. Same bit as three, but adjusting for seven sides, and applying a different filter. In this case, not a single cell is less than average! When considering the median size, about half of the cells (16 out of 29) are smaller.

5. I won't spoil the fun in determining whether a cell is near the border, or how to estimate doing so, but here's a relative frequency plot of the number of sides for the entire window.

    A nifty fact about 3-regular networks (meaning that the edges of the cells always meet in threes) is that we should expect cells to have (close to) six neighbors on average. However, when we look at the average number of neighbors, we get an average of 5.58. The problem here is the border cells. When you make your estimate for whether a cell is a border cell or not, you can test whether it's a good estimate by verifying that your interior cells (those not on the border) have an average number of cells sides near 6. There's a good chance that you'll be asked to estimate if a cell is interior for the main competition, so it's a good idea to write scripts which can easily adjust to different scenarios.