Home Work #1

1. Professor Neila Morfsram has stunned the world by claiming not only that some aliens were captured in the 'mysterious incident' in New Mexico in July 1947, but also that the strange powers they possessed are linked to a certain pattern in their DNA sequence: occurrences of the base G immediately followed by the base C. (A DNA sequence is a list of bases, each being one of {A, C, G, T}.) Now, she has asked for your help in exploring the relationship between the heights of their foreheads and the patterns.

   You can use her data file data-aliens.txt, which is a comma-separated file with one line per alien; each line contains an ID string, followed by the alien's *forehead height* in inches, followed by the DNA sequence.

   Create a file named `<yourfirstname>.py` containing a Python function `main()` plus supporting functions as appropriate (**without importing any libraries**) that

   (a) reads the data file;

   (b) iterates over the aliens;

   (c) for each alien, determines four attributes: whether the forehead height is small ($\leq$ 3.5 inches) or large($>$ 3.5 inches); p, the number of occurrences of a pattern (base *G* immediately followed by base *C*) in the DNA sequence; n, the number of bases that are either *C* or *G*; and the ratio $2p/n$ expressed as a percentage.

   (d) prints a table with a header row plus a row for each alien. The columns are *ID, Small, Large, Number of Patterns, Number C/G,* and *Pattern Frequency*.

   The values for *Small* are 1 if the forehead height is determined to be small, and 0 otherwise (and similarly for *Large*). The values for the last three columns are p, n, and the percentage corresponding to $2p/n$ correct to 2 decimal places.

   (e) prints a table with a header row plus two rows: one for each group based on the forehead height (small and large). The columns are *Forehead Size Category*, *Count*, *Total number of Patterns*, *Total number C/G,* and *Aggregate Pattern Frequency*. *Count* stands for the total number of aliens possessing that forehead size; The *Total*s are P = sum of p and N = sum of n over the respective groups; and the last column is $2P/N$ (as percent, correct to two decimal places).

   (f) exports the last table into a CSV file named `<yourfirstname>-result.csv` containing one header line plus two lines (one for 'S', the other for 'L'):

   ```
   ForeheadCategory,Count,TotalNumPattern,TotalNumC/G,AggPatFreq
   S,...
   L,...
   ```

   - Your tables must have neatly aligned columns.
   - Add a boilerplate execution control via `__name__` (next class).
   - **Your code must be executable in Python 3.**