Assignment Genetics CSIS-2530

Learning Objectives:

- Practics reading and writing to files
- Extract data from a csv file
- Formulate algorithms

Turn in:

Turn in the zipped up code via Canvas

Description:

Based on the data provided in a csv file (csv) create two new files that include subsets of the original data.
You can implement this assignment as a **console** application **or** as a **WPF** application**.**

- Allow the user to specify the path of the input file.
  If you write a console app allow the user to re-enter the file path multiple times – until s/he gets it right.
- Save the output files in the same directory where the input file is located (*Hint: check out class Path*)

Input

The input is provided in file **Mutations.csv**. It includes data from a genetics research lab. You might have noticed that all Genes start with A. That is the case because this is only a small part of a much larger file. (The original inclucded more than 150,000 rows)

Each row include information about a mutation found on a given gene. It also includes a column for each of the test subjects (persons) to indicate whether that particular mutation was present for the given person. (those are the last 15 columns before the very last column – U to AI)
Each gene can have multiple mutations, which means there can be multiple rows for one gene.

Output:

The output should be 2 new files: **SignificantGenes.csv** and **Summary.csv**

SignificantGenes.csv:

This files should include:

1. The same header row as the original file
2. A subset of data rows from the original file. Only include the data of genes, that have at least 3 mutations (i.e. rows, where at least n rows have the same gene value in column Gene)

Summary.csv:

Create a file called Summary.csv. It includes summarized data from SignificantGenes.csv (the file you just created)

While in SignificantGenes.csv each gene is listed in at least n rows and each row is a copy of the original input file, in the file Summary.csv each gene from SignificantGenes.csv is listed only once (**one row per gene**).
Also: rows in the summary file have fewer columns.
Here is what each row in Summary.csv should include:

1. A header row with the header names **exactly as specified in the sample file**
2. Rows of this form:
   - Gene
   - Number of test subjects with mutations in the given gene
   - Number of mutations of that given gene (i.e. number of rows of this gene in SignificantGenes.csv)
   - One column for each test subjects (listed in the same order as in the original file)
     1 .. test subject has at least one of the mutations of the given gene
     0 .. test subject has none of the mutations of the given gene