**Project 1 – Cleaning Up Brad’s Short Sell File**

**Graphical user interface, text, application

Description automatically generated**

**Background.** Brad Hennes was a data science major that graduated in 2019. While completing his degree, he also worked as an intern for DMR, Inc.; where he now works as a full stack developer. We asked Brad if there were any skills he wished we had covered, and he emphatically answered: cleaning up messy text files. I asked Brad for some examples, and he (and his boss) graciously provided a number of examples. Your job in this project is to clean-up and parse one of these files.

**Project Overview.** This project will be split into a number of parts, that are summarized below.

1. **Part 1 – Triage (20 points).** In the first part of the project, you will inspect the sell\_short.txt file and diagnosis as many problems as you can, then suggest a strategy to clean up this file.
2. **Part 2 – Removing Extra Lines and Combining Data Blocks (40 points).** As with the example presented in class, the data for each case are spread across multiple rows, with many extraneous lines that need to be removed. You will fix these problems in the second part of the project. This task is a little more complicated than the last, as the data blocks have a variable number of lines.
3. **Part 3 – Reordering and Extracting Data (40 points).** Finally, you will extract the data for each case, placing the results in a clean csv file. Note that there are a number of issues to deal with here, such as data not always in the same order or columns that run together.

**Evaluation and Assessment.** Each part of your project will be evaluated based on your performance in a number of areas.

1. **Part 1 – Triage (20 points).** This part of the process will be evaluated on (a) how many issues you identify and (b) the quality of your strategies for solving said problems.
2. **Part 2 – Removing Extra Lines and Combining Data Blocks (40 points).** The second part of the project will be assessed based on (a) the readability/quality of your code and (b) the correctness of the resulting output file.
3. **Part 3 – Reordering and Extracting Data (40 points** The second part of the project will be assessed based on (a) the readability/quality of your code and (b) the correctness of the resulting output file.