Data Science Problem Assignment

You are a data science consultant!

Your client has made a terrible mistake. They have acquired a company from a bankruptcy proceeding. The acquired company was called Vanish and sold weight loss products. They had only one asset of value, their receivables, roughly $40 million in total, which are recorded in a ledger. The ledger file has a column for each of the following fields

first name

last name

address street number

street name

city

state

amount owed

age

This is enough information to bill the people and collect the money. What could go wrong with that?

The trouble is that the ledger file has been corrupted. What appears to have happened is the following. The file was written out one buffer at a time. A buffer is fixed number of lines. A bug seems to have created the situation where before every buffer-write, there is some probability of a random mutation occurring. The effect of that mutation is that the location of two randomly chosen columns gets swapped from the previous state. Most often, there is no change but sometimes two columns are swapped. There are never more than two columns swapped at a buffer boundary.

For example:

If the column ordering starts out correctly corresponding to the header row as:

first\_name, last\_name, street\_number, city, state, amount, age

and after a few buffer-writes, the first column (first\_name) and fourth column (city) might get swapped and so the ordering now becomes:

city, last\_name, street\_number, first\_name, state, amount, age

A few buffer-writes later the third column (street\_number) and last column (age) might get swapped and so the ordering now becomes:

city, last\_name, age, first\_name, state, amount, street\_number

Before long, the order is completely shuffled from what the header row says the column should mean. The column ordering is however fixed within a buffer write.

Your job as a data scientist consultant is to find a way of correcting the ledger as best you can so that your client can mail the bills to the right people at the right address with the right amount in order to recover value from their investment.

You'll find that sorting out the text strings are not too difficult but that the three numbers: age, amount and street number, have overlapping ranges and so determining which is which at any point in the ledger requires some clever strategy. You can assume the people are all adults (18+).

**Instructions**

When you are done, send us the following:

1) Your corrected ledger file in the same csv format as the corrupted file

2) Your code

The code should take the ledger\_corrupted.csv file and produce the corrected one that you submitted. Also include instructions for installing and running your code and running any tests. We prefer that you use python but any language is OK.

3) An explanation file

This should be a short write-up on the method you used. Be sure to include the following:

A description on how you proceeded on the problem. What did you try first? How did you improve on that? What would you do next? What do you suppose is the optimal way of solving this problem were you to have a lot more time and incentive to continue? How would you refactor your code to improve code quality were you to put this in production? How did you validate that your solution actually works?

We'd also love to get some feedback on the problem as well. Did you enjoy working on it? The correct answer is yes ;)

We will give you four days to complete the assignment but if you need a little more time, we can accommodate. In total, we don't want you putting much more than about 8 hours of total work into this. But there is no time limit. It’s not a race. We can judge the quality of your thought process and coding style without a finished solution but it would be nice if you at least have some solution with a plan to improve it.

If you don’t have time to get much code written in this time, at least send us what you have and your explanation. What we’re mostly looking for is your thought process, how you solve problems from analyzing data and what mental tools and coding skills you have to solve complex problems like this. When you write code, is it well structured, testable and maintainable?

Please keep the problem private so that it does not get exposed to the public or any future candidates.

**Scoring**

The final scoring will depend on two things. It depends on the accuracy of your result; the percentage of lines that are correct. The record is 99.5% in case you’re wondering but that was the person who created the problem and we’d be surprised if you get anywhere close to that. It also depends on the quality of your code and description.

Regarding code quality:

We're looking for the usual things that would be expected if this were a software developer role. However, data science code can be a little harder to test particularly since you often need to try many different things and so we aren't expecting full test coverage or production quality code. We do expect some tests and that you follow standard coding practices and that the code is relatively well-structured, easy enough to read etc.

**Data provided**

Along with the corrupted ledger file, ledger\_corrupted.csv, there are five files that

may be helpful.

first\_names\_boys.csv

first\_names\_girls.csv

last\_names.csv

city.csv

states.csv

These name files have the name and the frequency of occurrence from the US Social Security Agency. You are on your own with regard to the other fields but you shouldn’t need any other external data.

Good luck!