Homework 1

Flat Files & Physical Data

This exercise introduces you to the problems encountered with storing data in "flat file" databases, and the drawbacks of coupling data to its physical structure on disk. Your goal is to write a program that reads data from a comma-separated value (CSV) file, make changes to the data, and think about the impact to your program when the structure of the data changes.

Scenario

Imagine that you work as a developer within the analytics division of a company that sells shoes online. The e-commerce team has provided you with a CSV file containing millions of sales transactions, and your boss, a slightly smarmy, yet somehow charming, CEO, asks you to answer the following questions:

1. How many of our customers are named Amanda?
2. What is the average sale amount of the transactions?
3. Can we have a version of the data that shows USA instead of United States? (Because `murica.)

Your goal is to write a program that helps answer these questions.

Step 1: Read and Print the Data

Using the programming language and libraries of your choice, create a program that reads data from a provided CSV file, and displays the data on the screen. Display the transactions ordered by "Product".

Step 2: Number of Amandas

Add a new feature to your program that displays the number of customers with the name "Amanda". After this goal is complete, think about the logic in your program that is necessary for completing this goal. What issues did you have to handle?

Step 3: Average Transaction Amount

Add a new feature to your program that displays the average transaction amount in the database. What issues did you have to handle?

Step 4: USA! USA!

Add a new feature to your program that changes the name of the country "United States" to "USA". Write the resulting data to a new CSV file on disk. In other words, your program should create a new CSV file that is identical to the one that it reads, but with every occurrence "United States" changed to "USA". How many records had to change?

Step 5: Data Structure Changes

In the original CSV file, you'll notice that the "State" column can contain a US state or the region within a non-US country. Imagine that the e-commerce team decides to change the sales database such that they add a new "region" column, so that the "state" column only contains US states, and the "region" column stores the name of a region within a country.

This change is represented in the second CSV file provided with this assignment. Without making any code changes (other than, perhaps, the filename used in the program), re-run your program using this second CSV file.

What breaks? Why? What needs to change in your program now that the structure of the data has changed?