STAT 29000 Project 3

Topics: Review about Topics in R and Awk

Motivation: We learned many concepts about R and Awk in STAT 19000… and reviewing these concepts is worthwhile. In this project, we remind ourselves how to use the tapply function. We also compare the speed of doing the same types of operations in R and Awk. By reviewing some concepts that we learned in STAT 19000, we prepare ourselves for writing more bash scripts, using more vectorized operations in R, and for thinking about larger computations on sets of data that are much larger in magnitude.

Context: Recall that the tapply function enables us to apply a function to pieces of a vector, according to the data stored in another vector. Remember that it works like this: `tapply(the data to work on, the way that the data is organized, a function to run on each piece of data)`

Also recall that we want to use the right tool for the right task. R is a wonderful tool, but it is a high-level tool, and it cannot match the speed of low-level tools like Awk. We compare the speed and effectiveness of working in R and Awk.

Scope: It is worthwhile to think ahead to how we organize the data analysis. We will often chop the data into pieces using UNIX tools, so that we import only the relevant data into R. Such division of labor among our data analysis tools does not happen automatically. It is useful to think about how to organize our computations, and how long they will take to run. We use examples from the grocery store transaction data, and from the airline data set.

Question 1: Introduction to Grocery Store Customer Insights, using the `tapply` function

Try the following *inside* the Scholar environment. This question deals with data from 84.51. They use data science to provide customer insights to Kroger. Using the 84.51 data set called `5000_transactions.csv`, found in the folder `The_Complete_Journey_2_Master`, consider the following questions:

1a. Which basket contains the largest total dollar amount of food purchased?

1b. Which basket contains the largest number of items purchased? Please note that a product might be purchased in multiple quantities (i.e., you need to pay attention to the number of units purchased).

1c. Which product was purchased most often (again, you need to take into account the number of units purchased)?

Question 2: Speed Test / Reminder: Comparing the usage of R and Awk: Analysis of Flight Data with Multiple Tools

How many miles have been flown altogether (i.e., the sum of the flight distances) during 1987 to 2019?

2a. First solve this question in R.

2b. Then solve this question in Awk.

The solution should be the same with both tools… but which solution is faster? Why?

Question 3: Indiana Remix of the Speed Test

3a./3b. Same questions as 2a./2b. but now focus only on flights that depart from Indianapolis.

Project Submission:

Submit your solutions for the project at this URL: https://classroom.github.com/a/qsUSiy2A using the instructions found in the GitHub Classroom instructions folder on Blackboard.