STAT 19000 Project 6

Topics: More functions in R

Motivation: The ability to split, merge, and take subsets of a data frame is essential. These are some of the most fundamental manipulations of data frames.

Context: It is possible to use more sophisticated transformations of data frames, using the `tidyverse` suite of R packages… but it is also very important to understand the ways to manipulate data frames using the built-in functions from R. These allow us to break a data frame into many different data frames; to combine data frames of different sizes; and to extract key pieces of information from data frames.

Scope: This project focuses on reading the documentation about these three functions: split, merge, subset, and seeing how they can be used with some data frames about grocery store purchases… but we emphasize that (in general) these functions are broadly applicable.

As always, in addition to solving the questions in the project, this is only the beginning. We encourage you to experiment and try things out for yourself, and learn how to go deeper into your knowledge of these tools and their applications.

Question 1: Introduction to the split function

1a. Read in the `5000_transactions.csv` data (from 8451) into a data frame to be called `myDF`.

1b. Split the data frame `myDF`, using the `STORE_R` column, and store the results of the split into a new variable called `myresults`. Use the split command to achieve this. Remember that we can read about the split command using: `?split`

1c. What is the class of `myresults`? What is the length of `myresults`? What are the names of `myresults`? (Use class, length, and names on myresults.)

1d. Check the dimensions (`dim`) and the head of `myresults[["CENTRAL"]]`.

1e. Now manually make a data frame that has all of the same columns as `myDF` but only has rows for which `myDF$STORE_R` is equal to `"CENTRAL"`:

```
centralresults <- myDF[myDF$STORE_R == "CENTRAL", ]
```

Verify that the `dim` and `head` of `myresults[["CENTRAL"]]` and `centralresults` look the same.

Question 2: Introduction to the merge function

2a. Read in the `5000_products.csv` data (from 8451) into a data frame to be called `myproducts`.

2b. Merge the data frames `myDF` and `myproducts`, according to the `"PRODUCT_NUM"` column (which is common to both data frames). Store the results of the merge into a new variable called `mybigDF`. Remember that we can read about the merge command using: `?merge`. Hint: You can use `by="PRODUCT_NUM"`

Question 3: Introduction to the subset function

3a. Take a subset of the data frame `myDF` that shows all of data about the purchases made on 23 December 2017. You do not need to store the results of the subset function anywhere. Remember that we can read about the subset command using: `?subset`

3b. Take a subset of the data frame `myDF` that shows only the dollar amounts of the purchases made on 23 December 2017.

3c. Take a subset of the data frame `myDF` that shows only the dates and dollar amounts of the purchases made on 23 December 2017.

3d. Take a subset of the data frame `myDF` that shows only the dates and dollar amounts and stores of the purchases made on 23 December 2017.

3e. On December 23, 2017, which store had the largest total amount (in dollars) of purchases? Hint: Use the `tapply` function.

Project Submission:

Submit your solutions for the project at this URL: https://classroom.github.com/a/594R_pB8 using the instructions found in the GitHub Classroom instructions folder on Blackboard.