STAT 19000 Project 5

Topics: Getting familiar with tapply

Motivation: R can apply a function to smaller pieces of our data sets in very easy ways, once we get used to these methods. It helps to try the tapply function using our examples, and then to try some examples of your own. In our family, we do not say "practice makes perfect" but rather we say "practice makes progress". We become strong in a new skill by practicing.

Context: Sometimes we desire to add all entries in a column of a data frame: we just add everything together. On the other hand, sometimes we want to add a few of the elements at a time, basically breaking the column into groups, according to some method of grouping the items The tapply function (think to yourself: "table apply") allows us to do so. There are many other apply functions in R, but tapply is perhaps the easiest to learn at the start. Indeed, tapply allows us to apply any function to data that is grouped together in the way that we want.

Scope: tapply 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)

It is best understood by consider the examples given in the file below.

Question 1: Largest Number of Reviews on AirBnB

Before you start question 1, please consider the examples given here:

`/class/datamine/data/examples/project5examples.R`

These will help you to better understand some of the functions that we will use for this project.

For this question, import the data about the AirBnB listings from Los Angeles:

`/class/datamine/data/airbnb/united-states/ca/los-angeles/2019-07-08/visualisations/listings.csv`

1a. Which host_id has received the largest number of reviews?

1b. Which neighbourhood has received the largest number of reviews?

Question 2: Understanding Airplane Flight Delays

Import that data from the 2019 airplane flights.

2a. Paste together the columns for the 3-letter codes for the origin and destination of each flight, and store the result in a new column of the data frame, called myflightpath.

2b. Which flight path has the longest average departure delay?

Question 3: Analysis of the 2020 Presidential Election Campaign Contributions

Import the data about the federal campaign contributions that have been given (so far) during the 2020 election season. The data is described here:

https://www.fec.gov/campaign-finance-data/contributions-individuals-file-description/

When you import the data, use the `read.delim` command, with the argument `sep="|"` because the file is not comma-separated. It has the symbol `|` between the elements of the data.

```
myDF <- read.delim("/class/datamine/data/election/itcont2020.txt", sep="|")
```

3a. Use the new "location" column, which you created in Question 3b of Project 4. Classify each (joint) city, state "location" according to the total amount donated by residents of that city, state.

3b. Which six campaign committees have received the largest total dollar amount of donations (so far) in the 2020 election campaign?

Project Submission:

Submit your solutions for the project at this URL: https://classroom.github.com/a/0zQUvToF using the instructions found in the GitHub Classroom instructions folder on Blackboard. Please submit R code (in a .R file) and make #comments throughout, to explain to a stranger (like the graders, for instance!) what you are doing with each part of the code.