STAT 29000 Project 9

Topics: Introduction to tidyverse

Motivation: The use of a suite of packages referred to as the `tidyverse` is popular with many R users. It is useful to gain some familiarity with this collection of packages, in case you run into a situation where these packages are needed. You have already learned about the `ggmap` package. A related package called `ggplot2` is in `tidyverse`. This package is one of the best visualization tools out there. It is extremely useful for creating beautiful, customized plots.

Context: We now have a solid foundation using R and its excellent tooling. You should be able to do almost anything you need, using base R. Nonetheless, some people are big proponents of a suite of packages referred to as the `tidyverse`. Exposure to reading and writing R code that uses `tidyverse` will almost certainly be useful to you, either during your courses, your research experiences, or during your future employment.

Scope: `Tidyverse` lets us do common data wrangling tasks in an organized manner. Additionally, `ggplot2` provides some of the best tools to create highly customizable, good-looking, insightful graphics.

A good resource for examples and learning more about `tidyverse` is https://r4ds.had.co.nz/tidy-data.html.

Question 1: read_csv, tibbles

1a. Explore the `tidyverse` website https://www.tidyverse.org/. What is the name of the package used to read rectangular data (i.e., data in a spreadsheet shape)?

1b. Read the file `/class/datamine/data/forest/SURVEY.csv` using the standard `read.csv` function. Save the output to a variable named `df1`. Now install the `readr` package. Then use the `read_csv` function from the `readr` package, to read the same dataset into a variable named `df2`. What are some of the differences that you see between `df1` and `df2`?

1c. A `tibble` is essentially a rebranded data.frame. Load the 2019 flight data into a `data.frame` called `myDF` using the `read.csv` command. Then read in the 2019 flight data again, this time into a `tibble` called `mytibble`, using the `read_csv` command. Consider the classes of 110 columns in `myDF` and `mytibble`. How many of these columns have the same classes? How many of these columns have different classes? Briefly describe a few of the differences that you observe between `myDF` and `mytibble`.

Question 2: A 'couple' of packages: dplyr & tidyr

For the following questions, please limit yourself to using the `tidyverse` packages. First use the `read_csv` function to load the data found at https://raw.githubusercontent.com/fastforwardlabs/couples-lime/master/couples.csv into a `tibble` named `couples`.

2a. Look at this data a little bit differently than it is presented in this dataset: Take the data from the `age` column, break it into three categories, `[0,20)`, `[20-65)`, `[65+]` and put the results into a new column called `age_category`.

2b. Now take the data from the `age` column, and break it into ten categories, `[0,10)`, `[10-20)`, ..., `[90-100)`, and put the results into a new column called `age_decade`.

For parts 2c and 2d, do not slice your tibbles.

2c. Create a `tibble` for all male, hispanic individiuals, which contains only the variables related to `age`, `age_category`, `age_decade`, `age_diff_abs`, `partner_age`, and `partner_education`.

2d. For each `age_category` group in the `tibble` in question 2c, what is the mean `partner_age` (within that group)?

Question 3: Thinking outside the box

3a. Use ggplot2 to create a side-by-side `geom_boxplot` where the x-axis shows the three age categories (from `age_category` in the previous question) and the y-axis shows the absolute age difference between couples. Label the x-axis `Age group` and the y-axis to be `Absolute age difference`. Lastly, vary the boxplot color using the 'fill' argument to aes(). Label the legend `Age group`.

3b. As you can see, there is 1 extreme outlier that is skewing our view of the data. Identify the outlier in your dataset, and re-do the plot without the 1 extreme outlier. The rest of the outliers are represented as dots (outside the primary graphic). Make the rest of the outliers be colored red. Add a label to the right of each outlier where the age difference is greater than 25 or equal to 25 years. The label should be the `caseid_new`. Please use `check_overlap=T` to remove overlapping values.

The resulting plot is depicted at https://datamine.purdue.edu/seminars/fall2019/stat29000project9question3.jpg

Helpful Links/Resources

Q2a, Q2b. R4DS link:

- 5.5 Add new variables with `mutate()`,

https://r4ds.had.co.nz/transform.html#add-new-variables-with-mutate

Q2c. R4DS links:

- 5.2 Filter rows with `filter()`,

https://r4ds.had.co.nz/transform.html#filter-rows-with-filter

- 5.4 Select columns with `select()`,

https://r4ds.had.co.nz/transform.html#select

- 5.6.1 Combining multiple operations with the pipe,

https://r4ds.had.co.nz/transform.html#combining-multiple-operations-with-the-pipe

Q2d. R4DS link:

- 5.6 Grouped summaries with `summarise()`,

https://r4ds.had.co.nz/transform.html#grouped-summaries-with-summarise

Q3. R4DS link:

- 3 Data visualisation,

https://r4ds.had.co.nz/data-visualisation.html

Q3a. Links:

- ggplot2: Modify labels,

https://ggplot2.tidyverse.org/reference/labs.html

- ggplot2: `geom_boxplot()`,

https://ggplot2.tidyverse.org/reference/geom_boxplot.html

Q3b Links:

- Add labels, an example on Stack Overflow

https://stackoverflow.com/questions/33524669/labeling-outliers-of-boxplots-in-r/33525389

Project Submission:

Submit your solutions for the project at this URL: https://classroom.github.com/a/6y5ZMUzx using the instructions found in the GitHub Classroom instructions folder on Blackboard.