Exercise: Grouping and Summarizing Data Frames

Contents

Overview

For this exercise, you will use a data frame and select variables to subset the data, filter cases/observations, group, and and summarize the data. I would recommend creating a script for this exercise.

Loading Libraries and Defining Functions

Load the `{tidyverse}` ecosystem and source your functions directory so that you have your tools in place.

Data

For this exercise, we will not formally read a data file using read functions. Instead, you will use the `ggplot2::diamonds` data set. You will pipe this data set to the functions needed to operate on the data frame. Take a look a the structure of the data frame so that you have an understanding of the variables therein. Make note of any variables that may require some modifying.

Taking Inventory of Variables

Examine the structure of your data frame so that you know variables that you might summarize and the variables you might group by.

Make a note of whether the data frame is *grouped* and thus contains a"grouping structure".

Mutating, Grouping, and Summarizing Data Frames

Grouping Data Frames

Take the data frame and pipe it to a *grouping* variable (e.g., `cut`) and then pipe to examine the *structure.* Make note of how the tibble's grouping structure has changed.

Mutating Variables without Grouping

Without using a grouping, *add a new variable* to the existing data frame that represents the mean of a *numeric* variable. Code your function to exclude `NA` values when computing the mean. Do not overwrite the data frame.

Describe what the mean represents:

Grouping and Mutating Data Frames

Using that same grouping structure used earlier (e.g., `cut`), *group* the data and then and *add a new variable* to the data frame that represents the mean of a *numeric* variable. Code your function to exclude `NA` values. Do not overwrite the data frame.

Describe what the mean represents:

Compare the values of the mean variable added to the data frame with and without the grouping. Is the calculated variable in one of the data frames more similar to what you expected when you mutated the variable? If so, which one?

Grouping and Summarizing Data Frames

Using that same grouping structure that you just used, rather than *add a new variable* to the data frame that represents the mean of the numeric variable, *summarize* the data frame. Code your function to exclude `NA` values. Do not overwrite the data frame.

Describe what the mean represents:

Grouping, Mutating, and Summarizing

Using that same grouping structure that you just used, *add a new variable* to the data frame that represents the mean of some numeric variable and *then summarize* the data frame by that *newly added numeric variable.* Code your function to exclude `NA` values using one of the approaches covered in the module. Do not overwrite the data frame.

Grouping, Summarizing, Grouping, and Summarizing

In many cases, you may need to group data frames in one way in order to obtain data summaries which will you will further summarize at a more general level. For example, you may need to aggregate your data in order to obtain average performance at a participant level so that you can further aggregate individuals within a group in service of obtaining group-level summaries.

In order to understand the difference in aggregation techniques, we will group the data two ways.

1. Take your data frame and (a) *group* by `cut` and (b) *summarize* the data frame so that the modified data frame contains the mean of some numeric variable but grouped by the grouping variable. Do not overwrite the data frame.
2. Next, take your data frame and (a) *group* by `cut` and `clarity`, (b) *summarize* the data frame so that your new data frame contains the mean of your numeric variable for each `cut` and `clarity`, (c) *group again* but only by `cut`, (d) *summarize by this new variable* so that your new data frame contains the mean of your numeric variable for each `cut`. Do not overwrite the data frame.

Describe what the differences in the summaries and why they might exist.

Practicing with Git

Remember that your Git commands will be in the terminal and not the RStudio console.

Save your File

Save the file you are working in. Make note of it's location and name.

Checkout your Feature Branch

If you are not working on a feature branch check it out. To verify your branch, use `git branch` in your terminal. If you are no *main*, checkout your feature branch using `git checkout`.

Stage your File Change

Use `git add <path to file>`

Commit Changes with a Message

Use `git commit -m "message to describe your change"`

Push Changes

Use `git push`

Bonus

1. Practice summarizing the data using different metrics (e.g., standard deviation, standard error of the mean, median, etc.).
2. Practice summarizing the data using different variables.
3. Practice summarizing the data by grouping the data different ways.