MATH 3190 Homework 1

Focus: Notes 1-3

Due February 10, 2024

Now its time to practice what we have learned in class and learn even more! Note that your homework should be completed in R Markdown or Quarto (you can just add your answers to this document in the appropriate part) and Knitted to an html document or pdf document. You will 'turn in' this homework by uploading to your GitHub Math_3190_Assignment repository in the Homework/Homework_1 directory.

Problem 1 (25 points)

Part a (20 points)

Write two functions called `ghist` and `gbox` that are similar to my `ggraph` function that you put in your myplots.in package from Lab 2. Remember that the "in" should be replaced with your initials. The `ghist` function should create a ggplot histogram of a variable that is given as a vector. The `gbox` function should create a ggplot box plot when a single numeric vector is given or it should create side-by-side box plots if one numeric and one categorical variables are given. Allow the user to indicate whether it should be horizontal or vertical box plots. Be sure to properly document these functions.

Part b (3 points)

Add those functions to your myplots.in package. Then run the `devtools::document()` function, update the DESCRIPTION file, and install your package to verify those functions work.

Part c (2 points)

Update your GitHub myplots.in repo with the updated package. This is only worth 2 points, but I cannot verify you did part a without this, so it is actually worth much more.

Problem 2 (60 points)

Part a (9 points)

Learn about the `read.fwf()` function for use in downloading data from a URL into **R**. Learn about tools for downloading files from external servers. The `widths` and `strip.white` options will be especially useful here. Use this function to download the scores for all college basketball games for the 2023-2024 season (http://kenpom.com/cbbga24.txt) and then convert it to a tibble (load the `tidyverse` package first). The second team listed per line is the home team. It is not clear what the numbers, letters, or city names indicate after the second listed score. Notice that this is a "live" file that gets updated every day! So, your tibble size may change if you work on this assignment over the course of several days. That's fine. Give the code you used to download these data.

Now lets practice using our tidy data/tidyverse tools! Using your `cbbga24` tibble, try doing the following:

Part b (2 points)

Use `rename()` to rename all of your variables to names that make sense.

Part c (2 points)

Use `mutate()` to create a new column that gives the score differences (team1−team2).

Part d (2 points)

Use `arrange()` to sort the data set by the home team.

Part e (2 points)

Use `select()` to remove the extra variable(s) that had that irrelevant information at the end of each line. Note: you can select every variable except one by using the "!".

Part f (2 points)

Put parts a-e all together in one piping expression (with 5 pipes) and save this as a new object in **R**.

Part g (3 points)

Use `filter()` to reduce the data down to only games played in 2023 (you could use the `lubridate` package for this, since it specializes in dealing with dates, but some base **R** packages will also work). Save this in a new tibble. **We will use this tibble with only the 2023 years from here on out**.

Part h (4 points)

Write a function that will filter the tibble to only games played by a given team. Demonstrate your function by displaying games played by SUU.

Part i (7 points)

Use `summarize()` to extract SUU's win/loss record and winning percentage for their 2023 games. Hint: using the `case_when()` function inside of a `mutate()` function to create a new variable that indicates whether SUU won or lost is helpful.

Part j (7 points)

Generalize this by writing a function that will do this for a given team, and create a tibble with this information for all teams. Arrange this tibble by winning percentage (descending). The `add_row()` function may be useful here.

Part k (8 points)

Write two functions that generate appropriate graphs for the basketball data. These two graphs could be anything you'd like and should use `ggplot2` and they should show something meaningful.

Part l (12 points)

Create an **R** package that contains your functions from Parts h, j, and k and your tibble that contains all the games from 2023. You can use the the `write_csv()` function to save your tibble as a .csv file and put it in a `data-raw` folder in your package. Make sure the functions are properly documented. Upload this package to your GitHub page and indicate here what you called this package.

Problem 3 (15 points)

Repeat parts b-f of Problem 2 using Python in R Markdown (or Quarto). First, pass the original object that you read in from the website to Python without any changes to it (you do not need to read the file from the web in Python, but you can if you'd like) and then use `pandas` to rename the columns as indicated in part b,

add the columns specified in part c, arrange the data as in part d, drop the "garbage" column as in part e, and filter it down as in part f. The `pandas` functions `rename`, `assign` (instead of `mutate`), `drop` (instead of `select`) and `str.contains` (used to select the right rows) will be useful here. Be sure to follow the guide in Notes 2 to properly install Python, install the `pandas` library and to load it in **R**.