ESM 206 – ASSIGNMENT 1

Get, wrangle, and visualize data in R

Due Wednesday 2019-10-09 at 8:00am (submit .R scripts online through GauchoSpace)

Read through the entire assignment before starting. **Follow instructions carefully.** There is a very detailed "What you will submit" section at the end of the assignment.
Submit exactly and only what is requested.

For Assignment 1, you will:

- Create a GitHub account (see instructions below)
- Read Broman & Woo and Lowndes et al. papers (links below)
- Complete 4 separate "tasks" in R, in which you will either create or modify an R script to practice basic data reading, wrangling and visualization skills

General notes for Assignment 1:

- You will not submit anything for the readings in this assignment, but you **are responsible for understanding** the reading material
- You will create a separate R project (.Rproj) and .R script for each task
- Do not change, simplify, or update the datasets **in any way before reading them into R**
- The instructions for mini-tasks will vary in detail to encourage creativity and independent coding, but you should **maintain good coding practices throughout (e.g. even if you aren't told so explicitly, your code should have a clear descriptive header, thorough annotation and nice formatting/spacing, etc.)**
- You will submit your assignment by uploading your 4 task .R scripts to GauchoSpace

A. CREATE A GITHUB ACCOUNT:

- Go to github.com
- Choose a good username. Yes, it matters. Read and follow the suggestions here: Jenny Bryan's *Happy Git with R*
- Create your GitHub account
- Remember or store your GitHub account information (username & password) – you're going to need it

B. READINGS:

Karl W. Broman and Kara H. Woo (2018). Data organization in spreadsheets. The American Statistician 72 (1): 2 – 10.

Lowndes, JS et al. (2017). Our path to better science in less time using open data science tools. Nature Ecology & Evolution.

C. R TASKS

Task 1. Global glacial volume loss and sea level rise

For Task 1, you will be creating a new .R script to wrangle and visualize global glacial volume loss from 1961 - 2003 using curated data (glacial_loss.csv) from the National Snow and Ice Data Center:

Dataset Publisher: NSIDC (National Snow and Ice Data Center)

Online Resource: http://nsidc.org/data/g10002.html

Dataset DOI: http://dx.doi.org/10.7265/N52N506F

Column descriptions:

- **year:** calendar year
- **europe - antarctica:** change in glacial volume (km^3) in each region that year
- **global_glacial_volume_change:** cumulative global glacial volume change (km^3), starting in 1961
- **annual_sea_level_rise:** annual rise in sea level (mm)
- **cumulative_sea_level_rise:** cumulative rise in sea level (mm) since 1961

To complete Task 1:

A. On your computer, create a folder where you will store all of your assignments for ESM 206. Within that folder, add a subfolder for Assignment 1.

B. Create a new R project (you might want to call this project something like *a1_task1*), that exists within the Assignment 1 folder you created above

C. Copy and paste the *glacial_loss.csv* file into your Task 1 .Rproj folder

D. Create a new R script within the Task 1 project and save the script as '*a1_task1_**firstname_lastname**.R*' (e.g. *a1_task1_allison_horst.R*)

E. Write a well-formatted, thoroughly annotated, clearly spaced and organized R script in which you:
    a. Add a well-formatted, descriptive header (e.g. name, date, brief title)
    b. Attach the tidyverse
    c. Read in *glacial_loss.csv* data, and assign it as *glacial_data*
    d. Create a new subset of *glacial_data* called *glacial_rise* that only includes the columns *year* and *cumulative_sea_level_rise*
    e. With the pipe operator (%>%), add a line of code to your code from (d) so that you only keep observations for years from 1961 to 1990

f. Create a basic scatterplot graph using ggplot2 depicting the cumulative rise in sea level due to global glacial volume loss from 1961 to 1990. In your graph code, you should at least (but feel free to try other customization):
    i. Update x- and y-axis labels, and add a main title that includes your last name at the end, like "Sea level rise due to glacial loss, 1961 – 1990 (HORST)"
    ii. Change the color of the graph points
    iii. Update to a non-default theme (e.g. theme_bw, theme_minimal, etc.)

F. Functionality check:
    a. **Save your script**
    b. Close the Task 1 project (File > Close Project)
    c. Reopen the project (by double clicking the .Rproj file in the project folder)
    d. Open the script if it doesn't come up automatically (by clicking on it in the 'Files' tab in RStudio)
    e. Run the entire script with Command + Shift + Return
    f. Was your graph recreated? **Yay! Your Task 1 .R script is ready to submit.**

TASK 2: Blood lead levels in children, St. Louis, MO (2010 – 2015 data)

For Task 2, you will explore relationships between race and elevated lead levels in blood tested from St. Louis, Missouri, children (observations taken 2010 – 2015).

Information and sources, compiled and shared by Chris Prener:

- Data and variables summary
- Reuters reporting on lead exposure and 2015 5-year American Community Survey estimates for City of St. Louis Census Tracts via American Fact Finder

A. Create a new project called '*a1_task2*' in the folder you created for Assignment 1. *Note: if you are already working in an R project, you can choose "Open project in a new session" to have multiple projects open simultaneously. Otherwise the one you were working in will close, but can be easily reopened by double-clicking the .Rproj file!*

B. Copy and paste the 'stl_lead.csv' data into your newly created *a1_task2* project folder

C. Create a new R script within your project, and save as '*a1_task2_**firstname_lastname**.R*'

D. Create a well-formatted, annotated script (with a header) in which you:
    a. Attach the tidyverse

b. Read in the stl_lead.csv data
c. Only keep columns for *Percent Elevated* (**pctElevated**: “percentage of children tested who had blood lead levels greater than or equal to 5 micrograms per deciliter” – considered “elevated” lead levels in blood), *Total Population* (**totalPop**: estimated total census tract population in 2015), and *Black Population* (**black**: estimate for black population in each census tract in 2015).
d. Add a new column in which you calculate the estimated percentage of each census tract population that black residents comprise (based on 2015 data)
e. Create a ggplot graph of the “% black population” variable you calculated above, vs. the % of children in each census tract with “elevated” blood lead levels.
f. Customize your graph by at least updating:
    i. X- and y- axis labels, and a brief and descriptive graph title (with your last name at the end)
    ii. The size and color of the points
g. As comments in your script below the graph code, describe in 1 – 2 sentences any general trends and interesting patterns that you learn from the graph.

E. Functionality check your script (to make sure it will run from scratch with Command + Shift + Enter when you reopen it, like in Task 1). Does it work? **Great, your Task 2 .R script is ready to submit!**

TASK 3: Troubleshooting R ‘diamonds’

Troubleshooting is a huge part of coding, and learning (a) what to look for, (b) how to find help online, and (c) how to read/use error messages is an important skill.

A. Create a new R project ‘*a1_task3’* in your Assignment 1 folder

B. Download the ‘error_script.R’ script, and copy into your project folder

C. Open error_script.R. Add YOUR NAME where indicated at the top of the script. Save the script as *a1_task3_**firstname_lastname**.R* (within the project folder).

D. Troubleshoot the code entirely, and where prompted at the end of problematic lines with `# ERROR(S):` , add a brief description of what the problem was (e.g. `# ERROR(S): was missing end parenthesis`).

E. Ensure that the summary table and graph are successfully created, and that you can run the entire corrected script (Command + Shift + Enter) with no errors.

F. Save your script, do a functionality check. Still working correctly? **Great! Your .R script for Task 3 is ready to submit.**

TASK 4: US Atlantic salmon imports

The 'atl_salmon_imports.csv' data contains estimates for U.S. imports of Atlantic salmon by volume (10^3 pounds) from different countries between 1989 and 2018.

Data source: [USDA Economic Research Service Aquaculture Trade Tables]

A. Create a new project for Task 4 (using the same naming system as in previous tasks), and start a new script *a1_task4_**firstname_lastname**.R*. Copy and paste the atl_salmon_imports.csv file into your new Task 4 project folder.

B. In your well-organized and annotated script, write code to complete the following:

   a. Calculate the total volume of Atlantic salmon that the US has imported from each country (10^3 pounds) for the 10 most recent years of data (2009 - 2018). *Note: remember to include an `na.rm = TRUE` argument when you calculate the group totals, so that countries with NA observations are still included in the summary table.
   b. For only the **top 5 total source countries** by total imported salmon volume from 2009 - 2018 (which you calculated in (a)), create a line graph to visualize annual US salmon imports **only from those top 5 source countries from 1989 - 2018**. Update axis labels, and add a title with your last name at the end.

C. Functionality check your script. Still works? **Great, your Task 4 .R script is ready to submit!**

What you will submit for Assignment 1:

Your four complete, organized, well-annotated and functionality-checked .R scripts.
That's IT. We will run your scripts in R to make sure they work - *which means it is very important that you DO NOT update the .csv files before reading them into R, otherwise we will be trying to run your script on files we don't have.*

What you will be graded on for Assignment 1:

The four tasks are worth 7 points each for a total of 28 points. Scores for each task are based on functionality (5 pts) and organization/good coding habits (2 pts).

- Functionality:
    - Works correctly = 5 points
    - Does not work correctly = 0 points
- Organization/good coding habits:
    - No organization/annotation = 0 points
    - Average organization/annotation = 1 point
    - Excellent organization/annotation = 2 points

How to submit your Assignment 1 scripts:

Upload your four scripts to the 'Submit Assignment 1' link on GauchoSpace. You are allowed to upload four files - those should be four .R scripts. Once you've submitted, congratulations! You're done with your first ESM 206 assignment.