Week 2 In-class Assignment

CHL8010: Statistical Programming and Computation in Health Data

In this assignment, we will build on the GitHub repo you created last week. Use version control to perform the following tasks. After performing all the tasks, pull+commit+push the changes to your repo.

Begin R project

1. Create an R project folder with a name that reflects the paper, such as `armed_conflict`. The name of the folder automatically will become the name of the GitHub repo. You can change the name later but it will be easier if you select a name that works best.
2. In the main directory, create subfolders to store different types of files. For now, create subfolders for `original`, `data` and `R`.
3. Download `maternalmortality.csv` from Quercus and save it in the `original` subfolder.
4. Write an R script that reads in the `maternalmortality.csv` data and does the following manipulations.
    a. Use the `select()` function in the `dplyr` package (which is one of the packages in the `tidyverse` bundle) to subset the data to have only the variables `Country.Name`, `X2000` – `X2019`.
    b. The data set is currently in a **wide** format. Use the `pivot_longer()` function to convert the data set into a **long** format. So that the first and last 20 rows of the resulting data set look like this. Hint: You need to select the columns `X2000` to `X2019`, **remove the prefix X** from them, change the **name of the variable** to `Year`, change the **values** to `MatMor`. Finally, make sure the `year` variable is stored as numeric.

```
# A tibble: 20 x 3
   Country.Name  year MatMor
   <chr>        <dbl>  <int>
 1 Aruba         2000     NA
 2 Aruba         2001     NA
```

```
 3 Aruba         2002     NA
 4 Aruba         2003     NA
 5 Aruba         2004     NA
 6 Aruba         2005     NA
 7 Aruba         2006     NA
 8 Aruba         2007     NA
 9 Aruba         2008     NA
10 Aruba         2009     NA
11 Aruba         2010     NA
12 Aruba         2011     NA
13 Aruba         2012     NA
14 Aruba         2013     NA
15 Aruba         2014     NA
16 Aruba         2015     NA
17 Aruba         2016     NA
18 Aruba         2017     NA
19 Aruba         2018     NA
20 Aruba         2019     NA
```

```
# A tibble: 20 x 3
   Country.Name  year MatMor
   <chr>        <dbl>  <int>
 1 Zimbabwe      2000    579
 2 Zimbabwe      2001    629
 3 Zimbabwe      2002    666
 4 Zimbabwe      2003    680
 5 Zimbabwe      2004    686
 6 Zimbabwe      2005    685
 7 Zimbabwe      2006    680
 8 Zimbabwe      2007    671
 9 Zimbabwe      2008    657
10 Zimbabwe      2009    632
11 Zimbabwe      2010    598
12 Zimbabwe      2011    557
13 Zimbabwe      2012    528
14 Zimbabwe      2013    509
15 Zimbabwe      2014    494
16 Zimbabwe      2015    480
17 Zimbabwe      2016    468
18 Zimbabwe      2017    458
19 Zimbabwe      2018     NA
20 Zimbabwe      2019     NA
```

c. We will complete the rest of the code next week.

Push to GitHub

3. If you haven't done so already, introduce yourself to `Git` by using the `usethis::use_git_config()` function.
4. Then use the `usethis::use_git()` and `usethis::use_github()` functions to push your project to a GitHub.
5. Go to your GitHub account and check out your new repo!
6. Close R studio and delete the folder that contains the R project from your local machine.

Use version control

7. Open R studio and follow the steps on slide 23 to open the repo as a new project.
8. Download `disaster.csv` from the course website and save it in the `original` folder. Create a new R script that reads in the data and does the following manipulations.
    a. Use the `filter()` function to subset the data set to only include years 2000–2019 and the disaster types "Earthquake" and "Drought"
    b. Subset the data set to only include the following variables: `Year`, `ISO`, `Disaster.type`.
    c. Create a **dummy variable** `drought` and another dummy variable `earthquake` such that:

```
# A tibble: 890 x 5
# Groups:   year, ISO [600]
    year ISO   Disaster.Type drought earthquake
   <int> <chr> <chr>           <dbl>      <dbl>
 1  2000 AFG   Drought             1          0
 2  2000 ARM   Drought             1          0
 3  2000 BIH   Drought             1          0
 4  2000 AZE   Earthquake          0          1
 5  2000 AZE   Drought             1          0
 6  2000 BGD   Earthquake          0          1
 7  2000 BGR   Drought             1          0
 8  2000 BOL   Drought             1          0
 9  2000 CHN   Earthquake          0          1
10  2000 CHN   Earthquake          0          1
# i 880 more rows
```

d. Notice that some countries that had more than one earthquakes/droughts a year have multiple entries in some years. Use the `group_by()` and `summarize()` functions to create a data set where only one row of observation exists for each country and each year, such that:

```
# A tibble: 600 x 4
    year ISO   drought earthquake
   <int> <chr>   <dbl>      <dbl>
 1  2000 AFG         1          0
 2  2000 ARM         1          0
 3  2000 AZE         1          1
 4  2000 BGD         0          1
 5  2000 BGR         1          0
 6  2000 BIH         1          0
 7  2000 BOL         1          0
 8  2000 CHN         1          1
 9  2000 COL         0          1
10  2000 CUB         1          0
# i 590 more rows
```

e. We will finish the code next week.

Pull + Commit + Push

9. Follow the steps on Slide 24 to update your GitHub repo.
10. You can delete/archive the R project folder on your local machine.

Enter your GitHub username on this spreadsheet