Assignment 5 - Data Acquisition, Cleaning, Organizing, and Exploring

MA 615/415

26Feb2018 / revised 4Mar2018

This is the final project for the TidyVerse section of the course. There are two parts to this assignment. In both parts, your job is to obtain data, clean and organize it, and explore it. Then, produce a slide presentaion and a Shiny application about what you have learned.

To put the assignment into perspective, imagine that you have been given the data – more or less as you have gotten it for this assignment. Your job is to prepare the data for analysis.

Please keep in mind that the purpose of this assignment is for you to exercise, extend, and demonstrate your ability to use the ideas and tools of the Tidyverse. You're working in groups. Help your teammates. If you or your group is stuck or has a question, contact Brian or Haviland after you have tried to work it out for yourself.

Over the last week of classes on Feb 26, Feb 28, and Mar 2, we have discussed and demonstrated:

- Access and manipulation of data in *veg1.xlsx*, the dats used in Part 2 of this assignment,
- Manipulation of date and time data,
- Using *dplyr* two table verbs to produce a full range of joins,
- A new package of Shiny UI widgets, *Shinydashboards*. See Package 'shinydashboard', the RStudio shiny dashboard tutorial, and the examples we worked on in class.

As you work on the assignment, be sure you review and use these and other tidyverse tools and methods we have discussed in this second section of the course.

Part 1 (50%)

Find the webside for NOAA Weather Station buoy 46035 at 57.026 N 177.738 W in the NOAA National Data Buoy Center.

Read, Clean, and organize the data to produce a time series composed of 30 years of daily **Air Temperature** and **Sea Temperature** readings recorded at noon for the time zone of Station 46035.

Visualize and explore the data. It seems reasonable that air temperature and sea temperature are related. Are they?

Some data may be missing. What should you do about missing data? See "An introduction to data cleaning with R" by Edwin de Jonge and Mark van der Loo in the **Contributed Documentation** section at the bottom of the CRAN R page.

Has the mean temperature changed over the past 30 years? What statistical methods can you use to test if the change is statistically significant?

You been instructed to use only one sample per day day out of 24 daily hourly temperature readings. Has your sampling affected your evaluation of temperature change? In what way? Explain and demonstrate.

Assemble your work on the Station 45035 data into an R-Driven slide presentation (no google slides or powerpoint, please). Also make a Shiny Dashboard that lets visitors to your dashboard explore what you have learned about buoy observations of air and water temperature in the southern Bering Sea.

Part 2 (50%)

The attached file *veg1.xlsx* was created using the USDA QuickStats system using the following parameters

- Program: Survey
- Sector: Environmental
- Group: Vegetables
- Commodity: Vegetables Totals, Vegetables Other, Broccoli, Brussels Sprouts, Cauliflower

All other parameters were left open.

The data was collected to gain insight about chemical treatments applied to food crops as fertilizer, insecticides, etc. Clean, visualize, explore the data. Produce a second slide presentation and Shiny dashboard for the *veg1* dataset.

Keep in mind that the work we did in class with the *veg1* dataset was intended to prepare you for this assignment. There is still work remaining to be done.

In class, we noted that some of the chemicals used on our food are classified RESTRICTED USE CHEMICALS. We isolated these chemicals in the *veg1* dataset and found technical information about their toxicity in ECOTOX, the Beta version of ECOTOX, and the EPA Chenical Dashboard.

Make a table of toxicity measurements (at least LD50 for a single experimental animal). Use this table and what you know about dplyr joins to augment your evaluation of chemical treatments applied to vegetables.

Rubric

Rubric: Assignment 5

	Item	Not Good	OK	Good	Great	Totals
		0.5	0.7	0.85	1	
50	Tidy data	Gross violation of tidy data definition.	Tidy. But, code does not use tidy data tools.	Tidy. Good use of tidy data tools.	Tidy. Functions and commentary indicate readiness for multiple similar tasks.	
20	Code	Code doesn't run.	Messy, inadequate comments, not reproducible, doesn't follow the style guide.	Acceptable appearance. Commentary merely adequate, not reproducible.	Easy to read. Reproducible.	
10	Statistics	Statistical analysis missing, flawed, or incomprehensible.	Descriptive graphics with lables.	Descriptive statistics with graphics.	Descriptive statistics with graphics and explanatory commentary.	
10	Slides	slides not readible, no logical flow, or fail to display.	Slides display in browser.	Slides display, have logical flow, good use of graphics.	Slides you want to stand in front of.	
10	Shiny	Missing or doesn't work.	Works from URL.	Would function well as a webpage that presents the data and the results of EDA.	Engaging.	