ECOLOGICAL DATA ANALYSIS IN R
DESCRIPTION - WEEK 4
Spring 2020-1

Dr. Luis Malpica Cruz
lmalpica@uabc.edu.mx

Week 4: Exercises on Tidy Data & Data Transformation
This week we will go full hands-on R, and use the tidying, transforming, and summarizing data skills that we learned last week through the tidyverse package.

Class 4.1: Workshop Activities
Tidying with dplyr
Basic data manipulation with dplyr

Class 4.2: Workshop Activities
Transforming and Summarizing data with dplyr
Handling dates in R, the lubridate package

Resources

- Wickham, Hadley (2014). "Tidy data". In: Journal of Statistical Software 59.1, pp. 1–23. DOI:10.18637/jss.v059.i10.

Also, chapters 11 and 12 of:

- Wickham, Hadley and Garrett Grolemund (2017). R for data science: visualize, model, transform, tidy, and import data. O'Reilly Media, p. 518. ISBN: 978-1491910399. http://r4ds.had.co.nz/index.html

Finally, see these cheat sheets as reference:

- https://github.com/rstudio/cheatsheets/raw/master/data-import.pdf
- https://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf
- https://github.com/rstudio/cheatsheets/raw/master/data-transformation.pdf

Details - Week 4
Class 4.1: Workshop Activities
We will try to integrate all that we learned over the past three weeks. Firstly you will create an R project in your computer and link it to a GitHub repository. Then you will look at one table with messy data (see lecture notes) your task is to make it tidy in R, you will have all the class time to work on it, team work and collaboration is encouraged (and bonus points if you ask around through Slack sharing MRE code)! I will be around to help if you get stuck (use the orange sticky note!). Try to avoid the use of excel!

Background:
You collected data of fish species in the Caribbean and have it in a long data format (messy) but need to be able to manipulate it further, mainly you need data on common names of fish species and counts from 2011, also, you do not need to know on which month were the data taken.

Suggested Step-by-Step Goals:

1) Create a GitHub repository and link it to a new R Project
2) Create the necessary subfolders and description files
3) Create a new Rmarkdown file (annotate & comment as necessary)
4) Explore and if necessary clean up the main data set provided (i.e. import to R, verify types of data, data structure, rename variables, etc.): "CompDec2011_2015.csv"

>>> Challenge: This file has a couple of bugs, hint: you can fix them on excel or explore the 'na.strings' option of 'read.csv'

5) Transform from wide to long format
6) One of your new columns will have two values, the first corresponds to a site #, the second to a year, split this column and name appropriately (hint, use: names_sep = "\\." within your 'pivot' function)
7) Eliminate all the rows that have no values (hint: use the values_drop_na option of 'pivot')
8) You need to select data from 2011 only (hint: remember 'filter'?)
9) You need to select all but the month column (hint: remember 'select'?)

>>> Challenge: pipe it out!

>>>Extra challenge, if you finish with time to spare, explore the true power of R and see how easy it is to implement this same routine on the other similar data sets available in the folder.

Class 4.2: Workshop Activities

Background:
You will again import and manipulate fish data. These were observations on different small coral reefs in the Caribbean. You need to estimate the total number of fish on any given date, per site, per species smaller than 20cm TL. You have three distinct data sets to do so, the first one 'FishAb.csv' contains counts and total lenght of fish per species and transect number; 'FishAbTrscts.csv' contains different type of data, specifically: Date of observation, site name/code, number of dive, name of observer, number of transect per site, time of dive, reef depth, temperature in C, benthic substrate of transect, weather conditions, visibility in ft, and current conditions; 'SppCodes.csv' has common names and a numeric code for every species. You would need to combine the three data sets by their common variables to be able to link all data used and estimate the total number of fish per date, sites, and species.

Goals:

1) Add a new Rmarkdown script to your existing RProject - GitHub repository
2) Load & explore the three new data sets: FishAb.csv, FishAbTrscts.csv, & SppCodes.csv
3) Join useful data sets to have species comon name, sites & date (FishAb.csv & FishAbTrscts.csv)

>>> Hint: remember '_join' functions in dplyr for this)

4) Create a summary table where we sum up all fish observations (counts or frequencies) per date, site and species common name

>>> (Hint: I recommend you follow this order of variables and use 'summarise' and 'group_by')