NSci253 - Lesson Plan 1

updated on 2/19/2018

Charles Jason Tinant

1/22/2018

Orientation

Hydrology is the study or knowledge of water. Understanding our societal need of water is important because at every point in time there are places in the world that experience excesses or deficits of water that have an adverse impact on society and ecosystems. Two growing areas of concern are an increasing human population that is stressing water resources, and global warming that is having an increasing impact on global food and water supplies.

Hydrology uses statistics to understand the average amount and variation of water as precipitation, soil water, ground water, stream flow, and water supply. Other uses of statistics in hydrology are to evaluate potential impacts, and to design and implement strategies to protect people and ecosystems. Statistical analysis is done using computer software. Rather than teach how to incorrectly use spreadsheets for statistics, in this course you will learn some of the basics of the R statistical programming language, and apply R to solve hydrology problems.

This week we are going to use global population data to learn some of the basic features of R. You can think of R as the statistical engine, and R Studio as seats, steering wheel, and windshield of your statistical vehicle to hydrology.

Weekly Readings

- Read syllabus
- Read Chapter 1; pages 1-12, 16, 20, 26-29.
- Read the r4beginners_v3.pdf in the Google Drive. You can also download the article from https://www.computerworld.com/article/2497143/business-intelligence/business-intelligence-beginner-s-guide-to-r-introduction.html

Homework Assignment

You can answer these questions right from RStudio in the Rmd file. Make sure you save your answers!

Question 1a: What are the variables (columns) of the world_pop dataset? Question 1b: What are the variables (columns) of the world_pop2 dataset? Question 1c: What are the variables (columns) of the world_pop3 dataset?

The question about variables is important because variables are groups of observations. The world_pop dataset variables are the different decades. In this case, observations are continents.

Question 2a: What is the mean population in 2000?
Question 2b: What is the median population in 2000? Question 2c: Why are they different? Hint: take a look at the graph. Question 2d: What is the mean population of all of the continents in all of the years? Hint: you will need to look at a different summary than the 'world_pop' dataset.

Question 3: What continent, other than Asia, is projected to grow substantially between 2000 and 2050? Question 4: What are some concerns that you might have as a water resources manager that this continent's population is growing so rapidly?

Reminders

- Make sure you keep up to date on attendance and homework (see syllabus)
- The textbook for this class is required (see syllabus)
- Please bring your laptop to class. Your own laptop is preferred over borrowing a computer.

Class Timeline

9:00 - Introductions / Discuss syllabus
10:20 - 10-minute break
10:30 - Set up Google Drive accounts
10:50 - Follow the r4beginners_v3.pdf to install R & R-Studio
11:40 - Summary discussion
11:50 - Feedback to instructor

Lecture Notes - Summary

We will be going over the 'r4beginners_v3.pdf' handout in Google Drive

Steps for Week 1 In-class and Homework

This set of steps sets up a working directory for your work.

1. If you have not done so already, follow the directions in 'Your first step' in the r4beginners.pdf document in the shared Google Drive class folder to install R and then RStudio. The document is in a folder has a sub-folder named 'Readings and Programs'. The folder also contains the most recent version of R and RStudio.
2. Copy the contents of the shared NSci253-students folder into a folder you will be using for class. If you are using your own computer (preferred), you can create a new folder in your own Google Drive. This will be your working directory. You should install the 'Backup and Sync' application from the Google Drive website to sync your Google Drive with a folder on your computer (if you do this, you might have to wait a little bit to have the cloud sync with your computer).
3. You are going to set your working directory in this step. Open R-Studio. Next, click on the 'Files' tab on the lower left panel in RStudio. Navigate to the place you moved the files in step 2. Set the working directory by clicking on 'More' -> 'Set as Working Directory'.
4. Some other things you can do from the 'Files' tab: You can copy files within RStudio by checking the box to the left of the 'Data' file. Next, click on the 'More' icon and then 'Copy'. You can move files within RStudio by checking the box to the left of the 'Data' file. Next, click on the 'More' icon and then 'Move'. You can rename files within RStudio by checking the box to the left of the 'Data' file. Next, click on the 'More' icon and then 'Rename'.
5. Make a copy the 'NSci253_LesPlan01.Rmd' file by checking the box next to the file, clicking 'More' -> 'Copy' and naming the copy 'your-name_LesPlan01'.
6. Open the 'your-name_LesPlan01' file you copied in the last step.
7. Next, run the code chunk below to install the 'tidyverse' package by putting your cursor on 'install.packages("tidyverse")' below and pressing Ctrl+enter. The package should download for you. The r4beginners_v3.pdf has more information on installing packages.

```
# This is the command to install a new package.
# install.packages("tidyverse") # only need to do this once.
```

The gray areas above and below are code chunks. They are the places you send requests to the R-engine.

```
# <- this is a hash-tag.  It tells R that you are going to make a comment.
# R will ignore comments, which is the text after the hash-tag.

# Packages are libraries of new commands.  This opens the 'Tidyverse' package.
library(tidyverse)

# Now, take a look at the tidyverse by using the dictionary.  One way to
# get to the dictionary is by using a '?' as shown below.  Ctrl+enter on
# '?tidyverse' below to see the description of the package.
# ?tidyverse

# You can also use the 'Help' tab in the lower left panel and type in the
# command or package your wondering about in the search bar.
```

11. Next, we are going to read in an existing dataset. While other types of data sets can also be read into R, most people (myself included) prefer the .csv format. The dataset is world population projections from 2000 to 2050. I chose this dataset to highlight that the population of the world is growing, but our water resources are remaining constant. The variable 'pop' in the data is the population in millions of people.

```
# This is a general statement to read in a csv file
# mydata <- read.csv("path to filename.csv")

# we are going to bring in three different looks of the same
# world population data.  The benefit of setting a working directory
# is that you don't need to specify a long path to tell R where
# your data is at.

# Ctrl+Enter to read in the three looks at world population data.
world_pop <- read_csv("world_pop.csv")
world_pop2 <- read_csv("world_pop2.csv")
world_pop3 <- read_csv("world_pop3.csv")
```

12. Now that you have your data tidied and into R (which is often the hardest part) you can take a look at the data, and summarize it.

This is for 'world_pop'.

```
head(world_pop) # shows the head (top) of the dataset
```

```
## # A tibble: 6 x 7
##   Region        `2000` `2010` `2020` `2030` `2040` `2050`
##   <chr>          <int>  <int>  <int>  <int>  <int>  <int>
## 1 Asia            3691   4133   4531   4841   5049   5167
## 2 Africa           803   1015   1261   1532   1827   2138
## 3 Europe           730    734    731    718    698    671
## 4 North America    486    539    595    648    695    739
## 5 South America    348    396    440    477    504    520
## 6 Oceania           30     35     39     43     46     49
```

```
str(world_pop) # shows the types of varables in the dataset
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    6 obs. of  7 variables:
##  $ Region: chr  "Asia" "Africa" "Europe" "North America" ...
##  $ 2000  : int  3691 803 730 486 348 30
```

```
##  $ 2010  : int  4133 1015 734 539 396 35
##  $ 2020  : int  4531 1261 731 595 440 39
##  $ 2030  : int  4841 1532 718 648 477 43
##  $ 2040  : int  5049 1827 698 695 504 46
##  $ 2050  : int  5167 2138 671 739 520 49
##  - attr(*, "spec")=List of 2
##   ..$ cols   :List of 7
##   .. ..$ Region: list()
##   .. .. ..- attr(*, "class")= chr  "collector_character" "collector"
##   .. ..$ 2000  : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ 2010  : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ 2020  : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ 2030  : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ 2040  : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   .. ..$ 2050  : list()
##   .. .. ..- attr(*, "class")= chr  "collector_integer" "collector"
##   ..$ default: list()
##   .. ..- attr(*, "class")= chr  "collector_guess" "collector"
##   ..- attr(*, "class")= chr "col_spec"
```

```r
# You can also view the data in a spreadsheet format
# by uncommenting the line below.
# View(world_pop) # shows the data in a 'spreadsheet' format

summary(world_pop) # creates a summary of the data
```

```
##     Region              2000            2010            2020
##  Length:6           Min.   :  30.0   Min.   :  35.0   Min.   :  39.0
##  Class :character   1st Qu.: 382.5   1st Qu.: 431.8   1st Qu.: 478.8
##  Mode  :character   Median : 608.0   Median : 636.5   Median : 663.0
##                     Mean   :1014.7   Mean   :1142.0   Mean   :1266.2
##                     3rd Qu.: 784.8   3rd Qu.: 944.8   3rd Qu.:1128.5
##                     Max.   :3691.0   Max.   :4133.0   Max.   :4531.0
##       2030            2040            2050
##  Min.   :  43.0   Min.   :  46.0   Min.   :  49.0
##  1st Qu.: 519.8   1st Qu.: 551.8   1st Qu.: 557.8
##  Median : 683.0   Median : 696.5   Median : 705.0
##  Mean   :1376.5   Mean   :1469.8   Mean   :1547.3
##  3rd Qu.:1328.5   3rd Qu.:1544.8   3rd Qu.:1788.2
##  Max.   :4841.0   Max.   :5049.0   Max.   :5167.0
```

Below is for 'world_pop2'.

```r
world_pop2 <- read.csv("world_pop2.csv")
head(world_pop2) # shows the head of the dataset
```

```
##   year Africa Asia Europe North.America Oceania South.America
## 1 2000    803 3691    730           486      30           348
## 2 2010   1015 4133    734           539      35           396
## 3 2020   1261 4531    731           595      39           440
## 4 2030   1532 4841    718           648      43           477
```

```
## 5 2040   1827 5049    698           695      46           504
## 6 2050   2138 5167    671           739      49           520
```

```
str(world_pop2) # shows the types of varables in the dataset
```

```
## 'data.frame':    6 obs. of  7 variables:
##  $ year         : int  2000 2010 2020 2030 2040 2050
##  $ Africa       : int  803 1015 1261 1532 1827 2138
##  $ Asia         : int  3691 4133 4531 4841 5049 5167
##  $ Europe       : int  730 734 731 718 698 671
##  $ North.America: int  486 539 595 648 695 739
##  $ Oceania      : int  30 35 39 43 46 49
##  $ South.America: int  348 396 440 477 504 520
```

```
# View(world_pop2) # shows the data in a 'spreadsheet' format

summary(world_pop2) # creates a summary of the data
```

```
##       year          Africa          Asia          Europe
##  Min.   :2000   Min.   : 803   Min.   :3691   Min.   :671.0
##  1st Qu.:2012   1st Qu.:1076   1st Qu.:4232   1st Qu.:703.0
##  Median :2025   Median :1396   Median :4686   Median :724.0
##  Mean   :2025   Mean   :1429   Mean   :4569   Mean   :713.7
##  3rd Qu.:2038   3rd Qu.:1753   3rd Qu.:4997   3rd Qu.:730.8
##  Max.   :2050   Max.   :2138   Max.   :5167   Max.   :734.0
##  North.America      Oceania      South.America
##  Min.   :486.0   Min.   :30.00   Min.   :348.0
##  1st Qu.:553.0   1st Qu.:36.00   1st Qu.:407.0
##  Median :621.5   Median :41.00   Median :458.5
##  Mean   :617.0   Mean   :40.33   Mean   :447.5
##  3rd Qu.:683.2   3rd Qu.:45.25   3rd Qu.:497.2
##  Max.   :739.0   Max.   :49.00   Max.   :520.0
```

Below is for 'world_pop3'.

```
world_pop3 <- read.csv("world_pop3.csv")
head(world_pop3) # shows the head of the dataset
```

```
##   region year  pop
## 1 Africa 2000  803
## 2 Africa 2010 1015
## 3 Africa 2020 1261
## 4 Africa 2030 1532
## 5 Africa 2040 1827
## 6 Africa 2050 2138
```

```
str(world_pop3) # shows the types of varables in the dataset
```

```
## 'data.frame':    36 obs. of  3 variables:
##  $ region: Factor w/ 6 levels "Africa","Asia",..: 1 1 1 1 1 1 2 2 2 2 ...
##  $ year  : int  2000 2010 2020 2030 2040 2050 2000 2010 2020 2030 ...
##  $ pop   : int  803 1015 1261 1532 1827 2138 3691 4133 4531 4841 ...
```

```
# View(world_pop3) # shows the data in a 'spreadsheet' format

summary(world_pop3) # creates a summary of the data
```

```
##           region       year           pop
```

```
##  Africa       :6   Min.   :2000   Min.   :  30.0
##  Asia         :6   1st Qu.:2010   1st Qu.: 467.8
##  Europe       :6   Median :2025   Median : 696.5
##  North America:6   Mean   :2025   Mean   :1302.8
##  Oceania      :6   3rd Qu.:2040   3rd Qu.:1328.8
##  South America:6   Max.   :2050   Max.   :5167.0
```

```r
# This is a plot using a package called ggplot.
# ggplot uses grammer of graphics.  I <3 ggplot!

ggplot(world_pop3, aes(year, pop, color = region)) +
  geom_point() +
  geom_smooth(method = 'lm') +
  theme_bw()
```