WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Center for Limnology
University of Wisconsin–Madison

A Short Introduction into Data Science with R

Robert Ladwig
University of Wisconsin-Madison

rladwig2@wisc.edu

@hydrobert

Disclaimer

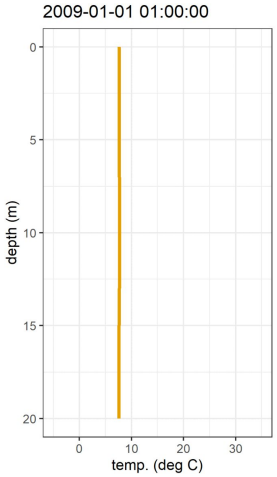
- material is based on the AWESOME workshop by Rachel Pilla (now postdoc at Oak Ridge National Lab, @rmpilla) “Introduction to R” (*Computer Science in Modern Biology* at Miami University, Ohio)
- with help by Andrew Cannizzarro, Alva Strand and Nicole Berry

About myself

- **Physical limnologist**
  - use mathematical models to explore mixing dynamics and water quality in lakes

Aquatic ecosystem modeling

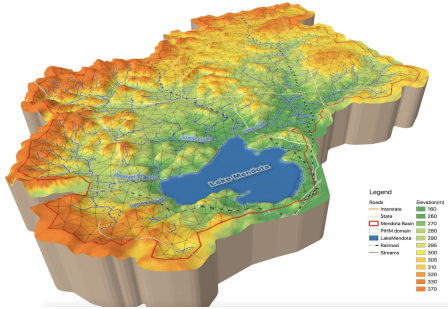
mechanistic understanding of cause and effects



$$\frac{\partial T}{\partial t} = \frac{1}{A}\frac{\partial}{\partial z}\left(A(\upsilon_t^{`} + \upsilon^{`})\frac{\partial T}{\partial z}\right) + \frac{1}{\rho_0 c_p}\frac{\partial H_{sol}}{\partial z} + \frac{dA}{dz}\frac{H_{geo}}{A\rho_0 c_p}$$

$$\frac{\partial [O_2]}{\partial t} = NEP = GPP - ER \pm D$$

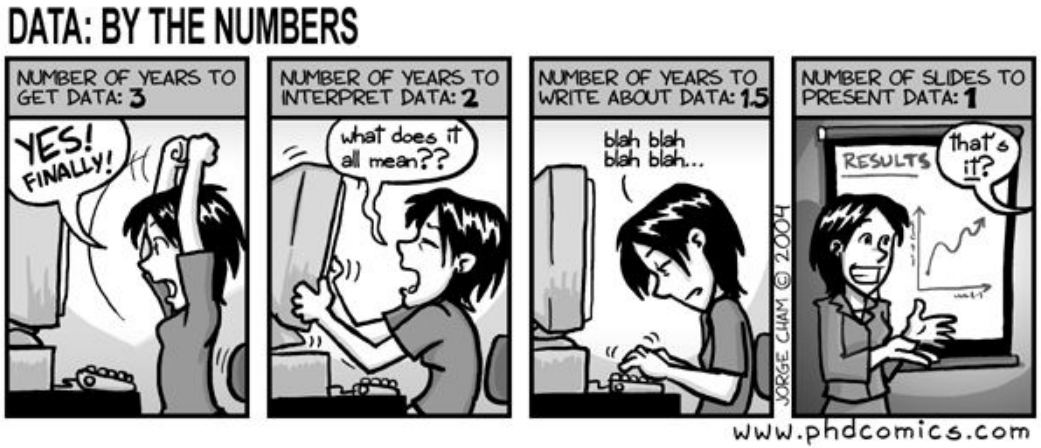
$$\frac{\partial P}{\partial t} = D\frac{\partial^2 P}{\partial z^2} + \lambda P(1 - \frac{P}{K})$$



- developing and applying open-source and open-access software in **R, Python and Matlab**
- background in civil engineering and geology

What's data science?

- combining math & statistics, programming and machine learning to explore any kind of data
- data is often noisy or unstructured → cleaning is big part of the job
- buzzwords like big data

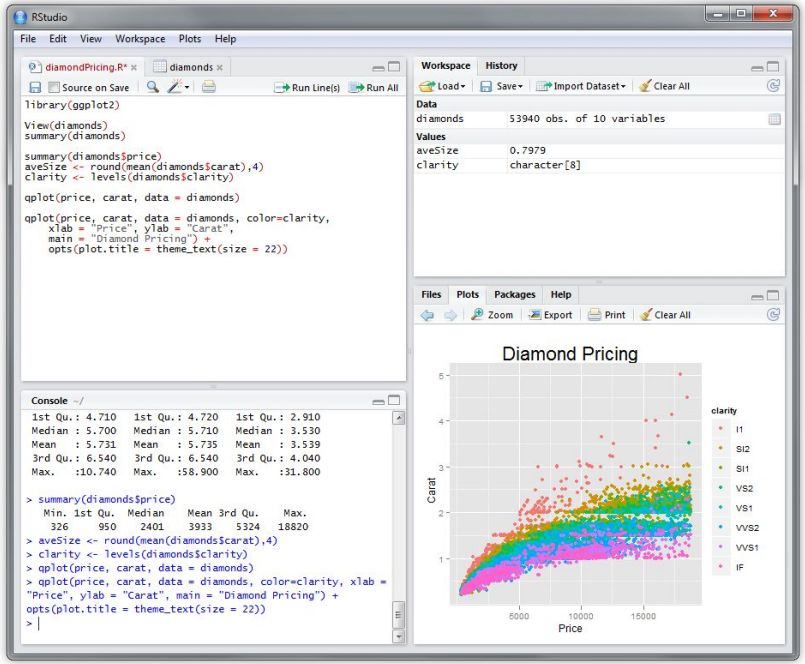


What's the plan for today?

- short introduction into the R scripting language
  - all material on github: https://github.com/robertladwig/introdatasciencer
  - GitHub: uses Git (version control), fantastic for coding and collaborations
- hands-on coding to introduce you to data analysis
- **at the end**: able to know basic R commands, load data, analyze data, plot data

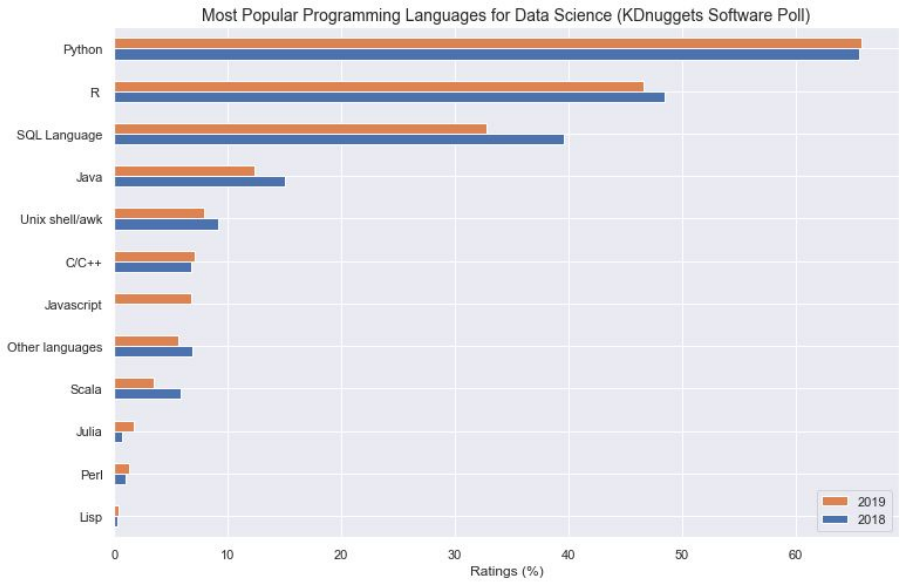
Install R and RStudio

- Install R from here (language): https://www.r-project.org/
- Install RStudio (editor and GUI): https://rstudio.com/products/rstudio/download/#download



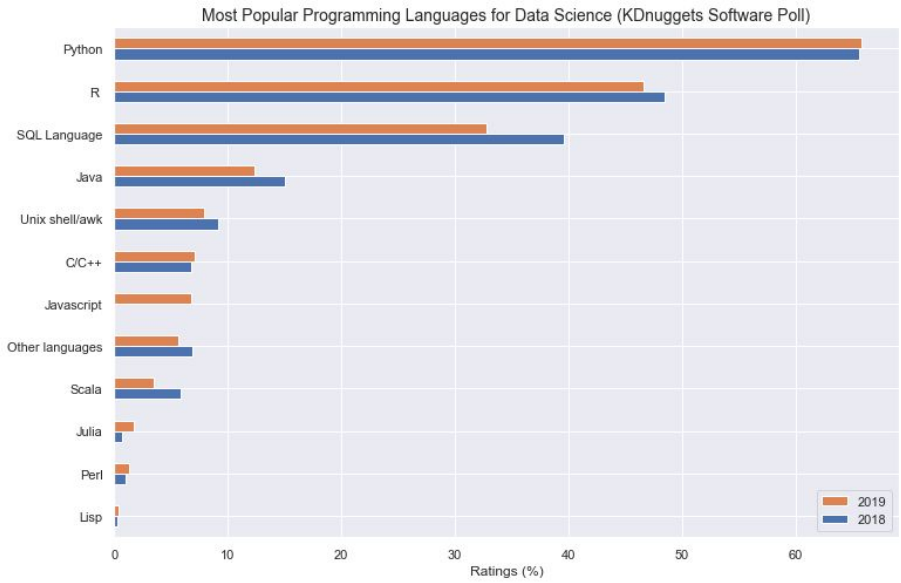
What's R?

- FREE and OPEN SOURCE statistical and computational software



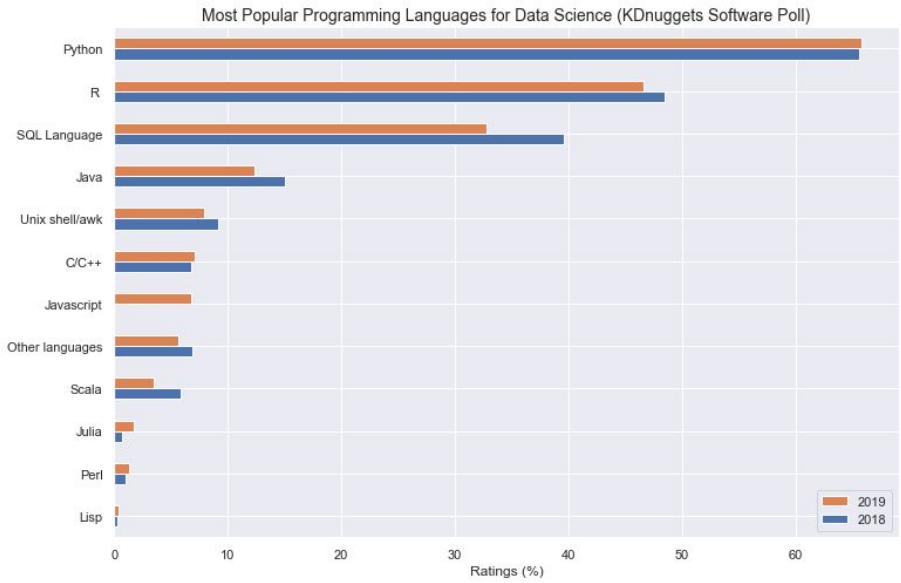
What's R?

- FREE and OPEN SOURCE statistical and computational software
- big community: easy to find solutions and troubleshooting online



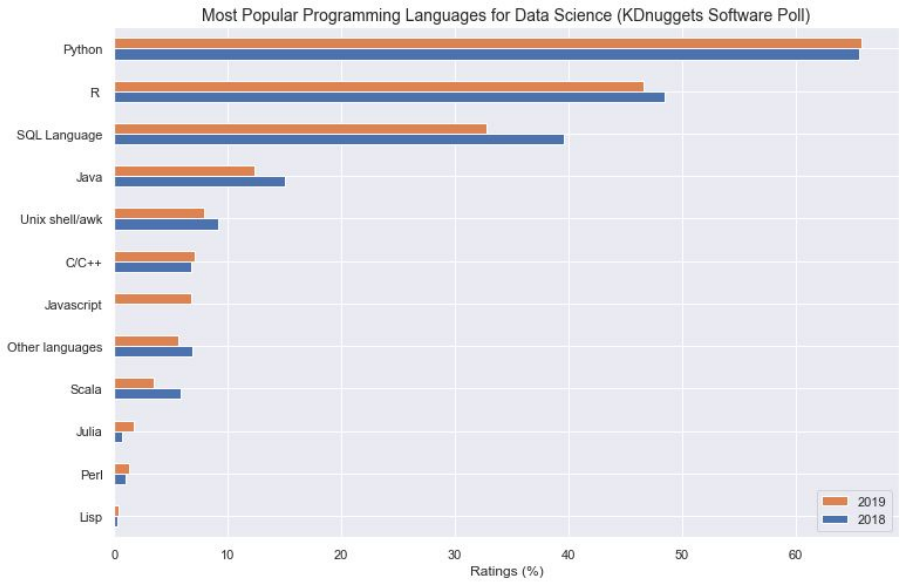
What's R?

- FREE and OPEN SOURCE statistical and computational software
- big community: easy to find solutions and troubleshooting online
- widely used in sciences (esp. statistics and visualization) and rapidly growing in popularity



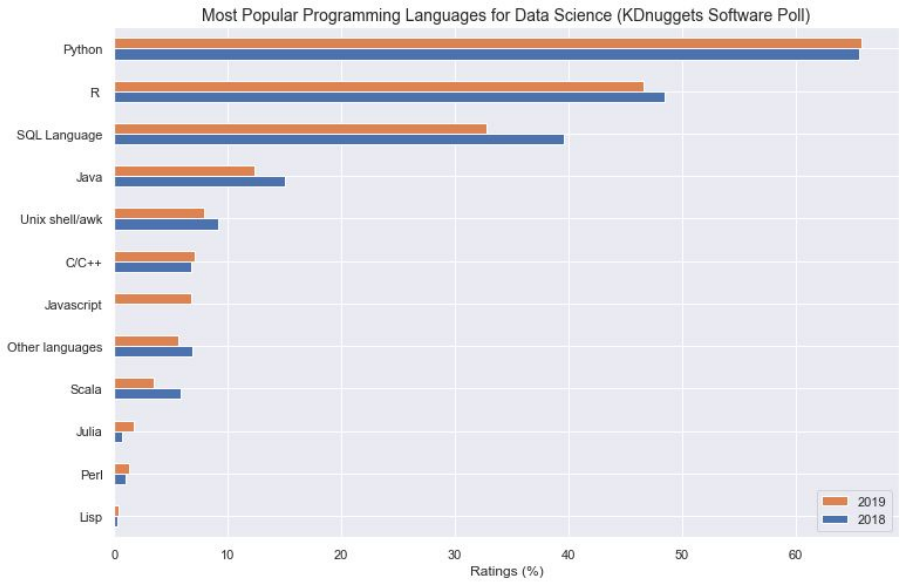
What's R?

- FREE and OPEN SOURCE statistical and computational software
- big community: easy to find solutions and troubleshooting online
- widely used in sciences (esp. statistics and visualization) and rapidly growing in popularity
- can handle more advanced computations, statistical analyses and bigger data files than Excel

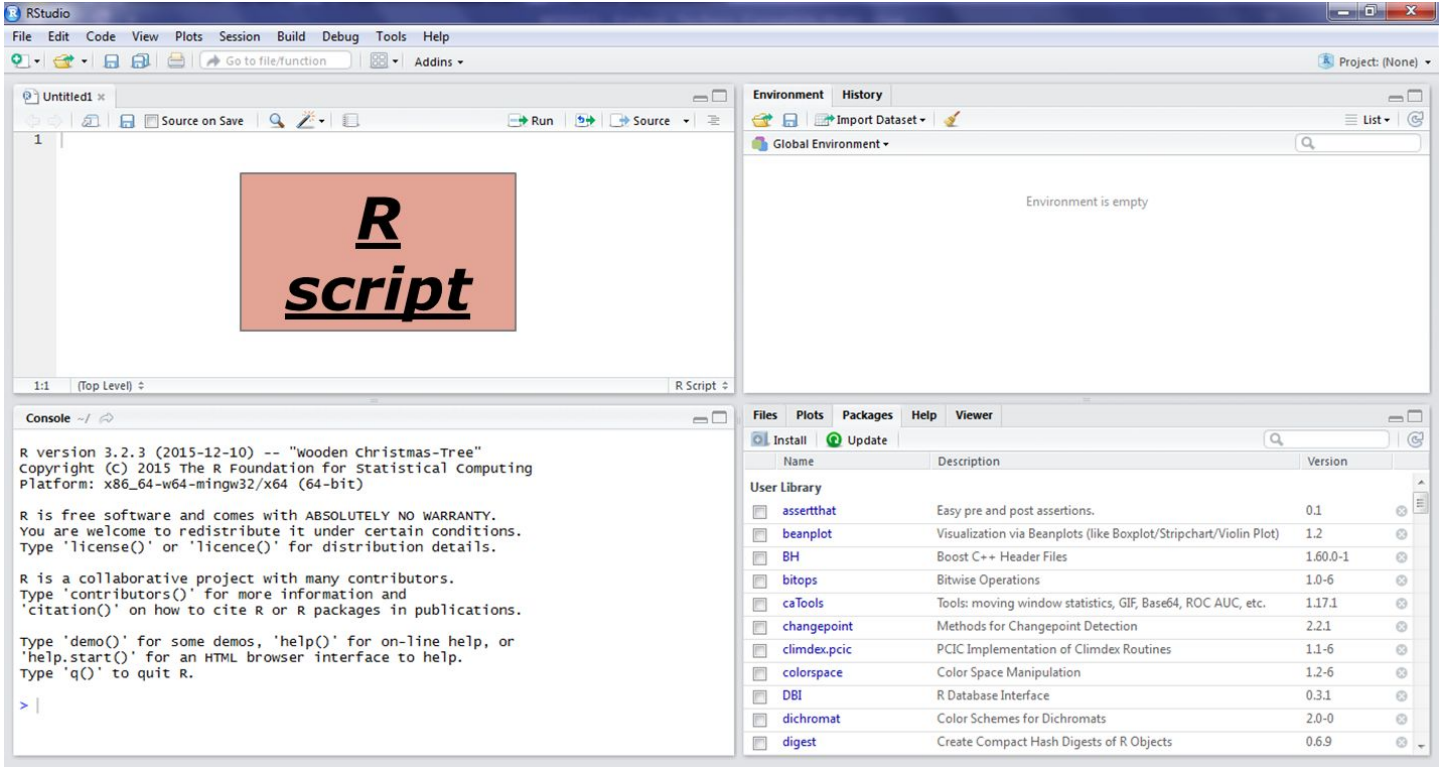


What's R?

- FREE and OPEN SOURCE statistical and computational software
- big community: easy to find solutions and troubleshooting online
- widely used in sciences (esp. statistics and visualization) and rapidly growing in popularity
- can handle more advanced computations, statistical analyses and bigger data files than Excel
- lots of styles for coding



Open RStudio on your laptop now



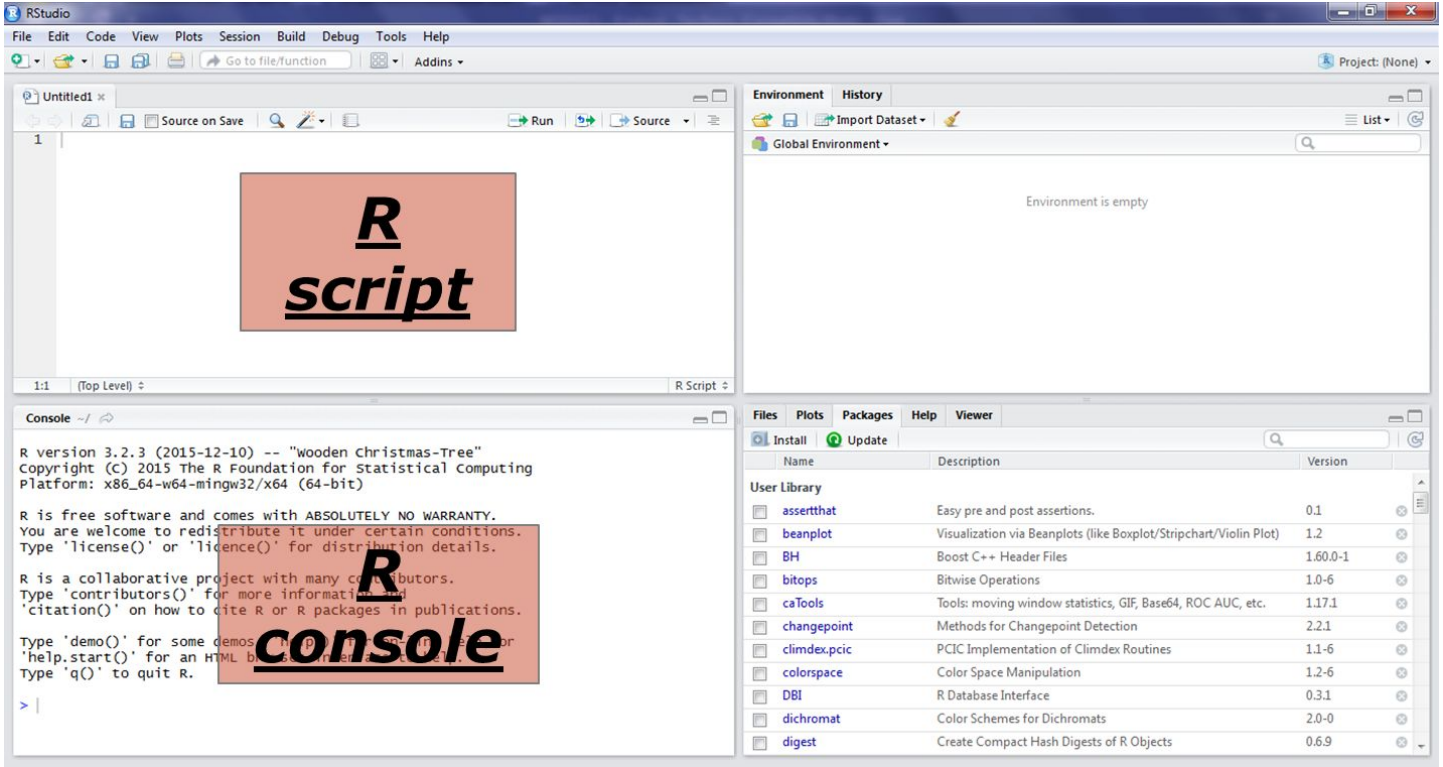
R script:

write and edit code

color-coding for easier identification

automatically fills in parentheses and quotations

Open RStudio on your laptop now

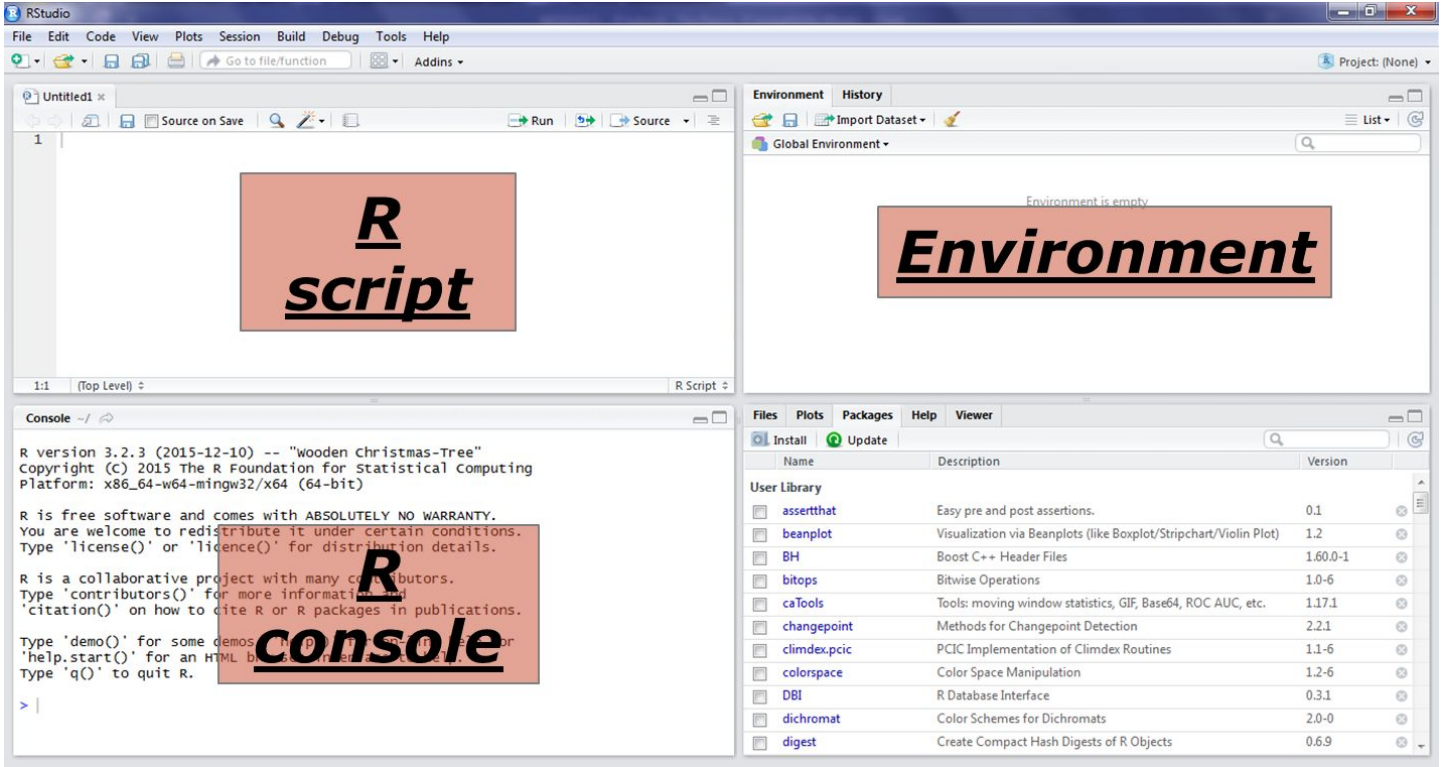


R console:

commands are run and results appear

> means ‘ready to work’

Open RStudio on your laptop now

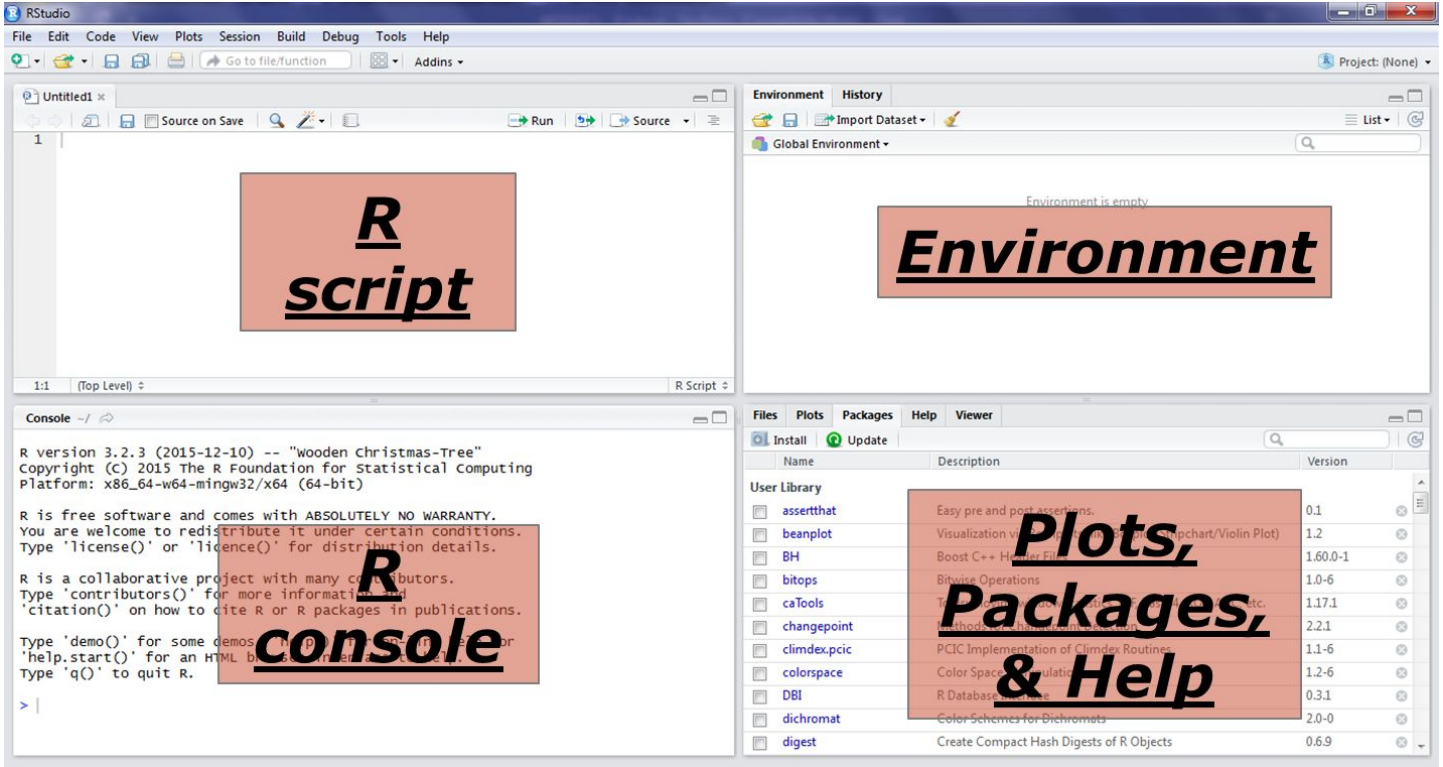


Environment:

see which objects were created

type of objects (double, integers, etc.), size and dimensions

Open RStudio on your laptop now



Plots/Packages/ Help:

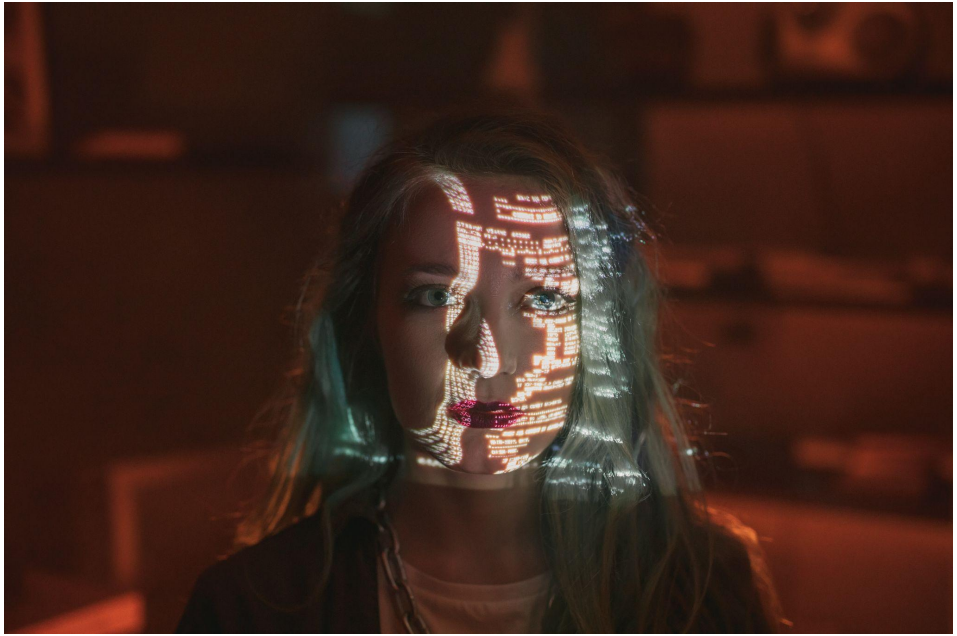
plots will appear

see which packages are installed

help functions

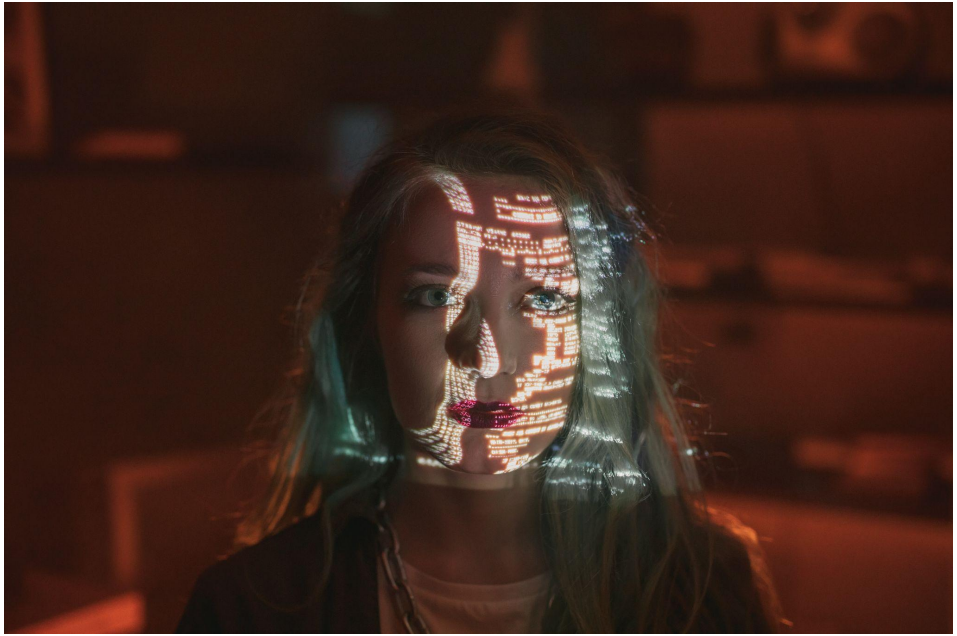
What are packages?

- bundles of tools and functions that others have developed to be used in R



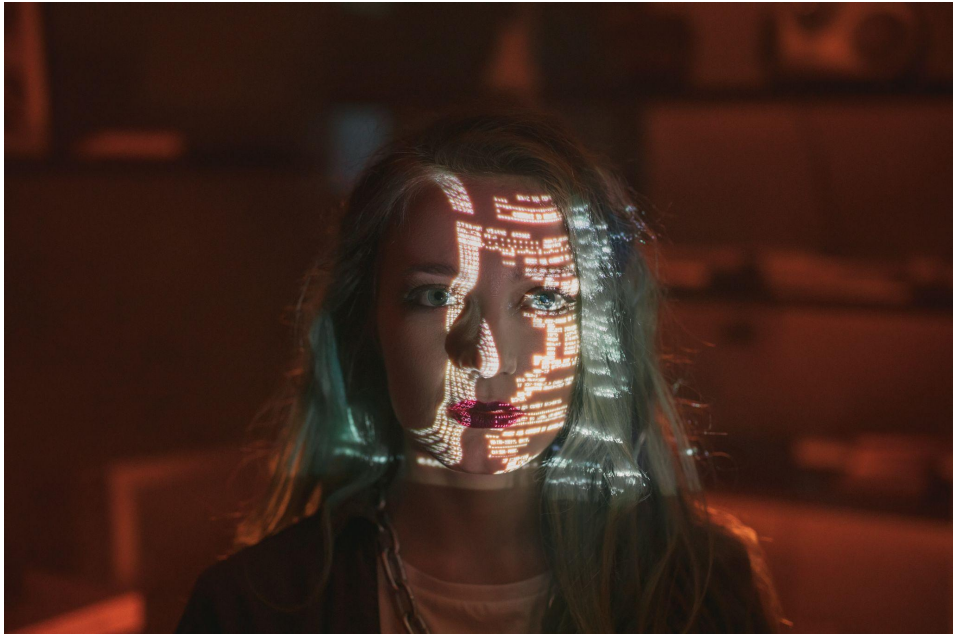
What are packages?

- bundles of tools and functions that others have developed to be used in R
- grouped to specific types of functions, analyses or datasets



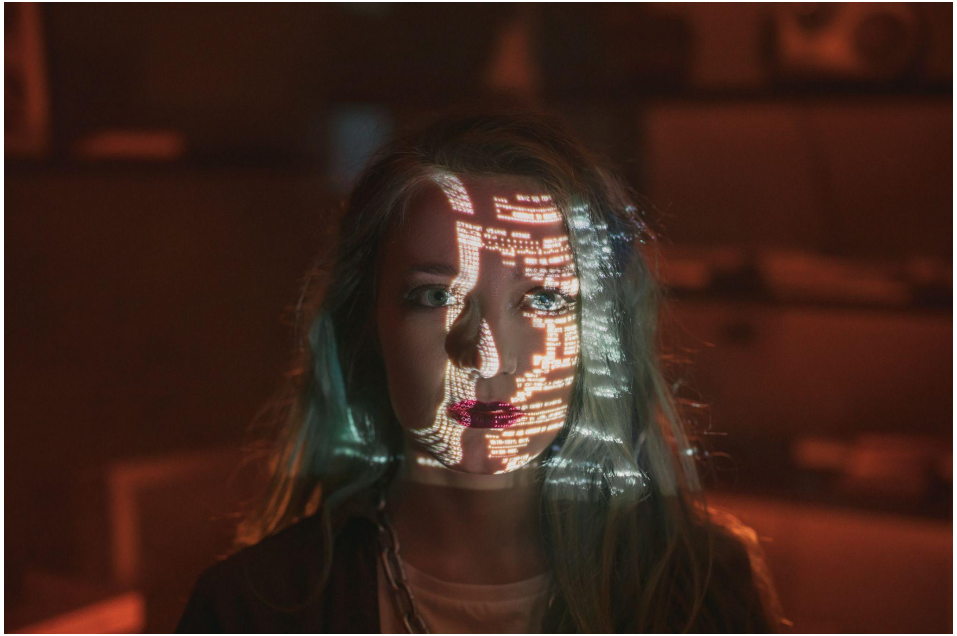
What are packages?

- bundles of tools and functions that others have developed to be used in R
- grouped to specific types of functions, analyses or datasets
- you can also create your own packages for specific tasks!

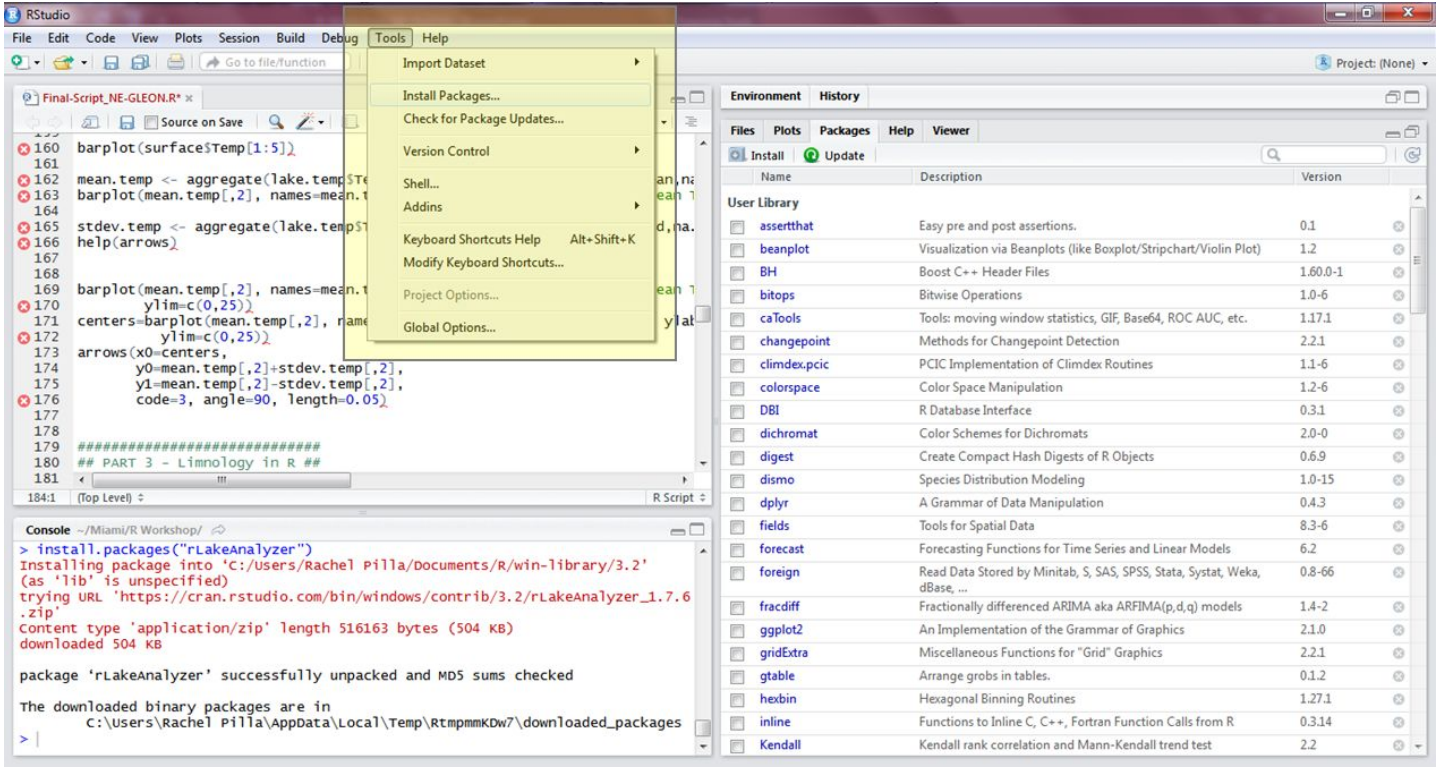


What are packages?

- bundles of tools and functions that others have developed to be used in R
- grouped to specific types of functions, analyses or datasets
- you can also create your own packages for specific tasks!
- currently, above 16,000 packages are available



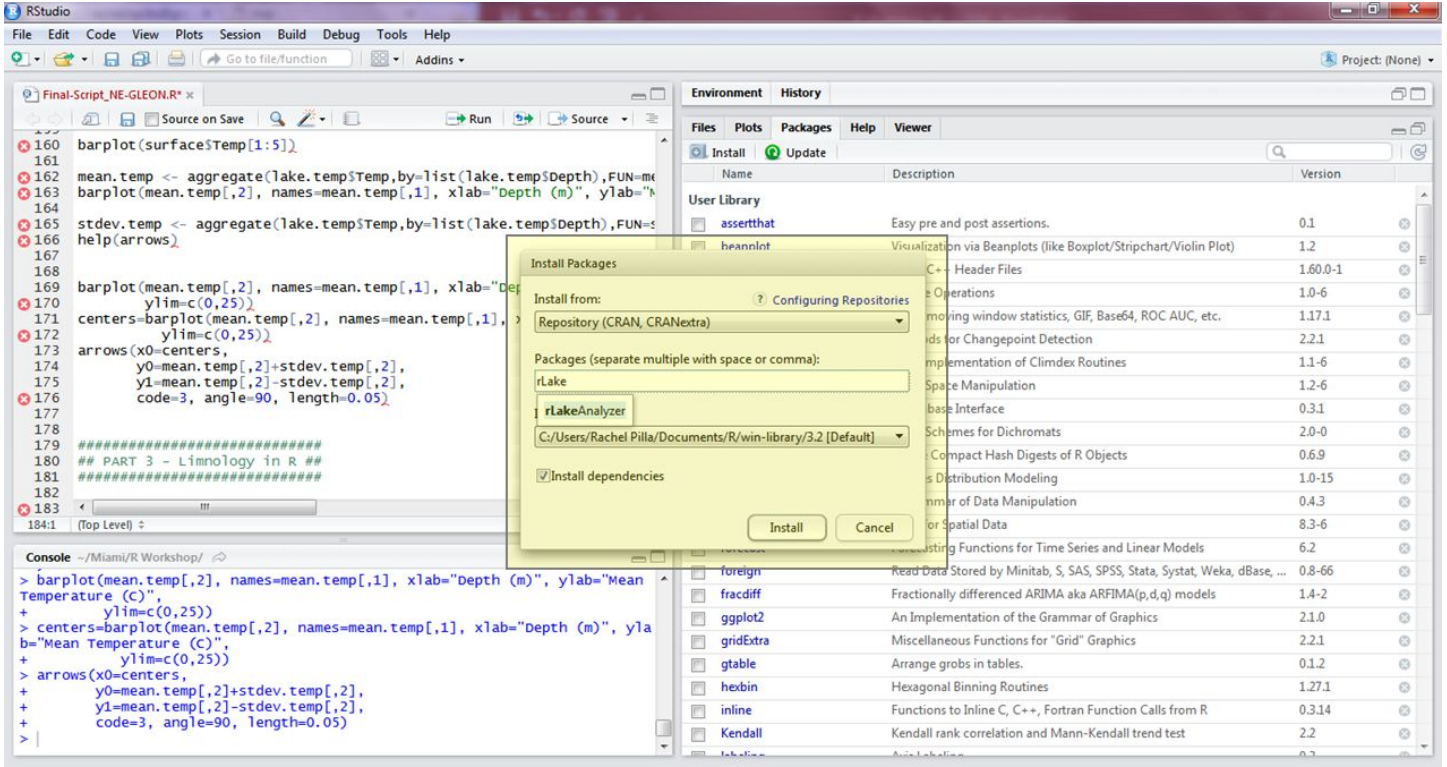
Install one package



we need:

tidyverse

Install one package



we need:

tidyverse

Install one package

RStudio

File Edit Code View Plots Session Build Debug Tools Help

Final-Script_NE-GLEON.R*

```
160 barplot(surface$Temp[1:5])
161
162 mean.temp <- aggregate(lake.temp$Temp,by=list(lake.temp$Depth),FUN=mean,na
163 barplot(mean.temp[,2], names=mean.temp[,1], xlab="Depth (m)", ylab="Mean T
164
165 stdev.temp <- aggregate(lake.temp$Temp,by=list(lake.temp$Depth),FUN=sd,na.
166 help(arrows)
167
168
169 barplot(mean.temp[,2], names=mean.temp[,1], xlab="Depth (m)", ylab="Mean T
170         ylim=c(0,25))
171 centers=barplot(mean.temp[,2], names=mean.temp[,1], xlab="Depth (m)", ylab
172           ylim=c(0,25))
173 arrows(x0=centers,
174        y0=mean.temp[,2]+stdev.temp[,2],
175        y1=mean.temp[,2]-stdev.temp[,2],
176        code=3, angle=90, length=0.05)
177
178
179 ############################
180 ## PART 3 - Limnology in R ##
181
```

184:1 (Top Level) R Script

Console ~/Miami/R Workshop/

```
> install.packages("rLakeAnalyzer")
Installing package into 'C:/Users/Rachel Pilla/Documents/R/win-library/3.2'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.2/rLakeAnalyzer_1.7.6
.zip'
Content type 'application/zip' length 516163 bytes (504 KB)
downloaded 504 KB

package 'rLakeAnalyzer' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\Rachel Pilla\AppData\Local\Temp\RtmpmmKDw7\downloaded_packages
>
```

Environment History

Files Plots Packages Help Viewer

Install Update

Name	Description	Version
User Library		
assertthat	Easy pre and post assertions.	0.1
beanplot	Visualization via Beanplots (like Boxplot/Stripchart/Violin Plot)	1.2
BH	Boost C++ Header Files	1.60.0-1
bitops	Bitwise Operations	1.0-6
caTools	Tools: moving window statistics, GIF, Base64, ROC AUC, etc.	1.17.1
changepoint	Methods for Changepoint Detection	2.2.1
climdex.pcic	PCIC Implementation of Climdex Routines	1.1-6
colorspace	Color Space Manipulation	1.2-6
DBI	R Database Interface	0.3.1
dichromat	Color Schemes for Dichromats	2.0-0
digest	Create Compact Hash Digests of R Objects	0.6.9
dismo	Species Distribution Modeling	1.0-15
dplyr	A Grammar of Data Manipulation	0.4.3
fields	Tools for Spatial Data	8.3-6
forecast	Forecasting Functions for Time Series and Linear Models	6.2
foreign	Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, Weka, dBase, ...	0.8-66
fracdiff	Fractionally differenced ARIMA aka ARFIMA(p,d,q) models	1.4-2
ggplot2	An Implementation of the Grammar of Graphics	2.1.0
gridExtra	Miscellaneous Functions for "Grid" Graphics	2.2.1
gtable	Arrange grobs in tables.	0.1.2
hexbin	Hexagonal Binning Routines	1.27.1
inline	Functions to Inline C, C++, Fortran Function Calls from R	0.3.14
Kendall	Kendall rank correlation and Mann-Kendall trend test	2.2

we need:

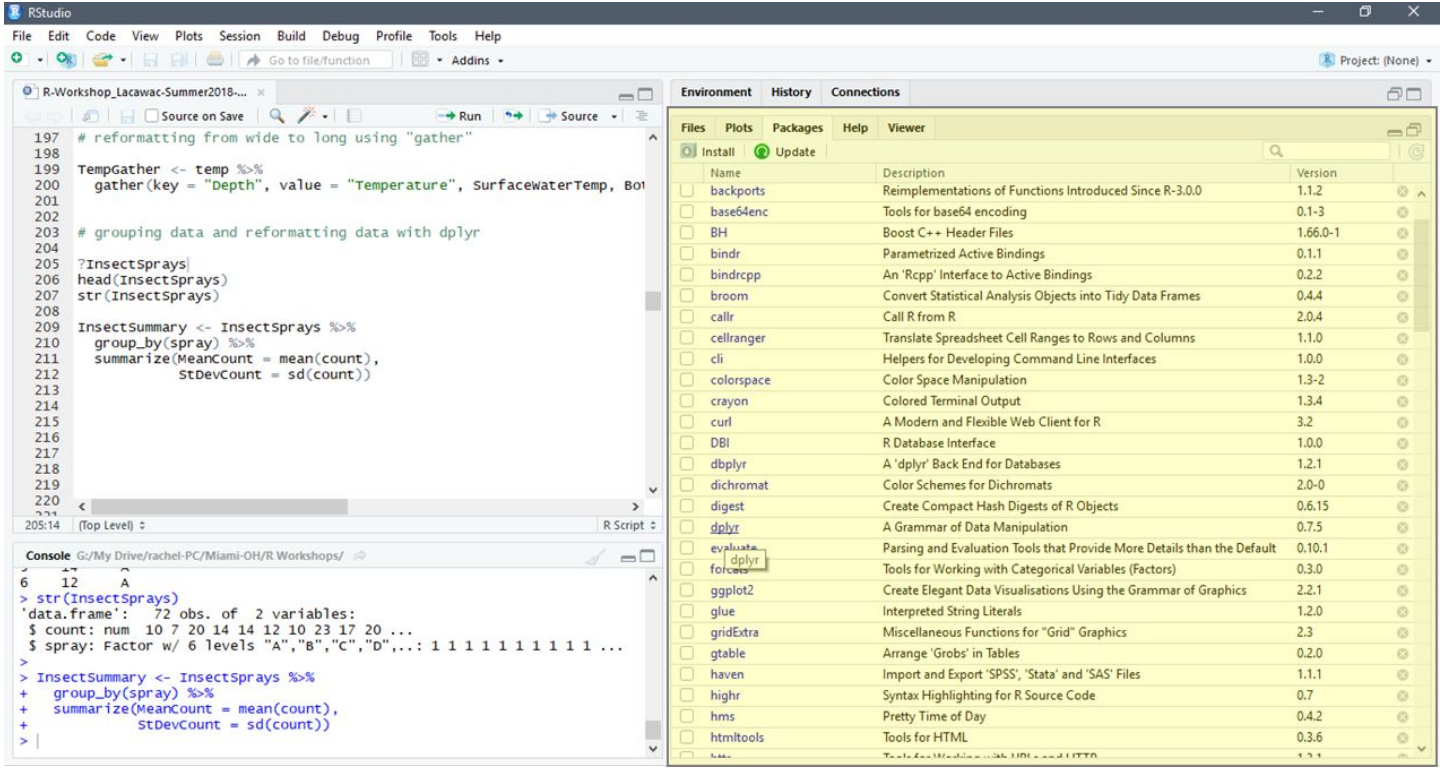
tidyverse

Installing a package

- packages only need to be installed ONCE
- but, if you want to use it you'll need to load it:

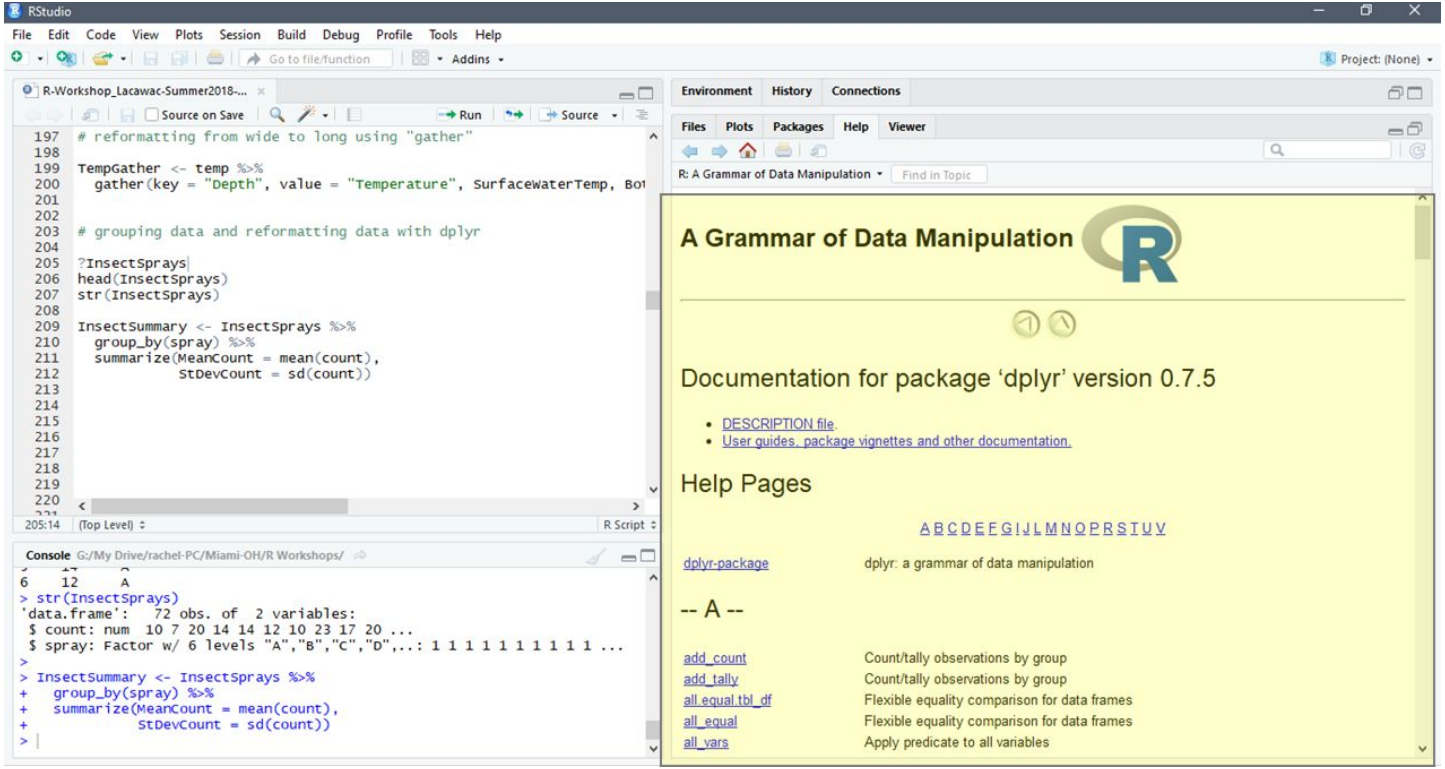
```
library(package name)
```

Package information



under packages, click on **dplyr**

Package information



this lists all available functions

click on any function to see the help file

Coding in R

- write commands in the script (upper left panel)
  - save it, edit it, revisit it later, etc.

Coding in R

- write commands in the script (upper left panel)
  - save it, edit it, revisit it later, etc.
- code NOT automatically run when you hit Enter

Coding in R

- write commands in the script (upper left panel)
  - save it, edit it, revisit it later, etc.
- code NOT automatically run when you hit Enter
- to run it:
  - “Run” button in upper right corner
  - “CTRL + Enter” (Windows)
  - “Command + Enter” (Mac)

Coding in R

- add comments using #
  - additional information
  - will be ignored

Coding in R

- add comments using #
  - additional information
  - will be ignored
- R is case sensitive
  - “Mean” ≠ “mean”

Coding in R

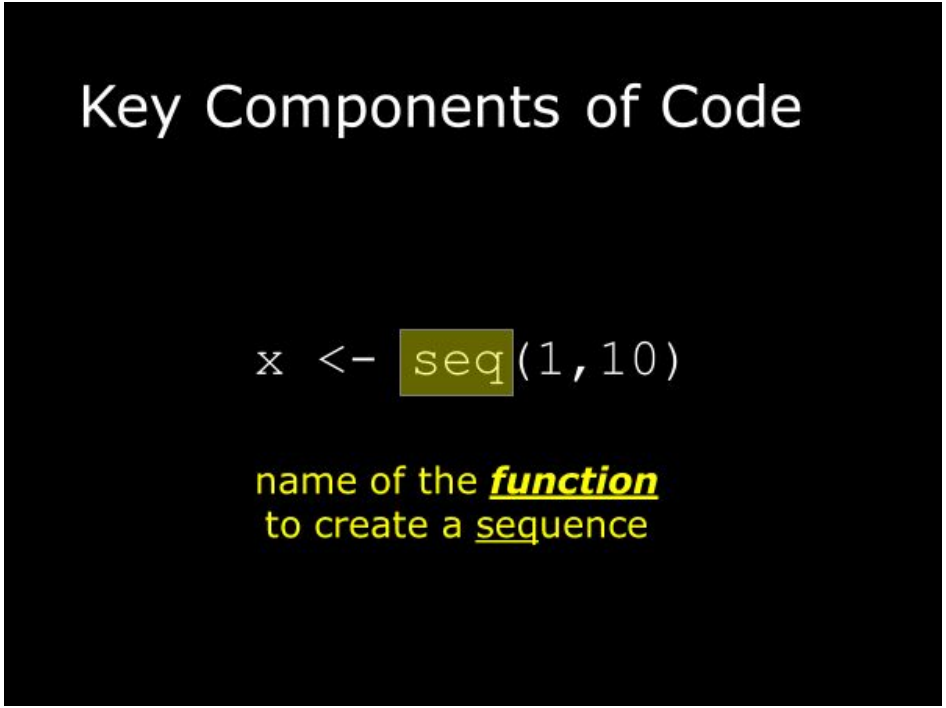
- add comments using #
  - additional information
  - will be ignored
- R is case sensitive
  - “Mean” ≠ “mean”
- R doesn’t care about spaces and tabs

Coding in R

- add comments using #
  - additional information
  - will be ignored
- R is case sensitive
  - “Mean” ≠ “mean”
- R doesn’t care about spaces and tabs
- need to close all parentheses and quotations

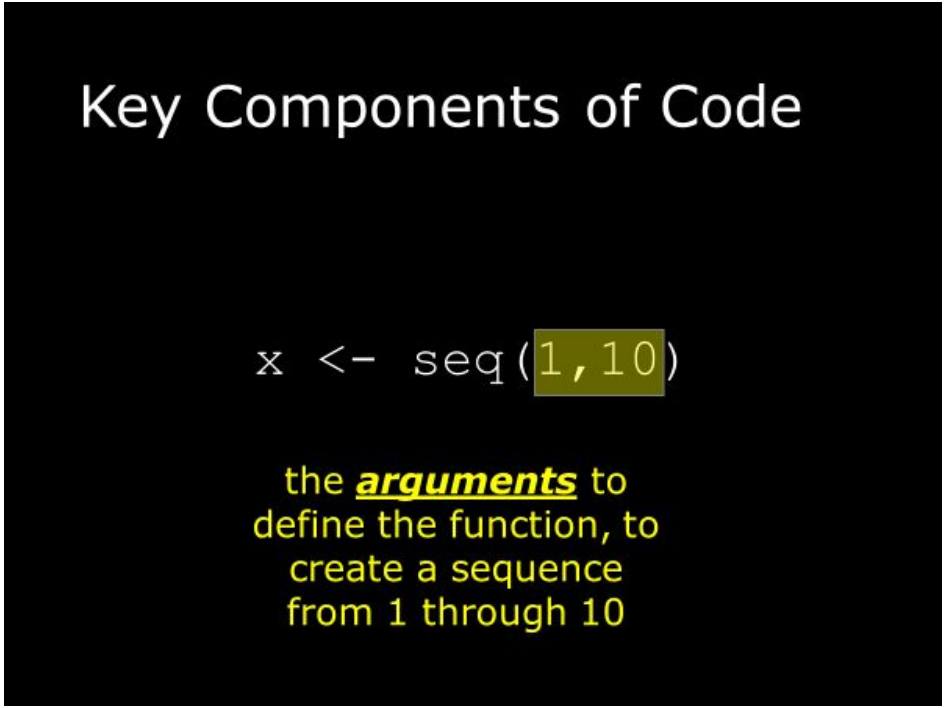
Key components

- **functions**: allow you to manipulate data, apply calculations, run statistical analysis, much more!



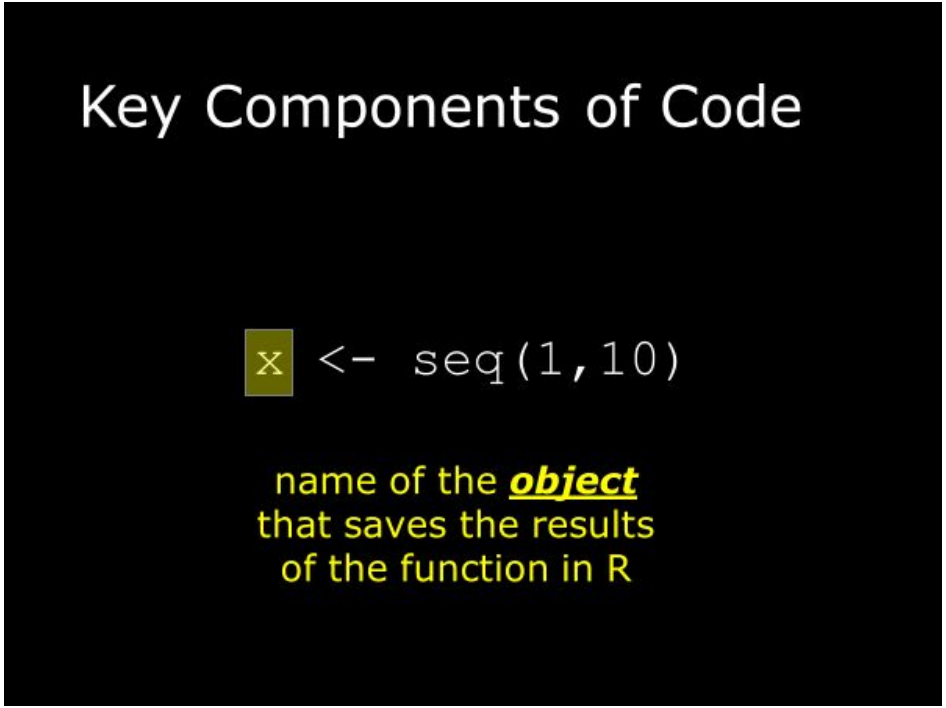
Key components

- **functions**: allow you to manipulate data, apply calculations, run statistical analysis, much more!
- **arguments**: defining information for functions, “customize” it



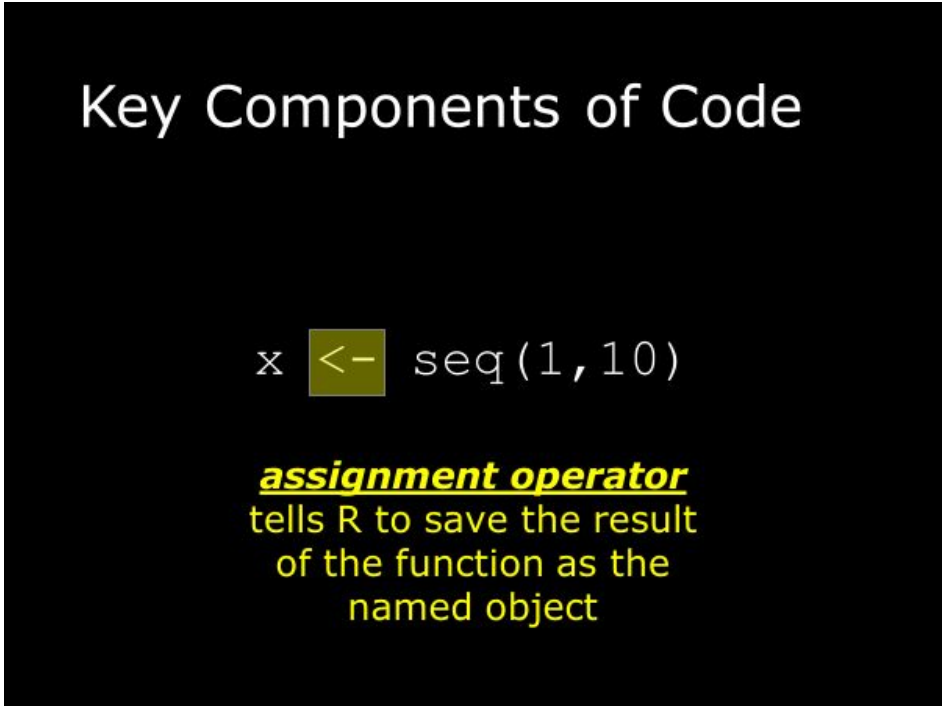
Key components

- **functions**: allow you to manipulate data, apply calculations, run statistical analysis, much more!
- **arguments**: defining information for functions, “customize” it
- **objects**: pieces of data saved in R, can be called up, reused and manipulated



Key components

- **functions**: allow you to manipulate data, apply calculations, run statistical analysis, much more!
- **arguments**: defining information for functions, “customize” it
- **objects**: pieces of data saved in R, can be called up, reused and manipulated



Types of data

- **objects (data types):**
  - numeric (2.334, 3.14159)
  - integer (2, 2405, 54)
  - characters (“hello world”)
  - logical (TRUE/FALSE)
  - complex (1+4*i*)

- **data structures:**
  - vector (1-D object with same data type)
  - matrix (2-D object with same data type
  - data frame (2-D object with different data types per column if needed, very useful!)

Live Coding time!