"Techné" skillshare meetup

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 Alva Strand and Nicole Berry

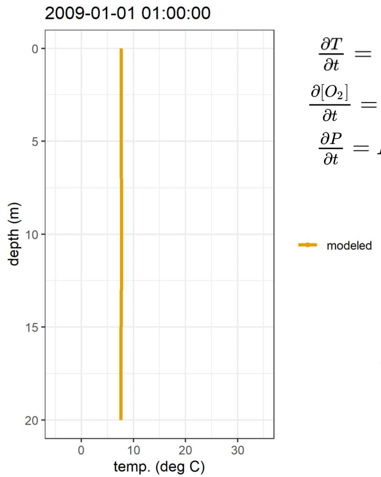
About myself

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

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

Aquatic ecosystem modeling

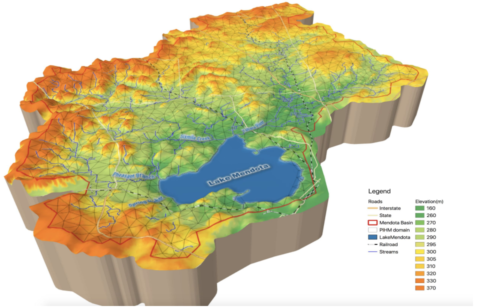
mechanistic understanding of cause and effects,



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

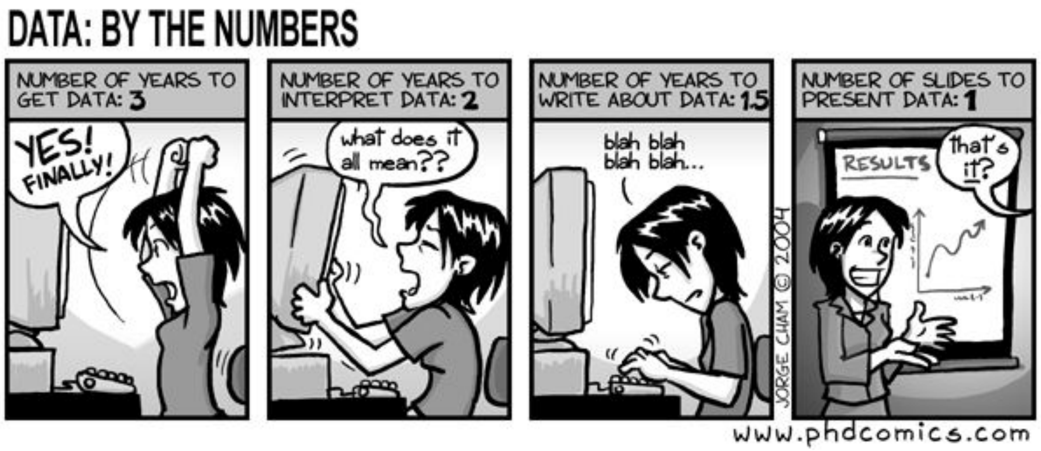
$$\frac{\partial [O_2]}{\partial t} = NEP = GPP - ER \pm D$$

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



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


DATA: BY THE NUMBERS
NUMBER OF YEARS TO GET DATA: 3
YES! FINALLY!
NUMBER OF YEARS TO INTERPRET DATA: 2
what does it all mean??
NUMBER OF YEARS TO WRITE ABOUT DATA: 1.5
blah blah blah blah...
NUMBER OF SLIDES TO PRESENT DATA: 1
RESULTS
that's it?
JORGE CHAM © 2004
www.phdcomics.com

phdcomics.com

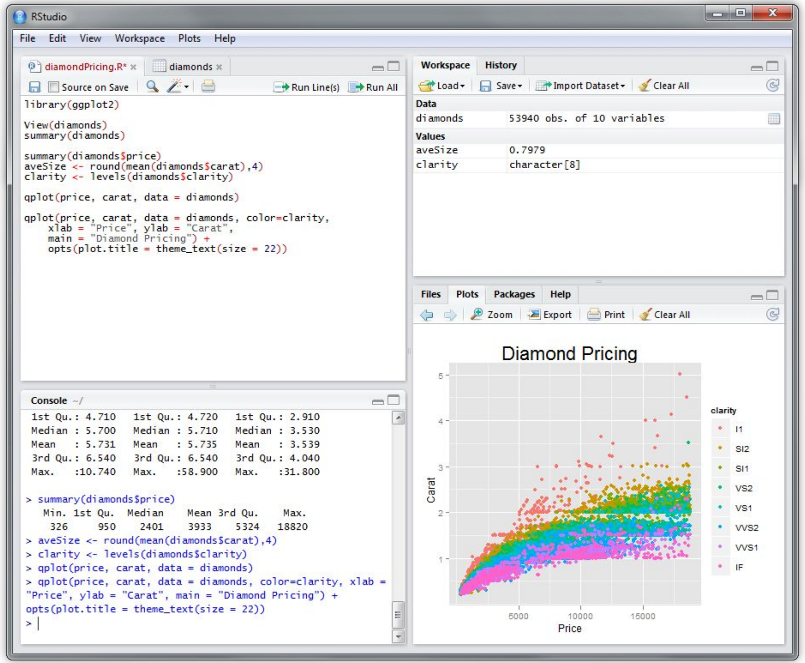
What’s the plan for today?

- 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, analyse data, plot data

Markus Spiske

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



```r
library(ggplot2)

View(diamonds)
summary(diamonds)

summary(diamonds$price)
aveSize <- round(mean(diamonds$carat),4)
clarity <- levels(diamonds$clarity)

qplot(price, carat, data = diamonds)

qplot(price, carat, data = diamonds, color=clarity,
    xlab = "Price", ylab = "Carat",
    main = "Diamond Pricing") +
    opts(plot.title = theme_text(size = 22))
```

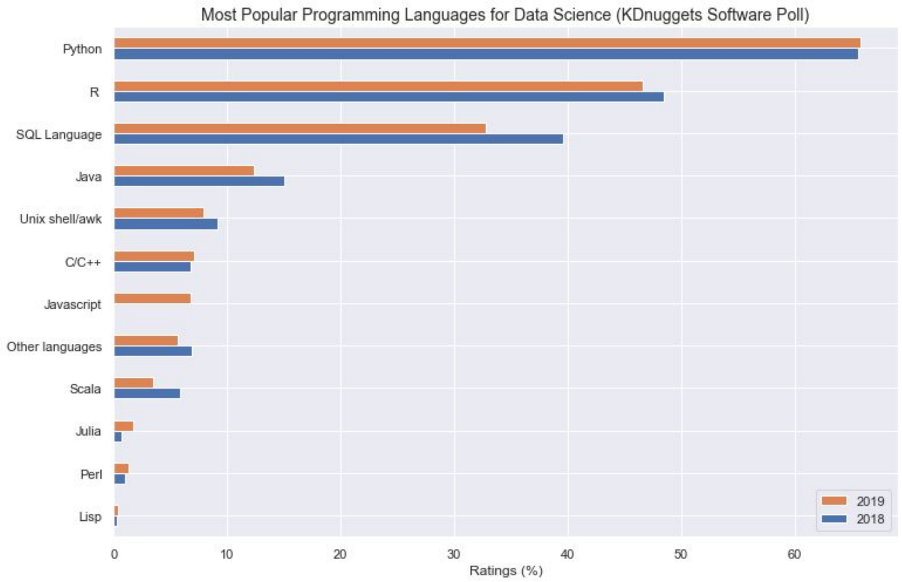
```
Data
diamonds      53940 obs. of 10 variables
Values
aveSize       0.7979
clarity       character[8]
```

```
 1st Qu.: 4.710   1st Qu.: 4.720   1st Qu.: 2.910
 Median : 5.700   Median : 5.710   Median : 3.530
 Mean   : 5.731   Mean   : 5.735   Mean   : 3.539
 3rd Qu.: 6.540   3rd Qu.: 6.540   3rd Qu.: 4.040
 Max.   :10.740   Max.   :58.900   Max.   :31.800

> summary(diamonds$price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    326     950    2401    3933    5324   18820
> aveSize <- round(mean(diamonds$carat),4)
> clarity <- levels(diamonds$clarity)
> qplot(price, carat, data = diamonds)
> qplot(price, carat, data = diamonds, color=clarity, xlab =
"Price", ylab = "Carat", main = "Diamond Pricing") +
opts(plot.title = theme_text(size = 22))
> |
```

RStudio interface

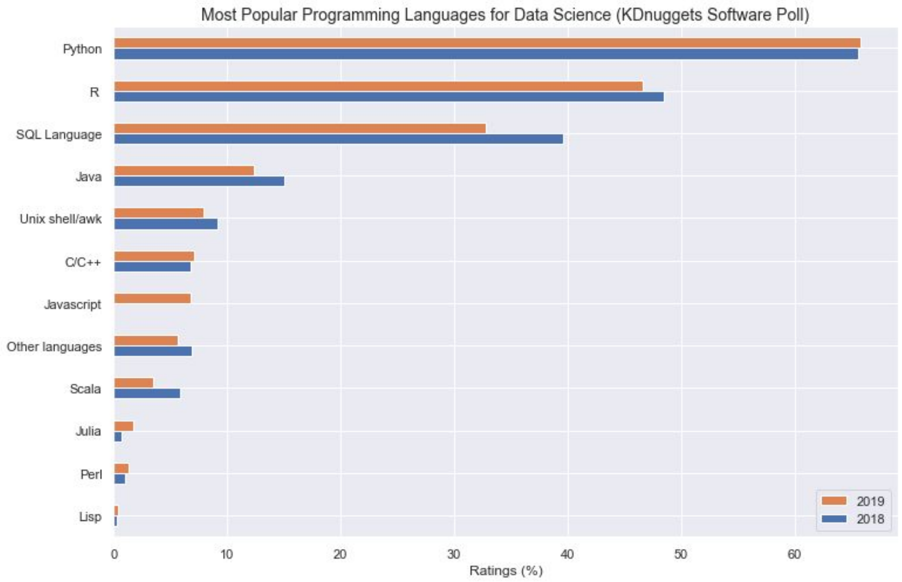
What's R?

- FREE and OPEN SOURCE statistical and computational software



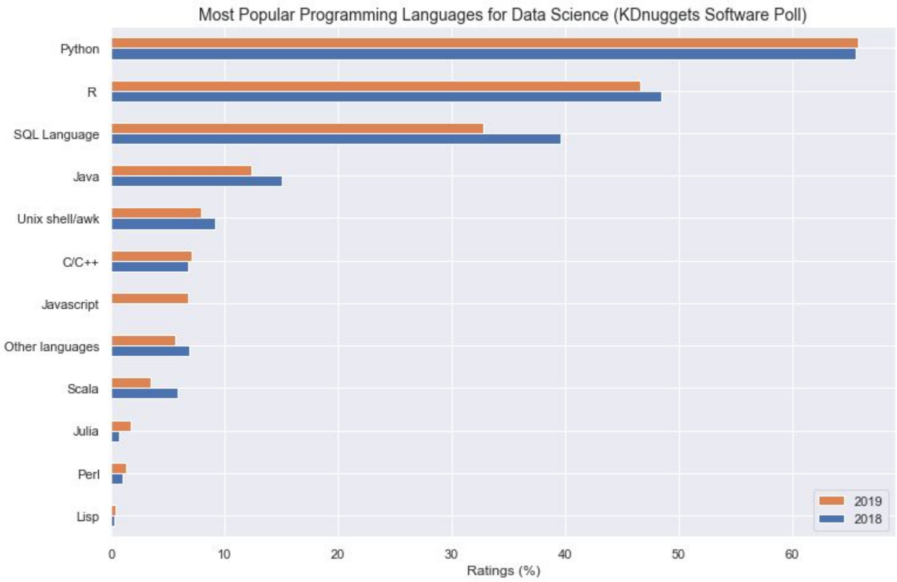
What's R?

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



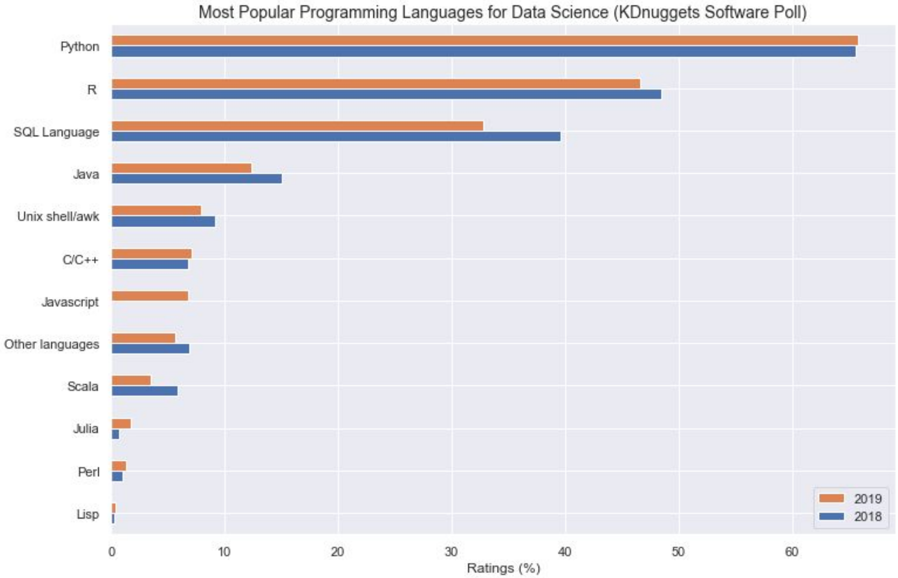
What's R?

- 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



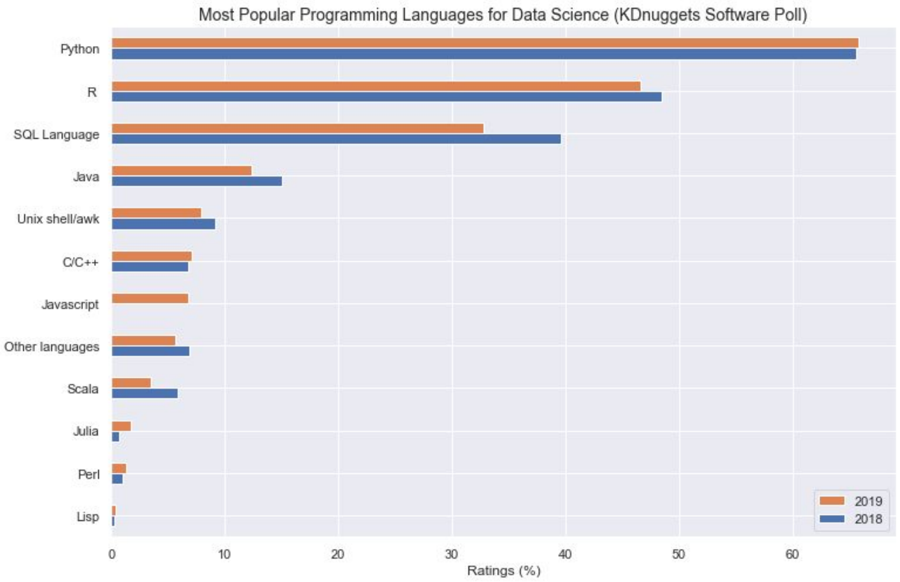
What’s R?

- 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

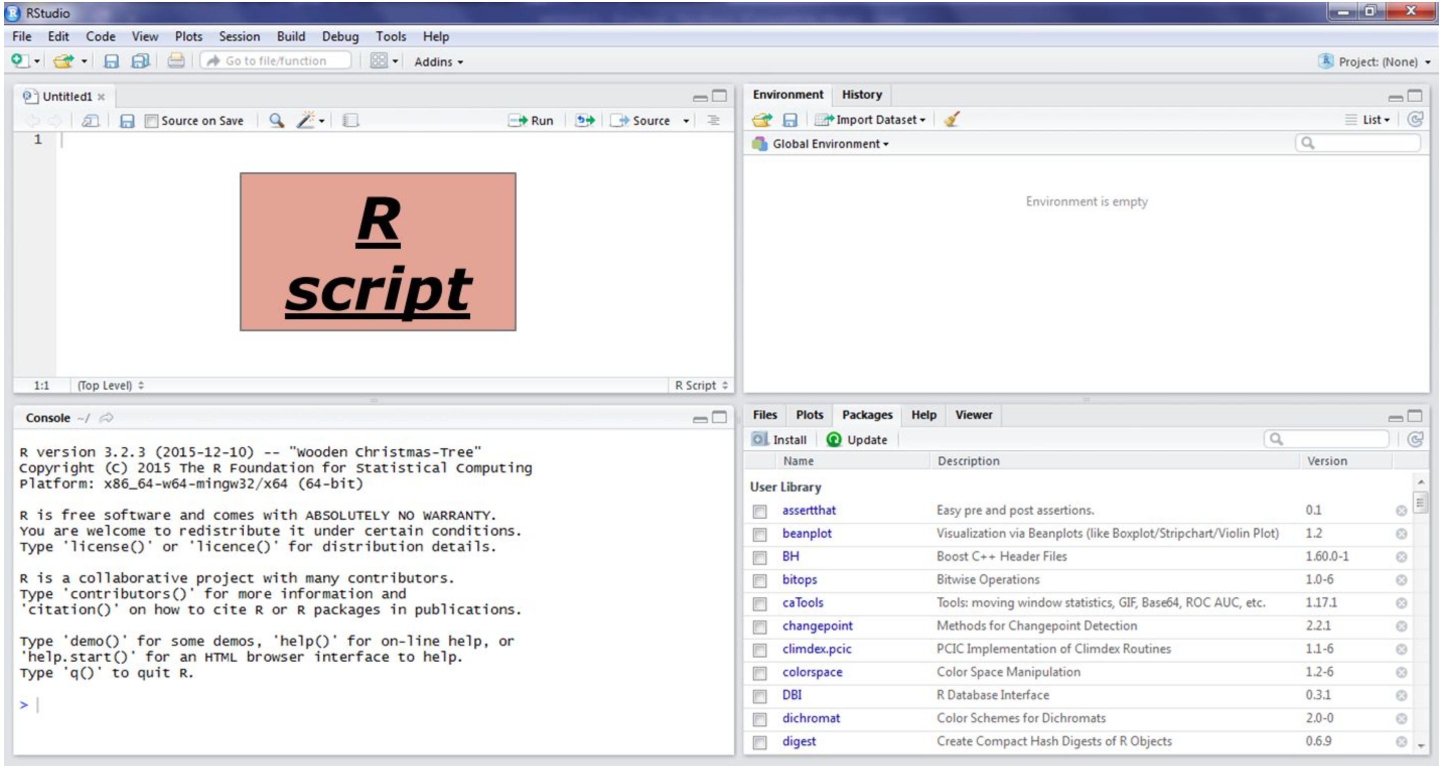


What’s R?

- 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 version 3.2.3 (2015-12-10) -- "Wooden Christmas-Tree"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

	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

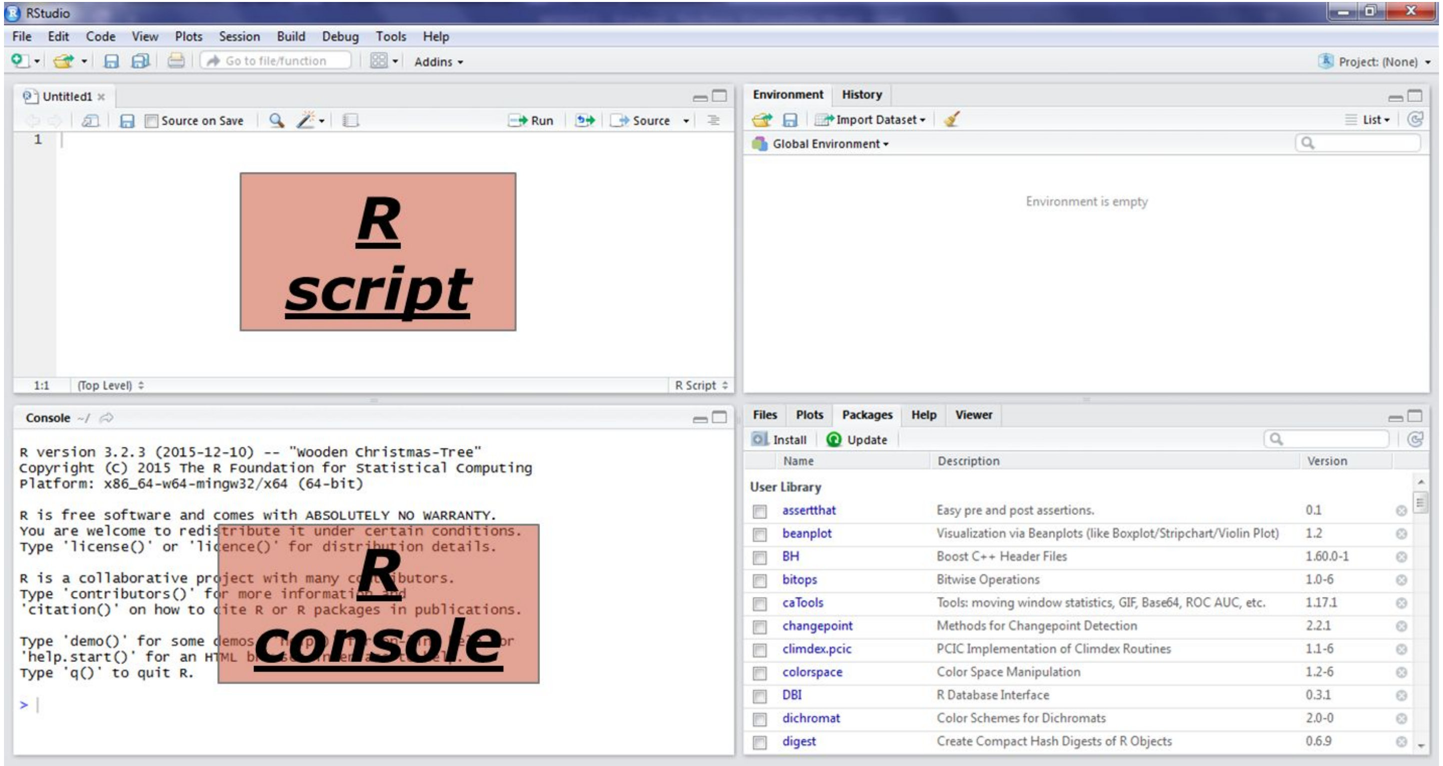
R script:

write and edit code

color-coding for easier identification

automatically fills in parentheses and quotations

Open RStudio on your laptop now



```
R version 3.2.3 (2015-12-10) -- "Wooden Christmas-Tree"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```

User Library

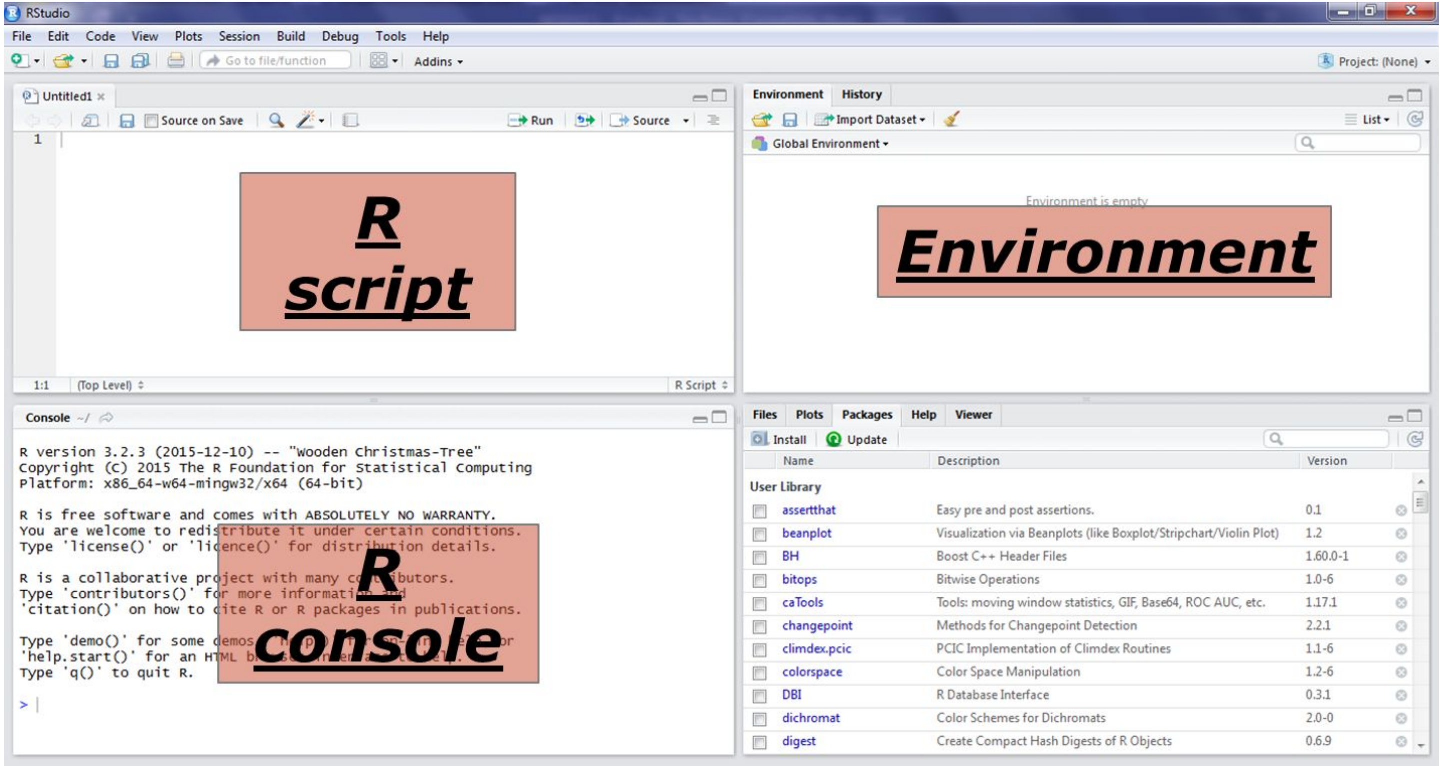
	Name	Description	Version
□	assertthat	Easy pre and post assertions.	0.1
□	beanplot	Visualization via Beanplots (like Boxplot/Stripchart/Violin Plot)	1.2
□	BH	Boost C++ Header Files	1.60.0-1
□	bitops	Bitwise Operations	1.0-6
□	caTools	Tools: moving window statistics, GIF, Base64, ROC AUC, etc.	1.17.1
□	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

R console:

commands are run and results appear

> means ‘ready to work’

Open RStudio on your laptop now



```
R version 3.2.3 (2015-12-10) -- "Wooden Christmas-Tree"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> 
```

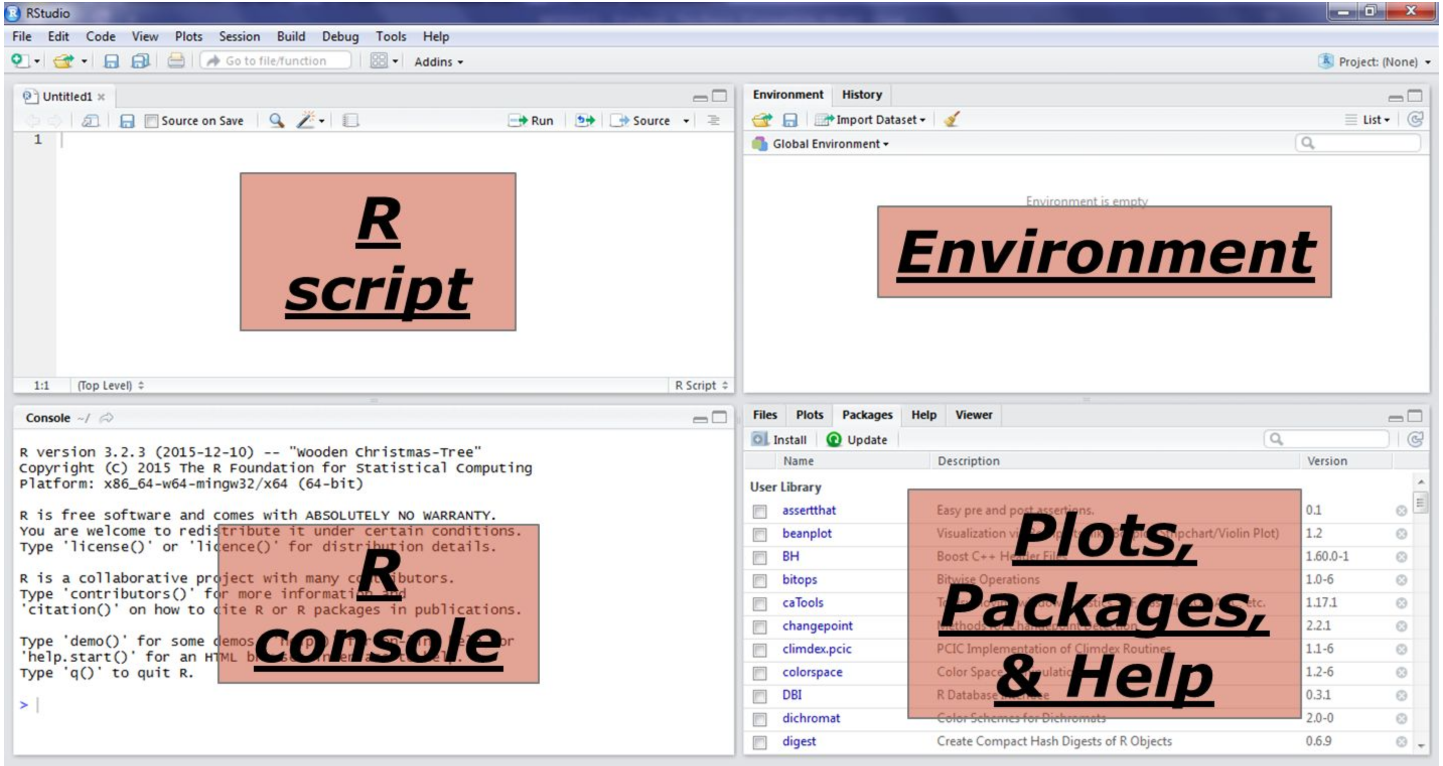
	Name	Description	Version
	User Library		
□	assertthat	Easy pre and post assertions.	0.1
□	beanplot	Visualization via Beanplots (like Boxplot/Stripchart/Violin Plot)	1.2
□	BH	Boost C++ Header Files	1.60.0-1
□	bitops	Bitwise Operations	1.0-6
□	caTools	Tools: moving window statistics, GIF, Base64, ROC AUC, 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

Environment:

see which objects were created

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

Open RStudio on your laptop now



```
R version 3.2.3 (2015-12-10) -- "Wooden Christmas-Tree"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

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

Plots/Packages/ Help:

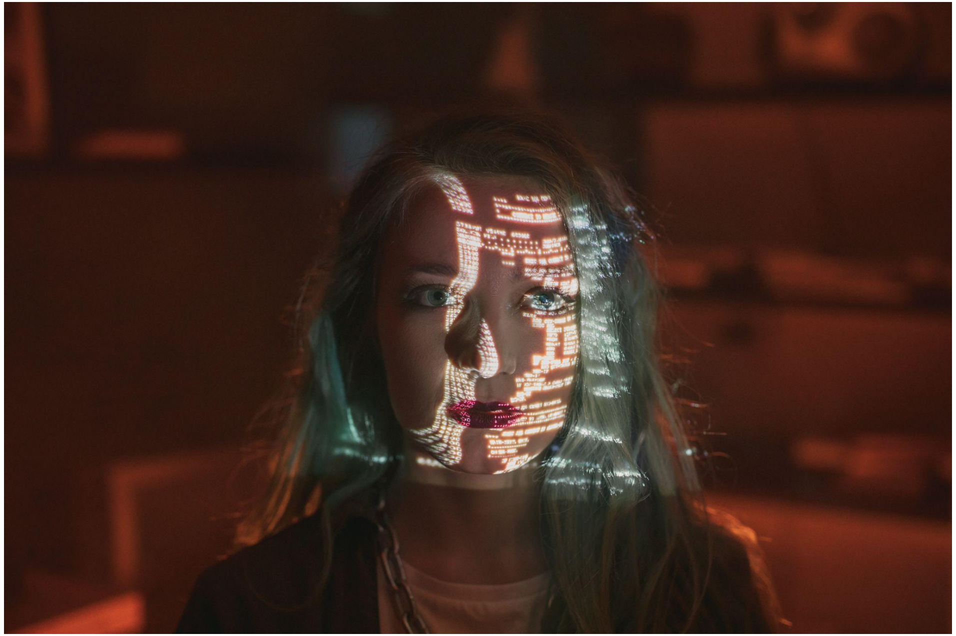
plots will appear

see which packages are installed

help functions

What are packages?

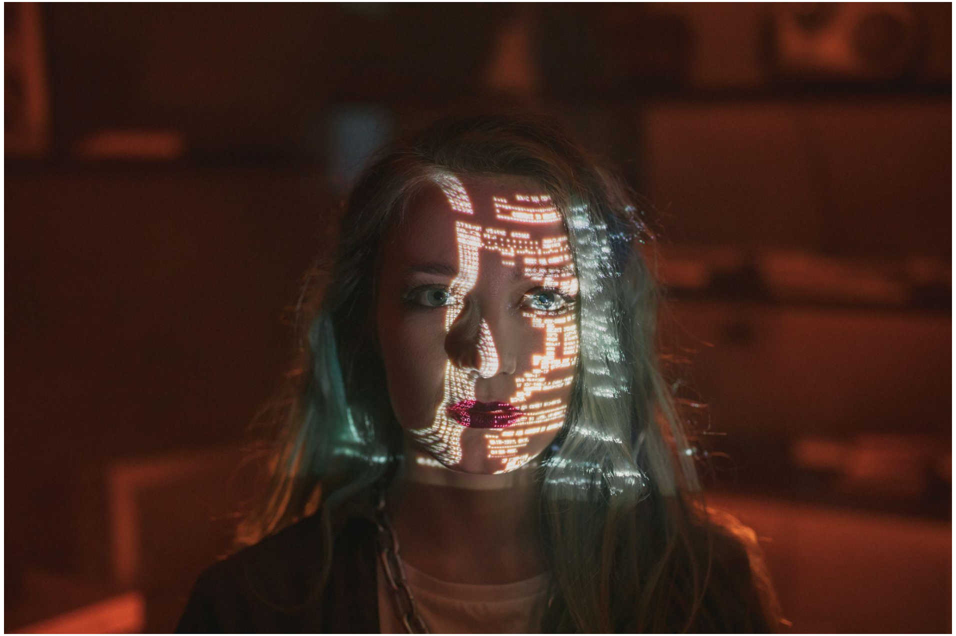
- bundles of tools and functions that others have developed to be used in R



Cottonbro

What are packages?

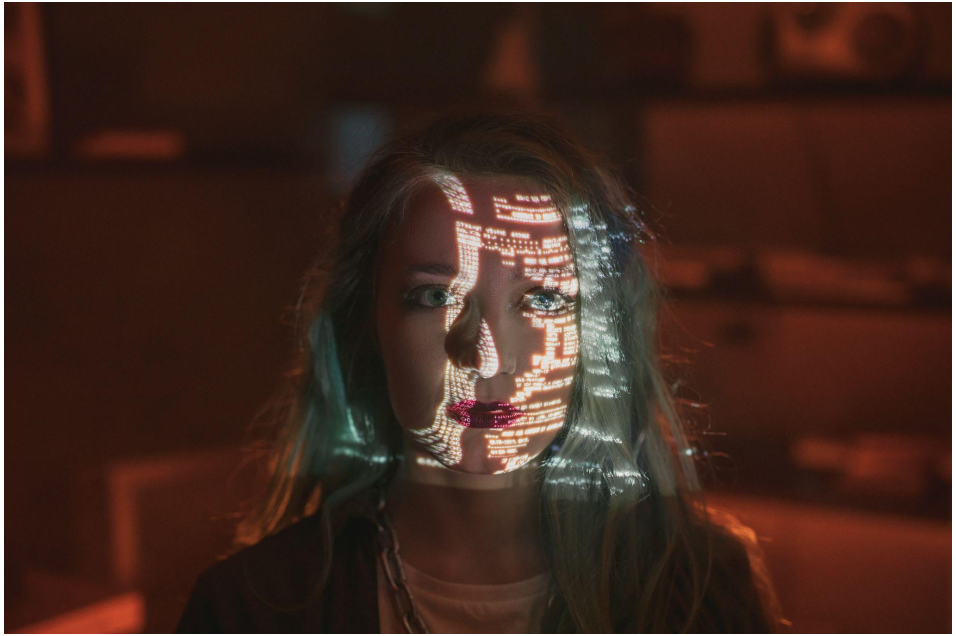
- bundles of tools and functions that others have developed to be used in R
- grouped to specific types of functions, analyses or datasets



Cottonbro

What are packages?

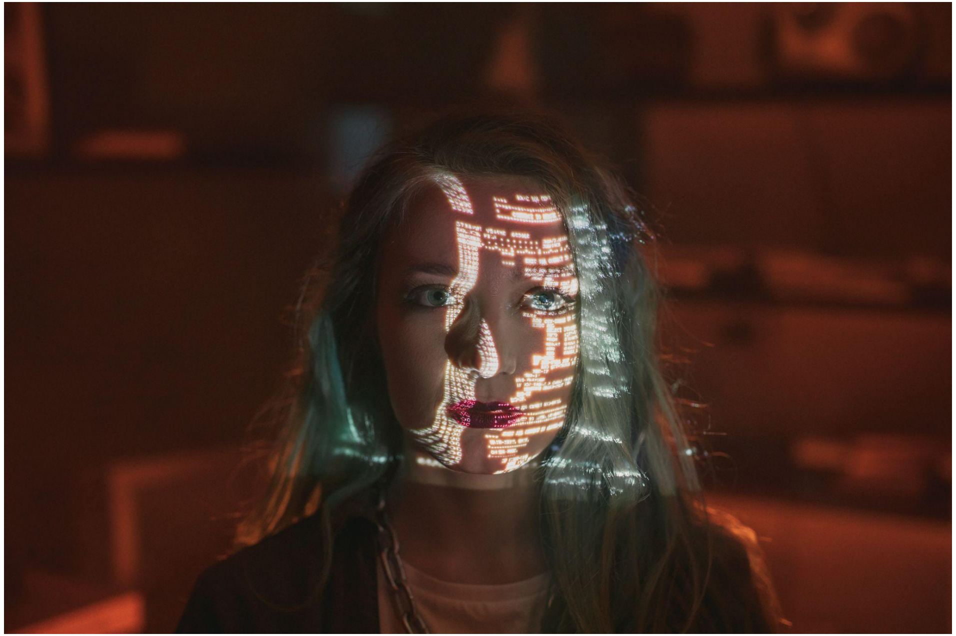
- bundles of tools and functions that others have developed to be used in R
- grouped to specific types of functions, analyses or datasets
- you can also create your own packages for specific tasks!



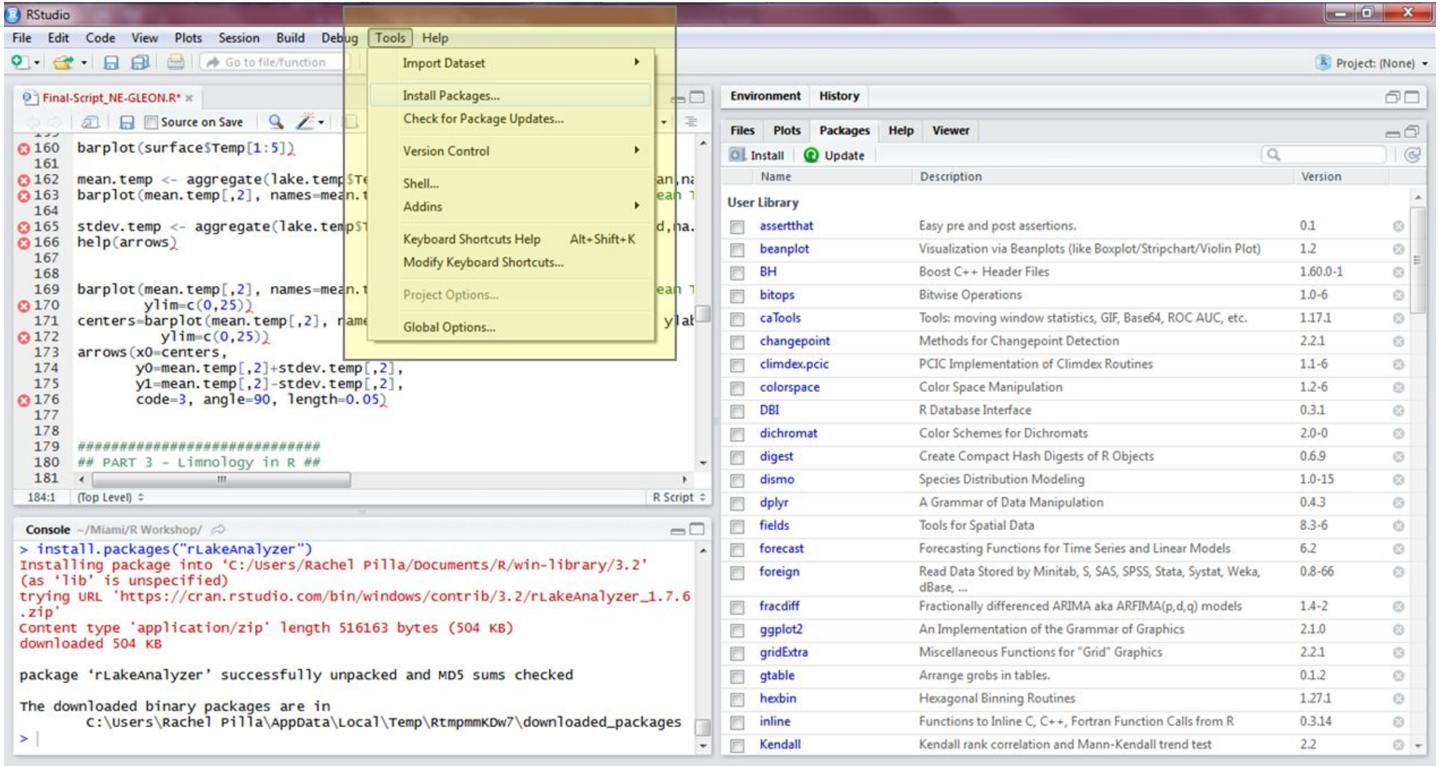
Cottonbro

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



```r
barplot(surface$Temp[1:5])

mean.temp <- aggregate(lake.temp$Te                                  an,na
barplot(mean.temp[,2], names=mean.t                                  ean T

stdev.temp <- aggregate(lake.temp$T                                  d,na.
help(arrows)


barplot(mean.temp[,2], names=mean.t                                  ean T
         ylim=c(0,25))
centers=barplot(mean.temp[,2], name                                  ylab
           ylim=c(0,25))
arrows(x0=centers,
       y0=mean.temp[,2]+stdev.temp[,2],
       y1=mean.temp[,2]-stdev.temp[,2],
       code=3, angle=90, length=0.05)


##############################
## PART 3 - Limnology in R ##
```

```
> 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
> |
```

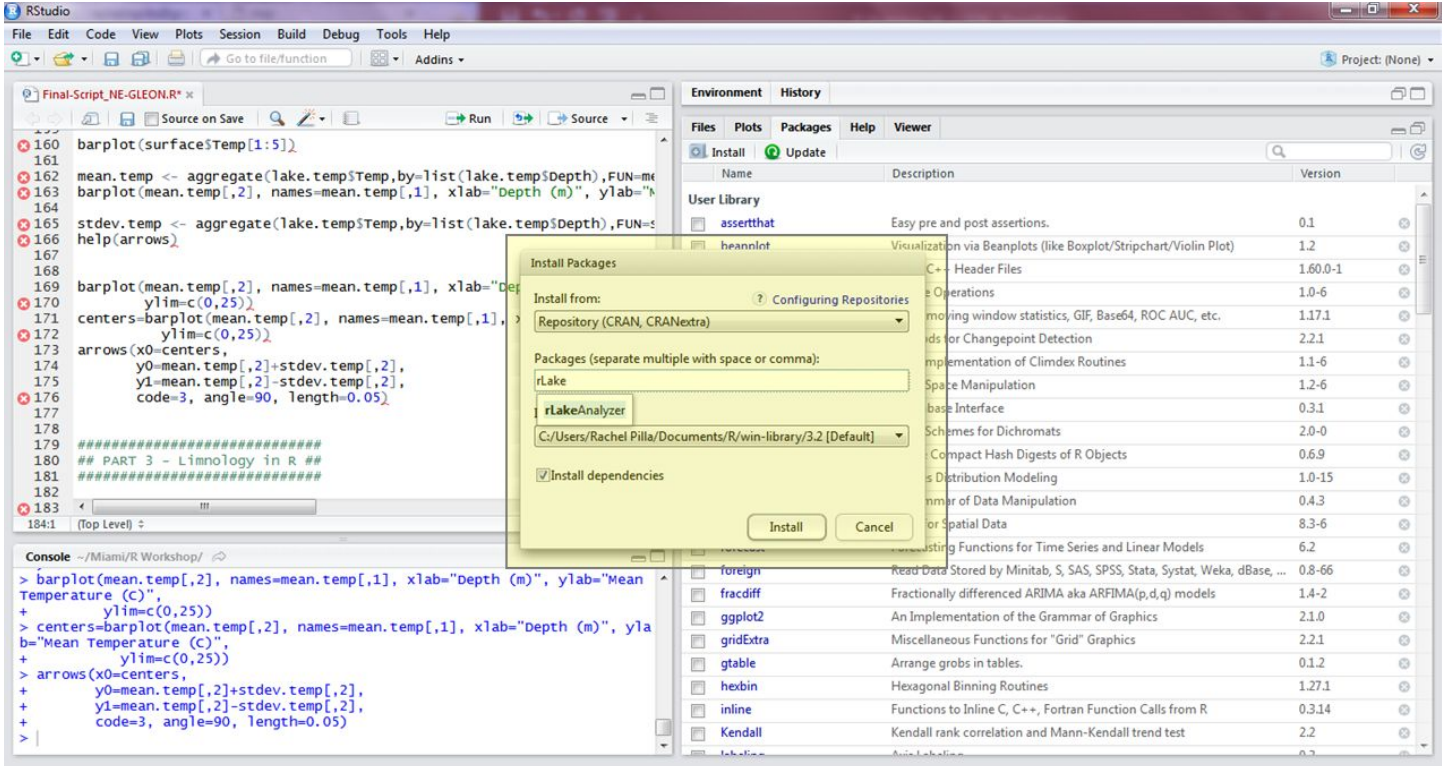
User Library

	Name	Description	Version
□	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

Install one package



```
barplot(surface$Temp[1:5])

mean.temp <- aggregate(lake.temp$Temp,by=list(lake.temp$Depth),FUN=me
barplot(mean.temp[,2], names=mean.temp[,1], xlab="Depth (m)", ylab="M

stdev.temp <- aggregate(lake.temp$Temp,by=list(lake.temp$Depth),FUN=s
help(arrows)


barplot(mean.temp[,2], names=mean.temp[,1], xlab="Dep
         ylim=c(0,25))
centers=barplot(mean.temp[,2], names=mean.temp[,1], x
           ylim=c(0,25))
arrows(x0=centers,
        y0=mean.temp[,2]+stdev.temp[,2],
        y1=mean.temp[,2]-stdev.temp[,2],
        code=3, angle=90, length=0.05)


###############################
## PART 3 - Limnology in R ##
###############################
```

184:1 (Top Level)

```
> barplot(mean.temp[,2], names=mean.temp[,1], xlab="Depth (m)", ylab="Mean
Temperature (C)",
+          ylim=c(0,25))
> centers=barplot(mean.temp[,2], names=mean.temp[,1], xlab="Depth (m)", yla
b="Mean Temperature (C)",
+           ylim=c(0,25))
> arrows(x0=centers,
+        y0=mean.temp[,2]+stdev.temp[,2],
+        y1=mean.temp[,2]-stdev.temp[,2],
+        code=3, angle=90, length=0.05)
> |
```

Install Packages

Install from: Repository (CRAN, CRANextra) — ? Configuring Repositories

Packages (separate multiple with space or comma): rLake

rLakeAnalyzer

I C:/Users/Rachel Pilla/Documents/R/win-library/3.2 [Default]

☑ Install dependencies

Install | Cancel

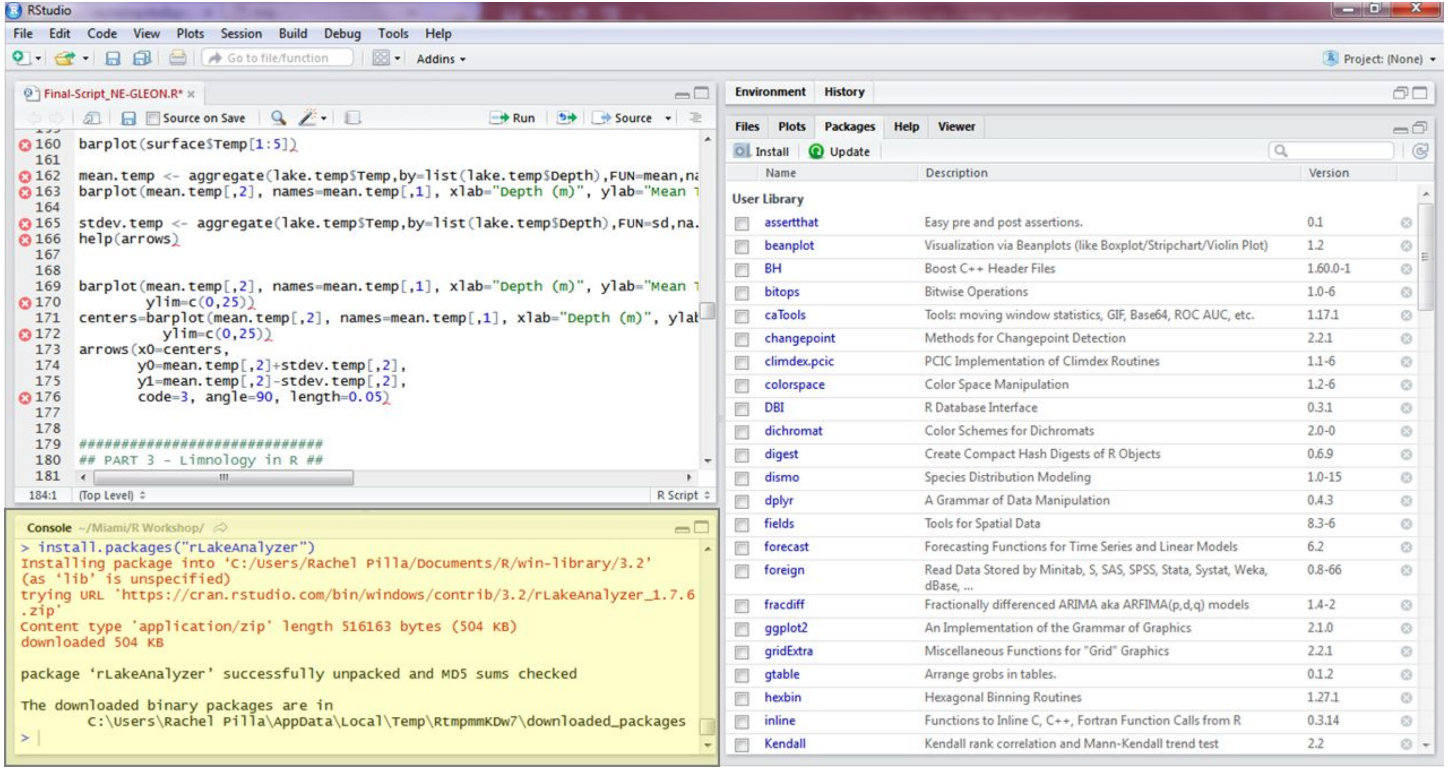
User Library

	Name	Description	Version
□	assertthat	Easy pre and post assertions.	0.1
□	beanplot	Visualization via Beanplots (like Boxplot/Stripchart/Violin Plot)	1.2
		C++ Header Files	1.60.0-1
		Operations	1.0-6
		moving window statistics, GIF, Base64, ROC AUC, etc.	1.17.1
		ods for Changepoint Detection	2.2.1
		mplementation of Climdex Routines	1.1-6
		Space Manipulation	1.2-6
		base Interface	0.3.1
		Schemes for Dichromats	2.0-0
		Compact Hash Digests of R Objects	0.6.9
		s Distribution Modeling	1.0-15
		mmar of Data Manipulation	0.4.3
		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

Install one package



```
barplot(surface$Temp[1:5])

mean.temp <- aggregate(lake.temp$Temp,by=list(lake.temp$Depth),FUN=mean,na
barplot(mean.temp[,2], names=mean.temp[,1], xlab="Depth (m)", ylab="Mean T

stdev.temp <- aggregate(lake.temp$Temp,by=list(lake.temp$Depth),FUN=sd,na.
help(arrows)


barplot(mean.temp[,2], names=mean.temp[,1], xlab="Depth (m)", ylab="Mean T
         ylim=c(0,25))
centers=barplot(mean.temp[,2], names=mean.temp[,1], xlab="Depth (m)", ylab
           ylim=c(0,25))
arrows(x0=centers,
        y0=mean.temp[,2]+stdev.temp[,2],
        y1=mean.temp[,2]-stdev.temp[,2],
        code=3, angle=90, length=0.05)


##############################
## PART 3 - Limnology in R ##
```

```
> 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
>
```

	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

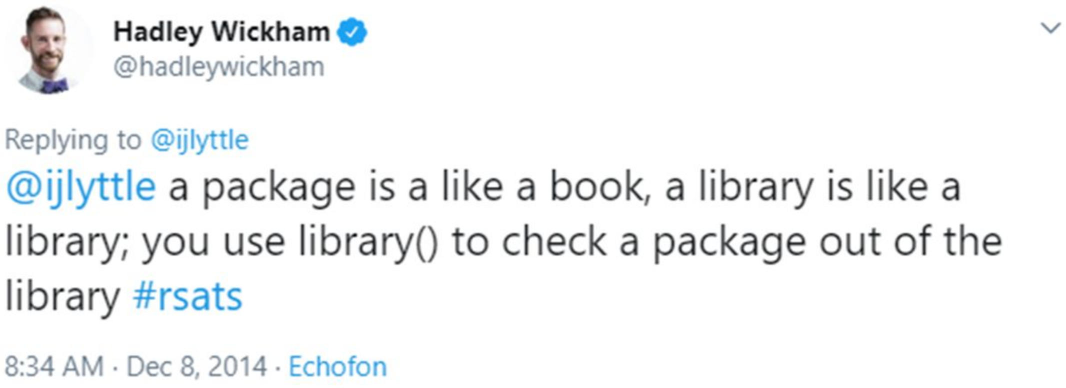
we need:

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)
```



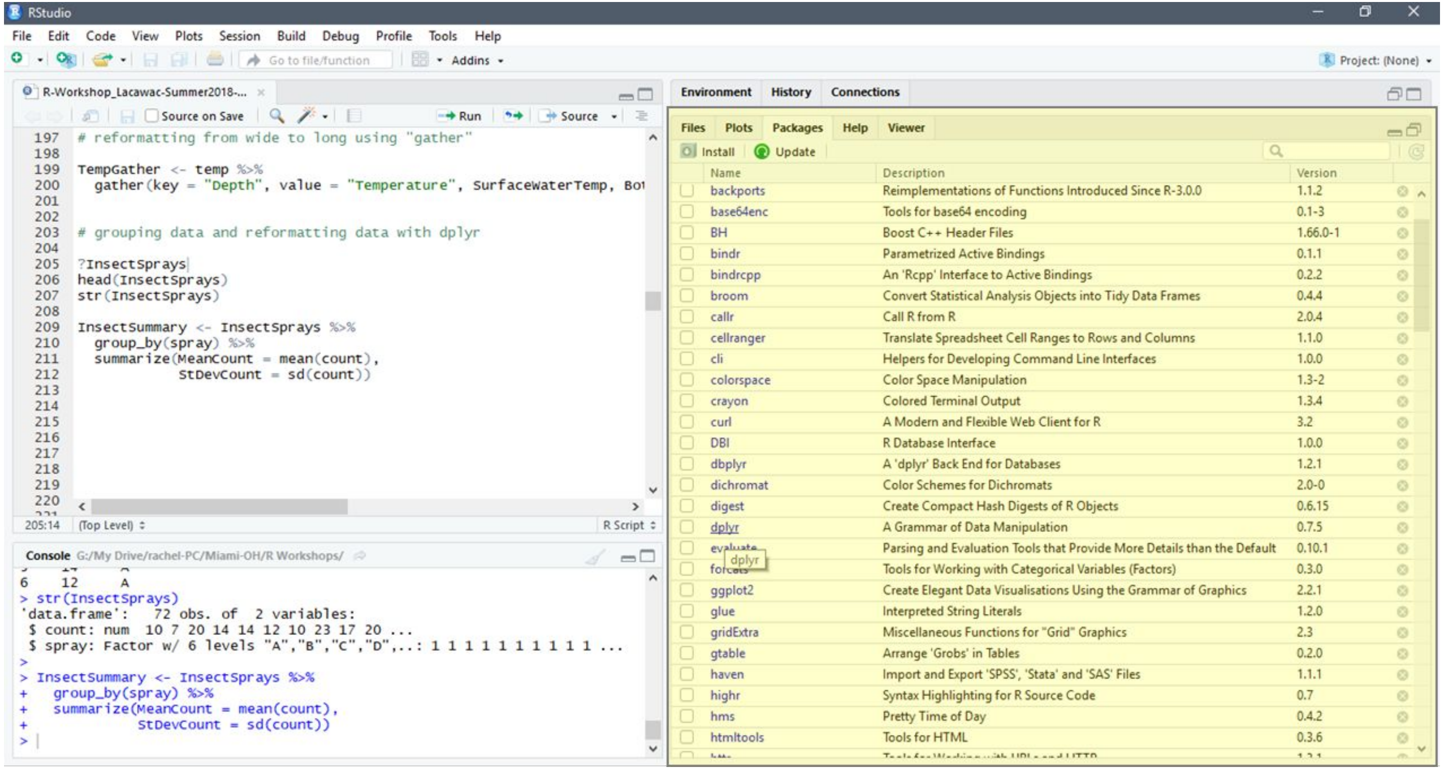
Hadley Wickham
@hadleywickham

Replying to @ijlyttle

@ijlyttle a package is a like a book, a library is like a library; you use library() to check a package out of the library #rsats

8:34 AM · Dec 8, 2014 · Echofon

Package information


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
R-Workshop_Lacawac-Summer2018-...
Source on Save
Run
Source
205:14
(Top Level)
R Script
Console G:/My Drive/rachel-PC/Miami-OH/R Workshops/
Environment History Connections
Files Plots Packages Help Viewer
Install Update
dplyr

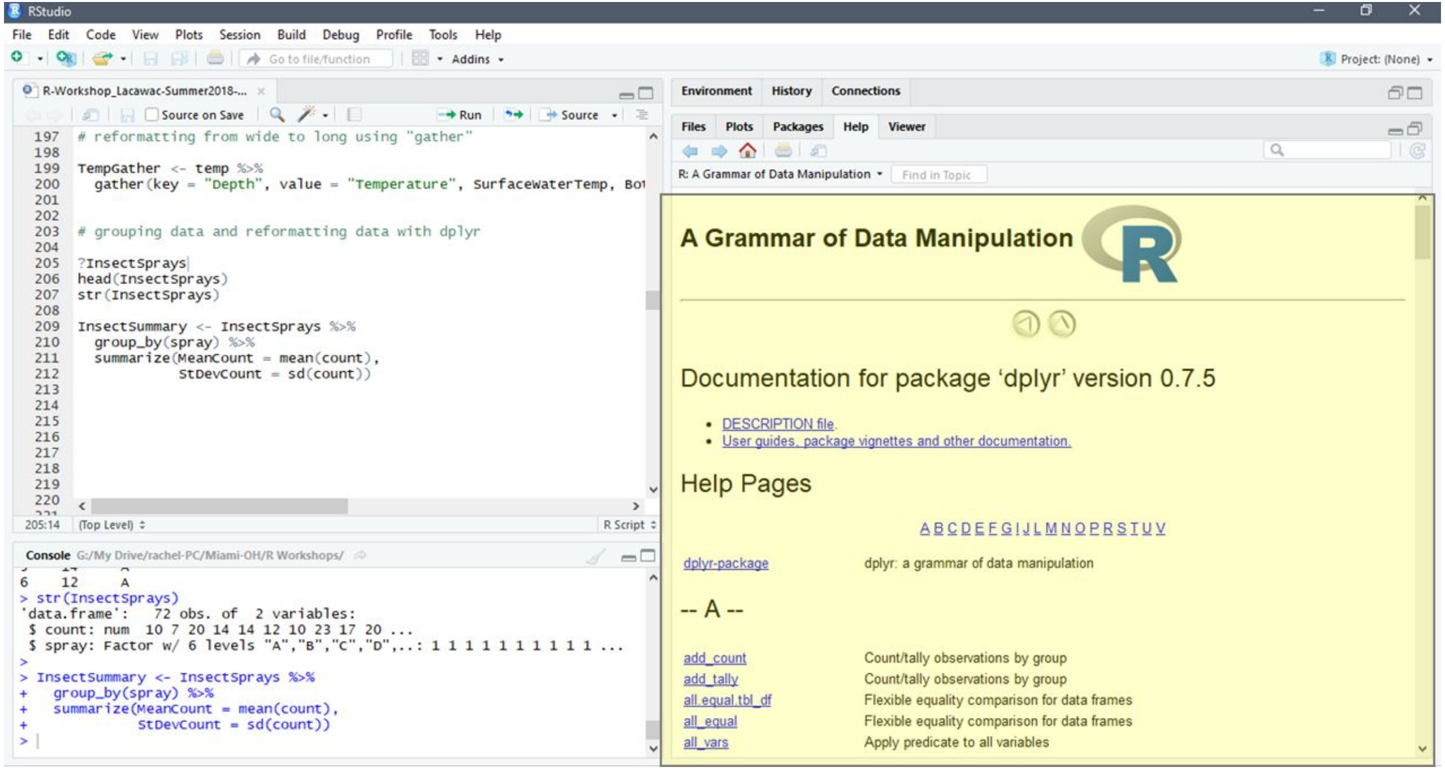
```
# reformatting from wide to long using "gather"

TempGather <- temp %>%
  gather(key = "Depth", value = "Temperature", SurfaceWaterTemp, Bot


# grouping data and reformatting data with dplyr

?InsectSprays
head(InsectSprays)
str(InsectSprays)

InsectSummary <- InsectSprays %>%
  group_by(spray) %>%
  summarize(MeanCount = mean(count),
            StDevCount = sd(count))
```

```
6    12      A
> str(InsectSprays)
'data.frame':   72 obs. of  2 variables:
 $ count: num  10 7 20 14 14 12 10 23 17 20 ...
 $ spray: Factor w/ 6 levels "A","B","C","D",..: 1 1 1 1 1 1 1 1 1 1 ...
>
> InsectSummary <- InsectSprays %>%
+   group_by(spray) %>%
+   summarize(MeanCount = mean(count),
+             StDevCount = sd(count))
> |
```

Name	Description	Version
backports	Reimplementations of Functions Introduced Since R-3.0.0	1.1.2
base64enc	Tools for base64 encoding	0.1-3
BH	Boost C++ Header Files	1.66.0-1
bindr	Parametrized Active Bindings	0.1.1
bindrcpp	An 'Rcpp' Interface to Active Bindings	0.2.2
broom	Convert Statistical Analysis Objects into Tidy Data Frames	0.4.4
callr	Call R from R	2.0.4
cellranger	Translate Spreadsheet Cell Ranges to Rows and Columns	1.1.0
cli	Helpers for Developing Command Line Interfaces	1.0.0
colorspace	Color Space Manipulation	1.3-2
crayon	Colored Terminal Output	1.3.4
curl	A Modern and Flexible Web Client for R	3.2
DBI	R Database Interface	1.0.0
dbplyr	A 'dplyr' Back End for Databases	1.2.1
dichromat	Color Schemes for Dichromats	2.0-0
digest	Create Compact Hash Digests of R Objects	0.6.15
dplyr	A Grammar of Data Manipulation	0.7.5
evaluate	Parsing and Evaluation Tools that Provide More Details than the Default	0.10.1
forcats	Tools for Working with Categorical Variables (Factors)	0.3.0
ggplot2	Create Elegant Data Visualisations Using the Grammar of Graphics	2.2.1
glue	Interpreted String Literals	1.2.0
gridExtra	Miscellaneous Functions for "Grid" Graphics	2.3
gtable	Arrange 'Grobs' in Tables	0.2.0
haven	Import and Export 'SPSS', 'Stata' and 'SAS' Files	1.1.1
highr	Syntax Highlighting for R Source Code	0.7
hms	Pretty Time of Day	0.4.2
htmltools	Tools for HTML	0.3.6

under packages, click on **dplyr**

Package information


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
R-Workshop_Lacawac-Summer2018-...
Source on Save
Run
Source
205:14
(Top Level)
R Script
Console G:/My Drive/rachel-PC/Miami-OH/R Workshops/
Environment History Connections
Files Plots Packages Help Viewer
R: A Grammar of Data Manipulation
Find in Topic

```
# reformatting from wide to long using "gather"

TempGather <- temp %>%
  gather(key = "Depth", value = "Temperature", SurfaceWaterTemp, Bot


# grouping data and reformatting data with dplyr

?InsectSprays
head(InsectSprays)
str(InsectSprays)

InsectSummary <- InsectSprays %>%
  group_by(spray) %>%
  summarize(MeanCount = mean(count),
            StDevCount = sd(count))
```

```
6    12      A
> str(InsectSprays)
'data.frame':    72 obs. of  2 variables:
 $ count: num  10 7 20 14 14 12 10 23 17 20 ...
 $ spray: Factor w/ 6 levels "A","B","C","D",..: 1 1 1 1 1 1 1 1 1 1 ...
>
> InsectSummary <- InsectSprays %>%
+   group_by(spray) %>%
+   summarize(MeanCount = mean(count),
+             StDevCount = sd(count))
> 
```

A Grammar of Data Manipulation

Documentation for package ‘dplyr’ version 0.7.5

- DESCRIPTION file.
- User guides, package vignettes and other documentation.

Help Pages

A B C D E F G I J L M N O P R S T U V

dplyr-package	dplyr: a grammar of data manipulation

-- A --

add_count	Count/tally observations by group
add_tally	Count/tally observations by group
all.equal.tbl_df	Flexible equality comparison for data frames
all_equal	Flexible equality comparison for data frames
all_vars	Apply predicate to all variables

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

```
$result !== false) {
  $distArray = array();
  $row = mysqli_fetch_assoc($result);
  $correctAnswer = $row['Correct'];
  $distArray['A'] = $row['Anum'];
  $distArray['B'] = $row['Bnum'];
  $distArray['C'] = $row['Cnum'];
  $distArray['D'] = $row['Dnum'];
  $distArray['Correct'] = $correctAnswer;
  $distArray['Answer'] = rtrim($row[$correctAnswer], ".");
  $distArray['Query'] = "SELECT * FROM TechTerms WHERE Date='$date'";
  return $distArray;
} else {
  $distArray['Error'] = 'Quiz load query failed';
  return $distArray;
}
```

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

Mikhail Nilov

Coding in R

```
$result !== false) {
 $distArray = array();
 $row = mysqli_fetch_assoc($result);
 $correctAnswer = $row['Correct'];
 $distArray['A'] = $row['Anum'];
 $distArray['B'] = $row['Bnum'];
 $distArray['C'] = $row['Cnum'];
 $distArray['D'] = $row['Dnum'];
 $distArray['Correct'] = $correctAnswer;
 $distArray['Answer'] = rtrim($row[$correctAnswer], ".");
 $distArray['Query'] = "SELECT * FROM TechTerms WHERE Date='$date'";
 return $distArray;
else {
 $distArray['Error'] = 'Quiz load query failed';
 return $distArray;
}
```

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

Mikhail Nilov

Coding in R

```
$result !== false) {
 $distArray = array();
 $row = mysqli_fetch_assoc($result);
 $correctAnswer = $row['Correct'];
 $distArray['A'] = $row['Anum'];
 $distArray['B'] = $row['Bnum'];
 $distArray['C'] = $row['Cnum'];
 $distArray['D'] = $row['Dnum'];
 $distArray['Correct'] = $correctAnswer;
 $distArray['Answer'] = rtrim($row[$correctAnswer], ".");
 $distArray['Query'] = "SELECT * FROM TechTerms WHERE Date='$date'";
 return $distArray;
else {
 $distArray['Error'] = 'Quiz load query failed';
 return $distArray;
}
```

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

Mikhail Nilov

Coding in R

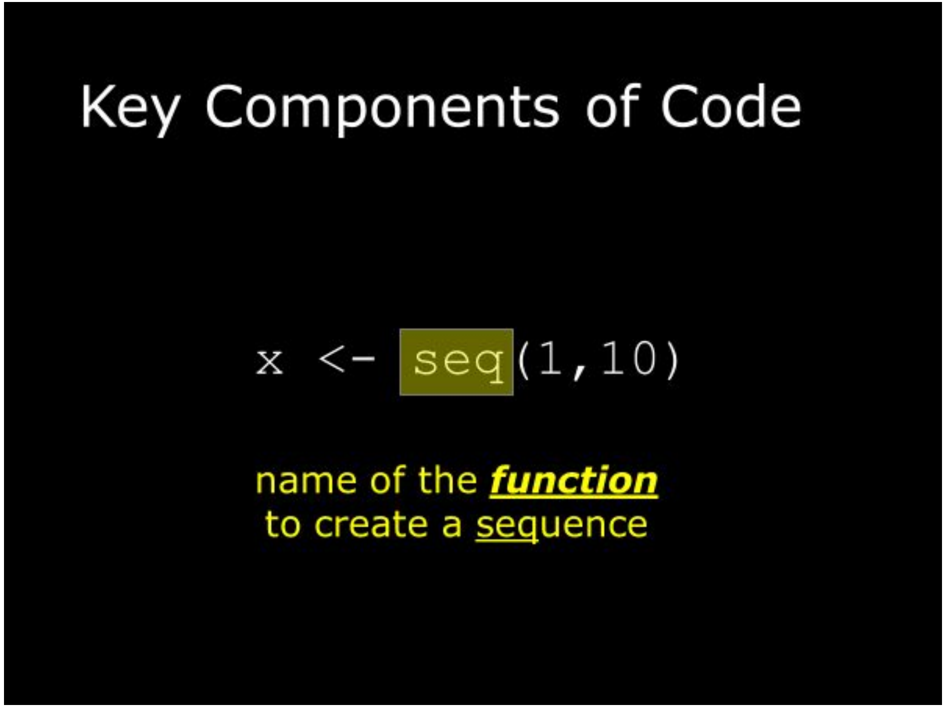
```
$result !== false) {
  $distArray = array();
  $row = mysqli_fetch_assoc($result);
  $correctAnswer = $row['Correct'];
  $distArray['A'] = $row['Anum'];
  $distArray['B'] = $row['Bnum'];
  $distArray['C'] = $row['Cnum'];
  $distArray['D'] = $row['Dnum'];
  $distArray['Correct'] = $correctAnswer;
  $distArray['Answer'] = rtrim($row[$correctAnswer], ".");
  $distArray['Query'] = "SELECT * FROM TechTerms WHERE Date='$date'";
  return $distArray;
else {
  $distArray['Error'] = 'Quiz load query failed';
  return $distArray;
}
```

- 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

Mikhail Nilov

Key components

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



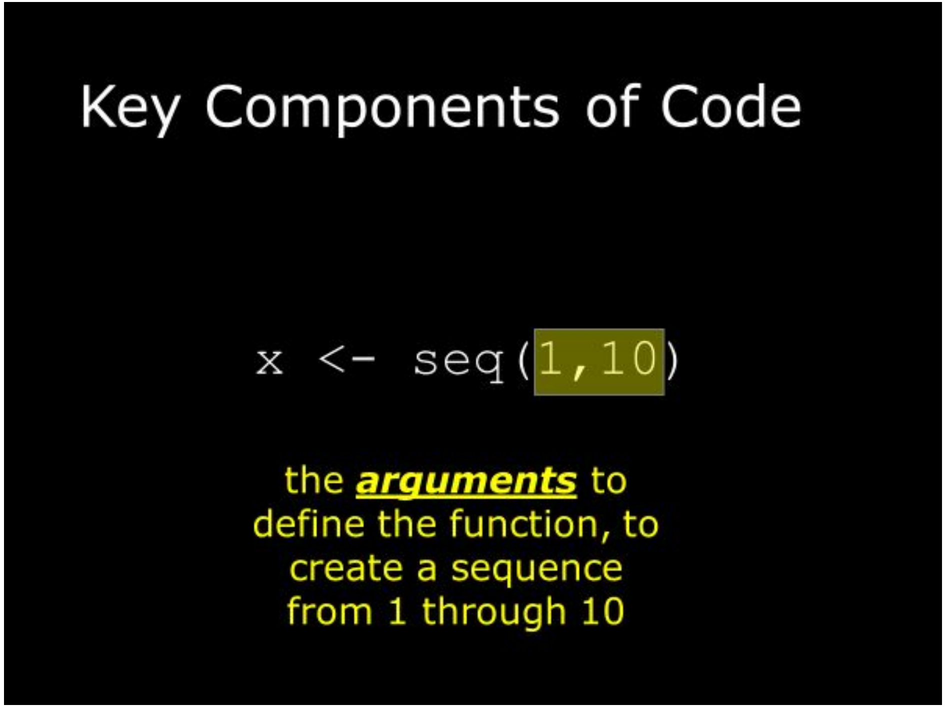
Key Components of Code

```
x <- seq(1,10)
```

name of the ***function***
to create a sequence

Key components

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



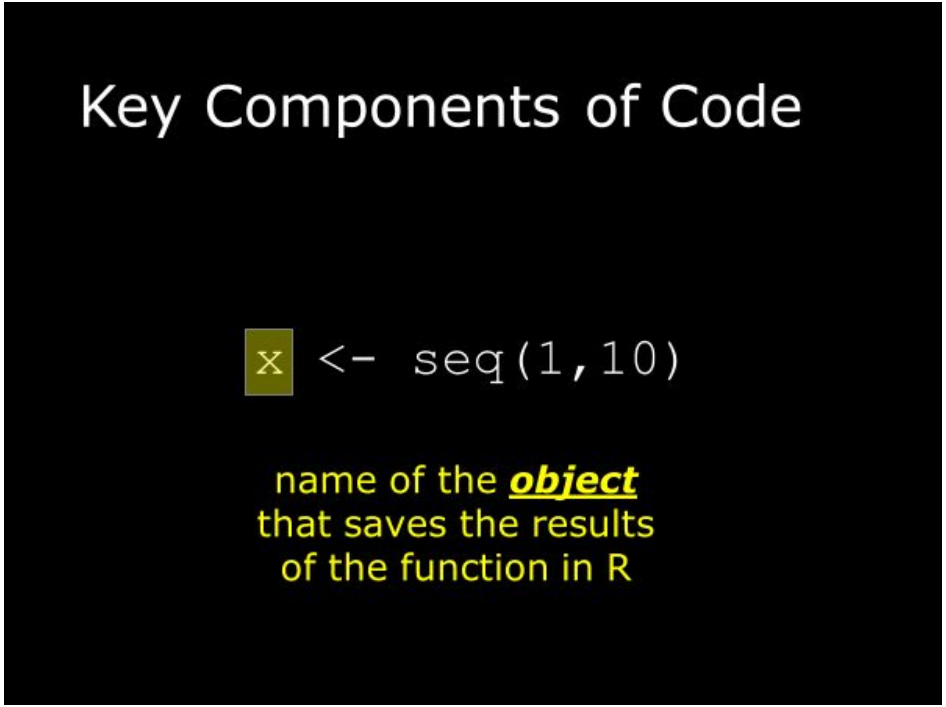
Key Components of Code

```
x <- seq(1,10)
```

the ***arguments*** to define the function, to create a sequence from 1 through 10

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



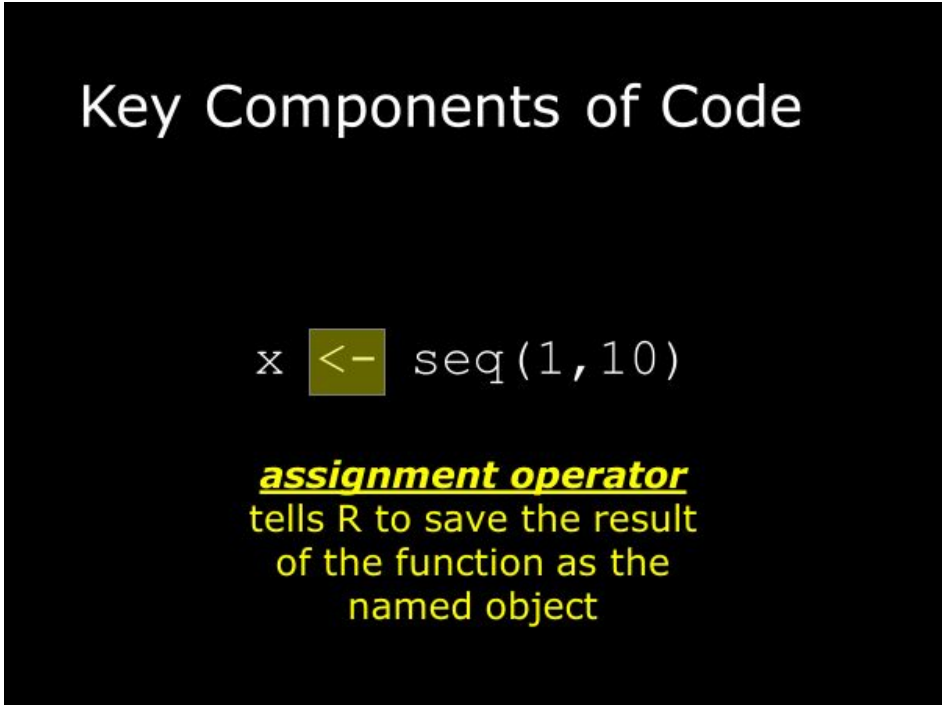
Key Components of Code

```
x <- seq(1,10)
```

name of the ***object***
that saves the results
of the function in R

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 of Code

```
x <- seq(1,10)
```

assignment operator
tells R to save the result of the function as the named object

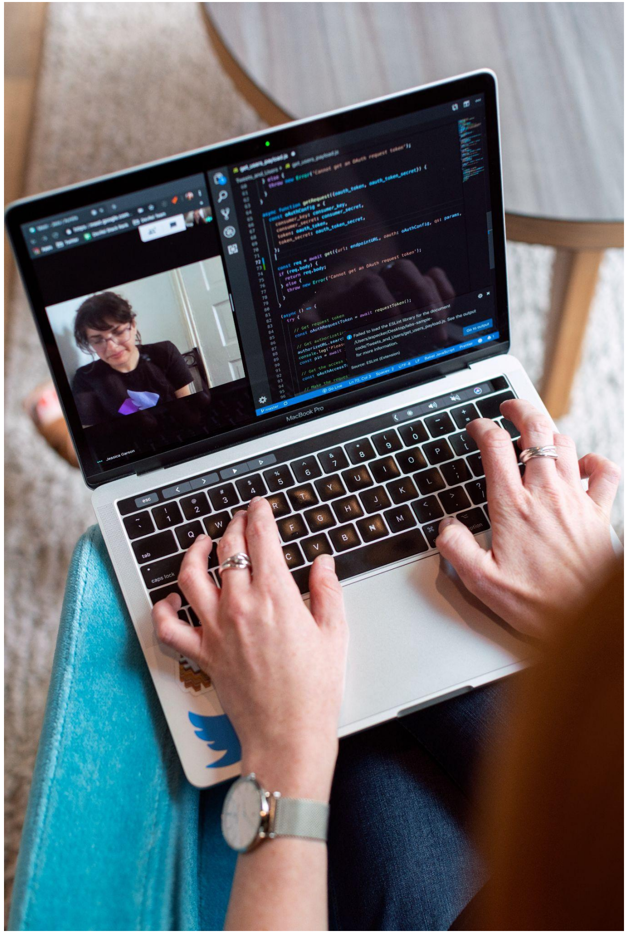
Types of data

- **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!

RStudio®



ThisIsEngineering