



A Short Introduction into Data Science with R



Robert Ladwig
University of Wisconsin-Madison



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

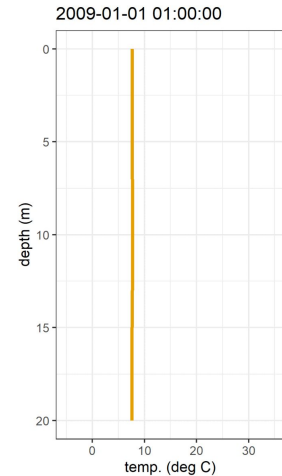


About myself

- **Physical limnologist**
 - use mathematical models to explore mixing dynamics and water quality in lakes
- developing and applying open-source and open-access software in **R**, **Python** and **Matlab**
- background in civil engineering and geology

Aquatic ecosystem modeling

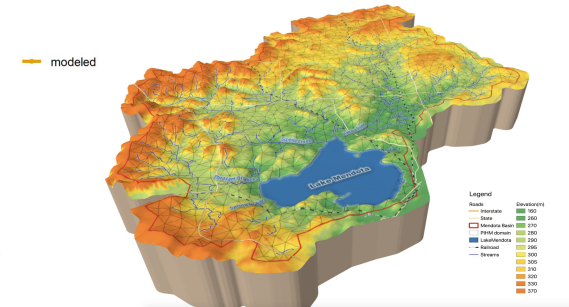
mechanistic understanding of cause and effects



$$\frac{\partial T}{\partial t} = \frac{1}{A} \frac{\partial}{\partial z} \left(A (v_t' + v') \frac{\partial T}{\partial z} \right) + \frac{1}{\rho_0 c_p} \frac{\partial H_{sol}}{\partial z} + \frac{dA}{dz} \frac{H_{geo}}{A \rho_0 c_p}$$

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

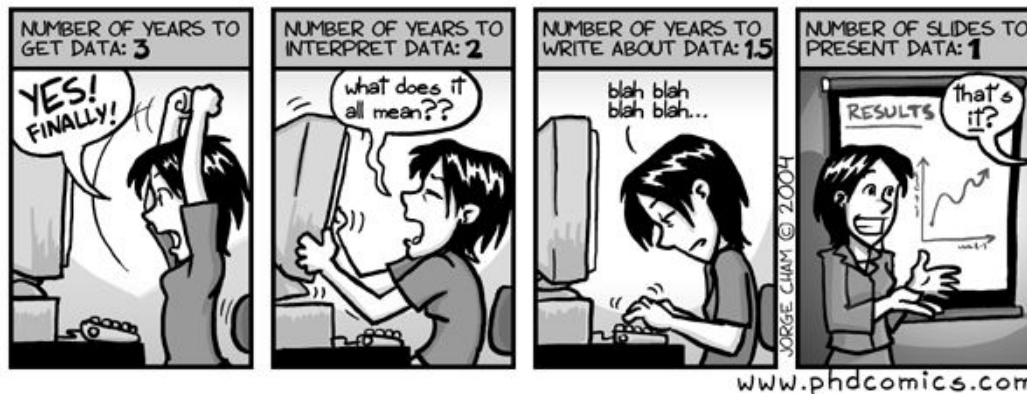
$$\frac{\partial P}{\partial t} = D \frac{\partial^2 P}{\partial z^2} + \lambda P \left(1 - \frac{P}{K} \right)$$



What's data science?

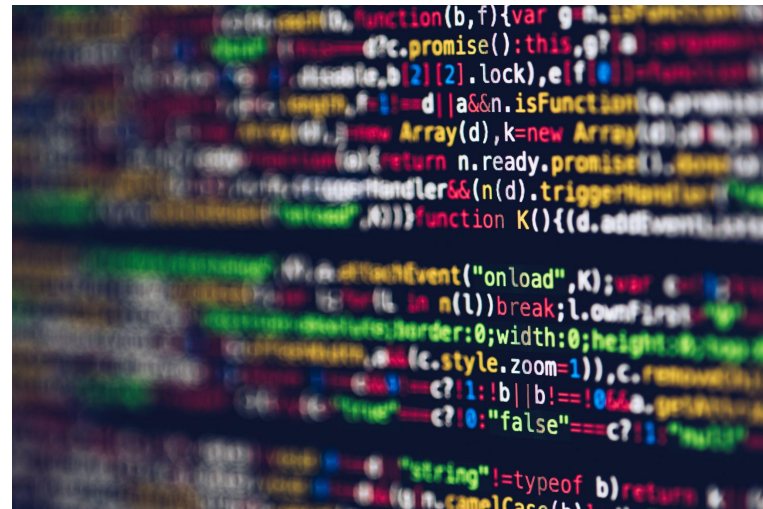
- combining math & statistics, programming and machine learning to explore any kind of data
- data is often noisy or unstructured → cleaning is big part of the job
- buzzwords like big data

DATA: BY THE NUMBERS



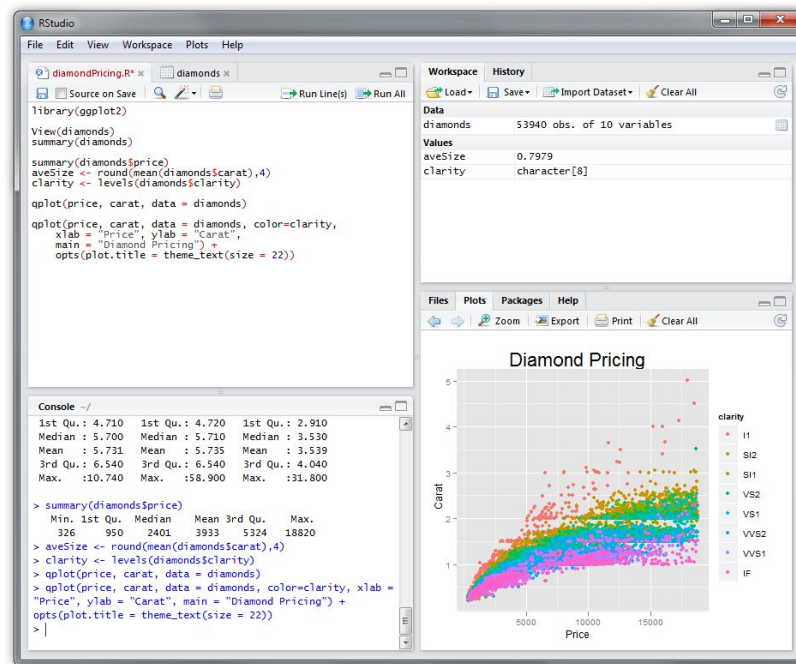
What's the plan for today?

- short introduction into the R scripting language
 - all material on github:
<https://github.com/robertladwig/intro-atasciencer>
 - 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



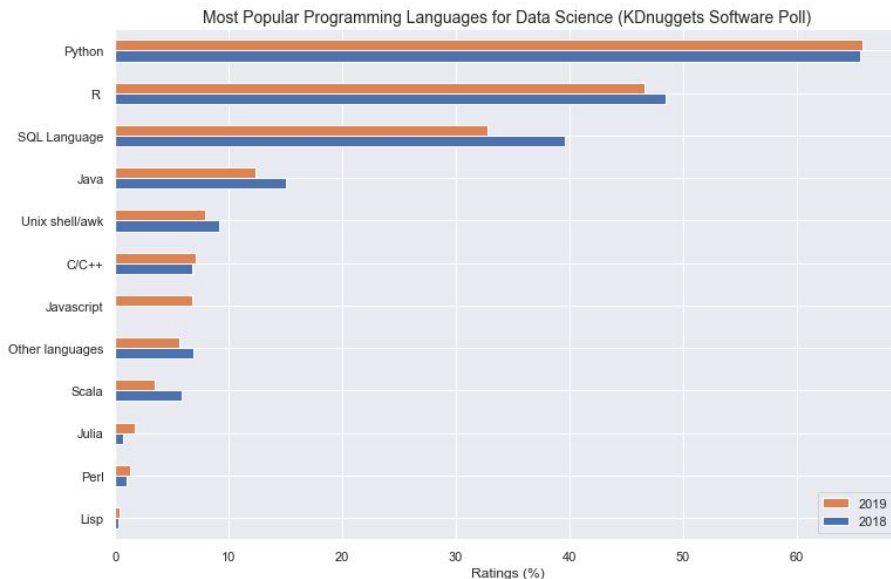
Install R and RStudio

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



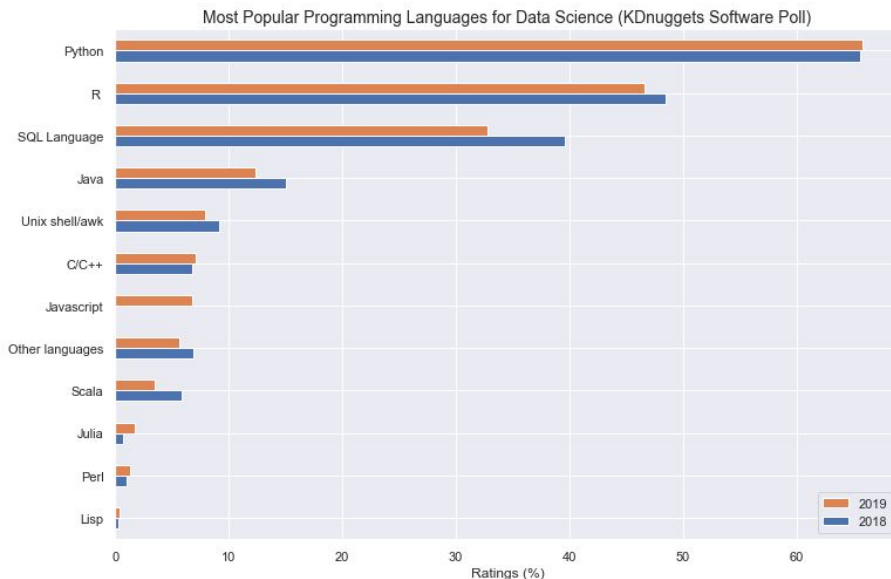
What's R?

- FREE and OPEN SOURCE statistical and computational software



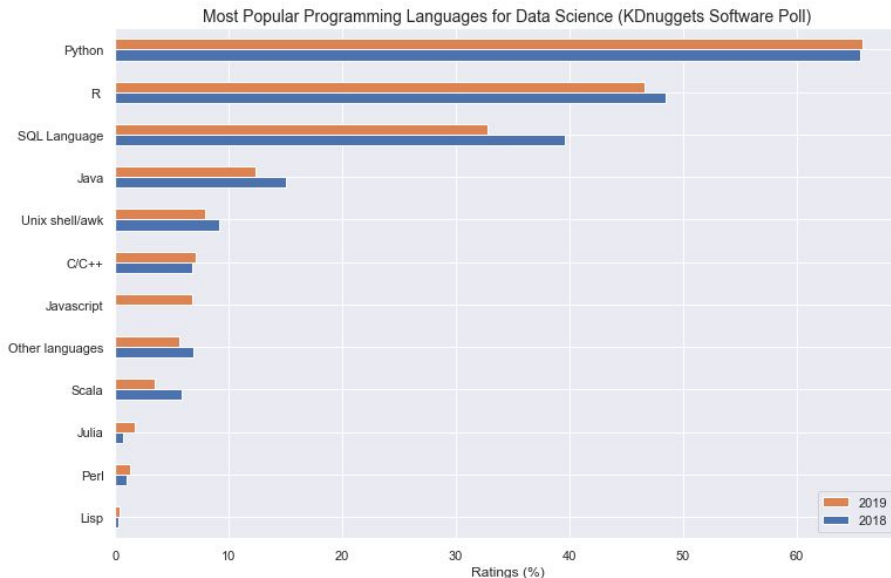
What's R?

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



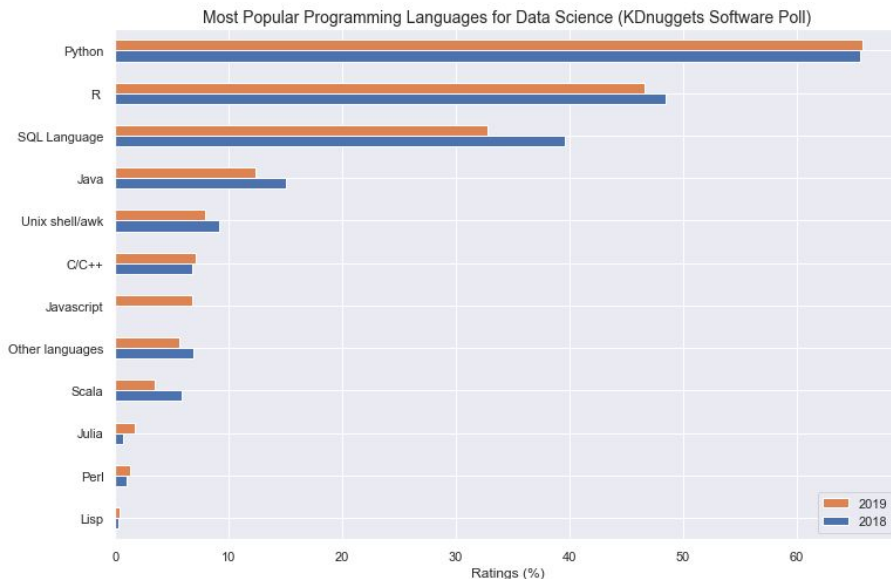
What's R?

- 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



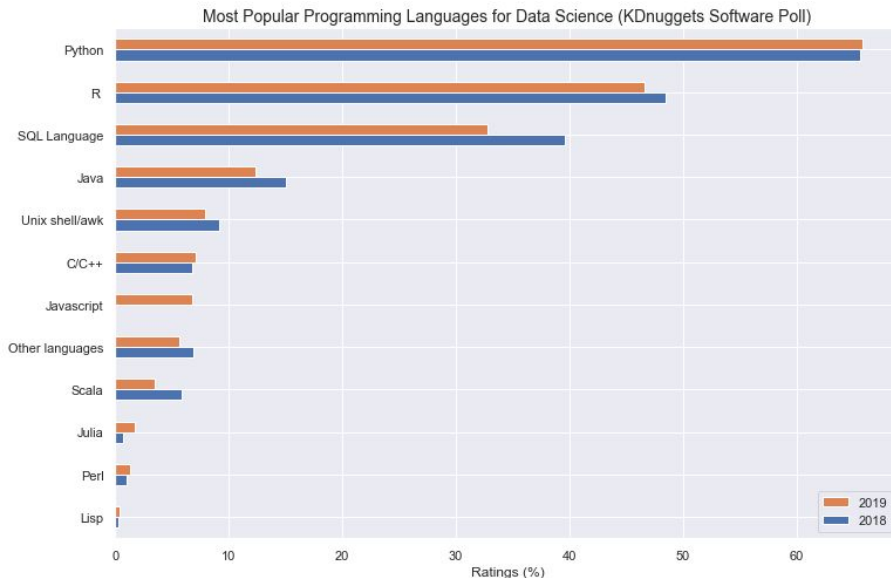
What's R?

- 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

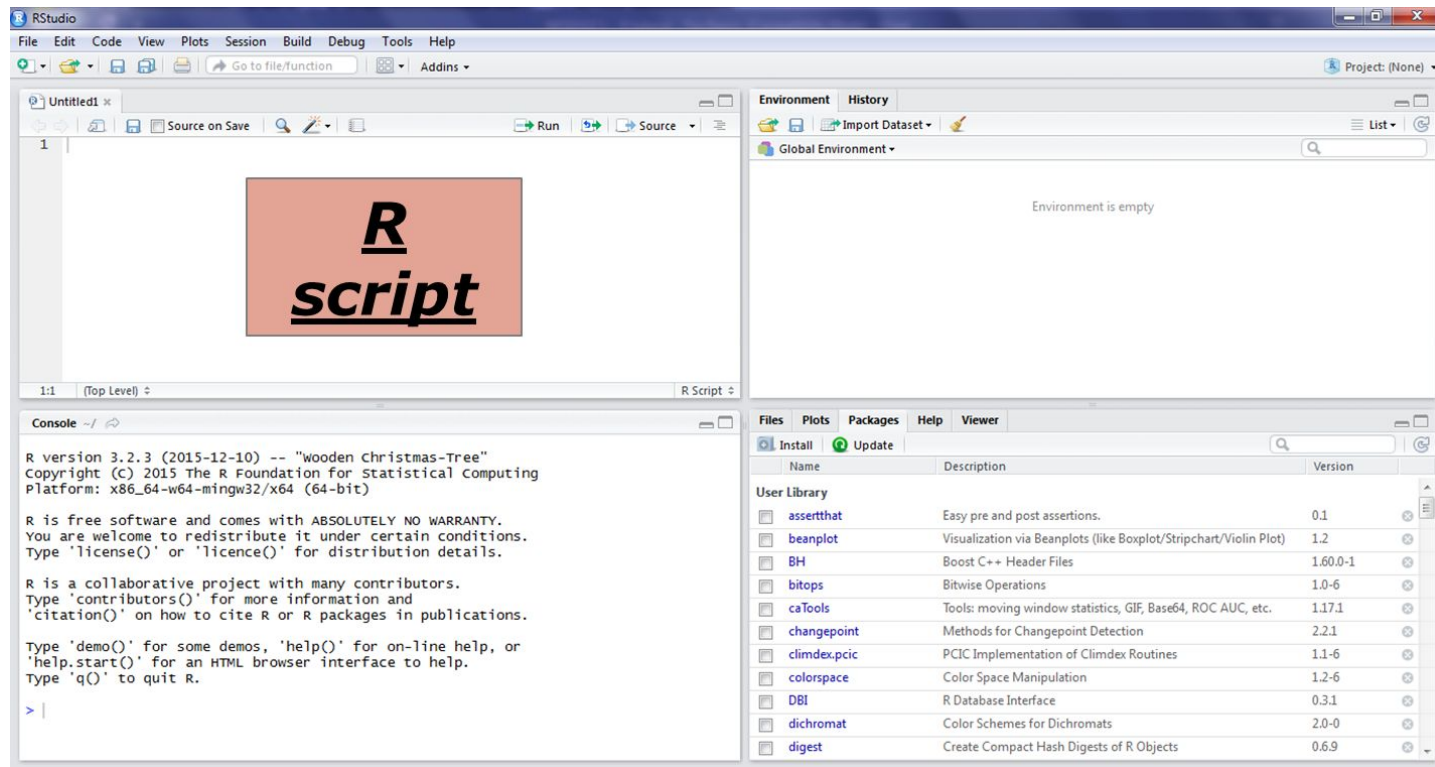


What's R?

- 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



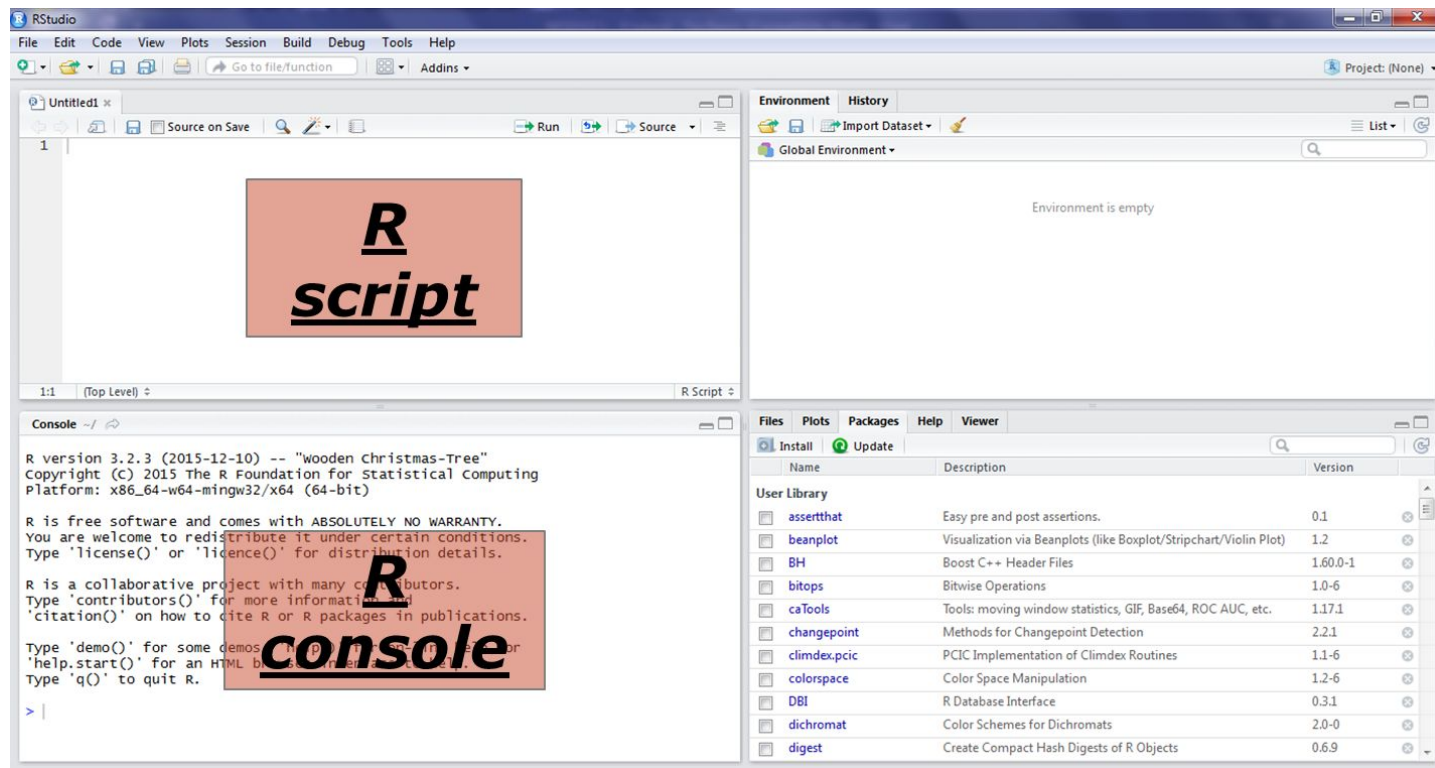
R script:

write and edit
code

color-coding for
easier
identification

automatically fills
in parentheses
and quotations

Open RStudio on your laptop now

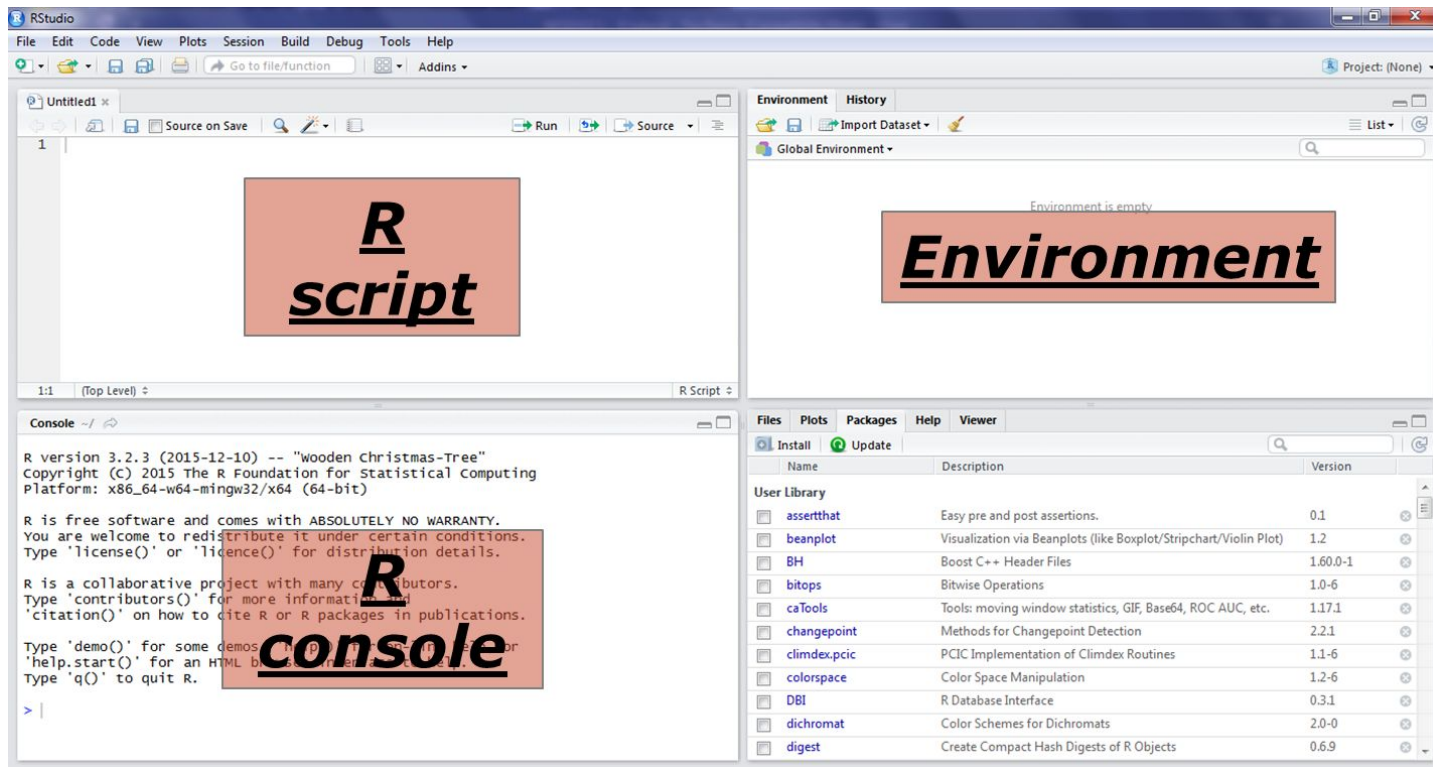


R console:

commands are
run and results
appear

> means 'ready to
work'

Open RStudio on your laptop now

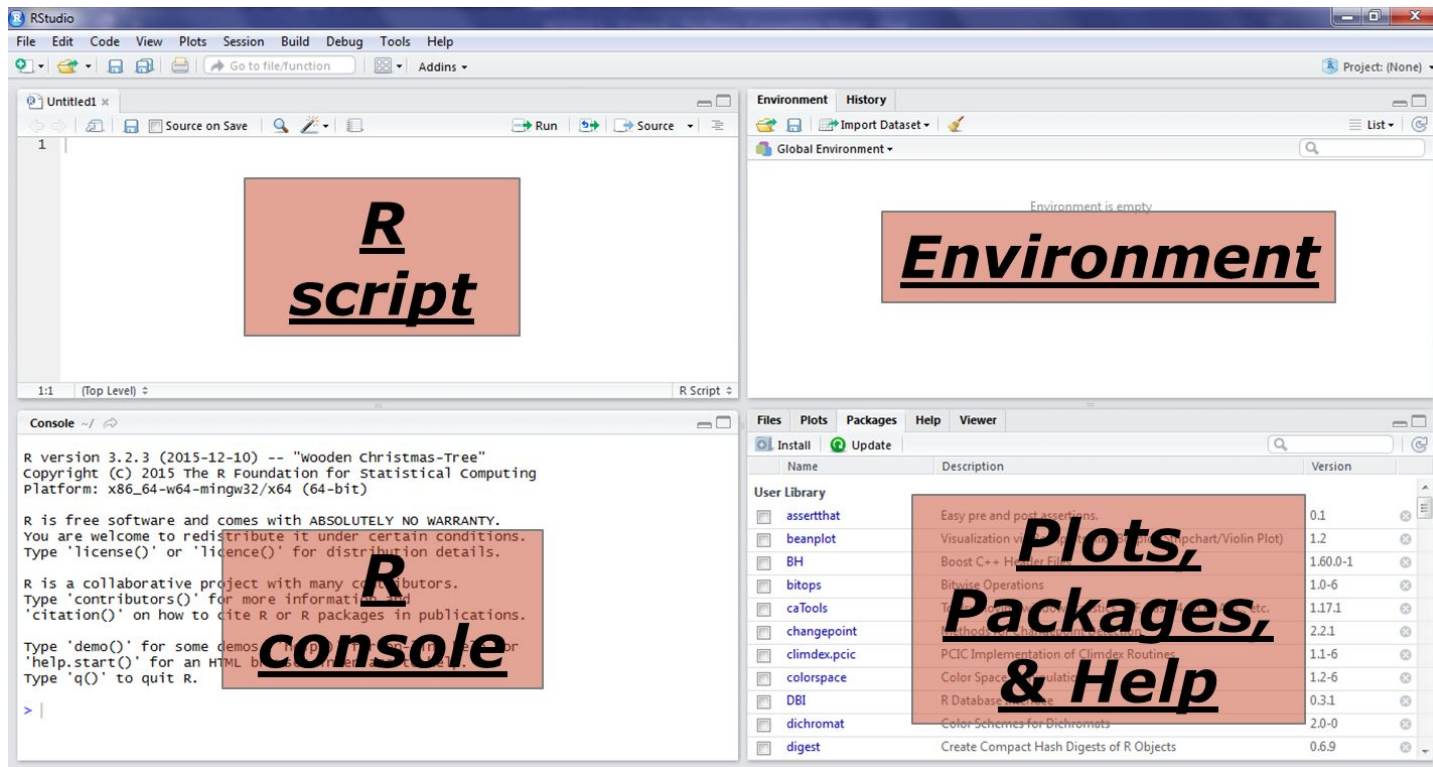


Environment:

see which objects were created

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

Open RStudio on your laptop now



**Plots/Packages/
Help:**

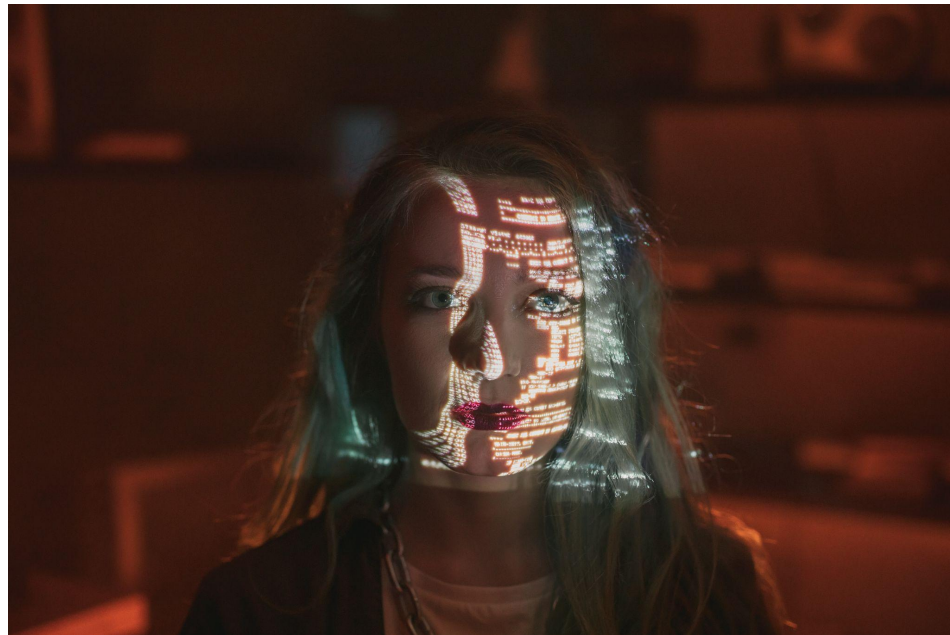
plots will appear

see which
packages are
installed

help functions

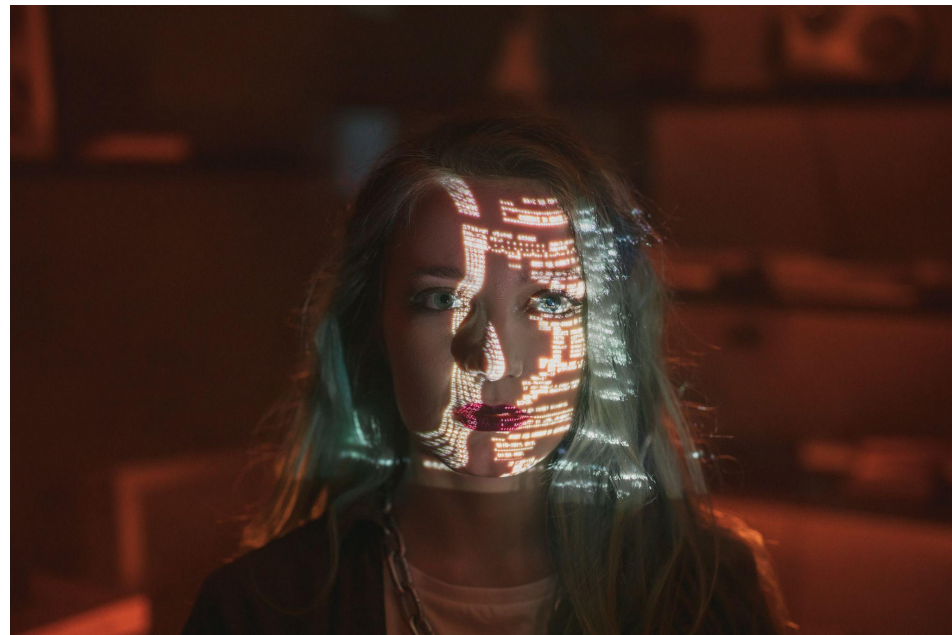
What are packages?

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



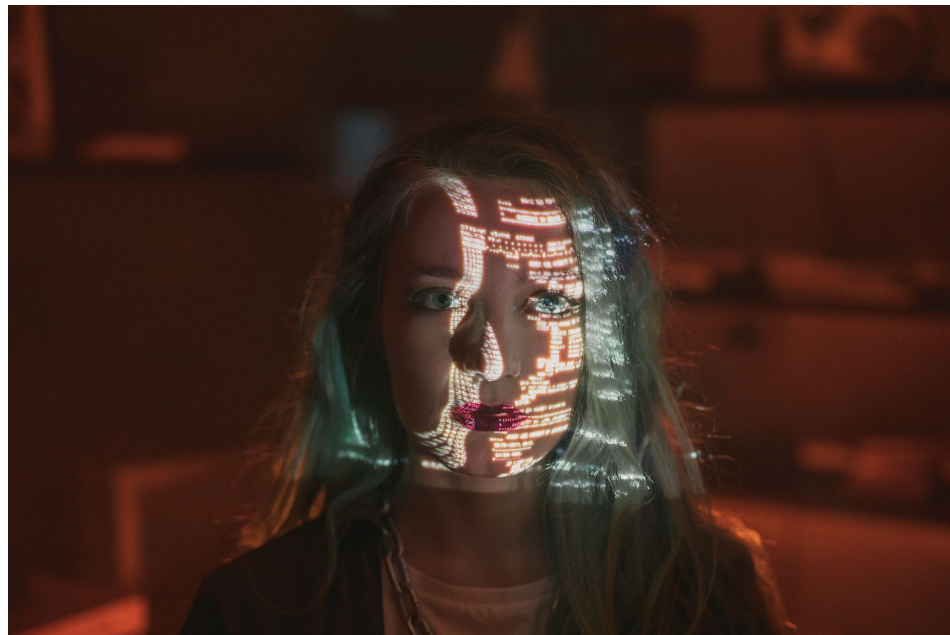
What are packages?

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



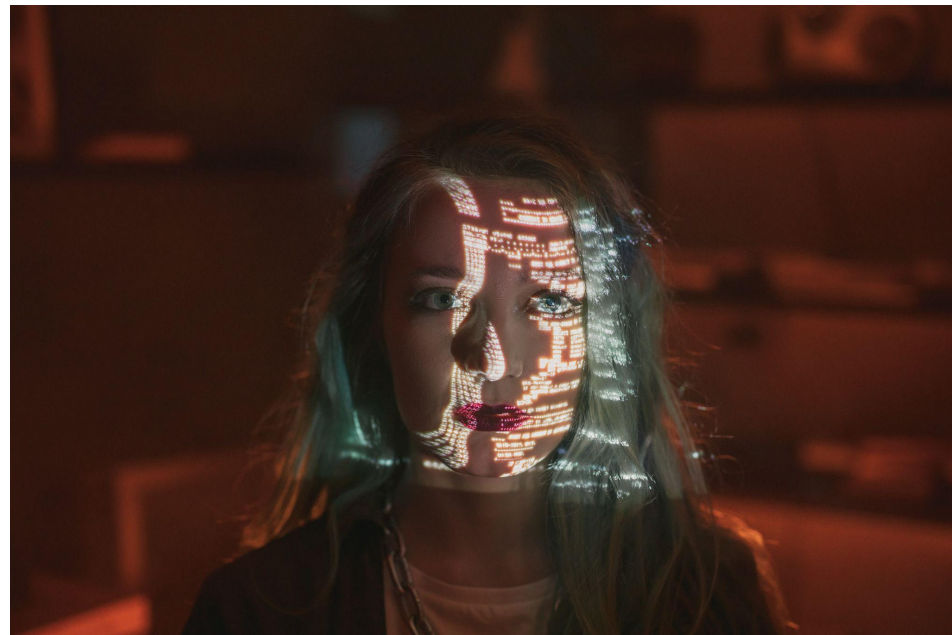
What are packages?

- 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

The screenshot shows the RStudio interface with the 'Tools' menu open, highlighting 'Install Packages...'. The console window at the bottom shows the following output:

```
> 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

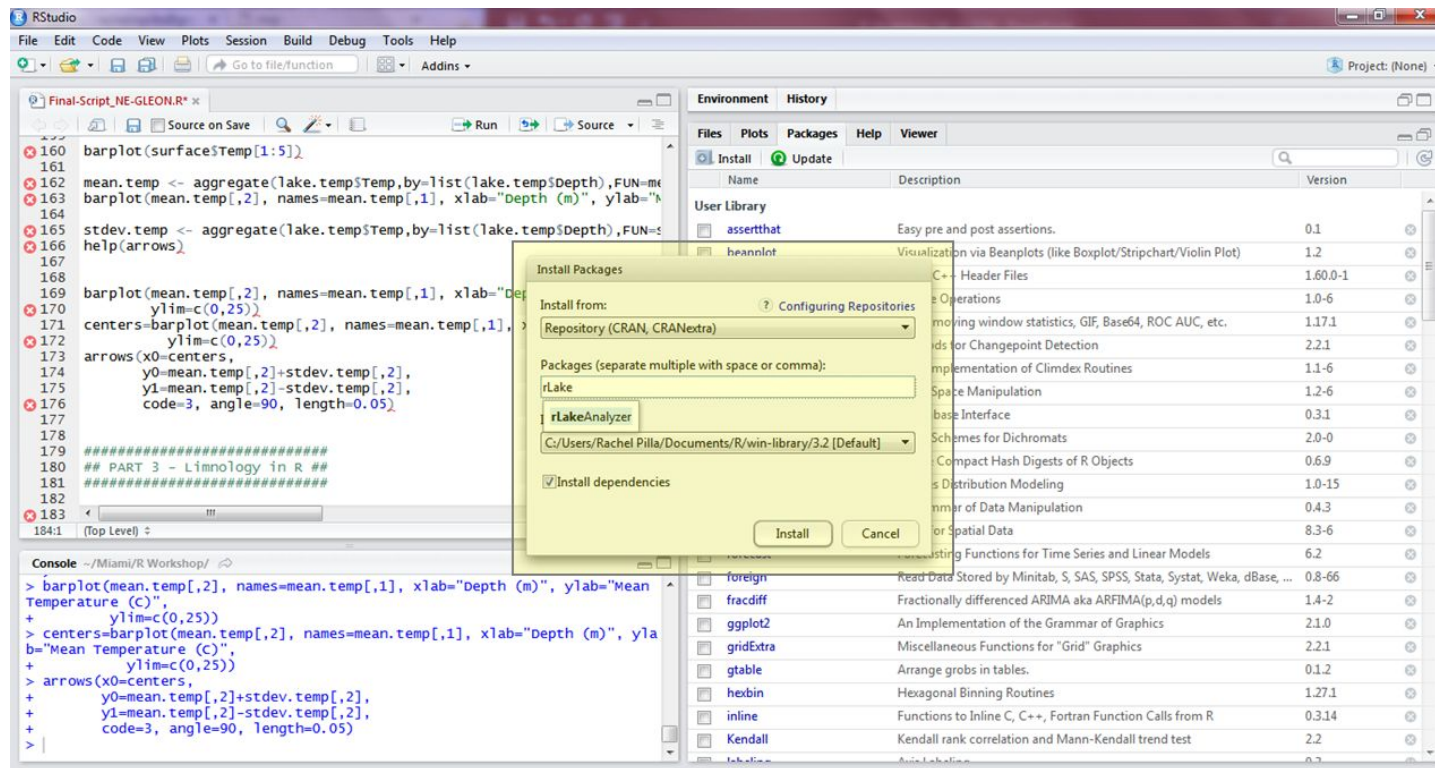
package 'rLakeAnalyzer' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
  C:\Users\Rachel Pilla\AppData\Local\Temp\RtmpmmkDw7\downloaded_packages
> |
```

we need:

tidyverse

Install one package



The screenshot shows the RStudio interface with the following components:

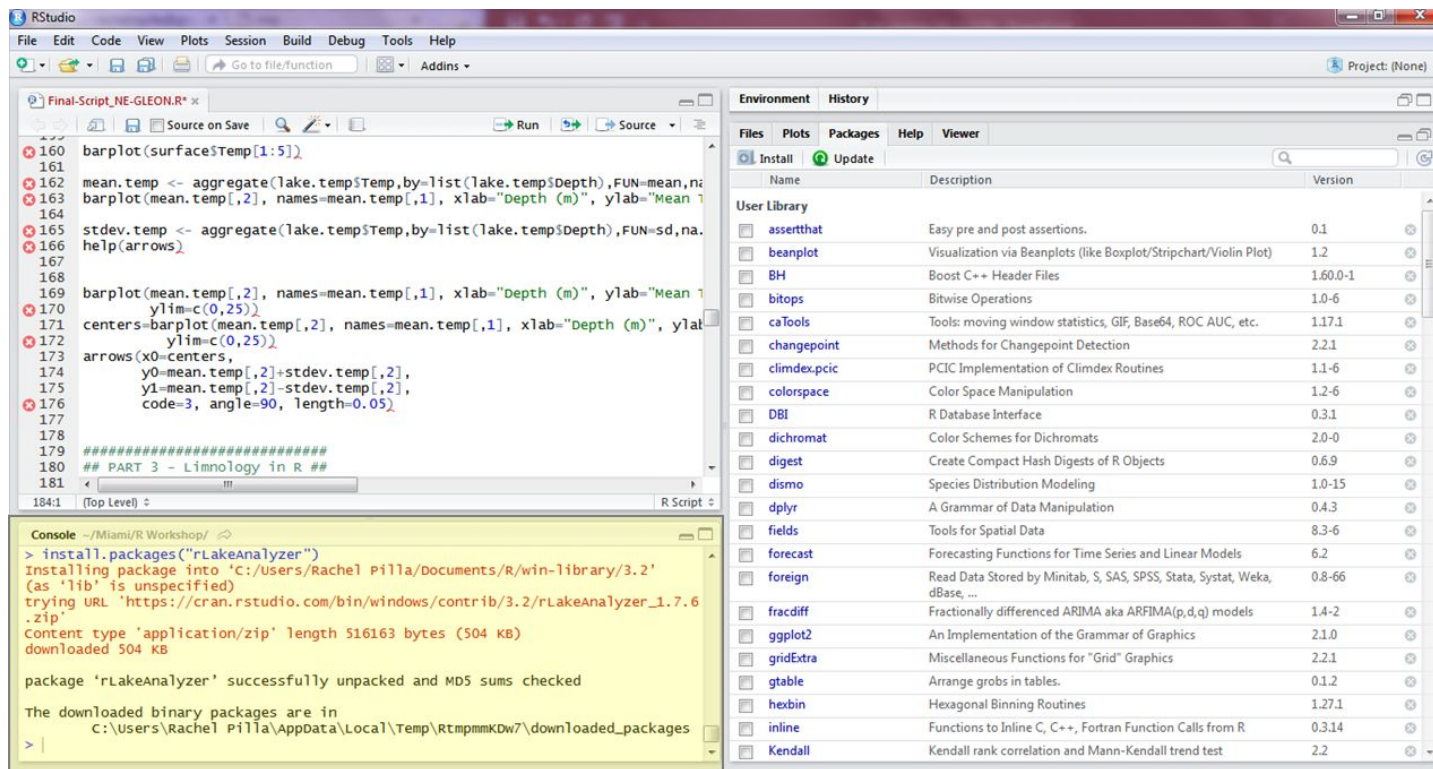
- Source Editor:** Contains R code for creating a barplot and arrows. The code is as follows:

```
barplot(surface$Temp[1:5])  
  
mean.temp <- aggregate(lake.temp$Temp, by=list(lake.temp$Depth), FUN=mean)  
barplot(mean.temp[,2], names=mean.temp[,1], xlab="Depth (m)", ylab="Mean Temperature (C)")  
  
stdev.temp <- aggregate(lake.temp$Temp, by=list(lake.temp$Depth), FUN=sd)  
help(arrows)  
  
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)", ylab="Mean Temperature (C)",  
               ylim=c(0,25))  
arrows(x0=centers[,1], y0=mean.temp[,2]+stdev.temp[,2],  
       x1=centers[,1], y1=mean.temp[,2]-stdev.temp[,2],  
       code=3, angle=90, length=0.05)
```
- Environment Pane:** Shows the User Library with various installed packages. The packages listed include: assertthat, beanint, C++ Header Files, C++ Operations, moving window statistics, GIF, Base64, ROC AUC, etc., vds for Changepoint Detection, implementation of Climdex Routines, Space Manipulation, base Interface, Schemes for Dichromats, Compact Hash Digests of R Objects, s Distribution Modeling, nmer of Data Manipulation, or Spatial Data, and Testing Functions for Time Series and Linear Models.
- Install Packages Dialog:** A dialog box is open, showing the repository (CRAN) and the package name (rLakeAnalyzer). The dialog also shows the installation path (C:/Users/Rachel Pilla/Documents/R/win-library/3.2 [Default]) and the option to install dependencies.

we need:

tidyverse

Install one package



The screenshot shows the RStudio interface with the following components:

- Script Editor:** Contains R code for plotting data. The code includes comments and function calls like `barplot`, `aggregate`, and `sd`.
- Console:** Shows the output of the `install.packages("rLakeAnalyzer")` command. The output indicates that the package was successfully installed into the user's library.
- Environment/History:** Shows a list of installed packages, including `assertthat`, `beanplot`, `BH`, `bitops`, `caTools`, `changepoint`, `climdex.pcic`, `colorspace`, `DBI`, `dichromat`, `digest`, `dismo`, `dplyr`, `fields`, `forecast`, `foreign`, `fracdiff`, `ggplot2`, `gridExtra`, `gtable`, `hexbin`, `inline`, and `Kendall`.

Console Output:

```
> 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\RtmpmmKw7\downloaded_packages
> |
```

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

The screenshot shows the RStudio interface with the following components:

- Source Editor:** Contains R code for data manipulation using `gather` and `dplyr`.


```

197 # reformatting from wide to long using "gather"
198
199 TempGather <- temp %>%
200   gather(key = "Depth", value = "Temperature", SurfaceWaterTemp, Bot
201
202
203 # grouping data and reformatting data with dplyr
204
205 ?InsectSprays|
206 head(InsectSprays)
207 str(InsectSprays)
208
209 InsectSummary <- InsectSprays %>%
210   group_by(spray) %>%
211   summarize(MeanCount = mean(count),
212             StDevCount = sd(count))
213
214
215
216
217
218
219
220
221
222
      
```
- Console:** Shows the output of the executed code.


```

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))
>
      
```
- Environment/History/Connections:** Shows a list of installed R packages.

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
gttable	Arrange 'Grob's' 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

The screenshot shows the RStudio interface with three main panels:

- Source Editor (Left):** Contains R code for data manipulation using `dplyr`. The code includes comments and function calls like `gather`, `group_by`, and `summarize`.
- Console (Bottom Left):** Shows the execution of the code, displaying the structure of the `InsectSprays` data frame and the resulting `InsectSummary` data frame.
- Help Viewer (Right):** Displays the documentation for the `dplyr` package version 0.7.5. It includes the title "A Grammar of Data Manipulation", a list of links for description and user guides, and a table of help pages.

Help Pages Table:

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



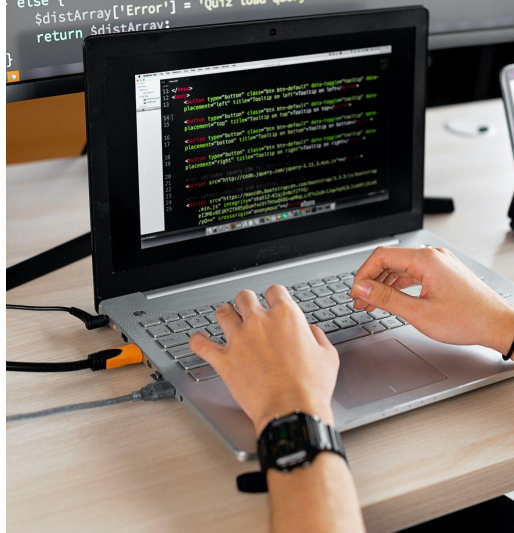
Coding in R

- 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

```
if(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

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 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+4i)
- **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!

