

R for Statistical Analysis

Introduction to R

- What R is
 - Statistical programming language
 - Command line interface (CLI)
 - Rstudio popular Graphical User Interface (GUI)
 - Open Source
 - Vector Mathematics
 - Most popular statistical language
 - Minitab, Spss, Jump, Excel competitors
 - Most have little programming language or difficult one
 - Layered Graphical Output
 - Allows multiple types of graphics on same output
 - Popular for cleaning and analyzing data
 - Tidyverse
 - Large active user base
 - Can be extended with packages
 - Often used for Machine Learning, AI

Introduction to R

- What R is not
 - No replacement for Hadoop or other such cloud data crunchers
 - Data must be in memory
 - Can extend with Big Memory package or MS R Open
 - Not a spreadsheet program
 - Not a general purpose language
 - C#, perl, etc.

Setup R

- Installing R from CRAN
 - Both 32 bit and 64 bit installs



CRAN
[Mirrors](#)
[What's new?](#)
[Task Views](#)
[Search](#)

About R
[R Homepage](#)
[The R Journal](#)

Software
[R Sources](#)
[R Binaries](#)
[Packages](#)
[Other](#)

Documentation
[Manuals](#)
[FAQs](#)
[Contributed](#)

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- [Download R for Linux](#)
- [Download R for \(Mac\) OS X](#)
- [Download R for Windows](#)

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2021-03-31, Shake and Throw) [R-4.0.5.tar.gz](#), read [what's new](#) in the latest version.
- Sources of [R alpha and beta releases](#) (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are [available here](#). Please read about [new features and bug fixes](#) before filing corresponding feature requests or bug reports.
- Source code of older versions of R is [available here](#).
- Contributed extension [packages](#)

Questions About R

- If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

What are R and CRAN?

R is 'GNU S', a freely available language and environment for statistical computing and graphics which provides a wide variety of statistical and graphical techniques: linear and nonlinear modelling, statistical tests, time series analysis, classification, clustering, etc. Please consult the [R project homepage](#) for further information.

CRAN is a network of ftp and web servers around the world that store identical, up-to-date, versions of code and documentation for R. Please use the CRAN [mirror](#) nearest to you to minimize network load.

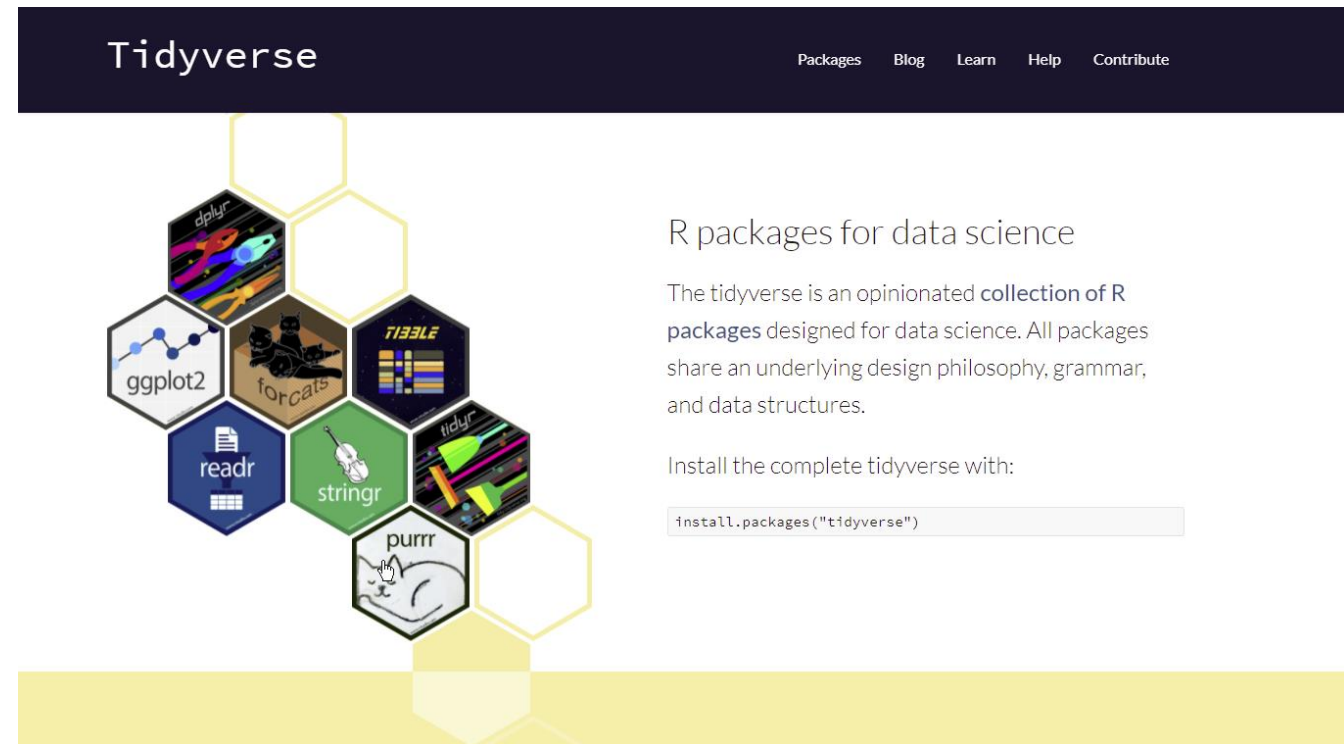
Setup R

- Install from Microsoft
 - R Open, mran.microsoft.com

The screenshot shows the Microsoft R Application Network (MRAN) website. At the top, there is a cookie consent banner. Below it, the header includes the title "Microsoft R Application Network" and navigation links for Home, About R, Microsoft R Open, R Packages, R Community, and R Tools. A search bar is located on the right side of the header. The main content area features a yellow banner with the text "Welcome to MRAN" and "Download Microsoft R Open 4.0.2 now." Below this is a large image of a person looking at a screen. The main content is divided into three columns: "Microsoft R Open" with a description and a "Download Now" button; "R Packages" with a description and an "Explore Packages" button; and "CRAN Time Machine" with a description and a "Browse Snapshots" button.

Setup R

- Install Rstudio
 - www.RStudio.com
 - Free
- Install packages to add capabilities to base R
 - TidyVerse
 - www.tidyverse.org
 - Packages on CRAN or through Rstudio console



The screenshot shows the Tidyverse website. At the top, the word "Tidyverse" is displayed in white on a dark blue background. To the right of the logo are navigation links: "Packages", "Blog", "Learn", "Help", and "Contribute". Below the header is a grid of hexagonal icons representing various R packages. The visible icons include: "dplyr" (a colorful abstract shape), "ggplot2" (a network diagram), "readr" (a document icon), "forcats" (a black cat), "stringr" (a violin), "purrr" (a white cat), "tidyr" (a colorful abstract shape), and "TIBBLE" (a colorful abstract shape). The website has a clean, modern design with a white background and a dark blue header.

R packages for data science

The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures.

Install the complete tidyverse with:

```
install.packages("tidyverse")
```

~/Flexible Nursing Home - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

data sampling.R x

Package triangle required but is not installed. Install Don't Show Again

```

1 RawData<-rnorm(1000,mean=15,sd=5.3)
2 Sample<-sample(RawData,size=35)
3 hist(Sample,col="yellow",xlim = c(0,30))
4 cBreaks<-seq(-5,35,1)
5 hist(RawData,col="yellow",xlim=c(-5,35),breaks=cBreaks)
6 require(triangle)
7 Sample<-as.data.frame(Sample)
8 Triage<-rltriangle(n=1000,a=7.9,b=25.8,c=15)
9 hist(Triage,col="yellow")
10

```

Environment History Connections

Global Environment

Data

- CorrRaw 610243 obs. of 9 variables
- data 2215 obs. of 34 variables
- data2 70880 obs. of 4 variables
- data3 70880 obs. of 4 variables
- Drive 1000 obs. of 1 variable
- fit Large lm (13 elements, 190.9 Mb)
- Sample 35 obs. of 1 variable
- sds List of 4
- sds1 List of 4
- sds2 List of 4
- sds3 List of 4
- Triage 1000 obs. of 1 variable

Files Plots Packages Help Viewer

Install Update Packrat

Name	Description	Version
User Library		
<input type="checkbox"/> abind	Combine Multidimensional Arrays	1.4-5
<input type="checkbox"/> acepack	ACE and AVAS for Selecting Multiple Regression Transformations	1.4.1
<input type="checkbox"/> adehabitatMA	Tools to Deal with Raster Maps	0.3.14
<input type="checkbox"/> AER	Applied Econometrics with R	1.2-9
<input type="checkbox"/> afex	Analysis of Factorial Experiments	0.28-1
<input type="checkbox"/> agricolae	Statistical Procedures for Agricultural Research	1.3-3
<input type="checkbox"/> akima	Interpolation of Irregularly and Regularly Spaced Data	0.6-2.1
<input type="checkbox"/> ALEPlot	Accumulated Local Effects (ALE) Plots and Partial Dependence (PD) Plots	1.1
<input type="checkbox"/> AlgDesign	Algorithmic Experimental Design	1.2.0
<input type="checkbox"/> alphavantage	Lightweight R Interface to the Alpha Vantage API	0.1.2
<input type="checkbox"/> anytime	Anything to 'POSIXct' or 'Date' Converter	0.3.9
<input type="checkbox"/> ash	David Scott's ASH Routines	1.0-15
<input type="checkbox"/> askpass	Safe Password Entry for R, Git, and SSH	1.1
<input type="checkbox"/> assertthat	Easy Pre and Post Assertions	0.2.1
<input type="checkbox"/> AUC	Threshold independent performance measures for probabilistic classifiers.	0.3.0
<input type="checkbox"/> backports	Reimplementations of Functions Introduced Since R-3.0.0	1.2.1
<input type="checkbox"/> base64enc	Tools for base64 encoding	0.1-3
<input type="checkbox"/> BayesFactor	Computation of Bayes Factors for Common Designs	0.9.12-4.2
<input type="checkbox"/> bayesplot	Plotting for Bayesian Models	1.8.0
<input type="checkbox"/> bayestestR	Understand and Describe Bayesian Models and Posterior Distributions	0.8.2
<input type="checkbox"/> bbmle	Tools for General Maximum Likelihood Estimation	1.0.23.1
<input type="checkbox"/> bdsmatrix	Routines for Block Diagonal Symmetric Matrices	1.3-4

1:1 (Top Level) R Script

Console Jobs

~/Flexible Nursing Home/

```

Natural language support but running in an English locale

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.

Microsoft R Open 4.0.2
The enhanced R distribution from Microsoft
Microsoft packages Copyright (C) 2020 Microsoft Corporation

Using the Intel MKL for parallel mathematical computing (using 4 cores).

Default CRAN mirror snapshot taken on 2020-07-16.
See: https://mran.microsoft.com/.

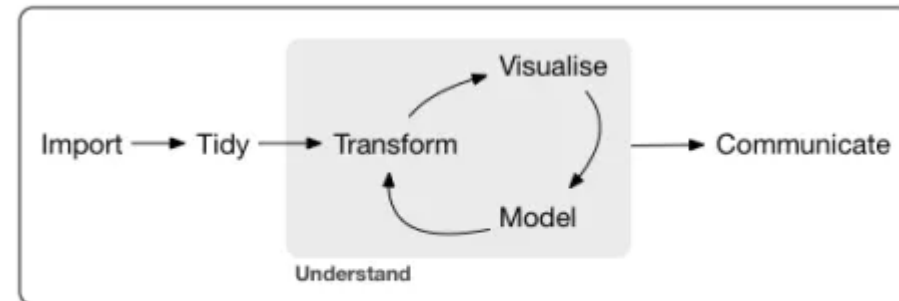
[Workspace loaded from ~/Flexible Nursing Home/.RData]
>

```

Using R to Clean and Prep Data

- **The Tidyverse**

- **ggplot2**: data visualization
- **dplyr**: data wrangling
- **readr**: reading data
- **tibble**: modern data frames
- **stringr**: string manipulation
- **forcats**: dealing with factors
- **tidyr**: data tidying
- **purrr**: functional programming



Program

The “tidy” workflow from *R for Data Science*

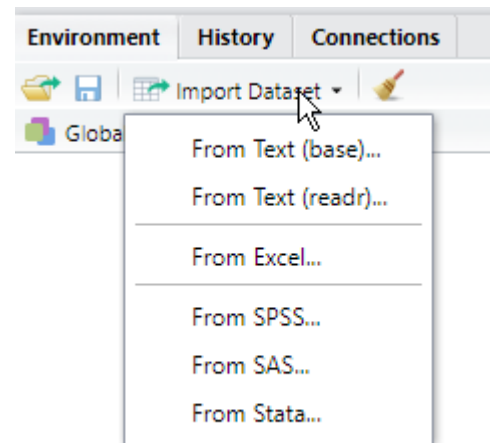
In tidy data:

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.

Wickham, Hadley “Tidy Data”, Journal of Statistical Software, Vol 59, 2014.

Using R to Clean and Prep Data

- R can take in almost any data format and from most sources
 - Supports excel, csv, json, etc.
 - Import local file or connect to web source
 - Can connect to databases through SQL
 - Many natively as well such as MySQL, MS Access, MS SQLServer, etc.
 - Can connect to Hadoop as a data source
- TidyVerse makes this easy



Using R to Clean and Prep Data

- Data Wrangle Demo in RStudio

Data Visualization

- Ggplot in Tidyverse
 - Graphical language
 - Multiple layers
- Histogram
 - With density fit
- Boxplot
 - Easy statistical significance comparison
- Time Series

Data Modeling

- Tidyverse include modelr
 - The modelr package, provides a few useful functions that are wrappers around base R's modeling functions.
- Base modeling functions
- Caret package

- Simple Linear modeling example

Data Modeling

- Caret is a modeling Suite
 - Includes tools to apportion data into train and validation sections
 - Will do bootstrapping and K folding
 - Caret support 233 different types of models
 - Caret supports ensemble modeling
 - As long as the base models are diverse and *independent*, the prediction error of the model decreases when the ensemble approach is used.
 - Different models for different sections of data
 - Different models for the whole data set
 - It's important to note that behind the scenes, caret is not actually performing the statistics/modeling – this job is left to individual R packages.
 - Has tools to compare between models as well

