Data visualization and analysis in R

Kristian Schultz

15.04.2025

www.sbi.uni-rostock.de

SYSTEMS BIOLOGY
BIOINFORMATICS
ROSTOCK

www.sbi.uni-rostock.de

Universität
Rostock

Traditio et Innovatio

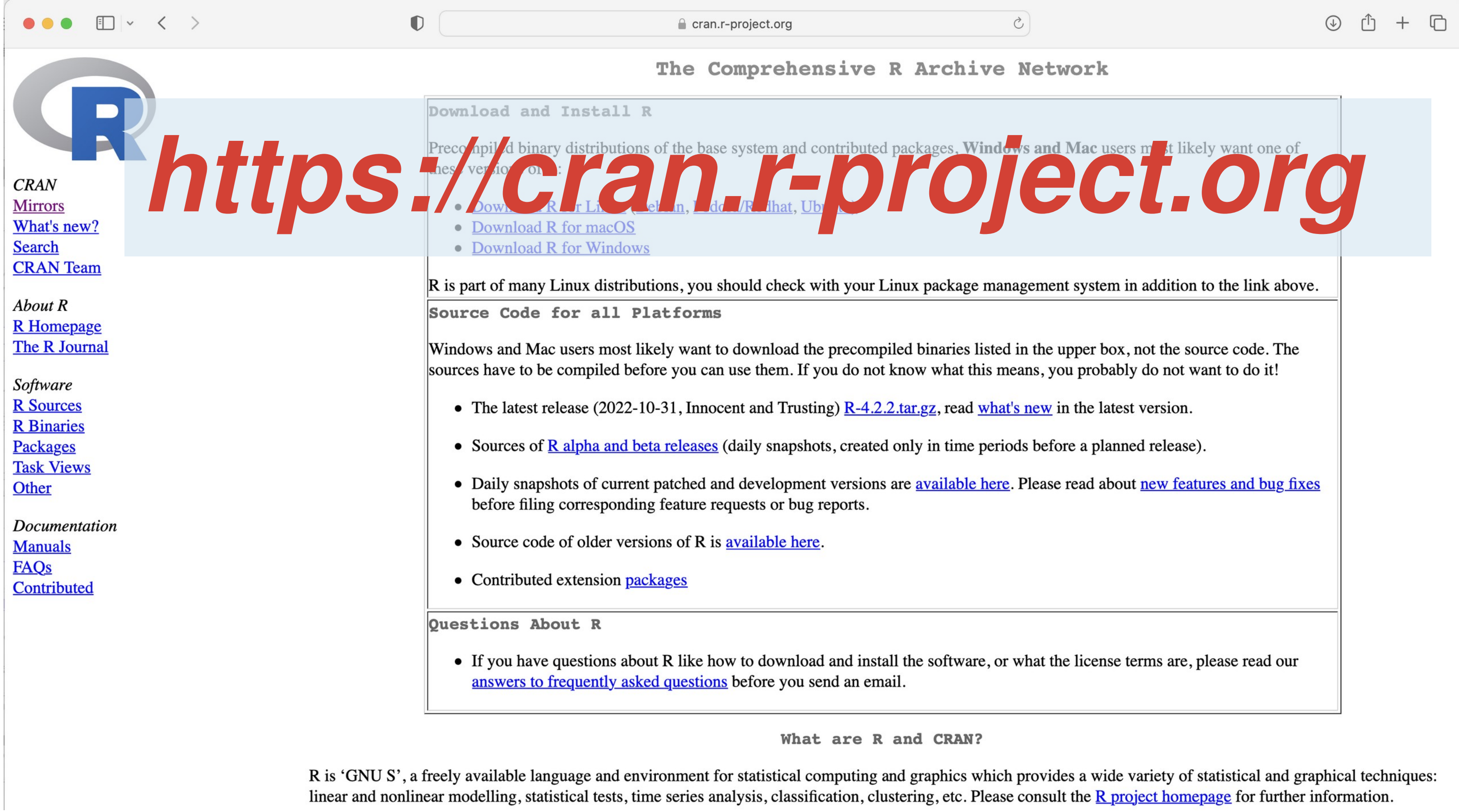
Data analysis with R

- *1992: Ross Ihaka and Robert Gentleman in the University of Auckland, New Zealand developed R.*
- *1995: The first version was released*
- *1997: CRAN (Comprehensive R Archive Network) was started*
- *2000: A stable beta version was released*
- *R has now thousands of packages, designed, maintained, and widely used by statisticians, biostatisticians, and geneticists*

A software environment used to analyze **statistical information**, **graphical representation**, **reporting**, and **data modeling**

Installing R

Installing R



https://cran.r-project.org

The Comprehensive R Archive Network

CRAN
Mirrors
What's new?
Search
CRAN Team

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Task Views
Other

Documentation
Manuals
FAQs
Contributed

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 (Debian, Fedora/Redhat, Ubuntu)
- Download R for macOS
- 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 (2022-10-31, Innocent and Trusting) R-4.2.2.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

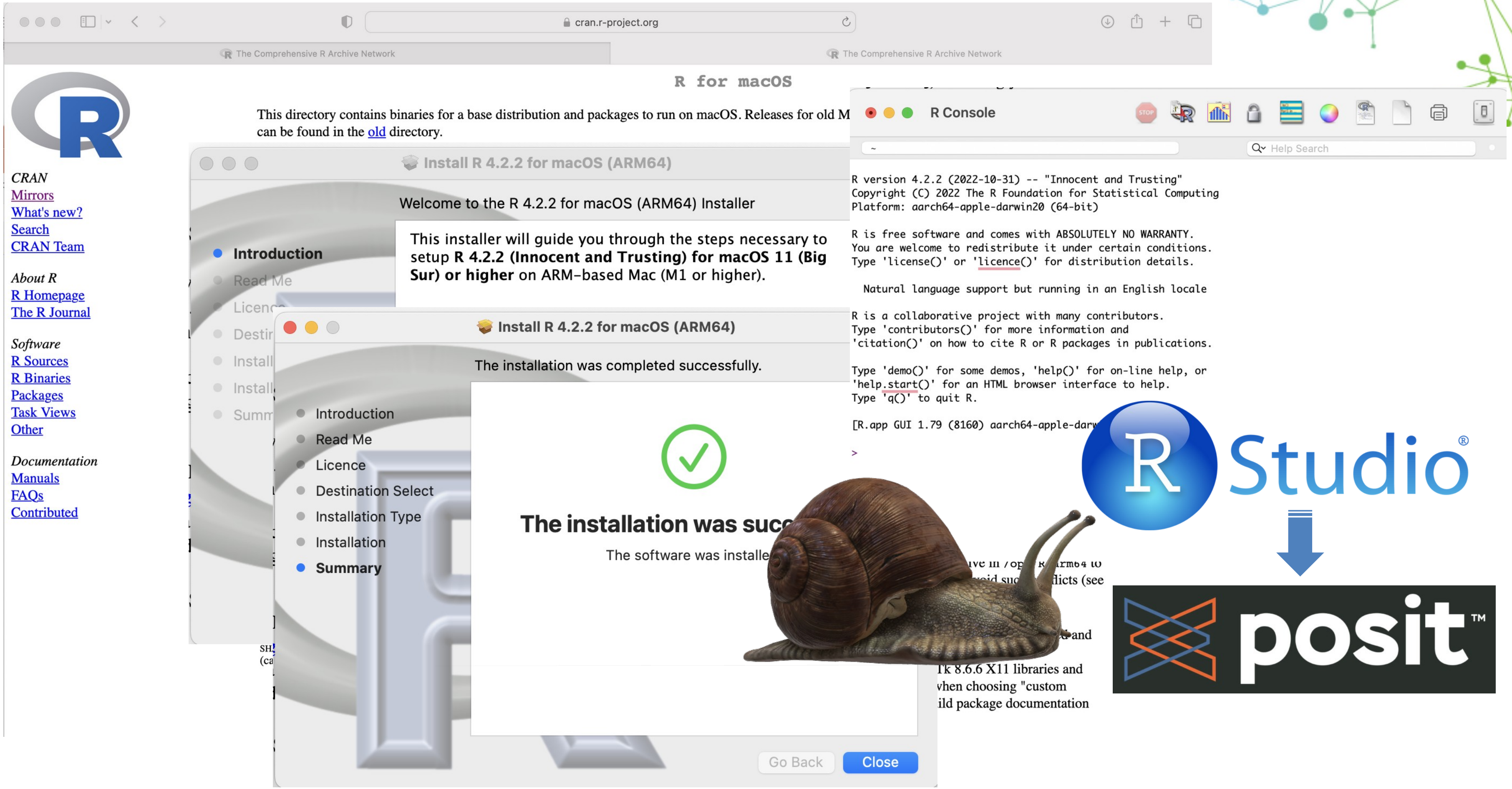
Questions About R

- 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.

Installing R



cran.r-project.org
The Comprehensive R Archive Network
The Comprehensive R Archive Network

R for macOS

This directory contains binaries for a base distribution and packages to run on macOS. Releases for old M
can be found in the old directory.

Install R 4.2.2 for macOS (ARM64)

Welcome to the R 4.2.2 for macOS (ARM64) Installer

This installer will guide you through the steps necessary to setup **R 4.2.2 (Innocent and Trusting) for macOS 11 (Big Sur) or higher** on ARM-based Mac (M1 or higher).

Install R 4.2.2 for macOS (ARM64)

The installation was completed successfully.

The installation was succ

The software was installe

R Console

```
R version 4.2.2 (2022-10-31) -- "Innocent and Trusting"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: aarch64-apple-darwin20 (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.

  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.

[R.app GUI 1.79 (8160) aarch64-apple-darw

>
```

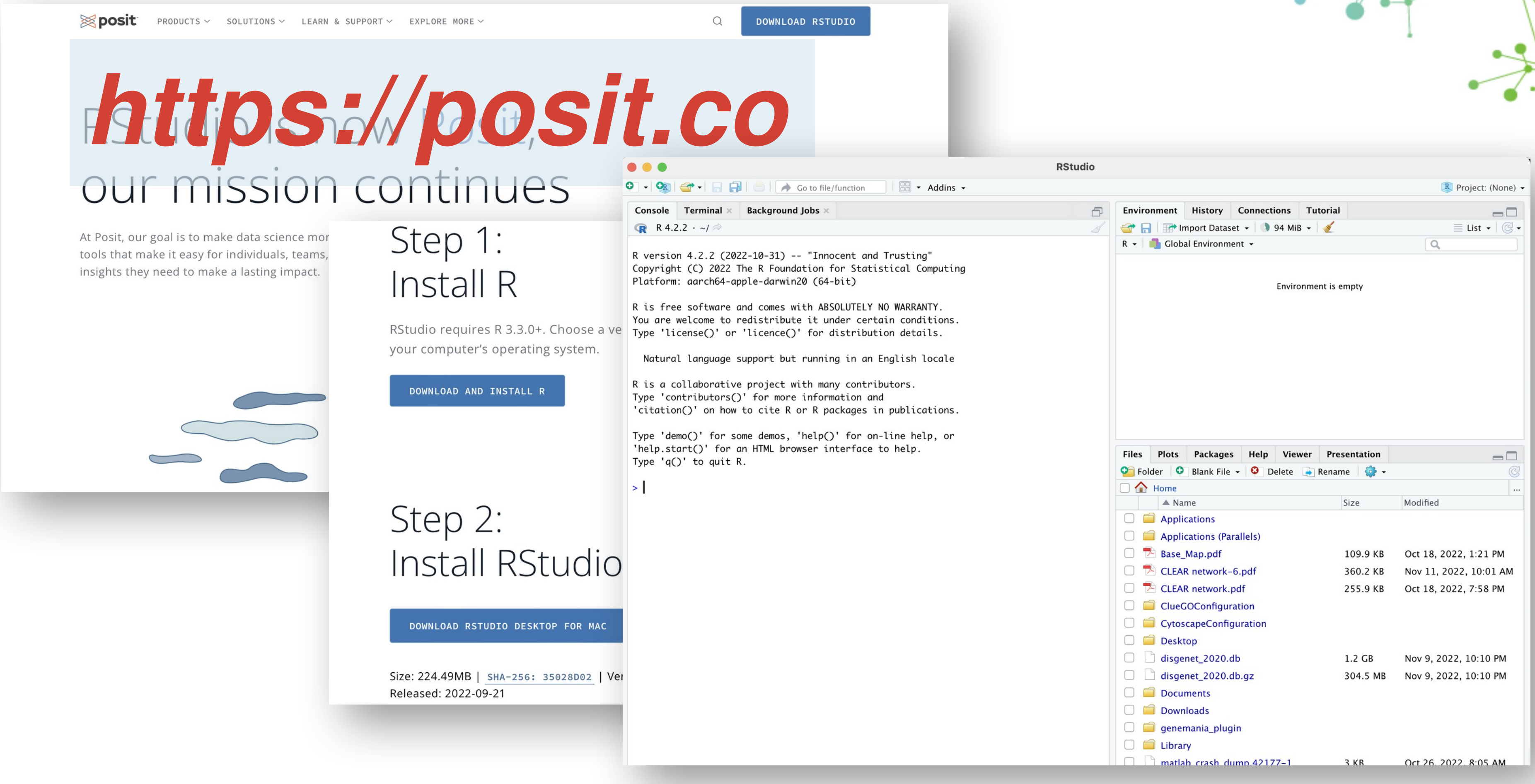
ve in /opt/R/arm64 to
void suc   dicts (see

and

Tk 8.6.6 X11 libraries and
when choosing "custom
ild package documentation

Installing RStudio Desktop

https://posit.co



RStudio is now Posit, our mission continues

At Posit, our goal is to make data science mor tools that make it easy for individuals, teams, insights they need to make a lasting impact.

Step 1: Install R

RStudio requires R 3.3.0+. Choose a ve your computer's operating system.

Step 2: Install RStudio

Size: 224.49MB | SHA-256: 35028D02 | Ver
Released: 2022-09-21

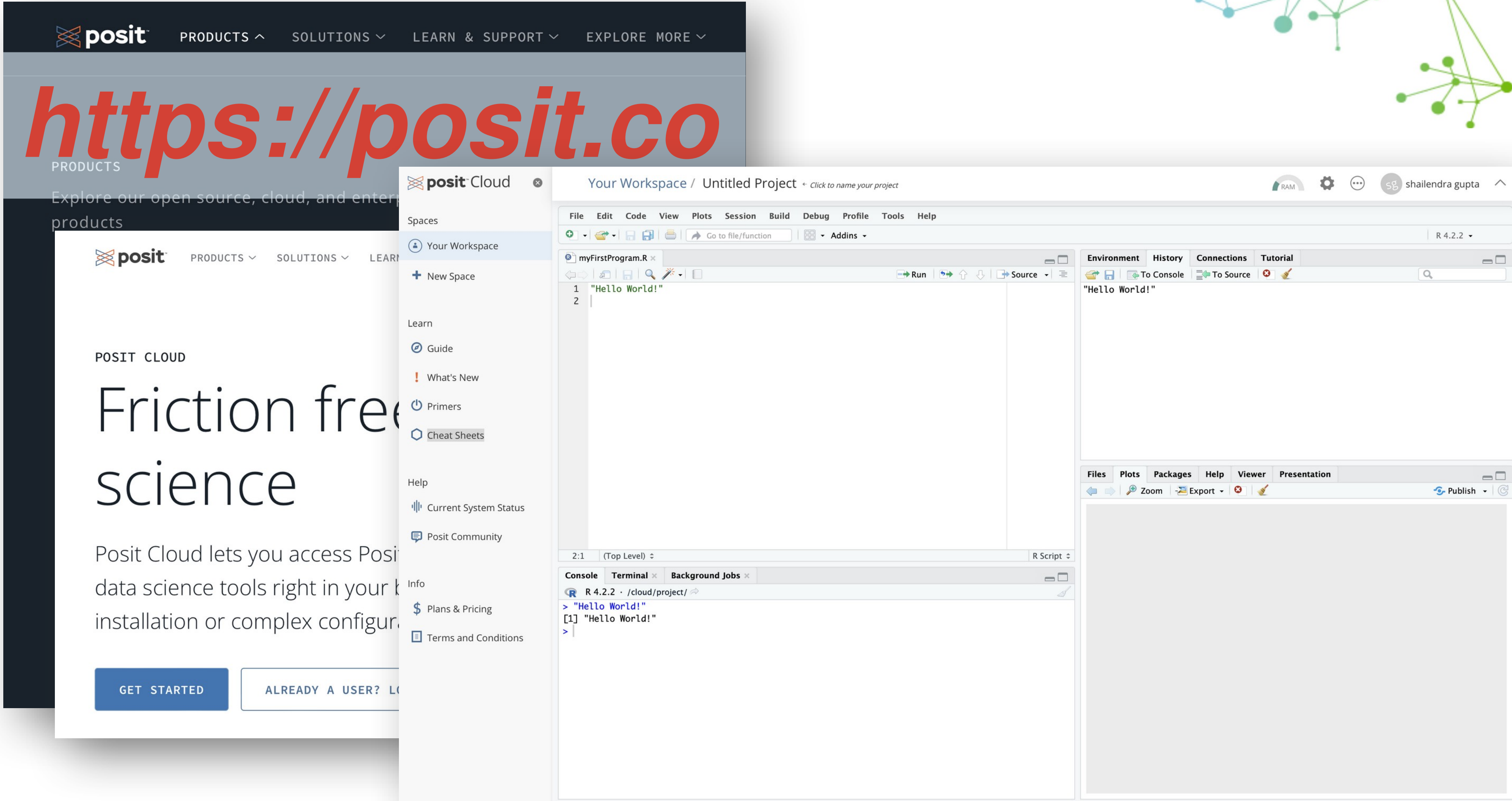
```
R version 4.2.2 (2022-10-31) -- "Innocent and Trusting"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: aarch64-apple-darwin20 (64-bit)

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.

  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.

> |
```

Environment is empty

Name	Size	Modified
Applications		
Applications (Parallels)		
Base_Map.pdf	109.9 KB	Oct 18, 2022, 1:21 PM
CLEAR network-6.pdf	360.2 KB	Nov 11, 2022, 10:01 AM
CLEAR network.pdf	255.9 KB	Oct 18, 2022, 7:58 PM
ClueGOConfiguration		
CytoscapeConfiguration		
Desktop		
disgenet_2020.db	1.2 GB	Nov 9, 2022, 10:10 PM
disgenet_2020.db.gz	304.5 MB	Nov 9, 2022, 10:10 PM
Documents		
Downloads		
genemania_plugin		
Library		
matlab_crash_dump.42177-1	3 KB	Oct 26, 2022, 8:05 AM

Running R on Posit Cloud

https://posit.co



Explore our open source, cloud, and enter

products

POSIT CLOUD

Friction free

science

Posit Cloud lets you access Posi

data science tools right in your b

installation or complex configur

Untitled Project · Click to name your project

myFirstProgram.R

```
"Hello World!"
```

History:

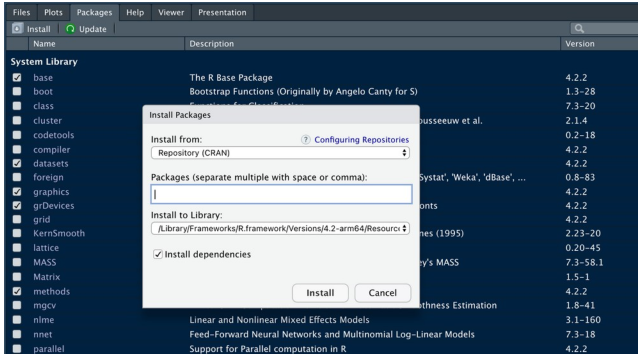
```
"Hello World!"
```

Console: R 4.2.2 · /cloud/project/

```
> "Hello World!"
[1] "Hello World!"
>
```

Packages in R

- Collection of R functions, data and compiled code in a well-defined format to perform specific task.
- Allow to expend the functionality available to you in R programming.
- Packages are stored in a directory called the library.
- Installing packages:
  - using gui
  - using command line in console e.g., install.packages(“ggplot2”)
- Activate install package
  - library(<name of the package>) e.g., library(ggplot2)



	Name	Description	Version
System Library			
☑	base	The R Base Package	4.2.2
□	boot	Bootstrap Functions (Originally by Angelo Canty for S)	1.3-28
□	class		7.3-20
□	cluster	ousseeuw et al.	2.1.4
□	codetools		0.2-18
□	compiler		4.2.2
☑	datasets		4.2.2
□	foreign	Systat', 'Weka', 'dBase', ...	0.8-83
☑	graphics		4.2.2
☑	grDevices	onts	4.2.2
□	grid		4.2.2
□	KernSmooth	nes (1995)	2.23-20
□	lattice		0.20-45
□	MASS	ey's MASS	7.3-58.1
□	Matrix		1.5-1
☑	methods		4.2.2
□	mgcv	othness Estimation	1.8-41
□	nlme	Linear and Nonlinear Mixed Effects Models	3.1-160
□	nnet	Feed-Forward Neural Networks and Multinomial Log-Linear Models	7.3-18
□	parallel	Support for Parallel computation in R	4.2.2

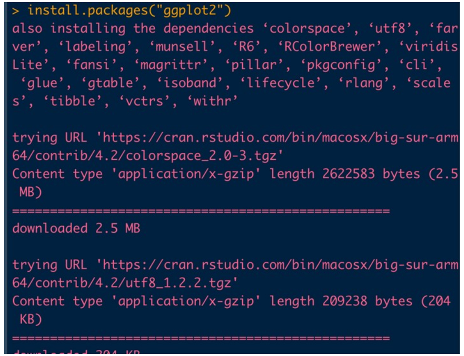
Install Packages

Install from: Repository (CRAN)

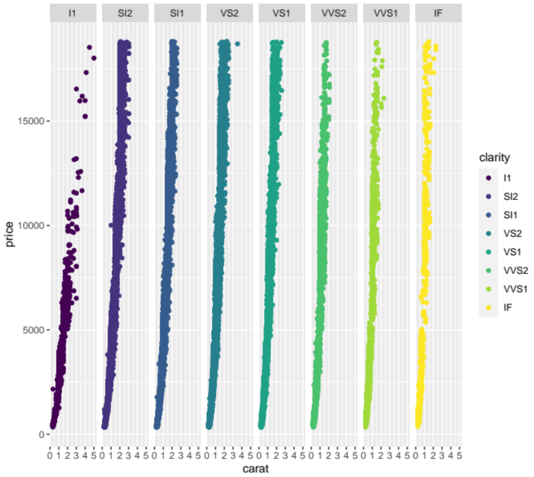
Packages (separate multiple with space or comma):

Install to Library: /Library/Frameworks/R.framework/Versions/4.2-arm64/Resourc

☑ Install dependencies

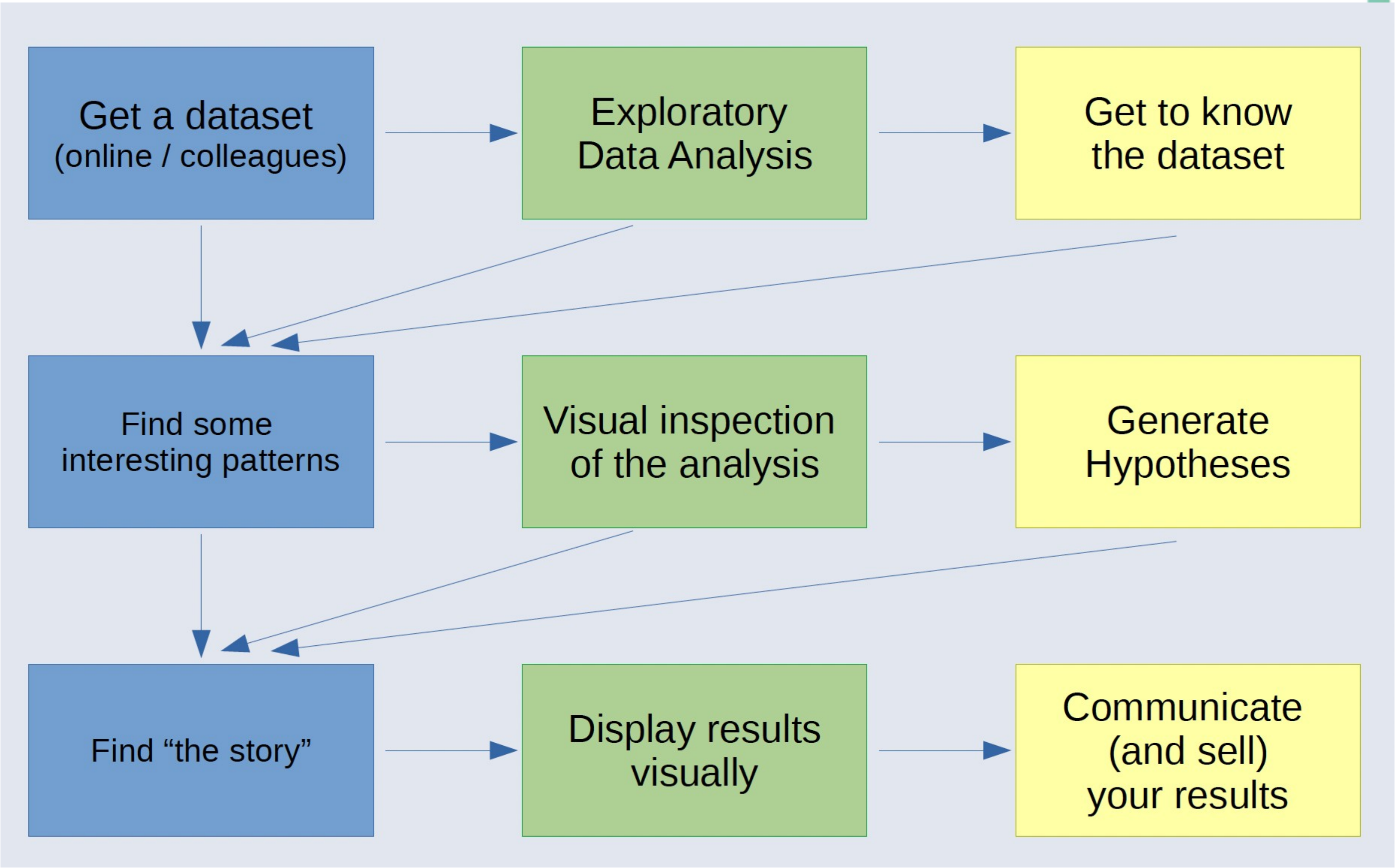


```
> install.packages("ggplot2")
also installing the dependencies ‘colorspace’, ‘utf8’, ‘far
ver’, ‘labeling’, ‘munsell’, ‘R6’, ‘RColorBrewer’, ‘viridis
Lite’, ‘fansi’, ‘magrittr’, ‘pillar’, ‘pkgconfig’, ‘cli’,
 ‘glue’, ‘gtable’, ‘isoband’, ‘lifecycle’, ‘rlang’, ‘scale
s’, ‘tibble’, ‘vctrs’, ‘withr’

trying URL 'https://cran.rstudio.com/bin/macosx/big-sur-arm
64/contrib/4.2/colorspace_2.0-3.tgz'
Content type 'application/x-gzip' length 2622583 bytes (2.5
 MB)
==================================================
downloaded 2.5 MB

trying URL 'https://cran.rstudio.com/bin/macosx/big-sur-arm
64/contrib/4.2/utf8_1.2.2.tgz'
Content type 'application/x-gzip' length 209238 bytes (204
 KB)
==================================================
```



- Example from ggplot2 (ggplots comes with few example datasets, e.g., diamonds, iris)
  - qplot(data=diamonds, carat, price, colour=clarity, facets=.~clarity)

The daily life of a data scientist..



Get a dataset (online / colleagues)

Exploratory Data Analysis

Get to know the dataset

Find some interesting patterns

Visual inspection of the analysis

Generate Hypotheses

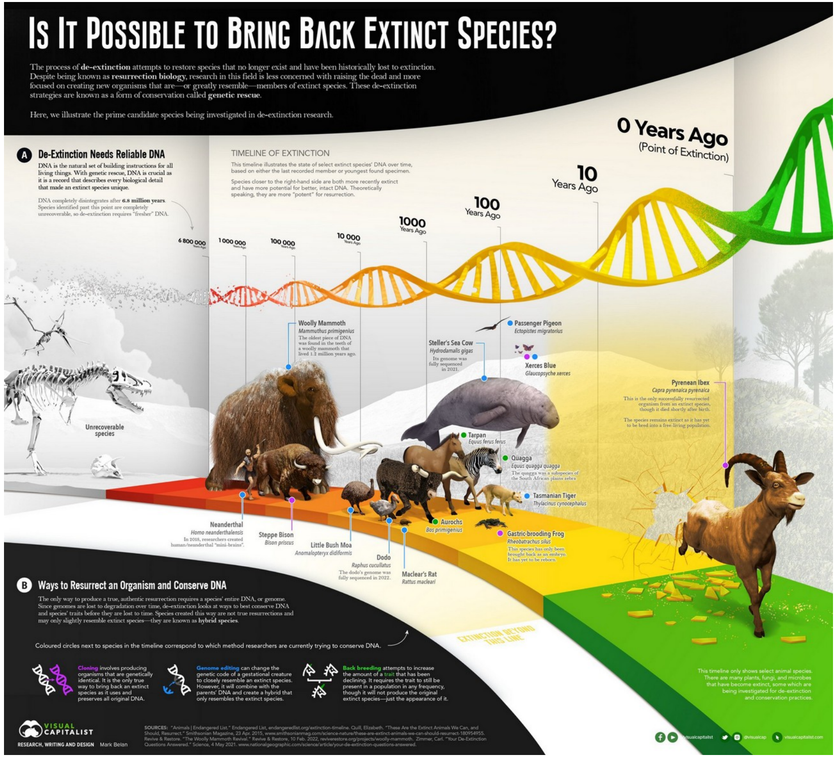
Find “the story”

Display results visually

Communicate (and sell) your results

Data visualization in R

Not this



Is It Possible to Bring Back Extinct Species?

The process of **de-extinction** attempts to restore species that no longer exist and have been historically lost to extinction. Despite being known as **resurrection biology**, research in this field is less concerned with raising the dead and more focused on creating new organisms that are—or greatly resemble—members of extinct species. These de-extinction strategies are known as a form of conservation called **genetic rescue**.

Here, we illustrate the prime candidate species being investigated in de-extinction research.

A De-Extinction Needs Reliable DNA

DNA is the natural set of building instructions for all living things. With genetic rescue, DNA is crucial as it is a record that describes every biological detail that made an extinct species unique.

DNA completely disintegrates after **6.8 million years**. Species identified past this point are completely unrecoverable, so de-extinction requires "fresher" DNA.

TIMELINE OF EXTINCTION

This timeline illustrates the state of select extinct species' DNA over time, based on either the last recorded member or youngest found specimen.

Species closer to the right-hand side are both more recently extinct and have more potential for better, intact DNA. Theoretically speaking, they are more "potent" for resurrection.

Woolly Mammoth *Mammuthus primigenius*: The oldest piece of DNA was found in the teeth of a woolly mammoth that lived 1.2 million years ago.

Steller's Sea Cow *Hydrodamalis gigas*: Its genome was fully sequenced in 2021.

Pyrenean Ibex *Capra pyrenaica pyrenaica*: This is the only successfully resurrected organism from an extinct species, though it died shortly after birth. The species remains extinct as it has yet to be bred into a free-living population.

Quagga *Equus quagga quagga*: The quagga was a subspecies of the South African plains zebra

Gastric-brooding Frog *Rheobatrachus silus*: This species has only been brought back as an embryo. It has yet to be reborn.

Neanderthal *Homo neanderthalensis*: In 2018, researchers created human/neanderthal "mini-brains".

Dodo *Raphus cucullatus*: The dodo's genome was fully sequenced in 2022.

B Ways to Resurrect an Organism and Conserve DNA

The only way to produce a true, authentic resurrection requires a species' entire DNA, or genome. Since genomes are lost to degradation over time, de-extinction looks at ways to best conserve DNA and species' traits before they are lost to time. Species created this way are not true resurrections and may only slightly resemble extinct species—they are known as **hybrid species**.

Coloured circles next to species in the timeline correspond to which method researchers are currently trying to conserve DNA.

Cloning involves producing organisms that are genetically identical. It is the only true way to bring back an extinct species as it uses and preserves all original DNA.

Genome editing can change the genetic code of a gestational creature to closely resemble an extinct species. However, it will combine with the parents' DNA and create a hybrid that only resembles the extinct species.

Back breeding attempts to increase the amount of a trait that has been declining. It requires the trait to still be present in a population in any frequency, though it will not produce the original extinct species—just the appearance of it.

This timeline only shows select animal species. There are many plants, fungi, and microbes that have become extinct, some which are being investigated for de-extinction and conservation practices.

VISUAL CAPITALIST

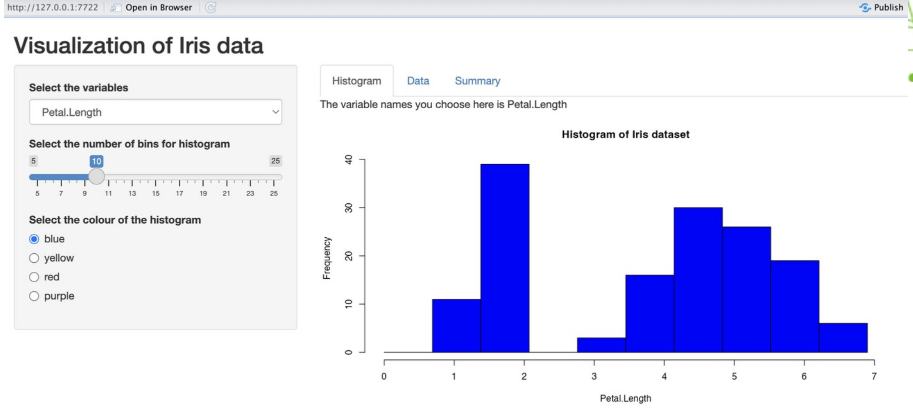
RESEARCH, WRITING AND DESIGN Mark Belan

SOURCES: "Animals | Endangered List." Endangered List, endangeredlist.org/extinction-timeline. Quill, Elizabeth. "These Are the Extinct Animals We Can, and Should, Resurrect." Smithsonian Magazine, 23 Apr. 2015, www.smithsonianmag.com/science-nature/these-are-extinct-animals-we-can-should-resurrect-180954955. Revive & Restore. "The Woolly Mammoth Revival." Revive & Restore, 10 Feb. 2022, reviverestore.org/projects/woolly-mammoth. Zimmer, Carl. "Your De-Extinction Questions Answered." Science, 4 May 2021. www.nationalgeographic.com/science/article/your-de-extinction-questions-answered.

Twitter

@markabelan

But this



http://127.0.0.1:7722 | Open in Browser | Publish

Visualization of Iris data

Select the variables

Petal.Length

Select the number of bins for histogram

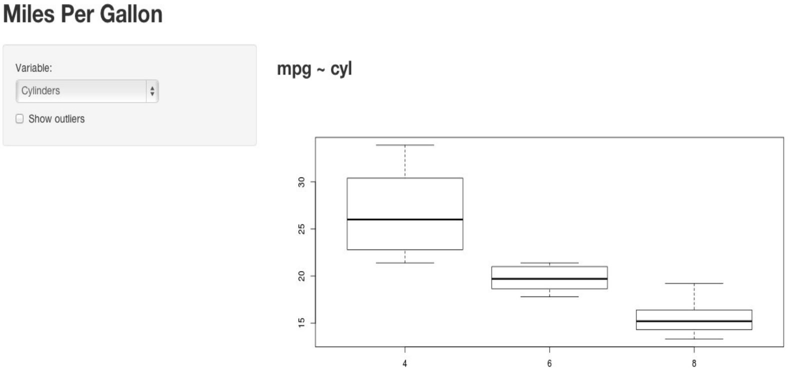
10

Select the colour of the histogram

- ● blue
- ○ yellow
- ○ red
- ○ purple

Histogram | Data | Summary

The variable names you choose here is Petal.Length



Miles Per Gallon

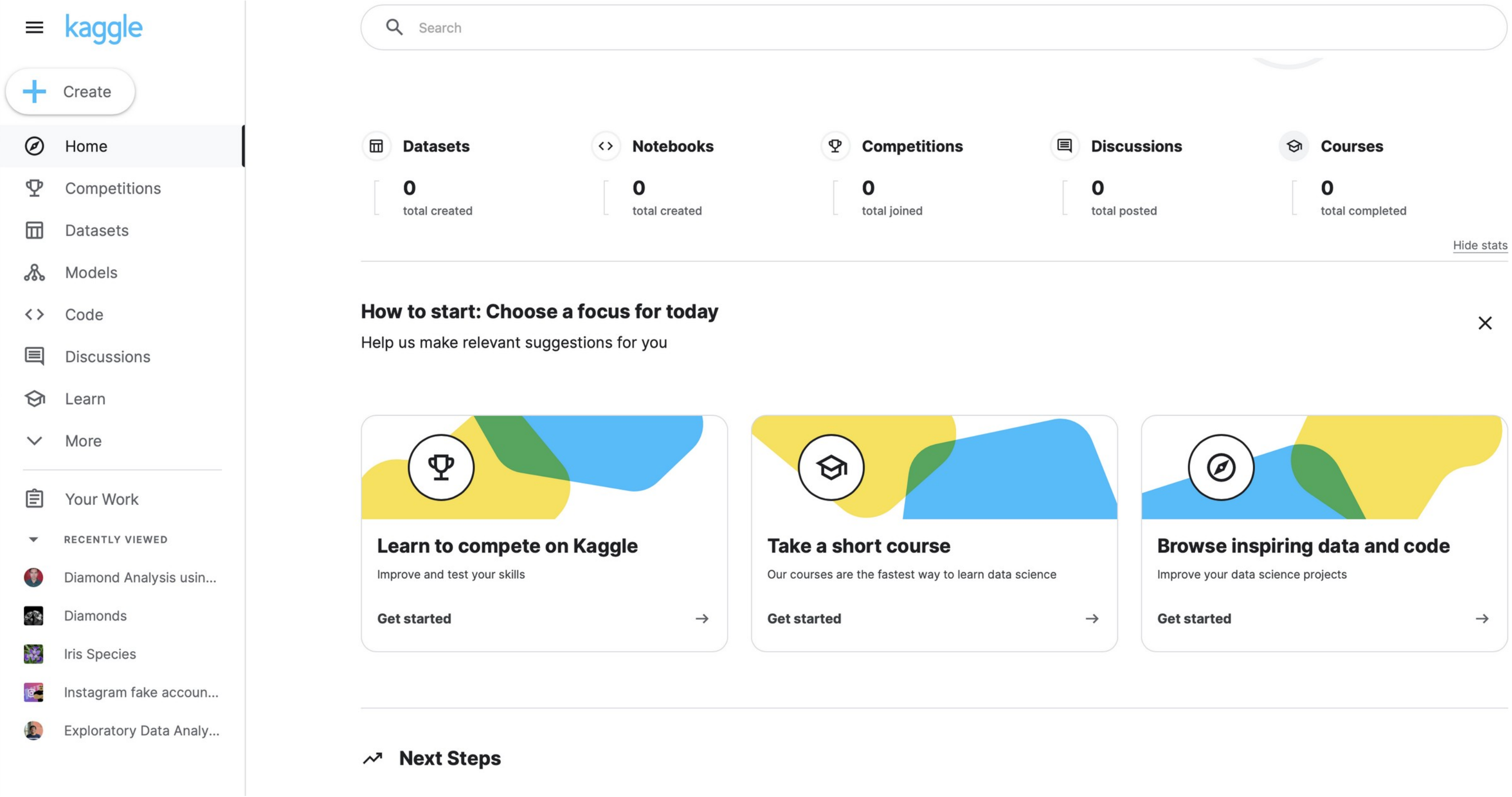
Variable:

Cylinders

☐ Show outliers

mpg ~ cyl

Getting the data



kaggle

Create

Home

Competitions

Datasets

Models

Code

Discussions

Learn

More

Your Work

RECENTLY VIEWED

Diamond Analysis usin...

Diamonds

Iris Species

Instagram fake accoun...

Exploratory Data Analy...

Search

Datasets
0
total created

Notebooks
0
total created

Competitions
0
total joined

Discussions
0
total posted

Courses
0
total completed

Hide stats

How to start: Choose a focus for today

Help us make relevant suggestions for you

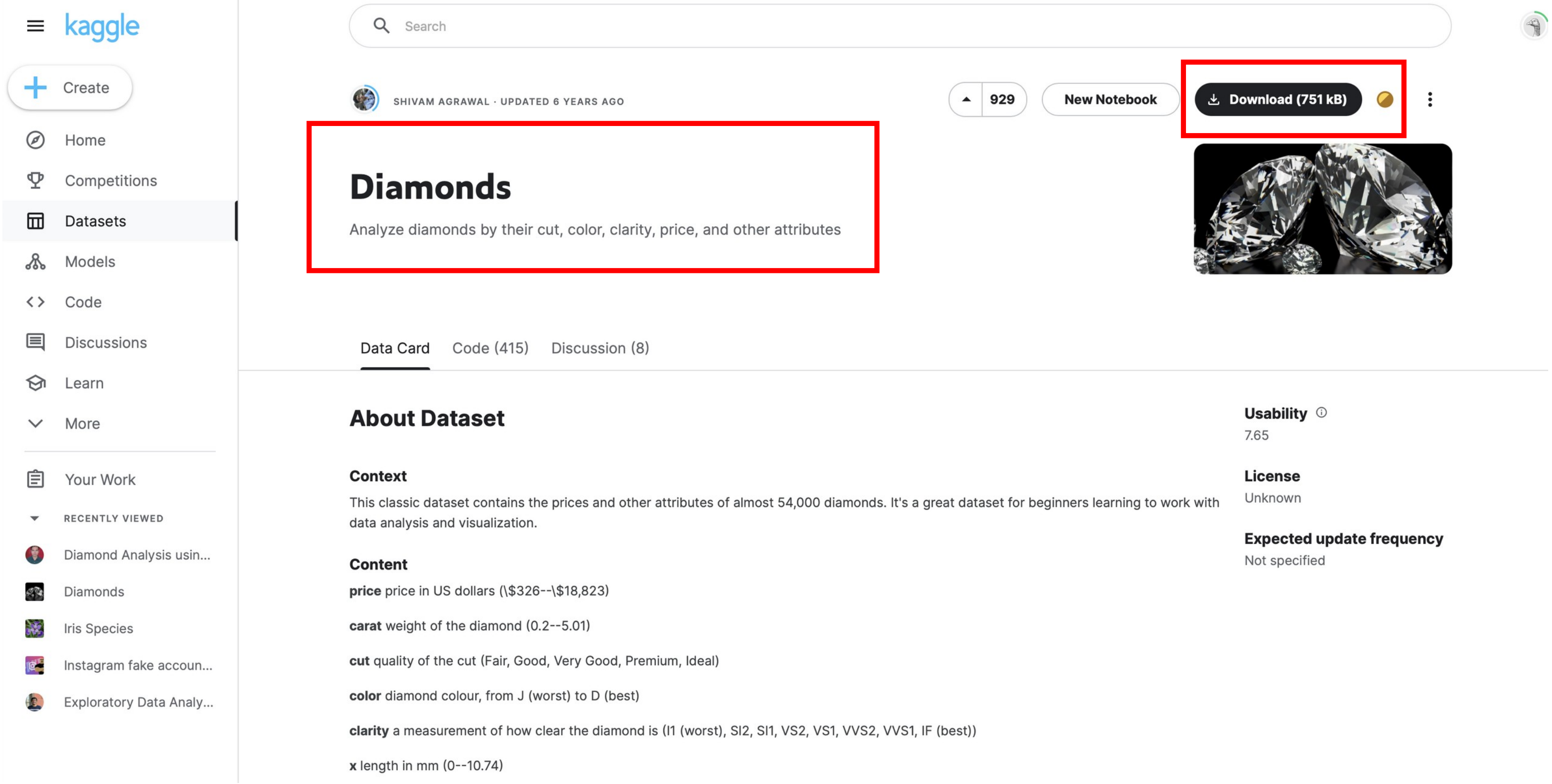
Learn to compete on Kaggle
Improve and test your skills
Get started →

Take a short course
Our courses are the fastest way to learn data science
Get started →

Browse inspiring data and code
Improve your data science projects
Get started →

Next Steps

Getting the data



Diamonds

Analyze diamonds by their cut, color, clarity, price, and other attributes

About Dataset

Context

This classic dataset contains the prices and other attributes of almost 54,000 diamonds. It's a great dataset for beginners learning to work with data analysis and visualization.

Content

price price in US dollars (\$326--\$18,823)

carat weight of the diamond (0.2--5.01)

cut quality of the cut (Fair, Good, Very Good, Premium, Ideal)

color diamond colour, from J (worst) to D (best)

clarity a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))

x length in mm (0--10.74)

Usability
7.65

License
Unknown

Expected update frequency
Not specified

Exploratory Data Analysis in R

Exploratory Data Analysis

- Exploratory Data Analysis (EDA) is an approach to analyzing datasets to summarize their main characteristics, often employing visual methods.

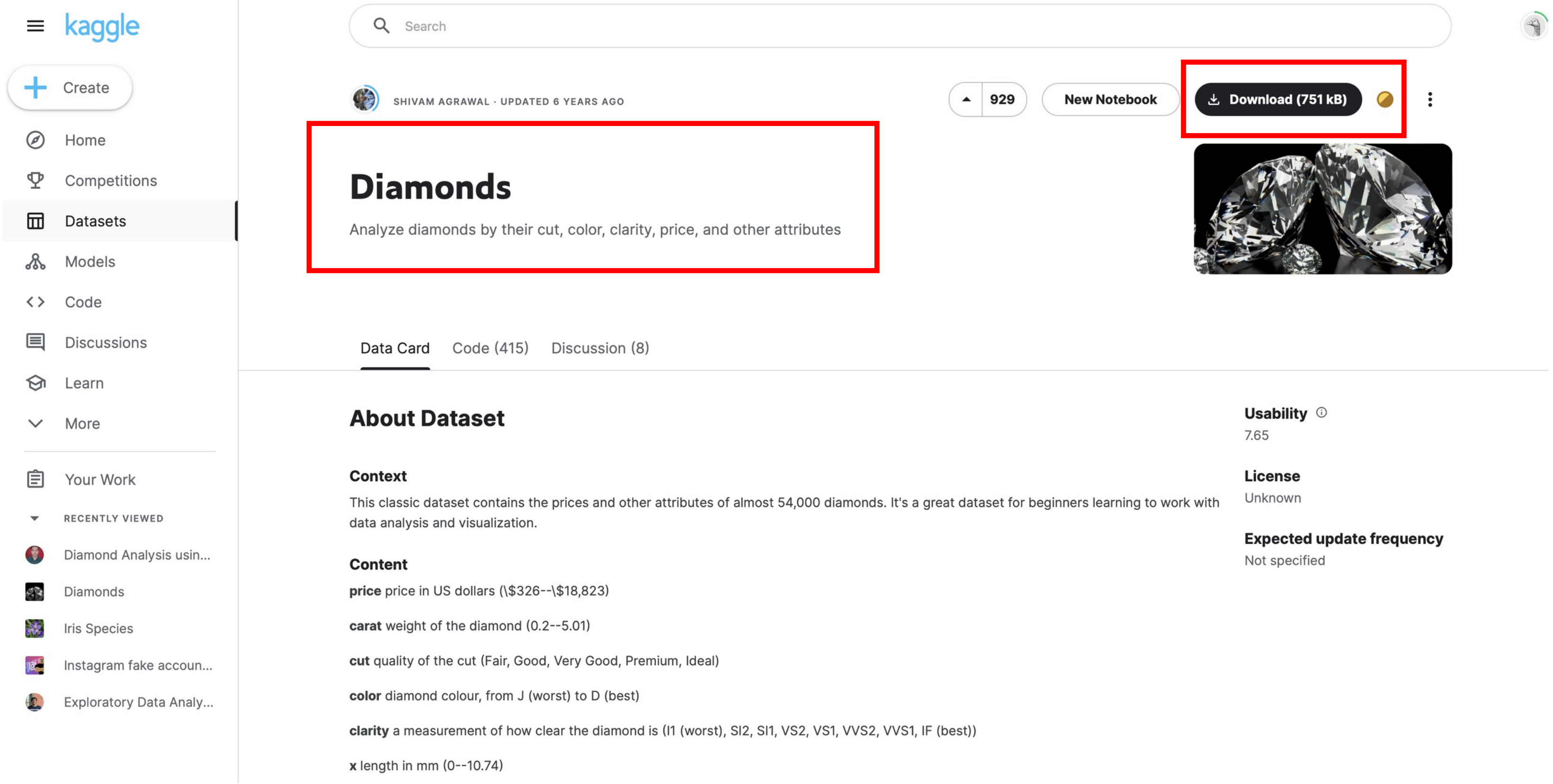
- The primary goal of EDA is to understand the underlying patterns, distributions, and relationships within the data.

- It involves techniques to identify outliers, detect patterns, test assumptions, and summarize the main features of the dataset.

Exploratory vs Confirmatory Data Analysis

EDA	CDA
• No hypothesis at first	• Start with hypothesis
• Generate hypothesis	• Test the null hypothesis
• Uses graphical methods (mostly)	• Uses statistical models

Getting the data



SHIVAM AGRAWAL · UPDATED 6 YEARS AGO

Diamonds

Analyze diamonds by their cut, color, clarity, price, and other attributes

About Dataset

Context

This classic dataset contains the prices and other attributes of almost 54,000 diamonds. It's a great dataset for beginners learning to work with data analysis and visualization.

Content

price price in US dollars (\$326--\$18,823)

carat weight of the diamond (0.2--5.01)

cut quality of the cut (Fair, Good, Very Good, Premium, Ideal)

color diamond colour, from J (worst) to D (best)

clarity a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))

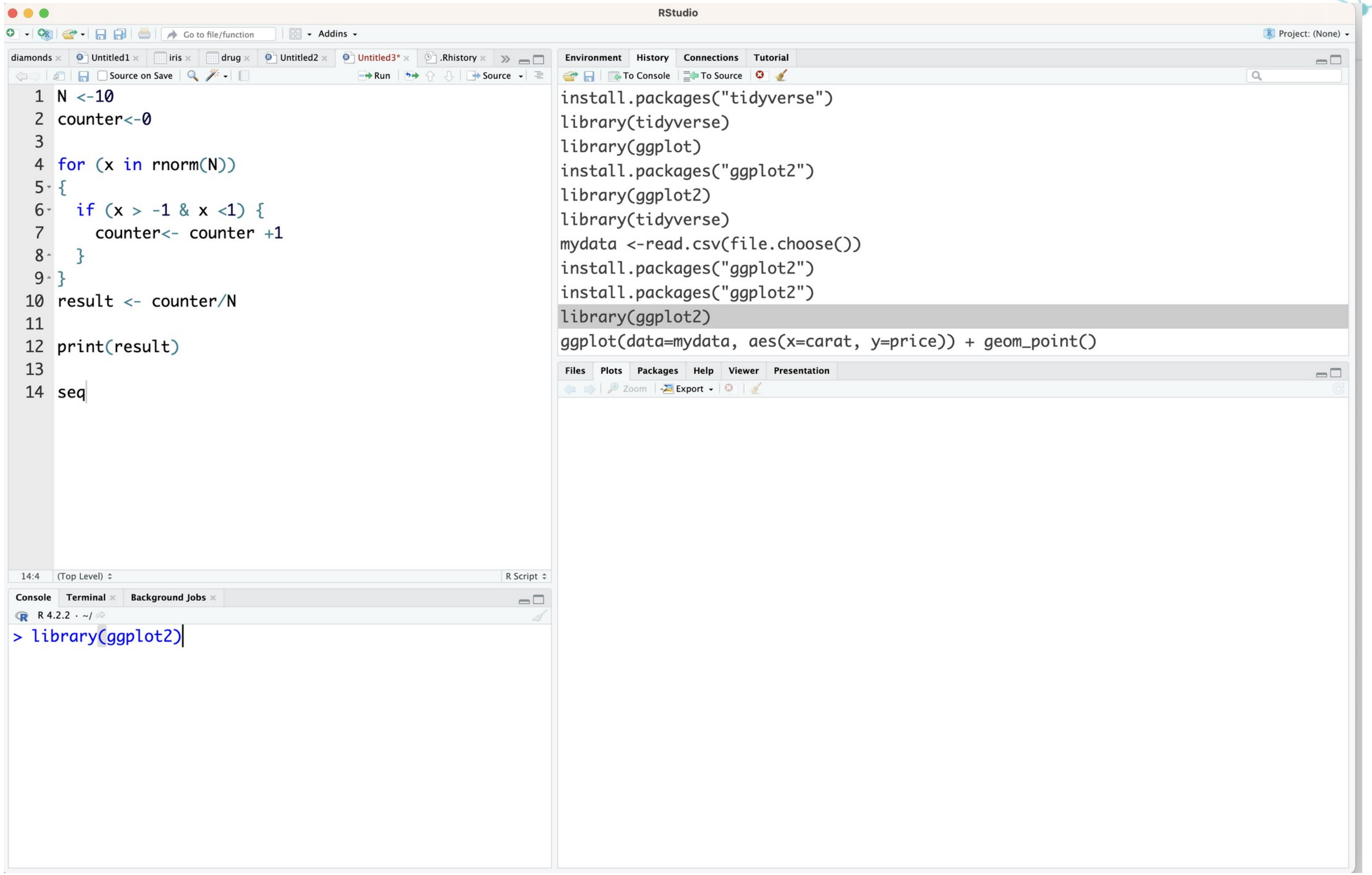
x length in mm (0--10.74)

Usability
7.65

License
Unknown

Expected update frequency
Not specified

Data visualization in R

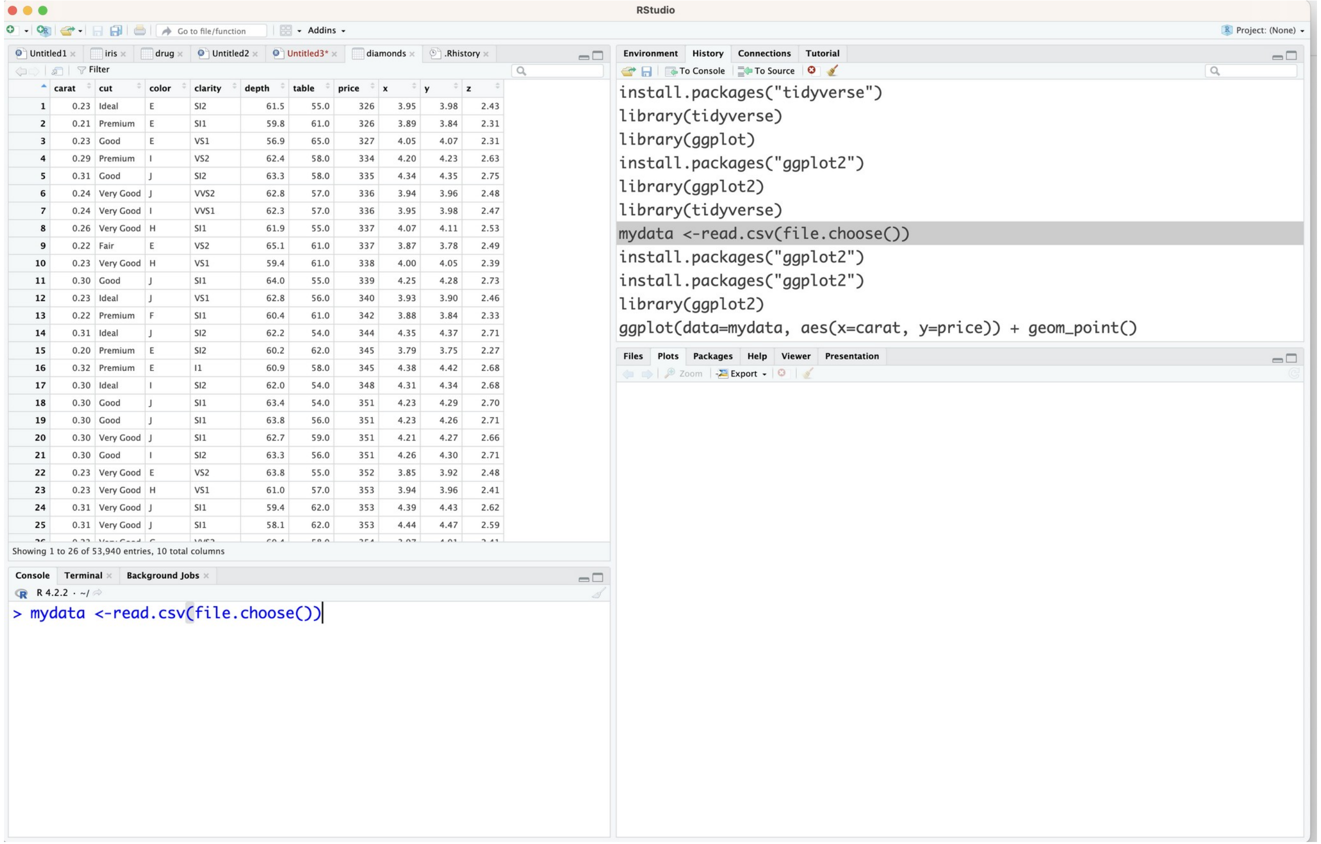


```r
N <-10
counter<-0

for (x in rnorm(N))
{
  if (x > -1 & x <1) {
    counter<- counter +1
  }
}
result <- counter/N

print(result)

seq
```

```r
install.packages("tidyverse")
library(tidyverse)
library(ggplot)
install.packages("ggplot2")
library(ggplot2)
library(tidyverse)
mydata <-read.csv(file.choose())
install.packages("ggplot2")
install.packages("ggplot2")
library(ggplot2)
ggplot(data=mydata, aes(x=carat, y=price)) + geom_point()
```

```
> library(ggplot2)
```

Data visualization in R



	carat	cut	color	clarity	depth	table	price	x	y	z
1	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
2	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
3	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
4	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
5	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75
6	0.24	Very Good	J	VVS2	62.8	57.0	336	3.94	3.96	2.48
7	0.24	Very Good	I	VVS1	62.3	57.0	336	3.95	3.98	2.47
8	0.26	Very Good	H	SI1	61.9	55.0	337	4.07	4.11	2.53
9	0.22	Fair	E	VS2	65.1	61.0	337	3.87	3.78	2.49
10	0.23	Very Good	H	VS1	59.4	61.0	338	4.00	4.05	2.39
11	0.30	Good	J	SI1	64.0	55.0	339	4.25	4.28	2.73
12	0.23	Ideal	J	VS1	62.8	56.0	340	3.93	3.90	2.46
13	0.22	Premium	F	SI1	60.4	61.0	342	3.88	3.84	2.33
14	0.31	Ideal	J	SI2	62.2	54.0	344	4.35	4.37	2.71
15	0.20	Premium	E	SI2	60.2	62.0	345	3.79	3.75	2.27
16	0.32	Premium	E	I1	60.9	58.0	345	4.38	4.42	2.68
17	0.30	Ideal	I	SI2	62.0	54.0	348	4.31	4.34	2.68
18	0.30	Good	J	SI1	63.4	54.0	351	4.23	4.29	2.70
19	0.30	Good	J	SI1	63.8	56.0	351	4.23	4.26	2.71
20	0.30	Very Good	J	SI1	62.7	59.0	351	4.21	4.27	2.66
21	0.30	Good	I	SI2	63.3	56.0	351	4.26	4.30	2.71
22	0.23	Very Good	E	VS2	63.8	55.0	352	3.85	3.92	2.48
23	0.23	Very Good	H	VS1	61.0	57.0	353	3.94	3.96	2.41
24	0.31	Very Good	J	SI1	59.4	62.0	353	4.39	4.43	2.62
25	0.31	Very Good	J	SI1	58.1	62.0	353	4.44	4.47	2.59

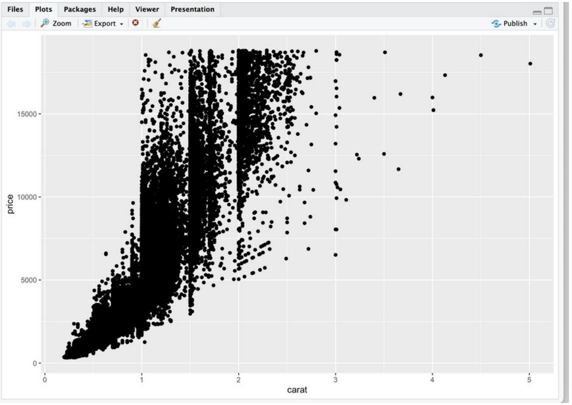
Showing 1 to 26 of 53,940 entries, 10 total columns

```
install.packages("tidyverse")
library(tidyverse)
library(ggplot)
install.packages("ggplot2")
library(ggplot2)
library(tidyverse)
mydata <-read.csv(file.choose())
install.packages("ggplot2")
install.packages("ggplot2")
library(ggplot2)
ggplot(data=mydata, aes(x=carat, y=price)) + geom_point()
```

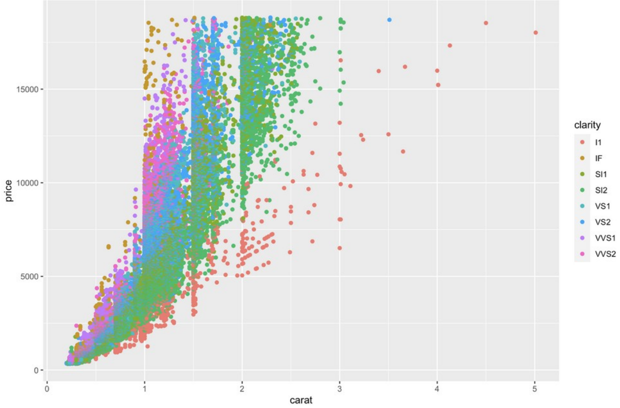
```
> mydata <-read.csv(file.choose())
```

Data visualization in R

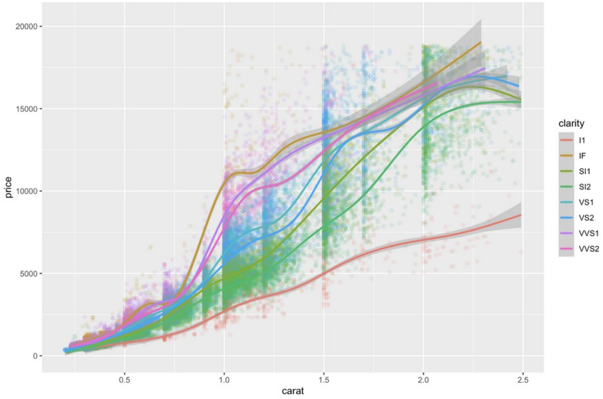
```
ggplot(data=mydata, aes(x=carat, y=price)) + geom_point()
```



```
ggplot(data=mydata, aes(x=carat,
y=price, color=clarity)) + geom_point()
```



```
ggplot(data=mydata[mydata$carat<2.5,], aes(x=carat, y=price,
color=clarity)) + geom_point(alpha=0.1) + geom_smooth()
```



Additional reading

https://carpentries-incubator.github.io/open-science-with-r/

https://marianattestad.com/blog

https://www.stats.ox.ac.uk/~evans/Rprog/LectureNotes.pdf

Thank you for your attention!