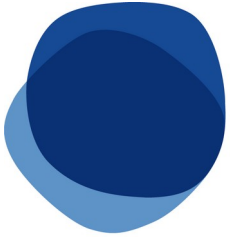
# *Data visualization and analysis in R*

Kristian Schultz

15.04.2025

www.sbi.uni-rostock.de



SYSTEMS BIOLOGY
BIOINFORMATICS
ROSTOCK

www.sbi.uni-rostock.de

Universität
Rostock
Traditio et Innovatio

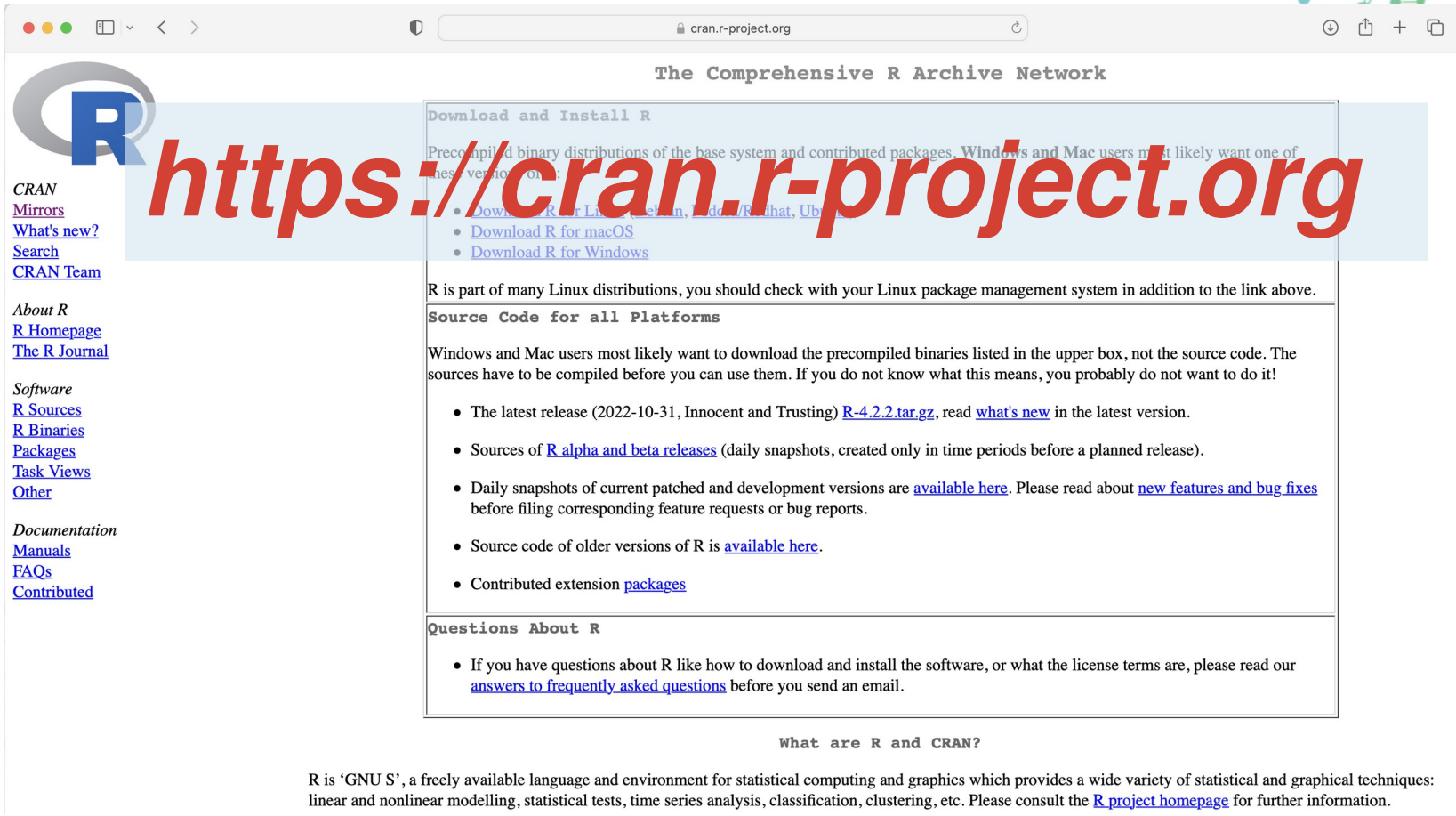
# *Data analysis with R*

- *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*

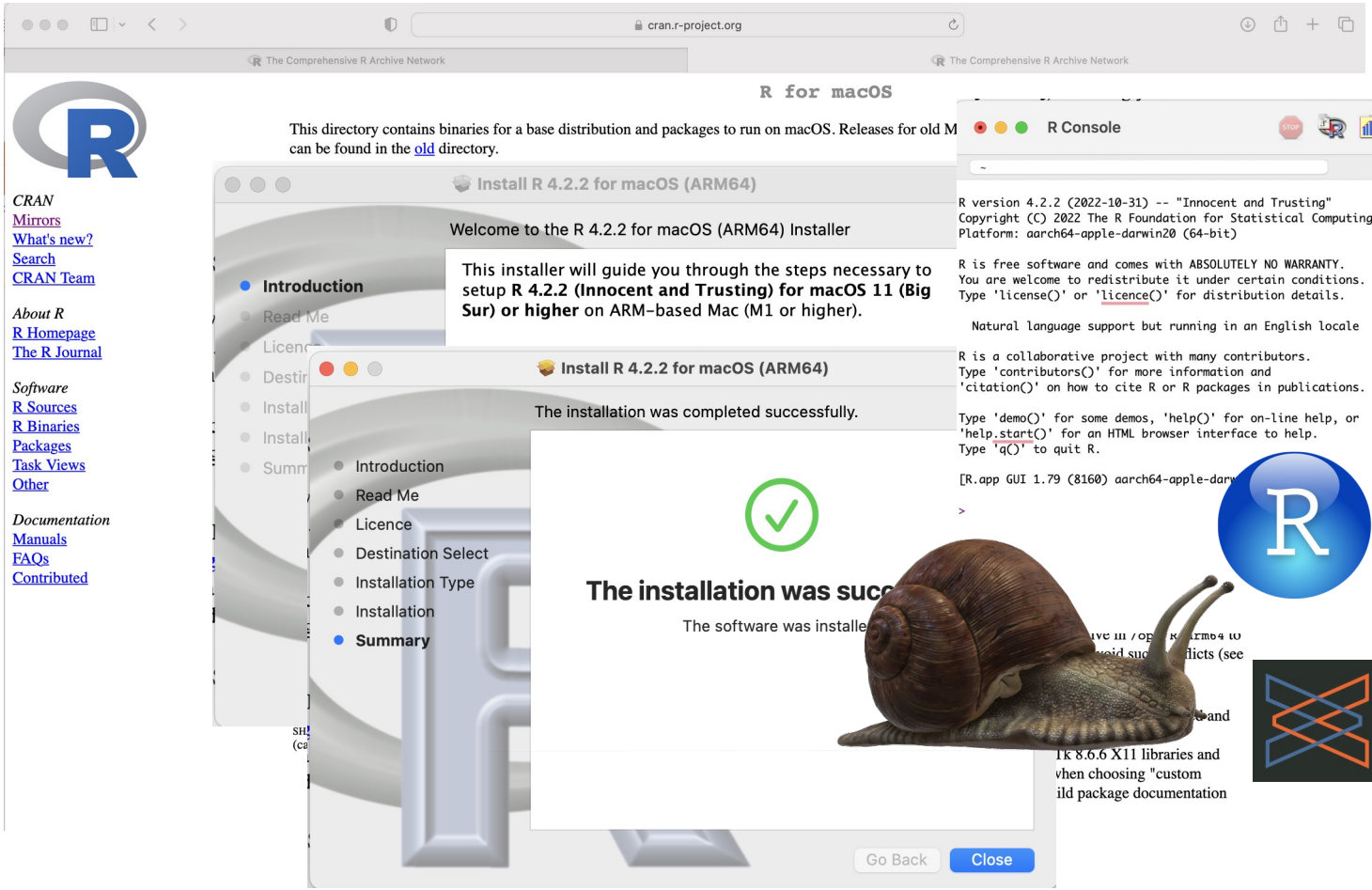
A software environment used to analyze **statistical information**, **graphical representation**, **reporting**, and **data modeling**

# *Installing*

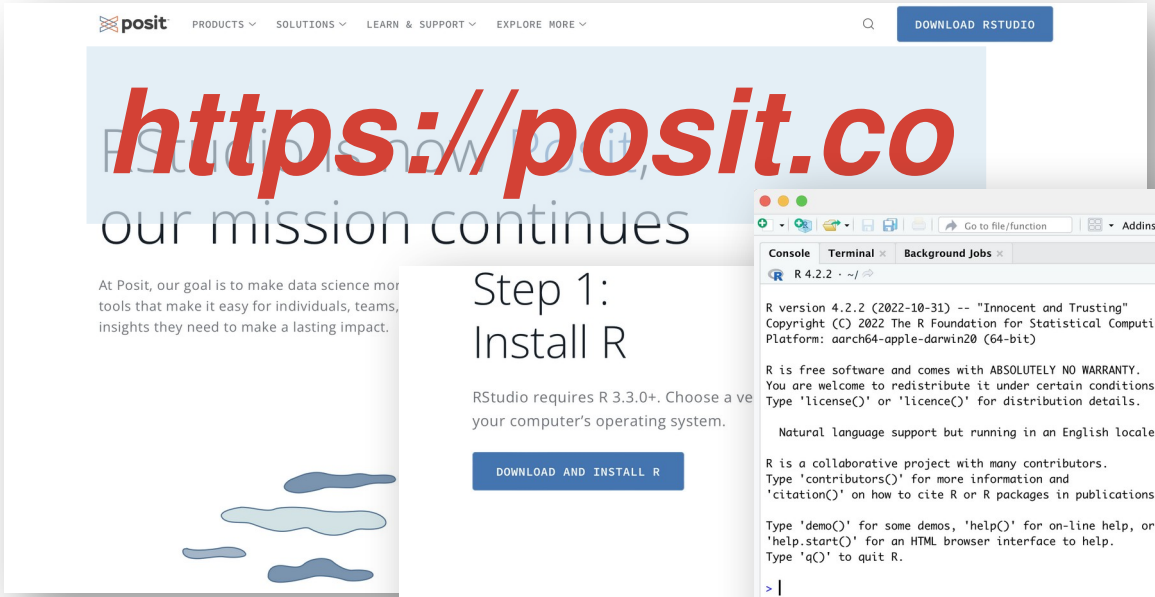
# Installing R

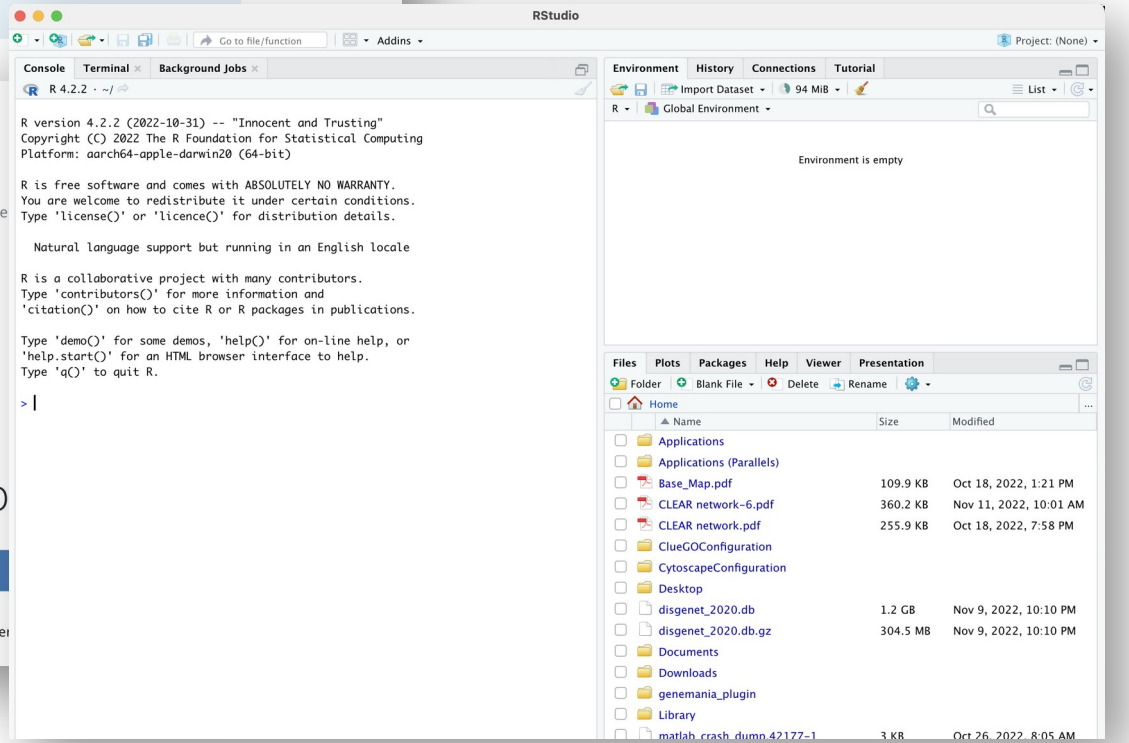


# Installing R

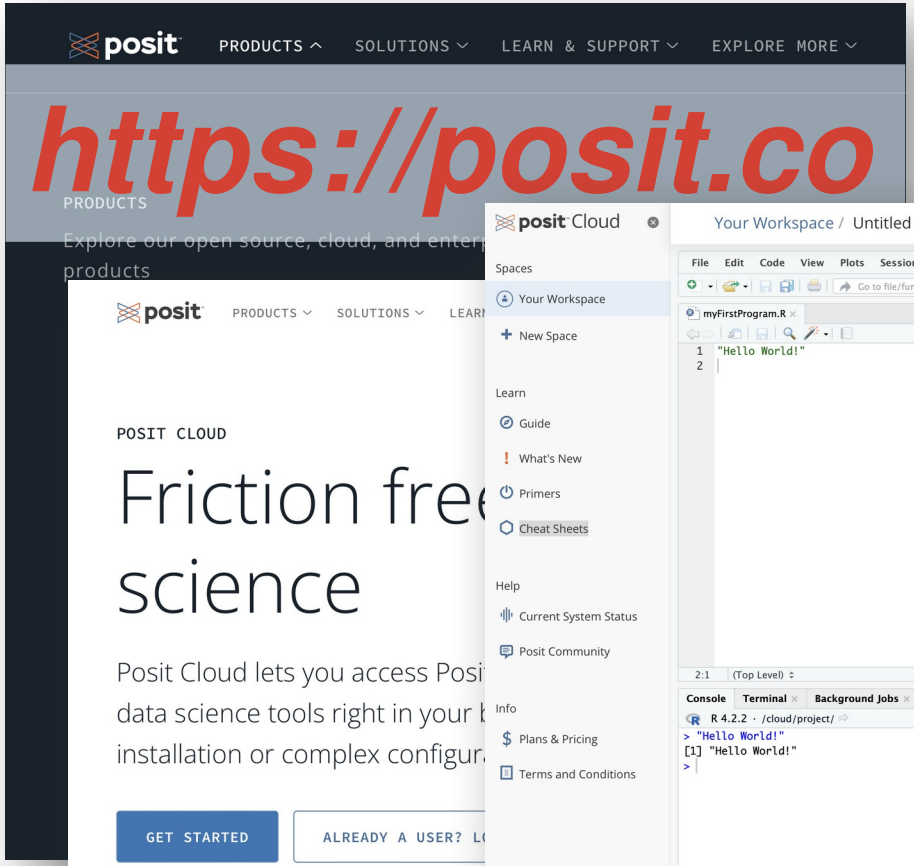


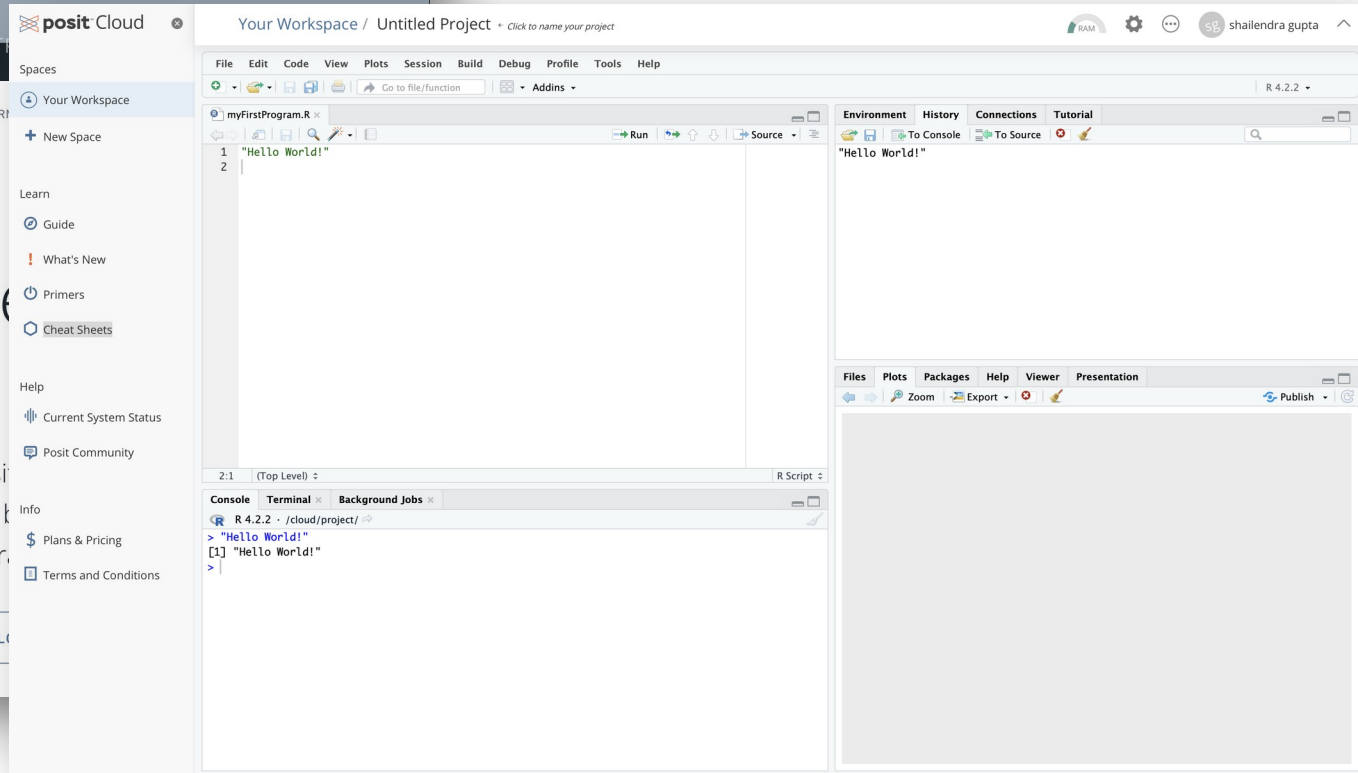
# Installing RStudio Desktop



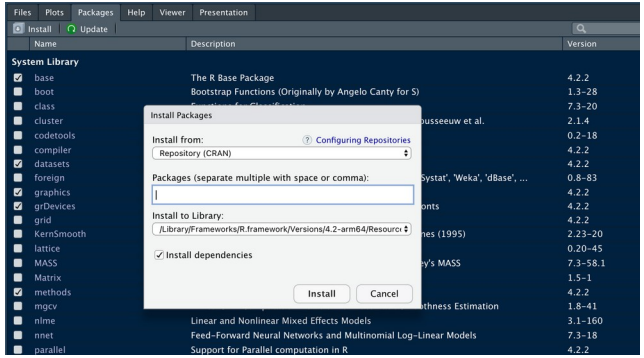
# Running R on Posit Cloud

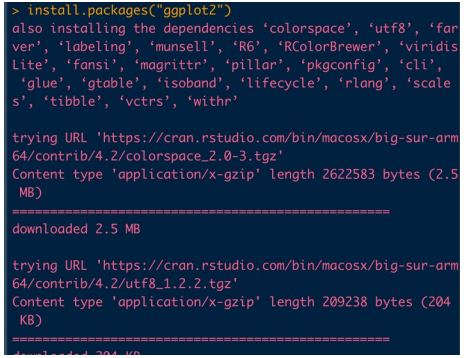


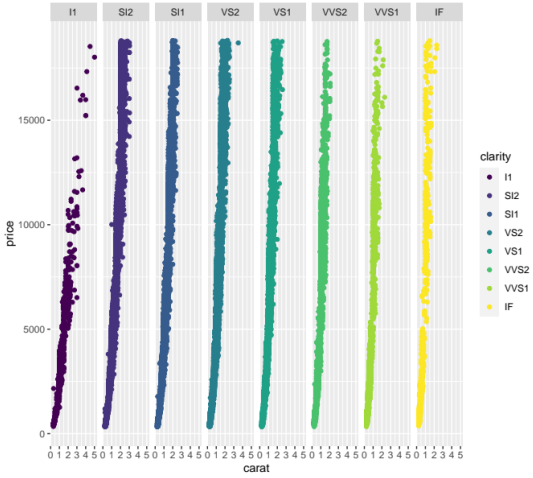
# Packages in R

- 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)

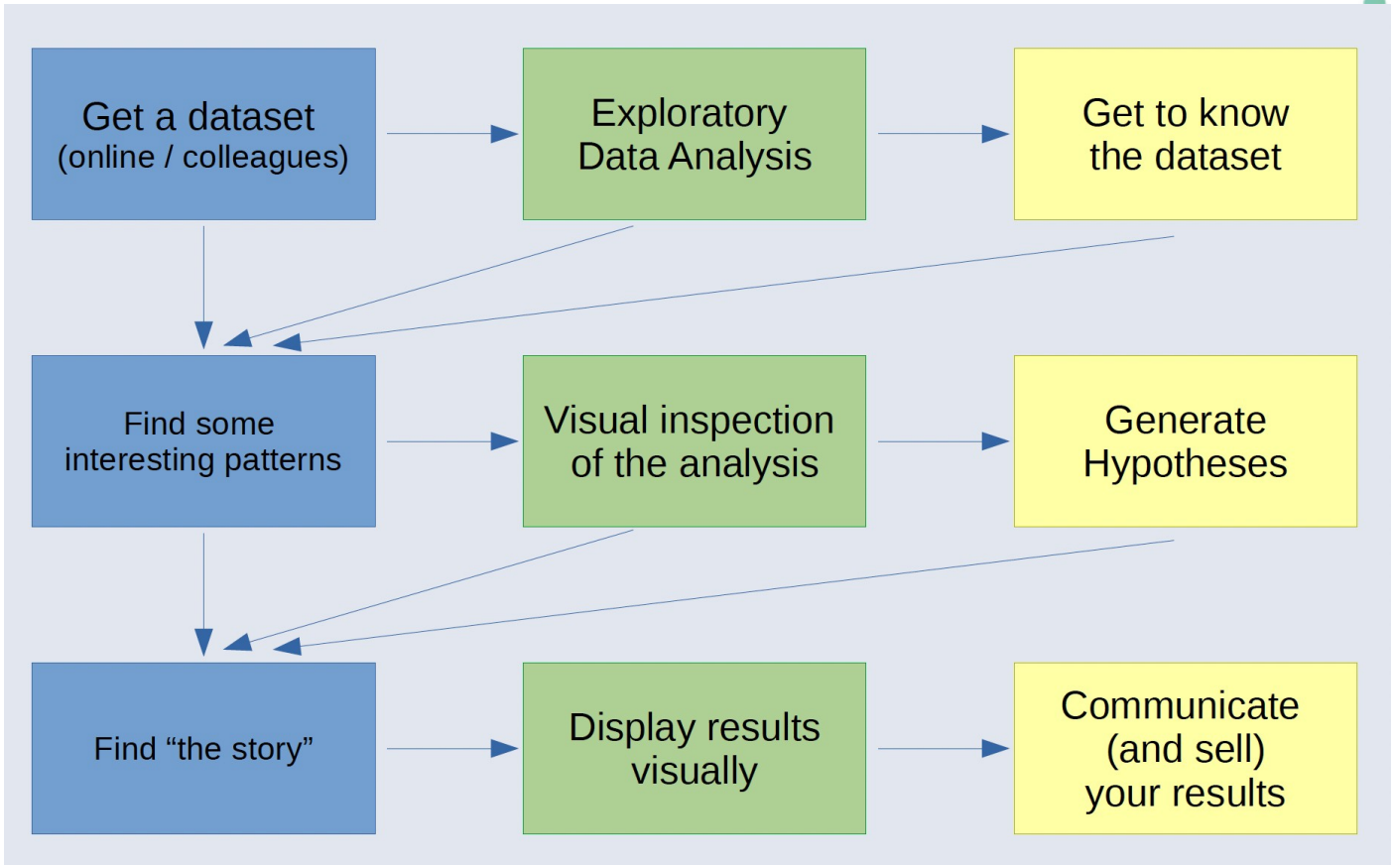



- Example from ggplot2 (ggplots comes with few example datasets, e.g., diamonds, iris)
  - qplot(data=diamonds, carat, price, colour=clarity, facets=.~clarity)

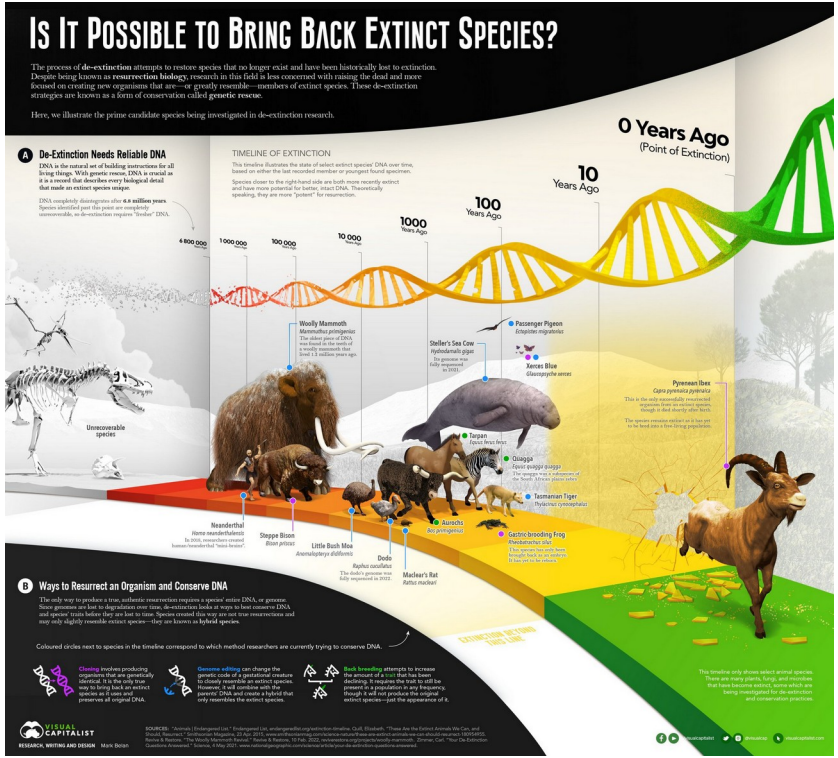


# *The daily life of a data scientist..*



# *Data visualization in R*

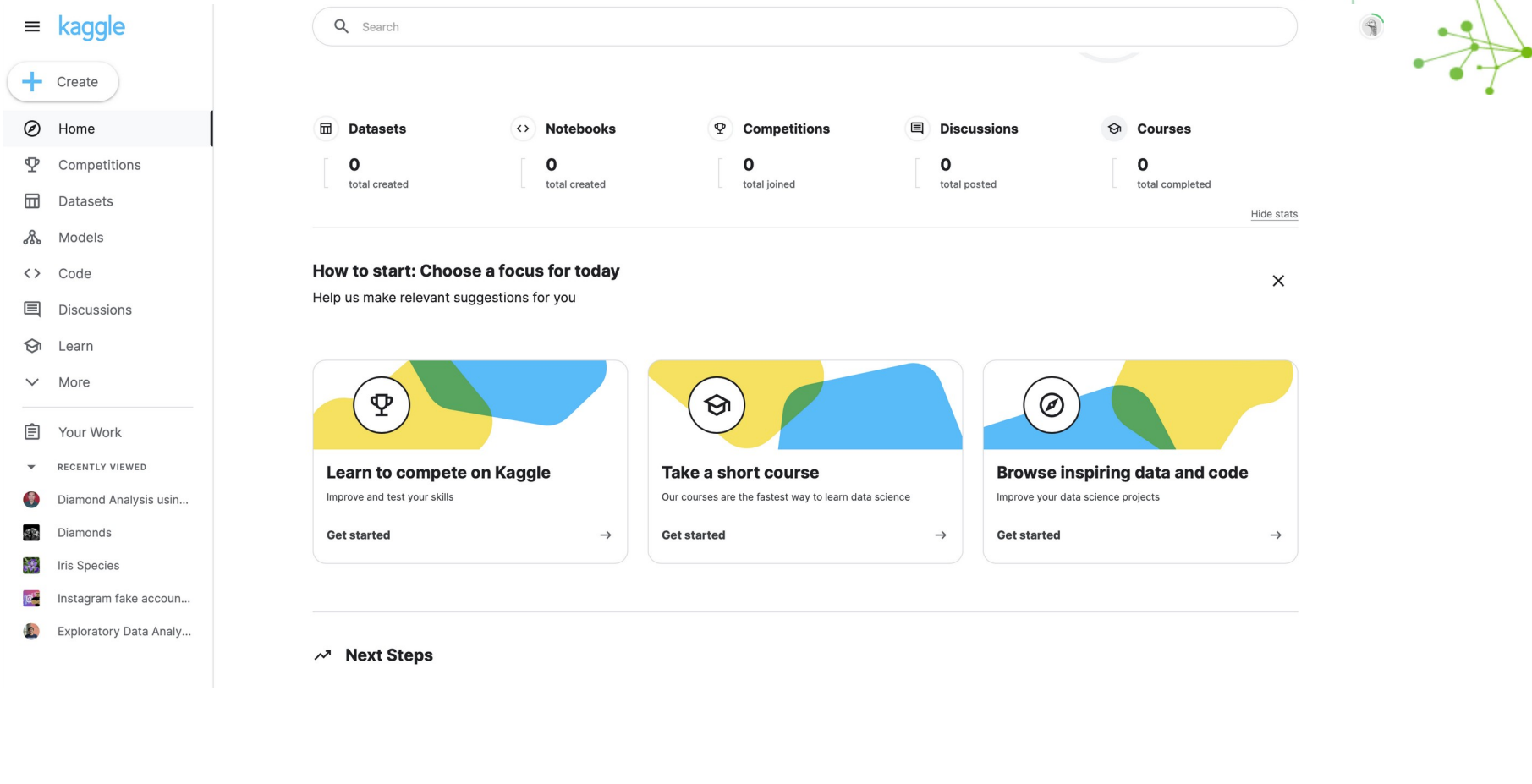
Not this




*Twitter*
*@markabelan*

But this

# Getting the data



# Getting the data

kaggle

Search

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...

SHIVAM AGRAWAL · UPDATED 6 YEARS AGO

929

New Notebook

Download (751 kB)

## Diamonds

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

Data Card | Code (415) | Discussion (8)

### 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*

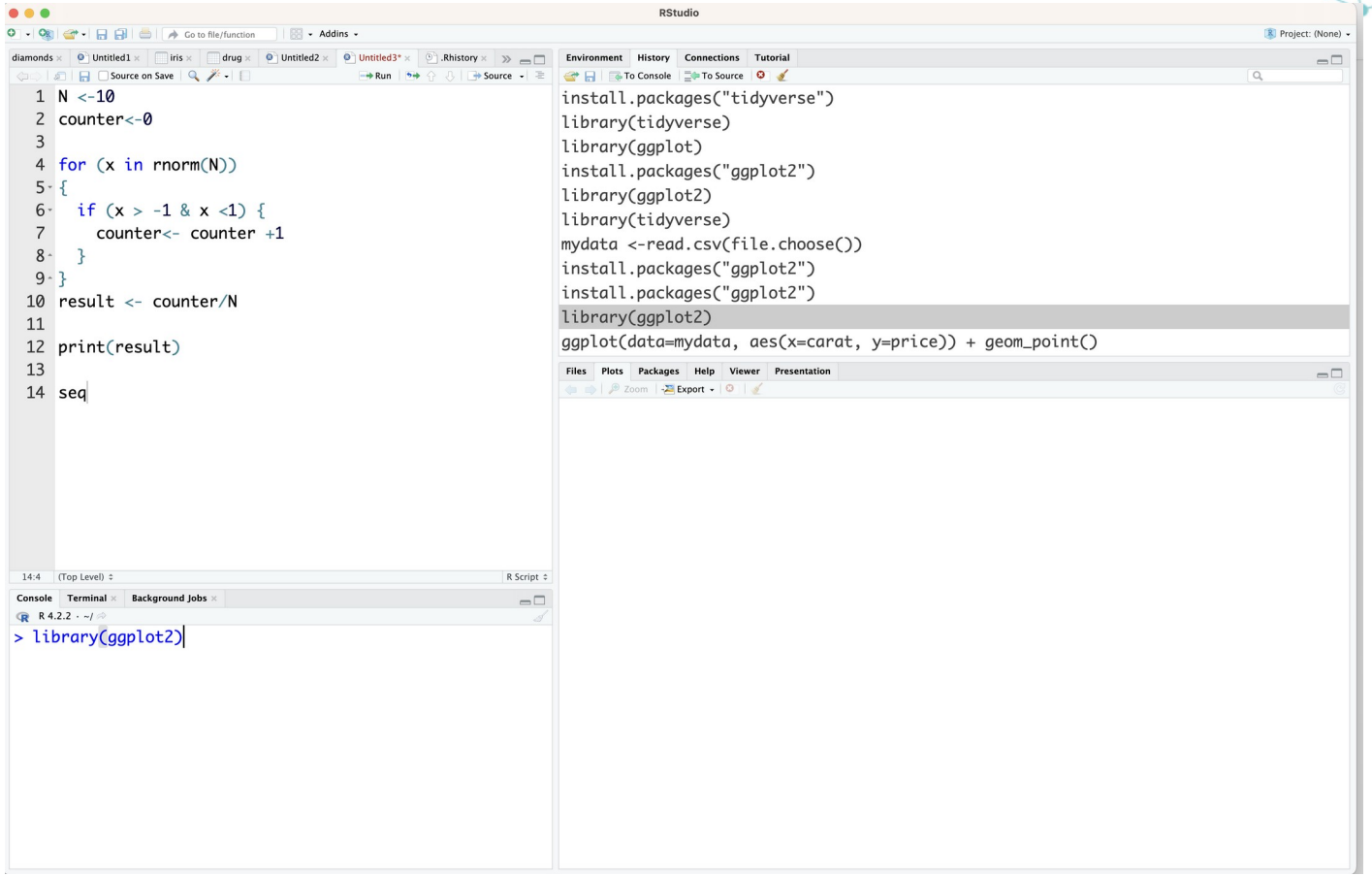
## *Exploratory Data Analysis*

- 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.

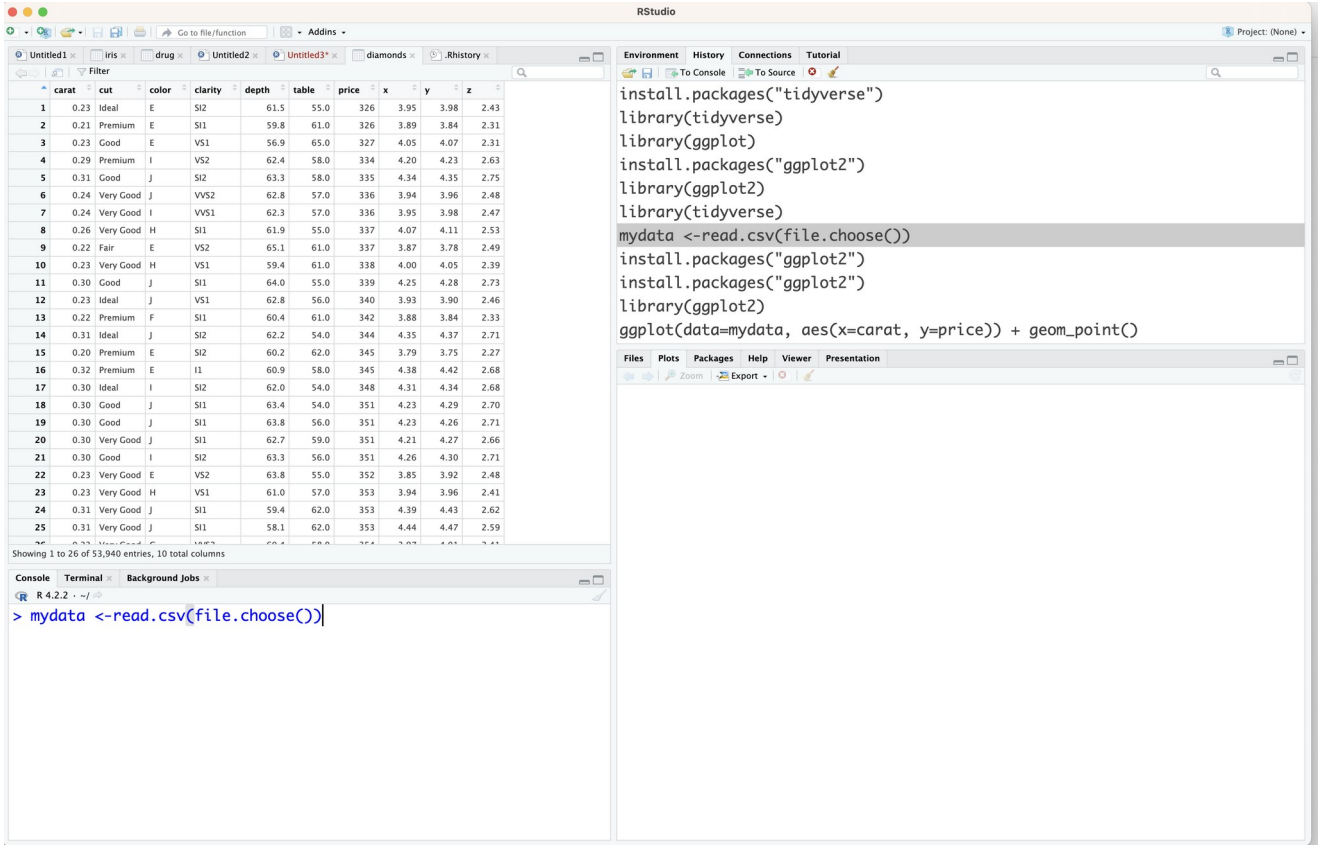
### Exploratory vs Confirmatory Data Analysis

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

# Data visualization in R

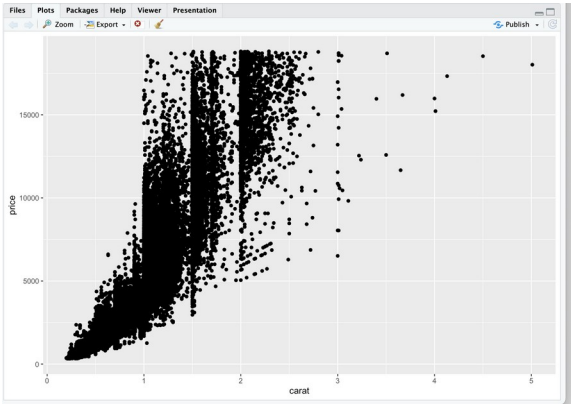


# Data visualization in R

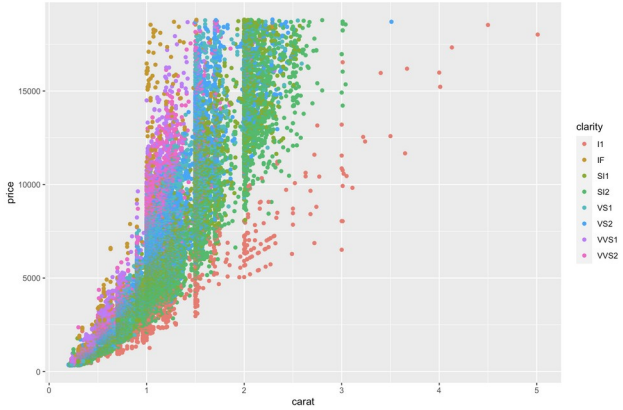


# *Data visualization in R*

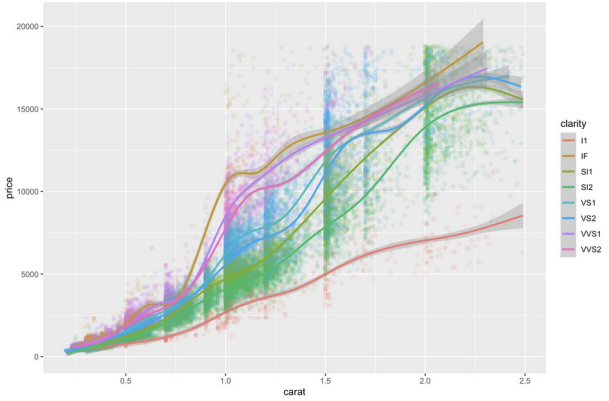
ggplot(data=mydata, aes(x=carat, y=price)) + geom_point()




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!*