

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

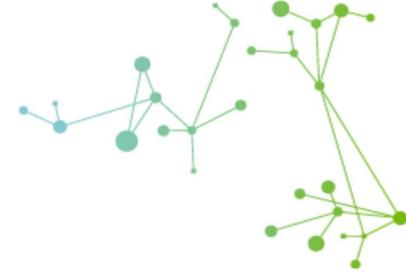


Data analysis with R



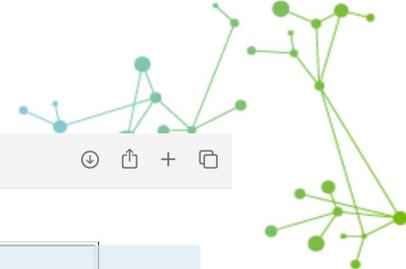
- *1992: Ross Ihaka and Robert Gentleman in the University of Auckland, New Zealand developed R.*
- *1995: The first version was released*
- *1997: CRAN (Comprehensive R Archive Network) was started*
- *2000: A stable beta version was released*
- *R has now thousands of packages, designed, maintained,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



The screenshot shows the CRAN website with the following content:

- 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](#)

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: [Linux](#), [Ubuntu](#), [Fedora](#), [CentOS](#), [Rocky Linux](#), [AlmaLinux](#), [Oracle Linux](#), [SUSE Linux Enterprise Server](#), [SUSE Linux Enterprise Desktop](#), [SUSE Linux Enterprise Server for SAP](#), [SUSE Linux Enterprise Server for Realtime](#), [SUSE Linux Enterprise Server for High Performance Computing](#), [SUSE Linux Enterprise Server for Cloud](#), [SUSE Linux Enterprise Server for Containers](#), [SUSE Linux Enterprise Server for Microservices](#), [SUSE Linux Enterprise Server for Edge](#), [SUSE Linux Enterprise Server for IoT](#), [SUSE Linux Enterprise Server for Industry](#), [SUSE Linux Enterprise Server for Manufacturing](#), [SUSE Linux Enterprise Server for Retail](#), [SUSE Linux Enterprise Server for Telecommunications](#), [SUSE Linux Enterprise Server for Utilities](#), [SUSE Linux Enterprise Server for Virtualization](#), [SUSE Linux Enterprise Server for Web](#), [SUSE Linux Enterprise Server for Cloud Managed Services](#), [SUSE Linux Enterprise Server for Cloud Managed Services for Containers](#), [SUSE Linux Enterprise Server for Cloud Managed Services for Edge](#), [SUSE Linux Enterprise Server for Cloud Managed Services for IoT](#), [SUSE Linux Enterprise Server for Cloud Managed Services for Industry](#), [SUSE Linux Enterprise Server for Cloud Managed Services for Manufacturing](#), [SUSE Linux Enterprise Server for Cloud Managed Services for Retail](#), [SUSE Linux Enterprise Server for Cloud Managed Services for Telecommunications](#), [SUSE Linux Enterprise Server for Cloud Managed Services for Utilities](#), [SUSE Linux Enterprise Server for Cloud Managed Services for Virtualization](#), [SUSE Linux Enterprise Server for Cloud Managed Services for Web](#).

- [Download R for Linux \(CentOS, Fedora, Ubuntu, etc.\)](#)
- [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](#)

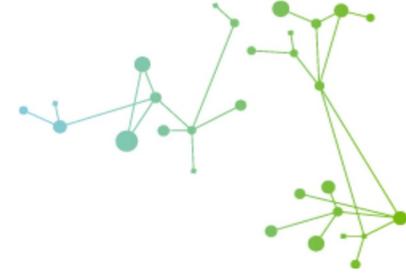
Questions About R

- 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 RStudio Desktop



posit PRODUCTS SOLUTIONS LEARN & SUPPORT EXPLORE MORE



DOWNLOAD RSTUDIO

https://posit.co

our mission continues

At Posit, our goal is to make data science more 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 version for your computer's operating system.

DOWNLOAD AND INSTALL R

Step 2: Install RStudio

DOWNLOAD RSTUDIO DESKTOP FOR MAC

Size: 224.49MB | SHA-256: 35028002 | Version: 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.

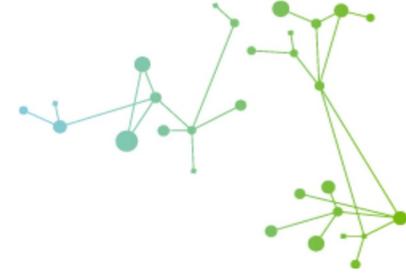
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |

Environment: Global Environment (94 MiB)

Files: Applications, Applications (Parallels), Base_Map.pdf, CLEAR network-6.pdf, CLEAR network.pdf, ClueGOConfiguration, CytoscapeConfiguration, Desktop, disgenet_2020.db, disgenet_2020.db.gz, Documents, Downloads, genemania_plugin, Library, matlab_crash_dump_42177-1

Running R on Posit Cloud



posit PRODUCTS ^ SOLUTIONS v LEARN & SUPPORT v EXPLORE MORE v

https://posit.co

PRODUCTS

Explore our open source, cloud, and enterprise products

posit PRODUCTS v SOLUTIONS v LEARN & SUPPORT v EXPLORE MORE v

POSIT CLOUD

Friction free science

Posit Cloud lets you access Posit data science tools right in your browser without installation or complex configuration.

[GET STARTED](#) [ALREADY A USER? LOGIN](#)

posit Cloud Your Workspace / Untitled Project - Click to name your project

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

myFirstProgram.R

```
1 "Hello World!"
2
```

Environment History Connections Tutorial

To Console To Source

Files Plots Packages Help Viewer Presentation

Zoom Export Publish

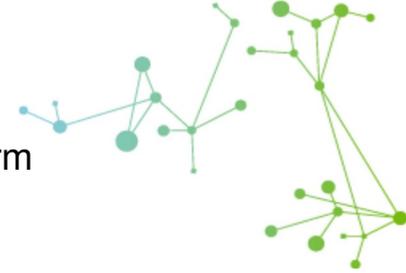
2:1 (Top Level) R Script

Console Terminal Background Jobs

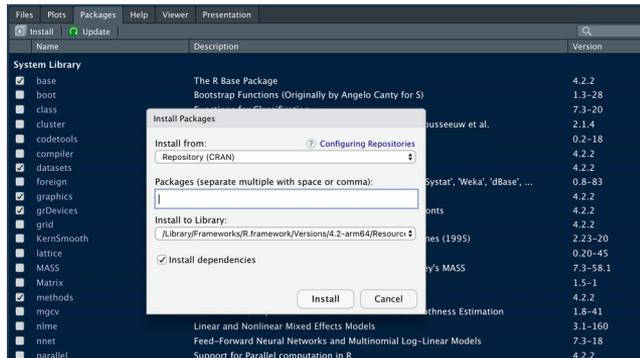
R 4.2.2 - /cloud/project

```
> "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)`

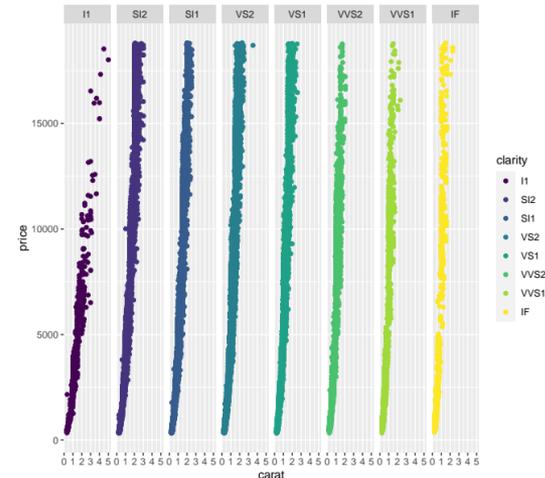


```
> install.packages("ggplot2")
also installing the dependencies 'colorspace', 'utf8', 'fan
ver', 'labeling', 'munsell', 'R6', 'RColorBrewer', 'viridis
Lite', 'fansi', 'magrittr', 'pillar', 'pkgconfig', 'cli',
'glue', 'gtable', 'isoband', 'lifecycle', 'rlang', 'scale
s', 'tibble', 'vctrs', 'withr'

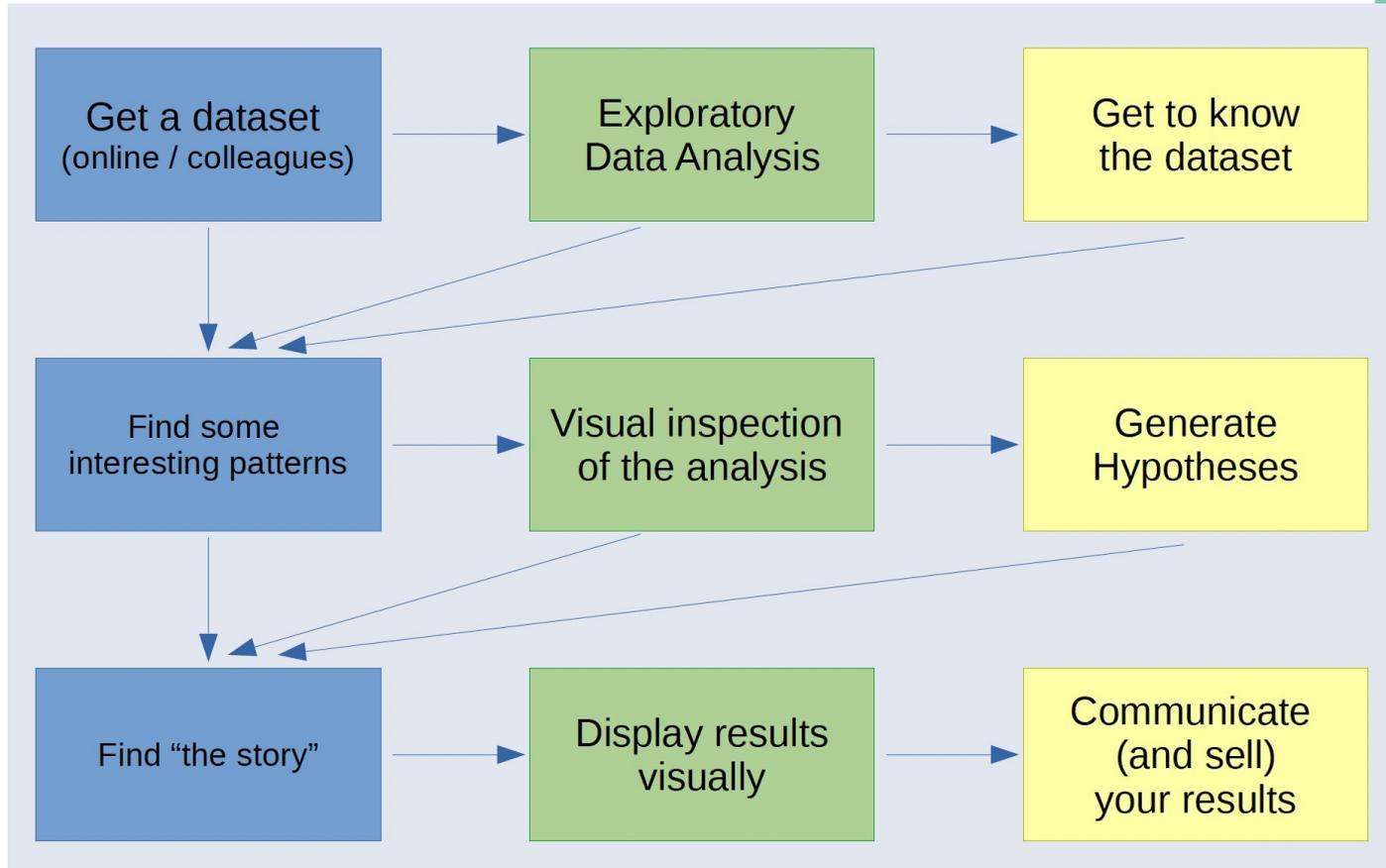
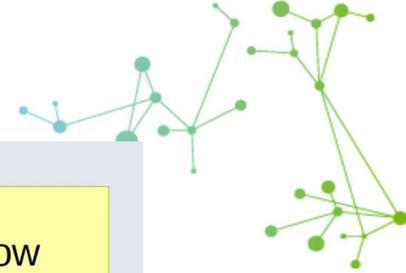
trying URL 'https://cran.rstudio.com/bin/macosx/big-sur-arm
64/contrib/4.2/colorspace_2.0-3.tgz'
Content type 'application/x-gzip' length 2622583 bytes (2.5
MB)
=====
downloaded 2.5 MB

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)
=====
downloaded 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..



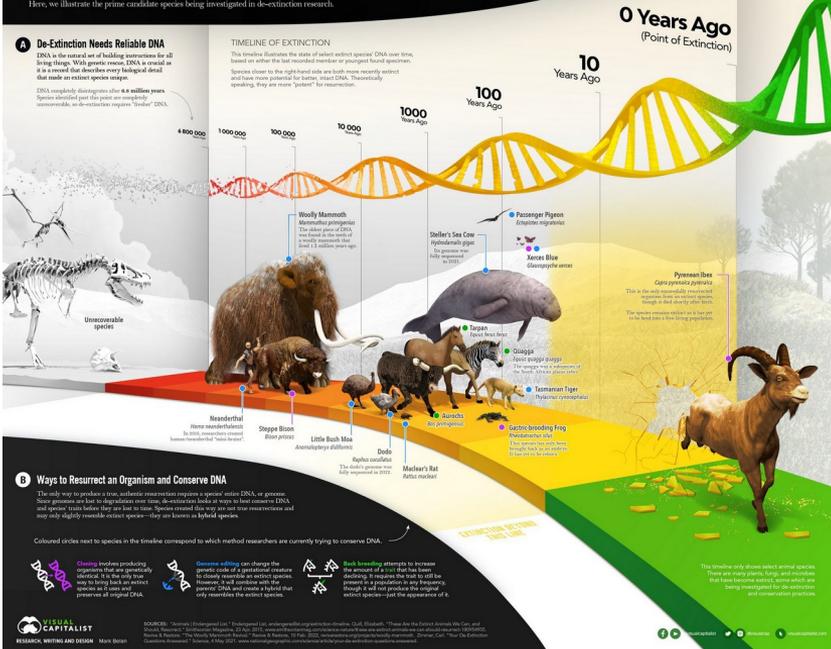
Data visualization in R

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—on a genetic level—more or less similar to 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.



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

5 10 25

Select the colour of the histogram

blue

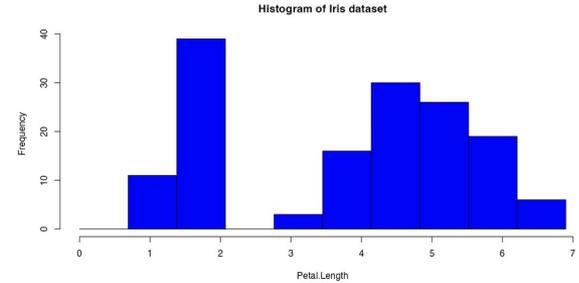
yellow

red

purple

Histogram Data Summary

The variable names you choose here is Petal.Length



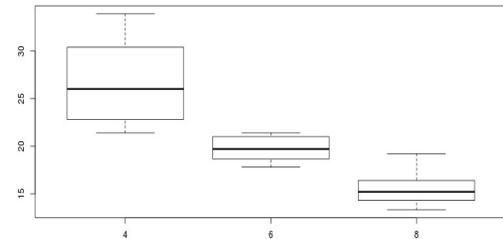
Miles Per Gallon

Variable:

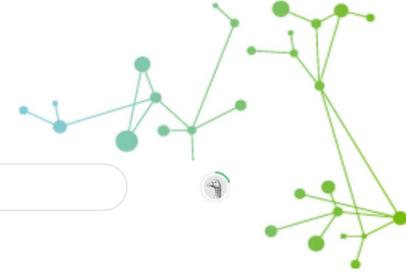
Cylinders

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 account...

👤 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

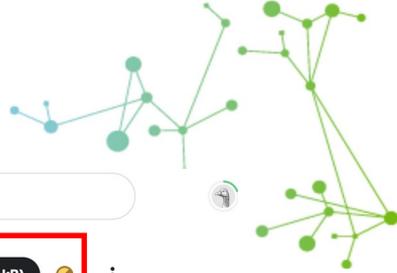
Improve your data science projects

Get started



↗ Next Steps

Getting the data



☰ kaggle

+ Create

🏠 Home

🏆 Competitions

📄 Datasets

🤖 Models

🔍 Code

🗨️ Discussions

🎓 Learn

∨ More

📁 Your Work

∨ RECENTLY VIEWED

📄 Diamond Analysis usin...

📄 Diamonds

📄 Iris Species

📄 Instagram fake account...

📄 Exploratory Data Analy...

🔍 Search



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

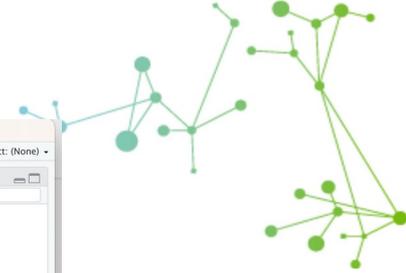


- Exploratory Data Analysis (EDA) is an approach to analyzing datasets to summarize their main characteristics, often employing visual methods.
- The primary goal of EDA is to understand the underlying patterns, distributions, and relationships within the data.
- It involves techniques to identify outliers, detect patterns, test assumptions, and summarize the main features of the dataset.

Exploratory vs Confirmatory Data Analysis

EDA	CDA
<ul style="list-style-type: none">• No hypothesis at first• Generate hypothesis• Uses graphical methods (mostly)	<ul style="list-style-type: none">• Start with hypothesis• Test the null hypothesis• Uses statistical models

Data visualization in R



The screenshot shows the RStudio environment. The main editor window contains the following R code:

```
1 N <-10
2 counter<-0
3
4 for (x in rnorm(N))
5 {
6   if (x > -1 & x <1) {
7     counter<- counter +1
8   }
9 }
10 result <- counter/N
11
12 print(result)
13
14 seq
```

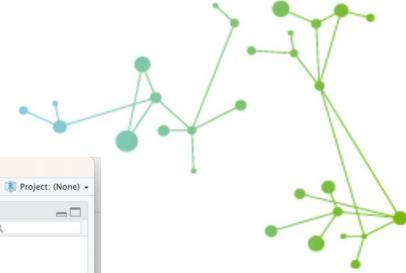
The console window at the bottom left shows the command:

```
> library(ggplot2)
```

The script editor on the right contains the following code:

```
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()
```

Data visualization in R

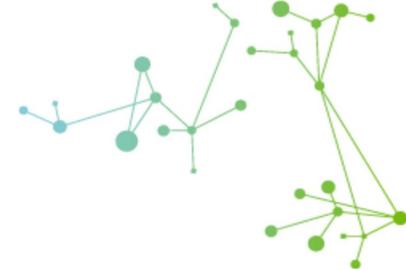
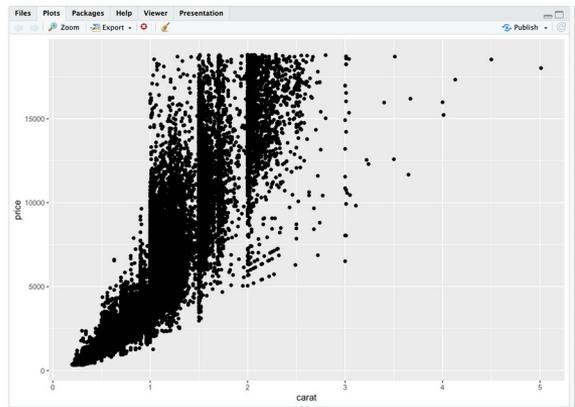


The screenshot displays the RStudio environment. On the left, a data table is visible with columns: carat, cut, color, clarity, depth, table, price, x, y, z. The table contains 25 rows of diamond data. Below the table, a console window shows the command `> mydata <- read.csv(file.choose())`. On the right, the R script editor contains the following code:

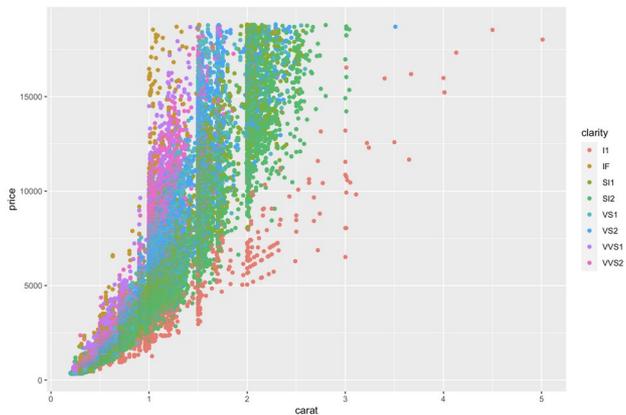
```
install.packages("tidyverse")
library(tidyverse)
library(ggplot)
install.packages("ggplot2")
library(ggplot2)
library(tidyverse)
mydata <- read.csv(file.choose())
install.packages("ggplot2")
install.packages("ggplot2")
library(ggplot2)
ggplot(data=mydata, aes(x=carat, y=price)) + geom_point()
```

Data visualization in R

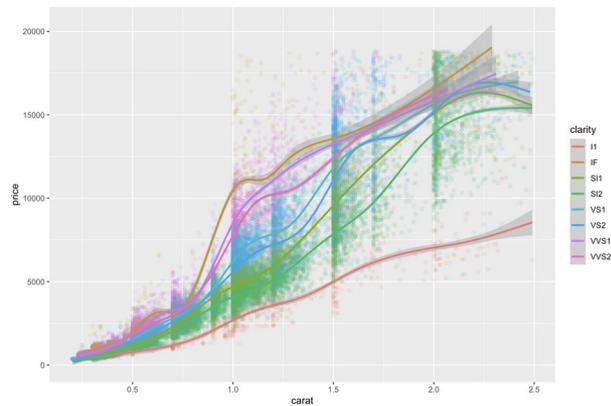
```
ggplot(data=mydata, aes(x=carat, y=price)) + geom_point()
```



```
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!