DATA MINING

19.08.2024 - 06.12. 2024

Rabu, 17.30 – 19.15 WIB

Pertemuan pertama
28-08-2024

1. Kontrak perkuliahan
2. Pembentukan Kelompok
3. Pendahuluan : apa, tugas, proses *data mining ?*
4. Pengenalan *software* yang dipakai : R
5. NEXT : *exploring data with R*

Deskripsi mata kuliah

- Kuliah ini dimaksudkan untuk memberikan kemampuan memberi arti data yang terdiri dari explorasi data dan visualisasi data, klasifikasi (*classification*) dengan menggunakan *random forest* dan teknik alternatif klasifikasi, *Association Analysis and Prediction using Multiple Non Linear Regression and Generalized Linear Models, Cluster Analysis : Basic Concepts and Algorithms, Additional Issues and Algorithms, Anomaly Detection*.

- Pendekatan perkuliahan merupakan kombinasi kajian analitik dan komputasi menggunakan *software R dan Matlab*

ADS+ DBN+BBS + FAL

Setelah mengikuti kuliah ini diharapkan mahasiswa mampu

1. *Identify data types and distinguish data mining application.*
2. *Describe data mining algorithms.*
3. *Describe applicability of data mining.*
4. *Suggest appropriate solution to data mining problems.*
5. *Analyze data mining algorithms and techniques.*
6. *Work as a team in solving challenging data mining problems.*

E. Tahap pembelajaran:

Pert-ke	Kemampuan Akhir	Materi Pembelajaran	Metode Pembelajaran	Waktu (menit)	Indikator	Penilaian
0	▪ Memiliki wawasan yang luas tentang Paket Program R	**Pengantar Paket Program R (FALDY)**	• Presentasi • Diskusi	150'	• Memberikan informasi tentang perintah-perintah yang digunakan dalam paket program R .	.
1	▪ Memiliki wawasan yang luas tentang Data Mining dan Aplikasinya dalam Industri dan Bisnis ▪ Mampu mengidentifikasi berbagai jenis data dan kemungkinan menambang informasi dari data	**Introduction (BBS)** **Data & Exploring Data (BBS)**	• Presentasi • Diskusi • Presentasi • Demostrasi • Diskusi • Praktek (dengan paket program R/R Studio atau R/R Studio under clouds)	100'	• Memberikan informasi tentang posisi bidang ilmu data mining dan manfaatnya . • Memberikan informasi tentang berbagai jenis data dan bagaimana melakukan eksplorasi data untuk memperoleh informasi yang penting. • Diskusi dan Praktikum	**Tugas :** Membuat laporan Praktikum.
2	▪ Mampu melakukan visualisasi berbagai jenis data	**Visualisasi Data (BBS)**	• Presentasi • Demostrasi • Diskusi • Praktek (dengan paket program R/R Studio atau R/R Studio under clouds) •	100'	• Demonstrasi tentang bagaimana melakukan visualisasi berbagai data. • Memperoleh informasi dari hasil visualisasi data • Diskusi dan Praktikum •	**Tugas :** Membuat laporan Praktikum.

PESERTA KELAS

:: SEMESTER 1 TA 2024 - 2025

:: MR456 - DATA MINING

:: RABU 19:00-21:00

No	NIM	Nama Mahasiswa	B/U/P
1.	662021001	NI KADEK ANGELITA MEILANI	B
2.	662021003	MUHAMMAD ANAS RIFAI	B
3.	662021005	DUTA ERLANGGA SAPUTRA	B
4.	662021009	JOY ANGELIO STEFANO RATAR	B
5.	662021010	ANGGA PRASETIA MULIA	B
6.	662021012	FADHIIL DHIAURRAHMAN ATHARIQ	B
7.	662021013	ANANDA AKHMAD AULIA ULIL ABSHOR	B
8.	662022002	MALIKA JASMINE NAZHIIF	B
9.	662022003	EVANGELIA NELVINA ISKANDAR	B
10.	662022009	MICHAEL BIMA KURNIANDI	B
11.	662022011	MUTIARA TYAS PUTRI ANDARIYANTO	B
12.	662022016	MADHURI NARA SARIRA NABABAN	B

PESERTA KELAS

:: SEMESTER 1 TA 2024 - 2025

:: BD001 - DATA MINING

:: RABU 16:00-19:00

No	NIM	Nama Mahasiswa	B/U/P
1.	632024001	HARFELY LEIPARY	B
2.	632024003	YULIUS ARDITA DWI NUGROHO	B
3.	632024005	DINA NOVIANI SAMALUKANG	B

Presensi Cetak * Untuk mencetak, silahkan matikan fitur Popup Blocker di browser Anda.

Madhu

Muti

Malika

Bima

Eva

Fai

Ulil

Angga

Joy

Fadhiil

Duta

Angel

Office hours...

- Selasa – Kamis (09-12)
- Perjanjian dulu via WA (085 865 227 961)
- Jadwal *make up class* :

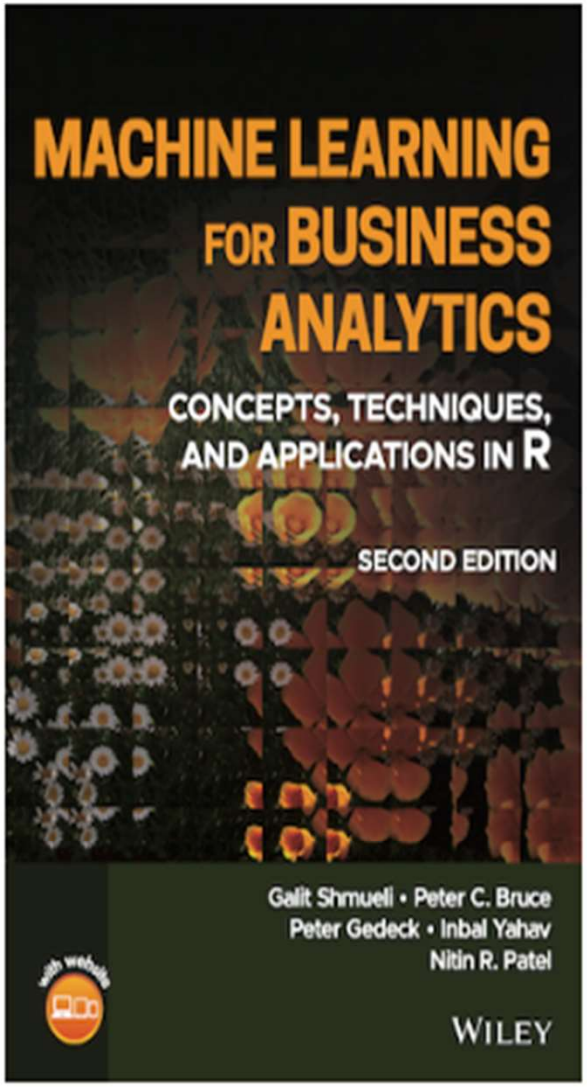
  Jumat, mulai 17.30 -> 27 SEPTEMBER 2024

Distribusi Penilaian

- 10 % Presensi
- 30 % Laporan Praktikum
- 60% TAS (“project” + presentasi)

Referensi utama

gedeck/mlba-R-code: Repository contains R and Rmd files for the 2nd edition of Machine Learning for Business Analytics (github.com)

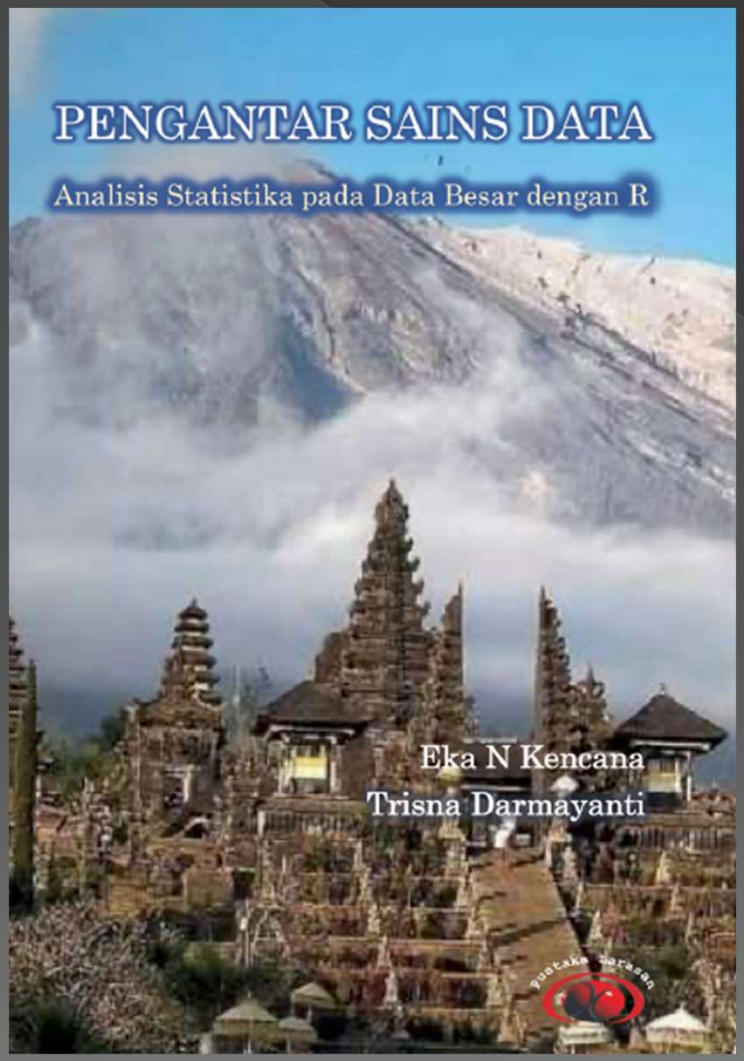


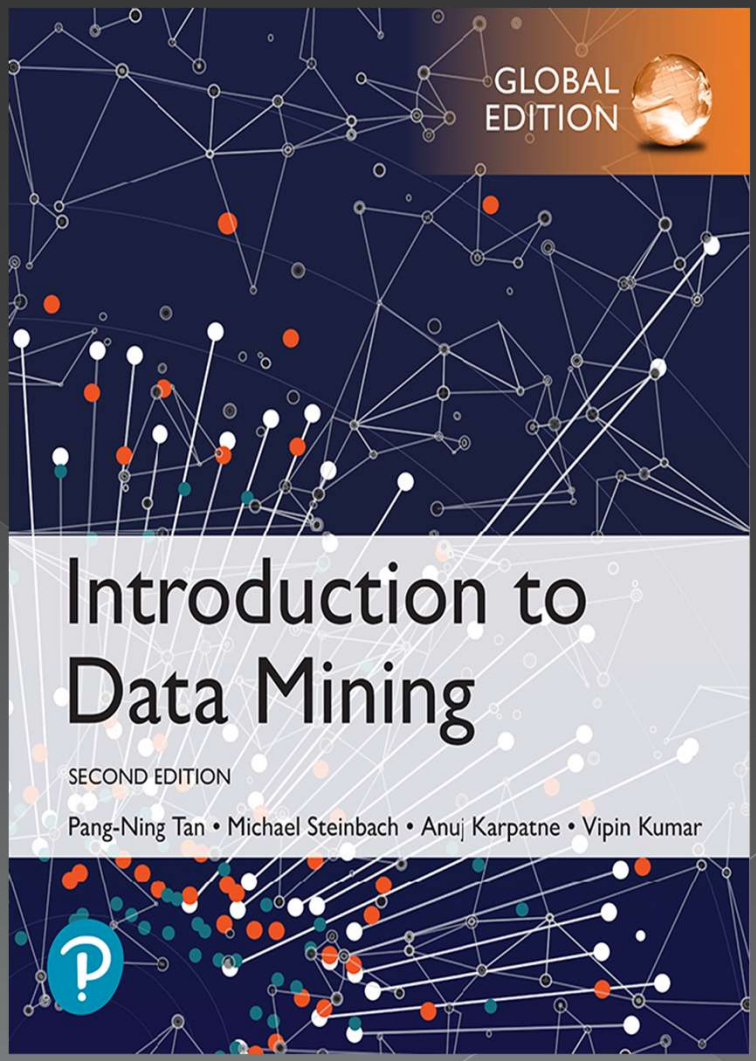
Machine Learning for Business Analytics Concepts, Techniques, and Applications in R

by Galit Shmueli, Peter C. Bruce, Peter Gedeck, Inbal Yahav, Nitin R. Patel

Publisher: Wiley; 2nd edition (February, 2023) ISBN: 978-1-118-83517-2 Buy at Amazon or Wiley

P. Tan, M. Steinback and V. Kumar, *Introduction to Data Mining*, Addison Wesley, Second Edition, 2016





Kontrak Perkuliahan

0% complete

Kuliah ini dimaksudkan untuk memberikan kemampuan memberi arti data yang terdiri dari explorasi data dan visualisasi data, klasifikasi (classification) dengan menggunakan random forest dan teknik alternatif klasifikasi, Association Analysis and Prediction using Multiple Non Linear Regression and Generalized Linear Models, Cluster Analysis : Basic Concepts and Algorithms, Additional Issues and Algorithms, Anomaly Detection.

RPS

Mark as done

Terlampir Rencana Pembelajaran Semester

Ebook-referensi

Mark as done

PPT-PERTEMUAN-PERTAMA

Mark as done

Sumber : https://github.com/mhahsler/Introduction_to_Data_Mining_R_Examples

ACUAN PENGAYAAN

- https://rafalab.dfci.harvard.edu/dsbook/
- Michael Hahsler. *An R Companion for Introduction to Data Mining*. 2021 https://mhahsler.github.io/Introduction_to_Data_Mining_R_Examples/book/

Data

- https://www.yourdictionary.com/data
- https://www.simplilearn.com/what-is-data-article
- https://www.ibm.com/cloud/learn/data-warehouse

What is Data?

- Collection of data objects and their attributes
- An attribute is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, or feature
- A collection of attributes describe an object
  - Object is also known as record, point, case, sample, entity, or instance

Attributes

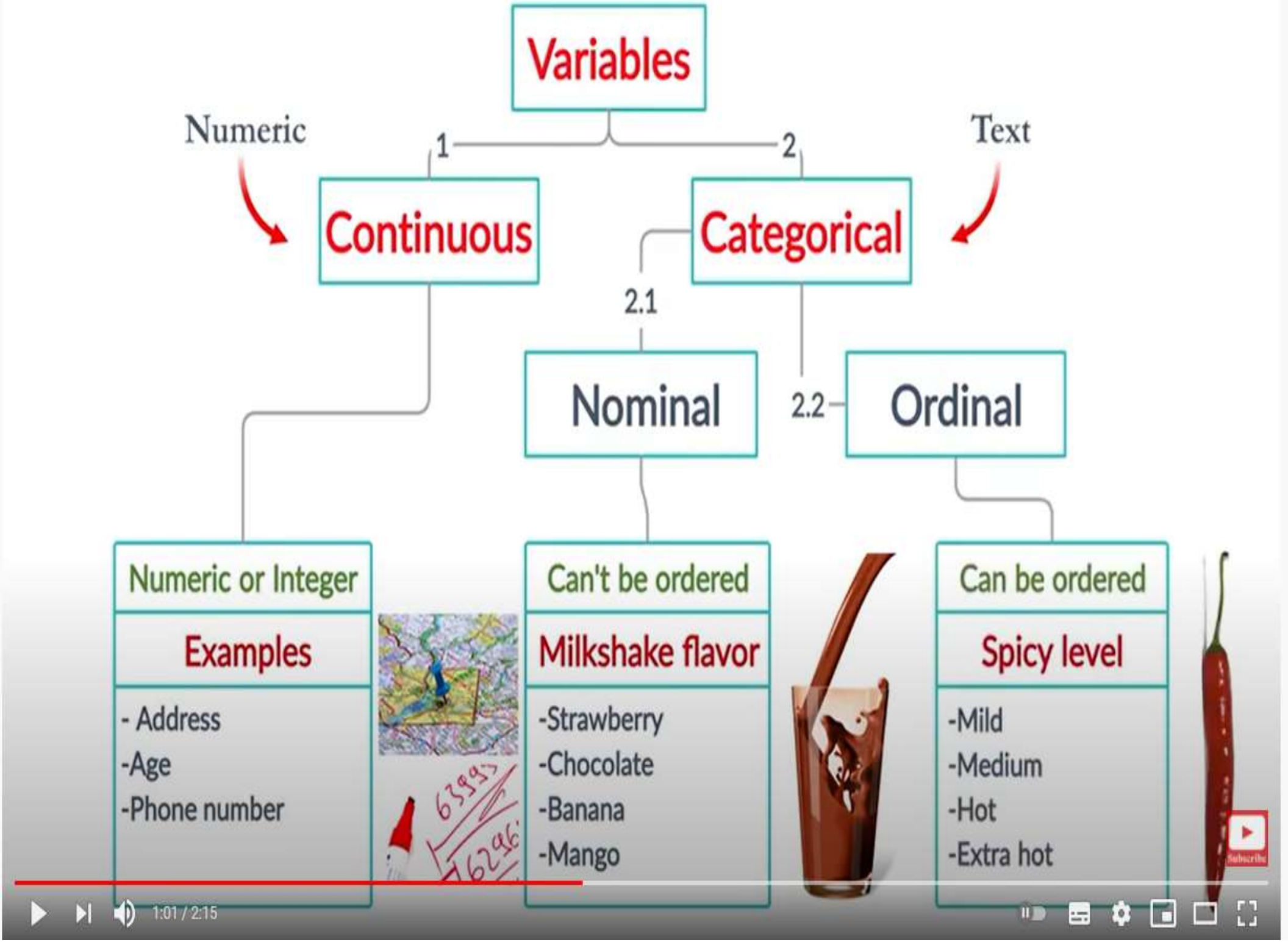
Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Attribute Values

- Attribute values are numbers or symbols assigned to an attribute

- Distinction between attributes and attribute values
  - Same attribute can be mapped to different attribute values
    - Example: height can be measured in feet or meters
  - Different attributes can be mapped to the same set of values
    - Example: Attribute values for ID and age are integers
    - But properties of attribute values can be different
      - ID has no limit but age has a maximum and minimum value


Variables
Numeric
1
2
Text
Continuous
Categorical
2.1
Nominal
2.2
Ordinal
Numeric or Integer
Examples
- Address
-Age
-Phone number
Can't be ordered
Milkshake flavor
-Strawberry
-Chocolate
-Banana
-Mango
Can be ordered
Spicy level
-Mild
-Medium
-Hot
-Extra hot
1:01 / 2:15

Types of Data

Structured: Data that has predefined structures. e.g. tables, spreadsheets, or relational databases.

Unstructured Data: Data with no predefined structure, comes in any size or form, cannot be easily stored in tables. e.g. blobs of text, images, audio

Quantitative Data: Numerical. e.g. height, weight

Categorical Data: Data that can be labeled or divided into groups. e.g. race, sex, hair color.

Big Data: Massive datasets, or data that contains greater *variety* arriving in increasing *volumes* and with ever-higher *velocity* (3 Vs). Cannot fit in the memory of a single machine.

Data Sources/Fomats

Most Common Data Formats CSV, XML, SQL, JSON, Protocol Buffers

Data Sources Companies/Proprietary Data, APIs, Government, Academic, Web Scraping/Crawling

Tahun 2009

oleh lab Fei-Fei Li

1,2 juta pasangan gambar dan label

untuk klasifikasi 1000 jenis gambar (image classification)

Internet cepat

buat apa?

Buat download dataset

1:08 / 3:05

3 alasan kenapa AI baru booming sekarang

Example of Data Volumes

Unit	Value	Example
Kilobytes (KB)	1,000 bytes	a paragraph of a text document
Megabytes (MB)	1,000 Kilobytes	a small novel
Gigabytes (GB)	1,000 Megabytes	Beethoven's 5th Symphony
Terabytes (TB)	1,000 Gigabytes	all the X-rays in a large hospital
Petabytes (PB)	1,000 Terabytes	half the contents of all US academic research libraries
Exabytes (EB)	1,000 Petabytes	about one fifth of the words people have ever spoken
Zettabytes (ZB)	1,000 Exabytes	as much information as there are grains of sand on all the world's beaches
Yottabytes (YB)	1,000 Zettabytes	as much information as there are atoms in 7,000 human bodies

https://www.eecis.udel.edu/~amer/Table-Kilo-Mega-Giga---YottaBytes.html

What is Data Mining?

- Combination of AI and statistical analysis to discover information that is "hidden" in the data
  - associations (e.g. linking purchase of pizza with beer)
  - sequences (e.g. tying events together: marriage and purchase of furniture)
  - classifications (e.g. recognizing patterns such as the attributes of customers that are most likely to quit)
  - forecasting (e.g. predicting buying habits of customers based on past patterns)

https://www.youtube.com/shorts/omxv0VjnxUs

AI bisa belajar? | Pengenalan Machine Learning

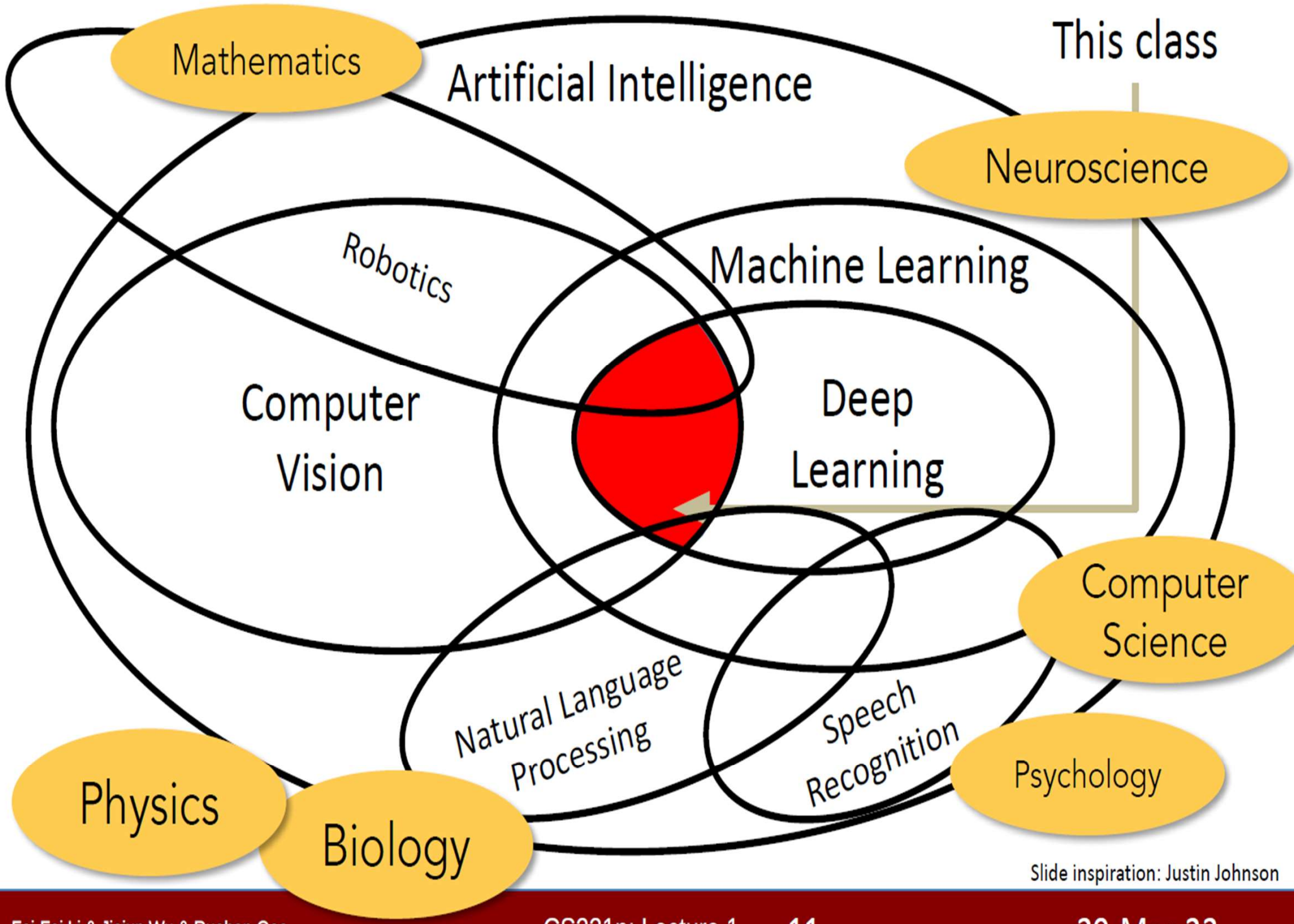
Anak AI
5,9 rb subscriber

Disubscribe

76

Bagikan

Download



Slide inspiration: Justin Johnson

https://www.youtube.com/watch?v=h82NuHDNhKI&t=711s

Apa itu Data Mining?

- Disiplin ilmu yang mempelajari metode untuk mengekstrak pengetahuan atau menemukan pola dari suatu data yang besar

- Ekstraksi dari data ke pengetahuan:
  1. Data: fakta yang terekam dan tidak membawa arti
  2. Informasi: Rekap, rangkuman, penjelasan dan statistik dari data
  3. Pengetahuan: pola, rumus, aturan atau model yang muncul dari data

- Nama lain data mining:
  - Knowledge Discovery in Database (KDD)
  - Big data
  - Business intelligence
  - Knowledge extraction
  - Pattern analysis
  - Information harvesting

RSW INSTITUTE

P. Tan, M. Steinback and V. Kumar, *Introduction to Data Mining.p25*

1.1 What Is Data Mining?

Data mining is the process of automatically discovering useful information in large data repositories. Data mining techniques are deployed to scour large data sets in order to find novel and useful patterns that might otherwise remain unknown. They also provide the capability to predict the outcome of a future observation, such as the amount a customer will spend at an online or a brick-and-mortar store.

Not all information discovery tasks are considered to be data mining. Examples include queries, e.g., looking up individual records in a database or finding web pages that contain a particular set of keywords. This is because such tasks can be accomplished through simple interactions with a database management system or an information retrieval system. These systems rely on traditional computer science techniques, which include sophisticated indexing structures and query processing algorithms, for efficiently organizing and retrieving information from large data repositories. Nonetheless, data mining techniques have been used to enhance the performance of such systems by improving the quality of the search results based on their relevance to the input queries.

P. Tan, M. Steinback and V. Kumar, *Introduction to Data Mining*

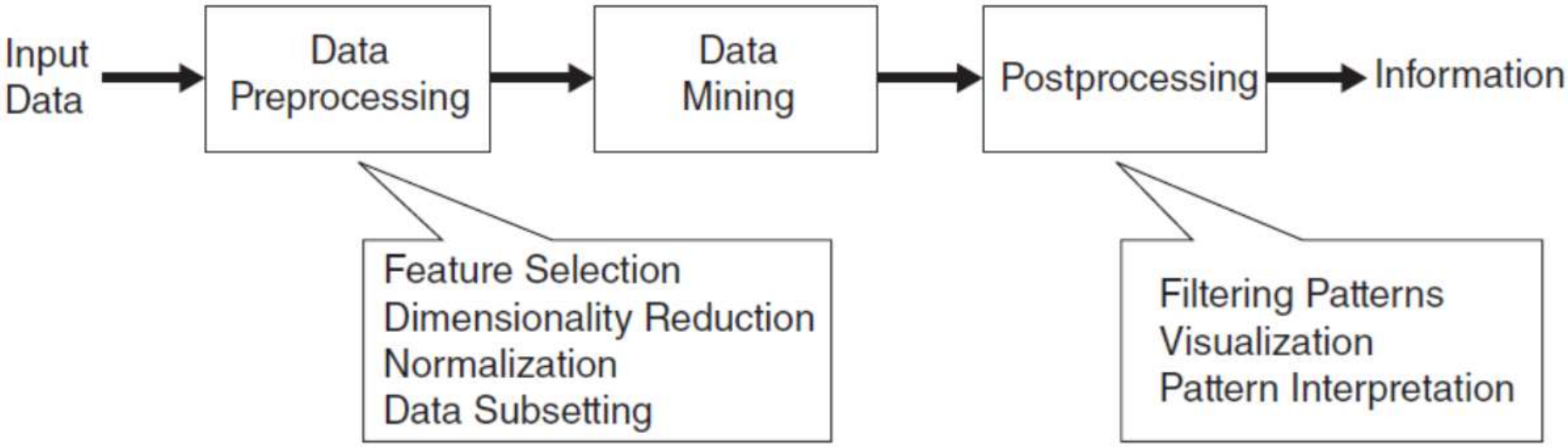


Figure 1.1. The process of knowledge discovery in databases (KDD).

Data mining

- https://www.upgrad.com/blog/what-is-data-mining-key-concepts-how-does-it-work/
- https://www.ibm.com/cloud/learn/data-mining
- https://www.youtube.com/watch?v=h82NuHDNhKI&t=209s

Definisi Data Mining

- Melakukan ekstraksi untuk mendapatkan informasi penting yang sifatnya implisit dan sebelumnya tidak diketahui, dari suatu data *(Witten et al., 2011)*

- Kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola dan hubungan dalam set data berukuran besar *(Santosa, 2007)*

- Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data *(Han et al., 2011)*

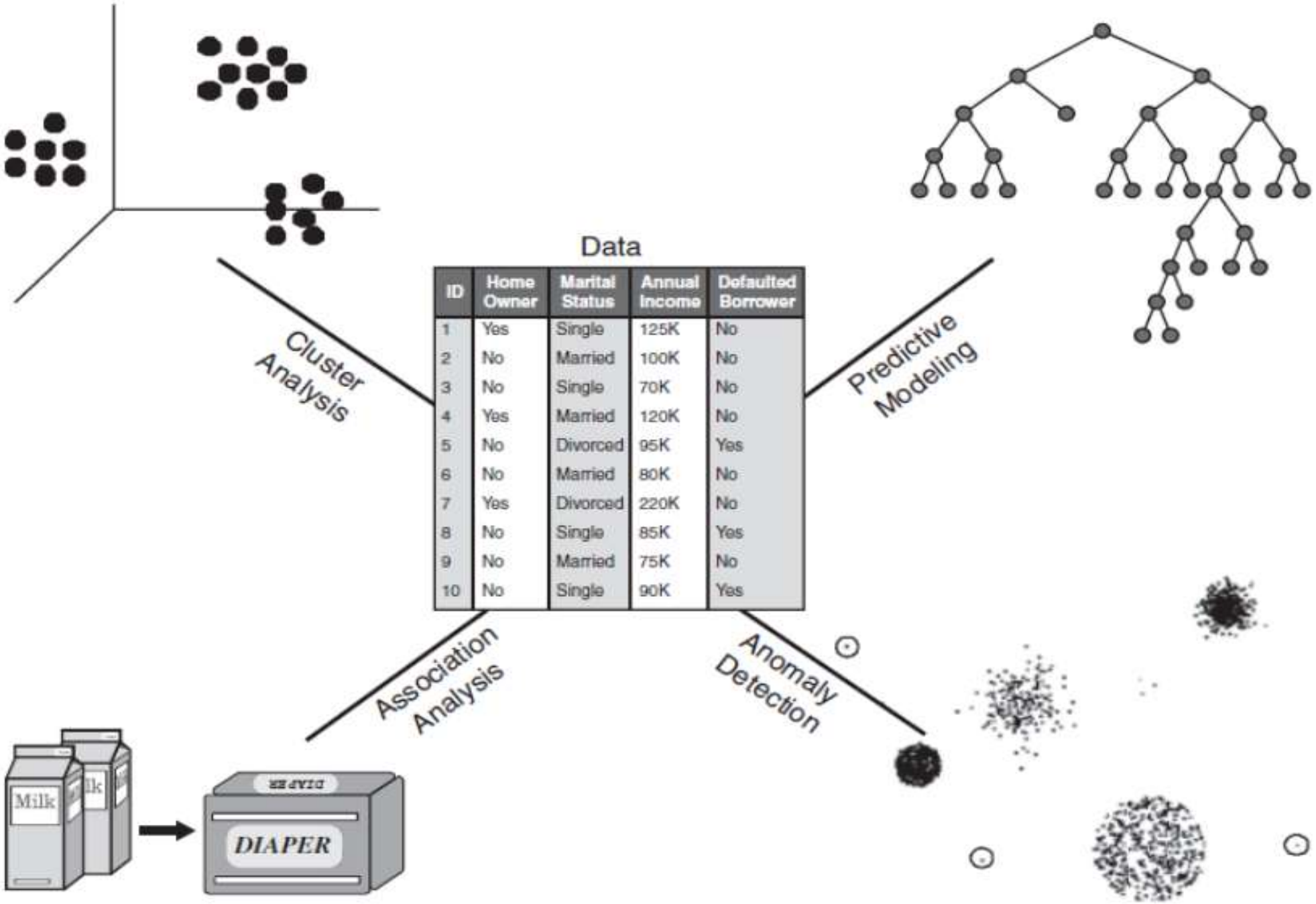
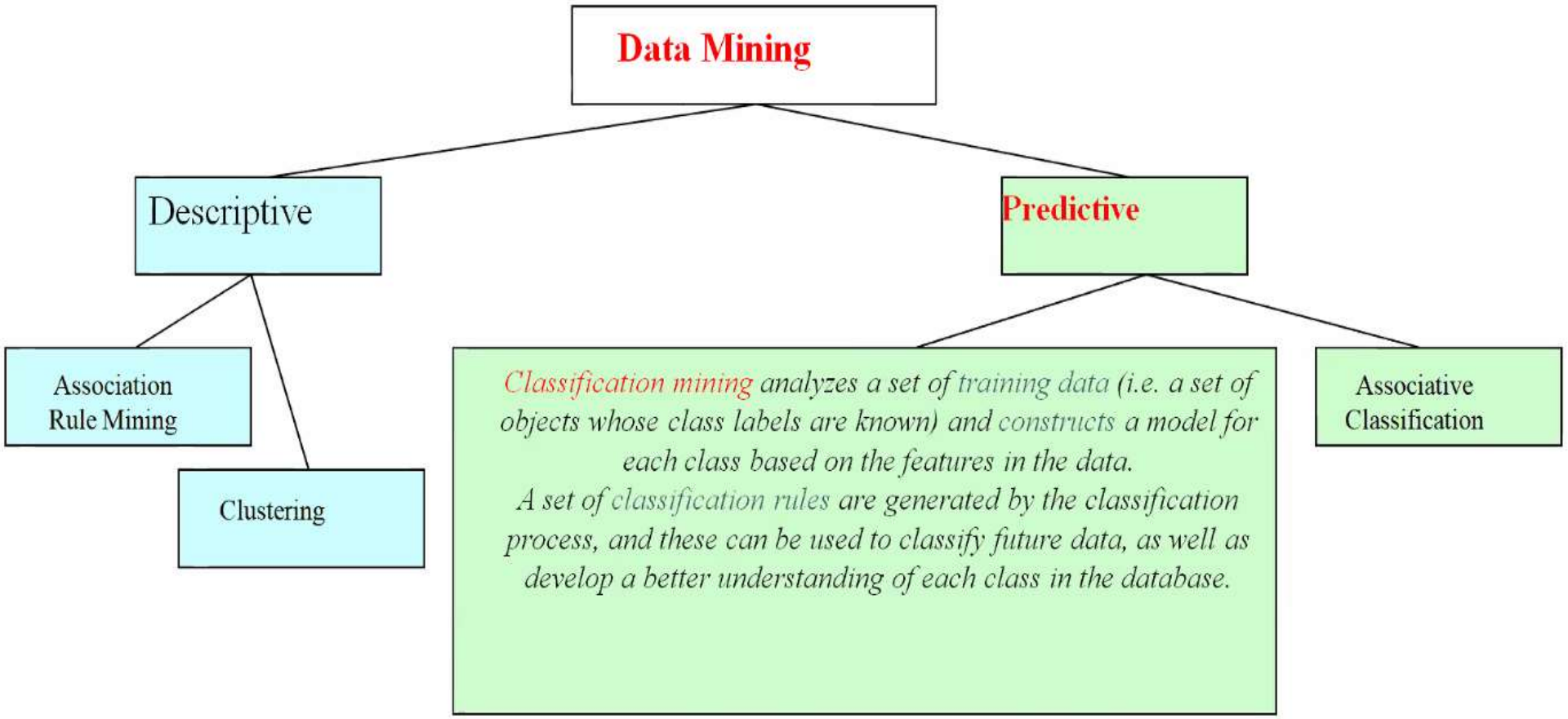


Figure 1.3. Four of the core data mining tasks.

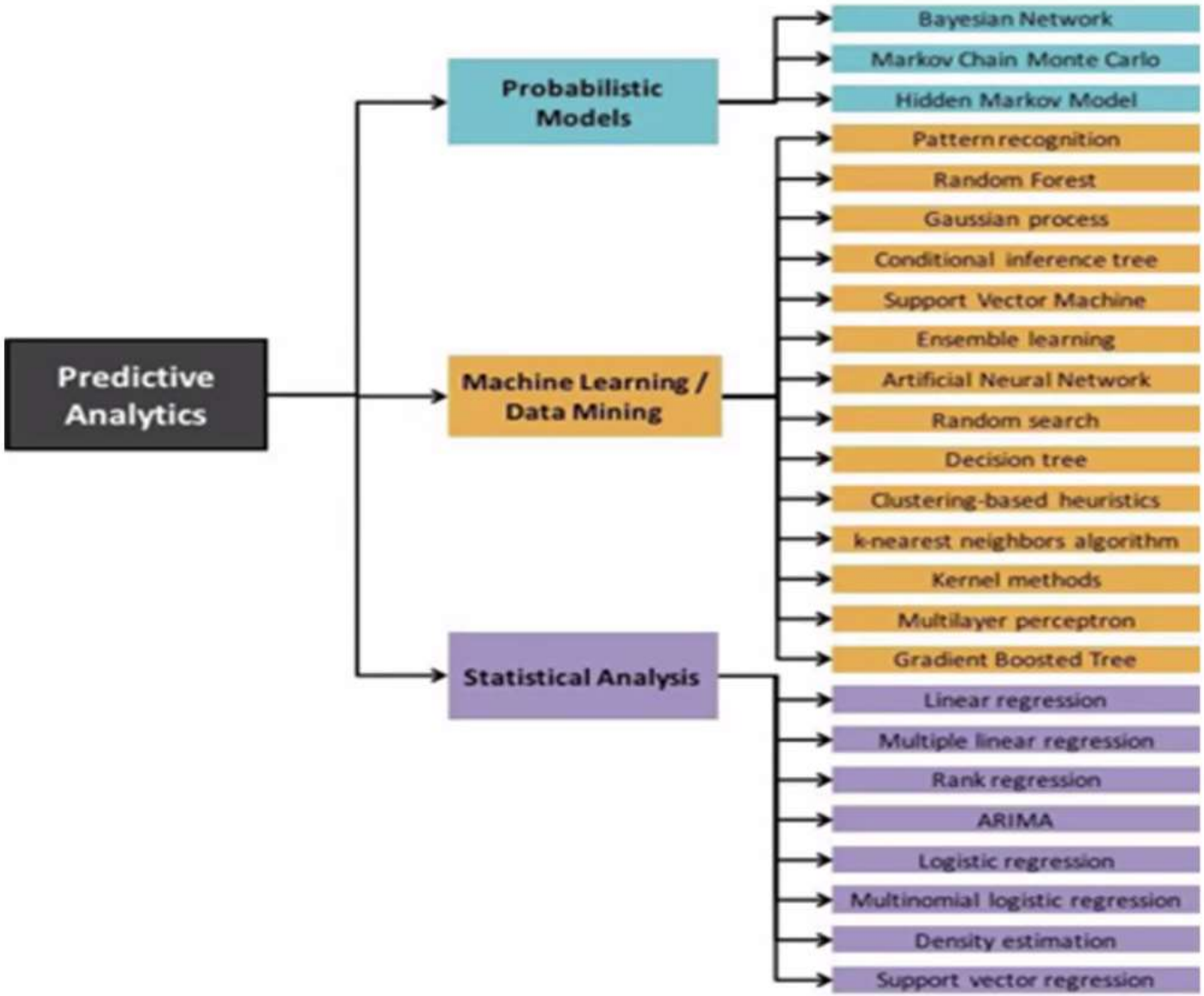

Data Mining
Descriptive
Predictive
Association Rule Mining
Clustering
Classification mining analyzes a set of training data (i.e. a set of objects whose class labels are known) and constructs a model for each class based on the features in the data.
A set of classification rules are generated by the classification process, and these can be used to classify future data, as well as develop a better understanding of each class in the database.
Associative Classification

Descriptive & Predictive Analytics

- Both the approaches uses past / historical data to get 'insight'
  - **Descriptive** approach finds many '**hidden**' trend in past data which were not noticed but have significant **business value.**
    - The **Sales of Noodles** have been mostly followed by the Sales of Tomato-Ketchup.
      - Association Rules of **Support value 75%, Confidence Value 80%.**
  - **Descriptive Analytics** though has latent predictive aspects as well.
    - **Future Sales of Noodles** will also be followed by the Sales of Tomato-Ketchup.as there are ***strongly associated*** items.
  - **Predictive Approach** typically answers the Question **" What will be sales figures of this items in coming month"** and **also Prescriptive approach is supposed to answer the Question " What should be best strategy of my company if the sales figures of this items is this much in coming month".**
  - **Prescriptive Analytics** would prescribe or recommend the action or set of actions that should be taken **not only for one predictions but for all possible predictions in the given context.**

Machine Learning & Data Mining

Categories of Predictive Analytics Methods



Dr. Suryasatria Trihandaru

Apa Keunggulan dan Kekurangan R?

Keunggulan

- Cepat dan gratis
- Ahli statistika dapat mengembangkan metode dengan membuat *package*
- Kemampuan grafik yang baik
- Unggul untuk simulasi dan analisis yang membutuhkan pemrograman secara intensif
- Mendorong kita berpikir tentang analisis data

Kekurangan

- Tidak *user friendly*
- Bagaimana menggunakan suatu fungsi dapat membuat kita frustasi
- Mudah membuat kesalahan. *Error* sudah menjadi hal yang biasa
- Penyiapan data dapat menjadi suatu yang sulit

Matlab vs R

Description	MATLAB	R
$a+b$, $a-b$, ab, a/b	`a+b`, `a-b`, `a*b`, `a/b`	`a+b`, `a-b`, `a*b`, `a/b`
$\sqrt{a}$	`sqrt(a)`	`sqrt(a)`
a^b	`a^b`	`a^b`
$\|a\|$ (note: for complex arguments, this computes the modulus)	`abs(a)`	`abs(a)`
e^a	`exp(a)`	`exp(a)`
$\ln(a)$	`log(a)`	`log(a)`
$\log_2(a)$, $\log_{10}(a)$	`log2(a)`, `log10(a)`	`log2(a)`, `log10(a)`
$\sin(a)$, $\cos(a)$, $\tan(a)$	`sin(a)`, `cos(a)`, `tan(a)`	`sin(a)`, `cos(a)`, `tan(a)`
$\sin^{-1}(a)$, $\cos^{-1}(a)$, $\tan^{-1}(a)$	`asin(a)`, `acos(a)`, `atan(a)`	`asin(a)`, `acos(a)`, `atan(a)`
$\sinh(a)$, $\cosh(a)$, $\tanh(a)$	`sinh(a)`, `cosh(a)`, `tanh(a)`	`sinh(a)`, `cosh(a)`, `tanh(a)`
$\sinh^{-1}(a)$, $\cosh^{-1}(a)$, $\tanh^{-1}(a)$	`asinh(a)`, `acosh(a)`, `atanh(a)`	`asinh(a)`, `acosh(a)`, `atanh(a)`
n MOD k (modulo arithmetic)	`mod(n,k)`	`n %% k`
Round to nearest integer	`round(x)`	`round(x)` (Note: R uses IEC 60559 standard, rounding 5 to the even digit — so e.g. `round(0.5)` gives 0, not 1.)
Round down to next lowest integer	`floor(x)`	`floor(x)`
Round up to next largest integer	`ceil(x)`	`ceiling(x)`
Round toward zero	`fix(x)`	`trunc(x)`

Matlab vs R

Description	MATLAB	R
Vector dot product $\vec{x} \cdot \vec{y} = \vec{x}^T \vec{y}$	`dot(x,y)`	`sum(x*y)`
Vector cross product $\vec{x} \times \vec{y}$	`cross(x,y)`	Not in base R, but you can use `cross(x,y)` after loading the **pracma** package (see item 348 for how to install/load packages)
Matrix multiplication AB	`A * B`	`A %*% B`
Element-by-element multiplication of A and B	`A .* B`	`A * B`
Transpose of a matrix, A^T	`A'` (This is actually the complex conjugate (i.e. Hermitian) transpose; use `A.'` for the non-conjugate transpose if you like; they are equivalent for real matrices.)	`t(A)` for transpose, or `Conj(t(A))` for conjugate (Hermitian) transpose
Solve $A\vec{x} = \vec{b}$	`A\b` Warning: if there is no solution, MATLAB gives you a least-squares "best fit." If there are many solutions, MATLAB just gives you one of them.	`solve(A,b)` Warning: this only works with square invertible matrices.
Reduced echelon form of A	`rref(A)`	R does not have a function to do this
Determinant of $\mathbf{A}$	`det(A)`	`det(A)`
Inverse of $\mathbf{A}$	`inv(A)`	`solve(A)`
Trace of $\mathbf{A}$	`trace(A)`	`sum(diag(A))`
AB^{-1}	`A/B`	`A %*% solve(B)`
Element-by-element division of A and B	`A ./ B`	`A / B`
$A^{-1}B$	`A\B`	`solve(A,B)`
Square the matrix A	`A^2`	`A %*% A`
Raise matrix A to the k^{th} power	`A^k`	(No easy way to do this in R other than repeated multiplication `A %*% A %*% A...`)
Raise each element of A to the k^{th} power	`A.^k`	`A^k`
Rank of matrix A	`rank(A)`	`qr(A)$rank`
Set $\mathbf{w}$ to be a vector of eigenvalues of $\mathbf{A}$, and $\mathbf{V}$ a matrix containing the corresponding eigenvectors	`[V,D]=eig(A)` and then `w=diag(D)` since MATLAB returns the eigenvalues on the diagonal of $\mathbf{D}$	`tmp=eigen(A); w=tmp$values; V=tmp$vectors`

R vs Python	
R is a programming language and software environment for statistical computing, graphics representation and reporting.	Python is an interpreted high-level programming language for general purpose programming.
Developed By	
R is supported by the R Foundation for Statistical Computing.	Python is supported by the Python Software Foundation.
Data Structures	
R supports data structures such as vectors, lists, matrices, arrays, factors and data frames.	Python supports data structure such as lists, dictionaries and tuples.
Switch Statement	
R supports switch statement.	Python does not support switch statement.
Scripts	
R scripts end with. R extension.	Python scripts end with .py extension.
IDE	
The common IDE for R programming is RStudio.	The common IDEs for Python programming are PyCharm and Eclipse.
Applications	
R can be used for statistical computing, machine learning and data analytics.	Python can be used for multiple applications such as machine learning, web development, networking, scientific computing, automation, natural language processing, etc.

Belajar Bahasa R
https://www.youtube.com/watch?v=rcwlltbsl7k&list=PLIeJsyt_FUfJS6o2fMNuI7hGeBpqRkYMS&index=1

Berikut daftar video Data Science Masterclass dengan Menggunakan R:

1. Kurikulum: https://youtu.be/rcwlltbsl7k
2. Cara Install Program R: https://youtu.be/yByyAFHxH1U
3. Jenis Data Statistik: https://youtu.be/Q_Nw0UxjS3U
4. Dasar-dasar Bahasa R: https://youtu.be/uO9SDhW0PgY
5. Belajar Matriks di R: https://youtu.be/qiLTYtHQ6uk
6. Data Kategori di R: https://youtu.be/3ATWuBnyKH8
7. Data Frame di R: https://youtu.be/-aCG_1yOFDE
8. List di R: https://youtu.be/O-BvHXXXiC0
9. Export Import Data di R: https://youtu.be/e79EzxD2yQE
10. Logical Operators di R: https://youtu.be/m--594V4F7k
11. Conditionals di R: https://youtu.be/1oFF5TgOuzc
12. Looping di R: https://youtu.be/V9kUC5tRELw
13. Belajar Fungsi di R: https://youtu.be/mhMo_7kqiEs
14. Penggunaan Fungsi Tingkat Mahir: https://youtu.be/23oOtidxi20
15. Analisis Data Waktu: https://youtu.be/uJF6ydAKvIA
16. ...

Integrated Development Environment

RStudio

file:///C:/Program%20Files/RStudio/www/docs/keyboard.htm

RStudio Desktop 1.2.5042 - Release Notes

1. Install R. RStudio requires R 3.0.1+.

2. Download RStudio Desktop. Recommended for your system:

DOWNLOAD RSTUDIO FOR MAC
1.2.5042 | 126.89MB

Requires macOS 10.13+ (64-bit)

All Installers

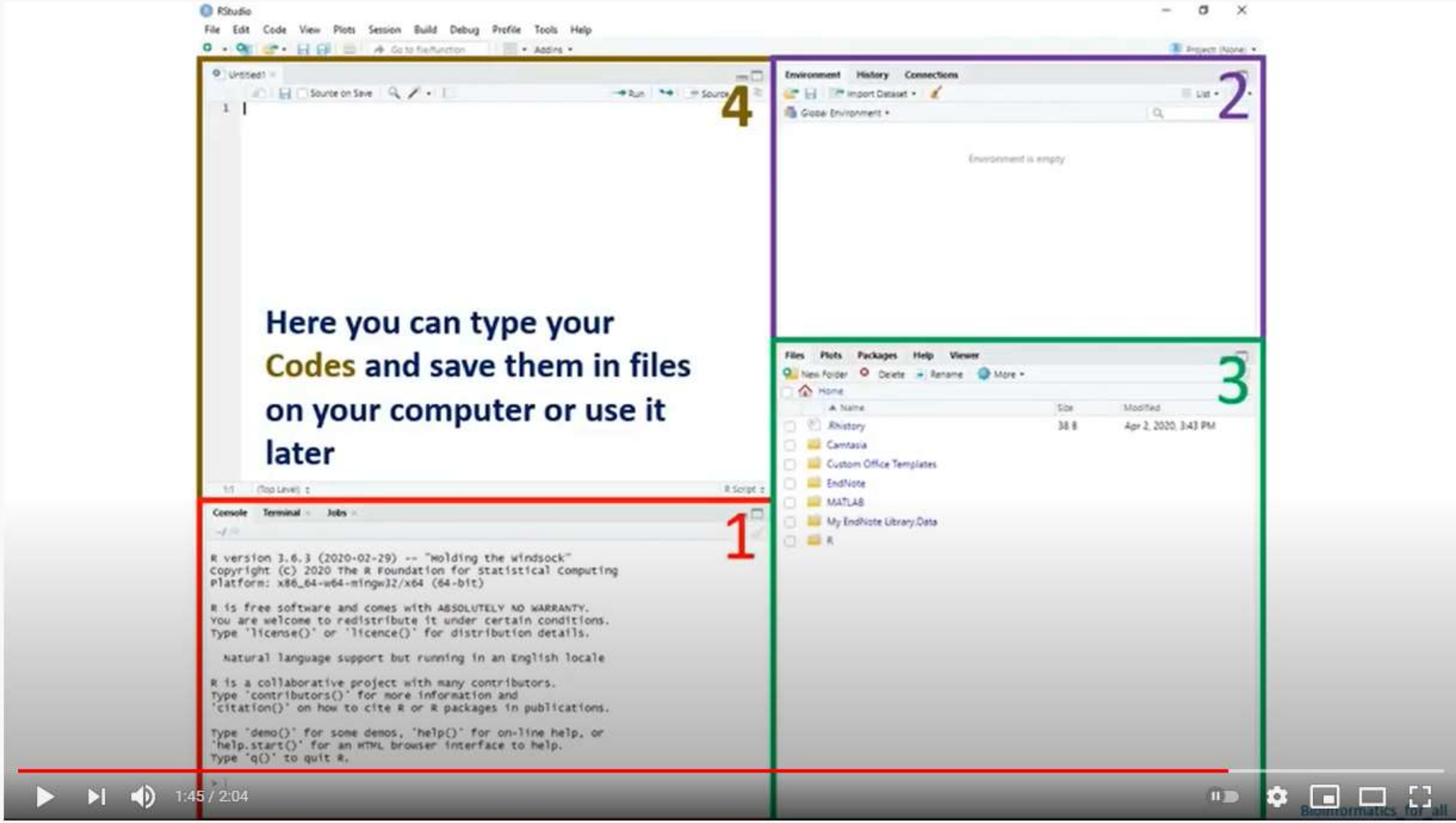
Linux users may need to import RStudio's public code-signing key prior to installation, depending on the operating system's security policy.

RStudio 1.2 requires a 64-bit operating system. If you are on a 32 bit system, you can use an older version of RStudio.

OS	Download	Size	SHA-256
Windows 10/8/7	RStudio-1.2.5042.exe	149.84 MB	

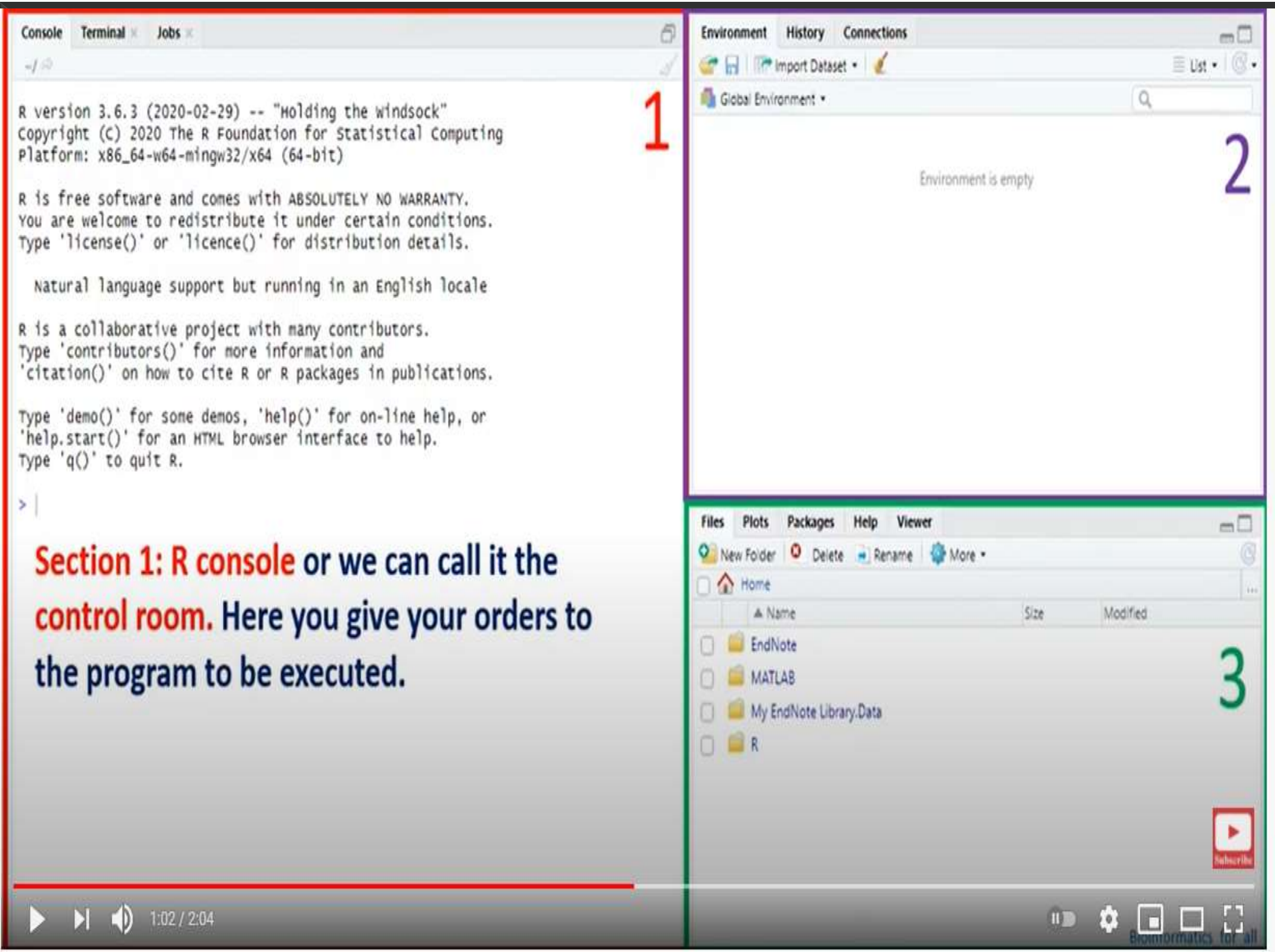
11:37 / 15:08

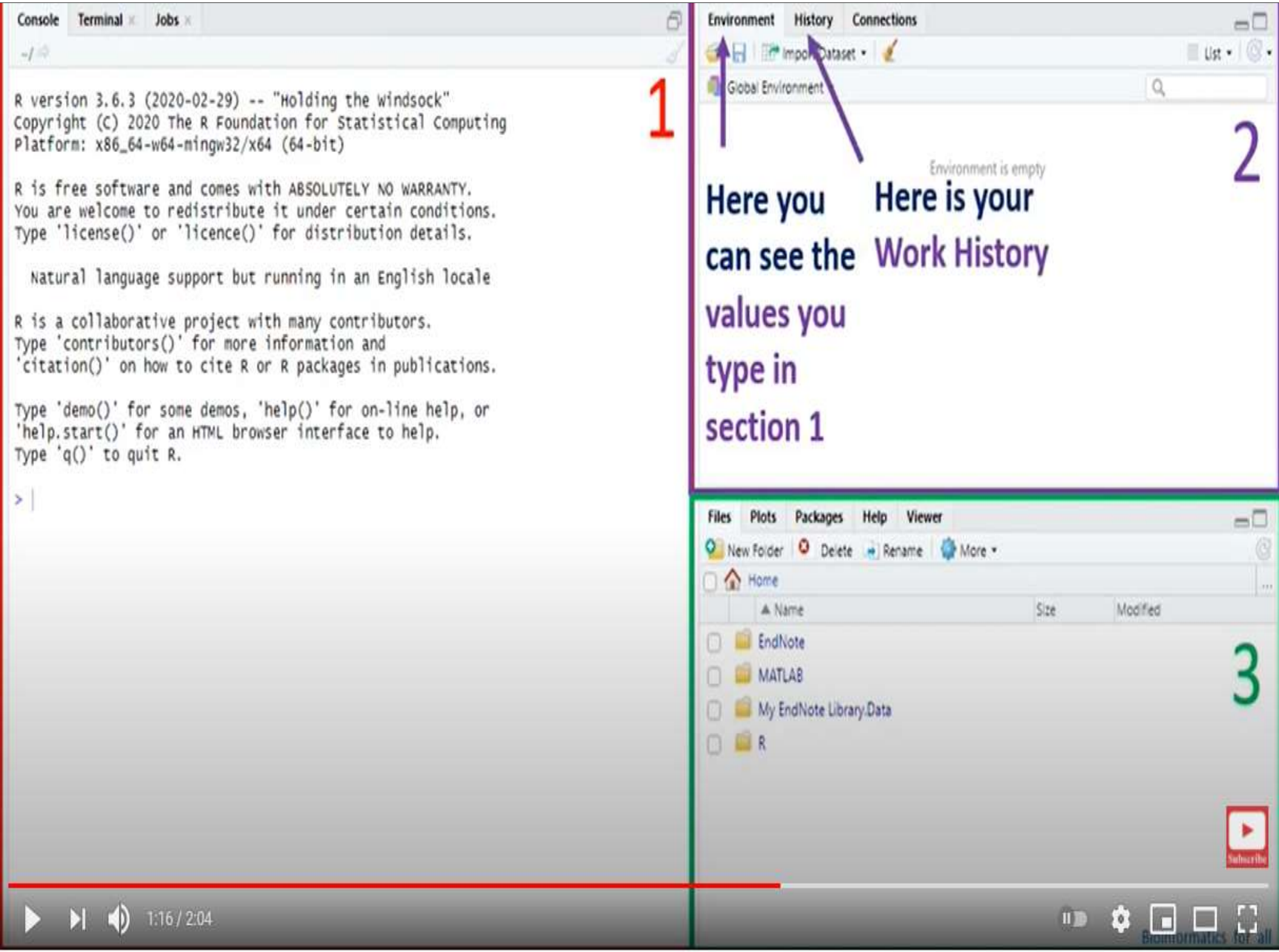
Pengenalan RStudio

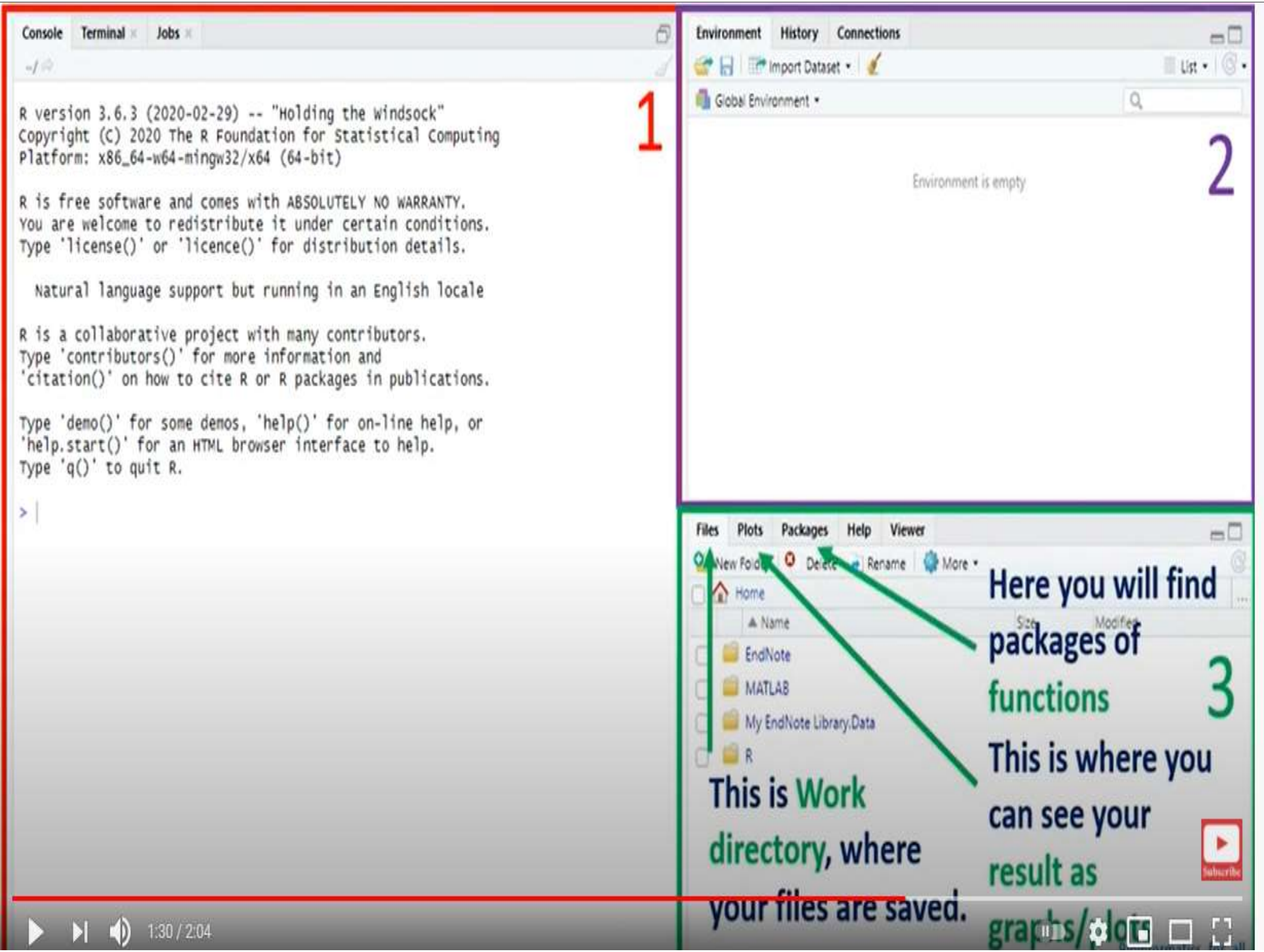


#R #RStudio #learnR

R Series #1 Install and run R & RStudio in very simple steps


Console
Terminal
Jobs
R version 3.6.3 (2020-02-29) -- "Holding the Windsock"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (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.
>
1
Environment
History
Connections
Import Dataset
List
Global Environment
Environment is empty
2
Files
Plots
Packages
Help
Viewer
New Folder
Delete
Rename
More
Home
Name
Size
Modified
EndNote
MATLAB
My EndNote Library.Data
R
3
Section 1: R console or we can call it the control room. Here you give your orders to the program to be executed.
1:02 / 2:04


Console
Terminal
Jobs
R version 3.6.3 (2020-02-29) -- "Holding the Windsock"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (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.
>
1
Environment
History
Connections
Import Dataset
List
Global Environment
Environment is empty
2
Here you can see the values you type in section 1
Here is your Work History
Files
Plots
Packages
Help
Viewer
New Folder
Delete
Rename
More
Home
Name
Size
Modified
EndNote
MATLAB
My EndNote Library.Data
R
3
Subscribe
1:16 / 2:04
Bioinformatics for all

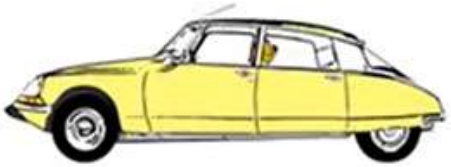

Console
Terminal
Jobs
R version 3.6.3 (2020-02-29) -- "Holding the Windsock"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type '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.
>
1
Environment
History
Connections
Import Dataset
List
Global Environment
Environment is empty
2
Files
Plots
Packages
Help
Viewer
New Folder
Delete
Rename
More
Home
Name
Size
Modified
EndNote
MATLAB
My EndNote Library.Data
R
Here you will find
packages of
functions
3
This is Work
directory, where
your files are saved.
This is where you
can see your
result as
graphs/plots
1:30 / 2:04

https://www.youtube.com/watch?v=qWJJG8yM98g

Data structure in R

Scalar

Vector



Matrix

Array

Data types in R

Numeric

Logical

Character

Data types

Numeric

Logical

Character

Data structures

	Homogeneous	Heterogeneous
1D	 Scalar  Vector	 List
2D	 Matrix	 Table/Data frame
nD	 Array	

Base R
Cheat Sheet

Getting Help

Accessing the help files

?mean
Get help of a particular function.
help.search('weighted mean')
Search the help files for a word or phrase.
help(package = 'dplyr')
Find help for a package.

More about an object

str(iris)
Get a summary of an object's structure.
class(iris)
Find the class an object belongs to.

Using Libraries

install.packages('dplyr')
Download and install a package from CRAN.

library(dplyr)
Load the package into the session, making all its functions available to use.

dplyr::select
Use a particular function from a package.

data(iris)
Load a built-in dataset into the environment.

Working Directory

getwd()
Find the current working directory (where inputs are found and outputs are sent).

setwd('C://file/path')
Change the current working directory.

Use projects in RStudio to set the working directory to the folder you are working in.

Vectors

Creating Vectors

c(2, 4, 6)	2 4 6	Join elements into a vector
2:6	2 3 4 5 6	An integer sequence
seq(2, 3, by=0.5)	2.0 2.5 3.0	A complex sequence
rep(1:2, times=3)	1 2 1 2 1 2	Repeat a vector
rep(1:2, each=3)	1 1 1 2 2 2	Repeat elements of a vector

Vector Functions

sort(x)
Return x sorted.
rev(x)
Return x reversed.
table(x)
See counts of values.
unique(x)
See unique values.

Selecting Vector Elements

By Position

x[4]	The fourth element.
x[-4]	All but the fourth.
x[2:4]	Elements two to four.
x[-(2:4)]	All elements except two to four.
x[c(1, 5)]	Elements one and five.

By Value

x[x == 10]	Elements which are equal to 10.
x[x < 0]	All elements less than zero.
x[x %in% c(1, 2, 5)]	Elements in the set 1, 2, 5.

Named Vectors

x['apple']	Element with name 'apple'.

Programming

For Loop

```
for (variable in sequence){
   Do something
}
```

Example

```
for (i in 1:4){
   j <- i + 10
   print(j)
}
```

While Loop

```
while (condition){
   Do something
}
```

Example

```
while (i < 5){
   print(i)
   i <- i + 1
}
```

If Statements

```
if (condition){
   Do something
} else {
   Do something different
}
```

Example

```
if (i > 3){
   print('Yes')
} else {
   print('No')
}
```

Functions

```
function_name <- function(var){
   Do something
   return(new_variable)
}
```

Example

```
square <- function(x){
   squared <- x*x
   return(squared)
}
```

Reading and Writing Data

Input	Ouput	Description
df <- read.table('file.txt')	write.table(df, 'file.txt')	Read and write a delimited text file.
df <- read.csv('file.csv')	write.csv(df, 'file.csv')	Read and write a comma separated value file. This is a special case of read.table/ write.table.
load('file.RData')	save(df, file = 'file.Rdata')	Read and write an R data file, a file type special for R.

Conditions

a == b	Are equal	a > b	Greater than	a >= b	Greater than or equal to	is.na(a)	Is missing
a != b	Not equal	a < b	Less than	a <= b	Less than or equal to	is.null(a)	Is null

Types

Converting between common data types in R. Can always go from a higher value in the table to a lower value.

`as.logical`	`TRUE, FALSE, TRUE`	Boolean values (TRUE or FALSE).
`as.numeric`	`1, 0, 1`	Integers or floating point numbers.
`as.character`	`'1', '0', '1'`	Character strings. Generally preferred to factors.
`as.factor`	`'1', '0', '1', levels: '1', '0'`	Character strings with preset levels. Needed for some statistical models.

Maths Functions

`log(x)`	Natural log.	`sum(x)`	Sum.
`exp(x)`	Exponential.	`mean(x)`	Mean.
`max(x)`	Largest element.	`median(x)`	Median.
`min(x)`	Smallest element.	`quantile(x)`	Percentage quantiles.
`round(x, n)`	Round to n decimal places.	`rank(x)`	Rank of elements.
`signif(x, n)`	Round to n significant figures.	`var(x)`	The variance.
`cor(x, y)`	Correlation.	`sd(x)`	The standard deviation.

Variable Assignment

```
> a <- 'apple'
> a
[1] 'apple'
```

The Environment

`ls()`	List all variables in the environment.
`rm(x)`	Remove x from the environment.
`rm(list = ls())`	Remove all variables from the environment.

You can use the environment panel in RStudio to browse variables in your environment.

Matrixes

`m <- matrix(x, nrow = 3, ncol = 3)`
Create a matrix from x.

`m[2, ]` - Select a row

`m[ , 1]` - Select a column

`m[2, 3]` - Select an element

`t(m)`
Transpose

`m %*% n`
Matrix Multiplication

`solve(m, n)`
Find x in: m * x = n

Lists

`l <- list(x = 1:5, y = c('a', 'b'))`
A list is collection of elements which can be of different types.

`l[[2]]`	`l[1]`	`l$x`	`l['y']`
Second element of l.	New list with only the first element.	Element named x.	New list with only element named y.

Also see the **dplyr** library.

Data Frames

`df <- data.frame(x = 1:3, y = c('a', 'b', 'c'))`
A special case of a list where all elements are the same length.

x	y
1	a
2	b
3	c

List subsetting

`df$x` 

`df[[2]]` 

Understanding a data frame

`View(df)`	See the full data frame.
`head(df)`	See the first 6 rows.

Matrix subsetting

`df[ , 2]` 

`df[2, ]` 

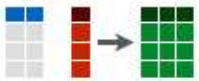
`df[2, 2]` 

`nrow(df)`
Number of rows.

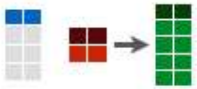
`ncol(df)`
Number of columns.

`dim(df)`
Number of columns and rows.

`cbind` - Bind columns.



`rbind` - Bind rows.



Strings

Also see the **stringr** library.

`paste(x, y, sep = ' ')`	Join multiple vectors together.
`paste(x, collapse = ' ')`	Join elements of a vector together.
`grep(pattern, x)`	Find regular expression matches in x.
`gsub(pattern, replace, x)`	Replace matches in x with a string.
`toupper(x)`	Convert to uppercase.
`tolower(x)`	Convert to lowercase.
`nchar(x)`	Number of characters in a string.

Factors

`factor(x)`
Turn a vector into a factor. Can set the levels of the factor and the order.

`cut(x, breaks = 4)`
Turn a numeric vector into a factor but 'cutting' into sections.

Statistics

`lm(x ~ y, data=df)`
Linear model.

`glm(x ~ y, data=df)`
Generalised linear model.

`summary`
Get more detailed information out a model.

`t.test(x, y)`
Preform a t-test for difference between means.

`pairwise.t.test`
Preform a t-test for paired data.

`prop.test`
Test for a difference between proportions.

`aov`
Analysis of variance.

Distributions

	Random Variates	Density Function	Cumulative Distribution	Quantile
Normal	`rnorm`	`dnorm`	`pnorm`	`qnorm`
Poison	`rpois`	`dpois`	`ppois`	`qpois`
Binomial	`rbinom`	`dbinom`	`pbinom`	`qbinom`
Uniform	`runif`	`dunif`	`punif`	`qunif`

Plotting

Also see the **ggplot2** library.

`plot(x)`
Values of x in order.

`plot(x, y)`
Values of x against y.

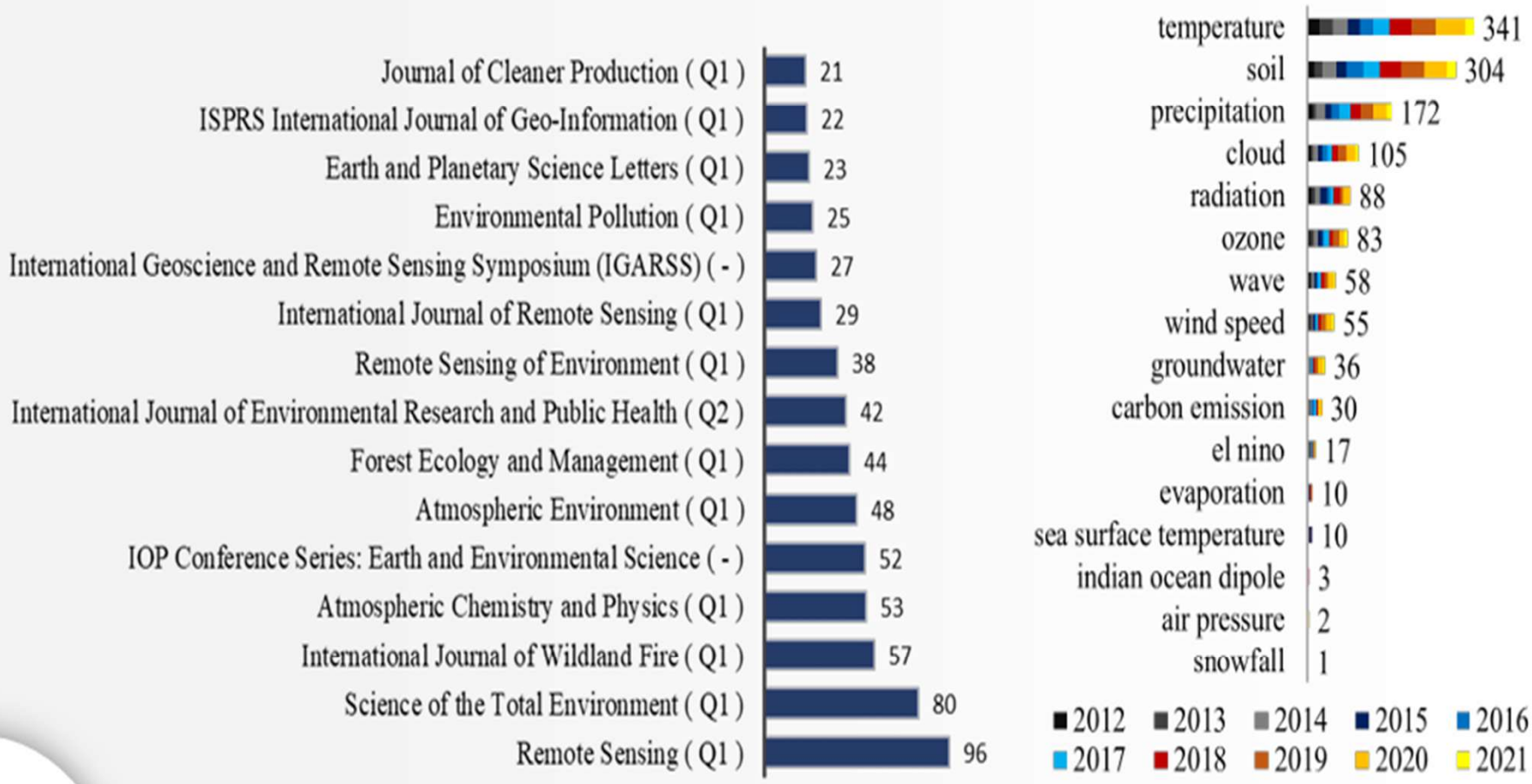
`hist(x)`
Histogram of x.

Dates

See the **lubridate** library.

Systematic Literature Review

Indikator iklim yang digunakan dan tempat publikasi



Gambar 3. SLR pada Penelitian Kebakaran Hutan: meta-data hingga Maret 2021

IPB University
— Bogor Indonesia —

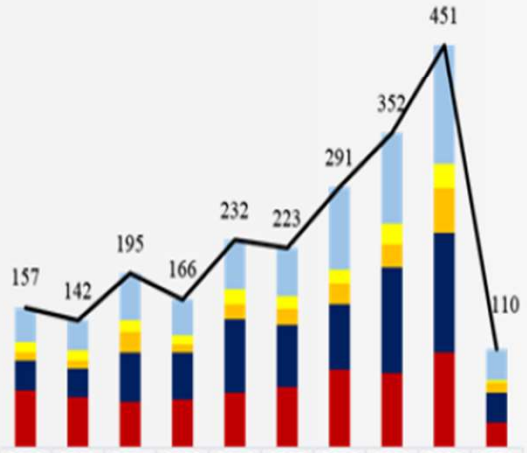
Computational Mathematics for Earth System Science

Prof. Dr. Ir. Sri Nurdiati, M.Sc.

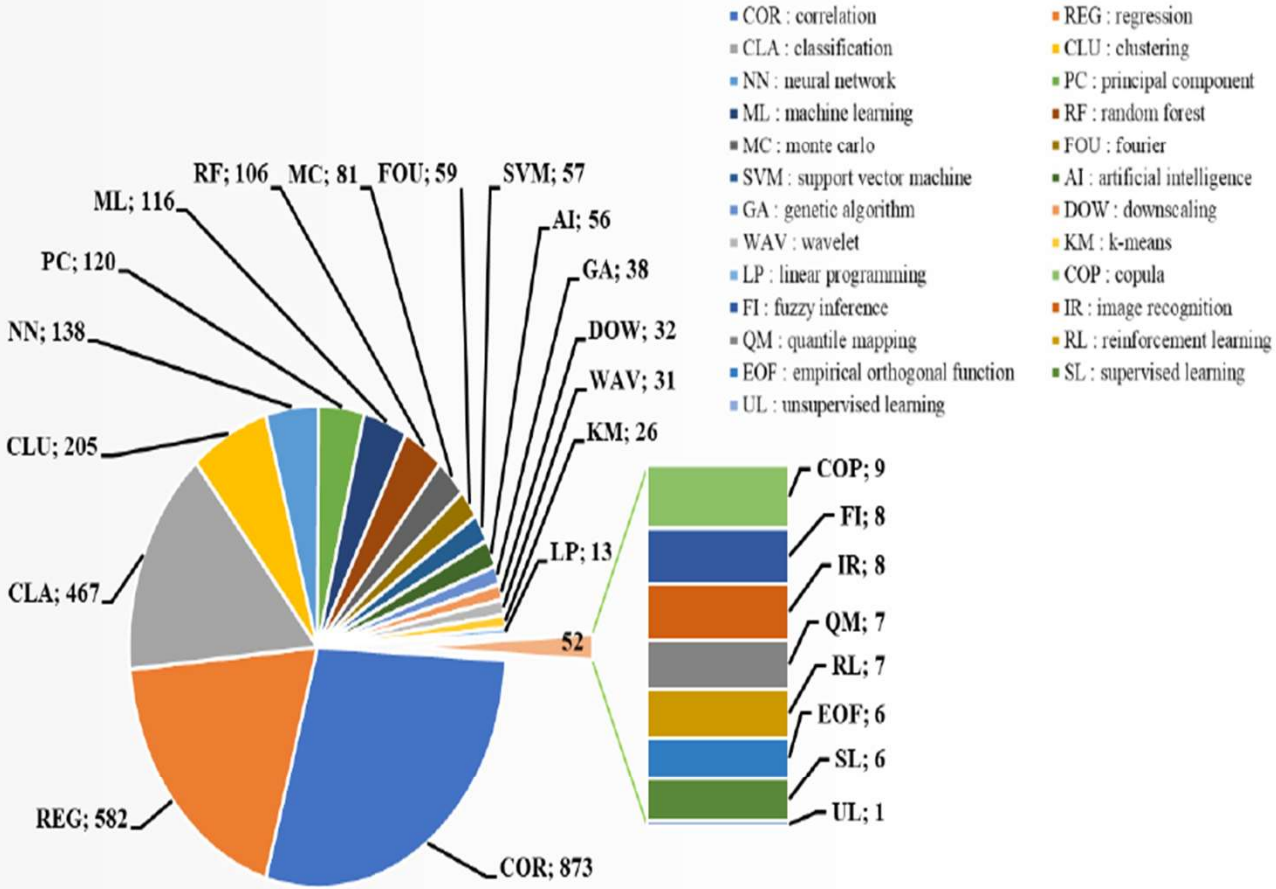
19 Maret 2024

Systematic Literature Review

Teknik-teknik dalam bidang Matematika, Statistika, dan Ilmu Komputer sangat berguna dalam analisis pada data Iklim.



	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Unranked	39	33	52	40	55	53	92	102	134	34
Q4	11	11	14	11	17	15	16	23	27	3
Q3	9	10	23	9	16	18	23	26	50	11
Q2	34	32	55	52	83	69	73	118	134	34
Q1	64	56	51	54	61	68	87	83	106	28
Total	157	142	195	166	232	223	291	352	451	110



Gambar 3. SLR pada Penelitian Kebakaran Hutan: meta-data hingga Maret 2021

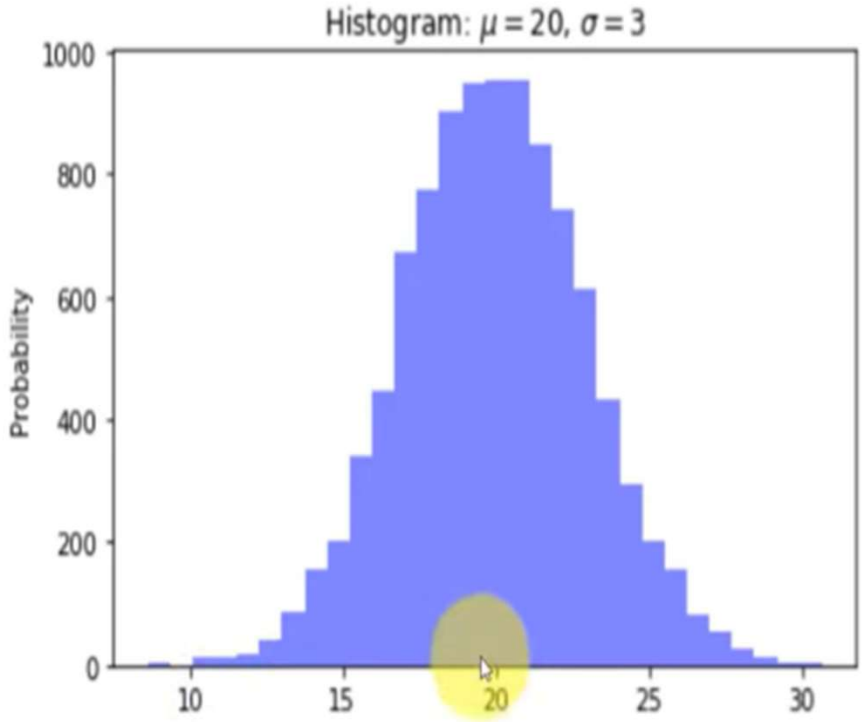
Types of Visualizations

- What do you want your visualization to show about your data?

- **Distribution:** how a variable or variables in the dataset distribute over a range of possible values.
- **Relationship:** how the values of multiple variables in the dataset relate
- **Composition:** how the dataset breaks down into subgroups
- **Comparison:** how trends in multiple variable or datasets compare

DISTRIBUTIONS

HISTOGRAMS



BOX PLOT

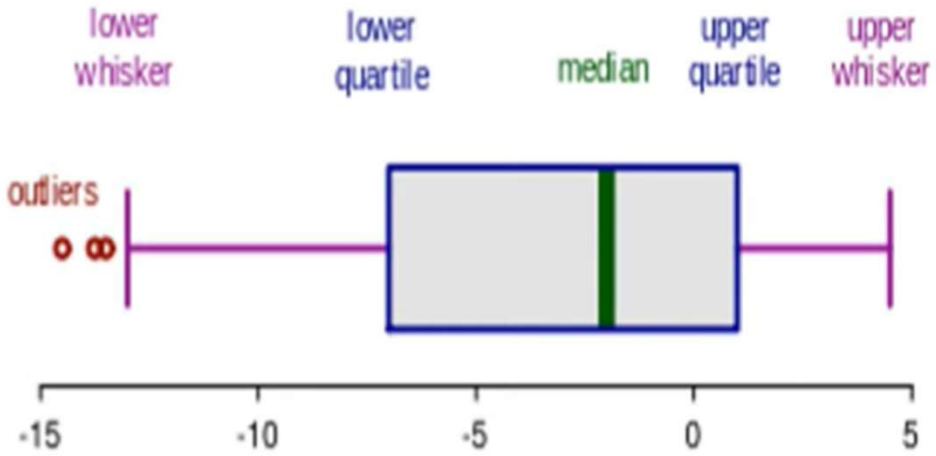
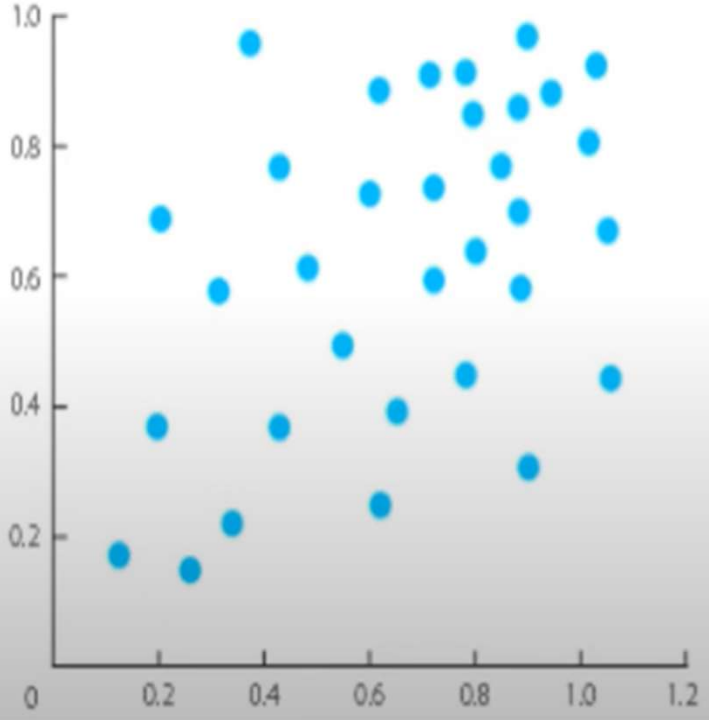


Photo Credit: https://commons.wikimedia.org/wiki/File:Elements_of_a_boxplot_en.svg

RELATIONSHIPS

SCATTERPLOT

"Scatterplot demonstrates the relationship between two variables (X, Y)"



BUBBLE CHART

"Bubble chart demonstrates the relationship between three variables (X, Y, Bubble Size)"

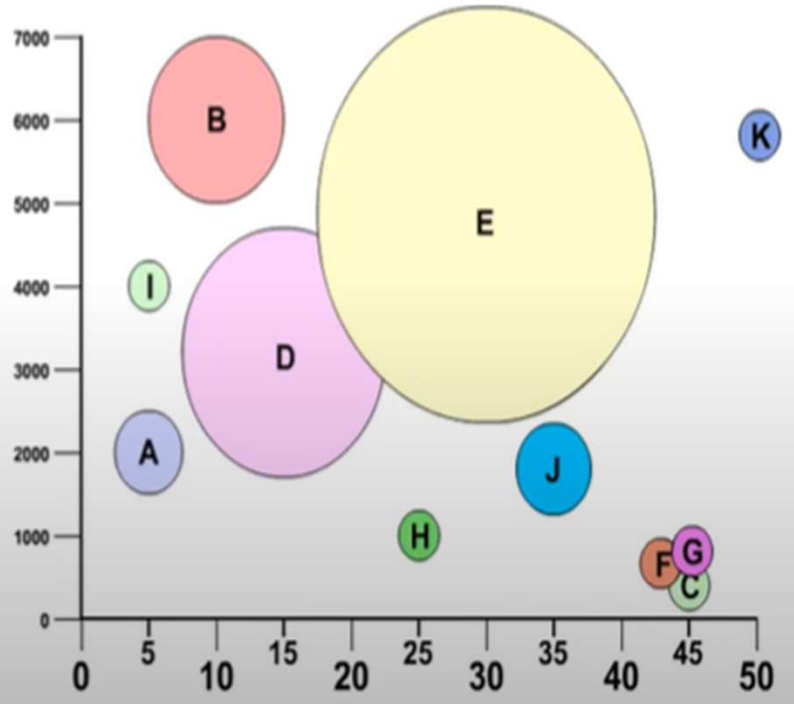


Photo Credit: https://commons.wikimedia.org/wiki/File:Example_of_Scatter_Plot.jpg

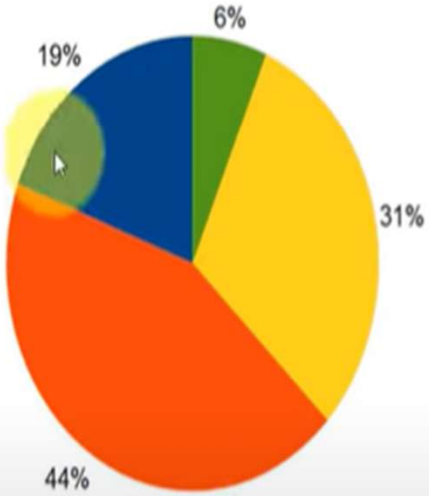
Photo Credit: https://commons.wikimedia.org/wiki/File:Bubble_chart.jpg

0:14 / 7:08

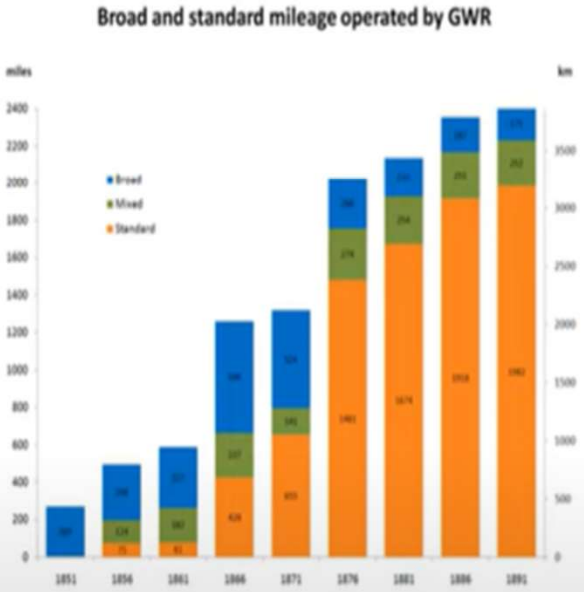
Science of Data Visualization | Bar, scatter plot, line, histograms, pie, box plots, bubble chart

COMPOSITIONS

PIE CHART



STACKED BAR CHART



STACKED AREA CHART

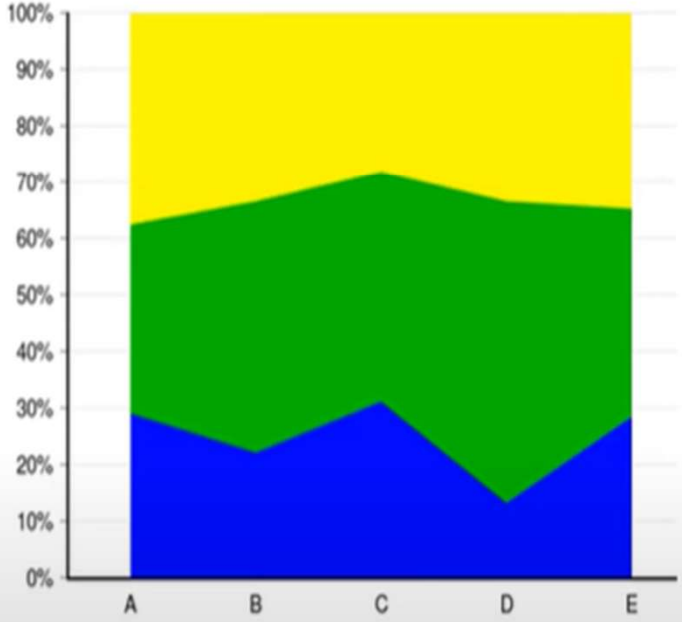


Photo Credit: https://commons.wikimedia.org/wiki/File:Broad_and_standard_mileage_operated_by_GWR.png
Photo Credit: https://commons.wikimedia.org/wiki/File:Charts_SVG_Example_12_-_Stacked_100%25_Area_Chart.svg
Photo Credit: https://commons.wikimedia.org/wiki/File:Pie-chart.jpg

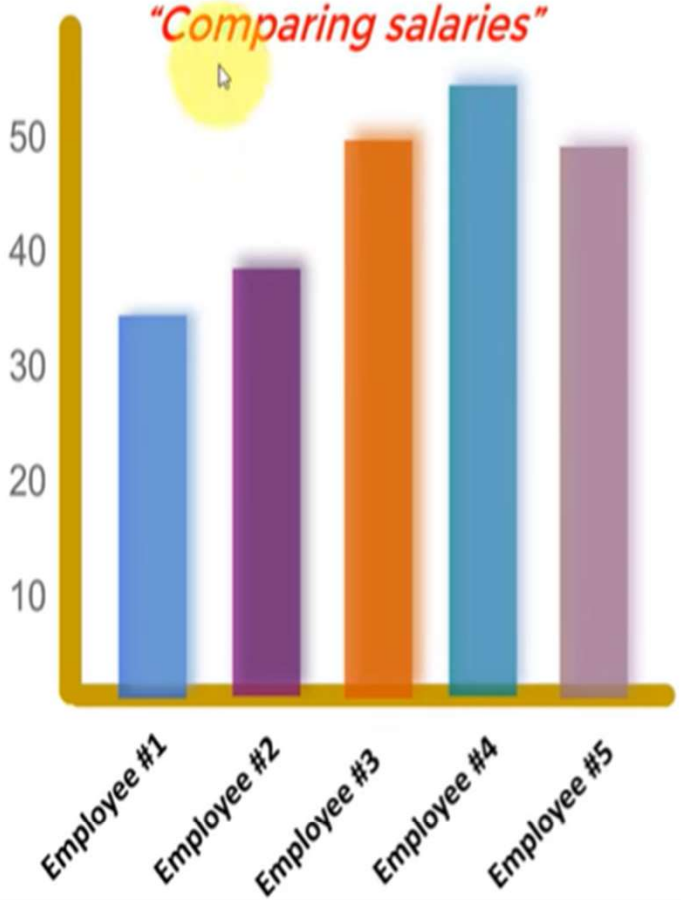
5:40 / 7:08

Science of Data Visualization | Bar, scatter plot, line, histograms, pie, box plots, bubble chart

COMPARISONS

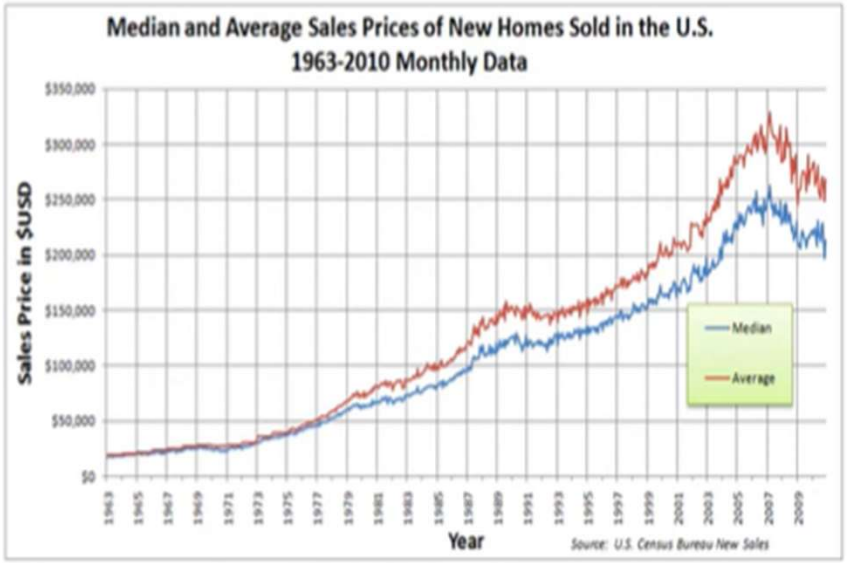
BAR CHART

"Comparing salaries"



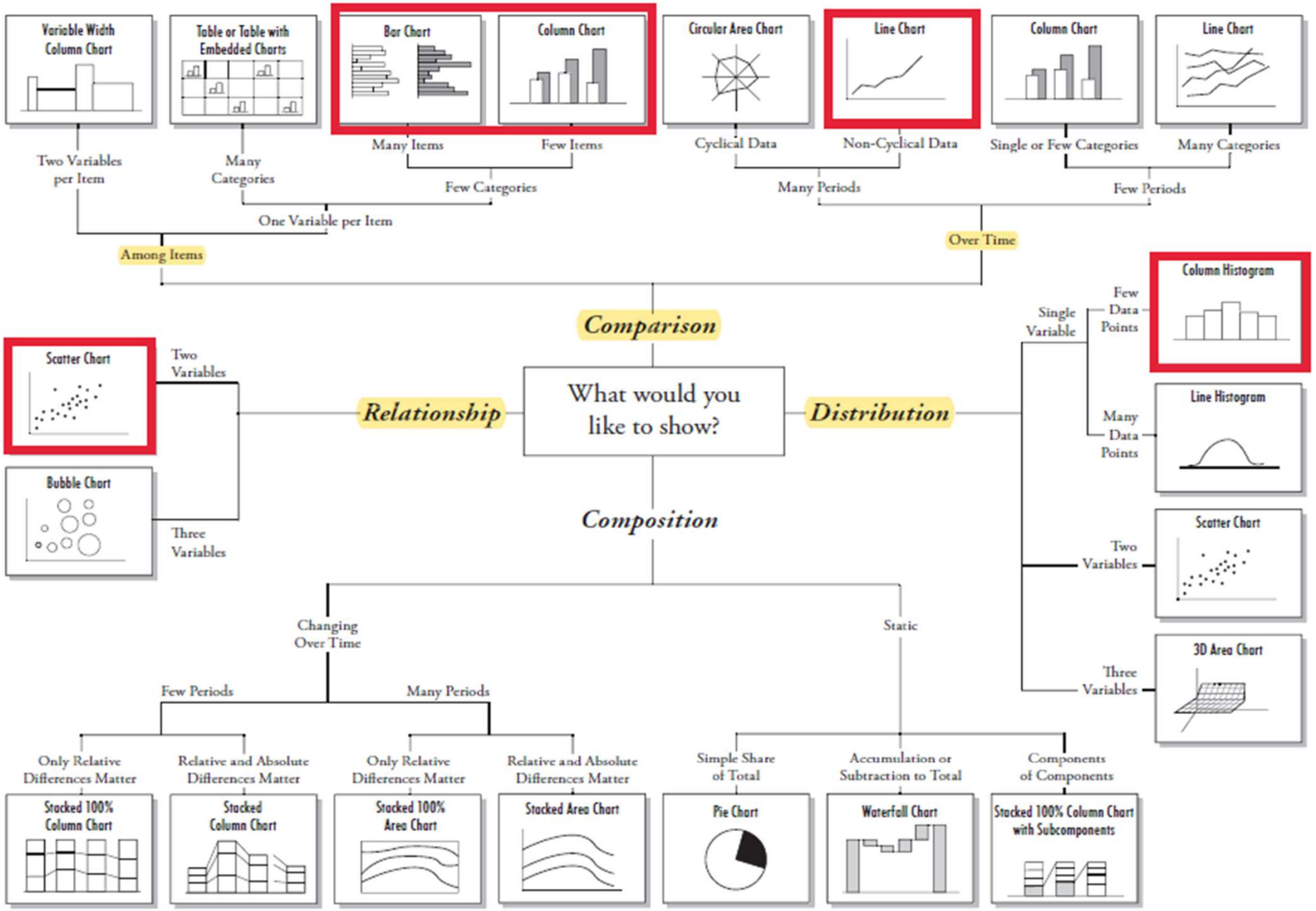
LINE CHART

"Comparing median and average House prices over the years"



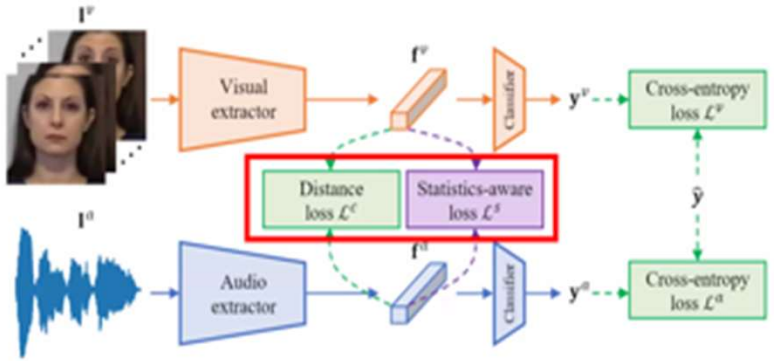
https://www.needpix.com/photo/89660/productivity-statistics-bar-chart-chart-graph-diagram-results
https://commons.wikimedia.org/wiki/File:Median_and_Average_Sales_Prices_of_New_Homes_Sold_in_the_US_1963-2010_Monthly.png

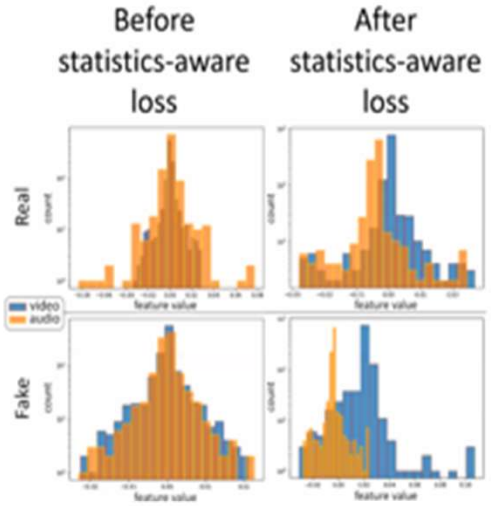
Chart Suggestions—A Thought-Starter



Network – To focus on artifacts

- Audio-visual feature matching → finding audio-visual inconsistencies
  - Real: audio features match visual features
  - Fake: audio features don't match visual features





Source: **Astrid, M.**, Ghorbel, E. and Aouada, D., 2024, October. Statistics-aware Audio-visual Deepfake Detector. In *IEEE International Conference on Image Processing (ICIP)*. (Accepted)

Materi- pengayaan

- file:///C:/Program%20Files/RStudio/www/docs/keyboard.htm
- https://www.youtube.com/watch?v=qWJJG8yM98g&list=PLnQYPiUVWCT49In0-pQ7taaxralUM8UKY&index=4
- Memahami dasar2 R
- https://www.youtube.com/watch?v=uO9SDhW0PgY

INFO TTG PYTHON

- Machine Learning Archives - SAINSDATA.ID
- STATS 220 Data Technology (earo.me)

- Tutorial Pemrograman Python - SAINSDATA.ID
- Booklet Python - SAINSDATA.ID

R For Data Science *Cheat Sheet*

Tidyverse for Beginners

Learn More R for Data Science Interactively at www.datacamp.com

Tidyverse

The **tidyverse** is a powerful collection of R packages that are actually data tools for transforming and visualizing data. All packages of the tidyverse share an underlying philosophy and common APIs.

The core packages are:

- **ggplot2**, which implements the grammar of graphics. You can use it to visualize your data.
- **dplyr** is a grammar of data manipulation. You can use it to solve the most common data manipulation challenges.
- **tidyr** helps you to create tidy data or data where each variable is in a column, each observation is a row end each value is a cell.
- **readr** is a fast and friendly way to read rectangular data.
- **purrr** enhances R's functional programming (FP) toolkit by providing a complete and consistent set of tools for working with functions and vectors.
- **tibble** is a modern re-imaginging of the data frame.
- **stringr** provides a cohesive set of functions designed to make working with strings as easy as posssible
- **forcats** provide a suite of useful tools that solve common problems with factors.

You can install the complete tidyverse with:

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

Then, load the core tidyverse and make it available in your current R session by running:

```
> library(tidyverse)
```

Note: there are many other tidyverse packages with more specialised usage. They are not loaded automatically with library(tidyverse), so you'll need to load each one with its own call to `library()`.

Useful Functions

`> tidyverse_conflicts()`	Conflicts between tidyverse and other packages
`> tidyverse_deps()`	List all tidyverse dependencies
`> tidyverse_logo()`	Get tidyverse logo, using ASCII or unicode characters
`> tidyverse_packages()`	List all tidyverse packages
`> tidyverse_update()`	Update tidyverse packages

Loading in the data

`> library(datasets)`	Load the datasets package
`> library(gapminder)`	Load the gapminder packa
`> attach(iris)`	Attach iris data to the R se

Start

dplyr

Filter

`filter()` allows you to select a subset of rows in a data frame.

`> iris %>%` `  filter(Species=="virginica")`	Select iris data of species "virginica"

`> iris %>%`
`  filter(Species=="virginica",`
`  Sepal.Length > 6)`
	Select iris data of species "virginica" and sepal length greater than 6.

Arrange

`arrange()` sorts the observations in a dataset in ascending or descending order based on one of its variables.

`> iris %>%` `  arrange(Sepal.Length)`	Sort in ascending order of sepal length
`> iris %>%` `  arrange(desc(Sepal.Length))`	Sort in descending order of sepal length

Combine multiple `dplyr` verbs in a row with the pipe operator `%>%`:

`> iris %>%`
`  filter(Species=="virginica") %>%`
`  arrange(desc(Sepal.Length))`
	Filter for species "virginica" then arrange in descending order of sepal length

Mutate

`mutate()` allows you to update or create new columns of a data frame.

`> iris %>%` `  mutate(Sepal.Length=Sepal.Length*10)`	Change Sepal.Length to be in millimeters
`> iris %>%` `  mutate(SLMm=Sepal.Length*10)`	Create a new column called SLMm

Combine the verbs `filter()`, `arrange()`, and `mutate()`:

```
> iris %>%
  filter(Species=="Virginica") %>%
  mutate(SLMm=Sepal.Length*10) %>%
  arrange(desc(SLMm))
```

Summarize

`summarize()` allows you to turn many observations into a single data point.

`> iris %>%` `  summarise(medianSL=median(Sepal.Length))`	Summarize to find the median sepal length

`> iris %>%`
`  filter(Species=="virginica") %>%`
`  summarise(medianSL=median(Sepal.Length))`
	Filter for virginica then summarize the median sepal length

You can also summarize multiple variables at once:

```
> iris %>%
  filter(Species=="virginica") %>%
  summarise(medianSL=median(Sepal.Length),
            maxSL=max(Sepal.Length))
```

`group_by()` allows you to summarize within groups instead of summarizing the entire dataset:

`> iris %>%`
`  group_by(Species) %>%`
`  summarise(medianSL=median(Sepal.Length),`
`  maxSL=max(Sepal.Length))`
	Find median and max sepal length of each species

`> iris %>%`
`  filter(Sepal.Length>6) %>%`
`  group_by(Species) %>%`
`  summarise(medianPL=median(Petal.Length),`
`  maxPL=max(Petal.Length))`
	Find median and max petal length of each species with sepal length > 6

ggplot2

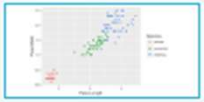
Scatter plot

Scatter plots allow you to compare two variables within your data. To do this with ggplot2, you use `geom_point()`

`> iris_small <- iris %>%`
`  filter(Sepal.Length > 5)`
`> ggplot(iris_small, aes(x=Petal.Length,`
`  y=Petal.Width)) +`
`  geom_point()`
	Compare petal width and length

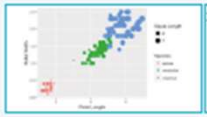
Additional Aesthetics

- Color



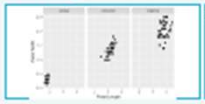
```
> ggplot(iris_small, aes(x=Petal.Length,
                         y=Petal.Width,
                         color=Species)) +
  geom_point()
```

- Size



```
> ggplot(iris_small, aes(x=Petal.Length,
                         y=Petal.Width,
                         color=Species,
                         size=Sepal.Length)) +
  geom_point()
```

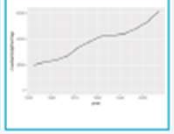
Faceting



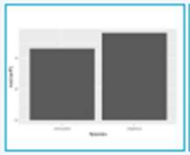
```
> ggplot(iris_small, aes(x=Petal.Length,
                         y=Petal.Width)) +
  geom_point()+
  facet_wrap(~Species)
```

Line Plots

```
> by_year <- gapminder %>%
  group_by(year) %>%
  summarize(medianGdpPerCap=median(gdpPercap))
> ggplot(by_year, aes(x=year,
                      y=medianGdpPerCap))+
  geom_line()+
  expand_limits(y=0)
```



Bar Plots

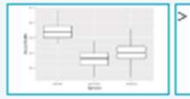


```
> by_species <- iris %>%
  filter(Sepal.Length>6) %>%
  group_by(Species) %>%
  summarise(medianPL=median(Petal.Length))
> ggplot(by_species, aes(x=Species,
                         y=medianPL)) +
  geom_col()
```

Histograms

```
> ggplot(iris_small, aes(x=Petal.Length))+
  geom_histogram()
```

Box Plots



```
> ggplot(iris_small, aes(x=Species,
                         y=Sepal.Width))+
  geom_boxplot()
```

DataCamp
Learn R for Data Science Interactively

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (None)

ISTA.R | Untitled1*

Source on Save | Run | Source

```r
data = rnorm(100)
p = ecdf(data)
plot(p, xlab='x', ylab='CDF', main='CDF of Data')

cumprob <- function(y) {
  fun <- function(y, x) length(y[y<x])/length(y)
  prob<-sapply(y, fun, y=y)
  data<- data.frame(value=unique(y[order(y)]), prob=unique(prob[order(prob)]))
}
cp<-cumprob(data)
plot(cp)
head(cp)

```

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (None)

Untitled1* | R data sets | Untitled2 | R data sets

```
Data sets in package 'datasets':

AirPassengers               Monthly Airline Passenger Numbers 1949-1960
BJsales                     Sales Data with Leading Indicator
BJsales.lead (BJsales)      Sales Data with Leading Indicator
BOD                         Biochemical Oxygen Demand
CO2                         Carbon Dioxide Uptake in Grass Plants
ChickWeight                 Weight versus age of chicks on different diets
DNase                       Elisa assay of DNase
EuStockMarkets              Daily Closing Prices of Major European Stock Indices, 1991-1998
Formaldehyde                Determination of Formaldehyde
HairEyeColor                Hair and Eye Color of Statistics Students
Harman23.cor                Harman Example 2.3
Harman74.cor                Harman Example 7.4
Indometh                    Pharmacokinetics of Indomethacin
InsectSprays                Effectiveness of Insect Sprays
JohnsonJohnson              Quarterly Earnings per Johnson & Johnson Share
LakeHuron                   Level of Lake Huron 1875-1972
LifeCycleSavings            Intercountry Life-Cycle Savings Data
Loblolly                    Growth of Loblolly pine trees
Nile                        Flow of the River Nile
Orange                      Growth of Orange Trees
OrchardSprays               Potency of Orchard Sprays
PlantGrowth                 Results from an Experiment on Plant Growth
Puromycin                   Reaction Velocity of an Enzymatic Reaction
Seatbelts                   Road Casualties in Great Britain 1969-84
Theoph                      Pharmacokinetics of Theophylline
Titanic                     Survival of passengers on the Titanic
ToothGrowth                 The Effect of Vitamin C on Tooth Growth in Guinea Pigs
UCBAdmissions               Student Admissions at UC Berkeley
UKDriverDeaths              Road Casualties in Great Britain 1969-84
UKgas                       UK Quarterly Gas Consumption
```

Console | Terminal | Jobs

```
> library(scatterplot3d)
> scatterplot3d(iris$Petal.Width, iris$Sepal.Length, iris$Sepal.Width)
>
```

Environment | History | Connections | Tutorial

Import Dataset | List

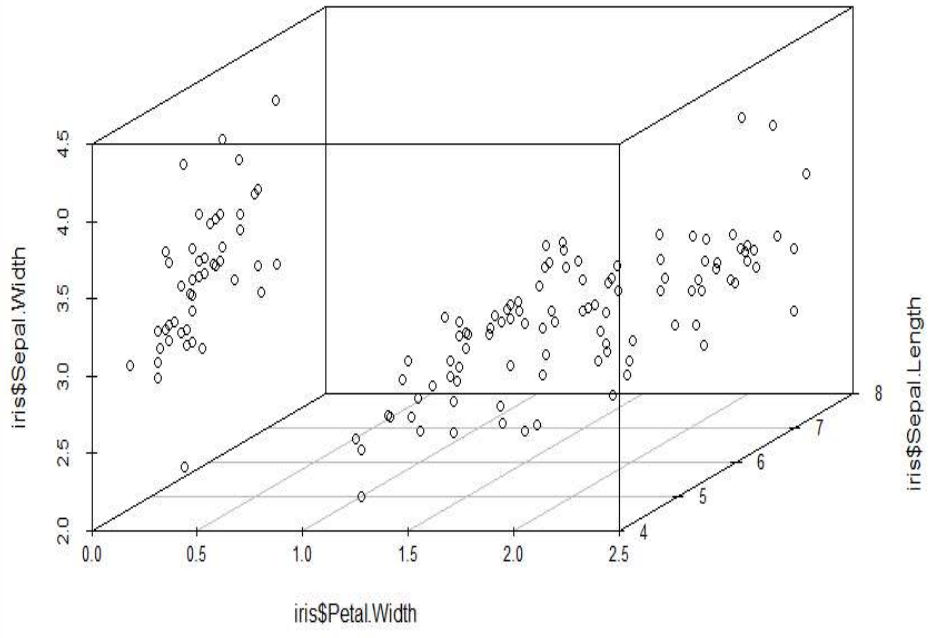
Global Environment

Data

iris	150 obs. of 5 variables

Files | Plots | Packages | Help | Viewer

Zoom | Export | Publish



R- code

- mlba-R-code/R at main · gedeck/mlba-R-code · GitHub
- R Archives - SAINSDATA.ID
- https://r4ds.hadley.nz/
- https://aditya-dahiya.github.io/RfDS2solutions/Chapter2.html
- https://mine-cetinkaya-rundel.github.io/r4ds-solutions/data-visualize.html

Pertemuan kedua
04-09-2024

1. Review syntax R dan RStudio
2. Pengantar : *exploring data with R*

https://r4ds.had.co.nz/introduction.html

Tipe dan struktur data di R

https://www.rdatamining.com/

Data Types and Structures

- Data types
  - Integer
  - Numeric
  - Character
  - Factor
  - Logical
- Data structures
  - Vector
  - Matrix
  - Data frame
  - List

Vector

```
## integer vector
x <- 1:10
print(x)
##  [1]  1  2  3  4  5  6  7  8  9 10

## numeric vector, generated randomly from a uniform distribution
y <- runif(5)
y
## [1] 0.5923091 0.6782441 0.5266127 0.1358263 0.7433572

## character vector
(z <- c("abc", "d", "ef", "g"))
## [1] "abc" "d"   "ef"  "g"
```

The values of features can be measured on several scales[1] ranging from simple labels all the way to numbers. The scales come in four levels.

Scale Name	Description	Operations	Statistics	R
Nominal	just a label (e.g., red, green)	$==, !=$	counts	`factor`
Ordinal	label with order (e.g., small, med., large)	$<, >$	median	`ordered`
Interval	difference between two values is meaningful (regular number)	$+, -$	mean, sd	`numeric`
Ratio	has a natural zero (e.g., count, distance)	$/, *$	percent	`numeric`

Matrix

```
## create a matrix with 4 rows, from a vector of 1:20
m <- matrix(1:20, nrow = 4, byrow = T)
m
##      [,1] [,2] [,3] [,4] [,5]
## [1,]    1    2    3    4    5
## [2,]    6    7    8    9   10
## [3,]   11   12   13   14   15
## [4,]   16   17   18   19   20

## matrix subtraction
m - diag(nrow = 4, ncol = 5)
##      [,1] [,2] [,3] [,4] [,5]
## [1,]    0    2    3    4    5
## [2,]    6    6    8    9   10
## [3,]   11   12   12   14   15
## [4,]   16   17   18   18   20
```

1. Create the vectors:

(a) $(1, 2, 3, \ldots, 19, 20)$

(b) $(20, 19, \ldots, 2, 1)$

(c) $(1, 2, 3, \ldots, 19, 20, 19, 18, \ldots, 2, 1)$

(d) $(4, 6, 3)$ and assign it to the name `tmp`.

For parts (e), (f) and (g) look at the help for the function `rep`.

(e) $(4, 6, 3, 4, 6, 3, \ldots, 4, 6, 3)$ where there are 10 occurrences of 4.

(f) $(4, 6, 3, 4, 6, 3, \ldots, 4, 6, 3, 4)$ where there are 11 occurrences of 4, 10 occurrences of 6 and 10 occurrences of 3.

(g) $(4, 4, \ldots, 4, 6, 6, \ldots, 6, 3, 3, \ldots, 3)$ where there are 10 occurrences of 4, 20 occurrences of 6 and 30 occurrences of 3.

2. Create a vector of the values of $e^x \cos(x)$ at $x = 3, 3.1, 3.2, \ldots, 6$.

3. Create the following vectors:

(a) $(0.1^3 0.2^1, 0.1^6 0.2^4, \ldots, 0.1^{36} 0.2^{34})$

(b) $\left(2, \frac{2^2}{2}, \frac{2^3}{3}, \ldots, \frac{2^{25}}{25}\right)$

4. Calculate the following:

(a) $\sum_{i=10}^{100} \left(i^3 + 4i^2\right)$.

(b) $\sum_{i=1}^{25} \left(\frac{2^i}{i} + \frac{3^i}{i^2}\right)$

1. Suppose

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 3 \\ 5 & 2 & 6 \\ -2 & -1 & -3 \end{bmatrix}$$

(a) Check that $\mathbf{A}^3 = \mathbf{0}$ where $\mathbf{0}$ is a 3×3 matrix with every entry equal to 0.

(b) Replace the third column of $\mathbf{A}$ by the sum of the second and third columns.

2. Create the following matrix $\mathbf{B}$ with 15 rows:

$$\mathbf{B} = \begin{bmatrix} 10 & -10 & 10 \\ 10 & -10 & 10 \\ \cdots & \cdots & \cdots \\ 10 & -10 & 10 \end{bmatrix}$$

Calculate the 3×3 matrix $\mathbf{B}^T\mathbf{B}$. (Look at the help for `crossprod`.)

Solve the following system of linear equations in five unknowns

$$x_1 + 2x_2 + 3x_3 + 4x_4 + 5x_5 = 7$$

$$2x_1 + x_2 + 2x_3 + 3x_4 + 4x_5 = -1$$

$$3x_1 + 2x_2 + x_3 + 2x_4 + 3x_5 = -3$$

$$4x_1 + 3x_2 + 2x_3 + x_4 + 2x_5 = 5$$

$$5x_1 + 4x_2 + 3x_3 + 2x_4 + x_5 = 17$$

by considering an appropriate matrix equation $\mathbf{Ax} = \mathbf{y}$.

Anscombe's Data

- The following four data sets comprise the Anscombe's Quartet; all four sets of data have identical simple summary statistics.

G. E. M. Anscombe
FBA

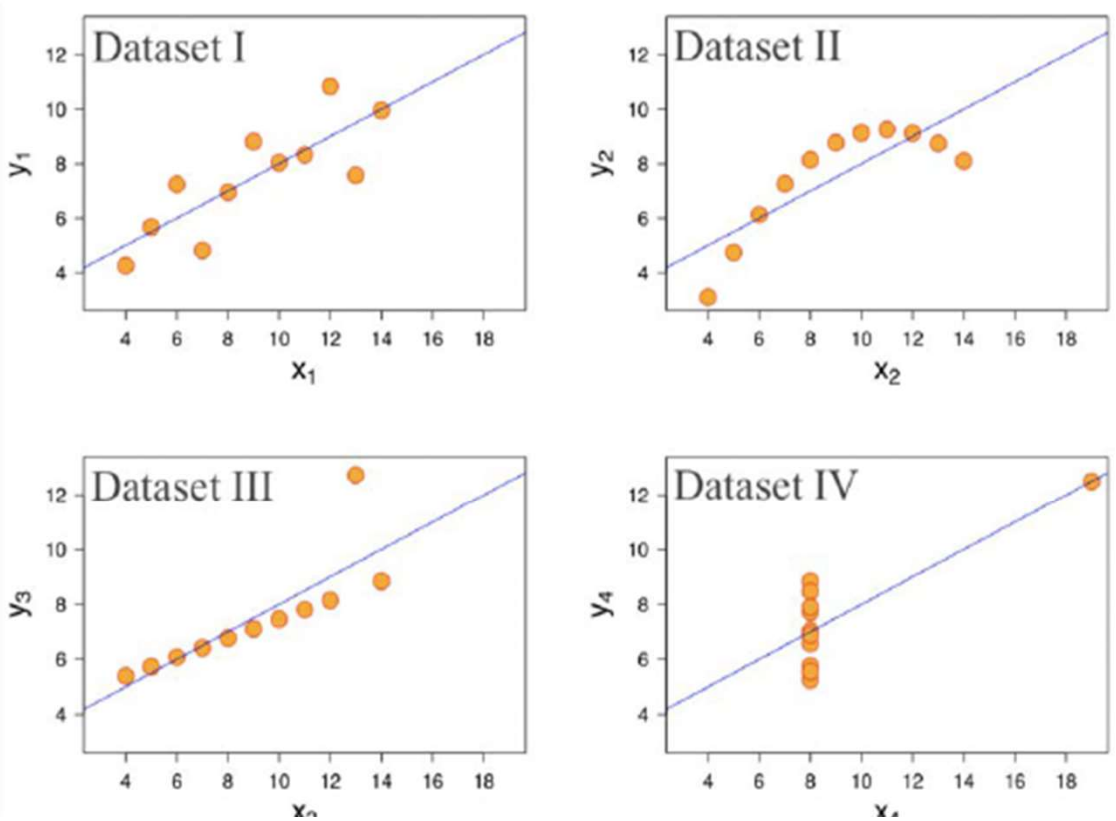
Anscombe as a young woman

Born	Gertrude Elizabeth Margaret Anscombe 18 March 1919 Limerick, Ireland
Died	5 January 2001 (aged 81) Cambridge, England

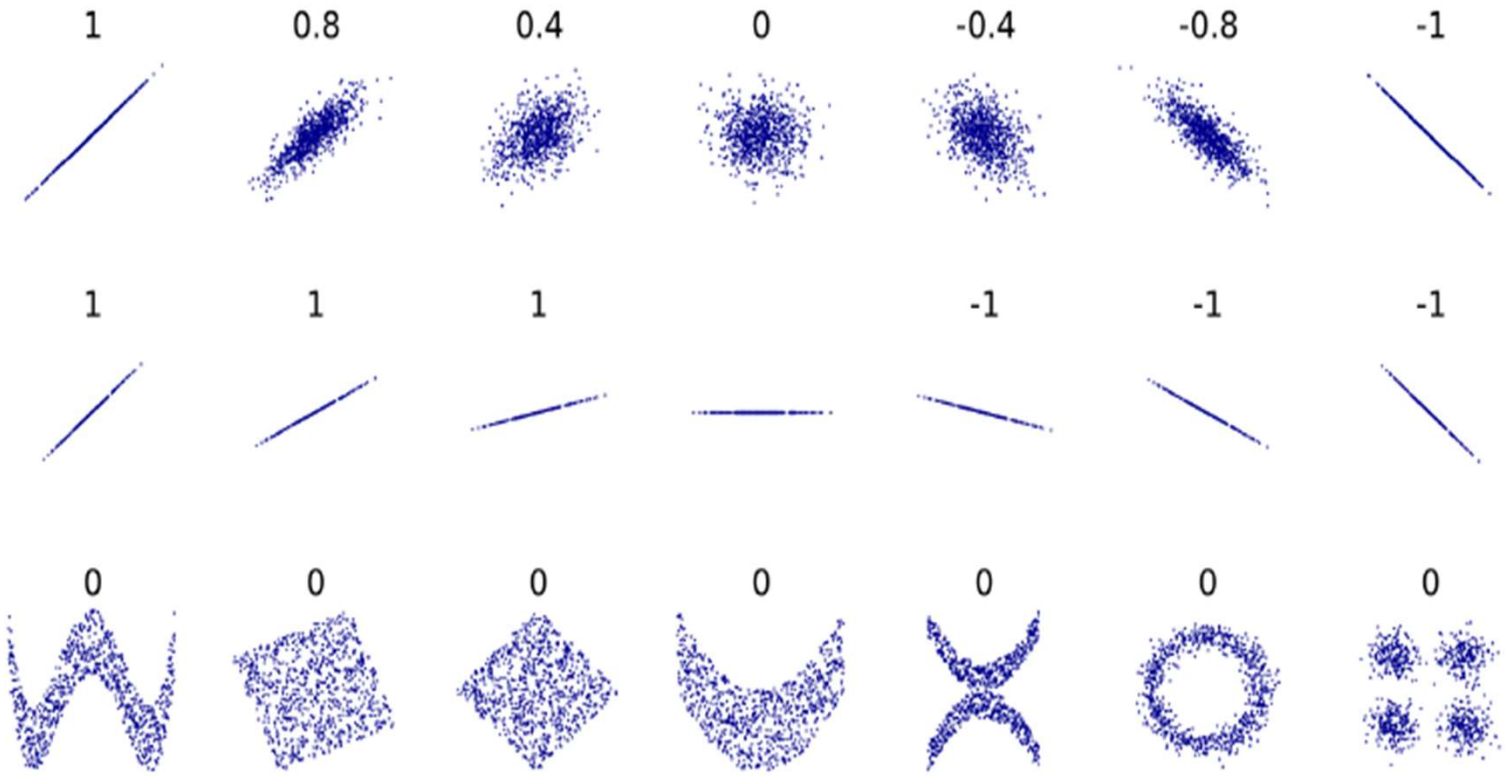
	Dataset I		Dataset II		Dataset III		Dataset IV	
	x	y	x	y	x	y	x	y
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.95	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.74	13	12.74	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.1	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.1	4	5.39	19	12.5
	12	10.84	12	9.13	12	8.15	8	5.56
	7	4.82	7	7.26	7	6.42	8	7.91
	5	5.68	5	4.74	5	5.73	8	6.89
Sum:	99.00	82.51	99.00	82.51	99.00	82.51	99.00	82.51
Avg:	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
Std:	3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03

Anscombe's Data (cont.)

- Summary statistics clearly don't tell the story of how they differ. But a picture can be worth a thousand words:



No correlation does not mean no relationship:



Correlation

The Greek symbol ρ is often used to denote correlation which we do not know and try to infer, and r the correlation between a sample which contains information on ρ

- $|\rho| \leq 1$: correlation is always between -1 and 1
- $\rho = 1$: unit increase in x predicts unit increase in y on average
- $\rho = -1$: unit increase in x predicts unit decrease in y on average

PERTEMUAN-KETIGA
11-09-2024

- https://michael.hahsler.net/SMU/DS_Workshop_Intro_R/slides/3b_ggplot.html
- https://rgraphgallery.blogspot.com/
- https://plot.ly/r/#basic-charts

Data Exploration Techniques

- Descriptive Statistics: Summarizing data
  - Mean, median, mode, variance, standard deviation
- Data Visualization: Graphical representation of data
  - Charts, graphs, plots
- Data Cleaning: Handling missing or noisy data
  - Imputation, outlier detection
- Correlation Analysis:
  -Identifying relationships btw var.
- Dimensionality Reduction: Reducing #variables
  - PCA, LDA

Chapter 2 Data | An R Companion for Introduction to Data Mining (mhahsler.github.io)

Intro to R - 4. Base R Plots

OIT/SMU Libraries Data Science Workshop Series

Michael Hahsler

OIT, SMU

World Changers Shaped Here

SMU

Chart Suggestions—A Thought-Starter

What would you like to show?

Comparison

Among Items

One Variable per Item

Two Variables per Item

Variable Width Column Chart

Many Categories

Table or Table with Embedded Charts

Few Categories

Many Items

Bar Chart

Few Items

Column Chart

Over Time

Many Periods

Cyclical Data

Circular Area Chart

Non-Cyclical Data

Line Chart

Few Periods

Single or Few Categories

Column Chart

Many Categories

Line Chart

Relationship

Two Variables

Scatter Chart

Three Variables

Bubble Chart

Distribution

Single Variable

Few Data Points

Column Histogram

Many Data Points

Line Histogram

Two Variables

Scatter Chart

Three Variables

3D Area Chart

Composition

Changing Over Time

Few Periods

Only Relative Differences Matter

Stacked 100% Column Chart

Relative and Absolute Differences Matter

Stacked Column Chart

Many Periods

Only Relative Differences Matter

Stacked 100% Area Chart

Relative and Absolute Differences Matter

Stacked Area Chart

Static

Simple Share of Total

Pie Chart

Accumulation or Subtraction to Total

Waterfall Chart

Components of Components

Stacked 100% Column Chart with Subcomponents

PENGAYAAN

- https://r4ds.hadley.nz/
- https://aditya-dahiya.github.io/RfDS2solutions/Chapter2.html
- https://mine-cetinkaya-rundel.github.io/r4ds-solutions/data-visualize.html

Graph Gallery:

A collection

2d (1) 3 vartiable plots (5) 3D plots (8) arch (1) area (1) association plot (4) bar (1) barchart (13) bean plot (1) beeswarm (1) binormial (1) biplot (1) box-percentile (2) box-whisker plot (1) boxplot (10) bubble plot (5) calendar (1) categorical data (6) centepede plot (1) circle (2) circular (1) cluster (4) color (2) colour (1) combination plot (10) countur (1) cross bar (1) cumulative (1) curve (3) dendogram (3) density (13) diagram (2) distribution (9) ditribution (1) dot plot (1) double axis (1) ellipse (2) error bar (6) factor plot (3) fluctutation diagram (1) google (1) grid plot (1) heatmap (20) hexabin plot (1) histogram (11) hive (1) kernel density (4) ladder plot (2) large data points (4) level plot (1) line plot (3) line range (1) manhattan plot (1) map (13) mosaic plot (1) normal (2) notched (1) parallel plot (1) pedigree plot (1) phylogentic tree (1) piechart (3) points (2) polar (1) Q-Q plot (1) raster (2) regression line (3) ribbon plot (1) rootogram (1) rugs (2) scale plot (1) scenes (1) shaded (1) spatial plot (2) sphere (1) spike histogram (1) Spine plot (1) stacked bar (1) Sunflower (1) ternary plot (1) text only (1) timeseries (6) trellis plot (8) two axis (1) vinn diagram (1) voilin plot (2) wireframe plot (1) xy barplot (4) xy line (10) xy points (25)

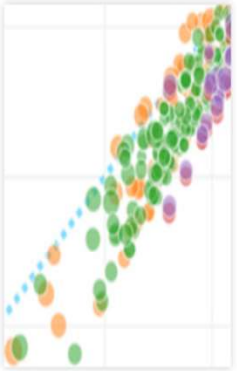
GitHub
community.plotly.com

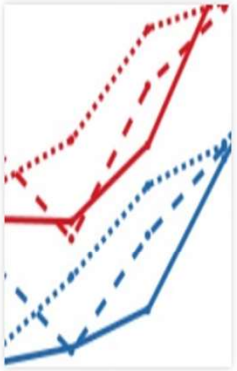
▼ Examples

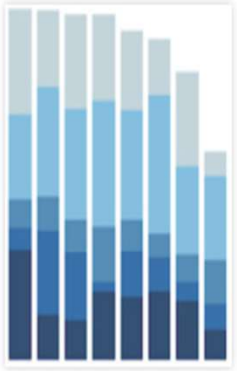
Fundamentals
Basic Charts
Statistical Charts
Scientific Charts
Financial Charts
Maps
AI and ML
3D Charts
Subplots
Animations
Advanced

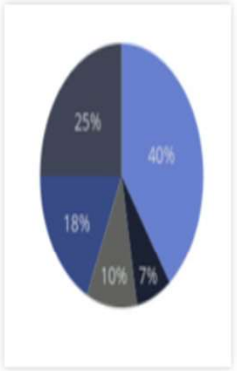
Basic Charts

More Basic Charts »


Scatter and Line Plots


Line Plots


Bar Charts

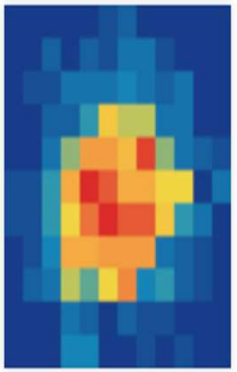


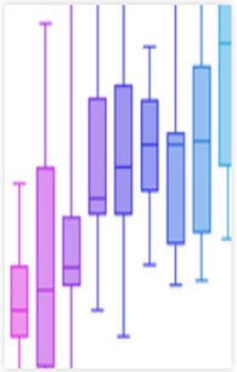
Pie Charts

Bubble Charts

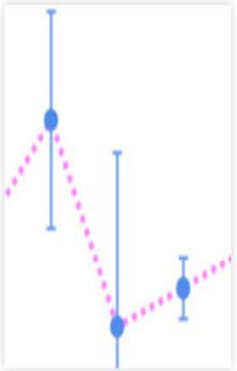
Statistical Charts

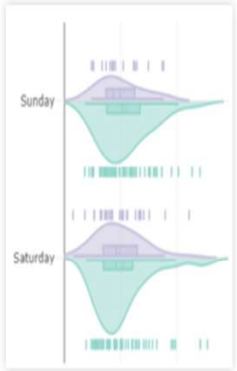
More Statistical Charts »


2D Histograms


Box Plots


Histograms


Error Bars



Violin Plots

Scientific Charts

More Scientific Charts »