

Pengenalan data sains dasar dan R

(Pelatihan data sains menggunakan R dan Gephi)

Ujang Fahmi

Pelajaran ke-1



Salam kenal dan selamat datang.

Semoga kita semua bisa saling berbagi pengalaman dan pengetahuan. Saya adalah Ujang Fahmi, Co-founder dan mentor Sadasa Academy.

Jika anda berada dan sedang membaca tutorial ini, maka kemungkinan anda adalah orang yang sedang ingin belajar data sains, atau mungkin ditugaskan untuk mempelajari R oleh institusi atau organisasi anda. Sama seperti saya dulu, dimana tanpa latar belakang engineering saya didiharuskan untuk belajar R, demi menyelesaikan tugas akhir dan akhirnya jadilah seperti saya sekarang ini.

Satu hal yang pasti, ini adalah langkah pertama dari banyak langkah yang harus dilalui, entah melalui lembaga resmi atau belajar secara mandiri. Jadi selamat belajar!!!

Ujang Fahmi,
Yogyakarta, 2021-09-18

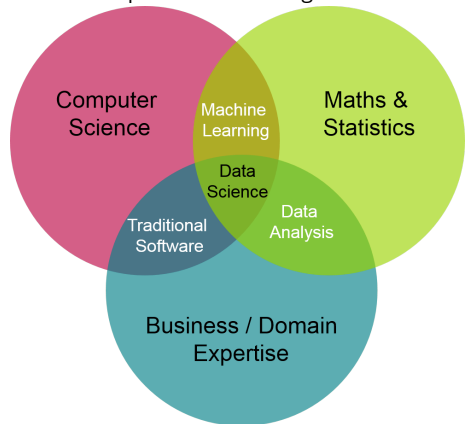
*Materi yang disampaikan disimpan dan dokumentasikan **disini***

Apa itu data sains?

Data Sains merupakan sebuah bidang inter atau bahkan multidisiplin yang menggunakan metode-metode saintifik, proses, algoritma, dan sistem untuk mengekstrak pengetahuan dari data yang terstruktur, tidak terstruktur dan bercabang.

- Memiliki keterampilan koding
- Mengetahui dan menguasai matematika dan statistik
- Memiliki domain knowledge spesifik yang kuat

Keterampilan data seorang data saintis



Social data sains?

Social data science is a new discipline combining the social sciences and computer science in which the analysis of big data is linked to social scientific theory and analysis.

<https://www.ox.ac.uk/>

[Home](#) › [Admissions](#) › [Graduate](#) › [Courses](#) › MSc in Social Data Science

MSc in Social Data Science

ABOUT

ENTRY REQUIREMENTS

RESOURCES

FUNDING AND COSTS

COLLEGE PREFERENCE

HOW TO APPLY

About the course

The multidisciplinary MSc in Social Data Science provides the social and technical expertise needed to collect, critique, and analyse unstructured heterogeneous data about human behaviour, thereby informing our understanding of the social world.

Figure 1: Program Social Data Science di Oxford

Alat-alat yang biasa digunakan

Umum

Bahasa pemrograman yang biasa digunakan untuk mengolah data

- R - statistical programming language
- Python - general programming language
- Julia - programming language untuk big data

Spesifik

Perangkat lunak yang biasa digunakan untuk mengolah data dengan tujuan spesifik

- Gephi - Network Analysis
- Nodexl - Network Analysis
- Orange - Data Mining

Harus menggunakan yang mana?

- Pilih yang sudah banyak digunakan oleh orang lain
- Pilih yang memiliki komunitas yang kuat baik di dunia maupun di negara kita
- Pilih yang sesuai dengan kebutuhan
- Pelajari semuanya, pilih salah satu untuk dikuasai

Kenapa menggunakan R dan Rstudio

- R adalah bahasa pemrogramannya
- R Studio adalah perangkat lunak yang menjadi interpreter R
- R Studio adalah merupakan salah satu lembaga yang berkontribusi besar dalam perkembangan R
- Apakah R bisa dijalankan di IDE lain? Ya, bisa. Misalnya di VsCode


	RStudio Desktop <small>Open Source License</small>	RStudio Desktop Pro <small>Commercial License</small>	RStudio Server <small>Open Source License</small>	RStudio Workbench <small>Commercial License</small>
	Free	\$995 <small>/year</small>	Free	\$4,975 <small>/year (5 Named Users)</small>
	DOWNLOAD <small>Learn more</small>	BUY <small>Learn more</small>	DOWNLOAD <small>Learn more</small>	BUY <small>Evaluation Learn more</small>
Integrated Tools for R	✓	✓	✓	✓
Priority Support		✓		✓
Access via Web Browser			✓	✓
RStudio Professional Drivers		✓		✓
Connect to RStudio Workbench  remotely		✓		

Figure 2: Pilihan R Studio yang bisa didapat

Mendapatkan R dan Rstudio

- R bisa didapatkan di <https://www.r-project.org/>
- R Studio bisa didapatkan di <https://www.rstudio.com/products/rstudio/download/>



Figure 3: Pilihan R Studio yang bisa didapat

Bisa apa saja dengan R dan Rstudio

Saat ini dengan menggunakan R dan Rstudio kita hampir bisa melakukan semua kegiatan yang berkaitan dengan pengolahan data, misalnya:

- Mendapatkan data
- Melakukan manipulasi atau pra-pemerosesan
- Membuat analisis dengan statistik, Macine Learning dan Deep Learning
- Membuat laporan hasil pengolahan data
- Membuat dashboard hasil pengolahan data

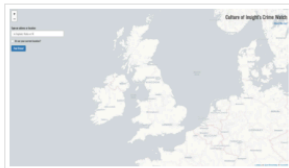
Gallery Shiny Dashboard

Government / Public sector

Mostly open data



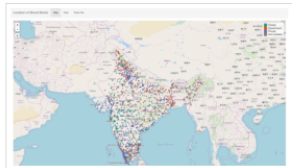
Voronovs - Understanding voters' profile in Brazilian elections



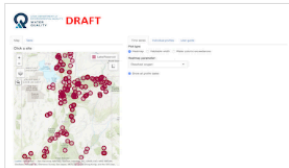
Crime Watch



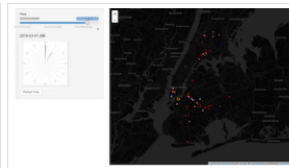
Pasture Potential Tool for improving dairy farm profitability and environmental impact



Locating Blood Banks in India



Utah Lake Water Quality Profile Dashboard



Animated NYC metro traffic

Figure 4: Shiny Dashboard salah satu output R

Menggunakan R dan Rstudio

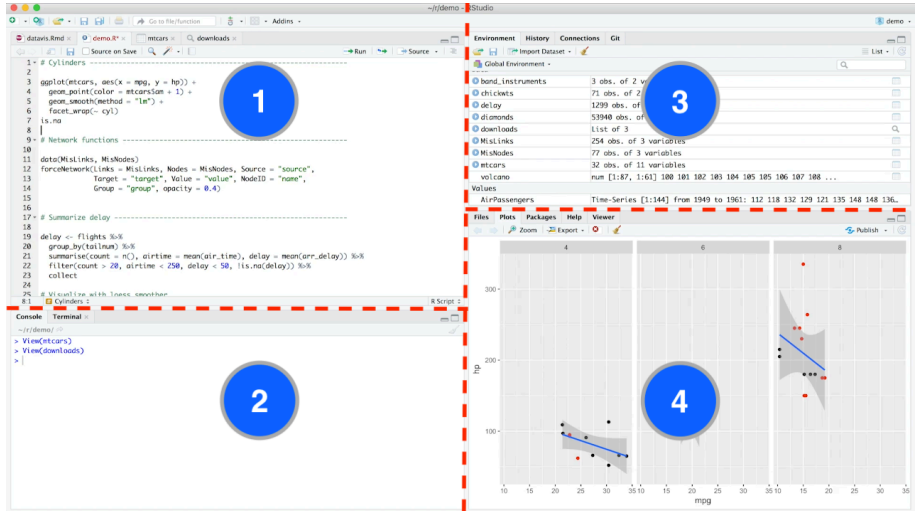


Figure 5: Tampilan RStudio

Membuat Proyek

Project merupakan sebuah folder seperti yang sudah sering kita buat. Di R folder tersebut difungsikan untuk menyimpan segala sesuatu yang kita buat di Rstudio secara otomatis kedalam folder tersebut. Keuntungan yang didapatkan adalah kita memiliki fokus folder yang sedang menjadi tempat kerja kita.

1. File
2. New Project
3. New Directory
4. New Project
5. Directory Name
6. Create Project

Kenapa?

Karena di R kita **hanya akan bisa mengolah data yang bisa diimpor kedalam R** atau benar-benar eksis dalam lingkungan R.

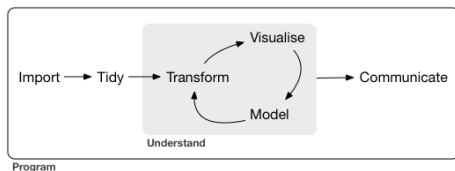


Figure 6: Proses Pengolahan data di R

Membuat dan menyimpan skrip/kumpulan perintah

Skrip adalah kumpulan perintah yang kita buat untuk menyelesaikan atau melakukan sesuatu dengan bahasa tertentu. Di R kita bisa membuat beberapa skrip sesuai dengan peruntukannya.

1. R Script (.R) umumnya digunakan untuk melakukan pengolahan data
2. R Markdown (.Rmd) umumnya digunakan untuk membuat laporan
3. Script-script lain yang biasa digunakan di pemerogaman seperti C, C++. CSS, dan lain sejenisnya

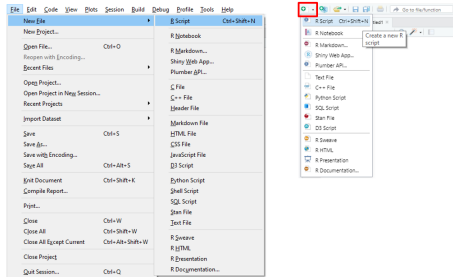


Figure 7: Menu untuk membuat skrip di Rstudio

Aturan menulis skrip

UMUM

1. Nama objek tidak boleh menggunakan spasi atau diawali dengan angka (1namaobjek, nama objek)
2. Setiap perintah yang didahului dengan tanda pagar (#) dibaca sebagai komentar
3. Komentar tidak akan dibaca sebagai perintah dan berfungsi untuk memberikan keterangan tambahan pada penulis atau yang menggunakan skrip
4. Objek di R bisa dibuat dengan menggunakan assignment. Misalnya `data1 = 12 * 19827` atau `data1 <-12 * 19827`

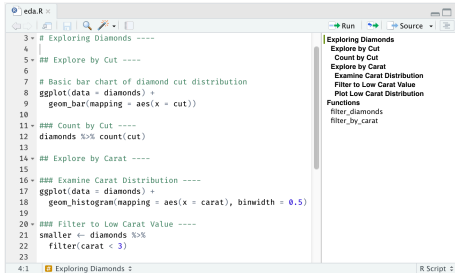
Khusus

Operator	Deskripsi
<	kurang dari
<=	kurang dari atau sama dengan
>	lebih dari
>=	lebih dari atau sama dengan
==	sama persis
!=	tidak sama persis
!x	bukan x
x & y	x AND Y
isTRUE(x)	test if X is TRUE

Figure 8: Operator/tanda yang digunakan di R

Jenis skrip di Rstudio

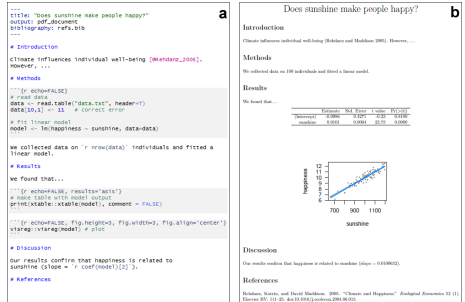
SKRIP R



```
3 # Exploring Diamonds ----
4
5 ## Explore by Cut ----
6
7 # Basic bar chart of diamond cut distribution
8 ggplot(data = diamonds) +
9   geom_bar(mapping = aes(x = cut))
10
11 ### Count by Cut ----
12 diamonds %>% count(cut)
13
14 ## Explore by Carat ----
15
16 ### Examine Carat Distribution ----
17 ggplot(data = diamonds) +
18   geom_histogram(mapping = aes(x = carat), binwidth = 0.5)
19
20 ### Filter to Low Carat Value ----
21 smaller <- diamonds %>%
22   filter(carat < 3)
23
4.1 Exploring Diamonds
```

Figure 9: Script R

R MARKDOWN



a

```
---
title: "Does sunshine make people happy?"
output: pdf_document
bibliography: refs.bib
---

# Introduction
Climate influences individual well-being [Mehdarez_2008]. However, ...
However, ...

# Methods
```{r echo=FALSE}
read data
data <- read.table("data.txt", header=T)
data[10,1] <- 11 # correct error
fit linear model
model <- lm(happiness ~ sunshine, data=data)
...

we collected data on "r nrow(data)" individuals and fitted a
linear model.

Results
We found that...
```{r echo=FALSE, results="asis"}
# make table with model output
print(xtable::xtable(model), comment = FALSE)
...

```{r echo=FALSE, fig.height=3, fig.width=3, fig.align="center"}
vitarreg::vitarreg(model) # plot
...

Discussion
Our results confirm that happiness is related to
sunshine (slope = r.coef(model)[2]).

References
```

**b**

Does sunshine make people happy?

Introduction

Climate influences individual well-being [Mehdarez and Mollison 2008]. However, ...


Methods

We collected data on 100 individuals and fitted a linear model.

Results

We found that...

	Estimate	Std. Error	t-value	P> t
(Intercept)	-0.0004	0.0201	-0.02	0.9818
sunshine	0.001	0.0001	20.75	0.0000



happiness

sunshine

Discussion

Our results confirm that happiness is related to sunshine (slope = 0.0010002).

References

Mehdarez, Katrien, and David Mollison. 2008. "Climate and Happiness." *Ecological Economics* 52 (1): Elsevier B.V. 111–25. doi:10.1016/j.ecolecon.2004.06.005.

Figure 10: R markdown

# Apa Library dan Package?



Figure 11: Library dianalogikan sebagai perpustakaan

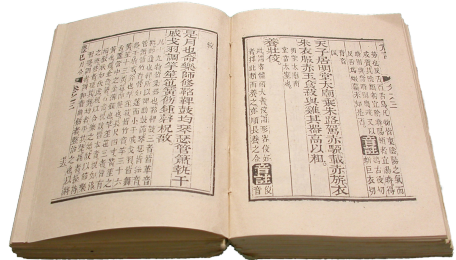


Figure 12: Package dianalogikan sebagai buku



# Mendapatkan Library

- Library atau package hanya perlu diinstall sekali
- Library dan package perlu dipanggil kembali dalam setiap sesi baru R
- Sebuah sesi baru di R dimulai dari saat membuka atau merstart R hingga menutup atau merestart kembali
- Semua Library dan package dibuat secara terbuka (seperti wikipedia)
- Library dan package dapat diinstall dengan menggunakan sintaks `install.packages(namaPackage)`
- Library dan package sama-sama dipanggil menggunakan sintak `library()`
- Semua Library dan package yang dikelola di kurasi oleh R dapat dilihat di:  
<https://cran.r-project.org/>

# Menggunakan fungsi dalam package

Fungsi dalam r terdiri dari nama\_fungsi(argumen1, argumen2, dst). Bisanya dibuat dengan cara sebagai berikut.

```
pembuatan fungsi
fungsi_modulo <- function(argumen1){
 hasil = argumen1%%2
 return(hasil)
}
```

```
penggunaan fungsi
fungsi_modulo(argumen1 = 7)
```

Setiap fungsi dari dalam pacakge juga memiliki format seperti, sehingga ketika kita akan menggunakannya kita perlu tahu terlebih dahulu argumen yang dibutuhkannya.

# Your Turn 1!

1. Buatlah sebuah project di Rstudio!
2. Buatlah sebuah skrip .R di project yang telah dibuat!
3. Tulislah perintah untuk menginstall package tidyverse, tidytext, dan igraph!

## Operator %>% (Kemudian)

Operator persen lebih dari persen disebut pipe dan bisa dibaca kemudian. Contohnya adalah sebagai berikut.

```
library(tidyverse)

daftar_siswa %>%
 select(1:4) %>%
 filter(kolom_1 >= 10)
```

Skrip diatas berarti kita menggunakan data, kemudian kita hanya memilih kolom 1 sampai 4 saja, dari daftar siswa tersebut kemudian kita hanya memilih yang nilai di kolom\_1 nya lebih dari atau sama dengan 10.

# Fungsi paste

Terdapat dua fungsi paste, yaitu `paste()` dan `paste0()`. Keduanya memiliki fungsi utama yang sama yaitu untuk meletakkan sebuah objek berdampingan dengan objek lainnya.

*Mirip dengan fungsi paste yang mungkin telah sering kita gunakan saat menggunakan komputer*

Contoh:

```
teks1 <- "aku adalah"
teks2 <- "raja rimba"

teks12 <- paste(teks1, teks2, sep = " ")
teks12

teks0 <- paste0(teks1, teks2)
teks0

teks <- paste0(teks1, " ", teks2)
teks
```

## Fungsi if dan else

Fungsi ini digunakan untuk memilih output sesuai dengan kondisi yang sudah ditentukan. Sering dinyatakan dengan JIKA ... MAKA ....  
Kondisinya Nilai Tim A = 10, sementara Nilai Tim B = 8, skrip di R bisa dibaca JIKA nilai tim A lebih dari tim B MAKA cetak tulisan TIM A adalah juaranya jika tidak MAKA cetak TIM B adalah juaranya

Contoh:

```
Tim_A = 10
Tim_B = 8

if (Tim_A > Tim_B) {
 print("Tim A adalah juaranya")
} else if(Tim_A == Tim_B){
 print("Seri")
} else {
 print("Tim B adalah juaranya")
}
```

# Fungsi for-Loop

Fungsi for-Loop juga disebut perulangan. Digunakan untuk melakukan hal-hal yang sama secara berulang sesuai dengan batas yang ditentukan.

Fungsi ini akan terasa manfaatnya jika kita memiliki data yang cukup banyak dan harus melakukan sebuah hal yang sama pada setiap observasi yang dimiliki.

Contoh:

```
for (value in vector){
 statements
}

vektor <- c(1:5)
loop
for(i in vektor){
 print(i)
}
```

# Fungsi-fungsi untuk melihat data (`str()`, `class()` dan `summary()`)

Fungsi-fungsi di atas digunakan untuk melihat struktur, kelas, dan rangkuman data. `str()` bisa dibaca struktur, `class()` atau kelas digunakan untuk mengetahui jenis data yang ada dalam data, apakah ia data frame, list, matrix atau lainnya. Sementara `summary` akan lebih banyak digunakan saat melakukan eksplorasi data.

```
library(tidyverse)

df1 = mtcars

str(df1)
class(df1)
summary(df1)
```



## Your Turn 2!

1. Data  $A = 6$  dan  $B = 187$ , jika  $A = \text{genap}$ , dan  $B = \text{Ganjil}$ , maka cetak "Ganjil Genap", jika  $A$  dan  $B$  genap, maka cetak "Genap", JIKA  $A$  dan  $B$  ganjil semua maka cetak Ganjil, JIKA  $A$  ganjil dan  $B$  genap, maka cetak "Ganjil genap", JIKA tidak memenuhi kondisi sebelumnya, maka cetak "genap".
2.  $A$  adalah sebuah distribusi angka antara 10 sampai dengan 1000 sebanyak 100, buatlah fungsi loop dimana setiap menemukan angka genap console mencetak "Genap" sementara jika menemukan angka ganjil tidak mencetak apapun.

# Vectors

Vector merupakan sebuah tipe data gabungan yang berisi beberapa elemen atomic berjenis sama. Untuk membuat sebuah vector, digunakan perintah kombinasi `c()`. Untuk mengakses vector bisa menggunakan tanda `namaVector[]`. Misalnya, `a[1]`, berarti mengakses vektor ke-1 dari `a`.

Contoh:

```
a <- c(1,2,5.3,6,-2,4)
b <- c("one","two","three")
c <- c(TRUE,TRUE,FALSE,TRUE,FALSE)
```

# Data Frame

Data frame juga disebut data flat atau rata, yaitu sebuah data yang setiap barisnya memiliki kolom yang sama, dan setiap kolomnya memiliki baris yang sama. Untuk mengakses data frame bisa menggunakan tanda dollar \$. Misalnya `datasiswa$name`, berarti mengakses kolom `name` dari data `datasiswa`. Sementara jika `datasiswa$name[3]`, berarti mengakses baris ke-3 dari kolom `name` dalam `datasiswa`.

Contoh:

```
id <- c(1,2,3,4)
name <- c("tom", "jerry",
 "dora", "emon")
score <- c(85.4,78.3,
 88.9,90)

membuat data frame dari kolom
datasiswa <- data.frame(id,
 name,
 score)
```

# Lists

List merupakan tipe data gabungan yang mirip dengan Vector, namun list memungkinkan penggunaan elemen-elemen dengan tipe data dan jenis variabel berbeda. Untuk mengakses list kita bisa menggunakan `namalist[[nama/indeks]]`. Misalnya `all_list[[1]]`, berarti mengakses list ke 1 dari `all_list`. Sementara jika untuk mengakses elemen spesifik dari sebuah indeks dalam list kita bisa menggunakan skema `namalist[[nama/indeks]][indeks]` dan seterusnya.

Contoh:

```
w <- list(name="Fred",
 age=25,
 height=159.7)
x <- list("saya",5.4,1,FALSE)

all_list = list(datasiswa, w, x)
```

## Your Turn 3!

1. Buatlah dua data frame. Data frame pertama memiliki 3 kolom dan 5 baris dan beri nama df1, lalu buat data frame kedua, dengan 5 kolom dan 4 baris.
2. Buatlah list dengan nama list\_df dari dua data frame yang sudah dibuat.

# Table of Contents

Pengenalan Data Sains

R dan Rstudio

Menggunakan R dan Rstudio

Library dan Package

Fungsi-fungsi Dasar

Jenis-jenis data yang umum digunakan