

Belajar Statistika dengan R

Krisna Gupta dan Donny Pasaribu

2020-08-10

Contents

1	Pendahuluan	5
1.1	Tentang Modul ini	5
1.2	Pengetahuan yang diperlukan	6
1.3	Tentang R	6
2	Memulai dengan R	11
3	Manipulasi Data	13
4	Visualisasi Data	15
4.1	Tentang Visualisasi Data	15
4.2	Tabel	17
4.3	Grafik	17
4.4	Latihan	18
5	Mean, Median, dan Modus	21
6	Tes Hipotesis	23
7	Analisis Regresi	25

Chapter 1

Pendahuluan

This is a guide to learn beginner statistics using R, written in Indonesian language. This is an open source book written using bookdown R package. Everyone feel free to use and share it with anyone, but please credit the authors. Please address your queries to krisna.gupta@anu.edu.au

1.1 Tentang Modul ini

1.1.1 Modul apa ini?

Modul ini merupakan panduan mahasiswa dalam mengaplikasikan teori tentang statistika dan analisis data ke dalam standar praktik di dunia usaha. Khususnya, modul ini bertujuan untuk membantu mahasiswa menggunakan R, sebuah bahasa pemrograman *open source* yang banyak digunakan oleh analis data untuk menganalisis data menggunakan statistik, dan melakukan visualisasi data. Lebih lengkap tentang R dapat ditemukan di <https://www.r-project.org/>

Modul ini dimaksudkan untuk membantu mahasiswa Indonesia belajar statistik dengan menggunakan R. Saat ini, penggunaan bahasa pemrograman untuk analisis data semakin menjamur, sementara itu *talent* di Indonesia yang sanggup menggunakannya masih sangat terbatas. Untuk itu modul ini dibuat untuk membantu siapapun yang memiliki kendala biaya maupun keterbatasan bahasa untuk belajar menggunakan R.

Modul ini adalah *open source*, dengan kata lain, modul ini dapat digunakan dan diunduh secara bebas oleh siapapun. Kami melarang penggunaan modul ini untuk siapapun yang bertujuan memperjual-belikan konten pada modul ini tanpa izin penulis.

1.1.2 Tentang penulis

Krisna Gupta adalah dosen Politeknik APP Jakarta, sebuah kampus di bawah naungan Kementerian Perindustrian. Saat ini tengah menempuh pendidikan doktoral di bidang ekonomi di Australian National University (ANU). Kontak dapat dilihat di blog pribadinya di <https://imedkrisna.github.io/about>.

1.2 Pengetahuan yang diperlukan

Modul ini mengasumsikan bahwa mahasiswa telah mampu menggunakan komputer dan mengakses internet. Mengetik dan *point-and-click windows interface* dianggap tidak perlu diajarkan. Modul ini juga mengasumsikan pengetahuan dasar mengenai statistik. Mahasiswa dianggap telah mengerti teori inferensi statistik (mean, median, dan modus), sampling, probabilitas, maupun regresi linear sederhana. Kemampuan menggunakan *spreadsheet* seperti Microsoft Excel atau Google Sheet merupakan pengetahuan yang sangat membantu, meskipun tidak diperlukan.

1.3 Tentang R

1.3.1 Kenapa R?

Ada beberapa aplikasi yang dapat digunakan untuk menganalisis data, seperti Microsoft Excel, EViews, SPSS, dan lain sebagainya. Beberapa alasan mengapa modul ini ditulis dengan menggunakan R:

1.3.1.1 Kebutuhan akan logika menulis kode

Beberapa aplikasi yang ada di pasaran saat ini seperti Microsoft Excel ataupun Google Sheet merupakan aplikasi yang cukup mudah dipelajari. Namun demikian, kebutuhan dunia usaha untuk tenaga kerja yang dapat menulis kode semakin meningkat. Menulis kode sendiri memiliki berbagai manfaat. Pertama, logika menulis kode akan mempermudah mahasiswa mempelajari bahasa lain yang lebih mendasar seperti Python. Bahasa-bahasa ini di masa depan dapat menjadi pilihan jika mahasiswa berniat meniti karir lebih khusus di dunia analisis data, di mana kebutuhan sumber daya manusia di sini masih sangat signifikan.

1.3.1.2 R termasuk cukup mudah dipelajari

R adalah bahasa yang memang ditulis untuk digunakan dalam analisis statistik. R adalah pintu masuk yang cukup baik dari statistisi menuju kemampuan lain

yang berhubungan dengan *coding* seperti *text mining* dan lain sebagainya.

Selain modul ini, ada sangat banyak sumber belajar di internet. R memiliki komunitas pengguna yang cukup luas dan beragam. Komunitas-komunitas ini tidak segan-segan berbagi dan menjawab pertanyaan anda. Beberapa permasalahan yang anda temui mungkin sudah dijawab orang lain, dan forum-forum seperti ini akan sangat membantu ketika sudah bekerja.

1.3.1.3 Open Source

R adalah *open source*, dengan kata lain, *user* dapat menggunakan R secara cuma-cuma (iya, gratis).

1.3.2 Instalasi R dan RStudio

Untuk menjalankan program-program yang ada di modul ini, anda membutuhkan sebuah komputer atau laptop, lalu anda akan memerlukan aplikasi bernama R dan Rstudio. Modul ini mengasumsikan anda menggunakan sistem operasi windows, namun melakukan instalasi di sistem operasi non-windows juga tidak kalah mudahnya. Langkah-langkah berikut ini juga dapat dengan mudah anda temukan di berbagai situs.

Anda juga membutuhkan kuota internet.

1.3.2.1 Menginstall R

R dapat didownload secara gratis di <https://cran.r-project.org/bin/windows/base/>. Kemudian anda dapat mengetuk pada tulisan “Download R X.x.x for windows (xx megabytes, 32/64 bit)” untuk memulai mengunduh R. Anda akan memulai mengunduh file dengan ekstensi .exe. Setelah unduhan selesai, silakan ketuk dua kali pada file .exe tersebut, dan klik next terus sampai instalasi dimulai.

1.3.2.2 Menginstall RStudio

Setelah R terinstall di komputer anda, silakan pergi ke <https://rstudio.com/products/rstudio/download/> untuk mengunduh RStudio. Pilih RStudio Desktop yang free, lalu ketuk “DOWNLOAD RSTUDIO FOR WINDOWS” untuk memulai pengunduhan. Setelah file selesai diunduh, silakan ketuk dua kali, klik next terus sampai instalasi dimulai.

1.3.3 Tampilan R

Setelah anda selesai menginstall RStudio, maka anda hanya perlu membuka RStudio dari desktop anda. Anda tidak perlu lagi membuka R. Karena itu, *shortcut* RStudio lebih penting daripada *shortcut* R.

RStudio memiliki tampilan utama berupa 4 jendela. Berikut adalah tampilan RStudio:

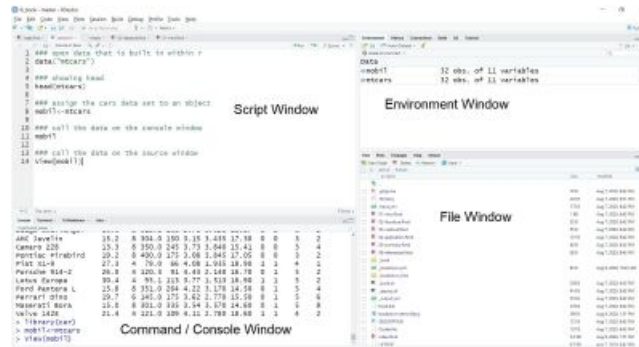


Figure 1.1: Tampilan RStudio

Secara garis besar, RStudio Memiliki 4(empat) jendela, yaitu script, console, environment dan file. kode yang ada di buku ini harus anda ketik di jendela '*console*', sementara script merupakan kumpulan kode.

Untuk saat ini, mengetahui nama-nama dari keempat jendela ini sudah cukup. Bagaimana menggunakannya akan diperjelas di bab-bab berikutnya.

1.3.4 update R

R merupakan aplikasi yang cukup sering mendapatkan update. Karena itu, anda harus rajin-rajin ngecek update ketika menggunakannya. Jangan lupa juga bahwa anda akan memerlukan kuota internet untuk melakukan update.

Ada banyak cara untuk melakukan update terhadap R, tapi berikut ini akan disampaikan cara tercepat untuk melakukannya dengan menggunakan '*console*' di RStudio.

pertama, instal paket bernama "*installr*":

```
install.packages("installr")
```

Instalasi paket ini hanya perlu dilakukan pertama kali anda menginstall R. setelah sekali diinstal, paket itu akan selalu ada di komputer anda. Setelah menginstal paket "*installr*", panggil paket tersebut dengan:


```
library(installr)
```

Fungsi library harus selalu dipanggil ketika akan menggunakan paket tersebut setiap kali anda memulai baru r. Setelah anda memanggil library tersebut, anda tinggal menggunakan fungsi “updateR()” pada console.

```
updateR()
```

Lalu di yes yes saja sampai update selesai.

Chapter 2

Memulai dengan R

under construction

Chapter 3

Manipulasi Data

under construction

Chapter 4

Visualisasi Data

4.1 Tentang Visualisasi Data

Visualisasi data yang baik amat diperlukan guna mengkomunikasikan data yang kita miliki kepada orang lain. Visualisasi data dapat berbentuk tabel, grafik garis, grafik batang, dan lain sebagainya. Kali ini kita akan belajar cara melakukan visualisasi data dengan R.

Pada modul ini, kita akan menggunakan paket bernama `ggplot2` dan paket bernama `dplyr`. untuk itu, jika belum ada paket ini di komputer anda, maka harus diinstal dengan `'install.packages(nama_paket)'`.

Jangan lupa bahwa kita harus terkoneksi dengan internet untuk menginstall paket tersebut. Kita juga harus memanggil paket tersebut dengan

```
library(ggplot2)
library(dplyr)
library(tidyverse)
library(pander)
library(lubridate)
```

Langkah pertama adalah mengambil dataset terlebih dahulu. Kita akan menggunakan data yang telah disiapkan di situs buku ini.

data tersebut merupakan nilai ekspor per bulan Indonesia yang didapat dari Statistik Ekonomi dan Keuangan Indonesia (SEKI), Bank Indonesia, setelah sebelumnya diolah terlebih dahulu untuk mendapatkan tabel seperti di atas. Data dalam ribu USD.

Kita tarik dari situs buku tersebut dengan kode berikut

```
# menarik data dari situs buku ini
dagang<-read.csv(url("https://imedkrisna.github.io/r/dataabi.csv"))

# gunakan paket libridate dan command dmy untuk membuat bulan menjadi waktu
dagang$bulan<-dmy(dagang$bulan)

# panggil 6 baris teratas
head(dagang)
```

```
##      bulan kopi  teh rempah tembakau coklat udang tanilain tekstil  kayu
## 1 2010-01-01 36312 11889 18540      6425 127962 54958 105325 836334 214095
## 2 2010-02-01 36858 11734 16752      4728  67117 60874 125776 814565 241661
## 3 2010-03-01 39535 14239 24954      8512 114329 68049 111998 916195 254687
## 4 2010-04-01 45247 13074 21440     10471  17260 69619 123098 870609 234767
## 5 2010-05-01 60271 12302 25826      6909 127798 63006 119479 888548 229760
## 6 2010-06-01 77593 12322 30163      7282  60956 83919 120270 981359 239067
##      sawit kimia  logam  alat semen kertas  karet minyak elpiji manufakturlain
## 1 565344 254486 593539 526702 2989 288466 505683 296280      NA      1899732
## 2 805257 256887 548909 581052 6039 298346 631605 325349      NA      1951675
## 3 928524 283782 1002479 683638 8087 350199 773076 246641      NA      2284053
## 4 600508 303844 678313 676969 8881 357734 772325 319986      NA      2314475
## 5 788746 332051 654923 677666 8809 357008 791302 389152      NA      2240443
## 6 864754 237749 681485 670858 5943 345368 780795 234197      NA      2385290
##      tembaga nikel batubara bauksit crude      gas gascair tambanglain  emas
## 1 263670 39735 1344304 24857 718273 1049665 754473      23575 58107
## 2 446785 34594 1348624 22999 826872 945454 689446      43447 55722
## 3 724134 50842 1520672 42212 913640 1026612 724647      32253 76236
## 4 322333 34815 1399697 33987 866674 1141495 830103      29723 120110
## 5 545442 48632 1277285 33129 903487 1136135 823827      25486 163462
## 6 341779 41563 1510298 43588 885780 1028868 752482      20895 146505
```

Seperti anda lihat, data tersebut kurang begitu cantik untuk dilihat. Ketika kita bekerja, kita tidak terlalu memusingkan bentuk dari tabel atau grafik, selama informasi yang diberikan oleh data tersebut dapat kita mengerti dengan mudah. Namun, klien atau bos kita di kantor mungkin akan memerlukan visualisasi data yang lebih representatif. Kita akan mencoba mengatur visualisasi data berikutnya.

Untuk keperluan sendiri, anda juga dapat menggunakan *command* ‘View(nama__dataset)’ untuk melihat datanya dengan lebih baik.

4.2 Tabel

Tabel adalah salah satu bentuk visualisasi data yang tidak terlalu populer untuk ditampilkan dalam berbagai publikasi populer. Namun, tabel dapat memberikan detail yang jauh lebih baik dibandingkan bentuk visualisasi data lainnya. Anda akan lebih sering melihat tabel di publikasi akademis atau laporan-laporan yang sifatnya lebih ke menunjukkan hal-hal detail seperti laporan keuangan atau ekspor-impor.

Untuk menulis tabel tersebut, kita akan menggunakan paket `pander` dan `tidyverse`.

```
# melihat tabel dengan hanya variabel tertentu saja
pander(head(dagang %>% select(bulan, udang, kopi, tembakau, sawit, tekstil, logam, kertas)))
```

bulan	udang	kopi	tembakau	sawit	tekstil	logam	kertas
2010-01-01	54958	36312	6425	565344	836334	593539	288466
2010-02-01	60874	36858	4728	805257	814565	548909	298346
2010-03-01	68049	39535	8512	928524	916195	1002479	350199
2010-04-01	69619	45247	10471	600508	870609	678313	357734
2010-05-01	63006	60271	6909	788746	888548	654923	357008
2010-06-01	83919	77593	7282	864754	981359	681485	345368

hanya dengan menggunakan `pander` saja, kita sudah mendapatkan tabel yang terlihat bagus dan representatif. Anda dapat melihat opsi-opsi untuk paket `pander` dengan menggunakan kode `?pander` atau lihat di google.

4.3 Grafik

Data serial waktu (*time series*) seperti ini memang paling cocok divisualisasikan dengan menggunakan grafik garis. Kita akan mencoba menggambar grafik ekspor biji kopi. Ada beberapa hal yang terjadi ketika kita menggambar grafik dengan menggunakan `ggplot`.

```
# perintah ggplot ke sebuah variabel yang saya namakan grafik
grafik<-ggplot(dagang,aes(x=bulan,y=kopi))
# ggplot telah meletakkan bulan di sumbu x dan kopi di sumbu y

# akan tetapi, untuk menggambarkan garisnya, kita perlu menambahkan perintah geoline()
grafik<-grafik+geom_line()+scale_x_date(date_breaks="1 year", date_labels = "%Y")

# panggil dengan print
print(grafik)
```

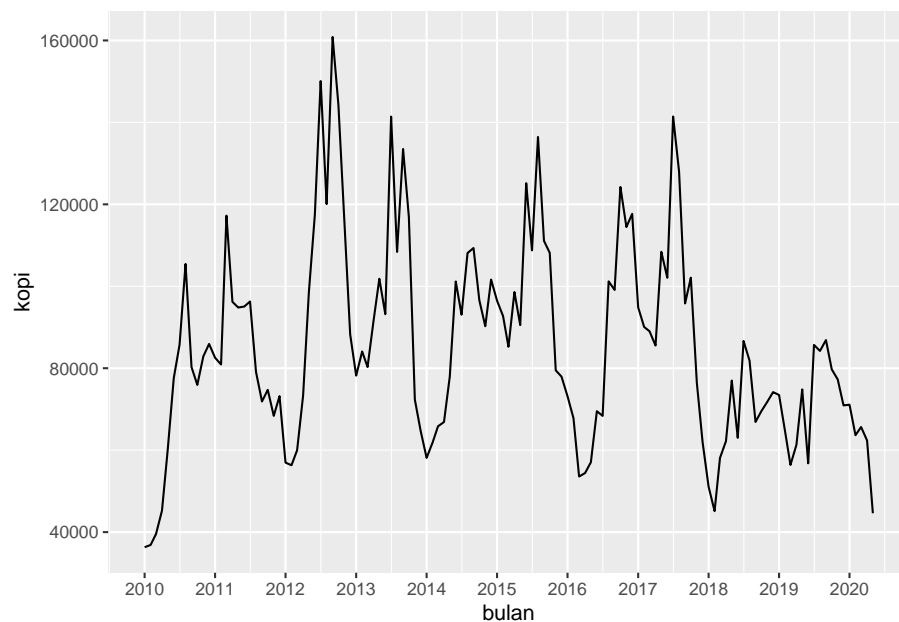


Figure 4.1: ekspor bulanan biji kopi

Grafik di atas menunjukkan perkembangan ekspor bulanan biji kopi Indonesia sejak Januari 2010 sampai Mei 2020.

perintah `geom_line()` juga dapat diatur opsinya untuk membuatnya jadi garis putus-putus, misalnya, atau mengubah warnanya. perintah `scale_x_date()` digunakan untuk membuat tanda x-nya menampilkan interval tahunan.

4.4 Latihan

Apakah anda dapat menggambar grafik yang sama untuk ekspor kertas? Buatlah menjadi garis putus-putus dan warna merah. Tuliskan kode-nya di bawah

ini

```
# awal kode
```

```
# akhir kode
```


Chapter 5

Mean, Median, dan Modus

under construction

Chapter 6

Tes Hipotesis

under construction

Chapter 7

Analisis Regresi

under construction