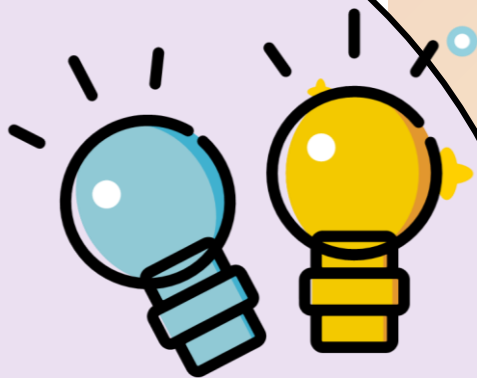


R for Young Data Analyst

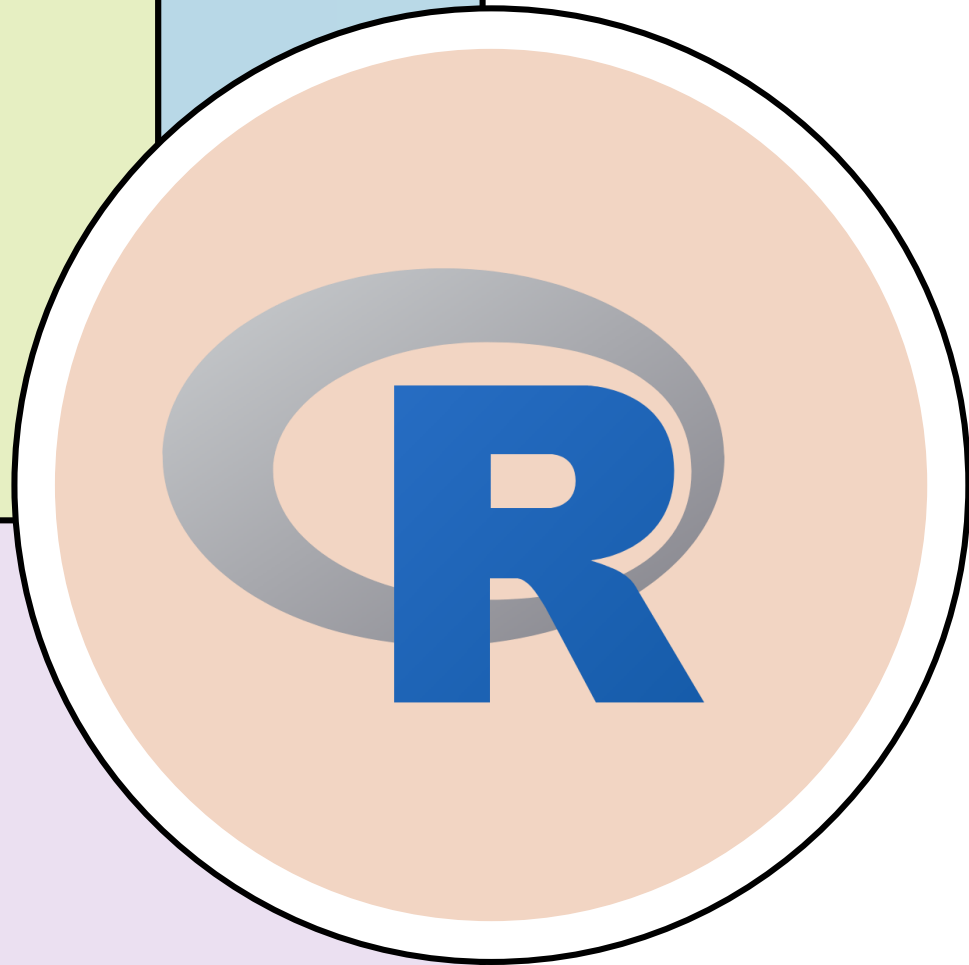
อ.ดร.กรรณิกาน์ หิรัญกลี
ภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากร



1

Getting start with R

Add a brief introduction of your section here: Let's dive in and get to know some interesting facts about animals!



What is R?



What is R?

- R เป็นภาษาโปรแกรม (programming language) สำหรับการคำนวณทางสถิติและการแสดงผลในรูปแบบกราฟ
- R เป็นโอเพนซอร์ส (open-source) ที่ใช้งานได้ฟรีและรองรับระบบปฏิบัติการ UNIX Windows และ Macintosh
- R มีระบบช่วยเหลือที่สามารถเรียกใช้งานได้เลย (built-in help system)
- R มีความสามารถในการแสดงผลการทำงานในรูปแบบกราฟ
- ภาษาของ R มีความง่ายที่จะเรียนรู้ไวยากรณ์ (syntax) ที่มาพร้อมกับฟังก์ชันทางสถิติมากมายที่สามารถเรียกใช้ได้เลย (built-in statistical functions)
- ภาษา R ง่ายต่อการเขียนฟังก์ชันที่ผู้ใช้งานกำหนดเอง (user-written functions)



Download and Install Software for R

Step 1 : Install R



R-4.2.2 for Windows

<https://cran.r-project.org/>

[Download R-4.2.2 for Windows](#) (76 megabytes, 64 bit)

[README on the Windows binary distribution](#)

[New features in this version](#)

Step 2 : Install RStudio Desktop

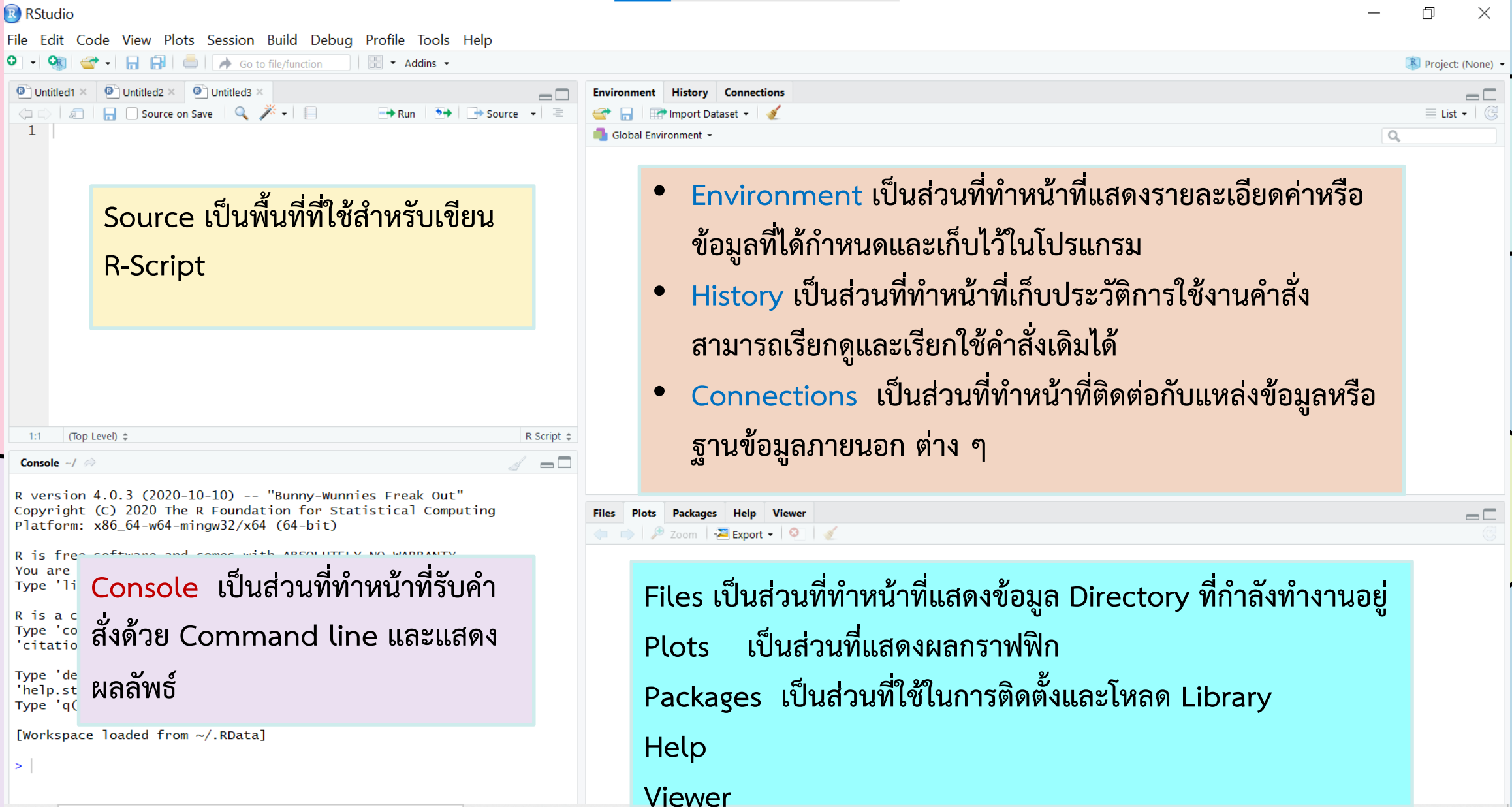


RStudio integrated development environment (IDE)

<https://posit.co/products/open-source/rstudio/>

DOWNLOAD RSTUDIO DESKTOP FOR WINDOWS

Size: 190.49MB | [SHA-256: B38BF925](#) | Version: 2022.07.2+576 | Released: 2022-09-21



Source เป็นพื้นที่ที่ใช้สำหรับเขียน R-Script

- **Environment** เป็นส่วนที่ทำหน้าที่แสดงรายละเอียดค่าหรือข้อมูลที่ได้กำหนดและเก็บไว้ในโปรแกรม
- **History** เป็นส่วนที่ทำหน้าที่เก็บประวัติการใช้งานคำสั่ง สามารถเรียกดูและเรียกใช้คำสั่งเดิมได้
- **Connections** เป็นส่วนที่ทำหน้าที่ติดต่อกับแหล่งข้อมูลหรือฐานข้อมูลภายนอก ต่าง ๆ

Console เป็นส่วนที่ทำหน้าที่รับคำสั่งด้วย Command line และแสดงผลลัพธ์

Files เป็นส่วนที่ทำหน้าที่แสดงข้อมูล Directory ที่กำลังทำงานอยู่
Plots เป็นส่วนที่แสดงผลกราฟฟิก
Packages เป็นส่วนที่ใช้ในการติดตั้งและโหลด Library
Help
Viewer

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function

Project: (None)

Environment History Connections

To Console To Source

5+3*8

3. คลิกที่ Tab “History” จะปรากฏคำสั่งที่ได้ run ไปแล้ว

1. พิมพ์โค้ดในหน้าต่าง Source แล้วคลิก icon Run หรือกดแป้นคีย์บอร์ด Ctrl+Enter

2. โค้ดที่ทำการรันและผลลัพธ์ของการรันจะถูกวางในหน้าต่างคอนโซล

Console

```
> 5+3*8  
[1] 29  
>
```

Files Plots Packages Help Viewer

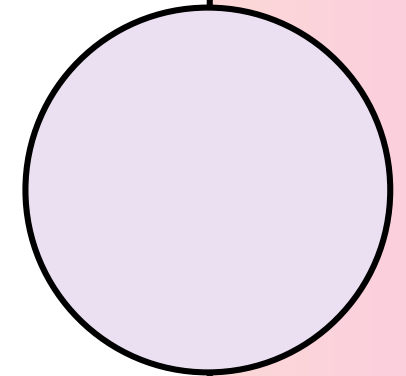
Zoom Export

Operator in R

- การดำเนินการเลขคณิต (Arithmetic operations) ของ R

ได้แก่

การดำเนินการ (operator)	คำอธิบาย (Description)
+	การบวก (addition)
-	การลบ (subtraction)
*	การคูณ (multiplication)
/	การหาร (division)
^ or **	การยกกำลัง (exponentiation)

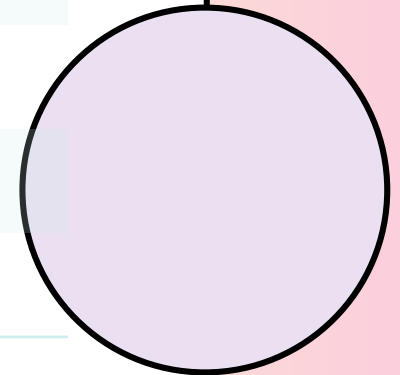


Operators in R

- การดำเนินการเชิงตรรกะ (Logical operations) ของ R ได้แก่

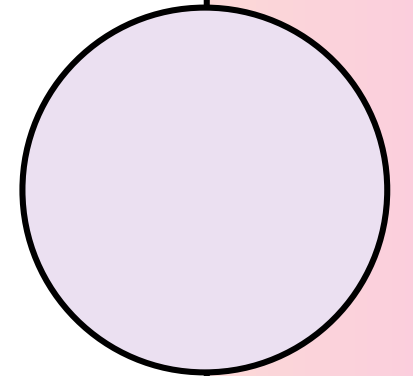
การดำเนินการ (operator)	คำอธิบาย (Description)
-------------------------	------------------------

>	มากกว่า (greater than)
>=	มากกว่าหรือเท่ากับ (greater than or equal to)
==	เท่ากับ (exactly equal to)
!=	ไม่เท่ากับ (not equal to)



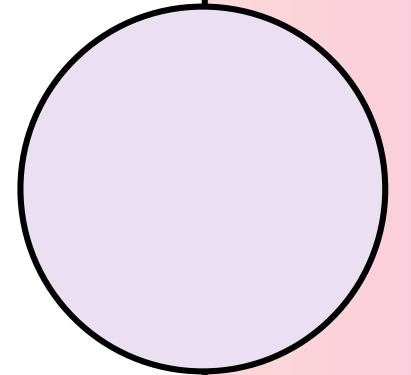
Data Types in R

- R มีชนิดข้อมูลที่หลากหลาย ประกอบด้วย
 - 1) เลขจำนวนจริง (numeric หรือ double)
 - 2) เลขจำนวนเต็ม (Integer)
 - 3) อักขระ (character) แสดงโดยเขียนชุดของอักขระภายในเครื่องหมายคำพูด “.....” (double quotes)
 - 4) ตรรกะ (logical) มีค่าได้ 2 แบบ คือ จริง (TRUE) หรือ เท็จ (FALSE)
 - 5) เลขจำนวนเชิงซ้อน (complex)



Data Structures in R

- โครงสร้างข้อมูลของ R ประกอบด้วย
 - 1) เวกเตอร์ (Vectors)
 - 2) เมทริกซ์ (Matrices)
 - 3) อาร์เรย์ (Arrays)
 - 4) ดาต้าเฟรม (Data frames)
 - 5) ลิสต์ (Lists)



Data Structures in R

vector

สร้างเวกเตอร์ v1 ที่มีสมาชิก 4 ค่า

```
v1 <- c(11, 12, 15, 18)
```

```
> v1
[1] 11 12 15 18
```

matrix

สร้างเมทริกซ์ m1 ขนาด 2x4

```
m1 <- matrix(11:18,nrow=2,ncol=4)
```

```
> m1 <- matrix(11:18,nrow=2,ncol=4)
> m1
     [,1] [,2] [,3] [,4]
[1,]  11  13  15  17
[2,]  12  14  16  18
```

array

สร้างอาร์เรย์ของเมทริกซ์ขนาด 3x3 จำนวน 2 เมทริกซ์

```
# Create two vectors of different lengths.
vector1 <- c(5,9,3)
vector2 <- c(10,11,12,13,14,15)
# Take these vectors as input to the array.
result <- array(c(vector1,vector2),dim = c(3,3,2))
result
```

```
> result
, , 1
```

```
     [,1] [,2] [,3]
[1,]    5   10   13
[2,]    9   11   14
[3,]    3   12   15
```

```
, , 2
```

```
     [,1] [,2] [,3]
[1,]    5   10   13
[2,]    9   11   14
[3,]    3   12   15
```

Data Structures in R

Data frame

สร้างดาต้าเฟรม data1 มีข้อมูล 3 ตัวแปร (คอลัมน์) และหน่วยสังเกต (observations) 5 แถว

```
# Create three vectors of equal lengths.
v1 <- c("001","002","003","004","005")
v2 <- c(11, 12, 15, 18, 20)
v3 <- c("Yes","NO","Yes","NO","NO")
# Take these vectors as columns to the data frame.
data1 <- data.frame(v1,v2,v3)
```

```
> data1
  v1 v2 v3
1 001 11 Yes
2 002 12 NO
3 003 15 Yes
4 004 18 NO
5 005 20 NO
```

List

สร้างลิสต์

```
# Create Two vectors
u1 <- 1:5
u2 <- c(T,T,F,F,T)
# Take these vectors as input to the list.
mylist <- list(numbers=u1, wrong=u2)
```

```
> mylist <- list(numbers=u1, wrong=u2)
> mylist
$numbers
[1] 1 2 3 4 5

$wrong
[1] TRUE TRUE FALSE FALSE TRUE
```

2

Data analysis

Import data set in R and Descriptive statistics

Import data set

- ไฟล์ข้อมูล covid-19.csv เป็นข้อมูลขององค์การอนามัยโลกเกี่ยวกับการแพร่ระบาดของไวรัสโคโรนา-19 (COVID-19) ประกอบด้วย ข้อมูลต่าง ๆ ดังนี้

	Field name	Type	Description
1.	Name	String	Country, territory, area (ประเทศ)
2.	WHO_region	String	WHO Region (ทวีป)
3.	Cases_cumulative total	Integer	Cumulative confirmed cases reported to WHO to date. (จำนวนผู้ติดเชื้อสะสมที่ได้รับการยืนยัน)
4.	Cases_cumulative total per 100000 population	Decimal	Cumulative confirmed cases reported to WHO to date per 100,000 population. (จำนวนผู้ติดเชื้อสะสมที่ได้รับการยืนยันต่อประชากรหนึ่งแสนคน)
5.	Cases_newly reported in last 7 days	Integer	New confirmed cases reported in the last 7 days. Calculated by subtracting previous cumulative case count (8 days prior) from current cumulative cases count. (จำนวนผู้ติดเชื้อรายใหม่ที่มีการยืนยันในช่วง 7 วันล่าสุด)

ที่มา: <https://covid19.who.int/WHO-COVID-19-global-data..csv> และ <https://covid19.who.int/who-data/vaccination-data.csv>

Import data set

- ไฟล์ข้อมูล covid-19.csv

	Field name	Type	Description
6.	Cases_newly reported in last 7 days per 100000 population	Decimal	New confirmed cases reported in the last 7 days per 100,000 population. (จำนวนผู้ติดเชื้อรายใหม่ที่มีการยืนยันในช่วง 7 วันล่าสุดต่อประชากรหนึ่งแสนคน)
7.	Cases_newly reported in last 24 hours	Integer	New confirmed cases reported in the last 24 hours. Calculated by subtracting previous cumulative case count from current cumulative cases count. (จำนวนผู้ติดเชื้อรายใหม่ที่มีการยืนยันในช่วง 24 ชม. ล่าสุด)
8.	Deaths_cumulative total	Integer	Cumulative confirmed deaths reported to WHO to date. (จำนวนผู้เสียชีวิตสะสมที่ได้รับการยืนยัน)
9.	Deaths_cumulative total per 100000 population	Decimal	Cumulative confirmed deaths reported to WHO to date per 100,000 population. (จำนวนผู้เสียชีวิตสะสมที่ได้รับการยืนยันต่อประชากรหนึ่งแสนคน)
10.	Deaths_newly reported in last 7 days	Integer	New confirmed deaths reported in the last 7 days. Calculated by subtracting previous cumulative death count (8 days prior) from current cumulative deaths count. (จำนวนผู้เสียชีวิตใหม่ที่ได้รับการยืนยันในช่วง 7 วันล่าสุด)

Import data set

- ไฟล์ข้อมูล covid-19.csv

	Field name	Type	Description
11.	Deaths_newly reported in last 7 days per 100000 population	Decimal	New confirmed deaths reported in the last 7 days per 100,000 population. (จำนวนผู้เสียชีวิตใหม่ที่ได้รับการยืนยันในช่วง 7 วันล่าสุดต่อประชากรหนึ่งแสนคน)
12.	Deaths_newly reported in last 24 hours	Integer	New confirmed deaths reported in the last 24 hours. Calculated by subtracting previous cumulative death count from current cumulative deaths count. (จำนวนผู้เสียชีวิตใหม่ที่ได้รับการยืนยันในช่วง 24 ชม.ล่าสุด)
13.	DATE_UPDATED	Date	Date of last update (วันที่ปรับปรุงข้อมูลล่าสุด)
14.	TOTAL_VACCINATIONS	Integer	Cumulative total vaccine doses administered (จำนวนโดสของวัคซีนทั้งหมดสะสมที่มีการบริหารจัดการ)
15.	PERSONS_VACCINATED_1 PLUS_DOSE	Decimal	Cumulative number of persons vaccinated with at least one dose (จำนวนผู้ที่ได้รับวัคซีนสะสมอย่างน้อยหนึ่งโดส)

Import data set

- ไฟล์ข้อมูล covid-19.csv

	Field name	Type	Description
16.	TOTAL_VACCINATIONS_PER100	Integer	Cumulative total vaccine doses administered per 100 population (จำนวนผู้ที่ได้รับวัคซีนสะสมอย่างน้อยหนึ่งโดสต่อประชากรหนึ่งร้อยคน)
17.	PERSONS_VACCINATED_1PLUS_DOSE_PER100	Decimal	Cumulative persons vaccinated with at least one dose per 100 population (จำนวนผู้ที่ได้รับวัคซีนสะสมอย่างน้อยหนึ่งโดสต่อประชากรหนึ่งร้อยคน)
18.	PERSONS_FULLY_VACCINATED	Integer	Cumulative number of persons fully vaccinated (จำนวนผู้ที่ได้รับวัคซีนเต็มโดสต่อประชากรหนึ่งแสนคน)
19.	PERSONS_FULLY_VACCINATED_PER100	Decimal	Cumulative number of persons fully vaccinated per 100 population (จำนวนผู้ที่ได้รับวัคซีนเต็มโดสต่อประชากรหนึ่งร้อยคน)
20.	VACCINES_USED	String	Combined short name of vaccine: "Company - Product name" (see below) (ชื่อวัคซีนโดยย่อ)

Import data set

- ไฟล์ข้อมูล covid-19.csv

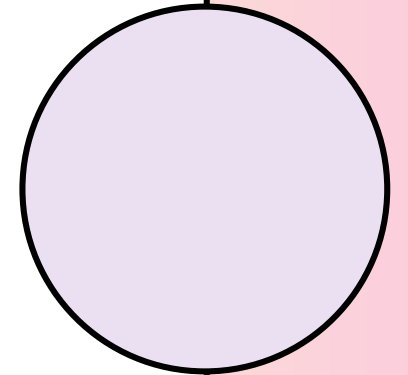
	Field name	Type	Description
21.	FIRST_VACCINE_DATE	Date	Date of first vaccinations. Equivalent to start/launch date of the first vaccine administered in a country. (วันแรกที่มีการใช้วัคซีน)
22.	NUMBER_VACCINES_TYPES_USED	Integer	Number of vaccine types used per country, territory, area (จำนวนชนิดของวัคซีนที่ใช้ในประเทศ)
23.	PERSONS_BOOSTER_ADD_DOSE	Integer	Persons received booster or additional dose (จำนวนผู้ที่ได้รับวัคซีนบูสเตอร์หรือเพิ่มโดส)
24.	PERSONS_BOOSTER_ADD_DOSE_PER100	Decimal	Persons received booster or additional dose per 100 population (จำนวนผู้ที่ได้รับวัคซีนบูสเตอร์หรือเพิ่มโดสต่อประชากรหนึ่งร้อยคน)

Import data set

- การนำเข้าข้อมูลที่เป็นไฟล์นามสกุล .csv ด้วยฟังก์ชัน read.csv()

```
covid19 <- read.csv("F:\\R for young data analysts\\covid data.csv", header=TRUE)
str(covid19)
View(covid19)
attach(covid19)
```

- ฟังก์ชัน str เป็นฟังก์ชันที่แสดงโครงสร้างภายในของวัตถุ สามารถใช้เพื่อแสดงรายละเอียดของวัตถุโดยสรุปอย่างสั้นๆ
- ฟังก์ชัน View แสดงรายละเอียดของดาต้าเฟรมในรูปแบบ spread sheet
- ฟังก์ชัน attach เป็นการแนบดาต้าเฟรมไว้ในพื้นที่ทำงานของ R (workspace) ซึ่งเมื่อดาต้าเฟรมถูกแนบไว้ในพื้นที่ทำงาน เราสามารถเรียกใช้งานหรืออ้างอิงตัวแปรในดาต้าเฟรมโดยไม่ต้องใส่ชื่อดาต้าเฟรมนำหน้าชื่อตัวแปร



Import data set

- การแสดงรายละเอียดของ data frame ด้วยฟังก์ชัน str()

```
> str(covid19)
```

```
'data.frame': 223 obs. of 24 variables:
```

```
$ Name : chr "Afghanistan" "Albania" "Algeria" "American Samoa" ...
$ WHO.Region : chr "Eastern Mediterranean" "Europe" "Africa" "Western Pacific" ...
$ Cases_cumulative.total : int 202993 331800 270836 8257 46535 103131 3866 9106 9717546 445242 ...
$ Cases_cumulative.total.per.100000.population : num 521 11530 618 14959 60228 ...
$ Cases_newly.reported.in.last.7.days : int 844 174 59 0 86 0 0 0 0 142 ...
$ Cases_newly.reported.in.last.7.days.per.100000.population : num 2.168 6.046 0.135 0 111.305 ...
$ Cases_newly.reported.in.last.24.hours : int 159 17 7 0 0 0 0 0 0 142 ...
$ Death_cumulative.total : int 7821 3593 6881 34 155 1917 12 146 129979 8709 ...
$ Deaths_cumulative.total.per.100000.population : num 20.1 124.9 15.7 61.6 200.6 ...
$ Deaths_newly.reported.in.last.7.days : int 3 1 0 0 0 0 0 0 0 3 ...
$ Deaths_newly.reported.in.last.7.days.per.100000.population : num 0.008 0.035 0 0 0 0 0 0 0 0.101 ...
$ Deaths_newly.reported.in.last.24.hours : int 0 0 0 0 0 0 0 0 0 3 ...
$ DATE_UPDATED : chr "10/24/2022" "10/16/2022" "9/4/2022" "8/23/2022" ...
$ TOTAL_VACCINATIONS : num 12055358 2991576 15267442 111316 154320 ...
$ PERSONS_VACCINATED_1PLUS_DOSE : int 11084618 1339100 7840131 44885 57898 14220830 10852 64290 41324100 1129669 ...
$ TOTAL_VACCINATIONS_PER100 : num 31 104 34.8 201.7 199.7 ...
$ PERSONS_VACCINATED_1PLUS_DOSE_PER100 : num 28.5 47.1 17.9 81.3 76 ...
$ PERSONS_FULLY_VACCINATED : int 10386823 1265900 6481186 41423 53482 7814121 10366 62384 37840119 985807 ...
$ PERSONS_FULLY_VACCINATED_PER100 : num 26.7 44.5 14.8 75 70.2 ...
$ VACCINES_USED : chr "AstraZeneca - Vaxzevria,Beijing CNBG - BBIBP-CorV,Bharat - Covaxin,CanSino -
Convidecia,Gamaleya - Gam-Covid-Va"| __truncated__ "AstraZeneca - Vaxzevria,Gamaleya - Gam-Covid-Vac,Pfizer BioNTech - Comirnaty,SII - Covishiel
d,Sinovac - CoronaVac" "Beijing CNBG - BBIBP-CorV,Gamaleya - Gam-Covid-Vac,SII - Covishield,Sinovac - CoronaVac" "Janssen - Ad26.COV 2-S,Moderna -
Spikevax,Pfizer BioNTech - Comirnaty" ...
$ FIRST_VACCINE_DATE : chr "2/22/2021" "1/13/2021" "1/30/2021" "12/21/2020" ...
$ NUMBER_VACCINES_TYPES_USED : int 11 5 4 3 3 1 2 6 7 8 ...
$ PERSONS_BOOSTER_ADD_DOSE : int NA 363122 575651 24160 42940 1127156 2998 9838 30810184 40725 ...
$ PERSONS_BOOSTER_ADD_DOSE_PER100 : num NA 12.76 1.31 43.77 56.37 ...
```

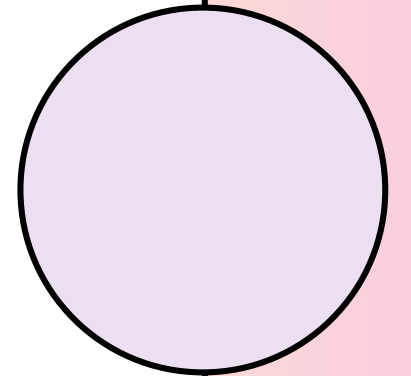
Descriptive statistics

- การทำตารางสรุปข้อมูล (Tabulation) ที่เป็นตัวแปรจำแนกประเภทด้วยฟังก์ชัน `table()`

```
table(WHO.Region)
```

```
> table(WHO.Region)
```

```
WHO.Region
      Africa      Americas Eastern Mediterranean      Europe
        47         51          21          59
South-East Asia Western Pacific
        10         35
```



สถิติพรรณนา (Descriptive Statistics)

การวัดแนวโน้มเข้าสู่ส่วนกลาง
(Measure of central tendency)

- ค่าเฉลี่ย (Mean)
- มัธยฐาน (Median)
- ฐานนิยม (Mode)

ค่าวัดตำแหน่งสัมพัทธ์
ในชุดข้อมูล

- ควอร์ไทล์ (Quartiles)
- เดไซส์ (Deciles)
- เพอร์เซ็นไทล์ (Percentiles)

การวัดการกระจาย
(Measures of dispersion)

- พิสัย (Range)
- ส่วนเบี่ยงเบนมาตรฐาน (standard deviation)

การวัดแนวโน้มเข้าสู่ส่วนกลาง

ค่าเฉลี่ย (Mean)

- ผลรวมของค่าสังเกตทุกค่าของข้อมูลชุดนั้นหารด้วยจำนวนค่าสังเกต

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

มัธยฐาน (Median)

- เป็นค่าที่แบ่งค่าสังเกตที่เรียงลำดับจากน้อยไปมากออกเป็น 2 ส่วน โดยมีจำนวนค่าสังเกตที่มีค่าน้อยกว่าหรือเท่ากับค่ามัธยฐานอยู่ 50% และมากกว่าค่ามัธยฐานอยู่ 50%

ฐานนิยม (Mode)

- เป็นค่าสังเกตที่มีความถี่สูงสุดในชุดข้อมูล
- ชุดข้อมูลอาจไม่มีฐานนิยมหรือมีมากกว่า 1 ค่า

ค่าวัดตำแหน่งสัมพัทธ์ในชุดข้อมูล

ควอร์ไทล์ (Quartiles)

- แบ่งข้อมูลออกเป็น 4 ส่วนเท่า ๆ กัน
- ข้อมูลชุดหนึ่งจะมี 3 ควอร์ไทล์ คือ Q_1 , Q_2 และ Q_3

เดไซล์ (Deciles)

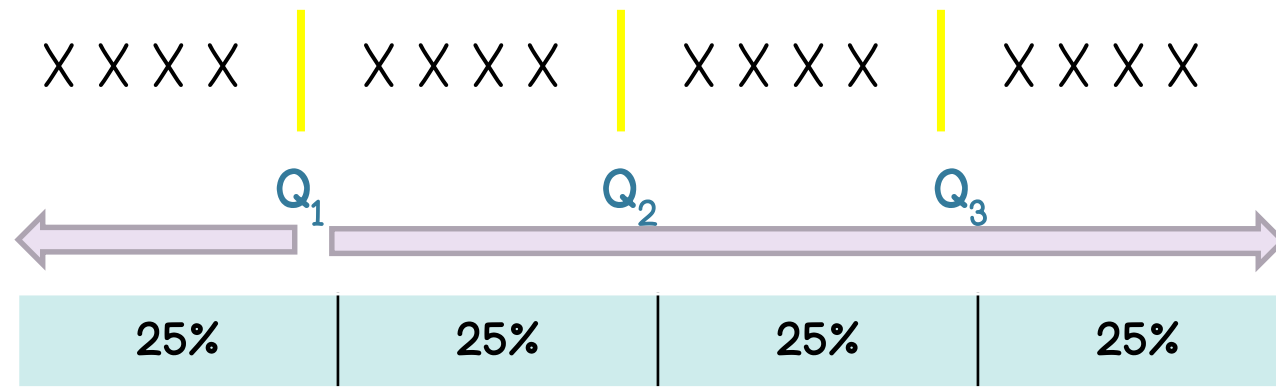
- แบ่งข้อมูลออกเป็น 10 ส่วนเท่า ๆ กัน
- ข้อมูลชุดหนึ่งจะมี 9 เดไซล์ คือ D_1, D_2, \dots, D_9

เปอร์เซ็นต์ไทล์ (Percentiles)

- เปอร์เซ็นต์ไทล์ แบ่งข้อมูลออกเป็น 100 ส่วนเท่า ๆ กัน
- ข้อมูลชุดหนึ่งจะมี 99 เปอร์เซ็นต์ไทล์ คือ P_1, P_2, \dots, P_{99}

ควอร์ไทล์ (Quartiles)

เรียงข้อมูลจากน้อยไปมาก



ควอร์ไทล์ที่หนึ่ง Q_1 คือ จำนวนที่แบ่งข้อมูลเป็น 25% ที่มีค่าน้อยกว่าและ 75% ที่มีค่ามากกว่า
 ควอร์ไทล์ที่สอง Q_2 คือ จำนวนที่แบ่งข้อมูลเป็น 50% ที่มีค่าน้อยกว่าและ 50% ที่มีค่ามากกว่า
 ควอร์ไทล์ที่สาม Q_3 คือ จำนวนที่แบ่งข้อมูลเป็น 75% ที่มีค่าน้อยกว่า และ 25% ที่มีค่ามากกว่า

เดไซส์ (Deciles)

เรียงข้อมูลจากน้อยไปมาก

x x	x x	x x	x x	x x	x x	x x	x x	x x	x x
D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	
10%	10%	10%	10%	10%	10%	10%	10%	10%	10%

เดไซส์ที่หนึ่ง D_1 คือ ค่าที่แบ่งข้อมูลออกเป็น 10% ที่มีค่าน้อยกว่าและ 90% ที่มีค่ามากกว่า

เดไซส์ที่สอง D_2 คือ ค่าที่แบ่งข้อมูลออกเป็น 20% ที่มีค่าน้อยกว่าและ 80% ที่มีค่ามากกว่า

⋮

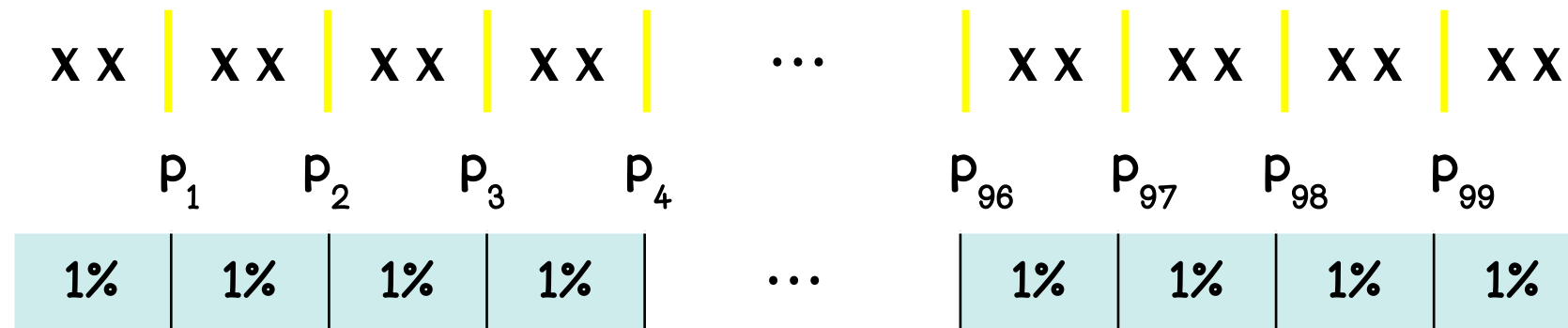
⋮

⋮

เดไซส์ที่เก้า D_9 คือ ค่าที่แบ่งข้อมูลออกเป็น 90% ที่มีค่าน้อยกว่าและ 10% ที่มีค่ามากกว่า

เปอร์เซ็นต์ไทล์ (Percentiles)

เรียงข้อมูลจากน้อยไปมาก



เปอร์เซ็นต์ไทล์ที่หนึ่ง p_1 คือ ค่าที่แบ่งข้อมูลออกเป็น 1% ที่มีค่าน้อยกว่าและ 99% ที่มีค่ามากกว่า

เปอร์เซ็นต์ไทล์ที่สอง p_2 คือ ค่าที่แบ่งข้อมูลออกเป็น 2% ที่มีค่าน้อยกว่าและ 98% ที่มีค่ามากกว่า

⋮

⋮

⋮

เปอร์เซ็นต์ไทล์ที่เก้าสิบเก้า p_{99} คือ ค่าที่แบ่งข้อมูลออกเป็น 99% ที่มีค่าน้อยกว่าและ 1% ที่มีค่ามากกว่า

การวัดการกระจาย

พิสัย (Range)

- ผลต่างระหว่างค่าสูงสุดและค่าต่ำสุดของข้อมูล

พิสัยระหว่างควอร์ไทล์ (Interquartile Range)

- ผลต่างระหว่างควอร์ไทล์ที่ 3 และควอร์ไทล์ที่ 1
$$IQR = Q_3 - Q_1$$

ความแปรปรวนของตัวอย่าง (Variance)

- ผลรวมของกำลังสองของระยะทางจากค่าสังเกตแต่ละค่าไปยังค่าเฉลี่ยที่หารด้วย $n - 1$

ส่วนเบี่ยงเบนมาตรฐานของตัวอย่าง (Standard deviation)

- ค่ารากที่สองของความแปรปรวน

ความแปรปรวน (Variance) และส่วนเบี่ยงเบนมาตรฐาน (Standard deviation)

ความแปรปรวนของตัวอย่างของชุดข้อมูล (ประกอบด้วยค่าสังเกต n ค่า) คือ ผลรวมของกำลังสองของระยะทางจากค่าสังเกตแต่ละค่าไปยังค่าเฉลี่ยหารด้วย $n - 1$ เขียนแทนด้วยสัญลักษณ์ S^2 คำนวณจากสูตรดังนี้

$$S^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

ส่วนเบี่ยงเบนมาตรฐานของตัวอย่าง คือ ค่ารากที่สองที่เป็นบวกของความแปรปรวนตัวอย่าง เขียนแทนด้วยสัญลักษณ์ S

$$S = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

สถิติพรรณนา (Descriptive statistics)

- ฟังก์ชัน R ที่ใช้ในการหาสถิติพรรณนา ได้แก่ `mean`, `sd`, `var`, `min`, `max`, `median`, `range`, and `quantile`

ฟังก์ชัน (Function)	คำอธิบาย (Description)
<code>mean(x, trim = 0, na.rm = FALSE)</code>	ค่าเฉลี่ยของสมาชิกในเวกเตอร์ x
<code>sd(x, na.rm = FALSE)</code>	ค่าเบี่ยงเบนมาตรฐานตัวอย่างของสมาชิกในเวกเตอร์ x
<code>var(x, na.rm = FALSE)</code>	ค่าความแปรปรวนตัวอย่างของสมาชิกในเวกเตอร์ x
<code>min(x, na.rm = FALSE)</code>	ค่าต่ำสุดของเวกเตอร์ x
<code>max(x, na.rm = FALSE)</code>	ค่าสูงสุดของเวกเตอร์ x
<code>median(x, na.rm = FALSE)</code>	ค่ามัธยฐานของสมาชิกในเวกเตอร์ x
<code>range(..., na.rm = FALSE)</code>	ฟังก์ชันส่งกลับคืนเป็นเวกเตอร์ที่มีค่าต่ำสุดและค่าสูงสุด
<code>quantile(x, probs = seq(0, 1, 0.25), na.rm = FALSE, names = TRUE, type = 7)</code>	ฟังก์ชันส่งกลับคืนเป็นเวกเตอร์ที่ประกอบด้วย ค่าต่ำสุด ควอร์ไทล์ที่ 1 ควอร์ไทล์ที่ 2 ควอร์ไทล์ที่ 3 และค่าสูงสุด

ฟังก์ชันต่าง ๆ ข้างต้น จะกำหนดค่าอาร์กิวเมนต์ของการเอาข้อมูลสูญหาย (NA) ออกจากข้อมูล `na.rm = FALSE` ซึ่งถ้ามีข้อมูล NA อยู่ด้วย ฟังก์ชันจะส่งกลับค่า NA

ฟังก์ชันระบุ Default ของ argument ที่มีชื่อ na.rm=FALSE

```
> mean(PERSONS_BOOSTER_ADD_DOSE)
[1] NA
> sd(PERSONS_BOOSTER_ADD_DOSE)
[1] NA
> min(PERSONS_BOOSTER_ADD_DOSE)
[1] NA
> max(PERSONS_BOOSTER_ADD_DOSE)
[1] NA
> median(PERSONS_BOOSTER_ADD_DOSE)
[1] NA
> range(PERSONS_BOOSTER_ADD_DOSE)
[1] NA NA
> quantile(PERSONS_BOOSTER_ADD_DOSE)
Error in quantile.default(PERSONS_BOOSTER_ADD_DOSE) :
  missing values and NaN's not allowed if 'na.rm' is FALSE
```

เปลี่ยน argument ที่มีชื่อ na.rm=TRUE ในฟังก์ชัน

```
> mean(PERSONS_BOOSTER_ADD_DOSE, na.rm=TRUE)
[1] 11391218
> sd(PERSONS_BOOSTER_ADD_DOSE, na.rm=TRUE)
[1] 58863044
> min(PERSONS_BOOSTER_ADD_DOSE, na.rm=TRUE)
[1] 0
> max(PERSONS_BOOSTER_ADD_DOSE, na.rm=TRUE)
[1] 776425498
> median(PERSONS_BOOSTER_ADD_DOSE, na.rm=TRUE)
[1] 499879
> range(PERSONS_BOOSTER_ADD_DOSE, na.rm=TRUE)
[1] 0 776425498
> quantile(PERSONS_BOOSTER_ADD_DOSE, na.rm=TRUE)
      0%      25%      50%      75%     100%
0.0    39403.5  499879.0 3813560.0 776425498.0
```


Descriptive statistics

สถิติพรรณนา (Descriptive statistics)

- `summary(...)` เป็นฟังก์ชัน R ที่ใช้ในการแสดงค่าสถิติต่าง ๆ ของข้อมูลแต่ละคอลัมน์ (summary statistics) ใน data frame

```
summary(covid19)
```

- `fivenum(x, na.rm = TRUE)` เป็นฟังก์ชัน R ที่ใช้ในการแสดงค่าสถิติ 5 ค่าของเวกเตอร์ x ได้แก่ ค่าต่ำสุด คออร์ไทล์ที่1 คออร์ไทล์ที่2 คออร์ไทล์ที่3 และค่าสูงสุด

```
fivenum(PERSONS_BOOSTER_ADD_DOSE)
```

```
> fivenum(PERSONS_BOOSTER_ADD_DOSE)
[1]      0.0    39403.5   499879.0  3813560.0 776425498.0
```

Descriptive statistics

> summary(covid19)

Name	WHO.Region	Cases_cumulative.total	Cases_cumulative.total.per.100000.population	Cases_newly.reported.in.last.7.days
Length:223	Length:223	Min. : 0	Min. : 0	Min. : 0.0
Class :character	Class :character	1st Qu.: 22752	1st Qu.: 1751	1st Qu.: 0.0
Mode :character	Mode :character	Median : 202993	Median : 12310	Median : 24.0
		Mean : 2728759	Mean : 18732	Mean : 8737.1
		3rd Qu.: 1256254	3rd Qu.: 31132	3rd Qu.: 740.5
		Max. : 95946824	Max. : 70926	Max. : 336437.0

Cases_newly.reported.in.last.7.days.per.100000.population	Cases_newly.reported.in.last.24.hours	Death_cumulative.total	Deaths_cumulative.total.per.100000.population
Min. : 0.00	Min. : 0.0	Min. : 0	Min. : 0.00
1st Qu.: 0.00	1st Qu.: 0.0	1st Qu.: 167	1st Qu.: 14.32
Median : 0.44	Median : 0.0	Median : 1968	Median : 78.40
Mean : 39.70	Mean : 354.8	Mean : 28946	Mean : 121.77
3rd Qu.: 15.64	3rd Qu.: 0.0	3rd Qu.: 14430	3rd Qu.: 200.32
Max. : 608.46	Max. : 33774.0	Max. : 1059255	Max. : 657.78

Deaths_newly.reported.in.last.7.days	Deaths_newly.reported.in.last.7.days.per.100000.population	Deaths_newly.reported.in.last.24.hours	DATE_UPDATED	TOTAL_VACCINATIONS
Min. : 0.00	Min. : 0.0000	Min. : 0.000	Length:223	Min. : 1.380e+02
1st Qu.: 0.00	1st Qu.: 0.0000	1st Qu.: 0.000	Class :character	1st Qu.: 4.710e+05
Median : 0.00	Median : 0.0000	Median : 0.000	Mode :character	Median : 4.125e+06
Mean : 23.53	Mean : 0.1271	Mean : 1.637		Mean : 5.700e+07
3rd Qu.: 4.00	3rd Qu.: 0.0360	3rd Qu.: 0.000		3rd Qu.: 2.007e+07
Max. : 549.00	Max. : 2.5480	Max. : 112.000		Max. : 3.458e+09
				NA's : 1

PERSONS_VACCINATED_1PLUS_DOSE	TOTAL_VACCINATIONS_PER100	PERSONS_VACCINATED_1PLUS_DOSE_PER100	PERSONS_FULLY_VACCINATED	PERSONS_FULLY_VACCINATED_PER100	VACCINES_USED	FIRST_VACCINE_DATE
Min. : 0.000e+00	Min. : 0.212	Min. : 0.00	Min. : 0.000e+00	Min. : 0.00	Length:223	Length:223
1st Qu.: 2.065e+05	1st Qu.: 74.117	1st Qu.: 42.15	1st Qu.: 1.923e+05	1st Qu.: 35.19	Class :character	Class :character
Median : 2.434e+06	Median : 152.482	Median : 67.14	Median : 2.144e+06	Median : 61.50	Mode :character	Mode :character
Mean : 2.402e+07	Mean : 147.081	Mean : 61.57	Mean : 2.204e+07	Mean : 56.23		
3rd Qu.: 9.325e+06	3rd Qu.: 215.918	3rd Qu.: 81.66	3rd Qu.: 8.077e+06	3rd Qu.: 77.26		
Max. : 1.307e+09	Max. : 364.741	Max. : 124.88	Max. : 1.277e+09	Max. : 122.94		
NA's : 1	NA's : 1	NA's : 1	NA's : 1	NA's : 1		

NUMBER_VACCINES_TYPES_USED	PERSONS_BOOSTER_ADD_DOSE	PERSONS_BOOSTER_ADD_DOSE_PER100
Min. : 1.000	Min. : 0	Min. : 0.000
1st Qu.: 3.000	1st Qu.: 39404	1st Qu.: 7.484
Median : 4.000	Median : 499879	Median : 28.959
Mean : 4.721	Mean : 11391218	Mean : 30.882
3rd Qu.: 6.000	3rd Qu.: 3813560	3rd Qu.: 51.863
Max. : 12.000	Max. : 776425498	Max. : 107.922
NA's : 4	NA's : 24	NA's : 24

Descriptive statistics

- ผลลัพธ์ของ summary(covid19) ข้างตัวแปร

PERSONS_VACCINATED_1PLUS_DOSE_PER100	PERSONS_FULLY_VACCINATED	
Min. : 0.00	Min. :0.000e+00	
1st Qu.: 42.15	1st Qu.:1.923e+05	
Median : 67.14	Median :2.144e+06	
Mean : 61.57	Mean :2.204e+07	
3rd Qu.: 81.66	3rd Qu.:8.077e+06	
Max. :124.88	Max. :1.277e+09	
NA's :1	NA's :1	
PERSONS_FULLY_VACCINATED_PER100	VACCINES_USED	FIRST_VACCINE_DATE
Min. : 0.00	Length:223	Length:223
1st Qu.: 35.19	Class :character	Class :character
Median : 61.50	Mode :character	Mode :character
Mean : 56.23		
3rd Qu.: 77.26		
Max. :122.94		
NA's :1		
NUMBER_VACCINES_TYPES_USED	PERSONS_BOOSTER_ADD_DOSE	PERSONS_BOOSTER_ADD_DOSE_PER100
Min. : 1.000	Min. : 0	Min. : 0.000
1st Qu.: 3.000	1st Qu.: 39404	1st Qu.: 7.484
Median : 4.000	Median : 499879	Median : 28.959
Mean : 4.721	Mean : 11391218	Mean : 30.882
3rd Qu.: 6.000	3rd Qu.: 3813560	3rd Qu.: 51.863
Max. :12.000	Max. :776425498	Max. :107.922
NA's :4	NA's :24	NA's :24

Descriptive statistics

- การหาค่าสถิติพรรณนาจำแนกตามกลุ่มย่อยที่สนใจ ด้วยฟังก์ชัน aggregate
- ตัวอย่าง การหาค่าเฉลี่ยของจำนวนคนที่ได้รับวัคซีนบูสเตอร์ ในแต่ละทวีป

```
> aggregate(PERSONS_BOOSTER_ADD_DOSE~WHO.Region,data=covid19,mean)
  WHO.Region PERSONS_BOOSTER_ADD_DOSE
1      Africa          845011.4
2    Americas          8615627.9
3 Eastern Mediterranean          6763861.2
4      Europe          5338640.7
5 South-East Asia          39434313.0
6 Western Pacific          29148361.2
> aggregate(PERSONS_BOOSTER_ADD_DOSE~WHO.Region,data=covid19,sd)
  WHO.Region PERSONS_BOOSTER_ADD_DOSE
1      Africa          1399370
2    Americas          23085874
3 Eastern Mediterranean          12376799
4      Europe          10771212
5 South-East Asia          68037691
6 Western Pacific          131182153
```

Descriptive statistics

- การหาค่าสถิติพรรณนาจำแนกตามกลุ่มย่อยที่สนใจ ด้วยฟังก์ชัน aggregate

```
aggregate(x, by, FUN, ..., simplify = TRUE, drop = TRUE)
aggregate(formula, data, FUN, subset, na.action = na.omit)
```

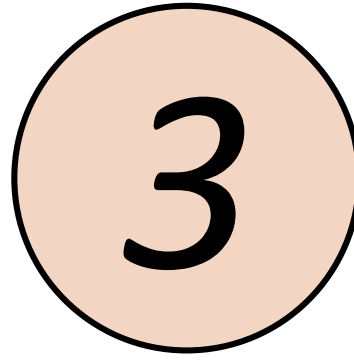
- ตัวอย่าง การหาค่าเฉลี่ยของ PERSONS_BOOSTER_ADD_DOSE (จำนวนคนที่ได้รับวัคซีนบูสเตอร์) ในแต่ละทวีป

```
> aggregate(PERSONS_BOOSTER_ADD_DOSE~WHO.Region,data=covid19,mean)
```

	WHO.Region	PERSONS_BOOSTER_ADD_DOSE
1	Africa	845011.4
2	Americas	8615627.9
3	Eastern Mediterranean	6763861.2
4	Europe	5338640.7
5	South-East Asia	39434313.0
6	Western Pacific	29148361.2

```
> aggregate(PERSONS_BOOSTER_ADD_DOSE~WHO.Region,data=covid19,sd)
```

	WHO.Region	PERSONS_BOOSTER_ADD_DOSE
1	Africa	1399370
2	Americas	23085874
3	Eastern Mediterranean	12376799
4	Europe	10771212
5	South-East Asia	68037691
6	Western Pacific	131182153



3

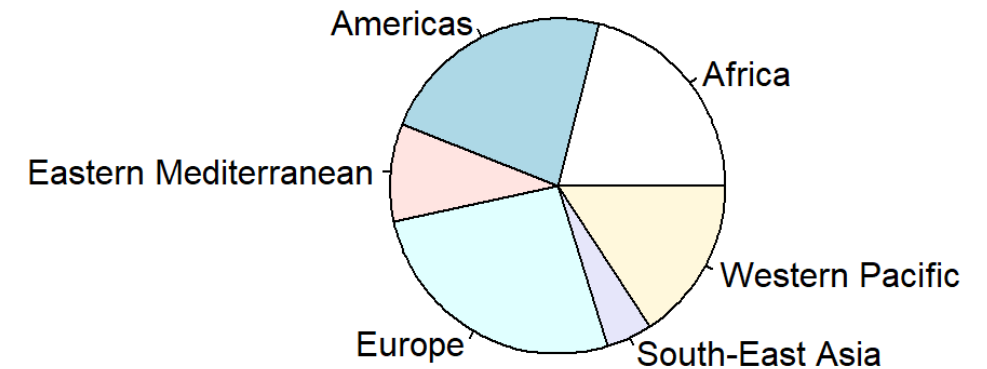
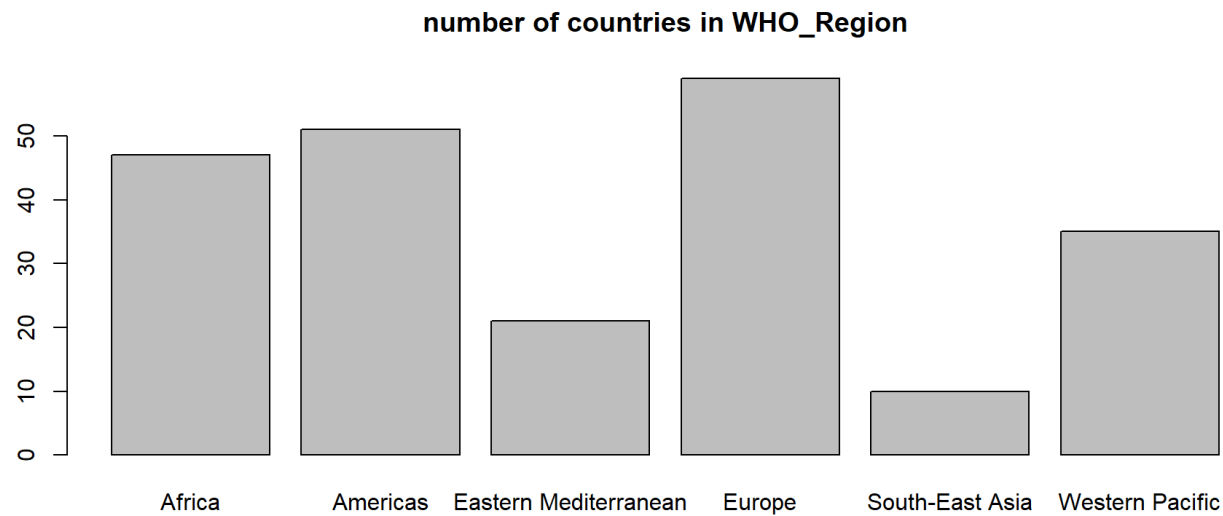
Data presentation

Bar chart / Pie chart/ Histogram / Boxplot

แผนภูมิแท่ง (Bar chart) และแผนภูมิวงกลม (Pie chart)

- การสร้างแผนภูมิแท่ง (Bar chart) และแผนภูมิวงกลม (Pie chart) ด้วยฟังก์ชัน `barplot()` และ `pie()`

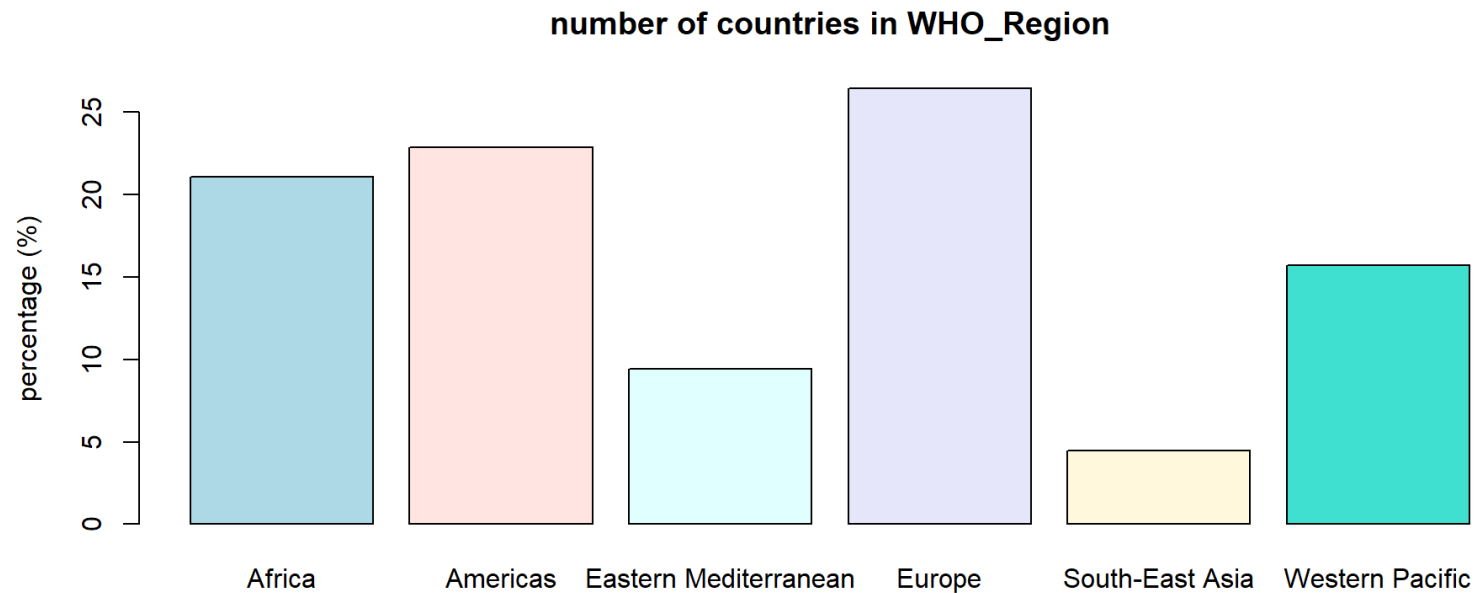
```
n.country <- table(WHO.Region)
barplot(n.country, main="number of country in WHO_Region")
pie(n.country)
```



แผนภูมิแท่ง (Bar chart) และแผนภูมิวงกลม (Pie chart)

- การสร้างแผนภูมิแท่ง (Bar chart) ด้วยฟังก์ชัน `barplot()`

```
percent <- n.country/sum(n.country)*100  
barplot(percent, main="number of countries in WHO_Region",ylab="percentage (%)",  
        col=c(c("lightblue", "mistyrose", "lightcyan",  
                "lavender", "cornsilk","turquoise")))
```

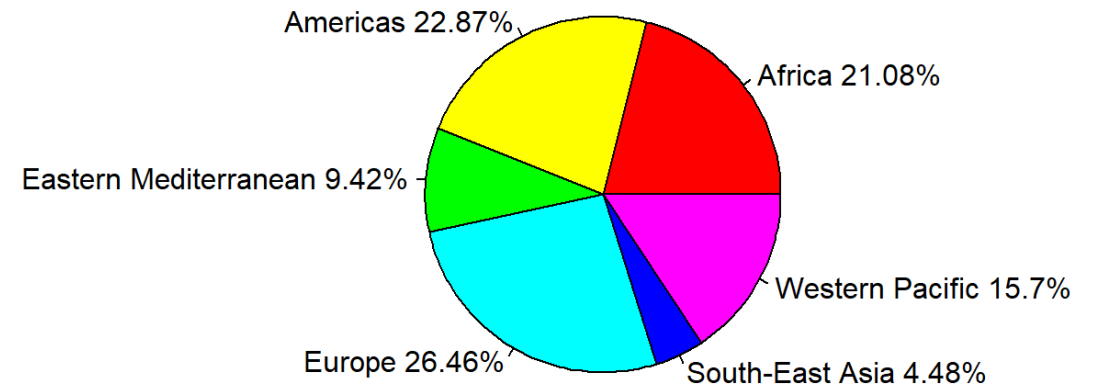


แผนภูมิแท่ง (Bar chart) และแผนภูมิวงกลม (Pie chart)

แผนภูมิวงกลม (Pie chart) ด้วยฟังก์ชัน `pie()` ที่แสดงชื่อกลุ่มและค่าร้อยละของแต่ละกลุ่ม

```
n.country <- table(WHO.Region)
lbls <- names(n.country)
pct <- round(n.country/sum(n.country)*100,2)
# add percents to labels
lbls <- paste(lbls, pct)
# add % to labels
lbls <- paste(lbls,"%",sep="")
pie(n.country,labels = lbls,
    col=rainbow(length(lbls)),
    main="Proportion of country in WHO.region")
```

Proportion of country in WHO.region



Histogram

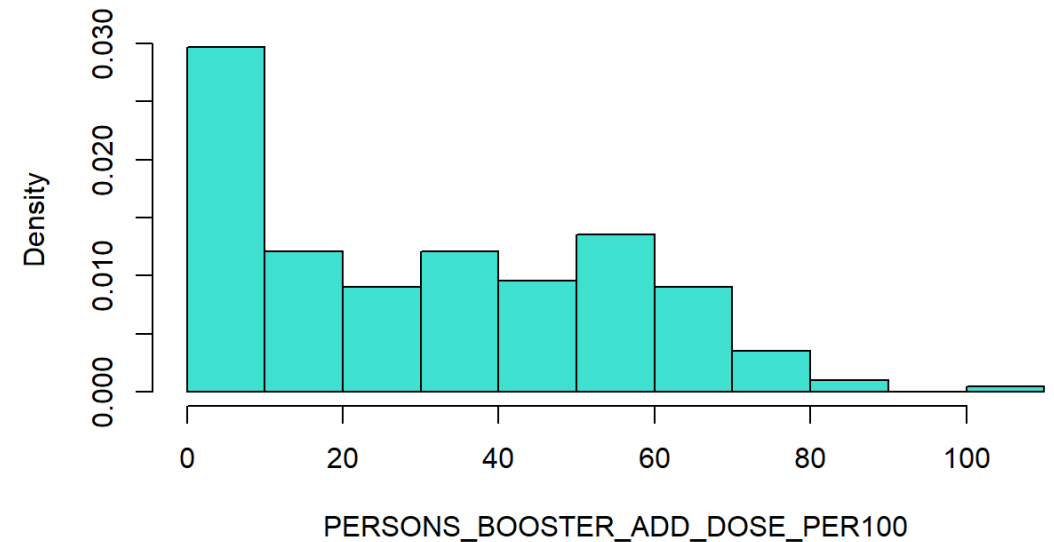
- การสร้างฮิสโทแกรมด้วยฟังก์ชัน hist()

```
hist(PERSONS_BOOSTER_ADD_DOSE_PER100,  
col = "mistyrose")
```

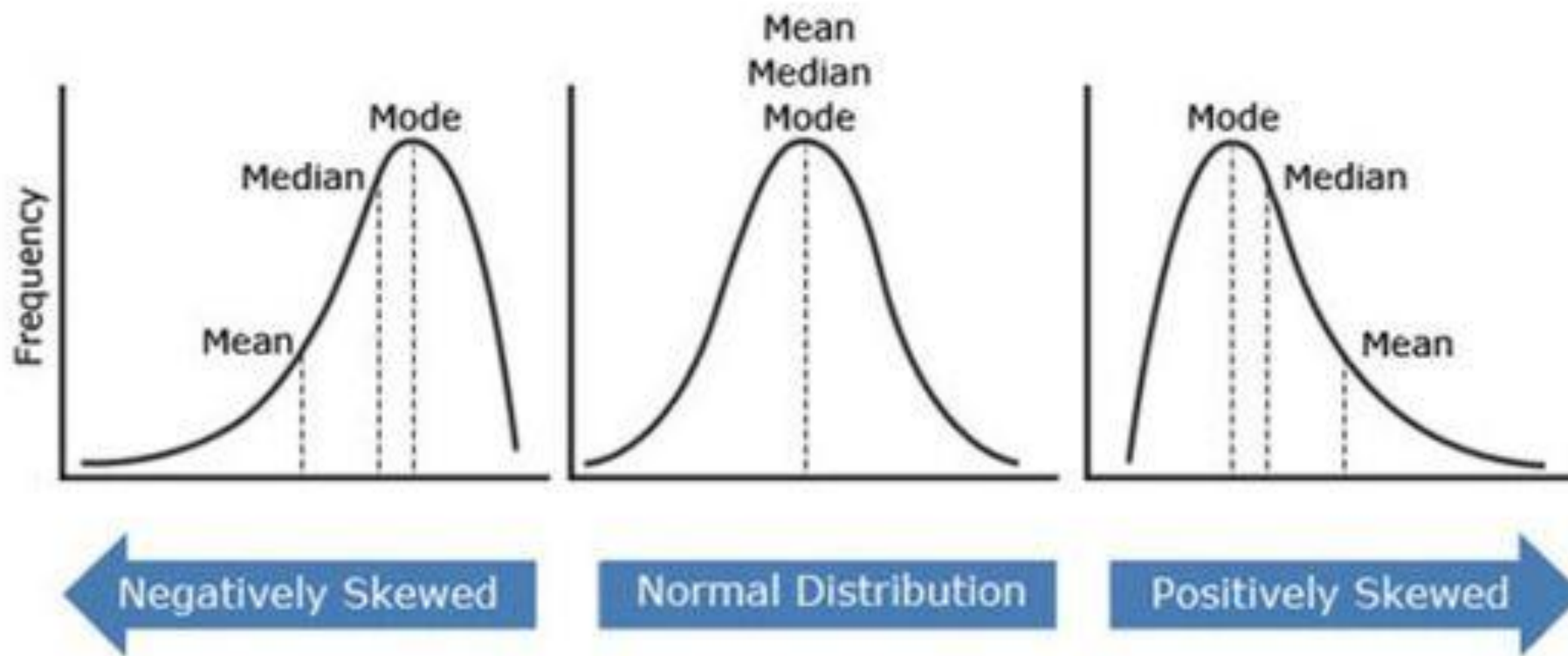
Histogram of PERSONS_BOOSTER_ADD_DOSE_PER100



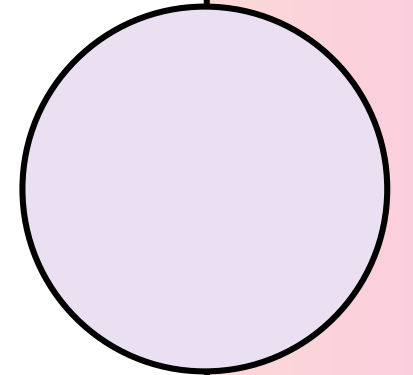
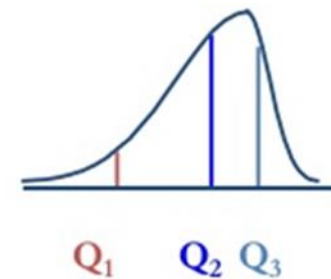
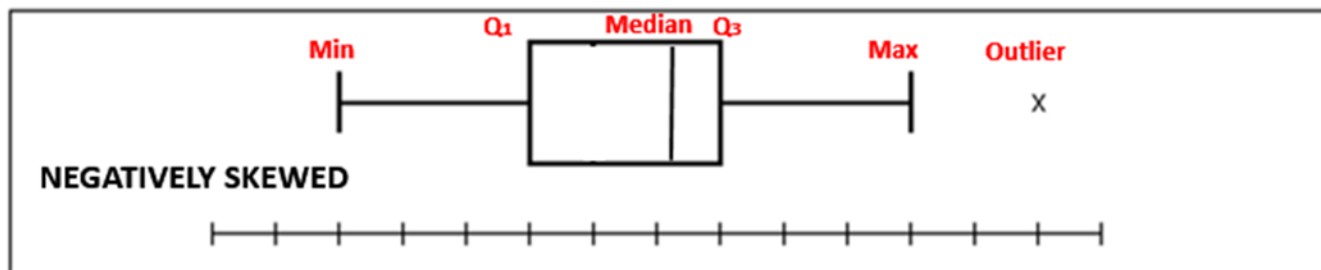
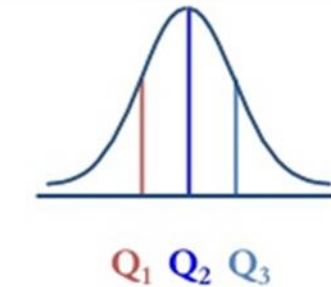
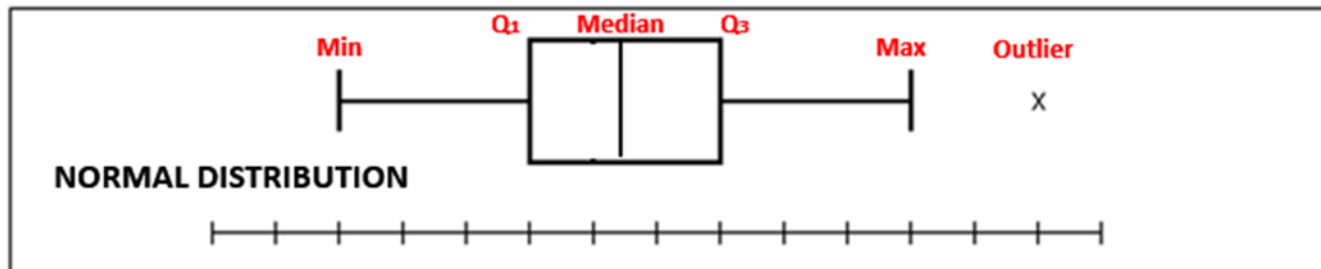
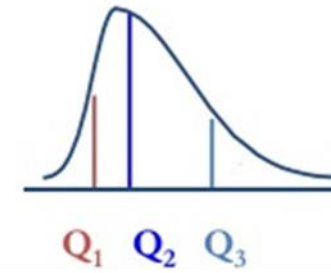
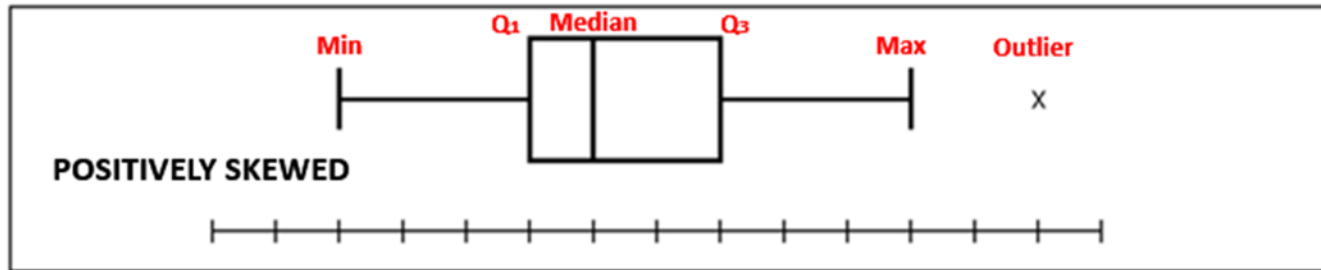
```
hist(PERSONS_BOOSTER_ADD_DOSE_PER100,  
freq = FALSE, col = "turquoise",main=NULL)
```



รูปร่างการแจกแจงของข้อมูล



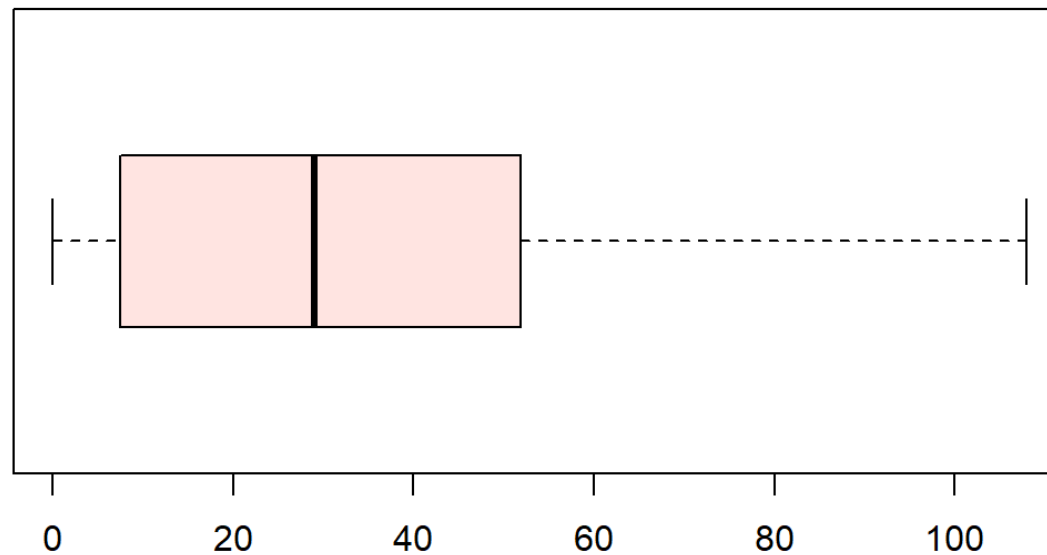
แผนภาพกล่อง (Boxplot)



แผนภาพกล่อง (Boxplot)

```
boxplot(PERSONS_BOOSTER_ADD_DOSE_PER100, horizontal = T, col="mistyrose",  
        main="PERSONS_BOOSTER_ADD_DOSE_PER100")
```

PERSONS_BOOSTER_ADD_DOSE_PER100



แผนภาพกล่อง (Boxplot)

- การแสดงค่าสถิติต่าง ๆ ของ boxplot ด้วยฟังก์ชัน `boxplot.stat()`
- ผลลัพธ์ของฟังก์ชัน `boxplot.stat` อยู่ในโครงสร้างข้อมูลแบบ `list`

```
> boxplot.stats(PERSONS_BOOSTER_ADD_DOSE_PER100)
$stats
[1]  0.0000   7.4835  28.9590  51.8630 107.9220
```

```
$n
[1] 199
```

```
$conf
[1] 23.98835 33.92965
```

```
$out
numeric(0)
```

```
> boxplot.stats(PERSONS_BOOSTER_ADD_DOSE_PER100)
$stats
[1]  0.0000   7.4835  28.9590  51.8630 107.9220
```

```
$n
[1] 199
```

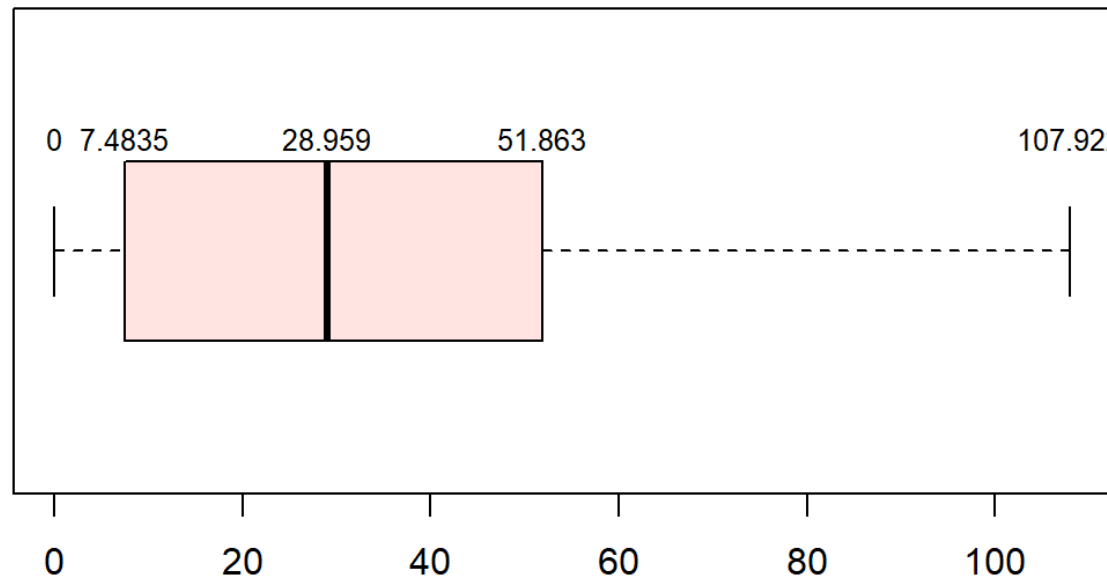
```
$conf
[1] 23.98835 33.92965
```

```
$out
numeric(0)
```

แผนภาพกล่อง (Boxplot)

```
boxplot(PERSONS_BOOSTER_ADD_DOSE_PER100, horizontal = T, col="mistyrose",  
        main="PERSONS_BOOSTER_ADD_DOSE_PER100")  
text(x = boxplot.stats(PERSONS_BOOSTER_ADD_DOSE_PER100)$stats,  
     labels = boxplot.stats(PERSONS_BOOSTER_ADD_DOSE_PER100)$stats, y = 1.25, cex=0.6)
```

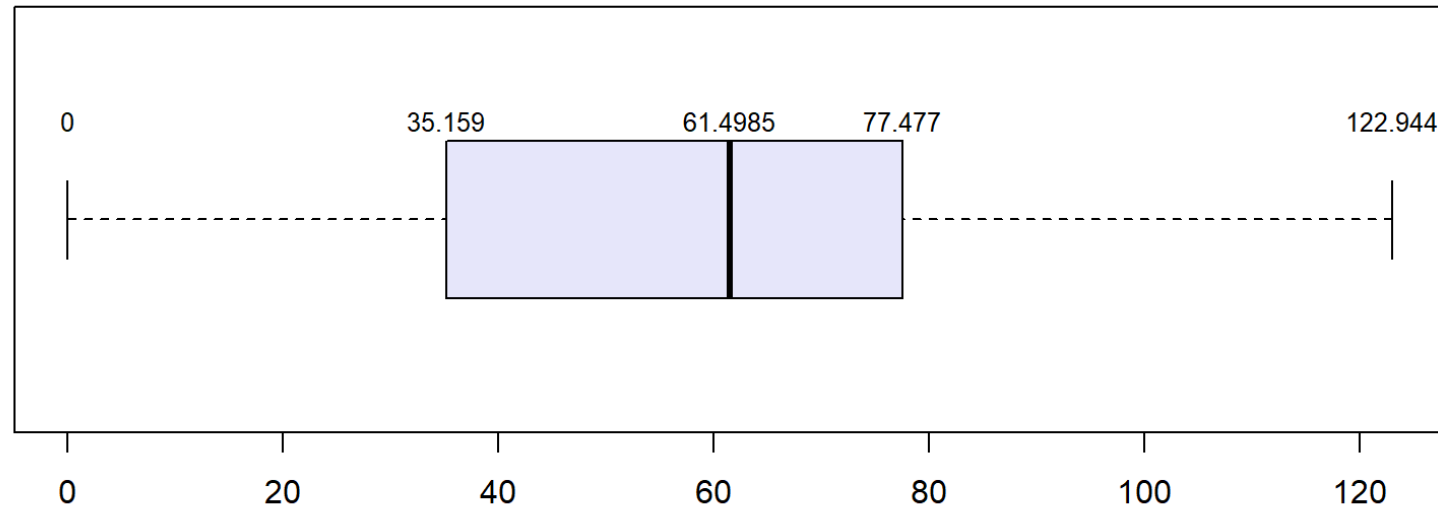
PERSONS_BOOSTER_ADD_DOSE_PER100



แผนภาพกล่อง (Boxplot)

```
boxplot(PERSONS_FULLY_VACCINATED_PER100, horizontal = T, col="lavender",  
        main="TOTAL_VACCINATIONS_PER100")  
text(x = boxplot.stats(PERSONS_FULLY_VACCINATED_PER100)$stats,  
     labels = boxplot.stats(PERSONS_FULLY_VACCINATED_PER100)$stats, y = 1.25, cex=0.8)
```

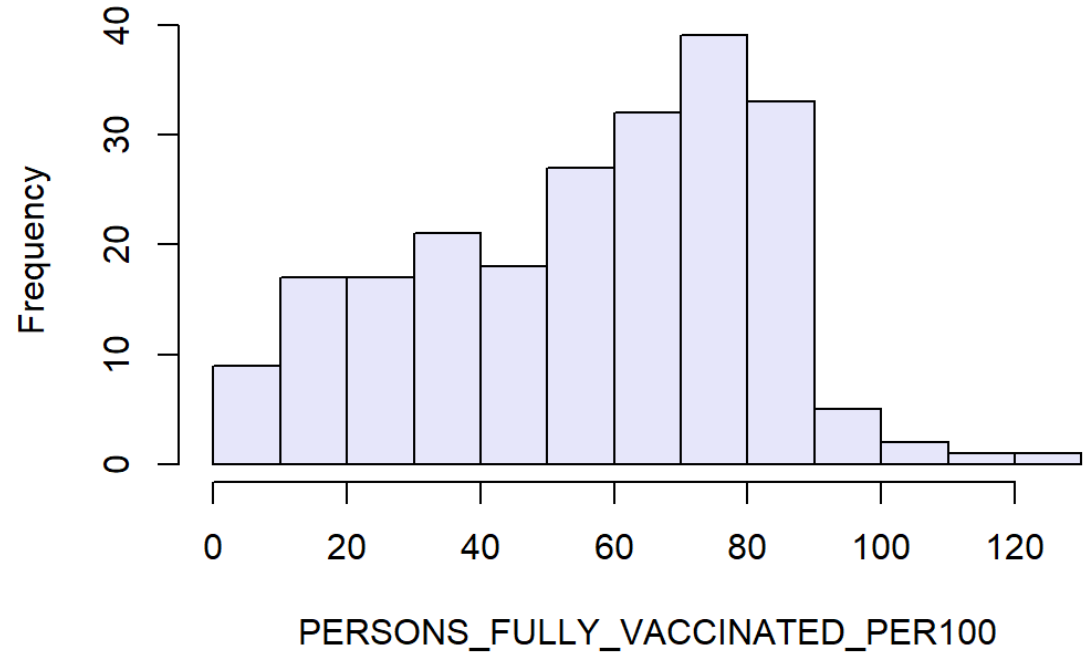
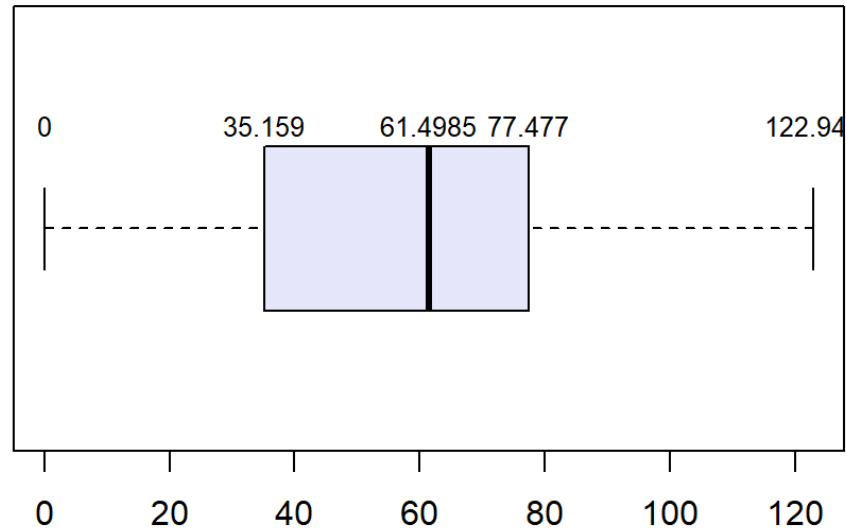
TOTAL_VACCINATIONS_PER100



แผนภาพกล่อง (Boxplot)

```
boxplot(PERSONS_FULLY_VACCINATED_PER100, horizontal = T, col="lavender",  
        main="TOTAL_VACCINATIONS_PER100")  
text(x = boxplot.stats(PERSONS_FULLY_VACCINATED_PER100)$stats,  
     labels = boxplot.stats(PERSONS_FULLY_VACCINATED_PER100)$stats, y = 1.25, cex=0.8)  
hist(PERSONS_FULLY_VACCINATED_PER100, col="lavender", main=NULL)
```

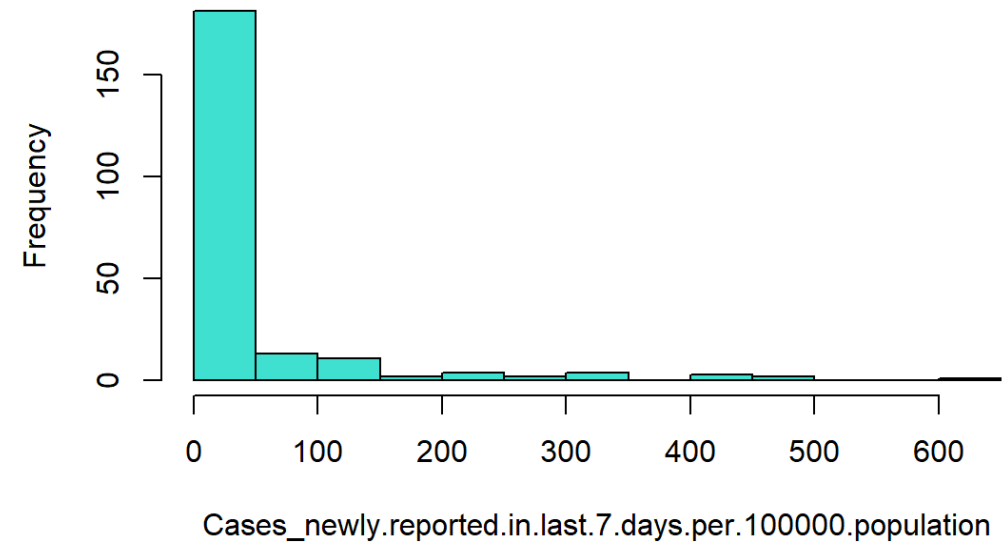
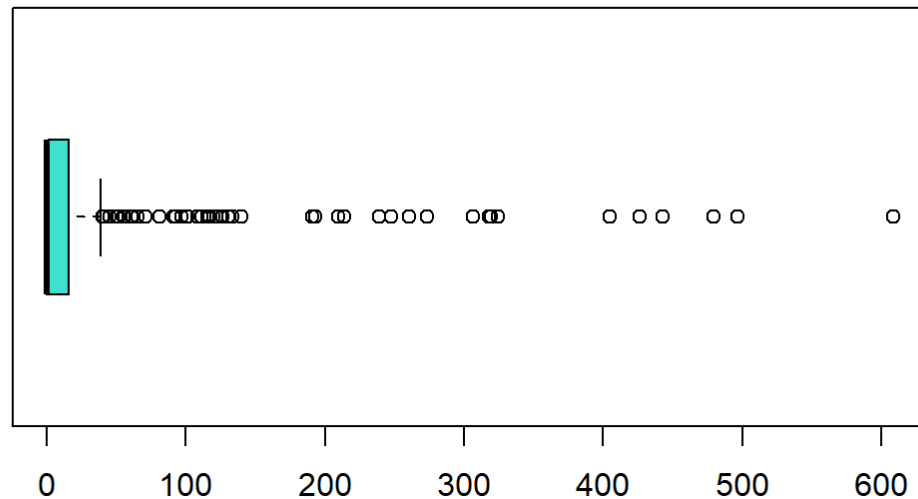
TOTAL_VACCINATIONS_PER100



แผนภาพกล่อง (Boxplot)

```
hist(Cases_newly.reported.in.last.7.days.per.100000.population,col="turquoise",main=NULL)  
boxplot(Cases_newly.reported.in.last.7.days.per.100000.population,horizontal = T,  
        col="turquoise",main="Cases_newly.reported.in.last.7.days.per.100000.population")
```

Cases_newly.reported.in.last.7.days.per.100000.population



แผนภาพกล่อง (Boxplot)

```
boxplot.stats(Cases_newly_reported.in.last.7.days.per.100000.population)
```

```
> boxplot.stats(Cases_newly_reported.in.last.7.days.per.100000.population)
```

```
$stats
```

```
[1] 0.0000 0.0000 0.4400 15.6365 38.6610
```

```
$n
```

```
[1] 223
```

```
$conf
```

```
[1] -1.214414 2.094414
```

```
$out
```

```
[1] 111.305 40.275 426.151 130.996 41.276 81.281 56.103 121.729 115.682 56.134  
[11] 317.678 52.029 117.075 90.891 260.864 404.533 496.512 319.538 92.359 65.103  
[21] 306.746 209.622 247.690 133.828 61.940 101.572 193.257 55.589 239.298 140.149  
[31] 71.171 61.180 442.575 119.964 108.328 479.270 214.133 324.120 97.261 608.460  
[41] 191.144 273.205 50.183 45.398 126.556
```

แผนภาพกล่อง (Boxplot)

```
par(cex.main=0.7) ; par(cex.axis=0.7)
boxplot(PERSONS_FULLY_VACCINATED_PER100~WHO.Region, horizontal = F, col=c("lightblue",
    "mistyrose", "lightcyan", "lavender", "cornsilk", "turquoise"))
```

