# R Programming Basics

# Objective

1. How to program in R and how to use R for effective data analysis

# Exercise(s)

1. Data analysis and visualization in R Studio
2. Case Study: Investigate dengue cases in Malaysia based on weekly statistics provided for each state

# Competencies

* Participants will be able to write program in R using the correct syntax
* Participants will be able to make use of R packages
* Participants will be able to visualize data in R

# Tools – R Studio

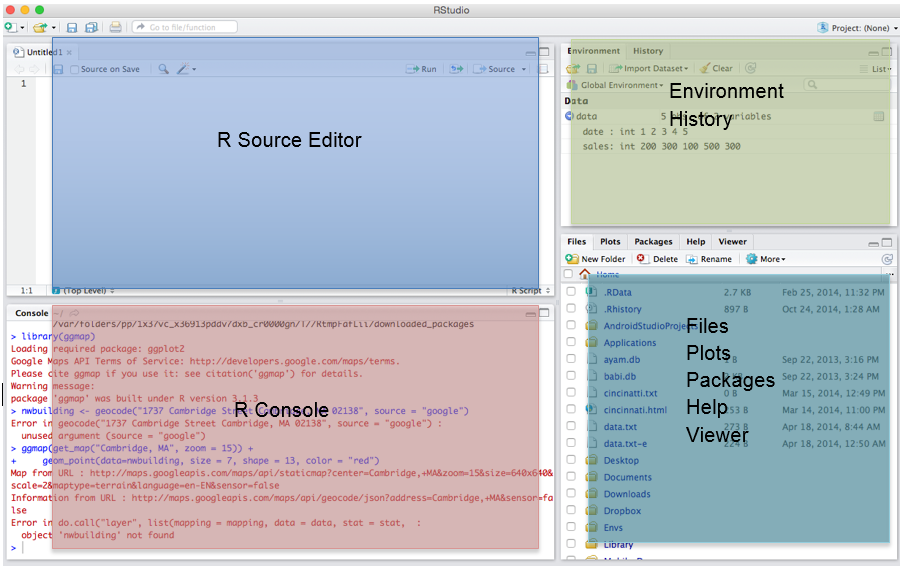
# Data source(s) and fields

1. Dengue\_Cases\_Malaysia.csv (Source: <http://data.gov.my/view.php?view=247>)
   1. State (NEGERI)
   2. Week (Minggu 1 – Minggu 52)

# Step by step guide

1. Download and install R and R Studio
2. Open “dengue.R” script in R Studio
3. Follow Steps 1-20 in order. Execute each code snippet by highlighting the lines and hitting the ‘Run’ button in the R Script Editor, or hitting the CTRL+ENTER buttons (Windows keyboard shortcut). The comments for each Step contain a description/ instruction about the code. View results of the code in the Console and Plots windows.
4. Complete Exercises 1-7 located at the bottom of the script. Each task will require copying and modifying code from the exercise. Instructions and hints are provided in the script comments.

## R Studio



## Import Data

1. Set Working Directory. *Windows computers must use back-slashes "/" not forward slashes.*

setwd("C:/Users/nadav.rindler/Documents/DataViz/DSPL Sept2017/Basic R")

1. Read CSV and view R object type

dengue\_data=read.csv("Dengue\_Cases\_Malaysia\_2011.csv", header = TRUE, sep = ",")

class(dengue\_data)

## Explore Data

1. GET dimension of the data frame - number of rows (1st) & columns (2nd).
2. GET column names.

dim(dengue\_data)

names(dengue\_data)

str(dengue\_data)

1. GET information about the data types for each column.
2. VIEW data set

View(dengue\_data)

1. GET summary of each column that shows minimum, maximum, and quartiles

summary(dengue\_data)

1. Practice selecting elemenets from data frame

# Select first column by index

dengue\_data[,1]

# Select first column by name

dengue\_data[,"NEGERI"]

# Select first row by index

dengue\_data[1,]

#Select rows where condition is TRUE

dengue\_data[dengue\_data$NEGERI == "PERLIS",]

1. COPY the data set, dropping first two columns: NEGERI, Year

weekly\_counts <- dengue\_data[,3:length(dengue\_data)]

1. CALCULATE 2011 total number of dengue cases for each state

dengue\_data$state\_total <- rowSums(weekly\_counts)

1. CALCULATE the highest case across all weeks and states

max(weekly\_counts, na.rm=TRUE)

1. CALCULATE the maximum # of cases for each week across all states

highest\_each\_week <- apply(weekly\_counts, MARGIN=2, max)

1. CHART total dengue cases in each state for the year 2011. Uncomment first line to install ggplot2 package, if needed.

#install.packages("ggplot2")

library(ggplot2)

q <- ggplot(data=dengue\_data, #Specify data set

mapping=aes(x=NEGERI, y=state\_total)) + #Set chart X and Y variable

geom\_bar(fill="red", stat="identity") + #Specify chart type (bar chart) and color

theme(axis.text.x = element\_text(angle = 90, hjust = 1)) + #Rotate X axis text 90 degrees

labs(title="Dengue Cases in Malaysia : 2011",

y="Total Cases", x="Negeri") #Set chart and axis labels

q

## Transform Data

1. TRANSPOSE data from wide to long

t\_matrix <- t(as.matrix(weekly\_counts)

dengue\_state <- as.data.frame(t\_matrix)

1. RENAME variable and adjust date type

names(dengue\_state) <- dengue\_data$NEGERI

1. CREATE and assign new date variable for the 52 weeks

weeks <- as.Date(paste(2011,1:52,1,sep="-"),format="%Y-%U-%u")

weeks

dengue\_state$date <- weeks

## Visualize Data

1. PLOT weekly data for Selangor using ggplot()

t <- ggplot(data=dengue\_state, #Specify data set

aes(x=date, y=SELANGOR)) + #Set chart's X and Y variables

geom\_line(color="red", size=1) + #Set chart type (line) and line color and size

theme(axis.text.x = element\_text(angle = 90, hjust = 1)) + #Rotate X axis text 90 degrees

labs(y="Total Cases",x="Week",title="Dengue Cases in SELANGOR : 2011") #Set chart and axis labels

t

1. PLOT first 10 weeks of data

r <- ggplot(data=dengue\_state[1:10,], #Specify data set -- SUBSET DATA -- take first 10 observations

aes(x=date, y=SELANGOR)) + #Set chart's X and Y variables

geom\_point(color="red", size=1) + #Set chart type (line) and line color and size

theme(axis.text.x = element\_text(angle = 90, hjust = 1)) + #Rotate X axis text 90 degrees

labs(y="Total Cases",x="Week",title="Dengue Cases in SELANGOR : 2011") #Set chart and axis labels

r

1. SUM weekly data in each state by month. Uncomment first two lines to install dplyr and lubridate packages, if needed.

#install.packages("dplyr")

#install.packages("lubridate")

library(dplyr)

library(lubridate)

dengue\_month <- dengue\_state %>%

group\_by(month\_count = month(date)) %>%

summarize\_each(funs(sum)) %>%

mutate(Month=as.Date(paste("2011",month\_count,"1",sep="-"),format="%Y-%m-%d"))

1. PLOT monthly Selangor dengue cases from aggregate data set

ggplot(data=dengue\_month, #Specify monthly data set

aes(x=Month, y=SELANGOR)) + #Set chart's X and Y variables

geom\_line(color="red", size=1) + #Set chart type (line) and line color and size

scale\_x\_date(date\_labels="%Y-%m", date\_breaks = "1 month") +

theme(axis.text.x = element\_text(angle = 90, hjust = 1)) + #Rotate X axis text 90 degrees

labs(y="Total Cases",x="Month",title="Dengue Cases in SELANGOR : 2011") #Set chart and axis labels

## Independent Exercises

1. CALCULATE total dengue cases in Malaysia for the year 2011
   1. Hint: Sum across the state totals calculated in Step #10
2. CALCULATE the maximum # of cases for each state across all weeks
   1. Hint: Modify code in Step #12
   2. Help page: ?apply
   3. Change the MARGIN argument in the apply function to apply over COLUMNS instead of ROWS
3. CALCULATE the average of dengue cases for each state for the whole year
   1. Hint: use the weekly\_counts object
   2. Hint: ?rowMeans
   3. Alternatively, can use the same apply function from Exercise #2, but change the function being applied
4. PLOT \*last\* 12 weeks of data for PERLIS state **as a LINE graph**
   1. Hint: Modify code from Step #18
   2. Make sure to update both the axis text and chart title
5. PLOT monthly dengue cases in PERLIS **as a BARPLOT**
   1. Hint: Modify code from Step #20 - change state in the Y Variable
   2. Hint: Change the chart type from geom\_line() to geom\_bar() -- see Step #13
   3. Make sure to update the chart title
   4. ADVANCED: Update the X axis labels to list every THIRD month (3 months)
      1. Hint: try modifying the ?scale\_x\_date function (See step #20)
6. ADVANCED: Which month in 2011 has the highest number of dengue cases across all states?
   1. Hint: Create monthly total in "dengue\_month" dataset using ?rowSums
   2. Hint: Use ?which.max to find which row in the monthly\_total column has the highest value
7. ADVANCED: For each month, list the state with the most dengue cases
   1. Hint: use dengue\_month dataset
   2. Hint: Combine ?apply with ?which.max across the state columns
   3. Hint: Use names(dengue\_month) to identify the column name associated with the column index

**See “dengue\_answer\_key.R” script for answers.**