**ปฏิบัติการที่ 1**

**การใช้โปรแกรม R เบื้องต้น**

**วัตถุประสงค์:** 1. เพื่อให้นักศึกษาฝึกเขียนคำสั่งต่าง ๆ ในโปรแกรม R

2. เพื่อให้นักศึกษาฝึกสำรวจข้อมูล

**Data file :** “childPattani\_practice.xlsx”, “nomotercycle\_practice.xlsx”

**สรุปเนื้อหา:**

การจัดเตรียมข้อมูล (Data preparation) และทำความสะอาดข้อมูล (Data cleansing) เป็นขั้นตอนสำคัญในกระบวนการวิเคราะห์ข้อมูล งานหลัก ๆ ในการจัดเตรียมข้อมูล ได้แก่ การทำความสะอาดข้อมูล เป็นการขจัดข้อมูลที่ผิดพลาด ไม่สมบูรณ์ หรือซ้ำซ้อน การจัดรูปแบบข้อมูล เป็นการแปลงข้อมูลให้อยู่ในรูปแบบที่สอดคล้องกัน เช่น แปลงวันที่ให้อยู่ในรูปแบบเดียวกัน การจัดการข้อมูลที่ขาดหายไป เป็นการประมาณค่าหรือเติมข้อมูลที่ขาดหายไป ส่วนใหญ่ใช้กับข้อมูลอนุกรมเวลา (Time series data) การจัดการค่าสุดโต่ง เป็นการตรวจจับและจัดการกับค่าที่ผิดปกติ ข้อมูลตัวเลขที่มีความเป็นไปได้น้อยมาก เช่น คนอายุ 120 ปี หรือ ส่วนสูง 230 ซม. (Outlier) การแปลงข้อมูลเป็นการแปลงข้อมูลให้เป็นรูปแบบที่เหมาะสมกับการวิเคราะห์ เช่น แปลงข้อมูลตัวอักษรเป็นตัวเลข การรวมข้อมูลเป็นการรวมข้อมูลจากหลายแหล่งเข้าด้วยกัน และการจัดระเบียบข้อมูลเป็นการจัดระเบียบข้อมูลให้เข้าใจง่าย ค้นหาง่าย

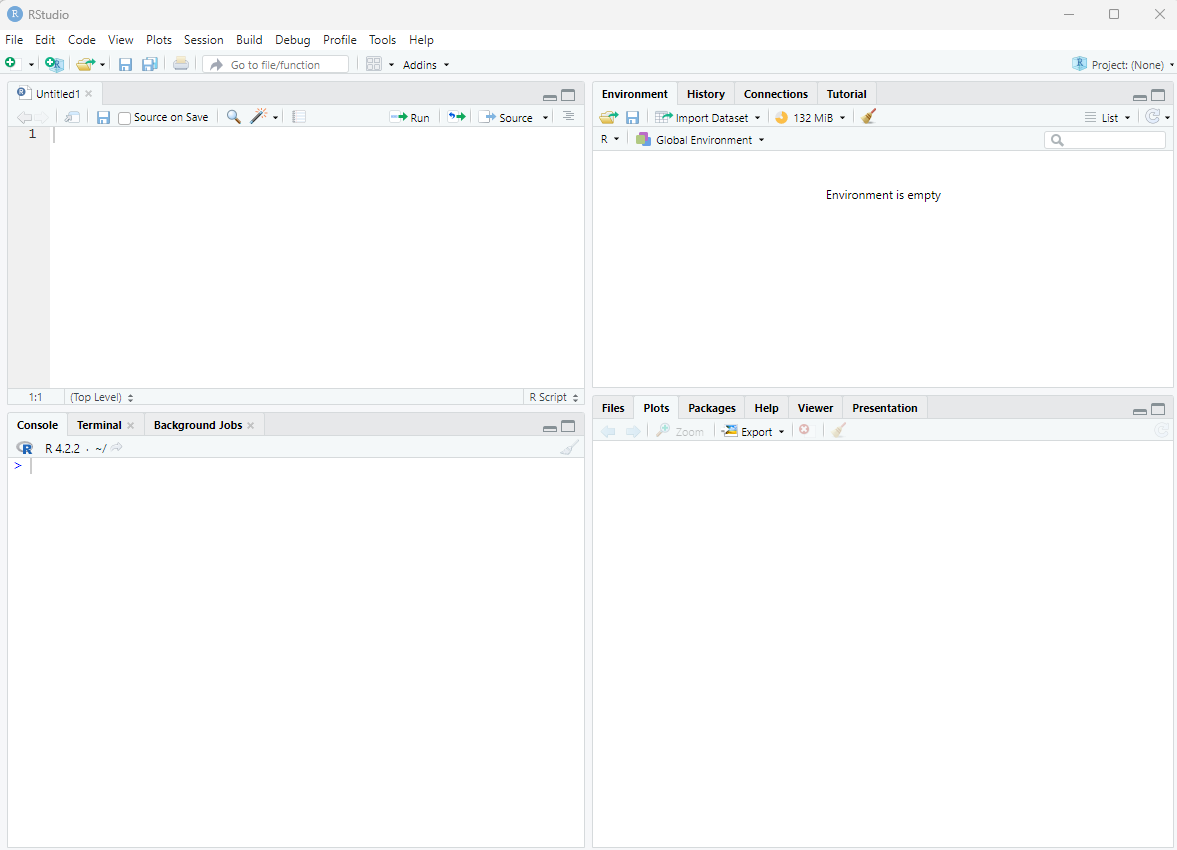
บทนี้จะให้นักศึกษาฝึกสำรวจข้อมูล เพื่อให้ข้อสังเกตว่าข้อมูลมีข้อผิดพลาด ไม่สมบูรณ์ หรือค่าที่ผิดปกติหรือไม่ อย่างไร โดยหนึ่งในเทคนิคที่ต้องใช้ในการสำรวจข้อมูล คือ นักศึกษาต้องเข้าประเภทของตัวแปรเพื่อให้ทำการสำรวจได้ง่ายขึ้น ในที่นี้เราแบ่งประเภทของข้อมูลออกเป็น 2 ประเภท คือ ตัวแปรกลุ่ม และตัวแปรต่อเนื่อง

ข้อมูลประเภทกลุ่ม (Categorical data) เป็นข้อมูลที่สามารถจำแนกออกเป็นกลุ่มหรือหมวดหมู่ที่แตกต่างกันได้ จัดเก็บด้วยอักขระ หรือตัวเลข (รหัส) และมีคำอธิบายรหัส เช่น สีผม (บลอนด์ น้ำตาม ดำ) ประเภทหนัง (ตลก แอคชั่น แฟนตาซี วิทยาศาสตร์) ข้อมูลประเภทกลุ่มสามารถแบ่งเป็น 2 ประเภท คือ นามบัญญัติ (Nominal data) และ เรียงอันดับ (Ordinal data) ข้อมูลนามบัญญัติไม่สามารถเปรียบเทียบค่าข้อมูลระหว่างกลุ่มเหมือนกันหรือแตกต่างกันได้ ไม่สามารถจัดเรียงลำดับได้ ไม่สามารถใช้การดำเนินการทางคณิตศาสตร์ได้ เช่น การบวก ลบ คูณ หาร ข้อมูลเรียงลำดับเป็นข้อมูลที่สามารถเปรียบเทียบค่าข้อมูลระหว่างกลุ่มเหมือนกันหรือแตกต่างกันได้ และจัดเรียงลำดับของข้อมูลได้ เช่น ระดับความพึงพอใจ (มากที่สุด มาก ปานกลาง น้อย น้อยที่สุด) สามารถใช้การดำเนินการทางคณิตศาสตร์ได้ เช่น การบวก ลบ คูณ หาร

ข้อมูลประเภทต่อเนื่อง (Continuous data) เป็นข้อมูลเชิงตัวเลขที่มาจากการวัด เช่น ส่วนสูง น้ำหนัก คะแนนสอบ เป็นต้น สามารถเปรียบเทียบและใช้การดำเนินการทางคณิตศาสตร์ได้

ให้นักศึกษาฝึกปฏิบัติดังนี้

1. เปิดหน้าต่างโปรแกรม RStudio



(1) เขียนและเก็บคำสั่ง (Command)

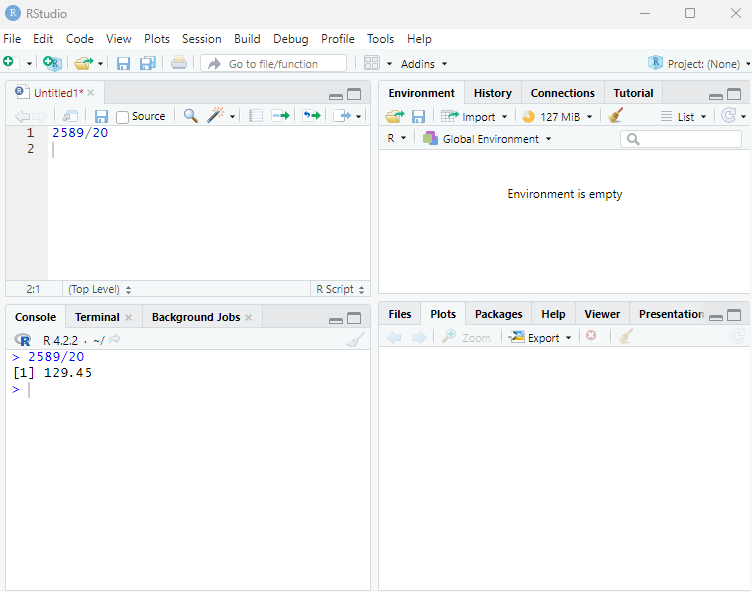
(2) รัน (Execute) คำสั่ง (Command)

(3) แสดงชื่อตัวแปร หรือ object ต่าง ๆ

(4) แสดงกราฟ

การรันคำสั่งเพียง 1 คำสั่ง

* พิมพ์คำสั่ง 2589/20 (ให้วางเคอร์เซอร์ตรงบรรทัดที่ต้องการรันคำสั่ง)
* กดปุ่ม Run  หรือกดปุ่มที่คีย์บอร์ด Ctrl+Enter



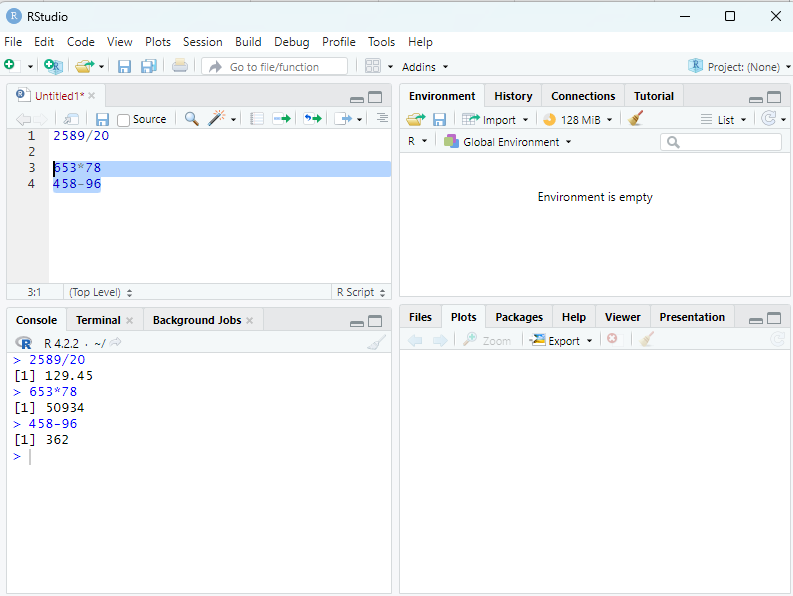
การรันคำสั่งมากกว่า 1 คำสั่ง

* พิมพ์คำสั่ง

653\*78

458-96

* คลุมทั้ง 2 บรรทัด กดปุ่ม Ctrl+Enter ให้สังเกตุผลลัพธ์



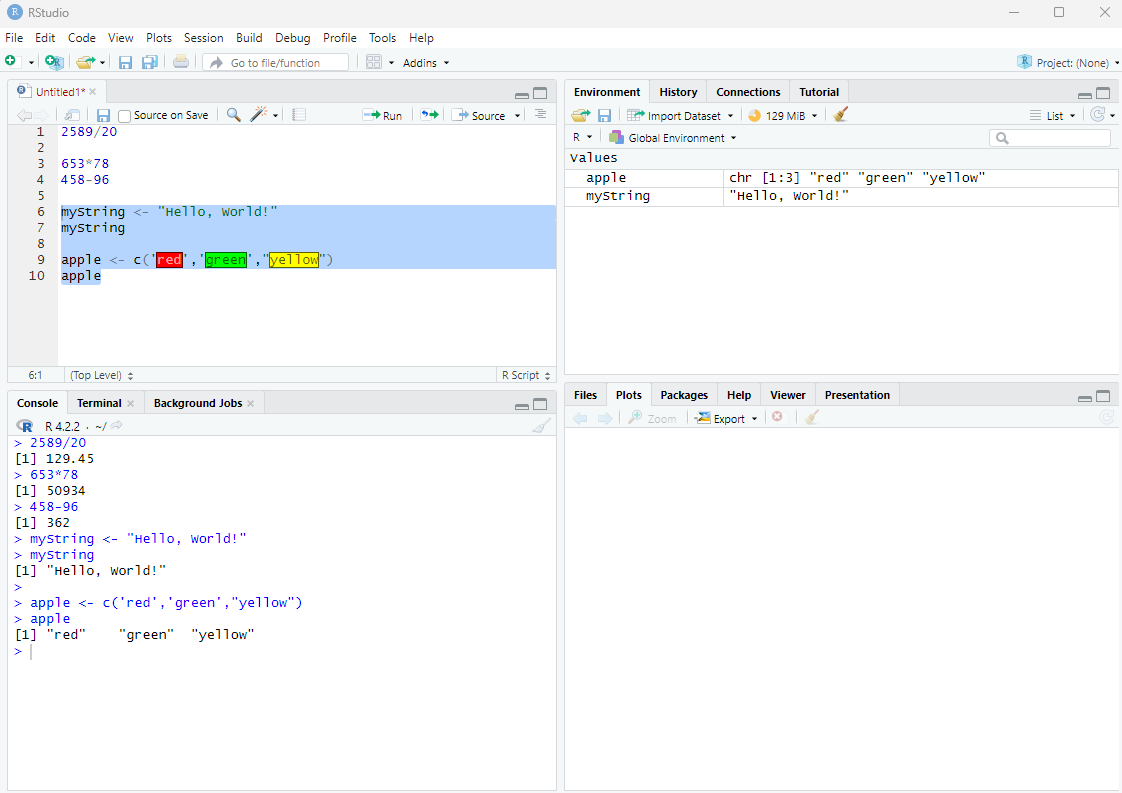
พิมพ์คำสั่ง

myString <- "Hello, World!"

myString

apple <- c('red','green',"yellow")

apple



1. ให้นักศึกษาหาคำตอบของข้อต่อไปนี้ โดยใช้โปรแกรม R

| **คำถาม** | **แสดงคำสั่ง (Commands) และคำตอบ (Output)** |
| --- | --- |
|  | 29.30119 |
|  | 697936 |
|  | 2 |
|  | 5.416174e+15 |
|  | 2.747271 |
|  | 1.193125 |
|  | 3.963474 |

1. ให้นักศึกษาหาคำตอบของข้อต่อไปนี้ โดยใช้โปรแกรม R

Vector

| **คำถาม** | **เขียนคำอธิบาย** |
| --- | --- |
| > vt1 <- c(1, 2, 3, 4, 5)  > vt1 | 1 2 3 4 5 |
| > vt2 <- c(5:10)  > vt2  > vt2[4] | 5 6 7 8 9 10  8 |
| > vt3 <- seq(from=1, to=10, by=2) | 1 3 5 7 9 |
| > x1 <- c(1:5) | 1 2 3 4 5 |
| > x2 <- c(11:15) | 11 12 13 14 15 |
| > xc1 <- cbind(x1,x2) | x1 x2  [1,] 1 11  [2,] 2 12  [3,] 3 13  [4,] 4 14  [5,] 5 15 |
| > xr1 <- rbind(x1,x2) | [,1] [,2] [,3] [,4] [,5]  x1 1 2 3 4 5  x2 11 12 13 14 15 |
| > x3 <- c(1:6) | 1 2 3 4 5 6 |
| > x4 <- c(11:18) | 11 12 13 14 15 16 17 18 |
| > xc2 <- cbind(x3,x4) | x3 x4  [1,] 1 11  [2,] 2 12  [3,] 3 13  [4,] 4 14  [5,] 5 15  [6,] 6 16  [7,] 7 17  [8,] 8 18 |
| xr2 <- rbind(x1,x2) | [,1] [,2] [,3] [,4] [,5]  x1 1 2 3 4 5  x2 11 12 13 14 15 |

List

| **คำถาม** | **เขียนคำอธิบาย** |
| --- | --- |
| > list1 <- list(“male”, 24, “single”, 72.5)  > list1  > class(list1) | [[1]]  [1] "male"  [[2]]  [1] 24  [[3]]  [1] "single"  [[4]]  [1] 72.5  "list" |
| > list2 <- list(c("female","male"), seq(1:10),c(24,57,29,15), "single", 72.5) | [[1]]  [1] "female" "male"  [[2]]  [1] 1 2 3 4 5 6 7 8 9 10  [[3]]  [1] 24 57 29 15  [[4]]  [1] "single"  [[5]]  [1] 72.5 |

Matrix

| **คำถาม** | **เขียนคำอธิบาย** |
| --- | --- |
| > v1 <- c(1:12)  > matrix1 <- matrix(v1, nrow=3)  > matrix1 | [,1] [,2] [,3] [,4]  [1,] 1 4 7 10  [2,] 2 5 8 11  [3,] 3 6 9 12 |
| > matrix2 <- matrix(v1, ncol=2)  > matrix2 | [,1] [,2]  [1,] 1 7  [2,] 2 8  [3,] 3 9  [4,] 4 10  [5,] 5 11  [6,] 6 12 |
| > class(matrix1);class(matrix2) | [1] "matrix" "array"  [1] "matrix" "array" |

Data frame

| **คำถาม** | **เขียนคำอธิบาย** |
| --- | --- |
| > x <- c(1,2,3)  > y <- c(10,20,30)  > z <- c(100,200,300)  > t1 <- data.frame(x,y,z)  > class(t1)  > t1 | "data.frame"  x y z  1 1 10 100  2 2 20 200  3 3 30 300 |

1. ให้นักศึกษาหาคำตอบของข้อต่อไปนี้ โดยใช้โปรแกรม R

| **คำถาม** | **แสดงคำสั่ง (Commands) และคำตอบ (Output)** |
| --- | --- |
| ให้นักศึกษาสร้างชุดข้อมูล (data frame) ชื่อ “student” จากตัวแปร studid, weight, age และ gender  > studid <- c(1:10)  > weight <- c(45, 50, 43, 60, 58, 40, 38, 30, 35, 32)  > age <- c(25, 18, 30, 45, 22, 50, 36, 42, 32, 28)  > gender <- c(1, 2, 1, 1, 1, 2, 2, 1, 1, 2) | studid weight age gender  1 1 45 25 1  2 2 50 18 2  3 3 43 30 1  4 4 60 45 1  5 5 58 22 1  6 6 40 50 2  7 7 38 36 2  8 8 30 42 1  9 9 35 32 1  10 10 32 28 2 |
| จาก data frame student แสดงข้อมูลเฉพาะตัวแปร studid | studid  1  2  3  4  5  6  7  8  9  10 |
| จาก data frame student แสดงข้อมูลแถวที่ 6 ของตัวแปรลำดับที่ 2 และ 3 | weight age  6 40 50 |
| จาก data frame student แสดงข้อมูลเฉพาะของผู้ชาย | studid weight age gender  1 45 25 1  3 43 30 1  4 60 45 1  5 58 22 1  8 30 42 1  9 35 32 1 |
| จาก data frame student แสดงข้อมูลเฉพาะคนที่มีอายุมากกว่า 30 ปี | studid weight age gender  3 3 43 30 1  4 4 60 45 1  6 6 40 50 2  7 7 38 36 2  8 8 30 42 1  9 9 35 32 1 |
| จาก data frame student แสดงข้อมูลสำหรับตัวแปร weightและ genderเฉพาะผู้หญิง | weight gender  2 50 2  6 40 2  7 38 2  10 32 2 |
| จาก data frame student แสดงข้อมูลทุกคอลัมน์เฉพาะผู้หญิงที่มีอายุน้อยกว่า 25 ปี | studid weight age gender  2 50 18 2 |
| จาก data frame student ลบข้อมูลของคนที่มีน้ำหนักมากกว่า 50 ปี | studid weight age gender  1 1 45 25 1  2 2 50 18 2  3 3 43 30 1  6 6 40 50 2  7 7 38 36 2  8 8 30 42 1  9 9 35 32 1  10 10 32 28 2 |

1. ให้ นศ. สร้างชุดข้อมูล (data frame) ข้างล่างนี้เก็บไว้ในตัวแปร pet

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **petid** | **petname** | **pettype** | **petweight** |  |
| 1 | Salapou | 1 | 12 | คำอธิบายตัวแปร |
| 2 | Namwan | 2 | 15 | petid คือ รหัสสัตว์เลี้ยง |
| 3 | Milo | 2 | 14 | petname คือ ชื่อสัตว์เลี้ยง |
| 4 | Jacky | 1 | 13.5 | pettype คือ ประเภทของสัตว์เลี้ยง โดยที่ 1 คือ cat และ 2 คือ dog |
| 5 | Fufu | 1 | 11.6 |
| 6 | Khaimook | 2 | 16 | petweight คือ น้ำหนักของสัตว์เลี้ยง |

| **คำถาม** | **แสดงคำสั่ง (Commands) และคำตอบ (Output)** |
| --- | --- |
| สร้าง data frame |  |
| กำหนดประเภทของตัวแปรต่าง ๆ เป็นดังนี้  petid -> integer  petname -> character  pettype -> factor พร้อมให้คำอธิบายของรหัส 1 และ 2  petweight -> numeric |  |
| แสดงข้อมูลของตัวแปร petname ทุกเรคคอร์ด |  |
| แสดงข้อมูลของตัวแปร petid และ petname  ทุกเรคคอร์ด (ใช้ 1 คำสั่ง) |  |
| แสดงข้อมูลเฉพาะแถวที่ 4 ทุกคอลัมน์ |  |
| แสดงข้อมูลเฉพาะแถวที่ 2, 3, 4, 6 ของทุกคอลัมน์ |  |
| เพิ่มตัวแปร petheight (ส่วนสูง) ดังนี้ (50, 60, 55, 57, 63, 57) |  |
| แก้ไขข้อมูล pettype ของ Fufu ให้เป็น 2 และ pettype ของ Khaimook ให้เป็น 1 |  |
| เพิ่มข้อมูล 1 แถว ดังนี้  petid = 7  petname = Slerpy  pettype = 1  petweight = 9  petheight = 77 |  |
| #ให้นักศึกษา backup ข้อมูลก่อน  > pet1 <- pet  ลบข้อมูลแถวที่ 1 แถวที่ 4 และแถวที่ 6 |  |
| ลบข้อมูลคอลัมน์ที่ 3 และคอลัมน์ที่ 4 |  |

1. ให้นักศึกษาฝึกสำรวจข้อมูลโดยใช้ชุดข้อมูล “childPattani\_practice.xlsx” มีคำอธิบายตัวแปรดังนี้

| ชื่อตัวแปร | คำอธิบาย |
| --- | --- |
| perid\_pre | รหัสบุคคล |
| titnme | คำนำหน้าชื่อ |
| age | อายุ |
| gender | เพศ |
| adr | ที่อยู่ |
| vllid | หมู่ที่ |
| tmb | ตำบล |
| amp | อำเภอ |
| prv | จังหวัด |
| plcid | รหัสที่อยู่ |
| hholdid | รหัสบ้าน |
| comment | ประเภทของบุคคล |

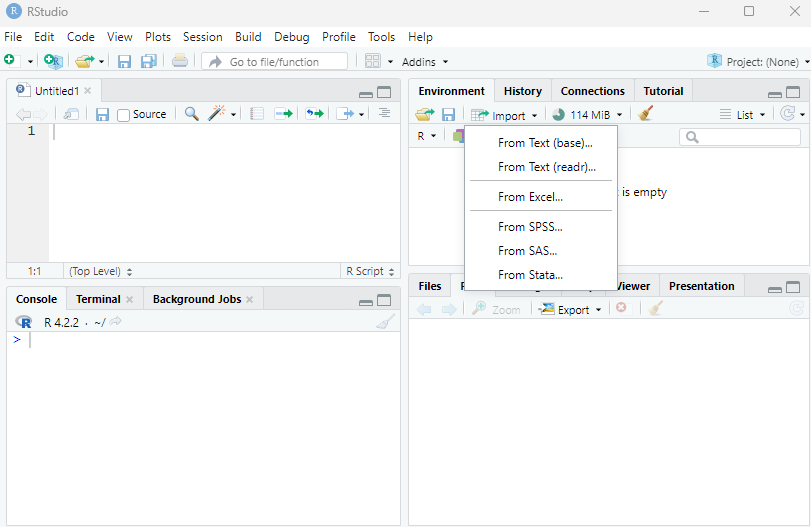
**คำถาม:** ให้ระบุประเภทของตัวแปร (ตัวแปรกลุ่ม หรือ ตัวแปรต่อเนื่อง) ต่อไปนี้

| ชื่อตัวแปร | ประเภทตัวแปร |
| --- | --- |
| age | ต่อเนือง |
| gender | กลุ่ม |
| tmb | กลุ่ม |
| amp | กลุ่ม |
| prv | กลุ่ม |
| comment | กลุ่ม |

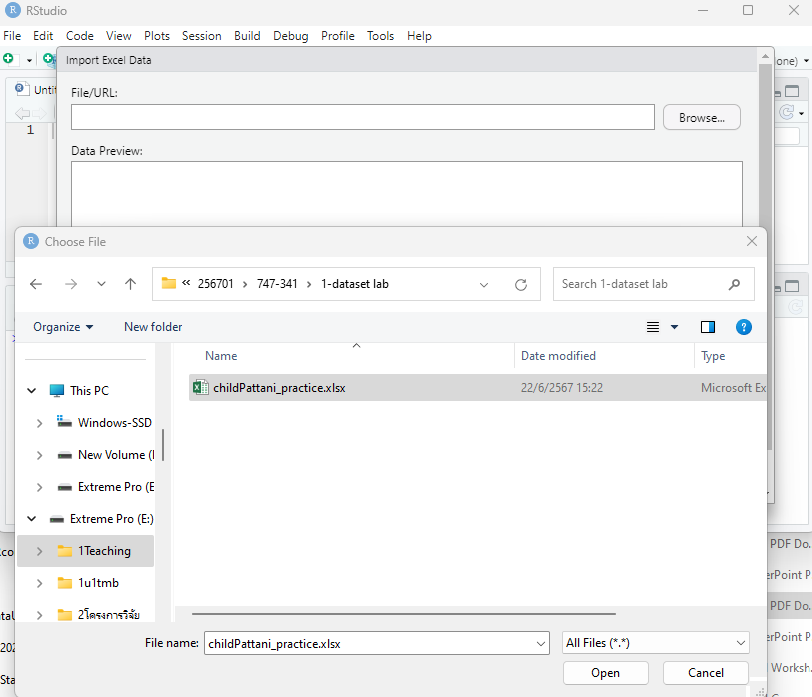
การสำรวจข้อมูล

นำเข้าข้อมูลจาก Excel “childPattani\_practice.xlsx”

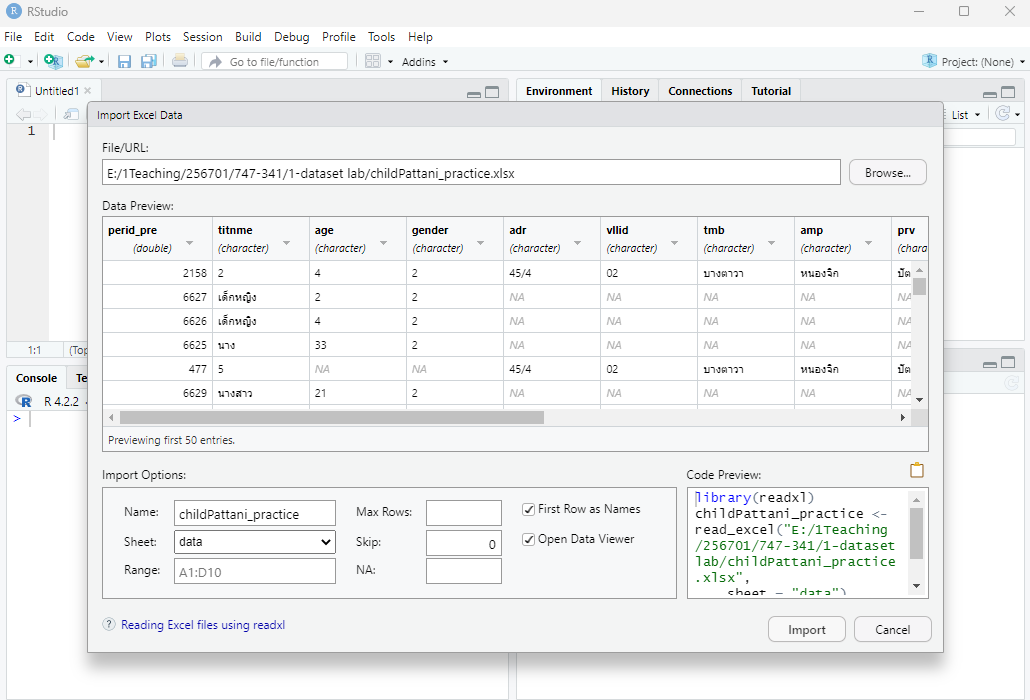
* ไปที่ Import -> From Excel…



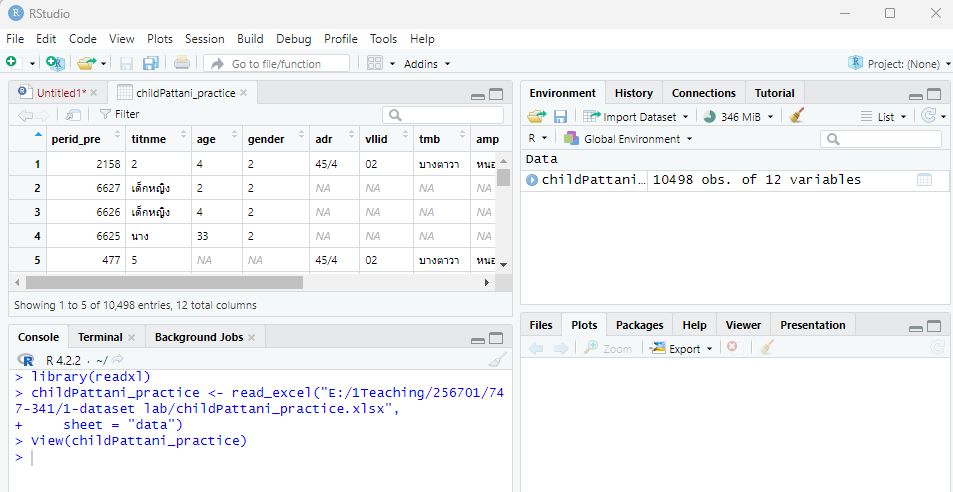
* Browse… --> ไปยังข้อมูลที่ต้องการนำเข้า
* Open



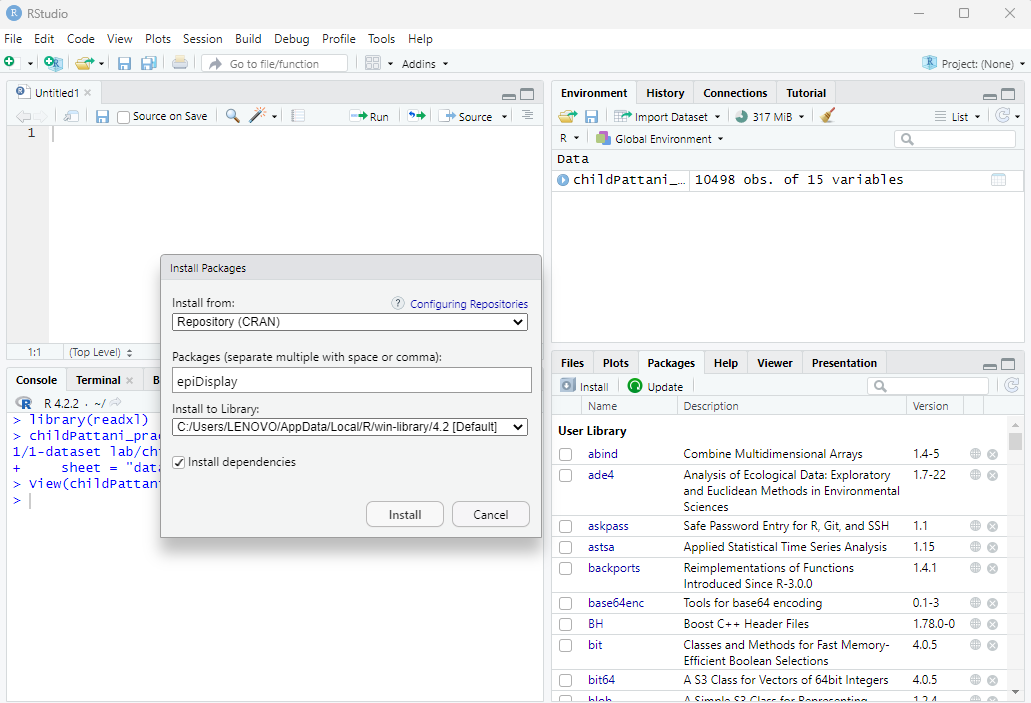
* Import

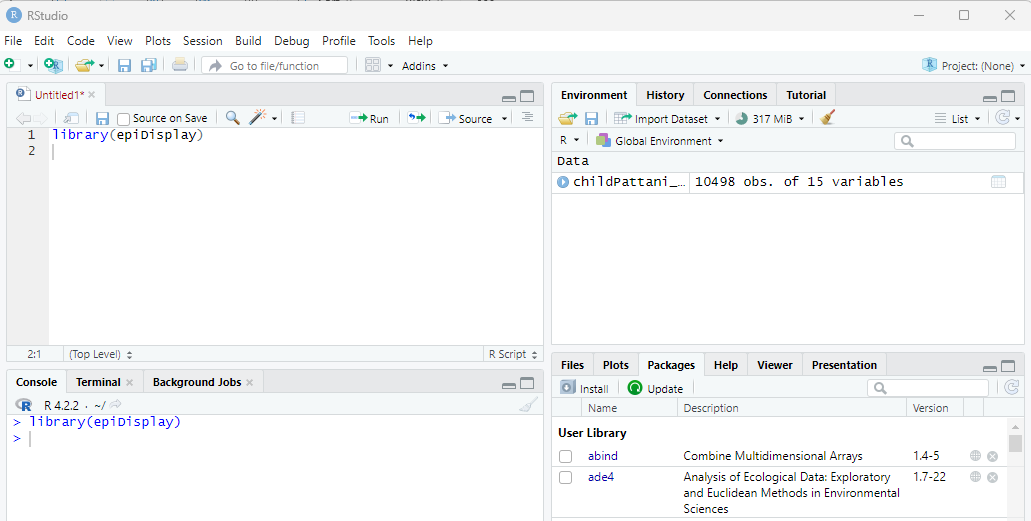


* ที่หน้าจอสำหรับพิมพ์คำสั่งจะแสดงข้อมูล สามารถปิดหน้าจอนี้ได้โดยคลิ๊กที่ x



* ให้ติดตั้ง Library(epiDisplay)
* เลือกแท็บ Packages 🡪 ในช่อง Packages ให้พิมพ์ epiDisplay
* คลิ๊ก Install



เรียกใช้ library epiDisplay ด้วยการพิมพ์คำสั่ง library(epiDisplay)  


ให้นักศึกษาตอบคำถามต่อไปนี้

ตั้งชื่อ data frame ใหม่

> dt <- childPattani\_practice

| **คำถาม** | **แสดง commands, ผลลัพธ์  และเขียนคำอธิบายผลลัพธ์** |
| --- | --- |
| ให้สำรวจข้อมูลอายุ (age) ด้วยคำสั่ง summ() และ summary()  > dt$age <- as.numeric(dt$age)  > summ(dt$age)  > summary(dt$age) | 'data.frame': 10498 obs. of 9 variables:  $ perid\_pre: num 2158 6627 6626 6625 477 ...  $ titnme : chr "2" "เด็กหญิง" "เด็กหญิง" "นาง" ...  $ age : num 4 2 4 33 NA 21 62 57 NA 27 ...  $ gender : chr "2" "2" "2" "2" ...  $ tmb : chr "บางตาวา" NA NA NA ...  $ amp : chr "หนองจิก" NA NA NA ...  $ prv : chr "ปัตตานี" NA NA NA ...  $ hholdid : num 0 0 0 0 0 0 0 0 0 0 ...  $ comment : chr "2เด็กอายุ 3-25 ปี" "3สมาชิกในครัวเรือน" "3สมาชิกใน  obs. mean median s.d. min. max.  9221 22.737 16 18.094 0 98  Min. 1st Qu. Median Mean 3rd Qu. Max. NA's  0.00 10.00 16.00 22.74 33.00 98.00 1277 |
| ให้สำรวจข้อมูลเพศ (gender) ด้วยคำสั่ง tab1()  > tab1(dt$gender) | Frequency %(NA+) %(NA-)  1 4474 42.6 47.5  2 4751 45.3 50.4  ชาย 109 1.0 1.2  หญิง 84 0.8 0.9  <NA> 1080 10.3 0.0  Total 10498 100.0 100.0 |
| ให้สำรวจข้อมูลตำบล (tmb) ด้วยคำสั่ง tab1() | Frequency %(NA+) %(NA-)  0 2 0.0 0.1  เมาะมะวี 1 0.0 0.0  เมาะมาวี 757 7.2 20.5  แม่ลาน 15 0.1 0.4  แหลมโพธ์ 5 0.0 0.1  แหลมโพธิ์ 198 1.9 5.4  ควน 36 0.3 1.0  ดอน 19 0.2 0.5  ตะโละ 172 1.6 4.7  ท่าน้ำ 13 0.1 0.4  นาเกตุ 410 3.9 11.1  บางตาวา 307 2.9 8.3  บ้านน้ำบ่อ 88 0.8 2.4  ป่าไร่ 856 8.2 23.2  ปากบางตาวา 4 0.0 0.1  ปุลากง 208 2.0 5.6  ยะหริ่ง 2 0.0 0.1  รุสะมิแล 3 0.0 0.1  รูสะมิแล 593 5.6 16.1  <NA> 6809 64.9 0.0  Total 10498 100.0 100.0 |
| ให้สำรวจข้อมูลอำเภอ (amp) | Frequency %(NA+) %(NA-)  0 2 0.0 0.1  เมือง 25 0.2 0.7  เมืองปัตตานี 571 5.4 15.5  แม่ลาน 871 8.3 23.6  โคกโพธิ์ 410 3.9 11.1  ปะนาเระ 156 1.5 4.2  ยะรัง 751 7.2 20.4  ยะหริ่ง 592 5.6 16.0  หนองจิก 311 3.0 8.4  <NA> 6809 64.9 0.0  Total 10498 100.0 100.0 |
| ให้สำรวจข้อมูลจังหวัด (prv) | Frequency %(NA+) %(NA-)  0 2 0.0 0.1  ปัตตานี 3687 35.1 99.9  <NA> 6809 64.9 0.0  Total 10498 100.0 100.0 |
| ให้สำรวจข้อมูลตัวแปร comment | Frequency Percent Cum. percent  1ผู้ให้ข้อมูล 999 9.5 9.5  2เด็กอายุ 3-25 ปี 2770 26.4 35.9  3สมาชิกในครัวเรือน 6729 64.1 100.0  Total 10498 100.0 100.0 |

ให้นักศึกษาเขียนสรุปการใช้คำสั่ง summ, summary และ tab1

|  |
| --- |
|  |

1. ให้นักศึกษาฝึกสำรวจข้อมูลโดยใช้ชุดข้อมูล “nomotercycle\_practice.xlsx”

จาก Data file “nomotercycle\_practice.xlsx” ให้นักศึกษาสำรวจตัวแปรดังนี้ DEAD\_YEAR, Age, Sex, BirthYear, NationalityId, RiskHelmet, RiskSafetyBelt, AccProv, และ Vehicle

| **คำถาม** | **แสดง commands, ผลลัพธ์  และเขียนคำอธิบายผลลัพธ์** |
| --- | --- |
| DEAD\_YEAR |  |
| Age |  |
| Sex |  |
| BirthYear |  |
| NationalityId |  |
| RiskHelmet |  |
| RiskSafetyBelt |  |
| AccProv |  |
| Vehicle |  |