**ปฏิบัติการที่ 3**

**การทำความสะอาดข้อมูลและจัดการข้อมูล**

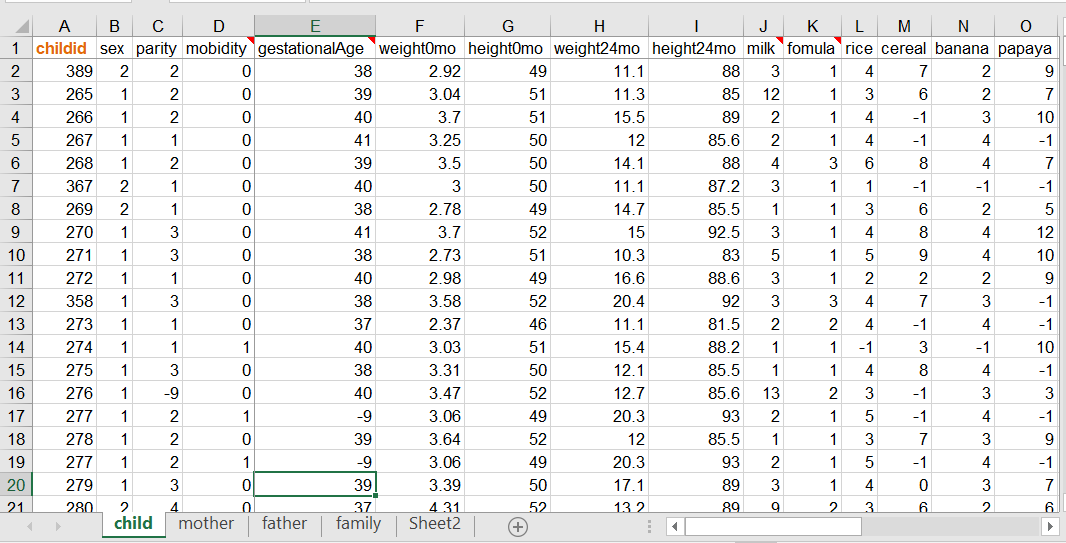
**วัตถุประสงค์:** 1. เพื่อให้นักศึกษาฝึกปฏิบัติทำความสะอาดและจัดการข้อมูลด้วยโปรแกรม R  
 2. เพื่อให้นักศึกษาฝึกปฏิบัติผสานข้อมูลจากหลาย ๆ ตารางได้

**Data file:** “childThepa\_practice.xlsx”, “childPattani\_practice.xlsx”, **“**teacherstress\_practice.csv**”, “**teachstress\_Q.docx**”,** “child140\_practice.xlsx”

**สรุปเนื้อหา:**

การจัดเตรียมข้อมูล ได้แก่ การทำความสะอาดข้อมูล เป็นการขจัดข้อมูลที่ผิดพลาด ไม่สมบูรณ์ หรือซ้ำซ้อน การจัดรูปแบบข้อมูล เป็นการแปลงข้อมูลให้อยู่ในรูปแบบที่สอดคล้องกัน เช่น แปลงวันที่ให้อยู่ในรูปแบบเดียวกัน การจัดการข้อมูลที่ขาดหายไป เป็นการประมาณค่าหรือเติมข้อมูลที่ขาดหายไป การแปลงข้อมูลเป็นการแปลงข้อมูลให้เป็นรูปแบบที่เหมาะสมกับการวิเคราะห์ เช่น แปลงข้อมูลตัวอักษรเป็นตัวเลข การรวมข้อมูลเป็นการรวมข้อมูลจากหลายแหล่งเข้าด้วยกัน นอกจากการตรวจสอบข้อมูลข้างต้นแล้ว ยังมีการคำนวณค่าต่าง ๆ เพื่อเตรียมข้อมูลในการวิเคราะห์ข้อมูลต่อไป โดยทั่วไปที่เราจะตรวจสอบเป็นอันดับต้น ๆ คือ ข้อมูลซ้ำซ้อน ข้อมูลสูญหายซึ่งมักจะรายงานเป็นเปอร์เซ็นต์ ข้อมูลที่มีค่าไม่เป็นตามคำอธิบายข้อมูลซึ่งค่าเหล่านี้มักจะแทนค่าด้วย NA หรือหากผู้วิจัยสามารถตรวจสอบค่าจากแหล่งข้อมูลได้ก็สามารถแก้ไขข้อมูลที่บันทึกผิดพลาดให้ถูกต้องได้ การจัดการข้อมูลในกรณีที่ตัวแปรตามมีหลายตัว เพื่อให้ง่ายต่อการวิเคราะห์ข้อมูลและง่ายในการอธิบายผลการวิเคราะห์ข้อมูล

1. **จาก data file “childThepa\_practice.xlsx” ให้นักศึกษาฝึกปฏิบัตทำความสะอาดและจัดการข้อมูล**



# กำหนด locale ให้อ่านภาษาไทย

> Sys.setlocale("LC\_ALL", "thai")

# กำหนดโฟลเดอร์ที่มี data file “childThepa\_practice.xlsx”

> setwd(“…………………………………………………”)

# อ่าน data file “childThepa\_practice.xlsx”

# install.packages("readxl") # ติดตั้งแค่ 1 ครั้ง

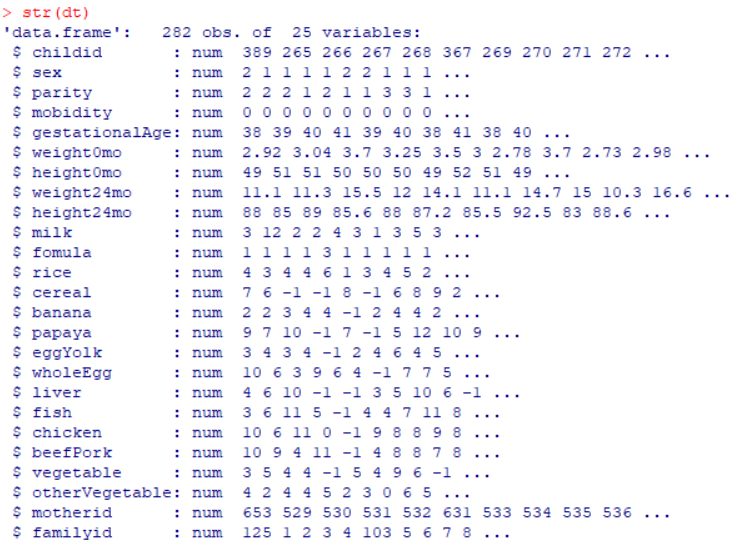
> library(readxl)

> dt <- read\_excel("childThepa\_practice.xlsx")

> str(dt)

> dt <- as.data.frame(dt)

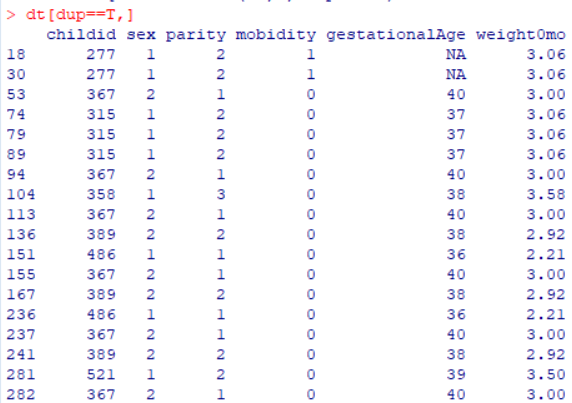
> str(dt)

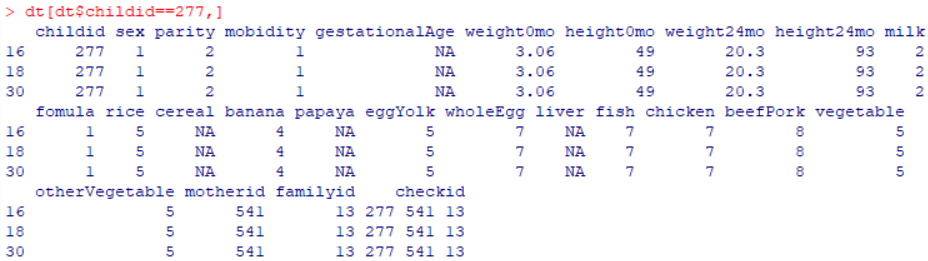


1. ตรวจสอบข้อมูลซ้ำซ้อนโดยใช้ตัวแปร childid, mothered, familyid ประกอบกัน  
   > dt$checkid <- paste(dt$childid, dt$motherid, dt$familyid)  
   > dup <- duplicated(dt$checkid)  
   > table(dup)

|  |
| --- |
| เขียนคำอธิบายพอสังเขป  ..........................................................................................................................................  ..........................................................................................................................................  .......................................................................................................................................... |

# แสดงข้อมูลที่ซ้ำ

> dt[dup==T,]  


> dt[dt$childid==277,]  


# ตัดข้อมูลที่ซ้ำออก

> dt1 <- dt[dup==F,]

> str(dt1)

|  |
| --- |
| เขียนคำอธิบายโดยเปรียบเทียบกับข้อมูลก่อนลบข้อมูลซ้ำ  ..........................................................................................................................................  .......................................................................................................................................... |

1. ตรวจสำรวจข้อมูลตัวแปรทั้งหมดด้วยคำสั่ง summary()  
   > summary(dt1)

|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์)    เขียนคำอธิบายพอสังเขป  ..........................................................................................................................................  ..........................................................................................................................................  ..........................................................................................................................................  ..........................................................................................................................................  ..........................................................................................................................................  .......................................................................................................................................... |

1. ทุกตัวแปรที่มีค่าเป็นลบให้แทนค่าด้วย NA

> for (i in c(2:23)) {

> dt1[,i] <- ifelse(dt1[,i]<0,NA,dt1[,i])

> }

> summary(dt1)

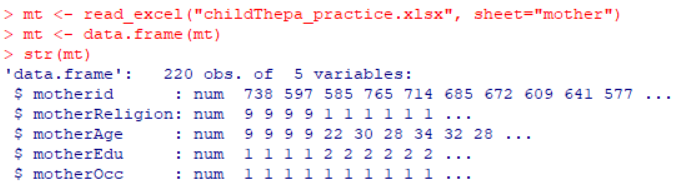
|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์)  เขียนคำอธิบายโดยเปรียบเทียบกับข้อมูลก่อนหน้า  .........................................................................................................................................  .........................................................................................................................................  .........................................................................................................................................  ......................................................................................................................................... |

1. กำหนดให้ตัวแปร sex และ mobidity เป็นตัวแปร factor พร้อมให้คำอธิบายรหัสของแต่ละตัวแปร (โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ)  
   # ตัวแปร sex  
   > dt1$sex <- factor(dt1$sex)  
   > levels(dt1$sex) <- c(“Male”,”Female”)

|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์) |

1. จัดกลุ่มตัวแปร milk, formula, rice, cereal, banana, papaya, eggYolk, wholeEgg, liver, fish, chicken, beefPork, vegetable, otherVegetable โดยทุกตัวแปร จัดให้เป็น 2 กลุ่ม ดังนี้ (โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ)  
   - 1: < 6  
   - 2: >=6

|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์) |

1. ให้ทำการผสาน (merge) ข้อมูล เด็ก และมารดา (sheet->mother)   
   # อ่าน data file “childThepa\_practice.xlsx” sheet mother  
   > mt <- read\_excel("childThepa\_practice.xlsx", sheet="mother")  
   > mt <- data.frame(mt)  
   > str(mt)  
     
     
   # Merger file dt1 and mt (เด็กและมารดา)  
   > dt2 <- merge(dt1, mt, by.x="motherid", by.y="motherid", all.x=T)  
   > str(dt2)

|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์) |

1. ให้ตรวจสอบ ทำความสะอาด และจัดการข้อมูลมารดาของตัวแปร motherReligion, motherAge, motherEdu และ motherOcc หากมีค่าที่ผิดพลาด ให้จัดการข้อมูล

ตัวแปร motherReligion (โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ)

|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์) |

ตัวแปร motherAge (โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ)

|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์) |

ตัวแปร motherEdu (โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ)

|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์) |

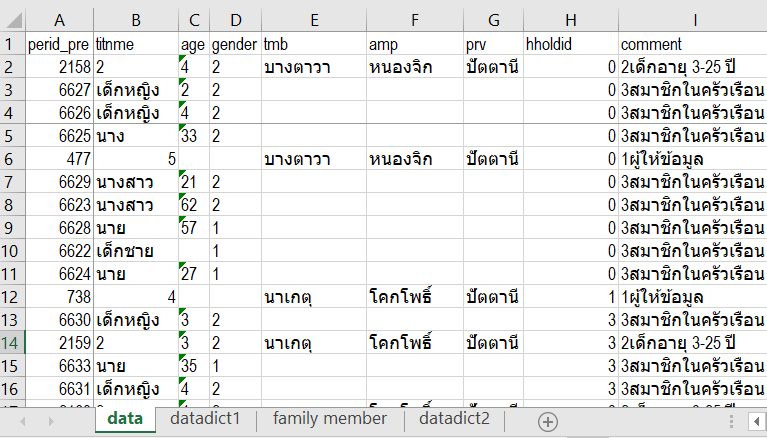
ตัวแปร motherOcc (โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ)

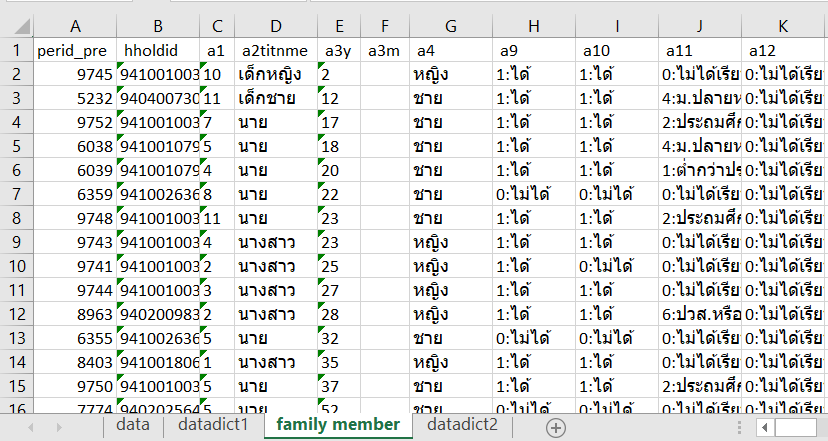
|  |
| --- |
|  |

1. ให้ผสานข้อมูลเด็กและครอบครัว (sheet->family) และตรวจสอบ ทำความสะอาดตัวแปร income และ SES หากมีค่าที่ผิดพลาด ให้จัดการข้อมูล (โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ)

|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์) |

1. **ให้นักศึกษาฝึกสำรวจ ทำความสะอาด และจัดการข้อมูลโดยใช้ชุดข้อมูล “childPattani\_practice.xlsx”**
2. อ่าน/นำเข้า data sheet “data” ตั้งชื่อเป็น “mem” ซึ่งเป็นข้อมูลของสมาชิกทั้งหมดในแต่ละครอบครัว และอ่าน/นำเข้า data sheet “family member” ตั้งชื่อเป็น “fam”





|  |
| --- |
| คำตอบ (แสดงคำสั่งและบางส่วนของผลลัพธ์) |

1. ให้ผสาน data frame “mem” และ “fam” ตั้งชื่อเป็น “dt”

|  |
| --- |
| คำตอบ (แสดงคำสั่งและบางส่วนของผลลัพธ์) |

1. ถ้าตัวแปร age เป็นค่าว่าง ให้แทนที่ด้วยค่าของข้อมูลจากตัวแปร a3y และถ้า gender เป็นค่าว่าง ให้แทนค่าของข้อมูลจากตัวแปร a4 (โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ)

|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์) |

1. ให้ตรวจสอบตัวแปร titnme แล้วแก้ไขข้อมูลดังนี้ (โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ)  
    เด็กชาย 🡪 1  
    เด็กหญิง 🡪 2  
    นาง 🡪 3  
    นาย 🡪 4  
    นางสาว 🡪 5

|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์) |

1. ให้ตรวจสอบตัวแปร gender แล้วแก้ไขข้อมูลดังนี้ (โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ)  
    ชาย 🡪 1  
    หญิง 🡪 2
2. ตรวจสอบคำนำหน้าชื่อกับเพศว่าสอดคล้องกันหรือไม่ หากไม่สอดคล้องให้ทำการแก้ไขข้อมูล (โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ)

|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์) |

1. ให้เรียงลำดับข้อมูลตามตัวแปร tmb จากมากไปน้อย

|  |
| --- |
| คำตอบ (แสดงคำสั่งและบางส่วนของผลลัพธ์) |

1. ให้ตรวจสอบข้อมูลซ้ำ โดยใช้ตัวแปร titnme, age, gender, hholdid เป็นเงื่อนไข และลบข้อมูลซ้ำ

|  |
| --- |
| คำตอบ (แสดงคำสั่งและบางส่วนของผลลัพธ์) |

1. ให้เลือกเฉพาะข้อมูลของสมาชิกที่มีอายุ 15-50 ปี เก็บไว้ใน data frame ใหม่ โดยตั้งชื่อเป็น “dt1”

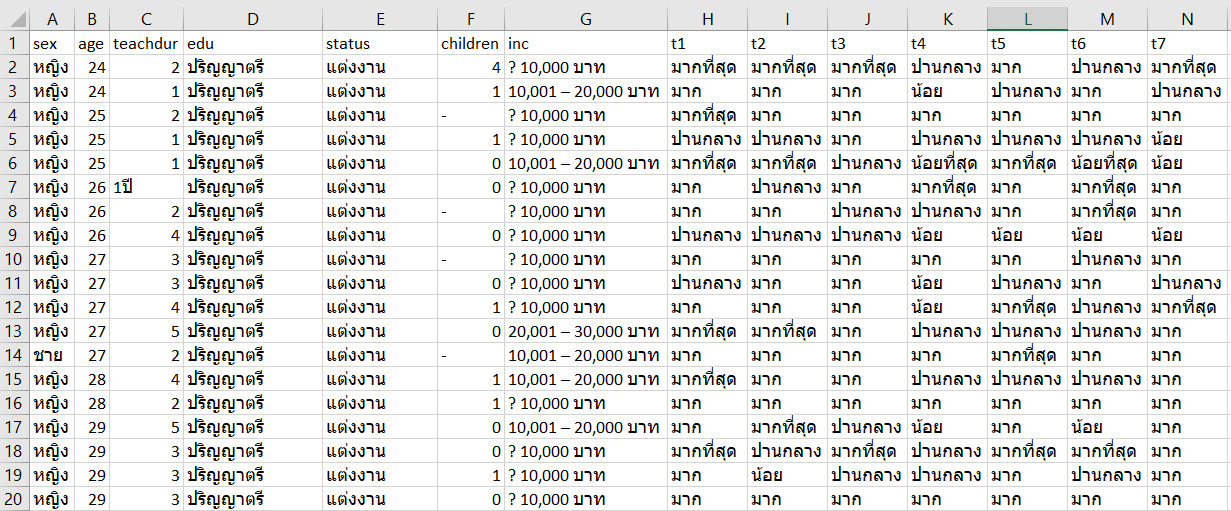
|  |
| --- |
| คำตอบ (แสดงคำสั่งและบางส่วนของผลลัพธ์) |

1. จาก data frame dt1 ให้ตรวจสอบว่าแต่ละคนประกอบอาชีพอะไรบ้าง (ตัวแปร a15c0, a15c1, a15c2, a15c3, a15c4, a15c5, a15c6, a15c7, a15c8, a15c9) (โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ)

|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์) |

1. **จากข้อมูล “teacherstress\_practice.xlsx” และคำอธิบายข้อมูล “teachstress\_Q.docx” ให้นักศึกษาฝึกสำรวจข้อมูลและจัดการข้อมูลของตัวแปรต่าง ๆ**

(รูปข้างล่างนี้ แสดงข้อมูลบางส่วนเท่านั้น)



1. ให้นักศึกษาระบุประเภทของตัวแปรต่อไปนี้ โดยทำเครื่องหมายถูก (√) ในช่องที่กำหนด

| ตัวแปร | คำตอบ | |
| --- | --- | --- |
| ตัวแปรกลุ่ม | ตัวแปรต่อเนื่อง |
| sex | √ |  |
| age |  | √ |
| teachdur |  | √ |
| edu | √ |  |
| stutus | √ |  |
| children |  | √ |
| inc | √ |  |
| t1-t10 | √ |  |
| m1-m3 | √ |  |
| st1-st10 | √ |  |

1. เพิ่มคอลัมน์ ID และกำหนดตัวเลขตั้งแต่ 1-n โดยที่ n คือ จำนวนข้อมูลทั้งหมด

|  |
| --- |
| คำตอบ (แสดงคำสั่งและบางส่วนของผลลัพธ์) |

1. สำรวจข้อมูลตัวแปร sex หากมีค่าที่ผิดพลาดให้ทำการจัดการข้อมูล (โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ)

|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์) |

1. สำรวจข้อมูลตัวแปร age หากมีค่าที่ผิดพลาดให้ทำการจัดการข้อมูล (โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ)

|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์) |

1. สำรวจข้อมูลตัวแปร teachdur หากมีค่าที่ผิดพลาดให้ทำการจัดการข้อมูล (โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ)

|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์) |

1. สำรวจข้อมูลตัวแปร edu หากมีค่าที่ผิดพลาดให้ทำการจัดการข้อมูล (โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ)

|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์) |

1. ตัวแปร teachdur และ edu ตรวจสอบความสอดคล้องของข้อมูล หากมีค่าที่ผิดพลาดให้ทำการจัดการข้อมูล (โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ)

|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์) |

1. สำรวจข้อมูลตัวแปร status หากมีค่าที่ผิดพลาดให้ทำการจัดการข้อมูล (โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ)

|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์) |

1. สำรวจข้อมูลตัวแปร children หากมีค่าที่ผิดพลาดให้ทำการจัดการข้อมูล (โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ)

|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์) |

1. ตัวแปร status และ children ตรวจสอบความสอดคล้องของข้อมูล หากมีค่าที่ผิดพลาดให้ทำการจัดการข้อมูล (โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ)

|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์) |

1. สำรวจข้อมูลตัวแปร inc หากมีค่าที่ผิดพลาดให้ทำการจัดการข้อมูล (โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ)

|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์) |

1. ตัวแปร t1-t10 และ m1-m2 ให้แก้ไขข้อมูลดังนี้ โดยใช้คำสั่ง for ในการทำงานให้เร็วขึ้น  
   มากที่สุด แก้ไขเป็น 5

มาก แก้ไขเป็น 4

ปานกลาง แก้ไขเป็น 3

น้อย แก้ไขเป็น 2

น้อยที่สุด แก้ไขเป็น 1

|  |
| --- |
| คำตอบ (แสดงคำสั่งและบางส่วนของผลลัพธ์) |

1. ตัวแปร st1-st10 ให้แก้ไขข้อมูลดังนี้ โดยใช้คำสั่ง for ในการทำงานให้เร็วขึ้น  
   ไม่เคย แก้ไขเป็น 0

ส่วนใหญ่ไม่เคย แก้ไขเป็น 1

บางครั้ง แก้ไขเป็น 2

ค่อนข้างบ่อย แก้ไขเป็น 3

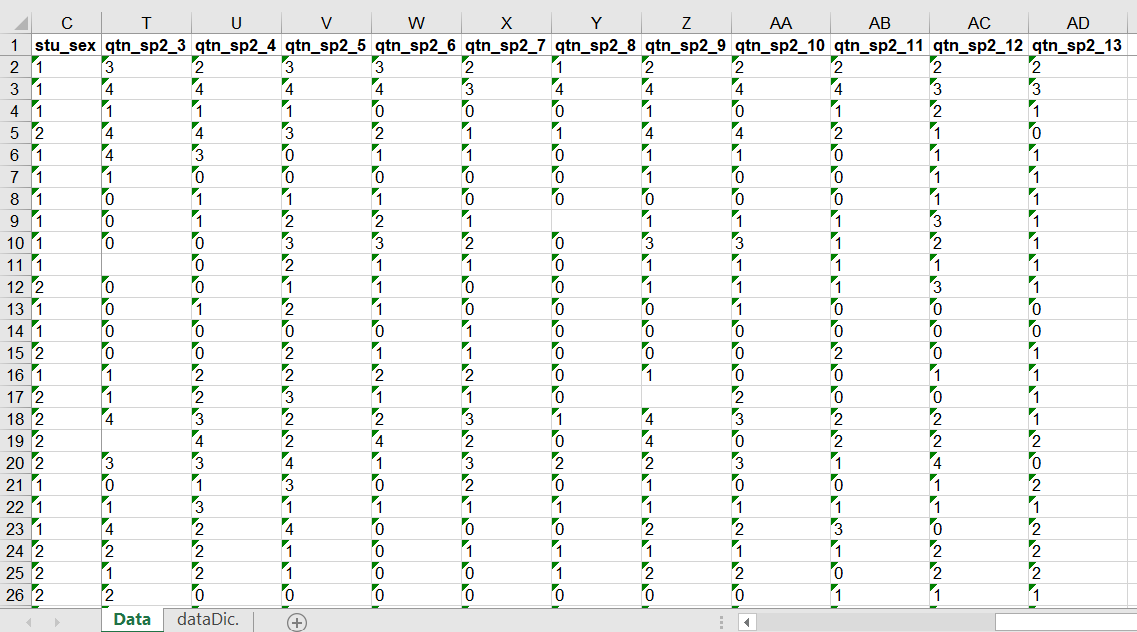
บ่อยมาก แก้ไขเป็น 4

|  |
| --- |
| คำตอบ (แสดงคำสั่งและบางส่วนของผลลัพธ์) |

1. ให้คำนวณค่าเฉลี่ยของระดับคะแนนของกลุ่มตัวอย่างแต่ละคน (row average) ดังนี้  
   ค่าเฉลี่ยของระดับคะแนนตัวแปร t1-t10 ตั้งชื่อเป็น t\_average  
   ค่าเฉลี่ยของระดับคะแนนตัวแปร m1-m3 ตั้งชื่อเป็น m\_average  
   ค่าเฉลี่ยของระดับคะแนนตัวแปร st1-st10 ตั้งชื่อเป็น st\_average  
   และให้แสดงผลลัพธ์ในรูปแบบตารางสำหรับตัวแปร t\_average, m\_average และ st\_average

|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์) |

1. **ให้นักศึกษาฝึกสำรวจข้อมูล และจัดการข้อมูล “child140\_practice.xlsx”**

****

1. ให้คำนวณค่าเฉลี่ยของระดับคะแนนตัวแปร qtn\_sp2\_1 - qtn\_sp2\_13 ของกลุ่มตัวอย่างแต่ละคน (row average) เก็บในตัวแปรใหม่ชื่อ behAvg โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ

|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์) |

1. ให้จัดกลุ่มน้ำหนักแรกเกิดเป็น 2 กลุ่ม ดังแสดงข้างล่างนี้ เก็บในตัวแปรใหม่ชื่อ bwgrp โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ  
    1: <3,000 gram  
    2: >= 3,000 gram

|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์) |

1. ให้แสดงค่าเฉลี่ยของตัวแปร behAvg โดยจำแนกแต่ละกลุ่มของ bwgrp ให้อธิบายความหมาย

|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์)    อธิบายความหมาย  ตัวแปร bwgrp และการจัดกลุ่มหาค่าเฉลี่ยตัวแปร behAvg  ผลลัพธ์ที่ได้สอดคล้องกับความเป็นจริง |

1. ให้คำนวณหาผลรวมของระดับคะแนนตัวแปร qtn\_sp2\_1 - qtn\_sp2\_13 ของกลุ่มตัวอย่างแต่ละคน เก็บในตัวแปรใหม่ชื่อ behSum โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ

|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์) |

1. ให้จัดกลุ่มข้อมูล behSum ดังแสดงข้างล่างนี้ ตั้งชื่อตัวแปรใหม่เป็น behgrp โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ  
    <3 🡪 0  
    >=3 🡪 1

|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์) |

1. ให้สำรวจข้อมูลโดยใช้ 2 ตัวแปร คือ เพศ และ behgrp โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ

|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์)      อธิบายความหมาย  ข้อมูลของเพศชายและหญิง สอดคล้องกับความป็นจริง |

1. ให้คำนวณค่า BMI (Body Mass Index) = น้ำหนัก (kg) / ส่วนสูง (m)2 ของกลุ่มตัวอย่างแต่ละคน และจัดกลุ่ม BMI ดังแสดงข้างล่างนี้ ตั้งชื่อตัวแปรใหม่เป็น bmigrp โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ

1: <18.5 ความหมาย ผอม

2: 18.5-24.9 ความหมาย สุขภาพดี

3: >24.9 8 ความหมาย น้ำหนักเกิน

|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์)      อธิบายความหมาย  มีค่า BMI ที่เกิน 30 ซึ่งไม่สอดคล้องกับความป็นจริง |

1. ให้สำรวจข้อมูลโดยใช้ 2 ตัวแปร คือ เพศ และ bmigrp โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ

|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์)      อธิบายความหมาย  ตัวแปร เพศ และ bmigrp มีข้อมูลที่ถูกต้อง |

1. ให้สำรวจข้อมูลโดยใช้ 2 ตัวแปร คือ bmigrp และ behgrp โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ

|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์)      อธิบายความหมาย  ค่า bmi มีค่าถูกต้อง |