**ปฏิบัติการที่ 4**

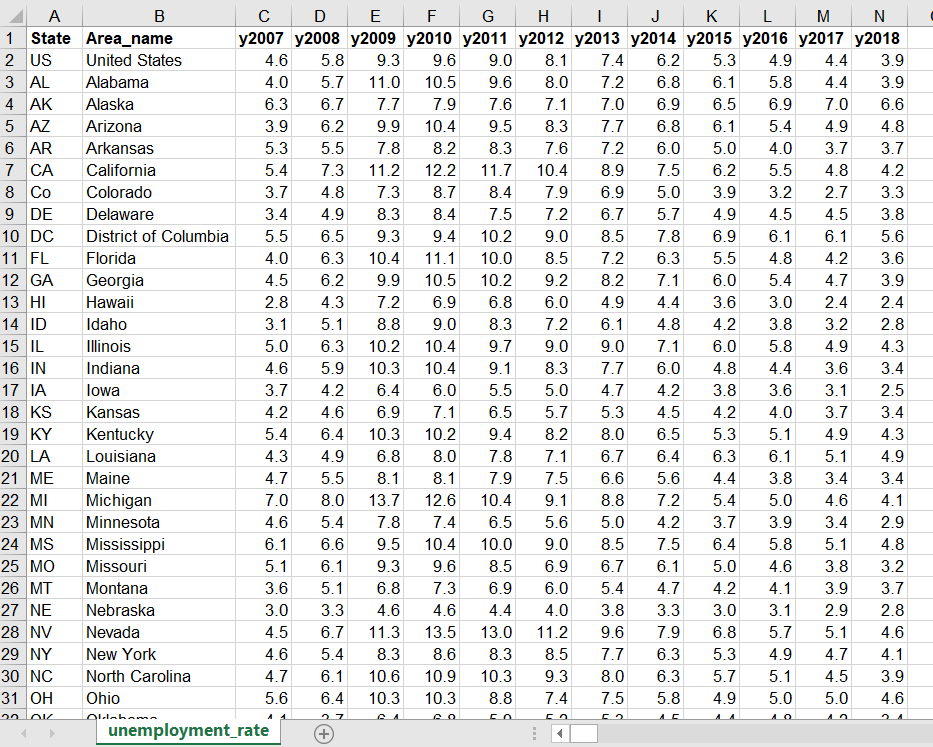
**การทำความสะอาดและจัดการข้อมูล**

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

**Data file:** “unemployment.xlsx”, “child140\_practice.xlsx”, “thai\_road\_accident\_practice.csv”,“CBR64-Nara1\_practice.xlsx” และ “CBR64-Nara2\_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. **ให้นักศึกษาฝึกจัดการโครงสร้างข้อมูลโดยใช้ data file “unemployment.xlsx”** 
2. ให้จัดข้อมูลจาก wide to long ดังตัวอย่างที่แสดงข้างล่างนี้

| state | areaname | year | unemrate |
| --- | --- | --- | --- |
| US | United States | 2007 | 4.6 |
| US | United States | 2008 | 5.8 |
| US | United States | 2009 | 9.3 |
| US | United States | 2010 | 9.6 |
| US | United States | 2011 | 9.0 |
| … | … | … | … |
| AL | Alabama | 2007 | 4.0 |
| AL | Alabama | 2008 | 5.7 |
| AL | Alabama | 2009 | 11.0 |
| AL | Alabama | 2010 | 10.5 |
| AL | Alabama | 2011 | 9.6 |
| … | … | … | … |

|  |
| --- |
| คำตอบ (แสดงคำสั่งและบางส่วนของผลลัพธ์)  Sys.setlocale("LC\_ALL", locale="Thai")  # install.packages("readxl")  > library(readxl)  # install.packages("epiDisplay")  > library(epiDisplay)  # install.packages("reshape2")  > library(reshape2)  > setwd("………………………………………………")  > dt <- read\_excel("unemployment.xlsx")  > dt <- as.data.frame(dt)  > str(dt)  # Reshape  > d <- melt(dt, id=c("State","Area\_name"))  > colnames(d) <- c("st","area","yr","unemp")  > d$yr <- sub("y","",d$yr)  > d <- d[order(d$st,d$area,d$yr),]  > View(d) |

1. ให้ตรวจสอบการแจกแจงของข้อมูลอัตราการว่างงาน unemp หากข้อมูลไม่มีการแจงแจงแบบปกติให้ทำการแปลงข้อมูลให้อยู่ในรูปแบบการแจงแจงข้อมูลแบบปกติ

|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์)  > hist(d$unemp) |

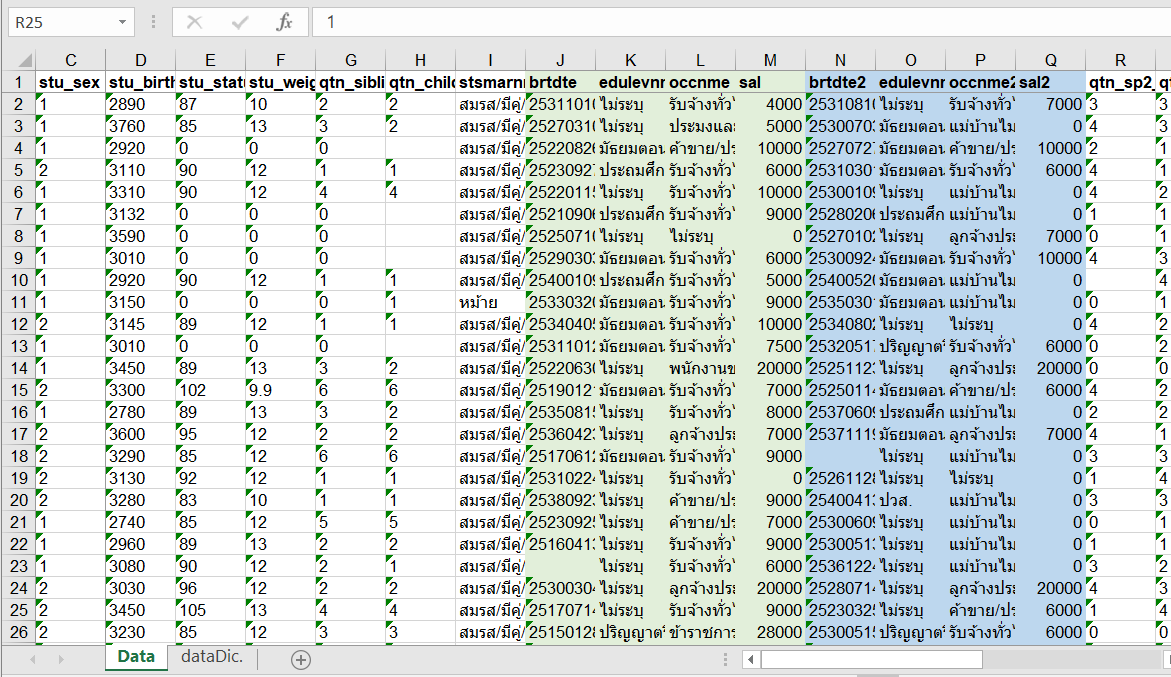
1. ให้ตรวจสอบข้อมูลแต่ละรัฐมีอัตราการว่างงานเป็นอย่างไรบ้าง โดยเฉลี่ยแล้วรัฐใดมีอัตราการว่างงานน้อยที่สุด และมากที่สุด

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

1. ให้บันทึกข้อมูลลงในโฟลเดอร์ โดยตั้งชื่อไฟล์เป็น “unemploy2017.csv”

|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์ที่ปรากฏในโฟลเดอร์)  > write.csv(d, file="unemploy2017.csv", row.names=F) |

1. **ให้นักศึกษาฝึกจัดการข้อมูลตัวแปรวันที่โดยใช้ชุดข้อมูล “child140\_practice.xlsx”**



# กำหนดโฟลเดอร์ที่มีข้อมูล “child140\_practice.xlsx”

> setwd("............................")

# เรียกใช้ library excel

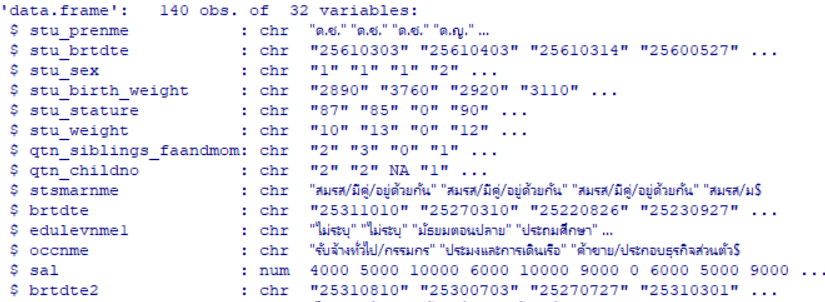
> library(readxl)

# อ่าน data file "child140\_practice.xlsx")

> dt <- read\_excel("child140\_practice.xlsx")

> dt <- data.frame(dt)

> str(dt)



1. ให้คำนวณอายุของกลุ่มตัวอย่างแต่ละคน (ตัวแปร stu\_brtdte) เก็บไว้ในตัวแปรใหม่ชื่อ stuage โดยแสดงในรูปแบบตาราง

|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์)  > dt$stubrtdte <- as.Date(dt$stu\_brtdte, "%Y%m%d")  > today <- as.Date("25670707","%Y%m%d")  > dt$stuage <- as.numeric(today - dt$stubrtdte)/365.25  > str(dt) |

1. ให้คำนวณอายุบิดาของกลุ่มตัวอย่างแต่ละคน (ตัวแปร brtdte) เก็บไว้ในตัวแปรใหม่ชื่อ faage โดยแสดงในรูปแบบตาราง

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

1. ให้คำนวณอายุมารดาของกลุ่มตัวอย่างแต่ละคน (ตัวแปร brtdte2) เก็บไว้ในตัวแปรใหม่ชื่อ momage โดยแสดงในรูปแบบตาราง

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

1. ให้จัดกลุ่มอายุบิดา ตามเงื่อนไขที่แสดงข้างล่างนี้ เก็บไว้ในตัวแปรใหม่ชื่อ faagegrp โดยแสดงในรูปแบบตารางหรือกราฟ  
    1: <=30

2: 31–40

3: >40

4 Undefined (กรณีไม่ระบุ/Missing value)

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

1. ให้จัดกลุ่มอายุมารดา ตามเงื่อนไขที่แสดงข้างล่างนี้ เก็บไว้ในตัวแปรใหม่ชื่อ momagegrp โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ

1: <=30

2: 31–40

3: >40

4 Undefined (กรณีไม่ระบุ/Missing value)

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

1. แต่ละกลุ่มอายุของบิดามีเงินเดือนเฉลี่ยเท่าไหร่ แสดงผลลัพธ์ในรูปแบบตาราง

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

1. แต่ละกลุ่มอายุของมารดามีเงินเดือนเฉลี่ยเท่าไหร่ แสดงผลลัพธ์ในรูปแบบตาราง

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

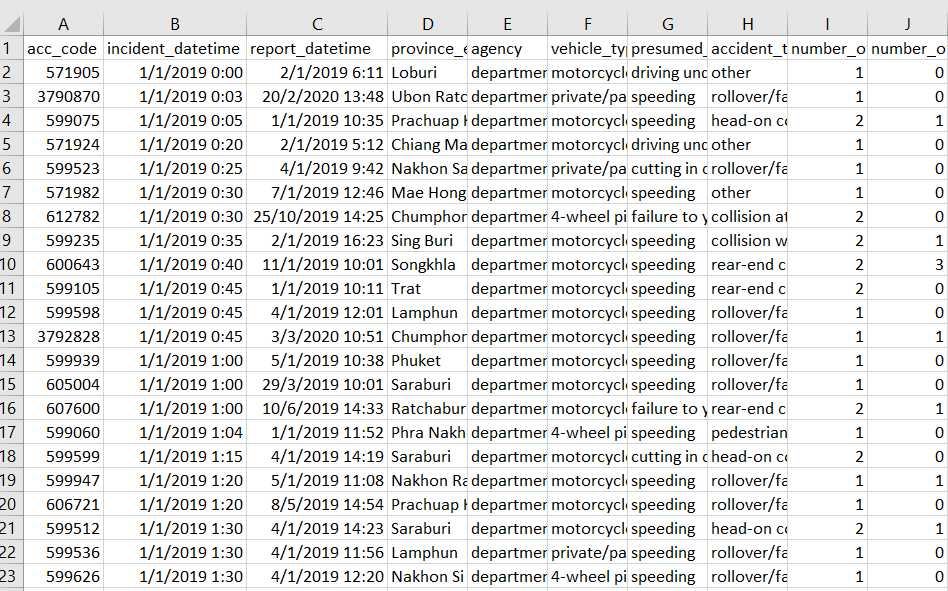
1. แต่ละกลุ่มอายุของบิดาประกอบอาชีพอะไรบ้าง แสดงผลลัพธ์ในรูปแบบตาราง

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

1. แต่ละกลุ่มอายุของมารดาจบการศึกษาระดับใดบ้าง แสดงผลลัพธ์ในรูปแบบตาราง

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

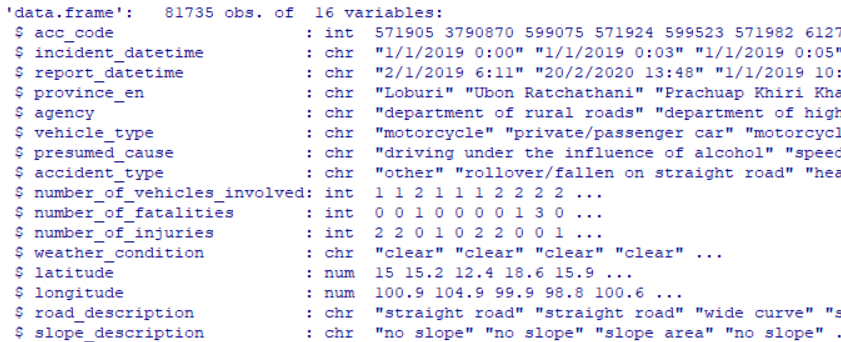
1. **ให้นักศึกษาฝึกสำรวจข้อมูลโดยใช้ชุดข้อมูล “thai\_road\_accident\_practice.csv”**



# อ่าน/นำเข้ามูล "thai\_road\_accident\_practice.csv")

> dt <- read.csv("thai\_road\_accident\_practice.csv")

> str(dt)



1. ให้นักศึกษาเขียนคำอธิบายตัวแปรโดยสังเกตจากข้อมูลของแต่ละตัวแปร

| ตัวแปร | คำอธิบาย |
| --- | --- |
| acc\_code | รหัสอุบัติเหตุ |
| incident\_datetime | เวลาเกิดเหตุ(วัน/เดือน/ปี) |
| report\_datetime | เวลารายงาน(วัน/เดือน/ปี) |
| province\_en | ชื่อจังหวัด |
| agency | เอเจนซี |
| vehicle\_type | ประเภทของยานพาหนะ |
| presumed\_cause | สาเหตุของการเกิดอุบัติเหตุ |
| accident\_type | ประเภทของอุบัติเหตุ |
| number\_of\_vehicles\_involved | จำนวนของยานพาหนะที่เกิดเหตุ |
| number\_of\_fatalities | จำนวนผู้เสียชีวิต |
| number\_of\_injuries | จำนวนผู้รับบาดเจ็บ |
| weather\_condition | สภาพอากาศ |
| latitude | วัดระยะทางเหนือ-ใต้ของเส้นศูนย์สูตร |
| longitude | วัดระยะทางตะวันออก-ตะวันตกของเส้นศูนย์สูตร |
| road\_description | คำอธิบายของถนน |
| slope\_description | คำอธิบายความลาดชัน |

1. ให้หาปีที่เกิดเหตุ (incident\_datetime) ของกลุ่มตัวอย่างแต่ละคน ข้อมูลถูกบันทึกในช่วงปีใด แต่ละปีมีจำนวนคนเกิดเหตุกี่ราย โดยแสดงผลลัพธ์เป็นตาราง

|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์)  > dt$incy <- format(as.Date(dt$incident\_datetime,"%d/%m/%Y"),"%Y")  > dt$incy <- as.numeric(dt$incy)  > summary(dt$incy)  > table(dt$incy) |

1. ให้หาเดือนที่เกิดเหตุ (incident\_datetime) ของกลุ่มตัวอย่างแต่ละคน แต่ละเดือนมีจำนวนคนเกิดเหตุกี่ราย โดยแสดงผลลัพธ์เป็นตาราง

|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์)  > dt$incm <- format(as.Date(dt$incident\_datetime,"%d/%m/%Y"),"%m")  > dt$incm <- as.numeric(dt$incm)  > summary(dt$incm)  > table(dt$incm) |

1. ในแต่ละปีส่วนใหญ่จะเกิดเหตุเดือนใด โดยแสดงผลลัพธ์เป็นตาราง

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

1. ให้หาช่วงเวลาที่เกิดเหตุ (incident\_datetime) ของกลุ่มตัวอย่างแต่ละคน และแบ่งกลุ่มช่วงเวลาดังแสดงข้างล่างนี้ เก็บในตัวแปรใหม่ชื่อ inctimegrp ให้นักศึกษาตรวจสอบว่าแต่ละช่วงเวลามีจำนวนคนเกิดเหตุกี่ราย โดยแสดงผลลัพธ์เป็นตาราง

1: 06.01-12.00

2: 12.01-18.00

3: 18.01-00.00

4: 00.01-06.00

|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์)  > dt$inctime <- as.POSIXct(dt$incident\_datetime, format = "%d/%m/%Y %H:%M")  > dt$inctime <- format(as.POSIXct(dt$inctime),format = "%H:%M")  > dt$inctime <- gsub("[:]", ".",dt$inctime)  > dt$inctime <- as.numeric(dt$inctime)  > summary(dt$inctime) |

1. ให้หาระยะห่างของวันนับตั้งแต่วันที่เกิดเหตุ (incident\_datetime) จนถึงวันที่รายงานการเกิดเหตุ (report\_datetime) ของกลุ่มตัวอย่างแต่ละคน โดยแสดงผลลัพธ์เป็นตาราง จำนวนวันที่สั้นที่สุดกี่วัน นานที่สุดกี่วัน โดยเฉลี่ยกี่วัน

|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์)  > dt$incdate <- as.Date(dt$incident\_datetime,"%d/%m/%Y")  > dt$rptdate <- as.Date(dt$report\_datetime,"%d/%m/%Y")  > dt$peri <- dt$rptdate- dt$incdate    คำอธิบาย |

1. ให้แสดงข้อมูลที่ผิดปกติในข้อข้างต้น โดยแสดงปี จังหวัด ประเภทของยานพาหนะ จำนวนผู้เสียชีวิต และจำนวนผู้บาดเจ็บ ให้หาระยะห่างของวันนับตั้งแต่วันที่เกิดเหตุ (incident\_datetime) จนถึงวันที่รายงานการเกิดเหตุ

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

1. ให้คำนวณแต่ละปีมีจำนวนผู้บาดเจ็บและเสียชีวิตจากอุบัติเหตุทางถนนโดยเฉลี่ยเท่าไหร่

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

1. ให้คำนวณแต่ละเดือนมีจำนวนผู้บาดเจ็บและเสียชีวิตจากอุบัติเหตุทางถนนโดยเฉลี่ยเท่าไหร่

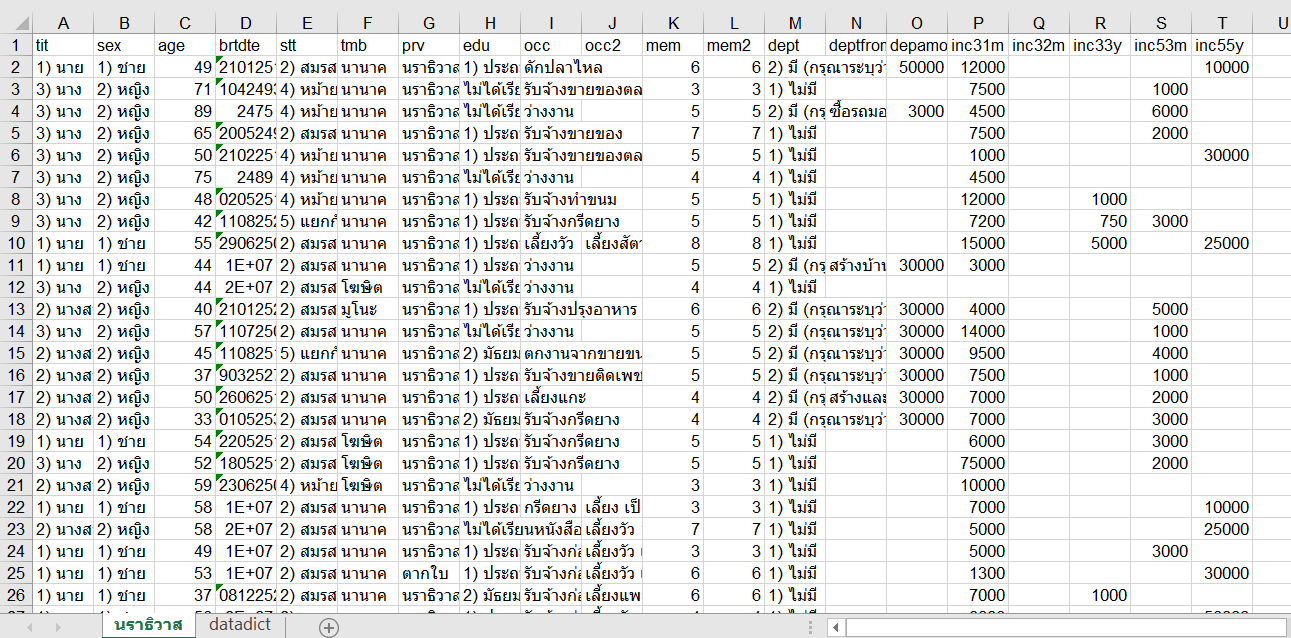
|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์) |

1. ให้หาในแต่ละปีโดยส่วนใหญ่ยานพาหนะประเภทใดที่ประสบอุบัติเหตุทางถนน

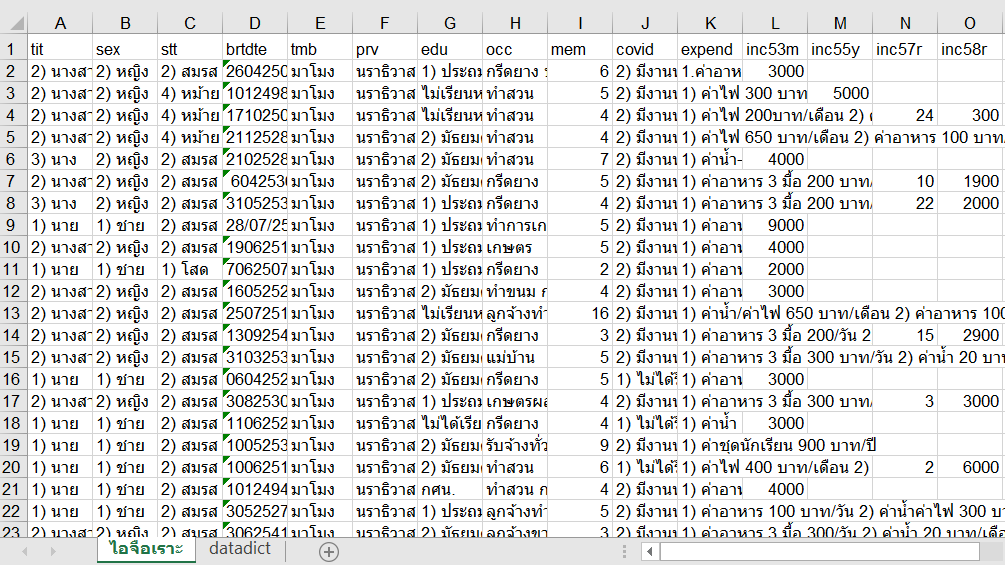
|  |
| --- |
| คำตอบ (แสดงคำสั่งและผลลัพธ์) |

1. **ให้นักศึกษาฝึกสำรวจข้อมูลโดยใช้ชุดข้อมูล “****CBR64-Nara1\_practice.xlsx” และ “CBR64-Nara2\_practice.xlsx” มีคำอธิบายตัวแปรดังนี้**

CBR64-Nara1\_practice.xlsx



CBR64-Nara2\_practice.xlsx



1. ให้ตรวจสอบข้อมูล CBR64-Nara1\_practice.xlsx และ CBR64-Nara2\_practice.xlsx มีจำนวนกี่แถวกี่คอลัมน์

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

1. ให้ทำการรวมข้อมูล CBR64-Nara1\_practice.xlsx และ CBR64-Nara2\_practice.xlsx หลังจากรวมแล้วข้อมูลมีจำนวนกี่แถวกี่คอลัมน์ (รวมทางแถว)

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

1. กลุ่มตัวอย่างมีอายุเฉลี่ยเท่าไหร่ มากที่สุดเท่าไหร่ และน้อยที่สุดเท่าไหร่

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

1. กลุ่มตัวอย่างมีจำนวนสมาชิกในบ้านโดยเฉลี่ยเท่าไหร่ มากที่สุดเท่าไหร่ และน้อยที่สุดเท่าไหร่

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

1. กลุ่มตัวอย่างมีรายได้ต่อเดือนโดยเฉลี่ยเท่าไหร่ มากที่สุดเท่าไหร่ และน้อยที่สุดเท่าไหร่

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