

ปฏิบัติการที่ 3

การทำความสะอาดข้อมูลและจัดการข้อมูล

- วัตถุประสงค์:**
1. เพื่อให้นักศึกษาฝึกปฏิบัติทำความสะอาดและจัดการข้อมูลด้วยโปรแกรม R
 2. เพื่อให้นักศึกษาฝึกปฏิบัติผานข้อมูลจากหลาย 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 การจัดการข้อมูลในกรณีนี้ที่ตัวแปรตามมีหลายตัวเพื่อให้ง่ายต่อการวิเคราะห์ข้อมูลและง่ายในการอธิบายผลการวิเคราะห์ข้อมูล

1. จาก data file “childThepa_practice.xlsx” ให้นักศึกษาฝึกปฏิบัติทำความสะอาดและจัดการข้อมูล

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	childid	sex	parity	mobidity	gestationalAge	weight0mo	height0mo	weight24mo	height24mo	milk	fomula	rice	cereal	banana	papaya
2	389	2	2	0	38	2.92	49	11.1	88	3	1	4	7	2	9
3	265	1	2	0	39	3.04	51	11.3	85	12	1	3	6	2	7
4	266	1	2	0	40	3.7	51	15.5	89	2	1	4	-1	3	10
5	267	1	1	0	41	3.25	50	12	85.6	2	1	4	-1	4	-1
6	268	1	2	0	39	3.5	50	14.1	88	4	3	6	8	4	7
7	367	2	1	0	40	3	50	11.1	87.2	3	1	1	-1	-1	-1
8	269	2	1	0	38	2.78	49	14.7	85.5	1	1	3	6	2	5
9	270	1	3	0	41	3.7	52	15	92.5	3	1	4	8	4	12
10	271	1	3	0	38	2.73	51	10.3	83	5	1	5	9	4	10
11	272	1	1	0	40	2.98	49	16.6	88.6	3	1	2	2	2	9
12	358	1	3	0	38	3.58	52	20.4	92	3	3	4	7	3	-1
13	273	1	1	0	37	2.37	46	11.1	81.5	2	2	4	-1	4	-1
14	274	1	1	1	40	3.03	51	15.4	88.2	1	1	-1	3	-1	10
15	275	1	3	0	38	3.31	50	12.1	85.5	1	1	4	8	4	-1
16	276	1	-9	0	40	3.47	52	12.7	85.6	13	2	3	-1	3	3
17	277	1	2	1	-9	3.06	49	20.3	93	2	1	5	-1	4	-1
18	278	1	2	0	39	3.64	52	12	85.5	1	1	3	7	3	9
19	277	1	2	1	-9	3.06	49	20.3	93	2	1	5	-1	4	-1
20	279	1	3	0	39	3.39	50	17.1	89	3	1	4	0	3	7
21	280	2	4	0	37	4.31	52	13.2	89	9	2	3	6	2	6

```
# 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)
> str(dt)
'data.frame': 282 obs. of 25 variables:
 $ childid      : num 389 265 266 267 268 367 269 270 271 272 ...
 $ sex          : num 2 1 1 1 1 2 2 1 1 1 ...
 $ parity       : num 2 2 2 1 2 1 1 3 3 1 ...
 $ mobility     : num 0 0 0 0 0 0 0 0 0 0 ...
 $ gestationalAge: num 38 39 40 41 39 40 38 41 38 40 ...
 $ weight0mo    : num 2.92 3.04 3.7 3.25 3.5 3 2.78 3.7 2.73 2.98 ...
 $ height0mo    : num 49 51 51 50 50 50 49 52 51 49 ...
 $ weight24mo   : num 11.1 11.3 15.5 12 14.1 11.1 14.7 15 10.3 16.6 ...
 $ height24mo   : num 88 85 89 85.6 88 87.2 85.5 92.5 83 88.6 ...
 $ milk         : num 3 12 2 2 4 3 1 3 5 3 ...
 $ formula      : num 1 1 1 1 3 1 1 1 1 1 ...
 $ rice         : num 4 3 4 4 6 1 3 4 5 2 ...
 $ cereal       : num 7 6 -1 -1 8 -1 6 8 9 2 ...
```

a) ตรวจสอบข้อมูลซ้ำซ้อนโดยใช้ตัวแปร childid, motherid, familyid ประกอบกัน

```
> dt$checkid <- paste(dt$childid, dt$motherid, dt$familyid)
> dup <- duplicated(dt$checkid)
> table(dup)
```

```
> dt$checkid <- paste(dt$childid, dt$motherid, dt$familyid)
> dup <- duplicated(dt$checkid)
> table(dup)
dup
FALSE TRUE
264    18
```

เขียนคำอธิบายพอสังเขป

.....

.....

.....

```
# แสดงข้อมูลซ้ำ
```

```
> dt[dup==T,]
```

```
> dt[dup==T,]
  childid sex parity mobility gestationalAge weight0mo
18      277   1     2         1             NA      3.06
30      277   1     2         1             NA      3.06
53      367   2     1         0             40      3.00
74      315   1     2         0             37      3.06
79      315   1     2         0             37      3.06
89      315   1     2         0             37      3.06
94      367   2     1         0             40      3.00
104     358   1     3         0             38      3.58
113     367   2     1         0             40      3.00
136     389   2     2         0             38      2.92
151     486   1     1         0             36      2.21
155     367   2     1         0             40      3.00
167     389   2     2         0             38      2.92
236     486   1     1         0             36      2.21
237     367   2     1         0             40      3.00
241     389   2     2         0             38      2.92
281     521   1     2         0             39      3.50
282     367   2     1         0             40      3.00
```

```
> dt[dt$childid==277,]
```

```
> dt[dt$childid==277,]
  childid sex parity mobility gestationalAge weight0mo height0mo weight24mo height24mo milk
16      277   1     2         1             NA      3.06      49      20.3      93      2
18      277   1     2         1             NA      3.06      49      20.3      93      2
30      277   1     2         1             NA      3.06      49      20.3      93      2
  fomula rice cereal banana papaya eggYolk wholeEgg liver fish chicken beefPork vegetable
16      1   5     NA      4     NA      5       7     NA      7       7       8       5
18      1   5     NA      4     NA      5       7     NA      7       7       8       5
30      1   5     NA      4     NA      5       7     NA      7       7       8       5
  otherVegetable motherid familyid checkid
16              5      541      13 277 541 13
18              5      541      13 277 541 13
30              5      541      13 277 541 13
```

```
# ตัดข้อมูลที่ซ้ำออก
```

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

```
> str(dt1)
```

```
> dt1 <- dt[dup==F,]
> str(dt1)
'data.frame': 264 obs. of 26 variables:
 $ childid      : num 389 265 266 267 268 367 269 270 271 272 ...
 $ sex          : num 2 1 1 1 1 2 2 1 1 1 ...
 $ parity       : num 2 2 2 1 2 1 1 3 3 1 ...
 $ mobility     : num 0 0 0 0 0 0 0 0 0 0 ...
 $ gestationalAge: num 38 39 40 41 39 40 38 41 38 40 ...
 $ weight0mo    : num 2.92 3.04 3.7 3.25 3.5 3 2.78 3.7 2.73 2.98 ...
 $ height0mo    : num 49 51 51 50 50 50 49 52 51 49 ...
 $ weight24mo   : num 11.1 11.3 15.5 12 14.1 11.1 14.7 15 10.3 16.6 ...
 $ height24mo   : num 88 85 89 85.6 88 87.2 85.5 92.5 83 88.6 ...
 $ milk         : num 3 12 2 2 4 3 1 3 5 3 ...
 $ fomula       : num 1 1 1 1 3 1 1 1 1 1 ...
```

เขียนคำอธิบายโดยเปรียบเทียบกับข้อมูลก่อนลบข้อมูลซ้ำ

b) ตรวจสอบข้อมูลตัวแปรทั้งหมดด้วยคำสั่ง summary()

```
> summary(dt1)
```

คำตอบ (แสดงคำสั่งและผลลัพธ์)

```
> summary(dt1)
  childid      sex      parity      mobility      gestationalAge      weight0mo
Min.   :265.0  Min.   :1.000  Min.   : -9.0000  Min.   :0.0000  Min.   : -9.00  Min.   :1.750
1st Qu.:330.8  1st Qu.:1.000  1st Qu.: 1.0000  1st Qu.:0.0000  1st Qu.:38.00  1st Qu.:2.900
Median :396.5  Median :1.000  Median : 2.0000  Median :0.0000  Median :39.00  Median :3.145
Mean   :396.5  Mean   :1.496  Mean   : 0.3939  Mean   :0.0303  Mean   :37.65  Mean   :3.179
3rd Qu.:462.2  3rd Qu.:2.000  3rd Qu.: 2.0000  3rd Qu.:0.0000  3rd Qu.:40.00  3rd Qu.:3.450
Max.   :528.0  Max.   :2.000  Max.   : 6.0000  Max.   :1.0000  Max.   :44.00  Max.   :5.220
 height0mo  weight24mo  height24mo      milk      fomula
Min.   : -9.00  Min.   : 8.40  Min.   : -21.00  Min.   : -9.000  Min.   : -1.000
1st Qu.:49.00  1st Qu.:11.20  1st Qu.: 84.00  1st Qu.: 1.000  1st Qu.: 1.000
Median :50.00  Median :12.10  Median : 86.00  Median : 3.000  Median : 1.000
Mean   :49.34  Mean   :12.75  Mean   : 85.83  Mean   : 5.409  Mean   : 2.432
3rd Qu.:51.00  3rd Qu.:13.82  3rd Qu.: 88.20  3rd Qu.: 6.000  3rd Qu.: 2.000
Max.   :55.00  Max.   :23.40  Max.   :98.00  Max.   :24.000  Max.   :14.000
   rice      cereal      banana      papaya      eggYolk
Min.   : -1.000  Min.   : -1.000  Min.   : -1.000  Min.   : -1.000  Min.   : -1.000
1st Qu.: 3.000  1st Qu.: 0.000  1st Qu.: 2.000  1st Qu.: 0.000  1st Qu.: 3.000
Median : 4.000  Median : 4.000  Median : 3.000  Median : 5.500  Median : 4.000
Mean   : 3.523  Mean   : 3.644  Mean   : 2.902  Mean   : 4.807  Mean   : 3.845
3rd Qu.: 4.000  3rd Qu.: 6.000  3rd Qu.: 4.000  3rd Qu.: 8.000  3rd Qu.: 4.250
Max.   : 9.000  Max.   :12.000  Max.   :10.000  Max.   :12.000  Max.   :10.000
 wholeEgg      liver      fish      chicken      beefPork
Min.   : -1.000  Min.   : -1.000  Min.   : -1.000  Min.   : -1.000  Min.   : -1.000
1st Qu.: 0.000  1st Qu.: 4.000  1st Qu.: 4.000  1st Qu.: 5.000  1st Qu.: 4.000
Median : 7.000  Median : 5.000  Median : 5.000  Median : 7.000  Median : 6.000
Mean   : 5.201  Mean   : 4.595  Mean   : 4.909  Mean   : 5.758  Mean   : 5.64
3rd Qu.: 8.000  3rd Qu.: 6.000  3rd Qu.: 7.000  3rd Qu.: 8.000  3rd Qu.: 8.000
Max.   :12.000  Max.   :11.000  Max.   :11.000  Max.   :12.000  Max.   :12.000
 vegetable  otherVegetable  motherid      familyid      checkid
Min.   : -1.000  Min.   : -1.000  Min.   :529.0  Min.   : 1.00  Length:264
1st Qu.: 4.000  1st Qu.: 4.000  1st Qu.:594.8  1st Qu.: 66.75  Class :character
Median : 5.000  Median : 5.000  Median :660.5  Median :132.50  Mode  :character
Mean   : 4.356  Mean   : 4.553  Mean   :660.5  Mean   :132.50
3rd Qu.: 5.000  3rd Qu.: 5.000  3rd Qu.:726.2  3rd Qu.:198.25
Max.   :11.000  Max.   :11.000  Max.   :792.0  Max.   :264.00
```

เขียนคำอธิบายพอสังเขป

.....

.....

.....

.....

.....

.....

c) ทุกตัวแปรที่มีค่าเป็นลบให้แทนค่าด้วย NA

```
> for (i in c(2:23)) {  
> dt1[,i] <- ifelse(dt1[,i]<0,NA,dt1[,i])  
> }  
> summary(dt1)
```

คำตอบ (แสดงคำสั่งและผลลัพธ์)

เขียนคำอธิบายโดยเปรียบเทียบกับข้อมูลก่อนหน้า

.....

.....

.....

.....

d) กำหนดให้ตัวแปร sex และ mobility เป็นตัวแปร factor พร้อมให้คำอธิบายรหัสของแต่ละตัวแปร (โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ)

```
# ตัวแปร sex
```

```
> dt1$sex <- factor(dt1$sex)
> levels(dt1$sex) <- c("Male","Female")
```

คำตอบ (แสดงคำสั่งและผลลัพธ์)

- e) 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)

- f) 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)
```

```
> mt <- read_excel("childThepa_practice.xlsx", sheet="mother")
> mt <- data.frame(mt)
> str(mt)
'data.frame': 220 obs. of 5 variables:
 $ motherid      : num  738 597 585 765 714 685 672 609 641 577 ...
 $ motherReligion: num   9  9  9  9  1  1  1  1  1  1 ...
 $ motherAge     : num   9  9  9  9 22 30 28 34 32 28 ...
 $ motherEdu     : num   1  1  1  1  2  2  2  2  2  2 ...
 $ motherOcc     : num   1  1  1  1  1  1  1  1  1  1 ...
```

```
# Merger file dt1 and mt (เด็กและมารดา)
```

```
> dt2 <- merge(dt1, mt, by.x="motherid", by.y="motherid", all.x=T)
```

```
> str(dt2)
```

คำตอบ (แสดงคำสั่งและผลลัพธ์)

- g) ให้ตรวจสอบ ทำความสะอาด และจัดการข้อมูลมารดาของตัวแปร motherReligion, motherAge, motherEdu และ motherOcc หากมีค่าที่ผิดพลาด ให้จัดการข้อมูลตัวแปร motherReligion (โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ)

คำตอบ (แสดงคำสั่งและผลลัพธ์)

ตัวแปร motherAge (โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ)

คำตอบ (แสดงคำสั่งและผลลัพธ์)

ตัวแปร motherEdu (โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ)

คำตอบ (แสดงคำสั่งและผลลัพธ์)

ตัวแปร motherOcc (โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ)

- h) ให้ฝึกลูกข้อมูลเด็กและครอบครัว (sheet->family) และตรวจสอบ ทำความสะอาดตัวแปร income และ SES หากมีค่าที่ผิดพลาด ให้จัดการข้อมูล (โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ)

คำตอบ (แสดงคำสั่งและผลลัพธ์)

2. ให้นักศึกษาฝึกสำรวจ ทำความสะอาด และจัดการข้อมูลโดยใช้ชุดข้อมูล

“childPattani_practice.xlsx”

- a) อ่าน/นำเข้า data sheet “data” ตั้งชื่อเป็น “mem” ซึ่งเป็นข้อมูลของสมาชิกทั้งหมดในแต่ละครอบครัว และอ่าน/นำเข้า data sheet “family member” ตั้งชื่อเป็น “fam”

	A	B	C	D	E	F	G	H	I
1	perid_pre	titme	age	gender	tmb	amp	prv	hholdid	comment
2	2158	2	4	2	บางดาวา	หนองจิก	ปัตตานี		0 2เด็กอายุ 3-25 ปี
3	6627	เด็กหญิง	2	2					0 3สมาชิกในครัวเรือน
4	6626	เด็กหญิง	4	2					0 3สมาชิกในครัวเรือน
5	6625	นาง	33	2					0 3สมาชิกในครัวเรือน
6	477		5		บางดาวา	หนองจิก	ปัตตานี		0 1ผู้ให้ข้อมูล
7	6629	นางสาว	21	2					0 3สมาชิกในครัวเรือน
8	6623	นางสาว	62	2					0 3สมาชิกในครัวเรือน
9	6628	นาย	57	1					0 3สมาชิกในครัวเรือน
10	6622	เด็กชาย		1					0 3สมาชิกในครัวเรือน
11	6624	นาย	27	1					0 3สมาชิกในครัวเรือน
12	738		4		นาเกตุ	โคกโพธิ์	ปัตตานี		1 1ผู้ให้ข้อมูล
13	6630	เด็กหญิง	3	2					3 3สมาชิกในครัวเรือน
14	2159	2	3	2	นาเกตุ	โคกโพธิ์	ปัตตานี		3 2เด็กอายุ 3-25 ปี
15	6633	นาย	35	1					3 3สมาชิกในครัวเรือน
16	6631	เด็กหญิง	4	2					3 3สมาชิกในครัวเรือน

	A	B	C	D	E	F	G	H	I	J	K
1	perid_pre	hholdid	a1	a2titnme	a3y	a3m	a4	a9	a10	a11	a12
2	9745	941001003	10	เด็กหญิง	2		หญิง	1:ได้	1:ได้	0:ไม่ได้เรียน	0:ไม่ได้เรียน
3	5232	940400730	11	เด็กชาย	12		ชาย	1:ได้	1:ได้	4:ม.ปลาย	0:ไม่ได้เรียน
4	9752	941001003	7	นาย	17		ชาย	1:ได้	1:ได้	2:ประถมศึกษา	0:ไม่ได้เรียน
5	6038	941001079	5	นาย	18		ชาย	1:ได้	1:ได้	4:ม.ปลาย	0:ไม่ได้เรียน
6	6039	941001079	4	นาย	20		ชาย	1:ได้	1:ได้	1:ต่ำกว่าป	0:ไม่ได้เรียน
7	6359	941002636	8	นาย	22		ชาย	0:ไม่ได้	0:ไม่ได้	0:ไม่ได้เรียน	0:ไม่ได้เรียน
8	9748	941001003	11	นาย	23		ชาย	1:ได้	1:ได้	2:ประถมศึกษา	0:ไม่ได้เรียน
9	9743	941001003	4	นางสาว	23		หญิง	1:ได้	1:ได้	0:ไม่ได้เรียน	0:ไม่ได้เรียน
10	9741	941001003	2	นางสาว	25		หญิง	1:ได้	0:ไม่ได้	0:ไม่ได้เรียน	0:ไม่ได้เรียน
11	9744	941001003	3	นางสาว	27		หญิง	1:ได้	1:ได้	0:ไม่ได้เรียน	0:ไม่ได้เรียน
12	8963	940200983	2	นางสาว	28		หญิง	1:ได้	1:ได้	6:ปวส.หรือ	0:ไม่ได้เรียน
13	6355	941002636	5	นาย	32		ชาย	0:ไม่ได้	0:ไม่ได้	0:ไม่ได้เรียน	0:ไม่ได้เรียน
14	8403	941001806	1	นางสาว	35		หญิง	1:ได้	1:ได้	0:ไม่ได้เรียน	0:ไม่ได้เรียน
15	9750	941001003	5	นาย	37		ชาย	1:ได้	1:ได้	2:ประถมศึกษา	0:ไม่ได้เรียน
16	7774	940202564	5	นาย	50		ชาย	0:ไม่ได้	0:ไม่ได้	0:ไม่ได้เรียน	0:ไม่ได้เรียน

คำตอบ (แสดงคำสั่งและบางส่วนของผลลัพธ์)

- b) ให้ผสาน data frame “mem” และ “fam” ตั้งชื่อเป็น “dt”

คำตอบ (แสดงคำสั่งและบางส่วนของผลลัพธ์)

- c) ถ้าตัวแปร age เป็นค่าว่าง ให้แทนที่ด้วยค่าของข้อมูลจากตัวแปร a3y และถ้า gender เป็นค่าว่าง ให้แทนค่าของข้อมูลจากตัวแปร a4 (โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ)

คำตอบ (แสดงคำสั่งและผลลัพธ์)

- d) ให้ตรวจสอบตัวแปร titime แล้วแก้ไขข้อมูลดังนี้ (โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ)

เด็กชาย → 1

เด็กหญิง → 2

นาง → 3

นาย → 4

นางสาว → 5

คำตอบ (แสดงคำสั่งและผลลัพธ์)

- e) ให้ตรวจสอบตัวแปร gender แล้วแก้ไขข้อมูลดังนี้ (โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ)

ชาย → 1

หญิง → 2

- f) ตรวจสอบค่านำหน้าชื่อกับเพศว่าสอดคล้องกันหรือไม่ หากไม่สอดคล้องให้ทำการแก้ไขข้อมูล (โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ)

คำตอบ (แสดงคำสั่งและผลลัพธ์)

- g) ให้เรียงลำดับข้อมูลตามตัวแปร tmb จากมากไปน้อย

คำตอบ (แสดงคำสั่งและบางส่วนของผลลัพธ์)

- h) ให้ตรวจสอบข้อมูลซ้ำ โดยใช้ตัวแปร titnme, age, gender, hholdid เป็นเงื่อนไข และลบข้อมูลซ้ำ

คำตอบ (แสดงคำสั่งและบางส่วนของผลลัพธ์)

- i) ให้เลือกเฉพาะข้อมูลของสมาชิกที่มีอายุ 15-50 ปี เก็บไว้ใน data frame ใหม่ โดยตั้งชื่อเป็น "dt1"

คำตอบ (แสดงคำสั่งและบางส่วนของผลลัพธ์)

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

คำตอบ (แสดงคำสั่งและผลลัพธ์)

3. จากข้อมูล "teacherstress_practice.xlsx" และคำอธิบายข้อมูล "teachstress_Q.docx" ให้
นักศึกษาฝึกสำรวจข้อมูลและจัดการข้อมูลของตัวแปรต่าง ๆ
(รูปข้างล่างนี้ แสดงข้อมูลบางส่วนเท่านั้น)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	sex	age	teachdur	edu	status	children	inc	t1	t2	t3	t4	t5	t6	t7
2	หญิง	24	2	ปริญญาตรี	แต่งงาน	4	? 10,000 บาท	มากที่สุด	มากที่สุด	มากที่สุด	ปานกลาง	มาก	ปานกลาง	มากที่สุด
3	หญิง	24	1	ปริญญาตรี	แต่งงาน	1	10,001 – 20,000 บาท	มาก	มาก	น้อย	ปานกลาง	มาก	ปานกลาง	ปานกลาง
4	หญิง	25	2	ปริญญาตรี	แต่งงาน	-	? 10,000 บาท	มากที่สุด	มาก	มาก	มาก	มาก	มาก	มาก
5	หญิง	25	1	ปริญญาตรี	แต่งงาน	1	? 10,000 บาท	ปานกลาง	ปานกลาง	มาก	ปานกลาง	ปานกลาง	ปานกลาง	น้อย
6	หญิง	25	1	ปริญญาตรี	แต่งงาน	0	10,001 – 20,000 บาท	มากที่สุด	มากที่สุด	ปานกลาง	น้อยที่สุด	มากที่สุด	น้อยที่สุด	น้อย
7	หญิง	26	1ปี	ปริญญาตรี	แต่งงาน	0	? 10,000 บาท	ปานกลาง	มาก	ปานกลาง	มาก	มากที่สุด	มาก	มากที่สุด
8	หญิง	26	2	ปริญญาตรี	แต่งงาน	-	? 10,000 บาท	มาก	มาก	ปานกลาง	ปานกลาง	มาก	มากที่สุด	มาก
9	หญิง	26	4	ปริญญาตรี	แต่งงาน	0	? 10,000 บาท	ปานกลาง	ปานกลาง	ปานกลาง	น้อย	น้อย	น้อย	น้อย
10	หญิง	27	3	ปริญญาตรี	แต่งงาน	-	? 10,000 บาท	มาก	มาก	มาก	มาก	มาก	ปานกลาง	มาก
11	หญิง	27	3	ปริญญาตรี	แต่งงาน	0	? 10,000 บาท	ปานกลาง	มาก	น้อย	ปานกลาง	มาก	ปานกลาง	ปานกลาง
12	หญิง	27	4	ปริญญาตรี	แต่งงาน	1	? 10,000 บาท	มาก	มาก	มาก	น้อย	มากที่สุด	ปานกลาง	มากที่สุด
13	หญิง	27	5	ปริญญาตรี	แต่งงาน	0	20,001 – 30,000 บาท	มากที่สุด	มากที่สุด	มาก	ปานกลาง	ปานกลาง	ปานกลาง	มาก
14	ชาย	27	2	ปริญญาตรี	แต่งงาน	-	10,001 – 20,000 บาท	มาก	มาก	มาก	มาก	มากที่สุด	มาก	มาก
15	หญิง	28	4	ปริญญาตรี	แต่งงาน	1	10,001 – 20,000 บาท	มากที่สุด	มาก	มาก	ปานกลาง	ปานกลาง	ปานกลาง	มาก

a) ให้นักศึกษาระบุประเภทของตัวแปรต่อไปนี้ โดยทำเครื่องหมายถูก (✓) ในช่องที่กำหนด

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

b) เพิ่มคอลัมน์ ID และกำหนดตัวเลขตั้งแต่ 1-n โดยที่ n คือ จำนวนข้อมูลทั้งหมด

คำตอบ (แสดงคำสั่งและบางส่วนของผลลัพธ์)

```
> dt <- dt[,c(30,1:29)]
> str(dt)
'data.frame': 230 obs. of 30 variables:
 $ id      : int  1 2 3 4 5 6 7 8 9 10 ...
 $ sex     : chr  "หญิง" "หญิง" "หญิง" "หญิง" ...
 $ age     : num  24 24 25 25 25 26 26 26 27 27 ...
 $ teachdur: chr   "2" "1" "2" "1" ...
 $ edu     : chr   "ปริญญาตรี" "ปริญญาตรี" "ปริญญาตรี" "ปริญญาตรี" ...
 $ status  : chr   "แต่งงาน" "แต่งงาน" "แต่งงาน" "แต่งงาน" ...
 $ children: chr   "4" "1" "0" "1" ...
 $ inc     : chr   "? 10,000 บาท" "10,001 – 20,000 บาท" "? 10,000 บาท" "? 10,000 บาท" ...
 $ t1      : chr   "มากที่สุด" "มาก" "มากที่สุด" "ปานกลาง" ...
 $ t2      : chr   "มากที่สุด" "มาก" "มากที่สุด" "ปานกลาง" ...
 $ t3      : chr   "มากที่สุด" "มาก" "มากที่สุด" "ปานกลาง" ...
 $ t4      : chr   "ปานกลาง" "น้อย" "มากที่สุด" "ปานกลาง" ...
 $ t5      : chr   "ปานกลาง" "มากที่สุด" "ปานกลาง" "ปานกลาง" ...
 $ t6      : chr   "ปานกลาง" "มากที่สุด" "ปานกลาง" "ปานกลาง" ...
 $ t7      : chr   "ปานกลาง" "ปานกลาง" "มากที่สุด" "น้อย" ...
 $ t8      : chr   "ปานกลาง" "ปานกลาง" "มากที่สุด" "ปานกลาง" ...
 $ t9      : chr   "มากที่สุด" "มากที่สุด" "ปานกลาง" "ปานกลาง" ...
```

```
dt <- read_excel("teacherstress_practice (1).xlsx")
dt <- data.frame(dt)
str(dt)

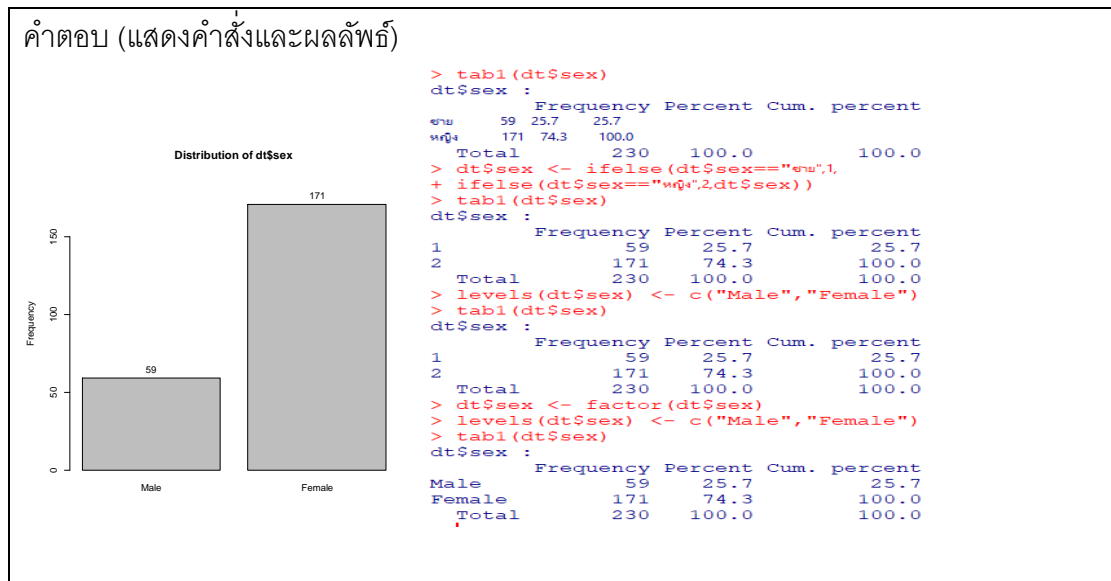
summary(dt)
tab1(dt$teachdur)

dt$id <- c(1:230)
str(dt)

dt <- dt[,c(30,1:29)]
str(dt)

tab1(dt$sex)
```

- c) สํารวจข้อมูลตัวแปร sex หากมีค่าที่ผิดพลาดให้ทำการจัดการข้อมูล (โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ)



- d) สํารวจข้อมูลตัวแปร age หากมีค่าที่ผิดพลาดให้ทำการจัดการข้อมูล (โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ)

คำตอบ (แสดงคำสั่งและผลลัพธ์)

```

> summary(dt$age)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
14.00  29.25   35.00   35.56  40.00   61.00
> dt$age <- ifelse(dt$age==14,NA,dt$age)
> summary(dt$age)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's
22.00  30.00   35.00   35.75  40.00   61.00      2
> |

```

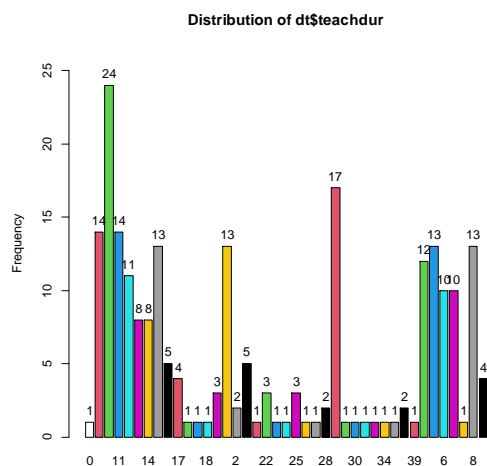
- e) สํารวจข้อมูลตัวแปร teachdur หากมีค่าที่ผิดพลาดให้ทำการจัดการข้อมูล (โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ)

คำตอบ (แสดงคำสั่งและผลลัพธ์)

```

> summary(dt$teachdur)
  Length      Class      Mode 
    230 character character 
> dt$teachdur <- ifelse(dt$teachdur=="7เดือน",0.7,dt$teachdur)
> summary(dt$teachdur)
  Length      Class      Mode 
    230 character character 
> dt$teachdur <- as.numeric(dt$teachdur)
> summary(dt$teachdur)
  Min. 1st Qu.  Median     Mean 3rd Qu.    Max. 
  0.0     4.0     10.0    10.3   14.0    39.0 
> dt$teachdur <- sub("ปี","",dt$teachdur)
> tab1(dt$teachdur)
dt$teachdur :
      Frequency Percent  Cum. percent
0             1      0.4         0.4
1            14      6.1         6.5
10           24     10.4        17.0
11           14      6.1        23.0
12           11      4.8        27.8
13            8      3.5        31.3
14            8      3.5        34.8
15           13      5.7        40.4
16            5      2.2        42.6
17            4      1.7        44.3
17            1      0.4        44.8
18            1      0.4        45.2
18            1      0.4        45.7
19            3      1.3        47.0
2            13      5.7        52.6
2             2      0.9        53.5
20            5      2.2        55.7
21            1      0.4        56.1
22            3      1.3        57.4
23            1      0.4        57.8
24            1      0.4        58.3
25            3      1.3        59.6

```



- f) สํารวจข้อมูลตัวแปร edu หากมีค่าที่ผิดพลาดให้ทำการจัดการข้อมูล (โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ)

คำตอบ (แสดงคำสั่งและผลลัพธ์)

```
> tab1(dt$edu)
dt$edu :
      Frequency Percent Cum. percent
ป.บัณฑิต      1   0.4    0.4
ปริญญาตรี     202  87.8    88.3
ปริญญาโท       23  10.0    98.3
ม.ต้น           1   0.4    98.7
สูงกว่าปริญญาโท  2   0.9    99.6
อนุปริญญา      1   0.4   100.0
Total          230  100.0   100.0
> dt$edu <- ifelse(dt$edu=="ม.ต้น",NA,dt$edu)
> dt$edu <- ifelse(dt$edu %in% c("ป.บัณฑิต","ปริญญาตรี","อนุปริญญา"),1,
+ ifelse(dt$edu %in% c("ปริญญาโท","สูงกว่าปริญญาโท"),2,dt$edu))
> dt$edu <- factor(dt$edu)
> levels(dt$edu) <- c("bsc","master+")
> tab1(dt$edu)
dt$edu :
      Frequency % (NA+) % (NA-)
bsc           204   88.7   89.1
master+        25  10.9   10.9
NA's           1   0.4    0.0
Total          230  100.0  100.0
.
```

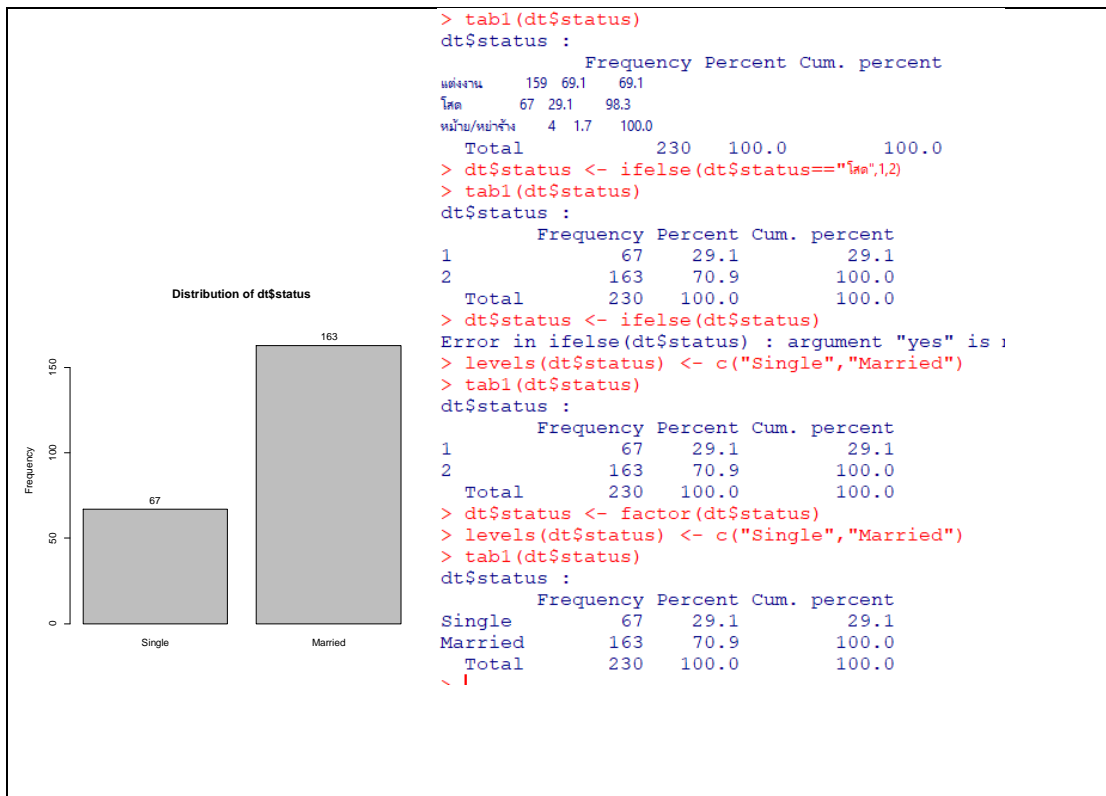
- g) ตัวแปร teachdur และ edu ตรวจสอบความสอดคล้องของข้อมูล หากมีค่าที่ผิดพลาดให้ทำการจัดการข้อมูล (โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ)

```
คำตอบ (แสดงคำสั่งและผลลัพธ์)
> tapply(dt$teachdur, dt$edu, mean, na.rm=T)
bsc master+
10.50739 9.44000
> tapply(dt$teachdur, dt$edu, min, na.rm=T)
bsc master+
1 1
> tapply(dt$teachdur, dt$edu, max, na.rm=T)
bsc master+
39 25
> summ(dt$teachdur, by=dt$edu, graph=F)
For dt$edu = bsc
obs. mean median s.d. min. max.
203 10.507 10 7.954 1 39

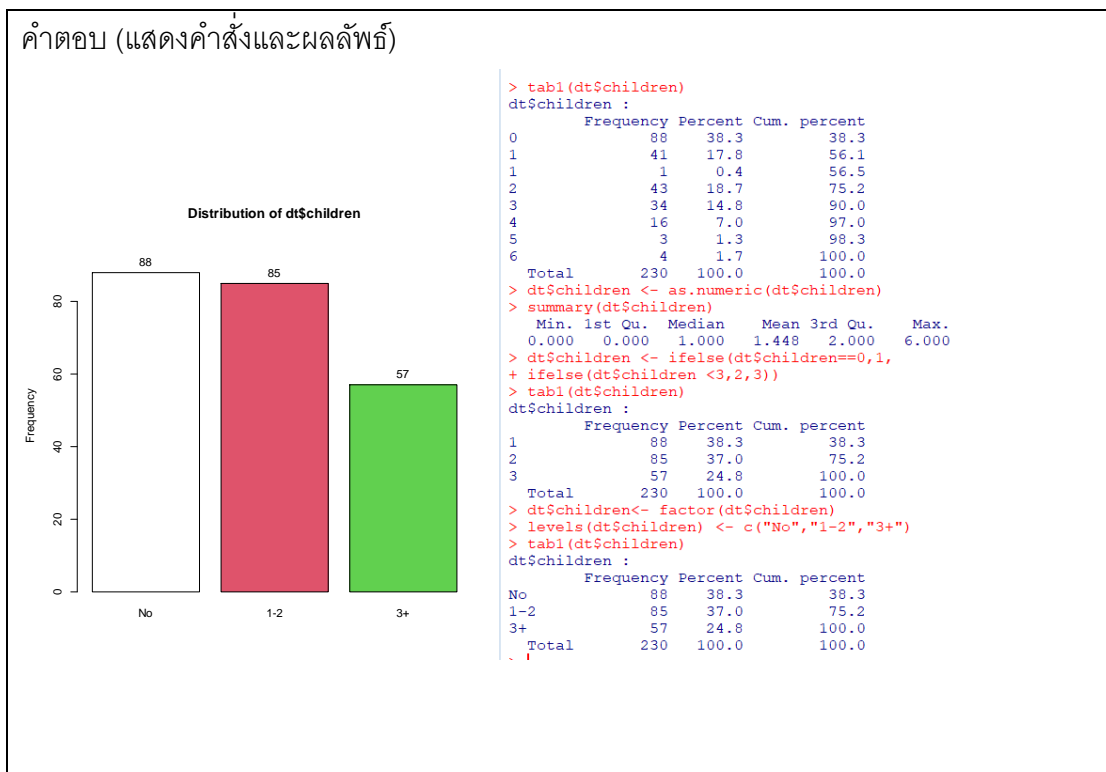
For dt$edu = master+
obs. mean median s.d. min. max.
25 9.44 8 5.355 1 25
```

- h) สํารวจข้อมูลตัวแปร status หากมีค่าที่ผิดพลาดให้ทำการจัดการข้อมูล (โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ)

```
คำตอบ (แสดงคำสั่งและผลลัพธ์)
```



- i) สํารวจข้อมูลตัวแปร children หากมีค่าที่ผิดพลาดให้ทำการจัดการข้อมูล (โดยแสดงผลพื้ในรูปแบบตารางหรือกราฟ)

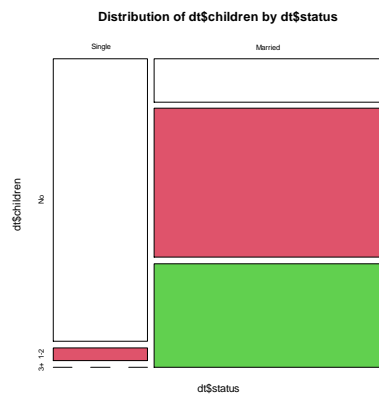


- j) ตัวแปร status และ children ตรวจสอบความสอดคล้องของข้อมูล หากมีค่าที่ผิดพลาดให้ทำการจัดการข้อมูล (โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ)

คำตอบ (แสดงคำสั่งและผลลัพธ์)

```
> table(dt$status, dt$children)
```

```
      No 1-2 3+
Single 64   3  0
Married 24  82 57
```

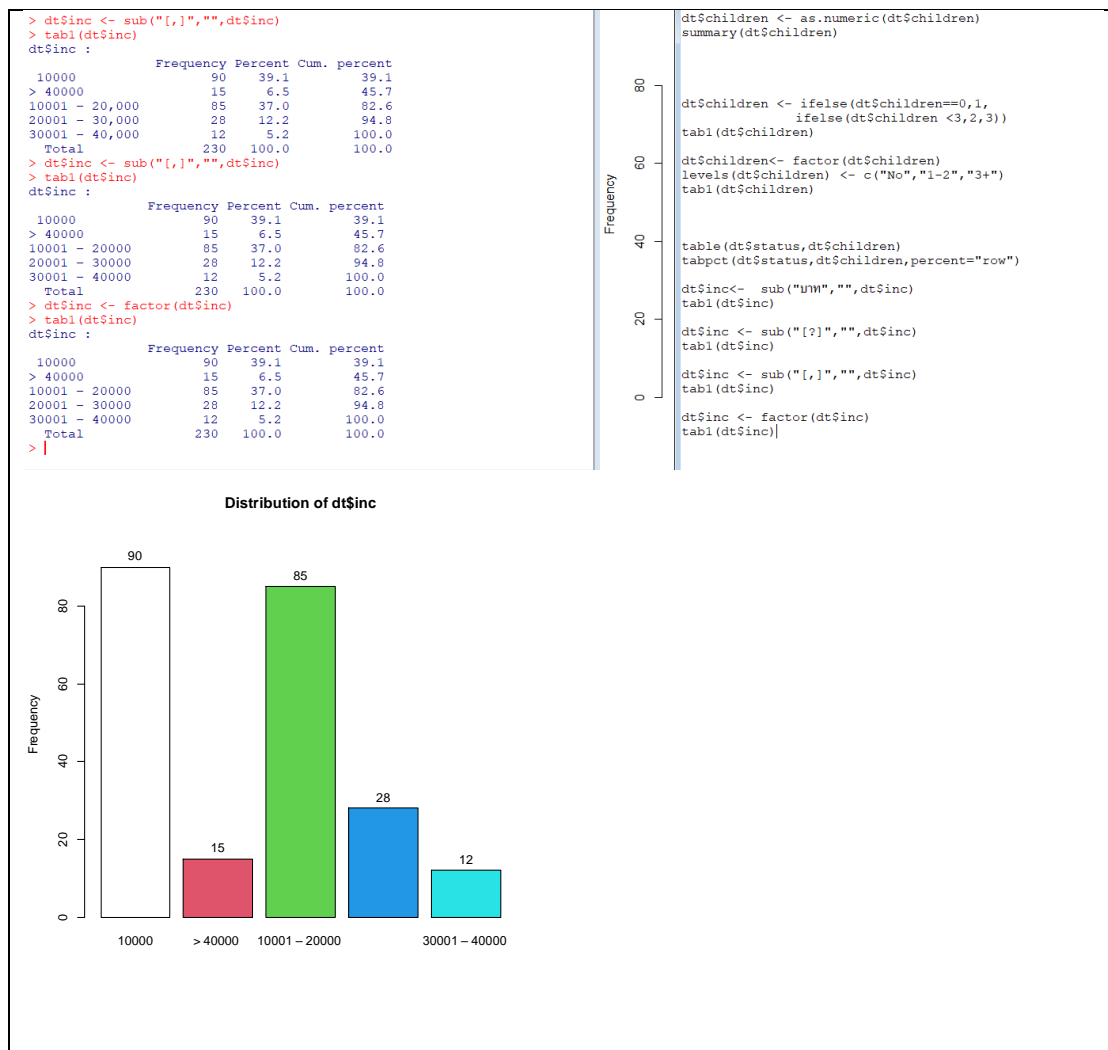


```
> tabpct(dt$status, dt$children, percent="row")
```

```
Row percent
      dt$children
dt$status      No      1-2      3+      Total
Single         64         3         0         67
              (95.5)    (4.5)    (0)    (100)
Married        24        82        57        163
              (14.7)   (50.3)   (35)   (100)
```

- k) สํารวจข้อมูลตัวแปร inc หากมีค่าที่ผิดพลาดให้ทำการจัดการข้อมูล (โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ)

คำตอบ (แสดงคำสั่งและผลลัพธ์)



- l) 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)

```

> str(dt)
'data.frame': 230 obs. of 30 variables:
 $ id      : int  1 2 3 4 5 6 7 8 9 10 ...
 $ sex     : Factor w/ 2 levels "Male","Female": 2 2 2 2 2 2 2 2 2 ...
 $ age     : num  24 24 25 25 25 26 26 26 27 27 ...
 $ teachdur: num  2 1 2 1 1 1 2 4 3 3 ...
 $ edu     : Factor w/ 2 levels "bsc","master+": 1 1 1 1 1 1 1 1 1 ...
 $ status  : Factor w/ 2 levels "Single","Married": 2 2 2 2 2 2 2 2 2 ...
 $ children: Factor w/ 3 levels "No","1-2","3+": 3 2 1 2 1 1 1 1 1 ...
 $ inc     : Factor w/ 5 levels "10000","> 40000",...: 1 3 1 1 3 1 1 1 1 ...
 $ t1      : num  5 4 5 3 5 4 4 3 4 3 ...
 $ t2      : num  5 4 4 3 5 3 4 3 4 4 ...
 $ t3      : num  5 4 4 4 3 4 3 3 4 4 ...
 $ t4      : num  3 2 4 3 1 5 3 2 4 2 ...
 $ t5      : num  4 3 4 3 5 4 4 2 4 3 ...
 $ t6      : num  3 4 4 3 1 5 5 2 3 4 ...
 $ t7      : num  5 3 4 2 2 4 4 2 4 3 ...
 $ t8      : num  5 3 4 3 5 5 4 2 4 4 ...
 $ t9      : num  4 4 4 3 1 4 4 4 4 4 ...
 $ t10     : num  2 1 4 4 1 5 1 1 3 1 ...
 $ m1      : num  3 3 4 2 2 4 1 1 4 1 ...
 $ m2      : num  4 3 4 3 1 4 2 1 5 2 ...
 $ st1     : chr  "มีงานแล้ว" "สองข้างบ่อ" "ไม่เลย" "มีงานแล้ว" ...
 $ st2     : chr  "มีงานแล้ว" "มีงานแล้ว" "ไม่เลย" "มีงานแล้ว" ...
 $ st3     : chr  "มีงานแล้ว" "มีงานแล้ว" "ไม่เลย" "มีงานแล้ว" ...
 $ st4     : chr  "สองข้างบ่อ" "สองข้างบ่อ" "ไม่เลย" "มีงานแล้ว" ...
 $ st5     : chr  "มีงานแล้ว" "สองข้างบ่อ" "ไม่เลย" "มีงานแล้ว" ...
 $ st6     : chr  "ไม่เลย" "มีงานแล้ว" "ไม่เลย" "มีงานแล้ว" ...
 $ st7     : chr  "มีงานแล้ว" "มีงานแล้ว" "ไม่เลย" "มีงานแล้ว" ...
 $ st8     : chr  "มีงานแล้ว" "มีงานแล้ว" "ไม่เลย" "มีงานแล้ว" ...
 $ st9     : chr  "มีงานแล้ว" "มีงานแล้ว" "ไม่เลย" "มีงานแล้ว" ...
 $ st10    : chr  "มีงานแล้ว" "มีงานแล้ว" "ไม่เลย" "มีงานแล้ว" ...
>

```

```

table(dt$status,dt$children)
tabpct(dt$status,dt$children,percent="row")

dt$inc<- sub("b","<=",dt$inc)
tabl(dt$inc)

dt$inc <- sub("[?]", "<=", dt$inc)

dt$inc <- sub("[,]", "", dt$inc)
dt$inc <- sub(" ", "", dt$inc)
dt$inc <- sub(" ", "", dt$inc)
tabl(dt$inc)

dt$inc <- factor(dt$inc)
tabl(dt$inc)

des(dt)

for(i in c(9:20)){
  dt[,i] <- ifelse(dt[,i]=="มากที่สุด",5,
    ifelse(dt[,i]=="มาก",4,
      ifelse(dt[,i]=="ปานกลาง",3,
        ifelse(dt[,i]=="น้อย",2,
          ifelse(dt[,i]=="น้อยที่สุด",1,dt[,i]
        ))))
  dt[,i] <- as.numeric(dt[,i])
}
str(dt)

```

m) ตัวแปร st1-st10 ให้แก้ไขข้อมูลดังนี้ โดยใช้คำสั่ง for ในการทำงานให้เร็วขึ้น

ไม่เคย แก้ไขเป็น 0

ส่วนใหญ่ไม่เคย แก้ไขเป็น 1

บางครั้ง แก้ไขเป็น 2

ค่อนข้างบ่อย แก้ไขเป็น 3

บ่อยมาก แก้ไขเป็น 4

คำตอบ (แสดงคำสั่งและบางส่วนของผลลัพธ์)

```

> str(dt)
'data.frame': 230 obs. of 30 variables:
 $ id      : int  1 2 3 4 5 6 7 8 9 10 ...
 $ sex     : Factor w/ 2 levels "Male","Female": 2 2 2 2 2 2 2 2 2 ...
 $ age     : num  24 24 25 25 25 26 26 26 27 27 ...
 $ teachdur: num  2 1 2 1 1 1 2 4 3 3 ...
 $ edu     : Factor w/ 2 levels "bsc","master+": 1 1 1 1 1 1 1 1 1 ...
 $ status  : Factor w/ 2 levels "Single","Married": 2 2 2 2 2 2 2 2 2 ...
 $ children: Factor w/ 3 levels "No","1-2","3+": 3 2 1 2 1 1 1 1 1 ...
 $ inc     : Factor w/ 5 levels "10000","> 40000",...: 1 3 1 1 3 1 1 1 1 ...
 $ t1      : num  5 4 5 3 5 4 4 3 4 3 ...
 $ t2      : num  5 4 4 3 5 3 4 3 4 4 ...
 $ t3      : num  5 4 4 4 3 4 3 3 4 4 ...
 $ t4      : num  3 2 4 3 1 5 3 2 4 2 ...
 $ t5      : num  4 3 4 3 5 4 4 2 4 3 ...
 $ t6      : num  3 4 4 3 1 5 5 2 3 4 ...
 $ t7      : num  5 3 4 2 2 4 4 2 4 3 ...
 $ t8      : num  5 3 4 3 5 5 4 2 4 4 ...
 $ t9      : num  4 4 4 3 1 4 4 4 4 4 ...
 $ t10     : num  2 1 4 4 1 5 1 1 3 1 ...
 $ m1      : num  3 3 4 2 2 4 1 1 4 1 ...
 $ m2      : num  4 3 4 3 1 4 2 1 5 2 ...
 $ st1     : num  2 3 0 2 2 1 3 4 2 2 ...
 $ st2     : num  1 2 0 1 2 0 2 4 2 3 ...
 $ st3     : num  2 1 0 1 2 1 3 4 2 2 ...
 $ st4     : num  3 3 0 2 4 0 1 4 2 2 ...
 $ st5     : num  3 3 0 2 3 1 1 2 2 3 ...
 $ st6     : num  0 1 0 2 1 1 2 4 2 3 ...
 $ st7     : num  4 2 0 3 1 1 2 4 2 2 ...
 $ st8     : num  2 2 0 3 4 0 2 3 2 2 ...
 $ st9     : num  2 2 0 1 0 1 2 4 2 2 ...
 $ st10    : num  1 1 0 1 0 0 2 4 2 2 ...
>

```

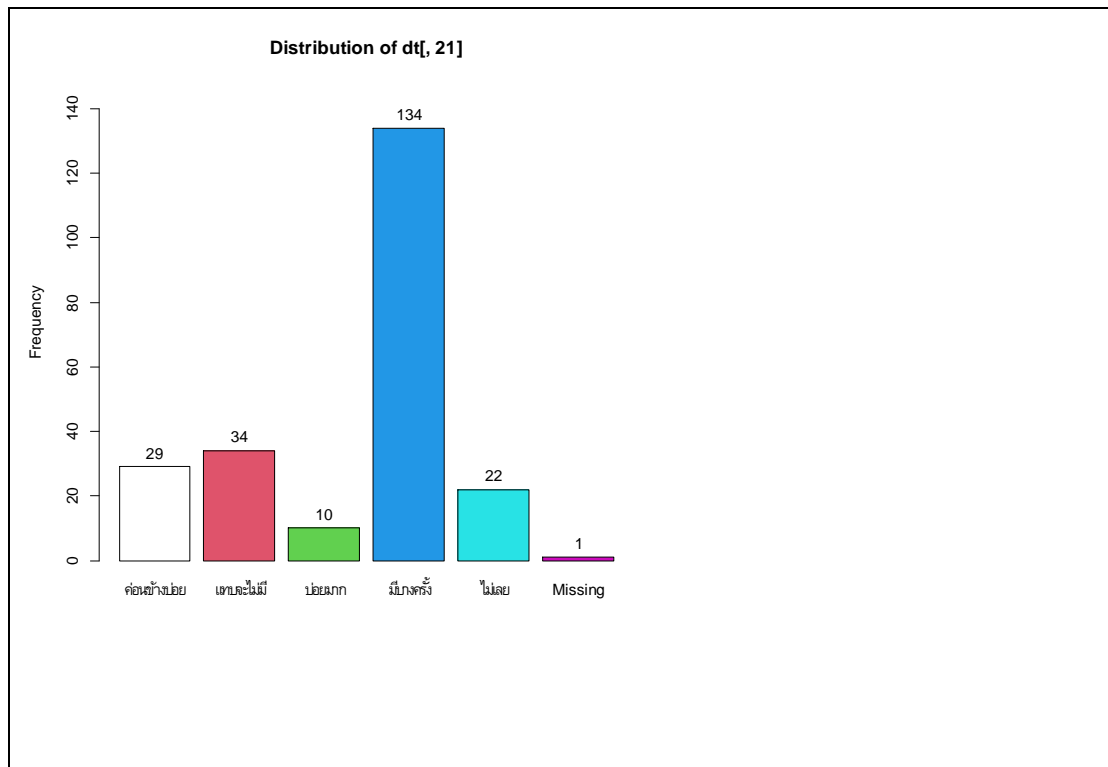
```

dt$inc <- factor(dt$inc)
tabl(dt$inc)

des(dt)

for(i in c(21:30)){
  dt[,i] <- ifelse(dt[,i]=="ไม่เลย",0,
    ifelse(dt[,i]=="แทบจะไม่",1,
      ifelse(dt[,i]=="มีบ้างครั้ง",2,
        ifelse(dt[,i]=="ค่อนข้างบ่อย",3,
          ifelse(dt[,i]=="บ่อยมาก",4,dt[,i]
        ))))
  dt[,i] <- as.numeric(dt[,i])
}
str(dt)

```



- n) ให้คำนวณค่าเฉลี่ยของระดับคะแนนของกลุ่มตัวอย่างแต่ละคน (row average) ดังนี้
- ค่าเฉลี่ยของระดับคะแนนตัวแปร t1-t10 ตั้งชื่อเป็น t_average
- ค่าเฉลี่ยของระดับคะแนนตัวแปร m1-m3 ตั้งชื่อเป็น m_average
- ค่าเฉลี่ยของระดับคะแนนตัวแปร st1-st10 ตั้งชื่อเป็น st_average
- และให้แสดงผลพื้ในรูปแบบตารางสำหรับตัวแปร t_average, m_average และ st_average

คำตอบ (แสดงคำสั่งและผลลัพธ์)

```

m1      m2      st1      st2
Min.    1.000    Min.    1.000    Min.    0.000    Min.    0.000
1st Qu.:2.000    1st Qu.:2.000    1st Qu.:2.000    1st Qu.:1.000
Median :3.000    Median :3.000    Median :2.000    Median :2.000
Mean   :2.935    Mean   :2.978    Mean   :1.873    Mean   :1.672
3rd Qu.:4.000    3rd Qu.:4.000    3rd Qu.:2.000    3rd Qu.:2.000
Max.   :5.000    Max.   :5.000    Max.   :4.000    Max.   :4.000
NA's    :1        NA's    :1        NA's    :1

st3      st4      st5      st6
Min.     0.00    Min.     0.00    Min.     0.00    Min.     0.00
1st Qu.:1.00    1st Qu.:1.00    1st Qu.:1.00    1st Qu.:1.00
Median :2.00    Median :2.00    Median :2.00    Median :2.00
Mean   :1.86    Mean   :1.974    Mean   :1.86    Mean   :1.572
3rd Qu.:2.00    3rd Qu.:2.00    3rd Qu.:2.00    3rd Qu.:2.00
Max.   :4.00    Max.   :4.00    Max.   :4.00    Max.   :4.00
NA's    :1        NA's    :1        NA's    :1

st7      st8      st9      st10
Min.     0.000    Min.     0.000    Min.     0.000    Min.     0.000
1st Qu.:1.000    1st Qu.:1.000    1st Qu.:1.000    1st Qu.:1.000
Median :2.000    Median :2.000    Median :2.000    Median :2.000
Mean   :1.943    Mean   :1.934    Mean   :1.511    Mean   :1.515
3rd Qu.:2.000    3rd Qu.:2.000    3rd Qu.:2.000    3rd Qu.:2.000
Max.   :4.000    Max.   :4.000    Max.   :4.000    Max.   :4.000
NA's    :1        NA's    :1        NA's    :1

tavg      mavg      stavg
Min.     2.300    Min.     0.000    Min.     0.000
1st Qu.:3.300    1st Qu.:1.500    1st Qu.:1.500
Median :3.700    Median :1.900    Median :1.900
Mean   :3.676    Mean   :1.772    Mean   :1.772
3rd Qu.:4.100    3rd Qu.:2.100    3rd Qu.:2.100
Max.   :5.000    Max.   :4.000    Max.   :4.000
NA's    :1        NA's    :1        NA's    :1
> |

```

```

        ifelse(dt[,i]=="มาก",4,
        ifelse(dt[,i]=="ปานกลาง",3,
        ifelse(dt[,i]=="น้อย",2,
        ifelse(dt[,i]=="น้อยที่สุด",1,dt[,i]
        ))))
    dt[,i] <- as.numeric(dt[,i])
  }
  str(dt)
  tab1(dt[,21])
  des(dt)

  for(i in c(21:30)){
    dt[,i] <- ifelse(dt[,i]=="ไม่เลย",0,
      ifelse(dt[,i]=="แทบจะไม่",1,
        ifelse(dt[,i]=="มีบางส่วน",2,
          ifelse(dt[,i]=="ค่อนข้างน้อย",3,
            ifelse(dt[,i]=="น้อยมาก",4,dt[,i]
          ))))
    dt[,i] <- as.numeric(dt[,i])
  }
  str(dt)
  dt$stavg <- rowMeans(dt[,c(9:18)],na.rm=T)
  str(dt)
  dt$mavg <- rowMeans(dt[,c(19:20)],na.rm=T)
  str(dt)
  dt$stavg <- rowMeans(dt[,c(21:30)],na.rm=T)
  str(dt)
  summary(dt)

```

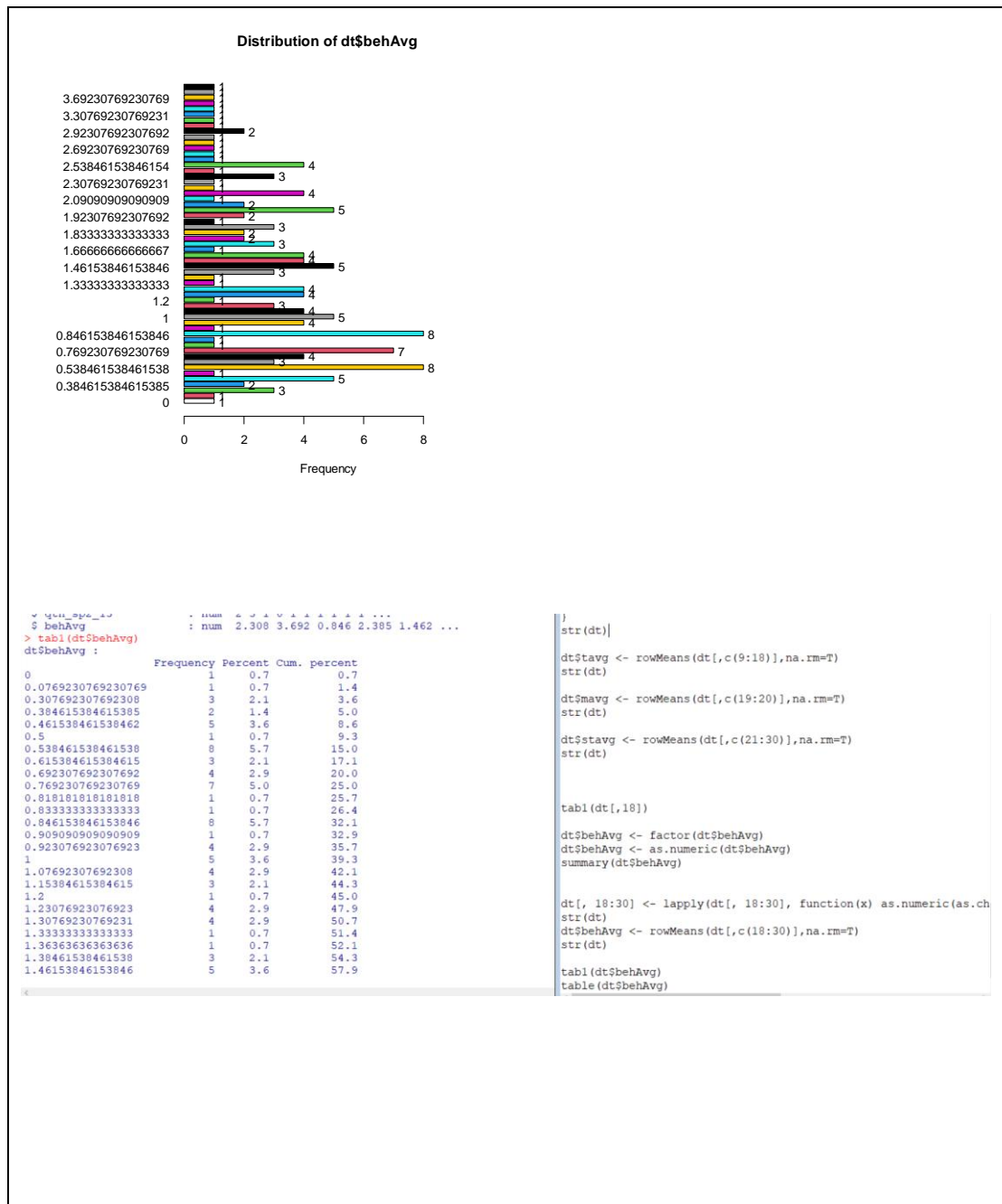


4. ให้นักศึกษาฝึกสำรวจข้อมูล และจัดการข้อมูล “child140_practice.xlsx”

	C	T	U	V	W	X	Y	Z	AA	AB	AC	AD
1	stu_sex	qtn_sp2_3	qtn_sp2_4	qtn_sp2_5	qtn_sp2_6	qtn_sp2_7	qtn_sp2_8	qtn_sp2_9	qtn_sp2_10	qtn_sp2_11	qtn_sp2_12	qtn_sp2_13
2	1	3	2	3	3	2	1	2	2	2	2	2
3	1	4	4	4	4	3	4	4	4	4	3	3
4	1	1	1	1	0	0	0	1	0	1	2	1
5	2	4	4	3	2	1	1	4	4	2	1	0
6	1	4	3	0	1	1	0	1	1	0	1	1
7	1	1	0	0	0	0	0	1	0	0	1	1
8	1	0	1	1	1	0	0	0	0	0	1	1
9	1	0	1	2	2	1		1	1	1	3	1
10	1	0	0	3	3	2	0	3	3	1	2	1
11	1		0	2	1	1	0	1	1	1	1	1
12	2	0	0	1	1	0	0	1	1	1	3	1
13	1	0	1	2	1	0	0	0	1	0	0	0
14	1	0	0	0	0	1	0	0	0	0	0	0
15	2	0	0	2	1	1	0	0	0	2	0	1
16	1	1	2	2	2	2	0	1	0	0	1	1
17	2	1	2	3	1	1	0	2	2	0	0	1
18	2	4	3	2	2	3	1	4	3	2	2	1
19	2		4	2	4	2	0	4	0	2	2	2
20	2	3	3	4	1	3	2	2	3	1	4	0
21	1	0	1	3	0	2	0	1	0	0	1	2
22	1	1	3	1	1	1	1	1	1	1	1	1
23	1	4	2	4	0	0	0	2	2	3	0	2
24	2	2	2	1	0	1	1	1	1	1	2	2
25	2	1	2	1	0	0	1	2	2	0	2	2
26	2	2	0	0	0	0	0	0	0	1	1	1

- a) ให้คำนวณค่าเฉลี่ยของระดับคะแนนตัวแปร qtn_sp2_1 - qtn_sp2_13 ของกลุ่มตัวอย่างแต่ละคน (row average) เก็บในตัวแปรใหม่ชื่อ behAvg โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ

คำตอบ (แสดงคำสั่งและผลลัพธ์)

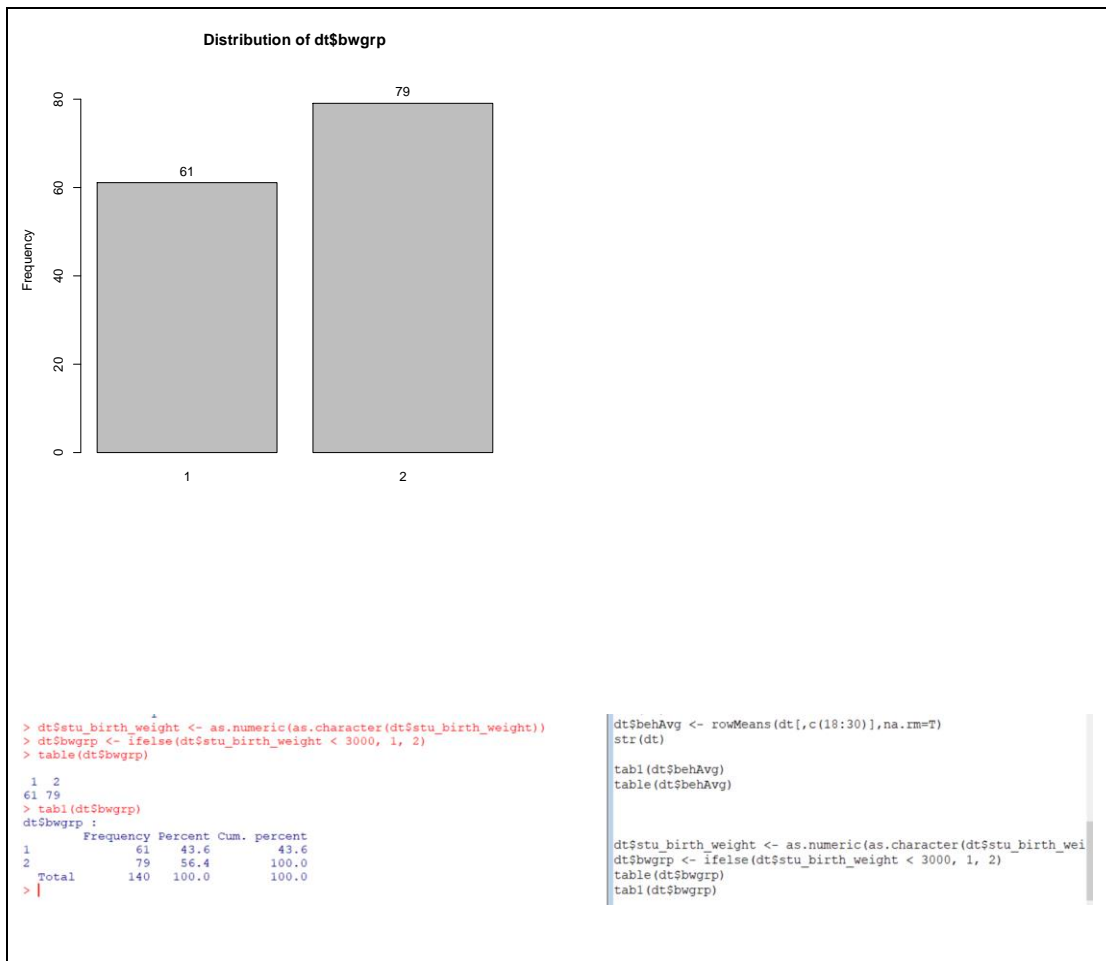


b) ให้จัดกลุ่มน้ำหนักแรกเกิดเป็น 2 กลุ่ม ดังแสดงข้างล่างนี้ เก็บในตัวแปรใหม่ชื่อ bwgrp โดยแสดงผลในรูปแบบตารางหรือกราฟ

1: <3,000 gram

2: >= 3,000 gram

คำตอบ (แสดงคำสั่งและผลลัพธ์)



c) ให้แสดงค่าเฉลี่ยของตัวแปร behAvg โดยจำแนกแต่ละกลุ่มของ bwgrp ให้อธิบายความหมาย

คำตอบ (แสดงคำสั่งและผลลัพธ์)

```

> avg<- aggregate(behAvg ~ bwgrp, data = dt, FUN = mean, na.rm = TRUE)
> str(avg)
'data.frame':  2 obs. of  2 variables:
 $ bwgrp : num  1  2
 $ behAvg: num  1.54 1.41
> avg
  bwgrp behAvg
1     1 1.535177
2     2 1.408737
> |

dt$bwgrp <- ifelse(dt$stu_birth_weight < 3000, 1, 2)
table(dt$bwgrp)
tab1(dt$bwgrp)

avg<- aggregate(behAvg ~ bwgrp, data = dt, FUN = mean, na.rm = TRUE)
avg

```

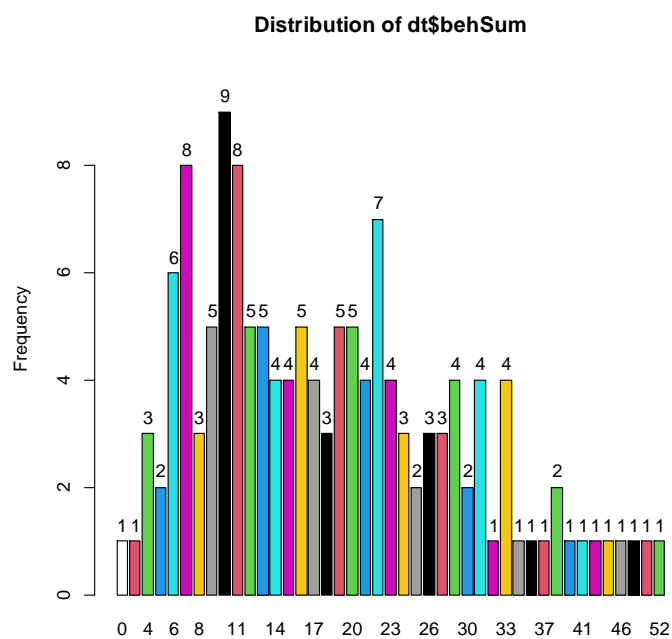
อธิบายความหมาย

ตัวแปร bwgrp และการจัดกลุ่มหาค่าเฉลี่ยตัวแปร behAvg

ผลลัพธ์ที่ได้สอดคล้องกับความเป็นจริง

- d) ให้คำนวณหาผลรวมของระดับคะแนนตัวแปร qtn_sp2_1 - qtn_sp2_13 ของกลุ่มตัวอย่างแต่ละคน เก็บในตัวแปรใหม่ชื่อ behSum โดยแสดงผลพื้ในรูปแบบตารางหรือกราฟ

คำตอบ (แสดงคำสั่งและผลลัพธ์)



```
> dt[, 18:30] <- lapply(dt[, 18:30], function(x) as.numeric(as.character(x)))
> dt$behSum <- rowSums(dt[, 18:30], na.rm = TRUE)
> summary(dt$behSum)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.00   10.00   17.00   18.73   25.00   52.00
> tab1(dt$behSum)
dt$behSum :
Frequency Percent Cum. percent
0          1     0.7           0.7
1          1     0.7           1.4
4          3     2.1           3.6
5          2     1.4           5.0
6          6     4.3           9.3
7          8     5.7          15.0
8          3     2.1          17.1
9          5     3.6          20.7
10         9     6.4          27.1
11         8     5.7          32.9
12         5     3.6          36.4
13         5     3.6          40.0
14         4     2.9          42.9
15         4     2.9          45.7
```

```
avg <- aggregate(behAvg ~ bwgrp, data = dt, FUN = mean, na.rm =
avg
```

```
dt[, 18:30] <- lapply(dt[, 18:30], function(x) as.numeric(as.ch
dt$behSum <- rowSums(dt[, 18:30], na.rm = TRUE)
summary(dt$behSum)
tab1(dt$behSum)
```

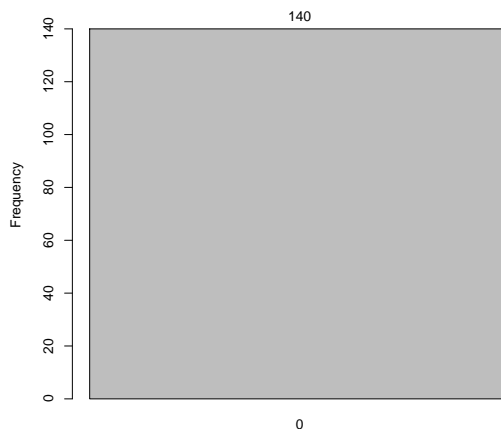

- e) ให้จัดกลุ่มข้อมูล behSum ดังแสดงข้างล่างนี้ ตั้งชื่อตัวแปรใหม่เป็น behgrp โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ

<3 → 0

>=3 → 1

คำตอบ (แสดงคำสั่งและผลลัพธ์)

Distribution of dt\$behgrp



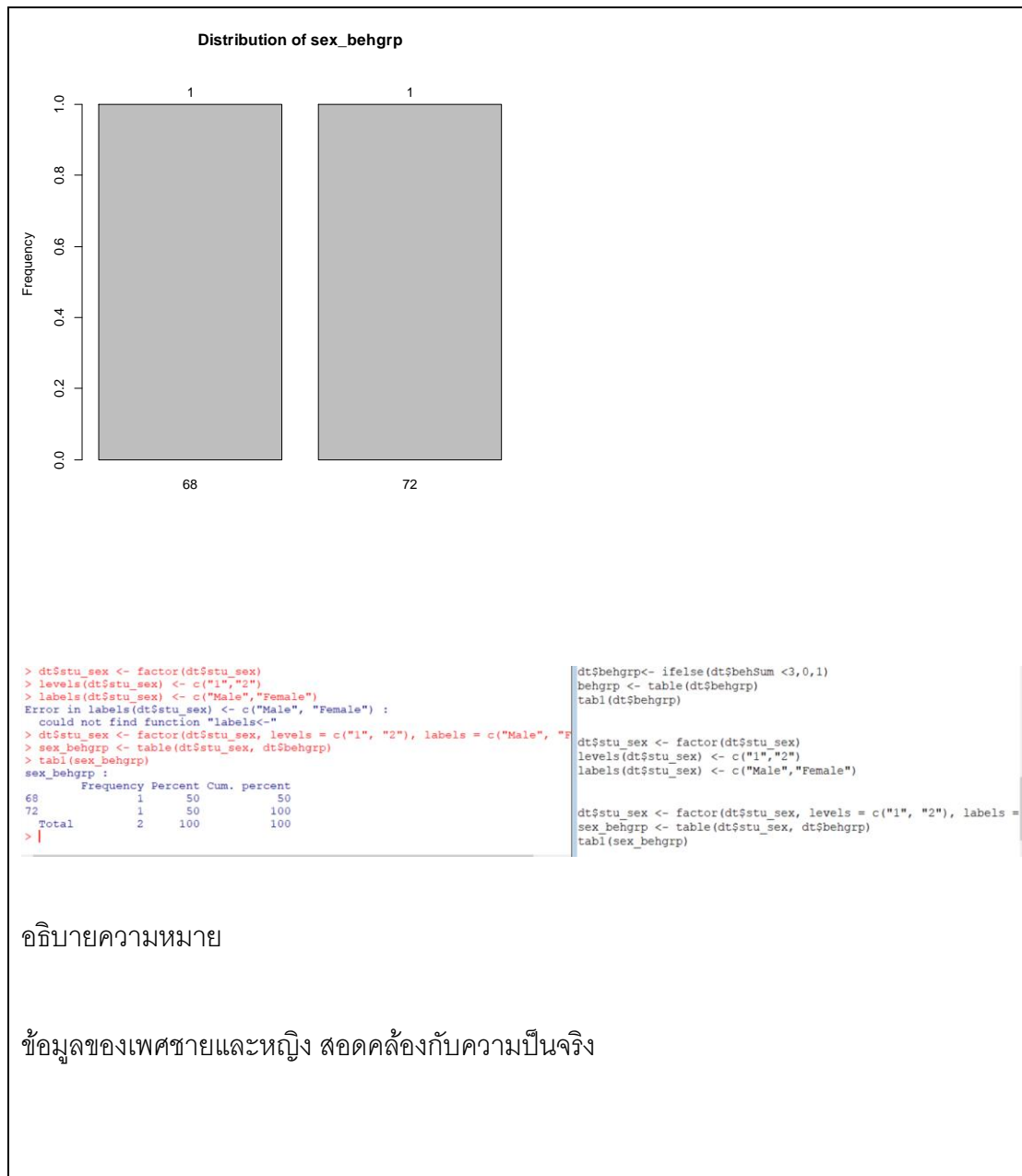
```
> dt$behSum <- ifelse(dt$behSum <3,0,1)
> behgrp <- table(dt$behSum)
> tab1(dt$behgrp)
Error in if (is.na(names(table.to.plot)[length(names(table.to.plot))]) | :
  argument is of length zero
> dt$behgrp
NULL
> dt$behgrp<- ifelse(dt$behSum <3,0,1)
> behgrp <- table(dt$behgrp)
> tab1(dt$behgrp)
dt$behgrp :
  Frequency Percent Cum. percent
0          140         100         100
Total          140         100         100
> |
```

```
dt[, 18:30] <- lapply(dt[, 18:30], function(x) as.numeric(as.ch
dt$behSum <- rowSums(dt[, 18:30], na.rm = TRUE)
summary(dt$behSum)
tab1(dt$behSum)
```

```
dt$behgrp<- ifelse(dt$behSum <3,0,1)
behgrp <- table(dt$behgrp)
tab1(dt$behgrp)
```

- f) ให้สำรวจข้อมูลโดยใช้ 2 ตัวแปร คือ เพศ และ behgrp โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ

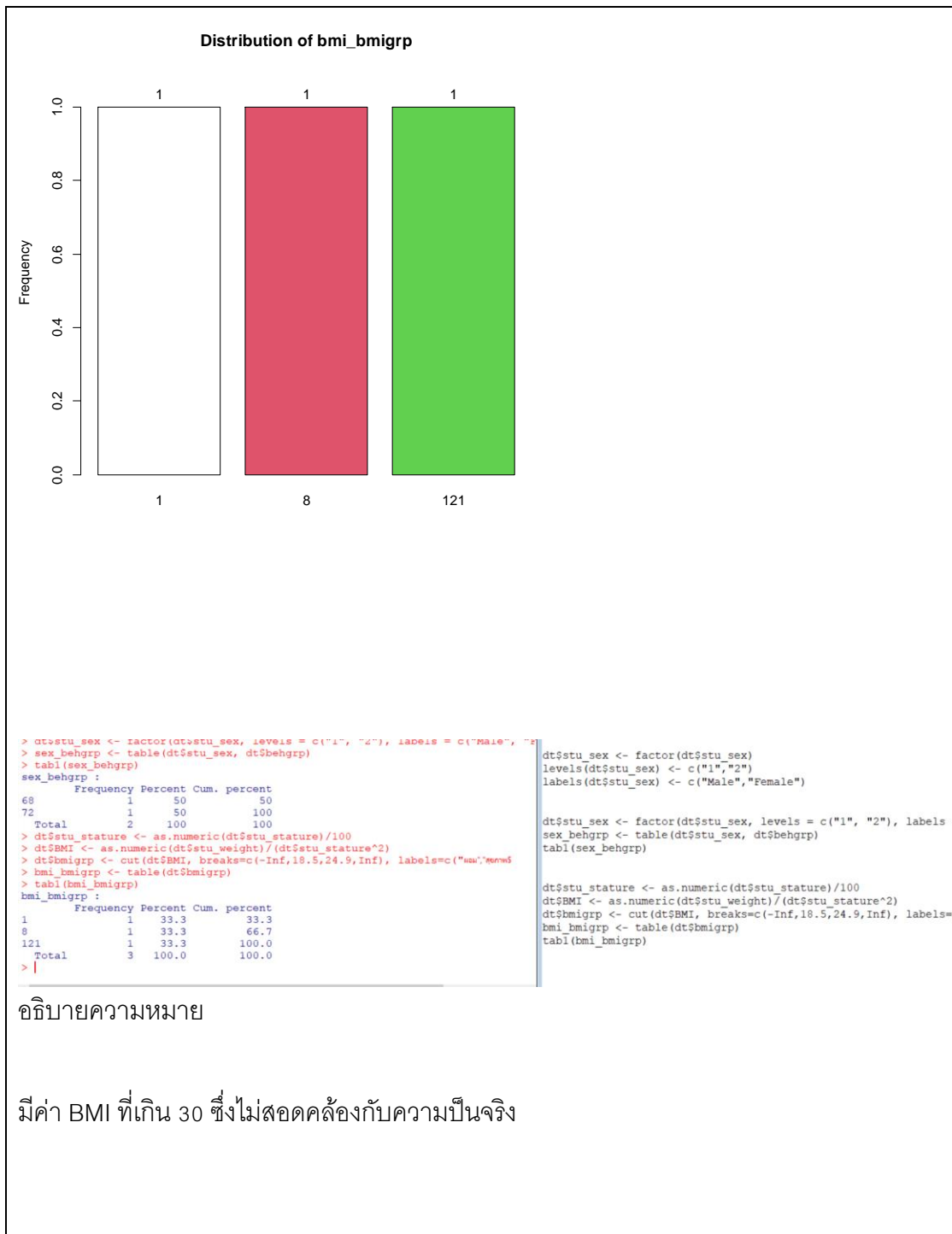
คำตอบ (แสดงคำสั่งและผลลัพธ์)



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

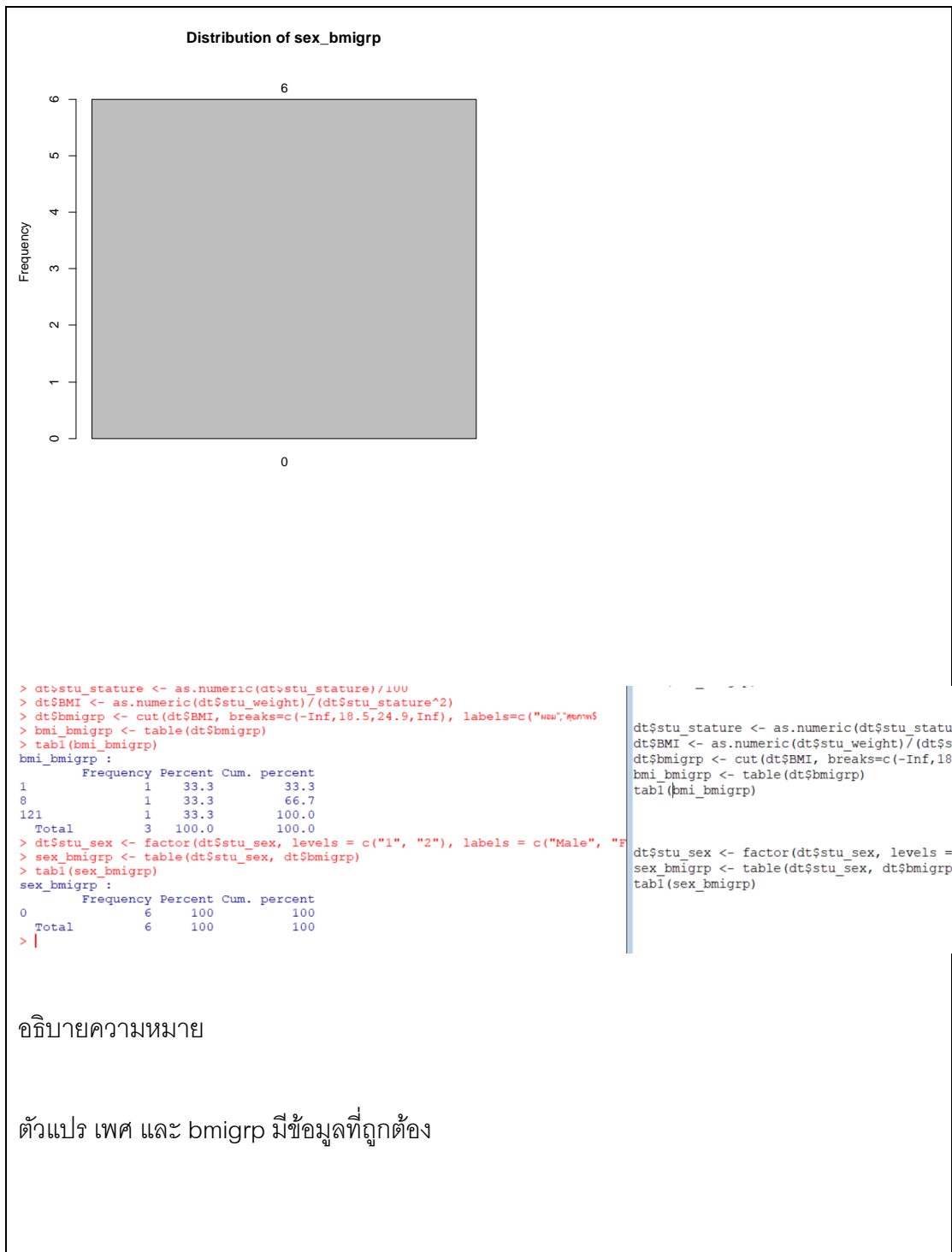
- 1: <18.5 ความหมาย ผอม
- 2: 18.5-24.9 ความหมาย สุขภาพดี
- 3: >24.9 ความหมาย น้ำหนักเกิน

คำตอบ (แสดงคำสั่งและผลลัพธ์)



- h) ให้สำรวจข้อมูลโดยใช้ 2 ตัวแปร คือ เพศ และ bmigrp โดยแสดงผลลัพธ์ในรูปแบบตารางหรือกราฟ

คำตอบ (แสดงคำสั่งและผลลัพธ์)



- i) ให้สำรวจข้อมูลโดยใช้ 2 ตัวแปร คือ bmigrp และ behgrp โดยแสดงผลในรูปแบบตารางหรือกราฟ

คำตอบ (แสดงคำสั่งและผลลัพธ์)

