

Classwork

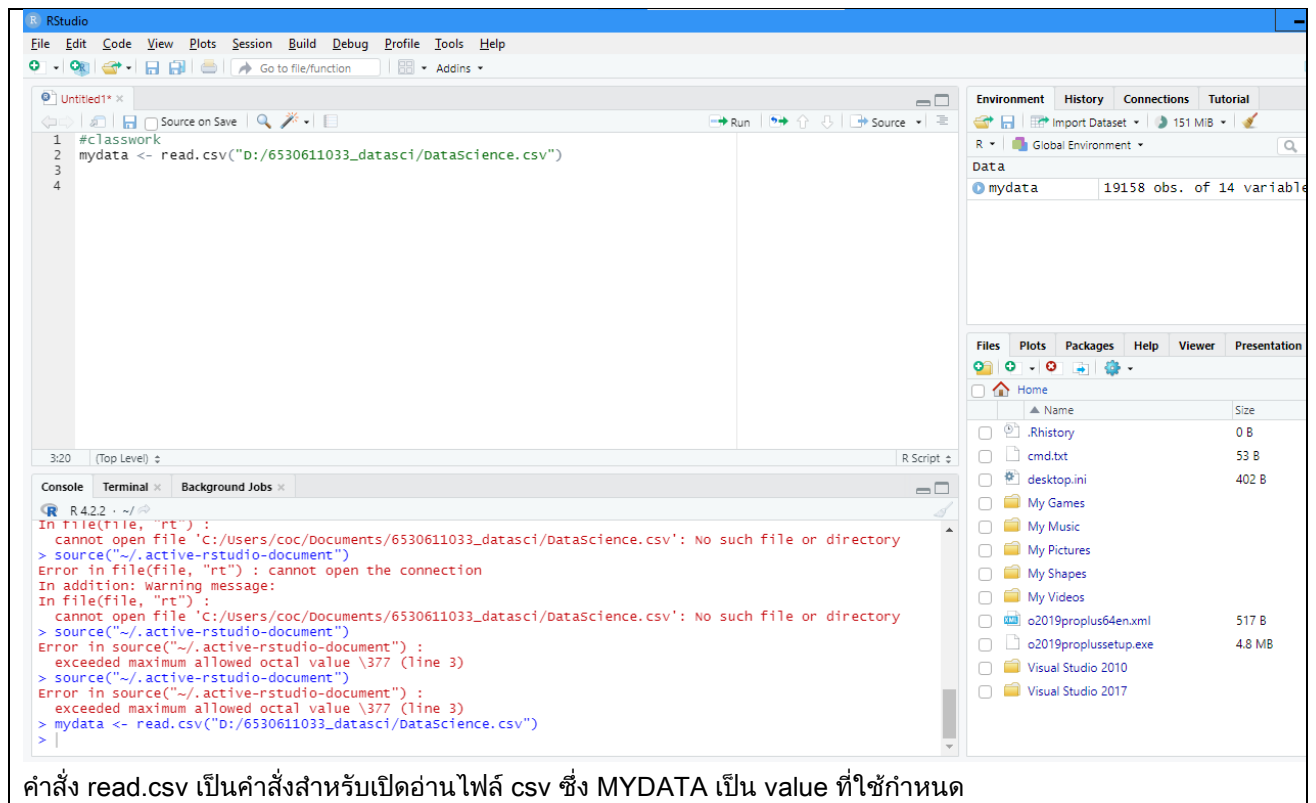
968-252 Data Science

Exploratory Data Analysis: General Information and Summary Statistics

ชื่อ-นามสกุล.....นาย.....ธนวัฒน์ วิริยธรรมโสภณ.....รหัส.....6530611033.....สาขา.....COMP.....

โหลดข้อมูลตัวอย่างไฟล์ชื่อ DataScience.csv จงดำเนินการคำสั่งต่อไปนี้ แล้วอธิบายว่าคำสั่งต่อไปนี้ใช้ทำอะไร พร้อมแสดงตัวอย่างผลลัพธ์ของจอที่ได้

- 1) `mydata <- read.csv("~/Desktop/Sample/Data/DataScience.csv")` *สามารถปรับตำแหน่งของไฟล์ที่เก็บได้"



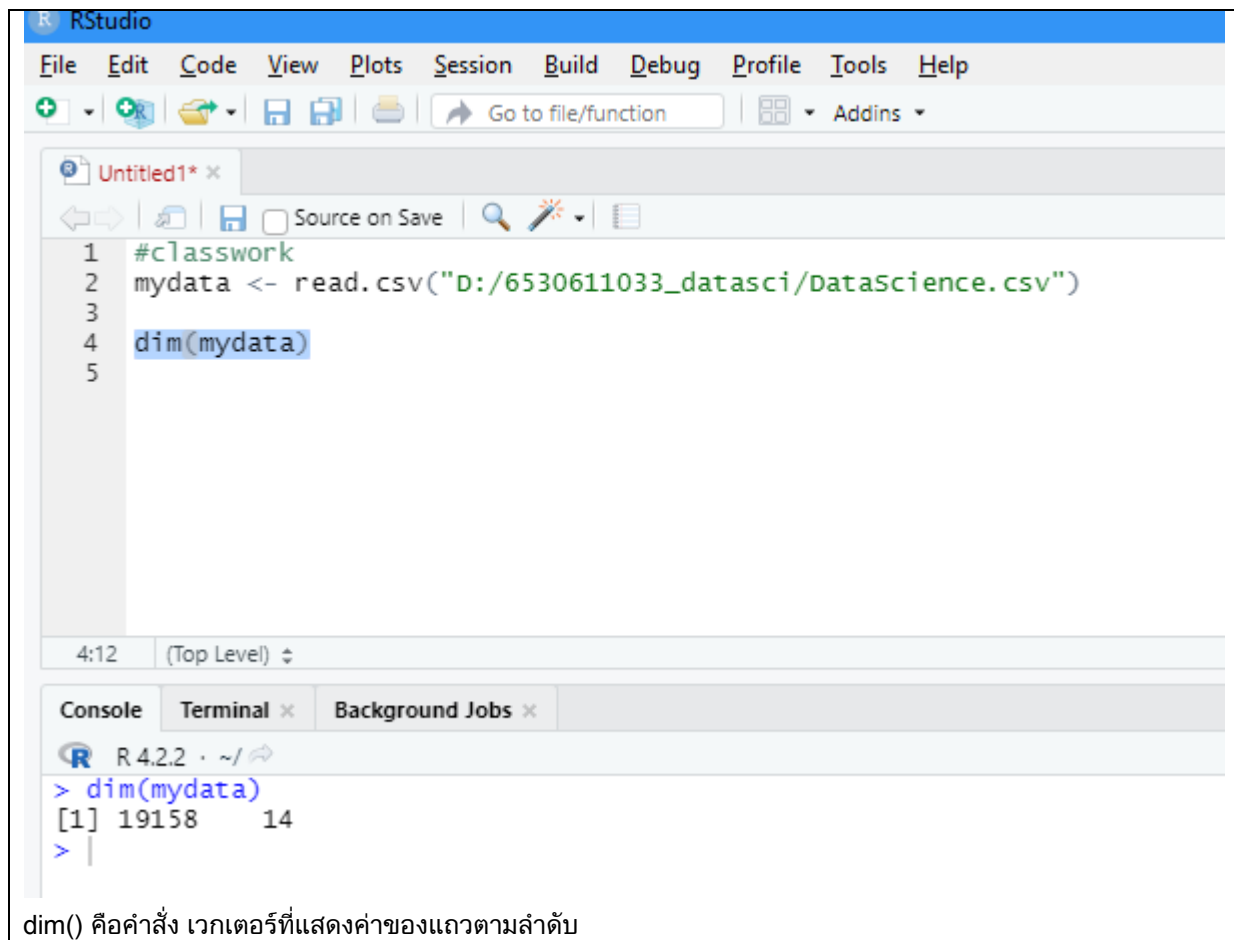
The screenshot shows the RStudio interface. The console displays the following error message:

```
R 4.2.2 ~\>  
In file(file, "rt") :  
cannot open file 'C:/users/coc/Documents/6530611033_datasci/DataScience.csv': No such file or directory  
> source("~/active-rstudio-document")  
Error in file(file, "rt") : cannot open the connection  
In addition: warning message:  
In file(file, "rt") :  
cannot open file 'C:/users/coc/Documents/6530611033_datasci/DataScience.csv': No such file or directory  
> source("~/active-rstudio-document")  
Error in source("~/active-rstudio-document") :  
exceeded maximum allowed octal value \377 (line 3)  
> source("~/active-rstudio-document")  
Error in source("~/active-rstudio-document") :  
exceeded maximum allowed octal value \377 (line 3)  
> mydata <- read.csv("D:/6530611033_datasci/DataScience.csv")  
>
```

The Environment pane shows the variable 'mydata' with 19158 observations and 14 variables. The Files pane shows the directory structure of the user's home folder.

คำสั่ง `read.csv` เป็นคำสั่งสำหรับเปิดอ่านไฟล์ csv ซึ่ง MYDATA เป็น value ที่ใช้กำหนด

2) `dim(mydata)`



The screenshot shows the RStudio interface. The source editor contains the following R code:

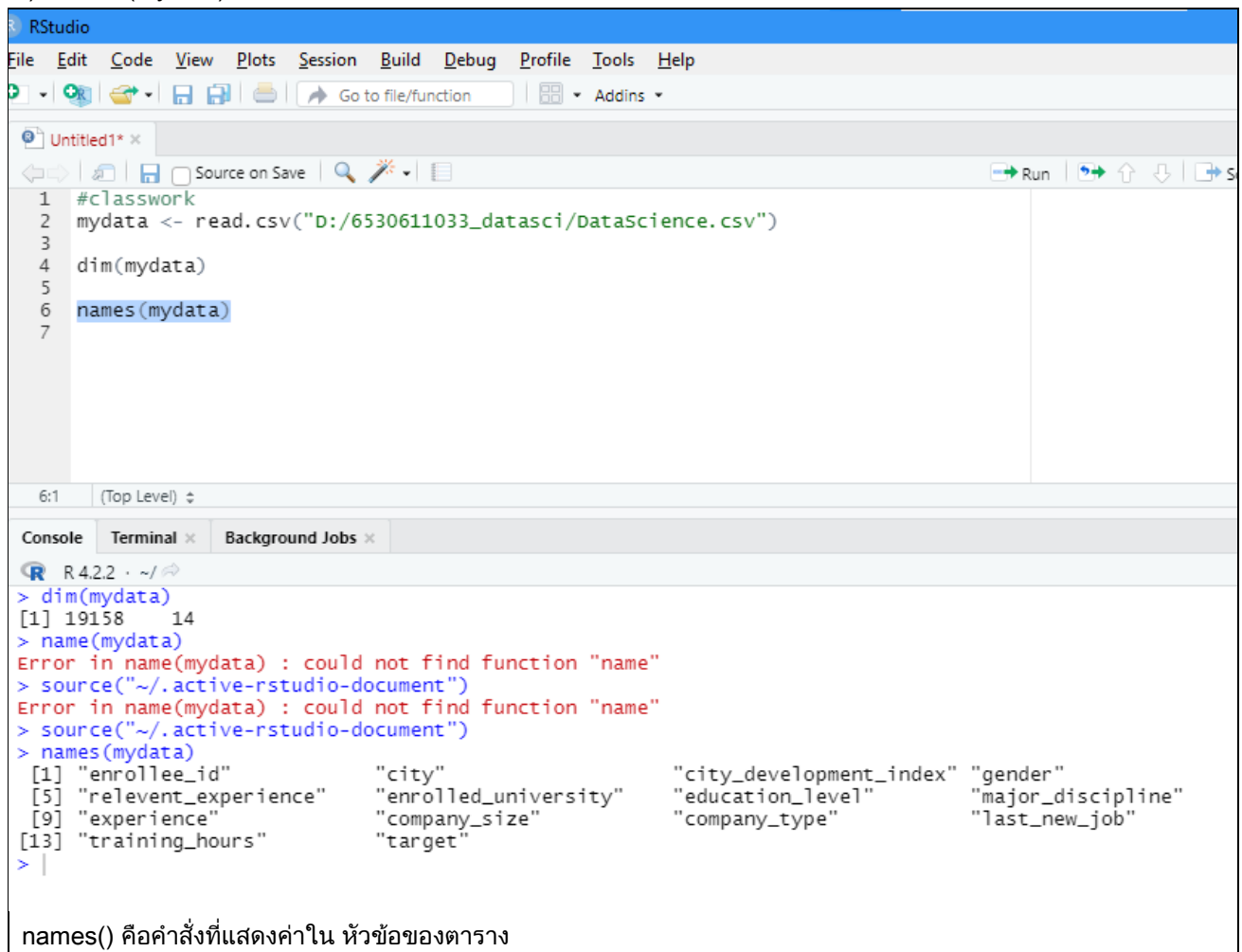
```
1 #classwork
2 mydata <- read.csv("D:/6530611033_datasci/DataScience.csv")
3
4 dim(mydata)
5
```

The console shows the output of the `dim(mydata)` command:

```
> dim(mydata)
[1] 19158 14
> |
```

Below the console, the text `dim()` คือคำสั่ง เวกเตอร์ที่แสดงค่าของแถวตามลำดับ

3). names(mydata)



The screenshot shows the RStudio interface. In the source editor, the following code is entered:

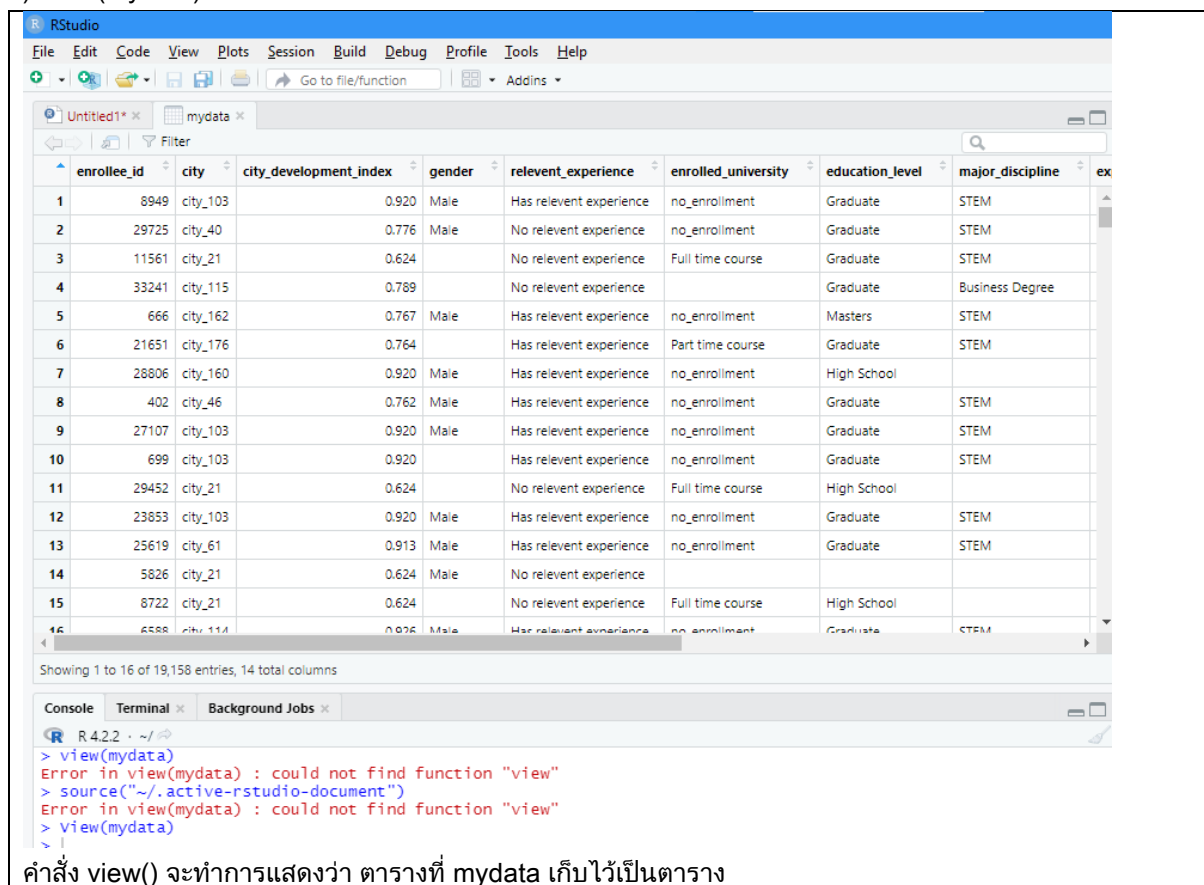
```
1 #classwork
2 mydata <- read.csv("D:/6530611033_datasci/DataScience.csv")
3
4 dim(mydata)
5
6 names(mydata)
7
```

The console output shows the result of the `names(mydata)` command:

```
> dim(mydata)
[1] 19158 14
> name(mydata)
Error in name(mydata) : could not find function "name"
> source("~/active-rstudio-document")
Error in name(mydata) : could not find function "name"
> source("~/active-rstudio-document")
> names(mydata)
 [1] "enrollee_id"      "city"              "city_development_index" "gender"
 [5] "relevent_experience" "enrolled_university" "education_level"        "major_discipline"
 [9] "experience"        "company_size"       "company_type"          "last_new_job"
[13] "training_hours"    "target"
>
```

names() คือคำสั่งที่แสดงค่าใน หัวข้อของตาราง

4) View(mydata)



The screenshot shows the RStudio interface. In the source editor, the following code is entered:

```
> view(mydata)
Error in view(mydata) : could not find function "view"
> source("~/active-rstudio-document")
Error in view(mydata) : could not find function "view"
> View(mydata)
>
```

The console output shows the result of the `View(mydata)` command:

```
> view(mydata)
Error in view(mydata) : could not find function "view"
> source("~/active-rstudio-document")
Error in view(mydata) : could not find function "view"
> View(mydata)
>
```

คำสั่ง view() จะทำการแสดงว่า ตารางที่ mydata เก็บไว้เป็นตาราง

5) summary(mydata)

```

10 summary(mydata)
11

```

10:1 (Top Level) ▾

Console Terminal × Background Jobs ×

R 4.2.2 · ~/

```

enrollee_id      city      city_development_index  gender      relevent_experience
Min.   :    1  Length:19158  Min.   :0.4480  Length:19158  Length:19158
1st Qu.: 8554  Class :character  1st Qu.:0.7400  Class :character  Class :character
Median :16983  Mode  :character  Median :0.9030  Mode  :character  Mode  :character
Mean   :16875                                Mean   :0.8288                                Mode  :character
3rd Qu.:25170                                3rd Qu.:0.9200                                Max.   :0.9490
Max.   :33380                                Max.   :0.9490

enrolled_university education_level  major_discipline  experience  company_size
Length:19158      Length:19158      Length:19158      Length:19158  Length:19158
Class :character   Class :character   Class :character   Class :character  Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character  Mode  :character

company_type      last_new_job      training_hours      target
Length:19158      Length:19158      Min.   : 1.00  Min.   :0.0000
Class :character   Class :character  1st Qu.: 23.00  1st Qu.:0.0000
Mode  :character   Mode  :character  Median : 47.00  Median :0.0000
                                Mean   : 65.37  Mean   :0.2493
                                3rd Qu.: 88.00  3rd Qu.:0.0000
                                Max.   :336.00  Max.   :1.0000

```

> |

summary() เป็นคำสั่งใช้สรุปค่าทั้งหมดให้เป็น max min median โดยดูจากแถวคอลัมและข้อมูลทั้งหมดสรุป

6) head(mydata)

```

> head(mydata)

```

Console Terminal × Background Jobs ×

R 4.2.2 · ~/

```

enrollee_id      city      city_development_index  gender      relevent_experience  enrolled_university
1      8949  city_103              0.920  Male  Has relevent experience  no_enrollment
2      29725  city_40              0.776  Male  No relevent experience  no_enrollment
3      11561  city_21              0.624              No relevent experience  Full time course
4      33241  city_115             0.789              No relevent experience
5         666  city_162             0.767  Male  Has relevent experience  no_enrollment
6      21651  city_176             0.764              Has relevent experience  Part time course

education_level  major_discipline  experience  company_size  company_type  last_new_job  training_hou
1      Graduate              STEM              >20              50-99      Pvt Ltd              >4
2      Graduate              STEM              15              50-99      Pvt Ltd              never
3      Graduate              STEM              5              50-99      Pvt Ltd              never
4      Graduate  Business Degree              <1              50-99      Pvt Ltd              never
5      Masters              STEM              >20              50-99      Funded Startup              4
6      Graduate              STEM              11              50-99      Funded Startup              1

target
1      1
2      0
3      0
4      1
5      0
6      1

```

head() คือคำสั่งใช้ แสดงแถวในตารางทั้งหมด

7) head(mydata, n=20)

14 head(mydata, n=20)

14:1 (Top Level) ⬇

Console Terminal x Background Job

R 4.2.2 · ~/

```
target
1      1
2      0
3      0
4      1
5      0
6      1
7      0
8      1
9      1
10     0
11     1
12     0
13     0
14     0
15     0
16     0
17     0
18     0
19     1
20     1
> |
```

head (x, n=20) ใช้สำหรับแสดงค่าข้อมูลในตารางของแต่ละหัวข้อให้แสดงข้อมูล 20 ตัว

8) tail(mydata)

15
16 tail(mydata)
17

16:1 (Top Level) ⬇

Console Terminal x Background Jobs x

R 4.2.2 · ~/

```
enrollee_id  city  city_development_index  gender  relevent_experience  enrolled_university
19153      29754  city_103                      0.920  Female  Has relevent experience  no_enrollment
19154      7386   city_173                      0.878  Male   No relevent experience  no_enrollment
19155      31398  city_103                      0.920  Male   Has relevent experience  no_enrollment
19156      24576  city_103                      0.920  Male   Has relevent experience  no_enrollment
19157       5756  city_65                       0.802  Male   Has relevent experience  no_enrollment
19158      23834  city_67                       0.855           No relevent experience  no_enrollment
education_level  major_discipline  experience  company_size  company_type  last_new_job  training_hours
19153      Graduate      Humanities      7          10/49  Funded Startup      1
19154      Graduate      Humanities     14
19155      Graduate      STEM          14
19156      Graduate      STEM          >20       50-99      Pvt Ltd      4
19157      High School
19158      Primary School      2          500-999      Pvt Ltd      2      1
target
19153      0
19154      1
19155      1
19156      0
19157      0
19158      0
> |
```

tail() ใช้สำหรับแสดงข้อมูลท้ายตาราง

9) tail(mydata, n = 20)

ใช้สำหรับแสดงข้อมูลท้ายตาราง 20 ตัว

```

15
16 fail(mydata)
17
16:1 (Top Level) R Script

```

enrollee_id	city	city_development_index	gender	relevant_experience	enrolled_university
19153	29754 city_103	0.920	Female	Has relevant experience	no_enrollment
19154	7386 city_173	0.878	Male	No relevant experience	no_enrollment
19155	31398 city_103	0.920	Male	Has relevant experience	no_enrollment
19156	24576 city_103	0.920	Male	Has relevant experience	no_enrollment
19157	5756 city_65	0.802	Male	Has relevant experience	no_enrollment
19158	23834 city_67	0.855		No relevant experience	no_enrollment

education_level	major_discipline	experience	company_size	company_type	last_new_job	training_hours
19153	Graduate	Humanities	7	10/49 Funded Startup	1	25
19154	Graduate	Humanities	14		1	42
19155	Graduate	STEM	14		4	52
19156	Graduate	STEM	>20	50-99 Pvt Ltd	4	44
19157	High School		<1	500-999 Pvt Ltd	2	97
19158	Primary School		2		1	127

target	
19153	0
19154	1
19155	1
19156	0
19157	0
19158	0

```

> |

```

10) mydata[1:10,]

```

> mydata[1:10, ]
  enrollee_id    city city_development_index gender  relevant_experience enrolled_universit
1      8949 city_103          0.920   Male Has relevant experience      no_enrollmer
2     29725 city_40          0.776   Male  No relevant experience      no_enrollmer
3     11561 city_21          0.624           No relevant experience      Full time cours
4     33241 city_115         0.789           No relevant experience
5       666 city_162          0.767   Male Has relevant experience      no_enrollmer
6     21651 city_176          0.764           Has relevant experience      Part time cours
7     28806 city_160          0.920   Male Has relevant experience      no_enrollmer
8       402 city_46          0.762   Male Has relevant experience      no_enrollmer
9     27107 city_103          0.920   Male Has relevant experience      no_enrollmer
10      699 city_103          0.920           Has relevant experience      no_enrollmer

```

education_level	major_discipline	experience	company_size	company_type	last_new_job	training
1	Graduate	STEM	>20			1
2	Graduate	STEM	15	50-99 Pvt Ltd		>4
3	Graduate	STEM	5			never
4	Graduate	Business Degree	<1		Pvt Ltd	never
5	Masters	STEM	>20	50-99 Funded Startup		4
6	Graduate	STEM	11			1
7	High School		5	50-99 Funded Startup		1
8	Graduate	STEM	13	<10 Pvt Ltd		>4
9	Graduate	STEM	7	50-99 Pvt Ltd		1

mydata[1:10,] แสดงข้อมูลตัวที่ 1 - 10

11) mydata[1:10, 1:3]

```

Console Terminal Background Jobs
R 4.2.2 ~
> mydata[1:10, 1:3]
  enrollee_id    city city_development_index
1      8949 city_103          0.920
2     29725 city_40          0.776
3     11561 city_21          0.624
4     33241 city_115         0.789
5       666 city_162          0.767
6     21651 city_176          0.764
7     28806 city_160          0.920
8       402 city_46          0.762
9     27107 city_103          0.920
10      699 city_103          0.920
> |

```

mydata[1:10, 1:3]
แสดงตัวที่ 1 – 10 พร้อมทั้งตารางที่ 1 และ 2 และ 3

12) `unique(mydata[c("gender")])`

```
10 1033 City_103
> unique(mydata[c("gender")])
gender
1   Male
3
20 Female
48 other
> |
```

`unique()` ใช้สำหรับกำหนดค่าที่มีความเป็นเอกลักษณ์ โดยที่ ภายใน [] คือตัวที่กำหนดหัวข้อว่าจะเลือก unique ตัวไหนโดยจะลบค่าที่มีซ้ำออกและหยิบออกมาเพียงตัวเดียว

13) `table(mydata$gender)`

```
23
24 table(mydata$gender)
24:1 (Top Level)
Console Terminal x Background Jobs x
R 4.2.2 · ~/
> source("~/active-rstudio-docum
> table(mydata$gender)

      Female   Male   other
4508    1238  13221    191
> |
```

`table()` ใช้สำหรับ แสดงรายการค่าทั้งหมดของตัวแปรพร้อมความถี่ของ gender

14) `genderEducation <- table(mydata$education_level, mydata$gender)`

mydata	19158 obs. of 14 variables
values	
genderEducation	'table' int [1:6, 1:4] 201 2569 522 1038 83 95 8 773 67 339 ...

กำหนดตัวแปร `gendereducation` ให้เก็บค่าของ `table()` (ที่มี `education_lvl` และ `gender`)

15) `addmargins(genderEducation)`

```
Console Terminal x Background Jobs x
R 4.2.2 · ~/
> genderEducation <- table(mydata$education_level, mydata$gender)
> addmargins(genderEducation)

      Female   Male   other   Sum
Graduate    201      8    242      9    460
High school 2569   773   8144   112 11598
Masters     522     67   1395     33   2017
Phd         1038   339   2957     27   4361
Primary school 83     47    280      4    414
Sum         4508  1238 13221    191 19158
> |
```

`addmargins()` คำสั่งใช้เพิ่มผลรวมให้กับตัวแปร `genderEducation` ที่ได้ค่าจากตัวเลขข้อก่อนหน้า

16) prop.table(genderEducation)

ConsoleTerminalBackground Jobs

R 4.2.2 · ~/ ↗

> prop.table(genderEducation)

		Female	Male	Other
		0.0104917006	0.0004175801	0.0126317987
Graduate		0.1340954171	0.0403486794	0.4250965654
High school		0.0272471030	0.0034972335	0.0728155340
Masters		0.0541810210	0.0176949577	0.1543480530
Phd		0.0043323938	0.0024532832	0.0146153043
Primary school		0.0049587640	0.0002087901	0.0105960956

> |

prop.table คือคำสั่งที่ใช้ ฟังก์ชันใน R สามารถใช้เพื่อคำนวณค่าของแต่ละเซลล์ในตารางเป็นสัดส่วนของค่าทั้งหมด

17) testdata <- xtabs(~gender+education_level+major_discipline, data = mydata)

Data	
mydata	19158 obs. of 14 variables
values	
genderEducation	'table' int [1:6, 1:4] 201 2569 522 1038 83 95 8 773 67 339 ...
testdata	'xtabs' int [1:4, 1:6, 1:7] 201 8 242 9 15 0 7 0 522 67 ...
<p>กำหนดตัวแปร testdata เพื่อใช้เก็บ คำสั่ง xtab ที่เป็นคำสั่งในการใช้สร้างตารางไขว้หรือตารางฉุกเฉิน 3 ตารางจาก gender education และ major</p>	

18) ftable(testdata)

		major_discipline	Arts	Business	Degree	Humanities	No Major	Other	STEM
gender	education_level								
Female	Graduate	201	0		0	0	0	0	0
	High School	15	39		50	62	35	62	2306
	Masters	522	0		0	0	0	0	0
	Phd	2	10		23	43	6	27	927
	Primary School	0	1		0	9	0	2	71
		95	0		0	0	0	0	0
Male	Graduate	8	0		0	0	0	0	0
	High School	0	25		16	70	14	28	620
	Masters	67	0		0	0	0	0	0
	Phd	0	10		9	42	2	8	268
	Primary School	0	1		0	6	0	0	40
		4	0		0	0	0	0	0
Other	Graduate	242	0		0	0	0	0	0
	High School	7	137		169	283	143	176	7229
	Masters	1395	0		0	0	0	0	0
	Phd	4	22		54	138	20	65	2654
	Primary School	0	1		3	8	0	6	262
		203	0		0	0	0	0	0
<p>ftable เป็นคำสั่ง ftable สร้างตารางฉุกเฉิน 'แบน' เช่นเดียวกับตารางฉุกเฉินทั่วไป ตารางเหล่านี้ประกอบด้วย การนับของแต่ละระดับของตัวแปร () ที่เกี่ยวข้อง</p>									

19) `install.packages("pastecs")`

`library(pastecs)`

`stat.desc(mydata)`

`Install.packages("pastecs")` ติดตั้ง packages

`library(pastecs)` ใช้งาน library()

```
35
36 install.packages("pastecs")
37 library(pastecs)
38 stat.desc(mydata)
39
```

38:1 (Top Level) ↕

Console Terminal × Background Jobs ×

R 4.2.2 · ~/

	training_hours	target
coef.var	NA	
nbr.val	1.915800e+04	1.915800e+04
nbr.null	0.000000e+00	1.438100e+04
nbr.na	0.000000e+00	0.000000e+00
min	1.000000e+00	0.000000e+00
max	3.360000e+02	1.000000e+00
range	3.350000e+02	1.000000e+00
sum	1.252299e+06	4.777000e+03
median	4.700000e+01	0.000000e+00
mean	6.536690e+01	2.493475e-01
SE.mean	4.339095e-01	3.125779e-03
CI.mean.0.95	8.505007e-01	6.126801e-03
var	3.607019e+03	1.871831e-01
std.dev	6.005846e+01	4.326466e-01
coef.var	9.187902e-01	1.735115e+00

`Stat.desc` เป็นคำสั่งที่ให้ค่าทาง static ทั้งหมดของ `mydata` โดยจาก desc มากไปน้อย

20) `stat.desc(mydata[,c("city_development_index","training_hours")])`

```
39
40 stat.desc(mydata[,c("city_development_index","training_hours")])
41
```

40:1 (Top Level) ↕

Console Terminal × Background Jobs ×

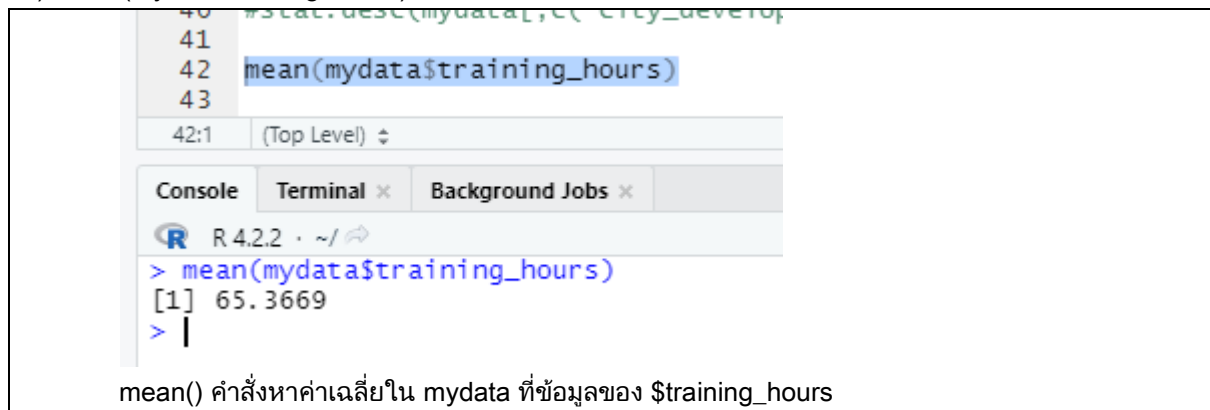
R 4.2.2 · ~/

```
> stat.desc(mydata[,c("city_development_index","training_hours")])
```

	city_development_index	training_hours
nbr.val	1.915800e+04	1.915800e+04
nbr.null	0.000000e+00	0.000000e+00
nbr.na	0.000000e+00	0.000000e+00
min	4.480000e-01	1.000000e+00
max	9.490000e-01	3.360000e+02
range	5.010000e-01	3.350000e+02
sum	1.587907e+04	1.252299e+06
median	9.030000e-01	4.700000e+01
mean	8.288480e-01	6.536690e+01
SE.mean	8.912621e-04	4.339095e-01
CI.mean.0.95	1.746952e-03	8.505007e-01
var	1.521812e-02	3.607019e+03
std.dev	1.233618e-01	6.005846e+01
coef.var	1.488352e-01	9.187902e-01

ให้ `stat.desc` คำนวณค่าของ "city_development_index","training_hours"

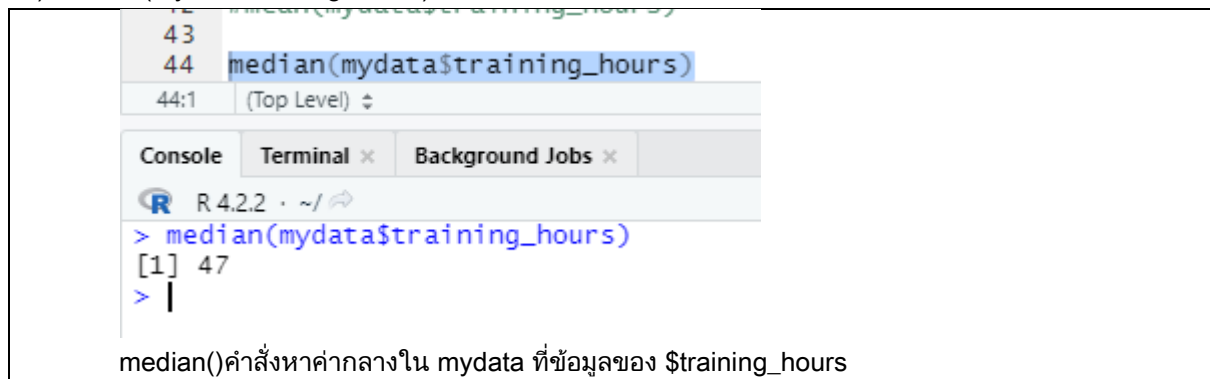
21) mean(mydata\$training_hours)



```
40 #stat.desc(mydata[,c("city_develop",
41
42 mean(mydata$training_hours)
43
42:1 (Top Level) ⚡
Console Terminal × Background Jobs ×
R 4.2.2 · ~/
> mean(mydata$training_hours)
[1] 65.3669
> |
```

mean() คำสั่งหาค่าเฉลี่ยใน mydata ที่ข้อมูลของ \$training_hours

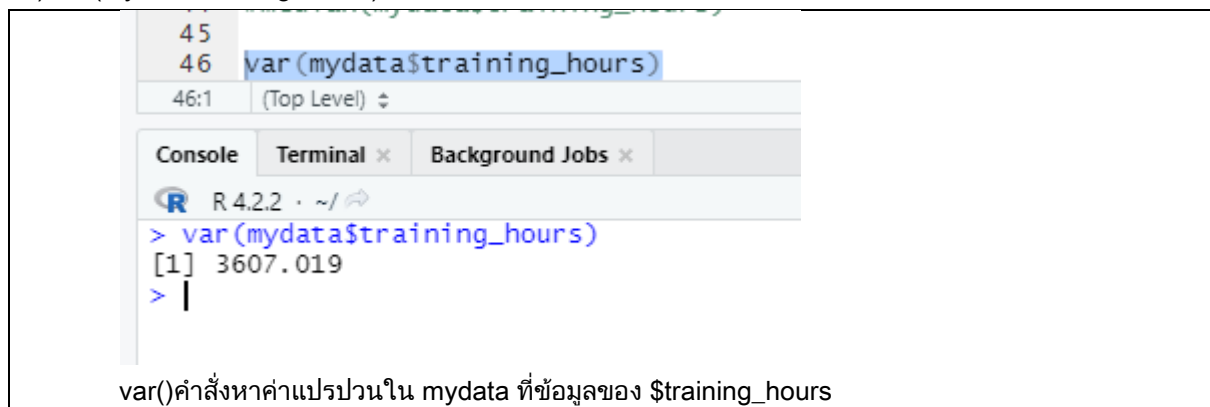
22) median(mydata\$training_hours)



```
43
44 median(mydata$training_hours)
44:1 (Top Level) ⚡
Console Terminal × Background Jobs ×
R 4.2.2 · ~/
> median(mydata$training_hours)
[1] 47
> |
```

median() คำสั่งหาค่ากลางใน mydata ที่ข้อมูลของ \$training_hours

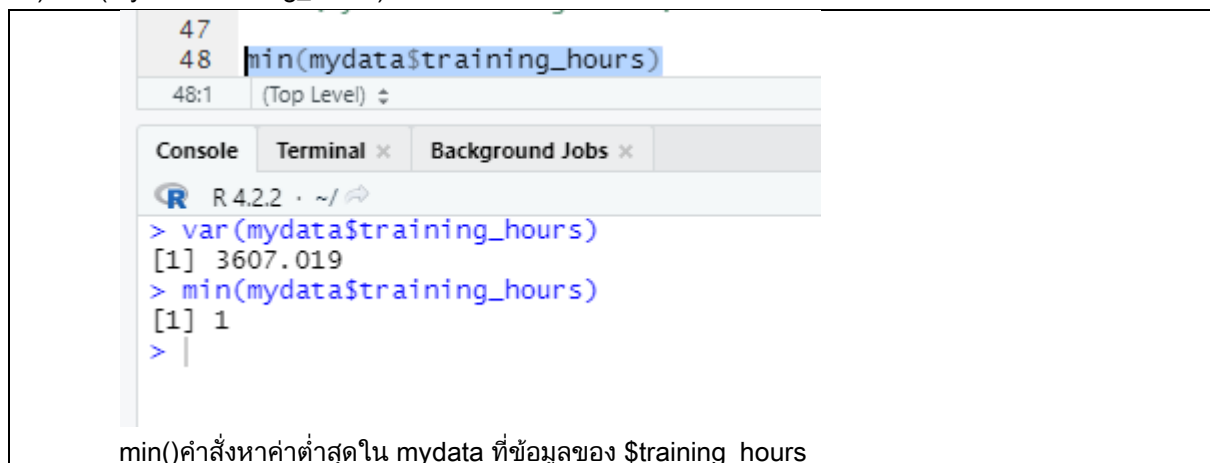
23) var(mydata\$training_hours)



```
45
46 var(mydata$training_hours)
46:1 (Top Level) ⚡
Console Terminal × Background Jobs ×
R 4.2.2 · ~/
> var(mydata$training_hours)
[1] 3607.019
> |
```

var() คำสั่งหาค่าแปรปรวนใน mydata ที่ข้อมูลของ \$training_hours


24) min(mydata\$training_hours)



```
47
48 min(mydata$training_hours)
48:1 (Top Level) ⚡
Console Terminal × Background Jobs ×
R 4.2.2 · ~/
> var(mydata$training_hours)
[1] 3607.019
> min(mydata$training_hours)
[1] 1
> |
```


min() คำสั่งหาค่าต่ำสุดใน mydata ที่ข้อมูลของ \$training_hours

25) `max(mydata$training_hours)`

```
49  
50 max(mydata$training_hours)  
50:1 (Top Level) ⚡  
Console Terminal × Background Jobs ×  
R 4.2.2 · ~/   
> max(mydata$training_hours)  
[1] 336  
> |
```


`max()` คำสั่งหาค่ามากสุดใน `mydata` ที่ข้อมูลของ `$training_hours`

26) `range(mydata$training_hours)`

```
51  
52 range(mydata$training_hours)  
52:1 (Top Level) ⚡  
Console Terminal × Background Jobs ×  
R 4.2.2 · ~/   
> range(mydata$training_hours)  
[1] 1 336  
> |
```


`range()` คำสั่งหาช่วงของข้อมูลใน `mydata` ที่ข้อมูลของ `$training_hours` คือ min and max

27) `quantile(mydata$training_hours)`

```
54 quantile(mydata$training_hours)  
55  
54:1 (Top Level) ⚡  
Console Terminal × Background Jobs ×  
R 4.2.2 · ~/   
> quantile(mydata$training_hours)  
0% 25% 50% 75% 100%  
1 23 47 88 336  
>
```

`quantile ()` คำสั่งหา quantile ใน `mydata` ที่ข้อมูลของ `$training_hours` ตั้งแต่ 1 ที่เป็น 0 จน 336 ที่เป็น 100 จุดตัดแบ่งพิสัยของการแจกแจงความน่าจะเป็นเป็นช่วงติดต่อกันที่มีโอกาสเท่า ๆ กัน

28) `max(table(mydata$gender))`

```
54 quantile(mydata$training_hours)  
55  
56 max(table(mydata$gender))  
57  
56:1 (Top Level) ⚡  
Console Terminal × Background Jobs ×  
R 4.2.2 · ~/   
> max(table(mydata$gender))  
[1] 13221  
>
```

`max(table(mydata$gender))`
ใช้สำหรับแสดงค่า max ตารางของ `mydata` ของหัวข้อ `gender`