# Data Process in R

Class 2

**Kai Yao**

**Apr, 2016**

# Outline

- Review of Last Class
- Homework 1 Explanation
- Read and Write Data files
- Brief Description Analysis
- Conclusions

# 作业提交说明

- 1. 访问github上面课程的文件夹https://github.com/jasonyaopku/Data-Processing-in-R.git，然后进入课程作业的目录Homeworks中下载对应的作业
- 2. 请大家将作业答案保存到word文件中另存为pdf，然后发送到邮箱 jasonyaopku@gmail.com
- 3提交作业的邮件标题和word文件名：DSJJYB_姓名_专业_HW*，务必按照这个方式，否则可能会遗漏大家的邮件造成减分。
- 4. 请每个人独立完成，可以相互交流，但不要在微信群里讨论，如果抄袭将记为0分。

# Outline

- Review of Last Class
- Homework 1 Explanation
- Read and Write Data files
- Brief Description Analysis
- Conclusions

# Review of Class 1

- Common Data Structures
  - NCL: (numeric, character, logical)
  - VDFM: (vector, factor, dataframe, matrix)
- Variable Definition
- Assignment, Indexing, Operations
- Conditional Execution and Loops

# Supplements

- Variable Definition
- Calculation Order
- Temporary Variable
- Vector Computation
- Multi Conditions
- Multi Loops

# Variable Definition

- Letters
- Letters+special symbols
- Letters+number
- Letters+[number, special symbols, letters]

- A and a are not the same

# Calculation Order

- Operators have different priority
- Same to the common calculation process in Math
- a+b/c does not equal to (a+b)/c
- a=1;b=2;c=3;
- a+b+c
- a*b+c
- a*(b+c)

# Temporary Variables

- The program will generate some implicit variables
- We should have a new mode of thinking
- a=1;b=2;c=3;
- d=a+b+c
- d=a+b/c

# Vector Computation

- Vectors can be computed with single number or vectors
- a=c(1,2,3,4,5,6);
- b=2;
- a/b?
- b/a?
- a*b?
- c=c(2,4)

# Multi Conditions

- AND (&)
  - TRUE, only when both of the two conditions are TRUE
- OR ( | )
  - TRUE, as long as one of the two conditions is TRUE

# Multi Loops

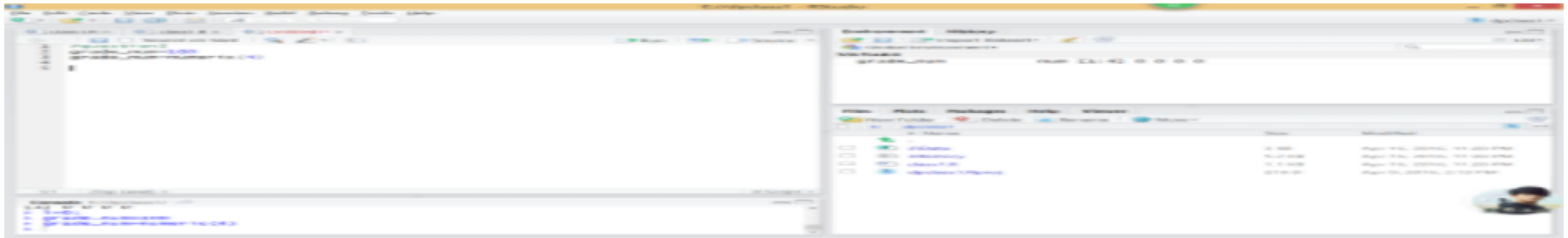
```
for(var1 in vector1)

{

    for(var2 in vector2)

    {

        ……

    }

}
```

# Dataframe

- names = c(“zhangsan”,”lisi”);
- ages = c(18,19);
- df.test = data.frame(names,ages);
- Element df.test$names does not equal variable names

# Outline

- Review of Last Class
- Homework 1 Explanation
- Read and Write Data files
- Brief Description Analysis
- Conclusions

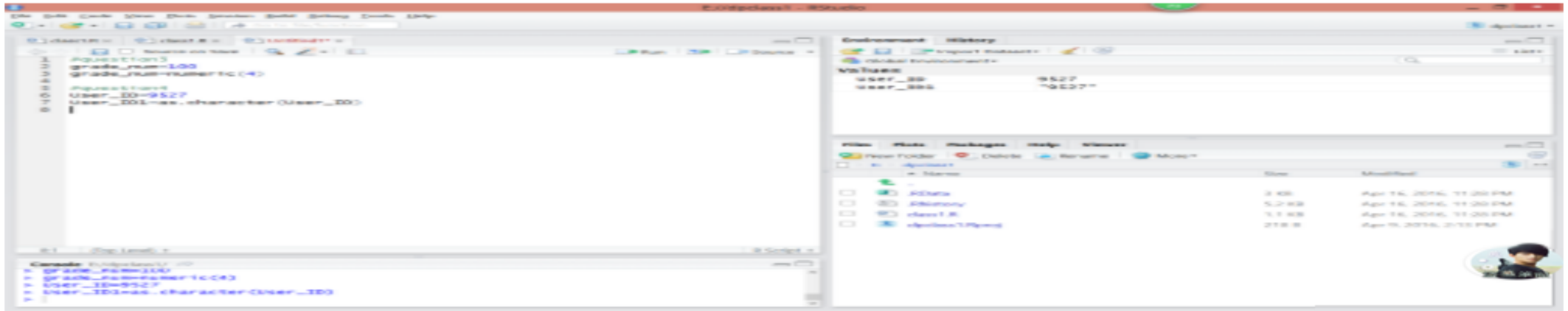


```
#question3
grade_num=100
grade_num=numeric(4)
```

```
> grade_num=100
> grade_num=numeric(4)
```

```
grade_num    num [1:4] 0 0 0 0
```

4、

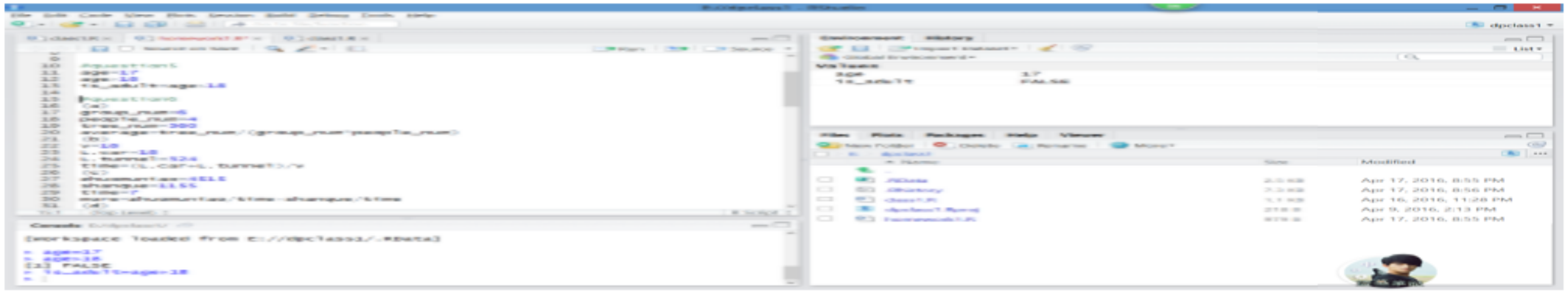


```
#question3
grade_num=100
grade_num=numeric(4)

#question4
user_ID=9527
user_ID1=as.character(user_ID)
```

```
> grade_num=100
> grade_num=numeric(4)
> user_ID=9527
> user_ID1=as.character(user_ID)
```

```
Values
  user_ID       9527
  user_ID1      "9527"
```

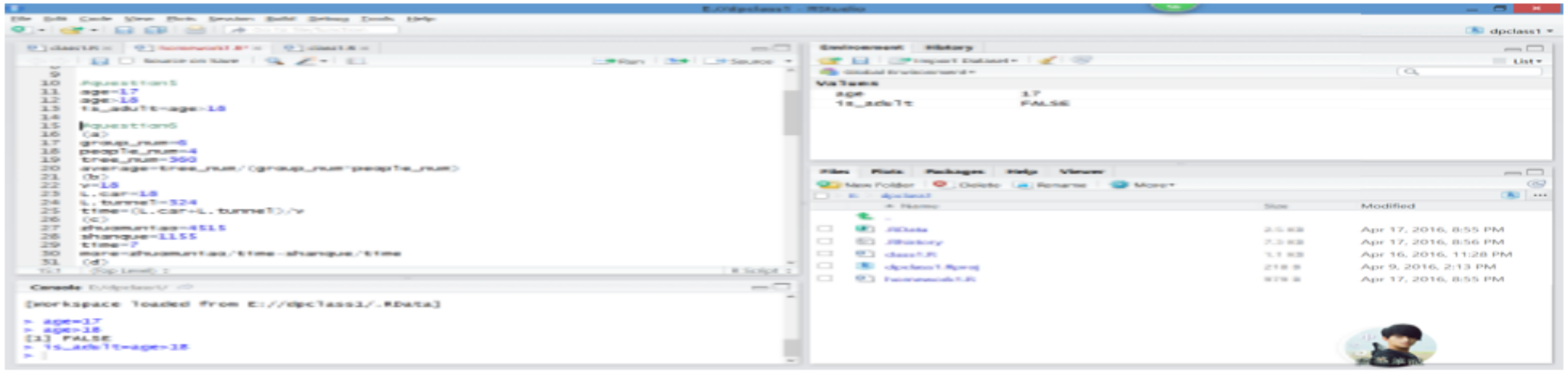
5、



```
#question5
age=17
age>18
is_adult=age>18

#question6
(a)
group_num=6
people_num=4
tree_num=360
average=tree_num/(group_num*people_num)
(b)
v=18
L.car=18
L.tunnel=524
time=(L.car+L.tunnel)/v
(c)
zhuzumuntao=4515
shanque=1155
time=7
more=zhuzumuntao/time-shanque/time
(d)
```

```
[Workspace loaded from E://dpclass1/.RData]

> age=17
> age>18
[1] FALSE
> is_adult=age>18
>
```

```
Values
  age          17
  is_adult     FALSE
```

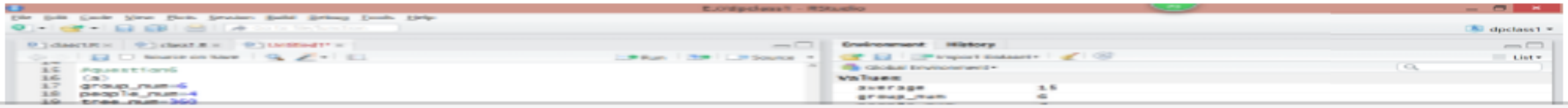


```
#question5
age=17
age>18
is_adult=age>18

#question6
(a)
group_num=6
people_num=4
tree_num=360
average=tree_num/(group_num*people_num)
(b)
v=18
L.car=18
L.tunnel=524
time=(L.car+L.tunnel)/v
(c)
zhuzumuntao=4515
shanque=1155
time=7
more=zhuzumuntao/time-shanque/time
(d)
```

```
[Workspace loaded from E://dpclass1/.RData]

> age=17
> age>18
[1] FALSE
> is_adult=age>18
>
```

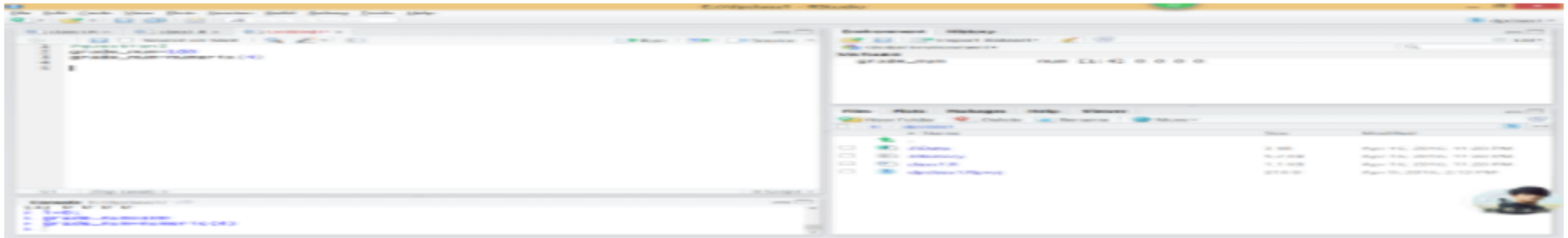
```
Values
  age          17
  is_adult     FALSE
```

6、(a)



```
#question6
(a)
group_num=6
people_num=4
tree_num=360
```

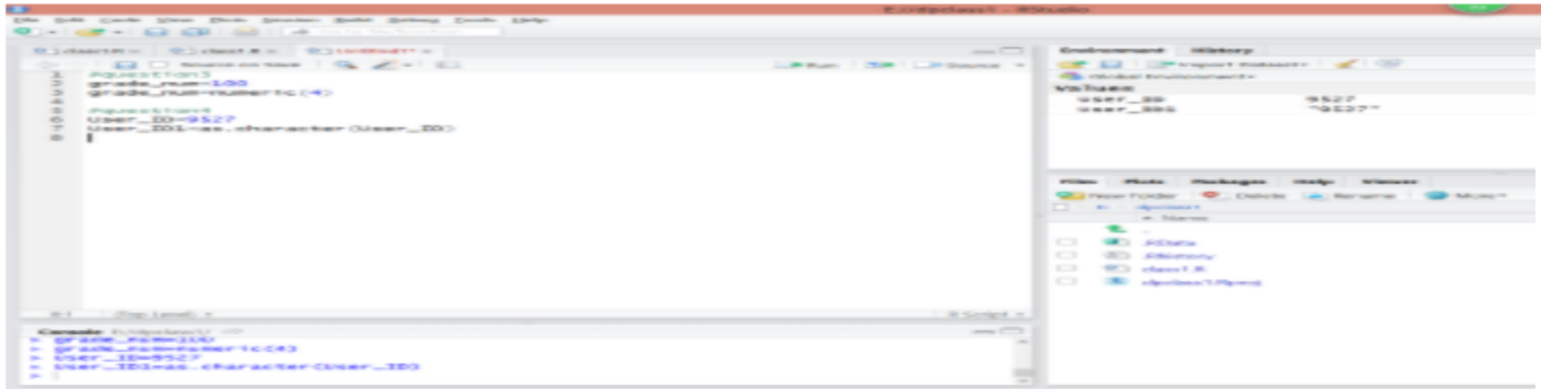
```
Values
  average      15
  group_num    6
```



```
#question3
grade_num=100
grade_num=numeric(4)
```

```
> grade_num=100
> grade_num=numeric(4)
>
```

4、

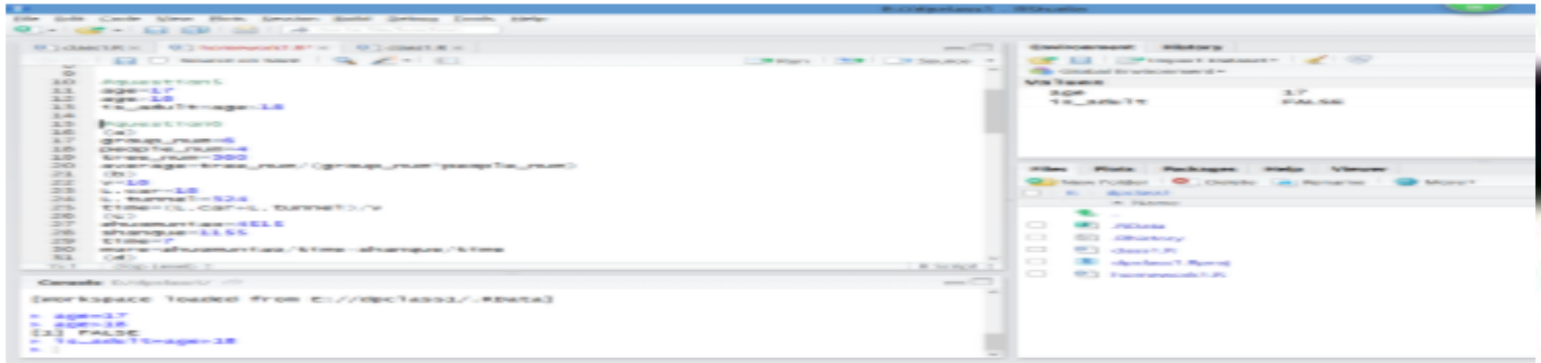


```
#question3
grade_num=100
grade_num=numeric(4)

#question4
User_ID=9527
User_ID1=as.character(User_ID)
```

```
> grade_num=100
> grade_num=numeric(4)
> User_ID=9527
> User_ID1=as.character(User_ID)
>
```

```
Values
  User_ID          9527
  User_ID1         "9527"
```

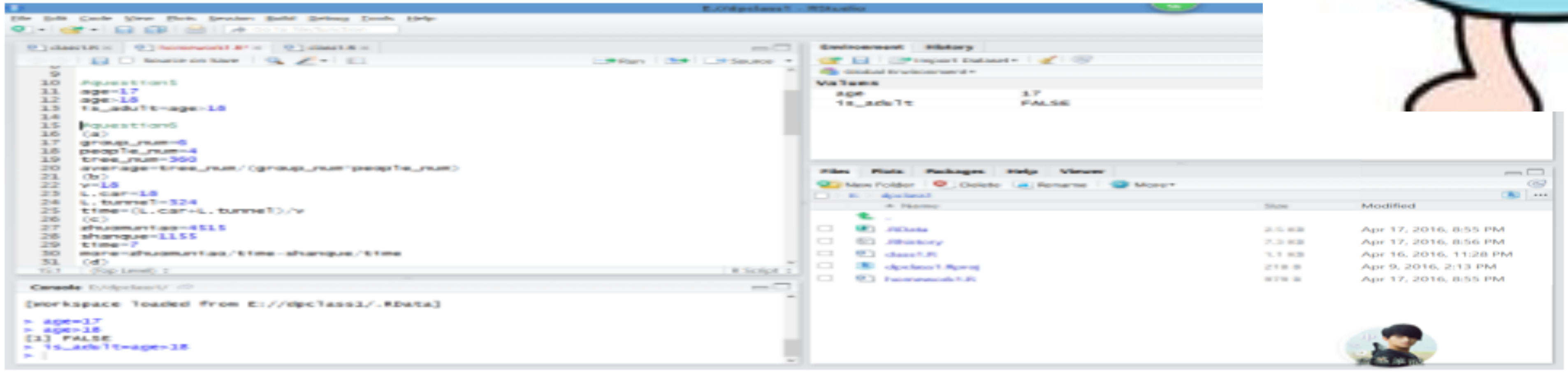
5、



```
#question5
age=17
age>18
is_adult=age>18

#question6
(a)
group_num=6
people_num=4
tree_num=360
average=tree_num/(group_num*people_num)
(b)
v=18
L.car=18
L.tunnel=524
time=(L.car+L.tunnel)/v
(c)
zhuamuntao=4515
shangue=1155
time=7
more=zhuamuntao/time-shangue/time
(d)
```

```
[Workspace loaded from E://dpclass1/.RData]

> age=17
> age>18
[1] FALSE
> is_adult=age>18
>
```

```
Values
  age              17
  is_adult         FALSE
```



```
#question5
age=17
age>18
is_adult=age>18

#question6
(a)
group_num=6
people_num=4
tree_num=360
average=tree_num/(group_num*people_num)
(b)
v=18
L.car=18
L.tunnel=524
time=(L.car+L.tunnel)/v
(c)
zhuamuntao=4515
shangue=1155
time=7
more=zhuamuntao/time-shangue/time
(d)
```

```
[Workspace loaded from E://dpclass1/.RData]

> age=17
> age>18
[1] FALSE
> is_adult=age>18
>
```

```
Values
  age              17
  is_adult         FALSE
```

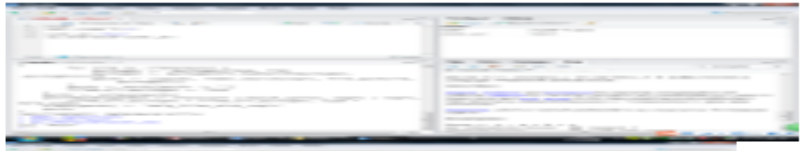
6、(a)

```
#question6
(a)
group_num=6
people_num=4
tree_num=360
```

```
Values
  average          15
  group_num        6
```

```
3. NUM<-numeric(4)
```

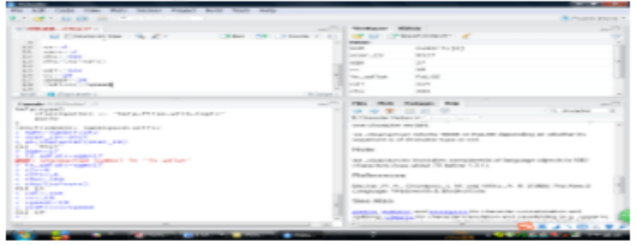
```
4. User_ID<-9527
 As.character(User_ID)
```

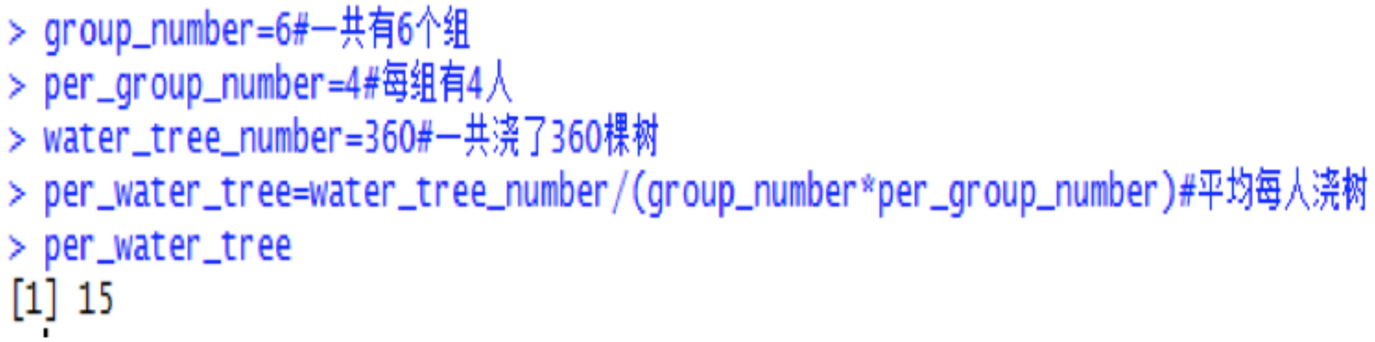


```
5. age<-17
 Is_adlut<-age>17
```

```
6. (1) xz<-6
       xzrs<-4
       shu<-360
       shu/(xz*xzrs)
```

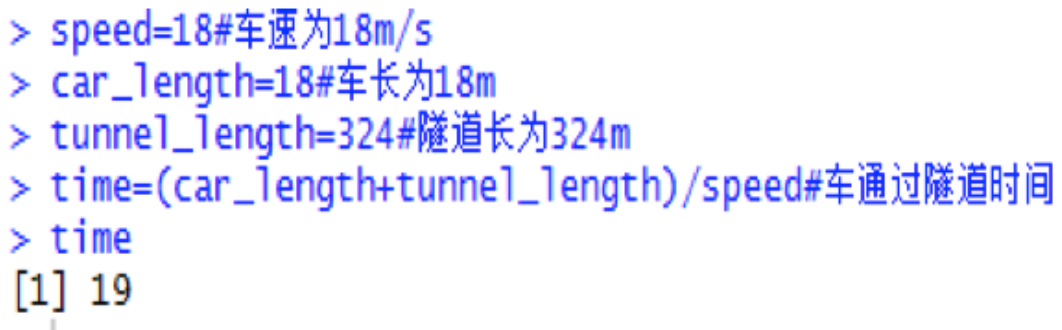
```
(2) sdl<-324
cc<-18
speed<-18
(sdl+cc)/speed
```





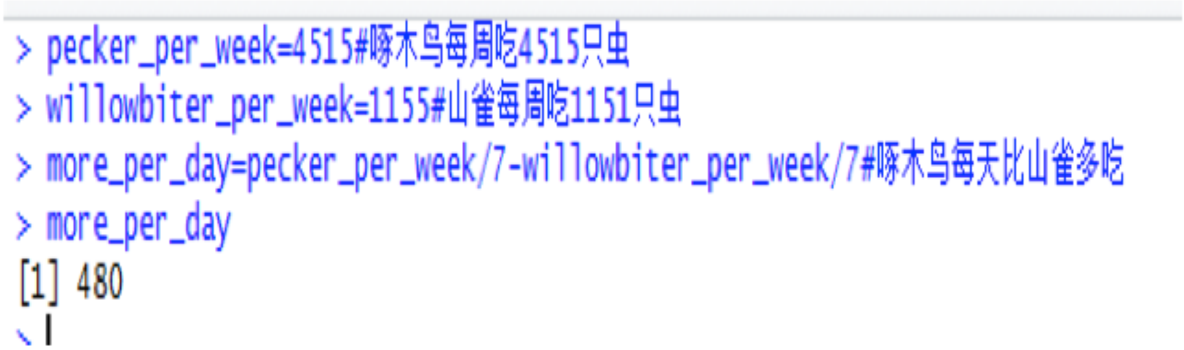
```
> group_number=6#一共有6个组
> per_group_number=4#每组有4人
> water_tree_number=360#一共浇了360棵树
> per_water_tree=water_tree_number/(group_number*per_group_number)#平均每人浇树
> per_water_tree
[1] 15
```

（2）



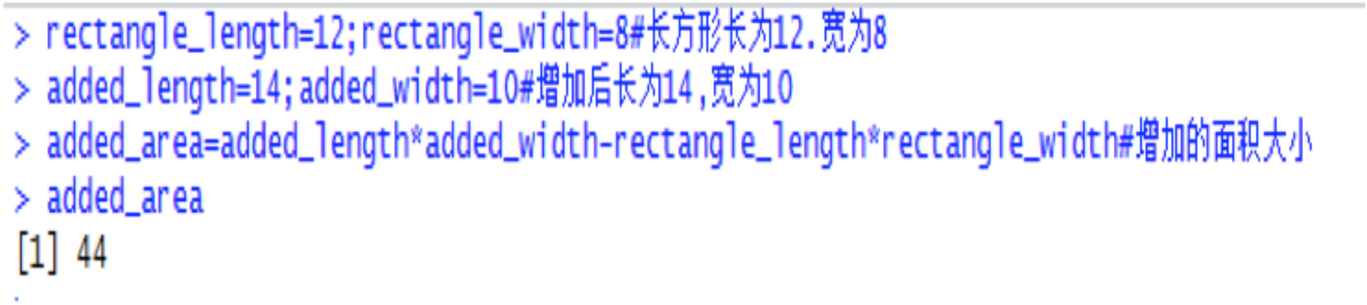
```
> speed=18#车速为18m/s
> car_length=18#车长为18m
> tunnel_length=324#隧道长为324m
> time=(car_length+tunnel_length)/speed#车通过隧道时间
> time
[1] 19
```

（3）



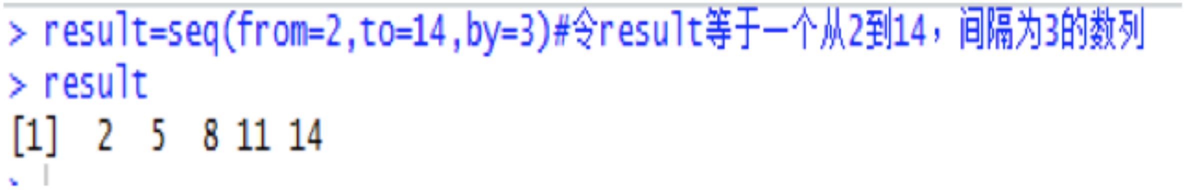
```
> pecker_per_week=4515#啄木鸟每周吃4515只虫
> willowbiter_per_week=1155#山雀每周吃1151只虫
> more_per_day=pecker_per_week/7-willowbiter_per_week/7#啄木鸟每天比山雀多吃
> more_per_day
[1] 480
```

（4）

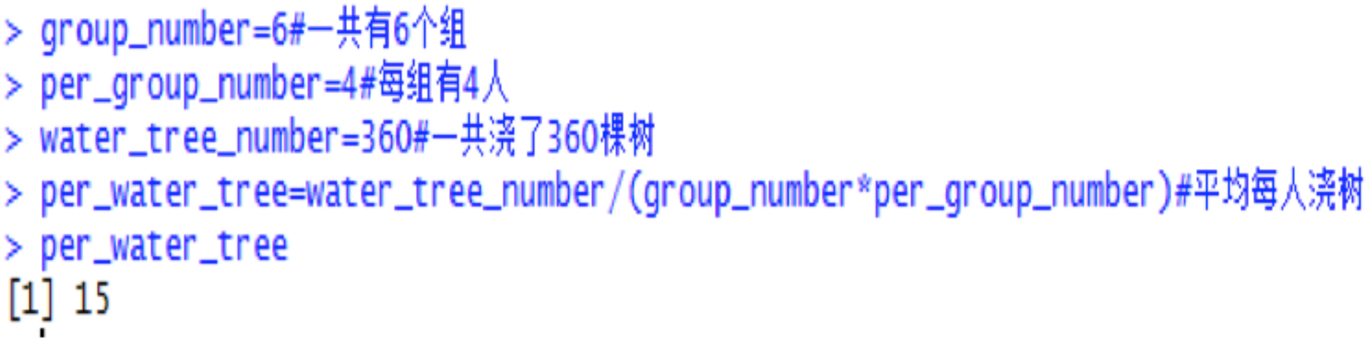


```
> rectangle_length=12;rectangle_width=8#长方形长为12.宽为8
> added_length=14;added_width=10#增加后长为14,宽为10
> added_area=added_length*added_width-rectangle_length*rectangle_width#增加的面积大小
> added_area
[1] 44
```

**7.**

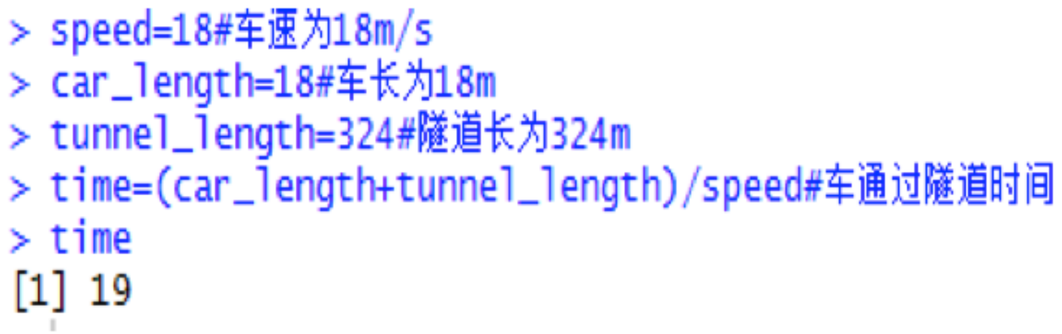


```
> result=seq(from=2,to=14,by=3)#令result等于一个从2到14，间隔为3的数列
> result
[1]  2  5  8 11 14
```



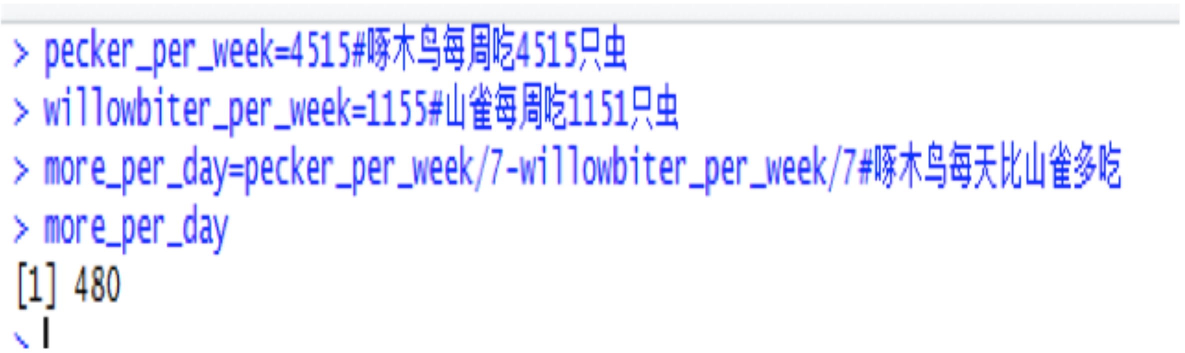
```
> group_number=6#一共有6个组
> per_group_number=4#每组有4人
> water_tree_number=360#一共浇了360棵树
> per_water_tree=water_tree_number/(group_number*per_group_number)#平均每人浇树
> per_water_tree
[1] 15
```

（2）



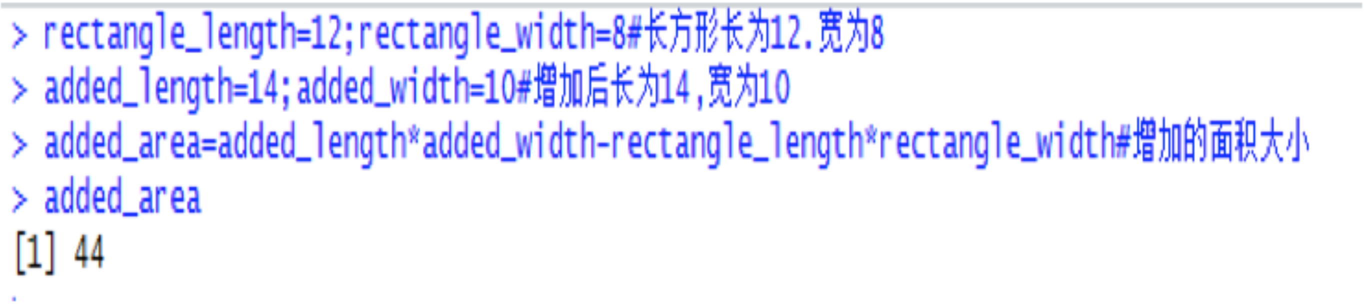
```
> speed=18#车速为18m/s
> car_length=18#车长为18m
> tunnel_length=324#隧道长为324m
> time=(car_length+tunnel_length)/speed#车通过隧道时间
> time
[1] 19
```

（3）



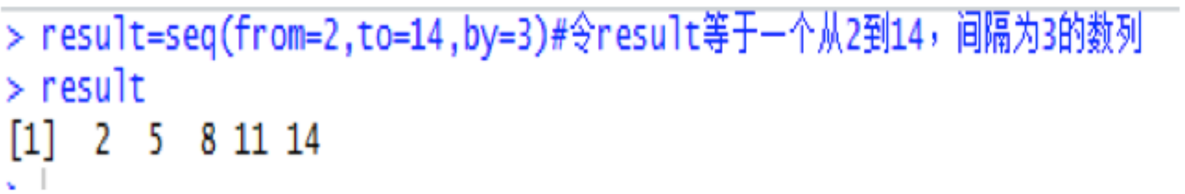
```
> pecker_per_week=4515#啄木鸟每周吃4515只虫
> willowbiter_per_week=1155#山雀每周吃1151只虫
> more_per_day=pecker_per_week/7-willowbiter_per_week/7#啄木鸟每天比山雀多吃
> more_per_day
[1] 480
```

（4）



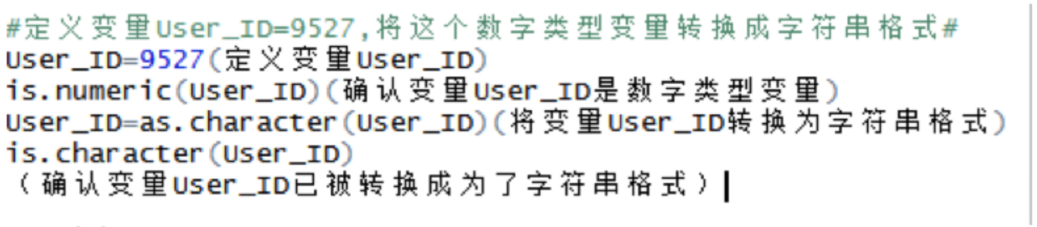
```
> rectangle_length=12;rectangle_width=8#长方形长为12.宽为8
> added_length=14;added_width=10#增加后长为14,宽为10
> added_area=added_length*added_width-rectangle_length*rectangle_width#增加的面积大小
> added_area
[1] 44
```

**7.**



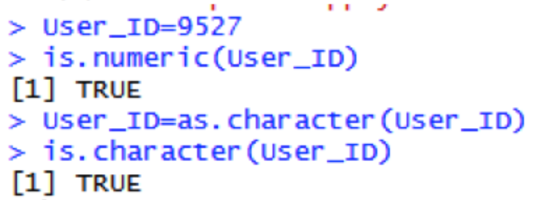
```
> result=seq(from=2,to=14,by=3)#令result等于一个从2到14，间隔为3的数列
> result
[1]  2  5  8 11 14
```

4. (1) code



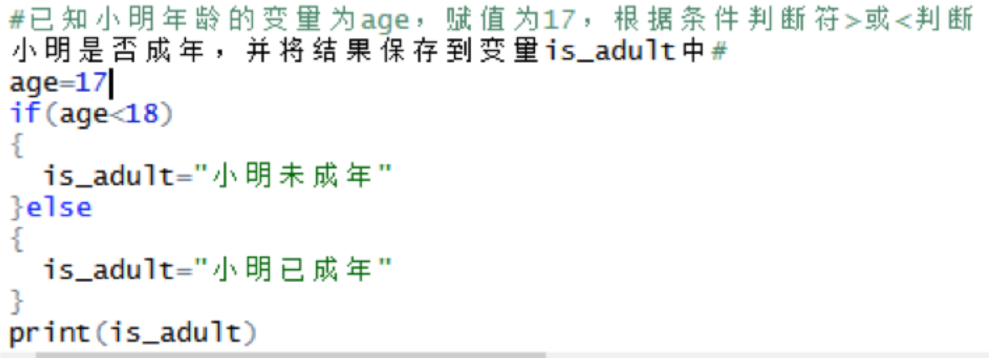
```
#定义变量User_ID=9527,将这个数字类型变量转换成字符串格式#
User_ID=9527(定义变量User_ID)
is.numeric(User_ID)(确认变量User_ID是数字类型变量)
User_ID=as.character(User_ID)(将变量User_ID转换为字符串格式)
is.character(User_ID)
（确认变量User_ID已被转换成为了字符串格式）
```

(2) result



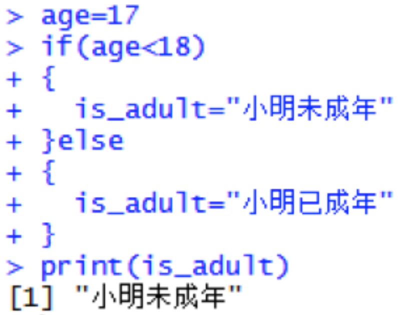
```
> User_ID=9527
> is.numeric(User_ID)
[1] TRUE
> User_ID=as.character(User_ID)
> is.character(User_ID)
[1] TRUE
```

5. (1) code



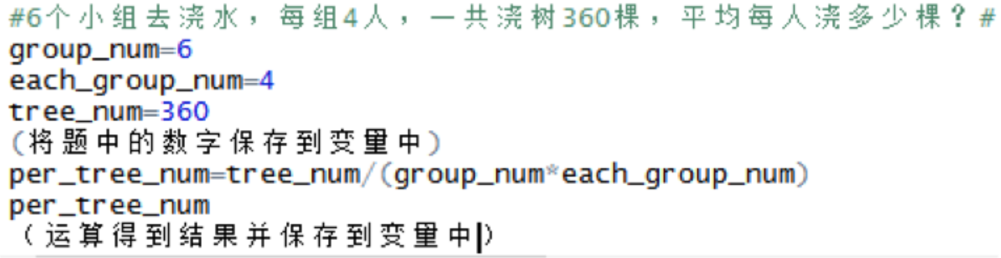
```
#已知小明年龄的变量为age，赋值为17，根据条件判断符>或<判断
小明是否成年，并将结果保存到变量is_adult中#
age=17
if(age<18)
{
  is_adult="小明未成年"
}else
{
  is_adult="小明已成年"
}
print(is_adult)
```

(2) result



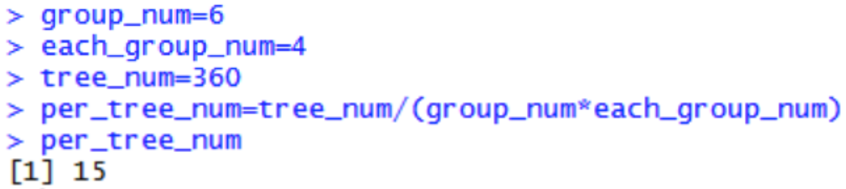
```
> age=17
> if(age<18)
+ {
+   is_adult="小明未成年"
+ }else
+ {
+   is_adult="小明已成年"
+ }
> print(is_adult)
[1] "小明未成年"
```

6.a. (1) code



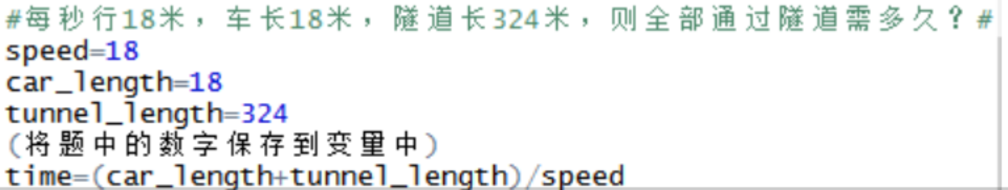
```
#6个小组去浇水，每组4人，一共浇树360棵，平均每人浇多少棵？#
group_num=6
each_group_num=4
tree_num=360
(将题中的数字保存到变量中)
per_tree_num=tree_num/(group_num*each_group_num)
per_tree_num
（运算得到结果并保存到变量中)
```

(2) result



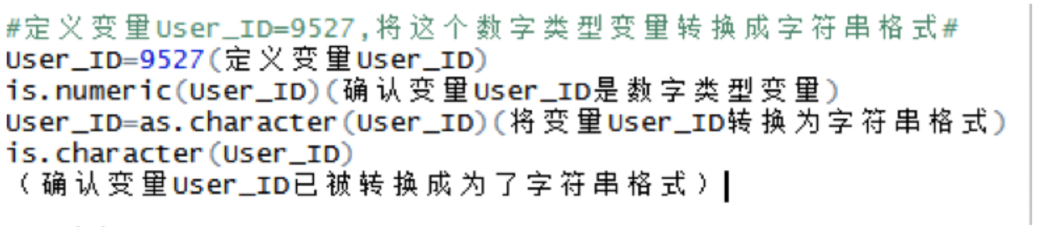
```
> group_num=6
> each_group_num=4
> tree_num=360
> per_tree_num=tree_num/(group_num*each_group_num)
> per_tree_num
[1] 15
```

b. (1) code



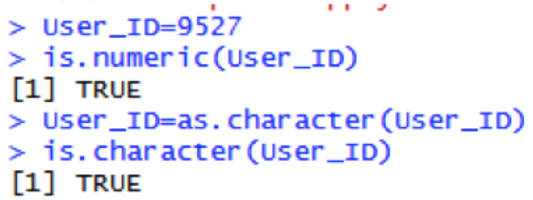
```
#每秒行18米，车长18米，隧道长324米，则全部通过隧道需多久？#
speed=18
car_length=18
tunnel_length=324
(将题中的数字保存到变量中)
time=(car_length+tunnel_length)/speed
```

4. (1) code



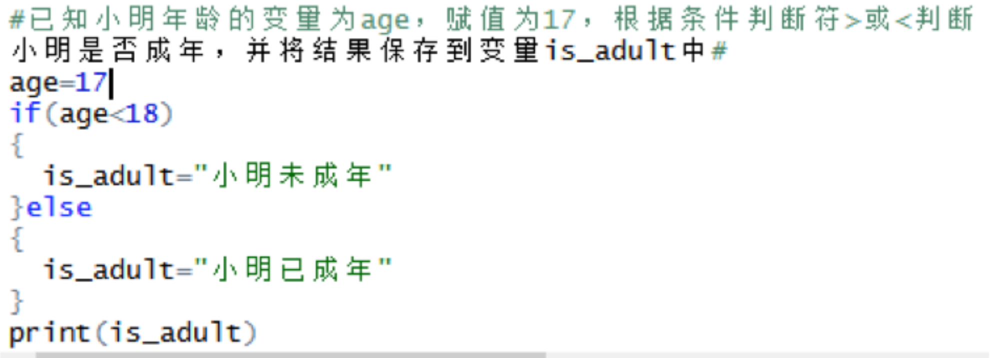
```
#定义变量User_ID=9527,将这个数字类型变量转换成字符串格式#
User_ID=9527(定义变量User_ID)
is.numeric(User_ID)(确认变量User_ID是数字类型变量)
User_ID=as.character(User_ID)(将变量User_ID转换为字符串格式)
is.character(User_ID)
（确认变量User_ID已被转换成为了字符串格式）
```

(2) result



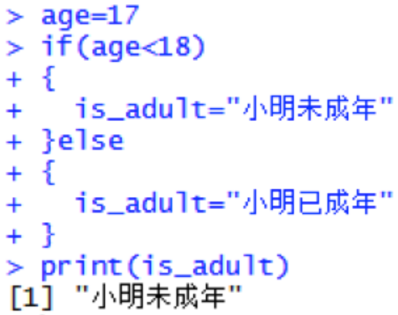
```
> User_ID=9527
> is.numeric(User_ID)
[1] TRUE
> User_ID=as.character(User_ID)
> is.character(User_ID)
[1] TRUE
```

5. (1) code



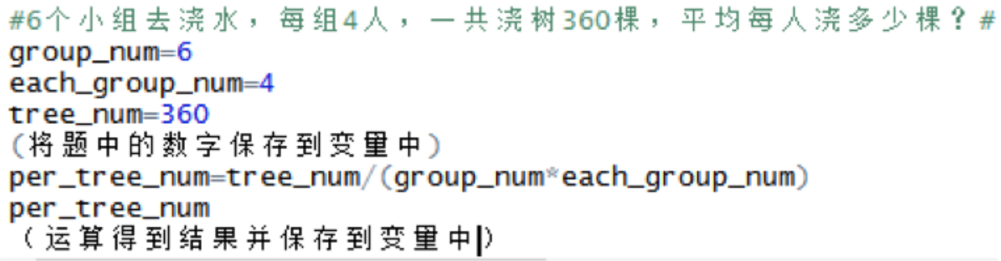
```
#已知小明年龄的变量为age，赋值为17，根据条件判断符>或<判断
小明是否成年，并将结果保存到变量is_adult中#
age=17
if(age<18)
{
  is_adult="小明未成年"
}else
{
  is_adult="小明已成年"
}
print(is_adult)
```

(2) result



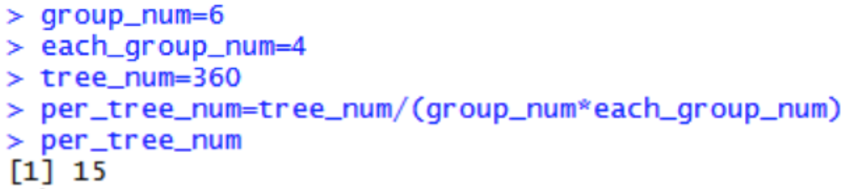
```
> age=17
> if(age<18)
+ {
+   is_adult="小明未成年"
+ }else
+ {
+   is_adult="小明已成年"
+ }
> print(is_adult)
[1] "小明未成年"
```

6.a. (1) code



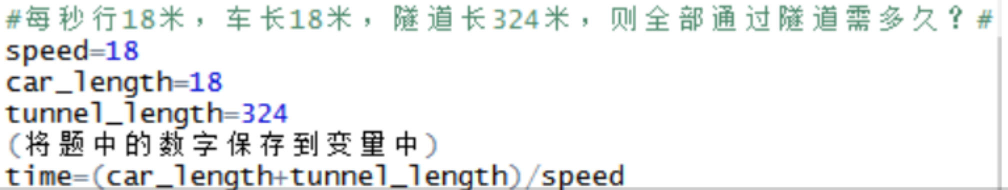
```
#6个小组去浇水，每组4人，一共浇树360棵，平均每人浇多少棵？#
group_num=6
each_group_num=4
tree_num=360
(将题中的数字保存到变量中)
per_tree_num=tree_num/(group_num*each_group_num)
per_tree_num
（运算得到结果并保存到变量中)
```

(2) result



```
> group_num=6
> each_group_num=4
> tree_num=360
> per_tree_num=tree_num/(group_num*each_group_num)
> per_tree_num
[1] 15
```

b. (1) code

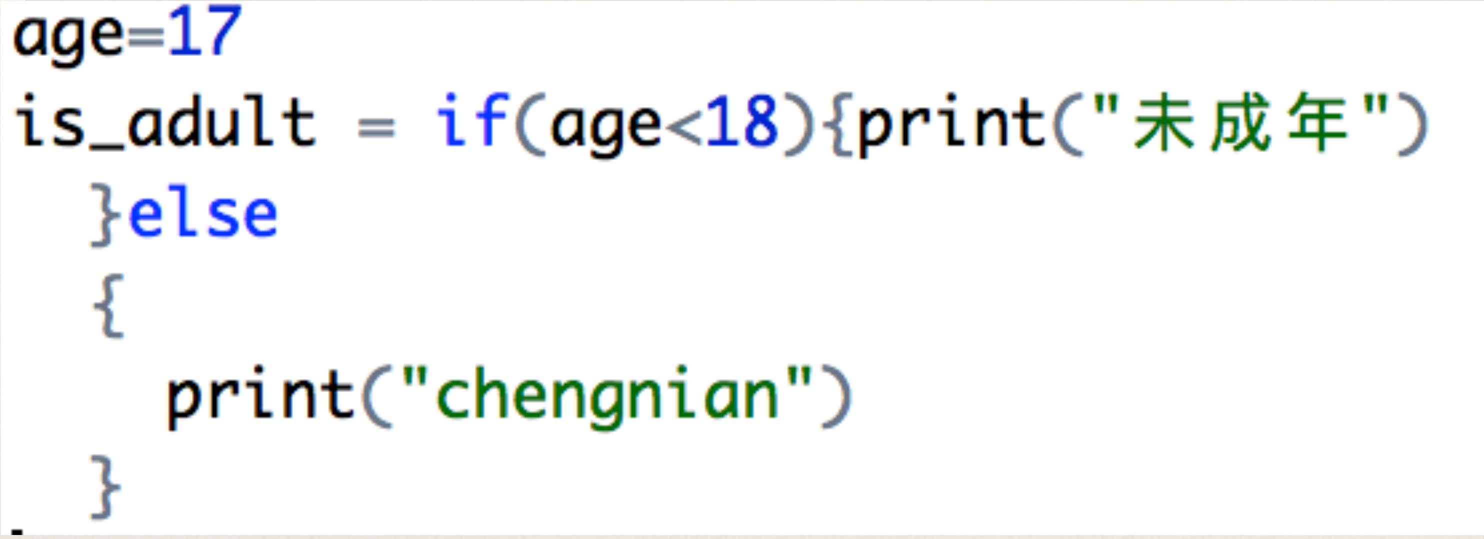


```
#每秒行18米，车长18米，隧道长324米，则全部通过隧道需多久？#
speed=18
car_length=18
tunnel_length=324
(将题中的数字保存到变量中)
time=(car_length+tunnel_length)/speed
```

# Make Your Document Readable!

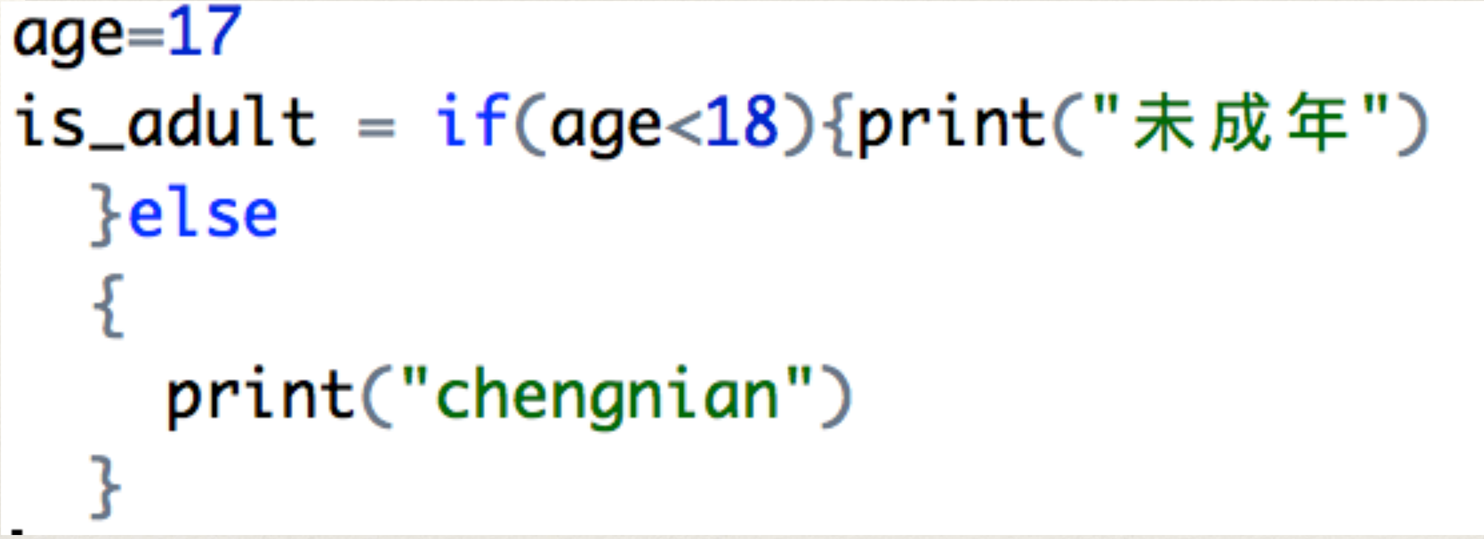
# Question 5

# Question 5

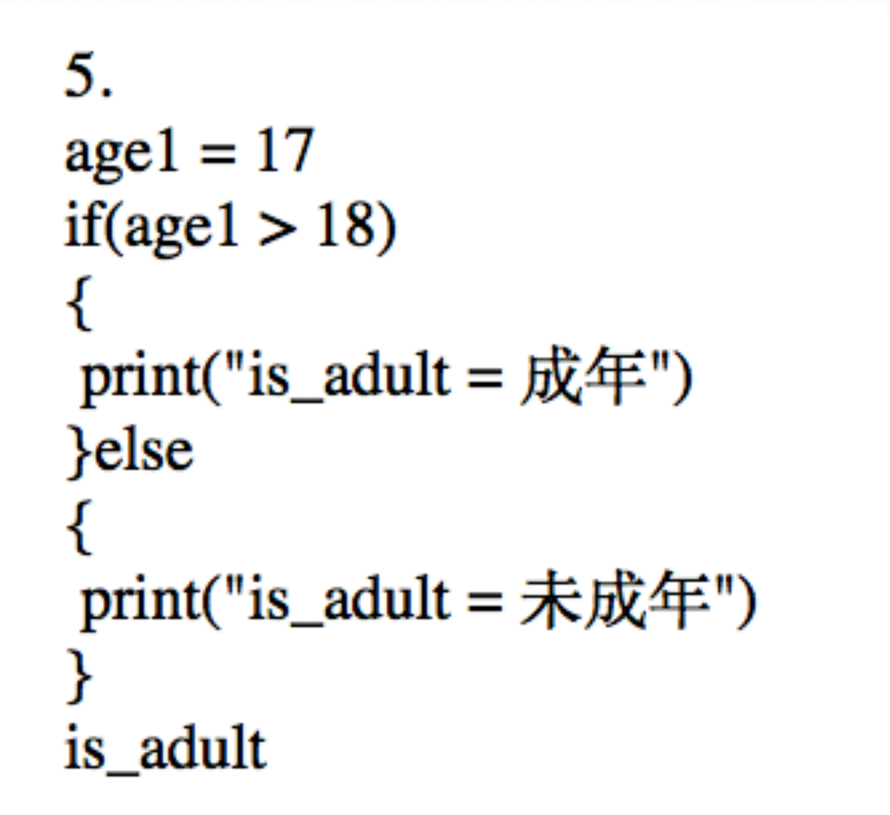


```
age=17
is_adult = if(age<18){print("未成年")
  }else
  {
    print("chengnian")
  }
```

# Question 5

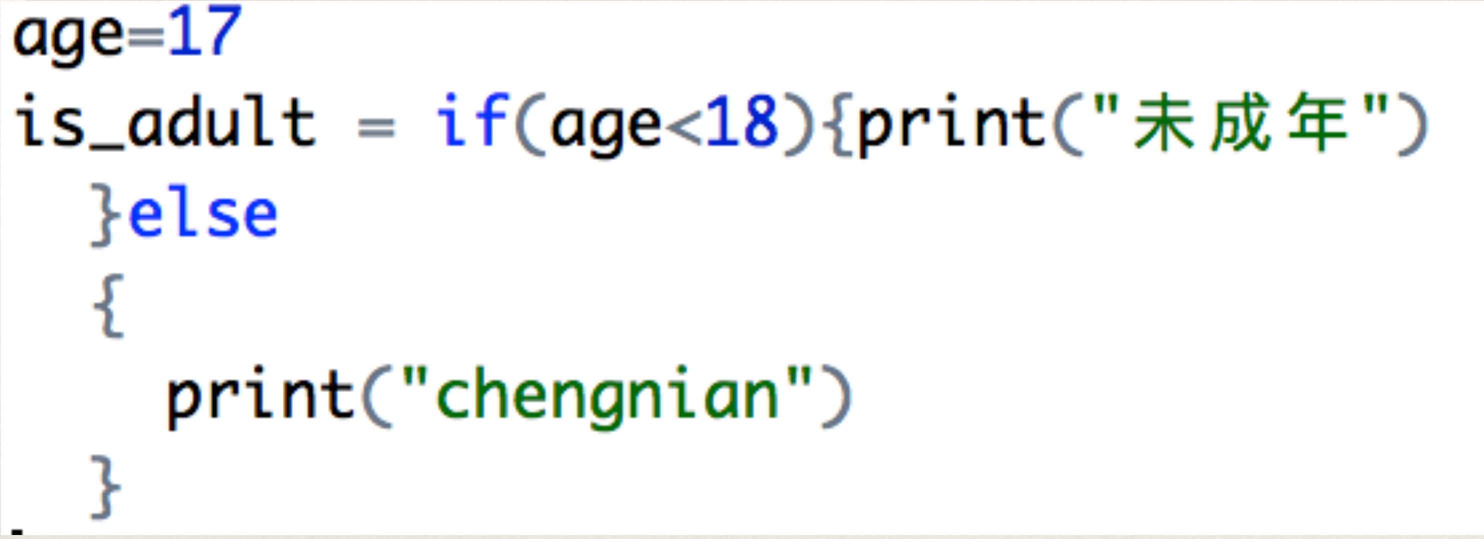


```
age=17
is_adult = if(age<18){print("未成年")
  }else
  {
    print("chengnian")
  }
```

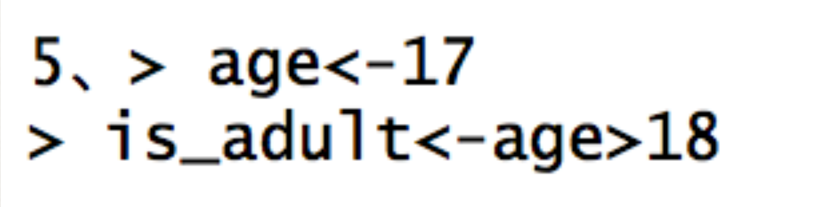


```
5.
age1 = 17
if(age1 > 18)
{
 print("is_adult = 成年")
}else
{
 print("is_adult = 未成年")
}
is_adult
```

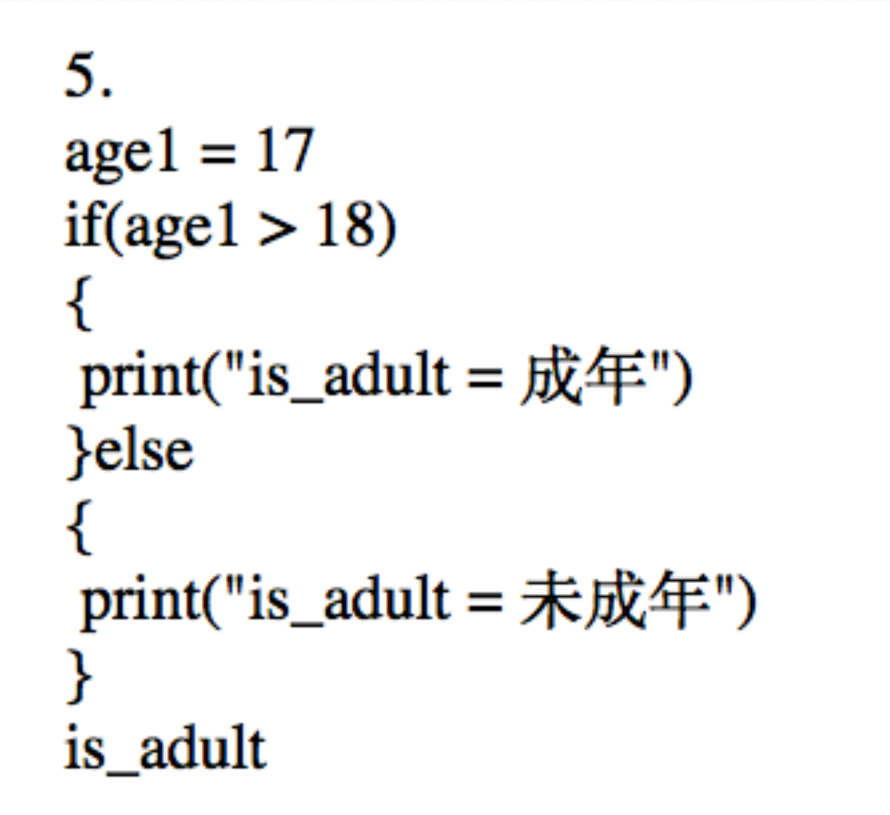
# Question 5



```
age=17
is_adult = if(age<18){print("未成年")
  }else
  {
    print("chengnian")
  }
```

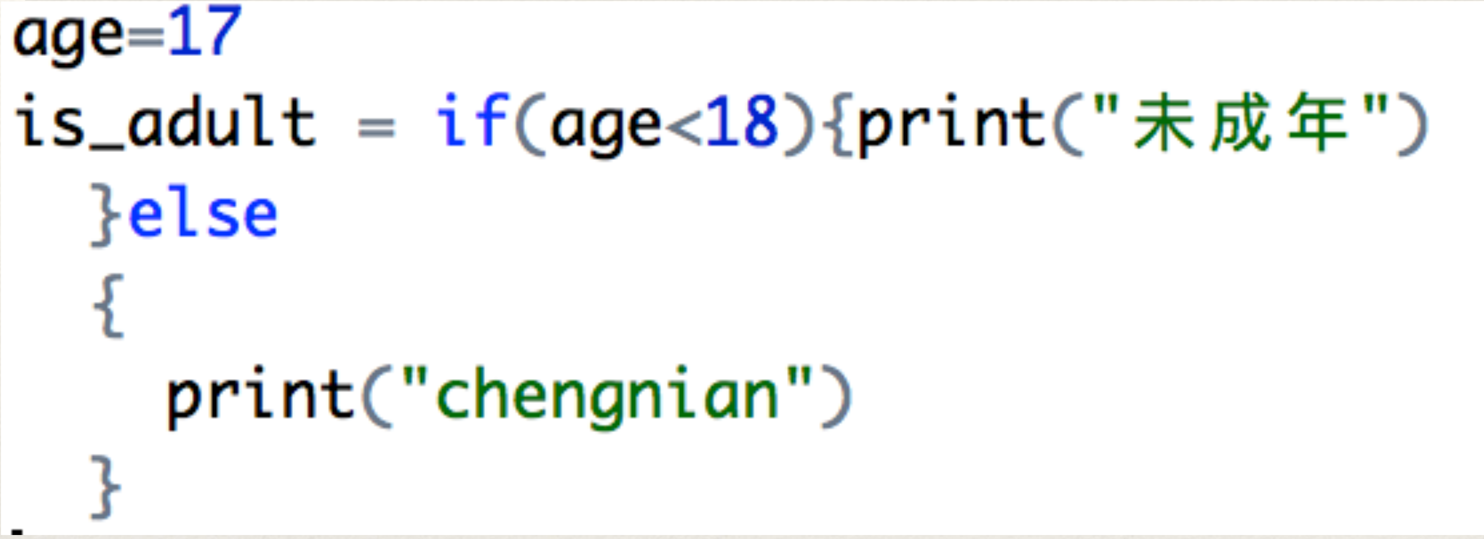


```
5、> age<-17
> is_adult<-age>18
```

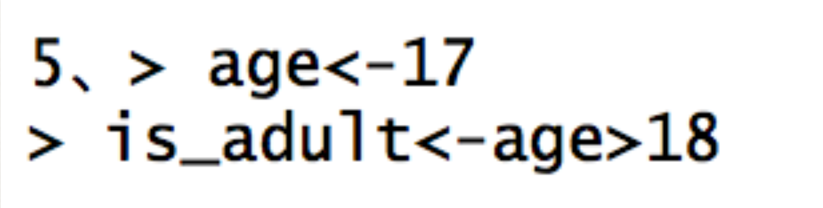


```
5.
age1 = 17
if(age1 > 18)
{
 print("is_adult = 成年")
}else
{
 print("is_adult = 未成年")
}
is_adult
```

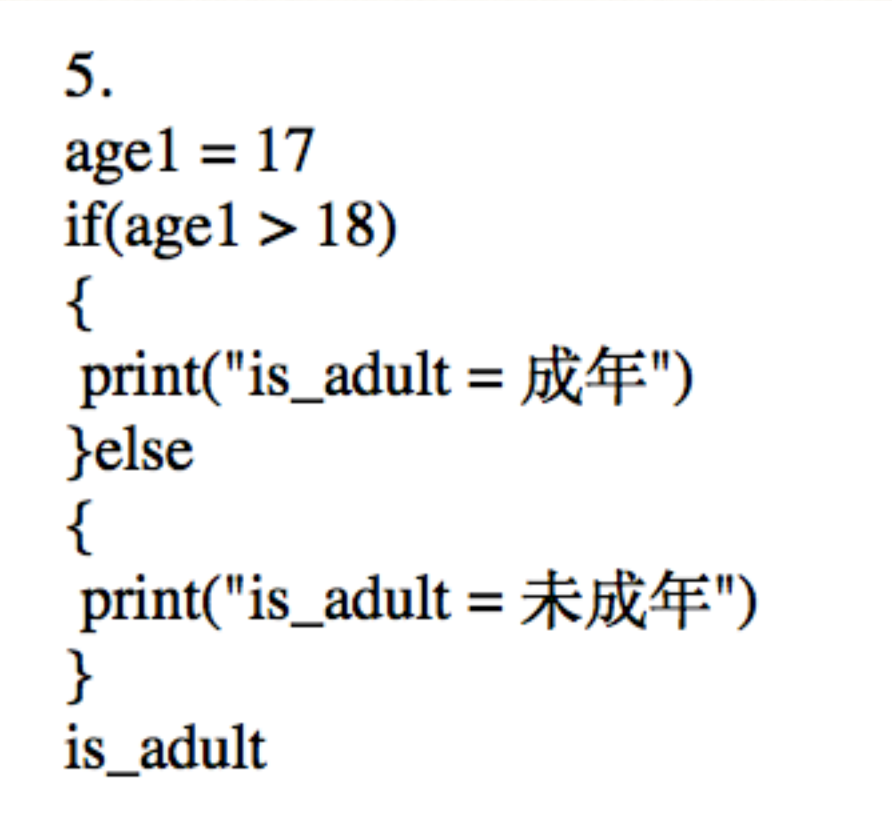
# Question 5



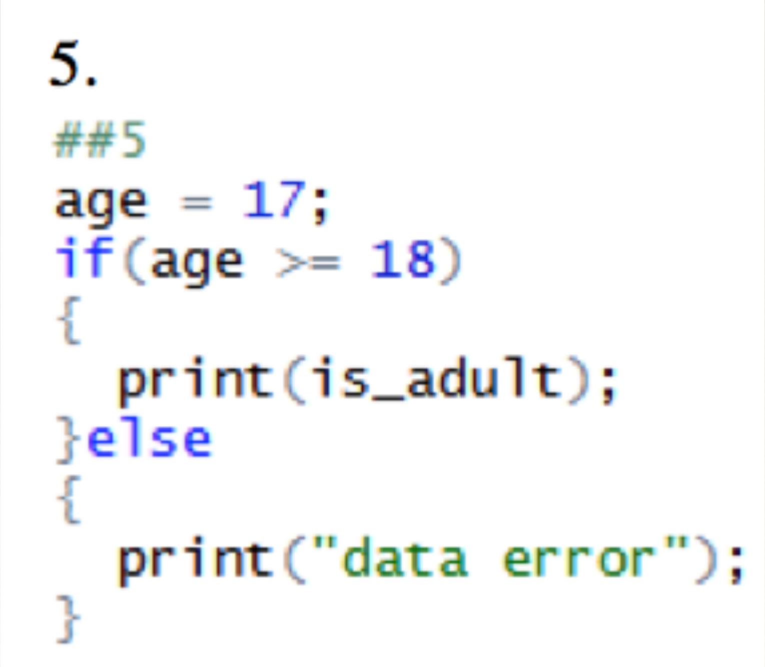
```
age=17
is_adult = if(age<18){print("未成年")
  }else
  {
    print("chengnian")
  }
```



```
5、> age<-17
> is_adult<-age>18
```

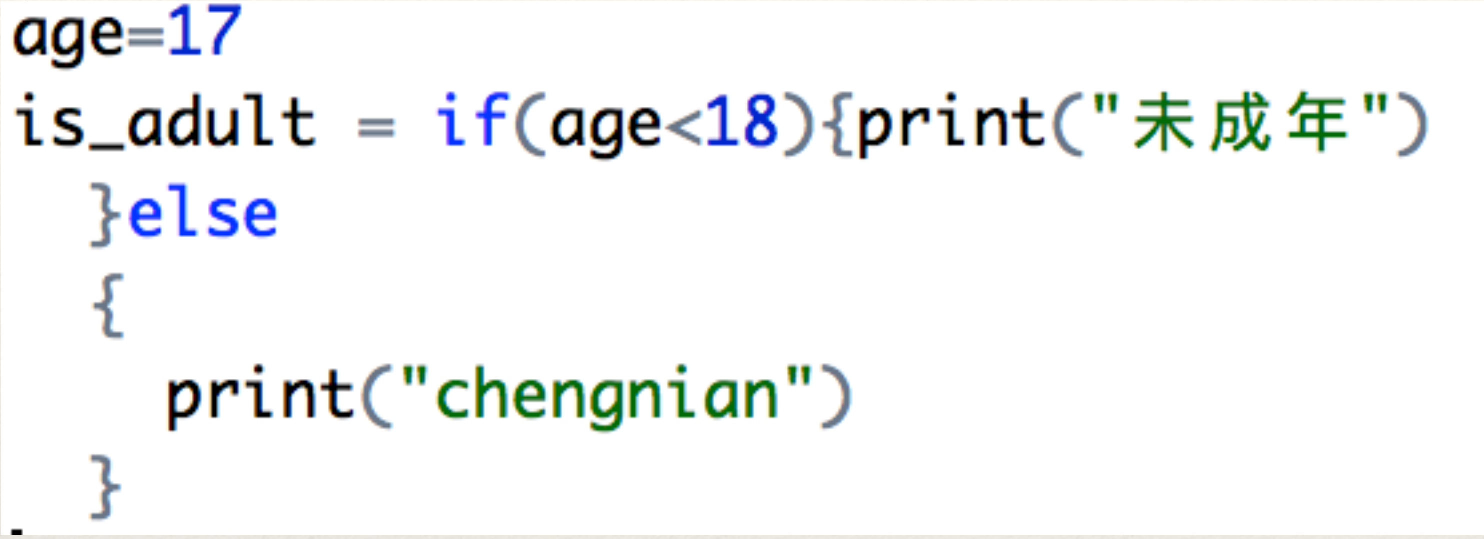


```
5.
age1 = 17
if(age1 > 18)
{
 print("is_adult = 成年")
}else
{
 print("is_adult = 未成年")
}
is_adult
```

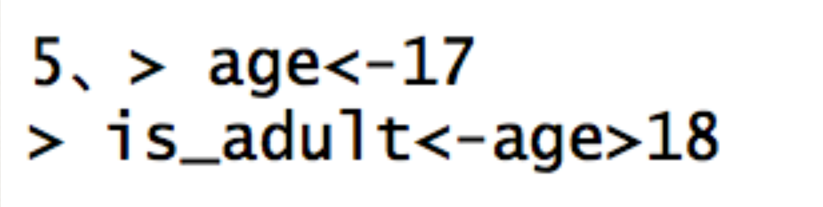


```
5.
##5
age = 17;
if(age >= 18)
{
  print(is_adult);
}else
{
  print("data error");
}
```

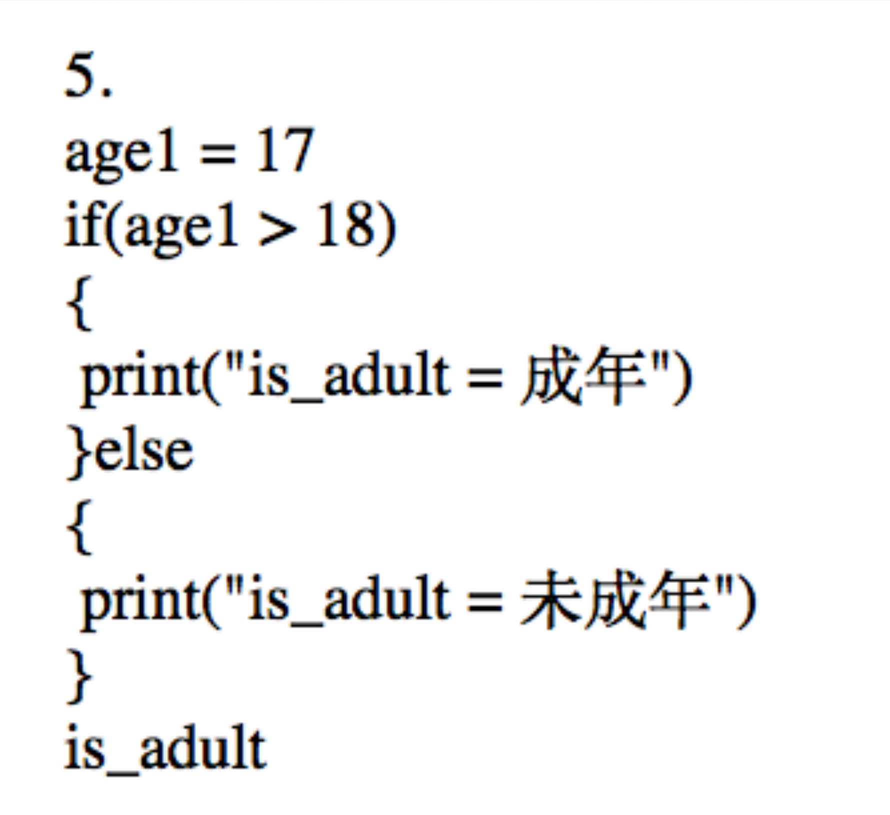
# Question 5



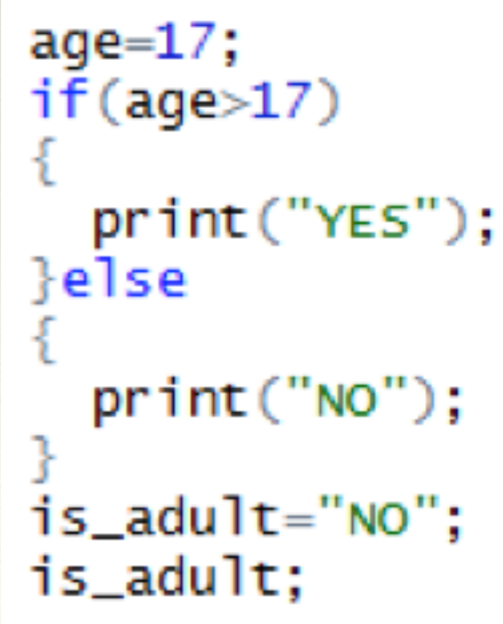
```
age=17
is_adult = if(age<18){print("未成年")
  }else
  {
    print("chengnian")
  }
```



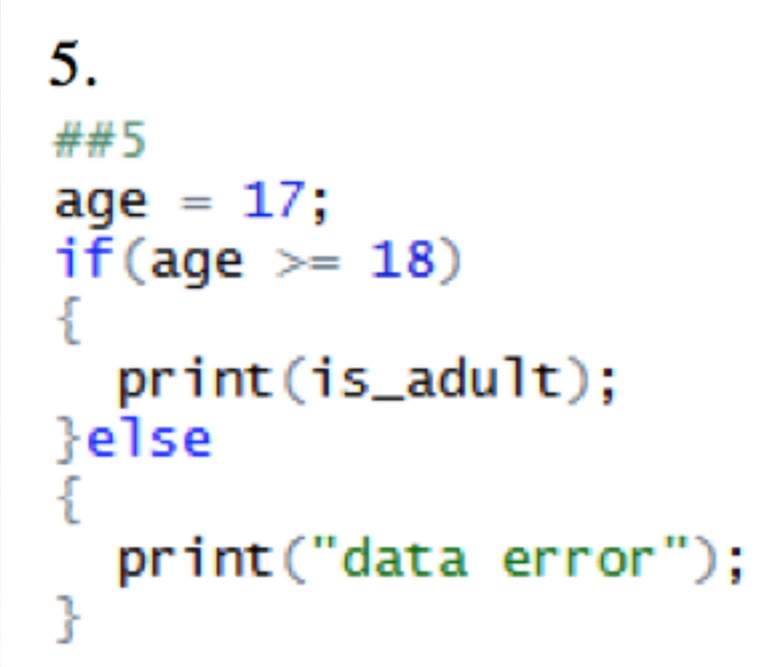
```
5、> age<-17
> is_adult<-age>18
```



```
5.
age1 = 17
if(age1 > 18)
{
 print("is_adult = 成年")
}else
{
 print("is_adult = 未成年")
}
is_adult
```

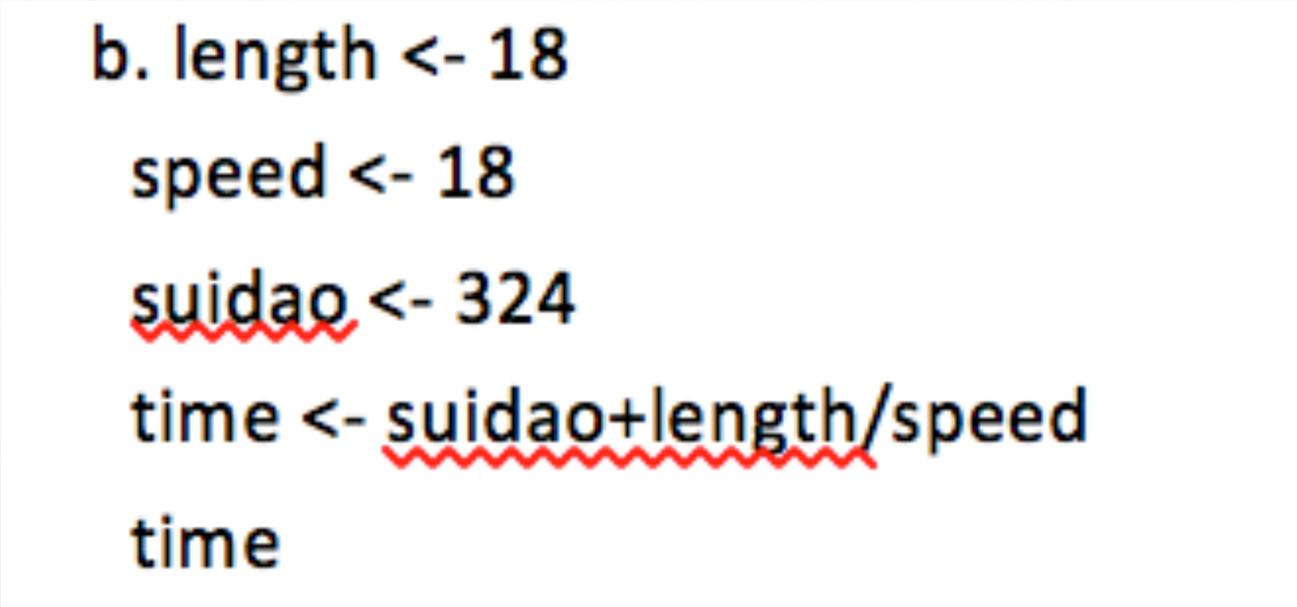


```
age=17;
if(age>17)
{
  print("YES");
}else
{
  print("NO");
}
is_adult="NO";
is_adult;
```



```
5.
##5
age = 17;
if(age >= 18)
{
  print(is_adult);
}else
{
  print("data error");
}
```

# Question 6



```
b. length <- 18
  speed <- 18
  suidao <- 324
  time <- suidao+length/speed
  time
```

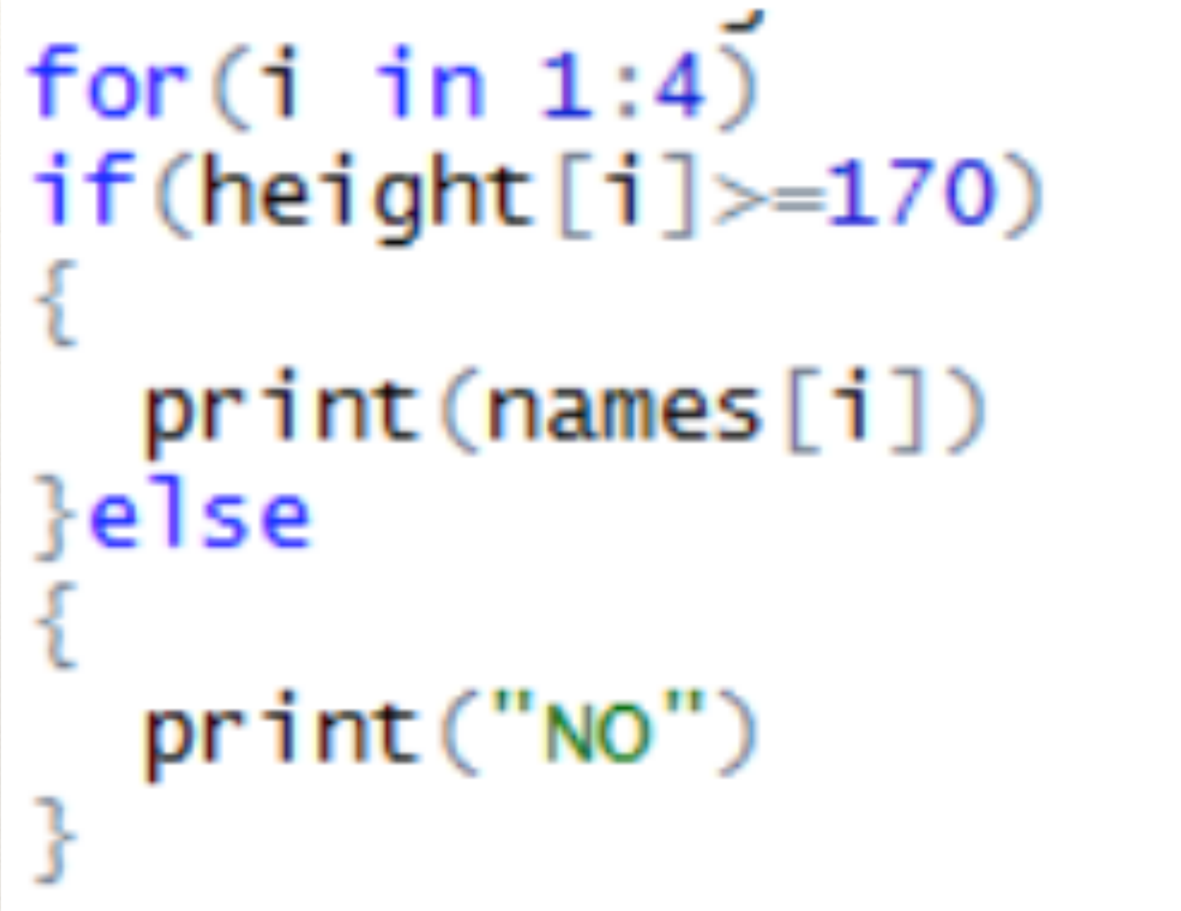
# Question 7

# Question 8

# Question 9

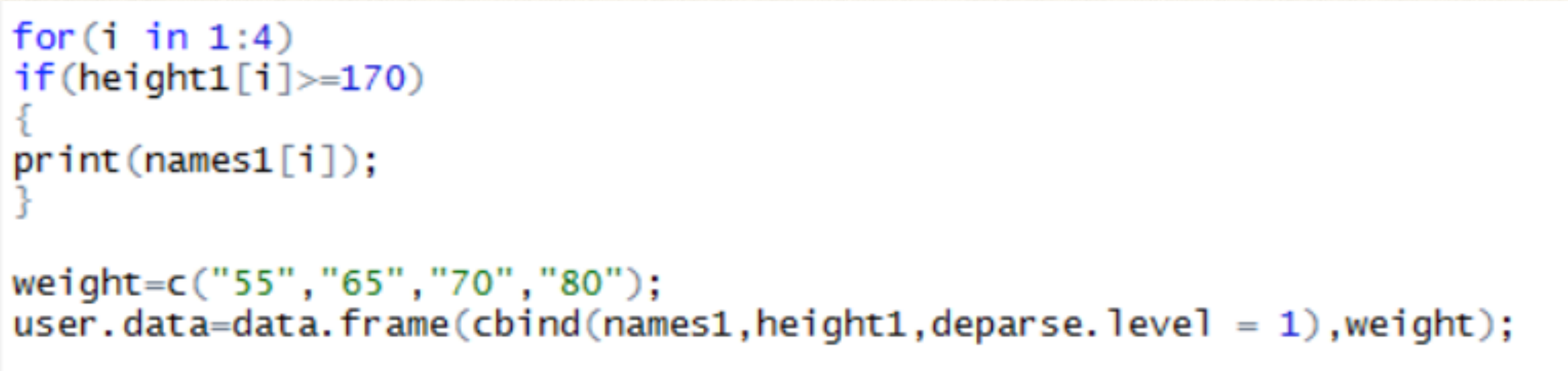
# Question 10

# Question 11

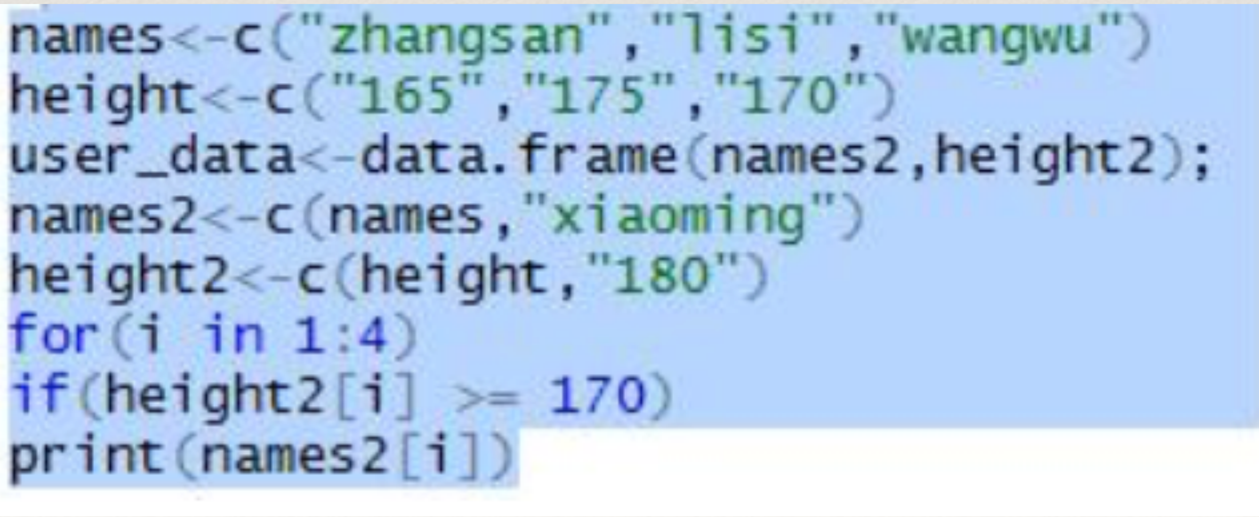


```
for(i in 1:4)
if(height[i]>=170)
{
  print(names[i])
}else
{
  print("NO")
}
```

# Question 12



```
for(i in 1:4)
if(height1[i]>=170)
{
print(names1[i]);
}

weight=c("55","65","70","80");
user.data=data.frame(cbind(names1,height1,deparse.level = 1),weight);
```



```
names<-c("zhangsan","lisi","wangwu")
height<-c("165","175","170")
user_data<-data.frame(names2,height2);
names2<-c(names,"xiaoming")
height2<-c(height,"180")
for(i in 1:4)
if(height2[i] >= 170)
print(names2[i])
```

# Good Homework

- DIY
- Readable
- Unique
- Insightful
- Don’t need to be all right

# Functional Area

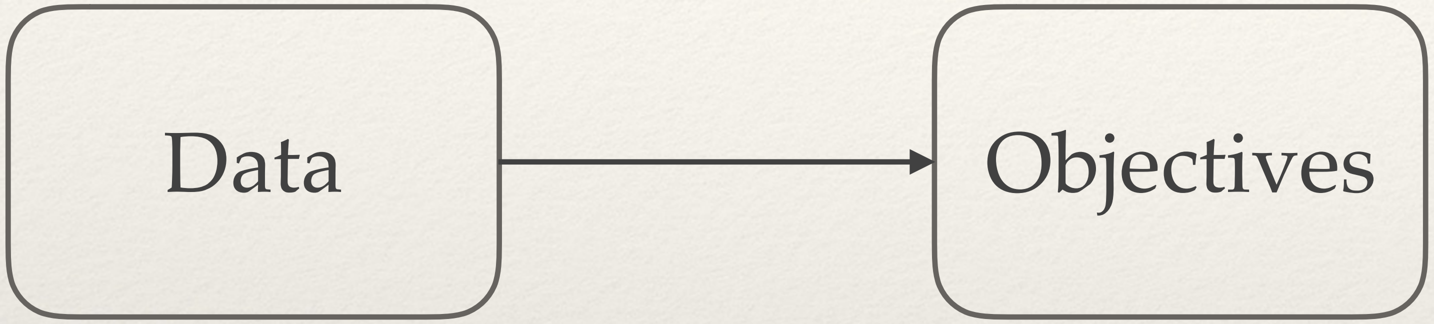
- For, If
  - we can ignore {}, only if there is one line program
- Keep a good programming style
- KISS (keep it stay simple), but not Wrong

# Data Process

# Data Process

Data

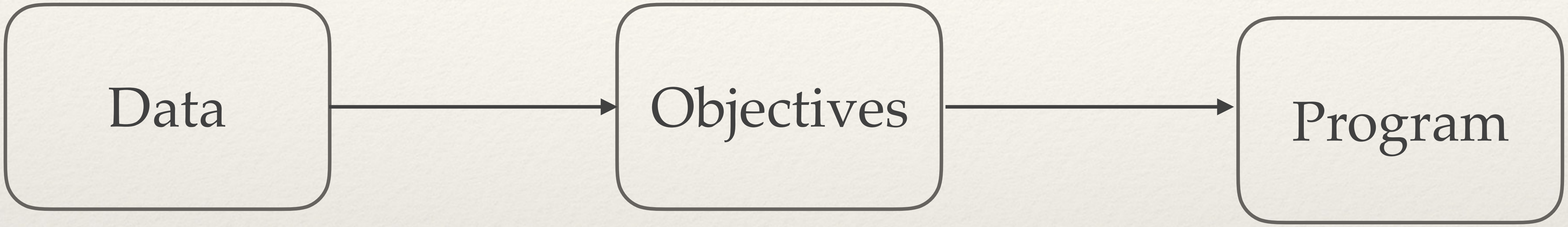
# Data Process



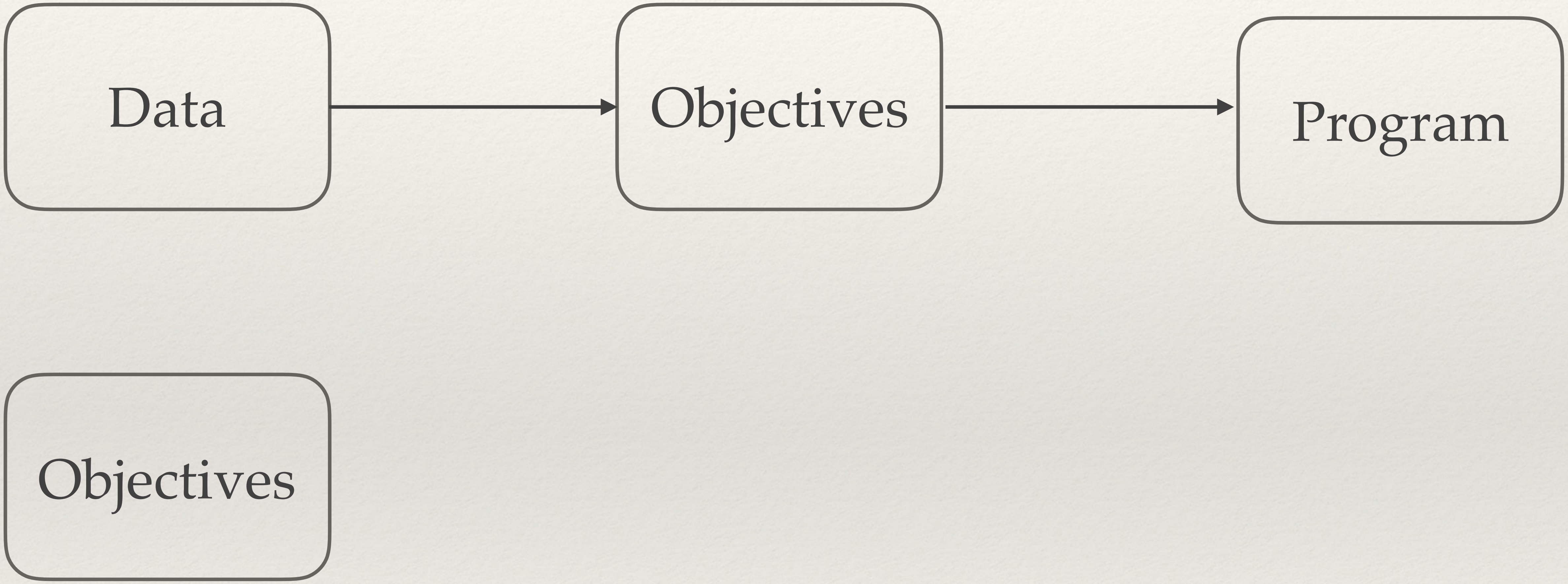
Data

Objectives

# Data Process



# Data Process



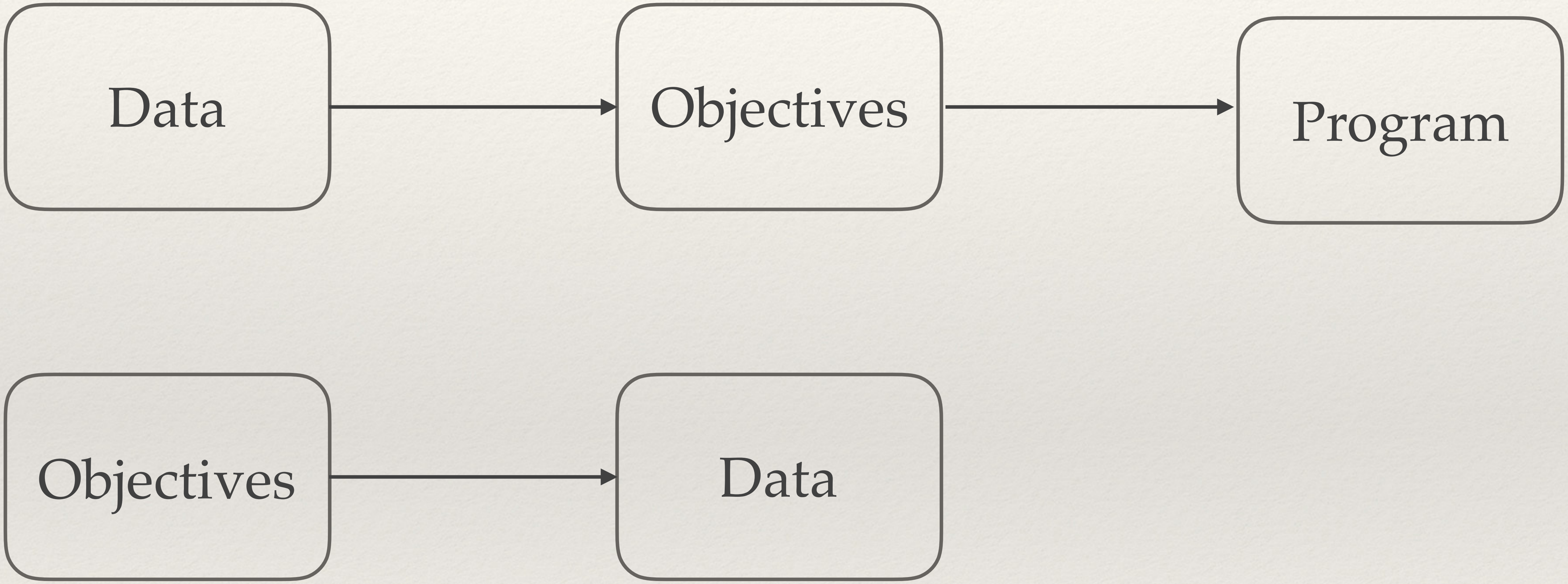
Data

Objectives

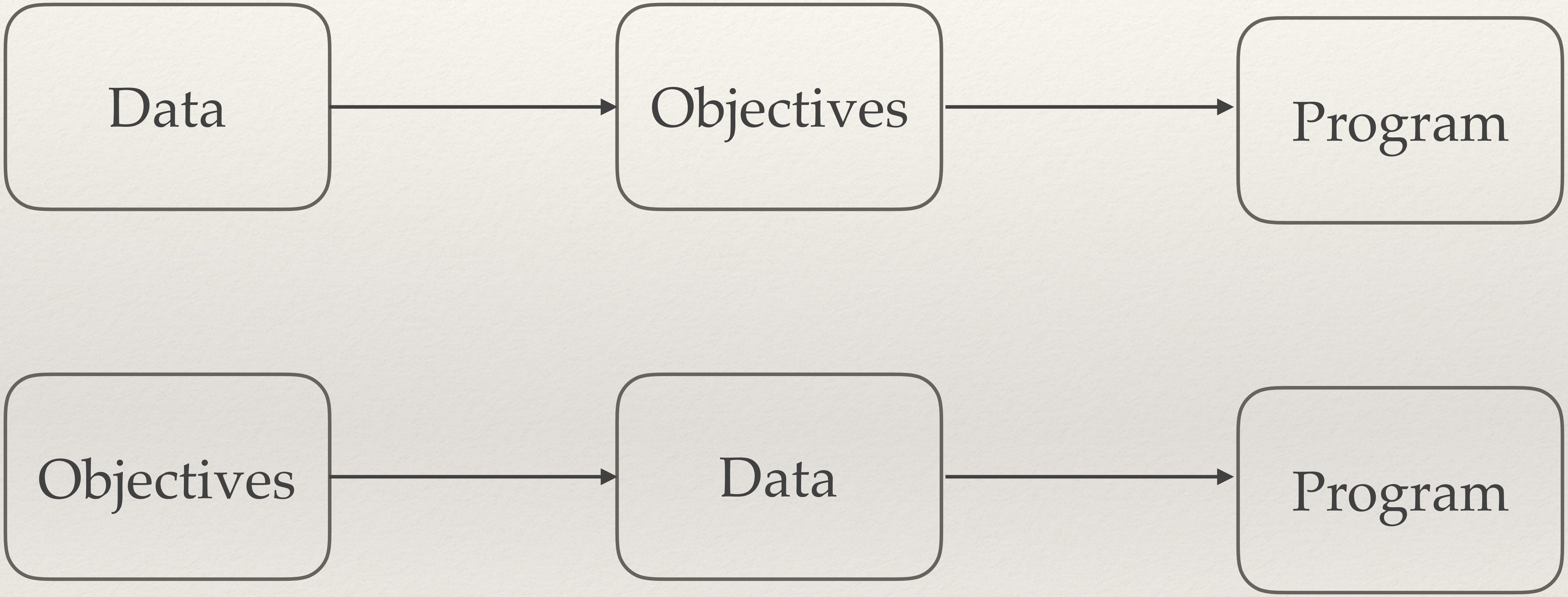
Program

Objectives

# Data Process



# Data Process



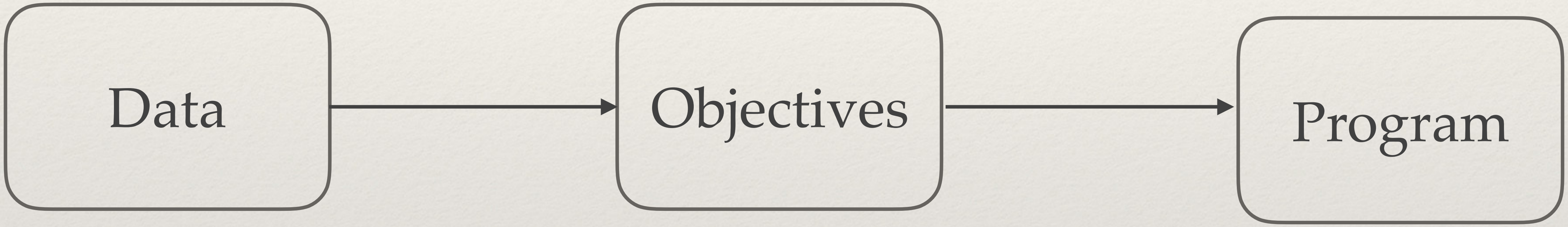
# Example

- After the final exam, the director wants to know which class performs better?

- Who need to diet?

# Example

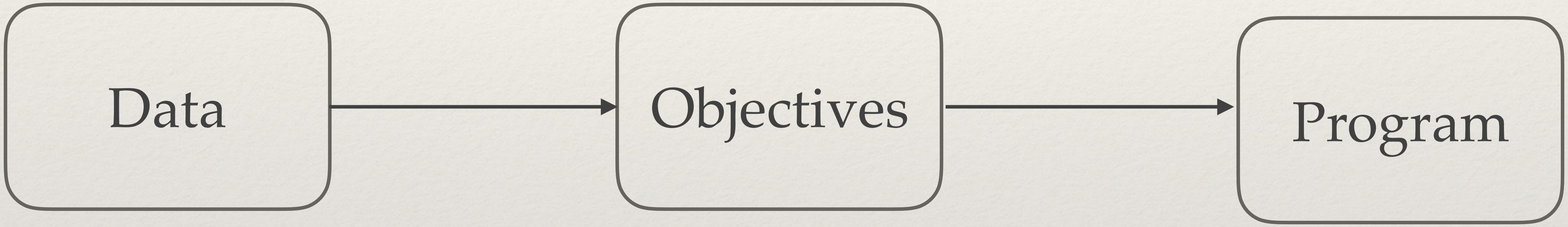
- After the final exam, the director wants to know which class performs better?



- Who need to diet?

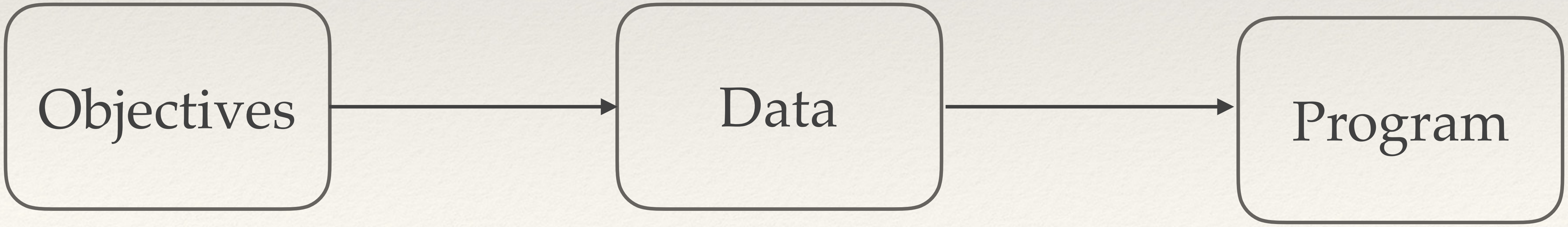
# Example

- After the final exam, the director wants to know which class performs better?



Data → Objectives → Program

- Who need to diet?



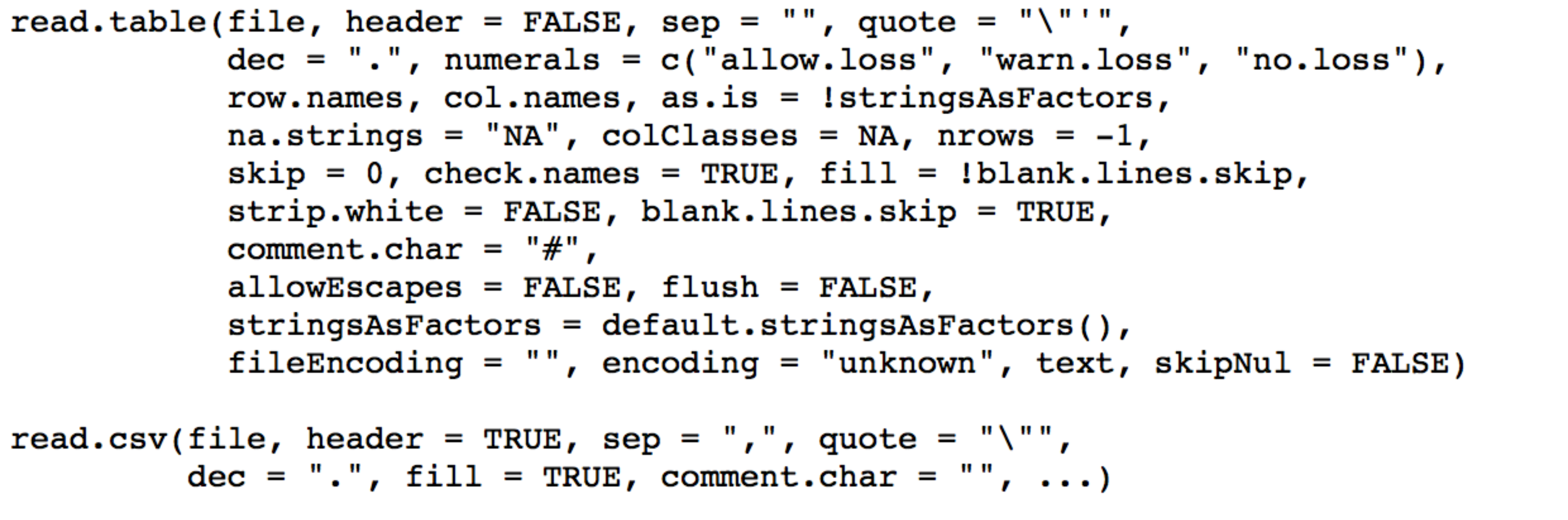
Objectives → Data → Program

# Outline

- Review of Last Class
- Homework 1 Explanation
- Read and Write Data files
- Brief Description Analysis
- Conclusions

# Read

- File type: .txt, .csv
- Functions: read.table, read.csv



```
read.table(file, header = FALSE, sep = "", quote = "\"'",
           dec = ".", numerals = c("allow.loss", "warn.loss", "no.loss"),
           row.names, col.names, as.is = !stringsAsFactors,
           na.strings = "NA", colClasses = NA, nrows = -1,
           skip = 0, check.names = TRUE, fill = !blank.lines.skip,
           strip.white = FALSE, blank.lines.skip = TRUE,
           comment.char = "#",
           allowEscapes = FALSE, flush = FALSE,
           stringsAsFactors = default.stringsAsFactors(),
           fileEncoding = "", encoding = "unknown", text, skipNul = FALSE)

read.csv(file, header = TRUE, sep = ",", quote = "\"",
         dec = ".", fill = TRUE, comment.char = "", ...)
```

# Brief Introduction of Function

- Function name
- Input
  - Data
  - Parameter (default parameter)
- Output
  - Return value
  - Can be considered as temporary variable

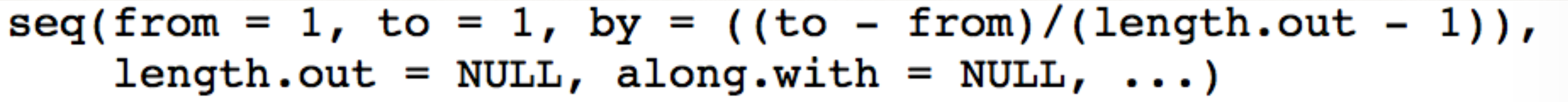
# Example

- a=c(1,2,3,4)
- b=min(a)
- c=seq(from=1,to=10,by=2)

- d=seq(to=10,by=4)

# Example

- a=c(1,2,3,4)
- b=min(a)
- c=seq(from=1,to=10,by=2)

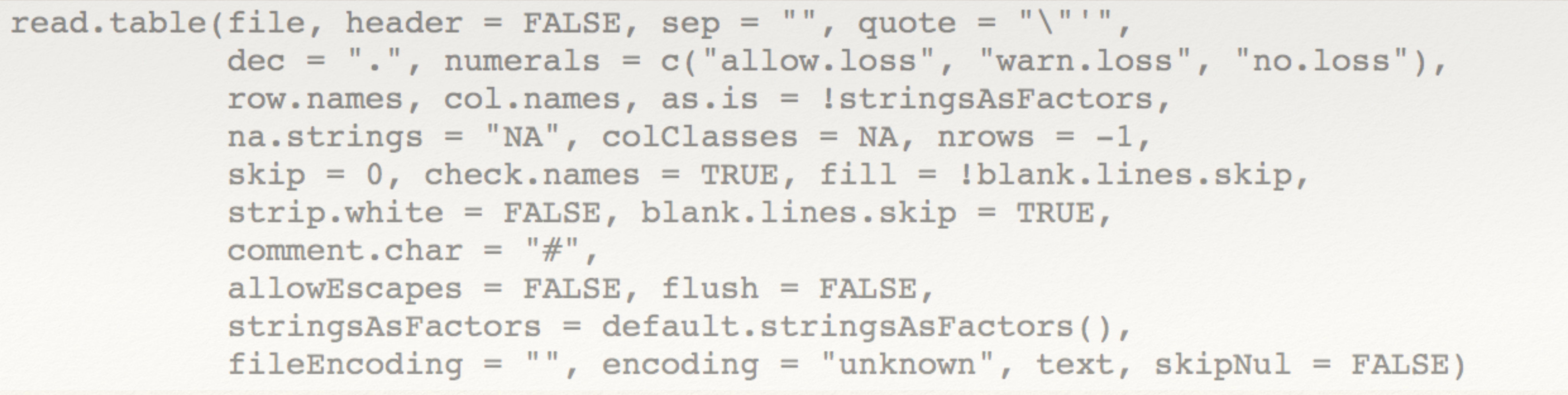


```
seq(from = 1, to = 1, by = ((to - from)/(length.out - 1)),
    length.out = NULL, along.with = NULL, ...)
```

- d=seq(to=10,by=4)

# Read – Key Parameters

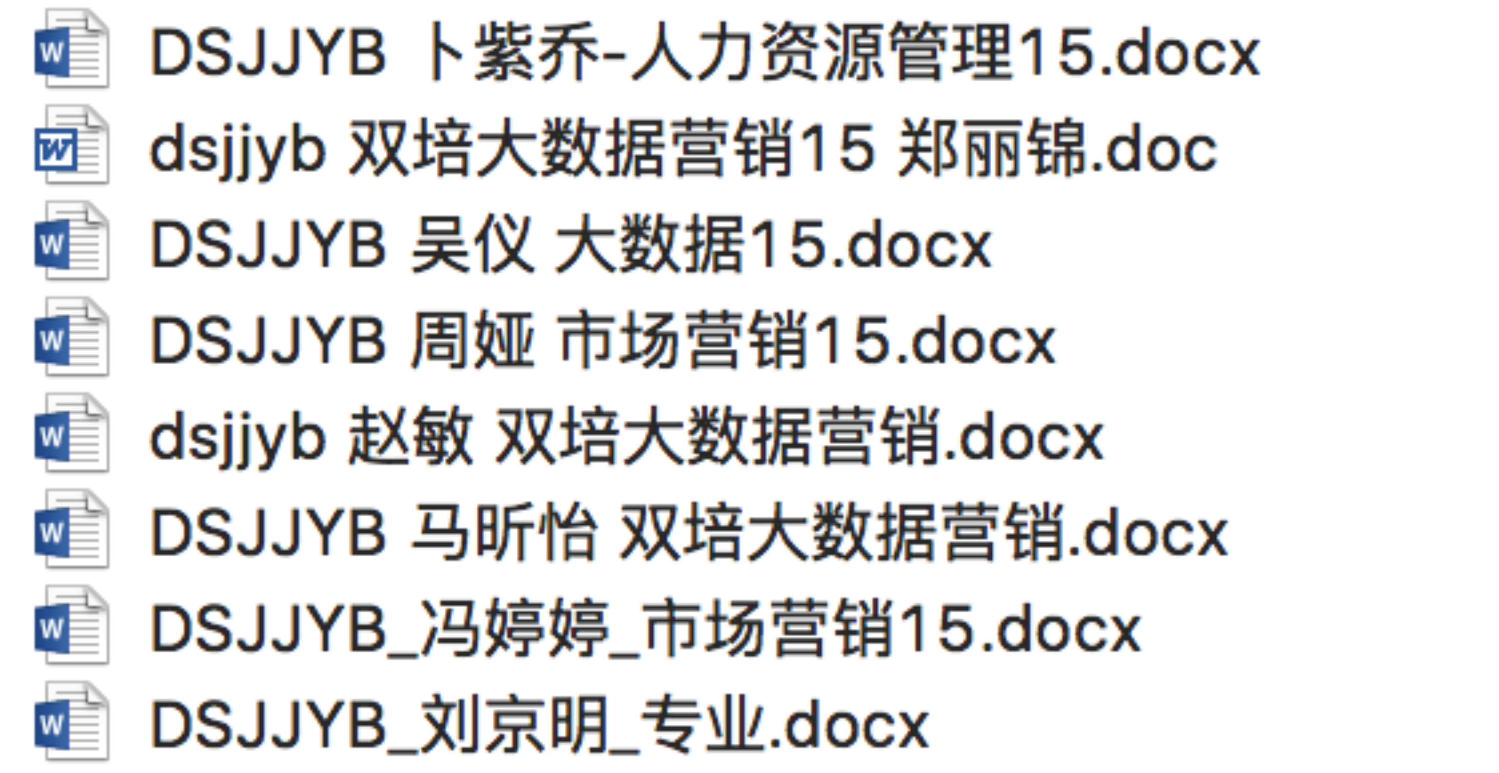
- file: file name (notice the directory)
- header: first row (column names)
- sep: separator (“ ”, “,”, “;”,”\t”)
- col.names & colClasses



```
read.table(file, header = FALSE, sep = "", quote = "\"'",
           dec = ".", numerals = c("allow.loss", "warn.loss", "no.loss"),
           row.names, col.names, as.is = !stringsAsFactors,
           na.strings = "NA", colClasses = NA, nrows = -1,
           skip = 0, check.names = TRUE, fill = !blank.lines.skip,
           strip.white = FALSE, blank.lines.skip = TRUE,
           comment.char = "#",
           allowEscapes = FALSE, flush = FALSE,
           stringsAsFactors = default.stringsAsFactors(),
           fileEncoding = "", encoding = "unknown", text, skipNul = FALSE)
```

# Example

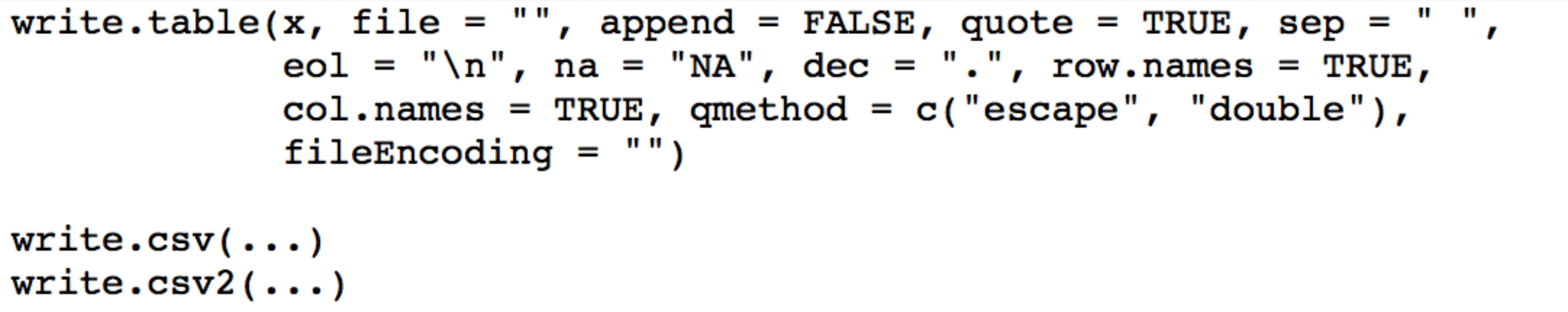
- Search the homework folder
- Find who submit the homework
- Whether they in a right way



DSJJYB 卜紫乔-人力资源管理15.docx
dsjjyb 双培大数据营销15 郑丽锦.doc
DSJJYB 吴仪 大数据15.docx
DSJJYB 周娅 市场营销15.docx
dsjjyb 赵敏 双培大数据营销.docx
DSJJYB 马昕怡 双培大数据营销.docx
DSJJYB_冯婷婷_市场营销15.docx
DSJJYB_刘京明_专业.docx

# Write

- txt, csv
- x: (output data, notice the format)
- file: file name of the output data (notice the directory)
- row.names: whether keep names of rows
- col.names: whether keep names of columns

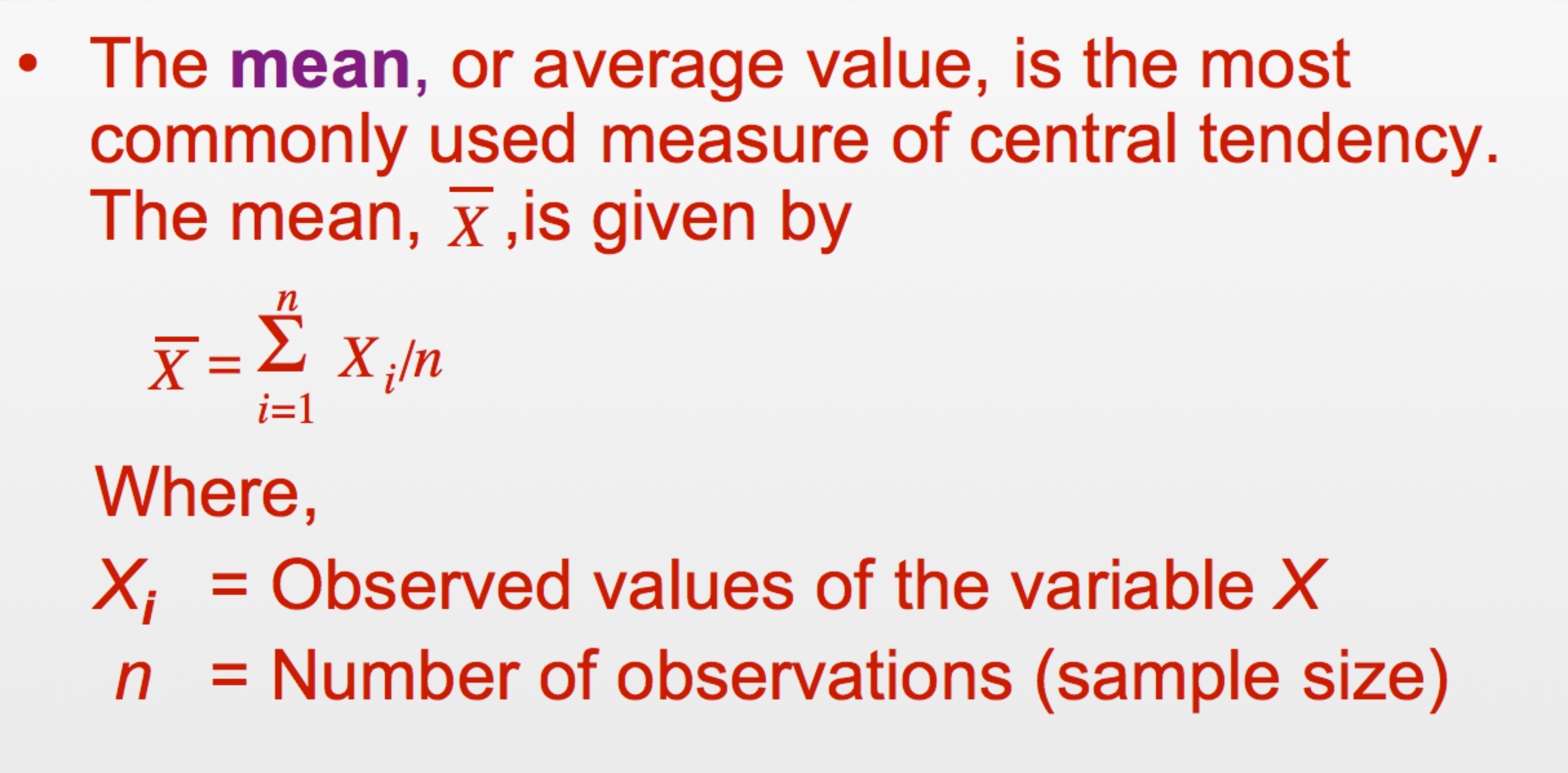


```
write.table(x, file = "", append = FALSE, quote = TRUE, sep = " ",
            eol = "\n", na = "NA", dec = ".", row.names = TRUE,
            col.names = TRUE, qmethod = c("escape", "double"),
            fileEncoding = "")

write.csv(...)
write.csv2(...)
```

# Outline

- Review of Last Class
- Homework 1 Explanation
- Read and Write Data files
- Brief Description Analysis
- Conclusions

# Frequency Distribution



- The **mean**, or average value, is the most commonly used measure of central tendency. The mean, $\overline{X}$ ,is given by

$$\overline{X} = \sum_{i=1}^{n} X_i/n$$

Where,

$X_i$ = Observed values of the variable $X$

$n$ = Number of observations (sample size)

Refer to Ji's Class

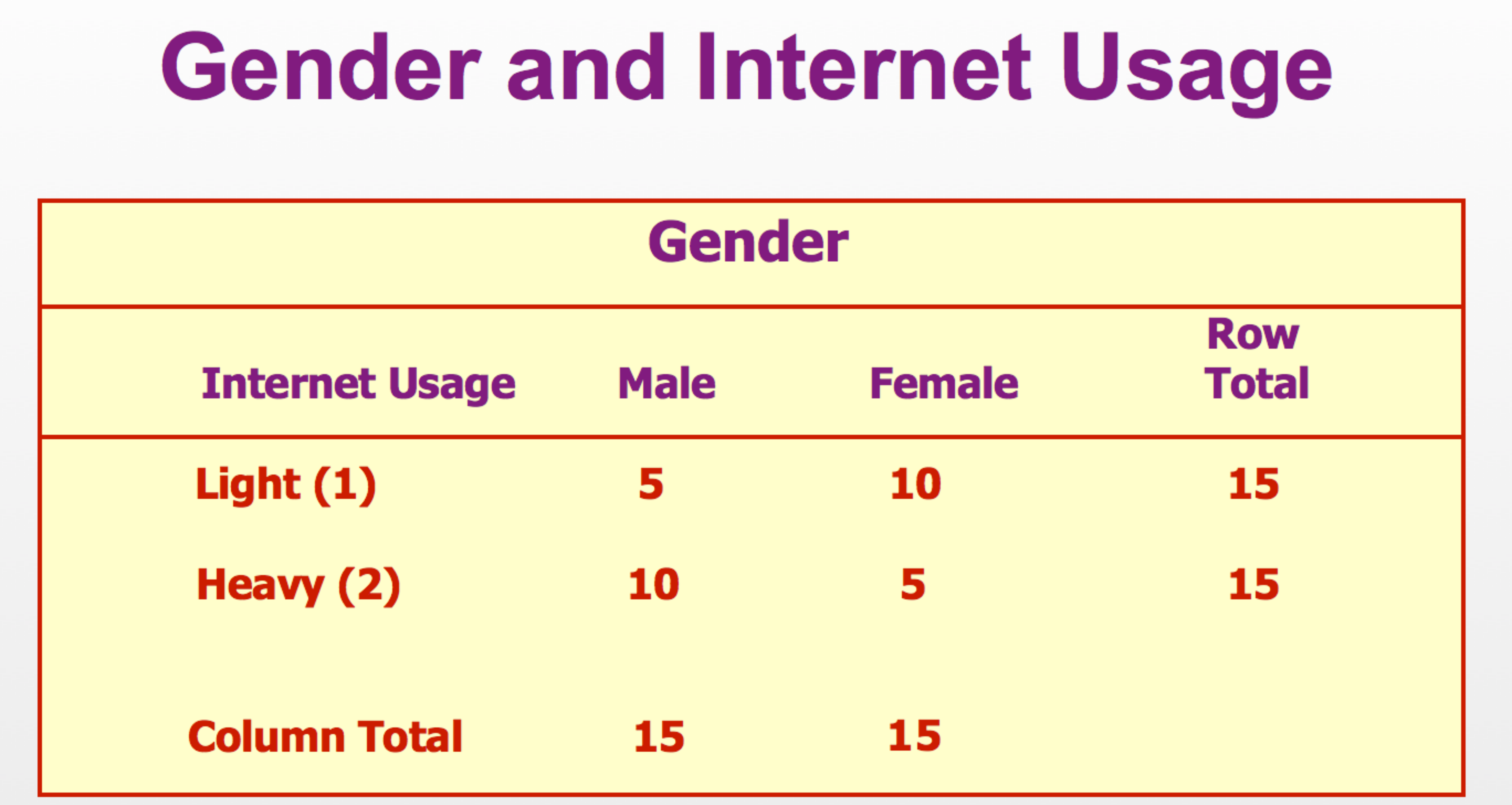
# Frequency Distribution

- Min
- Max
- Median
- Mean

# Cross-Tabulation

- While a frequency distribution describes one variable at a time, a **cross-tabulation** describes two or more variables simultaneously.

- Cross-tabulation results in tables that reflect the joint distribution of two or more variables with a limited number of categories or distinct values.



# Gender and Internet Usage

| | Gender | | |
|---|---|---|---|
| Internet Usage | Male | Female | Row Total |
| Light (1) | 5 | 10 | 15 |
| Heavy (2) | 10 | 5 | 15 |
| Column Total | 15 | 15 | |

# Outline

- Review of Last Class
- Homework 1 Explanation
- Read and Write Data files
- Brief Description Analysis
- Conclusions

# Conclusions

- Review of Class 1
- Supplements for Class 1
- What's a good homework
- Read and Write data file
- Description Analysis

# Next Class

- Data Visualization
- More Operations in Data Frame
    - Add
    - Delete
    - Revise
    - Search
- In Class Test