R 語言與資料工程

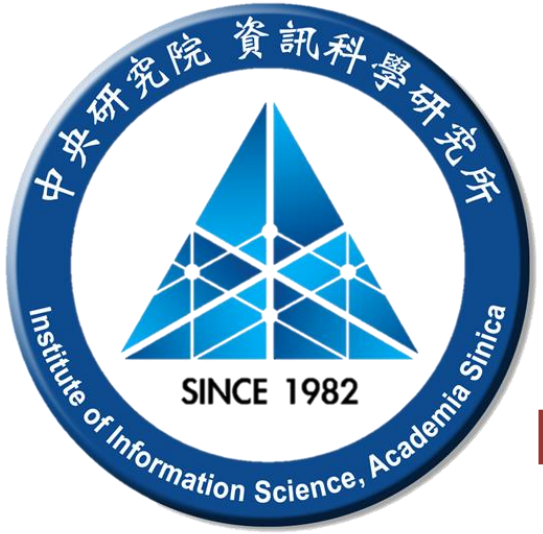
許懷中 Hwai-Jung Hsu
中央研究院
資訊科學研究所

*The material is powered by **Wush Wu** (吳齊軒).

我的背景

雲端計算、軟體工程

+



陳昇瑋研究員

資料洞察實驗室

我的資料科學相關經驗

線上遊戲玩家
黏著度分析

K-15 學生
快速程度評量

虛寶銷售
預測與成因分析

中華職棒票房
預測與成因分析

企業、法人與政府
資料科學人才培訓

真正的重點！

R 語言與資料工程

課程與 R 環境安裝

安裝 R

請至 http://cran.csie.ntu.edu.tw/ 下載 3.2 版以上的 R

For Windows Users

https://www.youtube.com/watch?v=FsOHPGUIDZU
注意影片下載的是 3.0.2版，請安裝最新版 (3.2版以上)

For Mac Users

https://www.youtube.com/watch?v=72MYRBNo5Bk
感謝中華 R 軟體學會的李明昌老師提供影片

For Ubuntu Users

請參照下列說明
http://cran.csie.ntu.edu.tw/bin/linux/ubuntu/README.html

安裝 R Studio

R Studio 為 R 的 IDE 環境

圖形化介面，完整支援 R 的編輯、繪圖以及文件說明

具備「自動完成」功能

在 Windows 下支援 UTF-8 的檔案編碼

請到 R Studio 官方網站

https://www.rstudio.com/products/RStudio/

下載並安裝 R Studio Desktop Open Source Edition

安裝課程

請在 Rstudio 環境中執行

```
> source("http://hjhsu.github.io/r_course/init-swirl.R")
> library(swirl)
> swirl()
```

```
| Please choose a course, or type 0 to exit swirl.

1: DSC2016-R
2: Take me to the swirl course repository!

Selection: 1

| Please choose a lesson, or type 0 to return to course menu.

 1: 01-DataObservation-01-SingleVariable
 2: 02-DataObservation-02-MultiVariables
 3: 03-RDataEngineer-01-Loading n Parsing
 4: 04-RDataEngineer-02-DataManipulation
 5: 05-RDataEngineer-03-Join
 6: 06-RDataMining-01-Association-Rule
 7: 07-RDataMining-02-Clustering
 8: 08-RDataMining-03-Classification
 9: X1-Optional-01-ggplot2
10: X2-Challenge-01-ChineseEncoding
11: X3-Challenge-02-PirateVisualization
12: X4-RDataMining-04-Text-Mining

Selection:
```

R 語言與資料工程

R 語言的資料結構

資料的種類

名目資料 (Nomnial)
順序資料 (Ordinal)
區間資料 (Interval)
比值資料 (Ratio)

名目資料 (Nominal)

Name

畢業學校：交大、台大、清大……

車輛廠牌：Toyota, VW, Benz

分類

性別：男、女、其他

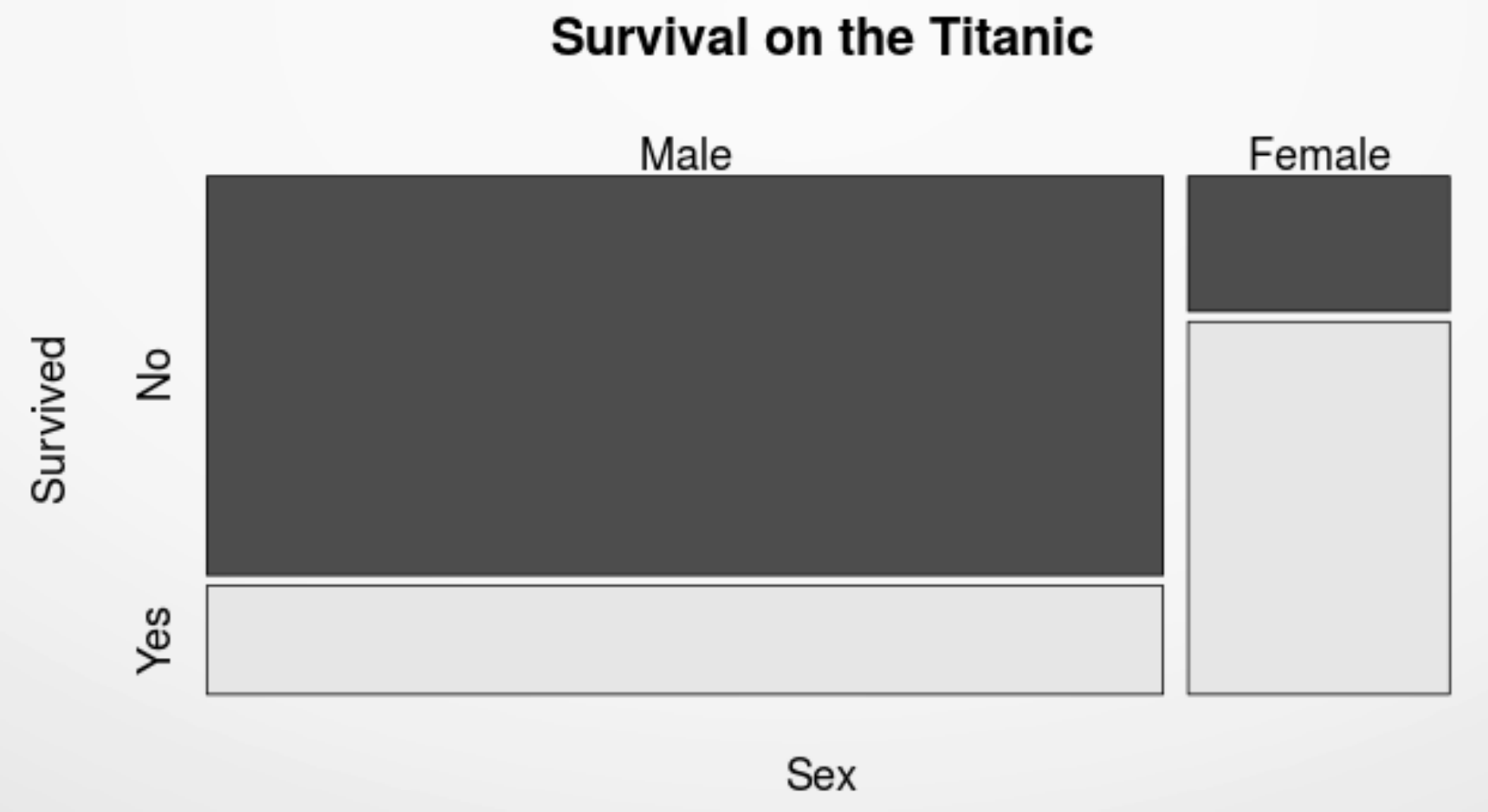
產業：金融、資訊、製造……

屬性上的有無

年收入>100萬：是、否

資料的先後沒有意義

名目資料 (cont.)



順序資料 (Ordinal)

先後的意義

硬度表：(1) 滑石, …(7) 石英,…(10) 金剛石

戰績排名：(1) Lamigo (2) 中信 (3) 義大……

有序的名目資料

不確定間隔的意義

區間資料 (Interval)

溫度

　攝氏溫標: 0ºC, 100ºC

時間

　秒、分、時、日、年

各式度量衡

　長度：公尺、公寸、英尺、英吋

　重量：公斤、磅

具有**固定間隔**的順序資料

比值資料 (Ratio)

絕對溫度

克氏溫標: 0°K = -273.15°C

股價

票面 10 元

公司營收

具有**參考點(零點)**的區間資料

比值資料 (cont.)


TWII
[2015-10-01/2016-01-13]
Las: 7824.609863
8800
8600
8400
8200
8000
7800
Volume (millions):
0
1.0
0.5
0.0
-0.5
-1.0
10月 01 2015
10月 19 2015
11月 02 2015
11月 16 2015
11月 30 2015
12月 14 2015
12月 28 2015
1月 11 2016

R 語言與資料工程

利用 R 語言的視覺化工具觀察資料

Let’s Roll

接下來我會帶著各位同學進行下面的實作課程

1: 01-DataObservation-01-SingleVariable

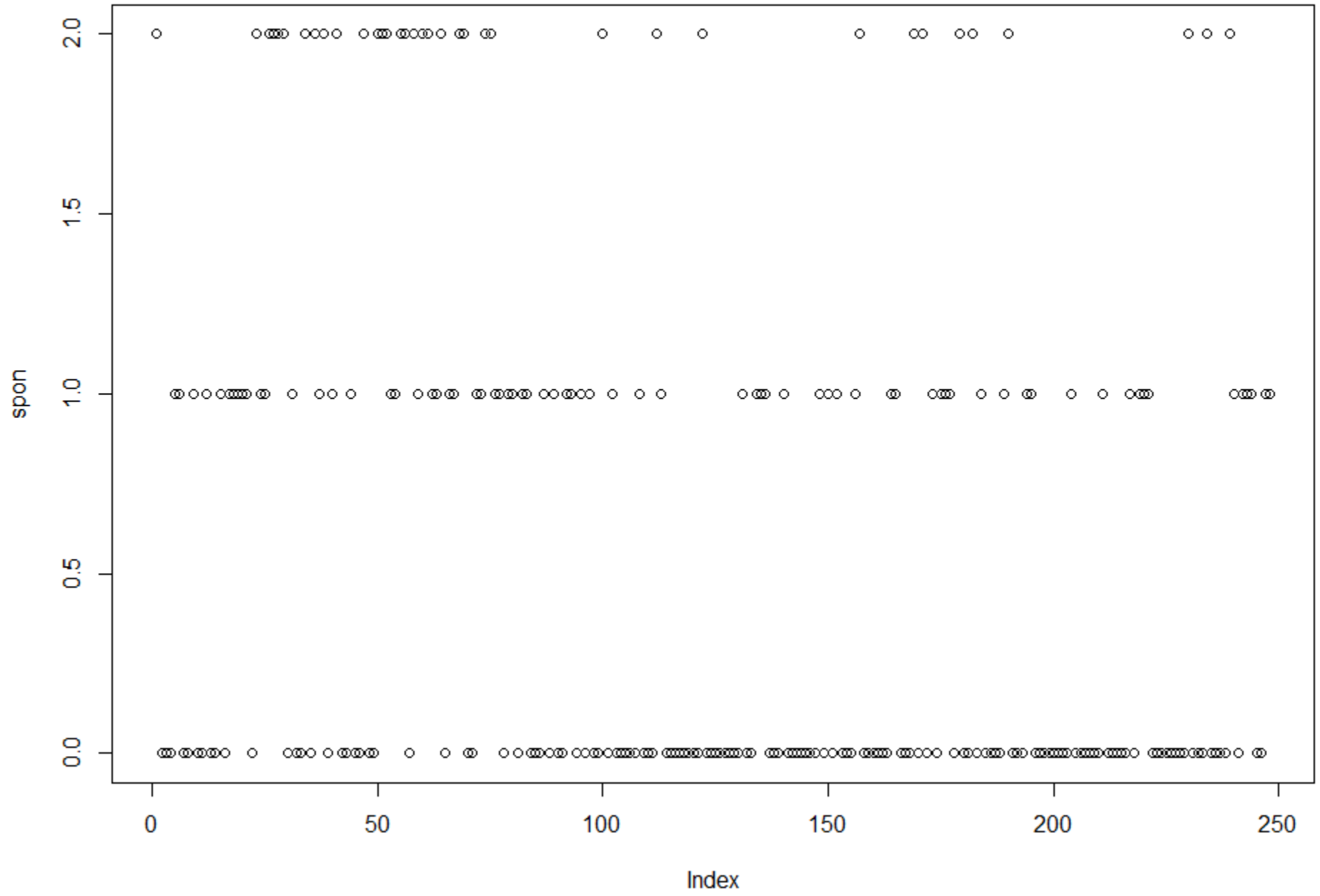
請各位同學搭配講課的進度，操作 swirl 課程

Infert 資料集

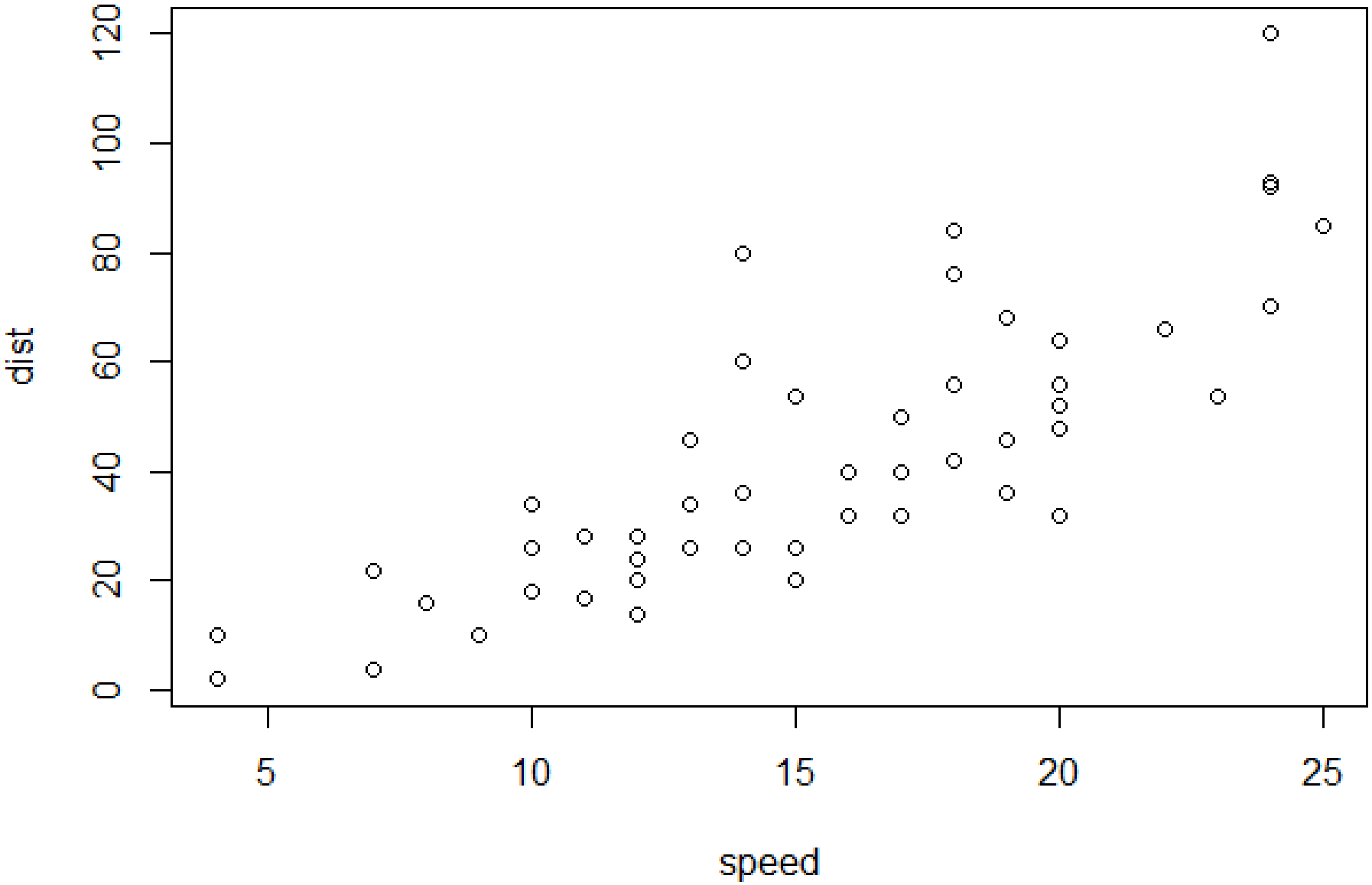
Infertility after Spontaneous and Induced Abortion

Education	教育程度
Age	年齡
Paruty	生育子女數
Induced	人工流產次數
Case	實驗組或對照組 (是否不孕)
Spontaneous	自然流產次數

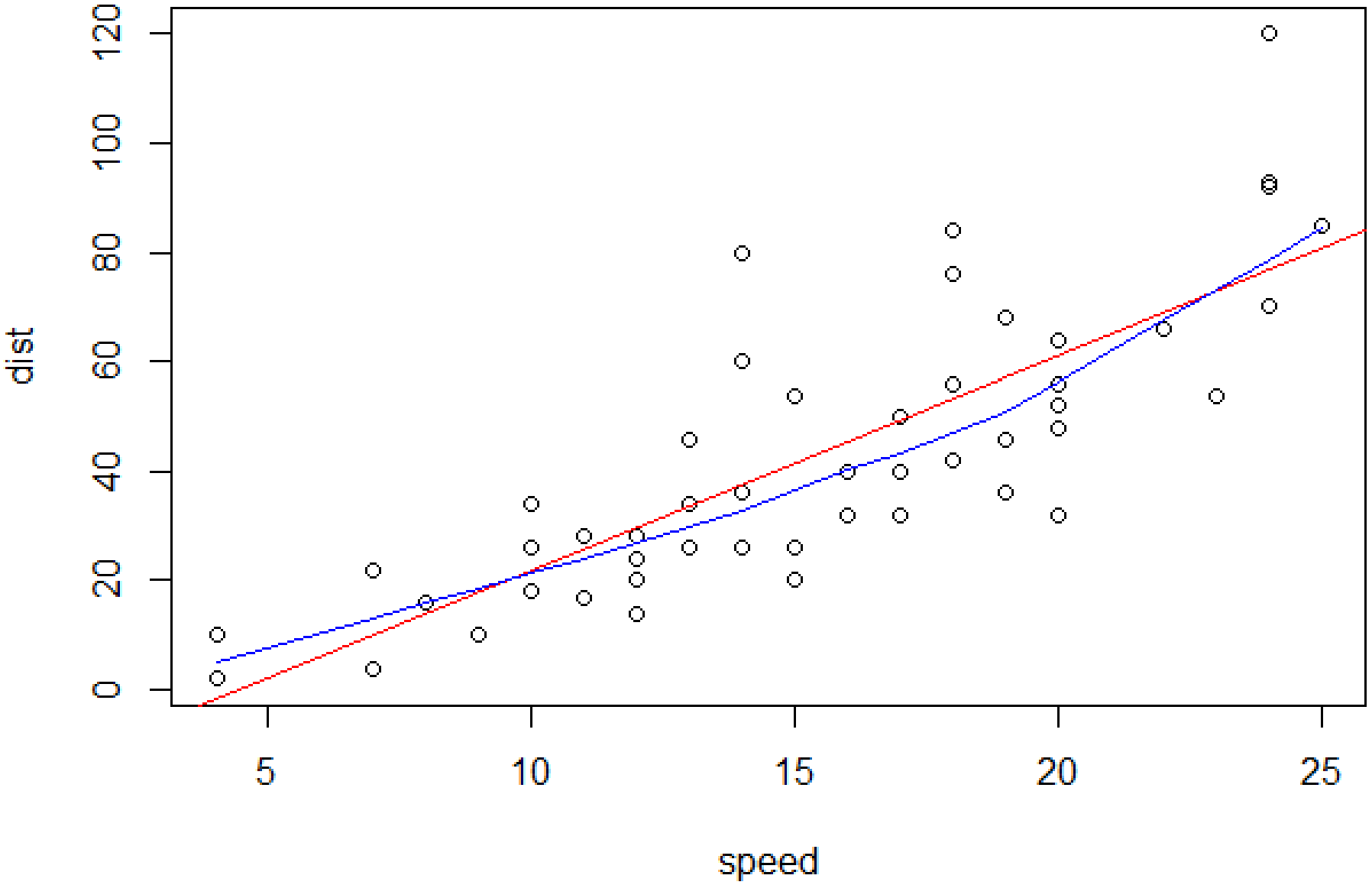
```
> plot(infert$spontaneous)
```



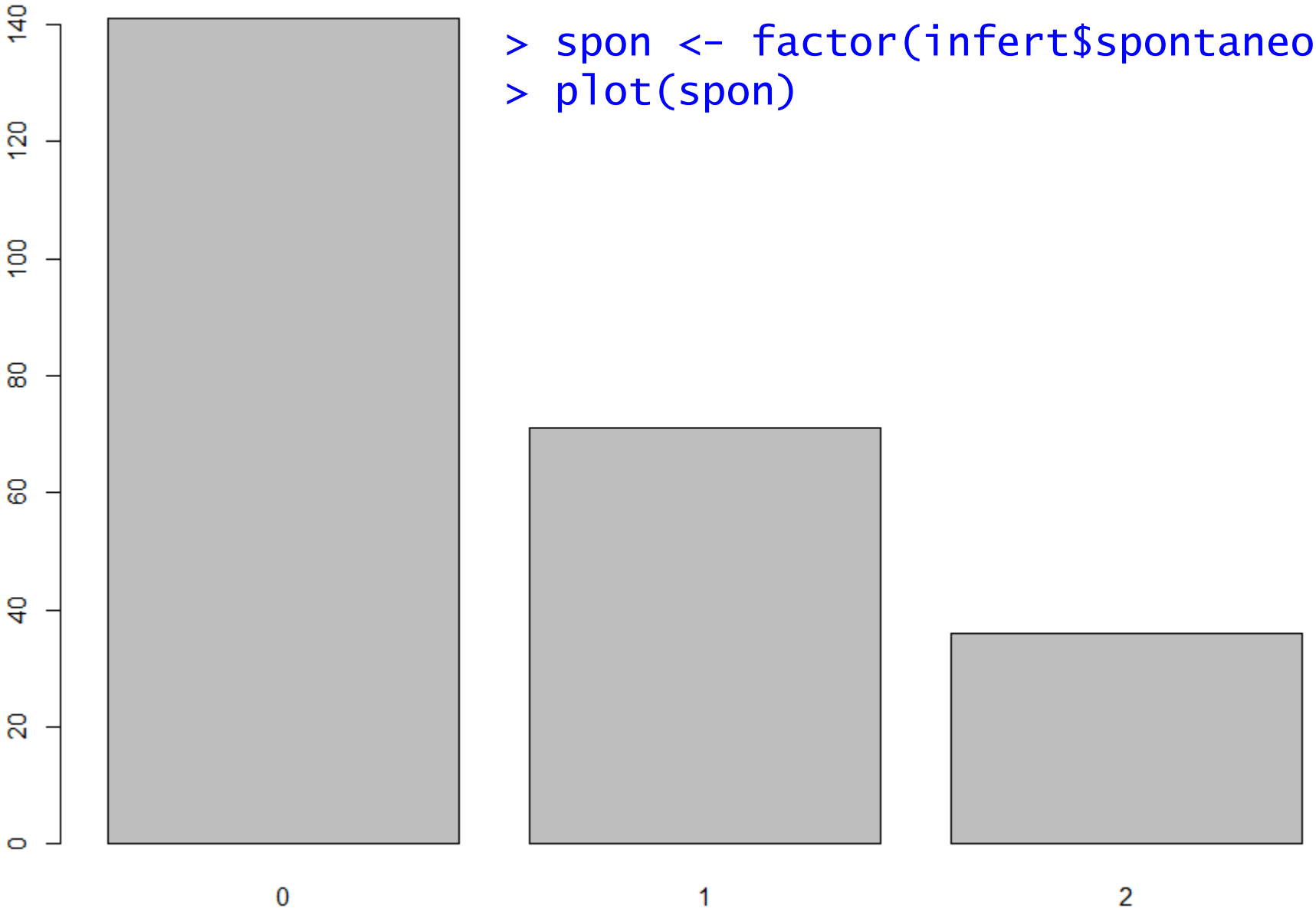
```
> plot(dist~speed, cars)
```



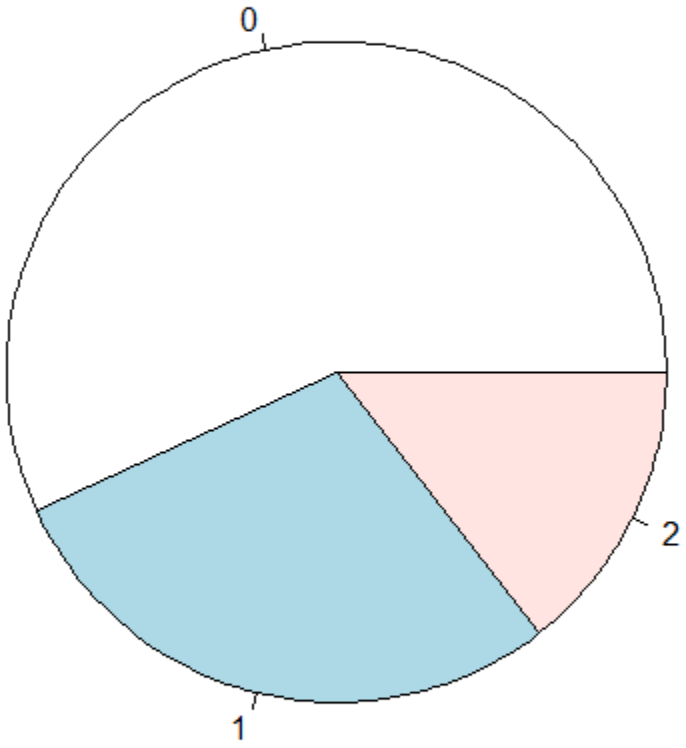
```
> abline(lm(dist~speed, cars), col="red)
> lines(lowess(cars), col="blue")
```



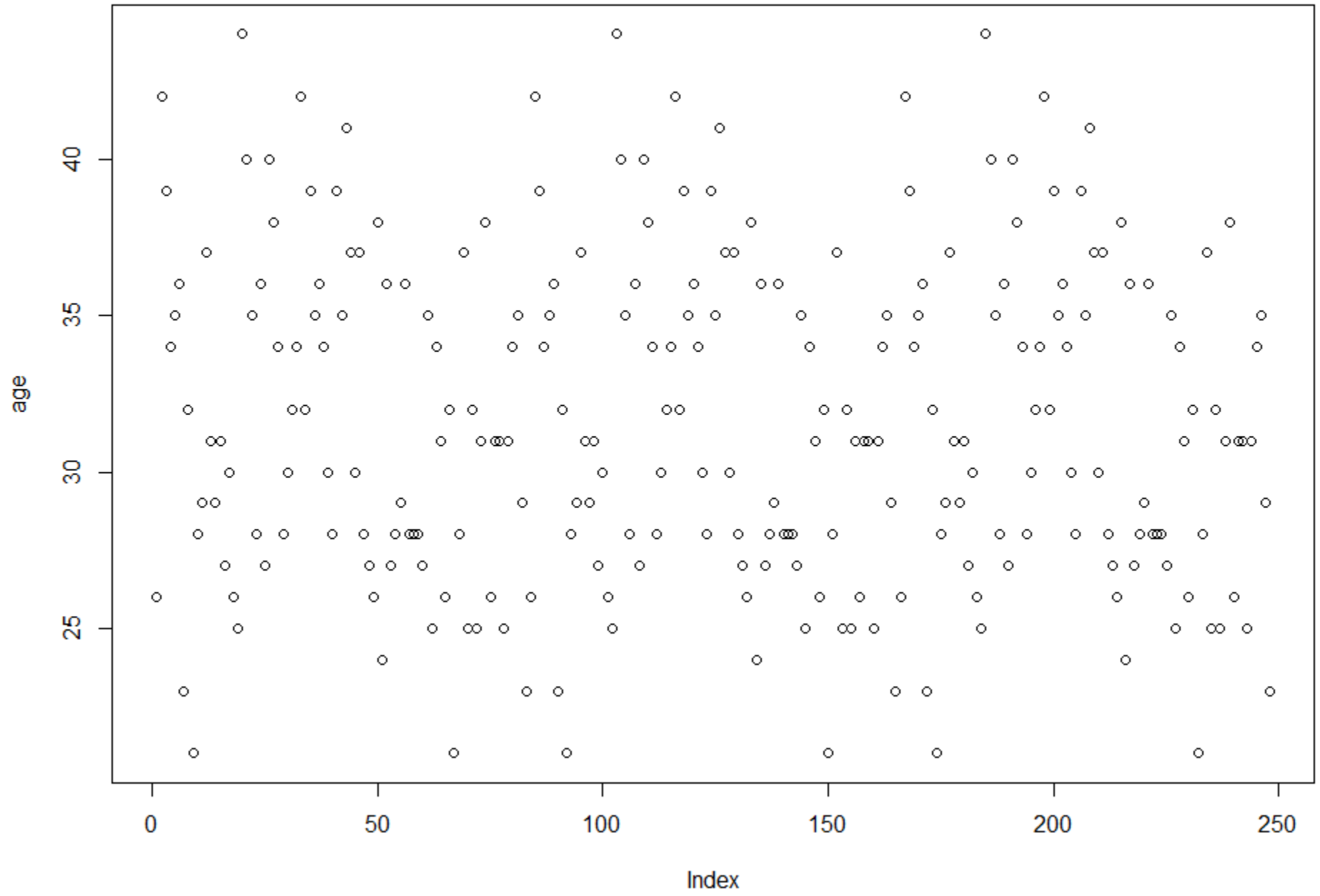
```
> spon <- factor(infert$spontaneous)
> plot(spon)
```



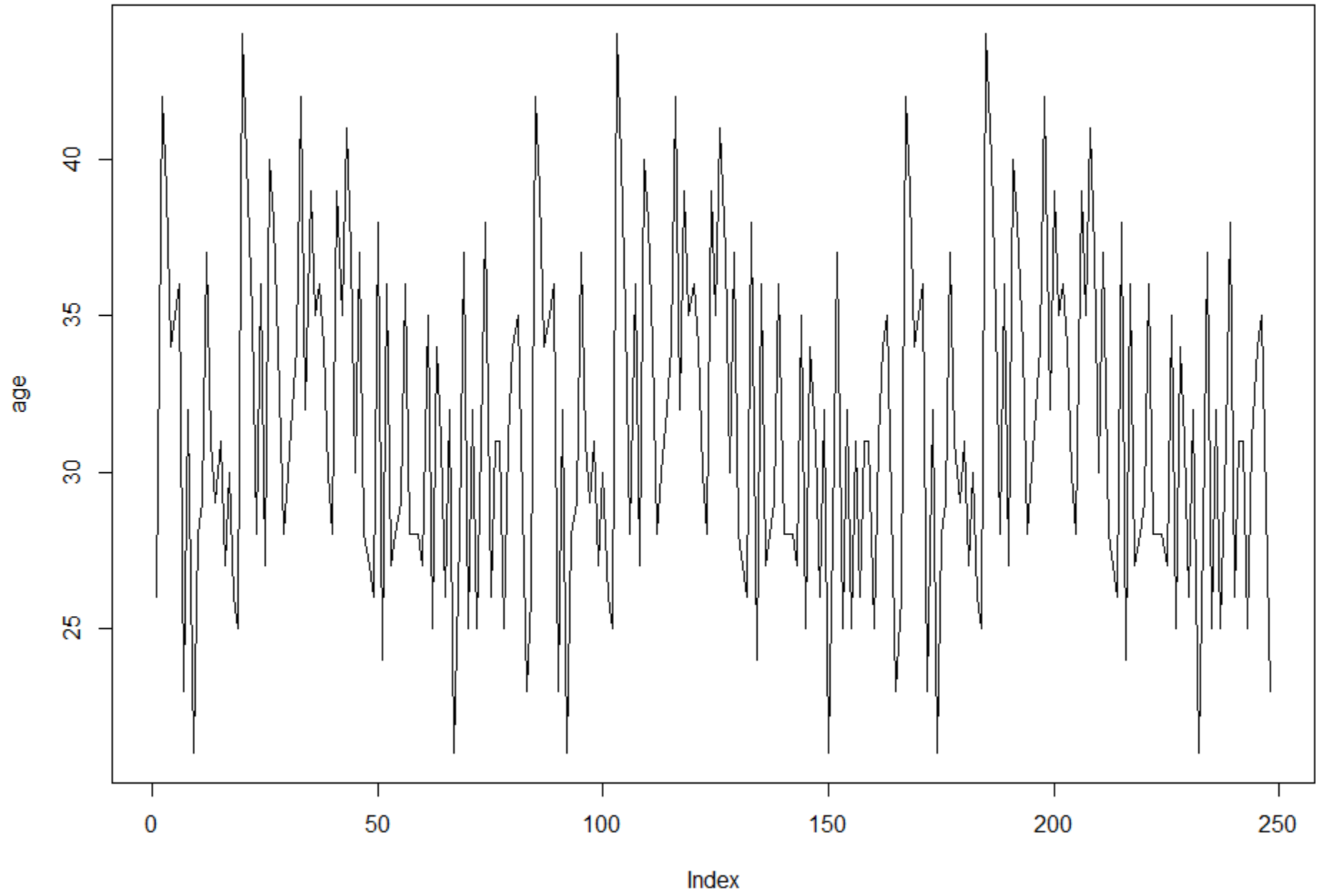
```
> pie(table(spon))
```

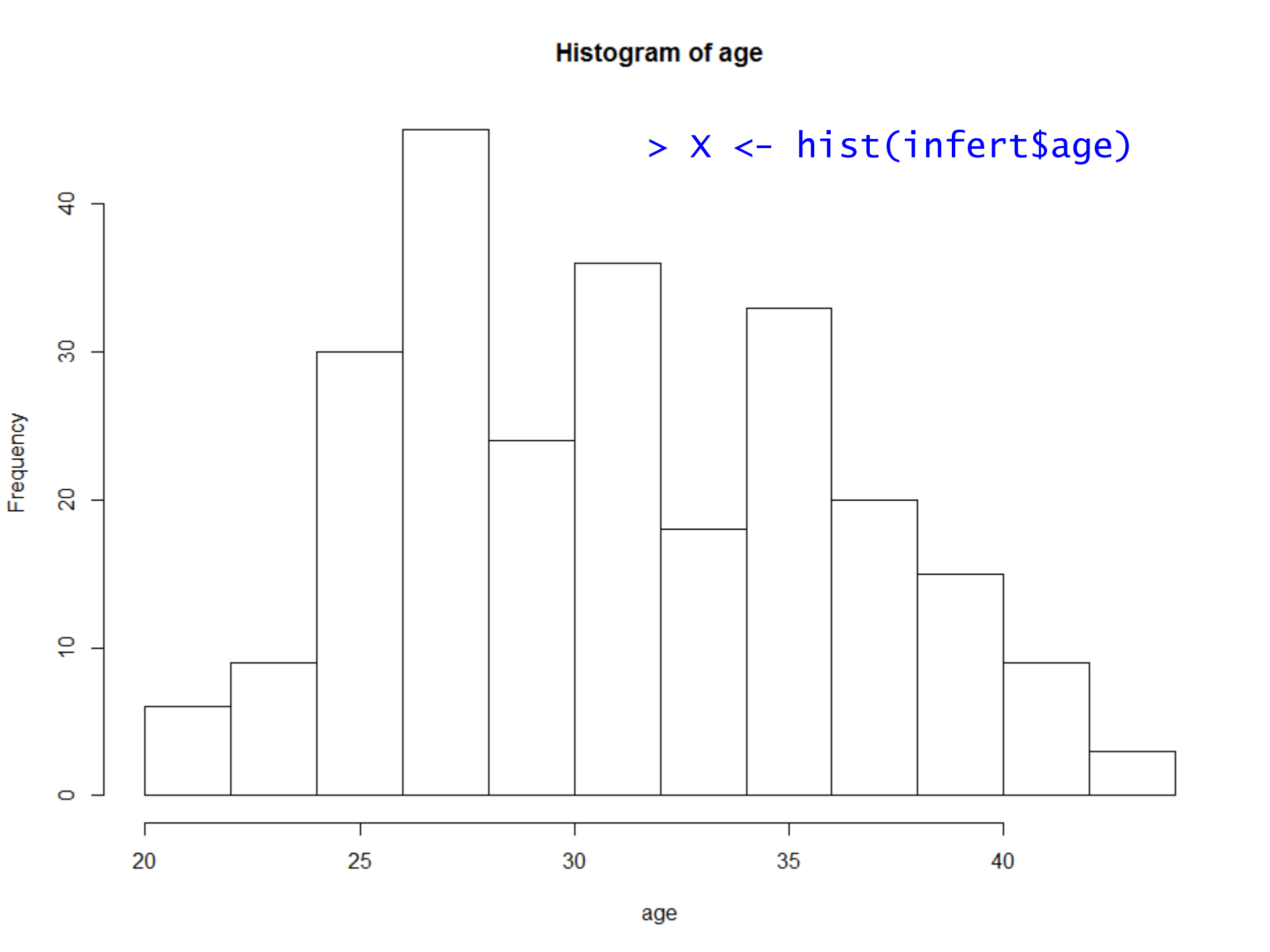


```
> plot(infert$age)
```

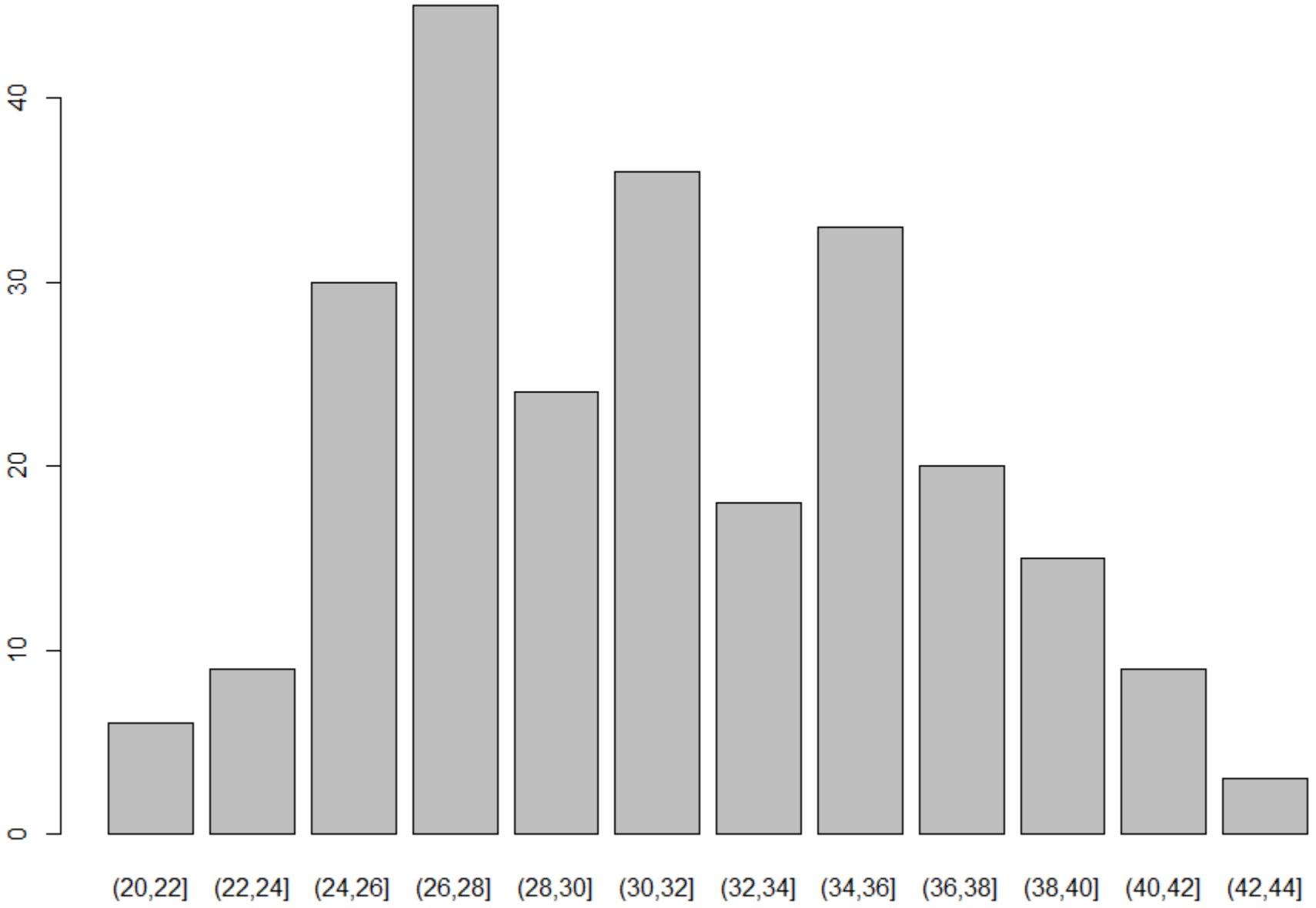


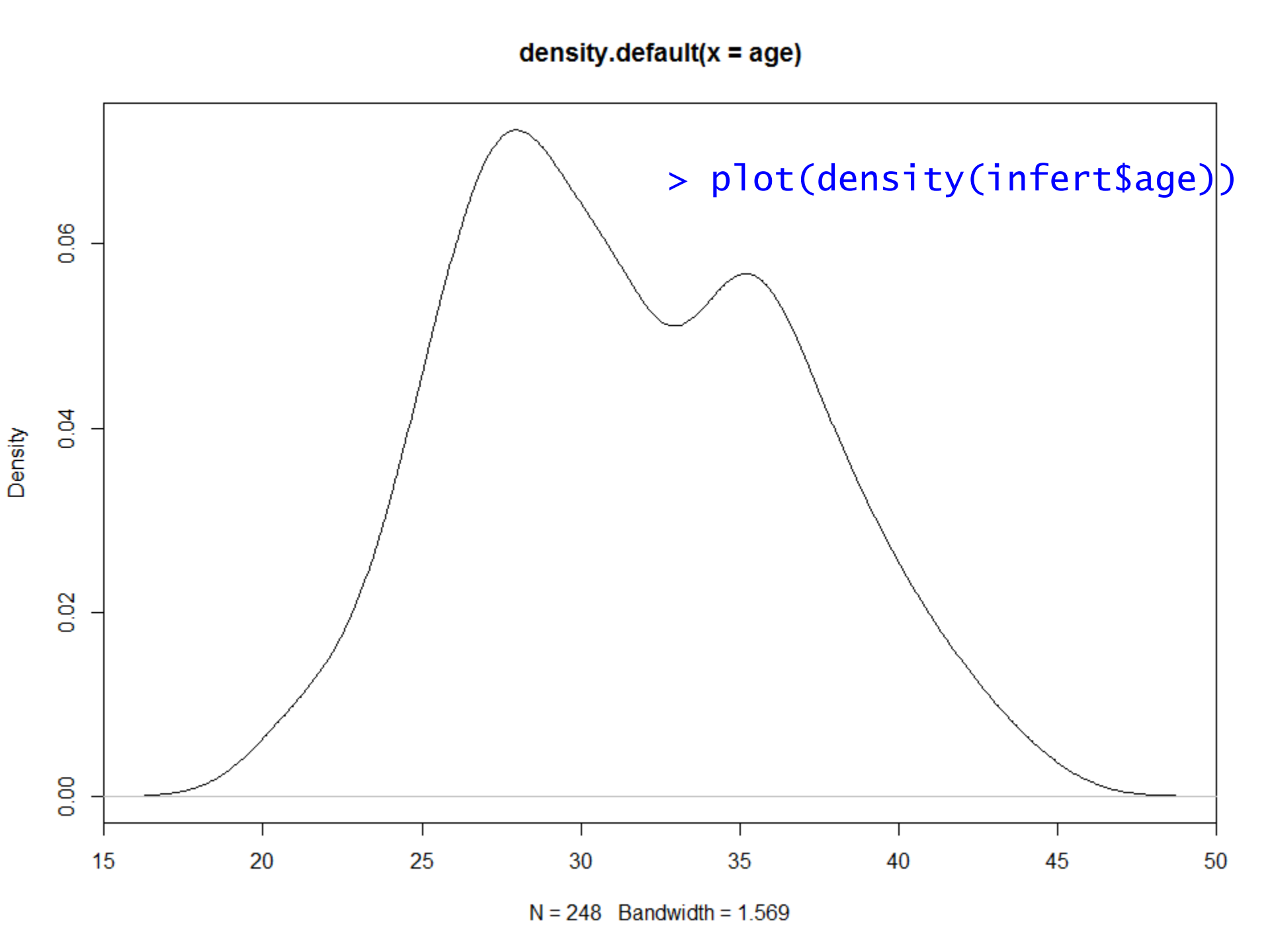
```
> plot(infert$age, type="l")
```

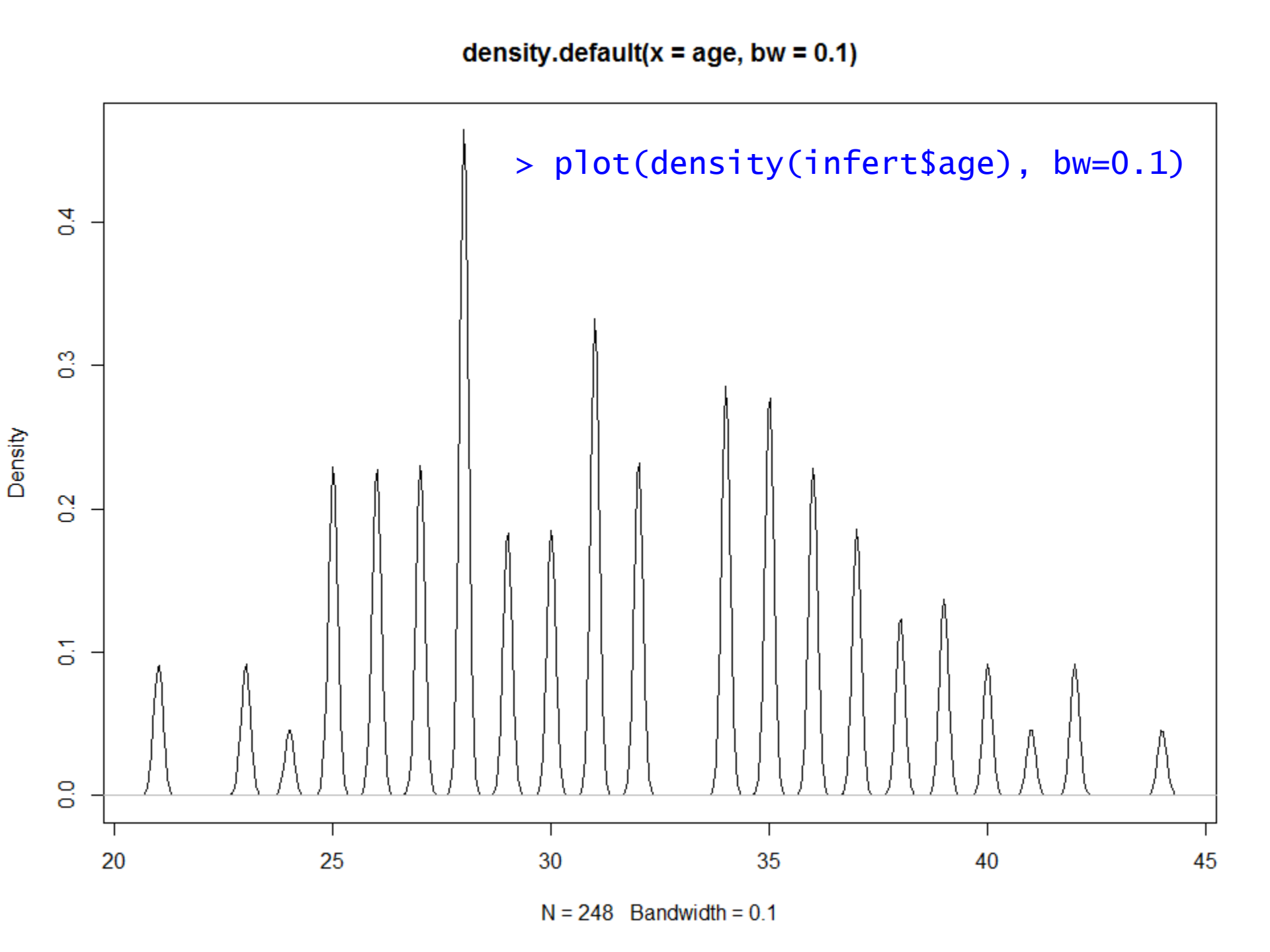


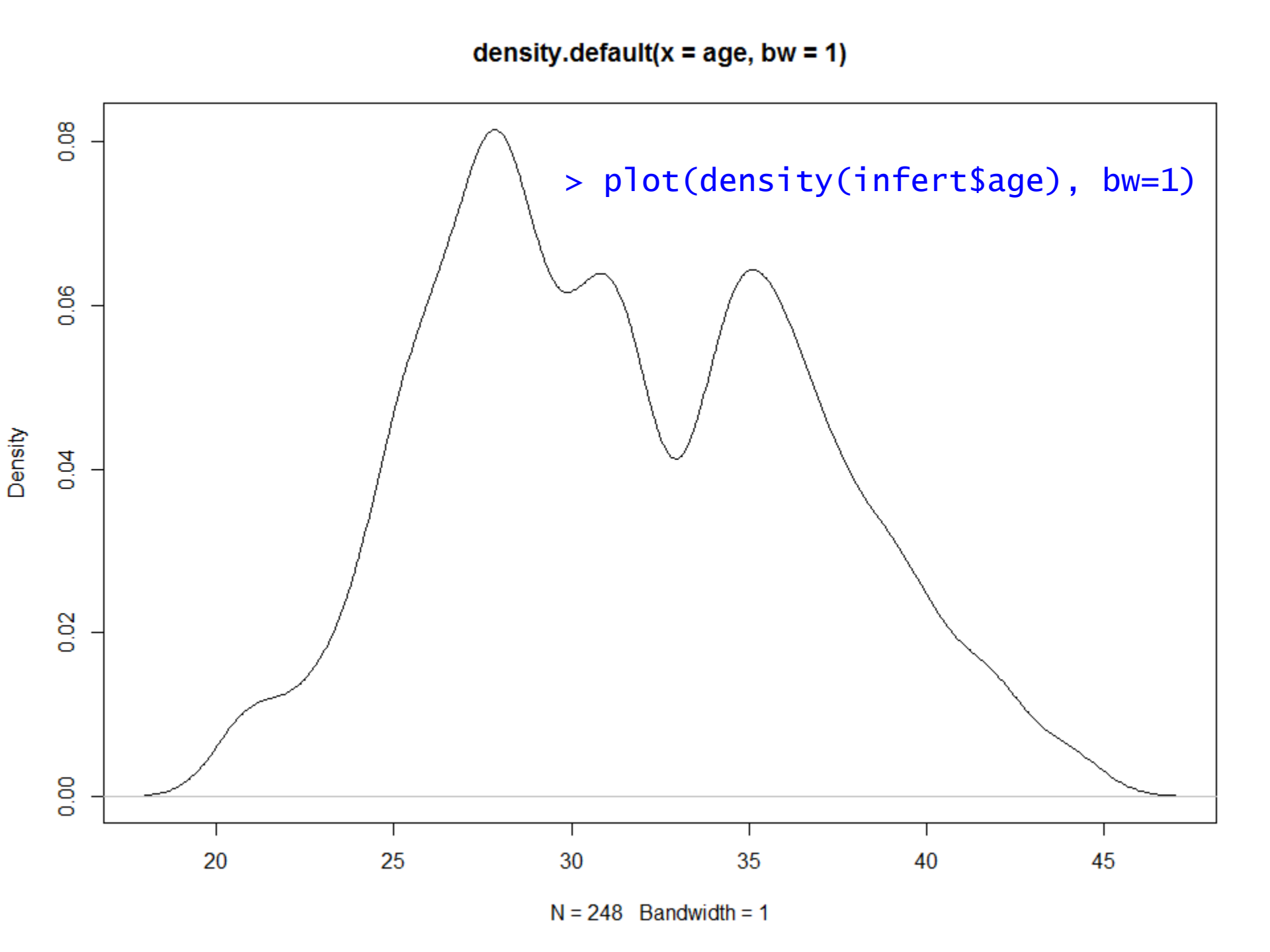

Histogram of age
> X <- hist(infert$age)
Frequency
0
10
20
30
40
20
25
30
35
40
age

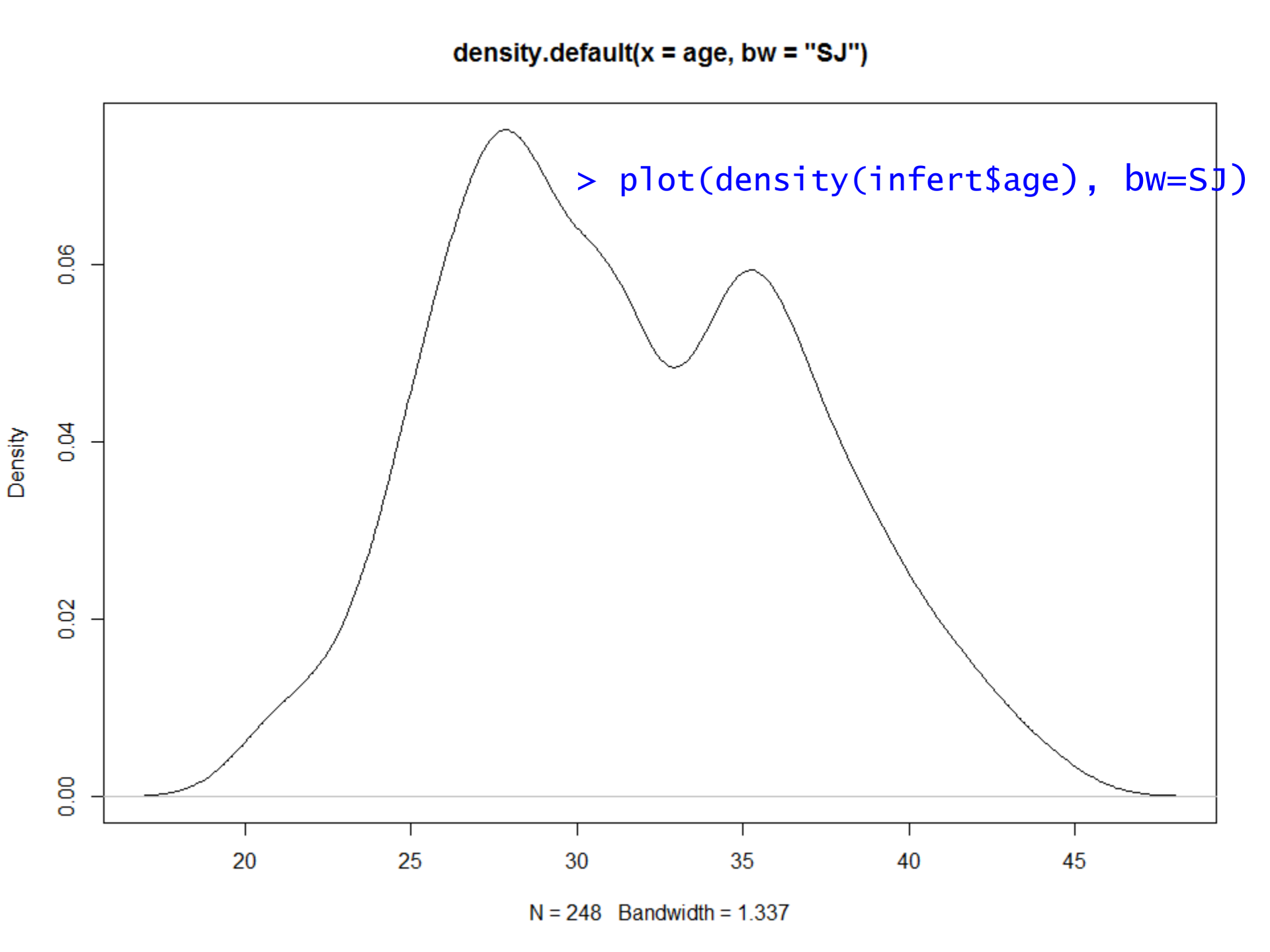
```
> plot(cut(infert$age, breaks = x$breaks))
```



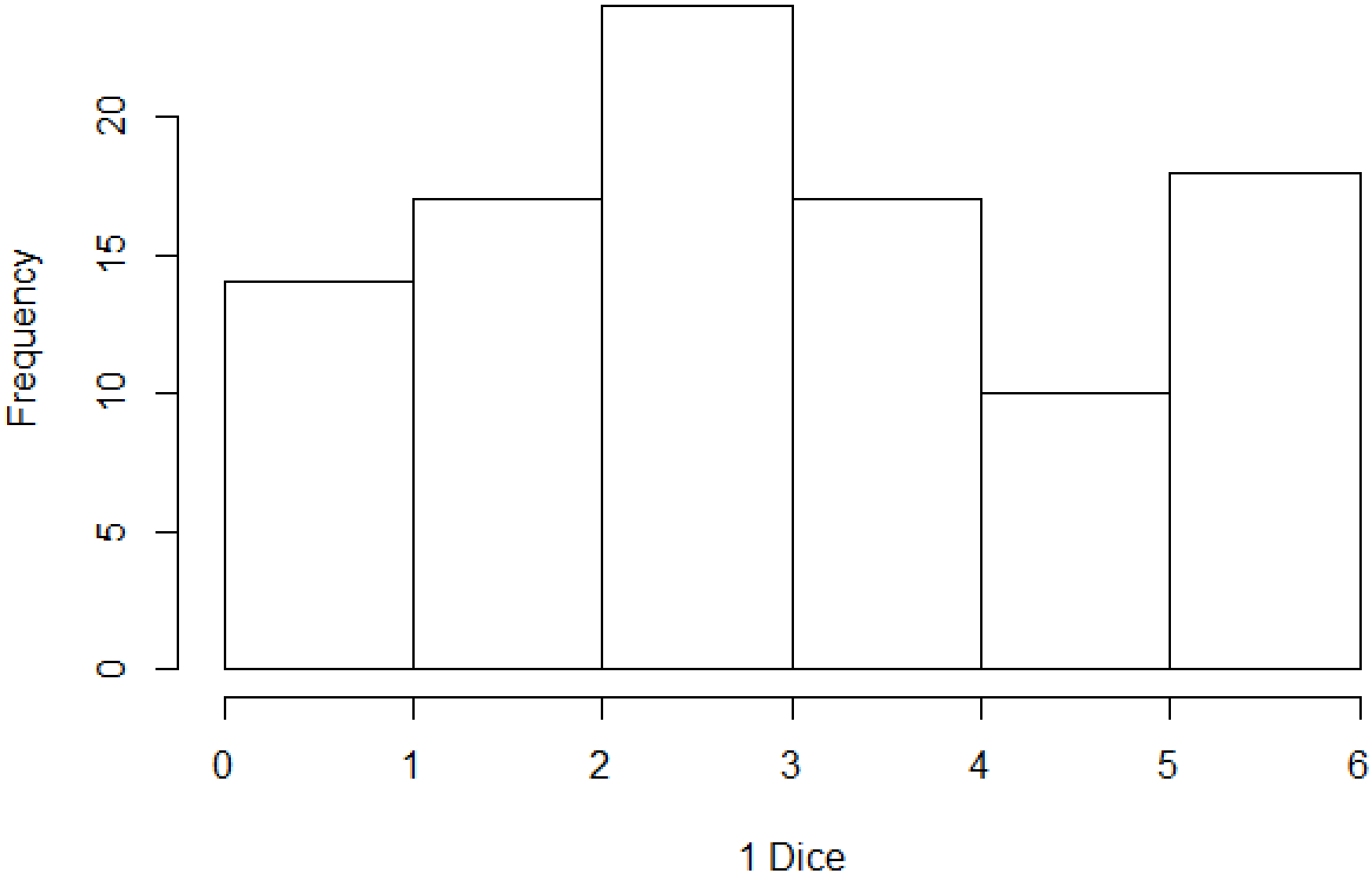

density.default(x = age)
> plot(density(infert$age))
Density
0.00
0.02
0.04
0.06
15
20
25
30
35
40
45
50
N = 248 Bandwidth = 1.569


density.default(x = age, bw = 0.1)
> plot(density(infert$age), bw=0.1)
Density
0.0
0.1
0.2
0.3
0.4
20
25
30
35
40
45
N = 248 Bandwidth = 0.1

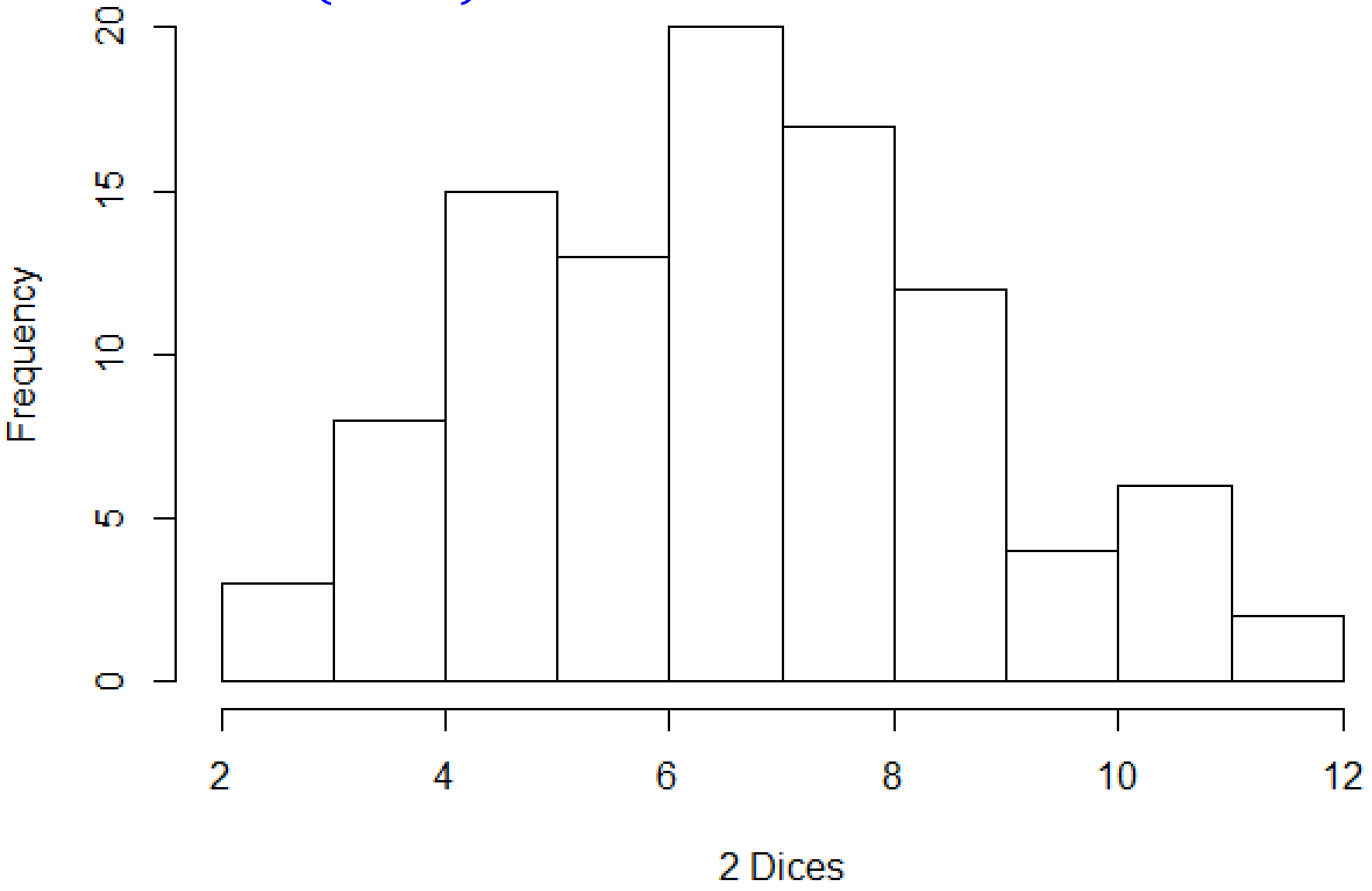

density.default(x = age, bw = 1)
> plot(density(infert$age), bw=1)
Density
0.00
0.02
0.04
0.06
0.08
20
25
30
35
40
45
N = 248 Bandwidth = 1


density.default(x = age, bw = "SJ")
> plot(density(infert$age), bw=SJ)
Density
0.00
0.02
0.04
0.06
20
25
30
35
40
45
N = 248 Bandwidth = 1.337

```
> d1 <- sample(1:6, 100, replace=TRUE)
> hist(d1)
```

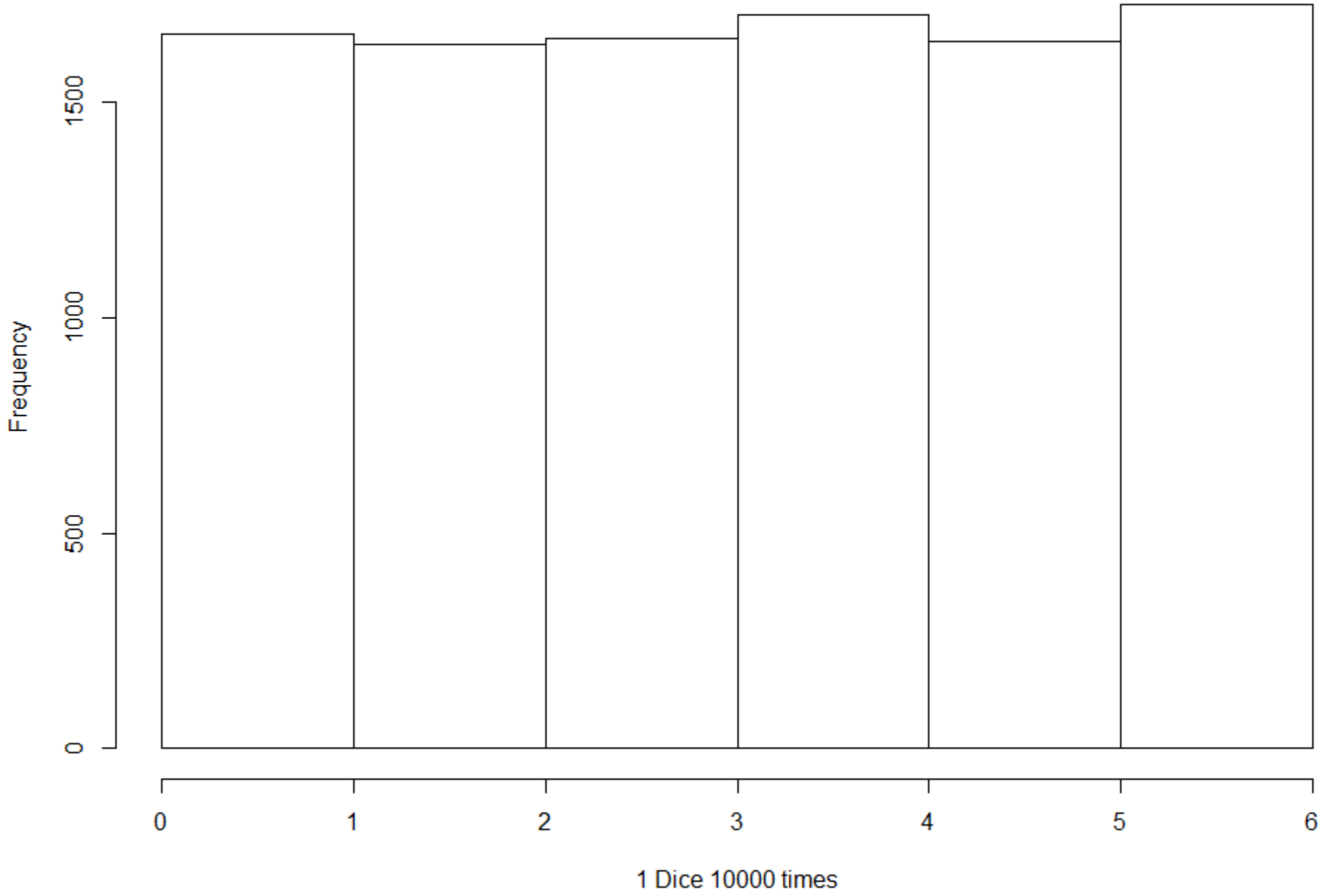


```
> d1 <- sample(1:6, 100, replace=TRUE)
> d1 <- sample(1:6, 100, replace=TRUE)
> hist(d1+d2)
```

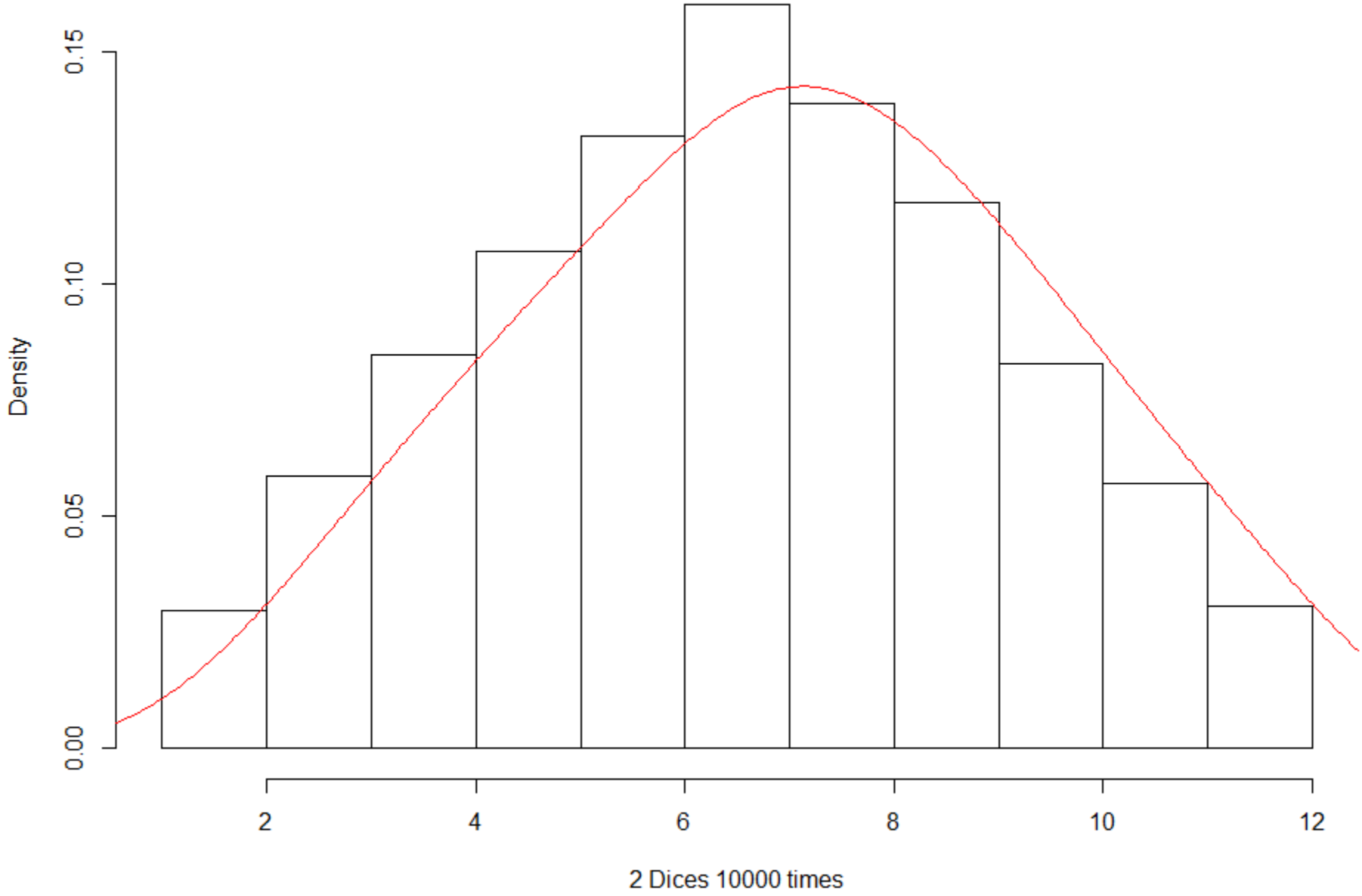


```
> hist(d1+d2, prob=TRUE)
> plot(density(d1+d2, bw="SJ"), col="red")
```

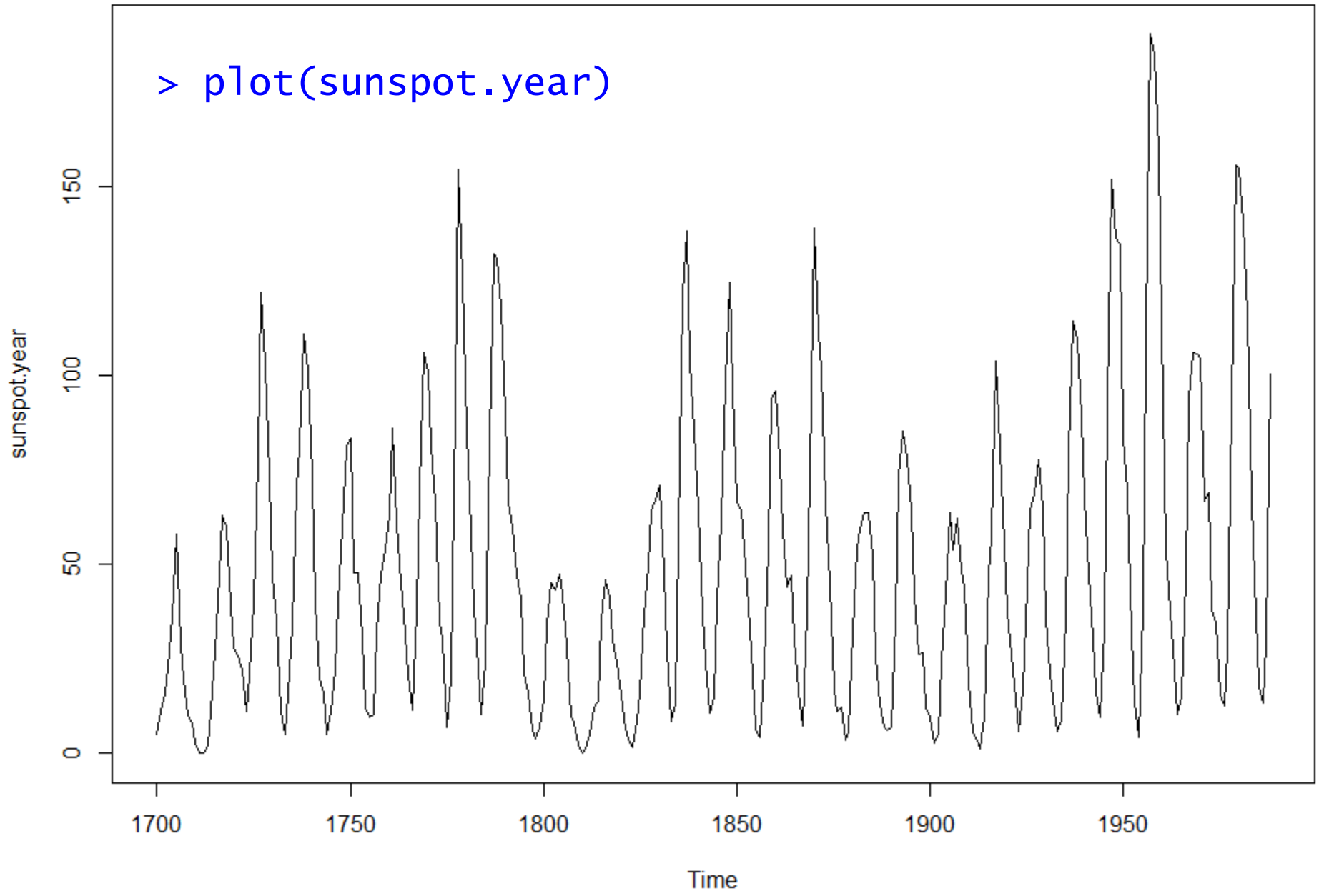
```
> d1 <- sample(1:6, 10000, replace=TRUE)
> hist(d1)
```



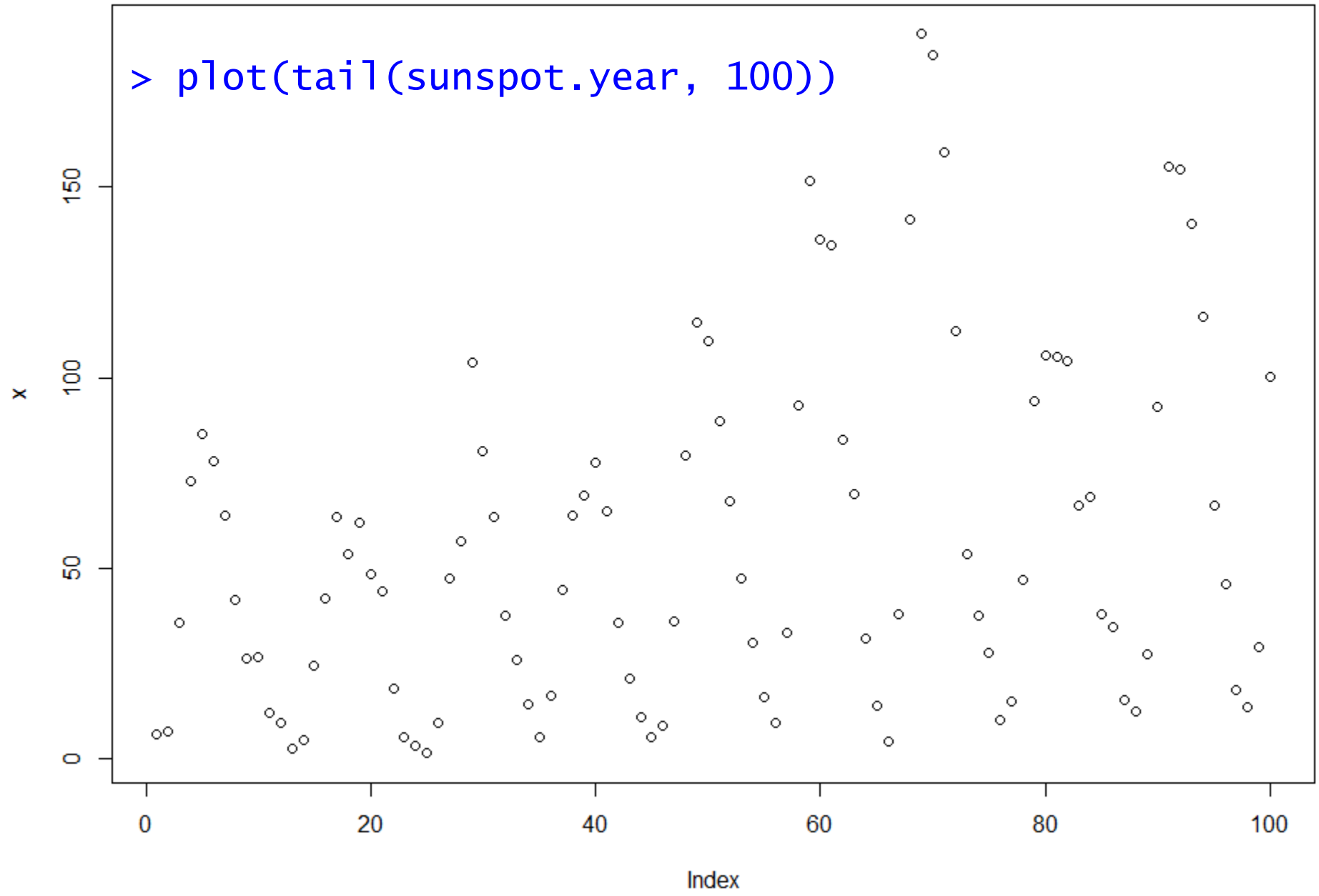
```
> hist(d1+d2, prob=TRUE)
> plot(density(d1+d2, bw=1), col="red")
```

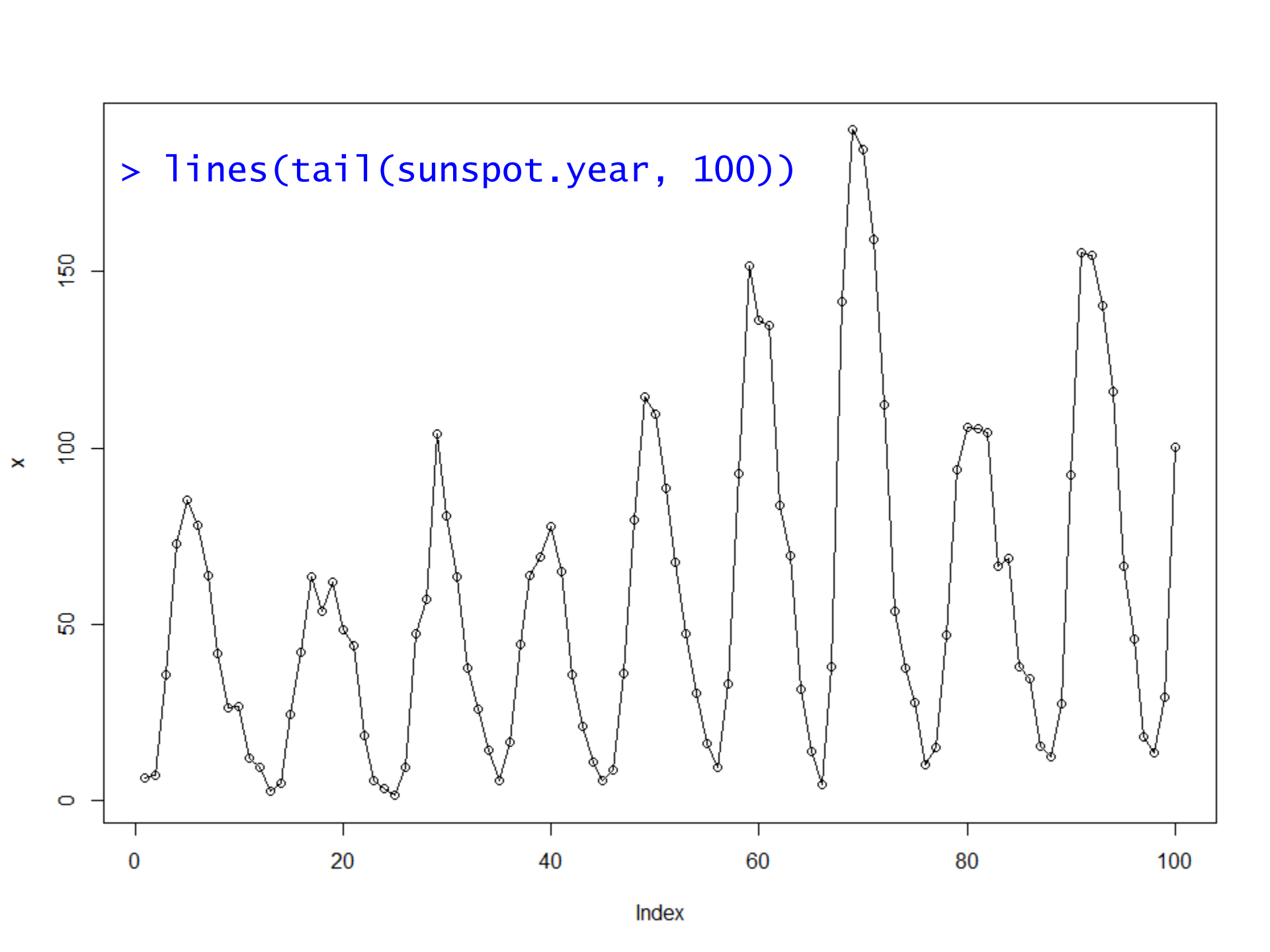


```
> plot(sunspot.year)
```

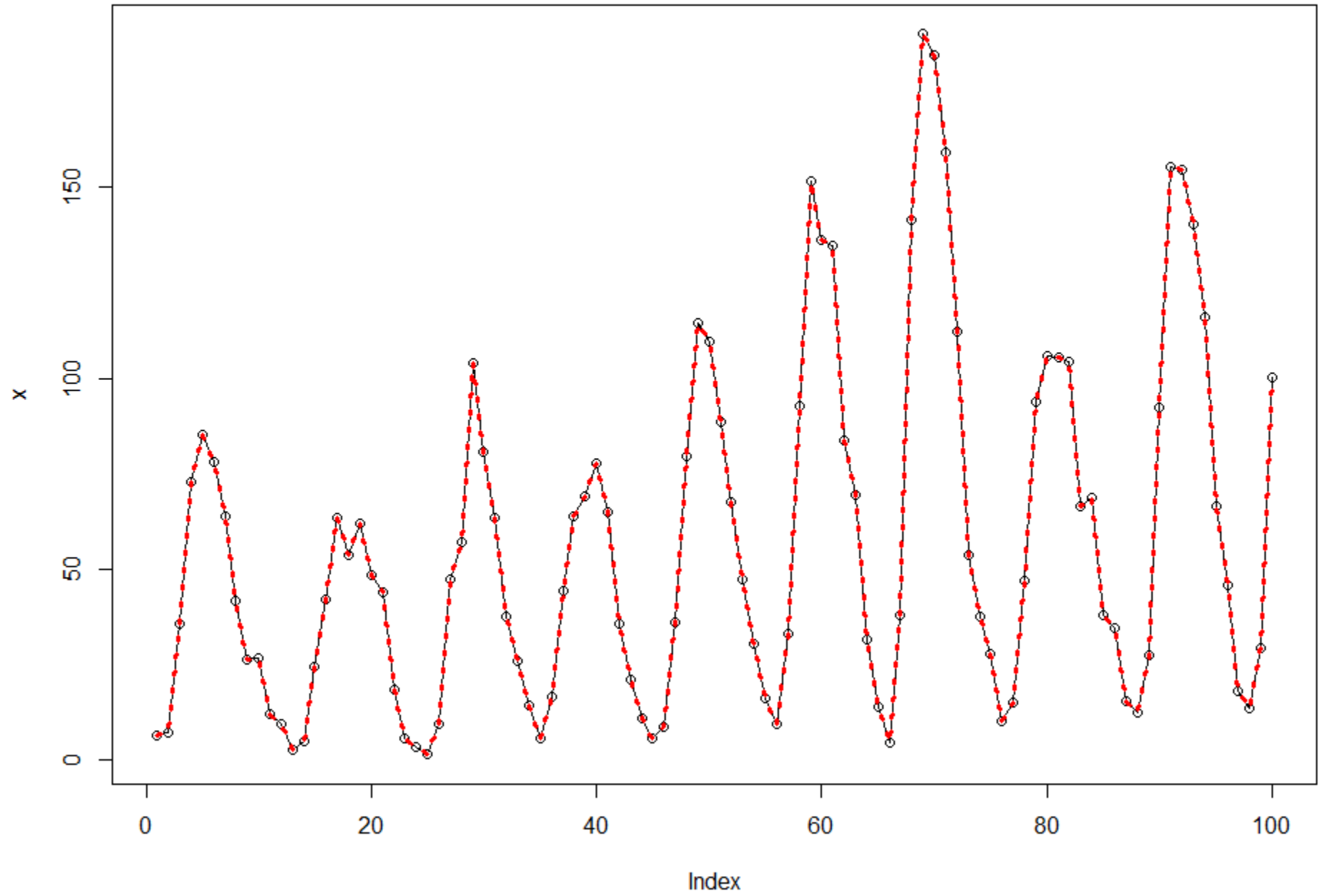


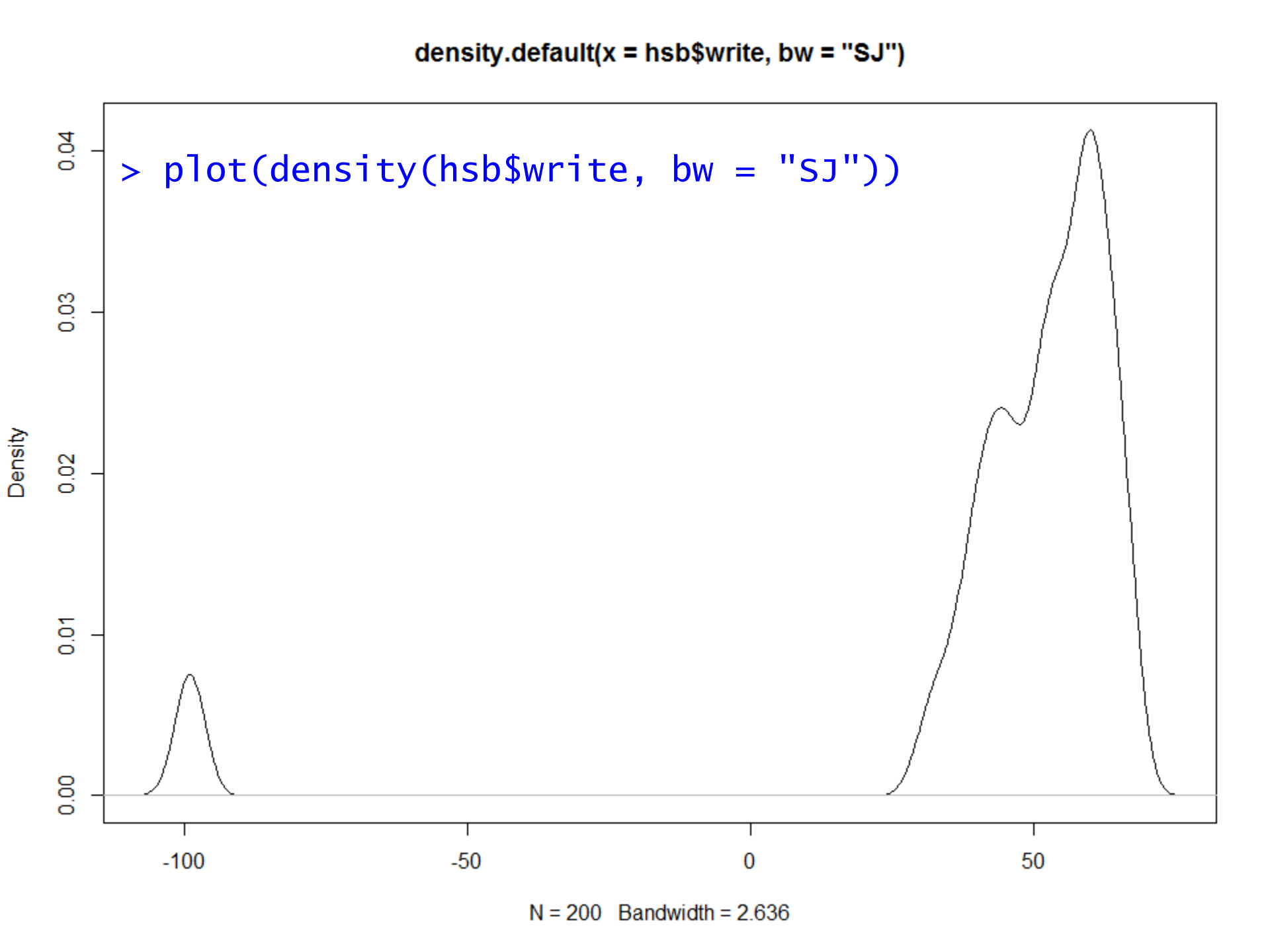
```
> plot(tail(sunspot.year, 100))
```



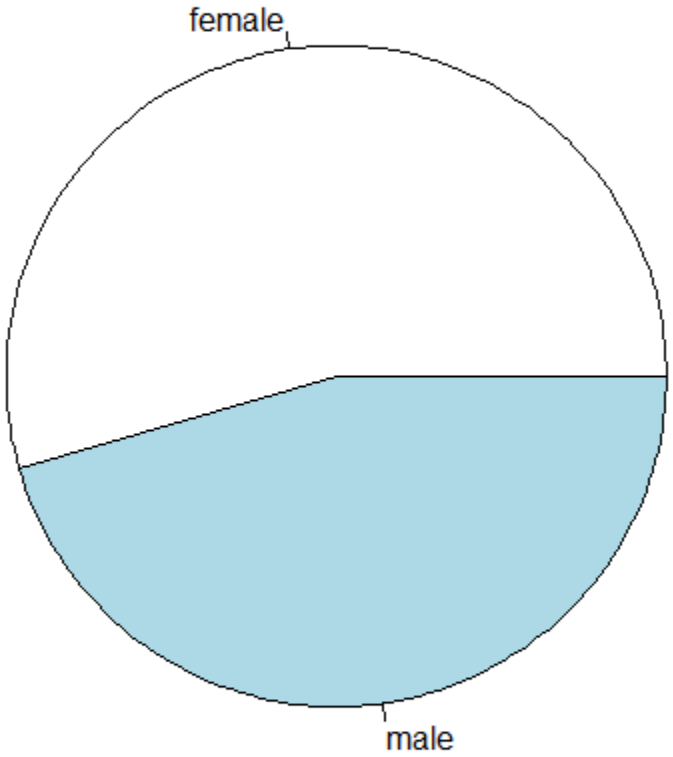

> lines(tail(sunspot.year, 100))
x
0
50
100
150
0
20
40
60
80
100
Index

```
> lines(tail(sunspot.year, 100), lty=3, lwd=3, col=2)
```




density.default(x = hsb$write, bw = "SJ")
> plot(density(hsb$write, bw = "SJ"))
Density
0.00
0.01
0.02
0.03
0.04
-100
-50
0
50
N = 200 Bandwidth = 2.636

```
> pie(hsb$sex)
```



各章節的習題

依照註解指示完成程式碼

輸入 `> submit()` 遞交習題

通過方可繼續進行課程

目前的章節為示範章節，請見現場展示

Let’s Roll

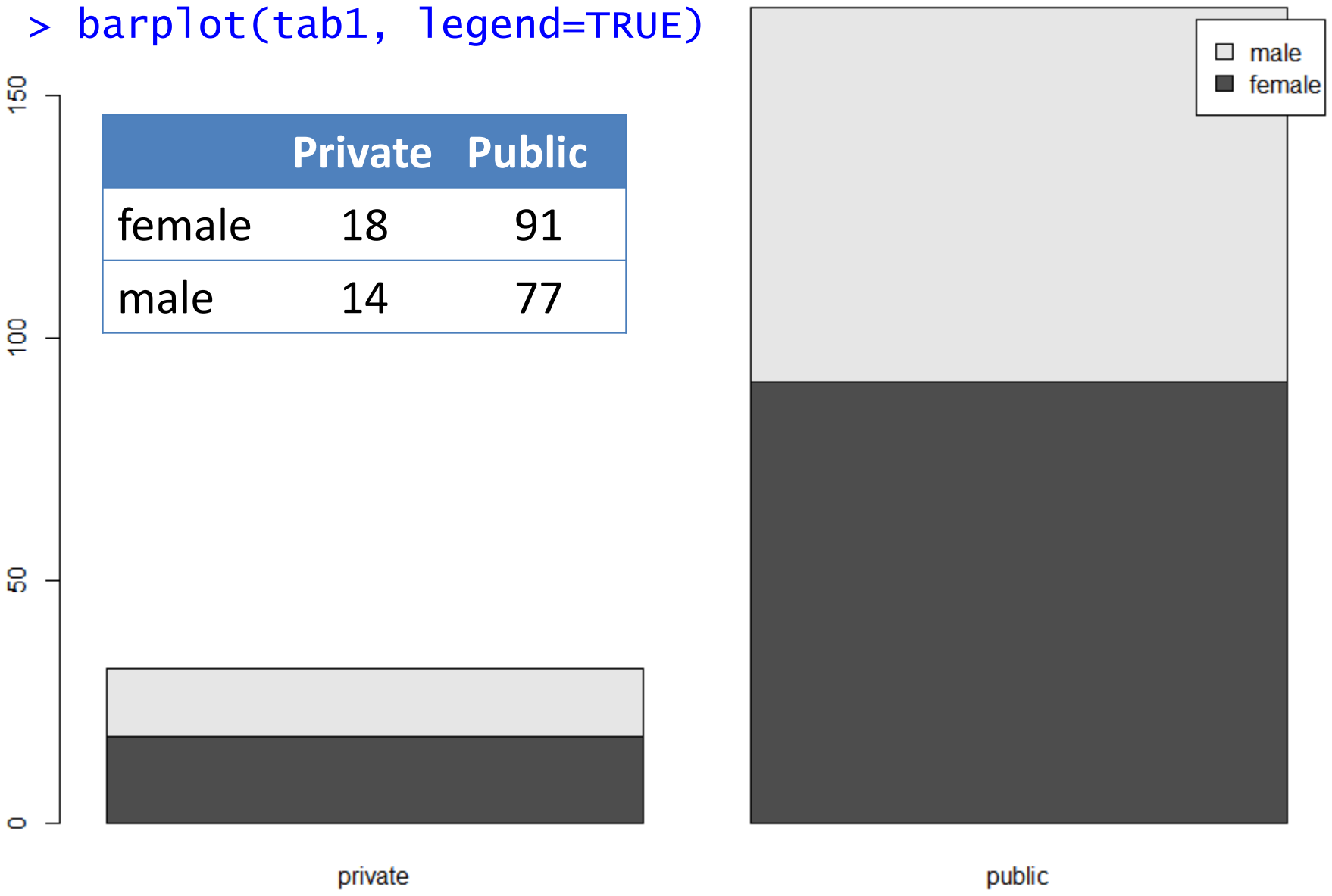
接下來我會帶著各位同學進行下面的實作課程

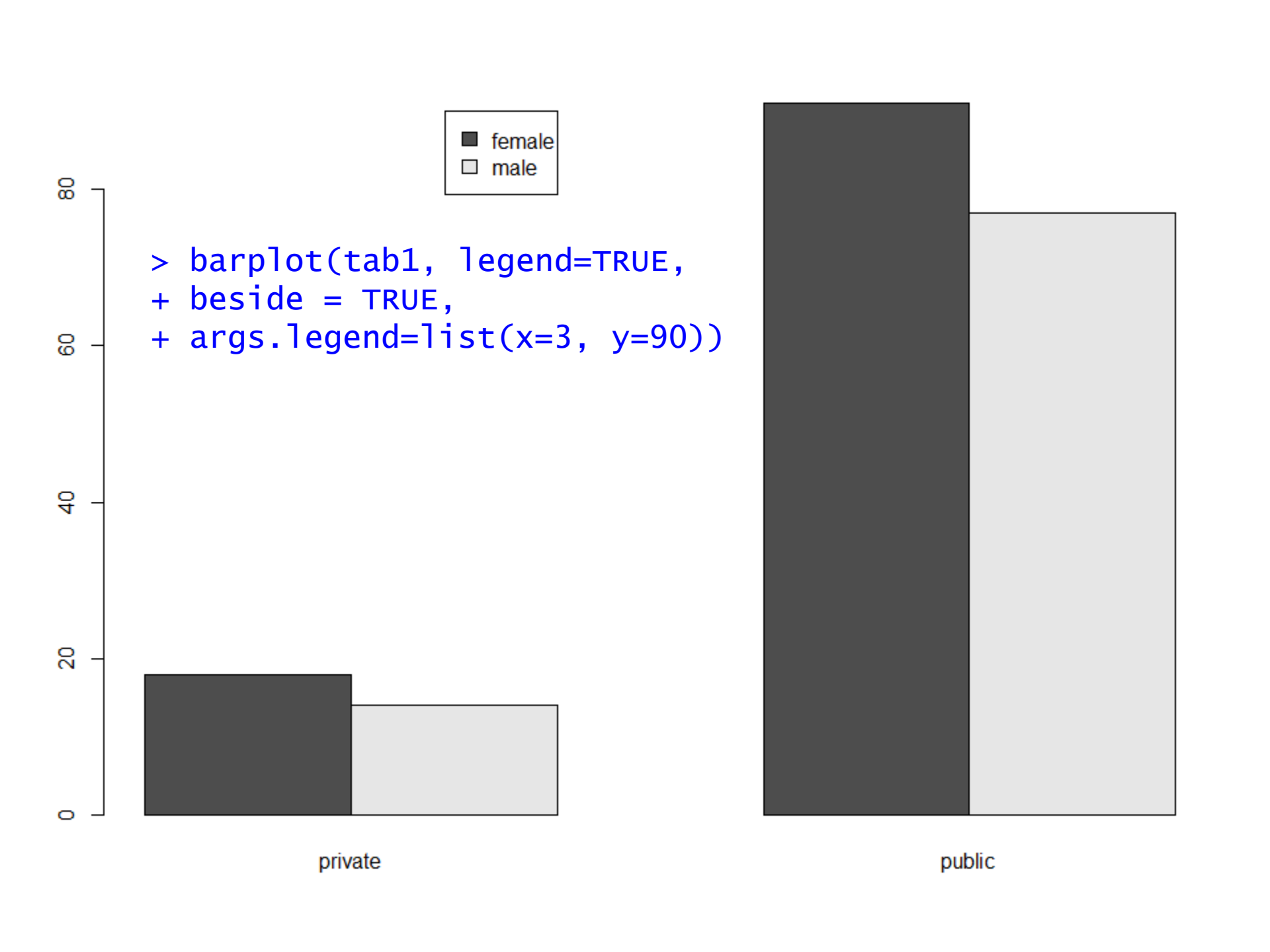
2: 02-DataObservation-02-MultiVariables

請各位同學搭配講課的進度，操作 swirl 課程

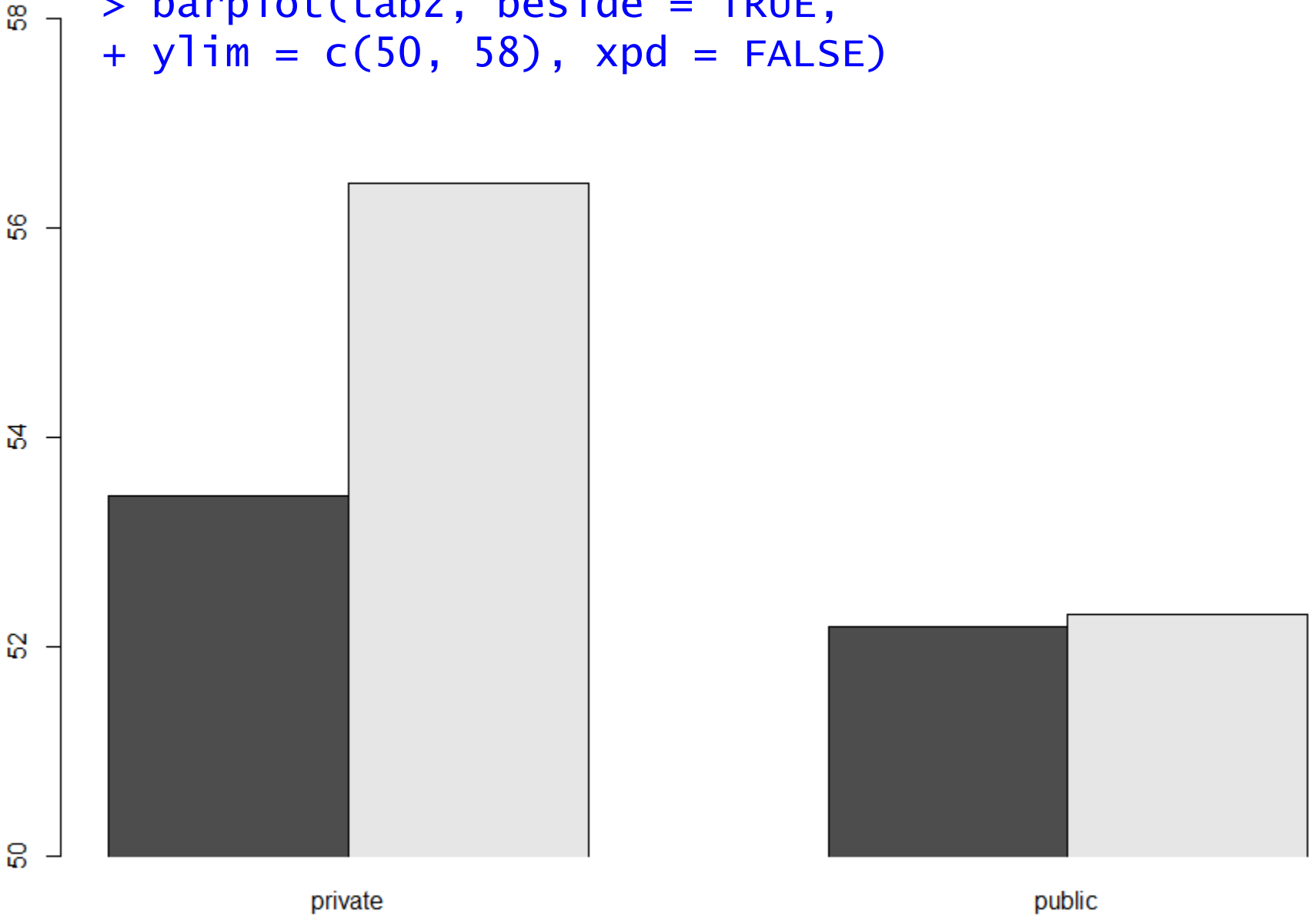
```
> barplot(tab1, legend=TRUE)
```

	Private	Public
female	18	91
male	14	77

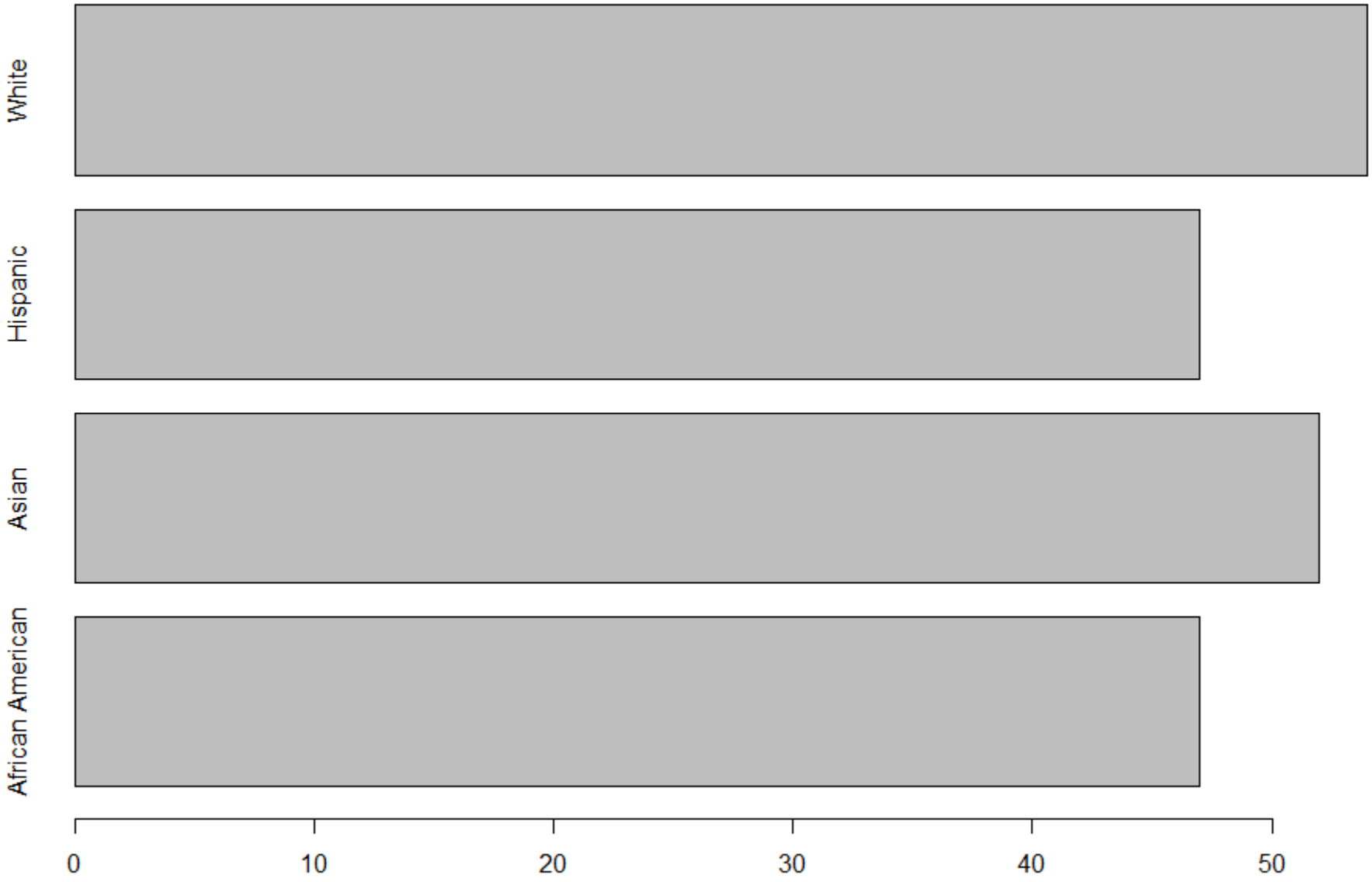



female
male
> barplot(tab1, legend=TRUE,
+ beside = TRUE,
+ args.legend=list(x=3, y=90))
0
20
40
60
80
private
public

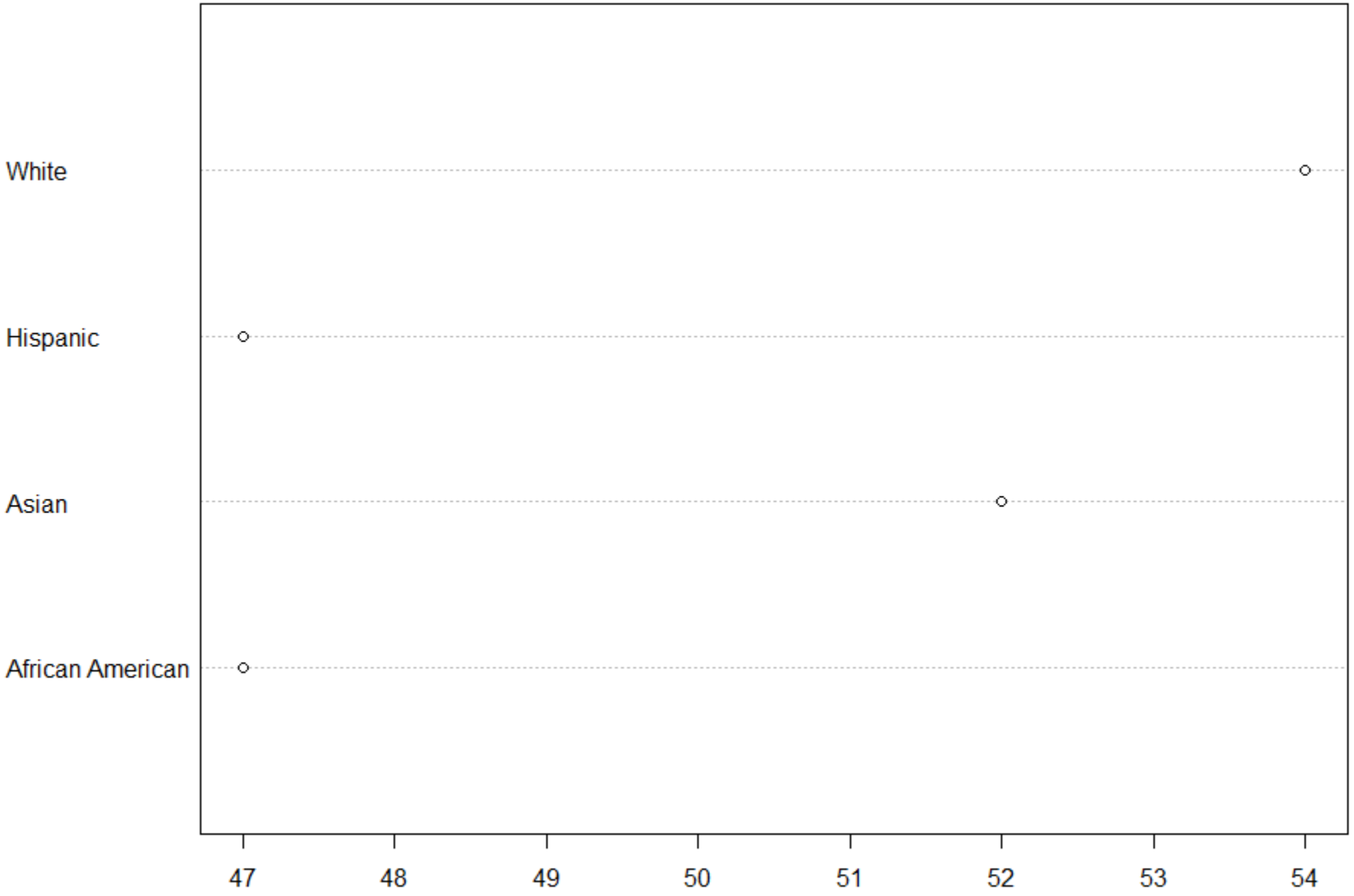
```
> barplot(tab2, beside = TRUE,
+ ylim = c(50, 58), xpd = FALSE)
```



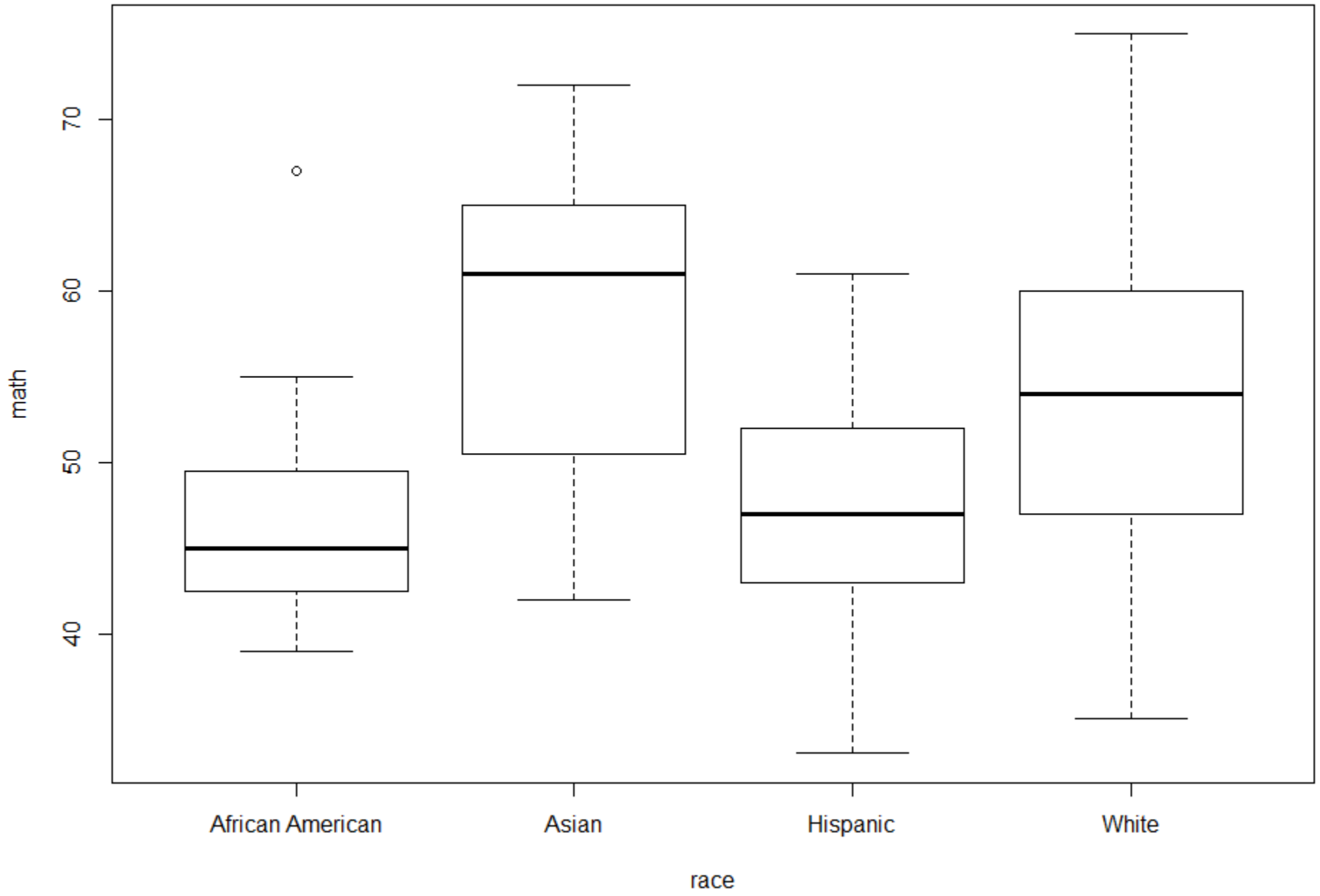
```
> barplot(dat3$read.med, names.arg=dat3$race, horiz=TRUE)
```

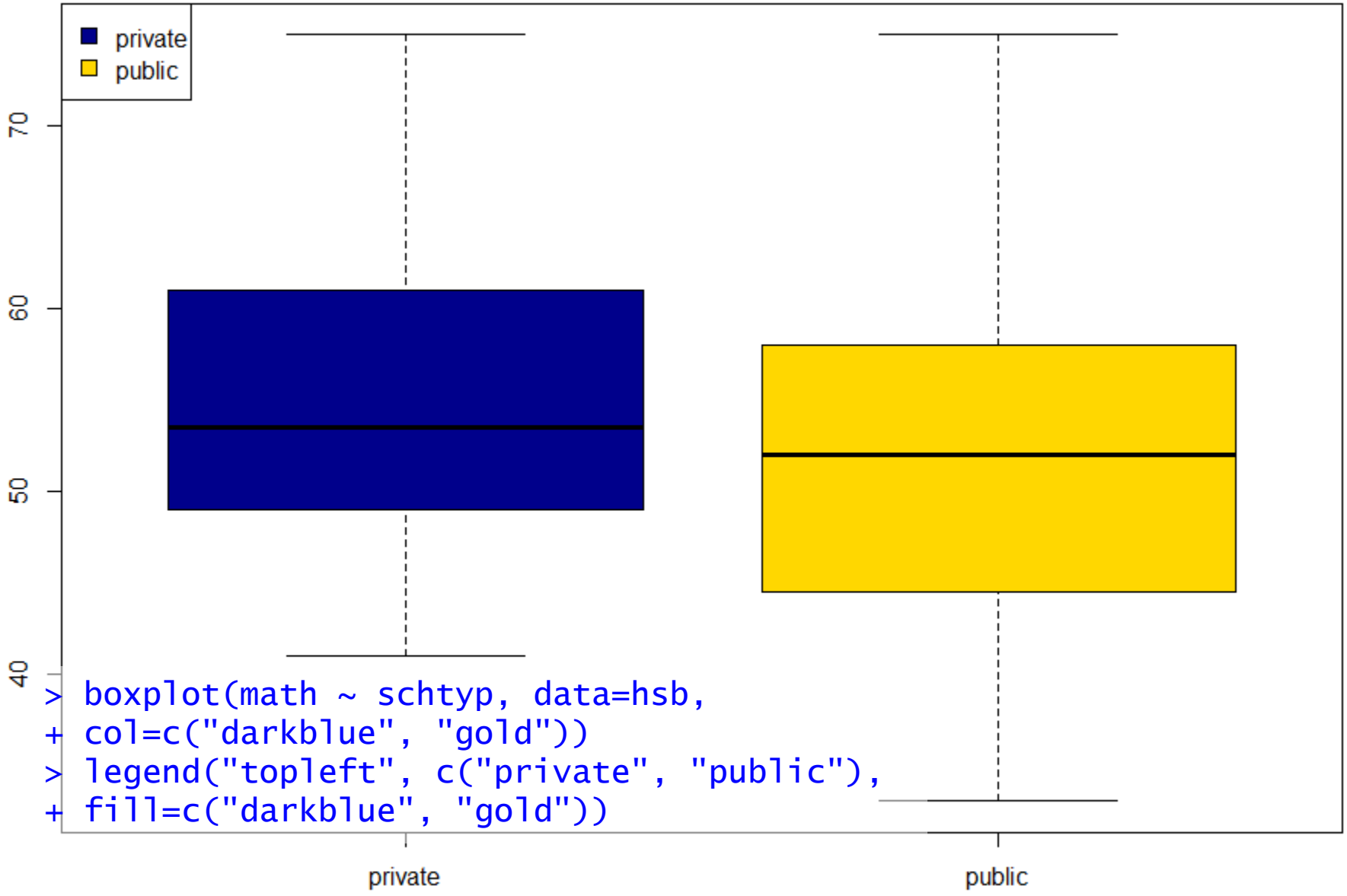


```
> dotchart(dat3$read.med, labels = dat3$race)
```



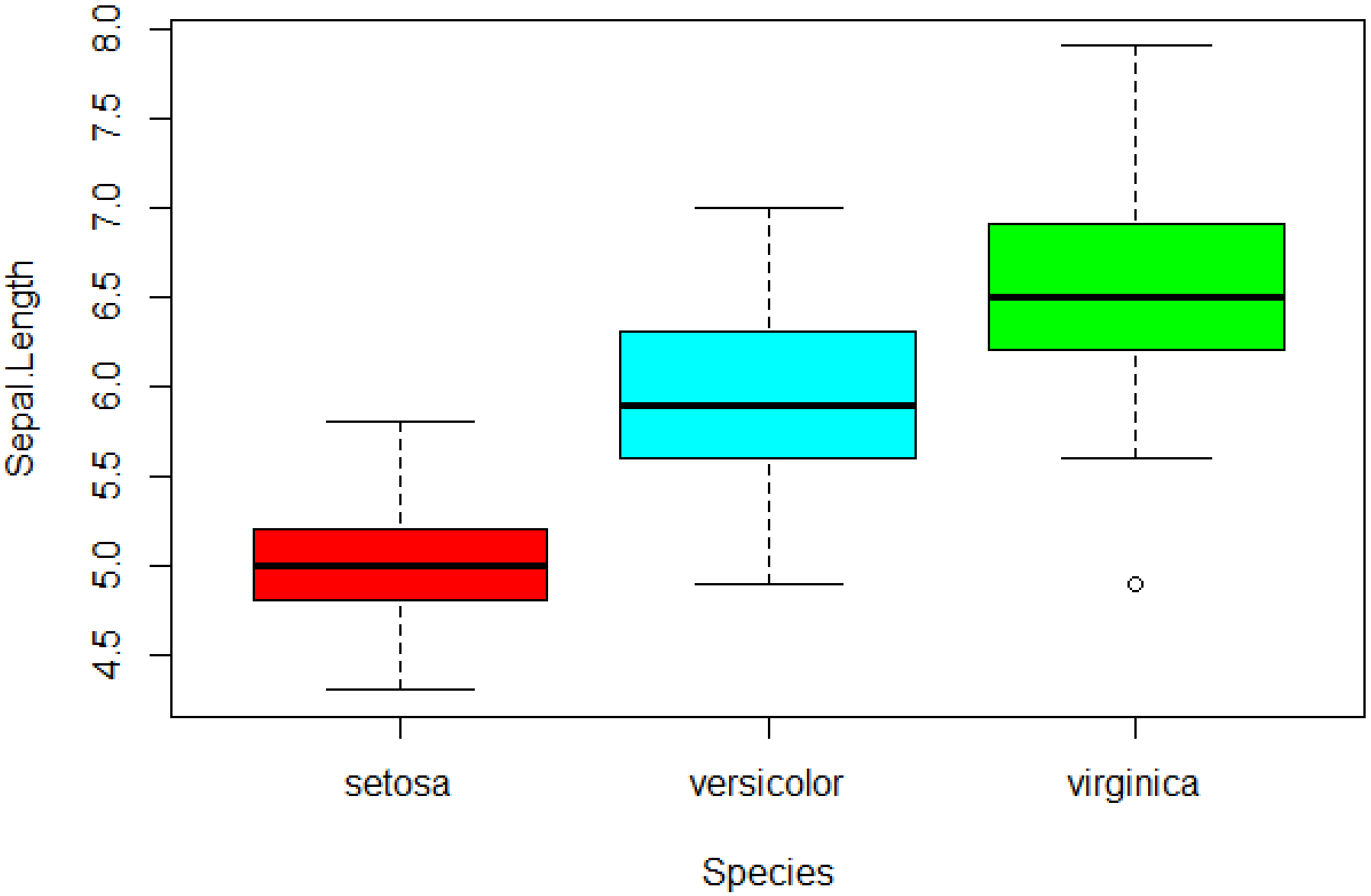
```
> plot(math ~ race, data=hsb)
```

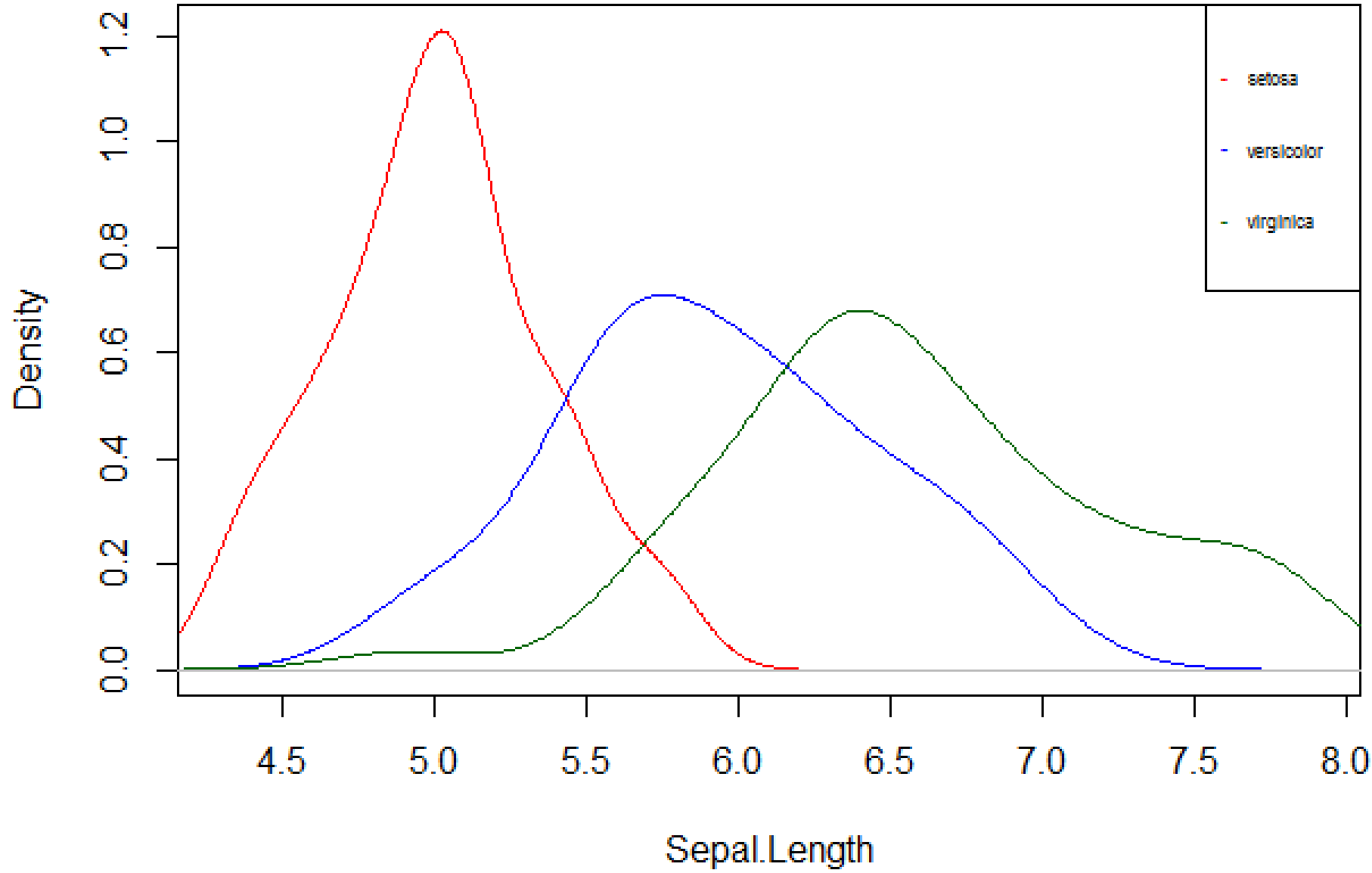


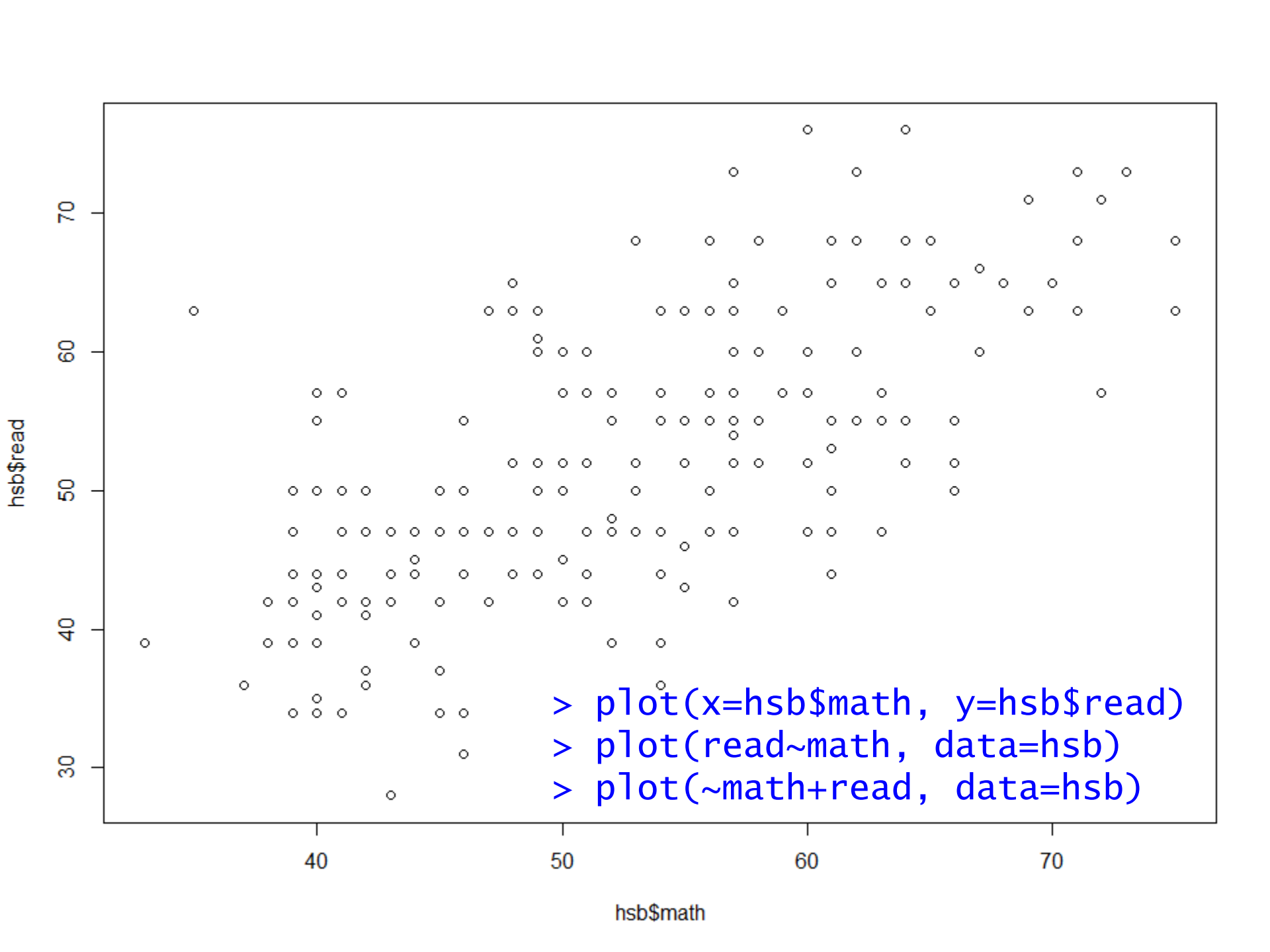


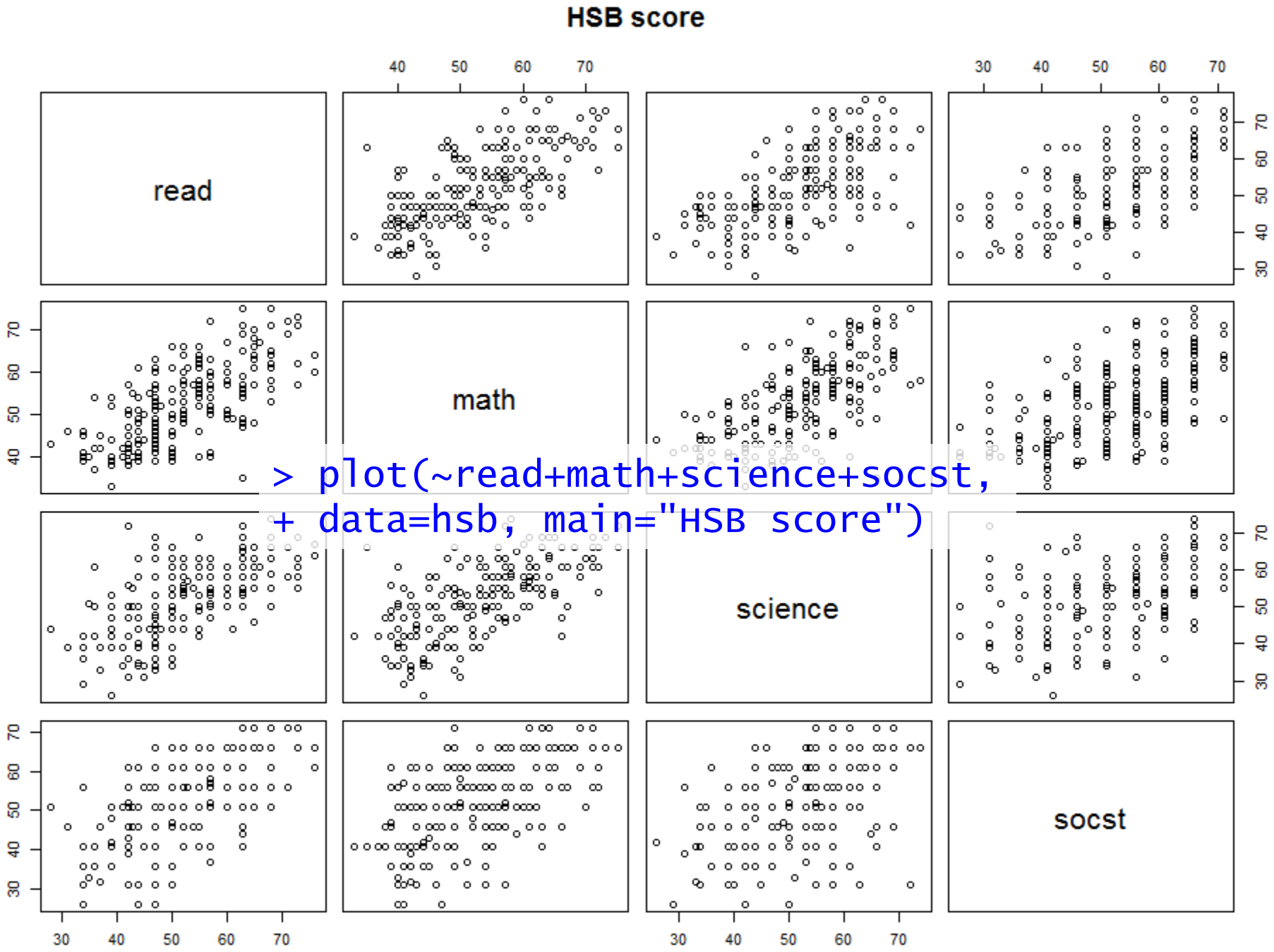

```
> boxplot(math ~ schtyp, data=hsb,
+ col=c("darkblue", "gold"))
> legend("topleft", c("private", "public"),
+ fill=c("darkblue", "gold"))
```

```
> plot(Sepal.Length~Species, iris)
```

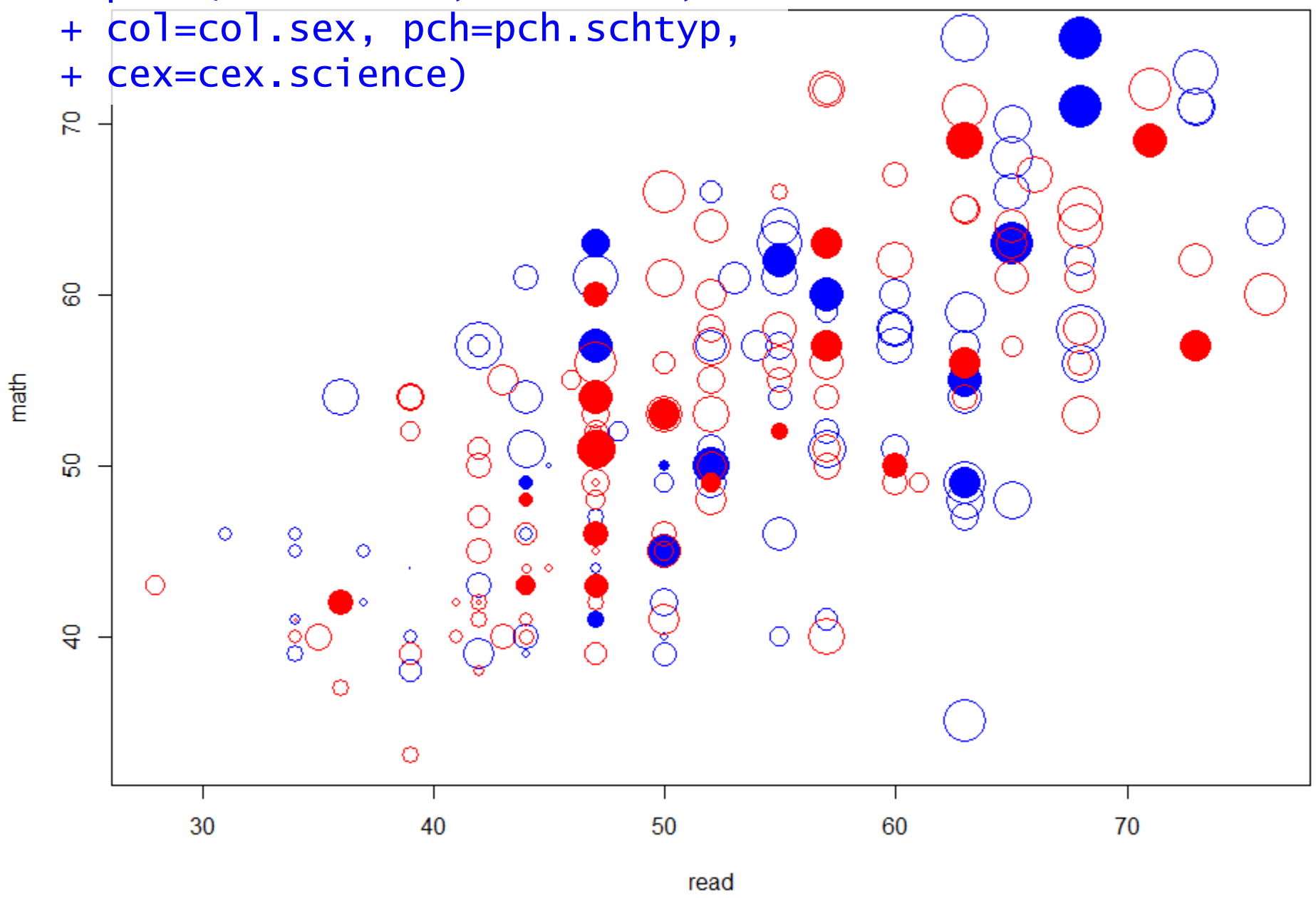



Density
0.0
0.2
0.4
0.6
0.8
1.0
1.2
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
Sepal.Length
setosa
versicolor
virginica


hsb$read
30
40
50
60
70
40
50
60
70
hsb$math
> plot(x=hsb$math, y=hsb$read)
> plot(read~math, data=hsb)
> plot(~math+read, data=hsb)


HSB score
read
math
science
socst
40
50
60
70
30
40
50
60
70
> plot(~read+math+science+socst,
+ data=hsb, main="HSB score")

```
> plot(~read+math, data=hsb,
+ col=col.sex, pch=pch.schtyp,
+ cex=cex.science)
```



R 語言視覺化工具

內建的基本繪圖 API

用於繪製各式統計圖形

以**直角座標系**為基礎，進行**幾何圖形**的繪製

ggplot2

Hadley Wickham 所開發

以**圖層**與 **Functional Language** 概念為基礎的繪圖 API

同學可以自行練習 **9: X1-Optional-01-ggplot2**

R 語言與資料工程

用 R 處理資料

R 是一套完整的資料科學解決方案

資料的收集

資料的處理

資料的視覺化

關於讀取資料的互動式課程

讀取儲存於磁碟的資料

3: 03-RDataEngineer-01-Loading n Parsing

處理與進一步整理結構化資料的技術

4: 04-RDataEngineer-02-DataManipulation

將不同的表格資料融合運用

5: 05-RDataEngineer-03-Join

R 語言與資料處理

將磁碟中的資料讀進 R 環境

R 預設的讀取資料技術

套件中的資料

- 預設的資料集，如 iris
- 載入套件帶來的資料集 如 library(Lahman)

CSV (Comma Separated Values)

- 以逗號區隔的結構化資料 (structured data)
- 每一列都有同樣多的資料欄
- TSV（ Tab Separated Values)

中文編碼

CSV 的眉眉角角

Year,Make,Model,
1997,Ford,E350
2000,Mercury,Cougar

"1997","Ford","E350"

1997,Ford,E350,"super, luxurious truck"

1997,Ford,E350,"super, ""luxurious"" truck"

1997,Ford,E350

1997, Ford, E350

1997,Ford,E350,4.9

1997;Ford;E350;4,9

字符編碼

人類習慣的是 **10** 進位 (Decimal)

0, 1, 2, 3, …, 10, …, 15, 16, …, 175

電腦用的是 **2** 進位 (Binary)

0, 1, 10, 11, …, 1010, …, 1111, 10000, …, 10101111

16 進位 (Hexadecimal, Hex)

0, 1, 2, 3, …, A, …, F, 10, …, AF

1 Byte (位元組) = **8 bit** (位元) 是目前電腦計算記憶體的基本單位

1 Byte 可以表達 0 – 255 的值

正好可以用兩個 Hex Code 表達 ($16 = 2^4$)

00000001 ==> 0000,0001 ==> 01

10101111 ==> 1010,1111 ==> AF

資料在電腦中是如何被儲存的

現今 32位元的電腦，會用 4 Byte儲存一個整數 (4 x 8 = 32)

因此 整數 0L 在記憶體中看起來是 00 00 00 00

那 “0” 要怎麼儲存呢？

還記得 **factor** 嗎？

我們說過 factor 其實是一種**字串的編碼**

將字串 mapping 成數字

文字在電腦中也是這樣儲存的

"0"=> 在電腦中以 Hex Code 來看是 30

"A": 41, "B": 42, …, "Z": 5A, …, "a": 97

Enter(\r): 0D, 換行(\n): 0A

⬅ **ASCII 編碼**

中文編碼

BIG5
"中" : A4 A4
"文" : A4 E5

UTF-8
"中" : E4 B8 AD
"文" : E6 96 87

以 UTF-8 編碼寫成的「中文」二字，在電腦看來是

E4 B8 AD E6 96 87

若是以 BIG5 編碼讀入，會變成「鍗剜 」

讀取中文檔案時，必須先**確定編碼**，否則無法正確讀取

處理編碼問題

iconv()

R 環境中轉換字元的函式

請以 ?iconv 閱讀其說明文件

encoding, fileEncoding

R 環境中用來讀取資料的函式常見的參數，用來設定來源資料的編碼型態

Sys.getlocale(), Sys.setlocale(locale = “cht”)

R 環境用來設定語境的環境變數，可以減少部分編碼帶來的困擾

R 語言與資料處理

從字串中獲取資料 - PARSING

何謂非結構化資料？

非結構化資料範例

64.242.88.10 - - [07/Mar/2004:16:05:49 -0800] "GET /twiki/bin/edit/Main/Double_bounce_sender?topicparent=Main.ConfigurationVariables HTTP/1.1" 401 12846

64.242.88.10 - - [07/Mar/2004:16:06:51 -0800] "GET /twiki/bin/rdiff/TWiki/NewUserTemplate?rev1=1.3&rev2=1.2 HTTP/1.1" 200 4523

什麼是 Parsing

告訴電腦分拆非結構化資料的規則

Domain Knowledge，例如 ip位址: 168.95.192.1

字元在字串中的**位置**，如 121**E**25**N**

分隔符號，如逗號、分號、冒號等

Regular Expression 正規表示式

身分證字號的正規表示式:

^[A-Z]{1}[1-2]{1}[0-9]{8}$

Let’s do it

在 R 環境中，讀取檔案，並且將字串轉換成可處理的資料，請同學們完成

3: 03-RDataEngineer-01-Loading n Parsing

What's More

在 R 環境中，克服中文編碼帶來的麻煩，同學可以自行練習

10: X2-Challenge-01-ChineseEncoding

R 語言與資料工程

處理以及操作結構化資料

R 的結構化資料來源

內部: `data.frame`、`data.table`

外部: 關聯式資料庫、格式健全良好的 CSV 檔案

整理結構化資料

分類報表

男性、女性、重度消費者、輕度消費者

分時報表

月報、季報、年報

從 raw data 中計算指標

安打/打數 = 打擊率、總得分/場次 = 平均得分

利用 dplyr 套件進一步整理結構化資料

函式以整理資料用的動詞命名，簡化整理資料的思考邏輯，方便程式撰寫

命名邏輯與 SQL 類似，習慣 SQL 語法的工程師可以快速上手

優化過的效能

整理資料的動作

filter(): 取出符合條件的**資料列** (**過濾**不符合條件的資料)

arrange(): 依照需求**安排**(排序)資料

select(): **揀選**需要的**欄位**

distinct(): 找出**獨一無二**的值

mutate(): 結合原有欄位，計算**新欄位**，例如比例、類型.....

group_by: 將資料依類別**集合**做各別的子 data.frame

summarise(): 將整個 data.frame 以一個值**概括**

sample_n() & sample_frac(): 依數量或比例進行**取樣**

Let’s do it

實際利用 dplyr 整理結構化資料，請同學們完成

4: 04-RDataEngineer-02-DataManipulation

程式碼壓縮與可讀性

```
> x1 <- filter(flights, ...)
> x2 <- select(x1, ...)
> x3 <- summarise(x2, ...)
```

Nested Function Call

```
> x3 <- summarise(select(filter(flights, ...) , ...) , ...)
```

Pipeline Operation

```
> x3 <-
+      filter(flights, ...) %>%
+      select(...) %>%
+      summarise(...)
```

R 語言與資料工程

結合不同的資料源

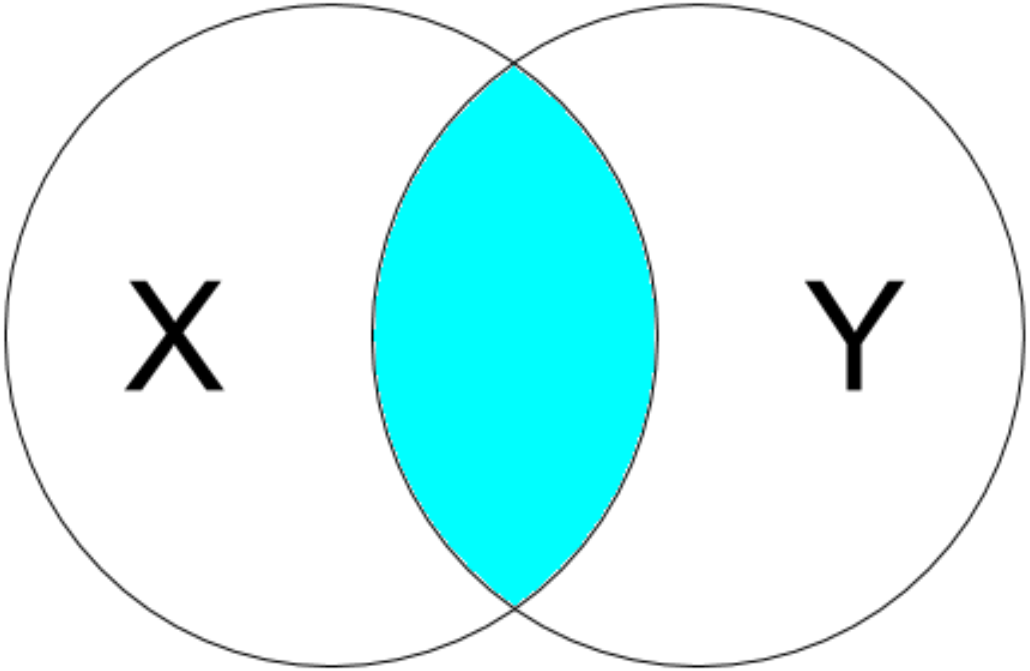
多資料源的價值

flights

flights + weather

flights + weather + airports

inner_join



All columns from both X and Y
Duplicate if multiple matching

superheroes

name	alignment	gender	publisher
Magneto	bad	male	Marvel
Storm	good	female	Marvel
Mystique	bad	female	Marvel
Batman	good	male	DC
Joker	bad	male	DC
Catwoman	bad	female	DC
Hellboy	good	male	Dark Horse Comics

publishers

publisher	yr_founded
DC	1934
Marvel	1939
Image	1992

inner_join(x = superheroes, y = publishers)

name	alignment	gender	publisher	yr_founded
Magneto	bad	male	Marvel	1939
Storm	good	female	Marvel	1939
Mystique	bad	female	Marvel	1939
Batman	good	male	DC	1934
Joker	bad	male	DC	1934
Catwoman	bad	female	DC	1934

*from https://stat545-ubc.github.io/bit001_dplyr-cheatsheet.html

publishers

publisher	yr_founded
DC	1934
Marvel	1939
Image	1992

superheroes

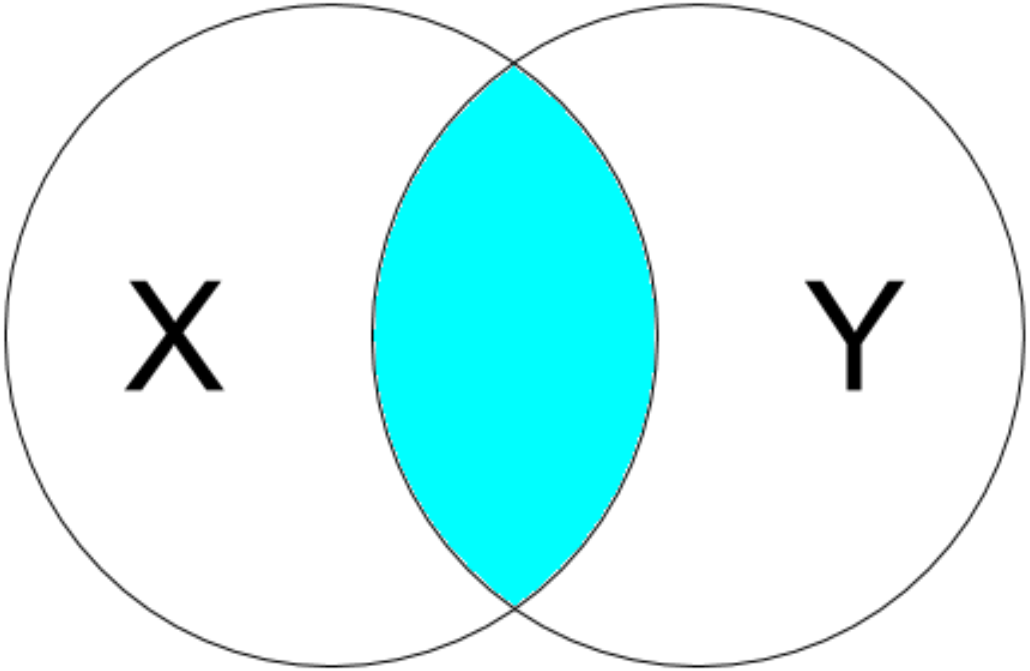
name	alignment	gender	publisher
Magneto	bad	male	Marvel
Storm	good	female	Marvel
Mystique	bad	female	Marvel
Batman	good	male	DC
Joker	bad	male	DC
Catwoman	bad	female	DC
Hellboy	good	male	Dark Horse Comics

inner_join(x = publishers, y = superheroes)

publisher	yr_founded	name	alignment	gender
DC	1934	Batman	good	male
DC	1934	Joker	bad	male
DC	1934	Catwoman	bad	female
Marvel	1939	Magneto	bad	male
Marvel	1939	Storm	good	female
Marvel	1939	Mystique	bad	female

*from https://stat545-ubc.github.io/bit001_dplyr-cheatsheet.html

semi_join



Only columns from X
No duplication

superheroes

name	alignment	gender	publisher
Magneto	bad	male	Marvel
Storm	good	female	Marvel
Mystique	bad	female	Marvel
Batman	good	male	DC
Joker	bad	male	DC
Catwoman	bad	female	DC
Hellboy	good	male	Dark Horse Comics

publishers

publisher	yr_founded
DC	1934
Marvel	1939
Image	1992

semi-join(x = superheroes, y = publishers)

name	alignment	gender	publisher
Batman	good	male	DC
Joker	bad	male	DC
Catwoman	bad	female	DC
Magneto	bad	male	Marvel
Storm	good	female	Marvel
Mystique	bad	female	Marvel

*from https://stat545-ubc.github.io/bit001_dplyr-cheatsheet.html

publishers

publisher	yr_founded
DC	1934
Marvel	1939
Image	1992

superheroes

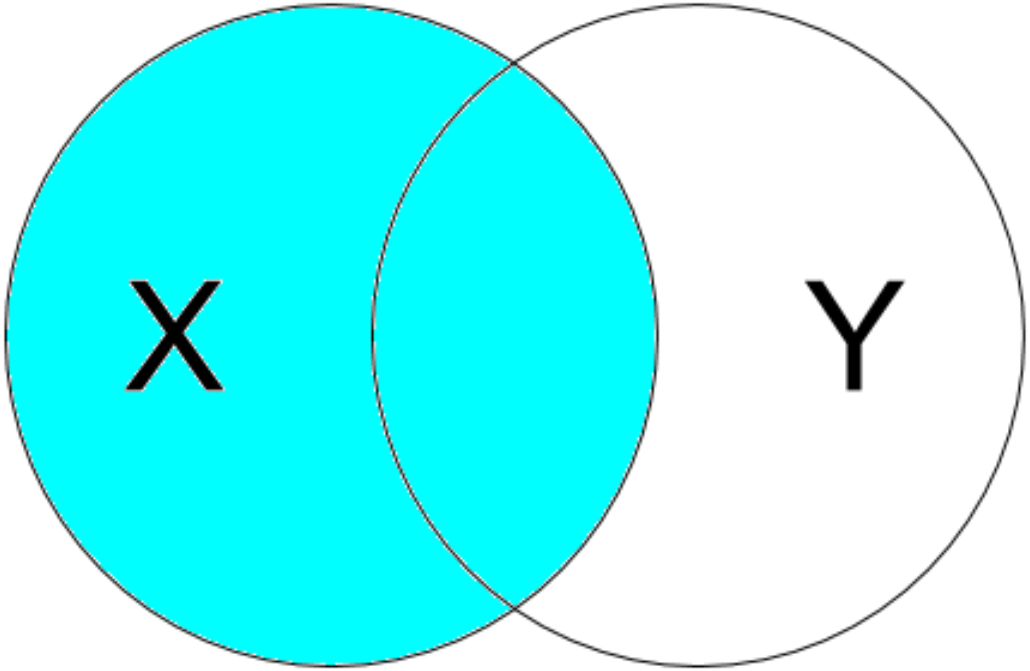
name	alignment	gender	publisher
Magneto	bad	male	Marvel
Storm	good	female	Marvel
Mystique	bad	female	Marvel
Batman	good	male	DC
Joker	bad	male	DC
Catwoman	bad	female	DC
Hellboy	good	male	Dark Horse Comics

semi-join(x = publishers, y = superheroes)

publisher	yr_founded
Marvel	1939
DC	1934

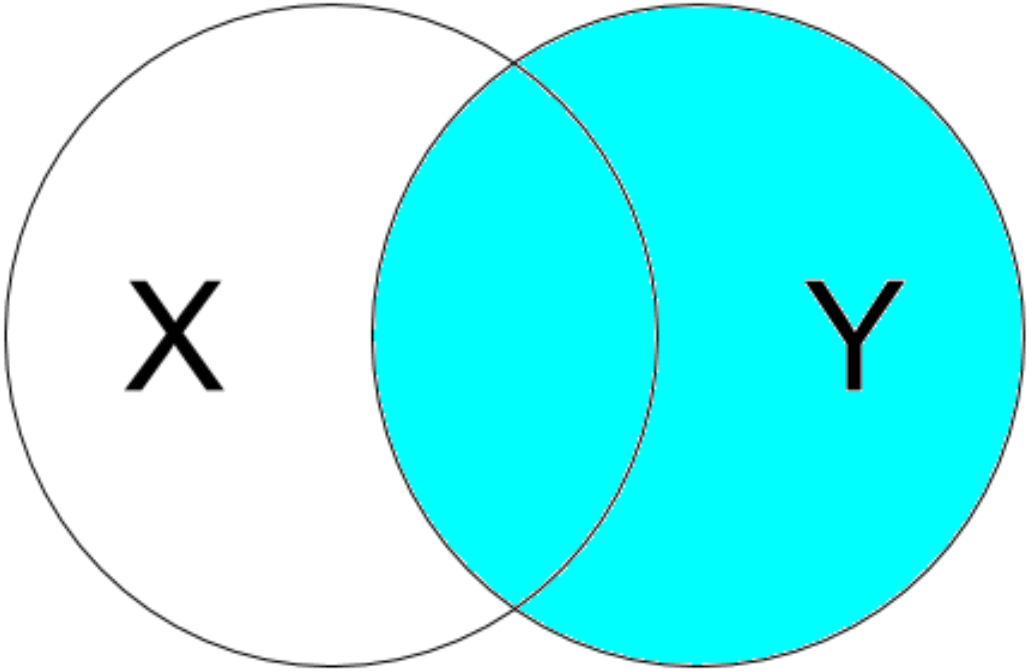
*from https://stat545-ubc.github.io/bit001_dplyr-cheatsheet.html

left_join



All columns from both X and Y
NA's for missing values

right_join



All columns from both X and Y

NA's for missing values

superheroes

name	alignment	gender	publisher
Magneto	bad	male	Marvel
Storm	good	female	Marvel
Mystique	bad	female	Marvel
Batman	good	male	DC
Joker	bad	male	DC
Catwoman	bad	female	DC
Hellboy	good	male	Dark Horse Comics

publishers

publisher	yr_founded
DC	1934
Marvel	1939
Image	1992

left_join(x = superheroes, y = publishers)

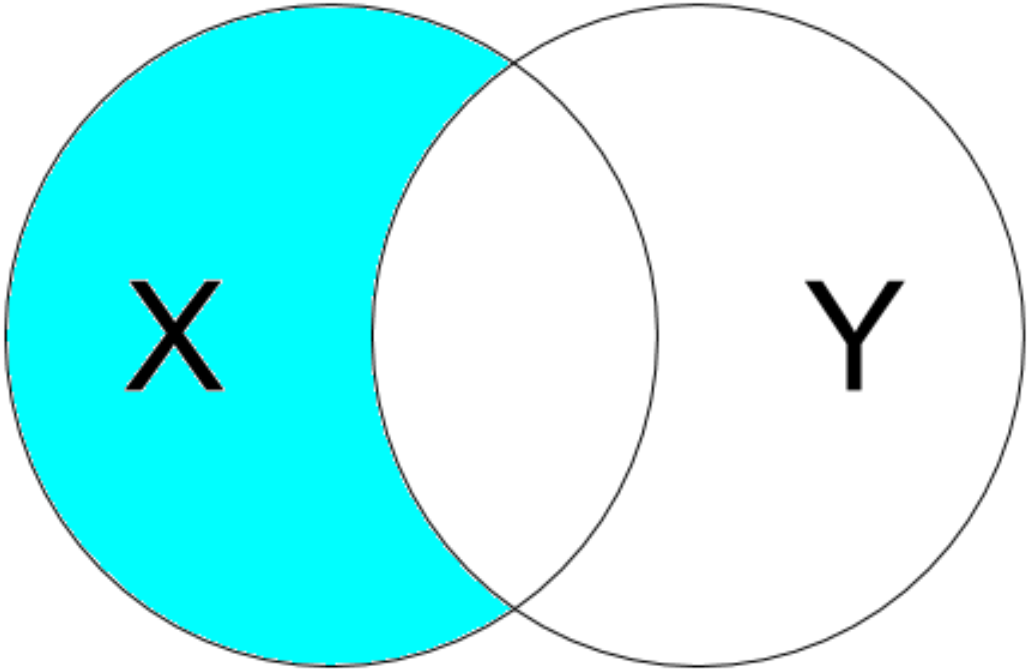
name	alignment	gender	publisher	yr_founded
Magneto	bad	male	Marvel	1939
Storm	good	female	Marvel	1939
Mystique	bad	female	Marvel	1939
Batman	good	male	DC	1934
Joker	bad	male	DC	1934
Catwoman	bad	female	DC	1934
Hellboy	good	male	Dark Horse Comics	NA

*from https://stat545-ubc.github.io/bit001_dplyr-cheatsheet.html

publishers		superheroes				left_join(x = publishers, y = superheroes)				
publisher	**yr_founded**	**name**	**alignment**	**gender**	**publisher**	**publisher**	**yr_founded**	**name**	**alignment**	**gender**
DC	1934	Magneto	bad	male	Marvel	DC	1934	Batman	good	male
Marvel	1939	Storm	good	female	Marvel	DC	1934	Joker	bad	male
Image	1992	Mystique	bad	female	Marvel	DC	1934	Catwoman	bad	female
		Batman	good	male	DC	Marvel	1939	Magneto	bad	male
		Joker	bad	male	DC	Marvel	1939	Storm	good	female
		Catwoman	bad	female	DC	Marvel	1939	Mystique	bad	female
		Hellboy	good	male	Dark Horse Comics	Image	1992	NA	NA	NA

*from https://stat545-ubc.github.io/bit001_dplyr-cheatsheet.html

anti_join



Only columns from X

superheroes

name	alignment	gender	publisher
Magneto	bad	male	Marvel
Storm	good	female	Marvel
Mystique	bad	female	Marvel
Batman	good	male	DC
Joker	bad	male	DC
Catwoman	bad	female	DC
Hellboy	good	male	Dark Horse Comics

publishers

publisher	yr_founded
DC	1934
Marvel	1939
Image	1992

anti_join(x = superheroes, y = publishers)

name	alignment	gender	publisher
Hellboy	good	male	Dark Horse Comics

*from https://stat545-ubc.github.io/bit001_dplyr-cheatsheet.html

publishers

publisher	yr_founded
DC	1934
Marvel	1939
Image	1992

superheroes

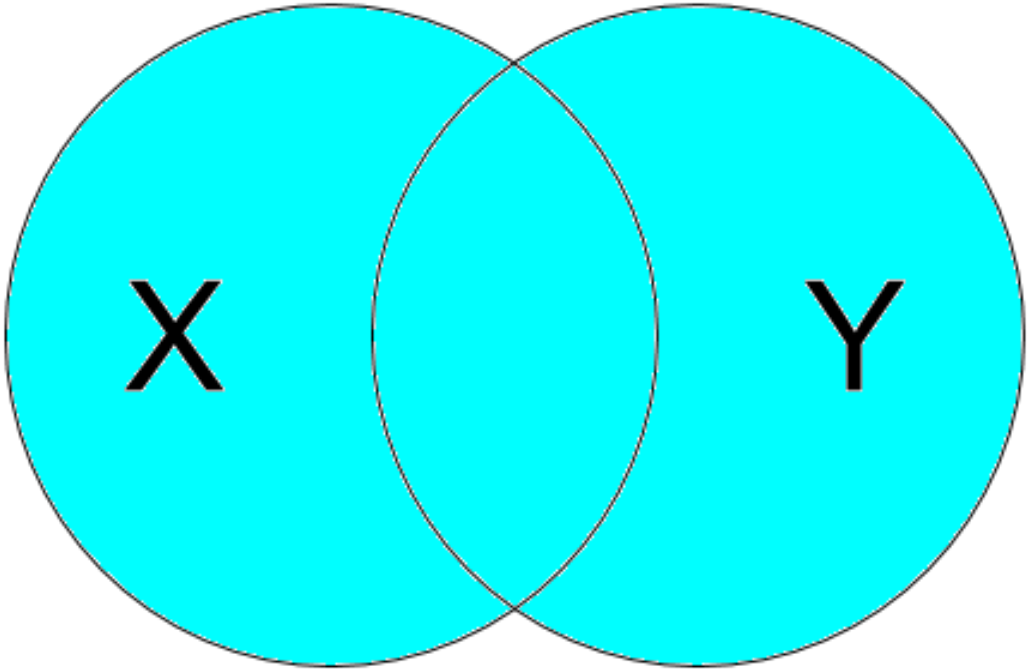
name	alignment	gender	publisher
Magneto	bad	male	Marvel
Storm	good	female	Marvel
Mystique	bad	female	Marvel
Batman	good	male	DC
Joker	bad	male	DC
Catwoman	bad	female	DC
Hellboy	good	male	Dark Horse Comics

anti_join(x = publishers, y = superheroes)

publisher	yr_founded
Image	1992

*from https://stat545-ubc.github.io/bit001_dplyr-cheatsheet.html

full_join



All columns from both X and Y

NA's for missing values

superheroes

name	alignment	gender	publisher
Magneto	bad	male	Marvel
Storm	good	female	Marvel
Mystique	bad	female	Marvel
Batman	good	male	DC
Joker	bad	male	DC
Catwoman	bad	female	DC
Hellboy	good	male	Dark Horse Comics

publishers

publisher	yr_founded
DC	1934
Marvel	1939
Image	1992

full_join(x = superheroes, y = publishers)

name	alignment	gender	publisher	yr_founded
Magneto	bad	male	Marvel	1939
Storm	good	female	Marvel	1939
Mystique	bad	female	Marvel	1939
Batman	good	male	DC	1934
Joker	bad	male	DC	1934
Catwoman	bad	female	DC	1934
Hellboy	good	male	Dark Horse Comics	NA
NA	NA	NA	Image	1992

*from https://stat545-ubc.github.io/bit001_dplyr-cheatsheet.html

Let’s do it

實際利用 dplyr 整理來自多個資料源的資料，請同學們完成

5: 05-RDataEngineer-03-Join

What's More

處理麻煩的真實資料，並且與結合圖資做視覺化呈現，同學可以自行練習

11: X3-Challenge-02-PirateVisualization

課程筆記

https://hjhsu.github.io/r_course/01-DataObservation-01-SingleVariable.html

https://hjhsu.github.io/r_course/02-DataObservation-02-MultiVariables.html

https://hjhsu.github.io/r_course/03-RDataEngineer-01-Loading_n_Parsing.html

https://hjhsu.github.io/r_course/04-RDataEngineer-02-DataManipulation.html

https://hjhsu.github.io/r_course/05-RDataEngineer-03-Join.html

https://hjhsu.github.io/r_course/X1-Optional-01-ggplot2.html

https://hjhsu.github.io/r_course/X2-Challenge-01-ChineseEncoding.html

https://hjhsu.github.io/r_course/X3-Challenge-02-PirateVisualization.html

R 語言與資料工程

資料科學的 HELLO WORLD

資料科學團隊的第一步

Dashboard 資料儀表板

建立 Dashboard 的意義

確認資料已經可以**正確**的被取出，並且**視覺化**呈現

讓組織中的不同團隊**共享**資料、刺激想法

減少資料科學團隊產生報表的需求，**減少內耗**

建立**信任感**，初步展現資料科學團隊的價值

一個資料源、一個 Dashboard

價值隨資料源的多元而增加

檢驗對不同資料源的想法

- 需要解決的問題
- 設計團隊的 KPI

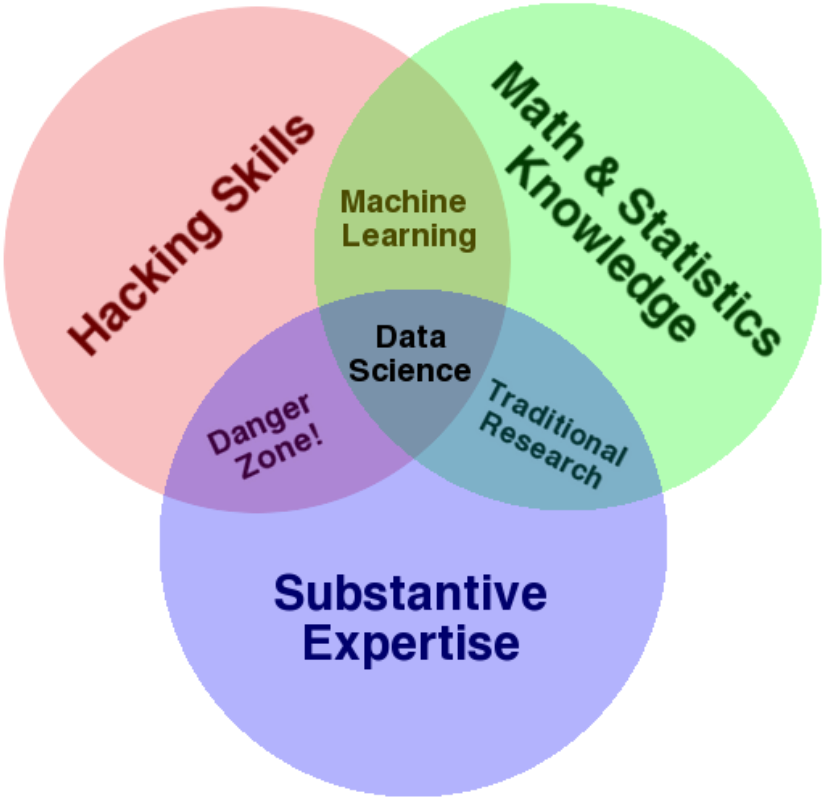
降低知覺的複雜度，幫助跨資料源的整合分析

邁出資料科學的第一步

政府採購資料 v.s. 公司基本資料

各里開票結果 v.s. 各里收入中位數

登革熱病例變化 v.s. 電子發票



謝 謝 各 位
Q & A

許懷中 Hwai-Jung Hsu
hjhsu@iis.sinica.edu.tw
https://tw.linkedin.com/in/hjhsu