R 軟體統計分析應用 (二)
- 資料彙整與演算法運用

日期 2016 07/05

地點 校務研究辦公室

MongoDB 巨量資料分析 01

```
> data <- read.table(header=TRUE, text='
+  subject sex size
+        1   M    7
+        2   F    NA
+        3   F    9
+        4   M   11
+  ')
>
> write.csv(data, "c:/users/user/data.csv", row.names=FALSE)
> save(data, file="c:/users/user/data.RData")
```

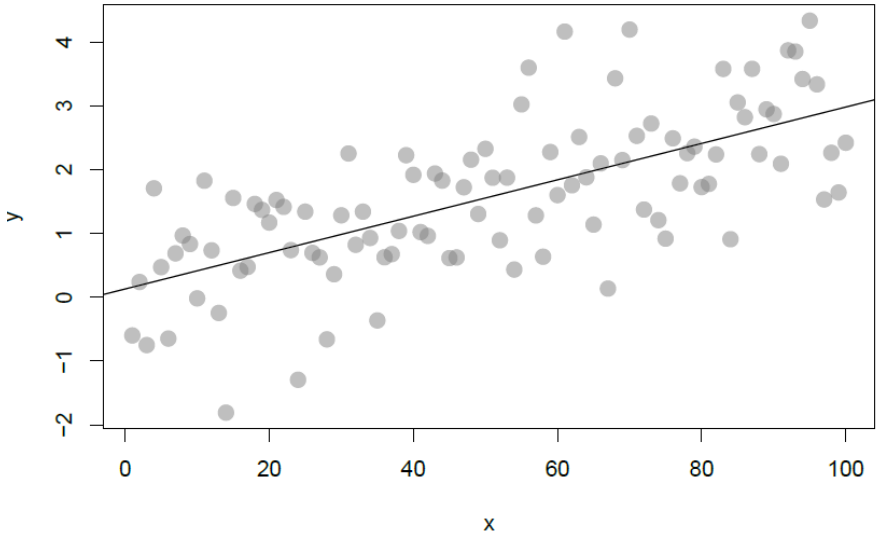
```
> load("c:/users/user/data.RData")
> dim(data)
[1] 4 3
> head(data)
  subject sex size
1       1   M    7
2       2   F   NA
3       3   F    9
4       4   M   11
>
```

```
data <- read.table(header=TRUE, text='
 subject sex size
      1   M    7
      2   F    NA
      3   F    9
      4   M   11
')
```

MongoDB 巨量資料分析 02

```
> set.seed(1)
> x <- 1:100
> y <- 0.029*x + rnorm(100)
> pdf("c:/users/user/sample.pdf", 7,5)
> plot(x, y, pch=19, col=rgb(0.5, 0.5, 0.5, 0.5), cex=1.5)
> abline(lm(y~x))
> dev.off()
```

pdf 與 png 都必須搭配 dev.off() 關閉檔案




```
getwd()
setwd("c:/users/user")
ls(data)
rm(data)


set.seed(1)
x <- 1:100
y <- 0.029*x + rnorm(100)
pdf("c:/users/user/sample.pdf", 7,5)
plot(x, y, pch=19, col=rgb(0.5, 0.5, 0.5, 0.5), cex=1.5)
abline(lm(y~x))
dev.off()
```

MongoDB 巨量資料分析 03

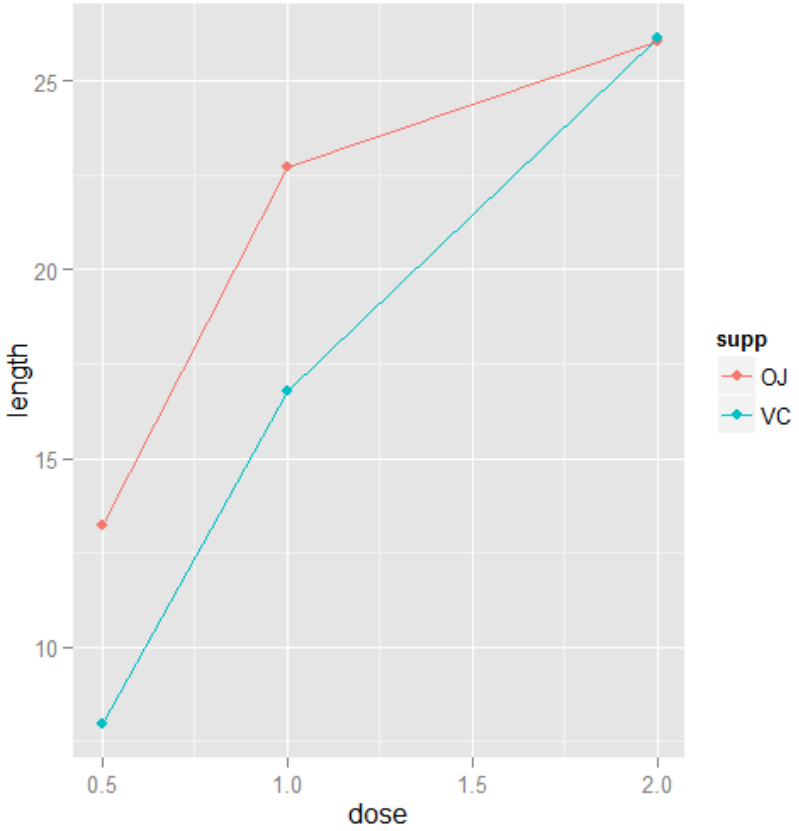
filepath <- "https://dl.dropbox.com/u/1648032/ggplot2_tutorial_dataset.txt"
mydata <- read.table(file=url(filepath), header=T, sep="\t")
mydata

```
> filepath <- "https://dl.dropbox.com/u/1648032/ggplot2_tutorial_dataset.txt"
> mydata  <-  read.table(file=url(filepath), header=T, sep="\t")
https:// URLs are not supported by the default method: using "wininet"
> mydata
           Tribe Hab     BM var1
1   Aepycerotini   L  56.25 36.5
2   Aepycerotini   L  56.25 40.9
3   Aepycerotini   L  56.25 37.0
4   Aepycerotini   L  56.25 36.2
5   Aepycerotini   L  56.25 36.6
6   Aepycerotini   L  56.25 37.7
7   Aepycerotini   L  56.25 37.3
8   Aepycerotini   L  56.25 39.0
9   Aepycerotini   L  56.25 37.7
10  Aepycerotini   L  56.25 35.3
11   Alcelaphini   O 161.00 45.3
12   Alcelaphini   O 161.00 47.7
```

MongoDB 巨量資料分析 04

必須安裝 ggplot2 套件

```
datn <- read.table(header=TRUE, text='
supp dose length
  OJ  0.5  13.23
  OJ  1.0  22.70
  OJ  2.0  26.06
  VC  0.5   7.98
  VC  1.0  16.77
  VC  2.0  26.14
')

ggplot(data=datn, aes(x=dose, y=length, group=supp,
colour=supp)) +
    geom_line() +
    geom_point()
```

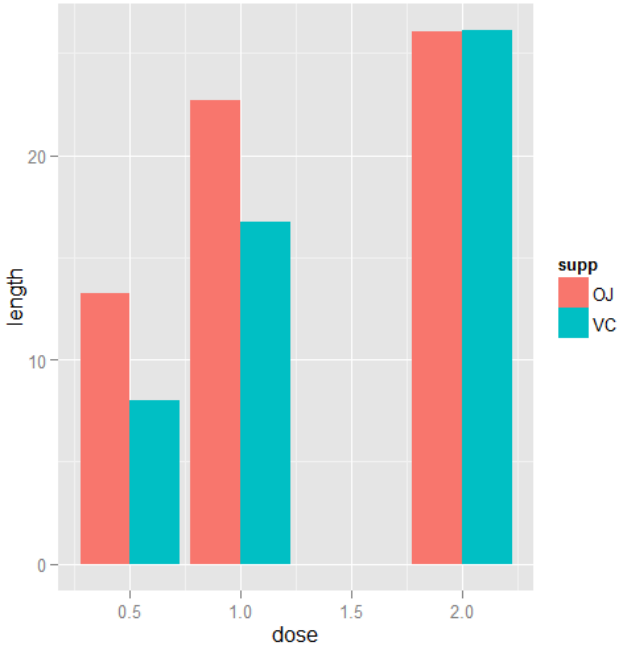



```
> datn <- read.table(header=TRUE, text='
+ supp dose length
+   OJ  0.5  13.23
+   OJ  1.0  22.70
+   OJ  2.0  26.06
+   VC  0.5   7.98
+   VC  1.0  16.77
+   VC  2.0  26.14
+ ')
> ggplot(data=datn, aes(x=dose, y=length, group=supp, colour=supp)) +
+     geom_line() +
+     geom_point()
> |
```

MongoDB 巨量資料分析 05

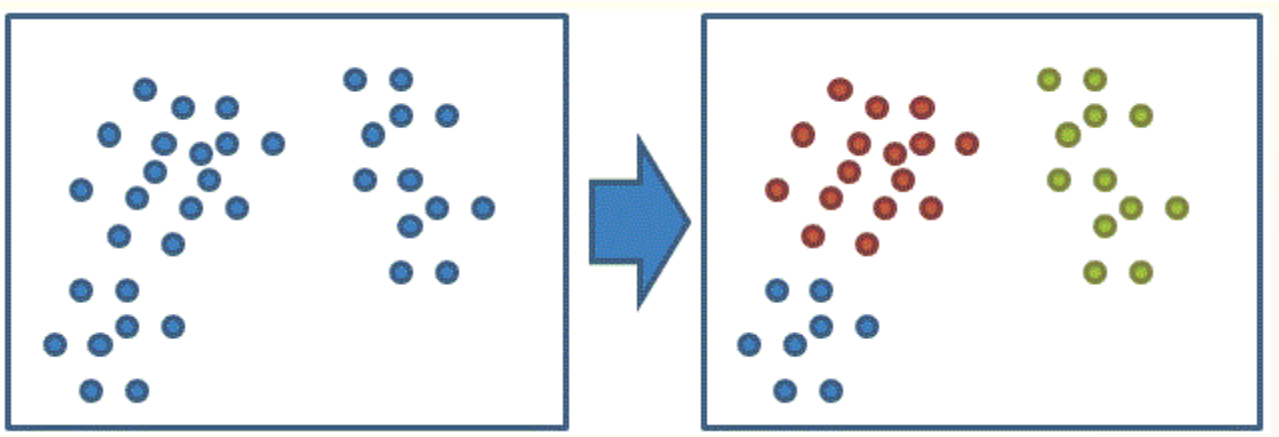
```
datn2 <- datn
ggplot(data=datn2, aes(x=dose, y=length, fill=supp)) +
    geom_bar(stat="identity", position=position_dodge())
```

```
> datn2 <- datn
> ggplot(data=datn2, aes(x=dose, y=length, fill=supp)) +
+     geom_bar(stat="identity", position=position_dodge())
> |
```



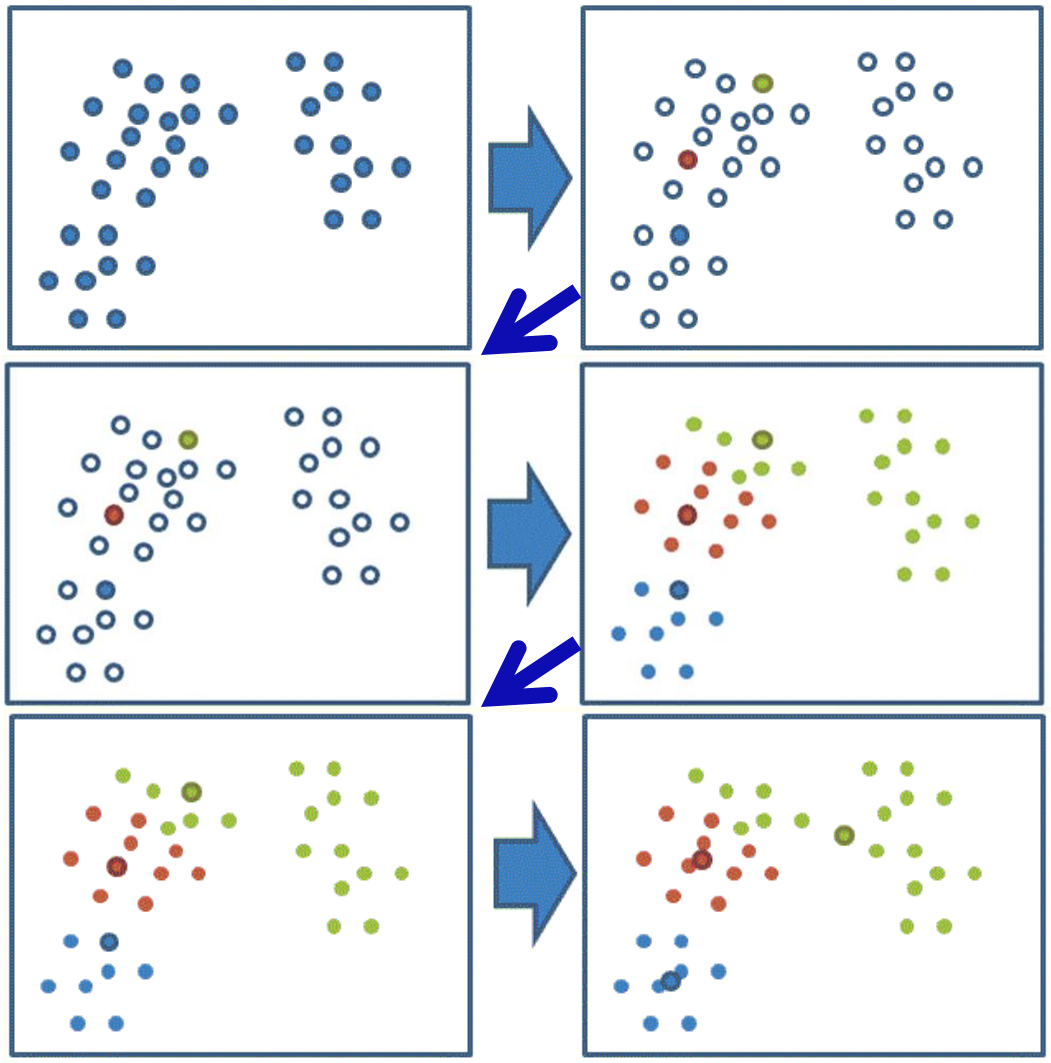
K-means 分群演算法

$$J = \sum_{i=1}^{k} \sum_{x_j \in S_i} \left(x_j - \mu_i \right)^2$$



資料來源: http://www.dotblogs.com.tw/dragon229/archive/2013/02/04/89919.aspx

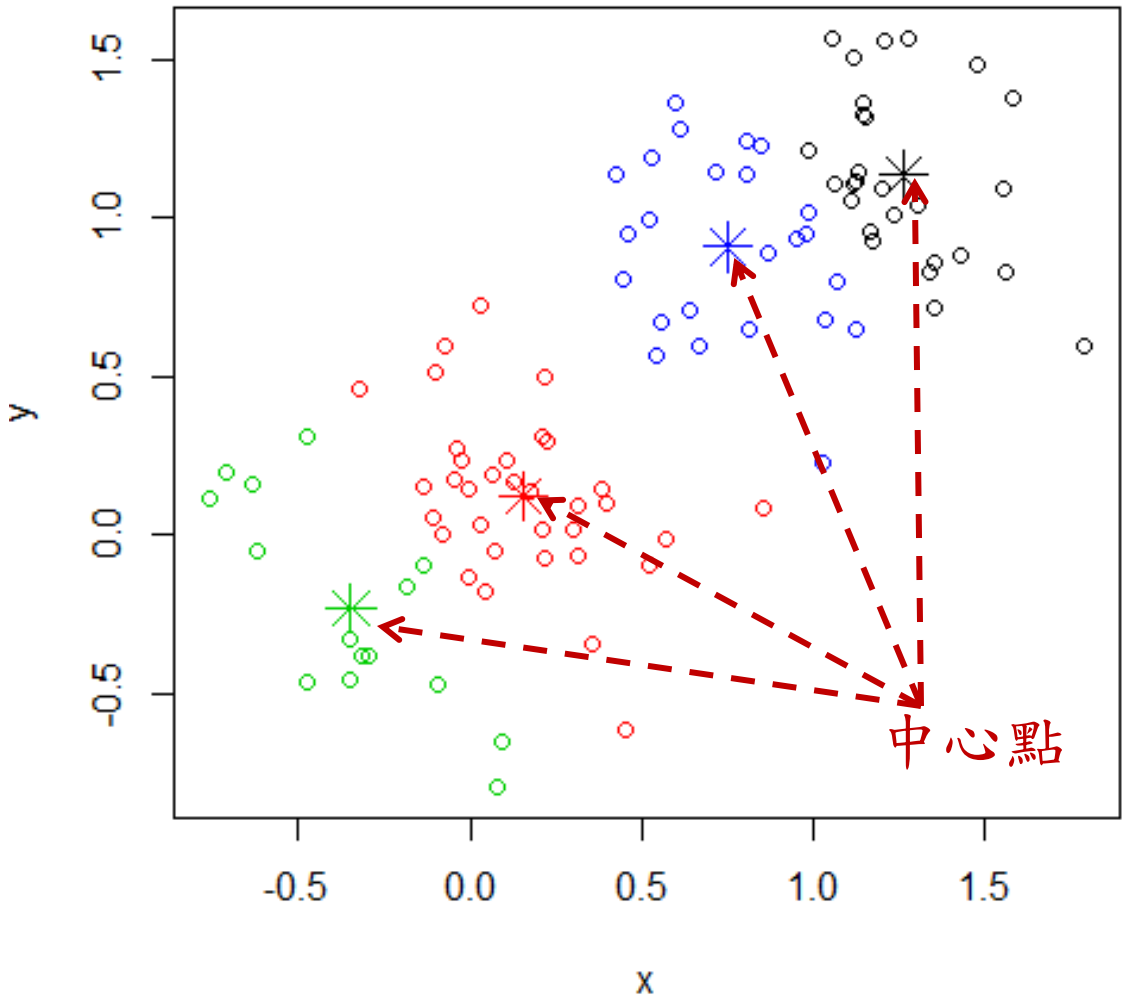
MongoDB 巨量資料分析 07



MongoDB 巨量資料分析 08

```
help(rnorm), help(rbind), help(matrix)


## a 2-dimensional example
x <- rbind(matrix(rnorm(100, sd = 0.3), ncol = 2),
           matrix(rnorm(100, mean = 1, sd = 0.3), ncol = 2))
colnames(x) <- c("x", "y")

(cl <- kmeans(x, 4))

plot(x, col = cl$cluster)
points(cl$centers, col = 1:4, pch = 8, cex=2)

## random starts do help here with too many clusters
(cl <- kmeans(x, 4, nstart = 25))   25組初始亂數中心點
plot(x, col = cl$cluster)
points(cl$centers, col = 1:4, pch = 8, cex=2)
```

MongoDB 巨量資料分析 09



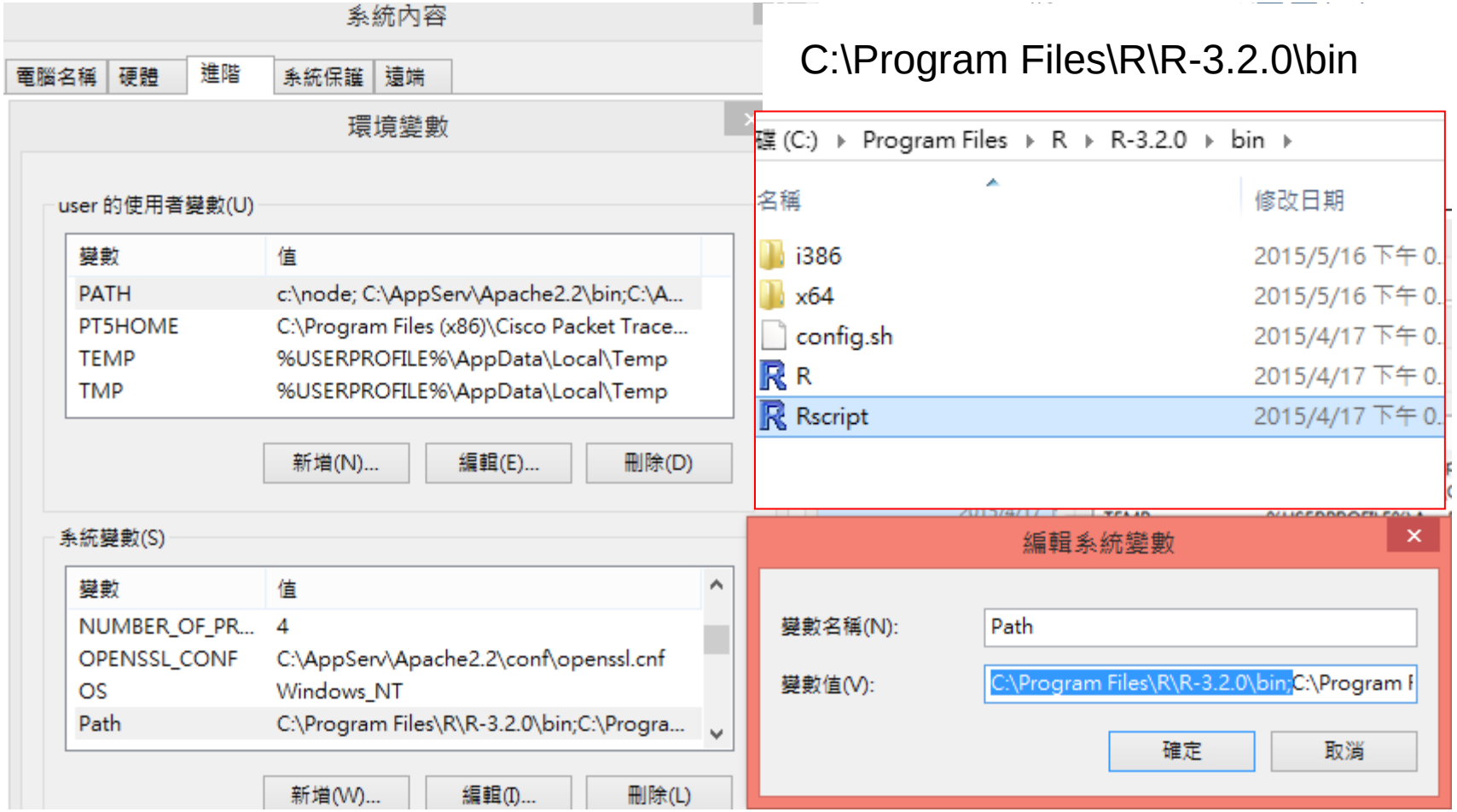
MongoDB 巨量資料分析 10

r1.R

```
args <- commandArgs(TRUE)

N <- args[1]
x <- rnorm(N,0,1)

png(filename="temp.png", width=500, height=500)
hist(x, col="lightblue")
dev.off()
```

MongoDB 巨量資料分析 11

須重新登入後才會生效



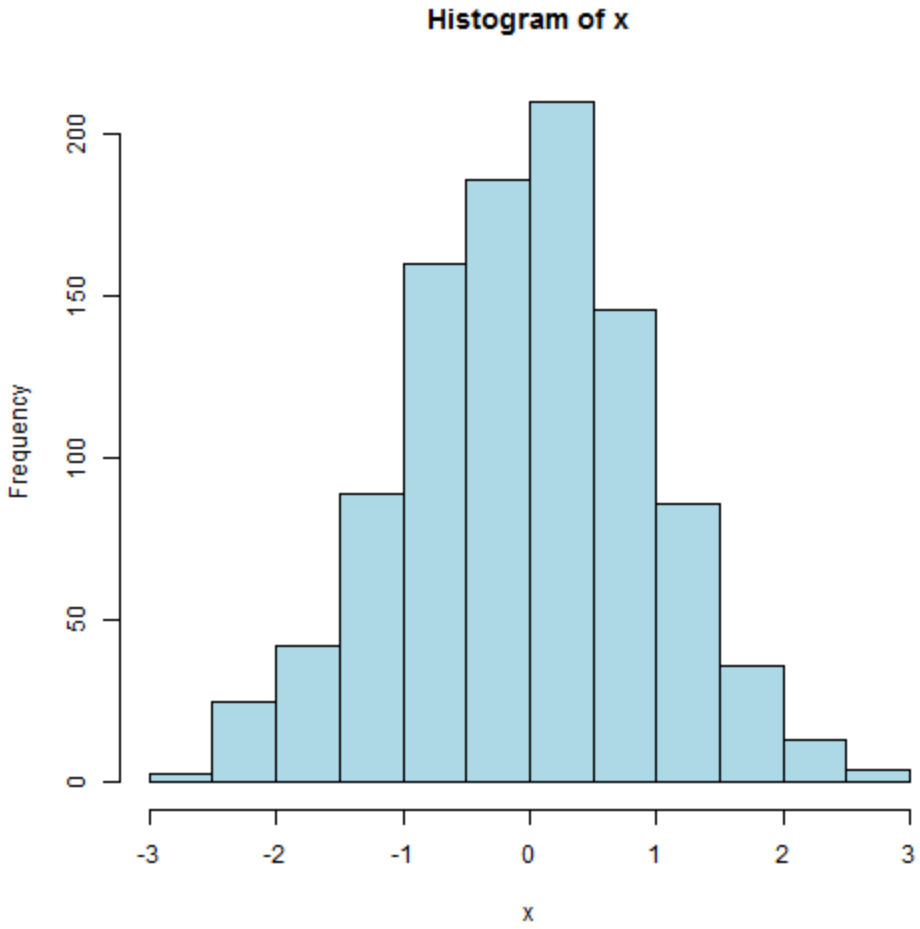
MongoDB 巨量資料分析 12

r1.php

```
<?php
echo "<form action='r1.php' method='get'>";
echo "Number values to generate: <input type='text' name='N' />";
echo "<input type='submit' />";
echo "</form>";

if(isset($_GET['N']))
{
        $N = $_GET['N'];
        shell_exec("rscript  r1.R  $N");
        echo("<img src='temp.png' />");
}
?>
```

MongoDB 巨量資料分析 13

Number values to generate: 1000 送出查詢



MongoDB 巨量資料分析 14

data1 <- read.table("C:/Users/user/customer.txt", header=T, sep=",")

```
> data1 <- read.table("C:/Users/user/customer.txt", header=T, sep=",")
> names(data1)
 [1] "region"         "gender"         "age"            "edcat"          "jobcat"         "employ"         "income"
 [8] "jobsat"         "marital"        "pets_cats"      "pets_dogs"      "pets_birds"     "pets_small"     "pets_saltfish"
[15] "pets_freshfish" "homeown"        "cardspent"      "card2spent"
> dim(data1)
[1] 100  18
> head(data1)
  region gender age edcat jobcat employ income jobsat marital pets_cats pets_dogs pets_birds pets_small pets_saltfish pets_freshfish
1      1      1  20     3      1      0     31      1       0         0         0          0          0            0             0
2      5      0  22     4      2      0     15      1       0         0         0          0          0            0             6
3      3      1  67     2      2     16     35      4       1         2         1          0          0            0             0
4      4      0  23     3      2      0     20      2       1         0         0          0          0            0             0
5      2      0  26     3      2      1     23      1       1         0         0          0          0            0             0
6      4      0  64     4      3     22    107      2       0         1         1          0          2            0             7
  homeown cardspent card2spent
1       0     81.66      67.80
2       1     42.60      34.94
3       1    184.22     175.75
4       1    340.99      18.42
5       0    255.10     252.73
6       1    228.27       0.00
```

MongoDB 巨量資料分析 15

```
> install.packages("rmongodb")
Installing package into 'C:/Users/user/Documents/R/win-library/3.2'
(as 'lib' is unspecified)
--- Please select a CRAN mirror for use in this session ---
also installing the de

嘗試 URL 'http://cran.
Content type 'applicat
downloaded 3.0 MB

嘗試 URL 'http://cran.
Content type 'applicat
downloaded 986 KB

嘗試 URL 'http://cran.
Content type 'applicat
downloaded 1.1 MB

嘗試 URL 'http://cran.
Content type 'applicat
downloaded 1.1 MB

package 'Rcpp' success
package 'jsonlite' suc
package 'plyr' success
package 'rmongodb' suc

The downloaded binary
        C:\Users\user\AppData\Local\Temp\Rtmp2rKJHk\downloaded_packages
>
```

```
> library("rmongodb")
> mongoabc <- mongo.create(db="blog", username="user01", password="1qaz")
> mongoabc
[1] 0
attr(,"mongo")
<pointer: 0x000000002194b2d0>
attr(,"class")
[1] "mongo"
attr(,"host")
[1] "127.0.0.1"
attr(,"name")
[1] ""
attr(,"username")
[1] "user01"
attr(,"password")
[1] "1qaz"
attr(,"db")
[1] "blog"
attr(,"timeout")
[1] 0
> mongo.is.connected(mongoabc)
[1] TRUE
>
```

1. 新增帳號密碼
2. MongoDB 必須切換到 MONGODB-CR 認證方式

MongoDB 巨量資料分析 16

MongoDB

mongoimport -u user01 -p 1qaz --db blog --collection jsonTable3 --file C:\Users\user\zips.json

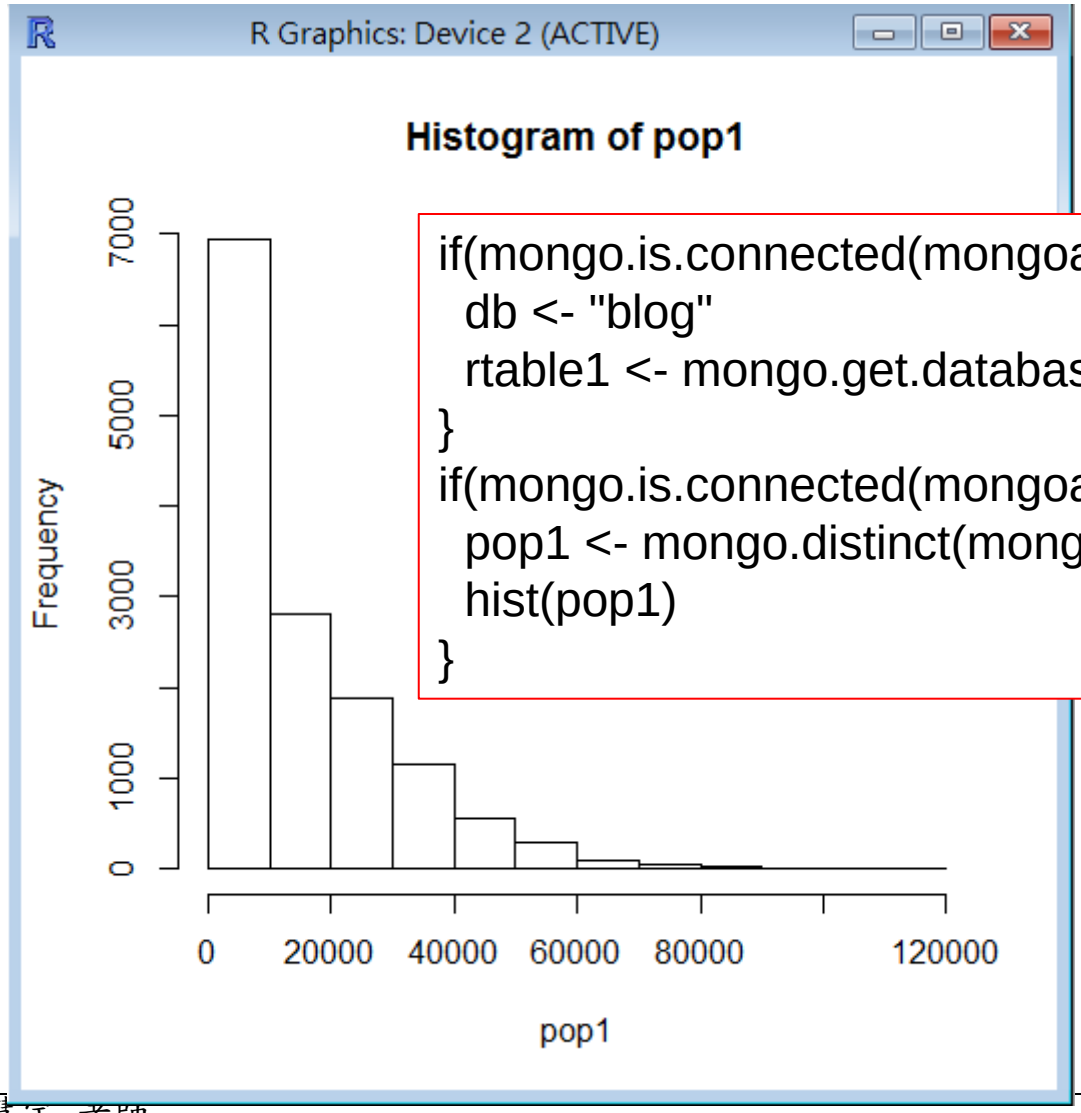
```
C:\Program Files\MongoDB\Server\3.0\bin>mongoimport -u user01 -p 1qaz --db blog
--collection jsonTable3 --file C:\Users\user\zips.json
2015-11-08T23:29:05.454+0800    connected to: localhost
2015-11-08T23:29:08.344+0800    [########################] blog.jsonTable3
3.0 MB/3.0 MB (100.0%)
2015-11-08T23:29:08.364+0800    imported 29353 documents

C:\Program Files\MongoDB\Server\3.0\bin>
```

R

```
if(mongo.is.connected(mongoabc) == TRUE) {
  db <- "blog"
  mongo.get.database.collections(mongoabc, db)
}
```

```
> if(mongo.is.connected(mongoabc) == TRUE) {
+   db <- "blog"
+   mongo.get.database.collections(mongoabc, db)
+ }
[1] "blog.jsonTable3"
>
```

MongoDB 巨量資料分析 17




```
if(mongo.is.connected(mongoabc) == TRUE) {
  db <- "blog"
  rtable1 <- mongo.get.database.collections(mongoabc, db)
}
if(mongo.is.connected(mongoabc) == TRUE) {
  pop1 <- mongo.distinct(mongoabc, rtable1, "pop")
  hist(pop1)
}
```

謝 謝 指 教