HW2 Answer

Boyie Chen

10/12/2019

```
head(data) #take a glance
```

```
##   DogLover BirthYear BirthMonth BirthDay  IQ Difficulty NumFamilyMemb
## 1        1      2000          8       11   0          5             4
## 2        0      2000          7       31 150          5             2
## 3        0      2000          8       30  78          1            12
## 4        0      2000          7       21 200          5             3
## 5        1      1999          7       25 100          3             4
## 6        0      1999          7       21 200          1            20
##   pi_millionth_digit GuessNumber
## 1                  9           0
## 2                  3          87
## 3                  6           8
## 4                  8          68
## 5                  7          73
## 6                  8          45
```

1.

資料表中是否有無法處理的欄位？請試著處理 NA 值，並另外儲存你的資料表為 csv 檔，最後附於附錄中

Method 1: Drop the NAs row by row in each column

```
data2 = data[!is.na(data$DogLover),] #we first drop NA in the first vlb
data2 = data2[!is.na(data2$NumFamilyMemb),] #then the second
data2 = data2[!is.na(data2$pi_millionth_digit),] #no more NAs
```

Method 1 Alternative way: Drop All the rows that includes NAs

```
df = na.omit(data)
```

Method 2:

We know there are NAs in `DogLover`, `NumFamilyMemb` and `pi_millionth_digit`. We'll use functions such as `sum(x, na.rm = T)` to avoid NAs that interfere the numeric calculation

eg. use `na.rm = T`

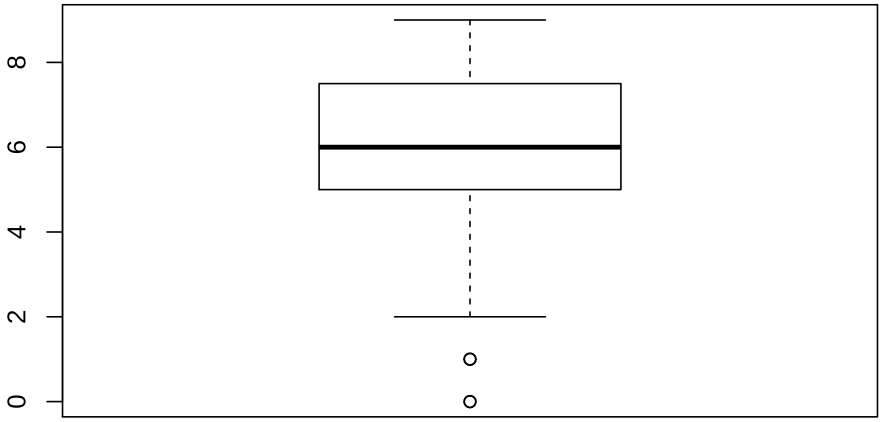
```
#The following functions with argument 'na.rm = T' won't be affected by the NAs
sum(data$DogLover, na.rm = T)
```

```
## [1] 48
```

```
mean(data$DogLover, na.rm = T)
```

```
## [1] 0.5783133
```

```
boxplot(data$pi_millionth_digit, na.rm = T)
```



2.

請將資料表的 row 按照出生年月日排序（由時點遠到近）

hint: 先按照年排序，接著將同一年份的 row(sub dataframe) 宣告為另一 data frame。在其內排序後取代原先的 row，如此便排序完年與月，重複此步驟即可排序年月日。

Method 1: follow the hint

Cons: 產生很多中間變數

```
#Year
data2 = data2[order(data2$BirthYear),] # 先排完年

#Month: 區分成兩年做
data_1999 = data2[data2$BirthYear == 1999,] # 取同一年的子資料表
data_1999 = data_1999[order(data_1999$BirthMonth),] # 在其內排序月
data2[data2$BirthYear == 1999,] = data_1999 # 丟回去原資料表，現在是年排序好且 1999 年的月也排序好的狀態

data_2000 = data2[data2$BirthYear == 2000,] # 取 2000 年的子資料表
data_2000 = data_2000[order(data_2000$BirthMonth),] # 在 2000 年內排序月
data2[data2$BirthYear == 2000,] = data_2000 # 丟回，現在`data2`是年月都排好的狀態

#Day: 區分成兩年的各兩個月做-> 要做四次
data_9907 = data2[data2$BirthYear == 1999 & data2$BirthMonth == 7,] # 把 1999 年 7 月的 row 取出來
data_9907 = data_9907[order(data_9907$BirthDay),] # 按照日排序
data2[data2$BirthYear == 1999 & data2$BirthMonth == 7,] = data_9907 # 丟回

data_9908 = data2[data2$BirthYear == 1999 & data2$BirthMonth == 8,] # 把 1999 年 8 月的 row 取出來
data_9908 = data_9908[order(data_9908$BirthDay),] # 按照日排序
data2[data2$BirthYear == 1999 & data2$BirthMonth == 8,] = data_9908 # 丟回

data_0007 = data2[data2$BirthYear == 2000 & data2$BirthMonth == 7,] # 把 2000 年 7 月的 row 取出來
data_0007 = data_0007[order(data_0007$BirthDay),] # 按照日排序
data2[data2$BirthYear == 2000 & data2$BirthMonth == 7,] = data_0007 # 丟回

data_0008 = data2[data2$BirthYear == 2000 & data2$BirthMonth == 8,] # 把 2000 年 8 月的 row 取出來
data_0008 = data_0008[order(data_0008$BirthDay),] # 按照日排序
data2[data2$BirthYear == 2000 & data2$BirthMonth == 8,] = data_0008 # 丟回
```

```r
head(data2)
```

```
##    DogLover BirthYear BirthMonth BirthDay  IQ Difficulty NumFamilyMemb
## 5         1      1999          7        2 200          5             4
## 6         0      1999          7        3 150          5             4
## 8         1      1999          7        7 150          4             3
## 9         0      1999          7       10 180          5             6
## 10        1      1999          7       12   0          1             4
## 11        1      1999          7       12   0          1             4
##    pi_millionth_digit GuessNumber
## 5                   7          67
## 6                   6          30
## 8                   9          15
## 9                   9          15
## 10                  5          74
## 11                  5          74
```

Method 2: Sort by date-month-year

```r
data = na.omit(data)
df2 = data[order(data$BirthDay),]
df2 = df2[order(df2$BirthMonth),]
df2 = df2[order(df2$BirthYear),]
head(df2)
```

```
##    DogLover BirthYear BirthMonth BirthDay  IQ Difficulty NumFamilyMemb
## 8         1      1999          7        2 200          5             4
## 54        0      1999          7        3 150          5             4
## 53        1      1999          7        7 150          4             3
## 42        0      1999          7       10 180          5             6
## 72        1      1999          7       12   0          1             4
## 83        1      1999          7       12   0          1             4
##    pi_millionth_digit GuessNumber
## 8                   7          67
## 54                  6          30
## 53                  9          15
## 42                  9          15
## 72                  5          74
## 83                  5          74
```

Method 3: Actually `order()` can simply do the job...

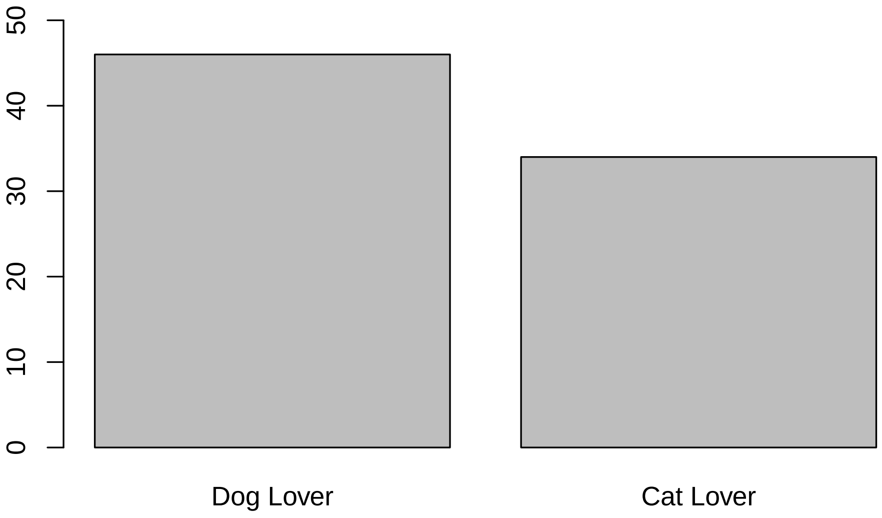
```r
df = df[order(df$BirthYear, df$BirthMonth, df$BirthDay),]
head(df) #exactly the same with above
```

```
##    DogLover BirthYear BirthMonth BirthDay  IQ Difficulty NumFamilyMemb
## 8         1      1999          7        2 200          5             4
## 54        0      1999          7        3 150          5             4
## 53        1      1999          7        7 150          4             3
## 42        0      1999          7       10 180          5             6
## 72        1      1999          7       12   0          1             4
## 83        1      1999          7       12   0          1             4
##    pi_millionth_digit GuessNumber
## 8                   7          67
## 54                  6          30
## 53                  9          15
## 42                  9          15
## 72                  5          74
## 83                  5          74
```

3.

請畫出貓派與狗派人數的 barplot

```r
barplot(c(sum(df$DogLover), length(df$DogLover) - sum(df$DogLover)),
        names.arg =c('Dog Lover','Cat Lover'), ylim = c(0,50))
```

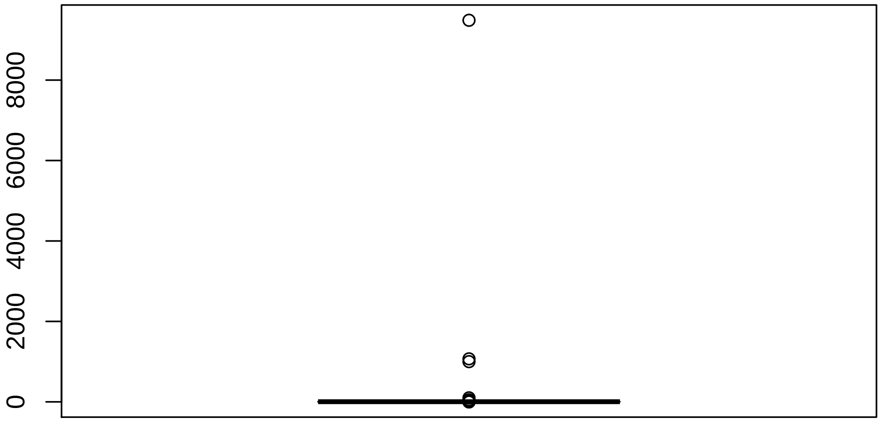


4.

畫出「家中有幾位成員」的 boxplot，mean, Q1, Q3 為何？

Method 1: show the unmodified boxplot

```r
boxplot(df$NumFamilyMemb)
```

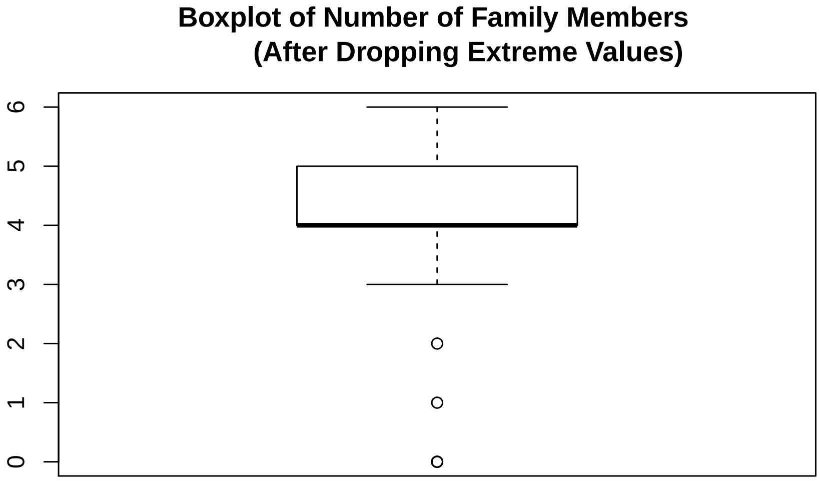


Method 2: drop the extreme values

```
summary(df$NumFamilyMemb) #take a look at the distribution.
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     4.0     4.0   152.1     6.0  9487.0
```

```
#say the extreme values are those exceed median+1.5*IQR
upperLimit = 4+1.5*(6-4)
NumFamilyMemb = df$NumFamilyMemb[df$NumFamilyMemb < 7] #drop the extreme values
boxplot(NumFamilyMemb,
        main = 'Boxplot of Number of Family Members
        (After Dropping Extreme Values)')
```

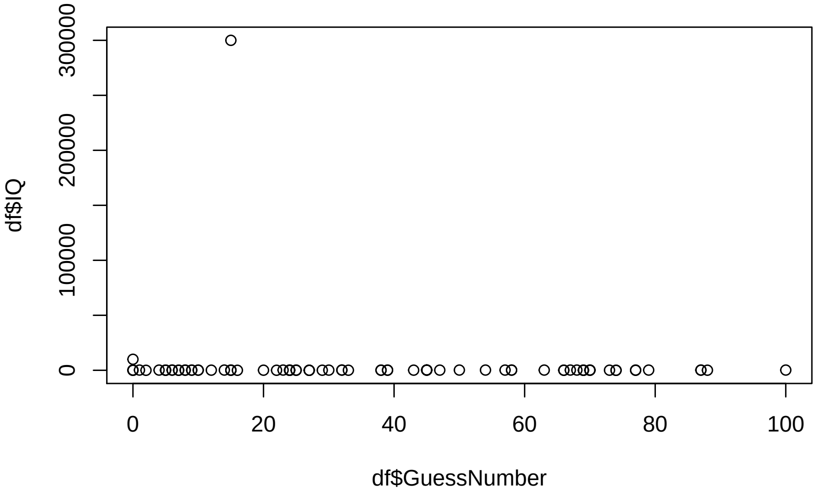


5.

在「終極密碼」題中，繪製以終極密碼答案為橫軸，智商為縱軸的 scatter plot

Method 1: show the unmodified scatter plot

```
plot(df$GuessNumber, df$IQ)
```

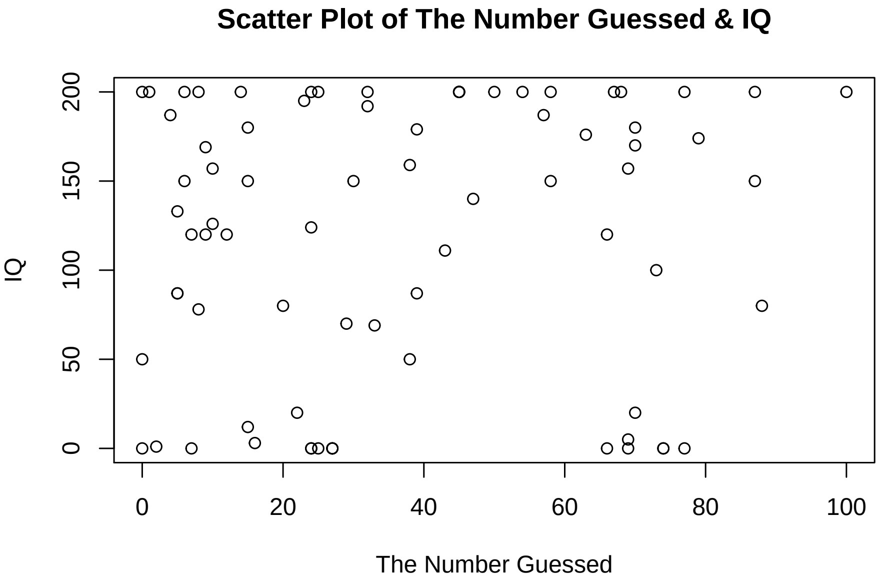


Method 2: drop the extreme values

```
summary(df$IQ)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##       0      50     150    3990     200  300000
```

```
#say the extreme values are those exceed median+1.5*IQR
upperLimit = 150+1.5*(200-50)
df2 = df[df$IQ <= upperLimit,] #drop the extreme values
plot(df2$GuessNumber, df2$IQ,
     xlab = 'The Number Guessed', ylab = 'IQ',
     main = 'Scatter Plot of The Number Guessed & IQ')
```


Scatter Plot of The Number Guessed & IQ
IQ
The Number Guessed