Homework1

陳凱騫(H24101222)

2024-09-19

Table of contents

Import data 1
Data Analysis 2
    Type of variables(Q1) 2
    Visualize the data(Q2) 2
    Summarize data(Q3) 6
    Symmetric or not ?(Q4) 7
    Outliers(Q5) 8

Import data

```r
data <- read.csv(file = "iris.csv")
head(data)
```

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

Data Analysis

Type of variables(Q1)

```
for (x in 1:length(data)) {
  print(class(data[,x]))
}
```

```
[1] "numeric"
[1] "numeric"
[1] "numeric"
[1] "numeric"
[1] "character"
```

```
data$Species <- factor(data$Species)
# change the class of the variable of the species
class(data$Species)
```

```
[1] "factor"
```

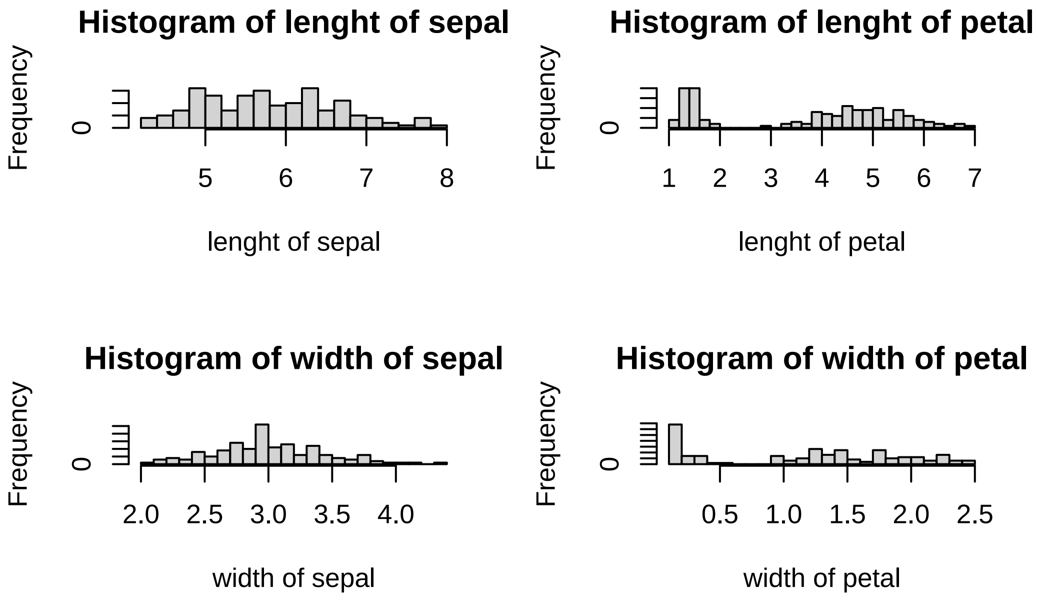
從上方的程式和table可以知道其各個variables如下表:

Variables	Type
Sepal.Length	Cardinal
Sepal.Width	Cardinal
Petal.Length	Cardinal
Petal.Width	Cardinal
Species	Nominal

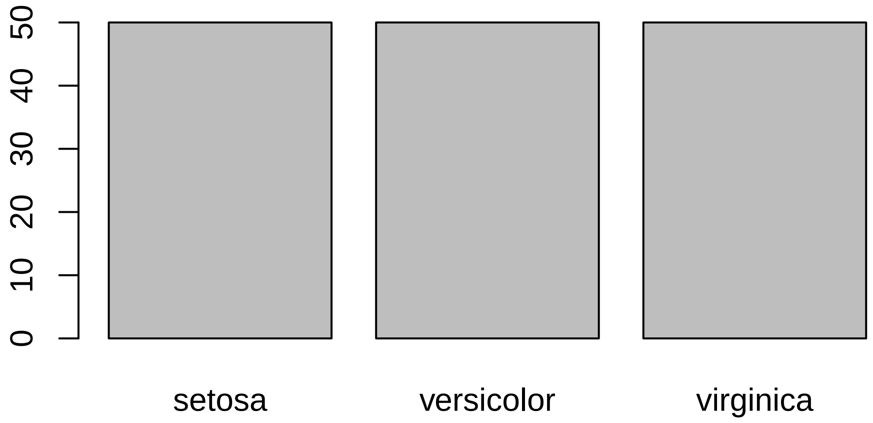
Visualize the data(Q2)

```
par(mfcol=c(2,2))
hist(data$Sepal.Length,main = "Histogram of lenght of sepal"
     ,xlab = "lenght of sepal",breaks = 25)
hist(data$Sepal.Width,main = "Histogram of width of sepal",
     xlab = "width of sepal",breaks = 25)
```

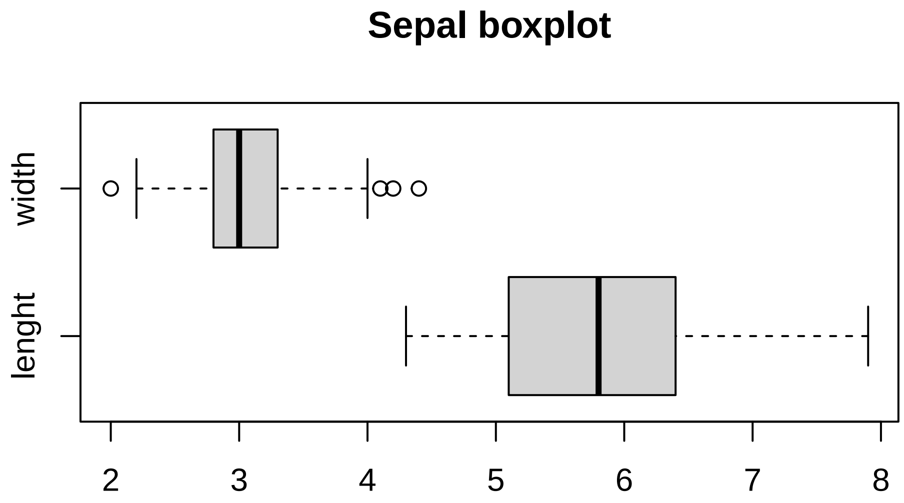
```
hist(data$Petal.Length,main = "Histogram of lenght of petal",
     xlab = "lenght of petal",breaks = 25)
hist(data$Petal.Width,main = "Histogram of width of petal",
     xlab = "width of petal",breaks = 25)
```



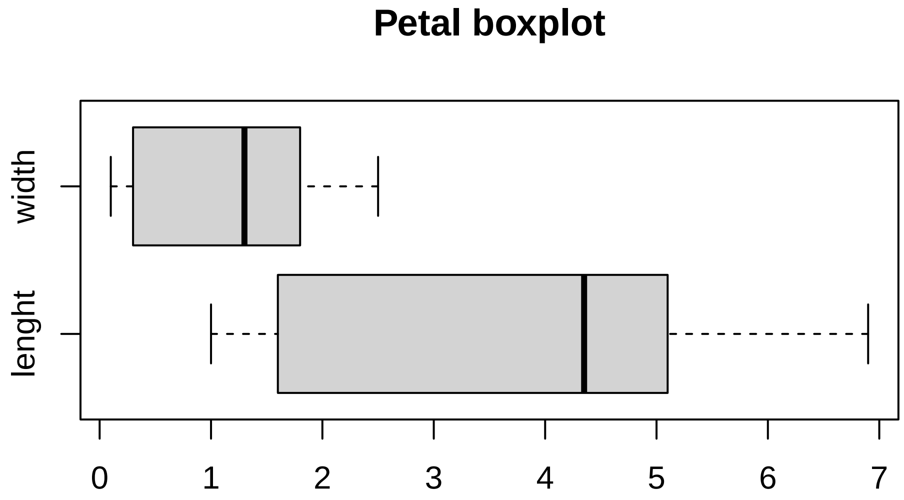
```
par(mfcol=c(1,1))
plot(data$Species)
```



```
# change columns name
sepal <- data[,1:2]
names(sepal) <- c("lenght","width")
petal <- data[,3:4]
names(petal) <- c("lenght","width")
# boxplot
boxplot(sepal,horizontal = T,main="Sepal boxplot")
```



```
boxplot(petal,horizontal = T,main="Petal boxplot")
```



Summarize data(Q3)

- 只做cardinal資料

```
# library
library(datasets)
library(e1071)  # calculate skewness and kurtosis


# only cardinal data be considered
numeric_vars <- data[, 1:4]

# create table
summary_table <- data.frame(
  Mean = apply(numeric_vars, 2, mean),
  Median = apply(numeric_vars, 2, median),
  Variance = apply(numeric_vars, 2, var),
  Std_Dev = apply(numeric_vars, 2, sd),
  Range = apply(numeric_vars, 2, function(x) diff(range(x))),
  IQR = apply(numeric_vars, 2, IQR),
  Skewness = apply(numeric_vars, 2, skewness),
  Kurtosis = apply(numeric_vars, 2, kurtosis)
)
# print result
print(summary_table)
```

```
                 Mean Median  Variance   Std_Dev Range IQR   Skewness
Sepal.Length 5.843333   5.80 0.6856935 0.8280661   3.6 1.3  0.3086407
Sepal.Width  3.057333   3.00 0.1899794 0.4358663   2.4 0.5  0.3126147
Petal.Length 3.758000   4.35 3.1162779 1.7652982   5.9 3.5 -0.2694109
Petal.Width  1.199333   1.30 0.5810063 0.7622377   2.4 1.5 -0.1009166
               Kurtosis
Sepal.Length -0.6058125
Sepal.Width   0.1387047
Petal.Length -1.4168574
Petal.Width  -1.3581792
```

- nominal data only need to know mode and we know that all of three species are the same, 50.

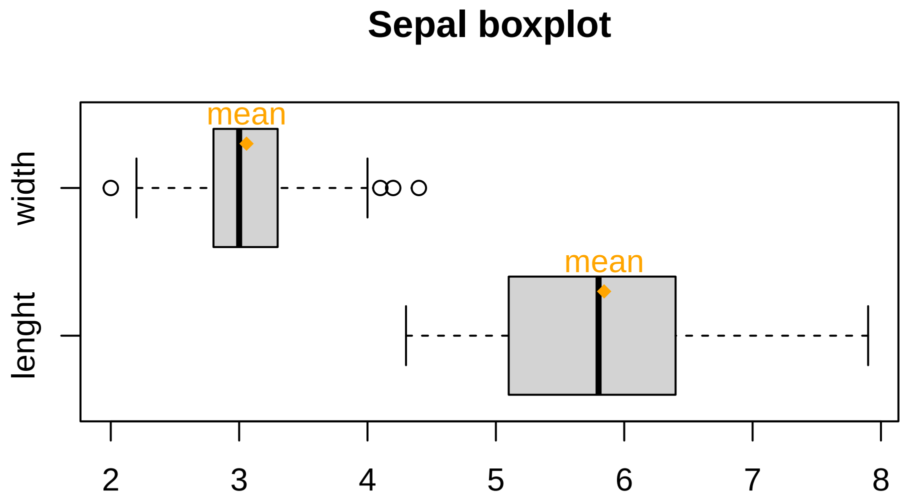
```
summary(data$Species)
```

```
    setosa versicolor  virginica 
        50         50         50 
```

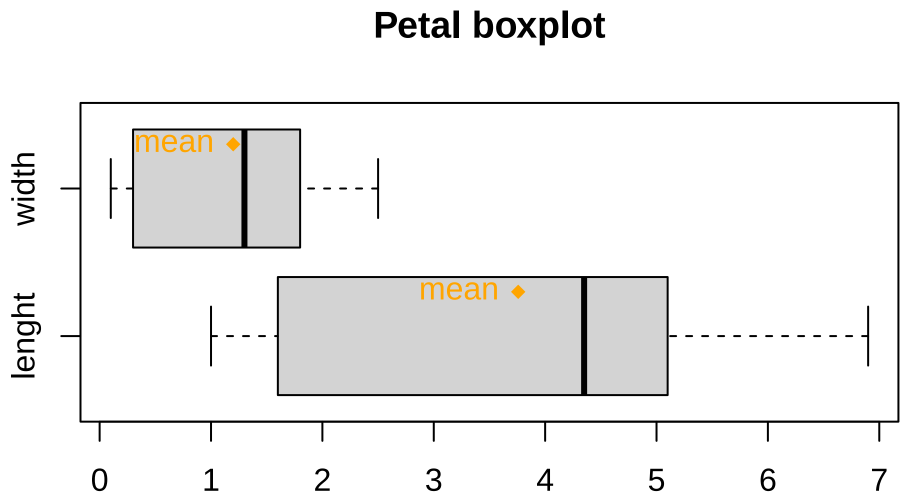
Symmetric or not ?(Q4)

使用boxplot來決定是否對稱

```
# boxplot
means <- apply(sepal, 2, mean)
meanp <- apply(petal, 2, mean)
bs <- boxplot(sepal,horizontal = T,main="Sepal boxplot")
xi <- 0.3 + seq(bs$n)
points(means,xi , col = "orange", pch = 18)
text(means,xi,"mean",pos=3,col = "orange")
```



```
bp <- boxplot(petal,horizontal = T,main="Petal boxplot")
xi <- 0.3 + seq(bp$n)
points(meanp,xi , col = "orange", pch = 18)
text(meanp,xi,"mean",pos=2,col = "orange")
```



從上方幾張**boxplot**可知，**sepal**的lenght和width基本上為**對稱**，至於**petal**的lenght和width**不對稱**，原因為mean和median之間有差距，且median看起來似乎也比較不在資料的中心。而從visualize data 那裡的histogram也可發現資料(petal)似乎有一點雙峰的感覺。

Outliers(Q5)

從上方Sepal boxplot中可以發現sepal.width有4個極端值。

```
outliers <- c()
outliers <- c(outliers,sort(data$Sepal.Width)[1])
for (i in 0:2) {
  outliers <- c(outliers,sort(data$Sepal.Width)[length(data$Sepal.Width)-i])
}
outliers
```

```
[1] 2.0 4.4 4.2 4.1
```

從中可知其outliers為:

Outliers
2
4.4
4.2
4.1