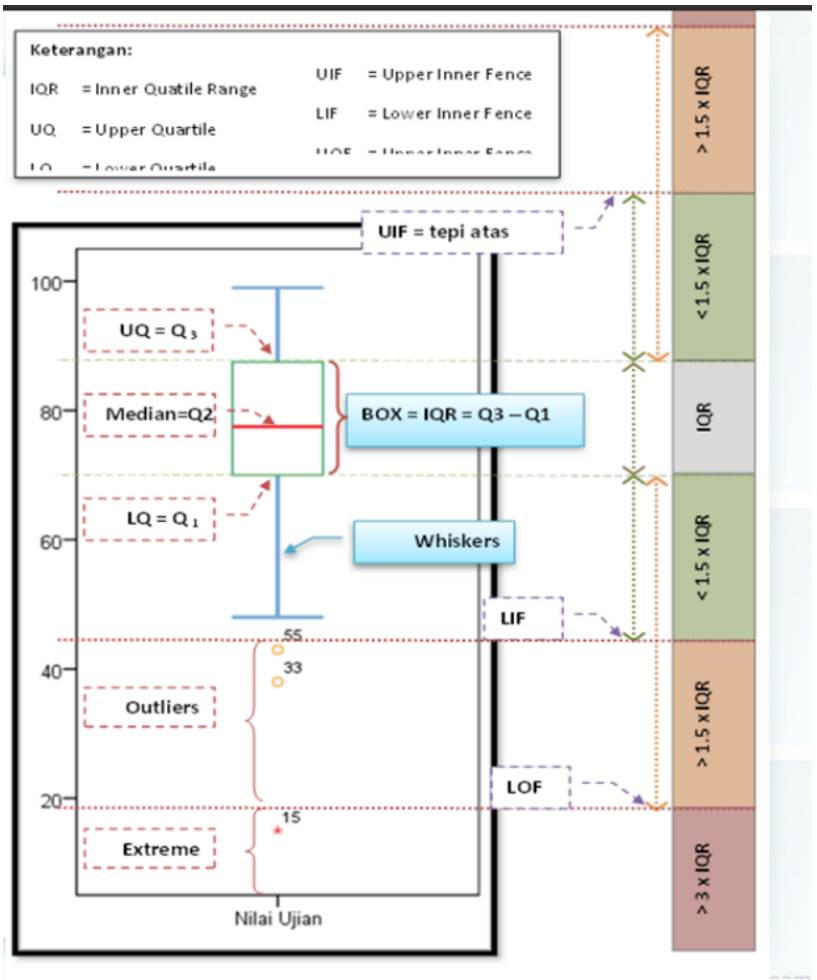
Praktikum Eksplorasi dan Visualisasi Data
Pertemuan 3
PENGGUNAAN RINGKASAN NUMERIK: BOXPLOT DAN STANDARDISASI

Ringkasan numerik merupakan ringkasan dari data, yang merupakan harga-harga yang penting dari data, atau yang dapat memberikan gambaran dari data. Seperti yang telah diketahui sebelumnya, ringkasan numerik terdiri dari ukuran pusat dan ukuran sebaran. Ringkasan numerik inilah yang nantinya akan kita gunakan dalam membuat boxplot serta melakukan standardisasi.

A. BOXPLOT

Boxplot adalah diagram kotak dan titik yang menyajikan ringkasan numerik data. Tujuannya untuk membandingkan beberapa angkatan melalui bentuk diagramnya (ringkasan numerik, bentuk distribusi, dan sebaran data/kesimetrisan/juraian), serta mendeteksi adanya outlier (nilai ekstrim). Berdasarkan jumlah angkatan, boxplot dibedakan menjadi dua, yakni boxplot untuk satu angkatan dan boxplot untuk lebih dari satu angkatan. Boxplot dibuat menggunakan ringkasan numerik lima angka yaitu nilai maksimum, Q3, median, Q1, dan minimum.

Struktur dari boxplot dapat dilihat seperti gambar dibawah :



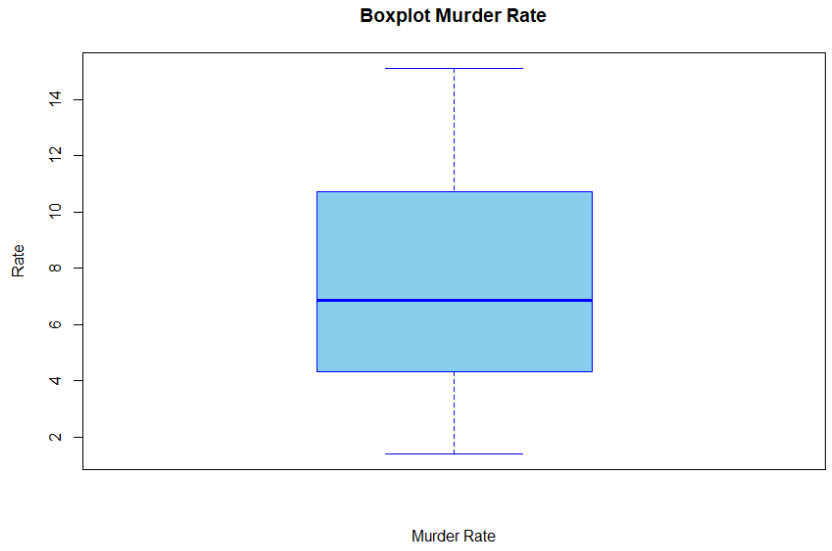
Terdapat 5 ukuran statistik yang bisa kita baca dari boxplot, yaitu:

1. Nilai minimum: nilai observasi terkecil
2. Q1: kuartil terendah atau kuartil pertama
3. Q2: median atau nilai pertengahan
4. Q3: kuartil tertinggi atau kuartil ketiga
5. Nilai maksimum: nilai observasi terbesar.

BOXPLOT SATU ANGKATAN

```
#menyiapkan data

help("state.x77")

data=state.x77

data=as.data.frame(data)

#membuat boxplot

boxplot(data$Murder, xlab="Murder Rate", ylab="Rate",

        main="Boxplot Murder Rate",col="skyblue",border = "blue")
```

NB. Syntax utama `boxplot(data)`

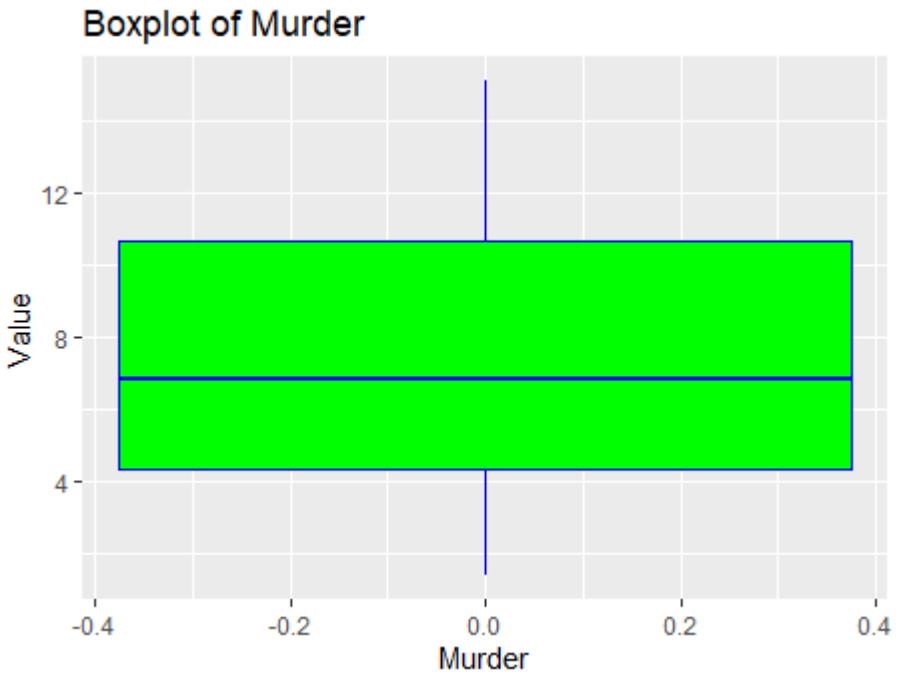


Boxplot dengan package ggplot2

```
library(ggplot2)

ggplot(data, aes(y=Murder))+

  geom_boxplot(fill="green",color="blue")+
```

```
  ggtitle("Boxplot of Murder")+
  xlab("Murder")+
  ylab("Value")
```

Syntax utama `ggplot(data, aes(y=nama_kolom))+ geom_boxplot()`

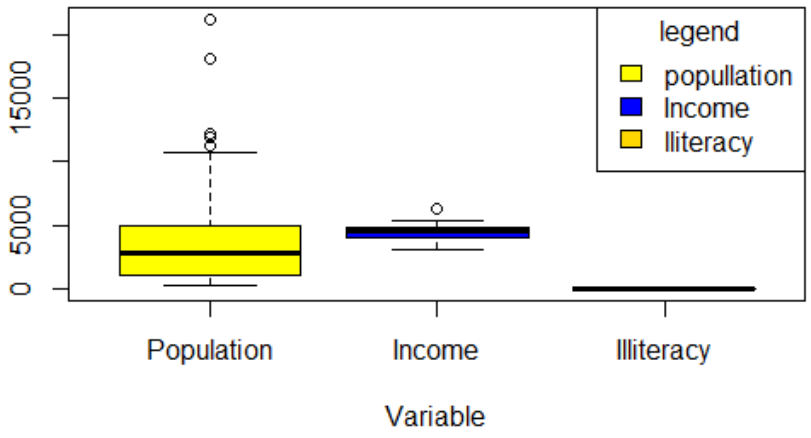


Boxplot dengan package plotly

```
#boxplot dengan menggabungkan data

#menyiapkan data

library(tidyverse)

data1=select(data,Population,Area)

data2=data1 %>% gather(key=variabel,

                  value=Value)

#membuat boxplot

boxplot(data2$Value, xlab="Murder Rate", ylab="Rate",

         main="Boxplot Murder Rate",col="skyblue",border =
"blue")
```

Boxplot Lebih dari Satu Angkatan

```
#Dengan Library Standar

boxplot(data[1:3],xlab="Variable", col=c("yellow","blue","gold"))

legend("topright",legend =c("popullation","Income","lliteracy"),

       fill=c("yellow","blue","gold"),cex=1,title = "legend")
```



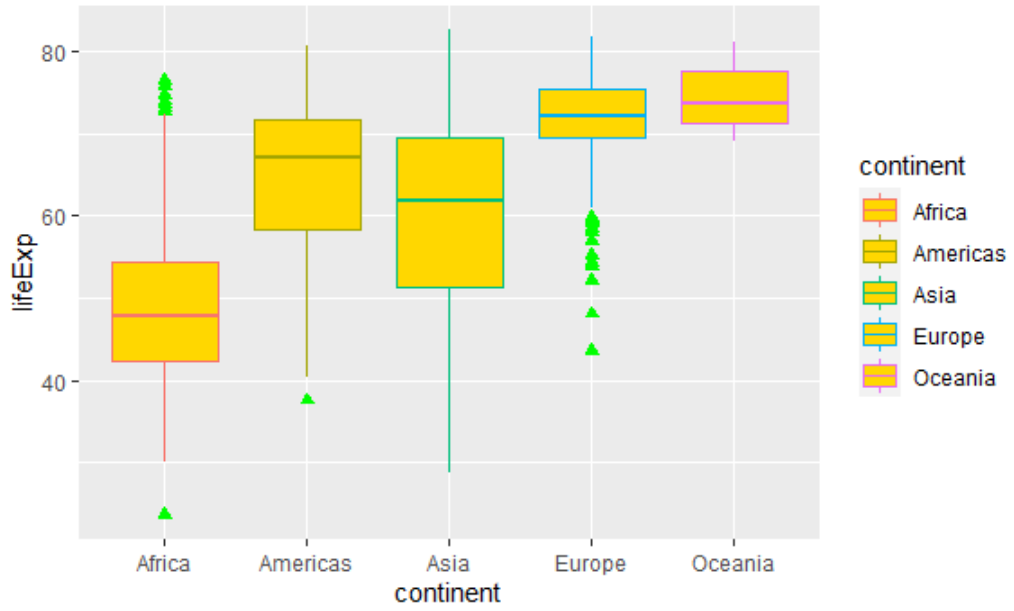
#Dengan Paket ggplot2

#menyiapkan data

```
library(gapminder)

mydat=gapminder
```

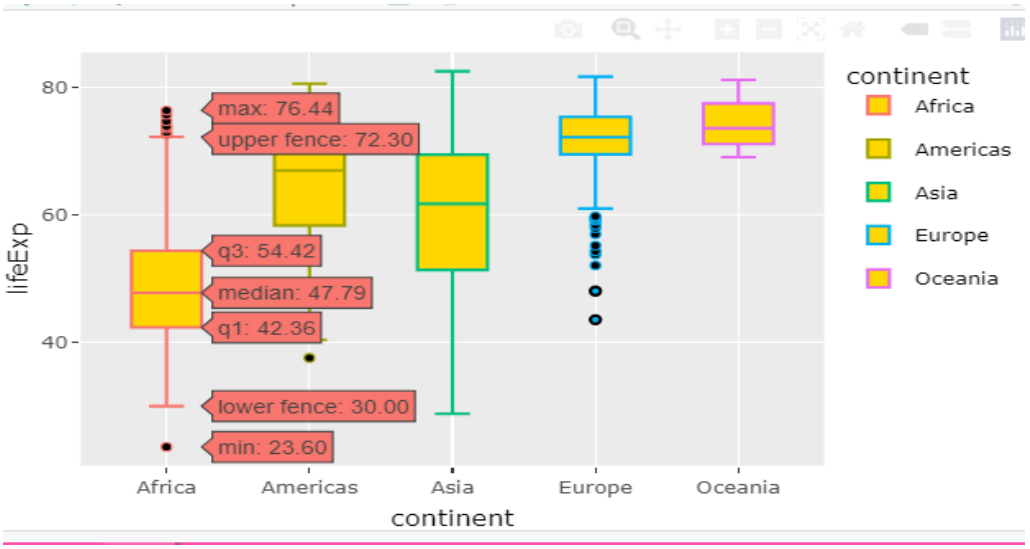
#membuat boxplot

```
ggplot(mydat,aes(continent,lifeExp,color=continent))+
  geom_boxplot(fill="gold",outlier.color = "green",
               outlier.size=2,outlier.shape=17)
```



#plotly

```
yeayy=ggplot(mydat,aes(continent,lifeExp,color=continent))+
  geom_boxplot(fill="gold",outlier.color = "green",
               outlier.size=2,outlier.shape=17)
ggplotly(yeayy)
```



Interpretasi

- Ringkasan numerik (tergantung paket yang digunakan),
- Ada tidaknya outlier,
- bentuk distribusi (normal atau menceng), dan sebaran datanya (juraian)
- Kehomogenan data

B. STANDARISASI

Dalam statistika inferensi salah satu asumsi yang kerap diperlukan adalah bahwa angkatan berdistribusi normal, dimana melalui analisis data eksploratif kita ketahui bahwa bentuk angkatan memegang peranan penting dalam menentukan apakah suatu angkatan berdistribusi normal atau tidak. Akan tetapi, terkadang bentuk ini tidak terlalu terlihat karena tertutup oleh pusat dan sebaran, oleh karena itu dilakukan standardisasi. Standardisasi adalah proses mengeluarkan pusat dan sebaran observasi.

Tujuan dilakukannya standardisasi:

a. Mempermudah dalam melihat bentuk angkatan
b. Memudahkan dalam membandingkan beberapa angkatan.

Proses standarisasi adalah dengan mengurangkan pusat dari tiap observasi dalam angkatan,lalu membaginya dengan sebarannya. Misalkan kita memiliki data dengan tiap observasidilambangkan dengan x maka, setelah dilakukan standarisasi kita memiliki observasi baru Z melalui cara berikut:

$$z = \frac{x - Pusat}{Sebaran}$$

Nantinya setelah dilakukan standarisasi akan dimiliki angkatan baru yang memiliki pusat 0 dan sebaran 1. Pasangan ukuran pusat dan sebaran harus digunakan secara bersama agar observasi baru memiliki pusat 0 dan sebaran 1. Pasangan ukuran pusat dan sebaran adalah sebagai berikut:

Ukuran Pusat	**Sebaran**
Rata-rata	Standar Deviasi
Median	Range
Median	IQR
Trirata	Range
Trirata	IQR

Contoh Pembuktian

Dimiliki angkatan dengan tiap observasi dilambangkan dengan x, akan dilakukan standarisasi dengan mengurangi tiap observasi dengan rata-ratanya, lalu membaginya dengan standar deviasinya.

$$Zi = \frac{x_i - \overline{x}}{sd(x)}$$

Pusat baru dari angkatan

$$\bar{Z} = \frac{\sum x_i}{n}$$

$$\bar{Z} = \frac{\sum \frac{x_i - \bar{x}}{sd(x)}}{n}$$

$$\bar{Z} = \frac{1}{sd(x)n}\left(\sum x_i - \bar{x}\right)$$

$$\bar{Z} = \frac{1}{sd(x)n}\left(\sum x_i - \sum \bar{x}\right)$$

$$\bar{Z} = \frac{1}{sd(x)n}\left(\sum x_i - n\frac{\sum x_i}{n}\right)$$

$$\bar{Z} = \frac{1}{sd(x)n}\left(\sum x_i - \sum x_i\right)$$

$$\bar{Z} = 0$$

Sebaran baru dari angkatan

$$Var(Z) = \frac{\sum(Z_i - \bar{Z})^2}{n-1}$$

$$Var(Z) = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{\sqrt{var(x)}} - 0\right)^2$$

$$Var(Z) = \frac{1}{(n-1)var(x)}\sum(x_i - \bar{x})^2$$

$$Var(Z) = \frac{1}{var(x)}\frac{\sum(x_i - \bar{x})^2}{n-1}$$

$$Var(Z) = \frac{var(x)}{var(x)}$$

$$Var(Z) = 1$$

$$SD(Z) = \sqrt{var(Z)} = 1$$

Maka, didapatkan observasi baru dengan pusat 0 dan sebaran 1.
Langkah-langkah dalam melakukan standarisasi adalah sebagai berikut:

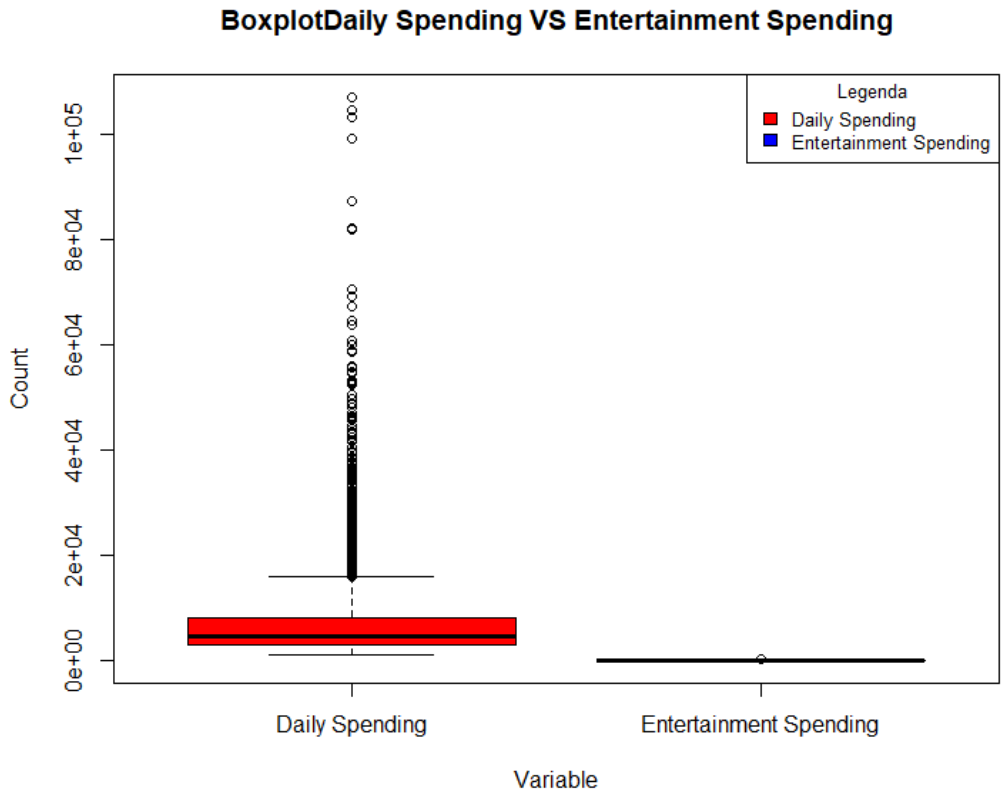
1. Tentukan pasangan ukuran pusat dan sebaran yang akan digunakan dalam melakukan standardisasi. Ukuran pusat dan ukuran sebaran ini dapat dicari menggunakan cara sebelumnya yaitu:

Syntax :

```
# Read data
library(readxl)
dataku <- read_excel("D:\\Kuliah\\Praktikum Eksplorasi dan Visualisasi Data\\Pertemuan 3\\Pertemuan 4.xlsx")

# Check Data
head(dataku)


boxplot(dataku,xlab="Variable",ylab="Count",main="BoxplotDaily Spending VS Entertainment Spending",
      col = c("red","blue"))

legend(legend=c("Daily Spending","Entertainment Spending"),
     col = c("red","blue"),'topright',
     cex=0.8,fill=c("red","blue"),title="Legenda")

# Summary
summary(dataku)
IQR(dataku$`Daily Spending`)
IQR(dataku$`Entertainment Spending`)
```

Output :

```
> head(dataku)
# A tibble: 6 x 2
  `Daily Spending` `Entertainment Spending`
             <dbl>                    <dbl>
1             3800                     19
2             3200                     15.7
3            29800                     30
4             1100                      5.6
5             6100                     31.7
6             5900                     29.4
```




```
> summary(dataku)
 Daily Spending   Entertainment Spending
 Min.   :   900   Min.   : 4.20
 1st Qu.:  2800   1st Qu.:14.00
 Median :  4500   Median :22.30
 Mean   :  6951   Mean   :30.21
 3rd Qu.:  8000   3rd Qu.:39.65
 Max.   :107000   Max.   :99.90
> IQR(dataku$`Daily Spending`)
[1] 5200
> IQR(dataku$`Entertainment Spending`)
[1] 25.65
```

2. Setelah didapatkan ukuran pusat dan sebaran yang akan digunakan barulah kita dapat melakukan standardisasi, menggunakan software R :

   Syntax :

```
# Standarisasi
standarisasi <- function(data,pusat,sebaran){
 z=(data-pusat)/sebaran
 return(z)
}
```

```
# Standarisasi Daily Spending
std_Dailyspd <- standarisasi(dataku$`Daily Spending`,
                  median(dataku$`Daily Spending`),
                  IQR(dataku$`Daily Spending`))
head(std_Dailyspd)

# Standarisasi Entertainment Spending
std_Entairspd <- standarisasi(dataku$`Entertainment Spending`,
                  median(dataku$`Entertainment Spending`),
                  IQR(dataku$`Entertainment Spending`))
head(std_Entairspd)

# Menggunakan Packages
######## Warning Tidak Semua Pusat dan Sebaran Digunakan ###########
library(robustHD)
std_data <- standardize(dataku, centerFun = median, scaleFun = IQR)

head(std_data)
```

Output :

```
> head(std_Dailyspd)
[1]   -0.1346154   -0.2500000      4.8653846   -0.6538462
0.3076923  0.2692308

> head(std_Entairspd)
[1]   -0.1286550   -0.2573099      0.3001949   -0.6510721
0.3664717  0.2768031

> head(std_data)
  Daily Spending Entertainment Spending
1     -0.1346154             -0.1286550
2     -0.2500000             -0.2573099
3      4.8653846              0.3001949
4     -0.6538462             -0.6510721
5      0.3076923              0.3664717
6      0.2692308              0.2768031
```

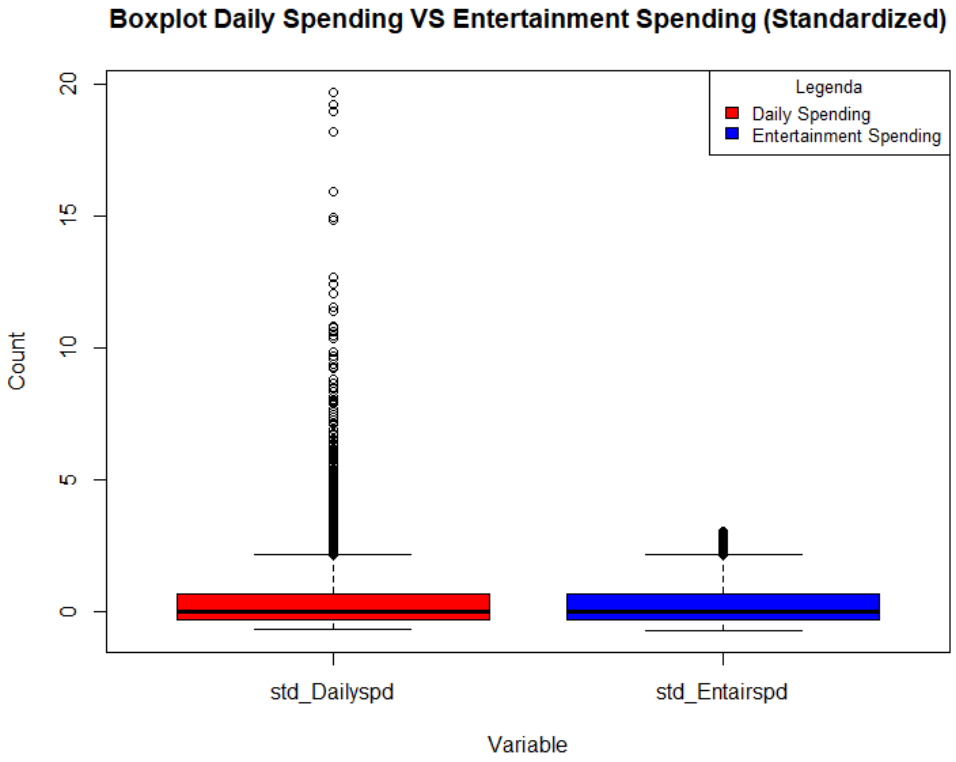
3. Setelah selesai, kita dapat membuat boxplot untuk melihat bentuk angkatan setelah dilakukan standardisasi.

Syntax :

```
# Boxplot Hasil Standarisasi
data_terstd <- data.frame(std_Dailyspd,std_Entairspd)
```

```
boxplot(data_terstd,xlab="Variable",ylab="Count",
      main="Boxplot Daily Spending VS Entertainment Spending (Standardized)",
      col = c("red","blue"))

legend(legend=c("Daily Spending","Entertainment Spending"),
     col = c("red","blue"),'topright',
     cex=0.8,fill=c("red","blue"),title="Legenda")
```

Output :



Catatan : *Dalam praktikum ini karena kita menggunakan boxplot untuk menggambarkan sebaran data, maka standardisasi yang akan lebih banyak kita gunakan adalah dengan mengurangi dengan median dan membagi dengan IQR (Interquartile Range).*

Latihan :

1. Disajikan data tingkat kematian laki-laki karena penyakit jantung di beberapa kota di Jawa Tengah dan DIY sebagai berikut :

Kota	Usia 20-29 tahun	Usia 30-39 tahun	Usia 40-49 tahun	Usia 50-59 tahun
Bantul	24	30	39	46

Sleman	7	7	19	29
Brebes	20	19	26	39
Tegal	29	39	44	52
Kendal	15	27	34	40
Batang	27	35	50	54
Cilacap	47	66	82	83
Blora	6	10	14	30
Rembang	7	10	14	31
Klaten	25	28	37	56
Sragen	4	6	27	45
Pati	26	40	48	48
Kudus	21	32	34	62
Kebumen	8	11	19	26
Demak	19	22	33	41

Buatlah boxplot terhadap keempat angkatan tersebut, tentukan ukuran pusat dan ukuran numerik yang akan digunakan untuk standarisasi, berikan alasannya. Kemudian lakukan standardisasi. Bandingkan hasil boxplot sebelum dan sesudah dilakukan standardisasi !

2. Disajikan data jumlah kasus demam berdarah disuatu kota dari bulan Januari sampai April, ***data berada di file P-3.xlsx.*** Dari data tersebut lakukan standardisasi untuk membandingkan jumlah kasus demam berdarah perbulan. Kemudian jawablah pertanyaan berikut,
    a. Jelaskan distribusi masing-masing angkatan dari data jumlah kasus demam berdarah disuatu kota dari bulan Januari sampai April.
    b. Bulan apa yang memiliki jumlah kasus demam berdarah yang paling heterogen? Jelaskan alasanmu
    c. Bulan apa yang memiliki jumlah kasus demam berdarah yang paling homogen? Jelaskan alasanmu

Nb. Jawaban pertanyaan bisa ditinjau dari boxplot setelah standardisasi