Praktikum Eksplorasi dan Visualisasi Data
Pertemuan 4
Transformasi Data

A. Transformasi

Transformasi data adalah suatu cara untuk Membuat distribusi angkatan mendekati distribusi normal jika dilihat dari bentuk sebarannya, agar memenuhi asumsi yang kerap diperlukan dalam inferensi statistik atau analisis data konfirmasi.

Ciri Distribusi Normal

- Berbentuk seperi lonceng (bell shaped curve) dengan kedua ekor yang
- Bentuk distribusi yang simetris terhadap ukuran pusatnya (mean atau median).
- Mean, median, dan modus sama.
- 95% dari data jatuh dalam 2 standar deviasi dari mean.

Dalam melihat kenormalan distribusi dari suatu angkatan, selain menggunakan diagram batang dan daun, juga dapat menggunakan boxplot. **Suatu angkatan sendiri dapat dianggap simetris apabila:**

1. Median berada di tengah-tengah sebaran data.
2. Nisbah (ratio)

$$Nisbah = \frac{Q3 - Med}{IQR} \approx \frac{Med - Q1}{IQR} \approx 0.5$$

Jenis transformasi:

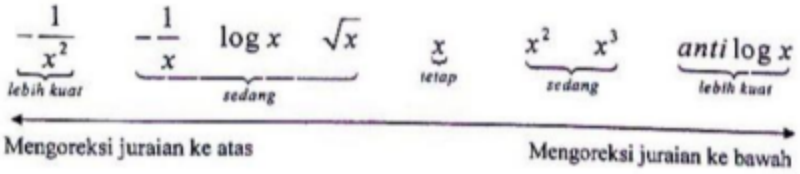
- Transformasi satu angkatan
- Transformasi beberapa angkatan

Beberapa metode transformasi:

- Metode dengan acuan tangga transformasi Tukey
- Metode dengan perkiraan nisbah
- Metode boxcox

B. Transformasi Satu Angkatan

Transformasi satu angkatan dapat dilakukan dengan metode coba-coba dengan acuan tangga transformasi Tukey



Syntax di R

- Transformasi $-\frac{1}{x^2}$

```
tfdata = -1/data$variabel^2
tfdata
```

- Transformasi $-\frac{1}{x}$

```
tfdata = -1/data$variabel
tfdata
```

- Transformasi $log(x)$

```
tfdata = log10(data$variabel)
tfdata
```

- Transformasi $\sqrt{x}$

```
tfdata = sqrt(data$variabel)
tfdata
```

- Transformasi x^2

```
tfdata = data$variabel^2
tfdata
```

- Transformasi $antilog(x)$

```
tfdata = 10^(data$variabel)
tfdata
```

Contoh Soal

CONTOH SOAL TRANSFORMASI SATU ANGKATAN

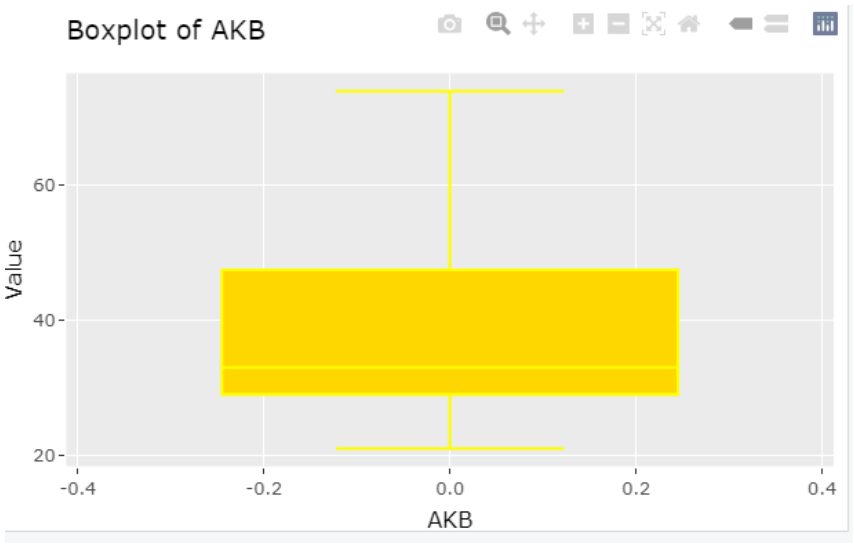
Diberikan data Angka Kematian Bayi per 1000 Kelahiran Hidup menurut Provinsi tahun 2012.

Provinsi	AKB	Provinsi	AKB
ACEH	47	NUSA TENGGARA BARAT	57
SUMATERA UTARA	40	NUSA TENGGARA TIMUR	45
SUMATERA BARAT	27	KALIMANTAN BARAT	31
RIAU	24	KALIMANTAN TENGAH	49
JAMBI	34	KALIMANTAN SELATAN	44
SUMATERA SELATAN	29	KALIMANTAN TIMUR	21
BENGKULU	29	SULAWESI UTARA	33
LAMPUNG	30	SULAWESI TENGAH	58
KEP. BANGKA BELITUNG	27	SULAWESI SELATAN	25
KEP. RIAU	35	SULAWESI TENGGARA	45
DKI JAKARTA	22	GORONTALO	67
JAWA BARAT	30	SULAWESI BARAT	60
JAWA TENGAH	32	MALUKU	36
DI YOGYAKARTA	25	MALUKU UTARA	62
JAWA TIMUR	30	PAPUA BARAT	74
BANTEN	32	PAPUA	54
BALI	29		

Lakukanlah analisis tentang kesimetrisan data. Jika diperlukan lakukan transformasi sehingga diperoleh transformasi yang optimal.

Langkah di R

```
#menyiapkan data
library(openxlsx)
data=read.xlsx("D:\\KULIAH\\Asprak\\P-4.xlsx")

#boxplot sebelum transformasi
library(plotly)
data_boxplot=ggplot(data,aes(y=AKB))+
  geom_boxplot(fill="gold",color="yellow")+
  ggtitle("Boxplot of AKB")+
  xlab("AKB")+ylab("Value")
ggplotly(data_boxplot)
```



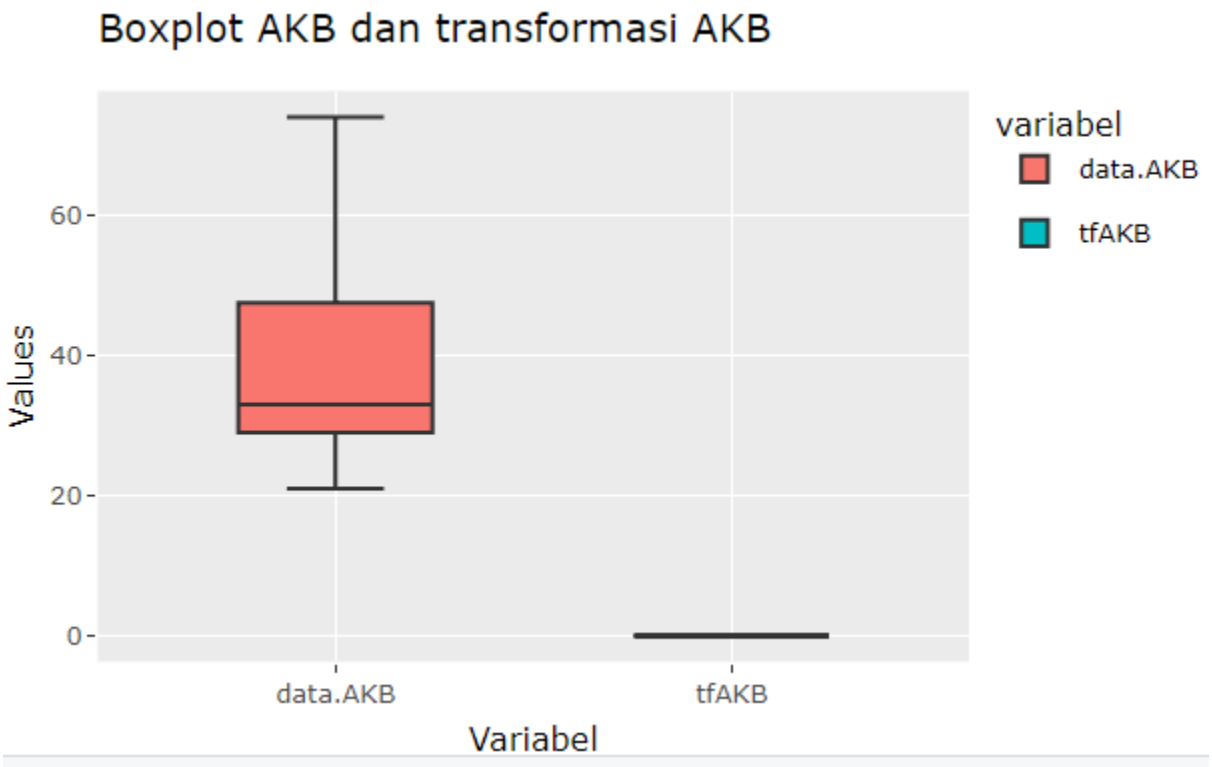
Interpretasi : Dari boxplot didapatkan beberapa nilai. Kuartil atas (Q1) dari data tersebut adalah 29. Median data tersebut adalah 33. Kuartil atas (Q3) data adalah 48. Interquartil range data adalah 19. Secara umum, bentuk boxplot tidak simetris, dengan median tidak tepat berada di tengah boxplot. Jarak median ke kuartil bawah (Q1) lebih dekat daripada jarak median ke kuartil atas (Q3). Atau dengan kata lain, nilai tinggi lebih menyebar dari nilai rendah, keadaan ini disebut "menjurai ke atas". Dari boxplot tersebut, tidak terdapat outliers (nilai ekstrim). Dalam hal ini, data belum bisa dikatakan berdistribusi normal. Sehingga perlu dilakukan transformasi

Transformasi Data AKB

Metode coba – coba dengan acuan tangga transformasi Tukey. Berdasarkan analisis kesimetrisan, disimpulkan bahwa bentuk distribusi data Angka Kematian Bayi menjurai ke atas. Sehingga dapat dipilih transformasi $\sqrt{x}$ atau log x atau $-1/x$ atau $-1/x^2$. Karena boxplot diperoleh cukup jauh menjurai ke atas, dapat dipilih metode Tukey yang lebih kuat.

```
#Membuat transformasi AKB
tfAKB=-1/data$AKB^2

#Membuat dataframe data sebelum dan setelah transformasi
transformasi=data.frame(data$AKB,tfAKB)

#membuat boxplot gabungan dengan ringkasan 5 angka setelah
transformasi
```

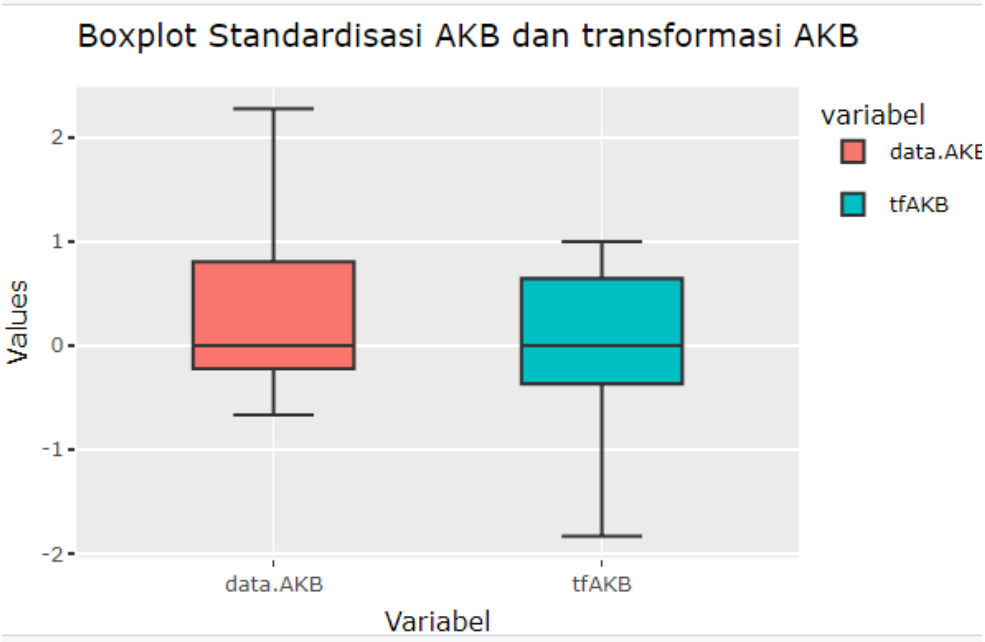
```
#menyatukan kolom data menggunakan data wrangling dengan
dataframe transformasi

library(tidyverse)
data2 = gather(transformasi,key="variabel", value="value")
data2
boxplot_tf = ggplot(data2, aes(x=variabel, y=value,
                    fill=variabel))+geom_boxplot()+
                    ggtitle("Boxplot AKB dan transformasi
AKB")+
                    xlab("Variabel")+ylab("Values")
ggplotly(boxplot_tf)
```



Karena bentuk kesimetrisan belum terlihat jelas, maka dibandingkan hasil standardisasi keduanya.

```
#standarisasi
     library(robustHD)
     standardisasi=standardize(transformasi,centerFun          =
median,scaleFun = IQR)

#boxplot setelah ditransformasi
#menjadikan data menjadi 1 kolom menggunakan data wrangling
```

```
data3 = gather(standardisasi,key="variabel", value="value")
data3
boxplotstd = ggplot(data3, aes(x=variabel, y=value,
                fill=variabel))+geom_boxplot()+
                     ggtitle("Boxplot Standardisasi AKB dan
transformasi AKB")+
                xlab("Variabel")+ylab("Values")
ggplotly(boxplotstd)
```



Interpretasi : Dari boxplot tersebut, dapat dilihat bahwa setelah ditransformasi −1/AKB^2 angkatan menjadi lebih simetris. Terlihat bahwa boxplot sebelum transformasi menjurai ke atas, dengan nilai median yang lebih dekat dengan kuartil bawah (Q3). Sedangkan boxplot setelah transfromasi, mendekati distribusi normal. Oleh karena itu dapat disimpulkan bahwa transformasi −1/AKB^2 optimal, karena distribusi data mendekati normal. Tingkat kesimetrisan dapat dilihat dari nisbahnya. Karena angka desimal dalam boxplot terbatas, dapat membuat summary data untuk menghitung nisbah

```
> #Cek nisbah
> summary(tfAKB)
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-0.0022676 -0.0011891 -0.0009183 -0.0009142 -0.0004527 -0.0001826
```

$$\text{Nisbah} = \frac{Q3 - Median}{IQR} = \frac{-0.0004527 - (-0.0009183)}{0.0007363671} = 0{,}6323$$

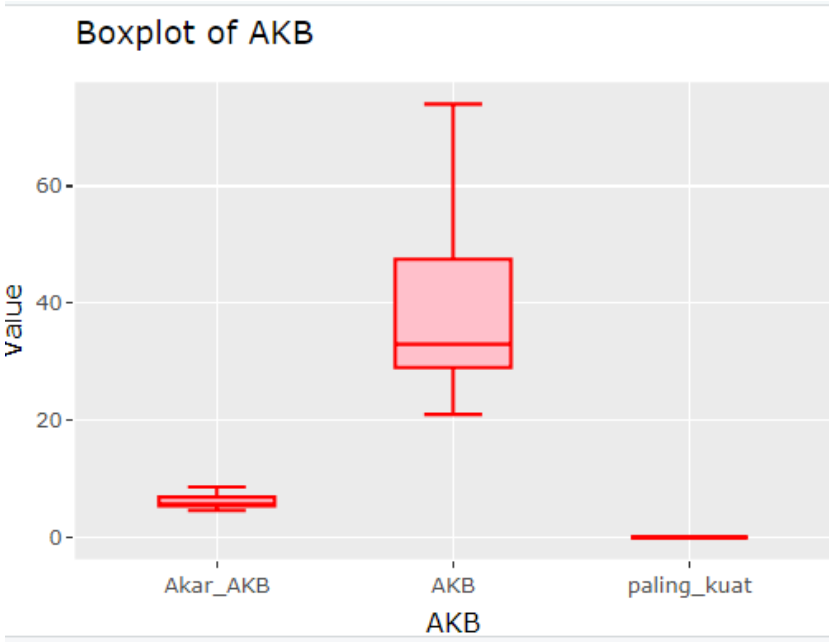
Berdasarkan transformasi yang dilakukan, dapat disimpulkan bahwa transformasi yang optimal yaitu −1/AKB^2, karena distribusi data setelah transformasi mendekati normal.

Jika Ingin tidak dicoba satu-satu???

Langkahnya sama pertama cek juraian distribusinya dengan boxplot, selanjutnya perhatikan syntax berikut

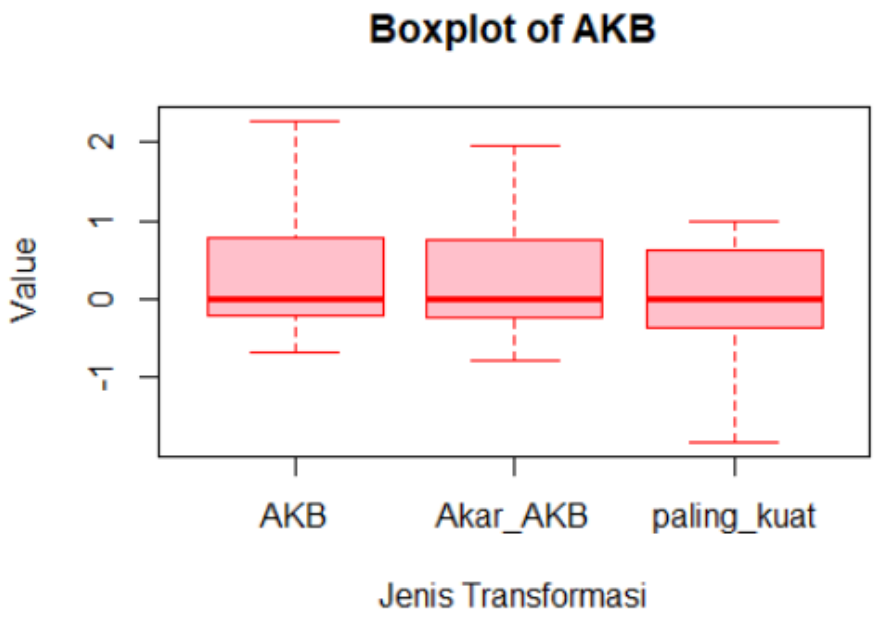
```
#menyiapkan data
library(openxlsx)
data=read.xlsx("D:\\KULIAH\\Asprak\\P-4.xlsx")

#membuat boxplot
library(plotly)
data_boxplot=ggplot(data,aes(y=AKB))+
  geom_boxplot(fill="gold",color="yellow")+
  ggtitle("Boxplot of AKB")+
  xlab("AKB")+ylab("Value")
ggplotly(data_boxplot)
#menjurai ke atas

#transformasi tukey
data1=transform(data,Akar_AKB=sqrt(data$AKB),paling_kuat=-1
/data$AKB^2)
Data1

#membuat boxplot hasil transformasi
#menyiapkan data
library(tidyverse)
data_transformasi=data1%>%
  select(-Provinsi)
data2=data1%>%
  select(-Provinsi)%>%
  gather(data1)

#boxplot
data_boxplot1=ggplot(data2,aes(x=data1,y=value))+
```

```
  geom_boxplot(fill="pink",color="red")+
  ggtitle("Boxplot of AKB")+
  xlab("AKB")+ylab("Value")
ggplotly(data_boxplot1)
```



Untuk memperjelas perbandingan angkatan digunakan standardisasi

```
#standardisasi
library(robustHD)
std_data=standardize(data_transformasi,centerFun  =  median,
scaleFun = IQR)

#boxplot setelah standardisasi
boxplot(std_data[1:3],col="pink",border="red",
        main="Boxplot of AKB", xlab="Jenis Transformasi",
        ylab="Value")

#bisa dicek nisbah ya!
```



Bisa diinterpretasi ya

C. Transformasi Beberapa Angkatan (Nisbah)

Untuk 2 angkatan :

$$Nisbah = \frac{log\ IQR(A) - log\ IQR(B)}{logMed(A) - logMed(B)}$$

Untuk beberapa angkatan secara umum :

Titik A dan B didapatkan dari plot log median vs log IQR. Kemudian titik A dan B dihubungkan hingga membentuk garis yang mendekati semua titik. Tabel jenis transformasi menggunakan nisbah.

Nisbah Kira-Kira	Transformasi
-2	x^3
-1	x^2
0	x
1/2	$\sqrt{x}$
1	log x
3/2	$-1/x$

2	$-1/x^2$

Langkah-Langkah Transformasi

1) Tentukan masing-masing median dan IQR dari setiap angkatan.
2) Plot log median sebagai sumbu (X) vs log IQR sebagai sumbu Y.
3) Cek hasil nisbah berdasarkan nilai
4) Cek hasil nisbah berdasarkan nilai kemiringan dari persamaan regresi yang didapat, kemudian cocokan dengan tabel transformasi nisbah kira-kira.

Langkah-Langkah di R :

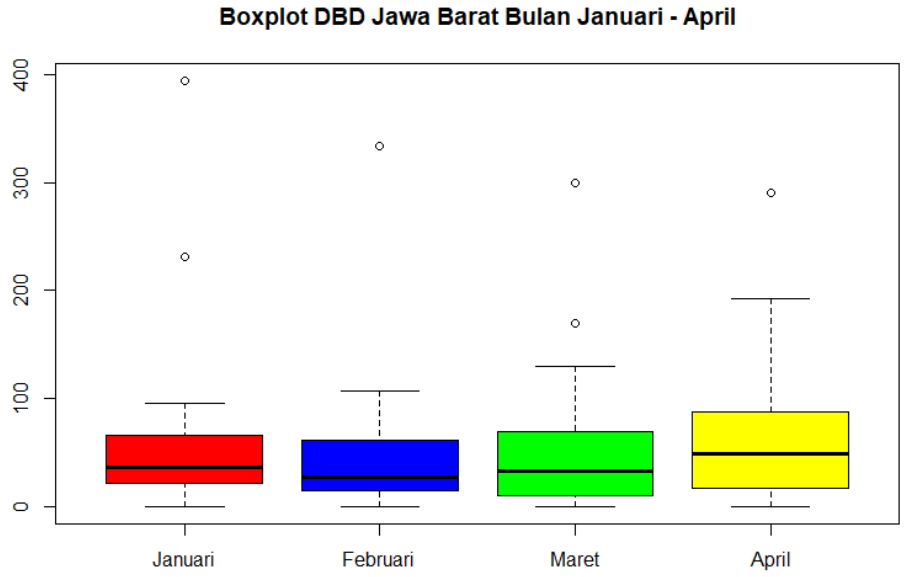
1. Baca data dan buat boxplotnya

   Syntax :

```
# Panggil Library yang dibutuhkan
library(readxl)

# Read data
data <- read_excel(file.choose())

# Check data
head(data)

# pilih variabel yg digunakan
mydata <- data[,-1]

# buat boxplot
boxplot(mydata,col = c("red","blue","green","yellow"),
        main="Boxplot DBD Jawa Barat Bulan Januari -
April")
```

   Output :

```
> head(data)
# A tibble: 6 x 5
  Kabupaten   Januari Februari Maret April
  <chr>         <db1>    <db1> <db1> <db1>
1 Bogor           231      105   130    66
2 Sukabumi         42       47    13   290
3 Cianjur          18       11     8    11
4 Bandung          25        0   169   114
5 Garut            30       18    33    49
6 Tasikmalaya      27       11     8    43
```



2. Tentukan masing masing median dan IQR dari setiap angkatan

Syntax :

```
# Tentukan masing median dan IQR dari setiap angkatan
library(tidyverse)
ringkas <- mydata %>%
  gather(key="Bulan",value = "Counts") %>%
  group_by(Bulan) %>%
  summarise(iqr=IQR(Counts),Med=median(Counts))

ringkas
```

Output :

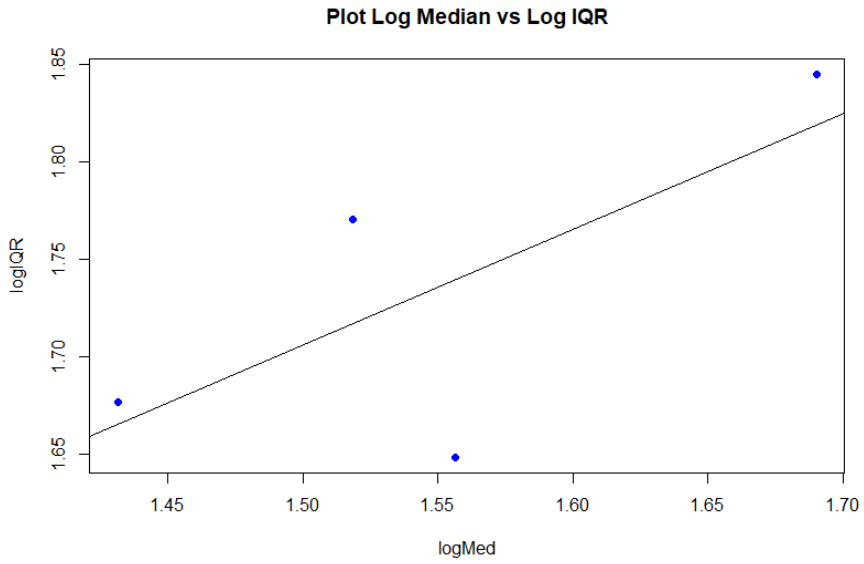
```
> ringkas
# A tibble: 4 x 3
  Bulan      iqr   Med
  <chr>    <dbl> <dbl>
1 April     70      49
2 Februari  47.5    27
3 Januari   44.5    36
4 Maret     59      33
```

3. Plot log median sebagai sumbu (X) vs log IQR sebagai sumbu Y

Syntax :

```
# Plot log median sebagai sumbu (X) vs log IQR sebagai
sumbu Y.
ringkas$logIQR <- log(ringkas$iqr,base = 10)
ringkas$logMed <- log(ringkas$Med,base = 10)

ringkas

plot(ringkas$logMed,ringkas$logIQR,pch = 16, cex =
1,col="blue",main = "Plot Log Median vs Log
IQR",xlab="logMed",ylab="logIQR")

abline(lm(ringkas$logIQR~ringkas$logMed))
```

Output :



4. Cek hasil nisbah berdasarkan nilai

Syntax :

```
#Cek hasil nisbah berdasarkan nilai
```

```
summary(lm(ringkas$logIQR~ringkas$logMed))
```

Output :

```
> summary(lm(ringkas$logIQR~ringkas$logMed))

Call:
lm(formula = ringkas$logIQR ~ ringkas$logMed)

Residuals:
       1        2        3        4
 0.02590  0.01149 -0.09118  0.05379

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.8136     0.6451   1.261    0.334
ringkas$logMed    0.5949     0.4157   1.431    0.289

Residual standard error: 0.07749 on 2 degrees of freedom
Multiple R-squared:  0.506,     Adjusted R-squared:  0.259
F-statistic: 2.049 on 1 and 2 DF,  p-value: 0.2887
```

Didapatkan persamaannya yaitu :

$$\log(\text{IQR}) = 0.8136 + 0.5949 \log(\text{Median})$$

Sehingga diketahui besar kemiringan atau nisbah adalah sebesar 0.5949, dimana ini berada diantara 0.5 dan 1, maka dengan mengacu pada tabel transformasi nisbah kira-kira, akan digunakan transformasi $\sqrt{x}$,

5. Lakukan transformasi

   Syntax :

```
# lakukan transformasi
mydataTRF <- sqrt(mydata)
head(mydataTRF)
```

   Output :

```
> head(mydataTRF)
# A tibble: 6 x 4
  Januari Februari Maret April
    <dbl>    <dbl> <dbl> <dbl>
1   15.2     10.2  11.4   8.12
2    6.48     6.86  3.61 17.0
3    4.24     3.32  2.83  3.32
4    5        0    13    10.7
5    5.48     4.24  5.74  7
6    5.20     3.32  2.83  6.56
```

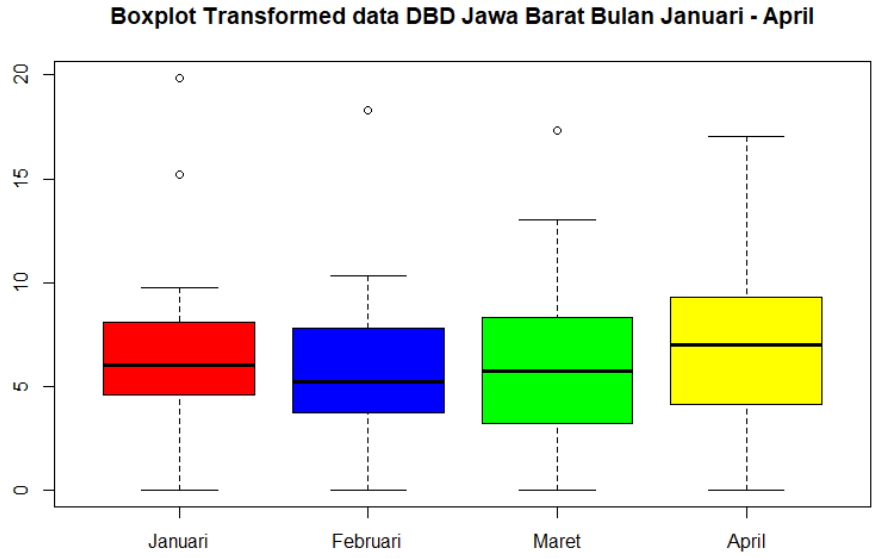
6. Buat Boxplot hasil transformasi

   Syntax :

```
# Buat Boxplot hasil transformasi
```

```
boxplot(mydataTRF,col=c("red","blue","green","yellow")
,main="Boxplot Transformed data DBD Jawa Barat Bulan
Januari - April")
```

Output :



Dapat dilihat bahwa angkatan menjadi lebih simetris dengan sebaran yang relatif sama jika dibandingkan dengan kondisi saat sebelum ditransformasi.

D. Transformasi Beberapa Angkatan (Boxcox)

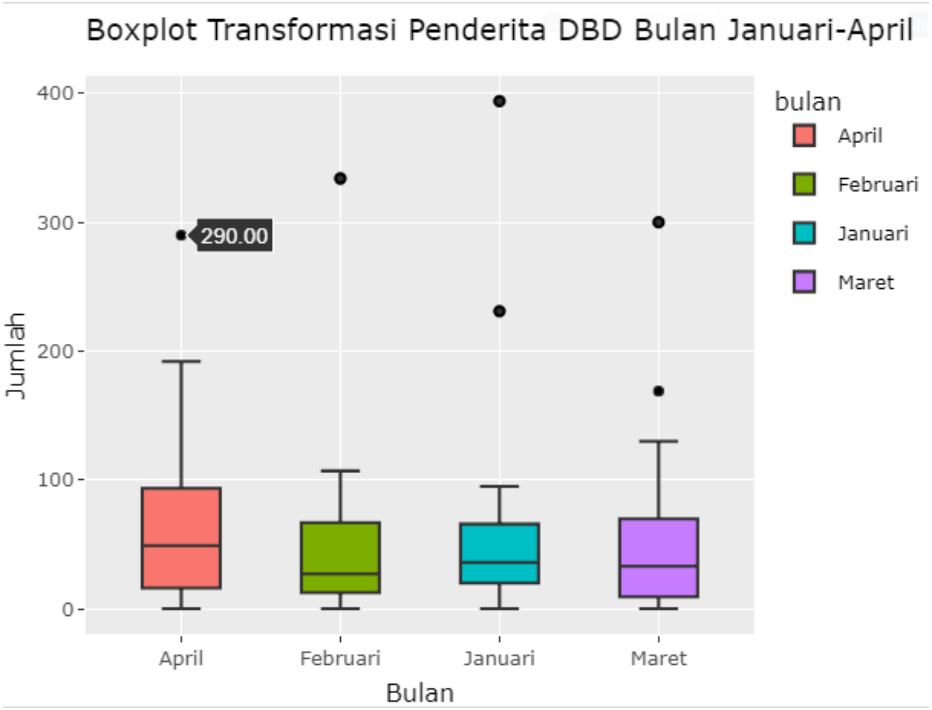
Transformasi Box-cox dari data proses dapat membantu memperbaiki kondisi ketika bentuk juraian masing-masing angkatan berbeda (sebaran angkatan relatif beda), dimana beberapa angkatan menjurai ke atas sedangkan yang lain menjurai ke bawah.

Syarat Transformasi: Data harus bernilai > 0, maka jika ada data yang bernilai 0 harus diganti menjadi 0,0001. Selain itu, jika data < 0 digunakan transformasi lain yang tidak dibahas dipraktikum ini.

Contoh Soal

Dengan Data ***P-4.xlxs Sheet="DBD"***, Lakukan analisis kesimetrisan data. tentang kesimetrisan data. Jika diperlukan lakukan transformasi sehingga diperoleh transformasi yang optimal.

```
#menyiapkan data
library(openxlsx)
data_box=read.xlsx("D:\\KULIAH\\Asprak\\P-4.xlsx",sheet=2)
library(tidyverse)

#Mengganti variabel bernilai nol
seleksi=function(x){
  for(i in 1:length(x)){
    if(x[i]==0){
      x[i]=0.0001
    }
  }
  return(x)
}

#menyiapkan data untuk dibuat boxplot
dataku=data_box%>%
  gather(key="bulan",
         value="penderita",
         -Kabupaten)

databaru=select(dataku,-Kabupaten)
databaru$penderita=seleksi(databaru$penderita)

#boxplot sebelum transformasi
library(plotly)
a = ggplot(databaru, aes(x=bulan,
y=penderita,fill=bulan))+geom_boxplot()+
  ggtitle("Boxplot Transformasi Penderita DBD Bulan
Januari-April")+
  xlab("Bulan")+ylab("Jumlah")
ggplotly(a)
```



Berdasarkan angkatan diatas terlihat bahwa data pada angkatan belum simetris, sehingga perlu dilakukan transformasi data.
Note. di laprak/tugas/kuis lebih jelaskan ya

```
#Mencari Nilai Lamda
#Ada beberapa cara
```

#1. Transformasi Box-cox dengan library AID

```
install.packages("AID")
library(AID)
out = boxcoxnc(databaru$penderita, method = "mle",
               lambda = seq(-2,2,0.001), verbose = F, plot
= F)
out$lambda.hat
```

#2. Transformasi Box-cox dengan library MASS

```
install.packages("MASS")
library(MASS)
out = boxcox(databaru$penderita~1, lambda =
seq(-2,2,0.0001), plotit = F)
out$x[which.max(out$y)]
```

#3.Transformasi Box-cox dengan library car

```
library(car)
```

```
out = powerTransform(databaru$penderita, family =
"bcPower")
out$lambda
```

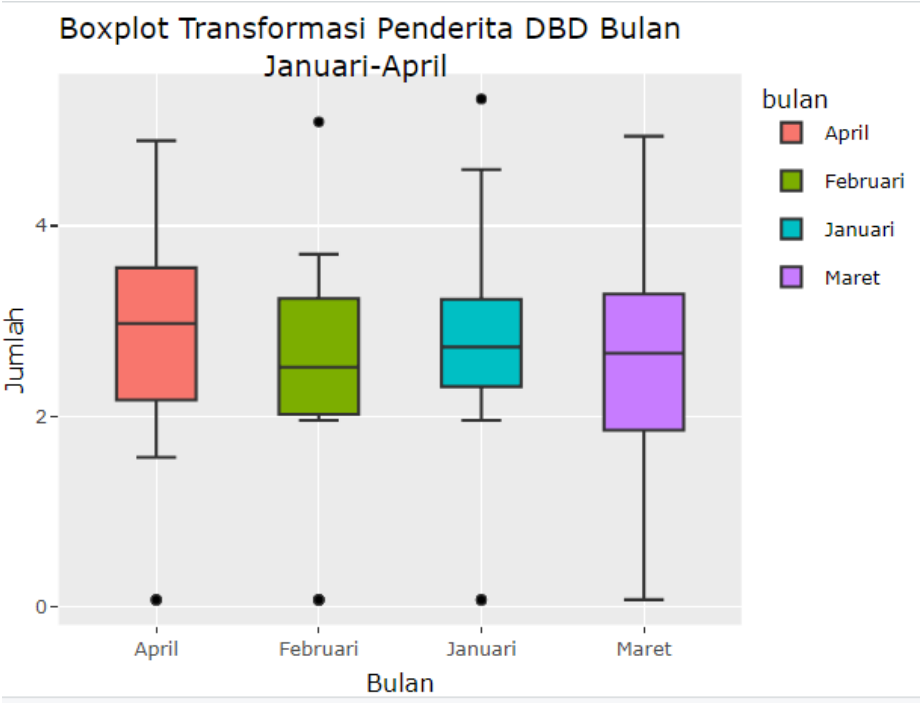
```
> library(AID)
> out = boxcoxnc(databaru$penderita, method = "mle",
+                lambda = seq(-2,2,0.001), verbose = F, plot = F)
> out$lambda.hat
[1] 0.282
```

```
> library(MASS)
> out = boxcox(databaru$penderita~1, lambda = seq(-2,2,0.0001), plotit = F)
> out$x[which.max(out$y)]
[1] 0.2818
> #Transformasi Box-cox dengan library car
> library(car)
> out = powerTransform(databaru$penderita, family = "bcPower")
> out$lambda
databaru$penderita
         0.2817606
```

Berdasarkan hasil di atas, didapatkan nilai λ = 0,28 sehingga akan digunakan untuk transformasi. Transformasi dilakukan dengan cara memangkatkan setiap datum pada data dengan 0,28.

```
#Transformasi data dengan Metode Box-Cox
#Melakukan transformasi terhadap data
transformasi.bc=function(data, lambda){
  for(i in 1:length(data)){
    data[i]=data[i]^lambda
  }
  return(data)
}
databaru.bc=databaru %>%
  mutate(tfbc=transformasi.bc(databaru$penderita,0.28))
hasil=select(databaru.bc, bulan, tfbc)

#membuat boxplot setelah ditransformasi
library(plotly)
y = ggplot(databaru.bc, aes(x=bulan, y=tfbc,
            fill=bulan))+geom_boxplot()+
                ggtitle("Boxplot Transformasi Penderita DBD
Bulan
```

```
Januari-April")+xlab("Bulan")+ylab("Jumlah")
ggplotly(y)
```



Dari tranformasi Box-cox, didapatkan bahwa nilai λ = 0,28 yang berarti data yang ada paling baik dilakukan transformasi dengan X 0,28. Dari boxplot hasil transformasi Box-cox tersebut, terlihat bahwa bentuk boxplot menjadi lebih simetris dari boxplot data sebelum ditransformasi. Sehingga, data setelah transformasi dianggap mendekati distribusi normal

```
#Untuk melihat lebih jelas distribusi angkatan bisa distandardisasi kemudian dibuat boxplot lagi
```

LATIHAN SOAL

1. Diketahui Data Hasil Panen Padi per-Kabupaten di Provinsi Daerah Istimewa Yogyakarta sebagai berikut.

Jogja	Bantul	Sleman	Kulon Progo	Gunung Kidul
55	164	91	239	921
13	70	187	295	629

3	**96**	**37**	**525**	**645**
6	**221**	**61**	**175**	**666**
64	**84**	**4**	**133**	**571**
50	**5**	**20**	**259**	**257**
33	**3**	**396**	**749**	**461**
10	**6**	**41**	**43**	**535**
33	**22**	**108**	**370**	**840**
	77	**398**	**321**	**857**
	250	**172**	**474**	**402**
	167	**111**	**327**	**57**
	17	**328**	**225**	**299**
	30	**7**	**346**	**644**
		28	**455**	
		224	**143**	
		63	**256**	
		146	**190**	
		69	**436**	
		254	**62**	
		192	**528**	
		136	**79**	
		77	**350**	
		25	**342**	
		88	**828**	

		977	**259**	
		254	**361**	
		273	**363**	
		13	**444**	
		28	**540**	
		148	**354**	
			399	
			545	
			669	

a. Periksalah bentuk sebaran dari hasil panen padi menggunakan boxplot, lalu bandingkan antar kabupaten
b. Jika tidak simetri, lakukan transformasi dengan metode perkiraan nisbah
c. Ulangi soal b) dengan metode Box-cox