6-Visualisasi-Data

Ashari Ramadhan

11/20/2020

Visualisasi Data

Singkatnya, visualisasi data dipakai untuk mempresentasikan data yang terstruktur ataupun tidak dengan grafik. Tujuan utama dari visualisasi data adalah untuk mengkomunikasikan informasi secara jelas dan efisien kepada pengguna lewat grafik informasi.

Figure 1: https://www.finereport.com/en/data-visualization/visualisasi-data.html

R memiliki library untuk visualisasi, baik fungsi built in, ggplo2, plotly, highcharter dan lain-lain.

Dataset

R menyediakan banyak dataset untuk dapat kita gunakan. Untuk melihat daftar dataset yang telah tersedia deafult di R, kita bisa menggunakan script berikut

```
data()
```

```
Data sets in package 'datasets':
```

AirPassengers	Monthly Airline Passenger Numbers 1949-1960
BJsales	Sales Data with Leading Indicator
BJsales.lead (BJsales)	Sales Data with Leading Indicator
BOD	Biochemical Oxygen Demand
CO2	Carbon Dioxide Uptake in Grass Plants
ChickWeight	Weight versus age of chicks on different diets
DNase	Elisa assay of DNase
EuStockMarkets	Daily Closing Prices of Major European Stock Indices, 1991-1998
Formaldehyde	Determination of Formaldehyde
HairEyeColor	Hair and Eye Color of Statistics Students
Harman23.cor	Harman Example 2.3
Harman74.cor	Harman Example 7.4
Indometh	Pharmacokinetics of Indomethacin
InsectSprays	Effectiveness of Insect Sprays
JohnsonJohnson	Quarterly Earnings per Johnson & Johnson Share
LakeHuron	Level of Lake Huron 1875-1972
LifeCycleSavings	Intercountry Life-Cycle Savings Data
Loblolly	Growth of Loblolly pine trees
Nile	Flow of the River Nile
Orange	Growth of Orange Trees
OrchardSprays	Potency of Orchard Sprays
PlantGrowth	Results from an Experiment on Plant Growth
Puromycin	Reaction Velocity of an Enzymatic Reaction
Seatbelts	Road Casualties in Great Britain 1969-84
Theoph	Pharmacokinetics of Theophylline
Titanic	Survival of passengers on the Titanic
ToothGrowth	The Effect of Vitamin C on Tooth Growth in Guinea Pigs
UCBAdmissions	Student Admissions at UC Berkeley
UKDriverDeaths	Road Casualties in Great Britain 1969-84
UKgas	UK Quarterly Gas Consumption
USAccDeaths	Accidental Deaths in the US 1973-1978
USArrests	Violent Crime Rates by US State
USJudgeRatings	Lawyers' Ratings of State Judges in the US Superior Court
USPersonalExpenditure	Personal Expenditure Data
UScitiesD	Distances Between European Cities and Between US Cities
VADeaths	Death Rates in Virginia (1940)
WWWusage	Internet Usage per Minute
WorldPhones	The World's Telephones
ability.cov	Ability and Intelligence Tests
airmiles	Passenger Miles on Commercial US Airlines, 1937-1960
airquality	New York Air Quality Measurements
anscombe	Anscombe's Quartet of 'Identical' Simple Linear Regressions
attenu	The Joyner-Boore Attenuation Data
attitude	The Chatterjee-Price Attitude Data
austres	Quarterly Time Series of the Number of Australian Residents
beaver1 (beavers)	Body Temperature Series of Two Beavers
beaver2 (beavers)	Body Temperature Series of Two Beavers
cars	Speed and Stopping Distances of Cars

chickwts	Chicken Weights by Feed Type
co2	Mauna Loa Atmospheric CO2 Concentration
crimtab	Student's 3000 Criminals Data
discoveries	Yearly Numbers of Important Discoveries
esoph	Smoking, Alcohol and (O)esophageal Cancer
euro	Conversion Rates of Euro Currencies
euro.cross (euro)	Conversion Rates of Euro Currencies
eurodist	Distances Between European Cities and Between US Cities
faithful	Old Faithful Geyser Data
fdeaths (UKLungDeaths)	Monthly Deaths from Lung Diseases in the UK
freeny	Freeny's Revenue Data
freeny.x (freeny)	Freeny's Revenue Data
freeny.y (freeny)	Freeny's Revenue Data
infert	Infertility after Spontaneous and Induced Abortion
iris	Edgar Anderson's Iris Data
iris3	Edgar Anderson's Iris Data
islands	Areas of the World's Major Landmasses
ldeaths (UKLungDeaths)	Monthly Deaths from Lung Diseases in the UK
lh	Luteinizing Hormone in Blood Samples
longley	Longley's Economic Regression Data
lynx	Annual Canadian Lynx trappings 1821-1934
mdeaths (UKLungDeaths)	Monthly Deaths from Lung Diseases in the UK
morley	Michelson Speed of Light Data
mtcars	Motor Trend Car Road Tests
nhtemp	Average Yearly Temperatures in New Haven
nottem	Average Monthly Temperatures at Nottingham, 1920-1939
npk	Classical N, P, K Factorial Experiment
occupationalStatus	Occupational Status of Fathers and their Sons
precip	Annual Precipitation in US Cities
presidents	Quarterly Approval Ratings of US Presidents
pressure	Vapor Pressure of Mercury as a Function of Temperature
quakes	Locations of Earthquakes off Fiji
randu	Random Numbers from Congruential Generator RANDU
rivers	Lengths of Major North American Rivers
rock	Measurements on Petroleum Rock Samples
sleep	Student's Sleep Data
stack.loss (stackloss)	Brownlee's Stack Loss Plant Data
stack.x (stackloss)	Brownlee's Stack Loss Plant Data
stackloss	Brownlee's Stack Loss Plant Data
state.abb (state)	US State Facts and Figures
state.area (state)	US State Facts and Figures
state.center (state)	US State Facts and Figures
state.division (state)	US State Facts and Figures
state.name (state)	US State Facts and Figures
state.region (state)	US State Facts and Figures
state.x77 (state)	US State Facts and Figures
sunspot.month	Monthly Sunspot Data, from 1749 to "Present"
sunspot.year	Yearly Sunspot Data, 1700-1988
sunspots	Monthly Sunspot Numbers, 1749-1983
swiss	Swiss Fertility and Socioeconomic Indicators (1888) Data
treering	Yearly Treering Data, -6000-1979
trees	Diameter, Height and Volume for Black Cherry Trees

```
uspop                       Populations Recorded by the US Census
volcano                     Topographic Information on Auckland's Maunga Whau
                            Volcano
warpbreaks                  The Number of Breaks in Yarn during Weaving
women                       Average Heights and Weights for American Women
```

Kita akan menggunakan beberapa dataset yang telah tersedia, salah satunya data iris. Dataset Iris merupakan dataset multivariat yang diperkenalkan oleh ahli statistika dan biologi Inggris, Ronald Fisher, dalam paper-nya tahun 1936. Dataset ini terdiri dari 3 spesies iris (Iris Setosa, Iris virginica, dan Iris versicolor) dan tiap spesies memiliki 50 sampel. Empat fitur yang diukur dari masing-masing sampel yaitu panjang dan lebar sepal dan kelopak, dalam sentimeter (Petal Length, Petal Width, Sepal Length, Sepal Width).

Melihat data iris

```
head(iris)
```

```
##   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
```

Melihat stuktur data iris

```
str(iris)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

Dari output di atas, diketahui data iris terdiri dari 150 observasu dan 5 variabel yang terdiri dari "Sepal.Length", "Sepal.Width", "Petal.Length", "Petal.Width", "Species"

Ggplot2

Ggplot2 merupakan Packages yang diciptakan oleh Hadley Wickham dengan kelebihannya dalam pembuatan gambar yang elegan dan kompleks. Popularitas ggplot2 di komunitas R tidak diragukan lagi. Ggplot2 memungkinkan anda untuk membuat grafik yang merepresentasikan data numerik dan kategorik baik univariat maupun multivariat secara simultan. Pengelompokan yang dapat diwakili oleh warna, simbol, ukuran dan ketebalan point. Ggplot2 mempunyai banyak fungsi dan pilihan untuk plot yang akan ditampilkan.

Instalasi dan load paket ggplot2

```
install.packages("ggplot2")
```

```
library(ggplot2)
```

Paket ggplot siap digunakan

Konsep ggplot2

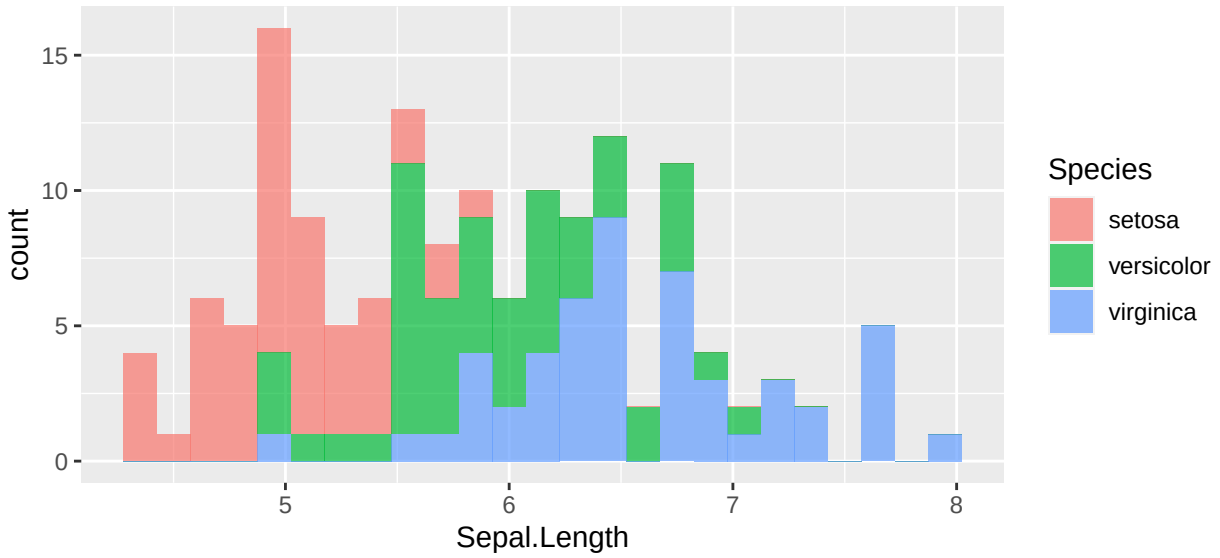
Konsep di balik ggplot2 membagi plot menjadi tiga bagian dasar yang berbeda: Plot = data + Estetika + Geometri.

Komponen utama dari setiap plot dapat didefinisikan sebagai berikut:

- data adalah kerangka data
- Aesthetics (aes) digunakan untuk menunjukkan variabel x dan y. Ini juga dapat digunakan untuk mengontrol warna, ukuran atau bentuk titik, ketinggian batang, dll... ..
- Geometri (geom_) mendefinisikan jenis grafik (histogram, boxplot, line, density, scatter plot, ...)

Contoh

```
ggplot(iris, aes(x=Sepal.Length, fill = Species)) +
  geom_histogram(bins = 25, alpha = 0.7)
```



Penjelesan

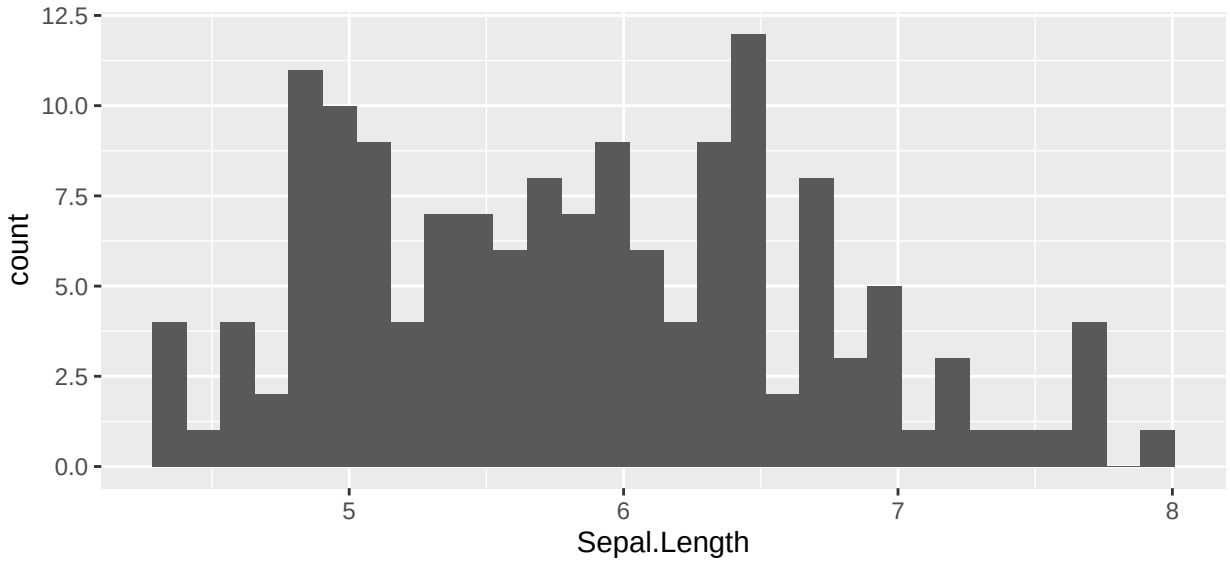
- ggplot() adalah fungsi untuk membuat grafik
- iris merupakan data
- 'x=Sepal.Length', 'fill = Species' adalah bagian dari aesthetic
- geom_histogram, adalah membuat isi dari aesthetic dipresentasikan sebagai histogram

Histogram

Untuk membuat histogram gunakan geom_hist

```
ggplot(iris, aes(x=Sepal.Length)) +
  geom_histogram()
```

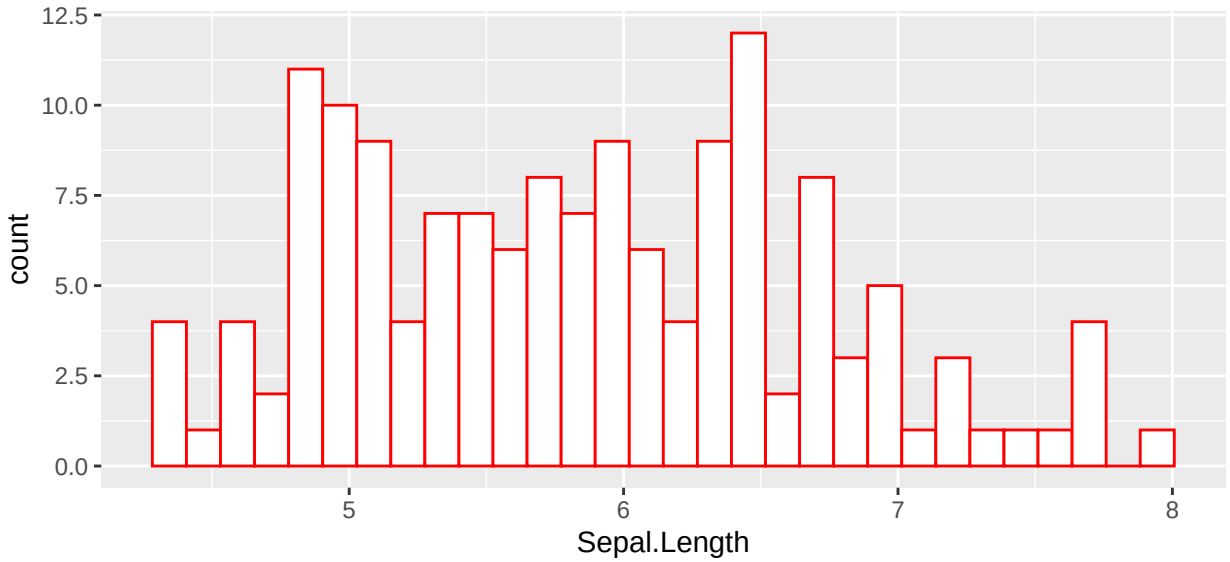
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Memberi warna garis dan batang

```
ggplot(iris, aes(x=Sepal.Length)) +
  geom_histogram(color = "red", # memberi warna garis batang merah
                 fill = "white") # memberi warna batang histogram putih
```

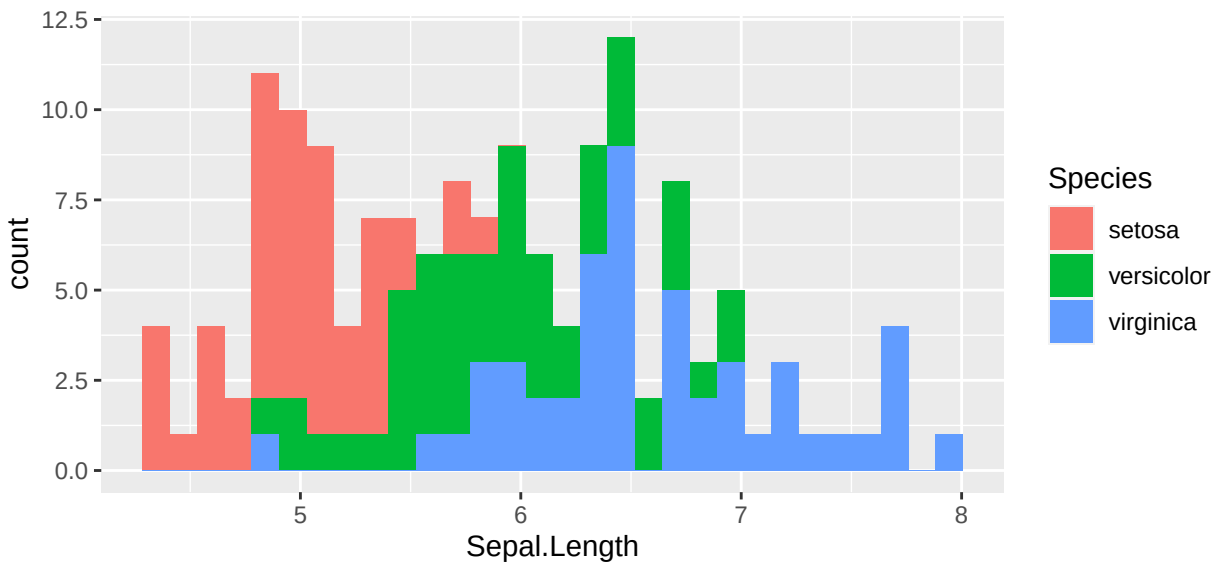
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Memberi warna sesuai dengan jenis species

```
ggplot(iris, aes(x=Sepal.Length, fill = Species)) +
  geom_histogram()
```

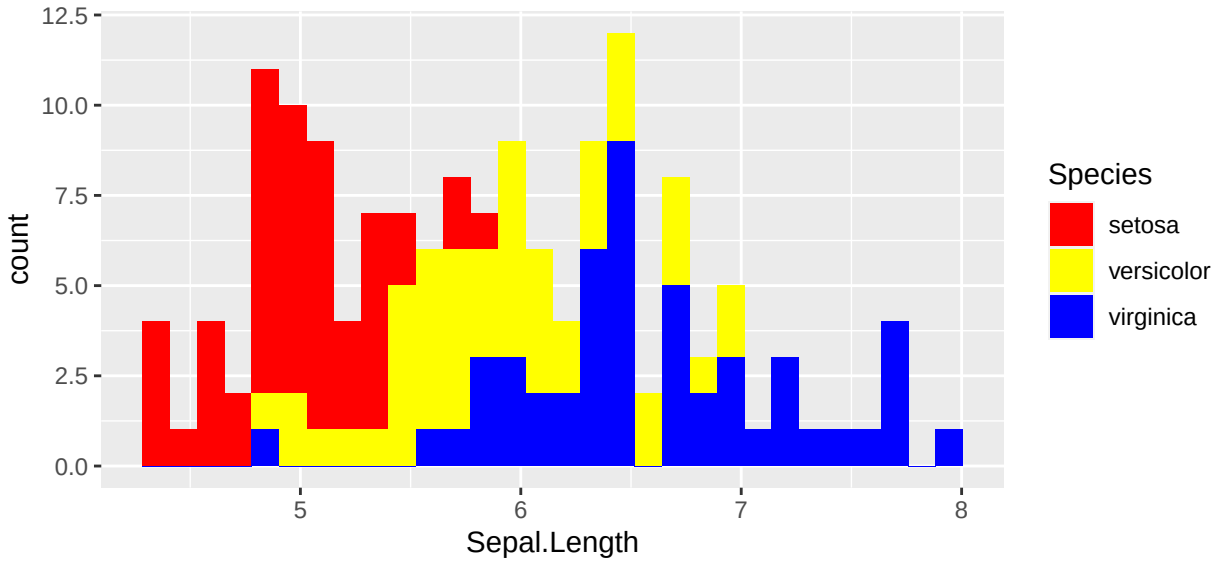
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Kostum warna

```
ggplot(iris, aes(x=Sepal.Length, fill = Species)) +
  geom_histogram() +
  scale_fill_manual(values=c("red", "yellow", "blue")) # Kostum warna
```

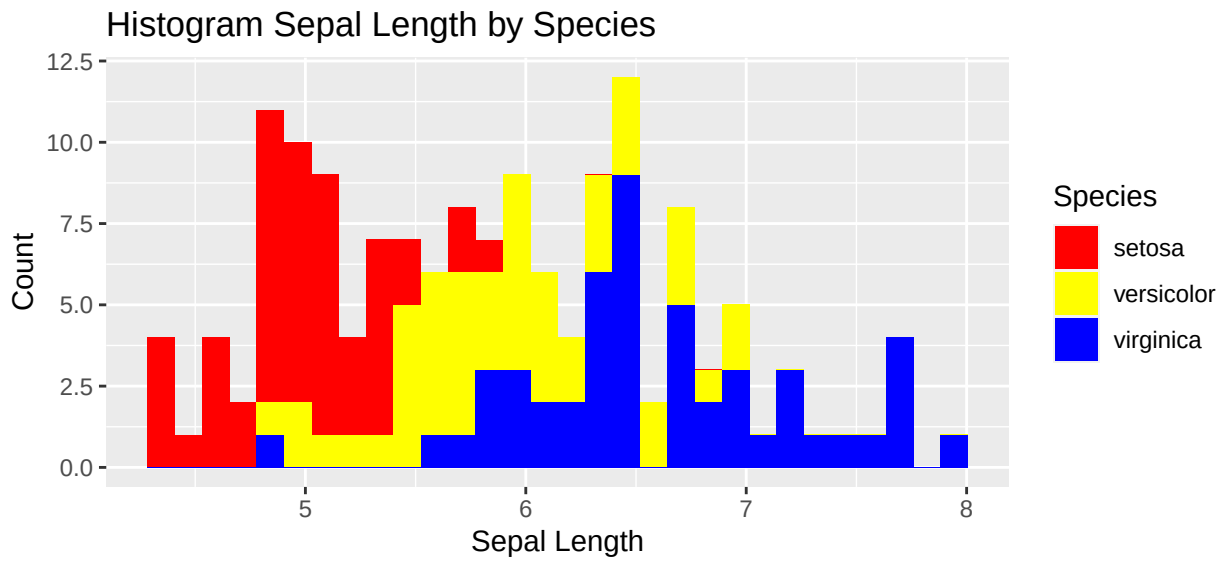
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Memberi judul utama, label X dan Y

```
ggplot(iris, aes(x = Sepal.Length, fill = Species)) +
   geom_histogram() +
   scale_fill_manual(values=c("red", "yellow", "blue")) +
   labs(title="Histogram Sepal Length by Species") + xlab("Sepal Length") + ylab("Count")
```

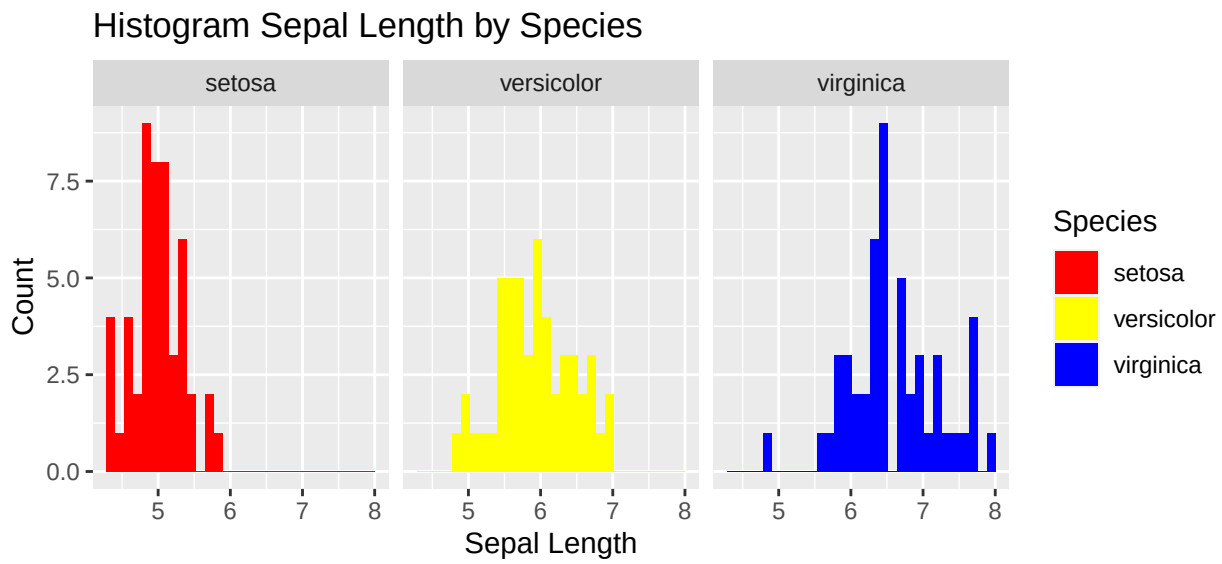
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Memisah histogam berdasarkan species dengan fungsi facet_wrap()

```
#density plot
ggplot(iris, aes(x = Sepal.Length, fill = Species)) +
   geom_histogram() +
   scale_fill_manual(values=c("red", "yellow", "blue")) +
   labs(title="Histogram Sepal Length by Species") + xlab("Sepal Length") + ylab("Count") +
  facet_wrap(~Species)
```

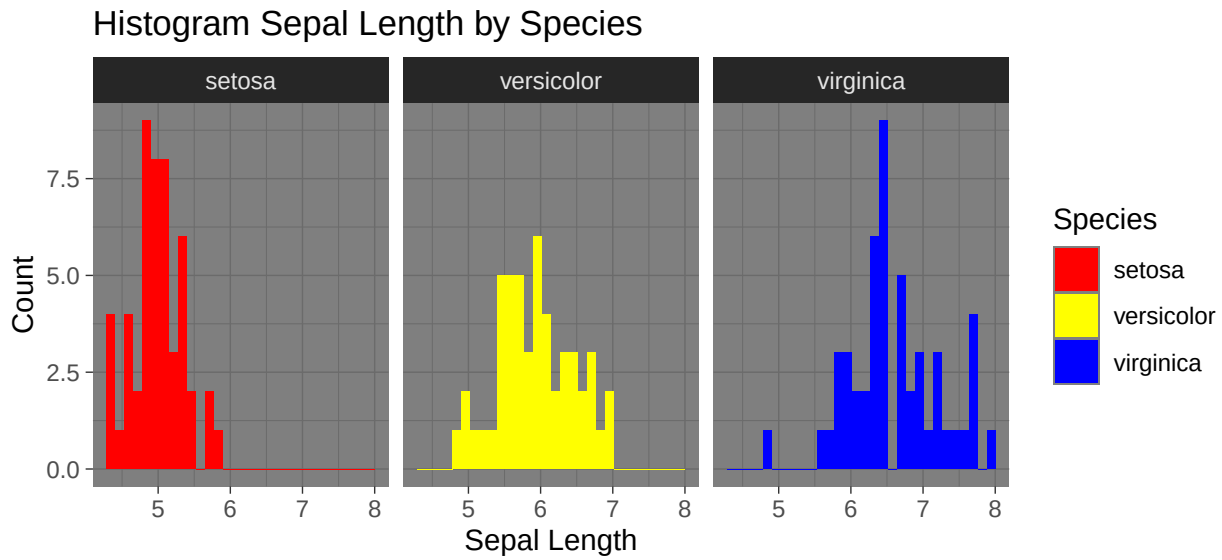
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Mengganti theme

```
#density plot
ggplot(iris, aes(x = Sepal.Length, fill = Species)) +
   geom_histogram() +
   scale_fill_manual(values=c("red", "yellow", "blue")) +
   labs(title="Histogram Sepal Length by Species") + xlab("Sepal Length") + ylab("Count") +
  facet_wrap(~Species) +
  theme_dark() #mengganti tema
```

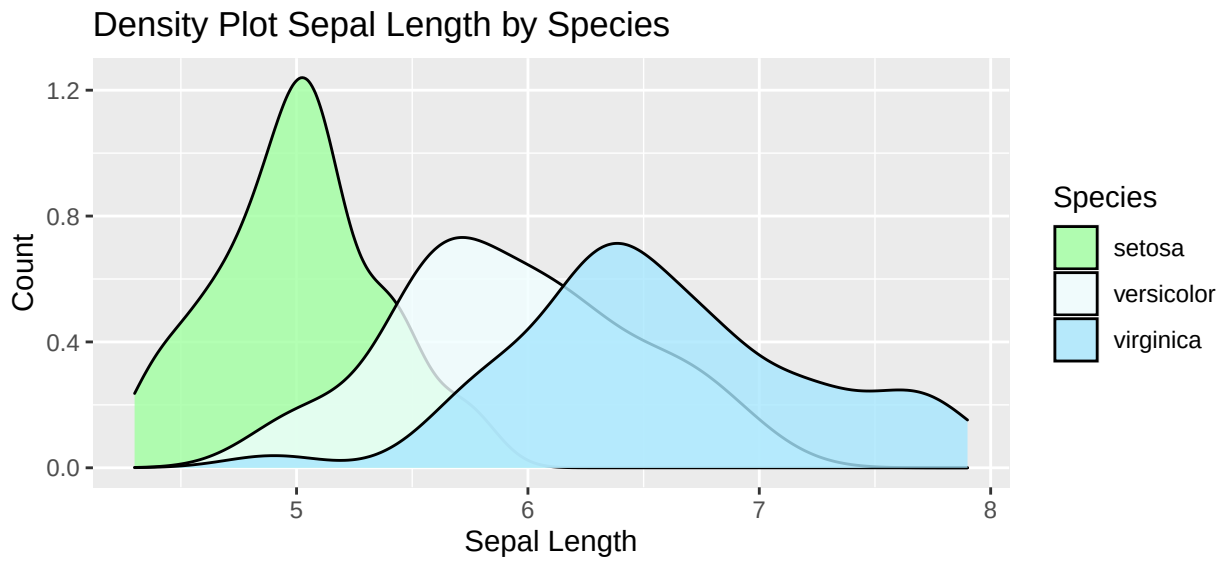
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



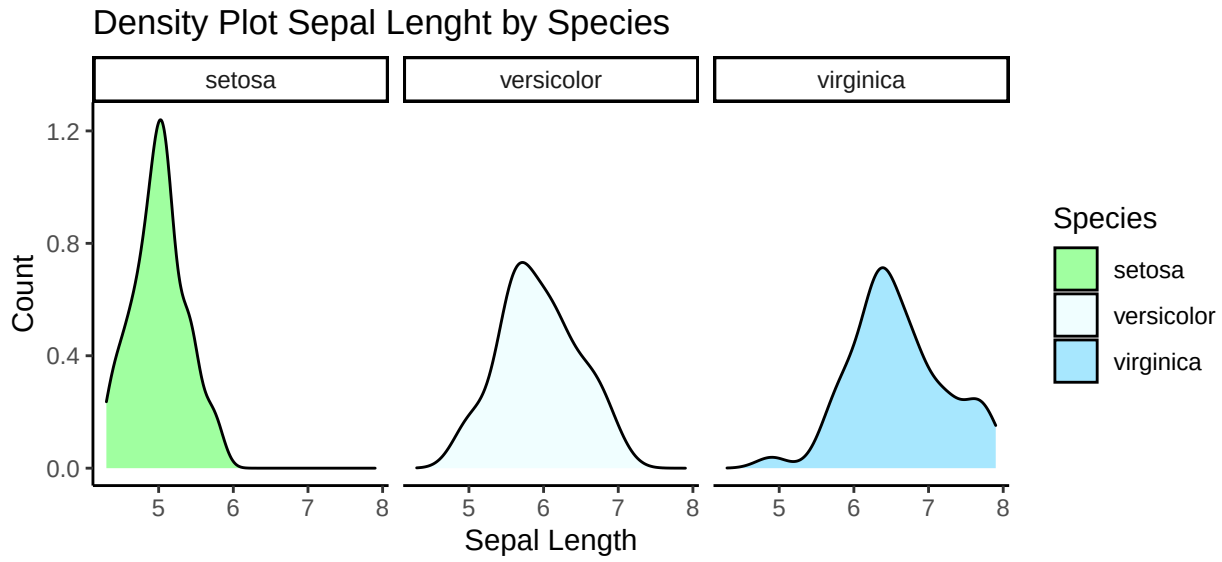
Density Plot

Membuat density plot persis membuat histogram, cukup ganti geom_historam menjadi menjadi geom_density

```
ggplot(iris, aes(x = Sepal.Length, fill = Species)) +
   geom_density(alpha = 0.8) +
   scale_fill_manual(values=c("#a0ffa0", "#f0feff", "#a7e7fe")) +
   labs(title="Density Plot Sepal Length by Species") + xlab("Sepal Length") + ylab("Count")
```



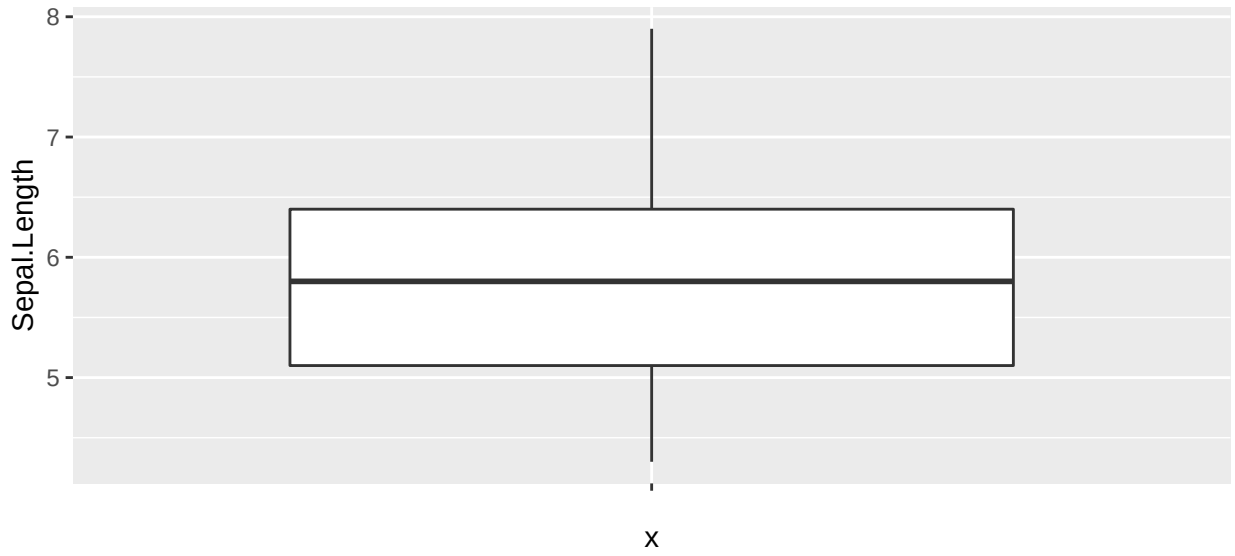

```
ggplot(iris, aes(x = Sepal.Length, fill = Species)) +
   geom_density() +
   scale_fill_manual(values=c("#a0ffa0", "#f0feff", "#a7e7fe")) +
   labs(title="Density Plot Sepal Lenght by Species") + xlab("Sepal Length") + ylab("Count") +
  facet_wrap(~Species) +
  theme_classic()
```



Boxplot

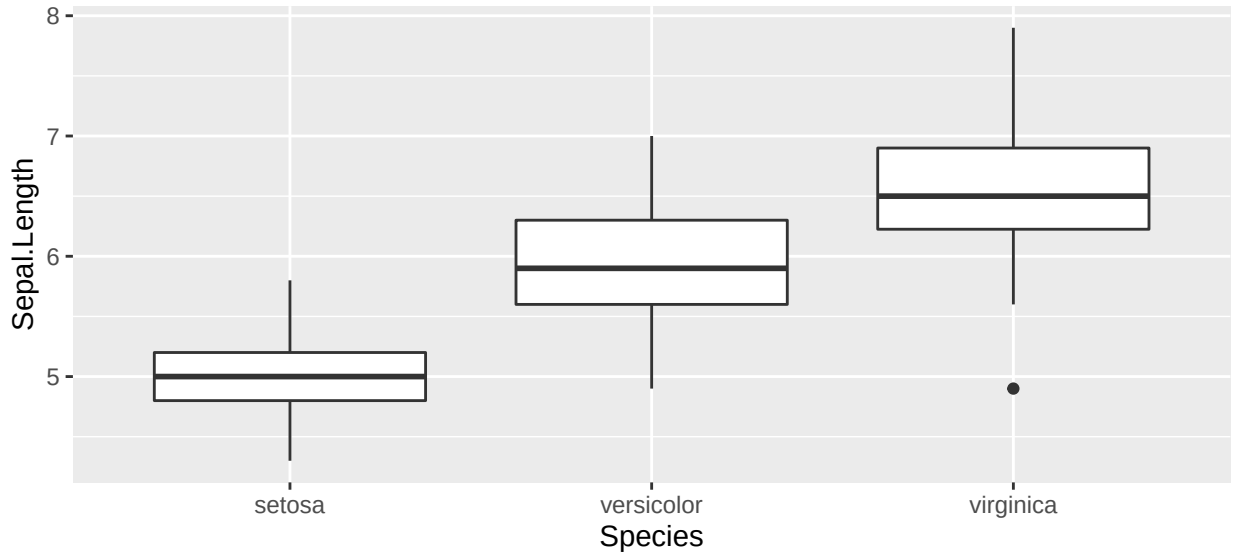
Boxplot pun demikian, ganti fungsi geom_histogram menjadi geom_boxplot. Jika kita ingin membuat boxplot satu variabel, parameter dalam fungsi aes() harus di akali menjadi aes(x = "", y = nama_variabel) sebab kita tidak bisa menghilangkan parameter x.

```
ggplot(iris, aes(x = "", y = Sepal.Length)) +
  geom_boxplot()
```

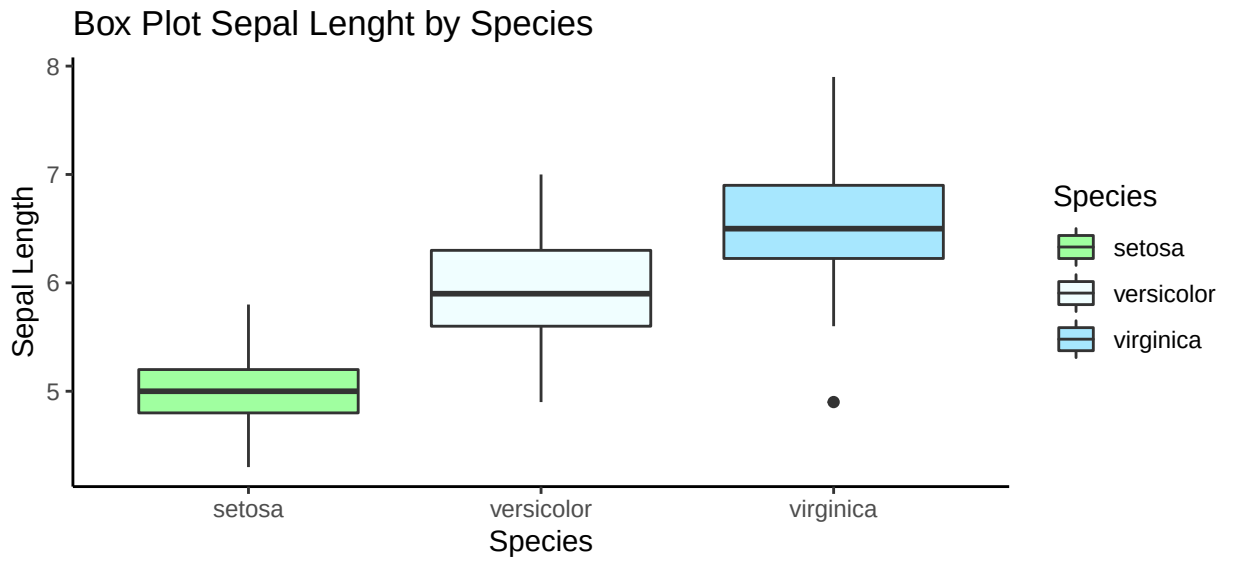


Jika ingin membuat boxplot Sepal.Length berdasarkan Species, masukkan Species sebagai parameter x pada fungsi aes().

```
ggplot(iris, aes(x = Species, y = Sepal.Length)) +
   geom_boxplot()
```



```
ggplot(iris, aes(x = Species, y = Sepal.Length, fill = Species)) +
   geom_boxplot() +
   scale_fill_manual(values=c("#a0ffa0", "#f0feff", "#a7e7fe")) +
   labs(title="Box Plot Sepal Lenght by Species") + xlab("Species") + ylab("Sepal Length") +
  theme(legeng.position = "none") +
  theme_classic()
```



Barplot

Gunakan geom_bar untuk membuat barplot. Misal kita punya data sebagai berikut:

```
survey <- data.frame(group=rep(c("Men", "Women"),each=6),
                     fruit=rep(c("Apple", "Kiwi", "Grapes", "Banana", "Pears", "Orange"),2),
                     people=c(22, 10, 15, 23, 12, 18, 18, 5, 15, 27, 8, 17))
survey
```

```
##    group  fruit people
## 1    Men  Apple     22
## 2    Men   Kiwi     10
## 3    Men Grapes     15
## 4    Men Banana     23
## 5    Men  Pears     12
## 6    Men Orange     18
## 7  Women  Apple     18
## 8  Women   Kiwi      5
## 9  Women Grapes     15
## 10 Women Banana     27
## 11 Women  Pears      8
## 12 Women Orange     17
```