Modul 03 - Introduction to Descriptive Analytics

Roni Yunis

17/09/2024

Pengantar

Secara sederhana, analisis deskriptif adalah analisis untuk memberikan gambaran tentang data dengan berbagai cara yang memungkinkan pengguna memahami situasi atau konteks dari data dengan cara yang jelas. Parameter statistik yang bisa digunakan seperti mean atau average, median, quartile, maximum, minimum, range, variance, dan standar deviasi. Untuk mendukung hasil analisis yang sudah dilakukan, biasanya akan di visualiasasikan dalam bentuk grafik (Plot). Plot dapat dibuat untuk menunjukkan hasil ringkasan data dari analisis statistik yang sudah dilakukan.

Analisis Deskriptif

Data

Data yang akan kita gunakan untuk pembahasa kali ini adalah dataset **insurance.csv**. Kita akan import data kedalam R dan kita simpan dalam objek *asuransi*

```
# import dataset
asuransi <- read.csv("data/insurance.csv")

# menampilkan 6 data teratas
head(asuransi)
```

```
##   age    sex    bmi children smoker    region   charges
## 1  19 female 27.900        0    yes southwest 16884.924
## 2  18   male 33.770        1     no southeast  1725.552
## 3  28   male 33.000        3     no southeast  4449.462
## 4  33   male 22.705        0     no northwest 21984.471
## 5  32   male 28.880        0     no northwest  3866.855
## 6  31 female 25.740        0     no southeast  3756.622
```

Langkah selanjutnya, Kita perlu melihat struktur data dari dataset agar kita bisa melihat variabel mana yang akan kita analisis.

```
# menampilkan struktur data
str(asuransi)
```

```
## 'data.frame':    1338 obs. of  7 variables:
##  $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
```

```
##  $ sex     : chr  "female" "male" "male" "male" ...
##  $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
##  $ children: int  0 1 3 0 0 0 1 3 2 0 ...
##  $ smoker  : chr  "yes" "no" "no" "no" ...
##  $ region  : chr  "southwest" "southeast" "southeast" "northwest" ...
##  $ charges : num  16885 1726 4449 21984 3867 ...
```

Dari data insurance.csv, bisa kita lihat bahwa data terdiri dari 1338 baris observasi dan 7 buah variabel. Untuk contoh kali kita akan menggunakan satu atau dua dari 4 buah variabel dengan type data numerik/integer yaitu **age**, **charges**, **bmi**, dan **children** yang nantinya akan dianalisis dengan pendekatan deskriptif.

Minimum dan Maximum

Untuk minimum dan maximum kita bisa menggunakan fungsi `min()` dan `max()`:

```
# nilai minimum dari Age
min(asuransi$age)
```

```
## [1] 18
```

```
# nilai maximum dari Age
max(asuransi$age)
```

```
## [1] 64
```

Bisa dilihat bahwa nilai minimum dari umur adalah 18 dan nilai maximum adalah 64

Latihan 1. Berapakah jumlah anak terkecil dan terbanyak dari variabel `children` 2. Berapakah body massa index yang terkecil dan terbesar dari variabel `bmi`

```
# your code
```

Hitunglah berapa nilai `charges` paling kecil dan paling besar?

```
# Melihat nilai charges paling kecil
min(asuransi$charges)
```

```
## [1] 1121.874
```

```
# Melihat nilai charges paling besar
max(asuransi$charges)
```

```
## [1] 63770.43
```

Jadi bisa dilihat bahwa nilai charges terkecil adalah 1121.874, dan nilai charges terbesar adalah 63770.43

Range

Fungsi selanjutnya adalah `range()` yang digunakan untuk melihat nilai minimum - maximum

```r
range(asuransi$age)
```

```
## [1] 18 64
```

Latihan 1. Berapakah range dari variabel `children` 2. Berapakah range dari variabel `bmi`

```r
# your code
```

hitunglah range dari variabel `charges`

```r
range(asuransi$charges)
```

```
## [1]  1121.874 63770.428
```

range dari variabel charges adalah 1121.874 - 63770.428

Mean

Fungsi selanjutnya adalah `mean()` yang digunakan untuk melihat nilai rata-rata.

```r
rata <- mean(asuransi$age)
rata
```

```
## [1] 39.20703
```

Jadi umur rata-rata adalah 39.20703

Latihan 1. Berapakah rata-rata dari variabel `children` 2. Berapakah rata-rata dari variabel `charges`

```r
# your code
```

```r
mean <- mean(asuransi$bmi)
mean
```

```
## [1] 30.6634
```

Kalau ada dalam data kita data *missing value (NA)*, maka fungsi `mean(asuransi$age, na.rm = TRUE)` bisa kita gunakan, artinya data NA itu tidak termasuk dalam rata-rata yang kita cari.

```r
rata_umur <- mean(asuransi$age, na.rm = TRUE)
rata_umur
```

```
## [1] 39.20703
```

```r
rata_bmi <- mean(asuransi$bmi, na.rm = TRUE)
rata_bmi
```

```
## [1] 30.6634
```

Median

Median atau nilai tengah dari ukuran pemusatan data, bisa menggunakan fungsi `median()`

```
median(asuransi$age)
```

```
## [1] 39
```

Jadi nilai tengah atau median dari umur adalah 39

Latihan 1. Berapakah nilai tengah dari variabel `children` 2. Berapakah nilai tengah dari variabel `charges` 3. Berapakah nilai tengah dari variabel `bmi`

```
# your code
```

Median juga bisa dihitung dengan fungsi `quantile()` dengan memasukkan nilai *quantile of ordernya*, yaitu nilainya 0.5 atau 50%. fungsi `quantile()` bisa digunakan seperti ini:

```
quantile(asuransi$age, 0.5)
```

```
## 50%
##  39
```

Jenis kuantil itu sangat tergantung pada kebutuhan dalam menentukan posisi sekumpulan data. Kuantil 2 disebut median karena data dibagi 2 (0.5). Kalau 4 disebut dengan *Kuartil*

Bagi kuartil = 4 bagian: 0,25, 0,5, 0,75, 100

```
quantile(asuransi$age, 0.25)
```

```
## 25%
##  27
```

```
quantile(asuransi$age, 0.75)
```

```
## 75%
##  51
```

Latihan 1. Berapakah nilai kuartil 75% dari variabel `children` 2. Berapakah nilai kuartil 25% dari variabel `charges` 3. Berapakah nilai kuantil 50% dari variabel `bmi`

```
# your code
```

Standar Deviasi dan Variance

Varians adalah ukuran dari seberapa jauh penyebaran data dari nilai rata-ratanya. Jika nilai varians semakin besar itu artinya semakin jauh menyebaran data dari nilai rata-ratanya. Standar Deviasi atau sering disebut dengan **Simpangan Baku** yaitu akar dari nilai varians. Tujuan dari Standar Deviasi adalah untuk mengetahui berapa banyak nilai atau jumlah data yang berbeda dari nilai rata-rata. Kalau kita sederhanakan bahwa Standar Deviasi itu mengukur data yang menyebar di sekitar Mean. Untuk menentukan Standar Deviasi dan Variance bisa menggunakan fungsi `sd()` dan `var()`.

```r
# Standar deviasi
sd <- sd(asuransi$age)
sd
```

```
## [1] 14.04996
```

```r
# Variance
var <- var(asuransi$age)
var
```

```
## [1] 197.4014
```

Latihan 1. Berapakah nilai sd dan var dari variabel `children` 2. Berapakah nilai sd dan var dari variabel `charges` 3. Berapakah nilai sd dan var dari variabel `bmi`

```r
# your code
```

Coefisien Varians

Koerfisien Variansi (CV) adalah rasio antara standar deviasi dengan nilai rata-rata. jadi bisa dihitung seperti ini.

```r
coefisien_varian <- sd(asuransi$age) / mean(asuransi$age)
coefisien_varian
```

```
## [1] 0.3583531
```

Bisa dilihat bahwa nilai koefisien varians dari umur adalah 0.358

```r
coefisien_varian_2 <- sd/rata_umur
coefisien_varian_2
```

```
## [1] 0.3583531
```

Latihan 1. Berapakah cv dari variabel `children` 2. Berapakah cv dari variabel `charges` 3. Berapakah cv dari variabel `bmi`

```r
# your code
```

Correlation

Korelasi digunakan untuk melihat hubungan antar 2 variabel, syarat dari korelasi adalah semua data harus bertype numerik. Kalau dilihat dari data yang kita miliki ada age, bmi, children, charges yang bertype numerik. Dalam contoh ini kita akan menghitung berapa korelasi antara *age* dengan *charges*, maka bisa menggunakan fungsi `cor()`

```r
cor(asuransi$age, asuransi$charges)
```

```
## [1] 0.2990082
```

Hitung korelasi dari Umur dengan BMI

```r
cor(asuransi$age, asuransi$bmi)
```

```
## [1] 0.1092719
```

Nilai korelasi dari kedua variabel tersebut adalah 0.29. Bisa kita simpulkan korelasi dari keduanya lemah. Nilai Korelasi berkisar antara 1 sampai -1. Jika nilai mendekati 1 atau -1 itu artinya hubungan antara 2 variabel kuat, tapi kalau nilai korelasi mendekati 0, artinya hubungan antara 2 variabel lemah.

Latihan Hitung Korelasi antar bmi - charges, children - bmi, age - children.

```r
# your code
```

Tabel Kontingensi

Tabel Kontingensi merupakan tabel yang dapat digunakan untuk mengukur hubungan/asosiasi antara 2 variabel yang kategorik, sehingga kita bisa rangkum frekuensi dari setiap kategori yang ada pada variabel. Misalnya varibel *sex* punya 2 kategori yaitu **male** dan **female**. Tabel *smoke* juga punya 2 kategori yaitu **yes** dan **no**. Jika kita ingin mengukur asosiasi antara *sex* dengan *smoke* maka hubungan itu bisa kita gambarkan seperti tabel 2 x 2.

Untuk melihat kategori dari variabel yang ada bisa menggunakan fungsi `table()`. Misalnya kita ingin melihat kategori yang ada pada variabel *sex*, maka fungsi `table()` bisa dituliskan seperti ini.

```r
table(asuransi$sex)
```

```
##
## female   male
##    662    676
```

Bisa dilihat bahwa jumlah laki-laki 676 dan perempuan 662.

```r
table(asuransi$smoker)
```

```
##
##   no  yes
## 1064  274
```

Bisa dilihat bahwa yang merokok 274 dan yang tidak 1064

Kemudian bagaimana kalau kita ingin melihat asosiasi antara 2 variabel yang berkategori contohnya antara *sex* dengan *smoke*, maka penulisan fungsi `table()` bisa seperti ini.

```r
table(asuransi$sex, asuransi$smoker)
```

```
##
##           no yes
##   female 547 115
##   male   517 159
```

Dari hubungan 2 kategori tersebut, bisa dilihat bahwa frekuensi jumlah laki-laki yang merokok 159 dan perempuan yang merokok 115.

Latihan Hitung frekuensi hubungan/asosiasi antar bmi - charges, children - bmi, age - children. Gunakan fungsi `table()`

```r
# your code
```

Untuk melihat frekuensi dan asosiasi antar 2 variabel ini kita juga bisa menggunakan fungsi `xtabs()`. penulisan fungsi tersebut bisa seperti ini.

```r
xtabs(~ asuransi$sex + asuransi$smoker)
```

```
##             asuransi$smoker
## asuransi$sex  no yes
##       female 547 115
##       male   517 159
```

Perbedaan dari kedua fungsi tersebut adalah, kalau pada fungsi `xtabs` menampilkan nama dari variabel.

Nah selanjutnya adalah bagaimana kita melihat besaran dari proporsi hubungan dari kedua variabel tersebut. Maka untuk mengukur besaran proporsi bisa menggunakan fungsi `prop.table()`. Penulisannya bisa dilakukan seperti ini.

```r
prop.table(table(asuransi$sex, asuransi$smoker))
```

```
##
##                  no        yes
##   female 0.40881913 0.08594918
##   male   0.38639761 0.11883408
```

Hutunglah hubungan sex dengan region, dan berapa nilai proporsi dari hubungan tersebut?

```r
# nilai frekuensi hubungan antara bmi dengan charger
sex_region <- table(asuransi$sex, asuransi$region)
sex_region
```

```
##
##          northeast northwest southeast southwest
##   female       161       164       175       162
##   male         163       161       189       163
```

```r
#nilai proporsi dari sex dengan region
prop_sex_region <- prop.table(sex_region)
prop_sex_region
```

```
##
##          northeast northwest southeast southwest
##   female 0.1203288 0.1225710 0.1307922 0.1210762
##   male   0.1218236 0.1203288 0.1412556 0.1218236
```

```r
prop.table(table(asuransi$age))
```

```
##
##         18         19         20         21         22         23         24
## 0.05156951 0.05082212 0.02167414 0.02092676 0.02092676 0.02092676 0.02092676
##         25         26         27         28         29         30         31
## 0.02092676 0.02092676 0.02092676 0.02092676 0.02017937 0.02017937 0.02017937
##         32         33         34         35         36         37         38
## 0.01943199 0.01943199 0.01943199 0.01868460 0.01868460 0.01868460 0.01868460
##         39         40         41         42         43         44         45
## 0.01868460 0.02017937 0.02017937 0.02017937 0.02017937 0.02017937 0.02167414
##         46         47         48         49         50         51         52
## 0.02167414 0.02167414 0.02167414 0.02092676 0.02167414 0.02167414 0.02167414
##         53         54         55         56         57         58         59
## 0.02092676 0.02092676 0.01943199 0.01943199 0.01943199 0.01868460 0.01868460
##         60         61         62         63         64
## 0.01718984 0.01718984 0.01718984 0.01718984 0.01644245
```

Kalau kita ingin menghitung proporsi pada setiap baris, karena setiap baris mewakili satu kategori, maka untuk mendapatkan proporsi yang benar kita bisa tambahkan *margin = 1*, jadi seperti ini.

```r
prop.table(table(asuransi$sex, asuransi$smoker), margin = 1)
```

```
##
##                 no       yes
##   female 0.8262840 0.1737160
##   male   0.7647929 0.2352071
```

Kalau kita ingin menambahkan margin kolom, maka kita bisa ganti *margin = 2*

```r
prop.table(table(asuransi$sex, asuransi$smoker), margin = 2)
```

```
##
##                 no       yes
##   female 0.5140977 0.4197080
##   male   0.4859023 0.5802920
```

Kita juga bisa membulatkan suatu bilangan desimal berkoma dengan fungsi `round()`. Misalnya kita ingin menghitung persentase per baris atau kolom pada proporsi sebelumnya, maka kita bisa tambahkan fungsi `round(prop.table(), 1),2)`. 1 untuk baris, dan 2 untuk kolom atau 2 digit dibelakang koma

```
round(prop.table(table(asuransi$sex, asuransi$smoker),1),2)
```

```
##
##            no  yes
##   female 0.83 0.17
##   male   0.76 0.24
```

Statistik Deskriptif dengan Fungsi `descr()`

Fungsi `descr()` merupakan fungsi untuk menampilkan statistik deskriptif secara langsung. Bisa menampilkan standar deviasi, minimum, maksimum, Q1, Q3, dan median sekaligus. Data yang ditampikan hanya bertype data numeric. Untuk menjalankan fungsi `descr()` kita harus menginstal terlebih dahulu `library (summarytools)`

```
library(summarytools)
```

Contoh penggunaan fungsi `descr()` pada objek asuransi yang menampung 4 buah variabel numeric yaitu age, bmi, charges, dan children adalah.

```
descr(asuransi,
      heading = TRUE,
      stats = "common")
```

```
## Non-numerical variable(s) ignored: sex, smoker, region

## Descriptive Statistics
## asuransi
## N: 1338
##
##                       age       bmi    charges   children
## --------------- --------- --------- ---------- ----------
##            Mean     39.21     30.66   13270.42       1.09
##         Std.Dev     14.05      6.10   12110.01       1.21
##             Min     18.00     15.96    1121.87       0.00
##          Median     39.00     30.40    9382.03       1.00
##             Max     64.00     53.13   63770.43       5.00
##         N.Valid   1338.00   1338.00    1338.00    1338.00
##       Pct.Valid    100.00    100.00     100.00     100.00
```

```
descr(asuransi)
```

```
## Non-numerical variable(s) ignored: sex, smoker, region

## Descriptive Statistics
## asuransi
## N: 1338
##
##                        age       bmi    charges   children
## ---------------- --------- --------- ---------- ----------
##             Mean     39.21     30.66   13270.42       1.09
```

```
##           Std.Dev     14.05      6.10   12110.01       1.21
##               Min     18.00     15.96    1121.87       0.00
##                Q1     27.00     26.29    4738.27       0.00
##            Median     39.00     30.40    9382.03       1.00
##                Q3     51.00     34.70   16657.72       2.00
##               Max     64.00     53.13   63770.43       5.00
##               MAD     17.79      6.20    7440.81       1.48
##               IQR     24.00      8.40   11899.63       2.00
##                CV      0.36      0.20       0.91       1.10
##          Skewness      0.06      0.28       1.51       0.94
##       SE.Skewness      0.07      0.07       0.07       0.07
##          Kurtosis     -1.25     -0.06       1.59       0.19
##           N.Valid   1338.00   1338.00    1338.00    1338.00
##         Pct.Valid    100.00    100.00     100.00     100.00
```

Visualisasi

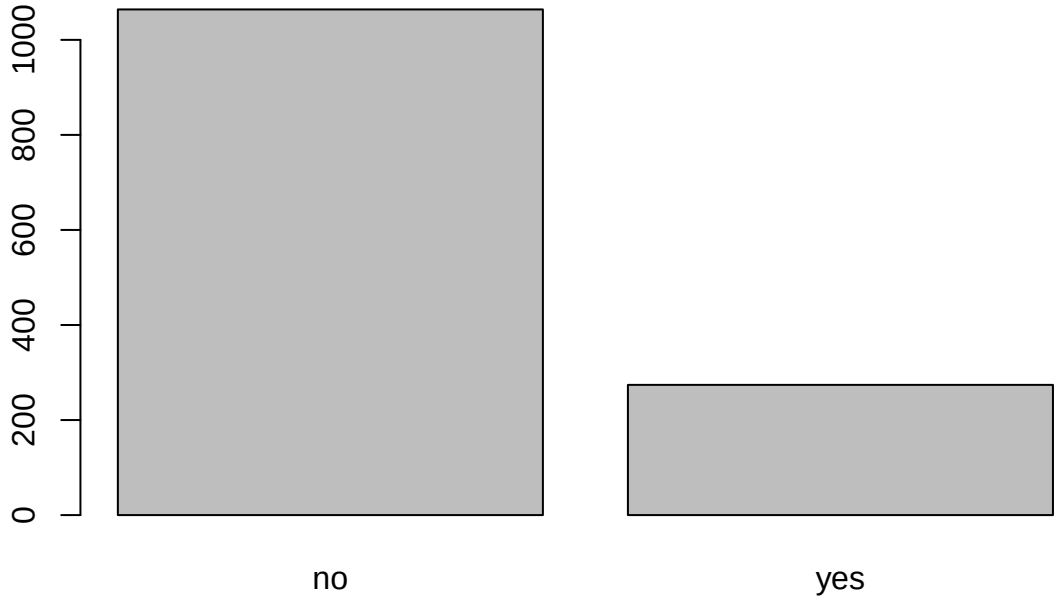
Untuk membuat visualiasi dari hasil analisis biasanya disajikan dalam berbagai bentuk grafik, grafik batang, garis, histogram, scatter, dll. Berikut ini akan dibahas beberapa contoh visualisasi data dengan grafik. Untuk memvisualisasikan data dalam R, dapat menggunakan fungsi yang sudah ada, dan bisa juga menggunakan sebuah packages/library yang khusus untuk visualisasi yaitu `library(ggplot2)`

Fungsi Visualisasi pada R

Barplot

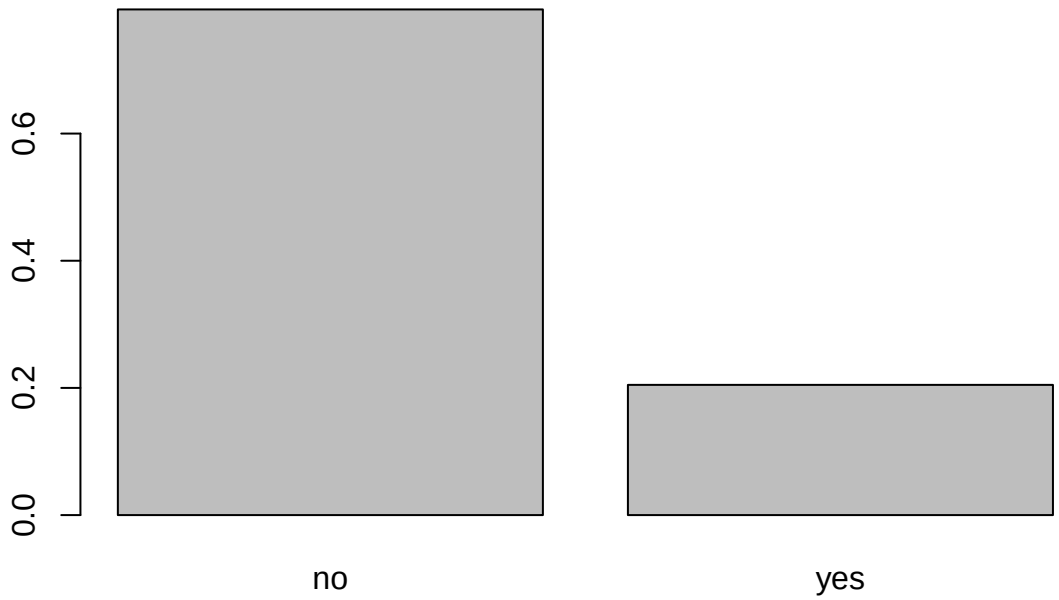
Barplot hanya dapat dilakukan untuk memvisualisasikan variabel yang kualitatif atau mengambarkan distribusi variabel kualitatif. Dalam contoh ini kita akan memvisualisasi variabel *smoker* yang berisi 2 kategori didalamnya yaitu "yes" dan "no". Untuk menampilkan barplot kita bisa menggunakan fungsi `barplot()`seperti ini.

```
barplot(table(asuransi$smoker))
```



Kalau nilai proporsi yang akan kita visualisasikan dengan barplot, maka bisa tulis seperti ini:

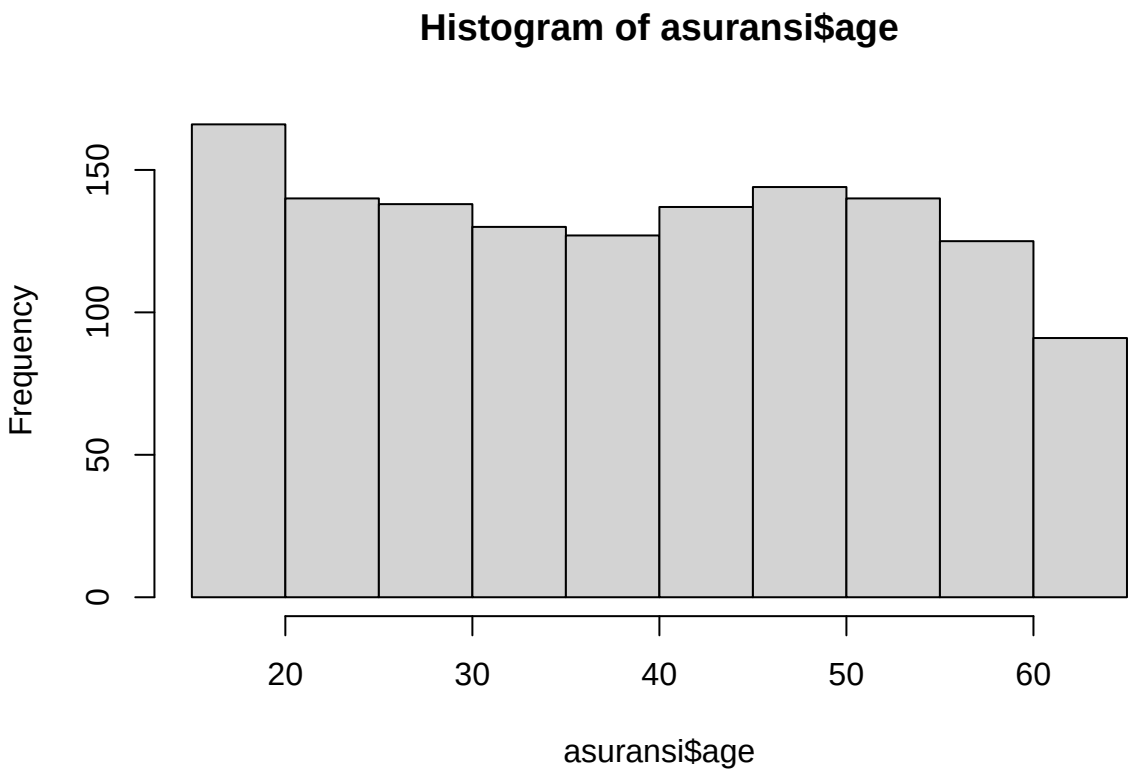
```
barplot(prop.table(table(asuransi$smoker)))
```



Histogram

Histogram biasanya digunakan untuk memberikan gambaran atau visualisasi distribusi variabel kualitatif. Dalam Histogram akan memecah rentang nilai menjadi interval dan akan menghitung berapa banyak observasi yang tepat pada setiap interval. Untuk menggambarkan Histogram pada R, mengguakan fungsi `hist()`. Dalam contoh kita akan menggambar histogram dari variabel "age".

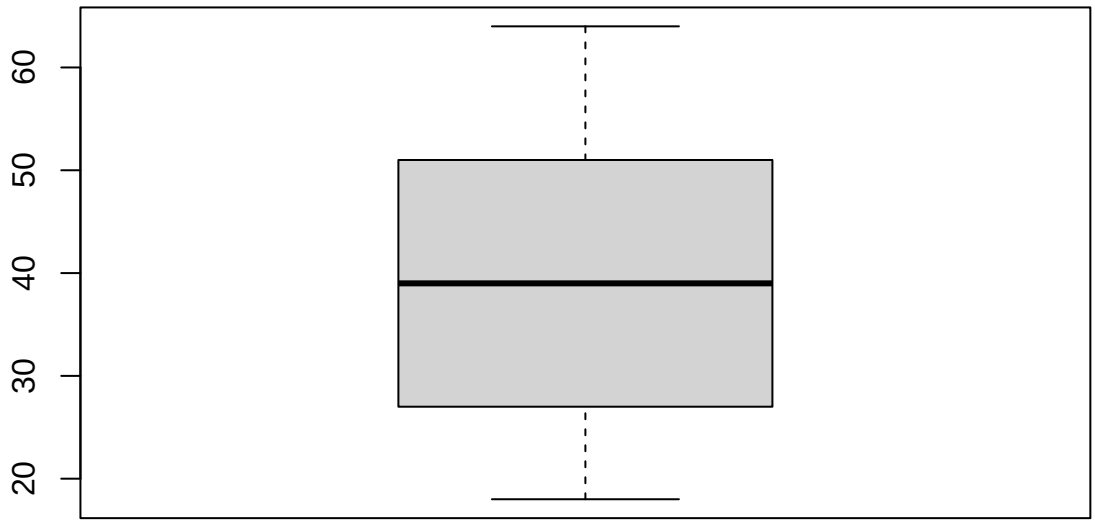
```
hist(asuransi$age)
```



Boxplot

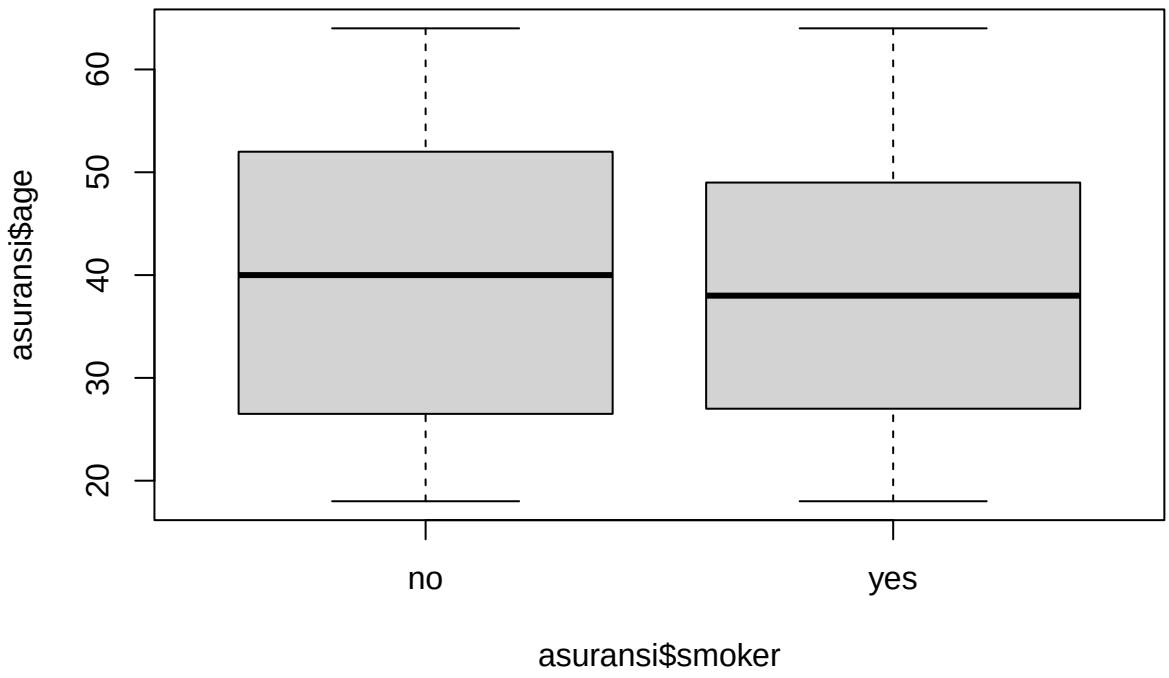
Boxplot ini sering digunakan dalam statistik deskriptif, biasanya diagram ini untuk menggambarkan distribusi variabel kuantitatif secara visual. Untuk menggambarkan Boxplot pada R, bisa menggunakan fungsi `boxplot()`

```
boxplot(asuransi$age)
```



Boxplot dapat disajikan berdampingan untuk membandingkan dan membedakan distribusi dari 2 atau lebih variabel. Misalnya dalam contoh ini kita akan membandingkan variabel “age” dan “smoker”.

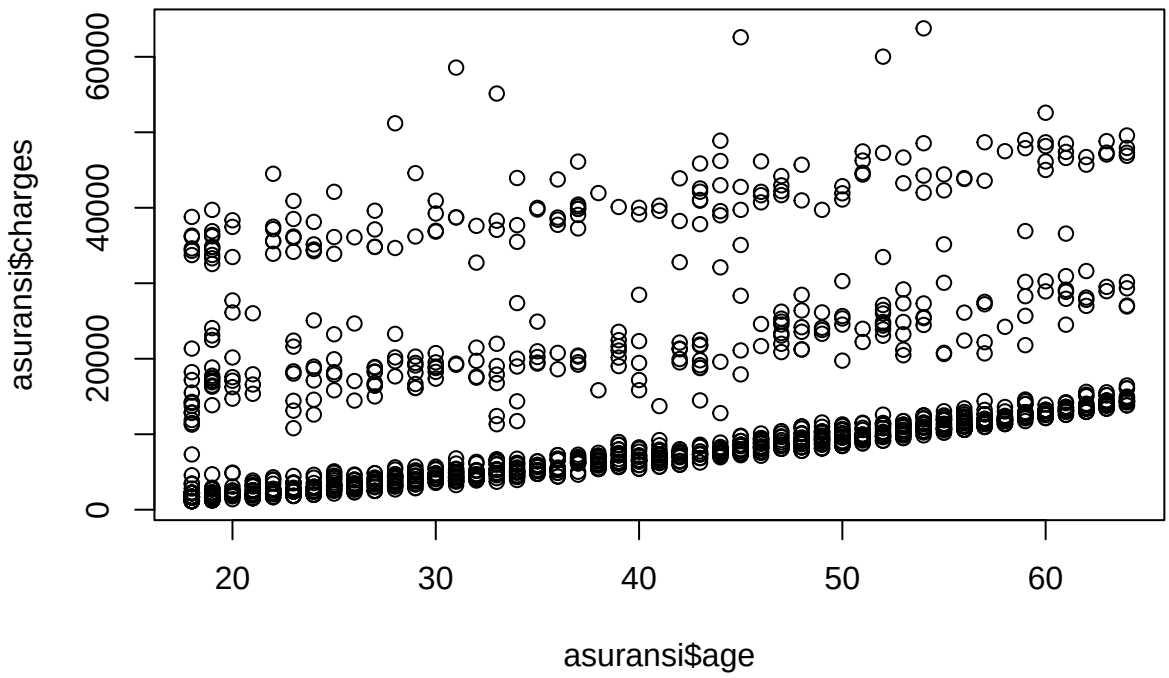
```
boxplot(asuransi$age ~ asuransi$smoker)
```



Scatterplot

Scatterplot sangat cocok digunakan untuk melihat distribusi 2 variabel kuantitatif, bisanya digunakan untuk melihat korelasi antar 2 variabel. Untuk menggambar Plot bisa menggunakan fungsi `plot()`. Dalam contoh ini kita akan menggambarkan korelasi 2 variabel yaitu antara "age" dan "charges".

```
plot(asuransi$age, asuransi$charges)
```

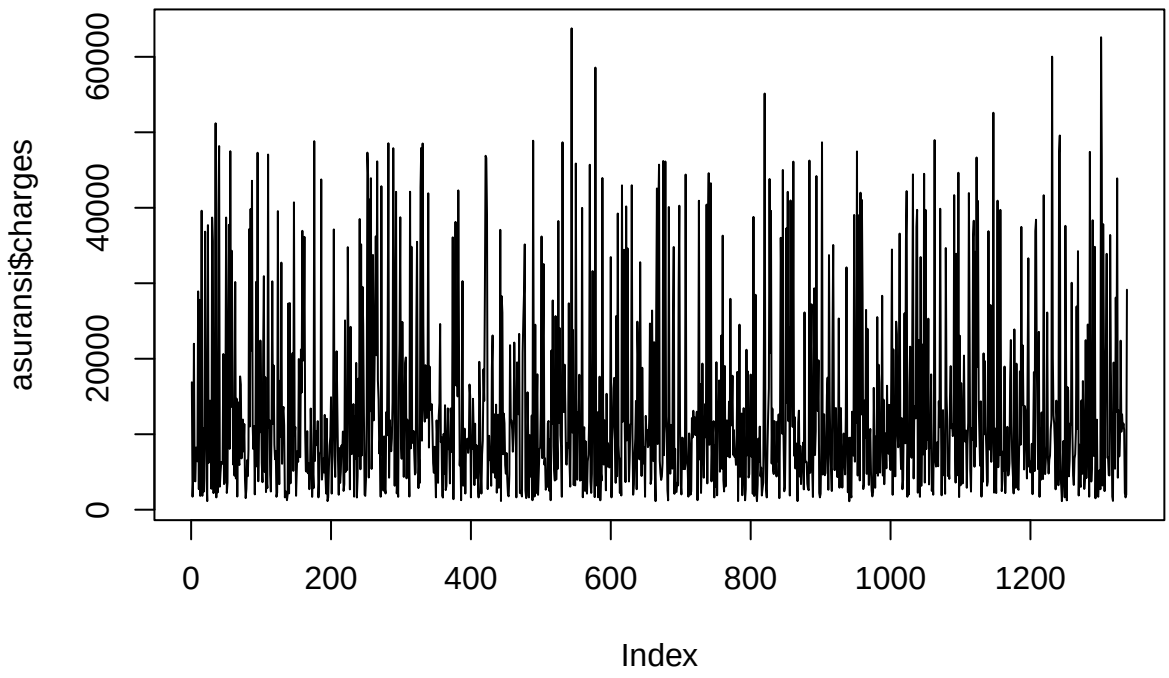


Kalau kita lihat dari kedua variabel tersebut, tampak adanya hubungan positif antara keduanya.

Line Plot

Line Plot, biasanya digunakan untuk mengambarkan data yang time series atau data yang disimpan dari waktu ke waktu, seperti data keuangan. Untuk menggambarkan lineplot bisa menambahkan `type = "l"` pada fungsi `plot()`.

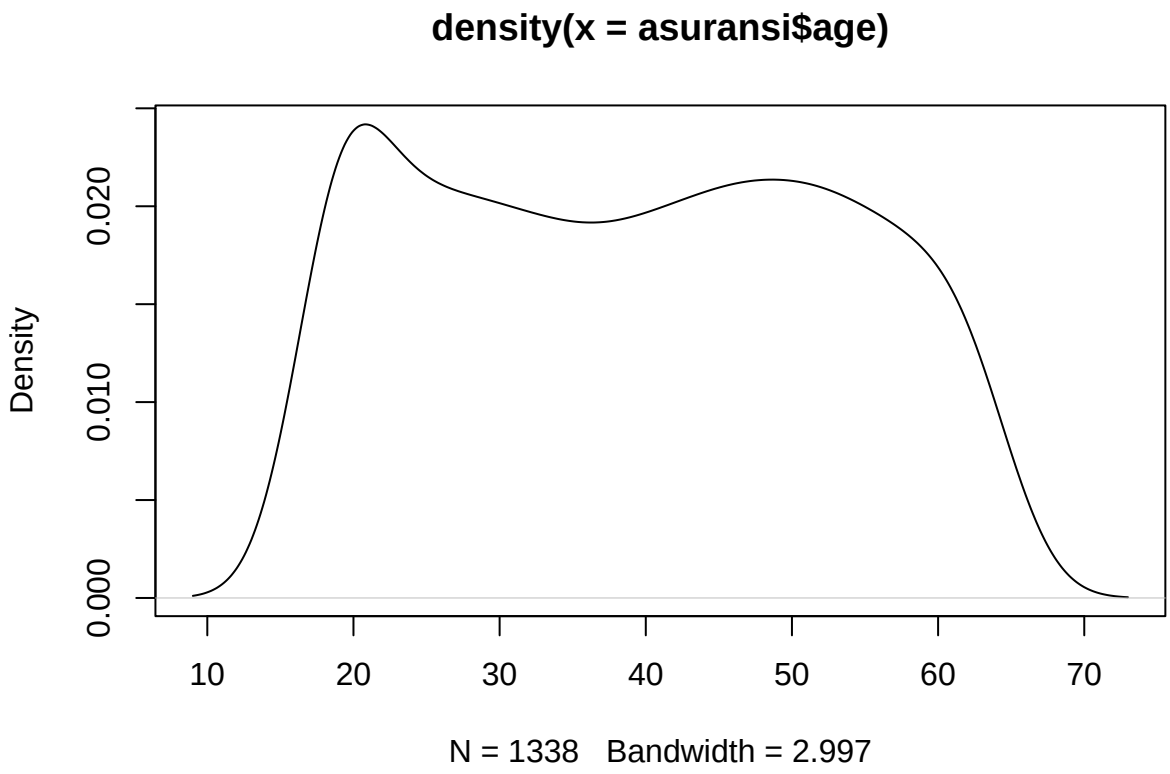
```
plot(asuransi$charges, type = "l")
```



Dessity Plot

Density plot merupakan bentuk lain dari histogram yang dibuat lebih halus (smooth), fungsi `density()` digunakan bersamaan dengan fungsi `plot()`

```
plot(density(asuransi$age))
```



Library ggplot2

Library `ggplot2` merupakan sebuah library yang dapat menggambarkan grafik lebih elegan dan komplek. Library ini sangat populer dikalangan komunitas R, dengan `ggplot2` kita bisa membuat grafik yang mempresentasikan data numerik dan kategorik secara simultan, yang dikelompokkan berdasarkan warna, simbol, ukuran dan ketebalan dari point. Disamping itu `ggplot2` memiliki banyak fungsi dan pilihan untuk plot yang akan ditampilkan.

Sebelum kita bahas beberapa contoh penerapannya, kita akan panggil library `ggplot2` terlebih dahulu.

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

# panggil library ggplot2
library(ggplot2)
```

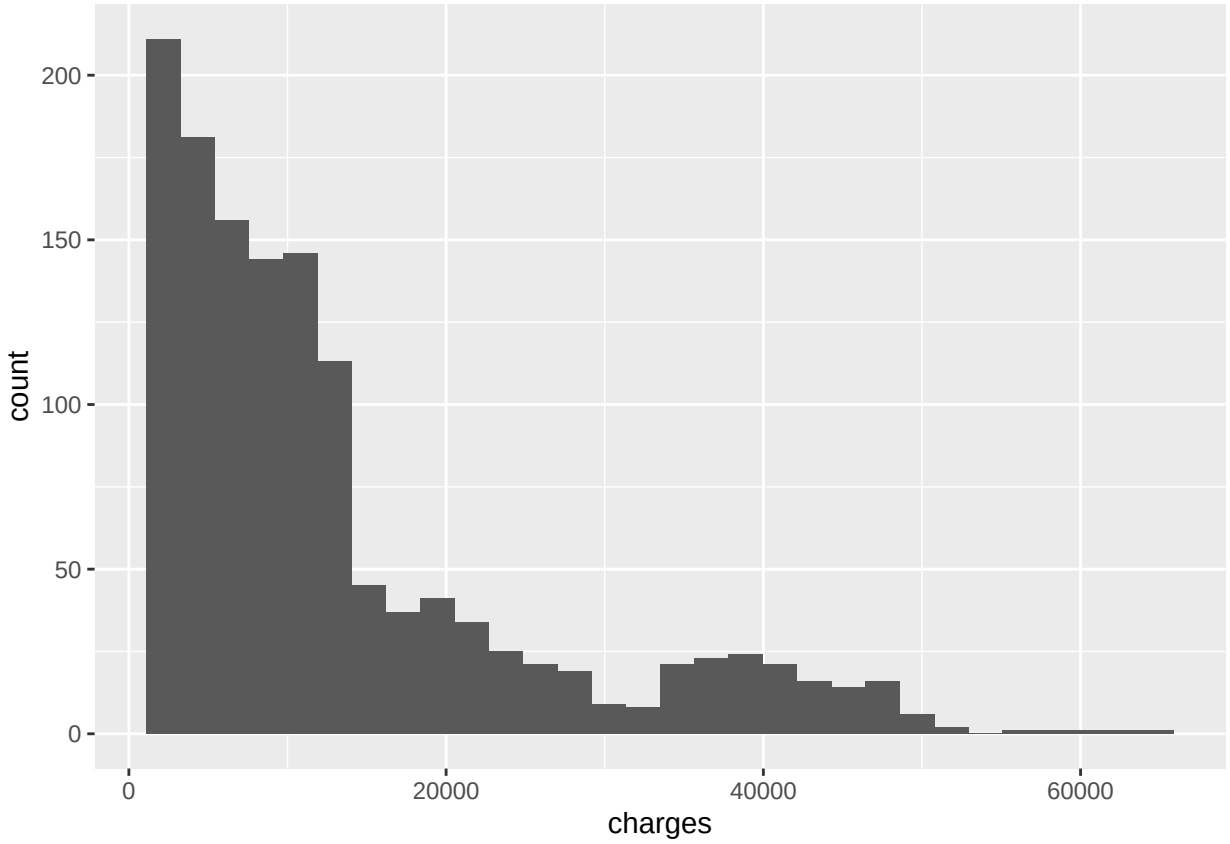
```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

Histogram

Berikut ini adalah contoh bagaimana histogram dengan fungsi `geom_histogram()` pada `ggplot2`

```
ggplot(asuransi, aes(x = charges))+ geom_histogram()
```

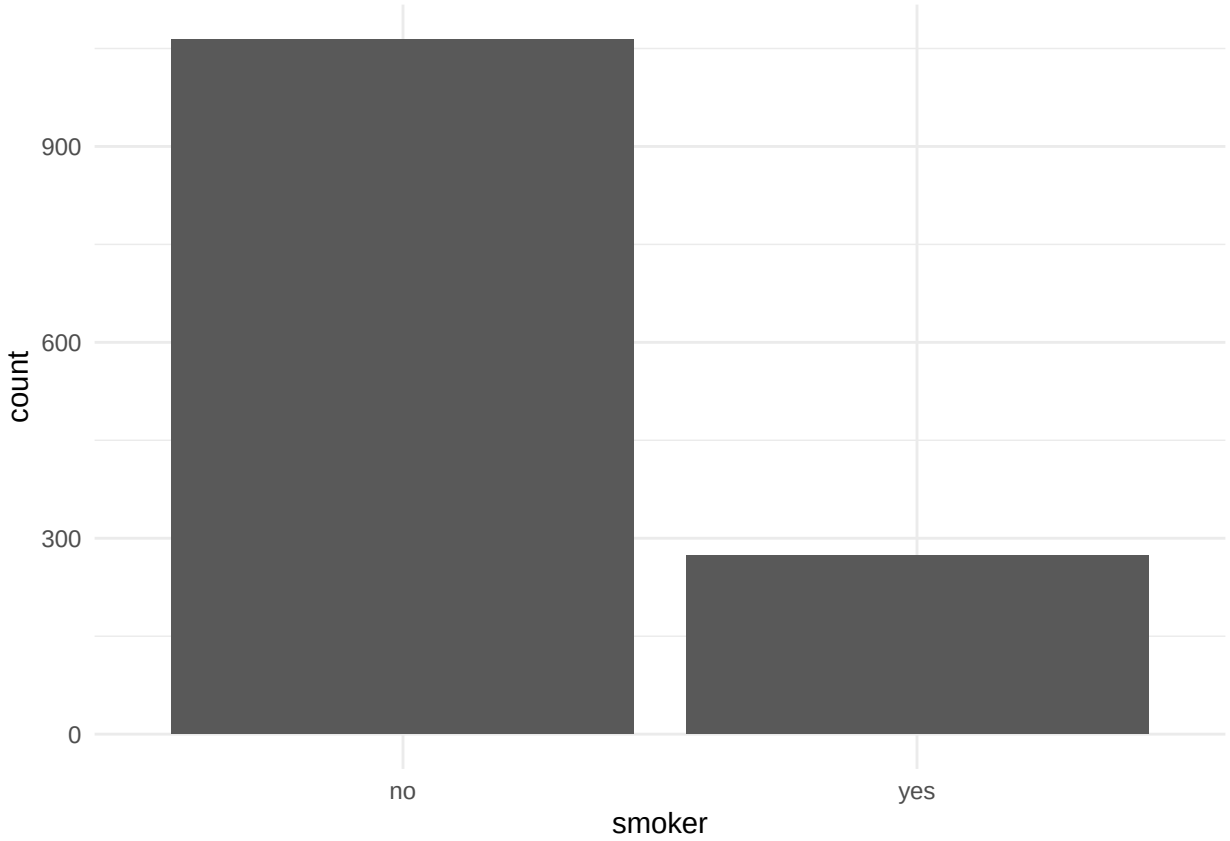
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Barplot

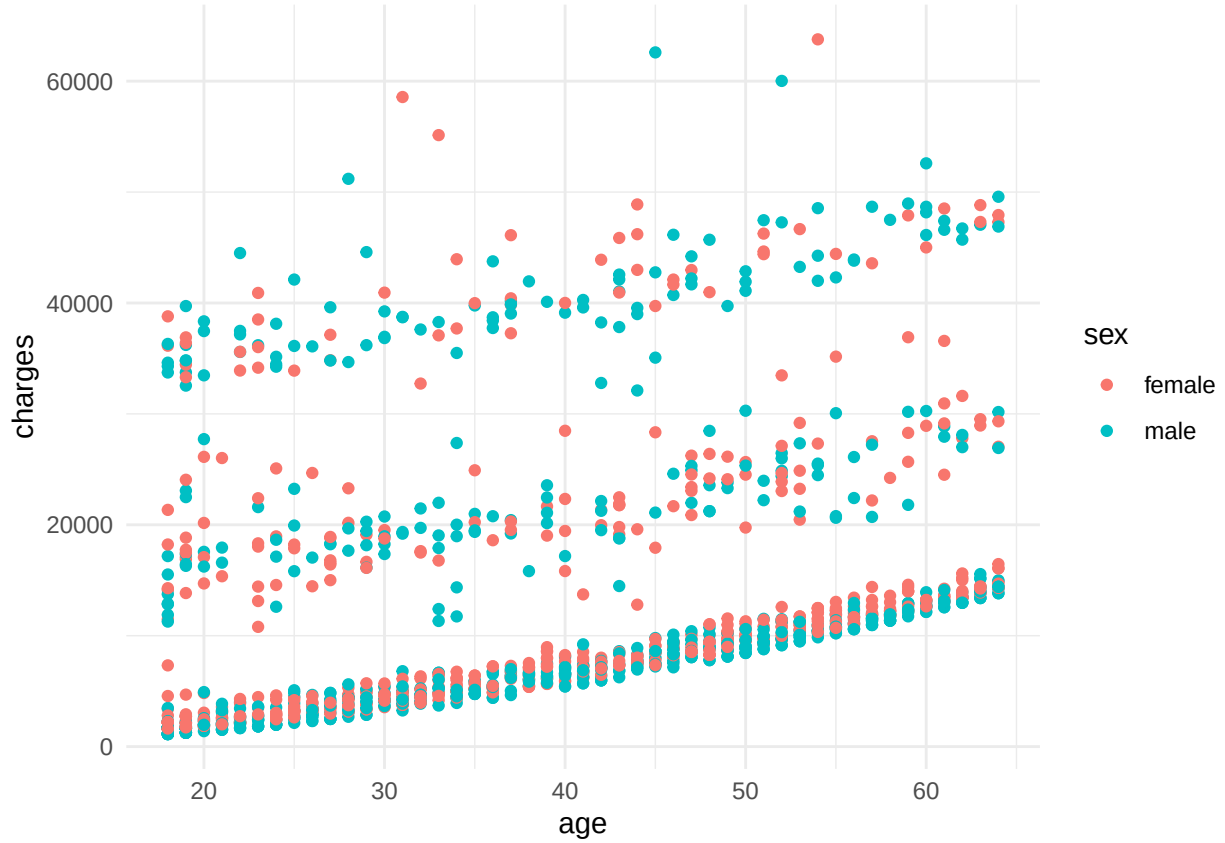
Berikut ini adalah contoh pengambaran barplot dengan fungsi `geom_bar()` dengan pada `ggplot2`

```
ggplot(asuransi) +
  aes(x = smoker) +
  geom_bar() +
  theme_minimal()
```



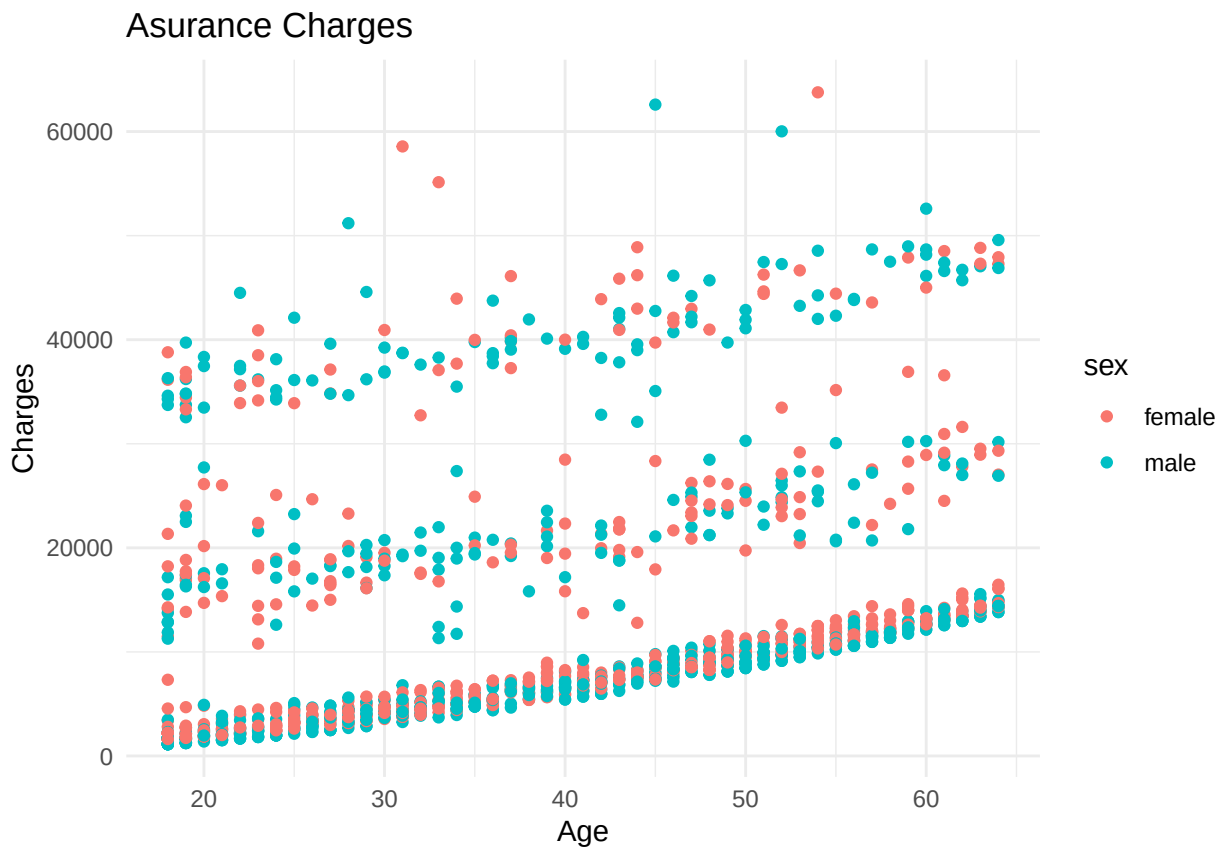
Scatterplot (Point) Berikut adalah contoh bagaimana pengabaran scatterplot dengan `ggplot2`

```
ggplot(asuransi)+
  aes(x = age, y = charges, colour = sex) +
  geom_point() +
  scale_color_hue() +
  theme_minimal()
```



Contoh lain dari ggplot dengan menambahkan judul dari grafik

```
ggplot(asuransi)+
  aes(x = age, y = charges, colour = sex) +
  geom_point() +
  scale_color_hue() +
  labs(x="Age",
       y="Charges",
       title="Asurance Charges") +
  theme_minimal()
```



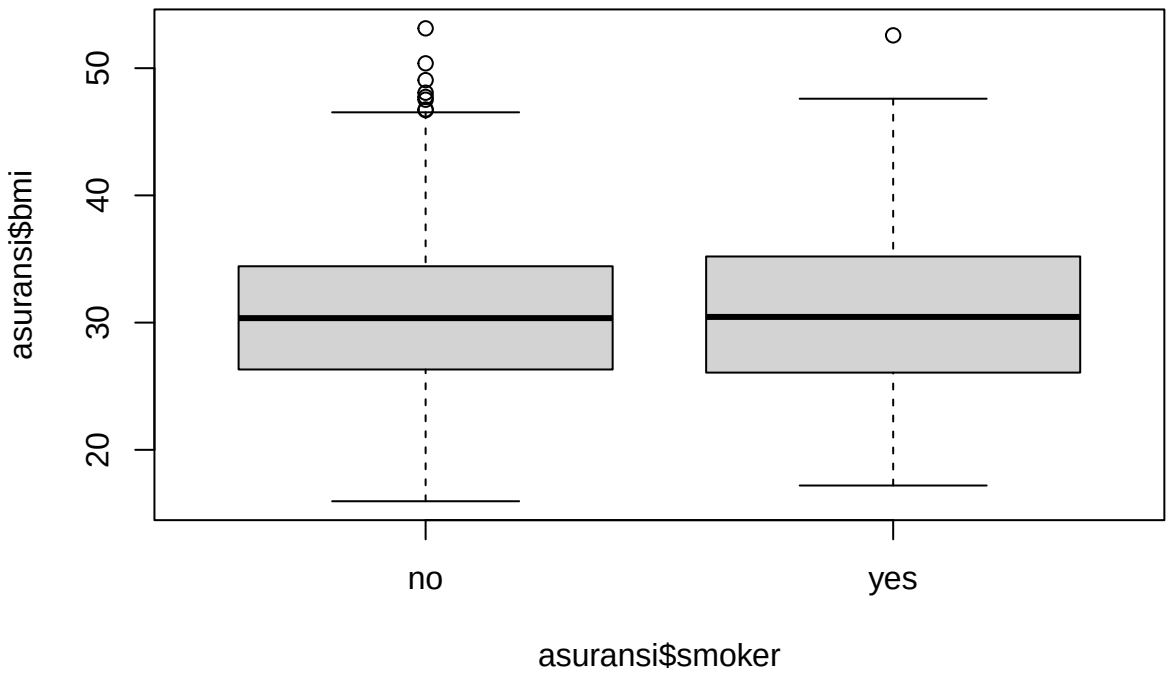
Latihan

```
# Korelasi variabel charges dengan variabel bmi
cor (asuransi$charges, asuransi$bmi)
```

```
## [1] 0.198341
```

Bisa dilihat bahwa nilai korelasinya 0,19, sehingga bisa disimpulkan bahwa hubungan kedua variabel lemah

```
# boxplot dari variabel bmi dan smoker
boxplot(asuransi$bmi ~ asuransi$smoker)
```


asuransi$bmi
20
30
40
50
no
yes
asuransi$smoker