



Universitas
Esa Unggul

**MODUL MATA KULIAH
DATA MINING
(MIK 620 SESI 10)**

Universitas
Esa Unggul

DISUSUN OLEH

NOVIANDI, M.Kom
NIDN. 0318018202

**PROGRAM STUDI MANAJEMEN INFORMASI KESEHATAN
FAKULTAS ILMU-ILMU KESEHATAN
UNIVERSITAS ESA UNGGUL
2018**

BAB II

DATA DAN EKSPLORASI DATA

Capaian pembelajaran

1. Mahasiswa mampu menjelaskan definisi data, menjelaskan tipe-tipe data dan melakukan eksplorasi data dengan menggunakan bahasa program R
2. Mahasiswa mampu memproses awal terhadap data hingga menjadi inputan dalam teknik data mining

Pendahuluan

Beberapa isu-isu yang terkait dengan data penting untuk suksesnya proses data mining, yaitu:

1. Tipe data
2. Kualitas data
3. Langkah preprocessing untuk membuat data lebih sesuai untuk data mining.
4. Menganalisis data dalam bentuk relasinya

Tipe Data

Sebuah data set dapat dipandang sebagai sebuah koleksi dari objek-objek data. Nama lain dari sebuah objek data adalah *record*, titik, vector, pola, *event*, *case*, *sample*, observasi atau entitas. Objek-objek data dijelaskan oleh sejumlah atribut yang menangkap karakteristik dasar dari sebuah objek.

Contoh: Informasi mahasiswa

Data set adalah sebuah file, dimana objek adalah *record-record* (baris) dalam file dan setiap *field* (kolom) berkaitan dengan sebuah atribut.

Tabel 2.1 *Data set* mahasiswa

Student ID	Year	Grade Point Average (GPA)
.....
1034262	Senior	3.24
1052663	Sophomore	3.51

1082246	Freshman	3.62
....

Selain *record-based*, *data set* juga bisa dalam bentuk *flat file* atau sistem basis data relasional.

Atribut dan Pengukuran

Atribut adalah sebuah sifat atau karakterik dari sebuah objek yang dapat bervariasi, baik dari satu objek ke objek lain atau dari satu waktu ke waktu yang lain.

Skala pengukuran adalah aturan (fungsi) yang menghubungkan nilai numerik atau simbolik dengan sebuah atribut dari sebuah objek. Proses pengukuran adalah penggunaan skala pengukuran untuk menghubungkan sebuah nilai dengan sebuah atribut tertentu dari sebuah objek.

Contoh:

Menghitung banyaknya kursi dalam sebuah ruangan untuk melihat apakah terdapat cukup empat duduk untuk semua orang yang akan datang pada sebuah pertemuan. Dalam kasus tersebut, nilai fisik dari sebuah atribut dari sebuah objek dipetakan ke sebuah nilai numerik atau simbolik.

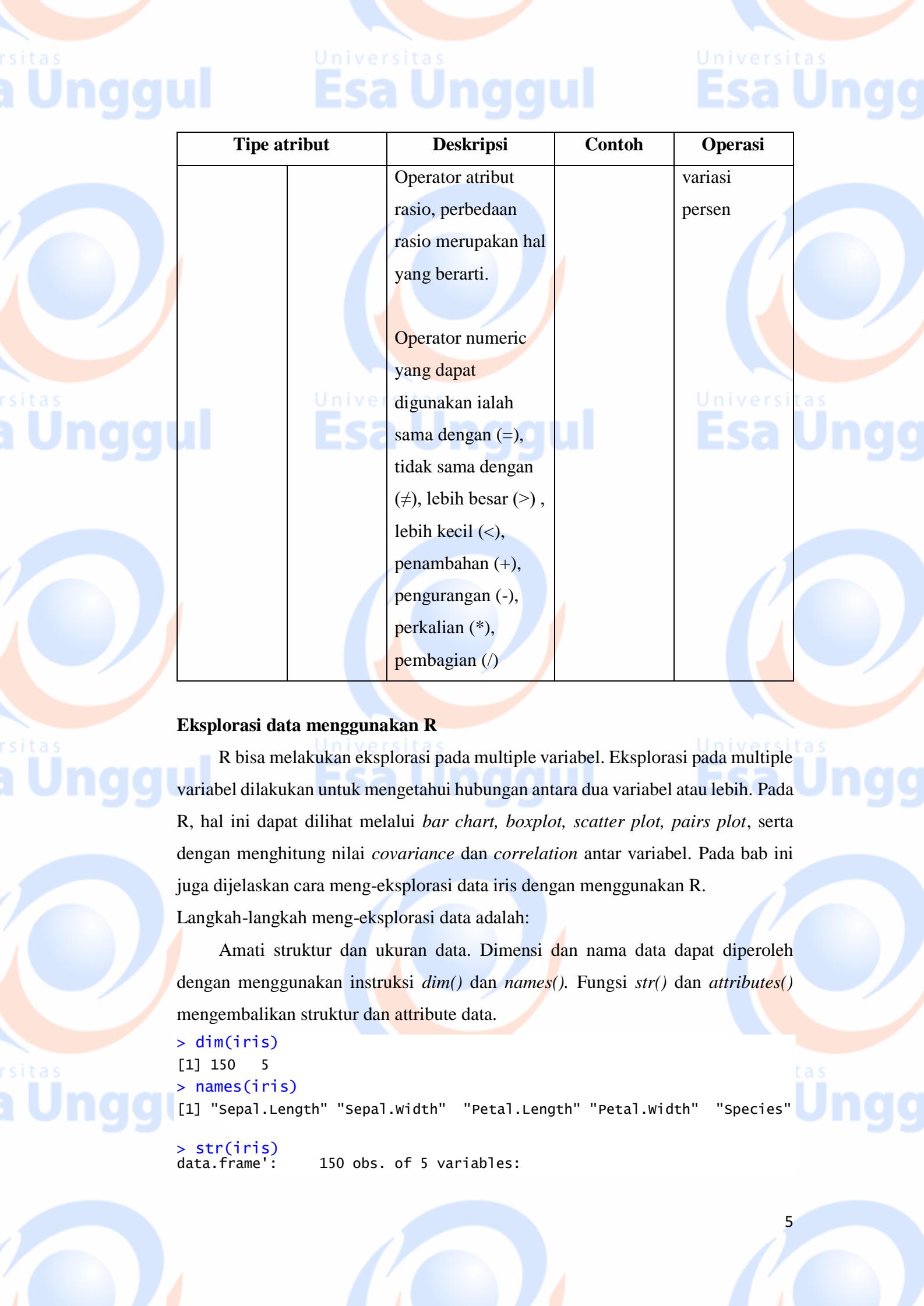
Tipe dari atribut

Tabel 2.2 Tipe-tipe atribut yang berbeda

Tipe atribut	Deskripsi	Contoh	Operasi
Kategori (Kualitatif)	Nominal <ul style="list-style-type: none">▪ Nilai dari atribut nominal adalah nama-nama yang berbeda, yaitu nilai nominal hanya menyediakan informasi yang	Kode pos, ID Number karywan, warna mata, jenis kelamin	Mode, entropy, contingency correlation, uji X^2

Tipe atribut	Deskripsi	Contoh	Operasi
	<p>cukup untuk membedakan satu objek dengan objek yang lain. (= dan ≠).</p> <ul style="list-style-type: none">▪ Atribut yang nilainya berupa simbol-simbol atau nama-nama benda dan nilainya tidak memiliki urutan yang memiliki arti.▪ Nilai pada atribut tidak dapat melakukan operasi matematika.		
Ordinal	<ul style="list-style-type: none">▪ Nilai dari atribut ordinal menyediakan informasi yang cukup mengurutkan objek (< , >)▪ Atribut dengan nilai-nilai yang kemungkinan memiliki	Nomor antrian, <i>grade</i> , kecepatan prosesor {lambat, cepat, sangat cepat}	Median, presentil, rank correlation, run test, sign test

Tipe atribut	Deskripsi	Contoh	Operasi
	urutan yang mempunyai arti atau tingkatan (rangking), akan tetapi jarak antara nilai tidak diketahui.		
Numerik (Kuantitatif)	Interval Dalam atribut interval perbedaan antar nilai merupakan sesuatu yang berarti, pada atribut interval terdapat unit pengukuran. Operator aritmatik yang dapat digunakan ialah sama dengan (=), tidak sama dengan (\neq), lebih besar ($>$), lebih kecil ($<$), penambahan (+), pengurangan (-).	Tanggal pada kalender, temperature	Rataan, simpangan baku, korelasi pearson, uji T dan F
	Ratio Dalam atribut rasio, perbedaan rasio merupakan hal yang berarti.	Umur, berat badan, tinggi badan	Rataan geometri, rataan harmonik,



Tipe atribut	Deskripsi	Contoh	Operasi
	<p>Operator atribut rasio, perbedaan rasio merupakan hal yang berarti.</p> <p>Operator numeric yang dapat digunakan ialah sama dengan (=), tidak sama dengan (\neq), lebih besar ($>$), lebih kecil ($<$), penambahan (+), pengurangan (-), perkalian (*), pembagian (/)</p>		variasi persen

Eksplorasi data menggunakan R

R bisa melakukan eksplorasi pada multiple variabel. Eksplorasi pada multiple variabel dilakukan untuk mengetahui hubungan antara dua variabel atau lebih. Pada R, hal ini dapat dilihat melalui *bar chart*, *boxplot*, *scatter plot*, *pairs plot*, serta dengan menghitung nilai *covariance* dan *correlation* antar variabel. Pada bab ini juga dijelaskan cara meng-eksplorasi data iris dengan menggunakan R.

Langkah-langkah meng-eksplorasi data adalah:

Amati struktur dan ukuran data. Dimensi dan nama data dapat diperoleh dengan menggunakan instruksi *dim()* dan *names()*. Fungsi *str()* dan *attributes()* mengembalikan struktur dan attribute data.

```
> dim(iris)
[1] 150 5
> names(iris)
[1] "Sepal.Length" "Sepal.Width" "Petal.Length" "Petal.Width" "Species"
> 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
1 ...

> attributes(iris)
$names
[1] "Sepal.Length" "Sepal.width" "Petal.Length" "Petal.width" "Species"

$row.names
[1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
[18] 19 20
[21] 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37
[38] 39 40
[41] 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57
[58] 59 60
[61] 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77
[78] 79 80
[81] 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97
[98] 99 100
[101] 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117
[118] 119 120
[121] 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137
[138] 139 140
[141] 141 142 143 144 145 146 147 148 149 150

$class
[1] "data.frame"
```

Selanjutnya, kita harus melihat lima baris data pertama. Baris data pertama atau terakhir dapat diambil dengan head () atau tail ().

```
> iris[1:5,]
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
```

Langkah selanjutnya juga dapat mengambil nilai dari satu kolom. Misalnya, mengambil 10 nilai pertama dari *sepal.length* dengan cara menggunakan satu kode dibawah ini.

```
> 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
```

```
> tail(iris)
Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
145          6.7         3.3         5.7         2.5 virginica
146          6.7         3.0         5.2         2.3 virginica
147          6.3         2.5         5.0         1.9 virginica
148          6.5         3.0         5.2         2.0 virginica
149          6.2         3.4         5.4         2.3 virginica
150          5.9         3.0         5.1         1.8 virginica
```

Cara mengambil 10 nilai pertama dari Sepal.Length dapat diambil dengan salah satu kode dibawah ini.

```
> iris[1:10,"Sepal.Length"]  
[1] 5.1 4.9 4.7 4.6 5.0 5.4 4.6 5.0 4.4 4.9
```

```
> iris$Sepal.Length[1:10]  
[1] 5.1 4.9 4.7 4.6 5.0 5.4 4.6 5.0 4.4 4.9
```

Langkah-langkah meng-eksplor *individual variables*

Distribusi setiap variabel numeric dapat diperiksa dengan fungsi *summary()*, dimana fungsi tersebut dapat menampilkan nilai minimum, maksimum, rata-rata, median, kuartil pertama (25%) dan kuartil ketiga (75%).

```
> summary(iris)  
Sepal.Length Sepal.Width Petal.Length Petal.Width Species  
Min. :4.300 Min. :2.000 Min. :1.000 Min. :0.100 setosa  
1st Qu.:5.100 1st Qu.:2.800 1st Qu.:1.600 1st Qu.:0.300 versicolor  
Median :5.800 Median :3.000 Median :4.350 Median :1.300 virginica  
3rd Qu.:6.400 3rd Qu.:3.300 3rd Qu.:5.100 3rd Qu.:1.800  
Max. :7.900 Max. :4.400 Max. :6.900 Max. :2.500
```

Untuk menampilkan nilai *mean*, *median* dan *range* juga dapat menggunakan perintah dibawah ini:

```
> quantile(iris$Sepal.Length)  
0% 25% 50% 75% 100%  
4.3 5.1 5.8 6.4 7.9
```

Keterangan:

dari nilai kuartil diatas, maka:

nilai min = 4.3
nilai Q1 = 5.1
nilai Q2 = 5.8
nilai Q3 = 6.4
nilai maksimum = 7.9

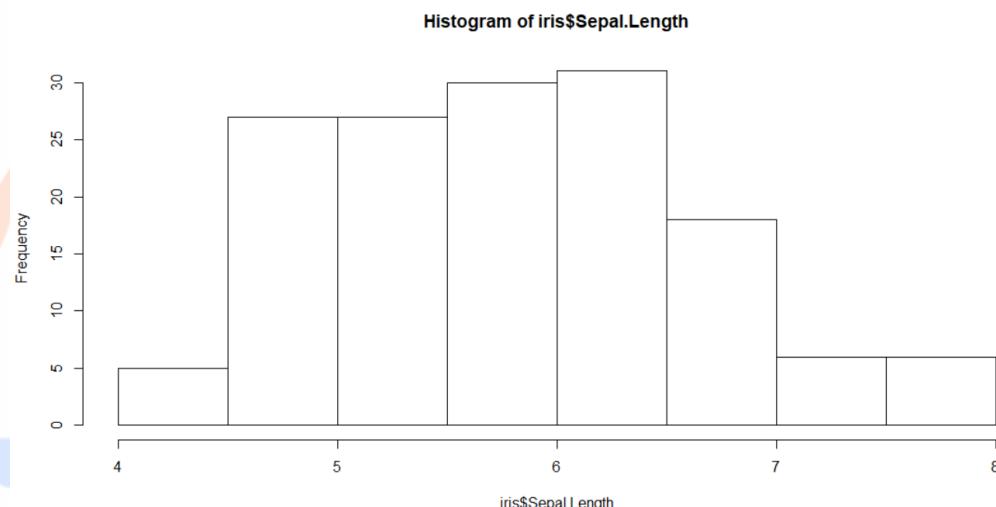
```
> quantile(iris$Sepal.Length, c(.1, .3, .65))  
10% 30% 65%  
4.80 5.27 6.20
```

Periksa *variance* dari Sepal.Length dengan *var()* dan juga memeriksa distribusi dengan *histogram* dan *density* dengan fungsi *hist()* dan *density()*

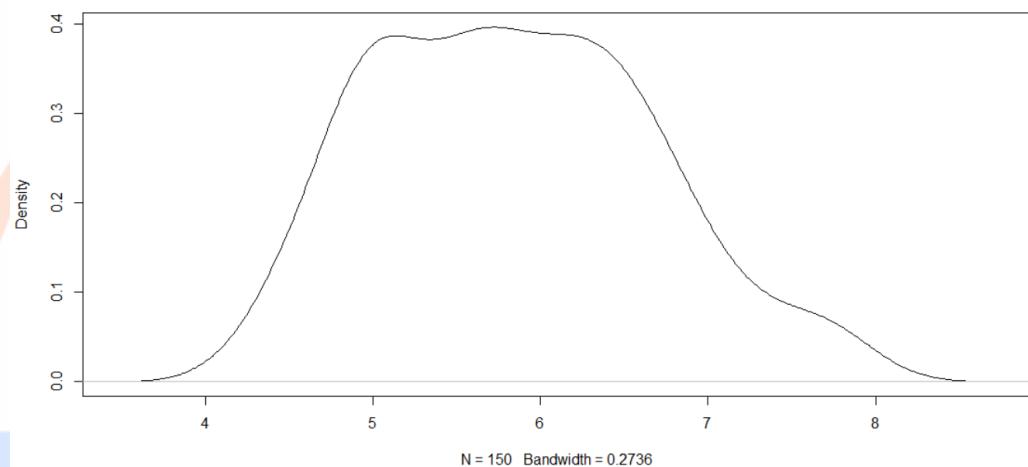
```
> var(iris$Sepal.Length)  
[1] 0.6856935
```

```
> hist(iris$Sepal.Length)
```

Plot Zoom



```
> plot(density(iris$Sepal.Length))  
density.default(x = iris$Sepal.Length)
```

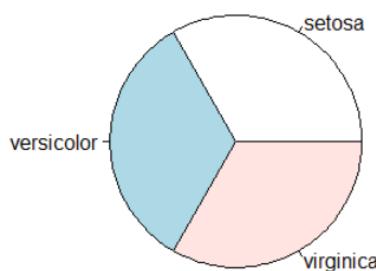


Frekuensi faktor dapat dihitung dengan fungsi `table()`, lalu diplot dalam bentuk pie chart dengan fungsi `pie()` atau bar chart dengan fungsi `barplot()`.

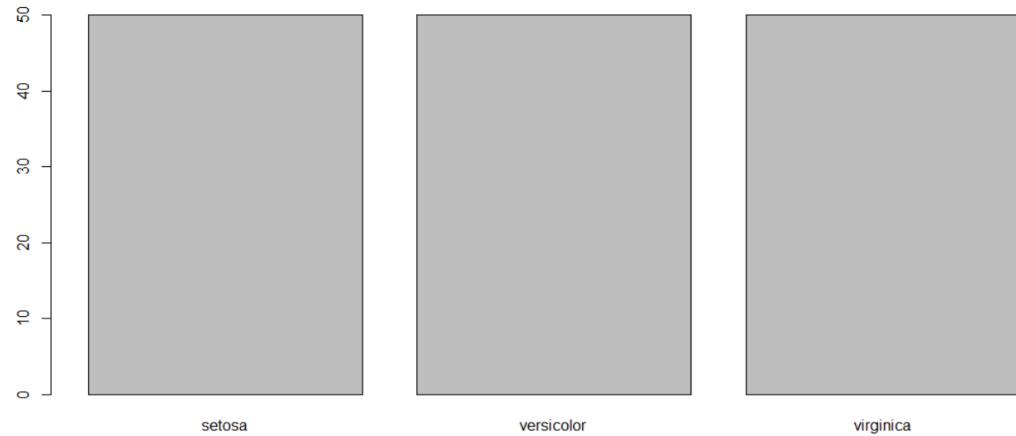
```
> table(iris$Species)
```

Species	Count
setosa	50
versicolor	50
virginica	50

```
> pie(table(iris$Species))
```



```
> barplot(table(iris$Species))
```



Explore Multiple Variable

Setelah memeriksa variabel individual, kemudian menyelidiki hubungan antara dua variabel. Berikut menghitung kovarians (digunakan untuk mengukur besarnya hubungan yang dapat terjadi antara dua variabel) dan korelasi antar variabel dengan menggunakan formula cov() dan cor().

```
> cov(iris$Sepal.Length, iris$Petal.Length)
[1] 1.274315
```

```
> cov(iris[,1:4])
Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    0.6856935 -0.0424340   1.2743154   0.5162707
Sepal.Width     -0.0424340   0.1899794  -0.3296564  -0.1216394
Petal.Length    1.2743154  -0.3296564   3.1162779   1.2956094
Petal.Width     0.5162707  -0.1216394   1.2956094   0.5810063
```

Berikutnya, menghitung Sepal.Length di setiap species dengan aggregate()

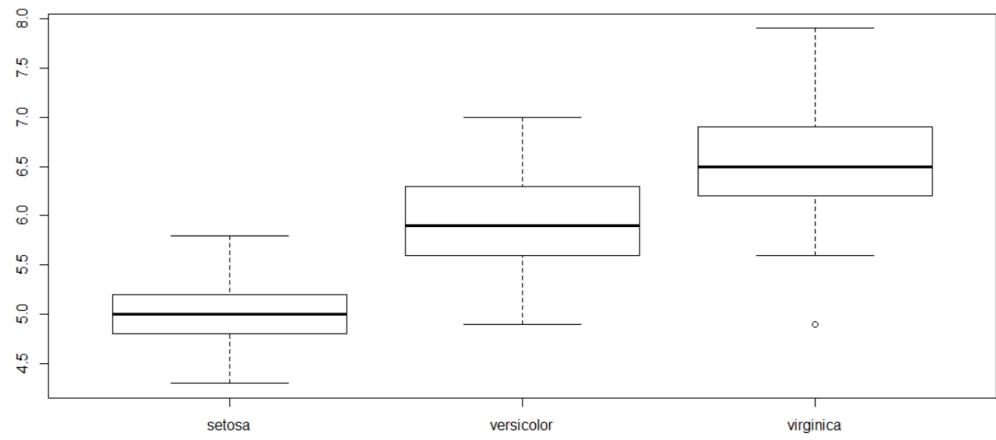
```
> aggregate(Sepal.Length ~ Species, summary, data=iris)
```

Species	Sepal.Length.Min.	Sepal.Length.1st Qu.	Sepal.Length.Median	Sepal.Length.Mean	Sepal.Length.3rd Qu.	Sepal.Length.Max.
1 setosa	4.300		4.800	5.000		
2 versicolor	4.900		5.600	5.900		
3 virginica	4.900		6.225	6.500		

Species	Sepal.Length.Mean	Sepal.Length.3rd Qu.	Sepal.Length.Max.
1	5.006	5.200	5.800
2	5.936	6.300	7.000
3	6.588	6.900	7.900

Langkah selanjut nya menggunakan fungsi boxplot untuk melihat nilai median, kuartil pertama, kuartil ketiga dari suatu distribusi dan nilai outlier.

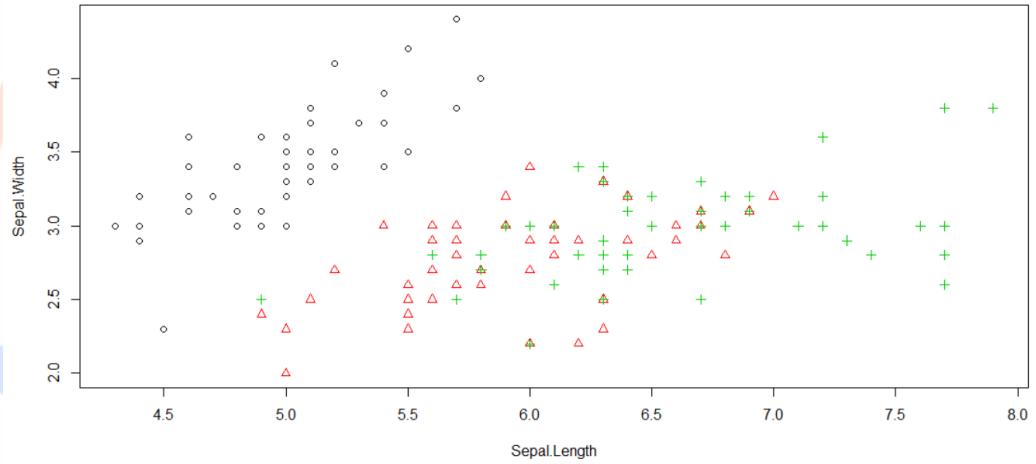
```
> boxplot(Sepal.Length~Species, data=iris)
```



Scatter plot untuk dua variabel numeric dapat ditarik dengan `plot()`, menggunakan fungsi `with()`, tanpa menambahkan “`iris$`” sebelum variabel names.

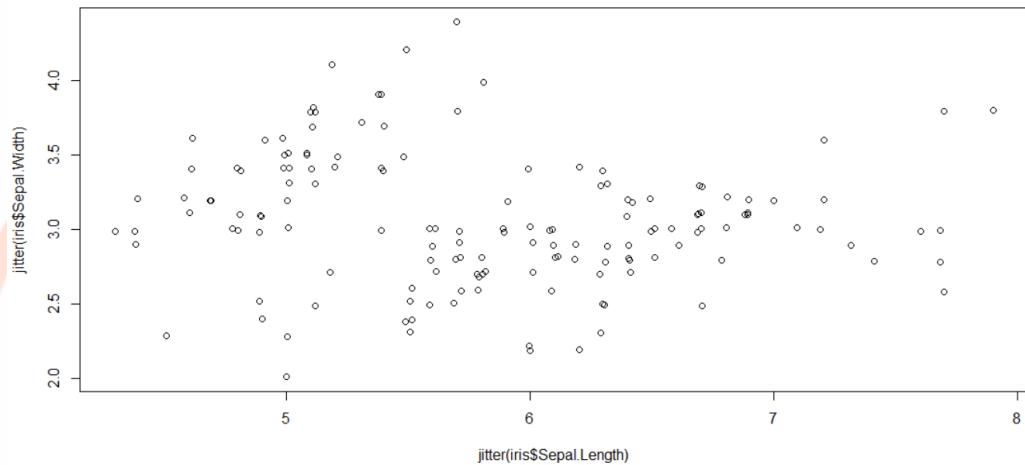
Menampilkan warna adalah `(col)` dan simbol `(pch)` yang diatur untuk `species`.

```
> with(iris, plot(Sepal.Length, Sepal.Width, col=Species, pch=as.numeric(Species)))
```



Jika ingin menampilkan satu jenis poin, maka gunakan fungsi `jitter()`, seperti dibawah ini.

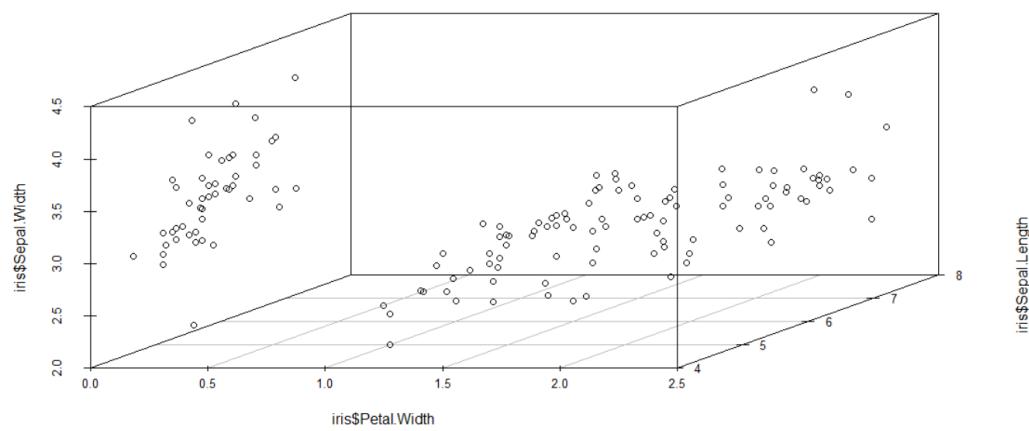
```
> plot(jitter(iris$Sepal.Length), jitter(iris$Sepal.width))
```



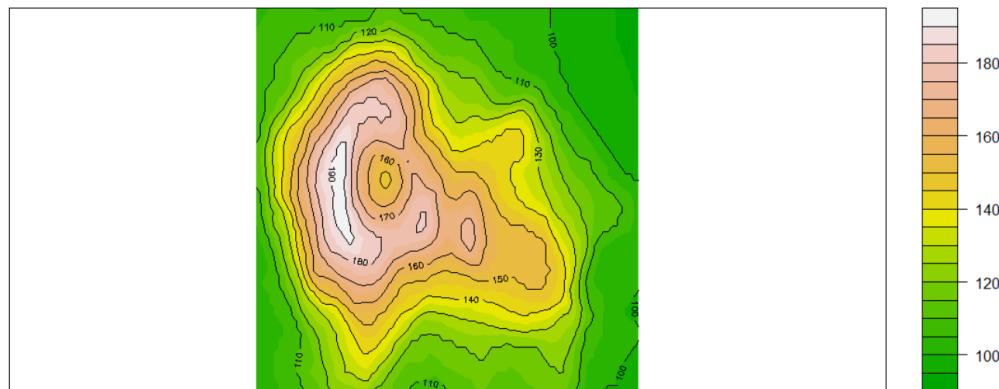
More Exploration

Selain bentuk grafik diatas, data iris juga dapat dieksplor atau disajikan dalam bentuk 3D, level plots, contour plots, interactive plots dan parallel coordinates. Scatter plot 3D dapat dihasilkan dengan menggunakan *package scatterplot3d* [Ligges and Machler, 2003].

```
> library("scatterplot3d", lib.loc="~/R/win-library/3.3")
> scatterplot3d(iris$Petal.Width, iris$Sepal.Length, iris$Sepal.Width)
```



```
> library("lattice", lib.loc="C:/Program Files/R/R-3.3.3/library")
> filled.contour(volcano, color=terrain.colors, asp=1, plot.axes = contour(volcano, add = T))
```



Cara lain untuk mengilustrasikan matriks numerik adalah plot permukaan 3D yang ditunjukkan seperti di bawah ini, yang dihasilkan dengan fungsi persp().

```
> persp(volcano, theta=25, phi=30, expand=0.5, col="lightblue")
```

