EXPLORATORY DATA ANALYSIS

What is Exploratory Data Analysis?

- EDA is an approach for data analysis using variety of techniques to gain insights about the data.
- Basic steps in any exploratory data analysis:
  - ✓Cleaning and preprocessing
  - ✓Statistical Analysis
  - ✓Visualization for trend analysis, anomaly detection, outlier detection (and removal).

Importance of EDA

Improve understanding of variables by extracting averages, variance, mean, minimum, and maximum values, range, etc.

Discover errors, outliers, and missing values in the data.

Identify patterns by visualizing data in graphs such as box plot, bar graphs, pie chart, scatter plots, heatmaps, histograms, etc.

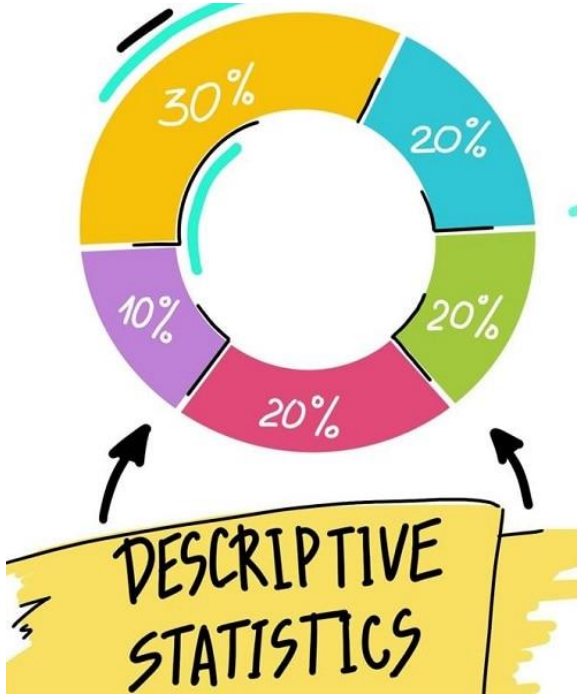
Statistika Deskriptif

- Statistika deskriptif mencakup **ringkasan data** yang dapat digunakan untuk memahami dan mengeksplorasi data.
- Ringkasan data dapat disajikan dengan nilai numerik untuk lokasi atau pusat data dan jumlah variabilitas yang ada.

Statistika Deskriptif

Ukuran Pemusatan

- *Mean*
- Median
- Modus



Ukuran Penyebaran

- Variansi
- *Standard deviation*
- *Range*
- *Inter Quartile Range* (IQR)
- *Skewness*
- Kurtosis

Ukuran Pemusatan Data

Mean

- Sampel : $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$
- Populasi : $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$

dengan *n* adalah ukuran sampel, *N* adalah ukuran populasi

- Mean merangkum semua informasi dalam data.
- Mean adalah satu titik yang dapat dilihat sebagai titik pusat massa data.

Mean memiliki beberapa **sifat matematika** yang membuatnya berguna dalam banyak konteks inferensi statistik.

- Mean lebih **rentan** terhadap data yang memuat ***outlier***, sedangkan median lebih tahan terhadap pengamatan ekstrim.

Median

- Median adalah **nilai tengah** data dalam artian separuh data berada di bawahnya dan separuh di atasnya. Median dapat dinotasikan Q_2.
- Menghitung persentil sampel:

1. Data diurutkan dari nilai terkecil hingga terbesar
2. Menghitung perkalian (*ukuran sampel)* x (*proporsi)* = *np*
   - *Jika np bukan bilangan integer, maka bulatkan ke atas dan temukan nilai pada urutan tersebut.*
   - *Jika np integer, disebut k, hitung rata-rata data ke-k dan (k+1)*

Q_1, Q_2, dan Q_3 masing-masing adalah persentil ke-25, 50, dan 75 (Richard and Bhattacharyya, 2010).

Modus

- Modus kumpulan data adalah nilai yang paling **sering muncul.**
- Nilai yang paling sering muncul terlihat dari frekuensi data.

Ukuran Penyebaran Data

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 \Big/ n}{n-1}$$

Range

selisih antara pengamatan terbesar dan pengamatan terkecil

Interquartile Range (IQR)

- IQR dapat dinyatakan sebagai selisih antara kuartil atas (Q_3) dan kuartil bawah (Q_1).

$$IQR = Q_3 - Q_1$$

- IQR lebih tahan terhadap pengamatan yang ekstrim daripada *range*.

Variansi

ialah deviasi kuadrat rata-rata dari titik data dari *mean-nya*. Variansi sampel s^2 dan variansi populasi .

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \qquad \sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

Deviasi Standar

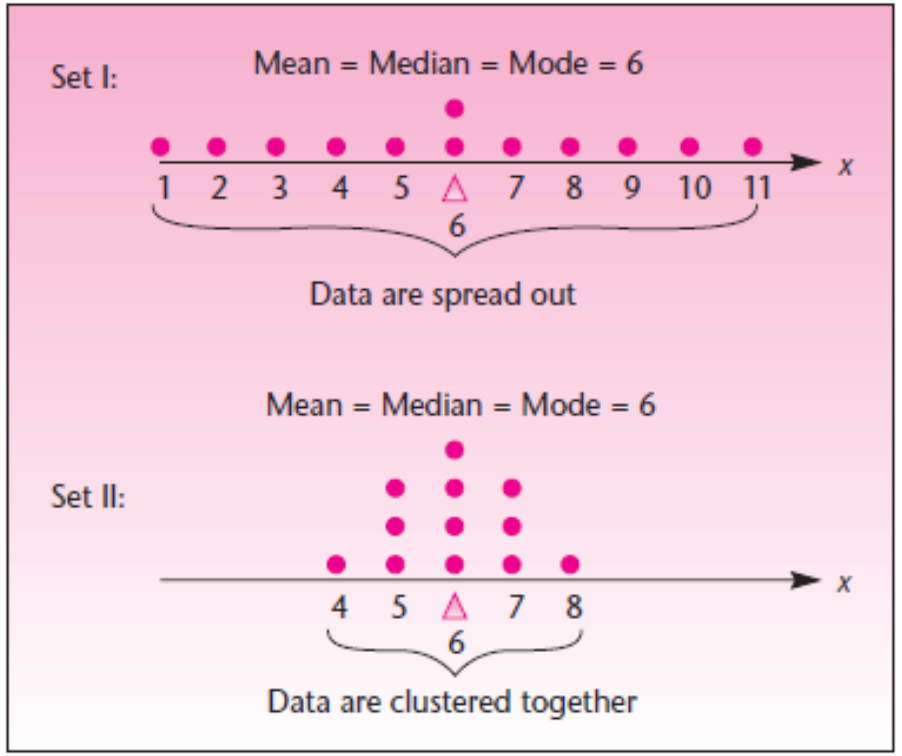
akar kuadrat dari variansi

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} \qquad \sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$

Variansi vs St. Dev vs Range vs IQR

- Variansi lebih suka digunakan oleh ahli statistika karena memiliki **sifat matematis** yang menyederhanakan perhitungan.
- Orang yang menerapkan statistika *(users)* lebih suka bekerja dengan deviasi standar karena lebih **mudah diinterpretasikan** (memiliki satuan ukuran yang sama dengan nilai mean).
- Variansi dan deviasi standar lebih berguna daripada *range* dan IQR karena seperti mean yang menggunakan **informasi** dari **semua pengamatan** dalam kumpulan data atau populasi.

Contoh Ukuran Pemusatan dan Penyebaran Data



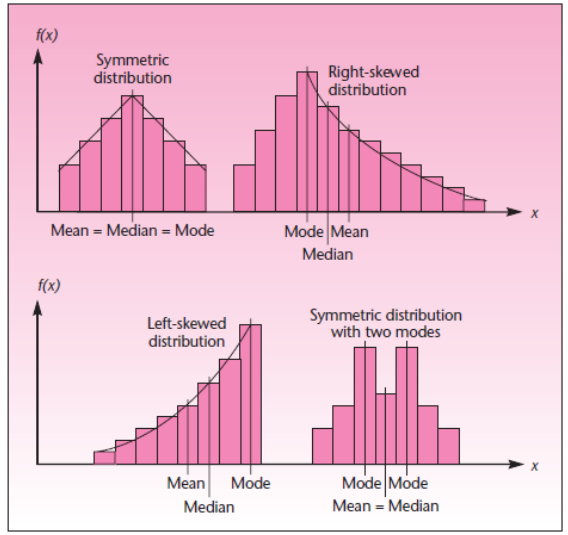
Aczel dan Sounderpandian (2008), hal.16

Secara visual, set data manakah yang lebih menyebar/ bervariasi?

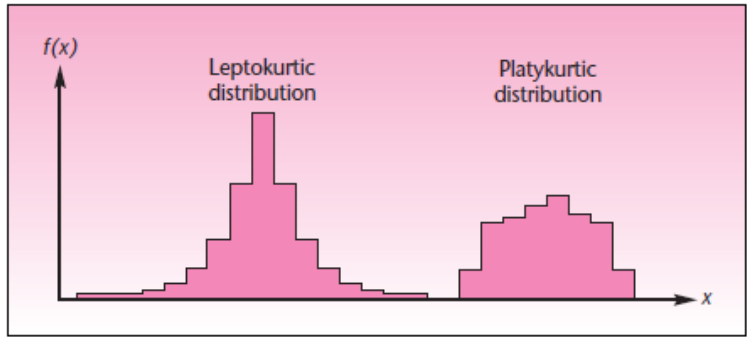
Bagaimana ukuran penyebaran kedua set data di samping?
Variansi = ?
St. dev = ?
Range = ?
IQR = ?

Skewness dan Kurtosis

- **Skewness** atau kemiringan adalah ukuran derajat asimetri suatu distribusi frekuensi.
- Distribusi miring ke kanan (**skewness positif**) ketika distribusi meregang/ condong ke kanan lebih dari ke kiri, disebut.
- Distribusi miring kiri (**skewness negatif**) adalah distribusi yang membentang secara asimetris ke kiri.
- Distribusi **simetris** (skewness nol) dengan modus tunggal memiliki median, mean, modus yang sama.



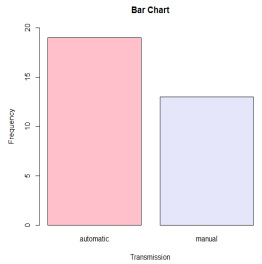
- **Kurtosis** adalah ukuran puncak suatu distribusi.
- Kurtosis dapat dibedakan menjadi nilai absolut atau relatif. Kurtosis absolut selalu berupa angka positif. Kurtosis absolut dari distribusi normal adalah 3. Nilai 3 ini diambil untuk menghitung kurtosis relatif.
- Kurtosis relatif bisa bernilai negatif. Kurtosis negatif menyiratkan distribusi yang lebih datar daripada distribusi normal atau disebut *platykurtic*. Kurtosis positif menyiratkan distribusi yang lebih memuncak atau runcing dibandingkan distribusi normal atau disebut leptokurtik.

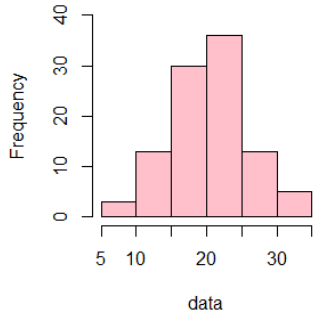


Aczel dan Sounderpandian (2008)

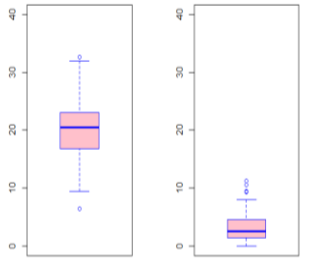
Visualization

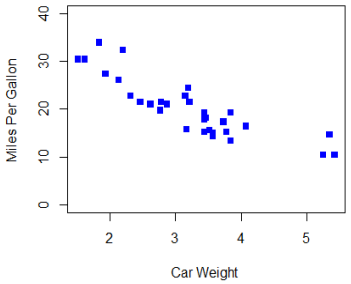
- Univariate: Looking at one variable/column at a time
  - Bar-graph
  - Histograms
  - Stem and Leaf plots
  - Boxplot
- Multivariate : Looking at relationship between two or more variables
  - Scatter plots
  - Time plots
  - Pie plots
  - Heatmaps (seaborn)

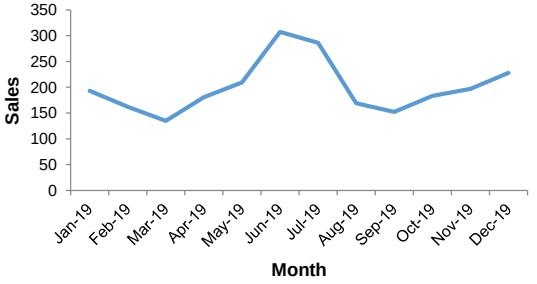


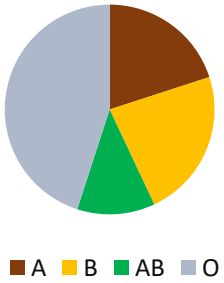



```
0 | 7
1 | 000122223444444
1 | 566667777777888999999999
2 | 000000011111111112222222222223333333
2 | 555555555677799
3 | 000223
```



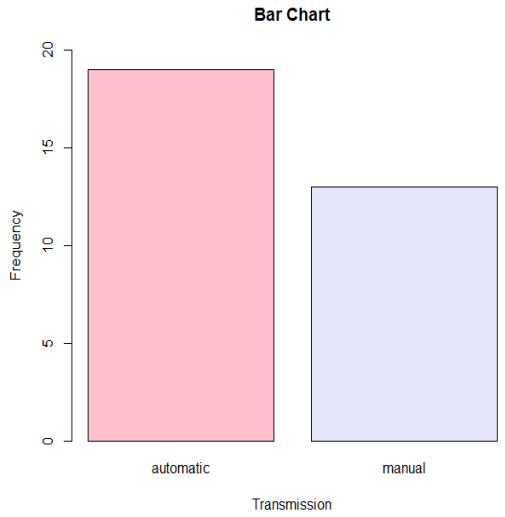






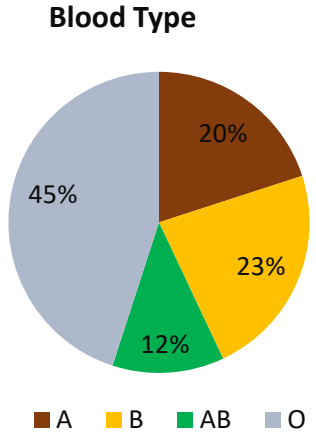
Bar Chart

- Menggunakan persegi panjang horizontal atau vertikal sering digunakan untuk menampilkan data kualitatif/kategorikal dan tidak ada penekanan persentase total yang diwakili setiap kategori.
- Panjang batang menunjukkan frekuensi data.
- Skala pengukurannya nominal atau ordinal.



Pie Chart

- Tampilan deskriptif sederhana dari data kualitatif yang menjumlahkan total tertentu.
- Luas lingkaran mewakili 100% (dari semua kategori)
- Ukuran setiap irisan adalah persentase dari total yang diwakili oleh kategori.
- Skala pengukuran bisa nominal atau ordinal.



Histogram

FIGURE 1–5 A Histogram of the Data in Example 1–7

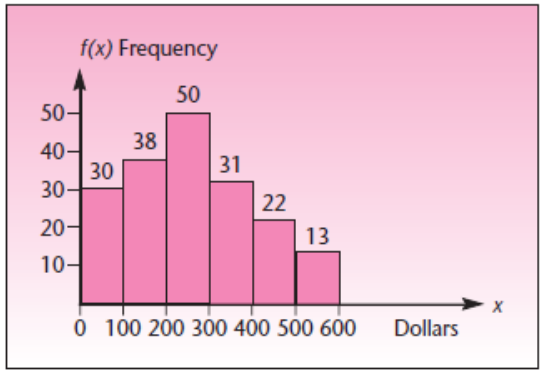
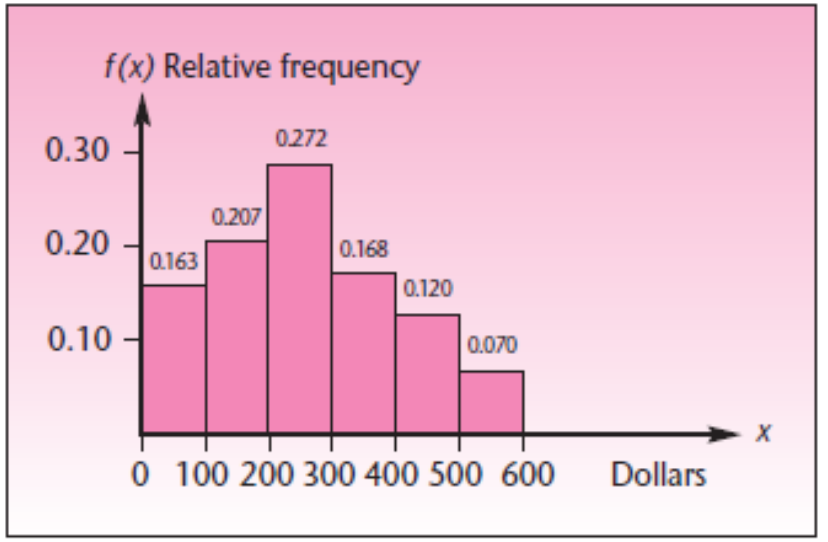


TABLE 1–6 Relative Frequencies for Example 1–7

x Class ($)	f(x) Relative Frequency
0 to less than 100	0.163
100 to less than 200	0.207
200 to less than 300	0.272
300 to less than 400	0.168
400 to less than 500	0.120
500 to less than 600	0.070
	1.000

Aczel dan Sounderpandian (2008)



- Histogram adalah bagan yang terbuat dari batang dengan ketinggian berbeda. Ketinggian setiap batang mewakili frekuensi nilai di kelas (kelompok) yang diwakili oleh batang tersebut.
- Histogram untuk memplot frekuensi data (frekuensi absolut atau jumlah titik data) yang dikelompokkan. Histogram juga dimungkinkan untuk memplot frekuensi relatif, yaitu jumlah titik data di kelas dibagi dengan jumlah total titik data.

Histogram


```
> mean(data)
[1] 20.20971
> median(data)
[1] 20.5279
> skewness(data)
[1] 0.01617045
> kurtosis(data)
[1] 3.037837
```

```
> mean(data2)
[1] 3.236877
> median(data2)
[1] 2.565741
> skewness(data2)
[1] 1.09224
> kurtosis(data2)
[1] 3.969029
```

- Histogram dapat menggambarkan pemusatan data, frekuensi data, sebaran data, dan kemiringan data. Namun, histogram menampilkan data dalam bentuk kelompok sehingga informasi tentang data mentah tidak tampak karena telah digabungkan dalam kelompok.

Stem and Leaf Plot

- *Stem-and-leaf* adalah penyajian data yang juga digunakan untuk melihat kumpulan data. Penyajian ini berisi beberapa fitur histogram, tetapi penyajiannya menghindarkan dari hilangnya informasi dalam histogram yang dihasilkan dari penggabungan data menjadi beberapa interval. *Stem* (batang) adalah angka tanpa digit paling kanan (*leaf* atau daun). Batang ditulis di sebelah kiri garis vertikal yang memisahkan batang dari daun.
- Misalnya, dimiliki angka 101, 102, 103, 104, 107, 108, maka ditampilkan sebagai

  10 | 123478

- Dengan kumpulan data yang lebih lengkap dengan nilai batang yang berbeda, digit terakhir dari setiap angka ditampilkan di tempat yang sesuai di sebelah kanan digit batangnya. Tampilan batang dan daun membantu mengidentifikasi angka dalam kumpulan data yang memiliki frekuensi tinggi dan menampilkan sebaran data, sama halnya dengan histogram.

Stem and Leaf Plot

```
> stem(data)

  The decimal point is 1 digit(s) to the right of the |

  0  | 7
  1  | 00012222344444
  1* | 566666777777788899999999
  2  | 00000000111111111122222222222333333
  2* | 555555555677799
  3  | 000223
```

STEM

LEAF

```
  0 | 7
  1 | 000122223444445666667777777888999999999
  2 | 00000000111111111222222222222333333355555555677799
  3 | 000223
```

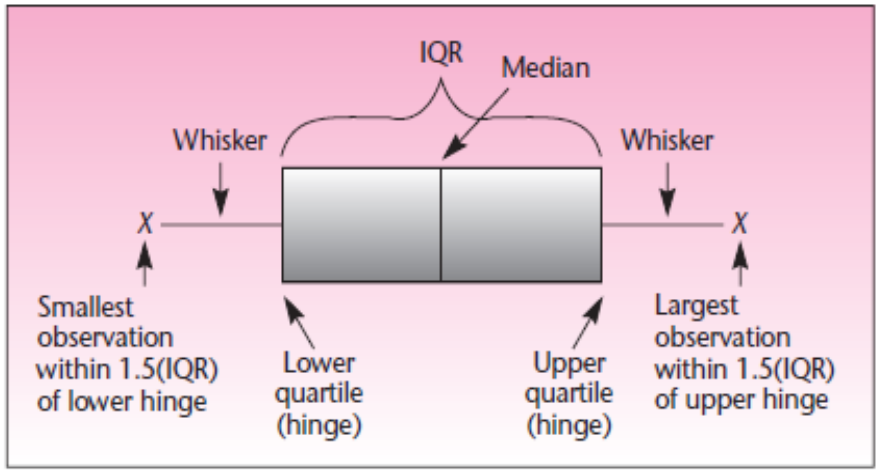
Stem and Leaf Plot

```
> stem(data2)

  The decimal point is at the |

   0 | 000002344677899022333344667888899999
   2 | 011122344455667889900124445667779 9
   4 | 34457799901256
   6 | 3455777155
   8 | 034
  10 | 63
```

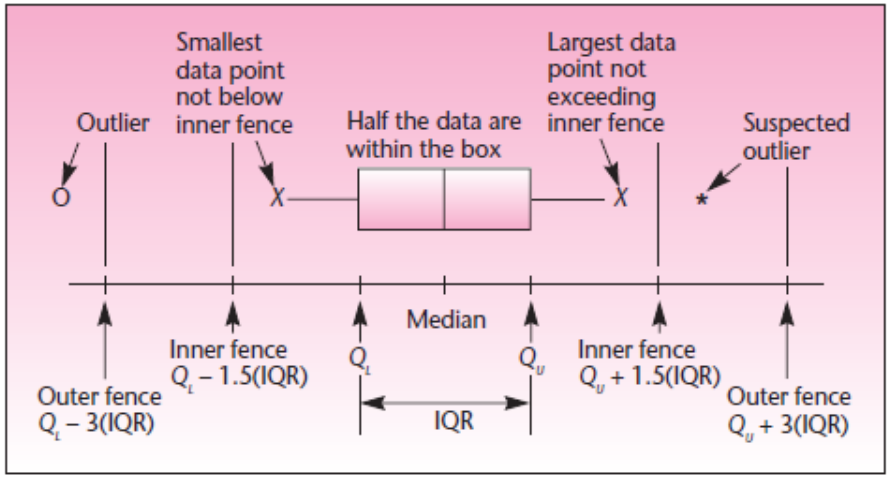
Box Plot

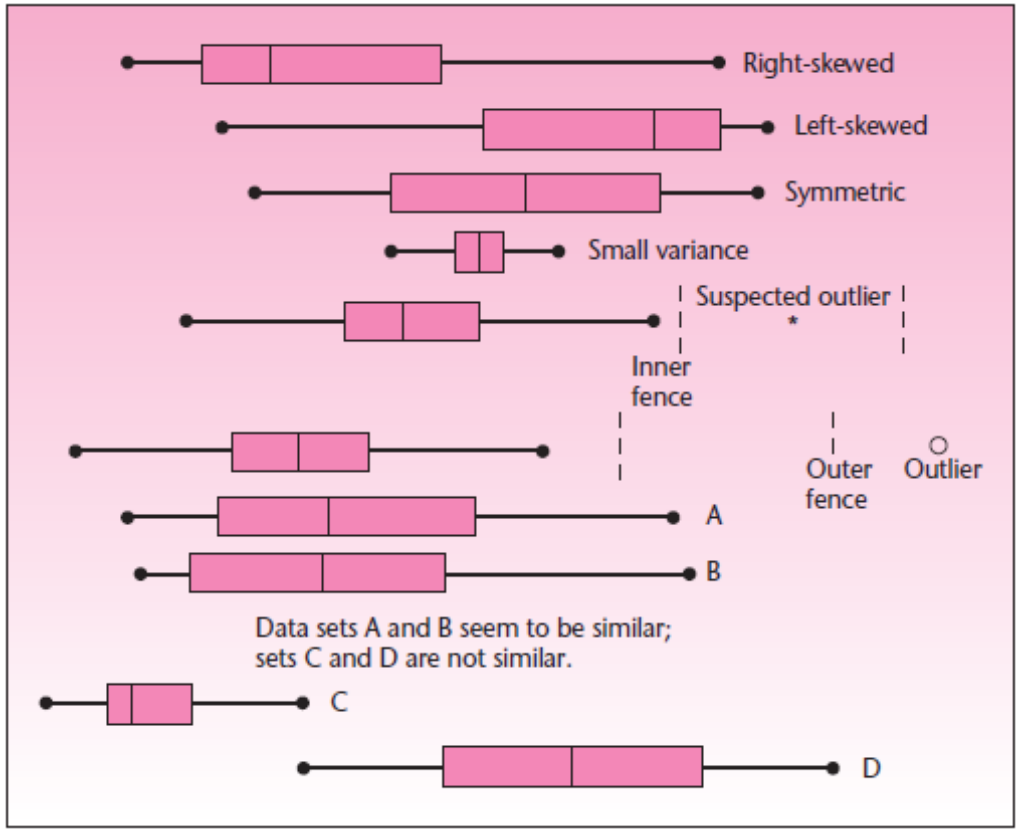
- *Box plot* dapat digunakan untuk menyajikan data dan menampilkan ukuran pemusatan data, penyebaran data, distribusi data, dan mendeteksi *outlier*.
- Garis yang berada di tengah *box* menunjukkan nilai median (Q_2)
- Ujung *box* bagian kiri atau bawah menunjukkan nilai *lower quartile* (Q_1) dan ujung *box* bagian kanan atau atas adalah nilai *higher quartile* (Q_3).
- Selisih antara Q_1 dengan Q_3 disebut *Interquartile Range* (IQR). Semakin besar nilai IQR, *box* akan semakin membesar berarti data semakin menyebar dari pemusatannya, begitupun sebaliknya.
- Garis yang memanjang dari *box* disebut *whiskers*, yaitu garis yang memanjang dari Q_3 hingga nilai observasi terbesar dan dari Q_1 hingga nilai observasi terkecil yang berada dalam jarak 1,5 IQR dari Q_3 dan Q_1. Jika terdapat observasi yang lebih dari 1,5 IQR, observasi tersebut disebut sebagai ***suspected outlier*** dan observasi yang lebih dari 3 IQR disebut sebagai ***outlier***.

Box Plot



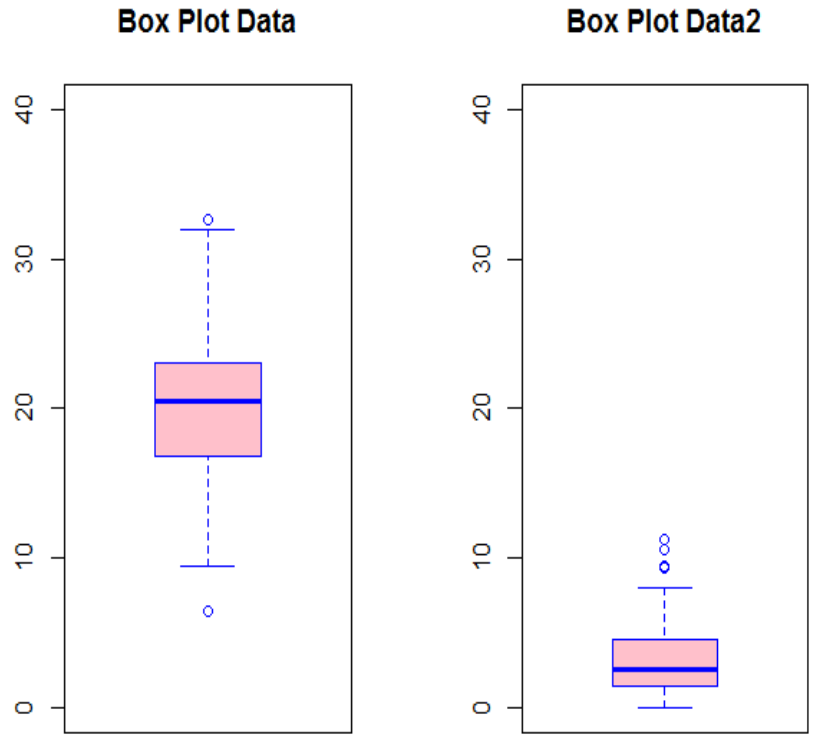
Komponen Box Plot





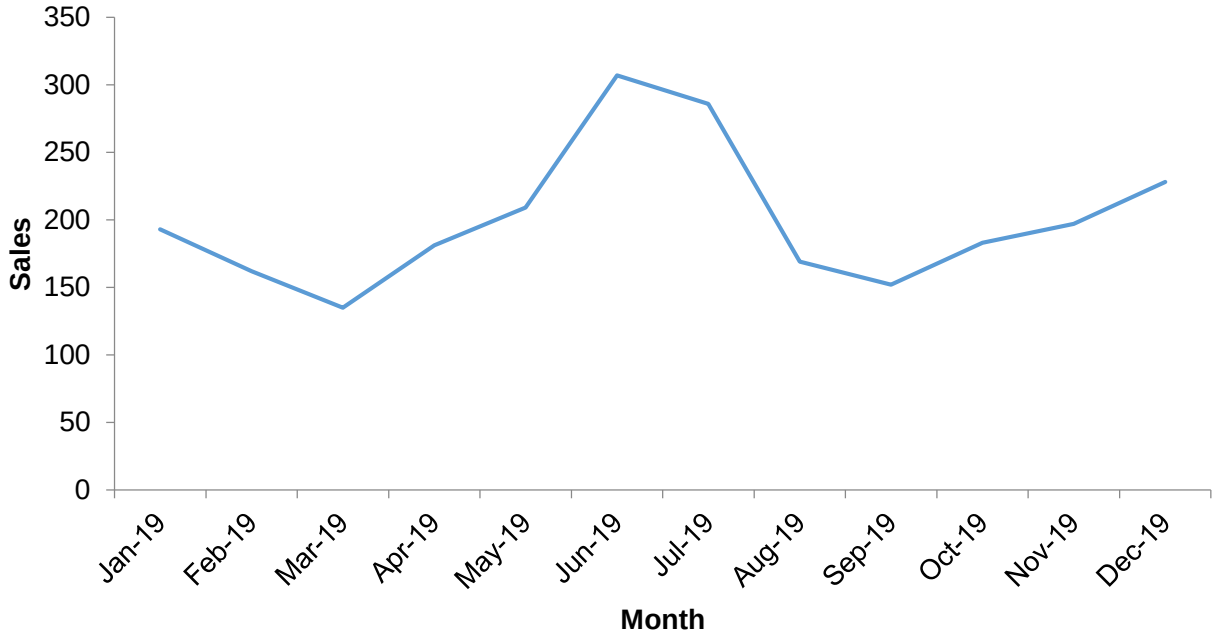
Aczel dan Sounderpandian (2008)

Box Plot



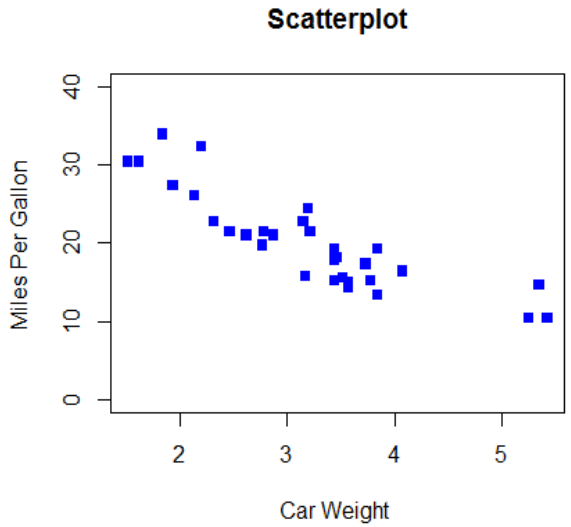

```
> summary(data)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  6.514  16.883  20.528  20.210  22.995  32.600
> summary(data2)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
 0.002601  1.555368  2.565741  3.236877  4.562966 11.300676
```

Time Plot



Grafik perubahan nilai variabel dari waktu ke waktu dapat disajikan dengan *time plot*. Dengan grafik ini, dapat diketahui pola variabel (data) secara periodik.

Scatter Plot



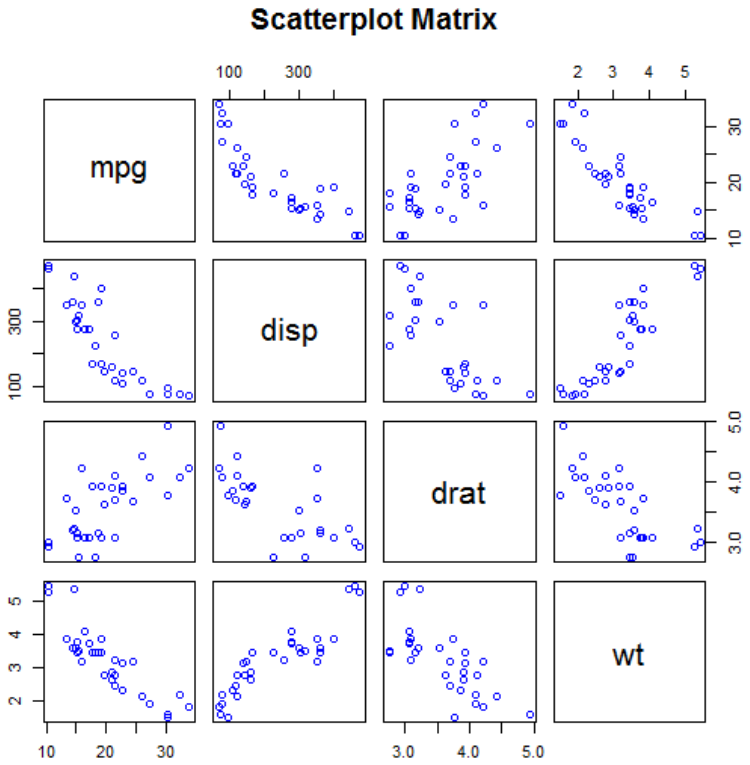

```
> cor(wt,mpg)
[1] -0.8676594
```

Data mtcars terdiri dari 11 variabel, 32 observasi

[, 1]	mpg	Miles/(US) gallon
[, 2]	cyl	Number of cylinders
[, 3]	disp	Displacement (cu.in.)
[, 4]	hp	Gross horsepower
[, 5]	drat	Rear axle ratio
[, 6]	wt	Weight (1000 lbs)
[, 7]	qsec	1/4 mile time
[, 8]	vs	V/S
[, 9]	am	Transmission (0 = automatic, 1 = manual)
[,10]	gear	Number of forward gears
[,11]	carb	Number of carburetors

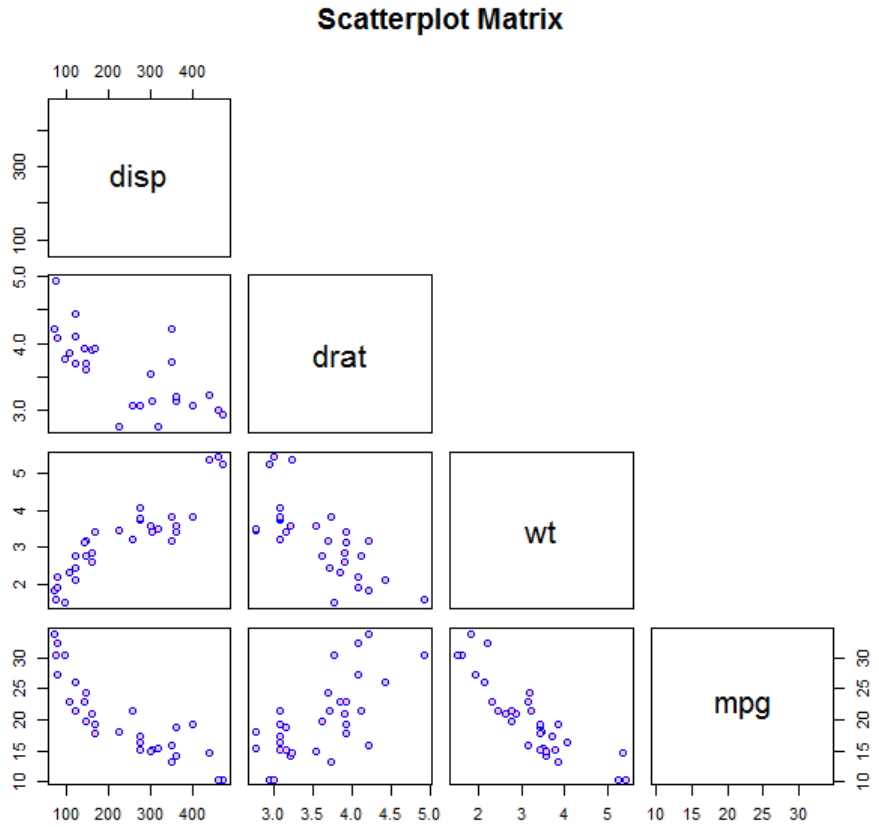
Scatter plot digunakan untuk mengidentifikasi dan menunjukkan hubungan di antara pasangan kumpulan data. Plot terdiri dari titik-titik yang tersebar mewakili pengamatan. Sebuah titik diberi tanda pada plot pada koordinat (x, y).

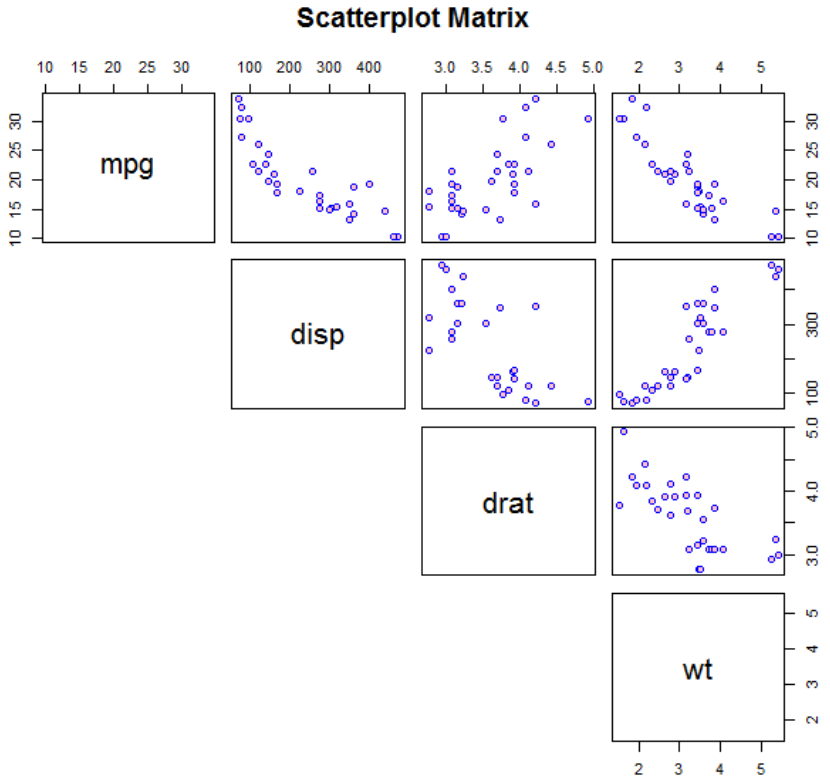
Membuat plot setiap pasang bisa menjadi kurang efisien, jadi akan lebih cepat dan mudah jika sekelompok *scatter plot* dibuat secara bersama.
Berikut *scatter plot matrix* dari empat variabel dan menghasilkan *scatter plot* untuk setiap pasangan variabel.



	mpg	disp	drat	wt
mpg	1.00000	-0.84755	0.68117	-0.86766
disp	-0.84755	1.00000	-0.71021	0.88798
drat	0.68117	-0.71021	1.00000	-0.71244
wt	-0.86766	0.88798	-0.71244	1.00000

Scatter Plot Matrix





Ringkasan

- Statistika deskriptif mencakup ringkasan data yang dapat disajikan dengan nilai numerik untuk lokasi atau pusat data dan jumlah variabilitas yang ada.
- Ukuran pemusatan data: mean, median, dan modus.
- Ukuran penyebaran data: *range*, *interquartile range*, variansi, dan deviasi standar.
- Penyajian data dengan menggunakan grafis untuk data kategorik dapat menggunakan diagram batang dan diagram lingkaran.
- Penyajian data untuk data terukur/ kuantitatif dapat menggunakan histogram, *stem-and-leaf, scatter plot, box plot,* dan *time plot*.

Daftar Pustaka

- Aczel, A. and Sounderpandian, J., 2008, *Complete Business Statistics*, 7th *Edition*, McGraw-Hill/Irwin, USA.
- Richard, A.J. and Bhattacharyya, G.K., 2010, *Statistics: Principles and Methods*, 6th *Edition*, John Wiley and Sons, USA.
- Mathur, Neha. *Exploratory Data Analysis.* Pace University.

Terima Kasih