Praktikum Eksplorasi dan Visualisasi Data
Pertemuan 2
BENTUK VISUAL DATA DAN RINGKASAN NUMERIK

---

1. **Bentuk Visual Data dan Ringkasan Numerik**
   - Membuat visualisasi data merupakan bentuk yang paling sederhana dari analisis data eksplorasi. Contoh bentuk visual data adalah daftar tally dan diagram batang dan daun (stem and leaf plot). Bentuk visual data dapat menunjukkan bentuk angkatan
   - Ringkasan numerik merupakan ringkasan dari data, harga-harga yang penting dari data, atau gambaran dari data yang berguna untuk estimasi nilai-nilai karakteristik data. Contoh ringkasan numerik adalah jumlah data, rata-rata, median, modus, range, variansi, dan standar deviasi.

2. **Daftar Tally**

   Daftar Tally adalah salah satu cara menyusun angkatan dengan menuliskan garis-garis pada interval-interval yang sudah dibuat sesuai nilai angkatan. Tujuannya untuk melihat pola, keteraturan, dan penyimpangan dari pola data. Berikut adalah kelebihan dan kelemahan dari daftar tally manual.

Kelebihan	Kelemahan
Dapat melihat bentuk distribusi data	Menghilangkan informasi dari data asli
Menghitung frekuensi dengan cepat	
Melihat ada tidaknya data ekstrim/*outlier*	

Langkah di R

Perlu menginstall package dplyr terlebih dahulu, dengan cara ketik `install.packages("dplyr")` kemudian run. Setelah itu panggil library dplyr dengan menuliskan syntax `library(dplyr)`.

```
>taly=mtcars %>% group_by(cyl)%>% tally()%>%
  mutate(Percent=n/sum(n)*100)%>%
  mutate(CumCnt=cumsum(n))%>%
  mutate(CumPct=cumsum(Percent))
>taly
```

```
# A tibble: 3 x 5
    cyl     n Percent CumCnt CumPct
  <dbl> <int>   <dbl>  <int>  <dbl>
1     4    11    34.4     11   34.4
2     6     7    21.9     18   56.2
3     8    14    43.8     32  100
```

Interpretasi

- Nilai minimum data …
- Nilai maksimum data …
- Modus …
- Median …
- Jangkauan …
- Jumlah data …
- Frekuensi Relatif per data (dilihat dari kolom Percent) …

3. **Stem and Leaf Plot**

Untuk mengatasi kelemahan yang ada pada daftar tally, digunakan diagram batang dan daun. Dalam Sofware R, Diagram batang dan daun memiliki dua komponen utama, yaitu sebagai berikut.

- Batang : Angka yang memiliki level lebih besar dari angka pada daun (biasanya puluhan/ratusan).
- Daun : Angka yang menunjukkan digit terakhir dari bilangan tersebut (biasanya satuan)

Langkah di R

Usage

```
stem(x, scale = 1, width = 80, atom = 1e-08)
```

Arguments

x = vector numerik
Scale = mengkontrol panjang plot
width = panjang plot yang diinginkan
atom= toleransi

Disarankan, argument yang perlu disesuaikan adalah scale.

Contoh, akan dibuat stem dan leaf plot dari data rivers. Pertama, perlu eksplorasi data dari data rivers.

```
#memanggil data rivers
rivers
#Apa itu data rivers
help("rivers")
#nilai minimum data rivers
```

```
min(rivers)
#nilai max data rivers
max(rivers)
```

```
> stem(rivers)

The decimal point is 2 digit(s) to the right of the |

 0 | 4
 2 | 011223334555566667778888899900001111223333344455555666688888999
 4 | 11122233344556679001233344567
 6 | 000112233578012234468
 8 | 045790018
10 | 04507
12 | 1471
14 | 56
16 | 7
18 | 9
20 |
22 | 25
24 | 3
26 |
28 |
30 |
32 |
34 |
36 | 1
```

Gambar 1. Stem and leaf

Gambar 1 kurang merepresentasikan data asli karena dari gambar 1. Tersebut nilai minimum data = 40, padahal aslinya 135. Jadi, perlu mencoba membuat stem and leaf plot yang sesuai.

```
> #scale yang besar dapat meningkatkan panjang batang
> stem(rivers, scale=2)
```

```
The decimal point is 2 digit(s) to the right of the |

 1 | 4
 2 | 011223334555566667778888899
 3 | 000011112233333444555556666888889999
 4 | 111222333445566779
 5 | 001233344567
 6 | 000112233578
 7 | 012234468
 8 | 04579
 9 | 0018
10 | 045
11 | 07
12 | 147
13 | 1
14 | 56
15 |
16 |
17 | 7
18 | 9
19 |
20 |
21 |
22 |
23 | 25
24 |
25 | 3
26 |
27 |
28 |
29 |
30 |
31 |
32 |
33 |
```

Gambar 2. Stem and leaf

Nb. untuk gambar 2. Sedikit kepotong

Gambar 2 lebih representatif data dengan nilai minimum dalam stem and leaf plot gambar 2. sebesar 140.

```
> #A scale between 0 and 1 will shorten the length of the stems.
> stem(rivers, scale = 0.5)

  The decimal point is 3 digit(s) to the right of the |

  0 | 12222222222333333333333333333333333333333333344444444444444444444
  0 | 55555555555555556666666666677777777778888999999
  1 | 0001122233
  1 | 5589
  2 | 33
  2 | 5
  3 |
  3 | 7
```

Gambar 3. Stem and leaf

Gambar 3. Juga representatif dengan nilai minimum dalam stem and leaf ini sebesar 100

Interpretasi

- Nilai minimum data …
- Nilai maksimum data …
- Jangkauan …
- Sebaran/Bentuk distribusi data ….
    a. Normal, simetris/mendekati simetris, data berbentuk lonceng
    b. Menceng kiri/menjurai kebawah, banyak nilai rendah yang menyebar
    c. Menceng kanan/mejurai keatas, banyak nilai tinggi yang menyebar

4. Ringkasan Numerik

Ringkasan numerik merupakan ringkasan dari data, harga-harga yang penting dari data, atau gambaran dari data yang berguna untuk estimasi nilai-nilai karakteristik data. Contoh ringkasan numerik adalah jumlah data, rata-rata, median, modus, range, variansi, dan standar deviasi.

A. Ukuran Pusat

Ukuran pusat menunjukkan letak dimana data berpusat. Pusat memberikan gambaran terhadap harga-harga suatu angkatan. Misal apabila pusatnya A, pastilah angka-angka angkatan tersebut kisar pada A, sebagian lebih dari A, dan sebagian lagi kurang dari A.

❖ Mean (Rata-rata)

Mean adalah ukuran pusat yang menjumlahkan semua datum kemudian dibagi banyaknya observasi yang sama banyak kemudian dibagi banyaknya observasi. Secara matematis ditulis sebagai berikut :

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Median

  Median adalah nilai yang membagi data menjadi 2 bagian dengan observasi yang sama banyak setelah data diurutkan dari kecil ke besar (sebaliknya). Secara matematis ditulis sebagai berikut :

$$Median\ (x) = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$$

- Kuartil

  Kuartil adalah harga yang membagi data menjadi 4 bagian dengan observasi sama banyak setelah data diurutkan dari kecil ke terbesar.

- Rata-rata tengah

  Rata-rata tengah adalah rata-rata dari observasi yang terletak di antara kuartil 1 dan kuartil 3 tidak termasuk kuartil 1 dan kuartil 3 tersebut.

- Modus

  Modus adalah harga yang muncul dengan frekuensi paling banyak. Suatu data bisa memiliki hanya satu modus, atau lebih dari 2 modus, bahkan tidak mempunyai modus atau dapat dikatakan semua observasi adalah modus.

B. Ukuran Sebaran

Ukuran sebaran menunjukkan sebaran atau penyimpangan data di sekitar pusat. Jika sebaran rendah berarti berarti data terletak di sekitar pusat dan jika sebaran tinggi berarti data terletak jauh dari data pusat yang artinya pusat kurang mewakili data dengan baik.

- Range (Jangkauan)

  Range adalah selisih bilangan terbesar dengan bilangan terkecil. Secara matematis ditulis sebagai berikut :

$$Range = x_A - x_B$$

dengan :

x_A : bilangan terbesar

x_B : bilangan terkecil

- Variansi atau ragam

  Dalam teori dan statistik dan statistika probabilitas, arti varians adalah pengukuran sebaran antar angka dalam suatu kumpulan data. Secara kasar variansi menyatakan besarnya ukuran data dilihat dari seberapa besar harga masing-masing observasi berbeda dari rata-ratanya. Secara matematis ditulis sebagai berikut :

$$Var\,(x) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

- Standar deviasi

  Standar deviasi adalah harga yang menunjukkan seberapa besar variasi (seperti penyebaran, penyebaran, penyebaran,) dari mean yang ada. Standar deviasi merupakan akar kuadrat dari variansi. Secara matematis ditulis sebagai berikut :

$$Sd = \sqrt{Var\,(x)}$$

Membuat Ringkasan Numerik dengan Software R

Sintaks :

```
# Panggil Library
library(openxlsx)

# Baca Data
mydata <- read.xlsx("D:\\Kuliah\\Praktikum Eksplorasi dan
Visualisasi Data\\Pertemuan 2\\Salary.xlsx")

# Cek sebagian data
head(mydata)
tail(mydata)
```

Output :

```
> head(mydata)
  Miami Chicago New_York Kansas San_Diego Seatle Alabama
Florida Hawaii Oklahoma
1  1669    5317     3969   6311      8880   9640    3298
5717   9855     9956
2  3942    7765     8928   8641      9142   3691    7645
7364   9300     4208
3  4052    7661     8910   2898      7285   9883    6244
2173   9180     4990
4  5815    9518     7802   4238      9242   9838    7887
5901   4019     9654
5  5988    8788     5567   6296      2805   2783    6415
9591   8892     8596
6  1735    9382     8899   6579      3430   8913    2709
6319   7598     4114
   Houston Philadelphia Vancouver Columbus Jacksonville
Quincy   Rialto    Tyler
1 5974.328    1407.9602  6529.535 6926.686     2034.812
5529.464 5009.362 2406.617
2 6006.458    1329.9584  6382.152 6999.062     1902.363
5345.986 5132.836 2365.163
3 5995.788    1207.1781  6606.464 7008.310     1997.279
5484.380 4983.634 2604.558
4 6044.749     969.3006  6610.144 6914.276     2134.064
5422.736 5020.320 2518.289
5 6005.720    1202.0131  6516.971 6999.237     1952.759
5539.946 5145.099 2359.494
6 6015.944    1336.4004  6611.654 7003.036     1956.609
5331.071 4926.597 2500.843
    Renton    Akron
1 5601.283 3867.284
2 5044.464 3360.151
3 5490.345 3664.315
4 5527.736 3344.308
5 5807.649 3600.703
6 5531.366 3441.459

> tail(mydata)
     Miami Chicago New_York Kansas San_Diego Seatle Alabama
Florida Hawaii Oklahoma
```

```
4495  4497    6025     7901   1622      8249   6215    8702
3633   3984     7872
4496  1076    5139     5711   2533      9384   4705    8029
6584   3576     9915
4497  4544    4801     6721   3423      5007   5448    4624
5797   5617     9301
4498  3540    6723     5843   7320      3179   9197    1970
1959   4246     7905
4499  1414    2035     5966   7464      3244   6207    4746
8485   9744     4729
4500  4505    5868     9326   2138      8335   7290    2181
9897   7701     8691
      Houston Philadelphia Vancouver Columbus Jacksonville
Quincy   Rialto    Tyler
4495 6085.390     1264.745  6624.423 6841.473     1935.848
5249.570 5073.881 2541.042
4496 5938.194     1227.942  6563.881 6942.651     1970.636
5310.978 4986.037 2617.913
4497 5965.747     1222.562  6466.868 7022.289     2015.304
5476.406 4985.743 2364.798
4498 6003.057     1335.931  6511.885 7013.256     1996.260
5406.180 4825.094 2375.920
4499 5956.670     1280.753  6523.053 6873.514     1976.089
5428.695 4896.370 2554.379
4500 6037.146     1337.640  6347.547 6963.469     1965.910
5391.792 4995.259 2559.574
       Renton    Akron
4495 5159.303 3452.357
4496 5388.325 3417.677
4497 5487.839 3550.556
4498 5538.823 3877.186
4499 5491.020 3410.037
4500 5239.734 3568.908

############################    Ukuran Pusat
##################################
# Mencari Mean
rataMiami <- mean(mydata$Miami)
rataMiami
```

Output :

```
> rataMiami
[1] 3504.3528888888891

# Mencari Median
medianMiami <- median(mydata$Miami)
medianMiami
```

Output :

```
> medianMiami
[1] 3495.5

# Mencari Kuartil
# Kuartil 1
Q1_Miami = quantile(mydata$Miami, prob=0.25)
Q1_Miami
```

Output :

```
> Q1_Miami
    25%
2272.75

# Kuartil 2
Q2_Miami = quantile(mydata$Miami, prob=0.50)
Q2_Miami
```

Output :

```
> Q2_Miami
   50%
3495.5

# Kuartil 3
Q3_Miami = quantile(mydata$Miami, prob=0.75)
Q3_Miami
```

Output :

```
> Q3_Miami
    75%
4746.25

# Modus
Modus <-function(x){
```

```
  u <- unique(x)
  tab <- tabulate(match(x,u))
  u[tab ==max(tab)]
}

Modus(mydata$Miami)
```

Output :

```
> Modus(mydata$Miami)
  [1] 1669 3942 4052 5815 5988 1735 2697 5434 4166 5105 4503
 [12] 4529 1784 2497 3998 2558 5925 3408 2548 2186 2818 5419
 [23] 1976 2387 2808 5863 3947 5972 1770 4628 3518 1529 3127
 [34] 4026 3215 1934 3993 4149 1375 5367 2989 3798 2636 5632
 [45] 1268 2596 2409 2104 2090 4848 5922 5761 2069 1462 5746
 [56] 2106 2617 1750 1856 1238 1326 2063 4332 3525 3080 4346
 [67] 5275 4801 2018 2888 5908 5152 1840 2961 4911 4284 4997
 [78] 4728 5812 3421 2623 1289 3848 4150 1388 5660 1176 3619
 [89] 1005 1403 4038 3894 1810 2611 3180 2405 3747 5760 2127
[100] 1589 4805 2435 4347 2516 5380 4130 5463 3431 2984 3244
[111] 4098 4287 3168 4673 4681 2362 1211 4920 4639 4761 3882
[122] 3489 1028 1153 1936 5946 5735 4162 1963 4029 1121 2663
[133] 3444 1079 4873 2809 1621 5033 2374 3141 4482 2954 1160
[144] 2211 2878 4976 2015 4879 4770 1977 5222 2217 1230 4593
[155] 3745 1466 3202 5203 2074 3125 1415 4384 1032 2072 5952
[166] 3115 1706 5109 3331 5785 4269 2438 1046 4169 5528 5605
[177] 4160 3475 3937 4869 1642 5051 2930 2308 5667 2202 3716
[188] 5845 4461 4577 1129 3396 2267 3502 5664 2142 2578 5241
[199] 1389 4843 2428 3414 4746 1181 4887 2275 5999 3351 5733
[210] 3289 4254 4560 5217 5702 4941 3876 5398 2396 5443 5251
[221] 4592 4237 5644 5803 1132 4798 3372 4687 5649 1518 5255
```

```
########################      Ukuran Sebaran
############################
# Range / Jangkauan
Jangkauan <- function(x){
  jangkauan = max(x)-min(x)
  return(jangkauan)
}
Jangkauan(mydata$Miami)
```

Output :

```
> Jangkauan(mydata$Miami)
[1] 4999
```

```
# Variansi
varian_Miami <- var(mydata$Miami)
```

```
options(digits = 20)
varian_Miami
```

Output :

```
> varian_Miami
[1] 2067827.339767158
```

```
# Standar Deviasi
sd_Miami <- sd(mydata$Miami)
sd_Miami
```

Output :

```
> sd_Miami
[1] 1437.9942071396388
```

```
country <- c("Miami")
ringkasan <- data.frame(country,
                        meanSallary = mean(mydata$Miami),
                        MedianSallary=median(mydata$Miami),
                        Q1 =
quantile(mydata$Miami,prob=0.25),

Q3=quantile(mydata$Miami,prob=0.75),
                        variansi = var(mydata$Miami),
                        Std_dev=sd(mydata$Miami),
                        IQR=IQR(mydata$Miami),
                        min=min(mydata$Miami),
                        max=max(mydata$Miami),
                        row.names = NULL)
Ringkasan
```

Output :

```
> ringkasan
  country meanSallary MedianSallary      Q1      Q3 variansi  Std_dev    IQR  min  max
1   Miami    3504.353        3495.5 2272.75 4746.25  2067827 1437.994 2473.5 1000 5999
```

##########

Latihan :

1. Salah satu ukuran pemusatan data adalah trirata. Trirata Disebut juga rata-rata berbobot karena menunjukkan bobot dari kuartil bawah dan kuartil atas dan median. Secara matematis ditulis sebagai berikut :

$$Trirata = \frac{Q1 + Q3 + 2\,Median}{4}$$

   Buatlah sintaks yang dapat menghitung trirata. Jalankan sintaks tersebut lalu interpretasikan hasilnya !

2. Syafa sedang melakukan evaluasi untuk melihat Salary yang didapatkan penduduk di 20 kota di negara Amerika selama 1 bulan terakhir. Oleh karena itu, ia mengumpulkan sebanyak 4500 sample dari setiap kotanya. Sebelum melakukan analisis ia ingin mengetahui ringkasan numerik dari pendapatan tiap kotanya. Ringkasan numerik yang dibutuhkan adalah rata-rata, median, Kuartil 1, Kuartil 2, Variansi, Standar Deviasi, IQR, Min, Max. bantulah Syafa untuk mengumpulkan ringkasan numerik dari pendapatan 20 kota tersebut.(data tersedia di *Salary.xlsx*) (bebas menggunakan library apapun selama hasil yang diharapkan terpenuhi)

   Output yang diharapkan :

```
> ringkas
# A tibble: 20 x 10
   Country   meanSallary MedianSallary    Q1    Q3 variansi Std_dev    IQR   min   max
   <chr>           <dbl>         <dbl> <dbl> <dbl>    <dbl>   <dbl>  <dbl> <dbl> <dbl>
 1 Akron           3648.         3646. 3518. 3779.   37498.   194.   261.  3010. 4236.
 2 Alabama         5935.         5952. 3828. 8016. 5729277.  2394.  4188.  1751  9999
 3 Chicago         6006.         5996. 4015. 8003. 5342661.  2311.  3988.  2000  9999
 4 Columbus        6954.         6955. 6917. 6992.    3043.    55.2   74.1 6771. 7153.
 5 Florida         5962.         5972  3917. 7945. 5387594.  2321.  4028.  1952  9999
 6 Hawaii          6241.         6248  4365. 8118. 4775113.  2185.  3754.  2451  9999
 7 Houston         6000.         6000. 5965. 6034.    2585.    50.8   68.4 5824. 6167.
 8 Jacksonv~       2004.         2004. 1970. 2038.    2437.    49.4   67.2 1831. 2164.
 9 Kansas          5721.         5706. 3612. 7840. 6014230.  2452.  4228.  1500  9996
10 Miami           3504.         3496. 2273. 4746. 2067827.  1438.  2474.  1000  5999
11 New_York        6753.         6766. 5110. 8379. 3530671.  1879.  3270.  3501  9999
12 Oklahoma        6532.         6514  4830. 8234. 3909055.  1977.  3405.  3126  9999
13 Philadel~       1246.         1245. 1162. 1331.   15337.   124.   168.   732. 1734.
14 Quincy          5422.         5421. 5362. 5483.    8567.    92.6  121.  5051. 5740.
15 Renton          5445.         5442. 5334. 5557.   27277.   165.   223.  4882. 6166.
16 Rialto          4992.         4993. 4938. 5046.    6349.    79.7  108.  4744. 5281.
17 San_Diego       6309.         6338  4411. 8208. 4734549.  2176.  3798.  2500  9999
18 Seatle          6170.         6184. 4225. 8102. 5029893.  2243.  3878.  2251  9999
19 Tyler           2457.         2455. 2396. 2519.    8298.    91.1  123.  2125. 2796.
20 Vancouver       6541.         6541. 6490. 6591.    5779.    76.0  101.  6244. 6843.
```

3. Jumlah kunjungan wisatawan nusantara adalah jumlah perjalanan kurang dari 6 bulan yang dilakukan oleh penduduk dalam wilayah Indonesia dengan tujuan bukan untuk bekerja atau sekolah. Indikator ini digunakan untuk mengetahui preferensi wisatawan domestik terhadap objek wisata domestik sebagai bentuk kontribusi dalam mendukung kemajuan sektor pariwisata Indonesia. Buatlah steam and leaf plot untuk mengetahui :
   a. Jumlah minimum dan maksimum perjalanan wisatawan
   b. Tentukan jangkauan dari data (selisih data maksimum dan minimum) kemudian dugalah apakah jumlah perjalanan antar provinsi dikatakan heterogen? Jelaskan alasanmu.

c. Bagaimana sebaran data dari jumlah kunjungan wisatawan di Indonesia tahun 2019? Apakah mendukung kesimpulan poin b? Jelaskan alasanmu

(*Digunakan data P-2.xlsx)*

4. Buatlah daftar tally pada data CO2 kolom conc. Conc adalah konsentrasi karbondioksida (mL/L). Kemudian jawab pertanyaan berikut,

- Nilai minimum data …
- Nilai maksimum data …
- Modus …
- Median …
- Jangkauan …
- Jumlah data …
- Frekuensi Relatif perdata (dilihat dari kolom Percent) …

NB. data CO2 adalah salah satu data yang disediakan R. untuk memanggil nya tinggal ketik CO2 atau juga bisa menyimpan CO2 ke dalam suatu objek.