

We will show how to visually summarize data and compute Descriptive Statistics in R via two examples.

1. In a study of job satisfaction, a series of tests were administered to 60 subjects and the following data was obtained (higher score represents higher job satisfaction) (data file Score.csv)

Score					
76	86	66	91	53	79
85	82	80	84	66	66
78	69	77	70	75	72
70	62	56	77	75	69
78	85	69	75	83	58
75	69	70	71	76	53
73	84	72	72	80	71
70	82	55	84	51	68
72	101	86	93	87	65
78	99	90	77	79	78

Compute descriptive statistics and graph the dot diagram, stem-and-leaf diagram, box plot and histogram for the data.

2. The following table shows economic data for 36 firms in Japan (data file Japan.csv)

Short Name	Market_Cap_Yen	Market Cap US \$	Book Equity	Revenues	Net Income	PE
CENTURY21 REAL	14496	\$123.72	1924.79	2452.09	463.24	31.29
BALS CORP	43231.88	\$369.01	3239	18994	464	93.17
MAINICHI COMNET	10723.2	\$91.69	2675	6449	464	23.11
MUTUAL	5990.63	\$51.42	6915	10100	464	12.91
NISSIN SHOJI CO	9728	\$83.18	17129	70529	464	20.97
TOW CO LTD	8030.752	\$68.94	3782	10705	465	17.27
CLEX CO LTD	10105.62	\$86.41	4455	9486	466	21.69
FENWAL						
CONTROLS	6334.975	\$54.37	4357	14930	466	13.59
FUJII SANGYO	8008	\$68.62	11392	46927	466	17.18
MOLITEC STEEL	10760.17	\$92.07	10221	18923	466	23.09
TAKACHIHO						
KOHEKI	19198.13	\$164.18	12108	22318	466	41.2
GMO HOSTING	67081.95	\$571.52	1247	3038	467	143.64
MIURA PRINTING	11011.9	\$94.15	9504	22136	467	23.58
TSUKEN CORP.	9278.147	\$79.65	14182	44581	468	19.83
KING CO LTD	14615.24	\$124.94	16822	22369	469	31.16
NIPPON RESIBON	5926.8	\$50.88	5760	15012	471	12.58
AVAL DATA CORP	10725.84	\$91.76	7682.24	7775.44	471.64	22.74
SOMAR CORP	15885.06	\$135.62	14781	36540	472	33.65
MISUMI CO LTD	8296.68399	\$71.22	10796	48836	473	17.54

Data Summarization in R

NICHIWA SANGYO HARADA INDUS CO	11144.49	\$95.32	15779	42124	473	23.56
IMV CORP MARUFUJI SHEET P	11388.83	\$97.43	6376	20771	474	24.03
DAIDO SIGNAL CO KANESHITA CONSTR	8478.508	\$72.75	2006	5088	476	17.81
KAKAKU.COM INC MR MAX CORP NAGOYA ELECTRIC	16108.48	\$137.51	24738	32758	476	33.84
ZOA CORP CHUO KAGAKU CO L	6954.948	\$59.69	8214	17632	477	14.58
PALTEK CORP RIX CORP ENSHU TOYAMA BANK LTD	17605.8	\$150.57	23205	19760	477	36.91
WAO CORP OIE SANGYO CO	71655.67	\$610.72	2223.01	2138.87	478.05	149.89
	28044.59	\$239.30	28391	86133	480	58.43
	6755.84	\$57.98	15568	15646	480	14.07
	6889.5	\$59.12	1629	17589	481	14.32
	28088.1	\$239.70	30825	82965	483	58.15
	6150.098	\$52.78	8711	19355	483	12.73
	8726.4	\$74.90	4978.53	27043.88	483.76	18.04
	22391.1	\$191.39	6409	35665	484	46.26
	17417.16	\$148.64	22493	7486	484	35.99
	6733.8	\$57.71	3577	15033	484.0665	NA
	11568.75	\$98.92	9021	52180	485	23.85

- a) Graph Book_equity vs. Revenue.
- b) Graph all variables against each other.
- c) Calculate the correlation coefficient matrix.
- d) Calculate descriptive statistics for PE by the following categories:
 - Market_Cap_US\$ 50 to 100
 - 100 to 200
 - 200 to 300
 - > 300
- e) Plot the histogram of Book_Equity.
- f)) Construct a crosstabulation for PE and Market_Cap_US using the following ranges:

PE range: 10 – 50, 50 – 100, > 100

Market_Cap_US range: 50 – 100 = 1, 100 – 200 = 2, 200 – 300 = 3, > 300 = 4.

1. Descriptives and Dot Diagram, Stem-and-leaf Diagram, and Histogram for SCORE data of Example 1

Read the Score data in R:

```
d1<-read.csv("g:/Stats24x7/R/Score.csv",header=TRUE)
```

The object d1 is a dataframe in R, with one variable 'Score', which can be accessed as d1\$Score.

In R, type `attach(d1)`

The variable Score can now be accessed as Score.

In R, load the library lattice:

```
library(epicalc)
```

(a) Compute the descriptive statistics.

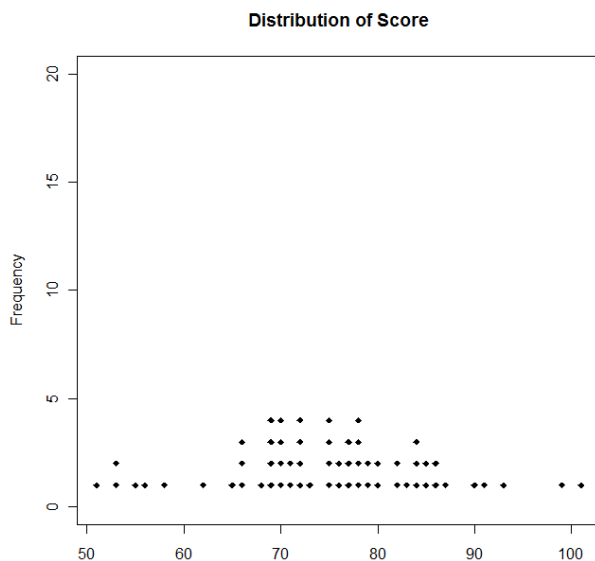
```
summary(Score)
```

```
Min. 1st Qu. Median Mean 3rd Qu. Max.
```

```
51.00 69.00 75.00 74.88 82.00 101.00
```

(b) Draw the dot plot:

```
dotplot(Score)
```



(c) Draw the stem-and-leaf diagram:

```
library(graphics)
```

```
stem(Score)
```

Output is shown below.

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

```
5 | 133568
```

```
6 | 2566689999
```

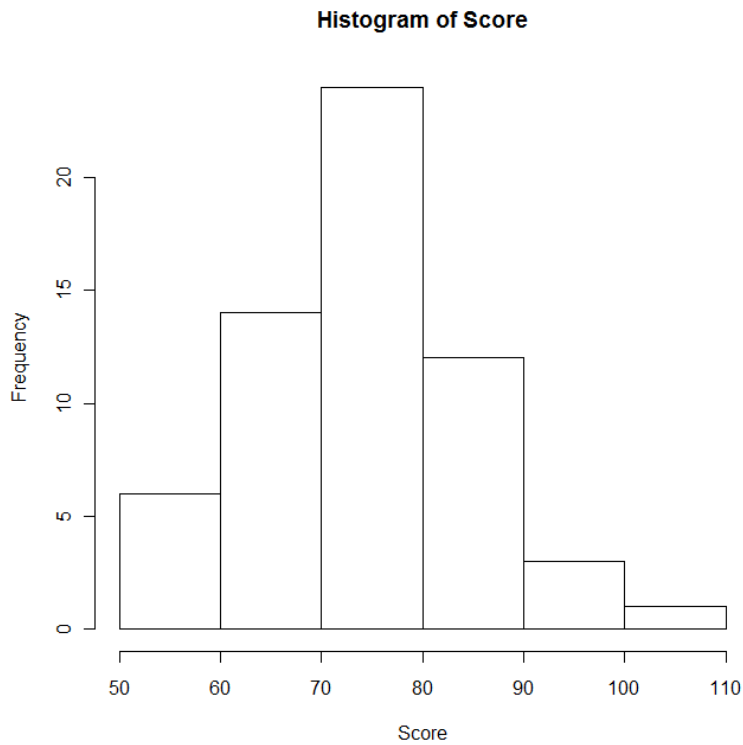
```
7 | 00001122223555566777888899
```

```
8 | 0022344455667
```

```
9 | 0139
```

```
10 | 1
```

(d) Draw the histogram:



Visual Summarization of Data of Example 2

Open the data file Japan.xlsx in MINITAB.

```
d2<-read.csv("g:/Stats24x7/R/Japan.csv",header=TRUE)
```

```
attach(d2)
```

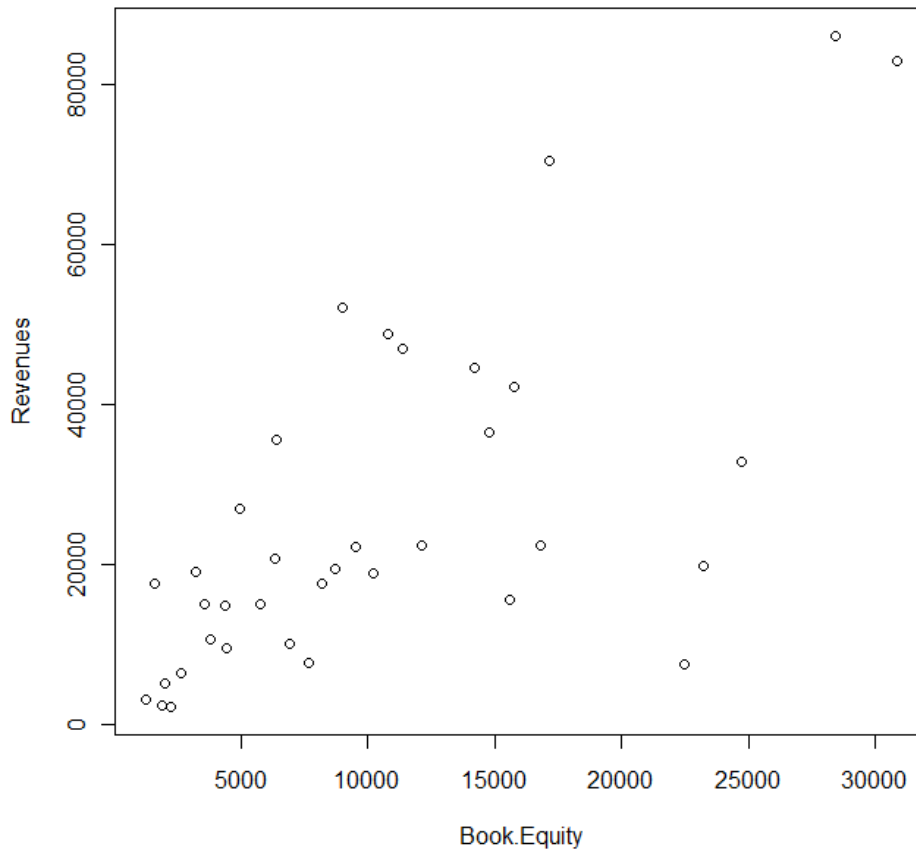
To see the variable names in the file d2 in R, type

```
names(d2)
```

```
[1] "Short.Name" "Market_Cap_Yen" "Market.Cap.US."  
"Book.Equity"  
[5] "Revenues" "Net.Income" "PE"
```

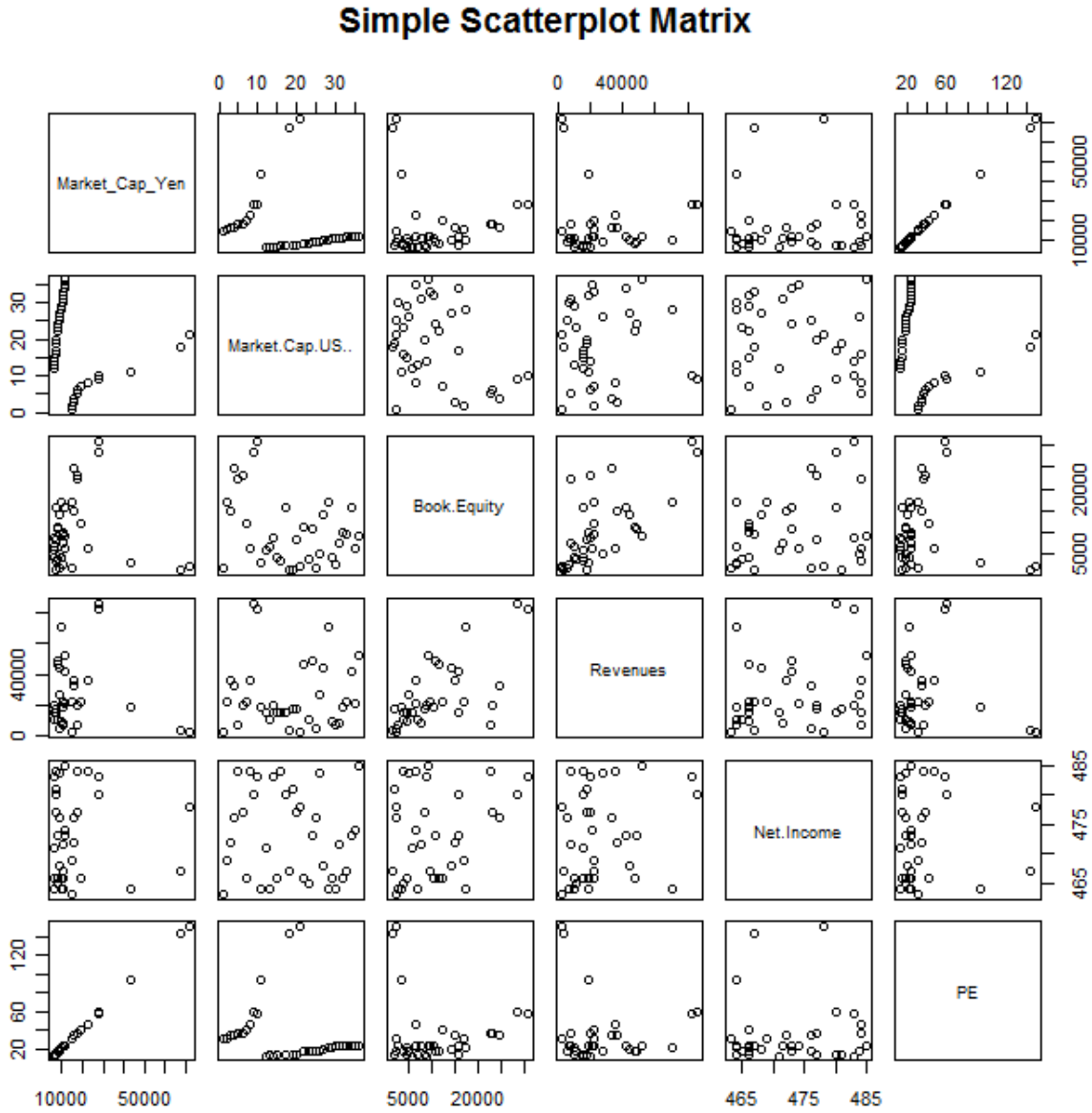
2(a): Scatter plot of Revenues vs. Book.Equity

```
plot(Revenues~Book.Equity)
```



2(b) Draw the Matrix Plot R:

```
pairs(~Market_Cap_Yen + Market.Cap.US.. + Book.Equity + Revenues +  
Net.Income + PE, main="Simple Scatterplot Matrix")
```



2(c) Correlation Matrix

The variable PE has a missing value (NA in R), which needs to be omitted in computing the correlation matrix.

```
cor(d2a, use="complete.obs")
```

```

                Market_Cap_Yen Market.Cap.US.. Book.Equity  Revenues
Market_Cap_Yen    1.00000000    0.99999952 -0.06109699 -0.07320537
Market.Cap.US..   0.99999952    1.00000000 -0.06111329 -0.07308374
Book.Equity       -0.06109699   -0.06111329  1.00000000  0.67265345
Revenues          -0.07320537   -0.07308374  0.67265345  1.00000000
Net.Income        0.03865069    0.03875115  0.32726516  0.25467834
PE                0.99977992    0.99977857 -0.06929141 -0.07978031
                Net.Income      PE
Market_Cap_Yen  0.03865069  0.99977992
Market.Cap.US.. 0.03875115  0.99977857
Book.Equity     0.32726516 -0.06929141
Revenues        0.25467834 -0.07978031
Net.Income      1.00000000  0.02244606
PE              0.02244606  1.00000000

```

2d. Descriptives by Grouping Variable

To calculate descriptive statistics for PE by the following categories:

```

Market_Cap_US$ 50 to 100
                100 to 200
                200 to 300
                > 300

```

First create a categorical variable market_cap_us_cat as follows:

```

d2$market_cap_us_cat[Market.Cap.US.>50 & Market.Cap.US.<= 100] <- 1
d2$market_cap_us_cat[Market.Cap.US.>100 & Market.Cap.US.<= 200] <- 2
d2$market_cap_us_cat[Market.Cap.US.>200 & Market.Cap.US.<= 300] <- 3
d2$[Market.Cap.US.>300] <- 4

```

```
attach(d2)
```

d2\$newvar
creates a
variable newvar
in the dataframe
d2 in R.

You must attach
d2 again since
d2 has changed.

Data Summarization in R

Use the R-package psych for summary by a group variable. If the package psych has not been installed earlier, then first install the package psyche, then load it by typing in R:

```
library(psych)
```

Then in R, type

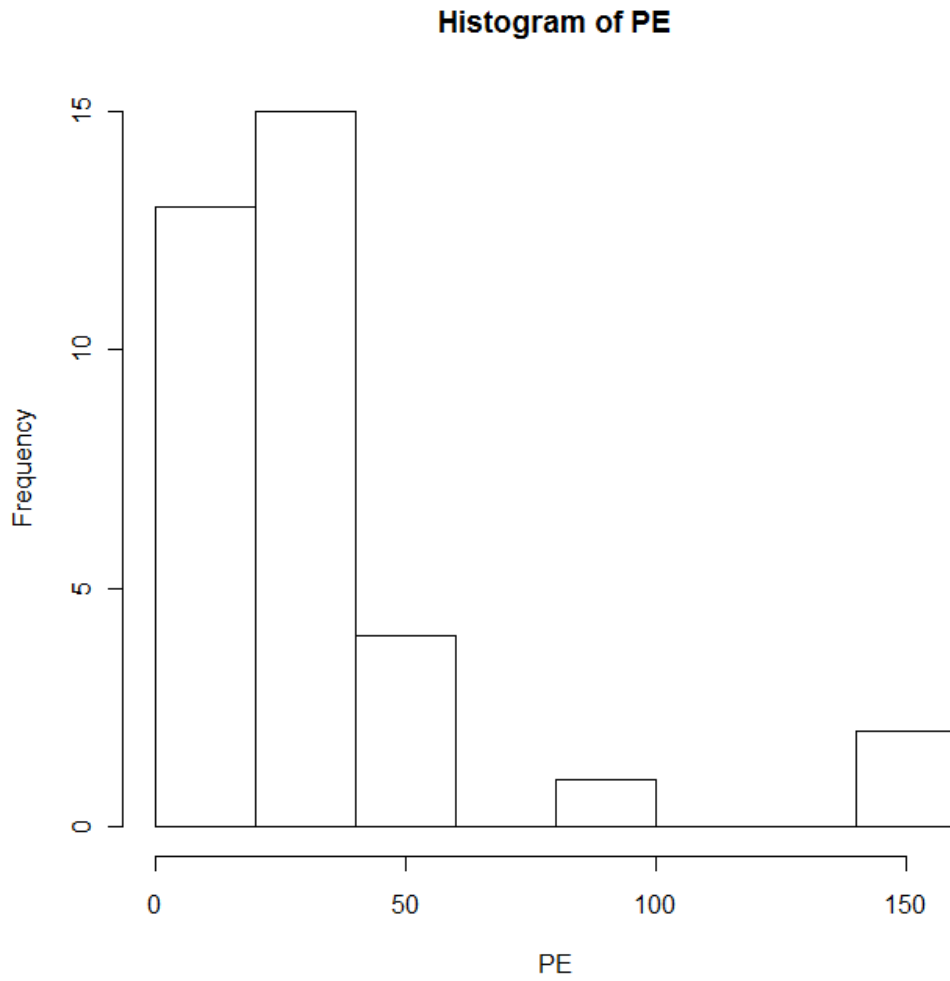
```
describe.by(PE, market_cap_us_cat)
```

The output from R is shown below.

```
group: 1
  var n mean  sd median trimmed  mad   min   max range  skew kurtosis  se
1   1 22 18.59 4.2  17.92   18.66 6.07 12.58 24.03 11.45 -0.07   -1.64 0.89
-----
group: 2
  var n mean  sd median trimmed  mad   min   max range  skew kurtosis  se
1   1  8 36.29 5.19  34.92   36.29 4.17 31.16 46.26 15.1  0.73   -0.95 1.83
-----
group: 3
  var n mean  sd median trimmed  mad   min   max range  skew kurtosis  se
1   1  2 58.29 0.2  58.29   58.29 0.21 58.15 58.43  0.28   0   -2.75 0.14
-----
group: 4
  var n mean  sd median trimmed  mad   min   max range  skew kurtosis  se
1   1  3 128.9 31.1 143.64   128.9 9.27 93.17 149.89 56.72 -0.37   -2.33 17.96
,
```

2e. Histogram of PE is obtained by typing in R:

```
hist(PE)
```



Data Summarization in R

2f)

To construct a crosstabulation for PE and Market_Cap_US using the following ranges:

PE range: 10 – 50, 50 – 100, > 100

Market_Cap_US range: 50 – 100 = 1, 100 – 200 = 2, 200 – 300 = 3, > 300 = 4,

create PECode using the above 3 ranges of PE as shown below:

```
d2$PE_cat[PE>10 & PE<= 50] <- 1
d2$PE_cat[PE>50 & PE<= 100] <- 2
d2$PE_cat[PE>100] <- 3
```

Once PECode has been created, type following in R:

```
ytable <- xtabs(~market_cap_us_cat+PE_cat, data=d2)
```

```
      PE_cat
market_cap_us_cat 1 2 3
1 22 0 0
2 8 0 0
3 0 2 0
4 0 1 2
```