STAT3622 Data Visualization (Lecture 2)

Exploratory Data Analysis

Dr. Aijun Zhang
The University of Hong Kong

4 February 2020

What's covered in this lecture?

I. Exploratory Data Analysis

- John Tukey
- Exploratory Data Analysis

II. Simple Base Graphics

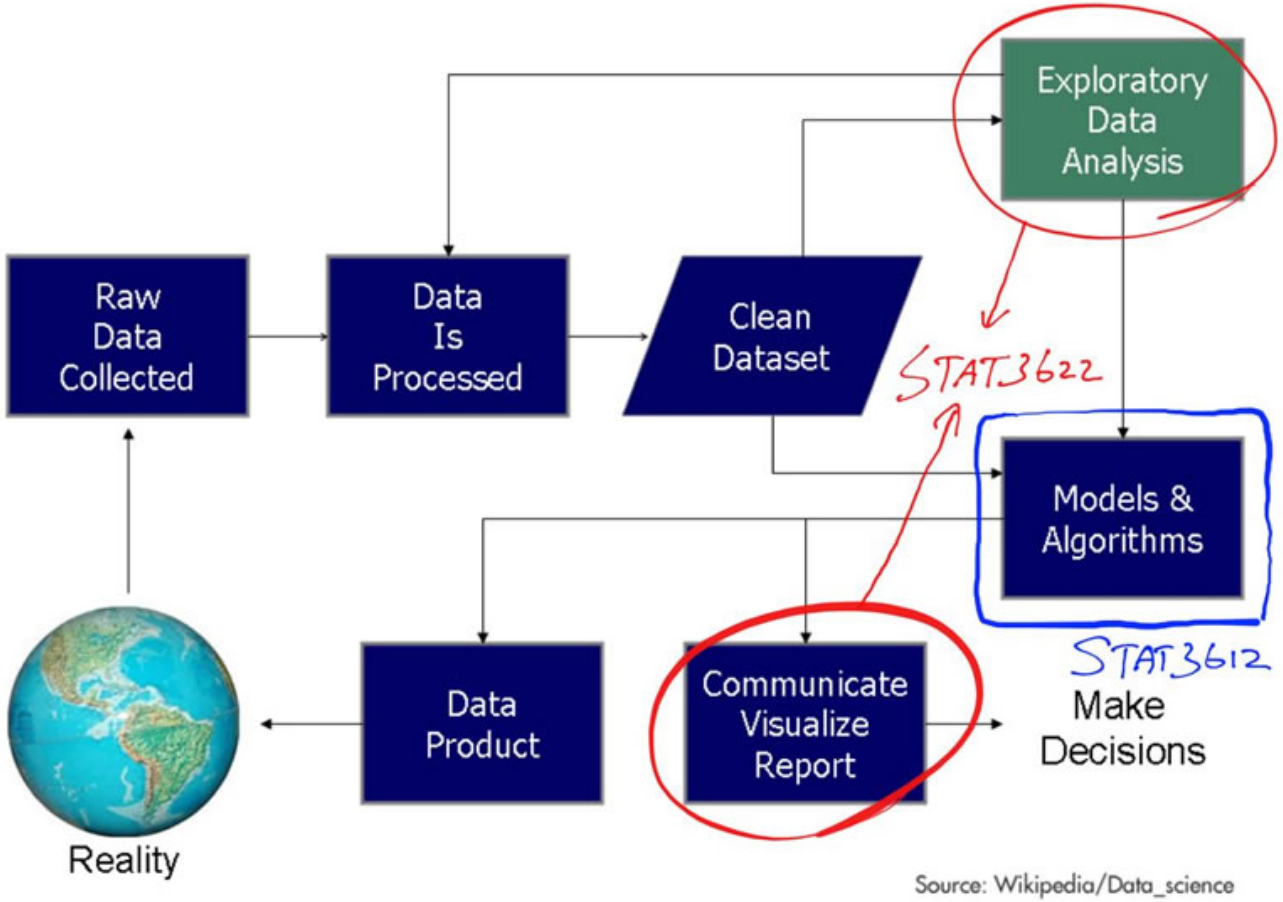
- Iris Dataset
- Basic R Plots

II. Using R:Lattice Package

- Conditioning and Grouping
- Cloud and Level Plots

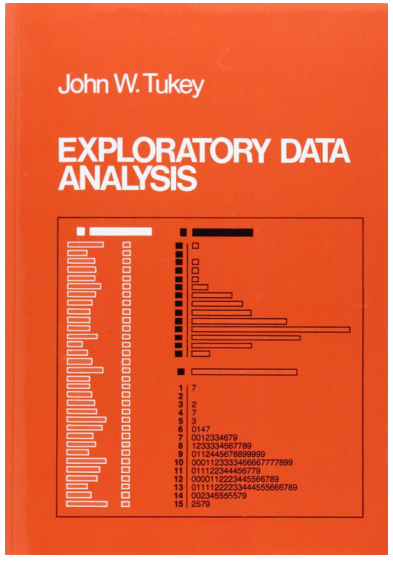
I. Exploratory Data Analysis

Review: Data Science Workflow



Roles of Data Visualization

- Role 1: Exploratory data analysis (pre stage);
- Role 2: Visual presentation of results (after stage).
- John W. Tukey (1977; Exploratory Data Analysis): "The greatest value of a picture is when it forces us to notice what we never expected to see."



John Tukey (1915-2000)

- Proposed “Exploratory Data Analysis”
- Coined terms: Boxplot, Stem-and-Leaf plot, ANOVA (Analysis of Variance)
- Coined terms “Bit” and “Software”
- Co-Developed Fast Fourier Transform algorithm, Projection Pursuit, Jackknife estimation
- Famous quote: “The best thing about being a statistician is that you get to play in everyone's backyard.”

- https://en.wikipedia.org/wiki/John_Tukey

John Tukey: The Future of Data Analysis (1962)

Excerpt from Donoho (2015) "50 years of Data Science"

3 *The Future of Data Analysis*, 1962

This paper was prepared for the John Tukey centennial. More than 50 years ago, John prophecied that something like today's Data Science moment would be coming. In "The Future of Data Analysis" [42], John deeply shocked his readers (academic statisticians) with the following introductory paragraphs:[16]

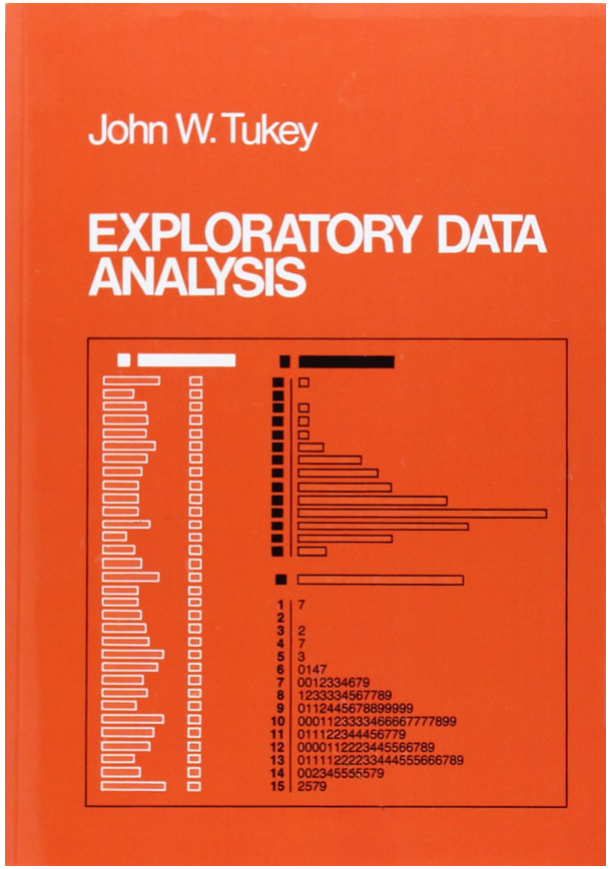
> *For a long time I have thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt. ... All in all I have come to feel that my central interest is in data analysis, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data*

This paper was published in 1962 in "The Annals of Mathematical Statistics", the central venue for mathematically-advanced statistical research of the day. Other articles appearing in that journal

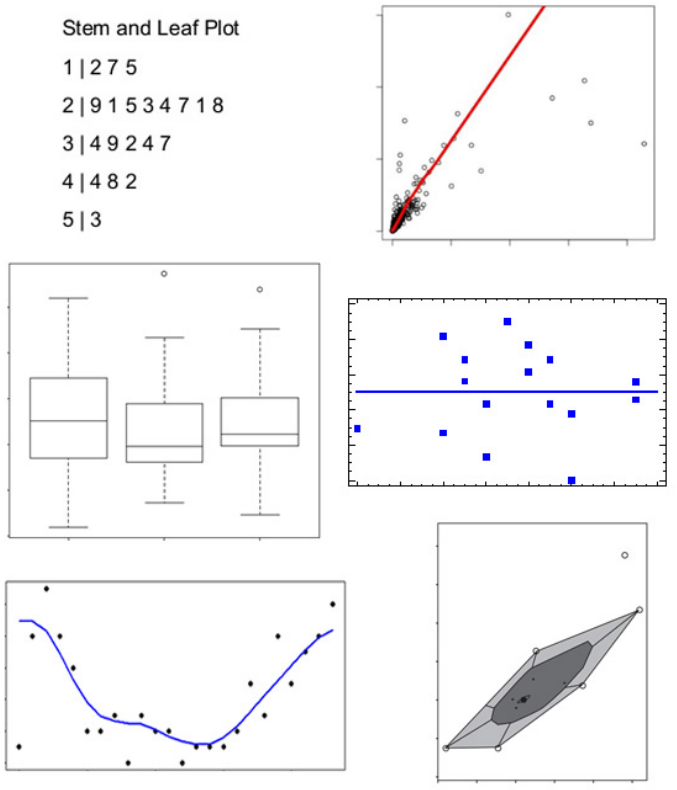
[16] One questions why the journal even allowed this to be published! ...

- Reference: Donoho, David (2017). "50 Years of Data Science" at *JCGS*, **26**(4), 745-766.

John Tukey: Exploratory Data Analysis (1977)



- Five-number summary
- Stem-and-Leaf plot
- Scatter plot
- Box-plot, Outliers
- Residual plot
- Smoother
- Bag plot



Example: Anscombe Dataset

```
## Anscombe Dataset:
```

x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Source: Anscombe, F. J. (1973). Graphs in statistical analysis. *American Statistician*, **27**, 17-21.

Example: Anscombe Dataset (Descriptive)

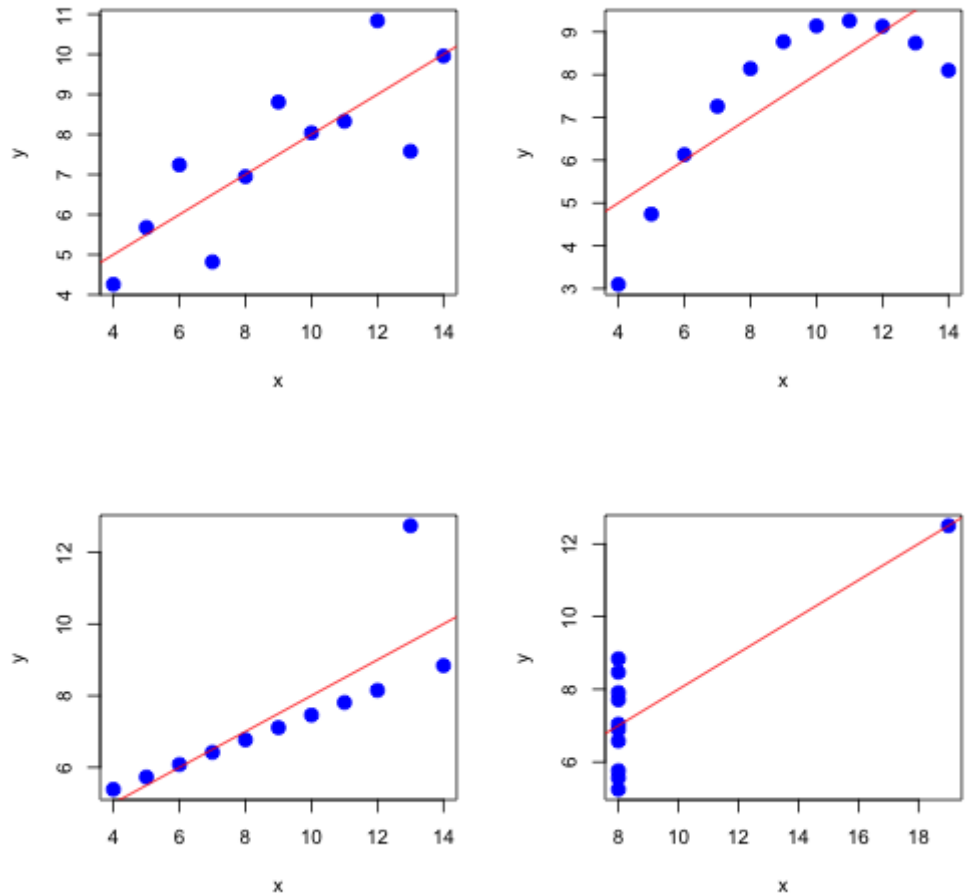
```
## Mean and standard deviation:
```

	x1	y1	x2	y2	x3	y3	x4	y4
mean	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
sd	3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03

```
## x-y correlation:
```

rho1	rho2	rho3	rho4
0.82	0.82	0.82	0.82

Example: Anscombe Dataset (Graphic)



Exploratory Data Analysis

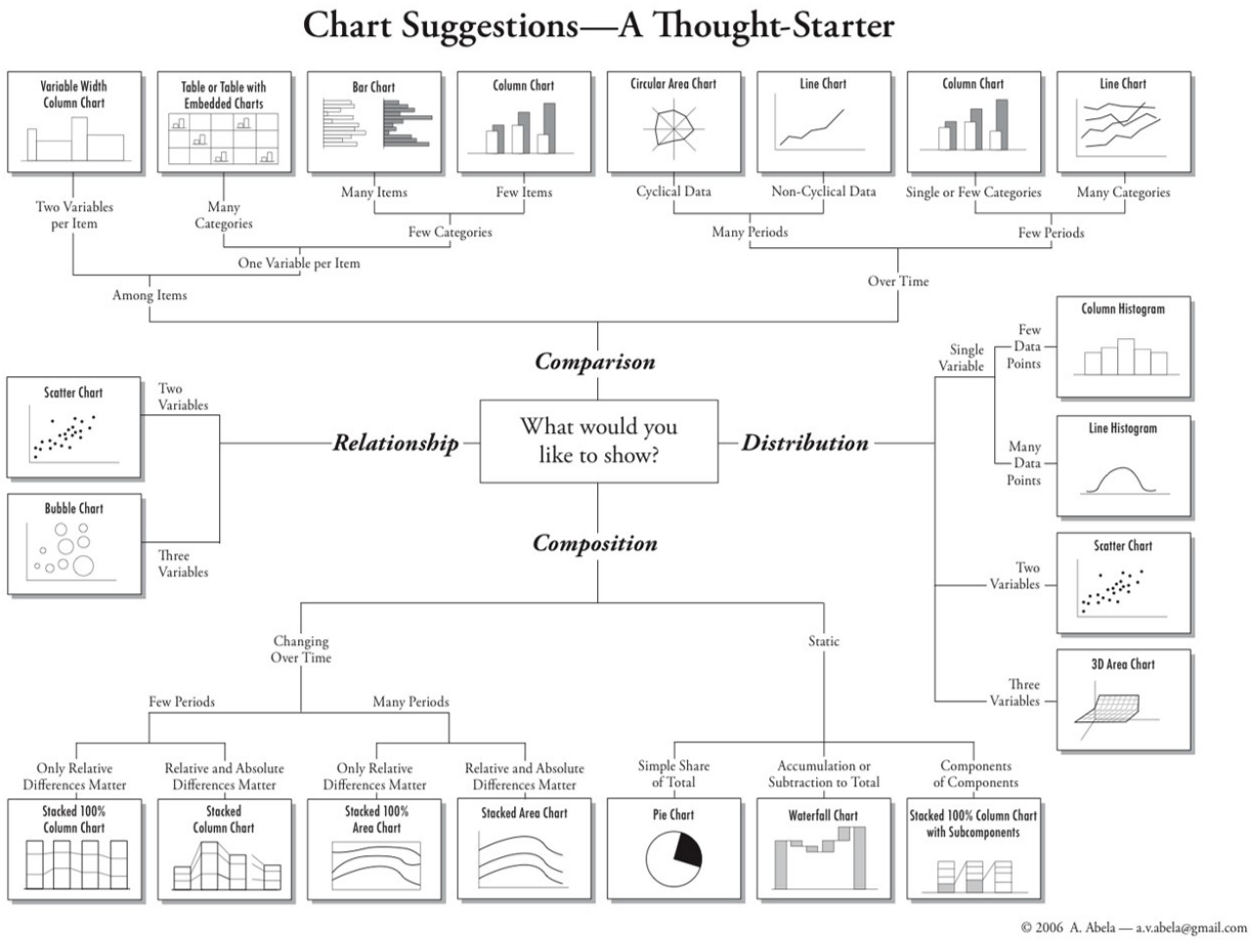
The EDA is a statistical approach to make sense of data by using a variety of techniques (mostly graphical). It may help us

- Assess assumption about variables distribution
- Identify relationship between variables
- Extract important variables
- Suggest use of appropriate models
- Detect problems of collected data (e.g. outliers, missing data, measurement errors)

Statistical Graphics

- **Univarite**
  - Histogram, Stem-and-Leaf, Dot, Q-Q, Density plots
  - Boxplot, Box-and-whisker
  - Bar, Pie, Polar, Waterfall charts
- **Bivariate**
  - XYplot, Line, Area, Scatter, Bubble charts
- **Trivariate**
  - 3D Scatter, Contour, Level/Heatmap, Surface plots

Which Chart to Use?



II. Simple Base Graphics

Iris Dataset

iris setosa	iris versicolor	iris virginica

```
DataX = iris   # ?iris
str(DataX)
```

```
## '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 ...
```

```
dim(DataX)
```

```
## [1] 150   5
```

```
head(DataX)  # tail
```

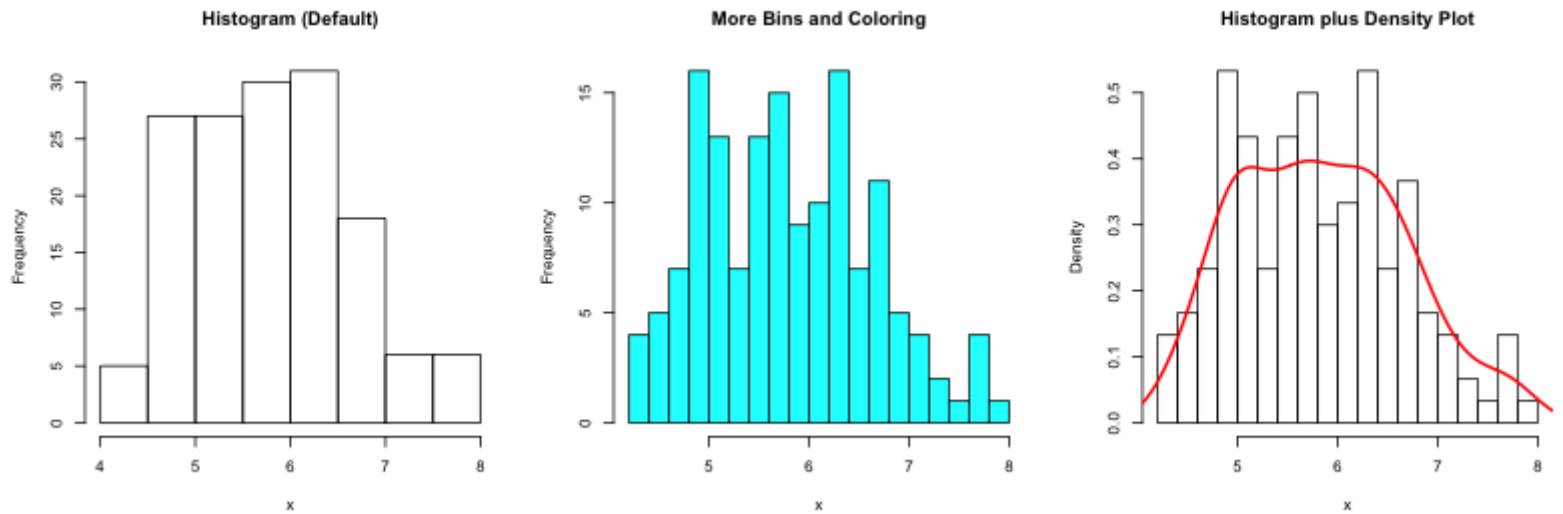
```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

```
summary(DataX)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##  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
```

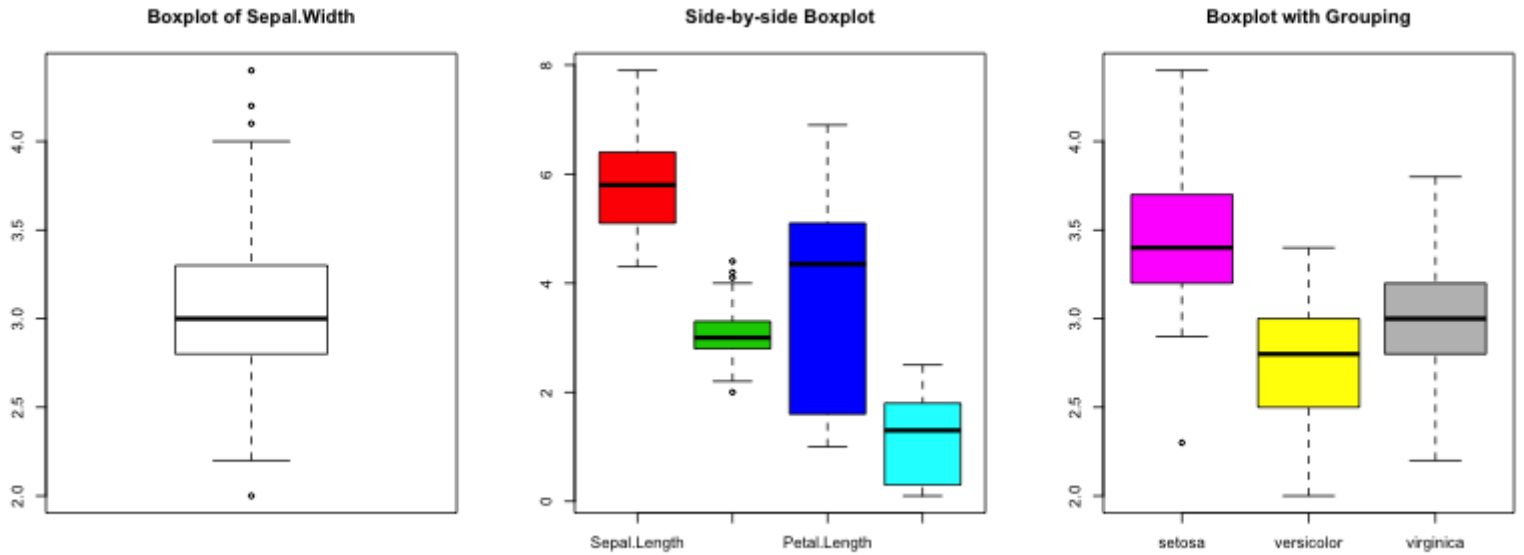
Basic R Plots: Histogram and Density Plot

```
x = DataX$Sepal.Length # a continuous variable
par(mfrow=c(1,3))
hist(x, main='Histogram (Default)')
hist(x, breaks=20, col=5, main='More Bins and Coloring')
hist(x, breaks=20, freq=F, main='Histogram plus Density Plot')   # using freq=FALSE
lines(density(x), col=2, lty=1, lwd=2)  #add the density curve
```



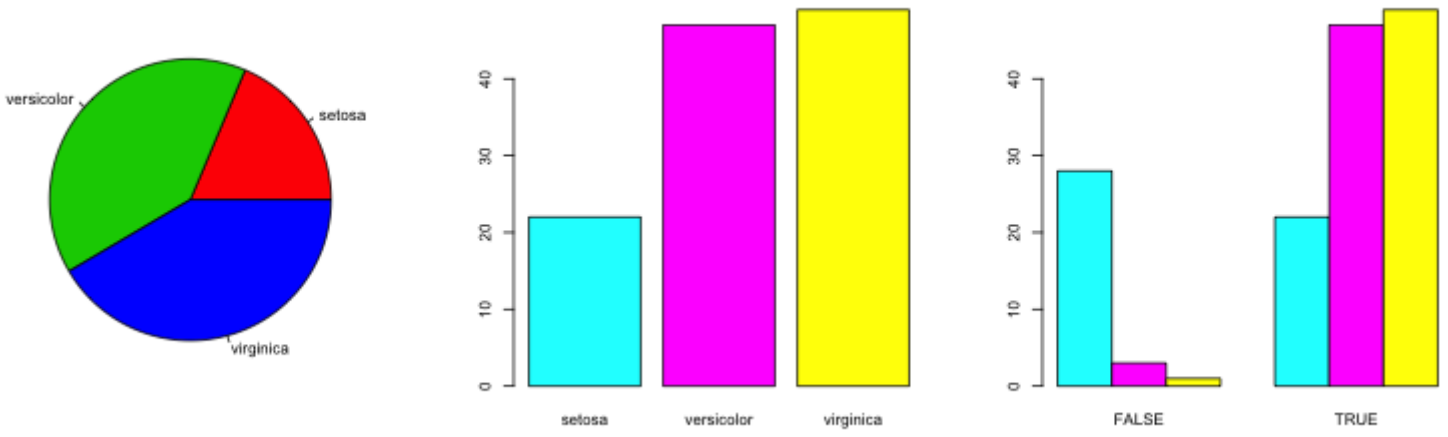
Basic R Plots: Boxplot

```
par(mfrow=c(1,3))
boxplot(DataX$Sepal.Width, main='Boxplot of Sepal.Width')  # Outliers
boxplot(DataX[,1:4], col=c(2,3,4,5), main='Side-by-side Boxplot')
boxplot(Sepal.Width~Species, DataX, col=c(6,7,8), main="Boxplot with Grouping")
```



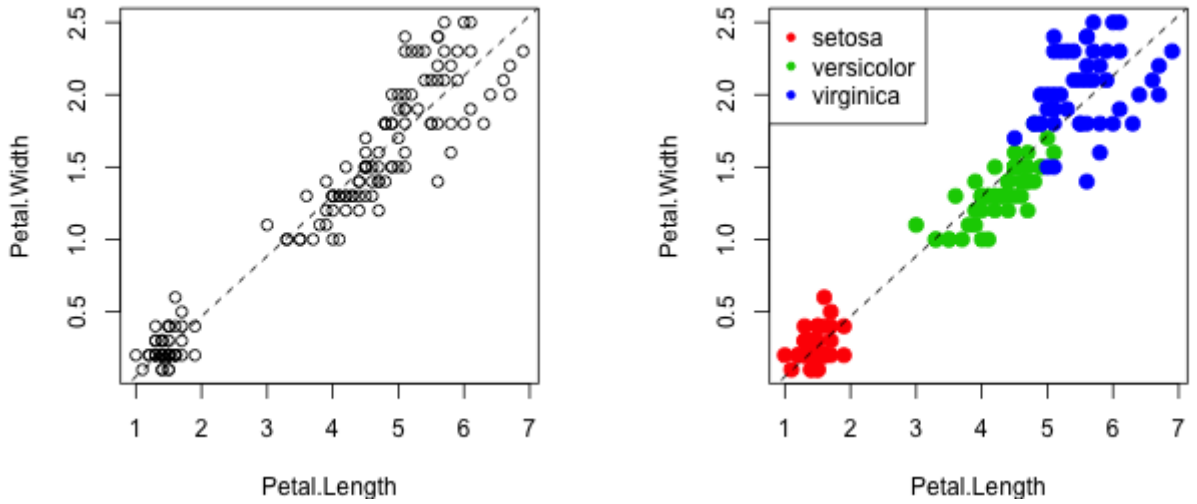
Basic R Plots: Pie and Bar Charts

```
DataX$Flag = DataX$Sepal.Length>5 # Create a binary flag
par(mfrow=c(1,3))
pie(table(DataX$Species[DataX$Flag]), col=c(2,3,4))
barplot(table(DataX$Species[DataX$Flag]), col=c(5,6,7))
barplot(table(DataX$Species, DataX$Flag), col=c(5,6,7), beside=T)
```



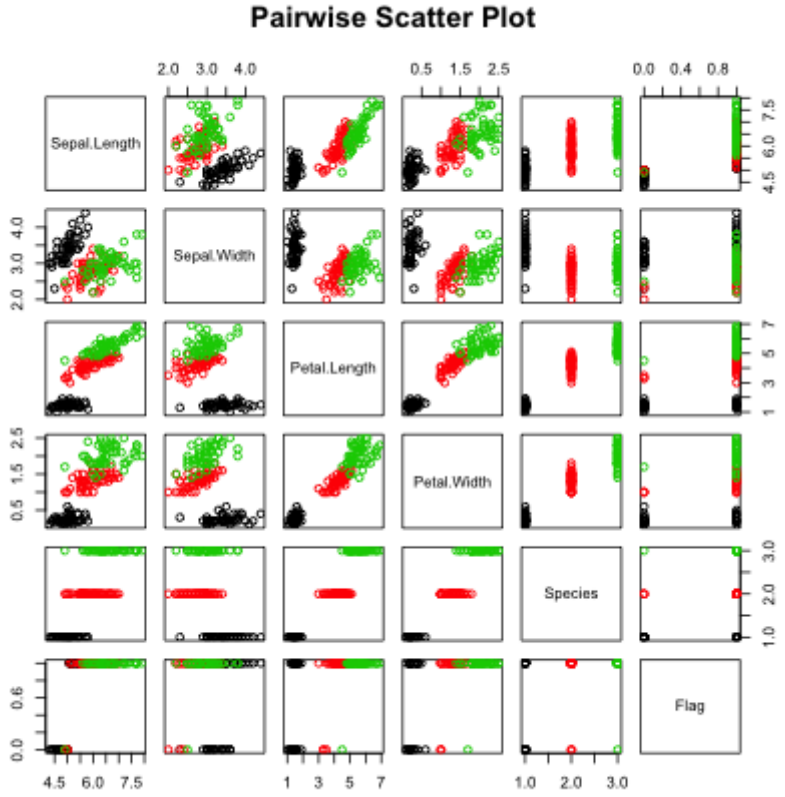
Relationship Between Variables

```
x = DataX$Petal.Length; y = DataX$Petal.Width; z = DataX$Species
par(mfrow=c(1,2)); par(mar=c(4,4,1,4))
plot(x, y, xlab="Petal.Length", ylab="Petal.Width")
abline(coef(lm(y~x)), col=1, lty=2)
plot(x, y, col=c(2,3,4)[z], pch=20, cex=2.0, xlab="Petal.Length", ylab="Petal.Width")
abline(lm(y~x), col=1, lty=2)
legend("topleft", levels(z), pch=20, col=c(2,3,4))
```

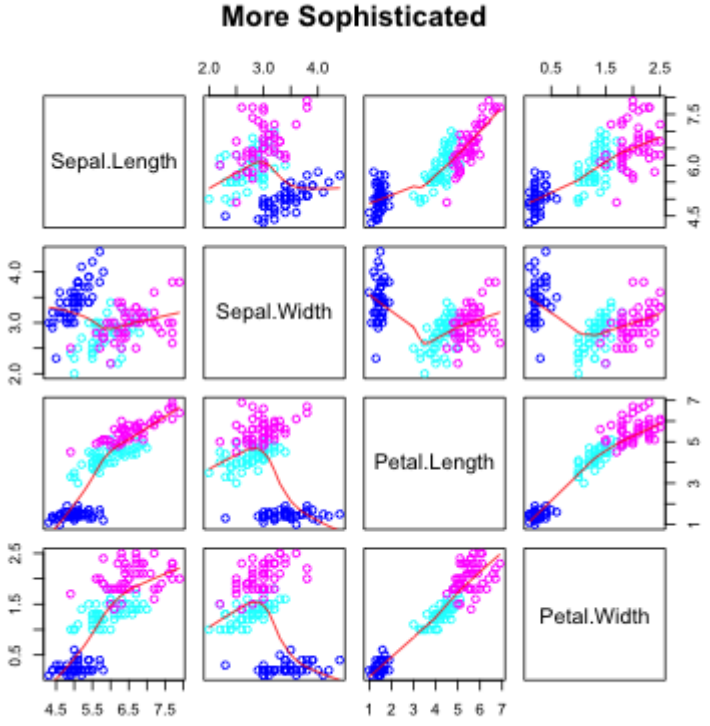


Pairwise Scatter Plot

```
plot(DataX, col=DataX$Species,
     main="Pairwise Scatter Plot")
```

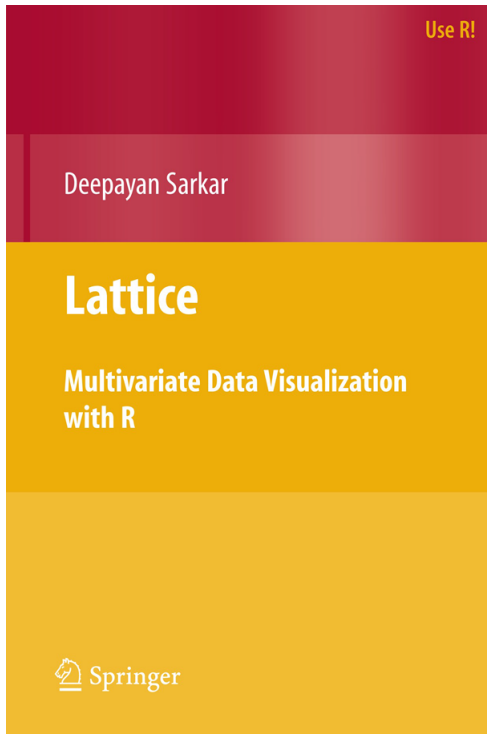



```
pairs(DataX[,1:4], panel = panel.smooth,
      col = c(4,5,6)[DataX$Species],
      main="More Sophisticated")
```



III. Using R:Lattice Package

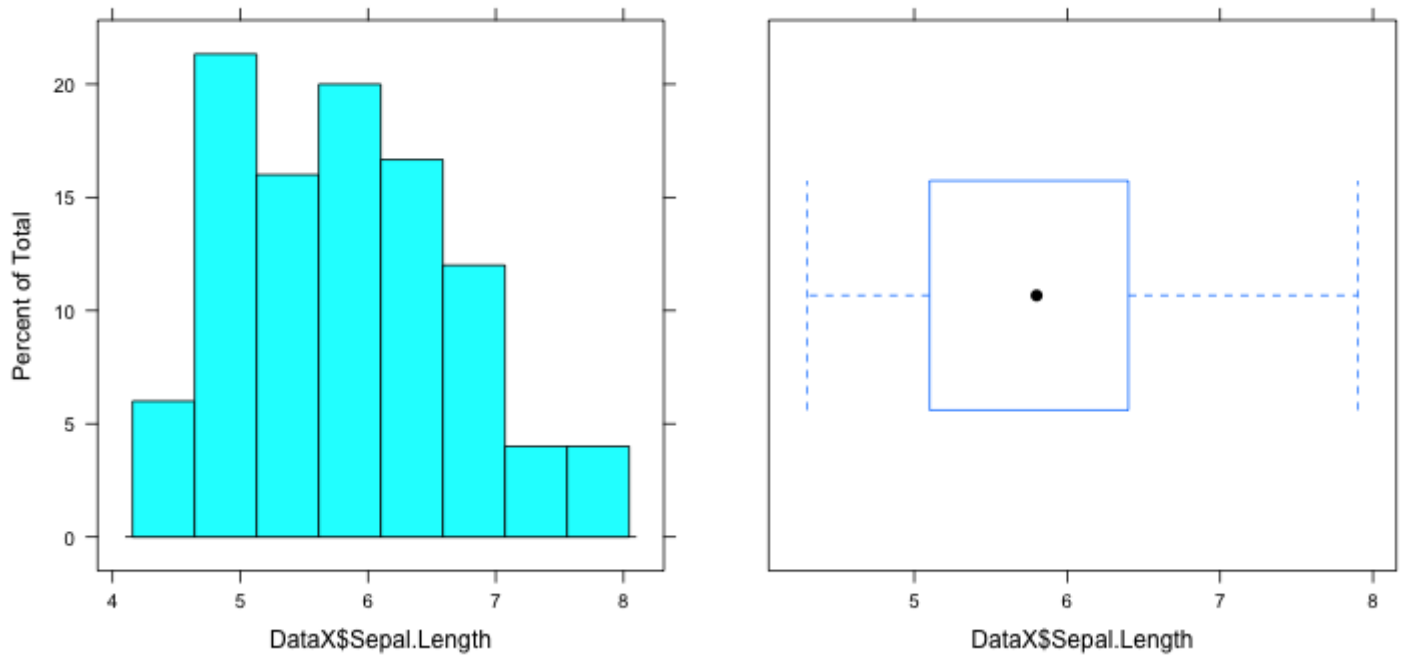
R:Lattice



- Using trellis graphs for multivariate data
- Multipanel conditioning and grouping
- Elegant high-level data visualization
- Covering most of statistical charts
- Figures and Codes can be found at http://lmdvr.r-forge.r-project.org/
- However, plot customization are not so straightforward

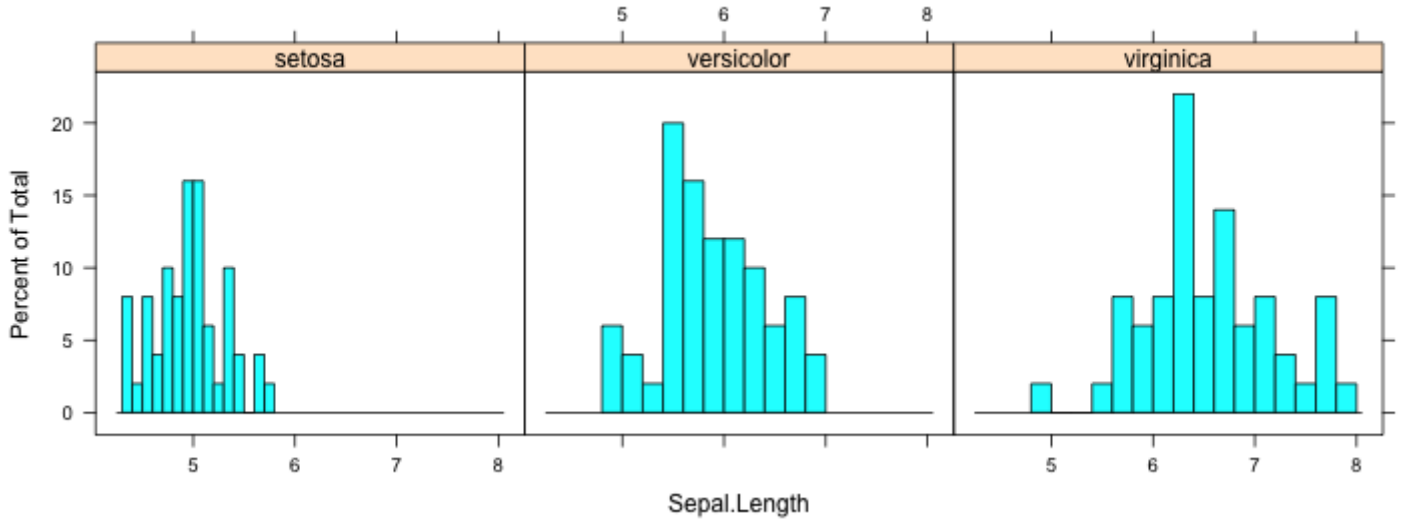
Univariate Distributions

```
library(lattice); library(gridExtra)
p1 = histogram(DataX$Sepal.Length)
p2 = bwplot(DataX$Sepal.Length)
grid.arrange(p1, p2, ncol=2)
```



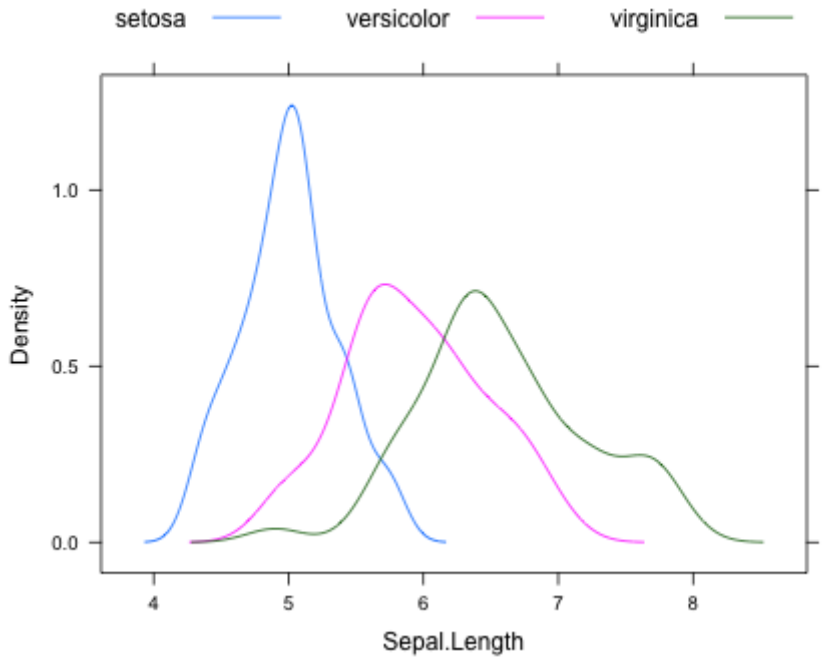
Histogram with Conditioning

```
histogram(data=DataX, ~Sepal.Length|Species, breaks=12, layout = c(3, 1))
```



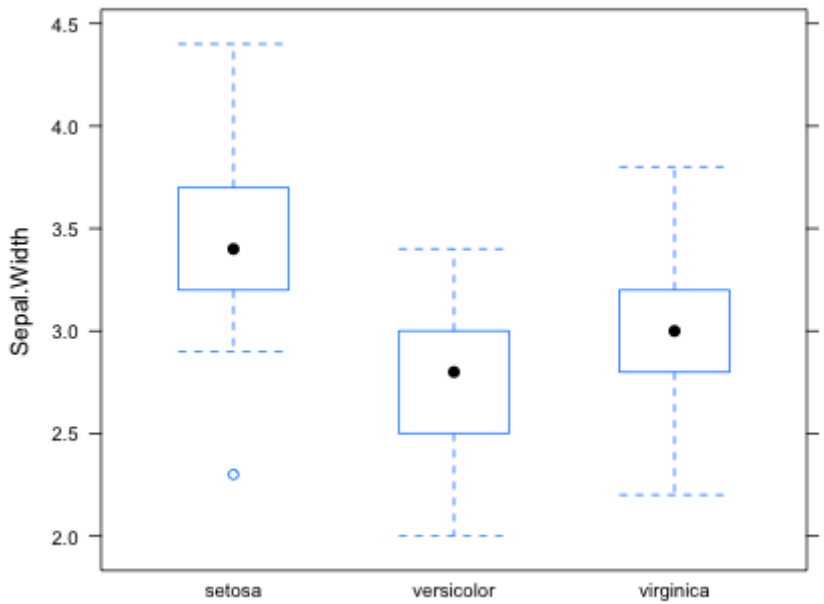
Density plot with Grouping

```
densityplot(data=DataX, ~Sepal.Length, groups=Species,
            plot.points=F, auto.key=list(space="top", columns=3))
```



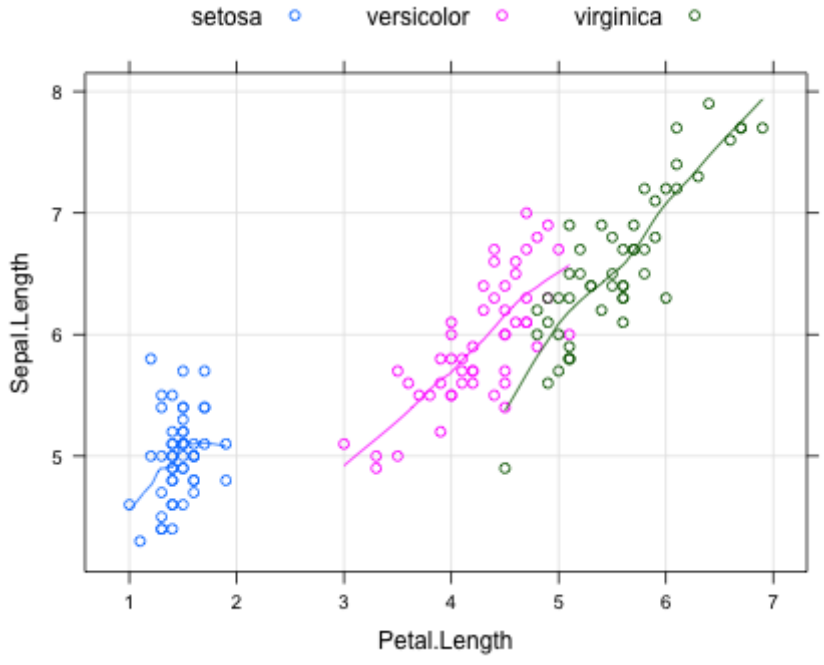
Boxplot with Grouping

```
bwplot(data=DataX, Sepal.Width~Species)
```



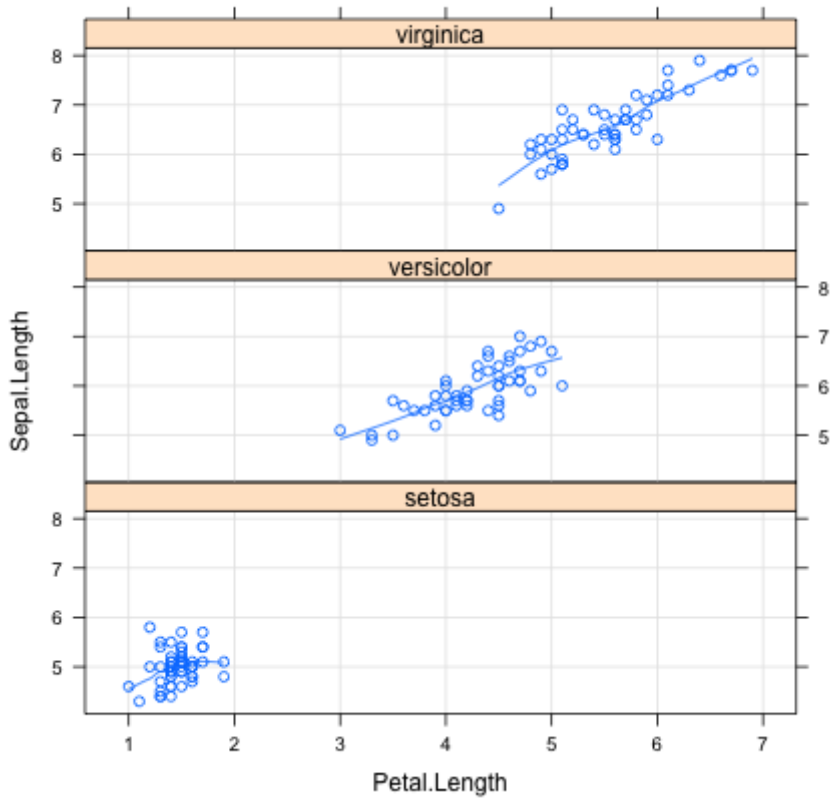
Bivariate plot with Grouping

```
xyplot(data=DataX, Sepal.Length ~ Petal.Length, groups = Species,
       type = c("p", "smooth",  "g"),
       auto.key = list(space="top", columns=3)) # grouping
```



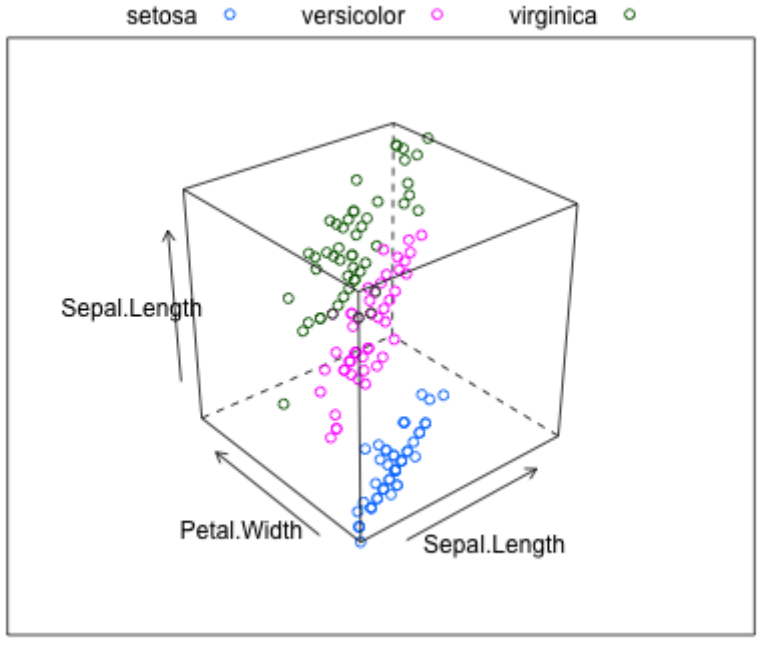
Bivariate plot with Conditioning

```
xyplot(data=DataX, Sepal.Length ~ Petal.Length|Species,
       type=c("p", "smooth",  "g"), layout=c(1,3)) # conditioning
```



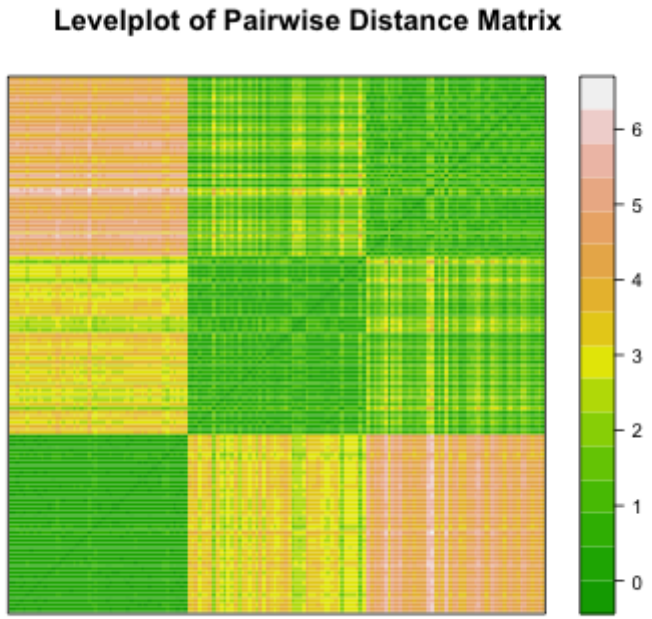
Trivariate 3D Plot

```
cloud(data=DataX, Sepal.Length ~ Sepal.Length * Petal.Width, groups = Species,
      auto.key = list(space="top", columns=3), panel.aspect = 0.8)
```



Trivariate Heatmap

```
dist = as.matrix(dist(DataX[,3:4]))
levelplot(dist, colorkey = T, col.regions = terrain.colors,
          scales = list(at=c(0,0),tck = c(0,0)),
          xlab="",ylab="",main="Levelplot of Pairwise Distance Matrix")
```



Thank you!

Q&A or Email ajzhang@hku.hk。