FACULTY OF ENGINEERING
AT SRI RACHA
DEPARMENT OF COMPUTER ENGINEERING

03603351 วิทยาศาสตร์ข้อมูลเบื้องต้น

Introduction to Data Science

การสำรวจข้อมูล

Data Exploration

ผศ.ดร. กุลวดี สมบูรณ์วิวัฒน์

kulwadee@eng.src.ku.ac.th

ภาคการศึกษาที่ 1 ปี 2565

การสำรวจข้อมูล (Data Exploration)

- Data คือ ผลจากการสังเกต หรือ ข้อเท็จจริงเกี่ยวกับสิ่งที่ต้องการศึกษา
- การสำรวจข้อมูล (data exploartion) หรือ Exploratory Data Analysis
- ทำให้เราเข้าใจคุณลักษณะพื้นฐานของข้อมูลมากขึ้น
- เตรียมข้อมูลให้อยู่ในรูปแบบที่เหมาะสมกับการวิเคราะห์ข้อมูลขั้นสูงในขั้นตอนต่อไป
- ทำให้ได้แนวทางในการเลือกใช้เครื่องมือทางสถิติและทางวิทยาศาสตร์ข้อมูลที่เหมาะสม
- แบ่งได้เป็น 2 ประเภทคือ **สถิติเชิงพรรณนา (descriptive statistics)** และ **การทำให้เป็นภาพ (data visualization)**

ประเภทของการสำรวจข้อมูล (Data Exploration)

- **สถิติเชิงพรรณนา (descriptive statistics)**

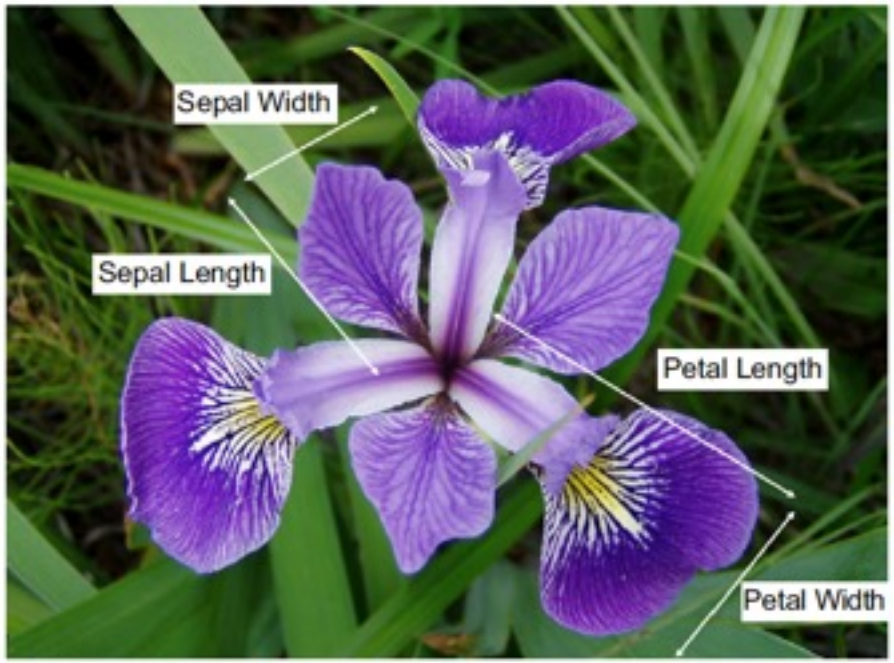
  การสรุปย่อคุณสมบัติของชุดข้อมูลโดยใช้ตัวชี้วัดเชิงตัวเลข เช่น ค่าเฉลี่ย มัธยฐาน ค่าเบี่ยงเบนมาตรฐาน เป็นต้น

- **การทำให้เป็นภาพ (data visualization)**

  การแสดงจุดข้อมูลลงบนระนาบหลายมิติ หรือ รูปเชิงนามธรรม เช่น scatterplot, histogram

ชุดข้อมูล (Dataset)

- Iris Dataset (Ronald Fisher: https://onlinelibrary.wiley.com/doi/10.1111/j.1469-1809.1936.tb02137.x)



- 150 observations of 3 species
- each observation consists of 4 numerical attributes and 1 class label
  - sepal length
  - sepal width
  - petal length
  - petal width
  - species

sepal_length	sepal_width	petal_length	petal_width	species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
...	...	...	...	...
7.0	3.2	4.7	1.4	versicolor
6.4	3.2	4.5	1.5	versicolor
6.9	3.1	4.9	1.5	versicolor
5.5	2.3	4.0	1.3	versicolor
6.5	2.8	4.6	1.5	versicolor
...	...	...	...	...
6.7	3.0	5.2	2.3	virginica
6.3	2.5	5.0	1.9	virginica
6.5	3.0	5.2	2.0	virginica
6.2	3.4	5.4	2.3	virginica
5.9	3.0	5.1	1.8	virginica

- 150 observations of 3 species
- each observation consists of 4 numerical attributes and 1 class label
  - sepal length
  - sepal width
  - petal length
  - petal width
  - species

ชนิดของข้อมูล

ชนิดข้อมูลเป็นตัวกำหนดโอเปอเรชั่นที่สามารถกระทำกับข้อมูลได้ ชนิดของข้อมูลแบ่งออกได้เป็น 2 ประเภทหลักคือ

- **ข้อมูลตัวเลข (Numeric or Continuous Data)** เช่น 10, 0.898
  - สามารถใช้โอเปอเรชั่นทางคณิตศาสตร์ (เช่น +, -, *, /) และการเปรียบเทียบ (เช่น <, >, =, !=) ได้
  - แบ่งเป็นสองกลุ่มคือ จำนวนเต็ม และ จำนวนจริง
- **ข้อมูลที่เป็นชื่อหรือสัญลักษณ์ (Nominal or Categorical)** เช่น species (setosa, virginica, versicolor) ช่วงอุณหภูมิ (hot, mild, cold) เป็นต้น
  - แบ่งเป็นสองกลุ่ม คือ unordered nominal (species: setosa, virginica, versicolor) และ ordered nominal (อุณหภูมิ hot, mild, cold)
- **เราสามารถแปลงรูปแบบข้อมูลจากตัวเลขไปเป็นสัญลักษณ์ หรือกลับกันก็ได้**

Descriptive Statistics

การศึกษาเกี่ยวกับปริมาณเชิงรวมของชุดข้อมูล แบ่งเป็นสองชนิดคือ

- Univariate Exploration: ศึกษาแอทริบิวต์เพียงตัวเดียว
- Multivariate Exploration ศึกษาความสัมพันธ์ระหว่างแอทริบิวต์ตั้งแต่สองตัวขึ้นไป

Characteristics of the Data Set	Measurement Technique
Center of the data set	Mean, median, and mode
Spread of the data set	Range, variance, and standard deviation
Shape of the distribution of the data set	Symmetry, skewness, and kurtosis

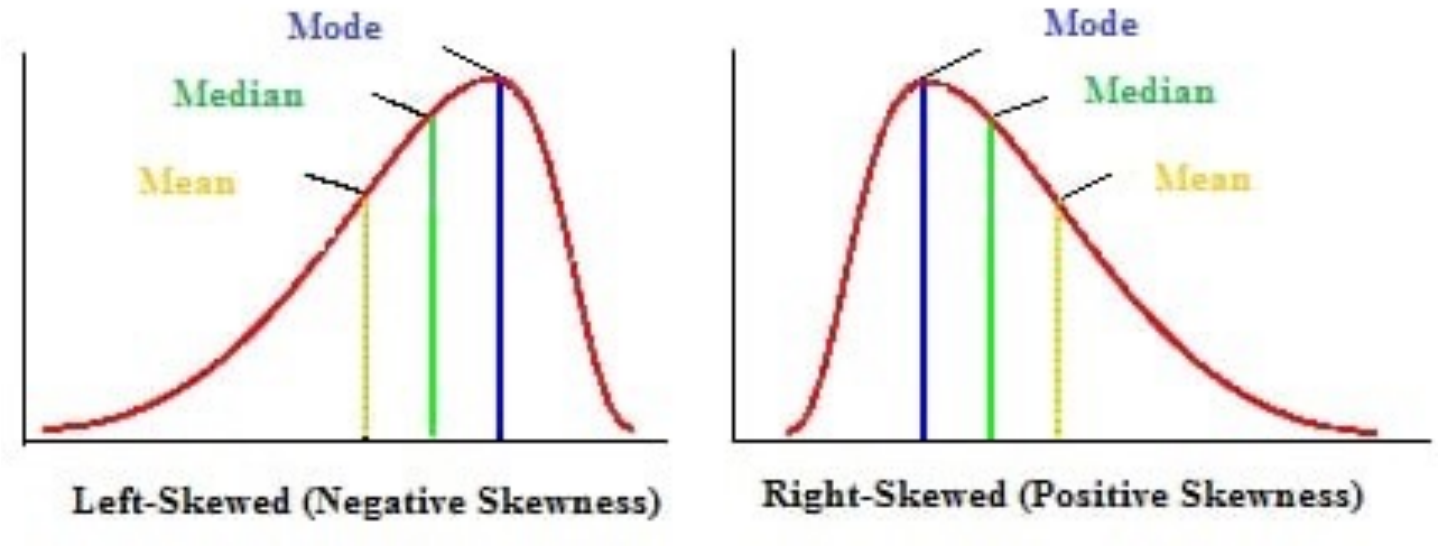
Univariate Descriptive Statistics

Table 3.1 Iris Data Set and Descriptive Statistics (Fisher, 1936)

Observation	Sepal Length	Sepal Width	Petal Length	Petal Width
1	5.1	3.5	1.4	0.2
2	4.9	3.1	1.5	0.1
...	...	...	...	...
49	5	3.4	1.5	0.2
50	4.4	2.9	1.4	0.2
Statistics	**Sepal Length**	**Sepal Width**	**Petal Length**	**Petal Width**
Mean	5.006	3.418	1.464	0.244
Median	5.000	3.400	1.500	0.200
Mode	5.100	3.400	1.500	0.200
Range	1.500	2.100	0.900	0.500
Standard Deviation	0.352	0.381	0.174	0.107
Variance	0.124	0.145	0.030	0.011

Measure of Central Tendency

- การสรุปคุณสมบัติเชิงปริมาณของชุดข้อมูลโดยใช้แนวโน้มสู่ศูนย์กลาง
  - ค่าเฉลี่ย (Mean)
  - มัธยฐาน (Medain)
  - ฐานนิยม (Mode)
- ค่าเฉลี่ย มัธยฐาน และฐานนิยม อาจมีค่าแตกต่างกันได้ ขึ้นกับรูปร่างของการกระจายของข้อมูล
- ค่าผิดปกติ (outlier) ส่งผลกระทบต่อค่าเฉลี่ย แต่ไม่กระทบต่อค่ามัธยฐาน



source: https://www.statisticshowto.com/pearson-mode-skewness/

Measure of Spread

- พิสัย (Range) = max - min
  สนใจเฉพาะค่ามากที่สุดและน้อยที่สุด
  ค่าผิดปกติส่งผลกระทบต่อค่าพิสัย

- ความเบี่ยงเบน (Deviation)
  วัดการกระจายของข้อมูลโดยพิจารณาจากจุดข้อมูลทุกจุดในชุดข้อมูล

For samples:

$$\text{variance} = s^2 = \frac{\Sigma (x - \bar{x})^2}{n - 1}$$

$$\text{standard deviation} = s = \sqrt{s^2}$$

Calculating Formula

$$s^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n - 1}$$

For populations:

$$\text{variance} = \sigma^2 = \frac{\Sigma (x - \bar{x})^2}{n}$$

$$\text{standard deviation} = \sigma = \sqrt{\sigma^2}$$

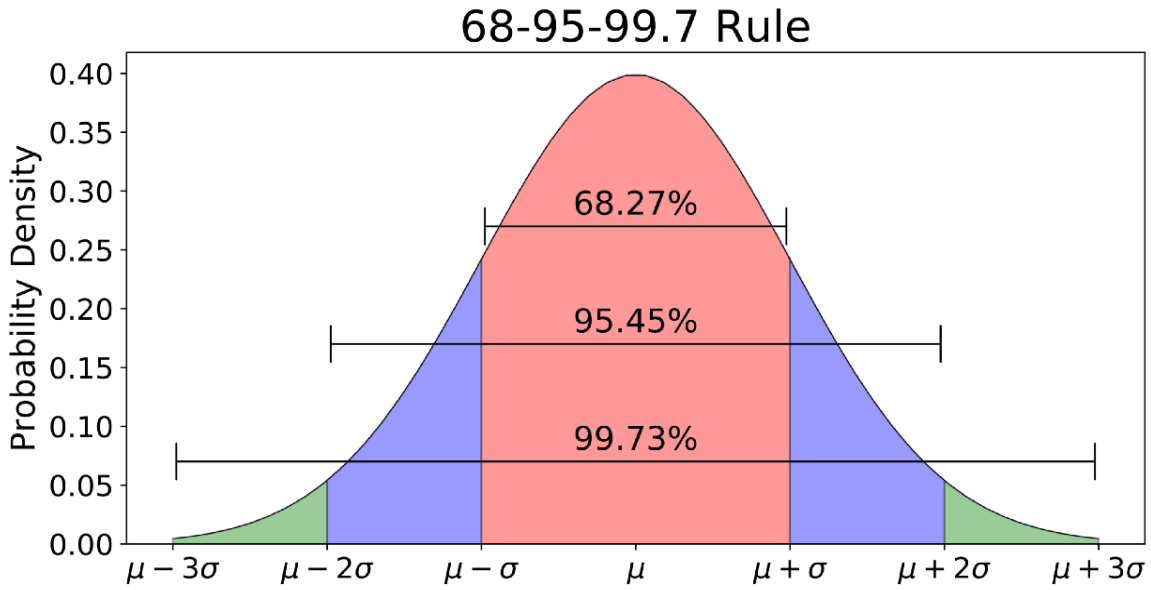
Calculating Formula

$$\sigma^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n}$$

Where

- X is individual one value
- N is size of population
- $\bar{x}$ is the mean of population

source: http://makemeanalyst.com/explore-your-data-variance-and-standard-deviation/



$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Normal Distribution

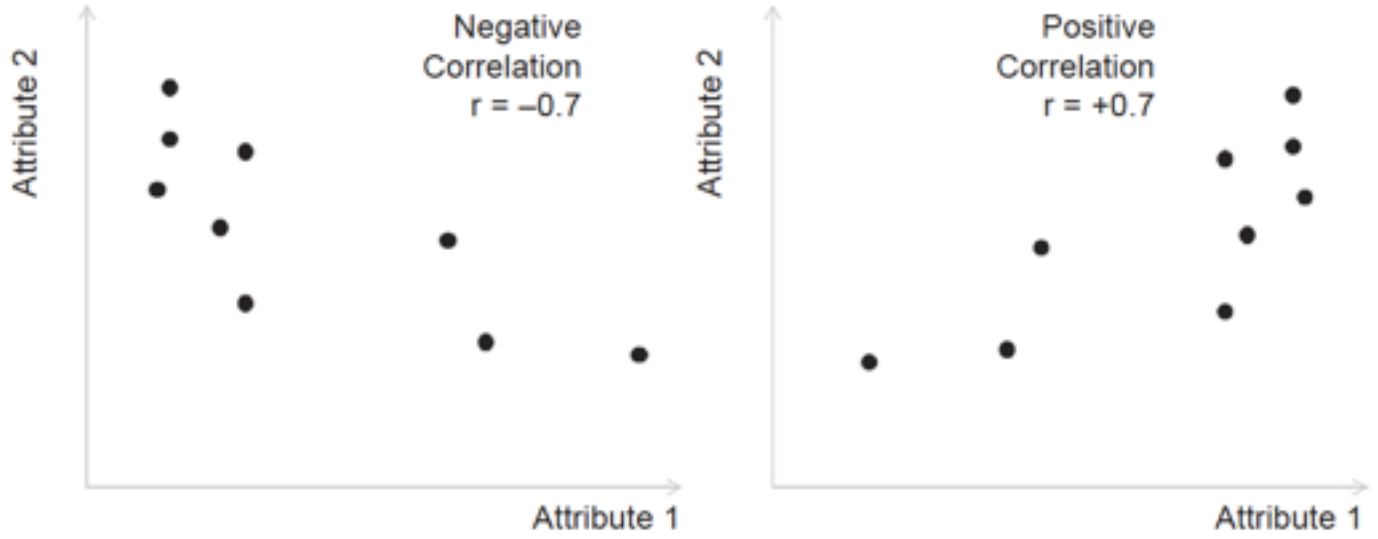
source: https://towardsdatascience.com/understanding-the-68-95-99-7-rule-for-a-normal-distribution-b7b7cbf760c2

Multivariate Descriptive Statistics

Central datapoint

observation i: {sepal length, sepal width, petal length, petal width}

Correlation



Pearson Correlation

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

$$= \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{N * S_x * S_y}$$

ข้อจำกัดของสถิติเชิงพรรณนา

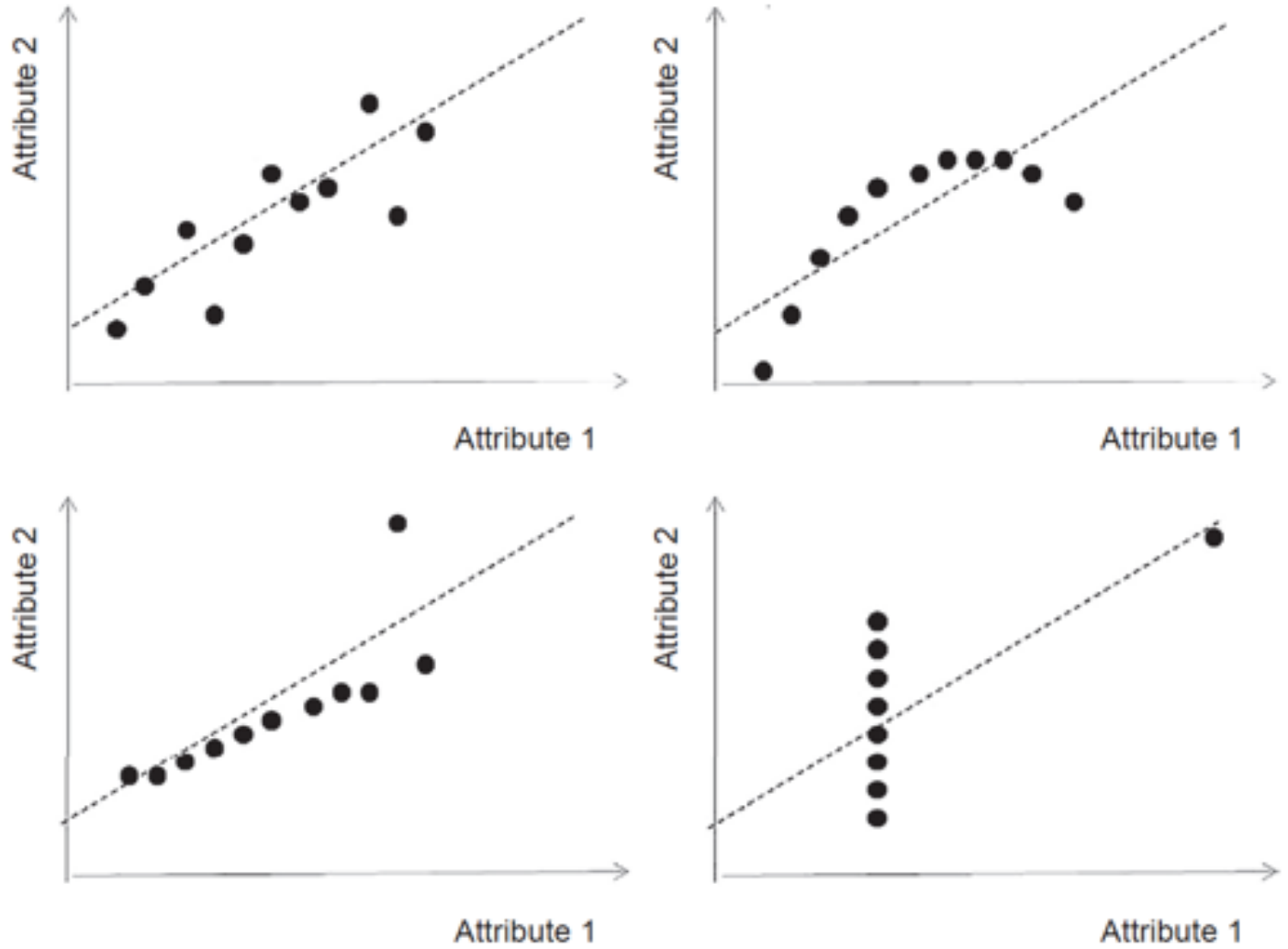


FIGURE 3.4

Anscombe's Quartet: descriptive statistics vs. visualization (Anscombe, F. J., 1973. Graphs in Statistical Analysis, *American Statistician 27* (1), pp. 19–20.)

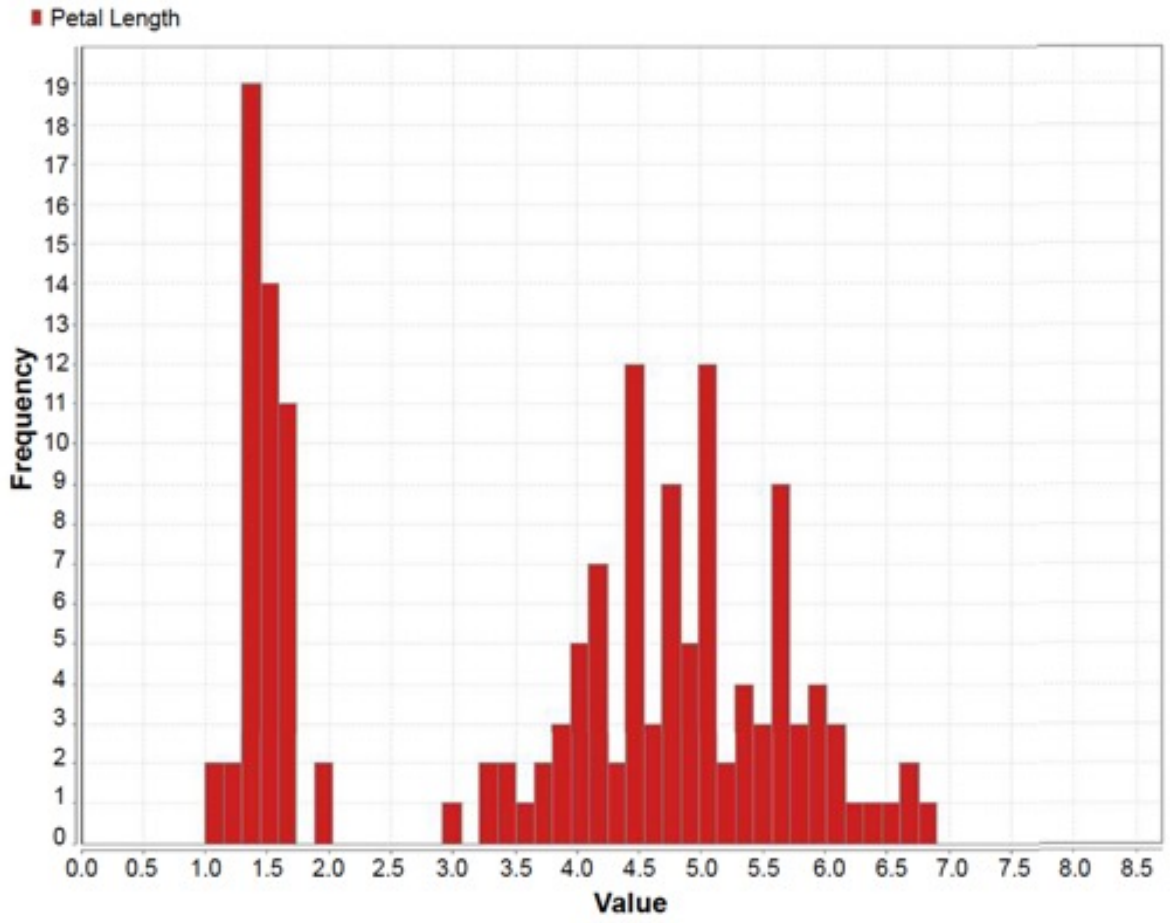
ชุดข้อมูลทั้งสี่ มีค่าเฉลี่ย ค่าความแปรปรวน และ ค่าสัมประสิทธิ์สหสัมพันธ์เท่ากัน!!

Data Visualization

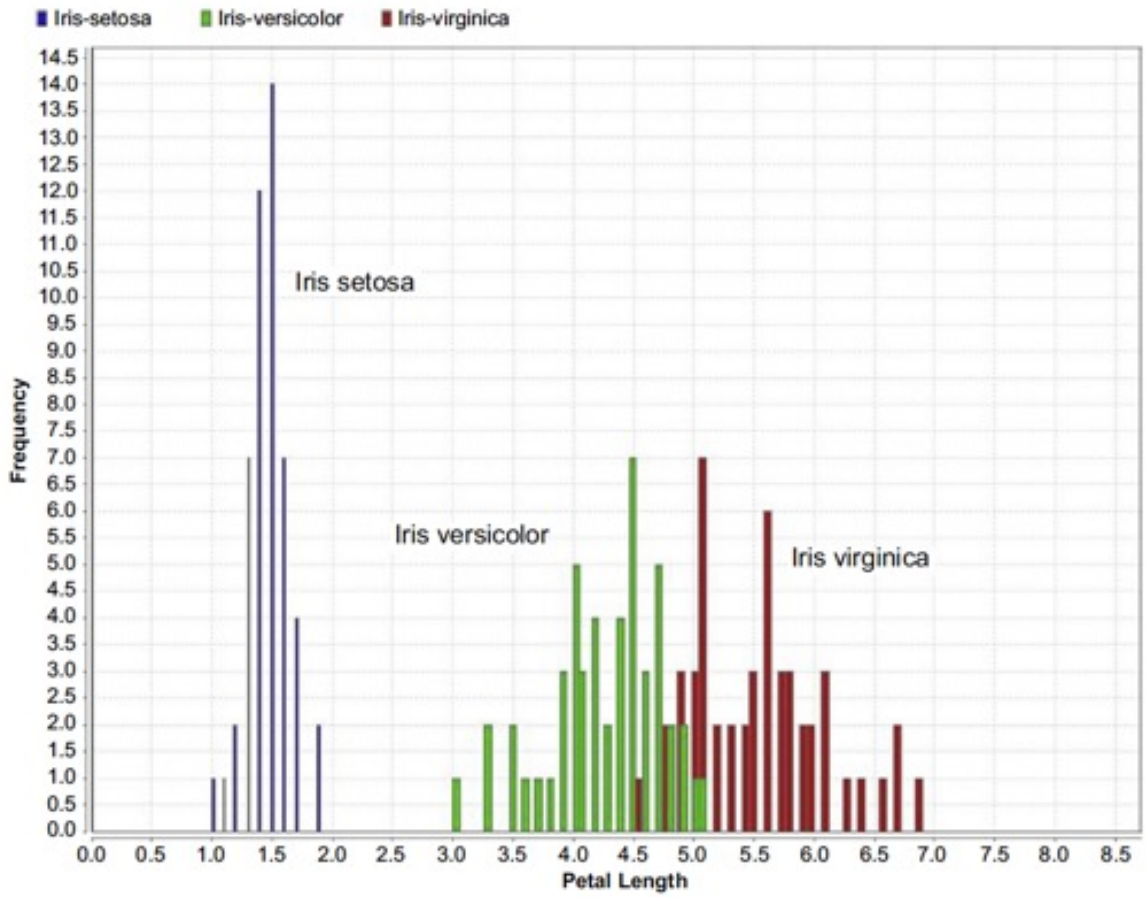
- Univariate
  - Histogram
  - Quartile (box whisker plot)
  - Distribution Chart

- Multivariate
  - Scatterplot
  - Scatter Multiple
  - Scatter matrix
  - Bubble Chart
  - Density Chart
  - Parallel Chart
  - Deviation Chart
  - Andrew Curves

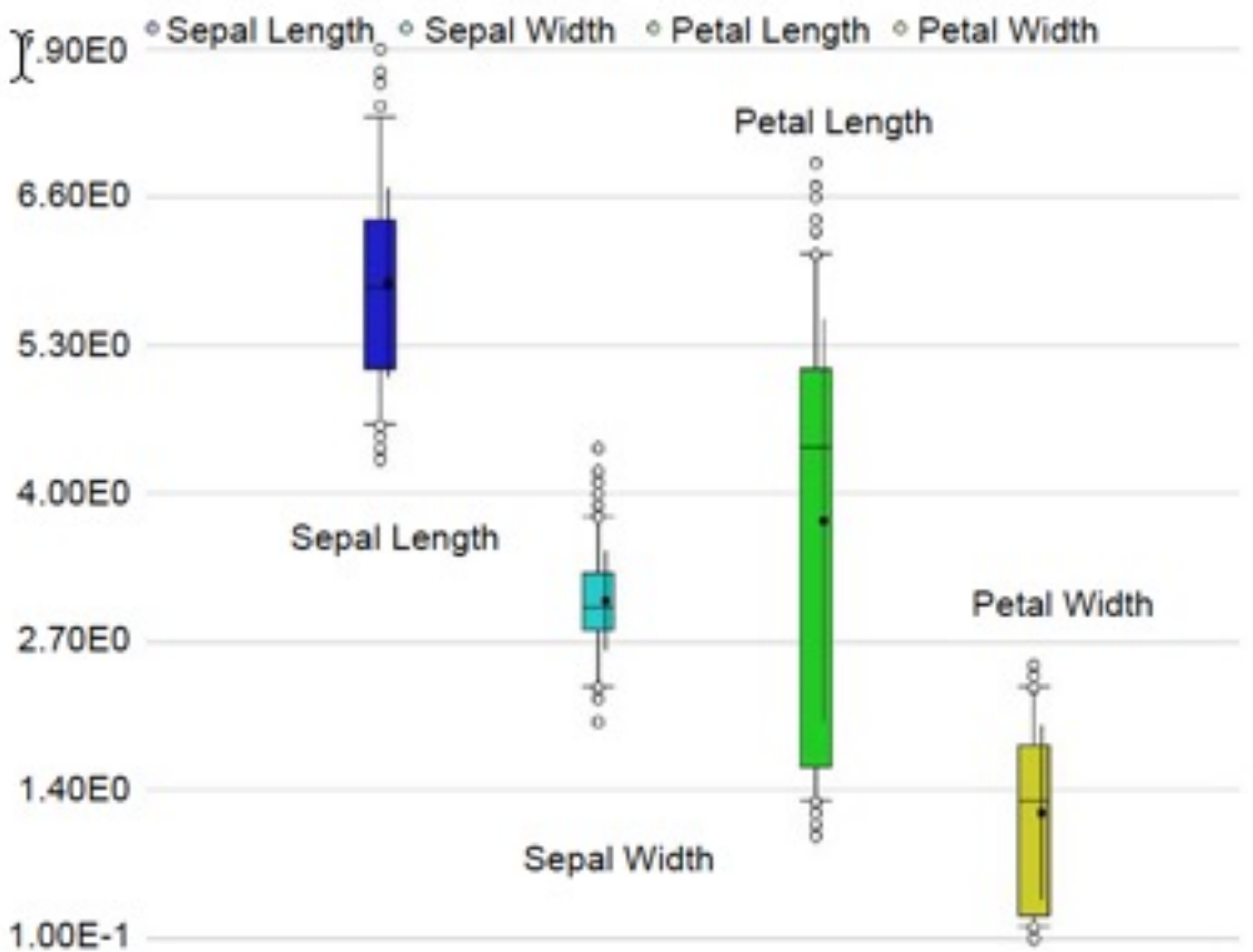
Histogram



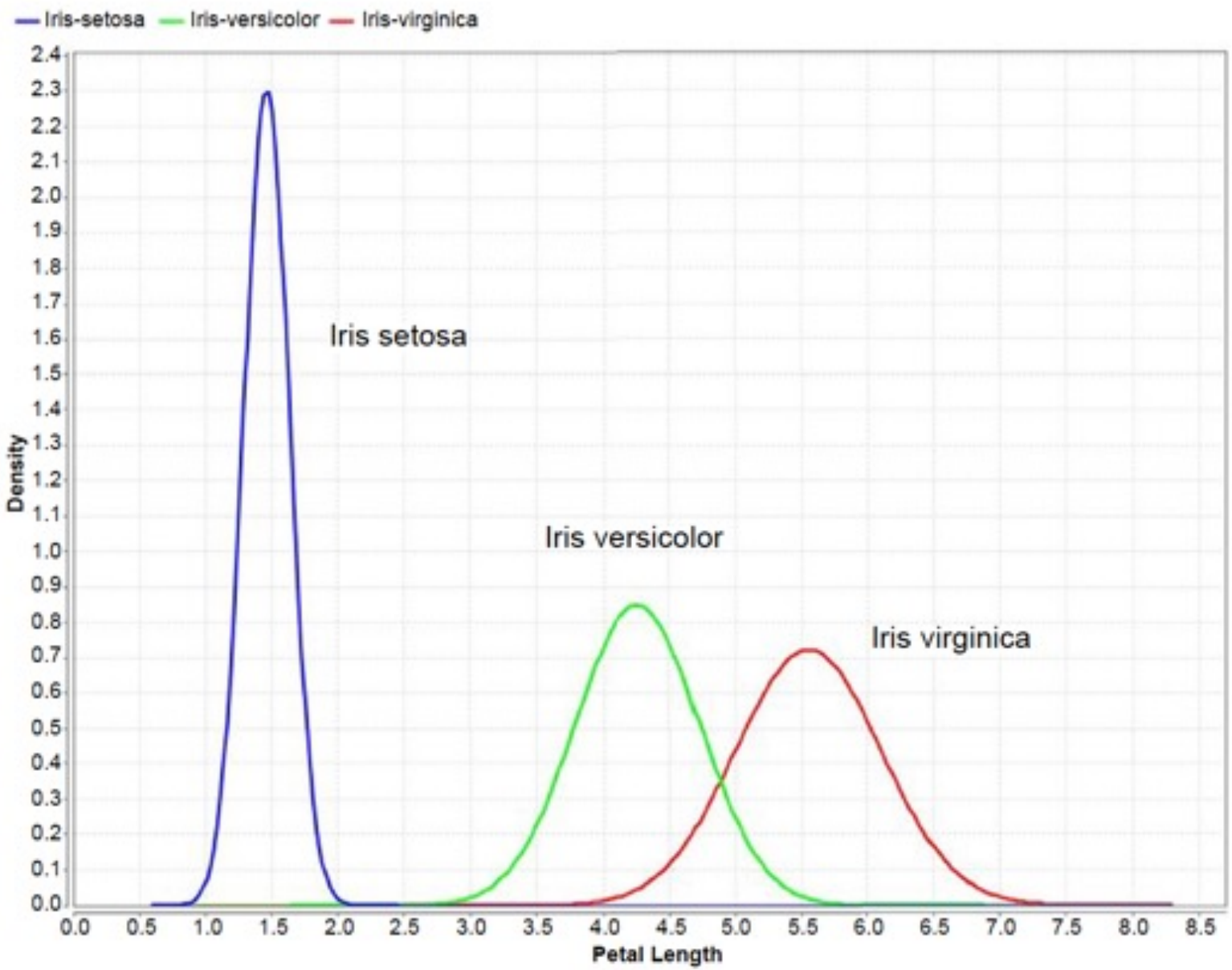
Class Stratified Histogram



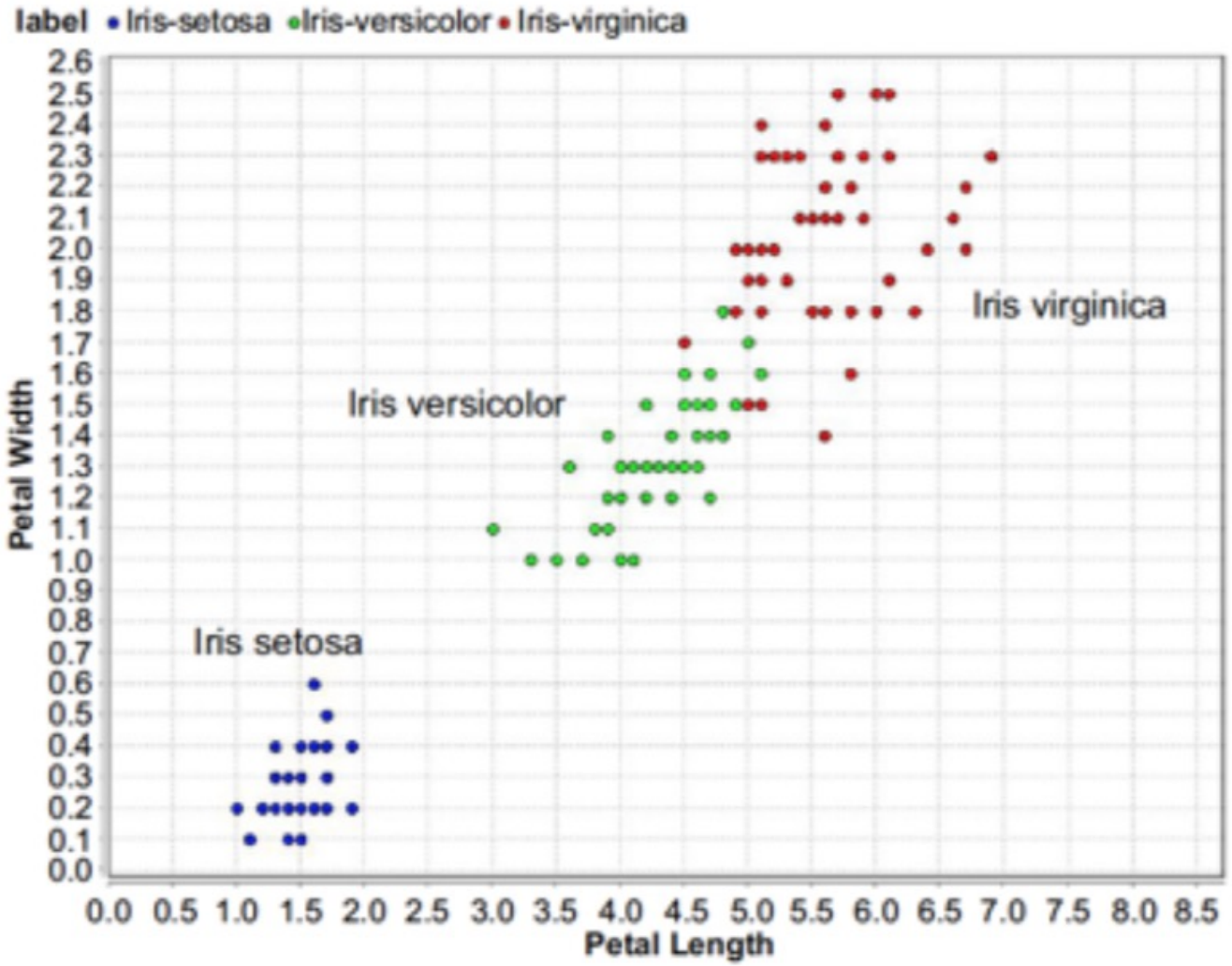
Quartile plot (Box Whisker Plot)



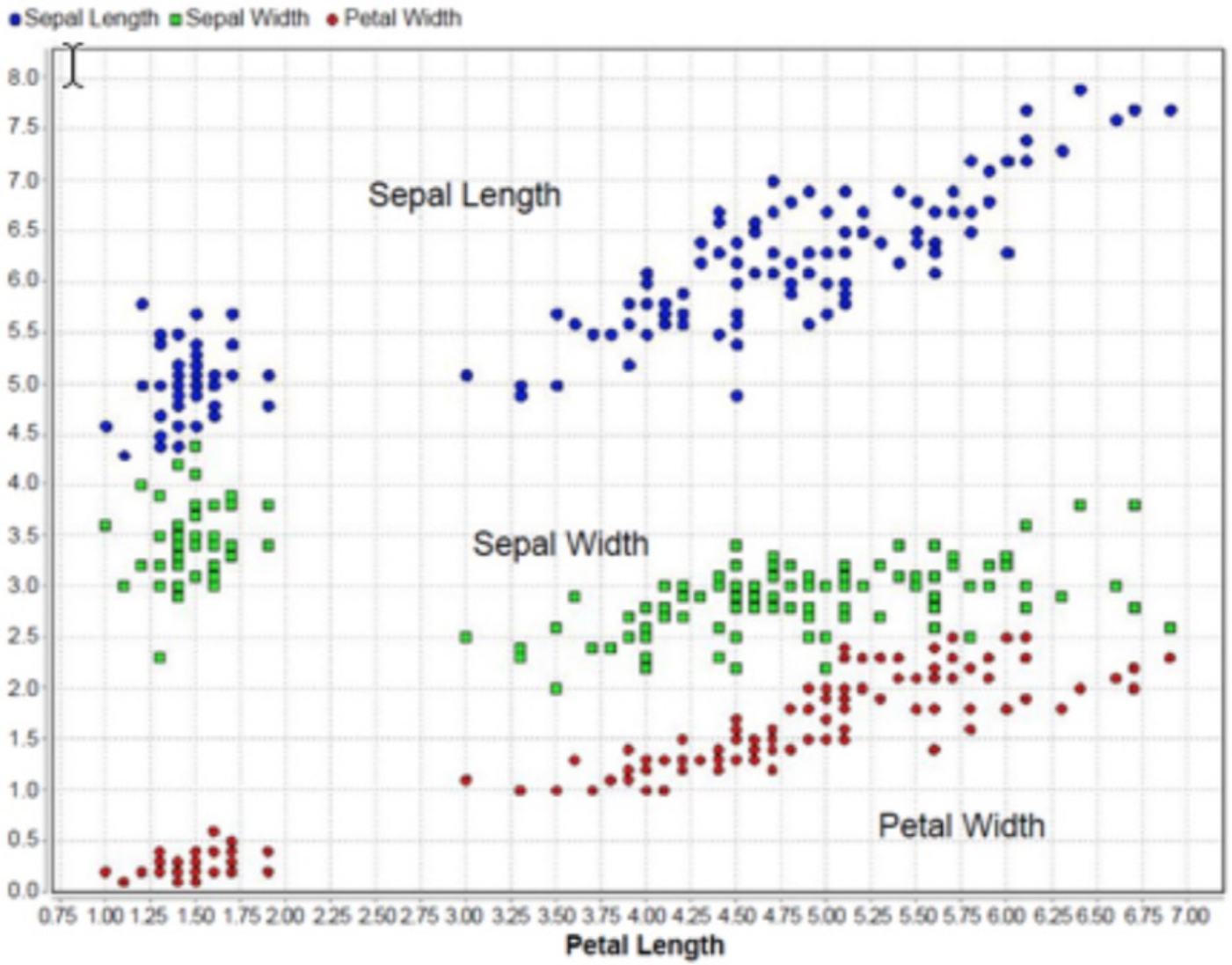
Distribution plot



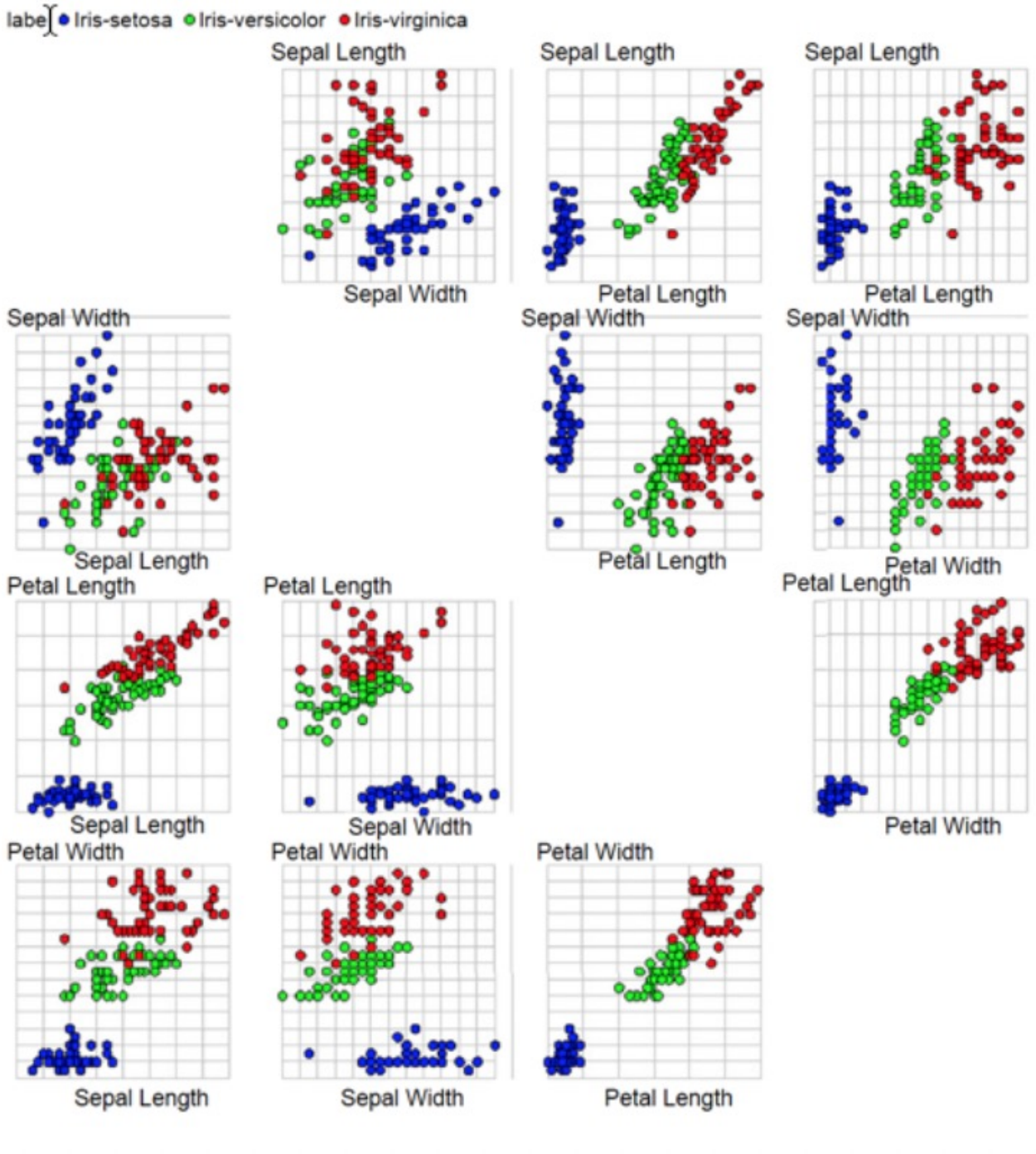
Scatter plot



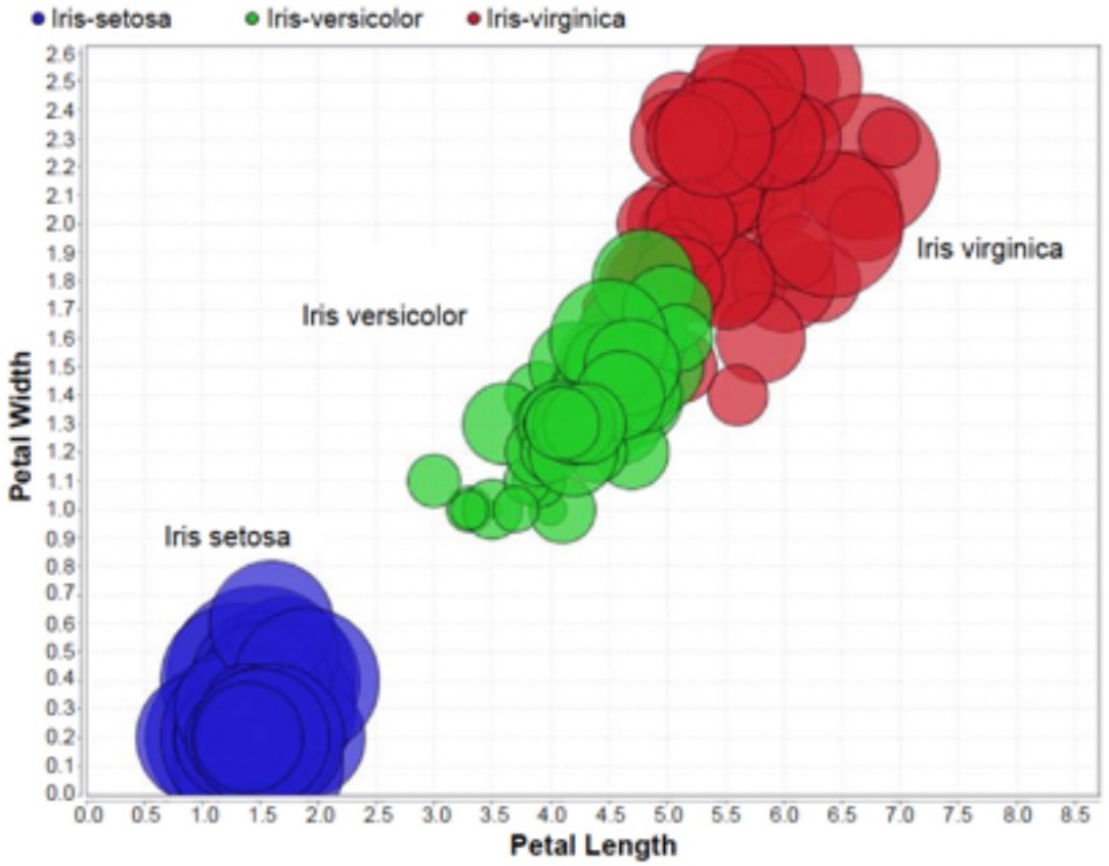
Scatter Multiple



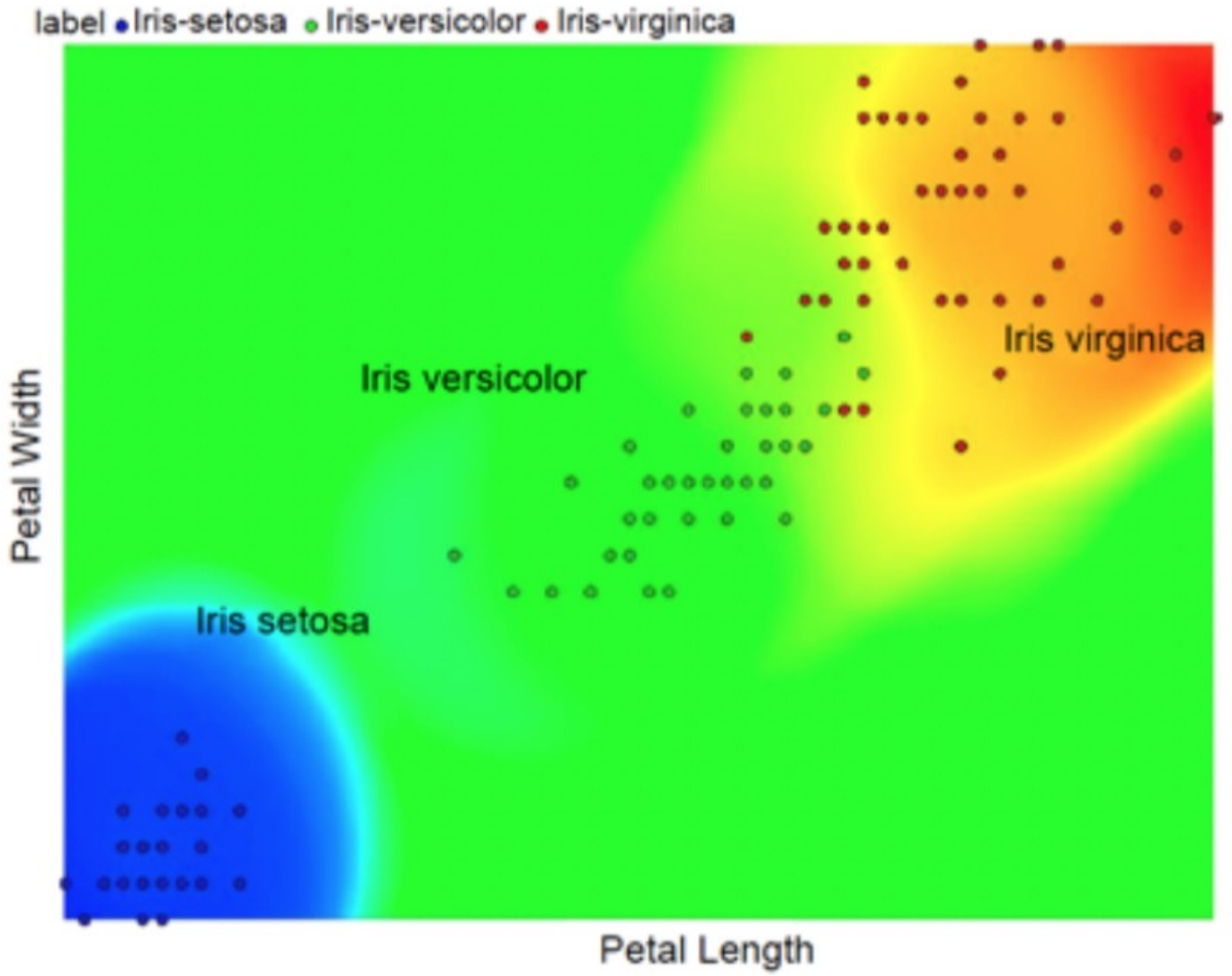
Scatter matrix



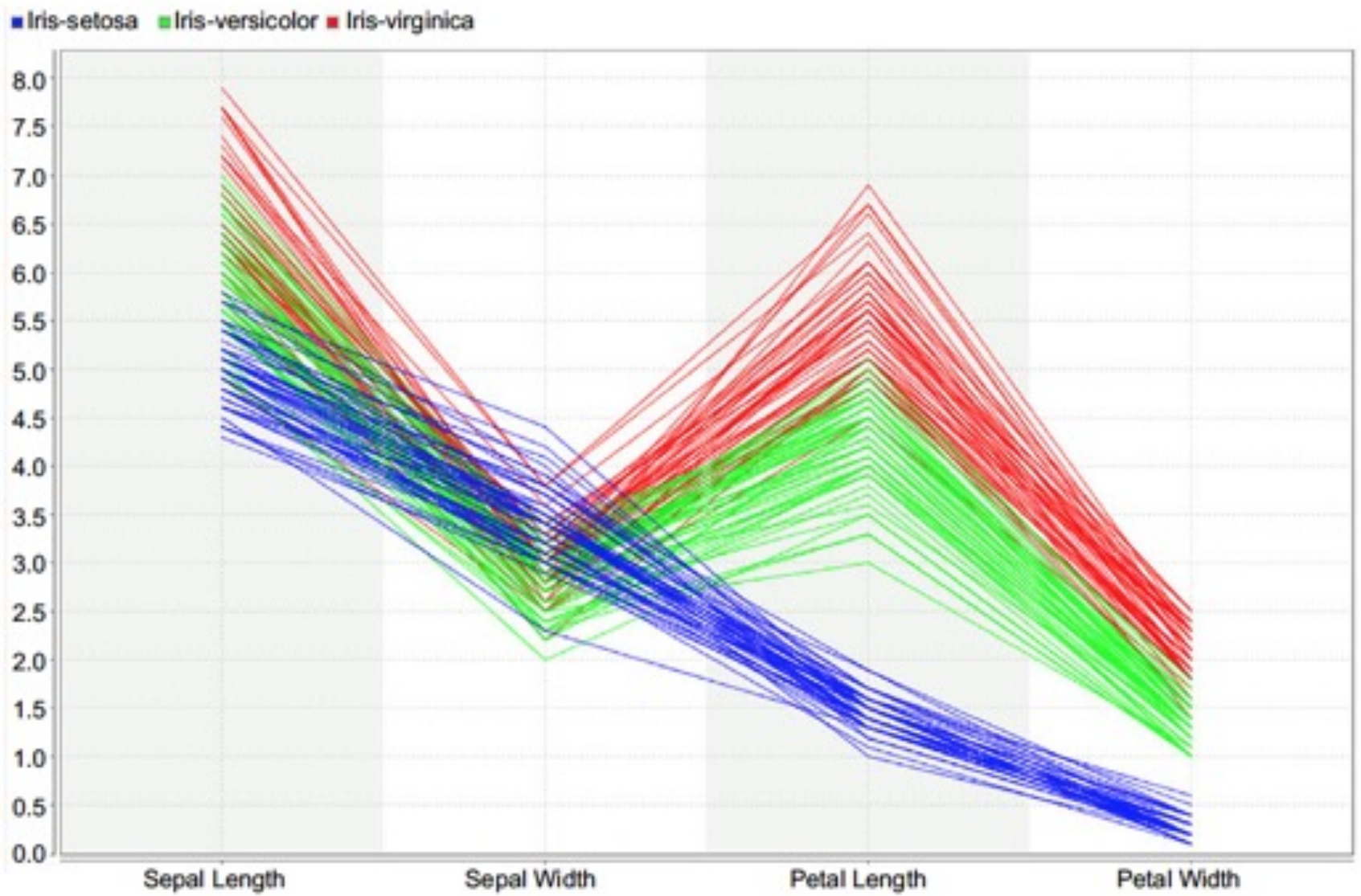
Bubble plot


Iris-setosa
Iris-versicolor
Iris-virginica
Iris virginica
Iris versicolor
Iris setosa
Petal Width
Petal Length
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.0 2.1 2.2 2.3 2.4 2.5 2.6
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5 6.0 6.5 7.0 7.5 8.0 8.5

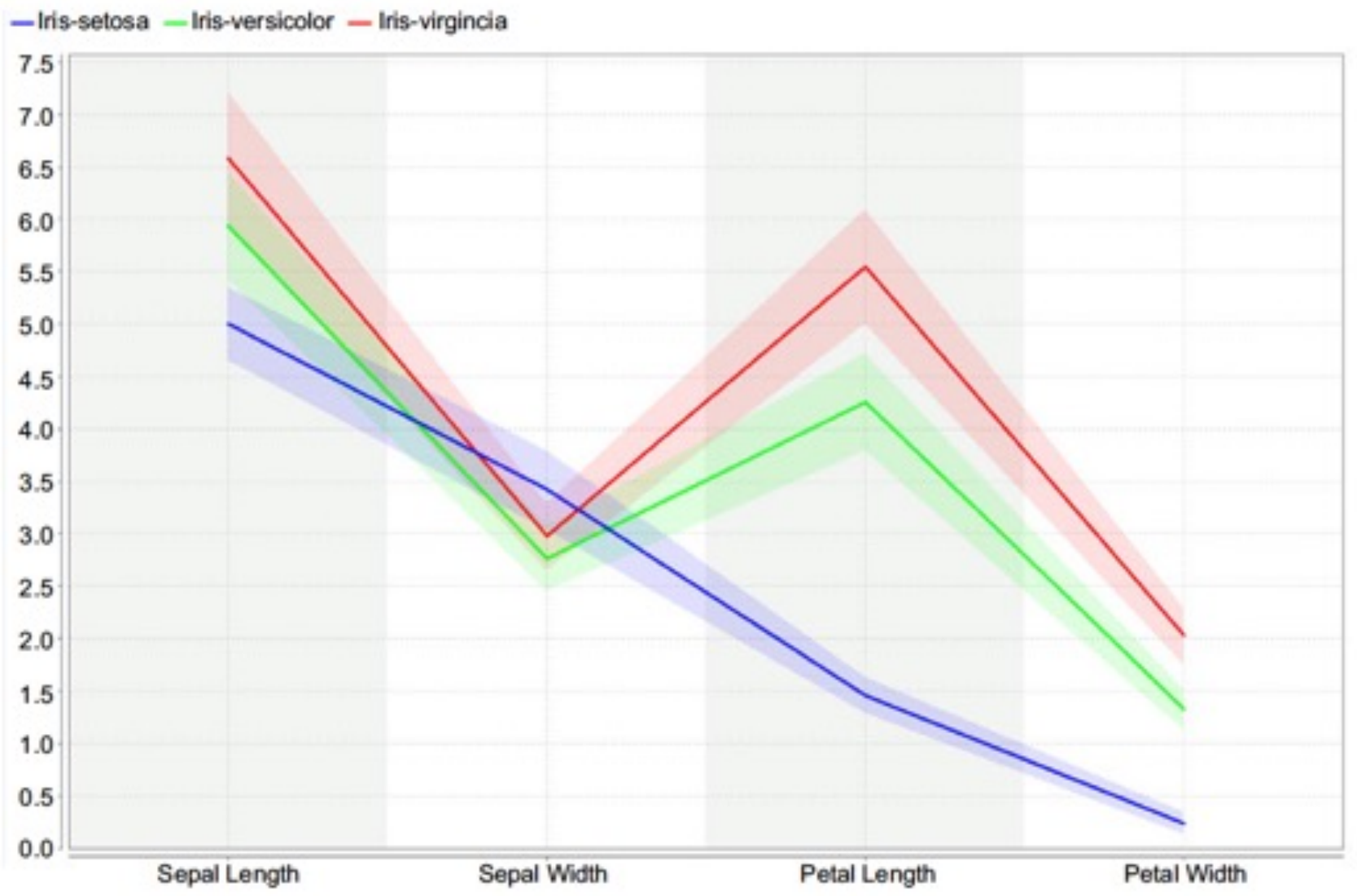
Density Chart



Parallel Chart

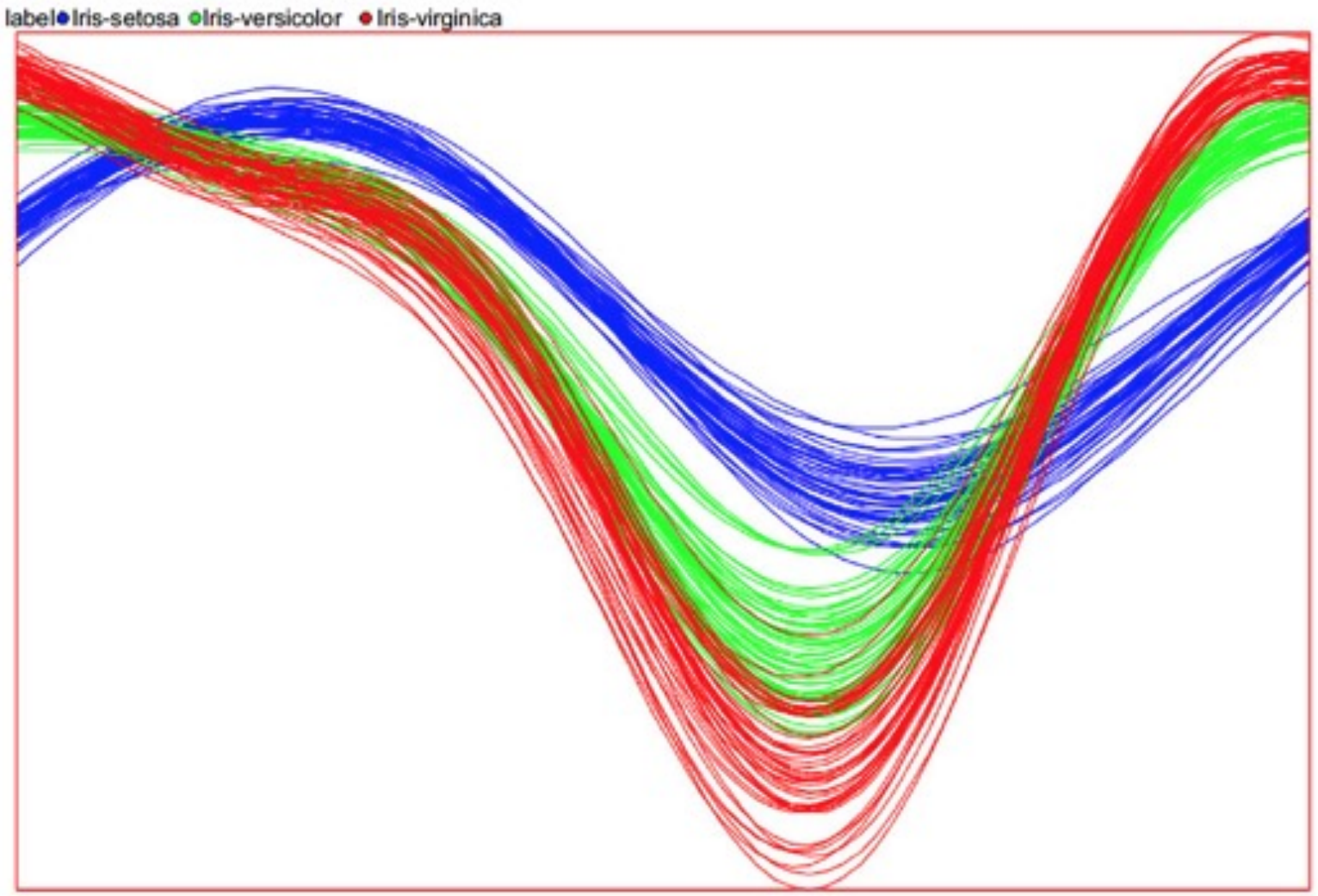


Deviation Chart



Andrews Curve

$$f_x(t) = \frac{x_1}{\sqrt{2}} + x_2 \sin(t) + x_3 \cos(t) + x_4 \sin(2t) + x_5 \cos(2t) + \cdots$$



แนวทางการสำรวจข้อมูล

- จัดรูปแบบข้อมูล
- หาค่าแนวโน้มสู่ศูนย์กลางของข้อมูล
- ทำความเข้าใจการกระจายของข้อมูล
- สร้างแผนภาพการกระจายข้อมูลของแอทริบิวต์แต่ละตัว
- ตรวจสอบค่าที่เป็นไปได้แต่ละค่าของแต่ละแอทริบิวต์
- ตรวจสอบค่าผิดปกติ
- วิเคราะห์ความสัมพันธ์ระหว่างแอทริบิวต์แต่ละตัว