

# What is today's agenda?

Today we are going to learn following things :

- Data Visualization

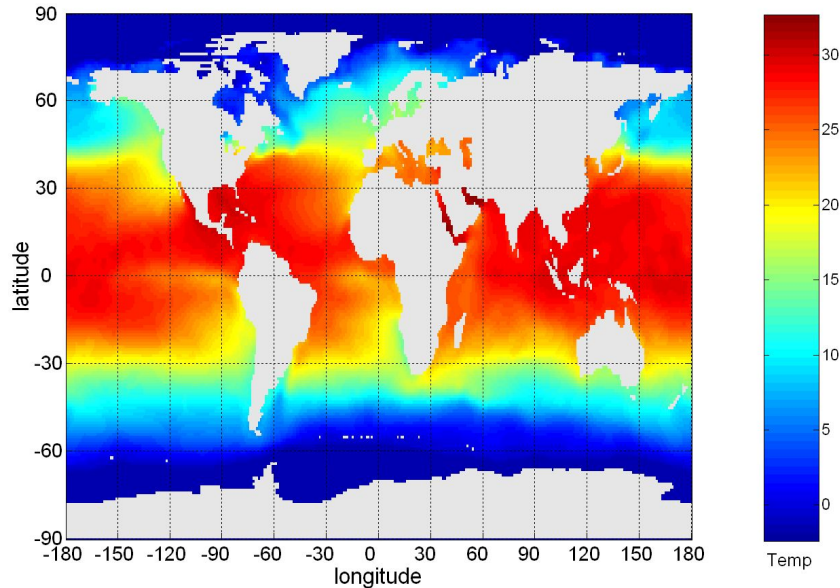
# Visualization

Visualization is the conversion of data into a visual or tabular format so that the characteristics of the data and the relationships among data items or attributes can be analyzed or reported.

- Visualization of data is one of the most powerful and appealing techniques for data exploration.
  - Humans have a well developed ability to analyze large amounts of information that is presented visually
  - Can detect general patterns and trends
  - Can detect outliers and unusual patterns

# Example : Sea Surface Temperature

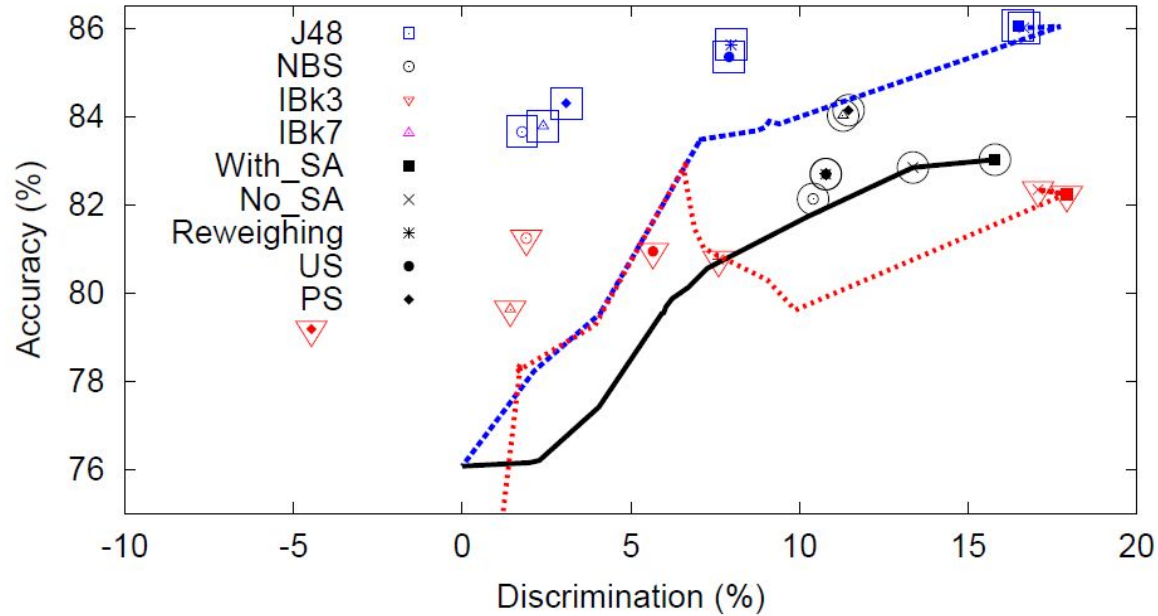
- The following shows the Sea Surface Temperature (SST) for July 1982
  - Tens of thousands of data points are summarized in a single figure



# Representation

- Is the mapping of information to a visual format
- Data objects, their attributes, and the relationships among data objects are translated into graphical elements such as points, lines, shapes, and colors.
- Example:
  - Objects are often represented as points
  - Their attribute values can be represented as the position of the points or the characteristics of the points, e.g., color, size, and shape
  - If position is used, then the relationships of points, i.e., whether they form groups or a point is an outlier, is easily perceived.

# Representation



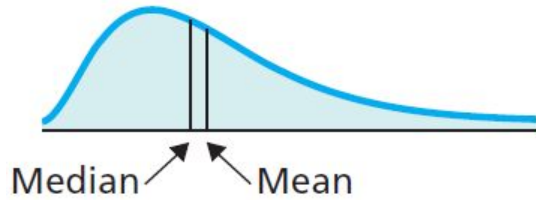
# Arrangement

- Is the placement of visual elements within a display
- Can make a large difference in how easy it is to understand the data
- Example:

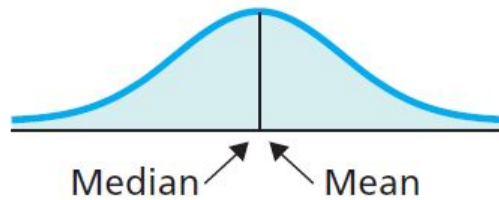
	1	2	3	4	5	6
1	0	1	0	1	1	0
2	1	0	1	0	0	1
3	0	1	0	1	1	0
4	1	0	1	0	0	1
5	0	1	0	1	1	0
6	1	0	1	0	0	1
7	0	1	0	1	1	0
8	1	0	1	0	0	1
9	0	1	0	1	1	0

	6	1	3	2	5	4
4	1	1	1	0	0	0
2	1	1	1	0	0	0
6	1	1	1	0	0	0
8	1	1	1	0	0	0
5	0	0	0	1	1	1
3	0	0	0	1	1	1
9	0	0	0	1	1	1
1	0	0	0	1	1	1
7	0	0	0	1	1	1

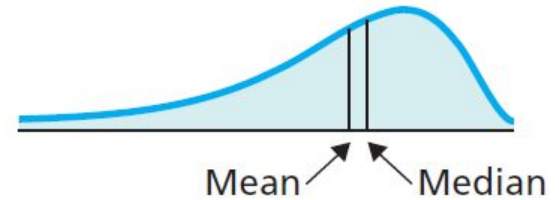
# Data Distribution Shapes



(a) Right skewed



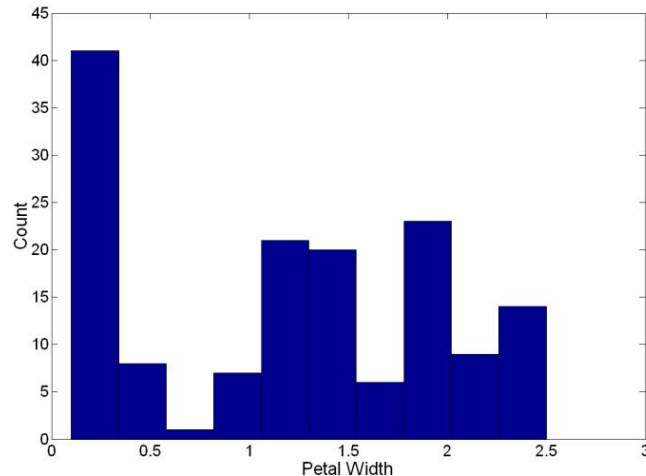
(b) Symmetric



(c) Left skewed

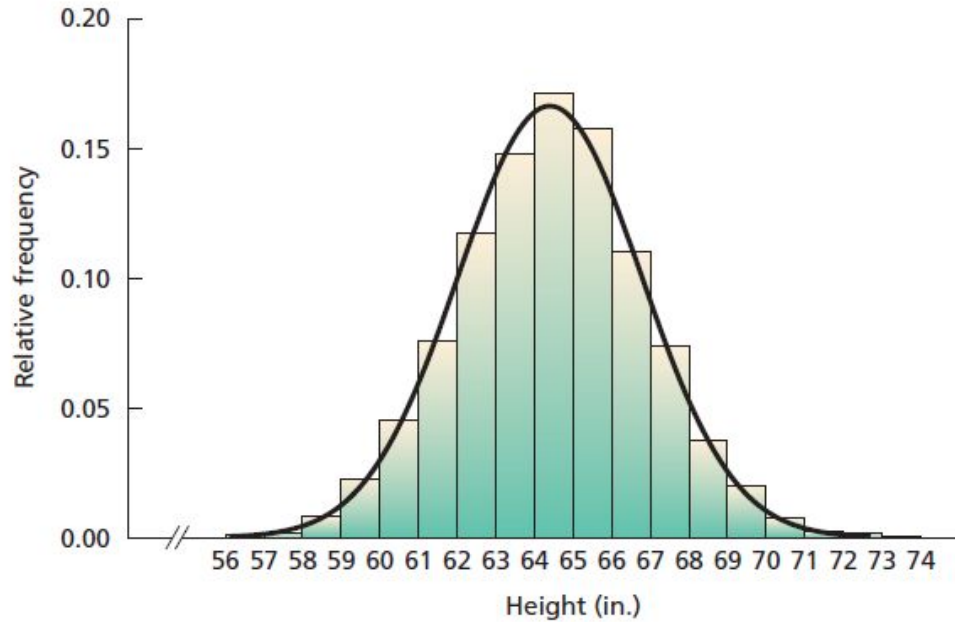
# Visualization Techniques : Histograms

- Histogram
  - Usually shows the distribution of values of a single variable
  - Divide the values into bins and show a bar plot of the number of objects in each bin.
  - The height of each bar indicates the number of objects
  - Shape of histogram depends on the number of bins
- Example: Petal Width



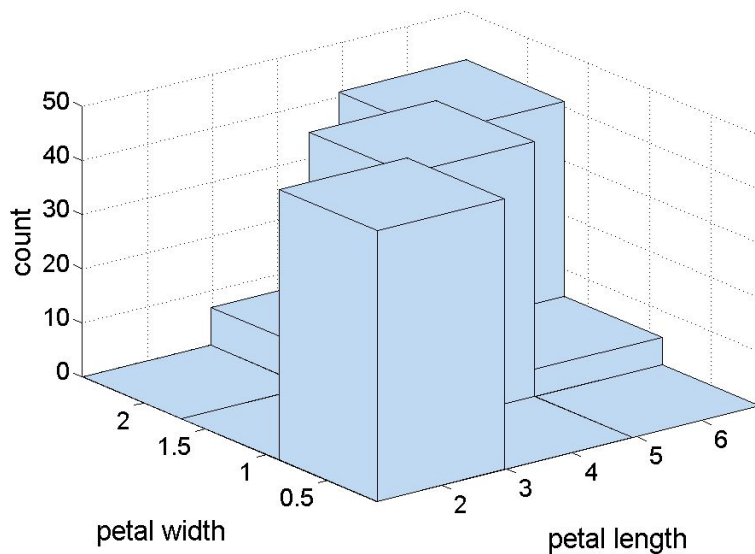


# Visualization Techniques : Histograms



# Two - Dimensional Histograms

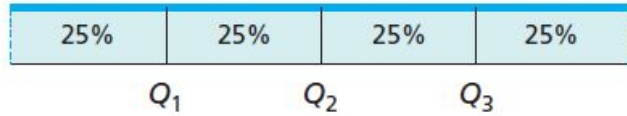
- Show the joint distribution of the values of two attributes
- Example: petal width and petal length
  - What does this tell us?



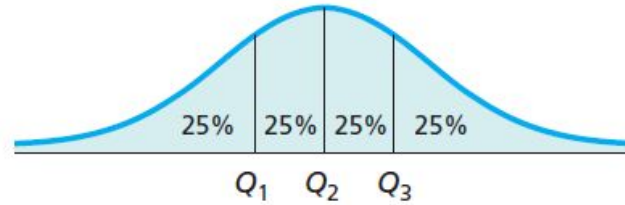
# Visualization Techniques : Quartiles

- **Percentile:** divides the data into hundredths (100 equal parts)  $P_1, P_2, \dots, P_{99}$
- **Deciles:** divides the data into tenths (10 equal parts)
- **Quintiles:** divides the data into fifths (5 equal parts)
- **Quartiles:** divides the data into quarters (4 equal parts)  $Q_1, Q_2, Q_3$

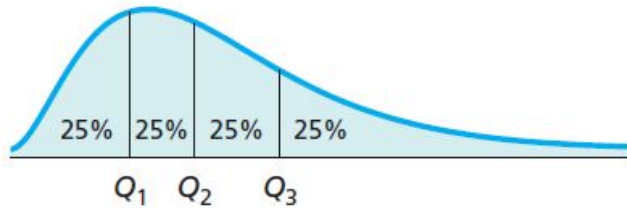
# Visualization Techniques : Quartiles



(a) Uniform



(b) Bell shaped



(c) Right skewed



(d) Left skewed

# Five Number Summary

- **Five-Number Summary:** min,  $Q_1$ ,  $Q_2$ ,  $Q_3$ , Max
- **Interquartile range (IQR):**  
$$IQR = Q_3 - Q_1$$
- **Limits of the dataset:**
  - Lower limit =  $Q_1 - 1.5 \times IQR$
  - Upper limit =  $Q_3 + 1.5 \times IQR$
- **Outliers:** The objects below the lower limit and above the upper limit are potential outliers.

# Five Number Summary

Find the 5 Number Summary of the following numbers:

3    12    7    40    9    14    18    15    17

# Five Number Summary

Find the 5 Number Summary of the following numbers:

3    12    7    40    9    14    18    15    17

- **Step 1:** Sort the numbers from lowest to highest

# Five Number Summary

Find the 5 Number Summary of the following numbers:

3    12    7    40    9    14    18    15    17

- **Step 1:** Sort the numbers from lowest to highest

3    7    9    12    14    15    17    18    40



# Five Number Summary

Find the 5 Number Summary of the following numbers:

3    12    7    40    9    14    18    15    17

- **Step 1:** Sort the numbers from lowest to highest

3    7    9    12    14    15    17    18    40

- **Step 2:** Identify the Median

# Five Number Summary

Find the 5 Number Summary of the following numbers:

3    12    7    40    9    14    18    15    17

- **Step 1:** Sort the numbers from lowest to highest

3    7    9    12    14    15    17    18    40

- **Step 2:** Identify the Median

3    7    9    12    14    15    17    18    40

# Five Number Summary

Find the 5 Number Summary of the following numbers:

3    12    7    40    9    14    18    15    17

- **Step 1:** Sort the numbers from lowest to highest

3    7    9    12    14    15    17    18    40

- **Step 2:** Identify the Median

3    7    9    12    14    15    17    18    40

- **Step 3:** Identify the Smallest and Largest numbers

# Five Number Summary

Find the 5 Number Summary of the following numbers:

3    12    7    40    9    14    18    15    17

- **Step 1:** Sort the numbers from lowest to highest

3    7    9    12    14    15    17    18    40

- **Step 2:** Identify the Median

3    7    9    12    14    15    17    18    40

- **Step 3:** Identify the Smallest and Largest numbers

3    7    9    12    14    15    17    18    40

# Five Number Summary

Find the 5 Number Summary of the following numbers:

3    12    7    40    9    14    18    15    17

- **Step 1:** Sort the numbers from lowest to highest

3    7    9    12    14    15    17    18    40

- **Step 2:** Identify the Median

3    7    9    12    14    15    17    18    40

- **Step 3:** Identify the Smallest and Largest numbers

3    7    9    12    14    15    17    18    40

- **Step 4:** Identify the Median between the smallest number and the Median for the entire set of data, and between that Median and the largest number in the set.

# Five Number Summary

Find the 5 Number Summary of the following numbers:

3    12    7    40    9    14    18    15    17

- **Step 1:** Sort the numbers from lowest to highest

3    7    9    12    14    15    17    18    40

- **Step 2:** Identify the Median

3    7    9    12    14    15    17    18    40

- **Step 3:** Identify the Smallest and Largest numbers

3    7    9    12    14    15    17    18    40

- **Step 4:** Identify the Median between the smallest number and the Median for the entire set of data, and between that Median and the largest number in the set.

3    7    9    12    14    15    17    18    40

# Five Number Summary

These are the five numbers in the 5 Number Summary

3    7    9    12    14    15    17    18    40

3 - Smallest number in the set

9 - Median between the smallest number  
and the median

14 - Median of the entire set

17 - Median between the largest number  
and the median

40 - Largest number in the set

# Five Number Summary

42
16
38
50
24
29
41
36
18
4
33
37
24
27
45



# Five Number Summary

42
16
38
50
24
29
41
36
18
4
33
37
24
27
45

4
16
18
24
24
27
29
33
36
37
38
41
42
45
50

# Five Number Summary

42
16
38
50
24
29
41
36
18
4
33
37
24
27
45

4
16
18
24
24
27
29
33
36
37
38
41
42
45
50



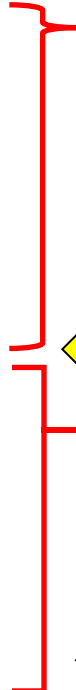
Median

33

# Five Number Summary

42
16
38
50
24
29
41
36
18
4
33
37
24
27
45

4
16
18
24
24
27
29
33
36
37
38
41
42
45
50



Smallest

4

Median

33

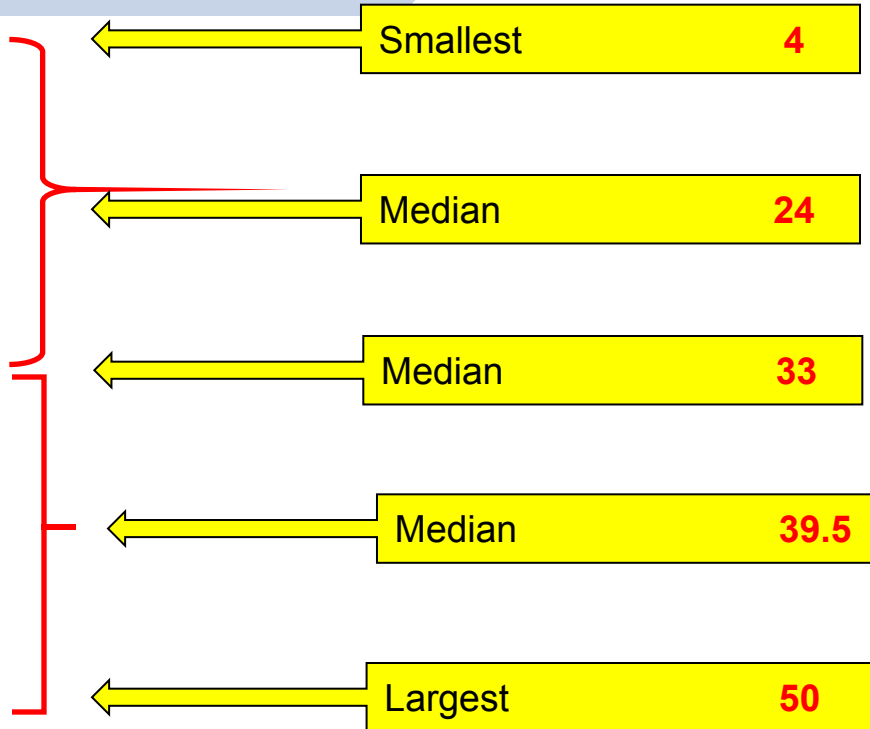
Largest

50

# Five Number Summary

42
16
38
50
24
29
41
36
18
4
33
37
24
27
45

4
16
18
24
24
27
29
33
36
37
38
41
42
45
50



# Five Number Summary

4
8
2
19
11
6
21
13
5
7
10
20
14
15
18
3



**Smallest**

**?**



**Median**

**?**



**Median**

**?**



**Median**

**?**



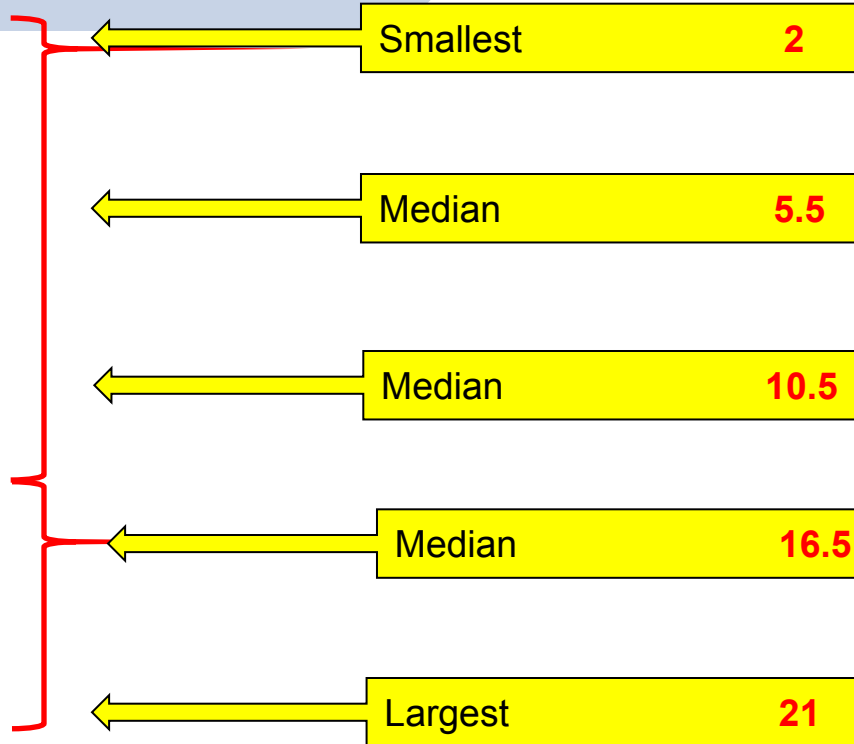
**Largest**

**?**

# Five Number Summary

4
8
2
19
11
6
21
13
5
7
10
20
14
15
18
3


2
3
4
5
6
7
8
10
11
13
14
15
18
19
20
21



## Five Number Summary

A 5 Number Summary divides your data into four quarters.

3      7      9                      12      14                                      15      17    18      40



40

# Five Number Summary

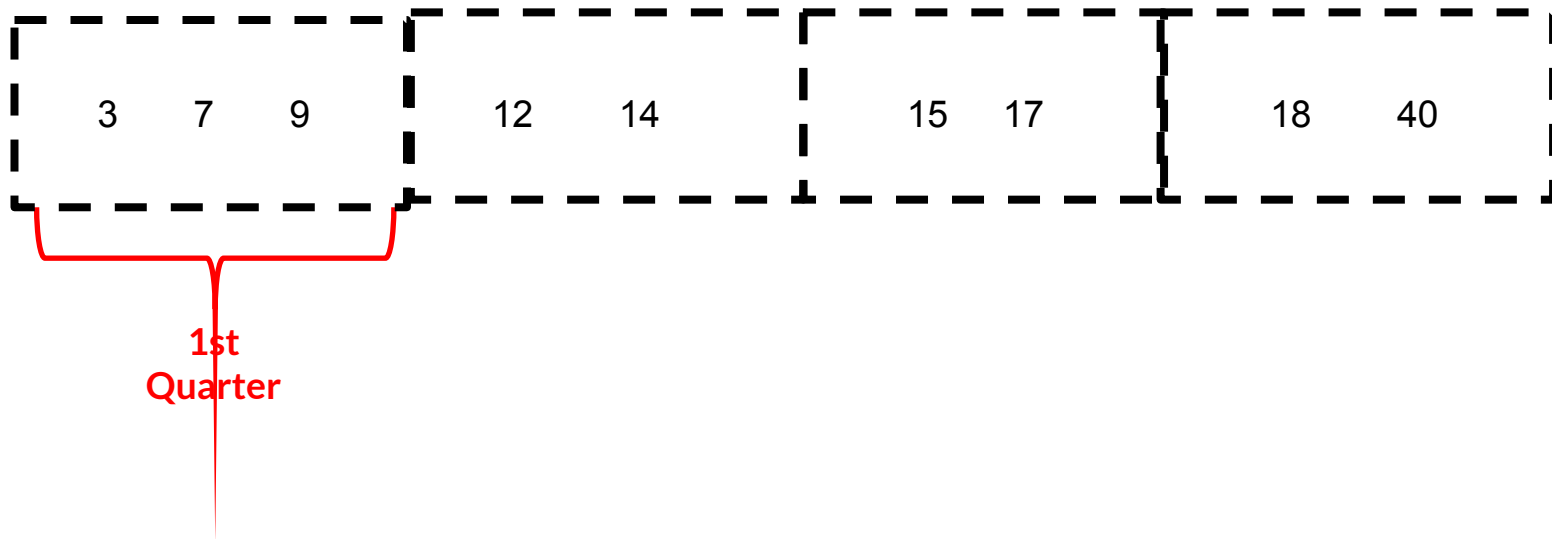
A 5 Number Summary divides your data into four quarters.

3	7	9	12	14	15	17	18	40
---	---	---	----	----	----	----	----	----



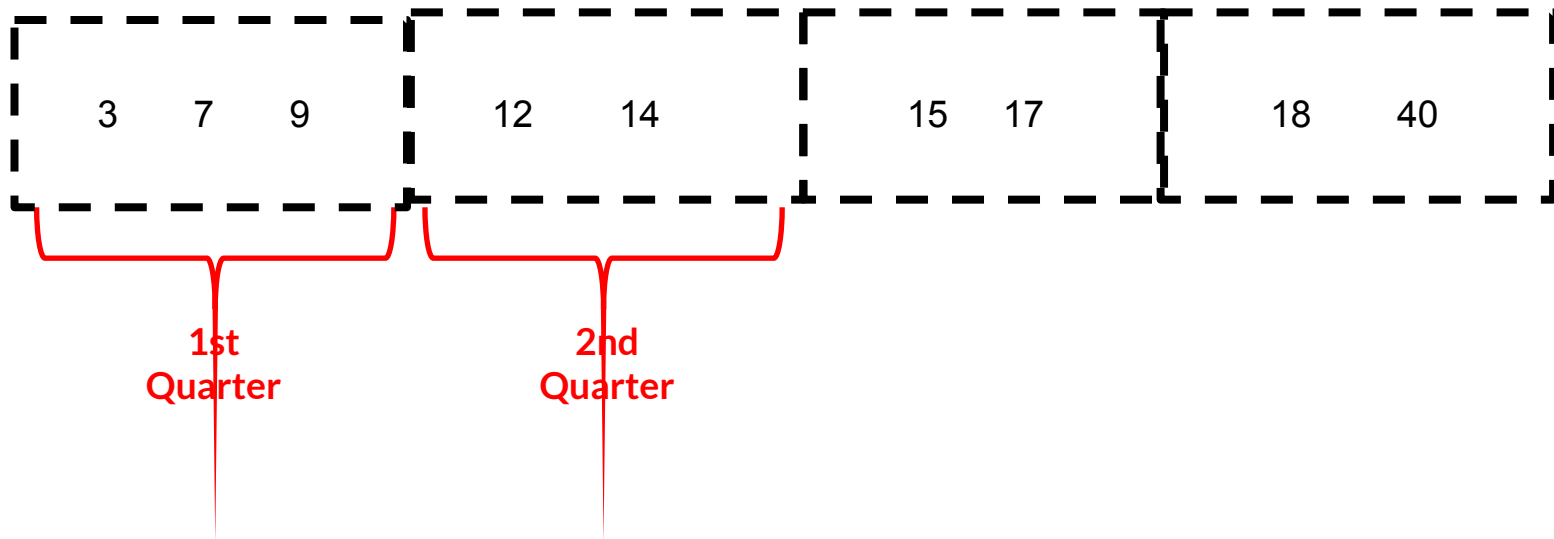
# Five Number Summary

A 5 Number Summary divides your data into four quarters.



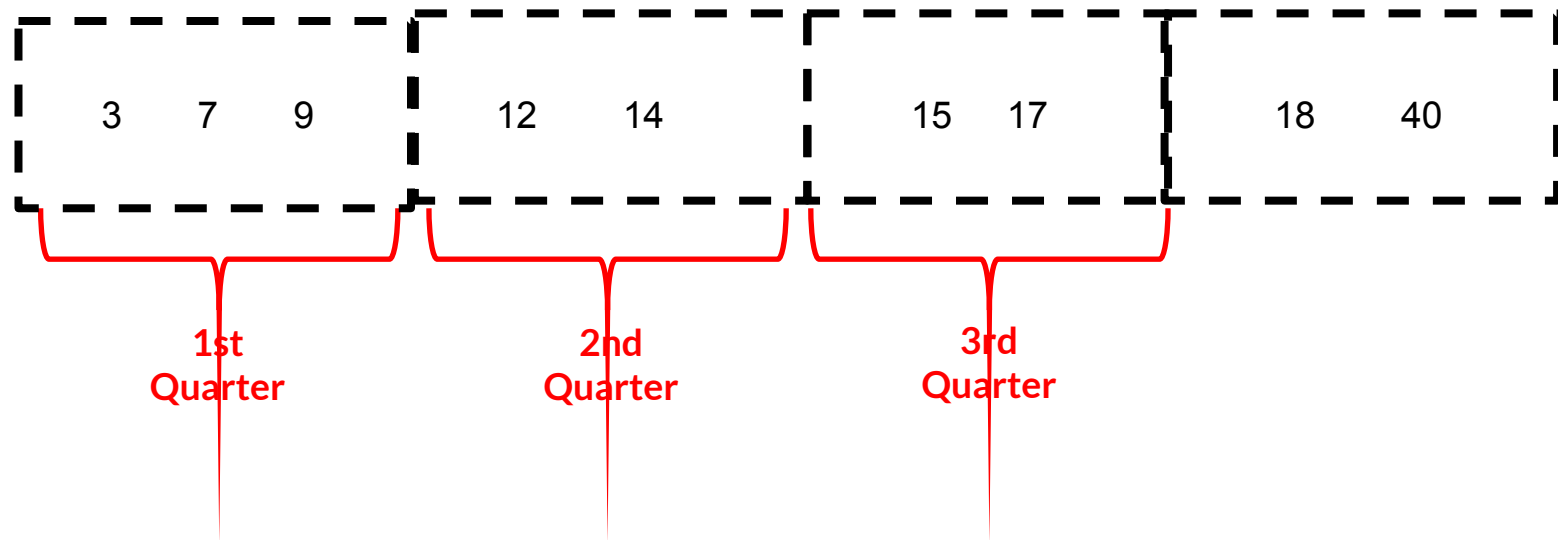
# Five Number Summary

A 5 Number Summary divides your data into four quarters.



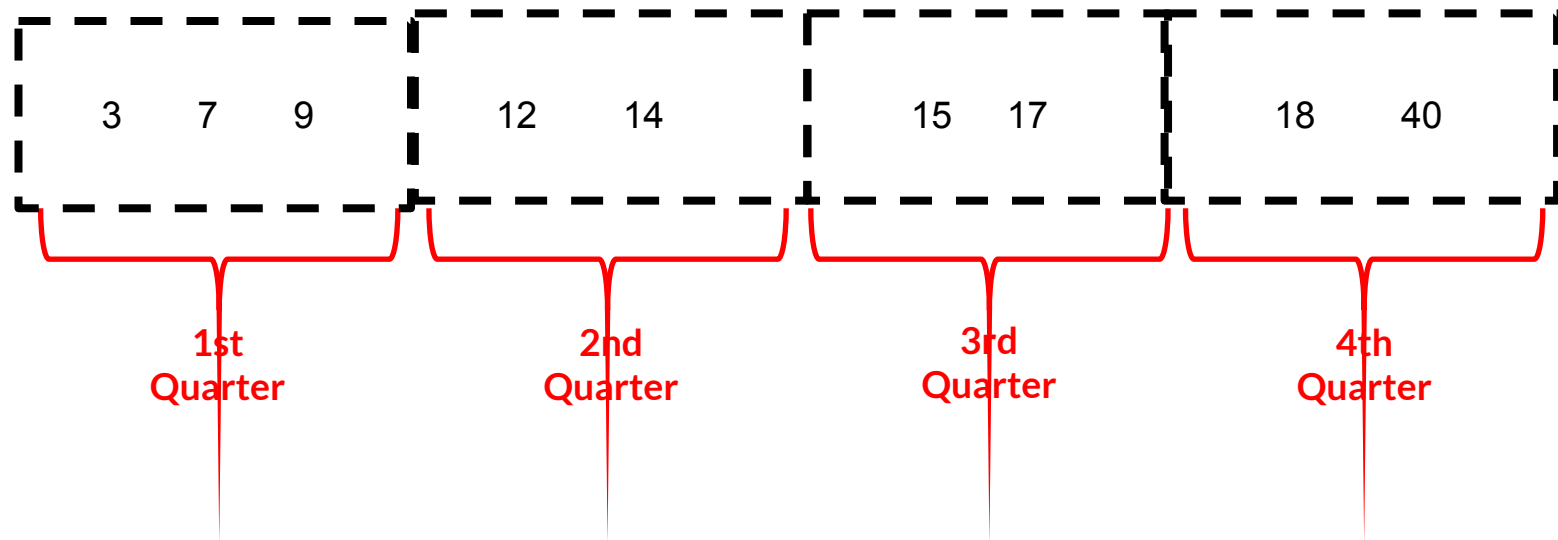
# Five Number Summary

A 5 Number Summary divides your data into four quarters.



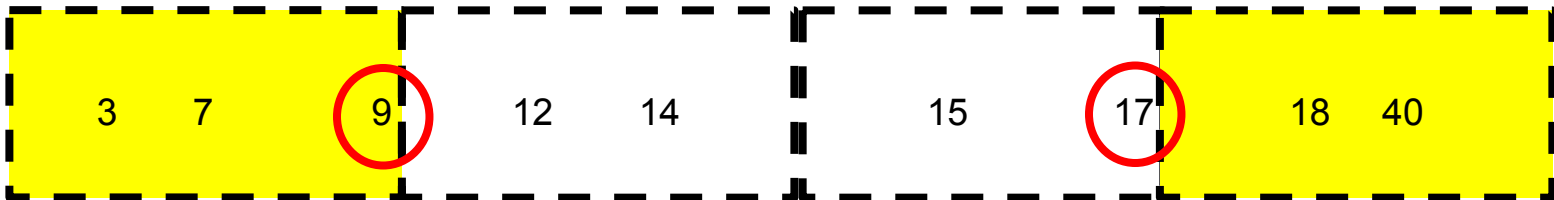
# Five Number Summary

A 5 Number Summary divides your data into four quarters.



# InterQuartile Range

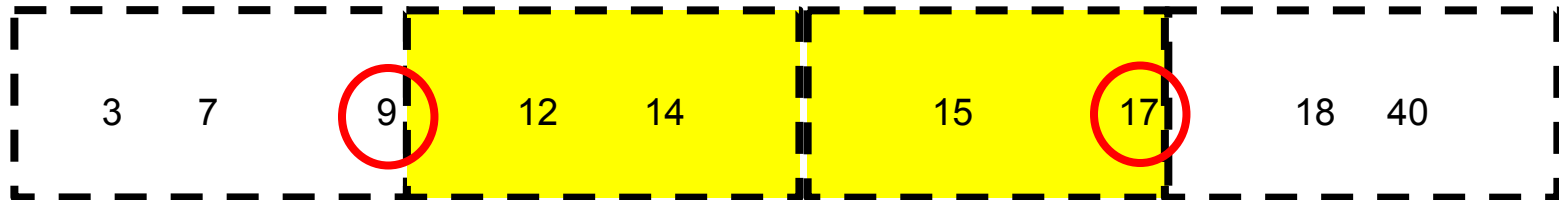
- The **Lower Quartile (Q1)** is the second number in the 5 Number Summary
  - 25% of all the numbers in the set are smaller than Q1



- The **Upper Quartile (Q3)** is the fourth number in the 5 Number Summary
  - 25% of all the numbers in the set are larger than Q3

# InterQuartile Range

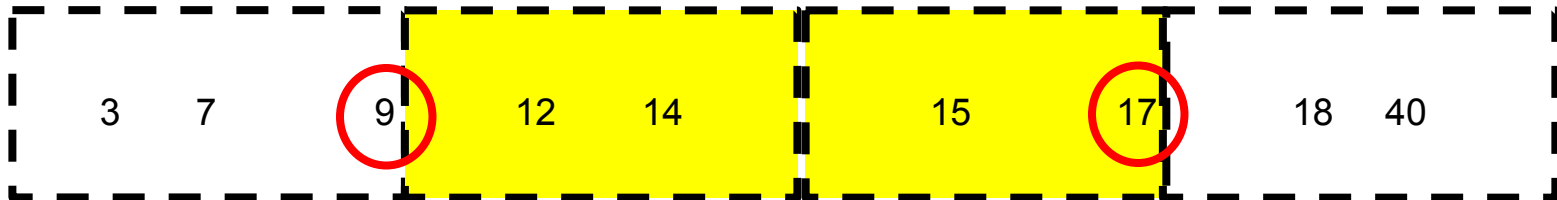
- What percent of all the numbers are between Q1 and Q3?
  - 50% of all the numbers are between Q1 and Q3



- This is called the Inter-Quartile Range (IQR)
  - The size of the IQR is the distance between Q1 and Q3
  - $17 - 9 = 8$

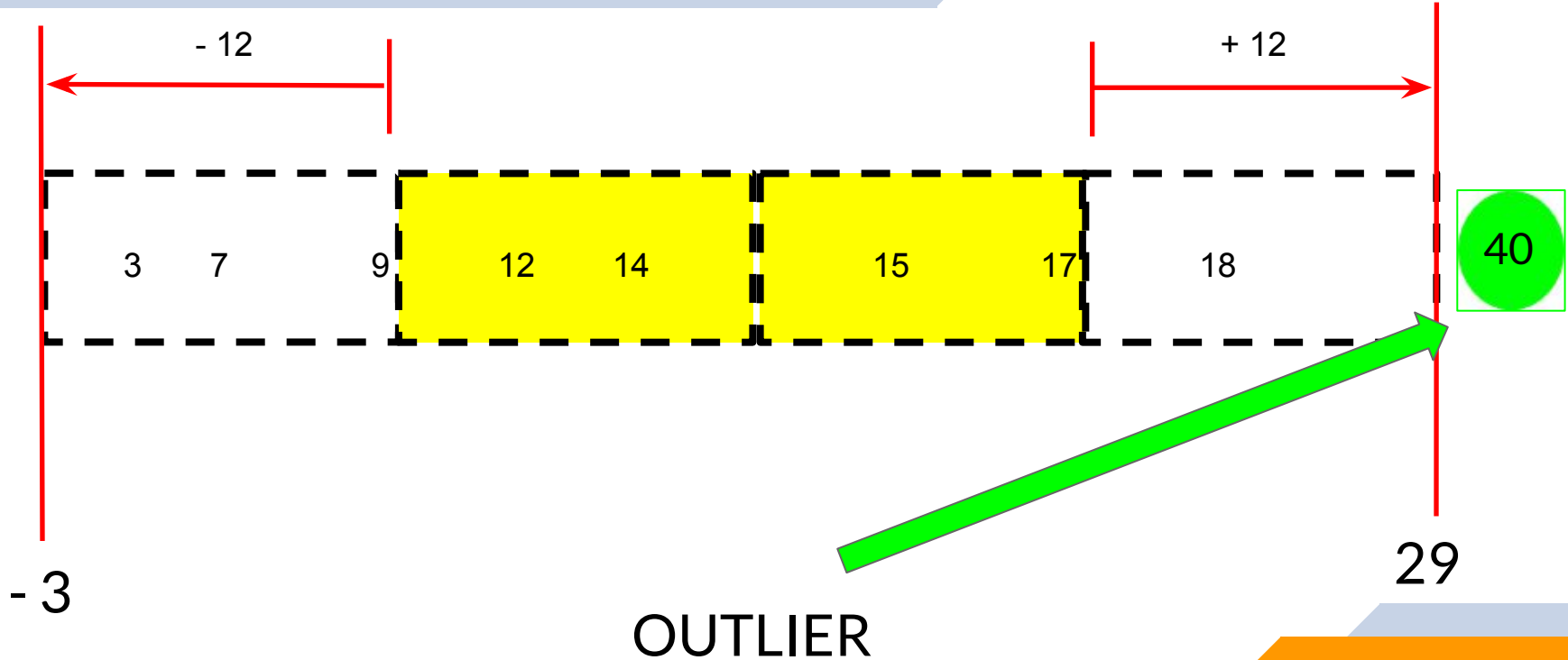
# Outlier Detection Using IQR

- To determine if a number is an outlier, multiply the IQR by 1.5
  - $8 \bullet 1.5 = 12$  where 8 is IQR



- An outlier is any number that is 12 less than Q1 or 12 more than Q3

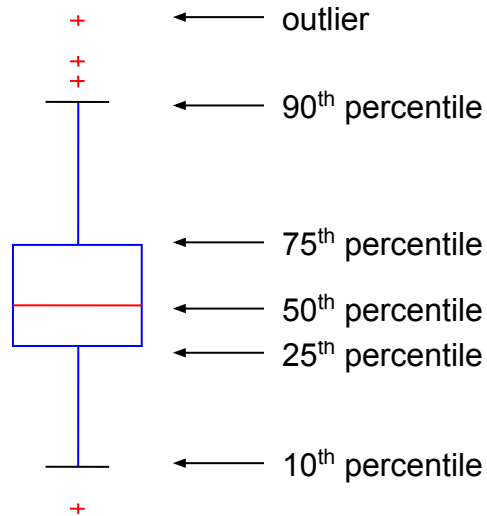
# Outlier Detection Using IQR





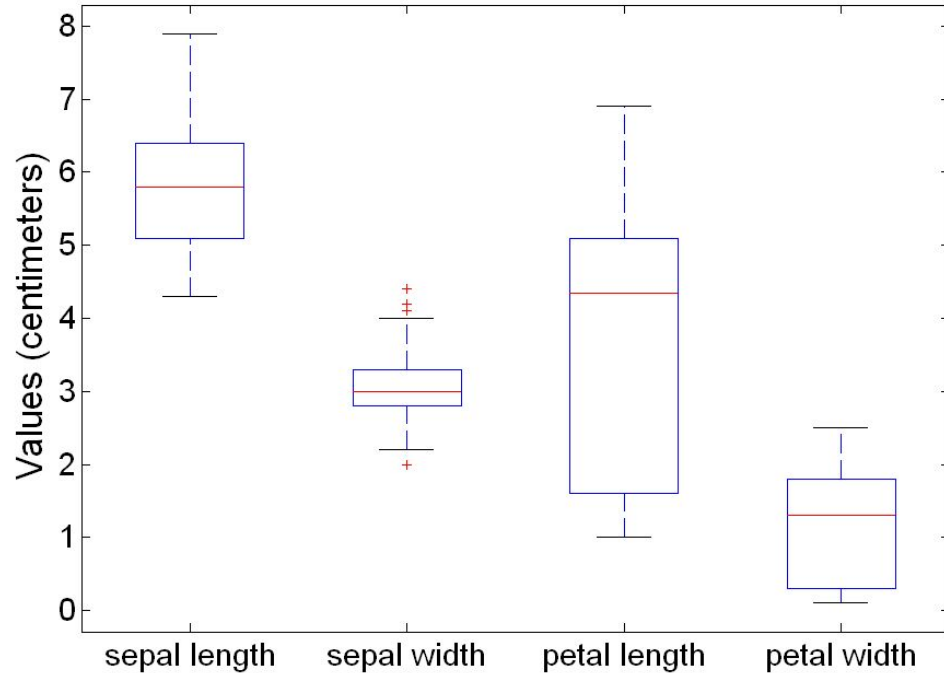
# Visualization Techniques : Box Plots

- Box Plots
  - Invented by J. Tukey
  - Another way of displaying the distribution of data
  - Following figure shows the basic part of a box plot

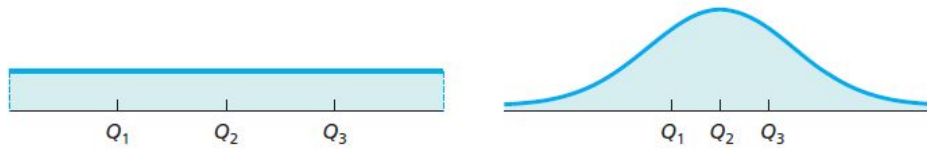


# Example of Box Plots

- Box plots can be used to compare attributes

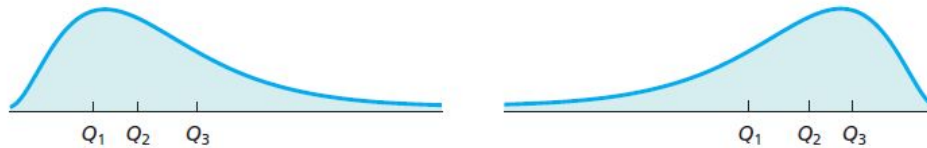


# Comparing Data By Box Plots



(a) Uniform

(b) Bell shaped



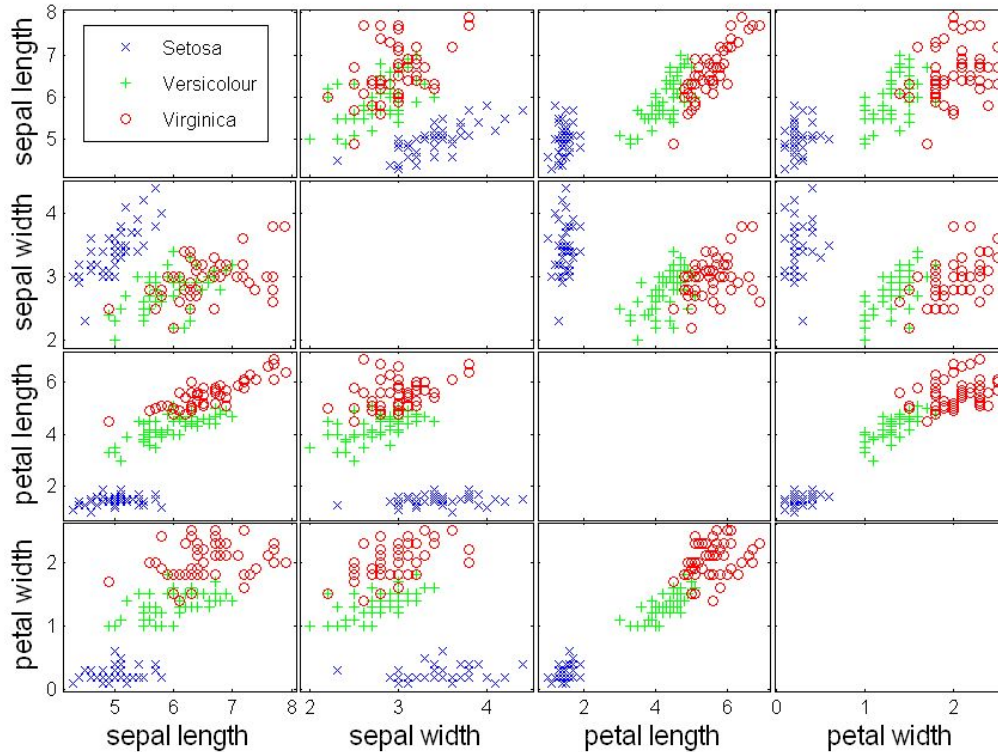
(c) Right skewed

(d) Left skewed

# Visualization Techniques : Scatter Plots

- Scatter plots
  - Attributes values determine the position
  - Two-dimensional scatter plots most common, but can have three-dimensional scatter plots
  - Often additional attributes can be displayed by using the size, shape, and color of the markers that represent the objects
  - It is useful to have arrays of scatter plots can compactly summarize the relationships of several pairs of attributes
  - ◆ See example on the next slide

# Scatter Plot of Iris Attributes



“

*Questions ?*