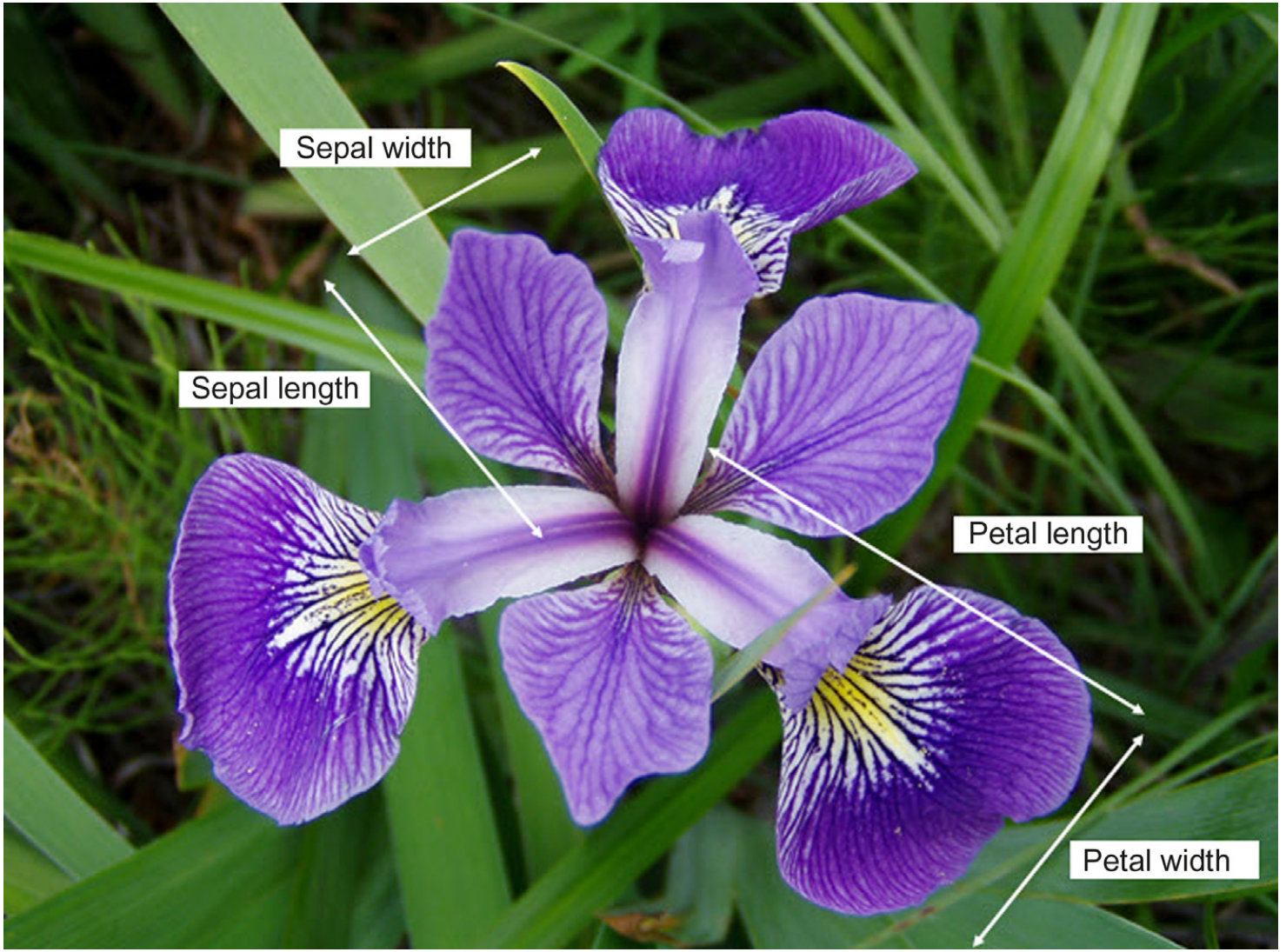
1.DATA SETS

- The most popular datasets used to learn data science is probably the Iris dataset, introduced by Ronald Fisher, in his seminal work on discriminant analysis,
- “The use of multiple measurements in taxonomic problems” (Fisher, 1936)

- The Iris dataset contains 150 observations of three different species
- Iris setosa, Iris virginica, and I. versicolor, with 50 observations each.
- Each observation consists of four attributes: sepal length, sepal width, petal length, and petal width.
- The fifth attribute, the label, is the name of the species observed.


Sepal width
Sepal length
Petal length
Petal width

1.1 Types of Data

- Data come in different formats and types
- For example, the temperature in weather data can be expressed as any of the following formats:
  - Numeric centigrade (31 C, 33.3 C) or Fahrenheit (100 F, 101.45 F) or on the Kelvin scale
  - Ordered labels as in hot, mild, or cold
  - Number of days within a year below 0 C (10 days in a year below freezing)

Types of Data Contd...

- Numerical or Continous

  Continuous values can be denoted by numbers and take an infinite number of values between digits.

  - An integer is a special form of the numeric data type which does not have decimals in the value or more precisely does not have infinite values between consecutive numbers.
  - If a zero point is defined, numeric data become a ratio or real data type.Examples include temperature in Kelvin scale, bank account balance, and income.

Categorical or Nominal

- Categorical data types are attributes treated as distinct symbols or just names. The color of the iris of the human eye is a categorical data type because it takes a value like black, green, blue, gray, etc.
- An ordered nominal data type is a special case of a categorical data type where there is some kind of order among the values.
- An example of an ordered data type is temperature expressed as hot, mild, cold.

- Not all data science tasks can be performed on all data types.
- For example,the neural network algorithm does not work with categorical data.
- However,one data type can be converted to another using a type conversion process,but this may be accompanied with possible loss of information.
- For example, credit scores expressed in poor, average, good, and excellent categories can be converted to either 1, 2, 3, and 4

2.DESCRIPTIVE STATISTICS

- Some examples of descriptive statistics include average annual income, medium home price in a neighborhood, range of credit scores of a population, etc.
- Descriptive statistics can be broadly classified into univariate and multivariate exploration depending on the number of attributes under analysis.

2.1 Univariate Exploration

- Univariate data exploration denotes analysis of one attribute at a time

•Measure of Central Tendency

- The objective of finding the central location of an attribute is to quantify the dataset with one central or most common number.

  Mean: The mean is the arithmetic average of all observations in the dataset. It is calculated by summing all the data points and dividing by the number of data points.

- Median: The median is the value of the central point in the distribution.The median is calculated by sorting all the observations from small to large and selecting the mid-point observation in the sorted list. If the number of data points is even, then the average of the middle two data points is used as the median.
- Mode: The mode is the most frequently occurring observation. In the dataset, data points may be repetitive, and the most repetitive data point is the mode of the dataset.

Measure of Spread

- Range: The range is the difference between the maximum value and the minimum value of the attribute.
- The range is simple to calculate and articulate but has shortcomings as it is severely impacted by the presence of outliers and fails to consider the distribution of all other data points in the attributes.
- Deviation: The variance and standard deviation measures the spread, by considering all the values of the attribute.
- Deviation is simply measured as the difference between any given value (x_i) and the mean of the sample (μ). The variance is the sum of the squared deviations of all data points divided by the number of data points.

- Standard deviation is the square root of the variance. Since the standard deviation is measured in the same units as the attribute, it is easy to understand the magnitude of the metric.
- High standard deviation means the data points are spread widely around the central point.
- Low standard deviation means data points are closer to the central point.
- If the distribution of the data aligns with the normal distribution, then 68% of the data points lie within one standard deviation from the mean.

2.2 Multivariate Exploration

- Multivariate exploration is the study of more than one attribute in the dataset simultaneously. This technique is critical to understanding the relationship between the attributes.
- Similar to univariate explorations, the measure of central tendency and variance in the data will be discussed.

2.2.1 Central Data Point

- In the Iris dataset, each data point as a set of all the four attributes can be expressed:observation i: {sepal length, sepal width, petal length, petal width}
- For example, observation one: {5.1, 3.5, 1.4, 0.2}.
- This observation point can also be expressed in four-dimensional Cartesian coordinates and can beplotted in a graph (although plotting more than three dimensions in a visualgraph can be challenging).
- In this way, all 150 observations can be expressed in Cartesian coordinates. If the objective is to find the most “typical” observation point, it would be a data point made up of the mean of each attribute in the dataset independently. For the Iris dataset, the central mean point is {5.006, 3.418, 1.464, 0.244}.
- *This data point may not be an actual observation. It will be a hypothetical data point with the most typical attribute values.*

2.2.2 Correlation

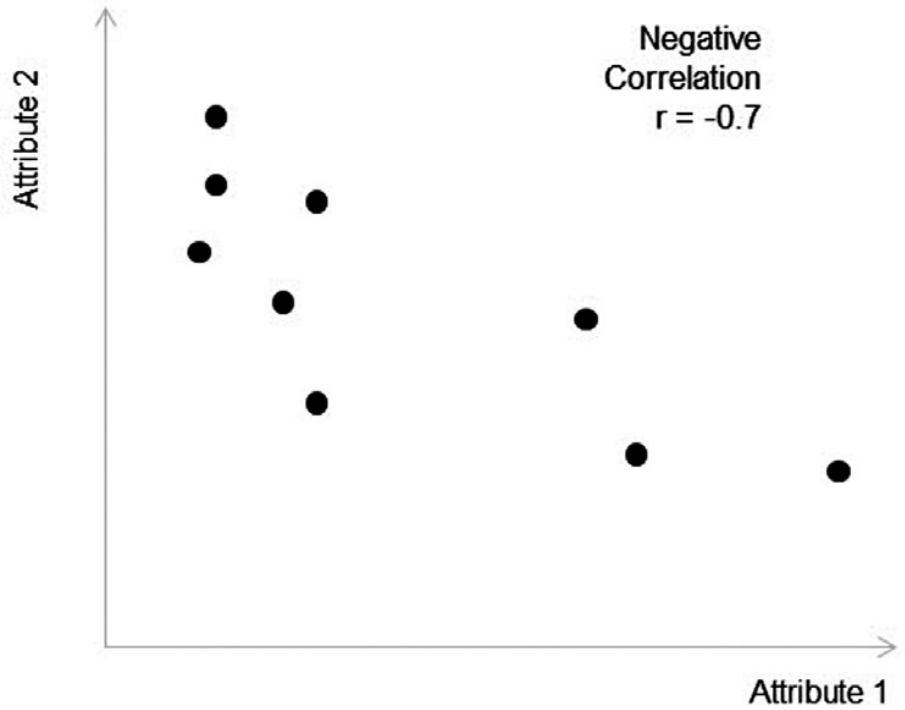
- Correlation measures the statistical relationship between two attributes, particularly dependence of one attribute on another attribute. When two attributes are highly correlated with each other, they both vary at the same rate with each other either in the same or in opposite directions.

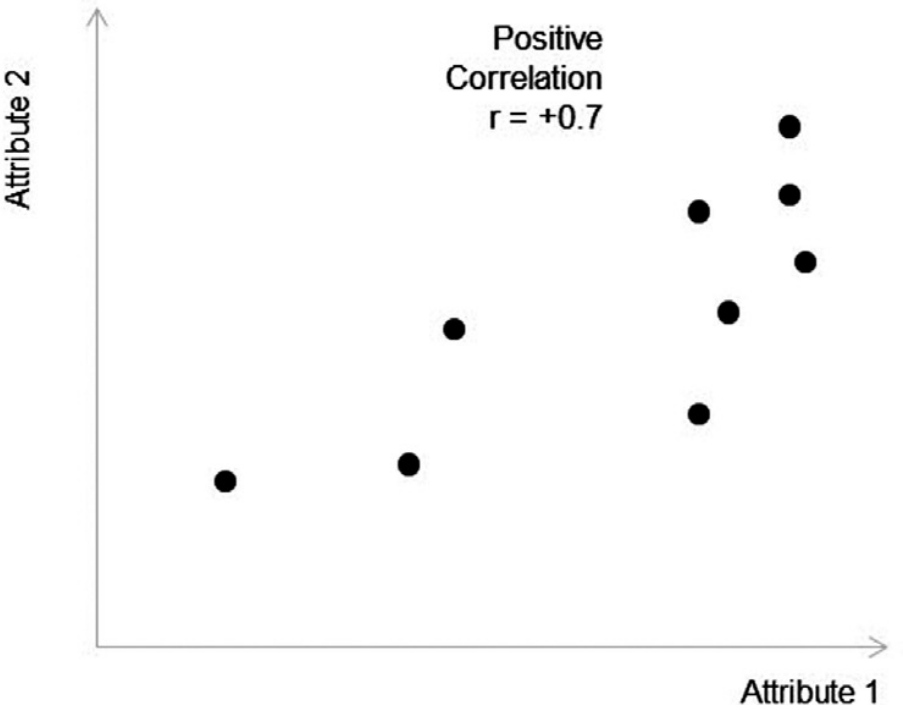
•Pearson correlation coefficient

- Correlation between two attributes is commonly measured by the Pearson correlation coefficient (r), which measures the strength of linear dependence.
- Correlation coefficients take a value from

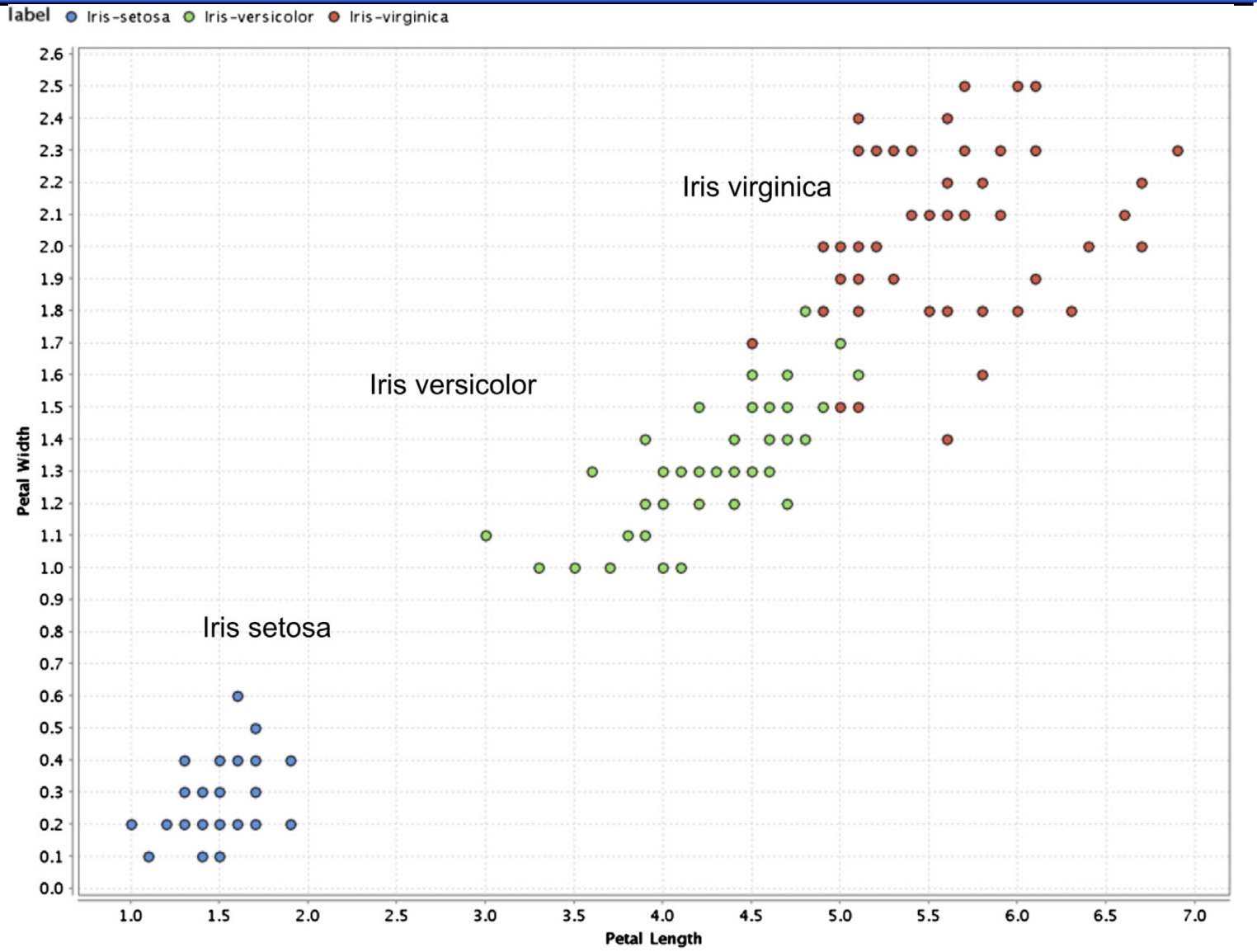
-1 <= r <= 1

- A value closer to 1 or -1 indicates the two attributes are highly correlated, with perfect correlation at 1 or -1.A correlation value of 0 means there is no linear relationship between two attributes

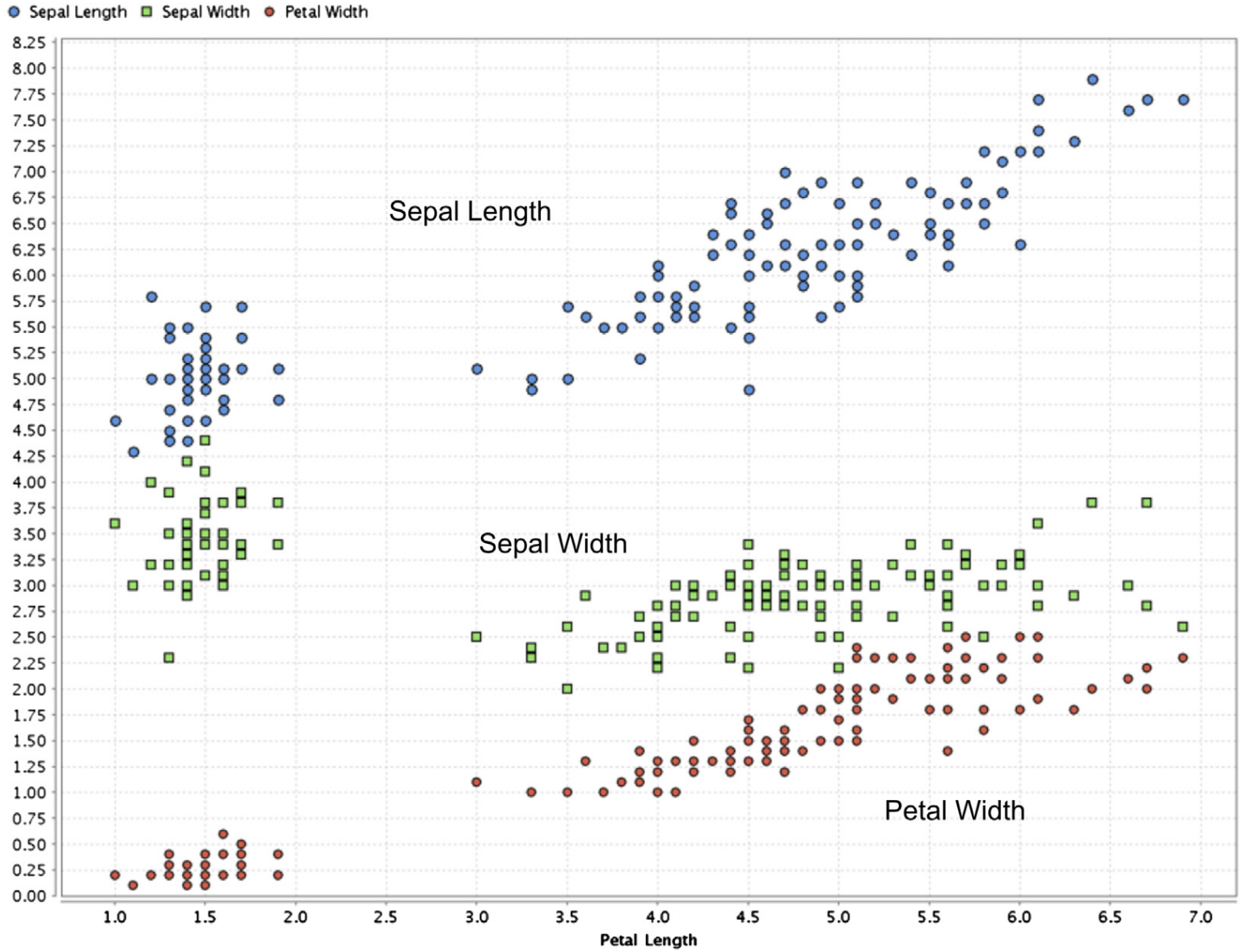

Negative
Correlation
r = -0.7
Attribute 2
Attribute 1


Positive
Correlation
r = +0.7
Attribute 2
Attribute 1

Scatter Plot



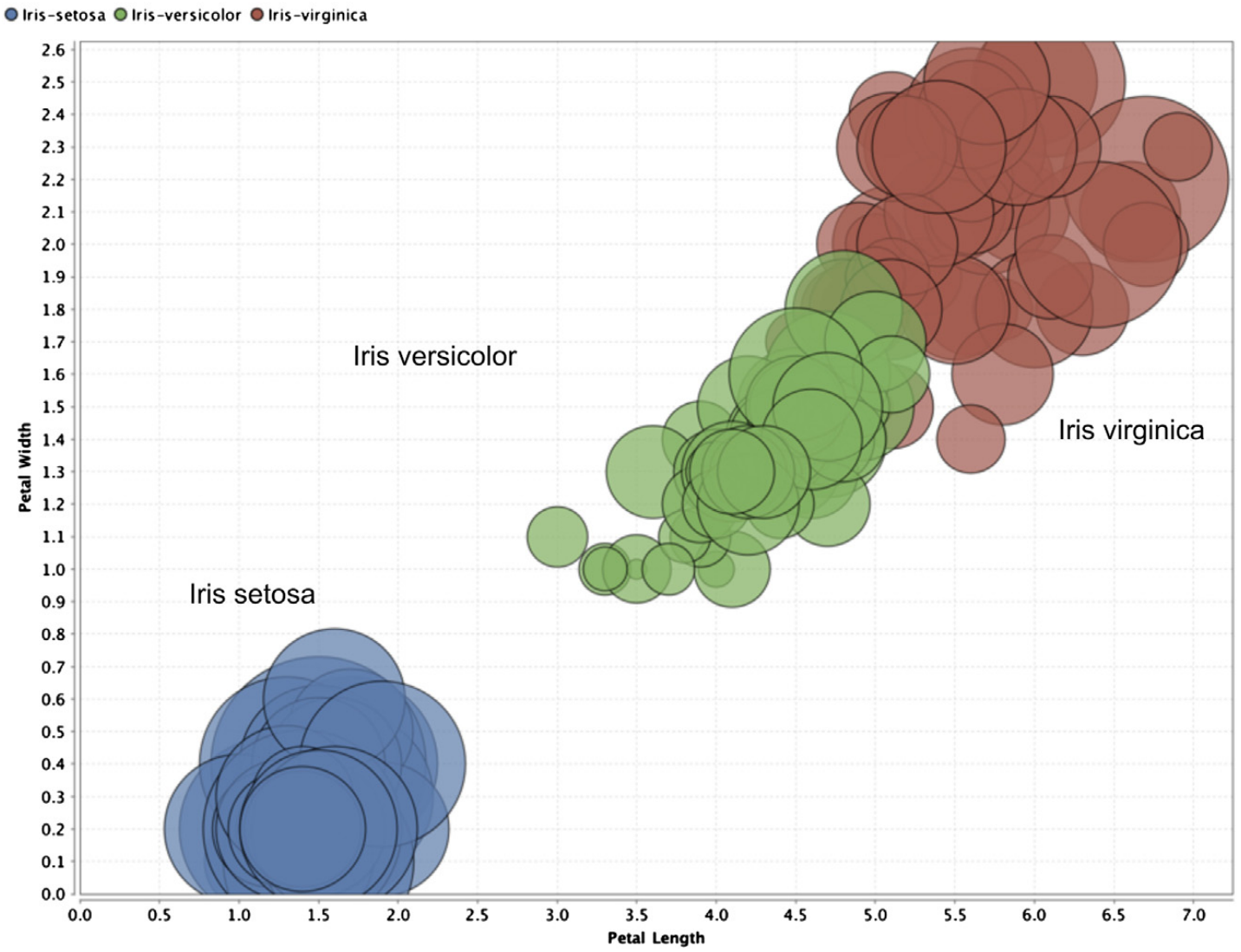
Scatter multiple plot of Iris dataset



•Bubble Chart

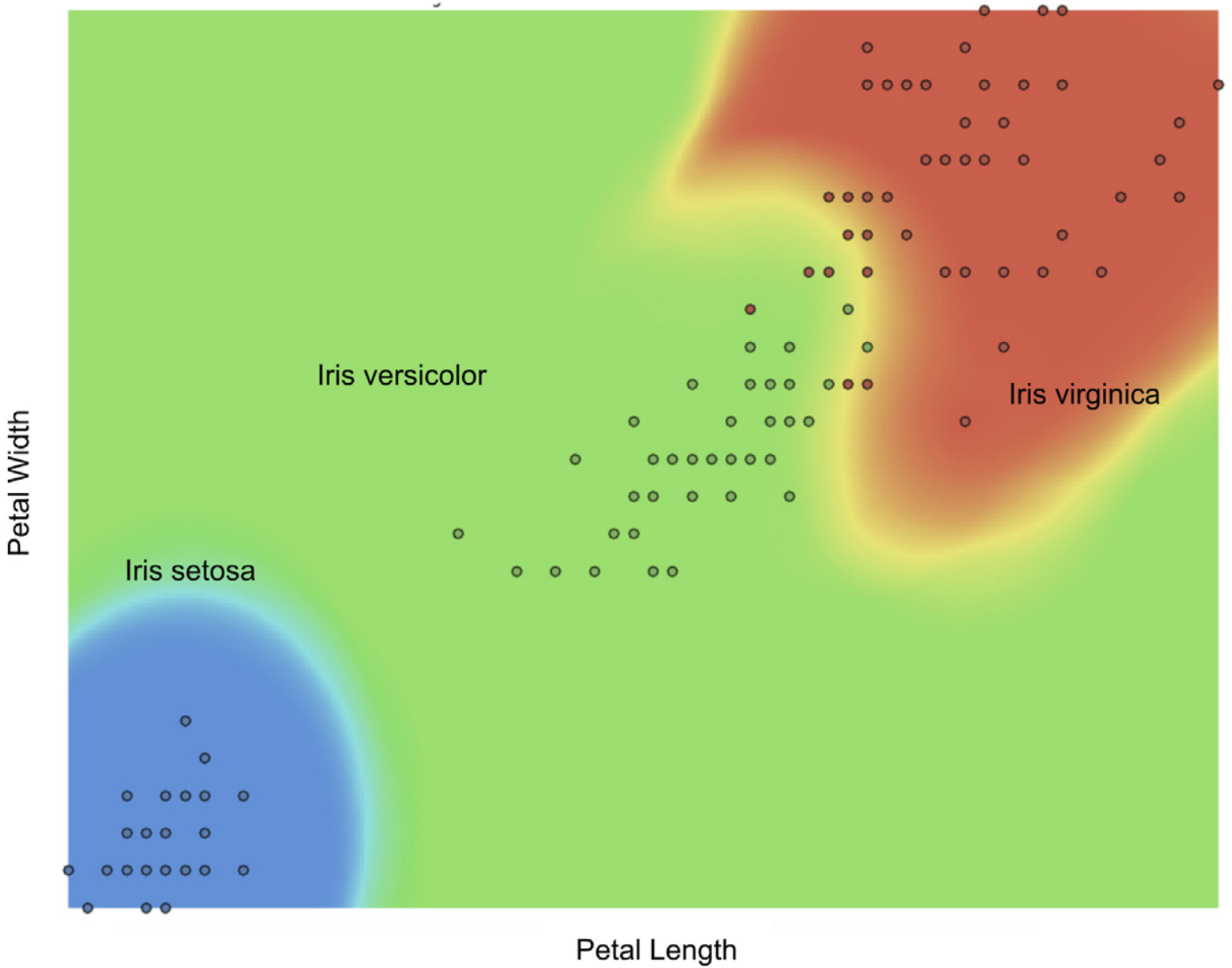
- A bubble chart is a variation of a simple scatterplot with the addition of one more attribute, which is used to determine the size of the data point.
- In the Iris dataset, petal length and petal width are used for x and y-axis, respectively and sepal width is used for the size of the data point. The color of the data point represents a species class label

Bubble Chart of Iris Data



Density Chart

- Density charts are similar to the scatterplots, with one more dimension included as a background color. The data point can also be colored to visualize one dimension, and hence, a total of four dimensions can be visualized in a density chart.
- In the example in Fig. petal length is used for the x-axis, sepal length for the y-axis, sepal width for the background color, and class label for the data point color.


Iris versicolor
Iris virginica
Iris setosa
Petal Width
Petal Length

Distribution Chart

- For continuous numeric attributes like petal length, instead of visualizing the actual data in the sample, its normal distribution function can be visualized instead.
- If a dataset exhibits normal distribution, then 68.2% of data points will fall within one standard deviation from the mean; 95.4% of the points will fall within 2σ and 99.7% within 3σ of the mean.

Distribution Chart