

DATA EXPLORATION

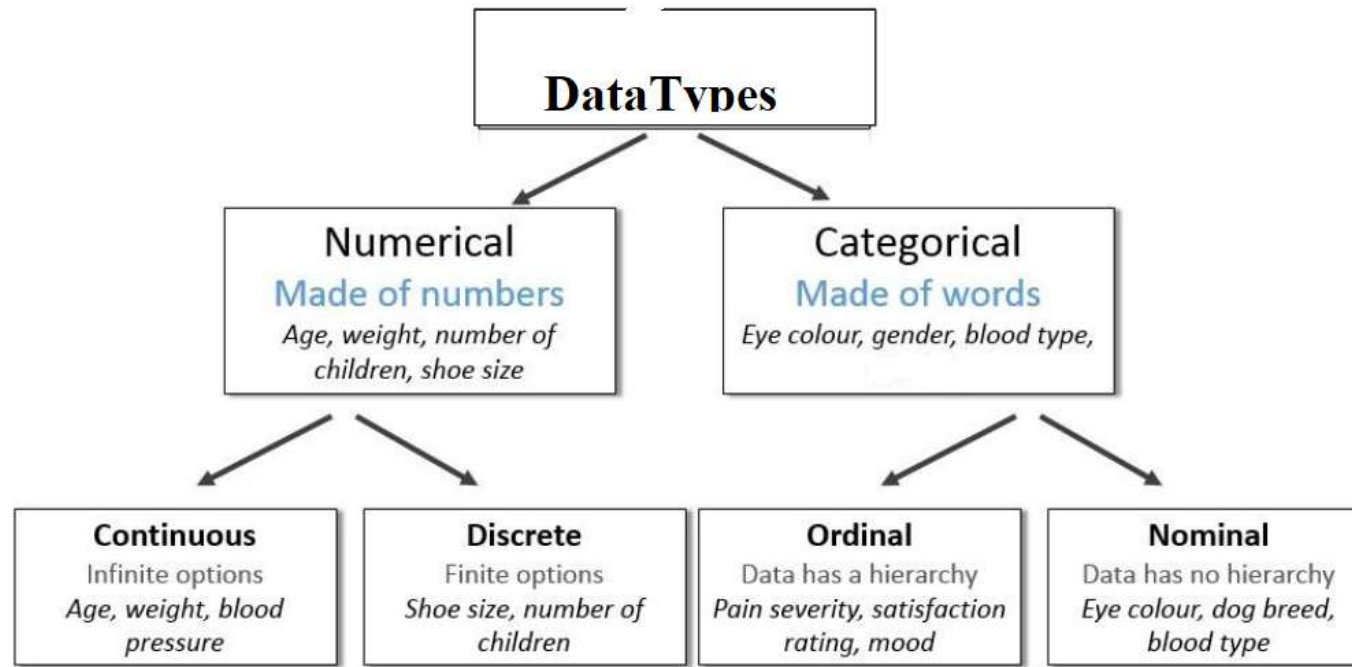
- Data exploration, also known as exploratory data analysis, provides a set of tools to obtain fundamental understanding of a dataset.
- Data exploration also provides guidance on applying the right kind of further statistical and data science treatment.
- Data exploration can be broadly classified into two types—
 1. descriptive statistics
 2. data visualization.
- Descriptive statistics is the process of condensing key characteristics of the dataset into simple numeric metrics. Some of the common quantitative metrics used are mean, standard deviation, and correlation.
- Visualization is the process of projecting the data, or parts of it, into multi-dimensional space.

DATA SET

A dataset (example set) is a collection of data with a defined structure.

Types of the data

Data come in different formats and types.



DESCRIPTIVE STATISTICS

Descriptive statistics refers to the study of the aggregate quantities of a dataset.

These measures are some of the commonly used notations in everyday life. Some examples of descriptive statistics include average annual income, median home price in a neighborhood, range of credit scores of a population, etc.

Characteristics of the Dataset	Measurement Technique
Center of the dataset	Mean, median, and mode
Spread of the dataset	Range, variance, and standard deviation
Shape of the distribution of the dataset	Symmetry, skewness, and kurtosis

Descriptive statistics can be broadly classified into

univariate

and

multivariate exploration depending on the number of attributes under analysis.

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 but has shortcomings as it is severely impacted by the presence of outliers.

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.

For a dataset with N observations, the variance is given by the following equation:

$$\text{Variance} = s^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

Standard deviation is the square root of the variance. 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.

Table 3.1 Iris Dataset 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

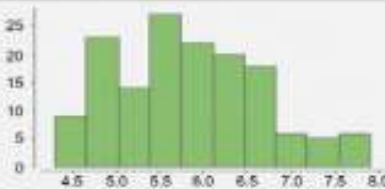
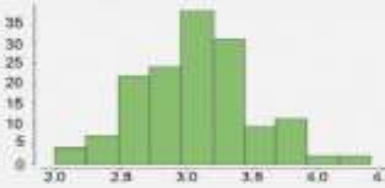
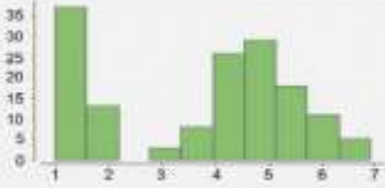
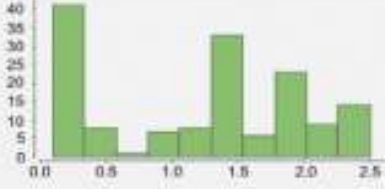
^ Sepal Length	Real	0	 Open chart	Min 4.300	Max 7.900	Average 5.843	Deviation 0.828
^ Sepal Width	Real	0	 Open chart	Min 2	Max 4.400	Average 3.054	Deviation 0.434
^ Petal Length	Real	0	 Open chart	Min 1	Max 6.900	Average 3.759	Deviation 1.764
^ Petal Width	Real	0	 Open chart	Min 0.100	Max 2.500	Average 1.199	Deviation 0.763

FIGURE 3.2

Descriptive statistics for the Iris dataset.

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, which is central to data science methods.

Central Data Point

Data point made up of the mean of each attribute in the dataset independently.

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}.

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}.

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.

For example, consider average temperature of the day and ice cream sales.

Statistically, the two attributes that are correlated are dependent on each other and one may be used to predict the other. If there are sufficient data, future sales of ice cream can be predicted if the temperature forecast is known.

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.

A correlation value of 0 means there is no linear relationship between two attributes.

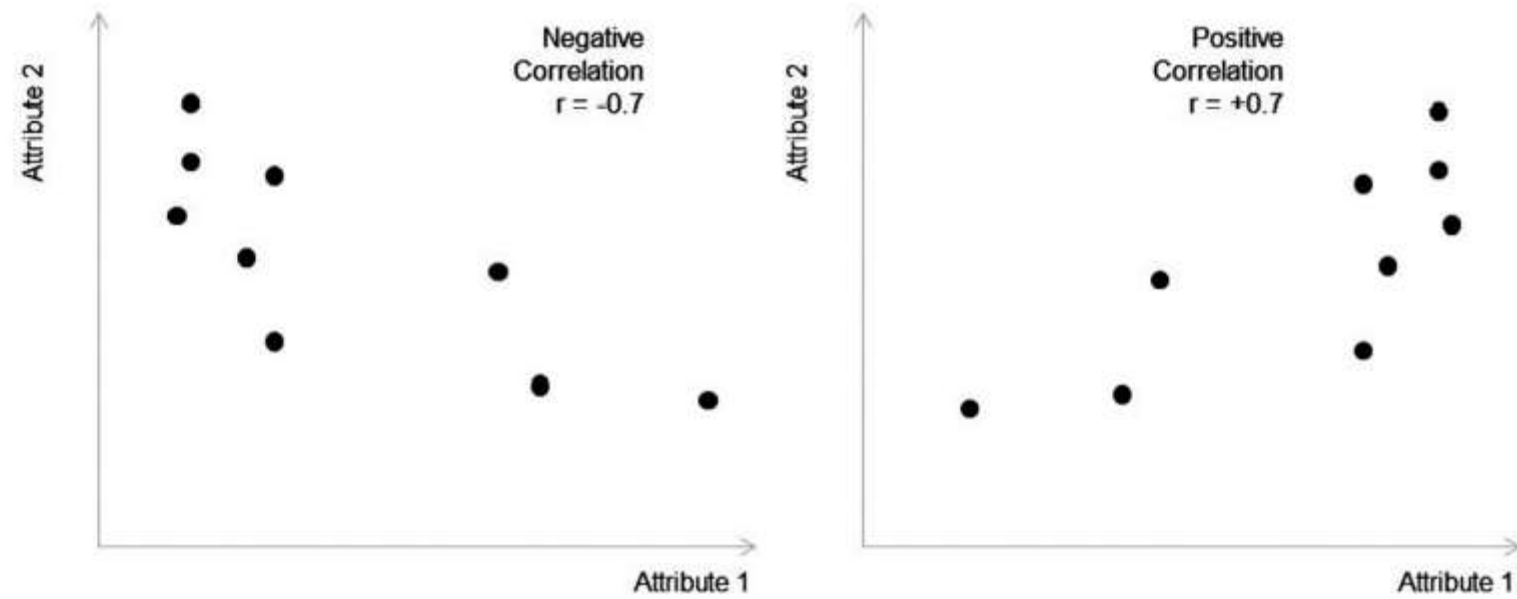


FIGURE 3.3
Correlation of attributes.

The Pearson correlation coefficient between two attributes x and y is calculated with the formula:

$$\begin{aligned} r_{xy} &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \\ &= \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{N \times s_x \times s_y} \end{aligned}$$

where s_x and s_y are the standard deviations of random variables x and y , respectively.

DATA VISUALIZATION

Visualizing data is one of the most important techniques of data discovery and exploration.

The visual representation of data provides easy comprehension of complex data with multiple attributes and their underlying relationships.

Univariate Visualization

Visual exploration starts with investigating one attribute at a time using univariate charts.

1. Histogram

A histogram is one of the most basic visualization techniques to understand the frequency of the occurrence of values.

It shows the distribution of the data by plotting the frequency of occurrence in a range.

In a histogram, the attribute under inquiry is shown on the horizontal axis and the frequency of occurrence is on the vertical axis.

Histograms are used to find the central location, range, and shape of distribution.

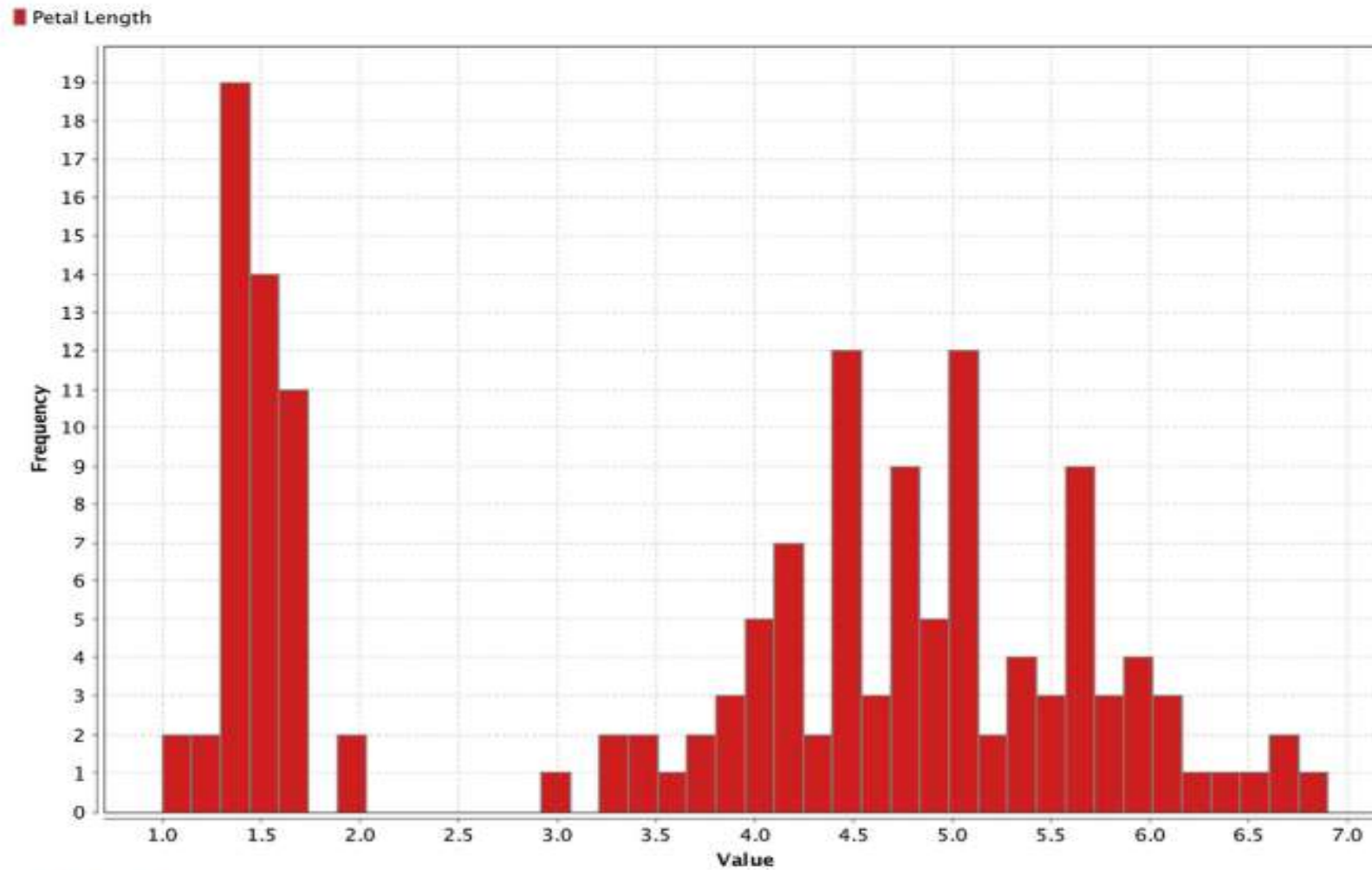


FIGURE 3.5

Histogram of petal length in Iris dataset.

2. Quartile

A box whisker plot is a simple visual way of showing the distribution of a continuous variable with information such as quartiles, median, and outliers, overlaid by mean and standard deviation.

The quartiles are denoted by Q1, Q2, and Q3 points, which indicate the data points with a 25% bin size. In a distribution, 25% of the data points will be below Q1, 50% will be below Q2, and 75% will be below Q3.

The Q1 and Q3 points in a box whisker plot are denoted by the edges of the box. The Q2 point, the median of the distribution, is indicated by a cross line within the box. The outliers are denoted by circles at the end of the whisker line.

Diagram shows that the quartile charts for all four attributes of the Iris dataset are plotted side by side. Petal length can be observed as having the broadest range and the sepal width has a narrow range, out of all of the four attributes.

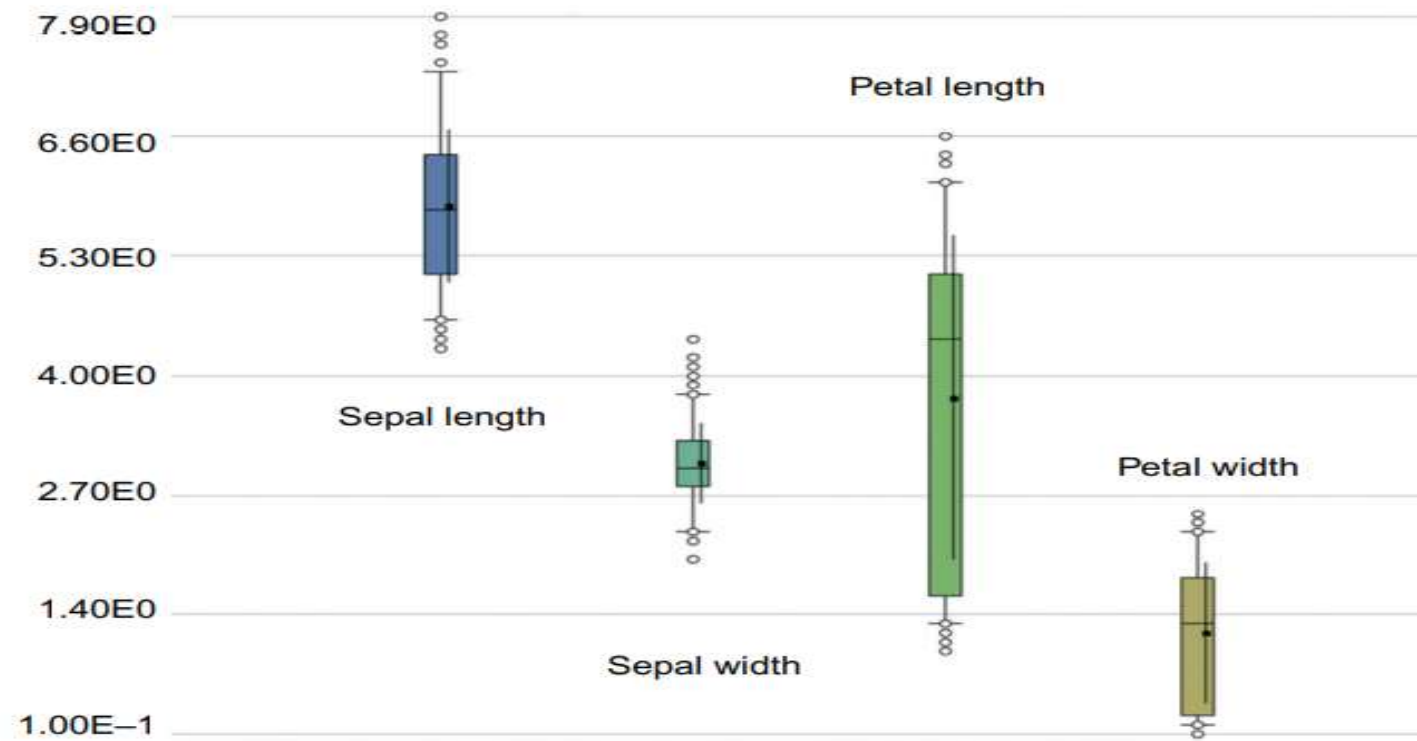


FIGURE 3.7
Quartile plot of Iris dataset.

3.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.

The normal distribution function of a continuous random variable is given by the formula:

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

where μ is the mean of the distribution and σ is the standard deviation of the distribution.

The normal distribution is also called the Gaussian distribution or “bell curve” due to its bell shape.

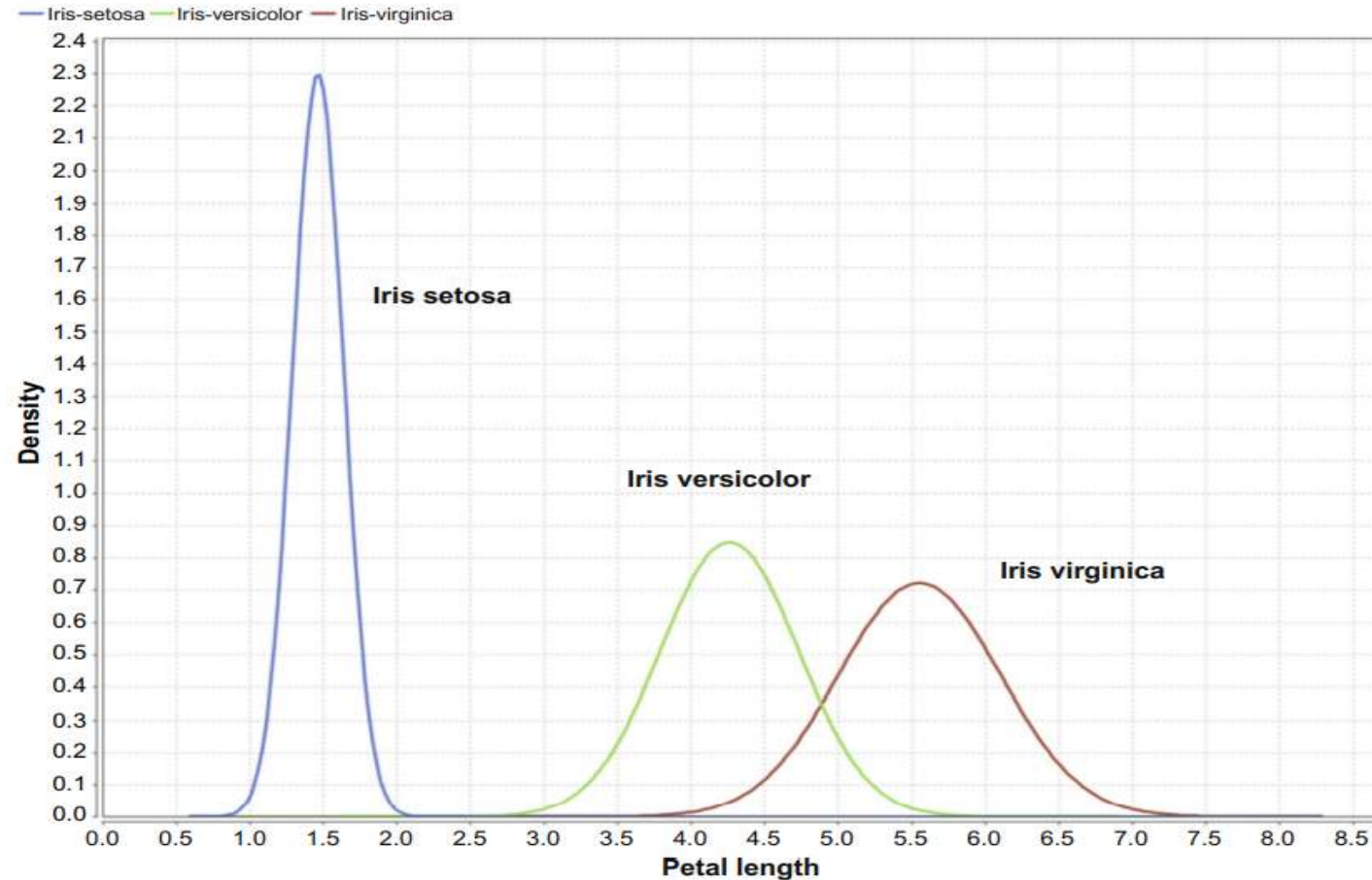


FIGURE 3.9
Distribution of petal length in Iris dataset.

Multivariate Visualization

The multivariate visual exploration considers more than one attribute in the same visual.

It focus on the relationship of one attribute with another attribute.

These visualizations examine two to four attributes simultaneously.

1. Scatterplot

A scatterplot is one of the most powerful yet simple visual plots available.

In a scatterplot, the data points are marked in Cartesian space with attributes of the dataset aligned with the coordinates.

The attributes are usually of continuous data type.

One of the key observations that can be concluded from a scatterplot is the existence of a relationship between two attributes.

If the attributes are linearly correlated, then the data points align closer to an imaginary straight line; if they are not correlated, the data points are scattered.

Apart from basic correlation, scatterplots can also indicate the existence of patterns or groups of clusters in the data and identify outliers in the data.

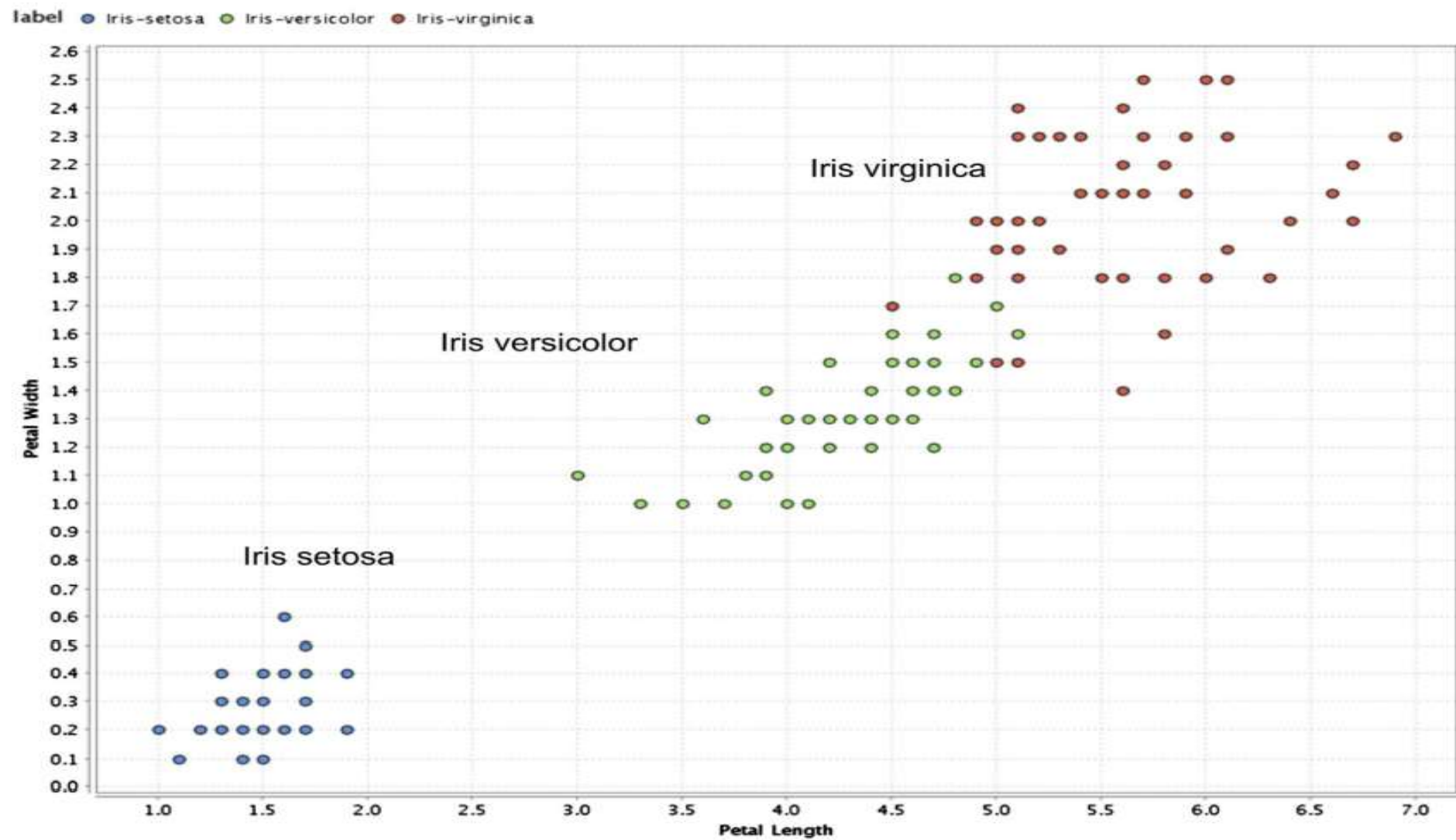


FIGURE 3.10
Scatterplot of Iris dataset.

2.A 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.

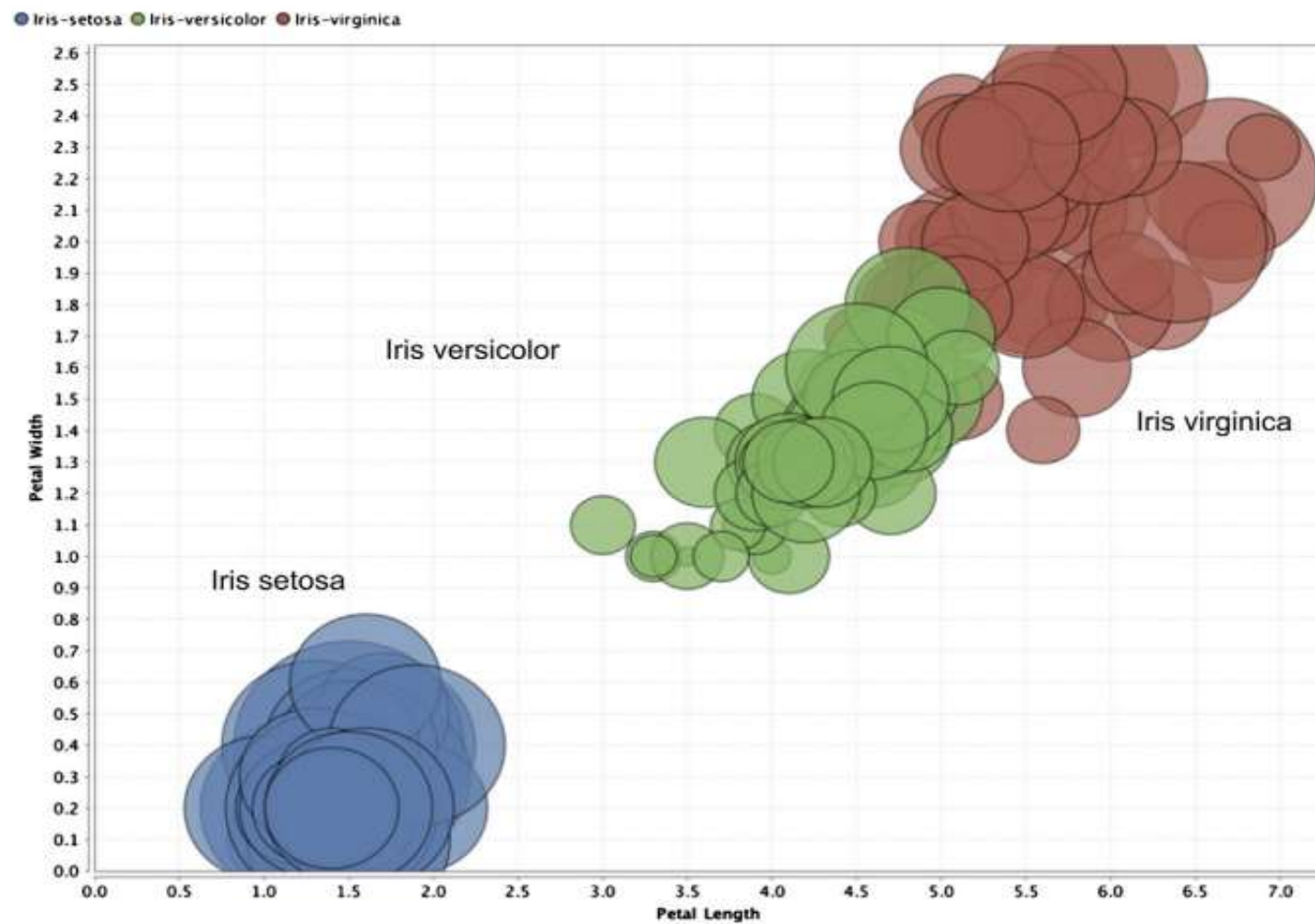


FIGURE 3.13

Bubble chart of Iris dataset.

3.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, 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.

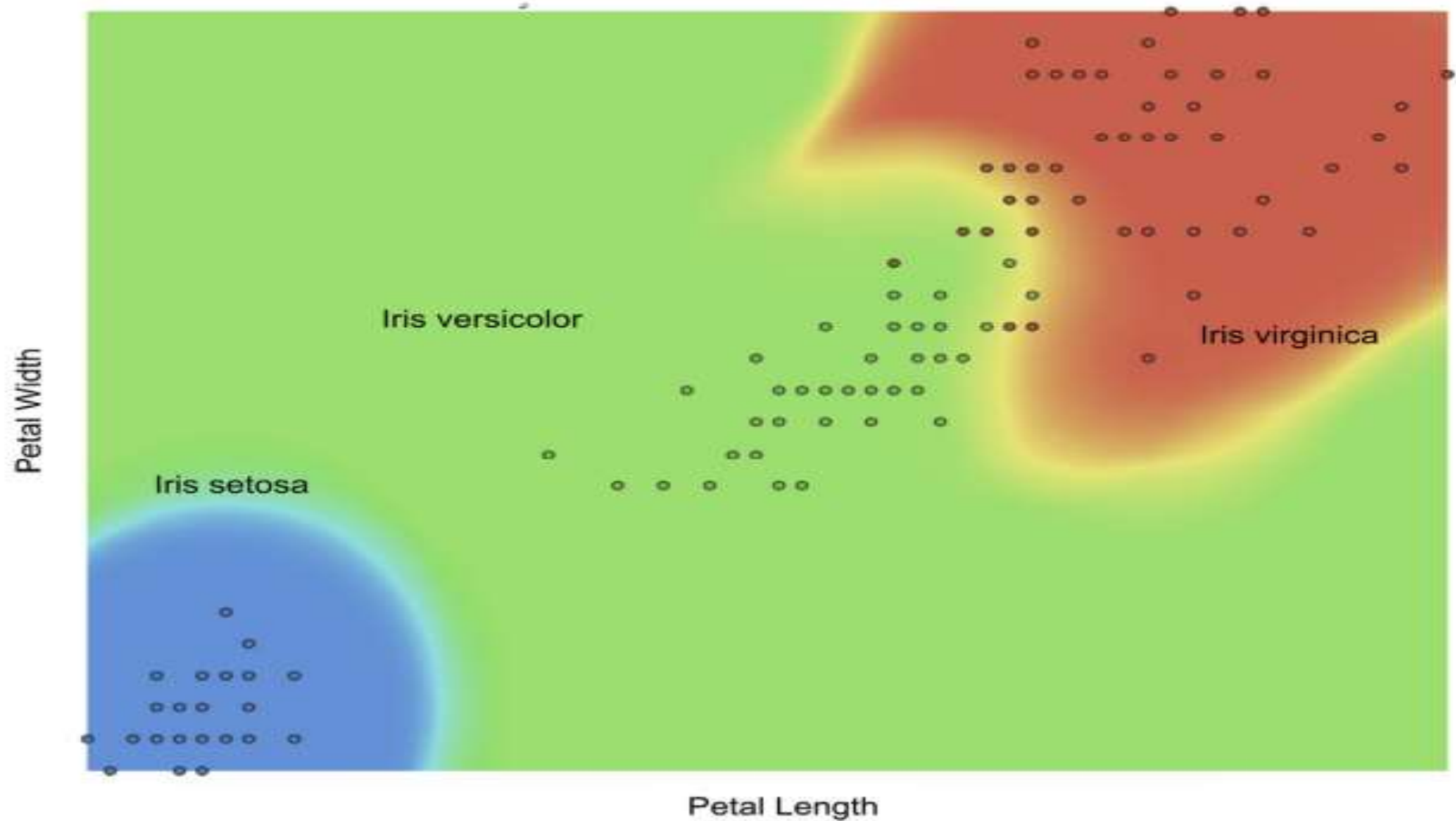


FIGURE 3.14

Density chart of a few attributes in the Iris dataset.