Data Exploration

Module 1

Introduction

- Data exploration, also known as exploratory data analysis, provides a set of tools to obtain fundamental understanding of a dataset. Before going to any advanced analysis of data, it is essential to perform basic data exploration to study the basic characteristics of a dataset. It helps to understand the data better, so that makes advanced analysis possible.

- The results of data exploration can be extremely powerful in
  - grasping the structure of the data
  - the distribution of the values
  - the presence of extreme values
  - the interrelationships between the attributes in the dataset.
  - helps to choose the right kind of further statistical and data science treatment.

Introduction…

- Data exploration can be broadly classified into two types—
  - Descriptive statistics
  - Data visualization

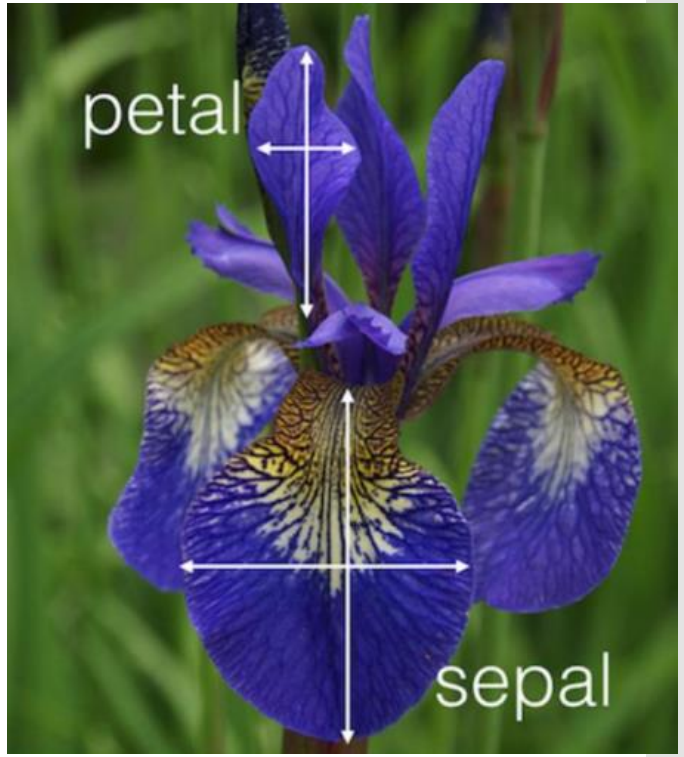
Descriptive statistics is the process of condensing key characteristics of the dataset into simple numeric metrics. Some of the common quantitative metrics used are mean, standard deviation, and correlation. Visualization is the process of projecting the data, or parts of it, into multi-dimensional space or abstract images. All the useful (and adorable) charts fall under this category.

Objectives

- **Data Understanding**: Data exploration provides a high-level overview of each attribute in the dataset and the interaction between the attributes.
- **Data Preparation:** Before applying the data science algorithm, the dataset has to be prepared for handling any of the anomalies that may be present in the data. These anomalies include outliers, missing values, or highly correlated attributes and are removed through data preparation.
- **Data Science Tasks**: Basic data exploration can sometimes substitute the entire data science process. In some cases, scatterplots can identify clusters in low-dimensional data or can help to develop regression or classification models with simple visual rules.
- **Interpreting the results**: Finally, data exploration is used in understanding the prediction, classification, and clustering of the results of the data science process.

3. DATASETS

- Data science techniques work on datasets. 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.
- Iris is a flowering plant that is found widely, across the world and has more than 300 different species. Each species exhibits different physical characteristics like shape and size of the flowers and leaves.
- The Iris dataset contains 150 observations of three different species, Iris setosa, Iris virginica, and Iris 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, takes the values setosa, virginica, and versicolor.

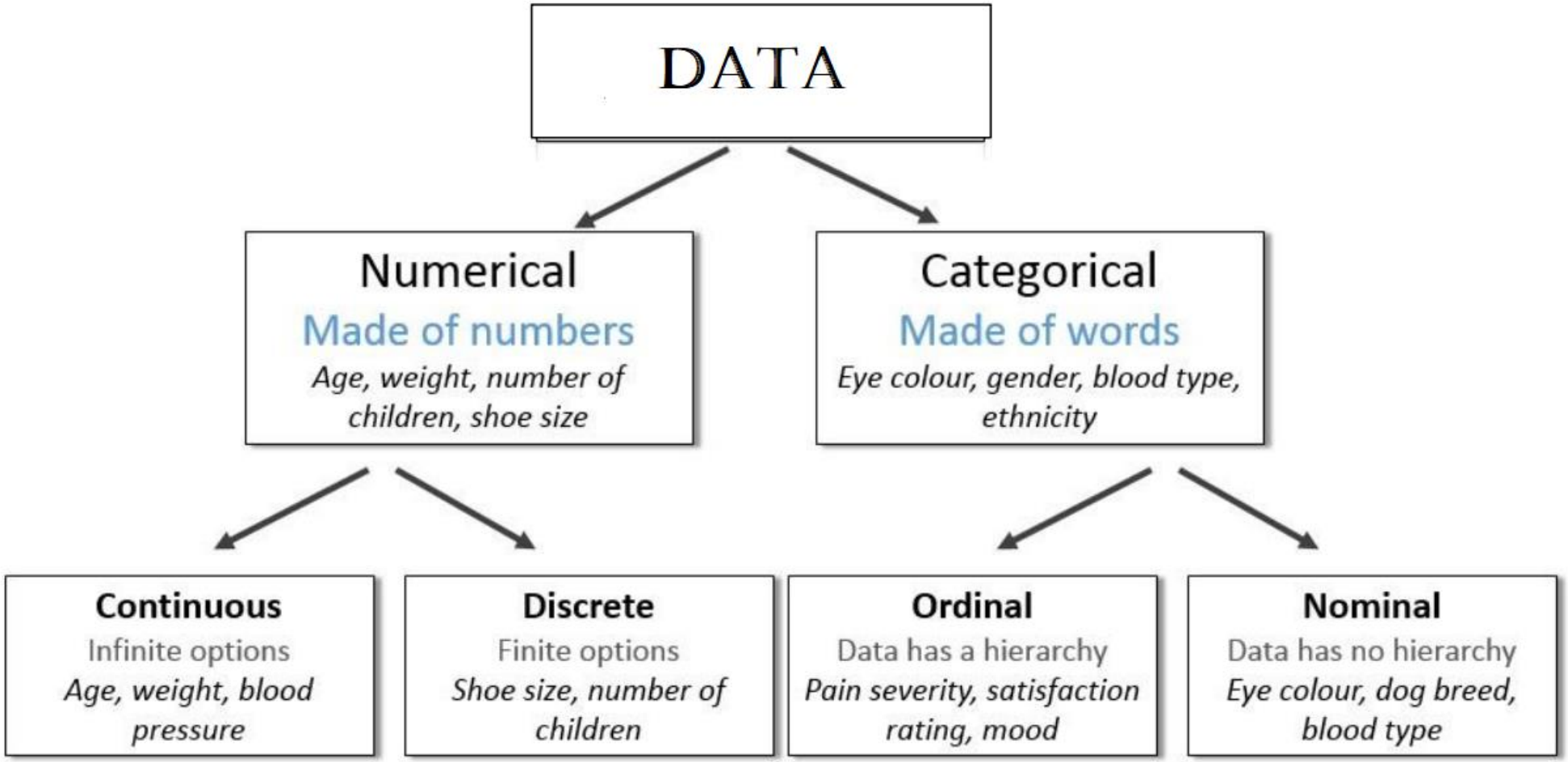


Sample Dataset-Iris

sepal.length	sepal.width	petal.length	petal.width	variety
5.1	3.5	1.4	.2	Setosa
4.9	3	1.4	.2	Setosa
4.7	3.2	1.3	.2	Setosa
4.6	3.1	1.5	.2	Setosa
5	3.6	1.4	.2	Setosa
5.4	3.9	1.7	.4	Setosa
4.6	3.4	1.4	.3	Setosa
5	3.4	1.5	.2	Setosa
5.7	2.9	4.2	1.3	Versicolor
6.2	2.9	4.3	1.3	Versicolor
5.1	2.5	3	1.1	Versicolor
5.7	2.8	4.1	1.3	Versicolor
6.3	3.3	6	2.5	Virginica
5.8	2.7	5.1	1.9	Virginica
7.1	3	5.9	2.1	Virginica
6.3	2.9	5.6	1.8	Virginica

3.1 Types of Data…

- Data come in different formats and types. Understanding the properties of each attribute provides information about what kind of operations can be performed on that attribute.
  - 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)
- All of these attributes indicate temperature in a region, but each have different data types. A few of these data types can be converted from one to another.


DATA
Numerical
Made of numbers
Age, weight, number of children, shoe size
Categorical
Made of words
Eye colour, gender, blood type, ethnicity
Continuous
Infinite options
Age, weight, blood pressure
Discrete
Finite options
Shoe size, number of children
Ordinal
Data has a hierarchy
Pain severity, satisfaction rating, mood
Nominal
Data has no hierarchy
Eye colour, dog breed, blood type

3.1.1 Numeric or Continuous

- Temperature expressed in Centigrade or Fahrenheit is numeric and continuous because it can be denoted by numbers and take an infinite number of values between digits. Additive and subtractive mathematical operations and logical comparison operators can be applied. Both integer and ratio data types are categorized as a numeric data type in most data science tools.

3.1.2 Categorical or Nominal

- Categorical data types are attributes treated as distinct symbols or just names. Temperature marked as hot, mild, or cold is an example. They are also called a nominal or polynominal data type. Some data science algorithm does not work with categorical data. A type conversion process can be used to convert one data to another and may occur possible loss of information.

3.2 Descriptive Statistics

- Descriptive statistics refers to the study of the aggregate quantities of a dataset. These include average, median, range etc. In general, descriptive analysis covers the following characteristics of the sample:

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.

3.2.1 Univariate Exploration

- Univariate data exploration denotes analysis of one attribute at a time. The example Iris dataset for one species, *setosa*, has 50 observations and 4 attributes. some of the descriptive statistics for sepal length attribute is given in the following table.

Iris Dataset and Descriptive Statistics

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

Measures of Central Tendency

- *The objective is to find the central location of an attribute.*
- **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. The mean for sepal length in centimeters is 5.0060.
- **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. The median for sepal length is in centimeters is 5.0000.
- **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. In this example, the mode in centimeters is 5.1000.

If the dataset has outliers, the mean will get affected while in most cases the median will not.

Measures of Spread

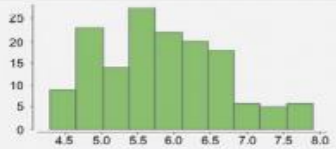
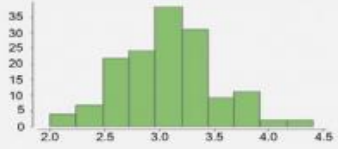
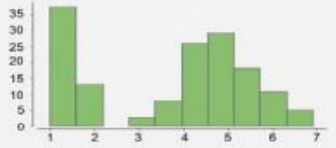
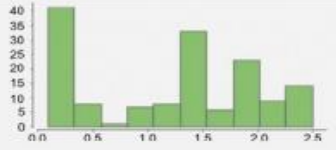
- There are two common metrics to quantify spread.
- **Range**: The range is the difference between the maximum value and the minimum value of the attribute. The range is simple to calculate, but has shortcomings as it is severely impacted by the presence of outliers.

- **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 (xi) 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

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

Measures of Spread

- 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.
- Following figure provides the univariate summary of the Iris dataset with all 150 observations,

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

Measures of Spread- Problems

1. The tensile strength in megapascals for 15 samples of tin were determined and found to be: 34.61, 34.57, 34.40, 34.63, 34.63, 34.51, 34.49, 34.61, 34.52, 34.55, 34.58, 34.53, 34.44, 34.48 and 34.40. Calculate the mean and standard deviation from the mean for these 15 values, correct to 4 significant figures.

Mean, $m = \dfrac{\text{sum of the terms}}{\text{number of terms}}$

= (34.61+ 34.57+ 34.40+ 34.63+ 34.63+ 34.51+34.49+34.61+34.52+ 34.55+34.58+ 34.53+ 34.44+ 34.48 + 34.40)/15

= **34.53** Megapascals

Standard Deviation, $\sigma = \sqrt{\dfrac{\sum(x_i - \mu)^2}{N}}$

Measures of Spread- Problems

Standard Deviation, $\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$

X	34.61	34.57	34.4	34.63	34.63	34.51	34.49	34.61	34.52	34.55	34.58	34.53	34.44	34.48	34.4
X- mean	0.08	0.04	-0.13	0.1	0.1	-0.02	-0.04	0.08	-0.01	0.02	0.05	0	-0.09	-0.05	-0.13
(X-mean)**2	0.0064	0.0016	0.0169	0.0100	0.0100	0.0004	0.0016	0.0064	0.0001	0.0004	0.0025	0.0000	0.0081	0.0025	0.0169

Sum of Deviation = 0.0838

SD= $\sqrt{0.0838}/15$

= $\sqrt{0.005587}$ = **0.07474**

Measures of Spread- Problems

2. **Find the mean of the first 10 odd integers.**

First 10 odd integers: 1, 3, 5, 7, 9, 11, 13, 15, 17, 19
Mean = Sum of the first 10 odd integers/Number of such integers
= (1 + 3 + 5 + 7 + 9 + 11 + 13 + 15 + 17 + 19)/10
= 100/10
= 10

Therefore, the **mean of the first 10 odd integers is 10.**

Measures of Spread- Problems

3. **What is the median of the following data set?**
32, 6, 21, 10, 8, 11, 12, 36, 17, 16, 15, 18, 40, 24, 21, 23, 24, 24, 29, 16, 32, 31, 10, 30, 35, 32, 18, 39, 12, 20

The ascending order of the given data set is:
6, 8, 10, 10, 11, 12, 12, 15, 16, 16, 17, 18, 18, 20, 21, 21, 23, 24, 24, 24, 29, 30, 31, 32, 32, 32, 35, 36, 39, 40
Number of values in the data set = n = 30
n/2 = 30/2 = 15
15th data value = 21
(n/2) +1 = 16
16th data value = 21

Median = [(n/2)th observation + {(n/2)+1}th observation]/2
= (15th data value + 16th data value)/2
= (21 + 21)/2
= **21**

Measures of Spread- Problems

3. **Identify the mode for the following data set:**
21, 19, 62, 21, 66, 28, 66, 48, 79, 59, 28, 62, 63, 63, 48, 66, 59, 66, 94, 79, 19 94

Write the given data set in ascending order as follows:
19, 19, 21, 21, 28, 28, 48, 48, 59, 59, 62, 62, 63, 63, 66, 66, 66, 66, 79, 79, 94, 94

Here, we can observe that the number 66 occurred the maximum number of times.

Thus, the **mode of the given data set is 66.**

3.2.2 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.
- **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}

Eg: {5.1, 3.5, 1.4, 0.2}

. 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, 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.

3.2.2 Multivariate Exploration…

- **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.
- 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 $-1 \leq r \leq 1$. 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.

3.2.2 Multivariate Exploration…

- **Correlation**

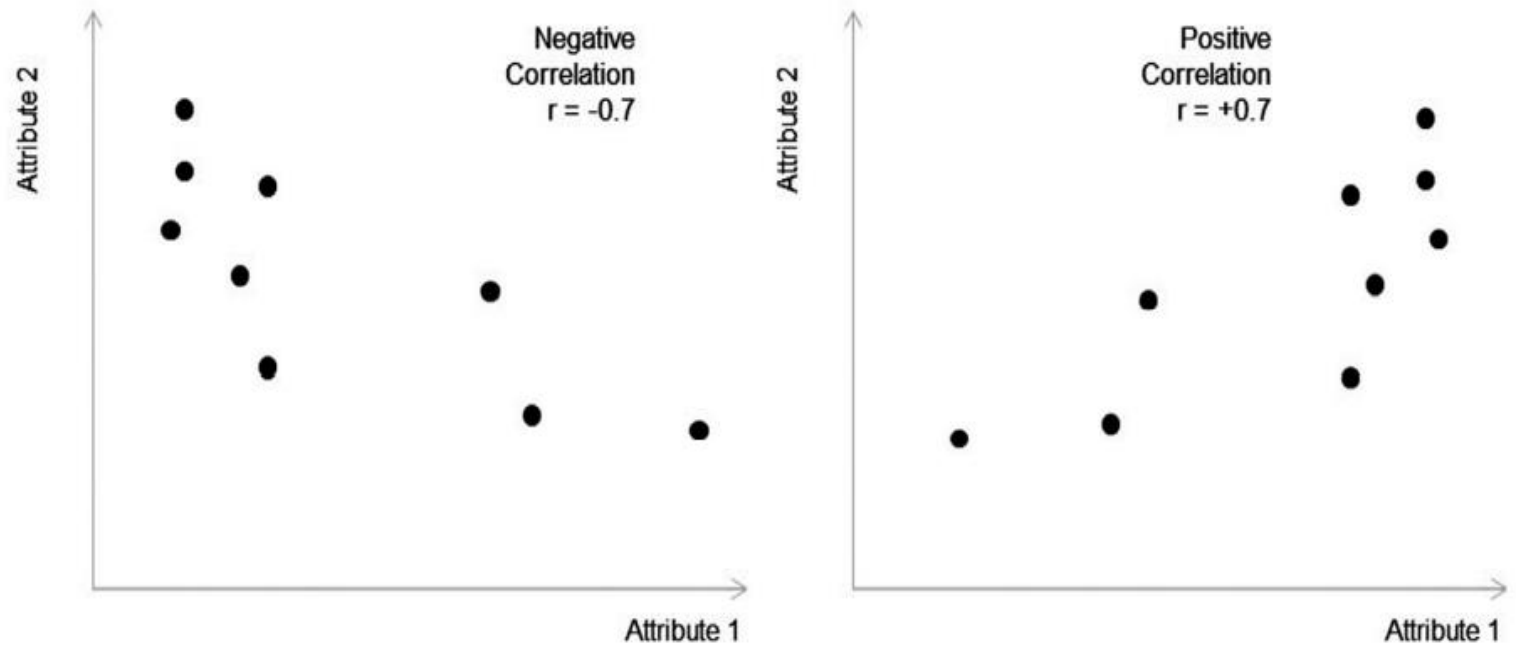


Fig: Correlation Of Attributes

3.3 DATA VISUALISATION

- Visualizing data is one of the most important techniques of data discovery and exploration. The discipline of data visualization includes the methods of expressing data in an abstract visual form. The visual representation of data provides easy comprehension of complex data with multiple attributes and their underlying relationships.

3.3.1 Univariate Visualization

- Visual exploration starts with investigating one attribute at a time using univariate charts. Methods are
  - **Histogram; Quartile; Distribution Chart;**

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.

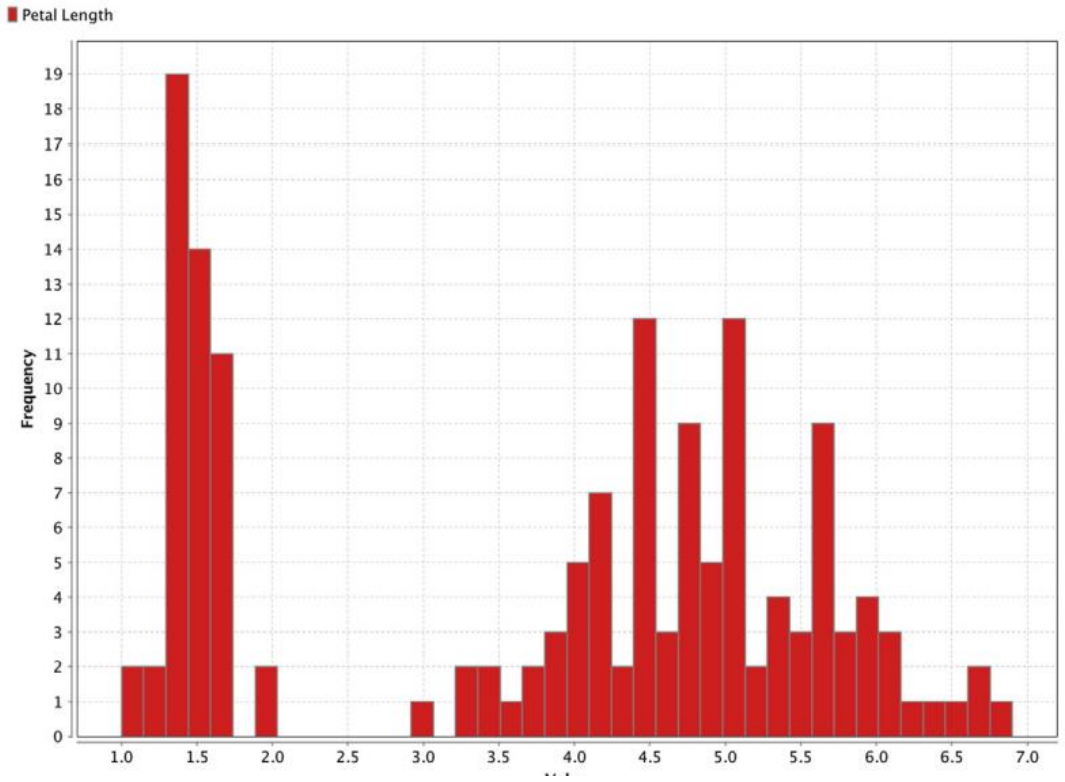


Fig: Histogram of petal length in Iris dataset

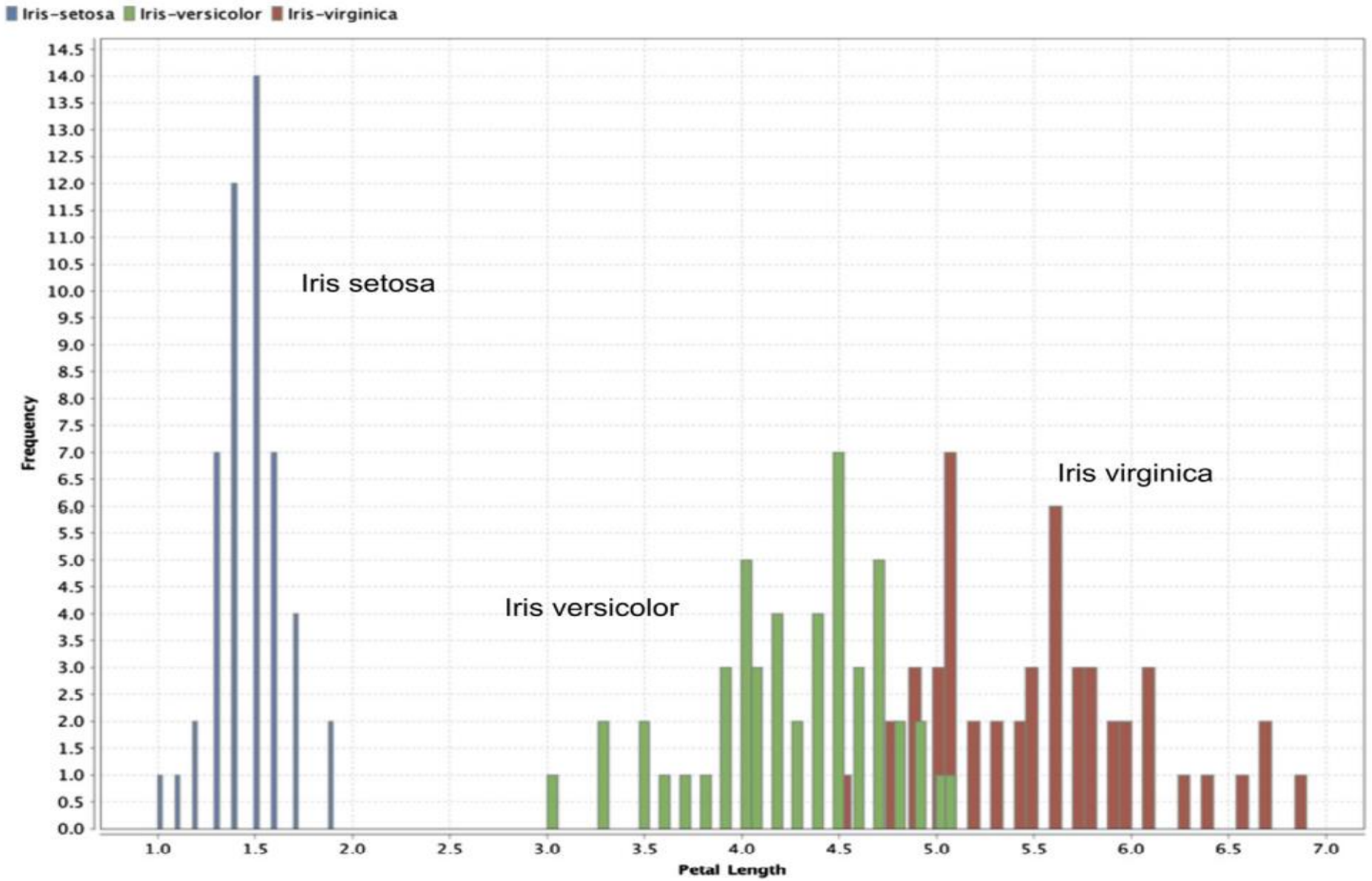


Fig: Class-stratified histogram of petal length in Iris dataset

Quartile

- A box and whisker plot is defined as a graphical method of displaying variation in a set of data.

- In most cases, a histogram analysis provides a sufficient display. At the same time, a box and whisker plot can provide additional details like multiple sets of data to be displayed in the same graph.

- The advantage is that box and whisker plots are
  - They are very effective and easy to read, as they can summarize data from multiple sources and display the results in a single graph.
  - Box and whisker plots allow for comparison of data from different categories for easier, more effective decision-making.

Quartile- When to use

- Use box and whisker plots when you have multiple data sets from independent sources that are related to each other in some way. Examples include:
  - Test scores between schools or classrooms
  - Data from before and after a process change
  - Similar features on one part, such as camshaft lobes
  - Data from duplicate machines manufacturing the same products

Quartile- How to construct

- The procedure to develop a box and whisker plot comes from the five statistics below.

  - **Minimum value:** The smallest value in the data set
  - **Second quartile:** The value below which the lower 25% of the data are contained
  - **Median value:** The middle number in a range of numbers
  - **Third quartile:** The value above which the upper 25% of the data are contained
  - **Maximum value:** The largest value in the data set

Quartile- How to construct

- For example, given the following 20 data points, the five required statistics are displayed.

Number	Data	
1	113	**Minimum value: 113**
2	116	
3	119	
4	121	
5	124	
		2nd quartile: 124
6	124	
7	125	
8	126	
9	126	
10	126	
		Median value: 126.5
11	127	
12	127	
13	128	
14	129	
15	130	
		3rd quartile: 130
16	130	
17	131	
18	132	
19	133	
20	136	**Maximum value: 136**

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 main attraction of box whisker or quartile charts is that distributions of multiple attributes can be compared side by side and the overlap between them can be deduced.
- 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.

Quartile…

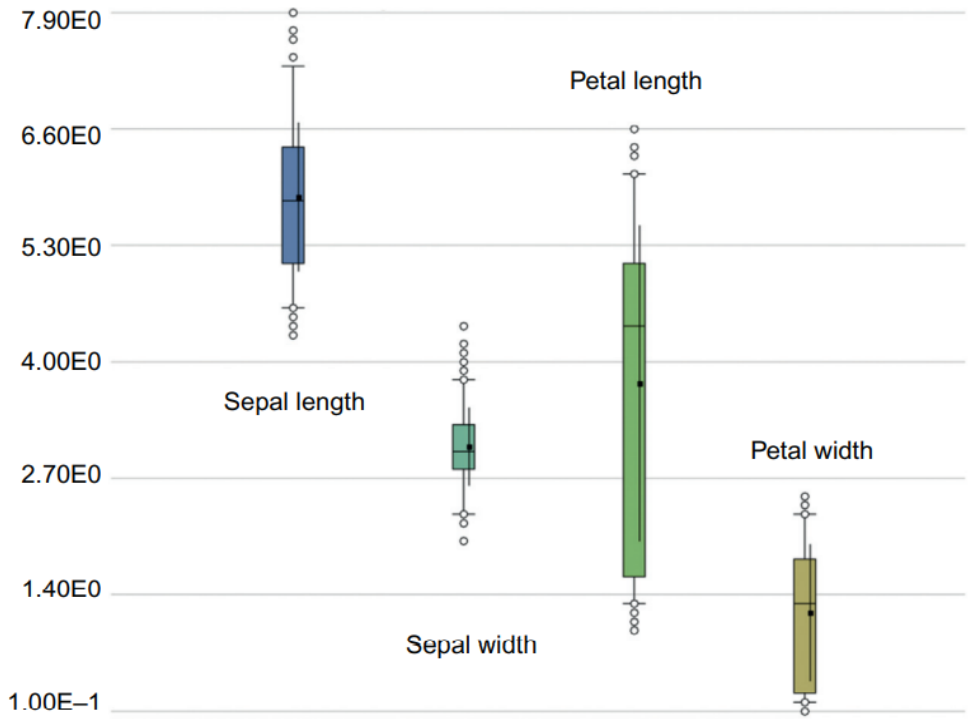


Fig1: Quartile plot of Iris dataset

Petal length having the broadest range and the sepal width has a narrow range, out of all of the four attributes

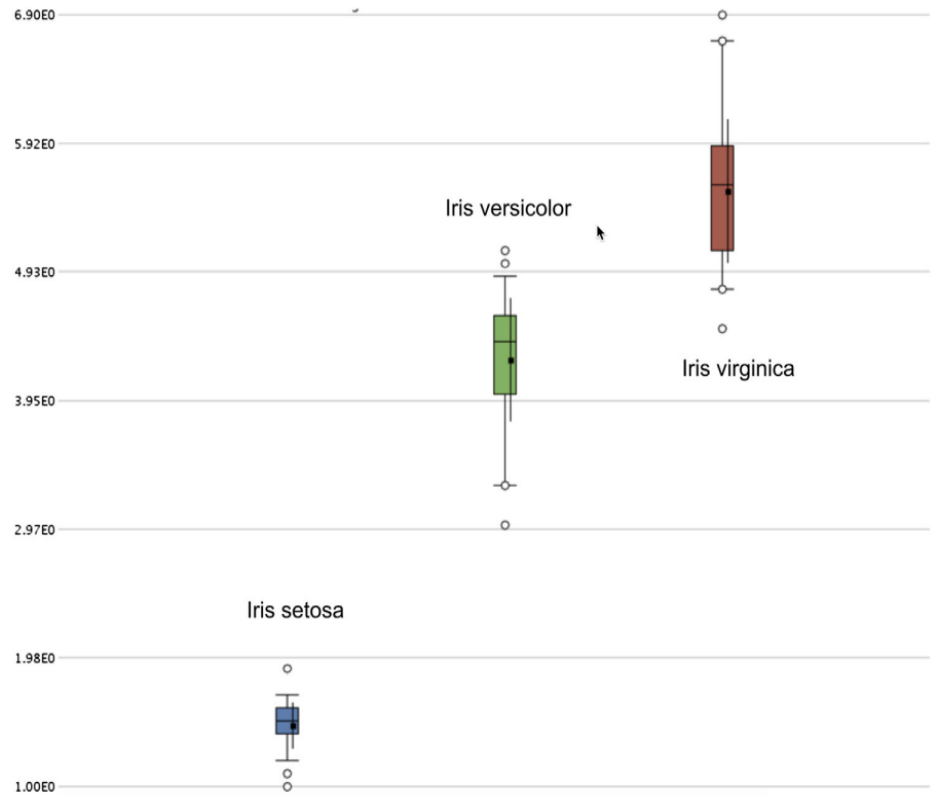


Fig2: Class-stratified quartile plot of petal length in Iris dataset

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.

Distribution Chart

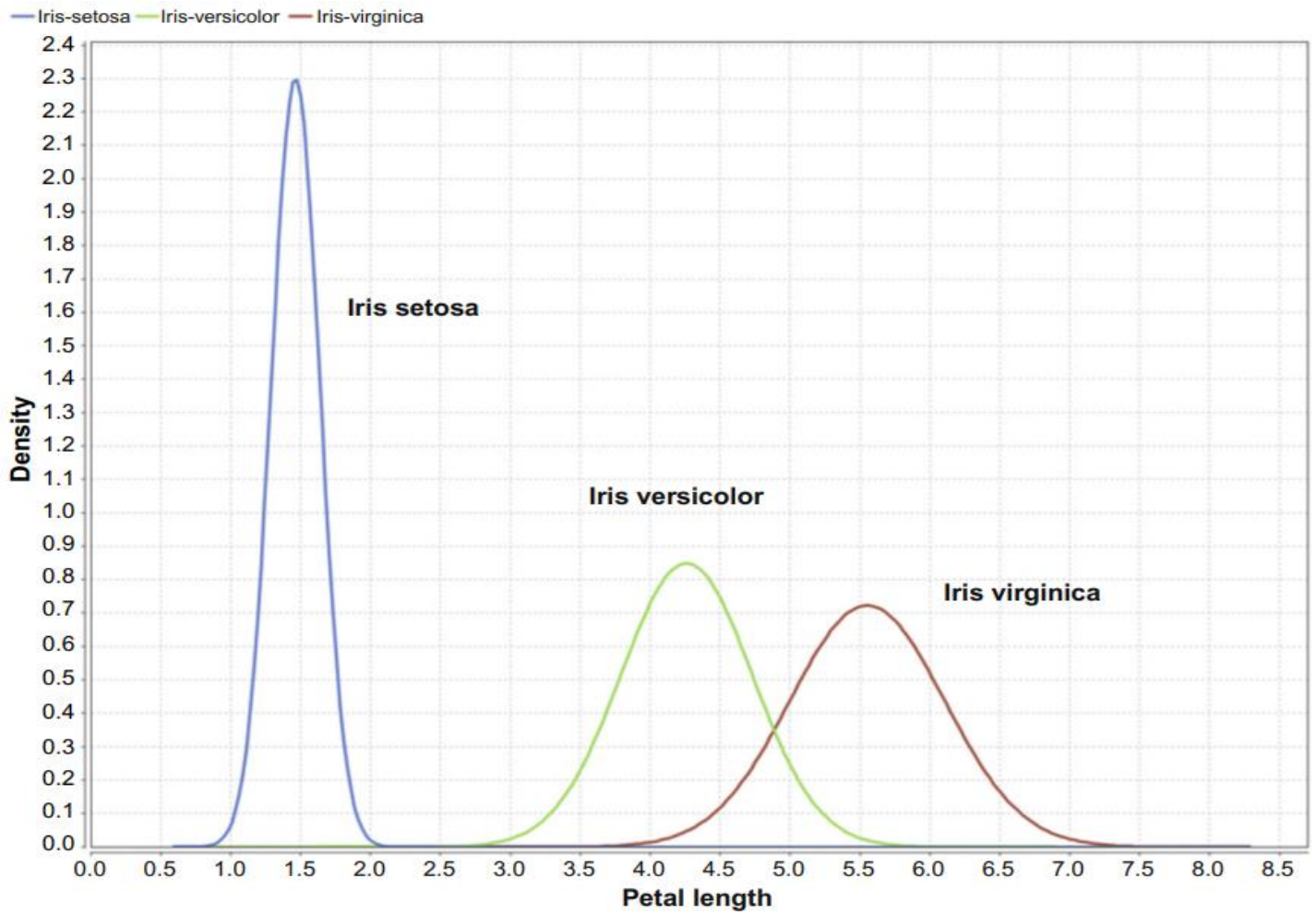


Fig: Distribution of petal length in Iris dataset.

Here petal length follow the normal distribution.

3.3.2 Multivariate Visualization

- The multivariate visual exploration considers more than one attribute in the same visual.
  - **Scatter plot; Scatter Multiple; Scatter Matrix; Bubble chart; Density Chart.**

Scatter Plot

- A scatterplot is one of the most powerful and simple visual plot. 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.
- The key observation obtained 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. Scatterplots can also indicate the existence of patterns or groups of clusters in the data and identify outliers in the data. This is particularly useful for low-dimensional datasets.

Scatter Plot

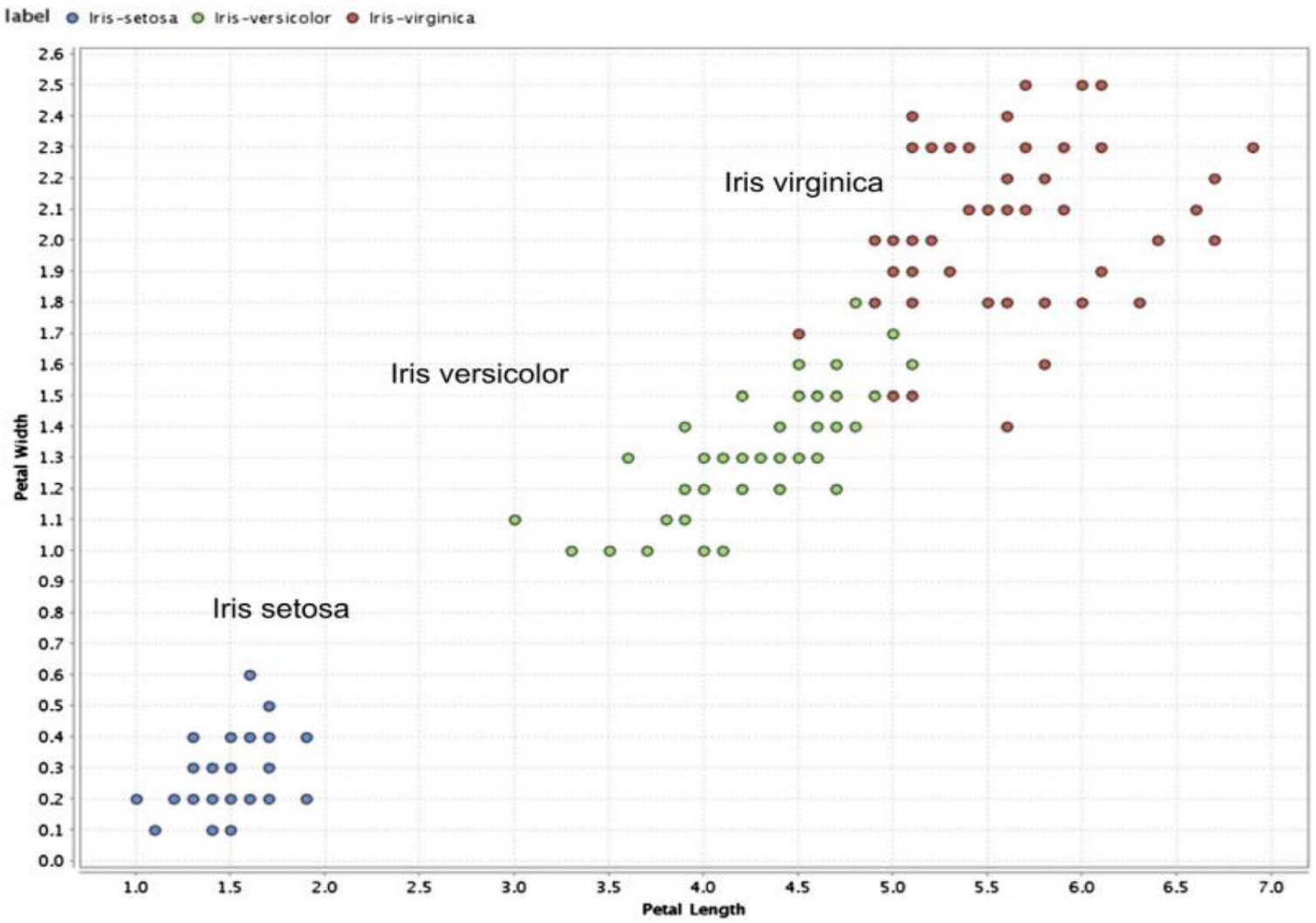
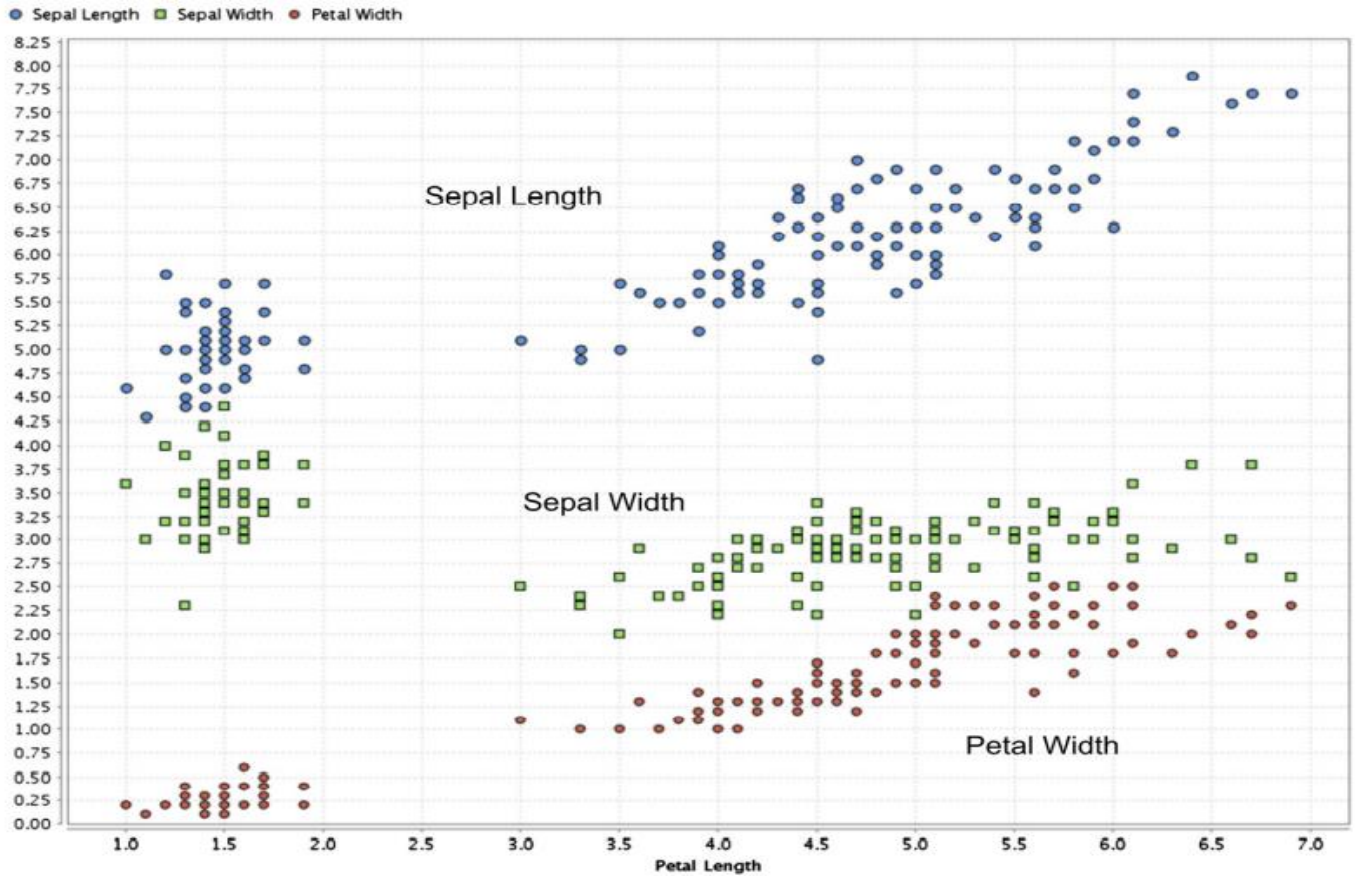


Fig: Scatterplot of Iris dataset.

Scatter Multiple

A scatter multiple is an enhanced form of a simple scatterplot where more than two dimensions can be included in the chart and studied simultaneously. The primary attribute is used for the x-axis coordinate. The secondary axis is shared with more attributes or dimensions.



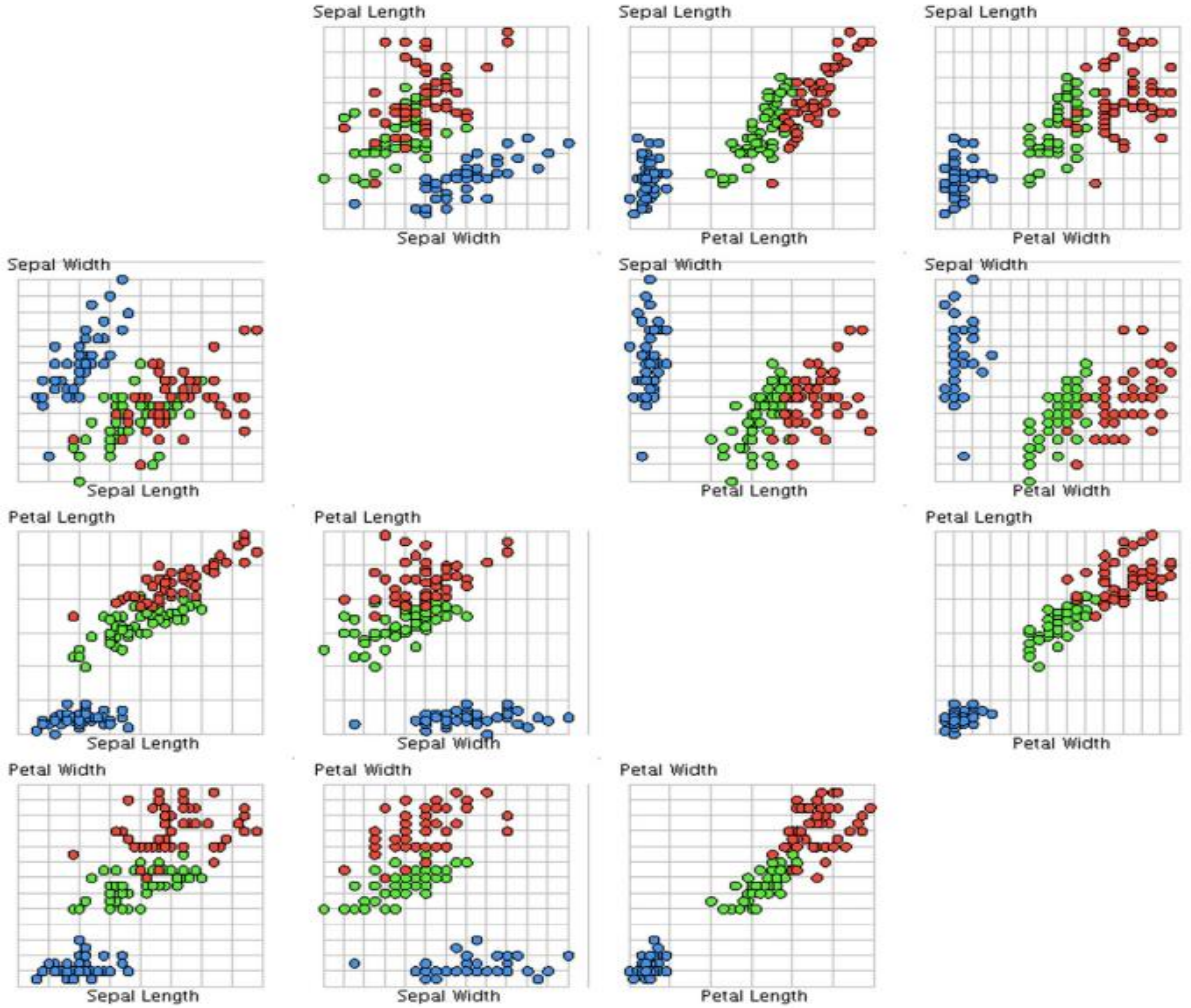
Scatter Multiple

In the graph, the values on the y-axis are shared between sepal length, sepal width, and petal width. Here, sepal length is represented by data points occupying the topmost part of the chart, sepal width occupies the middle portion, and petal width is in the bottom portion.

The data points are duplicated for each attribute in the y-axis, while the x-axis is anchored with one attribute— petal length. All the attributes sharing the y-axis should be of the same unit or normalized.

Scatter Matrix

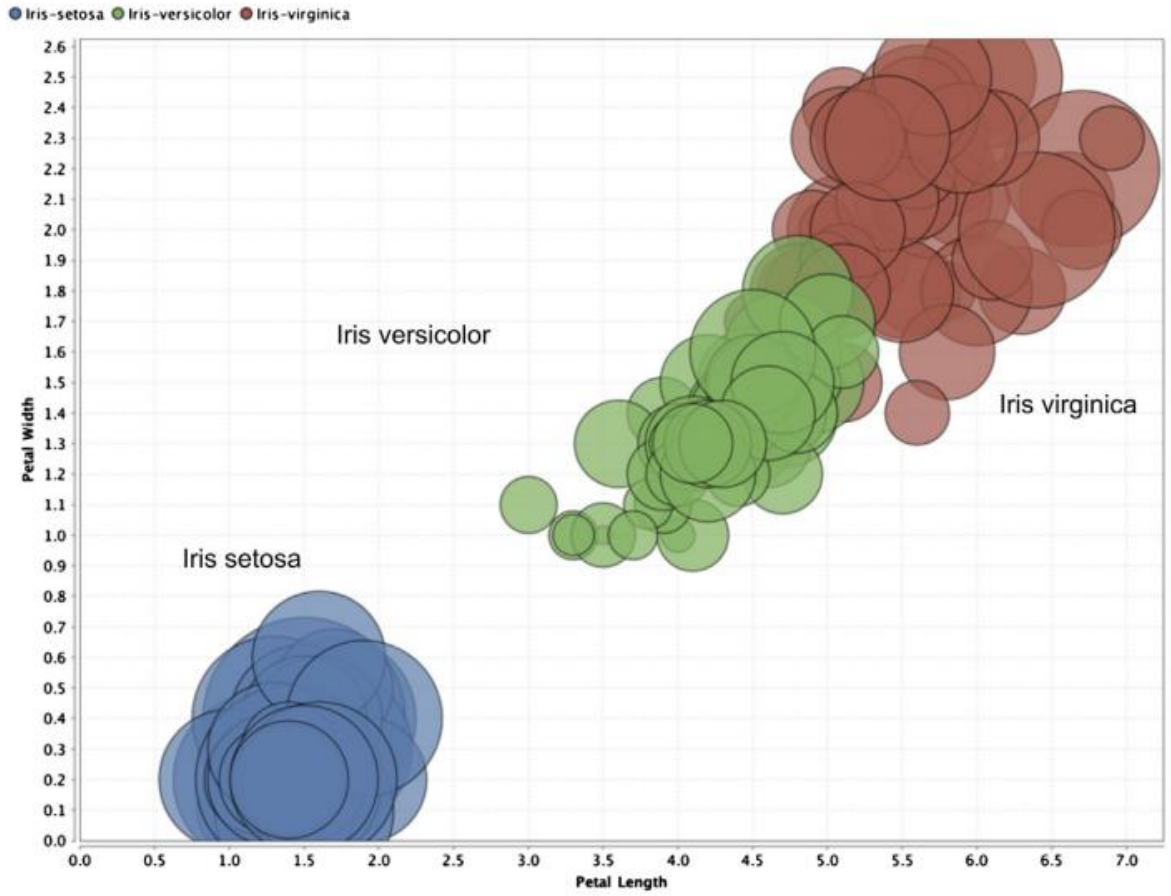
A scatter matrix displays all combinations of attributes with individual scatterplots and arranging these plots in a matrix.



Scatter matrices provide an effective visualization of comparative, multivariate, and high-density data.

Bubble Chart

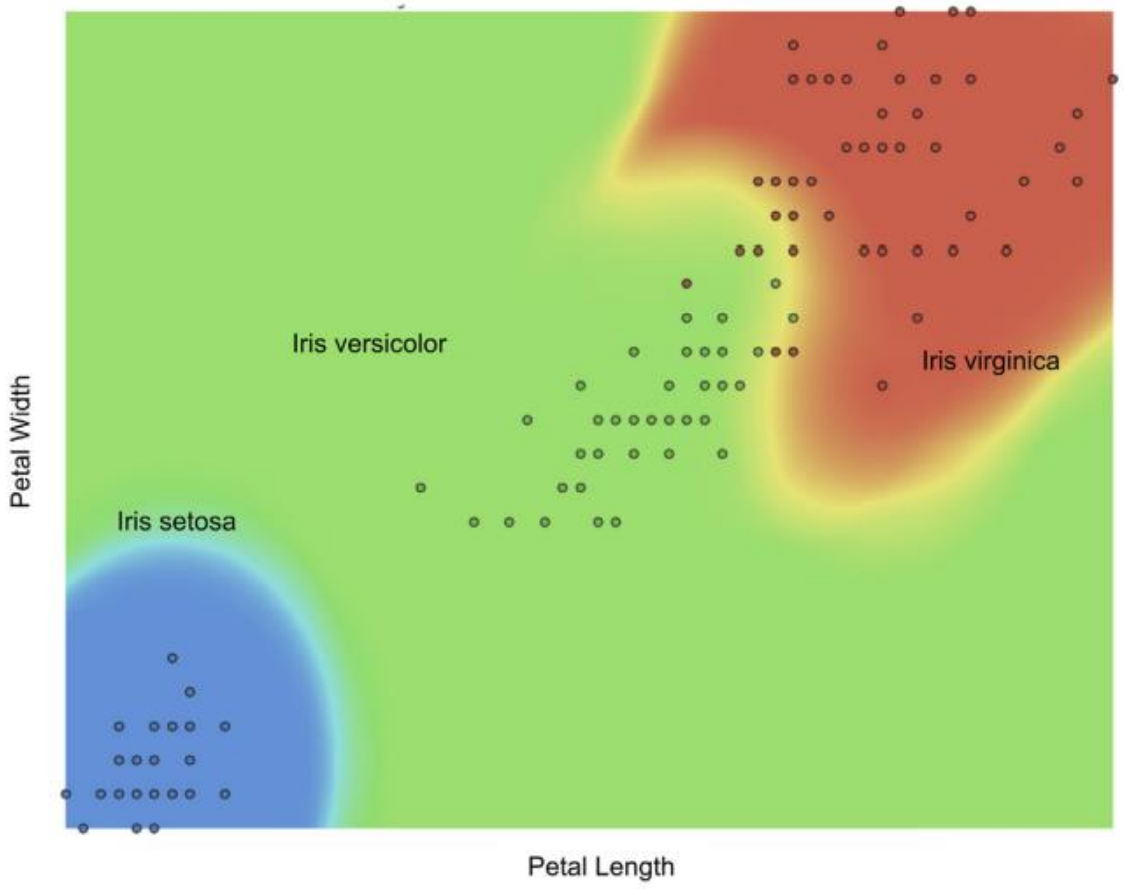
A bubble chart is a variation of a simple scatterplot with the addition of one more attribute, which is used to determine the size of the data point.



In this graph, petal length and petal width are used for x and y-axis, and sepal width is used for the size of the data point. The color of the data point represents a species class label

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 graph, petal length is used for the x-axis, petal width for the y-axis, sepal width for the background color, and class label for the data point color.