3. Data Exploration

Data exploration

- The word “data” is derived from the Latin word dare, which means “something given”— an observation or a fact about a subject.
- Data exploration helps with understanding data better, to prepare the data in a way that makes advanced analysis possible, and sometimes to get the necessary insights from the data faster than using advanced analytical techniques

Data exploration

- 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 of Data Exploration

Understanding data : Data exploration provides a high-level overview of each attribute (also called variable) in the dataset and the interaction between the attributes

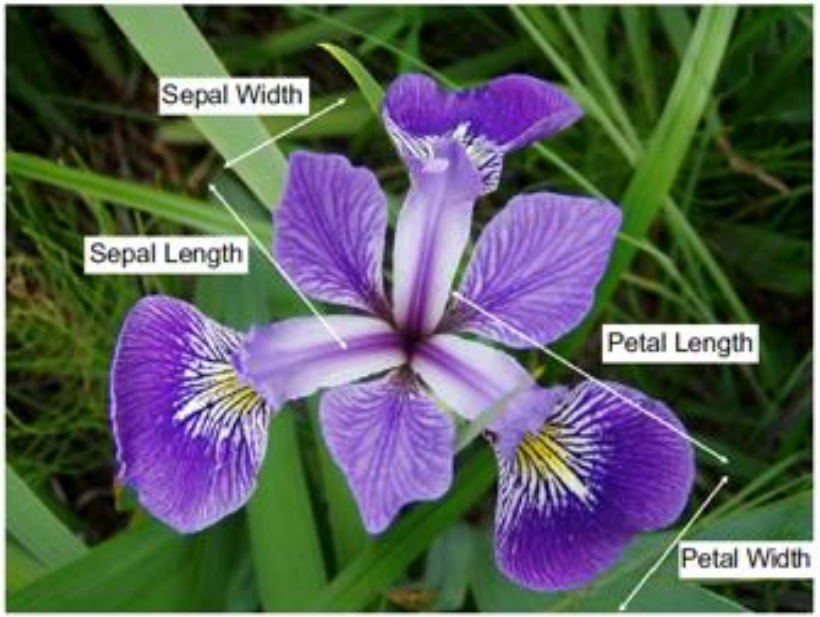
Data preparation : Before applying the data science algorithm, the dataset has to be prepared for handling any of the anomalies that may be present in the data.

Data mining tasks : Basic data exploration can sometimes substitute the entire data science process. For example, scatterplots can identify clusters in low-dimensional data

Interpreting data mining results : data exploration is used in understanding the prediction, classification, and clustering of the results of the data science process

Data Sets

Throughout the rest of this chapter (and the book) a few classic datasets, which are simple to understand, easy to explain, and can be used commonly across many different data science techniques



1http://commons.wikimedia.org/wiki/File:Iris_versicolor_3.jpg#mediaviewer/File:Iris_versicolor_3.jpg

Descriptive Statistics - Univariate

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

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

Descriptive Statistics - Univariate

Types of Data

- 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.
- Categorical or Nominal: Categorical data types are attributes treated as distinct symbols or just names. The colour of the iris of the human eye is a categorical data type.

Descriptive Statistics - Univariate

1: Univariate Exploration : Univariate data exploration denotes analysis of one attribute at a time

Eg: 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.

Descriptive Statistics - Univariate

- 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.

2.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
- 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)

Descriptive Statistics - Multivariate

- 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**

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 be plotted in a graph (although plotting more than three dimensions in a visual graph can be challenging).

Descriptive Statistics - Multivariate

- **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 <= r <= 1.

Descriptive Statistics - Multivariate

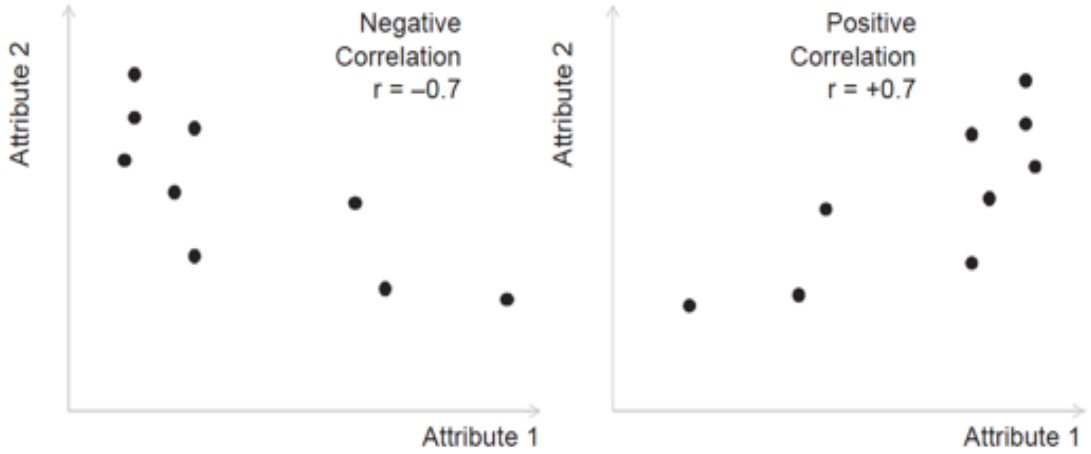
Central datapoint

Correlation

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

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

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



Descriptive Statistics - Multivariate

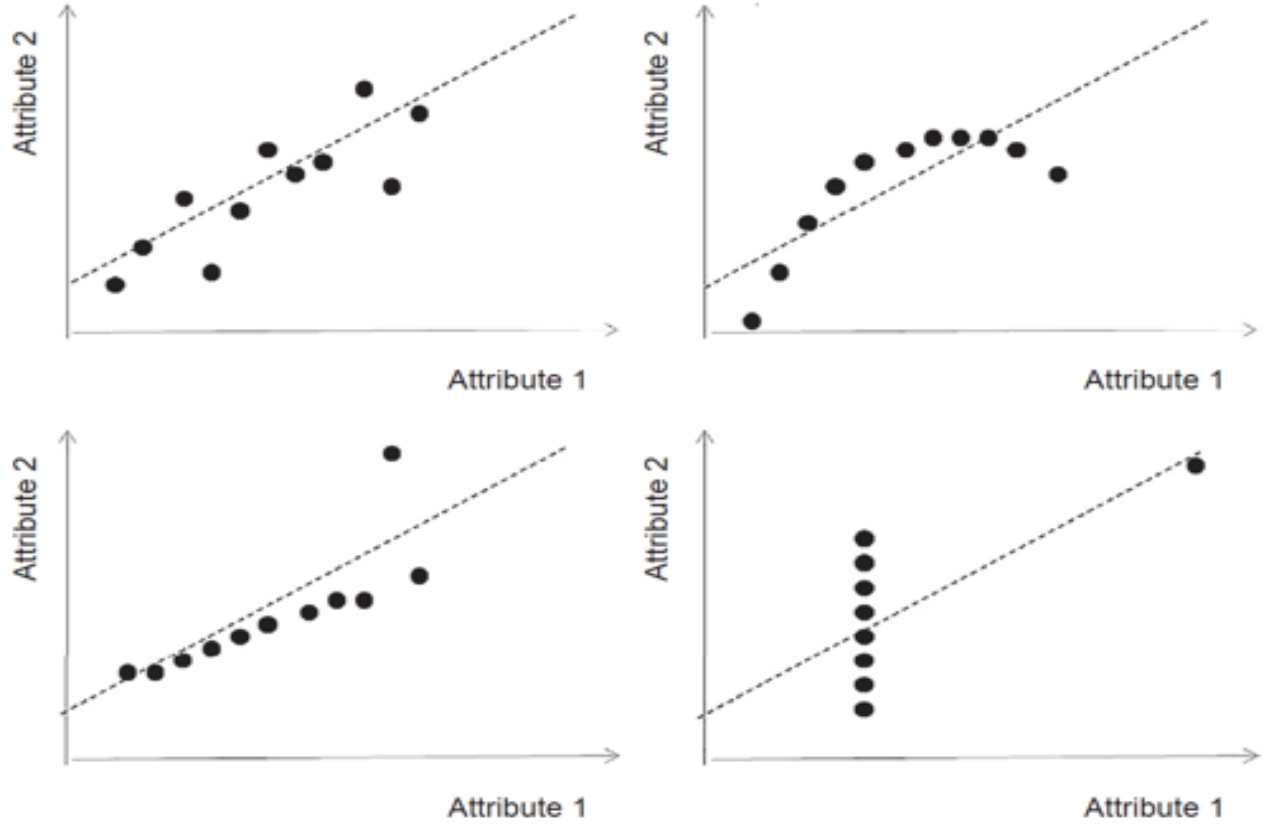


FIGURE 3.4

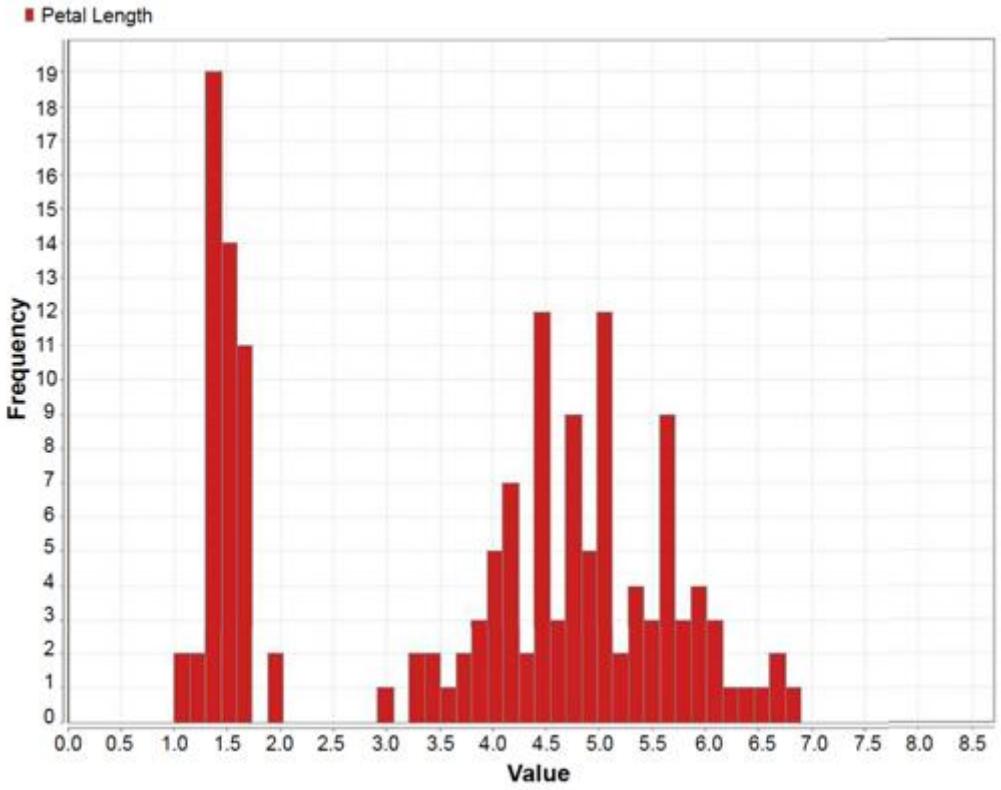
Data Visualization

- Visualizing data is one of the most important techniques of data discovery and exploration
- Comprehension of dense information: A simple visual chart can easily include thousands of data points. By using visuals, the user can see the big picture, as well as longer term trends that are extremely difficult to interpret purely by expressing data in numbers
- Relationships: Visualizing data in Cartesian coordinates enables exploration of the relationships between the attributes.

Data Visualization - Univariate

Histogram

It shows the distribution of the data by plotting the frequency of occurrence in a range. In a histogram, the attribute under inquiry is on the horizontal axis and the frequency of occurrence is on the vertical axis.



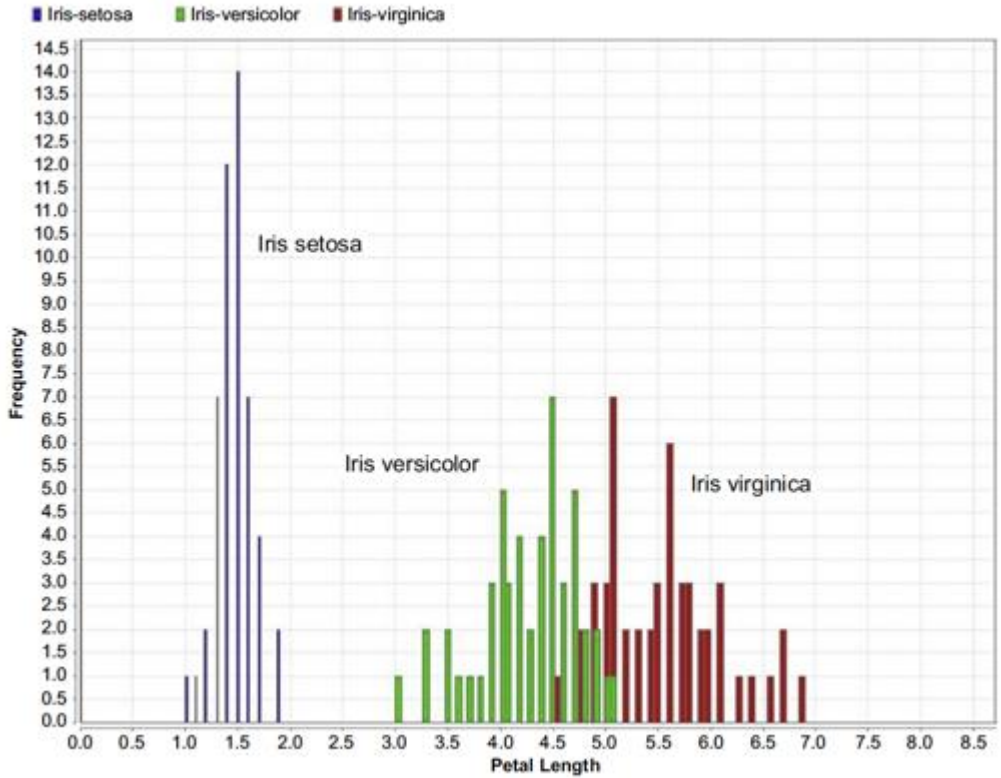
Data Visualization - Univariate

- One attribute can also be selected—petal length—and explored further using quartile charts by introducing a class label.
- In the plot in Fig. 3.8, we can see the distribution of three species for the petal length measurement. Similar to the previous comparison, the distribution of multiple species can be compared

Data Visualization

Class stratified Histogram

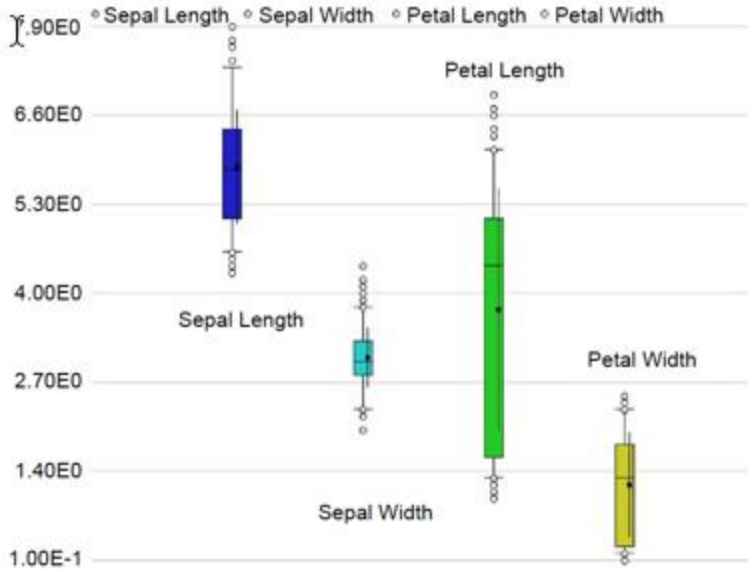
. A histogram can be stratified
to include different classes
in order to gain more insight



Data Visualization

Quantile plot

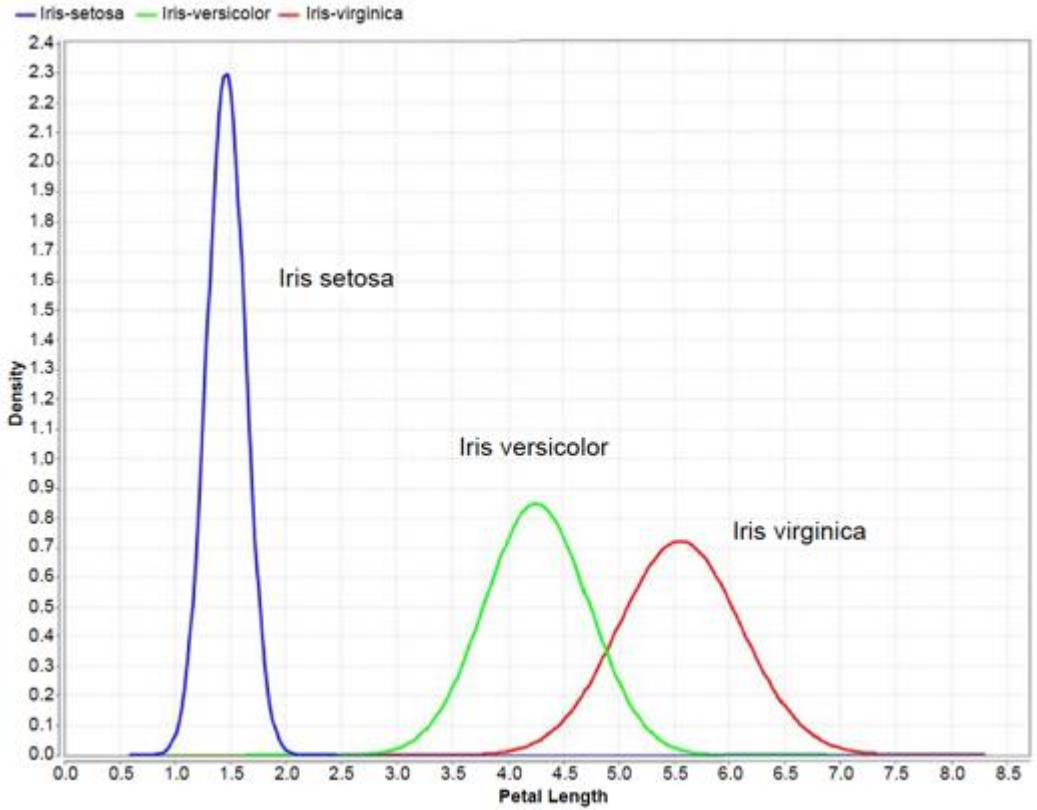
A box whisker plot is a simple visual way of showing the distribution of a continuous variable with information such as quartiles, median, and outliers, overlaid by mean and standard deviation.



Data Visualization

Distribution plot

For continuous numeric attributes like petal length, instead of visualizing the actual in the sample,its normal Distribution function can be visualized

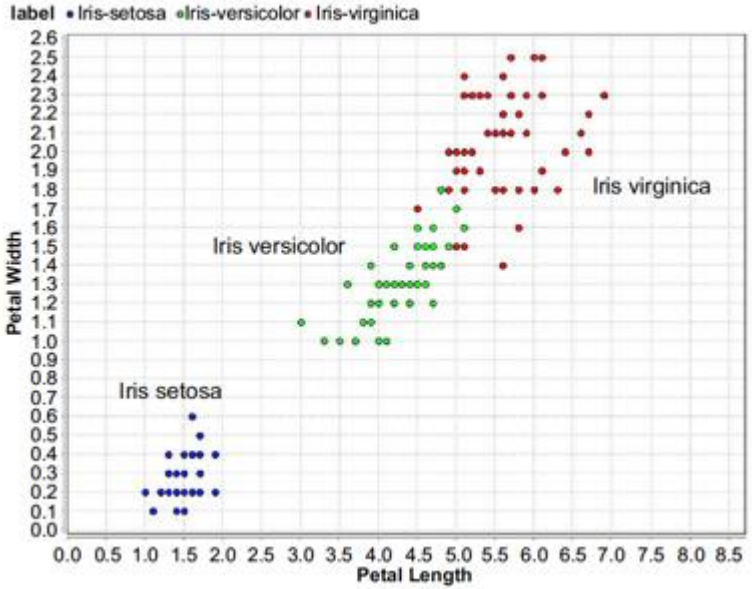


Data Visualization multivariate

Scatter plot

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

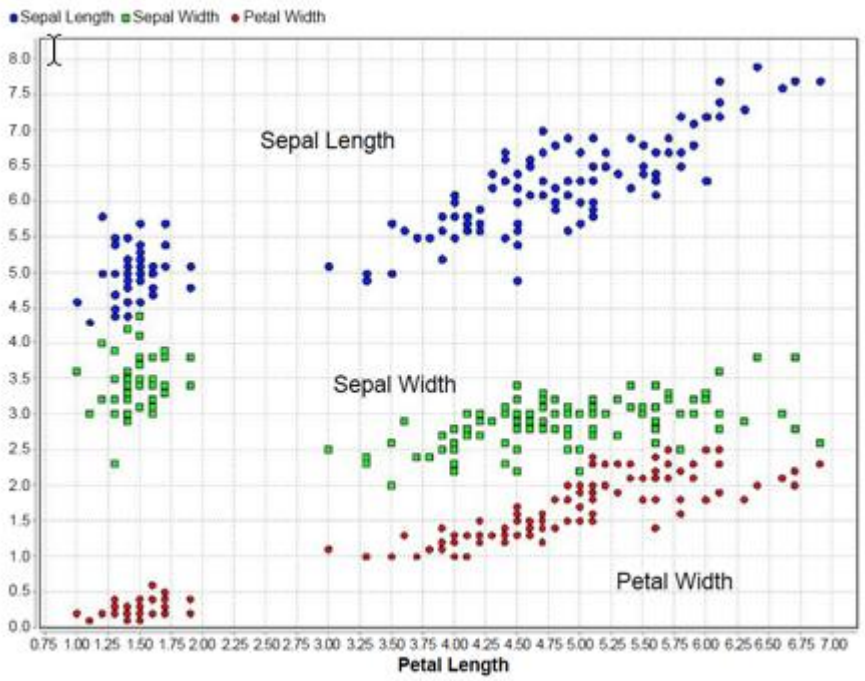
the dataset aligned with the coordinates. The attributes are of continuous data type



Data Visualization

Scatter multiple

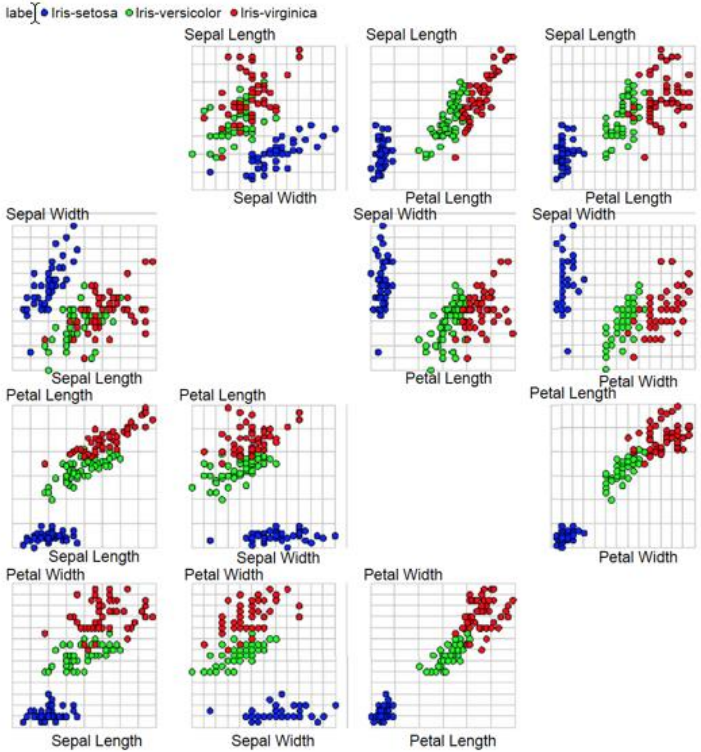
A scatter multiple is an enhanced form of a simple scatterplot where more than two dimensions can be included in the chart and studied simulta neously. The primary attribute is used for the x-axis coordinate. The secondary axis is shared with more attributes or dimensions.



Data Visualization

Multiple Scatter matrix

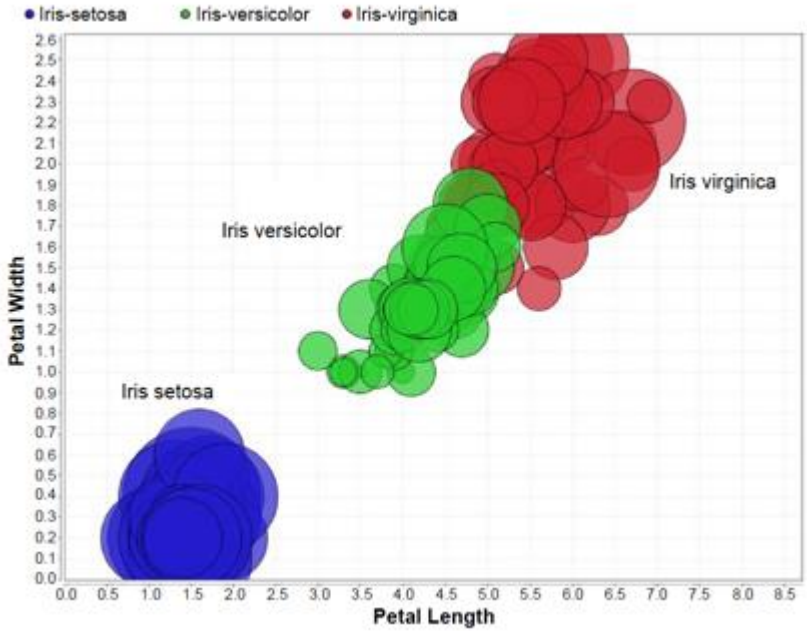
If the dataset has more than two attributes, it is important to look at combi nations of all the attributes through a scatterplot. A scatter matrix solves this need by comparing all combinations of attributes with individual scatterplots and arranging these plots in a matrix.



Data Visualization

Bubble plot

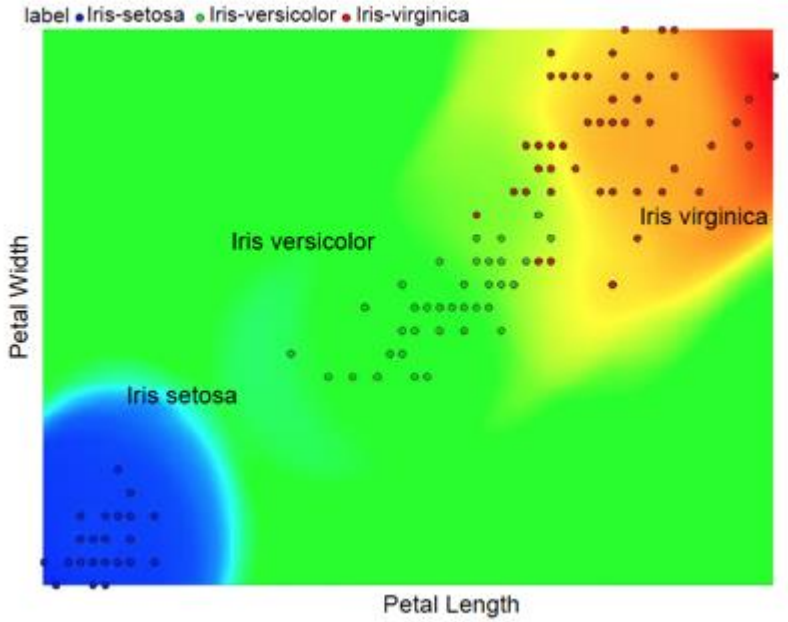
A bubble chart is a variation of a simple scatterplot with the addition of one more attribute, which is used to determine the size of the data point.



Data Visualization

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 visual ize one dimension, and hence, a total of four dimensions can be visualized in a density chart



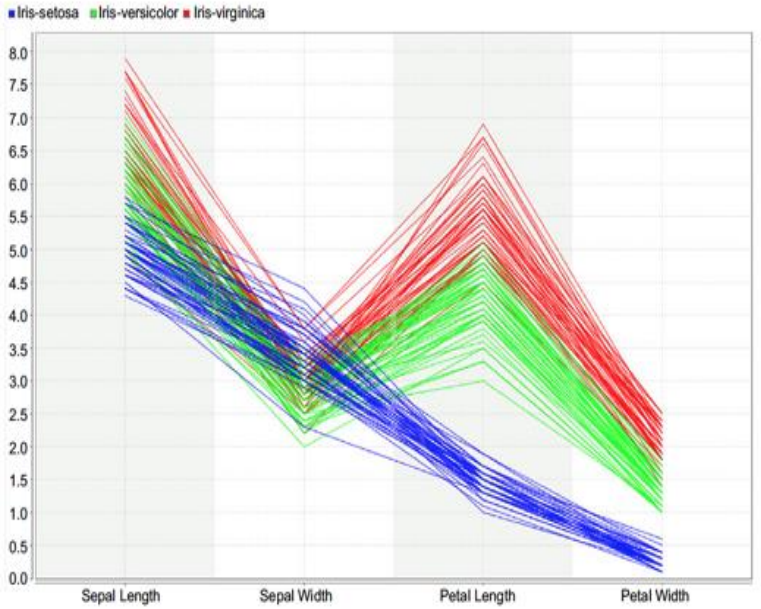
Data Visualization

- Visualizing more than three attributes on a two-dimensional medium (like a paper or screen) is challenging. This limitation can be overcome by using transformation techniques to project the high-dimensional data points into parallel axis space. In this approach, a Cartesian axis is shared by more than one attribute

Data Visualization

Parallel chart

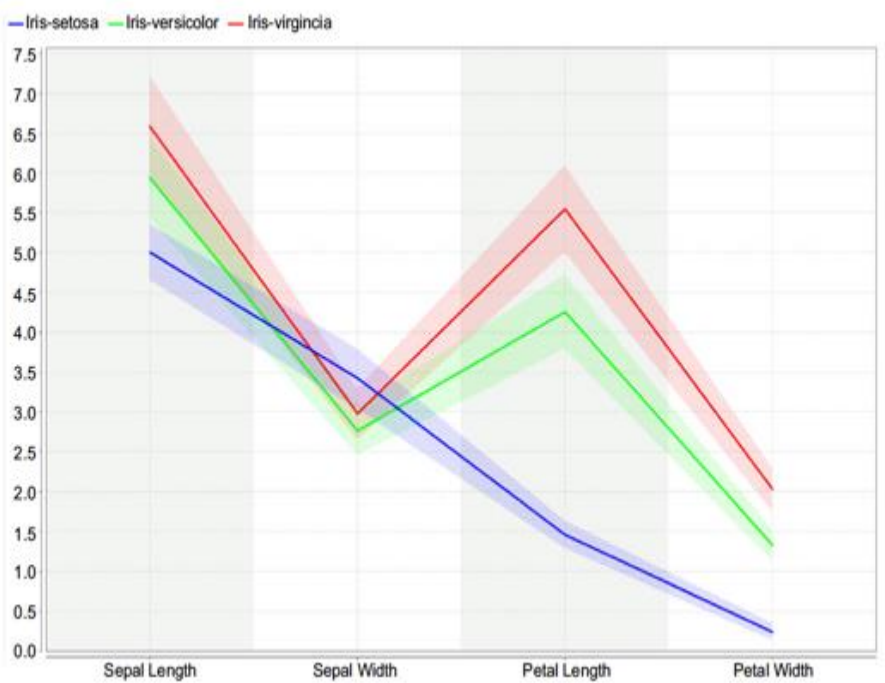
A parallel chart visualizes a data point quite innovatively by transforming or projecting multi-dimensional data into a two-dimensional chart medium. In this chart, every attribute or dimension is linearly arranged in one coordinate (x-axis) and all the measures are arranged in the other coordinate (y-axis). Since the x-axis is multivariate, each data point is represented as a line in a parallel space.



Data Visualization

Deviation chart

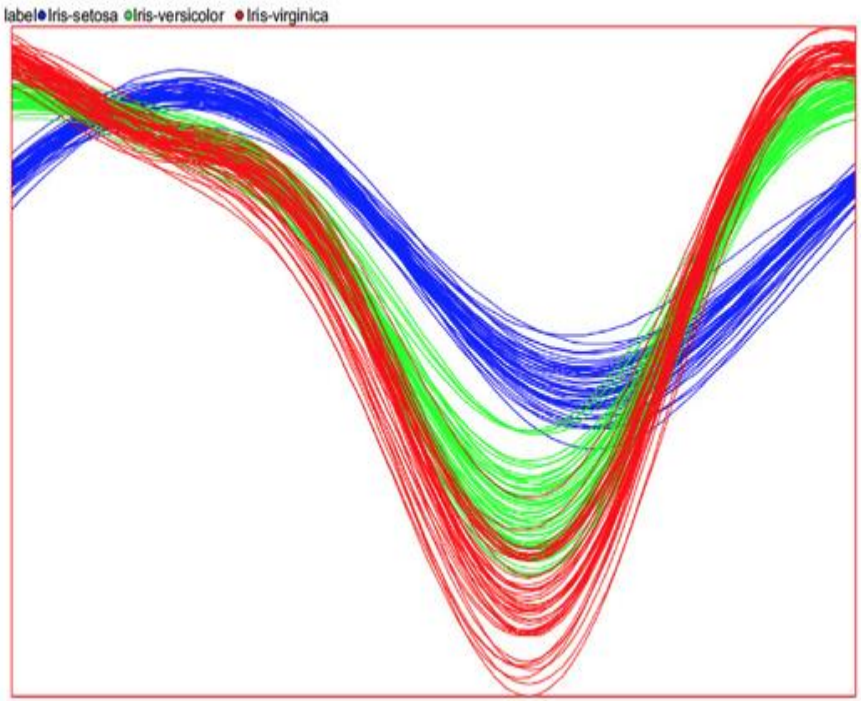
Instead of plotting all data lines, deviation charts only show the mean and standard deviation statistics.



Data Visualization

Andrews curves

An Andrews plot belongs to a family of visualization techniques where the high-dimensional data are projected into a vector space so that each data point takes the form of a line or curve.



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

- Andrews curves

In an Andrews plot, each data point X with d dimensions, X 5 (x1, x2, x3, ..., xd), takes the form of a Fourier series:

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

This function is plotted for -π < t < π for each data point. Andrews plots are useful to determine if there are any outliers in the data and to identify potential patterns within the data points

Roadmap for data exploration

1. Organize the data set
2. Find the central point for each attribute:
3. Understand the spread of the attributes:
4. Visualize the distribution of each attributes:
5. Pivot the data:
6. Watch out for outliers:
7. Understanding the relationship between attributes:
8. Visualize the relationship between attributes:
9. Visualization high dimensional data sets:

Kotu, V., & Deshpande, B. (2014). *Predictive analytics and data mining: concepts and practice with rapidminer*. Morgan Kaufmann.