Data Exploration

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, also known as exploratory data analysis, provides a set of tools to obtain fundamental understanding of a dataset.

The results of data exploration can be extremely powerful in

- Grasping the structure of the data
- Distribution of the values
- Presence of extreme values and
- The interrelationships between the attributes in the 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

- ✓ **Descriptive statistics**
- ✓ **Data visualization**.

Descriptive statistics is the process of converting 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. Data exploration in the context of data science uses both descriptive statistics and visualization techniques.

2.1 OBJECTIVES OF DATA EXPLORATION

1. Data understanding: Data exploration provides a high-level overview of each attribute (also called variable) in the dataset and the interaction between the attributes. Data exploration helps answers the questions like what is the typical value of an attribute or how much do the data points differ from the typical value, or presence of extreme values.

2. 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. Some data science algorithms do not work well when input attributes are correlated with each other. Thus, correlated attributes need to be identified and removed.

3. Data science tasks: Basic data exploration can sometimes substitute the entire data science process. For example, scatterplots can identify clusters in low-dimensional data or can help develop regression or classification models with simple visual rules.

4. Interpreting the results: Finally, data exploration is used in understanding the prediction, classification, and clustering of the results of the data science process. Histograms help to comprehend the distribution of the attribute and can also be useful for visualizing numeric prediction, error rate estimation, etc.

2.2 DATASETS

A data set is an ordered collection of data. collection of information obtained through observations, measurements, study, or analysis is referred to as data. It could include information such as facts, numbers, figures, names, or even basic descriptions of objects.

The most popular datasets used to learn data science is probably the Iris dataset, introduced by Ronald Fisher. The Iris dataset is used for learning data science mainly because it is simple to understand, explore, and can be used to illustrate how different data science algorithms approach the problem on the same standard dataset.

Iris is a flowering plant that is found widely, across the world. The genus of Iris contains 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 I. versicolor**, with **50 observations** each. Each observation consists of four attributes: **sepal length, sepal width, petal length, and petal width**. The petals are the brightly colored inner part of the flowers and the sepals form the outer part of the flower and are usually green in color. However, in an Iris flower, both sepals and petals are bright purple in color, but can be distinguished from each other by differences in the shape All four attributes in the Iris dataset are numeric continuous values measured in centimeters. One of the species, I. setosa, can be easily distinguished from the other two using simple rules like the petal length is less than 2.5 cm. Separating the virginica and versicolor classes requires more complex rules that involve more attributes. This

dataset and other datasets used in this book can be accessed from the book companion website: www.IntroDataScience.com

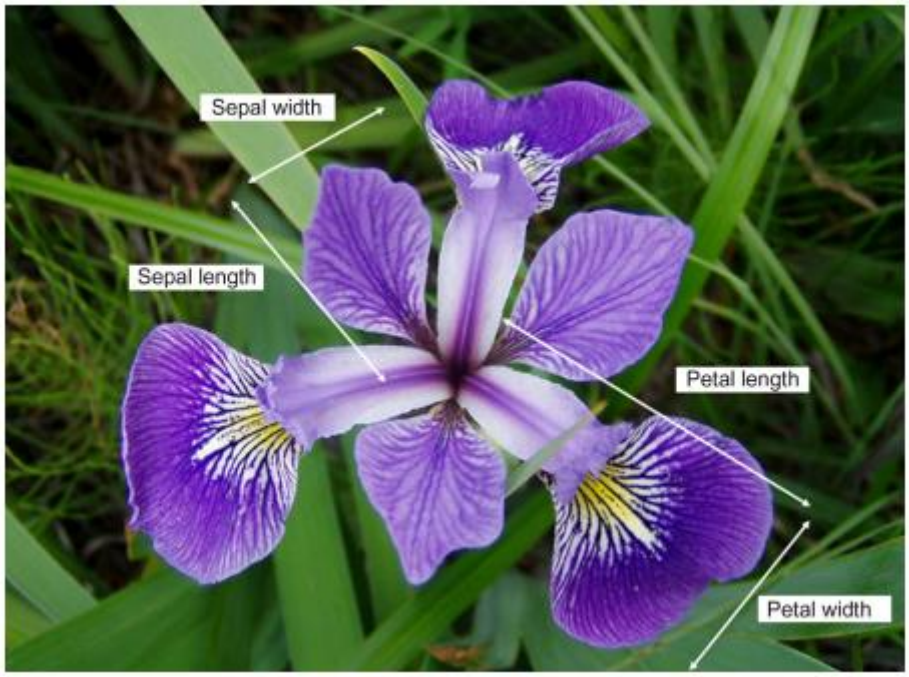


FIGURE 2.1 Iris versicolor

2.2.1 Types of Data

Data come in different formats and types. Understanding the properties of each attribute or feature 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 (31^0C, 33.3^0C) or Fahrenheit (100^0F, 101.45^0F) or on the Kelvin scale

● Ordered labels as in hot, mild, or cold

● Number of days within a year below 0^0C (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.

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. Values are ordered and calculating the difference between the values makes sense. Hence, additive and subtractive mathematical operations and logical comparison operators like greater than, less than, and equal to, operations can be applied.

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. Usually, they denote a count of something, number of days with temperature less than 0^0C, number of orders, number of children in a family, etc.

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. Along with additive and logical operations, ratio operations can be performed with this data type. Both integer and ratio data types are categorized as a numeric data type in most data science tools.

2. 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. There is no direct relationship among the data values, and hence, mathematical operators except the logical or "is equal" operator cannot be applied. They are also called a nominal or polynominal data type.

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 or average underlying numerical scores like 400, 500, 600, and 700 (scores here are just an example). In this type conversion, there is no loss of information. However, conversion from numeric credit score to categories (poor, average, good, and excellent) does incur loss of information.

2.3 DESCRIPTIVE STATISTICS

Descriptive statistics refers to the study of the aggregate quantities of a dataset. In general, descriptive analysis covers the following characteristics of the sample or population dataset.

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.

2.3.1 Univariate Exploration

Univariate data exploration denotes analysis of one attribute at a time. The example Iris dataset for one species, I. setosa, has 50 observations and 4 attributes, as shown in Table. Here some of the descriptive statistics for sepal length attribute are explored.

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

a) 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. 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.

In an attribute, the mean, median, and mode may be different numbers, and this indicates the shape of the distribution. If the dataset has outliers, the mean will get affected while in most cases the median will not. The mode of the distribution can be different from the mean or median, if the underlying dataset has more than one natural normal distribution.

b) Measure of Spread

In desert regions, it is common for the temperature to cross above 110^0F during the day and drop below 30^0F during the night while the average temperature for a 24-hour period is around 70^0F. Obviously, the experience of living in the desert is not the same as living in a tropical region with the same average daily temperature around 70^0F, where the temperature within the day is between

60^0F and 80^0F. What matters here is not just the central location of the temperature, but the spread of the temperature. 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 it 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.

In the example, the range for the temperature in the desert is 80^0F and the range for the tropics is 20^0F. The desert region experiences larger temperature swings as indicated by the range.

- **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 ($\mathbf{x}_i$) and the mean of the sample ($\boldsymbol{\mu}$).

The variance is the sum of the squared deviations of all data points divided by the number of data points.

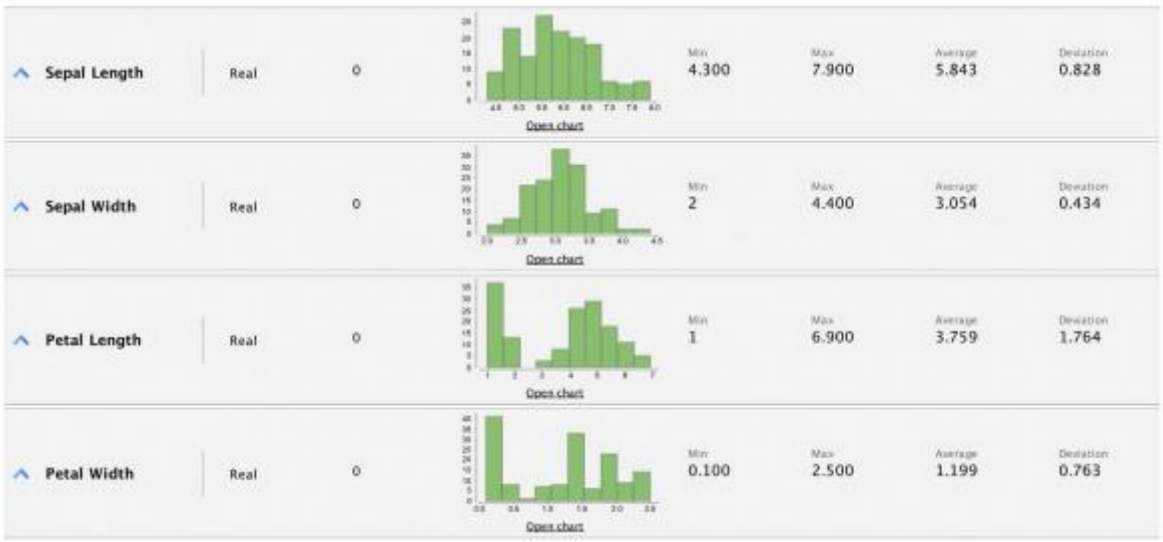
For a dataset with N observations, the variance is given by the following equation:

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

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. 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. Fig. 3.2 provides the univariate summary of the Iris dataset with all 150 observations, for each of the four numeric attributes.



2.3.2 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, which is central to data science methods.

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.

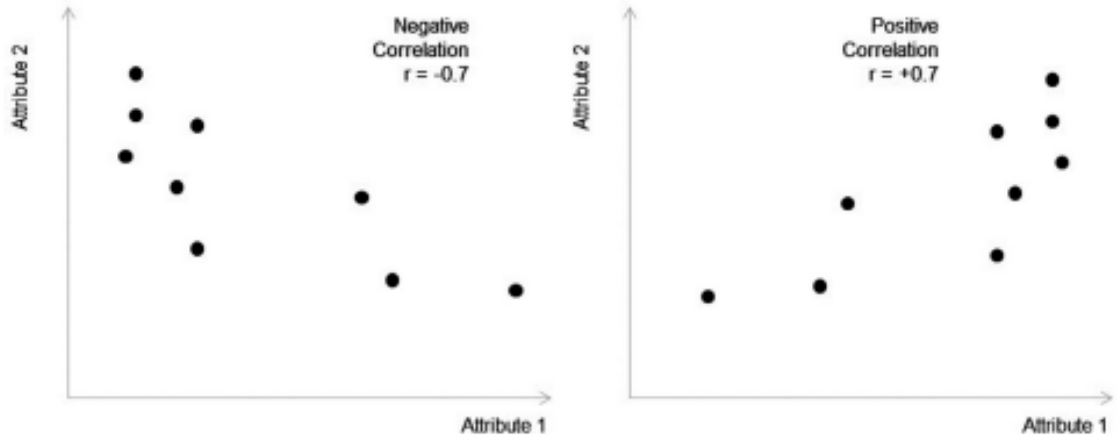
For example, consider average temperature of the day and ice cream sales. Statistically, the two attributes that are correlated are dependent on each other and one may be used to predict the other. However, correlation between two

attributes does not imply causation, that is, one doesn't necessarily cause the other. The ice cream sales and the shark attacks are correlated, however there is no causation. Both ice cream sales and shark attacks are influenced by the third attribute—the summer season. Generally, ice cream sales spikes as temperatures rise. As more people go to beaches during summer, encounters with sharks become more probable.

Correlation between two attributes is commonly measured by the Pearson correlation coefficient (r), which measures the strength of linear dependence (Fig. 3.3). 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, with perfect correlation at 1 or -1.

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



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

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

where Sx and Sy are the standard deviations of random variables x and y, respectively.

The Pearson correlation coefficient has some limitations in quantifying the strength of correlation. When datasets have more complex nonlinear relationships like quadratic functions, only the effects on linear relationships are considered and quantified using correlation coefficient. The presence of outliers can also skew the measure of correlation. Visually, correlation can be observed using scatterplots with the attributes in each Cartesian coordinate (Fig. 3.3). In fact, visualization should be the first step in understanding correlation because it can identify nonlinear relationships and show any outliers in the dataset. Anscombe's quartet clearly illustrates the limitations of relying only on the correlation coefficient to understand the data (Fig. 3.4). The quartet consists of four different datasets, with two attributes (x, y). All four datasets have the same mean, the same variance for x and y, and the same correlation coefficient between x and y, but look drastically different when plotted on a chart. This evidence illustrates the necessity of visualizing the attributes instead of just calculating statistical metrics.

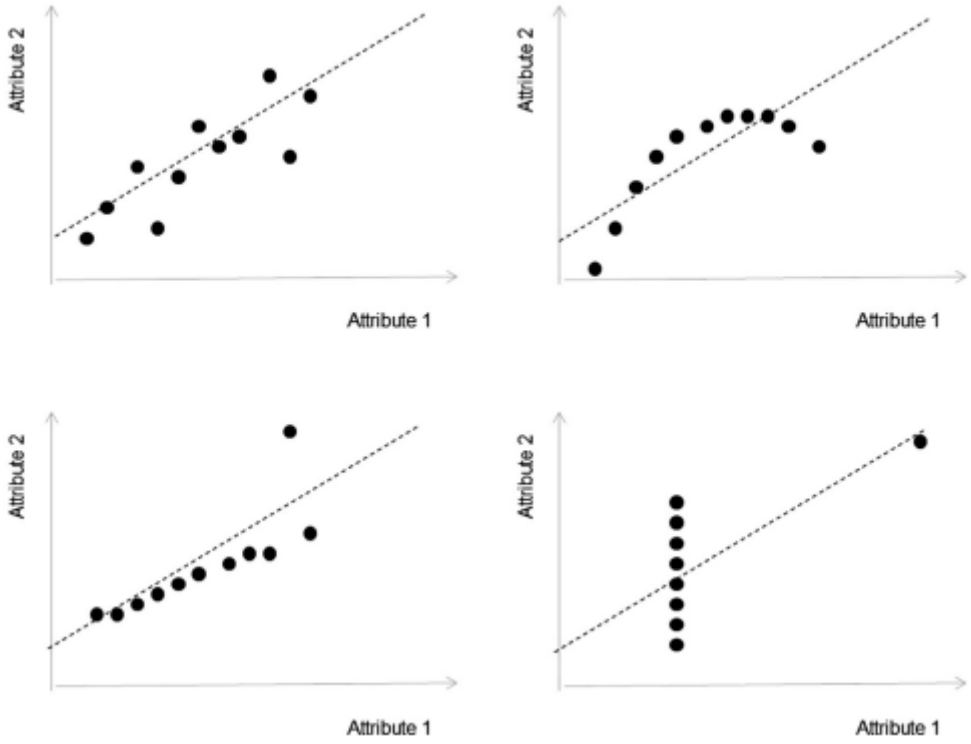


FIGURE 3.4 Anscombe's Quartet: descriptive statistics versus visualization.

2.4 DATA VISUALIZATION

Visualizing data is one of the most important techniques of data discovery and exploration. Though visualization is not considered a data science technique, terms like visual mining or pattern discovery based on visuals are increasingly used in the context of data science, particularly in the business world. The visual representation of data provides easy comprehension of complex data with multiple attributes and their underlying relationships. The motivation for using data visualization includes:

● **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.

Vision is one of the most powerful senses in the human body. As such, it is intimately connected with cognitive thinking. Human vision is trained to discover patterns and anomalies even in the presence of a large volume of data. However, the effectiveness of the pattern detection depends on how effectively the information is visually presented.

Visualization techniques are also categorized into:

- **univariate visualization**
- **multivariate visualization and**
- **visualization of a large number of attributes using parallel dimensions.**

2.4.1 Univariate Visualization

Visual exploration starts with investigating one attribute at a time using univariate charts. The techniques discussed in this section give an idea of how the attribute values are distributed and the shape of the distribution.

a. 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. For a continuous numeric data type, the range or binning value to group a range of values need to be specified.

There is no optimal number of bins or bin width that works for all the distributions. If the bin width is too small, the distribution becomes more precise but reveals the noise due to sampling. A general rule of thumb is to have a number of bins equal to the square root or cube root of the number of data points.

Histograms are used to find the central location, range, and shape of distribution. In the case of the petal length attribute in the Iris dataset, the data is multimodal (Fig. 3.5), where the distribution does not follow the bell curve pattern. Instead, there are two peaks in the distribution. This is due to the fact that there are 150 observations of three different species (hence, distributions) in the dataset.

The enhanced histogram with class labels shows the dataset is made of three different distributions (Fig. 3.6). I. setosa's distribution stands out with a mean

around 1.25 cm and ranges from 12 cm. I. versicolor and I. virginica's distributions overlap I. setosa's have separate means.

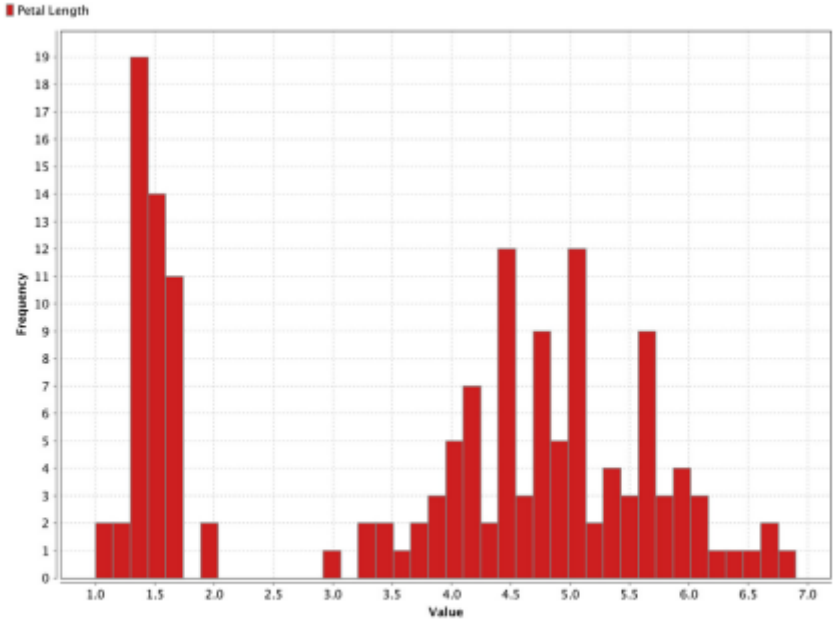


FIGURE 3.5
Histogram of petal length in Iris dataset.

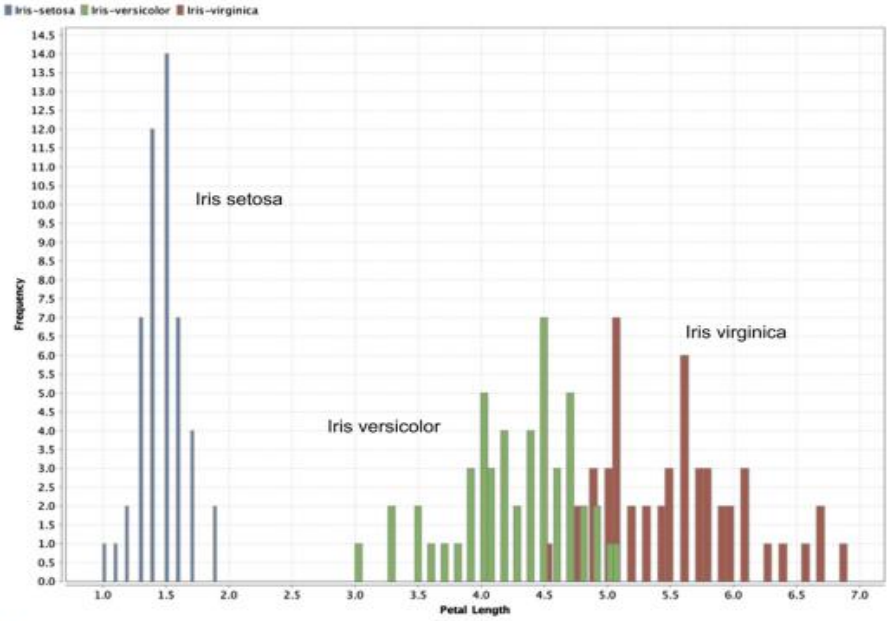


FIGURE 3.6
Class-stratified histogram of petal length in Iris dataset.

b. 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. In some cases, the mean point is denoted by a solid dot overlay followed by standard deviation as a line overlay.

Fig. 3.7 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.

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.

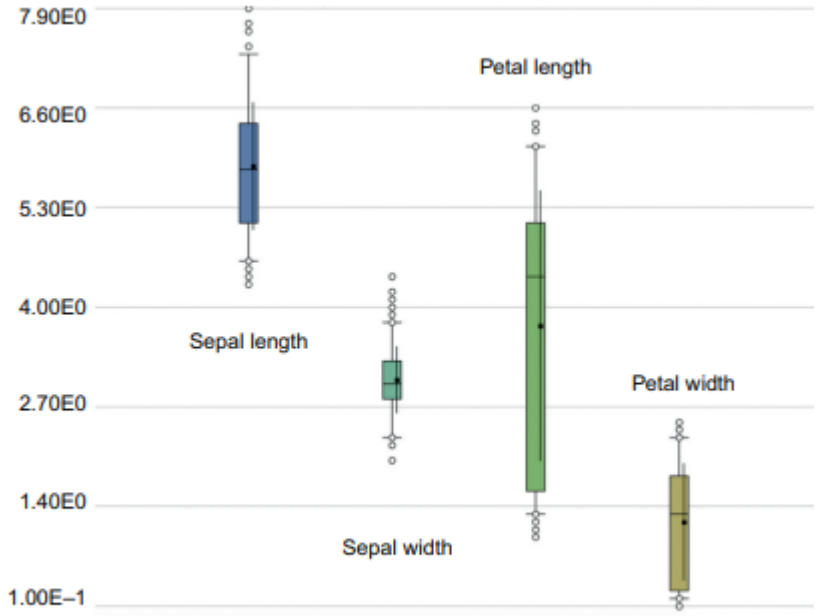


FIGURE 3.7
Quartile plot of Iris dataset.

c. 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. Here an inherent assumption is being made that the measurements of petal length (or any continuous variable) follow the normal distribution, and hence, its distribution can be visualized instead of the actual values. The normal distribution is also called the Gaussian distribution or “bell curve” due to its bell shape. The normal distribution fun shows the probability of occurrence of a data point within a range of values. 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. When the normal distribution curves are stratified by class type, more insight into the data can be gained. Fig. 3.9 shows the normal distribution curves for petal length measurement for each Iris species type. From the distribution chart, it can be inferred that the petal length for the I. setosa sample is more distinct and cohesive than I. versicolor and I. virginica. If there is an unlabeled measurement with a petal length of 1.5 cm, it can be predicted that the species is I. setosa. However, if the petal length measurement is 5.0 cm, there is no clear prediction, as the species could be either Iris versicolor and I. virginica.

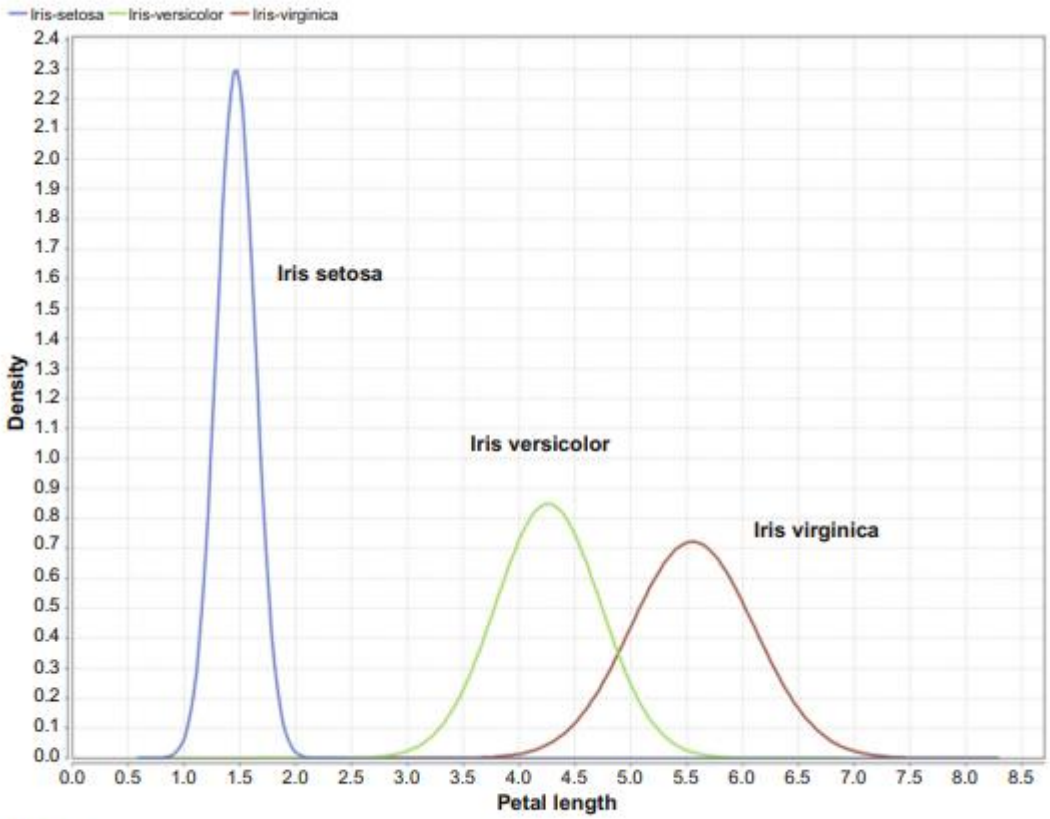


FIGURE 3.9
Distribution of petal length in Iris dataset.

3.4.2 Multivariate Visualization

The multivariate visual exploration considers more than one attribute in the same visual. The techniques discussed in this section focus on the relationship of one attribute with another attribute. These visualizations examine two to four attributes simultaneously

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 under inquiry. 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. This is particularly useful for low-dimensional datasets. Chapter 13: Anomaly detection, provides techniques for finding outliers in high-dimensional space.

Fig. 3.10 shows the scatterplot between petal length (x-axis) and petal width (y-axis). These two attributes are slightly correlated, because this is a measurement of the same part of the flower. When the data markers are colored to indicate different species using class labels, more patterns can be observed. There is a cluster of data points, all belonging to species I. setosa, on the lower left side of the plot. I. setosa has much smaller petals. This feature can be used as a rule to predict the species of unlabeled observations. One of the limitations of scatterplots is that only two attributes can be used at a time, with an additional

attribute possibly shown in the color of the data marker. However, the colors are usually reserved for class labels.

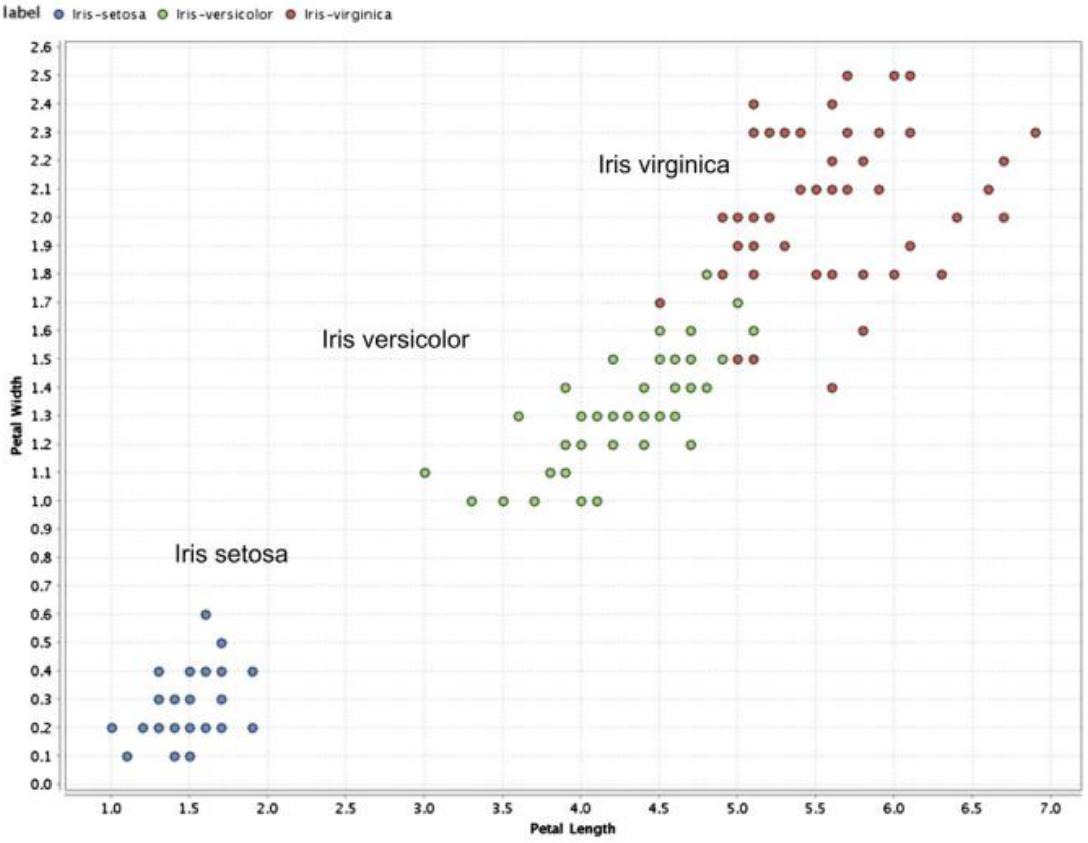


FIGURE 3.10
Scatterplot of Iris dataset.

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. In this example (Fig. 3.11), the values on the y-axis are shared between sepal length, sepal width, and petal width. The name of the attribute is conveyed by colors used in data markers. 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. Note that the data points are duplicated for each attribute in the y-axis. Data

points are color-coded for each dimension in 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.

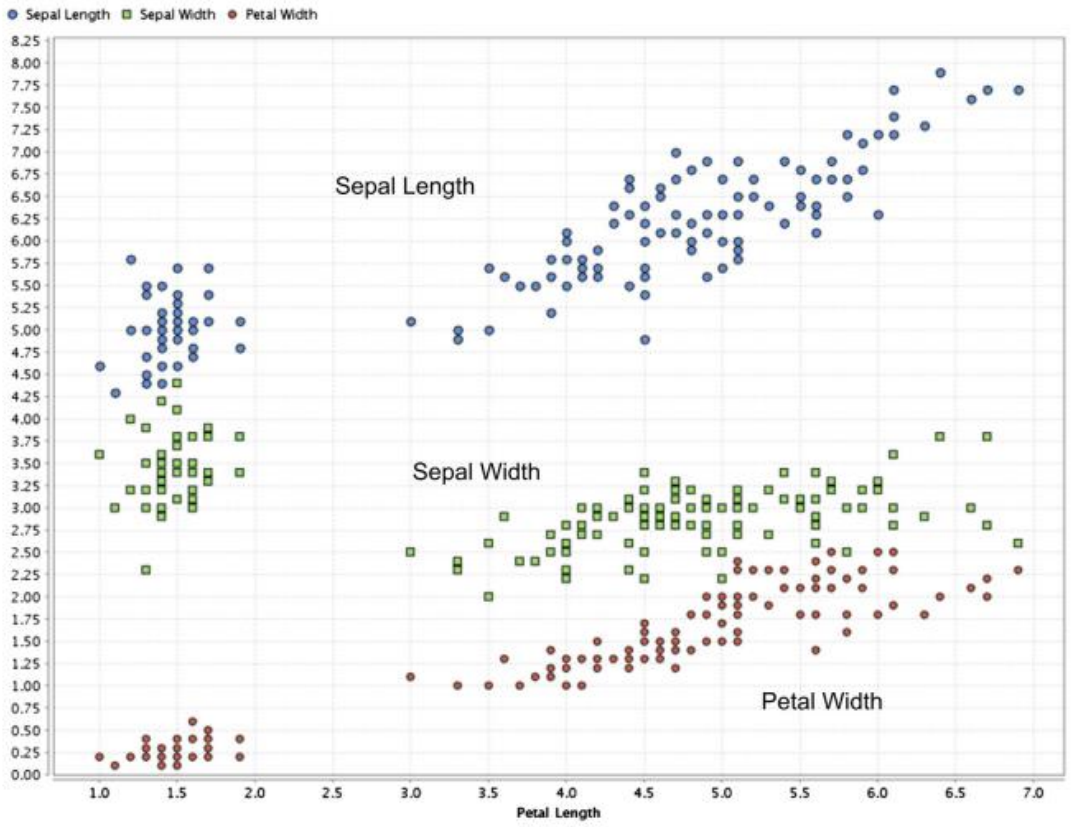


FIGURE 3.11
Scatter multiple plot of Iris dataset.

Scatter Matrix

If the dataset has more than two attributes, it is important to look at combinations 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.

A scatter matrix for all four attributes in the Iris dataset is shown in Fig. 3.12. The color of the data point is used to indicate the species of the flower. Since there are four attributes, there are four rows and four columns, for a total of 16 scatter charts. Charts in the diagonal are a comparison of the attribute with itself;

hence, they are eliminated. Also, the charts below the diagonal are mirror images of the charts above the diagonal. In effect, there are six distinct comparisons in scatter multiples of four attributes. Scatter matrices provide an effective visualization of comparative, multivariate, and high-density data displayed in small multiples of the similar scatterplots.

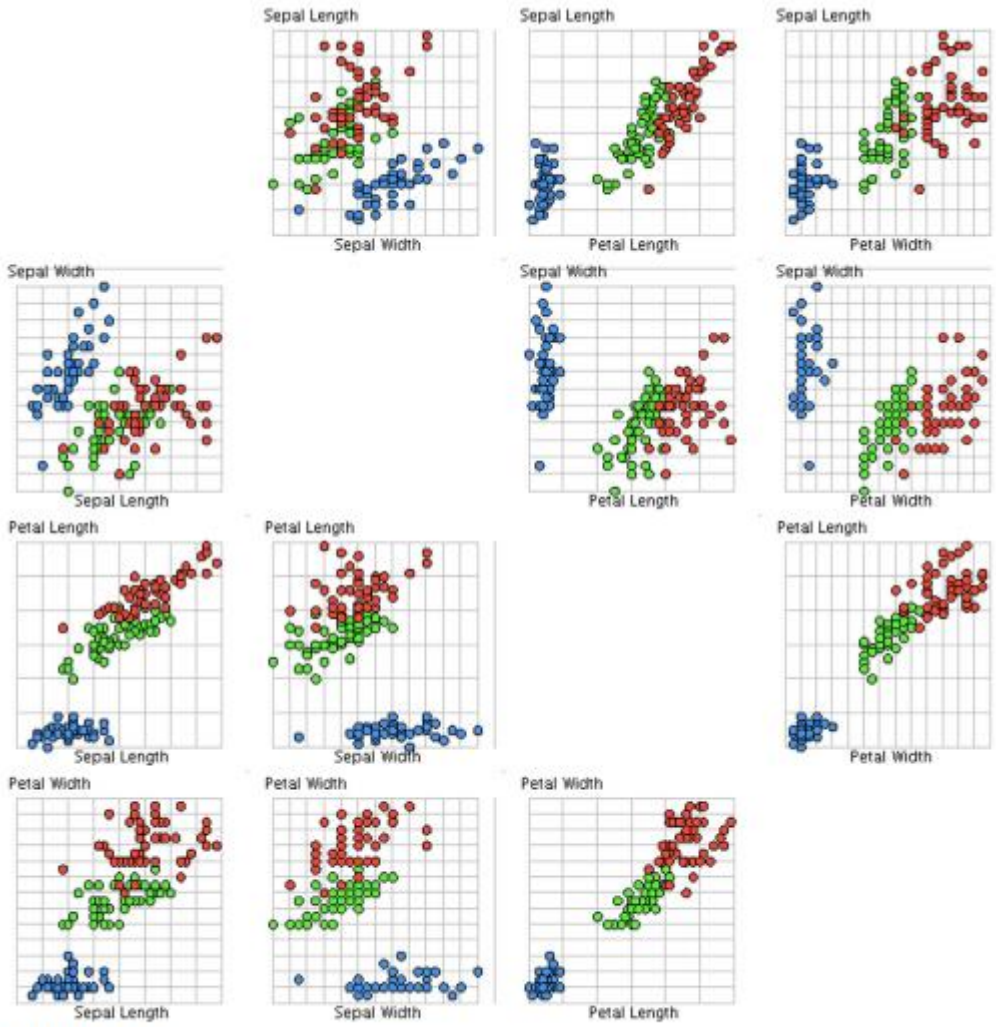


FIGURE 3.12
Scatter matrix plot of Iris dataset.

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 (Fig. 3.13).

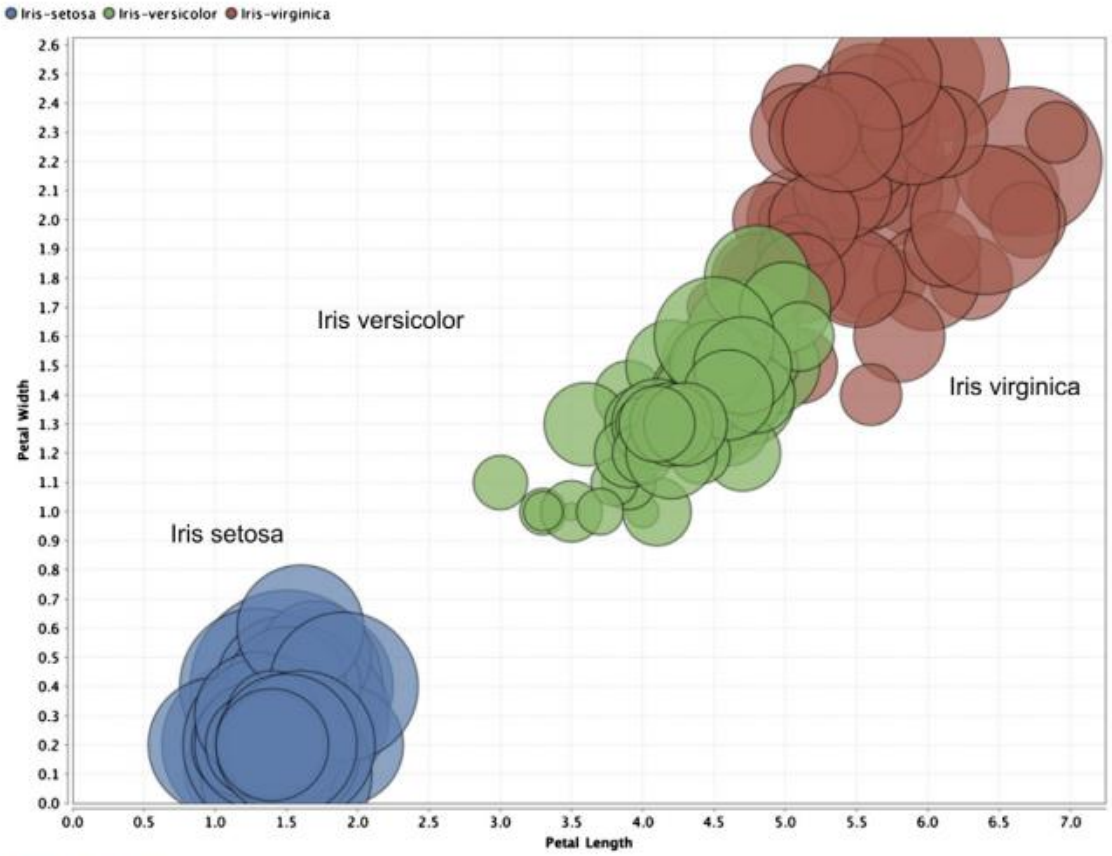


FIGURE 3.13
Bubble chart of Iris dataset.

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. 3.14, 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.

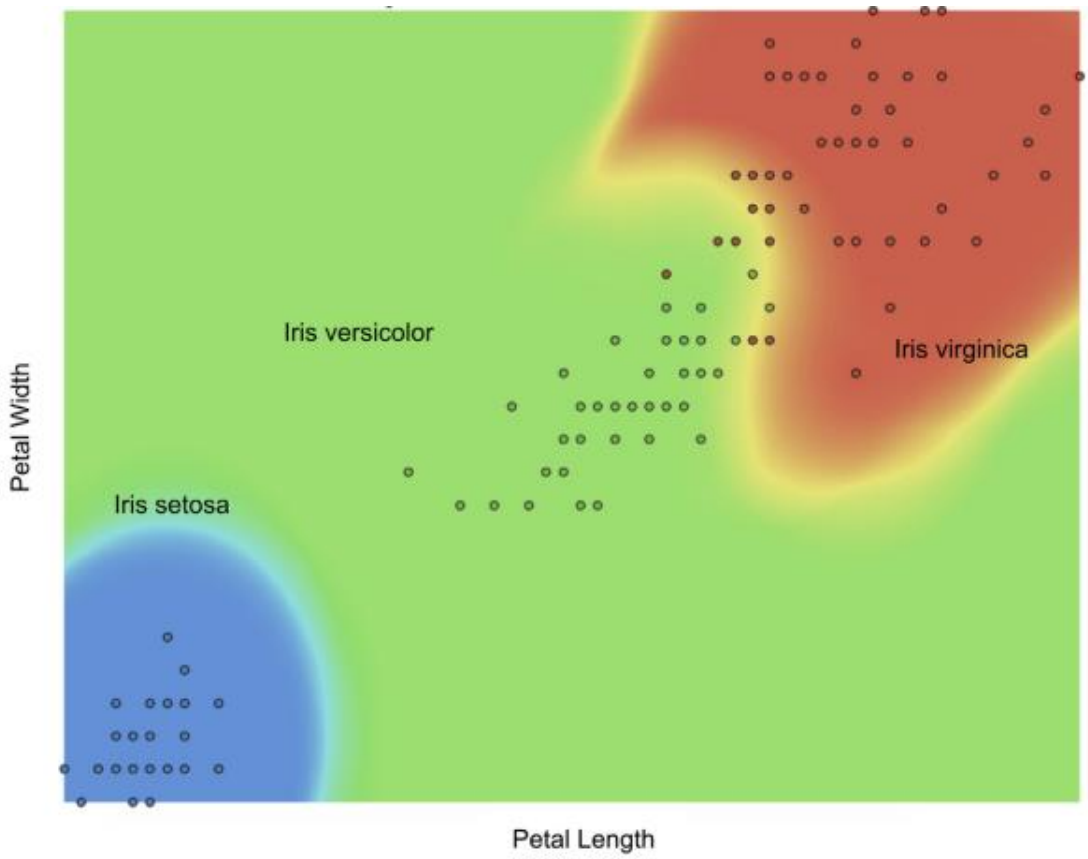


FIGURE 3.14
Density chart of a few attributes in the Iris dataset.

3.4.3 Visualizing High-Dimensional Data

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.

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. In the case of the Iris dataset, all four attributes are arranged along the x-axis. The y-axis represents a generic distance and it is “shared” by all these attributes on the x-axis. Hence, parallel charts work only when attributes share a common unit of numerical measure or when the attributes are normalized. This visualization is called a parallel axis because all four attributes are represented in four parallel axes parallel to the y-axis. In a parallel chart, a class label is used to color each data line so that one more dimension is introduced into the picture. By observing this parallel chart in Fig. 3.15, it can be noted that there is overlap between the three species on the sepal width attribute. So, sepal width cannot be the metric used to differentiate these three species. However, there is clear separation of species in petal length. No observation of I. setosa species has a petal length above 2.5 cm and there is little overlap between the I. virginica and I. versicolor species. Visually, just by knowing the petal length of an unlabeled observation, the species of Iris flower can be predicted. The relevance of this rule as a predictor will be discussed in the later chapter on Classification.

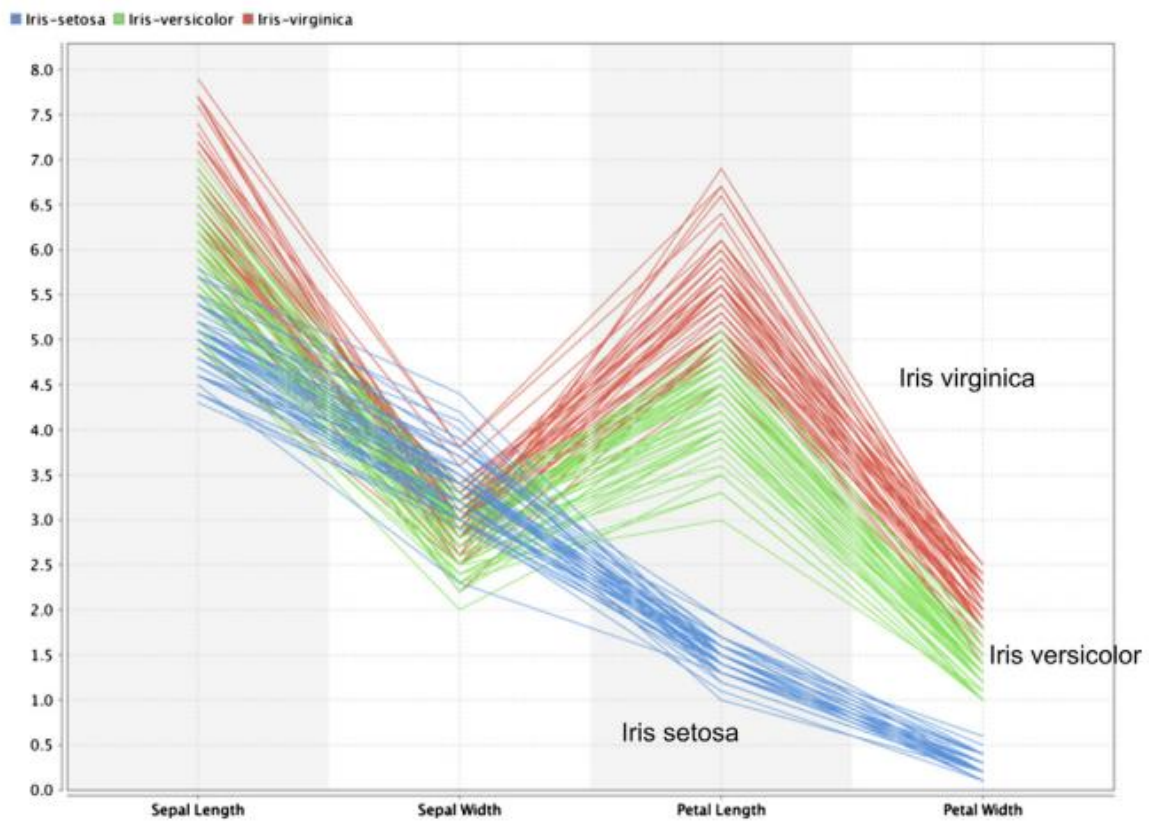


FIGURE 3.15
Parallel chart of Iris dataset.

Deviation Chart

A deviation chart is very similar to a parallel chart, as it has parallel axes for all the attributes on the x-axis. Data points are extended across the dimensions as lines and there is one common y-axis. Instead of plotting all data lines, deviation charts only show the mean and standard deviation statistics. For each class, deviation charts show the mean line connecting the mean of each attribute; the standard deviation is shown as the band above and below the mean line. The mean line does not have to correspond to a data point (line). With this method, information is elegantly displayed, and the essence of a parallel chart is maintained.

In Fig. 3.16, a deviation chart for the Iris dataset stratified by species is shown. It can be observed that the petal length is a good predictor to classify the species

because the mean line and the standard deviation bands for the species are well separated.

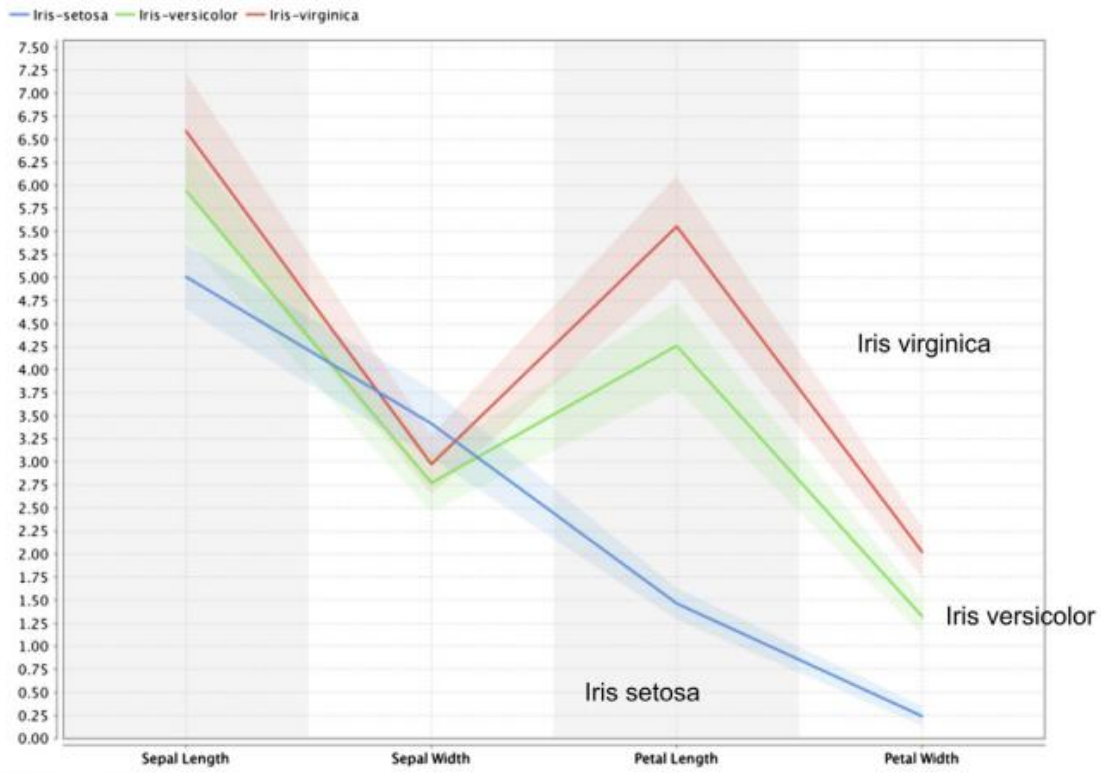


FIGURE 3.16
Deviation chart of Iris dataset.

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. 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 2 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 (Fig. 3.17). If two data points are similar, then the curves for the data points are closer to each other. If curves are far apart and belong to different classes, then this information can be used classify the data (Garcia-Osorio & Fyfe, 2005).

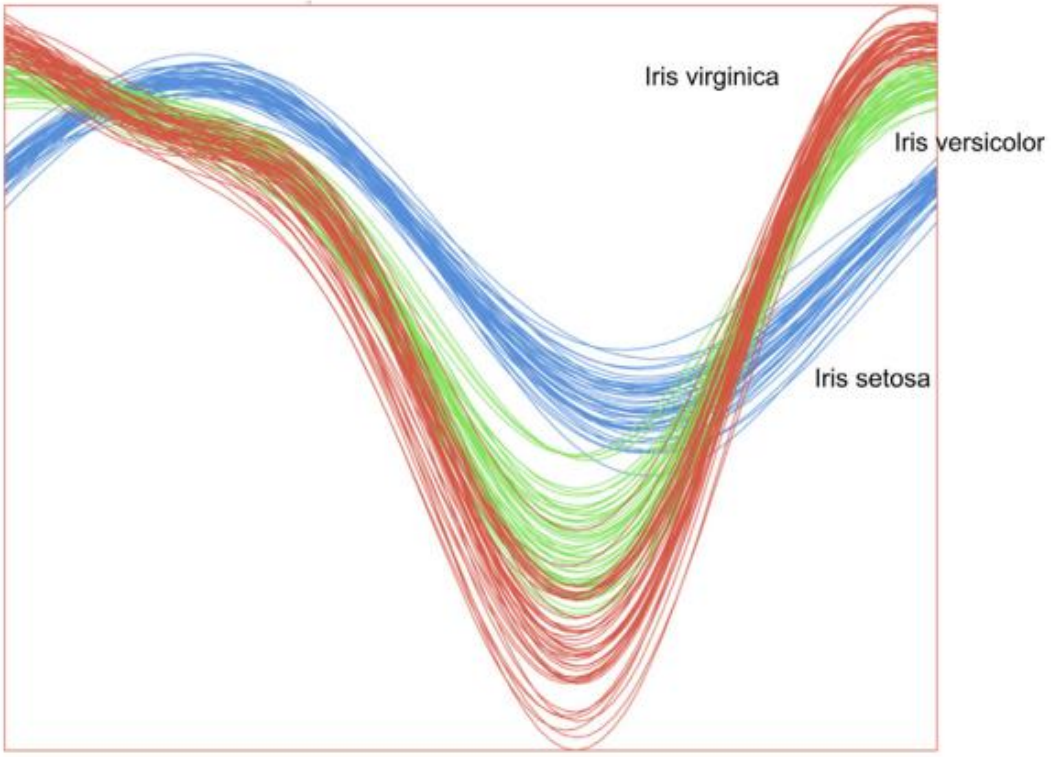


FIGURE 3.17
Andrews curves of Iris dataset.

3.5 ROADMAP FOR DATA EXPLORATION

If there is a new dataset that has not been investigated before, having a structured way to explore and analyze the data will be helpful. Here is a roadmap to inquire a new dataset. Not all steps may be relevant for every dataset and the

order may need to be adjusted for some sets, so this roadmap is intended as a guideline.

1. Organize the dataset: Structure the dataset with standard rows and columns. Organizing the dataset to have objects or instances in rows and dimensions or attributes in columns will be helpful for many data analysis tools. Identify the target or “class label” attribute, if applicable.

2. Find the central point for each attribute: Calculate mean, median, and mode for each attribute and the class label. If all three values are very different, it may indicate the presence of an outlier, or a multimodal or nonnormal distribution for an attribute.

3. Understand the spread of each attribute: Calculate the standard deviation and range for an attribute. Compare the standard deviation with the mean to understand the spread of the data, along with the max and min data points.

4. Visualize the distribution of each attribute: Develop the histogram and distribution plots for each attribute. Repeat the same for class-stratified histograms and distribution plots, where the plots are either repeated or color-coded for each class.

5. Pivot the data: Sometimes called dimensional slicing, a pivot is helpful to comprehend different values of the attributes. This technique can stratify by class and drill down to the details of any of the attributes. Microsoft Excel and Business Intelligence tools popularized this technique of data analysis for a wider audience.

6. Watch out for outliers: Use a scatterplot or quartiles to find outliers. The presence of outliers skews some measures like mean, variance, and range. Exclude outliers and rerun the analysis. Notice if the results change.

7. Understand the relationship between attributes: Measure the correlation between attributes and develop a correlation matrix. Notice what attributes are dependent on each other and investigate why they are dependent.

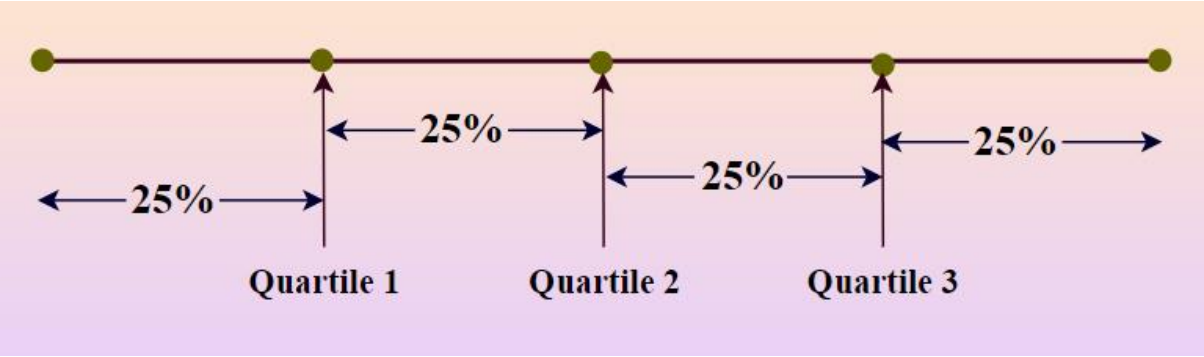
8. Visualize the relationship between attributes: Plot a quick scatter matrix to discover the relationship between multiple attributes at once. Zoom in on the attribute pairs with simple two-dimensional scatterplots stratified by class.

9. Visualize high-dimensional datasets: Create parallel charts and Andrews curves to observe the class differences exhibited by each attribute. Deviation charts provide a quick assessment of the spread of each class for each attribute.

appendix

The quartile measures the spread of values above and below the median by dividing the distribution into four groups.

They are grouped into four sections of 25% of the data, with the second and third groups representing the interquartile range.



Just like the median divides the data into half so that 50% of the measurement lies below the median and 50% lies above it, the quartile breaks down the data

into quarters so that 25% of the measurements are less than the lower quartile, 50% are less than the median, and 75% are less than the upper quartile.

There are three quartile values—a lower quartile, median, and upper quartile—which divide the data set into four ranges, each containing 25% of the data points:

- **First quartile**: The set of data points between the minimum value and the first quartile.
- **Second quartile**: The set of data points between the lower quartile and the median.
- **Third quartile**: The set of data between the median and the upper quartile.
- **Fourth quartile**: The set of data points between the upper quartile and the maximum value of the data set.

Calculating Quartiles Manually

Quartile manual calculation requires more effort as there are formulas involved. Using the same values as in the spreadsheet example:

- 59, 60, 65, 65, 68, 69, 70, 72, 75, 75, 76, 77, 81, 82, 84, 87, 90, 95, 98

Using the following formulas, you calculate each quartile:

- First Quartile (Q1) = (n + 1) x 1/4
- Second Quartile (Q2), or the median = (n + 1) x 2/4
- Third Quartile (Q3) = (n + 1) x 3/4

Where *n* is the number of integers in your dataset, and the result is the position of the number in the sequence dataset. So:

- First Quartile (Q1) = 20 x 1/4 = 5
- Second Quartile (Q2) = 20 x 2/4 = 10
- Third Quartile (Q3) = 20 x 3/4 = 15

Here, we have the Q1 (fifth) value of 68, the Q2 (tenth and the median) value of 75, and the Q3 (fifteenth) value of 84. The results differ slightly from the spreadsheet results because the spreadsheet calculates them differently. Your graph would then look like this: