Module 1
Introduction to Data Science

- Data science is a collection of techniques used to extract value from data
- It has become an essential tool for any organisation that collects, stores, and processes data as part of its operations
- Data science techniques find useful patterns, connections, and relationships within data
- Data science is also commonly referred to as knowledge discovery, machine learning, predictive analytics, and data mining
- However, each of these terms have a different meaning depending on the context

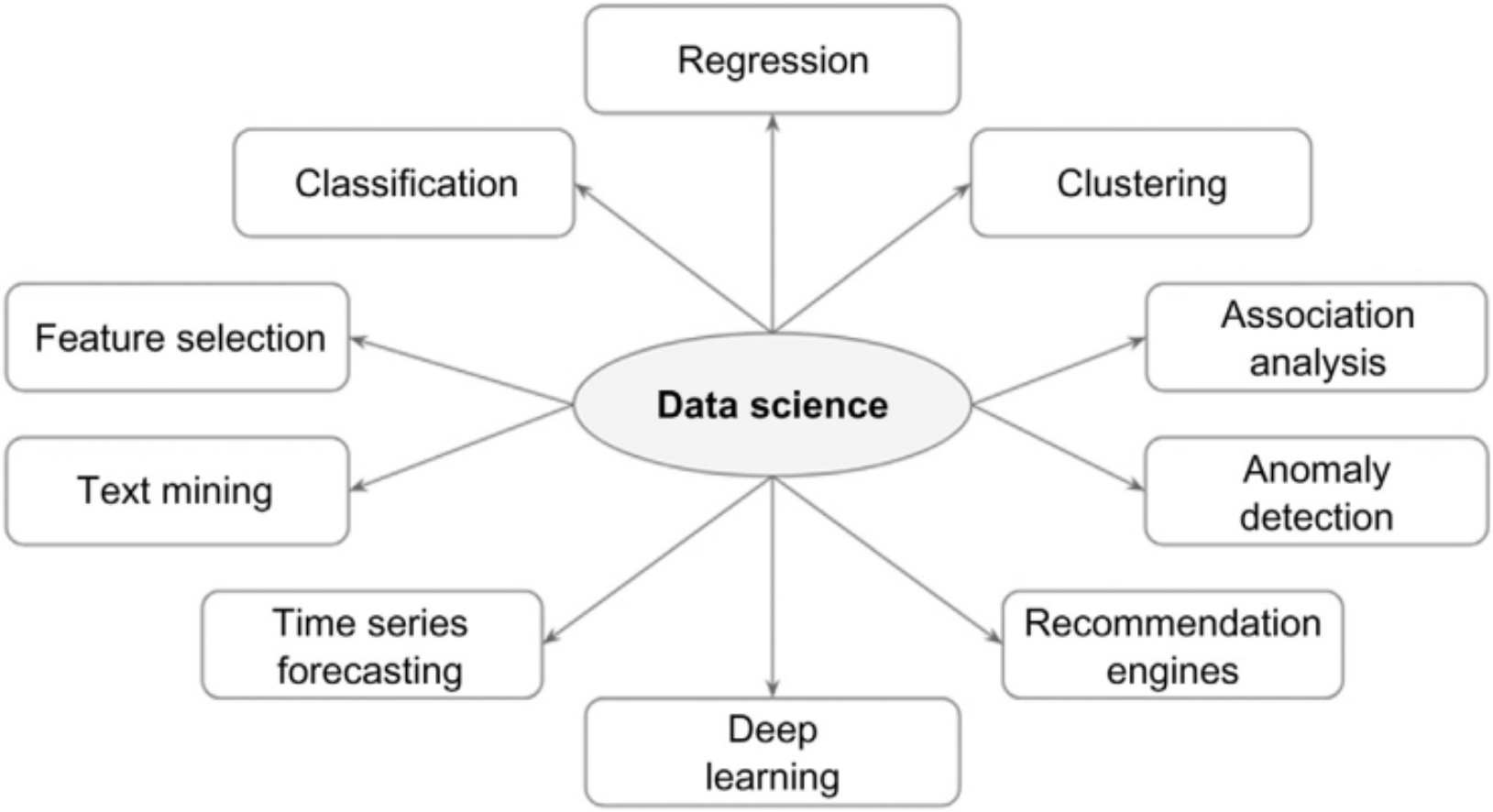
Data Science Classification

- Data Science problems are broadly classified into two types

1. Supervised Learning Model
   - In this model, training is given to the machine for solving problems
   - Simple Model
   - Highly Accurate
2. Unsupervised Learning Model
   - Here no training is given to the machine for solving problems
   - Complex Model
   - Less Accurate

Data Science Classification

- Data Science Tasks



Data Science Classification

- Classification and Regression techniques predict a target variable based on input variables
- The output variable which is predicted is called a target variable
- In classification, the target variable is a category such as 'yes', 'no', 'red', 'blue' etc.
- Example - Predicting whether monsoon will be normal this year
- In regression, the target variable is a numeric value
- Example - Predicting the age of a person
- Clustering is the process of identifying natural groupings within a data set
- Example - Grouping books in a library based on topics

Data Science Classification

- Association Analysis involves identifying associations or relationships within a data set
- Example - Finding which all items are bought together from a store
- Anomaly Detection is the process of identifying data points which are significantly different from other data points in a data set
- Example - Detecting fraudulent credit card transactions
- Recommendation Engines recommend items to users based on their preference / behaviour
- Example - Recommend items to users based on shopping behaviour
- Deep Learning is a machine learning model based on artificial neural networks, which are inspired by the functioning of human brain

Data Science Classification

- Time Series Forecasting involves predicting the future value of a variable based on its historical values
- Example - Predicting Temperature
- Text Mining is the process of retrieving information from text data such as documents, messages, email, web pages etc.
- Example - Web Search Engine
- Feature Selection is the process of selecting the most relevant features(attributes) to get the required output in the algorithm
- Example - Old Cars Data Set (Model, Year, Kms, Owner)
- Here we can discard 'Owner' details and select the other attributes for determining the cars to be crushed for spare parts

Data Science Process

- The method of discovering useful relationships and patterns in data is called the data science process
- Steps

1. Prior Knowledge
2. Data Preparation
3. Modelling
4. Application

Data Science Process

1. Prior Knowledge

- Here we define what problem is being solved
- We find out what data is needed for solving the problem
- Example - Consumer Loan Business
- Problem - If the interest rates and credit scores of past borrowers are known, can we predict the interest rate of a new borrower?
- Data - A sample data set of 10 data points with 3 attributes : borrower id, credit score, and interest rate

Data Science Process

1. Prior knowledge

Table 2.1 Dataset

Borrower ID	Credit Score	Interest Rate (%)
01	500	7.31
02	600	6.70
03	700	5.95
04	700	6.40
05	800	5.40
06	800	5.70
07	750	5.90
08	550	7.00
09	650	6.50
10	825	5.70

- Credit Score is a measure of the ability of a borrower to repay the loan
- Larger the credit score, greater the ability to repay the loan

Data Science Process

1. Prior knowledge

- A data set is a collection of data with a well defined structure
- Here, table is the data set
- A data point is a single instance of the data set
- Here, each record in the table is a data point
- An attribute is a single property of the data set
- Here, each column in the table is an attribute
- An identifier is a special attribute used for locating data points in a data set
- Here, Borrower Id is the identifier

Data Science Process

1. Prior knowledge

Table 2.2 New Data With Unknown Interest Rate

Borrower ID	Credit Score	Interest Rate
11	625	?

- A label is the special attribute to be predicted based on all the input attributes
- Here interest rate of the new borrower is to be predicted

Data Science Process

2. Data Preparation

- In this stage we prepare the whole data set needed for the data science task
- Steps

2.1 Data Exploration

- This involves in depth analysis of data to gain better understanding about it
- For this, statistical analysis and visualisation tools are used

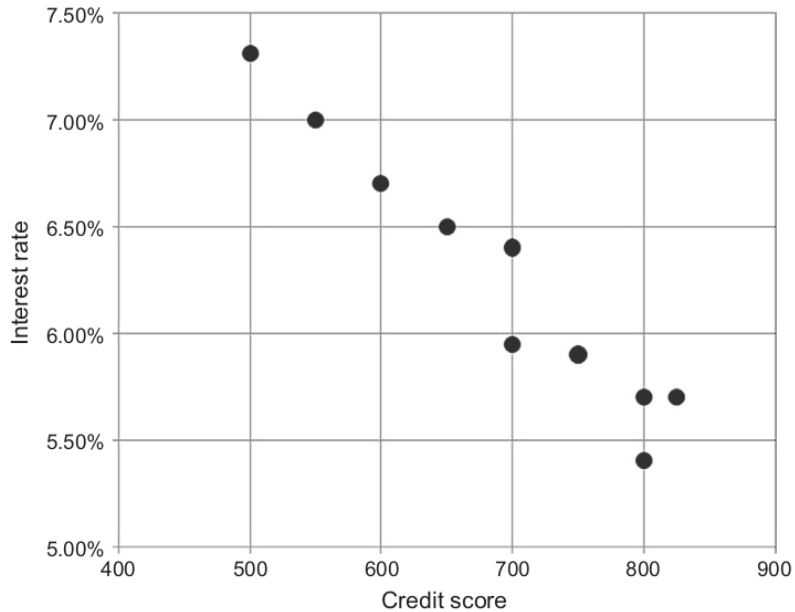


FIGURE 2.3
Scatterplot for interest rate dataset.

Data Science Process

2.2 Ensure Data Quality

- Data Cleansing techniques are used for ensuring data quality

1. Elimination of Duplicate Records
2. Dealing with Missing Values

- Example - Missing Credit Score
- It can be replaced with a credit score derived from the data set(mean)
- Alternatively, we can eliminate records with missing values

3. Data Type Conversion

- Depending on the requirement, we convert data from one type to another
- This depends on the data science algorithm we are using
- We can convert credit score to categorical values such as poor = 400, good = 600, excellent = 800

Data Science Process

2.2 Ensure Data Quality

- Data Cleansing techniques are used for ensuring data quality

4 Transformation of Attribute Ranges

- Different attributes have different ranges
- For example, range of income is larger compared to range of credit score
- For some data science algorithms, these ranges are normalised to a uniform scale from 0 to 1

5 Handling Outliers

- Outliers are anomalies in a data set
- Example - Human height as 1.73cm instead of 1.73m
- We need to correct these anomalies

Data Science Process

2.3 Feature Selection

- All the attributes in the data set may not be needed for solving the problem
- Reducing the number of attributes, without significant loss in the performance of the model, is called feature selection
- This leads to a simplified model

2.4 Data Sampling

- It involves selecting a subset of the original data set for analysis
- It reduces the amount of data to be processed
- It can speed up the process of analysis

Data Science Process

3 Modelling

- A model is the abstract representation of the data and the relationships in a given data set
- There are 2 kinds of data sets associated with a model
- The data set used to create the model is called a training data set
- The data set used to validate the model is called a test data set
- The entire data set is split into training data set and test data set
- A standard rule of thumb is two-thirds of the data are to be used as training and one-third as a test data set

Data Science Process

Table 2.3 Training Dataset

Borrower	Credit Score (*X*)	Interest Rate (*Y*) (%)
01	500	7.31
02	600	6.70
03	700	5.95
05	800	5.40
06	800	5.70
08	550	7.00
09	650	6.50

Table 2.4 Test Dataset

Borrower	Credit Score (*X*)	Interest Rate (*Y*)
04	700	6.40
07	750	5.90
10	825	5.70

Data Science Process

3 Modelling

- Now we will evaluate the model using the test data set
- We will be using simple linear regression technique for predicting interest rates of test data set

Table 2.5 Evaluation of Test Dataset

Borrower	Credit Score (*X*)	Interest Rate (*Y*) (%)	Model Predicted (*Y*) (%)	Model Error (%)
04	700	6.40	6.11	− 0.29
07	750	5.90	5.81	− 0.09
10	825	5.70	5.37	− 0.33

4 Application

- In this stage we present our findings to the world
- We can build an application that automatically updates reports, spreadsheets and presentation slides

Data Exploration

- Data exploration is the detailed analysis of data to gain better understanding of it
- It can be classified into two types

1. Descriptive Statistics

- Here we summarise the data using various statistical measures such as mean, median, mode etc.

2. Data Visualisation

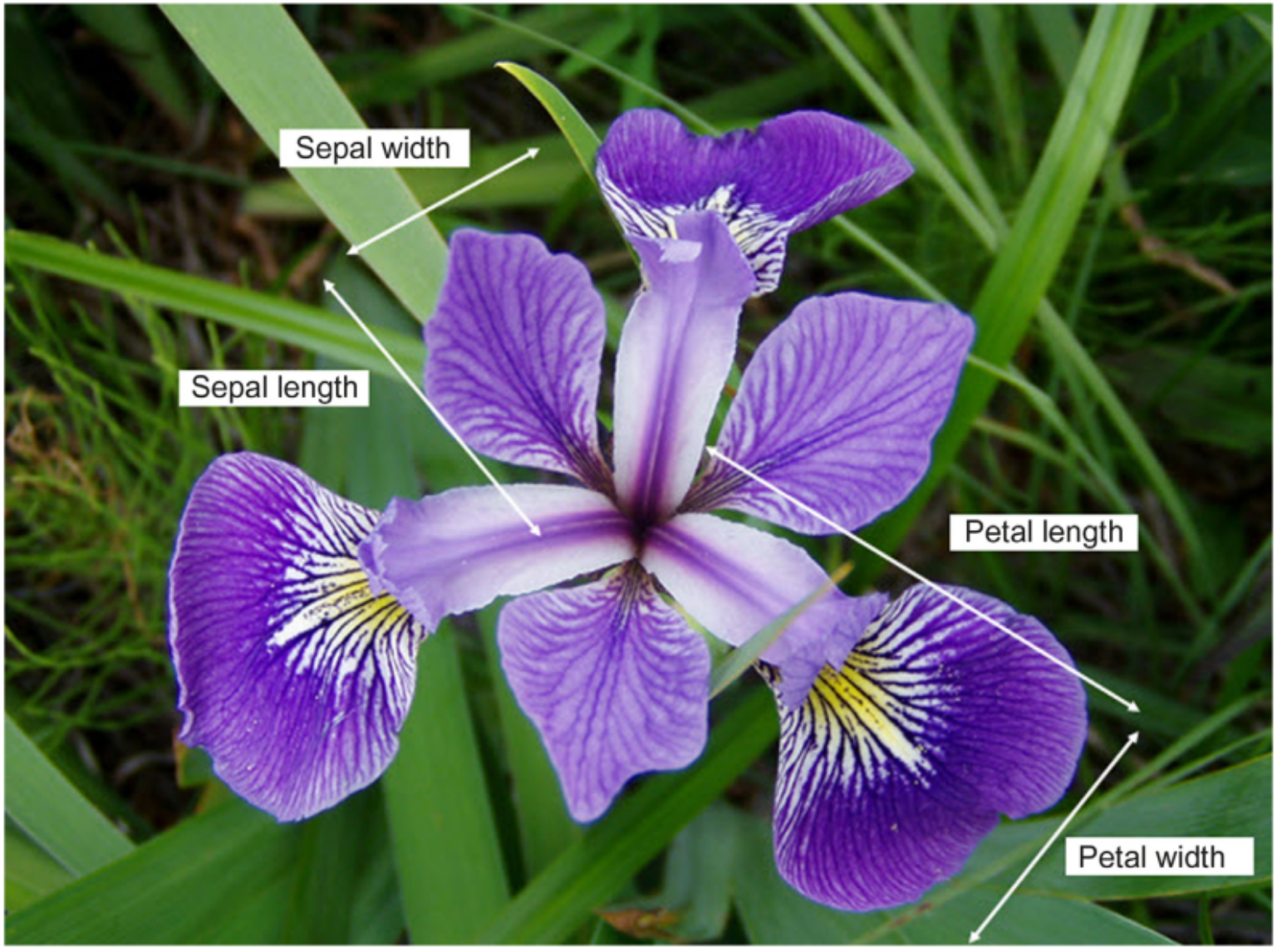
- Here we make visual representation of data using various charts like histogram, scatter plot, bubble chart etc.

Data Exploration

- Data Sets
- We will be considering a data set about Iris, a flowering plant
- Each observation has 4 attributes - sepal length, sepal width, petal length, petal width
- Sepal is the outer part of a flower
- Petal is the inner part of a flower
- Sepals support petals during their growth

Data Exploration

- Iris Flower



Data Exploration

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

Data Exploration

- Descriptive Statistics can be classified into two types

1 Descriptive Statistics for Univariate Data

- Here analysis of one attribute is done at a time

a Mean - It is the average of all observations in the data set

b Median - It is the value of the central point in the distribution

c Mode - It is the most frequently occurring observation

d Range - It is the difference between the maximum value and minimum value of the attribute

e Variance - It is a measure of how data points differ from the mean

f Standard Deviation - It is the square root of variance

Data Exploration

- Descriptive Statistics can be classified into two types
2. Descriptive Statistics for Multivariate Data
- Here analysis of multiple attributes is done at a time

a. Correlation

- It measures the statistical relationship between attributes
- It measures the dependence of one attribute on another
- Example - There is correlation between average temperature of a day and ice cream sales
- Correlation between two attributes is commonly measured by the Pearson correlation coefficient
- It ranges from -1 to 1, where negative values indicate negative correlation, positive values indicate positive correlation and 0 indicates no correlation
- -1 and 1 indicate perfect correlation

Data Visualisation

- For better understanding, visual representation of data is done using various charts
- Data Visualisation can also be classified into two types

1 Univariate Visualisation

- Here visualisation of one attribute is done at a time
- Charts - Histogram, Quartile Plot, Distribution Chart

2 Multivariate Visualisation

- Here visualisation of multiple attributes is done at a time
- Charts - Scatter Plot, Bubble Chart, Density Chart

Histogram

- It plots the frequency of occurrence of data within different ranges

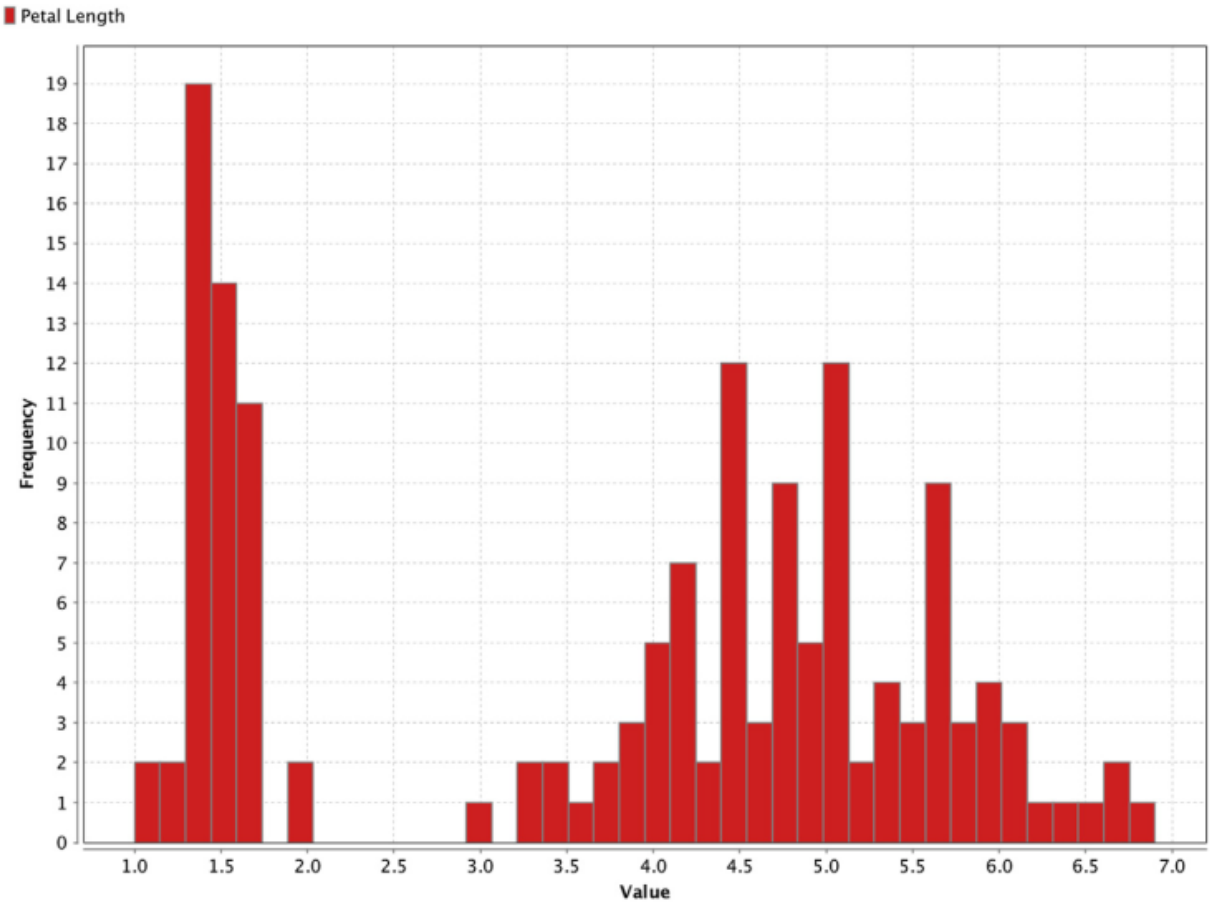


FIGURE 3.5

Histogram of petal length in Iris dataset.

Quartile Plot

- It plots the quartiles, outliers, mean and standard deviation
- The quartiles are denoted by Q1, Q2 and Q3
- 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 quartile 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
- The mean point is denoted by a solid dot overlay and standard deviation as a line overlay.

Quartile Plot

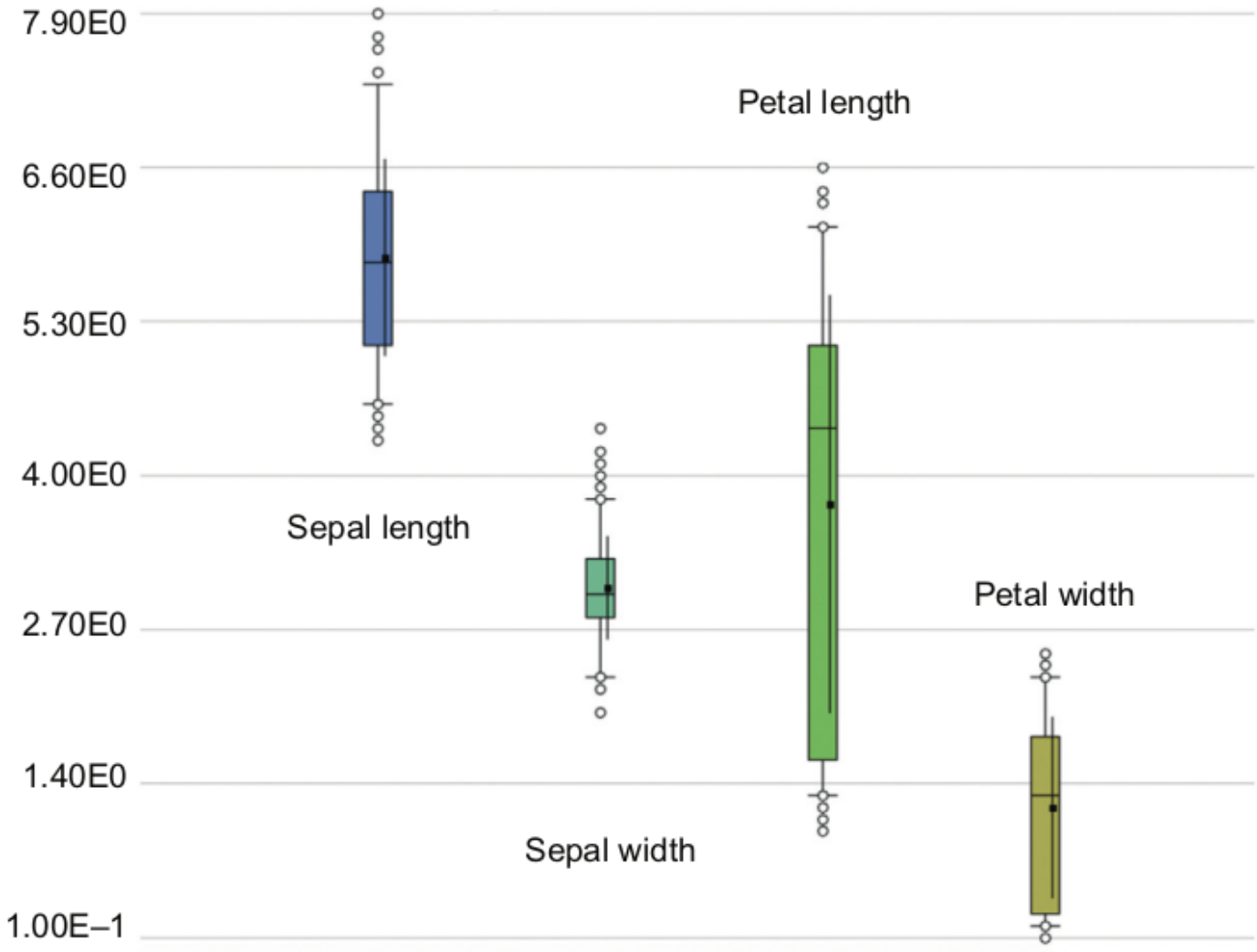


FIGURE 3.7
Quartile plot of Iris dataset.

Distribution Chart

- It shows the normal distribution function of the data
- It is also called the bell curve, due to its bell shape
- It shows the probability of occurrence of a data point within a range of values
- Example - Distribution charts of 3 different Iris species
- From the chart, we can predict that, an Iris flower of petal length 1.5 cm belongs to setosa species
- But we cannot predict the species of petal length 5 cm

Distribution Chart

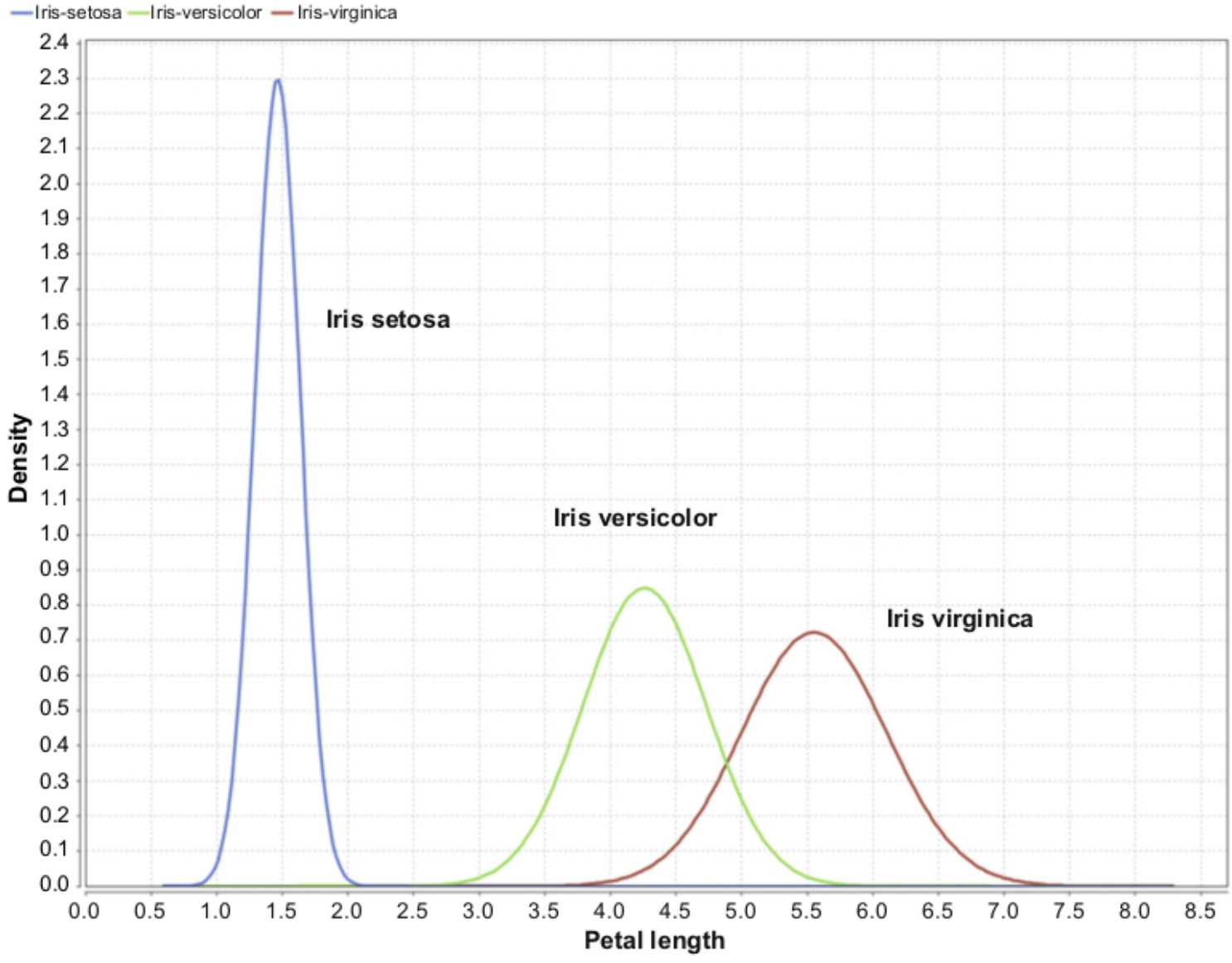


FIGURE 3.9

Distribution of petal length in Iris dataset.

Scatter Plot

- It shows the relationship between 2 attributes in the data set
- Each value in the scatter plot is represented using a dot
- Example - A scatter plot that shows the relationship between petal length and petal width of 3 different species of Iris data set

Scatter Plot

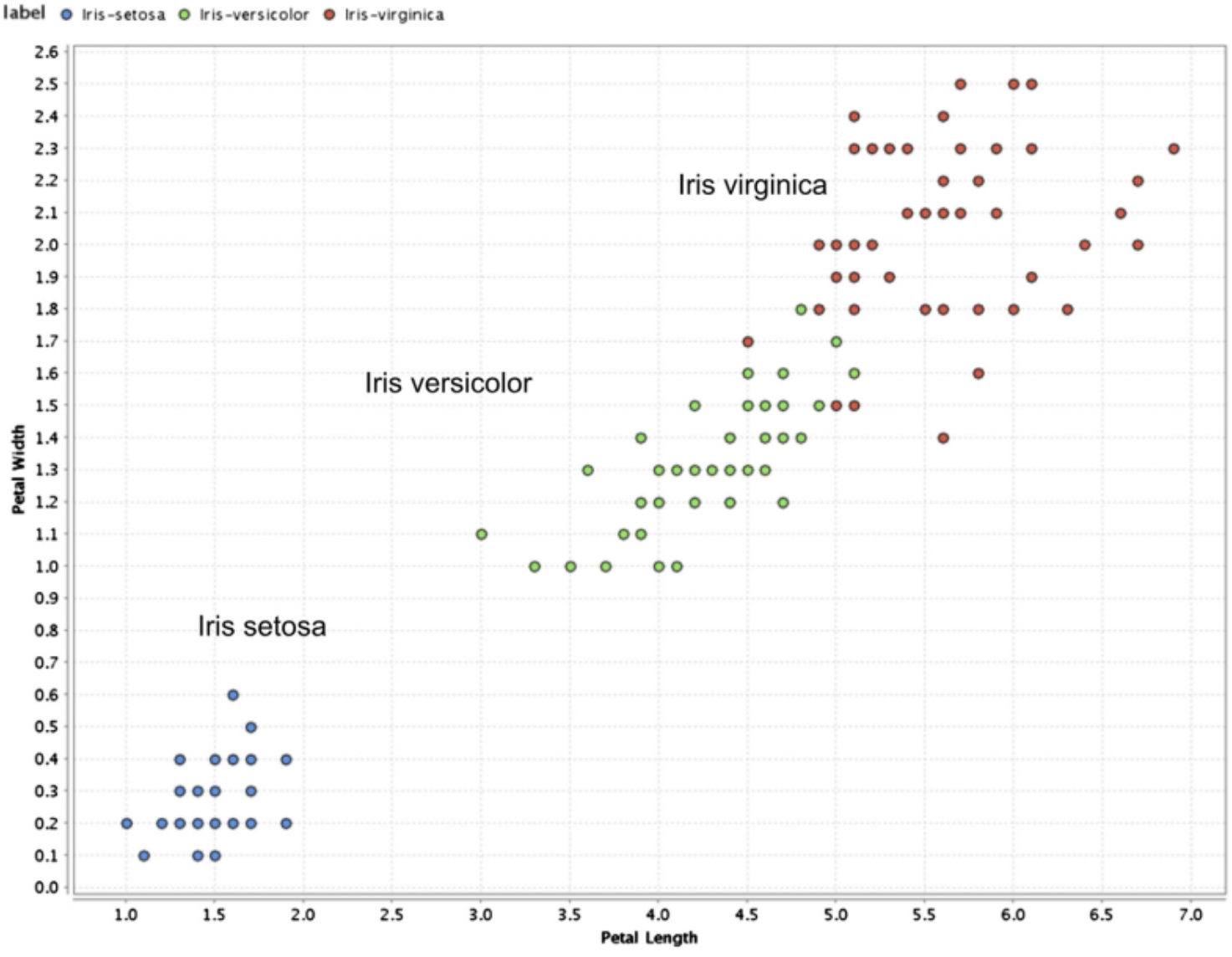


FIGURE 3.10
Scatterplot of Iris dataset.

Bubble Chart

- It also shows the relationship between 2 attributes in the data set
- Each value in the scatter plot is represented using a bubble
- Here, the size of the bubble is determined by a third attribute
- Example - A bubble chart that shows the relationship between petal length and petal width of 3 different species of Iris data set
- Here, the size of the bubble is determined by sepal width

Bubble Chart

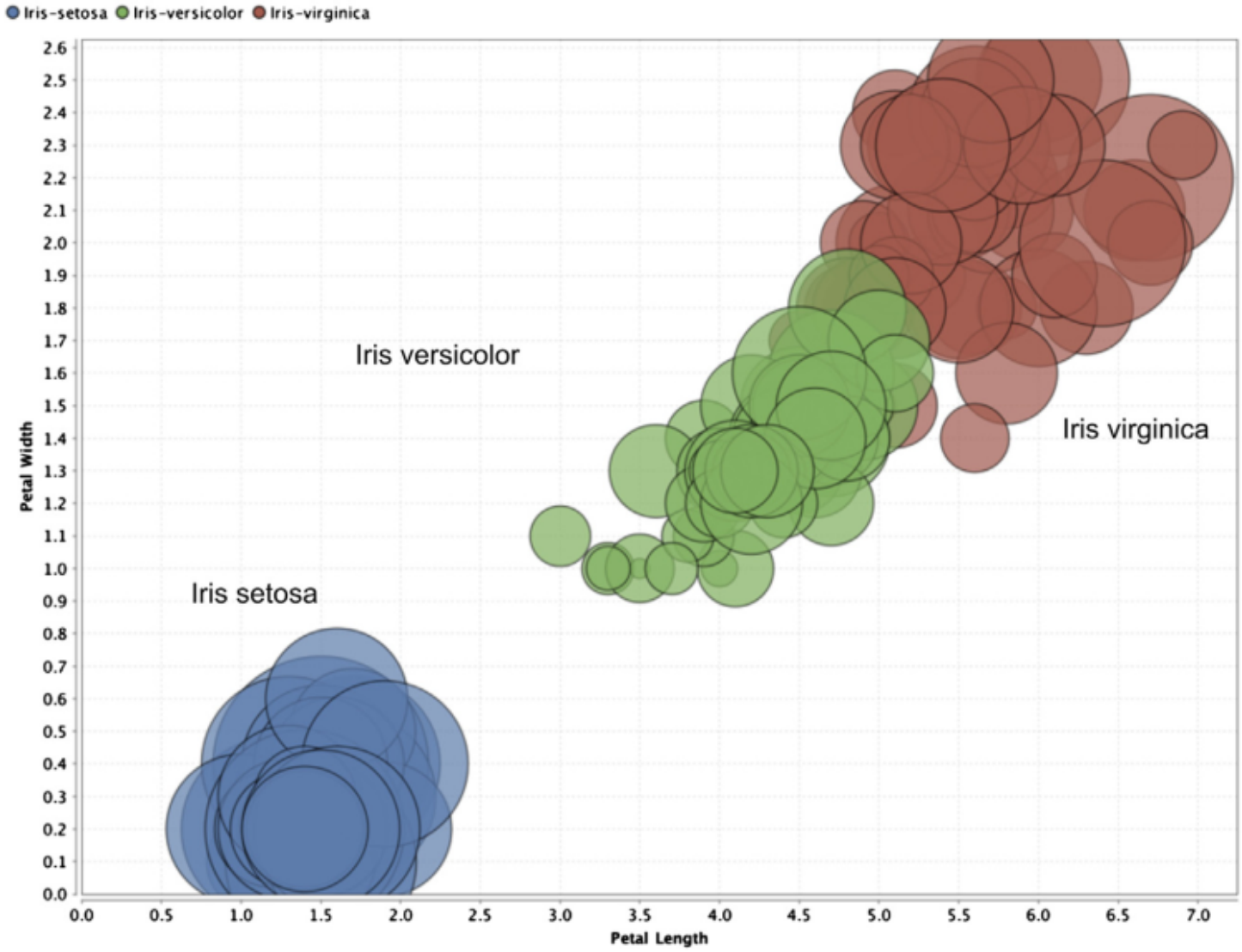


FIGURE 3.13

Bubble chart of Iris dataset.

Density Chart

- It also shows the relationship between 2 attributes in the data set
- Each value in the Density Chart is represented using a dot
- Here, the background colour is determined by a third attribute
- Example - A density chart that shows the relationship between petal length and petal width of 3 different species of Iris data set
- Here, the background colour is determined by sepal width

Density Chart

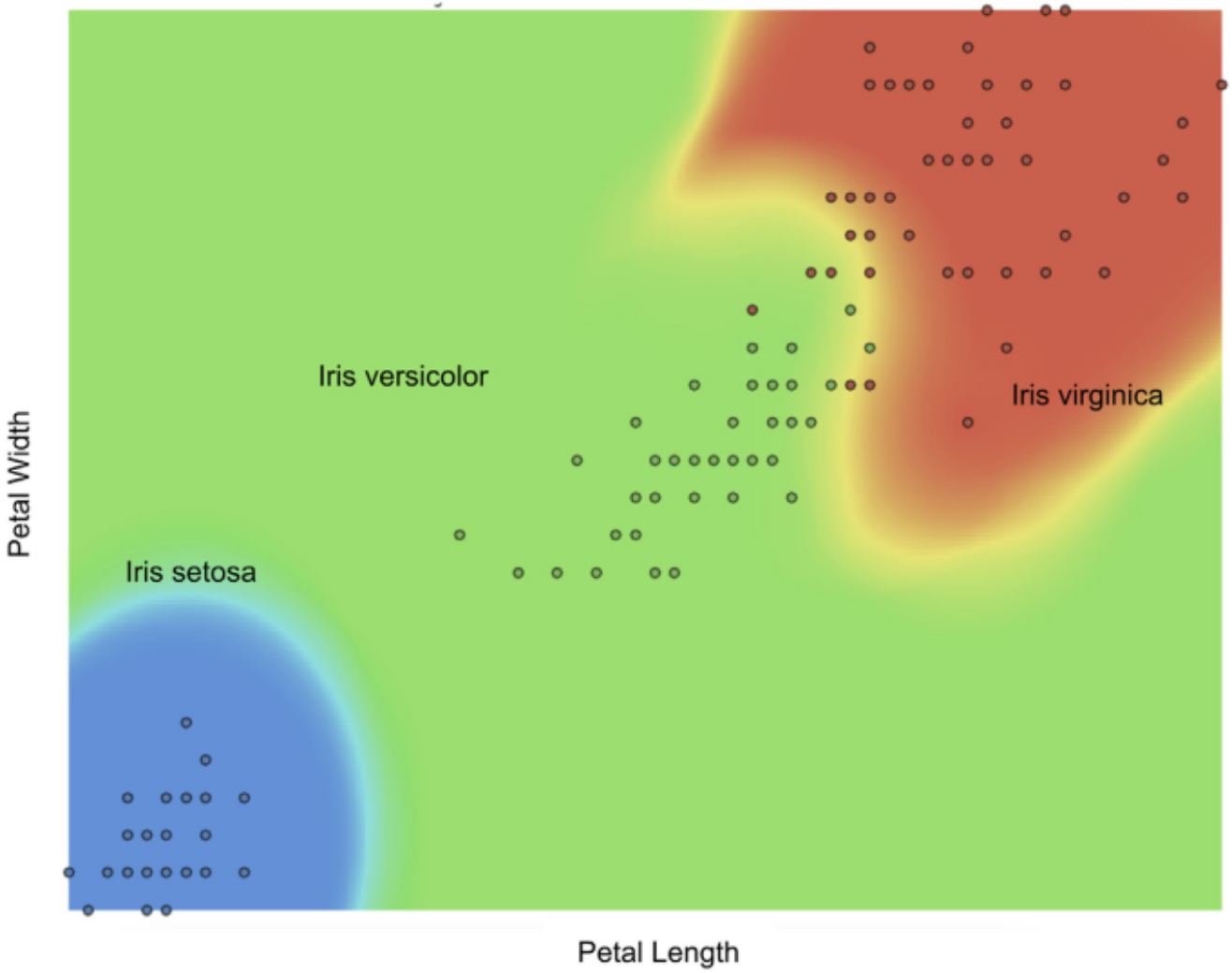


FIGURE 3.14
Density chart of a few attributes in the Iris dataset.

References

1. Vijay Kotu, Bala Deshpande, "Data Science Concepts and Practice", Morgan Kaufmann Publishers, 2018 (Module 1)
2. Brett Lantz, "Machine Learning with R", Second edition, PackT publishing, 2015 (Modules 2 to 5)