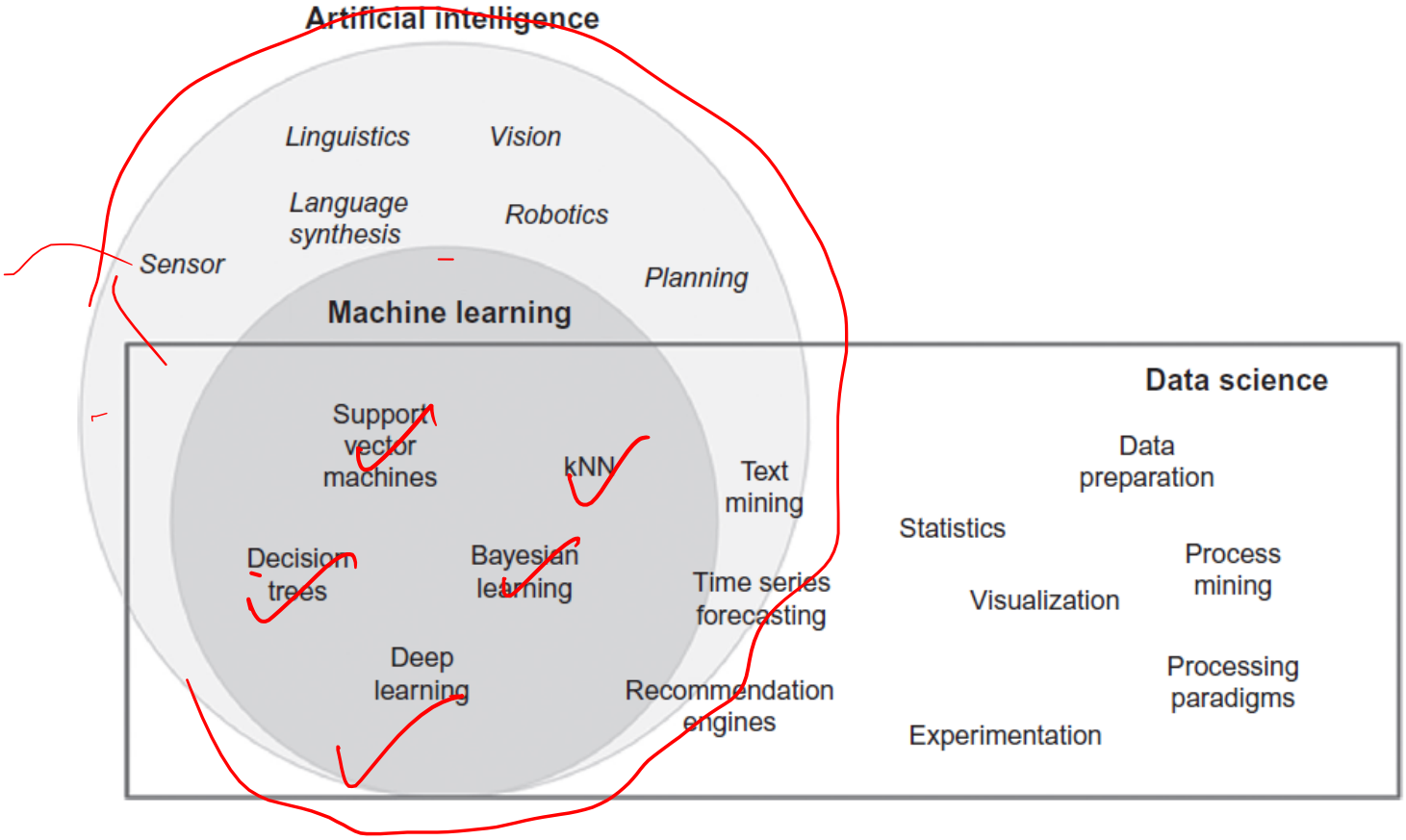
Data Science-Introduction

Dr. M. Brindha
Assistant Professor
Department of CSE
NIT, Trichy-15

Data Science

- Data science is a collection of techniques used to extract value from data.
- It has become an essential tool for any organization that collects, stores, and processes data as part of its operations.
- Data science techniques rely on finding useful patterns, connections, and relationships within data.
- Data science is also commonly referred to as knowledge discovery, machine learning, predictive analytics, and data mining.

AI, Machine Learning and Data Science



Artificial Intelligence (AI)

- Artificial intelligence is about giving machines the capability of mimicking human behaviour, particularly cognitive functions.
- Examples : facial recognition, automated driving, sorting mail based on postal code.
- Range of techniques that fall under artificial intelligence: linguistics, natural language processing, decision science, bias, vision, robotics, planning, etc.

Machine Learning

- Machine learning can either be considered a sub-field or one of the tools of artificial intelligence, is providing machines with the capability of learning from experience.
- Experience for machines comes in the form of data.
- Data that is used to teach machines is called training data.
- Machine learning turns the traditional programming model upside down.

Traditional program and machine learning

Data Science

- Data science is the business application of machine learning, artificial intelligence, and other quantitative fields like statistics, visualization, and mathematics.
- It is an interdisciplinary field that extracts value from data.
- Examples of data science user cases are: recommendation engines that can recommend movies for a particular user, a fraud alert model that detects fraudulent credit card transactions, find customers who will most likely churn next month, or predict revenue for the next quarter.

Data Science

- Data science starts with data, which can range from a simple array of a few numeric observations to a complex matrix of millions of observations with thousands of variables.
- Data science utilizes certain specialized computational methods in order to discover meaningful and useful structures within a dataset (Data Mining).
- The discipline of data science coexists and is closely associated with a number of related areas such as database systems, data engineering, visualization, data analysis, experimentation, and business intelligence (BI).

Data Science

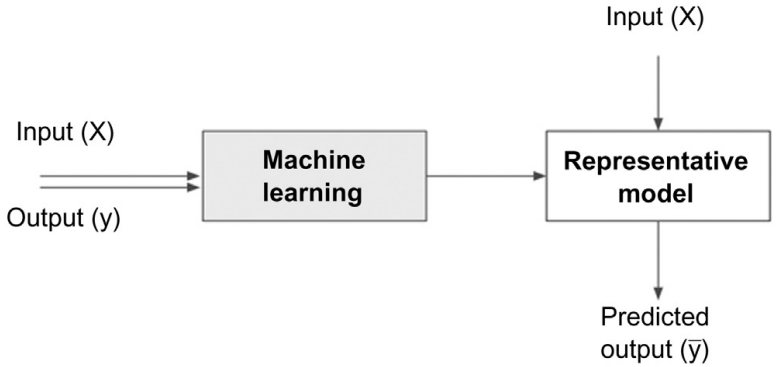
1. Extracting Useful Patterns

- Knowledge discovery in databases is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns or relationships within a dataset in order to make important decisions.
- Data science involves inference and iteration of many different hypotheses.
- One of the key aspects of data science is the process of generalization of patterns from a dataset.
- The generalization should be valid, not just for the dataset used to observe the pattern, but also for new unseen data.
- Data science is also a process with defined steps, each with a set of tasks. The term novel indicates that data science is usually involved in finding previously unknown patterns in data.
- The ultimate objective of data sscience is to find potentially useful conclusions that can be acted upon by the users of the analysis.

Data Science-

2. Building Representative Models

- In statistics, a model is the representation of a relationship between variables in a dataset.
- It describes how one or more variables in the data are related to other variables.
- Modeling is a process in which a representative abstraction is built from the observed dataset.
- For example, based on credit score, income level, and requested loan amount, a model can be developed to determine the interest rate of a loan.
- For this task, previously known observational data including credit score, income level, loan amount, and interest rate are needed.



- Once the representative model is created, it can be used to predict the value of the interest rate, based on all the input variables.

Data Science-

2. Building Representative Models

- Data science is the process of building a representative model that fits the observational data.
- This model serves two purposes: on the one hand, it predicts the output (interest rate) based on the new and unseen set of input variables (credit score, income level, and loan amount), and on the other hand, the model can be used to understand the relationship between the output variable and all the input variables.
- For example, does income level really matter in determining the interest rate of a loan?
- Does income level matter more than credit score?
- What happens when income levels double or if credit score drops by 10 points?
- A Model can be used for both predictive and explanatory applications.

Data Science- 3. Combination of Statistics, Machine Learning, and Computing

- In the pursuit of extracting useful and relevant information from large datasets, data science borrows computational techniques from the disciplines of statistics, machine learning, experimentation, and database theories.
- The algorithms used in data science originate from these disciplines but have since evolved to adopt more diverse techniques such as parallel computing, evolutionary computing, linguistics, and behavioral studies.
- One of the key ingredients of successful data science is substantial prior knowledge about the data and the business processes that generate the data, known as subject matter expertise.
- Like many quantitative frameworks, data science is an iterative process in which the practitioner gains more information about the patterns and relationships from data in each cycle.
- Data science also typically operates on large datasets that need to be stored, processed, and computed.
- This is where database techniques along with parallel and distributed computing techniques play an important role in data science.

Data Science- 4. Learning Algorithms

- Data science can also be defined as a process of discovering previously unknown patterns in data using automatic iterative methods.
- The application of sophisticated learning algorithms for extracting useful patterns from data differentiates data science from traditional data analysis techniques.
- Many of these algorithms were developed in the past few decades and are a part of machine learning and artificial intelligence.
- Some algorithms are based on the foundations of Bayesian probabilistic theories and regression analysis, originating from hundreds of years ago.
- These iterative algorithms automate the process of searching for an optimal solution for a given data problem.
- Based on the problem, data science is classified into tasks such as classification, association analysis, clustering, and regression.
- Each data science task uses specific learning algorithms like decision trees, neural networks, k-nearest neighbors (k-NN), and k-means clustering, among others.
- With increased research on data science, such algorithms are increasing, but a few classic algorithms remain foundational to many data science applications.

Data Science- 5. Associated Fields

- While data science covers a wide set of techniques, applications, and disciplines, there a few associated fields that data science heavily relies on.
- **Descriptive statistics:** Computing mean, standard deviation, correlation, and other descriptive statistics, quantify the aggregate structure of a dataset.
- **Exploratory visualization:** The process of expressing data in visual coordinates enables users to find patterns and relationships in the data and to comprehend large datasets.
- **Dimensional slicing:** Online analytical processing (OLAP) applications, which are prevalent in organizations, mainly provide information on the data through dimensional slicing, filtering, and pivoting.
- OLAP analysis is enabled by a unique database schema design where the data are organized as dimensions (e.g., products, regions, dates) and quantitative facts or measures (e.g., revenue, quantity).

Data Science- 5. Associated Fields

- Hypothesis testing: In confirmatory data analysis, experimental data are collected to evaluate whether a hypothesis has enough evidence to be supported or not.
- In general, data science is a process where many hypotheses are generated and tested based on observational data.
- Data engineering: Data engineering is the process of sourcing, organizing, assembling, storing, and distributing data for effective analysis and usage.
- Database engineering, distributed storage, and computing frameworks (e.g., Apache Hadoop, Spark, Kafka), parallel computing, extraction transformation and loading processing, and data warehousing constitute data engineering techniques.
- Business intelligence: Business intelligence helps organizations consume data effectively.
- It helps query the ad hoc data without the need to write the technical query command or use dashboards or visualizations to communicate the facts and trends.
- Business intelligence specializes in the secure delivery of information to right roles and the distribution of information at scale.

Case for Data Science

- In the past few decades, a massive accumulation of data has been seen with the advancement of information technology, connected networks, and the businesses it enables.
- This trend is also coupled with a steep decline in data storage and data processing costs.
- The applications built on these advancements like online businesses, social networking, and mobile technologies unleash a large amount of complex, heterogeneous data that are waiting to be analyzed.
- Traditional analysis techniques like dimensional slicing, hypothesis testing, and descriptive statistics can only go so far in information discovery.
- A paradigm is needed to manage the massive volume of data, explore the inter-relationships of thousands of variables, and deploy machine learning algorithms to deduce optimal insights from datasets.
- A set of frameworks, tools, and techniques are needed to intelligently assist humans to process all these data and extract valuable information.
- Data science is one such paradigm that can handle large volumes with multiple attributes and deploy complex algorithms to search for patterns from data.

Data Science Motivation- Volume

- The sheer volume of data captured by organizations is exponentially increasing.
- The rapid decline in storage costs and advancements in capturing every transaction and event, combined with the business need to extract as much leverage as possible using data, creates a strong motivation to store more data than ever.
- As data become more granular, the need to use large volume data to extract information increases.
- A rapid increase in the volume of data exposes the limitations of current analysis methodologies.

Data Science Motivation- Dimensions

- The three characteristics of the Big Data phenomenon are high volume, high velocity, and high variety.
- The variety of data relates to the multiple types of values (numerical, categorical), formats of data (audio files, video files), and the application of the data (location coordinates, graph data).
- Every single record or data point contains multiple attributes or variables to provide context for the record.
- For example, every user record of an ecommerce site can contain attributes such as products viewed, products purchased, user demographics, frequency of purchase, clickstream, etc.
- Determining the most effective offer for an ecommerce user can involve computing information across these attributes. Each attribute can be thought of as a dimension in the data space.
- The user record has multiple attributes and can be visualized in multidimensional space. The addition of each dimension increases the complexity of analysis techniques.

Data Science Motivation- Dimensions

- A simple linear regression model that has one input dimension is relatively easy to build compared to multiple linear regression models with multiple dimensions.
- As the dimensional space of data increase, a scalable framework that can work well with multiple data types and multiple attributes is needed.
- In the case of text mining, a document or article becomes a data point with each unique word as a dimension.
- Text mining yields a dataset where the number of attributes can range from a few hundred to hundreds of thousands of attributes.

Data Science Motivation- Complex Questions

- As more complex data are available for analysis, the complexity of information that needs to get extracted from data is increasing as well.
- If the natural clusters in a dataset, with hundreds of dimensions, need to be found, then traditional analysis like hypothesis testing techniques cannot be used in a scalable fashion.
- The machine-learning algorithms need to be leveraged in order to automate searching in the vast search space.
- Traditional statistical analysis approaches the data analysis problem by assuming a stochastic model, in order to predict a response variable based on a set of input variables.
- A linear regression is a classic example of this technique where the parameters of the model are estimated from the data.
- These hypothesis-driven techniques were highly successful in modeling simple relationships between response and input variables.

Data Science Motivation- Complex Questions

- Machine learning approaches the problem of modeling by trying to find an algorithmic model that can better predict the output from input variables.
- The algorithms are usually recursive and, in each cycle, estimate the output and "learn" from the predictive errors of the previous steps.
- This route of modeling greatly assists in exploratory analysis since the approach here is not validating a hypothesis but generating a multitude of hypotheses for a given problem.
- In the context of the data problems faced today, both techniques need to be deployed.

Data Science Classification

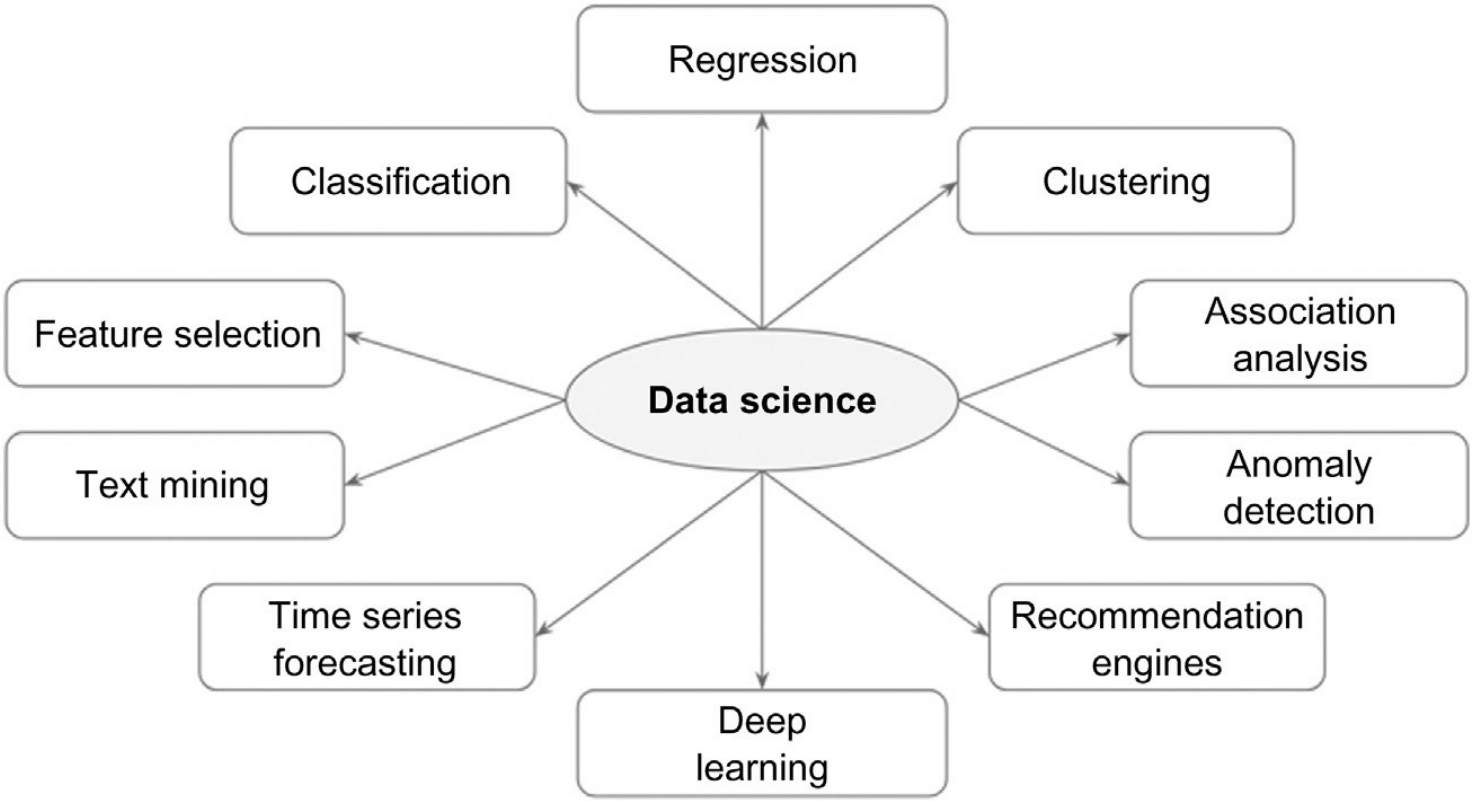
- Data science problems can be broadly categorized into supervised or unsupervised learning models.
- Supervised or directed data science tries to infer a function or relationship based on labeled training data and uses this function to map new unlabeled data.
- Supervised techniques predict the value of the output variables based on a set of input variables.
- To do this, a model is developed from a training dataset where the values of input and output are previously known.
- The model generalizes the relationship between the input and output variables and uses it to predict for a dataset where only input variables are known.
- The output variable that is being predicted is also called a class label or target variable.
- Supervised data science needs a sufficient number of labeled records to learn the model from the data.

Data Science Classification

- Unsupervised or undirected data science uncovers hidden patterns in unlabeled data.
- In unsupervised data science, there are no output variables to predict.
- The objective of this class of data science techniques, is to find patterns in data based on the relationship between data points themselves.
- An application can employ both supervised and unsupervised learners.

Data Science Classification

- Data science problems can also be classified into tasks such as: classification, regression, association analysis, clustering, anomaly detection, recommendation engines, feature selection, time series forecasting, deep learning, and text mining.



Data Science Classification

- Classification and regression techniques predict a target variable based on input variables. The prediction is based on a generalized model built from a previously known dataset.
- In regression tasks, the output variable is numeric (e.g., the mortgage interest rate on a loan).
- Classification tasks predict output variables, which are categorical or polynomial (e.g., the yes or no decision to approve a loan).
- Deep learning is a more sophisticated artificial neural network that is increasingly used for classification and regression problems.
- Clustering is the process of identifying the natural groupings in a dataset. For example, clustering is helpful in finding natural clusters in customer datasets, which can be used for market segmentation.
- Since this is unsupervised data science, it is up to the end user to investigate why these clusters are formed in the data and generalize the uniqueness of each cluster.

Data Science Classification

- In retail analytics, it is common to identify pairs of items that are purchased together, so that specific items can be bundled or placed next to each other.
- This task is called market basket analysis or association analysis, which is commonly used in cross selling.
- Recommendation engines are the systems that recommend items to the users based on individual user preference.
- Anomaly or outlier detection identifies the data points that are significantly different from other data points in a dataset.
- Credit card transaction fraud detection is one of the most prolific applications of anomaly detection.

Data Science Classification

- **Time series forecasting** is the process of predicting the future value of a variable (e.g., temperature) based on past historical values that may exhibit a trend and seasonality.
- **Text mining** is a data science application where the input data is text, which can be in the form of documents, messages, emails, or web pages.
- To aid the data science on text data, the text files are first converted into document vectors where each unique word is an attribute.
- Once the text file is converted to document vectors, standard data science tasks such as classification, clustering, etc., can be applied.
- **Feature selection** is a process in which attributes in a dataset are reduced to a few attributes that really matter.

Data Science Classification

- A complete data science application can contain elements of both supervised and unsupervised techniques.
- Unsupervised techniques provide an increased understanding of the dataset and hence, are sometimes called descriptive data science.
- As an example of how both unsupervised and supervised data science can be combined in an application, consider the following scenario.
- In marketing analytics, clustering can be used to find the natural clusters in customer records.
- Each customer is assigned a cluster label at the end of the clustering process.
- A labeled customer dataset can now be used to develop a model that assigns a cluster label for any new customer record with a supervised classification technique.

Data Science Algorithms

- In data science, Algorithm is the blueprint for how a particular data problem is solved.
- Many of the learning algorithms are recursive, where a set of steps are repeated many times until a limiting condition is met.
- Some algorithms also contain a random variable as an input and are aptly called randomized algorithms.
- A classification task can be solved using many different learning algorithms such as decision trees, artificial neural networks, k-NN, and even some regression algorithms.
- The choice of which algorithm to use depends on the type of dataset, objective, structure of the data, presence of outliers, available computational power, number of records, number of attributes, and so on.
- Data science tools or statistical programming tools -R, RapidMiner, Python, SAS Enterprise Miner

Data Science Algorithms

Table 1.1 Data Science Tasks and Examples

Tasks	Description	Algorithms	Examples
Classification	Predict if a data point belongs to one of the predefined classes. The prediction will be based on learning from a known dataset	Decision trees, neural networks, Bayesian models, induction rules, *k*-nearest neighbors	Assigning voters into known buckets by political parties, e.g., soccer moms Bucketing new customers into one of the known customer groups
Regression	Predict the numeric target label of a data point. The prediction will be based on learning from a known dataset	Linear regression, logistic regression	Predicting the unemployment rate for the next year Estimating insurance premium
Anomaly detection	Predict if a data point is an outlier compared to other data points in the dataset	Distance-based, density-based, LOF	Detecting fraudulent credit card transactions and network intrusion
Time series forecasting	Predict the value of the target variable for a future timeframe based on historical values	Exponential smoothing, ARIMA, regression	Sales forecasting, production forecasting, virtually any growth phenomenon that needs to be extrapolated
Clustering	Identify natural clusters within the dataset based on inherit properties within the dataset	*k*-Means, density-based clustering (e.g., DBSCAN)	Finding customer segments in a company based on transaction, web, and customer call data
Association analysis	Identify relationships within an item set based on transaction data	FP-growth algorithm, a priori algorithm	Finding cross-selling opportunities for a retailer based on transaction purchase history
Recommendation engines	Predict the preference of an item for a user	Collaborative filtering, content-based filtering, hybrid recommenders	Finding the top recommended movies for a user

LOF, *local outlier factor;* ARIMA, *autoregressive integrated moving average;* DBSCAN, *density-based spatial clustering of applications with noise;* FP, *frequent pattern.*

Data Science Process

- The methodical discovery of useful relationships and patterns in data is enabled by a set of iterative activities collectively known as the data science process.

The standard data science process involves

(1) understanding the problem,

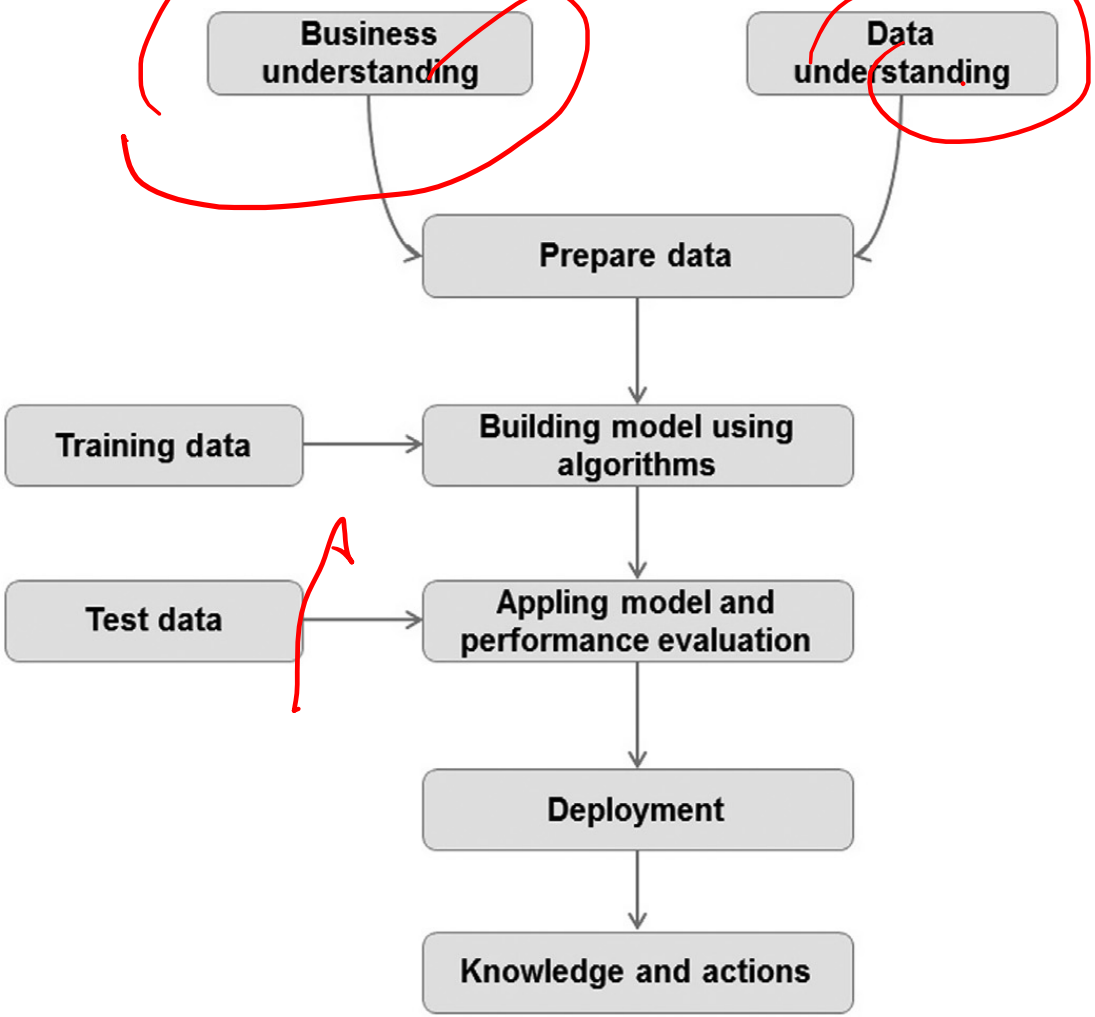
(2) preparing the data samples,

(3) developing the model,

(4) Applying the model on a dataset to see how the model may work in the real world,

(5) deploying and maintaining the models.

Data Science Process



1. Prior Knowledge
2. Preparation
3. Modeling
4. Application
5. Knowledge

Data Science Process-Prior Knowledge

- Prior knowledge refers to information that is already known about a subject.
- The data science problem doesn't emerge in isolation; it always develops on top of existing subject matter and contextual information that is already known.
- The prior knowledge step in the data science process helps to define what problem is being solved, how it fits in the business context, and what data is needed in order to solve the problem.

Data Science Process-Prior Knowledge

Objective

- The data science process starts with a need for analysis, a question, or a business objective.
- Without a well-defined statement of the problem, it is impossible to come up with the right dataset and pick the right data science algorithm.
- As an iterative process, it is common to go back to previous data science process steps, revise the assumptions, approach, and tactics.
- The data science process is going to be explained using a hypothetical use case.

For example, take the consumer loan business

- The business objective of this hypothetical case is: If the interest rate of past borrowers with a range of credit scores is known, can the interest rate for a new borrower be predicted?

Data Science Process-Prior Knowledge

Subject Matter

- The process of data science uncovers hidden patterns in the dataset by exposing relationships between attributes.
- But the problem is that it uncovers a lot of patterns.
- The false or spurious signals are the major problem in the data science process.
- It is up to the practitioner to sift through the exposed patterns and accept the ones that are valid and relevant to the answer of the objective question.
- Hence, it is essential to know the subject matter, the context, and the business process generating the data.

Data Science Process-Prior Knowledge

Subject Matter

- If the objective is to predict the lending interest rate, then it is important to know how the lending business works, why the prediction matters, what happens after the rate is predicted, what data points can be collected from borrowers, what data points cannot be collected because of the external regulations and the internal policies, what other external factors can affect the interest rate, how to verify the validity of the outcome, and so forth.
- Understanding current models and business practices lays the foundation and establishes known knowledge.
- Analysis and mining the data provides the new knowledge that can be built on top of the existing knowledge.

Data Science Process-Prior Knowledge

Data

- Similar to the prior knowledge in the subject area, prior knowledge in the data can also be gathered.
- Understanding how the data is collected, stored, transformed, reported, and used is essential to the data science process.
- This part of the step surveys all the data available to answer the business question and narrows down the new data that need to be sourced.
- There are quite a range of factors to consider: quality of the data, quantity of data, availability of data, gaps in data, does lack of data compel the practitioner to change the business question, etc.
- The objective of this step is to come up with a dataset to answer the business question through the data science process.
- It is critical to recognize that an inferred model is only as good as the data used to create it.

Data Science Process-Prior Knowledge

Data

- For the lending example, a sample dataset of ten data points with three attributes has been put together: identifier, credit score, and interest rate.

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

Table 2.2 New Data With Unknown Interest Rate

Borrower ID	Credit Score	Interest Rate
11	625	?

Data Science Process-Prior Knowledge

Data

- A **dataset (example set)** is a collection of data with a defined structure. Sometimes referred to as a "data frame". E.g. Table
- A **data point (record, object or example)** is a single instance in the dataset(row). Each instance contains the same structure as the dataset.
- An **attribute** (feature, input, dimension, variable, or predictor) is a single property of the dataset (column). Attributes can be numeric (credit score and the interest rate), categorical, date-time, text, or Boolean data types.
- A **label** (class label, output, prediction, target, or response) is the special attribute to be predicted based on all the input attributes. E.g. the interest rate is the output variable.
- **Identifiers** are special attributes that are used for locating or providing context to individual records. E.g. the attribute ID is the identifier.

Data Science Process-Prior Knowledge

Causation Versus Correlation

- Suppose the business question is inverted: Based on the data, can the credit score of the borrower be predicted based on interest rate? The answer is yes—but it does not make business sense.
- From the existing domain expertise, it is known that credit score influences the loan interest rate.
- Predicting credit score based on interest rate inverses the direction of the causal relationship.
- This question also exposes one of the key aspects of model building.
- The correlation between the input and output attributes doesn't guarantee causation.
- Hence, it is important to frame the data science question correctly using the existing domain and data knowledge.
- In this data science example, the interest rate of the new borrower with an unknown interest rate will be predicted based on the pattern learned from known data.

Data Science Process-Data Preparation

- Preparing the dataset to suit a data science task is the most time-consuming part of the process.
- It is extremely rare that datasets are available in the form required by the data science algorithms.
- Most of the data science algorithms would require data to be structured in a tabular format with records in the rows and attributes in the columns.
- If the data is in any other format, the data would need to be transformed by applying pivot, type conversion, join, or transpose functions, etc., to condition the data into the required structure.

Data Science Process-Data Preparation

Data Exploration

- Data exploration, also known as exploratory data analysis, provides a set of simple tools to achieve basic understanding of the data.
- Data exploration approaches involve computing descriptive statistics and visualization of data.
- They can expose the structure of the data, the distribution of the values, the presence of extreme values, and highlight the inter-relationships within the dataset.
- Descriptive statistics like mean, median, mode, standard deviation, and range for each attribute provide an easily readable summary of the key characteristics of the distribution of data.
- On the other hand, a visual plot of data points provides an instant grasp of all the data points condensed into one chart.

Data Science Process-Data Preparation

Data Exploration

- The scatterplot of credit score vs. loan interest rate and it can be observed that as credit score increases, interest rate decreases.

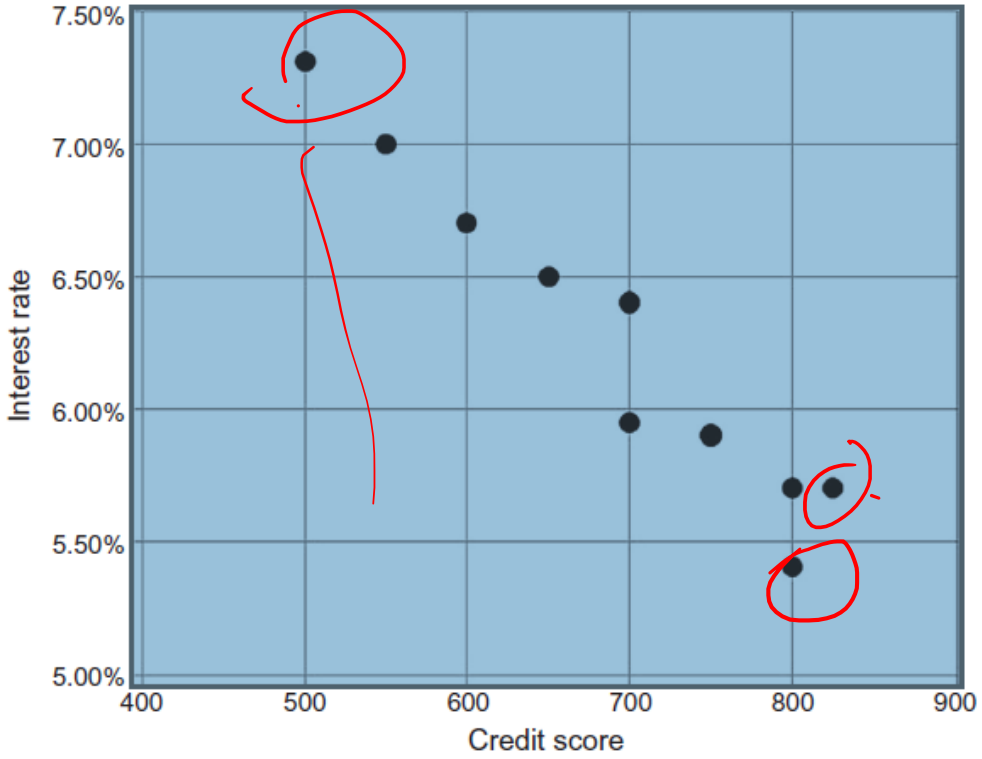


FIGURE 2.3
Scatterplot for interest rate dataset.

Data Science Process-Data Preparation

Data Quality

- Data quality is an ongoing concern wherever data is collected, processed, and stored.
- In the interest rate dataset, how does one know if the credit score and interest rate data are accurate?
- What if a credit score has a recorded value of 900 (beyond the theoretical limit) or if there was a data entry error? Errors in data will impact the representativeness of the model.
- Organizations use data alerts, cleansing, and transformation techniques to improve and manage the quality of the data and store them in companywide repositories called **data warehouses.**
- The data cleansing practices include elimination of duplicate records, quarantining outlier records that exceed the bounds, standardization of attribute values, substitution of missing values, etc.

Data Science Process-Data Preparation

Missing Values

- Missing attribute values. For example, a credit score may be missing in one of the records. Mitigation methods are there.
- Managing missing values- find the reason behind why the values are missing.
- Tracking the data lineage (provenance) of the data source can lead to the identification of systemic issues during data capture or errors in data transformation.
- Knowing the source of a missing – find mitigation methodology to use.
- The missing value can be substituted with a range of artificial data.
- Missing credit score values can be replaced with a credit score derived from the dataset (mean, minimum, or maximum value, depending on the characteristics of the attribute).
- This method is useful if the missing values occur randomly and the frequency of occurrence is quite rare.
- Alternatively, to build the representative model, all the data records with missing values or records with poor data quality can be ignored. This method reduces the size of the dataset.

Data Science Process-Data Preparation

Missing Values

- Missing attribute values. For example, a credit score may be missing in one of the records. Mitigation methods are there.
- Managing missing values- find the reason behind why the values are missing. Knowing the source of a missing – find mitigation methodology to use.
- Tracking the of the data source - identification of systemic issues during data capture or errors in data transformation.
- The missing value can be substituted with a range of artificial data.
- Missing credit score values can be replaced with a credit score derived from the dataset (mean, minimum, or maximum value, depending on the characteristics of the attribute). Useful if the missing values occur randomly and the frequency of occurrence is quite rare.
- Alternatively, to build the representative model, all the data records with missing values or records with poor data quality can be ignored. This method reduces the size of the dataset.
- k-nearest neighbor (k-NN) algorithm for classification-robust with missing values. Neural network models for classification- do not perform well with missing attributes, the data preparation step is essential for developing neural network models.

Data Science Process-Data Preparation

Data Types and Conversion

- The attributes in a dataset can be of different types, such as continuous numeric (interest rate), integer numeric (credit score), or categorical.
- For example, the credit score can be expressed as categorical values (poor, good, excellent) or numeric score.
- Linear regression models-The input attributes have to be numeric.
- If the available data are categorical, they must be converted to continuous numeric attribute.
- A specific numeric score can be encoded for each category value, such as poor-400, good-600, excellent-700, etc.
- Similarly, numeric values can be converted to categorical data types by a technique called binning, where a range of values are specified for each category, for example, a score between 400 and 500 can be encoded as "low" and so on.

Data Science Process-Data Preparation

Transformation

- Data science algorithms like k-NN, the input attributes are expected to be numeric and normalized, because the algorithm compares the values of different attributes and calculates distance between the data points.
- Normalization prevents one attribute dominating the distance results because of large values.
- For example, consider income (expressed in USD, in thousands) and credit score (in hundreds).
- The distance calculation will always be dominated by slight variations in income.
- One solution is to convert the range of income and credit score to a more uniform scale from 0 to 1 by normalization.
- This way, a consistent comparison can be made between the two different attributes with different units.

Data Science Process-Data Preparation

Outliers

- Outliers are anomalies in a given dataset.
- Outliers may occur because of correct data capture (few people with income in tens of millions) or erroneous data capture (human height as 1.73 cm instead of 1.73 m).
- Regardless, the presence of outliers needs to be understood and will require special treatments.
- The purpose of creating a representative model is to generalize a pattern or a relationship within a dataset and the presence of outliers skews the representativeness of the inferred model.
- Detecting outliers may be the primary purpose of some data science applications, like fraud or intrusion detection.

Data Science Process-Data Preparation

Feature Selection

- Table has one attribute or feature—the credit score—and one label—the interest rate.
- Many data science problems involve a dataset with hundreds to thousands of attributes. In text mining applications, every distinct word in a document forms a distinct attribute in the dataset.
- Not all the attributes are equally important or useful in predicting the target.
- The presence of some attributes might be counterproductive. Some of the attributes may be highly correlated with each other, like annual income and taxes paid.
- A large number of attributes in the dataset significantly increases the complexity of a model and may degrade the performance of the model due to the curse of dimensionality.
- More detailed information is desired in data science for discovering nuggets of a pattern in the data. But, as the number of dimensions in the data increase, data becomes sparse in high-dimensional space.
- Reducing the number of attributes, without significant loss in the performance of the model, is called feature selection.

Data Science Process-Data Preparation

Data Sampling

- Sampling is a process of selecting a subset of records as a representation of the original dataset for use in data analysis or modeling.
- The sample data serve as a representative of the original dataset with similar properties. Sampling reduces the amount of data that need to be processed and speeds up the build process of the modeling.
- In the build process for data science applications, it is necessary to segment the datasets into training and test samples. The training dataset is sampled from the original dataset using simple sampling or class label specific sampling.
- Consider the example cases for predicting anomalies in a dataset (e.g., predicting fraudulent credit card transactions).
- The objective of anomaly detection is to classify the outliers in the data. These are rare events and often the dataset does not have enough examples of the outlier class.
- Stratified sampling is a process of sampling where each class is equally represented in the sample; this allows the model to focus on the difference between the patterns of each class that is, normal and outlier records.
- In classification applications, sampling is used create multiple base models, each developed using a different set of sampled training datasets. These base models are used to build one meta model, called the ensemble model, where the error rate is improved.

Data Science Process-Modelling

- A model is the abstract representation of the data and the relationships in a given dataset.
- A simple rule of thumb like "mortgage interest rate reduces with increase in credit score" is a model.

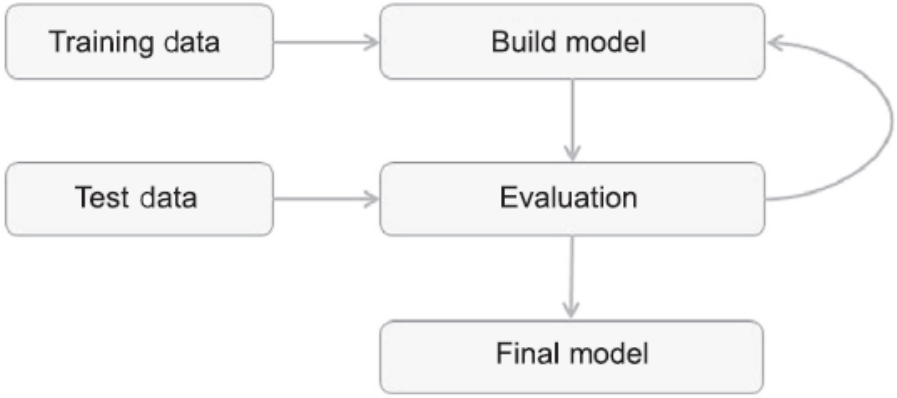


FIGURE 2.4
Modeling steps.

- Classification and regression tasks are predictive techniques because they predict an outcome variable based on one or more input variables.
- Predictive algorithms require a prior known dataset to learn the model.
- Association analysis and clustering are descriptive data science techniques where there is no target variable to predict; hence, there is no test dataset. However, both predictive and descriptive models have an evaluation step.

Data Science Process-Modelling

Training and Testing Dataset

- The dataset used to create the model, with known attributes and target, is called the training dataset.
- The validity of the created model will also need to be checked with another known dataset called the test dataset or validation dataset.
- To facilitate this process, the overall known dataset can be split into a training dataset and a test dataset.
- A standard rule of thumb is two-thirds of the data are to be used as training and one-third as a test dataset.

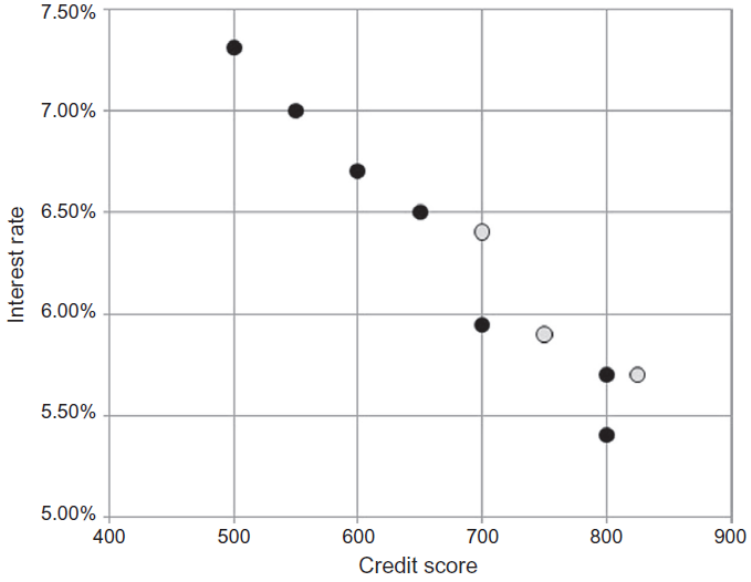


FIGURE 2.5
Scatterplot of training and test data.

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-Modelling

Learning Algorithms

- The business question and the availability of data will dictate what data science task (association, classification, regression, etc.,) can be used.
- The practitioner determines the appropriate data science algorithm within the chosen category.
- For example, within a classification task many algorithms can be chosen from: decision trees, rule induction, neural networks, Bayesian models, k-NN, etc. Likewise, within decision tree techniques, there are quite a number of variations of learning algorithms like classification and regression tree (CART), CHi-squared Automatic Interaction Detector (CHAID) etc.
- It is not uncommon to use multiple data science tasks and algorithms to solve a business question.

Data Science Process-Modelling

Learning Algorithms

- Interest rate prediction is a regression problem.
- A simple linear regression technique will be used to model and generalize the relationship between credit score and interest rate.
- The training set of seven records is used to create the model and the test set of three records is used to evaluate the validity of the model.
- The objective of simple linear regression can be visualized as fitting a straight line through the data points in a scatterplot.

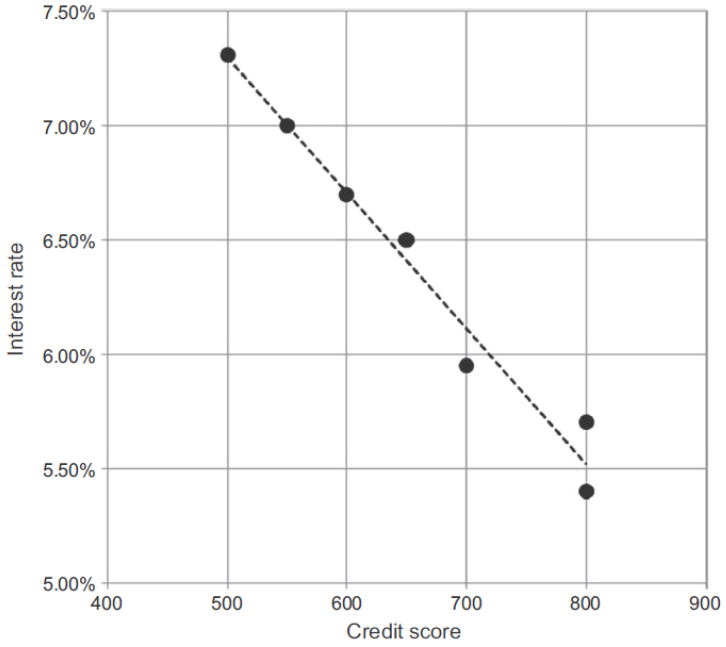


FIGURE 2.6
Regression model.

Data Science Process-Modelling

Learning Algorithms

- The line has to be built in such a way that the sum of the squared distance from the data points to the line is minimal.
- The line can be expressed as: $y = a * x + b$

where y is the output or dependent variable, x is the input or independent variable, b is the y-intercept, and a is the coefficient of x.

- The values of a and b can be found in such a way so as to minimize the sum of the squared residuals of the line.
- The line shown in Eq. serves as a model to predict the outcome of new unlabeled datasets.
- For the interest rate dataset, the simple linear regression for the interest rate (y) has been calculated as,

$$y = 0.1 + \frac{6}{100,000}x$$

$$\text{Interest rate} = 10 - \frac{6 \times \text{credit score}}{1000}$$

Data Science Process-Modelling

Evaluation of the Model

- The model generated in the form of an equation is generalized and synthesized from seven training records.
- The credit score in the equation can be substituted to see if the model estimates the interest rate for each of the seven training records.
- The estimation may not be exactly the same as the values in the training records.
- A model should not memorize and output the same values that are in the training records.
- The phenomenon of a model memorizing the training data is called **overfitting**.
- An overfitted model just memorizes the training records and will underperform on real unlabeled new data.
- The model should generalize or learn the relationship between credit score and interest rate.
- To evaluate this relationship, the validation or test dataset, which was not previously used in building the model, is used for evaluation.

Data Science Process-Modelling

Evaluation of the Model

- Table provides the three testing records where the value of the interest rate is known; these records were not used to build the model.
- The actual value of the interest rate can be compared against the predicted value using the model, and thus, the prediction error can be calculated.
- As long as the error is acceptable, this model is ready for deployment.
- The error rate can be used to compare this model with other models developed using different algorithms like neural networks or Bayesian models, etc.

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

Data Science Process-Modelling

Ensemble Modelling

- Ensemble modeling is a process where multiple diverse base models are used to predict an outcome.
- The motivation for using ensemble models is to reduce the generalization error of the prediction.
- As long as the base models are diverse and independent, the prediction error decreases when the ensemble approach is used.
- Even though the ensemble model has multiple base models within the model, it acts and performs as a single model.
- Most of the practical data science applications utilize ensemble modeling techniques.

Data Science Process-Modelling

At the end of the modeling stage of the data science process, one has

(1) analyzed the business question;

(2) sourced the data relevant to answer the question;

(3) selected a data science technique to answer the question;

(4) picked a data science algorithm and prepared the data to suit the algorithm;

(5) split the data into training and test datasets;

(6) built a generalized model from the training dataset;

(7) validated the model against the test dataset.

Data Science Process-Application

- Deployment is the stage at which the model becomes production ready or live.
- The model deployment stage has to deal with: assessing model readiness, technical integration, response time, model maintenance, and assimilation.

Production readiness

- The production readiness part of the deployment determines the critical qualities required for the deployment objective.
- Consider two business use cases: determining whether a consumer qualifies for a loan and determining the groupings of customers for an enterprise by marketing function.

Data Science Process-Application

Production readiness- Example 1

- The consumer credit approval process is a real-time endeavor.
- Either through a consumer-facing website or through a specialized application for frontline agents, the credit decisions and terms need to be provided in real-time as soon as prospective customers provide the relevant information.
- It is optimal to provide a quick decision while also proving accurate.
- The decision-making model has to collect data from the customer, integrate third-party data like credit history, and make a decision on the loan approval and terms in a matter of seconds.
- The critical quality of this model deployment is real-time prediction.

Data Science Process-Application

Production readiness- Example 2

- Segmenting customers based on their relationship with the company is a thoughtful process where signals from various customer interactions are collected.
- Based on the patterns, similar customers are put in cohorts and campaign strategies are devised to best engage the customer.
- For this application, batch processed, time lagged data would suffice.
- The critical quality in this application is the ability to find unique patterns amongst customers, not the response time of the model.
- The business application informs the choices that need to be made in the data preparation and modeling steps.

Data Science Process-Application

Technical Integration

- Currently, it is quite common to use data science automation tools or coding using R or Python to develop models.
- Data science tools save time as they do not require the writing of custom codes to execute the algorithm.
- This allows the analyst to focus on the data, business logic, and exploring patterns from the data.
- The models created by data science tools can be ported to production applications by utilizing the Predictive Model Markup Language (PMML) or by invoking data science tools in the production application.
- PMML provides a portable and consistent format of model description which can be read by most data science tools.
- This allows the flexibility for practitioners to develop the model with one tool (e.g., RapidMiner) and deploy it in another tool or application.
- Some models such as simple regression, decision trees, and induction rules for predictive analytics can be incorporated directly into business applications and business intelligence systems easily.
- These models are represented by simple equations and the "if-then" rule, hence, they can be ported easily to most programming languages.

Data Science Process-Application

Response Time

- Data science algorithms, like k-NN, are easy to build, but quite slow at predicting the unlabeled records.
- Algorithms such as the decision tree take time to build but are fast at prediction.
- There are trade-offs to be made between production responsiveness and modeling build time.
- The quality of prediction, accessibility of input data, and the response time of the prediction remain the critical quality factors in business application.

Data Science Process-Application

Model Refresh

- The key criterion for the ongoing relevance of the model is the representativeness of the dataset it is processing.
- It is quite normal that the conditions in which the model is built change after the model is sent to deployment.
- For example, the relationship between the credit score and interest rate change frequently based on the prevailing macroeconomic conditions.
- Hence, the model will have to be refreshed frequently.
- The validity of the model can be routinely tested by using the new known test dataset and calculating the prediction error rate.
- If the error rate exceeds a particular threshold, then the model has to be refreshed and redeployed.
- Creating a maintenance schedule is a key part of a deployment plan that will sustain a relevant model.

Data Science Process-Application

Assimilation

- In the descriptive data science applications, deploying a model to live systems may not be the end objective.
- The objective may be to assimilate the knowledge gained from the data science analysis to the organization.
- For example, the objective may be finding logical clusters in the customer database so that separate marketing approaches can be developed for each customer cluster.
- Then the next step may be a classification task for new customers to bucket them in one of known clusters.
- The association analysis provides a solution for the market basket problem, where the task is to find which two products are purchased together most often.
- The challenge for the data science practitioner is to articulate these findings, establish relevance to the original business question, quantify the risks in the model, and quantify the business impact.

Data Science Process-Application

Assimilation

- In the descriptive data science applications, deploying a model to live systems may not be the end objective.
- The objective may be to assimilate the knowledge gained from the data science analysis to the organization.
- For example, the objective may be finding logical clusters in the customer database so that separate marketing approaches can be developed for each customer cluster.
- Then the next step may be a classification task for new customers to bucket them in one of known clusters.
- The association analysis provides a solution for the market basket problem, where the task is to find which two products are purchased together most often.
- The challenge for the data science practitioner is to articulate these findings, establish relevance to the original business question, quantify the risks in the model, and quantify the business impact.

Data Science Process-Knowledge

- The data science process provides a framework to extract nontrivial information from data.
- With the advent of massive storage, increased data collection, and advanced computing paradigms, the available datasets to be utilized are only increasing.
- To extract knowledge from these massive data assets, data science algorithms are employed in addition to standard business intelligence reporting or statistical analysis.
- Though many of these algorithms can provide valuable knowledge, it is up to the practitioner to skillfully transform a business problem to a data problem and apply the right algorithm.
- The data science process starts with prior knowledge and ends with posterior knowledge, which is the incremental insight gained.
- Not all discovered patterns lead to incremental knowledge.
- Again, it is up to the practitioner to invalidate the irrelevant patterns and identify the meaningful information.
- It is not meant to be used as a set of rigid rules, but as a set of iterative, distinct steps that aid in knowledge discovery.

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.
- Simple pivot table functions, computing statistics like mean and deviation, and plotting data as a line, bar, and scatter charts are part of data exploration techniques that are used in everyday business settings.
- 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, the distribution of the values, and the presence of extreme values and the interrelationships between the attributes in the dataset.

Data Exploration

- Data exploration can be broadly classified into two types—descriptive statistics and 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

- **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.
- **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.

Objectives of Data Exploration

- **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.

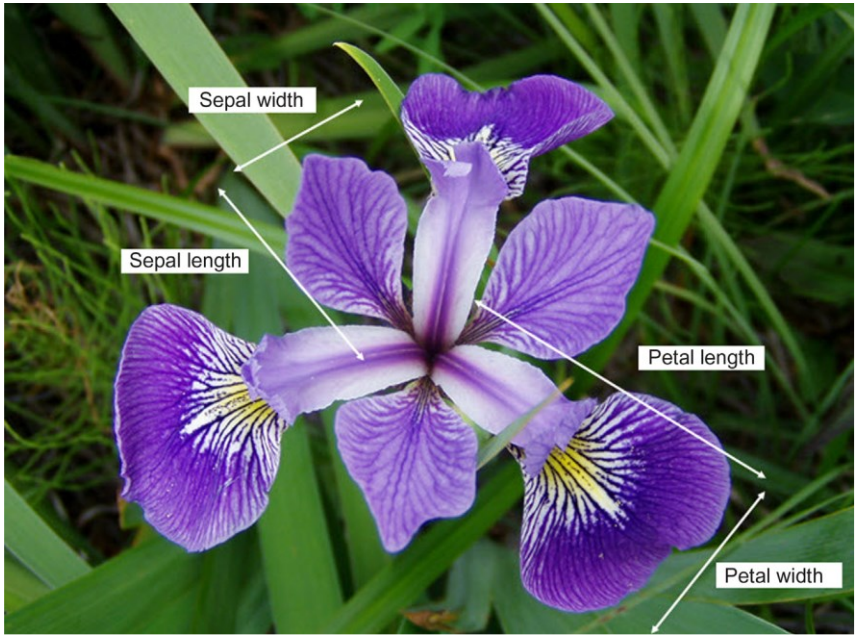
- **Interpreting the results**: Finally, data exploration is used in understanding the prediction, classification, and clustering of the results of the data science process.
- Histograms help to comprehend the distribution of the attribute and can also be useful for visualizing numeric prediction, error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

Data sets

- The most popular datasets used to learn data science is probably the **Iris 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. → class label
- The fifth attribute, the label, is the name of the species observed, which takes the values I. setosa, I. virginica, and I. versicolor.

Data sets

- 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.



Data sets

- 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.
- 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.
- The dataset extends beyond two dimensions, with three class labels, of which one class is easily separable (I. setosa) just by visual exploration, while classifying the other two classes is slightly more challenging.
- It helps to reaffirm the classification results that can be derived based on visual rules, and at the same time sets the stage for data science to build new rules beyond the limits of visual exploration.

Data sets-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°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 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.

Data sets-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.

- 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 $0^{o}C$, 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.

Data sets-Types of data

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.
- 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.
- 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).

Descriptive Statistics

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

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

- Descriptive statistics can be broadly classified into univariate and multivariate exploration depending on the number of attributes under analysis.

Descriptive Statistics-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.

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

Descriptive Statistics-Univariate Exploration

Measure of Central Tendency

- **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.
- **Mean, median, Mode** denote 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.

Descriptive Statistics-Univariate Exploration

Measure of Spread

- In desert regions, it is common for the temperature to cross above 110°F during the day and drop below 30°F during the night while the average temperature for a 24-hour period is around 70°F.
- 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°F, where the temperature within the day is between 60°F and 80°F.
- What matters here is not just the central location of the temperature, but the spread of the temperature.
- **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 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°F and the range for the tropics is 20°F.

Descriptive Statistics-Univariate Exploration

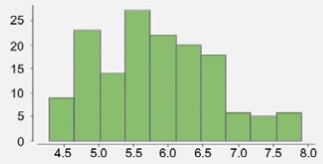
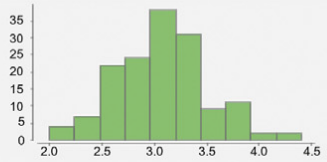
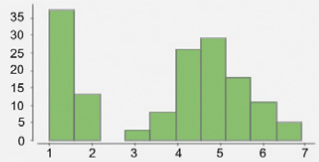
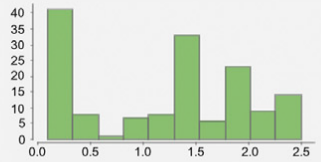
Measure of Spread

- **Deviation:** The variance and standard deviation measures the spread, by considering all the values of the attribute.
- Deviation is simply measured as the difference between any given value (x_i) and the mean of the sample (μ).
- The variance is the sum of the squared deviations of all data points divided by the number of data points.
- For a dataset with N observations, the variance is $\text{Variance} = s^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$
- Standard deviation is the square root of the variance.
- Since the standard deviation is measured in the same units as the attribute, it is easy to understand the magnitude of the metric.
- 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.

Descriptive Statistics-Univariate Exploration

Univariate summary of the Iris dataset with all 150 observations

I. Setosa

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

Descriptive Statistics-Multivariate Exploration

- Multivariate exploration is the study of more than one attribute in the dataset simultaneously.
- This technique is critical to understanding the relationship between the attributes, which is central to data science methods.

Central Data Point

- 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).
- In this way, all 150 observations can be expressed in Cartesian coordinates. 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 shown in Table , the 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.

Descriptive Statistics-Multivariate Exploration

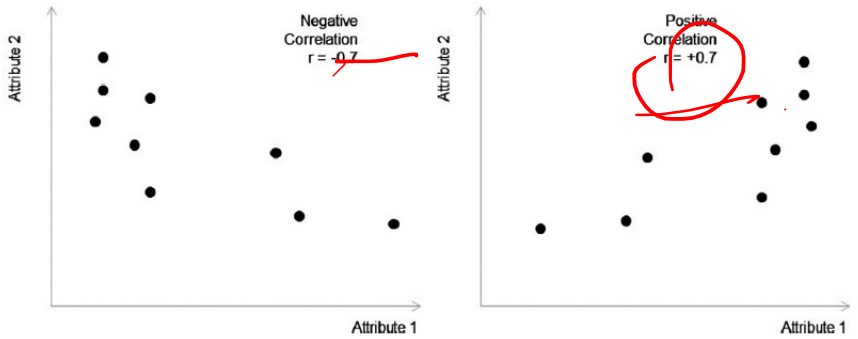
Correlation

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

Descriptive Statistics-Multivariate Exploration

Correlation

- Correlation between two attributes is commonly measured by the Pearson correlation coefficient (r), which measures the strength of linear dependence.



- Correlation coefficients take a value from -1<=r<=1. A value closer to 1 or -1 indicates the two attributes are highly correlated, 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:

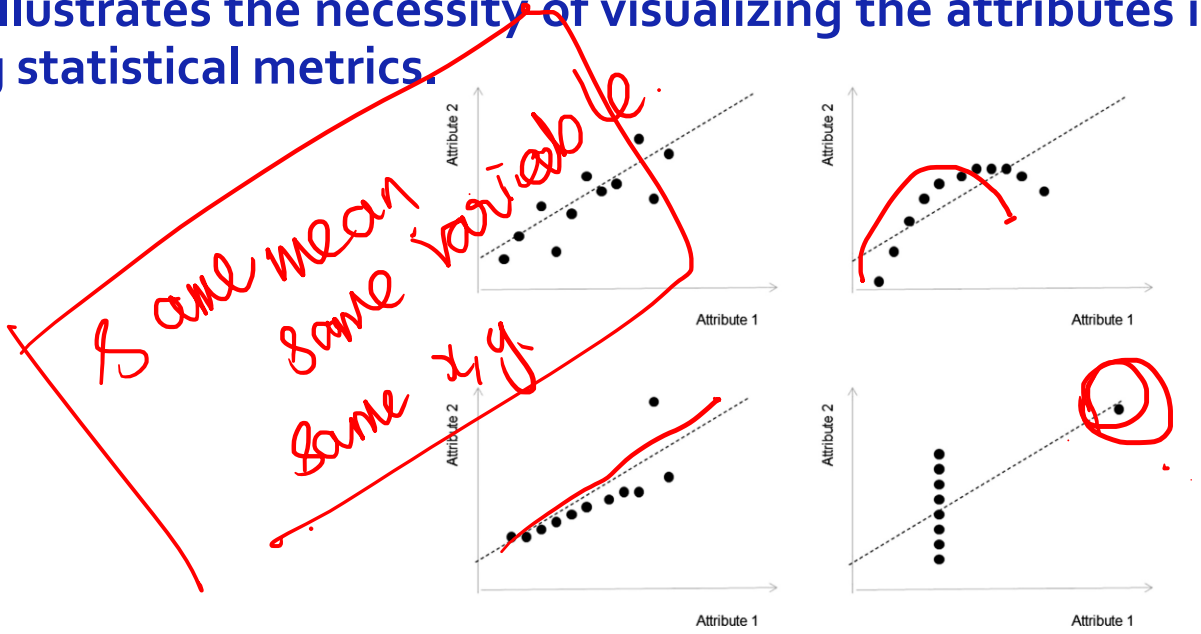
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{N\times s_x\times s_y}$$

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

Descriptive Statistics-Multivariate Exploration

Visualization and Correlation

- 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.
- 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.



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.
- **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.
- Although representing more than three attributes on the x, y, and z-axes is not feasible in Cartesian coordinates, there are a few creative solutions available where more than two attributes are used in a two-dimensional medium.

Data Visualization-Univariate Visualization

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

one attribute -

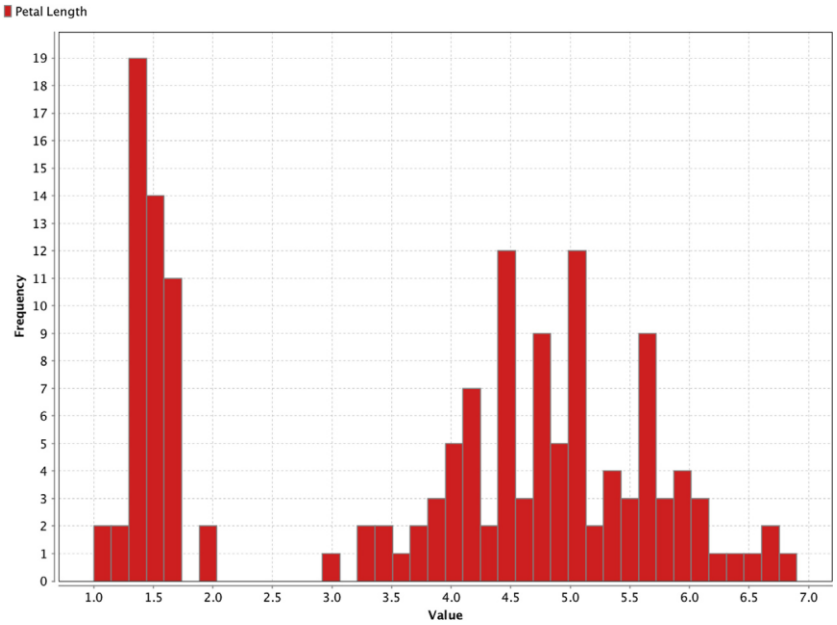
Histogram

- 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.
- For example, in the case of human height in centimeters, all the occurrences between 152.00 and 152.99 are grouped under 152.
- 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.

Data Visualization-Univariate Visualization

Histogram

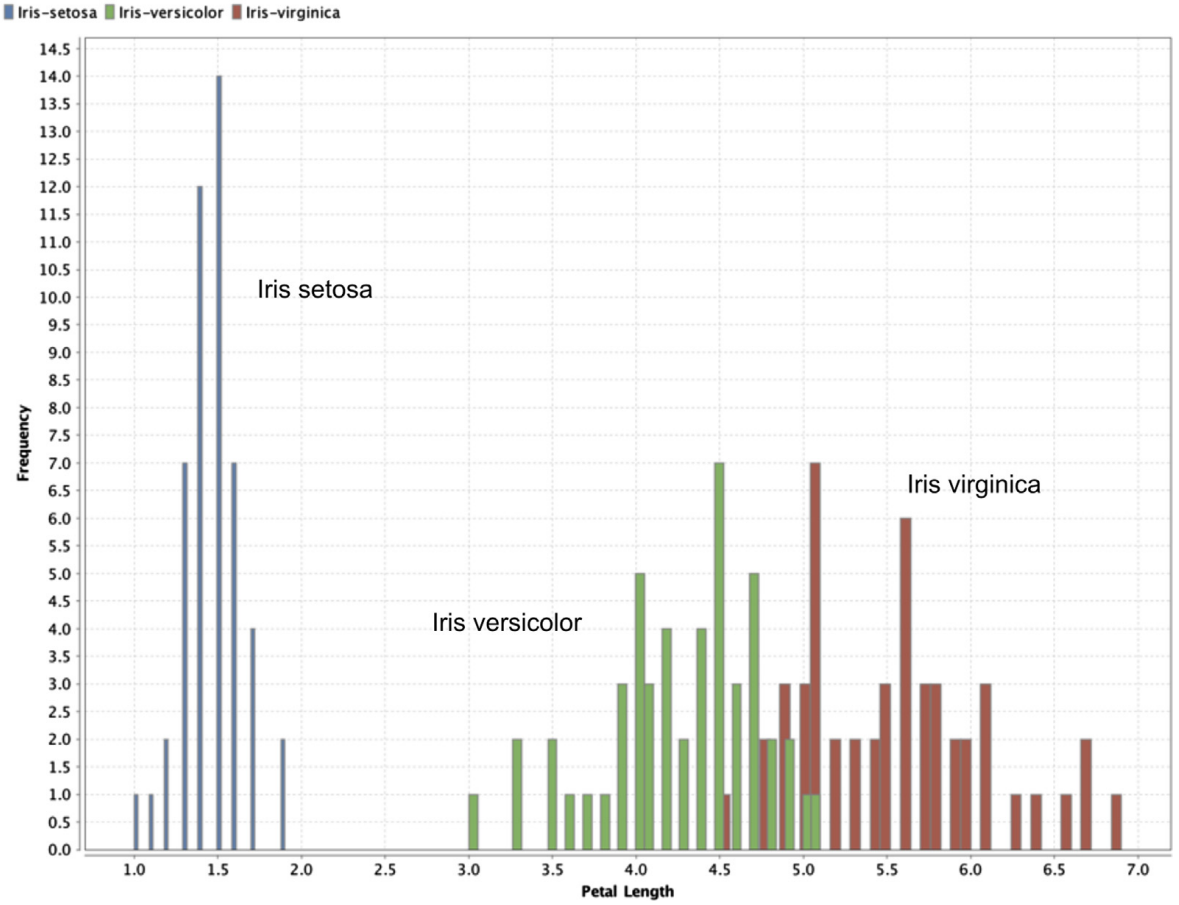
- Histograms are used to find the central location, range, and shape of distribution.
- In the case of the petal length attribute in the Iris dataset, the data is multimodal , where the distribution does not follow the bell curve pattern.
- Instead, there are two peaks in the distribution. This is due to the fact that there are 150 observations of three different species (hence, distributions) in the dataset.
- A histogram can be stratified to include different classes in order to gain more insight.



Data Visualization-Univariate Visualization

Histogram

- The enhanced histogram with class labels shows the dataset is made of three different distributions. I. setosa's distribution stands out with a mean around 1.25 cm and ranges from 1-2 cm. I. versicolor and I. virginica's distributions overlap I. setosa's have separate means.



Data Visualization-Univariate Visualization

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.

Data Visualization-Univariate Visualization

Quartile

- Figure 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.

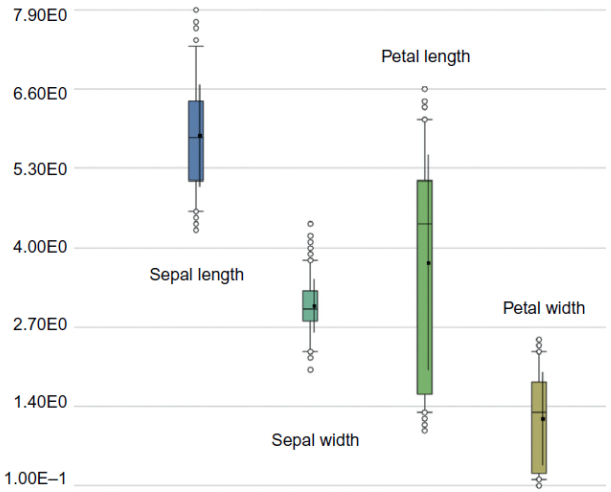


FIGURE 3.7
Quartile plot of Iris dataset.

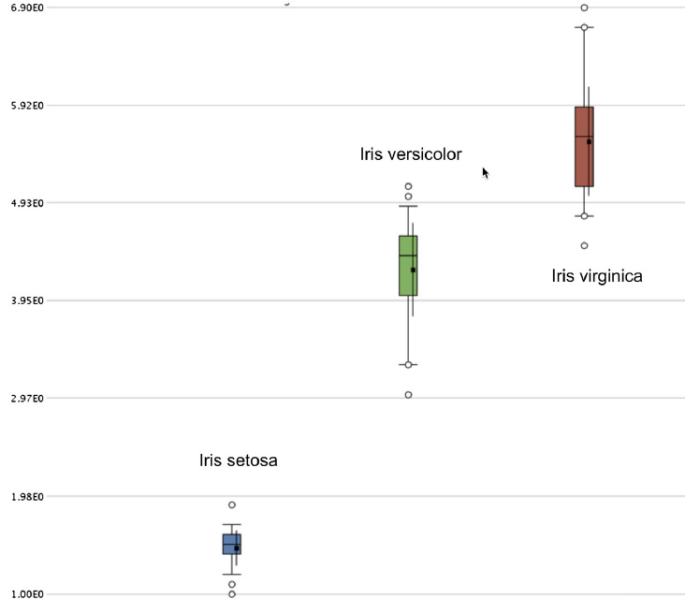


FIGURE 3.8
Class-stratified quartile plot of petal length in Iris dataset.

Data Visualization-Univariate Visualization

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 function 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.

Data Visualization-Univariate Visualization

Distribution chart

- Figure 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.

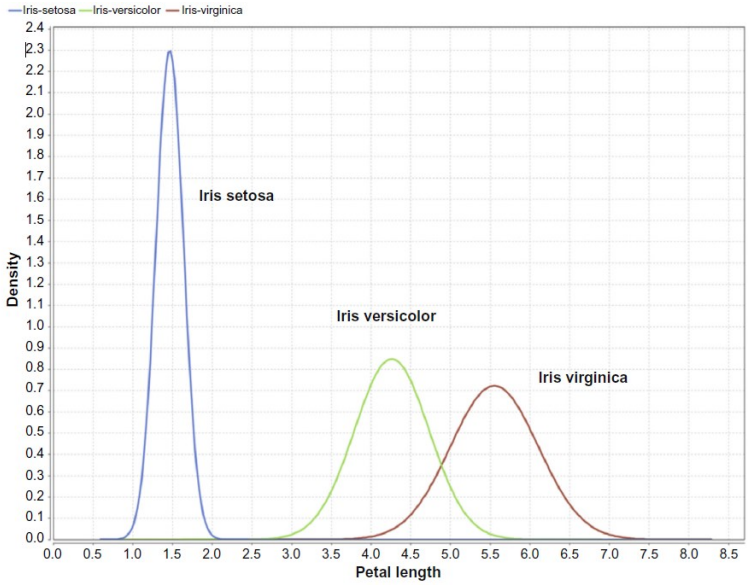


FIGURE 3.9
Distribution of petal length in Iris dataset.

Data Visualization-Multivariate Visualization

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

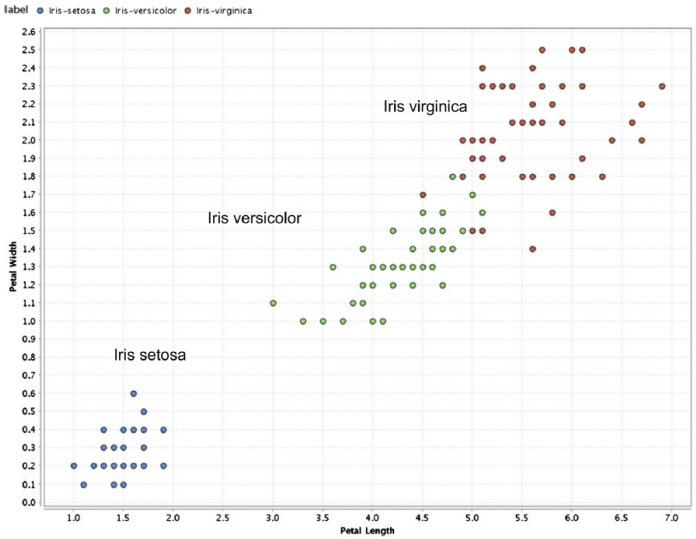
Scatterplot

- 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.

Data Visualization-Multivariate Visualization

Scatterplot

- Figure 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.



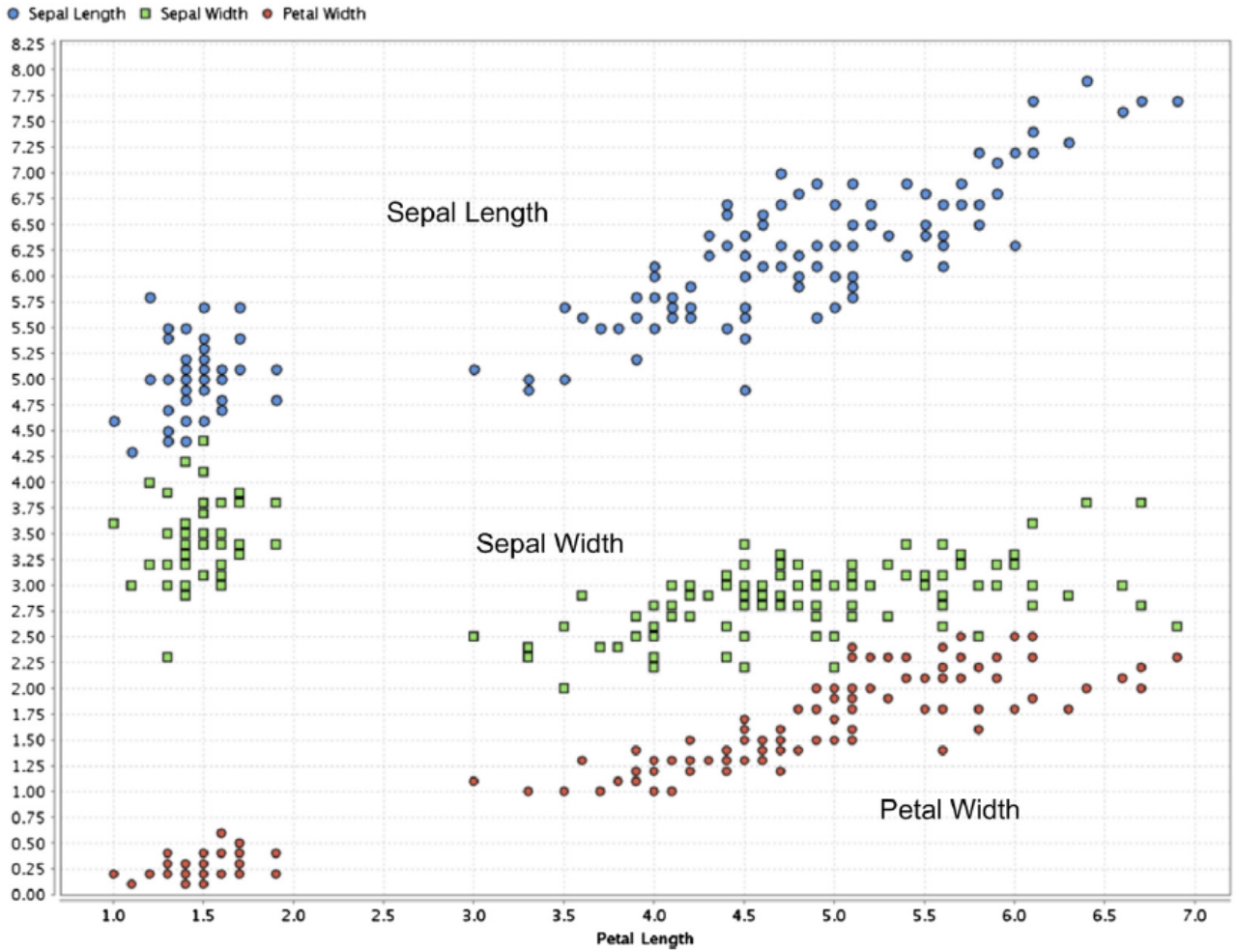
Data Visualization-Multivariate Visualization

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 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.

Data Visualization-Multivariate Visualization

Scatter Multiple



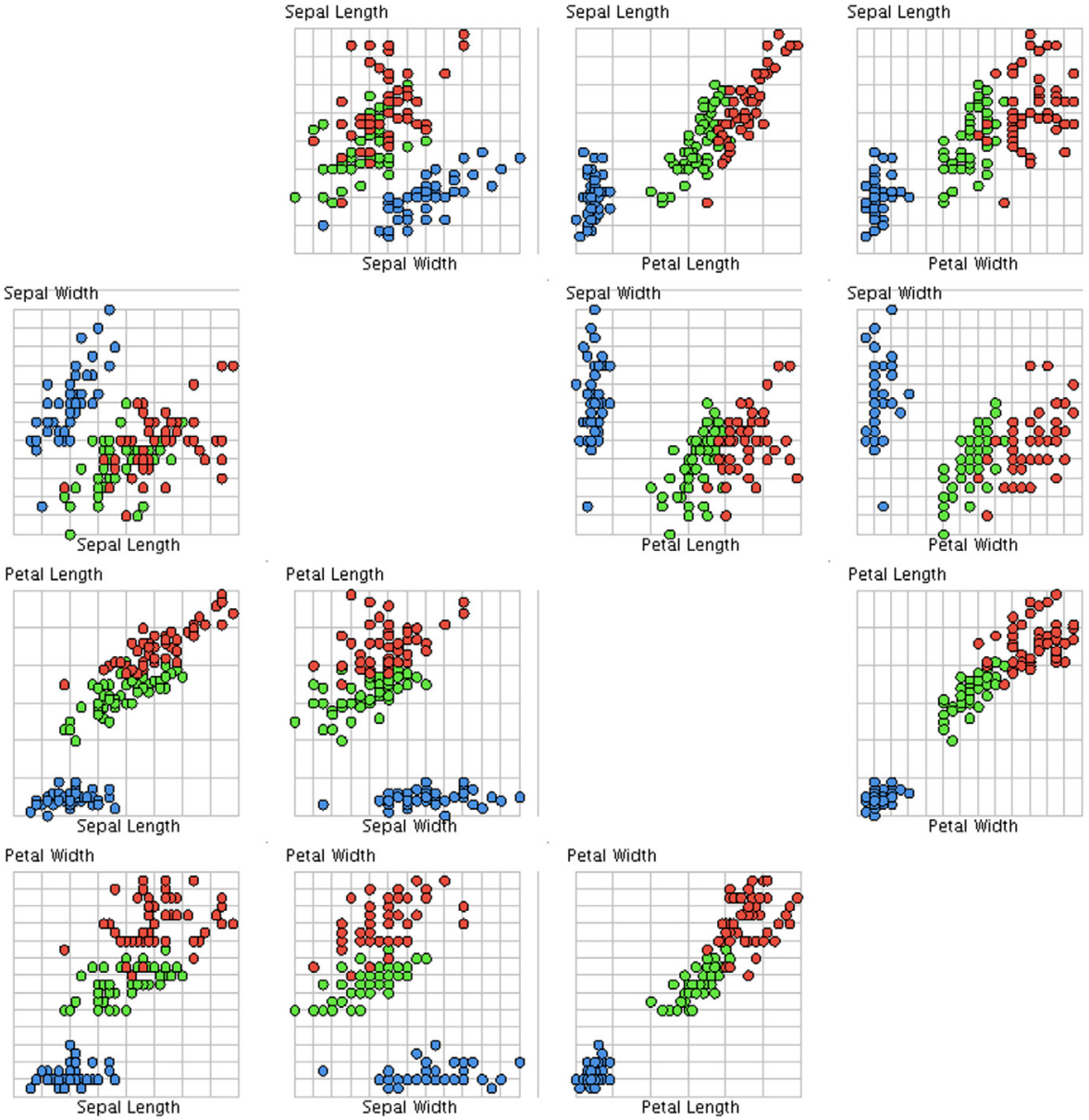
Data Visualization-Multivariate Visualization

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 figure.
- 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.

Data Visualization-Multivariate Visualization

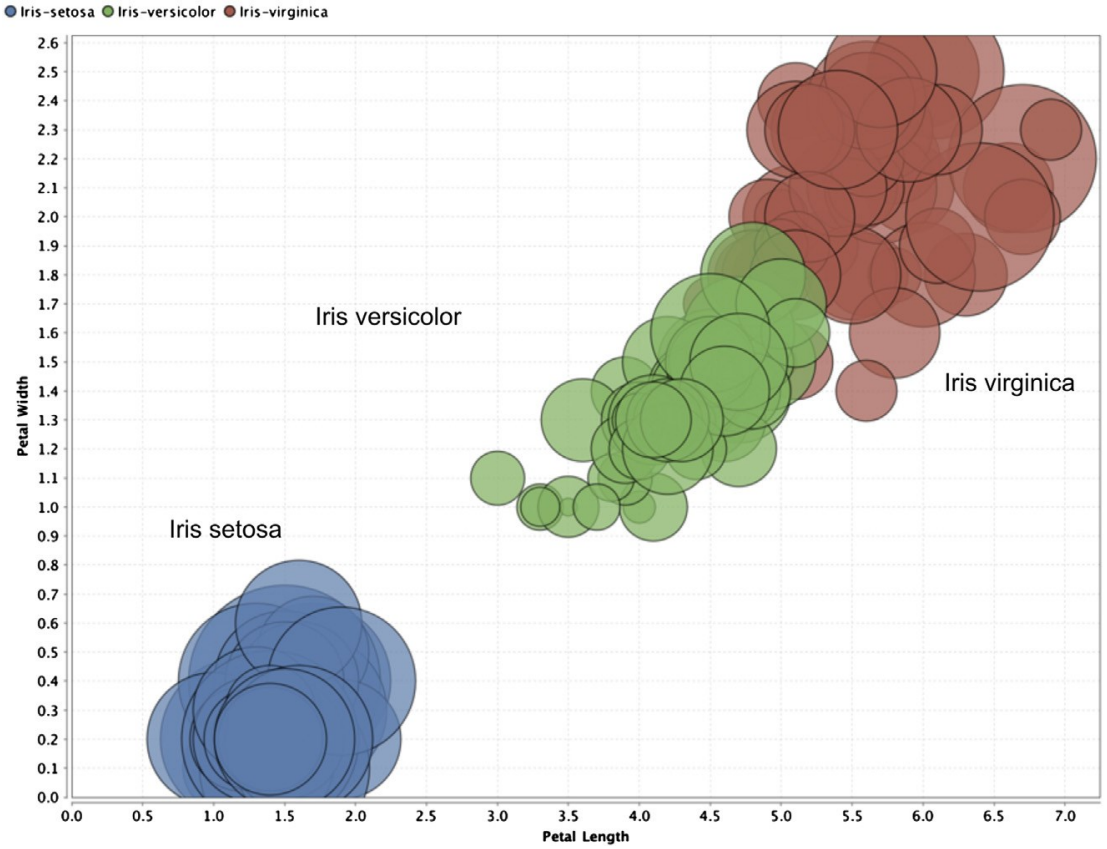
Scatter Matrix



Data Visualization-Multivariate Visualization

Bubble Chart

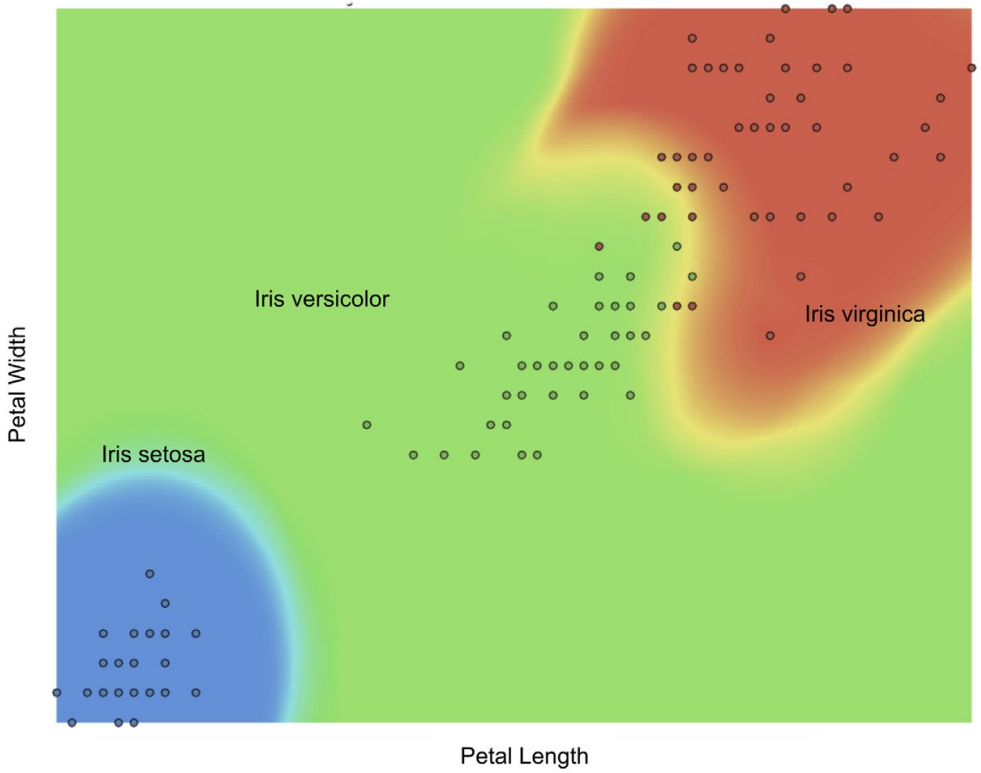
- A bubble chart is a variation of a simple scatterplot with the addition of one more attribute, which is used to determine the size of the data point.
- In the Iris dataset, petal length and petal width are used for x and y-axis, respectively and sepal width is used for the size of the data point.
- The color of the data point represents a species class label.



Data Visualization-Multivariate 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 visualize one dimension, and hence, a total of four dimensions can be visualized in a density chart.
- In the example, petal length is used for the x-axis, sepal length for the y-axis, sepal width for the background color, and class label for the data point color.



Data Visualization-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.

Data Visualization-Visualizing High-Dimensional Data

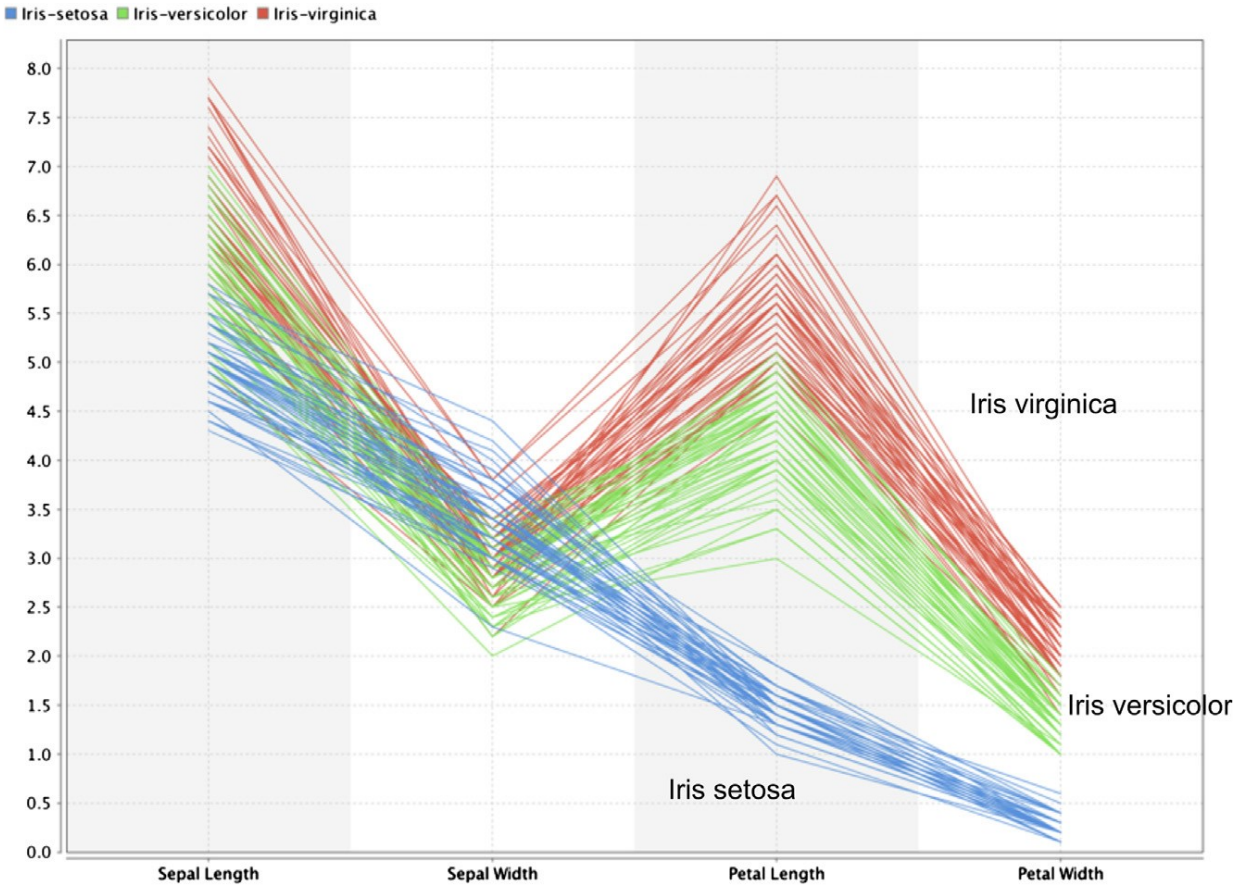
Parallel Chart

- 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 Figure, 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.

Data Visualization-Visualizing High-Dimensional Data

Parallel Chart

- 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.



Data Visualization-Visualizing High-Dimensional Data

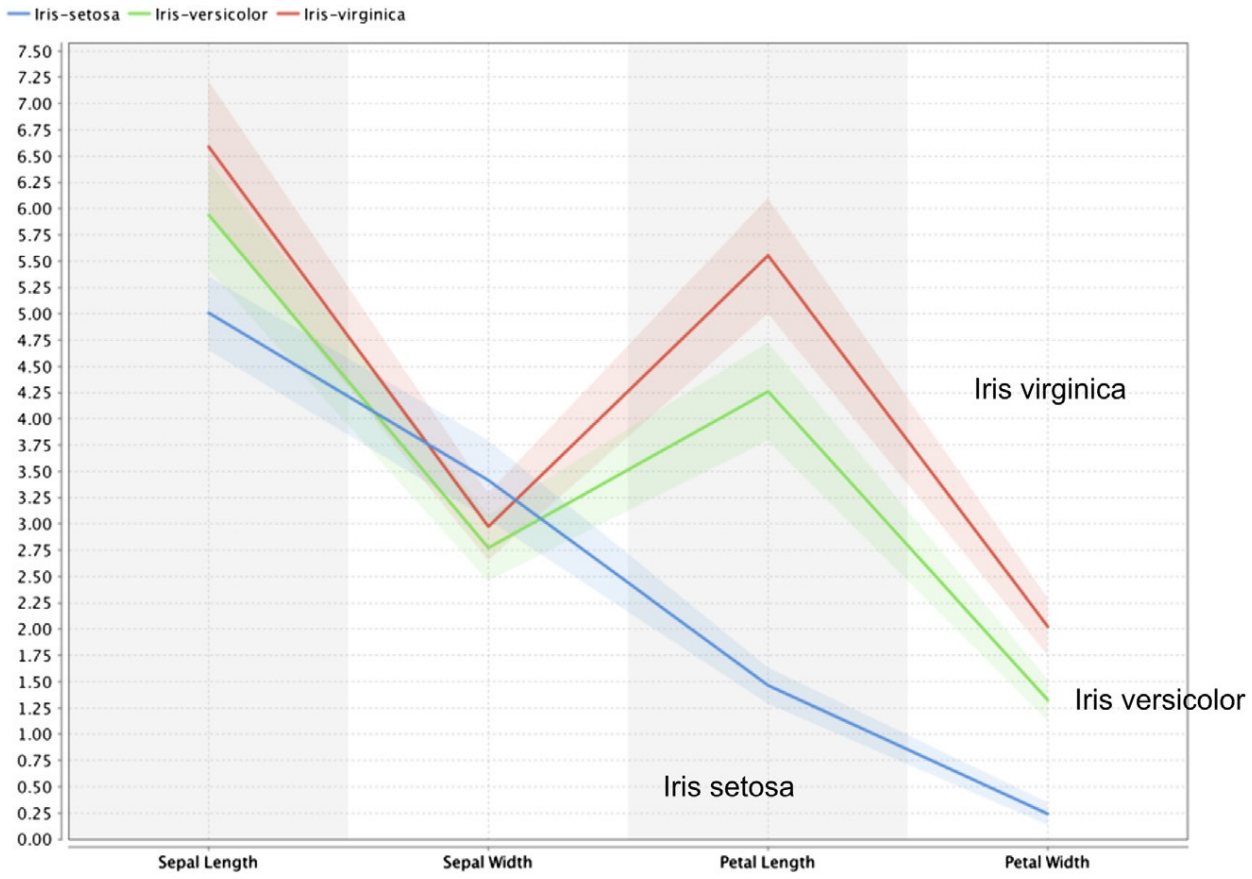
Deviation Chart

- 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.

Data Visualization-Visualizing High-Dimensional Data

Deviation Chart

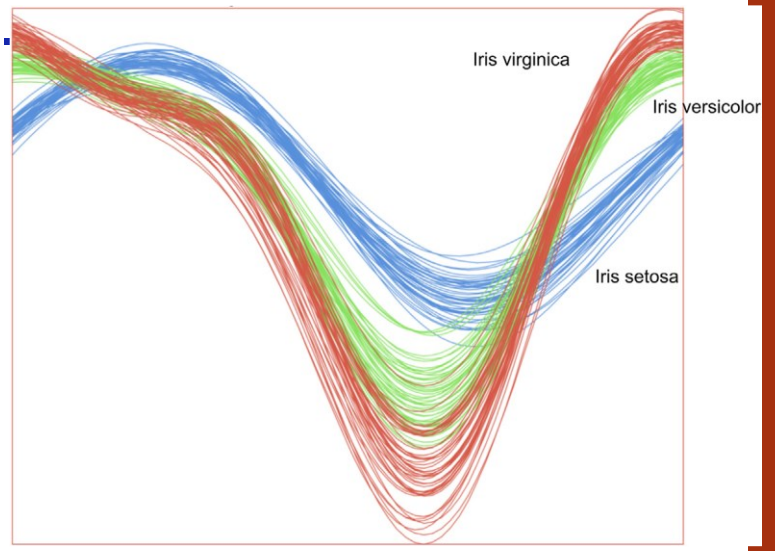
- In Figure, a deviation chart for the Iris dataset stratified by species is shown.
- It can be observed that the petal length is a good predictor to classify the species because the mean line and the standard deviation bands for the species are well separated.



Data Visualization-Visualizing High-Dimensional Data

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=(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 2π , 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.
- 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.



Road Map for Data Visualization

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.

Road Map for Data Visualization

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.

Thank You!!!