20MCA201 DATA SCIENCE & MACHINE LEARNING

Introduction to data science, Data science classification, Data science process - Prior knowledge, Data preparation, Modelling, Application, Data exploration - Data sets, Descriptive statistics for univariate and multivariate data, Data visualization – Histogram, Quartile plot, Distribution chart, Scatter plot, Bubble chart, Density chart

Data Science??

- It is an inter-disciplinary field about scientific methods, processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured.
- Examples of data science user case are: recommendation engine which recommend movies for user,a fraud alert model that detect fraudulent credit card transactions.
- We can define data science by investigating some of its key features:

- Data Exploration:

❖It is the most important step.

❖Around 70 per cent of the time is spent on data exploration.

❖The main ingredient for data science is data

❖When we get data, it is seldom that data is in a correct structured form.

❖Sampling and transformation of data

- Modeling:

❖Data is prepared and ready to go.

❖Model is the representation of a relationship between variables in a dataset.

❖Actually fit the data into the model.

❖The selection of a model depends on the type of data we have and the business requirement.

❖Modeling is a process in which representative abstraction is build from observed data set

- Eg:- Based on credit score, income level and requested loan amount, a model can be developed to determine the interest rate of a loan.
- Once representative model is created, it can be used to predict the value interest rate based on all the input variables.
- Model serves two purpose:-

❖ Predict output based on new and unseen set of input data

❖Model can be used to understand the relationship between output variables and all input variables

- **Combination of statistics, Machine Learning and Computing:**

❖ For extracting useful and relevant information from large dataset, data science borrows computational techniques from the discipline of statistics, machine learning, experimentation and database theories.

❖Algorithm used in the data science originated from these disciplines.

❖One of the key ingredients of successful data science is substantial prior knowledge about data and the business process that generate data known as subject matter expertise.

Learning Algorithms:

- The application of sophisticated learning algorithms for extracting useful patterns from data different data science from traditional data analysis techniques.
- Many of these algorithm were developed in past few decades and are part of machine learning and artificial intelligence.
- Some algorithm are based on foundation of Bayesian Probabilistic theories and regression analysis.

- Based on the problem, data science is classifies into task such as classification, association analysis, clustering and regression.
- Each data science task use special learning algorithm like decision tree, neural networks,k-nearest neighbor etc.

Associated Fields

Technique used in the step of data science process and in conjunction with the term data science are:-

- Descriptive statistics:

Computational mean, standard deviation, correlation and other descriptive quantify the aggregate structure of dataset. This is essential information to understand any dataset.

- Exploratory visualization:

The process of expressing data in visual coordinates enables user to find pattern and relationships in the data and comprehend large dataset.

- Dimensional Slicing:

Online analytical processing(OLAP) application which are prevalent in the organization, mainly provide information on data through dimensional slicing, filtering and pivoting.

- Hypothesis testing:

Data science is the process where many hypothesis are generated and tested based on observational data.

- Data Engineering:

It is the process of sourcing, organizing, assembling, storing and distributing data for effective analysis and usage.

Data Engineering helps source and prepare for data science learning algorithms.

- Business Intelligence:

It helps organizations consume data effectively.

BI specializes in the secure delivery of information to right roles and distribution of information at scale.

It can hold and distribute the result of data science.

- A massive accumulation of data has been seen with the advancement of information technology, connected networks, and the business it enables.
- This trend also coupled with a steep decline in data storage and data processing costs.
- Traditional analysis techniques can only for 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 algorithm to deduce optimal insights from datasets

- Data Science is one such paradigm that can handle large volumes with multiple attributes and deploy complex algorithm to search for patterns from the data.

- Key Motivation:

✓**Volume:**

- A rapid increase in the volume of data exposes the limitation of current analysis methodologies.
- In few implementations, time to create generalization model is critical and data volume plays a major part in determining the time frame of the development and deployment.

Dimensions

- Three character of big data phenomenon are:
- High volume
- High velocity
- High variety(relates to multiple types of values)
- Every single records or data point contain multiple attributes or variables to provide context for the record.

Complex Questions

- As more complex data are available for analysis, the complexity of information that need to be extracted from the data is increasing as well.
- Machine learning algorithm need to be leveraged in order to automate searching in the vast search space.
- ML approaches the problem of modelling by trying to find an algorithm model that can predict the output from input variables

Data Science Classification

- Supervised /Direct Science:

It tries to infer a function or relationship based on labeled training data and use this function to map new unlabeled data.

- 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 variable and use it to predict for a dataset where only input variable are known

Unsupervised/Undirect Data Science:

- Uncover the hidden patterns in unlabeled data.
- Objective of this class is to find pattern in data based on relationship between data points temselves

- Classification:
- Supervised
- It predict the output variable which are categorical or plynominal

Data Preparation

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

- **Steps:**

✓Data Exploration

✓Data Quality

✓Missing Values

✓Data Types and Conversion

✓Transformation

✓Outliers

✓Feature selection

✓Data Sampling

Data Exploration:-

- Data preparation starts with an in-depth exploration of data and gaining a better understanding of dataset.
- 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

✓Visualization of Data

➢These two approaches 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 a readable summary of the key characteristics of the distribution of data.

- Visual plot of data points provide an instant grasp of all data points condensed into one chart.

- Data Quality:

✓It is an ongoing concern, wherever data is collected, processed and stored.

❖How does one know if the data are accurate?

❖Data entry error

❖Error in data will impact the representativeness of the model

❖Organization use data alerts, cleansing and transformation techniques to improve and manage the quality of the data and stored them in company wide repositories called ***datawarehouse.***

- Data sourced from well maintained data warehouse have higher quality.

✓Missing Values:

- One of the most common data quality issues is that some records have missing attribute values
- Eg:-Credit Score may be missing in one of the records.
- The first step of managing missing values is

❖ To understand the reason behind *why the values are missing.*

- Tracking the data lineage of the data source can lead to the identification of systemic issues during data capture or errors in data transformation.
- Knowing the source of missing value will often guide which mitigation methodology to use.
- One way to manage missing value is it can be substituted with range of artificial data. This method is useful if the missing value 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 reduce the size of the datasets.

- Some data science algorithm are good at handling the missing values, while other expect the data preparation step to handle it before model is inferred.
- Eg:-K-NN algorithm for classification tasks are often robust with missing values. Neural network model for classification tasks do not perform well with missing attributes, thus data preparation is essential for preparing neural networks model.

- Data Types and Conversion:

✓The attributes in the dataset can be of different types, such as continuous numeric, integer numeric or categorical

✓Different data science algorithm impose different restrictions on the attribute data types

✓Eg: In case of linear regression models, the input attributes have to be numeric. If the available data are categorical, they must be converted to continuous numeric attribute

- Numeric values can be converted to categorical data type by a technique called binning, where range of value are specified for each category.

❖Transformation:

- In some data science algorithms like k-NN, the input attributes are expected to be numeric and normalized, the algorithm compares the values of different attributes and calculates distance between the data points

- Normalization prevents one attribute dominating the distance result because of large value.

❖Outliers:

- Outliers are anomalies in a given dataset.
- It may occur because of correct data capture or erroneous data capture(human height as 1.73cm instead of 1.73m)

❖Feature Selection:

- Not all the attributes are equally important or useful in predicting the target.
- A large number of attributes in the dataset significantly increases the complexity of a model and may degrade the performance of the model
- ***Reducing the number of attributes without significantly loss in the performance of model, is called feature selection.***
- It leads to a more simplified model and helps to synthesize a more effective explanation of the model

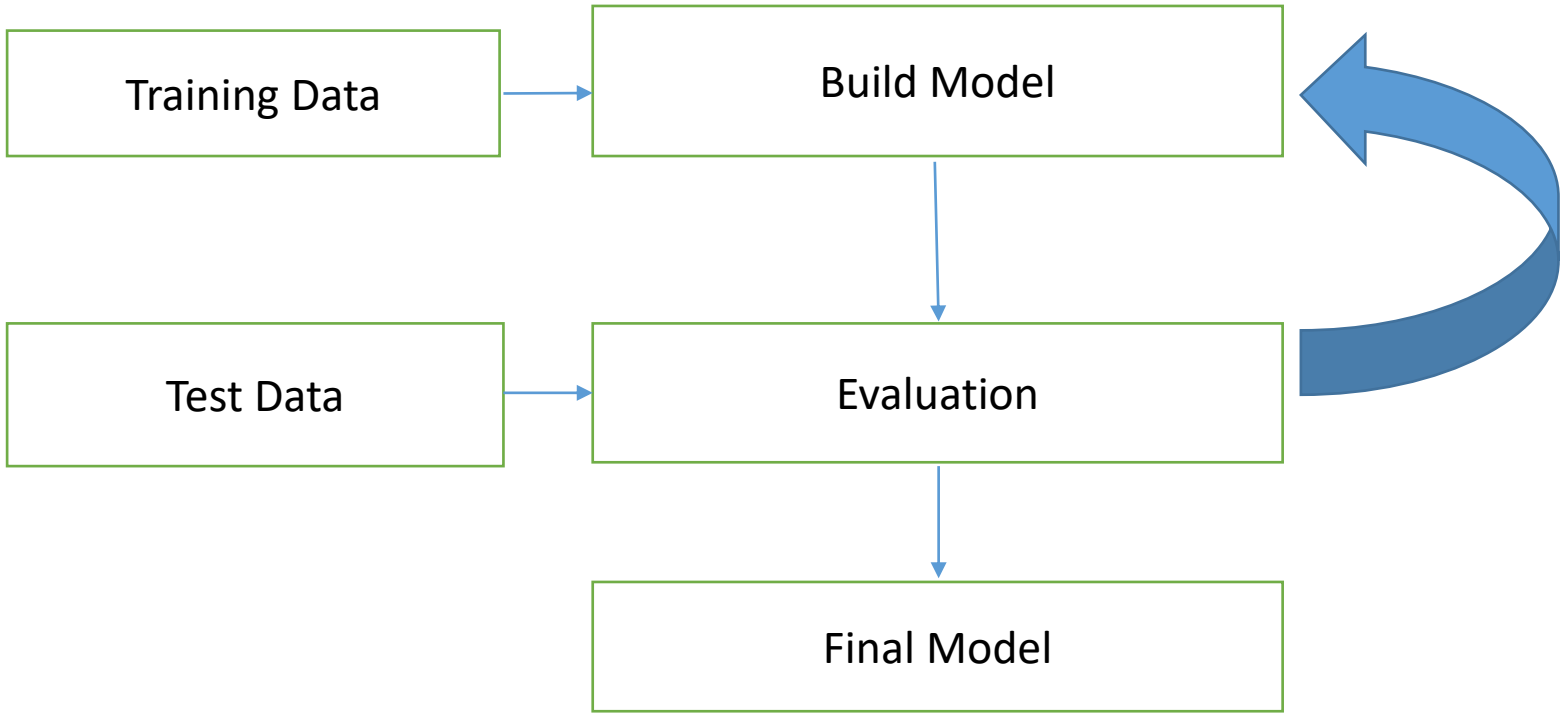
❖Data sampling:

- Sampling is the process of selecting subset of records as a representation of original dataset for use in data analysis or modeling.
- Sampling reduces the amount of data that need to be processed and speedup the build process of the modeling.
- In the build process of data science applications, it is necessary to segment dataset into training and test samples.
- Training dataset is sampled from the original dataset using simple sampling or class label specific sampling.

- Eg: Predicting anomalies in a dataset.
- 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 pattern of each class that is,normal and outlier records.

Modeling

- It is an abstract representation of the data and the relationships in a given dataset
- Classification and regression task are predictive techniques because they predict an outcome variable based on one or more input variables.
- Predictive algorithm requires a prior known dataset to learn model.
- Association and clustering are descriptive data science techniques where there is no target variable to predict.


Training Data
Build Model
Test Data
Evaluation
Final Model

✓Training and Testing Datasets:

- Overall dataset can be split into a training and test dataset. Two- third of data are to be used as training and one-third as test dataset
- The dataset used to create the model , with known attributes and target is called the training dataset.
- The validity of created model will also need to checked with another known dataset called the test dataset or validation dataset.

✓Learning Algorithm:

- The business question and the availability of data will dictate what data science task can be used.

✓Evaluation of the Model:

- The phenomenon of a model memorizing the training dataset is called overfitting.
- An overfitted model is just memorizes the training records and will underperform on real unlabeled new data.
- To evaluate the relationship, the validation or test dataset is used.

Ensemble Modeling

- It is a process where multiple diverse base model are used to predict the outcome.
- The prediction error decreases when the ensemble approach is used.
- Even though the ensemble model has multiple base model within the the model,it acts and perform as a single model.

- Applications:

✓**Production Readiness:**

- It determines the critical quality required for the deployment objective.
- Consider two business use cases:

1. Determining whether the consumer qualifies for a loan
2. Determining the grouping of customers for an enterprise by marketing function.

1. **Determining whether the consumer qualifies for a loan:**

- The consumer credit approval process is a real-time endeavor.
- The credit decision and terms need to be provided in real-time as soon as prospective customer provide relevant information.
- It is optimal to provide quick decision while also proving accurate.
- The decision making model collect data from the customer, integrate third party data like credit history, and make a decision on the loan approval and terms in matter of seconds.

- The critical quality of this model deployment is real-time prediction.

2. **Determining the grouping of customers for an enterprise by marketing function.**

- Segmenting customers based on their relationship with the company is 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

- The critical quality in this application is the ability to find unique pattern amongst customers, not the response time of the model.

✓Technical Integration:

- Data science tools save time as they do not require the writing of custom codes to execute the algorithm.
- This allows 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 application by utilizing the Predictive Model Markup Language(PMML).

- Response Time:

✓Data Science algorithm like k-NN are easy to build, but quite slow at predicting the unlabeled records.

✓Algorithm such as decision tree take time to build to build but are fast at prediction.

✓There is trade-off 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.

- Model refresh:

✓It is quite normal that the condition in the model is build change after the the model is sent to deployment.

✓Eg: 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 refreshed and redeployed.

- Assimilation:

✓In the descriptive data science applications, deploying a model to live system may not be the end objective.

✓The objective may be to assimilate the knowledge gained from the data science analysis to the organization.

Data Exploration

- Objectives:-

✓*Data Understanding*:-Data Exploration provides a high level overview of each attribute in the dataset and interaction between the attributes. Data exploration helps answer the question like what is the typical value of an attribute, presence of extreme values .

✓Data Preparation:-Before applying data science algorithm, 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 attribute.

- Data science task: Basic data exploration can sometimes substitute the entire data science process. Eg: 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 result of data science process.

- Dataset:-

✓Most popular dataset used to learn data science is probably the *iris dataset.*

✓*The genus* of iris contains more than 300 different species.

✓Each species exhibits different physical characteristics like shape and size of the flower and leaves.

✓The iris dataset contains 150 observations of three different species,iris setosa,iris virginica, and versicolor with 50 observations each.

- Each observation consists of four attributes: sepal length, sepal width, petal length, petal width.
- All four attributes in the Iris dataset are numeric continuous values measured in centimeters.
- The iris dataset is used for learning data science mainly because it is to understand, explore and can be used to illustrate how different data science algorithms approach the problem on the same standard dataset

- Types of Data:

✓Data come in different format and types.

✓Eg; The temperature in weather data can be expressed as any of the following format

- Numeric centigrade or Fahrenheit or on the kelvin scale.
- Ordered labels as in hot, mild or cold
- Number of days within a year below zero degree Celsius.

✓Numeric or Continuous:

- Temperature expressed in centigrade or Fahrenheit is numeric and continuous because it is denoted by numbers.
- Additive and subtractive mathematical operations and logical comparison operators like greater than,less than and equal to , operations can be applied
- If zero point is defined , numeric data become a ratio real data types

  eg:-temperature in kelvin scale,bank account balance etc.

- Categorical or Nominal:

❖Categorical data types are attributes treated as distinct symbols or just names.eg:-Color of iris of human eyes is a categorical data type because it takes a value like black, green, blue, gray etc.

❖An ordered nominal data type is a special case of categorical data type because it takes values like black, green,blue,gray etc.

Descriptive Statistics

- It refers to the study of the aggregate quantities of a dataset.
- It is broadly classifies into

✓Univariate

✓Multivariate

- Univariate Exploration:

✓ It denotes the analysis of one attribute at a time.

✓Descriptive statistics used:

- Measure of Central Tendency:

➢Mean: It is calculated by summing all the data points and dividing by number of data points

➢Median

➢Mode

- Measure of spread:
- Range: It is the difference between the maximum value and minimum value of the attribute. It 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
- Deviation: Deviation is simply measured as the difference between any given value and the mean of the sample. The variance is the sum of the squared deviations of all data points divided by the number of data points.

- Multivariate Exploration:

✓It is the study of more than one attribute in the dataset simultaneously.

✓This technique is critical to understanding the relationship between the attribute

✓Central Data Point:

- In the Iris Dataset, each data point as a set of all four attributes can be expressed:

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

- Eg:- Observation one:{5.1,3.5,1.4,0.2}
- This observation can also be expressed in four dimension Cartesian coordinates and plotted in graph.
- If the objective is to find the most “typical” observation point, it would be the data point made up of the mean of each attribute in the dataset independently.
- Eg: In Iris dataset, the central mean point is {5.006,3.418,1.464,0.244}

- ✓Correlation:
- ✓It measures the statistical relationship between two attributes, particularly dependence of one attribute on another attribute.
- ✓When two attribute are highly correlated with each other, they both vary at the same rate with each other either in the same or in opposite directions

  eg: Average temperature and ice cream sales.
- ✓ Statistically, two attributes that are correlated are dependent on each other and one may be used to predict the other.

- Correlation between two attributes is commonly measured by the Pearson correlation coefficient(r), which means strength of linear dependence.
- Correlation take a value from $-1 \leq r \leq 1$.
- A value closer to 1 or -1 indicates the two attributes are highly correlated with perfect correlation at 1 or-1.
- A correlation value of 0 means there is no linear relationship between two attributes.

Data visualization

- Visualizing data is one of the most important techniques of data discovery and exploration.
- The discipline of data visualization encompasses the method of expressing data in an abstract visual form.
- The motivation for using data visualization include:

✓**Comprehension of dense information**: A simple visual chart can easily include thousands of data points.

✓**Relationships**:

- Visualizing data in Cartesian coordinates enables exploration of the relationships between the attributes.

- Representing more than three attributes on x , y, z-axes is not feasible in Cartesian coordinates, there are few creative solutions available by changing the properties like size, color and shape of data markers or using flow maps.

Univariate Visualization

- Histogram:

✓It is the one of the most basic visualization technique to understand the frequency of the occurrence of value.

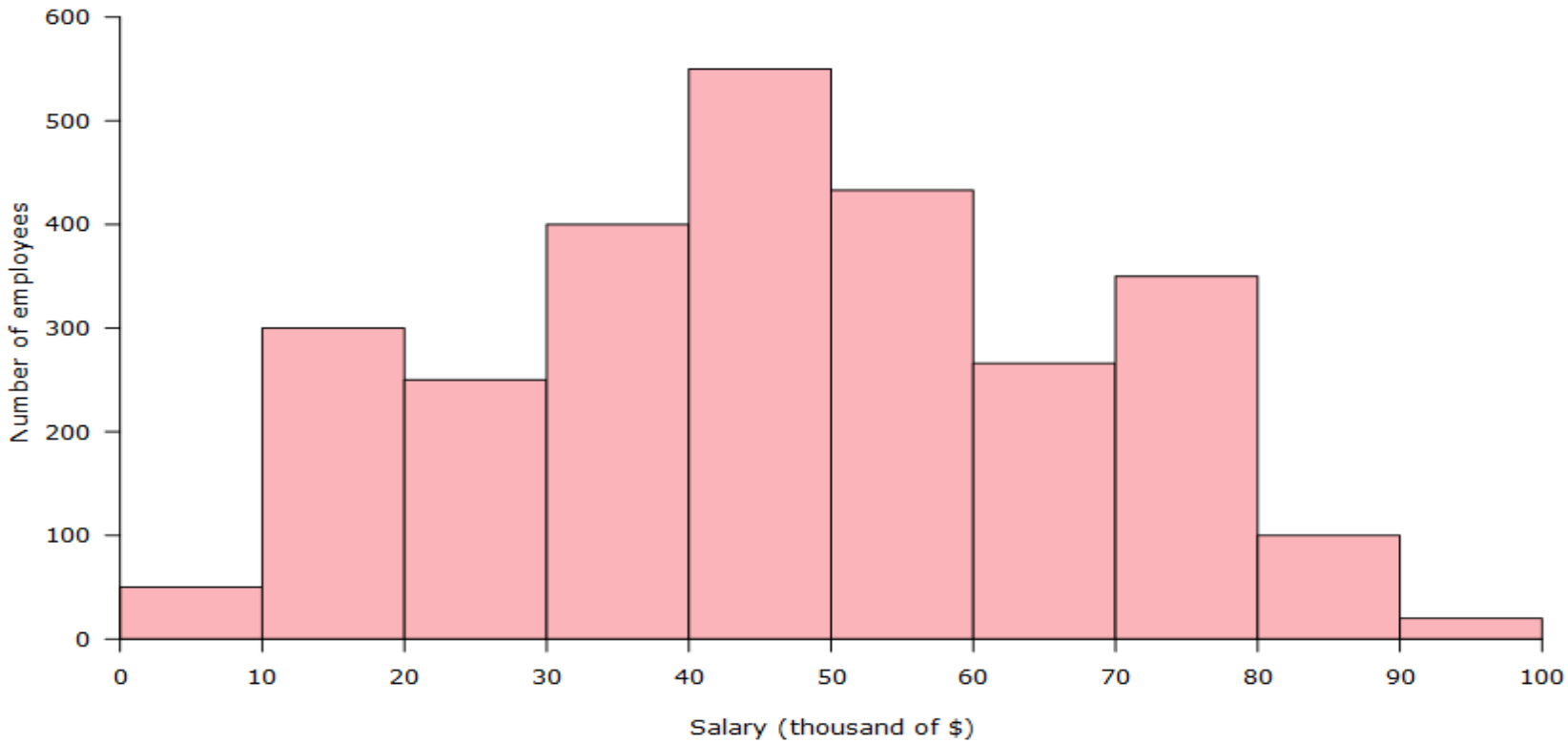
✓It shows the distribution of the data by plotting the frequency of occurrence in a range.

✓In histogram, attributes under inquiry is shown on the horizontal axis and frequency of occurrence on vertical axis.

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

- Eg: In case of human height in centimeters, all the occurrence between 152.00 and 152.99 are grouped under 152.
- Histogram are used to find the central location, range and shape of distribution.

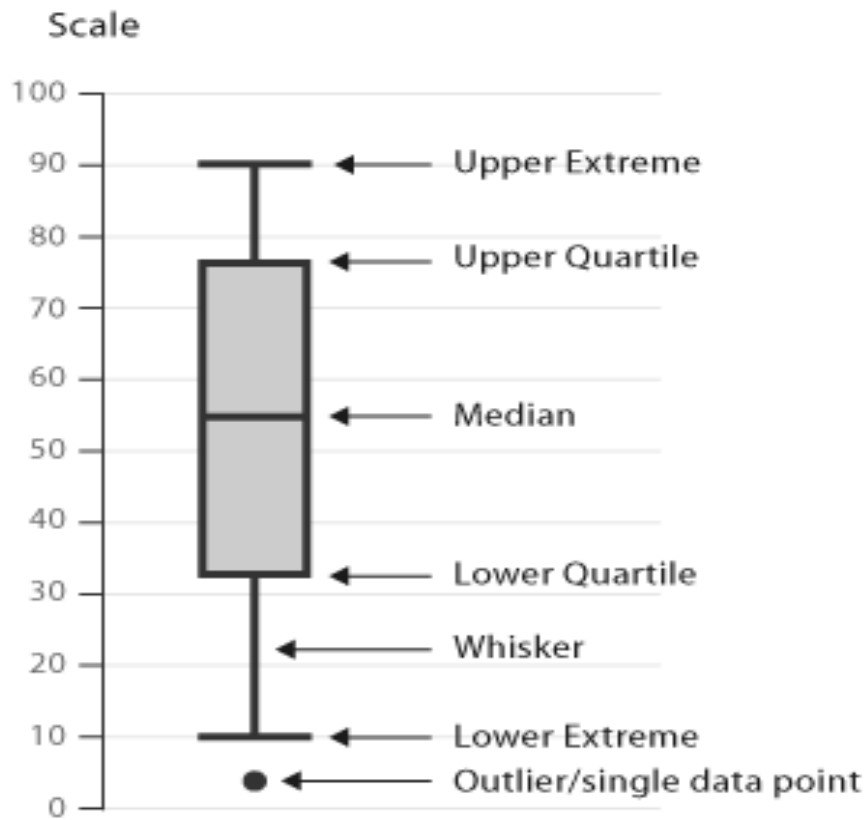
Chart 5.7.1
Distribution of salaries of the employees of ABC Corporation



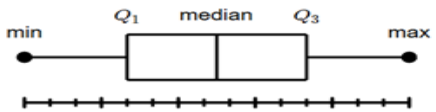
Quartile:

- The quartile are denoted by Q1,Q2 and Q3 points, which indicates 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.
- A box whisker plot is a simple visual way of showing the distribution of a continuous variable with information such as quartile, median and outlier overlaid by mean and standard deviation.

- The main attraction of box whisker or quartile charts is that distribution of multiple attributes can be compared side by side and the overlap between them can be deduced.
- The Q1 and Q3 points in a box whisker plot are denoted by the edges of the box.
- The Q2 point is the median of the distribution.
- The outliers are denoted by circles at the end of whisker line.


Scale
100
90
80
70
60
50
40
30
20
10
0
Upper Extreme
Upper Quartile
Median
Lower Quartile
Whisker
Lower Extreme
Outlier/single data point

- The box plot displays five number summary of a set of data.
- The five number summary is the minimum, first quartile,median,third quartile and maximum.



Example: Finding the five-number summary

A sample of 10 boxes of raisins has these weights (in grams):

$25, 28, 29, 29, 30, 34, 35, 35, 37, 38$

Make a box plot of the data.

Step 1: Order the data from smallest to largest.

Our data is already in order.

$25, 28, 29, 29, 30, 34, 35, 35, 37, 38$

Step 2: Find the median.

The median is the mean of the middle two numbers:

$25, 28, 29, 29, \mathbf{30}, \mathbf{34}, 35, 35, 37, 38$

$$\frac{30 + 34}{2} = 32$$

The median is 32.

Step 3: Find the quartiles.

The first quartile is the median of the data points to the *left* of the median.

$25, 28, \mathbf{29}, 29, 30$

$$Q_1 = 29$$

The third quartile is the median of the data points to the *right* of the median.

34, 35, $\mathbf{35}$, 37, 38

$Q_3 = 35$

Step 4: Complete the five-number summary by finding the min and the max.

The min is the smallest data point, which is 25.

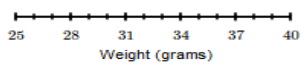
The max is the largest data point, which is 38.

The five-number summary is 25, 29, 32, 35, 38.

Example (continued): Making a box plot

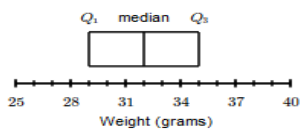
Let's make a box plot for the same dataset from above.

Step 1: Scale and label an axis that fits the five-number summary.



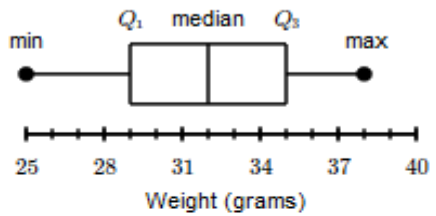
Step 2: Draw a box from Q_1 to Q_3 with a vertical line through the median.

Recall that $Q_1 = 29$, the median is 32, and $Q_3 = 35$.

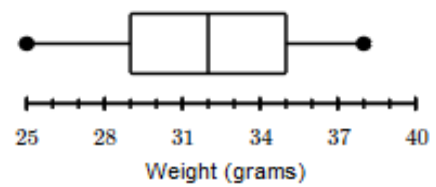


Step 3: Draw a whisker from Q_1 to the min and from Q_3 to the max.

Recall that the min is 25 and the max is 38.



We don't need the labels on the final product:



- **Distribution Chart:**
- For continuous numeric attributes, instead of visualizing the actual data in a sample, its normal distribution function can be visualized instead.
- The normal distribution function of a continuous random variable is given by the formula:

$$f(x)=\frac{1}{\sqrt{2\pi\sigma}}\, e^{(x-\mu)^2/2\sigma^2}$$

- The normal distribution is called the Gaussian distribution.
- The normal distribution function shows the probability of occurrence of data points within the range of values.
- If the dataset exhibits normal distribution, then 68.2% of data points will fall within one standard deviation from the mean,95.4% of the data points will fall within 2 standard deviation from the mean, 99.7% within 3 standard deviation of the mean.

- When normal distribution curves are stratified by class type, more insight into the data can be gained.

Multivariate Visualization

- This visualization examine two to four attributes simultaneously.
- **Scatterplot:**

✓In Scatterplot, data points are marked in Cartesian space with attributes of dataset are aligned within the coordinates.

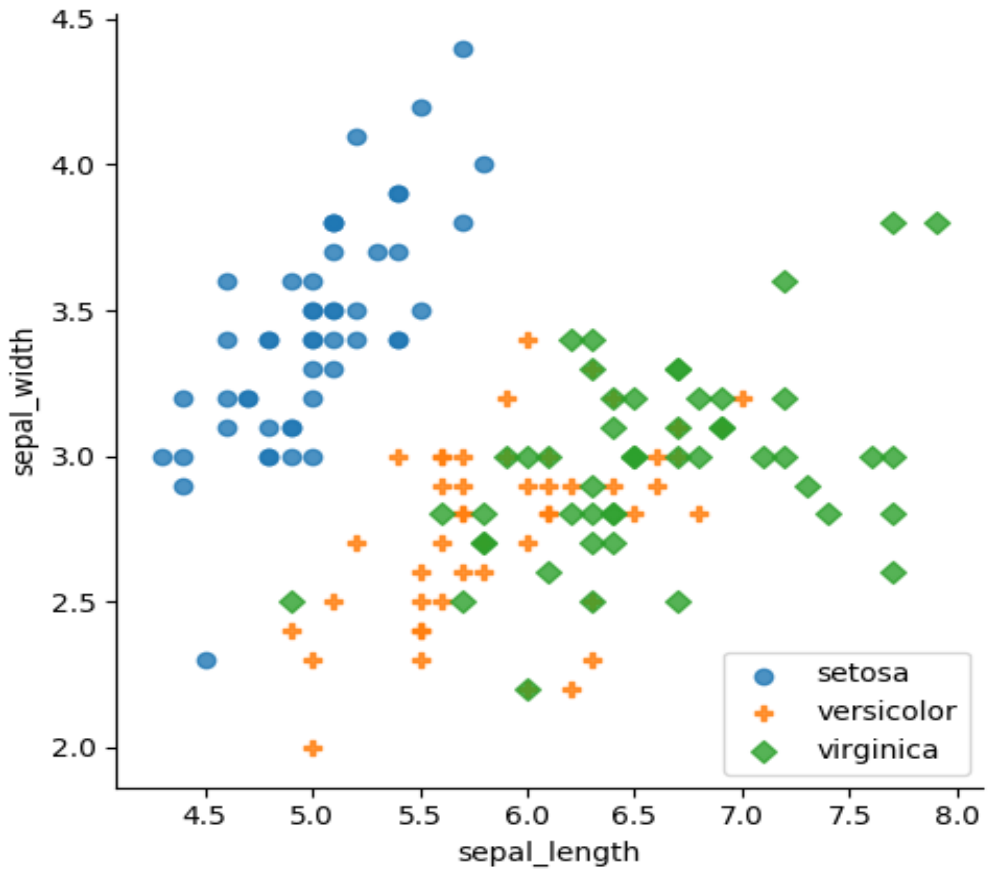
✓The attributes are usually of continuous data type.

✓If the attributes are not correlated, the data points are scattered.

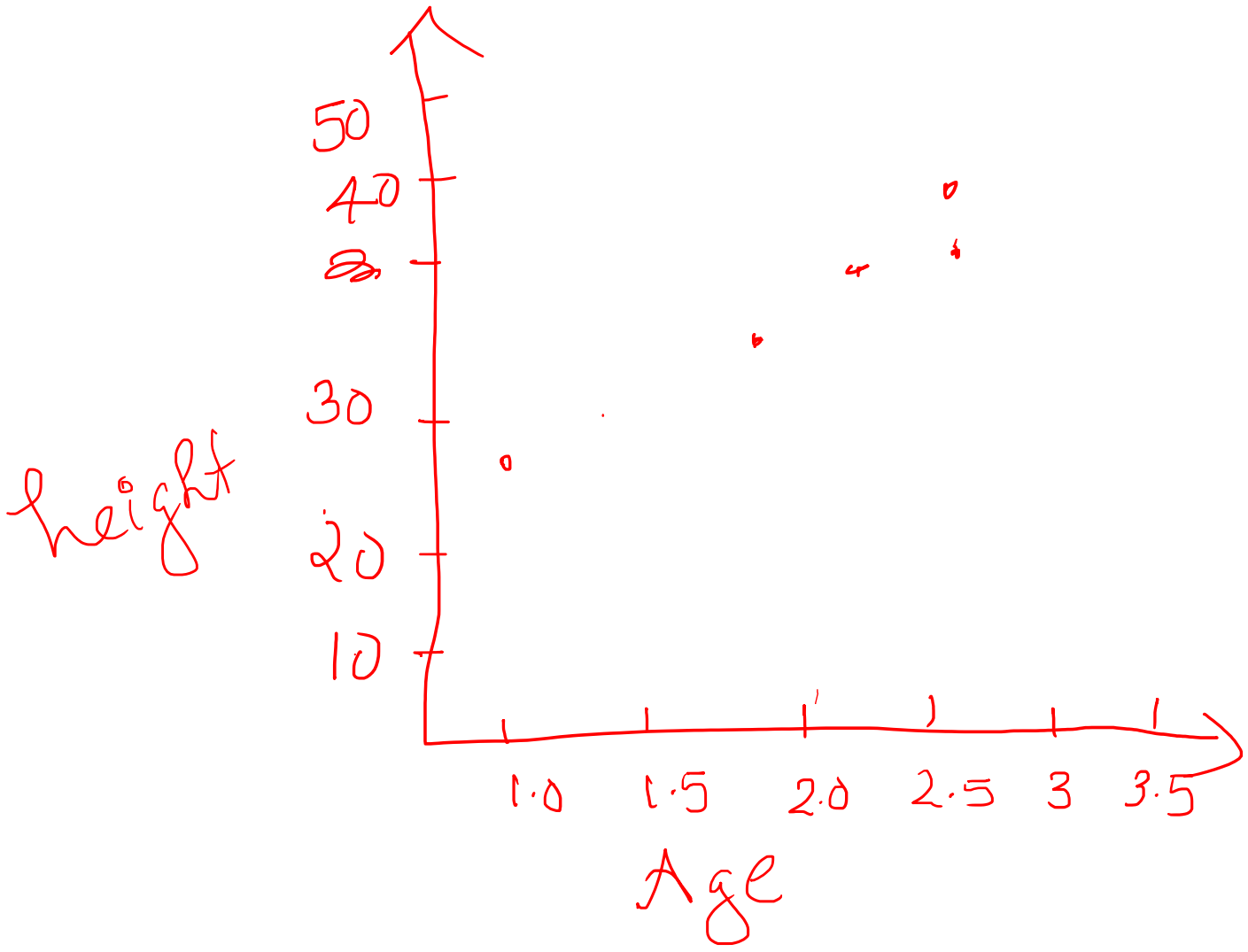
✓Scatterplot can also indicate the existence of pattern or group of clusters in the data and identify outliers in the data.

✓This is particularly useful for low-dimensional dataset.

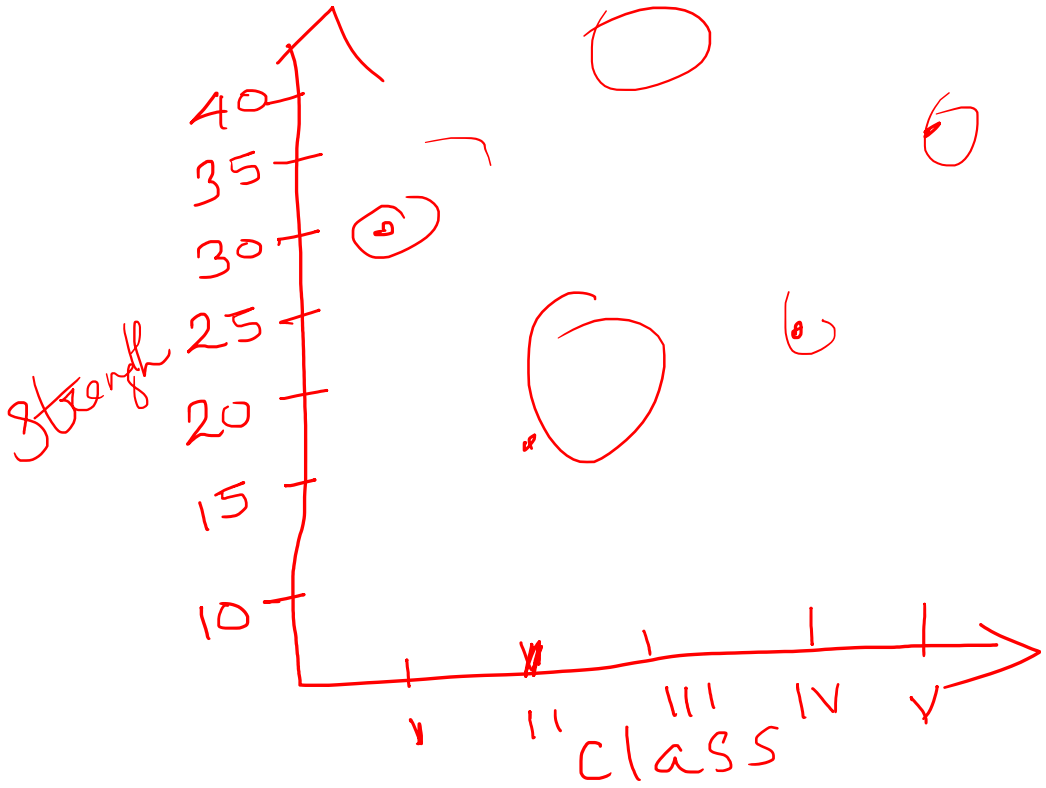
- The limitation of scatterplots is that only two attributes can be used at a time, with an additional attribute possibly shown in the color of the data marker.
- But colors are usually reserved for class labels.


4.5
4.0
3.5
3.0
2.5
2.0
sepal_width
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
sepal_length
setosa
versicolor
virginica

Age	Height
1.0	29.4
1.5	30.0
2.0	33.7
2.5	34.3
3.0	35.0
3.5	37.5
4.0	39.0
4.5	43.9



class	Strength
I	30
II	15
III	40
IV	20
V	34



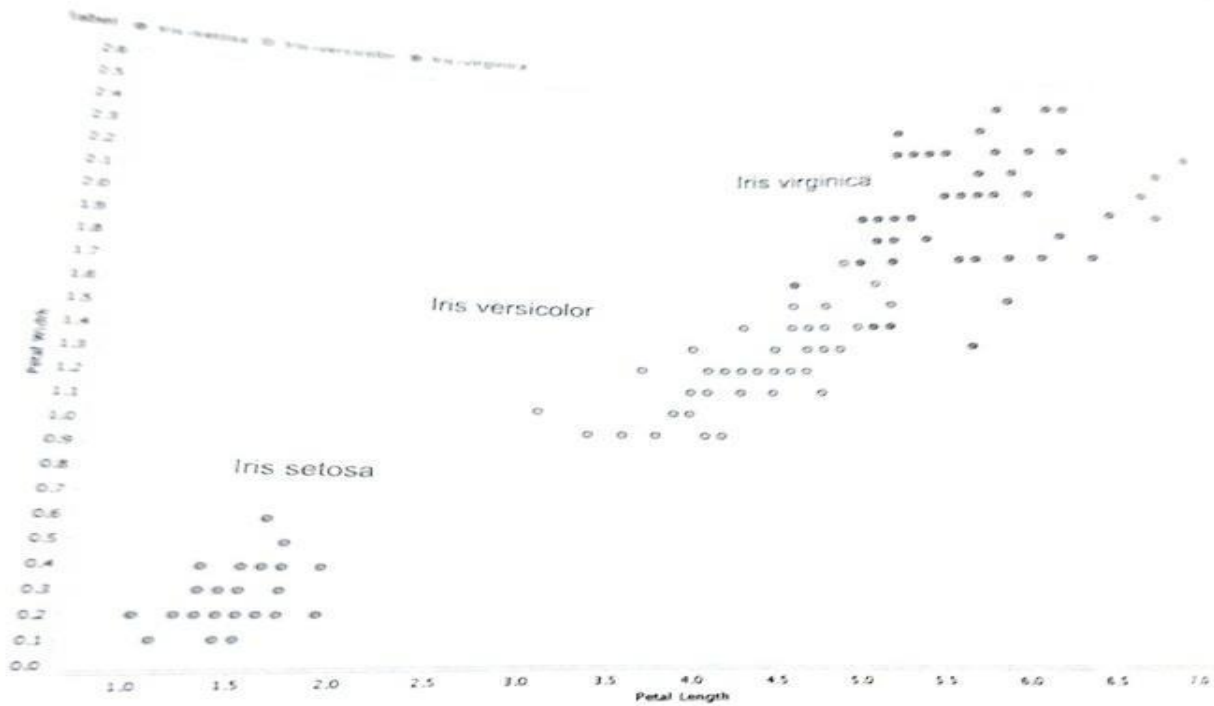


FIGURE 3.10
Scatterplot of Iris dataset.

- Scatter Multiple:

✓Enhanced form of a simple scatterplot

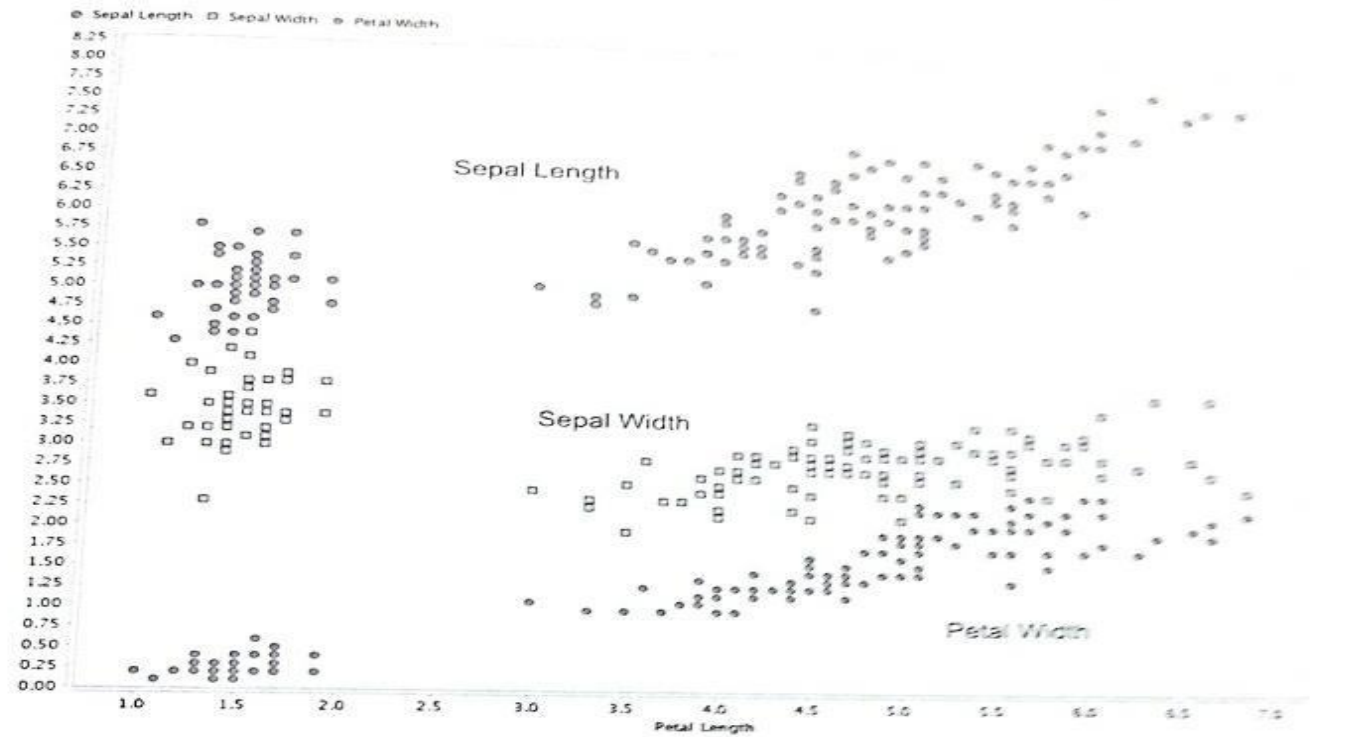
✓More than dimensions can be included in the chart and studied simultaneously.

✓The primary attribute is used for x-axis coordinate.

✓The secondary axis is shared with more attributes or dimensions.

✓Data points are color-coded for each dimension.

✓All the attributes sharing y-axis should be of same unit or normalized.



GURE 3.11

atter multiple plot of Iris dataset.

- Scatter Matrix

✓If the dataset has more than two attributes, it is important to look at the combination of all attributes through a scatterplot.

✓Scatter matrix solves this need by comparing all combination of attributes with individual scatterplot and arranging these plots in the matrix.

✓In Iris dataset, there are four attributes, so there are four rows and columns, for total of 16 scatter charts.

- ✓Charts in the diagonal are a comparison of the attribute with itself; hence they are eliminated.
- ✓Charts below the diagonal are mirror images of charts above the diagonal.
- ✓In effect, there are six distinct comparisons in scatter multiples of four attributes.
- ✓It provides an effective visualization of comparative, multivariate and high density data displayed in small multiples of the similar scatterplots.

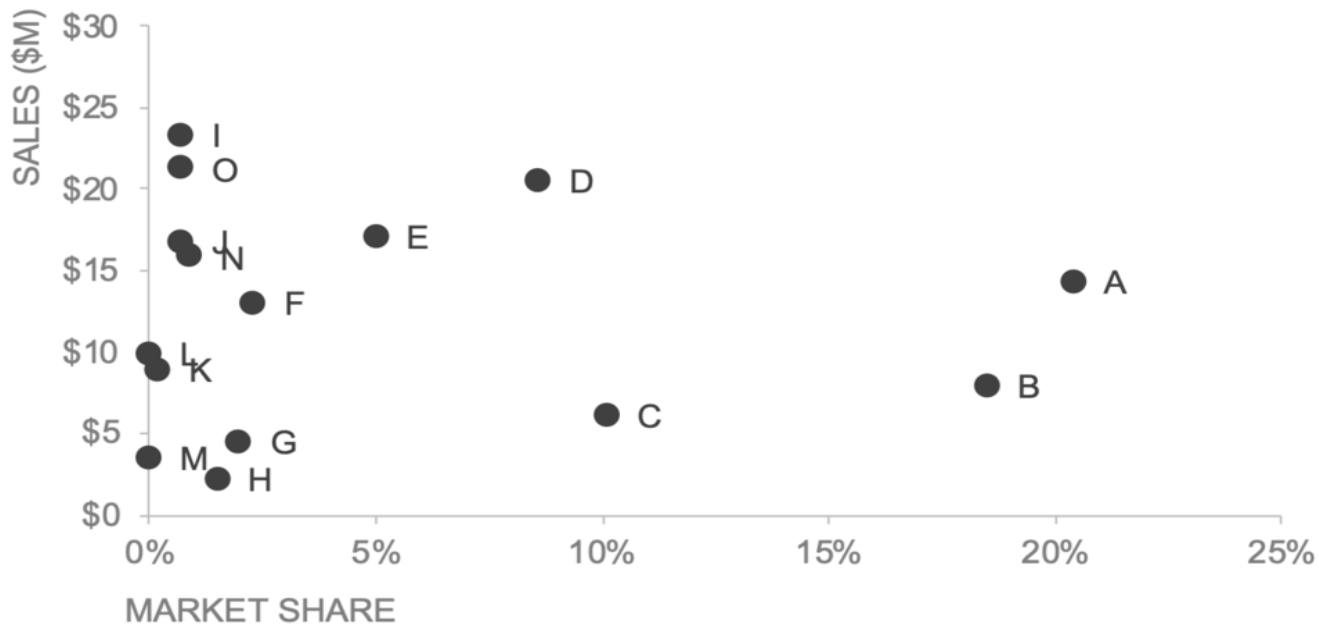
- Bubble Chart:

✓Variation of simple scatterplot with the addition of one more attribute, which is used to determine the size of the data point.

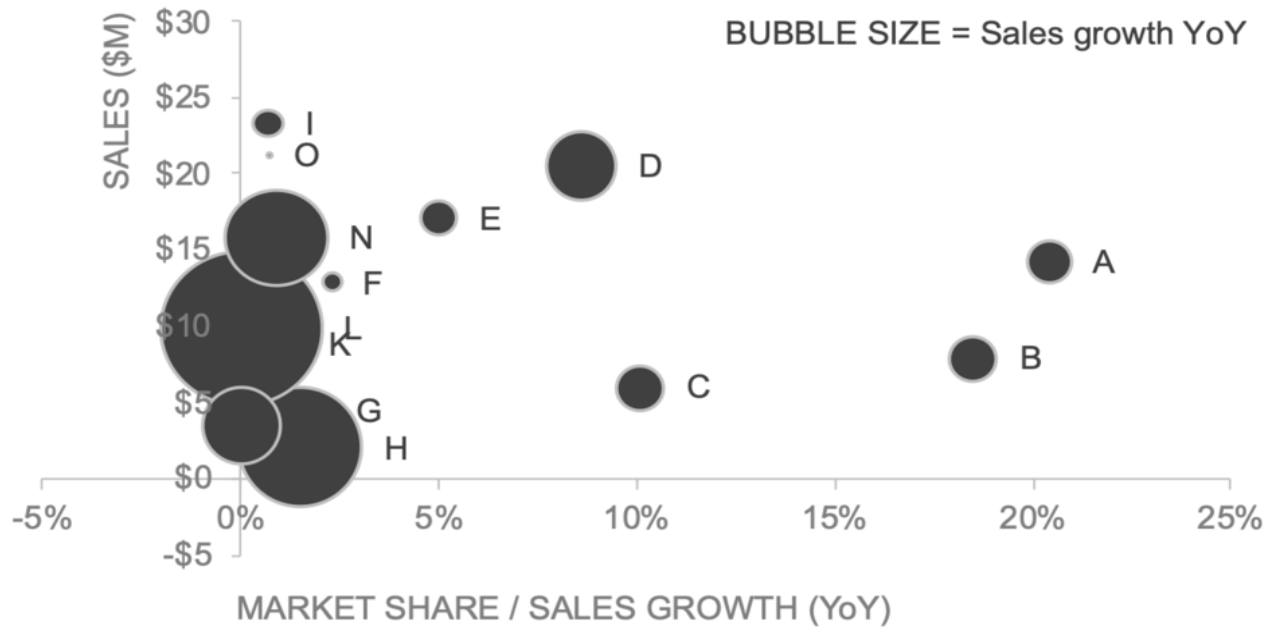
Competitive landscape

COMPETITOR	Market share	Sales ($M)	Sales growth	Region
A	20.4%	$14.2	9.0%	NA
B	18.5%	$7.9	10.0%	NA
C	10.1%	$6.0	9.8%	NA
D	8.6%	$20.5	15.0%	EMEA
E	5.0%	$17.1	7.6%	EMEA
F	2.3%	$12.9	3.3%	NA
G	2.0%	$4.4	5.6%	EMEA
H	1.5%	$2.1	26.8%	APAC
I	0.7%	$23.3	5.9%	EMEA
J	0.7%	$16.7	-4.2%	EMEA
K	0.2%	$8.8	24.8%	APAC
L	0.0%	$9.9	35.1%	APAC
M	0.0%	$3.5	17.0%	NA
N	0.9%	$15.8	21.8%	NA
O	0.7%	$21.2	0.8%	NA

Competitive landscape



Competitive landscape



- Here, the third dimension gives a visual sense of how much the competitors differ from each other with respect to their sales change: the higher the growth, the larger the bubble.

- Density Chart:

✓It is similar to scatter plot, with one more dimensions included as a background color.

✓The data point can also colored to visualize one dimension and hence, a total of four dimension can be visualized in a density chart.