# Machine learning

# UNIT-I

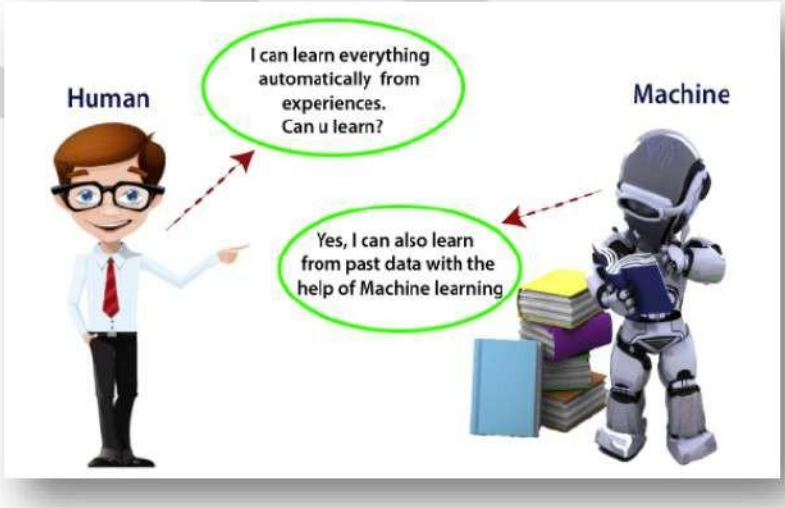
## INTRODUCTION TO MACHINE LEARNING
## &
## PREPARING TO MODEL

### Introduction:

- It has been more than 20 years since a computer program defeated the reigning world champion in a game which is considered to need a lot of intelligence to play.
- The computer program was **IBM's Deep Blue** and it defeated world chess champion, **Gary Kasparov.**
- That was the time, probably, when the most number of people gave serious attention to a fast-evolving field in computer science or more specifically artificial intelligence – i.e. **Machine learning (ML).**

### Definition:

In the real world, we are surrounded by humans who can learn everything from their experiences with their learning capability, and we have computers or machines which work on our instructions. But can a machine also learn from experiences or past data like a human does? So here comes the role of **Machine Learning**.



Machine Learning is said as a subset of **artificial intelligence** that is mainly concerned with the development of algorithms which allow a computer to learn from the data and past experiences on their own.

- The term machine learning was first introduced by **Arthur Samuel** in **1959**. We can define it in a summarized way as:

**"Machine learning enables a machine to automatically learn from data, improve performance from experiences, and predict things without being explicitly programmed".**

## Evolution of Machine learning:

- The foundation of machine learning started in the 18th and 19th centuries. The first related work dates back to 1763. In that year, Thomas Bayes's work '**An Essay towards solving a Problem in the Doctrine of Chances'** was published two years after his death. This is the work underlying Bayes Theorem, a fundamental work on which a number of algorithms of machine learning is based upon.
- In 1812, the **Bayes theorem** was actually formalized by the French mathematician **Pierre-Simon Laplace**. The method of least squares, which is the foundational concept to solve regression problems, was formalized in 1805.
- In 1913, Andrey Markov came up with the concept of **Markov chains.**
- However, the real start of focused work in the field of machine learning is considered to be Alan Turing's seminal work in 1950.

  Turing posed the question **'Can machines think?'** or in other words, **'Do machines have intelligence?'** He was the first to propose that machines can 'learn' and become artificially intelligent.
- In 1952, Arthur Samuel of IBM laboratory started working on machine learning programs, and **first developed programs that could play Checkers**.
- In 1957, Frank Rosenblatt designed the **first neural network program** simulating the human brain. From then on, for the next 50 years, the journey of machine learning has been fascinating. A number of machine learning algorithms were formulated by different researchers,

Ex: the nearest neighbour algorithm in 1969, recurrent neural network in 1982, support vector machines and random forest algorithms in 1995.

- The latest feather in the cap of machine learning development has been **Google's Alpha Go program**, which has beaten professional human Go player using machine learning techniques.

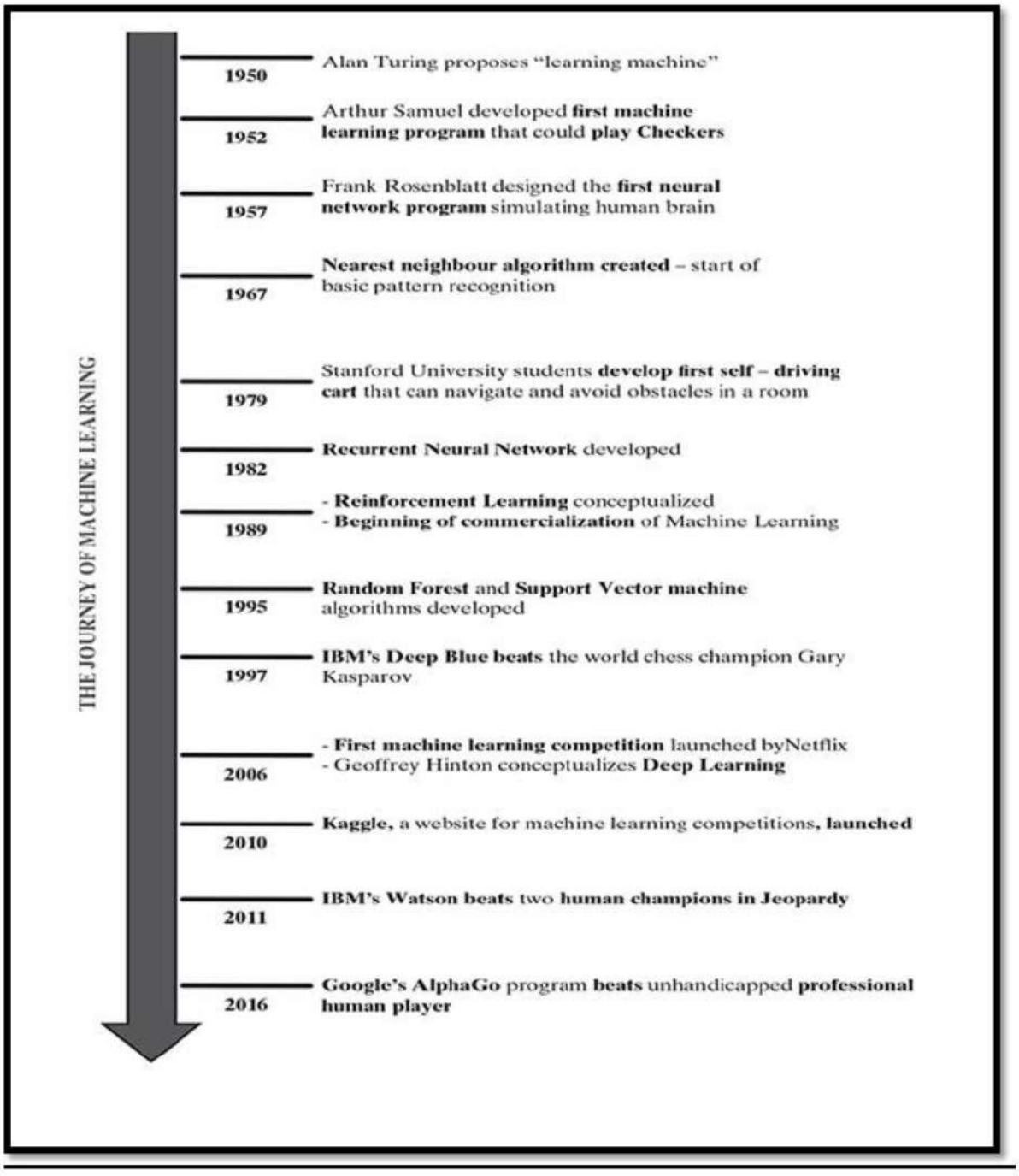



**Fig: Evolution of Machine Learning**

## WHAT IS HUMAN LEARNING?

- In cognitive science, learning is typically referred to as the process of gaining information through observation.
- To do a task in a proper way, we need to have prior information on one or more things related to the task.
- Also, as we keep learning more or acquiring more information, the efficiency in doing the tasks keep improving.
  - ✓ **For example**, i) with more knowledge, the ability to do homework with less number of mistakes increases.
  - ✓ ii) In the same way, information from past rocket launches helps in taking the right precautions and makes more successful rocket launch.

  **Thus, with more learning, tasks can be performed more efficiently.**

## TYPES OF HUMAN LEARNING:

Human learning happens in one of the three ways –

1. Either somebody who is an expert in the subject directly teaches us
2. We build our own notion indirectly based on what we have learnt from the expert in the past
3. We do it ourselves, may be after multiple attempts, some being unsuccessful.

The first type of learning falls under the category of learning directly under expert guidance, the second type falls under learning guided by knowledge gained from experts and the third type is learning by self or self-learning.

### 1. Learning under expert guidance:

- An infant may inculcate certain traits and characteristics, learning straight from its guardians. He calls his hand, a 'hand', because that is the information he gets from his parents.
- When the baby starts going to school. In school, he starts with basic familiarization of alphabets and digits. Then the baby learns how to form words from the alphabets and numbers from the digits. Slowly more complex learning happens in the form of sentences,

paragraphs, complex mathematics, science, etc. The baby is able to learn all these things from his teacher who already has knowledge on these areas.

- In all phases of life of a human being, there is an element of guided learning. This learning is imparted by someone, purely because of the fact that he/she has already gathered the knowledge by virtue of his/her experience in that field.

So **guided learning is the process of gaining information from a person having sufficient knowledge due to the past experience.**

### 2. Learning guided by knowledge gained from experts:

- An essential part of learning also happens with the knowledge which has been imparted by teacher or mentor at some point of time in some other form/context.
- For example, a baby can group together all objects of same colour even if his parents have not specifically taught him to do so. He is able to do so because at some point of time or other his parents have told him which colour is blue, which is red, which is green, etc
- In a professional role, a person is able to make out to which customers he should market a campaign from the knowledge about preference that was given by his boss long back.

In all these situations, there is no direct learning. It is some past information shared on some different context, which is used as a learning to make decisions.

### 3. Learning by self:

- In many situations, humans are left to learn on their own.
- A classic example is a baby learning to walk through obstacles. He bumps on to obstacles and falls down multiple times till he learns that whenever there is an obstacle, he needs to cross over it. He faces the same challenge while learning to ride a cycle as a kid or drive a car as an adult.
- Not all things are taught by others. A lot of things need to be learnt only from mistakes made in the past. We tend to form a check list on things that we should do, and things that we should not do, based on our experiences.

## WHAT IS MACHINE LEARNING?

Tom M. Mitchell, Professor of Machine Learning Department, School of Computer Science, Carnegie Mellon University has defined machine learning as…

**'A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.'**

- In the context of the learning to **play checkers**, E represents the experience of playing the game, T represents the task of playing checkers and P is the performance measure indicated by the percentage of games won by the player.
- In context of **image classification**, E represents the past data with images having labels or assigned classes (for example whether the image is of a class cat or a class dog or a class elephant etc.), T is the task of assigning class to new, unlabelled images and P is the performance measure indicated by the percentage of images correctly classified.

## HOW DO MACHINES LEARN?

The basic machine learning process can be divided into three parts.

**1. Data Input:** Past data or information is utilized as a basis for future decision-making

**2. Abstraction:** The input data is represented in a broader way through the underlying algorithm

**3. Generalization:** The abstracted representation is generalized to form a framework for making decisions.


**Fig:** Process of machine learning

## Abstraction:

During the machine learning process, knowledge is fed in the form of input data. However, the data cannot be used in the original shape and form. Abstraction helps in deriving a conceptual map based on the input data. This map, or a model as it is known in the machine learning paradigm, is summarized knowledge representation of the raw data.

The model may be in any one of the following forms

- Computational blocks like if/else rules
- Mathematical equations

- Specific data structures like trees or graphs
- Logical groupings of similar observations

The choice of the model used to solve a specific learning problem is a human task. The decision related to the choice of model is taken based on multiple aspects, some of which are listed below:

- **The type of problem to be solved:** Whether the problem is related to forecast or prediction, analysis of trend, understanding the different segments or groups of objects, etc.
- **Nature of the input data:** How exhaustive the input data is, whether the data has no values for many fields, the data types, etc.
- **Domain of the problem:** If the problem is in a business critical domain with a high rate of data input and need for immediate inference, e.g. fraud detection problem in banking domain.

## Generalization:

- The other key part is to tune up the abstracted knowledge to a form which can be used to take future decisions called generalization. This part is quite difficult to achieve. This is because the model is trained based on a finite set of data, which may possess a limited set of characteristics. But when we want to apply the model to take decision on a set of unknown data, usually termed as test data, we may encounter two problems.
  - The trained model is aligned with the training data too much, hence may not portray the actual trend.
  - The test data possess certain characteristics apparently unknown to the training data.

Hence, a precise approach of decision-making will not work. An approximate or heuristic approach, much like gut-feeling-based decision-making in human beings, has to be adopted. This approach has the risk of not making a correct decision – quite obviously because certain assumptions that are made may not be true in reality. But just like machines, same mistakes can be made by humans too when a decision is made based on intuition or gut-feeling – in a situation where exact reason-based decision-making is not possible.

**Well-posed learning problem:**

For defining a new problem, which can be solved using machine learning, a simple framework, highlighted below, can be used. This framework also helps in deciding whether the problem is a right candidate to be solved using machine learning.

The framework involves answering three questions:

**Step 1:** What is the problem? Describe the problem informally and formally and list assumptions and similar problems.

**Step 2:** Why does the problem need to be solved? List the motivation for solving the problem, the benefits that the solution will provide and how the solution will be used.

**Step 3:** How would I solve the problem? Describe how the problem would be solved manually to flush domain knowledge.

## TYPES OF MACHINE LEARNING:

Machine learning can be classified into three broad categories:

1. **Supervised learning –** Also called **predictive learning**. A machine predicts the class of unknown objects based on prior class-related information of similar objects.
2. **Unsupervised learning –** Also called **descriptive learning**. A machine finds patterns in unknown objects by grouping similar objects together.
3. **Reinforcement learning –** A machine learns to act on its own to achieve the given goals.

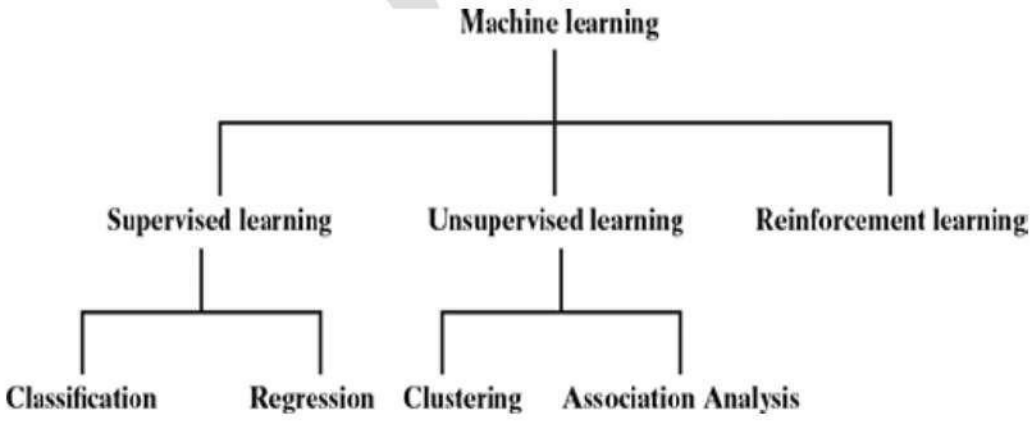



**Fig: Types of machine learning**

## 1. Supervised learning:

- The major motivation of supervised learning is to learn from past information. It is the information about the task which the machine has to execute. In context of the definition of machine learning, this past information is the experience.

**Example:** Say a machine is getting images of different objects as input and the task is to segregate the images by either shape or colour of the object. If it is by shape, the images which are of round-shaped objects need to be separated from images of triangular-shaped objects, etc. If the segregation needs to happen based on colour, images of blue objects need to be separated from images of green objects. But how can the machine know what is round shape, or triangular shape? Same way, how can the machine distinguish image of an object based on whether it is blue or green in colour?

- A machine needs the basic information to be provided to it.
- This basic input, or the experience in the paradigm of machine learning, is given in the form of **training data** .
- Training data is the past information on a specific task. In context of the image segregation problem, training data will have past data on different aspects or features on a number of images, along with a tag on whether the image is round or triangular, or blue or green in colour.
- The tag is called ' **label**' and we say that the training data is labelled in case of supervised learning

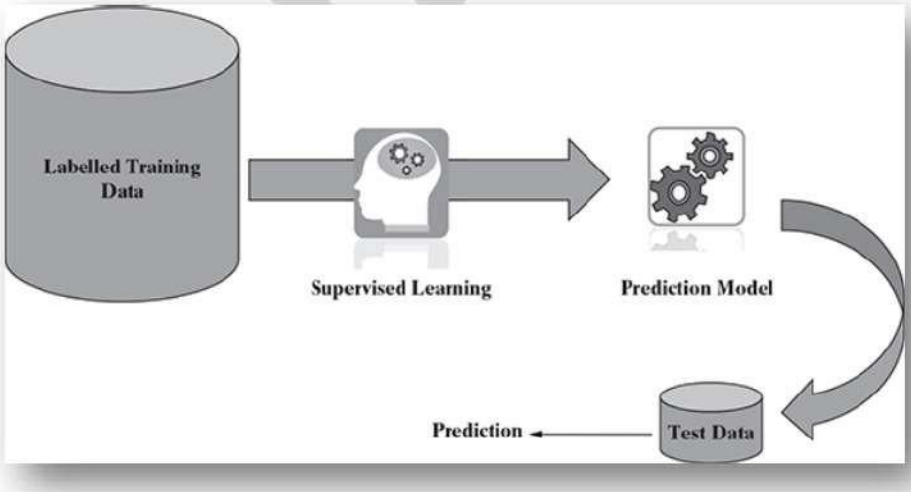



**Fig: Supervised Learning**

- Labelled training data containing past information comes as an input. Based on the training data, the machine builds a predictive model that can be used on test data to assign a label for each record in the test data.

**Examples of supervised learning are**

- Predicting the results of a game
- Predicting whether a tumour is malignant or benign
- Predicting the price of domains like real estate, stocks, etc.
- Classifying texts such as classifying a set of emails as spam or non spam.

- Supervised machine learning is as good as the data used to train it. If the training data is of poor quality, the prediction will also be far from being precise.

**The two areas of supervised learning, i.e. classification and regression.**

- When we are trying to predict a categorical or nominal variable, the problem is known as a **classification problem**. Whereas when we are trying to predict a real valued variable, the problem falls under the category of **regression.**

## i. CLASSIFICATION:

✓ Let's discuss how to segregate the images of objects based on the shape .

✓ If the image is of a round object, it is put under one category, while if the image is of a triangular object, it is put under another category.

✓ In which category the machine should put an image of unknown category, also called a test data in machine learning parlance, depends on the information it gets from the past data, which we have called as **training data.**

✓ Since the training data has a label or category defined for each and every image, the machine has to map a new image or test data to a set of images to which it is similar to and assign the same label or category to the test data.

So we observe that the whole problem revolves around assigning a label or category or class to a test data based on the label or category or class information that is imparted by the training data. Since the target objective is to assign a class label, this type of problem as classification problem.

There are number of popular machine learning algorithms which help in solving classification problems. To name a few, Naïve Bayes, Decision tree, and k-Nearest Neighbour algorithms are adopted by many machine learning practitioners.

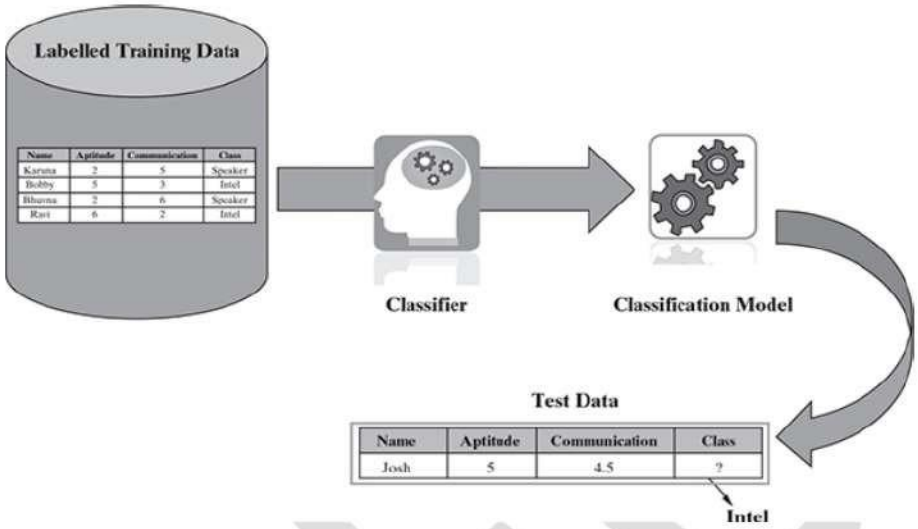



**Fig: Classification**

**Example:**

A critical classification problem in context of banking domain is **identifying potential fraudulent transactions**.

Since there are millions of transactions which have to be scrutinized and assured whether it might be a fraud transaction, it is not possible for any human being to carry out this task. Machine learning is effectively leveraged to do this task and this is a classic case of classification. Based on the past transaction data, specifically the ones labelled as fraudulent, all new incoming transactions are marked or labelled as normal or suspicious. The suspicious transactions are subsequently segregated for a closer review.

Classification is a type of supervised learning where a target feature, which is of type categorical, is predicted for test data based on the information imparted by training data. The target categorical feature is known as class.

**Some typical classification problems include:**

- ✓ Image classification
- ✓ Prediction of disease
- ✓ Win–loss prediction of games
- ✓ Prediction of natural calamity like earthquake, flood, etc.
- ✓ Recognition of handwriting

## ii. REGRESSION:

- In linear regression, the objective is to predict numerical features like real estate or stock price, temperature, marks in an examination, sales revenue, etc.
- The underlying predictor variable and the target variable are continuous in nature.
- In case of linear regression, a straight line relationship is 'fitted' between the predictor variables and the target variables, using the statistical concept of least squares method.
- As in the case of least squares method, the sum of square of error between actual and predicted values of the target variable is tried to be minimized.
- In case of simple linear regression, there is only one predictor variable whereas in case of multiple linear regression, multiple predictor variables can be included in the model.

**Example:**

Let's take the example of yearly budgeting exercise of the sales managers. They have to give sales prediction for the next year based on sales figure of previous years visà-vis investment being put in. Obviously, the data related to past as well as the data to be predicted are continuous in nature. In a basic approach, a simple linear regression model can be applied with investment as predictor variable and sales revenue as the target variable.

The following figure shows a typical simple regression model, where regression line is fitted based on values of target variable with respect to different values of predictor variable. A typical linear regression model can be represented in the form –

$$y = \alpha + \beta x$$

where 'x' is the predictor variable and 'y' is the target variable.

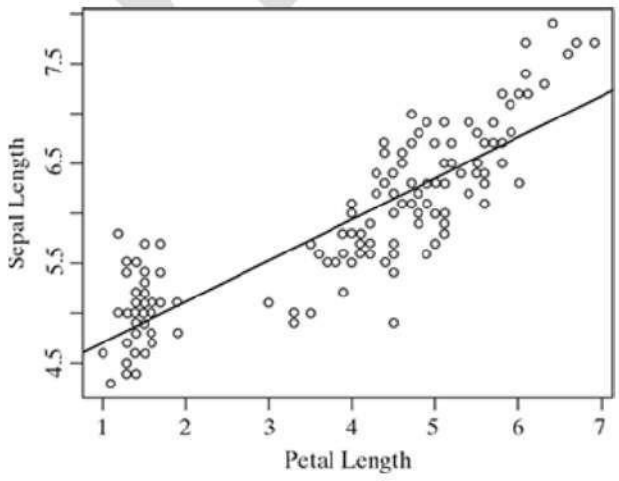



**Fig: Regression**

**Example - The input data come from a famous multivariate data set named Iris introduced by the British statistician and biologist Ronald Fisher.** The data set consists of 50 samples from each of three species of Iris – Iris setosa, Iris virginica, and Iris versicolor. Four features were measured for each sample – sepal length, sepal width, petal length, and petal width. These features can uniquely discriminate the different species of the flower.

The Iris data set is typically used as a training data for solving the classification problem of predicting the flower species based on feature values. However, we can also demonstrate regression using this data set, by predicting the value of one feature using another feature as predictor. In above figure, petal length is a predictor variable which, when fitted in the simple linear regression model, helps in predicting the value of the target variable sepal length.

**Typical applications of regression are:**

- Demand forecasting in retails
- Sales prediction for managers
- Price prediction in real estate
- Weather forecast
- Skill demand forecast in job market

## 2. UNSUPERVISED LEARNING:

- Unlike supervised learning, in unsupervised learning, there is no labelled training data to learn from and no prediction to be made.
- In unsupervised learning, the objective is to take a dataset as input and try to find natural groupings or patterns within the data elements or records.
- Therefore, unsupervised learning is often termed as descriptive model and the process of unsupervised learning is referred as pattern discovery or knowledge discovery.
- One critical application of unsupervised learning is customer segmentation.

i) **CLUSTERING** is the main type of unsupervised learning.

- It intends to group or organize similar objects together. For that reason, objects belonging to the same cluster are quite similar to each other while objects belonging to different clusters are quite dissimilar.
- Hence, the objective of clustering to discover the intrinsic grouping of unlabelled data and form clusters, as depicted in Figure.

Different measures of similarity can be applied for clustering. One of the most commonly adopted similarity measure is distance. Two data items are considered as a part of the same cluster if the distance between them is less. In the same way, if the distance between the data items is high, the items do not generally belong to the same cluster.

This is also known as **distance-based clustering**. The following Figure depicts the process of clustering at a high level.

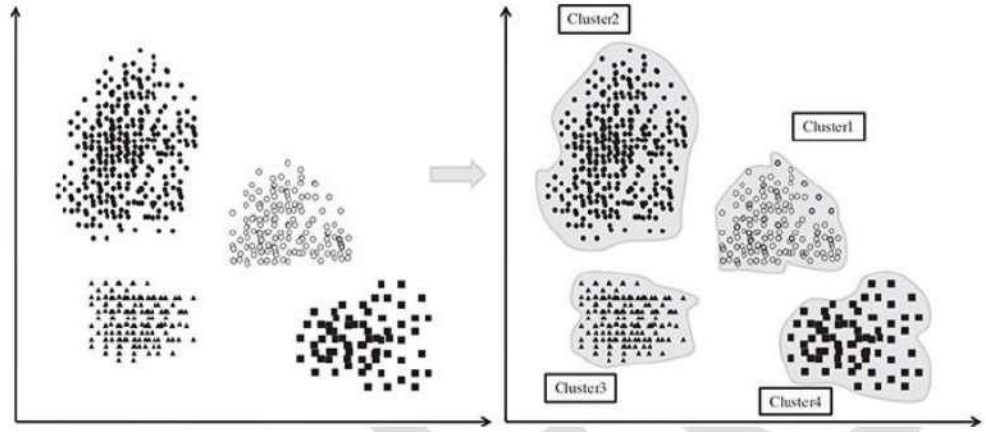



**Fig: Distance-based clustering**

**ii)** Other than clustering of data and getting a summarized view from it, one more variant of unsupervised learning is **ASSOCIATION** analysis.

- As a part of association analysis, the association between data elements is identified.
- Let's try to understand the approach of association analysis in context of one of the most common examples, i.e. market basket analysis as shown in below Figure.
- From past transaction data in a grocery store, it may be observed that most of the customers who have bought item A, have also bought item B and item C or at least one of them.
- This means that there is a strong association of the event 'purchase of item A' with the event 'purchase of item B', or 'purchase of item C'. Identifying these sorts of associations is the goal of association analysis.
- This helps in boosting up sales pipeline, hence a critical input for the sales group. Critical applications of association analysis include market basket analysis and recommender systems.

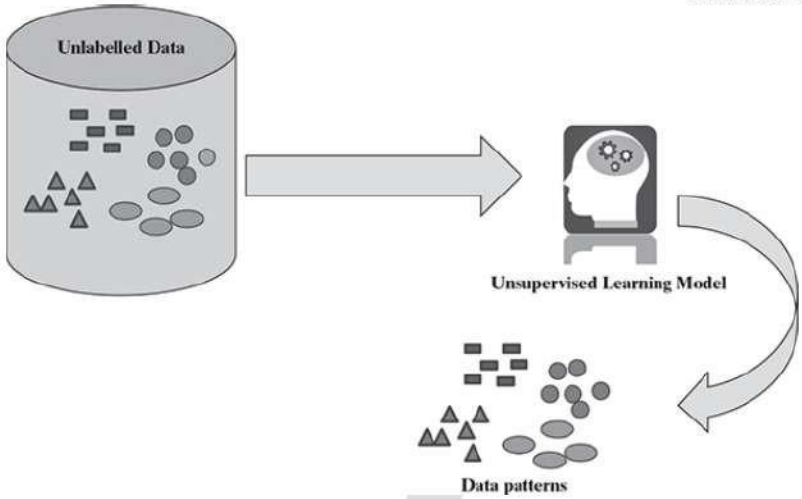



**Fig: Unsupervised Learning**

| TransID | Items Bought |
|---|---|
| 1 | {Butter, Bread} |
| 2 | {Diaper, Bread, Milk, Beer} |
| 3 | {Milk, Chicken, Beer, Diaper} |
| 4 | {Bread, Diaper, Chicken, Beer} |
| 5 | {Diaper, Beer, Cookies, Ice cream} |
| ... | ... |

Market Basket transactions
Frequent itemsets → (Diaper, Beer)
Possible association: Diaper → Beer

**Fig: Market based Analysis**

## 3. REINFORCEMENT LEARNING:

**Example:** While Babies walking, sometimes they fall down hitting an obstacle, whereas other times they are able to walk smoothly avoiding bumpy obstacles. When they are able to walk overcoming the obstacle, their parents are elated and appreciate the baby with loud claps / or may be a chocolates. When they fall down while circumventing an obstacle, obviously their parents do not give claps or chocolates. Slowly a time comes when the babies learn from mistakes and are able to walk with much ease.

- In the same way, machines often learn to do tasks autonomously.
- Let's try to understand in context of the example of the child learning to walk. The action tried to be achieved is walking, the child is the agent and the place with hurdles on which the child is trying to walk resembles the environment.
- It tries to improve its performance of doing the task.
- When a sub-task is accomplished successfully, a reward is given.

- When a sub-task is not executed correctly, obviously no reward is given. This continues till the machine is able to complete execution of the whole task.

  This process of learning is known as reinforcement learning.

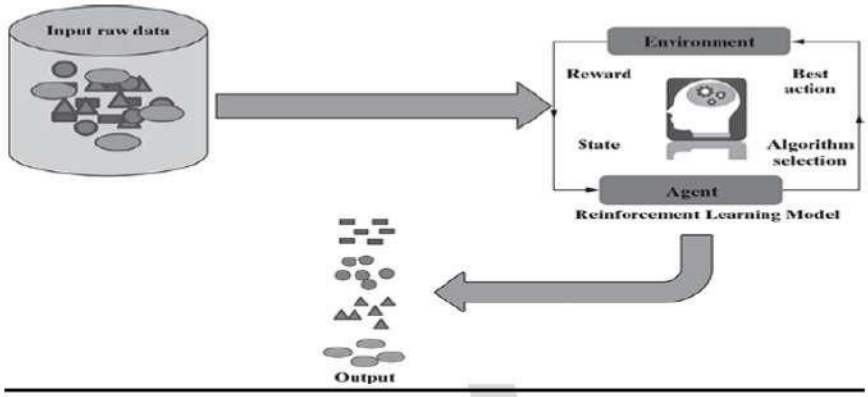



**Fig: Reinforcement Learning**

- One contemporary example of reinforcement learning is self-driving cars. The critical information which it needs to take care of speed and speed limit in different road segments, traffic conditions, road conditions, weather conditions, etc. The tasks that have to be taken care of are start/stop, accelerate/decelerate, turn to left / right, etc.

| SUPERVISED | UNSUPERVISED | REINFORCEMENT |
|---|---|---|
| This type of learning is used when you know how to classify a given data, or in other words classes or labels are available. | This type of learning is used when there is no idea about the class or label of a particular data. The model has to find pattern in the data. | This type of learning is used when there is no idea about the class or label of a particular data. The model has to do the classification – it will get rewarded if the classification is correct, else get punished. |
| Labelled training data is needed. Model is built based on training data. | Any unknown and unlabelled data set is given to the model as input and records are grouped. | The model learns and updates itself through reward/ punishment. |
| The model performance can be evaluated based on how many misclassifications have been done based on a comparison between predicted and actual values. | Difficult to measure whether the model did something useful or interesting. Homogeneity of records grouped together is the only measure. | Model is evaluated by means of the reward function after it had some time to learn. |
| There are two types of supervised learning problems – classification and regression. | There are two types of unsupervised learning problems – clustering and association. | No such types. |
| Simplest one to understand. | More difficult to understand and implement than supervised learning. | Most complex to understand and apply. |
| Standard algorithms include<br>• Naïve Bayes<br>• *k*-nearest neighbour (kNN)<br>• Decision tree<br>• Linear regression<br>• Logistic regression<br>• Support Vector Machine SVM), etc. | Standard algorithms are<br>• *k*-means<br>• Principal Component Analysis (PCA)<br>• Self-organizing map (SOM)<br>• Apriori algorithm<br>• DBSCAN etc. | Standard algorithms are<br>• Q-learning<br>• Sarsa |
| Practical applications include<br>• Handwriting recognition<br>• Stock market prediction<br>• Disease prediction<br>• Fraud detection, etc. | Practical applications include<br>• Market basket analysis<br>• Recommender systems<br>• Customer segmentation, etc. | Practical applications include<br>• Self-driving cars<br>• Intelligent robots<br>• AlphaGo Zero (the latest version of DeepMind's AI system playing Go) |

**Fig: Comparison – supervised, unsupervised, and reinforcement learning**

## PROBLES NOT TO BE SOLVED USING MACHINE LEARNING:

1. Machine learning should not be applied to tasks in which humans are very effective or frequent human intervention is needed.

   For example, air traffic control is a very complex task needing intense human involvement. At the same time, for very simple tasks which can be implemented using traditional programming paradigms, there is no sense of using machine learning. For example, simple rule-driven or formula-based applications like price calculator engine, dispute tracking application, etc. do not need machine learning techniques.
2. Machine learning should be used only when the business process has some lapses. If the task is already optimized, incorporating machine learning will not serve to justify the return on investment.
3. Machine learning cannot be used effectively For situations where training data is not sufficient. This is because, with small training data sets, the impact of bad data is exponentially worse. For the quality of prediction or recommendation to be good, the training data should be sizeable.

## APPLICATIONS OF MACHINE LEARNING:

Wherever there is a substantial amount of past data, machine learning can be used to generate actionable insight from the data. Though machine learning is adopted in multiple forms in every business domain, we have covered below three major domains just to give some idea about what type of actions can be done using machine learning.

i) **Banking and finance:**

In the banking industry, fraudulent transactions, especially the ones related to credit cards, are extremely prevalent. Since the volumes as well as velocity of the transactions are extremely high, high performance machine learning solutions are implemented by almost all leading banks across the globe. The models work on a real-time basis, i.e. the fraudulent transactions are spotted and prevented right at the time of occurrence. This helps in avoiding a lot of operational hassles in settling the disputes that customers will otherwise raise against those fraudulent transactions. Customers of a bank are often offered lucrative proposals by other competitor banks. Proposals like higher bank interest, lower processing charge of loans, zero balance savings accounts, no overdraft penalty, etc. are offered to customers, with the intent that the customer switches over to the competitor bank. Also,

sometimes customers get demotivated by the poor quality of services of the banks and shift to competitor banks. Machine learning helps in preventing or at least reducing the customer churn. Both descriptive and predictive learning can be applied for reducing customer churn. Using descriptive learning, the specific pockets of problem, i.e. a specific bank or a specific zone or a specific type of offering like car loan, may be spotted where maximum churn is happening. Quite obviously, these are troubled areas where further investigation needs to be done to find and fix the root cause. Using predictive learning, the set of vulnerable customers who may leave the bank very soon, can be identified. Proper action can be taken to make sure that the customers stay back.

ii) **Insurance:**

Insurance industry is extremely data intensive. For that reason, machine learning is extensively used in the insurance industry. Two major areas in the insurance industry where machine learning is used are risk prediction during new customer onboarding and claims management. During customer onboarding, based on the past information the risk profile of a new customer needs to be predicted. Based on the quantum of risk predicted, the quote is generated for the prospective customer. When a customer claim comes for settlement, past information related to historic claims along with the adjustor notes are considered to predict whether there is any possibility of the claim to be fraudulent. Other than the past information related to the specific customer, information related to similar customers, i.e. customer belonging to the same geographical location, age group, ethnic group, etc., are also considered to formulate the model.

iii) **Healthcare:**

Wearable device data form a rich source for applying machine learning and predict the health conditions of the person real time. In case there is some health issue which is predicted by the learning model, immediately the person is alerted to take preventive action. In case of some extreme problem, doctors or healthcare providers in the vicinity of the person can be alerted. Suppose an elderly person goes for a morning walk in a park close to his house. Suddenly, while walking, his blood pressure shoots up beyond a certain limit, which is tracked by the wearable. The wearable data is sent to a remote server and a machine learning algorithm is constantly analyzing the streaming data. It also has the history of the elderly person and persons of similar age group. The model predicts some

fatality unless immediate action is taken. Alert can be sent to the person to immediately stop walking and take rest. Also, doctors and healthcare providers can be alerted to be on standby.

Machine learning along with computer vision also plays a crucial role in disease diagnosis from medical imaging.

## STATE-OF-THE-ART LANGUAGES/TOOLS IN MACHINE LEARNING:

- The algorithms related to different machine learning tasks are known to all and can be implemented using any language/platform.
- It can be implemented using a Java platform or C / C++ language or in .NET. However, there are certain languages and tools which have been developed with a focus for implementing machine learning. Few of them, which are most widely used, are covered below.

1. **Python:**
   - ✓ Python is one of the most popular, open source programming language widely adopted by machine learning community.
   - ✓ It was designed by Guido van Rossum and was first released in 1991. The reference implementation of Python, i.e.CPython, is managed by Python Software Foundation, which is a non-profit organization.
   - ✓ Python has very strong libraries for advanced mathematical functionalities (NumPy), algorithms and mathematical tools (SciPy) and numerical plotting (matplotlib).
   - ✓ Built on these libraries, there is a machine learning library named scikit-learn, which has various classification, regression, and clustering algorithms embedded in it.
2. R:

R is a language for statistical computing and data analysis. It is an open source language, extremely popular in the academic community – especially among statisticians and data miners. R is considered as a variant of S, a GNU project which was developed at Bell Laboratories. Currently, it is supported by the R Foundation for statistical computing.

R is a very simple programming language with a huge set of libraries available for different stages of machine learning. Some of the libraries standing out in terms of popularity are plyr/dplyr (for data transformation), caret ('Classification and Regression Training' for classification), RJava (to facilitate integration with Java), tm (for text mining), ggplot2 (for data

visualization). Other than the libraries, certain packages like Shiny and R Markdown have been developed around R to develop interactive web applications, documents and dashboards on R without much effort.

### 3. Matlab:

MATLAB (matrix laboratory) is a licenced commercial software with a robust support for a wide range of numerical computing. MATLAB has a huge user base across industry and academia. MATLAB is developed by MathWorks, a company founded in 1984. Being proprietary software, MATLAB is developed much more professionally, tested rigorously, and has comprehensive documentation.

MATLAB also provides extensive support of statistical functions and has a huge number of machine learning algorithms in-built. It also has the ability to scale up for large datasets by parallel processing on clusters and cloud.

### 4. SAS:

SAS (earlier known as 'Statistical Analysis System') is another licenced commercial software which provides strong support for machine learning functionalities. Developed in C by SAS Institute, SAS had its first release in the year 1976.

SAS is a software suite comprising different components. The basic data management functionalities are embedded in the Base SAS component whereas the other components like SAS/INSIGHT, Enterprise Miner, SAS/STAT, etc. help in specialized functions related to data mining and statistical analysis.

### Other languages/tools:

There are a host of other languages and tools that also support machine learning functionalities. Owned by IBM, SPSS (originally named as Statistical Package for the Social Sciences) is a popular package supporting specialized data mining and statistical analysis. Originally popular for statistical analysis in social science (as the name reflects), SPSS is now popular in other fields as well.

Released in 2012, Julia is an open source, liberal licence programming language for numerical analysis and computational science. It has baked in all good things of MATLAB, Python, R, and

other programming languages used for machine learning for which it is gaining steady attention from machine learning development community. Another big point in favour of Julia is its ability to implement high-performance machine learning algorithms.

## ISSUES IN MACHINE LEARNING:

Machine learning is a field which is relatively new and still evolving. Also, the level of research and kind of use of machine learning tools and technologies varies drastically from country to country. The laws and regulations, cultural background, emotional maturity of people differ drastically in different countries. All these factors make the use of machine learning and the issues originating out of machine learning usage are quite different.

**The biggest fear and issue arising out of machine learning is related to privacy and the breach of it.** The primary focus of learning is on analyzing data, both past and current, and coming up with insight from the data. This insight may be related to people and the facts revealed might be private enough to be kept confidential. Also, different people have a different preference when it comes to sharing of information. While some people may be open to sharing some level of information publicly, some other people may not want to share it even to all friends and keep it restricted just to family members. Classic examples are a birth date (not the day, but the date as a whole), photographs of a dinner date with family, educational background, etc. Some people share them with all in the social platforms like Facebook while others do not, or if they do, they may restrict it to friends only. When machine learning algorithms are implemented using those information, inadvertently people may get upset. For example, if there is a learning algorithm to do preference-based customer segmentation and the output of the analysis is used for sending targeted marketing campaigns, it will hurt the emotion of people and actually do more harm than good. In certain countries, such events may result in legal actions to be taken by the people affected. Even if there is no breach of privacy, there may be situations where actions were taken based on machine learning may create an adverse reaction. Let's take the example of knowledge discovery exercise done before starting an election campaign. If a specific area reveals an ethnic majority or skewness of a certain demographic factor, and the campaign pitch carries a message keeping that in mind, it might actually upset the voters and cause an adverse result. So a very critical consideration before applying machine learning is that proper human judgement should be exercised before using any outcome from machine learning. Only then the decision taken will be beneficial and also not result in any adverse impact.

# PREPARING TO MODEL

This chapter gives a detailed view of how to understand the incoming data and create basic understanding about the nature and quality of the data. This information, in turn, helps to select and then how to apply the model. So, this chapter helps a beginner take the first step towards effective modeling and solving a machine learning problem.

## INTRODUCTION:

- It all started as a proposition from the renowned computer scientist Alan Turing – machines can 'learn' and become artificially intelligent.
- Gradually, through the next few decades path-breaking innovations came in from Arthur Samuel, Frank Rosenblatt, John Hopfield, Christopher Watkins, Geoffrey Hintonand many other computer scientists.
- They shaped up concepts of Neural Networks, Recurrent Neural Network, Reinforcement Learning, Deep Learning, etc. which took machine learning to new heights.

While development in machine learning technology has been extensive and its implementation has become widespread,, we need to gain some basic understanding. We need to understand how to apply the array of tools and technologies available in the machine learning to solve a problem. In fact, that is going to be very specific to the kind of problem that we are trying to solve.

## MACHINE LEARNING ACTIVITIES:

- The first step in machine learning activity starts with data.
- In case of supervised learning, it is the labelled training data set followed by test data which is not labelled.
- In case of unsupervised learning, there is no question of labelled data but the task is to find patterns in the input data.
- A thorough review and exploration of the data is needed to understand the type of the data, the quality of the data and relationship between the different data elements. Based on that, multiple pre-processing activities may need to be done on the input data before we can go ahead with core machine learning activities.

Following are the typical preparation activities done once the input data comes into the machine learning system:

- Understand the type of data in the given input data set.
- Explore the data to understand the nature and quality.
- Explore the relationships amongst the data elements, e.g. inter-feature relationship.
- Find potential issues in data.
- Do the necessary remediation, e.g. impute missing data values, etc., if needed.
- Apply pre-processing steps, as necessary.
- Once the data is prepared for modelling, then the learning tasks start off. As a part of it, do the following activities:
  - The input data is first divided into parts – the training data and the test data (called holdout). This step is applicable for supervised learning only.
  - Consider different models or learning algorithms for selection.
  - Train the model based on the training data for supervised learning problem and apply to unknown data. Directly apply the chosen unsupervised model on the input data for unsupervised learning problem.
  - After the model is selected, trained (for supervised learning), and applied on input data, the performance of the model is evaluated. Based on options available, specific actions can be taken to improve the performance of the model, if possible.

The Following figure depicts the four-step process of machine learning.

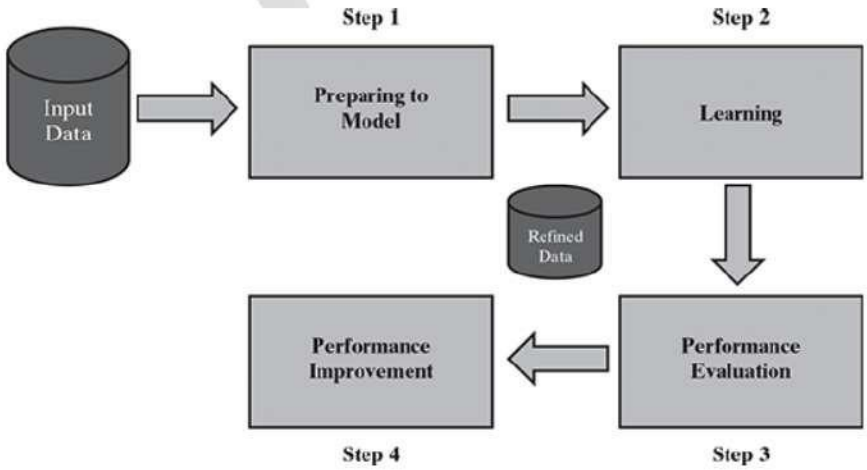



**Fig: Detailed process of machine learning**

The following Table contains a summary of steps and activities involved:

| Step # | Step Name | Activities Involved |
|---|---|---|
| Step 1 | Preparing to Model | • Understand the type of data in the given input data set<br>• Explore the data to understand data quality<br>• Explore the relationships amongst the data elements, e.g. inter-feature relationship<br>• Find potential issues in data<br>• Remediate data, if needed<br>• Apply following pre-processing steps, as necessary:<br>✓ Dimensionality reduction<br>✓ Feature subset selection |
| Step 2 | Learning | • Data partitioning/holdout<br>• Model selection<br>• Cross-validation |
| Step 3 | Performance evaluation | • Examine the model performance, e.g. confusion matrix in case of classification<br>• Visualize performance trade-offs using ROC curves |
| Step 4 | Performance improvement | • Tuning the model<br>• Ensembling<br>• Bagging<br>• Boosting |

## BASIC TYPES OF DATA IN MACHINE LEARNING:

A data set is a collection of related information or records. The information may be on some entity or some subject area.

**Example:** we may have a data set on students in which each record consists of information about a specific student. Again, we can have a data set on student performance which has records providing performance, i.e. marks on the individual subjects.

- Each row of a data set is called a record.
- Each data set also has multiple attributes, each of which gives information on a specific characteristic.

For example, in the data set on **students**, there are four attributes namely **Roll Number, Name, Gender, and Age**, each of which understandably is a specific characteristic about the student entity. Attributes can also be termed as feature, variable, dimension or field.

- Both the data sets, Student and Student Performance, are having four features or dimensions; hence they are told to have four-dimensional data space.
- A row or record represents a point in the four-dimensional data space as each row has specific values for each of the four attributes or features.

- Value of an attribute, quite understandably, may vary from record to record.

For example, if we refer to the first two records in the Student data set, the value of attributes Name, Gender, and Age are different.

**Student data set:**

| Roll Number | Name | Gender | Age |
|---|---|---|---|
| 129/011 | Mihir Karmarkar | M | 14 |
| 129/012 | Geeta Iyer | F | 15 |
| 129/013 | Chanda Bose | F | 14 |
| 129/014 | Sreenu Subramanian | M | 14 |
| 129/015 | Pallav Gupta | M | 16 |
| 129/016 | Gajanan Sharma | M | 15 |

**Student performance data set:**

| Roll Number | Maths | Science | Percentage |
|---|---|---|---|
| 129/011 | 89 | 45 | 89.33% |
| 129/012 | 89 | 47 | 90.67% |
| 129/013 | 68 | 29 | 64.67% |
| 129/014 | 83 | 38 | 80.67% |
| 129/015 | 57 | 23 | 53.33% |
| 129/016 | 78 | 35 | 75.33% |

Fig: Examples of data set

| Roll Number | Name | Gender | Age |
|---|---|---|---|
| 129/011 | Mihir Karmarkar | M | 14 |
| 129/012 | Geeta Iyer | F | 15 |

Fig: Data set records and attributes

Data can broadly be divided into following two types:

1. Qualitative data
2. Quantitative data

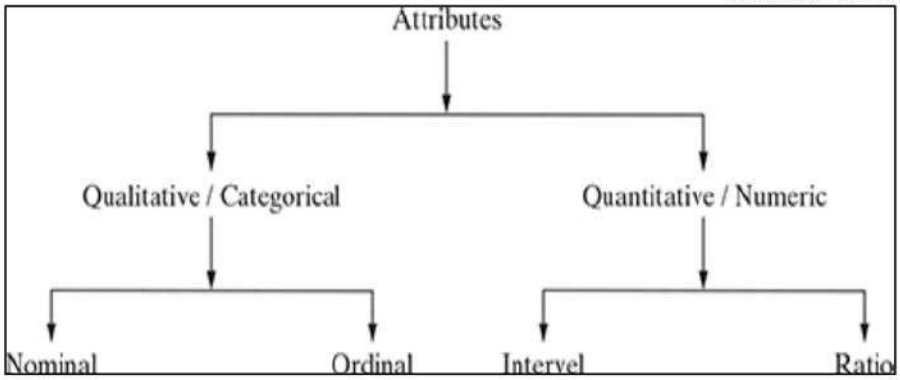



**Fig: Types of data**

## 1. Qualitative data :

- Qualitative data provides information about the quality of an object or information which cannot be measured.

**For example**:

- if we consider the quality of performance of students in terms of 'Good', 'Average', and 'Poor', it falls under the category of qualitative data.
- Also, name or roll number of students are information that cannot be measured using some scale of measurement. So they would fall under qualitative data.

Qualitative data is also called categorical data. Qualitative data can be further subdivided into two types as follows:

(i). Nominal data

(ii). Ordinal data

**(i). Nominal data:**

- Nominal data is one which has no numeric value, but a named value.
- It is used for assigning named values to attributes.
- Nominal values cannot be quantified.

**Examples:**

1. Blood group: A, B, O, AB, etc.

2. Nationality: Indian, American, British, etc.

3. Gender: Male, Female, Other

Mathematical operations such as addition, subtraction, multiplication, etc. cannot be performed on nominal data. For that reason, statistical functions such as mean, variance, etc. can also not be applied on nominal data. However, a basic count is possible.

So mode, i.e. most frequently occurring value, can be identified for nominal data.

**(ii). Ordinal data:**

- Ordinal data, in addition to possessing the properties of nominal data, can also be naturally ordered.
- This means ordinal data also assigns named values to attributes but unlike nominal data, they can be arranged in a sequence of increasing or decreasing value so that we can say whether a value is better than or greater than another value.

**Examples:**

1. Customer satisfaction: 'Very Happy', 'Happy', 'Unhappy', etc.
2. Grades: A, B, C, etc.
3. Hardness of Metal: 'Very Hard', 'Hard', 'Soft', etc

Like nominal data, basic counting is possible for ordinal data. Hence, the mode can be identified. Since ordering is possible in case of ordinal data, median, and quartiles can be identified in addition. Mean can still not be calculated.

**2. Quantitative data:**

Quantitative data relates to information about the quantity of an object – hence it can be measured. For example, if we consider the attribute 'marks', it can be measured using a scale of measurement. Quantitative data is also termed as numeric data.

There are two types of quantitative data:

(i) Interval data

(ii) Ratio data

**(i) Interval data :**

- Interval data is numeric data for which not only the order is known, but the exact difference between values is also known.
- An ideal example of interval data is Celsius temperature. The difference between each value remains the same in Celsius temperature.

For example, the difference between 12°C and 18°C degrees is measurable and is 6°C as in the case of difference between 15.5°C and 21.5°C.

- Other examples include date, time, etc. For interval data, mathematical operations such as addition and subtraction are possible. For that reason, for interval data, the central tendency can be measured by mean, median, or mode. Standard deviation can also be calculated.
- However, interval data do not have something called a 'true zero' value. For example, there is nothing called '0 temperature' or 'no temperature'. Hence, only addition and subtraction applies for interval data. The ratio cannot be applied. This means, we can say a temperature of 40°C is equal to the temperature of 20°C + temperature of 20°C. However, we cannot say the temperature of 40°C means it is twice as hot as in temperature of 20°C.

**(ii) Ratio data:**

Ratio data represents numeric data for which exact value can be measured. Absolute zero is available for ratio data. Also, these variables can be added, subtracted, multiplied, or divided. The central tendency can be measured by mean, median, or mode and methods of dispersion such as standard deviation. Examples of ratio data include height, weight, age, salary, etc.

Attributes can also be categorized into types based on a number of values that can be assigned. The attributes can be either discrete or continuous based on this factor.

**Discrete attributes** can assume a finite or countably infinite number of values.

- Nominal attributes such as roll number, street number, pin code, etc. can have a finite number of values.
- Numeric attributes such as count, rank of students, etc. can have countably infinite values.
- Binary attribute A special type of discrete attribute which can assume two values only.
- Examples of binary attribute include male/ female, positive/negative, yes/no, etc.

**Continuous attributes** can assume any possible value which is a real number. Examples of continuous attribute include length, height, weight, price, etc.

## EXPLORING STRUCTURE OF DATA:

The approach of exploring numeric data is different than the approach of exploring categorical data. In case of a standard data set, we may have the data dictionary available for reference.

**Data dictionary** is a metadata repository, **i.e. the repository of all information related to the structure of each data element contained in the data set**. The data dictionary gives detailed information on each of the attributes – the description as well as the data type and other relevant details. In case the data dictionary is not available, we need to use standard library function of the machine learning tool that we are using and get the details.

## Example:

The data set that we take as a reference is the Auto MPG data set available in the UCI repository.

| mpg | cylinder | displace-ment | horse-power | weight | accel-eration | model year | origin | car name |
|---|---|---|---|---|---|---|---|---|
| 18 | 8 | 307 | 130 | 3504 | 12 | 70 | 1 | Chevrolet chev-elle malibu |
| 15 | 8 | 350 | 165 | 3693 | 11.5 | 70 | 1 | Buick skylark 320 |
| 18 | 8 | 318 | 150 | 3436 | 11 | 70 | 1 | Plymouth satellite |
| 16 | 8 | 304 | 150 | 3433 | 12 | 70 | 1 | Amc rebel sst |
| 17 | 8 | 302 | 140 | 3449 | 10.5 | 70 | 1 | Ford torino |
| 15 | 8 | 429 | 198 | 4341 | 10 | 70 | 1 | Ford galaxie 500 |
| 14 | 8 | 454 | 220 | 4354 | 9 | 70 | 1 | Chevrolet impala |
| 14 | 8 | 440 | 215 | 4312 | 8.5 | 70 | 1 | Plymouth fury iii |
| 14 | 8 | 455 | 225 | 4425 | 10 | 70 | 1 | Pontiac catalina |
| 15 | 8 | 390 | 190 | 3850 | 8.5 | 70 | 1 | Amc acbassador dpl |
| 15 | 8 | 383 | 170 | 3563 | 10 | 70 | 1 | Dodge challenger se |
| 14 | 8 | 340 | 160 | 3609 | 8 | 70 | 1 | Plymouth ' cuda 340 |
| 15 | 8 | 400 | 150 | 3761 | 9.5 | 70 | 1 | Chevrolet monte carlo |
| 14 | 8 | 455 | 225 | 3086 | 10 | 70 | 1 | Buick estate wagon (sw) |
| 24 | 4 | 113 | 95 | 2372 | 15 | 70 | 3 | Toyota corona mark ii |
| 22 | 6 | 198 | 95 | 2933 | 15.5 | 70 | 1 | Plymouth duster |
| 18 | 6 | 199 | 97 | 2774 | 15.5 | 70 | 1 | Amc hornet |

**FIG: Auto MPG data set**

- As is quite evident from the data, the attributes such as 'mpg', 'cylinders', 'displacement', 'horsepower', 'weight', 'acceleration', 'model year', and 'origin' are all numeric.
- Out of these attributes, 'cylinders', 'model year', and 'origin' are discrete in nature as the only finite number of values can be assumed by these attributes.

- The remaining of the numeric attributes, i.e. 'mpg', 'displacement', 'horsepower', 'weight', and 'acceleration' can assume any real value.

  Hence, these attributes are continuous in nature.

- The only remaining attribute 'car name' is of type categorical, or more specifically nominal.

  This data set is regarding prediction of fuel consumption in miles per gallon, i.e. the numeric attribute 'mpg' is the target attribute.

  Here we will discuss

  i. **Exploring numerical data**
     - **Plotting and Exploring Numerical Data**
  ii. **Exploring Categorical data**
  iii. **Exploring relationship between variables.**

## 1. Exploring numerical data:

There are two most effective mathematical plots to explore numerical data – **box plot and histogram.**

### i) Understanding central tendency:

- To understand the nature of numeric variables, we can apply the measures of central tendency of data, i.e. **mean and median.**
- In statistics, measures of central tendency help us understand the central point of a set of data.
- **Mean,** by definition, **is a sum of all data values divided by the count of data elements.**

**For example**, mean of a set of observations – 21, 89, 34, 67, and 96 is calculated as below.

$$\text{Mean} = \frac{21 + 89 + 34 + 67 + 96}{5} = 61.4$$

If the above set of numbers represents marks of 5 students in a class, the mean marks, or the falling in the middle of the range is 61.4.

- **Median, on contrary, is the value of the element appearing in the middle of an ordered list of data elements.**
- If we consider the above 5 data elements, the ordered list would be – 21, 34, 67, 89, and 96. Since there are 5 data elements, the 3rd element in the ordered list is considered as the median. Hence, the median value of this set of data is 67.

So, in the context of the Auto MPG data set, let's try to find out for each of the numeric attributes the values of mean and median. We can also find out if the deviation between these values is large. In below figure the comparison between mean and median for all the attributes has been shown. We can see that for the attributes such as 'mpg', 'weight', 'acceleration', and 'model.year' the deviation between mean and median is not significant which means the chance of these attributes having too many outlier values is less. However, the deviation is significant for the attributes 'cylinders', 'displacement' and 'origin'.

So, we need to further drill down and look at some more statistics for these attributes. Also, there is some problem in the values of the attribute 'horsepower' because of which the mean and median calculation is not possible.

| | mpg | cylin-ders | dis-place-ment | horse-power | weight | accel-eration | model year | origin |
|---|---|---|---|---|---|---|---|---|
| Median | 23 | 4 | 148.5 | ? | 2804 | 15.5 | 76 | 1 |
| Mean | 23.51 | 5.455 | 193.4 | ? | 2970 | 15.57 | 76.01 | 1.573 |
| Deviation | 2.17 | 26.67% | 23.22% | | 5.59% | 0.45% | 0.01% | 36.43% |
| | Low | High | High | | Low | Low | Low | High |

**Fig: Mean vs. Median for Auto MPG**

With a bit of investigation, we can find out that the problem is occurring because of the 6 data elements, as shown in below figure, do not have value for the attribute 'horsepower'.

| mpg | cylinders | displace-ment | horse-power | weight | accel-eration | model year | origin | car name |
|---|---|---|---|---|---|---|---|---|
| 25 | 4 | 98 | ? | 2046 | 19 | 71 | 1 | Ford pinto |
| 21 | 6 | 200 | ? | 2875 | 17 | 74 | 1 | Ford maverick |
| 40.9 | 4 | 85 | ? | 1835 | 17.3 | 80 | 2 | Renault lecar deluxe |
| 23.6 | 4 | 140 | ? | 2905 | 14.3 | 80 | 1 | Ford mustang cobra |
| 34.5 | 4 | 100 | ? | 2320 | 15.8 | 81 | 2 | Renault 18i |
| 23 | 4 | 151 | ? | 3035 | 20.5 | 82 | 1 | Amc concord di |

**Fig: Missing values of attribute 'horsepower' in Auto MPG**

For that reason, the attribute 'horsepower' is not treated as a numeric. That's why the operations applicable on numeric variables, like mean or median, are failing. So we have to first remediate the missing values of the attribute 'horsepower' before being able to do any kind of exploration.

### *ii.* Understanding data spread:

Let's look closely at those attributes, To drill down more, we need to look at the entire range of values of the attributes, though not at the level of data elements as that may be too vast to review manually. So we will take a granular view of the data spread in the form of

- Dispersion of data
- Position of the different data values

### Measuring data dispersion:

Consider the data values of two attributes

**Attribute 1 values :** 44, 46, 48, 45, and 47

**Attribute 2 values :** 34, 46, 59, 39, and 52

- Both the set of values have a mean and median of 46. However, the first set of values that is of attribute 1 is more concentrated or clustered around the mean/median value whereas the second set of values of attribute 2 is quite spread out or dispersed.
- To measure the extent of dispersion of a data, or to find out how much the different values of a data are spread out, the variance of the data is measured.

The variance of a data is measured using the formula given below:

$$\text{Variance}, \quad (x) = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2$$

where x is the variable or attribute whose variance is to be measured
n is the number of observations or values of variable x.

Standard deviation of a data is measured as follows:

$$\text{Standard deviation } (x) = \sqrt{\text{Variance } (x)}$$

**Larger value of variance or standard deviation indicates more dispersion in the data and vice versa.**

In the above example, let's calculate the variance of attribute 1 and that of attribute 2.

For attribute 1,

$$\text{Variance} = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2$$

$$= \frac{44^2 + 46^2 + 48^2 + 45^2 + 47^2}{5} - \left(\frac{44 + 46 + 48 + 45 + 47}{5}\right)^2$$

$$= \frac{1936 + 2116 + 2304 + 2025 + 2209}{5} - \left(\frac{230}{5}\right)^2 = \frac{10590}{5} - (46)^2 = 2$$

For attribute 2,

$$\text{Variance} = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2$$

$$= \frac{34^2 + 46^2 + 59^2 + 39^2 + 52^2}{5} - \left(\frac{34 + 46 + 59 + 39 + 52}{5}\right)^2$$

$$= \frac{1156 + 2116 + 3481 + 1521 + 2704}{5} - \left(\frac{230}{5}\right)^2 = \frac{10978}{5} - (46)^2 = 79.6$$

So it is quite clear from the measure that attribute 1 values are quite concentrated around the mean while attribute 2 values are extremely spread out. Since this data was small, a visual inspection and understanding were possible and that matches with the measured value.

## Measuring data value position:

- When the data values of an attribute are arranged in an increasing order, we know that median gives the central data value, which divides the entire data set into two halves.
- Similarly, if the first half of the data is divided into two halves so that each half consists of one-quarter of the data set, then that median of the first half is known as first quartile or Q1.
- In the same way, if the second half of the data is divided into two halves, then that median of the second half is known as third quartile or Q3. The overall median is also known as second quartile or Q2.
- So, any data set has five values - minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum.

Let's review these values for the attributes 'cylinders', 'displacement', and 'origin'. The below Figure captures a summary of the range of statistics for the attributes.

If we take the example of the attribute 'displacement', we can see that the difference between minimum value and Q1 is 36.2 and the difference between Q1 and median is 44.3. On the contrary, the difference between median and Q3 is 113.5 and Q3 and the maximum value is 193. In other words, the larger values are more spread out than the smaller ones. This helps in understanding why the value of mean is much higher than that of the median for the attribute 'displacement'.

Similarly, in case of attribute 'cylinders', we can observe that the difference between minimum value and median is 1 whereas the difference between median and the maximum value is 4. For the attribute 'origin', the difference between minimum value and median is 0 whereas the difference between median and the maximum value is 2.

| | cylinders | displacement | origin |
|---|---|---|---|
| Minimum | 3 | 68 | 1 |
| Q1 | 4 | 104.2 | 1 |
| Median | 4 | 148.5 | 1 |
| Q3 | 8 | 262 | 2 |
| Maximum | 8 | 455 | 3 |

Fig: Attribute value drill-down for Auto MPG

However, we still cannot ascertain whether there is any outlier present in the data. For that, we can better adopt some means to visualize the data. Box plot is an excellent visualization medium for numeric data.

## 2. Plotting and exploring numerical data:

### i). Box plots:

- ✓ A box plot is an extremely effective mechanism to get a one-shot view and understand the nature of the data.
- ✓ The box plot (also called box and whisker plot) gives a standard visualization of the five-number summary statistics of a data, namely minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum.

**Detailed interpretation of a box plot:**

- The central rectangle or the box spans from first to third quartile (i.e. Q1 to Q3), thus giving the inter-quartile range (IQR).
- Median is given by the line or band within the box.
- The lower whisker extends up to 1.5 times of the inter-quartile range (or IQR) from the bottom of the box, i.e. the first quartile or Q1.
- However, the actual length of the lower whisker depends on the lowest data value that falls within (Q1 − 1.5 times of IQR).

For example. for a specific set of data, Q1 = 73, median = 76 and Q3 = 79. Hence, IQR will be 6 (i.e. Q3 – Q1).So, lower whisker can extend maximum till (Q1 – 1.5 × IQR) = 73 – 1.5 × 6 = 64.

- However, say there are lower range data values such as 70, 63, and 60. So, the lower whisker will come at 70 as this is the lowest data value larger than 64.
- The upper whisker extends up to 1.5 as times of the inter-quartile range (or IQR) from the top of the box, i.e. the third quartile or Q3.
- Similar to lower whisker, the actual length of the upper whisker will also depend on the highest data value that falls within (Q3 + 1.5 times of IQR).

Let's try to understand this with an example. For the same set of data mentioned in the above point, upper whisker can extend maximum till (Q3 + 1.5 × IQR) = 79 + 1.5 × 6 = 88.

- If there is higher range of data values like 82, 84, and 89. So, the upper whisker will come at 84 as this is the highest data value lower than 88.
- The data values coming beyond the lower or upper whiskers are the ones which are of unusually low or high values respectively. These are the outliers, which may deserve special consideration.

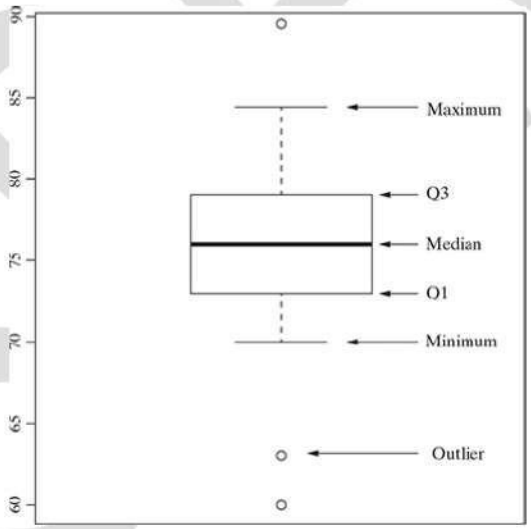



**Fig: Box plot**

Let's visualize the box plot for the three attributes - 'cylinders', 'displacement', and 'origin'. We will also review the box plot of another attribute in which the deviation between mean and median is very little and see what the basic difference in the respective box plots is. Figure below presents the respective box plots.

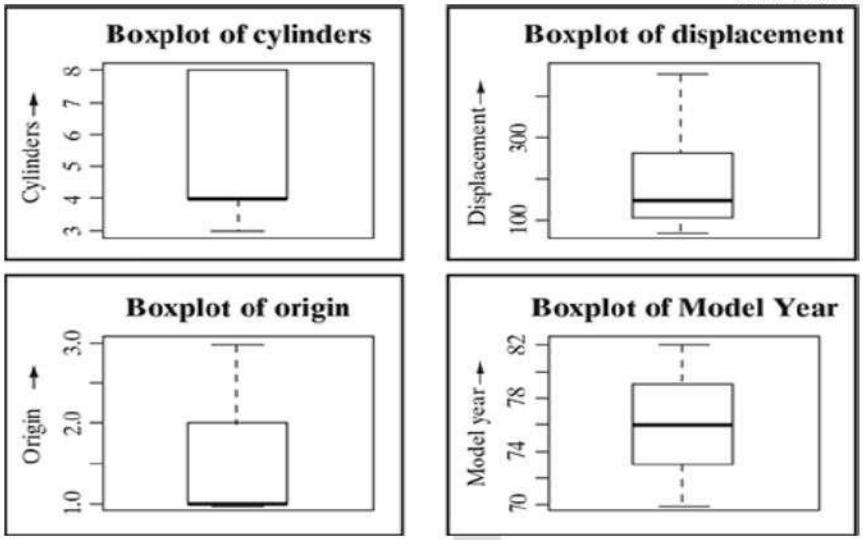



Fig: Box plot of Auto MPG attributes

i. **Analysing box plot for 'cylinders'**
ii. **Analysing box plot for 'origin'**
iii. **Analysing box plot for 'displacement'**
iv. **Analysing box plot for 'model Year'**

### Analysing box plot for 'cylinders' :

The box plot for attribute 'cylinders' looks pretty weird in shape. The upper whisker is missing, the band for median falls at the bottom of the box, even the lower whisker is pretty small compared to the length of the box! Is everything right?

The answer is a big YES, and you if you delve a little deeper into the actual data values of the attribute. The attribute 'cylinders' is discrete in nature having values from 3 to 8. Table 2.2 captures the frequency and cumulative frequency of it.

| Cylinders | Frequency | Cumulative Frequency |
|---|---|---|
| 3 | 4 | 4 |
| 4 | 204 | 208 (= 4 + 204) |
| 5 | 3 | 211 (= 208 + 3) |
| 6 | 84 | 295 (= 211 + 84) |
| 7 | 0 | 295 (= 295 + 0) |
| 8 | 103 | 398 (= 295 + 103) |

Table 2.2 Frequency of "Cylinders" Attribute

- As can be observed in the table, the frequency is extremely high for data value 4.
- Two other data values where the frequency is quite high are 6 and 8. So now if we try to find the quartiles, since the total frequency is 398, the first quartile (Q1), median (Q2), and third quartile (Q3) will be at a cumulative frequency 99.5 (i.e. average of 99th and

100th observation), 199 and 298.5 (i.e. average of 298th and 299th observation), respectively.

- This way Q1 = 4, median = 4 and Q3 = 8. Since there is no data value beyond 8, there is no upper whisker.
- Also, since both Q1 and median are 4, the band for median falls on the bottom of the box.
- Same way, though the lower whisker could have extended till -2 (Q1 – 1.5 × IQR = 4 – 1.5 × 4 = –2), in reality, there is no data value lower than 3. Hence, the lower whisker is also short.
- In any case, a value of cylinders less than 1 is not possible.

**Analysing box plot for 'origin' :**

Like the box plot for attribute 'cylinders', the box plot for attribute 'cylinders' also looks pretty weird in shape. Here the lower whisker is missing and the band for median falls at the bottom of the box! Let's verify if everything right?

Just like the attribute 'cylinders', attribute 'origin' is discrete in nature having values from 1 to 3. Table 2.3 captures the frequency and cumulative frequency (i.e. a summation of frequencies of all previous intervals) of it.

| origin | Frequency | Cumulative Frequency |
|---|---|---|
| 1 | 249 | 249 |
| 2 | 70 | 319 (= 249 + 70) |
| 3 | 79 | 398 (= 319 + 79) |

Table 2.3 Frequency of "Origin" Attribute

As can be observed in the table, the frequency is extremely high for data value 1. Since the total frequency is 398, the first quartile (Q1), median (Q2), and third quartile (Q3) will be at a cumulative frequency 99.5 (i.e. average of 99th and 100th observation), 199 and 298.5 (i.e. average of 298th and 299th observation), respectively. This way Q1 = 1, median = 1, and Q3 = 2. Since Q1 and median are same in value, the band for median falls on the bottom of the box. There is no data value lower than Q1. Hence, the lower whisker is missing.

**Analysing box plot for 'displacement'**

The box plot for the attribute 'displacement' looks better than the previous box plots. However, still, there are few small abnormalities, the cause of which needs to be reviewed. Firstly, the lower whisker is much smaller than an upper whisker. Also, the band for median is closer to the bottom of the box.

Let's take a closer look at the summary data of the attribute 'displacement'. The value of first quartile, Q1 = 104.2, median = 148.5, and third quartile, Q3 = 262. Since (median – Q1) = 44.3 is greater than (Q3 – median) = 113.5, the band for the median is closer to the bottom of the box (which represents Q1). The value of IQR, in this case, is 157.8. So the lower whisker can be 1.5

times 157.8 less than Q1. But minimum data value for the attribute 'displacement' is 68. So, the lower whisker at 15% [(Q1 – minimum)/1.5 × IQR = (104.2 – 68) / (1.5 × 157.8) = 15%] of the permissible length. On the other hand, the maximum data value is 455. So the upper whisker is 81% [(maximum – Q3)/1.5 × IQR = (455 – 262) / (1.5 × 157.8) = 81%] of the permissible length. This is why the upper whisker is much longer than the lower whisker.

**Analysing box plot for 'model Year'**

The box plot for the attribute 'model. year' looks perfect. Let's validate is it really what expected to be.

For the attribute 'model.year':

First quartile, Q1 = 73

Median, Q2 = 76

Third quartile, Q3 = 79

So, the difference between median and Q1 is exactly equal to Q3 and median (both are 3). That is why the band for the median is exactly equidistant from the bottom and top of the box.

I QR = Q3 9 Q1 = 7 9 - 7 3 = 6

Difference between Q1 and minimum data value (i.e. 70) is also same as maximum data value (i.e. 82) and Q3 (both are 3). So both lower and upper whiskers are expected to be of the same size which is 33% [3 / (1.5 × 6)] of the permissible length.

## ii) Histogram:

Histogram is another plot which helps in effective visualization of numeric attributes. It helps in understanding the distribution of a numeric data into series of intervals, also termed as 'bins'. **The important difference between histogram and box plot is**

- The focus of histogram is to plot ranges of data values (acting as 'bins'), the number of data elements in each range will depend on the data distribution. Based on that, the size of each bar corresponding to the different ranges will vary.
- The focus of box plot is to divide the data elements in a data set into four equal portions, such that each portion contains an equal number of data elements.

Histograms might be of different shapes depending on the nature of the data, e.g. skewness. The Below figure provides a depiction of different shapes of the histogram that are generally created. These patterns give us a quick understanding of the data and thus act as a great data exploration tool.

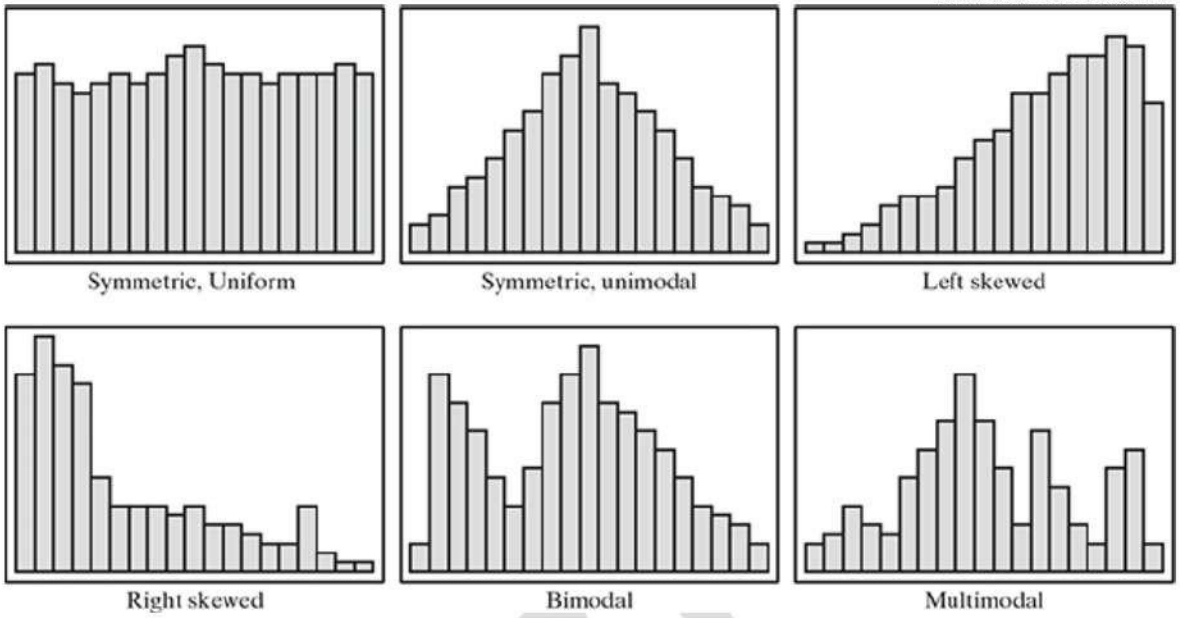



**Fig :General Histogram shapes**

Let's now examine the histograms for the different attributes of Auto MPG data set. The histograms for 'mpg' and 'weight' are right-skewed. The histogram for 'acceleration' is symmetric and unimodal, whereas the one for 'model.year' is symmetric and uniform. For the remaining attributes, histograms are multimodal in nature.

Now let's dig deep into one of the histograms, say the one for the attribute 'acceleration'. The histogram is composed of a number of bars, one bar appearing for each of the 'bins'. The height of the bar reflects the total count of data elements whose value falls within the specific bin value, or the frequency. Talking in context of the histogram for acceleration, each 'bin' represents an acceleration value interval of 2 units. So the second bin, e.g., reflects acceleration value of 10 to 12 units. The corresponding bar chart height reflects the count of all data elements whose value lies between 10 and 12 units. Also, it is evident from the histogram that it spans over the acceleration value of 8 to 26 units. The frequency of data elements corresponding to the bins first keep on increasing, till it reaches the bin of range 14 to 16 units. At this range, the bar is tallest in size. So we can conclude that a maximum number of data elements fall within this range. After this range, the bar size starts decreasing till the end of the whole range at the acceleration value of 26 units.

Please note that when the histogram is uniform, as in the case of attribute 'model. year', it gives a hint that all values are equally likely to occur.

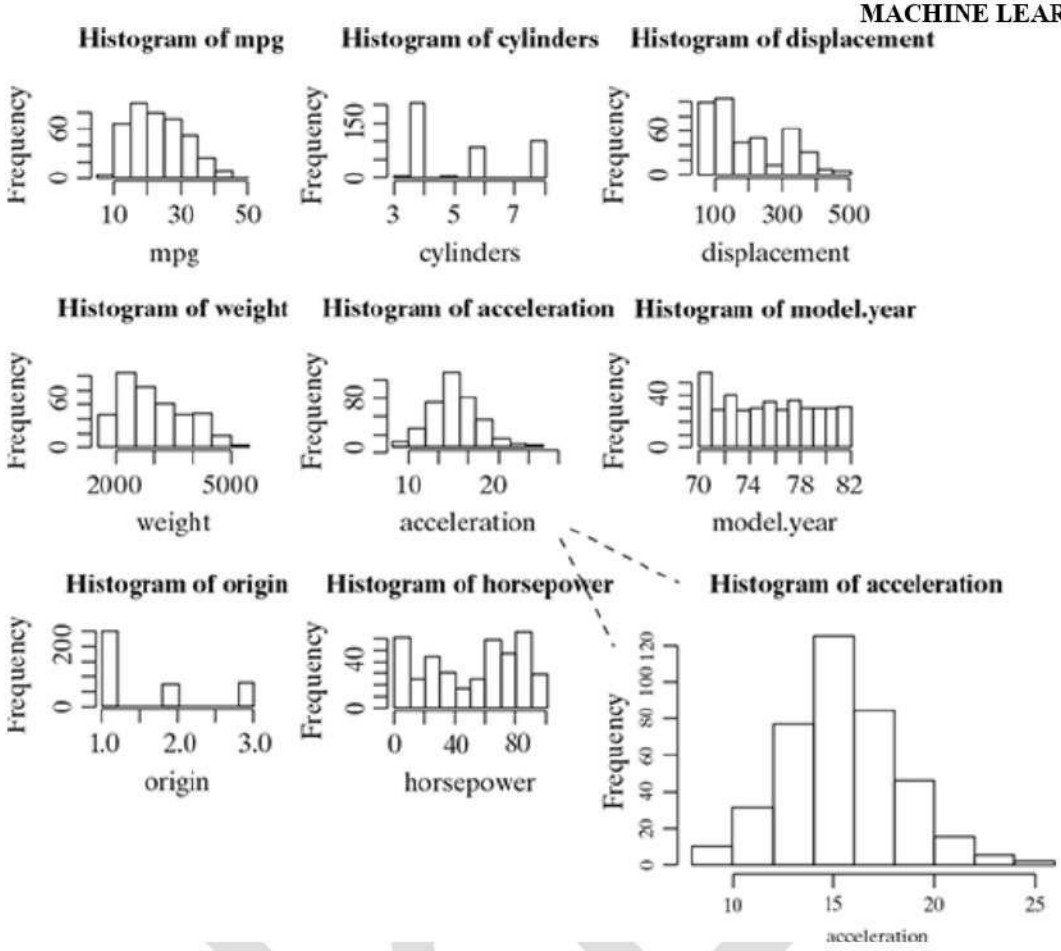



**Fig: Histogram Auto MPG attributes**

## 3 .Exploring categorical data:

We have seen there are multiple ways to explore numeric data. However, there are not many options for exploring categorical data. In the Auto MPG data set, attribute 'car.name' is categorical in nature. Also, as we discussed earlier, we may consider 'cylinders' as a categorical variable instead of a numeric variable.

The first summary which we may be interested in noting is how many unique names are there for the attribute 'car name' or how many unique values are there for 'cylinders' attribute. We can get this as follows:

For attribute 'car name'

1. Chevrolet chevelle malibu
2. Buick skylark 320
3. Plymouth satellite
4. Amc rebel sst
5. Ford torino
6. Ford galaxie 500

7. Chevrolet impala
8. Plymouth fury iii
9. Pontiac catalina
10. Amc ambassador dpl

For attribute 'cylinders'

8 4 6 3 5

We may also look for a little more details and want to get a table consisting the categories of the attribute and count of the data elements falling into that category. Tables 2.4 and 2.5 contain these details.

For attribute 'car name'

Table : Count of Categories for 'car name' Attribute

| **Attribute Value** | amc ambas-sador brougham | amc ambas-sador dpl | amc ambassa-dor sst | amc concord | amc concord d/l | amc con-cord dl 6 | amc gremlin | ... |
|---|---|---|---|---|---|---|---|---|
| **Count** | 1 | 1 | 1 | 1 | 2 | 2 | 4 | ... |

For attribute "cylinders"

| **Attribute Value** | 3 | 4 | 5 | 6 | 8 |
|---|---|---|---|---|---|
| **Count** | 4 | 204 | 3 | 84 | 103 |

Table : Count of Categories for 'Cylinders' Attribute

In the same way, we may also be interested to know the proportion (or percentage) of count of data elements belonging to a category. Say, e.g., for the attributes 'cylinders', the proportion of data elements belonging to the category 4 is 204 ÷ 398 = 0.513, i.e. 51.3%. Tables 2.6 and 2.7 contain the summarization of the categorical attributes by proportion of data elements.

For attribute 'car name'

| **Attribute Value** | Amc ambas-sador brougham | Amc ambassa-dor dpl | Amc ambassa-dor sst | Amc concord | Amc concord d/l | Amc concord dl 6 | Amc gremlin | ... |
|---|---|---|---|---|---|---|---|---|
| **Count** | 0.003 | 0.003 | 0.003 | 0.003 | 0.005 | 0.005 | 0.01 | ... |

Table: Proportion of Categories for '"Cylinders' Attribute

For attribute 'cylinders'

Table: Proportion of Categories for "Cylinders" Attribute

| Attribute Value | 3 | 4 | 5 | 6 | 8 |
|---|---|---|---|---|---|
| Count | 0.01 | 0.513 | 0.008 | 0.211 | 0.259 |

Last but not the least, as we have read in the earlier section on types of data, statistical measure "mode" is applicable on categorical attributes. As we know, like mean and median, mode is also a statistical measure for central tendency of a data. Mode of a data is the data value which appears most often. In context of categorical attribute, it is the category which has highest number of data values. Since mean and median cannot be applied for categorical variables, mode is the sole measure of central tendency.

Let's try to find out the mode for the attributes 'car name' and 'cylinders'. For cylinders, since the number of categories is less and we have the entire table listed above, we can see that the mode is 4, as that is the data value for which frequency is highest. More than 50% of data elements belong to the category 4. However, it is not so evident for the attribute 'car name' from the information given above. When we probe and try to find the mode, it is found to be category 'ford pinto' for which frequency is of highest value 6.

An attribute may have one or more modes. Frequency distribution of an attribute having single mode is called 'unimodal', two modes are called 'bimodal' and multiple modes are called 'multimodal'.

### 4. Exploring relationship between variables:

One of the most important angle of data exploration is to explore relationship between attributes. There are multiple plots to enable us explore the relationship between variables. The basic and most commonly used plot is scatter plot.

## Scatter plot:

A scatter plot helps in visualizing bivariate relationships, i.e. relationship between two variables. It is a two-dimensional plot in which points or dots are drawn on coordinates provided by values of the attributes.

For example, in a data set there are two attributes – attr_1 and attr_2. We want to understand the relationship between two attributes, i.e. with a change in value of one attribute, say attr_1, how does the value of the other attribute, say attr_2, changes. We can draw a scatter plot, with attr_1 mapped to x-axis and attr_2 mapped in y-axis. So, every point in the plot will have value of attr_1 in the x-coordinate and value of attr_2 in the y-coordinate. As in a two-dimensional plot, attr_1 is said to be the independent variable and attr_2 as the dependent variable.

Let's take a real example in this context. In the data set Auto MPG, there is expected to be some relation between the attributes 'displacement' and 'mpg'. Let's try to verify our intuition using the scatter plot of 'displacement' and 'mpg'. Let's map 'displacement' as the x-coordinate and 'mpg' as the y-coordinate. The scatter plot comes as in Figure 2.13.

As is evident in the scatter plot, there is a definite relation between the two variables. The value of 'mpg' seems to steadily decrease with the increase in the value of 'displacement'. It may come in our mind that what is the extent of relationship? Well, it can be reviewed by calculating the correlation between the variables. One more interesting fact to notice is that there are certain data values which stand-out of the others. For example, there is one data element which has a mpg of 37 for a displacement of 250. This record is completely different from other data elements having similar displacement value but mpg value in the range of 15 to 25. This gives an indication that of presence of outlier data values.

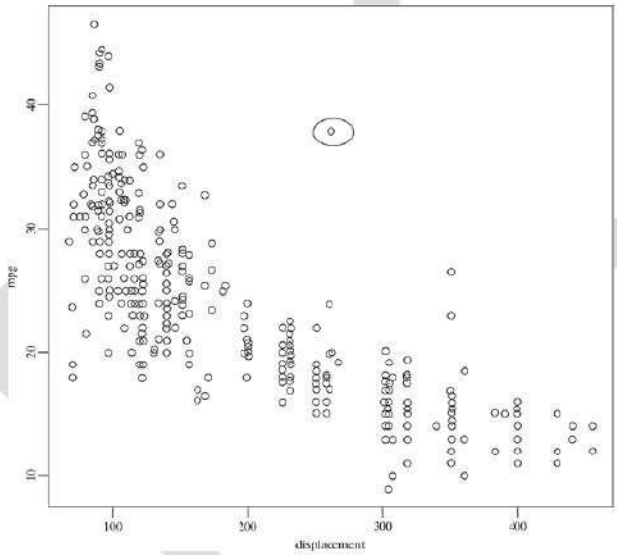



**Fig: Scatter plot of 'displacement' and 'mpg'**

In Below Figure, the pair wise relationship among the features – 'mpg', 'displacement', 'horsepower', 'weight', and 'acceleration' have been captured. As you can see, in most of the cases, there is a significant relationship between the attribute pairs. However, in some cases, e.g. between attributes 'weight' and 'acceleration', the relationship doesn't seem to be very strong.

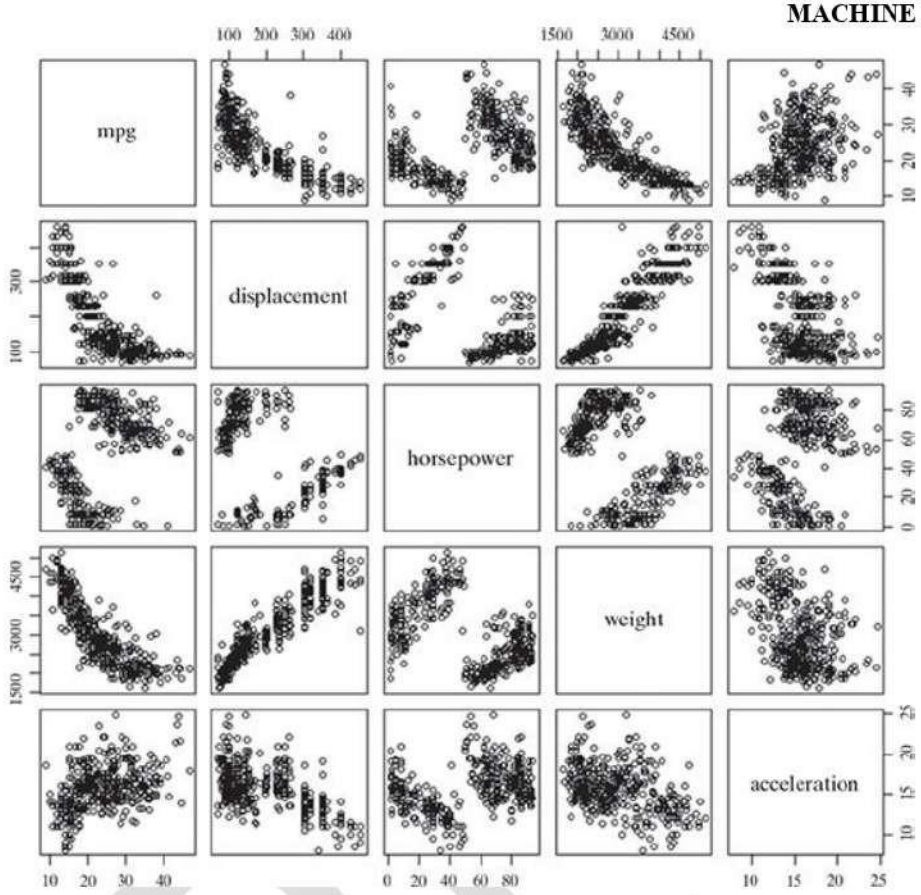



**Fig: Pair wise scatter plot between different attributes of Auto MPG**

## Two-way cross-tabulations:

Two-way cross-tabulations (also called cross-tab or contingency table) are used to understand the relationship of two categorical attributes in a concise way. It has a matrix format that presents a summarized view of the bivariate frequency distribution. A cross-tab, very much like a scatter plot, helps to understand how much the data values of one attribute changes with the change in data values of another attribute. Let's try to see with examples, in context of the Auto MPG data set.

Let's assume the attributes 'cylinders', 'model.year', and 'origin' as categorical and try to examine the variation of one with respect to the other. As we understand, attribute 'cylinders' reflects the number of cylinders in a car and assumes values 3, 4, 5, 6, and 8. Attribute 'model.year' captures the model year of each of the car and 'origin' gives the region of the car, the values for origin 1, 2, and 3 corresponding to North America, Europe, and Asia. Below are the cross-tabs. Let's try to understand what information they actually provide.

The first cross-tab, i.e. the one showing relationship between attributes 'model. year' and 'origin' help us understand the number of vehicles per year in each of the regions North America,

Europe, and Asia. Looking at it in another way, we can get the count of vehicles per region over the different years. All these are in the context of the sample data given in the Auto MPG data set.

Moving to the second cross-tab, it gives the number of 3, 4, 5, 6, or 8 cylinder cars in every region present in the sample data set. The last cross-tab presents the number of 3, 4, 5, 6, or 8 cylinder cars every year.

We may also want to create cross-tabs with a more summarized view like have a cross-tab giving a number of cars having 4 or less cylinders and more than 4 cylinders in each region or by the years. This can be done by rolling up data values by the attribute 'cylinder'. Tables 2.8–2.10 present cross-tabs for different attribute combinations.

'Model year' vs. 'origin'

| Origin \ Model Year | 70 | 71 | 72 | 73 | 74 | 75 | 76 | 77 | 78 | 79 | 80 | 81 | 82 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 22 | 20 | 18 | 29 | 15 | 20 | 22 | 18 | 22 | 23 | 7 | 13 | 20 |
| 2 | 5 | 4 | 5 | 7 | 6 | 6 | 8 | 4 | 6 | 4 | 9 | 4 | 2 |
| 3 | 2 | 4 | 5 | 4 | 6 | 4 | 4 | 6 | 8 | 2 | 13 | 12 | 9 |

Table:Cross-tab for 'Model year' vs. 'Origin'

'Cylinders' vs. 'Origin'

| Cylinders \ Origin | 1 | 2 | 3 |
|---|---|---|---|
| 3 | 0 | 0 | 4 |
| 4 | 72 | 63 | 69 |
| 5 | 0 | 3 | 0 |
| 6 | 74 | 4 | 6 |
| 8 | 103 | 0 | 0 |

Cross-tab for 'Cylinders' vs. 'Origin'

'Cylinders' vs. 'Model year'

| Cylinders \ Model Year | 70 | 71 | 72 | 73 | 74 | 75 | 76 | 77 | 78 | 79 | 80 | 81 | 82 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| 4 | 7 | 13 | 14 | 11 | 15 | 12 | 15 | 14 | 17 | 12 | 25 | 21 | 28 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 |
| 6 | 4 | 8 | 0 | 8 | 7 | 12 | 10 | 5 | 12 | 6 | 2 | 7 | 3 |
| 8 | 18 | 7 | 13 | 20 | 5 | 6 | 9 | 8 | 6 | 10 | 0 | 1 | 0 |

Table :Cross-tab for 'Cylinders' vs. 'Model year'

# DATA QUALITY AND REMEDIATION:

## 1. DATA QUALITY:

Success of machine learning depends largely on the quality of data. A data which has the right quality helps to achieve better prediction accuracy, in case of supervised learning. However, it is not realistic to expect that the data will be flawless. We have already come across at least two types of problems:

i. Certain data elements without a value or data with a missing value.
ii. Data elements having value surprisingly different from the other elements, which we term as outliers.

There are multiple factors which lead to these data quality issues. Following are some of them:

**Incorrect sample set selection:** The data may not reflect normal or regular quality due to incorrect selection of sample set.

For example, if we are selecting a sample set of sales transactions from a festive period and trying to use that data to predict sales in future.

In this case, the prediction will be far apart from the actual scenario, just because the sample set has been selected in a wrong time.

Similarly, if we are trying to predict poll results using a training data which doesn't comprise of a right mix of voters from different segments such as age, sex, ethnic diversities, etc., the prediction is bound to be a failure. It may also happen due to incorrect sample size.

For example, a sample of small size may not be able to capture all aspects or information needed for right learning of the model.

**Errors in data collection:** resulting in outliers and missing values

In many cases, a person or group of persons are responsible for the collection of data to be used in a learning activity.

In this manual process, there is the possibility of wrongly recording data either in terms of value (say 20.67 is wrongly recorded as 206.7 or 2.067) or in terms of a unit of measurement (say cm. is wrongly recorded as m. or mm.).

This may result in data elements which have abnormally high or low value from other elements. Such records are termed as outliers.

It may also happen that the data is not recorded at all.

In case of a survey conducted to collect data, it is all the more possible as survey responders may choose not to respond to a certain question.

So the data value for that data element in that responder's record is missing.

## 2.DATA REMEDIATION:

The issues in data quality, need to be remediated, if the right amount of efficiency has to be achieved in the learning activity. Out of the two major areas mentioned, the first one can be remedied by proper sampling technique. This is a completely different area – covered as a specialized subject area in statistics.

However, human errors are bound to happen, no matter whatever checks and balances we put in. Hence, proper remedial steps need to be taken for the second area mentioned above.

### i. Handling outliers:

Outliers are data elements with an abnormally high value which may impact prediction accuracy, especially in regression models. Once the outliers are identified and the decision has been taken to amend those values, you may consider one of the following approaches. However, if the outliers are natural, i.e. the value of the data element is surprisingly high or low because of a valid reason, then we should not amend it.

**Remove outliers**: If the number of records which are outliers is not many, a simple approach may be to remove them.

***Imputation*:** One other way is to impute the value with mean or median or mode. The value of the most similar data element may also be used for imputation.

***Capping***: For values that lie outside the 1.5|×| IQR limits, we can cap them by replacing those observations below the lower limit with the value of 5th percentile and those that lie above the upper limit, with the value of 95th percentile.

If there is a significant number of outliers, they should be treated separately in the statistical model. In that case, the groups should be treated as two different groups, the model should be built for both groups and then the output can be combined.

### ii. Handling missing values

In a data set, one or more data elements may have missing values in multiple records. It can be caused by omission on part of the surveyor or a person who is collecting sample data or by the responder, primarily due to his/her unwillingness to respond or lack of understanding needed to provide a response.

It may happen that a specific question (based on which the value of a data element originates) is not applicable to a person or object with respect to which data is collected.

There are multiple strategies to handle missing value of data elements. Some of those strategies have been discussed below.

**Eliminate records having a missing value of data elements:**

In case the proportion of data elements having missing values is within a tolerable limit, a simple but effective approach is to remove the records having such data elements. This is possible if the quantum of data left after removing the data elements having missing values is sizeable.

In the case of Auto MPG data set, only in 6 out of 398 records, the value of attribute 'horsepower' is missing. If we get rid of those 6 records, we will still have 392 records, which is definitely a substantial number. So, we can very well eliminate the records and keep working with the remaining data set.

However, this will not be possible if the proportion of records having data elements with missing value is really high as that will reduce the power of model because of reduction in the training data size.

**Imputing missing values:**

Imputation is a method to assign a value to the data elements having missing values. Mean/mode/median is most frequently assigned value. For quantitative attributes, all missing values are imputed with the mean, median, or mode of the remaining values under the same attribute

. For qualitative attributes, all missing values are imputed by the mode of all remaining values of the same attribute. However, another strategy may be identify the similar types of observations whose values are known and use the mean/median/mode of those known values.

For example, in context of the attribute 'horsepower' of the Auto MPG data set, since the attribute is quantitative, we take a mean or median of the remaining data element values and assign that to all data elements having a missing value. So, we may assign the mean, which is 104.47 and assign it to all the six data elements. The other approach is that we can take a similarity based mean or median. If we refer to the six observations with missing values for attribute 'horsepower' as depicted in Table 2.11, 'cylinders' is the attribute which is logically most connected to 'horsepower' because with the increase in number of cylinders of a car, the horsepower of the car is expected to increase. So, for five observations, we can use the mean of data elements of the 'horsepower' attribute having cylinders = 4; i.e. 78.28 and for one observation which has cylinders = 6, we can use a similar mean of data elements with cylinders = 6, i.e. 101.5, to impute value to the missing data elements.

Table 2.11 Missing Values for 'Horsepower' Attribute

| mpg | cylin-ders | dis-place-ment | horse-power | weight | accel-eration | model year | origin | car name |
|---|---|---|---|---|---|---|---|---|
| 25 | 4 | 98 | ? | 2046 | 19 | 71 | 1 | Ford pinto |
| 21 | 6 | 200 | ? | 2875 | 17 | 74 | 1 | Ford maverick |
| 40.9 | 4 | 85 | ? | 1835 | 17.3 | 80 | 2 | Renault lecar deluxe |
| 23.6 | 4 | 140 | ? | 2905 | 14.3 | 80 | 1 | Ford mustang cobra |
| 34.5 | 4 | 100 | ? | 2320 | 15.8 | 81 | 2 | Renault 18i |
| 23 | 4 | 151 | ? | 3035 | 20.5 | 82 | 1 | Amc concord dl |

**Estimate missing values:**

If there are data points similar to the ones with missing attribute values, then the attribute values from those similar data points can be planted in place of the missing value. For finding similar data points or observations, distance function can be used.

For example, let's assume that the weight of a Russian student having age 12 years and height 5 ft. is missing. Then the weight of any other Russian student having age close to 12 years and height close to 5 ft. can be assigned.

## DATA PRE-PROCESSING:

### 1 .Dimensionality reduction:

Till the end of the 1990s, very few domains were explored which included data sets with a high number of attributes or features. In general, the data sets used in machine learning used to be in few 10s. However, in the last two decades, there has been a rapid advent of computational biology like genome projects. These projects have produced extremely high-dimensional data sets with 20,000 or more features being very common. Also, there has been a wide-spread adoption of social networking leading to a need for text classification for customer behaviour analysis.

High-dimensional data sets need a high amount of computational space and time. At the same time, not all features are useful – they degrade the performance of machine learning algorithms. Most of the machine learning algorithms perform better if the dimensionality of data set, i.e. the number of features in the data set, is reduced.

a) Dimensionality reduction helps in reducing irrelevance and redundancy in features. Also, it is easier to understand a model if the number of features involved in the learning activity is less.

b) Dimensionality reduction refers to the techniques of reducing the dimensionality of a data set by creating new attributes by combining the original attributes.

The most common approach for dimensionality reduction is known as Principal Component Analysis (PCA).

- PCA is a statistical technique to convert a set of correlated variables into a set of transformed, uncorrelated variables called principal components
- . The principal components are a linear combination of the original variables. They are orthogonal to each other
- . Since principal components are uncorrelated, they capture the maximum amount of variability in the data.
- However, the only challenge is that the original attributes are lost due to the transformation.

Another commonly used technique which is used for dimensionality reduction is Singular Value Decomposition (SVD).

## 2. Feature subset selection:

Feature subset selection or simply called feature selection, both for supervised as well as unsupervised learning, try to find out the optimal subset of the entire feature set which significantly reduces computational cost without any major impact on the learning accuracy. It may seem that a feature subset may lead to loss of useful information as certain features are going to be excluded from the final set of features used for learning. However, for elimination only features which are not relevant or redundant are selected.

A feature is considered as irrelevant if it plays an insignificant role (or contributes almost no information) in classifying or grouping together a set of data instances. All irrelevant features are eliminated while selecting the final feature subset. A feature is potentially redundant when the information contributed by the feature is more or less same as one or more other features. Among a group of potentially redundant features, a small number of features can be selected as a part of the final feature subset without causing any negative impact to learn model accuracy.