Dr. SNS RAJALAKSHMI COLLEGE OF ARTS & SCIENCE
(AUTONOMOUS)

Accredited by NAAC (Cycle III) with 'A+' Grade
Affiliated to Bharathiar University
Coimbatore-641049

ESTD : 1999

DEPARTMENT OF COMPUTER APPLICATIONS

II BCA

ELECTIVE TRACK 2 - HIGHER EDUCATION

21UCU807 : DATA SCIENCE

UNIT-II

VISUALIZATION AND SIMPLE METRICS:

The field of data visualization covers a wide range of techniques and algorithms from the simple visual data representations to the complex three dimensions (3D) data animation applications. The goal of visual design is to create high quality, clear, easy to understand, and quick to perceive depictions. Therefore, the quality of the created visualization has always been a principal motivator for researchers. It is therefore crucial to use different types of judgments that could be applied to the visualization to reach the needed quality.

PIE CHARTS:

The pie chart is a pictorial representation of data that makes it possible to visualize the relationships between the parts and the whole of a variable. For example, it is possible to understand the industry count or percentage of a variable level from the division by areas or sectors. The recommended use for pie charts is two-dimensional, as three-dimensional use can be confusing.

Like a pie chart, the total of the data that make up the segments must equal 360°, or the sum of the values of the circumference must always be 100%. To calculate the percentage of a pie chart, it is necessary to: categorize the data, calculate the total, divide the categories, convert the rates and calculate the degrees. Thus, the formula for the pie chart is:

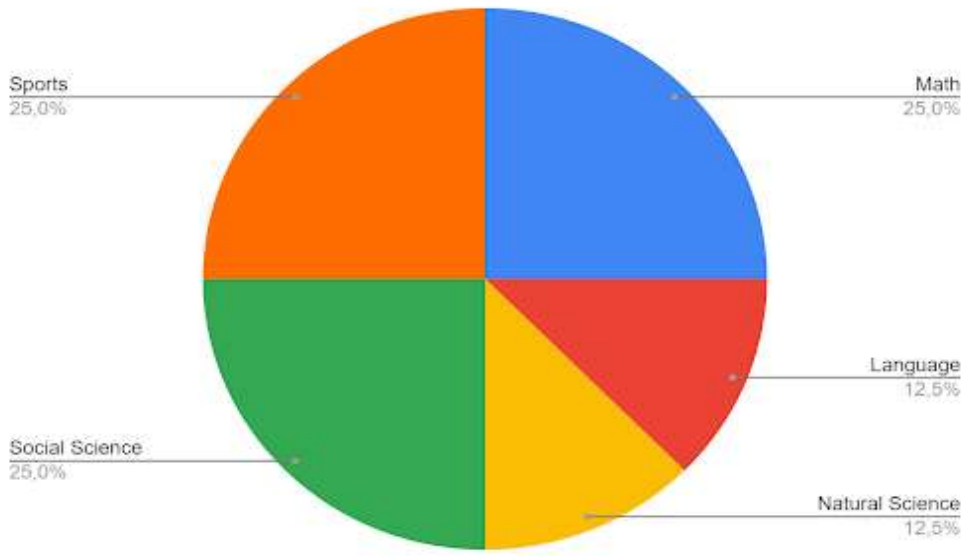
(Data given / total value of data) x 360

The dimensions form sectors of the measurement values; they can have one or two sizes and up to two measures. The first dimension is used to define the angle of each sector that makes up the chart and the second dimension optionally determines the radius of each sector. Additionally, these plots are useful for comparing data over a fixed period since they do not show changes over time. Therefore, their use should be considered if:

You are looking to categorize and compare a set of data.

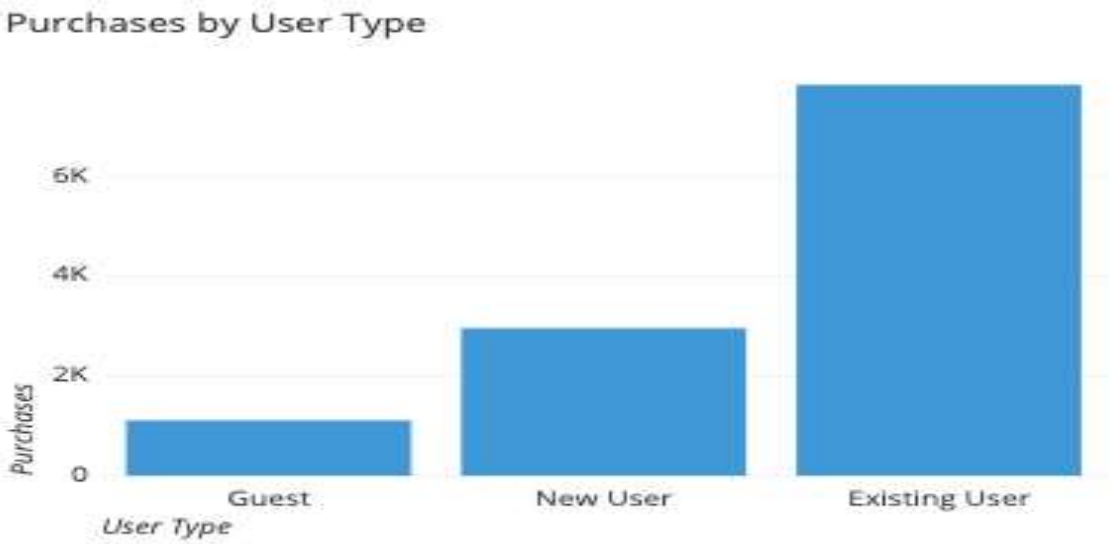
You only have positive values.

You have less than seven categories since a larger number can make it difficult to perceive each segment.



BAR CHART:

A bar chart (aka bar graph, column chart) plots numeric values for levels of a categorical feature as bars. Levels are plotted on one chart axis, and values are plotted on the other axis. Each categorical value claims one bar, and the length of each bar corresponds to the bar's value. Bars are plotted on a common baseline to allow for easy comparison of values.



This example bar chart depicts the number of purchases made on a site by different types of users. The categorical feature, user type, is plotted on the horizontal axis, and each bar's height corresponds to the number of purchases made under each user type. We can see from this chart that while there are about three times as many purchases from new users who create user accounts than those that do not create user accounts (guests), both are dwarfed by the number of purchases made by repeating users.

HISTOGRAM:

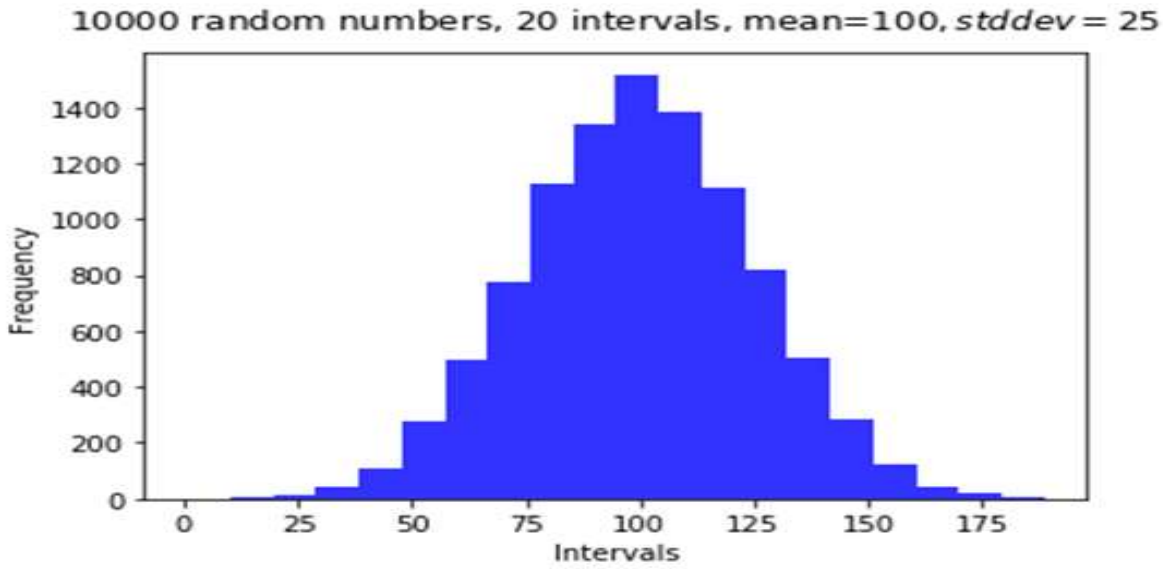
A histogram is a graphical representation of the distribution of a dataset. Although its appearance is similar to that of a standard bar graph, instead of making comparisons between different items or categories or showing trends over time, a histogram is a plot that lets you show the underlying frequency distribution or the **probability distribution** of a single **continuous numerical variable.**

Histograms are two-dimensional plots with two axes; the vertical axis is a frequency axis whilst the horizontal axis is divided into a range of numeric values (intervals or **bins**) or time intervals. The frequency of each bin is shown by the area of vertical rectangular bars. Each bar covers a range of continuous numeric values of the variable under study. The vertical axis shows frequency values derived from counts for each bin.

The midpoint value is the one that gives the name to the interval. When a numerical value corresponds exactly to one of the boundaries of the interval, it will be assigned to the left

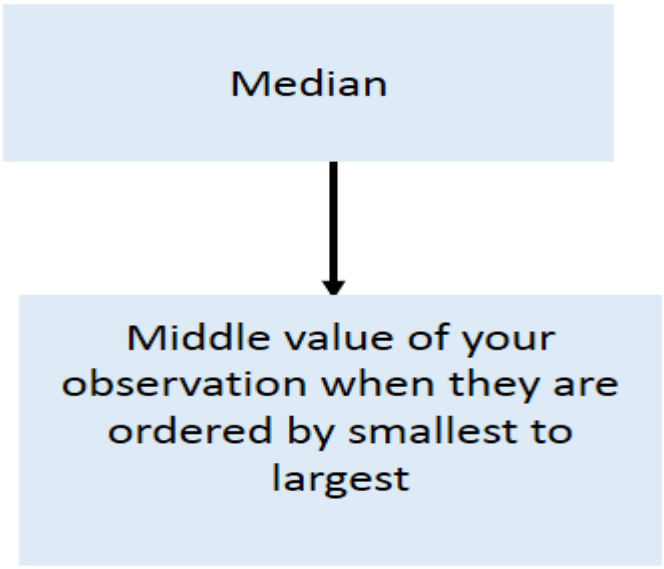
or right interval according to the default setting of the visualization tool. Some tools have the possibility to modify this default setting to accommodate it to the preferences or needs of the users.

Histograms sometimes have bars of unequal width. However, it is usual to plot them with the same width in order to represent equal ranges of data for each interval. As a counterexample the following case can be indicated: collect data from individuals in a population, split the data between bins of 10-year age ranges but accumulate in a single interval data from people over 75 years old. When the **bandwidth** is the same for all intervals, it is equivalent to replace the bar area with the bar length.



MEDIAN:

The second measure of central tendency is the median. The median is nothing more than the middle value of your observations when they are order from the smallest to the largest.

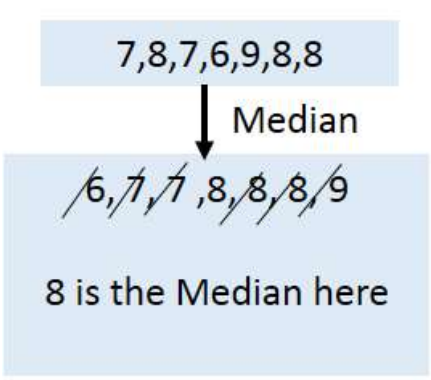


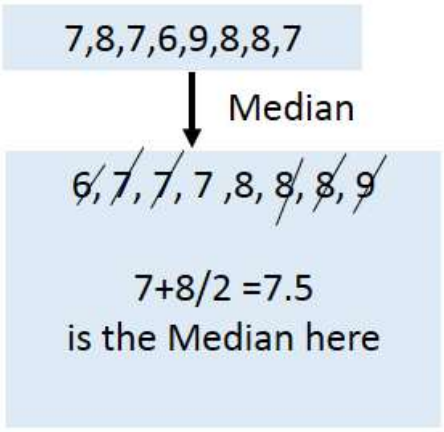
It involves two steps:

1. Oder your cases from smallest to largest
2. Find the middle Value

- If you have odd number of cases then finding middle value is easy. Let's think you have 5 cases. So, after ordering always 3rd position is the middle value.
- If you have even number no cases (let's think 6 cases). In this case there is no single middle value. Then how do we calculate median? Well, we just take the average of the two middle values.

Example:

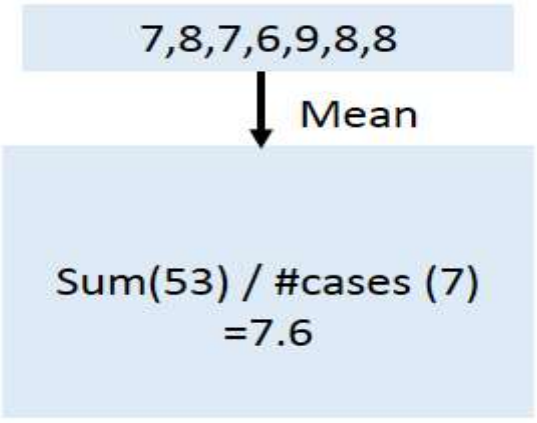




MEAN:

The third measure of central tendency is the most often used one, and also the one you most probably already know quite well: the mean. The mean is the sum of all the values divided by the number of observations. It is nothing but the average value.

$$\overline{X} = \frac{\Sigma X}{n}$$



- If data is Categorical (Nominal or Ordinal) it is impossible to calculate mean or median. So, go for mode.
- If your data is quantitative then go for mean or median. Basically, if your data is having some influential outliers or data is highly skewed then median is the best measurement for finding central tendency. Otherwise go for Mean.

STANDARD DEVIATION:

Standard deviation is a measure of the amount of variation or dispersion of a set of values or observations. Standard deviations are calculated as the square root of variance. Variance is a measure of how far the data points are compared to the mean. A low standard deviation indicates that the values are closer to the mean, whereas a high standard deviation is an indication of extreme values or skewness of the data. Standard deviation like dispersion is used as a metric to analyze the distribution of data.

Time to pick up a calculator and do the math. The formulas for variance and skewness are outlined on the right side of the image. Standard deviation can be used to answer a multitude of questions ranging from:

- Are we observing variation in sales? — Yes, observe the minimum and maximum figures
- How different are the sales figures? — The sales figures on an average lies between 38 and 54 which are +- 1 standard deviation of mean
- How many days observed such a trend mentioned above? — 79% of the sales figure lies between 38 and 54 which means although we observe a variation of sales, the variations aren't significant enough when compared to the distribution.

QUANTILES:

Quantiles are the set of values/points that divides the dataset into groups of equal size. For example, in the figure, there are nine values that splits the dataset. Those nine values are quantiles**.**

Quantiles are values that split sorted data or a probability distribution into equal parts. In general terms, a q-quantile divides sorted data into q parts. The most commonly used

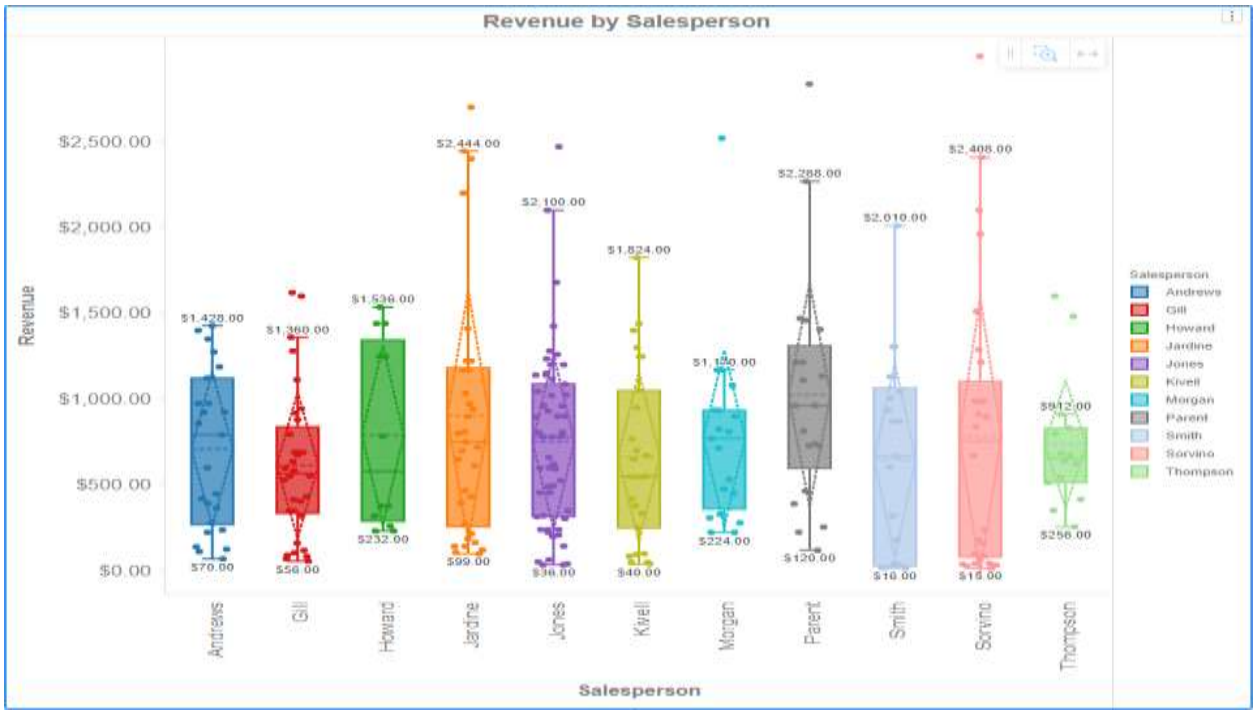
quantiles have special names: Quartiles (4-quantiles): Three quartiles split the data into four parts.

Quantiles are the set of values/points that divides the dataset into groups of equal size. For example, in the figure, there are nine values that splits the dataset. Those nine values are quantiles.

BOX PLOTS:

A box plot visualization allows you to examine the distribution of data. One box plot appears for each attribute element. Each box plot displays the minimum, first quartile, median, third quartile, and maximum values. In addition, you can choose to display the mean and standard deviation as dashed lines. Outliers appear as points in the visualization. You can adjust the spacing between points (that is, jitter) to avoid overlap. A box plot must include at least one metric and at least one attribute.

.



SCATTER PLOT:

A scatter plot is a type of data visualization that shows the relationship between different variables. This data is shown by placing various data points between an x- and y-axis.

Essentially, each of these data points looks “scattered” around the graph, giving this type of data visualization its name.

Scatter plots can also be known as scatter diagrams or x-y graphs, and the point of using one of these is to determine if there are patterns or correlations between two variables.

The two variables are the square footage of a home versus its price. We pulled a sample data set of a couple handfuls of homes to see if we could determine a relationship between these two variables.

As the x-axis goes from the smallest size to the largest, we can see that there is a slight positive correlation showing that as square footage increases, so does the price.

Of course there could be other factors contributing to this, like location or recent renovations, but we can see from this scatter diagram that there is a correlation between the square footage and home cost.

The patterns or correlations found within a scatter plot will have a few different features.

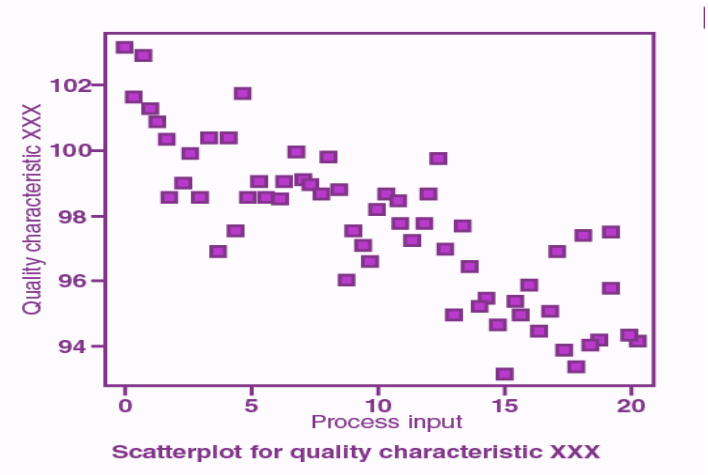
- **Linear or Nonlinear:**

  A linear correlation forms a straight line in its data points while a nonlinear correlation might have a curve or other form within the data points.
- **Strong or Weak:**

  A strong correlation will have data points close together while a weak correlation will have data points that are further apart.
- **Positive or Negative:**

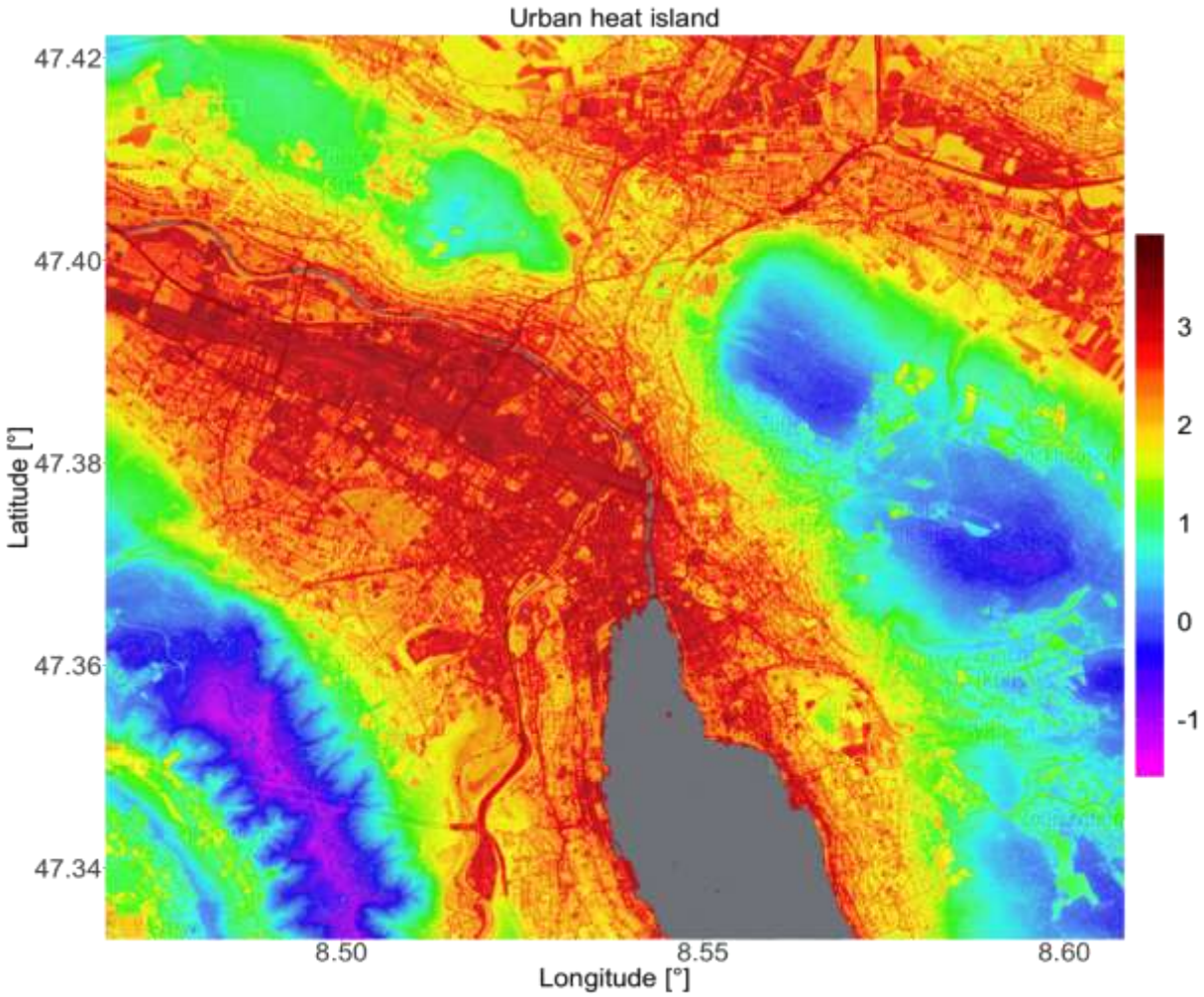
  A positive correlation will point up (i.e., the x- and y-values are both increasing) while a negative correlation will point down (i.e., the x-values are increasing while the corresponding y-values are decreasing).

However, if you don't see any of these features present within your graph that means there’s no correlation between your data.



HEAT MAPS:

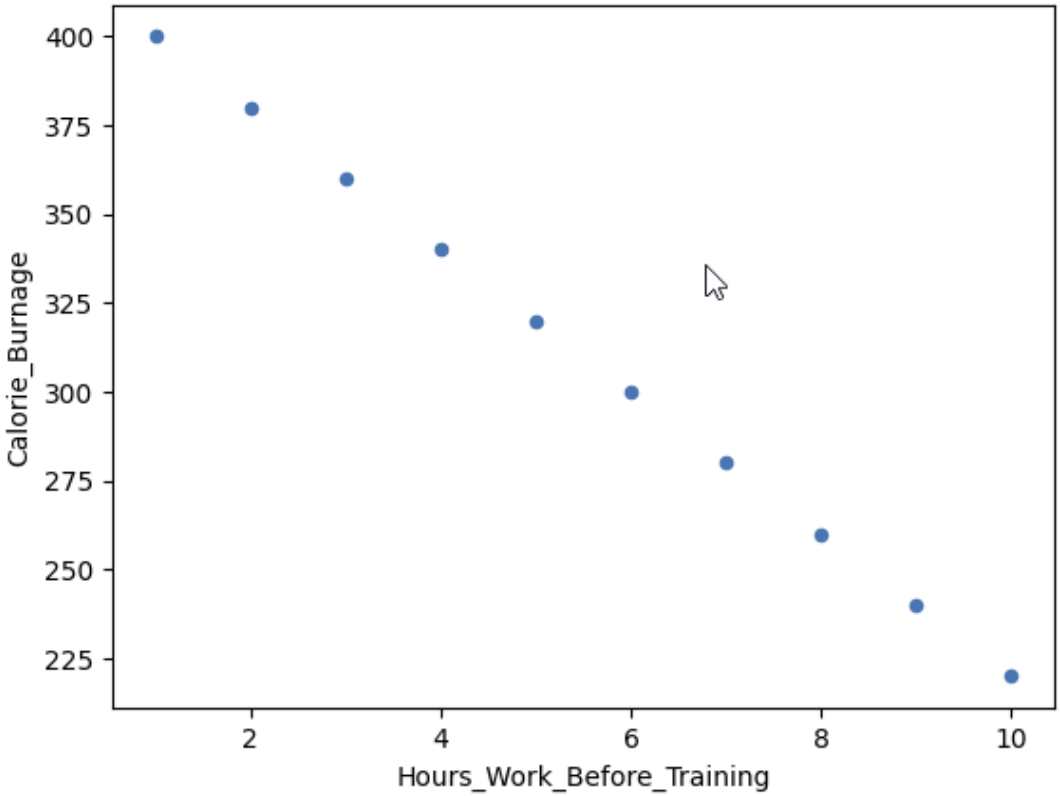
Heat map is a graphical way to visualize visitor behaviour data in the form of hot and cold spots employing a warm-to-cool colour scheme. The warm colours indicate sections with the most visitor interaction, red being the area of highest interaction, and the cool colours point to the sections with the lowest interaction.



Heat map analysis is the process of reviewing and analyzing heat map data to gather insights about user interaction and behavior as they engage with your product. This data analysis can lead to improved site designs with lower bounce rates, reduced churn, fewer drop-offs, more page views, and better conversion rates.

CORRELATION:

Correlation measures the relationship between two variables. We mentioned that a function has a purpose to predict a value, by converting input (x) to output (f(x)). We can say also say that a function uses the relationship between two variables for prediction.



TYPES OF CORRELATION:

Positive Linear Correlation. There is a positive linear correlation when the variable on the x -axis increases as the variable on the y -axis increases. ...

Negative Linear Correlation. ...

Non-linear Correlation (known as curvilinear correlation) ...

No-Correlation.

TIME SERIES:

Time series analysis is a specific way of analyzing a sequence of data points collected over an interval of time. In time series analysis, analysts record data points at consistent intervals over a set period of time rather than just recording the data points intermittently or randomly.

Weather records, economic indicators and patient health evolution metrics — all are time series data. Time series data could also be server metrics, application performance monitoring, network data, sensor data, events, clicks and many other types of analytics data.

The 4 major components:

- Trend component.
- Seasonal component.
- Cyclical component.
- Irregular component.

Time series is a machine learning technique that forecasts target value based solely on a known history of target values. It is a specialized form of regression, known in the literature as auto-regressive modelling. The input to time series analysis is a sequence of target values.

Time series are used in statistics, signal processing, pattern recognition, econometrics, mathematical finance, weather forecasting, earthquake prediction, electroencephalography, control engineering, astronomy, communications engineering, and largely in any domain of applied science and engineering which involves...

MACHINE LEARNING OVERVIEW:

Machine learning is a subfield of artificial intelligence, which is broadly defined as the capability of a machine to imitate intelligent human behavior. Artificial intelligence systems are used to perform complex tasks in a way that is similar to how humans solve problems.

Definition of Machine Learning: Arthur Samuel, an early American leader in the field of computer gaming and artificial intelligence, coined the term "Machine Learning" in 1959 while at IBM. He defined machine learning as "the field of study that gives computers the ability to learn without being explicitly programmed ". However, there is no universally accepted definition for machine learning. Different authors define the term differently. We give below two more definitions.

Machine learning is programming computers to optimize a performance criterion using example data or past experience. We have a model defined up to some parameters, and learning is the execution of a computer program to optimize the parameters of the model using the training data or past experience. The model may be predictive to make predictions in the future, or descriptive to gain knowledge from data.

The field of study known as machine learning is concerned with the question of how to construct computer programs that automatically improve with experience.

Machine learning is a subfield of artificial intelligence that involves the development of algorithms and statistical models that enable computers to improve their performance in tasks through experience. These algorithms and models are designed to learn from data and make predictions or decisions without explicit instructions. There are several types of machine learning, including supervised learning, unsupervised learning, and reinforcement learning. Supervised learning involves training a model on labelled data, while unsupervised learning involves training a model on unlabeled data. Reinforcement learning involves training a model through trial and error. Machine learning is used in a wide variety of applications, including image and speech recognition, natural language processing, and recommender systems.

HISTORICAL CONTEXT:

Machine learning, an application of artificial intelligence (AI), has some impressive capabilities. A machine learning algorithm can make software capable of unsupervised learning. Without being explicitly programmed, the algorithm can seemingly grow "smarter," and become more accurate at predicting outcomes, through the input of historical data.

THE HISTORY AND FUTURE OF MACHINE LEARNING:

Machine learning was first conceived from the mathematical modelling of neural networks. A paper by logician Walter Pitts and neuroscientist Warren McCulloch, published in 1943, attempted to mathematically map out thought processes and decision making in human cognition.

In 1950, Alan Turning proposed the Turing Test, which became the litmus test for which machines were deemed "intelligent" or "unintelligent." The criteria for a machine to receive status as an "intelligent" machine, was for it to have the ability to convince a human being that it, the machine, was also a human being. Soon after, a summer research program at Dartmouth College became the official birthplace of AI.

From this point on, "intelligent" machine learning algorithms and computer programs started to appear, doing everything from planning travel routes for salespeople, to playing board games with humans such as checkers and tic-tac-toe.

Intelligent machines went on to do everything from using speech recognition to learning to pronounce words the way a baby would learn to defeat a world chess champion at his own game. The info graphic below shows the history of machine learning and how it grew from mathematical models to sophisticated technology.

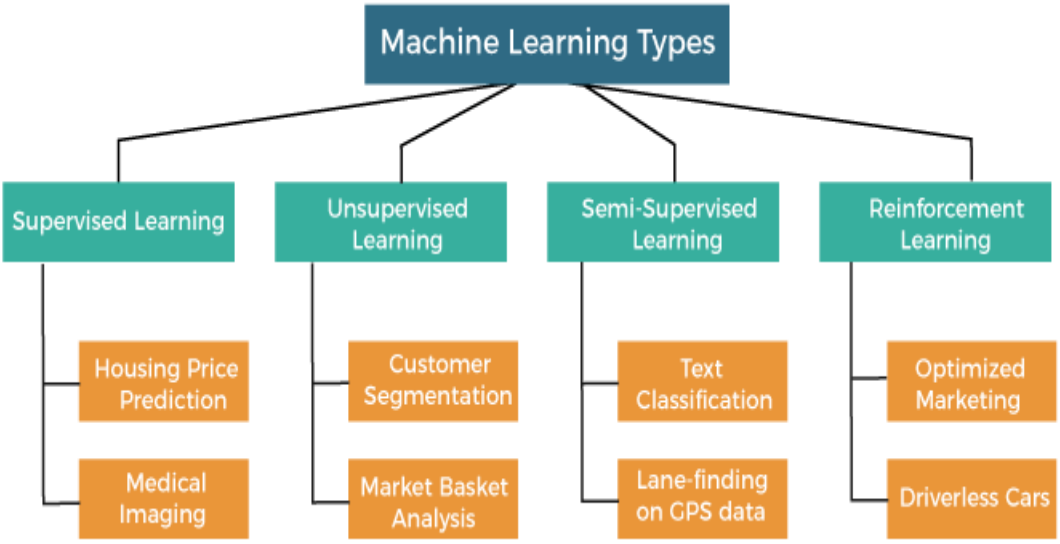
TYPES OF MACHINE LEARNING:

Machine learning is a subset of AI, which enables the machine to automatically learn from data, improve performance from past experiences, and make predictions. Machine learning contains a set of algorithms that work on a huge amount of data. Data is fed to these algorithms to train them, and on the basis of training, they build the model & perform a specific task.

These ML algorithms help to solve different business problems like Regression, Classification, Forecasting, Clustering, and Associations, etc.

Based on the methods and way of learning, machine learning is divided into mainly four types, which are:

1. Supervised Machine Learning
2. Unsupervised Machine Learning
3. Semi-Supervised Machine Learning
4. Reinforcement Learning



In this topic, we will provide a detailed description of the types of Machine Learning along with their respective algorithms:

1. SUPERVISED MACHINE LEARNING:

As its name suggests, supervised machine learning is based on supervision. It means in the supervised learning technique, we train the machines using the "labelled" dataset, and based on the training, the machine predicts the output. Here, the labelled data specifies that some of the inputs are already mapped to the output. More preciously, we can say; first, we train the machine with the input and corresponding output, and then we ask the machine to predict the output using the test dataset.

Let's understand supervised learning with an example. Suppose we have an input dataset of cats and dog images. So, first, we will provide the training to the machine to understand the images, such as the **shape & size of the tail of cat and dog, Shape of eyes, colour, height (dogs are taller, cats are smaller), etc.** After completion of training, we input the picture of a cat and ask the machine to identify the object and predict the output. Now, the machine is well trained, so it will check all the features of the object, such as height, shape, colour, eyes, ears, tail, etc., and find that it's a cat. So, it will put it in the Cat category. This is the process of how the machine identifies the objects in Supervised Learning.

The main goal of the supervised learning technique is to map the input variable(x) with the output variable(y). Some real-world applications of supervised learning are **Risk Assessment, Fraud Detection, Spam filtering,** etc.

Categories of Supervised Machine Learning:

Supervised machine learning can be classified into two types of problems, which are given below:

- **Classification**
- **Regression**

a) Classification:

Classification algorithms are used to solve the classification problems in which the output variable is categorical, such as "**Yes" or No, Male or Female, Red or Blue, etc**. The classification algorithms predict the categories present in the dataset. Some real-world examples of classification algorithms are **Spam Detection, Email filtering, etc.**

Some popular classification algorithms are given below:

- **Random Forest Algorithm**
- **Decision Tree Algorithm**
- **Logistic Regression Algorithm**
- **Support Vector Machine Algorithm**

b) Regression:

Regression algorithms are used to solve regression problems in which there is a linear relationship between input and output variables. These are used to predict continuous output variables, such as market trends, weather prediction, etc.

Some popular Regression algorithms are given below:

- **Simple Linear Regression Algorithm**
- **Multivariate Regression Algorithm**
- **Decision Tree Algorithm**
- **Lasso Regression**

Advantages and Disadvantages of Supervised Learning:

Advantages:

- Since supervised learning work with the labelled dataset so we can have an exact idea about the classes of objects.
- These algorithms are helpful in predicting the output on the basis of prior experience.

Disadvantages:

- These algorithms are not able to solve complex tasks.
- It may predict the wrong output if the test data is different from the training data.
- It requires lots of computational time to train the algorithm.

Applications of Supervised Learning:

Some common applications of Supervised Learning are given below:

- **Image-Segmentation:**
  Supervised Learning algorithms are used in image segmentation. In this process, image classification is performed on different image data with pre-defined labels.
- **Medical-Diagnosis:**
  Supervised algorithms are also used in the medical field for diagnosis purposes. It is done by using medical images and past labelled data with labels for disease conditions. With such a process, the machine can identify a disease for the new patients.
- **Fraud Detection -** Supervised Learning classification algorithms are used for identifying fraud transactions, fraud customers, etc. It is done by using historic data to identify the patterns that can lead to possible fraud.
- **Spam detection -** In spam detection & filtering, classification algorithms are used. These algorithms classify an email as spam or not spam. The spam emails are sent to the spam folder.
- **Speech Recognition -** Supervised learning algorithms are also used in speech recognition. The algorithm is trained with voice data, and various identifications can be done using the same, such as voice-activated passwords, voice commands, etc.

2. UNSUPERVISED MACHINE LEARNING:

Unsupervised learning is different from the supervised learning technique; as its name suggests, there is no need for supervision. It means, in unsupervised machine learning, the machine is trained using the unlabelled dataset, and the machine predicts the output without any supervision.

In unsupervised learning, the models are trained with the data that is neither classified nor labelled, and the model acts on that data without any supervision.

The main aim of the unsupervised learning algorithm is to group or categories the unsorted dataset according to the similarities, patterns, and differences. Machines are instructed to find the hidden patterns from the input dataset.

Let's take an example to understand it more preciously; suppose there is a basket of fruit images, and we input it into the machine learning model. The images are totally unknown to the model, and the task of the machine is to find the patterns and categories of the objects.

So, now the machine will discover its patterns and differences, such as colour difference, shape difference, and predict the output when it is tested with the test dataset.

Categories of Unsupervised Machine Learning:

Unsupervised Learning can be further classified into two types, which are given below:

- **Clustering**
- **Association**

1) Clustering:

The clustering technique is used when we want to find the inherent groups from the data. It is a way to group the objects into a cluster such that the objects with the most similarities remain in one group and have fewer or no similarities with the objects of other groups. An example of the clustering algorithm is grouping the customers by their purchasing behaviour.

Some of the popular clustering algorithms are given below:

- **K-Means Clustering algorithm**
- **Mean-shift algorithm**
- **DBSCAN Algorithm**
- **Principal Component Analysis**
- **Independent Component Analysis**

2) Association:

Association rule learning is an unsupervised learning technique, which finds interesting relations among variables within a large dataset. The main aim of this learning algorithm is to find the dependency of one data item on another data item and map those variables accordingly so that it can generate maximum profit. This algorithm is mainly applied in **Market Basket analysis, Web usage mining, continuous production**, etc.

Some popular algorithms of Association rule learning are **Apriori Algorithm, Eclat, FP-growth algorithm.**

Advantages and Disadvantages of Unsupervised Learning Algorithm:

Advantages:

- These algorithms can be used for complicated tasks compared to the supervised ones because these algorithms work on the unlabeled dataset.
- Unsupervised algorithms are preferable for various tasks as getting the unlabeled dataset is easier as compared to the labelled dataset.

Disadvantages:

- The output of an unsupervised algorithm can be less accurate as the dataset is not labelled, and algorithms are not trained with the exact output in prior.
- Working with Unsupervised learning is more difficult as it works with the unlabeled dataset that does not map with the output.

Applications of Unsupervised Learning:

- **Network Analysis:** Unsupervised learning is used for identifying plagiarism and copyright in document network analysis of text data for scholarly articles.
- **Recommendation Systems:** Recommendation systems widely use unsupervised learning techniques for building recommendation applications for different web applications and e-commerce websites.
- **Anomaly Detection:** Anomaly detection is a popular application of unsupervised learning, which can identify unusual data points within the dataset. It is used to discover fraudulent transactions.
- **Singular Value Decomposition:** Singular Value Decomposition or SVD is used to extract particular information from the database. For example, extracting information of each user located at a particular location.

TRAINING DATA:

The training data is the biggest (in -size) subset of the original dataset, which is used to train or fit the machine learning model. Firstly, the training data is fed to the ML algorithms, which lets them learn how to make predictions for the given task.

For example, for training a sentiment analysis model, the training data could be as below:

Input	Output (Labels)
The New UI is Great	Positive
Update is really Slow	Negative

The training data varies depending on whether we are using Supervised Learning or Unsupervised Learning Algorithms.

For Unsupervised learning, the training data contains unlabeled data points, i.e., inputs are not tagged with the corresponding outputs. Models are required to find the patterns from the given training datasets in order to make predictions.

On the other hand, for supervised learning, the training data contains labels in order to train the model and make predictions.

The type of training data that we provide to the model is highly responsible for the model's accuracy and prediction ability. It means that the better the quality of the training data, the better will be the performance of the model. Training data is approximately more than or equal to 60% of the total data for an ML project.

TESTING DATA:

Once we train the model with the training dataset, it's time to test the model with the test dataset. This dataset evaluates the performance of the model and ensures that the model can generalize well with the new or unseen dataset. The test dataset is another subset of original data, which is independent of the training dataset. However, it has some similar types of features and class probability distribution and uses it as a benchmark for model evaluation once the model training is completed. Test data is a well-organized dataset that contains data for each type of scenario for a given problem that the model would be facing when used in the real world. Usually, the test dataset is approximately 20-25% of the total original data for an ML project.

At this stage, we can also check and compare the testing accuracy with the training accuracy, which means how accurate our model is with the test dataset against the training dataset. If the accuracy of the model on training data is greater than that on testing data, then the model is said to have over fitting.

The testing data should:

- Represent or part of the original dataset.
- It should be large enough to give meaningful predictions.

OVER FITTING IN MACHINE LEARNING:

Over fitting refers to a model that models the training data too well.

Over fitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data. This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model. The problem is that these concepts do not apply to new data and negatively impact the models ability to generalize.

Over fitting is more likely with nonparametric and nonlinear models that have more flexibility when learning a target function. As such, many nonparametric machine learning

algorithms also include parameters or techniques to limit and constrain how much detail the model learns.

For example, decision trees are a nonparametric machine learning algorithm that is very flexible and is subject to over fitting training data. This problem can be addressed by pruning a tree after it has learned in order to remove some of the detail it has picked up.

UNDER FITTING IN MACHINE LEARNING:

Under fitting refers to a model that can neither model the training data nor generalize to new data.

An under fit machine learning model is not a suitable model and will be obvious as it will have poor performance on the training data.

Under fitting is often not discussed as it is easy to detect given a good performance metric. The remedy is to move on and try alternate machine learning algorithms. Nevertheless, it does provide a good contrast to the problem of over fitting.

A GOOD FIT IN MACHINE LEARNING:

Ideally, you want to select a model at the sweet spot between under fitting and over fitting.

This is the goal, but is very difficult to do in practice.

To understand this goal, we can look at the performance of a machine learning algorithm over time as it is learning a training data. We can plot both the skill on the training data and the skill on a test dataset we have held back from the training process.

Over time, as the algorithm learns, the error for the model on the training data goes down and so does the error on the test dataset. If we train for too long, the performance on the training dataset may continue to decrease because the model is over fitting and learning the irrelevant detail and noise in the training dataset. At the same time the error for the test set starts to rise again as the model's ability to generalize decreases.

The sweet spot is the point just before the error on the test dataset starts to increase where the model has good skill on both the training dataset and the unseen test dataset.

You can perform this experiment with your favourite machine learning algorithms. This is often not useful technique in practice, because by choosing the stopping point for training using the skill on the test dataset it means that the test set is no longer "unseen" or a standalone

objective measure. Some knowledge (a lot of useful knowledge) about that data has leaked into the training procedure.

There are two additional techniques you can use to help find the sweet spot in practice: resampling methods and a validation dataset.

MACHINE LEARNING CLASSIFICATION:

Machine learning, classification is a predictive modelling problem where THE class label is anticipated for a specific example of input data. For example, in determining handwriting characters, identifying spam, and so on, the classification requires training data with a large number of datasets of input and output.

Classification is a natural language processing task that depends on machine learning algorithms. There are many different types of classification tasks that you can perform, the most popular being sentiment analysis.

CLASSIFIERS:

In machine learning, a classifier is an algorithm that automatically assigns data points to a range of categories or classes. Within the classifier category, there are two main models: supervised and unsupervised. In the supervised model, classifiers train to make distinctions between labelled and unlabeled data. This training allows them to recognize patterns and ultimately operate autonomously without using labels. Unsupervised algorithms use pattern recognition to classify unlabeled datasets, progressively becoming more accurate.

AI applications are becoming an increasingly vital part of business operations, and classification algorithms are important because they're an integral part of these platforms. Many businesses rely on extensive data collection operations to enhance their processes, and it can be challenging to collect and analyze data effectively on this scale. AI tools with classification functionality make this process simpler by automating the process of analysis and classification. This reduces work for employees and allows companies to expand their data operations without straining resources or losing productivity.

There are a wide variety of tasks that classifiers can complete. These are some examples of how a company might use AI classification:

- To separate important emails from spam
- To separate customer complaints from other comments

- To locate named entities on the internet
- To extract contact information for marketing, recruiting and sales
- To assign customers to different market segments for targeted marketing
- To identify fraudulent financial transactions

5 TYPES OF CLASSIFIERS IN MACHINE LEARNING:

There are a wide variety of classification algorithms used in AI and each one uses a different mechanism to analyze data. These are five common types of classification algorithms:

1. NAIVE BAYES CLASSIFIER:

Naive Bayes classifiers use probability to predict whether an input will fit into a certain category. The Naive Bayes algorithm family includes a range of different classifiers based on a theorem of probability. These classifiers can determine the probability of an input fitting into one or more categories.

In multiple category scenarios, the algorithm reviews the probability that a data point fits into each classification. After comparing the probability of a match in each category, it outputs the category that is most likely to match the given text. Many companies use this type of algorithm to assign tags to text segments like email subject lines, customer comments and articles.

2. DECISION TREE:

A decision tree is a classification algorithm that uses a process of division to split data into increasingly specific categories. It's called a decision tree because the classification process resembles a tree's branches when represented graphically. The algorithm works on a supervised model and requires high-quality data to produce good results.

Since the primary goal of a decision tree is to make increasingly specific distinctions, it has to continuously learn new classification rules. It learns these rules by applying if-then logic to training data. The algorithm continues the classification process until it reaches a designated stopping condition.

3. ARTIFICIAL NEURAL NETWORKS:

Artificial neural networks (ANNs) are computing frameworks made up of many individual algorithms. Their mechanism of action mimics how human brains work, and

includes a collection of artificial neurons that transmit signals. This makes artificial neural networks capable of solving extremely complex problems that involve multiple layers. Because of their complexity, it can be challenging to train and adjust ANNs, and it often requires large amounts of training data. However, a fully trained ANN can perform tasks that would be impossible for single algorithms.

There are many types of artificial neural networks, including:

- Feed forward neural network
- Feedback neural network
- Recurrent neural network
- Classification-prediction network
- Radial basis function network
- Dynamic neural network
- Modular neural network

4. SUPPORT VECTOR MACHINE:

A support vector machine (SVM) is a simple algorithm that professionals can use for classification or regression activities. They work by finding hyper planes within a data distribution, which you can visualize as a line separating two different classes of data. There are often many hyper planes capable of separating the data, and the algorithm will select the optimum line of separation. In the SVM model, the optimum hyper plane is the dividing line that offers the greatest margin between the different classes.

SVMs are capable of working in multiple dimensions if they are unable to find an ideal hyper plane to separate the data into two dimensions. This makes them extremely effective for creating classifications from complicated data distributions. The more complex the data inputs are, the more accurate the SVM becomes, making them excellent machine learning tools.

5. K-NEAREST NEIGHBOUR:

K-nearest neighbor (KNN) is a supervised lazy learner algorithm used in machine learning. This means that it stores the training data that supervisors present and compares it to other data to make predictions. While the training period for these algorithms is often shorter than for "eager learners," they're often slower to make predictions.

After storing its training data, a KNN algorithm compares it with test data and measures the degree of similarity between them. It then stores all instances that correspond with the

training data. Next, the algorithm attempts to predict the likelihood that future data will correspond to the dataset it compiled. While this algorithm is common in classification, many professionals also use it to complete regression tasks.