**0Data visualization**

* Data visualization is the graphical representation of information and data.
* Include a variety of visual tools such as:
* **Charts**: Bar charts, line charts, pie charts, etc.
* **Graphs**: Scatter plots, histograms, etc.
* **Maps**: Geographic maps, heat maps, etc.
* **Dashboards**: Interactive platforms that combine multiple visualizations.
* Data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.
* Successful visualization requires that the data (information) be converted into a visual format so that the characteristics of the data and the relationships among data items or attributes can be analyzed or reported.
* The **goal** of visualization is the interpretation of the visualized information by a person and the formation of a mental model of the information.

**Visual data mining (VDM)**

* **Visual data mining** (VDM) is a process that combines data mining techniques with data visualization to discover patterns, relationships, and insights from large and complex datasets.
* By leveraging the human visual system's ability to recognize patterns and anomalies, visual data mining makes it easier to interpret and understand the results of data mining processes.

**General Concepts**

**Representation: Mapping Data to Graphical Elements**

**Object Representation**

The process of data visualization involves translating data objects, their attributes, and relationships into graphical elements such as points, lines, shapes, and colors.

Objects representation can be categorized in three main ways:

1.Single Categorical Attribute: Objects are grouped into categories based on a single attribute and displayed in a table or chart (e.g., bar chart).

2. Multiple Attributes: Objects are shown as rows/columns in a table or as lines in a graph.

3. Point in Space: Objects are represented as points in 2D or 3D space, using geometric figures.

**Attributes Representation:**

- Ordinal and Continuous Attributes: Mapped to continuous, ordered graphical features like axis location, intensity, color, or size.

- Categorical Attributes: Each category is represented by distinct features like position, color, or shape.

**Relationships Representation:**

- Explicit Representation: Common in graph data where nodes and links are used (e.g., cities and highways).

- Implicit Representation: Mapping objects and attributes often inherently maps relationships. For example, relative positions of cities can reflect actual geographic locations.

* Ensuring that the visualization effectively displays relationships among data elements is challenging, as implicit relationships may not always be easily observable.
* The goal of visualization is to choose techniques that highlight these relationships clearly.

**Arrangement**

The proper choice of visual representation of objects and attributes is essential for good visualization.

The arrangement of items within the visual display is also crucial.

Eg 1:

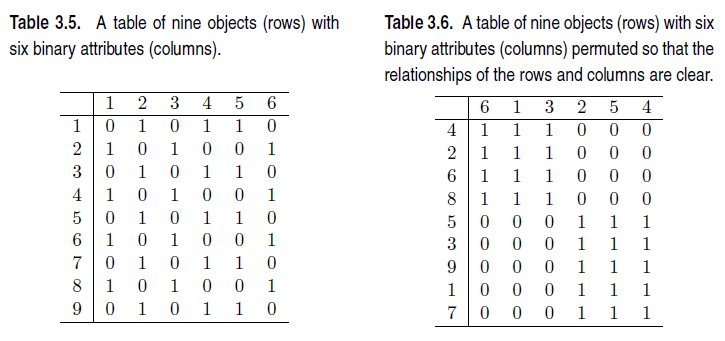


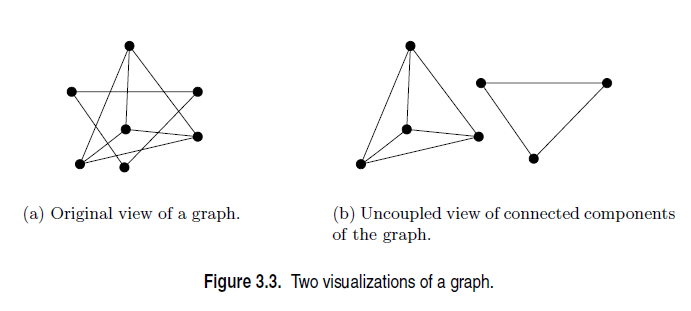
Table 3.5: There is no clear relationship between objects and attributes.

If the rows and columns of this table are permuted, however, as shown in Table 3.6, then it is clear that there are really only two types of objects in the table—one that has all ones for the first three attributes and one that has only ones for the last three attributes.

**Eg 2:**

Consider Figure 3.3(a), which shows a visualization of a graph.

If the connected components of the graph are separated, as in Figure 3.3(b), then the relationships between nodes and graphs become much simpler to understand.



**Selection**

* **Selection** involves reducing or de-emphasizing certain objects and attributes to manage complexity.
* For data with few dimensions, straightforward two- or three-dimensional representations are feasible.
* However, for data with many attributes or objects, visualizing everything at once can result in overcrowded displays.

### Strategies for Handling Many Attributes:

1. **Subset Selection**: Display a subset of attributes, typically two at a time.
2. **Matrix of Bivariate Plots**: Construct a matrix of scatter plots for all pairs of attributes, allowing simultaneous viewing (e.g., Iris data set scatter plot matrix).
3. **Sequential 2D Plots**: Use a visualization program to automatically show a series of two-dimensional plots, directed by the user or a predefined strategy, to provide a comprehensive view of the data.

These approaches aim to simplify visualization and make the data more comprehensible by focusing on smaller, more manageable subsets.

**Techniques**

Visualization techniques are often tailored to the specific type of data being analyzed, with new methods continuously developed to address emerging data types and tasks.

Despite this specialization, visualization techniques can be generically classified in several ways:

1. Number of Attributes: Techniques vary based on whether they involve 1, 2, 3, or many attributes.

2. Data Characteristics: Special characteristics, such as hierarchical or graph structures, influence visualization methods.

3. Type of Attributes: Techniques differ based on the nature of the attributes involved.

4. Application Type: Visualization can be categorized into scientific, statistical, or information visualization.

**Three main categories:**

1. Visualization of a small number of attributes.

2. Visualization of data with spatial and/or temporal attributes.

3. Visualization of data with many attributes.

**Visualization of a small number of attributes**

**Stem and Leaf plot**

A **stem and leaf plot** is a method of organizing numerical data based on the place value of the numbers.

Each number is split into two parts:

The first digit(s) form the **stem**.

The last digit forms the **leaf**.

The number of digits for the stem can vary, but the number of digits for the leaf should only ever contain a single digit.

For example, if the numerical data contains all whole numbers, the numbers in the tens place and greater would make up the stem values while all numbers in the ones place would make up the leaf values.

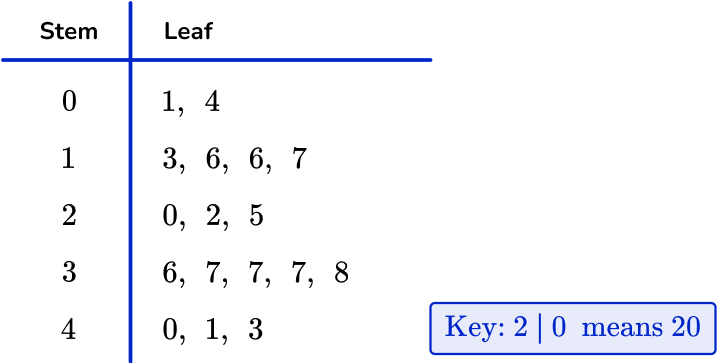
Each piece of numerical data will be written with the last digit to the right of a vertical line and all other digits to the left of the vertical line. For example, the number 154 would be split into two parts where the digits **15** would be the **stem** and **4** would be the **leaf.**



To set up a stem and leaf plot you need to:

* Organize the data into **ascending order**, smallest to largest.
* Determine how the numbers are split into 2 parts by writing a **key** for the stem and leaf plot.
* Write the values for the ‘stem’ into the stem and leaf plot.
* Write the values for the ‘leaf’ into the stem and leaf plot.

**NOTE:** A stem and leaf plot must have a **key** (sometimes referred to as a **legend**). This explains how to convert the digits in the stem and leaf plot into a single data point.



You can use stem and leaf plots to calculate averages like the median, the mode and the mean, and to calculate measures of spread like the range and the interquartile range.

**Histograms**

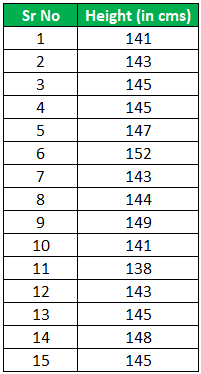
* Histogram is a graphical representation that condenses data series into easy-to-understand numerical data by grouping them into logical ranges of varying heights, often known as **bins.**
* It summarises discrete or continuous data.
* Histogram is a tool for visualising the distribution of data across a continuous interval or period.
* In a histogram data is grouped into continuous number ranges and each range corresponds to a vertical bar.
* **Horizontal axis** displays the number range.
* **Vertical axis** (frequency) represents the amount of data present in each range.
* It allows us to assess where the values are concentrated, what the extremes are, and whether there are any gaps or anomalous values.
* A Histogram is similar to a vertical bar graph, however, the distinction is that the Histogram has no space between the bars, but a bar graph has.

Differences between the bar graph and the histogram is given below:

|  |  |
| --- | --- |
| **Histogram** | **Bar Graph** |
| It is a two-dimensional figure | It is a one-dimensional figure |
| The frequency is shown by the area of each rectangle | The height shows the frequency and the width has no significance. |
| It shows rectangles touching each other | It consists of rectangles separated from each other with equal spaces. |

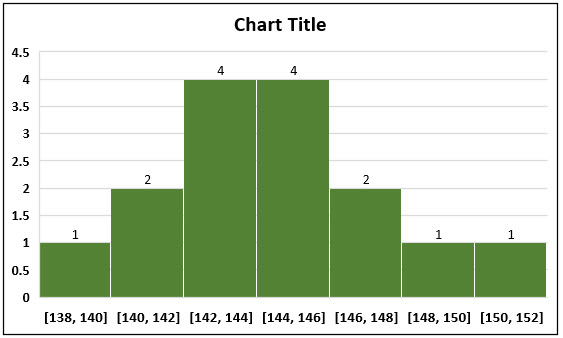
#### Histogram Example

Mr. Larry, a famous doctor, is researching the height of the students studying in the 8th standard. He has gathered 15 students but wants to know which maximum category is where they belong.



**Solution:**

We have created a histogram using 6 bins with 6 different frequencies, as seen in the chart below. The Y-axis is the average number of students falling in a particular category. In addition, on the X-axis, we have a range of heights. For example, the 1st bin range is 138 cms to 140 cms. We can note that the count is 1 for that category from the table, as seen in the below graph.



Here, we can see the students’ average heights range from 142 cm to 146 cm for the 8th standard. And also, one can note that one side of the average also falls on the other side of the average, which is a sign of [**normal distribution**](https://www.wallstreetmojo.com/normal-distribution/).

**Box Plots**

* Box plots are another method for showing the distribution of the values of a single numerical attribute.
* Box plot is a type of chart that depicts a group of numerical data through their quartiles.
* Box plot is also known as a whisker plot, box-and-whisker plot, or simply a box-and whisker diagram.
* Box plot is a graphical representation of the distribution of a dataset.
* It displays key summary statistics such as the median, quartiles, and potential outliers in a concise and visual manner.
* By using Box plot you can provide a summary of the distribution, identify potential and compare different datasets in a compact and visual manner.

A box plot gives a five-number summary of a set of data which is-

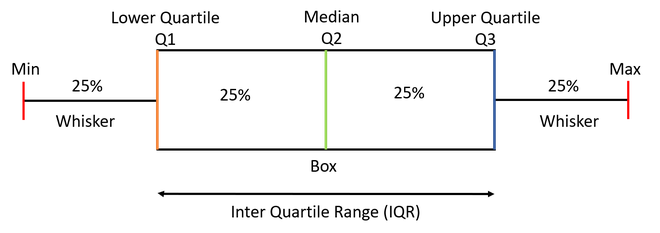
**Minimum** – It is the minimum value in the dataset excluding the outliers.

**First Quartile (Q1**) – 25% of the data lies below the First (lower) Quartile.

**Median (Q2)** – It is the mid-point of the dataset. Half of the values lie below it and half above.

**Third Quartile (Q3)** – 75% of the data lies below the Third (Upper) Quartile.

**Maximum** – It is the maximum value in the dataset excluding the outliers.



The area inside the box (50% of the data) is known as the Inter Quartile Range. The IQR is calculated as –

**IQR = Q3-Q1**

Outlies

are the data points below and above the lower and upper limit. The lower and upper limit is calculated as –

**Lower Limit = Q1 - 1.5\*IQR**

**Upper Limit = Q3 + 1.5\*IQR**

The values below and above these limits are considered outliers and the minimum and maximum values are calculated from the points which lie under the lower and upper limit.

How to create a box plots?

Here are the runs scored by a cricket team in a league of 12 matches – 100, 120, 110, 150, 110, 140, 130, 170, 120, 220, 140, 110.

To draw a box plot for the given data first we need to arrange the data in ascending order and then find the minimum, first quartile, median, third quartile and the maximum.

Ascending Order

100, 110, 110, 110, 120, 120, 130, 140, 140, 150, 170, 220

Median (Q2) = (120+130)/2 = 125; Since there were even values

To find the First Quartile we take the first six values and find their median.

Q1 = (110+110)/2 = 110

For the Third Quartile, we take the next six and find their median.

Q3 = (140+150)/2 = 145

Note: If the total number of values is odd then we exclude the Median while calculating Q1 and Q3. Here since there were two central values we included them. Now, we need to calculate the Inter Quartile Range.

IQR = Q3-Q1 = 145-110 = 35

We can now calculate the Upper and Lower Limits to find the minimum and maximum values and also the outliers if any.

Lower Limit = Q1-1.5\*IQR = 110-1.5\*35 = 57.5

Upper Limit = Q3+1.5\*IQR = 145+1.5\*35 = 197.5

So, the minimum and maximum between the range [57.5,197.5] for our given data are –

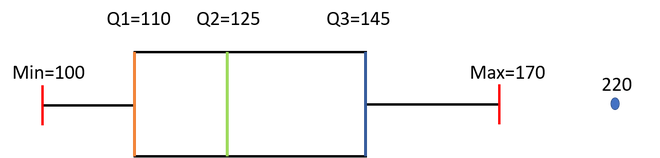
Minimum = 100

Maximum = 170

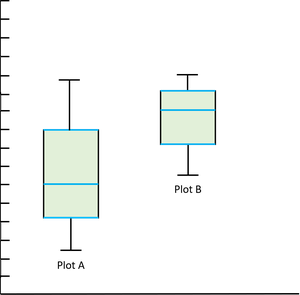
The outliers which are outside this range are –

Outliers = 220

We can draw the box plot which is as below:



## **How to compare box plots?**



* **Compare the Medians —**If the median line of a box plot lies outside the box of the other box plot with which it is being compared, then we can say that there is likely to be a difference between the two groups. Here the Median line of the plot B lies outside the box of Plot A.
* **Compare the Dispersion or Spread of data —**The Inter Quartile range (length of the box) gives us an idea about how dispersed the data is. Here Plot A has a longer length than Plot B which means that the dispersion of data is more in plot A as compared to plot B. The length of whiskers also gives an idea of the overall spread of data. The extreme values (minimum &maximum) give the range of data distribution. Larger the range more scattered the data. Here Plot A has a larger range than Plot B.
* **Comparing Outliers —**The outliers give the idea of unusual data values which are distant from the rest of the data. More number of Outliers means the prediction will be more uncertain. We can be more confident while predicting the values for a plot which has less or no outliers.
* **Compare Skewness —**[Skewness](https://www.geeksforgeeks.org/skewness-measures-and-interpretation/) gives us the direction and the magnitude of the lack of symmetry. Here Plot A is Positive or Right Skewed and Plot B is Negative or Left Skewed.

# **Dot Plot**

A dot plot is used to encode data in a dot or small circle.

The dot plot is shown on a number line that displays the distribution of numerical variables where a value is defined by each dot.

## **What is Dot Plot?**

A dot plot is used to represent any [data](https://www.cuemath.com/data/) in the form of dots or small circles. It is similar to a simplified [histogram](https://www.cuemath.com/data/histograms/) or a [bar graph](https://www.cuemath.com/data/bar-graphs/) as the height of the bar formed with dots represents the numerical value of each variable.

Dot plots are used to represent small amounts of data.

For **example**, a dot plot can be used to collect the vaccination report of newborns in an area, which is represented in the following table.

## **Types of Dot Plot**

There are two types of dot plot: Wilkinson dot plot and Cleveland dot plot.

### Wilkinson Dot Plot

### Cleveland Dot Plot

| **Colony** | **A** | **B** | **C** | **D** |
| --- | --- | --- | --- | --- |
| Number of babies vaccinated | 7 | 3 | 5 | 1 |

### dot plot data Wilkinson Dot Plot

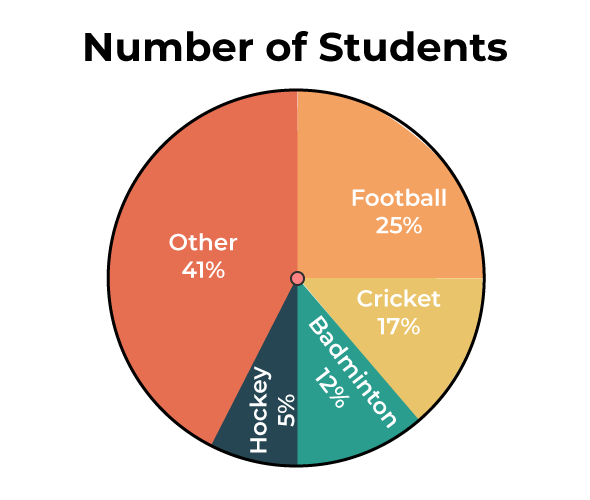
### Cleveland dot plot graphCleveland Dot Plot

**Pie Chart**

* A pie chart is one of the types of charts in which the data is represented in a circular shape.
* The pie chart circle is further divided into multiple sectors/slices; those sectors show the different parts of the data from the whole.
* i.e; A pie chart uses the relative area of a circle to indicate relative frequency.
* Pie charts, also known as circle graphs or pie diagrams, are very useful in representing and interpreting data.
* The data can be compared easily with the help of a pie chart.
* Pie charts are used less frequently in technical publications because the size of relative areas can be hard to judge.

**Pie Chart Examples**

. In a class of 200 students, a survey was done to collect each student’s favorite sports. The pie chart of the data is given below:



Since the pie chart is provided and the total number of students is given, we can easily take the original data out for each sport.

* **Cricket = 17/100 × 200 = 34 students**
* **Football = 25/100 × 200 = 50 students**
* **Badminton = 12/100 × 200 = 24 students**
* **Hockey = 5/100 × 200 = 10 students**
* **Other = 41/100 × 200 = 82 students**

The original data for the pie chart shown above is given below:

| **Sport** | **Number of Students** |
| --- | --- |
| **Cricket** | **34** |
| **Football** | **50** |
| **Badminton** | **24** |
| **Hockey** | **10** |
| **Other** | **82** |

**Pie Chart Formula**

* The total value or percentage of the pie is 100% always.
* Here it contains different sectors and segments in which each sector or segment of the chart corresponds to a certain portion of the net or total percentage (or data).
* **The total or sum of all the data can be summed up to 360 degrees.**

Converting the data into degrees on a pie chart. The formula for a pie chart can be summed up as:

***(Given Data / Total Value of Data) × 360°***

Calculating the percentage of each sector from degrees in a pie chart.

**To work out with degrees in a pie chart, we need to follow the following steps:**

* First, we need to measure every slice of the chart.
* Then we need to divide it by 360°.
* Finally, multiply the obtained result by 100.

**The pie chart formula is given below:**

***(Frequency)/(Total Frequency) × 100***

**Calculating Number of Sectors on a Pie Chart**

To calculate the total number of slices or sectors on a pie chart, we need to multiply the sector’s percentage by the total value of the data and finally divide the result by 100.

**How to Make Pie Chart**

Eg:

A teacher surveyed a group of students to see what their favorite hobby of each student is. Let’s take a look at the pie chart example with an explanation. The data collected is listed as follows:

| **Hobbies** | **Number of students** |
| --- | --- |
| Singing | 16 |
| Reading books | 20 |
| Dancing | 10 |
| Painting | 30 |
| Others | 24 |

Now we will see how to construct a pie chart step by step.

**Step 1: The first step requires us to write down the available data in tabular form as follows:**

| **Singing** | **Reading Books** | **Dancing** | **Painting** | **Others** |
| --- | --- | --- | --- | --- |
| 16 | 20 | 10 | 30 | 24 |

**Step 2: Now find the sum of all the given data. Here, the Sum of All Data = (16 + 20 + 10 + 30 + 24) = 100**

**Step 3:** **Now, calculate the percentage of each sector. We need to divide each sector value by the sum or total and then multiply it by 100.**

| **Singing** | **Reading Books** | **Dancing** | **Painting** | **Others** |
| --- | --- | --- | --- | --- |
| (16/100) × 100  = 16% | (20/100) × 100  = 20% | (10/100) × 100  = 10% | (30/100) × 100  = 30% | (24/100) × 100  = 24% |

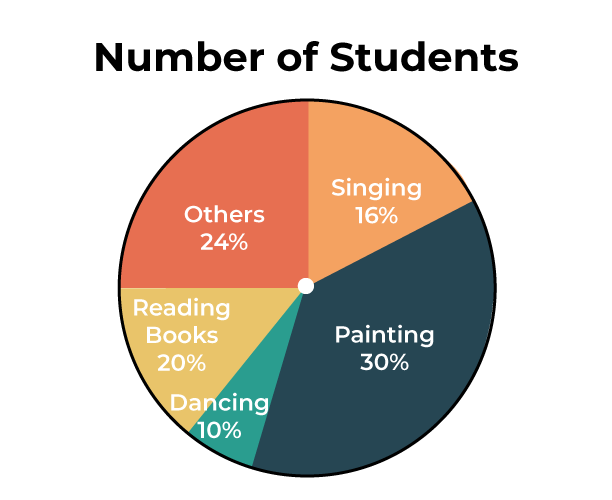
**Step 4: Next step is to calculate the degrees corresponding to each slice. The values can be calculated as:**

***Central Angle of Each Component = (Given Data / Total Value of Data) × 360***

**Hence, The values are as follows:**

| **Singing** | **Reading Books** | **Dancing** | **Painting** | **Others** |
| --- | --- | --- | --- | --- |
| (16/100) × 360  = 57.6 | (20/100) × 360  = 72 | (10/100) × 360  = 36 | (30/100) × 360  = 108 | (24/100) × 360  = 86.4 |

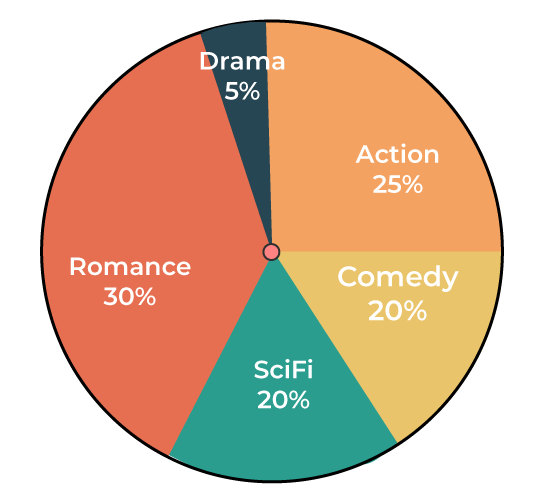
**Step 5:** **Now, with the help of a protractor, we will measure each angle from a single point or central point and draw the circle’s sectors. The resultant pie chart will be:**



**How to Read Pie Chart**

In order to read a pie chart, the first thing to notice is the data presented in the pie chart. If the data is given in percentage, it should be converted accordingly in order to analyze and interpret the data

**Example: In a survey done among 300 people, it was observed which type of genre each person prefers. The pie chart of the same is mentioned below. Analyze and interpret the pie chart accordingly to find the original data.**



**Solution:**

*While observing the pie chart, it came to notice that the data is present in percentage. Let’s convert the data to obtain the original value.*

* *Number of people who like comedy = 20/100 × 300 = 60 people.*
* *Number of people who like action = 25/100 × 300 = 75 people.*
* *Number of people who like romance = 30/100 × 300 = 90 people.*
* *Number of people who like drama = 5/100 × 300 = 15 people.*
* *Number of people who like sci-fi = 20/100 × 300 = 60 people.*

**Pie Chart Vs Bar Graph**

**The key difference between pie chart and bar graph are listed in the following table:**

| **Aspect** | **Pie Chart** | **Bar Graph** |
| --- | --- | --- |
| **Representation** | Circular display of data | rectangular display of data |
| **Purpose** | Shows parts of a whole | Compares discrete categories |
| **Data presentation** | Depicts percentages or proportions | Shows exact values or quantities. |
| **Number of variables** | Typically one variable | Can represent multiple variables. |
| **Visualization** | Easily shows relative proportions | Effective for comparing quantities. |
| **Comparison** | Might be difficult to compare precise values | Allows for easy comparison between categories. |
| **Data complexity** | Works well with simple datasets | Suitable for complex datasets |
| **Interpretation** | Provides a holistic view | Allows for detailed analysis. |
| **Space efficiency** | Not efficient with large datasets | Efficient for displaying large datasets |

**Pie Chart Advantages**

Pie Chart is very useful for finding and representing data. Various advantages of the pie chart are,

* Pie chart is easily understood and comprehended.
* Visual representation of data in a pie chart is done as a fractional part of a whole.
* Pie chart provides an effective mode of communication to all types of audiences.
* Pie chart provides a better comparison of data for the audience.

**Pie Chart Disadvantages**

There are some disadvantages also of using pie charts and some of them are added below,

* In the case of too much data, this presentation becomes less effective using a pie chart.
* For multiple data sets, we need a series to compare them.
* For analyzing and Assimilating the data in a pie chart, it is difficult for readers to comprehend.

**Uses of Pie Chart**

Whenever a fraction or fractions are represented as a part of the whole, pie charts are used. Pie charts are used to compare the data and to analyze which data is bigger or smaller. Hence, while dealing with discrete data, pie charts are preferred. Let’s take a look at the uses of the pie chart:

* Pie charts are used to compare the profit and loss in businesses.
* In schools, the grades can be easily compared using a pie chart.
* The relative sizes of data can be compared using a pie chart.
* The marketing and sales data can be compared using a pie chart.

**Scatter Plots**

* Scatter plots are the graphs that present the relationship between **two** variables in a data-set.
* It represents data points on a two-dimensional plane or on a Cartesian system. Each data object is plotted as a point in the plane using the values of the two attributes as *x* and *y* coordinates.
* The independent variable or attribute is plotted on the X-axis, while the dependent variable is plotted on the Y-axis.
* It is assumed that the attributes are either integer- or real-valued.
* These plots are often called scatter graphs, scatter chart, scattergram or scatter diagrams.

**When to use a scatter plot**?

Scatter plots are used in either of the following situations.

* When we have paired numerical data
* When there are multiple values of the dependent variable for a unique value of an independent variable
* In determining the relationship between variables in some scenarios, such as identifying potential root causes of problems, checking whether two products that appear to be related both occur with the exact cause and so on.

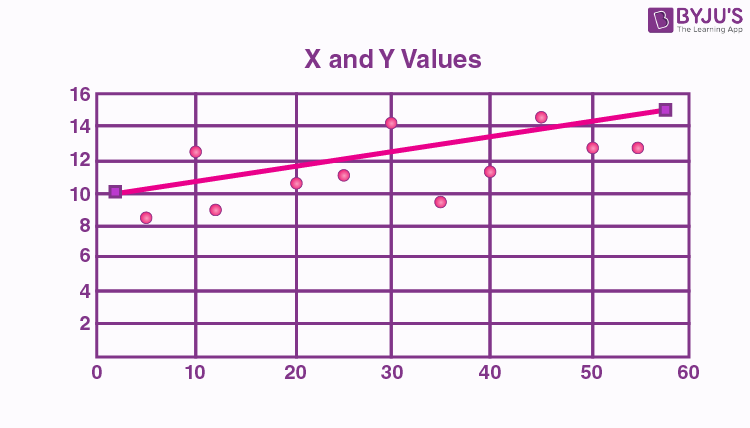
**Scatter Plot Uses and Examples**

Scatter plots instantly report a large volume of data. It is beneficial in the following situations –

* For a large set of data points given
* Each set comprises a pair of values
* The given data is in numeric form



The line drawn in a scatter plot, which is near to almost all the points in the plot is known as “**line of best fit**” or “**trend line**“. See the graph below for an example.



**Scatter plot Correlation**

Correlation is a statistical measure of the relationship between the two variables’ relative movements.

If the variables are correlated, the points will fall along a line or curve.

The better the correlation, the closer the points will touch the line.

**Types of correlation**

The scatter plot explains the correlation between two attributes or variables.

It represents how closely the two variables are connected.

There can be three such situations to see the relation between the two variables –

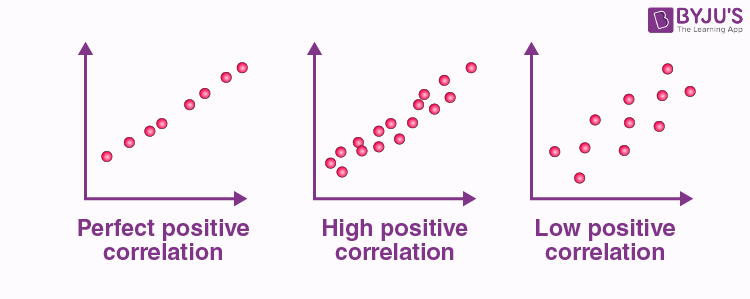
1. Positive Correlation
2. Negative Correlation
3. No Correlation

**Positive Correlation**

When the points in the graph are rising, moving from left to right, then the scatter plot shows a positive correlation. It means the values of one variable are increasing with respect to another.

Positive correlation can further be classified into three categories:

* **Perfect Positive**– Which represents a perfectly straight line
* **High Positive** – All points are nearby
* **Low Positive** – When all the points are scattered



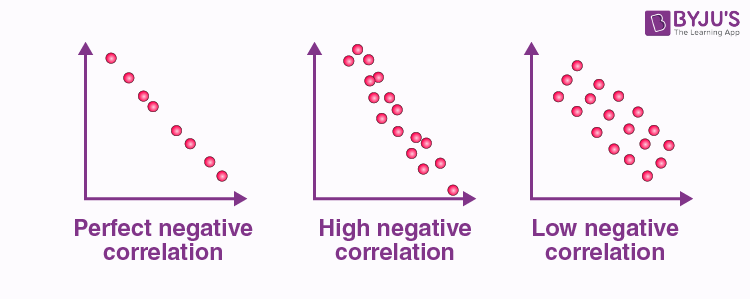
**Negative Correlation**

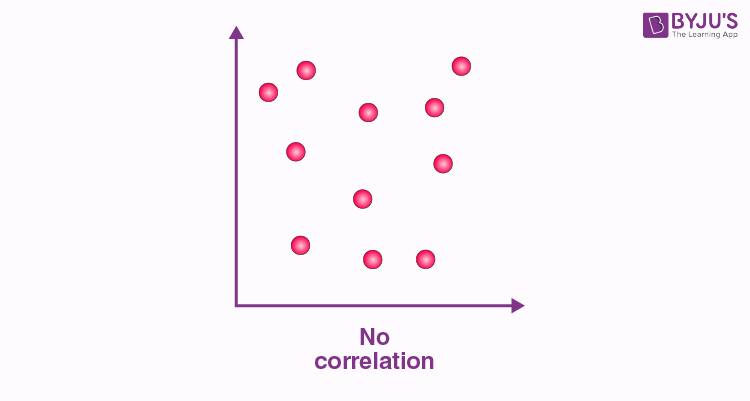
When the points in the scatter graph fall while moving left to right, then it is called a negative correlation.

It means the values of one variable are decreasing with respect to another.

These are also of three types:

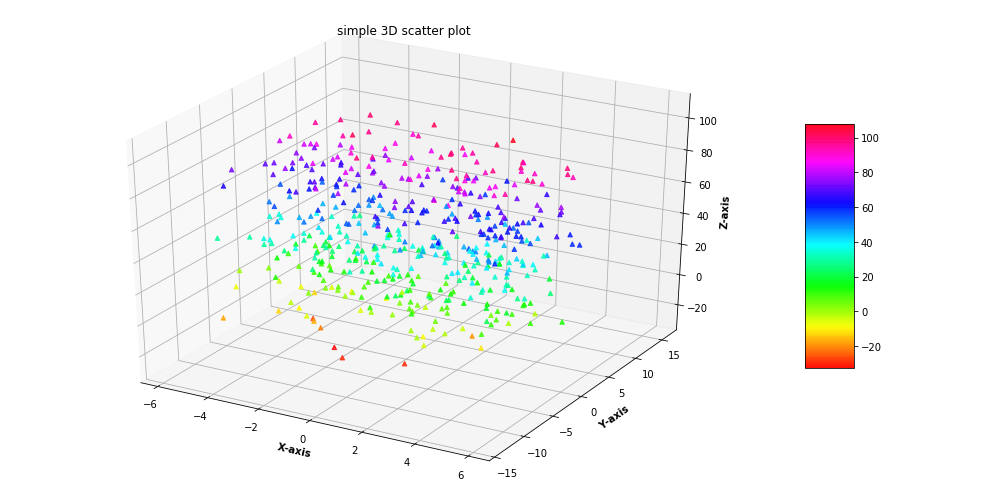
* **Perfect Negative** – Which form almost a straight line
* **High Negative** – When points are near to one another
* **Low Negative** – When points are in scattered form



**No Correlation**

When the points are scattered all over the graph and it is difficult to conclude whether the values are increasing or decreasing, then there is no correlation between the variables.

* A 3-D scatter plot uses three axes in a Cartesian coordinate system.
* If it also uses color, it can display up to 4-D data points.
* For data sets with more than four dimensions, scatter plots are usually ineffective. The scatter-plot matrix becomes less effective as the dimensionality increases.



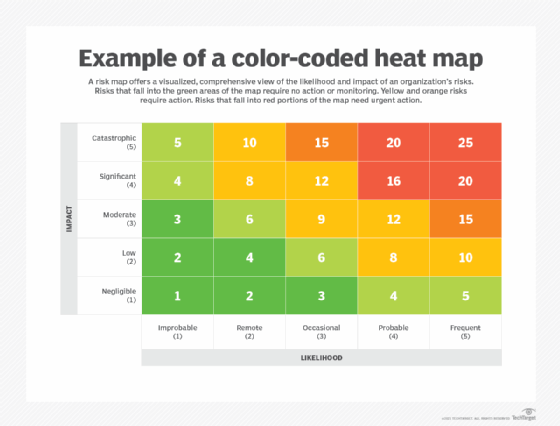
* The scatter-plot matrix technique is a useful extension to the scatterplot. For an n-dimensional dataset, a scatter-plot matrix is an n × n grid of 2-D scatter plots that provides a visualization of each dimension with every other dimension.

## **Heat map**

A heat map is a two-dimensional representation of [data](https://www.techtarget.com/searchdatamanagement/definition/data) in which various values are represented by colors.

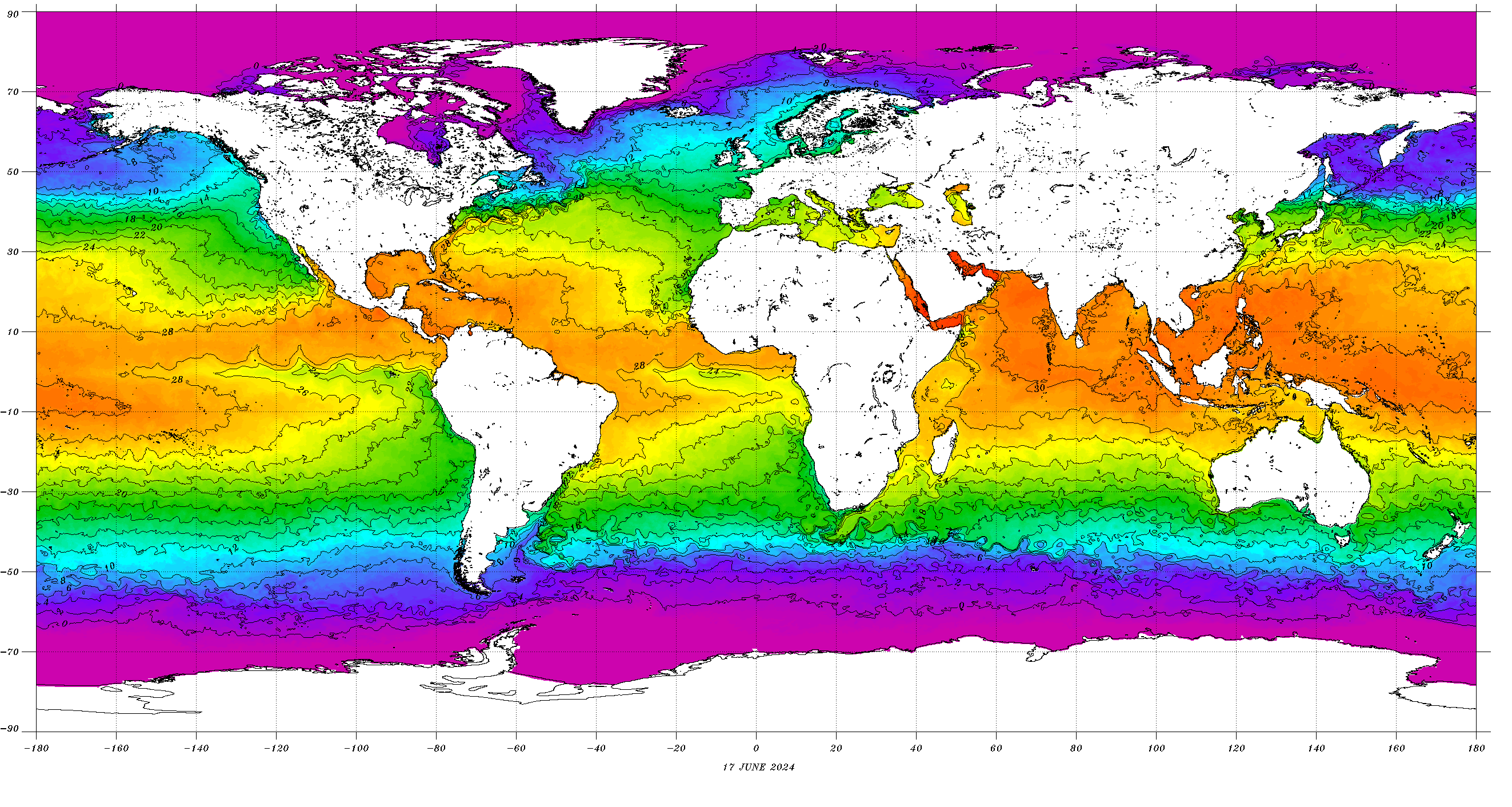
A simple heat map provides an immediate visual summary of information across two axes, allowing users to quickly grasp the most important or relevant [data points](https://www.techtarget.com/whatis/definition/data-point). More elaborate heat maps allow the viewer to understand complex [data sets](https://www.techtarget.com/whatis/definition/data-set).

* A heat map is a way to represent data points in a data set in a visual manner.
* All heat maps share one thing in common -- they use different colors or different shades of the same color to represent different values and to communicate the relationships that may exist between the variables plotted on the [x-axis and y-axis](https://www.techtarget.com/whatis/definition/x-and-y-coordinates).
* Usually, a darker color or shade represents a higher or greater quantity of the value being represented in the heat map.

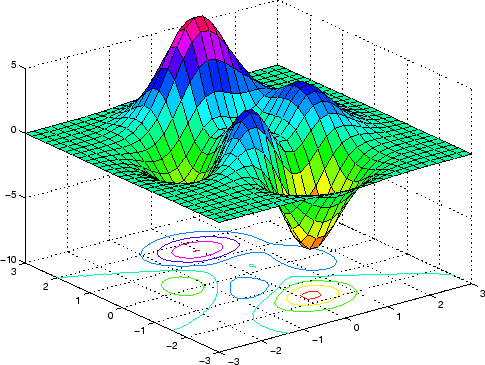


**Contour Plots**

* For some three-dimensional data, two attributes specify a position in a plane, while the third has a continuous value, such as temperature or elevation.
* A useful visualization for such data is a **contour plot**, which breaks the plane into separate regions where the values of the third attribute (temperature, elevation) are roughly the same.
* A common example of a contour plot is a contour map that shows the elevation of land locations.



**Surface Plots**

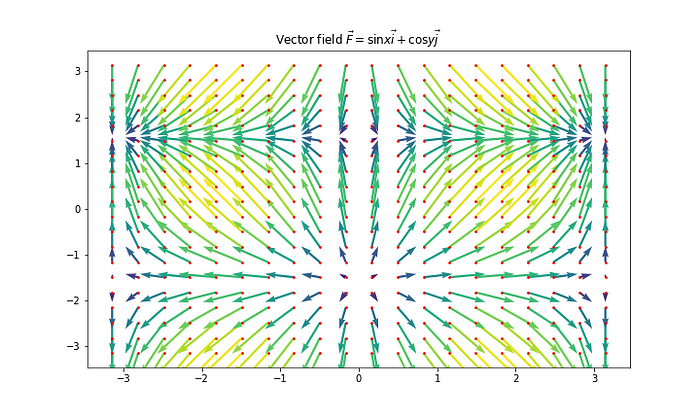
* Like contour plots, **surface plots** use two attributes for the *x* and *y* coordinates.
* The third attribute is used to indicate the height above the plane defined by the first two attributes.
* While such graphs can be useful, they require that a value of the third attribute be defined for all combinations of values for the first two attributes, at least over some range.
* Also, if the surface is too irregular, then it can be difficult to see all the information, unless the plot is viewed interactively.
* Thus, surface plots are often used to describe mathematical functions or physical surfaces that vary in a relatively smooth manner.

**Vector Field Plots**

In some data, a characteristic may have both a magnitude and a direction associated with it.

For example, consider the flow of a substance or the change of density with location.

In these situations, it can be useful to have a plot that displays both direction and magnitude. This type of plot is known as a **vector plot**.



**Visualizing Higher-Dimensional Data**

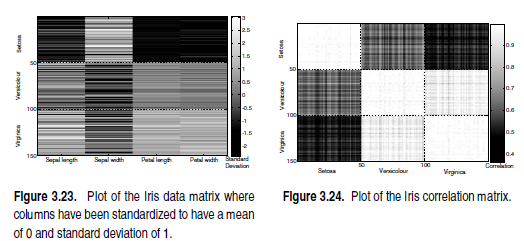
<https://www.kaggle.com/datasets/uciml/iris>

Visualization techniques that can display more than the handful of dimensions that can be observed.

However, even these techniques are somewhat limited in that they only show some aspects of the data.

**Matrices**

* An image can be regarded as a rectangular array of pixels, where each pixel is characterized by its color and brightness.
* A data matrix is a rectangular array of values.
* A data matrix can be visualized as an image by associating each entry of the data matrix with a pixel in the image.
* The brightness or color of the pixel is determined by the value of the corresponding entry of the matrix.

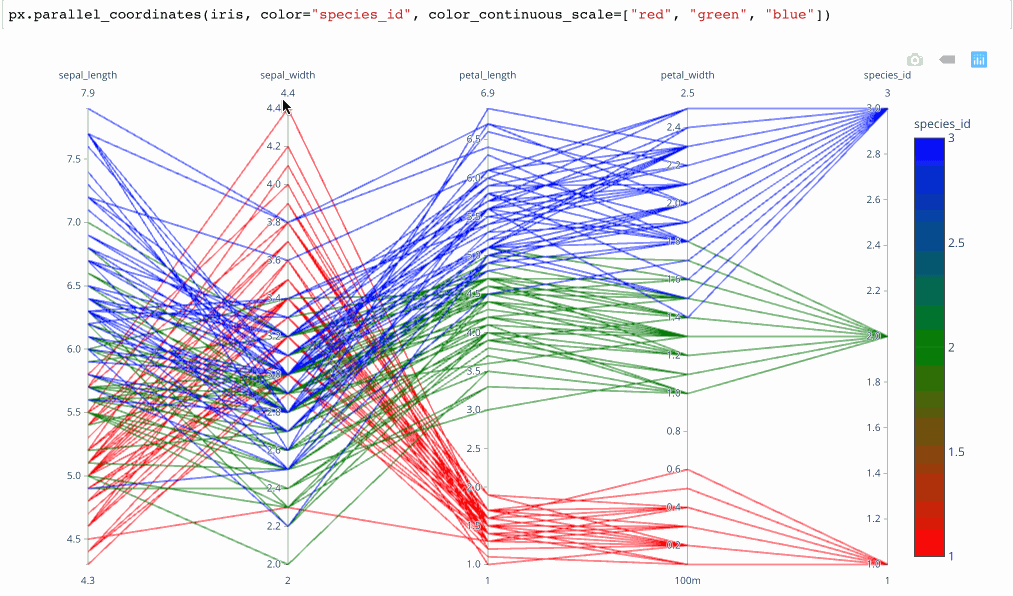


Some important practical considerations:

* If class labels are known, then it is useful to reorder the data matrix so that all objects of a class are together. This makes it easier, for example, to detect if all objects in a class have similar attribute values for some attributes.
* If different attributes have different ranges, then the attributes are often standardized to have a mean of zero and a standard deviation of 1. This prevents the attribute with the largest magnitude values from visually dominating the plot.

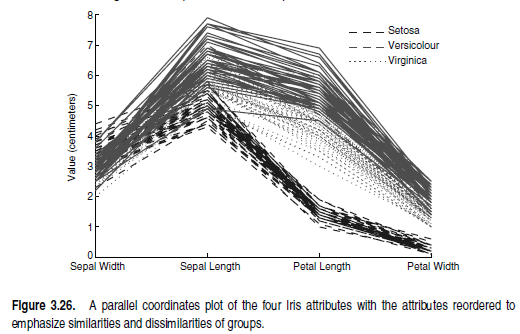
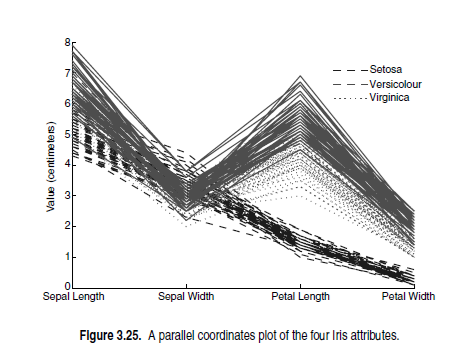
**Parallel Coordinates**

* Parallel coordinates have one coordinate axis for each attribute, but the different axes are parallel to one other instead of perpendicular, as is traditional.
* An object is represented as a line instead of as a point.
* Specifically, the value of each attribute of an object is mapped to a point on the coordinate axis associated with that attribute, and these points are then connected to form the line that represents the object.
* In many cases, objects tend to fall into a small number of groups, where the points in each group have similar values for their attributes.
* If so, and if the number of data objects is not too large, then the resulting parallel coordinates plot can reveal interesting patterns.



<https://www.serendipidata.com/posts/visualizing-high-dimensional-data>

* One of the drawbacks of parallel coordinates is that the detection of patterns in such a plot may depend on the order.
* For instance, if lines cross a lot, the picture can become confusing, and thus, it can be desirable to order the coordinate axes to obtain sequences of axes with less crossover.



**Star Coordinates and Chernoff Faces**

Another approach to displaying multidimensional data is to encode objects

as **glyphs** or **icons**—symbols that impart information non-verbally.

More specifically, each attribute of an object is mapped to a particular feature of a glyph, so that the value of the attribute determines the exact nature of the

feature. Thus, at a glance, we can distinguish how two objects differ.

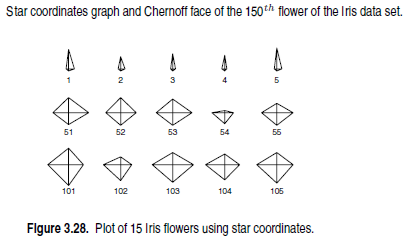
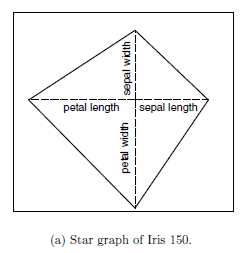
**Star coordinates**

* This technique uses one axis for each attribute. These axes all radiate from a center point, like the spokes of a wheel, and are evenly spaced.
* Typically, all the attribute values are mapped to the range [0,1].
* An object is mapped onto this star-shaped set of axes using the following

process:

1. Each attribute value of the object is converted to a fraction that represents its distance between the minimum and maximum values of the attribute.
2. This fraction is mapped to a point on the axis corresponding to this attribute.
3. Each point is connected with a line segment to the point on the axis preceding or following its own axis; this forms a polygon.

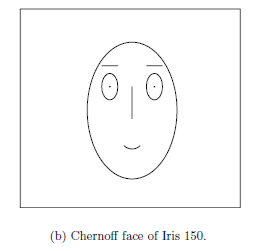
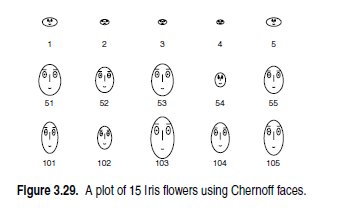
* The size and shape of this polygon gives a visual description of the attribute values of the object.
* For ease of interpretation, a separate set of axes is used for each object. In other words, each object is mapped to a polygon.

**Example of a star coordinates**: plot of flower 150 

**Chernoff faces**

* It is also possible to map the values of features to those of more familiar objects, such as faces.
* This technique is named **Chernoff faces** for its creator, Herman Chernoff.
* In this technique, each attribute is associated with a specific feature of a face, and the attribute value is used to determine the way that the facial feature is expressed.
* The shape of the face may become more elongated as the value of the corresponding data feature increases.

Example: A Chernoff face for flower 150 is given in Figure 3.27(b).



The program that we used to make this face mapped the features to the four features listed below. Other features of the face, such as width between the eyes and length of the mouth, are given default values.

|  |  |
| --- | --- |
| **Data Feature** | **Facial Feature** |
| sepal length | size of face |
| sepal width | forehead/jaw relative arc length |
| petal length | shape of forehead |
| petal width | shape of jaw |