Unit-1
Classification of Data

For performing statistical analysis, various kinds of data are gathered by the investigator or analyst. The information gathered is usually in raw form which is difficult to analyze. To make the analysis meaningful and easy, the raw data is converted or classified into different categories based on their characteristics. This grouping of data into different categories or classes with similar or homogeneous characteristics is known as the **Classification of Data**. Each division or class of the gathered data is known as a Class.

The different basis of classification of statistical information are Geographical, Chronological, Qualitative (Simple and Manifold), and Quantitative or Numerical.

For example, if an investigator wants to determine the poverty level of a state, he/she can do so by gathering the information of people of that state and then classifying them on the basis of their income, education, etc.

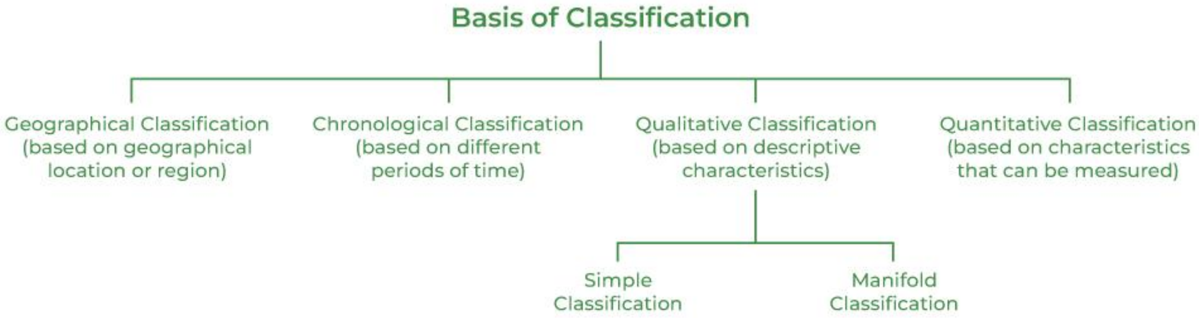
According to **Conner**, "Classification is the process of arranging things (either actually or notionally) in groups or classes according to their resemblances and affinities, and gives expression to the unity of attributes that may exist amongst a diversity of individuals."

The main objectives of Classification of Data are as follows:

- Explain similarities and differences of data
- Simplify and condense data's mass
- Facilitate comparisons
- Study the relationship
- Prepare data for tabular presentation
- Present a mental picture of the data

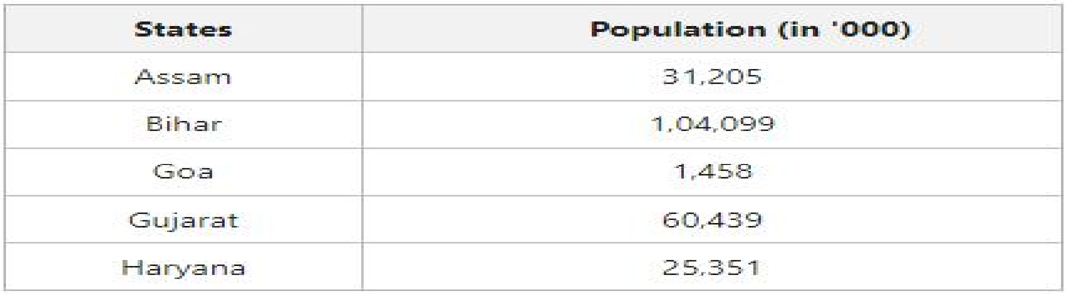
Basis of Classification of Data

The classification of statistical data is done after considering the scope, nature, and purpose of an investigation and is generally done on four bases; viz., geographical location, chronology, qualitative characteristics, and quantitative characteristics.


Basis of Classification
Geographical Classification (based on geographical location or region)
Chronological Classification (based on different periods of time)
Qualitative Classification (based on descriptive characteristics)
Quantitative Classification (based on characteristics that can be measured)
Simple Classification
Manifold Classification

1. Geographical Classification

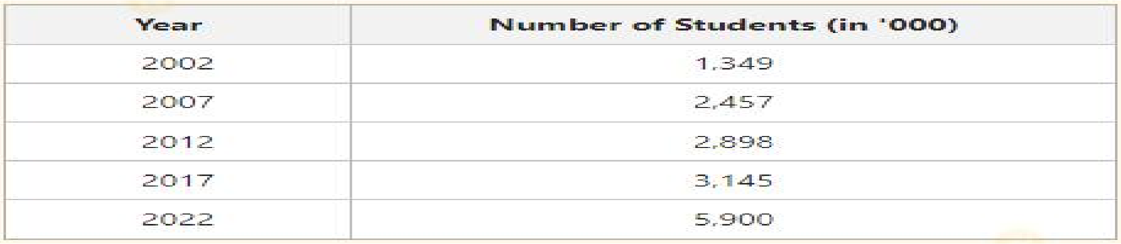
The classification of data on the basis of geographical location or region is known as **Geographical** or **Spatial Classification. For example,** presenting the population of different states of a country is done on the basis of geographical location or region.



States	Population (in '000)
Assam	31,205
Bihar	1,04,099
Goa	1,458
Gujarat	60,439
Haryana	25,351

2. Chronological Classification

The classification of data with respect to different time periods is known as **Chronological** or **Temporal Classification. For example,** the number of students in a school in different years can be presented on the basis of a time period.



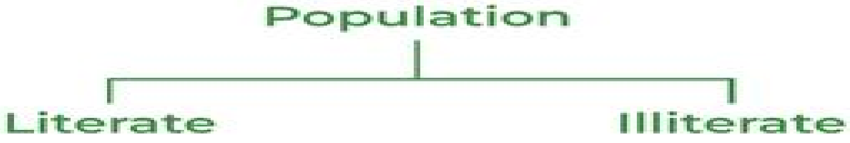
Year	Number of Students (in '000)
2002	1,349
2007	2,457
2012	2,898
2017	3,145
2022	5,900

3. Qualitative Classification

The classification of data on the basis of descriptive or qualitative characteristics like region, caste, sex, gender, education, etc., is known as **Qualitative Classification.** A qualitative classification can not be quantified and can be of two types; viz., **Simple Classification** and **Manifold Classification.**

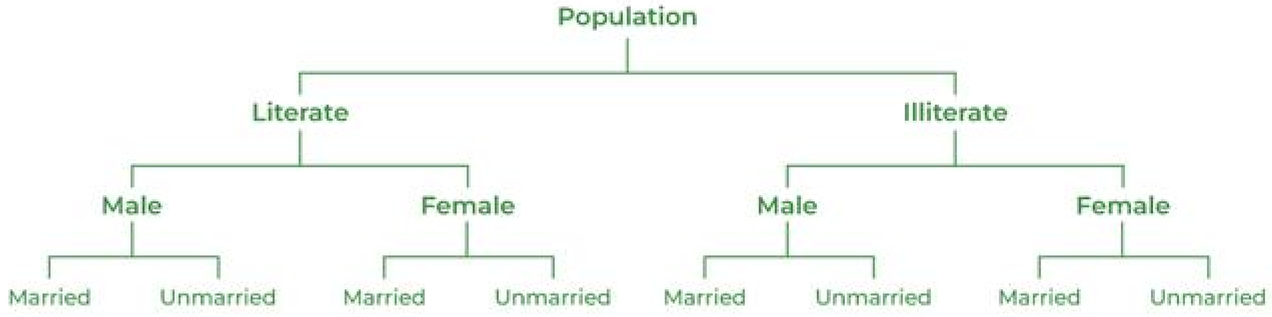
Simple Classification

When based on only one attribute, the given data is classified into two classes, which is known as **Simple Classification**. **For example,** when the population is divided into literate and illiterate, it is a simple classification.



Manifold Classification

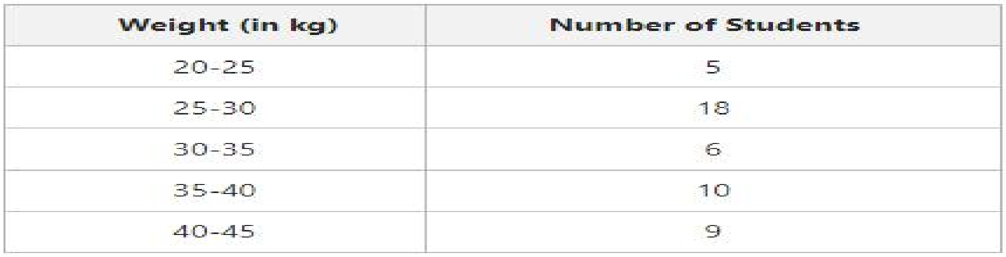
When based on more than one attribute, the given data is classified into different classes, and then sub-divided into more sub-classes, which is known as **Manifold Classification. For example,** when the population is divided into literate and illiterate, then sub-divided into male and female, and further sub-divided into married and unmarried, it is a manifold classification.



4. Quantitative Classification

The classification of data on the basis of the characteristics, such as age, height, weight, income, etc., that can be measured in quantity is known as **Quantitative Classification. For example,** the weight of students in a class can be classified as quantitative classification.

Classification. For example, the weight of students in a class can be classified as quantitative classification.

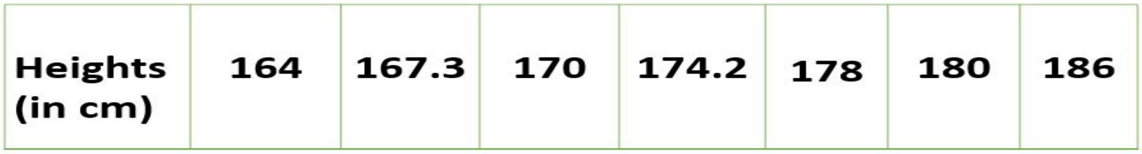


Weight (in kg)	Number of Students
20-25	5
25-30	18
30-35	6
35-40	10
40-45	9

Univariate, Bivariate and Multivariate data and its analysis

1. Univariate data –

This type of data consists of **only one variable**. The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes. It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it. The example of a univariate data can be height.

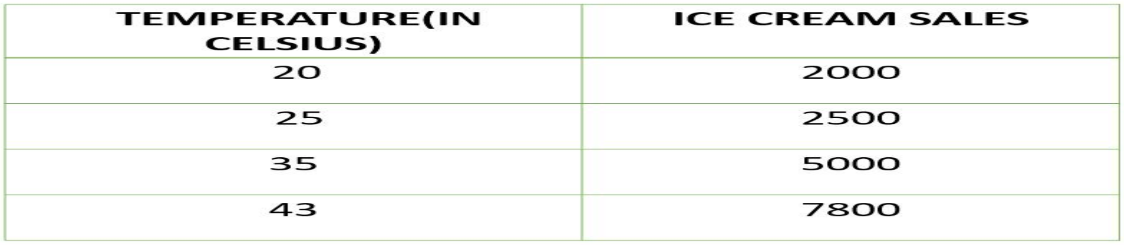


Heights (in cm)	164	167.3	170	174.2	178	180	186

Suppose that the heights of seven students of a class is recorded(figure 1),there is only one variable that is height and it is not dealing with any cause or relationship. The description of patterns found in this type of data can be made by drawing conclusions using central tendency measures (mean, median and mode), dispersion or spread of data (range, minimum, maximum, quartiles, variance and standard deviation) and by using frequency distribution tables, histograms, pie charts, frequency polygon and bar charts.

2. Bivariate data

This type of data involves **two different variables**. The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship among the two variables. Example of bivariate data can be temperature and ice cream sales in summer season.



TEMPERATURE(IN CELSIUS)	ICE CREAM SALES
20	2000
25	2500
35	5000
43	7800

Suppose the temperature and ice cream sales are the two variables of a bivariate data(figure 2). Here, the relationship is visible from the table that temperature and sales are directly proportional to each other and thus related because as the temperature increases, the sales also increase. Thus bivariate data analysis involves comparisons, relationships, causes and explanations. These variables are often plotted on X and Y axis on the graph for better understanding of data and one of these variables is independent while the other is dependent.

3. Multivariate data

When the data involves **three or more variables**, it is categorized under multivariate. Example of this type of data is suppose an advertiser wants to compare the popularity of four advertisements on a website, then their click rates could be measured for both men and women and relationships between variables can then be examined. It is similar to bivariate but contains more than one dependent variable. The ways to perform analysis on this data depends on the goals to be achieved. Some of the techniques are regression analysis, path analysis, factor analysis and multivariate analysis of variance (MANOVA).

There are a lots of different tools, techniques and methods that can be used to conduct your analysis. You could use software libraries, visualization tools and statistic testing methods. However, this blog we will be compare Univariate, Bivariate and Multivariate analysis.

Measures of Central Tendency in Statistics

Central Tendencies in Statistics are the numerical values that are used to represent mid-value or central value a large collection of numerical data. These obtained numerical values are called **central** or **average values** in Statistics. A central or average value of any statistical data or series is the value of that variable that is representative of the entire data or its associated frequency distribution. Such a value is of great significance because it depicts the nature or characteristics of the entire data, which is otherwise very difficult to observe.

Measures of Central Tendency Meaning

The representative value of a data set, generally the central value or the most occurring value that gives a general idea of the whole data set is called the Measure of Central Tendency.

Measures of Central Tendency

Some of the most commonly used measures of central tendency are:

- Mean
- Median
- Mode

Mean

Mean in general terms is used for the arithmetic mean of the data, but other than the arithmetic mean there are geometric mean and harmonic mean as well that are calculated using different formulas. Here in this article, we will discuss the arithmetic mean.

Median

The Median of any distribution is that value that divides the distribution into two equal parts such that the number of observations above it is equal to the number of observations below it. Thus, the median is called the central value of any given data either grouped or ungrouped.

Mode

The Mode is the value of that observation which has a maximum frequency corresponding to it. In other, that observation of the data occurs the maximum number of times in a dataset.