MODULE II

CHAPTER 2

Data Exploration and Data Pre-Processing

University Prescribed Syllabus w.e.f Academic Year 2021-2022

Types of Attributes, Statistical Description of Data, Measuring Data Similarity and Dissimilarity.

Why Preprocessing? Data Cleaning, Data Integration, Data Reduction: Attribute Subset Selection, Histograms, Clustering, Sampling, Data Cube aggregation, Data transformation and Data Discretization: Normalization, Binning, Histogram Analysis

Self-learning Topics Data Visualization, Concept hierarchy generation.

2.1 DATA EXPLORATION

- Data exploration refers to the initial step in data analysis in which data analysts use data visualization and statistical techniques to describe dataset characterizations, such as size, quantity, and accuracy, in order to better understand the nature of the data.
- Data exploration techniques include both manual analysis and automated data exploration software solutions that visually explore and identify relationships between different data variables, the structure of the dataset, the presence of outliers, and the distribution of data values in order to reveal patterns and points of interest, enabling data analysts to gain greater insight into the raw data.
- Data is often gathered in large, unstructured volumes from various sources and data analysts must first understand and develop a comprehensive view of the data before extracting relevant data for further analysis, such as univariate, bivariate, multivariate, and principal components analysis.
- Manual data exploration methods entail either writing scripts to analyze raw data or manually filtering data into spreadsheets. Automated data exploration tools, such as data visualization software, help data scientists easily monitor data sources and perform big data exploration on otherwise overwhelmingly large datasets. Graphical displays of data, such as bar charts and scatter plots, are valuable tools in visual data exploration.
- A popular tool for manual data exploration is Microsoft Excel spreadsheets, which can be used to create basic charts for data exploration, to view raw data, and to identify the correlation between variables.

2.1.1 Types of Attributes

Data Objects

- Data objects comprise to form data sets.
- A data object represents an entity, e.g. in a university database, the objects may be courses, professors and students.
- Data objects are typically described by attributes.
- If stored in a database, the data objects are referred to as data tuples. That is, the rows of the database correspond to the data objects and the columns correspond to the attributes.

Attribute Types

- An attribute represents the characteristic or feature of a data object. E.g., attributes for customer object can include customer_ID, name, and address.
- The type of an attribute is determined by the set of possible values the attribute can have. **They can be nominal, binary, ordinal, numeric, discrete or continuous.**

Module 2

(i) Nominal Attribute

- It is a qualitative attribute related to names.
- The values of a nominal attribute are names of things, some kind of symbols.
- Values of nominal attributes represents some category or state and thus, nominal attributes are also referred as **categorical attributes** and there is no order (rank, position) among values of the nominal attribute.
- **Example :**

Own House :	1. Yes 2. No
Marital status :	1. Unmarried 2. Married

(ii) Binary Attribute

- It is also a qualitative attribute.
- Binary data has only 2 values/states. For example, yes or no, affected or unaffected, true or false.
- Symmetric Binary Attribute: Both values are equally important (e.g. Gender).
- Asymmetric Binary Attribute: Both values are not equally important (e.g. Result).
- **Example :**

Gender :	Male, Female
Cancer Detected:	Yes, No
Result	Pass, Fail

(iii) Ordinal Attribute

- It is also a qualitative attribute.
- The Ordinal Attributes contains values that have a meaningful sequence or ranking(order) between them, but the magnitude between values is not actually known.
- The order of values shows what is important but don't indicate how important it is.
- **Example :**

Grade:	A, B, C, D, E, F, O
Income:	Low, Medium, High
Product Rating:	0, 1, 2, 3, 4, 5

(iv) Numeric Attribute

- A numeric attribute is quantitative because, it is a measurable quantity, represented in integer or real values.
- Numerical attributes are of 2 types, interval and ratio.
- An **interval-scaled attribute** has values, whose differences are interpretable, but the numerical attributes do not have the correct reference point, or we can call zero points. Data can be added and subtracted at an interval scale but cannot be multiplied or divided. Consider an example of temperature in degrees Centigrade. If a day's temperature of one day is twice of the other day, we cannot say that one day is twice as hot as another day.
- A **ratio-scaled attribute** is a numeric attribute with a fix zero-point. If a measurement is ratio-scaled, we can say of a value as being a multiple (or ratio) of another value. The values are ordered, and we can also compute the difference between values, and the mean, median, mode, Quantile-range, and Five number summary can be given.

(v) Discrete Attribute

- It is also a quantitative attribute.
- It can be numerical and can also be in categorical form.
- These attributes have finite or countably infinite set of values.
- **Example :**

Profession:	Principal, Teacher, Clerk, Peon
Zipcode:	400050, 400051, 400052

(vi) Continuous Attribute

- It is also a quantitative attribute.
- It can take any value between two specified values.
- **Example :**

Height:	5.2, 5.4, 5.6, ...
Weight:	50.33, ...

2.2 STATISTICAL DESCRIPTION AND DESCRIPTIVE DATA SUMMARIZATION

UQ Suppose that the data for analysis includes the attribute age. The age values for data tuples are (in increasing order): 13, 15, 16, 16, 19, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

(i) What is mean of data? What is median of data?

(ii) What is mode of data? Comment on data's modality.

(iii) What is mid-range of data?

(iv) Give the five-point summary of the data.

(v) Show boxplot of the data.

MU - May 2019, Dec. 2019

- It is essential to have an overall picture of the data, if data preprocessing is to be made successful.
- Statistical description of data is useful in identifying the properties of the data and highlight which data value should be treated at noise or outliers.
- Following are the basic statistical description of data :

2.2.1 Measures of Central Tendency

- A measure of central tendency is a number used to represent the center or middle of a set of data values.
- The mean, median, mode and midrange are commonly used measures of central tendency.

2.2.1(A) Mean

- The mean, or average, of n numbers is the sum of the numbers divided by n.

- The mean is denoted by $\bar{x}$ and is read as "x-bar".
- For the data set $x_1, x_2, \ldots, x_n$, the mean is

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

- Sometimes, weights are associated with the value x_i. The weights reflect the importance, significance or frequency of occurrence to their respective values. The weighted arithmetic mean or the weighted average is computed as

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} = \frac{w_1 x_1 + w_2 x_2 + \ldots + w_n x_n}{w_1 + w_2 + \ldots + w_n}$$

- Mean has one limitation; it is highly sensitive to outliers. Under such condition, median would be a better measure of central tendency.

2.2.1(B) Median

- The median of n numbers is the middle number when numbers are written in order.
- If n is even, the median is the mean of the two middle numbers.
- When we have large number of observations, the median is expensive to compute.
- In such case, we can approximate the median of the entire data set by interpolation using the **formula :**

$$\text{median} = L_1 + \left(\frac{\frac{n}{2} - (\Sigma \text{freq})_1}{\text{freq}_{\text{median}}}\right) \text{width}$$

where,

L_1 is the lower boundary of the median interval,

n is the number of values in the entire data set,

$(\Sigma \text{ freq})_1$ is the sum of frequencies of all of the intervals that are lower than the median interval

$\text{Freq}_{\text{median}}$ is the frequency of the median interval and

width is the width of the median interval.

2.2.1(C) Mode

- The mode of n numbers is the number or numbers that occur most frequently.
- There may be one mode, no mode or more than one mode.
- For unimodal numeric data that are asymmetrical, we have the following empirical relation :

mean − median = 3 × (mean − median)

2.2.1(D) Midrange

It is the average of the largest and smallest values in the set.

Ex. 2.2.1 : The data set below gives the waiting time (in minutes) of several people having the oil changed in their car at an auto mechanics shop. 22, 18, 25, 21, 28, 26, 20, 28, 20. Find the mean, median, mode and the midrange of the data set.

Soln. : Data set : 22, 18, 25, 21, 28, 26, 20, 28, 20

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

$$= \frac{22 + 18 + 25 + 21 + 28 + 26 + 20 + 28 + 20}{9}$$

$$= 23.11$$

- To find median, arrange the values in order.

18, 20, 20, 21, 22, 25, 26, 28, 28

- There are total 9 values i.e. n is odd. Thus, the median is the middle value.

Median = 22

- Mode is the number or numbers that occur most frequently. Here, 20 and 28 are repeated twice. Thus, data set is bimodal with values 20 and 28.

$$\text{Midrange} = \frac{\text{largest value} + \text{smallest value}}{2} = \frac{28 + 18}{2} = 23$$

2.2.2 Dispersion of Data

- A measure of dispersion is a statistic that tells you how dispersed, or spread out, data values are.
- The measures include range, quantiles, quartiles, percentiles, the interquartile range, and the five-number summary displayed as a boxplot, variance, and standard deviation.

2.2.2 (A) Quartiles

- Quartiles are values that divide your data into quarters.

- However, quartiles are not shaped like pizza slices; instead they divide your data into four segments according to which the numbers fall on the number line.
- The four quarters that divide a data set into quartiles are:
  a. The lowest 25% of numbers. Also called the 1st quartile (Q_1) or 25th percentile.
  b. The next lowest 25% of numbers (up to the median). Also called the 2nd quartile (Q_2) or 50th percentile.
  c. The second highest 25% of numbers (above the median). Also called the 3rd quartile (Q_3) or 75th percentile.
  d. The highest 25% of numbers. Also called the 4th quartile (Q_4) or 100th percentile.
- As quartiles divide numbers up according to where their position is on the number line, you have to put the numbers in order before you can figure out where the quartiles are.

2.2.2 (B) Interquartile Range (IQR)

- Interquartile range is defined as the difference between the upper and lower quartile values in a set of data.
- It is commonly referred to as IQR and is used as a measure of spread and variability in a data set.
- $IQR = Q_3 - Q_1$

2.2.2 (C) Five Number Summary

- The five number summary gives you a rough idea about what your data set looks like.
- It includes five items: the minimum value, the first quartile (Q_1), the median, the third quartile (Q_3), the maximum value.
- In order for the five numbers to exist, your data set must meet these two requirements :
  a. Your data must be **univariate**. In other words, the data must be a single variable. For example, this list of weights is one variable: 120, 100, 130, 145. If you have a list of ages and you want to compare the ages to weights, it becomes bivariate data (two variables). For example: age 1 (25 pounds), 5 (60 pounds), 15 (129 pounds). The matching pairs makes it impossible to find a five number summary.
  b. Your data must be **ordinal, interval,** or **ratio**.
- Steps to Find a Five-Number Summary

Step 1 : Put your numbers in ascending order (from smallest to largest).

For example, consider the data set in order as: 1, 2, 5, 6, 7, 9, 12, 15, 18, 19, 27.

Step 2: Find the minimum and maximum for your data set.

In the example in step 1, the minimum (the smallest number) is 1 and the maximum (the largest number) is 27.

Step 3 : Find the median. The median is the middle number.

Step 4 : Place parentheses around the numbers above and below the median. (This is not technically necessary, but it makes Q_1 and Q_3 easier to find).

(1, 2, 5, 6, 7), 9, (12, 15, 18, 19, 27).

Step 5 : Find Q_1 and Q_3. Q_1 can be thought of as a median in the lower half of the data, and Q_3 can be thought of as a median for the upper half of data.

(1, 2, 5, 6, 7), 9, (12,15,18,19,27).

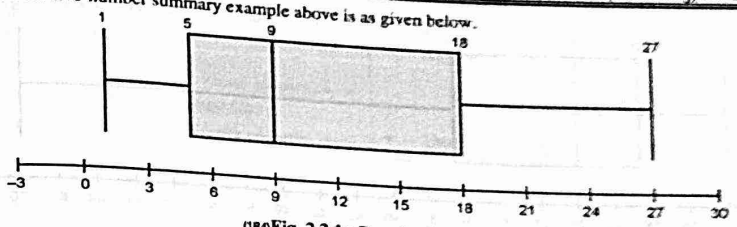
Step 6 : Write down your summary found in the above steps.

minimum = 1, Q_1 = 5, median = 9, Q_3 = 18 and maximum = 27.

2.2.2 (D) Boxplot

- A boxplot (or whisker plot) is defined as a graphical method of displaying variation in a set of data.
- A boxplot incorporates the five-summary as follows :
  a. The ends of the box are at the quartiles and the box length is the interquartile range.
  b. The median is marked by a line within a box.
  c. Two lines (called whiskers) outside the box extend to the minimum and maximum values in the data set.

- The boxplot for five-number summary example above is as given below.



(184)Fig. 2.2.1 : Boxplot Example

- Boxplots can be computed in $O(n\ log n)$ time.

2.2.2 (E) Outlier

It is a value higher or lower than 1.5 × IQR (**Inter-Quartile Range**)

2.2.2 (F) Variance and Standard Deviation

- Variance and standard deviation are measures of data dispersion. They indicate how spread out a data distribution is.
- For the data set $x_1, x_2, \ldots, x_n$, the variance is calculated as

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - (\bar{x})^2$$

where $\bar{x}$ is the mean value of the observation

- The standard deviation, σ, of the observations is the square root of the variance σ^2.
- A low standard deviation indicates that the data observations tend to be very close to the mean, while a high standard deviation indicates that the data observations are spread out over a large range of values.
- When all observations have the same value, $\sigma = 0$. Otherwise, $\sigma > 0$.

UEx. 2.2.2 **MU - May 2019, Dec. 2019**

Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order):

13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

(i) What is the mean of the data? What is the median?
(ii) What is the mode of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.)
(ii) What is the midrange of the data?
(iv) Give the five-point summary of the data.
(v) Show a boxplot of the data.

☑ **Soln. :**

(i) **Mean** = $\bar{x} = \sum_{i=1}^{n}\frac{x_i}{n} = \frac{809}{27} = 29.96 = 30$

The **median** (middle value of the ordered set, as the number of values in the set is odd) of the data is 25.

(ii) This data set has two values that occur with the same highest frequency and is, therefore, **bimodal**. The modes (values occurring with the greatest frequency) of the data are 25 and 35.

(iii) The **midrange** (average of the largest and smallest values in the data set) of the data is : $\frac{(70 + 13)}{2} = 41.5$

(iv) **Minimum** Value of the data : 13

The **first quartile** Q_1 (corresponding to the 25th percentile) of the data is : 20

The **median** of the data is : 25

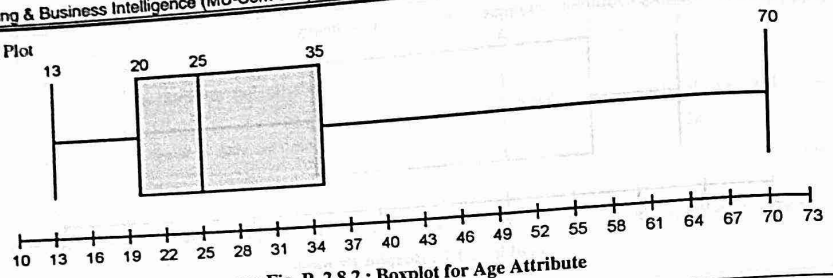
The **third quartile** Q_3 (corresponding to the 75th percentile) of the data is : 35

Maximum value of the data : 70

The five-point summary of the data :

{13, 20, 25, 35, 70}

(v) Box Plot



(185) **Fig. P. 2.8.2 : Boxplot for Age Attribute**

Ex. 2.2.3 : Suppose that the data for analysis includes the attribute salary. We have the following values for salary (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110.

(i) What are the mean, median, mode and midrange of the data?

(ii) Find the first quartile (Q_1) and the third quartile (Q_3) of the data.

(iii) Show the boxplot of the data.

☑ **Soln. :**

(i) Mean = $\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{696}{12} = 58$

(ii) The **median** (mean of the two middle values of the ordered set, as the number of values in the set is even) of the data is 54.

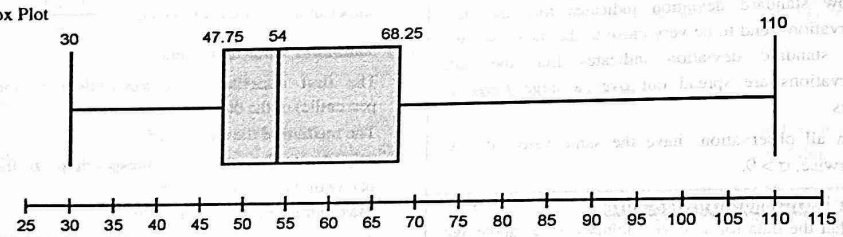
(iii) This data set has two values that occur with the same highest frequency and is, therefore, **bimodal**. The modes (values occurring with the greatest frequency) of the data are 52 and 70.

(iv) The **midrange** (average of the largest and smallest values in the data set) of the data is : $\frac{(110 + 30)}{2} = 70$

(v) The **first quartile** Q_1 (corresponding to the 25^{th} percentile) of the data is: 47.75.

The **third quartile** Q_3 (corresponding to the 75^{th} percentile) of the data is: 68.25.

(vi) Box Plot



(186) **Fig. P. 2.2.3 : Boxplot for Salary Attribute**

2.2.3 Graphic Displays of Basic Statistical Descriptions of Data

- Graphic displays are helpful for visual inspection of data, which is useful for data preprocessing.
- These include quantile plots, quantile-quantile plots, histograms and scatter plots.
- Quantile plots, quantile-quantile plots and histograms show univariate distributions.
- Scatter plots show bivariate distributions.

(a) Quantile Plot

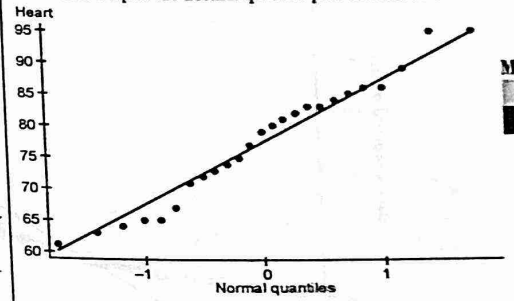
- A normal quantile plot (also known as a quantile-quantile plot or QQ plot) is a graphical way of checking whether your data are normally distributed.
- On one axis, you plot your data, sorted smallest to largest. On the other axis you plot the numbers you would expect to see if your data were normally distributed.
- If your data are normally distributed, you should see a nearly straight line.
- **Example:** Suppose we wish to know whether the resting heart rates of a sample of students are normally distributed.

Heart rate				
61	63	64	65	65
67	71	72	73	74
75	77	79	80	81
82	83	83	84	85
86	86	89	95	95

- We compute for each data value, the normal quantile value as

$$z = \frac{x_i - \bar{x}}{\rho}$$

- Then we plot the normal quantile plot of heart rate.

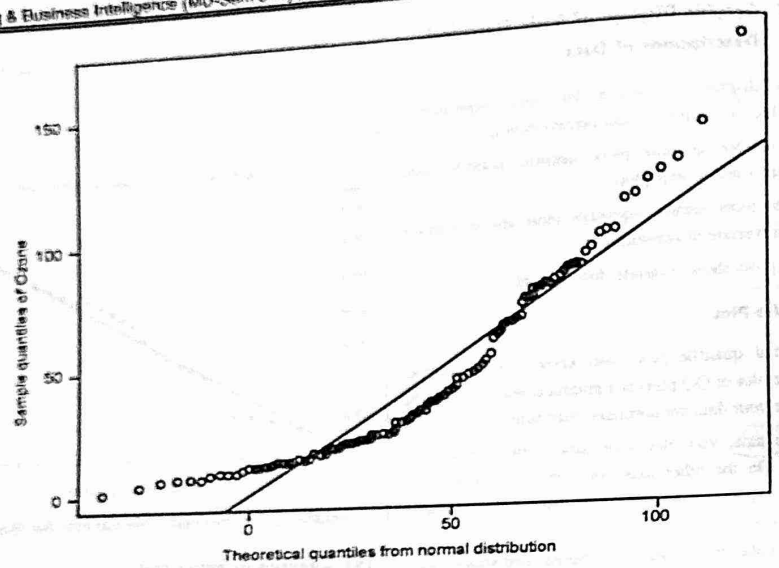


(187) **Fig. 2.2.2 : Normal Quantile Plot for Heart Rate**

(b) Quantile-quantile Plot

- Quantile-quantile Plots (Q-Q plots) are plots of two quantiles against each other.
- A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it.
- The purpose of Q-Q plots is to find out if two sets of data come from the same distribution.
- If the two data sets come from a common distribution, the points will fall on a single reference line.

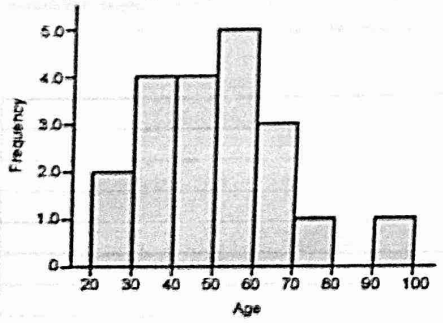
NOTES



(1B8)Fig. 2.2.3 : Quantile-Quantile Plot

(c) Histogram

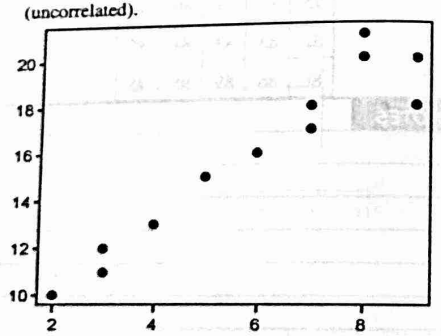
- Usually shows the distribution of values of a single variable.
- Divide the values into bins and show a bar plot of the number of objects in each bin.
- The height of each bar indicates the number of objects.
- Shape of histogram depends on the number of bins.



(1B9)Fig. 2.2.4 : Histogram

(d) Scatter Plot

- Scatter plot determines if there is a relationship, pattern, or trend existing between two numeric attributes.
- It also explores the possibility of correlation relationships between two attributes.
- Correlations can be positive, negative or zero (uncorrelated).



(1B10)Fig. 2.2.5 : Scatter Plot

2.3 DATA VISUALIZATION

READ

1. Visualization is the conversion of data into a visual or tabular format so that the characteristics of the data and the relationships among data items or attributes can be analyzed or reported.
2. Visualization of data is one of the most powerful and appealing techniques for data exploration.
3. Humans have a well-developed ability to analyze large amounts of information that is presented visually.
4. Can detect general patterns and trends.
5. Can detect outliers and unusual patterns.

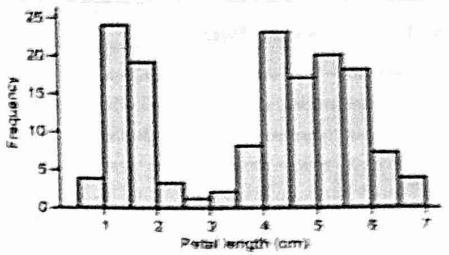
2.3.1 Visualization Techniques

2.3.1(A) Histogram

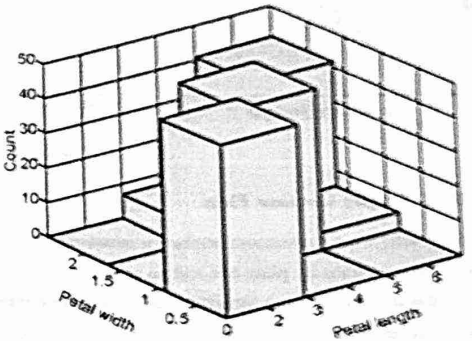
- Usually shows the distribution of values of a single variable.
- Divide the values into bins and show a bar plot of the number of objects in each bin.
- The height of each bar indicates the number of objects.
- Shape of histogram depends on the number of bins
- Example: Petal Width

Two-Dimensional Histograms

- Show the joint distribution of the values of two attributes
- **Example :** petal width and petal length



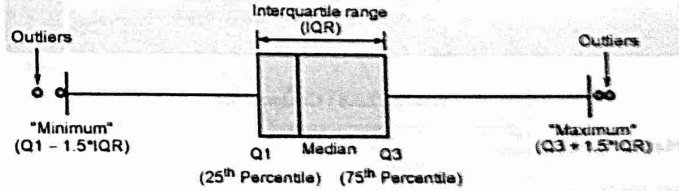
(1811)Fig. 2.3.1(a) : 1D Histogram



(1812)Fig. 2.3.1(b) : 2D Histogram

2.3.1(B) Boxplots

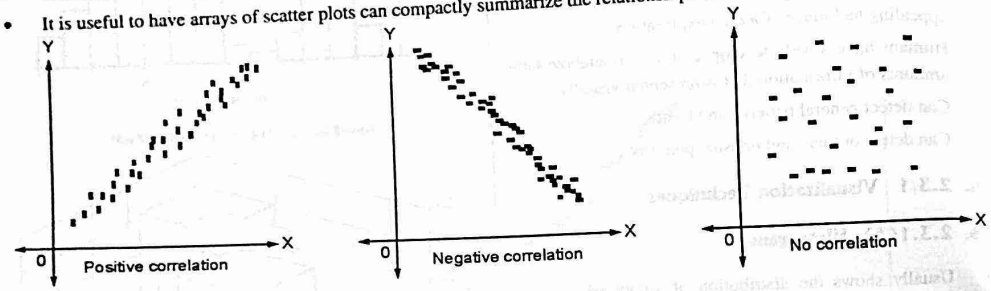
- Invented by J. Tukey.
- Another way of displaying the distribution of data.
- Following figure shows the basic part of a box plot.
- Box plots can be used to compare attributes.



(1813)Fig. 2.3.2: Boxplot

2.3.1(C) Scatter Plots

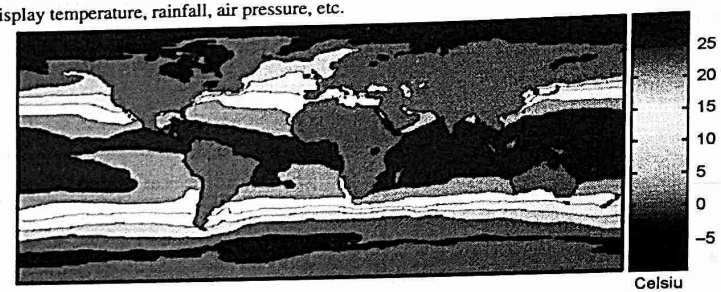
- Attributes values determine the position.
- Two-dimensional scatter plots most common, but can have three-dimensional scatter plots.
- Often additional attributes can be displayed by using the size, shape, and color of the markers that represent the objects.
- Often additional attributes can be displayed by using the size, shape, and color of the markers that represent the objects.
- It is useful to have arrays of scatter plots can compactly summarize the relationships of several pairs of attributes.



(1B14)Fig. 2.3.3: Scatter Plots

2.3.1(D) Contour Plots

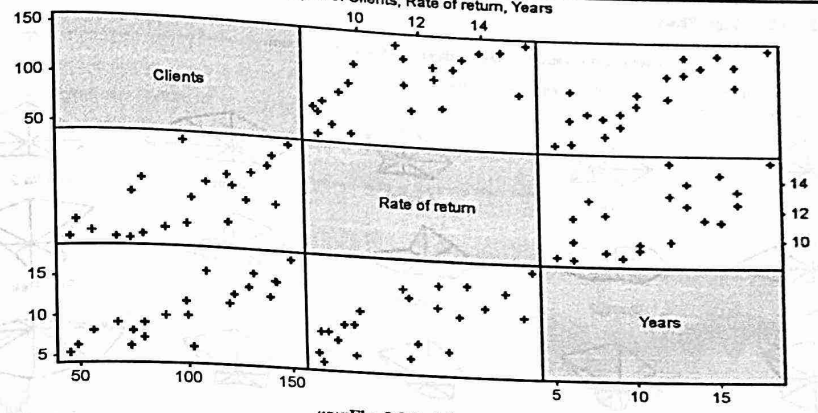
- Useful when a continuous attribute is measured on a spatial grid.
- They partition the plane into regions of similar values.
- The contour lines that form the boundaries of these regions connect points with equal values.
- The most common example is contour maps of elevation.
- Can also display temperature, rainfall, air pressure, etc.



(1B15)Fig. 2.3.4 : Contour Plot

2.3.1(E) Matrix Plots

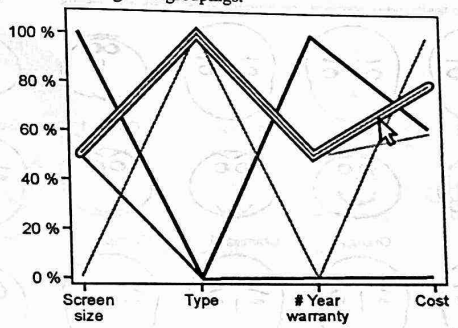
- Can plot the data matrix.
- This can be useful when objects are sorted according to class.
- Typically, the attributes are normalized to prevent one attribute from dominating the plot.
- Plots of similarity or distance matrices can also be useful for visualizing the relationships between objects.



(1B16)Fig. 2.3.5 : Matrix Plot

2.3.1(F) Parallel Coordinates

- Used to plot the attribute values of high-dimensional data.
- Instead of using perpendicular axes, use a set of parallel axes.
- The attribute values of each object are plotted as a point on each corresponding coordinate axis and the points are connected by a line.
- Thus, each object is represented as a line.
- Often, the lines representing a distinct class of objects group together, at least for some attributes.
- Ordering of attributes is important in seeing such groupings.

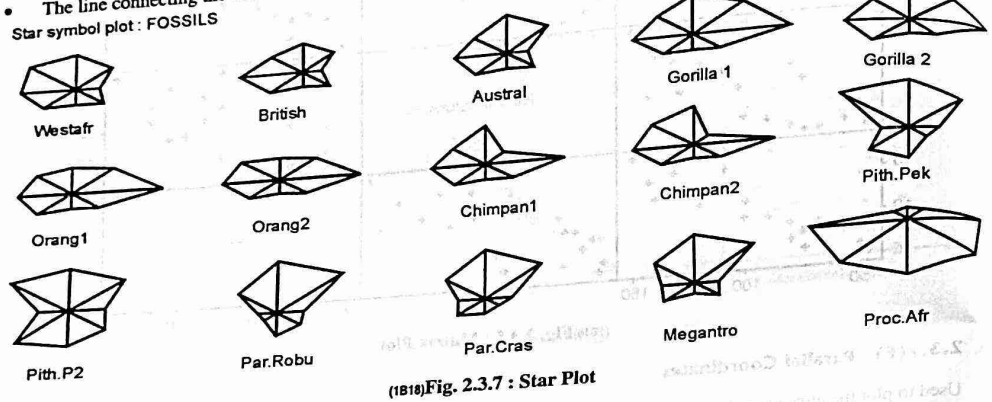


(1B17)Fig. 2.3.6 : Parallel Coordinates Plot

Module 2

2.3.1(G) Star Plots

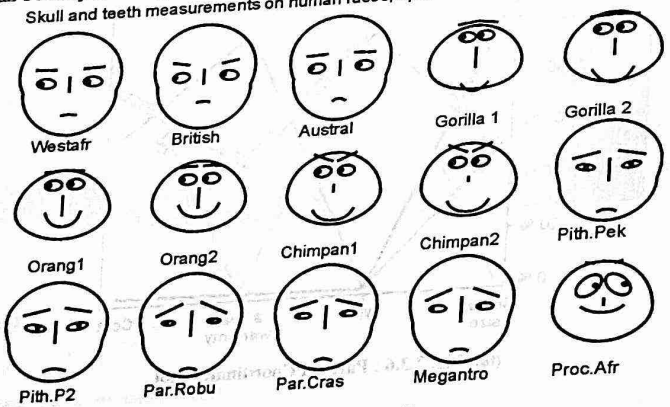
- Similar approach to parallel coordinates, but axes radiate from a central point.
- The line connecting the values of an object is a polygon.



(1B18)**Fig. 2.3.7 : Star Plot**

2.3.1(H) Chernoff Faces

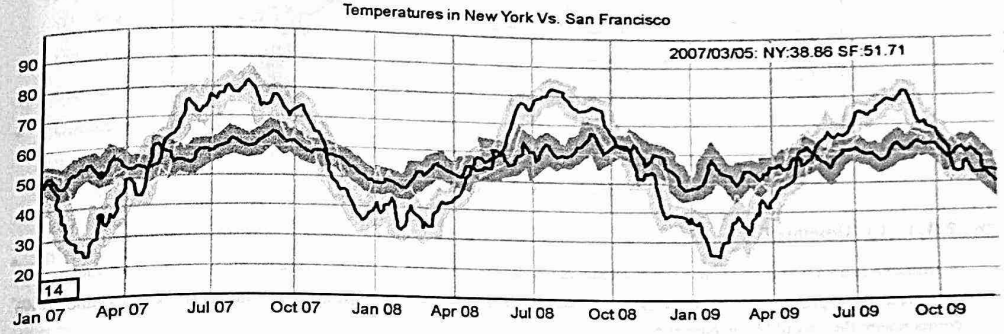
- Approach created by Herman Chernoff.
- This approach associates each attribute with a characteristic of a face.
- The values of each attribute determine the appearance of the corresponding facial characteristic.
- Each object becomes a separate face.
- Relies on human's ability to distinguish faces.



(1B19)**Fig. 2.3.8 : Chernoff Faces**

2.3.1(I) Dygraphs

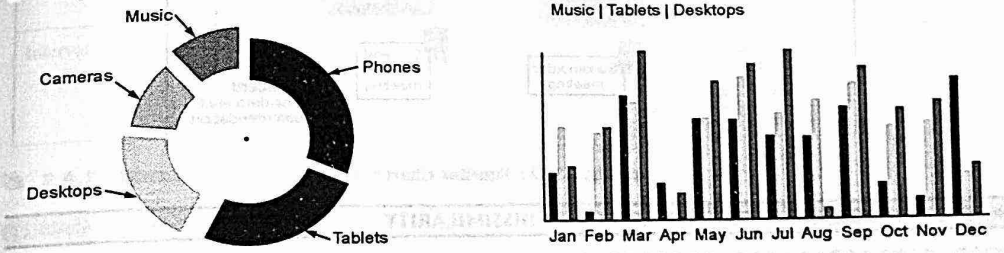
- Dygraphs is an open source JavaScript library that produces interactive, zoomable charts of time series.
- It is designed to display dense data sets and enable users to explore and interpret them.
- Another significant feature of the dygraphs library is the ability to display error bars around data series.
- dygraphs is purely client-side JavaScript. It does not send your data to any servers – the data is processed entirely in the client's browser.



(1B20)**Fig. 2.3.9 : Dygraphs**

2.3.1(J) Zing Chart

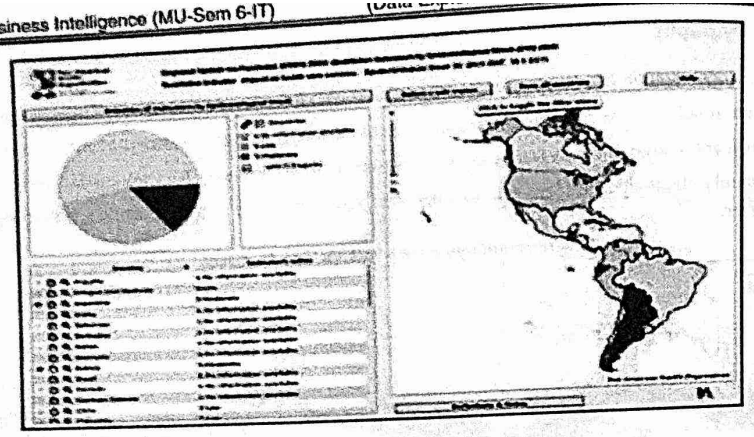
- Zing Chart is a JavaScript charting library that can help you manipulate data into visually appealing charts and graphs.
- First and foremost, Zing Chart is designed to handle big data and deliver fast results.
- Zing Chart is also mobile ready so data can adapt to screen size, or you can develop directly for mobile apps.



(1B21)**Fig. 2.3.10 : Zing Chart**

2.3.1(K) InstantAtlas

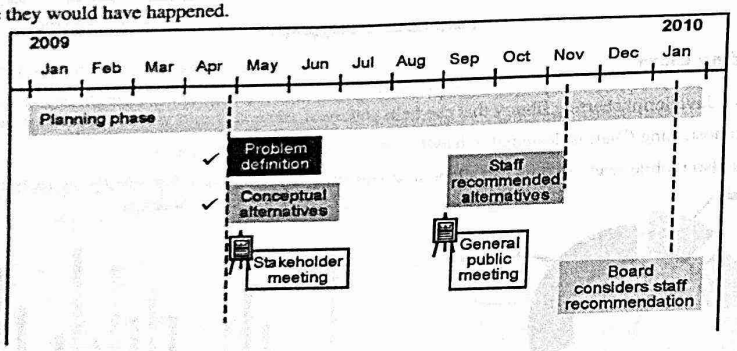
- Instant Atlas enables information analysts and researchers to create highly-interactive dynamic and profile reports that combine statistics and map data to improve data visualization, enhance communication, and engage people in more informed decision making.



(1B22)Fig. 2.3.11 : InstantAtlas

2.3.1 (L) Timeline

- A **timeline** is a way of displaying a list of events in chronological order, sometimes described as a project artifact.
- It is typically a graphic design showing a long bar labeled with dates alongside itself and usually events labeled on points where they would have happened.



(1B23)Fig. 2.3.12 : Timeline Chart

2.4 MEASURING DATA SIMILARITY AND DISSIMILARITY

UQ. Describe different types of attributes with example **MU - Dec. 2019**

- Distance or similarity measures are essential in solving many pattern recognition problems such as classification and clustering.
- Various distance/similarity measures are available in the literature to compare two data distributions.
- **Similarity** measure
  - (i) is a numerical measure of how alike two data objects are.
  - (ii) is higher when objects are more alike.
  - (iii) often falls in the range [0,1]

- Similarity might be used to identify
  - (i) duplicate data that may have differences due to typos.
  - (ii) equivalent instances from different data sets. E.g. names and/or addresses that are the same but have misspellings.
  - (iii) groups of data that are very close (clusters)
- **Dissimilarity** measure
  - (i) is a numerical measure of how different two data objects are
  - (ii) is lower when objects are more alike
  - (iii) minimum dissimilarity is often 0 while the upper limit varies depending on how much variation can be
- Dissimilarity might be used to identify
  - (i) Outliers
  - (ii) interesting exceptions, e.g. credit card fraud
  - (iii) boundaries to clusters
- **Proximity** refers to either a similarity or dissimilarity.

2.4.1 Single Attribute Similarity/Dissimilarity Measures

Attribute	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$	$s = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$
Ordinal	$d = \frac{\lvert x - y \rvert}{n - 1}$ Values are mapped to integers 0 to n – 1 where n is the number of values.	$s = 1 - d$
Interval or Ratio	$d = \lvert x - y \rvert$	$s = -d, s = \frac{1}{1+d}, s = e^{-d}$, $s = 1 - \frac{d - \min\_d}{\max\_d - \min\_d}$

2.4.2 Distance between Instances with Multiple Attributes

(1) Euclidean Distance

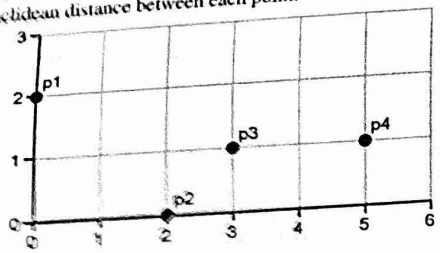
- The Euclidean distance is computed using formula

$$d(x, y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

where n is the number of dimensions (attributes) and x_k and y_k are, the *k-th* attributes (components) or data objects of x and y respectively.

- Standardization/normalization may be necessary to ensure an attribute does not skew the distances due to different scales.

Ex. 2.4.1 : Consider the data given below and compute the Euclidean distance between each point.



(1B24)Fig. P. 2.4.1

☑ **Soln. :** The x and y co-ordinate for each point is listed below.

p1(0,2), p2(2,0), p3(3,1) and p4(5,1)

The Euclidean distance formula is:

$$d(x, y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

The distance matrix between each point using the above formula is:

	p1	p2	p3	p4
p1	0			
p2	2.828	0		
p3	3.162	1.414	0	
p4	5.099	3.162	2	0

(2) Minkowski Distance

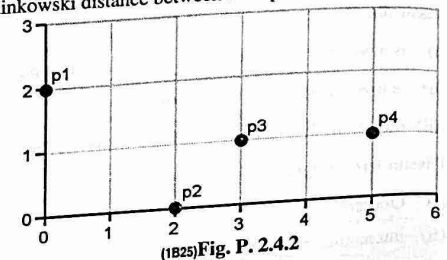
- It is a generalization of Euclidean distance.
- It is calculated using the formula :

$$d(x, y) = \sqrt[r]{\sum_{k=1}^{n} |x_k - y_k|^r}$$

where n is the number of dimensions (attributes) and x_k and y_k are, the *k-th* attributes (components) or data objects of x and y respectively.

- When r = 1, it is also called as City Block distance or Manhattan distance or L_1 norm distance.
- When r = 2, it is also called Euclidean distance or L_2 norm distance.
- When r = ∞, it is also called supremum or L_{max} norm or L_∞ norm distance. This is the maximum difference between any component of the vectors.

Ex. 2.4.2 : Consider the data given below and compute the Minkowski distance between each point.



(1B25)Fig. P. 2.4.2

☑ **Soln. :** The x and y co-ordinate for each point is listed below.

p1(0,2), p2(2,0), p3(3,1) and p4(5,1)

Minkowski distance is computed using formula

$$d(x, y) = \sqrt[r]{\sum_{k=1}^{n} |x_k - y_k|^r}$$

L_1 norm distance where r = 1 is shown in the matrix below.

L_1	p1	p2	p3	p4
p1	0			
p2	4	0		
p3	4	2	0	
p4	6	4	2	0

L_2 norm distance where r = 2 is shown in the matrix below.

L_2	p1	p2	p3	p4
p1	0			
p2	2.828	0		
p3	3.162	1.414	0	
p4	5.099	3.162	2	0

L_∞ norm distance is the maximum difference between any component of the vectors.

L_∞	p1	p2	p3	p4
p1	0			
p2	2	0		
p3	3	1	0	
p4	5	3	2	0

(3) Cosine Similarity

- Cosine similarity is a measure of similarity between two vectors. The data objects here are treated as vectors.
- Similarity is measured as the angle θ between the two vectors. Similarity is 1 when θ = 0 and 0 when θ = 90°.
- If $\mathbf{d_1}$ and $\mathbf{d_2}$ are two document vectors, then

$$\cos(d_1, d_2) = \frac{d_1 \cdot d_2}{\| d_1 \| \times \| d_2 \|}$$

where $d_1 \cdot d_2$ indicates inner product or vector dot product of vectors $\mathbf{d_1}$ and $\mathbf{d_2}$, and $\| d \|$ is the length of vector d calculated as

$$\| d \| = \sqrt{\sum_{k=1}^{n} d_k^2}$$

Module 2

Ex. 2.4.3 : Given two data objects

$\mathbf{d_1}$ = 3 2 0 5 0 0 0 2 0 0

$\mathbf{d_2}$ = 1 0 0 0 0 0 0 1 0 2

Calculate the cosine similarity.

☑ **Soln. :**

$\mathbf{d_1}$ = 3 2 0 5 0 0 0 2 0 0

$\mathbf{d_2}$ = 1 0 0 0 0 0 0 1 0 2

$\mathbf{d_1} \cdot \mathbf{d_2}$ = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5

$\| \mathbf{d_1} \|$ = $(3*3 + 2*2 + 0*0 + 5*5 + 0*0 + 0*0 + 0*0 + 2*2 + 0*0 + 0*0)^{1/2} = (42)^{1/2} = 6.481$

$\| \mathbf{d_2} \|$ = $(1*1 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 1*1 + 0*0 + 2*2)^{1/2} = (6)^{1/2} = 2.449$

$$\cos(d_1, d_2) = \frac{d_1 \cdot d_2}{\| d_1 \| \times \| d_2 \|} = \frac{5}{6.481 \times 2.449} = 0.315$$

Thus, the cosine similarity between d_1 and d_2 is 0.315

The dissimilarity between d_1 and d_2 is $1 - \cos(d_1, d_2) = 1 - 0.315 = 0.685$

(4) Jaccard Distance

- The Jaccard similarity index (sometimes called the Jaccard similarity *coefficient*) compares members for two sets to see which members are shared and which are distinct.
- It is a measure of similarity for the two sets of data, with a range from 0% to 100%.
- The formula to compute Jaccard index is:

$$J(x, y) = \frac{|x \cap y|}{|x \cup y|} = \frac{|x \cap y|}{|x| + |y| - |x \cap y|}$$

Ex. 2.4.4 : Given the record of users and movies viewed. Using Jaccard similarity measures, find similarity between {A-B, A-C, B-C}.

Users	Movie1	Movie2	Movie3	Movie4	Movie5
A	1	0	1	0	1
B	0	0	1	0	1
C	0	1	0	0	1

☑ **Soln. :**

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{2}{3} = 0.67$$

$$J(A, C) = \frac{|A \cap C|}{|A \cup B|} = \frac{1}{4} = 0.25$$

$$J(B, C) = \frac{|B \cap C|}{|B \cup C|} = \frac{1}{3} = 0.33$$

2.5 DATA PREPROCESSING

UQ. What is data preprocessing? Explain different methods. **(MU - June 2021)**

- Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, lacking in certain behaviors or trends, and is likely to contain many errors.
- Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing.
- Data preprocessing is used in database-driven applications such as customer relationship management and rule-based applications (like neural networks).
- In Machine Learning (ML) processes, data preprocessing is critical to encode the dataset in a form that could be interpreted and parsed by the algorithm.
- Data goes through a series of steps during preprocessing :

(1) Data Cleaning (2) Data Integration
(3) Data Transformation (4) Data Reduction
(5) Data Discretization (6) Data Sampling

- **(1) Data Cleaning :** Data is cleansed through processes such as filling in missing values or deleting rows with missing data, smoothing the noisy data, or resolving the inconsistencies in the data. Smoothing noisy data is particularly important for ML datasets, since machines cannot make use of data they cannot interpret. Data can be cleaned by dividing it into equal size segments that are thus smoothed (binning), by fitting it to a linear or multiple regression function (regression), or by grouping it into clusters of similar data (clustering). Data inconsistencies can occur due to human errors (the information was stored in a wrong field). Duplicated values should be removed through deduplication to avoid giving that data object an advantage (bias).
- **(2) Data Integration :** Data with different representations are put together and conflicts within the data are resolved.
- **(3) Data Transformation :** Data is normalized and generalized. Normalization is a process that ensures that no data is redundant, it is all stored in a single place, and all the dependencies are logical.
- **(4) Data Reduction :** When the volume of data is huge, databases can become slower, costly to access, and challenging to properly store. Data reduction step aims to present a reduced representation of the data in a data warehouse. There are various methods to reduce data. For example, once a subset of relevant attributes is chosen for its significance, anything below a given level is discarded. Encoding mechanisms can be used to reduce the size of data as well. If all original data can be recovered after compression, the operation is labelled as lossless. If some data is lost, then it's called a lossy reduction. Aggregation can also be used, for example, to condense countless transactions into a single weekly or monthly value, significantly reducing the number of data objects.
- **(5) Data Discretization :** Data could also be discretized to replace raw values with interval levels. This step involves the reduction of a number of values of a continuous attribute by dividing the range of attribute intervals.
- **(6) Data Sampling :** Sometimes, due to time, storage or memory constraints, a dataset is too big or too complex to be worked with. Sampling techniques can be used to select and work with just a subset of the dataset, provided that it has approximately the same properties of the original one.

2.6 CLEANING

UQ. What is noisy data? How to handle it? **MU - May 2019**

UQ. Explain Regression. Explain linear regression with example. **MU - Dec. 2019**

UQ. Partition the given data into 4 bins using binning method and perform smoothing by bin mean, by bin boundaries.
Data : 11, 13, 13, 15, 15, 16, 19, 20, 20, 20, 21, 21, 22, 23, 24, 30, 40, 45, 45, 45, 71, 72, 73, 75 **MU - June 2021**

- Data cleansing or data cleaning is the process of identifying and removing (or correcting) inaccurate records from a dataset, table, or database and refers to recognizing unfinished, unreliable, inaccurate, or non-relevant parts of the data and then restoring, remodelling, or removing the dirty or crude data.
- Data cleaning techniques may be performed as batch processing through scripting or interactively with data cleansing tools.
- After cleaning, a dataset should be uniform with other related datasets in the operation. The discrepancies identified or eliminated may have been basically caused by user entry mistakes, by corruption in storage or transmission, or by various data dictionary descriptions of similar items in various stores.
- Some data cleaning methods are explained in this section.

2.6.1 Missing Values

Imagine that you are asked to analyze a dataset. You find that there are many tuples having no recorded value for several attributes such as customer income. So the question arising here is how to fill in the missing values for this attribute. There are several methods as discussed here.

(1) **Ignore the tuple :** When the class label is missing, this technique is used. However, unless the tuple contains numerous attributes with missing values, this approach is not particularly useful.

(2) **Fill in the missing value manually :** This approach is effective on small data set with some missing values.

(3) **Use a global constant to fill in the missing value :** You can replace all missing attribute values with global constant, such as a label like "Unknown" or $-\infty$.

(4) **Use a measure of central tendency for attribute (e.g. the mean or median) to fill in the missing value :** For example, suppose customer average income is $25000, then you can use this value to replace missing value for income.

(5) **Use the attribute mean or median for all samples belonging to the same class as the given tuple :** For example, if you are classifying customers according to their credit_score, then you can replace the missing value with the mean income value for customers in the same credit_score category as that of the given tuple. If the data distribution for a given class is skewed, then use the median value.

(6) **Use the most probable value to fill in the missing value :** This can be determined using regression, Bayesian classification or decision-tree induction.

2.6.2 Noisy Data

- A random error or variance in a measured variable is referred to as noise.
- Noisy data can occur as a result of data gathering from malfunctioning instruments, data input issues, or technological limitations.
- Examples of noise in data:
  1. Encoding error: Gender has value, say, Z.
  2. Value beyond range: Age = 200
  3. Inconsistent entries: Date_of_joining = 32-Aug-2007

Techniques to Handle Noisy Data

(1) Binning

- Binning procedures smooth a sorted data item by examining its "neighborhood," or the values in its immediate vicinity.
- The sorted values are divided into several "buckets" or bins.

Different approaches of binning

(a) Smoothing by bin means : In smoothing by bin means, each value in a bin is replaced by the mean value of the bin.

(b) Smoothing by bin median : In this method each bin value is replaced by its bin median value.

(c) Smoothing by bin boundary : In smoothing by bin boundaries, the minimum and maximum values in a given bin are identified as the bin boundaries. Each bin value is then replaced by the closest boundary value.

Approach

(1) Sort the array of given data set.

(2) Divides the range into N intervals, each containing the approximately same number of samples (Equal-depth partitioning).

(3) Store mean/ median/ boundaries in each row.

Ex. 2.6.1 : Suppose a group of age records has been sorted as follows: 3, 7, 8,13, 22, 22, 22, 26, 26, 28, 30, 37. Partition them into three bins by equal-frequency (Equi-depth) partitioning method. Perform data smoothing by bin mean, and bin boundary.

☑ **Soln. :**

Smooth the data by equal frequency bins

Given three bins. There are total 12 observations. Hence, by equi-depth partitioning method, each bin will have 4 observations.

- Bin 1: 3, 7, 8, 13
- Bin 2: 22, 22, 22, 26
- Bin 3: 26, 28, 30, 37

Smooth the data by bin mean

We take average of each bin and replace each data value by mean value in corresponding bin.

- Bin 1: 8, 8, 8, 8
- Bin 2: 23, 23, 23, 23
- Bin 3: 30, 30, 30, 30

Smooth the data by bin boundary

We take difference of each data value and the bin boundaries. Each bin value is then replaced by the closest boundary value.

- Bin 1: 3, 3, 3, 13
- Bin 2: 22, 22, 22, 26
- Bin 3: 26, 26, 26, 37

Ex. 2.6.2 : Suppose a group of sales price records has been sorted as follows: 6, 9, 12, 13, 15, 25, 50, 70, 72, 92, 204, 232. Partition them into three bins by equal-frequency (Equi-depth) partitioning method. Perform data smoothing by bin mean.

☑ **Soln. : Smooth the data by equal frequency bins**

Given three bins.

There are total 12 observations. Hence, by equi-depth partitioning method, each bin will have 4 observations.

- Bin 1: 6, 9, 12, 13
- Bin 2: 15, 25, 50, 70
- Bin 3: 72, 92, 204, 232

Smooth the data by bin mean

We take average of each bin and replace each data value by mean value in corresponding bin.

- Bin 1: 10, 10, 10, 10
- Bin 2: 40, 40, 40, 40
- Bin 3: 150, 150, 150, 150

UEx. 2.6.3 **MU- June 2021**

Partition the given data into 4 bins using binning method and perform smoothing by bin mean, by median, by bin boundaries. **Data:** 11, 13, 13, 15, 15, 16, 19, 20, 20, 20, 21, 21, 22, 23, 24, 30, 40, 45, 45, 45, 71, 72, 73, 75

☑ **Soln. : Smooth the data by equal frequency bins**

Given four bins. There are total 12 observations. Hence, by equi-depth partitioning method, each bin will have 6 observations.

- Bin 1: 11, 13, 13, 15, 15, 16
- Bin 2: 19, 20, 20, 20, 21, 21
- Bin 3: 22, 23, 24, 30, 40, 45
- Bin 4: 45, 45, 71, 72, 73, 75

Smooth the data by bin mean

We take average of each bin and replace each data value by mean value in corresponding bin.

- Bin 1: 13.83, 13.83, 13.83, 13.83, 13.83, 13.83
- Bin 2: 20.16, 20.16, 20.16, 20.16, 20.16, 20.16
- Bin 3: 30.67, 30.67, 30.67, 30.67, 30.67, 30.67
- Bin 4: 63.50, 63.50, 63.50, 63.50, 63.50, 63.50

Smooth the data by bin median

We replace each value in the bin by its corresponding median value. Each bin contains 6 data values. So we take average of two middle values in corresponding bin and take it as median.

- Bin 1: 14, 14, 14, 14, 14, 14
- Bin 2: 20, 20, 20, 20, 20, 20
- Bin 3: 27, 27, 27, 27, 27, 27
- Bin 4: 71.50, 71.50, 71.50, 71.50, 71.50, 71.50

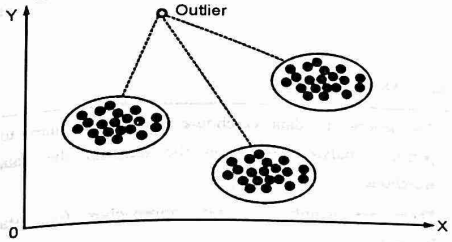
Smooth the data by bin boundary

We take difference of each data value and the bin boundaries. Each bin value is then replaced by the closest boundary value.

- Bin 1: 11, 11, 11, 16, 16, 16
- Bin 2: 19, 19, 19, 21, 21, 21
- Bin 3: 22, 22, 22, 22, 45, 45
- Bin 4: 45, 45, 75, 75, 75, 75

(2) Outlier analysis by clustering

- Outliers are nothing but an extreme value that deviates from the other observations in the dataset.
- Outlier Analysis is a process that involves identifying the anomalous observation in the dataset.
- Outliers may be detected by clustering where similar values are organized into groups, or clusters.
- The values that fall outside of the set of clusters may be considered outliers.



(1B26)**Fig. 2.6.1 : Outlier Analysis by Clustering**

(3) Regression

- Regression is a data mining technique used to predict a range of numeric values (also called *continuous values*), given a particular dataset. For example, regression might be used to predict the cost of a product or service, given other variables.
- Data smoothing can also be done using regression.
- Regression is of two types:

(a) Linear Regression : It finds the best line to fit two variables or attributes so that one variable can be used to predict the other variable.

Linear Regression : $y = \alpha + \beta x$

(b) Multiple Linear Regression : It is an extension of linear regression where more than two variables are involved and the data are fit to a multidimensional surface.

Multiple Linear Regression :

$y = \alpha + \beta_1 x_1 + \beta_1 x_2 + \ldots + \beta_n x_n$

Here, y is the variable that is to be predicted, x is the variable used to predict y, α is the intercept and β is the slope.

$$\beta = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Where, $\bar{x}$ and $\bar{y}$ are the mean of variables x and y respectively.

Ex. 2.6.3 : Develop a model to predict the salary of college graduates with 10 years of work experience using linear regression.

Years of Experience (x)	3	8	9	13	3	6	11	21	1	16
Salary in $100 (y)	30	57	64	72	36	43	59	90	20	83

☑ **Soln. :**

Linear Regression : $y = \alpha + \beta x$

where

$$\beta = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = 9.1 \quad \bar{y} = \frac{1}{10}\sum_{i=1}^{10} y_i = 55.4$$

X	Y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
3	30	– 6.1	– 25.4	154.94	37.21
8	57	– 1.1	1.6	– 1.76	1.21
9	64	– 0.1	8.6	– 0.86	0.01
13	72	3.9	16.6	64.74	15.21
3	36	– 6.1	– 19.4	118.34	37.21
6	43	– 3.1	– 12.4	38.44	9.61
11	59	1.9	3.6	6.84	3.61
21	90	11.9	34.6	411.74	141.61
1	20	– 8.1	– 35.4	286.74	65.61
16	83	6.9	27.6	190.44	47.61
				$\Sigma = 1269.6$	$\Sigma = 358.9$

$$\beta = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

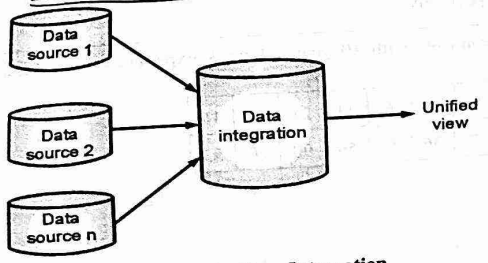
$$= 3.54$$

$$\alpha = \bar{y} - \beta \bar{x} = 55.4 - 3.54 \times 9.1 = 23.19$$

For x =10 years, $y = \alpha + \beta x = 23.19 + 3.54 \times 10 = 58.586$ (in $ 100)

2.7 DATA INTEGRATION

- Data Integration is a data preprocessing technique that combines data from multiple sources and provides users a unified view of these data.



(1B27)**Fig. 2.7.1 : Data Integration**

- These sources may include multiple databases, data cubes, or flat files. One of the most well-known implementation of data integration is building an enterprise's data warehouse.
- The benefit of a data warehouse enables a business to perform analyses based on the data in the data warehouse.
- There are mainly 2 major approaches for data integration:

(a) Tight Coupling :

In tight coupling data is combined from different sources into a single physical location through the process of ETL - Extraction, Transformation and Loading.

(b) Loose Coupling :

In loose coupling data only remains in the actual source databases. In this approach, an interface is provided that takes query from user and transforms it in a way the source database can understand and then sends the query directly to the source databases to obtain the result.

2.7.1 Data Integration Techniques

Below explained the different data integration techniques.

(1) Manual Integration
(2) Middleware Integration
(3) Application-Based Integration
(4) Uniform Access Integration
(5) Data Warehousing

(1) Manual Integration

- This technique avoids the use of automation during data integration. The data analyst himself collects the data, cleans it and integrate it to provide useful information.
- This technique can be implemented for a small organization with a small data set. But it would be tedious for the large, complex and recurring integration because it is a time consuming process as the entire process has to be done manually.

(2) Middleware Integration

- The middleware software is employed to collect the information from different sources, normalize the data and store into the resultant data set. This technique is adopted when the enterprise wants to integrate data from the legacy systems to modern systems.
- Middleware software act as an interpreter between the legacy systems and advanced systems. You can take an example of the adapter which helps in connecting two systems with different interfaces. It can be applied to some system only.

(3) Application-Based Integration

- This technique makes use of software application to extract, transform and load the data from the heterogeneous sources. This technique also makes the data from disparate source compatible in order to ease the transfer of the data from one system to another.
- This technique saves time and effort, but is little complicated as designing such an application requires technical knowledge.

(4) Uniform Access Integration

- This technique integrates data from a more discrepant source. But, here the location of the data is not changed, the data stays in its original location.
- This technique only creates a unified view which represents the integrated data. No separate storage is required to store the integrated data as only the integrated view is created for the end-user.

(5) Data Warehousing

- This technique loosely relates to the uniform access integration technique. But the difference is that the unified view is stored in certain storage. This allows the data analyst to handle more complex queries.
- Though this is a promising technique it has increased storage cost as the view or copy of the unified data needs separate storage and even it has an increase in maintenance cost.

Module 2

2.7.2 Issues in Data Integration

While integrating the data we have to deal with several issues which are discussed below.

(1) Entity Identification Problem
(2) Redundancy and Correlation Analysis
(3) Tuple Duplication
(4) Data Conflict Detection and Resolution

(1) Entity Identification Problem

- As we know, the data is unified from the heterogeneous sources; then how can we 'match the real-world entities from the data'. For example, we have customer data from two different data source. An entity from one data source has customer_id and the entity from the other data source has customer_number. Now how does the data analyst or the system would understand that these two entities refer to the same attribute?
- Well, here the **schema integration** can be achieved using metadata of each attribute. Metadata of an attribute incorporates its name, what does it mean in the particular scenario, what is its data type, up to what range it can accept the value. What rules does the attribute follow for the null value, blank, or zero? Analyzing this metadata information will prevent error in schema integration.
- **Structural integration** can be achieved by ensuring that the functional dependency of an attribute in the source system and its referential constraints matches the functional dependency and referential constraint of the same attribute in the target system.

- This can be understood with the help of an example. Suppose in the one system, the discount would be applied to an entire order but in another system, the discount would be applied to every single item in the order. This difference must be caught before the data from these two sources are integrated into the target system.

(2) Redundancy and Correlation Analysis

- Redundancy is one of the big issues during data integration. Redundant data is an unimportant data or the data that is no longer needed. It can also arise due to attributes that could be derived using another attribute in the data set.
- For example, one data set has the customer age and other data set has the customers date of birth, then age would be a redundant attribute as it could be derived using the date of birth.
- Inconsistencies in the attribute also raise the level of redundancy. The redundancy can be discovered using correlation analysis. The attributes are analyzed to detect their interdependency on each other thereby detecting the correlation between them.
- χ^2 (Chi-square) test is the test to analyze the correlation of nominal data.
- Correlation coefficient and covariance can be used to test the variation between the attributes of numeric data.

(3) Tuple Duplication

- Along with redundancies, data integration has also to deal with the duplicate tuples.
- Duplicate tuples may come in the resultant data if the denormalized table has been used as a source for data integration.

(4) Data Conflict Detection and Resolution

- Data conflict means the data merged from the different sources do not match. Like the attribute values may differ in different data sets. The difference maybe because they are represented differently in the different data sets.
- For example, the price of a hotel room may be represented in different currencies in different cities. This kind of issues are detected and resolved during data integration.

2.7.2(A) χ^2 (Chi-square) Test

- The Chi-square test is also known as the name of the "goodness of fit test".
- The chi-square test helps you to solve the problem in feature selection by testing the relationship between the features.
- χ^2 (Chi-square) test is defined by the formula

$$\chi^2_{df} = \sum \left[\frac{(O - E)^2}{E} \right]$$

where df = degrees of freedom, and

df = (number of rows – 1) × (number of columns – 1)

O = observed value(s)

E = expected value(s)

- Expected value in each cell is calculated as E = (row total × column total)/ table total

Steps in χ^2 (Chi-square) test

(1) Define the hypotheses.

Null Hypothesis (H_0) : Two variables are independent.

Alternate Hypothesis (H_1) : Two variables are not independent.

(2) Create the contingency table.

It is a table showing the distribution of one variable in rows and another in columns. It is used to study the relation between two variables. It contains the observed value O.

Degrees of freedom for contingency table is given as (r – 1) × (c – 1) where r, c are rows and columns.

(3) Find the expected value E.

(4) Calculate the χ^2 (Chi-square) value.

(5) For the given level of significance (or confidence level, α) and the degree of freedom computed in contingency table, find the critical value of χ^2 (Chi-square) from the distribution table.

(6) If the calculated χ^2 (Chi-square) value < critical χ^2 (Chi-square) value, then accept H_0 and reject H_1. Else, reject H_0 and accept H_1.

Ex. 2.7.1 : Assume that an app provides ratings to all the restaurants under 3 categories, good, okay, and not recommended. Now the challenge is to segregate restaurants under correct categories. They can be created under the name of the seating capacity of the restaurant. Small is for a restaurant with a sitting capacity of 20 people, the medium is for sitting capacity of 100 people and large is for sitting capacity of more than 100 people. Results are shown in the contingency table below.

Ratings	Restaurant Size		
	Small	Medium	Large
Good	30	10	20
Okay	8	10	12
Not Recommended	3	5	2

Level of significance is 0.05. Interpret the result.

Soln. :

(1) Define the hypotheses.

Null Hypothesis (H_0) : Ratings and size are independent.

Alternate Hypothesis (H_1) : Ratings and size are not independent.

(2) Create the contingency table.

It contains the observed value O.

Ratings	Restaurant Size			
	Small	Medium	Large	Column Total
Good	30	10	20	60
Okay	8	10	12	30
Not Recommended	3	5	2	10
Row Total	41	25	34	100

Degrees of freedom for contingency table is given as (r – 1) × (c – 1) where r, c are rows and columns.

df = (3 – 1) × (3 – 1) = 4

(3) Find the expected value E.

$$E = \frac{\text{(row total} \times \text{column total)}}{\text{table total}}$$

Ratings	Restaurant Size			
	Small	Medium	Large	Column Total
Good	24.6	15	20.4	60
Okay	12.3	7.5	10.2	30
Not Recommended	4.1	2.5	3.4	10
Row Total	41	25	34	100

(4) Calculate the χ^2 (Chi-square) value.

$$\chi^2_{df} = \sum \left[\frac{(O - E)^2}{E} \right]$$

$$\chi^2_4 = \frac{(30 - 24.6)^2}{24.6} + \frac{(10 - 15)^2}{15} + \frac{(20 - 20.4)^2}{20.4} + \frac{(8 - 12.3)^2}{12.3} + \frac{(10 - 7.5)^2}{7.5} + \frac{(12 - 10.2)^2}{10.2} + \frac{(3 - 4.1)^2}{4.1} + \frac{(5 - 2.5)^2}{2.5} + \frac{(2 - 3.4)^2}{3.4} = 8.88$$

(5) For $\alpha = 0.05$ and df = 4, find the critical value of χ^2 (Chi-square) from the distribution table.

$$\chi^2_{0.05,4} = 9.488$$

(6) Calculated χ^2 (Chi-square) value = 8.88 < critical χ^2 (Chi-square) value = 9.488.

Thus, accept H_0 and reject H_1. This states that, ratings and size are independent of each other.

Module 2

2.7.2(B) The Correlation Coefficient and Covariance

- Covariance and correlation are two measures that can tell you, statistically, whether or not a real relationship exists between two variables.
- **Covariance** is a statistical measure that shows whether two variables are related by measuring how the variables change in relation to each other. This could be **positive covariance**, meaning as one increases the other also increases, or **negative covariance**, meaning that as one increases the other decreases.
- **Correlation**, like covariance, is a measure of how two variables change in relation to each other, but it goes one step further than covariance in that correlation tells how strong the relationship is.
- If correlation > 0, then two variables are positively correlated. The higher value, the stronger correlation.

If correlation = 0, then two variables are independent.

If correlation < 0, then two variables are negatively correlated.

- The covariance between X and Y is defined as

$$\text{Cov}(X, Y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}$$

where, $\bar{x}$ and $\bar{y}$ are the mean value of x and y variables respectively.

- The correlation coefficient is given by,

$$r_{x,y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$$

where, $\sigma_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$,

also called the standard deviation of X.

Ex. 2.7.2 : Assume you are the new owner of a small ice-cream shop in a little village near the beach. You noticed that there was more business in the warmer months than the cooler months as shown in table below. Before you alter your purchasing pattern to match this trend, you want to be sure that the relationship is real. Compute covariance and correlation coefficient to support your assumption.

Temperature	Customers
98	15
87	12
90	10
85	10
95	16
75	7

Soln. :

$$\text{Cov}(X, Y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}$$

$\bar{x} = 88.33$, $\bar{y} = 11.67$

Temperature (x)	Customers (y)	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
98	15	9.67	3.33	93.51	11.09	32.20
87	12	– 1.33	0.33	1.77	0.19	– 0.44
90	10	1.67	– 1.67	2.79	2.79	– 2.79
85	10	– 3.33	– 1.67	11.09	2.79	5.56
95	16	6.67	4.33	44.49	18.75	28.88
75	7	– 13.33	– 4.67	177.69	21.81	62.25

$$\text{Cov}(X, Y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n} = \frac{125.66}{6} = 20.94$$

The covariance of this set of data is 20.94. The number is positive, so we can state that the two variables do have a positive relationship; as temperature rises, the number of customers in the store also rises.

$$\sigma_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}} = \sqrt{\frac{271.34}{6}} = 6.72$$

$$\sigma_x = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n}} = \sqrt{\frac{57.42}{6}} = 3.09$$

$$r_{x,y} = \frac{\text{Cov}(X,Y)}{\sigma_x \sigma_y} = \frac{20.94}{6.72 \times 3.09} = 1$$

Thus, there is a strong linear relation between temperature and number of customers.

2.8 DATA REDUCTION

- A database or date warehouse may store terabytes of data. So it may take very long to perform data analysis and mining on such huge amounts of data.
- Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume but still contain critical information.

2.8.1 Methods of Data Reduction

Different methods of data reduction are explained below.

(A) Data Cube Aggregation
(B) Dimensionality Reduction (Attribute Subset Selection)
(C) Data Compression
(D) Numerosity Reduction
(E) Data Transformation and Discretization

2.8.1 (A) Data Cube Aggregation

- This technique is used to aggregate data in a simpler form. For example, imagine that information you gathered for your analysis for the years 2012 to 2014, that data includes the revenue of your company every three months. They involve you in the annual sales, rather than the quarterly average.
- So we can summarize the data in such a way that the resulting data summarizes the total sales per year instead of per quarter. It summarizes the data.

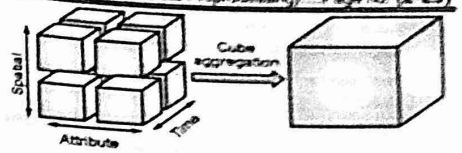


Fig. 2.8.1 : Data Cube Aggregation

2.8.1 (B) Dimensionality Reduction (Attribute Subset Selection)

Module 2

- Whenever we come across any data which is weakly important, then we use the attribute required for our analysis.
- It reduces data size as it eliminates outdated or redundant features.
- Dimensionality reduction can be performed using the below approaches.

(a) Step-wise Forward Selection

- The selection begins with an empty set of attributes.
- Later on we decide best of the original attributes on the set based on their relevance to other attributes.
- This is called as a p-value in statistics.
- **Example :**

Suppose there are the following attributes in the data set in which few attributes are redundant.

Initial attribute Set: {X1, X2, X3, X4, X5, X6}
Initial reduced attribute set: { }
Step-1 : {X1}
Step-2 : {X1, X2}
Step-3 : {X1, X2, X5}
Final reduced attribute set: {X1, X2, X5}

(b) Step-wise Backward Selection

This selection starts with a set of complete attributes in the original data and at each point, it eliminates the worst remaining attribute in the set.

Suppose there are the following attributes in the data set in which few attributes are redundant.

Initial attribute Set: {X1, X2, X3, X4, X5, X6}
Initial reduced attribute set: {X1, X2, X3, X4, X5, X6}
Step-1 : {X1, X2, X3, X4, X5}
Step-2 : {X1, X2, X3, X5}

Step-3 : {X1, X2, X5}
Final reduced attribute set: {X1, X2, X5}

(c) Combination of forwarding and Backward Selection

It allows us to remove the worst and select best attributes, saving time and making the process faster.

2.8.1 (C) Data Compression

The data compression technique reduces the size of the files using different encoding mechanisms (Huffman Encoding& Run Length Encoding). We can divide it into two types based on their compression techniques.

- **Lossless Compression :** Encoding techniques (Run Length Encoding) allows a simple and minimal data size reduction. Lossless data compression uses algorithms to restore the precise original data from the compressed data.
- **Lossy Compression :** Methods such as Discrete Wavelet Transform technique, PCA (principal component analysis) are examples of this compression. For e.g., JPEG image format is a lossy compression, but we can find the meaning equivalent to the original image. In lossy data compression, the decompressed data may differ from the original data, but they are useful enough to retrieve information from them.

(a) Wavelet Transform

- In the wavelet transform, a data vector X is transformed into a numerically different data vector X' such that both X and X' vectors are of the same length. Then how it is useful in reducing data?
- The data obtained from the wavelet transform can be truncated. The compressed data is obtained by retaining the smallest fragment of the strongest of wavelet coefficients.
- Wavelet transform can be applied to data cube, sparse data or skewed data.

(b) Principal Component Analysis

- Let us consider that we have a data set to be analyzed having tuples with n attributes, then the principal component analysis (PCA) identifies k independent tuples with n attributes that can represent the data set.
- In this way, the original data can be cast on a much smaller space and the dimensionality reduction can be achieved.
- Principal component analysis can be applied to sparse and skewed data.

2.8.1 (D) Numerosity Reduction

- Numerosity Reduction is a data reduction technique which replaces the original data by smaller form of data representation.
- There are two techniques for numerosity reduction: **Parametric** and **Non-Parametric** methods.

(a) Parametric Methods

- For parametric methods, data is represented using some model. The model is used to estimate the data, so that only parameters of data are required to be stored, instead of actual data.
- Regression and Log-Linear methods are used for creating such models.
- Regression and log-linear model can both be used on sparse data, although their application may be limited.

(i) Regression

- Regression can be a simple linear regression or multiple linear regression.
- When there is only single independent attribute, such regression model is called simple linear regression and if there are multiple independent attributes, then such regression models are called multiple linear regression.
- In linear regression, the data are modeled to a fit straight line. For example, a random variable y can be modeled as a linear function of another random variable x with the equation $y = ax + b$ where a and b (regression coefficients) specifies the slope and y-intercept of the line, respectively.
- In multiple linear regression, y will be modeled as a linear function of two or more predictor(independent) variables.

(ii) Log-Linear Model

- Log-linear model can be used to estimate the probability of each data point in a multidimensional space for a set of discretized attributes, based on a smaller subset of dimensional combinations.
- This allows a higher-dimensional data space to be constructed from lower-dimensional attributes.

(b) Non-Parametric Methods

- These methods are used for storing reduced representations of the data include histograms, clustering, sampling and data cube aggregation.
- Different techniques used for non-parametric methods are:

(i) Histograms

- Histogram is the data representation in terms of frequency. It uses binning to approximate data distribution and is a popular form of data reduction.
- A histogram partitions the data distribution into disjoint subsets, or buckets.
- If each bucket represents only a single attribute-value/frequency pair, the buckets are called singleton buckets.
- Singleton buckets are useful for storing outliers with high frequency.
- Histograms are highly effective at approximating both sparse and dense data, as well as highly skewed and uniform data.
- The histograms for single attributes can be extended for multiple attributes.
- Multidimensional histograms can capture dependencies between attributes.

(ii) Clustering

- Clustering divides the data into groups/clusters. This technique partitions the whole data into different clusters.
- In data reduction, the cluster representation of the data is used to replace the actual data.
- It also helps to detect outliers in data.

(iii) Sampling

- Sampling can be used for data reduction because it allows a large data set to be represented by a much smaller random data sample (or subset).
- There are four types of sampling data reduction methods:
  - **(a) Simple random sampling :** There is an equal probability of selecting any particular item.
  - **(b) Sampling without replacement :** Once an object is selected, it is removed from the population.
  - **(c) Sampling with replacement :** A selected object is not removed from the population.
  - **(d) Stratified sampling :** Partition the data set, and draw samples from each partition (proportionally, i.e., approximately the same percentage of the data). Used in conjunction with skewed data.

(iv) Data Cube Aggregation

- Data cube aggregation involves moving the data from detailed level to a fewer number of dimensions.
- The resulting data set is smaller in volume, without loss of information necessary for the analysis task.

Module 2

2.8.1 (E) Data Transformation and Discretization

- Data transformation in data mining is done for combining unstructured data with structured data to analyze it later. It is also important when the data is transferred to a new cloud data warehouse.
- When the data is homogeneous and well-structured, it is easier to analyze and look for patterns.
- For example, a company has acquired another firm and now has to consolidate all the business data. The smaller company may be using a different database than the parent firm. Also, the data in these databases may have unique IDs, keys and values. All this needs to be formatted so that all the records are similar and can be evaluated.
- This is why data transformation methods are applied. And, they are described below :

(i) Data Smoothing

- This method is used for removing the noise from a dataset. Noise is referred to as the distorted and meaningless data within a dataset.
- Smoothing uses algorithms to highlight the special features in the data.
- After removing noise, the process can detect any small changes to the data to detect special patterns.
- Any data modification or trend can be identified by this method.

(ii) Data Aggregation

- Aggregation is the process of collecting data from a variety of sources and storing it in a single format.

Here, data is collected, stored, analyzed and presented in a report or summary format.

- It helps in gathering more information about a particular data cluster. The method helps in collecting vast amounts of data.
- This is a crucial step as accuracy and quantity of data is important for proper analysis.
- Companies collect data about their website visitors. This gives them an idea about customer demographics and behaviour metrics. This aggregated data assists them in designing personalized messages, offers and discounts.

(iii) Discretization

- This is a process of converting continuous data into a set of data intervals. Continuous attribute values are substituted by small interval labels. This makes the data easier to study and analyze.
- If a continuous attribute is handled by a data mining task, then its discrete values can be replaced by constant quality attributes. This improves the efficiency of the task.
- This method is also called data reduction mechanism as it transforms a large dataset into a set of categorical data.
- Discretization can be done by Binning, Histogram Analysis, and Correlation Analyses.
- Discretization also uses decision tree-based algorithms to produce short, compact and accurate results when using discrete values.

(iv) Generalization

- In this process, low-level data attributes are transformed into high-level data attributes using concept hierarchies. This conversion from a lower level to a higher conceptual level is useful to get a clearer picture of the data.
- For example, age data can be in the form of (20, 30) in a dataset. It is transformed into a higher conceptual level into a categorical value (young, old).
- Data generalization can be divided into two approaches – data cube process (OLAP) and attribute oriented induction approach (AOI).

(v) Attribute construction

- In the attribute construction method, new attributes are created from an existing set of attributes.
- For example, in a dataset of employee information, the attributes can be employee name, employee ID and address.
- These attributes can be used to construct another dataset that contains information about the employees who have joined in the year 2019 only.
- This method of reconstruction makes mining more efficient and helps in creating new datasets quickly.

(vi) Normalization

- Also called data pre-processing, this is one of the crucial techniques for data transformation in data mining.
- Here, the data is transformed so that it falls under a given range. When attributes are on different ranges or scales, data modelling and mining can be difficult.
- Normalization helps in applying data mining algorithms and extracting data faster.

The popular normalization methods are:

(a) Min-max normalization

In this technique of data normalization, linear transformation is performed on the original data. Minimum and maximum value from data is fetched and each value is replaced according to the following formula:

$$v' = \frac{v - min_A}{max_A - min_A} (new\_max_A - new\_min_A) + new\_min_A$$

for 0 to 1 new range.

$$v' = \frac{v - min_A}{max_A - min_A} (1 - 0) + 0$$

Where A is the attribute data,

min_A, max_A are the minimum and maximum absolute value of A respectively,

v' is the new value of each entry in data,

v is the old value of each entry in data,

new_max_A, new_min_A is the max and min value of the range (i.e. boundary value of range required) respectively.

Example : Suppose the income range from $10,000 to $95,000 is normalized to [0.0, 1.0]. By min-max normalization, a value of $64,300 for income is transformed to

$$\frac{64300 - 10000}{95000 - 10000}(1.0 - 0.0) + 0.0 = 0.6388$$

(b) Z-score normalization

In this technique, values are normalized based on mean and standard deviation of the data A. The formula used is:

$$v' = \frac{v - \mu_A^{mean}}{\sigma_{A^{S.d.}}}$$

where v', v is the new and old value of each entry in data respectively. σ_A, μ_A is the standard deviation and mean of A respectively.

Example: If mean salary is $54,000 and standard deviation is $16,000, then the z-score value of salary $73,600 will be $\frac{73600 - 54000}{16000} = 1.225$

(c) Decimal scaling

It normalizes by moving the decimal point of values of the data. To normalize the data by this technique, we divide each value of the data by the maximum absolute value of data. The data value, v_i, of data is normalized to v_i' by using the formula below:

$$v_i' = \frac{v_i}{10^j}$$

where, j is the smallest integer such that $max(|v_i'|) < 1$.

Example :

Let the input data be: -10, 201, 301, – 401, 501, 601, 701

To normalize the above data,

- **Step 1 :**

Maximum absolute value in given data(m): 701

- **Step 2 :**

Divide the given data by 1000 (i.e. j = 3)

Result : The normalized data is :

– 0.01, 0.201, 0.301, – 0.401, 0.501, 0.601, 0.701

Ex. 2.8.1 : Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order): 13, 15, 16, 16, 19, 20, 23, 29, 33, 41, 44, 53, 62, 69, 72

Use min-max normalization to transform the value 45 for age onto the range [0.0, 1.0].

☑ **Soln. :**

By min-max normalization,

$$v' = \frac{v - min_A}{max_A - min_A} (new\_max_A - new\_min_A) + new\_min_A$$

where $min_A = 13$, $max_A = 72$, $v = 45$, $new\_max_A = 1.0$, $new\_min_A = 0.0$

$$v' = \frac{45 - 13}{72 - 13}(1.0 - 0.0) + 0.0 = 0.5423$$

Therefore, the value 45 for age is transformed to 0.5423

2.9 CONCEPT HIERARCHY GENERATION

- Concept hierarchies can be used to reduce the data by collecting and replacing low-level concepts with higher-level concepts. For example, the numeric value for age may be represented as Young, Middle-aged or Senior.
- In the multidimensional model, data are organized into multiple dimensions, and each dimension contains multiple levels of abstraction defined by concept hierarchies. This organization provides users with the flexibility to view data from different perspectives.
- Data mining on a reduced data set means fewer input/output operations and is more efficient than mining on a larger data set.
- Because of these benefits, discretization techniques and concept hierarchies are typically applied before data mining, rather than during mining.

Module 2