DMI COLLEGE OF ENGINEERING
DEPARTMENT OF INFORMATION TECHNOLOGY
UNIT1 INTRODUCTION
Part A & B Question and Answers

1.Define Data Science

Data science involves methods

(i)to **analyze massive amounts of data** and

(ii)extract the knowledge it contains.

2.Difference between Traditional data and Big data

Tradational Data	**Big Data**
Traditional data is **generated in enterprise level.**	Big data is generated **outside the enterprise level.**
Its volume ranges from **Gigabytes to Terabytes.**	Its volume ranges from **Petabytes to Zettabytes or Exabytes.**
Traditional database system **deals with structured data.**	Big data system deals with ***structured, semi-structured,database, and unstructured data.***

3.What is streaming data?

- While **streaming(Flowing) data** can take ***almost any of the previous forms***, it has an extra property
- The **data flows** into the system when an **event happens** instead of **being loaded into a data store in a batch**
- This isn't really ***a different type of data***,we treat it here as such because you need **to adapt your process to deal** with this type of information. Examples are the **"What's trending" on Twitter, live sporting or music events, and the stock market**

4.List the categories of the data

In **data science and big data** you'll come across many different types of data, and each of them tends **to require different tools and techniques**.

The **main categories** of data are these:

- Structured data
- Unstructured data
- Natural language
- Machine-generated
- Graph-based data
- Audio, video, and images
- Streaming data

5.List the steps of the data science process.

- Data science helps you to maximize your ***chances of success in a data science project at the lowest cost***.
- It also makes it **possible to take up a project as a team, with each team member focusing** on what **they do best.**
- This approach ***may not be suitable for every type of project*** or be the only way **to do good data science**

The data science process typically consists of six steps

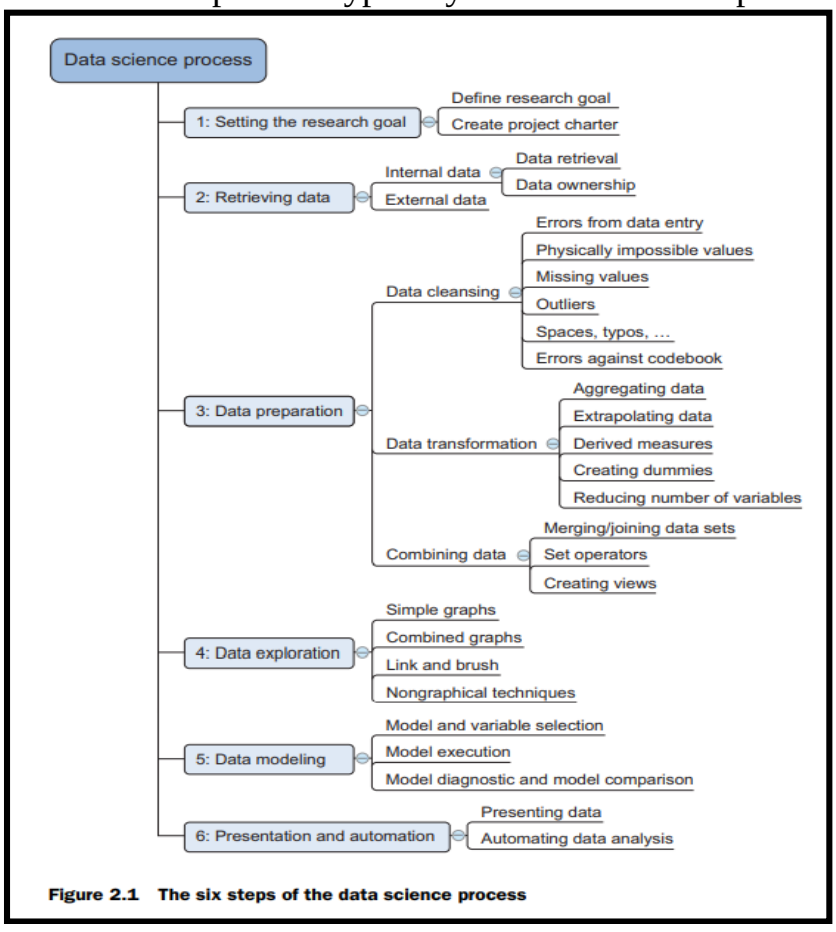


Figure 2.1 The six steps of the data science process

6.List the techniques to handle the missing data.

Table 2.4 An overview of techniques to handle missing data

Technique	Advantage	Disadvantage
Omit the values	Easy to perform	You lose the information from an observation
Set value to `null`	Easy to perform	Not every modeling technique and/or implementation can handle `null` values
Impute a static value such as 0 or the mean	Easy to perform You don't lose information from the other variables in the observation	Can lead to false estimations from a model
Impute a value from an estimated or theoretical distribution	Does not disturb the model as much	Harder to execute You make data assumptions
Modeling the value (nondependent)	Does not disturb the model too much	Can lead to too much confidence in the model Can artificially raise dependence among the variables Harder to execute You make data assumptions

7.Define view

Views in SQL are kind of **virtual tables**. A **view** also has rows and columns as they are in a real table in the **database**.

8.Define Box plot

- The **boxplot,** on the other hand, doesn't show how many observations are present but does offer an impression of the distribution within categories.
- It can **show the maximum, minimum, median,** and other characterizing measures at the same time.

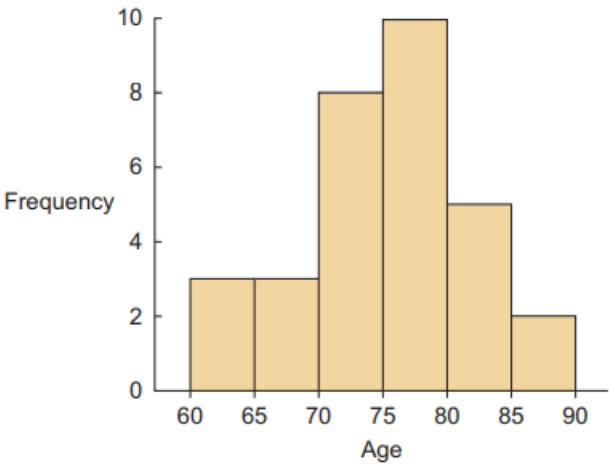


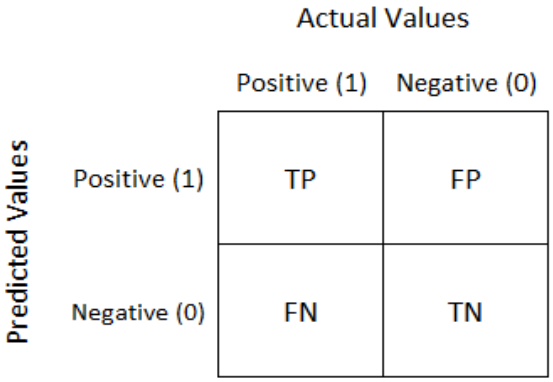
Figure 2.19 Example histogram: the number of people in the age-groups of 5-year intervals

9.Define Brushing and Linking

With brushing and linking ,we **combine and link different graphs and tables** (or views) so **changes in one graph** are **automatically transferred** to the **other graphs**

10.What is confusion matrix .Write the need of confusion matrix?

A confusion matrix is **a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known**



11.Define data mart

A data mart is a **subset of a data warehouse** focused on a particular line of business, department, or subject area.

12.List the open data providers to get the data

Table 2.1 A list of open-data providers that should get you started

Open data site	Description
Data.gov	The home of the US Government's open data
https://open-data.europa.eu/	The home of the European Commission's open data
Freebase.org	An open database that retrieves its information from sites like Wikipedia, MusicBrains, and the SEC archive
Data.worldbank.org	Open data initiative from the World Bank
Aiddata.org	Open data for international development
Open.fda.gov	Open data from the US Food and Drug Administration

13.Define ETL , Outliers

ETL(Extract,Transform and Load)

ETL stands for extract, transform, and load and is a traditionally accepted way for organizations to combine data from multiple systems into a single database, data store, data warehouse, or data lake.

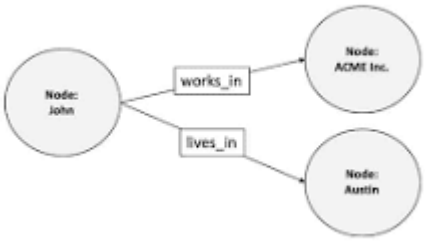
Outliers

The outlier is **an observation that lies an abnormal distance from other values in a random sample from a population**

14.Define Graph Database

A graph database stores nodes and relationships instead of tables, or documents.
Data is stored just like you might sketch ideas on a whiteboard.

Example: **Netflix uses Graph Database for its Digital Asset Management** because it is a perfect way to track which movies (assets) each viewer has already watched, and which movies they are allowed to watch (access management).



15.Exploratory data analysis

During exploratory data analysis you take a deep dive into the data

- Information becomes much **easier to grasp** when shown in a picture, therefore you mainly **use graphical techniques to gain an understanding of your data** and the interactions between variables.
- ❖ The **goal isn't to cleanse the data**, but it's common that **you'll still discover anomalies** you missed before, forcing you to take a step back and fix them.
- ❖ The **visualization techniques** you use in this **phase range from simple line graphs or histograms**, as shown in figure 2.15, to more complex diagrams such as Sankey and network graphs.
- ❖ Sometimes it's useful **to compose a composite graph** from **simple graphs to get even more insight into the data.** Other times the graphs can be animated or made interactive to make it easier and, let's admit it, way more fun.

16.Define DataWare House

A data warehouse is a **subject-oriented, integrated, time-variant and non-volatile collection** of data in support of management's decision making process.

Subject-Oriented: A data warehouse can be used to analyze a particular subject area. For example, "sales" can be a particular subject.

Integrated: A data warehouse integrates data from multiple data sources. For example, source A and source B may have different ways of identifying a product, but in a data warehouse, there will be only a single way of identifying a product.

Time-Variant: Historical data is kept in a data warehouse. For example, one can retrieve data from 3 months, 6 months, 12 months, or even older data from a data warehouse. This contrasts with a transactions system, where often only the most recent data is kept. For example, a transaction system may hold the most recent address of a customer, where a data warehouse can hold all addresses associated with a customer.

Non-volatile: Once data is in the data warehouse, it will not change. So, historical data in a data warehouse should never be altered.

17.Define Meta Data

Metadata is **"data that provides information about other data"**

18.What is Data Modeling?

Data modeling is **the process of analyzing and defining all the different data your business collects and produces, as well as the relationships between those bits of data**.

19.What is Data cleansing?

❖ Data cleansing is a **subprocess of the data science process** that focuses on **removing errors in your data so** your data becomes a **true and consistent(standard) representation of the processes** it originates from.

❖ By "true and consistent representation" we imply that at least **two types of errors** exist. **(i)Interpretation Error (ii)Inconsistencies between data *sources***.

20.Write the python code for k-nearest neighbor classification

```
from sklearn import neighbors
predictors = np.random.random(1000).reshape(500,2)
target = np.around(predictors.dot(np.array([0.4, 0.6])) +
         np.random.random(500))
clf = neighbors.KNeighborsClassifier(n_neighbors=10)
knn = clf.fit(predictors,target)
knn.score(predictors, target)
```

Imports modules.

Creates random predictor data and semi-random target data based on predictor data.

Fits 10-nearest neighbors model.

Gets model fit score: what percent of the classification was correct?

21.What is OLAP

OLAP(Online analytical Processing): OLAP is an approach to answering multi-dimensional analytical (MDA) queries swiftly. OLAP is part of the broader category of business intelligence, which also encompasses relational database, report writing and data mining. OLAP tools enable users to analyze multidimensional data interactively from multiple perspectives.

22.Define Data Mining

Data mining refers to extracting or mining knowledge from large amounts of data.

Properties of Data Mining

The key properties of datamining are

- Automatic discovery of patterns
- Prediction of likely outcomes
- Creation of actionable information
- Focus on large datasets and databases

PART B QUESTIONS

1.Explain the over view of data science process with neat diagram

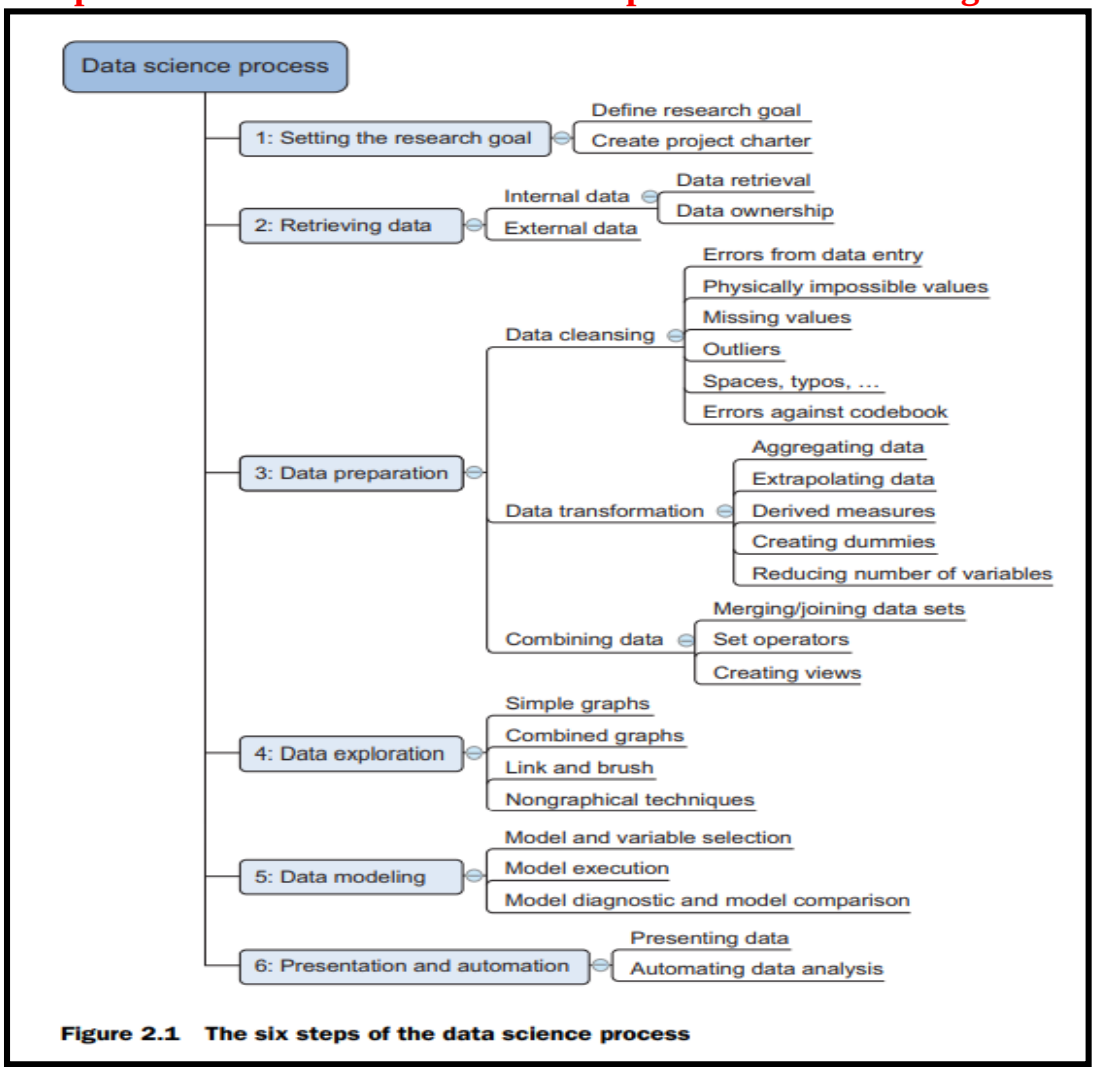
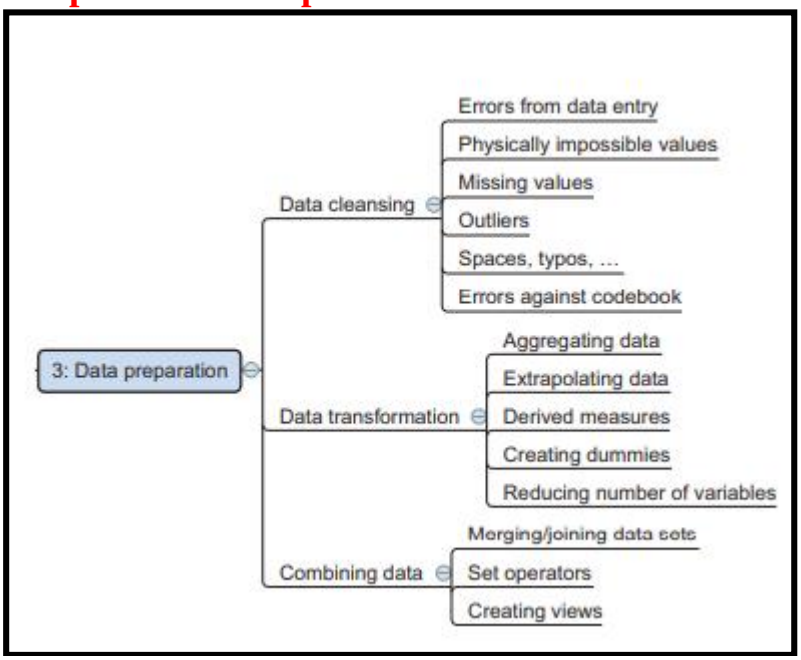


Figure 2.1 The six steps of the data science process

- Explanation for the six steps

2.Explain Data Preparation in detail



- Explanation for the data preparation steps.

3.What is Data Warehousing?Explain DataWarehousing with its architecture?

Definition

Data Warehouse Design Process:

A data warehouse can be built using a *top-down approach*, a *bottom-up approach*, or a *combination of both*.

- The top-down approach starts with the overall design and planning. It is useful in cases where the technology is mature and well known, and where the business problems that must be solved are clear and well understood.

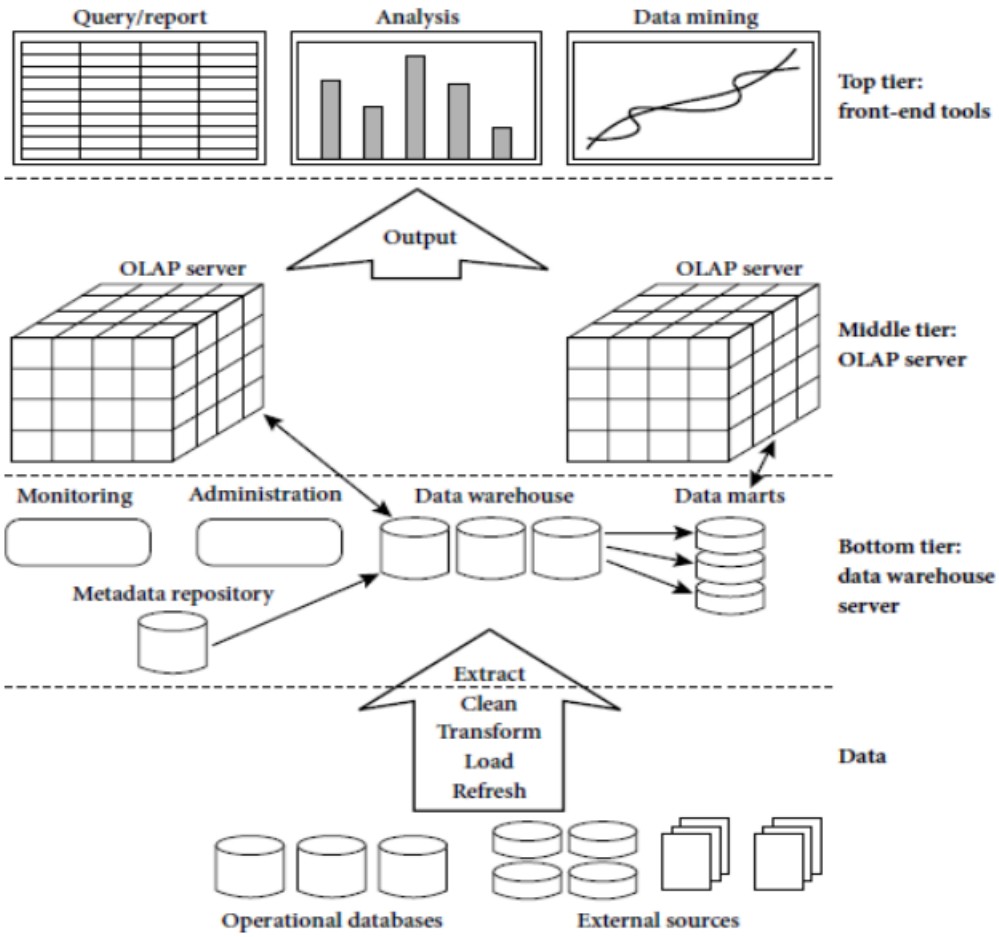
- The bottom-up approach starts with experiments and prototypes. This is useful in the early stage of business modeling and technology development. It allows an organization to move forward at considerably less expense and to evaluate the benefits of the technology before making significant commitments.

- In the combined approach, an organization can exploit the planned and strategic nature of the top-down approach while retaining the rapid implementation and opportunistic application of the bottom-up approach.

The warehouse design process consists of the following steps:

- Choose a business process to model, for example, orders, invoices, shipments, inventory, account administration, sales, or the general ledger. If the business process is organizational and involves multiple complex object collections, a data warehouse model should be followed. However, if the process is departmental and focuses on the analysis of one kind of business process, a data mart model should be chosen.
- Choose the grain of the business process. The grain is the fundamental, atomic level of data to be represented in the fact table for this process, for example, individual transactions, individual daily snapshots, and so on.
- Choose the dimensions that will apply to each fact table record. Typical dimensions are time, item, customer, supplier, warehouse, transaction type, and status.
- Choose the measures that will populate each fact table record. Typical measures are numeric additive quantities like dollars sold and units sold.

A Three Tier Data Warehouse Architecture:



Tier-1:

The bottom tier is a warehouse database server that is almost always a relationaldatabase system. Back-end tools and utilities are used to feed data into the bottomtier from operational databases or other external sources (such as customer profileinformation provided by external consultants). These tools and utilities performdataextraction, cleaning, and transformation (e.g., to merge similar data from differentsources into a unified format), as well as load and refresh functions to update thedata warehouse . The data are extracted using application programinterfaces known as gateways. A gateway is

supported by the underlying DBMS andallows client programs to generate SQL code to be executed at a server.

Examplesof gateways include ODBC (Open Database Connection) and OLEDB (Open Linkingand Embedding for Databases) by Microsoft and JDBC (Java Database Connection).

This tier also contains a metadata repository, which stores information aboutthe data warehouse and its contents.

Tier-2:

The middle tier is an OLAP server that is typically implemented using either a relational OLAP (ROLAP) model or a multidimensional OLAP.

- OLAP model is an extended relational DBMS thatmaps operations on multidimensional data to standard relational operations.
- A multidimensional OLAP (MOLAP) model, that is, a special-purpose server that directly implements multidimensional data and operations.

Tier-3:

The top tier is a front-end client layer, which contains query and reporting tools, analysis tools, and/or data mining tools (e.g., trend analysis, prediction, and so on).

1. Enterprise warehouse:

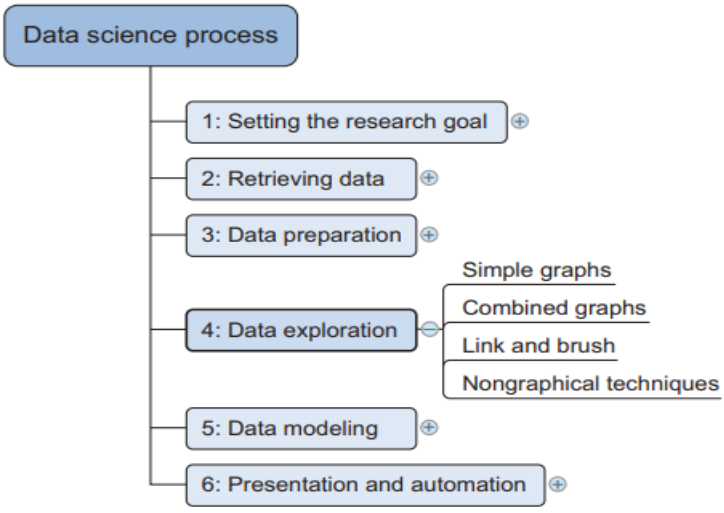
- An enterprise warehouse collects all of the information about subjects spanning the entire organization.
- It provides corporate-wide data integration, usually from one or more operational systems or external information providers, and is cross-functional in scope.
- It typically contains detailed data aswell as summarized data, and can range in size from a few gigabytes to hundreds of gigabytes, terabytes, or beyond.
- An enterprise data warehouse may be implemented on traditional mainframes, computer superservers, or parallel architecture platforms. It requires extensive business modeling and may take years to design and build.

2. Data mart:

- A data mart contains a subset of corporate-wide data that is of value to aspecific group of users. The scope is confined to specific selected subjects. For example,a marketing data mart may confine its subjects to customer, item, and sales. Thedata contained in data marts tend to be summarized.

- Data marts are usually implemented on low-cost departmental servers that areUNIX/LINUX- or Windows-based. The implementation cycle of a data mart ismore likely to be measured in weeks rather than months or years. However, itmay involve complex integration in the long run if its design and planning werenot enterprise-wide.

4.After cleaning ,how the data is explored explain in detail



- ❖ Graph diagram
- ❖ Brushing and linking
- ❖ Boxplot

5.Define Data Mining and write the steps in the process of knowledge discovery.

Define Data Mining

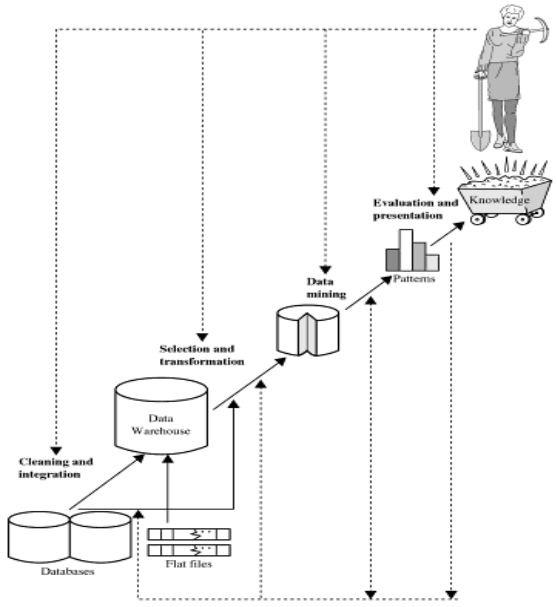


Figure Data mining as a step in the process of knowledge discovery.

The steps involved in datamining when viewed as process of knowledge discovery are as follows:

1.data cleaning (to remove noise or irrelevant data),

2.data integration (where multiple data sources may be combined)1,

3.data selection (where data relevant to the analysis task are retrieved from the database),

4. data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance)2,

5.data mining (an essential process where intelligent methods are applied in order to extract data patterns),

6.pattern evaluation

Pattern evaluation is used to identify the truly interesting patterns representing knowledge based

on some interesting measures

7.knowledge presentation

Knowledge representation techniques are used to present the mined knowledge to the user

6.Explain how data are presented and Analyzed

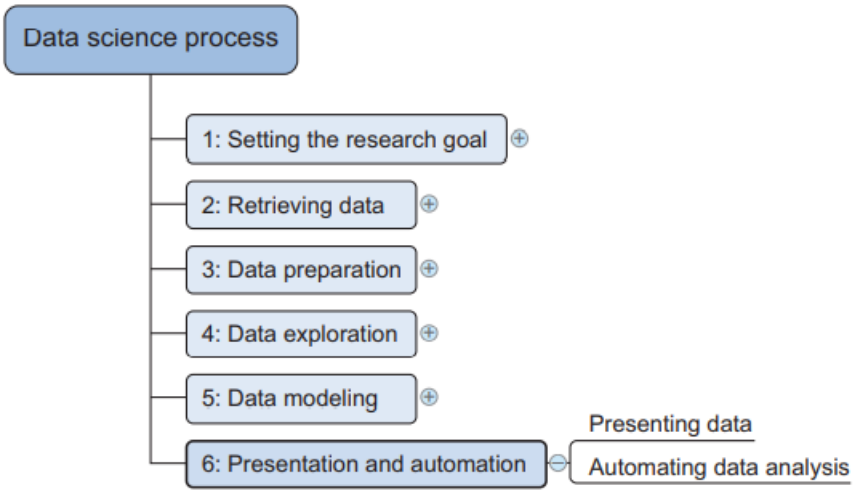


Figure 2.28 Step 6: Presentation and automation

7. Explain the Facets of data

The **main categories** of data are these:

- ❖ Structured
- ❖ Unstructured
- ❖ Natural language
- ❖ Machine-generated
- ❖ Graph-based
- ❖ Audio, video, and images
- ❖ Streaming

8.Explain how data modeling in done

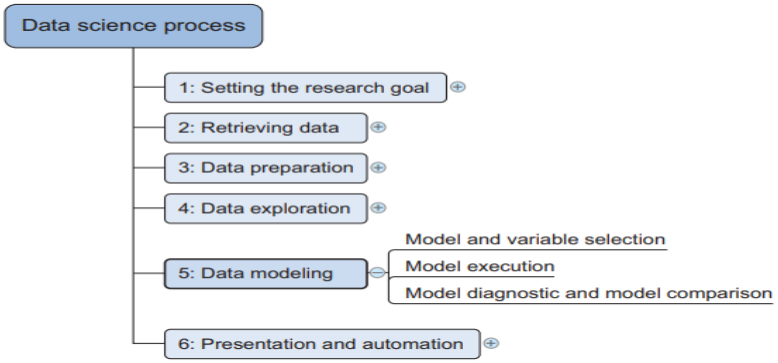


Figure 2.21 Step 5: Data modeling

DMI COLLEGE OF ENGINEERING

DEPARTMENT OF INFORMATION TECHNOLOGY

UNIT2--DESCRIBING DATA

Part A Question And B Answers

1. What is Descriptive and Inferential Statistics

Descriptive Statistics

The descriptive method of statistics is used to ***describe the data collected and summarize the data*** and its **properties** using the ***measures of central tendencies and the measures of dispersion***

Inferential Statistics

- Inferential statistics can be defined as a field of statistics that uses analytical tools for **drawing conclusions about a population** by examining random samples.
- The goal of inferential statistics is to make generalizations about a population. In inferential statistics, a statistic is taken from the sample data (e.g., the sample mean) that used to make inferences about the population parameter (e.g., the population mean).

2. Define Data and types of Data

Data is a collection of ***actual observations or scores*** in a survey or an experiment.

Statistical analysis often depends on ***whether data are qualitative, ranked, or quantitative.***

3.Define Variable and types of variable?

A variable is a ***characteristic or property*** that can take **on different values**

Constant

A **characteristic or property** that can take on **only one value.**

Independent Variable

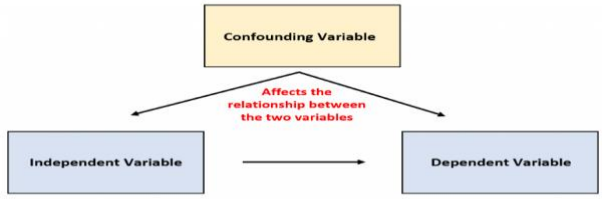
In an experiment, **an independent variable** is the ***treatment manipulated by the investigator***.

Dependent Variable

A variable that is believed to have **been influenced by the independent variable.**

4. What is confounding Variable?

An uncontrolled variable that compromises the interpretation of a study.



5.What is Frequency Distribution and list the uses of Frequency Distribution?

A **<u>frequency distribution is a collection of observations</u>** produced by **sorting observations** into ***<u>classes and showing their frequency (f) of occurrence</u>*** in each class.

<u>Uses of Frequency Distribution</u>

It is **widely used data reduction technique** in ***<u>descriptive statistics</u>***

6.What is stem and leaf display

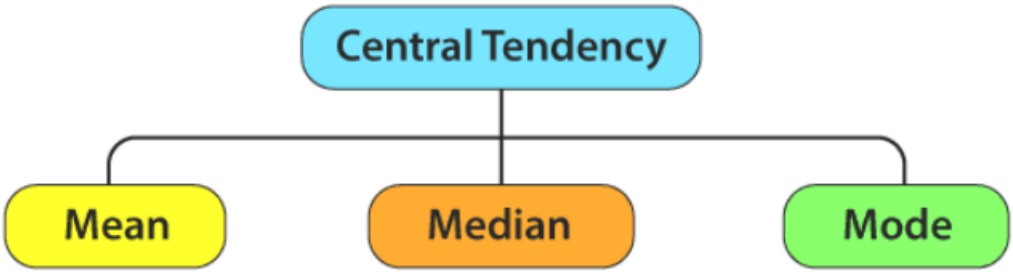
A device for sorting quantitative data on the basis of leading and trailing digits.

- It is a **technique for summarizing quantitative data**
- **Stem and leaf**displays are ***<u>ideal for summarizing distributions</u>***, such as that **for weight**
- **data, without destroying the identities of individual observations.**

7.What is Measures of Central Tendency .List the Measures of Central Tendency in detail

<u>Definition</u>

<u>Numbers or words that attempt to describe</u>, most generally, ***<u>the middle or typical value for a distribution.</u>***



8.Define variance and standard deviation

VARIANCE:

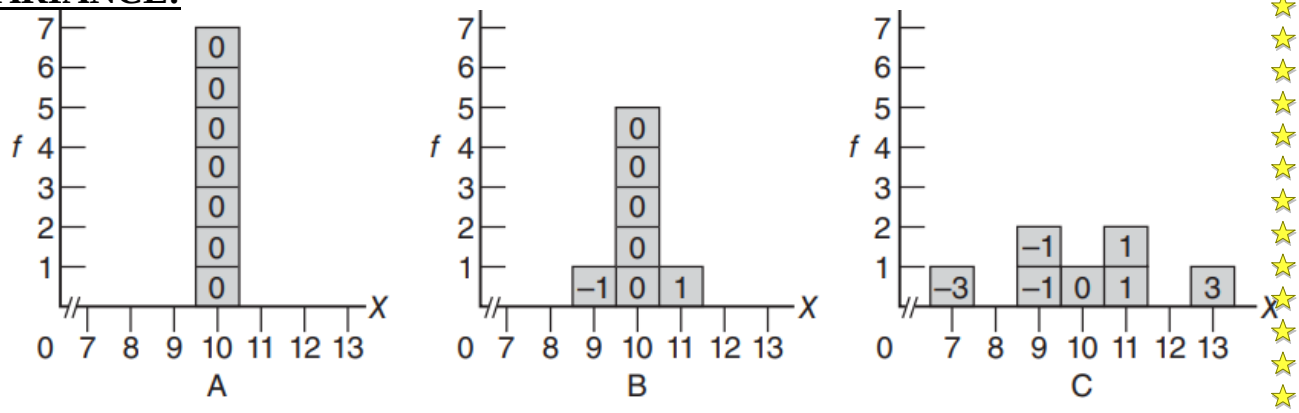


FIGURE 4.1

Three distributions with the same mean (10) but different amounts of variability. Numbers in the boxes indicate distances from the mean.

- Each **original score** is re-expressed ***as a distance or deviation*** from **the mean by subtracting the mean**.
- For each of the three distributions in Figure 4.1, the **face values of the seven original scores** (shown as numbers along the X axis) have been **re-expressed as deviation scores** from **their mean of 10** (shown as numbers in the boxes). For example, ***in distribution C, one score coincides*** with **the mean of 10**.

9.Define IQR and write the need of IQR

Table 4.6
CALCULATION OF THE IQR

A. INSTRUCTIONS

1. Order scores from least to most.
2. To determine how far to penetrate the set of ordered scores, begin at either end, then add 1 to the total number of scores and divide by 4. If necessary, round the result to the nearest whole number.
3. Beginning with the largest score, count the requisite number of steps (calculated in step 2) into the ordered scores to find the location of the third quartile.
4. The third quartile equals the value of the score at this location.
5. Beginning with the smallest score, again count the requisite number of steps into the ordered scores to find the location of the first quartile.
6. The first quartile equals the value of the score at this location.
7. The IQR equals the third quartile minus the first quartile.

B. EXAMPLE

1. 7, 9, 9, 10, 11, 11, 13
2. $(7 + 1)/4 = 2$
3. 7, 9, 9, 10, 11, 11, 13
   2 1
4. third quartile = 11
5. 7, 9, 9, 10, 11, 11, 13
   1 2
6. first quartile = 9
7. IQR = 11 − 9 = 2

Need of IQR

- IQR is its ***resistance to the distorting*** effect of ***extreme scores, or outliers.[To find the Outliers]***
- IQR used as the **measure of variability**, along ***with the median (or second quartile) as the measure of central tendency***

10.What is Normal curve or Normal Distribution and standard normal curve

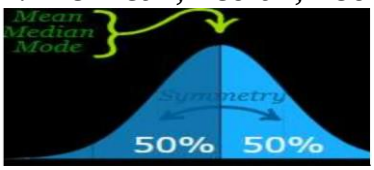
NORMAL CURVE:

Most of the **datas are clustered** around the central value (Mean) So it is called as Normal .

Properties of the Normal Curve

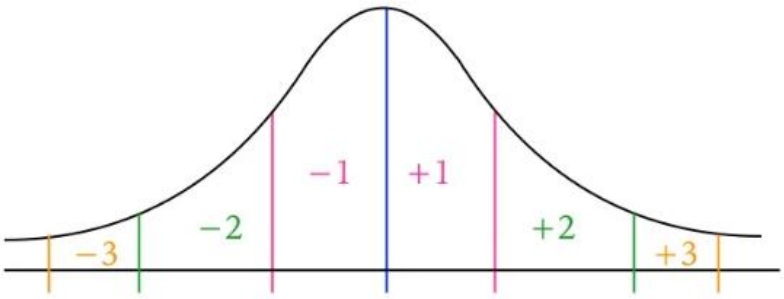
1.Normal curve is a **theoretical curve** defined for ***a continuous variable.***its **symmetrical bell-shaped** form.

2.The mean,median,mode are equal for normal distribution.



standard normal curve:

- The standard deviation is how spread out the numbers are from the middle value.
- Data is said to fall within a specific number of standard deviations when it is not the middle value.
- A normal distribution follows the 68-95-99.7 rule.



11.What is z-score

A z score is a ***unit-free, standardized score*** that, regardless of the **original units of measurement,** indicates ***how many standard deviations a score is above or below the mean of its distribution***

z SCORE

$$z = \frac{X - \mu}{\sigma}$$

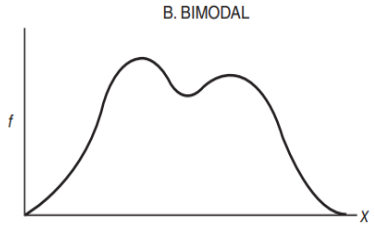
12.What is Bimodal Distribution

Bimodal Distribution: Two Peaks.

Data distributions in statistics can ***have one peak***, or they can **have several peaks**. The type of distribution you might be familiar with seeing is the **normal distribution, or bell curve, which has one peak.** The **bimodal distribution has two peaks**.

Definition

Bimodal(Bi Means two ,Modal is mode)



13,What is degrees of Freedom?

DEGREES OF FREEDOM (d f)

Degrees of freedom (df) refers to (i)**the number of values** that are **free to vary**, (ii)given one or **more mathematical restrictions**, in a **sample** being (iii)used **to estimate a population characteristic.**

14.Difference between Quantitative and Qualitative Data

Qualitative Data[Describes the quality of things]

- A ***set of observations*** where ***any single observation is a word, letter, or numerical code that represents a class or category.***

It cannot be measured in the form of numbers

Quantitative data[It talks about the quantity of things]

It consists of numbers (weights of 238, 170, . . . 185 lbs) that represent an **amount or a count**.

15.What is relative and cumulative Frequency Distribution?

RELATIVE FREQUENCY DISTRIBUTIONS

Relative frequency distributions shows**the frequency of each class** as a part or ***fraction of the total frequency for the entire distribution***.

Cumulative Frequency Distribution

A frequency distribution showing the total number of observations in each class and all lower-ranked classes.

Cumulative percentages are often referred to as ***percentile ranks***

16.What is histogram and list the features of histogram?

Definition

A bar-type graph for **quantitative data**. The common boundaries**between adjacent bars emphasize the continuity of the data, as with continuous *variables***

17,What is frequency polygon?

Definition

- A **line graph for quantitative data** that also emphasizes **the continuity of continuous variables.**
- **Frequency polygons** may be constructed **directly from frequency distributions.**

18.List the features of normal distribution?

features of the normal distribution:

- symmetrical shape
- **mode, median and mean** are the same and are together **in the centre of the curve**
- there can only be **one mode** (i.e. there is only one value which is most frequently observed)
- Most of the ***data are clustered around the centre***, while the more extreme values on either side of the centre become ***less rare as the distance from the centre increases*** (i.e. About 68% of values lie within one standard deviation (σ) away from the mean; about 95% of the values lie within two standard deviations; and about 99.7% are within three standard deviations. This is known as the *empirical rule* or the *3-sigma rule*.)

19.List the measures of variability?

Definition:

Descriptions of the amountby which scores are ***dispersed or scattered in a distribution***

Several **measures of variability**, including

The range
The interquartile range
The variance
and The standard deviation

20.How the Z-scores are interpreted?

Here is **how to interpret z-scores**:

- A **z-score of less than 0** represents an **element less than the mean**.
- A **z-score greater than 0** represents **an element greater than the mean**.
- A **z-score equal to 0** represents **an element equal to the mean.**
- A **z-score equal to 1** represents an element, **which is 1 standard deviation greater than the mean;** a **z-score equal to 2 signifies 2 standard deviations greater than the mean; etc.**
- A **z-score equal to -1** represents an element, which is **1 standard deviation less than the mean**; a z-score equal to **-2 signifies 2 standard deviations less than the mean; etc.**

- If the number of elements in the **set is large**, about **68% of the elements have a z-score between -1 and 1**; about **95% have a z-score between -2 and 2** and about **99% have a z-score between -3 and 3.**

21.Define nominal Measurement

Qualitative data consist of words, letters, or codes that represent only classes with nominal measurement.

22.Define Ordinal Measurement

Ranked data consist of numbers that represent relative standing with ordinal measurement.

23.What is Interval/ratio Measurement

Quantitative data consist of numbers that represent an amount or a count with interval/ratio measurement.

24.What is level of Measurement

Specifies the extent to which a number (or word or letter) actually represents some attribute and, therefore, has implications for the appropriateness of various arith metic operations and statistical procedures.

25.List the Distinctive properties of the three levels of measurement

Distinctive properties of the three levels of measurement are classification (nomi nal), order (ordinal), and equal intervals and true zero (interval/ratio).

26.What is Random sampling and Random Assignment

Random Sampling-A selection process that guarantees all potential observations in the population have an equal chance of being selected.

Random Assignment A procedure designed to ensure that each subject has an equal chance of being assigned to any group in an experiment

PART B QUESTIONS

1.Define Data and its types

- Data definition
- Qualitative Data
- Ranked Data
- Quantitative Data
- Example

2.Explain various types of variables

- Define variables
- Discrete and Continuous Variables
- Example
- Independent and Dependent Variable
- Confounding Variable

3.Explain how datas are described with Tables

- **FREQUENCY DISTRIBUTIONS FOR QUANTITATIVE DATA**
- **Define frequency disrtibutions**
- **Uses of Frequency Distribution**
- **Frequency Distribution for Grouped Data**
- **2GUIDELINES FOR FREQUENCY DISTRIBUTIONS**
  - **Gaps between Classes**
- **OUTLIER**
- **RELATIVE FREQUENCY DISTRIBUTIONS**
- **CUMULATIVE FREQUENCY DISTRIBUTIONS**
- **FREQUENCY DISTRIBUTIONS FOR QUALITATIVE (NOMINAL) DATA**

4. Explain Normal Distributions and Standard (z) Scores

Define Normal Distribution
Example
Zscore
Uses

5. Explain how datas are described with Graphs

- Define Histogram
- Uses of Histogram
- Frequency polygon
- Stem and leaf
- Shapes
- Steps for construction of Graphs

6. Explain Data with Average

- Measures of Central Tendency
- Mean
- Mode
- Medain
- Averages for qualitative and ranked data

DMI COLLEGE OF ENGINEERING

DEPARTMENT OF INFORMATION TECHNOLOGY

UNIT III DESCRIBING RELATIONSHIPS

PART A QUESTIONS

1. Define Correlation.Write a code to find correlation

Defintion

Correlation is a **statistical technique** used to ***analyze the relationships*** between variables.

It is used to ***measure the intensity of relationships*** between variables in your data

Python code:

```
import numpy as np

np.random.seed(100)

#create array of 50 random integers between 0 and 10
var1 = np.random.randint(0, 10, 50)

#create a positively correlated array with some random noise
var2 = var1 + np.random.normal(0, 10, 50)

#calculate the correlation between the two arrays
np.corrcoef(var1, var2)
```

2. **Define scatter plot and write the uses of scatter plot**

SCATTERPLOTS:

A scatterplot is a graph containing **a cluster of dots** that represents **all pairs of scores.**

Uses of Scatter Plot

- Scatter plots are used to observe ***relationships between variables***.
- When we have ***paired numerical data***
- When there are **multiple values of the dependent variable** for a ***unique value of an independent variable.***

3. What is correlation coefficient and list the properties of r

A C O R R E L AT I O N C O E F F I C I E N T FOR QUANTITATIVE DATA: r

- **Correlation Co-efficient** tells about the ***strength and direction of a relationship*** between **variables.**
- **Correlation coefficient (r), that describes the linear relationship *between pairs of variables for quantitative data.***

Properties of r

- 1. The **sign of r indicates** the type of ***linear relationship, whether positive or negative.***
- 2. The **numerical value of r,** without **regard to sign**, indicates ***the strength of the linear relationship.***

4. Define Regression and list the uses of regression

Definition

Regression is **a method to determine the statistical relationship between a *dependent variable and one or more independent variables (explanatory)*** y=f(x).y is dependent variable x is independent variable

Uses of Regression

- It indicates ***the significant relationships between dependent variable and independent variable.***
- It indicates **the strength of impact of multiple independent** variables on ***a dependent variable.***
- Regression helps **economists and financial analysts** in things **ranging from asset valuation to make predictions.**

5. Difference between Correlation and Regression

Correlation	Regression
In correlation analysis the **degree and direction** of relationship between the variables are studied.	In regression analysis, the **nature** of relationship is studied.
If **value of one variable** is known, the value of other variable **cannot** be estimated.	If **value of variable** is known, the value of other variable **can** be estimated using the functional relationships.
Correlation coefficient lies **between -1 and 1**.	Only one regression coefficient can be **greater than 1**.
Correlation coefficient is independent of change of origin and **scale**.	Regression coefficient is independent of change of origin **but not of scale**.
With the help of correlation coefficient and standard deviations of **two random variable(X,Y)** regression coefficient can be obtained.	With the help of regression coefficient, **correlation coefficient** can be obtained.
In Correlation analysis independent and dependent variables have **no practical significance.**	In regression analysis independent and dependent variable **have to be identified and practical significance.**

6. What is a regression line and write the equation for Y on X

Definition

❖A regression line **indicates a linear relationship** between ***the dependent variables on the y-axis(output) and the independent variables(Input)*** on the **x-axis.**

❖The correlation is established ***by analyzing the data pattern*** formed by the variables.

❖The regression line is ***plotted closest to the data points*** in a ***regression graph.***

The formula of the **regression line for Y on X**

$Y = a + bX + \varepsilon$

Here Y is the dependent variable, a is the Y-intercept, b is the slope of the regression line, X is the independent variable, and ε is the residual (error).

7. **State the multiple regression equation**.

Definition

Multiple regression is **a statistical technique that can be used to analyze the relationship between a single dependent variable and several independent variables**.

Example

$$Y' = .410(X_1) + .005(X_2) + .001(X_3) + 1.03$$

8. List the properties of correlation

Properties of Correlation

1. The **sign of r indicates** the type of ***linear relationship, whether positive or negative.***
2. The **numerical value of r,** without **regard to sign**, indicates ***the strength of the linear relationship.***

9. What is Curvilinear Relationship

A relationship that can be ***described best with a curved line***

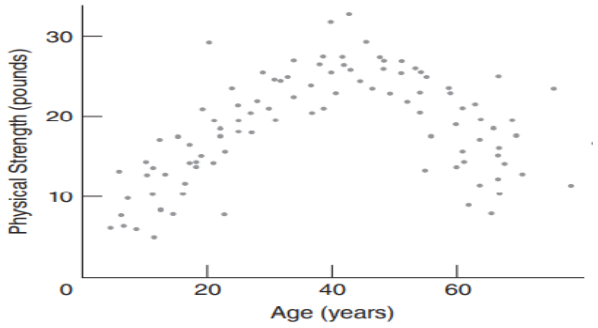


FIGURE 6.4
Curvilinear relationship.

:

10. What is dependent variable and Independent variable?

Independent Variable

A variable is said to be **independent**, which is stand alone variable whose **change influence another variable**,

Dependent Variable

If the ***variable is dependent, it will change in response*** to ***the change in some other variable.***

11. What is predicition error

Prediction Error

The **difference between the predicted values** made by **some model and the actual values.**

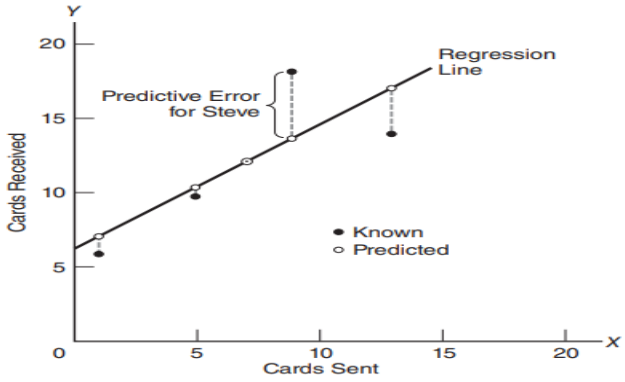


FIGURE
Predictive errors.

12.What is least square regression line

Regression Line has the **equation ŷ= a + b x**.

Definition

The Least Squares Regression Line is ***the line that makes the vertical distance*** from the ***data points to the regression line*** as small as possible.
It's called a **"least squares"** because the best line of fit is one that minimizes the variance (the sum of squares of the errors).
This can be a **bit hard to visualize** the **main point** is you are aiming to find the equation that fits the points as closely as possible.

13.What is correlation analysis

Correlation Analysis

Degree of Association is measured by ***Correlation Coefficient(r)(It is called as Pearsons correlation coefficient)***
It is a measure of linear association.

14.What is SEE and uses of standard error

STANDARD ERROR OF ESTIMATE, $S_{y|x}$

Definition

The **standard error of the estimate(SEE)** is **the estimation of the accuracy** of any predictions made by a ***regression model***.

- It is designed **to minimize predictive error**, the ***least squares equation*** does not eliminate it.Therefore, **our next task is to estimate** the ***amount of error associated*** with our predictions.

Uses of Standard Error

The standard error of the estimate gives us an idea of **how well a regression model fits a dataset.** In particular:

- The **smaller the value**, the **better the fit**.
- The **larger the value, the worse the fit**

15.List the regression assumptions

Regression assumptions

Linear regression makes several assumptions about the data, such as :
Linearity of the data. The relationship between the predictor (x) and the outcome (y) is assumed to be linear.
Normality of residuals. The residual errors are assumed to be normally distributed.
Homogeneity of residuals variance. The residuals are assumed to have a constant variance (**homoscedasticity**)

Independence of residuals error terms.

16.What is interpretation of r^2 or coefficient of determination

I N T E R P R E TAT I O N O F r^2

- ***R-squared (r^2 or Coefficient of Determination)*** is a **statistical measure** that indicates the ***variation in a dependent variable*** due to ***an independent variable.***

R-squared indicates the **variation in data** explained by the relationship between an **independent variable** and **a dependent variable**.

- It acts as a helpful tool for ***technical analysis***.
- It assesses the **performance of a dependent variable** with respect to a given **independent variable.**

17.List the Properties (or) Characteristics of the r-square

Properties (or) Characteristics of the R-square

- The ***coefficient of determination is the square of the correlation(r),*** thus it ranges **from 0 to 1.**
- With linear regression, the **coefficient of determination is equal to the square** of the ***correlation between the x and y variables***.
- If **r^2 =0,** then **the dependent variable cannot be predicted** from the ***independent variable***.
- If **r^2 =** 1, then **the dependent variable** can be **predicted from the independent variable without any error.**
- If **r^2 is between 0 and 1**, then it indicates ***the extent that the dependent variable can be predictable.***

18.Define Prediction

Prediction

Prediction is about **fitting** a shape that gets as **close** to the data as possible.

- ***Regression Equation*** is not able to construct if the relationship between X and Y scores varies because of lack of information.
- It is ***used to generate*** a ***customized prediction, Y'***

19. Difference between simple and multiple regression

SI.no	Simple Regression	Multiple Regression
1.	One dependent variable Y predicted from one independent variable X.	One dependent variable Y predicted from set of independent variable (X1,X2,…..X_k)
2.	One regression coefficient.	One regression coefficient for each independent variable.
3.	**r** ^{2}:Proposition of variation in dependent variable Y predicitable from X.	**R** ^{2}:Proposition of variation in dependent variable Y predicitable by set of independent variable from (X's)

20.What is residual analysis

Residual Analysis:

Residual analysis is important to check whether the assumption of regression modles have been satisfied.It is performed to check the following:

- ❖ The residual are normally distributed.
- ❖ The variance of residual is constant.
- ❖ The functional form of regression is correctly specified.
- ❖ If there are any Outliers

21.What is regression fallacy

Regression fallacy:

Regression fallacy assumes that a situation has returned to normal due to corrective actions having been taken while the situation was abnormal.It does not take into consideration normal fluctuations.

22.How is error calculated in linear regression?

Linear regression most often uses mean-square error (MSE) to calculate the error of the model. MSE is calculated by:

1. measuring the distance of the observed y-values from the predicted y-values at each value of x;
2. squaring each of these distances;
3. calculating the mean of each of the squared distances.

Linear regression fits a line to the data by finding the regression coefficient that results in the smallest MSE.

Part B Questions

1.Elaborate in detail the significance of correlation and the various types of correlation

- ❖ Definition
- ❖ Example
- ❖ Positive Relationship and Negative Relationship
- ❖ Properties of Correlation

2.What are scatterplot?Elabortae on the various types with suitable examples

Define

Uses of scatterplot

Positive, Negative, or Little or No Relationship

Strong or Weak Relationship

Perfect Relationship

Linear Relationship

Curvilinear Relationship

3.Hight light the significance of the correlation coefficient r.Compare the various correlation coefficients.

- ❖ Definition
- ❖ Properties
- ❖ Uses of Correlation coefficient
- ❖ Sign of r
- ❖ Numerical value of r

4.What is the significance of r^2?.Give a detailed interpretation of r^2?

- ❖ Definition
- ❖ R-Squared Formula
- ❖ Example
- ❖ Properties of R2

5. Discuss the importance of regression.Elaborate on types of regression?

- ❖ Definition of Regression Line
- ❖ Least squares regression line
- ❖ Least squares regression equation
- ❖ Multiple regression equation
- ❖ Example
- ❖ Features of Multiple Regression Equation
- ❖ Steps to perform Multiple regression

6.Elaborate regression towards mean ?Explain regression fallacy and state how it can be achieved?

Definition

Regression Fallacy

Xerox

DMI COLLEGE OF ENGINEERING

DEPARTMENT OF INFORMATION TECHNOLOGY

UNIT IV PYTHON LIBRARIES FOR DATA WRANGLING

1.Define list. How to create a list?

List:

The standard mutable multielement container in Python is the list.

Creating a Python list:

The list can be created using **either the list constructor or using square brackets [].**

Syntax

```
listName = [value1, value2, value3, value4]
```

2. List the properties of a list.

The properties of a list:

- **Mutable:** The elements of the list can be modified. We can add or remove items to the list after it has been created.
- **Ordered:** The items in the lists are ordered. Each item has a unique index value. The new items will be added to the end of the list.
- **Heterogenous**: The list can contain different kinds of elements i.e; they can contain elements of string, integer, boolean, or any type.
- **Duplicates:** The list can contain duplicates i.e., lists can have two items with the same values.

3. Difference between Array and list

Array	**List**
A single pointer to one contiguous block of data	It contains **a pointer to a block of pointers**, each of which in turn points to a **full Python object** like the Python integer
It is not flexible	It is **flexible** because each element is **a full structure** containing **both data and type information.**
We can store only ***homogeneous data***	We can store ***heterogeneous data***

4. Difference between Range and ARange

Range	ARange
It is a built-in python class	It is a function that belongs to **third-party library numpy.**
Range generates only integer values that can be accessed as list elements.	ARange function generates values that are stored in numpy arrays.
It is very slow because it occupies more space	It is fast it occupies less space
In [17]: `list(range(10))` Out[17]: `[0, 1, 2, 3, 4, 5, 6, 7, 8, 9]`	In [18]: `import numpy as np` `np.arange(10)` Out[18]: `array([0, 1, 2, 3, 4, 5, 6, 7, 8, 9])`

5. List the standard Numpy Data types.

NumPy Standard Data Types

- NumPy arrays contain values of a single type, so it is important to ***have detailed knowledge of those types and their limitations***.
- Because NumPy is built in C, the types will be familiar to users of C, FORTRAN, and other related languages.

Table 2-1. Standard NumPy data types

Data type	Description
`bool_`	Boolean (True or False) stored as a byte
`int_`	Default integer type (same as C `long`; normally either `int64` or `int32`)
`intc`	Identical to C `int` (normally `int32` or `int64`)
`intp`	Integer used for indexing (same as C `ssize_t`; normally either `int32` or `int64`)
`int8`	Byte (–128 to 127)
`int16`	Integer (–32768 to 32767)
`int32`	Integer (–2147483648 to 2147483647)
`int64`	Integer (–9223372036854775808 to 9223372036854775807)
`uint8`	Unsigned integer (0 to 255)
`uint16`	Unsigned integer (0 to 65535)
`uint32`	Unsigned integer (0 to 4294967295)
`uint64`	Unsigned integer (0 to 18446744073709551615)
`float_`	Shorthand for `float64`
`float16`	Half-precision float: sign bit, 5 bits exponent, 10 bits mantissa
`float32`	Single-precision float: sign bit, 8 bits exponent, 23 bits mantissa
`float64`	Double-precision float: sign bit, 11 bits exponent, 52 bits mantissa
`complex_`	Shorthand for `complex128`
`complex64`	Complex number, represented by two 32-bit floats
`complex128`	Complex number, represented by two 64-bit floats

6. Why NumPy Arrays are faster than Lists

1.NumPy Array uses **fixed memory to store data and less memory than Python lists**.

2.Contiguous memory allocation in NumPy Arrays.

7. List the categories of basic array manipulations

Attributes of arrays
Determining the size, shape, memory consumption, and data types of arrays

Indexing of arrays
Getting and setting the value of individual array elements

Slicing of arrays
Getting and setting smaller subarrays within a larger array

Reshaping of arrays
Changing the shape of a given array

Joining and splitting of arrays
Combining multiple arrays into one, and splitting one array into many

8.What is the use of reshape method()

The reshape method will use a no-copy view of theinitial array, but with noncontiguous memory buffers this is not always the case.

Another common reshaping pattern is the conversion of a one-dimensional arrayinto a two-dimensional row or column matrix.

```
In[39]: x = np.array([1, 2, 3])

        # row vector via reshape
        x.reshape((1, 3))

Out[39]: array([[1, 2, 3]])

In[40]: # row vector via newaxis
        x[np.newaxis, :]

Out[40]: array([[1, 2, 3]])
```

9.How concatenation of arrays or joining of arrays takes place

The routines **np.concatenate, np.vstack, and np.hstack. np.concatenate**

takes a tuple or list of arrays as its first argument.

```
In[43]: x = np.array([1, 2, 3])
        y = np.array([3, 2, 1])
        np.concatenate([x, y])

Out[43]: array([1, 2, 3, 3, 2, 1])
```

You can also concatenate more than two arrays at once:

```
In[44]: z = [99, 99, 99]
        print(np.concatenate([x, y, z]))

[ 1  2  3  3  2  1 99 99 99]
```

For working with arrays of mixed dimensions, it can be clearer to use the `np.vstack` (vertical stack) and `np.hstack` (horizontal stack) functions:

```
In[48]: x = np.array([1, 2, 3])
        grid = np.array([[9, 8, 7],
                         [6, 5, 4]])

        # vertically stack the arrays
        np.vstack([x, grid])

Out[48]: array([[1, 2, 3],
                [9, 8, 7],
                [6, 5, 4]])

In[49]: # horizontally stack the arrays
        y = np.array([[99],
                      [99]])
        np.hstack([grid, y])

Out[49]: array([[ 9,  8,  7, 99],
                [ 6,  5,  4, 99]])
```

10. What is Numpy'sUfunction

NumPy's UFuncs
Ufuncs exist in two flavors: *unary ufuncs*, which operate on a single input, and *binaryufuncs*, which operate on two inputs. We'll see examples of both these types of functions here.

Array arithmetic:

NumPy's ufuncs feel very natural to use because they make use of Python's native arithmetic operators. The standard addition, subtraction, multiplication, and division can all be used

Absolute value:

Just as NumPy understands Python's built-in arithmetic operators, it also understandsPython's built-in absolute value function:

In[11]: x = np.array([-2, -1, 0, 1, 2])

abs(x)

Out[11]: array([2, 1, 0, 1, 2])

Trigonometric functions:
NumPy provides a large number of useful ufuncs, and some of the most useful for the
data scientist are the trigonometric functions. We'll start by defining an array of angles:

```
In[15]: theta = np.linspace(0, np.pi, 3)
```

11.What are the two flavors are Ufuncs

Ufuncs exist in two flavors: unary ufuncs, which operate on a single input, and binary ufuncs, which operate on two inputs. We'll see examples of both these types of func- tions here.

12. List the Arithmetic operators implemented in NumPy

Array arithmetic:
NumPy's ufuncs feel very natural to use because they make use of Python's native
arithmetic operators. The standard addition, subtraction, multiplication, and division
can all be used:

```
In[7]: x = np.arange(4)
print("x =", x)
print("x + 5 =", x + 5)
print("x - 5 =", x - 5)
print("x * 2 =", x * 2)
print("x / 2 =", x / 2)
print("x // 2 =", x // 2) # floor division
```

Output:

x = [0 1 2 3]

x + 5 = [5 6 7 8]

x - 5 = [-5 -4 -3 -2]

x * 2 = [0 2 4 6]

x / 2 = [0. 0.5 1. 1.5]

x // 2 = [0 0 1 1]

13. List the specialized ufunctions

Specialized ufuncs:

NumPy has many more ufuncs available, including hyperbolic trig functions, bitwise arithmetic, comparison operators, conversions from radians to degrees, rounding and remainders, and much more. A look through the NumPy documentation reveals a lot of interesting functionality.

Another excellent source for more specialized and obscure ufuncs is the sub module scipy.special.

If you want to compute some obscure mathematical function on your data, chances are it is implemented in scipy.special.

14. List the . Aggregation functions available in NumPy

Aggregations: Min, Max, and Everything in Between:

Often when you are faced with a large amount of data, a first step is to compute summary

statistics for the data in question. Perhaps the most common summary statistics

are the mean and standard deviation, which allow you to summarize the "typical" values

in a dataset, but other aggregates are useful as well (the sum, product, median,

minimum and maximum, quantiles, etc.).

Table 2-3. Aggregation functions available in NumPy

Function Name	NaN-safe Version	Description
`np.sum`	`np.nansum`	Compute sum of elements
`np.prod`	`np.nanprod`	Compute product of elements
`np.mean`	`np.nanmean`	Compute median of elements
`np.std`	`np.nanstd`	Compute standard deviation
`np.var`	`np.nanvar`	Compute variance
`np.min`	`np.nanmin`	Find minimum value
`np.max`	`np.nanmax`	Find maximum value
`np.argmin`	`np.nanargmin`	Find index of minimum value
`np.argmax`	`np.nanargmax`	Find index of maximum value
`np.median`	`np.nanmedian`	Compute median of elements
`np.percentile`	`np.nanpercentile`	Compute rank-based statistics of elements
`np.any`	N/A	Evaluate whether any elements are true
`np.all`	N/A	Evaluate whether all elements are true

15.What is Broad casting and list the rulesof Broadcasting

Broadcasting:
Broadcasting is simply a set of rules for applying binary ufuncs (addition, subtraction, multiplication, etc.) on arrays of different sizes.
Rules of Broadcasting:
Broadcasting in NumPy follows a strict set of rules to determine the interaction between the two arrays:
• Rule 1: If the two arrays differ in their number of dimensions, the shape of the one with fewer dimensions is *padded* with ones on its leading (left) side.
• Rule 2: If the shape of the two arrays does not match in any dimension, the array with shape equal to 1 in that dimension is stretched to match the other shape.
• Rule 3: If in any dimension the sizes disagree and neither is equal to 1, an error is raised.

16.List the bitwise Boolean operators and their equivalent ufuncs

Bitwise Boolean operators:

Python's bitwise logic operators, &, |, ^, and ~. Like with the standard arithmetic

For example, we can address this sort of compound question as follows:

In[23]: np.sum((inches > 0.5) & (inches < 1))

Out[23]: 29

Combining comparison operators and Boolean operators on arrays can lead to a wide
range of efficient logical operations.
The following table summarizes the bitwise Boolean operators and their equivalent
ufuncs:

Operator	Equivalent ufunc
& | ^ ~	np.bitwise_and np.bitwise_or np.bitwise_xor np.bitwise_not

17. What is Fancy Indexing in Numpy

Fancy Indexing:

Fancy indexing is like the simple indexing we've already seen, but we pass

arrays of indices in place of single scalars. This allows us to very quickly access and

modify complicated subsets of an array's values.

18. What is combined indexing?

Combined Indexing:

For even more powerful operations, fancy indexing can be combined with the other indexing schemes we've seen:

In[9]: print(X)

[[0 1 2 3]

[4 5 6 7]

[8 9 10 11]]

We can combine fancy and simple indices:

In[10]: X[2, [2, 0, 1]]

Out[10]: array([10, 8, 9])

19. Write python code to create a Series by using an Array

Creating Series using Array

```
import pandas as pd

import numpy as np

info = np.array(['P','a','n','d','a','s'])

a = pd.Series(info)

print(a)
```

Output

```
0 P
1 a
2 n
```

3 d
4 a
5 s
dtype: object

20. List the two ways to find the missing data in the Data Frame

Detection of Missing Data

Two schemes to indicate the presence of missing data in a table or DataFrame:

- **Masking Approach**: The **mask that can be a separate** Boolean array
- **Sentinel Approach:** The sentinel value could be some

21. How to drop the null values

DROPPING NULL VALUES:

We use dropna() method on Series or DataFrame, which removes NaN values.

```
In [17]: data_pd.dropna()

Out[17]: 0    1.0
         2    2.0
         3    3.0
         dtype: float64
```

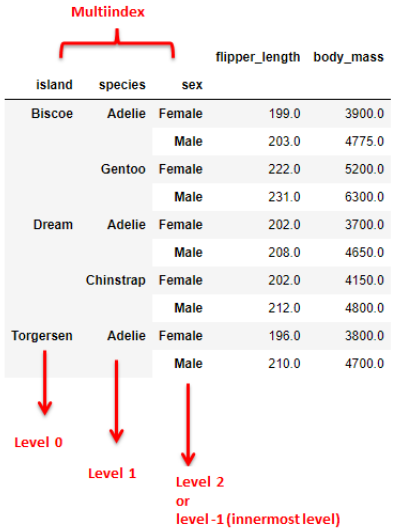
22. What is hierarchical Indexing?

Hierarchical Indexing or Multi-Indexing:

The index is like an address, that's how any data point across the data frame or series can be accessed.

Definition:

Hierarchical Indexes are also known as multi-indexing is setting more than one column name as the index.



23. What is stacking and unstacking in Indices

Stacking and unstacking indices:

The ***stack method*** turns ***column names into index values***, and the ***unstack method turns index values into column names***. You can see the data as a table ***with the unstack method.***

PART B QUESTIONS

1. Explain the Basic operations in array?

- **Definition**
- **Creating a Array**
- **NumPy Standard Data Types**
- **The Basics of NumPy Arrays**
  - **(i)** NumPy Array Attributes
  - ***(ii)*** Reshaping of Arrays
  - ***(iii)*** Array Indexing
  - ***(iv)*** Array Slicing
  - ***(v)*** Array Concatenation and Splitting
- **Computation on NumPy Arrays**
  - **(i)** Array arithmetic
  - **(ii)** Absolute value
  - **(iii)** Trigonometric functions
  - **(iv)** Exponents and logarithm

2. Explain the Aggregation operations of array?

- Definition
- Summing the Values in an Array
- Minimum and Maximum
- Multidimensional aggregates

3. How comparison and masking is done by array by using numpy?

- Definition
- Comparison Operators as ufuncs
- Working with Boolean Arrays
- Boolean operators
- Boolean Arrays as Masks

4. Explain Fancy Indexing in detail

- Definition
- Define RandomState
- Fancy Indexing –Multiple dimension
- Standard Indexing
- Combined Indexing

Fancy indexing can be combined with the other indexing schemes.

- (i)Fancy Index with simple Index
- (ii) Fancy indexing with slicing
- (iii) fancy indexing with masking
- Modifying Values with Fancy Indexing
- Arithmetic Operation

5. Explain Data Indexing and selection in Pandas

- Definition
- Data Selection in Series
- (i)Series as dictionary(Example)
- (ii)Series as one-dimensional array(Example)
- Data Selection in DataFrame
- (i)DataFrame as a dictionary(Example)
- (ii)DataFrame as a two-dimensional array(Example)

6. Explain Missing data in detail

- Definition
- Detection of Missing Data
- Handling Missing Data in Pyton(None,NaN)
- Operating on Null values
- (i)Detecting Null values(isnull,notnull)
- Example
- (ii)Dropping Null values
- Example
- (iii)Filling the Null values(forward and backward fill)
- Example

7. What is Hierarchical Indexing?

- Definition
- A Multiply Indexed Series(Example)
- Methods of MultiIndex Creation(Example)
- Creation of multi-index from arrays(Example)
- Creation of multi-index from DataFrame using pandas(Example)
- Indexing and Slicing a MultiIndex(Example)
- Rearranging Multi-Indices(Example)
- Index setting and resetting(Example)
- Data Aggrigations on Multi-Indices(Example)

8. What is grouping?.Explain the various operations in grouping?

- Definition
- Spilt
- Apply
- Combine
- A Visual representation of a groupby opration(diagram)
- Applying functions to a group
- Aggregation(Example)
- Filtering
- Transformation(Example)
- Method 1 (built-in function)
- Method 2(custom function)
- The apply()method(Example)

9 .How data sets are conbined?Explain concat,append ,join and mergein detail?

- Definition
- Combining Datasets
- Concat
- (i)Simple Concatention with pd.concat
- (ii)Duplicate indices
- (iii)Concatention with joins
- Merge(Example)
- Join
- Types of Join in Pandas
- (i)one_to_one
- (ii)Many to One
- (iii)Many-to-many joins
- Specification of the Merge key
- (i)On Keyword
- (ii)The left_on and right_on keywords
- Specifying Set Arithmetic for joins
- Outer(Example)
- Inner(Example)

10. What is pivot table? Explain in detail?

- Define Pivot Table
- Syntax of the Pivot Table creation
- Difference between groupby and pivot table
- Uses of Pivot table
- Example with pandas code for Pivot table creation
- Pivot table with single and multiple aggregation function
- Multiple features in pivot table
- Multi-level Index in Pivot Table

Unit – V
Two Marks

1. What is data visualization?

Data visualization is the graphical representation of information and data.

2. Which concept is used in data visualization?

   Data visualization based on two concepts:
   - Each attribute of training data is visualized in a separate part of screen.
   - Different class labels of training objects are represented by different colors.

3. List the benefits of data visualization.

   Constructing ways in absorbing information. Data visualization enables users to receive vast amounts of information regarding operational and business conditions.
   - Visualize relationships and patterns in businesses.
   - More collaboration and sharing of information.
   - More self-service functions for the end users.

4. Why big data visualization is important?

   Reasons:
   It provides clear knowledge about patterns of data.
   Detects hidden structures in data.
   Identify areas that need to be improved.
   Help us to understand which products to place where.
   Clarify factors which influence human behaviour.

5. Explain Matplotlib.

   Mathplotlib is a cross-platform, data visualization and graphical plotting library for Python and its numerical extension NumPy. Matplotlib is a comprehensive library for creating static, animated and interactive visualizations in Python. Matplotlib is a plotting library for the Python programming language. It allows to make quality charts in few lines of code. Most of the other python plotting library are build on top of Matplotlib.

6. What is contour plot?

   A contour line or isoline of a function of two variables is a curve along which the function has a constant values. It is a cross-section of the three-dimensional graph of the function f(x,y) parallel to the x,y plane. Contour lines are used e.g. in geography and meterology. In cartography, a contour

line joins points of equal height above a given level, such as mean sea level.

7. Explain legends.

   Plot legends give meaning to a visualization, assigning labels to the various plot elements. Legends are found in maps – describe the pictorial language or symbology of the map. Legends are used in line graphs to explain the function or the values underlying the different lines of the graph.

8. What is subplots?

   Subplots mean groups of axes that can exist in a single matplotlib figure. Subplots() function in the matplotlib library, helps in creating multiple layouts of subplots. It provides control over all the individual plots that are created.

9. What is use of tick?

   - A tick is a short line on an axis. For category axes, ticks separate each category. For value axes, ticks mark the major divisions and show the exact point on an axis that the axis label defines. Ticks are always the same color and line style as the axis.
   - Ticks are the markers denoting data points on axes. Matplotlib's default tick locators and formatters are designed to be generally sufficient in many common situations. Position and labels of ticks can be explicitly mentioned to suit specific requirements.

10. Describe in short Basemap.

    - Basemap is a toolkit under the Python visualization library Matplotlib. Its main function is to draw 2D maps, which are important for visualizing spatial data. Basemap itself does not do any plotting, but provides the ability to transform coordinates into one of 25 different map projections.
    - Matplotlib can also be used to plot contours, images, vectors, line or points in transormed coordinates. Basemap inclues the GSSH coastline dataset, as well as datasets from GMT for rivers, states and national boundaries.

11. What is Seaborn?

    - Seaborn is a Python data visualization library based on Matplotlob. It provides a high-level interface for drawing attractive and informative statistical graphics. Seaborn is an opensource Python library.

- Its dataset-oriented, declarative API. User should focus on what the different elements of your plots mean, rather than on the details of how to draw them.

12. What are the types of Ticks?

Ticks come in two types: Major and Minor

- Major ticks separate the axis into major units. On category axes, major ticks are the only ticks available. On values axes, one major tick appears for every major axis division.
- Minor ticks subdivide the major tick units. They can only appear on value axes. One minor tick appears for every minor axis division.

13. Difference between Matplotlib and Seaborn

Parameters	**Matplotlib**	**Seaborn**
Use cases	Matplotlib plots various graphs using Numpy and Pandas.	Seaborn is the extended version of Matplotlib which uses Matplotlib along with Numpy and Pandas for plotting
Syntax complity	It uses comparatively complex and lengthy syntax.	It uses comparatively simple syntax
Multiple figures	We can open multiple figures at a time.	Seaborn automates the creation of multiple figures which sometimes leads to out of memory issue.
Flexibility	It is highly customizable and powerful	Seaborn avoids ton of boilerplate by providing default themes which are commonly used.

14. What are the features of seaborn?

- Seaborn is a statistical plotting library
- It has beautiful default styles
- It also is designed to work very well with Pandas dataframe objects.

15. What are four important parameters that you must always use with annotate()?

text: This define the text label. Takes a string as a value.

xy: The place where you want your arrowhead to point to. In other words, the place you want to annotate. This is a tuple containing two values, x and y.

xytext: The coordinates for where you want to text to display.

arrowprops: A dictionary of key-value pairs which define various properties for the arrow, such as color, size and arrowhead type.

16. Define Histogram.

In a histogram, the data are grouped into ranges (e.g. 10-19, 20-29) and then plotted as connected bars. Each bar represents a range of data. The width of each bar is proportional to the width of each category, and the height is a proportional to the frequency or percentage of that category.

17. What are the different classes of color maps?

- Sequential
- Diverging
- Cyclic
- Quanlitative

18. What are the purpose for using Matplotlib functions?

- plt.contour for contour plots,
- plt.contourf for filled contour plots,
- plt.imshow for showing images.

19. How to add error bar in Matplotlib.

Adding the error bar in Matplotlib, Python. It's very simple, we just have to write the value of the error. We use the command:

Plt.errorbar(x, y, yerr=2, capsize =3)

20. What are the purpose of elements label, annotation & legend?

Label: Make it easy for the viewer to know the name or kind of data illustrated

Annonation: Help extend the viewer's knowledge of the data, rather than simply identify it.

Legend: Provides cues to make identification of the data group easier.

16 Marks

1. Explain Importing Matplotlib in details.

- Visualizing Information: Starting with Graph
- Line Plot
- Saving work to disk
- Setting the axis, ticks, grids
- Defining the line appearance and working with line style

Fig: Line style

- Adding markers
- Using Labels, Annotations and Legends
- Creating a legend

2. Explain in detail about Scatter Plots.
    - Comparing plt.scatter() amd plt.plot()
    - Creating Advanced Scatterplots
3. What are Density and Contour Plots?
    - Contour plot
    - Changing the colours and the line style
    - Different classes of color maps.
        - Sequential
        - Diverging
        - Cyclic
        - Quantitative
4. What is Visualization with Seaborn?
    - Key Features
    - Functionality that seaborn offers
    - Plot a Scatter Plot in Seaborn
    - Difference between Matplotlib and Seaborn
5. Explain about Histogram, Three Dimensional Plotting, Geographic Data with Basemap
    - Histogram
        - Fig. Histogram
        - Code for creating histogram with randomized data.
    - Three Dimensional Plotting
        - First import the library
        - Create figure and axes
    - Geographic Date with BaseMap
        - Examples objects in basemap
            a. contour()
            b. contour()
            c. imshow()
            d. pcolor()
            e. pcolormesh()
            f. plot()
            g. scatter()
            h. quiver()
            i. barbs()

j. drawgreatcircle()

- Basic benefit usage.