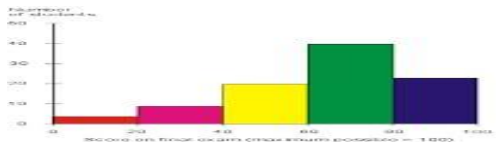
U23CBT41-Foundations of Data Science
QUESTION BANK

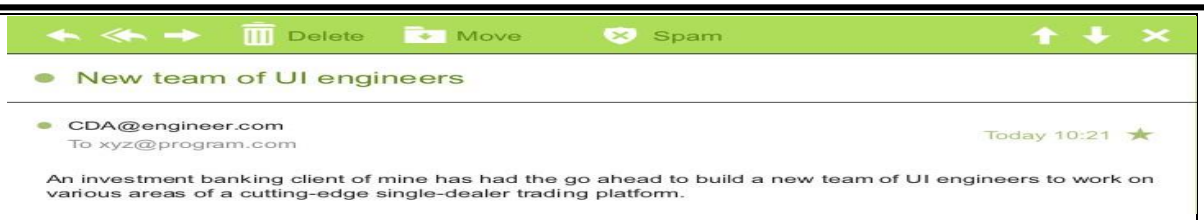
Sem/Year:V/III Year IT **Regulation:2023**

Staff Name:Mrs.K.VAIDEGI ,AP/IT Unit I-INTRODUCTION
PART A

1	**Define Data Science and Big data. [Nov/Dec 2022]** Data science is the study of working with a huge volume of data and enables data for prediction, prescriptive, and prescriptive analytical models. Big data is the study of collecting and analyzing a huge volume of data sets to find a hidden pattern that helps in stronger decision-making.
2	**List an overview of common errors in retrieving data and which cleansing solutions to be employed. [Nov/Dec 2022]** Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled.
3	**Outline the difference between structured data and unstructured data. [Apr/May 2023]** Structured data is standardized, clearly defined, and searchable data, while unstructured data is usually stored in its native format. Structured data is quantitative, while unstructured data is qualitative. Structured data is often stored in data warehouses, while unstructured data is stored in data lakes.
4	**Define Data mining. [Apr/May 2023]** Data mining refers to extracting or mining knowledge from large amounts of data. It is a process of discovering interesting patterns or Knowledge from a large amount of data stored either in databases, data warehouses, or other information repositories.
5	**Define Outlier.** Outlier detection is the process of detecting and subsequently excluding outliers from a given set of data. The easiest way to find outliers is to use a plot or a table with the minimum and maximum values.
6	**What is the use of Histogram?**

	The histogram is a popular graphing tool. It is used to summarize discrete or continuous data that are  measured on an interval scale. It is often used to illustrate the major features of the distribution of the data in a convenient form
7	**Define Project Charter** • A clear research goal • The project mission and context • How you're going to perform your analysis • What resources you expect to use
8	**How do measuring central Tendency?** The mode is the most frequent value. The median is the middle number in an ordered data set. The mean is the sum of all values divided by the total number of values.
9	**Write Steps for IQR with Example.** Order the data from least to greatest. Find the median. Calculate the median of both the lower and upper half of the data. The IQR is the difference between the upper and lower medians.
10	**Short notes on Streaming Data.** Streaming data is data that is generated continuously by thousands of data sources, which typically send in the data records simultaneously, and in small sizes (order of Kilobytes).

1	A	**Examine the different facets of data with the challenges in their Processing. [Nov/Dec 2022]** **Facets of Data** It is used to represent the various forms in which the data could be represented inside Big Data. The following are the various forms in which the data could be represented. **1.Structured**(Structured data is data that depends on a data model and resides in a fixed field within a record.) Example:Excel files. SQL , or Structured Query Language **2.Unstructured**(Unstructured data is data that isn't easy to fit into a data model because the content is context-specific or varying.) Example: Email



3.Natural Language(Natural language is a special type of unstructured data; it' challenging to process because it requires knowledge of specific data

PART B

science techniques and linguistics.)

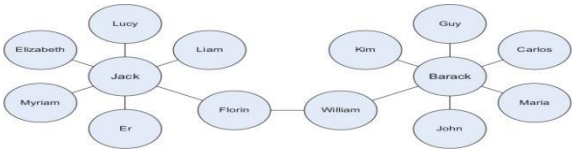
Example: Emails, mails, comprehensions, essays, articles etc..

4.Machine Generated(Machine-generated data is information that's automatically created by a computer, process, application, or other machine without human intervention.) Example:

```
CSIPERF:TXCOMMIT;313236
2014-11-28 11:36:13, Info                  CSI    00000153 Creating NT transaction (seq
69), objectname [6]"(null)"
2014-11-28 11:36:13, Info                  CSI    00000154 Created NT transaction (seq 6
result 0x00000000, handle @0x4e54
2014-11-28 11:36:13, Info                  CSI    00000155@2014/11/28:10:36:13.471
Beginning NT transaction commit...
```

5.Graph Based(In graph theory, a graph is a mathematical structure to model pair-wise relationships between objects.)

Example:Graph-based data is a natural way to represent social networks, and its structure allows you to calculate specific metrics



6.Audio, Video & Image Audio, image, and video are data types that pose specific challenges to a data scientist. Tasks that are trivial for humans, such as recognizing objects in pictures, turn out to be challenging for computers.

Examples: Youtube videos, podcast, music and lots more to add up to.

7.Streaming Data

The data flows into the system when an event happens instead of being loaded into a data store in a batch.

Examples: Video conferences and live telecasts all work on this basics.

2	R	**Elaborate about the steps in data process with a diagram or any 3 steps of it with suitable diagram and example . [Nov/Dec 2022] [Apr/May 2023]** **Data Science Process** Data science is an interdisciplinary field which is focused on extracting knowledge from Big Data, which are typically large, and applying the knowledge and actionable insights from data to solve problems in a wide range of application domains. **Characteristics of Big data Volume** - How much data is there? **Variety** - How diverse are different types of data? **Velocity** - At what speed is new data generated? **Veracity** - How accurate is the data? **Need for Data Science** • Big data is a huge collection of data with wide variety of different data set and in different formats. Data science involves using methods to analyse massive amounts of data and extract the knowledge it contains. **Benefits & uses of Data Science & Big Data** Data science and big data are used almost everywhere in both commercial and non-commercial settings. • Google AdSense, which collects data from internet users so relevant

commercial messages can be matched to the person
internet.

- Human resource professionals use people analytics
screen candidates, monitor the mood of employees, a
networks among coworkers.
- Financial institutions use data science to predict stock
the risk of lending money, and earn how to attract new clie

Application:

Gaming. ...
Image Recognition. ...
Recommendation Systems. ...
Fraud Detection. ...
Internet Search. ...
Speech recognition.

Process

Data science process consists of six stages :

1. Discovery or Setting the research goal
2. Retrieving data
3. Data preparation
4. Data exploration
5. Data modeling
6. Presentation and automation

- Fig. 1.3.1 showsdata science

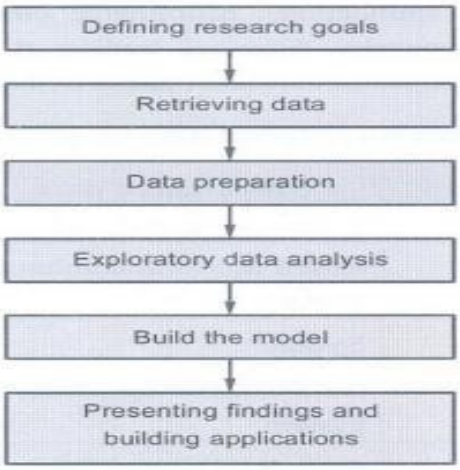


Fig. 1.3.1 : Data science design process

- **Step 1: Discovery or Defining research goal**

This step involves acquiring data from all the identified i
sources, which helps to answer the business question.

- **Step 2: Retrieving data**

It collection of data which required for project. This is the
a business understanding of the data user have and decipher
of data means. This could entail determining exactly what
the best methods for obtaining it. This also entails determ
the data points means in terms of the company. If we ha
from a client, for example, we shall need to know what eadesign process.

column and row represents.

• Step 3: Data preparation

Data can have many inconsistencies like missing values, blank columns, an incorrect data format, which needs to be cleaned. We need to process, explore and condition data before modeling. The cleandata, gives the better predictions.

- **Step 4: Data exploration**

Data exploration is related to deeper understanding of data. Try to understand how variables interact with each other, the distribution of the data and whether there are outliers. To achieve this use descriptive statistics, visual techniques and simple modeling. This steps is also called as Exploratory Data Analysis.

- **Step 5: Data modeling**

In this step, the actual model building process starts. Here, Data scientist distributes datasets for training and testing. Techniques like association, classification and clustering are applied to the training data set. The model, once prepared, is tested against the "testing" dataset.

- **Step 6: Presentation and automation**

Deliver the final baselined model with reports, code and technical documents in this stage. Model is deployed into a real-time production environment after thorough testing. In this stage, the key findings are communicated to all stakeholders. This helps to decide if the project results are a success or a failure based on the inputs from the model.

1.Data Preparation

- Data preparation means data cleansing, Integrating and transforming data.

Data Cleaning

- Data is cleansed through processes such as filling in missing values, smoothing the noisy data or resolving the inconsistencies in the data.
- Data cleaning tasks are as follows:

1. Data acquisition and metadata
2. Fill in missing values
3. Unified date format
4. Converting nominal to numeric
5. Identify outliers and smooth out noisy data
6. Correct inconsistent data

- Data cleaning is a first step in data pre-processing techniques which is used to find the missing value, smooth noise data, recognize outliers and correct inconsistent.
- **Missing value**: These dirty data will affects on miming procedure and led to unreliable and poor output. Therefore it is important for some data cleaning routines. For example, suppose that the average salary of staff is Rs. 65000/-. Use this value to replace the missing value for salary.
- **Data entry errors:** Data collection and data entry are error-prone processes. They often require human intervention and because humans are only human, they make typos or lose their concentration for a second and

		introduce an error into the chain. But data collected by machines or computers isn't free from errors either. Errors can arise from human

sloppiness, whereas others are due to machine or hardware failure. Examples of errors originating from machines are transmission errors or bugs in the extract, transform and load phase (ETL).

- **Whitespace error:** Whitespaces tend to be hard to detect but cause errors like other redundant characters would. To remove the spaces present at start and end of the string, we can use strip() function on the string in Python.
- **Fixing capital letter mismatches:** Capital letter mismatches are common problem. Most programming languages make a distinction between "Chennai" and "chennai".
- Python provides string conversion like to convert a string to lowercase, uppercase using lower(), upper().
- The lower() Function in python converts the input string to lowercase. The upper() Function in python converts the input string to uppercase.

Outlier

- Outlier detection is the process of detecting and subsequently excluding outliers from a given set of data. The easiest way to find outliers is to use a plot or a table with the minimum and maximum values.

2.Exploratory Data Analysis

- Exploratory Data Analysis (EDA) is a general approach to exploring datasets by means of simple summary statistics and graphic visualizations in order to gain a deeper understanding of data.
- EDA is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers user need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis or check assumptions.

- Box plots are an excellent tool for conveying location and variation information in data sets, particularly for detecting and illustrating location and variation changes between different groups of data.
- Exploratory data analysis is majorly performed using the following methods:

1. Univariate analysis: Provides summary statistics for each field in the raw data set (or) summary only on one variable. Ex : CDF,PDF,Box plot
2. Bivariate analysis is performed to find the relationship between each variable in the dataset and the target variable of interest (or) using two variables and finding relationship between them. Ex: Boxplot, Violin plot.
3. Multivariate analysis is performed to understand interactions between different fields in the dataset (or) finding interactions between variables more than 2.

- **A box plot is** a type of chart often used in explanatory data analysis to visually show the distribution of numerical data and skewness through displaying the data quartiles or percentile and averages.

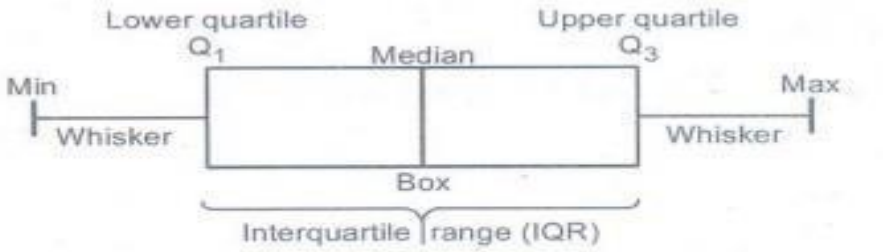


Fig. 1.7.1

3. **Build the Models**

• To build the model, data should be clean and understand the content properly. The components of model building are as follows:

a) Selection of model and variable

b) Execution of model

c) Model diagnostic and model comparison

• Building a model is an iterative process. Most models consist of the following main steps:

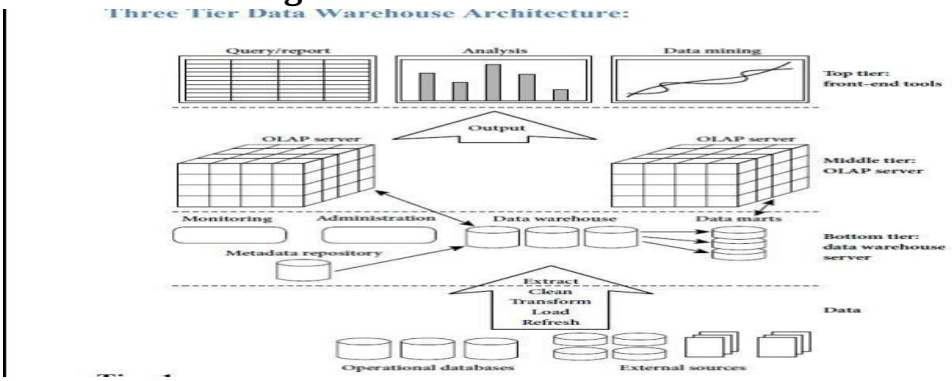
1. Selection of a modeling technique and variables to enter in the model
2. Execution of the model
3. Diagnosis and model comparison

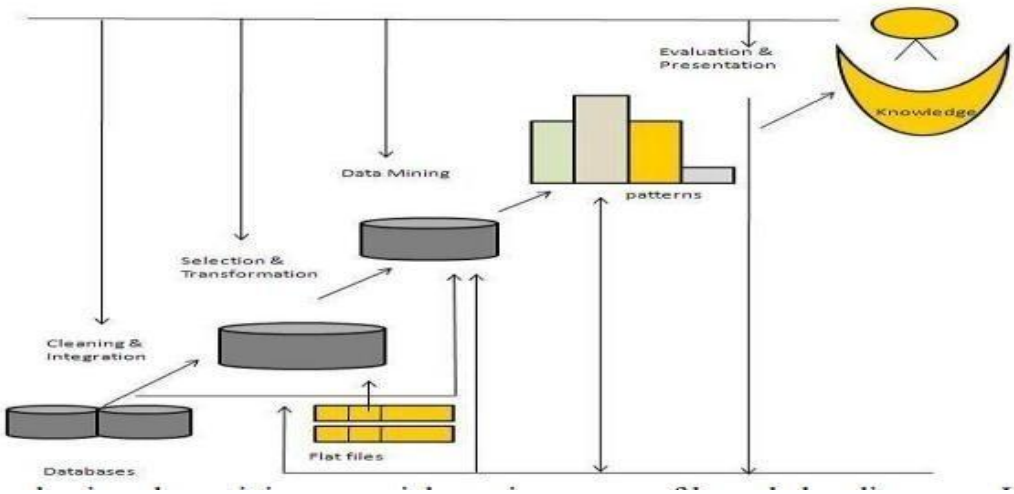
Model and Variable Selection

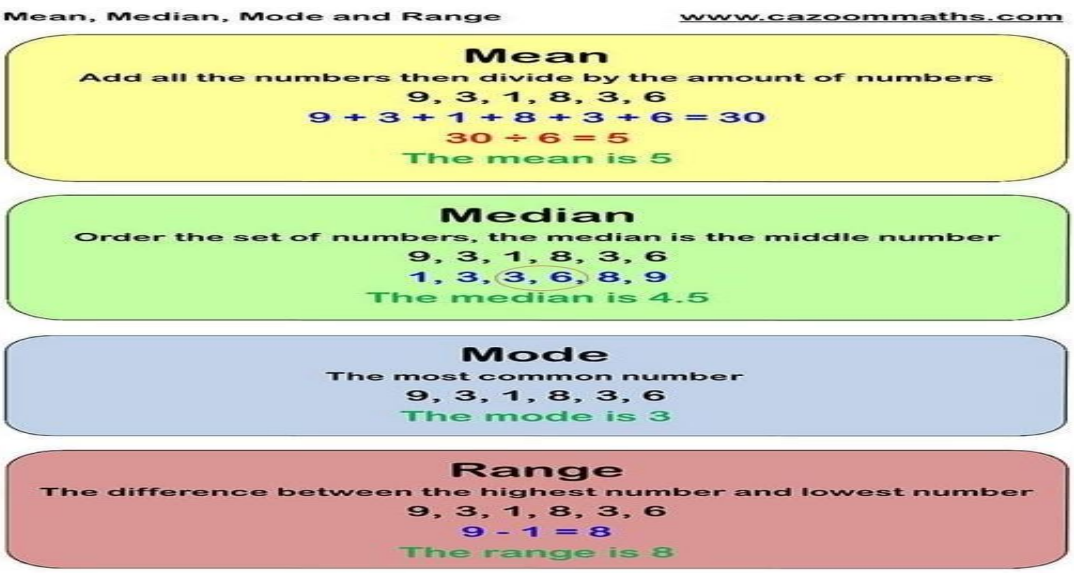
• For this phase, consider model performance and whether project meets all the requirements to use model, as well as other factors:

1. Must the model be moved to a production environment and, if so, would it be easy to implement?
2. How difficult is the maintenance on the model: how long will it remain relevantif left untouched?
3. Does the model need to be easy to explain? Model Execution

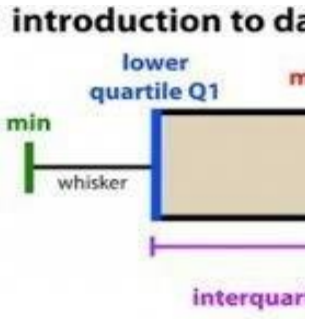
• Various programming language is used for implementing the model. For model execution, Python provides libraries like StatsModels or Scikit-learn. These packages use several of the most popular techniques.

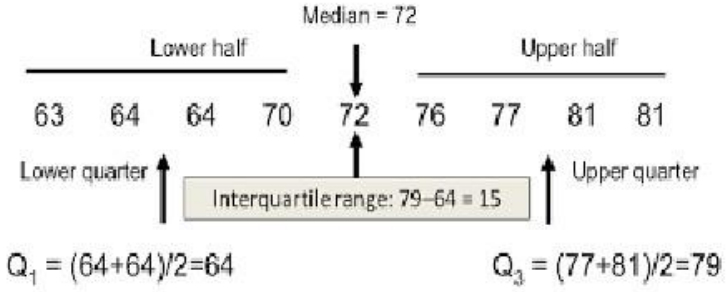
3	R	**What is Data Warehousing? Outline the architecture of Data Warehousing with neat diagram[Apr/May 2023]** **Data Warehousing**  **Tier-1:** The bottom tier is a warehouse database server that is almost always a relationaldatabase system. Back-end tools and utilities are used to feed data

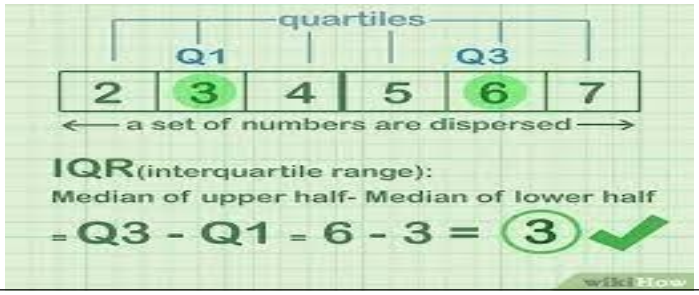
		into the bottomtier from operational databases or other external sources (such as customer profileinformationprovided by external consultants). **Tier-2:** The middle tier is an OLAP server that is typically implemented using either a relational OLAP (ROLAP) model or a multidimensional OLAP. OLAP model is an extended relational DBMS thatmaps operations on multidimensionaldata to standard relational operations. A multidimensional OLAP (MOLAP) model, that is, a specialpurpose 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)
4	R	**What is Data mining? Outline the architecture of Data Mining with neat diagram DataMining** Data mining refers to extracting or mining knowledge from large amounts of data. • **Data Cleaning** - In this step the noise and inconsistent data is removed. • **Data Integration** - In this step multiple data sources are combined. • **Data Selection** - In this step relevant to the analysis task are retrieved from the database. • **Data Transformation** - In this step data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations. • **Data Mining** - In this step intelligent methods are applied in order to extract datapatterns. • **Pattern Evaluation** - In this step, data patterns are evaluated. • **Knowledge Presentation** - In this step,knowledge is represented.The 
5	C	**Construct Statistical Description of data.**



IQR(Inter Quartile Range)







UNIT II-DESCRIBING DATA

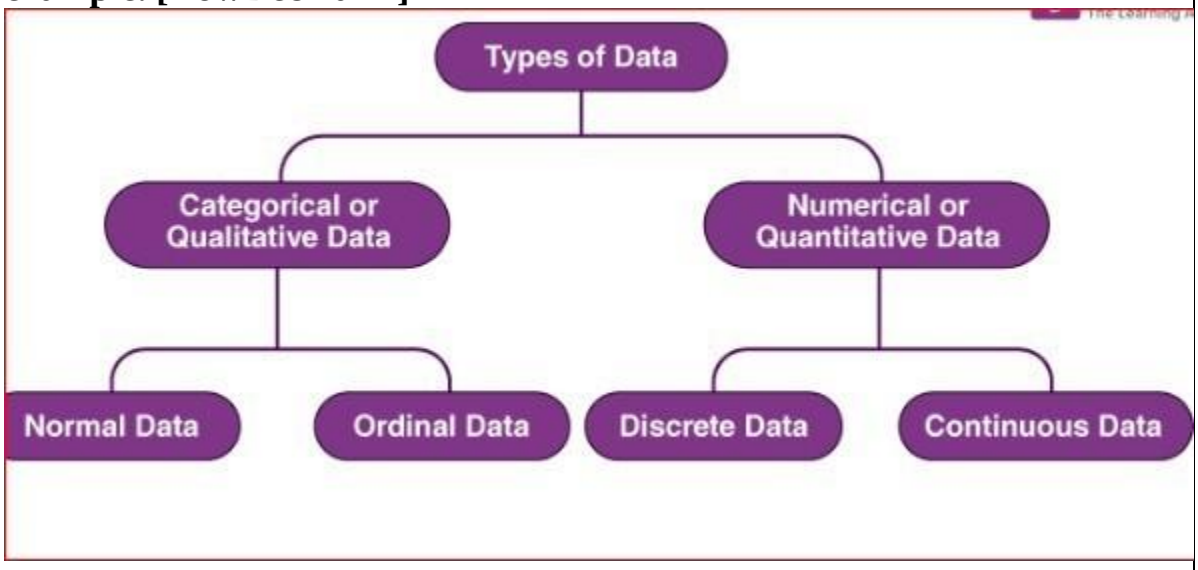
PART A

1	U	**Differentiate/compare Quantitative and Qualitative Data with example. [Apr/May 2023]** Quantitative data refers to any information that can be quantified, counted or measured, and given a numerical value. Qualitative data is descriptive in nature, expressed in terms of language rather than numerical values. Quantitative research is based on numeric data

2	R	**Define Ranked and Nominal Values.** Ordinal data is qualitative data that is categorized in a specific ranked order or hierarchy. Nominal data is qualitative data that is categorized based only on descriptive characteristics. This kind of data has no ranked order or hierarchy.
3	U	**Compare or Differentiate Continuous and Discrete variables with an example [Nov/Dec 2022] [Apr/May 2023]** Discrete and continuous variables are two types of quantitative variables: Discrete variables represent counts (e.g. the number of objects in a collection). Continuous variables represent measurable amounts (e.g. water volume or weight).
4	U	**Differentiate Grouped and Ungrouped data.** Ungrouped data is not classified or organized into different classes, whereas grouped data is organized into a number of classes. Ungrouped data is presented in the form of lists, whereas, frequency tables are used to express, grouped data.
5	A	**How to calculate Relative Frequency, Cumulative Frequency and percentile Rank with an example. [Nov/Dec 2022]** The cumulative relative frequencies are the cumulative frequencies divided by n. For example, the cumulative relative frequency on row [2] is the cumulative frequency 6 divided by n=15 to give 6/15=3/5=0.6.
6		**Classify the below list of data into their types.a)ethnic group b)age c)family size d)academic major e)IQ score f)networth g)gender h)Temperature. [Nov/Dec 2022]**
7	An	**Compute mean Mode and median for following 55,60,60,63,63,63,63,65,65.**
8	C	**Construct Histogram and Frequency Polygon for following Example.** VIEWING TIME — f 35–above — 2 30–34 — 5 25–30 — 29 20–22 — 60 15–19 — 60 10–14 — 34 5–9 — 31 0–4 — 29 Total — 250
9	R	**Define Misleading Graph.**
		A misleading graph, also known as a **distorted graph**, is a graph that misrepresents data, constituting a misuse of statistics and with the result that an incorrect conclusion may be derived from it.
10	An	**Calculate Stem and Leaf for following data 12,22,52,46,14,13,26,41,30,120,112,101,105**

Part B

1	R	**Differentiate Type of data and variable used in data analysis with an example. [Nov/Dec 2022]**



Qualitative or Categorical Data

Qualitative data, also known as the categorical data, describes the data that fits into the categories. Qualitative data are not numerical. The categorical information involves categorical variables that describe the features such as a person's gender, home town etc. Categorical measures are defined in terms of natural language specifications, but not in terms of numbers.

Sometimes categorical data can hold numerical values (quantitative value), but those values do not have a mathematical sense. Examples of the categorical data are birthdate, favourite sport, school postcode. Here, the birthdate and school postcode hold the quantitative value, but it does not give numerical meaning.

Nominal Data

Nominal data is one of the types of qualitative information which helps to label the variables without providing the numerical value. Nominal data is also called the nominal scale. It cannot be ordered and measured. But sometimes, the data can be qualitative and quantitative. Examples of nominal data are letters, symbols, words, gender etc.

The nominal data are examined using the grouping method. In this method, the data are grouped into categories, and then the frequency or the percentage of the data can be calculated. These data are visually represented using the pie charts.

Ordinal Data

Ordinal data/variable is a type of data that follows a natural order. The

significant feature of the nominal data is that the difference between the data values is not determined. This variable is mostly found in surveys, finance, economics, questionnaires, and so on.

The ordinal data is commonly represented using a bar chart. These data are investigated and interpreted through many visualisation tools. The information may be expressed using tables in which each row in the table shows the distinct category. Quantitative or Numerical Data

Quantitative data is also known as numerical data which represents the numerical value (i.e., how much, how often, how many). Numerical data gives information about the quantities of a specific thing. Some examples of numerical data are height, length, size, weight, and so on. The quantitative data can be classified into two different types based on the data sets. The two different classifications of numerical data are discrete data and continuous data.

Discrete Data

Discrete data can take only discrete values. Discrete information contains only a finite number of possible values. Those values cannot be subdivided meaningfully. Here, things can be counted in whole numbers.

Example: Number of students in the class

Continuous Data

Continuous data is data that can be calculated. It has an infinite number of probable values that can be selected within a given specific range.

Example: Temperature range

2 A & C

a.The number of friends by Face book users is summarized in the following frequency distribution. [Nov/Dec 2022]

Data	f
400-above	2
350-399	5
300-349	12
250-299	17
200-249	23
150-199	49
100-149	27
50-99	29
0-49	36
Total	**200**

i. What is the shape of this distribution?

ii. Find the relative Frequency and Cumulative Frequency. iii. Find the approximate percentile rank of interval 300-349 iv. Convert to a histogram

v. Why would it not be possible to convert to a stem and leaf display. b.What is relative frequency distribution ? the GRE scores for a group of graduate school applicants are distributed as follows. [Apr/May 2023]

			GRE Score	**Frequency**
			475-499	2
			500-524	4
			525-549	13
			550-574	27
			575-599	30
			600-624	42
			625-649	34
			650-774	30
			675-699	14
			700-724	3
			725-749	1
			Total	**200**
		convert a frequency distrib frequency distribution and presented in above table ght of the decimal point. istribution. Do not round numbers to two digits to		
3	C	What is a frequency distribution? Customers who have purchased a particular product rated the usability of the product on a 10 point scale, ranging from 1 (poor) to 10 (excellent) as follows. [Apr/May 2023] 3 7 2 7 8 3 1 4 10 3 2 5 3 5 8 9 7 6 3 7 8 9 7 3 6 Construct Frequency Distribution of each data.		
4	E	i)What is Median? Outline the steps to find the median and find the median for the following scores: first, set of five scores 2,8,2,7,6 and second, set of six scores 3,8,9,3,1,8 with steps. [Apr/May 2023] ii)What is mode? Can there be distribution with bo mode or more than one mode? The owner of new car conducts six gas mileage tests and obtain the following results, expressed in miles per gallon: 26.3, 28.7,27.4,26.6,27.4,26.9. Find the mode for these data. [Apr/May 2023] iii)Determine the values of the range and IQR for the following set of data. a)Retirement ages:60,63,45,63,65,70,55,63,60,65,63 b)Residence changes : 1,3,4,1,0,2,5,8,0,2,3,4,7,11,0,2,3,4 iv)Using computation formula for the sum of squares calculate the population standard deviation for the scores in (a) and sample standard deviation for the scores in (b) (a) 1,3,7,2,0,4,7,3 (b)10,8,5,0,1,1,7,9,2		

5	R & A	i) What is Z Score? Outline the steps to obtain a Z score. [Apr/May 2023] ii) Express each of the following scores as a Z Score: First Mary's intelligence quotient is 135,given a mean of 100 and standard deviation 15.Second, Mary obtained a score of 470 in the Competitive Examination conducted in April 2022, given a mean of 500 and a standard deviation 0f

UNIT III-DESCRIBING RELATIONSHIPS PART A

1	**What do You mean by least square method?** The least square method is the process of finding the best-fitting curve or line of best fit for a set of data points by reducing the sum of the squares of the offsets (residual part) of the points from the curve.
2	**Compare Correlation and Regression.** Correlation is a statistical measure that determines the association or corelationship between two variables. Regression describes how to numerically relate an independent variable to the dependent variable. To represent a linear relationship between two variables.
3	**What is Correlation and define Correlation coefficient?** [Nov/Dec 2022] he correlation coefficient is a statistical measure of the strength of a linear relationship between two variables. Its values can range from -1 to 1. A correlation coefficient of -1 describes a perfect negative, or inverse, correlation, with values in one series rising as those in the other decline, and vice versa.
4	**Define Interpretation R^2 with an Example. [Nov/Dec 2022]** The value of R-Squared is always between 0 to 1 (0% to 100%). A high R-Squared value means that many data points are close to the linear regression function line. A low R-Squared value means that the linear regression function line does not fit the data well.
5	**What is Scatterplots and its types and usage? [Apr/May 2023]** scatter plot (aka scatter chart, scatter graph) uses dots to represent values for two different numeric variables. The position of each dot on the horizontal and vertical axis indicates values for an individual data point. Scatter plots are used to observe relationships between variables.
6	**Consider Helen sent 10 greeting card to her friends and she received back 8 cards, what is the kind of relationship it is? Brief on it. [Nov/Dec 2022]** Negative Relationship
7	**Differentiate simple Regression and Multiple Regression.** Simple linear regression has only one x and one y variable. Multiple linear regression has one y and two or more x variables. For instance, when we predict rent based on square feet alone that is simple linear regression.

8	**What is Regression towards the mean with an example.** Regression toward the mean simply says that, following an extreme random event, the next random event is likely to be less extreme.

9		In studies dating back over 100 years, it's well established that regression toward the mean occurs between the heights of fathers and the heights of their adult sons. Indicate whether the following statements are true or false. (a) Sons of tall fathers will tend to be shorter than their fathers. (b) Sons of short fathers will tend to be taller than the mean for all sons. (c) Every son of a tall father will be shorter than his father. (d) Taken as a group, adult sons are shorter than their fathers. (e) Fathers of tall sons will tend to be taller than their sons. (f) Fathers of short sons will tend to be taller than their sons but shorter than the mean for all fathers. **Answers** (a) True (b) False. Sons of short fathers will tend to be taller than their fathers but still shorter than the mean for all sons. (c) False. Regression toward the mean is only a tendency, so there will be exceptions. (d) False. Taken as an entire group, adult sons will be as tall as their fathers. (In fact, a comparison of entire groups might reveal that sons tend to be slightly taller because of an improvement in nutrition across generations.) (e) False. Given the subset of tall sons, their fathers will tend to be shorter because of regression toward the mean. (f) True
10	**Define Regression line.**	A regression line is a straight line that describes how a response variable y changes as an explanatory variable x changes. ❖A regression line can be used to predict the value of y for a given value of x.

PART B

1	U	Explain about Scatter plot and Various types of Scatterplot with neat diagram. [Nov/Dec 2022]
2	An	Calculate the correlatio n co efficient for the heights of fathers(X) and their sons(y) with the data prese nted below. x 66 68 68 70 71 72 72 y 68 70 69 72 72 72 74
3	A	The values of x and their corresponding values of y are presented below. x 0.5 1.5 2.5 3.5 4.5 5.5 6.5 y 2.5 3.5 5.5 4.5 6.5 8.5 10.5 i) Find the Least square regression line y=ax+b. ii) Estimate the values of y when x=10.
4	E	Calculate Standard Error Estimate

Couple	X	Y
A	1	2
B	3	4
C	2	3
D	3	2
E	1	0

			F	2	3
5	An	Estimate whether the following pairs of scores for x and y a positive			

		relationship, negative relationship or no relationship x 64 40 30 71 55 31 61 42 57 y 66 79 98 65 76 83 68 80 72 a) Construct a scatterplot for x and y verify that scatter does not describe a pronounced curvilinear. b) Calculate r using the Computation formula.
	A & C	Each of the following pairs represents the number of licensed drivers (X) and the number of cars (Y) for seven house in my neighborhood. [Nov/Dec 2022] Drivers (X) Cars (Y) 5 4 5 3 2 2 2 2 3 2 1 1 2 2 1. Construct a scatterplot to verify a lack of pronounced Curvilinearity. 2. Determine the least squares equation for these data. (Remember, you will first have to calculate r,SSy and SSx) Determine the standard error of estimate, Sy/x given that n=7.
		PART C
1.	A	Consider the following dataset with one response variable y and two predictor variables x1 and x2. [Apr/May 2023]

Y	140	155	159	179	192	200	212	215
x1	60	62	67	70	71	72	75	78
x_2	22	25	24	20	15	14	14	11

Fit a multiple linear regression model to this dataset.

<table>
<tr><td>2.</td><td>An</td><td>i)Assume that an r=0.30 describe the relationship between education level and estimate number of hours spent reading each work
<table>
<tr><th>Education level(X)</th><th>Weekly Reading Time(Y)</th></tr>
<tr><td>X=13</td><td>Y =8</td></tr>
<tr><td>SSX=25</td><td>SSY=50</td></tr>
</table>
ii)Determine the least square equation for predicting weekly repot time from education level. iii)Faith’s education level is 15. What is her predicted reading time?
iv)Keegan’s education level is 11. What is his predicted reading time?
v)Calculate the standard error estimate based on n=35 pairs of observation.</td></tr>
<tr><td></td><td></td><td>vi)Supply a rough interpretation of standard error estimate.</td></tr>
<tr><td>3.</td><td>An</td><td>Assume that an of -.80 describe the strong negative relationship between years of heavy smoking (X) and life expectancy(Y). [Nov/Dec 2022]
Assume, furthermore that the distributions of heavy smoking and life expectancy each have the following means and sum of squares:5, 60,35, 70 x,y,SSx,SSy.
i)Determine the least square squares regression equation for predicting life expectancy from years of heavy smoking.(3) ii)Determine the standard error of estimate, SSy/x, assuming that the correlation of -.80 was based on n=50 pairs of observation.(3) iii)Supply a rough interpretation of SSy/x.(3)
iv)Predict the life expectancy for john ,who has smoked heavily for 8 yars.(3)
v)Predict the life expectancy for Katie ,who has never smokes heavily.(3)</td></tr>
</table>

UNIT IV-PYTHON LIBRARIES FOR DATA WRANGLING PART A

1	R	Define Numpy array and list the attributes of numpy array with example. [Nov/Dec 2022] NumPy, attributes are properties of NumPy arrays that provide information about the array's shape, size, data type, dimension, and so on. For example, to get the dimension of an array, we can use the ndim attribute.

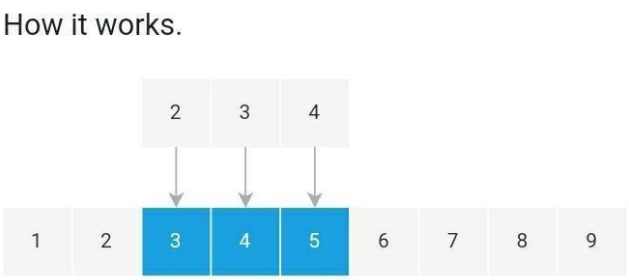
<table>
<tr><td>2</td><td>R</td><td>List Aggregate Function with Example.
Aggregate functions perform an operation on a set of values and
<table><tr><th>Functions</th><th>Description</th></tr><tr><td>np.sum()</td><td>Returns the sum of array elements over a given axis.</td></tr><tr><td>np.prod()</td><td>Returns the product of array elements over a given axis.</td></tr><tr><td>np.mean()</td><td>Computes the arithmetic mean along the specified axis.</td></tr><tr><td>np.std()</td><td>Computes the standard deviation along the specified axis.</td></tr><tr><td>np.var()</td><td>Computes the variance along the specified axis.</td></tr><tr><td>np.min()</td><td>Returns the indices of the minimum values along an axis.</td></tr><tr><td>np.max()</td><td>Returns the indices of the maximum values along an axis.</td></tr><tr><td>np.all()</td><td>Checks if all array elements along a given axis evaluate to True.</td></tr></table>produce a single result.</td></tr>
<tr><td>3</td><td>R</td><td>Define Data Wrangling.
Data wrangling is the process of transforming data from its original "raw" form into a more digestible format and organizing sets from various sources into a singular coherent whole for further processing.</td></tr>
<tr><td>4</td><td>R</td><td>Define Structure Array.
A structured Numpy array is an array of structures. As numpy arrays are homogeneous i.e. they can contain data of same type only. So,</td></tr>
<tr><td></td><td></td><td>instead of creating a numpy array of int or float, we can create numpy array of homogeneous structures too.</td></tr>
<tr><td>5</td><td>C</td><td>State the advantage of Using Numpy arrays. [Apr/May 2023]
NumPy arrays are faster and more compact than Python lists. An array consumes less memory and is convenient to use. NumPy uses much less memory to store data and it provides a mechanism of specifying the data types. This allows the code to be optimized even further.</td></tr>
<tr><td>6</td><td>R</td><td>Outline the two types of Numpy UFunc. [Apr/May 2023]
There are two types of ufuncs: unary ufuncs: take one array (ndarray) as the argument. binary ufuncs: take two arrays (ndarray) as arguments</td></tr>
<tr><td>7</td><td>R</td><td>What is Combining Data set?
With pandas, you can merge, join, and concatenate your datasets, allowing you to unify and better understand your data as you analyze it.
merge() for combining data on common columns or indices(df.merge()
join() for combining data on a key column or an index
concat() for combining DataFrames across rows or columns</td></tr>
</table>

8	R	List the Aggregate Pivot and Grouping function in Pandas. Groupby() is a powerful function in pandas that allows you to group data based on a single column or more. You can apply many operations to a groupby object, including aggregation functions like sum(), mean(), and count(), as well as lambda function and other custom functions using apply() The pivot function in pandas is used to reshape the given data frame based on specific columns. Specified columns act as pivots of the data frame. An important thing to note is that the pivot function does not support data aggregation. Instead, multiple columns will return the data frame, becoming multi-indexed
9	E	i)Convert a 1-D array into a 2-D array with 3 rows **Input:** exercise_2 = np.array([0, 1, 2, 3, 4, 5, 6, 7, 8]) **Sample Output:** [[0, 1, 2] [3, 4, 5] [6, 7, 8]] import numpy as np exercise_2 = np.array([0, 1, 2, 3, 4, 5, 6, 7, 8]) exercise_2.reshape(3,3) print(exercise_2) ii)How to combine many series to form a data frame? import pandas as pd sr1 = pd.Series(['php', 'python', 'java', 'c#', 'c++']) sr2 = pd.Series([1, 2, 3, 4, 5]) print("Original Series:") print(sr1) print(sr2) print(ombine above series to a dataframe:") ser_df = d.DataFrame(sr1, sr2).reset_index() f .head))

10	C	Create a data frame with key and data pairs as A-10,B-20,A-40,C=5,B=10,C=10.Find the sum of each key and display the results a each key group. [Nov/Dec 2022]
		import pandas as pd
		data = {
		
		"A": [10,40],
		"B": [20,10],
		"c" :[5,10]
		}
		df = pd.DataFrame(data)

```
df.sum()
```

PART B

1	U	**i)Describe about Fancy Indexing with Example.(7)** Fancy Indexing means passing an array of indices to access multiple array elements at once. Fancy indexing allows you to index a numpy array using the following: ○ Another numpy array ○ A Python list ○ A sequence of integers  Let's see the following example: import numpy as np a = np.arange(1, 10) print(a) indices = np.array([2, 3, 4]) print(a[indices]) Output: [1 2 3 4 5 6 7 8 9] [3 4 5] **ii)Explain about Comparision,Masks and Boolean Logic.(6)** NumPy also implements comparisonoperators such as < (less than) and > (greater than) as element-wise ufuncs.the result of these comparison operators is always an array with a Boolean data type. **All six of the standard comparison operations are available:** In[4]: x = np.array([1, 2, 3, 4, 5]) In[5]: x <3 *# less than* Out[5]: array([True, True, False, False, False], dtype=bool) In[6]: x >3 *# greater than* Out[6]: array([False, False, False, True, True], dtype=bool) In[7]: x <= 3 *# less than or equal* Out[7]: array([True, True, True, False, False], dtype=bool) In[8]: x >= 3 *# greater than or equal* Out[8]: array([False, False, True, True, True], dtype=bool) In[9]: x != 3 *# not equal* Out[9]: array([True, True, False, True, True], dtype=bool) In[10]: x == 3 *# equal* Out[10]: array([False, False, True, False, False], dtype=bool) **Boolean Arrays as Masks** A more powerful pattern is to use Boolean arrays as masks, to select particular subsets of the data themselves. X=array([[5, 0, 3, 3], [7, 9, 3, 5], [2, 4, 7, 6]])

We can obtain a Boolean array for this condition easily, as we've already seen:
x< 5
array([[False, True, True, True],
[False, False, True, False],
[True, True, False, False]], dtype=bool)
Now to select these values from the array, we can simply index on this Boolean array; this is known as a masking operation:

Boolean Logical Operators

Logical operators are used to combine conditional statements:

Example

x = 5
print(x > 3 and x < 10) **Output:** True x = 5
print(x > 3 or x < 4)
Outpu t True
x = 5
print(not(x > 3 and x < 10)) # returns False because not is used to reverse the result
Output
False

Operator	Description	Example
and	Returns True if both statements are true	x < 5 and x < 10
or	Returns True if one of the statements is true	x < 5 or x < 4
not	Reverse the result, returns False if the	not(x < 5 and x < 10)

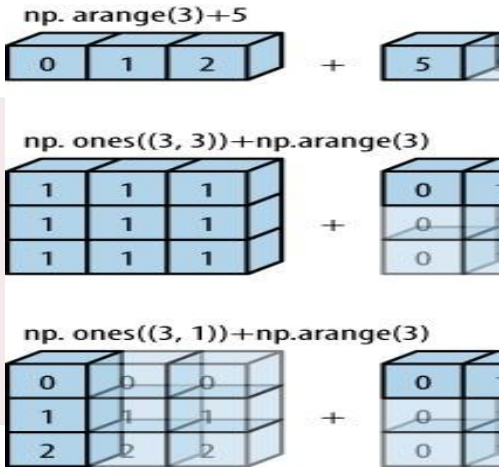
What is an aggregate function? Elaborate about the aggregate functions in numpy. [Apr/May 2023]

U The Python numpy aggregate functions are sum, min, max, mean, average, product, median, standard deviation, variance, argmin, argmax, percentile, cumprod, cumsum, and corrcoef.
Min: Input: x = min(5, 10) Output:5
Max: Input: x = max(5, 10) Output:10

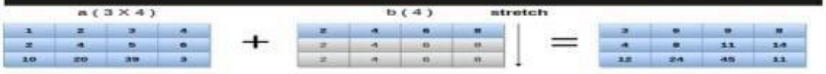
Mean,Mode,Std,Median:

```
import numpy as np
speed=[99,86,87,88,111,86,103,87,94,78,77,85,86]
x=np.mean(speed)
print(x)
```

o/p 89.7692307692307

2	U	**What is broadcasting and explain the rules with Example. [Apr/May 2023]**
		NumPy‘s broadcasting functionality. Broadcasting is simply aset of rules for applying binary ufuncs (addition, subtraction, multiplication, etc.) onarrays of different sizes. we can perform other operations such as subtraction, multiplication and division. **Consider the below example**
		import numpy as np
		x= np.array([(1,2,3),(3,4,5)])
		y= np.array([(1,2,3),(3,4,5)])
		print(x-y)
		print(x*y)
		print(x/y)
		Output – [[0 0 0] [0 0 0]]
		[[1 4 9] [9 16 25]]
		[[1. 1. 1.][1. 1. 1.]
		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.
		Example
		`import numpy as np`
		`a = np.array([[1,2,3,4],[2,4,5,6],[10,20,39,3]])`
		`b = np.array([2,4,6,8])`
		`print("\nprinting array a..")`
		`print(a)`
		`print("\nprinting array b..")`
		`print(b)`
		`print("\nAdding arrays a and b ..")`
		`c = a + b;`
		`print(c)`
		 

```
printing array a..
[[ 1  2  3  4]
 [ 2  4  5  6]
 [10 20 39  3]]

printing array b..
[2 4 6 8]

Adding arrays a and b ..
[[ 3  6  9 12]
 [ 4  8 11 14]
 [12 24 45 11]]
```



a = [[1, 2, 3, 4], [2, 4, 5, 6], [10, 20, 39, 3]]
b = [[2, 4, 6, 8]]

3 U

Describe the various methods of handling the missing data in Pandas In DataFrame sometimes many datasets simply arrive with missing data, either because it exists and was not collected or it never existed**.**

In Pandas missing data is represented by two value:

1.None: None is a Python singleton object that is often used for missing data in Python code.

2.NaN : NaN (an acronym for Not a Number), is a special floating-point value recognized by all systems that use the standard IEEE floating-point representation

```
In [1]: 1 import pandas as pd
        2 import numpy as np
```

```
In [2]: 1 s=pd.Series(["Sam",np.nan,"Tim","Kim"])
        2 s
```

```
Out[2]: 0    Sam
        1    NaN
        2    Tim
        3    Kim
        dtype: object
```

Funtion:

```
In [3]: 1 s.isnull()
```

```
Out[3]: 0   False
        1    True
        2   False
        3   False
        dtype: bool
```

1.isnull()

```
s.notnull()
```

```
True
False
True
True
pe: bool
```

2.notnull()

3.dropna():

4.fillna():

If you want to assign another value instead

		of missing data, you can use the fillna method. the dropna method removes rows with missing In [16]: `df.fillna(0)` Out[16]: 0 1 2 0 1.0 0.0 3.0 1 4.0 0.0 5.0 2 0.0 0.0 0.0

4	R	**i)Briefly explain about Hierarchical Indexing.**
		Hierarchical indexing is a method of creating structured group relationships in the dataset. Data frames can have hierarchical indexes. To show this, let me create a dataset.
		In [10]:
		`df=pd.DataFrame(`
		`np.arange(12).reshape(4,3),`
		`index=[["a","a","b","b"],`
		`[1,2,1,2]],`
		`columns=[["num","num","ver"],`
		`["math","stat","geo"]])`
		`df`
		Out[10]:
		num ver
		math stat geo
		a 1 0 1 2
		2 3 4 5
		b 1 6 7 8
		2 9 10 11
		ii)Demonstrate different ways of creating Pandas data frame.
		Create Pandas Dataframe in Python
		There are several ways to create a Dataframe in Pandas Dataframe. Here are some of the most common methods:
		• Create Pandas DataFrame from list of lists
		• Create Pandas DataFrame from dictionary of numpy array/list
		• Creating Dataframe from list of dicts
		• Create Pandas DataFrame from list of dictionaries
		• Create Pandas Dataframe from dictionary of Pandas Series
		• Creating DataFrame using zip() function
		• Creating a DataFrame by proving index label explicitly
		# Importing Pandas to create DataFrame import
		pandas as pd
		
		# Creating Empty DataFrame and Storing it in variable df df
		= pd.DataFrame()
		
		# Printing Empty DataFrame
		print(df)

5	C	i)Image you have a series of data that represents the amount of precipitation each day for a year in a given city. Load the daily rainfall statistics for the city of Chennai in 2021. Which is given in a csv file Chennai rainfall 2021.csv using Pandas generate a histogram for rainy days and find out the days that have high rainfall. [Nov/Dec 2022] Chennai rainfall 2021.csv

First five lines of rain dataset:

country	precip	area
Afghanistan	327.0	652.2
Albania	1485.0	27.4
Algeria	89.0	2381.7
American Samoa	NaN	0.2
Andorra	NaN	0.5
Angola	1010.0	1246.7
Antigua and Barbuda	1030.0	0.4
Argentina	591.0	2736.7
Armenia	562.0	28.5
Aruba	NaN	0.2

Program

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
rain = pd.read_csv(—Chennai rainfall2021.csv —)
 rain[‘country’].max()
rain.hist()
```

ii) Consider that an E commerce organization like Amazon have different region sales as Northsales, Southsales, Westsales.csv files. They want combine North and west region ales and south and east sales to find the aggregate sales of this collaborating region help them to do so using python code. [Nov/Dec 2022]

```
 import pandas as pd
ecom=pd.read_csv('../input/ecommerce-purchases-csv/Ecommerce Purchases.csv')
ecom.info( )
```

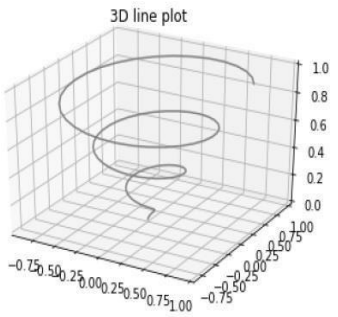
#	Column	Non-Null Count	Dtype
0	Address	10000 non-null	object
1	Lot	10000 non-null	object
2	AM or PM	10000 non-null	object
3	Browser Info	10000 non-null	object
4	Company	10000 non-null	object
5	Credit Card	10000 non-null	int64
6	CC Exp Date	10000 non-null	object
7	CC Security Code	10000 non-null	int64
8	CC Provider	10000 non-null	object
9	Email	10000 non-null	object
10	Job	10000 non-null	object
11	IP Address	10000 non-null	object
12	Language	10000 non-null	object
13	Purchase Price	10000 non-null	float64

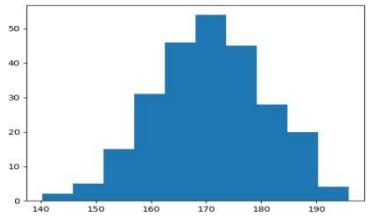
```
Ecom[‘purchase price’].max( )

Ecom[‘purchase price’].min( )
```

UNIT V-DATA VISUALIZATION

PART A

1	R	**What is the purpose of error bar function in Matplotlib? Give an example. [Nov/Dec 2022]** The errorbar() function in pyplot module of matplotlib library is used to plot y versus x as lines and/or markers with attached errorbars. Parameters: This method accept the following parameters that are described below: x, y: These parameter are the horizontal and vertical coordinates of the data points
2	C	**Write the command for Text annotations with Example.** Annotations are graphical elements, often pieces of text, that explain, add context to, or otherwise highlight some portion of the visualized data. annotate supports a number of coordinate systems for flexibly positioning data and annotations relative to each other and a variety of options of for styling the text.
3	R,An	**i)Define line Plot and Subplot.** A line graph—also known as a line plot or a line chart—is a graph that uses lines to connect individual data points. A line graph displays quantitative values over a specified time interval. A subplot is otherwise known as a minor story or a secondary plot which often runs parallel to the main plot. It can be about your main character(s) or about another character whose narrative interacts or impacts their narrative. **ii)How plt.scatter function differ from plt.flot function.[Apr/May 2023]** The primary difference of plt. scatter from plt. plot is that it can be used to create scatter plots where the properties of each individual point (size, face color, edge color, etc.) can be individually controlled or mapped to data.
4	R	**What is Legend & Color with Example?** A legend is an area describing the elements of the graph. In the matplotlib library, there's a function called **legend()** which is used to Place a lege import matplotlib.pyplot as plt import numpy as np y = np.array([35, 25, 25, 15]) mylabels = ["Apples", "Bananas", "Cherries", "Dates"] plt.pie(y, labels = mylabels) plt.legend(title = "Four Fruits:") plt.show() nd on the axes. **The colors parameter, if specified, must be an array with one value for each wedge**: import matplotlib.pyplot as plt import numpy as np y = np.array([35, 25, 25, 15])

		mylabels = ["Apples", "Bananas", "Cherries", "Dates"] mycolors = ["black", "hotpink", "b", "#4CAF50"] plt.pie(y, labels = mylabels, colors = mycolors) plt.show()
5	U	**Briefly explain Visualizing Error with example** errorbar() method is used to create a line plot with error bars. The two positional arguments supplied to ax. errorbar() are the lists or arrays of x, y data points. The two keyword arguments xerr= and yerr= define the error bar lengths in the x and y directions.
6	R	**What is the use of Seaborn?** Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas data structures. Seaborn helps you explore and understand your data.
7	C	**Showcase 3 dimensions drawing in matplotlib with corresponding Python code. [Nov/Dec 2022]** from mpl_toolkits import mplot3d import numpy as np import matplotlib.pyplot as plt fig = plt.figure() ax = plt.axes(projection='3d') z = np.linspace(0, 1, 100) x = z * np.sin(20 * z) y = z * np.cos(20 * z) ax.plot3D(x, y, z, 'gray') ax.set_title('3D line plot') plt.show() 
8	R	**Define Data Visualization**. Data visualization is the representation of data through use of common graphics, such as charts, plots, infographics, and even animations. These visual displays of information communicate complex data relationships and data-driven insights in a way that is easy to understand
9	R	**What functions to be used to draw the scatterplot?** It can simply use the scatter() function. This function is used to plot one dot for each observation. It accepts two arrays of the same length for the x and y-axis. Where x and y can be the NumPy arrays.

10	R	**What is Histogram with Example diagram?** A histogram is a graph showing *frequency* distributions. It is a graph showing the number of observations within each given interval Create Histogram import matplotlib.pyplot as plt import numpy as np x = np.random.normal(170, 10, 250) plt.hist(x) plt.show() 

PART B

1	C	Appraise the flowing Density & Contour Plot , Histograms and Binning with appropriate in python code. [Nov/Dec 2022]

Density & Contour Plot

The **matplotlib.pyplot.contour()** are usually useful when **Z = f(X, Y)** i.e Z changes as a function of input X and Y. A contourf() is also available which allows us to draw filled contours.

Syntax: *matplotlib.pyplot.contour([X, Y,] Z, [levels], **kwargs)*

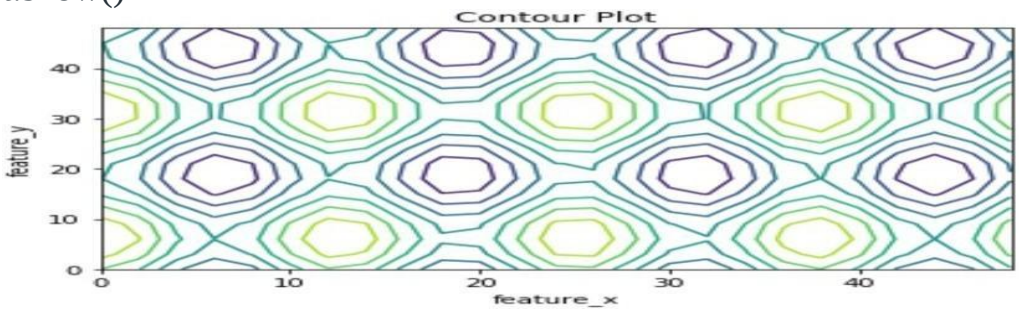
Parameters:

X, Y: *2-D numpy arrays with same shape as Z or 1-D arrays such that len(X)==M and len(Y)==N (where M and N are rows and columns of Z)*

Z: *The height values over which the contour is drawn. Shape is (M, N)*

levels: *Determines the number and positions of the contour lines / regions.*

```
# Implementation of matplotlib function
import matplotlib.pyplot as plt
import numpy as np feature_x
= np.arange(0, 50, 2) feature_y
= np.arange(0, 50, 3)
# Creating 2-D grid of features
[X, Y] = np.meshgrid(feature_x, feature_y)
fig, ax = plt.subplots(1, 1)
Z = np.cos(X / 2) + np.sin(Y / 4)
# plots contour lines
ax.contour(X, Y, Z)
ax.set_title('Contour Plot')
ax.set_ylabel('feature_y')
plt.show()
```



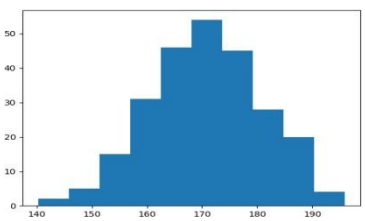
histogram

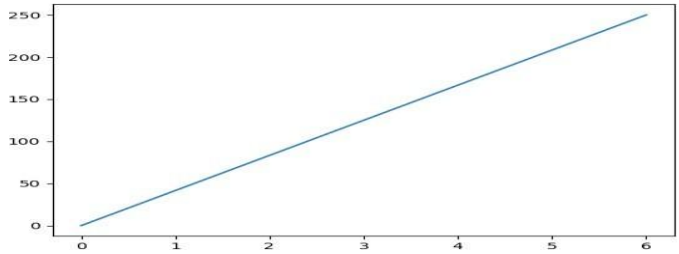
A histogram is a graph showing *frequency* distributions.

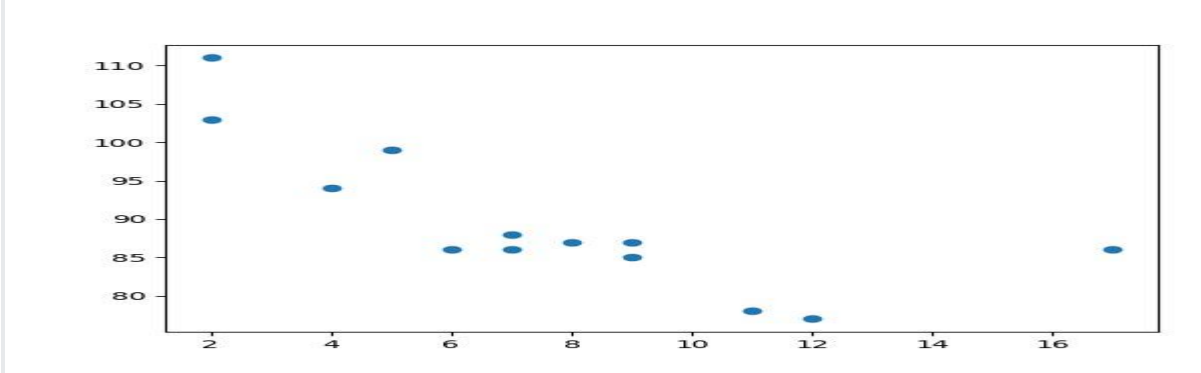
It is a graph showing the number of observations within each given interval

Create Histogram

```
import matplotlib.pyplot as plt
import numpy as npx
= np.random.normal(170, 10, 250)
plt.hist(x) plt.show()
```



		The towers or bars of a histogram are called bins. The height of each bin shows how many values from that data fall into that range. A histogram displays numerical data by grouping data into "bins" of equal width. Each bin is plotted as a bar whose height corresponds to how many data points are in that bin. Bins are also sometimes called "intervals", "classes", or "buckets".
2	U	**Explain about various visualization charts like line plots, scatter plts and histograms using Matplotlib with an example. [Apr/May 2023]** **line plots** Importing Matplotlib The Pyplot package can be referred to as plt. import matplotlib.pyplot as plt Example Draw a line in a diagram from position (0,0) to position (6,250): import matplotlib.pyplot as plt import numpy as np xpoints = np.array([0, 6]) ypoints = np.array([0, 250]) plt.plot(xpoints, ypoints) plt.show() Result:  **scatter plts** The scatter() function plots one dot for each observation. It needs two arrays of the same length, one for the values of the x-axis, and one for values on the y-axis: Example A simple scatter plot: import matplotlib.pyplot as plt import numpy as np x = np.array([5,7,8,7,2,17,2,9,4,11,12,9,6]) y = np.array([99,86,87,88,111,86,103,87,94,78,77,85,86]) plt.scatter(x, y) plt.show() Result:



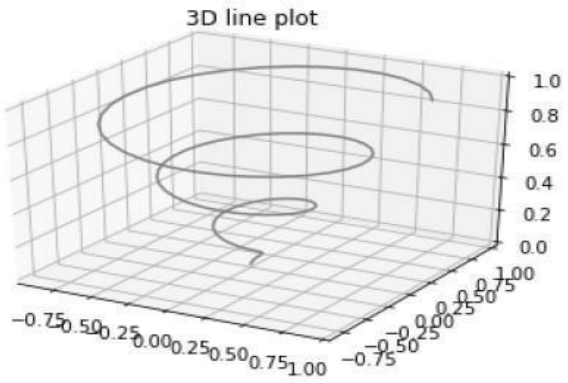
histogram

A histogram is a graph showing *frequency* distributions.

It is a graph showing the number of observations within each given interval

Create Histogram

```
import matplotlib.pyplot as plt
import numpy as np
x = np.random.normal(170, 10, 250)
plt.hist(x) plt.show()
```

3	R	**Briefly Explain 3 Dimensional Plotting with an example. [Apr/May 2023]** A three-dimensional axes can be created by passing the keyword projection='3d' to any of the normal axes creation routines. `from mpl_toolkits import mplot3d import numpy as np import matplotlib.pyplot as plt fig = plt.figure() ax = plt.axes(projection='3d') z = np.linspace(0, 1, 100) x = z * np.sin(20 * z) y = z * np.cos(20 * z) ax.plot3D(x, y, z, 'gray') ax.set_title('3D line plot') plt.show()` We can now plot a variety of three-dimensional plot types. The most basic three-dimensional plot is a **3D line plot** created from sets of (x, y, z) triples. This can be created using the ax.plot3D function. 
4	R	**Discuss about Geographic base map and Seaborn**. One common type of visualization in data science is that of geographic data. Matplotlib's main tool for this type of visualization is the Basemap toolkit, which is one of several Matplotlib toolkits which lives under the mpl_toolkits namespace. Admittedly, Basemap feels a bit clunky to use, and often even simple visualizations take much longer to render than you might hope. More modern solutions such as leaflet or the Google Maps API may be a better choice for more intensive map visualizations. Still, Basemap is a useful tool for Python users to have in their virtual toolbelts. In this section, we'll show several examples of the type of map visualization that is possible with this toolkit. Installation of Basemap is straightforward; if you're using conda you can type this and the package will be downloaded: `$ conda install basemap`

We add just a single new import to our standard boilerplate:

In [1]:

```
%matplotlib inline import
numpy as np import
matplotlib.pyplot as plt
from mpl_toolkits.basemap import Basemap
```

Once you have the Basemap toolkit installed and imported, geographic plots are just a few lines away (the graphics in the following also requires the PIL package in Python 2, or the pillow package in Python 3): In [2]:

```
plt.figure(figsize=(8, 8))
m = Basemap(projection='ortho', resolution=None, lat_0=50, lon_0=-100)
m.bluemarble(scale=0.5);
```

The meaning of the arguments to Basemap will be discussed momentarily.

5	C	**How text and image annotations are done using python? Give an example of your own with appropriate Python code. [Nov/Dec 2022]**

matplotlib.pyplot.annotate() Function

The **annotate() function** in pyplot module of matplotlib library is used to annotate the point xy with text s.

Syntax: angle_spectrum(x, Fs=2, Fc=0, window=mlab.window_hanning, pad_to=None, sides='default', **kwargs)

Parameters: This method accept the following parameters that are described below:

- **s:** This parameter is the text of the annotation.
- **xy:** This parameter is the point (x, y) to annotate.
- **xytext:** This parameter is an optional parameter. It is The position (x, y) to place the text at.
- **xycoords:** This parameter is also an optional parameter and contains the string value.
- **textcoords:** This parameter contains the string value.Coordinate system that xytext is given, which may be different than the coordinate system used for xy

- **arrowprops :** This parameter is also an optional parameter and contains dict type.Its default value is None.
- **annotation_clip :** This parameter is also an optional parameter and contains boolean value.Its default value is None which behaves as True.

```
# Implementation of matplotlib.pyplot.annotate()
# function
import matplotlib.pyplot as plt
import numpy as np fig,
geeeks = plt.subplots() t =
np.arange(0.0, 5.0, 0.001) s =
np.cos(3 * np.pi * t)
line = geeeks.plot(t, s, lw = 2)
# Annotation
geeeks.annotate('Local Max', xy =(3.3, 1),
                                xytext =(3, 1.8),
                arrowprops = dict(facecolor ='green', shrink = 0.05),)
geeeks.set_ylim(-2, 2)
# Plot the Annotation in the graph
plt.show()
```

OUTPUT

UNIT III Correlation and Regression

CALCULATION OF *r*: COMPUTATION FORMULA

A. COMPUTATIONAL SEQUENCE

Assign a value to n (1), representing the number of pairs of scores.
Sum all scores for X (2) and for Y (3).
Find the product of each pair of X and Y scores (4), one at a time, then add all of these products (5).
Square each X score (6), one at a time, then add all squared X scores (7).
Square each Y score (8), one at a time, then add all squared Y scores (9).
Substitute numbers into formulas (10) and solve for SP_{xy}, SS_x, and SS_y.
Substitute into formula (11) and solve for r.

B. DATA AND COMPUTATIONS

	CARDS		4	6	8
FRIEND	SENT, X	RECEIVED, Y	XY	X^2	Y^2
Doris	13	14	182	169	196
Steve	9	18	162	81	324
Mike	7	12	84	49	144
Andrea	5	10	50	25	100
John	1	6	6	1	36

1 $n = 5$ 2 $\Sigma X = 35$ 3 $\Sigma Y = 60$ 5 $\Sigma XY = 484$ 7 $\Sigma X^2 = 325$ 9 $\Sigma Y^2 = 800$

10 $$SP_{xy} = \Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n} = 484 - \frac{(35)(60)}{5} = 484 - 420 = 64$$

$$SS_x = \Sigma X^2 - \frac{(\Sigma X)^2}{n} = 325 - \frac{(35)^2}{5} = 325 - 245 = 80$$

$$SS_y = \Sigma Y^2 - \frac{(\Sigma Y)^2}{n} = 800 - \frac{(60)^2}{5} = 800 - 720 = 80$$

11 $$r = \frac{SP_{xy}}{\sqrt{SS_x SS_y}} = \frac{64}{\sqrt{(80)(80)}} = \frac{64}{80} = .80$$

DETERMINING THE LEAST SQUARES REGRESSION EQUATION

A. COMPUTATIONAL SEQUENCE

Determine values of SS_x, SS_y, and r (1) by referring to the original correlation analysis in Table 6.3.

Substitute numbers into the formula (2) and solve for b.

Assign values to $\overline{X}$ and $\overline{Y}$ (3) by referring to the original correlation analysis in Table 6.3.

Substitute numbers into the formula (4) and solve for a.

Substitute numbers for b and a in the least squares regression equation (5).

B. COMPUTATIONS

1 $SS_x = 80^*$

$SS_y = 80^*$

$r = .80$

2 $$b = r\sqrt{\frac{SS_Y}{SS_X}} = .80\sqrt{\frac{80}{80}} = .80$$

3 $\overline{X} = 7^{**}$

$\overline{Y} = 12^{**}$

4 $$a = \overline{Y} - (b)(\overline{X}) = 12 - (.80)(7) = 12 - 5.60 = 6.40$$

5 $$Y' = (b)(X) + a$$

$$= (.80)(X) + 6.40$$

$$Y' = .80(11) + 6.40$$

$$= 8.80 + 6.40$$

$$= 15.20$$

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

A. COMPUTATIONAL SEQUENCE

Assign values to SS_y and r (1) by referring to previous work with the least squares regression equation in Table 7.1.

Substitute numbers into the formula (2) and solve for $s_{y|x}$.

B. COMPUTATIONS

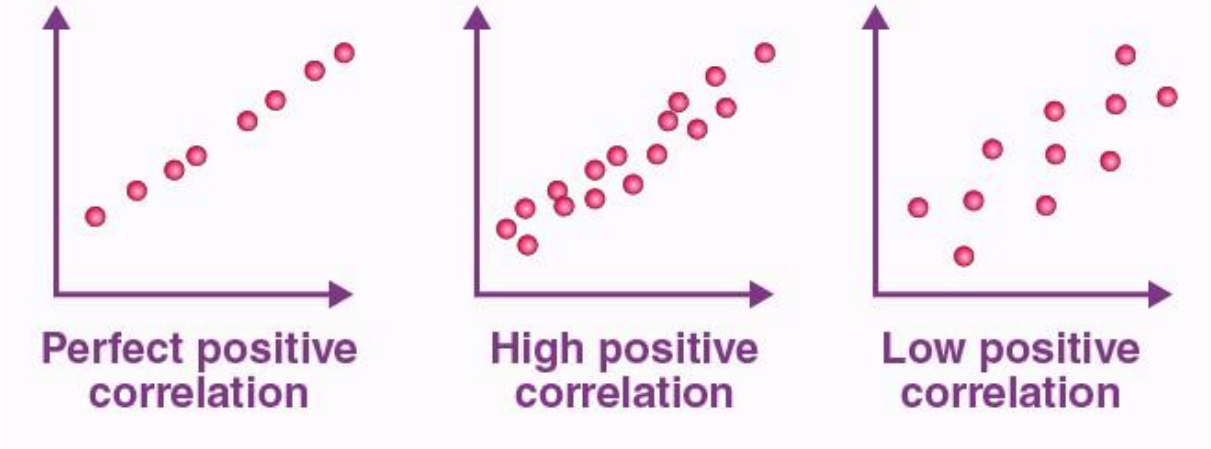
1 $SS_y = 80$

$r = .80$

2 $$s_{y|x} = \sqrt{\frac{SS_y(1-r^2)}{n-2}} = \sqrt{\frac{80\left(1-\left[.80\right]^2\right)}{5-2}} = \sqrt{\frac{80(.36)}{3}} = \sqrt{\frac{28.80}{3}} = \sqrt{9.60}$$

$$= 3.10$$

Scatterplot

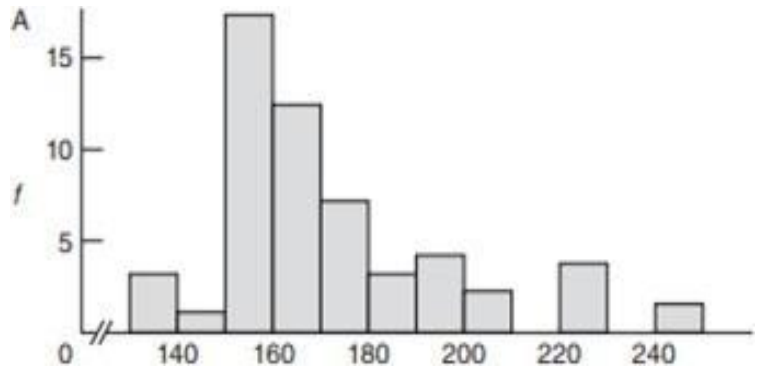


UNIT II

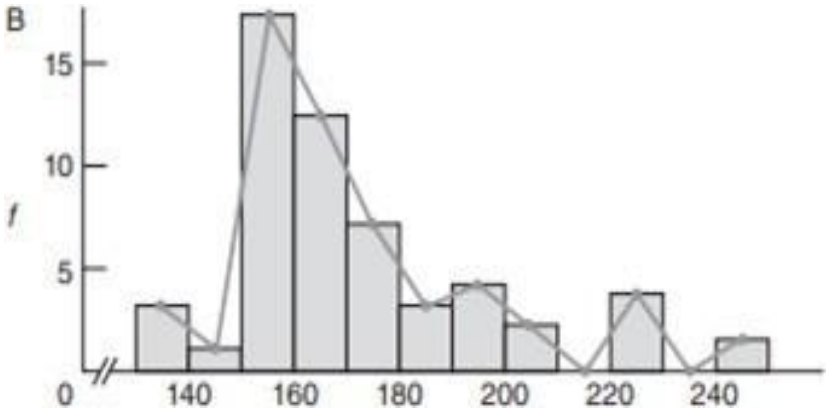
1.Grouped Data

WEIGHT	*f*	CUMULATIVE *f*	CUMULATIVE PERCENT
240–249	1	53	100
230–239	0	52	98
220–229	3	52	98
210–219	0	49	92
200–209	2	49	92
190–199	4	47	89
180–189	3	43	81
170–179	7	40	75
160–169	12	33	62
150–159	17	21	40
140–149	1	4	8
130–139	3	3	6
Total	53		

Histogram



Frequency Polygon



2. **Un grouped Date**

91	85	84	79	80
87	96	75	86	104
95	71	105	90	77
123	80	100	93	108
98	69	99	95	90
110	109	94	100	103
112	90	90	98	89

(a) Calculating the class width, $\frac{123 - 69}{10} \quad \frac{54}{10} \quad 5.4$

IQ	*f*
120–124	1
115–119	0
110–114	2
105–109	3
100–104	4
95–99	6
90–94	7
85–89	4
80–84	3
75–79	3
70–74	1
65–69	$\frac{1}{35}$
Total	