BETHLAHEM INSTITUTE OF ENGINEERING

KARUNGAL

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

III SEMESTER

CS3352 – Foundation of Data Science

Regulation – 2021

Academic Year 2022 –2023

Prepared by

Ms.G.Marly, Assistant Professor/CSE

CS3352- Foundation of Data Science

UNIT I	**INTRODUCTION**		
Data Science: Benefits and uses – facets of data - Data Science Process: Overview – Definingresearch goals – Retrieving data – Data preparation - Exploratory Data analysis – build the model–presenting findings and building applications - Data Mining - Data Warehousing – Basic Statistical descriptions of Data			
PART-A			
Q. No	**Questions**	**BT Level**	**Competence**
1	1. What is Data Science? Data Science is a combination of multiple disciplines that uses statistics, data analysis, and machine learning to analyze data and to extract knowledge and insights from it. Data Science is about data gathering, analysis and decision-making. Also, it is about finding patterns in data, through analysis, and make future predictions. Data science and big data are used almost everywhere in both commercial and non-commercial settings. By using Data Science, companies are able to make: • Better decisions (should we choose A or B) • Predictive analysis (what will happen next?) • Pattern discoveries (find pattern, or maybe hidden information in the data)	BTL1	Remember
2	2. What is big data? Big Data is a collection of data that is huge in volume, yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently. Big data is also a data but with huge size. The characteristics of big data are often referred to as the three Vs: Volume—How much data is there? Variety—How diverse are different types of data? Velocity—At what speed is new data generated?	BTL2	Understand
3	List the facets of data. Structured data Unstructured data Natural Language Machine-generated Graph-based Audio, video and images Streaming	BTL1	Remember
4	Applications of data Science. ● Fraud and Risk Detection	BTL3	Apply

	• Healthcare • Internet Search • Targeted Advertising • Website Recommendations • Advanced Image Recognition • Speech Recognition • Airline Route Planning • Gaming • Augmented Reality		
5	List the benefits and uses of data Science? - Increases business predictability - Ensures real-time intelligence - Favors the marketing and sales area - Improves data security - Helps interpret complex data - Facilitates the decision-making process - Study purpose	BTL1	Remember
6	What are all difference sources of unstructured data? - Web pages - Images (JPEG, GIF, PNG, etc.) - Videos - Memos - Reports - Emails - Surveys	BTL4	Analyze
7	What is NLP? - Natural Language Processing or NLP is a branch that focuses on teaching computers how to read and interpret the text in the same way as humans do. It is a field that is developing methodologies for filling the gap between Data Science and human languages. - Many areas like Healthcare, Finance, Media, Human Resources, etc are using NLP for utilizing the data available in the form of text and speech. Many text and speech recognition applications are built using NLP. - The natural language processing community has had success in entity recognition, topic recognition, summarization, text completion, and sentiment analysis, but models trained in one domain don't generalize well to other domains.	BTL1	Remember
8	What is Machine data? List the different types of Machine data. Machine data, also known as machine-generated data, is information that is created without human interaction as a result of a computer process or application activity. This means that data entered manually by an end-user is not recognized to be machine-generated. These data affect all industries that use computers in their daily operations, and individuals are increasingly generating this data inadvertently or causing it to be generated by the machine. The different types of machine data are, • Sensor Data • Computer or System Log Data • Geotag Data • Call Log Data	BTL1	Remember

	• Web Log Data		
9	What are the different steps involved in data science process? - Setting the research goal - Retrieving data - Data preparation - Data exploration - Data modelling - Presentation and automation	BTL1	Remember
10	List out the contents of project charter. ● A clear research goal ● The project mission and context ● Set a Budget ● Assess scope and Risks. ● How you're going to perform your analysis ● What resources you expect to use ● Proof that it's an achievable project, or proof of concepts ● Deliverables and a measure of success ● A timeline	BTL1	Remember
11	List the different types of data cleaning techniques. ● Remove duplicates or Data entry errors ● Remove irrelevant data ● Standardize capitalization ● Convert data type ● Clear formatting ● Fix errors ● Language translation ● Handle missing values	BTL1	Remember
12	What are the steps involved in building a data model? Building a model is an iterative process. The way for build the model depends on whether you go with classic statistics or the somewhat more recent machine learning. most of the models consist of the following main steps: •Selection of a modeling technique and variables to enter in the model •Execution of the model •Diagnosis and model comparison	BTL1	Remember
13	What is data mining? *Data mining is the process of sorting through large data sets to identify patterns and relationships that can help solve business problems through data analysis. Data mining techniques and tools enable enterprises to predict future trends and make more-informed business decisions. *Data mining is a key part of data analytic overall and one of the core disciplines in data science, which uses advanced analytic techniques to find useful information in data sets..	BTL5	Evaluate
14	List out the steps involved in data mining process. ● Data gathering ● Data preparation ● Mining the data ● Data analysis and interpretation	BTL5	Evaluate
15	Mention the different types of data mining techniques. ● Association rule mining. ● Classification. ● Clustering. ● Regression	BTL5	Evaluate

	• Sequence and path analysis • Neural networks.		
	List out the benefits of data mining. • More effective marketing and sales. • Better customer service. • Improved supply chain management. • Increased production uptime • Stronger risk management • Lower costs		
	What is data warehousing? A data warehouse is a central repository of information that can be analyzed to make more informed decisions. Data flows into a data warehouse from transaction systems, relational databases, and other sources, typically on a regular cadence. Business analysts, data engineers, data scientists, and decision makers access the data through business intelligence (BI) tools, SQL clients, and other analytic applications. Data and analytic have become indispensable to businesses to stay competitive. Business users rely on reports, dashboards, and analytic tools to extract insights from their data, monitor business performance, and support decision making.		
	What is statistical distribution of data? Mention the different types of distribution. The distribution provides a parameterized mathematical function that can be used to calculate the probability for any individual observation from the sample space. This distribution describes the grouping or the density of the observations, called the probability density function. We can also calculate the likelihood of an observation having a value equal to or lesser than a given value. A summary of these relationships between observations is called a cumulative density function. The different types of distribution includes, • Gaussian Distribution • Student's t-Distribution • Chi-Squared Distribution		

PART-B

Q. No	Questions	BT Level	Competence
1	How does a Data Scientist work? List out the various benefits and uses of data science.	BTL4	Analyze
2	Explain about different facets of data with it's sources and characteristic.	BTL3	Apply
3	Explain in details about setting the research goal step under data science process.	BTL1	Remember
4	Explain about retrieving data under data science process.	BTL1	Remember
5	Explain in detail about data preparation process.	BTL6	Create
6	Explain in detail about different data cleaning techniques.	BTL2	Understand
7	Explain in detail about data integration and transformation.	BTL5	Evaluate

PART-C

Q. No	Questions	BT Level	Competence
1	Explain about exploratory data analysis.	BTL4	Analyze
2	Explain about how to build a model. Also mention the importance of machine learning in building a model.	BTL3	Apply
3	What is data mining? Explain about different data mining techniques.	BTL1	Remember
4	Explain about different components of data ware housing with a diagram.	BTL1	Remember
5	Explain about three different statistical descriptions of data.	BTL6	Create
6	Explain about exploratory data analysis.	BTL2	Understand
7	Explain about how to build a model. Also mention the importance of machine learning in building a model.	BTL5	Evaluate

UNIT II DESCRIBING DATA

Types of Data - Types of Variables -Describing Data with Tables and Graphs –Describing Data with Averages - Describing Variability - Normal Distributions and Standard (z) Scores

PART-A

Q. No	Questions	BT Level	Competence
1	Define data. What are the types of data? Data is a collection of actual observations or scores in a survey or an experiment Any statistical analysis is performed on data.Data can be broadly classified into qualitative and quantitative.	BTL1	Remember
2	What is Qualitative data? Give example. Qualitative or Categorical data is a set of observations where any single observation is a word, letter, or numerical code that represents a class or a category Examples: Words Yes or No, Letters - Y or N, Numerical code - 0 or 1	BTL1	Remember
3	What is quantitative data? Give an example Quantitative Data is a set of observations where any single observation is a number that represents an amount or a count. It can be expressed in numerical values, which make it countable and includes statistical data analysis. It is also known as numerical data. Example: Weights: 35, 56. 70 kg	BTL3	Apply
4	What are the types of Frequency Distribution? : Grouped frequency distribution : Ungrouped frequency distribution : Cumulative frequency distribution : Relative frequency distribution :Relative cumulative frequency distribution	BTL 1	Remember
5	Define an outlier.	BTL 4	Analyze

	An outlier is a data point that differs significantly from other observations. An outlier can occur due to variability in the measurement or it may indicate an experimental error.		
6	What is percentile rank? Percentile Rank (PR) of an observation is the percentage of scores in the entire distribution with equal or smaller values than that score. Its mathematical formula is PR= CF-(0.5 F)/N x 100	BTL 4	Analyze
7	What are the measures of central tendency? Mean Median Mode	BTL 2	Understand
8	Define Mode. The mode represents the value of the most frequently occurring score.	BTL3	Apply
9	Define Median. Median represents the middle value when observations are ordered from least to most.	BTL1	Remember
10	What is a Positively Skewed Distribution? Positively Skewed Distribution is a distribution that includes a few extreme observations in the positive direction (to the right of the majority of observations	BTL 2	Understand
11	What is a Negatively Skewed Distribution? Negatively Skewed Distribution is a distribution that includes a few extreme observations in the negative direction (to the left of the majority of observations),	BTL1	Remember
12	What is z Score? A 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. Where, X is the original score, μ and a are the mean and the standard deviation.	BTL5	Evaluate
13	What is nominal data? A nominal data is the 1st level of measurement scale in which the members serve a “tags” or “labels” to classify or identify the objects. Nominal data is type of qualitative data. A nominal data usually deals with the non-numeric variables or the numbers that do not have any value. While developing statistical models, nominal data are usually transformed before building the model	BTL1	Remember
14	Describe ordinal data. Ordinal data is a variable in which the value of the data is captured from an ordered set, which is recorded in the order of magnitude. Ordinal represents the "order”. Ordinal data is known as qualitative data or categorical data. It can be grouped, named and also ranked.	BTL 6	Create
15	What is an interval data? Interval data corresponds to a variable in which the value is chosen from an interval set. It is defined as a quantitative measurement scale in which the difference between the two variables is meaningful. In other words, the variables are measured in an exact manner, not as in a relative way in which the presence of zero is arbitrary.	BTL2	Understand
16	What is frequency distribution? Frequency distribution is a representation, either in a graphical or tabular format, which displays the number of observations within a given interval. The interval size depends on the data being analysed and the goals of the analyst.		

17	What is cumulative frequency? A cumulative frequency distribution can be useful for ordered data (e.g. data arranged in intervals, measurement data, etc.). Instead of reporting frequencies, the recorded values are the sum of all frequencies for values less than and including the current value.		
18	What is Steam and Leaf diagram? Stem and leaf diagrams allow to display raw data visually. Each raw score is divided into a stem and a leaf. The leaf is typically the last digit of the raw value. The stem is the remaining digits of the raw value. Data points are split into a leaf (usually the ones digit) and a stem (the other digits).		

PART-B

Q. No	Questions	BT Level	Competence
1	Elaborate the different ways to describe or represent data using tables with suitable examples.		
2	Explain the various ways by which data can be represented or described using graphs with suitable examples and diagrams.		
3	Explain the different measures of central tendency and describe the suitable measures for the different types of data distribution		
4	Explain the different types of frequency distribution with suitable examples and diagrams		
5	Explain the various measures of variability with suitable examples. –		

PARTC

Q. No	Questions	BT Level	Competence
1	Construct the frequency table and draw bar graph, stem and leaf displays for the following data: 139, 145, 150, 145, 136, 150, 152, 144, 138, 138		
2	Construct the histogram and convert it to a frequency polygon for the following data: 138, 139, 139,145, 145, 150, 145, 136, 150, 152, 144, 138, 138, 150, 149, 133,134, 152, 155, 151**.**		
3	Compute the mean, median and mode for the following : 45, 55, 60, 60, 63, 63, 63, 63, 65. 65, 70 :26.9, 26.3. 28.7,27.4, 26.6, 27.4. 26.9, 26.9		
4	Using the computation formula for the sum of squares, calculate the population standard deviation and the sample standard deviation for the scores: 1. 3. 7, 2, 0, 4, 3, 7 10, 8, 5, 0, 1, 7, 9, 2, 1		
5	Consider the test scores approximating a normal curve with a mean of 500 and a standard deviation of 100.Sketch a normal curve and shade in the target area described by the following: more than 570 less than 515 between 520 and 540		

	Plan solutions for the target areas. Convert to z scores and find proportions that correspond to the target areas.		

UNIT III DESCRIBING RELATIONSHIPS			
Correlation –Scatter plots –correlation coefficient for quantitative data –computational formula for correlation coefficient – Regression –regression line –least squares regression line – Standard error of estimate – interpretation of r2 –multiple regression equations –regression towards the mean.			
PART-A			
Q. No	**Questions**	**BT Level**	**Competence**
1	What is correlation? Correlation refers to a relationship between two or more objects. In statistics, the word correlation refers to the relationship between two variables. Correlation exists between two variables when one of them is related to the other in some way.	BTL-1	Remember
2	Define positive and negative correlation. Positive correlation Association between variables such that high scores on one variable tends to have high scores on the other variable. A direct relation between the variables. Negative correlation: Association between variables such that high scores on one variable tends to have low scores on the other variable. An inverse relation between the variables.	BTL-1	Remember
3	What is cause and effect relationship? If two variables vary in such a way that movement in one are accompanied by movement in other, these variables are called cause and effect relationship.	BTL 2	Understand
4	Explain advantages of scatter diagram. a) It is a simple to implement and attractive method to find out the nature of correlation. b) It is easy to understand c) User will get rough idea about correlation (positive or negative correlation). d) Not influenced by the size of extreme item. e) First step in investing the relationship between two variables	BTL5	Evaluate
5	What is regression problem? For an input x, if the output is continuous, this is called a regression problem	BTL 2	Understand

6	What are assumptions of regression? The regression has five key assumptions: Linear relationship, Multivariate normality. No or little multi-collinearity and No auto-correlation.	BTL-1	Remember
7	What is regression analysis used for? Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable (s) (predictor). This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables.	BTL 2	Understand
8	What are the types of regressions? Types of regression are linear regression, logistic regression, polynomial regression, stepwise regression, ridge regression, lasso regression and elastic-net regression	BTL 4	Analyze
9	What do you mean by least square method? Least squares is a statistical method used to determine a line of best fit by minimizing the sum of squares created by a mathematical function. A "square" is determined by squaring the distance between a data point and the regression line or mean value of the data set.	BTL 5	Evaluate
10	What is correlation analysis? Correlation is a statistical analysis used to measure and describe the relationship between two variables. A correlation plot will display correlations between the values of variables in the dataset. If two variables are correlated, X and Y then a regression can be done in order to predict scores on Y from the scores on X.	BTL 6	Create
11	What is multiple regression equations? Multiple linear regression is an extension of linear regression, which allows a response variable, y to be modelled as a linear function of two or more predictor variables. In a multiple regression model, two or more independent variables, i.e. predictors are involved in the model. The simple linear regression model and the multiple regression models assume that the dependent variable is continuous.	BTL-1	Remember
12		BTL-1	Remember
13			

PART-B

Q. No	Questions	BT Level	Competence
1	Elaborate in detail the significance of correlation and the various types of correlation.	BTL 1	Remember

2	What are scatterplots? Elaborate on the various types with mutable examples	BTL 3	Apply
3	Highlight the significance of the correlation coefficient r. Compare the various correlation coefficients.	BTL5	Evaluate
4	What is the significance of r2 ? Give a detailed interpretation of r2 ?	BTL4	Analyze
5	Discuss the importance of regression. Elaborate on the types of Regression.	BTL6	Create
6	Explain the significance of regression line and Least squares regression equation.	BTL4	Analyze
7	Elaborate on multiple regression equations	BTL 1	Remember
8	Elucidate regression towards the mean. Explain regression fallacy and state how it can be avoided.		

PART-C

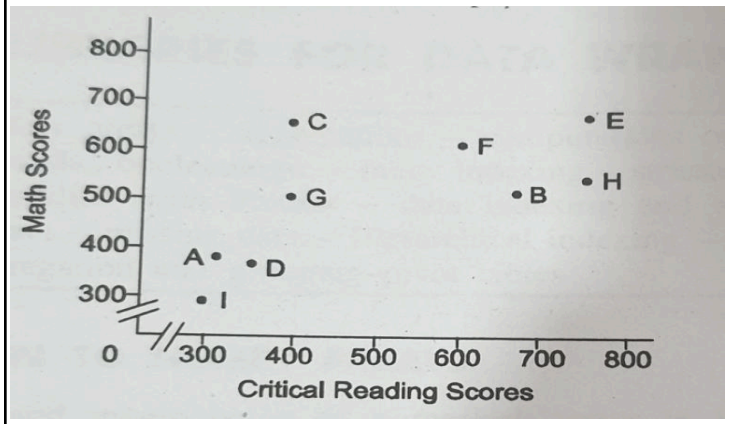
Q. No	Questions	BT Level	Competence
1	Calculate and analyse the correlation coefficient between the numbers of sady hours and the number of sleeping hours of different students.	BTL 1	Remember

Number of Study Hours	2	4	6	8	
Number of Sleeping Hours	10	9	8	7	

<table>
<tr>
<td>2</td>
<td>Find the standard error of the estimate of the mean weight of high school football players using the data given of weights of the players.

<table>
<tr><th>Player Number</th><th>Weight in Pounds</th></tr>
<tr><td>1</td><td>150</td></tr>
<tr><td>2</td><td>203</td></tr>
<tr><td>3</td><td>176</td></tr>
<tr><td>4</td><td>190</td></tr>
<tr><td>5</td><td>168</td></tr>
<tr><td>6</td><td>193</td></tr>
<tr><td>7</td><td>189</td></tr>
<tr><td>8</td><td>178</td></tr>
<tr><td>9</td><td>197</td></tr>
<tr><td>10</td><td>172</td></tr>
</table>
</td>
<td>BTL 3</td>
<td>Apply</td>
</tr>
<tr>
<td>3</td>
<td>Find the value of the correlation coefficient from the following table:

<table>
<tr><th>Subject</th><th>Age x`</th><th>Glucose Level</th></tr>
<tr><td>1</td><td>43</td><td>99</td></tr>
<tr><td>2</td><td>21</td><td>65</td></tr>
<tr><td>3</td><td>25</td><td>79</td></tr>
<tr><td>4</td><td>42</td><td>75</td></tr>
<tr><td>5</td><td>57</td><td>87</td></tr>
<tr><td>6</td><td>59</td><td>81</td></tr>
</table>
</td>
<td>BTL5</td>
<td>Evaluate</td>
</tr>
<tr>
<td>4</td>
<td>Critical reading and math scores on the SAT test for students A, BC. D. E. F. G. and H are shown in the following scatterplot</td>
<td>BTL4</td>
<td>Analyze</td>
</tr>
</table>

	 (a) Which student(s) scored about the same on both tests? (b) Which student(s) scored higher on the critical reading test than on the math test? (c) Which student(s) will be eligible for an honours program that requires minimum scores of 700 in critical reading and 500 in math? (d) Is there a negative relationship between the critical reading and math scores?		
5		BTL6	Create
6		BTL4	Analyze
7		BTL 1	Remember

UNIT IV PYTHON LIBRARIES FOR DATA WRANGLING
Basics of Numpy arrays –aggregations –computations on arrays –comparisons, masks, booleanlogic – fancy indexing – structured arrays – Data manipulation with Pandas – data indexing and selection – operating on data – missing data – Hierarchical indexing – combining datasets –aggregation and grouping – pivot tables
PART-A

Q. No	Questions	BT Level	Competence
1	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	BTL-1	Remember
2	What is Python? Python is a high-level scripting language which can be used for a wide variety of text processing, system administration and internet-related tasks. Python is a true object-oriented.Language and is available on a wide variety of platforms.	BTL-3	Apply
3	What is NumPy? NumPy, short for Numerical Python, is the core library for scientific computing in Python has been designed specifically for performing basic and advanced array operations. It primary supports multi-dimensional arrays and vectors for complex arithmetic operations	BTL1	Remember
4	What is an aggregation function ? An aggregation function is one which takes multiple individual values and returns a summary. In the majority of the cases, this summary is a single value. The most common aggregation functions are a simple average or summation of values.	BTL1	Remember
5	What is Structured Arrays? A structured Numpy array is an array of structures. As numpy arrays are homogeneous ie they can contain data of same type only. So, instead of creating a numpy array of int or float, we can create numpy array of homogeneous structures too.	BTL2	Understand
6	Describe Pandas. Pandas is a high-level data manipulation tool developed by Wes McKinney. It is built on the Numpy package and its key data structure is called the DataFrame. DataFrames allow you to store and manipulate tabular data in rows of observations and columns of variables. Pandas is built on top of the NumPy package, meaning a lot of the structure of NumPy is used or replicated in Pandas.	BTL2	Understand
7	How to Manipulating and Creating Categorical Variables? Categorical variable is one that has a specific value from a limited selection of values. The number of values is usually fixed. Categorical features can only take on a limited and usually fixed, number of possible values. For example, if a dataset is about information related to users, then user will typically find features like country, gender, age group, etc.	BTL1	Remember

	Alternatively, if the data we are working with is related to products, you will find features like product type, manufacturer, seller and so on.		
8	Explain Hierarchical Indexing. Hierarchical indexing is a method of creating structured group relationships in data. A MultiIndex or Hierarchical index comes in when our DataFrame has more than two dimensions. As we already know, a Series is a one-dimensional labelled NumPy array and a DataFrame is usually a two-dimensional table whose columns are Series. In some instances, in order to carry out some sophisticated data analysis and manipulation, our data is presented in higher dimensions.	BTL1	Remember
9	What is Pivot Tables? A pivot table is a similar operation that is commonly seen in spreadsheets and other programs that operate on tabular data. The pivot table takes simple column-wise data as input and groups the entries into a two-dimensional table that provides a multidimensional summarization of the data	BTL 3	Apply
10	What is NumPy? List its uses. NumPy is a general-purpose array-processing package with high-performance multidimensional array object, and tools. It is the fundamental package for scientific computing with Python. It provides N-dimensional array object supporting many sophisticated (broadcasting) functions. **Uses of NumPy:** NumPy is a package in Python used for Scientific Computing. NumPy packages used to perform different operations. The ndarray (NumPy Array) is a multidimensional array used to store values of same datatype. These arrays are indexed just like Sequences, starts with zero.	BTL4	Analyse
11	Where is NumPy used? NumPy is an open source numerical Python library. It provides a multi-dimentional array and matrix data structures. It can be utilised to perform mathematical operations on arrays such as trigonometric, statistical and algebraic routines.	BTL1	Remember
12	Write python code to create 1D, 2D and 3D NumPy arrays. ID array: import numpy as np al=np.array([1,2,3]) 2D array: a2=np.array([1.2.3)/2.2.211) 3D array: a3-np.array([[[1, 2], [3. 41],[[5, 6], [7, 81]]])		

13	Write short note on python array object. *The array module allows us to store a collection of numeric values. *To create an array of numeric values, we need to import the array module. *Indices are used to access elements of an array: *Slicing operator is used to access a range of items in an array		
14	How to perform slicing to access the element of a NumPy array? We can access a range of items in an array by using the slicing operator import array as arr numbers_list = [2, 5, 62, 5, 42, 52, 48, 5] numbers_array = arr.array('i, numbers_list) print(numbers_array[2:51) # 3rd to 5th print(numbers_array[:-5]) # beginning to 4th print(numbers_array[5:]) # 6th to end print(numbers_array[:]) # beginning to end		
15	Explain how to create a dictionary in python? Dictionaries are enclosed by curly braces ([]) and values can be assigned and accessed using square braces ([]). dict = {} dict['one'] = "This is one" dict[2] = "This is two" tiny dict = {'name': 'john',' code':6734, 'dept': 'sales'}		
16	What are universal Functions? * A universal function (or ufunc for short) is a function that operates on ndarrays in an element-by-element fashion. *It is a "vectorized" wrapper for a function that takes a fixed number of specific inputs and produces a fixed number of specific outputs. *These functions include standard trigonometric functions, functions for arithmetic operations, handling complex numbers, statistical functions, etc.		
17	What is fancyindexing? * With NumPy array fancy indexing, an array can be indexed with another NumPy array, a Python list, or a sequence of integers, whose values select elements in the indexed array. *Fancy indexing is like the simple indexing in which arrays are passed as indices in place of single scalars. *This allows us to very quickly access and modify complicated subsets of an array's values. *When using fancy indexing, the shape of the result replicates the shape of the index arrays not the shape of the array being indexed.		

PART-B

Q. No	Questions	BT Level	Competence
1	.Elaborate on indexing and slicing operations of Numpy Arrays.	BTL 2	Understand
2	Demonstrate on how vertical and horizontal splitting are done in ndarray	BTL 3	Apply
3	Discuss the array aggregation operations of Numpy arrays with example.	BTL5	Evaluate
4	What are ufuncs in Python? Explain with examples.	BTL 2	Understand
5	Explain comparison and masking operations.	BTL1	Remember
6	Assess the benefits of fancy indexing..	BTL 2	Understand
7	Explain about the pandas objects		
8	Discuss the approaches to combine datasets and identify the challenges		

PARTC-C

Q. No	Questions	BT Level	Competence
1	. Extract from the array np.array([3,4,6,10,24,89,45,43,46,99,100]) with Boolean masking all the number • which are not divisible by 3 • which are divisible by 5 • which are divisible by 3 and 5 • which are divisible by 3 and set them to 42	BTL 2	Understand
2	List the prime numbers between 0 and 100 by using a Boolean array.	BTL 3	Apply
3	Demonstrate different ways of creating pandasdataframe.	BTL5	Evaluate
4	How indexing is done in pandas? Explain.	BTL 2	Understand
5	Describe various methods of handling the missing data in pandas.	BTL1	Remember
6	Exhibit the benefits of multiple indexing.	BTL 2	Understand

UNIT V DATA VISUALIZATION			
Importing Matplotlib – Line plots – Scatter plots – visualizing errors – density and contour plots –Histograms – legends – colors – subplots – text and annotation – customization – three dimensionalplotting - Geographic Data with Basemap - Visualization with Seaborn.			
PART-A			
Q. No	**Questions**	**BT Level**	**Competence**
1	What is data visualization? Data visualization is the graphical representation of information and data	BTL-1	Remember
2	Which concept is used in data visualization? Data visualization based on two concepts: 1. Each attribute of training data is visualized in a separate part of screen. 2. Different class labels of training objects are represented by different colors.	BTL-2	Understand
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.	BTL5	Evaluate
4	Why big data visualization is important? - 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.	BTL-3	Apply
5	Explain Matplotlib. Matplotlib 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 value. It is a cross-section of the	BTL1	Remember

	three-dimensional graph of the function f(x, y) parallel to the(x,y) plane. Contour lines are used e.g. in geography and meteorology. In cartography, a contour e 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 Used in line graphs to explain the function or the values underlying the different lines of the graph.	BTL1	Remember
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.	BTL1	Remember
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.	BTL2	Understand
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, lines or points in transformed coordinates. Basemap includes the GSSH coastline dataset, as well as datasets from GMT for rivers, states and national boundaries.	BTL2	Understand
11	What is Seaborn? • Seaborn is a Python data visualization library based on Matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics. Seaborn is an open source 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	BTL1	Remember

12	List the application of line plot. A line plot is used to display a trend in data. It is used to express a between two variables. Ex: To see the performance of a company in the daily stock market for a you relation	BTL2	Understand
13	What is scatter plot? Scatter plots are used to observe relationship between variables. Scatter play a type of plot in which the points are represented individually with a dot, circle, other shape. The scatter () method in the matplotlib library is used to draw a sc plot. Scatter plots are used to visualize the relation among variables and how cha in one affects the other variable.		
14	What is histogram? A histogram is a graph showing frequency distributions. It shows the number of observations within each given interval. A simple histogram is useful in understanding a dataset. Matplotlib's histogram function creates a basic histogram in one line, once the normal boiler plate imports are done.		
15	Write code to plot sine wave using line plot. Line Plots are used to represent the relation between two data X and Y on a different axis. Use the ax.plot function to plot the data. In[ ]:Fig= plt.figure() ax = plt.axes() x=np.linspace(0, 10, 1000) ax.plot(x, np.sin(x));		
16	List the interfaces supported by matplotlib. Two Interfaces of Matplotlib: (a) MATLAB-style state-based interface: Python alternative for MATLAB users This interface is stateful: Keeps track of the "current" figure and axes. Stateful interface is fast and convenient for simple plots (b) Object-oriented interface The object-oriented interface is available for more complicated situations and provides more control over the figure. Object-oriented interface are more readable and explicit		

PART-B

Q. No	Questions	BT Level	Competence
1	Write Python program to plot Line chart by assuming your own data and explain the various attributes of line chart.	BTL5	Evaluate
2	Write python program to visualize the dataset using scatterplot and explain its parameters.	BTL 2	Understand
3	Elaborate the error visualization methods in pyplot.	BTL5	Evaluate
4	Explain different methods of showing three-dimensional surface on a two-dimensional plane with example.	BTL 2	Understand
5	Demonstrate the usage of histograms for data exploration and explain its attributes.	BTL4	Analyze
6	Elaborate the concept of subplots and its applications.		
7	Discuss in details about the three dimensional plotting functions of matplotlib module.		
8	Explain in details about the functions of mpl_tool kit for Geographic data visualization.		

PART-C

Q. No	Questions	BT Level	Competence
1	Write Python program to plot Line chart by assuming your own data and explain the various attributes of line chart.	BTL5	Evaluate
2	Write python program to visualize the dataset using scatterplot and explain its parameters.	BTL 2	Understand