CS3352 **FOUNDATIONS OF DATA SCIENCE** **L T P C**
3 0 0 3

COURSE OBJECTIVES:

- To understand the data science fundamentals and process.
- To learn to describe the data for the data science process.
- To learn to describe the relationship between data.
- To utilize the Python libraries for Data Wrangling.
- To present and interpret data using visualization libraries in Python

UNIT I **INTRODUCTION** **9**

Data Science: Benefits and uses – facets of data - Data Science Process: Overview – Defining research 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

UNIT II **DESCRIBING DATA** **9**

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

UNIT III **DESCRIBING RELATIONSHIPS** **9**

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

UNIT IV **PYTHON LIBRARIES FOR DATA WRANGLING** **9**

Basics of Numpy arrays –aggregations –computations on arrays –comparisons, masks, Boolean logic – 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

UNIT V **DATA VISUALIZATION** **9**

Importing Matplotlib – Line plots – Scatter plots – visualizing errors – density and contour plots – Histograms – legends – colors – subplots – text and annotation – customization – three dimensional plotting - Geographic Data with Basemap - Visualization with Seaborn

TOTAL: 45 PERIODS

TEXT BOOKS

1. David Cielen, Arno D. B. Meysman, and Mohamed Ali, "Introducing Data Science", Manning Publications, 2016. (Unit I)
2. Robert S. Witte and John S. Witte, "Statistics", Eleventh Edition, Wiley Publications, 2017. (Units II and III)
3. Jake VanderPlas, "Python Data Science Handbook", O'Reilly, 2016. (Units IV and V)

REFERENCES:

1. Allen B. Downey, "Think Stats: Exploratory Data Analysis in Python", Green Tea Press, 2014

COURSE OUTCOMES

Upon completion of the course, the student should be able to:

303.1	Define the data science process
303.2	Understand different types of data description for data science process
303.3	Gain knowledge on relationships between data
303.4	Use the Python Libraries for Data Wrangling
303.5	Apply visualization Libraries in Python to interpret and explore data

MAPPING BETWEEN CO AND PO, PSO WITH CORRELATION LEVEL 1/2/3

Course Outcomes	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2	PSO3	PSO4
303.1	**3**	**2**	**1**										**3**	**3**	**3**	
303.2	**3**	**2**	**1**										**3**	**3**	**3**	
303.3	**3**	**2**	**2**										**3**	**3**	**3**	
303.4	**3**	**3**	**3**										**3**	**3**	**3**	
303.5	**3**	**3**	**3**										**3**	**3**	**3**	

RELATION BETWEEN COURSE CONTENT WITH COS

UNIT I - INTRODUCTION

S. No.	Knowledge Level	Course Content	Course Outcomes
1.	R/U	Data Science: Benefits and uses	**303.1**
2.	R/U	Facets of data	
3.	R/U	Data Science Process: Overview, Defining research goals	
4.	R/U	Retrieving data	
5.	R/U/An	Data preparation	
6.	R/U/An	Exploratory Data analysis	
7.	R/U/An/Ap	Build the model, presenting findings and building applications	
8.	R/U/An	Data Mining	
9.	R/U/An	Data Warehousing	
10.	R/U/An	Basic Statistical descriptions of Data	

UNIT II - DESCRIBING DATA

S. No.	Knowledge Level	Course Content	Course Outcomes
1.	R/U	Types of Data	**303.2**
2.	R/U/An/Ap	Types of Variables	
3.	R/U/An/Ap	Describing Data with Tables and Graphs	
4.	R/U/An/Ap	Describing Data with Averages	

5.	R/U/An/Ap	Describing Variability	
6.	R/U/An/Ap	Normal Distributions and Standard (z) Scores	

UNIT III - DESCRIBING RELATIONSHIPS

S. No.	Knowledge Level	Course Content	Course Outcomes
1	R/U	Correlation	303.3
2	R/U/An	Scatter plots	
3	R/U/An	Correlation coefficient for quantitative data	
4	R/U/An	Computational formula for correlation coefficient	
5	R/U	Regression, regression line	
6	R/U/An	Least squares regression line	
7	R/U/An/Ap	Standard error of estimate	
8	R/U/An/Ap	Interpretation of r2	
9	R/U/An/Ap	Multiple regression equations	
10	R/U/An	Regression towards the mean	

UNIT IV - PYTHON LIBRARIES FOR DATA WRANGLING

S. No.	Knowledge Level	Course Content	Course Outcomes
1	R/U	Basics of Numpy arrays, aggregations	303.4
2	R/U/An	Computations on arrays	
3	R/U/An/Ap	Comparisons, masks, Boolean logic	
4	R/U/An/Ap	Fancy indexing	
5	R/U/An	Structured arrays	
6	R/U/An/Ap	Data manipulation with Pandas	
7	R/U/An	Data indexing and selection	
8	R/U/An/Ap	Operating on data , missing data	
9	R/U/An/Ap	Hierarchical indexing	
10	R/U/An/Ap	Combining datasets, aggregation and grouping	
11	R/U/An/Ap	Pivot tables	

UNIT V - DATA VISUALIZATION

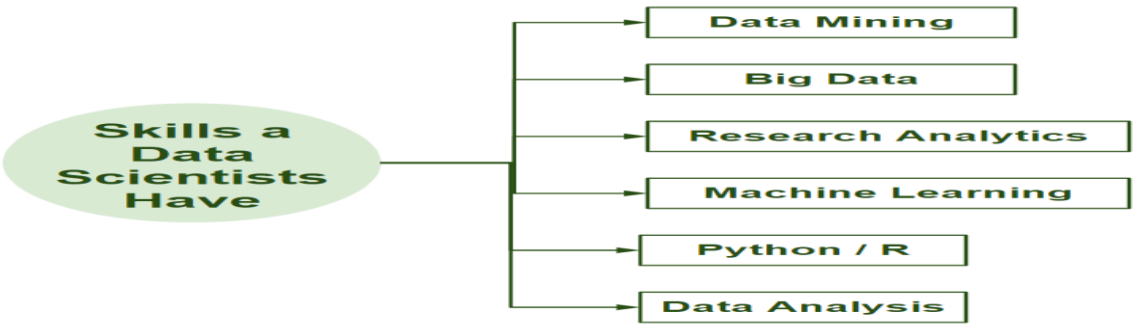
S. No.	Knowledge Level	Course Content	Course Outcomes
1	R/U/An/Ap	Importing Matplotlib	303.5
2	R/U/An/Ap	Line plots, Scatter plots	
3	R/U	Visualizing errors	
4	R/U/An	Density and contour plots	
5	R/U/An/Ap	Histograms	
6	R/U/An/Ap	Legends, colors , subplots	
7	R/U/An/Ap	Text and annotation	
8	R/U/An/Ap	Customization ,three dimensional plotting	
9	R/U/An/Ap	Geographic Data with Basemap	
10	R/U/An/Ap	Visualization with Seaborn	

UNIT I - INTRODUCTION

Data Science: Benefits and uses – facets of data - Data Science Process: Overview – Defining research 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

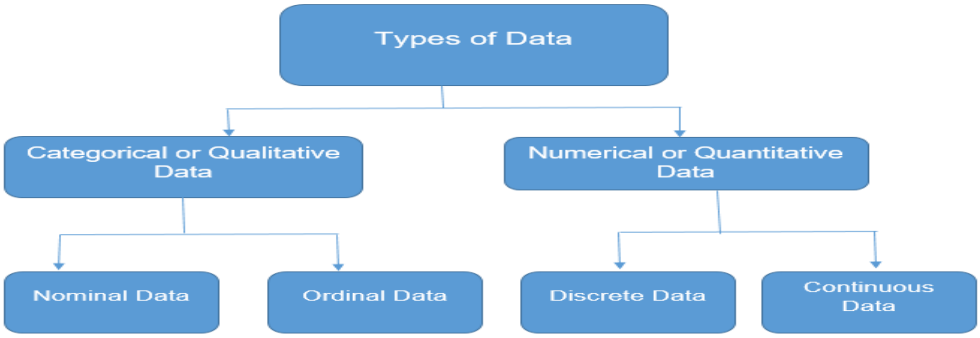
PART A

1	**What is Data Science?** • Data Science is the area of study which involves extracting insights from vast amounts of data using various scientific methods, algorithms, and processes. • It helps you to discover hidden patterns from the raw data. • Data Science is an interdisciplinary field that allows you to extract knowledge from structured or unstructured data. Data science enables you to translate a business problem into a research project and then translate it back into a practical solution.
2	**Why Data Science needed?** • It helps you to recommend the right product to the right customer to enhance your business • Allows to build intelligence ability in machines • It enables you to take better and faster decisions • Data Science can help you to detect fraud using advanced machine learning algorithms • It helps you to prevent any significant financial losses
3	**What are the components of data science?** • Domain expertise • Data engineering • Statistics • Visualization • Advanced computing
4	**List out the data science jobs.** Most prominent Data Scientist job titles are: • Data Scientist • Data Engineer • Data Analyst • Statistician • Data Architect • Data Admin • Business Analyst • Data/Analytics Manager
5	**List out the tools for Data Science.** **Data Analysis –** Python, R, Spark and SAS **Data Warehousing –** Hadoop, SQL **Data Visualization -** R, Tableau **Machine Learning –** Spark, Azure ML studio
6	**List out Some applications of Data Science.** • Internet Search Results (Google) • Recommendation Engine (Spotify) • Intelligent Digital Assistants (Google Assistant) • Autonomous Driving Vehicle (Waymo, Tesla) • Spam Filter (Gmail) • Abusive Content and Hate Speech Filter (Facebook) • Robotics (Boston Dynamics) • Automatic Piracy Detection (YouTube)

7	**What are the skills required to become the data scientist?** 
8	**What are the Challenges of Data Science Technology?** • A high variety of information & data is required for accurate analysis • Not adequate data science talent pool available • Management does not provide financial support for a data science team • Unavailability of/difficult access to data • Business decision-makers do not effectively use data Science results • Explaining data science to others is difficult • Privacy issues • Lack of significant domain expert • If an organization is very small, it can't have a Data Science team
9	**What is a Project Charter?** Clients like to know upfront what they are paying for, so after getting a good understanding of the business problem, try to get a formal agreement on the deliverables. All this information is collected in a project charter. The outcome should be a clear research goal, a good understanding of the context well-defined deliverables and a plan of action with a timetable. This information is then placed in a project charter.
10	**List the steps involved in the data cleansing** • Errors from data entry • Physically impossible values • Missing values • Outliers • Spaces and types • Errors against codebook
11	**What do you mean by Outliers?** An outlier is an observation that seems to be distant from other observations or, more specifically, one observation that follows a different logic or generative process than the other observations. The easiest way to find outliers is to use a plot or a table with the minimum and maximum values.
12	**What are the two operations used to combine information from different datasets?** • The first operation is joining: enriching an observation from one table with information from another table. • The second operation is appending or stacking: adding the observations of one table to those of another table.
13	**What do you mean by Exploratory data analysis?** ▸ Exploratory Data Analysis (EDA) is an approach to analyse the data using visual techniques. ▸ Information becomes much easier to grasp when shown in a picture, therefore we mainly use graphical techniques to gain an understanding of data and the interactions between variables. The visualization techniques used in this phase range from simple line graphs or histograms to more complex diagrams such as Sankey and network graphs.
14	**What is a Pareto diagram?** • A Pareto diagram is a combination of the values and a cumulative distribution. • A **Pareto chart** is a type of chart that contains both bars and a line graph, where individual values are represented in descending order by bars, and the cumulative total is represented by the line.
15	**What are the steps involved in building a model?** Building a model is an iterative process. 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
16	**What is data mining?** • Data mining is searching for knowledge (interesting patterns) in data. • Data mining is an essential step in the process of knowledge discovery. Data mining provides tools to discover knowledge from data and it turns a large collection of data into knowledge.
17	**What is a data warehouse?** • A data warehouse is a repository of information collected from multiple sources stored under a unified schema and usually residing at a single site. • Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing.
18	**What is a boxplot and what do we use it?** Boxplots are a popular way of visualizing a distribution. A boxplot incorporates the five-number summary as follows: • Typically, the ends of the box are at the quartiles so that the box length is the interquartile range. • The median is marked by a line within the box. • Two lines called whiskers outside the box extend to the smallest (Minimum) and largest (Maximum) observations.
19	**What do you mean by external data?** • Although data is considered an asset more valuable by certain companies, more and more governments and organizations share their data for free with the world. • This data can be of excellent quality and it depends on the institution that creates and manages it. • The information they share covers a broad range of topics in a certain region and its demographics.
20	**What is the need for basic statistical descriptions of data?** Basic statistical descriptions can be used to identify properties of the data. It highlights which data values should be treated as noise or outliers.
	PART B
1	Describe the Benefits and uses of data science?
2	Explain are the facets of data?
3	Describe the overview of the data science process
4	Explain the steps involved in the knowledge discovery process
5	Briefly describe the steps involved in Data Preparation.
6	What are the technologies used in data mining?
7	Explain in detail about Warehouse?
8	Explain the data exploration in detail.
9	What are the different sources of Warehouse?
10	Explain the Data Mining architecture.
11	Briefly discuss about the Internal and External Data?
12	Detail about Basic Statistical Data in Measuring the Central Tendency?
13	Explain in detail about build a model with example
	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
1	**What is qualitative data?** Qualitative data is defined as **the data that approximates and characterizes**. Qualitative data can be observed and recorded. This data type is non-numerical in nature. This type of data is collected through methods of observations, one-to-one interviews, conducting focus groups, and similar methods.

2 **What are the types of data?**



3 **What is quantitative data? Give some example.**

Quantitative data is data that can be counted or measured in numerical values. The two main types of quantitative data are discrete data and continuous data. **Height in feet, age in years, and weight in pounds** are examples of quantitative data.

4 **Differentiate quantitative and qualitative data.**

	Quantitative	Qualitative
Definition	Data that can be numerically analyzed and quantified into hard facts.	Non-numerical data that describes qualities, opinions, or feelings.
Collection Methods	Online, in-person, and phone interviews or surveys with closed-ended questions, controlled experiments, and more	Open-ended survey questions, unstructured interviews, focus groups, observation, and more
Best For	Drawing conclusions through larger-scale studies, conducting statistical analyses.	Formulating hypotheses and gathering detailed information from smaller groups
Analysis	Statistical analysis through charts, tables, and statistical programs.	Manual analysis through grouping of common themes and other methods.
Question Example	"Did you buy ice cream today? 1) Yes 2) No"	"Why did you buy ice cream today?"
Data Example	67% of respondents bought ice cream today.	"I saw ice cream on sale by the checkout and it was an impulse buy. I wanted to treat myself."

5 **What are the 4 types of variables?**

Variables in statistics are broadly divided into four categories such as **independent variables, dependent variables, categorical and continuous variables**.

6 **What is dependent variable in data science?**

- There are two types of data:
- Independent variables: Data that can be controlled directly.
- Dependent variables: **Data that cannot be controlled directly**

7 **What is an independent variable example?**

- The independent variable is the cause. Its value is **independent** of other variables in your study.
- The dependent variable is the effect. Its value **depends** on changes in the independent variable.

8 **What is the difference between a data table and a graph?**

Data are presented in a table to make it easier to compare and interpret them. The information in a table can be displayed using bars in a diagram called a bar graph or bar chart. The length of bars changes with the value of data.

9 **How do you write a data description in statistics?**

Step 1: Describe the size of your sample. Use N to know how many observations are in your sample. ...
Step 2: Describe the center of your data. ...
Step 3: Describe the spread of your data. ...
Step 4: Assess the shape and spread of your data distribution. ...
Step 5: Compare data from different groups.

10 **What are the 4 types of variation?**

	Types of variation include **direct, inverse, joint, and combined variation**.
11	**What are the four types of descriptive statistics?** Measures of Frequency: * Count, Percent, Frequency. ... Measures of Central Tendency. * Mean, Median, and Mode. ... Measures of Dispersion or Variation. * Range, Variance, Standard Deviation. ... Measures of Position. * Percentile Ranks, Quartile Ranks.
12	**What is an example of variability?** A simple measure of variability is **the range, the difference between the highest and lowest scores in a set**. For the example given above, the range of Drug A is 40 (100-60) and Drug B is 10 (85-75). This shows that Drug A scores are dispersed over a larger range than Drug B.
13	**Is z-score the same as standard normal distribution?** A standard normal distribution (SND). **A z-score, also known as a standard score**, indicates the number of standard deviations a raw score lays above or below the mean. When the mean of the z-score is calculated it is always 0, and the standard deviation (variance) is always in increments of 1.
14	**How do you find the z-score for a standard normal distribution?** **$z = (x – \mu) / \sigma$** Assuming a normal distribution, Your z score would be: $z = (x – \mu) / \sigma = (190 – 150) / 25 = 1.6$.
15	**What is the difference between normal distribution and standard distribution?** The standard normal distribution has a mean of 0 and a standard deviation of1, while a nonstandard normal distribution has a different value for one or both of those parameters.
16	**How do you interpret an average?** When the term 'average' is used in a mathematical sense, **it usually refers to the mean, especially when no other information is given**. Add the numbers together and divide by the number of numbers. (The sum of values divided by the number of values). Arrange the numbers in order, find the middle number
17	**What are nominal and ordinal variables?** There are two types of categorical variable, nominal and ordinal. **A nominal variable has no intrinsic ordering to its categories**. For example, gender is a categorical variable having two categories (male and female) with no intrinsic ordering to the categories. An ordinal variable has a clear ordering.
18	**What is z-score in statistics?** A Z-score is **a numerical measurement that describes a value's relationship to the mean of a group of values**. Z-score is measured in terms of standard deviations from the mean. If a Z-score is 0, it indicates that the data point's score is identical to the mean score.
	PART B
1	Explain the different types of data in statistics?
2	Explain the quantitative and qualitative data with examples?
3	What are variables in data science with examples?
4	How do you describe data with graphs? Give some examples
5	Explain the four types of descriptive statistics with examples?
6	How do you describe variability on a graph with examples?
7	How do you describe data with tables? give examples
8	How do you find the z-score with the mean and standard deviation of a normal distribution?
9	What is the relationship between normal distribution and standard normal distribution?
10	How do you describe charts with examples?
	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
1	**What is Correlation?** • A statistical measure that defines co-relationship or association of two variables • To describe a linear relationship between two variables

	• The objective is to find a value expressing the relationship between variable
2	**Define Scatter plots.** Scatter plots are the graphs that present the relationship between two variables in a data-set. It represents data points on a two-dimensional plane or on a Cartesian system. The independent variable or attribute is plotted on the X-axis, while the dependent variable is plotted on the Y-axis. These plots are often called scatter graphs or scatter diagrams.
3	**What is Correlation Coefficient?** The correlation coefficient (r) tells you the strength of the relationship between two variables. The value of r has a range of -1 to 1 (0 indicates no relationship). Values of r closer to -1 or 1 indicate a stronger relationship and values closer to 0 indicate a weaker relationship
4	**List out the types of Correlation Coefficient**

Correlation coefficient value	Correlation type	Meaning
1	Perfect positive correlation	When one variable changes, the other variables change in the same direction.
0	Zero correlation	There is no relationship between the variables.
-1	Perfect negative correlation	When one variable changes, the other variables change in the opposite direction.

5	**Define regression.** Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).
6	**List out some Real-world examples of linear regression models** • Forecasting sales: Organizations often use linear regression models to forecast future sales. ... • Cash forecasting: Many businesses use linear regression to forecast how much cash they'll have on hand in the future.
7	**What is the use of regression line?** • A regression line indicates a linear relationship between the dependent variables on the y-axis and the independent variables on the x-axis • The regression line is plotted closest to the data points in a regression graph. This statistical tool helps analyse the behaviour of a dependent variable y when there is a change in the independent variable x—by substituting different values of x in the regression equation.
8	**What is computational formula for correlation coefficient?** • There are several types of correlation coefficient formulas. • One of the most commonly used formulas is Pearson's correlation coefficient formula. $r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\,n\Sigma x^2 - (\Sigma x)^2\,][\,n\Sigma y^2 - (\Sigma y)^2\,]}}$
9	**How to find a regression line?** The formula of the regression line for Y on X is as follows: $\mathbf{Y = a + bX + \varepsilon}$ Here Y is the dependent variable, a is the Y-intercept, b is the slope of the regression line, X is the independent variable, and ε is the residual (error).
10	**What is the slope of a regression line?** The slope of a regression line is denoted by 'b,' which shows the variation in the dependent variable y brought out by changes in the independent variable x. The formula to determine the slope of the regression line for Y on X is as follows: $\mathbf{b = n(\sum XY\text{-})(\sum X)(\sum Y) / (n\sum X^2 – (\sum X)^2)}$
11	**What is the Least Squares Method?** • The least squares method is a form of mathematical regression analysis used to determine the line of best fit for a set of data, providing a visual demonstration of the relationship between the data points. • Each point of data represents the relationship between a known independent variable and an unknown dependent variable.
12	**What is least squares regression line?**

	• The Least Squares Regression Line is the line that makes the vertical distance from the data points to the regression line as small as possible. It's called a "least squares" because the best line of fit is one that minimizes the variance
13	**Define standard error of the estimate?** The standard error of the estimate is a way to measure the accuracy of the predictions made by a regression model. It is denoted as SEE
14	**What is R-Squared?** A statistical measure that determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, r-squared shows how well the data fit the regression model (the goodness of fit).
15	**What is Interpretation of R-Squared** The most common interpretation of r-squared is how well the regression model explains observed data. For example, an r-squared of 60% reveals that 60% of the variability observed in the target variable is explained by the regression model.
16	**What is Multiple regression?** Multiple regression is a statistical technique that can be used to analyse the relationship between a single dependent variable and several independent variables.
17	**What is the difference between MLR and SLR?** SLR examines the relationship between the dependent variable and a single independent variable. MLR examines the relationship between the dependent variable and multiple independent variables.
18	**What is regression toward the mean ?** In statistics, **regression toward the mean** is a concept that refers to the fact that if one sample of a random variable is extreme, the next sampling of the same random variable is likely to be closer to its mean.
	PART B
1	Explain about Correlation in detail
2	Describe about Scatter Plots with example
3	Explain the process of finding the correlation coefficient for quantitative data
4	Explain about Regression in detail with example
5	Differentiate Correlation and Regression
6	Discuss the computational formula for correlation coefficient
7	Describe about least squares regression line
8	Write short notes on Standard error of
9	How to interpret the value of r2 in detail
10	Discuss about multiple regression equations
11	How the regression used towards the mean?
	UNIT IV -PYTHON LIBRARIES FOR DATA WRANGLING Basics of Numpy arrays –aggregations –computations on arrays –comparisons, masks, Boolean logic – 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
1	**What are the two types of magic commands?** The two types of magic commands are **Line magics :** They are similar to command line calls. They start with % character. Rest of the line is its argument passed without parentheses or quotes. Line magic's can be used as expression and their return value can be assigned to variable. **Cell magics :** They have %% character prefix. Unlike line magic functions, they can operate on multiple lines below their call. They can make arbitrary modifications to the input they receive, which need not even be a valid Python code at all. They receive the whole block as a single string.

2	**What are the categories of basic array manipulation?** • Attributes of arrays • Determining the size, shape, memory consumption, and data types of arrays. • Indexing of arrays • Getting and setting the value of individual array elements. • Slicing of arrays • Getting and setting smaller sub arrays within a larger array • Reshaping of arrays • Changing the shape of a given array • Joining and splitting of arrays • Combining multiple arrays into one, and splitting one array into many
3	**What is the syntax for Numpy slicing?** The Numpy slicing syntax follows that of the standard Python list, to access a slice of an array x: **x[start:stop:step]** If any of these are unspecified, they default to the values start=0, stop=size of dimension, step=1. We can access sub-arrays in one dimension and in multiple dimensions.
4	**What will be the output for the below code:** x2 = array([[12, 5, 2, 4], [7, 6, 8, 8], [1, 6, 7, 7]]) print(x2[0, :]) **Output:** [12 5 2 4]
5	**What do you mean by ufuncs?** Ufuncs are the universal functions. The Vectorized operations in Numpy are implemented via ufuncs whose main purpose is to quickly execute repeated operations on values in Numpy arrays. NumPy's universal functions can be used to vectorize operations and thereby remove slow Python loops.
6	**What is the purpose of the axis keyword?** • The axis keyword specifies the dimension of the array that will be collapsed, rather than the dimension that will be returned. • So specifying axis=0 means that the first axis will be collapsed. For two-dimensional arrays, this means that values within each column will be aggregated.
7	**What are the rules for 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 are 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.
8	**What is fancy indexing?** • A style of array indexing is known as fancy indexing. • Fancy indexing is like the simple indexing but we pass arrays of indices in place of single scalars. This allows us to very quickly access and modify complicated subsets of an array's values.
9	**What is the difference between np.sort and np.argsort?** • np.sort is used to return a sorted version of the array without modifying the input. • np.argsort is used to return the indices of the sorted elements.
10	**What is the output of the given code?** data = np.zeros(4, dtype={'names':('name', 'age', 'weight'),'formats':('U10', 'i4', 'f8')}) print(data.dtype) **Output:** [('name', '<U10'), ('age', '<i4'), ('weight', '<f8')]
11	**What is the difference between Numpy array and pandas series?** • While the Numpy Array has an implicitly defined integer index used to access the values, the Pandas Series has an explicitly defined index associated with the values. • This explicit index definition gives the Series object additional capabilities. For example, the index need not be an integer but can consist of values of any desired type. For example we can use strings as an index.

12	**How the series object can be modified?** Series objects can be modified with a dictionary-like syntax. Just as we can extend a dictionary by assigning to a new key, we can extend a Series by assigning to a new index value.
13	**What is python none object?** The first sentinel value used by Pandas is None, a Python singleton object that is often used for missing data in Python code. Because it is a Python object, None cannot be used in any arbitrary Numpy/Pandas array, but only in arrays with data type 'object' i.e. arrays of Python objects.
14	**What is the use of multi-indexing?** • Multi-indexing is used to represent two-dimensional data within a one-dimensional Series. • We can also use it to represent data of three or more dimensions in a Series or Data Frame. Each extra level in a multi-index represents an extra dimension of data.
15	**What is pd.merge ( ) function?** The pd.merge ( ) function implements a number of types of joins: the one-to-one, many-to-one and many-to-many joins. All three types of joins are accessed via an identical call to the pd.merge () interface. The type of join performed depends on the form of the input data.
16	**What is describe ( ) method?** The method describe ( ) computes several common aggregates for each column and returns the result. We can use this method on the dataset for dropping rows with missing values.
17	**What is split, apply and combine?** • The split step involves breaking up and grouping a data frame depending on the value of the specified key. • The apply step involves computing some function usually an aggregate, transformation, or filtering within the individual groups. • The combine step merges the results of these operations into an output array.
18	**What is the use of get ( ) and slice ( ) operations?** • The get () and slice () operations enable vectorized element access from each array. • For example, we can get a slice of the first three characters of each array using str.slice (0, 3). • get () and slice() methods also let us access elements of arrays returned by split(). • For example, to extract the last name of each entry, we can combine split () and get().
19	**What do you mean by datetime and dateutil?** The datetime type is used to manually build a date. Using the dateutil module, we can parse dates from a variety of string formats. With datetime object, we can print the day of the week.
20	**What is the advantage of using numexpr library?** ● The Numexpr library gives the ability to compute compound expressions element by element without the need to allocate full intermediate arrays. ● Numexpr evaluates the expression in a way that does not use full-sized temporary arrays and can be much more efficient than Numpy, especially for large arrays. ● The Pandas eval() and query() tools are conceptually similar and depend on the Numexpr package
PART B	
1	Explain all the array manipulation functions with examples in Numpy.
2	Write short notes on Computation on Arrays.
3	Explain Aggregation Functions and Fancy Indexing with examples in Numpy.
4	Explain selection sort and other sorting methods used in Numpy with Examples
5	What are the Data Manipulation Techniques in Pandas.
6	Explain in detail the steps involved in constructing a pandas data frame
7	What are the steps involved in handling missing data in pandas.
8	Explain in detail about the aggregate, filter, transform and apply operations of the GroupBy object
9	Write short notes on dates and times in pandas with examples.
10	Explain in detail about the Pivot table?
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 dimensional plotting - Geographic Data	

with Basemap - Visualization with Seaborn	
1	**What is Matplotlib?** • Matplotlib is a python library used to create 2D graphs and plots by using python scripts. • It has a module named pyplot which makes things easy for plotting by providing feature to control line styles, font properties, formatting axes etc. • It supports a very wide variety of graphs and plots namely - histogram, bar charts, power spectra, error charts etc.
2	**What is the line plot?** • A Line plot can be defined as a graph that displays data as points or check marks above a number line, showing the frequency of each value.
3	**Define Scatter plots.** • Scatter plots are the graphs that present the relationship between two variables in a data-set. • It represents data points on a two-dimensional plane or on a Cartesian system. • The independent variable or attribute is plotted on the X-axis, while the dependent variable is plotted on the Y-axis. • These plots are often called scatter graphs or scatter diagrams.
4	**Define Error bars.** • Error bars function used as graphical enhancement that visualizes the variability of the plotted data on a Cartesian graph. • Error bars can be applied to graphs to provide an additional layer of detail on the presented data. As you can see in below graphs.
5	**How do you visualize error bars?** • Error bars are used to display either the standard deviation, standard error, confidence intervals or the minimum and maximum values in a ranged dataset. • To visualise this information, Error Bars work by drawing cap-tipped lines that extend from the centre of the plotted data point
6	**What is density plot?** • Density Plot is a type of data visualization tool. • It is a variation of the histogram that uses 'kernel smoothing' while plotting the values. It is a continuous and smooth version of a histogram inferred from a data.
7	**What are Contour plots?** • Contour plots (sometimes called Level Plots) are a way to show a three-dimensional surface on a two-dimensional plane. • It graphs two predictor variables X Y on the y-axis and a response variable Z as contours. These contours are sometimes called the z-slices or the iso-response values.
8	**Define histogram** • A histogram is the graphical representation of data where data is grouped into continuous number ranges and each range corresponds to a vertical bar. • The horizontal axis displays the number range. • The vertical axis (frequency) represents the amount of data that is present in each range.
9	**What are legends in data visualization?** • A legend is used to identify data in visualizations by its color, size, or other distinguishing features. • Legends identify the meaning of various elements in a data visualization and can be used as an alternative to labeling data directly
10	**Why is color important in data visualization?** • Color is important in data visualization because it allows you to highlight certain pieces of information and promote information recall. • Using different colors can separate and define different data points within visualization so that viewers can easily distinguish significant differences or similarities in values.
11	**What is the use of subplots () function?** • The subplots ( ) function in pyplot module of matplotlib library is used to create a figure and a set of subplots.

12	**What are Visualization Annotations?** • Annotations are text boxes that can be added on top of visualization. You can use them for various presentation purposes like explaining what is displayed in the visualization or informing of interesting findings. They can also be used for adding instructions or asking questions.
13	**Define figure and axes in matplotlib.** • Axes object is the region of the image with the data space. • Axes object is added to figure by calling the add_axes () method.
14	**What is matplotlib basemap?** • Basemap is a great tool for creating maps using python in a simple way. • It's a matplotlib extension, so it has got all its features to create data visualizations, and adds the geographical projections and some datasets to be able to plot coast lines, countries, and so on directly from the library.
15	**How do you create a contour plot?** • A contour plot can be created with the plt.contour function. • It takes three arguments: a grid of x values, a grid of y values, and a grid of z values. The x and y values represent positions on the plot, and the z values will be represented by the contour levels.
16	**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.
17	**What is the difference between Matplotlib and Seaborn?** • Seaborn is more comfortable in handling Pandas data frames. • It uses basic sets of methods to provide beautiful graphics in python. • Matplotlib works efficiently with data frames and arrays. It treats figures and axes as objects.
18	**How do you create an axes?** • The most basic method of creating an axes is to use the plt.axes function. • By default this creates a standard axes object that fills the entire figure. plt.axes also takes an optional argument that is a list of four numbers in the figure coordinate system. • These numbers represent [left, bottom, width, height] in the figure coordinate system, which ranges from 0 at the bottom left of the figure to 1 at the top right of the figure.

PART B

1	Explain the simple line plots and simple scatter plots
2	Explain density and contour plots with an example
3	Write short notes on histograms
4	Explain in detail about legends.
5	Write short notes on customization in matplotlib.
6	Explain the three dimensional plotting in matplotlib.
7	Write short notes on visualization with seaborn
8	Discuss about Geographic Data with Basemap in detail.
9	Write short notes on text and annotation
10	Explain about subplots with example