Unit – I Introduction
Two Marks Questions with Answers

1. What is data science?

Data Science is also known as data driven science. Data science is an interdisciplinary field that seeks to extract knowledge or insights from various forms of data.

2. Define structured data.

Structured data is arranged in rows and column format. It helps for application to retrieve and process data easily. Database management system is used for storing structured data. The term structured data refers to data that is identifiable because it is organized in a structure.

3. What is data?

Data is a Collection of raw fact. That is data is plain fact, unprocessed data. Data set is collection of related records or information. The information may be on some entity or some subject area.

4. What is unstructured data ?

Unstructured data is data that does not follow a specified format. Row and columns are not used for unstructured data. Therefore it is difficult to retrieve required information. Unstructured data has no identifiable structure.

5. What is machine - generated data ?

Machine-generated data is an 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.

6. Define 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).

7. List the stages of data science process.

Stages of data science process are as follows:

1. Discovery or Setting the research goal

2. Retrieving data
3. Data preparation
4. Data exploration
5. Data modeling
6. Presentation and automation

8. What are the advantages of data repositories?

Advantages are as follows:

i. Data is preserved and archived.
ii. Data isolation allows for easier and faster data reporting.
iii. Database administrators have easier time tracking problems.
iv. There is value to storing and analyzing data.

9. What is data cleaning?

Data cleaning means removing the inconsistent data or noise and collecting necessary information of a collection of interrelated data.

10. What is outlier detection?

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.

11. What is 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.

12. Define data mining.

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.

13. What are the three challenges to data mining regarding data mining methodology?

Challenges to data mining regarding data mining methodology include the following:

1. Mining different kinds of knowledge in databases,
2. Interactive mining of knowledge at multiple levels of abstraction,
3. Incorporation of background knowledge.

14. What is predictive mining?

Predictive mining tasks perform inference on the current data in order to make predictions. Predictive analysis provides answers of the future queries that move across using historical dataas the chief principle for decisions.

15. What is data cleaning?

Data cleaning means removing the inconsistent data or noise and collecting necessary information of a collection of interrelated data.

16. List the five primitives for specifying a data mining task.

1. The set of task-relevant data to be mined
2. The kind of knowledge to be mined
3. The background knowledge to be used in the discovery process
4. The interestingness measures and thresholds for pattern evaluation
5. The expected representation for visualizing the discovered pattern.

17. List the stages of data science 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

18. What is data repository?

Data repository is also known as a data library or data archive. This is a general term to refer to a data set isolated to be mined for data reporting and analysis. The data repository is a large database infrastructure, several databases that collect, manage and store data sets for data analysis, sharing and reporting.

19. List the data cleaning tasks.

Data cleaning 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

20. What is Euclidean distance ?

Euclidean distance is used to measure the similarity between observations. It is calculated as the square root of the sum of differences between each point.

Unit – II Describing Data
Two Marks Questions with Answers

1. Define qualitative data.
Qualitative data provides information about the quality of an object or information which cannot be measured. Qualitative data cannot be expressed as a number. Examples : gender, economic status and religious preference.

2. What is quantitative data?
Quantitative data is the one that focuses on numbers and mathematical calculations and can be calculated and computed. Quantitative data are anything that can be expressed as a number or quantified. Example : scores on achievement tests, weight of a student.

3. What is nominal data?
A nominal data is the 1st level of measurement scale in which the numbers serve as "tags" or "labels" to classify or identify the objects. Nominal data is type of qualitative data.

4. Define 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.

5. 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.

6. What do you mean observational study?
An observational study focuses on detecting relationships between variables not manipulated by the investigator. An observational study is used to answer a research question based purely on what the researcher observes. There is no interference or manipulation of the research subjects and no control and treatment groups.

7. What is frequency distribution?

Frequency distribution is a representation, either in a graphical or tabular format, that displays the number of observations within a given interval. The interval size depends on the data being analyzed and the goals of the analyst.

8. 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.

9. Define histogram.

A histogram is a special kind of bar graph that applies to quantitative data (discrete or continuous). The horizontal axis represents the range of data values. The bar height represents the frequency of data values falling within the interval formed by the width of the bar. The bars are also pushed together with no spaces between them.

10. What is goal of variability?

The goal for variability is to obtain a measure of how spread out the scores are in a distribution. A measure of variability usually accompanies a measure of central tendency as basic descriptive statistics for a set of scores.

11. How to calculate range?

The range is the total distance covered by the distribution, from the highest score to the lowest score (using the upper and lower real limits of the range).

Range = Maximum value - Minimum value

12. What is an Independent variables?

An independent variable is the variable that is changed or controlled in a scientific experiment to test the effects on the dependent variable.

13. What is an observational study?

An observational study focuses on detecting relationships between variables not manipulated by the investigator. An observational study is used to answer a research question based purely on what the researcher observes. There is no interference or manipulation of the research subjects and no control and treatment groups.

14. Define frequency polygon.

Frequency polygons are a graphical device for understanding the shapes of distributions. They serve the same purpose as histograms, but are especially helpful for comparing sets of data. Frequency polygons are also a good choice for displaying cumulative frequency distributions.

15. 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).

Unit – III Describing Relationships
Two Marks Questions with Answers

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.

2. Define positive correlation .

Positive correlation : Association between variables such that high scores on onevariable tends to have high scores on the other variable. A direct relation between the variables.

3. Define negative correlation.

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.

4. 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.

5. Write the advantages of scatter diagram.

1. It is a simple to implement and attractive method to find out the nature of correlation.
2. It is easy to understand.
3. User will get rough idea about correlation (positive or negative correlation).
4. Not influenced by the size of extreme item.
5. First step in investing the relationship between two variables.

5. What is regression problem?

For an input x, if the output is continuous, this is called a regression problem.

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.

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.

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.

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.

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.

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.

Unit – IV Python Libraries for Data Wrangling
Two Marks Questions with Answers

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.

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.

3. What is NumPy ?

NumPy means Numerical Python. It is used for scientific computing in Python. It hasbeen designed specifically for performing basic and advanced array operations. It primarily supports multi-dimensional arrays and vectors for complex arithmeticoperations.

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.

5. What is Structured Arrays?

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, instead of creating a numpy array of int or float, we can create numpy array of homogeneous structures too.

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.

7. What is 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.

8. Define 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.

9. What is Pivot Tables?

A pivot table is as same as 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.

Unit – V Data Visualization
Two Marks Questions with Answers

1. What is data visualization?

Data visualization is the graphical representation of information and data.

2. What 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.

3. List the benefits of data visualization.

- Visualize relationships and patterns in businesses.
- More collaboration and sharing of information.
- More self-service functions for the end users.

4. Why big data visualization is important?

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

5. What is 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.

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 three-dimensional graph of the function f(x, y) parallel to the x, y plane. Contour lines are used in geography and meteorology.

7. What is legends?

Legends are found in maps describe the pictorial language or symbology of the map. Legends are used in line graphs to explain the function or the values underlying the different lines of the graph.

8. What is subplots?

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

9. What is use of tick?

A tick is a short line on an axis. Ticks are the markers denoting data points on axes.
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.

10. Define Basemap.

Basemap is a toolkit under the Python visualization library Matplotlib. Its mainfunction 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.

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 opensource Python library.