DATA SCIENCE
(CSEN 5141)

Time Allotted : 2½ hrs **Full Marks : 60**

Figures out of the right margin indicate full marks.

Candidates are required to answer Group A and any 4 (four) from Group B to E, taking one from each group.

Candidates are required to give answer in their own words as far as practicable.

Group – A

1. Answer any twelve: **12 × 1 = 12**

Choose the correct alternative for the following

(i) Point out the correct statement
(a) Raw data is original source of data
(b) Preprocessed data is original source of data
(c) Raw data is the data obtained after processing steps
(d) None of the mentioned.

(ii) In statistics, a population consists of
(a) All people living in a country
(b) All people living in the area under study
(c) All subjects or objects whose characteristics are being studied
(d) None of the above.

(iii) Which of the following is performed by Data Scientist?
(a) Define the question (b) Create reproducible code
(c) Challenge results (d) All of the mentioned.

(iv) You asked five of your classmates about their height. On the basis of this information, you stated that the average height of all students in your university or college is 67 inches. This is an example of
(a) Descriptive statistics (b) Inferential statistics
(c) Parameter (d) Population.

(v) Which of the following allows you to find the relationship you didn't know about?
(a) Inferential (b) Exploratory
(c) Causal (d) None of the mentioned.

(vi) Which of the following is the common goal of statistical modelling?
(a) Inference (b) Summarizing
(c) Subsetting (d) None of the above.

(vii) Suppose that for a data set: there are m points and k clusters; half the points and clusters are in "more dense" regions; half the points and clusters are in "less dense" regions; the two regions are well-separated from each other. For the given data set, which of the following should occur in order to minimize the squared error when finding k clusters:
(a) Centroids should be equally distributed between more dense and less dense regions
(b) More centroids should be allocated to the less dense region
(c) More centroids should be allocated to the denser region
(d) None of the above.

(viii) One marble jar has several different coloured marbles inside it. It has 1 red, 2 green, 4 blue, and 8 yellow marbles. All the marbles are of the same size and shape. If Asha takes out one marble from the jar without looking, what is the probability that she will not choose a yellow marble?
(a) 7/15 (b) 8/15 (c) 7/8 (d) 5/8.

(ix) Which of the following is true about statistical methods?
(a) They are most useful for examining complex problems in the real world
(b) They are most useful for examining simple problems in the virtual world
(c) They are most useful for examining simple problems in the real world
(d) None of the above.

(x) Data visualization is also an element of the broader _____.
(a) deliver presentation architecture (b) data presentation architecture
(c) dataset presentation architecture (d) data process architecture

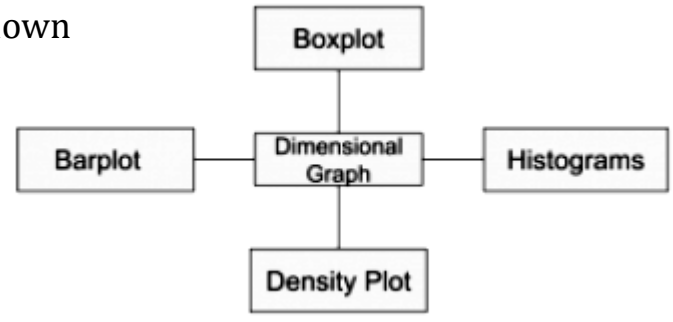
Fill in the blanks with the correct word

(xi) _____ is true about the regression analysis.

(xii) Removing false values from a data source and inconsistencies across data sources is know as _____.

(xiii) The type of regression used to model phenomena with binary outcomes is known as _____.

(xiv) _____ type graph is shown in the figure.



(xv) A _____ is a probability distribution of a statistic that is arrived at through repeated sampling from a specific population.

Group - B

2. (a) What are the four characteristics of big data? *[(CO1)(Remember/LOCQ)]*
(b) Mention and explain the main categories in the facets of data science. *[(CO2)(Understand/LOCQ)]*

(c) What is structured data? Explain with examples. *[(CO2)(Understand/LOCQ)]*

4 + 4 + 4 = 12

3. (a) What are the differences between supervised and unsupervised learning? *[(CO1)(Explain/LOCQ)]*

(b) You are given a data set consisting of variables with more than 30 percent missing values. How will you deal with them? *[(CO2)(Understand/LOCQ)]*

(c) What are dimensionality reduction and its benefits? *[(CO1)(Explain /LOCQ)]*

(d) How should you maintain a deployed model? *[(CO2)(Understand/LOCQ)]*

3 + 3 + 3 + 3 = 12

Group - C

4. Consider a Multiple-Choice Exam that contains 10 multiple-choice questions with 4 possible choices for each question, only one of which is correct.
Suppose a student is to select the answer for every question randomly. Let X be the number of questions the student answers correctly. Then, X has a binomial distribution with parameters n = 10 and p = 0.25. (Convince yourself that all assumptions for a binomial distribution are reasonable in this setting.) *[(CO4)(Demonstrate/IOCQ)]*

(i) What is the probability for the student to get no answer correct?

(ii) What is the probability for the student to get two answers correct?

(iii) What is the probability for the student to fail the test (i.e., to have less than 6 correct answers)?

(4 + 4 + 4) = 12

5. (a) What is classification? Explain with an example. *[(CO5)(Understand/LOCQ)]*

(b) Mention and explain each of the four possible types of results in binary classification. *[(CO5)(Understand/LOCQ)]*

(c) What is the simplest measure of classification accuracy? *[(CO5)(Understand/LOCQ)]*

(d) What is the difference between sensitivity and specificity? *[(CO5)(Understand/LOCQ)]*

4 + 4 + 2 + 2 = 12

Group - D

6. (a) One way to sparsify a proximity matrix is the following: For each object (row in the matrix), set all entries to 0 except for those corresponding to the objects k-nearest neighbors. However, the sparsified proximity matrix is typically not symmetric. *[(CO5)(Develop/HOCQ)]*

(i) If object ***a*** is among the ***k-nearest* neighbors** of object ***b***, why is ***b*** not guaranteed to be among the ***k-nearest* neighbors** of ***a***?

(ii) Suggest at least two approaches that could be used to make the sparsified proximity matrix symmetric.

(b) A study looks to model the relationship between the number of cookies left for Santa Claus and the number of presents received. Data collected over the last 5 Christmas is given in the table. *[(CO6)(Apply/HOCQ)]*

Year	**2015**	**2016**	**2017**	**2018**	**2019**
Cookies Left	2	4	2	6	8
Presents Received	1	1	4	5	5

(i) Using the method of Least Squares, find the linear equation, describing the number of presents received as a function of the number of cookies left, that best fits the given data.

(ii) Based on your linear model, if you want 7 presents this year, how many cookies should you leave Santa Claus?

6 + 6 = 12

7. (a) Suppose the fraction of undergraduate students who smoke is 15% and the fraction of graduate students who smoke is 23%. If one-fifth of the college students are graduate students and the rest are undergraduates, what is the probability that a student who smokes is a graduate student? *[(CO4)(Demonstrate/IOCQ)]*

(b) Given the information in part (a), is a randomly chosen college student more likely to be a graduate or undergraduate student? *[(CO4)(Demonstrate/IOCQ)]*

(c) Repeat part (b) assuming that the student is a smoker. *[(CO4)(Demonstrate/IOCQ)]*

(d) Suppose 30% of the graduate students live in a dorm but only 10% of the undergraduate students live in a dorm. If a student smokes and lives in the dorm, is he or she more likely to be a graduate or undergraduate student? You can assume independence between students who live in a dorm and those who smoke. *[(CO4)(Demonstrate/IOCQ)]*

3 + 3 + 3 + 3 = 12

Group - E

8. (a) What is k-means clustering? *[(CO5)(Remember/LOCQ)]*

(b) How are k-means clusters evaluated? *[(CO5)(Analyse/IOCQ)]*

(c) Identify and briefly explain two other clustering techniques different from k-means clustering. *[(CO5)(Understand/LOCQ)]*

4 + 4 + (2 + 2) = 12

9. (a) You are given the runs scored in each one-day international innings that Sachin Tendulkar batted in, and Virat Kohli batted in. Using this information, and all your knowledge of the data science process and techniques, design a study that will help you establish whether Kohli is a better batsman than Tendulkar. Write down the steps of the study. *[(CO6)(Create/HOCQ)]*

(b) What are the benefits of data visualization? *[(CO3)(Understand/LOCQ)]*

(c) Identify and briefly explain the three different types of data analytics. *[(CO3)(Understand/LOCQ)]*

6 + 3 + 3 = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	52.08	29.17	18.75

Course Outcome (CO):

After the completion of the course students will be able to

CSEN5141.1 Explain how data is collected, managed and stored for data science

CSEN5141.2 Understand the key concepts in data science, including their real-world applications and some of the popular techniques used by data scientists

CSEN5141.3 Build skills in data management

CSEN5141.4 Demonstrate proficiency with statistical analysis of data.

CSEN5141.5 Develop ability to build and assess data-based models.

CSEN5141.6 Apply data science concepts and methods to solve real-world problems.

**LOCQ: Lower Order Cognitive Question; IOCQ: Intermediate Order Cognitive Question; HOCQ: Higher Order Cognitive Question.*