SET-4

Series #CDBA

Q.P. Code 106

Roll No. | | | | | | | | |

Candidates must write the Q.P. Code on the title page of the answer-book.

- Please check that this question paper contains **11** printed pages.
- Please check that this question paper contains **21** questions.
- Q.P. Code given on the right hand side of the question paper should be written on the title page of the answer-book by the candidate.
- **Please write down the serial number of the question in the answer-book before attempting it.**
- 15 minute time has been allotted to read this question paper. The question paper will be distributed at 10.15 a.m. From 10.15 a.m. to 10.30 a.m., the students will read the question paper only and will not write any answer on the answer-book during this period.

DATA SCIENCE

Time allowed : 2 hours *Maximum Marks : 50*

General Instructions :

(i) *Please read the instructions carefully.*

(ii) *This question paper consists of **21** questions in **two** sections : **Section A** and **Section B**.*

(iii) *Section A has Objective Type Questions, whereas Section B contains Subjective Type Questions.*

(iv) ***Out of the given (5 + 16 =) 21 questions, the candidate has to answer (5 + 10 =) 15 questions in the allotted (maximum) time of 2 hours.***

(v) *All questions of a particular section must be attempted in the correct order.*

(vi) ***Section A*** *: Objective Type Questions (24 marks) :*

*(a) This section has **5** questions.*

(b) There is no negative marking.

(c) Do as per the instructions given.

(d) Marks allotted are mentioned against each question/part.

(vii) ***Section B*** *: Subjective Type Questions (26 marks) :*

*(a) This section has **16** questions.*

*(b) A candidate has to do **10** questions.*

(c) Do as per the instructions given.

(d) Marks allotted are mentioned against each question/part.

SECTION A

(Objective Type Questions) **(24 marks)**

1. Answer any ***4*** out of the given ***6*** questions on Employability Skills. *4×1=4*

(i) What is a common misconception about starting a business ?

(A) You need a large amount of money to get started

(B) Starting a business requires no financial investment

(C) You can start a business with whatever money you have

(D) Every business idea has to be unique

(ii) Which of the following is ***not*** true with respect to self-awareness abilities ?

(A) Being self-aware means that you can identify your strengths and weaknesses.

(B) Self-awareness helps to convert your strengths to weaknesses.

(C) Self-awareness helps to convert your strengths to exceptional talent.

(D) Analysing your strengths and weaknesses helps you to attain success in life.

(iii) Managing _____ is about making a plan to be able to cope effectively with daily pressures and to strike a balance between life, work, relationships, relaxation, and fun.

(A) Weight

(B) Stress

(C) Beliefs

(D) Interests

(iv) Which of the following is a type of verbal communication ?

(A) Gestures

(B) Body language

(C) Facial expressions

(D) Interpersonal communication

(v) What is the keyboard shortcut for copying selected text or files in most operating systems ?

(A) Ctrl + X (B) Ctrl + Z

(C) Ctrl + V (D) Ctrl + C

(vi) How many Sustainable Development Goals (SDGs) were adopted by the United Nations in 2015 ?

(A) 5 (B) 10

(C) 15 (D) 17

2. Answer any ***5*** out of the given ***6*** questions. *5×1=5*

(i) In data analysis, why should you create subsets of your data ?

(A) To increase the size of your dataset

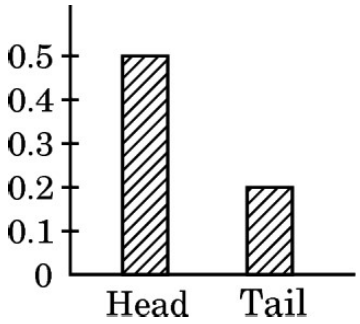
(B) To make the data easier to analyse

(C) To add extra information in your dataset

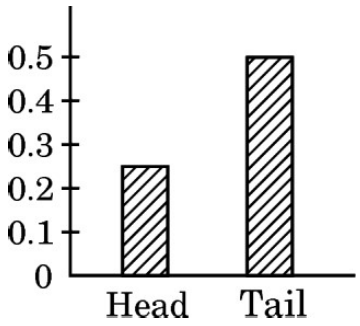
(D) To add noise to the data

(ii) Which of the following graphs shows the probability distribution for tossing a coin ?

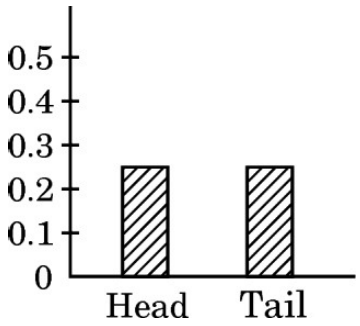
(A)



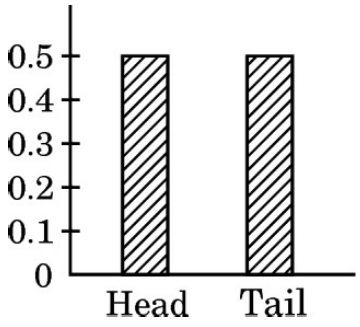
(B)



(C)



(D)

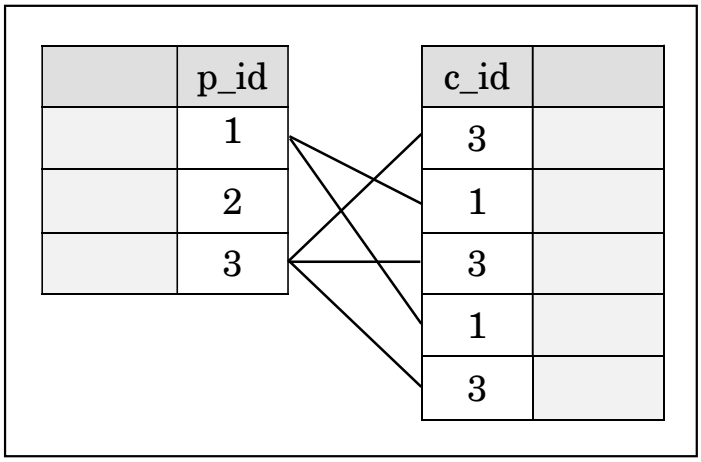


(iii) In Data Science, _____ is a deviation from the expected outcome in the data.

(A) Outlier (B) Standard Deviation

(C) Bias (D) Variance

(iv) Identify the type of join from the following diagram.



(A) Many to Many
(B) One to Many
(C) One to One
(D) Many to One

(v) State True or False.

If you soft delete a particular file, it gets deleted permanently and cannot be restored.

(vi) Statement I : Shredding and cutting are two ways to discard physical data.

Statement II : The shredded documents can be read easily but the documents that are cut down into small pieces cannot be read.

(A) Both statement I and Statement II are true
(B) Both statement I and Statement II are false
(C) Statement I is true, but Statement II is false
(D) Statement II is true, but Statement I is false

3. Answer any ***5*** out of the given ***6*** questions. *5×1=5*

(i) A _____ table represents the percentage of data points that fit in each category.

(A) Dummy
(B) Two-way frequency
(C) Two-way relative frequency
(D) Two-way direct frequency

(ii) The value of z-score is always _____ if the value of z-score lies above the mean.

(A) Positive (B) Negative

(C) Zero (0) (D) Infinite

(iii) What does a distribution represent in Data Science ?

(A) The exact values for a variable and how often they occur

(B) The probable values for a variable and how often they occur

(C) The average values for a variable and how often they occur

(D) The maximum values for a variable and how often they occur

(iv) In which bias do we assume that change in one quantity produces an equal and proportional change in another ?

(A) Linearity Bias (B) Confirmation Bias

(C) Recall Bias (D) Selection Bias

(v) Which of the following is an appropriate way of discarding physical copies of confidential data ?

(A) Crumbing the papers which contain confidential data and throwing them in the dustbin

(B) Burning the documents

(C) Leaving the confidential data in an unlocked drawer

(D) Using the papers containing confidential data for rough work.

(vi) Quartiles of a dataset partition the data in _____ equal parts.

(A) Three (B) Four

(C) Six (D) Eight

4. Answer any ***5*** out of the given ***6*** questions. *5×1=5*

(i) What security measure should you employ to ensure that even in the case of a data leak, the hackers are not able to read your data ?

(A) Backing up data to an external drive

(B) Using any password

(C) Encrypting the data

(D) Deleting the data

(ii) In a dataset of income levels for a small town's residents, the income data includes a few extremely high values (outliers) due to the presence of some wealthy individuals. In this scenario, which form of central tendency would be more appropriate to describe the typical income of the town's residents ?

(A) Mean (B) Median

(C) Mode (D) Range

(iii) What is the correct order of steps in the Statistical Problem Solving Process ?

(A) Collect/Consider the data, Formulate statistical investigative questions, Analyse the data, Interpret the results

(B) Analyse the data, Interpret the results, Collect/Consider the data, Formulate statistical investigative questions

(C) Formulate statistical investigative questions, Collect/Consider the data, Interpret the results, Analyse the data

(D) Formulate statistical investigative questions, Collect/Consider the data, Analyse the data, Interpret the results

(iv) If n is the total number of data in a sample, which of the following correctly represents the first decile ?

(A) $D_1 = 1 * (n+1) / 10^{th}$ Data

(B) $D_1 = 1 + (n+1) / 10^{th}$ Data

(C) $D_1 = 1 - (n+1) / 10^{th}$ Data

(D) $D_1 = 1 / (n+1) / 10^{th}$ Data

(v) If you have a dataset with 100 data points, what data point defines the first quartile (Q1) ?

(A) The 25^{th} percentile

(B) The 50^{th} percentile

(C) The 75^{th} percentile

(D) The 100^{th} percentile

(vi) Which type of bias is commonly associated with recommendation systems, polls, and personalised advertisements, where the sample data is not representative of the true future population of cases that the model will see ?

(A) Linearity Bias (B) Recall Bias

(C) Selection Bias (D) Survivor Bias

5. Answer any ***5*** out of the given ***6*** questions. *5×1=5*

(i) Statement I : In Data Science, we can perform data merging by implementing data joins on the databases in frame.

Statement II : There are five categories of data joins.

(A) Both statement I and statement II are correct

(B) Both statement I and statement II are incorrect

(C) Statement I is correct, but statement II is incorrect

(D) Statement I is incorrect, but statement II is correct

(ii) You have a dataset of ages for a group of 5 people. The ages are as follows: 25, 32, 29, 35 and 27. What is the median age for this dataset ?

(A) 27 (B) 28

(C) 29 (D) 30

(iii) You have a bag with two red marbles and two green marbles. If you randomly select one marble from the bag without looking, what is the probability that it will be a red marble ?

(A) 0·50 (B) 0·25

(C) 1·0 (D) 0·75

(iv) In statistical analysis, what minimum sample size is generally considered enough for the Central Limit Theorem to hold ?

(A) 5 (B) 10

(C) 20 (D) 30

(v) Deciles sort the data into how many equal parts ?

(A) 5 (B) 10

(C) 9 (D) 8

(vi) *Assertion (A) :* Privacy does not always mean confidentiality of data.

Reason (R) : Private data may need to be audited based on the relevant requirements.

(A) Both (A) and (R) are correct and (R) is the correct explanation of (A).

(B) Both (A) and (R) are correct, but (R) is ***not*** the correct explanation of (A).

(C) (A) is correct, but (R) is incorrect.

(D) (R) is correct, but (A) is incorrect.

SECTION B

(Subjective Type Questions) **(26 marks)**

*Answer any **3** out of the given **5** questions on Employability Skills. Answer each question in 20 – 30 words.* *3×2=6*

6. Give any two ways to overcome barriers of effective communication.

7. Differentiate between wage employed people and self-employed people.

8. Explain any two key strategies or initiatives that can contribute towards achieving Sustainable Development Goals (SDGs).

9. Define any ***two*** of the following terms :

(a) ICT

(b) Trojan horse

(c) Firewall

10. What does the acronym SMART stand for in the context of goal setting ? Why is goal setting an essential factor in your personal life ?

*Answer any **4** out of the given **6** questions in 20 – 30 words each.* *4×2=8*

11. Discuss any two ways of protecting confidential data that is stored in digital form.

12. Write steps to calculate Mean Absolute Deviation (MAD).

13. Explain the concept of Recall Bias in Statistics.

14. For each of the following scenarios, classify the data as either discrete or continuous :

(a) Result of an admission test

(b) Weight of apples at a grocery store

(c) Football score in a match

(d) Depth of a river

15. Explain one-to-one join with the help of an example.

16. In a dataset representing the ages of a group of people, the lower quartile (Q1) is 30 years, and the upper quartile (Q3) is 50 years. Calculate the interquartile range (IQR) for this dataset.

*Answer any **3** out of the given **5** questions in 50 – 80 words each.* *3×4=12*

17. What is a two-way frequency table ? Explain its features with a suitable example.

18. Explain the following components of the Statistical Problem Solving Process:

(a) Analyse the data

(b) Interpret the results

19. Explain Central Limit Theorem. Give any two real world scenarios in which it is used.

20. What is z-score in data science ? Write the formula to calculate z-score. Also explain the importance of z-score in data science.

21. You are a data analyst working for a healthcare organisation. Your team is responsible for analysing patient data to improve healthcare services and outcomes. Ethical considerations are crucial in handling this sensitive information. Recently, you received a dataset containing patient records, including medical history, personal information, and treatment details.

(i) What is the most important ethical consideration when handling patient's data ?

(A) Data accuracy and quality

(B) Data sharing with external organisations without patients' knowledge

(C) Data storage cost optimisation

(D) Patient's privacy protection

(ii) A colleague suggests using patients' data to train an artificial intelligence model without the patients' consent. What is your ethical stance on this suggestion ?

(A) Support the idea as it can lead to technological advancements

(B) Implement the idea without informing the patients

(C) Reject the idea and ensure patients' consent is obtained for any data usage

(D) Share the colleague's idea with other organisations for wider adoption

(iii) List any two ethical guidelines around data analysis.