MASTER OF SCIENCE
DATA SCIENCE AND ANALYTICS
(MSCDSA)

MSCDSA/ASSIGN/SEMESTER-II

ASSIGNMENTS

(July – 2026 & January – 2027)

MCS-066, MCS-067, MCS-224, MCS-068, MCSL-069, MCSL-070

SCHOOL OF COMPUTER AND INFORMATION SCIENCES
INDIRA GANDHI NATIONAL OPEN UNIVERSITY
MAIDAN GARHI, NEW DELHI – 110 068

CONTENTS

Important Notes

1. Submit your assignments to the Coordinator of your Study Centre on or before the due date.
2. Assignment submission before due dates is compulsory to become eligible for appearing in corresponding Term End Examinations. For further details, please refer to MSCDSA Programme Guide.
3. To become eligible for appearing the Term End Practical Examination for the lab courses, it is essential to fulfill the minimum attendance requirements as well as submission of assignments (on or before the due date). For further details, please refer to the MSCDSA Programme Guide.
4. The viva voce is compulsory for the assignments. For any course, if a student submitted the assignment and not attended the viva-voce, then the assignment is treated as not successfully completed and would be marked as ZERO.

Course Code	**:**	**MCS-066**
Course Title	**:**	**Mathematical Foundations for Data Science-II**
Assignment Number	**:**	**MSCDSA (II)/066/Assign/2026-27**
Maximum Marks	**:**	**100**
Weightage	**:**	**30%**
Last Date of Submission	**:**	**31st October, 2026 (for July session)** **30th April, 2027 (for January session)**

There are four questions in this assignment, which carry 80 marks. Each question carries 20 marks. Rest 20 marks are for viva voce. You may use illustrations and diagrams to enhance the explanations, if necessary.

Q1: (Covers Block 1)

(a) A retail company wants to study customer purchasing behaviour in its stores. What would be the population for this study be? How will you sample the data and draw inferences? Also, differentiate between descriptive and inferential statistics with suitable examples. **(2 Marks)**

(b) A researcher collects information on the following variables from 200 university students: **(2 Marks)**

- Gender
- Academic programme
- Annual family income
- Student satisfaction level (Very Low to Very High)
- CGPA

Identify the appropriate measurement scale (nominal, ordinal, interval or ratio) for each variable. Also, explain the importance of selecting the correct measurement scale during statistical analysis.

(c) The following marks (out of 100) were obtained by 30 students: **(5 Marks)**

42, 55, 67, 72, 48, 63, 59, 74, 81, 65, 58, 69, 77, 84, 71, 62, 57, 66, 73, 60, 75, 79, 68, 64, 53, 70, 61, 56, 76, 82

Construct:

- a frequency distribution,
- relative frequency table,
- histogram,
- ogive, and
- stem-and-leaf plot.

(d) Using the data given in Question 3, calculate: **(5 Marks)**

- Mean
- Median
- First Quartile (Q1)
- Third Quartile (Q3)

Draw a box plot and comment on the presence of any outliers.

(e) The monthly sales (in lakh rupees) of a company for twelve months are: **(4 Marks)**

35, 42, 38, 41, 45, 39, 47, 50, 43, 44, 46, 40

Calculate:

- Range
- Variance
- Standard deviation
- Coefficient of variation

Comment on the sales pattern.

(f) Explain the concepts of skewness and kurtosis. A data scientist observes two datasets with the same mean and variance but different skewness and kurtosis. Explain how these measures help in understanding the nature of the distributions. Also, explain the characteristics of a normally distributed dataset. **(2 Marks)**

Q2: (Covers Block 2)

(a) If the letters of the word "Probability" are arranged randomly, what is the probability that all the letter 'b' are together in the random arrangement. **(2 Marks)**

(b) A factory produces bulbs using two machines: **(3 Marks)**

- Machine A manufactures **60%** of the bulbs, and **2%** of the bulbs it produces are defective.
- Machine B manufactures **40%** of the bulbs, and **5%** of the bulbs produced by it are defective.

A randomly selected bulb is found to be defective. Using Bayes' theorem, determine the probability that the defective bulb was manufactured by Machine B.

(c) A random variable (X) denotes the number of defective items found in a shipment and has the following probability distribution: **(2 Marks)**

X	**0**	**1**	**2**	**3**
P(X)	0.25	0.40	0.25	0.10

Calculate:

- Expected value of finding the
- Variance

(d) Explain the concepts of joint probability mass function and marginal probability function for two discrete random variables with the help of an example. **(2 Marks)**

(e) A manufacturing process produces 2% defective products. A random sample of 25 products is selected. Perform the following tasks for this process: **(2 Marks)**

(i) Calculate the probability of obtaining:

- exactly one defective product,
- at most two defective products.

(ii) State the assumptions under which the binomial distribution is applicable.

(f) Explain the Bernoulli, Binomial, Poisson and Hypergeometric distributions. Compare their assumptions, parameters and applications. **(3 Marks)**

(g) The lifetime (in hours) of a sensor follows a normal distribution with a mean of 800 hours and a standard deviation of 60 hours. Compute the following for this sensor: **(3 Marks)**

- Probability that a sensor lasts more than 850 hours.
- Probability that its lifetime lies between 760 and 900 hours.

(h) Explain the Uniform, Normal, Chi-square, t and F distributions. Highlight the relationships among these distributions. **(3 Marks)**

Q3: (Covers Block 3)

(a) Explain the Central Limit Theorem and discuss its significance in statistical inference. Illustrate how the theorem helps in estimating population parameters when the population distribution is unknown. **(3 Marks)**

(b) A population has a mean of 80 and a standard deviation of 18. Random samples of size 64 are repeatedly selected and reported. Compute the following for this: **(3 Marks)**

- Mean of the sampling distribution of the sample mean.
- Standard error.
- Probability that the sample mean exceeds 84.

(c) A random sample of 64 observations has a sample mean of 48 and a sample standard deviation of 10. Find the following: **(4 Marks)**

- A 95% confidence interval for the population mean.
- Distinguish between point estimation and interval estimation.

(d) Explain the principles of Maximum Likelihood Estimation (MLE). Also, explain the characteristics of an estimator. **(3 Marks)**

(e) A company claims that the average battery life of its rechargeable battery is 18 hours. A random sample of 40 batteries gives a mean life of 17.2 hours with a standard deviation of 2.5 hours. Using a 5% significance level, test the manufacturer's claim. **(4 Marks)**

Clearly state the following while answering:

- Null and alternative hypotheses,
- Test statistic,
- Conclusion.

(f) Explain the concepts of significance level, Type-I error, Type-II error and p-value. Also, compare the z-test and t-test, and discuss the situations in which each test should be applied. **(3 Marks)**

Q4: (Covers Block 4)

(a) Three teaching methods are tested on three different groups of students. The examination scores are given below: **(4 Marks)**

- Method A: 72, 75, 70, 73, 74
- Method B: 78, 81, 79, 82, 80
- Method C: 69, 68, 71, 70, 67

Perform a one-way ANOVA at the 5% significance level and interpret the results.

(b) Explain the assumptions underlying one-way and two-way ANOVA. Also, explain the two-way ANOVA using a suitable example. **(3 Marks)**

(c) The following data show customer preference for payment method across two age groups: **(3 Marks)**

Age Group	UPI	Card	Cash
Below 40	60	45	25
40 and Above	35	40	55

Using the chi-square test of independence, determine whether payment preference is independent of age group at the 5% significance level.

(d) Explain the objectives and procedures of: **(4 Marks)**

- Chi-square goodness-of-fit test,
- Test of independence,
- Kolmogorov–Smirnov goodness-of-fit test.

(e) Explain the concepts of local minima, global minima, gradient descent and gradient ascent. Compare gradient descent and stochastic gradient descent. Discuss their advantages and limitations in machine learning applications. **(3 Marks)**

(f) What are pseudo-random numbers? Explain the methods of generating pseudo-random numbers. **(3 Marks)**