

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

Course Code	Assignment No.	Submission-Schedule		Page No.
		For July-December Session	For January-June Session	
MCS-066	MSCDSA(II)/066/Assign/26-27	31 st October, 2026	30 th April, 2027	3
MCS-067	MSCDSA(II)/067/Assign/26-27	31 st October, 2026	30 th April, 2027	7
MCS-224	MSCDSA(II)/224/Assign/26-27	31 st October, 2026	30 th April, 2027	10
MCS-068	MSCDSA(II)/068/Assign/26-27	31 st October, 2026	30 th April, 2027	13
MCSL-069	MSCDSA(II)/L-069/Assign/26-27	31 st October, 2026	30 th April, 2027	16
MCSL-070	MSCDSA(II)/L-070/Assign/26-27	31 st October, 2026	30 th April, 2027	17

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-067
Course Title	:	Data Wrangling and Visualization
Assignment Number	:	MSCDSA(II)/067/Assign/2026-27
Maximum Marks	:	100
Weightage	:	30%
Last Dates for 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. Algorithms, if asked, should be written in pseudocode that is closer to Python. You may use illustrations and diagrams to enhance the explanations, if necessary

Q1:

- (a) Explain data wrangling. Discuss the importance of data semantics, set-based profiling, summary statistics, and data profiling in preparing raw data for data analysis. Illustrate your answer with a suitable example. **(3 Marks)**
- (b) A retail company has collected customer transaction data containing the following attributes: **(3 Marks)**
- Customer ID
 - Age
 - Gender
 - City
 - Purchase Amount
 - Payment Mode

Describe how you would perform data profiling on this dataset. What are the summary statistics that you would perform on this data?

- (c) Make a CSV file that contains employee information. The file should have a few missing salary values, a few duplicate employee records, a few inconsistent department names, and some incorrect city names. **(3 Marks)**

Write a Python program with suitable comments using **pandas** that performs the following tasks after reading the file.

- identify missing values,
 - remove duplicate records,
 - replace inconsistent values,
 - fill missing salary values using an appropriate method, and
 - rename the column headings.
- (d) Explain different techniques used for handling missing data. Discuss situations where each technique is appropriate. **(3 Marks)**

- (e) A dataset contains the ages of 500 customers ranging from 18 to 80 years. Explain different ways to classify these customers into age groups using discretisation (binning). **(4 Marks)**
- (f) Write Python code to detect outliers using the interquartile range method. Also, explain the strategies for handling outliers in data with the help of suitable examples. **(4 Marks)**

Q2:

- (a) Two data files contain customer details and purchase details having the same format. These files may contain overlapped data. Write and explain the Python program to merge the two files. Make and state assumptions, if any **(3 Marks)**
- (b) Explain hierarchical indexing. Illustrate how hierarchical indexing can be used for reshaping datasets with suitable examples. **(3 Marks)**
- (c) A supermarket dataset contains the following fields: **(3 Marks)**
- Region
 - Product Category
 - Sales
 - Profit

Write a Python program to perform the following tasks:

- group data by Region,
 - calculate total sales,
 - calculate average profit.
- (d) Why is the need to create groups during data analysis? Explain with the help of an example. Explain the working of the **groupby()** function in pandas. Discuss grouping using dictionaries, functions and Series in Python with suitable examples. **(3 Marks)**
- (e) Make a synthetic dataset containing monthly sales of three different products for one year. Write a Python program using Matplotlib and/or Seaborn to generate a line plot and a scatter plot. Make suitable assumptions. **(4 Marks)**
- (f) Write a Python code to demonstrate the purpose of figures, axes, labels, legends, annotations and subplot arrangements with suitable examples. **(4 Marks)**

Q3:

- (a) What is a scatter plot matrix? Why is it needed? Write a Python program to make a scatter plot matrix. **(3 Marks)**
- (b) A weather department has recorded daily rainfall and temperature data for five years. Suggest suitable graphical techniques for visualising this dataset. Justify your choices with reference to time series, scatterplot matrices, and geographical maps. **(6 Marks)**

- (c) Using Python, create a dashboard consisting of multiple plots arranged in a single figure. Demonstrate how titles, legends, annotations and customised plot parameters improve the readability of the visualisation. **(3 Marks)**
- (d) Explain any four applications of graphs in data science with suitable examples. **(4 Marks)**
- (e) A social networking website wants to analyse the relationships among users. Describe how graph representations such as bipartite graphs, hierarchical trees and graph drawing techniques can be used for analysing the network. Illustrate your answer with suitable diagrams. **(4 Marks)**

Q4:

- (a) Explain different methods of visualising random forests. **(5 Marks)**
- (b) Explain the use of the bivariate histogram and kernel density estimation with the help of a suitable example. **(6 Marks)**
- (c) Explain the Trellis Paradigm. Why is it needed? **(3 Marks)**
- (d) Explain various regression visualisations with the help of an example. **(6 Marks)**