

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 : **MCSL-070**
Course Title : **Data Analysis Lab**
Assignment Number : **MSCDSA(II)/L-070/Lab_Assign/2026-27**
Maximum Marks : **100**
Weightage : **30%**
Last Date of Submission : **31st October, 2026 (for July session)**
30th April, 2027 (for January session)

Note: This assignment consists of two sections, each carrying 20 marks, making a total of 40 marks. Your Lab Records will carry 40 Marks. The remaining 20 marks are for viva voce. Prepare and present your assignment in accordance with the guidelines provided in the MSCDSA Programme Guide. For programming-based questions, submit the source code, documentation, output, and screenshots of the program execution. You may also use diagrams, tables, illustrations, and screenshots, wherever appropriate, to improve the clarity and presentation of your answers.

SECTION 1: Data Wrangling and Visualisation Lab

1. Write Python programs using NumPy, Pandas, Matplotlib and/or Seaborn wherever required.
2. You may use Google Colab or Jupyter Notebook or any Python IDE.
3. Make and state assumptions, if any.

Q1:

(3 Marks)

A company that does retail marketing has collected customer transaction data, as shown in the following table

CustomerID	Name	City	Age	Amount	Payment Mode
C101	Amit	Delhi	25	1250	UPI
C102	Neha	Mumbai	32	2300	Card
C103	Ravi	Delhi	29	NaN	Cash
C104	Priya	Bengaluru	NaN	1850	Card
C105	Amit	Delhi	25	1250	UPI
C106	Karan	Delhi	41	4200	Cash
C107	Sonia	Kolkata	35	2750	UPI
C108	Meena	Delhi	27	1500	Card
C108	Meena	Delhi	27	1500	Card
C109	Rohan	Mumbai	31	3150	UPI

Perform the following tasks for the data as above:

- a) Create a DataFrame using the above data.
- b) Display summary statistics of the dataset.
- c) Find missing values and show records with missing data.

- d) Fill the missing numerical data using the mean value of that column.
- e) Remove all the duplicate records.
- f) Finally, display the cleaned dataset.

Q2:

(3 Marks)

The following synthetic sales data is for supermarket sales.

OrderID	Region	Category	Sales	Profit
O101	South	Grocery	5200	650
O102	South	Grocery	6100	720
O103	North	Clothing	18000	3100
O104	North	Grocery	4500	480
O105	North	Clothing	24000	5200
O106	South	Clothing	8200	1100
O107	South	Grocery	3900	420
O108	North	Clothing	7600	980
O109	North	Clothing	9100	1250
O110	South	Grocery	21000	4700

Perform the following tasks for the data as above:

- a) Group the data by Region and calculate the total Sales.
- b) What is the average Profit for each Category?
- c) Apply multiple aggregation functions (sum, mean, maximum) on the Sales column.
- d) Sort the aggregated data in descending order of Sales.
- e) Display the final aggregated table.

Q3:

(4 Marks)

Consider the following two datasets:

Dataset A: Student Information

StudentID	Name	Department
S101	Asha	Data Science
S102	Vivek	AI
S103	Mohit	Data Science
S104	Sneha	Cyber Security
S105	Riya	AI

Dataset B: Semester Marks

StudentID	Marks in Python	Marks in Statistics
S101	82	78
S102	74	80
S103	91	88
S104	69	72
S105	85	81

Perform the following tasks for the synthetic data as above:

- Merge the two datasets on StudentID.
- Create a pivot table showing department-wise average marks.
- Set Department and StudentID as a hierarchical index.
- Rearrange the index levels.
- Display the final reshaped DataFrame.

Q4:

(5 Marks)

The monthly sales (in thousands Rs) of three products are given below:

Month	Product A	Product B	Product C
Jan	45	52	39
Feb	48	55	41
Mar	52	58	45
Apr	49	56	47
May	60	63	51
Jun	64	66	54
Jul	67	70	58
Aug	65	72	60
Sep	70	75	63
Oct	74	78	67
Nov	78	82	71
Dec	81	86	74

Make the following plots by using Matplotlib and Seaborn libraries:

- Line plot
- Grouped bar chart
- Histogram for Product A sales
- Scatter plot between Product A and Product B
- Add suitable titles, legends, axis labels and annotations.

Q5:

(5 Marks)

The following synthetic dataset shows the data of advertisement expenditure and monthly sales:

Advertisement cost in Thousand Rs	Sales in Rs (lakhs)
5	21
8	24
10	27
12	30
14	33
16	35
18	38
20	41
22	43
24	46
26	48
28	51

Perform the following tasks for the synthetic data as above:

- Create a scatter plot of Advertisement versus Sales.
- Fit a simple linear regression model and draw the regression line.
- Plot the residuals.
- Plot the distribution of the Sales variable using the Kernel Density Estimate (KDE)

SECTION 2:

Q1:

(5 Marks)

A car showroom wants to analyse the details of different cars using R programming. The company wants to store car-related information, identify the type of each variable, convert variables into suitable data types, filter records, and summarise the data for decision-making.

Create or import the following dataset in R as car_showroom.csv:

Car_ID	Brand	Model	Mileage	Engine_Type	Transmission	Price_Lakh
C101	Maruti	Swift	22.5	Petrol	Manual	7.2
C102	Hyundai	i20	20.1	Petrol	Automatic	9.5
C103	Tata	Nexon	17.8	Diesel	Manual	11.3
C104	Honda	City	18.4	Petrol	Automatic	14.8

Car_ID	Brand	Model	Mileage	Engine_Type	Transmission	Price_Lakh
C105	Mahindra	XUV300	16.9	Diesel	Manual	12.6
C106	Toyota	Glanza	21.2	Petrol	Manual	8.9

Perform the following tasks in R:

- Import the dataset and display its structure using suitable R functions.
- Identify the data type and class of each variable using `typeof()`, `class()`, and `str()`.
- Convert Brand, Engine_Type, and Transmission into factor variables.
- Filter and display all cars having mileage greater than **18 km/litre**.
- Find the average mileage and average price of the cars.
- Convert all brand names into uppercase using suitable string functions.
- Create a new column named Price_Category with values **Low**, **Medium**, and **High** based on price.
- Write a short interpretation of how basic R data types and variables help in business data analysis.

Q2:

(5 Marks)

An e-commerce company wants to integrate customer, order, and product-review data stored in different files and visualise customer behaviour. The company wants to understand region-wise sales, category-wise product performance, and customer satisfaction.

Create the following three datasets in R:

Dataset 1: customers.csv

Customer_ID	Customer_Name	Region	Membership
CU101	Alisha	North	Gold
CU102	Ravi	West	Silver
CU103	Neha	South	Gold
CU104	Salim	East	Bronze
CU105	Priya	North	Silver

Dataset 2: orders.csv

Order_ID	Customer_ID	Product_Category	Amount
O1001	CU101	Electronics	45000
O1002	CU102	Clothing	8000
O1003	CU103	Grocery	3500
O1004	CU104	Electronics	52000
O1005	CU105	Clothing	12000
O1006	CU101	Grocery	5000

Dataset 3: reviews.csv

Review_ID	Customer_ID	Rating	Feedback
R01	CU101	5	Excellent
R02	CU102	3	Average
R03	CU103	4	Good
R04	CU104	5	Excellent
R05	CU105	2	Poor

Perform the following tasks in R:

- Import all three datasets using suitable R functions.
- Merge customer and order data using Customer_ID.
- Merge the resulting dataset with review data.
- Display the final integrated dataset.
- Find total sales region-wise and product-category-wise.
- Draw a bar chart showing total sales by region.
- Draw a pie chart or bar chart showing product-category-wise sales.
- Draw a scatter plot between Amount and Rating.
- Add suitable title, axis labels, colour, and legend wherever required.
- Interpret which region and product category performed best.

Q3:

(5 Marks)

An automobile manufacturing company wants to study vehicle safety, fuel efficiency, and transmission type using R. The company wants to apply statistical tests and predictive modelling to support product planning.

Use the built-in R datasets **Cars93** from the MASS package and **mtcars**. Perform the following tasks:

- Load the MASS package and import the Cars93 dataset.
- Extract the variables Type and Airbags.
- Create a contingency table between car type and airbag availability.
- Apply the Chi-Square test using `chisq.test()` and interpret whether car type and airbag availability are associated.
- Load the mtcars dataset and build a simple linear regression model using:
`lm(mpg ~ wt, data = mtcars)`
- Interpret how vehicle weight affects mileage.
- Predict mileage for cars having weights **2.5, 3.0, and 3.5**.
- Build a multiple regression model using:
`lm(mpg ~ disp + hp + wt, data = mtcars)`
- Identify which variable has the strongest effect on fuel efficiency.
- Build a logistic regression model to classify transmission type using:
`glm(am ~ hp + wt + cyl, data = mtcars, family = binomial)`
- Interpret which vehicle specifications influence manual or automatic transmission.
- Present all results with proper R output, graphs, and interpretation.