DATA PREPROCESSING AND ANALYSIS
(CSEN 5231)

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) Pearson's Correlation Coefficient helps us find measure of strength of association between two variables between.
(a) 0 to 1 (b) -1 to 1 (c) -1 to 0 (d) None of these

(ii) The data that does not follow the tabular data structure models associated with relational databases or other data table is
(a) Semi structured Data (b) Structured Data
(c) Unstructured Data (d) All of these.

(iii) Data visualizations are used to
(a) create reproducible code
(b) perform data analytics and build predictive models
(c) explore a given dataset
(d) share biased representation of data.

(iv) Data that summarize all observations in a category are called _____ data.
(a) frequency (b) summarized
(c) raw (d) none of the mentioned

(v) Data Cleaning
(a) Degrades the quality of data (b) Improves the quality of data
(c) Doesn't detect or remove errors. (d) None of the above.

(vi) Which type of data is generated by POS terminal in a busy supermarket each day?
(a) Source (b) Processed (c) Synchronized (d) All of the mentioned.

(vii) Matplotlib was created by
(a) Daniel Johnson, a German physicist
(b) John Hunter, an American neurobiologist
(c) John Butler, an American psychologist
(d) Cleve Moler, an American mathematician and computer programmer

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

(ix) Measure of Central Tendency: In a set of observations the unusual lower and higher values are called?
(a) Outliers (b) Free liners (c) Central liners (d) Median liners.

(x) Which of the following best describes the purpose of using ANOVA in research?
(a) ANOVA is used to compare mean of two groups
(b) ANOVA is used to compare the mean of more than two groups.
(c) ANOVA is used to determine correlation between variables.
(d) ANOVA is used to determine the interaction effect between dependent variables.

Fill in the blanks with the correct word

(xi) Consistency checking is a mechanism for checking whether rules does not contain _____ Conflicting elements.

(xii) _____ is a method of storing and presenting data in summary format.

(xiii) The Backend, Artist, and Scripting Layers are the three layers that make up the _____ architecture.

(xiv) _____ are used to explore a given dataset and perform data analytics and build predictive models.

(xv) Brainstorm diagram is a _____ Representation of data that was born way before the term "Data Visualization "was termed. It reads from the centre of dashboard or page.

Group - B

2. (a) '*An example of structured data is picture of goods/service*' – Justify the statement. *[(CO3)(Apply/IOCQ)]*

(b) '*Unstructured data can come from Facebook, Twitter and Presentations*' - Justify the statement *[(CO3)(Apply/IOCQ)]*

(c) Explain why XML may be useful for the transfer of structured data from one database to another. Consider such aspects as to whether there is a standardized relational format for data transfer, existence of libraries, and human readability. *[(CO3)(Apply/IOCQ)]*

3 + 3 + 6 = 12

3. (a) What do you mean by parsing ? Illustrate the use of parser how the set of rules in a grammar needed to construct valid statement in a Language. *[(CO3)(Apply /IOCQ)]*

(b) Differentiate horizontal scaling from vertical scaling .What are the goals of scalable platforms? *[(CO2)(Understanding/LOCQ)]*

(3 + 3) + (3 + 3) = 12

Group - C

4. (a) Why Mean and Median are both important in Statistical Data?

Height	Weight	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
5	45	−0.14	−5	0.7	0.019	25
5.5	53	−0.36	3	−1.08	0.129	9
6	70	0.86	20	17.2	0.739	400
4.7	42	−0.44	−8	3.52	0.193	64
4.5	40	−0.64	−10	6.4	0.409	100

From the table (above), find 1) Sum (Height), 2) Sum (Weight), 3) Mean (Height), 4) Mean (Weight), and 5) Correlation between variables Height and Weight. *[(CO4)(Analyze/IOCQ)]*

(b) Computing the mode, median and mean for a single variable measured on the interval or ratio scale is useful. Why? *[(CO4)(Analyze/IOCQ)]*

(3 + 6) + 3 = 12

5. (a) Explain with the help of block diagram how Extraction ,Transformation and Loading to warehouse can be done when multi source input are concerned . *[(CO4)(Analyse/HOCQ)]*

(b) What are the different mechanisms to handle missing data?
Illustrate with an example how will you deal with Missing data for the following dataset

Table 1

S_Id	Age_of_Student	Exam_Percentage	Hours_of_Study
1001	21	85	11
1002	20	90	12
1003	?	75	10
1004	21	80	9

[(CO3)(Apply/IOCQ)]

(c) Name few transformation techniques. *[(CO1)(Remember /LOCQ)]*

4 + 6 + 2 = 12

Group - D

6. An insurance company wanted to understand the time to process an insurance claim. They timed a random sample of 47 claims and determined that it took on average 25 minutes per claim and the standard deviation was calculated to be 3. With a confidence level of 95%, what is the confidence interval? *[(CO5)(Evaluate/HOCQ)]*

12

7. (a) Under Comparative statistics comes Paired t-test and Unpaired t-test . *[(CO1)(Remember /LOCQ)]*

(b) Explain step by step implementation of The Mann –Whitney test by taking an Example. *[(CO2)(Understanding/IOCQ)]*

(3 + 3) + 6 = 12

Group - E

8. (a) What is data visualization ? Differentiate Numerical data from categorical data. *[(CO1)(Remember/IOCQ)]*

(b) Write python code to visualize time series data.
Read sales.csv; print the data types of the parameters:
Convert date object to string. Visualize the time series data by performing sum over the grouping on date_Block_Id, Item_count_daily.Plot the figure with x.label 'Time', ylabel = 'Sales'. *[(CO5)(Apply/HOCQ)]*

(3 + 3) + 6 = 12

9. (a) What is a scatter plot? For what type of data is scatter plot usually used for? *[(CO2)(Understand/LOCQ)]*

(b) When will you use a histogram and when will you use a bar chart? Explain with an example. *[(CO2)(Understand/LOCQ)]*

(c) When analyzing a histogram, what are some of the features to look for? *[(CO2)(Understand/LOCQ)]*

4 + 4 + 4 = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	27.08	50	22.92

Course Outcome (CO):

After the completion of the course students will be able to:

1. Acquire knowledge in a broad range of methods based on statistics and informatics for data pre-processing and analysis and tools for visualizing the main characteristics of data.
2. Understand the whole process line of gathering relevant data, pre-processing the data, performing exploratory analysis on the data and visualizing the implicit knowledge extracted from data.
3. .Apply suitable methods for unveiling the underlying structure of the data, testing underlying assumptions in various fields
4. Analyze the results of experiment with the help of various visualization tools and statistical tests
5. Evaluate the performance of not only a computational method after obtaining different results by using different parameter values in order to choose the correct parameter value, but also, all similar methods in order to find out the best performing algorithm for a dataset.
6. Get familiar with relevant literatures, derive theoretical properties of the existing methods and come up with novel approach or pipeline for analyzing data across various fields by solving assignment problems

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