DATA PREPROCESSING AND ANALYSIS
(CSEN 5231)

Time Allotted : 3 hrs. **Full Marks : 70**

Figures out of the right margin indicate full marks.

Candidates are required to answer Group A and any 5 (five) from Group B to E, taking at least one from each group.

Candidates are required to give answer in their own words as far as practicable.

Group – A
(Multiple Choice Type Questions)

1. Choose the correct alternative for the following: **10 × 1 = 10**

(i) Data visualizations are used to (check all that apply):
- (a) create reproducible code
- (b) perform data analytics and build predictive models
- (c) explore a given dataset
- (d) share biased representation of data

(ii) Point out the correct statement.
- (a) Data has only qualitative value
- (b) Data has only quantitative value
- (c) Data has both qualitative and quantitative value
- (d) None of the mentioned.

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

(iv) 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.

(v) Which of the following data is put into a formula to produce commonly accepted results?
- (a) Raw
- (b) Processed
- (c) Synchronized
- (d) All of the Mentioned.

(vi) 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

(vii) What are the layers that make up the Matplotlib architecture?
(a) Figure Layer, Artist Layer, and Scripting Layer
(b) Backend_Bases Layer, Artist Layer, Scripting Layer
(c) Backend Layer, Artist Layer, and Scripting Layer
(d) FigureCanvas Layer, Renderer Layer, and Artist Layer

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

(ix) Point out the wrong statement.
(a) Merging concerns combining datasets on the same observations to produce a result with more variables
(b) Data visualization is the organization of information according to preset specifications
(c) Subsetting can be used to select and exclude variables and observations
(d) All of the mentioned.

(x) Which of the following step is performed by data scientist after acquiring the data?
(a) Data Cleansing (b) Data Integration
(c) Data Replication (d) All of the mentioned.

Group – B

2. (a) (i) '*An example of structured data is a picture of goods/service*' – Justify the statement.
(ii) '*Unstructured data can come from Facebook, Twitter and Presentations*' - Justify the statement.

(b) 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.

(3 + 3) + 6 = 12

3. (a) An example of structured data is (i) age information, (ii) customer reviews – Justify the correctness of the statement.

(b) What is the difference between data parsing and data transformation? Regarding data scalability, explain horizontal and vertical scaling.

4 + (4 + 4) = 12

Group – C

4. (i) Create a new column by normalizing the Weight (kg) variable into range 0 to 1 using the ***min-max normalization***.

Name	Weight (kg)
P. Lee	50
R. Jones	115
J. Smith	96
A. Patel	41
M. Owen	79
S. Green	109
N. Cook	73
W. Hands	104
P. Rice	64
F. Marsh	136

(ii) Create a new column by binning the Weight variable into three categories: low (less than 60kg), medium (60-100 kg), and high (greater than 100 kg).

(6 × 2) = 12

5. (a) Computing the mode, median and mean for a single variable measured on the interval or ratio scale is useful. Why?

(b) 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.

3 + (3 + 6) = 12

Group – D

6. Table-1 presents the ages for a number of individuals: Calculate the following statistics for the variable **Age**:
   (i) Range
   (ii) Variance
   (iii) Standard Deviation
   (iv) Z-score
   (v) Skewness
   (vi) Kurtosis.

Table with variables **Name** and **Age**

Name	Age
P. Das	35
M. Roy	52
A.Sen	45
S. Giri	70
R. Paul	24
D. Rai	43
T. Mitra	68
N. Roy	77
L. Das	45
R. Pal	28

Table-1

(6 × 2) = 12

7. An insurance company wanted to understand the time to process an insurance claim. They timed a random sample of 53 claims and determined that it took on average 30 minutes per claim and the standard deviation was calculated to be 3.5. With a confidence level of 95%, what is the confidence interval?

12

Group – E

8. (a) What are the ten data design '*Dos*' and '*Don't*'.

   (b) Explain five steps to design Information Visualization.

6 + 6 = 12

9. (a) What is a scatter plot? For what type of data is scatter plot usually used for?

   (b) When will you use a histogram and when will you use a bar chart? Explain with an example.

   (c) When analyzing a histogram, what are some of the features to look for?

4 + 4 + 4 = 12

Department & Section	Submission Link
CSE	https://classroom.google.com/c/MzEyNTIxOTI0MjMz/a/MzcxNjA3NDM0MzMw/details