University of Roehampton London

| CMP020L012A | Data Analytics | 2022/23 |
|---|---|---|

# Week-5
# Data Processing and Visualization using Off-the Shelf Tools

## Answer the following Multiple-Choice Questions

1. Amongst which of the following option is the most appropriate activity to be performed when a Data Scientist/Data Analyst acquires data from multiple sources and store those data in a single storage
    a. Deletion
    b. Integration
    c. Replication
    d. Generalization
2. Which one(s) of the following is a/are outcome(s) of visualization of Data in Data Science?
    a. Meaningful Representation of Data
    b. User-friendly Representation of Data
    c. Both a and b
    d. None of the above
3. It is possible to store, manipulate, manage, analyze, and generate insightful outcomes from raw data/dataset using the tool:
    a. Microsoft Word
    b. Microsoft Excel
    c. Microsoft Outlook
    d. Microsoft Powerpoint

4. Which of the following is not a Data Pre-processing method?
    a. Data Discretization
    b. Data Visualization
    c. Data Cleaning
    d. Data Aggregation
5. Which amongst the following are Data Cleaning tasks:
    a. Removing Noisy Data
    b. Correcting Inconsistencies in Data
    c. Transformation of Data
    d. All of the above
6. Normalization and Aggregation are Data Transformation Processes.
    a. True
    b. False
7. Binning is a method that is used for handling noisy data:
    a. True
    b. False

Kimia Aksir

8. Data Discretization is a Data Reduction activity that is particularly useful for _____ .
    a. Numeric Data
    b. Text Data
    c. Audio Data
    d. Image Data
9. Same Attribute may have different names in different data sources, that needs to be made consistent when the sources are integrated in a single place.
    a. True
    b. False
10. Min-Max Normalization performs a linear transformation on the original data
    a. True
    b. False
11. Compression of "Jpeg" is a Lossy Compression:
    a. True
    b. False
12. Which of the following method of data reduction is used for data redundancy detection:
    a. Aggregation
    b. Compression
    c. Dimension Reduction
    d. None of the Above

# Answer the following Questions

Assume following is a part of dataset collected from an institution is given to you where the people are of 20 to 25 years old age range, works in three different departments. Now before you use the data available in this dataset for any analytical purposes, where are the places you think data cleaning is necessary and how can it be done?

| ID | Height | Age | Departments | Employment Level | Payroll (Annum) |
|---|---|---|---|---|---|
| 10123 | 160.4cm | 21.5 | Literature & Linguistics | Senior | £40,000 |
| 10573 | 162.5cm | 25 | Literature and Linguistics | Junior | $35,000 |
| 10567 | 5 feet 3 inches | 24.6 | CSE | Junior | $360,0000 |
| 10647 | 158cm | | Dance and Drama | Senior | £40,000 |
| 13490 | | | | | |
| 14377 | | | | | |
| 10452 | 165cm | 20 | Computer Science and Engineering | Senior | £40,000 |
| 10630 | 162.7cm | 23 | Literature and Linguistics | Junior | $35,000 |