Practical 4

**Data Pre-Processing Review Questions**

# What are we doing?

Using what you learned from lectures (Lecture 4) and relevant reading materials, you will answer some review questions. These questions are for your self-review on topics covered: Data exploration and pre-processing. You will need to review lecture and reading materials or seek for other resources (e.g. Googling), in order to answer questions.

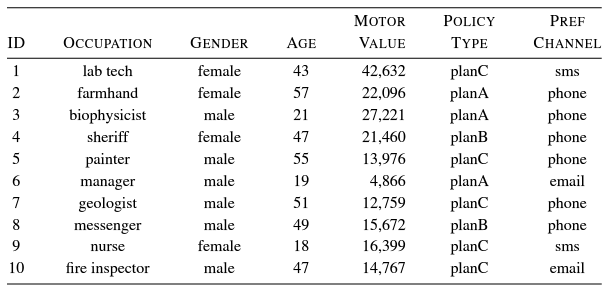
Submission:

You are required to submit one document containing your answers via the weekly-practical submission box (available on CP1407 LearnJCU)

For Laboratory questions, screen capture your computer screen after the completion of each task and include the captured image in your document to submit.

# Self-Review Questions

1. The table below shows a sample of a larger dataset containing details of policy holders at an insurance company. The descriptive features (attributes) included in the table describe each policy holders’ ID, occupation, gender, age, the value of their car, the type of insurance policy they hold, and their preferred contact channel.



1. State whether each descriptive feature contains numeric, interval, ordinal, categorical, binary, or textual data. (For example, the ‘Gender’ feature is categorical data.)
2. How many levels does each categorical and ordinal feature have? (For example, ‘Gender’ feature has 2 levels (male, female))
3. The table below contains sample data about the employees of an IT company.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Emp ID | Name | Year of Birth | Gender | Status | Salary |
| 100 | Smith | 1954 | M | Director | $200.000 |
| 125 | Jones | 1967 | F | Technician | $36,000 |
| 167 | Highley | 1975 | F | Senior Technician | $70,000 |
| 200 | Millins | 1987 | M | Technician | $32,000 |
| 205 | Dujevic | 1985 | M | Technician | $34,000 |
| 216 | Isovic | 1985 | F | Technician | $34,000 |
| 220 | Sun | 1986 | F | Senior Technician | $66,000 |
| 301 | Bean | 1955 | M | Deputy Director | $160,000 |

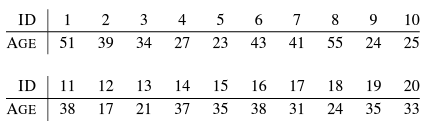
Answer the following questions.

* 1. An input data set consists of individual data objects, also known as data records, instances or examples. All data sets have the common properties: Type, Size, Dimensionality, Sparsity, Abstraction
* Abstraction: Data can be presented at different levels of abstraction. At the least abstract level, data are presented as they are recorded, known as the raw data. Summarisation and approximation on the raw data can produce a level of more abstract representation. For instance, in an analysis of student intake numbers over the years, individual student record details are not of interest but the total number of students each year is of interest. The latter value is summarised over the original data records and is used for further analysis.

Sum up the following characteristics of the table (as above) as a data set: type, size, dimensionality, sparsity and level of abstraction.

* 1. If the Salary attribute needs to be discretised into three pay bands, suggest a simple yet sensible solution for the discretisation backed with a valid argument.
  2. If Mr Dujevic’s salary were unknown and the unknown value needed to be imputed, what is a sensible replacement value and why?
  3. Among the employee records, which record can be considered as an outlier? What harm can an outlier object cause to the understanding of the data set?

1. A team of researchers conducts research into face recognition. The team has created a database of short colour video clips of frontal face images for 50 human subjects of different genders and ethnic groups. Each video clip consists of 100~120 frames (still images) of 120x80 pixels and is stored in a separate file. Describe how to compose a table from the video clips for classification purposes. In your description, you must state clearly what the rows and columns of the table represent. Which type of data set does the table belong to?
2. The table below shows the age of each employee at a cardboard box factory.



Based on this data calculate the following summary statistics for the AGE feature.

(This task is just to check your understanding about fundamental summary methods to describe the data set. You are not required to calculate actual numerical results. Instead, try to show how each statistical summary can be produced using the data given.)

1. Minimum, maximum and range
2. Mean and median
3. Variance and standard deviation
4. 1st quartile (25th percentile) and 3rd quartile (75th percentile)
5. Inter-quartile range
6. 12th percentile

# Laboratory Questions

1. The table below presents a data set about student homework and examination results.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Student ID | Home-work 1 | Home-work 2 | Home-work 3 | Exam |
| 1 |  | 94 | 34 | 42 |
| 2 | 35 | 94 | 85 | 45 |
| 3 | 31 | 46 | 22 | 48 |
| 4 | 46 | 90 | 60 | 50 |
| 5 | 52 | 94 | 49 | 50 |
| 6 | 58 | 94 | 30 | 51 |
| 7 | 47 | 90 |  | 52 |
| 8 | 37 | 94 | 25 | 52 |
| 9 | 35 | 94 | 45 | 54 |
| 10 | 57 | 94 | 100 | 54 |
| 11 | 51 | 94 | 5 | 54 |
| 12 | 45 | 94 | 33 | 55 |
| 13 | 44 | 0 | 35 | 55 |
| 14 | 52 | 95 | 36 | 56 |
| 15 | 35 | 94 |  | 57 |
| 16 | 57 | 97 | 57 | 57 |
| 17 | 45 | 90 | 71 | 57 |
| 18 | 39 | 94 | 54 | 57 |
| 19 | 31 | 94 | 63 | 57 |
| 20 | 45 | 94 |  | 59 |
| 21 | 35 | 90 | 84 | 59 |
| 22 | 37 | 90 | 40 | 61 |
| 23 | 83 | 97 | 26 | 61 |
| 24 | 68 | 97 | 55 | 62 |
| 25 | 50 | 95 | 56 | 62 |
| 26 | 77 | 93 |  | 63 |
| 27 | 84 | 48 | 18 | 63 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Student ID | Home-work 1 | Home-work 2 | Home-work 3 | Exam |
| 28 | 45 | 90 | 21 | 63 |
| 29 | 62 | 95 | 38 | 63 |
| 30 | 38 | 94 | 40 | 64 |
| 31 | 50 | 90 |  | 64 |
| 32 | 32 | 90 | 38 | 64 |
| 33 | 44 | 90 | 43 | 65 |
| 34 | 57 | 94 | 52 | 68 |
| 35 | 50 | 94 | 39 | 70 |
| 36 | 55 | 90 | 62 | 71 |
| 37 | 43 | 94 | 54 | 72 |
| 38 | 50 | 90 | 30 | 74 |
| 39 | 54 | 90 | 82 | 77 |
| 40 | 64 | 95 | 5 | 78 |
| 41 | 85 | 95 |  | 79 |
| 42 | 63 | 90 | 62 | 82 |
| 43 | 75 | 90 | 35 | 83 |
| 44 | 86 | 97 | 39 | 84 |
| 45 | 77 | 95 | 79 | 84 |
| 46 | 79 | 94 | 35 | 86 |
| 47 | 86 | 98 | 57 | 87 |
| 48 | 71 | 90 | 9 | 89 |
| 49 | 45 | 94 | 72 | 90 |
| 50 | 90 | 94 | 68 | 92 |
| 51 | 89 | 94 | 53 | 93 |
| 52 | 90 | 98 | 79 | 98 |
| 53 | 57 | 92 | 40 |  |
| 54 | 36 | 94 | 54 | 22 |

Use MS Excel and WEKA to create an ARFF file for the data set in the table above. (Store this example data in MS Excel 🡪 save as .csv file 🡪 open the csv file in WEKA 🡪 save as .arff file in WEKA)

Open the ARFF file in Weka, and perform the following tasks .

* 1. Observe the summary data for the data set and the histograms for all attributes on the Preprocess tap page. Use the Visualize tab page to view the scatter plots between the variables of the data set.
  2. Apply the unsupervised Discretize filter to the exam marks.
  3. Practise filling in missing values in Weka both manually in the Viewer window and by using filters.

1. [Extra Task: this task is optional and are not required to complete or submit]

Principal component analysis (PCA) is a useful tool to reduce dimensionality of a given data set. WEKA is equipped with a PCA filter, which can be used on the Select Attribute tab of the Explorer. Open one of the data sets provided in WEKA, such as *cpu.arff*. On the *Select Attribute* tab, press the *Choose* button and select the PricipalComponents filter. Press the Start button and observe the output in the Attribute Selection Output window. The window should show the new attributes in eigenvectors, each of which is a linear combination of the original attributes and the ranking among them according to their significance. Discuss your findings.