Artificial Intelligence 2 – Assessment 1

Data Quality Report

For this assignment you need to write Python code to generate the continuous and categorical feature tables in a Data Quality Report from and provide an analysis of the data issues in the dataset.

For the purposes of this assignment you **DO NOT** have to include the bar plots and histograms as part of the data quality report you submit.

Read the specification below carefully because marks will be deducted if the instructions are not followed.

SUBMISSION DEADLINE:

27th Feb (Monday) 8am

This assessment contributes 10% of the module mark. Marks will be deducted for late submissions – 10% for each day after the deadline.

SUBMISSION PROCESS:

Submit a zip file containing all the required files (described later) via webcourses. This zip file should be named STUDENTNUMBER.zip where the string STUDENTNUMBER has been replaced by your student number, for example: c123456789.zip

INPUT: Your code should expect the following files as **input** (note the paths to the files):

1. Your program should extract the names of the features in the dataset from a file called 'featureNames.txt'. Your program should expect to find this file using the following path './data/featureNames.txt'. Each line in this file will contain the name of one feature and the first line will contain the name of the feature in the first column in the dataset, the second line will contain the name of the second column in the dataset, and so on.

2. Your program should expect that the dataset is in a comma separated file called 'DataSet.txt' that is stored in a directory called 'data' that is a subdirectory of the directory your program is run from (in other words the path to the dataset file should be './data/DataSet.txt')

OUTPUT: Your code should **output** the following files (note the paths to the files):

- Your program should output the table for the continuous features to a comma separated file using the following path './data/studentnumberCONT.csv' and for the categorical featuers to a commas separated file with the following path './data/studentnumberCAT.csv'.

In other words, the comma separated files containing the tables should be written to the same directory that you get the input files from. In the file names, replace the string studentnumber with your student number.

- The format of this file should mirror the continuous feature table in the data quality report as presented in the notes (see below for more info on file formats).

  The first line in each file should be a header line - a comma separated listing the descriptive feature name for each column in the file (use the string FEATURENAME for the name of the first column). Each of the subsequent lines in the file should be a comma separated list with the name of the feature as the first element in the list and then the descriptive statistics in the subsequent commas separated elements in the list in the same order as they are listed in the notes.

WHAT YOU SHOULD SUBMIT:

1. The **Python source code** you wrote for the assignment. This source code should be in a file called **studentnumber.py** where studentnumber has been replaced with your student number. Also, include your name and student number at the top of the file as comments.
2. The **Data Quality Report table for the continuous descriptive features**—as identified by your code—in the dataset. This table should be in a comma separated file and the name of the file should be named: **studentnumberCONT.csv**
3. The **Data Quality Report table for the categorical descriptive features**—as identified by your code—in the dataset. This table should be in a comma separated file and the name of the file should be named: **studentnumberCAT.csv**
4. **A brief (1 page) description of the dataset** that describes your analysis of the dataset in terms of missing values, outliers, feature cardinality, etc. and your opinions as to what should be done to address these quality issues.