# CMPU 4011 Machine Learning for Predictive Analytics

**Deadline: 25th Oct 2019 (Friday)**

---

## 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:

**26 5h Oct (Friday)**

This assessment contributes 10% of the module mark.

(CA2 will be another 20%, with overall 30% of the module mark coming from Continuous Assessments, and 70% from Exam)

Marks will be deducted for late submissions – 10% for each day after the deadline.

### PYTHON INPUTS AND OUTPUT

Below are the expected inputs and outputs of your Python program:

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

- Your program should extract the names of the features in the dataset from a file called 'feature_names.txt'. Your program should expect to find this file using the following path './data/feature_names.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.

- Your program should expect that the dataset is in a comma separated file called 'dataset.csv' that is stored in a directory called 'data' that is a subdirectory of the directory your program is run from (so the path to the dataset file should be './data/dataset.csv')

**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 'studentnumberCONT.csv'
    - the table for the categorical features to a commas separated file'studentnumberCAT.csv'.

  where you replace the string studentnumber with your student number.

- The format of these files 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:**

Submit all the required files via Brightspace. Your submission should contain:

1. The **Python source code** you wrote for the assignment. Include your name and student number at the top of the file as a comment (Jupyter notebook or .py file)
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 types of features (categorical vs. continuous), potential data quality issues with the data, e.g. missing values, outliers, feature cardinality, etc., as well as your opinion as to what should be done to address these quality issues.

**PLAGIARISM:**

Plagiarism is a serious offence – do not use other people's code, as well as do not share your files with others and do not share them online (e.g. public github)!

**MARKING SCHEME:**

| | |
|---|---|
| Table: studentnumberCAT.csv | 40 marks |
| Table format | 10 |
| Column headings | 10 |
| Correct features | 10 |
| Correct values | 10 |
| Table: studentnumberCONT.csv | 40 marks |
| Table format | 10 |
| Column headings | 10 |
| Correct features | 10 |
| Correct values | 10 |
| Report | 20 marks |
| 15-20 marks | The student has clearly put a lot of thought and effort into the analysis and has covered all the major points with respect to data quality issues (missing values, outliers, cardinality issues). |
| 10-15 marks | The student has done a good job but hasn't really gone in depth beyond the basic analysis. |
| 5-10 marks | The report is very basic and minimal. |
| 0-5 marks | No effort has gone into the report (e.g. mentions generic concepts with no reference to the data set, or copy/paste from the notes). |