M.Sc. In Data Science – Programming for Data Scientists

Coursework Assignment 1

Improving Data Quality – Data Cleansing

Due date: November 3rd 2017 by 5:00pm (end of week 6). **Worth:** 25%.

Email your submission to paul.barry@itcarlow.ie

Coursework Description:

In David Beazley's PyData 2016 talk (see *Resources*, below), he identifies a number of problems with the "quality" of the data (most memorably the spelling on "McDonalds"). You are to examine the data used by Beazley to identify other data quality problems. Once identified, you are to devise data cleanup processes for the data which you are then to implement in Python.

Deliverables:

A Jupyter Notebook which provides a textual description of the data quality issues you identified in the data, together with a textual description of your proposed "fixes" for cleaning the data. Python code which performs each of your cleanups is also required.

The "output" from your notebook is a new CSV file with your cleanups applied.

Note: you are to use the techniques and tools provided by standard Python (which we have discussed in class). At this point in time, you are to avoid using any 3rd party tools (such as `pandas`).

Resources:

- David Beazley's "built-ins" talk is on YouTube: https://www.youtube.com/watch?v=j6VSAsKAj98
- The data David uses can be viewed (and downloaded) from here: https://data.cityofchicago.org/Health-Human-Services/Food-Inspections/4ijn-s7e5 - Click on the **Download** button, then select *download as CSV*.
- The Python Standard Library: https://docs.python.org/3/library/index.html

Some data cleansing “questions” for you to consider:

- Is there any obsolete or redundant data?
- What about missing data? How will you deal with missing data (if any)?
- Are there any obvious outliers?
- Is the data usage/encoding consistent? Spelling mistakes? Mislabelling? Multiple value for the same thing?
- What about duplicates? Are there any? What can you do about them?

Can you think of any other questions/issues? If so, document and implement them.

There is lots of good material on “data cleansing” on the web (try some Google searches), and a good text in this area is this one (especially Chapter 7):