Data Analytics Coursework 1

# Introduction

# Data Preparation

Before starting off, I had to first edit the original, dirty dataset to add headings. The entire dataset was lacking any headings which would be devastating for data cleaning and management.

## OpenRefine

### Facets

When performing operations on a dataset with OpenRefine, one of the most powerful tools at disposal are ‘Facets’. This tool allows for easy viewing of unique entries in fields when using the ‘Text Faucet’ tool, as well as easy viewing of numeric ranges with the ‘Numeric Faucet’ tool.

#### credit\_amount

|  |  |
| --- | --- |
| The *credit\_amount* attribute contained exceedingly large values, far outside the of anything else within the range.  The 7 *out-of-range* entries all seemed to contain a lot of 0’s at the end of them. Under the assumption that these were inserted as pennies, instead of full pounds, I removed the trailing two 0’s for each of those entries. |  |

#### class

The *class* attribute featured stand-out entries such as ‘1’ and ‘0’. I deferred to using the *computing* approach and assuming that ‘0’ means ‘bad’, ‘1’ means ‘good’.

#### purpose

The *purpose* attribute required the most work. It features many mistakes: typos, shortenings, format, etc. I fixed the typos for; *business, education, furniture/equipment, radio/tv, used car.* These were all obvious typos/shortenings and didn’t require much thought before editing. I followed through by cross-referencing the brief to see which values should be set as the ‘correct’ ones.

#### job

The *job* attribute had some incorrect entries such as the ‘good’ and ‘poor’ entries. I designated these to ‘skilled’ and ‘unskilled resident’ respectively. I also cleaned up and made the quotation marks consistent. [*(Although this is addressed anyway).*](#_Quotation_marks)

### Transformations

Transformations are done on the data by: clicking on the field title, Edit Cells -> Common transforms -> To text/number. This is done to establish what type of data that field is. This is important for later data processing so that numeric values aren’t treated as nominal values – (strings).

‘To number’ transformations are done on fields; *Case\_no, credit\_amount, age.* All the rest of the fields are also forcefully ‘To text’ transformed, just to be sure they are in the right format.

### Quotation marks

Because of issues, which are later seen during the data processing, I decided it would be best to remove all quotation marks from the dataset. This included essentially every string’s first and last character. I did this by going through each facet for each attribute and editing all instances.

## Data Transformation and Conversion

## Data Framework and Visualisation

## Scientific Analysis

# Appendix