Data Quality Issues

Some of the data quality issues handled by the current script are described below.

# Encoding of the file

1. By default, the pandas library reads csv in ‘UTF-8’ format. Before processing the file using pandas, detect the encoding of the file is the first part of the pipeline.
2. The current script finds the encoding and then reads it by specifying the detected format.

# Data Pre-processing

1. Since there were duplicate rows with same values, we removed them.
2. There were some special characters in the `**donor\_id** `column which were removed.
3. Whitespaces and indentation issues for `**donor\_id**` column was handled.
4. Quality Issue with `**postcode**` column as some values in postcode were of float type ‘nan’
5. Quality Issue with `**donor\_type**` column as some values in postcode were of float type ‘nan’
6. Gender column has different values.
7. The `**postcode** were extracted based on regex (4 digit).
8. **`birth\_date`** column was formatted to appropriate date type using pandas library.

# FIXES

1. Extract `**postcode`** with format 4-digit regex search.
2. **`Gender** `values were converted to one of the following ‘F’. ‘M’ or null. There were some values which were MALE or FEMALE. These are converted to M or F respectively. Others such as UNKOWN, UNK or misspellings were converted to null.
3. The `**donor\_id**` column values for digits less than 19 were appended zero.
4. All the values in postcode which had ‘nan’ are returned with null postcode values.
5. Similarly, all values in `**donor\_type**` which had nan float values are converted to **-1**.
6. The donor\_type column was cast to `**int32**` type before saving it as a table.