Section 1:

COMP 6214 Open Data Innovation

Course: MSc Computer Science

Student Name: Honggang Wang

Student ID: 31531857

Section 2: Open Data Cleaning

2.1 Tool Used for Data Cleaning

OpenRefine is an independent open-source desktop application for data cleaning and conversion to other formats. It is an application running on the local machine, which means that there is no need to upload large data sets to a web service. Besides, the advantage of this is that the data is still private[1].

The following screen is the import screen and shows a preview of what Refine thinks the dataset should look like.

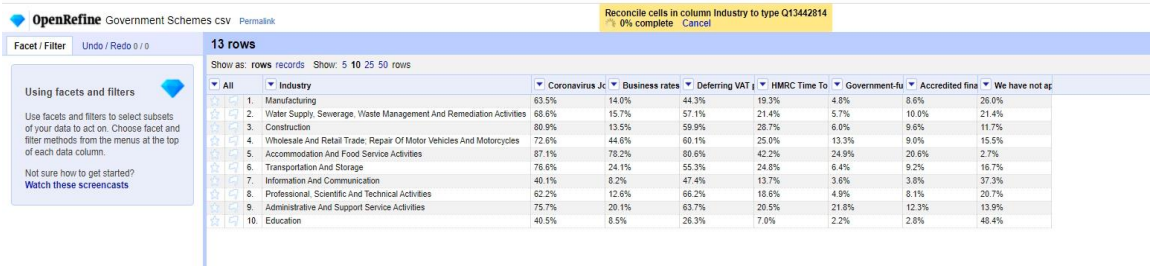

Figure 2.1: Example screenshot of the Government Scheme worksheet

2.2 List of Errors

Here is the list of error value and error type which found from the dataset.

	A	B	C	D	E
1	No.	Sheet Name	Cell	Error Value	Value After Correct
2	1	Sample	C20	7181.97	7180
3	2	Response Rates	C15	-59	59
4	3	Response Rates	H16	-43.40%	43.4%
5	4	Response Rates	J7	217.50%	27.5%
6	5	Trading Status	Column C	*	Recalculated, displayed as a value
7	6	Government Schemes	H16	1200.60%	12.6%
8	7	Government Schemes (2)	D32	46.1-%	46.1%
9	8	Government Schemes (2)	Multi	*	Recalculated, displayed as a value
10	9	Government Schemes (3)	B17	-36.80%	36.8%
11	10	Government Schemes (3)	H7	42.2+%	42.2%

Table 2.1: List of Errors(1)

No.	Type of Error	Analysis of Error
1	Mixed use of numerical scales	The sample size should not be a decimal number
2	Nonsensical data entries	The sample size should not be negative
3	Nonsensical data entries	The percentage should not be negative
4	Saturated data	The percentage should not be greater than 100% according to the contextual understanding
5	Missing value error, '*' does not match the data type of the subsequent visualization	*' is a character, which cannot be parsed by visualization tools
6	Saturated data	The percentage should not be greater than 100% according to the contextual understanding
7	Nonsensical data entries	Wrong character usage
8	Missing value error, '*' does not match the data type of the subsequent visualization	*' is a character, which cannot be parsed by visualization tools
9	Nonsensical data entries	The percentage should not be greater than 100% according to the contextual understanding
10	Nonsensical data entries	Wrong character usage

Table 2.2: List of Errors(2)

(a) Error 1

Since the data in this worksheet is the number of surveys, the data should be integers. After using OpenRefine to calculate the sum function for various industries where the workforce size is greater than 250, the integer 7180 is obtained.

(b) Error 2 & 3

In the worksheet named 'Response Rate', the number of responses is negative. This can be achieved by using column transformation for the same type of data in OpenRefine. The data will be distributed and sudden outliers can be found. The value is a negative number, so change this data. Similarly, the response rate will not be a negative number.

(c) Error 4 & 6

The data problem in this situation is saturated data. These two data exist in the worksheets 'Response Rate' and 'Government Scheme'. Both of these values are percentages, and the meaning expressed here is the percentage of the survey sample to the total. There will be no more than 100%. , So after the contextual comparison and calculation in the data table, the correct data value can be obtained.

(d) Error 5 & 8

The two data problems here are due to missing values in the table and replacement with the '*' symbol. According to the data before and after the table, it can be found that the sum is 100%, so the correct value can be obtained. Such values will affect the visualization of the data.

(e) Error 7, 9 & 10

Symbols appear in different positions of the three values here, and the values are no longer reasonable. In OpenRefine, use indexOf() to find the position where the unreasonable character appears and replace it.

2.3 Validation of the Resulting Cleaned-up Data

First, I used manual calculations to test these improved data one by one, to improve the reliability and readability of the data. And try to find out if there are missing unaltered wrong data. Secondly, use the CSV lint tool to verify the data file. This tool will detect and compare the data content and data type of each row and column, and try to find inappropriate data. Finally, use Excel to test the data in each table one by one. Excel has data verification functions, such as setting conditions for percentage data. The data in the column needs to meet the conditions of percentage and less than or equal to 100% and greater than or equal to 0%, otherwise, it will prompt that the data is invalid.

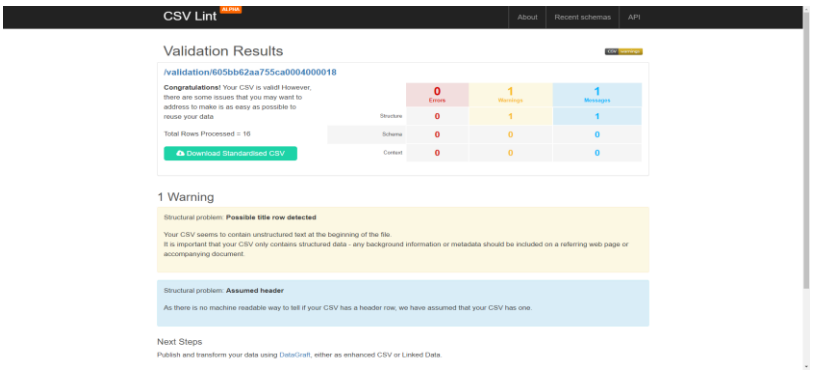

Figure 2.2: Validation result returned by CSV Lint

Section 3: Open Data Modelling

The URL of the RDF: honggang-soton.com/report

When you open this webpage you can find the files embedded on the website(Ontology and Linked Data they both share the same URL within the URL I attached above).

I have used a tool to create my ontology and modelling the open data by using Protégé.

Protégé is a free, open-source ontology editor and a knowledge management system. First, I created the ontology by modifying the entities, creating 6 classes each of them has several subclasses. At the same time, subclasses conclude the open data from the dataset which has already been cleaned up.

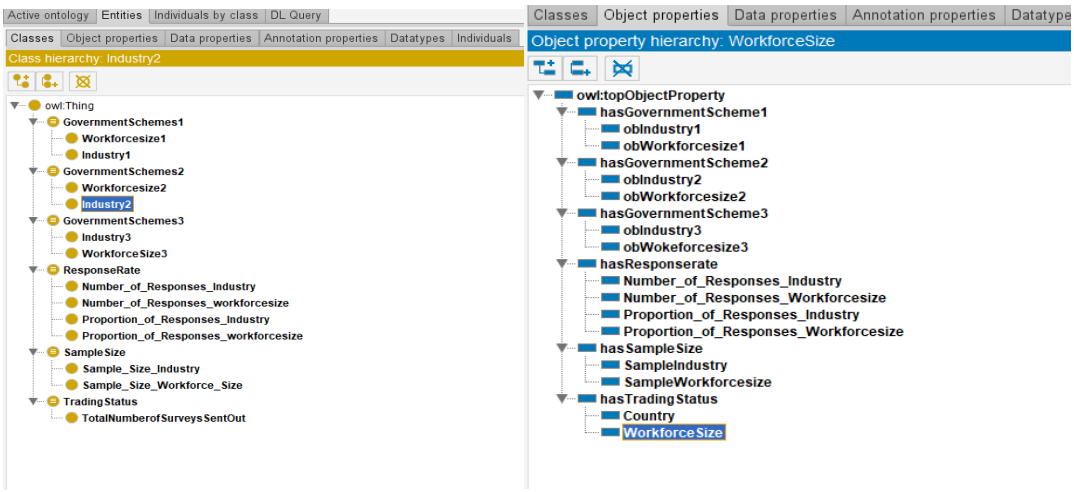

Figure 3.1: Creating Classes and Objects by Protégé

The object attributes include different attributes in different tables. In the industry data table, each industry (such as manufacturing) is an object belonging to its industry object. There are several different types in the data properties, there are numeric types, including percentages, integers, decimals, and String and Name types. After the data model is structured and described, it is necessary to use the function of *Cellfie* to fill the data, and use the domain-specific language to identify the data in the datasheet and fill it in the appropriate position. Different types of data in the data properties are used when filling, and the title type is used when filling the title of the table. In addition, when filling the value, it will correspond to the title and industry or workforce size, country and other data attributes to fill the data in the correct position.

The main purpose of ontology is to classify things according to semantics or meaning. In OWL, this is achieved through the use of classes and subclasses. Individuals who are members of a given OWL class are called their class extensions. While using OWL ontology, object-oriented thinking is used for modelling to make its data model more structured and shareable. In addition, using the Dublin Core to describe data, Dublin Core can create concise and descriptive records. The Dumb-down Principle can facilitate creation and maintenance, and it has commonly understood semantics.

Section 4: Open Data Visualization

First of all, here is the URL of the web application hosted with a set of open data visualizations.

URL for Visualization: honggang-soton.com

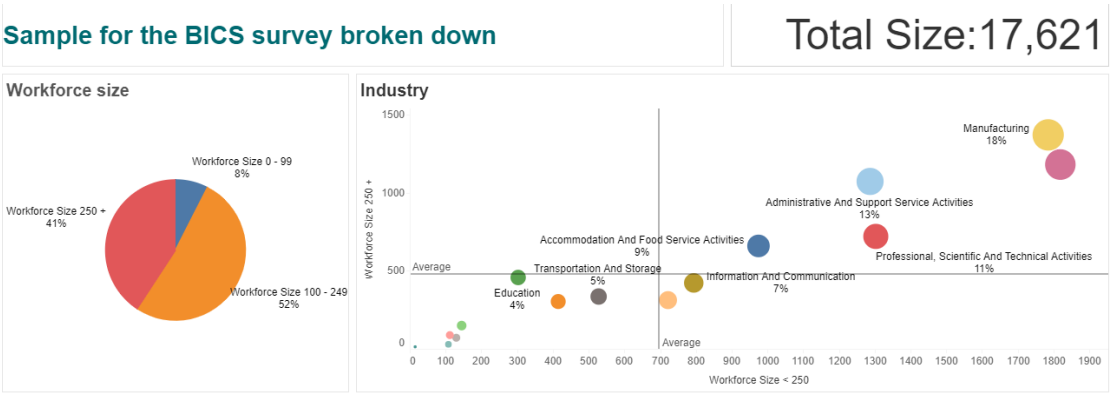

Figure 4.1: Visualization from hosted URL

I used several different charts and interactive methods to present my visualization. First of all, because there is some percentage of data, I used a pie chart to visualize. Because the pie chart can intuitively see the size of the percentage for each category. At the same time, I displayed the percentage value on a two-dimensional axis and use circular patterns of different sizes and colours on the axis to indicate the size of the data, combining the data from the industry and workforce size tables.

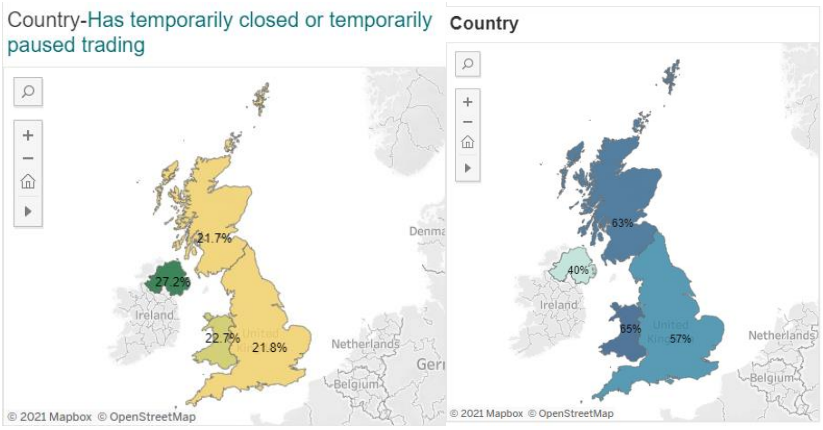

Figure 4.2: UK map for open data

In addition, I used a map for the United Kingdom to point out the different data in the four regions under different circumstances. And to enhance the interactivity and aesthetics, I added an optional menu next to it. You can select different attributes from several sets of data to present different charts.

Finally, using a box plot can not only elucidates the distribution of values along the axis, but also see its maximum, minimum, and quartile.