Open Data Innovation

Name: Zehao Xue

Student Number: 29176158

Open Data Cleaning

A. The tool used for data cleaning is Microsoft Excel. Through examining the data from the data set given, the size of dataset were fairly small it is easier to use spreadsheet tool like Excel to fix these errors.

B. A list of the error or error types you found in the dataset

1. Format issue across different spreadsheets, "*" is used to represent any value less than 1% however, this data comes in different format therefore "*" is replaced with 0.0%.
2. In sheet Government Scheme 1, row 16 and column H value is 1200.6%. This is incorrect as the it should not exceed 100%. This error have been modified and value of changed to 12.06%
3. In sheet "Response Rate" There is a semantic error, where the number of responses and proportion of responses were negative values. This is incorrect, as the minimum number of response can not be less than 0. Through examining the dataset, the data have been cleaned with the by removing the negative sign.
4. In sheet Government Scheme 2, Row 32 and column D, the format of this value is different compare to other records with an extra "-" sign. This is inconsistent with other record therefore This is solved by removing the - sign.
5. In Government Scheme 3, B17 the value were negative, and it is incorrect as the minimum percentage should not be below 0. This is fixed by removing the negatives sign.
6. In Sample Size, C20 and E28 have decimal places, this is inconsistent with the real world data since, the the total number of survey can only be integer values.

C. How you validated the resulting cleaned-up file

Some validation have been done, by inspecting the dataset and understanding the semantic of the data. For an example, for the data in "sample size" dataset the total number of survey is validated by calculating the total number of survey sent out to each industries and workforce size band.

Open Data Modelling

Process of Modelling the data

Producing Schema file

The schema file, is one of the two outputs from data modelling process. Its purpose is to define set of classes that represents the entities within this dataset. Inside of the schema file, consist of 6 classes each classifies different entity attributes within the dataset.

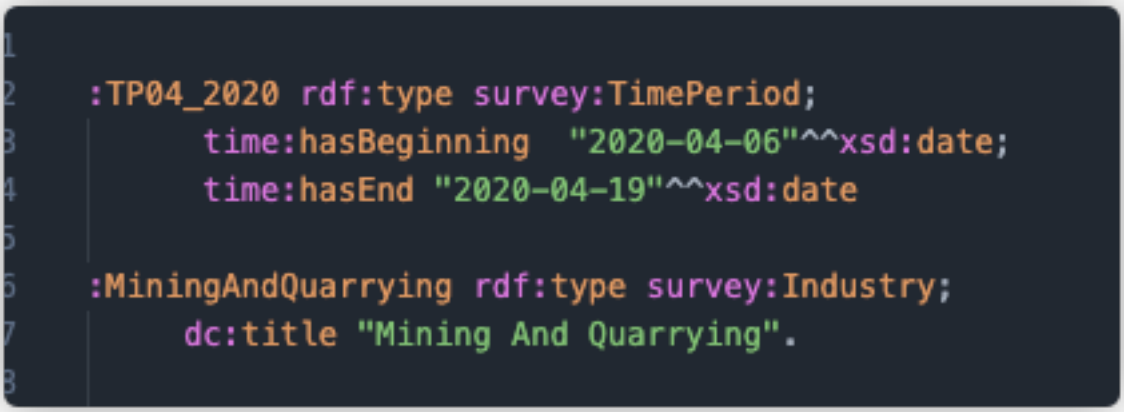
```
#Dimension - row
:Industry rdf:type owl:Class;
    rdfs:subClassOf qb:DimensionProperty;

#Dimensions - columns
:WorkForce rdf:type owl:Class;
    rdfs:subClassOf qb:DimensionProperty.
#Dimensions - row
:Country rdf:type owl:Class;
    rdfs:subClassOf qb:DimensionProperty.

:TradingStatus rdf:type owl:Class;
    rdfs:subClassOf qb:DimensionProperty.

:SchemeType rdf:type owl:Class;
    rdfs:subClassOf qb:DimensionProperty.

:TimePeriod rdf:type owl:Class;
    rdfs:subClassOf qb:DimensionProperty.      You, 21 h
```

The Industry class is used to represent different industry instances within the dataset, the “WorkForce” is used to represent to represent the different size band in terms of a company’s work force size. The “Country” class represents different countries the company is from. The “SchemeType” represents different government schemes and TimePeriod is representing the time period the survey was taken.

The schema file will be referenced under prefix name “survey” to populate class instances of the dataset attributes.

Producing Linked data

The linked data file, is populated by using python script. The linked data consist three parts including the instances of dimension attributes and instances of datasets e.g tables in the spread sheet and data points in the each datasets. The instances of attributes are populated using by reading the rows and columns of each tables and produce RDF triplet for each distinct attribute instance. Different tables may share common attributes, which would be classify as same instances when they have same semantics in the dataset. The attributes instance will be interpreted as a class instance defined in the schema file.

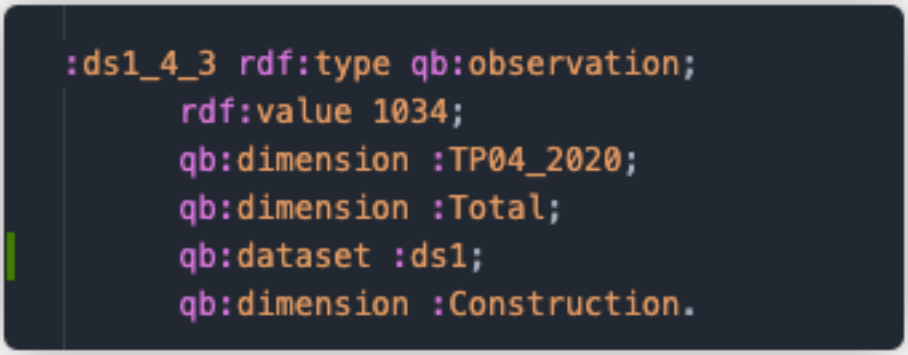


```
:TP04_2020 rdf:type survey:TimePeriod;
      time:hasBeginning  "2020-04-06"^^xsd:date;
      time:hasEnd "2020-04-19"^^xsd:date

:MiningAndQuarrying rdf:type survey:Industry;
     dc:title "Mining And Quarrying".
```

The tables as a “qb:dataset” where the dataset contains a “dc:title” that specifies the name of the table and “qb:observations” which represents each cell in a table.

```
:ds12 rdf:type qb:dataset;
      dc:title"Government Scheme3 By Industry";
    qb:observation :ds12_1_1;
    qb:observation :ds12_1_2;
    qb:observation :ds12_1_3;
    qb:observation :ds12_1_4;      You, 4 days ago •
    qb:observation :ds12_1_5;
```

Every cell in a table is modelled as a instance of “qb:observation”, where the instance consist of dimension properties that is defined in previous section for an example, a data point may have a dimesions of industry of “Construction” and dimesion of work force size of “250 and more.”

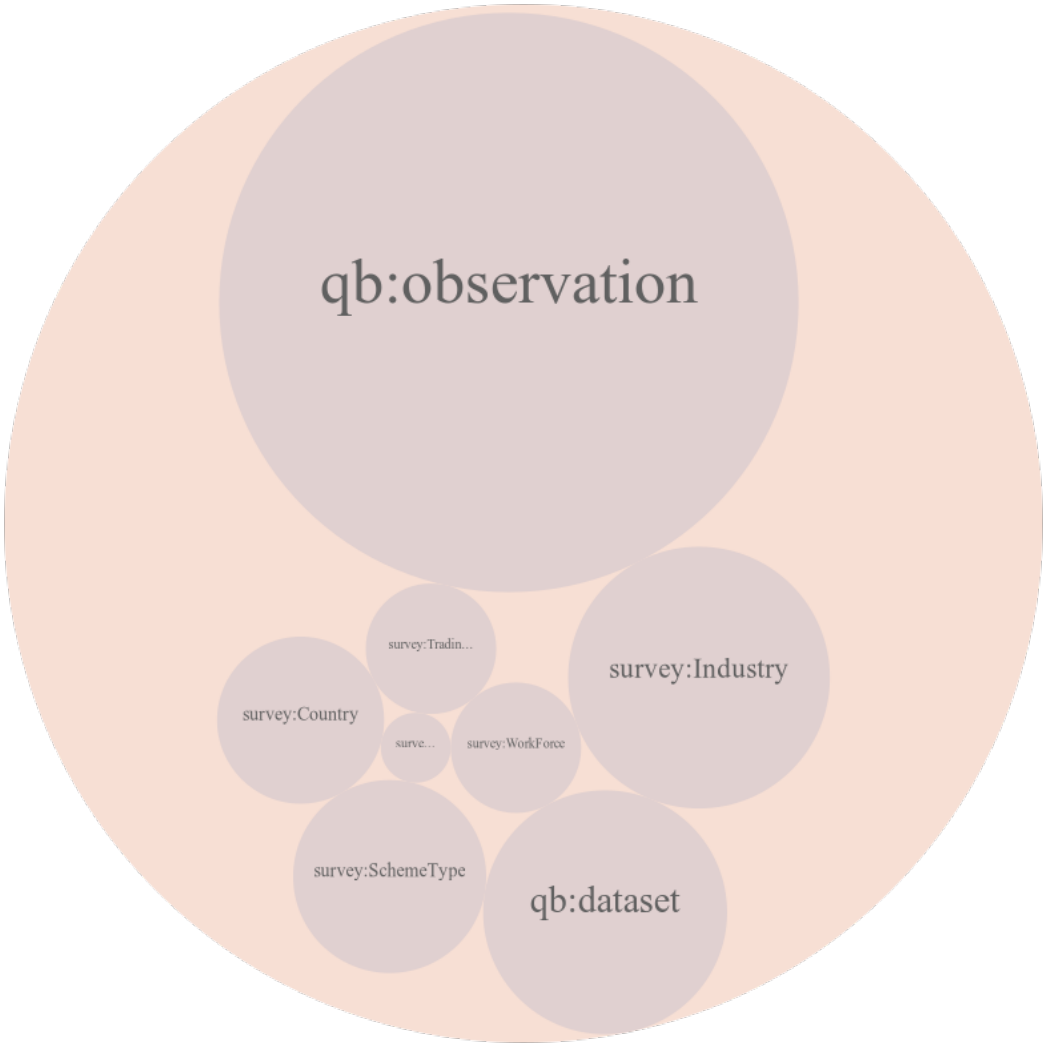


```
:ds1_4_3 rdf:type qb:observation;
      rdf:value 1034;
      qb:dimension :TP04_2020;
      qb:dimension :Total;
      qb:dataset :ds1;
      qb:dimension :Construction.
```

Ontologies Chosen

Prefix	URI	Reason for use
owl	http://www.w3.org/2002/07/owl#	owl is used to decribe classes
xsd	http://www.w3.org/2001/XMLSchema#	xsd is used to describe, the data type of the values in the dataset

Prefix	URI	Reason for use
time	http://www.w3.org/2006/time#	time is used to describe the time period of each dataset
qb	http://purl.org/linked-data/cube#	qb is used to describe relations between dataset components including qb:obseration for describing the data points. qb: dimension for describing the columns and rows of dataset
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#	rdf is used to describe relationships between entities in the turtle file for an example “mining and quarrying” is rdf:type Industry
dc	http://purl.org/dc/elements/1.1/	dc is an ontology for describing books, the element “title” is used in linked data to describe the title for each class entity.
survey	http://example.org/schema/survey/	Desinated ontology for this coursework, that defines set of owl:classes to group common properties of dataset.



The generated linked-data is validated by using SparQL to retrieve data from linked data. SparQL is also used to query necessary data for data visualisation usage.