COMP6214 Open Data Innovation Coursework 1

Assignment:	Open Data Innovation	**Lecturer:**	Tope Omitola too1y15@soton.ac.uk	**Weighting:**	40%
Deadlines:	26.March.2021 @16:00	**Feedback:**	14.May.2021	**Effort:**	45 hours

Instructions

This coursework has four parts:

i. the first requires you to **clean** the provided dataset,
ii. the second requires you to **model** the dataset in an Open Data format (RDF) and **populate** the model using the data from the datasets
iii. the third requires you to **create a visualisation using a Linked Data Visualization** tool (we will look at such a tool in the module, but you can also use a different but appropriate one) and host that visualisation
iv. the fourth requires you to describe and submit a **report** of your work

You must use the provided dataset(s) for both cleaning and generating your visualisation. The dataset(s) should not be considered "authentic" data. It has been heavily modified in order to evaluate your ability to clean, manipulate and visualise such data. The data is provided in CSV and can be found via the course website.

Relevant Learning Outcomes

1) Identify innovation opportunities for open data.
2) Be able to apply appropriate validation, cleaning and transformation to use, reuse and combine a multitude of complex datasets.
3) Be able to model data sets in open data format (RDF) and populate these models with data from the datasets
4) Critically evaluate a large range of infographics and interaction techniques suitable for different tasks.

Marking Scheme

Criterion	**Description**	**Outcomes**	**Mark**
Cleaning and manipulation of dataset	The student has identified a number of errors or different types of errors in the dataset. The student has applied suitable techniques to fix errors and manipulate the dataset ready to be visualised.	2	8

Modelling the dataset & Populating the Model	The student has described how they modelled the concepts (and their attributes) of the data sets and has populated these models with real data (from data sets).	1,3	12
Visualisation I: Implementation & function	The implementation is functional and runs without errors. The visualisation is hosted on a webpage which opens without errors. Good use is made of an appropriate library for presenting dynamically loading visualisations.	1, 2, 4	10
Visualisation II: Interactivity and innovation	The visualisation presents multi-dimensional data that is interactive; i.e. it allows features such as filtering, selection, zooming, and multi-view capability to explore the dataset. The choice of visualisation is appropriate to the data and audience. The visualisation is innovative and useful; it provides value to the intended audience beyond that of the raw data or simple non-interactive graphs.	2, 4	10
Total Marks			**40**

Please Note:

The **visualisation** (and the website it is hosted upon) is aesthetically appealing, intuitive, easy to navigate, has a good user experience. The purpose function and instructions for use of the visualisation are well communicated.

Written work is free from grammatical errors, offers a high level of readability, clarity of expression and communication and good sentence/paragraph structure.

WHAT YOU MUST DO:

Part 1: Clean the dataset

The dataset depicts impacts of Covid-19 on some UK businesses in the month of April 2020. You are required to clean the data in worksheet 2 to worksheet 7 (i.e. Sample Size, Response Rates, Trading Status, Government Scheme, Government Scheme (2), and Government Scheme (3); and perform simple manipulations such as formatting, fixing errors etc. to prepare the dataset for creating your visualisation. You will be assessed on your ability to identify and handle a number of different types of errors in the dataset. These errors should be accounted

for through pre-processing (using tools such as Open Refine or using your own scripts or code). You must provide a written description of your data cleaning and manipulation methods.

There are several errors and error types in the dataset, and you should look for **at least 6 errors**. It is **not** necessary to find and fix all of the errors in the CSV file(s) to be awarded the full marks, provided you have spotted, reported and outlined solutions for **at least 6** and have provided solutions for fixing them.

Part 2: Model your dataset and represent them in RDF

Any RDF serialisation type is adequate [RDF/XML, JSON-LD, TURTLE, etc.]. Populate your RDF model with the dataset (examples are given in class, and will be uploaded on the course website). **The model must have a minimum of 6 classes with each class having a minimum of 6 predicates**, (you can have more than 6 classes and 6 predicates, if it makes your model clearer).

Part 3: Create and Host your visualisation

Creating your visualisation
You must build a visualisation of the dataset. Your visualisation should have suitable interactivity that allows for manipulation, filtering, and detailed analysis of the data.

You should aim to develop a multidimensional (greater than 2 dimensions) visualisation that enables rich exploration of the data. Note that “multidimensional” refers to the dimensions of the data, not the visualisation, i.e. expected to use the values from at least 3 columns from the provided dataset to create your visualisation (from one or more worksheets). The visualisation should be appropriate to the dataset and appropriate for the target audience or use case of your choosing.

Hosting your visualisation
You must create a simple website or web page to host your visualisation. Most of the marks relate to the data cleaning and the quality of the visualisation itself, so there is no need to produce a complex website. You can use publicly available templates when creating your website/webpage provided you reference the source.

Part 4 What You Should Submit

A pdf file consisting of these parts:

1. Section 1, to have the Title of the Course, your name and your student number. (**Note: No Abstract and No Content Page**)

2. Section 2: Title: Open Data Cleaning.

This section will be a description of your cleaning and manipulation of the dataset(s). The description will have:

a. The tool(s) you used for the data cleaning
b. A list of the error or error types you found in the dataset
c. For each error type: solutions or transformations you have applied to clean the dataset
d. How you validated the resulting cleaned-up file

3. Section 3: Open Data Modelling

This is where you will report on your modelling. It will consist of:

a. description of how you modelled your data
b. ontologies you chose and why you chose them

4. Section 4: This will have the following URLs:
    - a URL of where your visualisation is hosted. Do make sure the URL can be accessed from a public facing host for marking and for possible external examiner inspection
    - URL(s) of your generated Linked Data file (in json-ld, rdfxml, ttl, n3, or any other RDF serialization format), e.g. http://myRootUrl/mySchemaFile1, http://myRootUrl/myLinkedDataFile2 ...
    - URL of your ontology, e.g. http://myRootUrl/myOntology (if you have designed/developed your own ontology)

(Note: for your hosting, iSolutions can provide free hosting. If you require an iSolutions-provided free hosting, do let me know, and I will contact them).

Part 5 Submission

Submit one pdf file, consisting of the contents in Part 4, to the C-BASS handin system (http://handin.ecs.soton.ac.uk), by the submission deadline stated above. The standard ECS late penalties apply, as detailed in the regulations (para. 4.1 of http://www.calendar.soton.ac.uk/sectionXII/ecs-ug.html). They are 10% per working day that a piece of work is overdue, up to a maximum of 5 days, after which the mark becomes zero.