Working with data types and structures in Python and R_B208593

Administrative Information

1) School or Institute

- CMVM - Usher Institute

Data management plan for 'Working with data types and structures in Python and R (2021-2022)' course.

2) Name and Contact details of supervisor(s)

Dr Mairead Bermingham, mairead.bermingham@ed.ac.uk

3) Project start date

2022-06-06

4) Project end date

2022-07-04

Data Collection

5) Data Collection

The data will be collected from existing NHS England accident and emergency attendances and admissions records from large type 1 emergency departments.
The data will be collected with a data capture tool created in Jupyter notebook (Python).
The qualitative data for this project to be collected:

- period (the month that this activity relates to, stored as a date (1st of each month)
- org_code (the Organisation data service (ODS) code for the organisation. The ODS code is a unique code created by the Organisation data service within NHS Digital, and used to identify organisations across health and social care)

The quantitative data to be collected:

- attendances (the number of attendances at type 1 A&E departments)
- breaches (the number of attendances that breached the four-hour target)
- performance (([1 - breaches]/attendances) calculated for type 1 A&E departments)

Initial raw data and the processed data will be stored in CSV format, the scripts will be stored in the appropriate RStudio and Jupyter notebook format. The exported report, data management plan and R markdown reports will be stored in PDF format.

Documentation & Metadata

6) Documentation & Metadata

RStudio will be used to process the raw data and to create a data dictionary (metadata).
Jupyther notebook (Python) will be used to create and validate the data capture tool.
The data collection tool should be validated by splitting our raw dataset into training and testing data sets and the testing data will be used to validate the data capture tool.
For future reproducibility the R markdown document can be used as overview for the process and all scripts will be available for future reference.

Ethics & Legal Compliance

7) Ethics & Legal Compliance

As the data is comprised NHS patients data we must be adhered to GDPR and obtain consent from the end-user to process and share the data collected with the data capture tool. The data must be stored securely and only accessible for by authorized personnel (tutor on the course).

Storage and Back-Up

8) Where will your data be stored and backed-up during the project?

All information (raw data files, R scripts, Jupyter notebook scripts, data dictionary) will be stored in a GitHub repository.
Storage of the data will be on a GitHub repository in hierarchical filling system in separate folders:

- RawData (original raw data, data dictionary)
- Data (processed, collected data)
- Rscripts (R scripts)
- ipynbScripts (Jupyter notebook scripts)
- Outputs (data management plan, R markdown reports, final project report)

Selection and Preservation

9) Where will the data be stored long-term?

For long-term storage a recognised research data repository (e.g. DataShare or DataVault at the UoE) and external hard drives can be used.

10) Which data will be retained long-term?

Raw data, R scripts, Jupyter notebook scripts, data dictionaries

Data Sharing

11) Will the data produced from your project be made open?

- Yes: go to 12

12) How will you maximize data discoverability & access?

Data collected and scripts used for this project will be available for sharing on a secured GitHub repository which can be accessed only authorized university tutors.

Responsibilities & Resources

14) Who will be responsible for the research data management of this project?

I will be responsible for the research data management of this project.

15) Will you require any training or resources to properly manage your research data throughout this project?

I will require further training as I am a first-year MSc student in Data Science.