Working with data types and structures in Python and R

(word count: 1975)

University of Edinburgh
MSc Data Science for Health and Social Care
Digital technologies in health and social care

(2021-2022)

Exam number: **B208593**

Working with data Types in R and Python

1. Description of data

The aim of this project is to analyse the performance of large (Type 1) accident and emergency (A&E) departments in England regarding breaches of four-hour attendance over the years (Nuffieldtrust, 2022). Type 1 emergency departments provide consultant led 24 hour service with full resuscitation facilities and designated accommodation for the reception of accident and emergency patients (NHS, 2019).

The data used for this project are from a dataset from the NHSRdatasets package: NHS England accident and emergency attendances and admissions (*ae_attendances*) which contains all reported attendances, four-hour breaches and admissions for all A&E departments in England between 2016 and 2019 (Hutson et al., 2021). I planned to investigate the four-hour waiting time performance for Type 1 departments over time, so I selected the variables shown in Table 1 for the data capture tool to collect.

VARIABLE	DATA TYPE		DESCRIPTION
	in RSTUDIO	in PYTHON	
period	character, date	string, character	the month that this activity relates to, stored as a date (1st of each month)
org_code	character	string, character	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
attendances	integer	numeric, integer	the number of attendances
breaches	integer	numeric, integer	the number of attendances that breached the four hour target
performance	numeric	numeric, float	([1 - breaches]/attendances) calculated for type 1 A&E departments

Table 1. Variables selected for the project

I used online versions of RStudio and Jupyter notebook (Python) on the Notable platform to process the data so the data types of the variables had to be determined for both platforms so the correct coding process could be used (Table 1).

To set up the data capture tool in the Jupyter notebook, the sub-setted data (containing data for only Type 1 departments) were needed to be split into training and testing data sets. An index column was also added to the raw data so that the partitioned data sets can be linked the to the raw data.

The reason that I have chosen this concept is because Type 1 A&E departments are dealing with the most severe medical emergencies, therefore they require the most resources. In addition, if there is suboptimal performance at Type 1 A&E departments, it is more likely to cause significant - potentially life-threatening - harm to patients (Prentice, 2022).

After the preliminary analysis of the dataset I found that the performance of these departments to meet the four-hour target became worse over time, which indicates that more resources are required to improve this (Figure 1). The outcome of this concept might deliver important message to healthcare providers to plan accordingly.

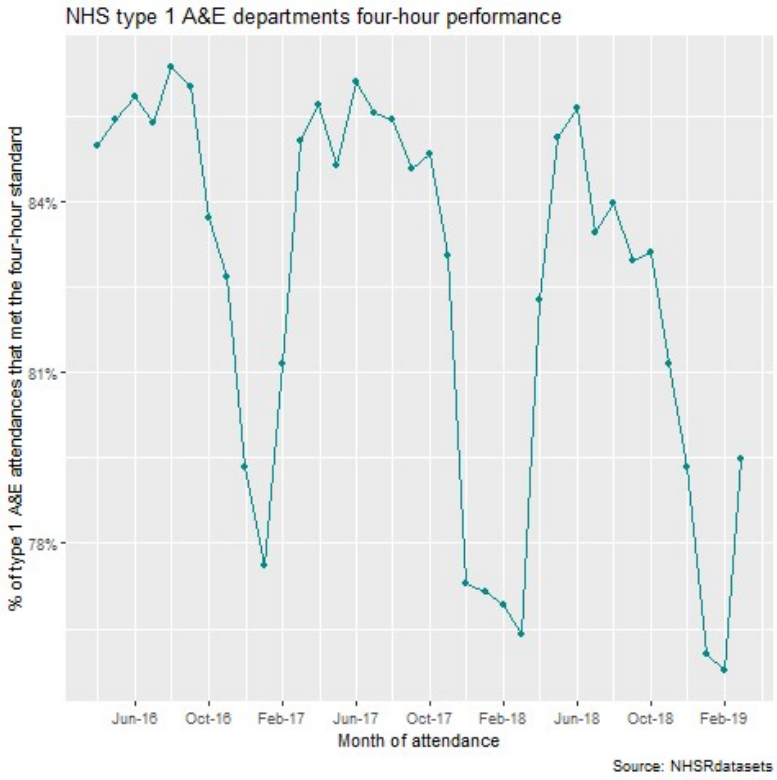

Figure 1. Four-hour waiting time performance of Type 1 departments between 2016-2019

2. Data capture tool

To set up the data capture tool (DCT) the Jupyter notebook web application was used which applies Python coding (Granger and Perez, 2021). Interactive Jupyter widgets (e.g., button, dropdown or textbox) were used to collect the data and pass on observations to the target audience (e.g., hospital management department) to convince them to share their data for further analysis (B208593, 2022).

The data type of the variables had to be determined so we can apply the appropriate widgets for them. Table 2 shows the variables and widgets used based on the data type of the variables.

VARIABLE	DATA TYPE	WIDGET
period	string, character	DatePicker
org_code	string, character	Selection
attendances	numeric, integer	Numeric (IntText)
breaches	numeric, integer	Numeric (IntText)
performance	numeric, float	Numeric (FloatText)
Consent	logical, boolean	Boolean (Checkbox)

Table 2. Data types of variables in Jupyter notebook (Python) and widgets used for the DCT

To validate the DCT the testing data set was used which was previously created using RStudio. With the use of embedded widgets (date picker, selection dropdown menu, numeric widgets) the data can be displayed in a user-friendly interface that helps the users (hospital management) to input the data easily.

Before collecting the data, according to General Data Protection Regulation (Tsohou et al., 2020), we needed to make sure that we get consent from the end-user to process and share the data we wish to collect with the DCT. For that purpose, a Boolean widget (checkbox) was used to allow the end-user to add consent (Figure 2).

At the end of the DCT a background was provided for our data collection (how we can persuade our target audience to share their data) and a reactive form was created for the end-user displaying only the widgets used for the data we wish to collect (Figure 2).

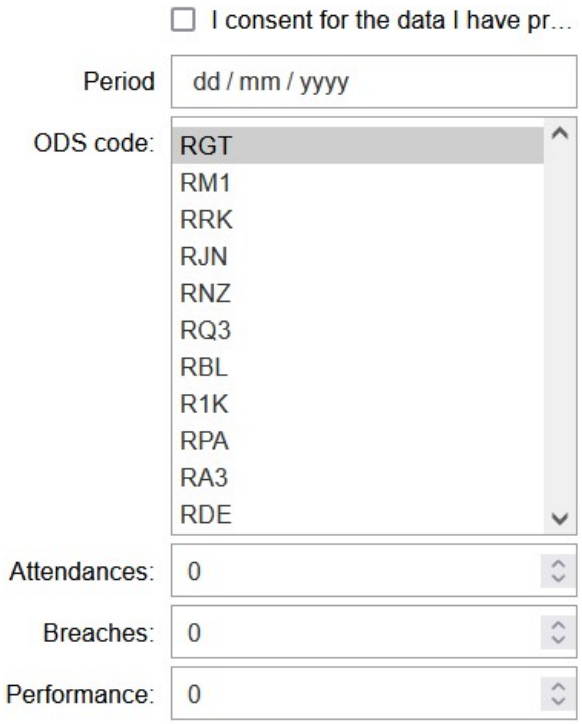

Figure 2. Final reactive form of the DCT showing the widgets for the data to be collected

3. Data management

A data management plan is essential during any research project to outline the management of the research data across the data lifecycle (Figure 3) (Stodden, 2020).

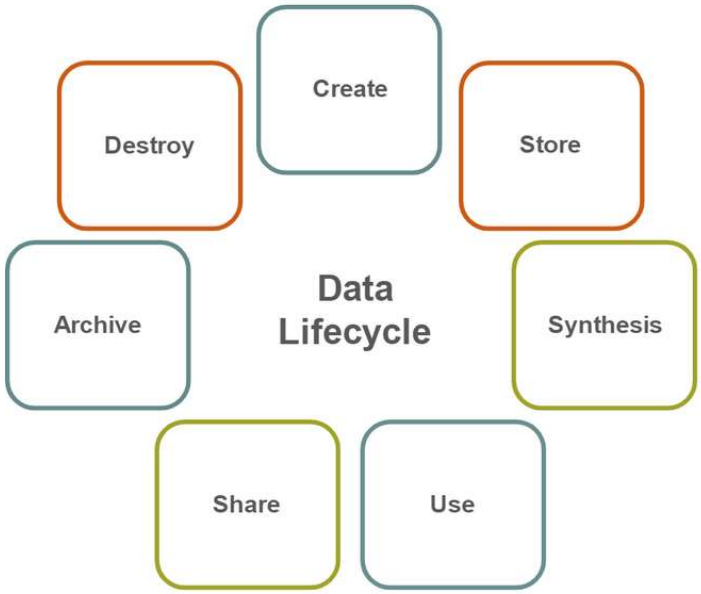

Figure 3. The data lifecycle of health and social care data

3. 1. Creation (data collection)

The dataset used are from the *NHSRdatasets* package which has been created to support skills development in the NHS-R community (Hutson et al., 2021). A free dataset from the package was chosen: NHS England accident and emergency attendances and admissions (*ae_attendances dataset)* containing 6 variables as shown in Table 3.

VARIABLE	DATA TYPE		DESCRIPTION
	in RSTUDIO	in PYTHON	
period	character, date	string, character	the month that this activity relates to, stored as a date (1st of each month)
org_code	character	string, character	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
type	character	string, character	the Department Type for this activity –1: Emergency departments are a consultant led 24 hour service with full resuscitation facilities –2: Consultant led mono specialty accident and emergency service –other: Other type of A&E/minor injury activity with designated accommodation for the reception of accident and emergency patients.
attendances	integer	numeric, integer	the number of attendances
breaches	integer	numeric, integer	the number of attendances that breached the four hour target

Table 3. Variables in the NHS England accident and emergency attendances and admissions data set

The format of raw data is comma-separated values (CSV) file. RScripts were used to process the raw data. The data for the project was collected with a DCT created in Jupyter notebook.

3. 2. Storage

A repository was created on the GitHub platform (B208593, 2022) and all relevant data and scripts are stored there in a hierarchical filling system in separate folders (Table 4).

FOLDERS	CONTENT
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

Table 4. Filing structure of the project data folder

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

3. 3. Synthesis

To create the data subset for the DCT RStudio was used to subset the raw data (*ae_attendances dataset)* with the chosen variables (Table 1) and to divide it into training and testing datasets. An index column was added to the raw data so the partitioned data sets can be linked to the raw data. The testing dataset was used to evaluate the data capture tool created in Jupyter notebook (Python). For the data collected with the DCT a data dictionary was generated in RStudio with description of the variables (metadata) for future reproducibility.

3. 4. Usage

The collected data will be used to study the monthly four-hour waiting time target performance at Type 1 Emergency departments at each organisation over the years. The outcome of this concept could deliver important message to healthcare providers regarding the different performances in different regions of England, and as a result they could plan accordingly with the necessary resources (staff, equipment, etc.).

3. 5. Sharing

Data collected and scripts used for this project will be available for sharing in a clear filing system (Table 4) on the secured GitHub repository described previously (B208593, 2022) which can be accessed by authorized university tutors.

3. 6. Archiving

Raw data, data dictionary and scripts for R and Python can be preserved for appropriate number of years in the university research data repository (e.g., DataShare or DataVault at the UoE).

3. 7. Destruction

Data will be kept until the end of the project and finally it will be destroyed according to UoE processes.

4. Coding

Reproducible data science research needs a straightforward workflow with repeatable and clear scripts that show a clear path from raw data to the final outputs (Wratten et al., 2021).

Git version control system was used to manage and track the changes made of the source codes online on Notable in RStudio and in Jupyter notebook. To locate and retrieve the processed data a hierarchical filing structure was set up in the RStudio so the data can be found in an easily identifiable folder. All files associated with the project were organized in properly labelled folders set up in RStudio (Figure 5).

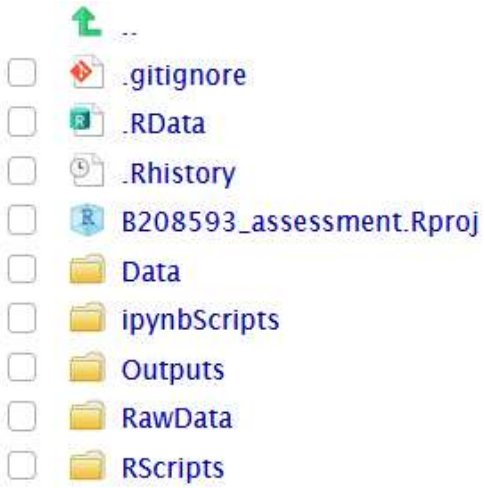

Figure 5. Filing structure of the working project folder

For this project several types of files had to be managed and placed in the appropriate folder: raw files, processed data files, codes and scripts, markdown and data management files, final descriptive report (Table 4).

All data and scripts were made available for markers on GitHub (B208593, 2022). This web-based platform allows management of Git repositories, version control, collaboration, sharing information and regular backup (Fylaktopoulos et al., 2016).

For reproducibility the documentation of changes made to the codes is also important (for others to be able to reproduce the same results. Therefore a detailed report was saved in an R Markdown document with the codes used for the data process in RStudio and Jupyter notebook using proper annotations.

All changes to the codes and scripts made in RStudio or Jupyter were synchronized with the data repository using version control. We used commit messages with detailed description of the changes to track the modifications made to the files in RStudio or in Jupyter notebook.

With detailed annotation and description of the codes and processes in the markdown document the whole process was made more understandable, repeatable and shareable.

5. Data storytelling

The DCT was set up for this project to gather the necessary information/data for the analysis of Type 1 emergency departments' four-hour waiting time target performance (Figure 2).

With the data collected we can compare the number of attendances and the four-hour performance in the different A&E departments across England.

Initial data analysis suggested that there is a worsening trend regarding the four-hour waiting of Type 1 A&E departments therefore further steps are required to make improvements (Figure 1).

6. Future work

This tool would be useful:

- to analyse the different performances across England and compare the different regions,

- to find out if there is a relationship between regional emergency departments,

- to identify the reason for the longer waiting times or common factors of the breaches,

- to allocate resources.

By analysing the collected data over time we can determine:

- if there is a pattern amongst different units and different regions,

- if there is a local need to increase resources to improve poor performance,

- if the additional resources materialized in sustainable better performance.

Reflective practice (400 words; 20%)

1. Data management process

During the first week of teaching in this course about data management I remembered the data lifecycle which was taught in the Introduction course in the first semester. Now I understand why this is so important during processing databased on the lectures and the feedback from tutors which helped me to improve my data management plan (DMP) and as time progressed, I became more aware of importance of this process.

2. Coding practice

When I started the first semester with the Intro course, I found it very useful that we have very well annotated R scripts which helped me a lot when I was learning the basic R programming. I was also using these scripts when I was doing R programming for other courses (e.g., Introduction to Statistics course) so I became more familiar with the environment and the annotations, and I learned how important the description of codes was for the reproducibility of the processes.

3. Response to feedback

Initial feedback for my discussion post on the board greatly helped me to further improve my narrative. Feedback from the markers was very useful to develop my data management plan by drawing the attention to several aspects which were missing from my DMP, and I realized how important to establish the basics for a data science project.

Continuous communication with tutors and peers has been essential for me not only because I received answers to my questions but by following the communication between others on the discussion board, I realized that we all faced similar problems. Initially I was only emailing with the course leader because I thought my questions might seem “silly” and didn’t want to post on the discussion board. However, she encouraged me to post my question on the board as well, so I realized that they were indeed relevant and meaningful queries.

4. Skills developed

Since the start of the course my programming skills in R have been gradually improving, not just learning from this particular course but from previous courses (e.g., Introduction to Statistics) as well. I have learned how to use coding for data visualisation and how to use RMarkdown for creating a document for good coding practice.

My communication skills are also improved as during this course I participated more on the discussion boards asking questions from tutors and interacting with other students. I learned a lot from others asking questions on the discussion board and I became more confident to ask myself. Other tasks for previous courses (e.g., Digital technologies in health and social care) also enhanced my communication and presentation skills as I had to make a recorded PowerPoint presentation which made me more comfortable speaking.

I also gained new insights into programming when using Python for the DCT. Although it was a new area for me, as the time progressed, I found it a very useful being all coding and the output displayed in one.

REFERENCES

B208593 2022. GitHub repository.

FYLAKTOPOULOS, G., GOUMAS, G., SKOLARIKIS, M., SOTIROPOULOS, A. & MAGLOGIANNIS, I. 2016. An overview of platforms for cloud based development. *Springerplus,* 5**,** 38.

GRANGER, B. E. & PEREZ, F. 2021. Jupyter: Thinking and Storytelling With Code and Data. *Computing in Science & Engineering,* 23**,** 7-14.

HUTSON, G., JEMMETT, T., MAINEY, C. & TURNER, Z. 2021. NHS and Healthcare-Related Data for Education and Training.

NHS 2019. A&E Attendances and Emergency Admissions Monthly Return Definitions.

NUFFIELDTRUST. 2022. *A&E waiting times* [Online]. Available: https://www.nuffieldtrust.org.uk/resource/a-e-waiting-times#background [Accessed].

PRENTICE, D. 2022. A lay perspective and commentary on the association between delays to patient admission from the emergency department and all-cause 30-day mortality. *Emerg Med J,* 39**,** 166-167.

STODDEN, V. 2020. The data science life cycle. *Communications of the ACM,* 63**,** 58-66.

TSOHOU, A., MAGKOS, E., MOURATIDIS, H., CHRYSOLORAS, G., PIRAS, L., PAVLIDIS, M., DEBUSSCHE, J., ROTOLONI, M. & GALLEGO-NICASIO CRESPO, B. 2020. Privacy, security, legal and technology acceptance elicited and consolidated requirements for a GDPR compliance platform. *Information & Computer Security,* 28**,** 531-553.

WRATTEN, L., WILM, A. & GOKE, J. 2021. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. *Nat Methods,* 18**,** 1161-1168.