Working with Data types and structures in Python and R

Data Collection

What data will you collect or create?

The NHS accident and emergency attendances and admissions data set which contains 12,765 rows and six columns on attendance for accidents and emergency departments across England for the years 2016/17 through 2018/19 (Apr-Mar) will be used. This is a freely accessible online dataset created by NHS for training purposes.

The data set contains one date variable (period), two factor variables (organization code and type of department), and three numeric (double precision) variables (attendance, breaches, and admissions).

I would like to investigate the monthly trend of attendances that result to admissions in emergency departments across the whole of England to determine periods of high and low burden. This would be crucial for planning and deploying resources.

From the data set, I will select the following variables: period, type, organization code and admissions. I will filter the dataset to select only type 1 which is emergency department. After selection and filtering, the new dataset will consist of 12,765 rows with four columns.

Because of the large size of the dataset, the CSV files format will be used to increase storage capacity as well as speed up the processing time.

How will the data be collected or created?

The data will be collected and analyze in R and Python. As a programme for statistical analysis and computing, R enhances the ability for reproducibility of results and can ensure version control using GitHub repository. Python is also a very popular and useful programme for data collection. Features such as widgets in python can enhance data collection efforts. A data capture tool using interactive widgets will be developed in python and used to collect the datase of interest.

To ensure findability, the datasets will be structured in separate folders as follows: RawData (for the original and collected datasets), Rscripts folder, Python scripts folder, Figures (for plots and graphs), Outputs for RmarkedDown files. Simple naming conventions that clearly described the data will be used throughout the developments of scripts in R and Python to enable ease of understanding, use, and reference.

GitHub will be used for version control and to ensure collaboration as well as opportunity for reproducibility of the results.

Documentation and Metadata

What documentation and metadata will accompany the data?

A data dictionary describing each of the variables in the data collection tool will be developed to enable clear documentation and consistent interpretation of the dataset. The data dictionary will be created in R and appended to the original dataset in the Raw Data folder for ease of access and reference. Metadata describing the variables in the dataset as well as the authors, timing of edit, and other relevant descriptions will be created and linked to the data dictionary.

Ethics and Legal Compliance

How will you manage any ethical issues?

Consent will be sought from the end-user for collecting, analyzing, and sharing of the data. A consent variable of boolean variable type will be collected for each user using the checkbox widget in python. No personal identifying information will be collected.

How will you manage copyright and Intellectual Property Rights (IPR) issues?

The data is owned by NHS England and has been freely available for practice. However, references will be provided to NHS England as the owner of the dataset.

Storage and Backup

How will the data be stored and backed up during the research?

All datasets will be stored online using GitHub repositories for purpose of backup and storage capacity. Changes on the scripts or datasets will be push to the GitHub repository.

How will you manage access and security?

The link to the data repository will be shared with collaborators to ensure access.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

The original dataset will be retained in the RawData folder to guarantee preservation and retrievability. None of the datasets will be destroyed.

What is the long-term preservation plan for the dataset?

The data will be held in GitHub repository.

Data Sharing

How will you share the data?

The data will be shared via the GitHub repository and freely available to everyone with interest.

Are any restrictions on data sharing required?

No restrictions will be placed on data sharing except where expressedly indicated by the end user.

Responsibilities and Resources

Who will be responsible for data management?

All managment activities pertaining to the date will be done by me.

What resources will you require to deliver your plan?

A critical resource would be sustainable access to internet.