Working with Data types and structures in Python and R
Assessment#3 Descriptive Report B210635

Overview

Emergency departments are critical in life-saving processes; thus, health decision-makers need to rationally allocate resources (human, medical supplies, and logistics) to ensure efficiency and economy. This report describes the monthly burden of admissions in emergency departments across England using the National Health Service (NHS) accident and emergency attendances and admissions datasets, one of five freely available datasets obtained from the NHSR datasets package. Exploring the monthly trend in the burden of emergency admissions is a useful way of demonstrating evidence-based approach to resource allocation during varying periods of the year.

In this report, the dataset is described in more detail, including a data capture tool developed along with the data management plan and coding practices. Insights from the data are provided, along with critical reflections garnered from feedback and thought processes.

NHS England accident and emergency attendances and admissions dataset

Exploration of the structure of the NHS accident and emergency attendances and admissions dataset shows 12,765 rows and six columns. The columns correspond to the following variable types: one date variable (period), two factor variables (org_code and type), and three numeric variables (attendances, breaches, and admissions).

The **'period'** variable is stored as date and refers to the month in which the activity (attendance, breaches, and admission) took place.

The **'org_code'** variable is stored as factor and refers to the Organisation data service (ODS) unique code created within NHS Digital, and used to identify organisations across health and social care.

The **'type'** variable is stored as factor and refers to the department type with following categories: 1: Emergency departments which are consultant led 24-hour service with full resuscitation facilities and designated accommodation for the reception of accident and emergency patients; 2: Consultant led mono specialty accident and emergency service (e.g. ophthalmology, dental) with designated accommodation for the reception of patients; Other: Other type of A&E/minor injury activity with designated accommodation for the reception of accident and emergency patients.

The **attendances** variable is stored as numeric and refers to the number of attendances for this department type at this organisation for this month.

The **breaches** variable is stored as numeric and refers to the number of attendances that breached the four-hour target.

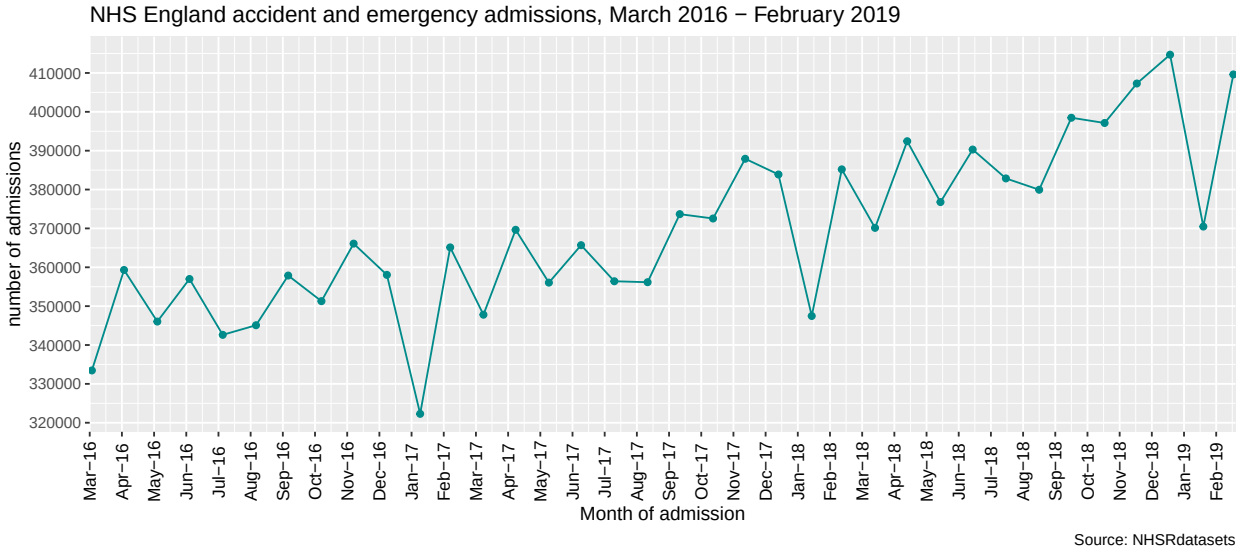
The **admissions** variable refers to the number of attendances that resulted in an admission to the hospital (Chris Mainey, 2021).

This report focuses on the monthly burden of emergency admissions across all organizations in England, as such, the following variables were selected from the original dataset: org_code, period, and admissions for analysis. To do this, an index column was first added to the original dataset before selecting the variables

and filtering the emergency department from the 'type' variable. Indexing the original dataset is a crucial first step in the development of a data capture tool.

The resulting dataset contains 4,932 rows with four columns corresponding to five variables as follows: index stored as integer, org_code stored as factor, period stored as date, and admissions stored as numeric.

Analysis of the data showed overall increasing trend despite monthly fluctuations in the number of admissions at emergency departments. Noticeable dips were seen yearly during the month of January.



The data capture tool

In order to develop the data capture tool, the provisional dataset was separated into training and testing datasets. Most of the dataset (n=4,920) was assigned to the training dataset while 12 records were assigned to the test dataset to reduce the time spent for data capture. One of the 12 rows from the test dataset was set aside for use by the markers to test and evaluate the data-capture tool.

The data capture tool was developed in Python using interactive jupyter widgets to collect the data. Widget is an intuitive feature consisting of graphical user interface elements, such as a button, dropdown menus, or textboxes to collect or input user data. The test dataset used for data collection consisted of eleven rows and four columns or variables with data types as follows: index (integer), org_code (string), period (date), and admission (integer). Because consent is crucial to ensure data protection compliance in line with data regulations standards, a boolean variable 'consent' was added to the data capture tool using a Boolean widget (checkbox widget) with values as 'True' or 'False'. True corresponds to consent provided by the end-user to analyze and share the collect dataset. A datepicker widget was used to collect data for the period variable as date. The org_code variable was displayed as list and selection widgets were used to select the org_code value from the list. Because the admissions variable is an integer, the *'IntText widget'* was used to input the value. Eleven iterations were performed for each variable to collect the data captured by the Juypter widgets to an empty data frame.

Data dictionary

A data dictionary of the data collected using the data capture tool was developed to provide detailed information about the variables and features of the collected dataset including its metadata. The collected dataset consisted of eleven rows and five columns with each column corresponding to variables of a specific data type. The dataset consisted of two quantitative variables (index and admissions) and three fixed values variables (org-code, period, and consent). The following metadata was included to describe attributes of the dataset: the main string indicating that the dataset describes NHS England accident and emergency

(A&E) admissions across emergency departments in England, the data dictionary with a description of each of the variables in the dataset, the total number of columns (5) and rows (11) in the dataset, the author's identification, and the last time it was edited.

Data management practices

The datasets were 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 was developed in python and used to collect the dataset. Because of the large size of the dataset, the CSV files format was used to speed up the processing time. To ensure findability, the datasets was 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 was used throughout the developments of scripts in R and Python to enable ease of understanding, use, and reference. GitHub was used for version control. A data dictionary describing each of the variables in the data collection tool was developed to enable clear documentation and consistent interpretation of the dataset. The data dictionary was 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 author, timing of edit, and other relevant descriptions was created and linked to the data dictionary. Consent was sought from the end-user for collecting, analyzing, and sharing of the data. A consent variable of boolean variable type was collected for each user using the checkbox widget in python. No personal identifying information was collected. 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. All datasets were stored online using GitHub repositories for purpose of backup and storage capacity. Changes on the scripts or datasets was push to the GitHub repository for version control monitoring. To ensure accessibility and file sharing with collaborators or course supervisors, the link to the data repository was included in the scripts and MarkedDown files. The original dataset will be retained in the RawData folder to guarantee preservation and retrievability. None of the datasets will be destroyed. The data will be held in GitHub repository and freely available to everyone with interest. No restrictions will be placed on data sharing except where expressly indicated by the end user. All management activities pertaining to the data will be done by me. Internet access would be a critical resource to ensure sustainability.

Good coding practices

The source code panel was used for the development of all codes. All codes developed or used in both R and Python were carefully annotated to ensure readibility and understanding by other users. Annotations were placed either on top of or beside the code to provide explanation. An RMarkedDown file was used to knit the scripts to a pdf or word document output. Lines of coded pertaining to different aspects of the work were separated as code chunks. A hashtag followed by text enquoted in double asterisk were used to set main headings apart. Include = FALSE in the heading of the code chunk was used to prevent printing of outputs.

Effectively communicating insights to non-technical audience

The above data capture and exploration is a typical example of how simplified by powerful messaging can be conveyed to non-technical audience to influence decision making. The trend in the monthly attendance that results in admissions at emergency department is increasing but clear yearly dips in January is pronounced. This can help decision-makers or managers planned the allocation of their scarce resources. Provision of complementary data on resource allocation during these period can provide further insight on the relationship betweeen the two and provide more convincing evidence for decision making.

Best practices to improve next iteration of data capture tool

Best practices to consider are the factors affecting the reproducibility of the data and these are mainly anchored on management of the data in line with the FAIR principles (Wilkinson et al. 2016; GO FAIR, 2022). The organization of the data into folders, use of simplified naming schemes, and committing changes to GitHub repository can be considered as best practices that ensure the datasets are easy to find, accessible, interoperable, and reusable. One critical aspect linked to data integrity is consent for collecting, using, and sharing the data. Consent was employed in the collection of the data as part of compliance to ethical standards. Also, the use of GitHub for version control system enabled collaborators or other users of the data to identify changes to the content.

Reflection

Introduction to data management plan has been truly helpful as I am beginning to see data management practices through this lens. I am so grateful that I have been exposed to the conceptual approach for developing a data management plan. The DMP online tool helped me to follow a systematic approach for developing a data management plan, albeit the need to strenghten my skills in this area. There were two important feedbacks received. One of them was a query on my data retention strategy and another was a caution to be a little more careful when handling sensitive data on github. I found the recommendation to use SSH key on Github to handle sensitive data quite useful, as this was totally unknown to me. Concerning the data retention strategy, I had to reflect a little more but thought that keeping the datasets in a secure environment online would guide against loss.

The learning in this course has been a little steep for me as new topics were explored, but I have followed through. I look forward to deepening my skills with more practice.