Descriptive Report

B209510

7/2/2022

# Narrative

The data derives from an artificially created dataset the LOS\_model with the purpose of studying statistical modelling (<https://cran.r-project.org/web/packages/NHSRdatasets/vignettes/LOS_model.html>) and it’s obtained via the NHSRdatasets package in R programming language. It is a data frame contains 300 observations from ten fictional hospitals and it has five variables: ID, age and LOS (Length of Hospital Stay) are continuous numerical variables and organization and death are categorical variables. Table 1 displays the first ten observations of the LOS\_model dataset. An initial observation of the dataset will be carried out to evaluate its characteristics and to assess if there is any missing data. For the scope of this report the variables of interest are numerical integer Age and LOS and the categorical variable Death this will allow exploration of Age and LOS regarding outcome of hospitalization. A data capture tool was developed to retain this information using Python programming language and available at GIT.COM. The collected data will be stored as a CSV file and as a data frame. An inspection of a possible link between the LOS or Age and outcome of hospitalization is the main purpose for this data to be collected however observing if Age is related to LOS will also be an intention. Getting information or even being able to predict outcome of a hospitalization or LOS provides powerful knowledge for policy makers and planning.

Table 1 : First ten observations of LOS model dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | Organisation | Age | LOS | Death |
| 1 | Trust1 | 55 | 2 | 0 |
| 2 | Trust2 | 27 | 1 | 0 |
| 3 | Trust3 | 93 | 12 | 0 |
| 4 | Trust4 | 45 | 3 | 1 |
| 5 | Trust5 | 70 | 11 | 0 |
| 6 | Trust6 | 60 | 7 | 0 |
| 7 | Trust7 | 25 | 4 | 0 |
| 8 | Trust8 | 48 | 4 | 0 |
| 9 | Trust9 | 51 | 7 | 1 |
| 10 | Trust10 | 81 | 1 | 0 |

The data capture tool provides a user interface to manually input the required information using Widgets as interactive controls from Python programming language for the integer numerical variables Age and LOS a IntText widget was used and for the categorical variable outcome a RadioButton widget was used. Index variable is inserted via a user input query and consent via CheckBox widget. The platform includes fields to be filled manually corresponding to the variables of interest and as a quality assurance it should not allow erroneous type of data to be entered such as characters when the required input is numeric this ensures optimal data entry. The data collected by this tool is displayed in table 1.

Table 2: First ten collected observations

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Index | Lenght of stay | Age | Outcome | Consent |
| 56 | 1 | 11 | Discharge | TRUE |
| 72 | 6 | 79 | Discharge | TRUE |
| 76 | 7 | 66 | Discharge | TRUE |
| 85 | 12 | 90 | Discharge | TRUE |
| 147 | 3 | 38 | Discharge | TRUE |
| 162 | 2 | 17 | Discharge | TRUE |
| 202 | 6 | 69 | Discharge | TRUE |
| 204 | 2 | 27 | Discharge | TRUE |
| 252 | 4 | 90 | Death | TRUE |
| 275 | 6 | 44 | Discharge | TRUE |

There are three variables of interest for this project as aforementioned (LOS, Age and Outcome) and two further variables were created the Index and Consent. The index will allow identification of an observation in the primary dataset thus enabling the investigator to reach the primary information if required. Despite the data is artificially created a consent checkbox was developed to ensure every individual has given consent for its data to be captured and processed this field may not be entirely necessary as data is artificially created but for the purposes of this project it was included in as it could be that in future this tool can be used to acquire non-artificially created data and thus consent is paramount. As artificially created data for training and teaching purposes there are no significant legal or ethical requirements to obey. Appendix 1 shows the user interface for data acquisition. The data for this project is obtained from the LOS Model (<https://cran.r-project.org/web/packages/NHSRdatasets/vignettes/LOS_model.html>) via NHSRdataset package in R programming language and it will be acquired using an interactive user interface developed in python programming language (….). This interface allows the data to be captured and retained for this project via user manual entries in the developed interface. The data captured will be stored in University of Edinburgh servers and it will be generated via Noteable platform and deposited in the git hub repository with a clear folder structure and file naming allowing users to easily access and use this data. A data dictionary and metadata will be available describing the data components in a detailed manner. An analysis of the data collected will be carried out following the necessary steps for this process such as data integration and manipulation. A visualization of the data acquired will be perform using the required functions in R and Python to confirm there are no data entry errors such as an impossible age. The git hub repository contains all components of this project with a detailed documentation with the description of the variables and all syntax for this analysis in R and Python programming languages making the data findable and accessible. This platform allows sharing the data with other users and keep track of any changes made keeping a version control for the documents. After completion of this project the data will be archived and destroyed in line with the University of Edinburgh data retention and destruction policy. (<https://noteable.edina.ac.uk/data-retention/>) Establishing a data management plan for the captured data will create a sustainable and organized data process and the data capture tool should provide the ability to incorporate newer data for analysis. The data management plan provides the framework and guidance throughout all the data life cycle since its creation to destruction. The initial unprocessed data will be stored in a raw data folder and preferentially as read – only to ensure there is always availability of it and processed data will be stored in a data folder and the other relevant elements of the data analysis process will be stored independently to guarantee good organization and security.

The data capture tool was developed in Python programming language and the data inspection and manipulation was performed in R and Python programming language. Categorical data had to be converted into factors in R programming language to ease future use of this variables for downstream analysis and the data was split in test and train datasets to allow validation of models created from the acquired data. An index was created to the main dataset letting the investigator reach the main observation if needed. These processes were well documented in the R script document and logically structured in such a way that others could easily understand this process thus promoting reproducibility and reusability. Several annotations in the syntax were made and meaningful naming of objects and variables in R and Python scripts were used to ensure readable code (<https://thedataist.com/ten-good-coding-practices-for-data-scientists/>) . The git hub repository contains all the necessary scripts to reproduce the data capture tool and data analysis/manipulation with a clear filling path where the results of this project can be stored, accessed, and run efficiently without any error. The code is structured to avoid performing several actions in one chunk or cell and collapses different operations into different cells or chunks as I found this methodology easier to find an error and to perform corrections. Using clear and descriptive variable and object naming in coding reduces the possibility of over annotations as this can be self-documenting code and others and my future self can understand what that represents and includes. The git hub platform allows version control of all files in its structure by keeping a track of every single modification noted in its files and folders permitting the main user and collaborators to be aware of any changes made. The modifications should be accompanied by a short and clear description of the alterations improving the traceability and shareability of the code and documents.

Collection of hospital activity data with information regarding length of admission, age of patient and outcome of hospitalization can provide the foundation for data-driven findings to improve healthcare delivery and planning by gathering insights to answer relevant questions such as the extent that older people stay longer in hospital when compared with someone young or how likely is an outcome of an hospitalization based on age or length of time spend in hospital. With an aging population and staff constrains observed in the healthcare setting this information can be very relevant for policymakers.

This project can be further developed by creating a web-based platform for the user to input the data and by introducing prompt error messages when the wrong time of data is entered for example characters instead of numbers. A confirmation prompt for the value in the numerical variables would add value to this tool by ensuring that no typo errors happen thus ensuring good data quality. A downstream analysis with statistical methods giving information regarding correlation, hypothesis testing and even modelling can improve the next iteration of this tool.

# Reflective practice

Developing a data management plan in the early stages of this project enhanced my learning about data organization throughout the data life cycle and its methodology for handling the data since its creation until the destruction. It allowed me to plan the data journey and I gained knowledge of all the processes involved in each of the stages of the data lifecycle. Based on the feedback from my markers I was able to reflect on the target audience and how to communicate with them the development of the data capture tool and the resulting outputs. After reflection all the stages of the data lifecycle are equally important when elaborating a data management plan and I should not focus on some of the steps but see the plan as consolidative document integrating all the components since the creation of the data until its destruction. I believe coding is a skill that can always be improved with this and the previous courses allowing me to continuously reflect on my coding performance and adopt several improvements such as descriptive annotations, meaningful variables and objects, structuring the code efficiently and try to make the code as simple as possible to be able to share with others proficiently and even for my own understanding in the future. I developed the analysis of my own coding practice by making sure it runs appropriately with the desired behaviour and then trying to make it simple and better structured or finding better methods to achieve the same result. The feedback and feedforward received on my discussion post guided me in improving my data management plan by highlighting some points that I did not focus on and would deserve more attention such as data creation, storage and destruction in the data lifecycle. This course allowed me to develop further my coding skills by practicing and acquiring new knowledge of different tools and methods such as widgets in python and expanding my understanding of R markdown in R studio. I feel this course gave me an opportunity to improve my organization and planning skills by implementing a structured filling structure and awareness of all stages of the data management in this project. The previous learning from introduction to software development in Python allowed me to add an user input query to retain the index number.