MyFirstMarkdownDocument

B203349

16/06/2022

The link to my git hub repository is as follows: Link to my git hub repository

Data Dictionary for test data

Loading the required packages and reading the data

Loading packages

```
library(dataMeta)
library (tidyverse)
```

```
## Warning: replacing previous import 'lifecycle::last_warnings' by
## 'rlang::last_warnings' when loading 'pillar'
```

```
## Warning: replacing previous import 'lifecycle::last_warnings' by
## 'rlang::last_warnings' when loading 'hms'
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
```

```
## Warning: package 'tibble' was built under R version 4.1.2
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(here)
```

```
## here() starts at /Users/matt/Desktop/Dropbox/Home/College/Edinburgh - MSc Data Science/Working with c
```

Reading the collected data

`read_csv()` function from the *readr* package is used to read my collected data from the Raw data folder.

```
CollectedData=read_csv(here("RawData", "CollectedDataFinal.csv"))
```

```
## Rows: 11 Columns: 8-- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): org_code, type
## dbl  (4): index, attendances, breaches, admissions
## lgl  (1): consent
## date (1): period
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Building a linker data frame

Variable descriptions

String vectors for the different variable descriptions

```
variable_description <- c("The index column that allows us to link the data collected to the original ae
"The month that this activity relates to, stored as a date (1st of each month).",
"The Organisation data service (ODS) code for the organisation. If you want to know the organisation ass
"The department type for this activity.",
"The number of attendances for this department type at this organisation for this month.",
"The number of attendances that breached the four-hour target.",
"The number of attendances that resulted in an admission to the hospital.",
"The consent from the end-user to process and share the data collected with the data capture tool.")
print(variable_description)
```

```
## [1] "The index column that allows us to link the data collected to the original ae_attendances data
## [2] "The month that this activity relates to, stored as a date (1st of each month)."
## [3] "The Organisation data service (ODS) code for the organisation. If you want to know the organisa
## [4] "The department type for this activity."
## [5] "The number of attendances for this department type at this organisation for this month."
## [6] "The number of attendances that breached the four-hour target."
## [7] "The number of attendances that resulted in an admission to the hospital."
## [8] "The consent from the end-user to process and share the data collected with the data capture tool
```

Variable types

A string vector is created representing the different variable types. It is a vector of integers with values 0 or 1. 0 is used for a variable with quantitative values (measured values) variables and 1 for fixed values (allowable values or codes) variables.

`glimpse()` function from *tibble* package is used to view the variable types in the CollectedData data frame

```
glimpse(CollectedData)
```

```
## Rows: 11
## Columns: 8
## $ index       <dbl> 1155, 2059, 3468, 4153, 4820, 7243, 8057, 8957, 10214, 103~
## $ period      <date> 2016-12-01, 2016-10-01, 2016-05-01, 2018-03-01, 2018-02-0~
## $ org_code    <chr> "C82010", "RDZ", "RVR", "RQM", "R1F", "RE9", "RQM", "RNL",~
```

```
## $ type        <chr> "other", "1", "2", "other", "other", "1", "1", "1", "other~
## $ attendances <dbl> 200, 6452, 417, 9376, 245, 5170, 15957, 7258, 3197, 2033, ~
## $ breaches    <dbl> 0, 360, 0, 112, 0, 235, 1309, 1374, 0, 8, 1
## $ admissions  <dbl> 0, 1814, 6, 0, 0, 1269, 3375, 1947, 0, 105, 0
## $ consent     <lgl> TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE~
```

This indicated that there are four quantitative values (measured values) variables and four fixed values (allowable values or codes) variables.

```
variable_type <- c(0, 1, 1, 1, 0, 0, 0, 1)
print(variable_type)
```

```
## [1] 0 1 1 1 0 0 0 1
```

Building the linker

`build_linker()` function from the *dataMeta* package is used to construct an intermediary (linker) data frame between the CollectedData and the data dictionary.

```
linker<-build_linker(CollectedData, variable_description, variable_type)
print(linker)
```

```
##      var_name
## 1       index
## 2      period
## 3    org_code
## 4        type
## 5 attendances
## 6    breaches
## 7  admissions
## 8     consent
##
## 1
## 2
## 3 The Organisation data service (ODS) code for the organisation. If you want to know the organisatio
## 4
## 5
## 6
## 7
## 8
##   var_type
## 1        0
## 2        1
## 3        1
## 4        1
## 5        0
## 6        0
## 7        0
## 8        1
```

Data dictionary

the `build_dict()` function from the *dataMeta* package is used to construct a data dictionary for a CollectedData data frame with the aid of the linker data frame.

```
dictionary <- build_dict(my.data = CollectedData, linker = linker)
```

```
## Enter description for variable 'admissions' and option '0 to 3375':
## Enter description for variable 'attendances' and option '200 to 15957':
## Enter description for variable 'breaches' and option '0 to 1374':
## Enter description for variable 'consent' and option 'TRUE':
## Enter description for variable 'index' and option '1155 to 11767':
## Enter description for variable 'org_code' and option 'C82010':
## Enter description for variable 'org_code' and option 'RDZ':
## Enter description for variable 'org_code' and option 'RVR':
## Enter description for variable 'org_code' and option 'RQM':
## Enter description for variable 'org_code' and option 'R1F':
## Enter description for variable 'org_code' and option 'RE9':
## Enter description for variable 'org_code' and option 'RNL':
## Enter description for variable 'org_code' and option 'RJ1':
## Enter description for variable 'org_code' and option 'RKB':
## Enter description for variable 'org_code' and option 'NLO12':
## Enter description for variable 'period' and option '17136':
## Enter description for variable 'period' and option '17075':
## Enter description for variable 'period' and option '16922':
## Enter description for variable 'period' and option '17591':
## Enter description for variable 'period' and option '17563':
## Enter description for variable 'period' and option '17348':
## Enter description for variable 'period' and option '17257':
## Enter description for variable 'period' and option '17928':
## Enter description for variable 'period' and option '17805':
## Enter description for variable 'period' and option '17683':
## Enter description for variable 'type' and option 'other':
## Enter description for variable 'type' and option '1':
## Enter description for variable 'type' and option '2':
```

```
glimpse(dictionary)
```

```
## Rows: 28
## Columns: 4
## $ `variable name`        <chr> "admissions", "attendances", "breaches", "conse~
## $ `variable description` <chr> "The number of attendances that resulted in an ~
## $ `variable options`     <chr> "0 to 3375", "200 to 15957", "0 to 1374", "TRUE~
## $ notes                  <chr> "", "", "", "", "", "", "", "", "", "", "", "",~
```

```
dictionary[6,4]<-"C82010: Prescribing Cost Centre - OAKHAM MEDICAL PRACTICE"
dictionary[7,4]<-"RDZ: NHS Trust - THE ROYAL BOURNEMOUTH AND CHRISTCHURCH HOSPITALS NHS FOUNDATION TRUST
dictionary[8,4]<-"RVR: NHS Trust - EPSOM AND ST HELIER UNIVERSITY HOSPITALS NHS TRUST"
dictionary[9,4]<-"RQM: NHS Trust - CHELSEA AND WESTMINSTER HOSPITAL NHS FOUNDATION TRUST"
dictionary[10,4]<-"R1F: NHS Trust - ISLE OF WIGHT NHS TRUST"
dictionary[11,4]<-"RE9: NHS Trust - SOUTH TYNESIDE NHS FOUNDATION TRUST"
dictionary[12,4]<-"RNL: NHS Trust - NORTH CUMBRIA UNIVERSITY HOSPITALS NHS TRUST"
```

```
dictionary[13,4]<-"RJ1: NHS Trust - GUY'S AND ST THOMAS' NHS FOUNDATION TRUST"
dictionary[14,4]<-"RKB: NHS Trust - UNIVERSITY HOSPITALS COVENTRY AND WARWICKSHIRE NHS TRUST"
dictionary[15,4]<-"NLO12: Independent Sector H/c Provider Site - OAKHAM URGENT CARE CENTRE"
dictionary[26,4] <-"other: Other types of A&E/minor injury activity with designated accommodation for tl
dictionary[27,4]<- "1: Emergency departments are a consultant-led 24-hour service with full resuscitati
dictionary[28,4] <- "2: Consultant-led mono speciality accident and emergency service (e.g. ophthalmolo
```

Save the data dictionary for CollectedData to the 'RawData' folder

```
glimpse(dictionary)
```

```
## Rows: 28
## Columns: 4
## $ `variable name`        <chr> "admissions", "attendances", "breaches", "conse~
## $ `variable description` <chr> "The number of attendances that resulted in an ~
## $ `variable options`     <chr> "0 to 3375", "200 to 15957", "0 to 1374", "TRUE~
## $ notes                  <chr> "", "", "", "", "", "C82010: Prescribing Cost C~
```

```
write_csv(dictionary, here("RawData", "CollectedData_DataDictionary.csv"))
```

Appending the data dictionary to data

Incorporate attributes as metadata to the CollectedData as metadata using the 'incorporate_attr()' function from the *dataMeta* package. This requires the CollectedData, dictionary, and main_string as inputs (main_string is a character string describing the CollectedData data frame)

Create main_string for attributes

```
main_string <- "This data describes the NHS England accident and emergency (A&E) attendances and breache
main_string
```

```
## [1] "This data describes the NHS England accident and emergency (A&E) attendances and breaches of fou
```

Incorporate attributes as metadata

```
complete_CollectedData <- incorporate_attr(my.data = CollectedData, data.dictionary = dictionary,
main_string = main_string)
attributes(complete_CollectedData)$author[1]<-"B203349"
complete_CollectedData
```

```
## # A tibble: 11 x 8
##    index period     org_code type  attendances breaches admissions consent
##  * <dbl> <date>     <chr>    <chr>       <dbl>    <dbl>      <dbl> <lgl>
##  1  1155 2016-12-01 C82010   other         200        0          0 TRUE
##  2  2059 2016-10-01 RDZ      1            6452      360       1814 TRUE
```

```
##  3  3468 2016-05-01 RVR      2             417        0          6 TRUE
##  4  4153 2018-03-01 RQM      other        9376      112          0 TRUE
##  5  4820 2018-02-01 R1F      other         245        0          0 TRUE
##  6  7243 2017-07-01 RE9      1            5170      235       1269 TRUE
##  7  8057 2017-04-01 RQM      1           15957     1309       3375 TRUE
##  8  8957 2019-02-01 RNL      1            7258     1374       1947 TRUE
##  9 10214 2018-10-01 RJ1      other        3197        0          0 TRUE
## 10 10328 2018-10-01 RKB      2            2033        8        105 TRUE
## 11 11767 2018-06-01 NLO12    other         336        1          0 TRUE
```

```
attributes(complete_CollectedData)
```

```
## $row.names
##  [1]  1  2  3  4  5  6  7  8  9 10 11
##
## $names
## [1] "index"       "period"      "org_code"    "type"        "attendances"
## [6] "breaches"    "admissions"  "consent"
##
## $spec
## cols(
##   index = col_double(),
##   period = col_date(format = ""),
##   org_code = col_character(),
##   type = col_character(),
##   attendances = col_double(),
##   breaches = col_double(),
##   admissions = col_double(),
##   consent = col_logical()
## )
##
## $problems
## <pointer: 0x7f88b6f49e00>
##
## $class
## [1] "spec_tbl_df" "tbl_df"      "tbl"         "data.frame"
##
## $main
## [1] "This data describes the NHS England accident and emergency (A&E) attendances and breaches of fou
##
## $dictionary
##    variable name
## 1     admissions
## 2    attendances
## 3       breaches
## 4        consent
## 5          index
## 6       org_code
## 7
## 8
## 9
## 10
## 11
## 12
```

```
## 13
## 14
## 15
## 16        period
## 17
## 18
## 19
## 20
## 21
## 22
## 23
## 24
## 25
## 26          type
## 27
## 28
##
## 1
## 2
## 3
## 4
## 5
## 6  The Organisation data service (ODS) code for the organisation. If you want to know the organisatio
## 7
## 8
## 9
## 10
## 11
## 12
## 13
## 14
## 15
## 16
## 17
## 18
## 19
## 20
## 21
## 22
## 23
## 24
## 25
## 26
## 27
## 28
##    variable options
## 1         0 to 3375
## 2      200 to 15957
## 3         0 to 1374
## 4              TRUE
## 5     1155 to 11767
## 6             C82010
## 7                RDZ
## 8                RVR
```

```
## 9               RQM
## 10              R1F
## 11              RE9
## 12              RNL
## 13              RJ1
## 14              RKB
## 15            NLO12
## 16            17136
## 17            17075
## 18            16922
## 19            17591
## 20            17563
## 21            17348
## 22            17257
## 23            17928
## 24            17805
## 25            17683
## 26            other
## 27                1
## 28                2
##
## 1
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9
## 10
## 11
## 12
## 13
## 14
## 15
## 16
## 17
## 18
## 19
## 20
## 21
## 22
## 23
## 24
## 25
## 26
## 27
## 28 2: Consultant-led mono speciality accident and emergency service (e.g. ophthalmology, dental) wit
##
## $last_edit_date
## [1] "2022-06-18 09:54:09 ADT"
##
## $author
```

```
## [1] "B203349"
```

Save the CollectedData with attributes

```
save_it(complete_CollectedData, here("RawData", "complete_CollectedData"))
```

Loading NHS Datasets

Data

The **NHS England accident and emergency attendances and admissions (ae_attendances)** dataset will be loaded from the NHSRdatasets package. This dataset reports on the attendances, four-hour breaches and admissions for all A&E departments in England for the years 2016/17 through 2018/19 (Apr-Mar)

Loading the required packages

Of note *tidyverse* and *here* packages are also required but have already been loaded earlier

```
library(NHSRdatasets)
library(knitr)
library(scales)
library(lubridate)
library(caret)
```

Loading ae_attendances data

ae_attendances data is loaded from the NHSdatasets package

```
data(ae_attendances)
ae<-ae_attendances
```

Viewing the data

The `head()` function is used to look at top n rows of a data frame (default n = 6 rows)

```
head(ae)
```

```
## # A tibble: 6 x 6
##   period     org_code type  attendances breaches admissions
##   <date>     <fct>    <fct>       <dbl>    <dbl>      <dbl>
## 1 2017-03-01 RF4      1           21289     2879       5060
## 2 2017-03-01 RF4      2             813       22          0
## 3 2017-03-01 RF4      other        2850        6          0
## 4 2017-03-01 R1H      1           30210     5902       6943
## 5 2017-03-01 R1H      2             807       11          0
## 6 2017-03-01 R1H      other       11352      136          0
```

Check for missing data

```
# Calculate how many NAs there are in each variable
ae %>%
  map(is.na) %>%
map(sum)
```

```
## $period
## [1] 0
##
## $org_code
## [1] 0
##
## $type
## [1] 0
##
## $attendances
## [1] 0
##
## $breaches
## [1] 0
##
## $admissions
## [1] 0
```

Data is complete

Add an index link column to ae_attendances data

The `rowid_to_column()` function is used to convert row identities to a column named index. This facilitates the linking of partitioned datasets to the raw data if required in the future.

```
ae <- rowid_to_column(ae, "index")
```

Tabiulate data for a report

```
ae %>%
  # Set the period column to show in month-year format
  mutate_at(vars(period), format, "%b-%y") %>%
  # Set the numeric columns to have a comma at the 1000's place
  mutate_at(vars(attendances, breaches, admissions), comma) %>%
  # Show the first 10 rows
  head(10) %>%
  # Format as a table
  kable()
```

index	period	org_code	type	attendances	breaches	admissions
1	Mar-17	RF4	1	21,289.0	2,879.0	5,060.0
2	Mar-17	RF4	2	813.0	22.0	0.0
3	Mar-17	RF4	other	2,850.0	6.0	0.0
4	Mar-17	R1H	1	30,210.0	5,902.0	6,943.0
5	Mar-17	R1H	2	807.0	11.0	0.0
6	Mar-17	R1H	other	11,352.0	136.0	0.0
7	Mar-17	AD913	other	4,381.0	2.0	0.0
8	Mar-17	RYX	other	19,562.0	258.0	0.0
9	Mar-17	RQM	1	17,414.0	2,030.0	3,597.0
10	Mar-17	RQM	other	7,817.0	86.0	0.0

Save the raw ae_attendances data to your 'RawData' folder

```
write_csv(ae, here("RawData", "ae_attendances.csv"))
```

Subsetting the data

The aim of my data capture tool is to help identify performance outliers in the with regard to the 4hr breaches. In order to do this effectively I will need use the entire dataset (i.e. the variables: index, period, attendances, breaches, organisation name, and organisation type).

Of note the **performance** variable was deliberately not created as this is a function of the **attendances** and **breaches** variables. The variable can be created after the collection of data at the data analysis stage.

Creating test and training datasets

Calculate the proportion (prop) of the raw data to assign to the training data.

The proportion of the raw that needs to be assigned to the training data to ensure there is only 10 to 15 records in the test data is:

```
prop<-(1-(15/nrow(ae)))
print(prop)
```

```
## [1] 0.9988249
```

Splitting the raw data

The createDataPartition() function from the *caret* package is used to splint the raw data into test and training data sets.

The 'set.seed()' function allows a random numner to be generated in a reproducable fashion. This ensures that every time the script is run the raw data will be partitioned into the same test and training datasets.

```
set.seed(333)
#Partitioning the raw data into the test and training data.
trainIndex <- createDataPartition(ae$index, p = prop,
```

```
                                  list = FALSE,
                                  times = 1)
head(trainIndex)
```

```
##      Resample1
## [1,]         1
## [2,]         2
## [3,]         3
## [4,]         4
## [5,]         5
## [6,]         6
```

Assign all records in the trainIndex to the training data

```
aeTrain <- ae[ trainIndex,]
nrow(aeTrain)
```

```
## [1] 12753
```

Save the training data to the to working data folder 'Data' as ae_attendances_train.csv

```
write_csv(aeTrain, here("Data", "ae_attendances_train.csv"))
```

Extract the test data (all records that are not in the trainIndex (`-trainIndex`) are assigned to the test data).

```
aeTest  <- ae[-trainIndex,]
nrow(aeTest)
```

```
## [1] 12
```

Set aside the fist record for markers to test and evaluate the data-capture tool.

```
aeTestMarker  <- aeTest[1,]
```

Save the marker test data to the to working data folder 'Data' as ae_attendances_test_marker.csv

```
write_csv(aeTestMarker, here("Data", "ae_attendances_test_marker.csv"))
```

The remaining records are set aside to test with the data capture tool. These are saved to the to working data folder 'Data' as ae_attendances_test.csv

```
aeTest  <- aeTest[2:nrow(aeTest),]
write_csv(aeTest, here("Data", "ae_attendances_test.csv"))
```

Data Capture Tool (Python)

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.