B205683_Assessment

B205683

2022/6/21

[Link to my git repository] (https://github.com/B205683/B205683_assessment.git)

Data dictionary for test data

```
## Warning in system("timedatectl", intern = TRUE): running command 'timedatectl'
## had status 1
```

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

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.6
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.0     v forcats 0.5.1
```

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

```
## here() starts at /home/jovyan/B205683/B205683_assessment
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

```
##        ID          Organisation      Age              LOS
##  Min.   :  1.00   Trust1 : 30   Min.   : 5.00   Min.   : 1.000
##  1st Qu.: 75.75   Trust2 : 30   1st Qu.:24.00   1st Qu.: 2.000
##  Median :150.50   Trust3 : 30   Median :54.00   Median : 4.000
##  Mean   :150.50   Trust4 : 30   Mean   :50.66   Mean   : 4.937
##  3rd Qu.:225.25   Trust5 : 30   3rd Qu.:75.25   3rd Qu.: 7.000
##  Max.   :300.00   Trust6 : 30   Max.   :95.00   Max.   :18.000
##                   (Other):120
##      Death
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.1767
##  3rd Qu.:0.0000
##  Max.   :1.0000
##
```

```
## # A tibble: 300 x 5
##       ID Organisation   Age   LOS Death
##    <int> <ord>        <int> <int> <int>
##  1     1 Trust1          55     2     0
##  2     2 Trust2          27     1     0
##  3     3 Trust3          93    12     0
```

```
##  4     4 Trust4          45     3     1
##  5     5 Trust5          70    11     0
##  6     6 Trust6          60     7     0
##  7     7 Trust7          25     4     0
##  8     8 Trust8          48     4     0
##  9     9 Trust9          51     7     1
## 10    10 Trust10         81     1     0
## # ... with 290 more rows
```

Variable descriptions

```
variable_description <- c("an integer value fictional patient ID number.",
"A factor,stands for 1of 10 fictional hospital trusts, e.g. "Trust1.","an integer, representing the age

print(variable_description)
```

```
## [1] "an integer value fictional patient ID number."
## [2] "A factor,stands for 1of 10 fictional hospital trusts, e.g. "Trust1."
## [3] "an integer, representing the age of fictional patient in years."
## [4] "'Length of Stay,' an integer representing the number of days a patient was in hospital."
## [5] "an integer flag of the status of patients (0 = survived, 1= died in hospital)."
```

Variable types

```
glimpse(LOS)
```

```
## Rows: 300
## Columns: 5
## $ ID           <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17~
## $ Organisation <ord> Trust1, Trust2, Trust3, Trust4, Trust5, Trust6, Trust7, T~
## $ Age          <int> 55, 27, 93, 45, 70, 60, 25, 48, 51, 81, 58, 16, 21, 82, 1~
## $ LOS          <int> 2, 1, 12, 3, 11, 7, 4, 4, 7, 1, 4, 3, 1, 9, 12, 1, 4, 3, ~
## $ Death        <int> 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, ~
```

We have one quantitative values (measured values) variables and four fixed values (allowable values or codes) variables.

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

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

linker

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

```
##       var_name
## 1           ID
## 2 Organisation
## 3          Age
## 4          LOS
## 5        Death
##                                                                          var_desc
## 1                                    an integer value fictional patient ID number.
```

```
## 2                    A factor,stands for 1of 10 fictional hospital trusts, e.g. "Trust1.
## 3                        an integer, representing the age of fictional patient in years.
## 4 'Length of Stay,' an integer representing the number of days a patient was in hospital.
## 5          an integer flag of the status of patients (0 = survived, 1= died in hospital).
##   var_type
## 1        0
## 2        1
## 3        0
## 4        1
## 5        0
```

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

```
## Enter description for variable 'Age' and option '5 to 95':
## Enter description for variable 'Death' and option '0 to 1':
## Enter description for variable 'ID' and option '1 to 300':
## Enter description for variable 'LOS' and option '2':
## Enter description for variable 'LOS' and option '1':
## Enter description for variable 'LOS' and option '12':
## Enter description for variable 'LOS' and option '3':
## Enter description for variable 'LOS' and option '11':
## Enter description for variable 'LOS' and option '7':
## Enter description for variable 'LOS' and option '4':
## Enter description for variable 'LOS' and option '9':
## Enter description for variable 'LOS' and option '5':
## Enter description for variable 'LOS' and option '10':
## Enter description for variable 'LOS' and option '14':
## Enter description for variable 'LOS' and option '8':
## Enter description for variable 'LOS' and option '6':
## Enter description for variable 'LOS' and option '15':
## Enter description for variable 'LOS' and option '18':
## Enter description for variable 'LOS' and option '13':
## Enter description for variable 'Organisation' and option 'Trust1':
## Enter description for variable 'Organisation' and option 'Trust2':
## Enter description for variable 'Organisation' and option 'Trust3':
## Enter description for variable 'Organisation' and option 'Trust4':
## Enter description for variable 'Organisation' and option 'Trust5':
## Enter description for variable 'Organisation' and option 'Trust6':
## Enter description for variable 'Organisation' and option 'Trust7':
## Enter description for variable 'Organisation' and option 'Trust8':
## Enter description for variable 'Organisation' and option 'Trust9':
## Enter description for variable 'Organisation' and option 'Trust10':
```

```
glimpse(dictionary)
```

```
## Rows: 29
## Columns: 4
## $ `variable name`        <chr> "Age", "Death", "ID", "LOS", " ", " ", " ", " "~
## $ `variable description` <chr> "an integer, representing the age of fictional ~
## $ `variable options`     <chr> "5 to 95", "0 to 1", "1 to 300", "2", "1", "12"~
## $ notes                  <chr> "", "", "", "", "", "", "", "", "", "", "", "",~
```

```
glimpse(dictionary)
```

```
## Rows: 29
## Columns: 4
```

```
## $ `variable name`        <chr> "Age", "Death", "ID", "LOS", " ", " ", " ", " "~
## $ `variable description` <chr> "an integer, representing the age of fictional ~
## $ `variable options`     <chr> "5 to 95", "0 to 1", "1 to 300", "2", "1", "12"~
## $ notes                  <chr> "", "", "", "", "", "", "", "", "", "", "", "",~
```

Append data dictionary to the CollectedData

```
main_string <- "This data describes an artificially generated hospital data. Fictional patients at 10 fi
main_string
```

Create main_string for attributes

```
## [1] "This data describes an artificially generated hospital data. Fictional patients at 10 fictional
```

```
complete_LOSData <- incorporate_attr(my.data = LOS, data.dictionary = dictionary,
main_string = main_string)
#Change the author name
attributes(complete_LOSData)$author[1]<-"B205687"
complete_LOSData
```

```
## # A tibble: 300 x 5
##       ID Organisation   Age   LOS Death
##   * <int> <ord>        <int> <int> <int>
##  1     1 Trust1          55     2     0
##  2     2 Trust2          27     1     0
##  3     3 Trust3          93    12     0
##  4     4 Trust4          45     3     1
##  5     5 Trust5          70    11     0
##  6     6 Trust6          60     7     0
##  7     7 Trust7          25     4     0
##  8     8 Trust8          48     4     0
##  9     9 Trust9          51     7     1
## 10    10 Trust10         81     1     0
## # ... with 290 more rows
```

```
attributes(complete_LOSData)
```

```
## $names
## [1] "ID"           "Organisation" "Age"          "LOS"          "Death"
##
## $row.names
##   [1]   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18
##  [19]  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36
##  [37]  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54
##  [55]  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72
##  [73]  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90
##  [91]  91  92  93  94  95  96  97  98  99 100 101 102 103 104 105 106 107 108
## [109] 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126
## [127] 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144
## [145] 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162
## [163] 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180
## [181] 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198
## [199] 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216
## [217] 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234
## [235] 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252
```

```
## [253] 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270
## [271] 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288
## [289] 289 290 291 292 293 294 295 296 297 298 299 300
##
## $class
## [1] "tbl_df"     "tbl"        "data.frame"
##
## $main
## [1] "This data describes an artificially generated hospital data. Fictional patients at 10 fictional
##
## $dictionary
##    variable name
## 1            Age
## 2          Death
## 3             ID
## 4            LOS
## 5
## 6
## 7
## 8
## 9
## 10
## 11
## 12
## 13
## 14
## 15
## 16
## 17
## 18
## 19
## 20  Organisation
## 21
## 22
## 23
## 24
## 25
## 26
## 27
## 28
## 29
##                                                                    variable description
## 1                           an integer, representing the age of fictional patient in years.
## 2             an integer flag of the status of patients (0 = survived, 1= died in hospital).
## 3                                                 an integer value fictional patient ID number.
## 4  'Length of Stay,' an integer representing the number of days a patient was in hospital.
## 5
## 6
## 7
## 8
## 9
## 10
## 11
## 12
```

```
## 13
## 14
## 15
## 16
## 17
## 18
## 19
## 20                    A factor,stands for 1of 10 fictional hospital trusts, e.g. "Trust1.
## 21
## 22
## 23
## 24
## 25
## 26
## 27
## 28
## 29
##    variable options notes
## 1           5 to 95
## 2            0 to 1
## 3          1 to 300
## 4                 2
## 5                 1
## 6                12
## 7                 3
## 8                11
## 9                 7
## 10                4
## 11                9
## 12                5
## 13               10
## 14               14
## 15                8
## 16                6
## 17               15
## 18               18
## 19               13
## 20           Trust1
## 21           Trust2
## 22           Trust3
## 23           Trust4
## 24           Trust5
## 25           Trust6
## 26           Trust7
## 27           Trust8
## 28           Trust9
## 29          Trust10
##
## $last_edit_date
## [1] "2022-06-21 05:57:02 UTC"
##
## $author
## [1] "B205687"
```

Calculate how many NAs there are in each variable

```
LOS %>%
  map(is.na) %>%
map(sum)
```

```
## $ID
## [1] 0
##
## $Organisation
## [1] 0
##
## $Age
## [1] 0
##
## $LOS
## [1] 0
##
## $Death
## [1] 0
```

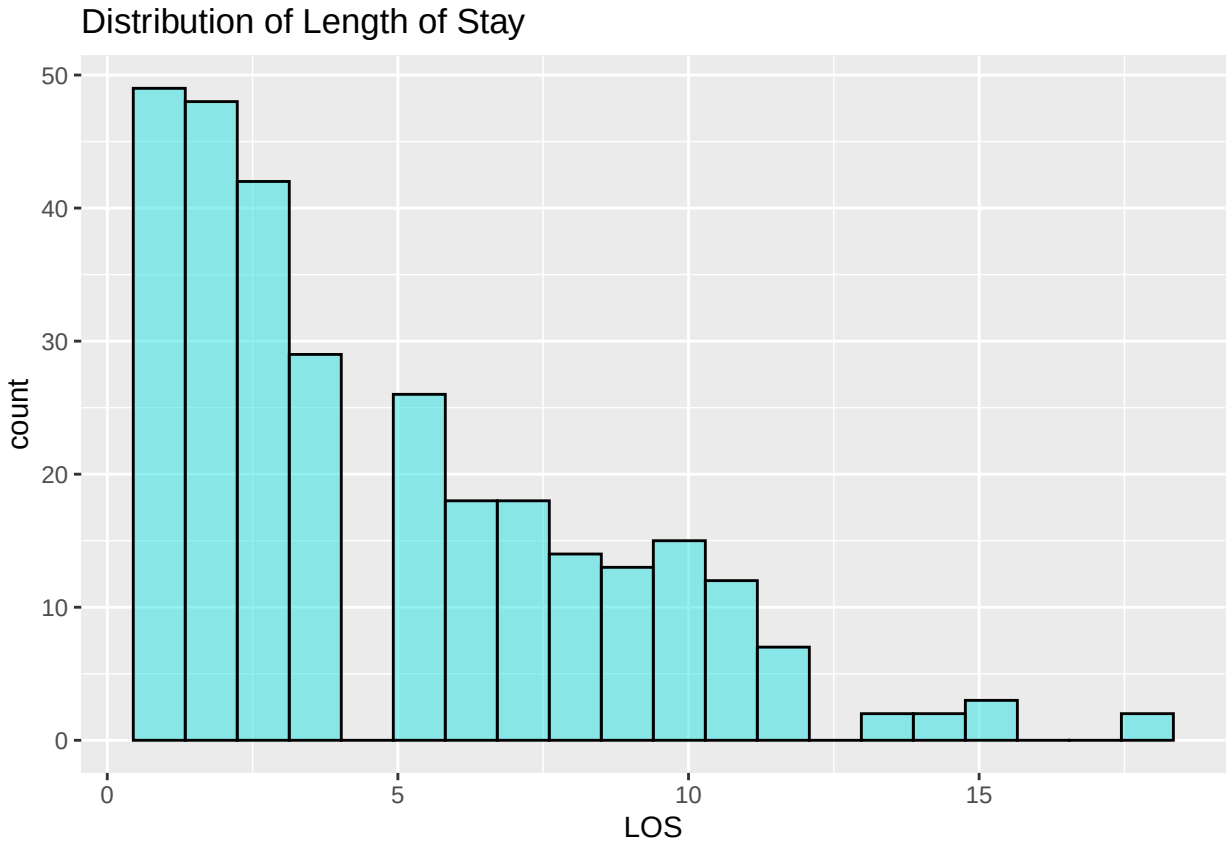
```
#The data is complete.

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

Including Plots

Look at the distribution of Length of Stay (LOS)

```
ggplot(LOS_model, aes(x=LOS)) +
  geom_histogram(alpha=0.5, col=1, fill=13, bins=20)+
  ggtitle("Distribution of Length of Stay")
```


Distribution of Length of Stay
count
LOS
0
10
20
30
40
50
0
5
10
15