P8106_midterm

lz2951

2024-03-28

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Import Data

```
load("recovery.RData")

str(dat)
```

```
## 'data.frame':    3000 obs. of  16 variables:
##  $ id           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ age          : num  56 70 57 53 59 60 56 58 60 60 ...
##  $ gender       : int  0 1 1 0 1 1 0 1 0 1 ...
##  $ race         : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 3 1 1 1 1 ...
##  $ smoking      : Factor w/ 3 levels "0","1","2": 3 2 1 1 3 2 1 1 2 1 ...
##  $ height       : num  170 170 168 167 174 ...
##  $ weight       : num  78.7 73.1 77.4 76.1 70.2 75.1 79.1 62.6 81.8 75.7 ...
##  $ bmi          : num  27.2 25.4 27.3 27.4 23.3 28.4 27.5 26.8 28.8 27.3 ...
##  $ hypertension : num  0 1 1 0 0 0 0 1 1 0 ...
##  $ diabetes     : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ SBP          : num  120 134 131 115 127 129 122 134 136 127 ...
##  $ LDL          : num  97 112 88 87 118 104 66 104 126 123 ...
##  $ vaccine      : int  0 0 1 0 1 0 0 1 1 1 ...
##  $ severity     : int  0 0 0 1 0 0 0 0 1 0 ...
##  $ study        : chr  "A" "A" "A" "A" ...
##  $ recovery_time: num  31 44 29 47 40 34 31 41 50 33 ...
```

```r
recovery = dat |>
  janitor::clean_names() |>
  mutate(gender = as.factor(gender),
         hypertension = as.factor(hypertension),
         diabetes = as.factor(diabetes),
         vaccine = as.factor(vaccine),
         severity = as.factor(severity),
         study = as.factor(study))

str(recovery)
```

```
## 'data.frame':    3000 obs. of  16 variables:
##  $ id           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ age          : num  56 70 57 53 59 60 56 58 60 60 ...
##  $ gender       : Factor w/ 2 levels "0","1": 1 2 2 1 2 2 1 2 1 2 ...
##  $ race         : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 3 1 1 1 1 ...
##  $ smoking      : Factor w/ 3 levels "0","1","2": 3 2 1 1 3 2 1 1 2 1 ...
##  $ height       : num  170 170 168 167 174 ...
##  $ weight       : num  78.7 73.1 77.4 76.1 70.2 75.1 79.1 62.6 81.8 75.7 ...
##  $ bmi          : num  27.2 25.4 27.3 27.4 23.3 28.4 27.5 26.8 28.8 27.3 ...
##  $ hypertension : Factor w/ 2 levels "0","1": 1 2 2 1 1 1 1 2 2 1 ...
##  $ diabetes     : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 2 1 1 1 ...
##  $ sbp          : num  120 134 131 115 127 129 122 134 136 127 ...
##  $ ldl          : num  97 112 88 87 118 104 66 104 126 123 ...
##  $ vaccine      : Factor w/ 2 levels "0","1": 1 1 2 1 2 1 1 2 2 2 ...
##  $ severity     : Factor w/ 2 levels "0","1": 1 1 1 2 1 1 1 1 2 1 ...
##  $ study        : Factor w/ 2 levels "A","B": 1 1 1 1 1 1 1 1 1 1 ...
##  $ recovery_time: num  31 44 29 47 40 34 31 41 50 33 ...
```

Exploratory analysis and data visualization

```r
skimr::skim(recovery)
```

Table 1: Data summary

Name	recovery
Number of rows	3000
Number of columns	16
Column type frequency:	
factor	8
numeric	8
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
gender	0	1	FALSE	2	0: 1544, 1: 1456
race	0	1	FALSE	4	1: 1967, 3: 604, 4: 271, 2: 158
smoking	0	1	FALSE	3	0: 1822, 1: 859, 2: 319
hypertension	0	1	FALSE	2	0: 1508, 1: 1492
diabetes	0	1	FALSE	2	0: 2537, 1: 463
vaccine	0	1	FALSE	2	1: 1788, 0: 1212
severity	0	1	FALSE	2	0: 2679, 1: 321
study	0	1	FALSE	2	A: 2000, B: 1000

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
id	0	1	1500.50	866.17	1.0	750.75	1500.50	2250.25	3000.0	
age	0	1	60.20	4.48	42.0	57.00	60.00	63.00	79.0	
height	0	1	169.90	5.97	147.8	166.00	169.90	173.90	188.6	
weight	0	1	79.96	7.14	55.9	75.20	79.80	84.80	103.7	
bmi	0	1	27.76	2.79	18.8	25.80	27.65	29.50	38.9	
sbp	0	1	130.47	7.97	105.0	125.00	130.00	136.00	156.0	
ldl	0	1	110.45	19.76	28.0	97.00	110.00	124.00	178.0	
recovery_time	0	1	42.17	23.15	2.0	31.00	39.00	49.00	365.0	