P8106_midterm

lz2951

2024-03-28

```r
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
```

```r
library(ggcorrplot)
library(pheatmap)
```

Import Data

```r
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) |>
  select(-numeric.hist)
```

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

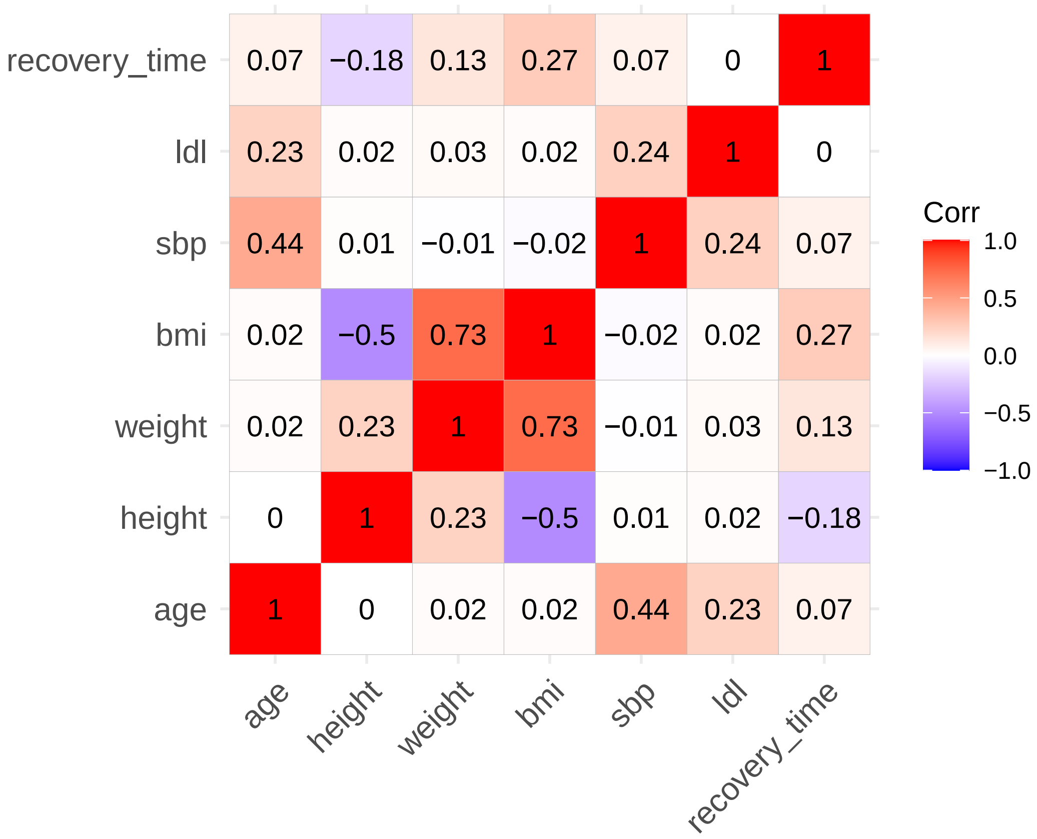
skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
gender	0	1	FALSE	2	0: 1544, 1: 1456
race	0	1	FALSE	4	1: 1967, 3: 604, 4: 271, 2: 158
smoking	0	1	FALSE	3	0: 1822, 1: 859, 2: 319
hypertension	0	1	FALSE	2	0: 1508, 1: 1492
diabetes	0	1	FALSE	2	0: 2537, 1: 463
vaccine	0	1	FALSE	2	1: 1788, 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
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

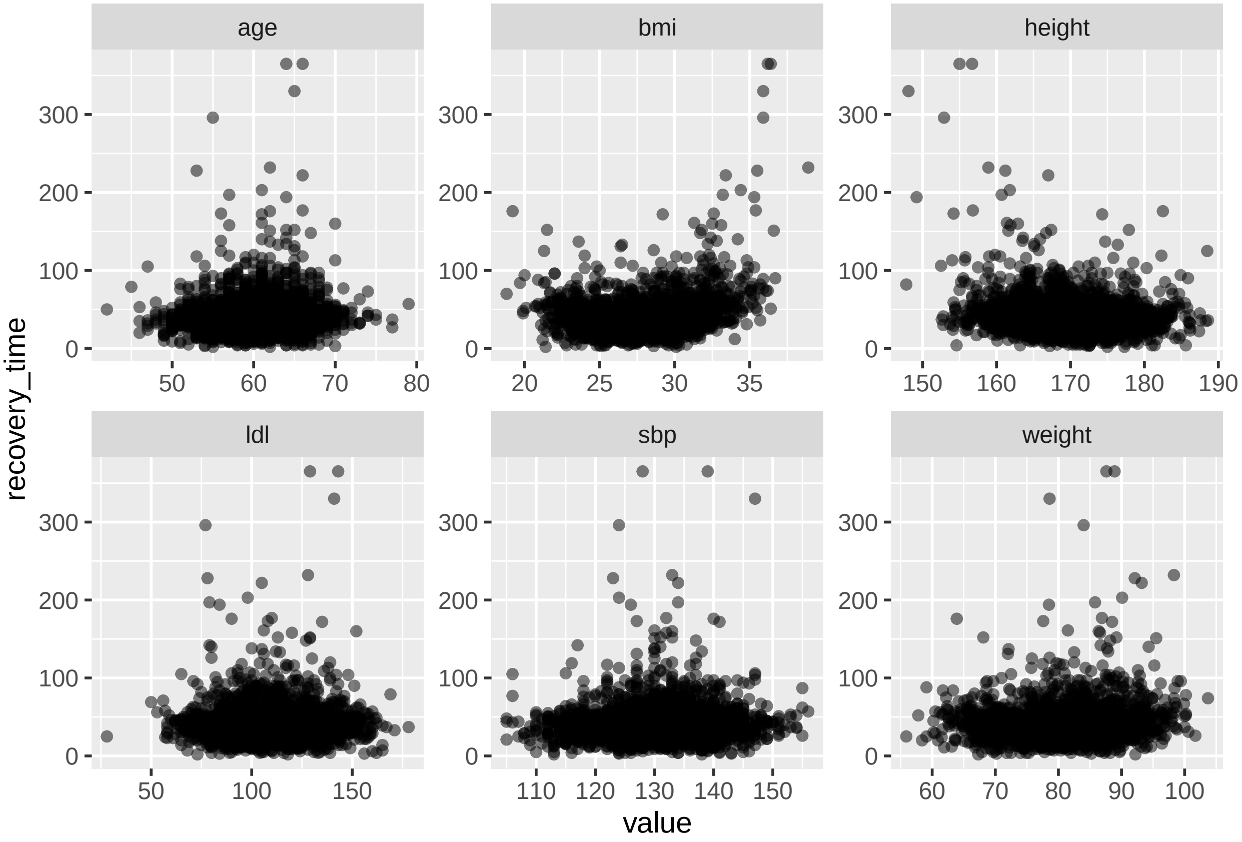
Analysis between numeric predictors

```r
recovery_numeric =
  recovery |>
  select(where(is.numeric), -id)

# recovery_numeric

ggcorrplot(cor(recovery_numeric), lab = T)
```



	age	height	weight	bmi	sbp	ldl	recovery_time
recovery_time	0.07	−0.18	0.13	0.27	0.07	0	1
ldl	0.23	0.02	0.03	0.02	0.24	1	0
sbp	0.44	0.01	−0.01	−0.02	1	0.24	0.07
bmi	0.02	−0.5	0.73	1	−0.02	0.02	0.27
weight	0.02	0.23	1	0.73	−0.01	0.03	0.13
height	0	1	0.23	−0.5	0.01	0.02	−0.18
age	1	0	0.02	0.02	0.44	0.23	0.07

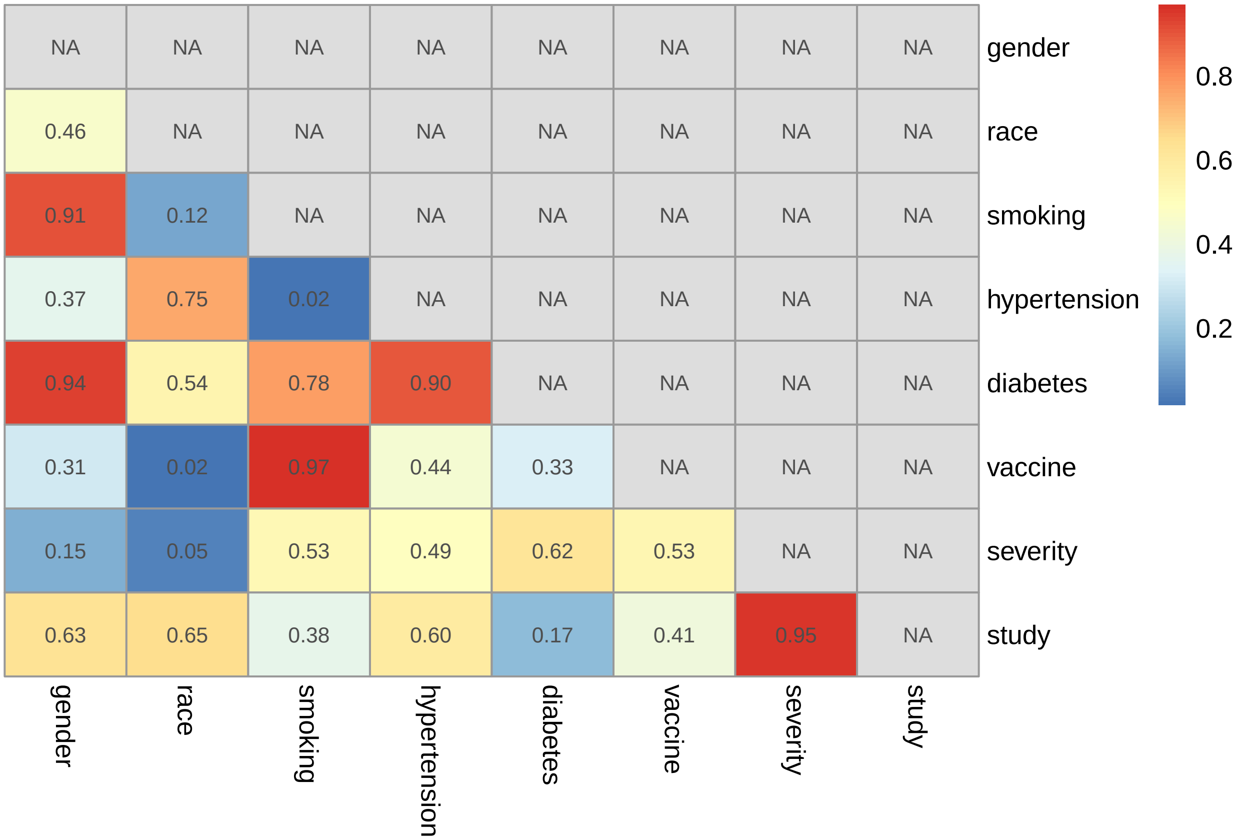
```r
recovery_numeric_long =
  recovery_numeric |>
  gather(key = "predictor", value = "value", -recovery_time)

# recovery_numeric_long

ggplot(recovery_numeric_long, aes(x = value, y = recovery_time)) +
  geom_point(alpha = 0.5) +
  facet_wrap(~predictor, scales = "free")
```



Analysis between factor predictors

```
recovery_factor =
  recovery |>
  select(where(is.factor), recovery_time)

# recovery_factor

recovery_factor_nonresp =
  recovery |>
  select(where(is.factor))

# recovery_factor_nonresp

chi_sq_matrix = matrix(NA, ncol = ncol(recovery_factor_nonresp), nrow = ncol(recovery_factor_nonresp))
for (i in 1:(ncol(recovery_factor_nonresp)-1)) {
  for (j in (i+1):ncol(recovery_factor_nonresp)) {
    cross_table = table(recovery_factor_nonresp[,i],
                        recovery_factor_nonresp[,j])
    chi_sq_matrix[i,j] = chisq.test(cross_table)$p.value
  }
}

rownames(chi_sq_matrix) = colnames(recovery_factor_nonresp)
colnames(chi_sq_matrix) = colnames(recovery_factor_nonresp)
```

```r
# chi_sq_matrix

chi_sq_matrix = t(chi_sq_matrix)

# chi_sq_matrix

pheatmap(chi_sq_matrix,
         cluster_rows = FALSE, cluster_cols = FALSE,
         show_rownames = TRUE, show_colnames = TRUE,
         legend = TRUE, display_numbers = TRUE)
```

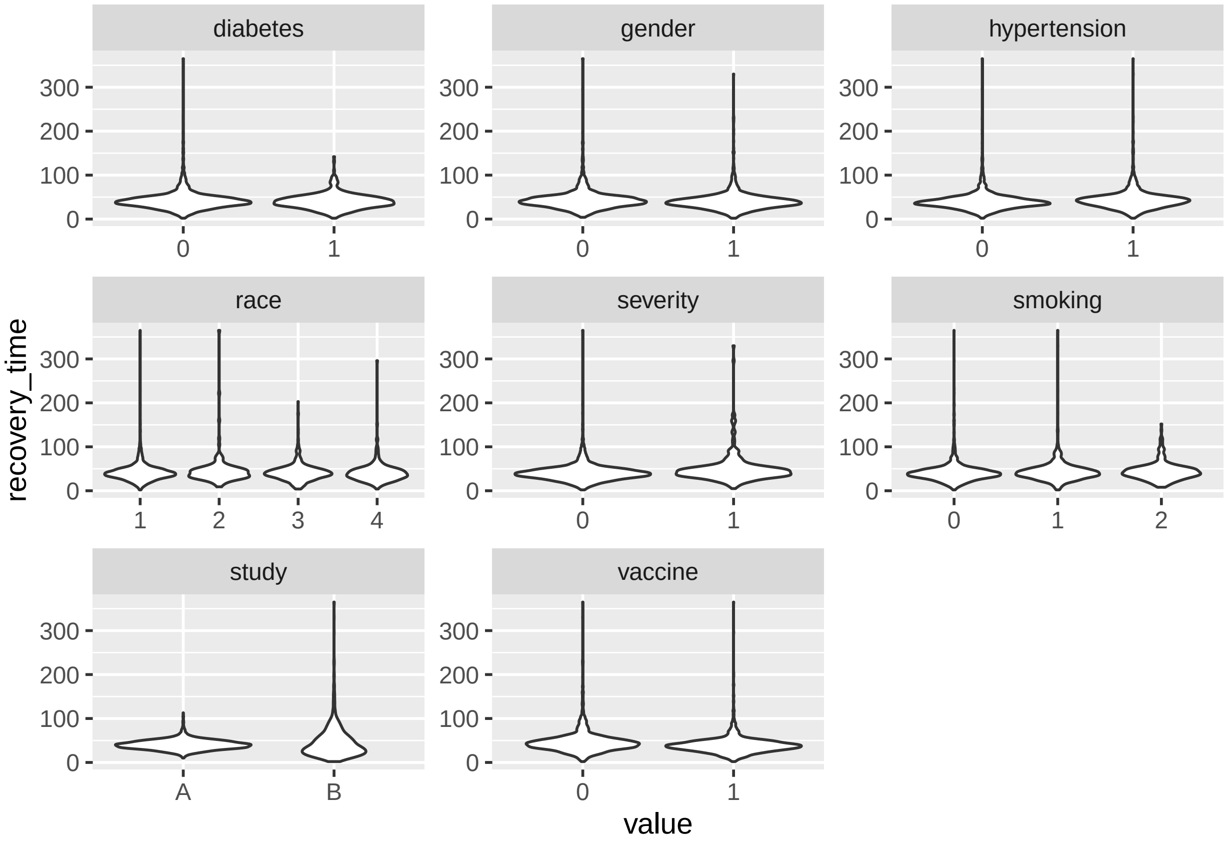


gender	race	smoking	hypertension	diabetes	vaccine	severity	study	
NA	NA	NA	NA	NA	NA	NA	NA	gender
0.46	NA	NA	NA	NA	NA	NA	NA	race
0.91	0.12	NA	NA	NA	NA	NA	NA	smoking
0.37	0.75	0.02	NA	NA	NA	NA	NA	hypertension
0.94	0.54	0.78	0.90	NA	NA	NA	NA	diabetes
0.31	0.02	0.97	0.44	0.33	NA	NA	NA	vaccine
0.15	0.05	0.53	0.49	0.62	0.53	NA	NA	severity
0.63	0.65	0.38	0.60	0.17	0.41	0.95	NA	study

```r
recovery_factor_long =
  recovery_factor |>
  gather(key = "predictor", value = "value", -recovery_time)
```

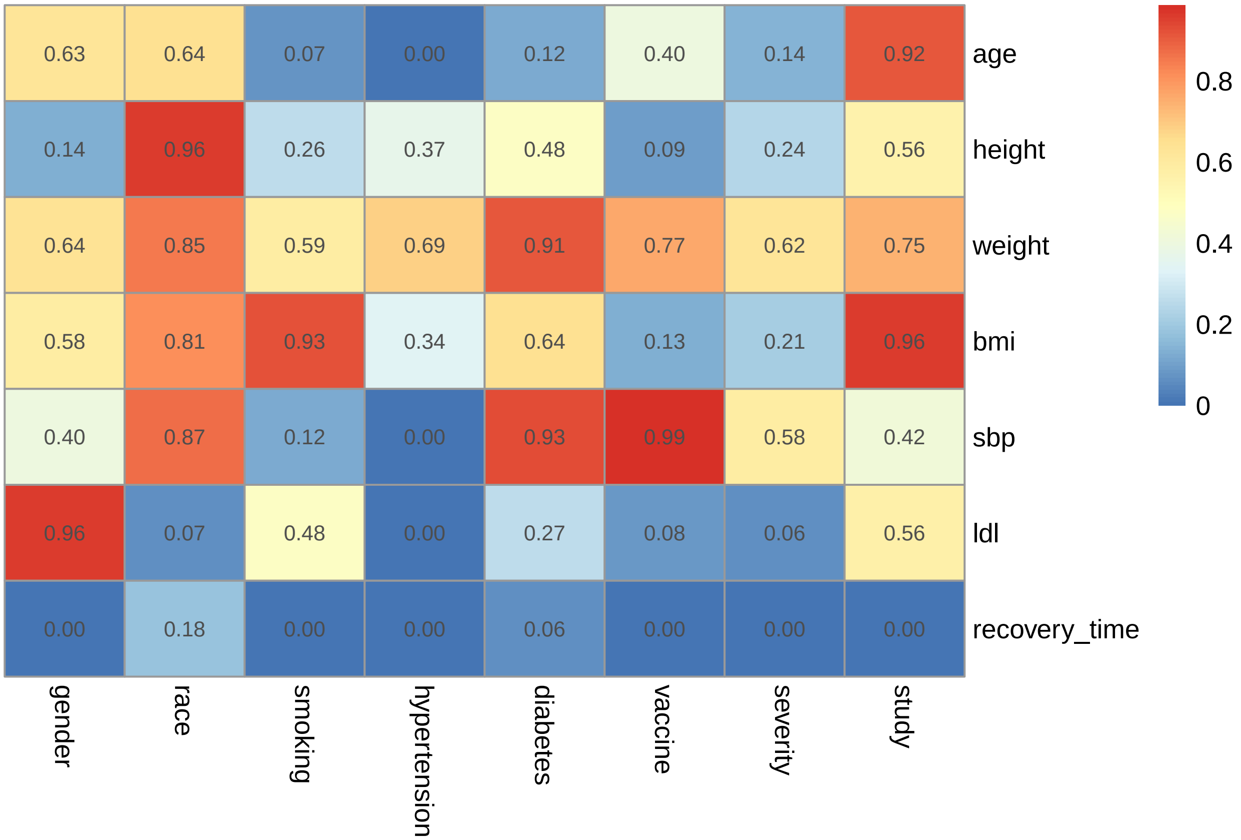
```
## Warning: attributes are not identical across measure variables; they will be
## dropped
```

```r
# recovery_factor_long

ggplot(recovery_factor_long, aes(x = value, y = recovery_time)) +
  geom_violin() +
  facet_wrap(~predictor, scales = "free")
```



Analysis between numeric and factor predictors

```r
anova_matrix = matrix(NA, ncol = ncol(recovery_factor_nonresp), nrow = ncol(recovery_numeric))
for (i in 1:(ncol(recovery_numeric))) {
  for (j in 1:ncol(recovery_factor_nonresp)) {
    cross_dat = data.frame(num = recovery_numeric[,i],
                           fac = recovery_factor_nonresp[,j])
    anova_matrix[i,j] = summary(aov(num ~ fac, data = cross_dat))[[1]]$"Pr(>F)"[[1]]
  }
}

# anova_matrix

rownames(anova_matrix) = colnames(recovery_numeric)
colnames(anova_matrix) = colnames(recovery_factor_nonresp)

pheatmap(anova_matrix,
         cluster_rows = FALSE, cluster_cols = FALSE,
         show_rownames = TRUE, show_colnames = TRUE,
         legend = TRUE, display_numbers = TRUE)
```



	gender	race	smoking	hypertension	diabetes	vaccine	severity	study
age	0.63	0.64	0.07	0.00	0.12	0.40	0.14	0.92
height	0.14	0.96	0.26	0.37	0.48	0.09	0.24	0.56
weight	0.64	0.85	0.59	0.69	0.91	0.77	0.62	0.75
bmi	0.58	0.81	0.93	0.34	0.64	0.13	0.21	0.96
sbp	0.40	0.87	0.12	0.00	0.93	0.99	0.58	0.42
ldl	0.96	0.07	0.48	0.00	0.27	0.08	0.06	0.56
recovery_time	0.00	0.18	0.00	0.00	0.06	0.00	0.00	0.00