data_cleaning_ravi

Ravi Brenner

2025-03-25

```
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
library(tidymodels)
```

```
## -- Attaching packages ------------------------------------- tidymodels 1.2.0 --
```

```
## v broom        1.0.6     v rsample      1.2.1
## v dials        1.3.0     v tibble       3.2.1
## v dplyr        1.1.4     v tidyr        1.3.1
## v infer        1.0.7     v tune         1.2.1
## v modeldata    1.4.0     v workflows    1.1.4
## v parsnip      1.2.1     v workflowsets 1.1.0
## v purrr        1.0.2     v yardstick    1.3.1
## v recipes      1.1.0
```

```
## -- Conflicts ---------------------------------------- tidymodels_conflicts() --
## x purrr::discard()         masks scales::discard()
## x dplyr::filter()          masks stats::filter()
## x dplyr::lag()             masks stats::lag()
## x purrr::lift()            masks caret::lift()
## x yardstick::precision()   masks caret::precision()
## x yardstick::recall()      masks caret::recall()
## x yardstick::sensitivity() masks caret::sensitivity()
## x yardstick::specificity() masks caret::specificity()
## x recipes::step()          masks stats::step()
## * Learn how to get started at https://www.tidymodels.org/start/
```

```
library(vtable)
```

```
## Loading required package: kableExtra
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
library(corrplot)
```

```
## corrplot 0.95 loaded
```

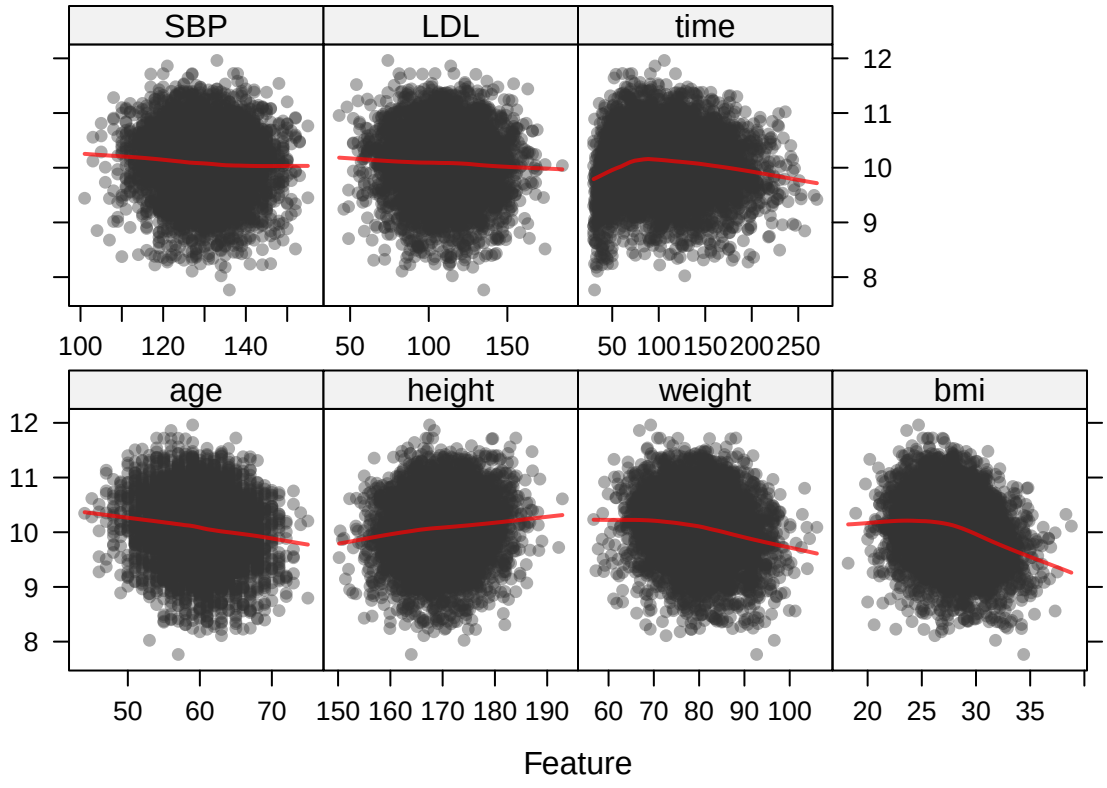
```
load("dat1.RData")
load("dat2.RData")
```

Go from labels to variable names

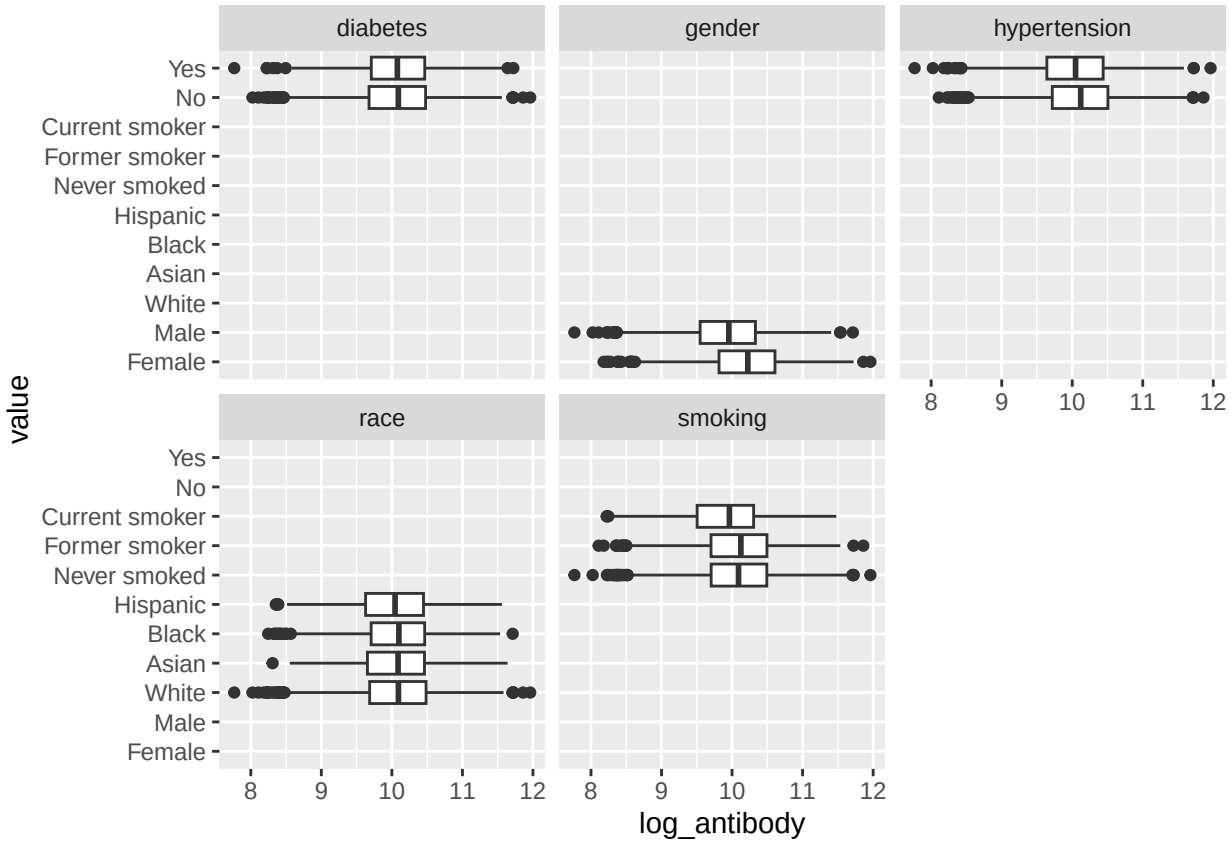
```
dat1 <- dat1 |>
  mutate(gender = factor(gender, levels = c(0,1),
                         labels = c("Female","Male")),
         race = factor(race, levels = c(1,2,3,4),
                       labels = c("White","Asian","Black","Hispanic")),
         smoking = factor(smoking, levels = c(0,1,2),
                          labels = c("Never smoked","Former smoker","Current smoker")),
         diabetes = factor(diabetes, levels = c(0,1),
                           labels = c("No","Yes")),
         hypertension = factor(hypertension, levels = c(0,1),
                               labels = c("No","Yes")),
         ) |>
  dplyr::select(-id)

dat2 <- dat2 |>
  mutate(gender = factor(gender, levels = c(0,1),
                         labels = c("Female","Male")),
         race = factor(race, levels = c(1,2,3,4),
                       labels = c("White","Asian","Black","Hispanic")),
         smoking = factor(smoking, levels = c(0,1,2),
                          labels = c("Never smoked","Former smoker","Current smoker")),
         diabetes = factor(diabetes, levels = c(0,1),
                           labels = c("No","Yes")),
         hypertension = factor(hypertension, levels = c(0,1),
                               labels = c("No","Yes")),
         ) |>
  dplyr::select(-id)
```

Use featureplot from caret to plot the training data

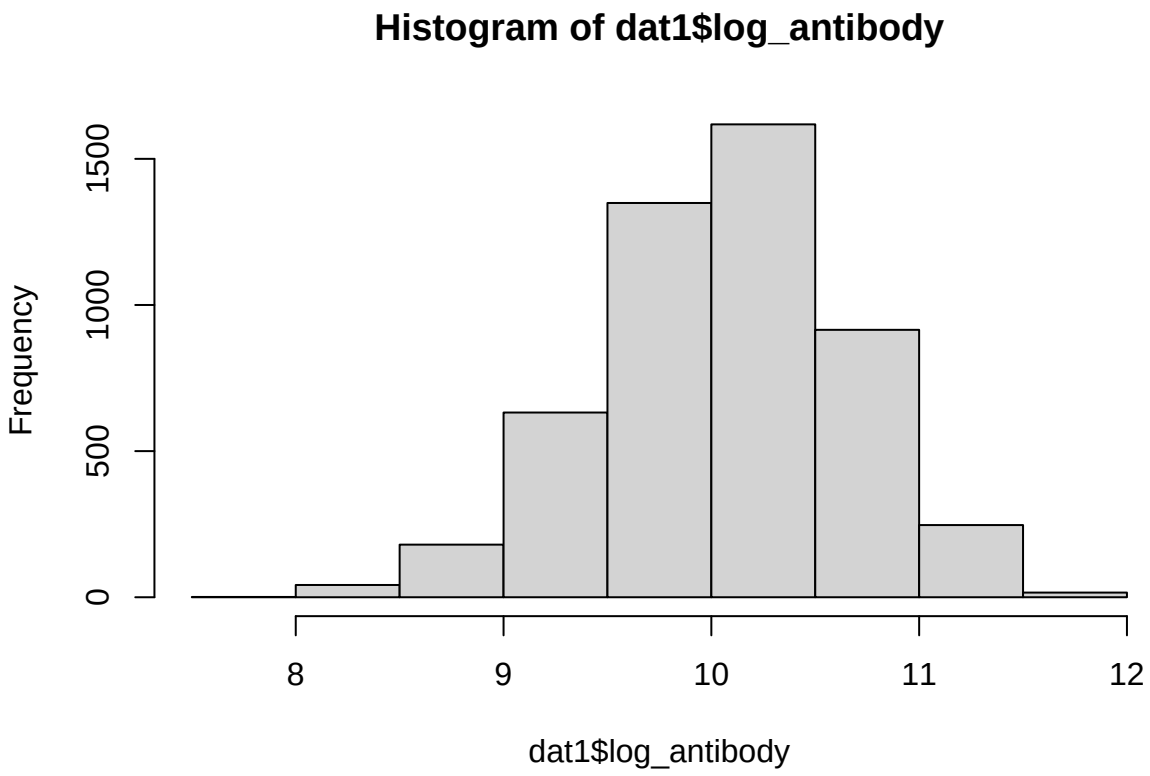
```
theme1 <- trellis.par.get()
theme1$plot.symbol$col = rgb(.2, .2, .2, .4)
theme1$plot.symbol$pch = 16
theme1$plot.line$col = rgb(1, 0, 0, .7)
theme1$plot.line$lwd <- 2
trellis.par.set(theme1)

featurePlot(x = dat1 |> dplyr::select(-log_antibody,
                                      -where(is.factor)),
            y = dat1$log_antibody,
            type = c("p","smooth"))
```



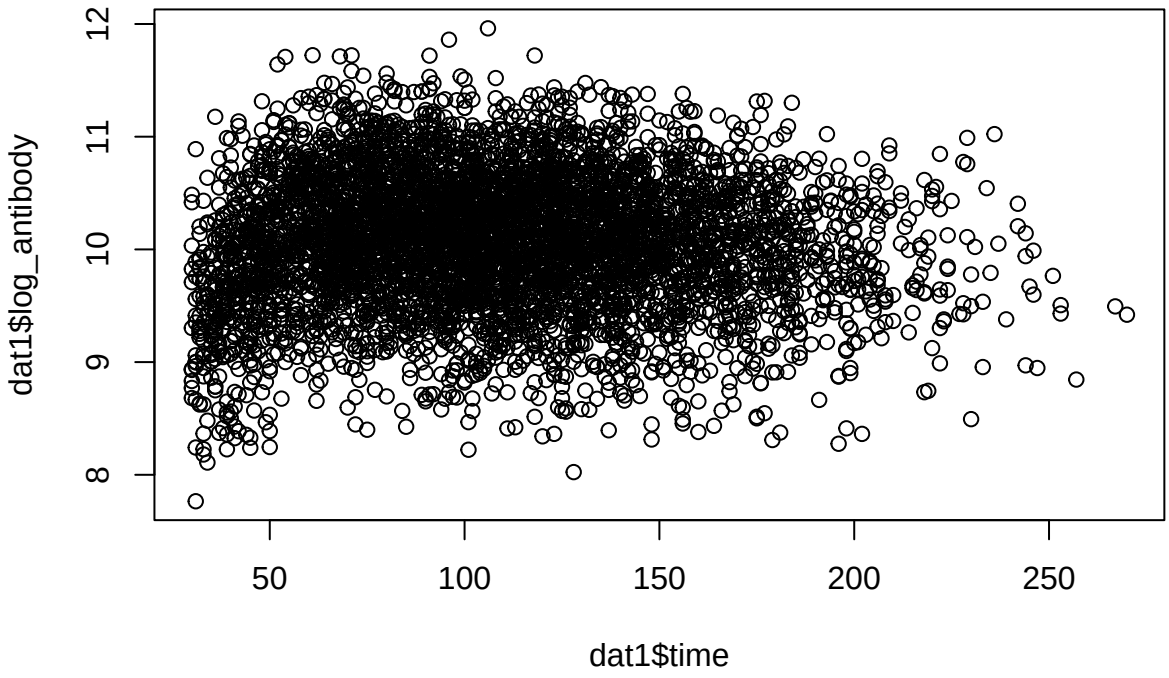

```
dat1 |>
  dplyr::select(log_antibody,gender,race, smoking,diabetes, hypertension) |>
  pivot_longer(cols = 2:6,
               names_to = "variable",
               values_to = "value") |>
  ggplot(aes(y = value, x = log_antibody)) +
  geom_boxplot() +
  facet_wrap(.~variable)
```



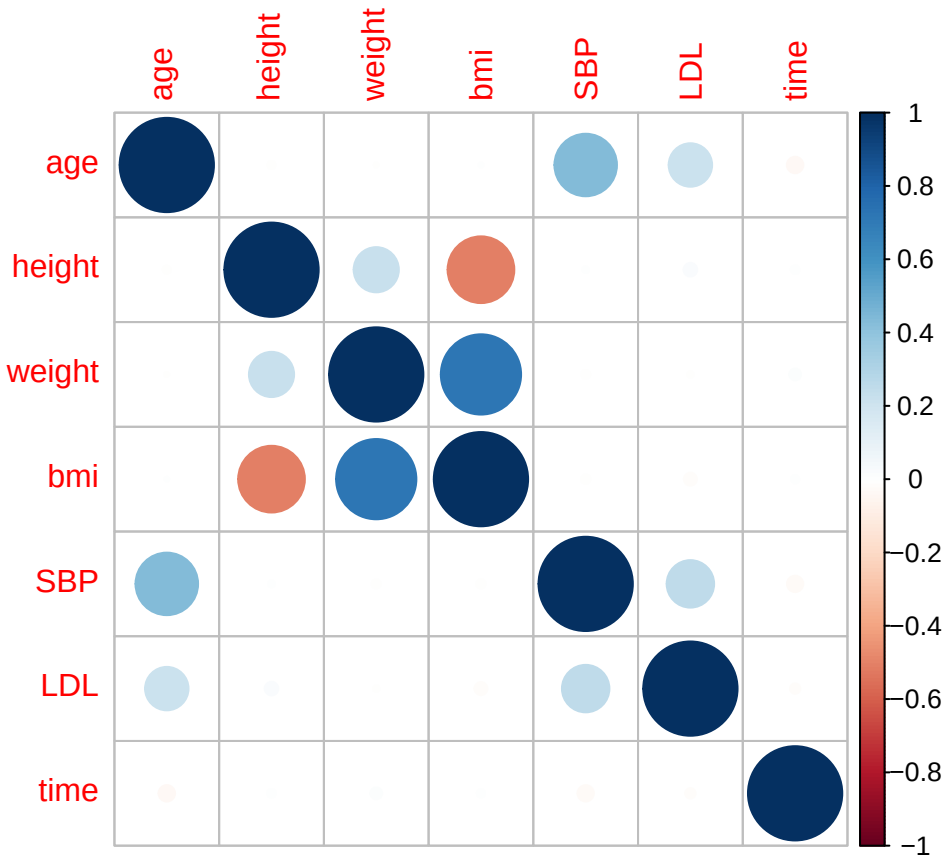

```
antibody_hist = hist(dat1$log_antibody)
```




```
antibody_scatter = plot(x = dat1$time, y = dat1$log_antibody)
```




```
summ_table = sumtable(dat1, out = 'return')
continuous = dat1[c(1,5:7,10:12)]
correlations = cor(continuous)
corr_plot = corrplot(correlations)
```




```
report_table = sumtable(dat1, out = 'kable')
report_table
```

Table 1: Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
age	5000	60	4.5	44	57	63	75
gender	5000						
... Female	2573	51%					
... Male	2427	49%					
race	5000						
... White	3221	64%					
... Asian	278	6%					
... Black	1036	21%					
... Hispanic	465	9%					
smoking	5000						
... Never smoked	3010	60%					
... Former smoker	1504	30%					
... Current smoker	486	10%					
height	5000	170	5.9	150	166	174	193
weight	5000	80	7.1	57	75	85	106
bmi	5000	28	2.8	18	26	30	39
diabetes	5000						
... No	4228	85%					
... Yes	772	15%					
hypertension	5000						
... No	2702	54%					
... Yes	2298	46%					
SBP	5000	130	8	101	124	135	155
LDL	5000	110	20	43	96	124	185
time	5000	109	43	30	76	138	270
log_antibody	5000	10	0.6	7.8	9.7	10	12