EDA

Rio Yan

2021-05-11

load data

```
student_df = read.table("./student-mat.csv", header = TRUE, sep = ";") %>%
  janitor::clean_names() %>%
  mutate(grade = round((g1+g2+g3)/3,2),
        letter_grade = case_when(grade >= 10 ~ "pass",
                                 grade < 10 ~ "fail")) %>%
  dplyr::select(-g1,-g2,-g3,-grade) %>%
  mutate_if(is.character, as.factor)

view(student_df)
```

```
skimr::skim(student_df)
```

Table 1: Data summary

Name	student_df
Number of rows	395
Number of columns	31
Column type frequency:	
factor	18
numeric	13
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
school	0	1	FALSE	2	GP: 349, MS: 46
sex	0	1	FALSE	2	F: 208, M: 187
address	0	1	FALSE	2	U: 307, R: 88
famsize	0	1	FALSE	2	GT3: 281, LE3: 114
pstatus	0	1	FALSE	2	T: 354, A: 41
mjob	0	1	FALSE	5	oth: 141, ser: 103, at_: 59, tea: 58
fjob	0	1	FALSE	5	oth: 217, ser: 111, tea: 29, at_: 20
reason	0	1	FALSE	4	cou: 145, hom: 109, rep: 105, oth: 36
guardian	0	1	FALSE	3	mot: 273, fat: 90, oth: 32
schoolsup	0	1	FALSE	2	no: 344, yes: 51
famsup	0	1	FALSE	2	yes: 242, no: 153
paid	0	1	FALSE	2	no: 214, yes: 181
activities	0	1	FALSE	2	yes: 201, no: 194

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
nursery	0	1	FALSE	2	yes: 314, no: 81
higher	0	1	FALSE	2	yes: 375, no: 20
internet	0	1	FALSE	2	yes: 329, no: 66
romantic	0	1	FALSE	2	no: 263, yes: 132
letter_grade	0	1	FALSE	2	pas: 231, fai: 164

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
age	0	1	16.70	1.28	15	16	17	18	22	
medu	0	1	2.75	1.09	0	2	3	4	4	
fedu	0	1	2.52	1.09	0	2	2	3	4	
traveltime	0	1	1.45	0.70	1	1	1	2	4	
studytime	0	1	2.04	0.84	1	1	2	2	4	
failures	0	1	0.33	0.74	0	0	0	0	3	
famrel	0	1	3.94	0.90	1	4	4	5	5	
freetime	0	1	3.24	1.00	1	3	3	4	5	
goout	0	1	3.11	1.11	1	2	3	4	5	
dalc	0	1	1.48	0.89	1	1	1	2	5	
walc	0	1	2.29	1.29	1	1	2	3	5	
health	0	1	3.55	1.39	1	3	4	5	5	
absences	0	1	5.71	8.00	0	0	4	8	75	

exploritory analysis

summary table for categorical

```
library(gtsummary)

mat_categorical =
  student_df %>%
  dplyr::select(-age, -medu, -fedu, -traveltime, -studytime, -failures, -famrel, -freetime, -goout, -dal

mat_categorical %>%
  tbl_summary()
```

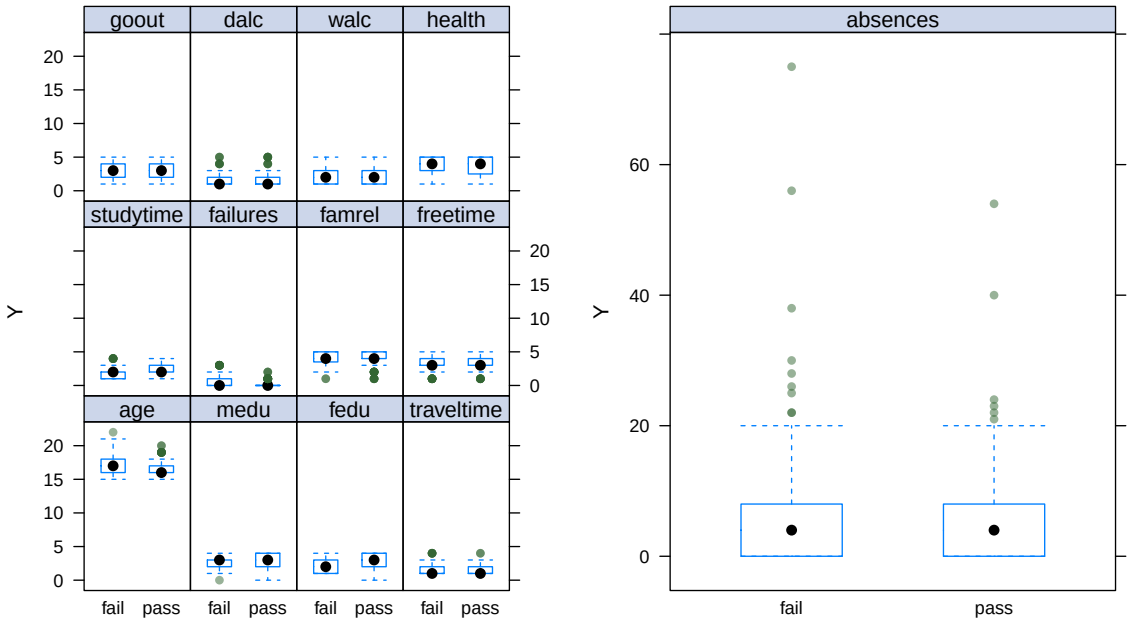
Characteristic	**N = 395**
school	
GP	349 (88%)
MS	46 (12%)
sex	
F	208 (53%)
M	187 (47%)
address	
R	88 (22%)
U	307 (78%)
famsize	
GT3	281 (71%)
LE3	114 (29%)
pstatus	

Characteristic	N = 395
A	41 (10%)
T	354 (90%)
mjob	
at_home	59 (15%)
health	34 (8.6%)
other	141 (36%)
services	103 (26%)
teacher	58 (15%)
fjob	
at_home	20 (5.1%)
health	18 (4.6%)
other	217 (55%)
services	111 (28%)
teacher	29 (7.3%)
reason	
course	145 (37%)
home	109 (28%)
other	36 (9.1%)
reputation	105 (27%)
guardian	
father	90 (23%)
mother	273 (69%)
other	32 (8.1%)
schoolsup	51 (13%)
famsup	242 (61%)
paid	181 (46%)
activities	201 (51%)
nursery	314 (79%)
higher	375 (95%)
internet	329 (83%)
romantic	132 (33%)
letter_grade	
fail	164 (42%)
pass	231 (58%)

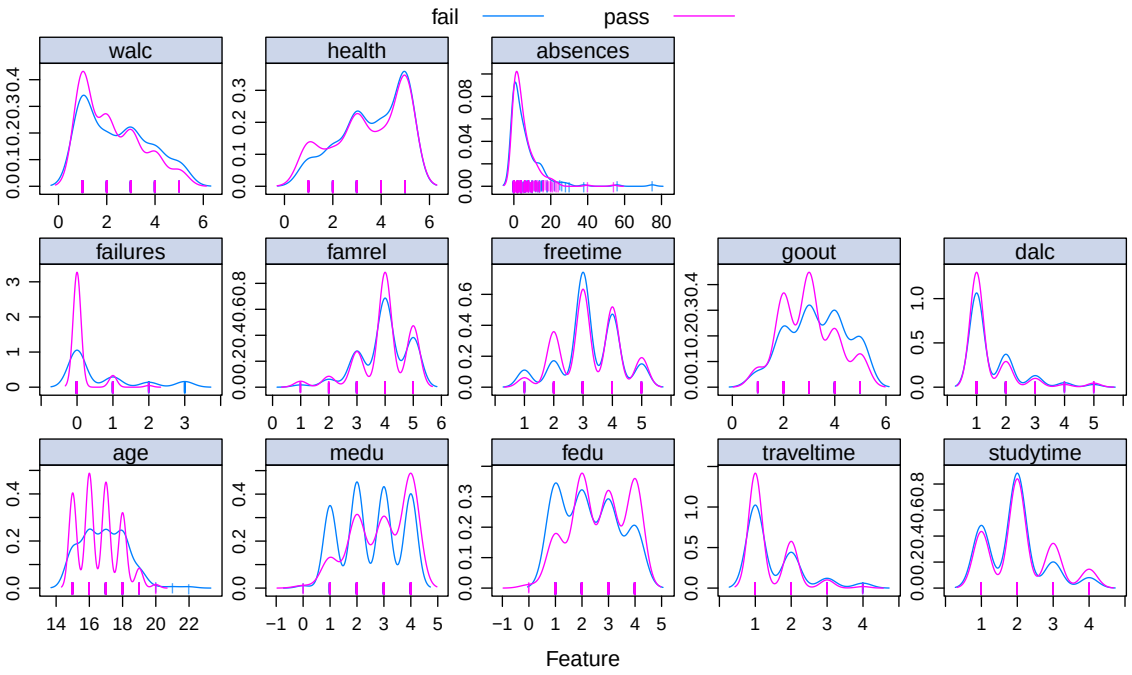
continuous variable plot

```
# df of predictors

x_corr = student_df %>%
  dplyr::select(-school, -sex, -address,-famsize, -pstatus, -mjob, -fjob, -reason, -guardian, -schoolsu


x = student_df %>%
  dplyr::select(absences)

# vector of response
y_corr = student_df$letter_grade

theme1 <- trellis.par.get()
theme1$plot.symbol$col <- rgb(.2, .4, .2, .5)
```

```
theme1$plot.symbol$pch <- 16
theme1$plot.line$col <- rgb(.8, .1, .1, 1)
theme1$plot.line$lwd <- 2
theme1$strip.background$col <- rgb(.0, .2, .6, .2)
trellis.par.set(theme1)
plot1 = featurePlot(x_corr, y_corr, plot = "boxplot", labels = c("","Y"),
            type = c("p"), layout = c(4, 3))
plot2 = featurePlot(x, y_corr, plot = "boxplot", labels = c("","Y"),
            type = c("p"), layout = c(1,1))

plot3 = featurePlot(dplyr::select_if(student_df, is.numeric),
                    y = y_corr,
                    scales = list(x =list(relation ="free"),
                                  y =list(relation ="free")),
                    plot = "density", pch = "|",
                    auto.key = list(columns = 2))
library(cowplot)
plot_grid(plot1, plot2)
```

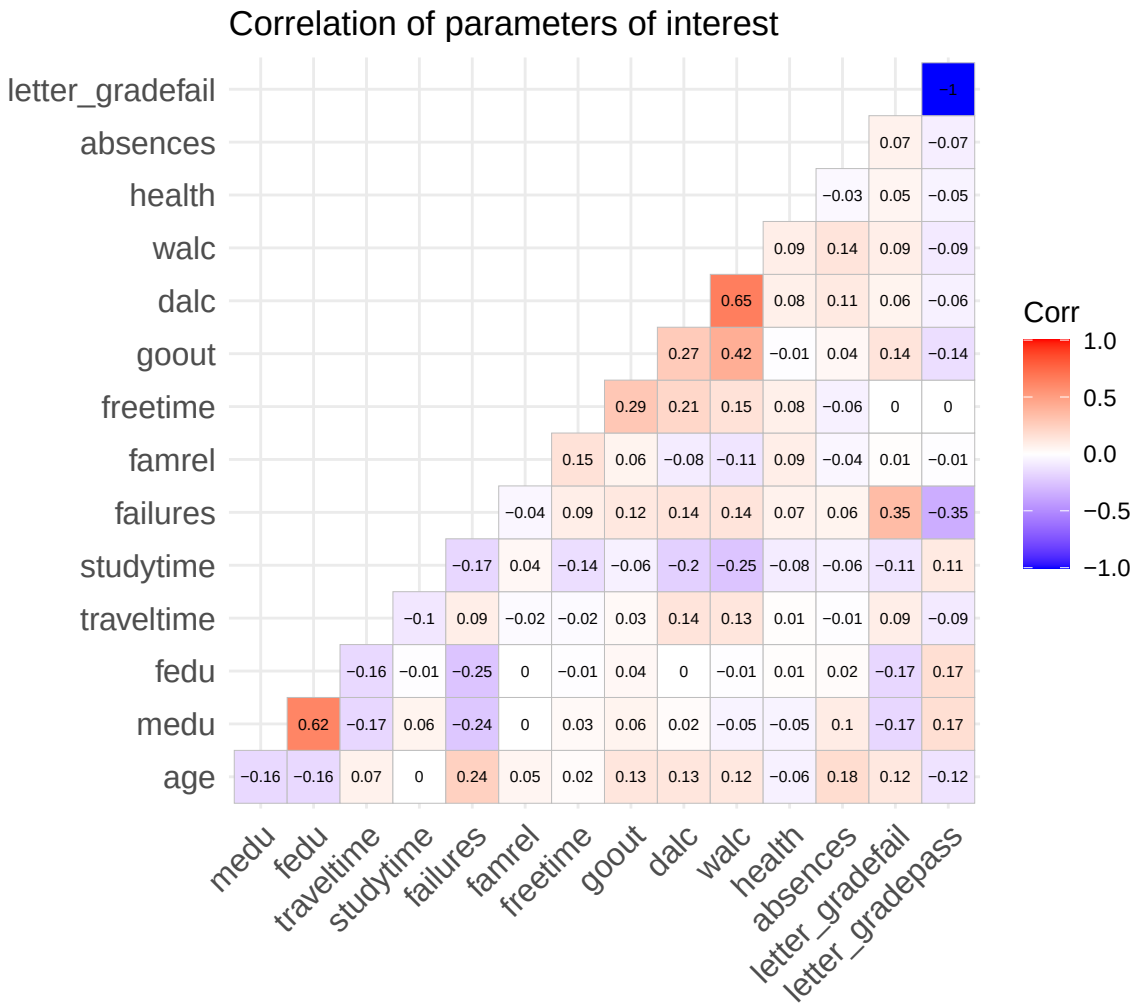



```
plot3
```

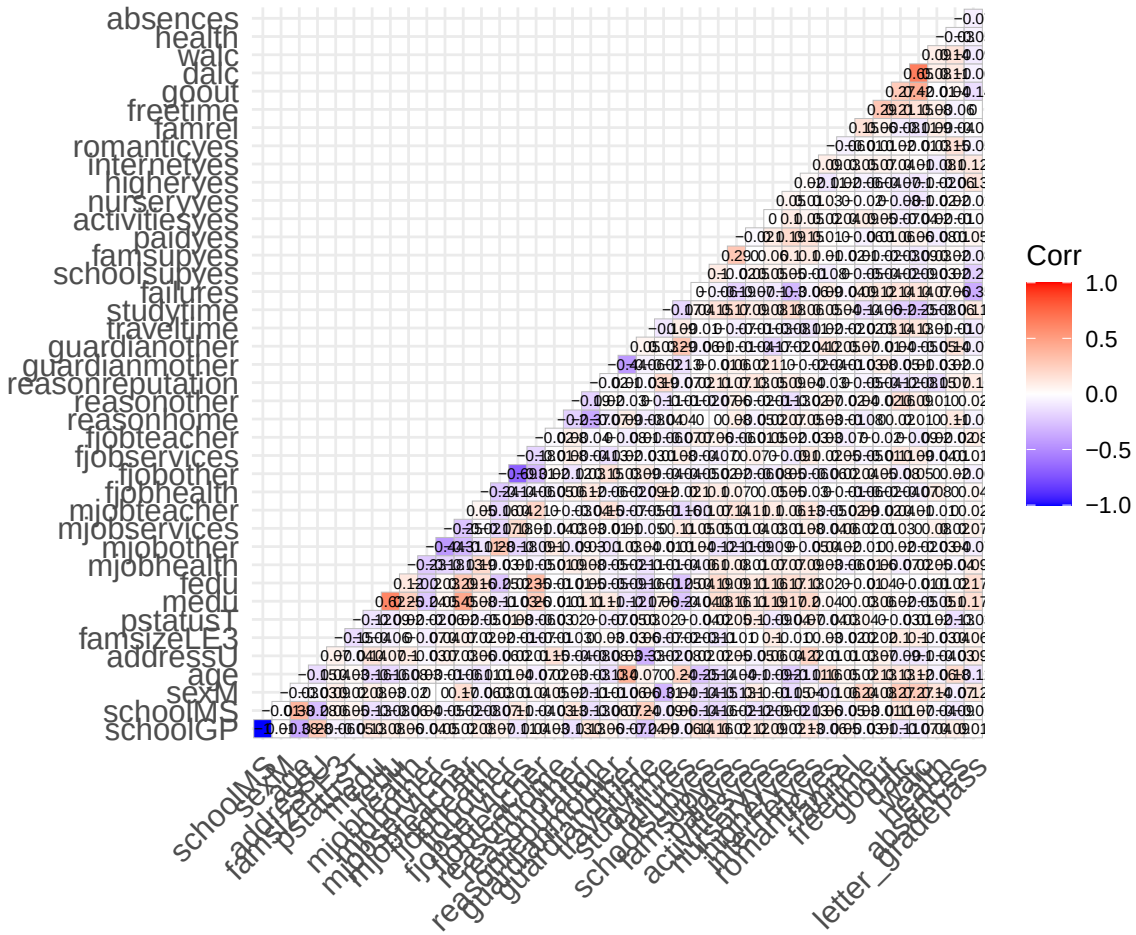


Correlation plot

```
library(ggcorrplot)

corr_df =
  student_df %>%
  select(age, medu, fedu, traveltime, studytime, failures, famrel, freetime, goout, dalc, walc, health,

# correlation plot with only continous var
model.matrix(~0+., data = corr_df) %>%
  cor(use = "pairwise.complete.obs") %>%
   ggcorrplot(show.diag = F, type="lower",
              lab=TRUE, lab_size=2,
              title = "Correlation of parameters of interest")
```




```
# correlation plot with all variables
model.matrix(~0+., data = student_df) %>%
  cor(use = "pairwise.complete.obs") %>%
   ggcorrplot(show.diag = F, type="lower",
              lab=TRUE, lab_size=2,
              title = "Correlation of parameters of interest")
```


Correlation of parameters of interest
absences
health
walc
dalc
goout
freetime
famrel
romanticyes
internetyes
higheryes
nurseryyes
activitiesyes
paidyes
famsupyes
schoolsupyes
failures
studytime
traveltime
guardianother
guardianmother
reasonreputation
reasonother
reasonhome
fjobteacher
fjobservices
fjobother
fjobhealth
mjobteacher
mjobservices
mjobother
mjobhealth
fedu
medu
pstatusT
famsizeLE3
addressU
age
sexM
schoolMS
schoolGP
Corr
1.0
0.5
0.0
−0.5
−1.0