

HW2

R code

數據一 謝宗佑

2025-03-15

Import package

```
library(table1)
library(dplyr)
library(purrr)
library(ggplot2)
library(GGally) # ggpair func. package
library(Hmisc) # describe func. package
```

Varibale definition

Variable	Data Type	Definition	Note
family	Nominal	String of the name of the family of mushroom species	23 types
name	Nominal	String of the mushroom species	
class	Nominal	Edibility classification (binary, p = poisonous(0), e = edible(1))	binary ,p = poisonous(0), e = edible(1)
cap-diameter	Metrical	Float number(s) in cm (two values = min max, one value = mean)	two variables (min-max) or one variable (mean)
cap-shape	Nominal	bell=b, conical=c, convex=x, flat=f, sunken=s, spherical=p, others=o	
cap-surface	Nominal	fibrous=i, grooves=g, scaly=y, smooth=s, shiny=h, leathery=l, silky=k, sticky=t, wrinkled=w, fleshy=e	

Variable	Data Type	Definition	Note
cap-color	Nominal	brown=n, buff=b, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y, blue=l, orange=o, black=k	
does-bruise-bleed	Nominal	bruises-or-bleeding=t, no=f	binary ,t = yes(1), f = no(0)
gill-attachment	Nominal	adnate=a, adnexed=x, decurrent=d, free=e, sinuate=s, pores=p, none=f	
gill-spacing	Nominal	close=c, distant=d, none=f	
gill-color	Nominal	See cap-color + none=f	same as cap-color, with 'none = f'
stem-height	Metrical	Float number(s) in cm (two values = min max, one value = mean)	two variables (min-max) or one variable (mean)
stem-width	Metrical	Float number(s) in mm (two values = min max, one value = mean)	two variables (min-max) or one variable (mean)
stem-root	Nominal	bulbous=b, swollen=s, club=c, cup=u, equal=e, rhizomorphs=z, rooted=r	
stem-surface	Nominal	See cap-surface + none=f	same as cap- surface, with 'none = f'
stem-color	Nominal	See cap-color + none=f	same as cap-color, with 'none = f'
veil-type	Nominal	partial=p, universal=u	

Variable	Data Type	Definition	Note
veil-color	Nominal	See cap-color + none=f	same as cap-color, with 'none = f'
has-ring	Nominal	ring=t, none=f	binary ,t = yes(1), f = no(0)
ring-type	Nominal	cobwebby=c, evanescent=e, flaring=r, grooved=g, large=l, pendant=p, sheathing=s, zone=z, scaly=y, movable=m, none=f	
spore-print-color	Nominal	See cap-color	
habitat	Nominal	grasses=g, leaves=l, meadows=m, paths=p, heaths=h, urban=u, waste=w, woods=d	
season	Nominal	spring=s, summer=u, autumn=a, winter=w	

Read csv

```
dat <- read.csv("D:/HW2/mushroom/primary_data.csv", sep = ";")
```

Describe(not processed)

```
# sink("describe_output.html")
# html(describe(dat), descript = "Descriptive Statistics",
# file = '', caption.placement = "top")
# sink()
latex(describe(dat), descript = "Descriptive Statistics",
file = '', caption.placement = "top")
```

23 Variables ^{dat} 173 Observations

family111			
n	missing	distinct		
173	0	23		
lowest : Amanita Family	Bolbitius Family	Bolete Family	Bracket Fungi	Chanterelle Family
highest: Russula Family	Saddle-Cup Family	Stropharia Family	Tricholoma Family	Wax Gill Family

name

n	missing	distinct
173	0	173

lowest :	Amethyst Deceiver	Aniseed Funnel Cap	Apricot Fungus	Bare-toothed Russula	Bay Bolete
highest:	Yellow-gilled Russula	Yellow-staining Mushroom	Yellow-stemmed Bell Cap	Yellow Swamp Russula	Yellow Wax cap

class

n	missing	distinct
173	0	2

Value	e	p
Frequency	77	96
Proportion	0.445	0.555

cap.diameter

n	missing	distinct
173	0	51

lowest :	[0.4, 1]	[0.5, 1.5]	[0.5, 1]	[0.7, 1.3]	[1, 1.5]
highest:	[8, 14]	[8, 15]	[8, 20]	[8, 25]	[8, 30]

cap.shape

n	missing	distinct
173	0	27

lowest :	[b, f, s]	[b, f]	[b, x, f]	[b, x]	[b]
highest:	[x, f]	[x, o]	[x, p]	[x, s]	[x]

Cap.surface

n	missing	distinct
133	40	40

lowest :	[d, e, y, i]	[d, k, s]	[d, k]	[d, s]	[d]
highest:	[t]	[w, t]	[w]	[y, s]	[y]

cap.color

n	missing	distinct
173	0	67

lowest :	[b, p, e, y]	[b, u]	[b]	[e, n, p, w]	[e, n, y]
highest:	[y, n]	[y, o, g, n, r]	[y, o, r, n]	[y, o]	[y]

does.bruise.or.bleed

n	missing	distinct
173	0	2

Value	[f]	[t]
Frequency	143	30
Proportion	0.827	0.173

gill.attachment

n	missing	distinct
145	28	8

Value	[a, d]	[a]	[d]	[e]	[f]	[p]	[s]	[x]
Frequency	8	32	25	16	10	17	16	21
Proportion	0.055	0.221	0.172	0.110	0.069	0.117	0.110	0.145

gill.spacing

	n	missing	distinct
	102	71	3

Value	[c]	[d]	[f]
Frequency	70	22	10
Proportion	0.686	0.216	0.098

gill.color

	n	missing	distinct
	173	0	59

lowest :	[b, p, w]	[b, u]	[b]	[e]	[f]
highest:	[y, o, e]	[y, r, k]	[y, r]	[y, w]	[y]

stem.height

	n	missing	distinct
	173	0	46

lowest :	[0]	[1, 2]	[1, 3]	[10, 12]	[10, 15]	highest:	[8, 12]	[8, 15]	[8, 20]	[8, 25]	[8, 30]
----------	-----	--------	--------	----------	----------	----------	---------	---------	---------	---------	---------

stem.width

	n	missing	distinct
	173	0	48

lowest :	[0.5, 1]	[0]	[1, 2]	[1, 3]	[1]	highest:	[7, 15]	[8, 12]	[8, 15]	[8, 18]	[8, 20]
----------	----------	-----	--------	--------	-----	----------	---------	---------	---------	---------	---------

stem.root

	n	missing	distinct
	27	146	5

Value	[b]	[c]	[f]	[r]	[s]
Frequency	9	2	3	4	9
Proportion	0.333	0.074	0.111	0.148	0.333

stem.surface

	n	missing	distinct
	65	108	14

Value	[f]	[g]	[h]	[i, s]	[i, t]	[i, y]	[i]	[k, s]	[k]	[s, h]	[s]	[t]
Frequency	3	5	1	1	1	1	11	1	4	1	15	7
Proportion	0.046	0.077	0.015	0.015	0.015	0.015	0.169	0.015	0.062	0.015	0.231	0.108

Value	[y, s]	[y]
Frequency	1	13
Proportion	0.015	0.200

stem.color

	n	missing	distinct
	173	0	41

lowest :	[b, u]	[e, n]	[e, u, y]	[e, y]	[e]
highest:	[w]	[y, e, n]	[y, n]	[y, o, k]	[y]

veil.type

	n	missing	distinct	value
	9	164	1	[u]

Value	[u]
Frequency	9
Proportion	1

veil.color

n missing distinct
21 152 7

Value	[e, n]	[k]	[n]	[u]	[w]	[y, w]	[y]
Frequency	1	1	1	1	15	1	1
Proportion	0.048	0.048	0.048	0.048	0.714	0.048	0.048

has.ring

n missing distinct
173 0 2

Value	[f]	[t]
Frequency	130	43
Proportion	0.751	0.249

ring.type

n missing distinct
166 7 13

Value	[e, g]	[e]	[f]	[g, p]	[g]	[l, e]	[l, p]	[l, r]	[l]	[m]	[p]	[r]
Frequency	1	6	137	2	2	1	1	2	2	1	2	3
Proportion	0.006	0.036	0.825	0.012	0.012	0.006	0.006	0.012	0.012	0.006	0.012	0.018

Value	[z]
Frequency	6
Proportion	0.036

Spore.print.color

n missing distinct
18 155 8

Value	[g]	[k, r]	[k, u]	[k]	[n]	[p, w]	[p]	[w]
Frequency	1	1	1	5	3	1	3	3
Proportion	0.056	0.056	0.056	0.278	0.167	0.056	0.167	0.167

habitat

n missing distinct
173 0 21

lowest :	[d, h]	[d]	[g, d, h]	[g, d]	[g, h, d]
highest:	[m, d]	[m, h]	[m]	[p, d]	[w]

season

n missing distinct
173 0 10

Value	[a, w]	[a]	[s, a, w]	[s, u, a, w]	[s, u, a]	[s, u]
Frequency	15	16	1	13	5	3
Proportion	0.087	0.092	0.006	0.075	0.029	0.017

Value	[s]	[u, a, w]	[u, a]	[u]
Frequency	1	12	106	1
Proportion	0.006	0.069	0.613	0.006

Table1(not processed)

```
table1(~ .|class, data = subset(dat, select = -c(name))) # only can use subset
```

	e	p	Overall
	(N=77)	(N=96)	(N=173)
family			
Amanita Family	3 (3.9%)	5 (5.2%)	8 (4.6%)
Bolbitius Family	1 (1.3%)	2 (2.1%)	3 (1.7%)
Bolete Family	11 (14.3%)	3 (3.1%)	14 (8.1%)
Bracket Fungi	1 (1.3%)	6 (6.3%)	7 (4.0%)
Chanterelle Family	3 (3.9%)	0 (0%)	3 (1.7%)
Entoloma Family	1 (1.3%)	6 (6.3%)	7 (4.0%)
Hydnum Family	1 (1.3%)	0 (0%)	1 (0.6%)
Ink Cap Family	6 (7.8%)	7 (7.3%)	13 (7.5%)
Lepiota Family	2 (2.6%)	1 (1.0%)	3 (1.7%)
Morel Family	1 (1.3%)	0 (0%)	1 (0.6%)
Mushroom Family	4 (5.2%)	1 (1.0%)	5 (2.9%)
Oyster Mushroom Family	2 (2.6%)	0 (0%)	2 (1.2%)
Pluteus Family	2 (2.6%)	0 (0%)	2 (1.2%)
Russula Family	11 (14.3%)	16 (16.7%)	27 (15.6%)
Stropharia Family	1 (1.3%)	7 (7.3%)	8 (4.6%)
Tricholoma Family	23 (29.9%)	20 (20.8%)	43 (24.9%)
Wax Gill Family	4 (5.2%)	4 (4.2%)	8 (4.6%)
Cortinarius Family	0 (0%)	11 (11.5%)	11 (6.4%)
Crepidotus Family	0 (0%)	1 (1.0%)	1 (0.6%)
Ear-Pick Family	0 (0%)	1 (1.0%)	1 (0.6%)
Jelly Discs Family	0 (0%)	1 (1.0%)	1 (0.6%)
Paxillus Family	0 (0%)	3 (3.1%)	3 (1.7%)
Saddle-Cup Family	0 (0%)	1 (1.0%)	1 (0.6%)
cap.diameter			
[0.5, 1.5]	2 (2.6%)	1 (1.0%)	3 (1.7%)
[1, 2]	3 (3.9%)	4 (4.2%)	7 (4.0%)
[1, 4]	2 (2.6%)	2 (2.1%)	4 (2.3%)
[10, 25]	1 (1.3%)	1 (1.0%)	2 (1.2%)
[12, 18]	1 (1.3%)	0 (0%)	1 (0.6%)
[12, 25]	1 (1.3%)	0 (0%)	1 (0.6%)
[2, 10]	1 (1.3%)	0 (0%)	1 (0.6%)
[2, 4]	1 (1.3%)	7 (7.3%)	8 (4.6%)
[2, 5]	7 (9.1%)	9 (9.4%)	16 (9.2%)
[2, 6]	2 (2.6%)	8 (8.3%)	10 (5.8%)
[2, 7]	1 (1.3%)	1 (1.0%)	2 (1.2%)
[2, 8]	1 (1.3%)	0 (0%)	1 (0.6%)
[3, 10]	4 (5.2%)	1 (1.0%)	5 (2.9%)
[3, 5]	1 (1.3%)	1 (1.0%)	2 (1.2%)
[3, 6]	4 (5.2%)	2 (2.1%)	6 (3.5%)
[3, 7]	1 (1.3%)	3 (3.1%)	4 (2.3%)
[3, 8]	2 (2.6%)	4 (4.2%)	6 (3.5%)
[4, 10]	4 (5.2%)	4 (4.2%)	8 (4.6%)
[4, 12]	3 (3.9%)	2 (2.1%)	5 (2.9%)
[4, 8]	5 (6.5%)	4 (4.2%)	9 (5.2%)
[4, 9]	2 (2.6%)	1 (1.0%)	3 (1.7%)
[5, 10]	4 (5.2%)	8 (8.3%)	12 (6.9%)
[5, 12]	3 (3.9%)	3 (3.1%)	6 (3.5%)
[5, 15]	7 (9.1%)	1 (1.0%)	8 (4.6%)
[5, 18]	1 (1.3%)	0 (0%)	1 (0.6%)
[5, 20]	1 (1.3%)	0 (0%)	1 (0.6%)
[50]	1 (1.3%)	0 (0%)	1 (0.6%)
[6, 10]	1 (1.3%)	1 (1.0%)	2 (1.2%)
[6, 12]	2 (2.6%)	3 (3.1%)	5 (2.9%)
[6, 14]	1 (1.3%)	0 (0%)	1 (0.6%)
[7, 15]	2 (2.6%)	1 (1.0%)	3 (1.7%)

Data processing method

```
continuous_process <- function(column, name){
  # column = dat$cap.diameter
  # name = "stem.height"
  column_ = column %>%
    gsub(pattern = "\\[|\\]", replacement = "")%>%
    strsplit(", ")

  n_element <- function(x, n = 2){
    length(x) == n
  }
  names = paste(name, c("min", "max", "mean"), sep = "_")

  output = data.frame(min = rep(0,173)) %>%
    mutate(min = map_if(column_, n_element, ~ .x[[1]], .else = ~ NA) %>% unlist %>% as.numeric(),
           max = map_if(column_, n_element, ~ .x[[2]], .else = ~ NA) %>% unlist %>% as.numeric(),
           # mean= map_if(column_, ~ n_element(.x), ~ mean(as.numeric(.x)), .else = ~ as.numeric(.x)) %>%
           mean = map_if(column_, n_element, ~ NA, .else = ~ as.numeric(.x)) %>% unlist()) %>%
    setNames(names)
  return(output)
}

continuous = dat %>% select(cap.diameter, stem.height, stem.width)
continuous_finish = mapply(continuous_process, continuous, names(continuous),SIMPLIFY = FALSE)
# continuous_finish

#####
make_dummy <- function(column, name){
  # column = dat$cap.shape
  # name = "cap.shape"
  column_ = column %>%
    gsub(pattern = "\\[|\\]", replacement = "")%>%
    strsplit(", ")
  types = column_%>%
    unlist()%>%
    unique
  names = paste0(rep(name,length(types)), "_", types)

  check_type <- function(list, type){
    as.integer(type %in% list)
  }

  output = lapply(types, function(type){
    rapply(column_, check_type, type = type)
  }) %>%
    do.call(what = cbind)%>%
    as.data.frame %>%
    setNames(names)

  return(output)
}
```



```

categorical = dat %>% select(-c(family , name, does.bruise.or.bleed, has.ring, class, cap.diameter, stem
categorical_finish = mapply(make_dummy, categorical, names(categorical))
# categorical_finish

dat$class <- ifelse(dat$class == "p", 0, 1)
dat$does.bruise.or.bleed <- ifelse(dat$does.bruise.or.bleed == "[f]", 0, 1)
dat$has.ring <- ifelse(dat$has.ring == "[f]", 0, 1)

new_data <- bind_cols(data.frame(family = dat$family, name = dat$name,
                                class = dat$class,
                                does.bruise.or.bleed = dat$does.bruise.or.bleed,
                                has.ring = dat$has.ring),
                      continuous_finish, categorical_finish)

new_data <- new_data %>% mutate(across(-c("cap.diameter_min","cap.diameter_max","cap.diameter_mean",
                                           "stem.width_min","stem.width_max","stem.width_mean"), as.factor))

head(new_data)

```

	family	name	class	does.bruise.or.bleed	has.ring	
1	Amanita Family	Fly Agaric	0	0	1	
2	Amanita Family	Panther Cap	0	0	1	
3	Amanita Family	False Panther Cap	0	0	1	
4	Amanita Family	The Blusher	1	1	1	
5	Amanita Family	Death Cap	0	0	1	
6	Amanita Family	False Death Cap	1	0	1	
	cap.diameter_min	cap.diameter_max	cap.diameter_mean	stem.height_min		
1	10	20	NA	15		
2	5	10	NA	6		
3	10	15	NA	10		
4	5	15	NA	7		
5	5	12	NA	10		
6	4	9	NA	5		
	stem.height_max	stem.height_mean	stem.width_min	stem.width_max		
1	20	NA	15	20		
2	10	NA	10	20		
3	12	NA	10	20		
4	15	NA	10	25		
5	12	NA	10	20		
6	7	NA	10	15		
	stem.width_mean	cap.shape_x	cap.shape_f	cap.shape_p	cap.shape_b	cap.shape_c
1	NA	1	1	0	0	0
2	NA	1	0	1	0	0
3	NA	1	1	0	0	0
4	NA	1	1	0	0	0
5	NA	1	1	0	0	0
6	NA	1	0	0	0	0
	cap.shape_s	cap.shape_o	Cap.surface_g	Cap.surface_h	Cap.surface_t	
1	0	0	1	1	0	
2	0	0	1	0	0	
3	0	0	0	0	0	
4	0	0	0	0	0	

5	0	0	0	1	0	
6	0	0	0	0	0	
	Cap.surface_y	Cap.surface_e	Cap.surface_s	Cap.surface_l	Cap.surface_d	
1	0	0	0	0	0	
2	0	0	0	0	0	
3	0	0	0	0	0	
4	0	0	0	0	0	
5	0	0	0	0	0	
6	0	0	0	0	0	
	Cap.surface_w	Cap.surface_i	Cap.surface_k	cap.color_e	cap.color_o	cap.color_n
1	0	0	0	1	1	0
2	0	0	0	0	0	1
3	0	0	0	0	0	1
4	0	0	0	0	0	1
5	0	0	0	0	0	0
6	0	0	0	0	0	0
	cap.color_g	cap.color_r	cap.color_w	cap.color_y	cap.color_p	cap.color_b
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	1	0	0	0	0	0
4	0	0	0	0	0	0
5	0	1	0	0	0	0
6	0	0	1	1	0	0
	cap.color_u	cap.color_l	cap.color_n	,w	cap.color_k	gill.attachment_e
1	0	0	0	0		1
2	0	0	0	0		1
3	0	0	0	0		1
4	0	0	0	0		0
5	0	0	0	0		0
6	0	0	0	0		1
	gill.attachment_a	gill.attachment_d	gill.attachment_s	gill.attachment_x		
1	0	0	0	0		0
2	0	0	0	0		0
3	0	0	0	0		0
4	0	0	0	0		0
5	0	0	0	0		0
6	0	0	0	0		0
	gill.attachment_p	gill.attachment_f	gill.spacing_c	gill.spacing_d		
1	0	0	0	0		0
2	0	0	0	0		0
3	0	0	0	0		0
4	0	0	0	0		0
5	0	0	1	0		0
6	0	0	0	0		0
	gill.spacing_f	gill.color_w	gill.color_n	gill.color_p	gill.color_u	
1	0	1	0	0	0	
2	0	1	0	0	0	
3	0	1	0	0	0	
4	0	1	0	0	0	
5	0	1	0	0	0	
6	0	1	0	0	0	
	gill.color_b	gill.color_g	gill.color_y	gill.color_r	gill.color_e	gill.color_o
1	0	0	0	0	0	0
2	0	0	0	0	0	0

3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0
6	0	0	0	0	0	0
	gill.color_k	gill.color_f	stem.root_s	stem.root_b	stem.root_r	stem.root_c
1	0	0	1	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	1	0	0
5	0	0	0	0	0	0
6	0	0	0	1	0	0
	stem.root_f	stem.surface_y	stem.surface_s	stem.surface_k	stem.surface_i	
1	0	1	0	0	0	
2	0	1	0	0	0	
3	0	0	0	0	0	
4	0	0	0	0	0	
5	0	0	0	0	0	
6	0	0	0	0	0	
	stem.surface_h	stem.surface_t	stem.surface_g	stem.surface_f	stem.color_w	
1	0	0	0	0	1	
2	0	0	0	0	1	
3	0	0	0	0	1	
4	0	0	0	0	1	
5	0	0	0	0	1	
6	0	0	0	0	1	
	stem.color_y	stem.color_n	stem.color_b	stem.color_u	stem.color_l	stem.color_r
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0
6	1	0	0	0	0	0
	stem.color_p	stem.color_e	stem.color_k	stem.color_g	stem.color_o	
1	0	0	0	0	0	
2	0	0	0	0	0	
3	0	0	0	0	0	
4	0	0	0	0	0	
5	0	0	0	0	0	
6	0	0	0	0	0	
	stem.color_w	stem.color_f	veil.type_u	veil.color_w	veil.color_y	veil.color_e
1	0	0	1	1	0	0
2	0	0	1	1	0	0
3	0	0	1	1	0	0
4	0	0	1	1	0	0
5	0	0	1	1	0	0
6	0	0	1	1	1	0
	veil.color_n	veil.color_u	veil.color_k	ring.type_g	ring.type_p	ring.type_e
1	0	0	0	1	1	0
2	0	0	0	0	1	0
3	0	0	0	1	0	1
4	0	0	0	1	0	0
5	0	0	0	1	1	0
6	0	0	0	1	0	0
	ring.type_l	ring.type_f	ring.type_m	ring.type_r	ring.type_z	

1	0	0	0	0	0		
2	0	0	0	0	0		
3	0	0	0	0	0		
4	0	0	0	0	0		
5	0	0	0	0	0		
6	0	0	0	0	0		
Spore.print.color_w Spore.print.color_p Spore.print.color_k							
1	0	0	0	0	0		
2	0	0	0	0	0		
3	0	0	0	0	0		
4	0	0	0	0	0		
5	0	0	0	0	0		
6	0	0	0	0	0		
Spore.print.color_r Spore.print.color_u Spore.print.color_n							
1	0	0	0	0	0		
2	0	0	0	0	0		
3	0	0	0	0	0		
4	0	0	0	0	0		
5	0	0	0	0	0		
6	0	0	0	0	0		
Spore.print.color_g habitat_d habitat_m habitat_g habitat_h habitat_l							
1	0	1	0	0	0	0	
2	0	1	0	0	0	0	
3	0	1	0	0	0	0	
4	0	1	0	0	0	0	
5	0	1	0	0	0	0	
6	0	1	0	0	0	0	
habitat_p habitat_w habitat_u season_u season_a season_w season_s							
1	0	0	0	1	1	1	0
2	0	0	0	1	1	0	0
3	0	0	0	1	1	0	0
4	0	0	0	1	1	0	0
5	0	0	0	1	1	0	0
6	0	0	0	1	1	0	0

Describe(processed)

```
# sink("describe_output.html")
# html(describe(new_data), descript = "Descriptive Statistics",
# file = '', caption.placement = "top")
# sink()
latex(describe(new_data), descript = "Descriptive Statistics",
file = '', caption.placement = "top")
```

new_data
128 Variables 173 Observations

family

n	missing	distinct
173	0	23

lowest : Amanita Family	Bolbitius Family	Bolete Family	Bracket Fungi	Chanterelle Family
highest: Russula Family	Saddle-Cup Family	Stropharia Family	Tricholoma Family	Wax Gill Family

name

n	missing	distinct
173	0	173

lowest : Amethyst Deceiver	Aniseed Funnel Cap	Apricot Fungus	Bare-toothed Russula	Bay Bolete
highest: Yellow-gilled Russula	Yellow-staining Mushroom	Yellow-stemmed Bell Cap	Yellow Swamp Russula	Yellow Wax cap

class

n	missing	distinct
173	0	2

Value	0	1
Frequency	96	77
Proportion	0.555	0.445

does.bruise.or.bleed

n	missing	distinct
173	0	2

Value	0	1
Frequency	143	30
Proportion	0.827	0.173

has.ring

n	missing	distinct
173	0	2

Value	0	1
Frequency	130	43
Proportion	0.751	0.249

cap.diameter_min

n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
172	1	13	0.976	3.776	3.5	2.533	1	1	2	3	5	7	8

Value	0.4	0.5	0.7	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	10.0	12.0
Frequency	2	4	1	17	39	24	26	29	11	4	9	4	2
Proportion	0.012	0.023	0.006	0.099	0.227	0.140	0.151	0.169	0.064	0.023	0.052	0.023	0.012

For the frequency table, variable is rounded to the nearest 0

cap.diameter_max

n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
172	1	19	0.991	9.199	8.5	6.147	2	3	5	8	12	15	20

Value	1.0	1.3	1.5	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10.0	12.0	14.0
Frequency	3	1	4	7	6	12	18	16	7	16	3	28	18	3
Proportion	0.017	0.006	0.023	0.041	0.035	0.070	0.105	0.093	0.041	0.093	0.017	0.163	0.105	0.017

Value	15.0	18.0	20.0	25.0	30.0
Frequency	15	3	5	5	2
Proportion	0.087	0.017	0.029	0.029	0.012

For the frequency table, variable is rounded to the nearest 0

cap.diameter_mean

n	missing	distinct	Info	Mean
1	172	1	0	50

Value	50
Frequency	1
Proportion	1

stem.height_min

n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
170	3	11	0.955	4.382	4	2.157	2	2	3	4	5	7	8

Value	1	2	3	4	5	6	7	8	10	12	15
Frequency	2	21	38	52	24	15	3	7	5	1	2
Proportion	0.012	0.124	0.224	0.306	0.141	0.088	0.018	0.041	0.029	0.006	0.012

For the frequency table, variable is rounded to the nearest 0

stem.height_max

n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
170	3	18	0.976	9.029	8.5	4.205	4.45	5.00	6.00	8.00	10.00	15.00	15.00

Value	2	3	4	5	6	7	8	9	10	11	12	14	15	18
Frequency	1	2	6	14	25	16	37	2	35	1	12	1	10	1
Proportion	0.006	0.012	0.035	0.082	0.147	0.094	0.218	0.012	0.206	0.006	0.071	0.006	0.059	0.006

Value	20	25	30	35
Frequency	4	1	1	1
Proportion	0.024	0.006	0.006	0.006

For the frequency table, variable is rounded to the nearest 0

stem.height_mean

n	missing	distinct	Info	Mean
3	170	1	0	0

Value	0
Frequency	3
Proportion	1

stem.width_min

n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
162	11	15	0.98	8.83	8	6.785	2	2	4	8	10	20	20

Value	0.5	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	10.0	12.0	15.0	20.0	30.0
Frequency	1	6	17	12	12	19	7	1	10	38	1	20	16	1
Proportion	0.006	0.037	0.105	0.074	0.074	0.117	0.043	0.006	0.062	0.235	0.006	0.123	0.099	0.006

Value	40.0
Frequency	1
Proportion	0.006

For the frequency table, variable is rounded to the nearest 0

stem.width_max

n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
162	11	20	0.991	16.58	15	13.51	3	4	8	15	20	30	40

Value	1	2	3	4	5	6	7	8	10	12	15	18	20	25
Frequency	1	5	10	9	5	3	3	17	15	11	19	4	26	10
Proportion	0.006	0.031	0.062	0.056	0.031	0.019	0.019	0.105	0.093	0.068	0.117	0.025	0.160	0.062

Value	30	40	50	60	80	100
Frequency	11	8	1	2	1	1
Proportion	0.068	0.049	0.006	0.012	0.006	0.006

For the frequency table, variable is rounded to the nearest 0

stem.width_mean

n	missing	distinct	Info	Mean	pMedian	Gmd
11	162	4	0.918	4.091	5	5.055

Value	0	1	2	10
Frequency	3	3	1	4
Proportion	0.273	0.273	0.091	0.364

For the frequency table, variable is rounded to the nearest 0

cap.shape_x

n	missing	distinct
173	0	2

Value	0	1
Frequency	63	110
Proportion	0.364	0.636

cap.shape_f

n	missing	distinct
173	0	2

Value	0	1
Frequency	99	74
Proportion	0.572	0.428

cap.shape_p

n	missing	distinct
173	0	2

Value	0	1
Frequency	158	15
Proportion	0.913	0.087

cap.shape_b

n	missing	distinct
173	0	2

Value	0	1
Frequency	150	23
Proportion	0.867	0.133

cap.shape_c

n	missing	distinct
173	0	2

Value	0	1
Frequency	165	8
Proportion	0.954	0.046

cap.shape_s

n	missing	distinct
173	0	2

Value	0	1
Frequency	137	36
Proportion	0.792	0.208

cap.shape_o

n	missing	distinct
173	0	2

Value	0	1
Frequency	161	12
Proportion	0.931	0.069

Cap.surface_g

n	missing	distinct
173	0	2

Value	0	1
Frequency	157	16
Proportion	0.908	0.092

Cap.surface_h

n	missing	distinct
173	0	2

Value	0	1
Frequency	147	26
Proportion	0.85	0.15

Cap.surface_t

n	missing	distinct
173	0	2

Value	0	1
Frequency	136	37
Proportion	0.786	0.214

Cap.surface_y

n	missing	distinct
173	0	2

Value	0	1
Frequency	150	23
Proportion	0.867	0.133

Cap.surface_e

n	missing	distinct
173	0	2

Value	0	1
Frequency	162	11
Proportion	0.936	0.064

Cap.surface_s

n	missing	distinct
173	0	2

Value	0	1
Frequency	140	33
Proportion	0.809	0.191

Cap.surface_l

n	missing	distinct
173	0	2

Value	0	1
Frequency	169	4
Proportion	0.977	0.023

Cap.surface_d

n	missing	distinct
173	0	2

Value	0	1
Frequency	155	18
Proportion	0.896	0.104

Cap.surface_w

n	missing	distinct
173	0	2

Value	0	1
Frequency	165	8
Proportion	0.954	0.046

Cap.surface_i

n	missing	distinct
173	0	2

Value	0	1
Frequency	164	9
Proportion	0.948	0.052

Cap.surface_k

n	missing	distinct
173	0	2

Value	0	1
Frequency	163	10
Proportion	0.942	0.058

cap.color_e

n	missing	distinct
173	0	2

Value	0	1
Frequency	148	25
Proportion	0.855	0.145

cap.color_o

n	missing	distinct
173	0	2

Value	0	1
Frequency	151	22
Proportion	0.873	0.127

cap.color_n

n	missing	distinct
173	0	2

Value	0	1
Frequency	64	109
Proportion	0.37	0.63

cap.color_g

n	missing	distinct
173	0	2

Value	0	1
Frequency	145	28
Proportion	0.838	0.162

cap.color_r

n	missing	distinct
173	0	2

Value	0	1
Frequency	160	13
Proportion	0.925	0.075

cap.color_w

n	missing	distinct
173	0	2

Value	0	1
Frequency	139	34
Proportion	0.803	0.197

cap.color_y

n	missing	distinct
173	0	2

Value	0	1
Frequency	129	44
Proportion	0.746	0.254

cap.color_p

n	missing	distinct
173	0	2

Value	0	1
Frequency	162	11
Proportion	0.936	0.064

cap.color_b

n	missing	distinct
173	0	2

Value	0	1
Frequency	166	7
Proportion	0.96	0.04

cap.color_u

n	missing	distinct
173	0	2

Value	0	1
Frequency	163	10
Proportion	0.942	0.058

cap.color_l

n	missing	distinct
173	0	2

Value	0	1
Frequency	167	6
Proportion	0.965	0.035

cap.color_n,w

n	missing	distinct
173	0	2

Value	0	1
Frequency	172	1
Proportion	0.994	0.006

cap.color_k

n	missing	distinct
173	0	2

Value	0	1
Frequency	164	9
Proportion	0.948	0.052

gill.attachment_e

n	missing	distinct
173	0	2

Value	0	1
Frequency	157	16
Proportion	0.908	0.092

gill.attachment_a

n	missing	distinct
173	0	2

Value	0	1
Frequency	133	40
Proportion	0.769	0.231

gill.attachment_d

n	missing	distinct
173	0	2

Value	0	1
Frequency	140	33
Proportion	0.809	0.191

gill.attachment_s

n	missing	distinct
173	0	2

Value	0	1
Frequency	157	16
Proportion	0.908	0.092

gill.attachment_x

n	missing	distinct
173	0	2

Value	0	1
Frequency	152	21
Proportion	0.879	0.121

gill.attachment_p

n	missing	distinct
173	0	2

Value	0	1
Frequency	156	17
Proportion	0.902	0.098

gill.attachment_f

n	missing	distinct
173	0	2

Value	0	1
Frequency	163	10
Proportion	0.942	0.058

gill.spacing_c

n	missing	distinct
173	0	2

Value	0	1
Frequency	103	70
Proportion	0.595	0.405

gill.spacing_d

n	missing	distinct
173	0	2

Value	0	1
Frequency	151	22
Proportion	0.873	0.127

gill.spacing_f

n	missing	distinct
173	0	2

Value	0	1
Frequency	163	10
Proportion	0.942	0.058

gill.color_w

n	missing	distinct
173	0	2

Value	0	1
Frequency	100	73
Proportion	0.578	0.422

gill.color_n

n	missing	distinct
173	0	2

Value	0	1
Frequency	126	47
Proportion	0.728	0.272

gill.color_p

n	missing	distinct
173	0	2

Value	0	1
Frequency	145	28
Proportion	0.838	0.162

gill.color_u

n	missing	distinct
173	0	2

Value	0	1
Frequency	166	7
Proportion	0.96	0.04

gill.color_b

n	missing	distinct
173	0	2

Value	0	1
Frequency	168	5
Proportion	0.971	0.029

gill.color_g

n	missing	distinct
173	0	2

Value	0	1
Frequency	150	23
Proportion	0.867	0.133

gill.color_y

n	missing	distinct
173	0	2

Value	0	1
Frequency	129	44
Proportion	0.746	0.254

gill.color_r

n	missing	distinct
173	0	2

Value	0	1
Frequency	165	8
Proportion	0.954	0.046

gill.color_e

n	missing	distinct
173	0	2

Value	0	1
Frequency	167	6
Proportion	0.965	0.035

gill.color_o

n	missing	distinct
173	0	2

Value	0	1
Frequency	160	13
Proportion	0.925	0.075

gill.color_k

n	missing	distinct
173	0	2

Value	0	1
Frequency	158	15
Proportion	0.913	0.087

gill.color_f

n	missing	distinct
173	0	2

Value	0	1
Frequency	163	10
Proportion	0.942	0.058

stem.root_s

n	missing	distinct
173	0	2

Value	0	1
Frequency	164	9
Proportion	0.948	0.052

stem.root_b

n	missing	distinct
173	0	2

Value	0	1
Frequency	164	9
Proportion	0.948	0.052

stem.root_r

n	missing	distinct
173	0	2

Value	0	1
Frequency	169	4
Proportion	0.977	0.023

stem.root_c

n	missing	distinct
173	0	2

Value	0	1
Frequency	171	2
Proportion	0.988	0.012

stem.root_f

n	missing	distinct
173	0	2

Value	0	1
Frequency	170	3
Proportion	0.983	0.017

stem.surface_y

n	missing	distinct
173	0	2

Value	0	1
Frequency	158	15
Proportion	0.913	0.087

stem.surface_s

n	missing	distinct
173	0	2

Value	0	1
Frequency	154	19
Proportion	0.89	0.11

stem.surface_k

n	missing	distinct
173	0	2

Value	0	1
Frequency	168	5
Proportion	0.971	0.029

stem.surface_i

n	missing	distinct
173	0	2

Value	0	1
Frequency	159	14
Proportion	0.919	0.081

stem.surface_h

n	missing	distinct
173	0	2

Value	0	1
Frequency	171	2
Proportion	0.988	0.012

stem.surface_t

n	missing	distinct
173	0	2

Value	0	1
Frequency	165	8
Proportion	0.954	0.046

stem.surface_g

n	missing	distinct
173	0	2

Value	0	1
Frequency	168	5
Proportion	0.971	0.029

stem.surface_f

n	missing	distinct
173	0	2

Value	0	1
Frequency	170	3
Proportion	0.983	0.017

stem.color_w

n	missing	distinct
173	0	2

Value	0	1
Frequency	100	73
Proportion	0.578	0.422

stem.color_y

n	missing	distinct
173	0	2

Value	0	1
Frequency	141	32
Proportion	0.815	0.185

stem.color_n

n	missing	distinct
173	0	2

Value	0	1
Frequency	103	70
Proportion	0.595	0.405

stem.color_b

n	missing	distinct
173	0	2

Value	0	1
Frequency	172	1
Proportion	0.994	0.006

stem.color_u

n	missing	distinct
173	0	2

Value	0	1
Frequency	166	7
Proportion	0.96	0.04

stem.color_l

n	missing	distinct
173	0	2

Value	0	1
Frequency	171	2
Proportion	0.988	0.012

stem.color_r

n	missing	distinct
173	0	2

Value	0	1
Frequency	169	4
Proportion	0.977	0.023

stem.color_p

n	missing	distinct
173	0	2

Value	0	1
Frequency	169	4
Proportion	0.977	0.023

stem.color_e

n	missing	distinct
173	0	2

Value	0	1
Frequency	162	11
Proportion	0.936	0.064

stem.color_k

n	missing	distinct
173	0	2

Value	0	1
Frequency	169	4
Proportion	0.977	0.023

stem.color_g

n	missing	distinct
173	0	2

Value	0	1
Frequency	159	14
Proportion	0.919	0.081

stem.color_o

n	missing	distinct
173	0	2

Value	0	1
Frequency	161	12
Proportion	0.931	0.069

stem.color_w

n	missing	distinct
173	0	2

Value	0	1
Frequency	172	1
Proportion	0.994	0.006

stem.color_f

n	missing	distinct
173	0	2

Value	0	1
Frequency	170	3
Proportion	0.983	0.017

veil.type_u

n	missing	distinct
173	0	2

Value	0	1
Frequency	164	9
Proportion	0.948	0.052

veil.color_w

n	missing	distinct
173	0	2

Value	0	1
Frequency	157	16
Proportion	0.908	0.092

veil.color_y

n	missing	distinct
173	0	2

Value	0	1
Frequency	171	2
Proportion	0.988	0.012

veil.color_e

n	missing	distinct
173	0	2

Value	0	1
Frequency	172	1
Proportion	0.994	0.006

veil.color_n

n	missing	distinct
173	0	2

Value	0	1
Frequency	171	2
Proportion	0.988	0.012

veil.color_u

n	missing	distinct
173	0	2

Value	0	1
Frequency	172	1
Proportion	0.994	0.006

veil.color_k

n	missing	distinct
173	0	2

Value	0	1
Frequency	172	1
Proportion	0.994	0.006

ring.type_g

n	missing	distinct
173	0	2

Value	0	1
Frequency	168	5
Proportion	0.971	0.029

ring.type_p

n	missing	distinct
173	0	2

Value	0	1
Frequency	168	5
Proportion	0.971	0.029

ring.type_e

n	missing	distinct
173	0	2

Value	0	1
Frequency	165	8
Proportion	0.954	0.046

ring.type_l

n	missing	distinct
173	0	2

Value	0	1
Frequency	167	6
Proportion	0.965	0.035

ring.type_f

n	missing	distinct
173	0	2

Value	0	1
Frequency	36	137
Proportion	0.208	0.792

ring.type_m

n	missing	distinct
173	0	2

Value	0	1
Frequency	172	1
Proportion	0.994	0.006

ring.type_r

n	missing	distinct
173	0	2

Value	0	1
Frequency	168	5
Proportion	0.971	0.029

ring.type_z

n	missing	distinct
173	0	2

Value	0	1
Frequency	167	6
Proportion	0.965	0.035

Spore.print.color_w

n	missing	distinct
173	0	2

Value	0	1
Frequency	169	4
Proportion	0.977	0.023

Spore.print.color_p

n	missing	distinct
173	0	2

Value	0	1
Frequency	169	4
Proportion	0.977	0.023

Spore.print.color_k

n	missing	distinct
173	0	2

Value	0	1
Frequency	166	7
Proportion	0.96	0.04

Spore.print.color_r

n	missing	distinct
173	0	2

Value	0	1
Frequency	172	1
Proportion	0.994	0.006

Spore.print.color_u

n	missing	distinct
173	0	2

Value	0	1
Frequency	172	1
Proportion	0.994	0.006

Spore.print.color_n

n	missing	distinct
173	0	2

Value	0	1
Frequency	170	3
Proportion	0.983	0.017

Spore.print.color_g

n	missing	distinct
173	0	2

Value	0	1
Frequency	172	1
Proportion	0.994	0.006

habitat_d

n	missing	distinct
173	0	2

Value	0	1
Frequency	22	151
Proportion	0.127	0.873

habitat_m

n	missing	distinct
173	0	2

Value	0	1
Frequency	156	17
Proportion	0.902	0.098

habitat_g

n	missing	distinct
173	0	2

Value	0	1
Frequency	135	38
Proportion	0.78	0.22

habitat_h

n	missing	distinct
173	0	2

Value	0	1
Frequency	160	13
Proportion	0.925	0.075

habitat_l

n	missing	distinct
173	0	2

Value	0	1
Frequency	155	18
Proportion	0.896	0.104

habitat_p

n	missing	distinct
173	0	2

Value	0	1
Frequency	171	2
Proportion	0.988	0.012

habitat_w

n	missing	distinct
173	0	2

Value	0	1
Frequency	172	1
Proportion	0.994	0.006

habitat_u

n	missing	distinct
173	0	2

Value	0	1
Frequency	172	1
Proportion	0.994	0.006

season_u

n	missing	distinct
173	0	2

Value	0	1
Frequency	33	140
Proportion	0.191	0.809

season_a

n	missing	distinct
173	0	2

Value	0	1
Frequency	5	168
Proportion	0.029	0.971

season_w

n	missing	distinct
173	0	2

Value	0	1
Frequency	132	41
Proportion	0.763	0.237

season_s

n	missing	distinct
173	0	2

Value	0	1
Frequency	150	23
Proportion	0.867	0.133

table1(processed)

```
new_data_selected <- new_data %>%
  select(cap.diameter_min, cap.diameter_max, cap.diameter_mean,
         stem.height_min, stem.height_max, stem.height_mean,
         stem.width_min, stem.width_max, stem.width_mean)

dat_selected <- dat %>%
  select(-c(stem.height, stem.width, cap.diameter))

merged_data <- cbind(new_data_selected, dat_selected)
# merged_data

table1(~ .| class, data = merged_data)
```

	0	1	Overall
	(N=96)	(N=77)	(N=173)
cap.diameter_min			
Mean (SD)	3.47 (2.27)	4.16 (2.38)	3.78 (2.34)
Median [Min, Max]	3.00 [0.400, 10.0]	4.00 [0.500, 12.0]	3.00 [0.400, 12.0]
Missing	0 (0%)	1 (1.3%)	1 (0.6%)
cap.diameter_max			
Mean (SD)	8.29 (5.58)	10.3 (5.76)	9.20 (5.73)
Median [Min, Max]	7.00 [1.00, 30.0]	10.0 [1.50, 30.0]	8.00 [1.00, 30.0]
Missing	0 (0%)	1 (1.3%)	1 (0.6%)
cap.diameter_mean			
Mean (SD)	NA (NA)	50.0 (NA)	50.0 (NA)
Median [Min, Max]	NA [NA, NA]	50.0 [50.0, 50.0]	50.0 [50.0, 50.0]
Missing	96 (100%)	76 (98.7%)	172 (99.4%)
stem.height_min			
Mean (SD)	4.27 (2.22)	4.52 (2.20)	4.38 (2.21)
Median [Min, Max]	4.00 [1.00, 15.0]	4.00 [2.00, 15.0]	4.00 [1.00, 15.0]
Missing	3 (3.1%)	0 (0%)	3 (1.7%)
stem.height_max			
Mean (SD)	8.57 (3.80)	9.58 (5.03)	9.03 (4.41)
Median [Min, Max]	8.00 [2.00, 20.0]	8.00 [3.00, 35.0]	8.00 [2.00, 35.0]
Missing	3 (3.1%)	0 (0%)	3 (1.7%)
stem.height_mean			
Mean (SD)	0 (0)	NA (NA)	0 (0)
Median [Min, Max]	0 [0, 0]	NA [NA, NA]	0 [0, 0]
Missing	93 (96.9%)	77 (100%)	170 (98.3%)
stem.width_min			
Mean (SD)	7.67 (5.65)	10.2 (6.90)	8.83 (6.36)
Median [Min, Max]	5.00 [0.500, 20.0]	10.0 [1.00, 40.0]	8.00 [0.500, 40.0]
Missing	7 (7.3%)	4 (5.2%)	11 (6.4%)
stem.width_max			
Mean (SD)	14.4 (11.8)	19.2 (15.9)	16.6 (13.9)
Median [Min, Max]	10.0 [1.00, 60.0]	15.0 [2.00, 100]	15.0 [1.00, 100]
Missing	7 (7.3%)	4 (5.2%)	11 (6.4%)
stem.width_mean			
Mean (SD)	2.00 (3.61)	7.75 (4.50)	4.09 (4.72)
Median [Min, Max]	1.00 [0, 10.0]	10.0 [1.00, 10.0]	1.00 [0, 10.0]
Missing	89 (92.7%)	73 (94.8%)	162 (93.6%)
family			
Amanita Family	5 (5.2%)	3 (3.9%)	8 (4.6%)
Bolbitius Family	2 (2.1%)	1 (1.3%)	3 (1.7%)
Bolete Family	3 (3.1%)	11 (14.3%)	14 (8.1%)
Bracket Fungi	6 (6.3%)	1 (1.3%)	7 (4.0%)
Cortinarius Family	11 (11.5%)	0 (0%)	11 (6.4%)
Crepidotus Family	1 (1.0%)	0 (0%)	1 (0.6%)
Ear-Pick Family	1 (1.0%)	0 (0%)	1 (0.6%)
Entoloma Family	6 (6.3%)	1 (1.3%)	7 (4.0%)
Ink Cap Family	7 (7.3%)	6 (7.8%)	13 (7.5%)
Jelly Discs Family	1 (1.0%)	0 (0%)	1 (0.6%)
Lepiota Family	1 (1.0%)	2 (2.6%)	3 (1.7%)
Mushroom Family	1 (1.0%)	4 (5.2%)	5 (2.9%)
Paxillus Family	3 (3.1%)	0 (0%)	3 (1.7%)
Russula Family	16 (16.7%)	11 (14.3%)	27 (15.6%)
Saddle-Cup Family	1 (1.0%)	0 (0%)	1 (0.6%)
Stropharia Family	7 (7.3%)	1 (1.3%)	8 (4.6%)
Tricholoma Family	20 (20.8%)	23 (29.9%)	43 (24.9%)

ggpairs(random select too much variables)

```
selected_vars <- c("class", "has.ring", "stem.width_mean", "cap.diameter_min",
  "stem.color_o", "does.bruise.or.bleed", "veil.color_n")
```

```
ggpairs(new_data,
  columns = selected_vars,
  mapping = aes(color = factor(class)))
```

