HW2

Create TableOne for the mushroom dataset

Bo-Cheng, Qiu

2025-03-20

目錄

Data Describtion

This dataset contains descriptions of hypothetical samples from various species of mushrooms families. The species in the dataset are classified into two categories: poisonous and edible. The classification is based on various characteristics such as cap shape, color, texture, stem properties, and habitat, similar to the categorization provided by the Audubon Society Field Guide to North American Mushrooms (1981). The edibility of the mushrooms is determined by the presence or absence of certain features, which are typical of either poisonous or edible varieties. This dataset was created for classification purposes, with the goal of building a model that can predict the edibility of a mushroom based on these features.

Variable	Data Type	Definition	Note
family	String	The name of the family to which the mushroom species belongs	Multinomial
name	String	The name of the mushroom species	Multinomial
class	String	Classifies the mushroom as poisonous (p) or edible (e)	Binary
cap-diameter	Float (cm)	Represents the diameter of the cap of the mushroom	Two values: min-max, or mean
cap-shape	Nominal	Describes the shape of the cap of the mushroom	bell = b, conical = c, convex = x, flat = f, sunken = s, spherical = p, others = o
cap-surface	Nominal	Describes the texture of the cap surface of the mushroom	fibrous = i, grooves = g, scaly = y, smooth = s, shiny = h, leathery = l, silky = k, sticky = t, wrinkled = w, fleshy = e
cap-color	Nominal	Describes the color of the mushroom cap	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	Indicates whether the mushroom bruises or bleeds when touched	bruises or bleeding = t, no = f
gill-attachment	Nominal	Describes how the gills are attached to the stem	adnate = a, adnexed = x, decurrent = d, free = e, sinuate = s, pores = p, none = f, unknown = ?
gill-spacing	Nominal	Describes the spacing between the gills	close = c, distant = d, none = f
gill-color	Nominal	Describes the color of the gills	Same as cap-color, none = f
stem-height	Float (cm)	Represents the height of the mushroom’s stem	Two values: min-max, or mean
stem-width	Float (mm)	Represents the width of the mushroom’s stem	Two values: min-max, or mean
stem-root	Nominal	Describes the type of the root of the mushroom’s stem	bulbous = b, swollen = s, club = c, cup = u, equal = e, rhizomorphs = z, rooted = r
stem-surface	Nominal	Describes the surface of the stem	Same as cap-surface, none = f
stem-color	Nominal	Describes the color of the stem	Same as cap-color, none = f
veil-type	Nominal	Describes the type of veil covering the mushroom’s stem	partial = p, universal = u
veil-color	Nominal	Describes the color of the veil	Same as cap-color, none = f
has-ring	Nominal	Indicates if the mushroom has a ring on the stem	ring = t, none = f
ring-type	Nominal	Describes the type of ring on the mushroom’s stem	cobwebby = c, evanescent = e, flaring = r, grooved = g, large = l, pendant = p, sheathing = s, zone = z, scaly = y, movable = m, none = f, unknown = ?
spore-print-color	Nominal	Describes the color of the spore print	Same as cap-color
habitat	Nominal	Describes the habitat where the mushroom is found	grasses = g, leaves = l, meadows = m, paths = p, heaths = h, urban = u, waste = w, woods = d
season	Nominal	Describes the season in which the mushroom is found	spring = s, summer = u, autumn = a, winter = w

Data Preprocessing

For two value(min, max) of numerical variables, we calculate the mean to present their value.

```
# install.packages("table1")
# install.packages("Hmisc")
# install.packages("glmnet")

library(readr)
library(dplyr)
library(stringr)
library(table1)
```

```r
#
file_path <- "primary_data.csv"
data <- read_delim(file_path,
                   delim = ";",
                   col_types = cols(.default = "c"))

#
clean_data <- data %>%
  mutate(across(everything(), ~ str_remove_all(.x, "\\[|\\]"))) %>%
  mutate(across(everything(), ~ str_squish(.x)))  #

colnames(clean_data) <- gsub("-", "_", colnames(clean_data))
# write_csv(clean_data, "primary_data_cleaned_R.csv")
# clean_data
# str(clean_data)
```

```r
#
calculate_average <- function(value) {
  if (grepl(",", value)) {
    #
    numbers <- as.numeric(strsplit(value, ",")[[1]])
    return(mean(numbers))
  } else {
    #
    return(as.numeric(value))
  }
}

clean_data$cap_diameter <- sapply(clean_data$cap_diameter, calculate_average)
clean_data$stem_height <- sapply(clean_data$stem_height, calculate_average)
clean_data$stem_width <- sapply(clean_data$stem_width, calculate_average)

# head(clean_data[c("cap_diameter", "stem_height", "stem_width")])

#
# write.csv(clean_data, "updated_data.csv", row.names = FALSE)
```

TableOne

```r
table1(~ cap_diameter + stem_height + stem_width +
         cap_shape + Cap_surface + cap_color + does_bruise_or_bleed +
         gill_attachment + gill_spacing + gill_color +
         stem_root + stem_surface + stem_color +
         veil_type + veil_color +
         has_ring + ring_type + Spore_print_color +
         habitat + season | class, data = clean_data)
```

	e	p	Overall
	(N=77)	(N=96)	(N=173)
cap_diameter			
Mean (SD)	7.81 (6.26)	5.88 (3.85)	6.74 (5.14)
Median [Min, Max]	6.50 [1.00, 50.0]	5.00 [0.700, 19.0]	6.00 [0.700, 50.0]
stem_height			
Mean (SD)	7.05 (3.48)	6.22 (3.05)	6.59 (3.26)
Median [Min, Max]	6.00 [2.50, 25.0]	5.50 [0, 17.5]	6.00 [0, 25.0]
stem_width			

	e	p	Overall
Mean (SD)	14.4 (10.8)	10.4 (8.66)	12.2 (9.86)
Median [Min, Max]	12.5 [1.00, 70.0]	7.50 [0, 40.0]	10.0 [0, 70.0]
cap_shape			
b	2 (2.6%)	8 (8.3%)	10 (5.8%)
b, f	2 (2.6%)	3 (3.1%)	5 (2.9%)
c	1 (1.3%)	2 (2.1%)	3 (1.7%)
c, x	1 (1.3%)	0 (0%)	1 (0.6%)
c, x, f	1 (1.3%)	0 (0%)	1 (0.6%)
f	4 (5.2%)	4 (4.2%)	8 (4.6%)
f, s	3 (3.9%)	5 (5.2%)	8 (4.6%)
f, x	1 (1.3%)	1 (1.0%)	2 (1.2%)
o	1 (1.3%)	7 (7.3%)	8 (4.6%)
p, b	1 (1.3%)	2 (2.1%)	3 (1.7%)
p, c, o	1 (1.3%)	0 (0%)	1 (0.6%)
p, f	2 (2.6%)	0 (0%)	2 (1.2%)
p, x	3 (3.9%)	1 (1.0%)	4 (2.3%)
p, x, f	2 (2.6%)	0 (0%)	2 (1.2%)
s	4 (5.2%)	5 (5.2%)	9 (5.2%)
s, o	2 (2.6%)	0 (0%)	2 (1.2%)
x	23 (29.9%)	25 (26.0%)	48 (27.7%)
x, f	14 (18.2%)	15 (15.6%)	29 (16.8%)
x, f, s	7 (9.1%)	6 (6.3%)	13 (7.5%)
x, p	1 (1.3%)	1 (1.0%)	2 (1.2%)
x, s	1 (1.3%)	2 (2.1%)	3 (1.7%)
b, f, s	0 (0%)	1 (1.0%)	1 (0.6%)
b, x	0 (0%)	3 (3.1%)	3 (1.7%)
b, x, f	0 (0%)	1 (1.0%)	1 (0.6%)
c, f	0 (0%)	2 (2.1%)	2 (1.2%)
p	0 (0%)	1 (1.0%)	1 (0.6%)
x, o	0 (0%)	1 (1.0%)	1 (0.6%)
Cap_surface			
d	4 (5.2%)	5 (5.2%)	9 (5.2%)
d, k	1 (1.3%)	1 (1.0%)	2 (1.2%)
d, s	1 (1.3%)	0 (0%)	1 (0.6%)
e	3 (3.9%)	2 (2.1%)	5 (2.9%)
e, y	1 (1.3%)	0 (0%)	1 (0.6%)
g	5 (6.5%)	7 (7.3%)	12 (6.9%)
g, s, h, t	1 (1.3%)	0 (0%)	1 (0.6%)
g, s, t	1 (1.3%)	0 (0%)	1 (0.6%)
h	3 (3.9%)	2 (2.1%)	5 (2.9%)
h, s, d	1 (1.3%)	0 (0%)	1 (0.6%)
h, t	6 (7.8%)	4 (4.2%)	10 (5.8%)
i, y	2 (2.6%)	0 (0%)	2 (1.2%)
l	2 (2.6%)	2 (2.1%)	4 (2.3%)
s	8 (10.4%)	5 (5.2%)	13 (7.5%)
s, d	1 (1.3%)	0 (0%)	1 (0.6%)
s, t	2 (2.6%)	2 (2.1%)	4 (2.3%)
s, y	1 (1.3%)	2 (2.1%)	3 (1.7%)
t	2 (2.6%)	10 (10.4%)	12 (6.9%)
t, h	1 (1.3%)	1 (1.0%)	2 (1.2%)
t, h, s	1 (1.3%)	0 (0%)	1 (0.6%)
w	2 (2.6%)	3 (3.1%)	5 (2.9%)
w, t	1 (1.3%)	0 (0%)	1 (0.6%)
y	7 (9.1%)	7 (7.3%)	14 (8.1%)
y, s	1 (1.3%)	0 (0%)	1 (0.6%)
d, e, y, i	0 (0%)	1 (1.0%)	1 (0.6%)
d, k, s	0 (0%)	1 (1.0%)	1 (0.6%)
e, k, s, h	0 (0%)	1 (1.0%)	1 (0.6%)
e, t, k	0 (0%)	1 (1.0%)	1 (0.6%)

	e	p	Overall
g, h	0 (0%)	1 (1.0%)	1 (0.6%)
g, s, d	0 (0%)	1 (1.0%)	1 (0.6%)
h, s, t	0 (0%)	1 (1.0%)	1 (0.6%)
h, t, w	0 (0%)	1 (1.0%)	1 (0.6%)
h, t, y	0 (0%)	1 (1.0%)	1 (0.6%)
i	0 (0%)	4 (4.2%)	4 (2.3%)
i, e	0 (0%)	1 (1.0%)	1 (0.6%)
k	0 (0%)	4 (4.2%)	4 (2.3%)
k, e	0 (0%)	1 (1.0%)	1 (0.6%)
s, h	0 (0%)	1 (1.0%)	1 (0.6%)
s, i	0 (0%)	1 (1.0%)	1 (0.6%)
t, w, d	0 (0%)	1 (1.0%)	1 (0.6%)
Missing	19 (24.7%)	21 (21.9%)	40 (23.1%)
cap_color			
b	1 (1.3%)	0 (0%)	1 (0.6%)
b, u	1 (1.3%)	0 (0%)	1 (0.6%)
e, n, y	2 (2.6%)	0 (0%)	2 (1.2%)
g, k	1 (1.3%)	1 (1.0%)	2 (1.2%)
g, n	6 (7.8%)	4 (4.2%)	10 (5.8%)
g, u, n, p	1 (1.3%)	0 (0%)	1 (0.6%)
k, n, w	1 (1.3%)	0 (0%)	1 (0.6%)
l, g, b, w	1 (1.3%)	0 (0%)	1 (0.6%)
l, r, w	1 (1.3%)	0 (0%)	1 (0.6%)
l, u, g, n	1 (1.3%)	0 (0%)	1 (0.6%)
l, y	1 (1.3%)	0 (0%)	1 (0.6%)
n	22 (28.6%)	16 (16.7%)	38 (22.0%)
n ,w	1 (1.3%)	0 (0%)	1 (0.6%)
n, b	1 (1.3%)	1 (1.0%)	2 (1.2%)
n, e	1 (1.3%)	4 (4.2%)	5 (2.9%)
n, g	3 (3.9%)	0 (0%)	3 (1.7%)
n, o	2 (2.6%)	2 (2.1%)	4 (2.3%)
n, o, e	1 (1.3%)	0 (0%)	1 (0.6%)
n, p, e	1 (1.3%)	1 (1.0%)	2 (1.2%)
n, r, u, y	1 (1.3%)	0 (0%)	1 (0.6%)
n, w	1 (1.3%)	3 (3.1%)	4 (2.3%)
n, y	3 (3.9%)	6 (6.3%)	9 (5.2%)
n, y, e	1 (1.3%)	0 (0%)	1 (0.6%)
n, y, w	1 (1.3%)	0 (0%)	1 (0.6%)
o, b	1 (1.3%)	0 (0%)	1 (0.6%)
o, n	1 (1.3%)	0 (0%)	1 (0.6%)
o, p, e	1 (1.3%)	0 (0%)	1 (0.6%)
u, k	1 (1.3%)	0 (0%)	1 (0.6%)
w	6 (7.8%)	6 (6.3%)	12 (6.9%)
w, g	1 (1.3%)	1 (1.0%)	2 (1.2%)
w, n	2 (2.6%)	2 (2.1%)	4 (2.3%)
w, p, o	1 (1.3%)	0 (0%)	1 (0.6%)
w, y	1 (1.3%)	1 (1.0%)	2 (1.2%)
y	6 (7.8%)	4 (4.2%)	10 (5.8%)
b, p, e, y	0 (0%)	1 (1.0%)	1 (0.6%)
e	0 (0%)	3 (3.1%)	3 (1.7%)
e, n	0 (0%)	2 (2.1%)	2 (1.2%)
e, n, p, w	0 (0%)	1 (1.0%)	1 (0.6%)
e, o	0 (0%)	1 (1.0%)	1 (0.6%)
e, o, k	0 (0%)	1 (1.0%)	1 (0.6%)
e, p, w	0 (0%)	1 (1.0%)	1 (0.6%)
e, u, y	0 (0%)	1 (1.0%)	1 (0.6%)
g	0 (0%)	1 (1.0%)	1 (0.6%)
g, n, k	0 (0%)	1 (1.0%)	1 (0.6%)
g, r, k, n	0 (0%)	1 (1.0%)	1 (0.6%)

	e	p	Overall
g, r, n	0 (0%)	2 (2.1%)	2 (1.2%)
g, u, n	0 (0%)	1 (1.0%)	1 (0.6%)
l, k	0 (0%)	1 (1.0%)	1 (0.6%)
n, e, y	0 (0%)	1 (1.0%)	1 (0.6%)
n, o, y, w	0 (0%)	1 (1.0%)	1 (0.6%)
o	0 (0%)	2 (2.1%)	2 (1.2%)
o, e, n, k	0 (0%)	1 (1.0%)	1 (0.6%)
o, y	0 (0%)	3 (3.1%)	3 (1.7%)
o, y, r	0 (0%)	1 (1.0%)	1 (0.6%)
p	0 (0%)	2 (2.1%)	2 (1.2%)
r	0 (0%)	1 (1.0%)	1 (0.6%)
r, l	0 (0%)	1 (1.0%)	1 (0.6%)
r, n	0 (0%)	1 (1.0%)	1 (0.6%)
r, p, y	0 (0%)	1 (1.0%)	1 (0.6%)
r, y	0 (0%)	1 (1.0%)	1 (0.6%)
u	0 (0%)	2 (2.1%)	2 (1.2%)
w, u	0 (0%)	1 (1.0%)	1 (0.6%)
w, y, g, n	0 (0%)	1 (1.0%)	1 (0.6%)
y, n	0 (0%)	3 (3.1%)	3 (1.7%)
y, o	0 (0%)	1 (1.0%)	1 (0.6%)
y, o, g, n, r	0 (0%)	1 (1.0%)	1 (0.6%)
y, o, r, n	0 (0%)	1 (1.0%)	1 (0.6%)
does_bruise_or_bleed			
f	63 (81.8%)	80 (83.3%)	143 (82.7%)
t	14 (18.2%)	16 (16.7%)	30 (17.3%)
gill_attachment			
a	11 (14.3%)	21 (21.9%)	32 (18.5%)
a, d	5 (6.5%)	3 (3.1%)	8 (4.6%)
d	9 (11.7%)	16 (16.7%)	25 (14.5%)
e	10 (13.0%)	6 (6.3%)	16 (9.2%)
f	4 (5.2%)	6 (6.3%)	10 (5.8%)
p	12 (15.6%)	5 (5.2%)	17 (9.8%)
s	7 (9.1%)	9 (9.4%)	16 (9.2%)
x	9 (11.7%)	12 (12.5%)	21 (12.1%)
Missing	10 (13.0%)	18 (18.8%)	28 (16.2%)
gill_spacing			
c	29 (37.7%)	41 (42.7%)	70 (40.5%)
d	13 (16.9%)	9 (9.4%)	22 (12.7%)
f	4 (5.2%)	6 (6.3%)	10 (5.8%)
Missing	31 (40.3%)	40 (41.7%)	71 (41.0%)
gill_color			
b	1 (1.3%)	0 (0%)	1 (0.6%)
b, u	1 (1.3%)	0 (0%)	1 (0.6%)
f	4 (5.2%)	6 (6.3%)	10 (5.8%)
g	3 (3.9%)	1 (1.0%)	4 (2.3%)
g, k	1 (1.3%)	1 (1.0%)	2 (1.2%)
g, n	1 (1.3%)	2 (2.1%)	3 (1.7%)
g, p	1 (1.3%)	0 (0%)	1 (0.6%)
g, w	2 (2.6%)	0 (0%)	2 (1.2%)
g, w, y	1 (1.3%)	0 (0%)	1 (0.6%)
k, n	2 (2.6%)	4 (4.2%)	6 (3.5%)
k, p, w	1 (1.3%)	0 (0%)	1 (0.6%)
n	3 (3.9%)	8 (8.3%)	11 (6.4%)
n, y	1 (1.3%)	1 (1.0%)	2 (1.2%)
o	2 (2.6%)	2 (2.1%)	4 (2.3%)
o, b	1 (1.3%)	0 (0%)	1 (0.6%)
o, e	1 (1.3%)	1 (1.0%)	2 (1.2%)
o, y	1 (1.3%)	4 (4.2%)	5 (2.9%)
p	3 (3.9%)	5 (5.2%)	8 (4.6%)

	e	p	Overall
p, n	1 (1.3%)	0 (0%)	1 (0.6%)
p, n, k	1 (1.3%)	0 (0%)	1 (0.6%)
p, w	3 (3.9%)	2 (2.1%)	5 (2.9%)
r	1 (1.3%)	0 (0%)	1 (0.6%)
u, w	1 (1.3%)	0 (0%)	1 (0.6%)
w	21 (27.3%)	15 (15.6%)	36 (20.8%)
w, n	3 (3.9%)	2 (2.1%)	5 (2.9%)
w, p	1 (1.3%)	2 (2.1%)	3 (1.7%)
w, p, y	1 (1.3%)	0 (0%)	1 (0.6%)
w, u, g, n	1 (1.3%)	0 (0%)	1 (0.6%)
w, y	3 (3.9%)	2 (2.1%)	5 (2.9%)
y	6 (7.8%)	7 (7.3%)	13 (7.5%)
y, e, n	1 (1.3%)	0 (0%)	1 (0.6%)
y, k	1 (1.3%)	0 (0%)	1 (0.6%)
y, n	1 (1.3%)	4 (4.2%)	5 (2.9%)
y, r	1 (1.3%)	0 (0%)	1 (0.6%)
b, p, w	0 (0%)	1 (1.0%)	1 (0.6%)
e	0 (0%)	1 (1.0%)	1 (0.6%)
g, n, u	0 (0%)	1 (1.0%)	1 (0.6%)
g, r, w	0 (0%)	1 (1.0%)	1 (0.6%)
g, u	0 (0%)	1 (1.0%)	1 (0.6%)
k, p	0 (0%)	1 (1.0%)	1 (0.6%)
n, e, y	0 (0%)	1 (1.0%)	1 (0.6%)
n, p	0 (0%)	2 (2.1%)	2 (1.2%)
n, r	0 (0%)	1 (1.0%)	1 (0.6%)
n, u	0 (0%)	1 (1.0%)	1 (0.6%)
n, w	0 (0%)	2 (2.1%)	2 (1.2%)
p, y	0 (0%)	1 (1.0%)	1 (0.6%)
p, y, r	0 (0%)	1 (1.0%)	1 (0.6%)
r, y	0 (0%)	1 (1.0%)	1 (0.6%)
w, b, n	0 (0%)	1 (1.0%)	1 (0.6%)
w, g	0 (0%)	1 (1.0%)	1 (0.6%)
w, g, k	0 (0%)	1 (1.0%)	1 (0.6%)
w, g, p, n	0 (0%)	1 (1.0%)	1 (0.6%)
w, g, u	0 (0%)	1 (1.0%)	1 (0.6%)
w, r	0 (0%)	1 (1.0%)	1 (0.6%)
w, y, g, n	0 (0%)	1 (1.0%)	1 (0.6%)
y, g, k	0 (0%)	1 (1.0%)	1 (0.6%)
y, o, e	0 (0%)	1 (1.0%)	1 (0.6%)
y, r, k	0 (0%)	1 (1.0%)	1 (0.6%)
y, w	0 (0%)	1 (1.0%)	1 (0.6%)
stem_root			
b	6 (7.8%)	3 (3.1%)	9 (5.2%)
s	4 (5.2%)	5 (5.2%)	9 (5.2%)
c	0 (0%)	2 (2.1%)	2 (1.2%)
f	0 (0%)	3 (3.1%)	3 (1.7%)
r	0 (0%)	4 (4.2%)	4 (2.3%)
Missing	67 (87.0%)	79 (82.3%)	146 (84.4%)
stem_surface			
i	4 (5.2%)	7 (7.3%)	11 (6.4%)
i, t	1 (1.3%)	0 (0%)	1 (0.6%)
k	1 (1.3%)	3 (3.1%)	4 (2.3%)
k, s	1 (1.3%)	0 (0%)	1 (0.6%)
s	9 (11.7%)	6 (6.3%)	15 (8.7%)
t	3 (3.9%)	4 (4.2%)	7 (4.0%)
y	4 (5.2%)	9 (9.4%)	13 (7.5%)
y, s	1 (1.3%)	0 (0%)	1 (0.6%)
f	0 (0%)	3 (3.1%)	3 (1.7%)
g	0 (0%)	5 (5.2%)	5 (2.9%)

	e	p	Overall
h	0 (0%)	1 (1.0%)	1 (0.6%)
i, s	0 (0%)	1 (1.0%)	1 (0.6%)
i, y	0 (0%)	1 (1.0%)	1 (0.6%)
s, h	0 (0%)	1 (1.0%)	1 (0.6%)
Missing	53 (68.8%)	55 (57.3%)	108 (62.4%)
stem_color			
b, u	1 (1.3%)	0 (0%)	1 (0.6%)
e, n	1 (1.3%)	2 (2.1%)	3 (1.7%)
e, y	1 (1.3%)	0 (0%)	1 (0.6%)
g	2 (2.6%)	0 (0%)	2 (1.2%)
g, n	1 (1.3%)	3 (3.1%)	4 (2.3%)
g, w	3 (3.9%)	0 (0%)	3 (1.7%)
k, n	1 (1.3%)	1 (1.0%)	2 (1.2%)
l, r, w	1 (1.3%)	0 (0%)	1 (0.6%)
n	15 (19.5%)	20 (20.8%)	35 (20.2%)
n, g	1 (1.3%)	1 (1.0%)	2 (1.2%)
n, o	1 (1.3%)	1 (1.0%)	2 (1.2%)
n, p, w	1 (1.3%)	0 (0%)	1 (0.6%)
n, w	2 (2.6%)	1 (1.0%)	3 (1.7%)
n, y	1 (1.3%)	1 (1.0%)	2 (1.2%)
o, e	1 (1.3%)	0 (0%)	1 (0.6%)
o, n	1 (1.3%)	0 (0%)	1 (0.6%)
o, y	1 (1.3%)	4 (4.2%)	5 (2.9%)
u	1 (1.3%)	1 (1.0%)	2 (1.2%)
w	32 (41.6%)	25 (26.0%)	57 (32.9%)
w, n	2 (2.6%)	1 (1.0%)	3 (1.7%)
w, o	1 (1.3%)	0 (0%)	1 (0.6%)
w, y	1 (1.3%)	2 (2.1%)	3 (1.7%)
y	5 (6.5%)	8 (8.3%)	13 (7.5%)
e	0 (0%)	1 (1.0%)	1 (0.6%)
e, u, y	0 (0%)	1 (1.0%)	1 (0.6%)
f	0 (0%)	3 (3.1%)	3 (1.7%)
g, r, n	0 (0%)	2 (2.1%)	2 (1.2%)
g, u, n	0 (0%)	1 (1.0%)	1 (0.6%)
k	0 (0%)	1 (1.0%)	1 (0.6%)
n, e	0 (0%)	2 (2.1%)	2 (1.2%)
n, p	0 (0%)	1 (1.0%)	1 (0.6%)
o	0 (0%)	1 (1.0%)	1 (0.6%)
p	0 (0%)	2 (2.1%)	2 (1.2%)
r, y	0 (0%)	1 (1.0%)	1 (0.6%)
u, e	0 (0%)	1 (1.0%)	1 (0.6%)
w, l, n	0 (0%)	1 (1.0%)	1 (0.6%)
w, u	0 (0%)	1 (1.0%)	1 (0.6%)
y, e, n	0 (0%)	1 (1.0%)	1 (0.6%)
y, n	0 (0%)	4 (4.2%)	4 (2.3%)
y, o, k	0 (0%)	1 (1.0%)	1 (0.6%)
veil_type			
u	3 (3.9%)	6 (6.3%)	9 (5.2%)
Missing	74 (96.1%)	90 (93.8%)	164 (94.8%)
veil_color			
w	7 (9.1%)	8 (8.3%)	15 (8.7%)
y	1 (1.3%)	0 (0%)	1 (0.6%)
y, w	1 (1.3%)	0 (0%)	1 (0.6%)
e, n	0 (0%)	1 (1.0%)	1 (0.6%)
k	0 (0%)	1 (1.0%)	1 (0.6%)
n	0 (0%)	1 (1.0%)	1 (0.6%)
u	0 (0%)	1 (1.0%)	1 (0.6%)
Missing	68 (88.3%)	84 (87.5%)	152 (87.9%)
has_ring			

	e	p	Overall
f	60 (77.9%)	70 (72.9%)	130 (75.1%)
t	17 (22.1%)	26 (27.1%)	43 (24.9%)
ring_type			
e	3 (3.9%)	3 (3.1%)	6 (3.5%)
f	61 (79.2%)	76 (79.2%)	137 (79.2%)
g	2 (2.6%)	0 (0%)	2 (1.2%)
l	1 (1.3%)	1 (1.0%)	2 (1.2%)
l, p	1 (1.3%)	0 (0%)	1 (0.6%)
l, r	2 (2.6%)	0 (0%)	2 (1.2%)
m	1 (1.3%)	0 (0%)	1 (0.6%)
p	1 (1.3%)	1 (1.0%)	2 (1.2%)
r	1 (1.3%)	2 (2.1%)	3 (1.7%)
e, g	0 (0%)	1 (1.0%)	1 (0.6%)
g, p	0 (0%)	2 (2.1%)	2 (1.2%)
l, e	0 (0%)	1 (1.0%)	1 (0.6%)
z	0 (0%)	6 (6.3%)	6 (3.5%)
Missing	4 (5.2%)	3 (3.1%)	7 (4.0%)
Spore_print_color			
g	1 (1.3%)	0 (0%)	1 (0.6%)
k	1 (1.3%)	4 (4.2%)	5 (2.9%)
p	1 (1.3%)	2 (2.1%)	3 (1.7%)
w	2 (2.6%)	1 (1.0%)	3 (1.7%)
k, r	0 (0%)	1 (1.0%)	1 (0.6%)
k, u	0 (0%)	1 (1.0%)	1 (0.6%)
n	0 (0%)	3 (3.1%)	3 (1.7%)
p, w	0 (0%)	1 (1.0%)	1 (0.6%)
Missing	72 (93.5%)	83 (86.5%)	155 (89.6%)
habitat			
d	47 (61.0%)	57 (59.4%)	104 (60.1%)
d, h	1 (1.3%)	3 (3.1%)	4 (2.3%)
g	1 (1.3%)	10 (10.4%)	11 (6.4%)
g, d	6 (7.8%)	4 (4.2%)	10 (5.8%)
g, d, h	1 (1.3%)	0 (0%)	1 (0.6%)
g, h, d	1 (1.3%)	2 (2.1%)	3 (1.7%)
g, l, m, d	1 (1.3%)	0 (0%)	1 (0.6%)
g, m	3 (3.9%)	2 (2.1%)	5 (2.9%)
g, m, d	1 (1.3%)	4 (4.2%)	5 (2.9%)
g, u, d	1 (1.3%)	0 (0%)	1 (0.6%)
l	1 (1.3%)	0 (0%)	1 (0.6%)
l, d	7 (9.1%)	6 (6.3%)	13 (7.5%)
l, d, h	1 (1.3%)	0 (0%)	1 (0.6%)
l, h	1 (1.3%)	0 (0%)	1 (0.6%)
m	1 (1.3%)	1 (1.0%)	2 (1.2%)
m, d	2 (2.6%)	1 (1.0%)	3 (1.7%)
w	1 (1.3%)	0 (0%)	1 (0.6%)
g, l, d	0 (0%)	1 (1.0%)	1 (0.6%)
h, d	0 (0%)	2 (2.1%)	2 (1.2%)
m, h	0 (0%)	1 (1.0%)	1 (0.6%)
p, d	0 (0%)	2 (2.1%)	2 (1.2%)
season			
a	5 (6.5%)	11 (11.5%)	16 (9.2%)
a, w	9 (11.7%)	6 (6.3%)	15 (8.7%)
s	1 (1.3%)	0 (0%)	1 (0.6%)
s, a, w	1 (1.3%)	0 (0%)	1 (0.6%)
s, u	2 (2.6%)	1 (1.0%)	3 (1.7%)
s, u, a	1 (1.3%)	4 (4.2%)	5 (2.9%)
s, u, a, w	7 (9.1%)	6 (6.3%)	13 (7.5%)
u, a	43 (55.8%)	63 (65.6%)	106 (61.3%)
u, a, w	8 (10.4%)	4 (4.2%)	12 (6.9%)

	e	p	Overall
u	0 (0%)	1 (1.0%)	1 (0.6%)

Visualization

Because the distribution of the categorical variables is so sparse, we choose numerical variables to visualize first.

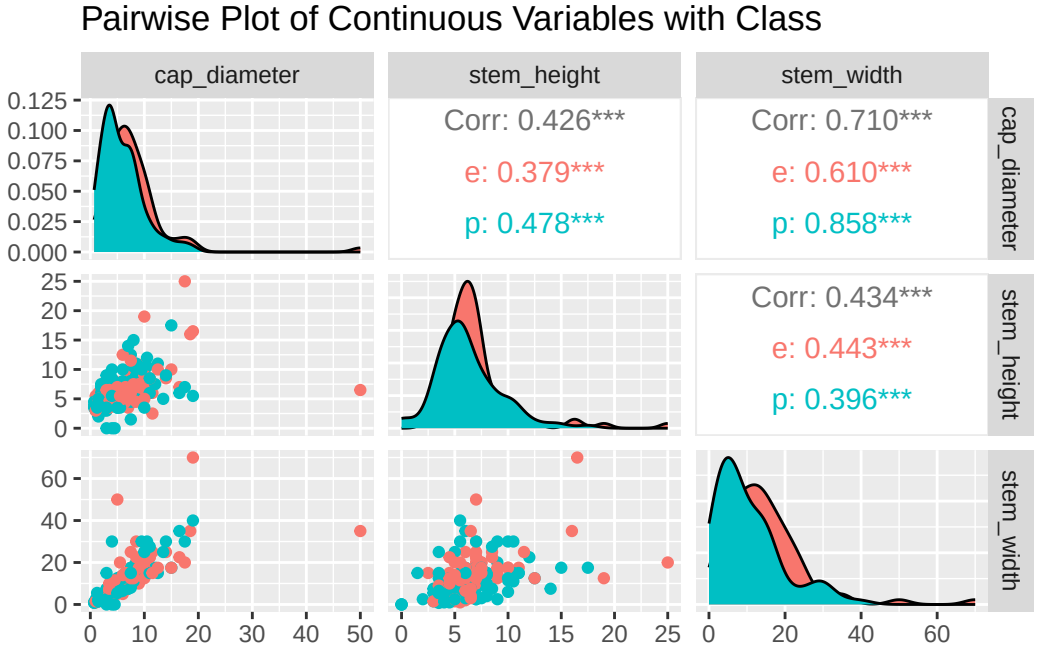
```
#
library(GGally)
library(ggplot2)

#
data <- read.csv("updated_data.csv")

#       ggpairs
numeric_data <- data[, c("cap_diameter", "stem_height", "stem_width")]

#
data$class <- as.factor(data$class)  #   class

#   ggpairs        class
ggpairs(numeric_data,
        mapping = ggplot2::aes(color = data$class),
        title = "Pairwise Plot of Continuous Variables with Class")
```



Lasso

We tried lasso to choose important variables; however, because the variables are so sparse that we can find each variable is insignificant to classify the mushroom is edible or not.

```
library(dplyr)
library(glmnet)

data <- read.csv("updated_data.csv")
data$class <- ifelse(data$class == 'p', 1, 0)
```

```r
numeric_columns <- c("cap_diameter", "stem_height", "stem_width")
categorical_columns <- c("cap_shape", "Cap_surface", "cap_color",
                         "does_bruise_or_bleed",
                         "gill_attachment", "gill_spacing", "gill_color",
                         "stem_root", "stem_surface", "stem_color",
                         "veil_type", "veil_color", "has_ring", "ring_type",
                         "Spore_print_color", "habitat", "season")

#           30%
missing_data <- colSums(is.na(data)) / nrow(data)
variables_to_keep <- names(missing_data[missing_data <= 0.30])
data <- data[, variables_to_keep]
numeric_columns <- intersect(numeric_columns, variables_to_keep)
categorical_columns <- intersect(categorical_columns, variables_to_keep)

#
for (col in categorical_columns) {
  mode_value <- names(sort(table(data[[col]]), decreasing = TRUE))[1]  #
  data[[col]] <- ifelse(is.na(data[[col]]), mode_value, data[[col]])
}
data[categorical_columns] <- lapply(data[categorical_columns], as.factor)


X <- data[, c(numeric_columns, categorical_columns)]
y <- data$class


X_dummies <- model.matrix(~ . - 1, data = X)  # One-hot encoding

# Lasso
lasso_model <- cv.glmnet(X_dummies, y, alpha = 1,
                         family = "binomial", type.measure = "class")

#   lambda
lasso_model$lambda.min
```

```
[1] 0.01351627
```

```r
#
selected_variables <- coef(lasso_model, s = "lambda.min")
selected_variables <- selected_variables[selected_variables != 0]
selected_variable_names <- rownames(selected_variables)[-1]
selected_variable_names
```

```
NULL
```