HW2

Yuzhen, Liu

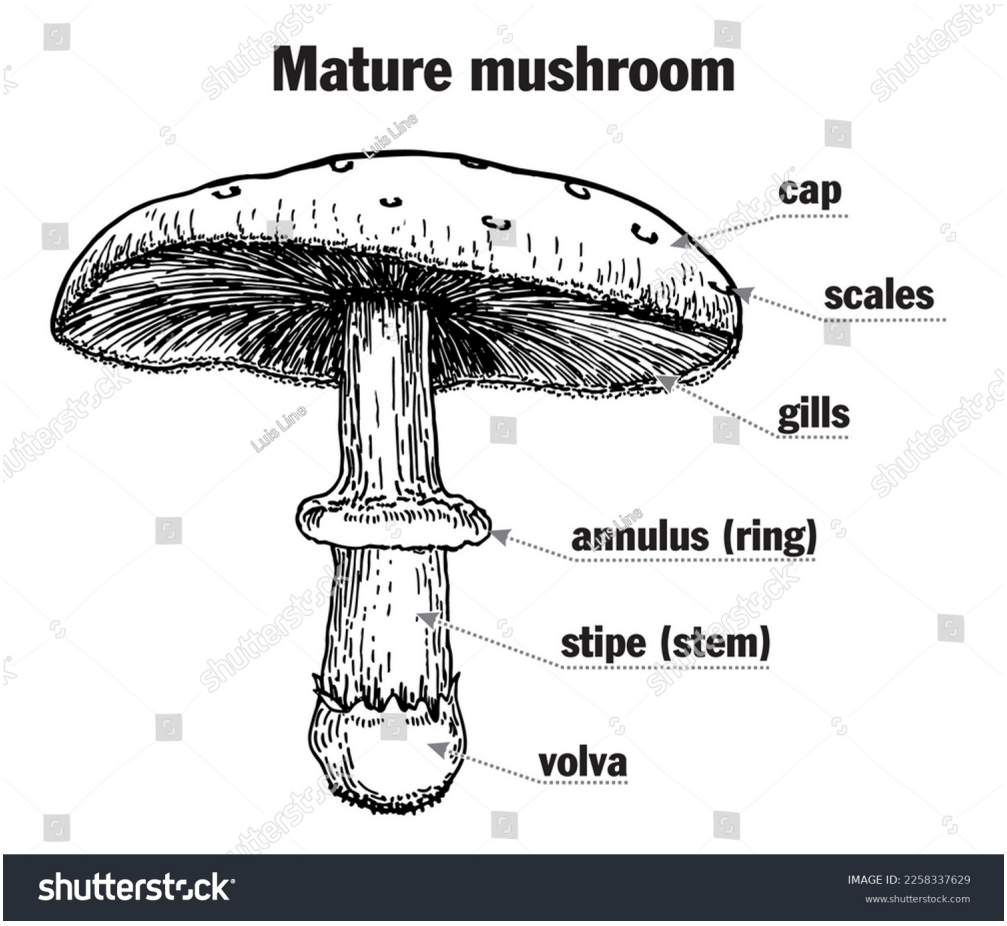
2025-03-20

目錄

0.1 Data information

Variable	Type	Annotation
family	Categorical	Mushroom family name
name	Categorical	Mushroom variety name
class	Binary	Edible (e) / Poisonous (p)
cap-diameter	Continuous	Minimum value, maximum value, or average value (cm)
cap-shape	Categorical	bell (b), conical (c), convex (x), flat (f), sunken (s), spherical (p), others (o)
cap-surface	Categorical	fibrous (i), grooves (g), scaly (y), smooth (s), shiny (h), leathery (l), silky (k), sticky (t), wrinkled (w), fleshy (e)
cap-color	Categorical	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	Categorical	bruises-or-bleeding (t), no (f)
gill-attachment	Categorical	adnate (a), adnexed (x), decurrent (d), free (e), sinuate (s), pores (p), none (f), unknown (?)
gill-spacing	Categorical	close (c), distant (d), none (f)
gill-color	Categorical	Same as cap-color + none (f)
stem-height	Continuous	Minimum value, maximum value, or average value (cm)
stem-width	Continuous	Minimum value, maximum value, or average value (mm)
stem-root	Categorical	bulbous (b), swollen (s), club (c), cup (u), equal (e), rhizomorphs (z), rooted (r)
stem-surface	Categorical	Same as cap-surface + none (f)
stem-color	Categorical	Same as cap-color + none (f)

veil-type	Categorical	partial (p), universal (u)
veil-color	Categorical	Same as cap-color + none (f)
has-ring	Categorical	ring (t), none (f)
ring-type	Categorical	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	Categorical	Same as cap-color
habitat	Categorical	grasses (g), leaves (l), meadows (m), paths (p), heaths (h), urban (u), waste (w), woods (d)
season	Categorical	spring (s), summer (u), autumn (a), winter (w)



0.2 Data pre-processing

To improve dataset readability and facilitate the observation of relationships between class and various characteristics, the following variables have been processed and structured accordingly:

- Cap
  - cap.diameter.min – Minimum cap diameter (cm).
  - cap.diameter.max – Maximum cap diameter (cm).

  - cap.diameter.avg – Observed average cap diameter (cm), if available.
  - cap.diameter.imputed – Estimated average cap diameter (cm), calculated as $(\text{min} + \text{max})/2$ if ‘cap.diameter.avg ‘ is missing.
  - cap.shape.X – One-hot encoded columns for each cap shape category.
  - cap.surface.X – One-hot encoded columns for each cap surface type.
  - cap.color.group – Categorized cap colors into:
    * dark: brown (‘n ‘), black (‘k ‘), gray (‘g ‘).
    * light: buff (‘b ‘), white (‘w ‘), yellow (‘y ‘).
    * warm: red (‘e ‘), orange (‘o ‘), pink (‘p ‘).
    * cool: green (‘r ‘), purple (‘u ‘), blue (‘l ‘).
    * other: includes missing values (‘NA ‘) and none (‘f ‘).

- Gill
  - gill.attachment.X – One-hot encoded columns for each gill attachment type.
  - gill.color.group – Categorized gill colors, see cap.color.group.

- Stem
  - stem.height.min, stem.height.max, stem.height.avg, stem.height.imputed – Same as cap.diameter.min, but for stem height (cm).
  - stem.width.min, stem.width.max, stem.width.avg, stem.width.imputed – Same as cap.diameter.min, but for stem width (mm).
  - stem.surface.X – One-hot encoded columns for each stem surface type.
  - stem.color.group – Categorized stem colors, see cap.color.group.

- Other
  - ring.type.X – One-hot encoded columns for each ring type.
  - spore.print.color.group – Categorized spore print colors, see cap.color.group.
  - veil.color.group – Categorized veil colors, see cap.color.group.
  - habitat.X – One-hot encoded columns for each habitat type.
  - season.X – One-hot encoded columns for each season type.

```
library(dplyr)
library(Hmisc)
library(tidyr)
library(stringr)


# Read original data
file_path <- "mushroom.txt"  # Replace with your file name
lines <- readLines(file_path)

# Process column names
columns <- unlist(strsplit(lines[1], ";"))

# Process data within `[]` to ensure they're treated as single units
process_line <- function(line) {
  # Use regex to preserve content within `[]` to avoid splitting by `;`
  line <- gsub("\\[([^\\]]+)\\]", "'\\1'", line)
```

```
  # Split by `;`
  split_line <- unlist(strsplit(line, ";"))

  split_line[split_line == ""] <- NA

  return(split_line)
}

# Process all data
data_list <- lapply(lines[-1], process_line)

# Convert to DataFrame
df <- as.data.frame(do.call(rbind, data_list), stringsAsFactors = FALSE)

# Set column names
colnames(df) <- columns

df<-read.csv("df.csv")
df <- df %>% rename(season = `season.............`)

df <- df %>%
  mutate(across(-c(X,cap.diameter, stem.height, stem.width), as.factor))

df <- df %>% select(-X)

df$class<-ifelse(df$class=="e","Edible","Poisonous")
df <- df %>%
  mutate(cap.diameter.clean = gsub("\\[|\\]", "", as.character(cap.diameter))) %>%
  mutate(value_count = sapply(strsplit(cap.diameter.clean, "\\s+"), length)) %>%
  separate(cap.diameter.clean, into = c("cap.diameter.min", "cap.diameter.max"), sep = "\\s+", fill = "r
  mutate(
    cap.diameter.avg = ifelse(value_count == 1, cap.diameter.min, NA),
    cap.diameter.min = ifelse(value_count == 2, cap.diameter.min, NA),
    cap.diameter.max = ifelse(value_count == 2, cap.diameter.max, NA)
  ) %>%
  select(-value_count)

df <- df %>%
  mutate(stem.height.clean = gsub("\\[|\\]", "", as.character(stem.height))) %>%
  mutate(stem.value_count = sapply(strsplit(stem.height.clean, "\\s+"), length)) %>%
  separate(stem.height.clean, into = c("stem.height.min", "stem.height.max"), sep = "\\s+", fill = "righ
  mutate(
    stem.height.avg = ifelse(stem.value_count == 1, stem.height.min, NA),
    stem.height.min = ifelse(stem.value_count == 2, stem.height.min, NA),
    stem.height.max = ifelse(stem.value_count == 2, stem.height.max, NA)
  ) %>%
  select(-stem.value_count)

df <- df %>%
  mutate(stem.width.clean = gsub("\\[|\\]", "", as.character(stem.width))) %>%
  mutate(stem.width.value_count = sapply(strsplit(stem.width.clean, "\\s+"), length)) %>%
  separate(stem.width.clean, into = c("stem.width.min", "stem.width.max"), sep = "\\s+", fill = "right",
  mutate(
```

```
    stem.width.avg = ifelse(stem.width.value_count == 1, stem.width.min, NA),
    stem.width.min = ifelse(stem.width.value_count == 2, stem.width.min, NA),
    stem.width.max = ifelse(stem.width.value_count == 2, stem.width.max, NA)
  ) %>%
  select(-stem.width.value_count)

df <- df %>%
  mutate(across(c(cap.diameter.min, cap.diameter.max, cap.diameter.avg,
                  stem.height.min, stem.height.max, stem.height.avg,
                  stem.width.min, stem.width.max, stem.width.avg),
                ~ na_if(., 0)))

df <- df %>%
  mutate(across(c(cap.diameter.min, cap.diameter.max, cap.diameter.avg,
                  stem.height.min, stem.height.max, stem.height.avg,
                  stem.width.min, stem.width.max, stem.width.avg),
                ~ as.numeric(as.character(.)))) %>%

  mutate(
    cap.diameter.imputed = ifelse(!is.na(cap.diameter.avg), cap.diameter.avg, (cap.diameter.min + cap.di
    stem.height.imputed = ifelse(!is.na(stem.height.avg), stem.height.avg, (stem.height.min + stem.heigh
    stem.width.imputed  = ifelse(!is.na(stem.width.avg), stem.width.avg, (stem.width.min + stem.width.ma
  )


df <- df %>%
  # Clean variables by removing `[`, `]`, and `\t`
  mutate(
    cap.shape.clean = str_replace_all(cap.shape, "[\\[\\]\\t]", ""),
    cap.surface.clean = str_replace_all(Cap.surface, "[\\[\\]\\t]", ""),
    stem.surface.clean = str_replace_all(stem.surface, "[\\[\\]\\t]", ""),
    ring.type.clean = str_replace_all(ring.type, "[\\[\\]\\t]", ""),
    habitat.clean = str_replace_all(habitat, "[\\[\\]\\t]", ""),
    season.clean = str_replace_all(season, "[\\[\\]\\t]", "")
  ) %>%

  # One-Hot Encoding for cap.shape
  separate_rows(cap.shape.clean, sep = " ") %>%
  mutate(value = 1) %>%
  pivot_wider(names_from = cap.shape.clean, values_from = value, values_fill = list(value = 0), names_pr
  mutate(across(starts_with("cap.shape."), as.factor)) %>%

  # One-Hot Encoding for cap.surface
  separate_rows(cap.surface.clean, sep = " ") %>%
  mutate(value = 1) %>%
  pivot_wider(names_from = cap.surface.clean, values_from = value, values_fill = list(value = 0), names_
  mutate(across(starts_with("cap.surface."), as.factor)) %>%

  # One-Hot Encoding for stem.surface
  separate_rows(stem.surface.clean, sep = " ") %>%
  mutate(value = 1) %>%
  pivot_wider(names_from = stem.surface.clean, values_from = value, values_fill = list(value = 0), names
  mutate(across(starts_with("stem.surface."), as.factor)) %>%
```

```r
  # One-Hot Encoding for ring.type
  separate_rows(ring.type.clean, sep = " ") %>%
  mutate(value = 1) %>%
  pivot_wider(names_from = ring.type.clean, values_from = value, values_fill = list(value = 0), names_pr
  mutate(across(starts_with("ring.type."), as.factor)) %>%

  # One-Hot Encoding for habitat
  separate_rows(habitat.clean, sep = " ") %>%
  mutate(value = 1) %>%
  pivot_wider(names_from = habitat.clean, values_from = value, values_fill = list(value = 0), names_pref
  mutate(across(starts_with("habitat."), as.factor)) %>%

  # One-Hot Encoding for season
  separate_rows(season.clean, sep = " ") %>%
  mutate(value = 1) %>%
  pivot_wider(names_from = season.clean, values_from = value, values_fill = list(value = 0), names_prefi
  mutate(across(starts_with("season."), as.factor))


library(forcats)

df <- df %>%
  # Clean color-related variables
  mutate(
    cap.color.clean = str_replace_all(cap.color, "[\\[\\]\\t]", ""),
    gill.color.clean = str_replace_all(gill.color, "[\\[\\]\\t]", ""),
    spore.print.color.clean = str_replace_all(Spore.print.color, "[\\[\\]\\t]", ""),
    stem.color.clean = str_replace_all(stem.color, "[\\[\\]\\t]", ""),
    veil.color.clean = str_replace_all(veil.color, "[\\[\\]\\t]", "")
  ) %>%

  # Group colors into meaningful categories
  mutate(across(c(cap.color.clean, gill.color.clean, spore.print.color.clean,
                  stem.color.clean, veil.color.clean),
                ~ case_when(
                  . %in% c("n", "k", "g") ~ "dark",
                  . %in% c("b", "w", "y") ~ "light",
                  . %in% c("e", "o", "p") ~ "warm",
                  . %in% c("r", "u", "l") ~ "cool",
                  TRUE ~ "other"
                ), .names = "{.col}.group")) %>%

  # Convert to factor and reorder "other" to the last level
  mutate(across(ends_with(".group"), ~ fct_relevel(as.factor(.), "other", after = Inf)))


df.a<-df %>%
  select(-matches("\\.clean$|\\.NA$"))
```

0.3 Data description

```
latex(describe(df.a), descript = "Descriptive Statistics (orginal)",
      file = '', caption.placement = "top")
```

df.a
86 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	Edible	Poisonous
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]
```

---

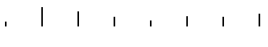
does.bruise.or.bleed

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\t 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

---

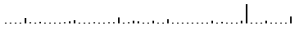
gill.spacing

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]

---

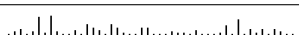
stem.height

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\t s]	[i\t t]	[i\t y]	[i]	[k\t s]	[k]	[s\t h]
Frequency	3	5	1	1	1	1	11	1	4	1
Proportion	0.046	0.077	0.015	0.015	0.015	0.015	0.169	0.015	0.062	0.015

Value	[s]	[t]	[y\t s]	[y]
Frequency	15	7	1	13
Proportion	0.231	0.108	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\t n]	[k]	[n]	[u]	[w]	[y\t 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\t g]	[e]	[f]	[g\t p]	[g]	[l\t e]	[l\t p]	[l\t r]	[l]	[m]
Frequency	1	6	137	2	2	1	1	2	2	1
Proportion	0.006	0.036	0.825	0.012	0.012	0.006	0.006	0.012	0.012	0.006

Value	[p]	[r]	[z]
Frequency	2	3	6
Proportion	0.012	0.018	0.036

Spore.print.color

n	missing	distinct
18	155	8

Value	[g]	[k\t r]	[k\t u]	[k]	[n]	[p\t 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

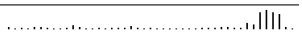
---

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	45

```
lowest : [a  w]                          [a    w]                          [a  w]                          [a    w]
highest: [u  a]                             [u    a]                             [u  a]
```

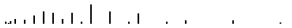
cap.diameter.min

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
172	1	13	0.976	3.776	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	Gmd	.05	.10	.25	.50	.75	.90	.95
172	1	19	0.991	9.199	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.avg

n	missing	distinct	Info	Mean	Gmd
1	172	1	0	50	NA

```
Value      50
Frequency   1
Proportion  1
```

stem.height.min

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
170	3	11	0.955	4.382	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	Gmd	.05	.10	.25	.50	.75	.90	.95
170	3	18	0.976	9.029	4.205	4.45	5.00	6.00	8.00	10.00	15.00	15.00

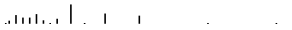
Value	2	3	4	5	6	7	8	9	10	11	12	14	15	18
Frequency	1	2	6	14	25	16	37	2	35	1	12	1	10	1
Proportion	0.006	0.012	0.035	0.082	0.147	0.094	0.218	0.012	0.206	0.006	0.071	0.006	0.059	0.006

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.width.min



n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
162	11	15	0.98	8.83	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	Gmd	.05	.10	.25	.50	.75	.90	.95
162	11	20	0.991	16.58	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.avg

n	missing	distinct	Info	Mean	Gmd
8	165	3	0.833	5.625	5.107

Value	1	2	10
Frequency	3	1	4
Proportion	0.375	0.125	0.500

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

---

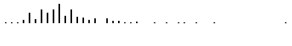
cap.diameter.imputed

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
173	0	35	0.997	6.739	4.755	1.5	2.0	3.5	6.0	8.5	11.4	15.0

lowest : 0.7 0.75 1 1.25 1.5 , highest: 16.5 17.5 18.5 19 50

---

stem.height.imputed



n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
170	3	29	0.993	6.706	3.105	3.50	3.50	5.00	6.00	7.50	10.05	12.27

lowest : 1.5 2 2.5 3 3.5 , highest: 16 16.5 17.5 19 25

---

stem.width.imputed

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
170	3	34	0.997	12.37	10.01	1.725	2.500	5.500	10.000	16.875	25.000	30.000

```
lowest : 0.75 1    1.5  2    2.5 , highest: 30   35   40   50   70
```

---

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

---

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

---

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

---

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

---

cap.color.clean.group

n	missing	distinct
173	0	5

Value	cool	dark	light	warm	other
Frequency	3	39	23	7	101
Proportion	0.017	0.225	0.133	0.040	0.584

---

gill.color.clean.group

n	missing	distinct
173	0	5

Value	cool	dark	light	warm	other
Frequency	1	15	50	13	94
Proportion	0.006	0.087	0.289	0.075	0.543

---

spore.print.color.clean.group

```
   n    missing    distinct
 173          0          4

Value       dark light  warm other
Frequency      9     3     3   158
Proportion 0.052 0.017 0.017 0.913
```

---

stem.color.clean.group

```
   n    missing    distinct
 173          0          5

Value       cool  dark light  warm other
Frequency      2    38    70     4    59
Proportion 0.012 0.220 0.405 0.023 0.341
```

---

veil.color.clean.group

```
   n    missing    distinct
 173          0          4

Value       cool  dark light other
Frequency      1     2    16   154
Proportion 0.006 0.012 0.092 0.890
```

Variables with all observations missing: `stem.height.avg`

0.4 Table1

Continuous variables (e.g., cap diameter, stem height, stem width) were analyzed using the Wilcoxon Rank-Sum Test, while categorical variables were assessed with Fisher□□s Exact Test to evaluate associations with mushroom edibility. To control for multiple comparisons, p-values were adjusted using the False Discovery Rate (FDR) correction via the Benjamini-Hochberg method.

Summary:

- Some numerical variables, such as cap diameter and stem width, initially showed significance but did not remain statistically significant after FDR adjustment.
- Stem root had a high proportion of missing values and could be considered negligible in the analysis.
- Cap shape (bell), cap surface (silky), cap color (cool), ring type (zone), and winter seasonality showed initial significance but did not hold after multiple testing correction.
- No variables showed strong evidence of association with edibility after FDR adjustment.

表 2: Characteristics of mushroom

	Overall	Edible	Poisonous	P.value	FDR
	(N=173)	(N=77)	(N=96)		
does.bruise.or.bleed					
[f]	143 (82.7%)	63 (81.8%)	80 (83.3%)	0.841	1.000
[t]	30 (17.3%)	14 (18.2%)	16 (16.7%)		
gill.attachment					

[a d]	8 (4.6%)	5 (6.5%)	3 (3.1%)	0.206	0.792
[a]	32 (18.5%)	11 (14.3%)	21 (21.9%)		
[d]	25 (14.5%)	9 (11.7%)	16 (16.7%)		
[e]	16 (9.2%)	10 (13.0%)	6 (6.3%)		
[f]	10 (5.8%)	4 (5.2%)	6 (6.3%)		
[p]	17 (9.8%)	12 (15.6%)	5 (5.2%)		
[s]	16 (9.2%)	7 (9.1%)	9 (9.4%)		
[x]	21 (12.1%)	9 (11.7%)	12 (12.5%)		
Missing	28 (16.2%)	10 (13.0%)	18 (18.8%)		
gill.spacing					
[c]	70 (40.5%)	29 (37.7%)	41 (42.7%)	0.358	0.899
[d]	22 (12.7%)	13 (16.9%)	9 (9.4%)		
[f]	10 (5.8%)	4 (5.2%)	6 (6.3%)		
Missing	71 (41.0%)	31 (40.3%)	40 (41.7%)		
stem.root					
[b]	9 (5.2%)	6 (7.8%)	3 (3.1%)	0.070.	0.459
[c]	2 (1.2%)	0 (0%)	2 (2.1%)		
[f]	3 (1.7%)	0 (0%)	3 (3.1%)		
[r]	4 (2.3%)	0 (0%)	4 (4.2%)		
[s]	9 (5.2%)	4 (5.2%)	5 (5.2%)		
Missing	146 (84.4%)	67 (87.0%)	79 (82.3%)		
has.ring					
[f]	130 (75.1%)	60 (77.9%)	70 (72.9%)	0.483	0.899
[t]	43 (24.9%)	17 (22.1%)	26 (27.1%)		
cap.diameter.imputed					
Mean (SD)	6.74 (5.14)	7.81 (6.26)	5.88 (3.85)	0.007**	0.207
Median [Min, Max]	6.00 [0.700, 50.0]	6.50 [1.00, 50.0]	5.00 [0.700, 19.0]		
stem.height.imputed					
Mean (SD)	6.71 (3.17)	7.05 (3.48)	6.42 (2.88)	0.122	0.600
Median [Min, Max]	6.00 [1.50, 25.0]	6.00 [2.50, 25.0]	6.00 [1.50, 17.5]		
Missing	3 (1.7%)	0 (0%)	3 (3.1%)		
stem.width.imputed					
Mean (SD)	12.4 (9.81)	14.4 (10.8)	10.7 (8.59)	0.005**	0.207
Median [Min, Max]	10.0 [0.750, 70.0]	12.5 [1.00, 70.0]	7.50 [0.750, 40.0]		
Missing	3 (1.7%)	0 (0%)	3 (3.1%)		
Cap Shape (Convex)					
0	63 (36.4%)	23 (29.9%)	40 (41.7%)	0.115	0.600
1	110 (63.6%)	54 (70.1%)	56 (58.3%)		
Cap Shape (Flat)					
0	99 (57.2%)	41 (53.2%)	58 (60.4%)	0.358	0.899
1	74 (42.8%)	36 (46.8%)	38 (39.6%)		
Cap Shape (Spherical)					
0	158 (91.3%)	67 (87.0%)	91 (94.8%)	0.102	0.600
1	15 (8.7%)	10 (13.0%)	5 (5.2%)		
Cap Shape (Bell)					
0	150 (86.7%)	72 (93.5%)	78 (81.3%)	0.023*	0.271
1	23 (13.3%)	5 (6.5%)	18 (18.8%)		
Cap Shape (Conical)					
0	165 (95.4%)	73 (94.8%)	92 (95.8%)	1.000	1.000
1	8 (4.6%)	4 (5.2%)	4 (4.2%)		

Cap Shape (Sunken)					
0	137 (79.2%)	60 (77.9%)	77 (80.2%)	0.711	0.941
1	36 (20.8%)	17 (22.1%)	19 (19.8%)		
Cap Shape (Other)					
0	161 (93.1%)	73 (94.8%)	88 (91.7%)	0.552	0.928
1	12 (6.9%)	4 (5.2%)	8 (8.3%)		
Cap Surface (Grooved)					
0	157 (90.8%)	70 (90.9%)	87 (90.6%)	1.000	1.000
1	16 (9.2%)	7 (9.1%)	9 (9.4%)		
Cap Surface (Shiny)					
0	147 (85.0%)	64 (83.1%)	83 (86.5%)	0.669	0.941
1	26 (15.0%)	13 (16.9%)	13 (13.5%)		
Cap Surface (Sticky)					
0	136 (78.6%)	62 (80.5%)	74 (77.1%)	0.710	0.941
1	37 (21.4%)	15 (19.5%)	22 (22.9%)		
Cap Surface (Scaly)					
0	150 (86.7%)	65 (84.4%)	85 (88.5%)	0.502	0.899
1	23 (13.3%)	12 (15.6%)	11 (11.5%)		
Cap Surface (Fleshy)					
0	162 (93.6%)	73 (94.8%)	89 (92.7%)	0.757	0.950
1	11 (6.4%)	4 (5.2%)	7 (7.3%)		
Cap Surface (Smooth)					
0	140 (80.9%)	59 (76.6%)	81 (84.4%)	0.243	0.792
1	33 (19.1%)	18 (23.4%)	15 (15.6%)		
Cap Surface (Leathery)					
0	169 (97.7%)	75 (97.4%)	94 (97.9%)	1.000	1.000
1	4 (2.3%)	2 (2.6%)	2 (2.1%)		
Cap Surface (Wrinkled)					
0	155 (89.6%)	69 (89.6%)	86 (89.6%)	1.000	1.000
1	18 (10.4%)	8 (10.4%)	10 (10.4%)		
Cap Surface (Waxy)					
0	165 (95.4%)	74 (96.1%)	91 (94.8%)	0.734	0.941
1	8 (4.6%)	3 (3.9%)	5 (5.2%)		
Cap Surface (Fibrous)					
0	164 (94.8%)	75 (97.4%)	89 (92.7%)	0.302	0.891
1	9 (5.2%)	2 (2.6%)	7 (7.3%)		
Cap Surface (Silky)					
0	163 (94.2%)	76 (98.7%)	87 (90.6%)	0.044*	0.371
1	10 (5.8%)	1 (1.3%)	9 (9.4%)		
Stem Surface (Scaly)					
0	158 (91.3%)	72 (93.5%)	86 (89.6%)	0.424	0.899
1	15 (8.7%)	5 (6.5%)	10 (10.4%)		
Stem Surface (Smooth)					
0	154 (89.0%)	66 (85.7%)	88 (91.7%)	0.231	0.792
1	19 (11.0%)	11 (14.3%)	8 (8.3%)		
Stem Surface (Silky)					
0	168 (97.1%)	75 (97.4%)	93 (96.9%)	1.000	1.000
1	5 (2.9%)	2 (2.6%)	3 (3.1%)		
Stem Surface (Fibrous)					
0	159 (91.9%)	72 (93.5%)	87 (90.6%)	0.582	0.928

1	14 (8.1%)	5 (6.5%)	9 (9.4%)		
Stem Surface (Shiny)					
0	171 (98.8%)	77 (100%)	94 (97.9%)	0.503	0.899
1	2 (1.2%)	0 (0%)	2 (2.1%)		
Stem Surface (Sticky)					
0	165 (95.4%)	73 (94.8%)	92 (95.8%)	1.000	1.000
1	8 (4.6%)	4 (5.2%)	4 (4.2%)		
Stem Surface (Grooved)					
0	168 (97.1%)	77 (100%)	91 (94.8%)	0.066.	0.459
1	5 (2.9%)	0 (0%)	5 (5.2%)		
Stem Surface (None)					
0	170 (98.3%)	77 (100%)	93 (96.9%)	0.255	0.792
1	3 (1.7%)	0 (0%)	3 (3.1%)		
Ring Type (Grooved)					
0	168 (97.1%)	75 (97.4%)	93 (96.9%)	1.000	1.000
1	5 (2.9%)	2 (2.6%)	3 (3.1%)		
Ring Type (Pendant)					
0	168 (97.1%)	75 (97.4%)	93 (96.9%)	1.000	1.000
1	5 (2.9%)	2 (2.6%)	3 (3.1%)		
Ring Type (Evanescent)					
0	165 (95.4%)	74 (96.1%)	91 (94.8%)	0.734	0.941
1	8 (4.6%)	3 (3.9%)	5 (5.2%)		
Ring Type (Large)					
0	167 (96.5%)	73 (94.8%)	94 (97.9%)	0.409	0.899
1	6 (3.5%)	4 (5.2%)	2 (2.1%)		
Ring Type (None)					
0	36 (20.8%)	16 (20.8%)	20 (20.8%)	1.000	1.000
1	137 (79.2%)	61 (79.2%)	76 (79.2%)		
Ring Type (Movable)					
0	172 (99.4%)	76 (98.7%)	96 (100%)	0.445	0.899
1	1 (0.6%)	1 (1.3%)	0 (0%)		
Ring Type (Flaring)					
0	168 (97.1%)	74 (96.1%)	94 (97.9%)	0.657	0.941
1	5 (2.9%)	3 (3.9%)	2 (2.1%)		
Ring Type (Zone)					
0	167 (96.5%)	77 (100%)	90 (93.8%)	0.034*	0.334
1	6 (3.5%)	0 (0%)	6 (6.3%)		
Habitat (woods)					
0	22 (12.7%)	8 (10.4%)	14 (14.6%)	0.494	0.899
1	151 (87.3%)	69 (89.6%)	82 (85.4%)		
Habitat (meadows)					
0	156 (90.2%)	69 (89.6%)	87 (90.6%)	1.000	1.000
1	17 (9.8%)	8 (10.4%)	9 (9.4%)		
Habitat (grasses)					
0	135 (78.0%)	62 (80.5%)	73 (76.0%)	0.580	0.928
1	38 (22.0%)	15 (19.5%)	23 (24.0%)		
Habitat (heaths)					
0	160 (92.5%)	72 (93.5%)	88 (91.7%)	0.775	0.953
1	13 (7.5%)	5 (6.5%)	8 (8.3%)		
Habitat (leaves)					

0	155 (89.6%)	66 (85.7%)	89 (92.7%)	0.143	0.603
1	18 (10.4%)	11 (14.3%)	7 (7.3%)		
Habitat (paths)					
0	171 (98.8%)	77 (100%)	94 (97.9%)	0.503	0.899
1	2 (1.2%)	0 (0%)	2 (2.1%)		
Habitat (waste)					
0	172 (99.4%)	76 (98.7%)	96 (100%)	0.445	0.899
1	1 (0.6%)	1 (1.3%)	0 (0%)		
Habitat (urban)					
0	172 (99.4%)	76 (98.7%)	96 (100%)	0.445	0.899
1	1 (0.6%)	1 (1.3%)	0 (0%)		
Summer					
0	33 (19.1%)	16 (20.8%)	17 (17.7%)	0.698	0.941
1	140 (80.9%)	61 (79.2%)	79 (82.3%)		
Autumn					
0	5 (2.9%)	3 (3.9%)	2 (2.1%)	0.657	0.941
1	168 (97.1%)	74 (96.1%)	94 (97.9%)		
Winter					
0	132 (76.3%)	52 (67.5%)	80 (83.3%)	0.019*	0.271
1	41 (23.7%)	25 (32.5%)	16 (16.7%)		
Spring					
0	150 (86.7%)	65 (84.4%)	85 (88.5%)	0.502	0.899
1	23 (13.3%)	12 (15.6%)	11 (11.5%)		
Cap color					
cool	3 (1.7%)	0 (0%)	3 (3.1%)	0.012*	0.236
dark	39 (22.5%)	22 (28.6%)	17 (17.7%)		
light	23 (13.3%)	13 (16.9%)	10 (10.4%)		
warm	7 (4.0%)	0 (0%)	7 (7.3%)		
other	101 (58.4%)	42 (54.5%)	59 (61.5%)		
Gill color					
cool	1 (0.6%)	1 (1.3%)	0 (0%)	0.240	0.792
dark	15 (8.7%)	6 (7.8%)	9 (9.4%)		
light	50 (28.9%)	28 (36.4%)	22 (22.9%)		
warm	13 (7.5%)	5 (6.5%)	8 (8.3%)		
other	94 (54.3%)	37 (48.1%)	57 (59.4%)		
Spore print color					
dark	9 (5.2%)	2 (2.6%)	7 (7.3%)	0.537	0.928
light	3 (1.7%)	2 (2.6%)	1 (1.0%)		
warm	3 (1.7%)	1 (1.3%)	2 (2.1%)		
other	158 (91.3%)	72 (93.5%)	86 (89.6%)		
Stem color					
cool	2 (1.2%)	1 (1.3%)	1 (1.0%)	0.139	0.603
dark	38 (22.0%)	17 (22.1%)	21 (21.9%)		
light	70 (40.5%)	37 (48.1%)	33 (34.4%)		
warm	4 (2.3%)	0 (0%)	4 (4.2%)		
other	59 (34.1%)	22 (28.6%)	37 (38.5%)		
Veil color					
cool	1 (0.6%)	0 (0%)	1 (1.0%)	0.646	0.941
dark	2 (1.2%)	0 (0%)	2 (2.1%)		
light	16 (9.2%)	8 (10.4%)	8 (8.3%)		

other	154 (89.0%)	69 (89.6%)	85 (88.5%)	