HW2

Xsin-Yu Huang

2025-03-18

目錄

一、Variable Definition

```r
library(dplyr)
library(Hmisc)
library(tibble)
library(knitr)
library(table1)
library(reticulate)
#
mushroom_data <- tibble::tribble(
  ~Variable, ~ "Data Type", ~"Values/Range",
  "family"," Categorical","familyname",
  "name","Categorical","Mushroomvarietyname",
  "class","Binary","poisonous=p, edibile=e",
  "cap-diameter", "Continuous", "float (cm)",
  "cap-shape", "Categorical", "{b, c, x, f, s, p, o}",
  "cap-surface", "Categorical", "{i, g, y, s, h, l, k, t, w, e}",
  "cap-color", "Categorical", "{n, b, g, r, p, u, e, w, y, l, o, k}",
  "does-bruise-bleed", "Categorical", "{t, f}",
  "gill-attachment", "Categorical", "{a, x, d, e, s, p, f, ?}",
  "gill-spacing", "Categorical", "{c, d, f}",
  "gill-color", "Categorical", "{cap-color + f}",
  "stem-height", "Continuous", "float (cm)",
  "stem-width", "Continuous", "float (mm)",
  "stem-root", "Categorical", "{b, s, c, u, e, z, r}",
  "stem-surface", "Categorical", "{cap-surface + f}",
  "stem-color", "Categorical", "{cap-color + f}",
  "veil-type", "Categorical", "{p, u}",
  "veil-color", "Categorical", "{cap-color + f}",
  "has-ring", "Categorical", "{t, f}",
  "ring-type", "Categorical", "{c, e, r, g, l, p, s, z, y, m, f, ?}",
  "spore-print-color", "Categorical", "{cap-color}",
  "habitat", "Categorical", "{g, l, m, p, h, u, w, d}",
  "season", "Categorical", "{s, u, a, w}"
)
kable(mushroom_data, format = "latex", booktabs = TRUE,escape = FALSE,caption="Mushroom Variable Table"
```

表 1: Mushroom Variable Table

Variable	Data Type	Values/Range
family	Categorical	familyname
name	Categorical	Mushroomvarietyname
class	Binary	poisonous=p, edibile=e
cap-diameter	Continuous	float (cm)
cap-shape	Categorical	b, c, x, f, s, p, o
cap-surface	Categorical	i, g, y, s, h, l, k, t, w, e
cap-color	Categorical	n, b, g, r, p, u, e, w, y, l, o, k
does-bruise-bleed	Categorical	t, f
gill-attachment	Categorical	a, x, d, e, s, p, f, ?
gill-spacing	Categorical	c, d, f
gill-color	Categorical	cap-color + f
stem-height	Continuous	float (cm)
stem-width	Continuous	float (mm)
stem-root	Categorical	b, s, c, u, e, z, r
stem-surface	Categorical	cap-surface + f
stem-color	Categorical	cap-color + f
veil-type	Categorical	p, u
veil-color	Categorical	cap-color + f
has-ring	Categorical	t, f
ring-type	Categorical	c, e, r, g, l, p, s, z, y, m, f, ?
spore-print-color	Categorical	cap-color
habitat	Categorical	g, l, m, p, h, u, w, d
season	Categorical	s, u, a, w

二、Data Preprocessing

```
library(dplyr)
library(Hmisc)
library(tibble)
library(tidyverse)
library(table1)
library(reticulate)

#   CSV
df <- read.csv("primary_data.csv", sep = ";", stringsAsFactors = FALSE)

#        []
df[] <- lapply(df, function(x) gsub("\\[|\\]", "", x))
df <- df %>%
  mutate_all(~ifelse(. == "", NA, .))

split_column <- function(data, column_name) {
  data %>%
    mutate(!!column_name := str_replace_all(.data[[column_name]], " ", "")) %>%  #
    separate(
      !!column_name, into = c(paste0(column_name, "_Min"), paste0(column_name, "_Max")),
      sep = ",", fill = "right"
    ) %>%
    mutate(
      across(c(paste0(column_name, "_Min"), paste0(column_name, "_Max")), as.numeric)
    ) %>%
    mutate(
      !!paste0(column_name, "_Max") := ifelse(
        is.na(.data[[paste0(column_name, "_Max")]]),
        .data[[paste0(column_name, "_Min")]],
        .data[[paste0(column_name, "_Max")]]
      )
    )
}

df_clean <- df %>%
  split_column("cap.diameter") %>%
  split_column("stem.height") %>%
  split_column("stem.width")

#
write.csv(df_clean, "cleaned_data.csv", row.names = FALSE)
```

三、Data Description

```
library(Hmisc)
library(knitr)
df <- read.csv("cleaned_data.csv")
library(dplyr)
df<-df %>%
 mutate(across(-c(cap.diameter_Min,cap.diameter_Max,stem.height_Min,stem.height_Max,stem.width_Min,stem.
df<-df %>%
 mutate(across(c(family,name),as.character))
df<-df %>%
 mutate(across(c(cap.diameter_Min,cap.diameter_Max,stem.height_Min,stem.height_Max,stem.width_Min,stem.w

desc_stats <- Hmisc::describe(df)
desc_stats
```

```
df

 26  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
highest: Yellow-gilled Russula    Yellow-staining Mushroom Yellow-stemmed Bell Cap  Yellow Swamp Russula
--------------------------------------------------------------------------------
class
       n  missing distinct
     173        0        2

Value          e     p
Frequency     77    96
Proportion 0.445 0.555
--------------------------------------------------------------------------------
cap.diameter_Min
       n  missing distinct     Info     Mean  pMedian      Gmd      .05
     173        0       14    0.976    4.043      3.5    3.038        1
     .10      .25      .50      .75      .90      .95
       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
Frequency      2     4     1    17    39    24    26    29    11     4     9
Proportion 0.012 0.023 0.006 0.098 0.225 0.139 0.150 0.168 0.064 0.023 0.052

Value       10.0  12.0  50.0
Frequency      4     2     1
Proportion 0.023 0.012 0.006

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
cap.diameter_Max
       n  missing distinct     Info     Mean  pMedian      Gmd      .05
     173        0       20    0.991    9.435      8.5    6.548        2
     .10      .25      .50      .75      .90      .95
       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
Frequency      3     1     4     7     6    12    18    16     7    16     3
Proportion 0.017 0.006 0.023 0.040 0.035 0.069 0.104 0.092 0.040 0.092 0.017

Value       10.0  12.0  14.0  15.0  18.0  20.0  25.0  30.0  50.0
Frequency     28    18     3    15     3     5     5     2     1
Proportion 0.162 0.104 0.017 0.087 0.017 0.029 0.029 0.012 0.006

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
cap.shape
       n  missing distinct
     173        0       27

lowest : b       b, f    b, f, s b, x    b, x, f
highest: x, f    x, f, s x, o    x, p    x, s
--------------------------------------------------------------------------------
Cap.surface
       n  missing distinct
     133       40       40

lowest : d          d, e, y, i d, k       d, k, s    d, s
highest: t, w, d    w          w, t       y          y, s
--------------------------------------------------------------------------------
cap.color
       n  missing distinct
     173        0       67

lowest : b             b, p, e, y    b, u          e             e, n
highest: y             y, n          y, o          y, o, g, n, r y, o, r, n
--------------------------------------------------------------------------------
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  a, d     d     e     f     p     s     x
Frequency     32     8    25    16    10    17    16    21
Proportion 0.221 0.055 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       b, p, w b, u    e       f
highest: y, n    y, o, e y, r    y, r, k y, w
--------------------------------------------------------------------------------
stem.height_Min
       n  missing distinct     Info     Mean  pMedian      Gmd      .05
     173        0       12    0.957    4.306        4    2.233      2.0
     .10      .25      .50      .75      .90      .95
     2.0      3.0      4.0      5.0      6.8      8.0 

Value          0     1     2     3     4     5     6     7     8    10    12
Frequency      3     2    21    38    52    24    15     3     7     5     1
Proportion 0.017 0.012 0.121 0.220 0.301 0.139 0.087 0.017 0.040 0.029 0.006

Value         15
Frequency      2
Proportion 0.012

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
stem.height_Max 
       n  missing distinct     Info     Mean  pMedian      Gmd      .05 
     173        0       19    0.977    8.873        8     4.37      4.0 
     .10      .25      .50      .75      .90      .95 
     5.0      6.0      8.0     10.0     14.8     15.0 

Value          0     2     3     4     5     6     7     8     9    10    11
Frequency      3     1     2     6    14    25    16    37     2    35     1
Proportion 0.017 0.006 0.012 0.035 0.081 0.145 0.092 0.214 0.012 0.202 0.006

Value         12    14    15    18    20    25    30    35
Frequency     12     1    10     1     4     1     1     1
Proportion 0.069 0.006 0.058 0.006 0.023 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  pMedian      Gmd      .05 
     173        0       16     0.98    8.529        8    6.804        1 
     .10      .25      .50      .75      .90      .95 
       2        4        8       10       19       20 

Value        0.0   0.5   1.0   2.0   3.0   4.0   5.0   6.0   7.0   8.0  10.0
Frequency      3     1     9    18    12    12    19     7     1    10    42
Proportion 0.017 0.006 0.052 0.104 0.069 0.069 0.110 0.040 0.006 0.058 0.243

Value       12.0  15.0  20.0  30.0  40.0
Frequency      1    20    16     1     1
Proportion 0.006 0.116 0.092 0.006 0.006

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
stem.width_Max 
       n  missing distinct     Info     Mean  pMedian      Gmd      .05 
     173        0       21    0.992    15.79       14    13.49        2 
     .10      .25      .50      .75      .90      .95 
       3        8       12       20       30       40 

lowest :   0   1   2   3   4, highest:  40  50  60  80 100
--------------------------------------------------------------------------------
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  i, s  i, t  i, y     k  k, s     s  s, h
Frequency      3     5     1    11     1     1     1     4     1    15     1
Proportion 0.046 0.077 0.015 0.169 0.015 0.015 0.015 0.062 0.015 0.231 0.015

Value          t     y  y, s
Frequency      7    13     1
Proportion 0.108 0.200 0.015
--------------------------------------------------------------------------------
stem.color
       n  missing distinct
     173        0       41

lowest : b, u    e       e, n    e, u, y e, y
highest: w, y    y       y, e, n y, n    y, o, k
--------------------------------------------------------------------------------
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  y, w
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  e, g     f     g  g, p     l  l, e  l, p  l, r     m     p
```

```
Frequency      6     1   137     2     2     2     1     1     2     1     2
Proportion 0.036 0.006 0.825 0.012 0.012 0.012 0.006 0.006 0.012 0.006 0.012

Value          r     z
Frequency      3     6
Proportion 0.018 0.036
--------------------------------------------------------------------------------
Spore.print.color
       n  missing distinct
      18      155        8

Value          g     k  k, r  k, u     n     p  p, w     w
Frequency      1     5     1     1     3     3     1     3
Proportion 0.056 0.278 0.056 0.056 0.167 0.167 0.056 0.167
--------------------------------------------------------------------------------
habitat
       n  missing distinct
     173        0       21

lowest : d       d, h    g       g, d    g, d, h
highest: m       m, d    m, h    p, d    w
--------------------------------------------------------------------------------
season
       n  missing distinct
     173        0       10

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

Value      s, u, a, w          u       u, a    u, a, w
Frequency          13          1        106         12
Proportion      0.075      0.006      0.613      0.069
--------------------------------------------------------------------------------
```

四、Table One

```
library(table1)
df$class<-ifelse(df$class=="e","Edible","Poisonous")
options(table1.longtable= TRUE)
table1(~ cap.diameter_Min+cap.diameter_Max+stem.height_Min+stem.height_Max+stem.width_Min+stem.width_Max
```

	Edible	Poisonous	Overall
	(N=77)	(N=96)	(N=173)
cap.diameter_Min			
Mean (SD)	4.75 (5.74)	3.47 (2.27)	4.04 (4.22)
Median [Min, Max]	4.00 [0.500, 50.0]	3.00 [0.400, 10.0]	3.00 [0.400, 50.0]
cap.diameter_Max			
Mean (SD)	10.9 (7.29)	8.29 (5.58)	9.44 (6.50)
Median [Min, Max]	10.0 [1.50, 50.0]	7.00 [1.00, 30.0]	8.00 [1.00, 50.0]
stem.height_Min			
Mean (SD)	4.52 (2.20)	4.14 (2.31)	4.31 (2.26)
Median [Min, Max]	4.00 [2.00, 15.0]	4.00 [0, 15.0]	4.00 [0, 15.0]
stem.height_Max			
Mean (SD)	9.58 (5.03)	8.30 (4.03)	8.87 (4.53)
Median [Min, Max]	8.00 [3.00, 35.0]	8.00 [0, 20.0]	8.00 [0, 35.0]
stem.width_Min			
Mean (SD)	10.1 (6.80)	7.26 (5.71)	8.53 (6.36)
Median [Min, Max]	10.0 [1.00, 40.0]	5.00 [0, 20.0]	8.00 [0, 40.0]
stem.width_Max			
Mean (SD)	18.6 (15.7)	13.5 (11.8)	15.8 (13.9)
Median [Min, Max]	15.0 [1.00, 100]	10.0 [0, 60.0]	12.0 [0, 100]