HW_2

summary of mushroom data

Wang Hao Yu - RE6131032

2025-03-09

目錄

Variable Information

Variable	Measurement	Values
cap-diameter	Quantitative	Float number in cm
cap-shape	Qualitative	bell=b, conical=c, convex=x, flat=f, sunken=s, spherical=p, others=o
cap-surface	Qualitative	fibrous=i, grooves=g, scaly=y, smooth=s, shiny=h, leathery=l, silky=k, sticky=t, wrinkled=w, fleshy=e

Variable	Measurement	Values
cap-color	Qualitative	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	Qualitative	bruises-or-bleeding=t, no=f
gill-attachment	Qualitative	adnate=a, adnexed=x, decurrent=d, free=e, sinuate=s, pores=p, unknown=?
gill-spacing	Qualitative	close=c, distant=d, none=f
gill-color	Qualitative	see cap-color
stem-height	Quantitative	Float number in cm
stem-width	Quantitative	Float number in mm
stem-root	Qualitative	bulbous=b, swollen=s, club=c, cup=u, equal=e, rhizomorphs=z, rooted=r
stem-surface	Qualitative	see cap-surface
stem-color	Qualitative	see cap-color
veil-type	Qualitative	partial=p, universal=u
veil-color	Qualitative	see cap-color
has-ring	Qualitative	ring=t
ring-type	Qualitative	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	Qualitative	see cap-color

Variable	Measurement	Values
habitat	Qualitative	grasses=g, leaves=l, meadows=m, paths=p, heaths=h, urban=u, waste=w, woods=d
season	Qualitative	spring=s, summer=u, autumn=a, winter=w

```
#install.packages("table1", repos = "https://cran.r-project.org/")
```

load data an package

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

```
#
library(ggplot2)
library(gridExtra)
library(dplyr)
library(table1)
```

```
head(mushroom_data)
```

```
          family              name class cap.diameter cap.shape Cap.surface
1 Amanita Family        Fly Agaric     p     [10, 20]    [x, f]       [g, h]
2 Amanita Family       Panther Cap     p      [5, 10]    [p, x]          [g]
3 Amanita Family False Panther Cap     p     [10, 15]    [x, f]
4 Amanita Family       The Blusher     e      [5, 15]    [x, f]
5 Amanita Family         Death Cap     p      [5, 12]    [x, f]          [h]
6 Amanita Family   False Death Cap     e       [4, 9]       [x]
  cap.color does.bruise.or.bleed gill.attachment gill.spacing gill.color
1    [e, o]                  [f]             [e]                    [w]
2       [n]                  [f]             [e]                    [w]
3    [g, n]                  [f]             [e]                    [w]
4       [n]                  [t]                                    [w]
5       [r]                  [f]                            [c]     [w]
6    [w, y]                  [f]             [e]                    [w]
  stem.height stem.width stem.root stem.surface stem.color veil.type veil.color
1    [15, 20]   [15, 20]       [s]          [y]        [w]       [u]        [w]
2     [6, 10]   [10, 20]                    [y]        [w]       [u]        [w]
3    [10, 12]   [10, 20]                               [w]       [u]        [w]
4     [7, 15]   [10, 25]       [b]                     [w]       [u]        [w]
5    [10, 12]   [10, 20]                               [w]       [u]        [w]
```

```
6      [5, 7]   [10, 15]       [b]                  [w, y]       [u]     [y, w]
  has.ring ring.type Spore.print.color habitat    season
1      [t]    [g, p]                       [d] [u, a, w]
2      [t]       [p]                       [d]    [u, a]
3      [t]    [e, g]                       [d]    [u, a]
4      [t]       [g]                       [d]    [u, a]
5      [t]    [g, p]                       [d]    [u, a]
6      [t]       [g]                       [d]    [u, a]
```

```
str(mushroom_data)
```

```
'data.frame':   173 obs. of  23 variables:
 $ family              : chr  "Amanita Family" "Amanita Family" "Amanita Family" "Amanita Family" ...
 $ name                : chr  "Fly Agaric" "Panther Cap" "False Panther Cap" "The Blusher" ...
 $ class               : chr  "p" "p" "p" "e" ...
 $ cap.diameter        : chr  "[10, 20]" "[5, 10]" "[10, 15]" "[5, 15]" ...
 $ cap.shape           : chr  "[x, f]" "[p, x]" "[x, f]" "[x, f]" ...
 $ Cap.surface         : chr  "[g, h]" "[g]" "" "" ...
 $ cap.color           : chr  "[e, o]" "[n]" "[g, n]" "[n]" ...
 $ does.bruise.or.bleed: chr  "[f]" "[f]" "[f]" "[t]" ...
 $ gill.attachment     : chr  "[e]" "[e]" "[e]" "" ...
 $ gill.spacing        : chr  "" "" "" "" ...
 $ gill.color          : chr  "[w]" "[w]" "[w]" "[w]" ...
 $ stem.height         : chr  "[15, 20]" "[6, 10]" "[10, 12]" "[7, 15]" ...
 $ stem.width          : chr  "[15, 20]" "[10, 20]" "[10, 20]" "[10, 25]" ...
 $ stem.root           : chr  "[s]" "" "" "[b]" ...
 $ stem.surface        : chr  "[y]" "[y]" "" "" ...
 $ stem.color          : chr  "[w]" "[w]" "[w]" "[w]" ...
 $ veil.type           : chr  "[u]" "[u]" "[u]" "[u]" ...
 $ veil.color          : chr  "[w]" "[w]" "[w]" "[w]" ...
 $ has.ring            : chr  "[t]" "[t]" "[t]" "[t]" ...
 $ ring.type           : chr  "[g, p]" "[p]" "[e, g]" "[g]" ...
 $ Spore.print.color   : chr  "" "" "" "" ...
 $ habitat             : chr  "[d]" "[d]" "[d]" "[d]" ...
 $ season              : chr  "[u, a, w]" "[u, a]" "[u, a]" "[u, a]" ...
```

```
#mushroom_data$season

#
extract_season <- function(x, position = "first") {
  #
  seasons <- unlist(strsplit(gsub("\\[|\\]", "", x), ", "))

  if (position == "first") {
    return(seasons[1])  #
  } else if (position == "last") {
    return(seasons[length(seasons)])  #
  } else {
    return(NA)  #
  }
}

#     sta_season   end_season
```

```r
mushroom_data$sta_season <- sapply(mushroom_data$season, extract_season, position = "first")
mushroom_data$end_season <- sapply(mushroom_data$season, extract_season, position = "last")
```

table one

- for numerical data ,we compute the mean first ,and create the table1

```r
mushroom_data$does.bruise.or.bleed <- factor(mushroom_data$does.bruise.or.bleed,
    levels = c("[t]", "[f]", ""),
    labels = c("Bruises or Bleeds", "No", "None"))

mushroom_data$gill.attachment <- factor(mushroom_data$gill.attachment,
    levels = c("[a]", "[x]", "[d]", "[e]", "[s]", "[p]", "[?]", ""),
    labels = c("Adnate", "Adnexed", "Decurrent", "Free", "Sinuate", "Pores", "Unknown", "None"))

mushroom_data$gill.spacing <- factor(mushroom_data$gill.spacing,
    levels = c("[c]", "[d]", "[f]", ""),
    labels = c("Close", "Distant", "None", "None"))

mushroom_data$stem.root <- factor(mushroom_data$stem.root,
    levels = c("[b]", "[s]", "[c]", "[u]", "[e]", "[z]", "[r]", ""),
    labels = c("Bulbous", "Swollen", "Club", "Cup", "Equal", "Rhizomorphs", "Rooted", "None"))

mushroom_data$veil.type <- factor(mushroom_data$veil.type,
    levels = c("[p]", "[u]", ""),
    labels = c("Partial", "Universal", "None"))

mushroom_data$veil.color <- factor(mushroom_data$veil.color,
    levels = c("[n]", "[b]", "[g]", "[r]", "[p]", "[u]", "[e]", "[w]", "[y]", "[l]", "[o]", "[k]", ""),
    labels = c("Brown", "Buff", "Gray", "Green", "Pink", "Purple", "Red",
               "White", "Yellow", "Blue", "Orange", "Black", "None"))

mushroom_data$has.ring <- factor(mushroom_data$has.ring,
    levels = c("[t]", "[f]"),
    labels = c("Ring", "[f]"))

mushroom_data$Spore.print.color <- factor(mushroom_data$Spore.print.color,
    levels = c("[n]", "[b]", "[g]", "[r]", "[p]", "[u]", "[e]", "[w]", "[y]", "[l]", "[o]", "[k]", ""),
    labels = c("Brown", "Buff", "Gray", "Green", "Pink", "Purple", "Red",
               "White", "Yellow", "Blue", "Orange", "Black", "None"))

mushroom_data$sta_season <- factor(mushroom_data$sta_season,
    levels = c("s", "u", "a", "w"),
    labels = c("Spring", "Summer", "Autumn", "Winter"))

mushroom_data$end_season <- factor(mushroom_data$end_season,
    levels = c("s", "u", "a", "w"),
    labels = c("Spring", "Summer", "Autumn", "Winter"))

mushroom_data$class <- factor(mushroom_data$class,
    levels = c("p", "e"),
    labels = c("Poisonous", "Edible"))
```

```
# table one of categorical data
table1(~ does.bruise.or.bleed+gill.attachment
        +gill.spacing+stem.root|class, data=mushroom_data)
```

	Poisonous	Edible	Overall
	(N=96)	(N=77)	(N=173)
does.bruise.or.bleed			
Bruises or Bleeds	16 (16.7%)	14 (18.2%)	30 (17.3%)
No	80 (83.3%)	63 (81.8%)	143 (82.7%)
None	0 (0%)	0 (0%)	0 (0%)
gill.attachment			
Adnate	21 (21.9%)	11 (14.3%)	32 (18.5%)
Adnexed	12 (12.5%)	9 (11.7%)	21 (12.1%)
Decurrent	16 (16.7%)	9 (11.7%)	25 (14.5%)
Free	6 (6.3%)	10 (13.0%)	16 (9.2%)
Sinuate	9 (9.4%)	7 (9.1%)	16 (9.2%)
Pores	5 (5.2%)	12 (15.6%)	17 (9.8%)
Unknown	0 (0%)	0 (0%)	0 (0%)
None	18 (18.8%)	10 (13.0%)	28 (16.2%)
Missing	9 (9.4%)	9 (11.7%)	18 (10.4%)
gill.spacing			
Close	41 (42.7%)	29 (37.7%)	70 (40.5%)
Distant	9 (9.4%)	13 (16.9%)	22 (12.7%)
None	46 (47.9%)	35 (45.5%)	81 (46.8%)
stem.root			
Bulbous	3 (3.1%)	6 (7.8%)	9 (5.2%)
Swollen	5 (5.2%)	4 (5.2%)	9 (5.2%)
Club	2 (2.1%)	0 (0%)	2 (1.2%)
Cup	0 (0%)	0 (0%)	0 (0%)
Equal	0 (0%)	0 (0%)	0 (0%)
Rhizomorphs	0 (0%)	0 (0%)	0 (0%)
Rooted	4 (4.2%)	0 (0%)	4 (2.3%)
None	79 (82.3%)	67 (87.0%)	146 (84.4%)
Missing	3 (3.1%)	0 (0%)	3 (1.7%)

```
table1(~ veil.type+veil.color+has.ring
        +Spore.print.color+sta_season+end_season|class, data=mushroom_data)
```

	Poisonous	Edible	Overall
	(N=96)	(N=77)	(N=173)
veil.type			
Partial	0 (0%)	0 (0%)	0 (0%)
Universal	6 (6.3%)	3 (3.9%)	9 (5.2%)
None	90 (93.8%)	74 (96.1%)	164 (94.8%)
veil.color			
Brown	1 (1.0%)	0 (0%)	1 (0.6%)
Buff	0 (0%)	0 (0%)	0 (0%)
Gray	0 (0%)	0 (0%)	0 (0%)
Green	0 (0%)	0 (0%)	0 (0%)
Pink	0 (0%)	0 (0%)	0 (0%)
Purple	1 (1.0%)	0 (0%)	1 (0.6%)
Red	0 (0%)	0 (0%)	0 (0%)
White	8 (8.3%)	7 (9.1%)	15 (8.7%)
Yellow	0 (0%)	1 (1.3%)	1 (0.6%)
Blue	0 (0%)	0 (0%)	0 (0%)
Orange	0 (0%)	0 (0%)	0 (0%)
Black	1 (1.0%)	0 (0%)	1 (0.6%)
None	84 (87.5%)	68 (88.3%)	152 (87.9%)
Missing	1 (1.0%)	1 (1.3%)	2 (1.2%)
has.ring			
Ring	26 (27.1%)	17 (22.1%)	43 (24.9%)
[f]	70 (72.9%)	60 (77.9%)	130 (75.1%)
Spore.print.color			
Brown	3 (3.1%)	0 (0%)	3 (1.7%)
Buff	0 (0%)	0 (0%)	0 (0%)
Gray	0 (0%)	1 (1.3%)	1 (0.6%)
Green	0 (0%)	0 (0%)	0 (0%)
Pink	2 (2.1%)	1 (1.3%)	3 (1.7%)
Purple	0 (0%)	0 (0%)	0 (0%)
Red	0 (0%)	0 (0%)	0 (0%)
White	1 (1.0%)	2 (2.6%)	3 (1.7%)
Yellow	0 (0%)	0 (0%)	0 (0%)
Blue	0 (0%)	0 (0%)	0 (0%)
Orange	0 (0%)	0 (0%)	0 (0%)
Black	4 (4.2%)	1 (1.3%)	5 (2.9%)
None	83 (86.5%)	72 (93.5%)	155 (89.6%)
Missing	3 (3.1%)	0 (0%)	3 (1.7%)
sta_season			
Spring	11 (11.5%)	12 (15.6%)	23 (13.3%)
Summer	68 (70.8%)	51 (66.2%)	119 (68.8%)
Autumn	17 (17.7%)	14 (18.2%)	31 (17.9%)
Winter	0 (0%)	0 (0%)	0 (0%)
end_season			
Spring	0 (0%)	1 (1.3%)	1 (0.6%)
Summer	2 (2.1%)	2 (2.6%)	4 (2.3%)
Autumn	78 (81.3%)	49 (63.6%)	127 (73.4%)
Winter	16 (16.7%)	25 (32.5%)	41 (23.7%)

```r
calculate_mean <- function(x) {
  nums <- as.numeric(unlist(strsplit(gsub("\\[|\\]", "", x), ",")))
  mean(nums, na.rm = TRUE)
}

mushroom_data$cap.diameter_num <- sapply(mushroom_data$cap.diameter, calculate_mean)
mushroom_data$stem.height_num <- sapply(mushroom_data$stem.height, calculate_mean)
mushroom_data$stem.width_num <- sapply(mushroom_data$stem.width, calculate_mean)

# table one of numerical data
table1(~ cap.diameter_num + stem.height_num + stem.width_num|class, data = mushroom_data)
```

	Poisonous	Edible	Overall
	(N=96)	(N=77)	(N=173)
cap.diameter_num			
Mean (SD)	5.88 (3.85)	7.81 (6.26)	6.74 (5.14)
Median [Min, Max]	5.00 [0.700, 19.0]	6.50 [1.00, 50.0]	6.00 [0.700, 50.0]
stem.height_num			
Mean (SD)	6.22 (3.05)	7.05 (3.48)	6.59 (3.26)
Median [Min, Max]	5.50 [0, 17.5]	6.00 [2.50, 25.0]	6.00 [0, 25.0]
stem.width_num			
Mean (SD)	10.4 (8.66)	14.4 (10.8)	12.2 (9.86)
Median [Min, Max]	7.50 [0, 40.0]	12.5 [1.00, 70.0]	10.0 [0, 70.0]

```r
mushroom_data$cap.diameter_num <- NULL
mushroom_data$stem.height_num <- NULL
mushroom_data$stem.width_num <- NULL
```

NA and value count

```r
# 1.    -
mushroom_summary <- function(data) {

  #      NA
  na_counts <- colSums(is.na(data))
  print("NA count of each variable:")
  print(na_counts)

  #
  cat("\nvalue count of categorical variable:\n")
  excluded_features <- c("cap-diameter", "stem-height", "stem-width", "name")

  for (col in names(data)) {
    if (!(col %in% excluded_features) && (is.factor(data[[col]]) || is.character(data[[col]]))) {
      n_categories <- length(unique(data[[col]][!is.na(data[[col]])]))
      cat(col, ": ", n_categories, "   \n", sep="")
    }
  }
}
```

```
mushroom_summary(mushroom_data)
```

```
[1] "NA count of each variable:"
              family                 name                class
                   0                    0                    0
       cap.diameter            cap.shape          Cap.surface
                   0                    0                    0
          cap.color does.bruise.or.bleed      gill.attachment
                   0                    0                   18
       gill.spacing           gill.color          stem.height
                   0                    0                    0
         stem.width            stem.root         stem.surface
                   0                    3                    0
         stem.color            veil.type           veil.color
                   0                    0                    2
           has.ring            ring.type    Spore.print.color
                   0                    0                    3
            habitat               season           sta_season
                   0                    0                    0
         end_season
                   0
```

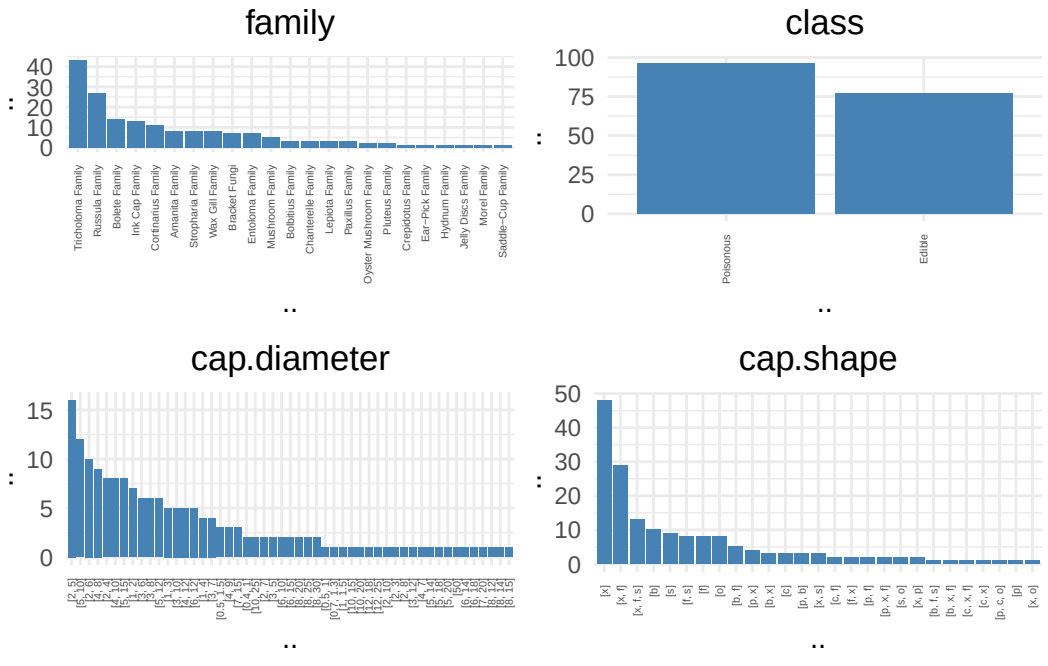
```
value count of categorical variable:
family: 23
class: 2
cap.diameter: 51
cap.shape: 27
Cap.surface: 41
cap.color: 67
does.bruise.or.bleed: 2
gill.attachment: 7
gill.spacing: 3
gill.color: 59
stem.height: 46
stem.width: 48
stem.root: 5
stem.surface: 15
stem.color: 41
veil.type: 2
veil.color: 6
has.ring: 2
ring.type: 14
Spore.print.color: 6
habitat: 21
season: 10
sta_season: 3
end_season: 4
```

Bar plot of data

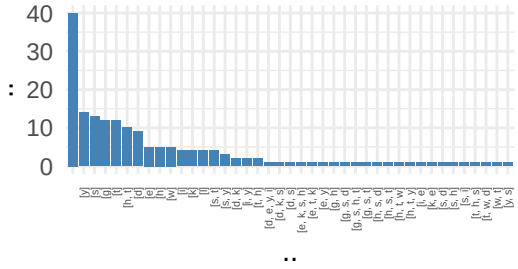
```r
plot_categorical_ggplots <- function(data) {
  #
  excluded_features <- c("cap-diameter", "stem-height", "stem-width", "name")

  #
  categorical_vars <- c()
  for (col in names(data)) {
    if (!(col %in% excluded_features) && (is.factor(data[[col]]) || is.character(data[[col]]))) {
      categorical_vars <- c(categorical_vars, col)
    }
  }

  plot_list <- list()

  for (i in 1:length(categorical_vars)) {
    var <- categorical_vars[i]

    #
    freq_df <- as.data.frame(table(data[[var]], useNA = "ifany"))
    colnames(freq_df) <- c("Category", "Count")

    #   NA
    freq_df$Category <- as.character(freq_df$Category)
    freq_df$Category[is.na(freq_df$Category)] <- "NA"

    #
    freq_df <- freq_df[order(-freq_df$Count),]

    #
    has_special_categories <- any(nchar(as.character(freq_df$Category)) > 10 |
                                  grepl("\\[.*,.*\\]", as.character(freq_df$Category)))

    #
    text_size <- ifelse(has_special_categories, 5, 6)

    #
    p <- ggplot(freq_df, aes(x = reorder(Category, -Count), y = Count)) +
      geom_bar(stat = "identity", fill = "steelblue") +
      labs(title = var, x = " ", y = " ") +
      theme_minimal() +
      theme(
        #   x
        axis.text.x = element_text(
          angle = 90,
          hjust = 1,
          vjust = 0.5,
          size = 4
        ),
        plot.title = element_text(hjust = 0.5),
        #
        plot.margin = margin(t = 5, r = 1, b = 5, l = 1)
      )
```

```
    plot_list[[i]] <- p
  }

  #     4  2 2
  plots_per_page <- 4
  n_pages <- ceiling(length(plot_list) / plots_per_page)

  #
  for (page in 1:n_pages) {
    start_idx <- (page - 1) * plots_per_page + 1
    end_idx <- min(page * plots_per_page, length(plot_list))

    #
    if (start_idx <= length(plot_list)) {
      current_plots <- plot_list[start_idx:end_idx]

      #
      grid.arrange(
        grobs = current_plots,
        ncol = 2,
        nrow = 2
      )
    }
  }
}
```

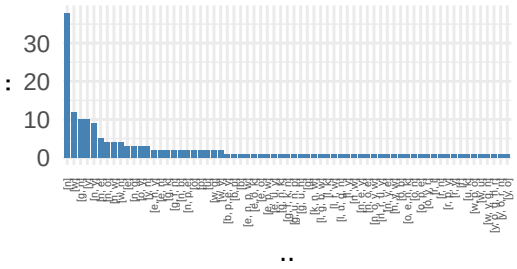
```
plot_categorical_ggplots(mushroom_data)
```



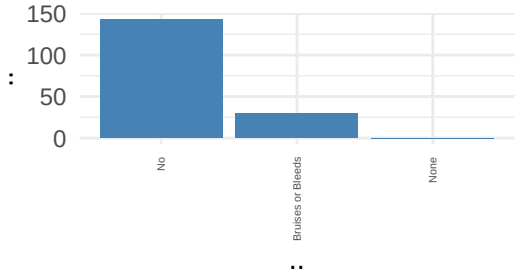
Cap.surface



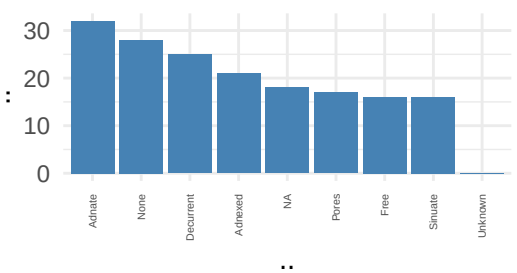
cap.color



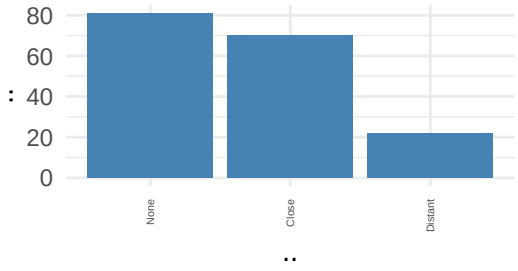
does.bruise.or.bleed



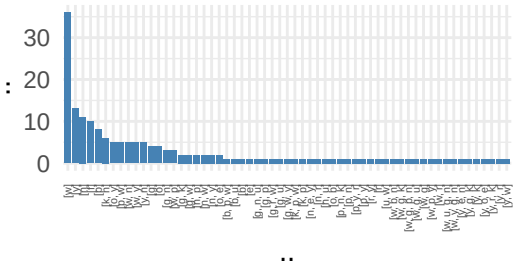
gill.attachment



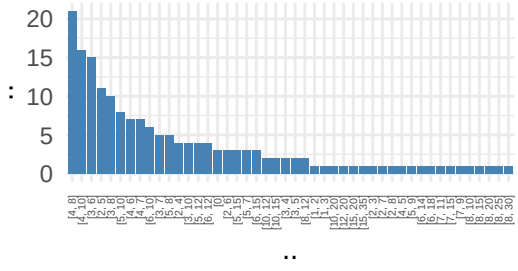
gill.spacing



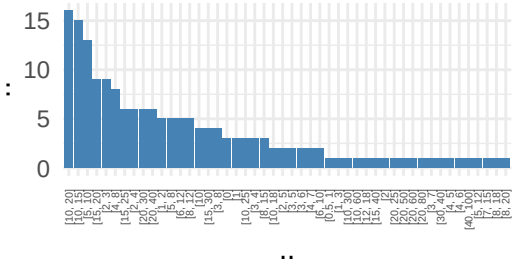
gill.color

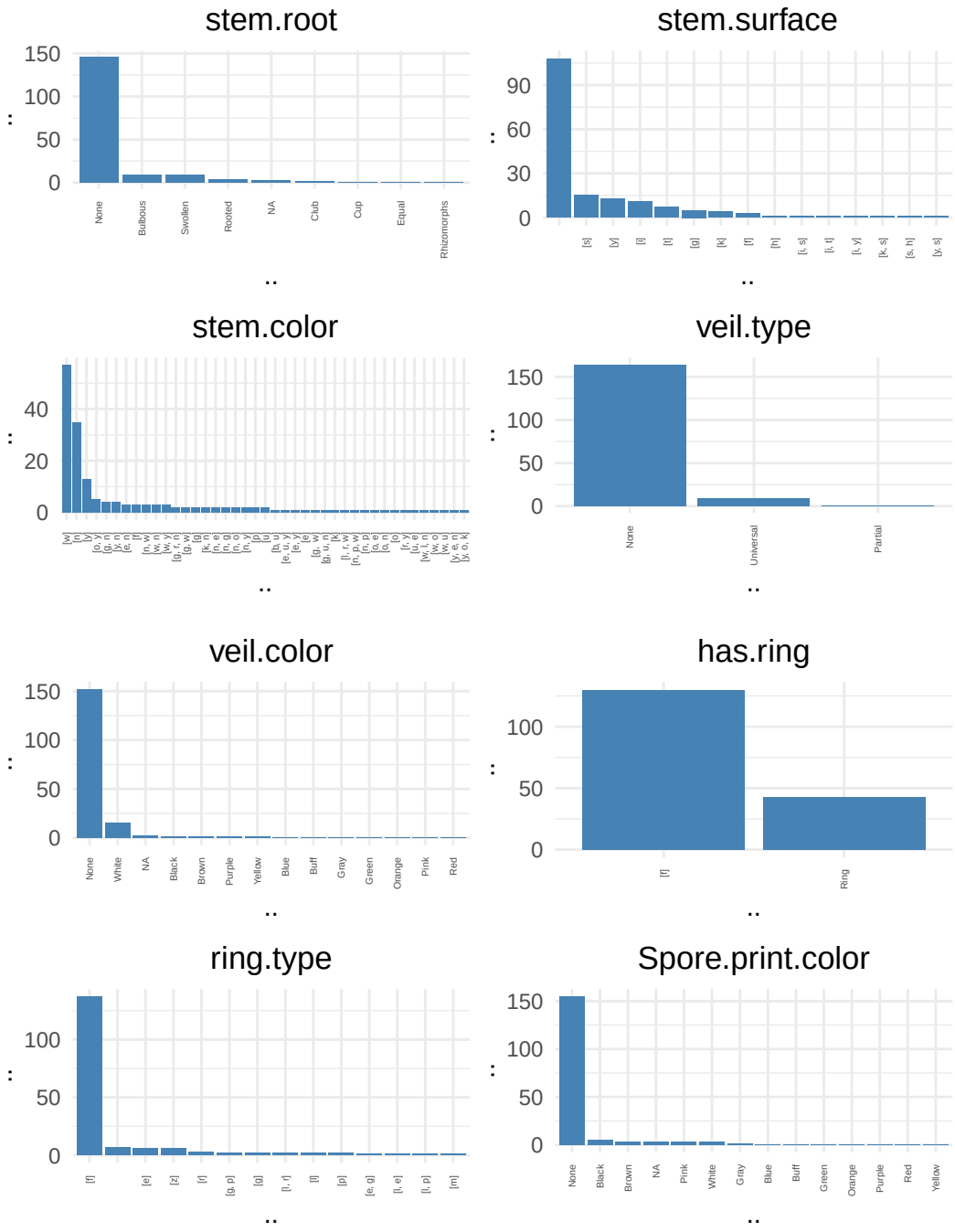


stem.height

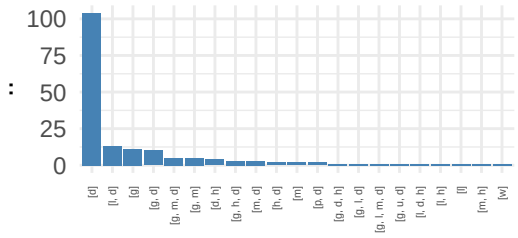


stem.width



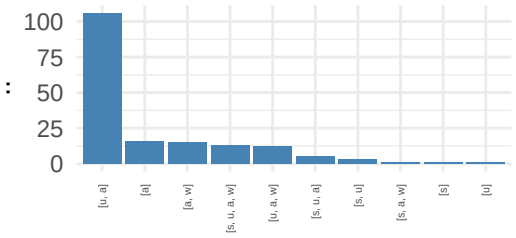

stem.root
150
100
50
0
None
Bulbous
Swollen
Rooted
NA
Club
Cup
Equal
Rhizomorphs
stem.surface
90
60
30
0
stem.color
40
20
0
veil.type
150
100
50
0
None
Universal
Partial
veil.color
150
100
50
0
None
White
NA
Black
Brown
Purple
Yellow
Blue
Buff
Gray
Green
Orange
Pink
Red
has.ring
100
50
0
[f]
Ring
ring.type
100
50
0
[f]
[e]
[z]
[r]
[g, p]
[g]
[l, r]
[l]
[p]
[e, g]
[l, e]
[l, p]
[m]
Spore.print.color
150
100
50
0
None
Black
Brown
NA
Pink
White
Gray
Blue
Buff
Green
Orange
Purple
Red
Yellow

habitat



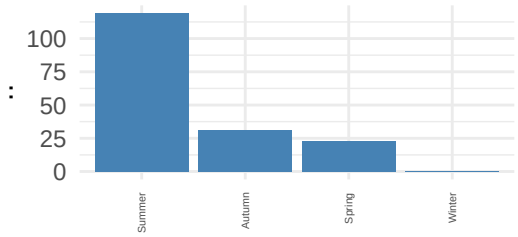
..

season



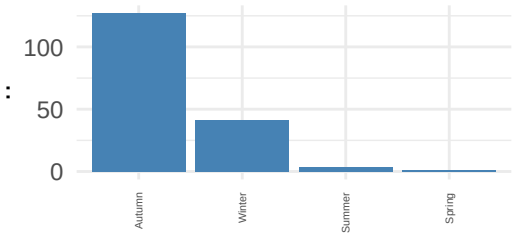
..

sta_season



..

end_season



..