HW2

Summary of the Mushroom Dataset

Cheng-Jun Kang

2025-03-19

Table of contents

I. Data Loading and Basic Statistical Analysis 1

II. Mushroom Dataset Statistical Analysis 8

I. Data Loading and Basic Statistical Analysis

```
library(tinytex)
library(tidyverse)
library(ggplot2)
library(reticulate)
library(Hmisc)
library(dplyr)
#library(tinytex)

#tinytex:::is_tinytex()  # Should return TRUE if properly installed

# Load the mushroom dataset
mushroom <- read.csv('C:\\Users\\cjkan\\OneDrive\\Desktop\\CJK\\113_2\\SC\\mushroom\\primary
#colnames(mushroom)
#names(mushroom)

# Display the basic structure of the dataset
str(mushroom)
```

```
'data.frame':   173 obs. of  23 variables:
 $ family              : chr  "Amanita Family" "Amanita Family" "Amanita Family" "Amanita Fa
 $ 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]" ...
```

```r
# Adjust code based on actual array names
mushroom_analysis <- mushroom %>%
  mutate(cap.shape = gsub("\\[|\\]", "", cap.shape),
         cap.shape = strsplit(cap.shape, ", ")) %>%
  unnest(cap.shape) %>%
  group_by(cap.shape, class) %>%
  summarise(count = n(), .groups = 'drop')

# Show dimensions
dim(mushroom)
```

```
[1] 173  23
```

```r
head(mushroom)
```

```
          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]
```

```r
# Display basic statistics
summary(mushroom)
```

```
   family             name              class           cap.diameter
Length:173         Length:173         Length:173         Length:173
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character
 cap.shape         Cap.surface         cap.color         does.bruise.or.bleed
Length:173         Length:173         Length:173         Length:173
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character
gill.attachment    gill.spacing        gill.color        stem.height
Length:173         Length:173         Length:173         Length:173
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character
 stem.width         stem.root         stem.surface        stem.color
Length:173         Length:173         Length:173         Length:173
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character
 veil.type          veil.color          has.ring          ring.type
Length:173         Length:173         Length:173         Length:173
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character
Spore.print.color    habitat             season
```

```
 Length:173         Length:173         Length:173
 Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character
```

```
describe(mushroom)
```

```
mushroom

 23  Variables      173  Observations
--------------------------------------------------------------------------------
family
       n  missing distinct
     173        0       23

lowest : Amanita Family     Bolbitius Family   Bolete Family      Bracket Fungi      Chanter
highest: Russula Family     Saddle-Cup Family  Stropharia Family  Tricholoma Family  Wax Gil
--------------------------------------------------------------------------------
name
       n  missing distinct
     173        0      173

lowest : Amethyst Deceiver        Aniseed Funnel Cap       Apricot Fungus           Bare-too
highest: Yellow-gilled Russula    Yellow-staining Mushroom Yellow-stemmed Bell Cap  Yellow S
--------------------------------------------------------------------------------
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]
Frequency       3      5      1      1      1      1     11      1      4
Proportion  0.046  0.077  0.015  0.015  0.015  0.015  0.169  0.015  0.062

Value      [s, h]    [s]    [t] [y, s]    [y]
Frequency       1     15      7      1     13
Proportion  0.015  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, 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]
Frequency       1      6    137      2      2      1      1      2      2
Proportion  0.006  0.036  0.825  0.012  0.012  0.006  0.006  0.012  0.012

Value         [m]    [p]    [r]    [z]
Frequency       1      2      3      6
Proportion  0.006  0.012  0.018  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]
Frequency            15           16            1           13            5
Proportion        0.087        0.092        0.006        0.075        0.029

Value            [s, u]          [s]    [u, a, w]       [u, a]          [u]
Frequency             3            1           12          106            1
Proportion        0.017        0.006        0.069        0.613        0.006
--------------------------------------------------------------------------------
```

```r
# Select main analysis variables
selected_vars <- c(
  "family",           # Mushroom family (multinomial)
  "class",            # Edibility: p=poisonous, e=edible (binary)
  "cap.shape",        # Cap shape: b=bell, c=conical, x=convex, f=flat, s=sunken, p=spherica
  "cap.color",        # Cap color: n=brown, w=white, y=yellow, etc.
  "does.bruise.or.bleed", # Whether it bruises/bleeds: t=yes, f=no
  "habitat",          # Growing environment: g=grasses, l=leaves, m=meadows, d=woods, etc.
  "season"            # Growing season: s=spring, u=summer, a=autumn, w=winter
)

# Check if selected variables exist in the dataset
all(selected_vars %in% colnames(mushroom))
```

```
[1] TRUE
```

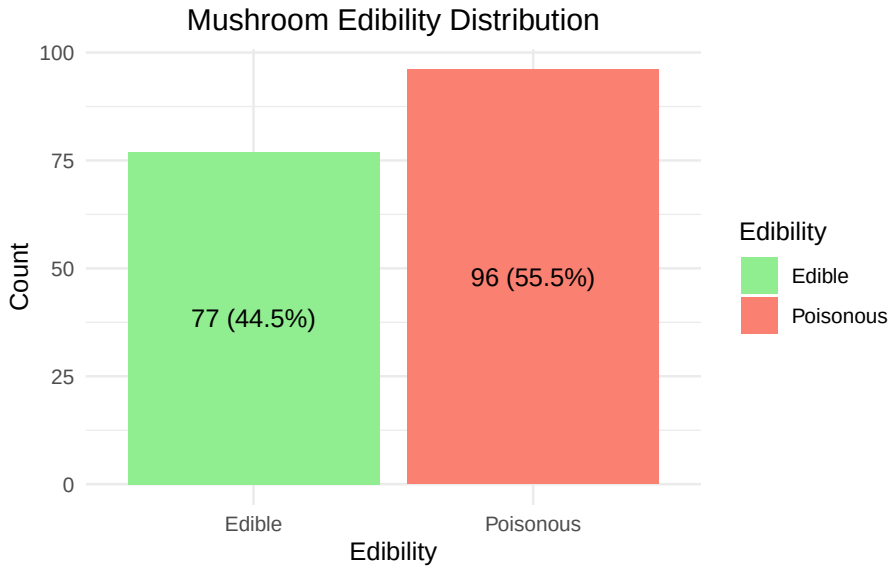
```r
# View basic information for each selected variable
sapply(mushroom[selected_vars], function(x) length(unique(x)))
```

```
             family               class           cap.shape
                 23                   2                  27
          cap.color does.bruise.or.bleed             habitat
                 67                   2                  21
             season
                 10
```

II. Mushroom Dataset Statistical Analysis

```r
# Calculate edibility distribution
edibility_count <- mushroom %>%
  group_by(class) %>%
  summarise(count = n()) %>%
  mutate(percentage = count / sum(count) * 100,
         class = factor(class, levels = c("e", "p"),
                        labels = c("Edible", "Poisonous")))

# Plot edibility distribution
ggplot(edibility_count, aes(x = class, y = count, fill = class)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = sprintf("%d (%.1f%%)", count, percentage)),
            position = position_stack(vjust = 0.5)) +
  scale_fill_manual(values = c("#90EE90", "#FA8072"),
                    name = "Edibility") +
  labs(title = "Mushroom Edibility Distribution",
       x = "Edibility",
       y = "Count") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```




```r
# Analyze mushroom family distribution
family_distribution <- mushroom %>%
  group_by(family) %>%
  summarise(count = n(),
```

```
            poisonous = sum(class == "p"),
            edible = sum(class == "e"),
            poisonous_rate = poisonous / count * 100) %>%
  arrange(desc(count))

# Plot distribution of major mushroom families (top 10)
top_families <- family_distribution %>%
  top_n(10, count)

ggplot(top_families, aes(x = reorder(family, count), y = count, fill = poisonous_rate)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = count), hjust = -0.2) +
  scale_fill_gradient(low = "#90EE90", high = "#FA8072", name = "Poisonous Rate (%)") +
  labs(title = "Distribution of Major Mushroom Families",
       x = "Family",
       y = "Count") +
  coord_flip() +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```