Palestine Technical University – Kadoorie

Faculty of Engineering and Technology

Computer Systems Department

Data Mining

Mushroom Dataset Analysis Report

By:

Omar Abbadi 201911226

Moawiah Ararawi 201910669

Mahmoud Khaled Qasem 201910708

Supervisor: Dr. Anas Melhem

1. Introduction

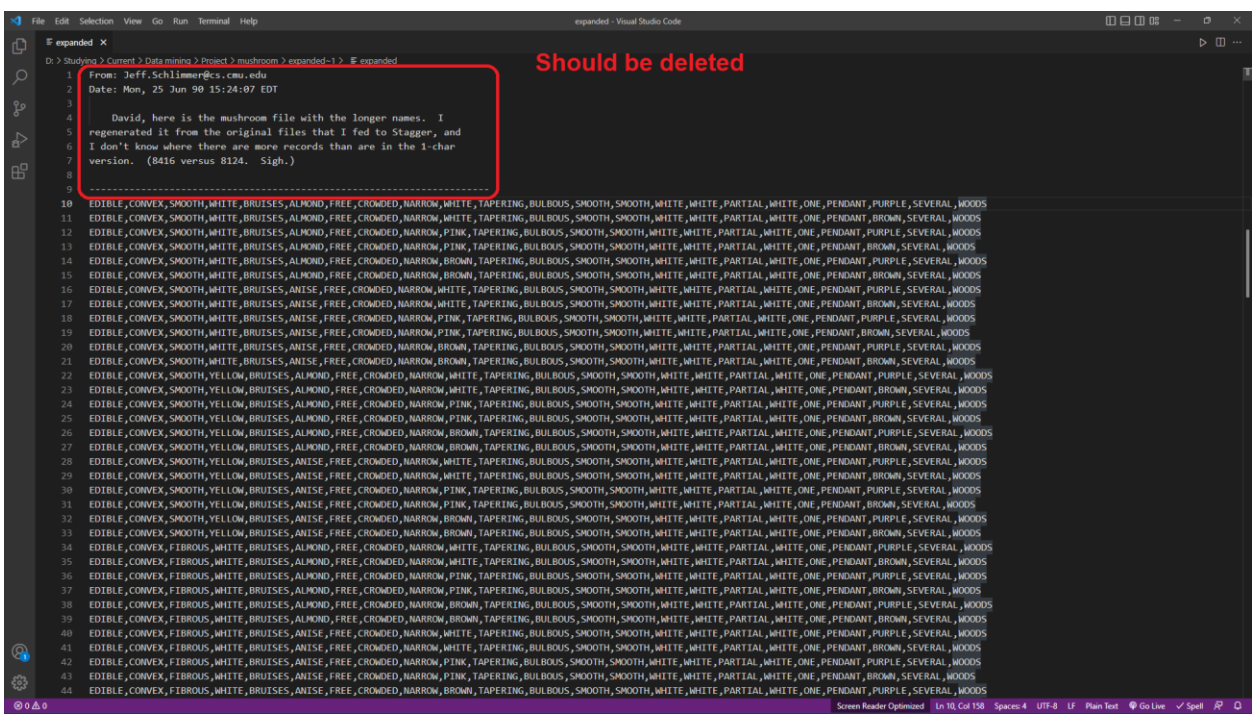
The purpose of this work is to analyze the "Mushroom" dataset using R language, this dataset is taken from UCI Machine Learning Repository, and the main aim of this study is to predict whether the Mushroom is safe to eat (edible) or poisoned (poisonous). In our analysis, first, we will prepare the data, find out missing values and how to deal with them, then, apply one of the classification data mining algorithms.

1.1 Dataset Description

- Number of Instances: **8415**
- Number of Attributes: **23 (all nominally valued)**
- Missing Attribute Values: **2480**

Dataset link: https://archive.ics.uci.edu/ml/datasets/Mushroom

When we installed the expanded file of the dataset (expanded.txt) and opened it, we found a paragraph and dashed lines that caused a problem when reading the file in the CSV format, so we deleted them and converted the file to CSV in order to load it to RStudio as shown below:



```
EDIBLE,KNOBBED,SMOOTH,BROWN,NO,NONE,ATTACHED,CLOSE,BROAD,BROWN,ENLARGING,?,SMOOTH,SMOOTH,ORANGE,ORANGE,PARTIAL,ORANGE,ONE,PENDANT,BUFF,CLUSTERED,LEAVES
EDIBLE,KNOBBED,SMOOTH,BROWN,NO,NONE,ATTACHED,CLOSE,BROAD,BROWN,ENLARGING,?,SMOOTH,SMOOTH,ORANGE,ORANGE,PARTIAL,BROWN,ONE,PENDANT,YELLOW,SEVERAL,LEAVES
EDIBLE,KNOBBED,SMOOTH,BROWN,NO,NONE,ATTACHED,CLOSE,BROAD,BROWN,ENLARGING,?,SMOOTH,SMOOTH,ORANGE,ORANGE,PARTIAL,BROWN,ONE,PENDANT,YELLOW,CLUSTERED,LEAVES
EDIBLE,KNOBBED,SMOOTH,BROWN,NO,NONE,ATTACHED,CLOSE,BROAD,BROWN,ENLARGING,?,SMOOTH,SMOOTH,ORANGE,ORANGE,PARTIAL,BROWN,ONE,PENDANT,BROWN,SEVERAL,LEAVES
EDIBLE,KNOBBED,SMOOTH,BROWN,NO,NONE,ATTACHED,CLOSE,BROAD,BROWN,ENLARGING,?,SMOOTH,SMOOTH,ORANGE,ORANGE,PARTIAL,BROWN,ONE,PENDANT,BROWN,CLUSTERED,LEAVES
EDIBLE,KNOBBED,SMOOTH,BROWN,NO,NONE,ATTACHED,CLOSE,BROAD,BROWN,ENLARGING,?,SMOOTH,SMOOTH,ORANGE,ORANGE,PARTIAL,BROWN,ONE,PENDANT,ORANGE,SEVERAL,LEAVES
EDIBLE,KNOBBED,SMOOTH,BROWN,NO,NONE,ATTACHED,CLOSE,BROAD,BROWN,ENLARGING,?,SMOOTH,SMOOTH,ORANGE,ORANGE,PARTIAL,BROWN,ONE,PENDANT,ORANGE,CLUSTERED,LEAVES
EDIBLE,KNOBBED,SMOOTH,BROWN,NO,NONE,ATTACHED,CLOSE,BROAD,BROWN,ENLARGING,?,SMOOTH,SMOOTH,ORANGE,ORANGE,PARTIAL,BROWN,ONE,PENDANT,BUFF,SEVERAL,LEAVES
EDIBLE,KNOBBED,SMOOTH,BROWN,NO,NONE,ATTACHED,CLOSE,BROAD,BROWN,ENLARGING,?,SMOOTH,SMOOTH,ORANGE,ORANGE,PARTIAL,BROWN,ONE,PENDANT,BUFF,CLUSTERED,LEAVES
-----------------------------------------------------------------
```

Also this

Code for loading data to Rstudio:

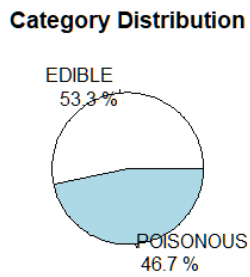
```r
# Loading the data (the expanded version) from storage
data <- read.csv('c:/DFrame/expanded.csv')

# Configuring columns names
columns <- c('class', 'cap-shape', 'cap-surface', 'cap-color', 'bruises', 'odor', 'gill-attachment', 'gill-spacing',
             'gill-size', 'gill-color', 'stalk-shape', 'stalk-root', 'stalk-surface-above-ring', 'stalk-surface-below-ring',
             'stalk-color-above-ring', 'stalk-color-below-ring', 'veil-type', 'veil-color', 'ring-number', 'ring-type',
             'spore-print-color', 'population', 'habitat')

# Setting the columns names into the data set
colnames(data) <- columns

# A summary of the current state of the data set
summary(data)
str(data)
```

Output:

```
     class             cap-shape          cap-surface         cap-color           bruises
 Length:8415        Length:8415        Length:8415        Length:8415        Length:8415
 Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character
      odor           gill-attachment    gill-spacing        gill-size          gill-color
 Length:8415        Length:8415        Length:8415        Length:8415        Length:8415
 Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character
 stalk-shape         stalk-root         stalk-surface-above-ring stalk-surface-below-ring
 Length:8415        Length:8415        Length:8415              Length:8415
 Class :character   Class :character   Class :character         Class :character
 Mode  :character   Mode  :character   Mode  :character         Mode  :character
 stalk-color-above-ring stalk-color-below-ring  veil-type          veil-color
 Length:8415            Length:8415             Length:8415        Length:8415
 Class :character       Class :character        Class :character   Class :character
 Mode  :character       Mode  :character        Mode  :character   Mode  :character
 ring-number         ring-type          spore-print-color   population          habitat
 Length:8415        Length:8415        Length:8415        Length:8415        Length:8415
 Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character
> str(data)
'data.frame':   8415 obs. of  23 variables:
 $ class                 : chr  "EDIBLE" "EDIBLE" "EDIBLE" "EDIBLE" ...
 $ cap-shape             : chr  "CONVEX" "CONVEX" "CONVEX" "CONVEX" ...
 $ cap-surface           : chr  "SMOOTH" "SMOOTH" "SMOOTH" "SMOOTH" ...
 $ cap-color             : chr  "WHITE" "WHITE" "WHITE" "WHITE" ...
 $ bruises               : chr  "BRUISES" "BRUISES" "BRUISES" "BRUISES" ...
 $ odor                  : chr  "ALMOND" "ALMOND" "ALMOND" "ALMOND" ...
 $ gill-attachment       : chr  "FREE" "FREE" "FREE" "FREE" ...
 $ gill-spacing          : chr  "CROWDED" "CROWDED" "CROWDED" "CROWDED" ...
 $ gill-size             : chr  "NARROW" "NARROW" "NARROW" "NARROW" ...
 $ gill-color            : chr  "WHITE" "PINK" "PINK" "BROWN" ...
 $ stalk-shape           : chr  "TAPERING" "TAPERING" "TAPERING" "TAPERING" ...
 $ stalk-root            : chr  "BULBOUS" "BULBOUS" "BULBOUS" "BULBOUS" ...
 $ stalk-surface-above-ring: chr  "SMOOTH" "SMOOTH" "SMOOTH" "SMOOTH" ...
 $ stalk-surface-below-ring: chr  "SMOOTH" "SMOOTH" "SMOOTH" "SMOOTH" ...
 $ stalk-color-above-ring  : chr  "WHITE" "WHITE" "WHITE" "WHITE" ...
 $ stalk-color-below-ring  : chr  "WHITE" "WHITE" "WHITE" "WHITE" ...
 $ veil-type             : chr  "PARTIAL" "PARTIAL" "PARTIAL" "PARTIAL" ...
 $ veil-color            : chr  "WHITE" "WHITE" "WHITE" "WHITE" ...
 $ ring-number           : chr  "ONE" "ONE" "ONE" "ONE" ...
 $ ring-type             : chr  "PENDANT" "PENDANT" "PENDANT" "PENDANT" ...
 $ spore-print-color     : chr  "BROWN" "PURPLE" "BROWN" "PURPLE" ...
 $ population            : chr  "SEVERAL" "SEVERAL" "SEVERAL" "SEVERAL" ...
 $ habitat               : chr  "WOODS" "WOODS" "WOODS" "WOODS" ...
```

1.2 Class Distribution

- Edible: 4485 (53.3%)
- Poisonous: 3930 (46.6%)



Code for finding class distribution:

```
# A summary of the class and plot it
str(data$class)
summary(data$class)
install.packages("plotrix")
library(plotrix)
x<-table(data$class)
pie(x)
# Calculate the percentage of each category
x_percents <- prop.table(x) * 100
# Create the pie chart
pie(x, labels=paste(names(x), "\n", round(x_percents, 1), "%"), main="Category Distribution")
```

2. Data Preparation (pre-processing)

In this section, we plotted a histogram for each attribute and their class to identify the count of observations according to edibility. The aim of doing this is to find the attributes which are exclusive only in either class. More exclusiveness means a stronger correlation between the attribute and the edibility of the mushroom.

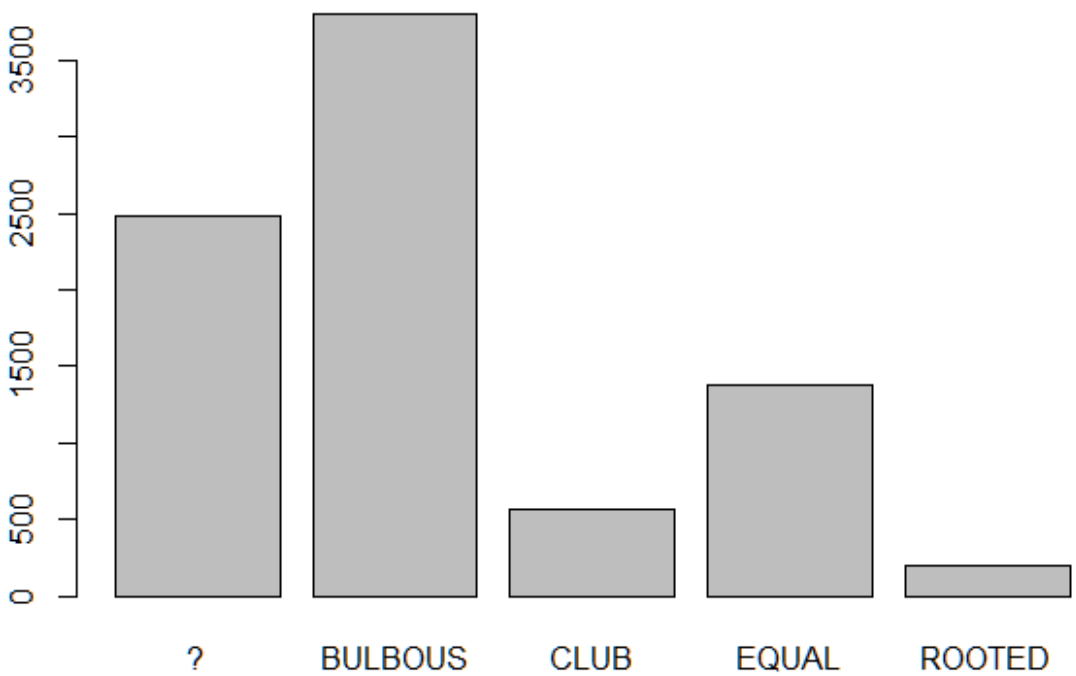
2.1 Finding out Missing Values

In our dataset, missing values are represented as '?' and are present in one column which is "stalk-root".

Code for counting missing values, and plotting the distribution of the "stalk-root" values with regard to the class.

```
# Missing values are represented by '?'
sum(data == '?')

# All missing values are in the 'stalk-root' attribute
# and this is their distribution
counts <- table(data$`stalk-root`)
counts # To make sure that 2480 question marks are only in this attribute
barplot(counts)

# To make sure there are 0 NA
sum(is.na(data)) # prints 0
```

Plot for the attribute "stalk-root" that has missing values:



2.2 Plots

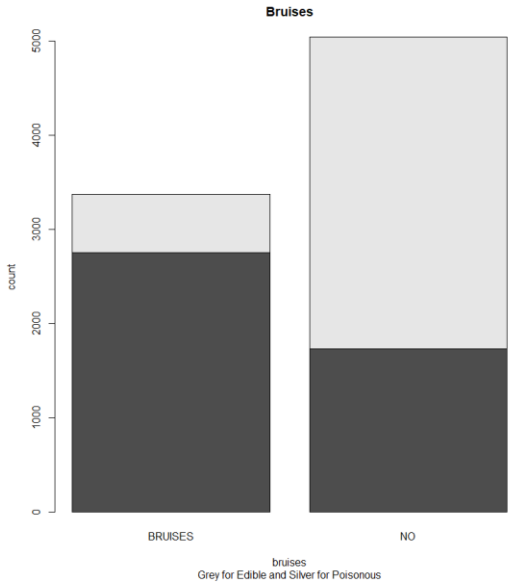
In this section we plotted each attribute with regard to the class to visualize data in a way that is easily understandable and to explore data and identify outliers, anomalies, or other unusual patterns, as following:

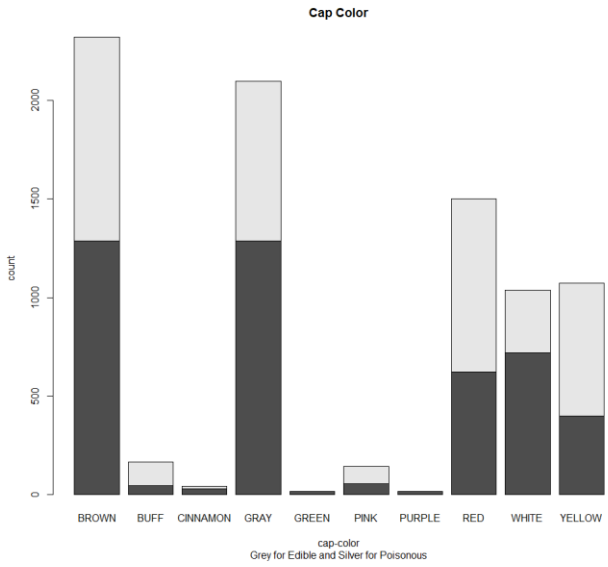
```r
library(dplyr)

### Plotting each attribute with regard to the class

# cap-shape
cap_shape_df <- select(data, class, `cap-shape`)
table(cap_shape_df)
barplot(table(cap_shape_df), xlab = 'cap-shape', ylab = 'count', sub = 'Grey for Edible and Silver for Poisonous', main = 'Cap Shape')

# cap-surface
cap_surface_df <- select(data, class, `cap-surface`)
table(cap_surface_df)
barplot(table(cap_surface_df), xlab = 'cap-surface', ylab = 'count', sub = 'Grey for Edible and Silver for Poisonous', main = 'Cap Surface')

# cap-color
cap_color_df <- select(data, class, `cap-color`)
table(cap_color_df)
barplot(table(cap_color_df), xlab = 'cap-color', ylab = 'count', sub = 'Grey for Edible and Silver for Poisonous', main = 'Cap Color')

# bruises
bruises_df <- select(data, class, bruises)
table(bruises_df)
barplot(table(bruises_df), xlab = 'bruises', ylab = 'count', sub = 'Grey for Edible and Silver for Poisonous', main = 'Bruises')

# odor
odor_df <- select(data, class, odor)
table(odor_df)
barplot(table(odor_df), xlab = 'odor', ylab = 'count', sub = 'Grey for Edible and Silver for Poisonous', main = 'Odor')

# gill-attachment
gill_attachment_df <- select(data, class, `gill-attachment`)
table(gill_attachment_df)
barplot(table(gill_attachment_df), xlab = 'gill-attachment', ylab = 'count', sub = 'Grey for Edible and Silver for Poisonous', main = 'Gill Attachment')

# gill-spacing
gill_spacing_df <- select(data, class, `gill-spacing`)
table(gill_spacing_df)
barplot(table(gill_spacing_df), xlab = 'gill-spacing', ylab = 'count', sub = 'Grey for Edible and Silver for Poisonous', main = 'Gill Spacing')
```

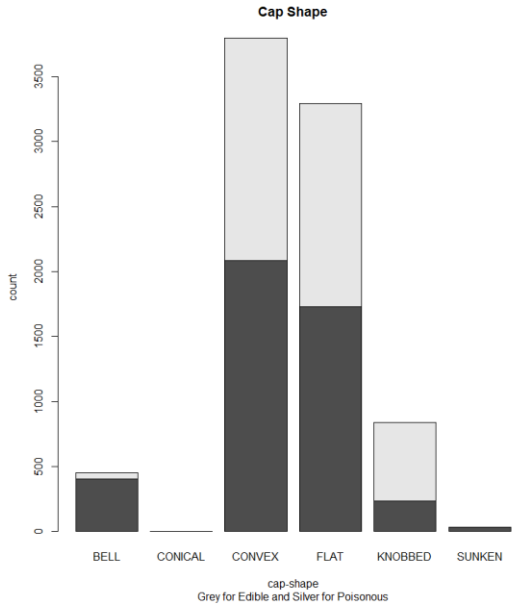
```r
# gill-size
gill_size_df <- select(data, class, `gill-size`)
table(gill_size_df)
barplot(table(gill_size_df), xlab = 'gill-size', ylab = 'count', sub = 'Grey for Edible and Silver for Poisonous', main = 'Gill Size')

# gill-color
gill_color_df <- select(data, class, `gill-color`)
table(gill_color_df)
barplot(table(gill_color_df), xlab = 'gill-color', ylab = 'count', sub = 'Grey for Edible and Silver for Poisonous', main = 'Gill Color')

# stalk-shape
stalk_shape_df <- select(data, class, `stalk-shape`)
table(stalk_shape_df)
barplot(table(stalk_shape_df), xlab = 'stalk-shape', ylab = 'count', sub = 'Grey for Edible and Silver for Poisonous', main = 'Stalk Shape')

# stalk-root
stalk_root_df <- select(data, class, `stalk-root`)
table(stalk_root_df)
barplot(table(stalk_root_df), xlab = 'stalk-root', ylab = 'count', sub = 'Grey for Edible and Silver for Poisonous', main = 'Stalk Root')

# stalk-surface-above-ring
stalk_surface_above_ring_df <- select(data, class, `stalk-surface-above-ring`)
table(stalk_surface_above_ring_df)
barplot(table(stalk_surface_above_ring_df), xlab = 'stalk-surface-above-ring',
        ylab = 'count', sub = 'Grey for Edible and Silver for Poisonous', main = 'Stalk Surface Above Ring')

# stalk-surface-below-ring
stalk_surface_below_ring_df <- select(data, class, `stalk-surface-below-ring`)
table(stalk_surface_below_ring_df)
barplot(table(stalk_surface_below_ring_df), xlab = 'stalk-surface-below-ring',
        ylab = 'count', sub = 'Grey for Edible and Silver for Poisonous', main = 'Stalk Surface Below Ring')

# stalk-color-above-ring
stalk_color_above_ring_df <- select(data, class, `stalk-color-above-ring`)
table(stalk_color_above_ring_df)
barplot(table(stalk_color_above_ring_df), xlab = 'stalk-color-above-ring',
        ylab = 'count', sub = 'Grey for Edible and Silver for Poisonous', main = 'Stalk Color Above Ring')
```

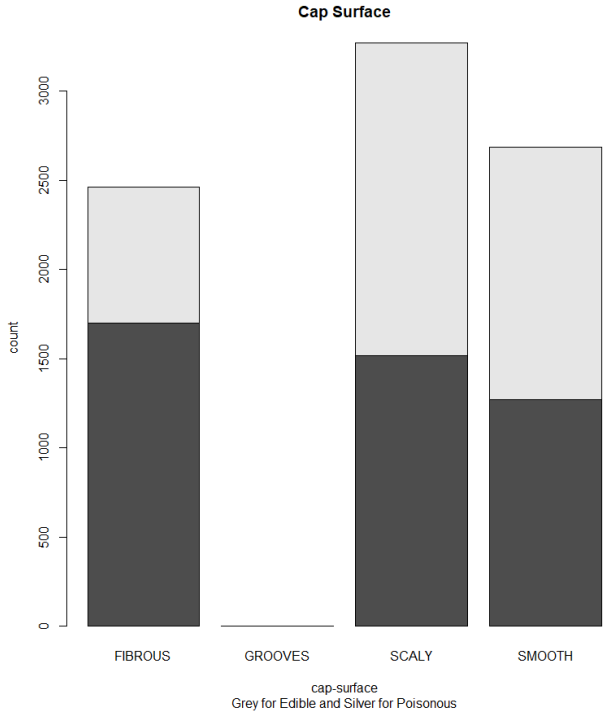
```r
# stalk-color-below-ring
stalk_color_below_ring_df <- select(data, class, `stalk-color-below-ring`)
table(stalk_color_below_ring_df)
barplot(table(stalk_color_below_ring_df), xlab = 'stalk-color-below-ring',
        ylab = 'count', sub = 'Grey for Edible and Silver for Poisonous', main = 'Stalk Color Below Ring')

# veil-type
veil_type_df <- select(data, class, `veil-type`)
table(veil_type_df)
barplot(table(veil_type_df), xlab = 'veil-type', ylab = 'count', sub = 'Grey for Edible and Silver for Poisonous', main = 'Veil Type')
unique(data$`veil-type`) # one value for this attribute -> should be deleted

# veil-color
veil_color_df <- select(data, class, `veil-color`)
table(veil_color_df)
barplot(table(veil_color_df), xlab = 'veil-color', ylab = 'count', sub = 'Grey for Edible and Silver for Poisonous', main = 'Veil Color')

# ring-number
ring_number_df <- select(data, class, `ring-number`)
table(ring_number_df)
barplot(table(ring_number_df), xlab = 'ring-number', ylab = 'count', sub = 'Grey for Edible and Silver for Poisonous', main = 'Ring Number')

# ring-type
ring_type_df <- select(data, class, `ring-type`)
table(ring_type_df)
barplot(table(ring_type_df), xlab = 'ring-type', ylab = 'count', sub = 'Grey for Edible and Silver for Poisonous', main = 'Ring Type')

# spore-print-color
spore_print_color_df <- select(data, class, `spore-print-color`)
table(spore_print_color_df)
barplot(table(spore_print_color_df), xlab = 'spore-print-color',
        ylab = 'count', sub = 'Grey for Edible and Silver for Poisonous', main = 'Spore Print Color')

# population
population_df <- select(data, class, `population`)
table(population_df)
barplot(table(population_df), xlab = 'population', ylab = 'count', sub = 'Grey for Edible and Silver for Poisonous', main = 'Population')

# habitat
habitat_df <- select(data, class, `habitat`)
table(habitat_df)
barplot(table(habitat_df), xlab = 'habitat', ylab = 'count', sub = 'Grey for Edible and Silver for Poisonous', main = 'Habitat')
```

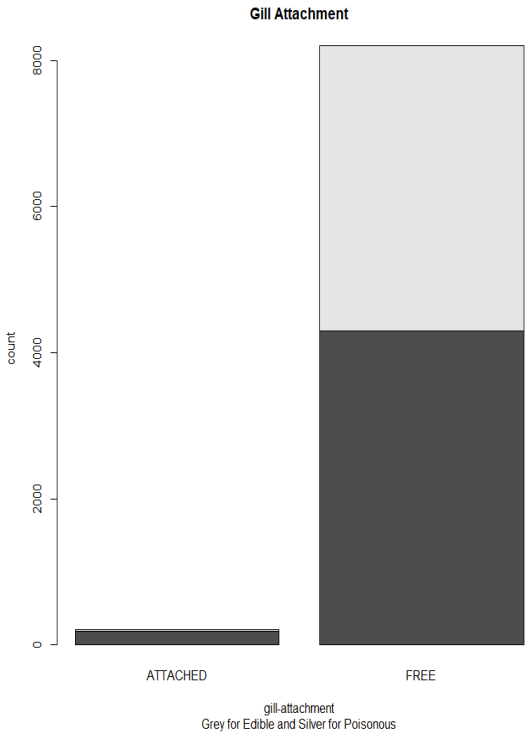
*Note: The plot of each attribute is attached in the ‘plots’ directory

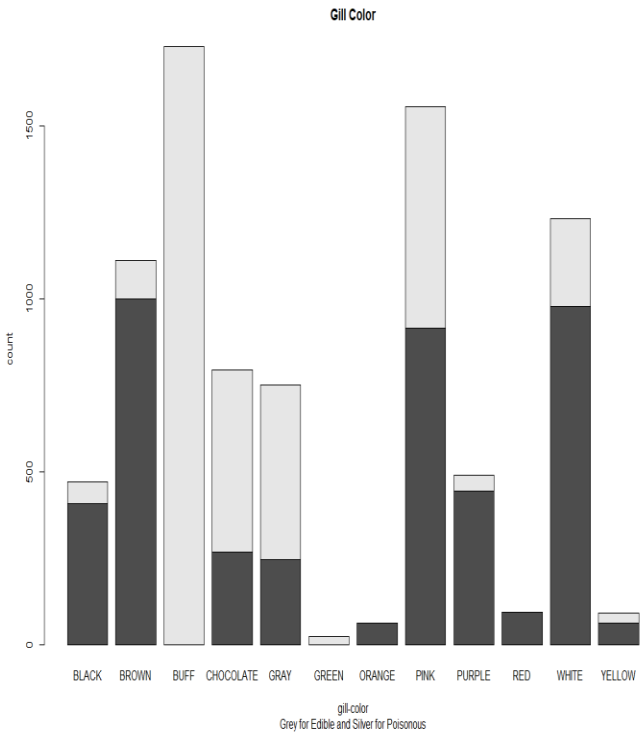

Bruises
count
0
1000
2000
3000
4000
5000
BRUISES
NO
bruises
Grey for Edible and Silver for Poisonous

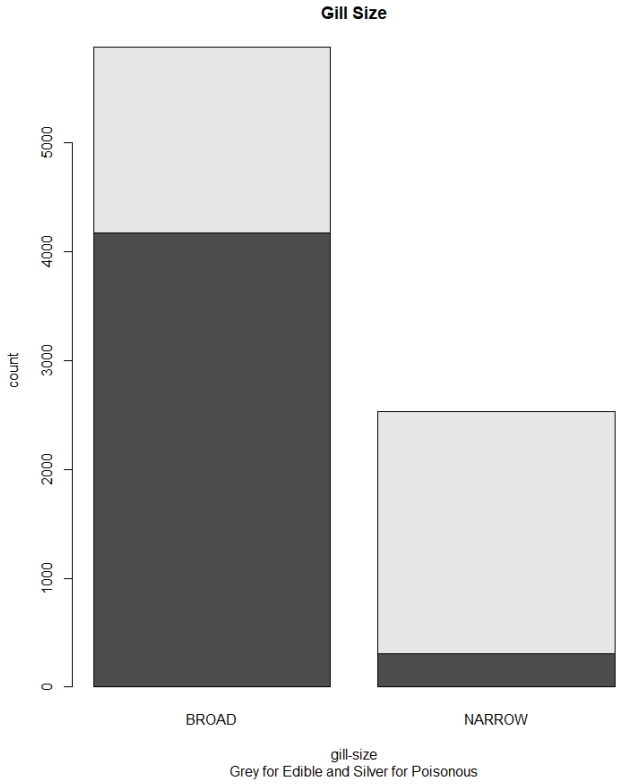

Cap Color
count
0
500
1000
1500
2000
BROWN
BUFF
CINNAMON
GRAY
GREEN
PINK
PURPLE
RED
WHITE
YELLOW
cap-color
Grey for Edible and Silver for Poisonous

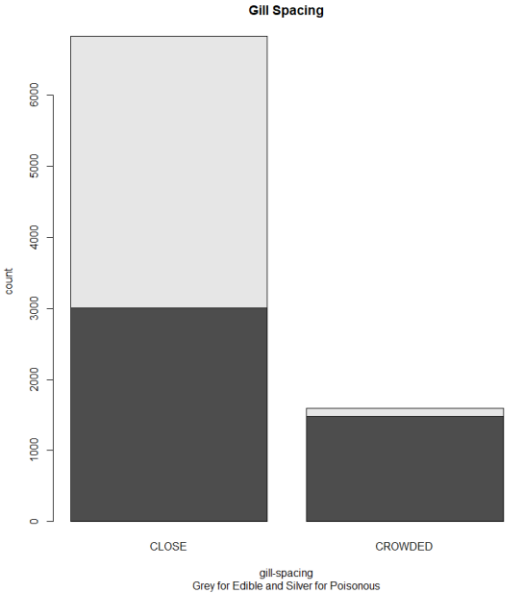

Cap Shape
count
0
500
1000
1500
2000
2500
3000
3500
BELL
CONICAL
CONVEX
FLAT
KNOBBED
SUNKEN
cap-shape
Grey for Edible and Silver for Poisonous

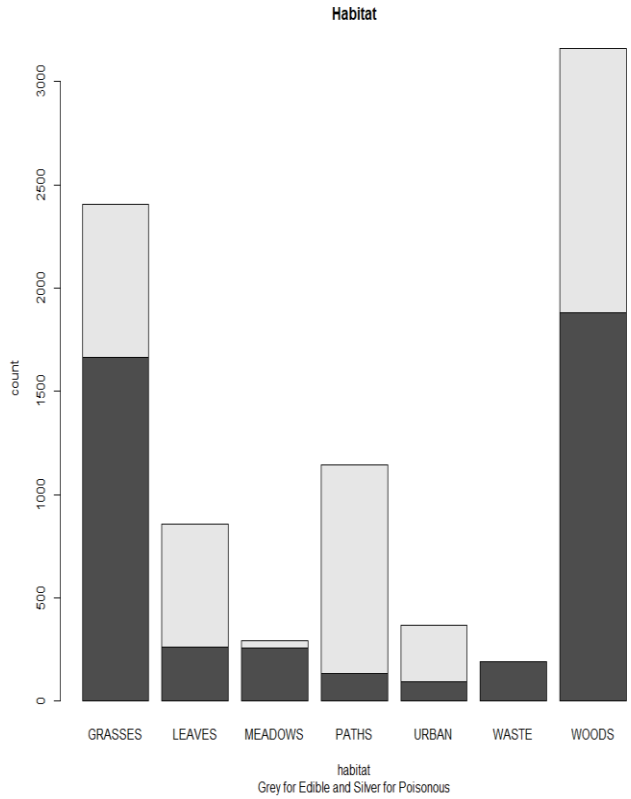

Cap Surface
count
0
500
1000
1500
2000
2500
3000
FIBROUS
GROOVES
SCALY
SMOOTH
cap-surface
Grey for Edible and Silver for Poisonous

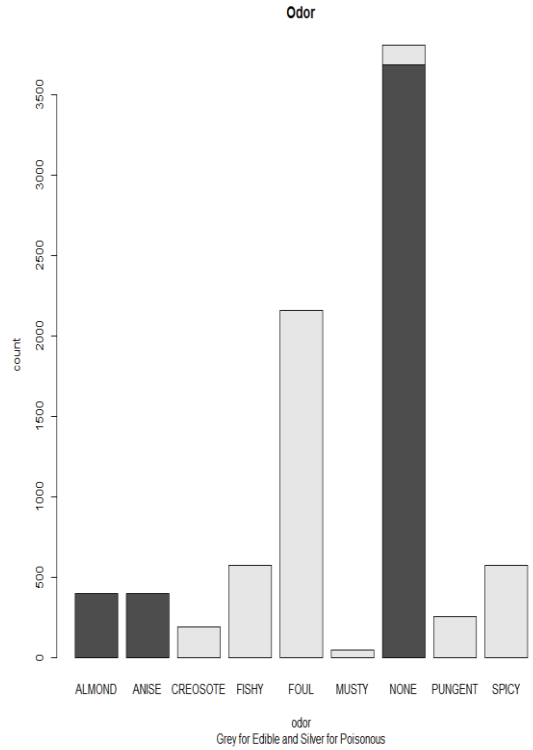

Gill Attachment
count
0
2000
4000
6000
8000
ATTACHED
FREE
gill-attachment
Grey for Edible and Silver for Poisonous

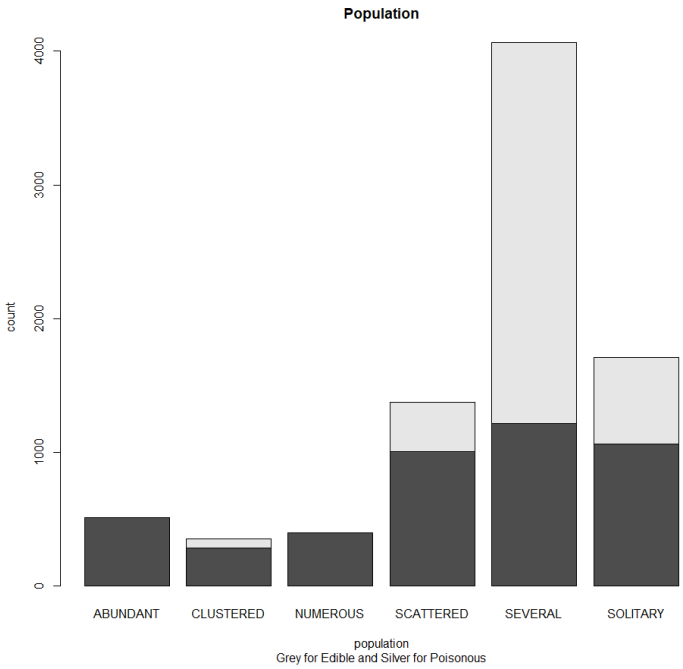

Gill Color
count
0
500
1000
1500
BLACK
BROWN
BUFF
CHOCOLATE
GRAY
GREEN
ORANGE
PINK
PURPLE
RED
WHITE
YELLOW
gill-color
Grey for Edible and Silver for Poisonous

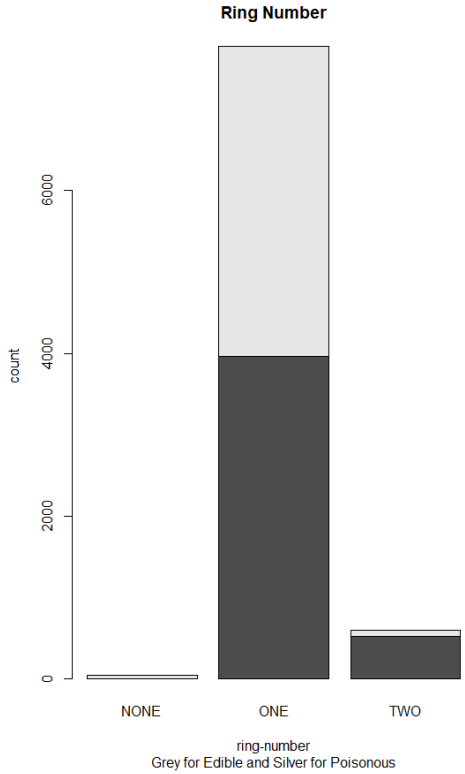

Gill Size
count
0
1000
2000
3000
4000
5000
BROAD
NARROW
gill-size
Grey for Edible and Silver for Poisonous

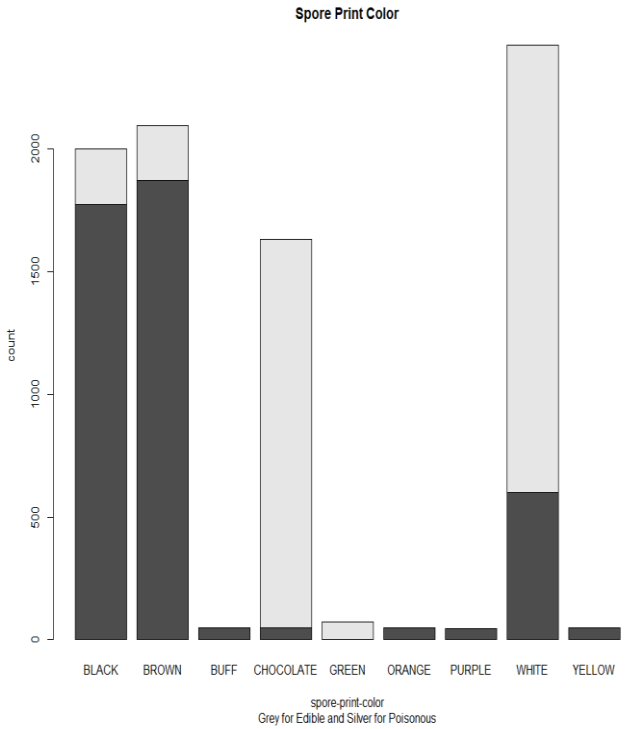

Gill Spacing
count
0
1000
2000
3000
4000
5000
6000
CLOSE
CROWDED
gill-spacing
Grey for Edible and Silver for Poisonous

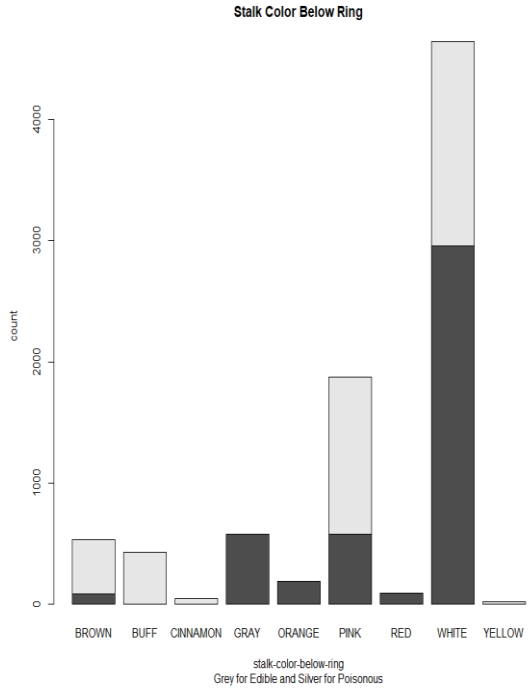

Habitat
count
0
500
1000
1500
2000
2500
3000
GRASSES
LEAVES
MEADOWS
PATHS
URBAN
WASTE
WOODS
habitat
Grey for Edible and Silver for Poisonous

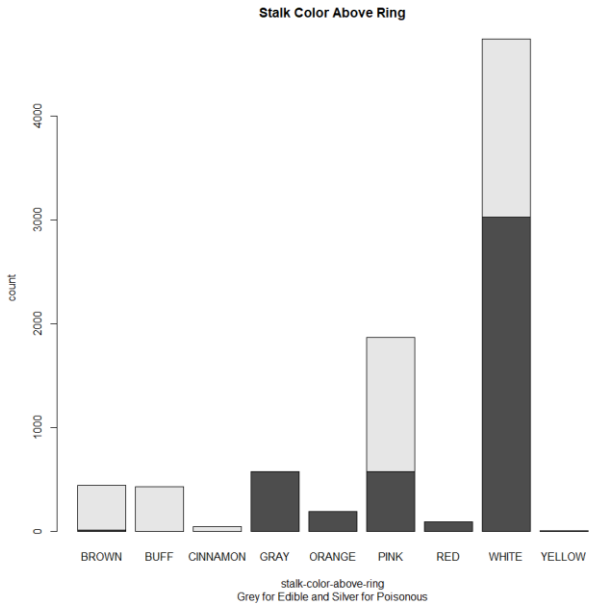

Odor
count
0
500
1000
1500
2000
2500
3000
3500
ALMOND
ANISE
CREOSOTE
FISHY
FOUL
MUSTY
NONE
PUNGENT
SPICY
odor
Grey for Edible and Silver for Poisonous

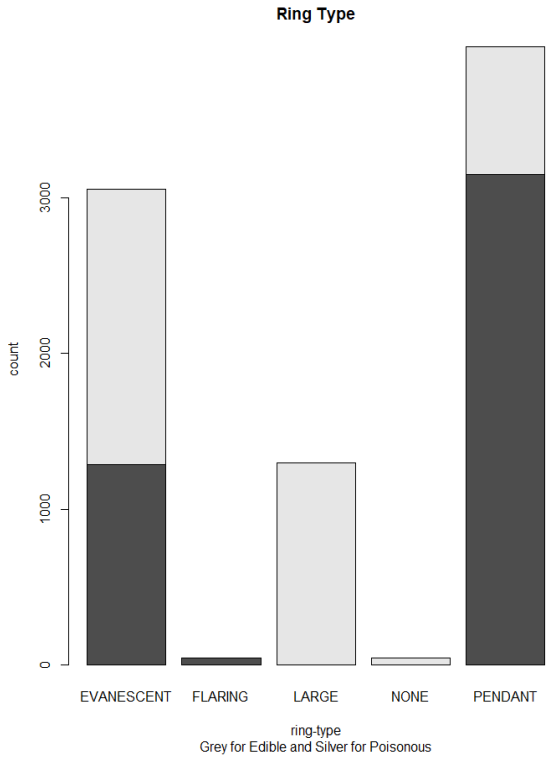

Population
count
0
1000
2000
3000
4000
ABUNDANT
CLUSTERED
NUMEROUS
SCATTERED
SEVERAL
SOLITARY
population
Grey for Edible and Silver for Poisonous

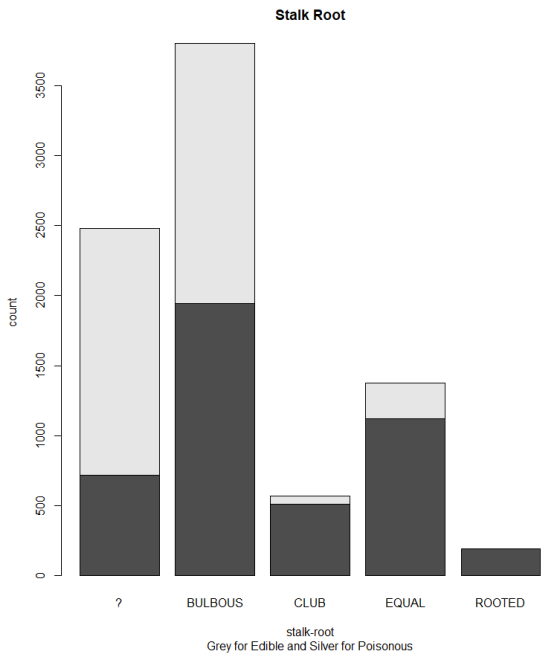

Ring Number
count
0
2000
4000
6000
NONE
ONE
TWO
ring-number
Grey for Edible and Silver for Poisonous

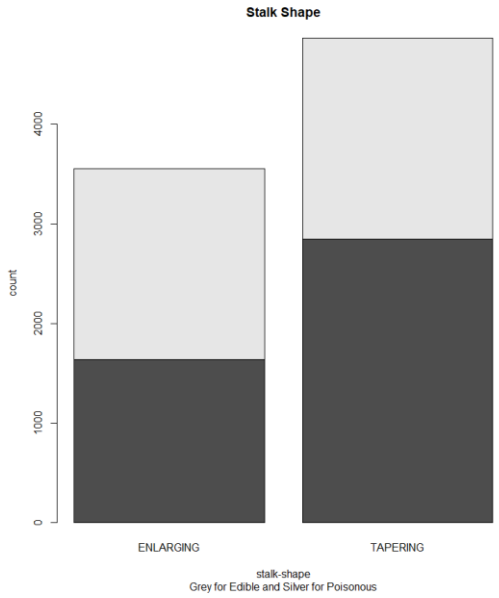

Spore Print Color
count
0
500
1000
1500
2000
BLACK
BROWN
BUFF
CHOCOLATE
GREEN
ORANGE
PURPLE
WHITE
YELLOW
spore-print-color
Grey for Edible and Silver for Poisonous

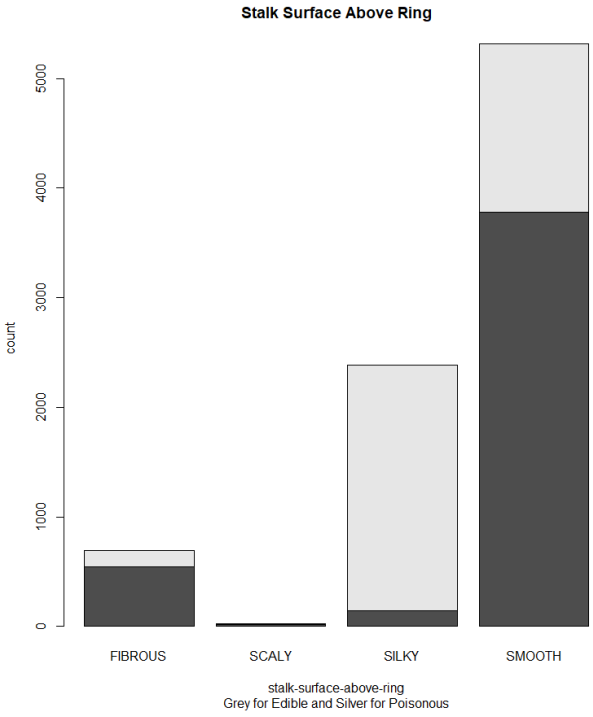

Stalk Color Below Ring
count
0
1000
2000
3000
4000
BROWN
BUFF
CINNAMON
GRAY
ORANGE
PINK
RED
WHITE
YELLOW
stalk-color-below-ring
Grey for Edible and Silver for Poisonous

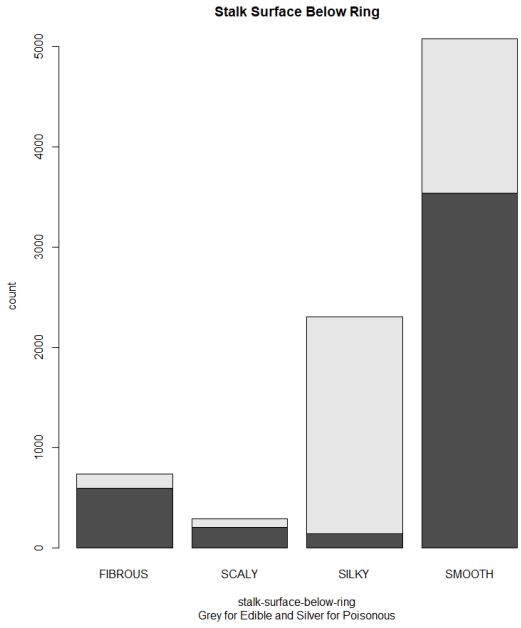

Stalk Color Above Ring
count
0
1000
2000
3000
4000
BROWN
BUFF
CINNAMON
GRAY
ORANGE
PINK
RED
WHITE
YELLOW
stalk-color-above-ring
Grey for Edible and Silver for Poisonous

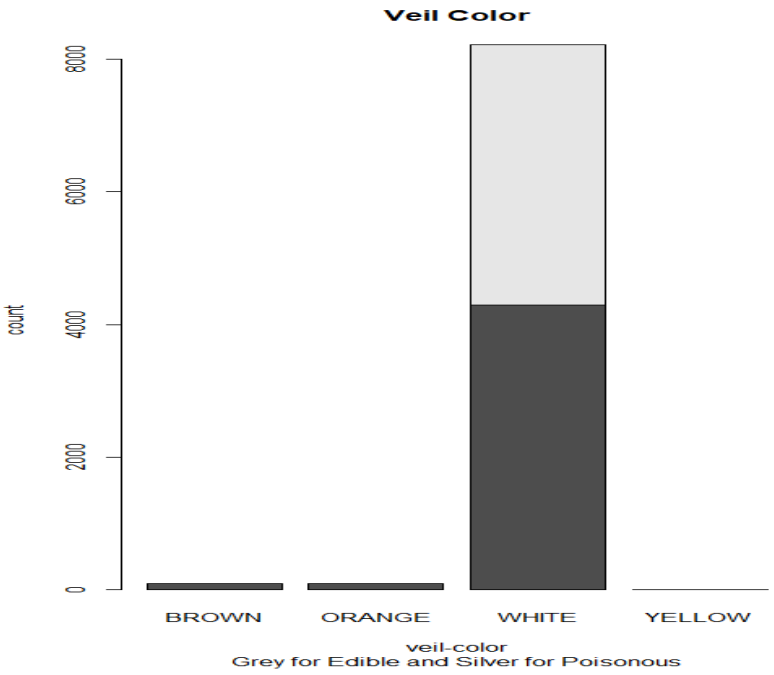

Ring Type
count
0
1000
2000
3000
EVANESCENT
FLARING
LARGE
NONE
PENDANT
ring-type
Grey for Edible and Silver for Poisonous

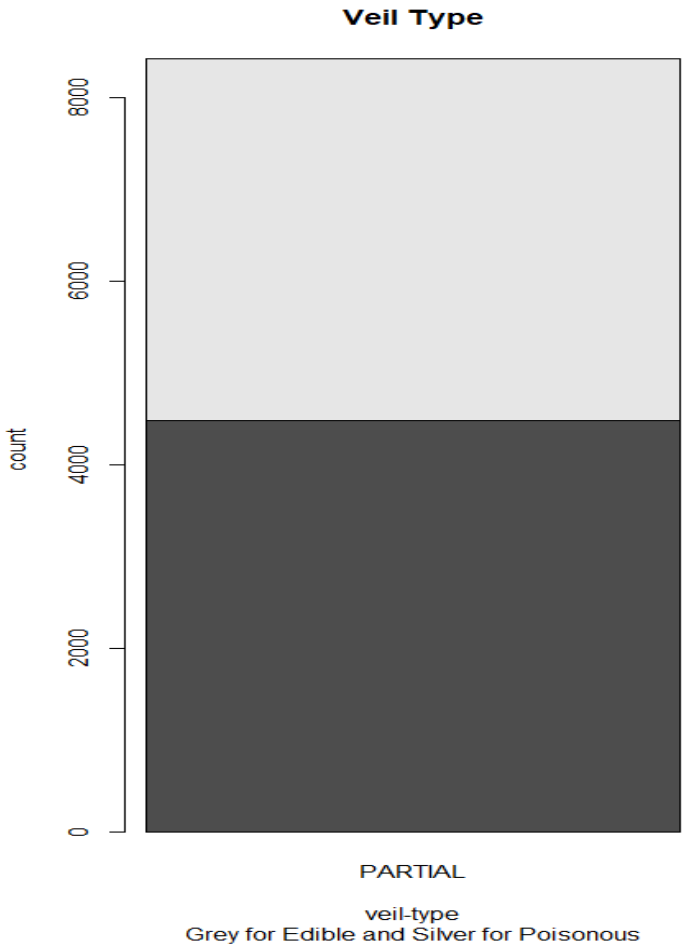

Stalk Root
count
0
500
1000
1500
2000
2500
3000
3500
?
BULBOUS
CLUB
EQUAL
ROOTED
stalk-root
Grey for Edible and Silver for Poisonous


Stalk Shape
count
0
1000
2000
3000
4000
ENLARGING
TAPERING
stalk-shape
Grey for Edible and Silver for Poisonous


Stalk Surface Above Ring
count
0
1000
2000
3000
4000
5000
FIBROUS
SCALY
SILKY
SMOOTH
stalk-surface-above-ring
Grey for Edible and Silver for Poisonous


Stalk Surface Below Ring
count
0
1000
2000
3000
4000
5000
FIBROUS
SCALY
SILKY
SMOOTH
stalk-surface-below-ring
Grey for Edible and Silver for Poisonous


Veil Color
count
0
2000
4000
6000
8000
BROWN
ORANGE
WHITE
YELLOW
veil-color
Grey for Edible and Silver for Poisonous


Veil Type
count
0
2000
4000
6000
8000
PARTIAL
veil-type
Grey for Edible and Silver for Poisonous

2.3 Dealing with Missing Values

In data mining, missing values can occur for a variety of reasons and can have a significant impact on the analysis of a dataset. Handling missing values is an important step in the data preprocessing stage because the presence of missing values can often lead to inaccurate or biased results.

There are several approaches to dealing with missing values, and the appropriate approach depends on the nature of the data, the cause of the missing values, and the goals of the analysis. In our study all missing values are in one attribute "stalk-root", and the mode value "BULBOUS" is almost equally distributed with regard to the class (as shown before in the plot), so we found that filling them with the mode value will be the best choice.

Code:

```
# We will convert all '?' to NA
marks <- data == '?'
is.na(data) <- marks

sum(is.na(data)) # prints 2480

# In the stalk-root attribute we will fill the missing values with the mode value which is 'BULBLOUS'
# Function to calculate the mode
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}
getmode(data$`stalk-root`) # prints "BULBOUS"

data$`stalk-root`[is.na(data$`stalk-root`)] <- getmode(data$`stalk-root`)

sum(is.na(data)) # prints 0
```

2.4 Columns Reduction

From the previous plots, we noticed that the *"veil-type"* column has one and only one value, which is *"partial"*, so, there is no need to keep it since it has nothing to do with classification.

Code:

```
# Since veil-type attribute has only one values it should be deleted
data <- subset(data, select = -`veil-type`)
```

2.5 Duplicate Data

We found that there near 300 records duplicated, so we removed them from our data so that we guarantee less biasing and more accurate results.

Code:

```
### Removing duplicate data
install.packages("tidyverse")
library(tidyverse)
sum(duplicated(data)) # Num of duplicates
data <- data[!duplicated(data), ]
```

I

2.6 Finding Relationships

After doing a preliminary analysis of the dataset, finding out the missing values, and fixing them, and because all attributes are categorical we want to calculate the correlation between attributes and class using the **Chi-Square** rule. Correlation in data mining refers to the strength and direction of the relationship between two variables and to see how closely attributes are related to the class.

According to the plots, we drew and after analyzing the dataset, the three attributes having an obvious strong correlation are in order: *"odor", "spore print color", and "gill color"*, so, we want to calculate Chi-square to see if that is true.

Code for Chi-square calculation:

```
###### finding relationship between attributes and class
library(MASS)
#show correlation between odor and class
tbl1<- table(data$class,data$odor)
chisq.test(tbl1)
#show correlation between gill-color and class
tbl3<- table(data$class,data$`gill-color`)
chisq.test(tbl3)
#show correlation between spore-print-color and class
tbl2<- table(data$class,data$`spore-print-color`)
chisq.test(tbl2)
```

Output:

```
> tbl1<- table(data$class,data$odor)
> chisq.test(tbl1)

        Pearson's Chi-squared test

data:  tbl1
X-squared = 7948.1, df = 8, p-value < 2.2e-16

> #show correlation between gill-color and class
> tbl3<- table(data$class,data$`gill-color`)
> chisq.test(tbl3)

        Pearson's Chi-squared test

data:  tbl3
X-squared = 3834.8, df = 11, p-value < 2.2e-16

> #show correlation between spore-print-color and class
> tbl2<- table(data$class,data$`spore-print-color`)
> chisq.test(tbl2)

        Pearson's Chi-squared test

data:  tbl2
X-squared = 4810.9, df = 8, p-value < 2.2e-16
```

After referring to the probability level table, we can accept the null theorem with a very low percentage, generally less than 0.01 (or less), meaning that, we are accepting that each one of these attributes with regard to the class is dependent by more than 99.99%.

---

3. Classification

The main objective of this analysis is to predict whether a given mushroom is edible or poisonous. To achieve that we split the entire dataset into two parts 70% for training and 30% for testing, then calculated the accuracy of our prediction, we used the Naive Bayes classifier with Laplace set to 1 to achieve the best accuracy.

Code:

```r
################################################ Classification using Naive Bayes
data$class <- as.factor(data$class)

install.packages('caret')
library(caret)
trainIndex <- createDataPartition(data$class, p = 0.7, list = FALSE)
train <- data[trainIndex,]
test <- data[-trainIndex,]

str(train)
table(train$class) # to make sure the data is split in a convenient way

install.packages("e1071")
library(e1071)

nv <- naiveBayes(class~., data, laplace = 1)

p <- predict(nv, test)
table(p, test$class)
prediction <- table(p, test$class)
##to calculate the accuracy for test part
accuracy = (sum(diag(prediction)) / sum(prediction)) * 100

accuracy
```

Output:

```
> str(train)
'data.frame':   5891 obs. of  22 variables:
 $ class                   : Factor w/ 2 levels "EDIBLE","POISONOUS": 1 1 1 1 1 1 1 1 1 1 ...
 $ cap-shape               : chr  "CONVEX" "CONVEX" "CONVEX" "CONVEX" ...
 $ cap-surface             : chr  "SMOOTH" "SMOOTH" "SMOOTH" "SMOOTH" ...
 $ cap-color               : chr  "WHITE" "WHITE" "WHITE" "WHITE" ...
 $ bruises                 : chr  "BRUISES" "BRUISES" "BRUISES" "BRUISES" ...
 $ odor                    : chr  "ALMOND" "ALMOND" "ALMOND" "ANISE" ...
 $ gill-attachment         : chr  "FREE" "FREE" "FREE" "FREE" ...
 $ gill-spacing            : chr  "CROWDED" "CROWDED" "CROWDED" "CROWDED" ...
 $ gill-size               : chr  "NARROW" "NARROW" "NARROW" "NARROW" ...
 $ gill-color              : chr  "PINK" "BROWN" "BROWN" "WHITE" ...
 $ stalk-shape             : chr  "TAPERING" "TAPERING" "TAPERING" "TAPERING" ...
 $ stalk-root              : chr  "BULBOUS" "BULBOUS" "BULBOUS" "BULBOUS" ...
 $ stalk-surface-above-ring: chr  "SMOOTH" "SMOOTH" "SMOOTH" "SMOOTH" ...
 $ stalk-surface-below-ring: chr  "SMOOTH" "SMOOTH" "SMOOTH" "SMOOTH" ...
 $ stalk-color-above-ring  : chr  "WHITE" "WHITE" "WHITE" "WHITE" ...
 $ stalk-color-below-ring  : chr  "WHITE" "WHITE" "WHITE" "WHITE" ...
 $ veil-color              : chr  "WHITE" "WHITE" "WHITE" "WHITE" ...
 $ ring-number             : chr  "ONE" "ONE" "ONE" "ONE" ...
 $ ring-type               : chr  "PENDANT" "PENDANT" "PENDANT" "PENDANT" ...
 $ spore-print-color       : chr  "PURPLE" "PURPLE" "BROWN" "PURPLE" ...
 $ population              : chr  "SEVERAL" "SEVERAL" "SEVERAL" "SEVERAL" ...
 $ habitat                 : chr  "WOODS" "WOODS" "WOODS" "WOODS" ...
> table(train$class) # to make sure the data is split in a convenient way

   EDIBLE POISONOUS
     3141      2750
>
> install.packages("e1071")
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into 'C:/Users/user/AppData/Local/R/win-library/4.2'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.2/e1071_1.7-12.zip'
Content type 'application/zip' length 663514 bytes (647 KB)
downloaded 647 KB

package 'e1071' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\user\AppData\Local\Temp\Rtmpe4ZMdz\downloaded_packages
```

```
package 'e1071' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\user\AppData\Local\Temp\Rtmpe4ZMdz\downloaded_packages
> library(e1071)
>
> nv <- naiveBayes(class~., data, laplace = 1)
>
> p <- predict(nv, test)
> table(p, test$class)

p           EDIBLE POISONOUS
  EDIBLE      1338       105
  POISONOUS      8      1073
> prediction <- table(p, test$class)
> ##to calculate the accuracy for test part
> accuracy = (sum(diag(prediction)) / sum(prediction)) * 100
>
> accuracy
[1] 95.52298
```

4. Conclusion

In conclusion, our classifier had high accuracy, generally, 95% - 96%, which is good for a sensitive case to determine the edibility of a mushroom even though the dataset wasn't enormous but it seems that it was correctly collected, prepared, and classified, resulting in satisfying results. Also, the Chi-Square rule shows that there is a strong dependency between the *"odor", "spore print color", "gill color"* attributes and the class, this information could be useful if published to the public to reduce the rates of poisoning cases between people every year.

5. References

- UCI Machine Learning Repository
- Data Mining Concepts and Techniques by: Jiawei Han, Micheline Kamber, Jian Pei