Mushroom Analysis

Soldatov Vadim

2024-12-05

Introduction

The Mushroom'dataset consists of 8124 observations with 23 columns. The details of the names of columns i.e. attributes in the dataset after loading the dataset are found.

Loading libraries

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.
0.0 ──
```

```
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
```

```
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
```

```
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
```

```
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.1
```

```
## ✓ purrr     1.0.2
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflict
s() ──
```

```
## ✖ dplyr::filter() masks stats::filter()
```

```
## ✖ dplyr::lag()    masks stats::lag()
```

```
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors
```

```
library(dslabs)
library(dplyr)
library(ggplot2)
library(dplyr)
library(vcd)
```

```
## Загрузка требуемого пакета: grid
```

```
library(psych)
```

```
##
```

```
## Присоединяю пакет: 'psych'
```

```
##
```

```
## Следующие объекты скрыты от 'package:ggplot2':
```

```
##
```

```
##     %+%, alpha
```

```
library(gridExtra)
```

```
##
## Присоединяю пакет: 'gridExtra'
##
## Следующий объект скрыт от 'package:dplyr':
##
##     combine
```

```
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```
library(MASS)
```

```
##
## Присоединяю пакет: 'MASS'
##
## Следующий объект скрыт от 'package:dplyr':
##
##     select
```

Mushroom dataset

```
mush <- readr::read_csv("https://raw.githubusercontent.com/Vadim77-AI/Mushroom-Analysis/refs/heads/main/mushrooms.csv")
```

```
## Rows: 8124 Columns: 23
## ── Column specification ────────────────────────────────────────────────────
## Delimiter: ","
## chr (22): class, cap-shape, cap-surface, cap-color, odor, gill-attachment, g...
## lgl  (1): bruises
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(mush)
```

```
## # A tibble: 6 × 23
##   class `cap-shape` `cap-surface` `cap-color` bruises odor  `gill-attachment`
##   <chr> <chr>       <chr>         <chr>       <lgl>   <chr> <chr>
## 1 p     x           s             n           TRUE    p     f
## 2 e     x           s             y           TRUE    a     f
```

```
## 3 e     b           s           w           TRUE    l     f

## 4 p     x           y           w           TRUE    p     f

## 5 e     x           s           g           FALSE   n     f

## 6 e     x           y           y           TRUE    a     f

## # ℹ 16 more variables: `gill-spacing` <chr>, `gill-size` <chr>,
## #   `gill-color` <chr>, `stalk-shape` <chr>, `stalk-root` <chr>,
## #   `stalk-surface-above-ring` <chr>, `stalk-surface-below-ring` <chr>,
## #   `stalk-color-above-ring` <chr>, `stalk-color-below-ring` <chr>,
## #   `veil-type` <chr>, `veil-color` <chr>, `ring-number` <chr>,
## #   `ring-type` <chr>, `spore-print-color` <chr>, population <chr>,
## #   habitat <chr>
```

Data Exploration

About this file:

Attribute Information: (classes: edible=e, poisonous=p)

cap-shape: bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s

cap-surface: fibrous=f,grooves=g,scaly=y,smooth=s

cap-color: brown=n,buff=b,cinnamon=c,gray=g,green=r,pink=p,purple=u,red=e,white=w, ellow=y

bruises: bruises=t,no=f

odor: almond=a,anise=l,creosote=c,fishy=y,foul=f,musty=m,none=n,pungent=p,spicy=s

gill-attachment: attached=a,descending=d,free=f,notched=n

gill-spacing: close=c,crowded=w,distant=d

gill-size: broad=b,narrow=n

gill-color: black=k,brown=n,buff=b,chocolate=h,gray=g,
green=r,orange=o,pink=p,purple=u,red=e,white=w,yellow=y

stalk-shape: enlarging=e,tapering=t

stalk-root: bulbous=b,club=c,cup=u,equal=e,rhizomorphs=z,rooted=r,missing=?

stalk-surface-above-ring: fibrous=f,scaly=y,silky=k,smooth=s

stalk-surface-below-ring: fibrous=f,scaly=y,silky=k,smooth=s

stalk-color-above-ring:
brown=n,buff=b,cinnamon=c,gray=g,orange=o,pink=p,red=e,white=w,yellow=y

stalk-color-below-ring:
brown=n,buff=b,cinnamon=c,gray=g,orange=o,pink=p,red=e,white=w,yellow=y

veil-type: partial=p,universal=u

veil-color: brown=n,orange=o,white=w,yellow=y

ring-number: none=n,one=o,two=t

ring-type: cobwebby=c,evanescent=e,flaring=f,large=l,none=n,pendant=p,sheathing=s,zone=z

spore-print-color:
black=k,brown=n,buff=b,chocolate=h,green=r,orange=o,purple=u,white=w,yellow=y

population: abundant=a,clustered=c,numerous=n,scattered=s,several=v,solitary=y

habitat: grasses=g,leaves=l,meadows=m,paths=p,urban=u,waste=w,woods=d

```
class(mush)
```

```
## [1] "spec_tbl_df" "tbl_df"      "tbl"         "data.frame"
```

```
head(mush)
```

```
## # A tibble: 6 × 23
##   class `cap-shape` `cap-surface` `cap-color` bruises odor  `gill-attachme
nt`
##   <chr> <chr>       <chr>         <chr>       <lgl>   <chr> <chr>

## 1 p     x           s             n           TRUE    p     f

## 2 e     x           s             y           TRUE    a     f

## 3 e     b           s             w           TRUE    l     f

## 4 p     x           y             w           TRUE    p     f

## 5 e     x           s             g           FALSE   n     f

## 6 e     x           y             y           TRUE    a     f

## # ℹ 16 more variables: `gill-spacing` <chr>, `gill-size` <chr>,
## #   `gill-color` <chr>, `stalk-shape` <chr>, `stalk-root` <chr>,
## #   `stalk-surface-above-ring` <chr>, `stalk-surface-below-ring` <chr>,
## #   `stalk-color-above-ring` <chr>, `stalk-color-below-ring` <chr>,
## #   `veil-type` <chr>, `veil-color` <chr>, `ring-number` <chr>,
## #   `ring-type` <chr>, `spore-print-color` <chr>, population <chr>,
## #   habitat <chr>
```

```
glimpse(mush)
```

```
## Rows: 8,124
## Columns: 23
## $ class                     <chr> "p", "e", "e", "p", "e", "e", "e", "e",
 "p"…
## $ `cap-shape`               <chr> "x", "x", "b", "x", "x", "x", "b", "b",
 "x"…
## $ `cap-surface`             <chr> "s", "s", "s", "y", "s", "y", "s", "y",
 "y"…
## $ `cap-color`               <chr> "n", "y", "w", "w", "g", "y", "w", "w",
 "w"…
```

```
## $ bruises                    <lgl> TRUE, TRUE, TRUE, TRUE, FALSE, TRUE, TR
UE, …
## $ odor                       <chr> "p", "a", "l", "p", "n", "a", "a", "l",
 "p"…
## $ `gill-attachment`          <chr> "f", "f", "f", "f", "f", "f", "f", "f",
 "f"…
## $ `gill-spacing`             <chr> "c", "c", "c", "c", "w", "c", "c", "c",
 "c"…
## $ `gill-size`                <chr> "n", "b", "b", "n", "b", "b", "b", "b",
 "n"…
## $ `gill-color`               <chr> "k", "k", "n", "n", "k", "n", "g", "n",
 "p"…
## $ `stalk-shape`              <chr> "e", "e", "e", "e", "t", "e", "e", "e",
 "e"…
## $ `stalk-root`               <chr> "e", "c", "c", "e", "e", "c", "c", "c",
 "e"…
## $ `stalk-surface-above-ring` <chr> "s", "s", "s", "s", "s", "s", "s", "s",
 "s"…
## $ `stalk-surface-below-ring` <chr> "s", "s", "s", "s", "s", "s", "s", "s",
 "s"…
## $ `stalk-color-above-ring`   <chr> "w", "w", "w", "w", "w", "w", "w", "w",
 "w"…
## $ `stalk-color-below-ring`   <chr> "w", "w", "w", "w", "w", "w", "w", "w",
 "w"…
## $ `veil-type`                <chr> "p", "p", "p", "p", "p", "p", "p", "p",
 "p"…
## $ `veil-color`               <chr> "w", "w", "w", "w", "w", "w", "w", "w",
 "w"…
## $ `ring-number`              <chr> "o", "o", "o", "o", "o", "o", "o", "o",
 "o"…
## $ `ring-type`                <chr> "p", "p", "p", "p", "e", "p", "p", "p",
 "p"…
## $ `spore-print-color`        <chr> "k", "n", "n", "k", "n", "k", "k", "n",
 "k"…
## $ population                 <chr> "s", "n", "n", "s", "a", "n", "n", "s",
 "v"…
## $ habitat                    <chr> "u", "g", "m", "u", "g", "g", "m", "m",
 "g"…
```

```
dim(mush)
```

```
## [1] 8124   23
```

```
summary(mush)
```

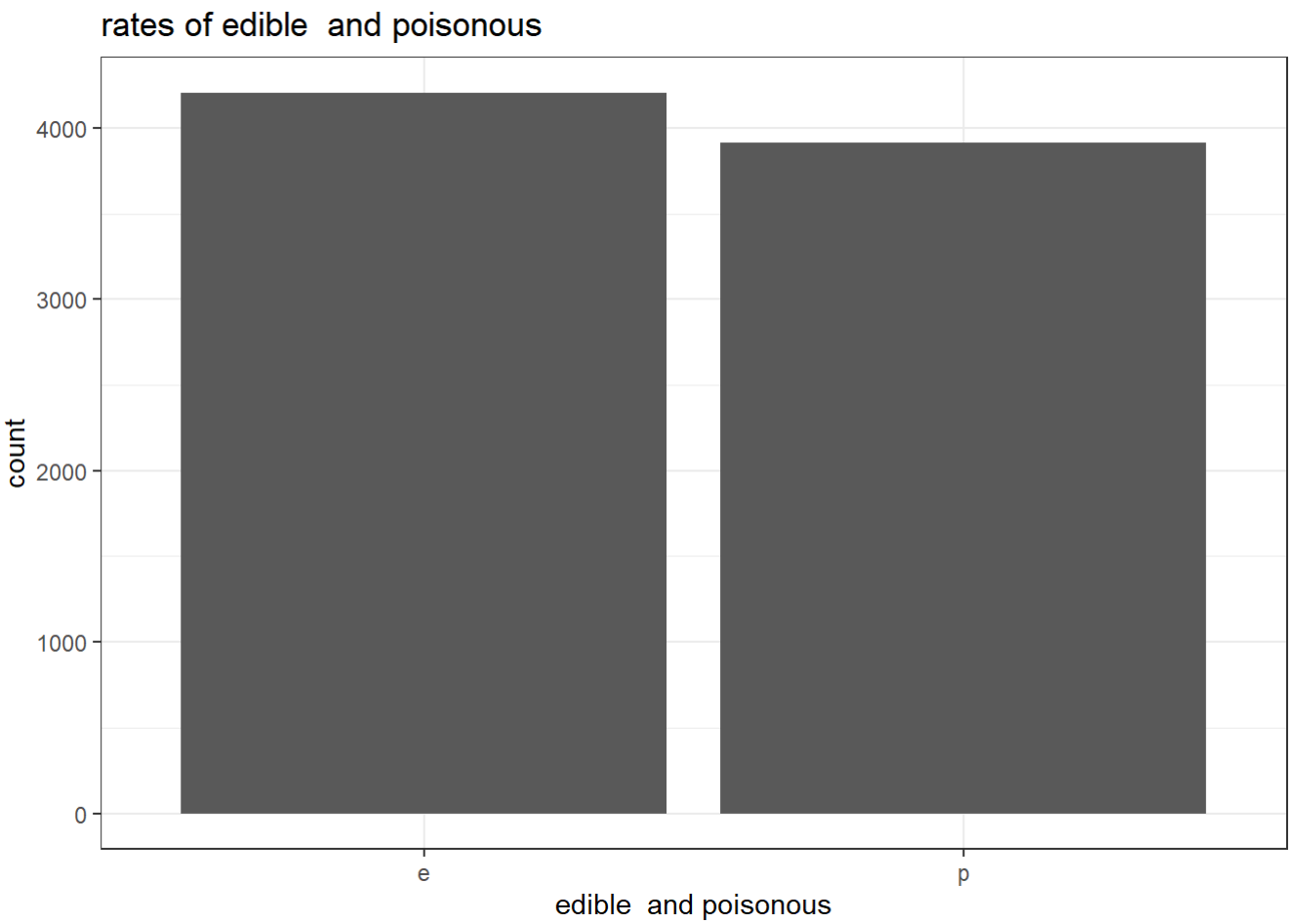
```
##     class            cap-shape         cap-surface         cap-color

##  Length:8124        Length:8124        Length:8124        Length:8124

##  Class :character   Class :character   Class :character   Class :character
```

```
##  Mode   :character   Mode   :character   Mode   :character   Mode   :character

##    bruises            odor           gill-attachment    gill-spacing
##  Mode :logical   Length:8124        Length:8124        Length:8124
##  FALSE:4748      Class :character   Class :character   Class :character
##  TRUE :3376      Mode  :character   Mode  :character   Mode  :character
##    gill-size          gill-color        stalk-shape          stalk-root

##  Length:8124        Length:8124        Length:8124        Length:8124

##  Class :character   Class :character   Class :character   Class :character

##  Mode  :character   Mode  :character   Mode  :character   Mode  :character

##  stalk-surface-above-ring stalk-surface-below-ring stalk-color-above-ring
##  Length:8124              Length:8124              Length:8124
##  Class :character         Class :character         Class :character
##  Mode  :character         Mode  :character         Mode  :character
##  stalk-color-below-ring  veil-type          veil-color
##  Length:8124            Length:8124        Length:8124
##  Class :character       Class :character   Class :character
##  Mode  :character       Mode  :character   Mode  :character
##  ring-number         ring-type         spore-print-color   population

##  Length:8124        Length:8124        Length:8124        Length:8124

##  Class :character   Class :character   Class :character   Class :character

##  Mode  :character   Mode  :character   Mode  :character   Mode  :character

##    habitat
##  Length:8124
##  Class :character
##  Mode  :character
```

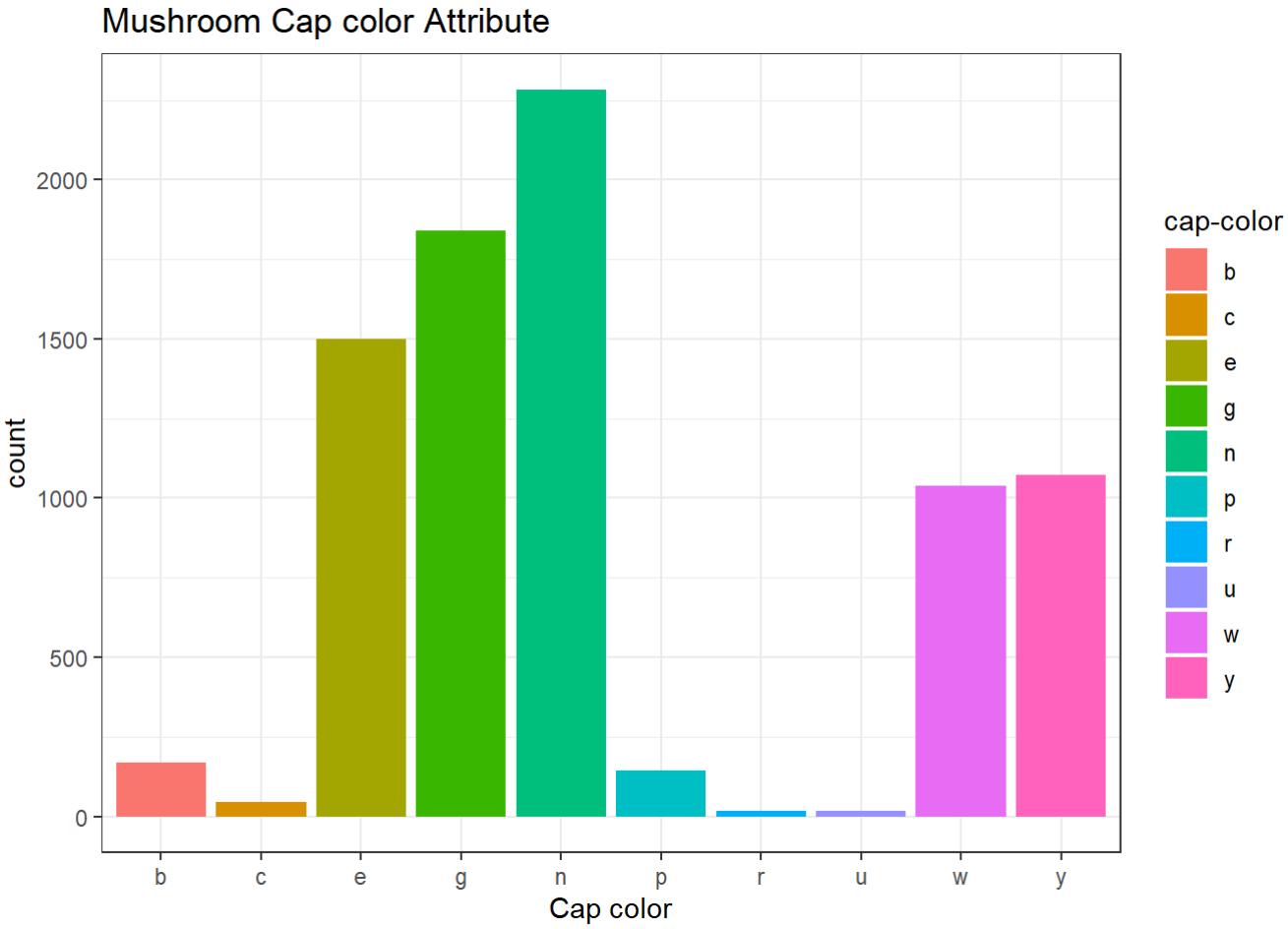
This plotted is histograms of each category and split them into two graphs according to their edibility. The objective to do is to find the attributes which are exclusive only in either class.

```
ggplot(mush, aes(x = class)) +
  theme_bw()+
  geom_bar()+
  labs(x = "edible  and poisonous",
       y = "count",
```

```
        title = "rates of edible  and poisonous")
```




```
ggplot(mush, aes(x = `cap-color`, fill = `cap-color`)) +
  theme_bw()+
  geom_bar()+
  labs(x = "Cap color",
       y = "count",
       title = "Mushroom Cap color Attribute")
```

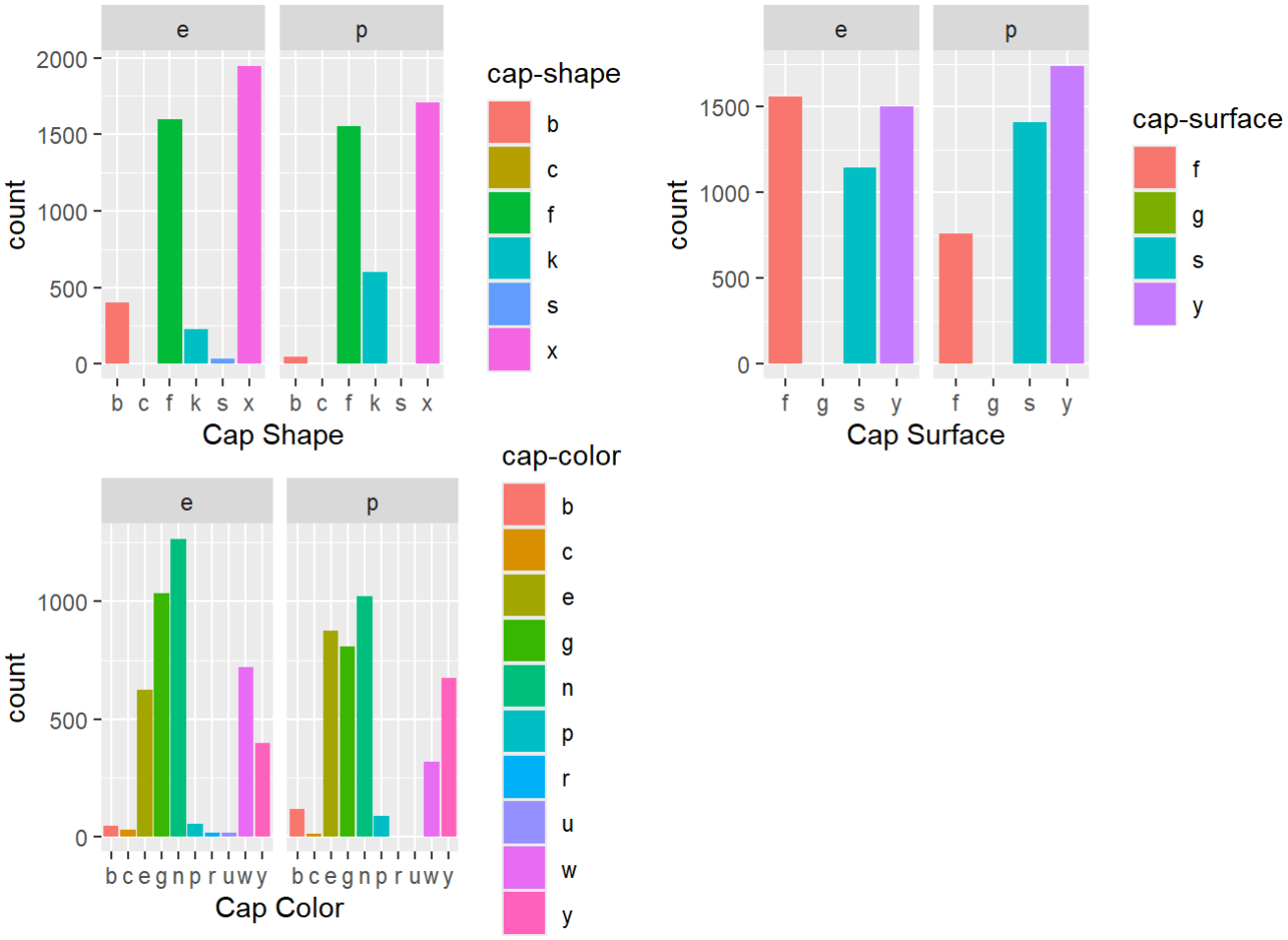


Summary of the section: Mushrooms are most likely to be brown, gray, red, white or yellow

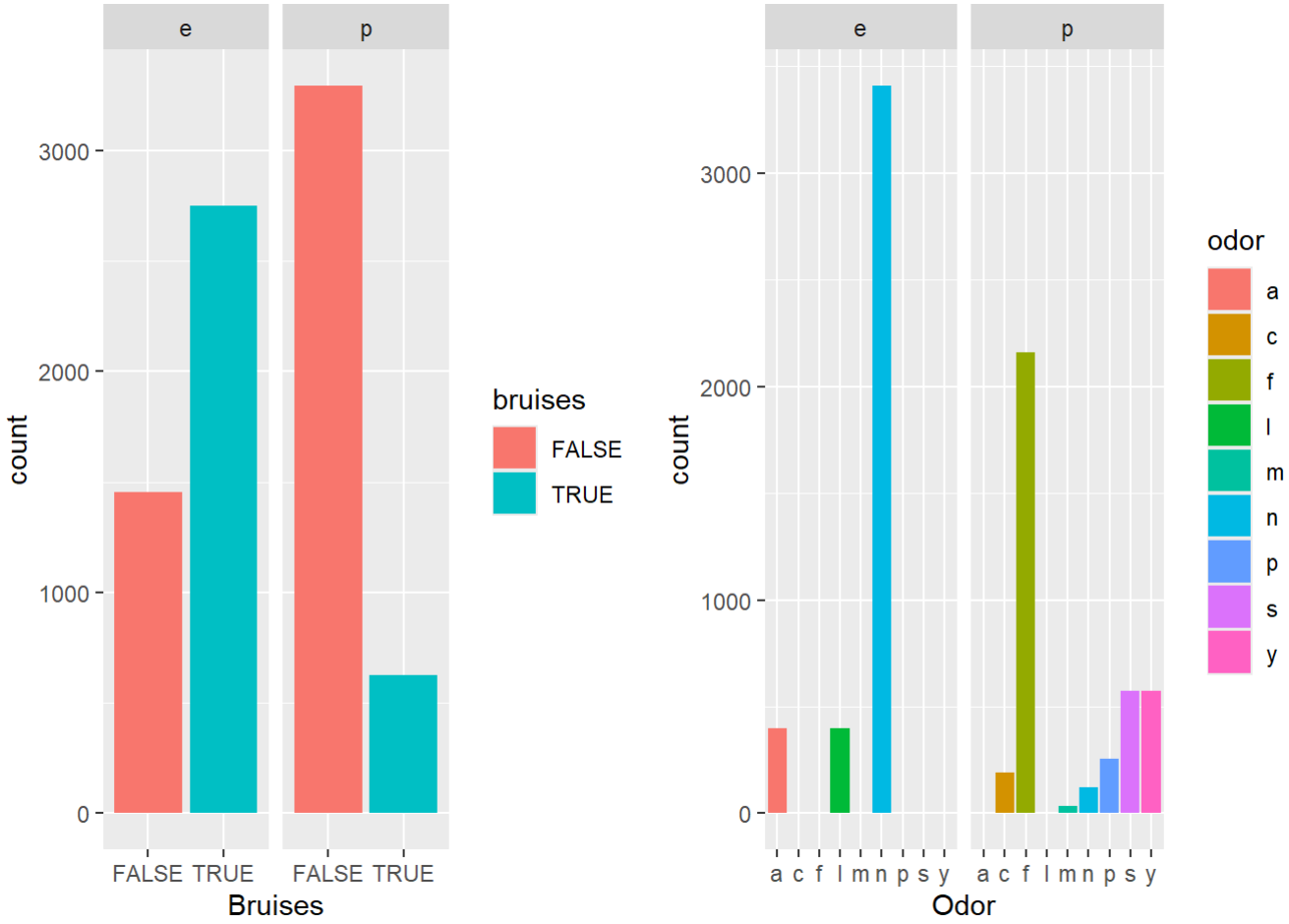
Below we can see the histograms of each category and split into two graphs according to their edibility. The objective is to find the attributes which are exclusive only in either class. The more exclusiveness hints towards a stronger correlation between the attribute and the edibility of the mushroom. The first three attributes - cap shape, cap surface and cap color are plotted below.

```
mush1 <- ggplot(aes(x = `cap-shape`, fill = `cap-shape`), data = mush) +
  geom_bar(stat="count") +
  facet_wrap(~class) +
  xlab("Cap Shape")

mush2 <- ggplot(aes(x = `cap-surface`, fill = `cap-surface`), data = mush) +
  geom_bar(stat="count") +
  facet_wrap(~class) +
  xlab("Cap Surface")

mush3 <- ggplot(aes(x = `cap-color`, fill = `cap-color`), data = mush) +
  geom_bar(stat="count") +
  facet_wrap(~class) +
  xlab("Cap Color")
```

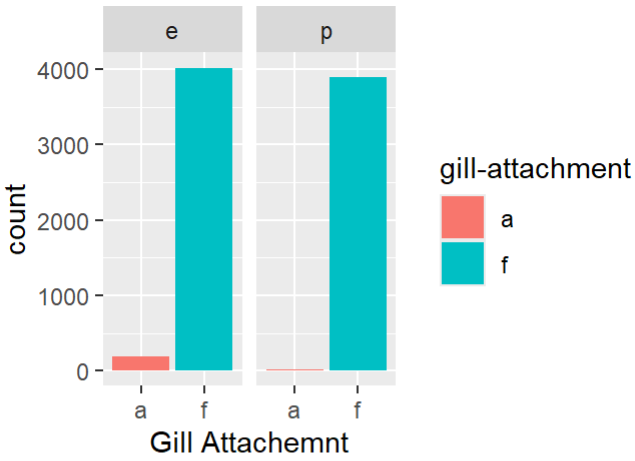
```
grid.arrange(mush1, mush2, mush3, ncol = 2)
```

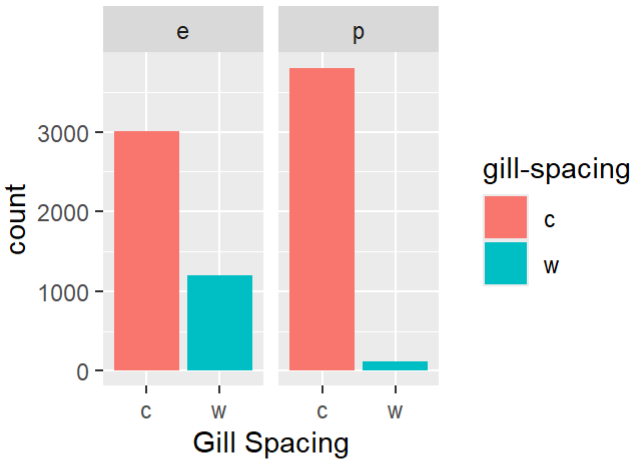


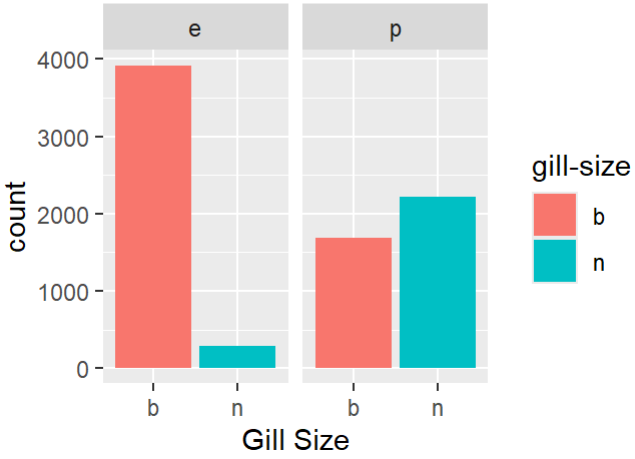

```
mush4 <- ggplot(aes(x = bruises, fill = bruises), data = mush) +
  geom_bar(stat = "count") +
  facet_wrap(~class) +
  xlab("Bruises")

mush5 <- ggplot(aes(x = odor, fill = odor), data = mush) +
  geom_bar(stat = "count") +
  facet_wrap(~class) +
  xlab("Odor")

grid.arrange(mush4, mush5, ncol = 2)
```



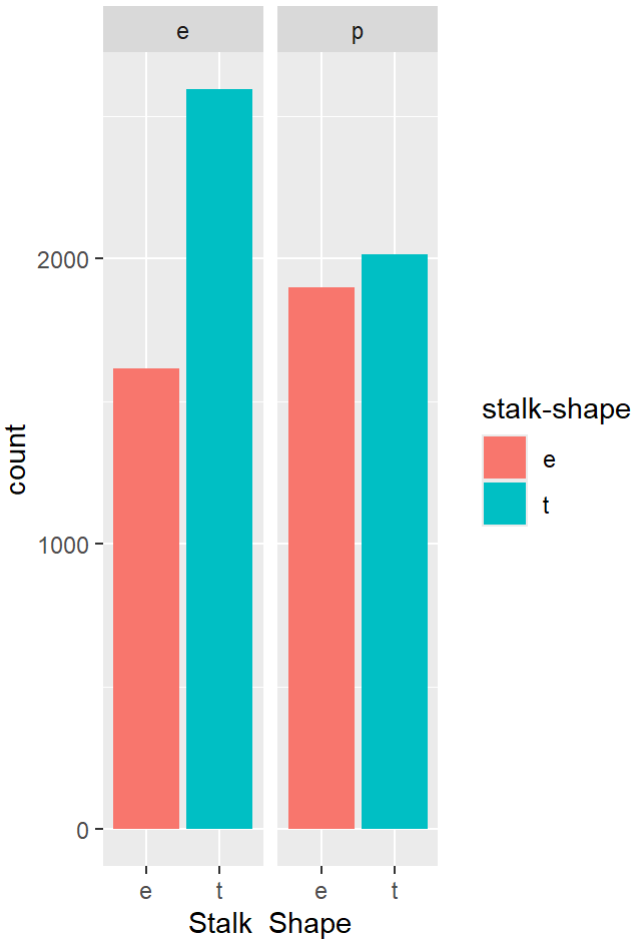

```
mush6 <- ggplot(aes(x = `gill-attachment`, fill = `gill-attachment`), data =
mush) +
  geom_bar(stat = "count") +
  facet_wrap(~class) +
  xlab("Gill Attachemnt")


mush7 <- ggplot(aes(x = `gill-spacing`, fill = `gill-spacing`), data = mush)
+
  geom_bar(stat = "count") +
  facet_wrap(~class) +
  xlab("Gill Spacing")


mush8 <- ggplot(aes(x = `gill-size`, fill = `gill-size`), data = mush) +
  geom_bar(stat = "count") +
  facet_wrap(~class) +
  xlab("Gill Size")


grid.arrange(mush6, mush7, mush8, ncol = 2)
```

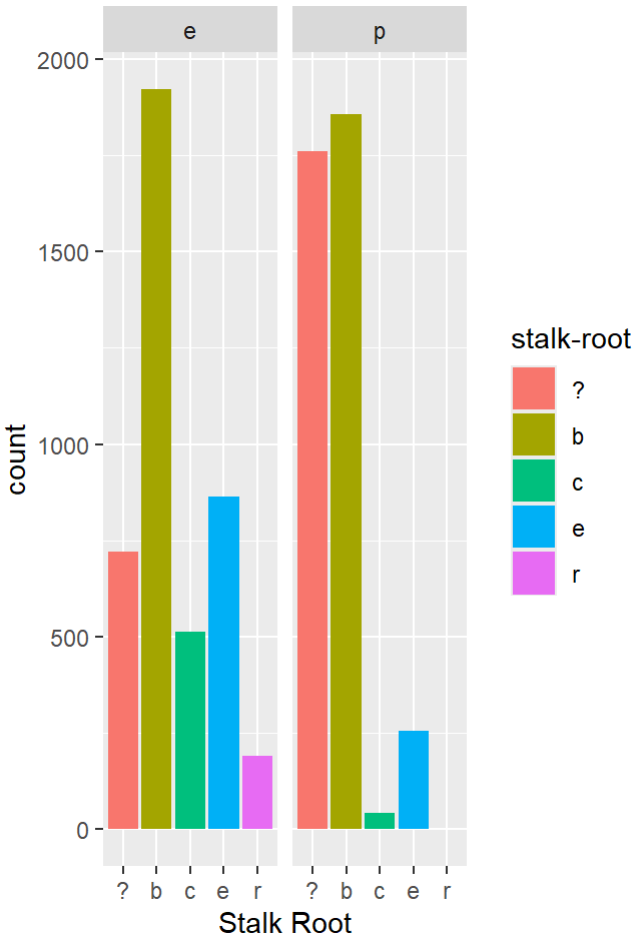




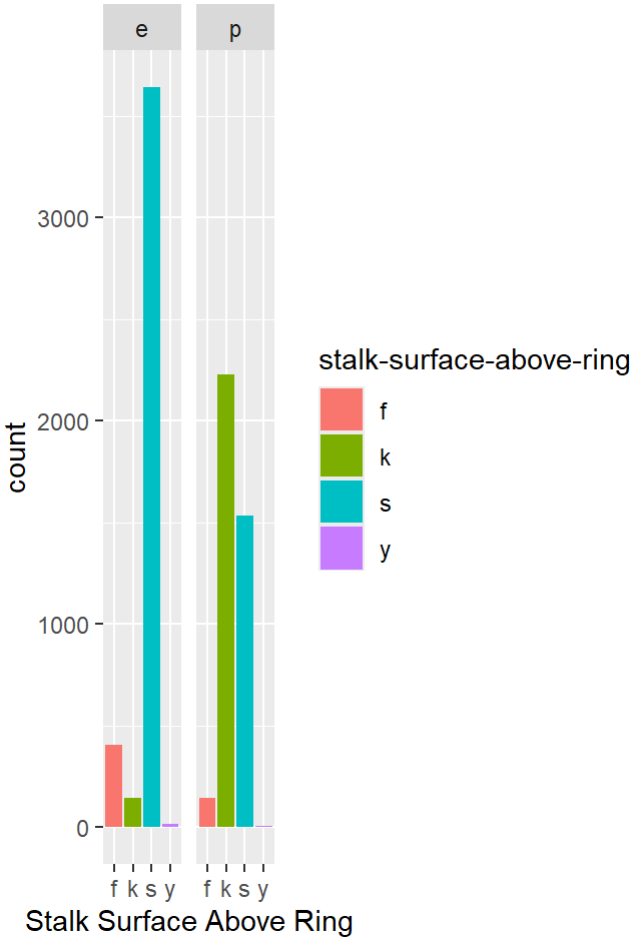



```
mush9 <- ggplot(aes(x = `stalk-shape`, fill = `stalk-shape`), data = mush) +
  geom_bar(stat = "count") +
  facet_wrap(~class) +
  xlab("Stalk  Shape")

mush10 <- ggplot(aes(x = `stalk-root`, fill = `stalk-root`), data = mush) +
  geom_bar(stat = "count") +
  facet_wrap(~class) +
  xlab("Stalk Root")

grid.arrange(mush9, mush10, ncol = 2)
```

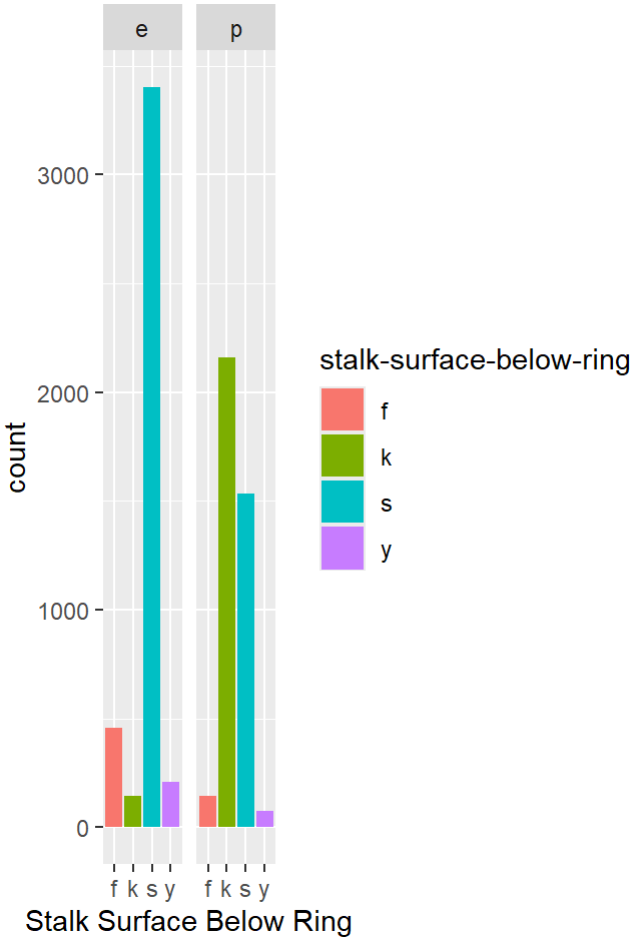





```
mush11 <- ggplot(aes(x = `stalk-surface-above-ring`, fill = `stalk-surface-ab
ove-ring`), data = mush) +
  geom_bar(stat = "count") +
  facet_wrap(~class) +
  xlab("Stalk Surface Above Ring")


mush12 <- ggplot(aes(x = `stalk-surface-below-ring`, fill = `stalk-surface-be
low-ring`), data = mush) +
  geom_bar(stat = "count") +
  facet_wrap(~class) +
  xlab("Stalk Surface Below Ring")


grid.arrange(mush11, mush12, ncol = 2)
```



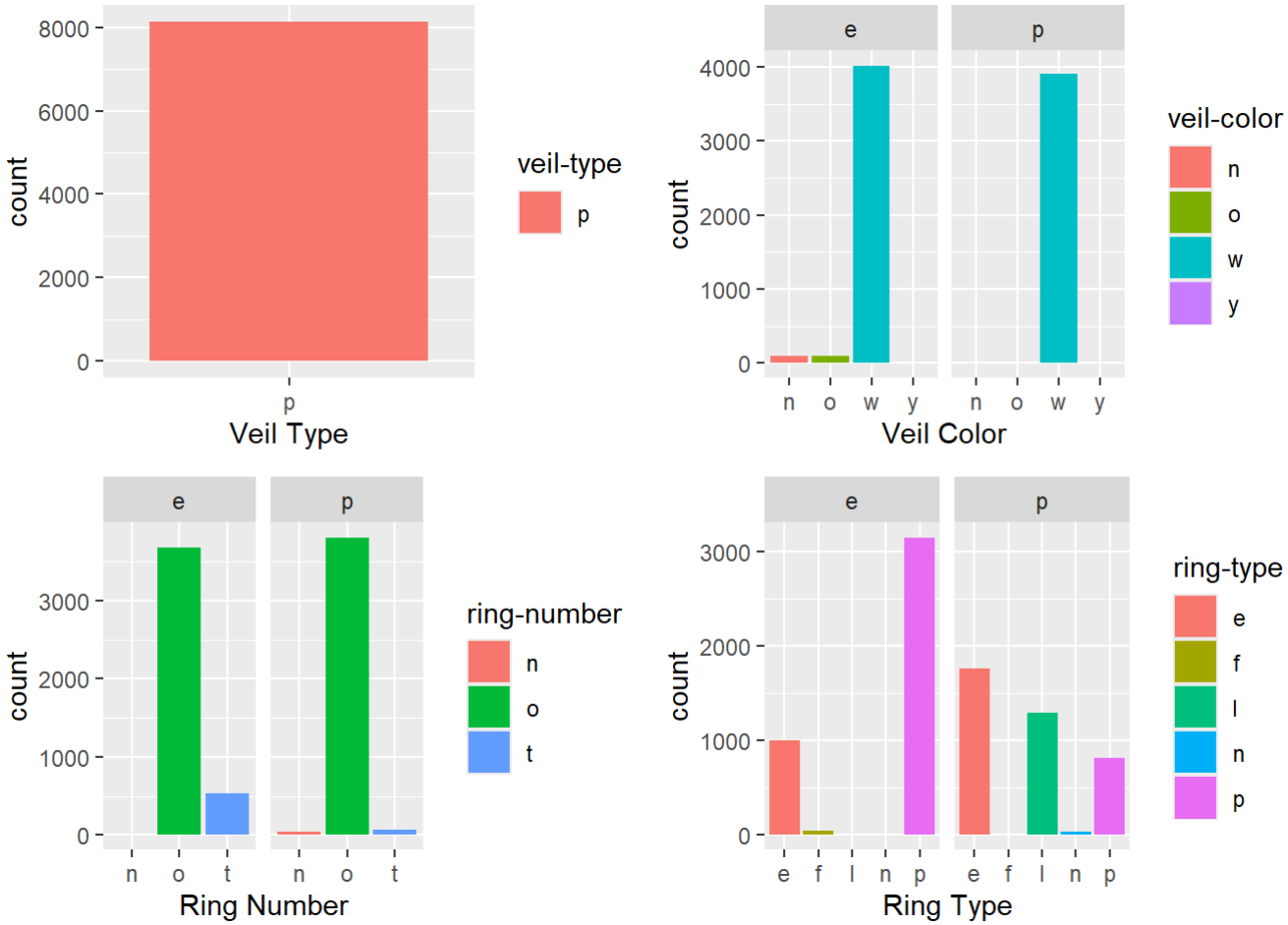
Stalk Surface Above Ring



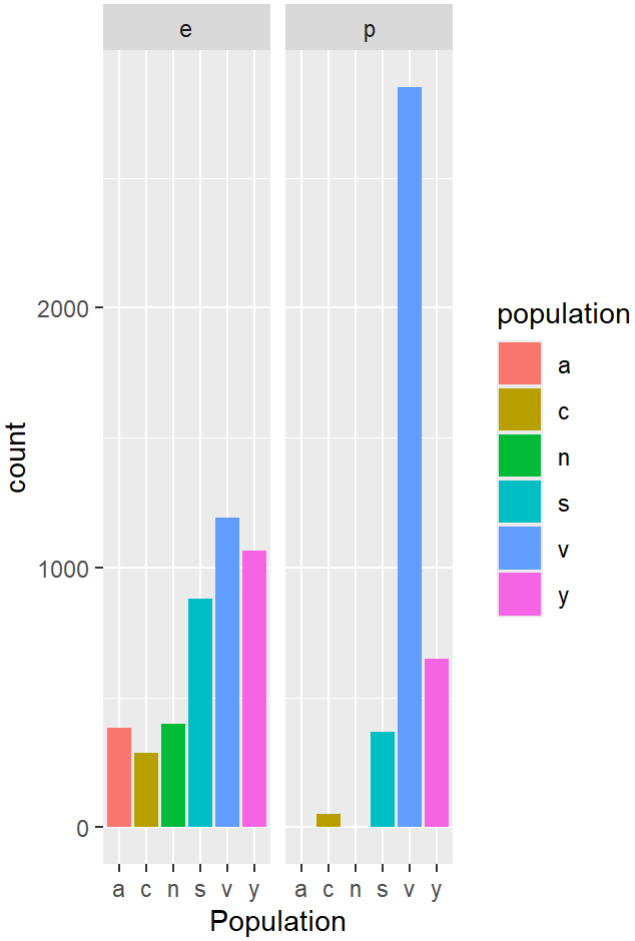
Stalk Surface Below Ring

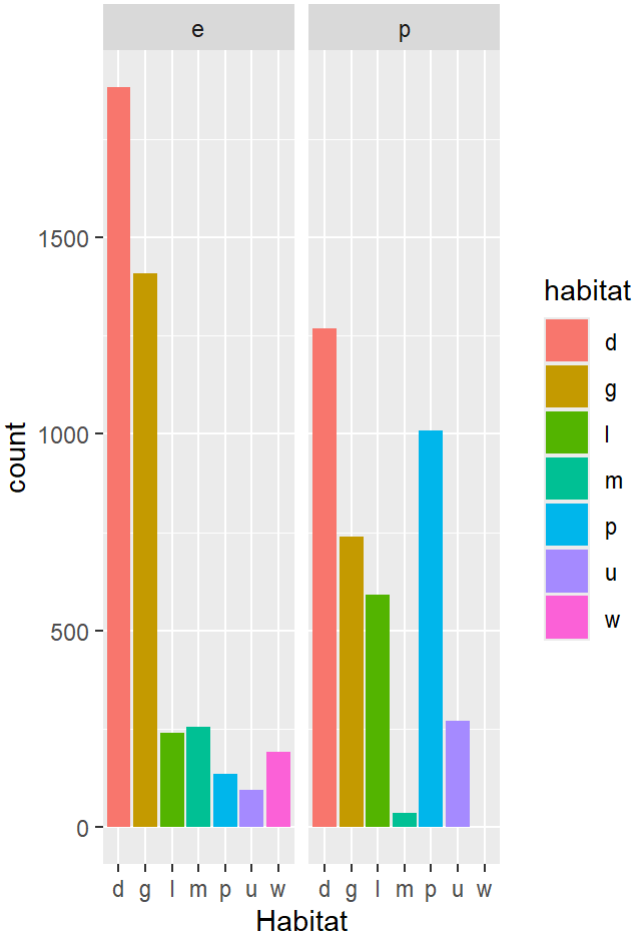
```
mush13 <- ggplot(aes(x = `veil-type`, fill = `veil-type`), data = mush) +
  geom_bar(stat = "count") +
  xlab("Veil Type")

mush14 <- ggplot(aes(x = `veil-color`, fill = `veil-color`), data = mush) +
  geom_bar(stat = "count") +
  facet_wrap(~class) +
  xlab("Veil Color")

mush15 <- ggplot(aes(x = `ring-number`, fill = `ring-number`), data = mush) +
  geom_bar(stat = "count") +
  facet_wrap(~class) +
  xlab("Ring Number")

mush16 <- ggplot(aes(x = `ring-type`, fill = `ring-type`), data = mush) +
  geom_bar(stat = "count") +
  facet_wrap(~class) +
  xlab("Ring Type")
```

```
grid.arrange(mush13, mush14, mush15, mush16, ncol = 2)
```




```
mush17 <- ggplot(aes(x = population, fill = population), data = mush) +
  geom_bar(stat = "count") +
  facet_wrap(~class) +
  xlab("Population")

mush18 <- ggplot(aes(x = habitat, fill = habitat), data = mush) +
  geom_bar(stat = "count") +
  facet_wrap(~class) +
  xlab("Habitat")

grid.arrange(mush17, mush18, ncol = 2)
```





Summary of the section:

Finding relationship

Next we will tried to find how closely the attributes are related to the class of the mushroom. The correlation between two categorical variables can be calculated by using the Chi-squared test.

```
tbl1 <- table(mush$class, mush$`cap-shape`)
chisq.test(tbl1)
```

```
## Warning in chisq.test(tbl1): аппроксимация на основе хи-квадрат может быть
## неправильной
##
##  Pearson's Chi-squared test
##
## data:  tbl1
## X-squared = 489.92, df = 5, p-value < 2.2e-16
```

```
tbl2 <- table(mush$class, mush$`cap-surface`)
chisq.test(tbl2)
```

```
## Warning in chisq.test(tbl2): аппроксимация на основе хи-квадрат может быть
## неправильной
##
```

```
##  Pearson's Chi-squared test
##
## data:  tbl2
## X-squared = 315.04, df = 3, p-value < 2.2e-16
tbl3 <- table(mush$class, mush$`cap-color`)
chisq.test(tbl3)
##
##  Pearson's Chi-squared test
##
## data:  tbl3
## X-squared = 387.6, df = 9, p-value < 2.2e-16
tbl4 <- table(mush$class, mush$bruises)
chisq.test(tbl4)
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tbl4
## X-squared = 2041.4, df = 1, p-value < 2.2e-16
tbl5 <- table(mush$class, mush$odor)
chisq.test(tbl5)
##
##  Pearson's Chi-squared test
##
## data:  tbl5
## X-squared = 7659.7, df = 8, p-value < 2.2e-16
tbl6 <- table(mush$class, mush$`gill-attachment`)
chisq.test(tbl6)
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tbl6
## X-squared = 133.99, df = 1, p-value < 2.2e-16
tbl7 <- table(mush$class, mush$`gill-spacing`)
chisq.test(tbl7)
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tbl7
## X-squared = 984.14, df = 1, p-value < 2.2e-16
```

```
tbl8 <- table(mush$class, mush$`gill-size`)
chisq.test(tbl8)
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tbl8
## X-squared = 2366.8, df = 1, p-value < 2.2e-16
tbl9 <- table(mush$class, mush$`stalk-shape`)
chisq.test(tbl9)
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tbl9
## X-squared = 84.142, df = 1, p-value < 2.2e-16
tbl10 <- table(mush$class, mush$`stalk-root`)
chisq.test(tbl10)
##
##  Pearson's Chi-squared test
##
## data:  tbl10
## X-squared = 1344.4, df = 4, p-value < 2.2e-16
tbl11 <- table(mush$class, mush$`stalk-surface-above-ring`)
chisq.test(tbl11)
##
##  Pearson's Chi-squared test
##
## data:  tbl11
## X-squared = 2808.3, df = 3, p-value < 2.2e-16
tbl12 <- table(mush$class, mush$`stalk-surface-below-ring`)
chisq.test(tbl12)
##
##  Pearson's Chi-squared test
##
## data:  tbl12
## X-squared = 2684.5, df = 3, p-value < 2.2e-16
tbl13 <- table(mush$class, mush$`veil-type`)
chisq.test(tbl13)
##
##  Chi-squared test for given probabilities
```

```
##
## data:  tbl13
## X-squared = 10.495, df = 1, p-value = 0.001197
```

```
tbl14 <- table(mush$class, mush$`veil-color`)
chisq.test(tbl14)
```

```
## Warning in chisq.test(tbl14): аппроксимация на основе хи-квадрат может быт
ь
## неправильной
##
##  Pearson's Chi-squared test
##
## data:  tbl14
## X-squared = 191.22, df = 3, p-value < 2.2e-16
```

```
tbl15 <- table(mush$class, mush$`ring-number`)
chisq.test(tbl15)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tbl15
## X-squared = 374.74, df = 2, p-value < 2.2e-16
```

```
tbl16 <- table(mush$class, mush$`ring-type`)
chisq.test(tbl16)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tbl16
## X-squared = 2956.6, df = 4, p-value < 2.2e-16
```

```
tbl17 <- table(mush$class, mush$population)
chisq.test(tbl17)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tbl17
## X-squared = 1929.7, df = 5, p-value < 2.2e-16
```

```
tbl18 <- table(mush$class, mush$habitat)
chisq.test(tbl18)
```

```
##
##  Pearson's Chi-squared test
##
```

```
## data:  tbl18
## X-squared = 1573.8, df = 6, p-value < 2.2e-16
```

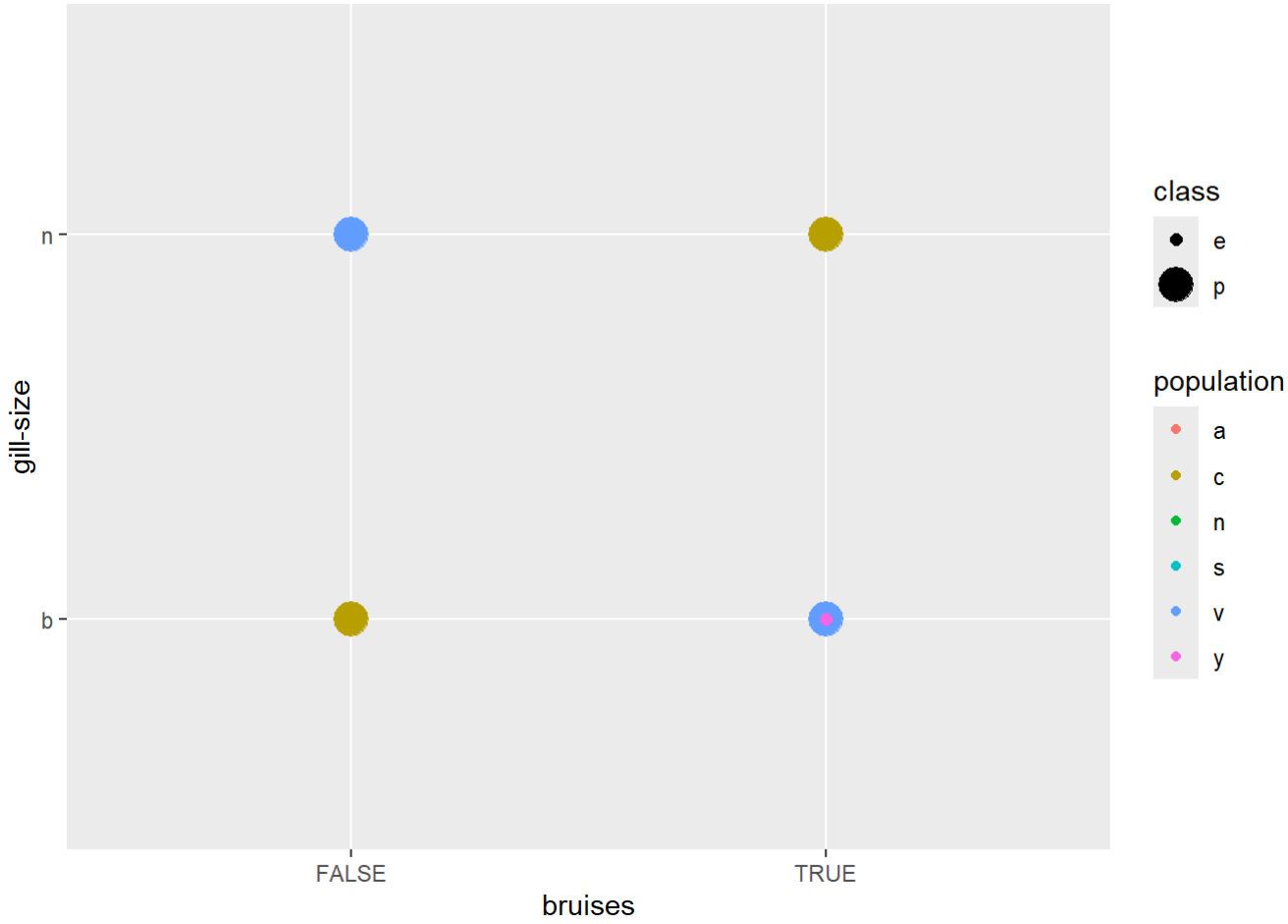
Based on correlation between the class of the mushroom and the other attributes based on the Chi-squared test, I have chosen for further analysis: stalk surface above ring, stalk surface below ring, gill size and bruises.

Multivariate analysis

We have taken into account two attributes along with class and how they contribute the the edibility of the mushrooms. The first graph is of bruises and gill.size. The second graph is of stalk.surface.above.ring and stalk.surface.below.ring. The choice of these two attributes is based on the Chi-squared test values for correlation. The purpose of these plots is to find the combined exclusivity of attributes in deciding the edibility of mushroom. The observations are noted after each graph.

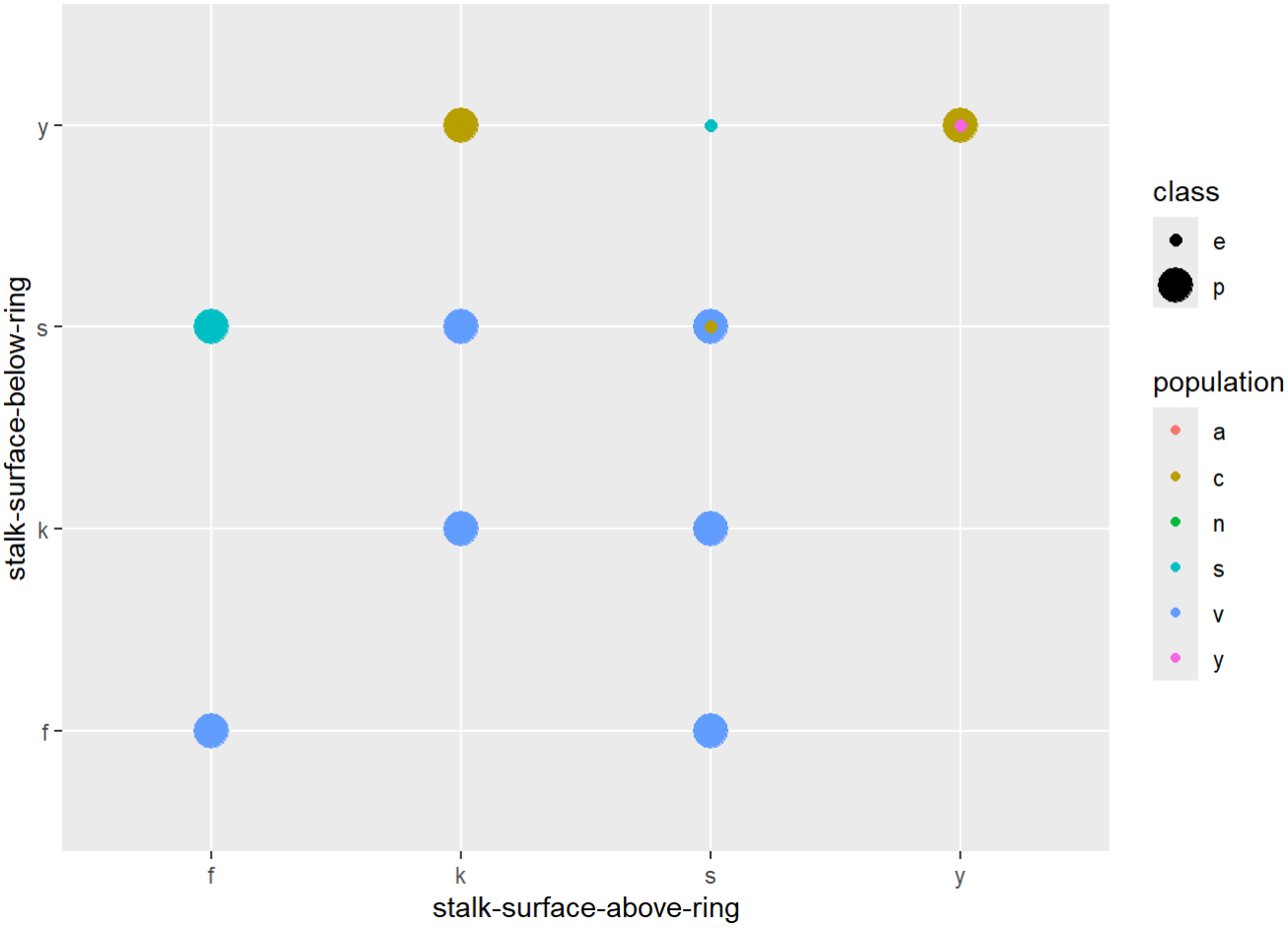
```
ggplot(mush, aes(bruises, `gill-size`, col = population, size = class)) +
  geom_point()
```

```
## Warning: Using size for a discrete variable is not advised.
```




```
ggplot(mush, aes(`stalk-surface-above-ring`, `stalk-surface-below-ring`, col
= population, size = class)) +
  geom_point()
```

```
## Warning: Using size for a discrete variable is not advised.
```



Summary of the section: Both of them cases when the combination of two attributes is taken into consideration, the class of the mushroom can be predicted. The combination of two can be extended to many possibilities since it is difficult to find the importance of every attribute in deciding the edibility of a given mushroom.

Logistic regression

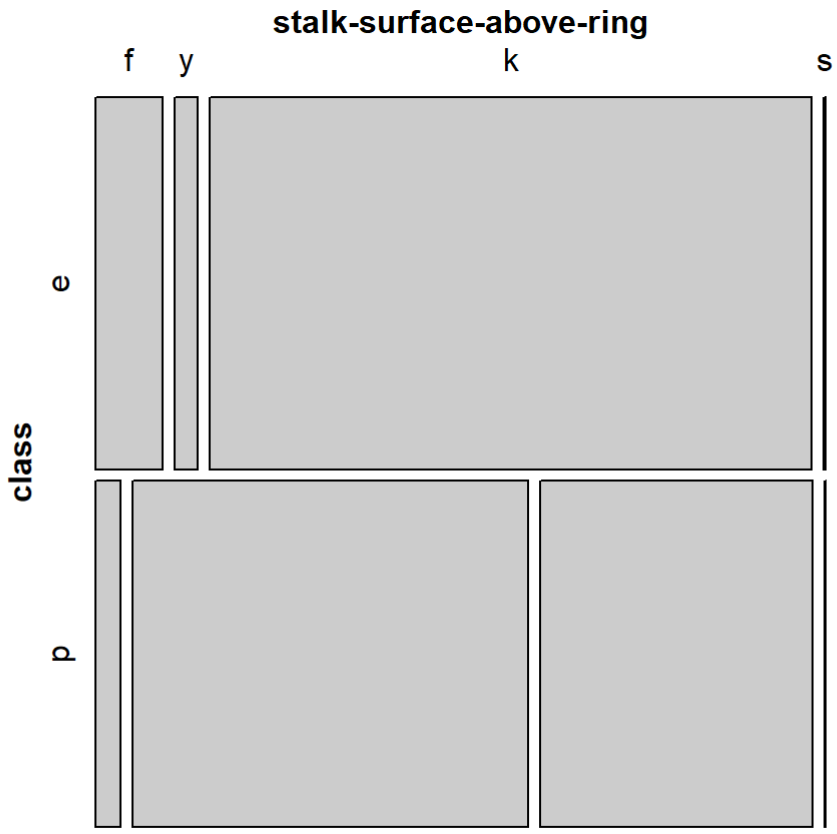
Task.

I want to predict the probability of getting a poisonous mushroom To do this, I built a regression model in which the dependent variable is whether the mushroom is poisonous or not. Predictors: stalk.surface.above.ring (fibrous=f, scaly=y, silky=k, smooth=s), gill size: (wide =b, narrow=n) and their interaction.

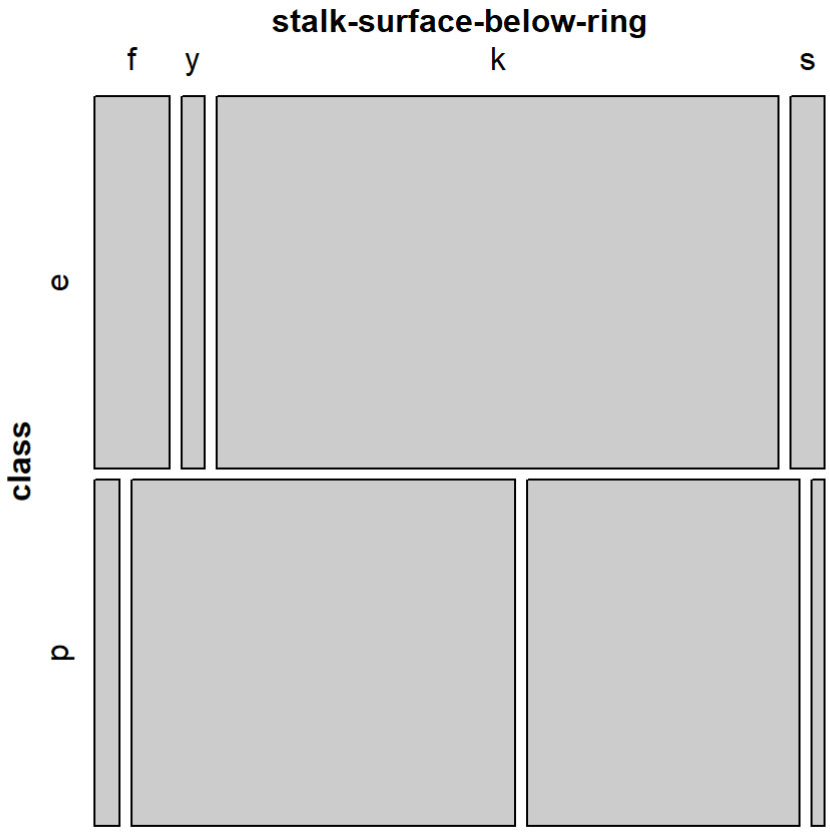
Let's build a mosaic graph

```
mush_train <- mutate(mush,
                     class = factor(class, labels = c("e", "p")),
                     `stalk-surface-above-ring` = factor(`stalk-surface-above
-ring`, labels = c("f","y","k","s")),
                     `stalk-surface-below-ring` = factor(`stalk-surface-below
-ring`, labels = c("f","y","k","s")),
                     `gill-size` = factor(`gill-size`, labels = c("b","n")),
                     bruises = factor(bruises, labels = c("t","f")))
```

```
mosaic(~ class + `stalk-surface-above-ring`, data=mush_train)
```

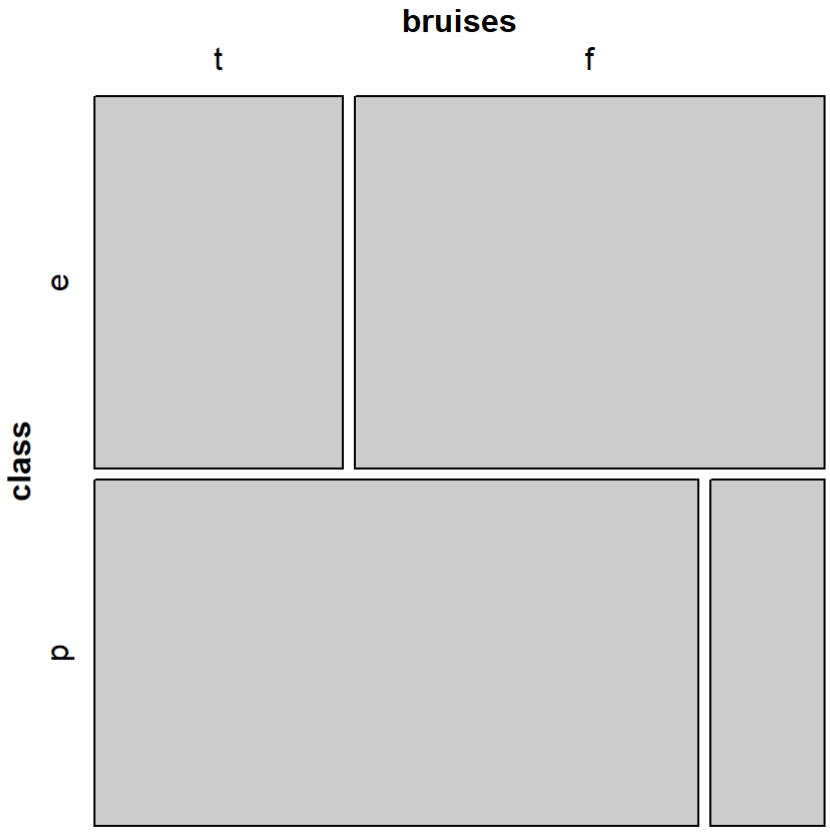



```
mosaic(~ class + `stalk-surface-below-ring`, data=mush_train)
```




```
mosaic(~ class + `gill-size`, data=mush_train)
```


```
mosaic(~ class + bruises, data=mush_train)
```



The figure shows three main hypotheses: a. hypothesis about the importance of the first factor - the edibility of mushrooms; b. hypothesis about the importance of the second factor - stalk-surface-above ring c. hypothesis about the importance of the third factor -gill-size d. hypothesis about the importance of the fourth factor - bruises; e. hypothesis about the interaction of these factors.

1. Intercept only model

```
mush_simple <- glm(class ~ 1, mush_train,  family = "binomial")
coef(mush_simple)
## (Intercept)
## -0.07191675
table(mush_train$class)
##
##    e    p
## 4208 3916
```

Intercept is the natural logarithm of the chances of a positive outcome To calculate the probability of a positive outcome, you need to calculate the indicator of the degree of “interception” of the output: “-0.07191675” is the probability of taking Mushroom’edible, which is significantly less than the probability that you will take Mushroom’poisonous.

2. A model with a single nominative predictor

```
mush_fit1 <- glm(class ~ `stalk-surface-above-ring`, mush_train, family = "binomial")
coef(mush_fit1)
##                  (Intercept) `stalk-surface-above-ring`y
##                   -1.0414539                   3.7805002
## `stalk-surface-above-ring`k `stalk-surface-above-ring`s
##                    0.1786518                   0.3483067
summary(mush_fit1)
##
## Call:
## glm(formula = class ~ `stalk-surface-above-ring`, family = "binomial",
##     data = mush_train)
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -1.04145    0.09693 -10.744   <2e-16 ***
## `stalk-surface-above-ring`y  3.78050    0.12957  29.177   <2e-16 ***
## `stalk-surface-above-ring`k  0.17865    0.10159   1.759   0.0787 .
## `stalk-surface-above-ring`s  0.34831    0.44373   0.785   0.4325
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11251.8  on 8123  degrees of freedom
## Residual deviance:  8045.1  on 8120  degrees of freedom
## AIC: 8053.1
##
## Number of Fisher Scoring iterations: 5
table(mush_train$class, mush_train$`stalk-surface-above-ring`)
##
##        f    y    k    s
##   e  408  144 3640   16
##   p  144 2228 1536    8
```

Intercept - the natural logarithm of the chances of Mushroom edible for fibrous=f. The stalk.surface.above.ringy, stalk.surface.above.ring and stalk.surface.above.rings are the logarithms of the ratio of the chances of Mushroom'edible for scaly=y,silky=k,smooth=s

3. A model with two categorical predictors

```
mush_fit2 <- glm(class ~ `stalk-surface-above-ring` * `gill-size`, mush_trai
n, family = "binomial")
coef(mush_fit2)
```

```
##                                (Intercept)
##                                 -0.9808293
##              `stalk-surface-above-ring`y
##                                  3.2054528
##              `stalk-surface-above-ring`k
##                                 -1.7683392
##              `stalk-surface-above-ring`s
##                                -17.5852393
##                               `gill-size`n
##                                -17.5852390
## `stalk-surface-above-ring`y:`gill-size`n
##                                 33.9266840
## `stalk-surface-above-ring`k:`gill-size`n
##                                 21.9438454
## `stalk-surface-above-ring`s:`gill-size`n
##                                 54.7173760
```

```
summary(mush_fit2)
```

```
##
## Call:
## glm(formula = class ~ `stalk-surface-above-ring` * `gill-size`,
##     family = "binomial", data = mush_train)
##
## Coefficients:
##                                           Estimate Std. Error z value Pr
(>|z|)
## (Intercept)                               -0.98083    0.09772 -10.037   <
2e-16
## `stalk-surface-above-ring`y                3.20545    0.13132  24.410   <
2e-16
## `stalk-surface-above-ring`k               -1.76834    0.12031 -14.698   <
2e-16
## `stalk-surface-above-ring`s              -17.58524 1630.65965  -0.011
0.991
## `gill-size`n                             -17.58524 1331.42801  -0.013
0.989
## `stalk-surface-above-ring`y:`gill-size`n  33.92668 1349.14181   0.025
0.980
## `stalk-surface-above-ring`k:`gill-size`n  21.94385 1331.42802   0.016
0.987
```

```
## `stalk-surface-above-ring`s:`gill-size`n   54.71738 3122.47550   0.018
0.986
##
## (Intercept)                                ***
## `stalk-surface-above-ring`y                ***
## `stalk-surface-above-ring`k                ***
## `stalk-surface-above-ring`s
## `gill-size`n
## `stalk-surface-above-ring`y:`gill-size`n
## `stalk-surface-above-ring`k:`gill-size`n
## `stalk-surface-above-ring`s:`gill-size`n
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11252  on 8123  degrees of freedom
## Residual deviance:  4623  on 8116  degrees of freedom
## AIC: 4639
##
## Number of Fisher Scoring iterations: 17
```

```
table(mush_train$class, mush_train$`stalk-surface-above-ring`, mush_train$`gi
ll-size`)
```

```
## , ,  = b
##
##
##        f    y    k    s
##   e  384  144 3376   16
##   p  144 1332  216    0
##
## , ,  = n
##
##
##        f    y    k    s
##   e   24    0  264    0
##   p    0  896 1320    8
```

Intercept - the natural logarithm of the chances of Mushroom edible for fibrous=f. The stalk.surface.above.ringy, stalk.surface.above.ring and stalk.surface.above.rings are the logarithms of the ratio of the chances of Mushroom'edible for scaly=y,silky=k,smooth=s The stalk.surface.above.ringy:gill.sizen, stalk.surface.above.ringk:gill.sizen,

and stalk.surface.above.rings:gill.sizen are the logarithms of the ratio of the chances of Mushroom'edible for scaly=y,silky=k,smooth=s wiht gill-size of broad.

4. model comparison

```
anova(mush_fit1, mush_fit2, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: class ~ `stalk-surface-above-ring`
## Model 2: class ~ `stalk-surface-above-ring` * `gill-size`
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      8120     8045.1
## 2      8116     4623.0  4   3422.1 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The quality indicator of the model when adding the interaction parameter with the gill.size, the quality indicator of the second model became lower than the first, and therefore the second model works better than the first.

5. predicting new data

```
mush_fit3 <- glm(class ~ bruises, mush_train, family = "binomial")
coef(mush_fit2)
```

```
##                                 (Intercept)
##                                   -0.9808293
##                `stalk-surface-above-ring`y
##                                    3.2054528
##                `stalk-surface-above-ring`k
##                                   -1.7683392
##                `stalk-surface-above-ring`s
##                                  -17.5852393
##                               `gill-size`n
##                                  -17.5852390
## `stalk-surface-above-ring`y:`gill-size`n
##                                   33.9266840
## `stalk-surface-above-ring`k:`gill-size`n
##                                   21.9438454
## `stalk-surface-above-ring`s:`gill-size`n
##                                   54.7173760
```

```
summary(mush_fit3)
```

```
##
```

```
## Call:
## glm(formula = class ~ bruises, family = "binomial", data = mush_train)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.81580    0.03147   25.92   <2e-16 ***
## bruisesf    -2.29974    0.05437  -42.30   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11251.8  on 8123  degrees of freedom
## Residual deviance:  9085.1  on 8122  degrees of freedom
## AIC: 9089.1
##
## Number of Fisher Scoring iterations: 4
```

```
new_mush <- data.frame(class = "e", bruises = "t")
predict(mush_fit3, newdata = new_mush)
```

```
##         1
## 0.8158023
```

The model is correctly predicts all test cases

6. Check a model with different variables

```
mush_fit4 <- glm(class ~ `stalk-surface-above-ring` + `gill-size` + bruises,
mush_train, family = "binomial")
summary(mush_fit4)
```

```
##
## Call:
## glm(formula = class ~ `stalk-surface-above-ring` + `gill-size` +
##     bruises, family = "binomial", data = mush_train)
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -1.7153     0.1224 -14.014   <2e-16 ***
## `stalk-surface-above-ring`y   3.9441     0.1504  26.225   <2e-16 ***
## `stalk-surface-above-ring`k  -2.1957     0.1451 -15.128   <2e-16 ***
## `stalk-surface-above-ring`s  -1.2585     0.8408  -1.497    0.134
```

```
## `gill-size`n                    5.2022     0.1483  35.083   <2e-16 ***
## bruisesf                        1.4404     0.1371  10.503   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11251.8  on 8123  degrees of freedom
## Residual deviance:  4646.1  on 8118  degrees of freedom
## AIC: 4658.1
##
## Number of Fisher Scoring iterations: 6
```

```
anova(mush_fit4, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: class
##
## Terms added sequentially (first to last)
##
##
##                            Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                         8123    11251.8
## `stalk-surface-above-ring`  3   3206.7      8120     8045.1 < 2.2e-16 ***
## `gill-size`                 1   3260.1      8119     4785.0 < 2.2e-16 ***
## bruises                     1    138.8      8118     4646.1 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All three variables (stalk.surface.above.ring + gill.size + bruises) allow us to significantly predict Mushroom'edible

Total resume: The mushroom dataset is analysed in three ways. The first this is histograms to explore the contribution of a single attribute in deciding the edibility of the mushroom. The second is calculation oriented based on contribution of single attribute towards the class of mushroom. The dataset has only categorical variables for all attributes. It has used the Chi-squared Test to determine the correlation between a given attribute and the class of mushroom. The correlation test helped in establishing relationship between each attribute and the class (edibility) of the mushroom. The higher X-squared implies higher correlation. The third it was drawing plots that would investigate the exclusiveness of two attributes taken together on classify them according to that class of the mushroom. For predicting it was use Logistic regression. We predicted the probability of getting a poisonous mushroom To do this, it was built a regression model in which the dependent variable is whether the mushroom is poisonous or not.