Predicting whether a Mushroom is Edible or Poisonous

MATH 2319 Machine Learning Applied Project Phase I

Wesley Paul Nderi (s3635870)

03/04/2018

1. Introduction

This dataset for the purposes of this assignment is sourced from Kaggle (https://www.kaggle.com/uciml/mushroom-classification) although it was originally cited in the UCI Machine Learning Repository. This dataset includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family Mushroom drawn from The Audubon Society Field Guide to North American Mushrooms (1981).

The goal in this project is to predict whether a mushroom is edible or poisonous based on its descriptive features.The project has two phases. **Phase I** focuses on data preprocessing and exploration, as covered in this report. We shall explore model building in **Phase II**.

The rest of this report is organised as follows. Section 2 describes the data sets and its attributes. Section 3 covers data pre-processing. In Section 4, we explore each attribute and their inter-relationships. The final section ends with a summary.

2. Data Set

It is important to highlight that the original dataset available on the UCI Repository has 22 attributes while this has 23 attributes. This is with the addition of the **class** attribute which serves as a target feature.

2.1 Target feature

The target feature is the class. It has two possible values either **edible** or **poisonous** and hence it is a binary classification problem.

2.2 Descriptive Features

This dataset has 22 attributes which we shall explore below:

1. **cap-shape**: bell=b, conical=c, convex=x, flat=f, knobbed=k, sunken=s
2. **cap-surface**: fibrous=f, grooves=g, scaly=y, smooth=s
3. **cap-color**: brown=n, buff=b, cinnamon=c, gray=g, green=r, pink=p,purple=u,red=e, white=w, yellow=y
4. **bruises**: bruises=t,no=f
5. **odor**: almond=a, anise=l, creosote=c, fishy=y, foul=f, musty=m, none=n, pungent=p, spicy=s
6. **gill-attachment**: attached=a, descending=d, free=f, notched=n
7. **gill-spacing**: close=c, crowded=w, distant=d
8. **gill-size**: broad=b,narrow=n

9. **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
10. **stalk-shape**: enlarging=e, tapering=t
11. **stalk-root**: bulbous=b, club=c, cup=u, equal=e, rhizomorphs=z, rooted=r, missing=?
12. **stalk-surface-above-ring**: fibrous=f, scaly=y, silky=k, smooth=s
13. **stalk-surface-below-ring**: fibrous=f, scaly=y, silky=k, smooth=s
14. **stalk-color-above-ring**: brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
15. **stalk-color-below-ring**: brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
16. **veil-type**: partial=p, universal=u
17. **veil-color**: brown=n, orange=o, white=w, yellow=y
18. **ring-number**: none=n, one=o, two=t
19. **ring-type**: cobwebby=c, evanescent=e, flaring=f, large=l, none=n, pendant=p, sheathing=s, zone=z
20. **spore-print-color**: black=k, brown=n, buff=b, chocolate=h, green=r, orange=o, purple=u, white=w, yellow=y
21. **population**: abundant=a, clustered=c, numerous=n, scattered=s, several=v, solitary=y
22. **habitat**: grasses=g, leaves=l, meadows=m, paths=p, urban=u, waste=w, woods=d
23. **class:** edible=e, poisonous=p

As we can see, the dataset has features which are entirely categorical in nature.

3. Data Preprocessing

3.1 Preliminaries

In this project we used the following R packages.

```
library(knitr)
library(GGally)
library(ggplot2)
library(dplyr)
library(mlr)
library(cowplot)
library(ggmosaic)
```

For consistency of the data labels during the preprocessing of the data, we manually renamed the columns.

```
mushroom1 <-read.csv("/Users/wes/Downloads/mushrooms.csv")
names(mushroom1) <- c("class","cap_shape", "cap_surface", "cap_color","bruises", "odo
r", "gill_attachment", "gill_spacing", "gill_size", "gill_color", "stalk_shape", "sta
lk_root", "stalk_surface_above_ring", "stalk_surface_below_ring", "stalk_color_above_
ring", "stalk_color_below_ring", "veil_type", "veil_color", "ring_number", "ring_typ
e", "spore_print_color", "population", "habitat")
```

3.2 Data Cleaning and Transformation

A quick summary to understand our data better.

```
summary(mushroom1)
```

```
##  class    cap_shape cap_surface   cap_color    bruises       odor
##  e:4208   b: 452    f:2320      n      :2284   f:4748   n      :3528
##  p:3916   c:   4    g:   4      g      :1840   t:3376   f      :2160
##           f:3152    s:2556      e      :1500            s      : 576
##           k: 828    y:3244      y      :1072            y      : 576
##           s:  32                w      :1040            a      : 400
##           x:3656                b      : 168            l      : 400
##                                 (Other): 220            (Other): 484
##  gill_attachment gill_spacing gill_size   gill_color   stalk_shape
##  a: 210          c:6812       b:5612    b      :1728   e:3516
##  f:7914          w:1312       n:2512    p      :1492   t:4608
##                                         w      :1202
##                                         n      :1048
##                                         g      : 752
##                                         h      : 732
##                                         (Other):1170
##  stalk_root stalk_surface_above_ring stalk_surface_below_ring
##  ?:2480     f: 552                   f: 600
##  b:3776     k:2372                   k:2304
##  c: 556     s:5176                   s:4936
##  e:1120     y:  24                   y: 284
##  r: 192
##
##
##  stalk_color_above_ring stalk_color_below_ring veil_type veil_color
##  w      :4464           w      :4384           p:8124    n:  96
##  p      :1872           p      :1872                     o:  96
##  g      : 576           g      : 576                     w:7924
##  n      : 448           n      : 512                     y:   8
##  b      : 432           b      : 432
##  o      : 192           o      : 192
##  (Other): 140           (Other): 156
##  ring_number ring_type spore_print_color population habitat
##  n:  36      e:2776    w      :2388      a: 384     d:3148
##  o:7488      f:  48    n      :1968      c: 340     g:2148
##  t: 600      l:1296    k      :1872      n: 400     l: 832
##              n:  36    h      :1632      s:1248     m: 292
##              p:3968    r      :  72      v:4040     p:1144
##                        b      :  48      y:1712     u: 368
##                        (Other): 144                 w: 192
```

We can also have a more concise look at the number of class of each variable.

```
b<-cbind.data.frame(Var=names(mushroom1), Total_Class=sapply(mushroom1,function(x){a
s.numeric(length(levels(x)))}))
print(b)
```

```
##                                           Var Total_Class
## class                                   class           2
## cap_shape                           cap_shape           6
## cap_surface                       cap_surface           4
## cap_color                           cap_color          10
## bruises                               bruises           2
## odor                                     odor           9
## gill_attachment               gill_attachment           2
## gill_spacing                     gill_spacing           2
## gill_size                           gill_size           2
## gill_color                         gill_color          12
## stalk_shape                       stalk_shape           2
## stalk_root                         stalk_root           5
## stalk_surface_above_ring stalk_surface_above_ring      4
## stalk_surface_below_ring stalk_surface_below_ring      4
## stalk_color_above_ring     stalk_color_above_ring      9
## stalk_color_below_ring     stalk_color_below_ring      9
## veil_type                           veil_type           1
## veil_color                         veil_color           4
## ring_number                       ring_number           3
## ring_type                           ring_type           5
## spore_print_color           spore_print_color           9
## population                         population           6
## habitat                               habitat           7
```

We note that only **veil_type** has one class. In order to improve our models, it would be advantageous to remove the variable from the data.

```
#Omitting the feature veil_type
mushroom1$veil_type<-NULL
```

The attribute **stalk_root** is the only attribute in the data-set that has a peculiarity. One of the levels in the attribute is marked as a **?** and described as missing. In the context of the data, this means that no observation could be made in regard to the stalk root. This presents a challenge as it could be a case of incorrect labelling and there are mushrooms that do not have this feature and hence constitutes a valid observation. Alternatively, it could be a case of missing values arising from invalid data intergration techniques or some other reason.

Let us explore this attribute **stalk_root** further as shown below.

```
stalk_root.tab<-table(mushroom1$class, mushroom1$stalk_root)
stalk_root.tab
```

```
##
##        ?    b    c    e    r
##   e  720 1920  512  864  192
##   p 1760 1856   44  256    0
```

We can see that the level **?** has quite a high frequency of both **edible** (e) and **poisonous** (p) which are the binary levels in our target feature.

In total this is:

```
margin.table(stalk_root.tab,2) #Column frequencies
```

```
##
##    ?    b    c    e    r
## 2480 3776  556 1120  192
```

It is the second most frequent level in this attribute.

For the purposes of this project, we shall treat this as a case of missing values arising from valid data exploration and consider that since the proportion of missing values is above 60% to omit this feature.

```
#Omitting the feature stalk_root
mushroom1$stalk_root<-NULL
```

The table below presents the summary statistics after data-preprocessing.

```
summarizeColumns(mushroom1) %>% kable( caption = 'Feature Summary statistics after Da
ta Preprocessing' )
```

Feature Summary statistics after Data Preprocessing

name	type	na	mean	disp	median	mad	min	max	nlevs
class	factor	0	NA	0.4820286	NA	NA	3916	4208	2
cap_shape	factor	0	NA	0.5499754	NA	NA	4	3656	6
cap_surface	factor	0	NA	0.6006893	NA	NA	4	3244	4
cap_color	factor	0	NA	0.7188577	NA	NA	16	2284	10
bruises	factor	0	NA	0.4155588	NA	NA	3376	4748	2
odor	factor	0	NA	0.5657312	NA	NA	36	3528	9
gill_attachment	factor	0	NA	0.0258493	NA	NA	210	7914	2
gill_spacing	factor	0	NA	0.1614968	NA	NA	1312	6812	2
gill_size	factor	0	NA	0.3092073	NA	NA	2512	5612	2
gill_color	factor	0	NA	0.7872969	NA	NA	24	1728	12
stalk_shape	factor	0	NA	0.4327917	NA	NA	3516	4608	2
stalk_surface_above_ring	factor	0	NA	0.3628754	NA	NA	24	5176	4
stalk_surface_below_ring	factor	0	NA	0.3924175	NA	NA	284	4936	4
stalk_color_above_ring	factor	0	NA	0.4505170	NA	NA	8	4464	9
stalk_color_below_ring	factor	0	NA	0.4603644	NA	NA	24	4384	9
veil_color	factor	0	NA	0.0246184	NA	NA	8	7924	4
ring_number	factor	0	NA	0.0782866	NA	NA	36	7488	3
ring_type	factor	0	NA	0.5115707	NA	NA	36	3968	5
spore_print_color	factor	0	NA	0.7060561	NA	NA	48	2388	9
population	factor	0	NA	0.5027080	NA	NA	340	4040	6
habitat	factor	0	NA	0.6125062	NA	NA	192	3148	7

```
#What are the classes of the variables in the dataset?
sapply(mushroom1,class)
```

```
##                    class                cap_shape              cap_surface
##                 "factor"                 "factor"                 "factor"
##                cap_color                  bruises                     odor
##                 "factor"                 "factor"                 "factor"
##          gill_attachment             gill_spacing                gill_size
##                 "factor"                 "factor"                 "factor"
##               gill_color              stalk_shape stalk_surface_above_ring
##                 "factor"                 "factor"                 "factor"
## stalk_surface_below_ring   stalk_color_above_ring   stalk_color_below_ring
##                 "factor"                 "factor"                 "factor"
##               veil_color              ring_number                ring_type
##                 "factor"                 "factor"                 "factor"
##        spore_print_color              population                  habitat
##                 "factor"                 "factor"                 "factor"
```

```
str(mushroom1)
```

```
## 'data.frame':    8124 obs. of  21 variables:
##  $ class                   : Factor w/ 2 levels "e","p": 2 1 1 2 1 1 1 1 2 1 ...
##  $ cap_shape               : Factor w/ 6 levels "b","c","f","k",..: 6 6 1 6 6 6 1
1 6 1 ...
##  $ cap_surface             : Factor w/ 4 levels "f","g","s","y": 3 3 3 4 3 4 3 4 4
3 ...
##  $ cap_color               : Factor w/ 10 levels "b","c","e","g",..: 5 10 9 9 4 10
9 9 9 10 ...
##  $ bruises                 : Factor w/ 2 levels "f","t": 2 2 2 2 1 2 2 2 2 2 ...
##  $ odor                    : Factor w/ 9 levels "a","c","f","l",..: 7 1 4 7 6 1 1
4 7 1 ...
##  $ gill_attachment         : Factor w/ 2 levels "a","f": 2 2 2 2 2 2 2 2 2 2 ...
##  $ gill_spacing            : Factor w/ 2 levels "c","w": 1 1 1 1 2 1 1 1 1 1 ...
##  $ gill_size               : Factor w/ 2 levels "b","n": 2 1 1 2 1 1 1 1 2 1 ...
##  $ gill_color              : Factor w/ 12 levels "b","e","g","h",..: 5 5 6 6 5 6 3
6 8 3 ...
##  $ stalk_shape             : Factor w/ 2 levels "e","t": 1 1 1 1 2 1 1 1 1 1 ...
##  $ stalk_surface_above_ring: Factor w/ 4 levels "f","k","s","y": 3 3 3 3 3 3 3 3 3
3 ...
##  $ stalk_surface_below_ring: Factor w/ 4 levels "f","k","s","y": 3 3 3 3 3 3 3 3 3
3 ...
##  $ stalk_color_above_ring  : Factor w/ 9 levels "b","c","e","g",..: 8 8 8 8 8 8 8
8 8 8 ...
##  $ stalk_color_below_ring  : Factor w/ 9 levels "b","c","e","g",..: 8 8 8 8 8 8 8
8 8 8 ...
##  $ veil_color              : Factor w/ 4 levels "n","o","w","y": 3 3 3 3 3 3 3 3 3
3 ...
##  $ ring_number             : Factor w/ 3 levels "n","o","t": 2 2 2 2 2 2 2 2 2 2
...
##  $ ring_type               : Factor w/ 5 levels "e","f","l","n",..: 5 5 5 5 1 5 5
5 5 5 ...
##  $ spore_print_color       : Factor w/ 9 levels "b","h","k","n",..: 3 4 4 3 4 3 3
4 3 3 ...
##  $ population              : Factor w/ 6 levels "a","c","n","s",..: 4 3 3 4 1 3 3
4 5 4 ...
##  $ habitat                 : Factor w/ 7 levels "d","g","l","m",..: 6 2 4 6 2 2 4
4 2 4 ...
```

```
sapply(mushroom1[sapply(mushroom1, is.factor)], table)
```

```
## $class
##
##    e    p
## 4208 3916
##
## $cap_shape
##
##    b    c    f    k    s    x
##  452    4 3152  828   32 3656
##
## $cap_surface
##
##    f    g    s    y
## 2320    4 2556 3244
##
## $cap_color
##
##    b    c    e    g    n    p    r    u    w    y
##  168   44 1500 1840 2284  144   16   16 1040 1072
##
## $bruises
##
##    f    t
## 4748 3376
##
## $odor
##
##    a    c    f    l    m    n    p    s    y
##  400  192 2160  400   36 3528  256  576  576
##
## $gill_attachment
##
##    a    f
##  210 7914
##
## $gill_spacing
##
##    c    w
## 6812 1312
##
## $gill_size
##
##    b    n
## 5612 2512
##
## $gill_color
##
##    b    e    g    h    k    n    o    p    r    u    w    y
## 1728   96  752  732  408 1048   64 1492   24  492 1202   86
##
## $stalk_shape
##
##    e    t
## 3516 4608
##
## $stalk_surface_above_ring
##
```

```
##    f    k    s    y
##  552 2372 5176   24
##
## $stalk_surface_below_ring
##
##    f    k    s    y
##  600 2304 4936  284
##
## $stalk_color_above_ring
##
##    b    c    e    g    n    o    p    w    y
##  432   36   96  576  448  192 1872 4464    8
##
## $stalk_color_below_ring
##
##    b    c    e    g    n    o    p    w    y
##  432   36   96  576  512  192 1872 4384   24
##
## $veil_color
##
##    n    o    w    y
##   96   96 7924    8
##
## $ring_number
##
##    n    o    t
##   36 7488  600
##
## $ring_type
##
##    e    f    l    n    p
## 2776   48 1296   36 3968
##
## $spore_print_color
##
##    b    h    k    n    o    r    u    w    y
##   48 1632 1872 1968   48   72   48 2388   48
##
## $population
##
##    a    c    n    s    v    y
##  384  340  400 1248 4040 1712
##
## $habitat
##
##    d    g    l    m    p    u    w
## 3148 2148  832  292 1144  368  192
```

We can see that all the variables are factors which shapes the kind of visualisations we shall use in the following section as we explore the relationships between the features.

4. Data exploration

4.1. Categorical features

4.1.1 Univariate Visualisations

In this section, each feature is explored individually and split by the classes of the target feature. It is also important to highlight that as the features are all categorical, we shall only focus on explorations suited to this kind of variable.

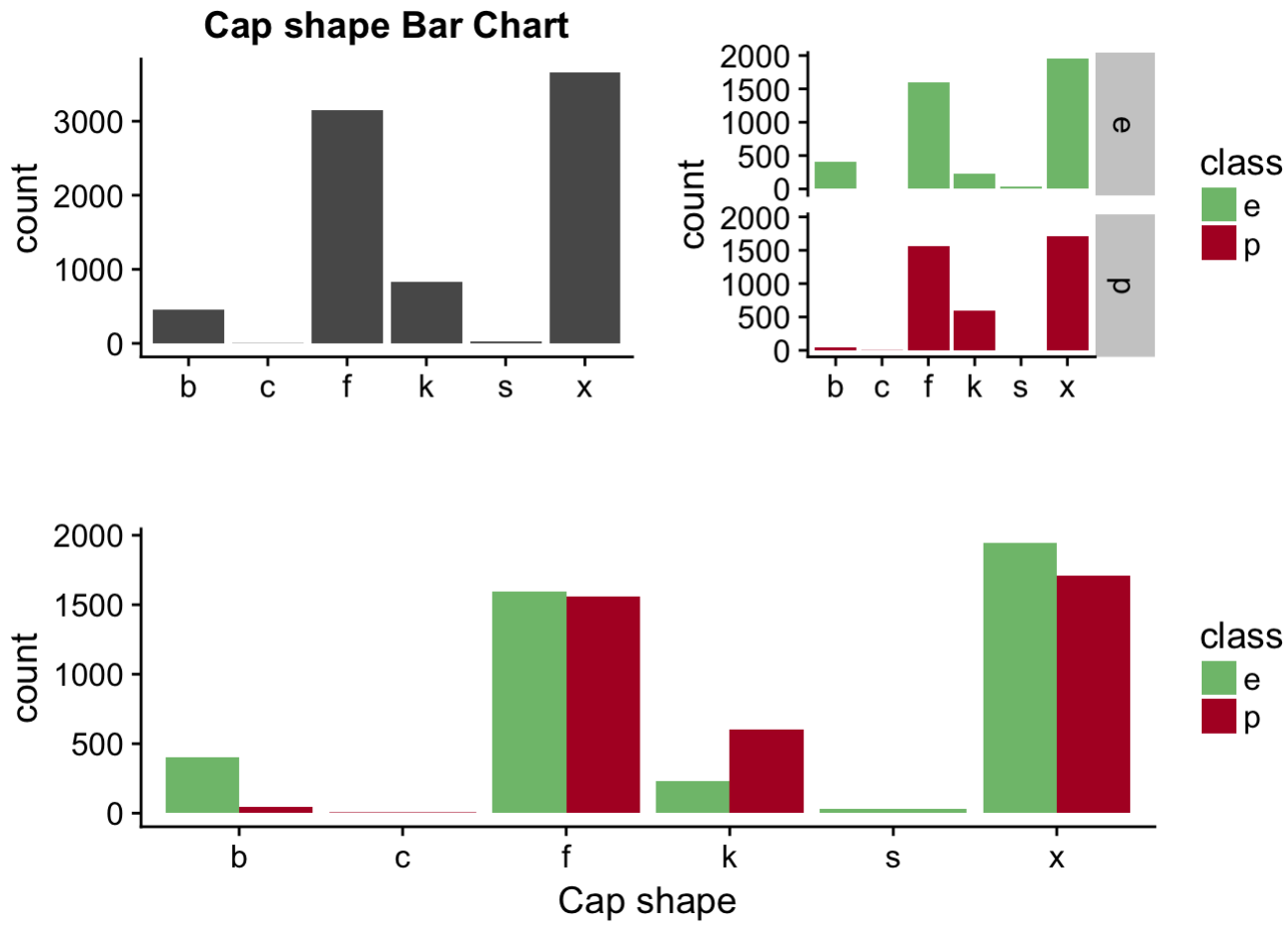
The color green is used to represent mushrooms that are edible while red is used to represent mushrooms that are poisonous.

Cap shape

The cap shape according to this data set can take any of the following descriptions and is marked by the letters alongside:

- **bell** = b
- **conical** = c
- **convex** = x
- **flat** = f
- **knobbed** = k
- **sunken** = s

These are illustrated below and compared against the target feature.



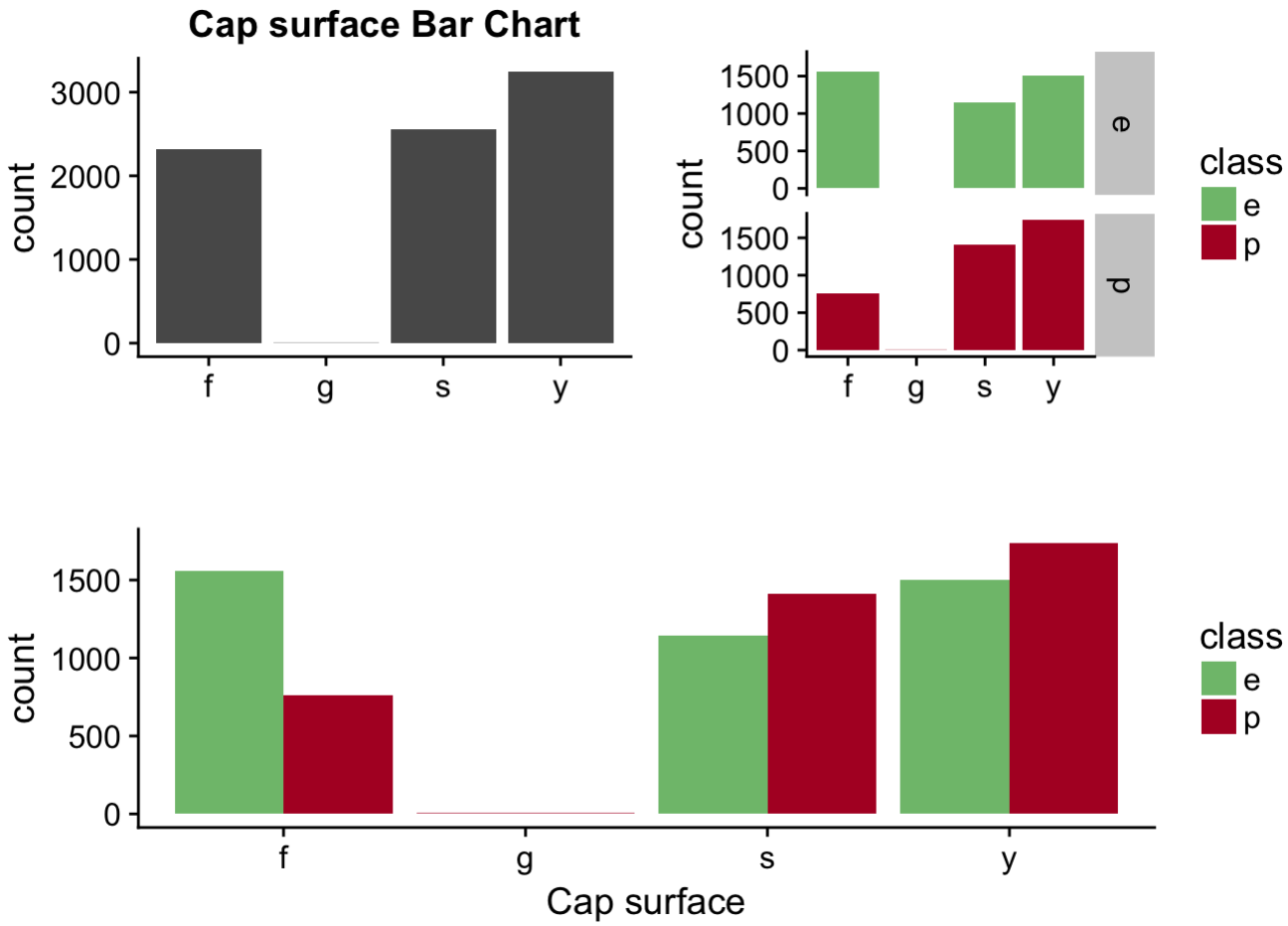
Most mushrooms appear to have a convex cap shape and there is almost an identical number of edible and poisonous ones. However, mushrooms with a bell shape are more likely to be edible. In contrast, mushrooms with a knobbed cap are more likely to be poisonous.

Comparatively, a conical shaped mushroom is almost certain to be poisonous while a sunken shaped one is almost certain to be edible. Both of these are relatively rare to find.

Cap surface

The cap surface according to this data set can take any of the following descriptions and is marked by the letters alongside:

- **fibrous** = f
- **grooves** = g
- **scaly** = y
- **smooth** = s



A mushroom with a fibrous cap surface is likely to be edible while those with a smooth or scaly surface are more likely to be poisonous. However, those with a scaly or smooth surface have an equal number of the alternative.

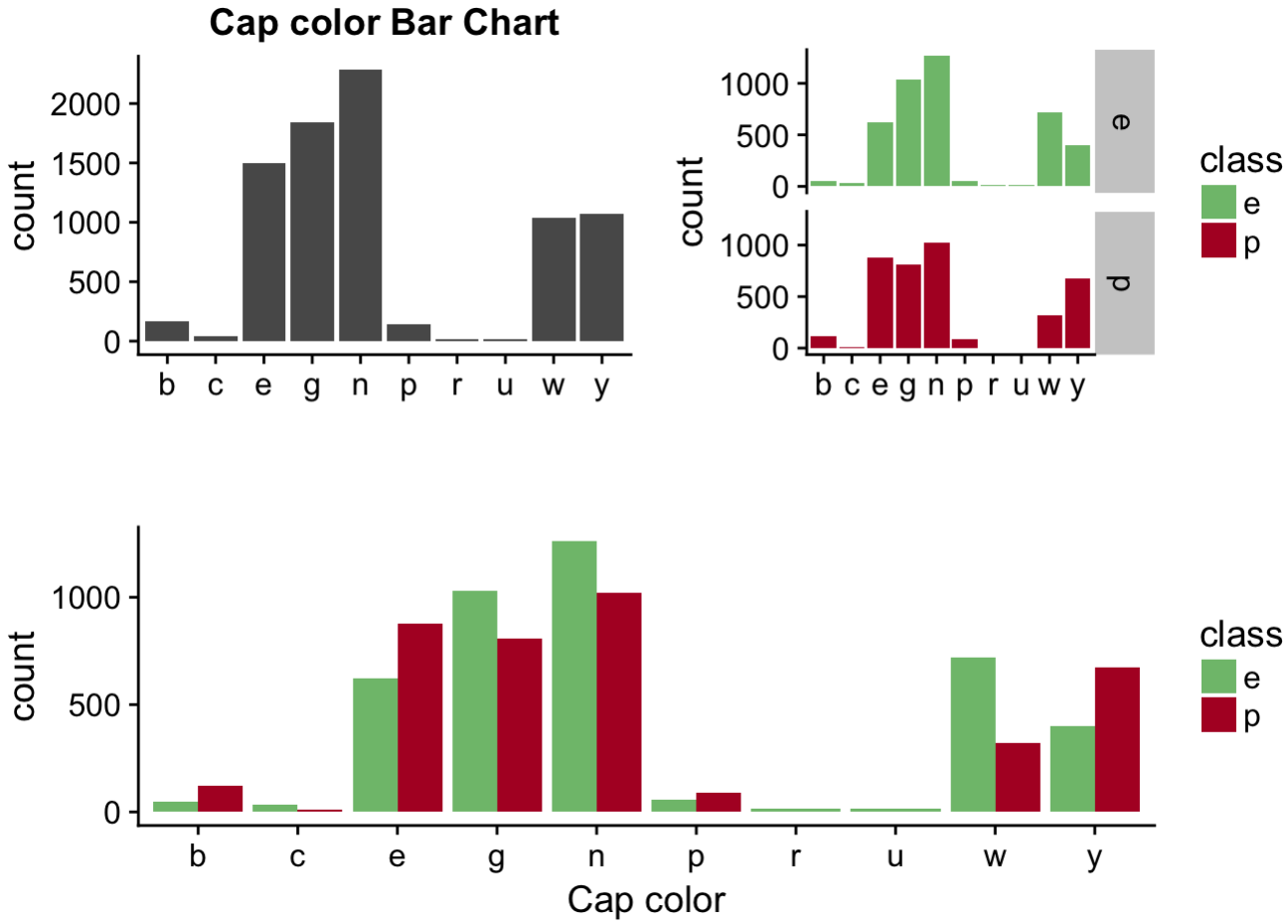
Comparatively, a mushroom with a surface with grooves is almost certain to be poisonous and is not commonly found.

Cap color

The cap color according to this data set can take any of the following descriptions and is marked by the letters alongside:

- **brown** = n
- **buff** = b
- **cinnamon** = c
- **red** = e
- **gray** = g

- **green** = r
- **pink** = p
- **purple** = u
- **white** = w
- **yellow** = y



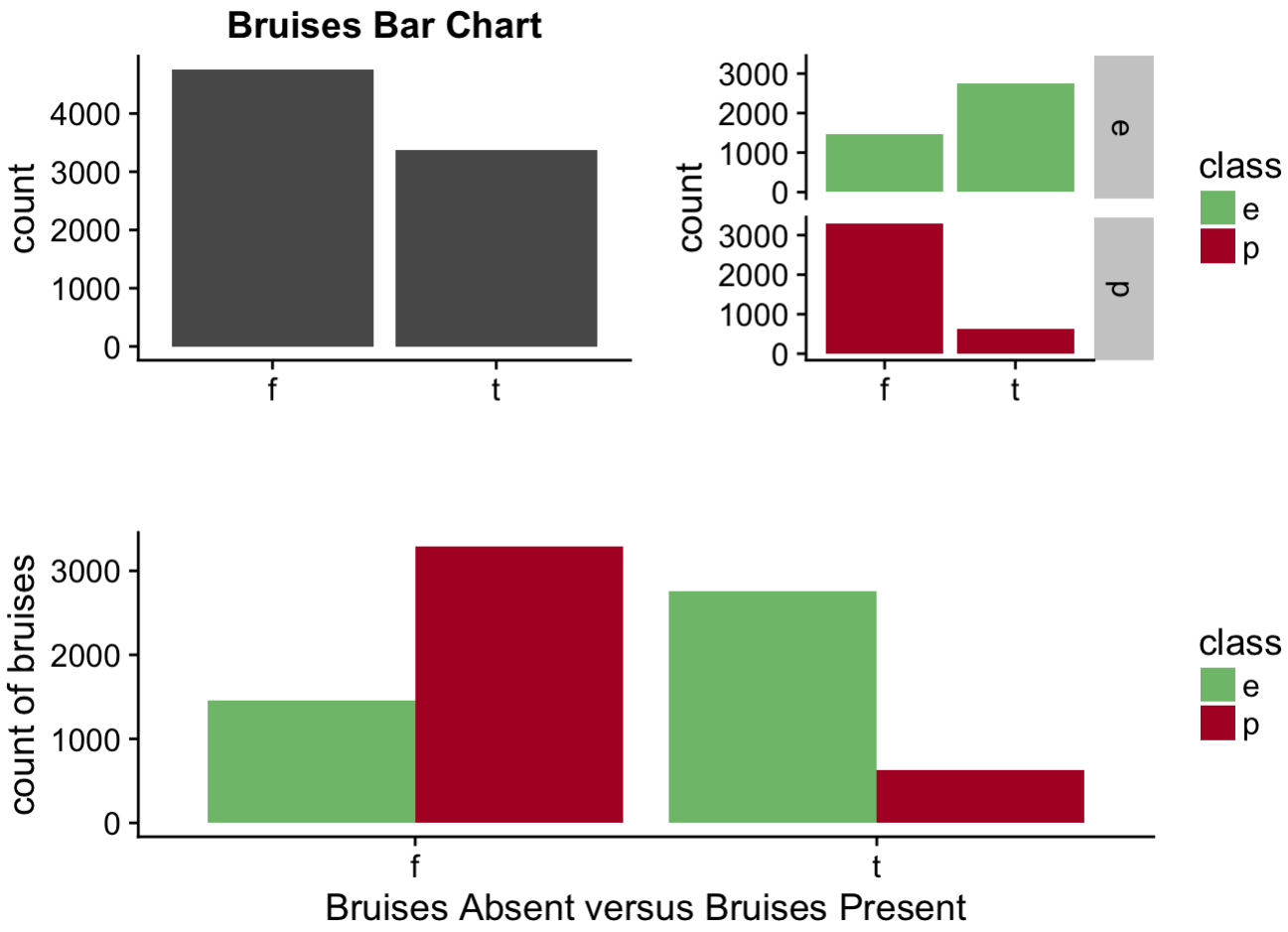
Most mushrooms appear to be brown and are more likely to be edible although there are also a good number that are poisonous. Similarly, cinnamon, gray and white mushrooms appear to be more edible than poisonous. On the other hand, buff, red, pink and yellow mushrooms are more likely to be poisonous.

However, green and purple mushrooms are rare to find and almost certain to be edible.

Bruises

A mushroom can appear to have bruises or not and this is marked by the letters alongside as shown below :

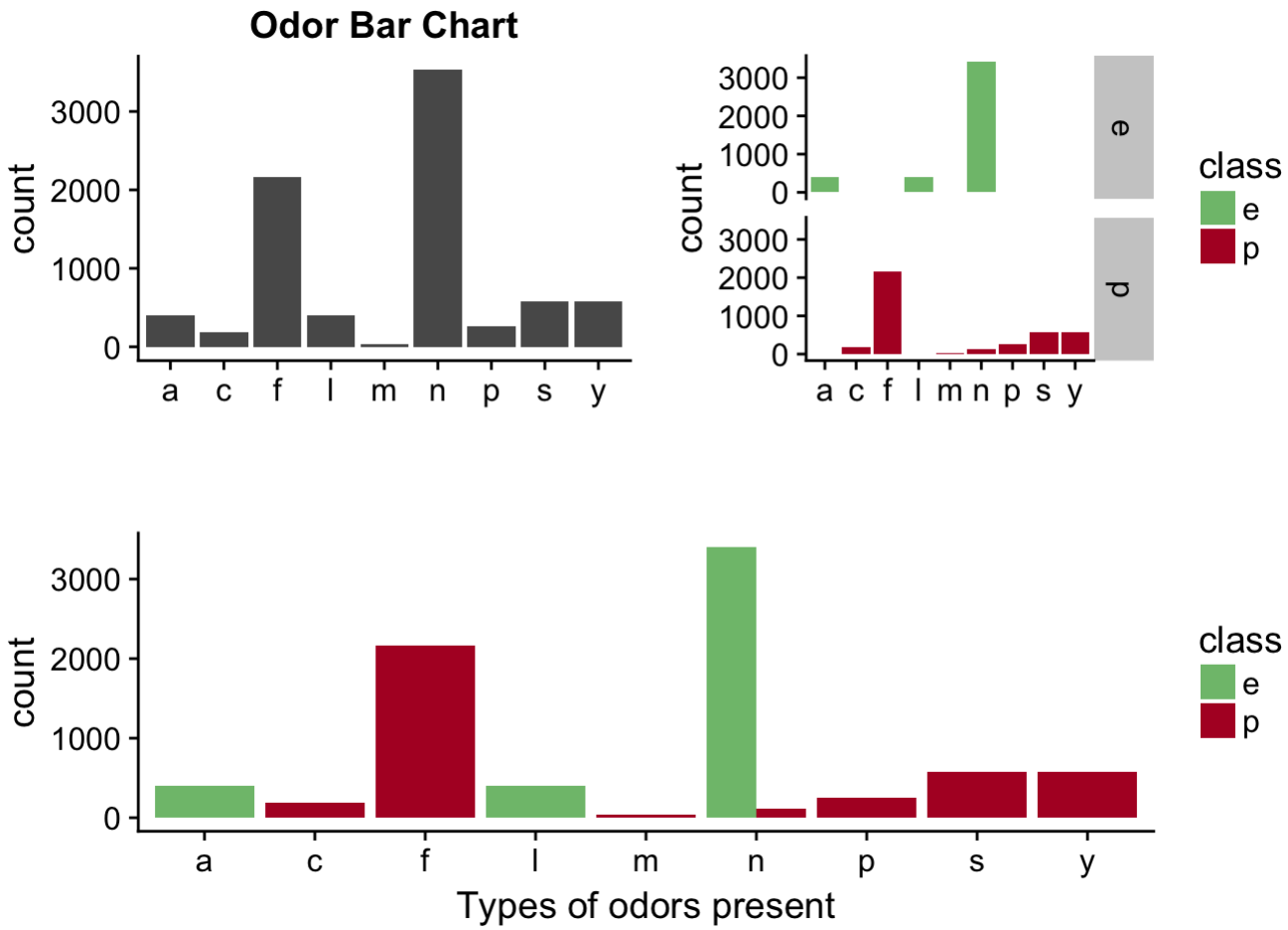
- **bruises** = t
- **no** = f



Odor

A mushroom can have certain odors and these are marked by the letters alongside as shown below :

- **almond** = a
- **anisel** = l
- **creosote** = c
- **fishy** = y
- **foul** = f
- **musty** = m
- **none** = n
- **pungent** = p
- **spicy** = s



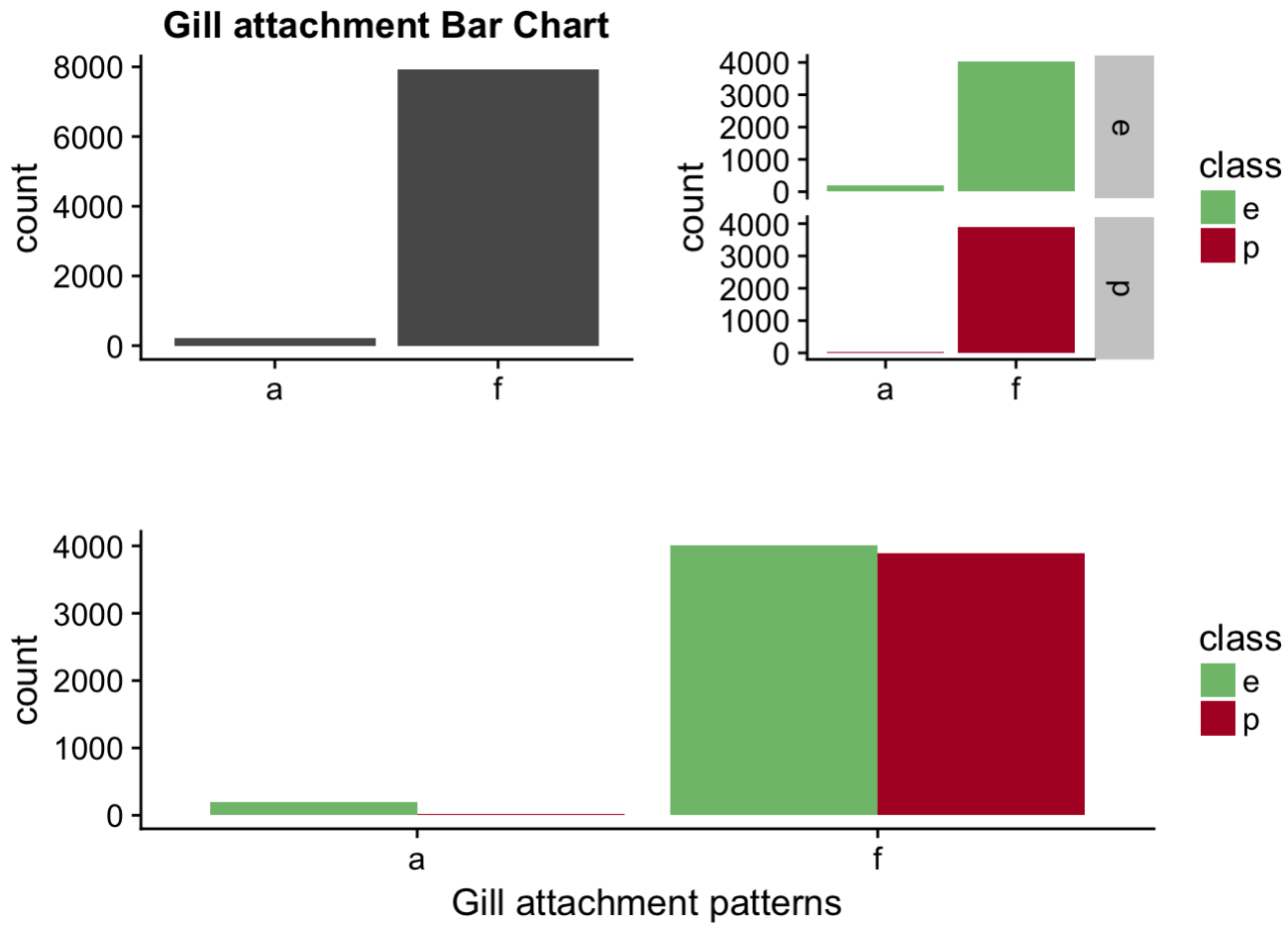
Most mushrooms do not have an odor and these are likely to be edible. In addition, mushrooms with an almond or anisel odor are almost certain to be edible. It is also fairly common to find a mushroom with a foul smell.

In contrast, mushrooms with a creosote, foul, fishy, musty, pungent or spicy smell are almost certain to be poisonous.

Gill attachment

A mushroom can have certain gill attachment patterns and these are marked by the letters alongside as shown below :

- **attached** = a
- **descending** = d
- **free** = f
- **notched** = n



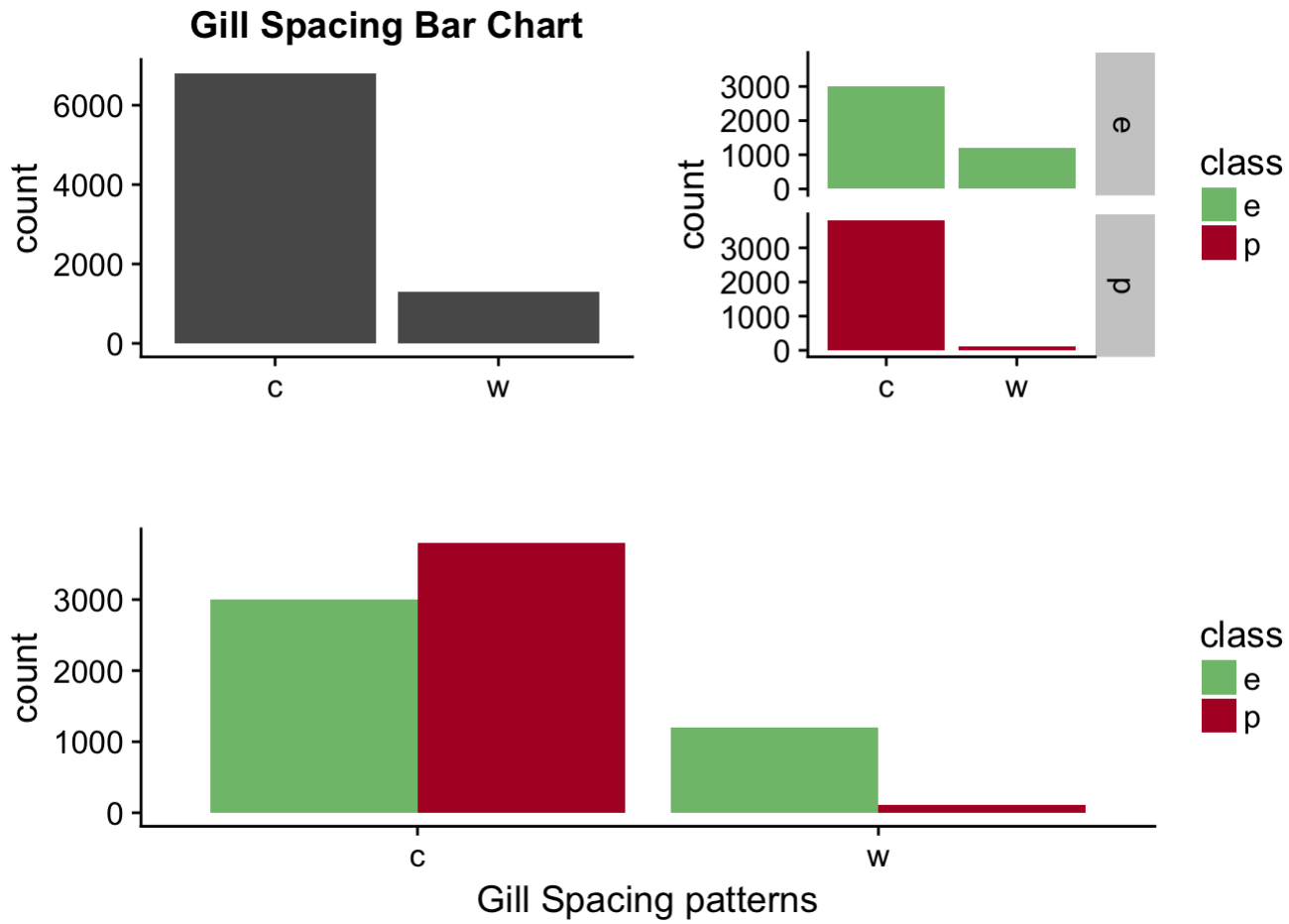
Most mushrooms appear to either have a free gill pattern or an attached gill pattern. Those with a free gill pattern are equally likely to be edible or poisonous while those with an attached gill pattern, are almost always edible.

There were no mushrooms found with a notched or descending gill pattern and this may be an indication that these are redundant levels.

Gill spacing

A mushroom can have certain gill spacing patterns and these are described below and marked by letters in the dataset indicated alongside as shown below :

- **close** = c
- **crowded** = w
- **distant** = d



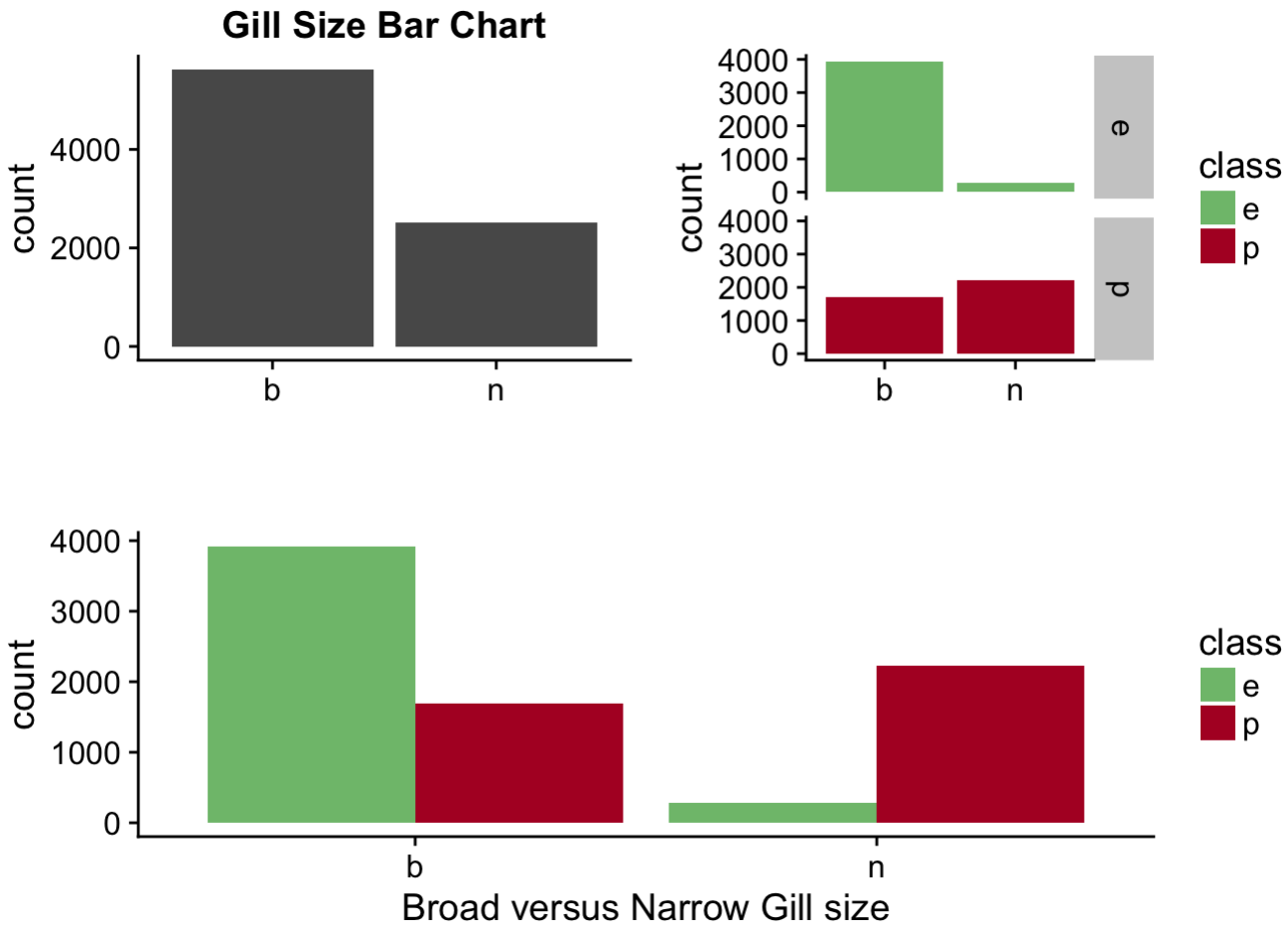
Most mushrooms either have a close or crowded gill spacing pattern. Mushrooms with a close gill spacing are more likely to be poisonous although a considerably large number are also edible. Contrastingly, those with a crowded gill pattern are almost always edible although can occasionally be poisonous.

There were no mushrooms found with a distant gill spacing and this may be an indication that this is a redundant level.

Gill size

A mushroom can have either have a broad or narrow gill size and these are described below and marked by letters in the dataset indicated alongside as shown below :

- **broad** = b
- **narrow** = n

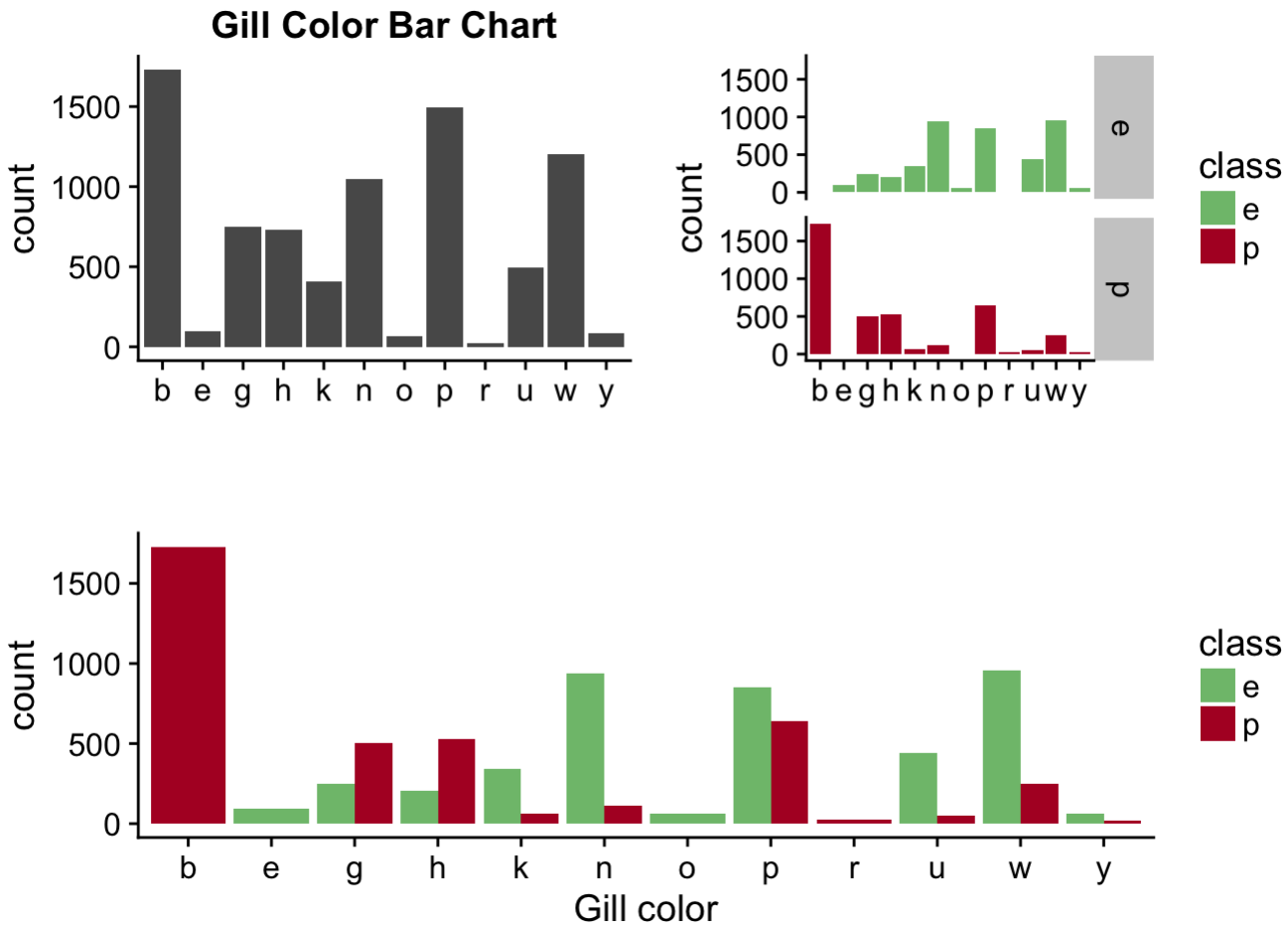


Broad gill sized mushrooms are more likely to be edible while narrow are more likely to be poisonous.

Gill color

A mushroom can have certain colors on its gill and these are described below and marked by letters in the dataset indicated alongside as shown below :

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



Most mushrooms appear to have a buff color along the gill. These mushrooms with this color are also almost always poisonous.

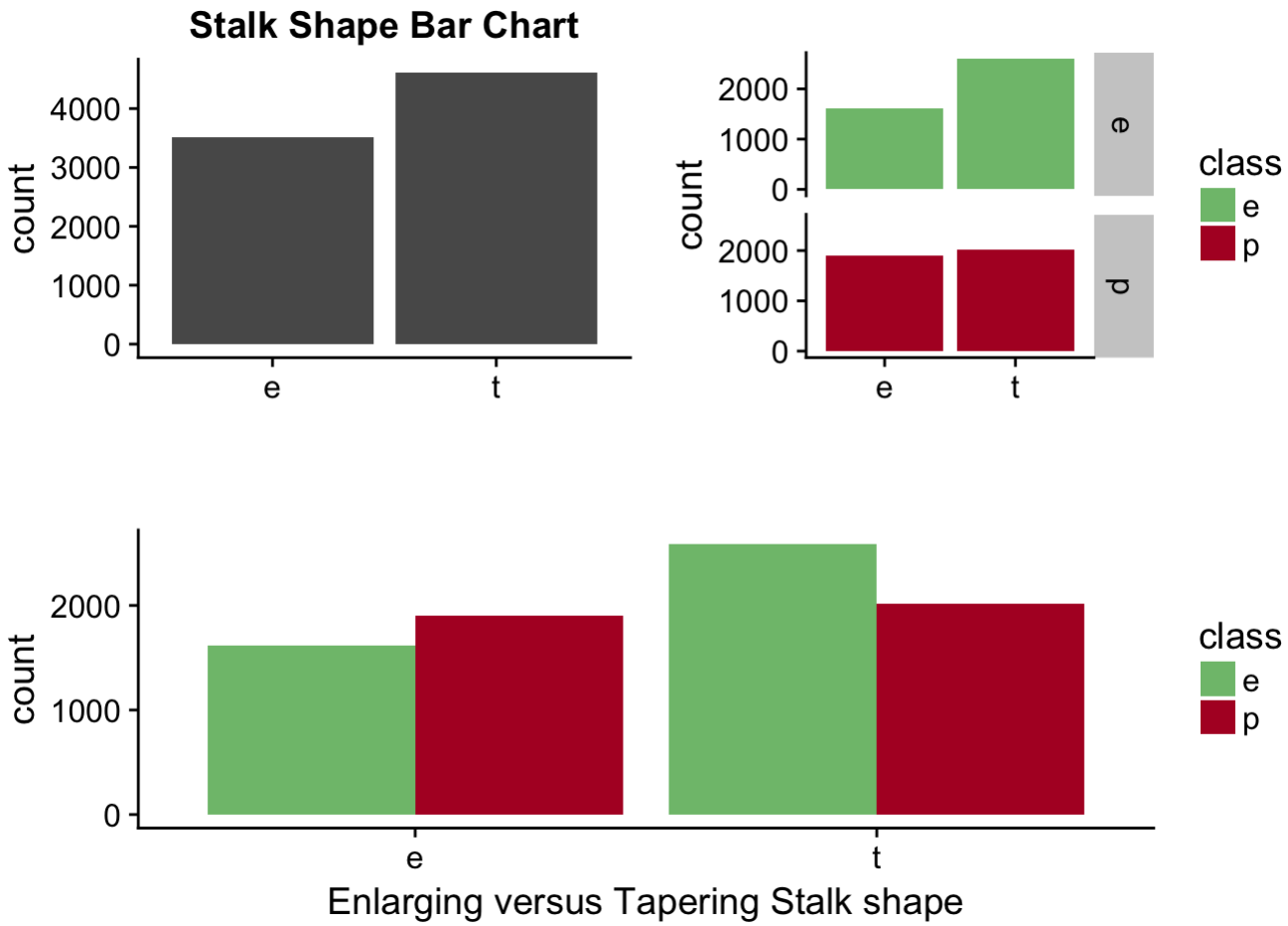
The other colors are relatively less common. Those with a brown, pink, purple, white, yellow color are more likely to be edible while those with a gray or chocolate color are more likely to be poisonous.

Orange and red colored gills certainly indicates that a mushroom is edible while the rare and green colored mushroom is a sure sign that the mushroom is posionous.

Stalk shape

A mushroom can either have a stalk shape that enlarges or tapers off. These are described below and marked by letters in the dataset indicated alongside as shown below :

- **enlarging** = e
- **tapering** = t

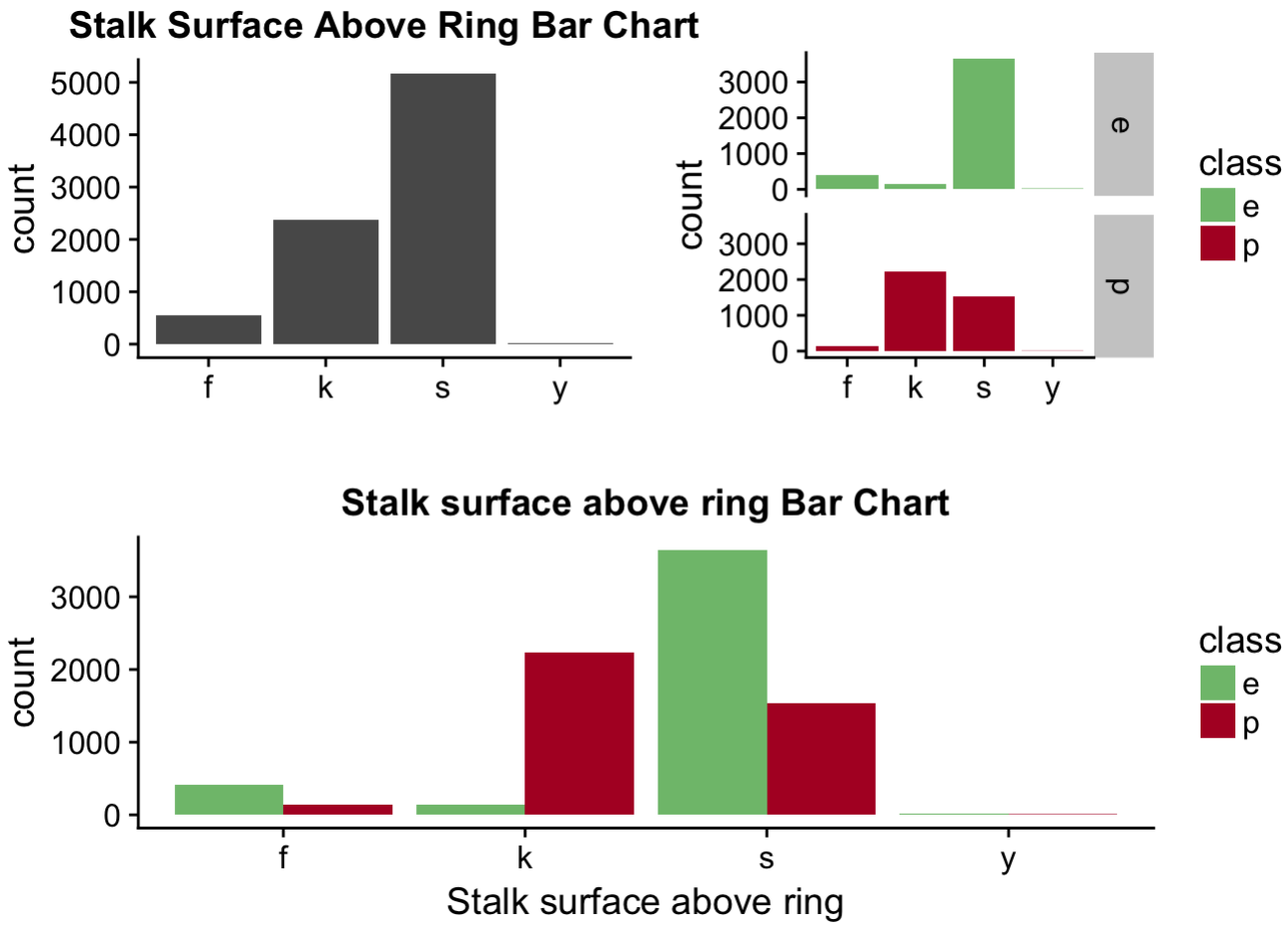


These two descriptions of the stalk shape appear to give very similar information. However, it is more likely that a mushroom with an enlarging stalk shape is poisonous while that with a tapering is edible.

Stalk surface above ring

The stalk surface above ring to this data set can take any of the following descriptions and is marked by the letters alongside:

- **fibrous** = f
- **silky** = k
- **scaly** = y
- **smooth** = s



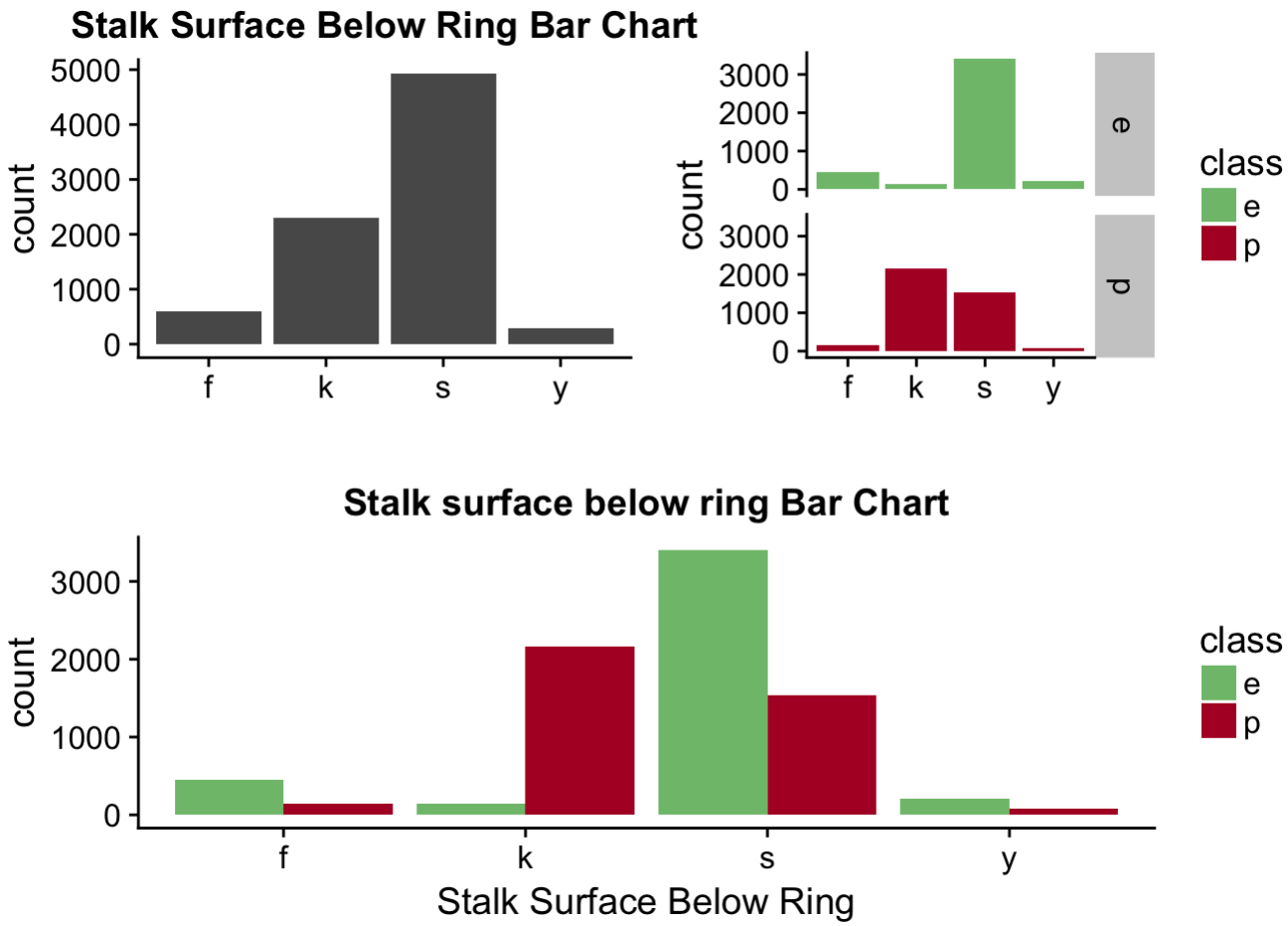
Most mushrooms seem to have a smooth stalk surface above its ring. These smooth stalked mushrooms are also more likely to be edible. It is also fairly common to observe a mushroom with a silky stalk surface above its ring although in contrast, it is more likely to be poisonous.

It is rare to observe mushrooms with a fibrous stalk surface above its ring although they are more likely to be edible.

In addition, it is extremely rare to find mushrooms with a scaly stalk above its ring and this may be an indication that this is a redundant level.

Stalk surface below ring

The stalk surface below the ring has a similar level description to the stalk surface above the ring because these attributes are only differentiated by their position. The descriptions are similar and hence not shown below.



There is an almost identical representation between this attribute(stalk surface below the ring) and the previous attribute(stalk surface above the ring).

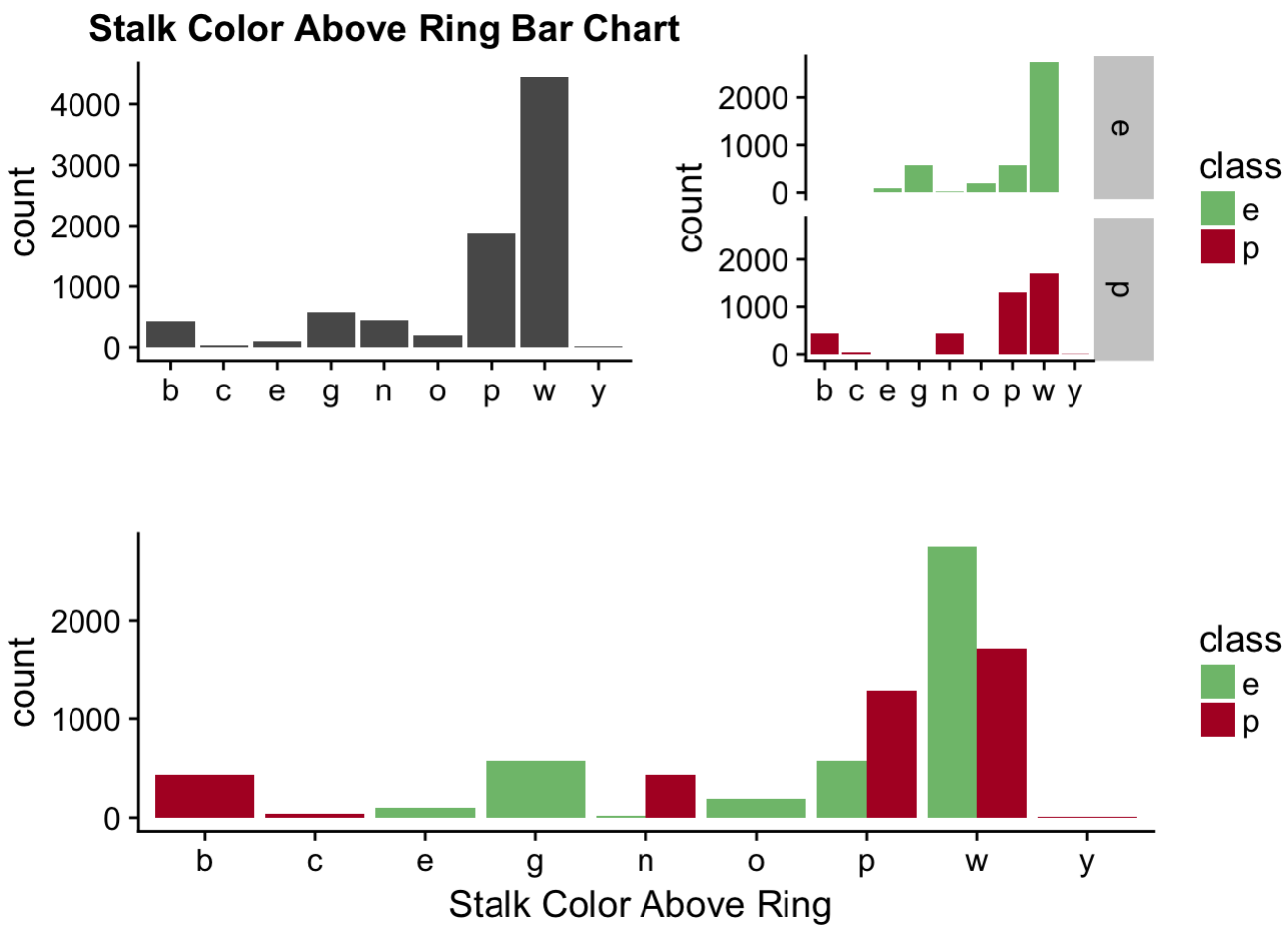
The inference is therefore that most mushrooms have the same surface above and below the stalk. The only exception is that there is an increase of mushrooms with a scaly stalk surface below the ring.

Both of these would not be useful for the model as they present the same information and I would merge the two creating a new attribute to reflect this.

Stalk color above ring

The stalk color above ring according to this data set can take any of the following descriptions and is marked by the letters alongside as shown below:

- **brown** = n
- **buff** = b
- **cinnamon** = c
- **red** = e
- **gray** = g
- **orange** = o
- **pink** = p
- **white** = w
- **yellow** = y



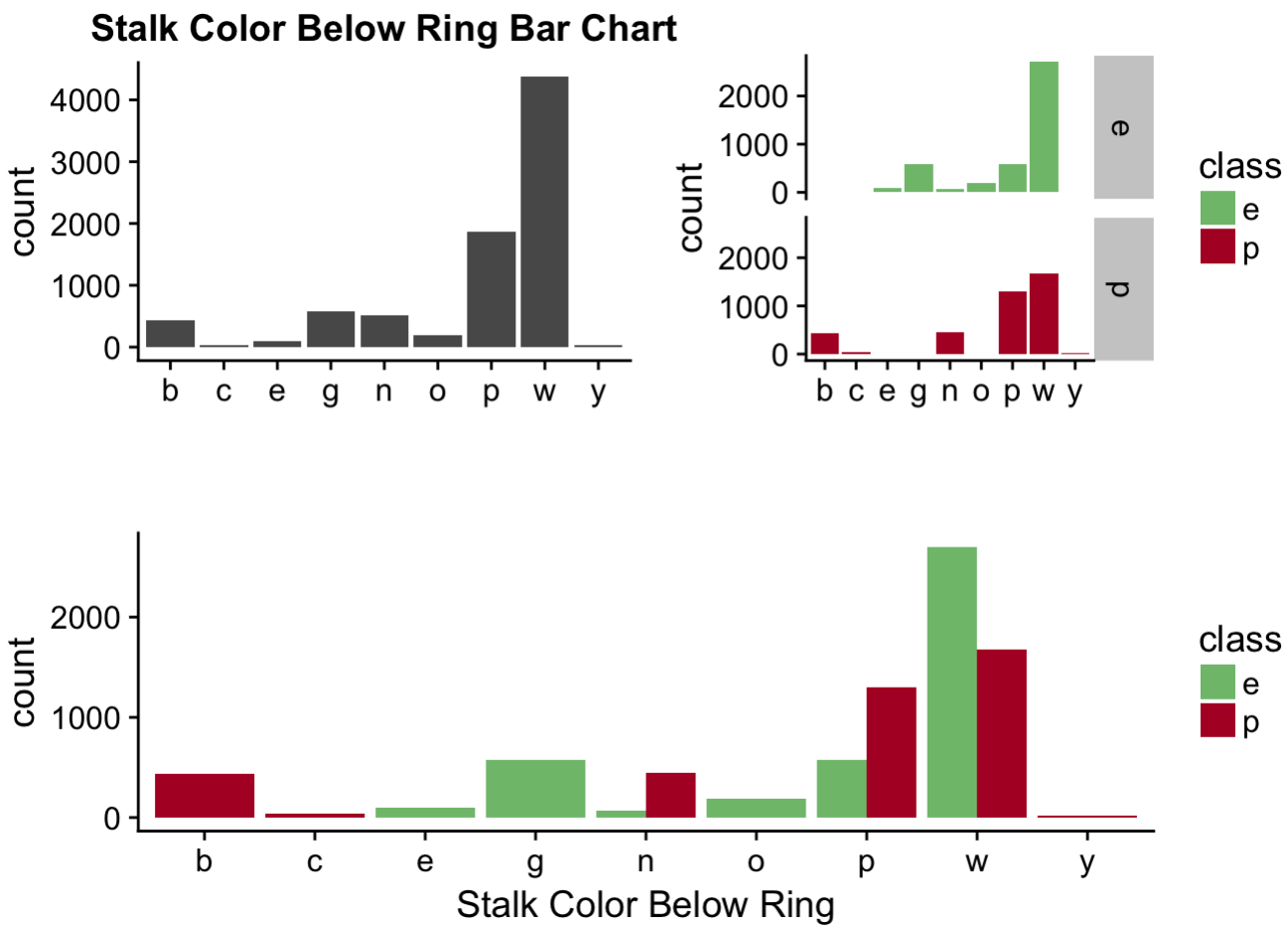
Most mushrooms have a white stalk above the ring and these mushrooms are more likely to be edible. It is fairly common to spot a mushroom with a pink colored stalk below the ring although it is likely to be poisonous.

Mushrooms with a buff or brown colored stalk below the ring are almost always certain to be poisonous. Although rare, cinnamon colored stalks below the ring also almost always poisonous.

In contrast, red, gray or orange colors on the stalk below the ring are a sure sign that the mushroom is edible.

Stalk color below ring

The stalk color below the ring has a similar level description to the stalk color above the ring because these attributes are only differentiated by their position. The descriptions are similar and hence not shown below.



There is an almost identical representation between this attribute(stalk color below the ring) and the previous attribute(stalk color above the ring).

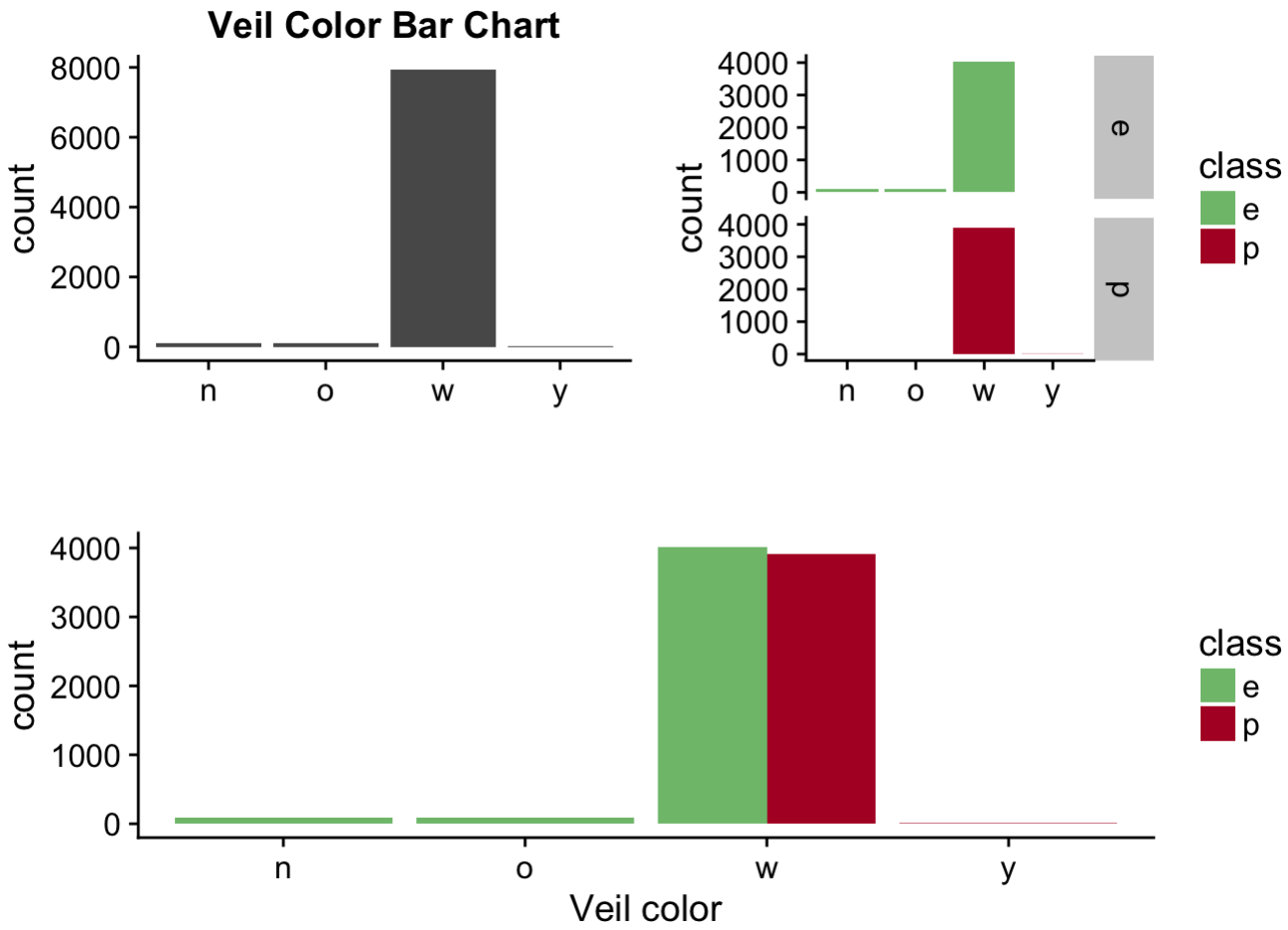
The inference is therefore that most mushrooms have the same color above and below the stalk. There is a slight increase in gray colored stalks below the ring.

Similarly and for reasons stated previously, I would merge these two features and create a new attribute to reflect this.

Veil color

The veil color according to this data set can take any of the following descriptions and is marked by the letters alongside as shown below:

- **brown** = n
- **orange** = o
- **white** = w
- **yellow** = y



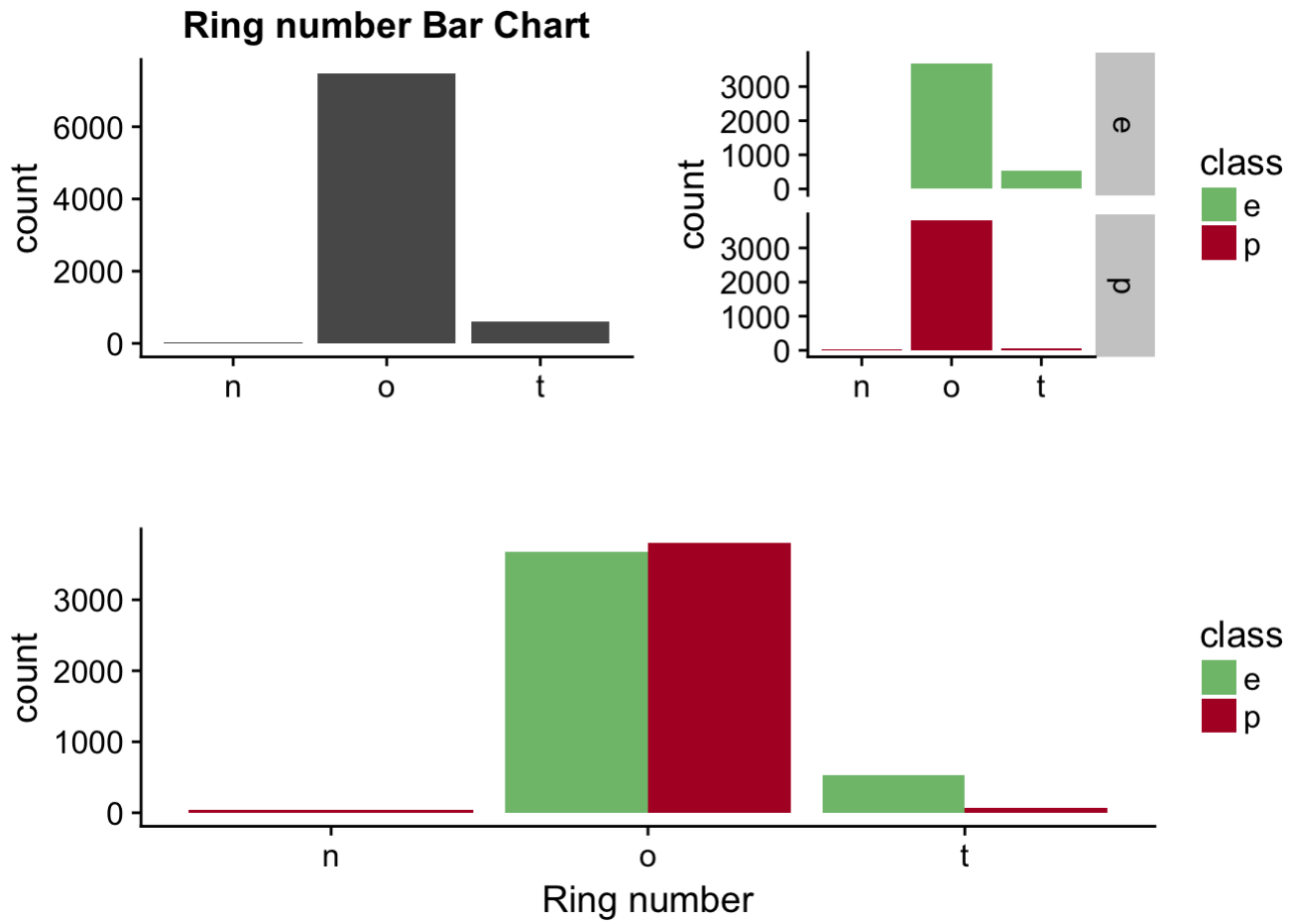
There is an almost equal number of edible and poisonous white veil mushrooms.

However, most mushrooms with a brown or orange colored veil are almost certain to be edible while those that are yellow are almost surely posionous.

Ring number

The ring number according to this data set can take any of the following values and is marked by the letters alongside as shown below:

- **none** = n
- **one** = o
- **two** = w

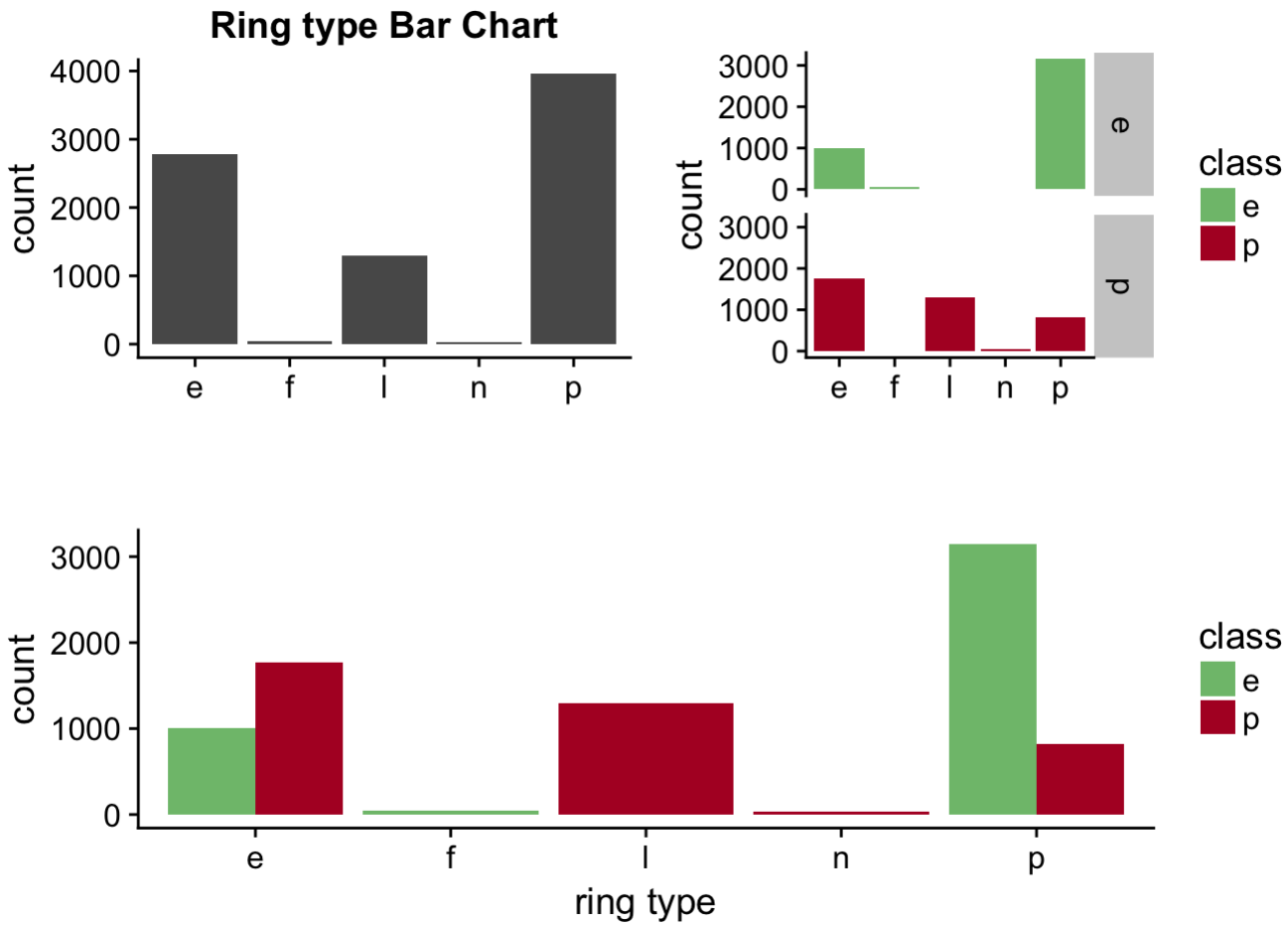


Majority of the mushrooms have one ring although these are equally likely to be edible as they are posionous. This is an indication that this is not a very descriptive level. In contrast, mushrooms with no rings are almost certain to be poisonous while those with two rings are more likely to be edible than poisonous.

Ring type

The ring type according to this data set can be described in any of the following ways and is marked by the letters alongside as shown below:

- **cobwebby** = c
- **evanescent** = e
- **flaring** = f
- **large** = l
- **none** = n
- **pendant** = p
- **sheathing** = s
- **zone** = z



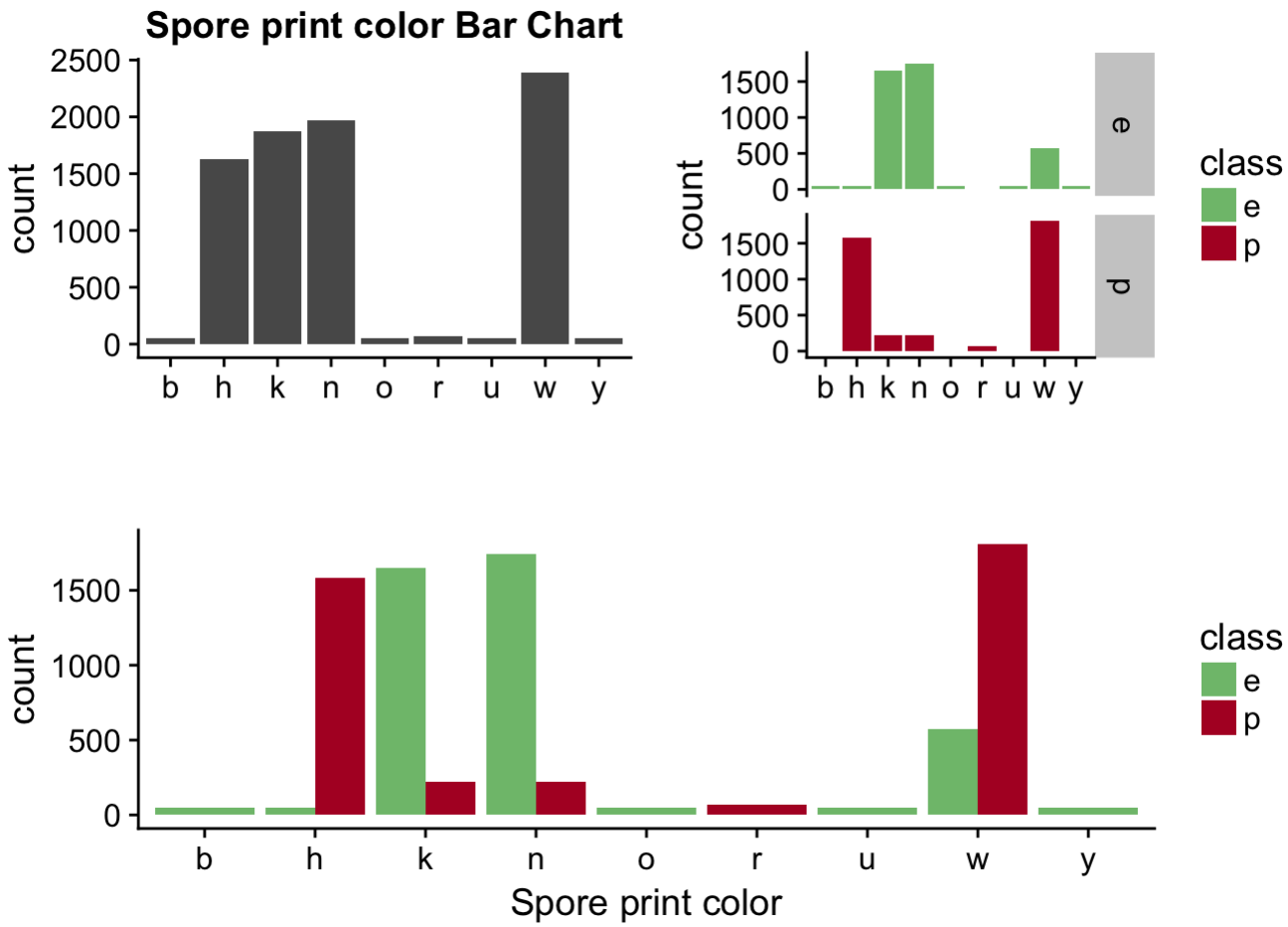
Pendant is the most common ring type and most of these mushrooms in this category are edible. Mushrooms with a flaring ring, although rare, type are certain to be edible. In contrast, those with no rings, similarly rare, are certain to be poisonous.

Mushrooms with a large ring are fairly common and certain to be poisonous.Those with an evanescent ring are more likely to be poisonous although can also occassionally be edible.

Spore print color

A mushroom can have certain colors on its spore print and these are described below and marked by letters in the dataset indicated alongside as shown below :

- **black** = k
- **brown** = n
- **buff** = b
- **chocolate** = h
- **green** = r
- **orange** = o
- **purple** = u
- **white** = w
- **yellow** = y



The most common colors to be observed in this category are white, brown, black and chocolate. White and chocolate colored spores are more likely to be poisonous while black and brown colored spres are more likely to be eligible.

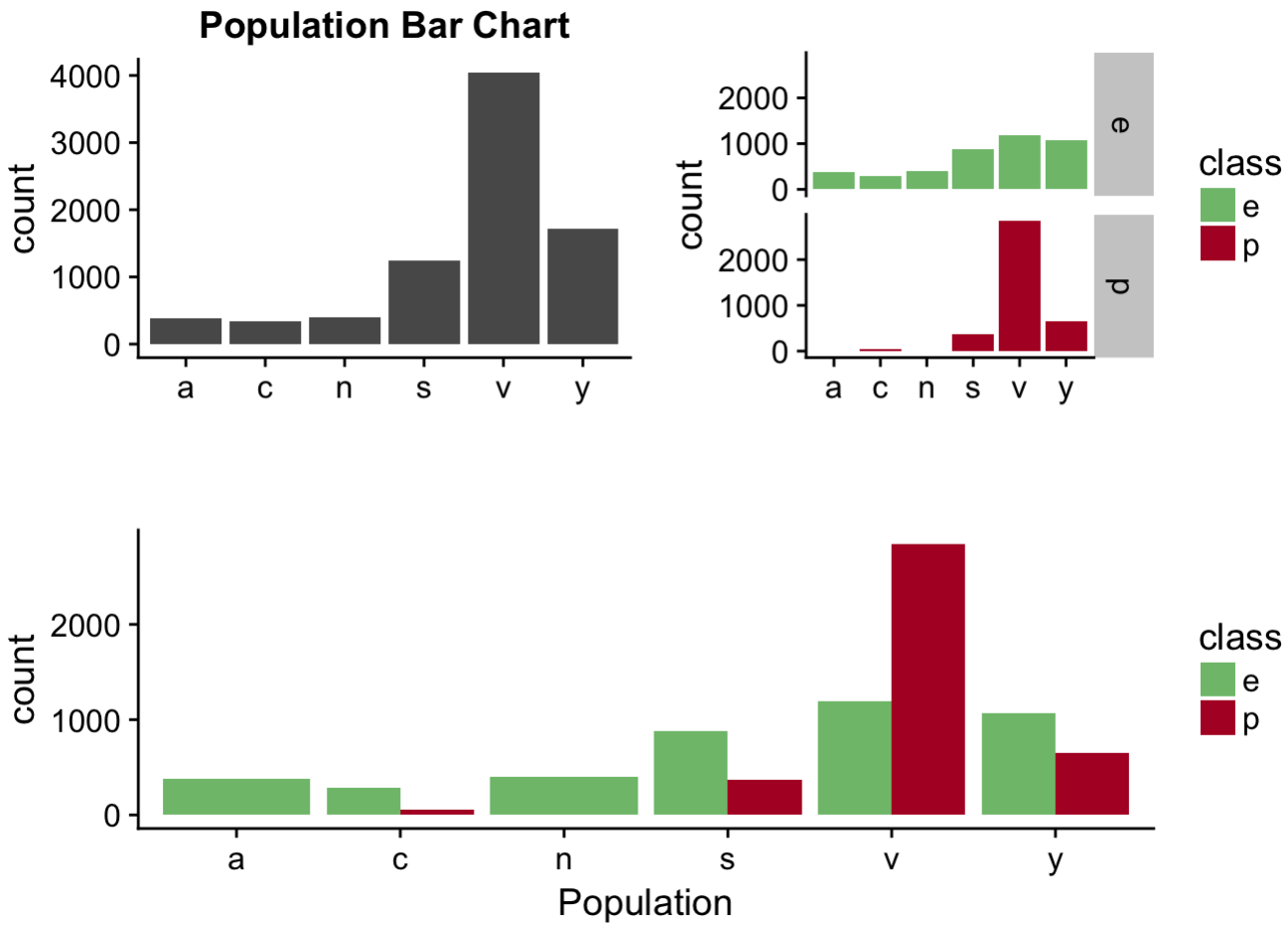
Buff, orange, purple and yellow colored spores are rare but always certain to be an indication that a mushroom is edible.

On the other hand, green colored spores are similarly rare but a certain indication that a mushroom is poisonous.

Population

The population a mushroom can be found in according to this data set can be described in any of the following ways and is marked by the letters alongside as shown below:

- **abundant** = a
- **clustered** = c
- **numerous** = n
- **scattered** = s
- **several** = v
- **solitary** = y



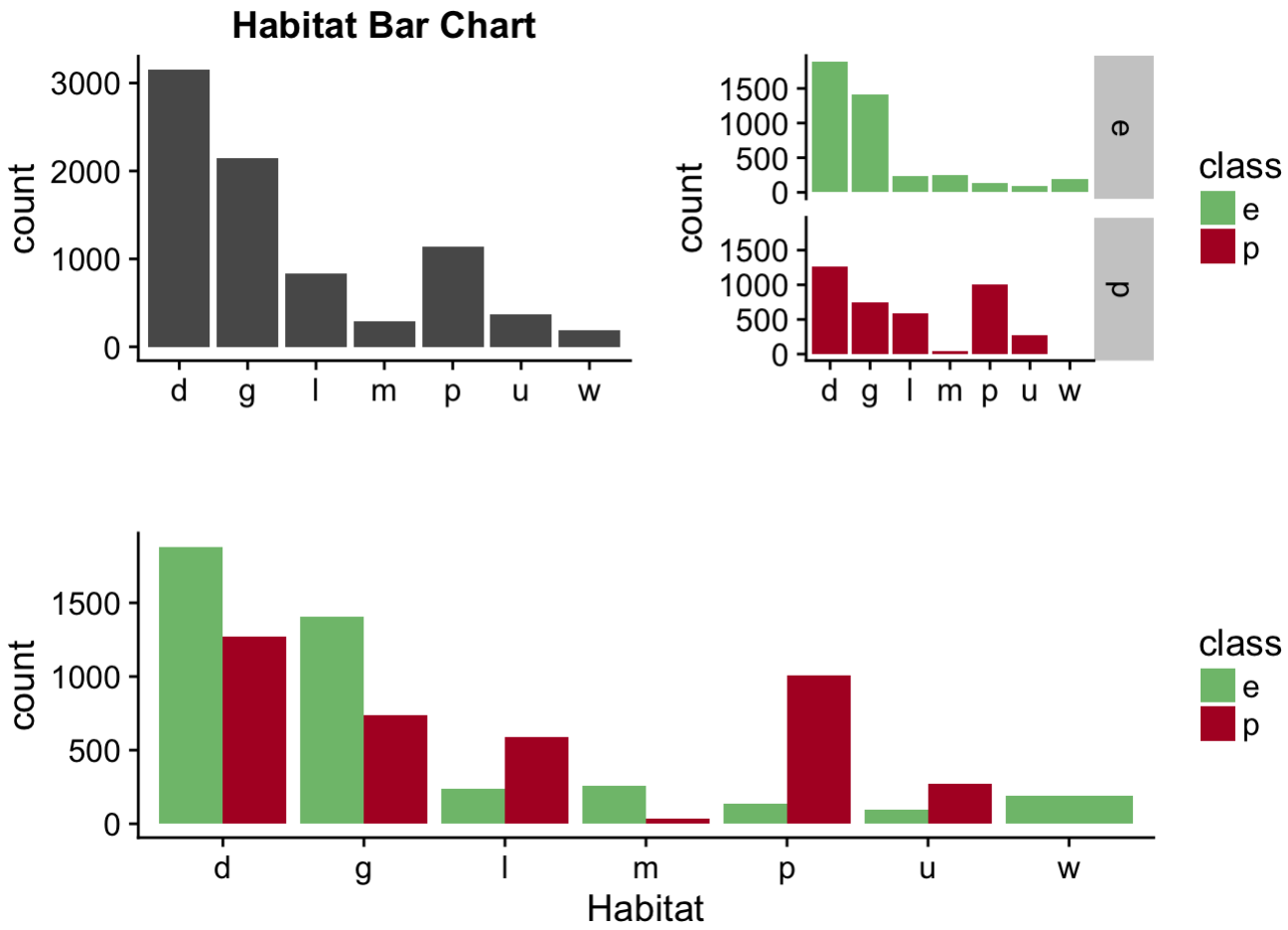
Most mushrooms are found in several populations and by virtue of this category are more likely to be poisonous. It is also common to find mushrooms in clustered,scattered and solitary populations and in all of these categories are more likely to be edible than poisonous.

Mushrooms appearing in abundant or numerous populations are certain to be edible.

Habitat

The habitat a mushroom can be found in according to this data set can be described in any of the following ways and is marked by the letters alongside as shown below:

- **grasses** = g
- **leaves** = l
- **meadows** = m
- **paths** = p
- **urban** = u
- **waste** = w
- **woods** = d



Most mushrooms can be found in woods and grasses and in both categories, are more likely to be edible. It is common to find mushrooms along paths and they are more likely to be poisonous.

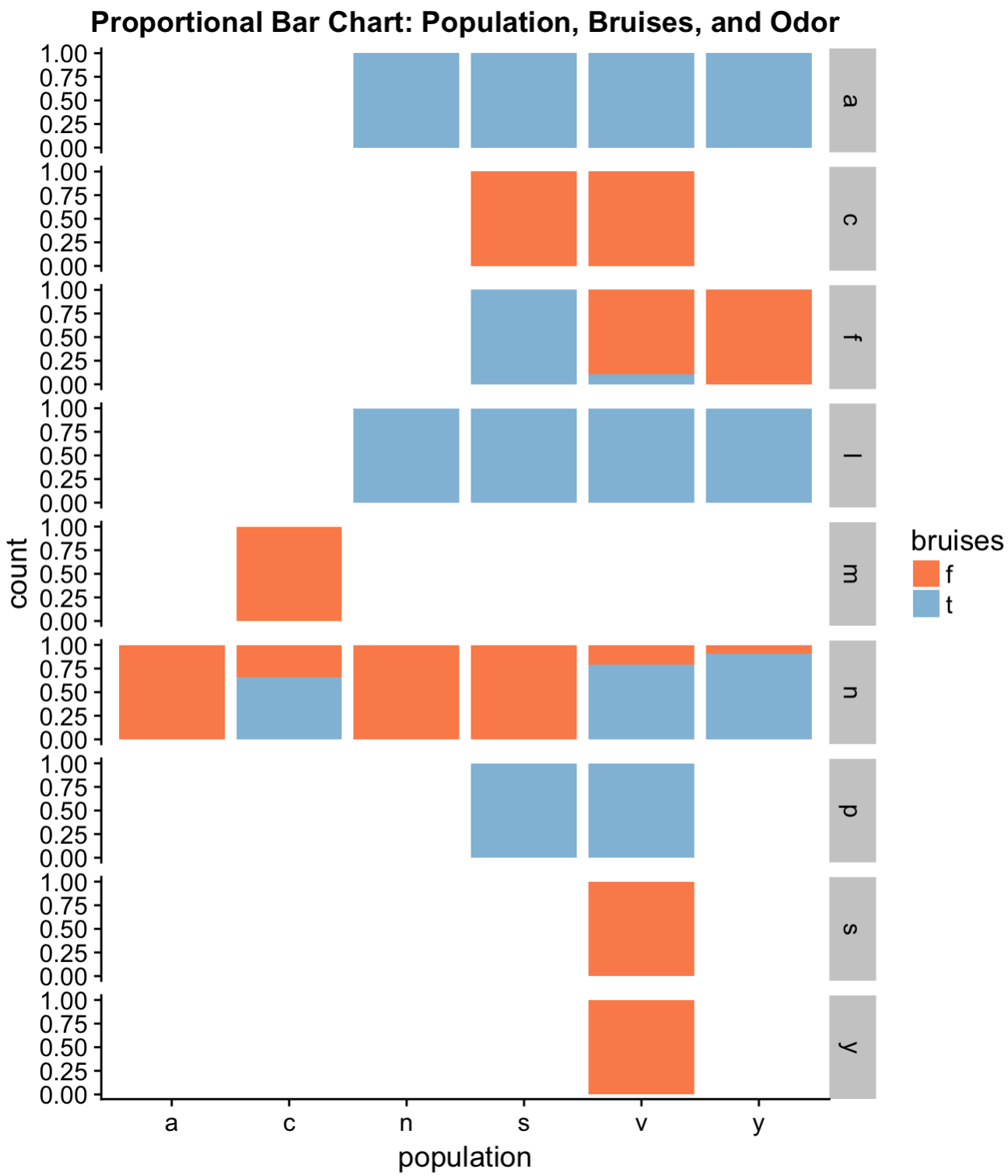
It is relatively to find mushrooms located in leaves and in urban areas and in both cases if existent, are likely to be poisonous. Similarly, it is difficult to spot mushrooms in meadows, although they are likely to be edible.

Mushrooms located in waste are rare but certain to edible.

4.1.2 Multivariate Visualisations

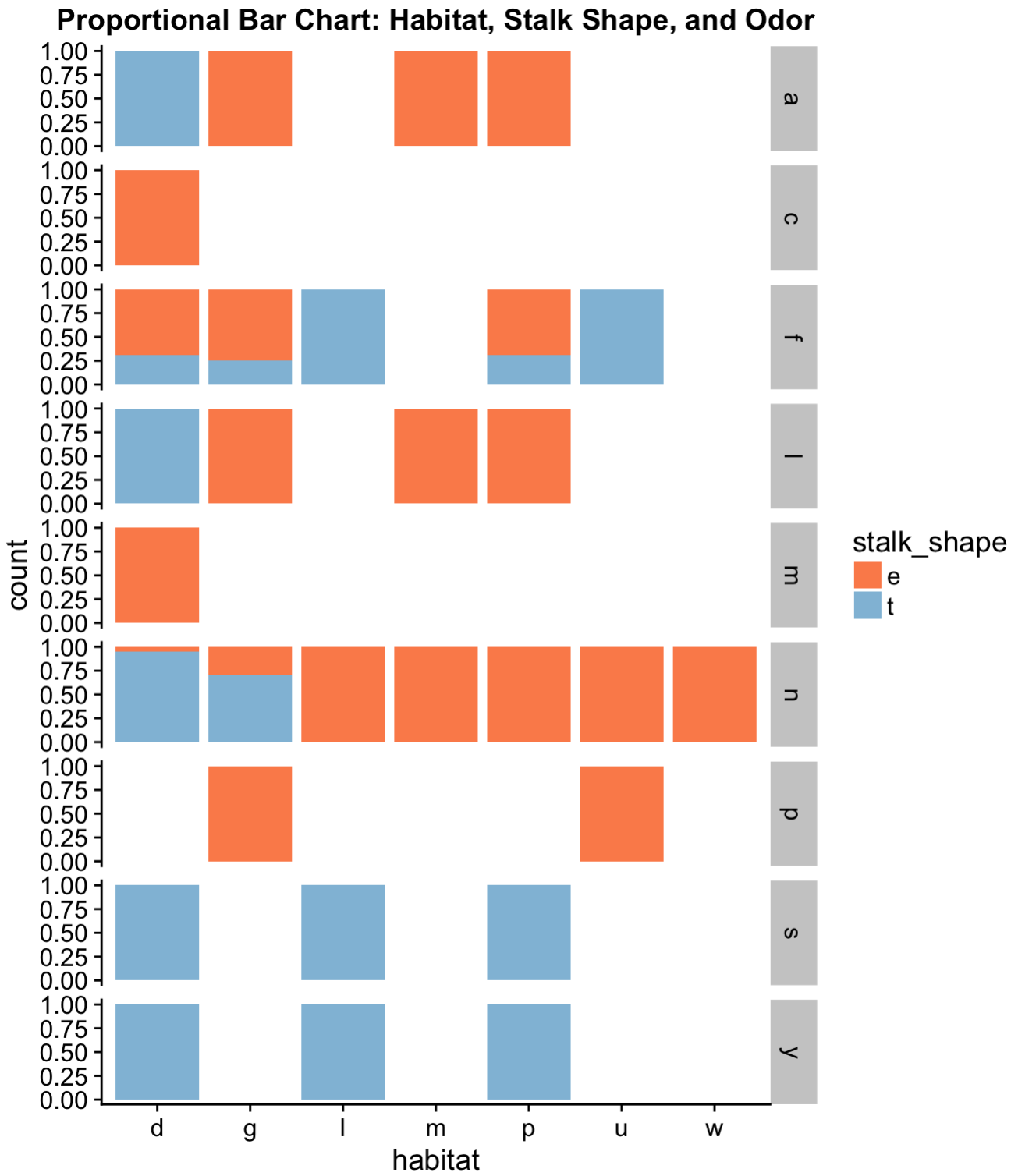
Population, Bruises, and Odor

The following visualisation depicts that overall, most mushrooms do not have an odor, are located in all the various kinds of populations and more often than not do not have bruises. Mushrooms with an almond or anisel odor are most likely to be bruised and equally appear in solitary, several, scattered or numerous populations In addition, mushrooms with a pungent odor are most likely bruised and appearing in several or scattered populations.



Habitat, Stalk Shape, and Odor

This proportional bar chart confirms once again, that mushrooms with no odor remain the most represented category, are widely distributed in all the habitats and this time are more likely to have an enlarged stalk shape. The woods have the most mushrooms and there is an almost equal mix of stalk types and the presence of almost every odor with the exception of pungent. The waste habitat as expected, bears mushrooms with a pungent odor and with an enlarged stalk.



Summary

This dataset only had categorical features. We omitted **stalk_root** and **veil_type** but other than that, we did not remove the original features of the dataset. From the exploration through visualisations, we see that the remaining features could potentially be useful in prediciting whether a mushroom was edible or poisonous.