A05_naivebayes

Verena Haunschmid

25 March 2016

Load library

The naive bayes algorithm is included in the library `e1071`. For visualisation I use `ggplot2`.

```
library(e1071)
library(ggplot2)
```

Data

The data was downloaded directly in R (most read methods in R do not only take file paths but also URLs). The second line sets a seed such that everytime the code is executed the test samples selected in the third line are the same. Approximately 10% of the samples are chosen as test samples.

```
load("data.RData")
```

```
mush_data<-read.csv("https://archive.ics.uci.edu/ml/machine-learning-databases/mushroom/agaricus-lepiota
set.seed(5678)
mush_test<-sample(1:nrow(mush_data), nrow(mush_data)*0.1)
mush_data_melt <- melt(mush_data, id.vars = "edible")
```

To chose test samples also the following code can be used. It randomly generates `n` numbers between 0 and 1 and each sample with a value below 0.1 is assigned to the test set.

```
which(runif(nrow(mush_data)) <= 0.1)
```

Sampling is important because the data might have been collected in a specified order.

Before applying any algorithm it's suggested to look into the data. Use `summary` and `head` to get some info about the data set.

```
summary(mush_data)
```

```
##  edible   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
```

```
head(mush_data)
```

```
##   edible cap-shape cap-surface cap-color bruises odor gill-attachment
## 1      p         x           s         n       t    p               f
## 2      e         x           s         y       t    a               f
## 3      e         b           s         w       t    l               f
## 4      p         x           y         w       t    p               f
## 5      e         x           s         g       f    n               f
## 6      e         x           y         y       t    a               f
##   gill-spacing gill-size gill-color stalk-shape stalk-root
## 1            c         n          k           e          e
## 2            c         b          k           e          c
## 3            c         b          n           e          c
## 4            c         n          n           e          e
## 5            w         b          k           t          e
## 6            c         b          n           e          c
##   stalk-surface-above-ring stalk-surface-below-ring stalk-color-above-ring
## 1                        s                        s                       w
## 2                        s                        s                       w
## 3                        s                        s                       w
## 4                        s                        s                       w
## 5                        s                        s                       w
## 6                        s                        s                       w
##   stalk-color-below-ring veil-type veil-color ring-number ring-type
## 1                      w         p          w           o         p
```

```
## 2                     w         p          w           o         p
## 3                     w         p          w           o         p
## 4                     w         p          w           o         p
## 5                     w         p          w           o         e
## 6                     w         p          w           o         p
##   spore-print-color population habitat
## 1                 k          s       u
## 2                 n          n       g
## 3                 n          n       m
## 4                 k          s       u
## 5                 n          a       g
## 6                 k          n       g
```

All values are categorical which is different from many datasets and makes it interesting for analysis. In the last section you can see plots for each feature.

Naive bayes on all features

```
training_features <- names(mush_data)[-1]
target <- "edible"

nb_mush <- naiveBayes(mush_data[-mush_test, training_features], mush_data[-mush_test, target])
pred_mush_test <- predict(nb_mush, mush_data[mush_test, training_features])
pred_mush_train <- predict(nb_mush, mush_data[-mush_test, training_features])
tab1 <- table(pred_mush_test, mush_data[mush_test, target])
tab2 <- table(pred_mush_train, mush_data[-mush_test, target])
```

Confusion matrix test data

```
kable(tab1, col.names = c("e - truth", "p - truth"))
```

	e - truth	p - truth
e	404	49
p	1	358

Accuracy

```
(tab1[1,1]+tab1[2,2])/sum(tab1)
```

```
## [1] 0.9384236
```

Confusion matrix training data

```
kable(tab2, col.names = c("e - truth", "p - truth"))
```

	e - truth	p - truth
e	3777	404
p	26	3105

Accuracy

```
(tab2[1,1]+tab2[2,2])/sum(tab2)
```

```
## [1] 0.9411926
```

Naive bayes on selected features (see below)

```
training_features <- c("bruises", "odor", "gill-spacing", "gill-size", "gill-color", "ring-type", "spore
nb_mush <- naiveBayes(mush_data[-mush_test, training_features], mush_data[-mush_test, target])
pred_mush_test <- predict(nb_mush, mush_data[mush_test, training_features])
pred_mush_train <- predict(nb_mush, mush_data[-mush_test, training_features])
tab1 <- table(pred_mush_test, mush_data[mush_test, target])
tab2 <- table(pred_mush_train, mush_data[-mush_test, target])
```

Confusion matrix test data

```
kable(table(pred_mush_test, mush_data[mush_test, target]), col.names = c("e - truth", "p - truth"))
```

	e - truth	p - truth
e	402	23
p	3	384

Accuracy

```
(tab1[1,1]+tab1[2,2])/sum(tab1)
```

```
## [1] 0.9679803
```

Confusion matrix training data

```
kable(table(pred_mush_train, mush_data[-mush_test, target]), col.names = c("e - truth", "p - truth"))
```

	e - truth	p - truth
e	3776	197
p	27	3312

Accuracy

```
(tab2[1,1]+tab2[2,2])/sum(tab2)
```

```
## [1] 0.9693654
```

Naive bayes on feature odor

```
training_features <- c("odor")
nb_mush <- naiveBayes(data.frame("odor" = mush_data[-mush_test, training_features]), mush_data[-mush_tes
pred_mush_test <- predict(nb_mush, data.frame("odor" = mush_data[mush_test, training_features]))
pred_mush_train <- predict(nb_mush, data.frame("odor" = mush_data[-mush_test, training_features]))
tab1 <- table(pred_mush_test, mush_data[mush_test, target])
tab2 <- table(pred_mush_train, mush_data[-mush_test, target])
```

Confusion matrix test data

```
kable(table(pred_mush_test, mush_data[mush_test, target]), col.names = c("e - truth", "p - truth"))
```

	e - truth	p - truth
e	405	11
p	0	396

Accuracy

```
(tab1[1,1]+tab1[2,2])/sum(tab1)
```

```
## [1] 0.9864532
```

Confusion matrix training data

```
kable(table(pred_mush_train, mush_data[-mush_test, target]), col.names = c("e - truth", "p - truth"))
```

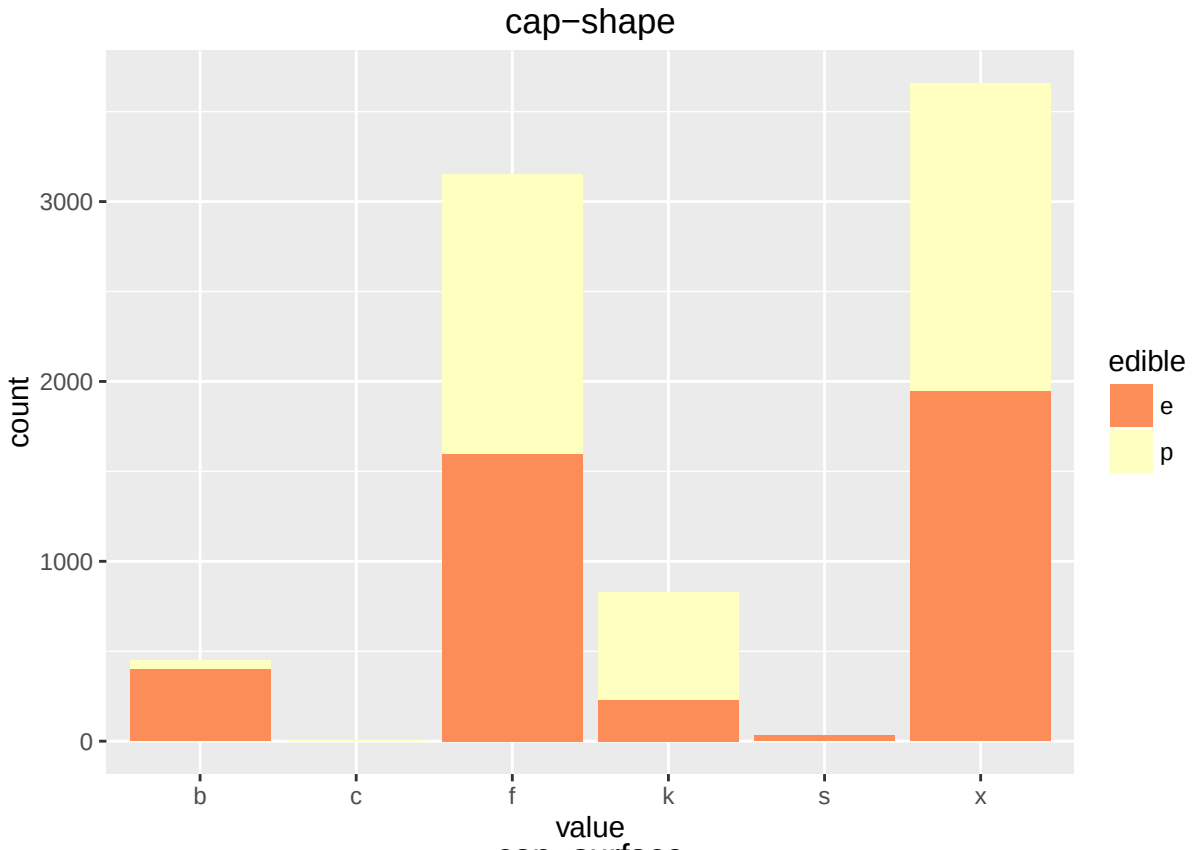
	e - truth	p - truth
e	3803	109
p	0	3400
####	Accuracy	

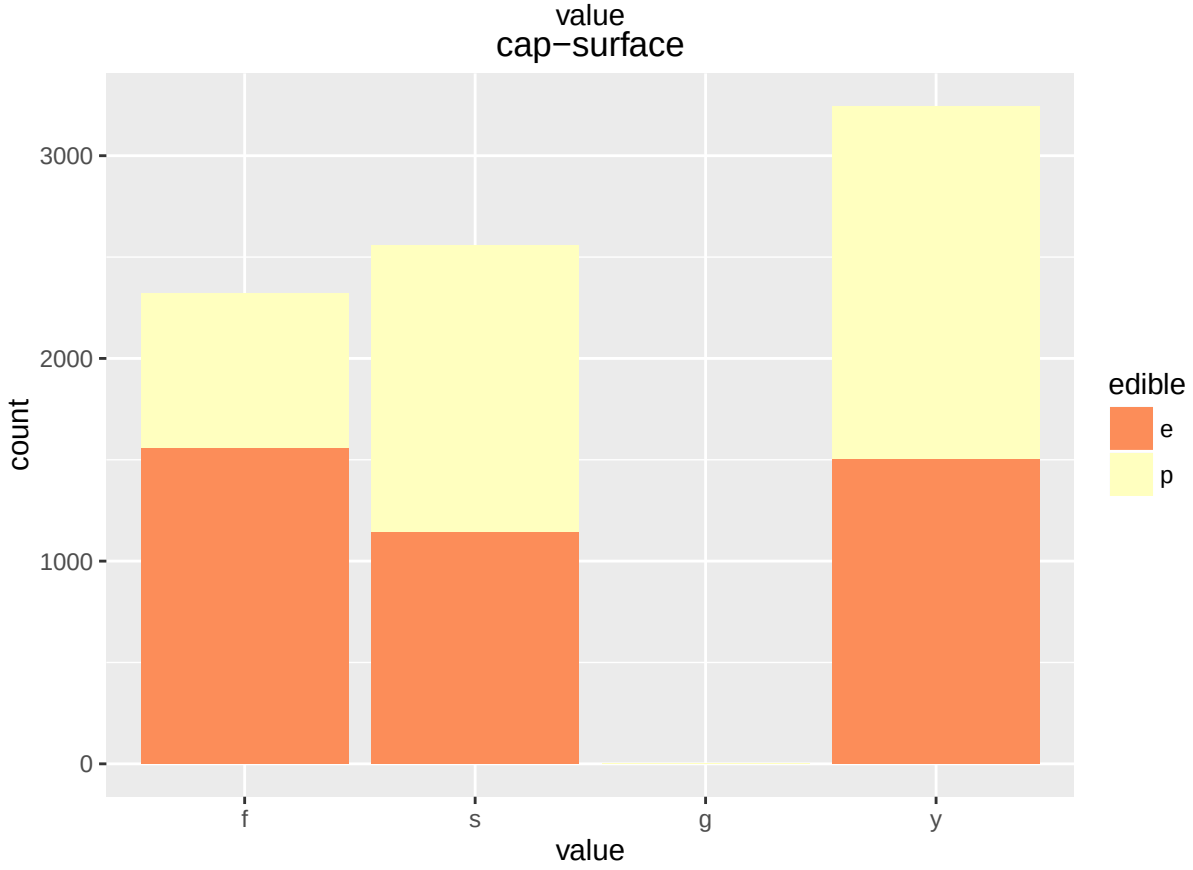
```
(tab2[1,1]+tab2[2,2])/sum(tab2)
```

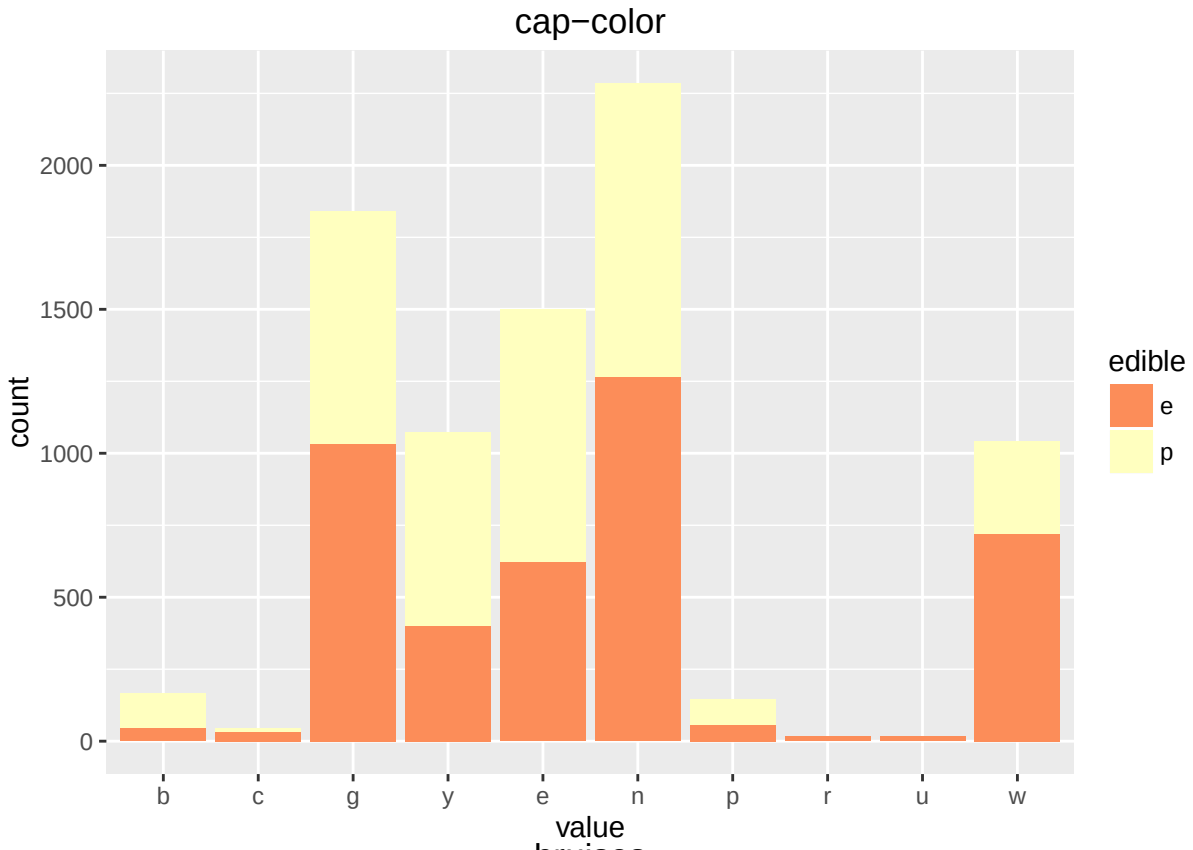
```
## [1] 0.985093
```

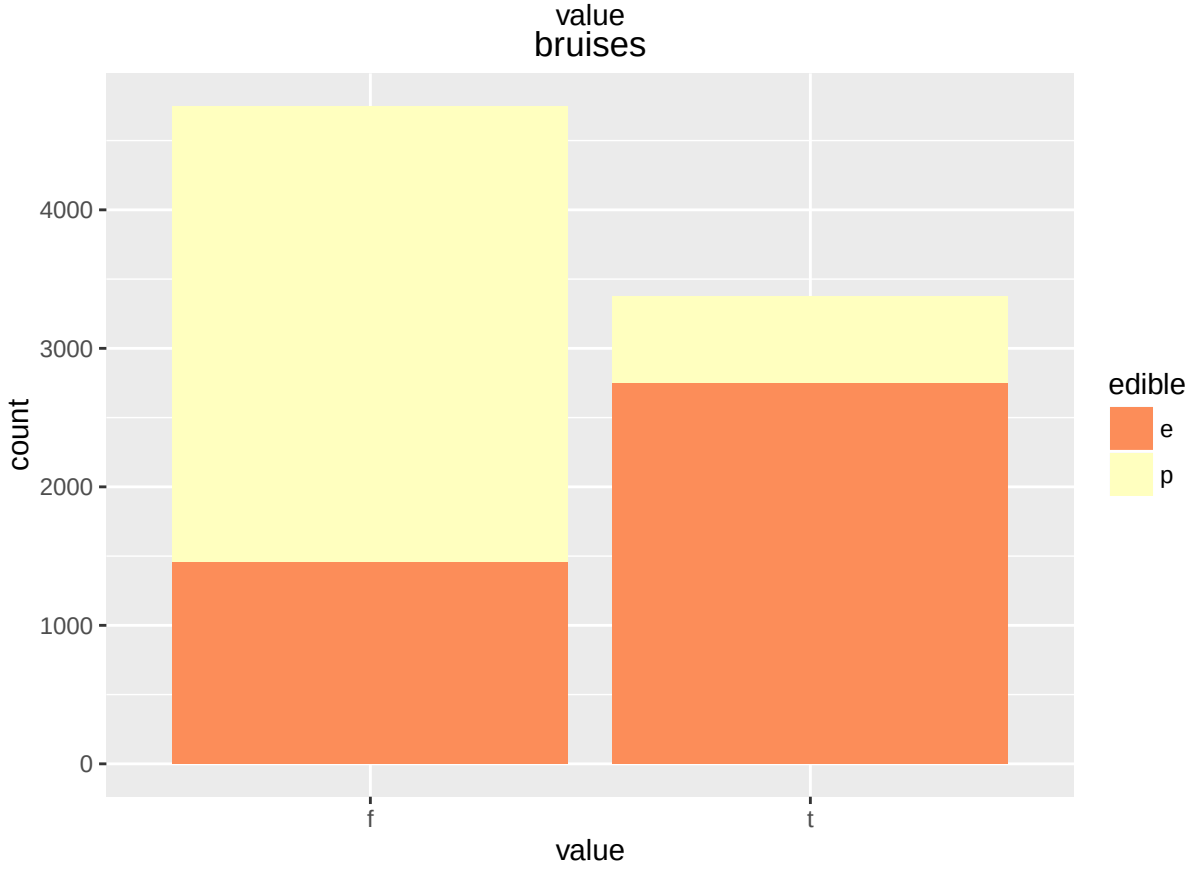
Visual feature selection

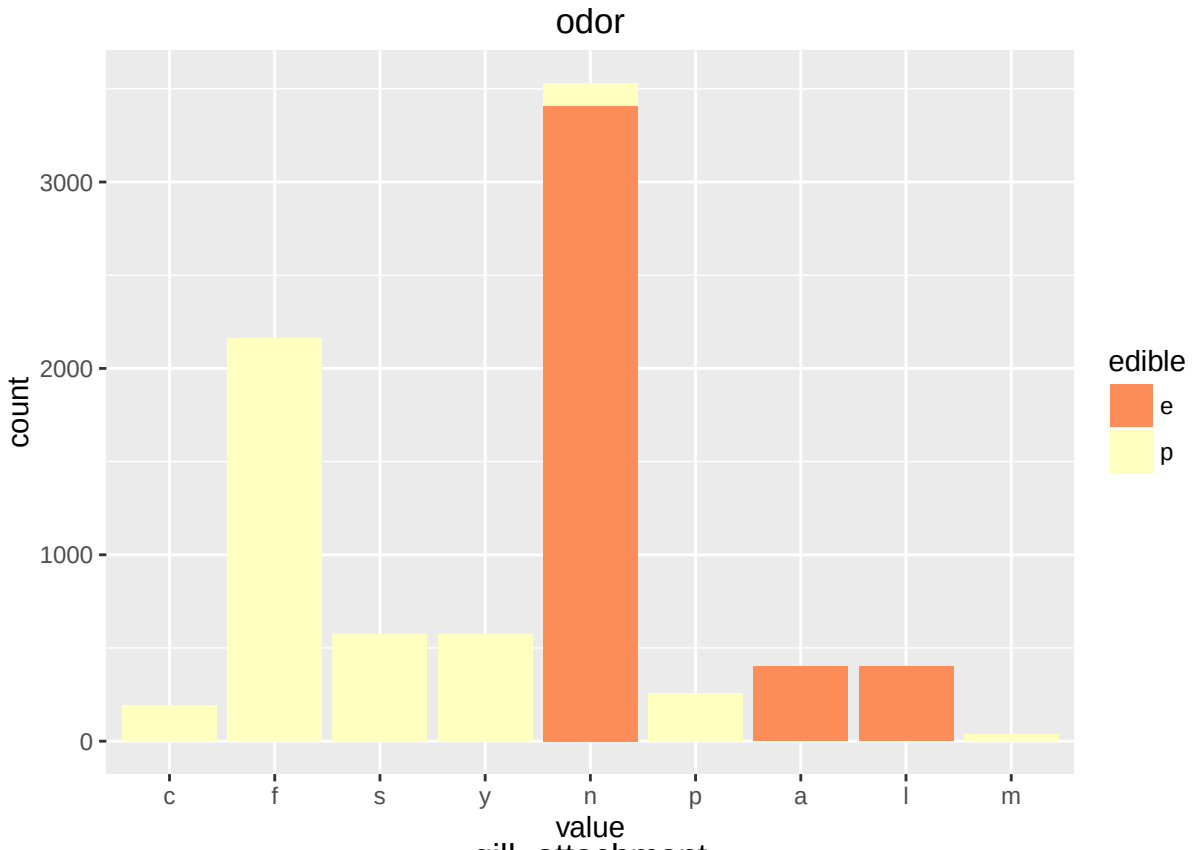
Since the dataset has many features I thought it would be useful to look at the features first and maybe it would be possible to preselect some features. So far I don't know of any method like correlation for numerical values that can also compute a correlation value for categorical data so I just chose features manually that I thought would be useful. This sections shows the plots that I created.

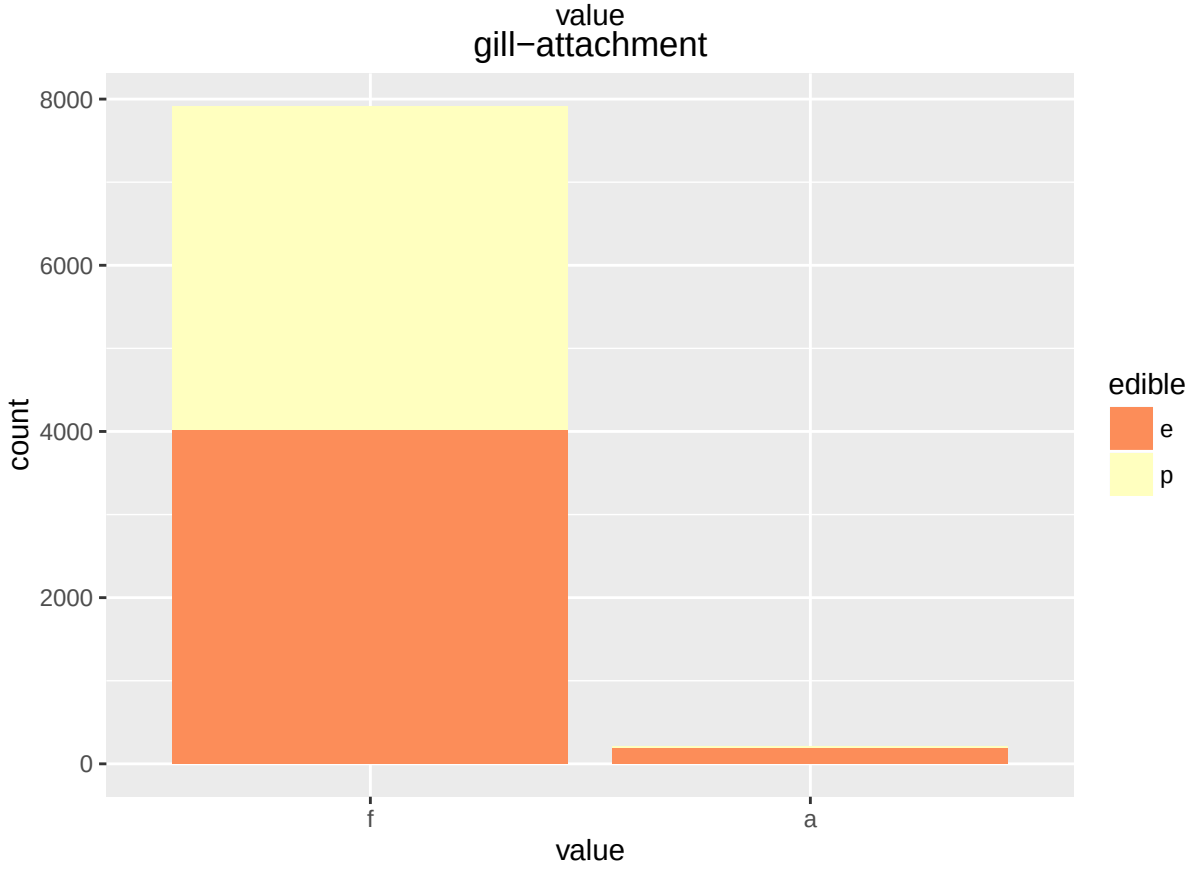

cap−shape
count
3000
2000
1000
0
b
c
f
k
s
x
value
edible
e
p

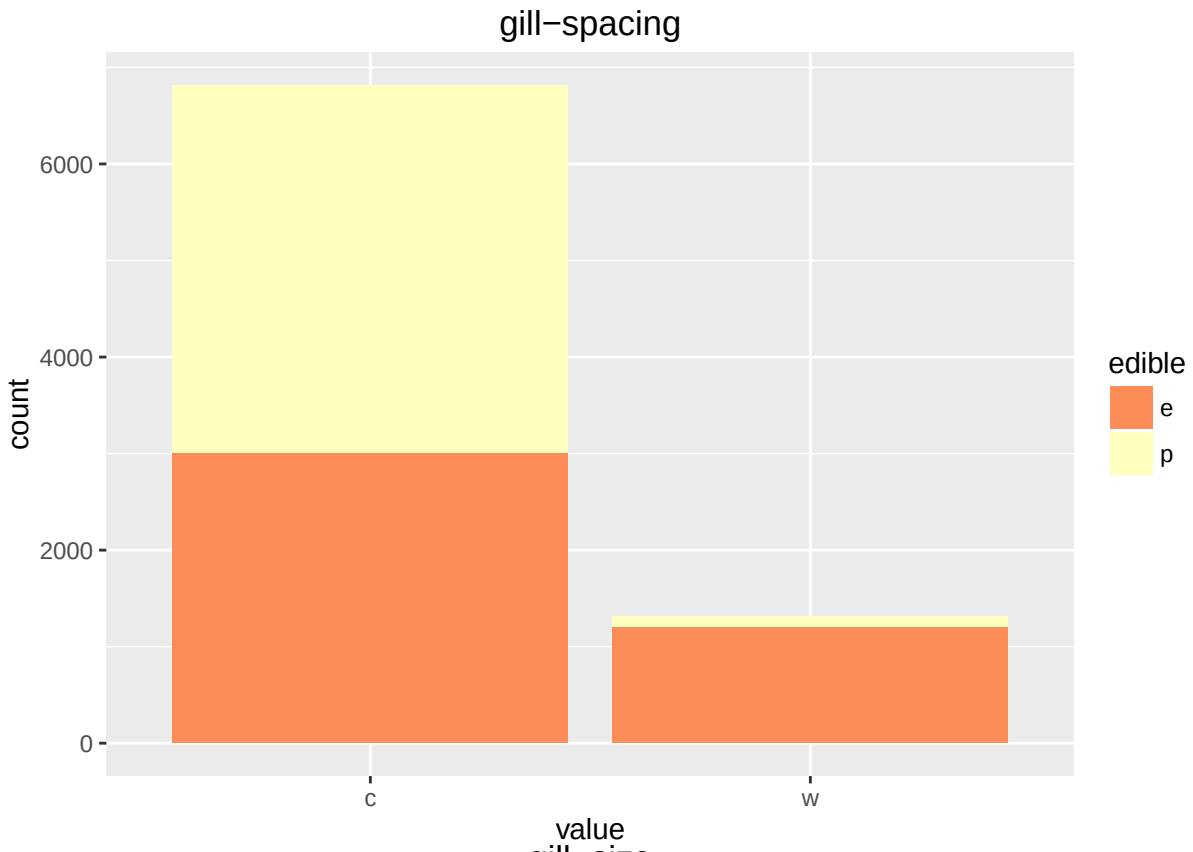

cap−surface
count
3000
2000
1000
0
f
s
g
y
value
edible
e
p

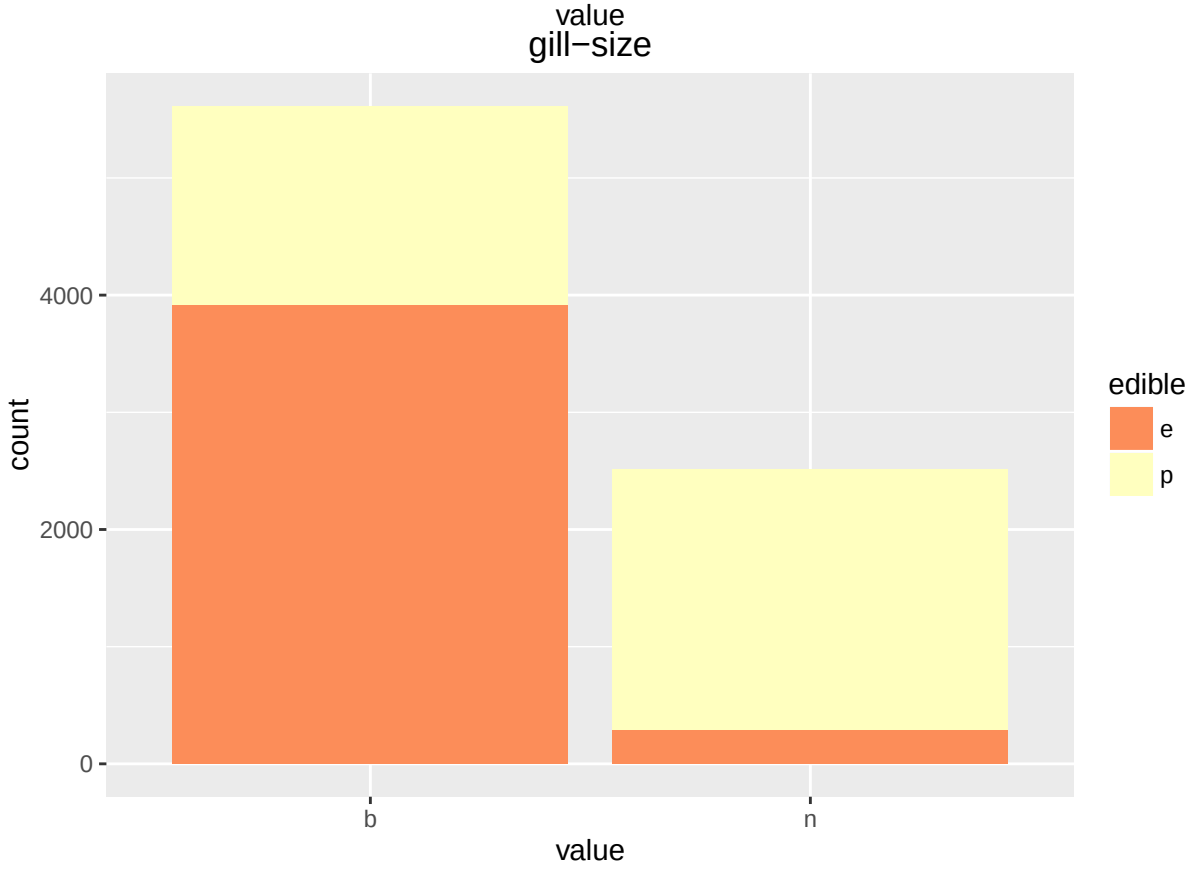

cap−color
count
2000
1500
1000
500
0
b
c
g
y
e
n
p
r
u
w
value
edible
e
p

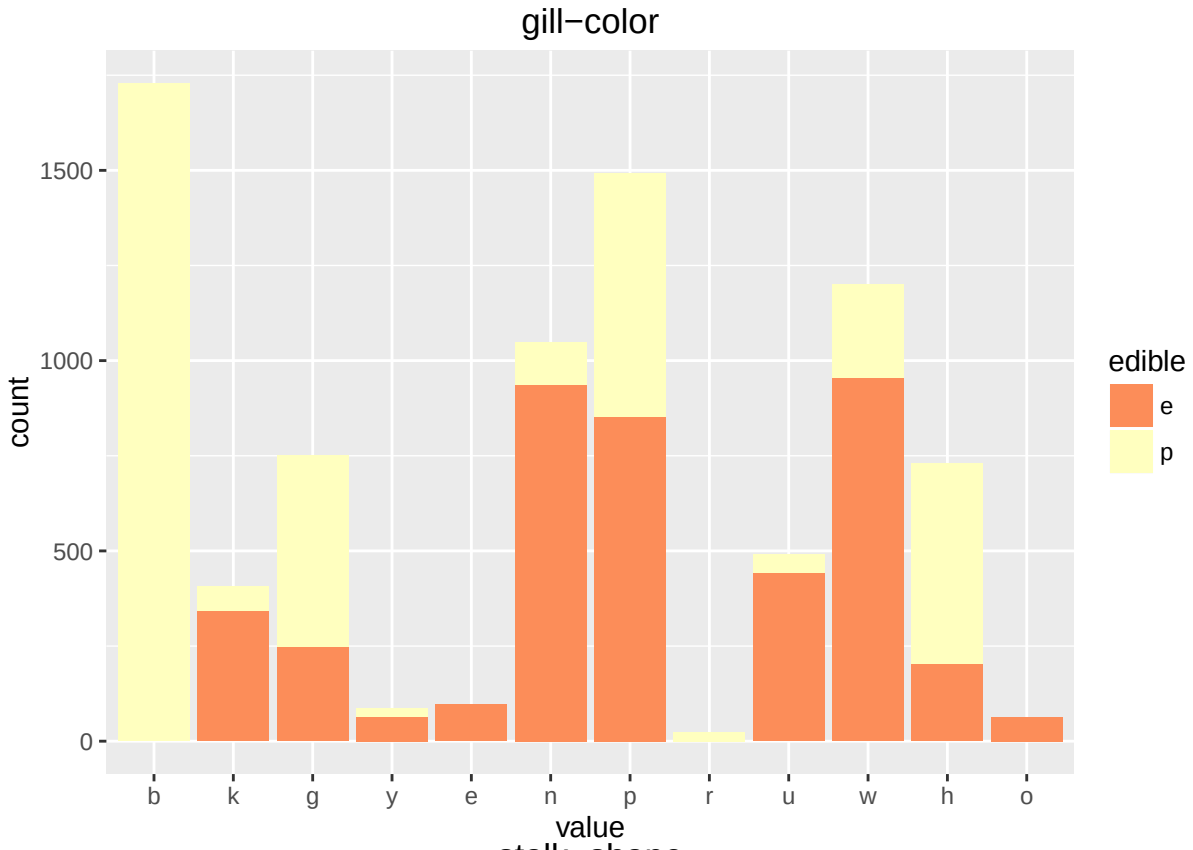

bruises
count
4000
3000
2000
1000
0
f
t
value
edible
e
p

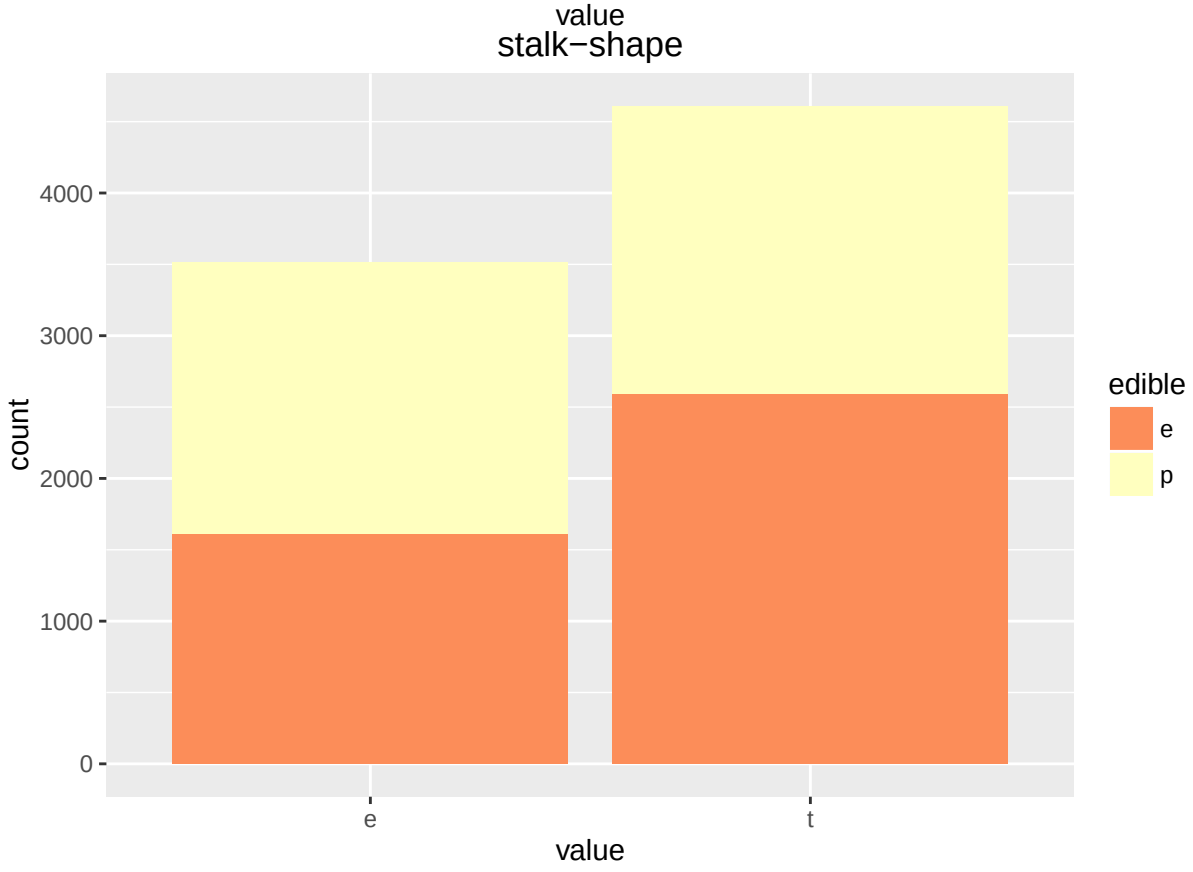

odor
count
3000
2000
1000
0
c
f
s
y
n
p
a
l
m
value
edible
e
p

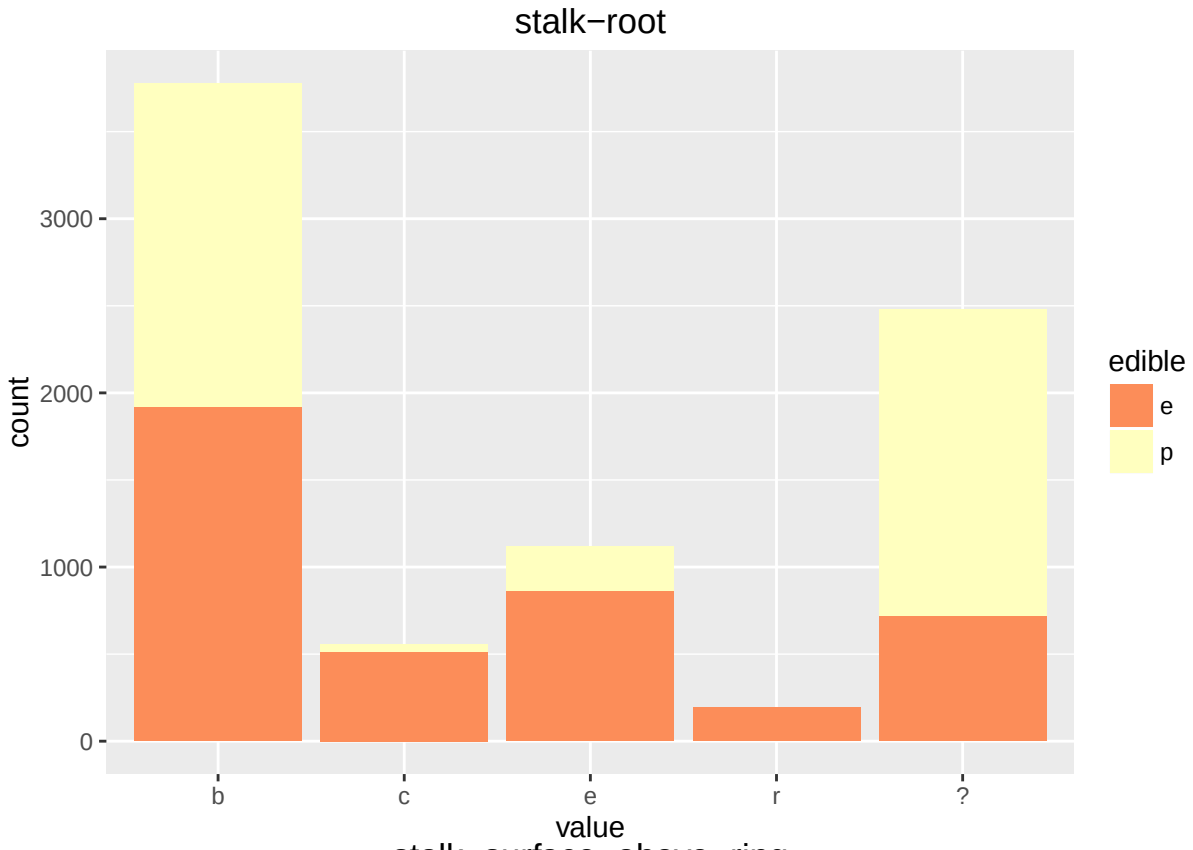

gill−attachment
count
8000
6000
4000
2000
0
f
a
value
edible
e
p

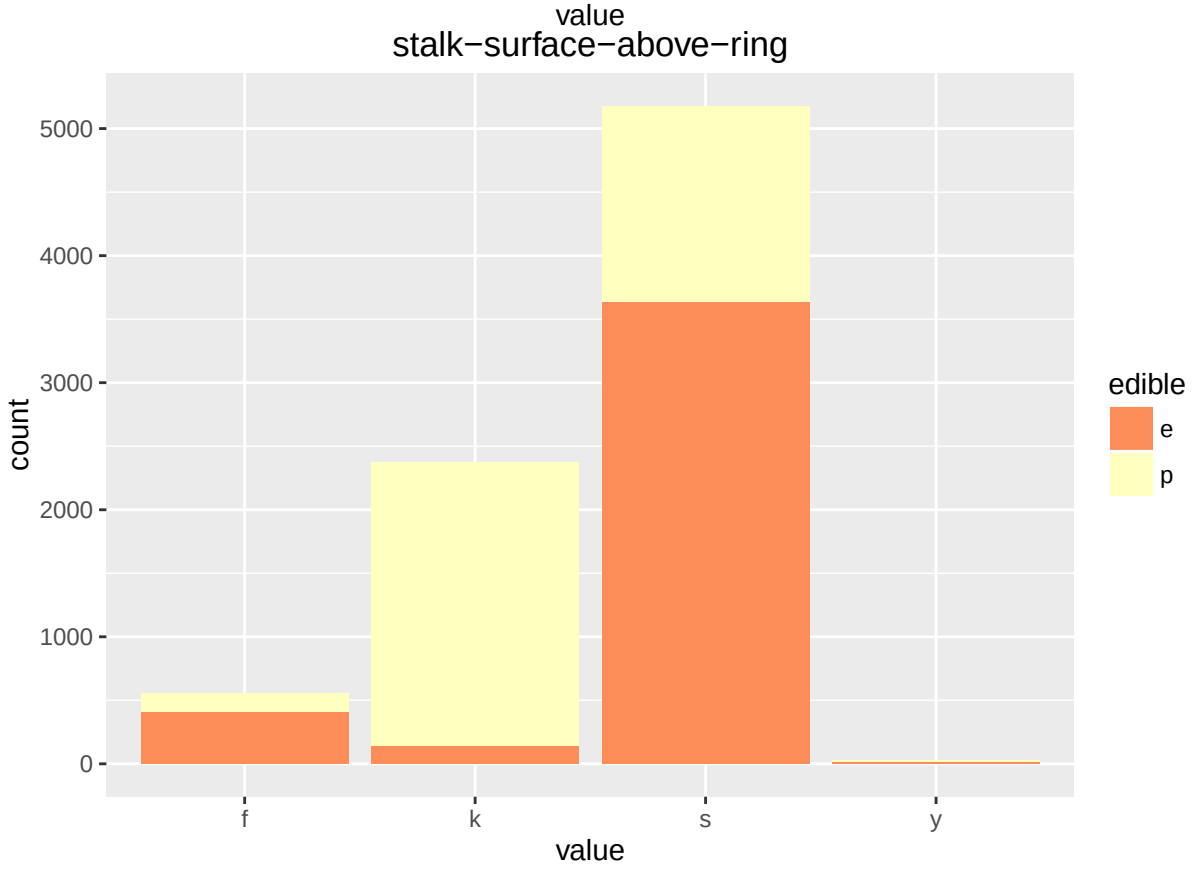

gill−spacing
6000
4000
2000
0
count
c
w
value
edible
e
p

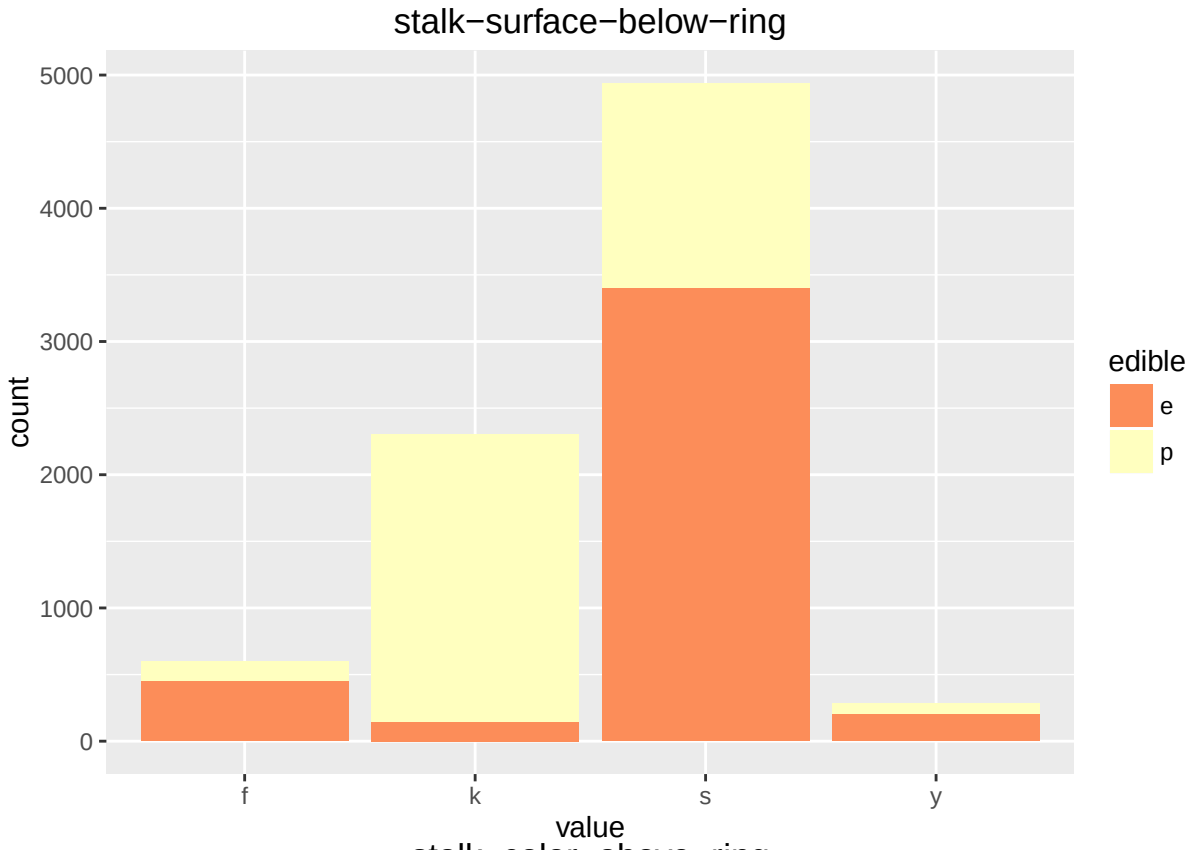

gill−size
4000
2000
0
count
b
n
value
edible
e
p

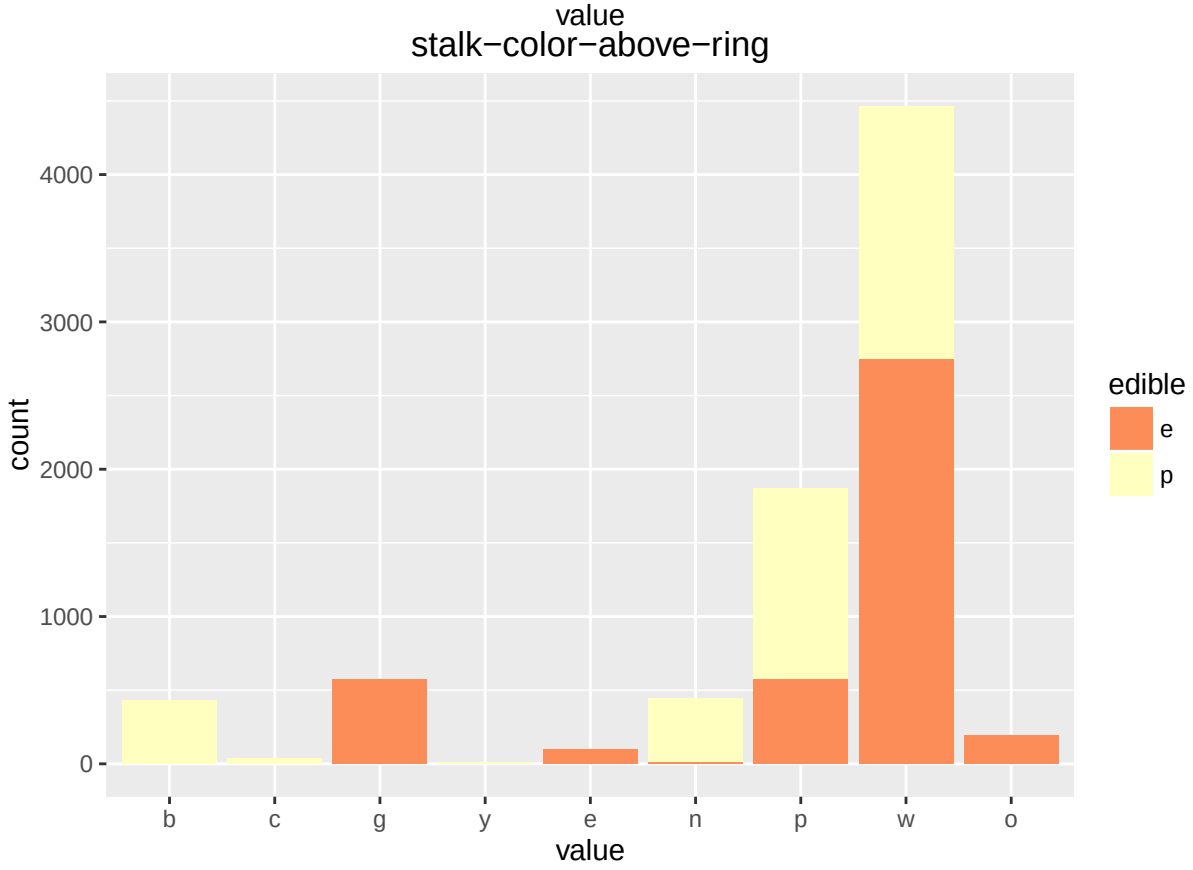

gill−color
count
1500
1000
500
0
b
k
g
y
e
n
p
r
u
w
h
o
value
edible
e
p

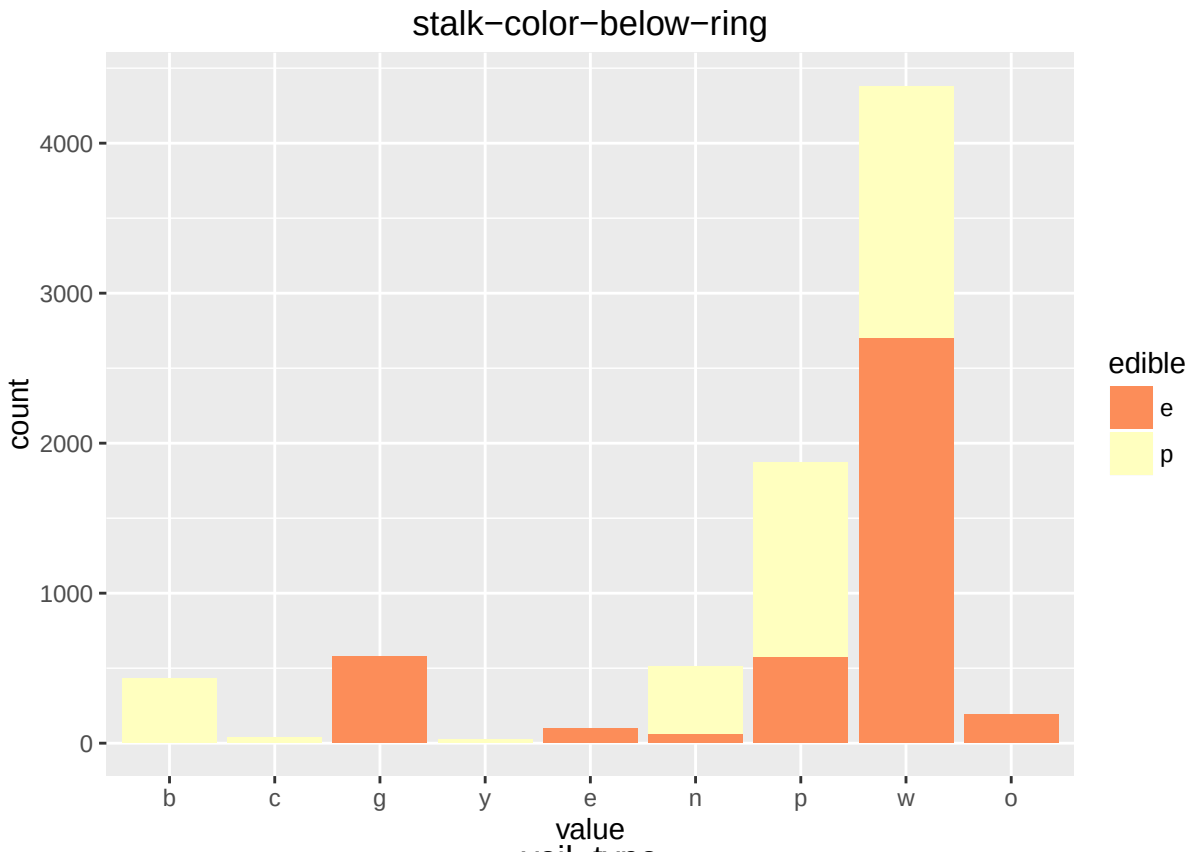

stalk−shape
count
4000
3000
2000
1000
0
e
t
value
edible
e
p

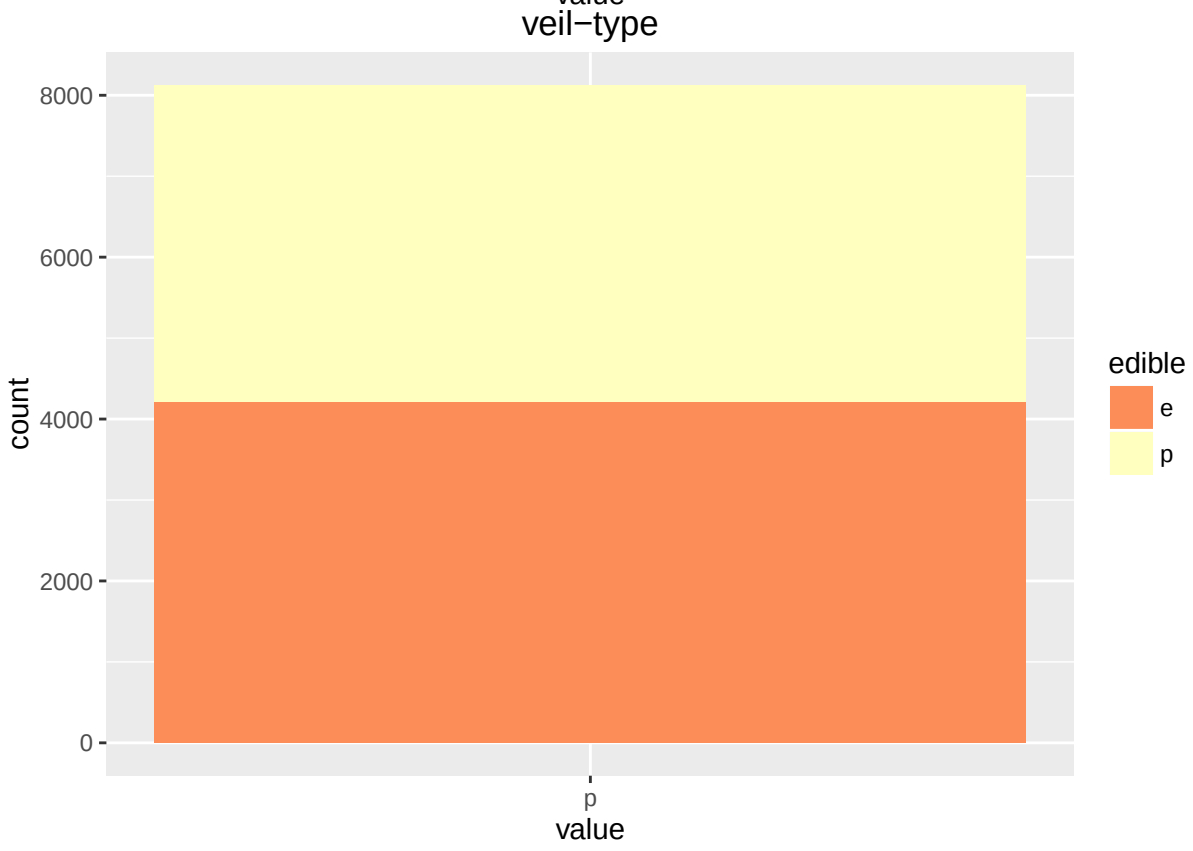

stalk−root
count
3000
2000
1000
0
b
c
e
r
?
value
edible
e
p

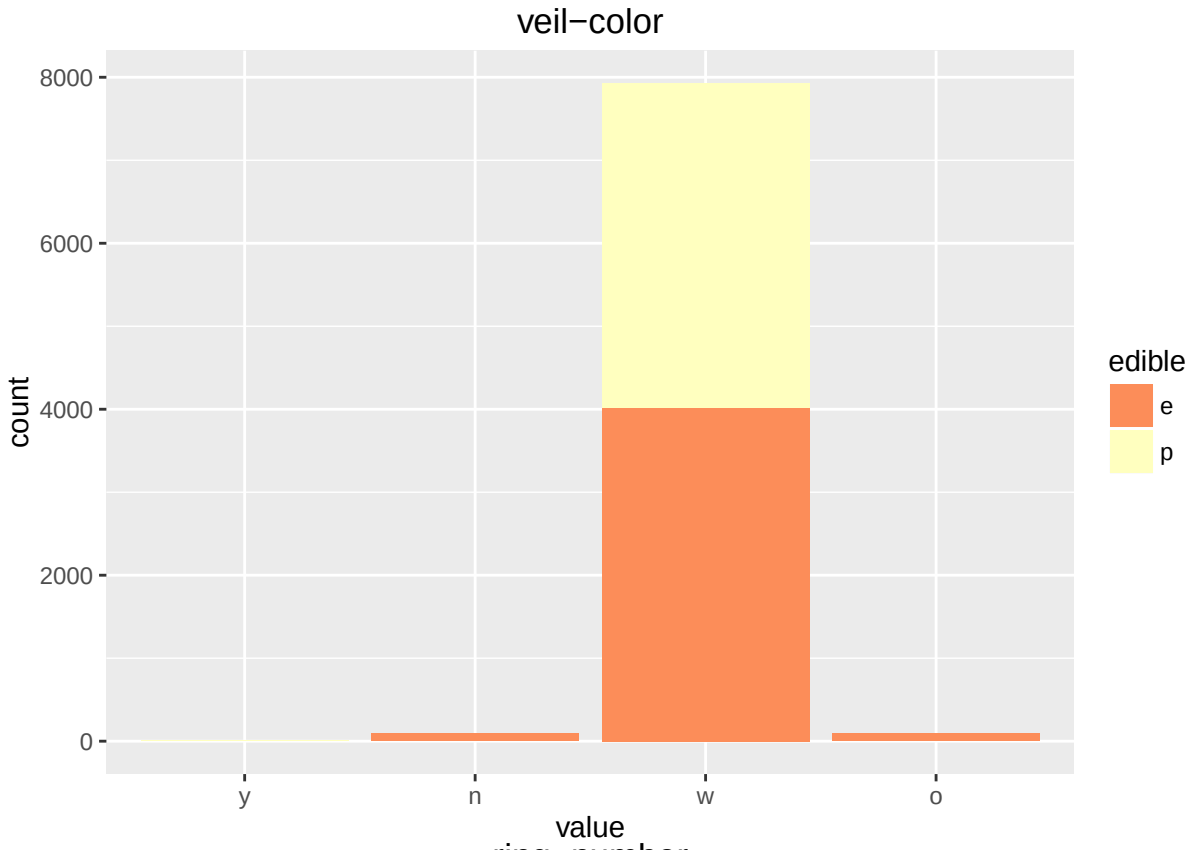

stalk−surface−above−ring
count
5000
4000
3000
2000
1000
0
f
k
s
y
value
edible
e
p

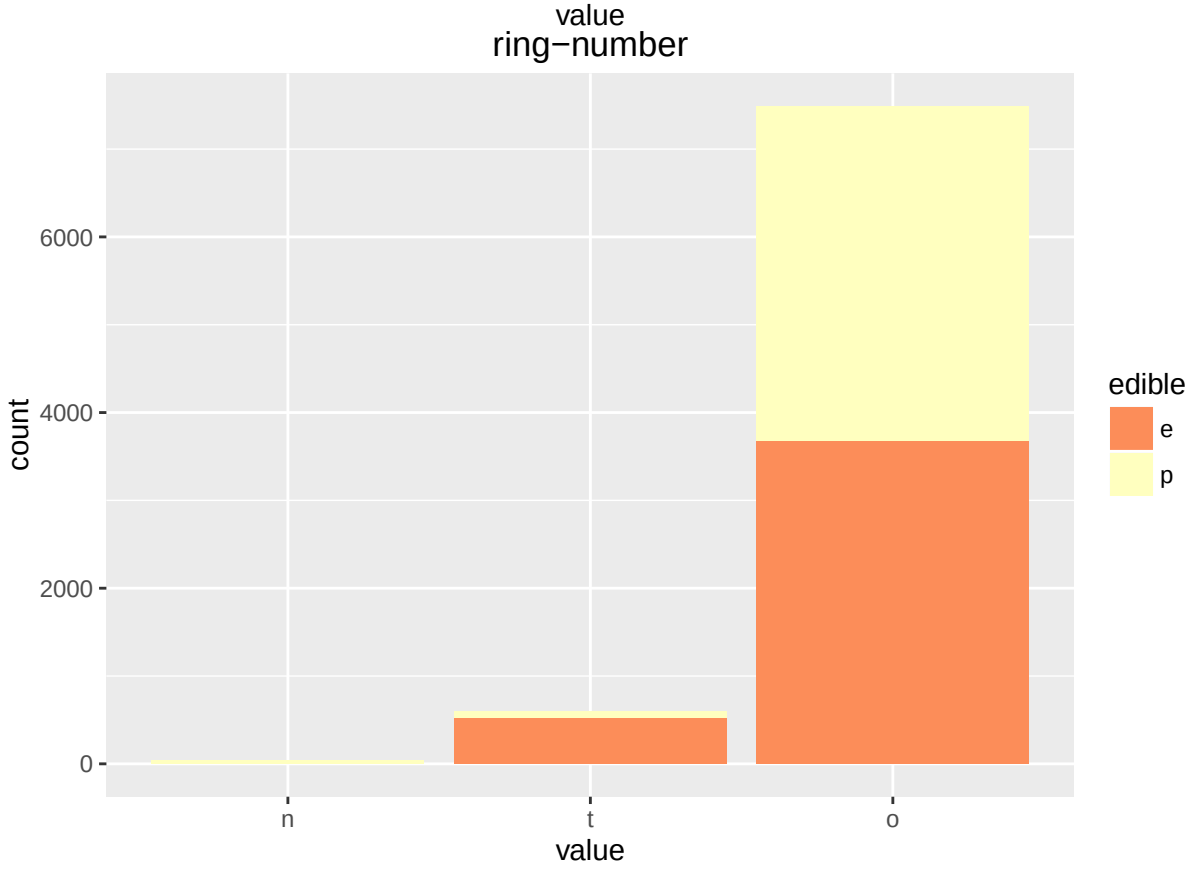

stalk−surface−below−ring
5000
4000
3000
2000
1000
0
count
f
k
s
y
value
edible
e
p

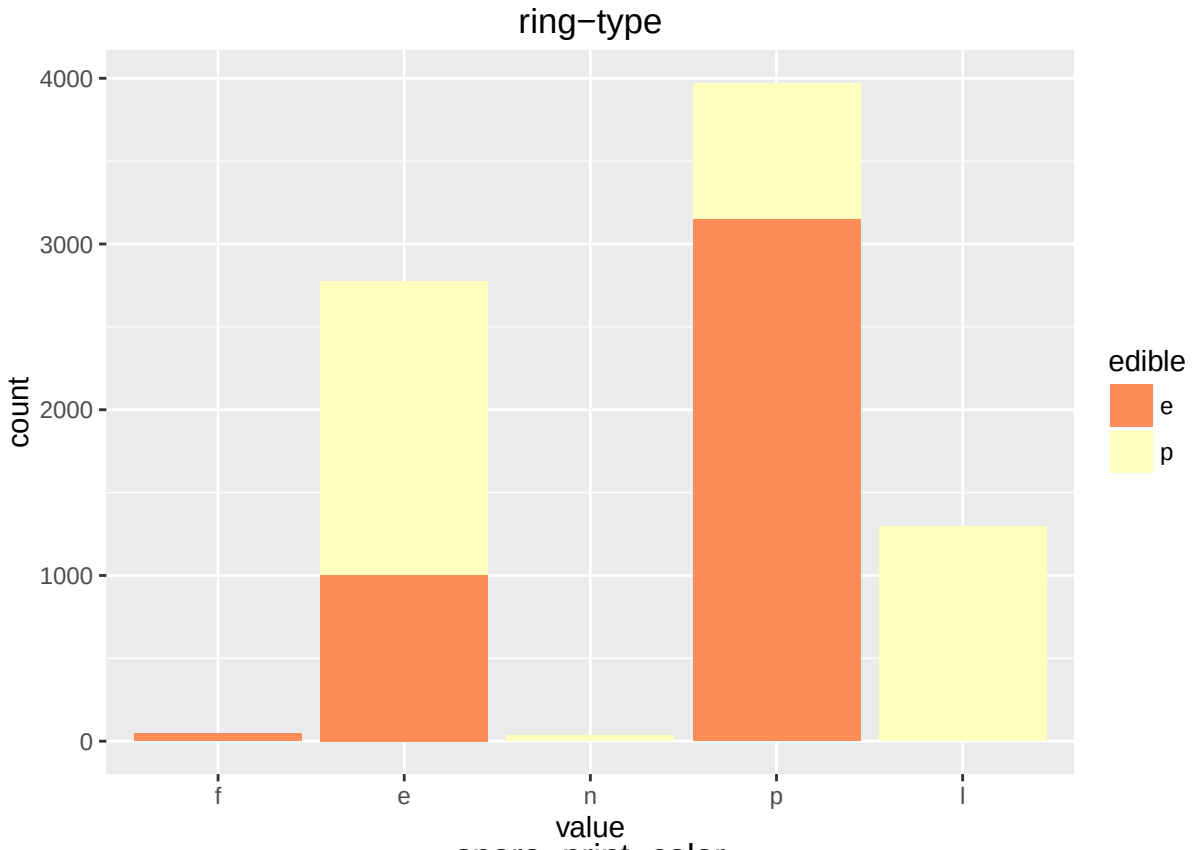

stalk−color−above−ring
4000
3000
2000
1000
0
count
b
c
g
y
e
n
p
w
o
value
edible
e
p

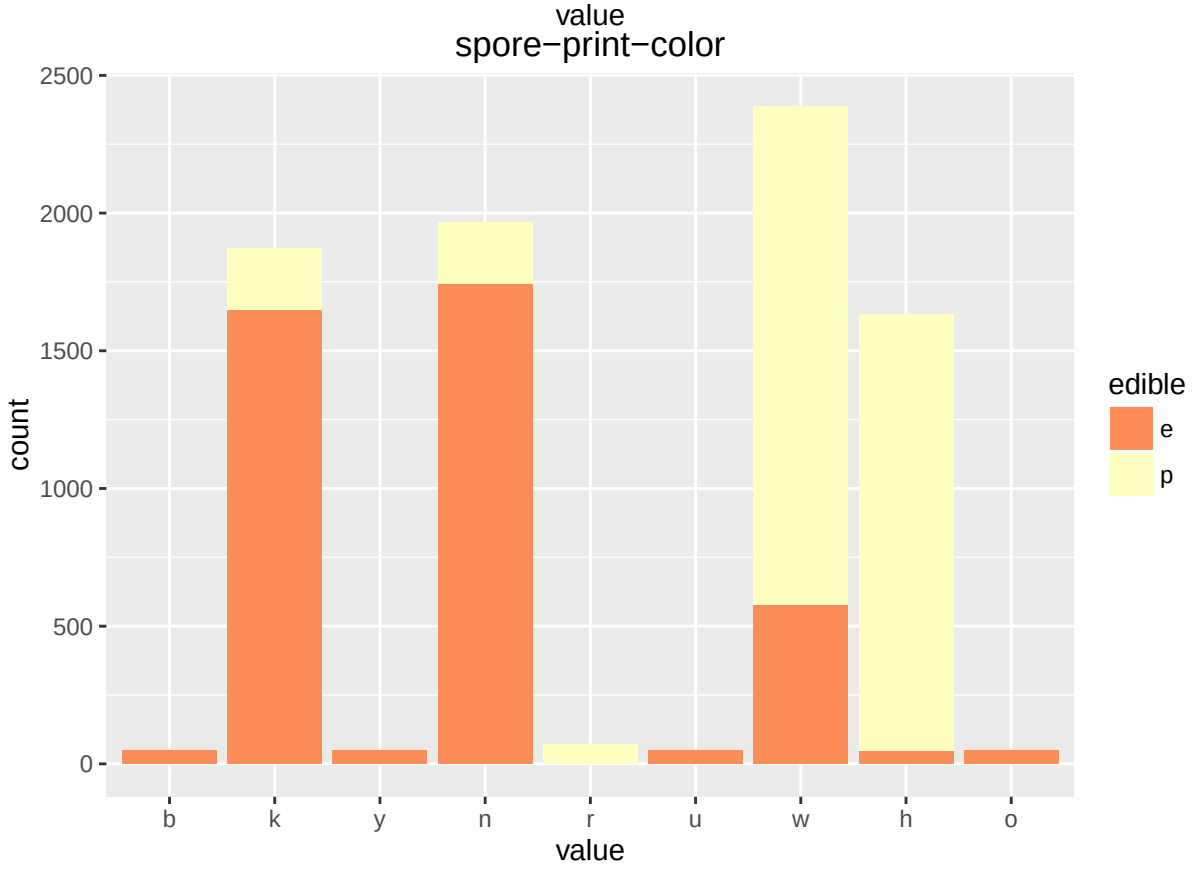

stalk−color−below−ring
count
4000
3000
2000
1000
0
b
c
g
y
e
n
p
w
o
value
edible
e
p

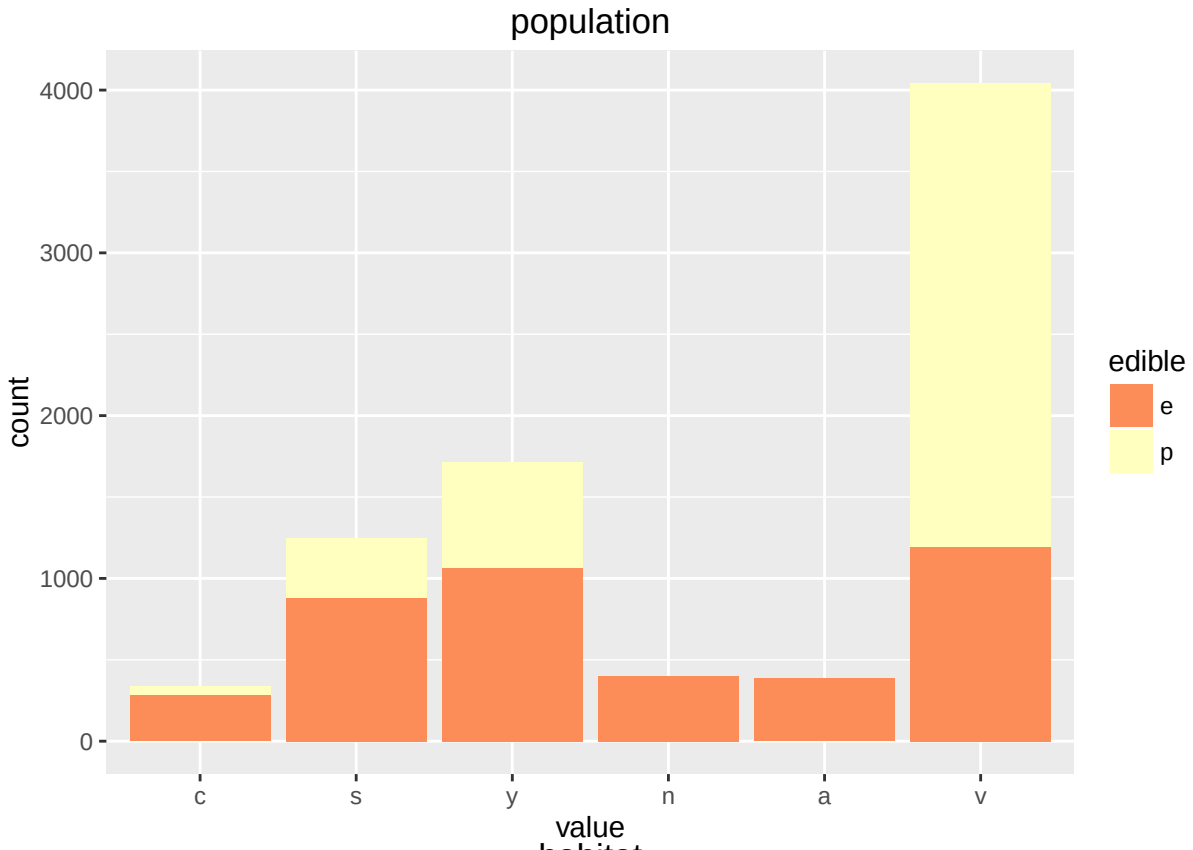

veil−type
count
8000
6000
4000
2000
0
p
value
edible
e
p

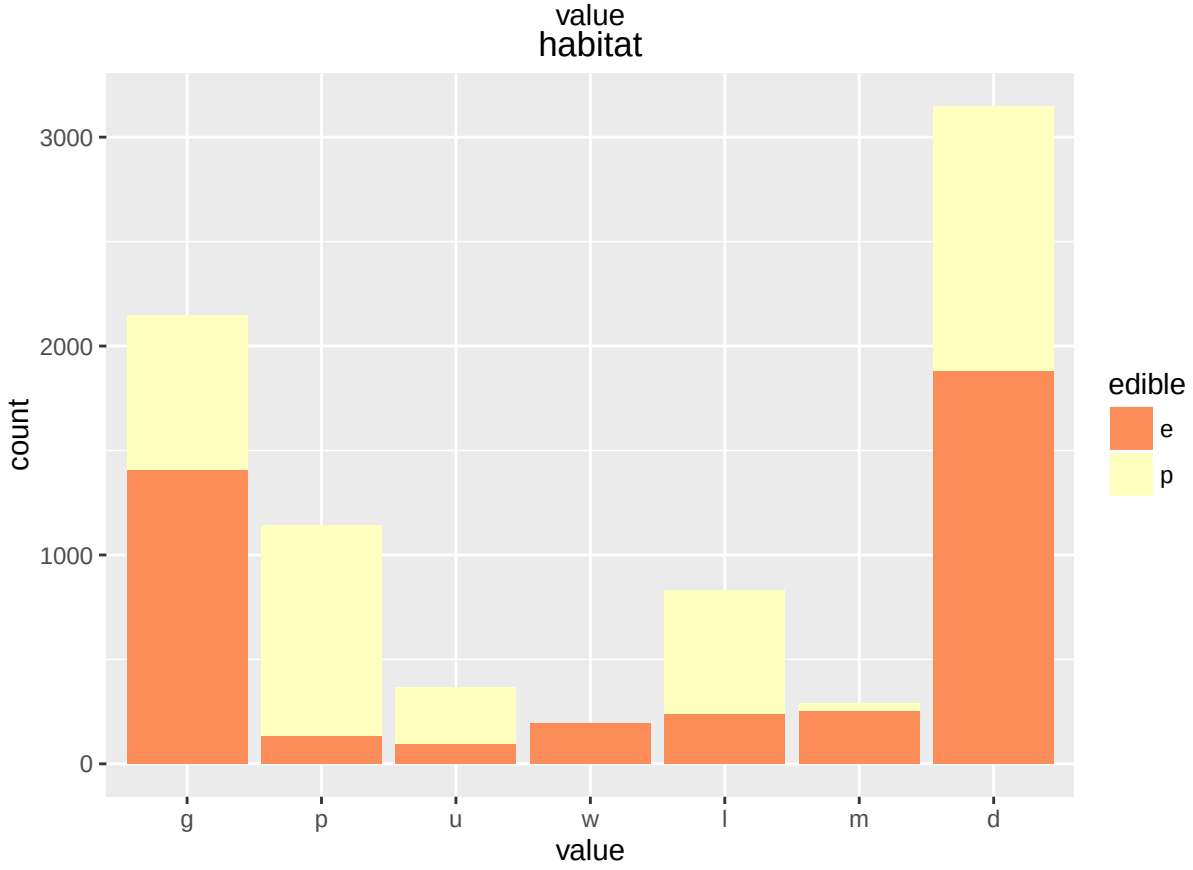

veil−color
8000
6000
4000
2000
0
count
y
n
w
o
value
edible
e
p


ring−number
6000
4000
2000
0
count
n
t
o
value
edible
e
p


ring−type
4000
3000
2000
1000
0
count
f
e
n
p
l
value
edible
e
p


spore−print−color
2500
2000
1500
1000
500
0
count
b
k
y
n
r
u
w
h
o
value
edible
e
p


population
count
4000
3000
2000
1000
0
c
s
y
n
a
v
value
edible
e
p


habitat
count
3000
2000
1000
0
g
p
u
w
l
m
d
value
edible
e
p

Conclusion

It's interesting to see that reducing the number of features increases the accuracy. Using only one feature, **odor**, yields the highest accuracy. I conclude that analysing the data first before performing algorithms is an important step.

It can also be seen that test and training accuracy are very close for all 3 models, from which I conclude that the Naive Bayes algorithm does not overfit on the training data.

Future work

There are still a few things that could/should be done. Not only accuracy, but also precision/recall should be computed and compared. Furthermore the split into training/test set should be performed several times to average over the model performance (Cross Validation).