Homework 8 - Data Visualiztion Techniques

Tymoteusz Barciński (Student ID number: 313469)

01.02.2022

1 Introduction

The following report summarises the exploratory data analysis of a dataset "mushrooms.csv" conducted for a class Data Visualization Techniques.

1.1 Dataset

This dataset includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family Mushroom. Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. This latter class was combined with the poisonous one. The dataset is available on the site: www.kaggle.com/uciml/mushroom-classification

1.2 Attribute Information

- classes: edible=e, poisonous=p
- 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,yellow=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

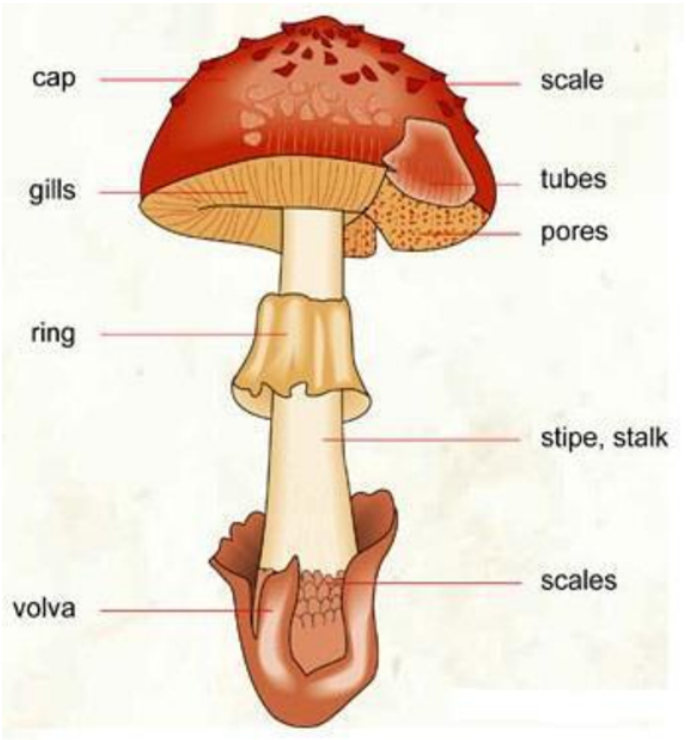


Fig. 2. Mushroom parts [24].

2 Introductory Analysis

In order to find out about different characteristics of the dataset the DataReporter package was used. The following tables come from a report generated by a function from this package. They neatly and accurately describe the nature of the dataset. As one can see, there are no quantitive variables. This fact can potentially limit the scope of the analysis. On a positive note, there are no missing observations.

	character	factor	labelled	haven labelled	numeric	integer	logical	Date
Identify miscoded missing values	×	×	×	×	×	×		×
Identify prefixed and suffixed whitespace	×	×	×	×				
Identify levels with < 6 obs.	×	×	×	×				
Identify case issues	×	×	×	×				
Identify misclassified numeric or integer variables	×	×	×	×				
Identify outliers					×	×		×

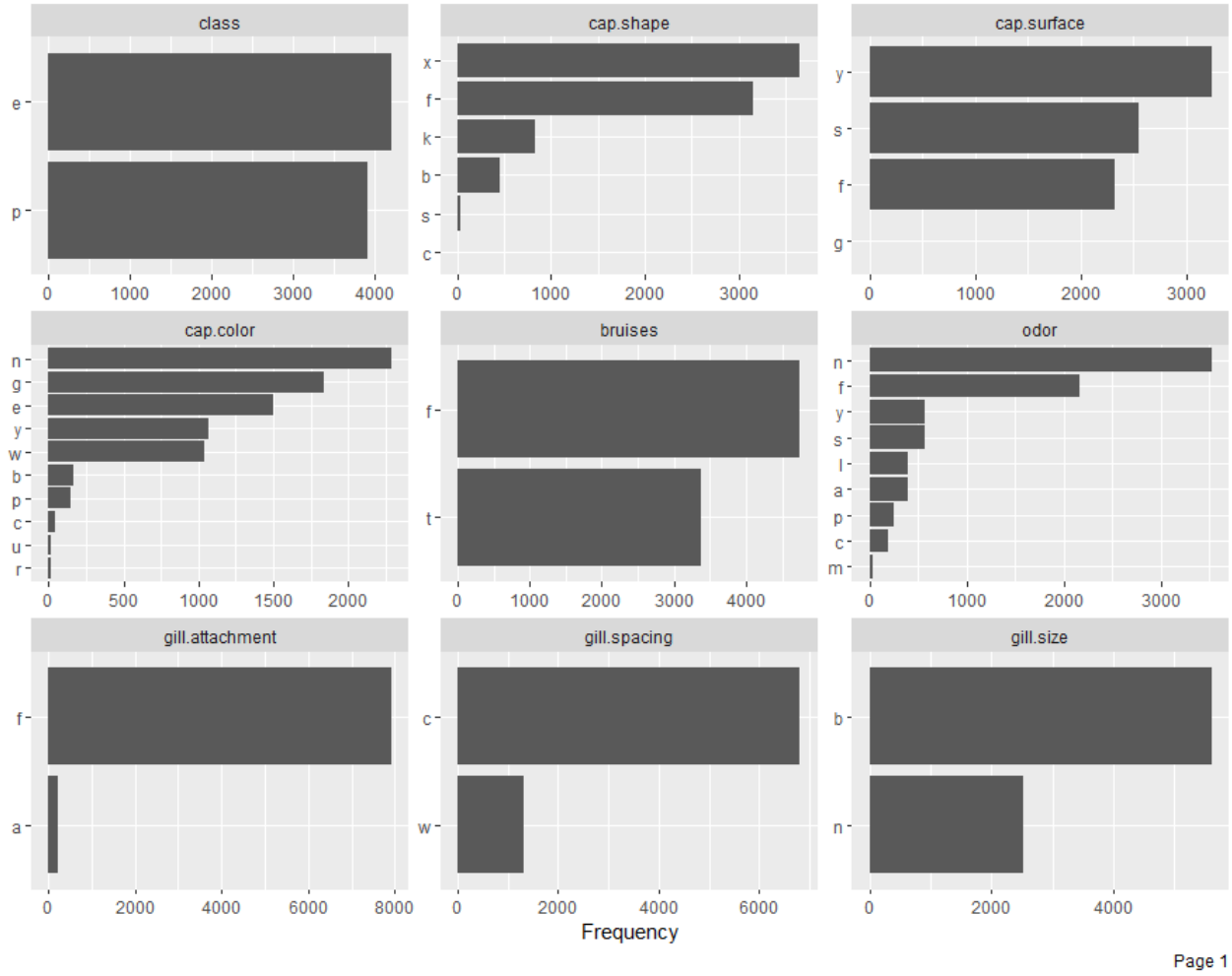
	Variable class	# unique values	Missing observations	Any problems?
class	character	2	0.00 %	
cap.shape	character	6	0.00 %	×
cap.surface	character	4	0.00 %	×
cap.color	character	10	0.00 %	
bruises	character	2	0.00 %	
odor	character	9	0.00 %	
gill.attachment	character	2	0.00 %	
gill.spacing	character	2	0.00 %	
gill.size	character	2	0.00 %	
gill.color	character	12	0.00 %	
stalk.shape	character	2	0.00 %	
stalk.root	character	5	0.00 %	
stalk.surface.above.ring	character	4	0.00 %	
stalk.surface.below.ring	character	4	0.00 %	
stalk.color.above.ring	character	9	0.00 %	
stalk.color.below.ring	character	9	0.00 %	
veil.type	character	1	0.00 %	×
veil.color	character	4	0.00 %	
ring.number	character	3	0.00 %	
ring.type	character	5	0.00 %	
spore.print.color	character	9	0.00 %	
population	character	6	0.00 %	
habitat	character	7	0.00 %	

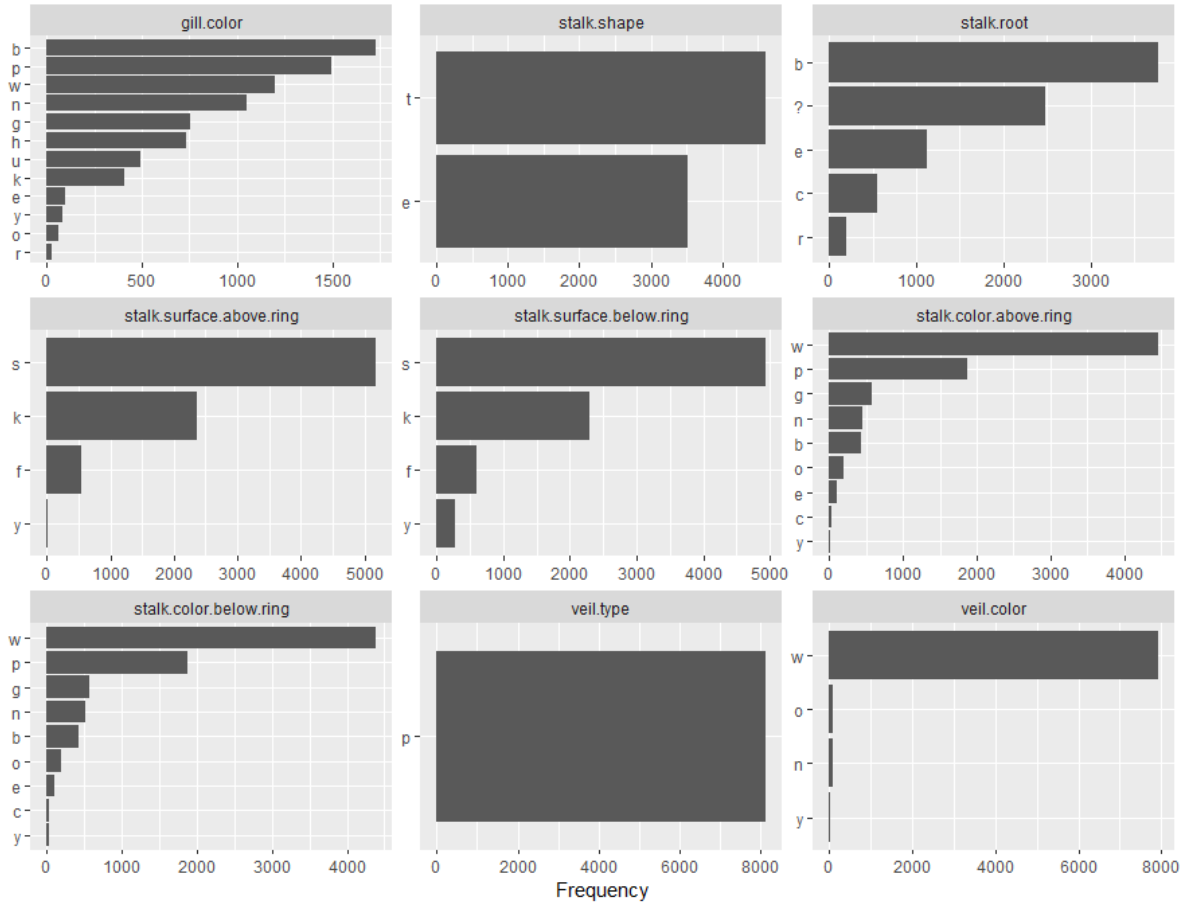
3 Introductory Plots

In this seciton, some basic bar plots will be presented. They help understand the distribution of every variable and categorize some variables as significant.

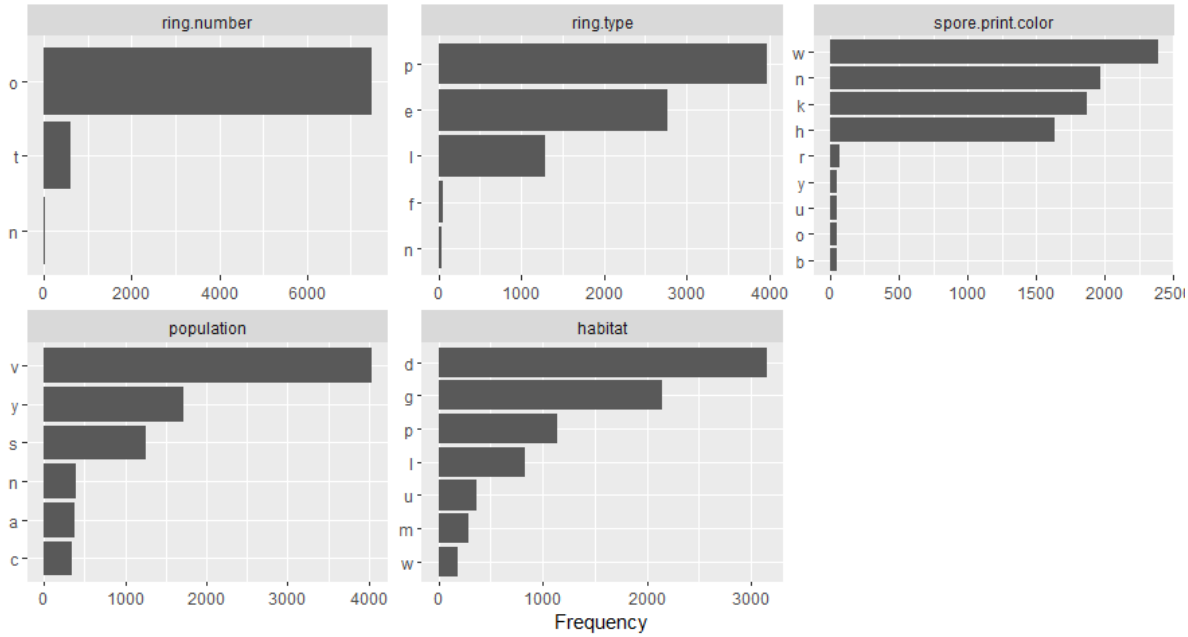
3.1 Basic Bar Plots

Some variables seem to have interesting distributions. However, at this point it should be determined what variable is most relevant for the analysis. Not surprisingly, poisonousness is a good candidate. Could one identify those variables and those trait that help determine which mushrooms have a high probability of being edible? In order to do so, we need to include the variable class in each graph. The next section does exactly that.




gill.color
b
p
w
n
g
h
u
k
e
y
o
r
0
500
1000
1500
stalk.shape
t
e
0
1000
2000
3000
4000
stalk.root
b
?
e
c
r
0
1000
2000
3000
stalk.surface.above.ring
s
k
f
y
0
1000
2000
3000
4000
5000
stalk.surface.below.ring
s
k
f
y
0
1000
2000
3000
4000
5000
stalk.color.above.ring
w
p
g
n
b
o
e
c
y
0
1000
2000
3000
4000
stalk.color.below.ring
w
p
g
n
b
o
e
c
y
0
1000
2000
3000
4000
veil.type
p
0
2000
4000
6000
8000
veil.color
w
o
n
y
0
2000
4000
6000
8000
Frequency

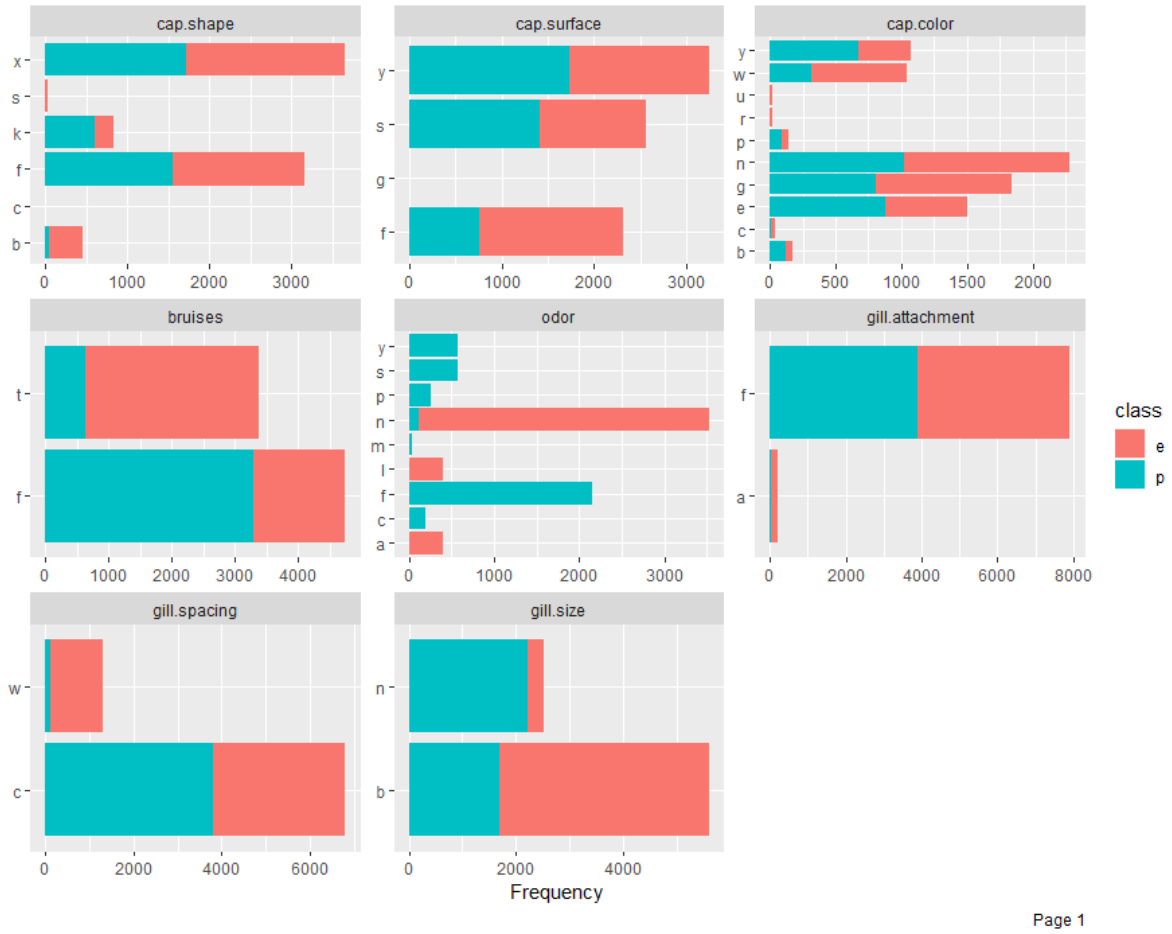
Page 2


ring.number
o
t
n
0
2000
4000
6000
ring.type
p
e
l
f
n
0
1000
2000
3000
4000
spore.print.color
w
n
k
h
r
y
u
o
b
0
500
1000
1500
2000
population
v
y
s
n
a
c
0
1000
2000
3000
4000
habitat
d
g
p
l
u
m
w
0
1000
2000
3000
Frequency

Page 3

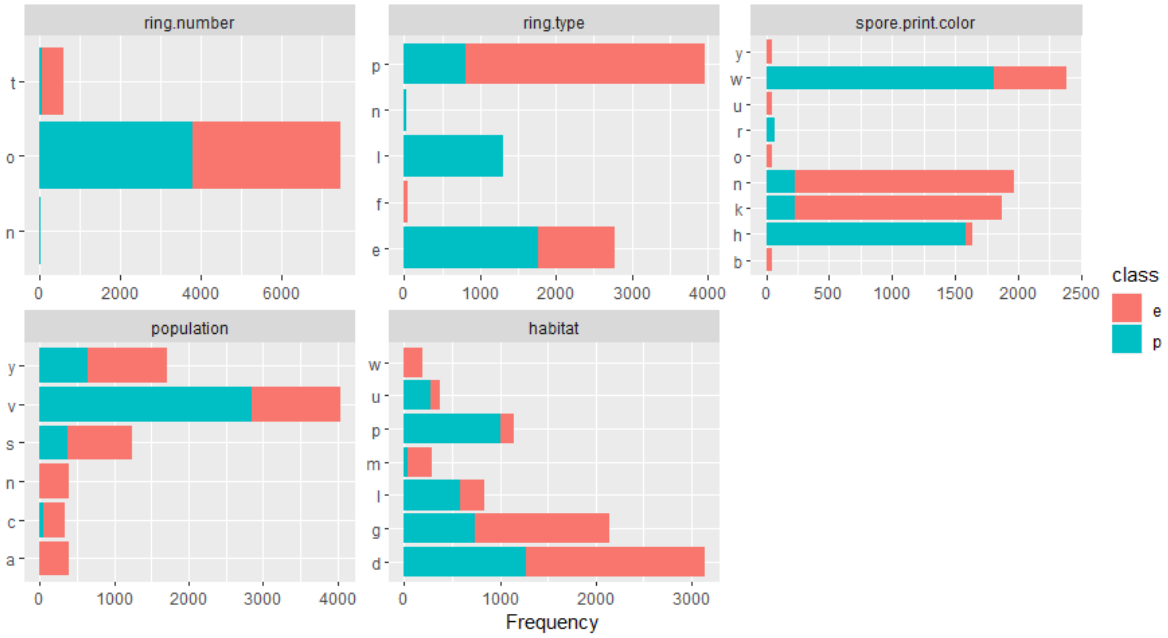
3.2 Bar Plots accounted for poisonousness

Some variables, such as cap surface or cap color, may not be the best predictors of edibleness. The majority of categories in those variables are divided equally between the variable class. Therefore, they won't help much a mushroom picker. Let's identify those variables that are more unambiguous. After close examination 4 variables can be identified. Namely: odor, gill color, ring type and spore print color.




gill.color
stalk.shape
stalk.root
stalk.surface.above.ring
stalk.surface.below.ring
stalk.color.above.ring
stalk.color.below.ring
veil.type
veil.color
class
e
p
Frequency

Page 2


ring.number
ring.type
spore.print.color
population
habitat
class
e
p
Frequency

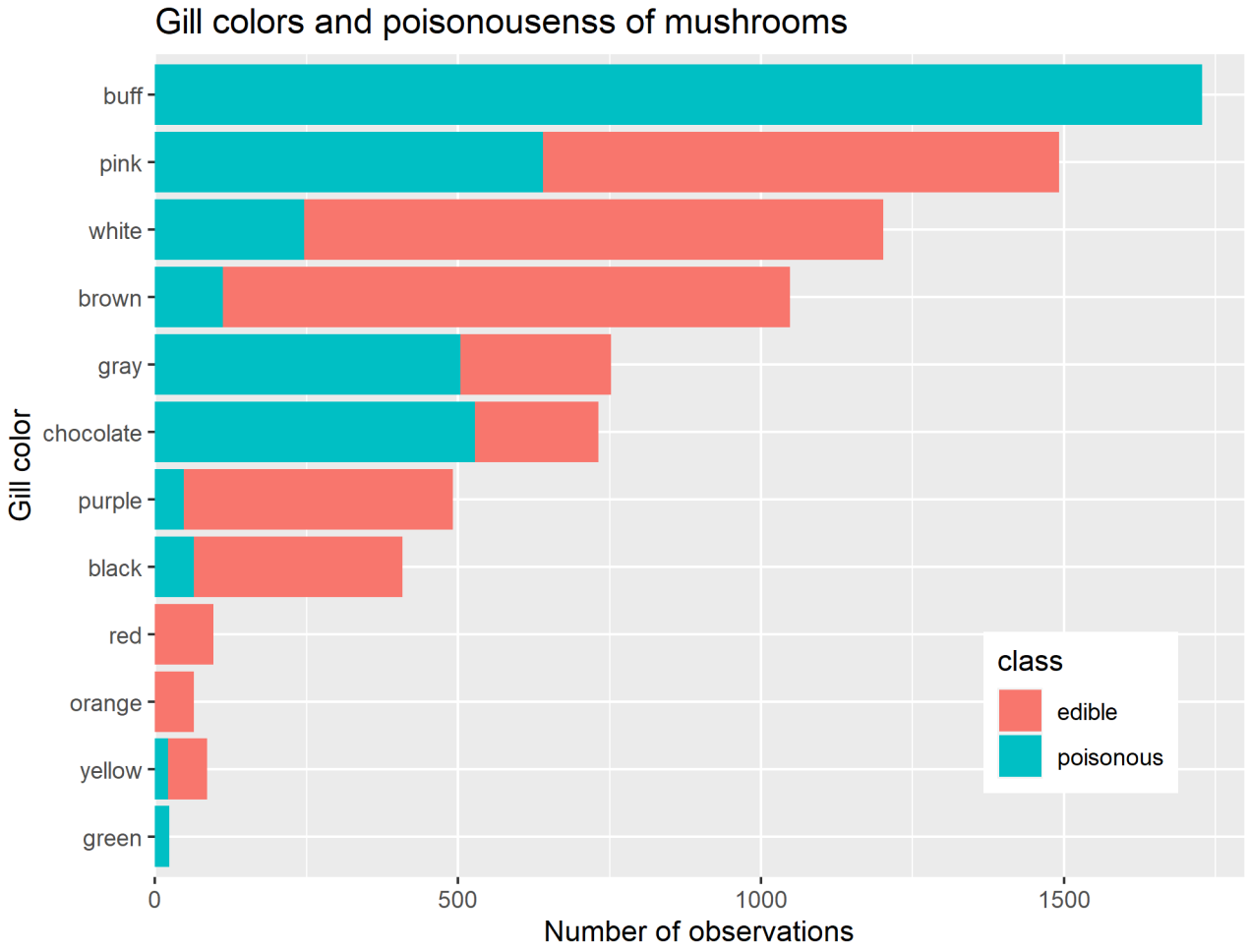
Page 3

4 Exploratory plots

In this section we will take a closer look at those 4 variables identified previously as highly indicative of edibleness.

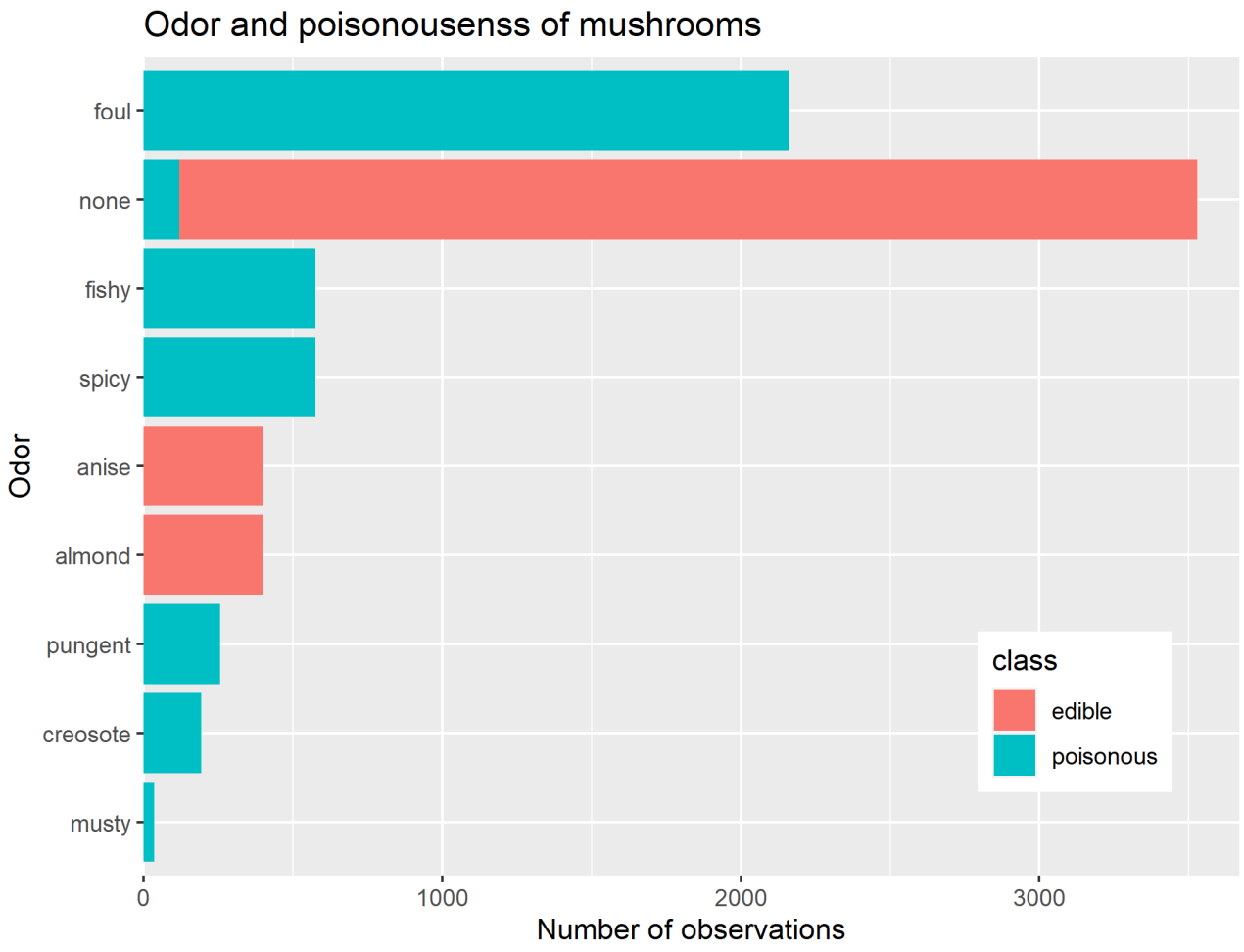
4.1 Gill Color

Mushrooms with gill that is either white or brown are much more likely to be edible than poisonous. Therefore, mushroom pickers should look out for mushrooms with those traits. At the same time, he should be beware of those with buff gill since they turn out to be only poisonous.



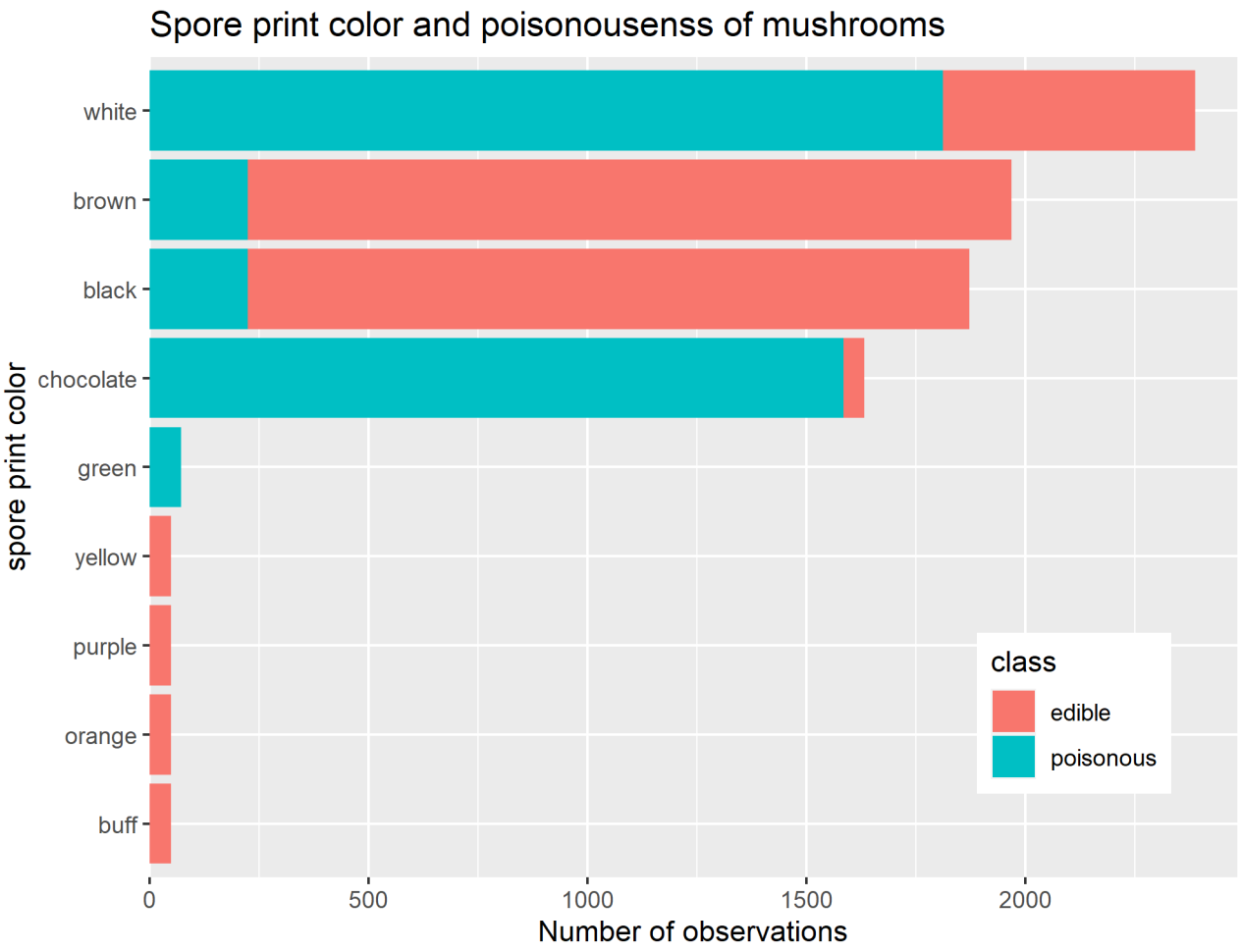
4.2 Odor

This variable seems to be the most significant one in our dataset when it comes to edibleness. A mushroom picker, will almost always find an eatable mushroom if it doesn't have a certain odor. In fact, this probability is around 97%. The problem here may lay somewhere else. Determining what kind of smell the mushroom gives out can be a challenging task.



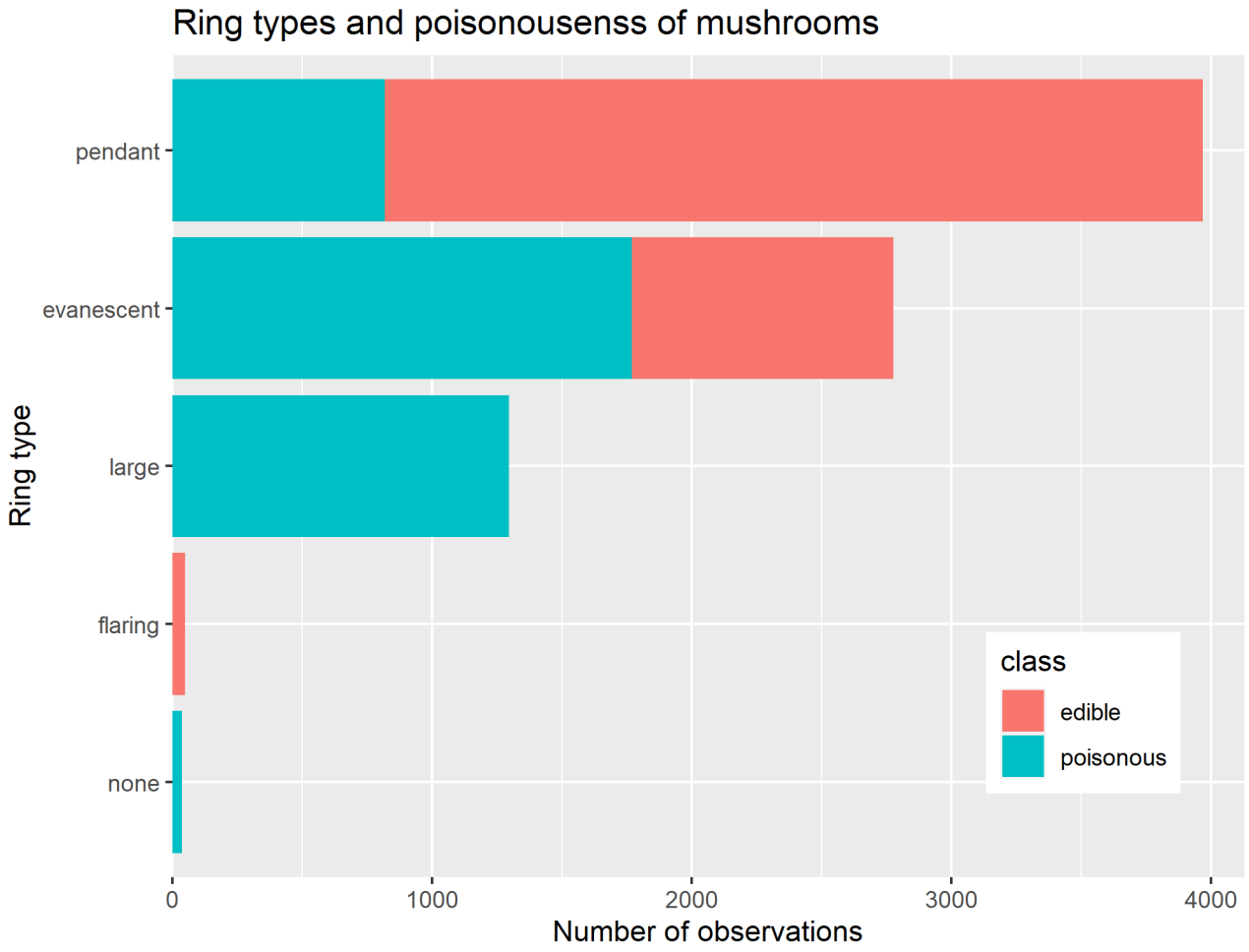
4.3 Spore print color

There is a tangible difference between spore prints that are black brown and chocolate. The first two colors are good indicators of edibleness. However, it might be hard to distinguish between brown and chocolate. If one picks a mushroom with chocolate spore print the consequences will almost certainly be deadly.



4.4 Ring type

As one can see, ring type might not be the clearest indicator of edibleness. However, it is a fairly distinguishable trait when picking mushrooms and therefore it was included. Nevertheless, having a pedant ring type indicates edibleness with roughly 80% accuracy. Then, however, more traits should be checked to ensure picking is not a poisonous mushroom.



5 Summary

This exploratory data analysis leads to a few interesting conclusions:

- Mushrooms with white or brown gill are almost always edible
- Mushrooms with no odor are almost always edible
- Mushrooms with spore print brown or black are almost always edible
- Mushrooms with pedant ring type are often edible