Group 8

Group Members

Dongwook Chun

Gahyun Lee

Kuan-Fen Liao

Ananthapadmanabhan Pillai

Juwon Song

**Mushroom Classification  
Safe to eat or deadly poison?**

BUAN 6356 Business Analytics with R

**Executive Summary**

This report provides analysis on mushroom’s traits, such as color, shape, and odors. We use 20 different traits to run the different analysis model to determine a specific mushroom is edible or poisonous. Methods of analysis include Logistic Regression, LDA, and Tree-based method. All models can be found in the R markdown file enclosed for the Models and Analysis section. Results of the data analyzed show our models present high accuracy and some of the traits bear high importance when determining whether the mushroom is toxic or not.

The report finds the relationship of different traits and edibility. The major area to be further consider is that not a single analysis has a 100% accuracy. Therefore, we must use more than one models to identify the toxic mushrooms.

**Project motivation/background**

Mushrooms are popular among Asia cuisines. People often ate wild mushrooms and got illness due to lack of knowledge in the mushroom. Though nowadays we can see different kind of mushrooms in the supermarket, we still have limited exploration on the thousand different kind of mushrooms.

From the data analytics perspective, we can divide a single mushroom species into different traits and use the data we collected to build the identification model. Our goal is to provide basic separation method to use as the fundamental knowledge in mushrooms.

**Data Description**

A mushroom, or toadstool, is the fleshy, spore-bearing fruiting body of a fungus. The summary and str give a basic understanding of the mushrooms dataset that contains 8,124 observations and 23 variables. The data is entirely nominal variables and the predictions are going to be on a field with only 2 values(classes)

Let us clean the data up a bit and then rename the factor levels so that they are more meaningful and look closer at their numbers

1. Class: The mushroom dataset is effectively created to answer one pressing need, that is to predict if the mushrooms are poisonous or edible, as indicated by the class variable

## Edible Poisonous

## 4208 3916

Clearly there is a large number of poisonous and edible mushrooms making this exercise more meaningful, the marks associated with the project also help with motivating us.

2-4. Mushroom Caps: the expanded, upper part of the mushroom; whose surface is the pileus

2. Mushroom Cap Shape: The six cap sizes in the dataset are flat, conical, convex, bell, sunken and knobbed

## Bell Conical Convex Flat Knobbed Sunken

## 452 4 3656 3152 828 32

Clearly a majority of the shapes are Convex or Flat

3. Mushroom Cap Surface: The four cap shapes in the dataset are scaly, fibrous, smooth and groves

## Fibrous Grooves Scaly Smooth

## 2320 4 3244 2556

This tells us that there are negligible numbers with Grooves but the remaining factor levels have significant populations.

4. Mushroom Cap Color: The ten cap colors in the dataset are yellow, buff, pink, gray, purple, green, red, cinnamon, white, brown.

## Brown Buff Cinnamon Green Grey Pink Purple Red

## 2284 168 44 16 1840 144 16 1500

## White Yellow

## 1040 1072

Brown has the largest share but Grey, Red, White and Yellow also have significant numbers.

5. Bruises indicates whether the mushroom tends to have bruises like (t=True, bruises=false)

## f t

## 4748 3376

There are large numbers with bruises and good numbers without bruises also.

6. Odor(smell): nine types of odor in the dataset are almond, anise, creosote, fishy, foul, musty, pungent, spicy and none.

## Almond Anise Creosote Fishy Foul Musty None Pungent

## 400 400 192 576 2160 36 3528 256

## Spicy

## 576

Most mushrooms have no smell and a large number are foul smelling, there are very few that are musky.

7-10. Mushroom Gills: a series of radially arranged (from the center) flat surfaces located on the underside of the cap.

7. Gill Attachments: the two types of gill attachments are attached and free for this dataset

## Attached Free

## 210 7914

A vast majority are free.

8. Gill Spacing: the two types of gill spacing are crowded and close

## Close Crowded

## 6812 1312

A majority are closed but a good number are crowded as well, there were no mushrooms with distant in this dataset.

9. Gill Size: Broad and Narrow are the two options in the dataset

## Broad Narrow

## 5612 2512

The broad to narrow ratio is upwards of 2:1.

10. Gill Size: Black, brown, buff, chocolate, green, pink, gray, orange, purple, red, white, yellow are the twelve options in the dataset. The following code give us better legibility

mush<-mush%>%mutate(gill.color=ifelse(gill.color=="n","Brown",

ifelse(gill.color=="b","Buff",

ifelse(gill.color=="h","Chocolate",

ifelse(gill.color=="g","Grey",

ifelse(gill.color=="r","Green",

ifelse(gill.color=="p","Pink",

ifelse(gill.color=="u","Purple",

ifelse(gill.color=="e","Red",

ifelse(gill.color=="w","White",

ifelse(gill.color=="k","Black",

ifelse(gill.color=="y","yellow","orange"))))))))))))

mush<-mush%>%mutate(gill.color=as.factor(gill.color))

## Black Brown Buff Chocolate Green Grey orange

## 408 1048 1728 732 24 752 64

## Pink Purple Red White yellow

## 1492 492 96 1202 86

A majority have brown, buff, pink or white colorations.

11-16. Stalk: the main support of the mushroom; it is topped by the cap.

11.Stalk shape: Two alternates in the dataset are enlarging and tapering

## Enlarging Tapering

## 3516 4608

There are good numbers with enlarging and tapering as well.

12.Stalk root: The four alternates in the dataset are bulbous, club, equal and rooted. There are also 2480 missing values.

## ? Bulbous Club Equal Rooted

## 2480 3776 556 1120 192

Bulbous and Equal to a lesser extend are the most occurring values in this field.

13.Stalk surface above ring: The four alternates in the data are fibrous, scaly, silky and smooth

## Fibrous Scaly Silky Smooth

## 552 24 2372 5176

An overwhelming majority are smooth but a significant number are silky as well

14.Stalk surface above ring: The four alternates in the data are fibrous, scaly, silky and smooth, like the previous field.

## Fibrous Scaly Silky Smooth

## 600 284 2304 4936

Like the last column, this column also has smooth as the majority with silky coming in a distant second

15.Stalk color above ring: Brown, buff, cinnamon, grey, pink, orange, red, yellow and white are the options in the dataset.

## Brown Buff Cinnamon Grey Orange Pink Red White

## 448 432 36 576 192 1872 96 4464

## Yellow

## 8

A majority are white with a good number being pink as well.

16.Stalk color below ring: Brown, buff, cinnamon, grey, pink, orange, red, yellow and white are the options in the dataset, like the previous field.

## Brown Buff Cinnamon Grey Orange Pink Red White

## 512 432 36 576 192 1872 96 4384

## Yellow

## 24

Echoing the values from the last column, white is the majority with pink in a distant second.

17-18. Veil: The tissue that connects the stem and the cap before the gills are exposed and the fruiting body develops.

17.Veil Type: This filed has only one value, partial and therefore will be eliminated from future analysis.

## Partial

## 8124

18.Veil.color: Brown, orange, white and yellow are the possible values for this field in the data.

## Brown Orange White Yellow

## 96 96 7924 8

An overwhelming majority are white.

19-20. Ring: A skirt-like ring of tissue circling the stem of mature mushrooms. The ring is the remnant of the veil.

19.Number of Rings: The data has records with 1,2 or no rings.

## 0 1 2

## 36 7488 600

An overwhelming majority have one ring.

20. Ring types: The fives values for ring type in the dataset are evanescent, flaring, large, pendant and none

## Evanescent Flaring Large None Pendant

## 2776 48 1296 36 3968

Pendant, Evanescent and Large are the significant groups in this field.

21. Spore Print Color: The nine values that this field takes in the data are black, brown, buff, yellow, white, orange, purple, chocolate and green.

## Black Brown Buff Chocolate Green Orange Purple

## 1872 1968 48 1632 72 48 48

## White Yellow

## 2388 48

Black, Brown, White, Chocolate have significant numbers in this field.

22. Population: Abundant, clustered, numerous, scattered, several and solitary are the different values for this field

## Abundant Clustered Numerous Scattered Several Solitary

## 384 340 400 1248 4040 1712

A majority occur in several and a surprising number are solitary.

23. Habitat: The seven possible values for this field in the data are grasses, leaves, waste, meadows, paths, waste and woods.

## Grasses Leaves Meadows Paths Urban Waste Woods

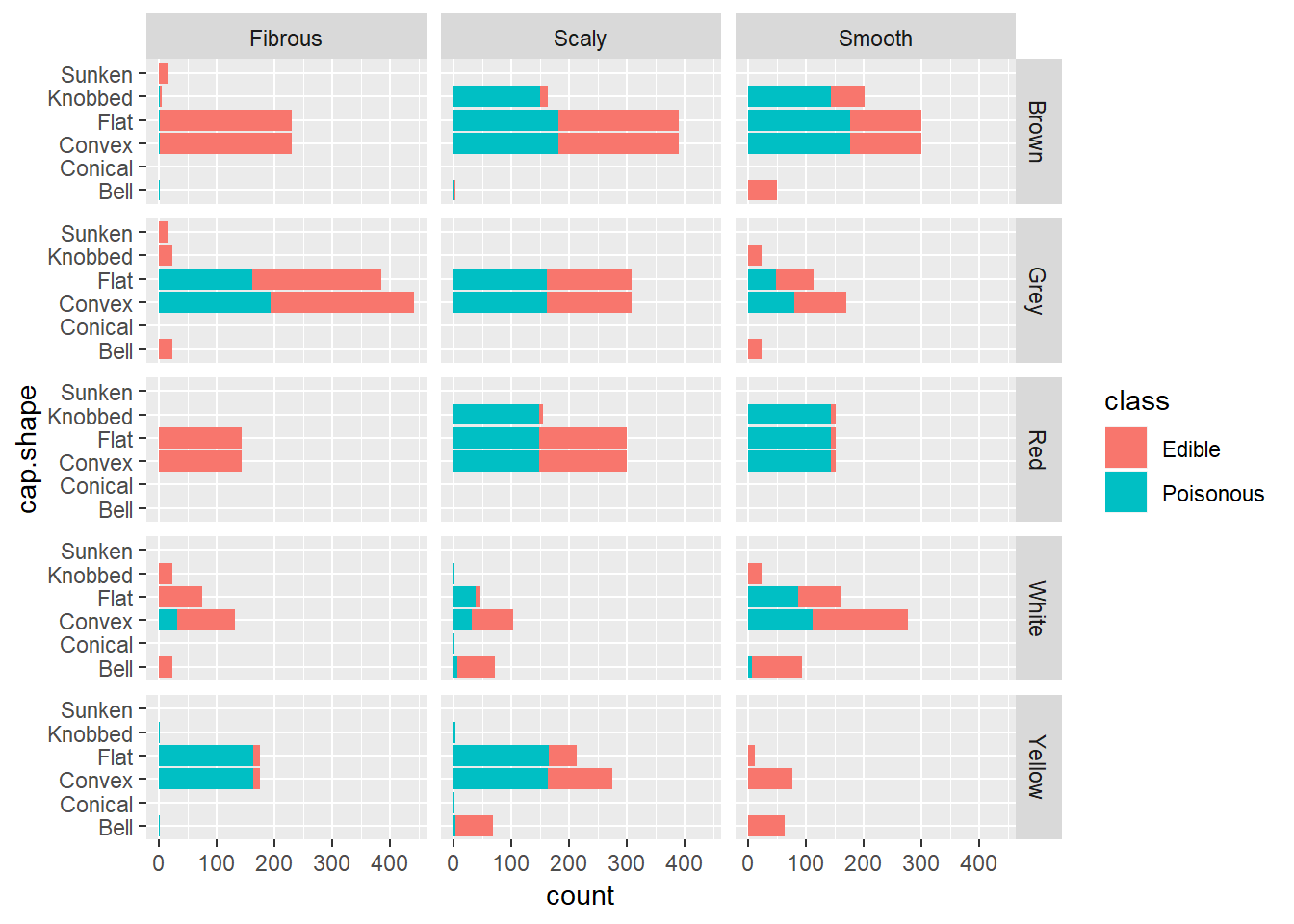
## 2148 832 292 1144 368 192 3148

Grasses and Woods seem to be the most favorable habitats.

**Data Exploration**

Grasses and Woods seem to be the most favorable habitats. Our primary concern is the class (whether the mushroom is poisonous or not) and hence this will be constantly there in all our explorations.

First let’s see if the cap size, cap shape and cap color



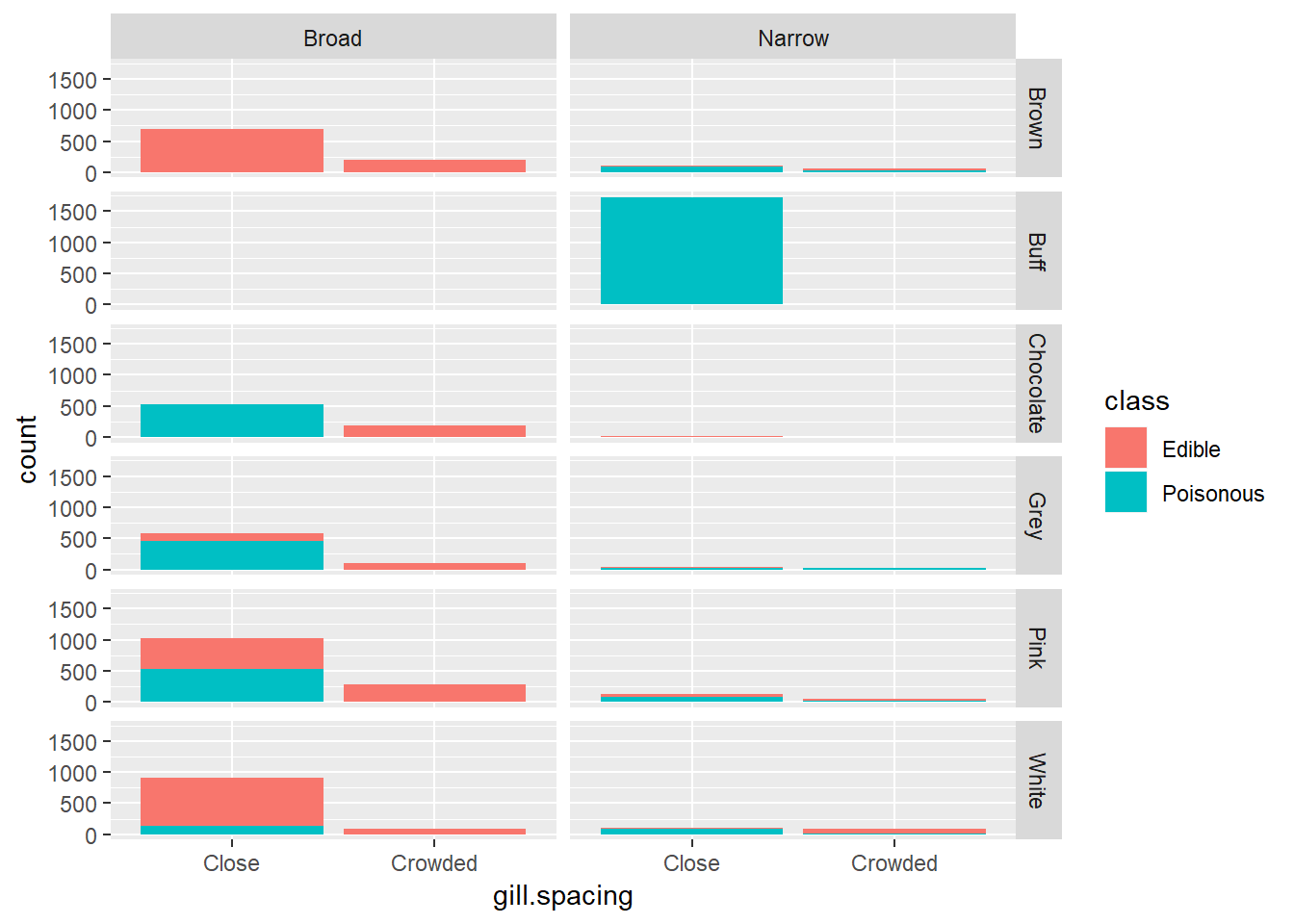
It is evident that mushrooms whose cap shape is bell like are much more likely to be edible, therefore cap shape should be considered when building a model.

The stark difference between fibrous and the other two types of cap surface in terms of poisonous mushrooms makes it evident that cap surface should be considered for further analysis too.

The case of cap shape being flat or convex adds complexity though as the trend it shows with fibrous cap surface and red or brown cap color of near zero probability of poison is bucked by the smooth case where the probability is near certain. This could possibly mean an interaction between the variables.

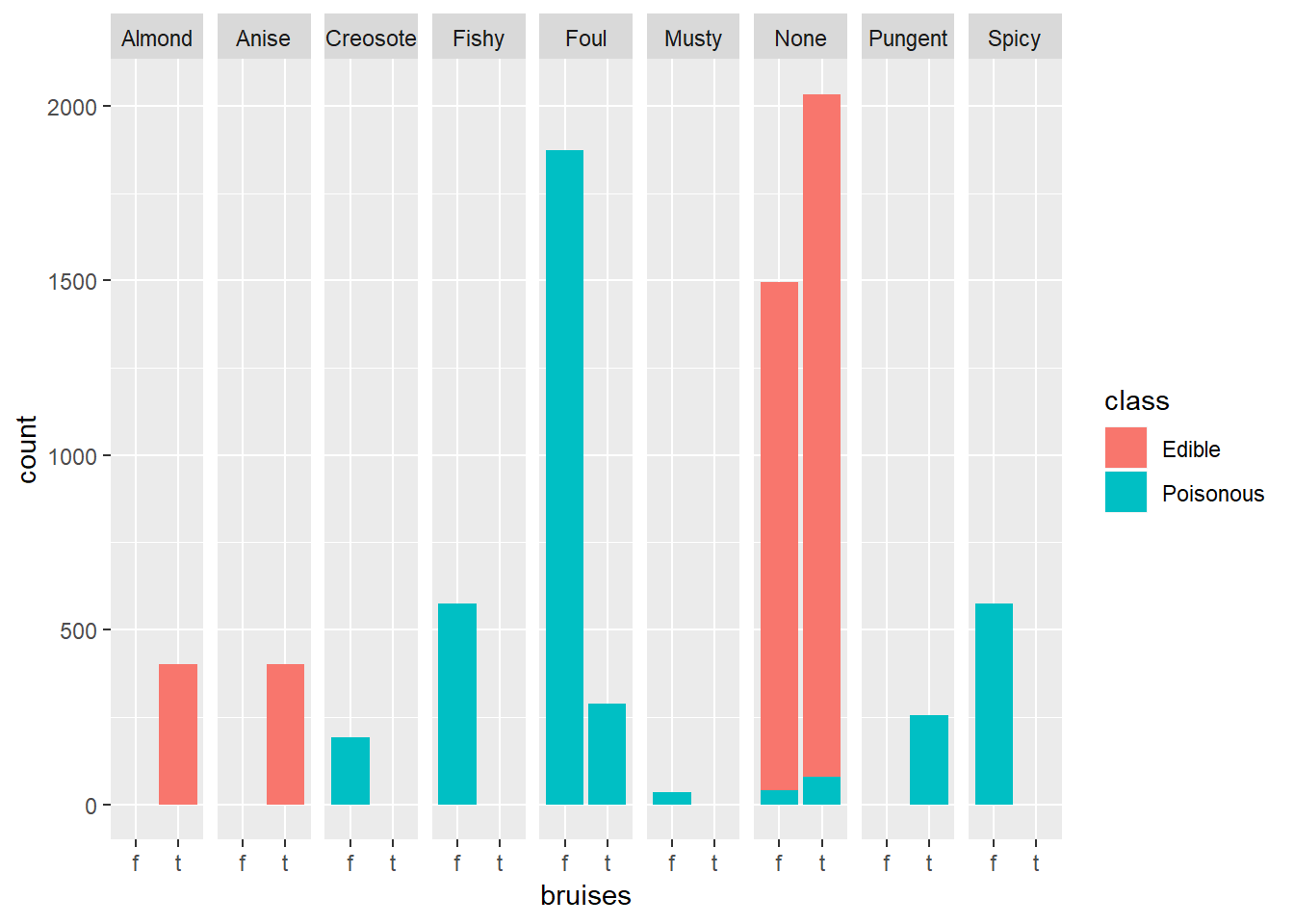
Although color doesn’t have any trends that stand out on its own, when it is combined other fields it is clearly trends and hence should not be eliminated.

Next, let’s look at the gill attributes.



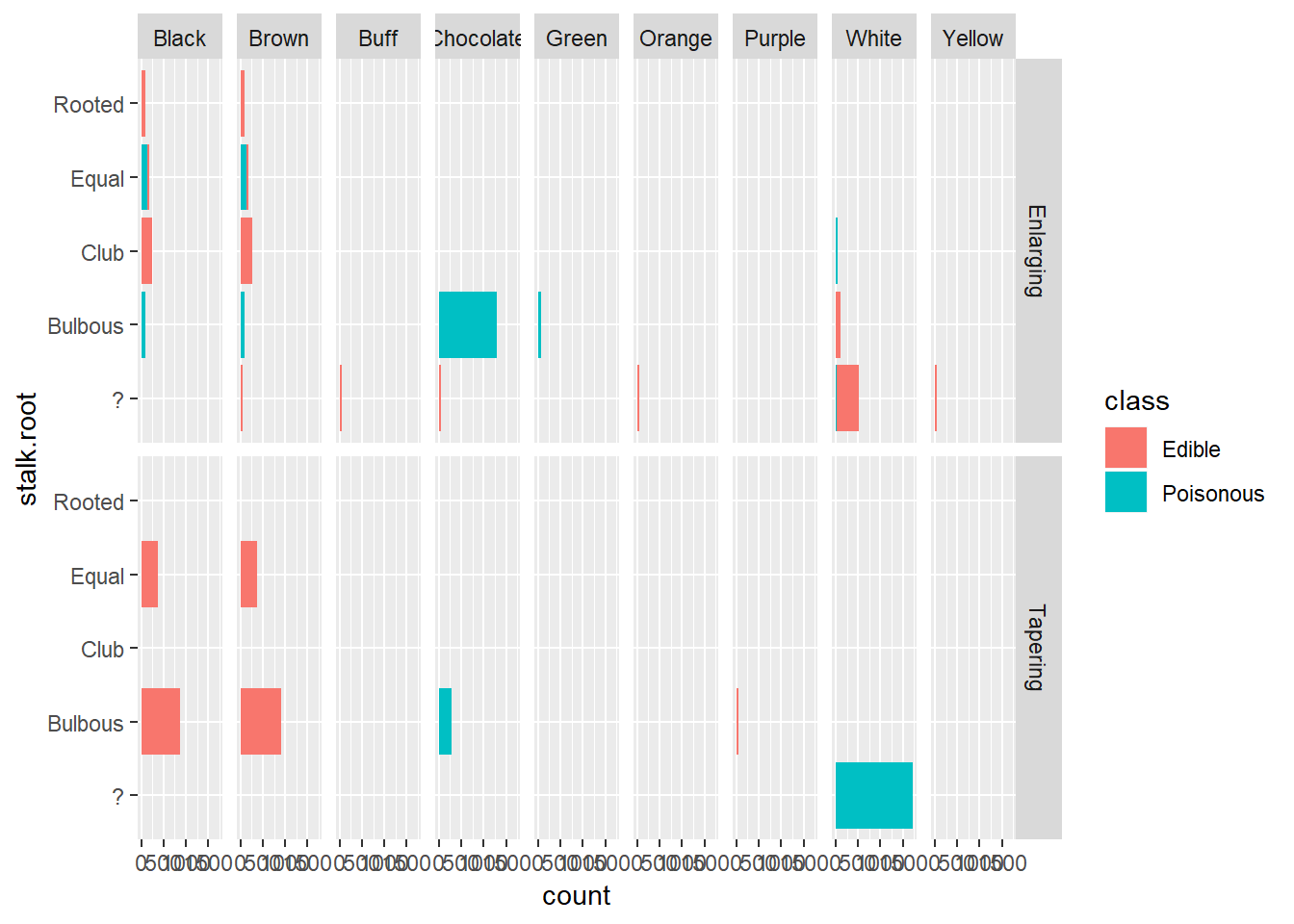
Broad gill size clearly seems to influence the edible/poisonous trait of the mushrooms. Buff colored mushrooms also seem to be much more likely to be poisonous. Crowded gill spacing also seems to have a positive correlation with edibility. Clearly, these factors should be considered for the model.

Next let’s look at bruises and odor



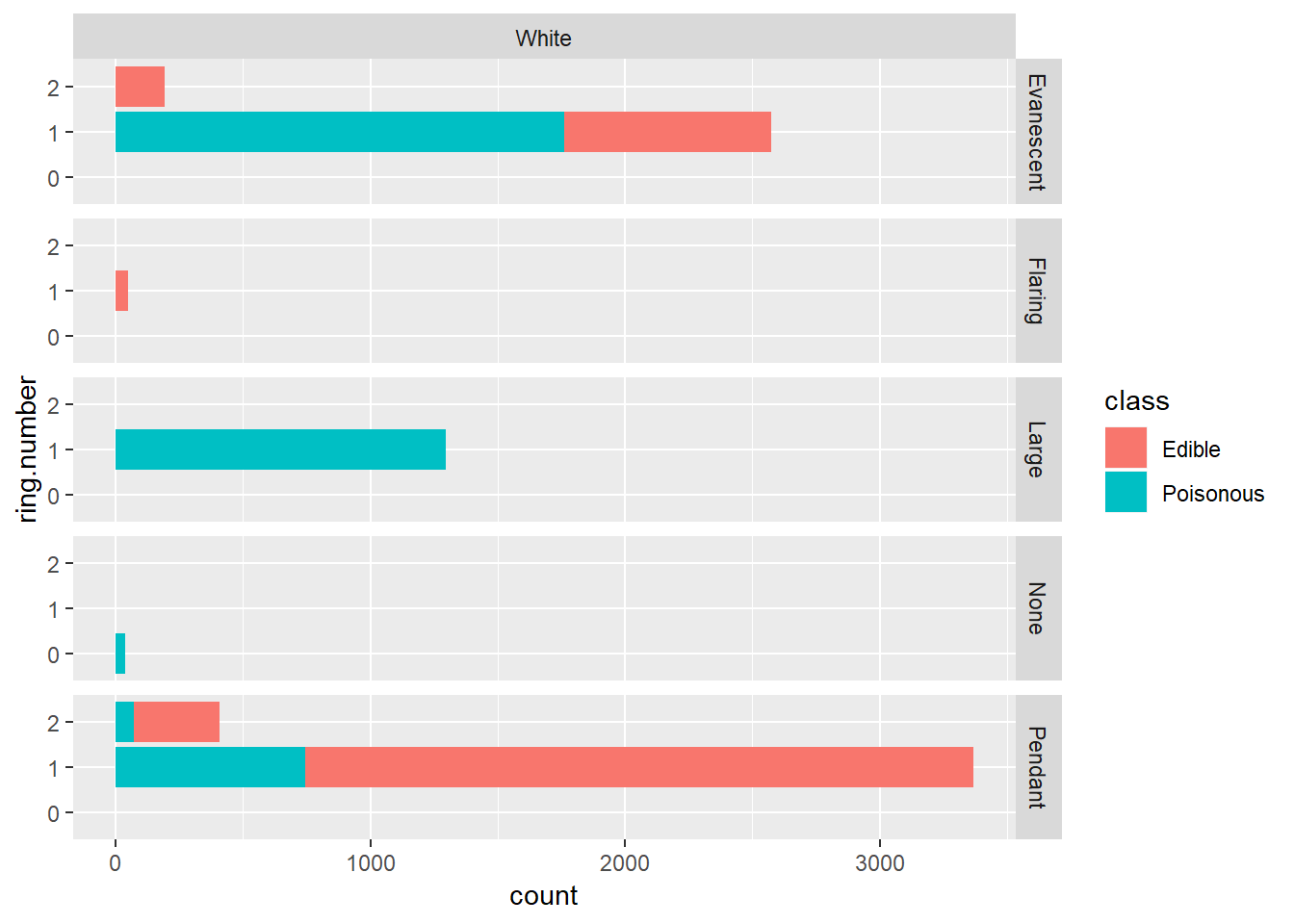
Foul smelling clearly seems to be poisonous, whereas those with no odor were more likely to be edible. Therefore, we can conclude that this variable should be used in model building. The type of order and their interaction also inform us before the model selection stage.

Let’s look at a new graph to explore the rest of stalk attributes with one other variables



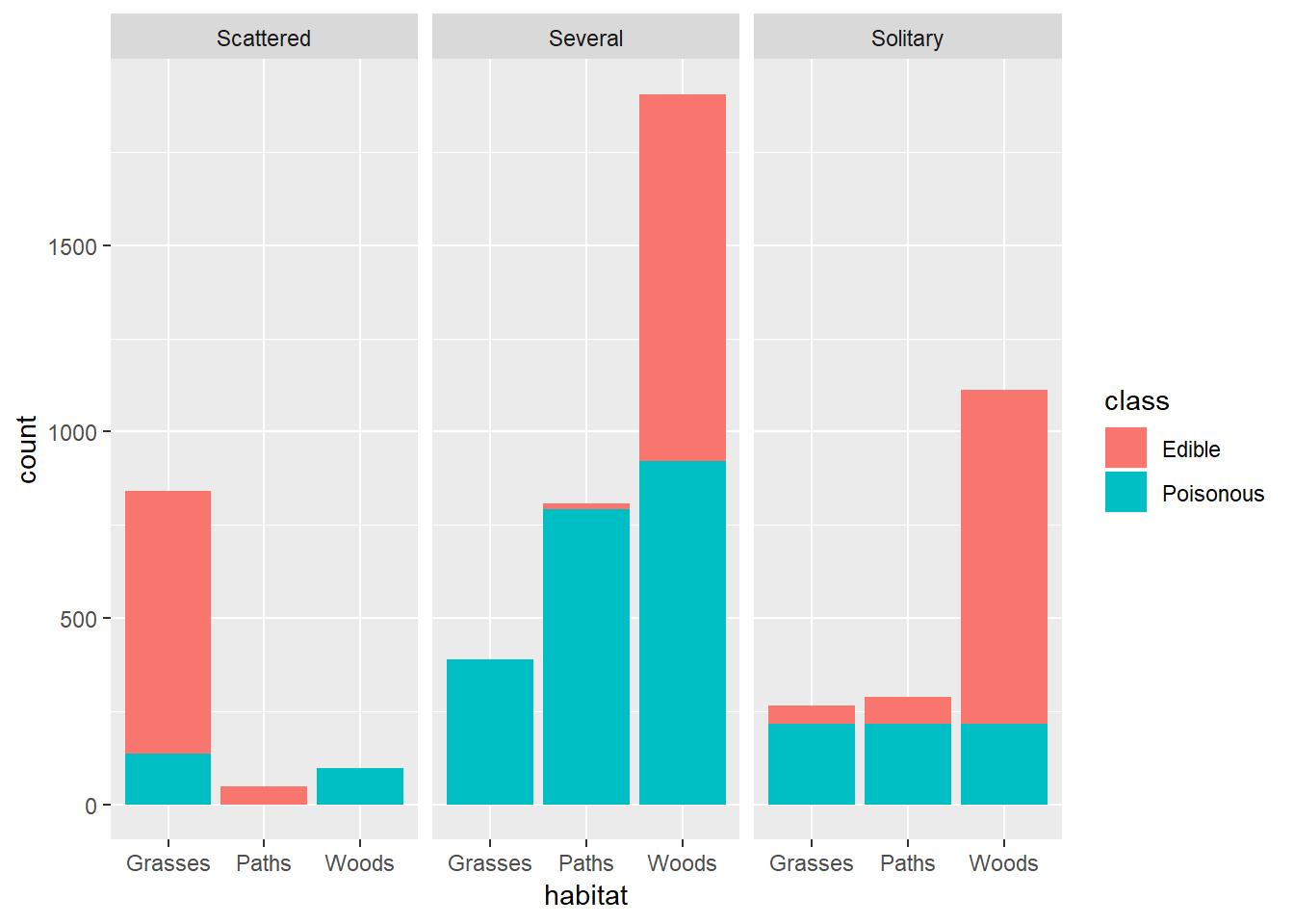
It is more than evident that bulbous mushrooms are more likely than not to be edible and the opposite holds true for stalks that are chocolate in color. It is also evident that the interaction of stalk shape with other variables seems to be significant, hence it is best if we hold on to these variables for future analysis.

Let’s next turn to the ring and veil attributes now



This graph looks like the previous ones and tells us to keep the variables for further review as it also shows clear patterns and doesn’t suggest that any of these are not useful.

Looking at population and habitat.



This graph indicates that scattered mushrooms in grassy areas are very likely to be edible, the same can be said for solitary mushrooms in the woods. This means that the variables should be considered for model building.

This is some of the preliminary exploratory analysis used to understand the data and prepare ourselves for the prediction model building stage.

**Model and analysis**

Refer attached R markdown file.

**Findings and managerial implications**

According to the models we have built, there were several findings about significant variable.

1. Logistic Regression Model
   1. Top 7 Important Variables are:  
      gill size  
      gill spacing  
      stalk root  
      stalk surface above ring  
      odor  
      ring type  
      population
2. Linear Discriminant Analysis
   1. Top 7 Important Variables are:  
      gill color   
      gill size

bruises  
ring type  
stalk surface above ring

stalk surface below ring  
stalk root

1. Decision Tree
   1. Top 5 Important Variables are:  
      odor  
      spore print color  
      gill color  
      stalk surface above ring  
      stalk surface below ring
2. Random Forest
   1. 6 Important Variables are:  
      odor  
      spore print color  
      stalk color below ring  
      stalk surface above ring  
      habitat  
      cap surface

We have found some information from the data exploration. For example, if a gill color of a mushroom is buff(b) and green(r), they are all poisonous. We, however, can make decisions more easily from looking at these important variables first. According to above findings, we now know what are more relevant to classify mushrooms.

The model with highest accuracy is random forest model. This method is not only considerably effective, but also gives us intuitive knowledge to decide whether a mushroom is edible or not, looking at the 6 specifications the mushroom has. For instance, if a mushroom’s odor is not among almond(a), anise(l), or none(n), which means that if the odor is creosote(c), foul(f), musty(m), pungent(p), spicy(s), and fishy(y), they are poisonous. If the odor is either a, l, or n, you can follow the stem and move to the next node. Repeat this process until you arrive any bottom of the graph and there is your answer.

This is especially handy when it comes to application. You can just bring the graph of random forest model with you when going for mushroom hunting: you can immediately identify the mushroom you are looking at and decide whether to pick or not. Although you cannot smell well, we highly recommend you should hire someone who can smell on behalf of you and take this method to make decision on whether to pick/eat the mushroom or not.

**Conclusion**

As written above, we tried to apply several models to classify mushroom. Each way has advantage and disadvantage all together. Because the dataset has all categorical variables, we have tried to change them to numerical variable to conduct models on regression models. In this process, there were several problems such as multicollinearity and distortion relationship between variables. Therefore, we recommend the model based on tree-method which can maintain the characteristic of categorical variable and present high model performance to this classification case.