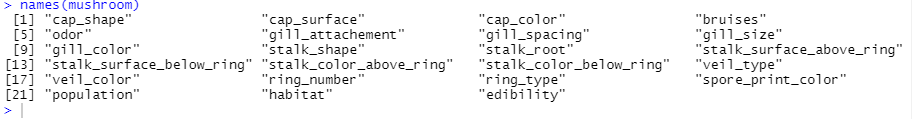
**Introduction**

A mushroom also known as toadstool is an enlarged fleshy fruiting body of a fungus usually produce on the ground or on its food source. Mushrooms are beneficial for human body but several mushrooms are not edible. It is true that several mushroom types may use to cure cancer, but some of them may be toxic and contain deadly viruses and leads to infectious diseases. Usually wild mushrooms can be highly dangerous and not edible. It is difficult to identify the deadly poisoned and edible mushrooms manually on the basis of few features. Therefore, automate this process and build a model can be helpful to identify and avoid poisonous mushrooms, on the basis of certain features and characteristics of the mushrooms. In this report, we will classify the mushroom as edible or poisonous by analyzing the different attributes and properties of a mushroom. We will analyze the behavioral features namely cap-shape, cap-surface and cap-color, gill and stalk, odor, population, veil-type and habitat of the mushrooms.

**Data Collection**

Mushrooms are utilized widely in cooking, in numerous foods, but identify the poisonous requires attention to detail. Therefore, automate the identification process using features of the mushrooms themselves and Machine Learning (ML) model, it can be helpful for people to stay safe from poisonous mushrooms. We will be using the mushroom dataset, which was provided for assessment. Download the file “mushrooms.csv” from Blackboard and save it the working directory, and then load the data in your R environment. By analyzing at the several attributes of a mushroom, we are going to predict either the mushroom is poisonous or edible. Total features in mushrooms dataset are shown in Fig. 1.

Fig. 1. Attributes in mushrooms dataset

The dataset consists 23 attributes and 8124 observations. The entire dataset entries are meaningful represented by only single character. The attribute and the acronyms are shown below in the table, each name is separated with comma (,). It will help to understand the meaning of each entry and attribute.

Attributes and observations Information:

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?:** yes=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

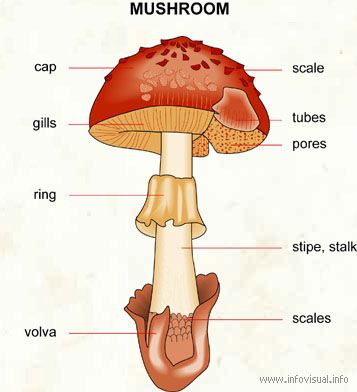
20. **spore-print-color:** black=k, brown=n, buff=b, chocolate=h, green=r, orange=o, purple=u, white=w, yellow=y

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. **Edibility:** Class of the mushroom; poisonous or edible. Edible = e, poisonous = p

Further, we are going to present some domain knowledge related to mushrooms which will increase the understandability of the dataset features and showing mushroom’s parts as mentioned above in dataset.



**Cap:** Top of the mushroom (often like a small umbrella). Mushroom caps may have different colors but usually are yellow, brown, and white.

**Gills, Pores, or Teeth:** These are structures similar to a fish gills and always appear under the mushroom's cap.

**Ring:** Membrane which circling the stem under the cap.

**Stem or Stipe:** The stem is the main body, usually holds the cap above ground tall structure that holds the cap high above the ground.

**Volva:** The volva is the protective veil that remains after the mushroom sprouted up from the ground. As the fungus grows, it breaks through the volva.

**Spores:** Microscopic seeds acting as reproductive agents; they are usually released into the air and fall on a substrate to produce a new mushroom (Kalač, 2013).

**Data Preparation**

One of the essential phase of any machine learning task is ensuring that the data is in the optimal state for processing, prior to building any models. Few standard tasks will be implemented for data preparation as follows:

* Data Cleaning: Identify errors, corrections and no trivial/redundant data in our dataset
* Data Transformation: Change the scaling of attributes.
* Feature Selection: Select important and most relevant features

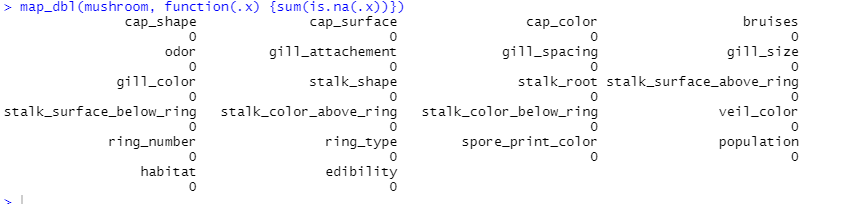
**Data Cleaning**

The initial task in data preprocessing is data cleaning. Discover how to transform messy data into clean by identifying outliers, and identifying and handling missing values with statistical and modeling techniques.

We can ensure that dataset has no missing data, redundant records, and outliers through coding in Rstudio.

**Remove Missing values**

We performed missing value analysis, no missing value found in any attribute.



**Remove Duplicate values**

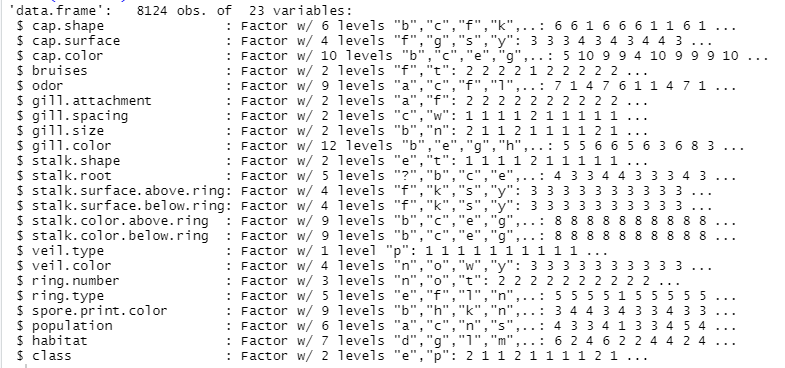
Duplicate values analysis test (distinct) shown that no duplicate values are in dataset, are attributes contains only distinct values.

**Outliers**

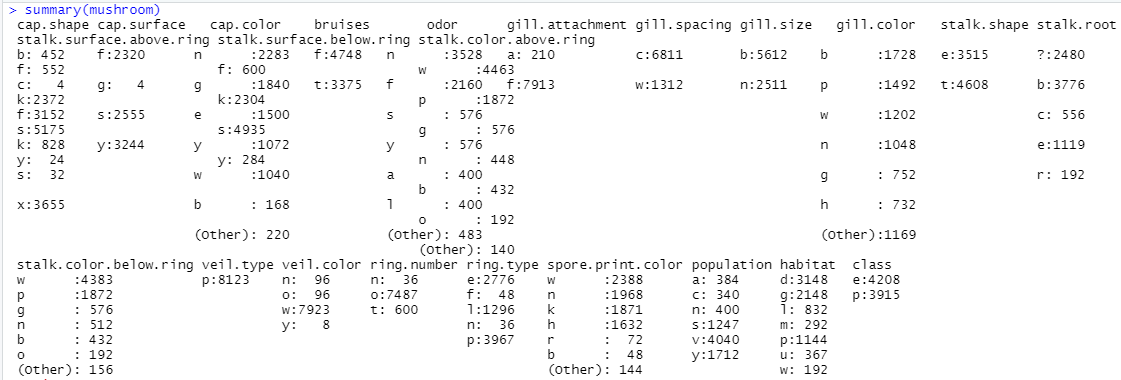
Outliers are extreme values that we come across, where they may be influential to the model or not. When it comes to categorical data (say Gender: as in male and female). There's no way of any outlier detection in that.

**Data Transformation**

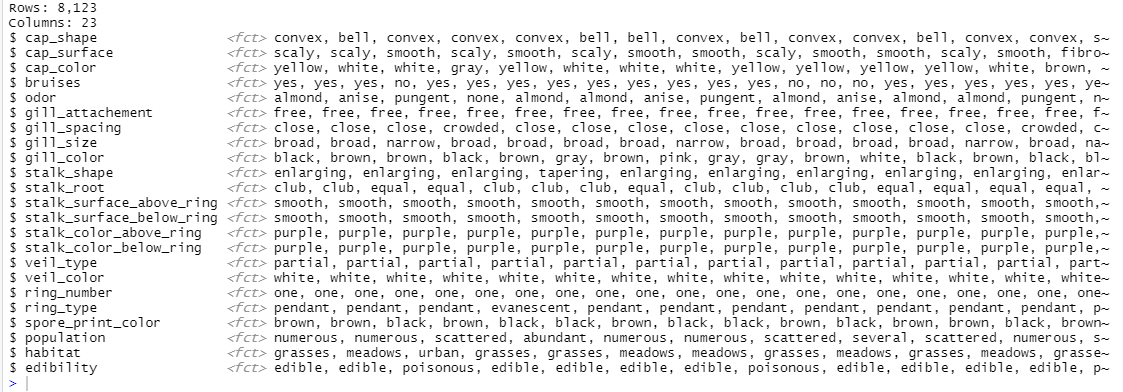
View the dataset, once loaded it in. we noticed that the data is primarily in string format, we need to convert it in to factor to process further.



Thereafter the summary of the dataset can be seen. Here, each attribute and the possible values of each category with the count are seen.



When we are checking the summary, it seems less meaningful with the initials of all values in features, we have seen above each name is contains meaningful information related to the mushrooms. Rename to each variable will enhance the understanding of data. Now dataset looks as follows:

We will run an exploratory analysis. This dataset about any mushroom is a safe to eat or deadly poison. I will examine the data and I will create a machine learning algorithm which is detects any mushroom is edible or poisonous due to its specifications like cap shape-color or gill color-size. Let's examine the data.

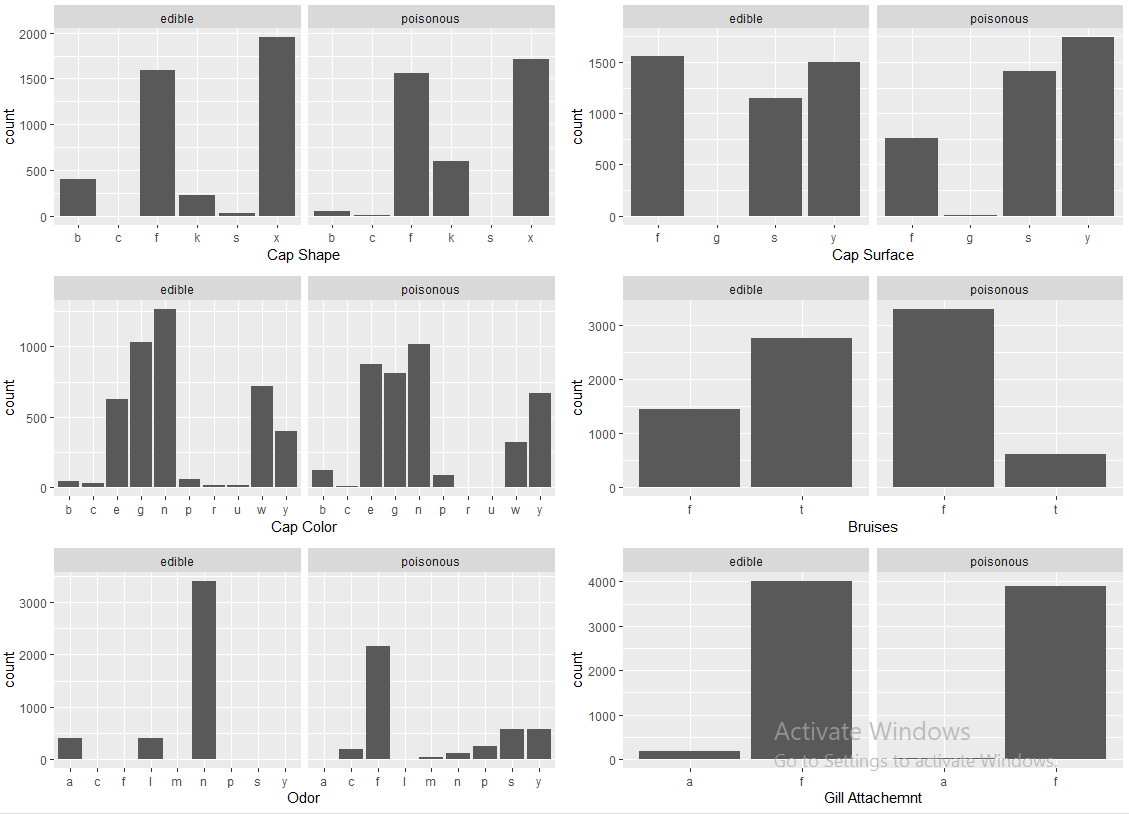
**Data Exploration**

Now, I have plotted the histograms of each category and split them into two graphs according to their edibility. The objective of doing do 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 five attributes - cap shape, cap surface and cap color, bruises and odor are plotted below.

The information of interest is finding the attributes which have count only in one of the charts - edible or poisonous. In Cap Shape, c (conical) is present only in poisonous whereas s (sunken) is only present in edible. Similarly, for Cap Surface, g(grooves) is present only in poisonous. In Cap Color, r(green) and u(purple) are only present in poisonous. This exclusiveness means that if the cap surface is conical the mushroom is poisonous. In summary, I have tried to find the deciding attributes that play a role in deciding whether a given mushroom is edible or poisonous.

The bruises attribute have distribution in both classes and does not contribute in decision. The odor attribute seems to have more exclusiveness as compared to any other attribute encountered so far. If the odor is a(almond), c(creosote), or l(anise) the mushroom is edible and if the odor is m(musty), p(pungent), s(spicy) or y(fishy) the mushroom belong to the poisonous class. This is the best example of what is strived for in this analysis. This attribute seems to have more weightage in deciding the class of the mushroom.

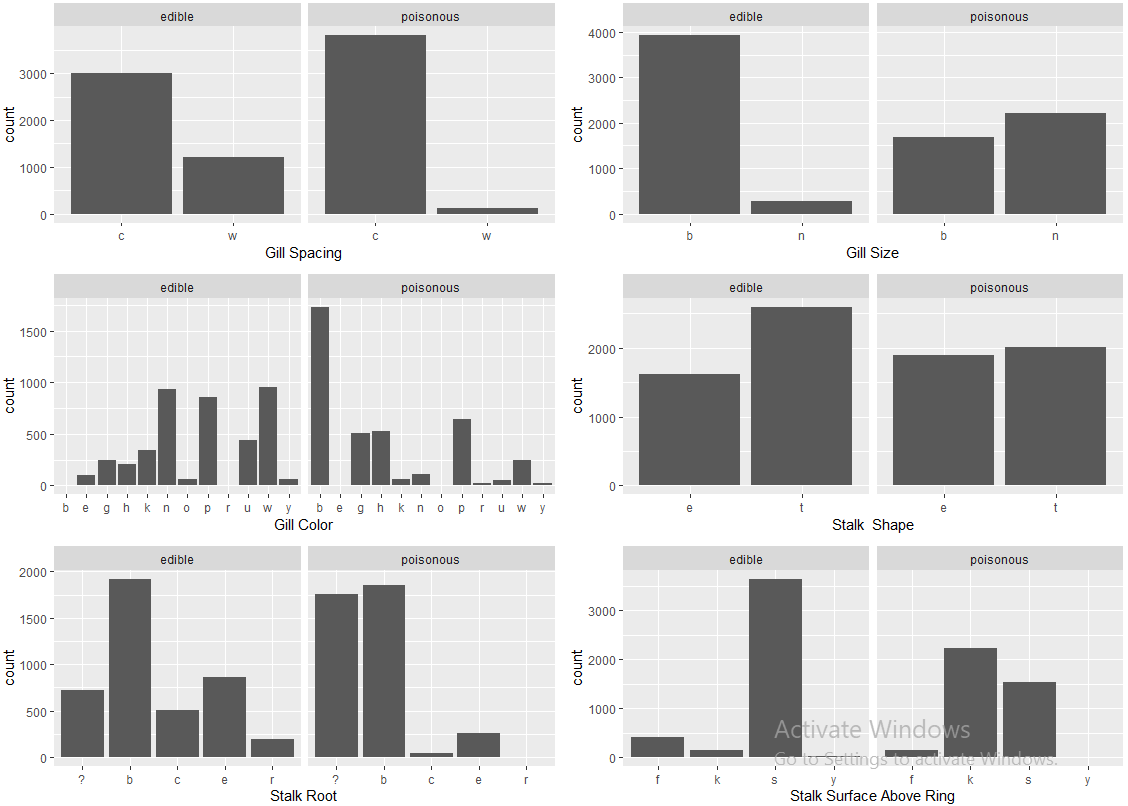
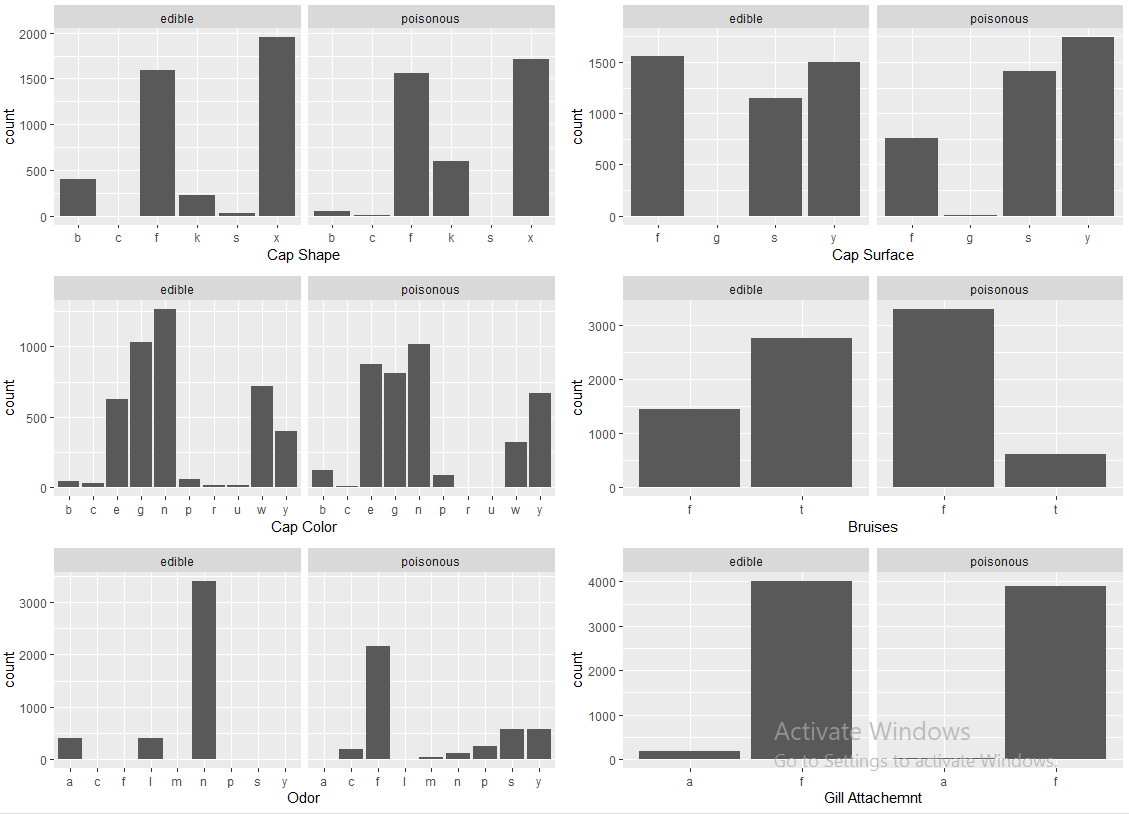
The Gill attachment, gill spacing and gill size seem to have distribution in both classes with no exclusiveness at all. As for gill color, there is exclusiveness as follows - the color b(buff), e(red) and o(orange) implies edible and r(green) implies poisonous. There is exclusiveness but there is distribution in other cases. However, the contribution of gill color cannot be ignored.

The stalk shape does not seem to contribute to decide the class of mushroom. On the other hand the stalk root having (rooted), is poisonous.

The attributes related to stalk surface - either above or below the ring have distribution in both the classes and there is no exclusiveness for both these parameters. The stalk color above ring and below do have exclusiveness. The exclusiveness is elaborated here.

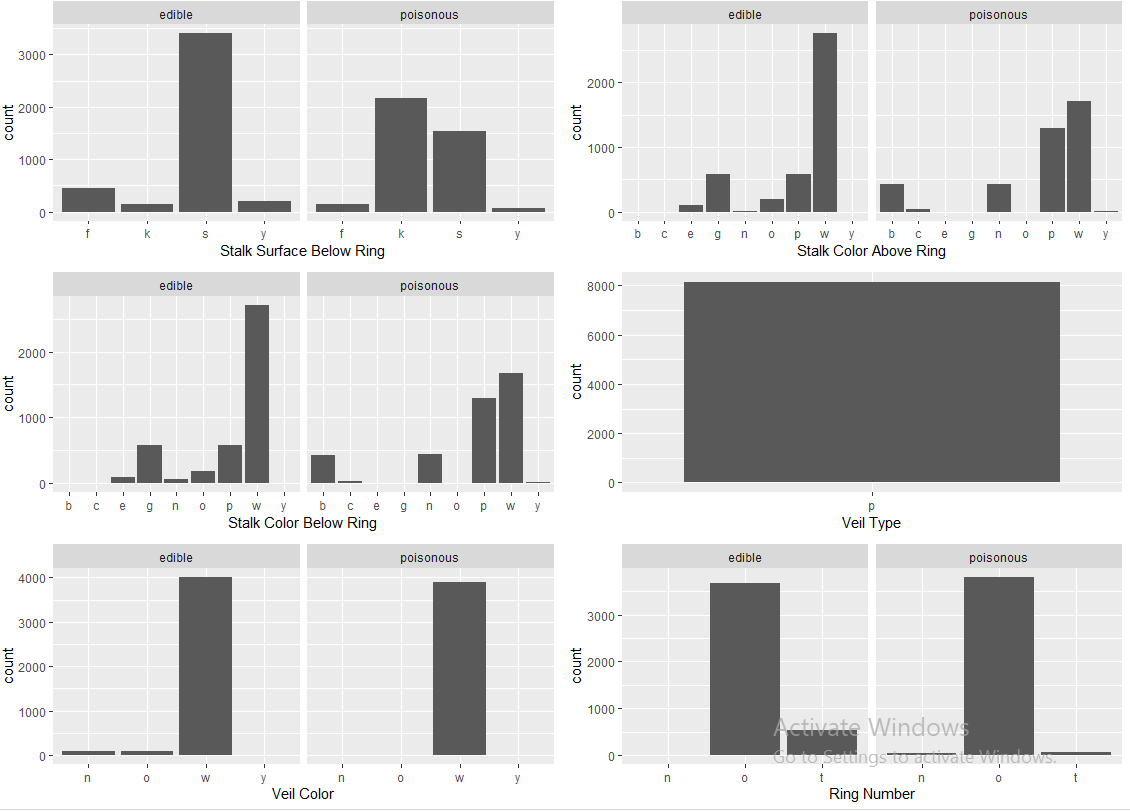
1. If the stalk color above ring is e(red), g(grey) or o(orange) implies an edible mushroom. On the other hand if the stalk color above ring is b(buff), c(cinnamon) or y(yellow) implies that the mushroom is poisonous.

2. If the stalk color below ring is e(red), g(grey) or o(orange) implies an edible mushroom. On the other hand if the stalk color below ring is b(buff), c(cinnamon) or y(yellow) implies that the mushroom is poisonous.

 The attributes related to stalk surface - either above or below the ring have distribution in both the classes and there is no exclusiveness for both these parameters. The stalk color above ring and below do have exclusiveness. The exclusiveness is elaborated here.

1. If the stalk color above ring is e(red), g(grey) or o(orange) implies an edible mushroom. On the other hand if the stalk color above ring is b(buff), c(cinnamon) or y(yellow) implies that the mushroom is poisonous.

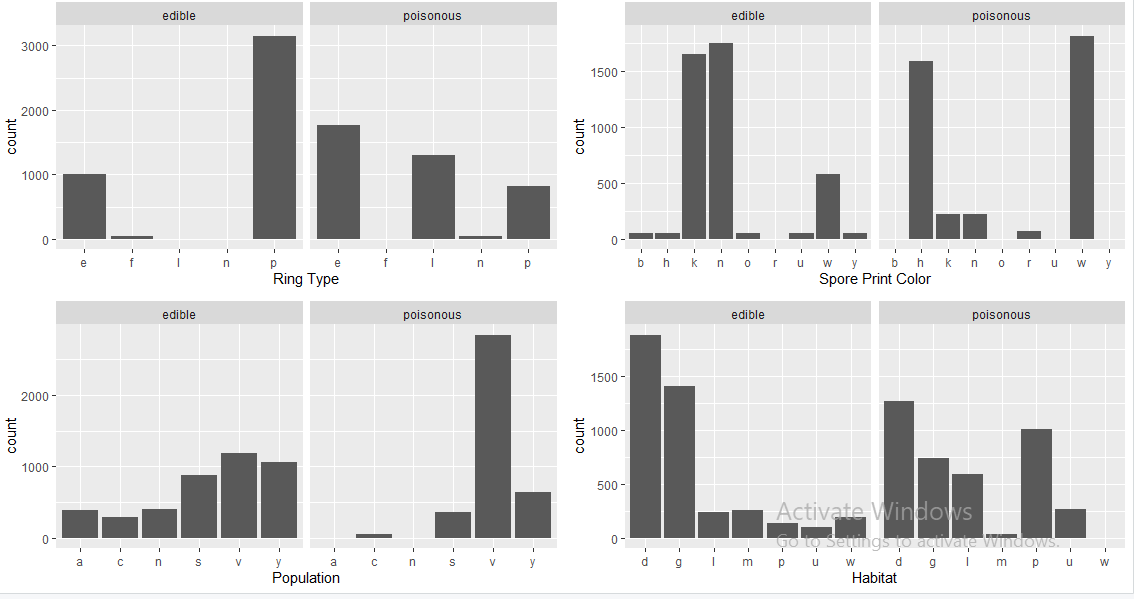
2. If the stalk color below ring is e(red), g(grey) or o(orange) implies an edible mushroom. On the other hand if the stalk color below ring is b(buff), c(cinnamon) or y(yellow) implies that the mushroom is poisonous.



The attributes related to stalk surface - either above or below the ring have distribution in both the classes and there is no exclusiveness for both these parameters. The stalk color above ring and below do have exclusiveness. The exclusiveness is elaborated here.

1. If the stalk color above ring is e(red), g(grey) or o(orange) implies an edible mushroom. On the other hand if the stalk color above ring is b(buff), c(cinnamon) or y(yellow) implies that the mushroom is poisonous.

2. If the stalk color below ring is e(red), g(grey) or o(orange) implies an edible mushroom. On the other hand if the stalk color below ring is b(buff), c(cinnamon) or y(yellow) implies that the mushroom is poisonous.



The veil type partial for all mushrooms as confirmed later and hence it had no contribution at all in deciding the class of the mushroom. This type of attributes is an excellent example of having data but of no use in analyzing the dataset. It would be better to get rid of such attributes, a concept called dimensionality reduction. Dimensionality reduction plays a very important role in speeding up the analysis when the datasets have large number of attributes and/or observations.

The veil color n(brown) and o(orange) suggests that the mushroom belongs to edible class and y(yellow) suggests that it belongs to poisonous class.

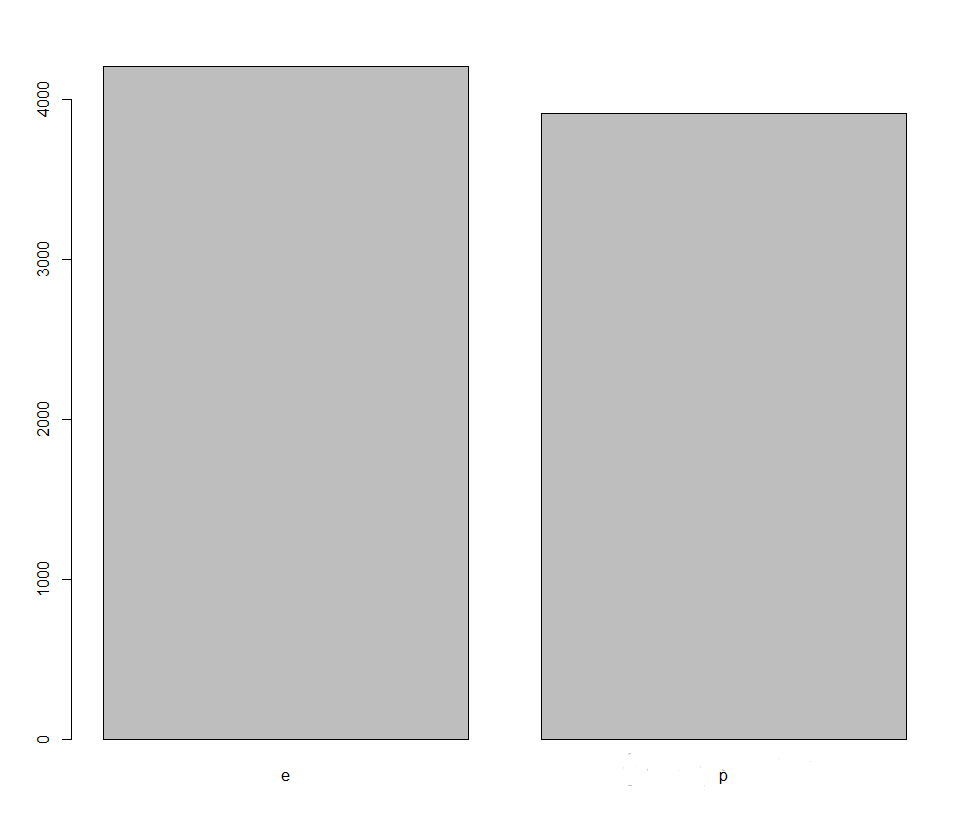
The number of ring has only one possible way to help in deciding.

The spore print color have exclusiveness. If the spore print color is b(black), o(orange), u(purple) or y(yellow) the mushroom belongs to edible class. Rest of the spore print colors are distributed in both types of classes.

The population attribute also has exclusiveness as follows - if the population attribute is a(abundant) or n(numerous), the mushroom is edible.

The last attribute of habitat has only one parameter that has exclusive nature. If the habitat is w(waste), then the mushroom is poisonous.

It is worth noting that the observations made from these histograms may be questionable. The reason for this being, the limitation of the histogram itself. It cannot be guaranteed from the histogram alone the absence for a particular bar. The count on the y-axis is considerably high and hence minor counts are expected to be so small that they almost do not appear. So, it is assumed here that the attributes which do not seem to have any count are considered that not even a single value falls under that category.

By summarizing the class distribution, we can understand that target attribute is almost, we don’t need to balance this attribute, according to the machine learning algorithms we have same probabilities to both classes for classifications. We only focus if we have imbalance class distributions!

**Multivariate Analysis**

I have taken in to account two attributes along with class and how they contribute the edibility of the mushrooms. The purpose of these plots is to find the combined exclusivity of two attributes in deciding the edibility of mushroom. The observations are noted after each graph.