DATASET LINK:

https://www.kaggle.com/uciml/mushroom-classification

A. Describe the dataset in your own words.

**Objective**

Classify whether if a mushroom is safe to eat or poisonous?

**Overview**

23 species of gilled mushrooms from the Agaricus and Lepiota family are observed in the dataset. The dataset contains 23 features of mushroom and the first feature is whether the mushroom is edible or poisonous.

**Attributes and observations**

The dataset contains 23 features and 8124 observations, all are categorical (factor) data.

**Missing values**

There are 2480 missing values in the dataset, where the missing values is denoted as “?” and are found in “stalk-root”, the 11th column.

**Distribution of edible/poisonous mushroom**

-- edible : 4208 observations (51.8%)

-- poisonous : 3916 observations (48.2%)

-- total : 8124 observations (100.0%)

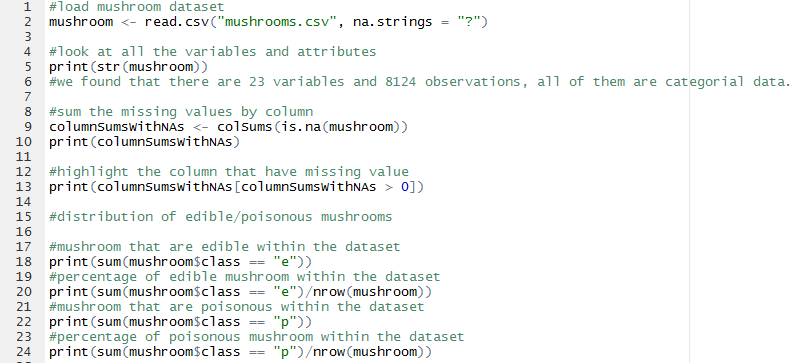
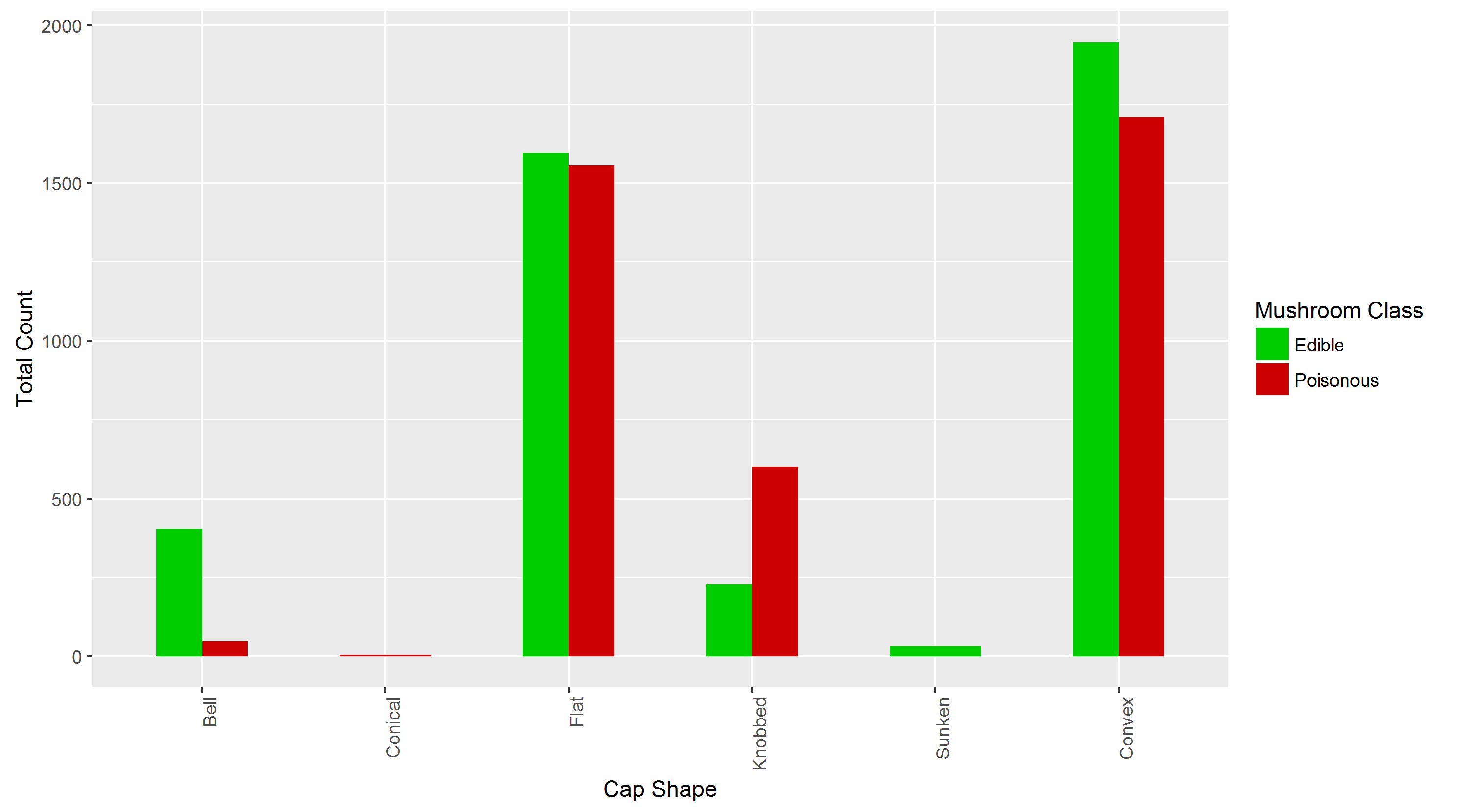
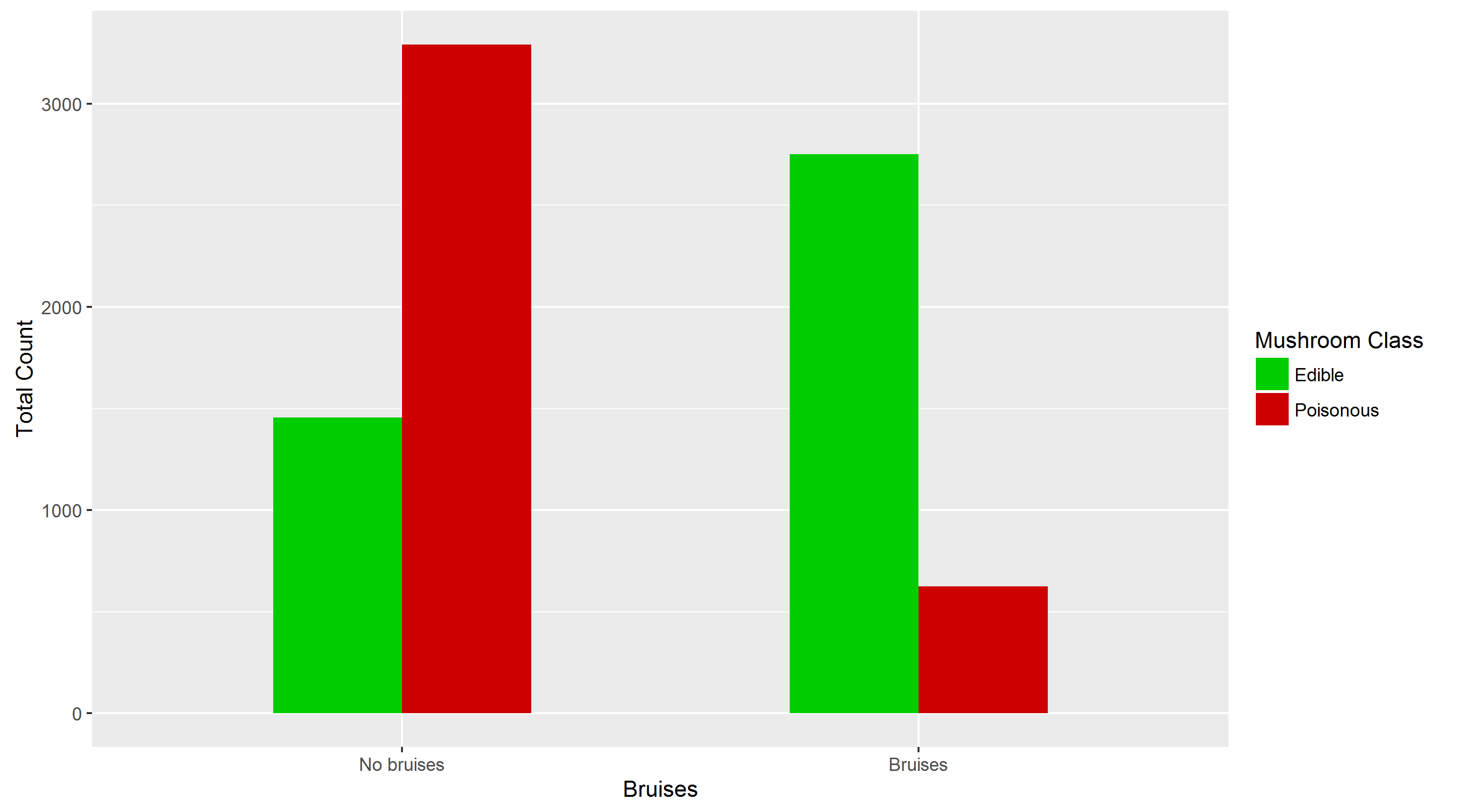


Figure 1. Code we used for describing the mushroom dataset.

B. What possible insights can be obtained from mining the chosen dataset?

To obtain common features that are found in edible and poisonous mushrooms for building a prediction model.

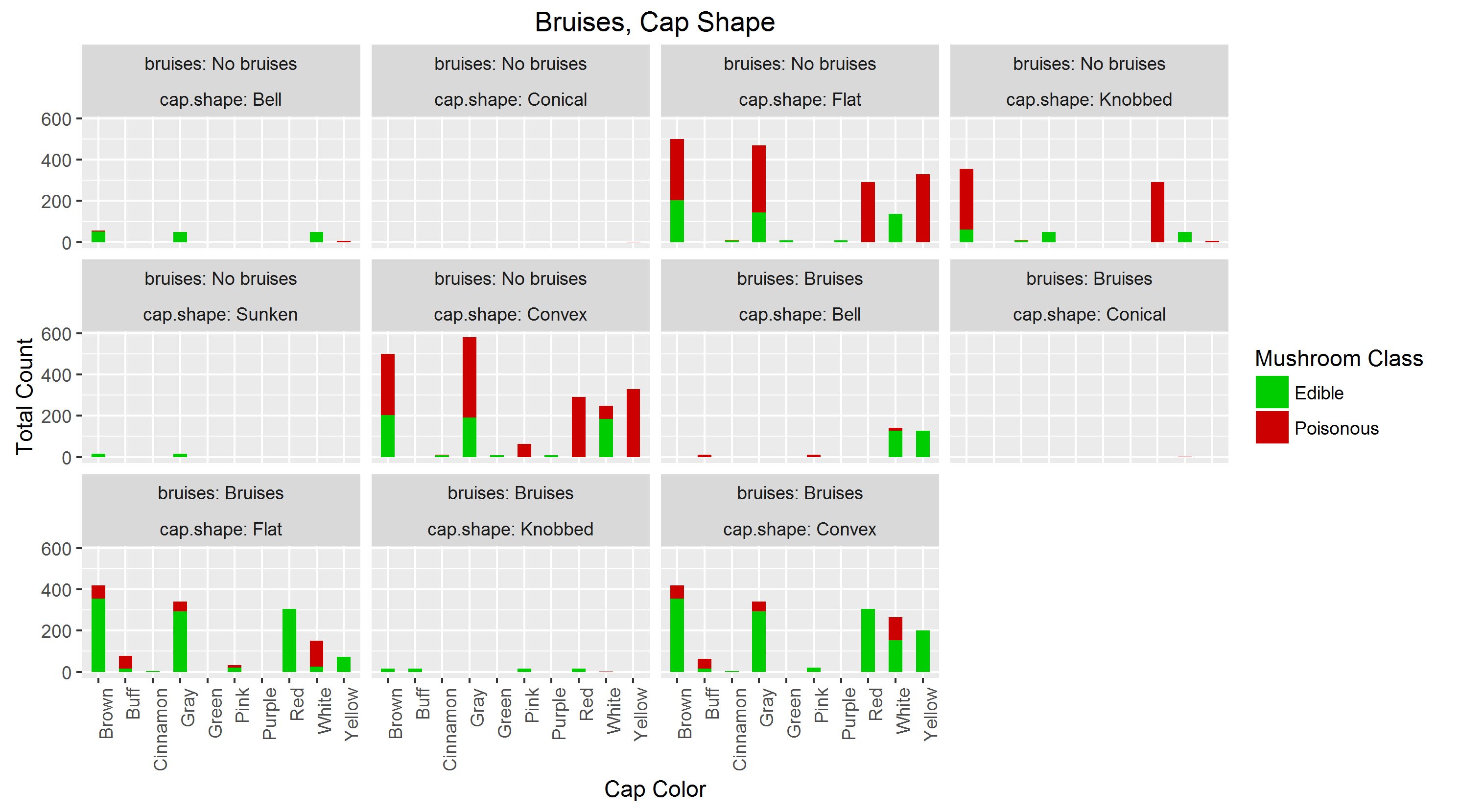
We tried to plot 2 dimensional barchart for each column with the class (poisonous or edible)



However as seen above, the graphs cannot be use for model prediction as some values of cap shape and bruises doesn’t give us enough details to accurately predict the output of whether if the mushroom is poisonous or edible. For example in cap shape, flat and convex mushroom have a 50~% rate of poisonous or edible. We need a more distinctive and clearer pattern to use for our classification model.

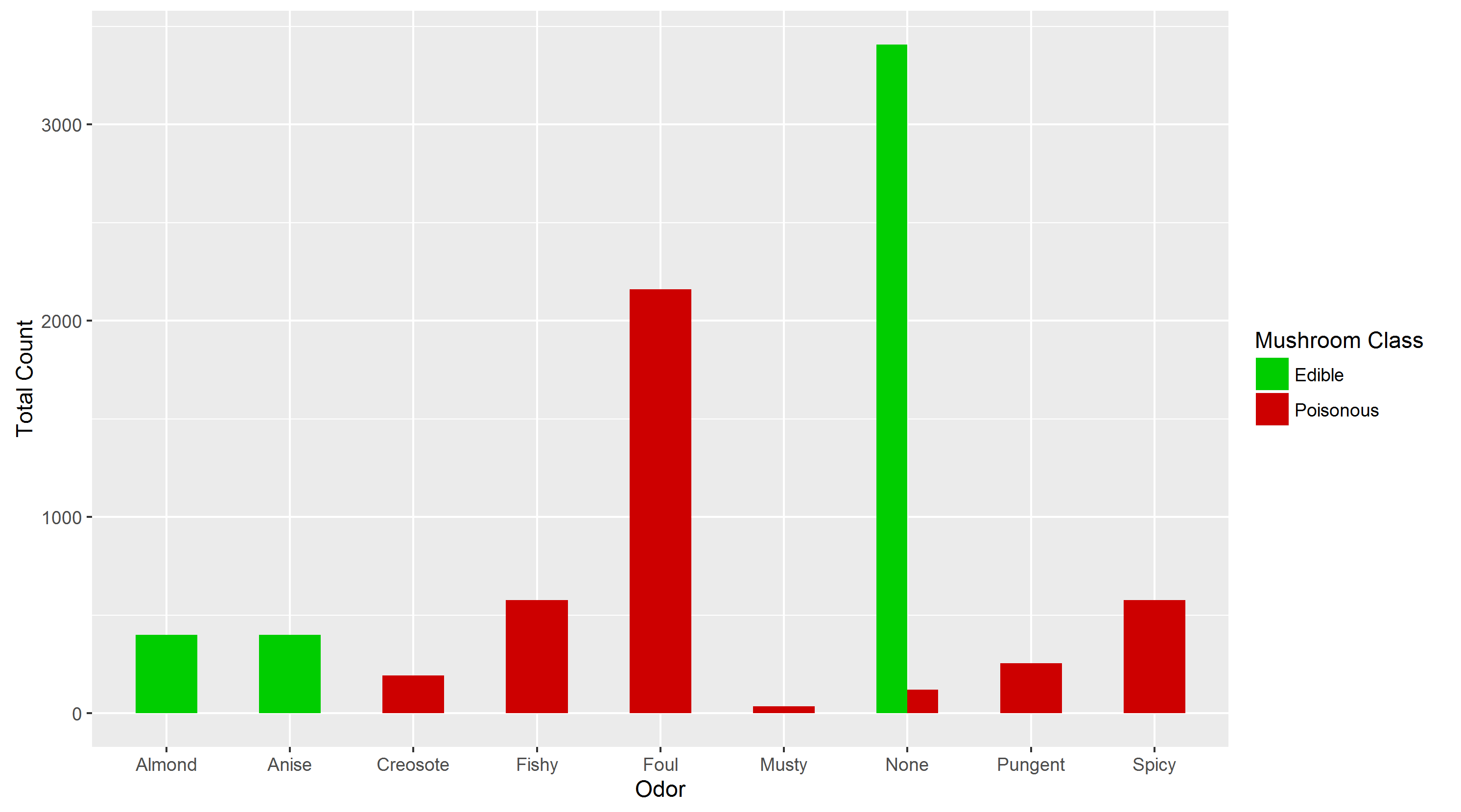
Therefore, we added more variables to the graph grouping them by different section.

This graph provided us a more in-depth understanding of edible and poisonous mushroom by further elaborating the cap shape, bruises and also colour.

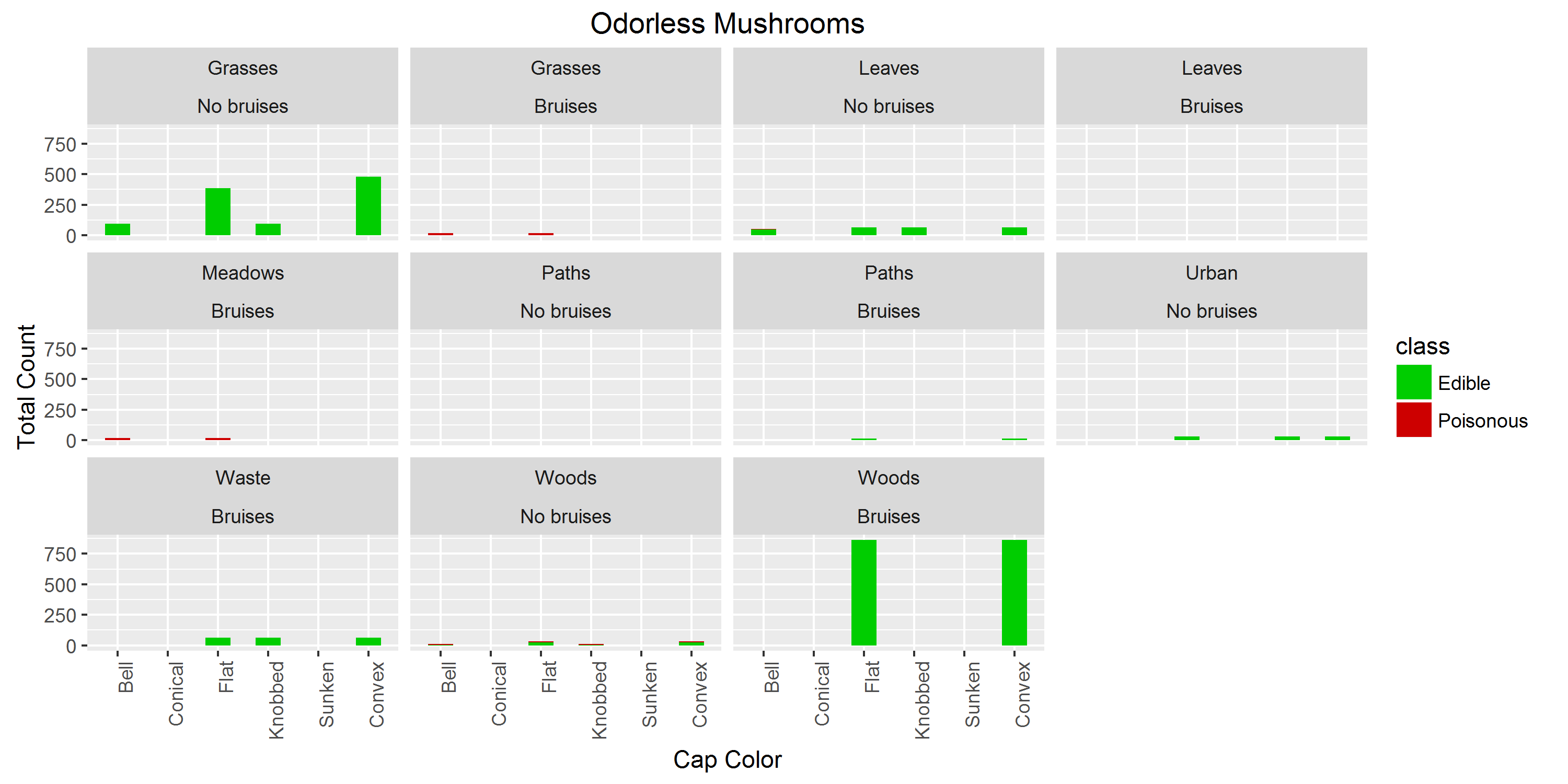


We’ve tried several different combinations of columns and joining them together and the most accurate and distinctive graph we got are shown below:

This graph is accurate enough for us to produce a decent classification model, however there are still room for improvement. We can further elaborate the value “None” by sub setting the data to only the rows that contains the value “None”. Then plot to analyse what makes an odorless mushroom edible or poisonous.



As you can see most of the odorless (“None”) are further elaborated by the habitat of the mushroom, bruises and also the shape. Now we can build an accurate model with an estimated accuracy of at least (90%) for whether if a mushroom is edible or poisonous. This is because we’ve broken the characteristics of what makes a mushroom poisonous. Note that the graph can be even further broken down until all values are distinct with no overlapping values. This will give us a close to 100% prediction model, however we found that it’s unnecessary to overfit our prediction model.



## C. What type of data mining technique (association rule mining, classification or clustering) would be relevant? Give an example, for example, if you think classification is suitable, describe what will be classified and what the possible classes are.

Association mining rule can be used to assist the classification process, to describe the common and pattern of poisonous or edible mushroom. Classes/Attributes that may be used for association mining rule are cap shape, bruises and color. “arules” and “arulesViz” in R can both be used for generating association mining rule and plotting the mining rules.

Classification - Classify whether a mushroom is poisonous or edible based on the association mining rule. Multiple or more association mining rule can be used to improve the accuracy of the classification model. “randomForest” library in R can be used for creating a classification model.

**Discovered disjunctive rules for mushroom datasets**

odor=NOT(almond.OR.anise.OR.none)

result: 120 poisonous cases missed, 98.52% accuracy

spore-print-color=green

result: 48 cases missed, 99.41% accuracy

odor=none.AND.stalk-surface-below-ring=scaly.AND.(stalk-color-above-ring=NOT.brown)

result: 8 cases missed, 99.90% accuracy

habitat=leaves.AND.cap-color=white

result: 100% accuracy

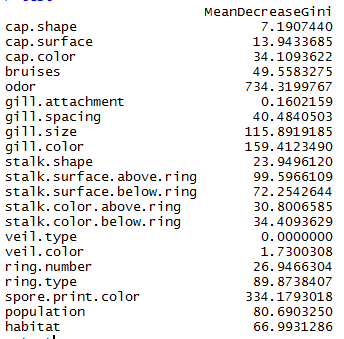
We are planning to use random forest as one of our classification model to predict if a mushroom is poisonous or not. The code below shows a **SIMPLE SAMPLE** that we may be using for part 2. Accuracy we got from random forest is pretty decent, out of few runs (4 different executions with different seed) our prediction model got 0.00 ~ 0.05% error rate. We also extracted which variable/attribute influences the model the most.

Figure 2. Importance (function) of randomForest package, to show which column influences the outcome of the prediction model the most.

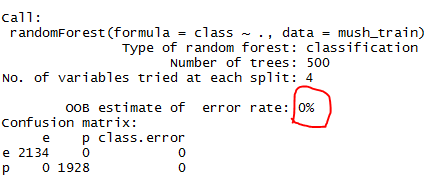


Figure 3. Result of the model

D.  Describe data quality issues, and be specific. Identify which attribute (column) has issues, or if the structure of the data has problems.

A dataset is said to be tidy if it satisfies all of the condition:

1. Observations are in rows
2. Variables are in columns
3. Contained in a single dataset for each type of observations

Mushroom dataset:

1. Each row is an observation of a single mushroom
2. Each column describes a variable of the mushroom
3. Each dataset contains each type of observation, where our type of observation is whether if the mushroom is edible or poisonous

Column 11 (stalk-root) has issues, it contains 2480 missing values and 5644 non-missing values.

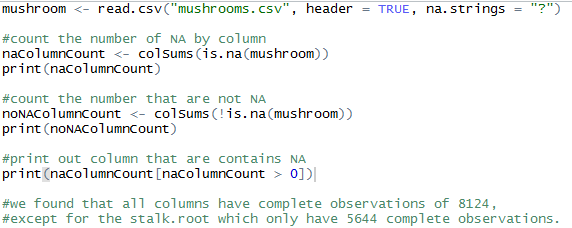


Figure 4. Script used to find whether if any column contains any missing values.

## E. Perform a pre-processing task on the dataset chosen.

“Editrules” one of the library that allows reading, manipulating and impose a set of rules (edit rules) on numerical and categorical data. We will be using this library along with our “pre-defined” rules for the mushroom dataset. This step ensure that we will know if there is any unknown values that isn’t specified in the mushroom’s data codebook. If there is any we can use library “deducorrect” to correct those unknown values.

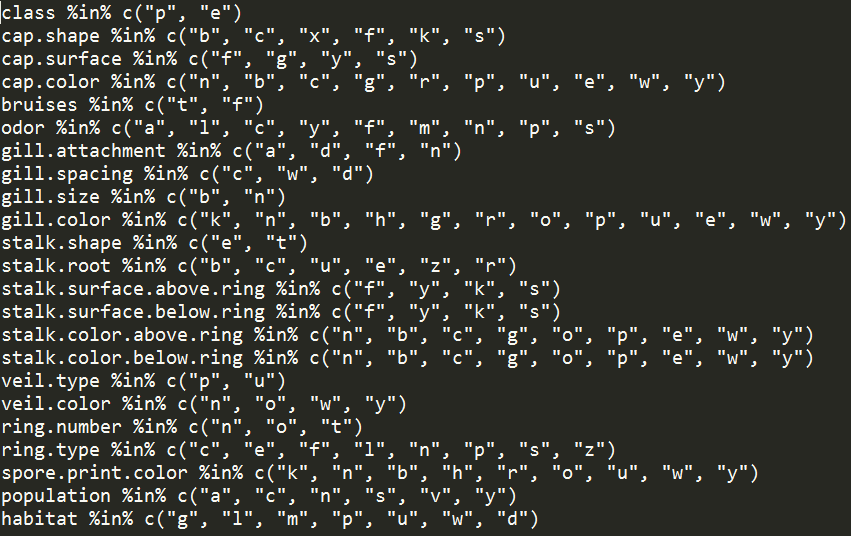


Figure 5. Rules of the mushroom dataset to check if there are any violation of rules (undefined values).



Figure 6. No violation found in the mushroom dataset

We found that the mushroom dataset does not have any violation of rules.