DATASET LINK:

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

GITHUB LINK:

https://github.com/jackwong95/TDS3301DataMining

A. Describe the dataset in your own words.

**Objective**

Classify whether 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 which are all categorical data, 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 data.

**Missing values**

There are 2480 missing values in the dataset which are denoted as “?”. The missing values are only found in “stalk-root”, the 11th column. However, “stalk-root” is a categorical data, therefore, missing values won’t be a problem.

**Distribution of edible/poisonous mushroom**

* Edible: 4208 observations (51.8%)
* Poisonous: 3916 observations (48.2%)
* Total: 8124 observations (100.0%)

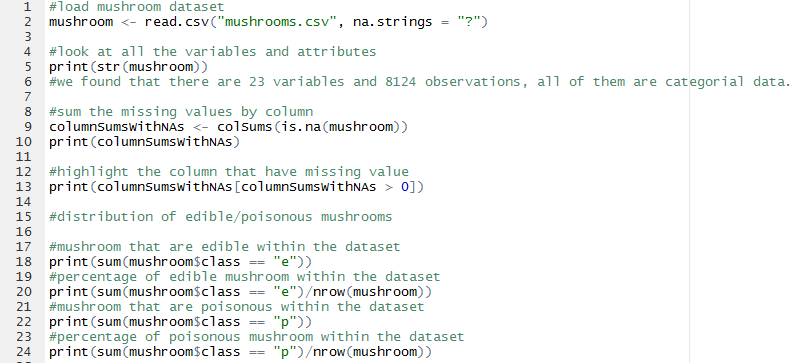
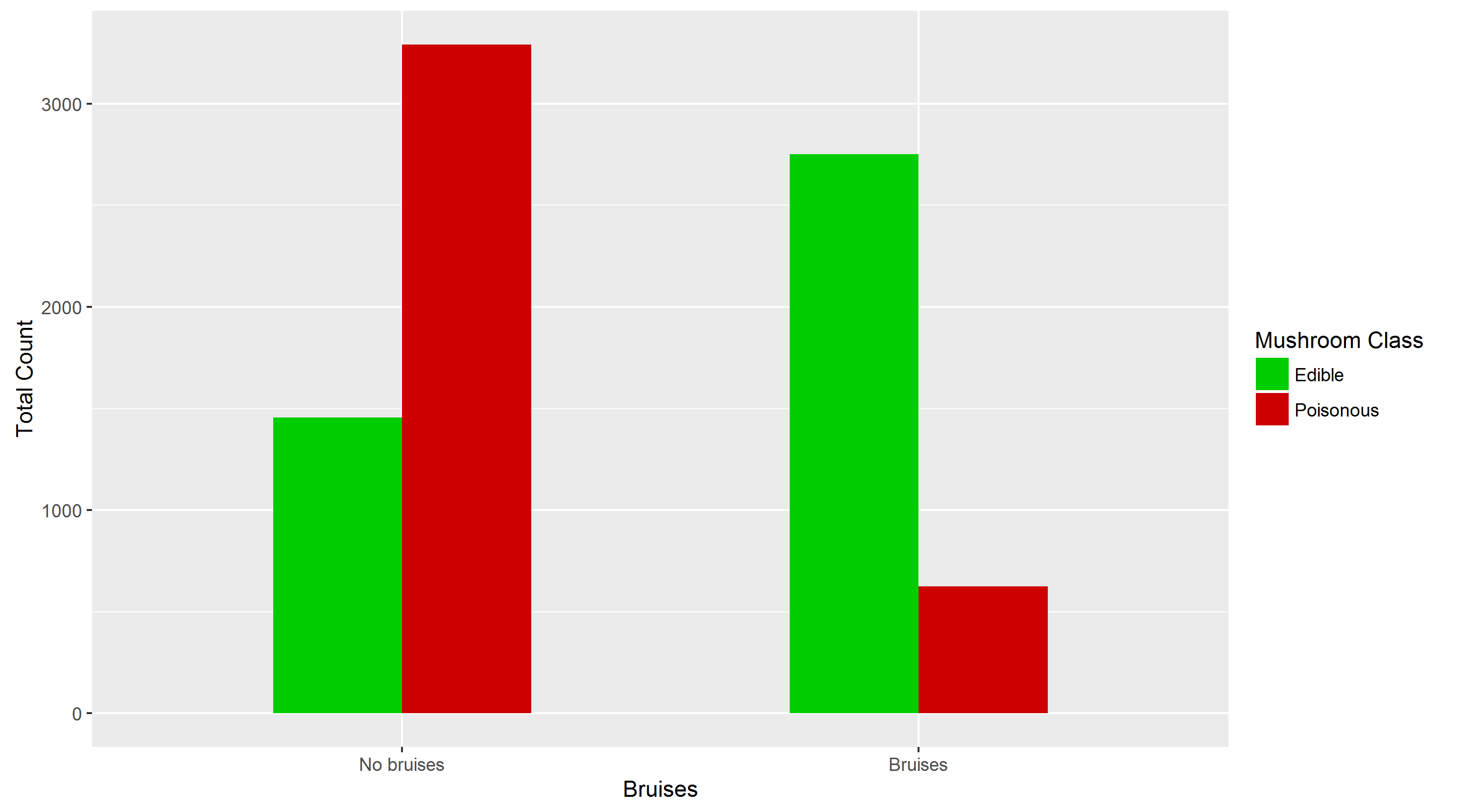


Figure . Code we used for describing the mushroom dataset.

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

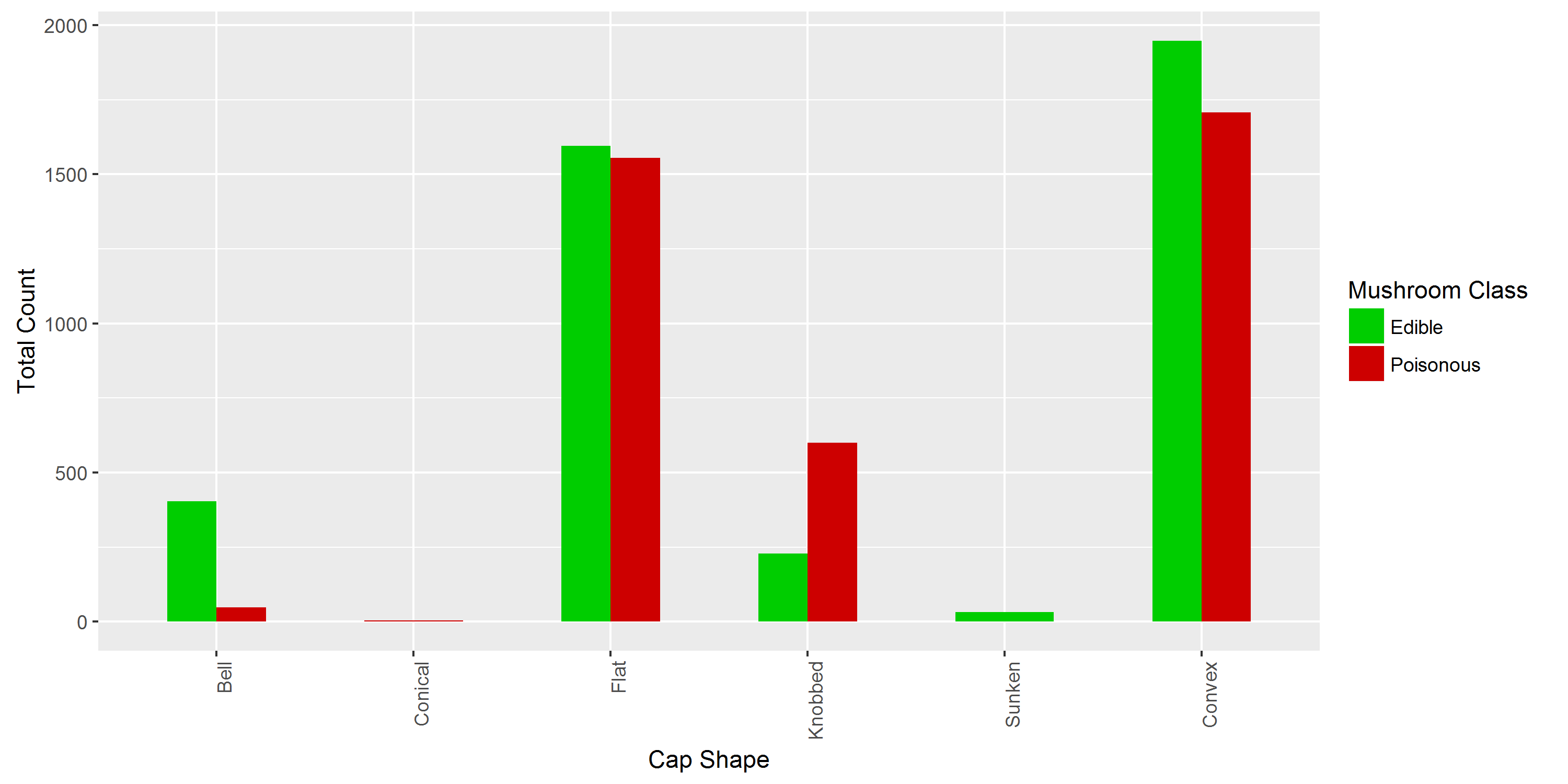
To identify which feature(s) can be a good predictor to classify which mushroom is edible or poisonous precisely.

As we don’t have a clear idea of what each feature mean, we started with bruises because it is a binary data, with or without bruises.



We've plotted a side by side bar chart to see the distribution of the mushrooms by bruises and further separated the bars by edibility of the mushroom. From the plot, we can say that mushrooms with bruises tends to be edible, and without bruises tends to be poisonous.

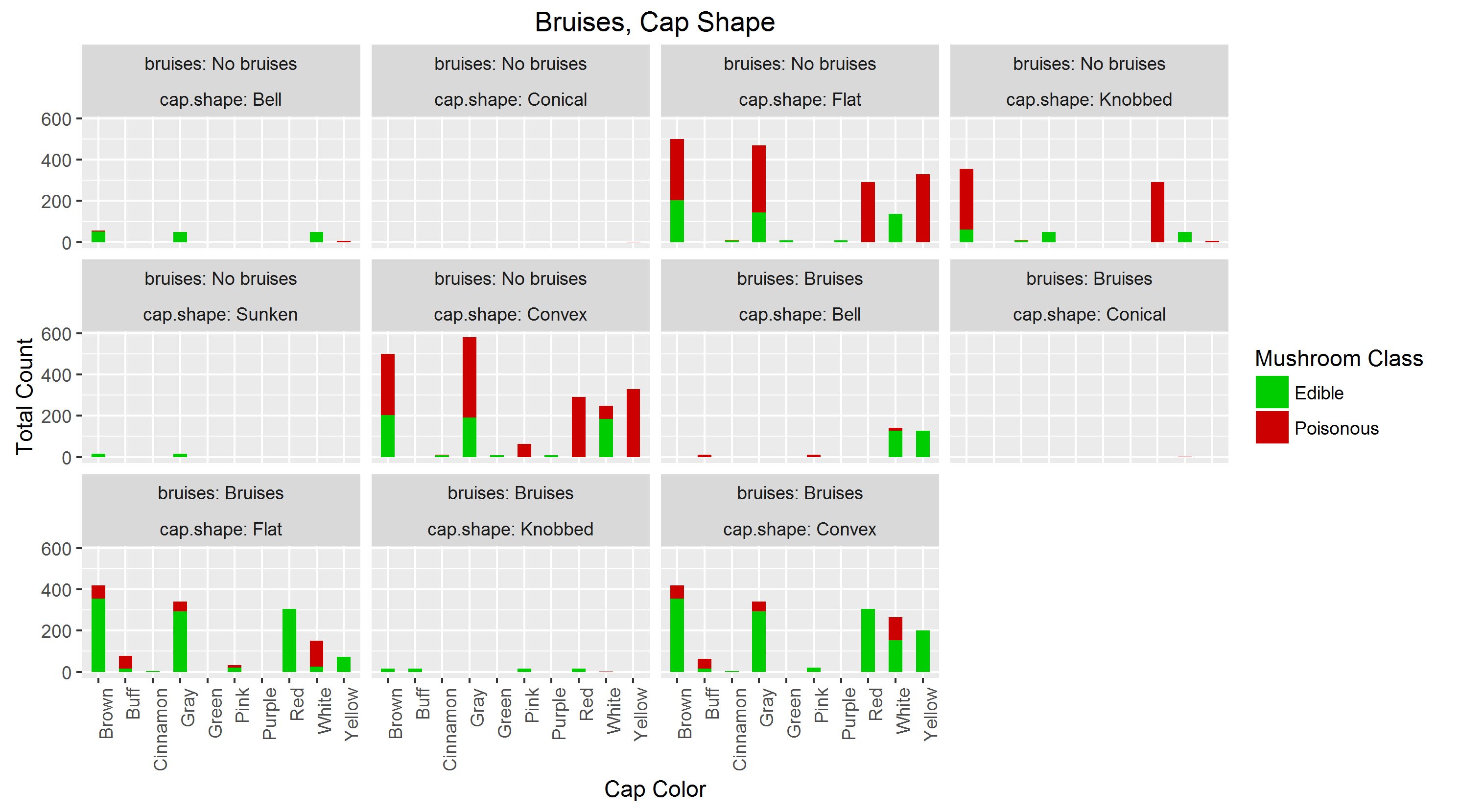
Next, we tried the same observation based on cap shape instead of bruises.



From the plot, there isn’t any clear pattern. Either cap shape alone isn’t enough or there isn’t a strong correlation with the edibility of the mushroom.

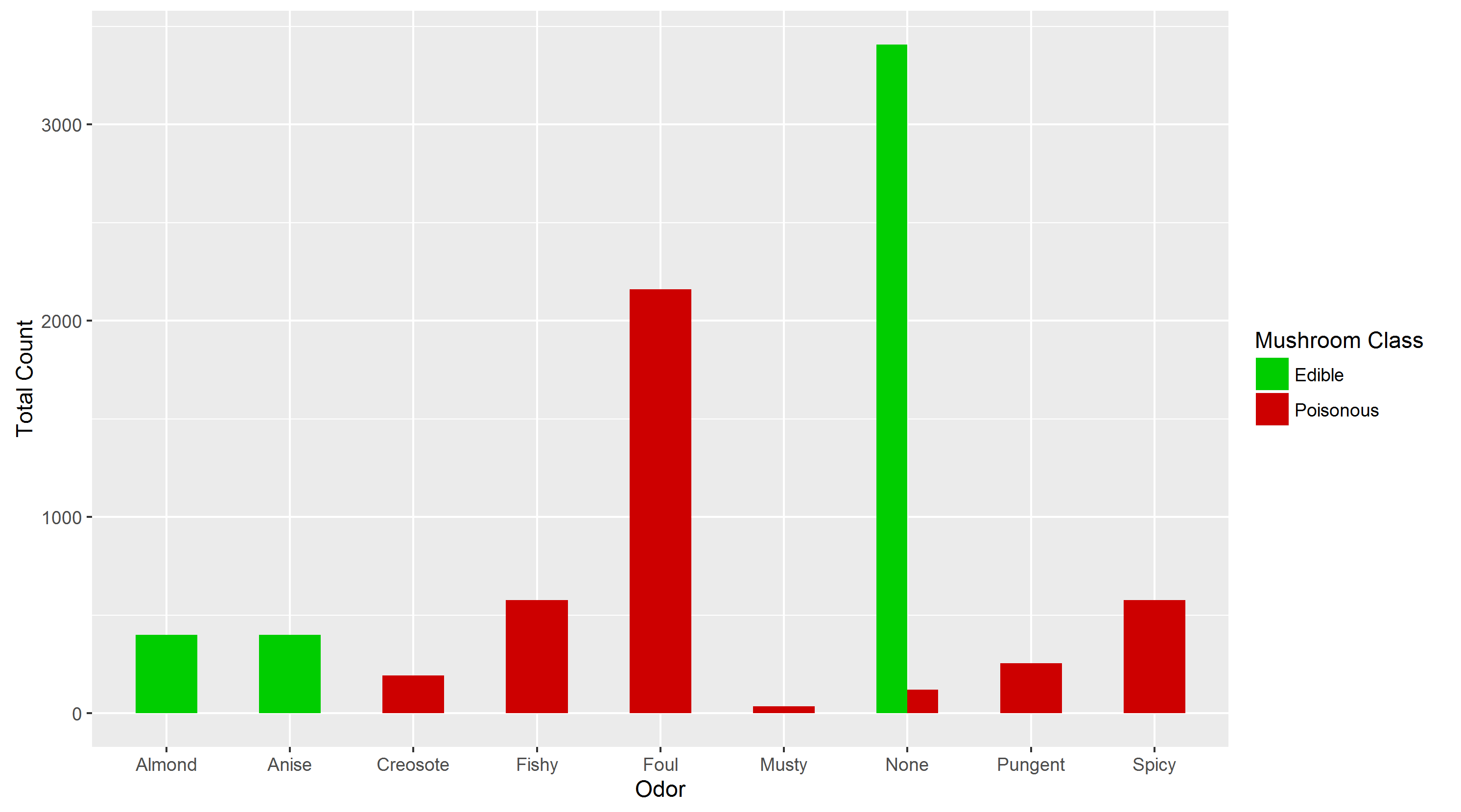
Therefore, we further split the bar chart by adding more features.

So, the new plot provided us a more in-depth understanding for the edibility of mushroom. We can observe that mushrooms with bruises and with flat or convex shape have a high edibility rate. However, there still isn’t a clear pattern. Furthermore, there are many bars with a low frequency which are not suitable to use as predictors.

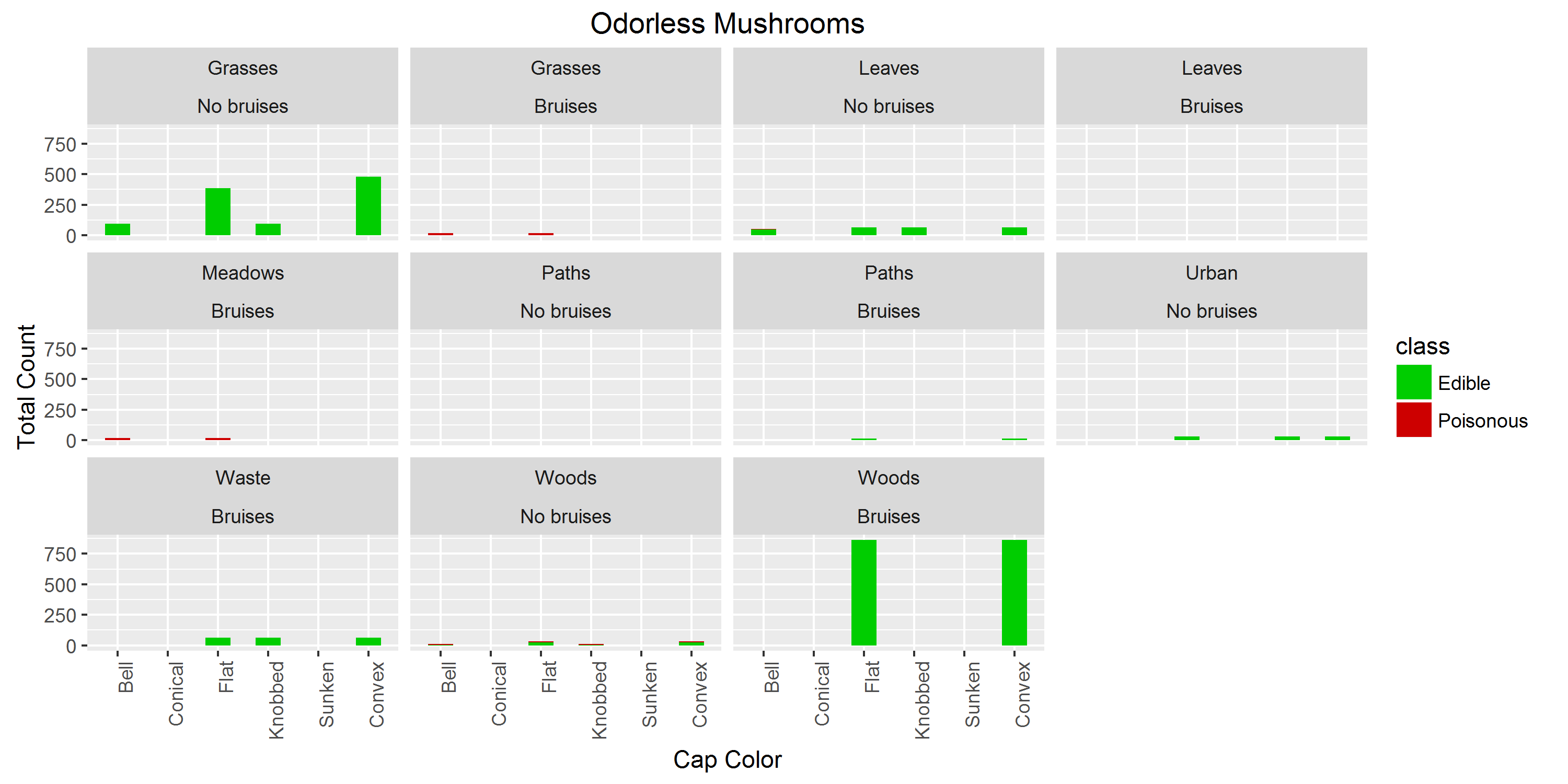


Therefore, instead of proceeding with the same trend. We tried to observe the distribution by odor.

According to the plot above, we can clearly see a distinct pattern. This plot alone is sufficient to produce an accurate classification model, however there are still room for improvement. We can further elaborate the value “None” by sub-setting the data. Then further analyse what makes an odorless mushroom edible or poisonous.



We observed that the red bars are very insignificant as compared to some green bars. Therefore, we can build an accurate model with just odor alone.



## 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 applied to assist the classification process, to describe the common pattern of poisonous or edible mushroom. Features 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

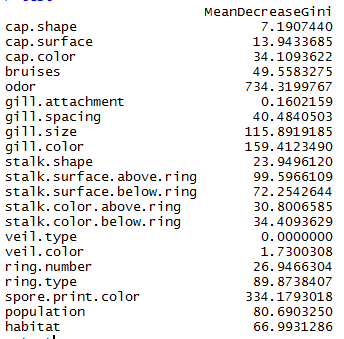
We’ve tried random forest classification to mushroom dataset. We first selected 50% of our original dataset by random sampling without replacement as our training set. The training set will be used to build our model. Then, another random sample as our test set. The test set will be used to judge our model. Random forest gave us an accuracy of 99.5%~100% accuracy.

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

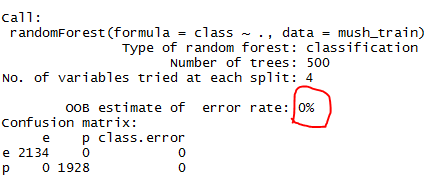


Figure 4. Result of the model

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. Each observation is one row.
2. Each variable is one column.
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. The dataset contains only one type of observation which is mushrooms.

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

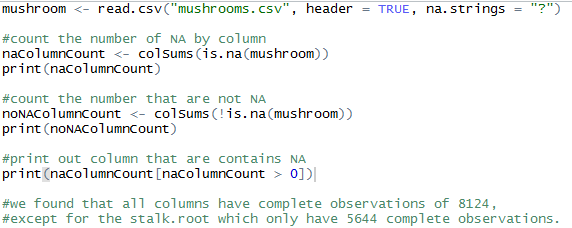


Figure 5. Script used to find whether if any column contains any missing values.

## 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.

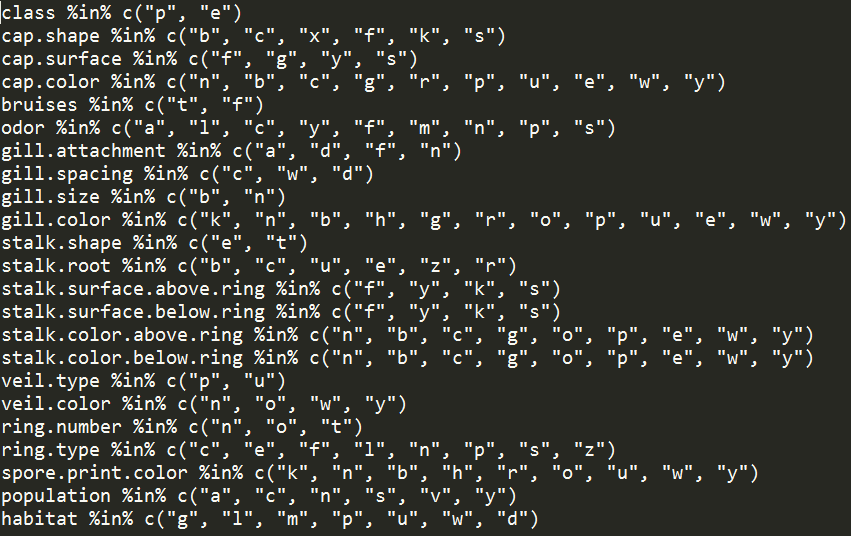


Figure 6. Rules of the mushroom dataset to check if there are any violation of rules (undefined values).



Figure 3. No violation found in the mushroom dataset

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