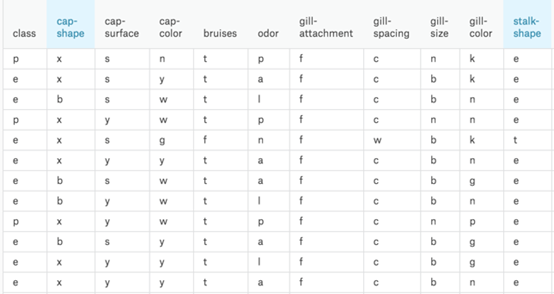
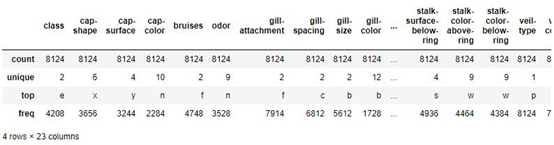
**Data**

***Data Description***

The raw data of our project is from Kaggle, it is a dataset with 23 species of gilled mushrooms. By checking the dataset in the below of figure 1 and 2, we can see that there are over 8000 variables and 23 attributes such as cap shape, cap surface, cap color and so on. Each attribute influences on the outcome of a single type of mushroom is edible or poisonous.



*Figure 1. Data set*

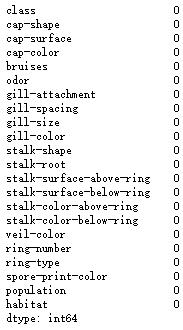


*Figure 2. Data check*

***Data Cleaning***

**Null check**

First, we check the dataset by null check. The reason of this data process is that dataset should have no null values. A null indicates that a variable doesn’t point to any object and holds no value. Our project should avoid null otherwise there will be problems in building the model. Fortunately, from figure 3 in the below, it is clearly that the null values of each attributes is zero.



*Figure 3. Null check*

**Encoding**

Second, in our mushroom dataset, each attribute is a string in python. Since the dataset has string values, it needed to convert all the unique values into integers. Furthermore, label encoding is performed on the data shown in the figure 4 as below. By doing so, each attribute has a sequence of numbers into a specialized format for efficient analyze in the future process of the project.

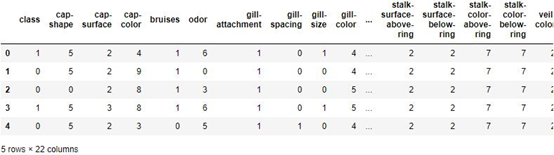
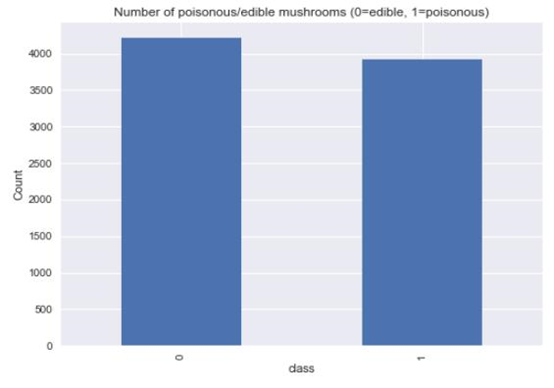


Figure 4. *Label encoding*

**Balance check**

Third, we use data balance check technique to make sure data prediction will be unbiased towards to the frequent class. Imbalance dataset is a special case for classification problem where the class distribution is not uniform among the classes. From the figure 5 shown as below, we can see that our dataset is basically balance between edible and poisonous.



*Figure 5. Data balance*

**Data Splitting**

The fourth and the last part of data cleaning, we split the data into train data and test data according to the ratio of 8 to 2 to build and test the model. Also, we set up a plot correlation matrix to check the correlation of each attribute to another. The blue blocks mean positive correlation and red blocks means negative correlation between two attributes. The darker color means stronger correlation. For example, from the figure 6 shown in below, veil-color and gill-attachment have the highest positive correlation which is 0.9.

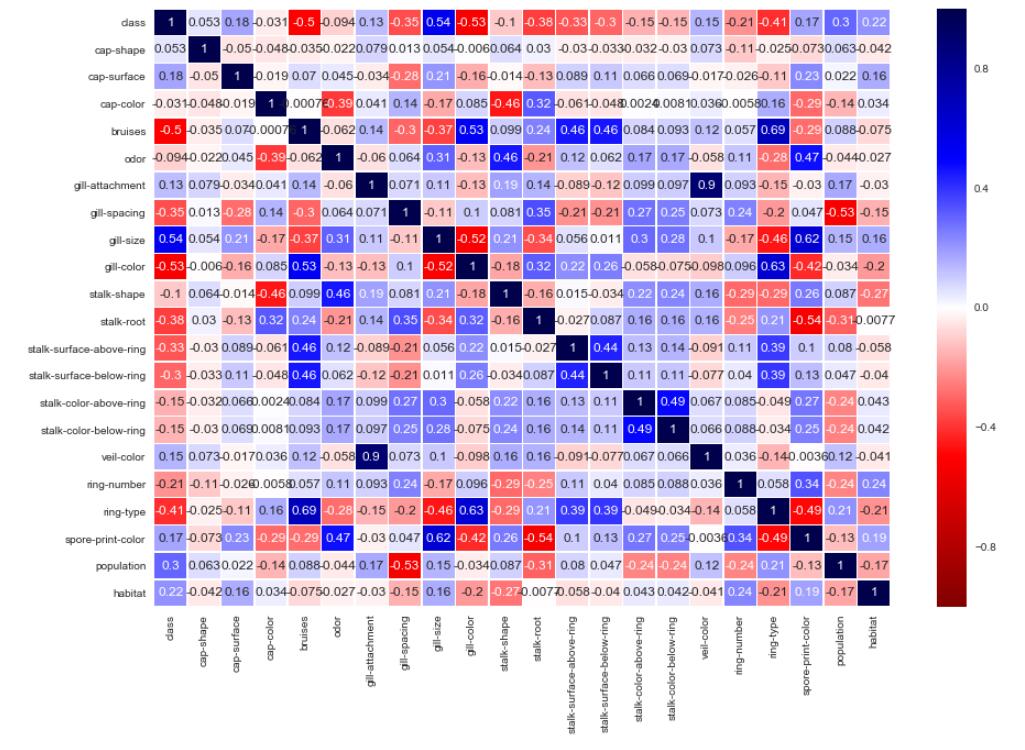


Figure 6. *Plot correlation matrix*