

2.data visualization (hw1_1.py,hw1_2.py):

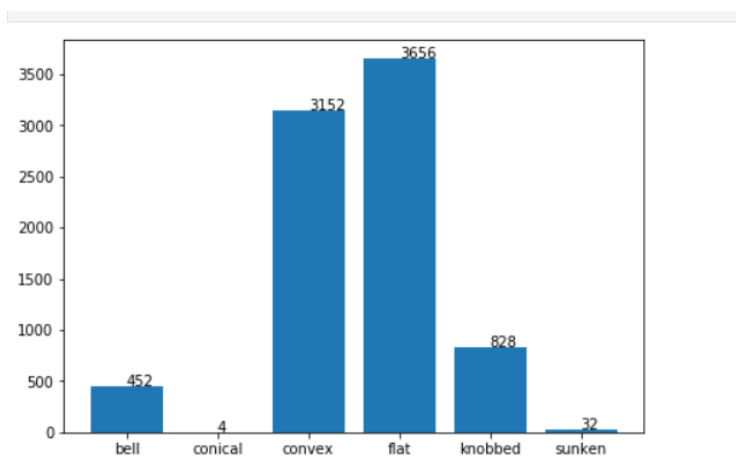
hw1_1.py 因為要畫圖過多，所以有 warning，但是圖都有畫出來，

在我的電腦上跑也只須等個幾秒就好。

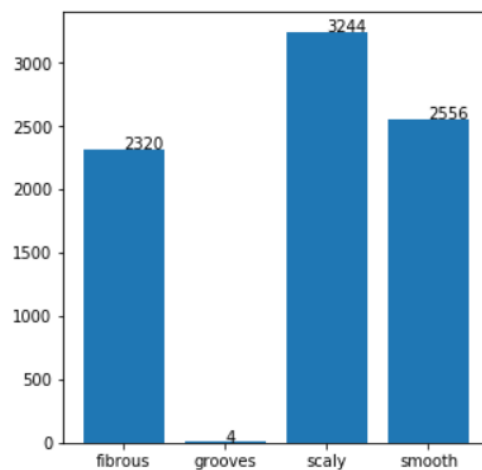
Mushroom dataset:

Value frequency of every feature:

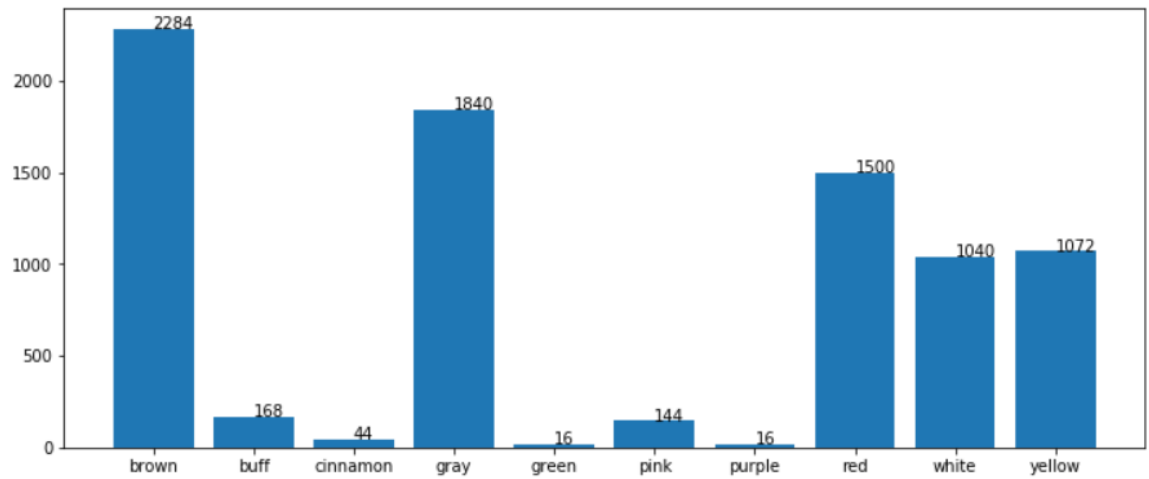
Cap_shape



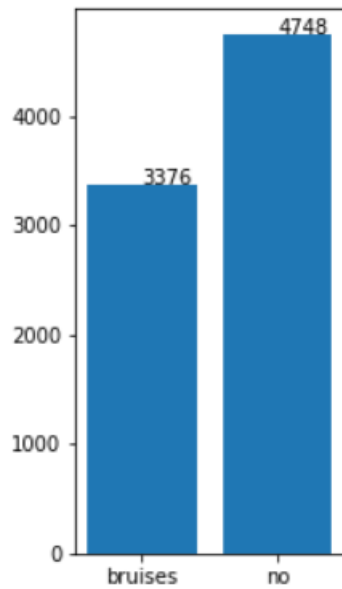
Cap_surface:



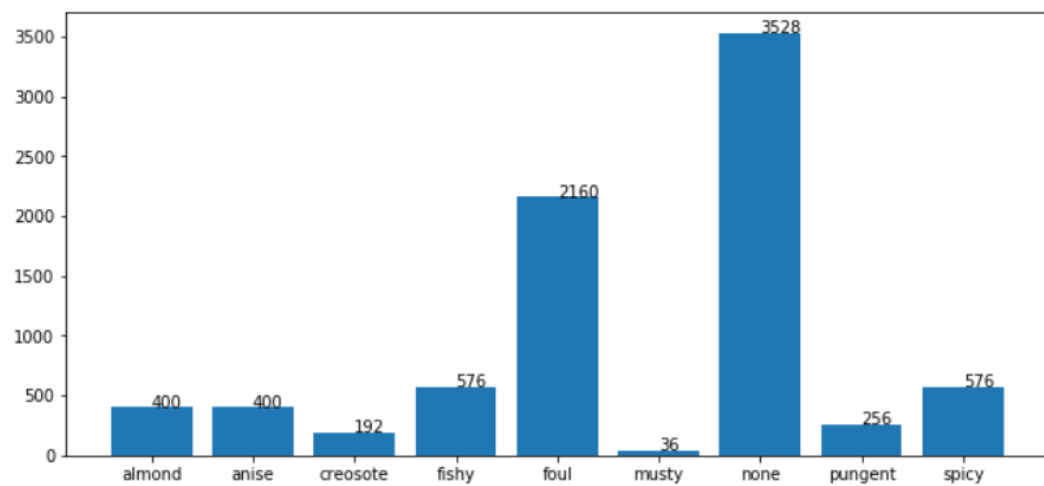
Cap_color:



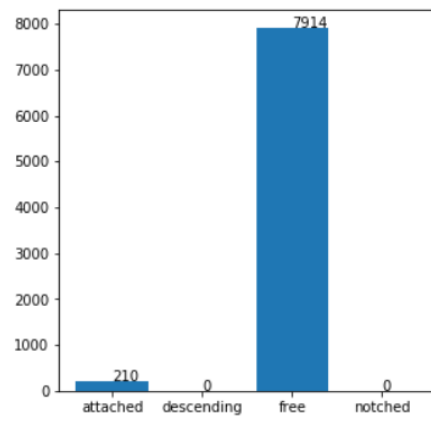
Bruises:



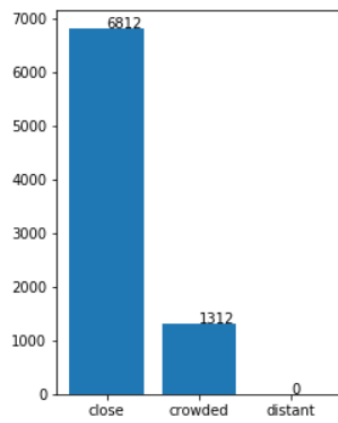
Odor:



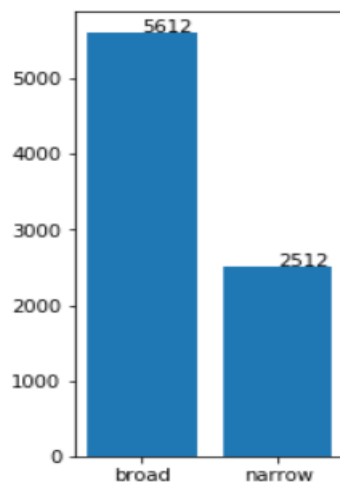
Gill_attachment:



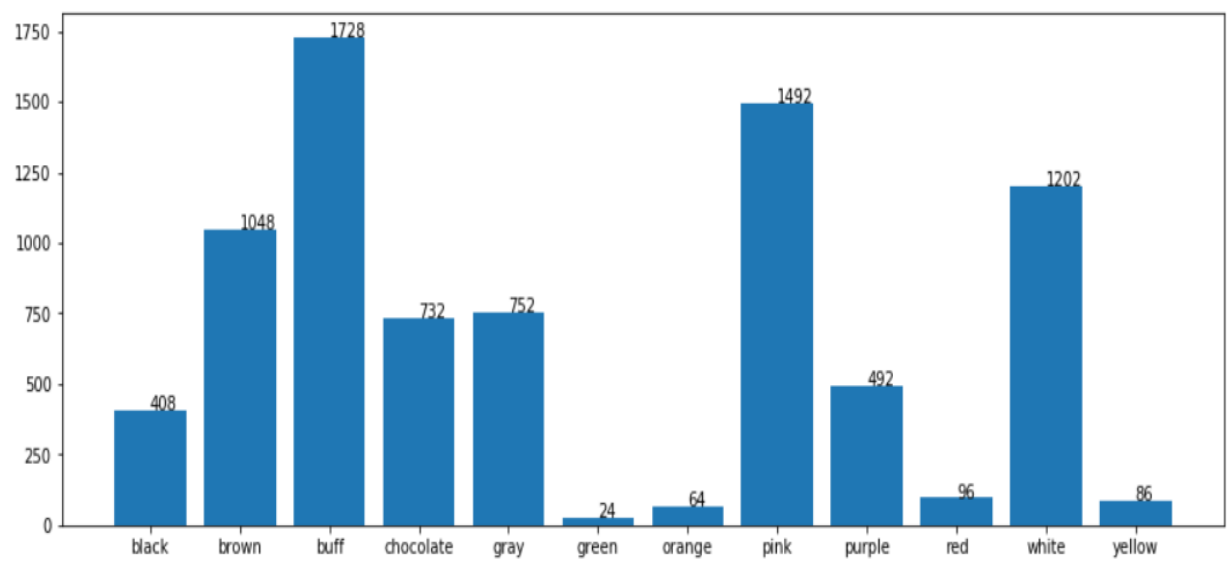
gill-spacing:



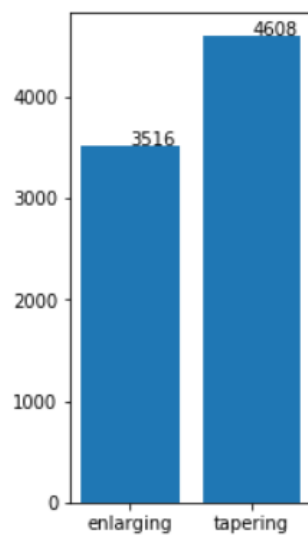
gill-size:



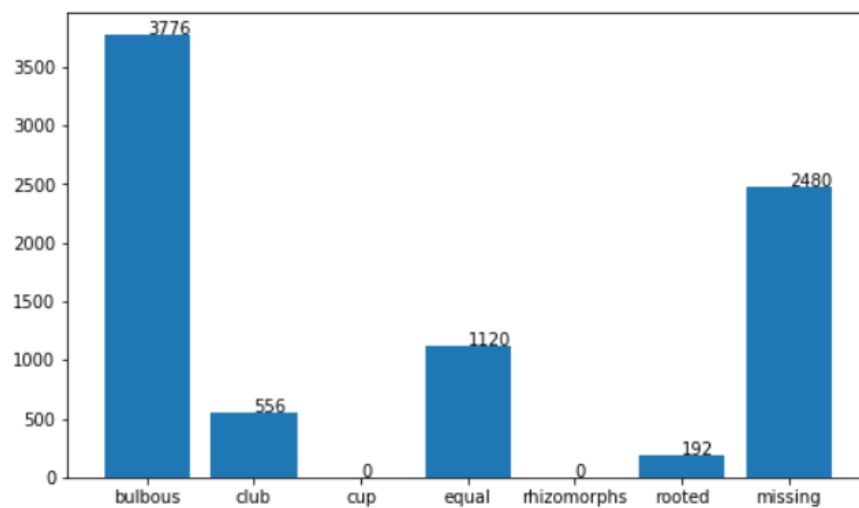
gill-color:



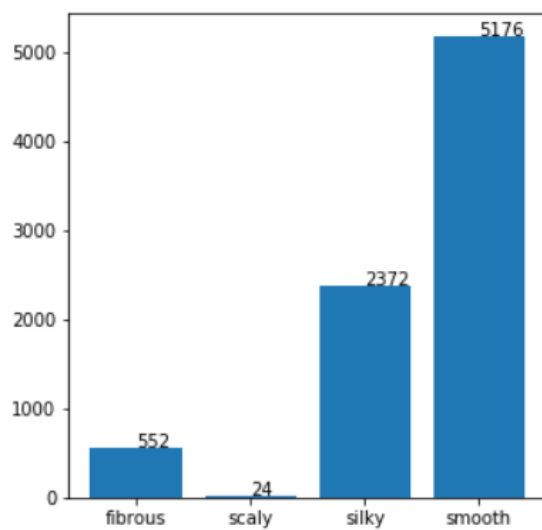
stalk-shape:



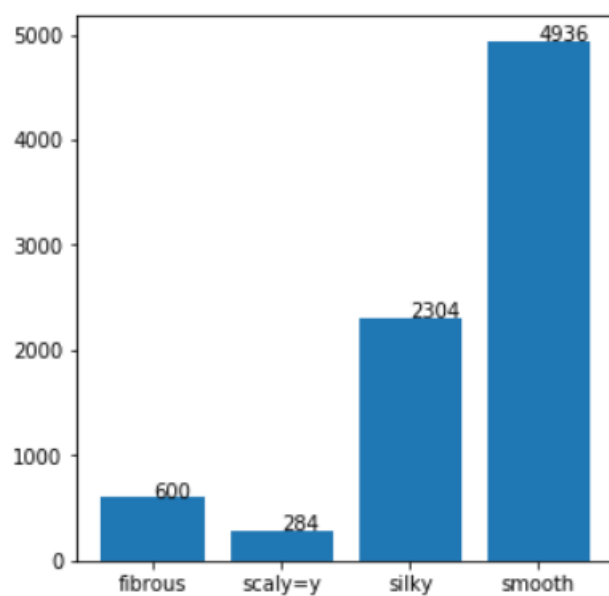
stalk-root:



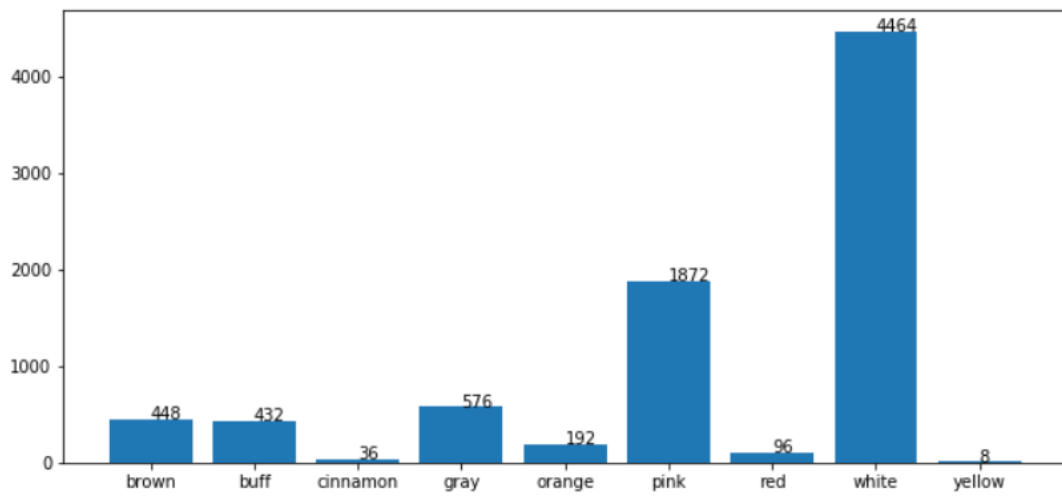
stalk-surface-above-ring:



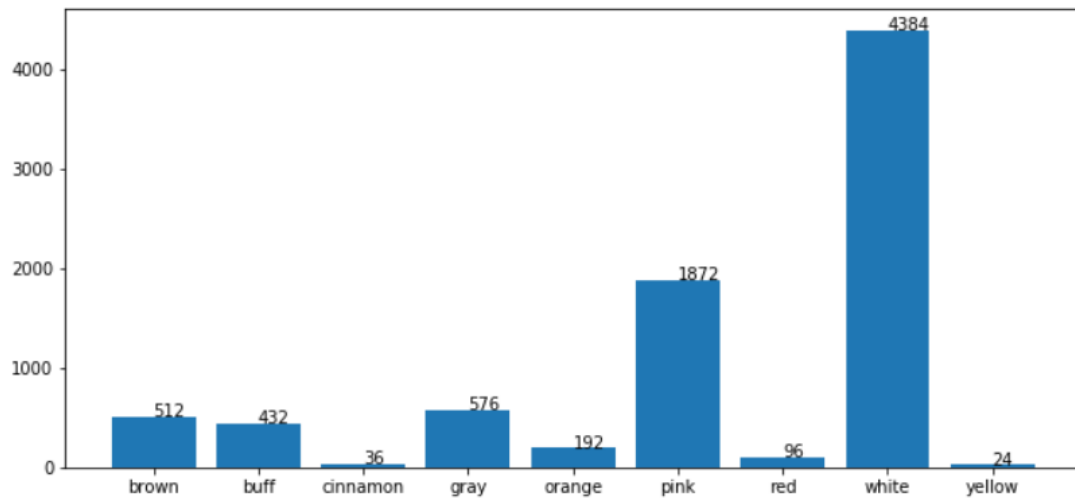
stalk-surface-below-ring:



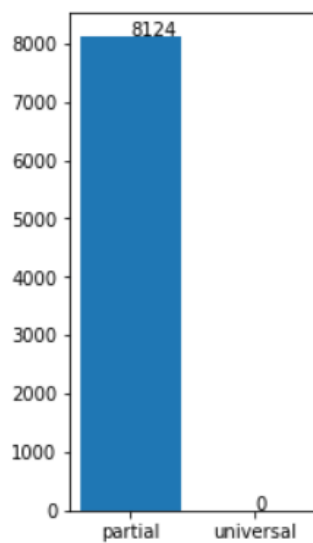
stalk-color-above-ring:



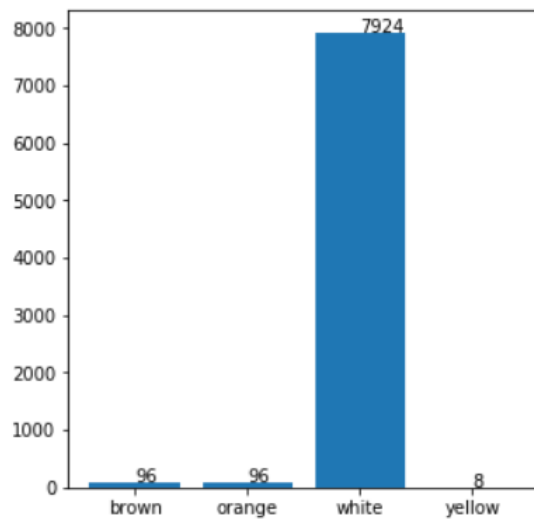
stalk-color-below-ring:



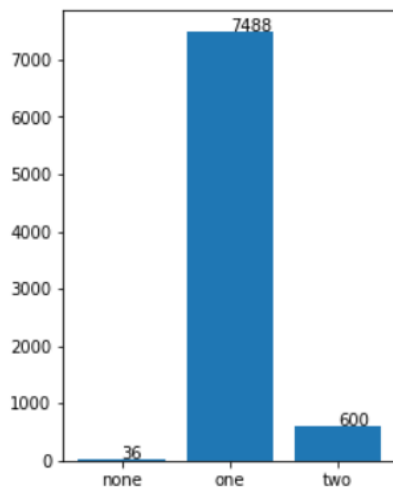
veil-type:



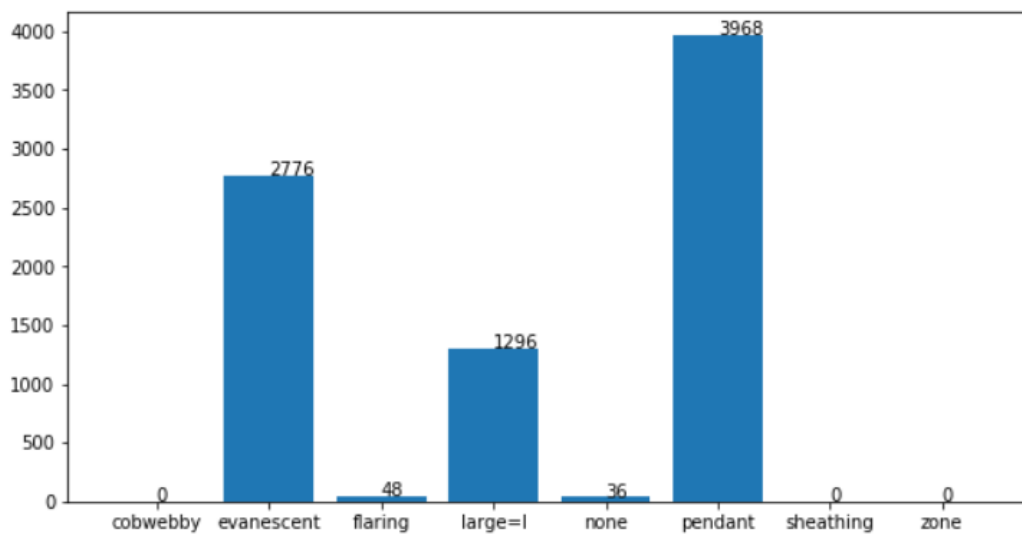
veil-color:



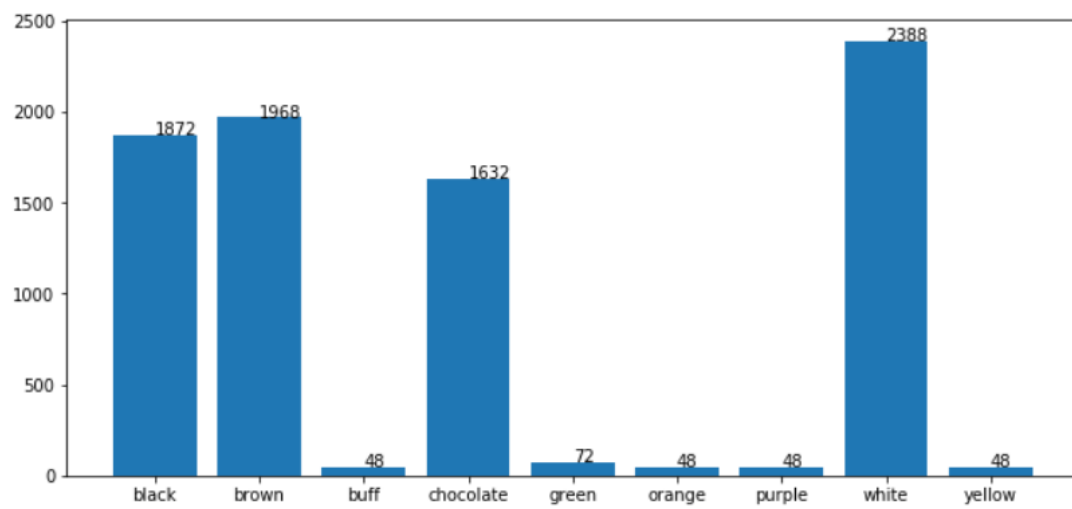
ring-number:



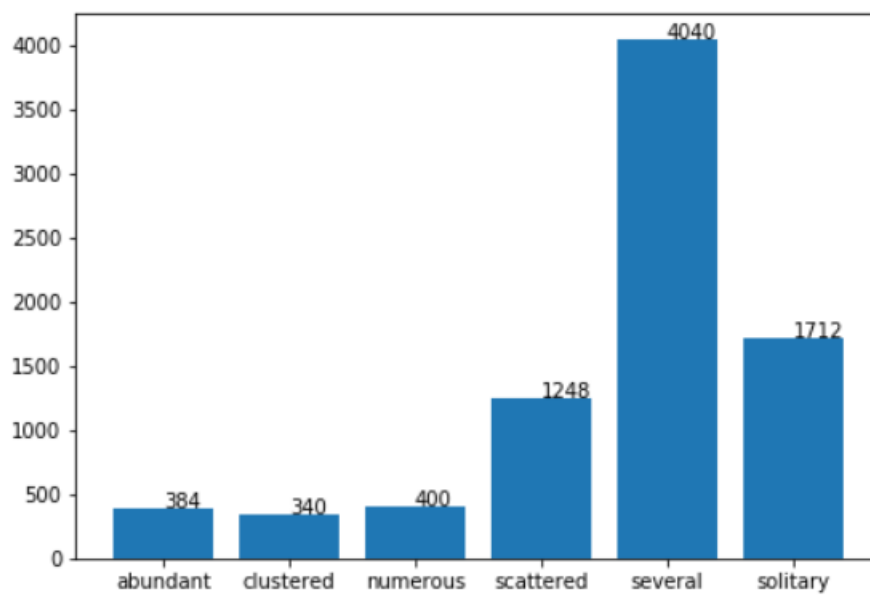
ring-type:



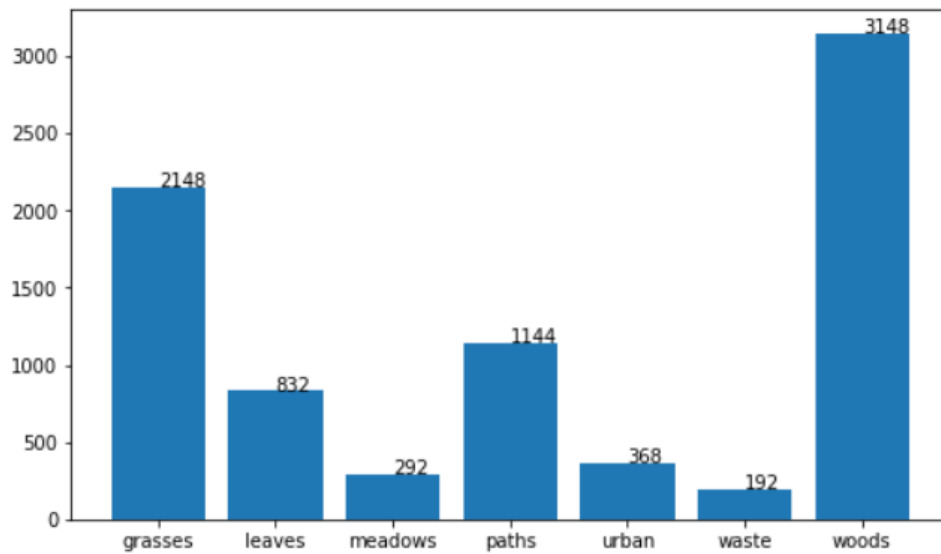
spore-print-color:



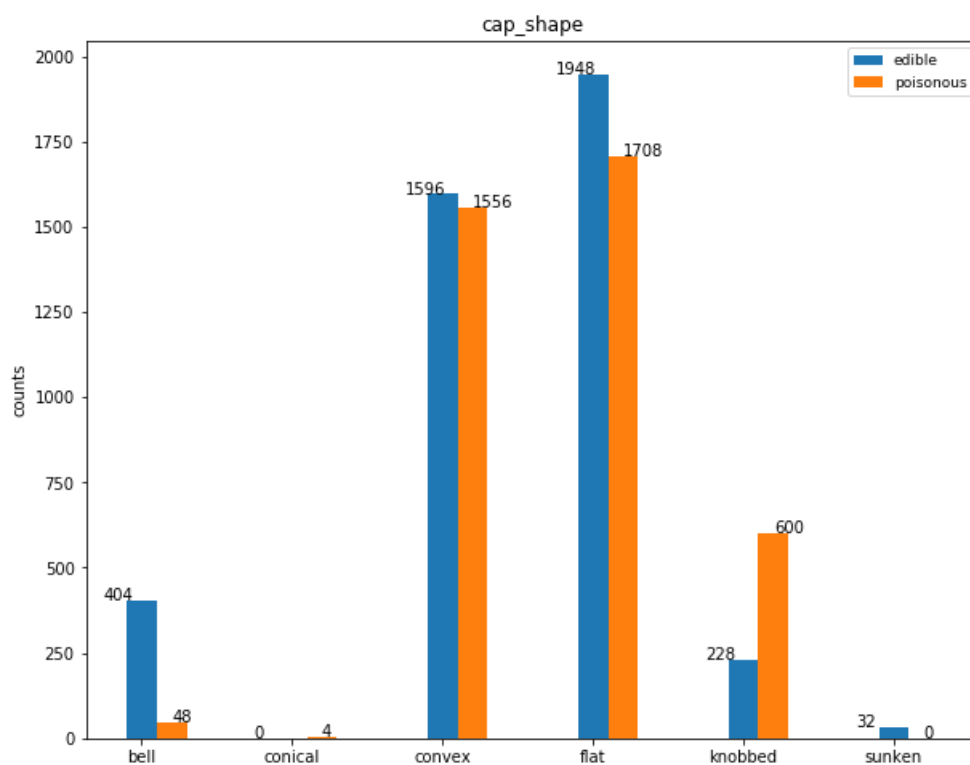
population:

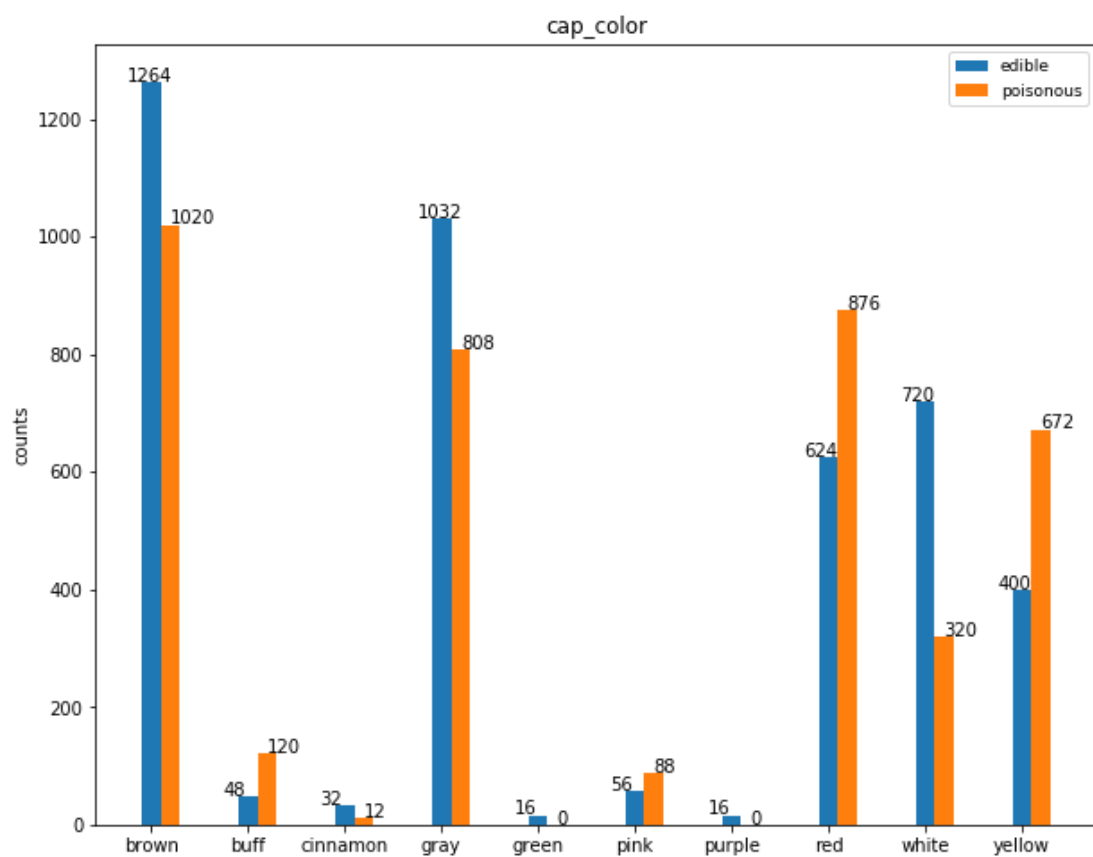
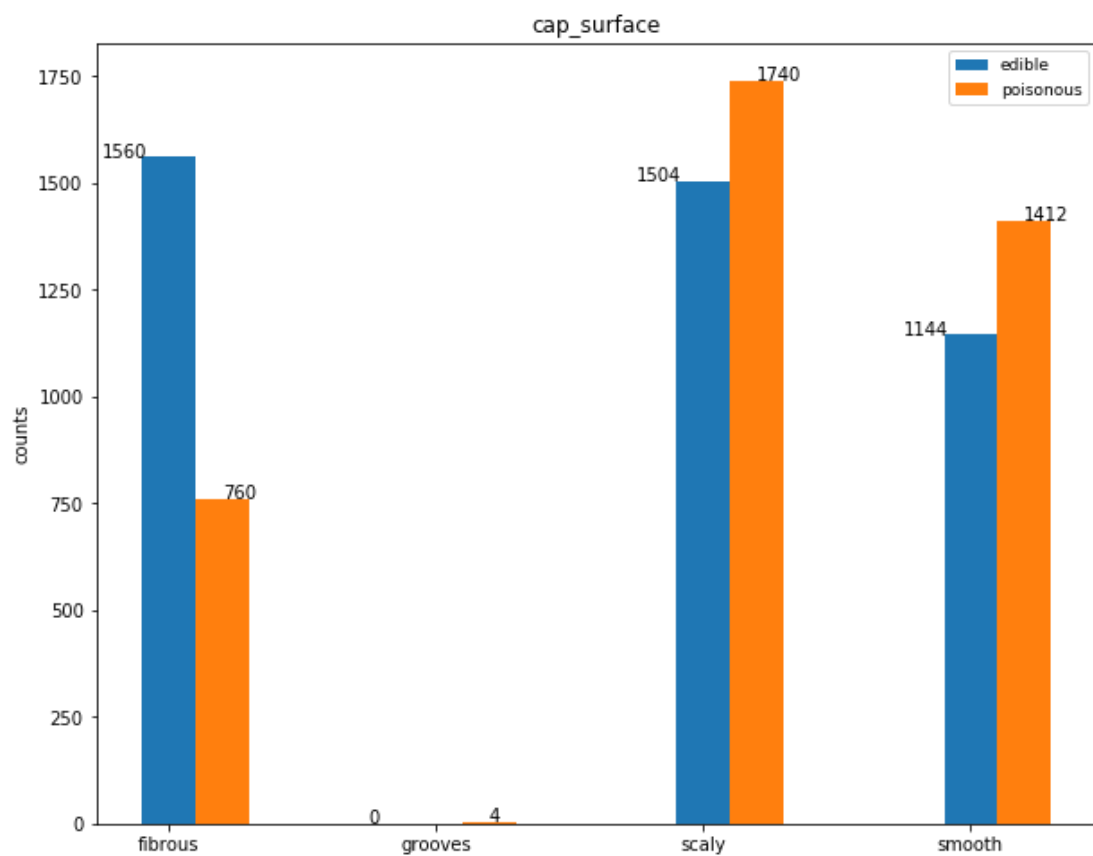


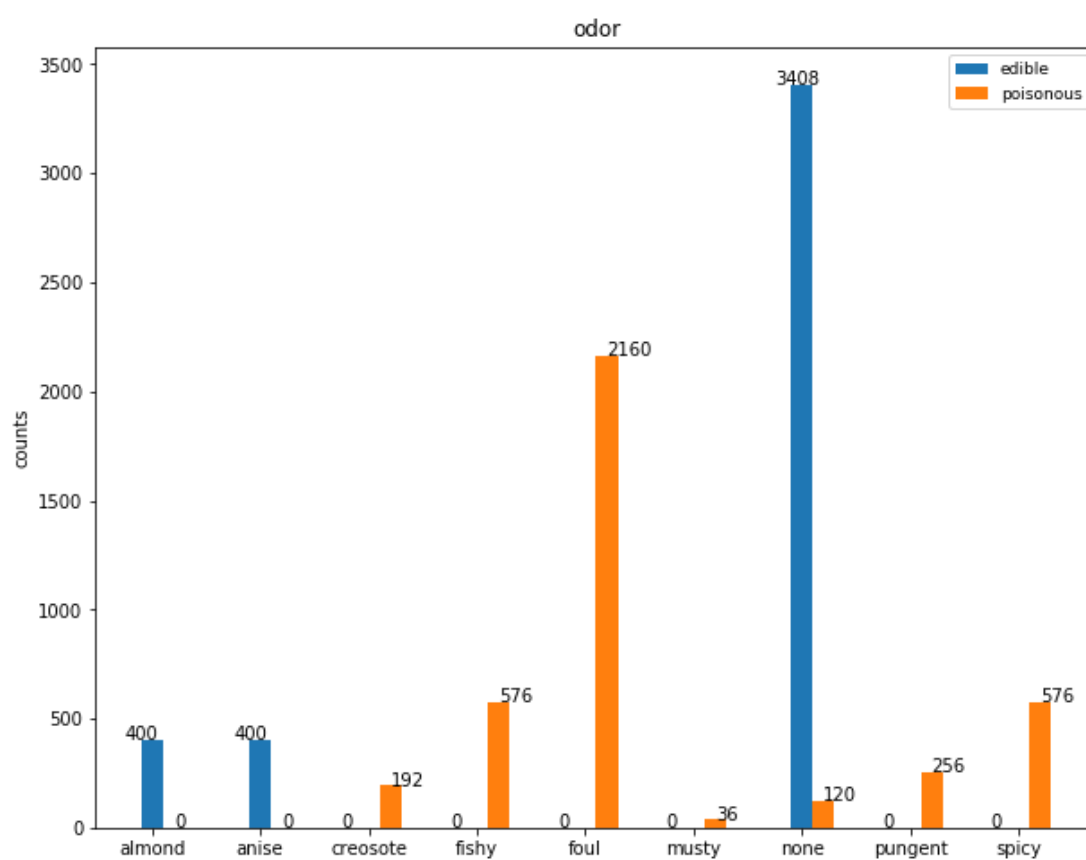
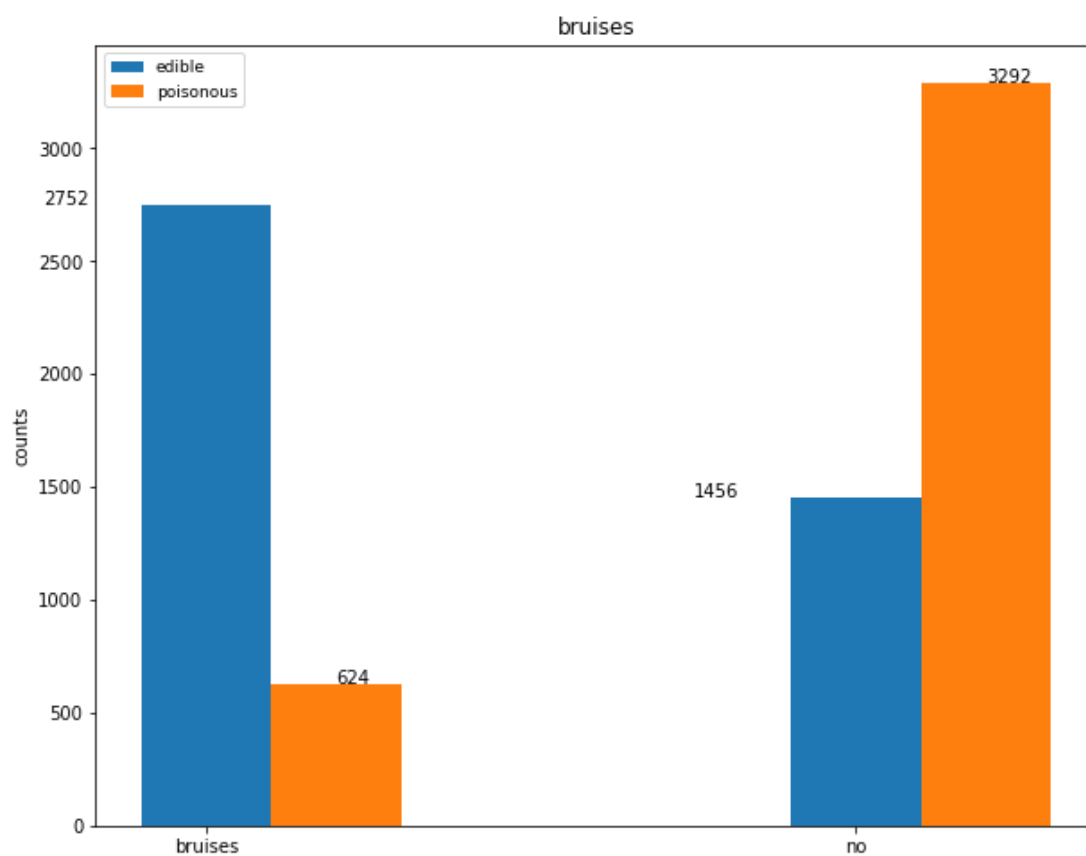
habitat:

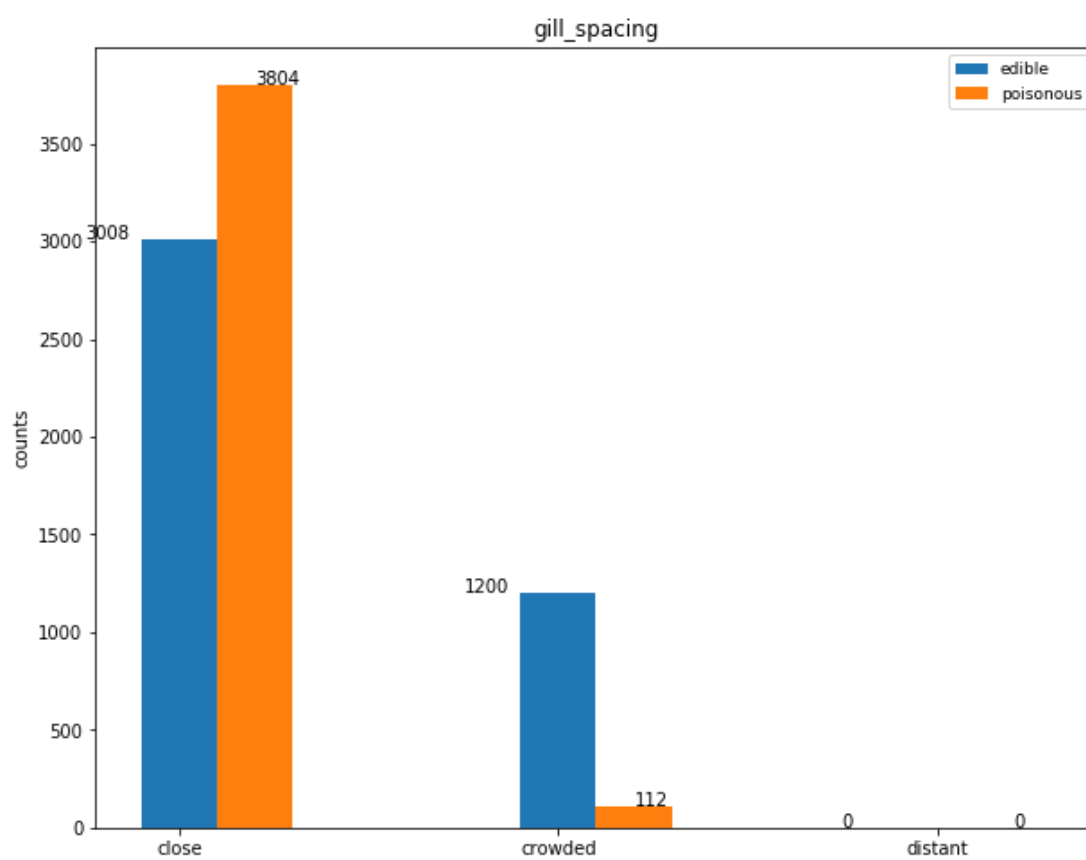
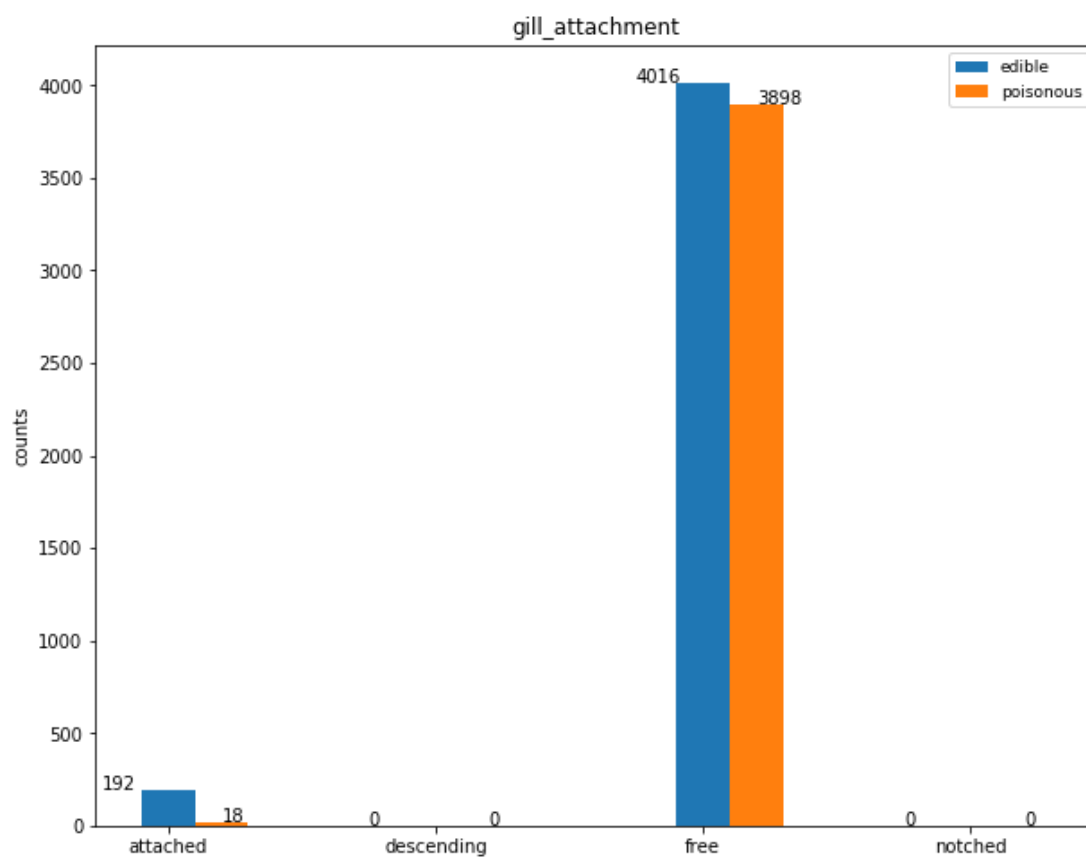


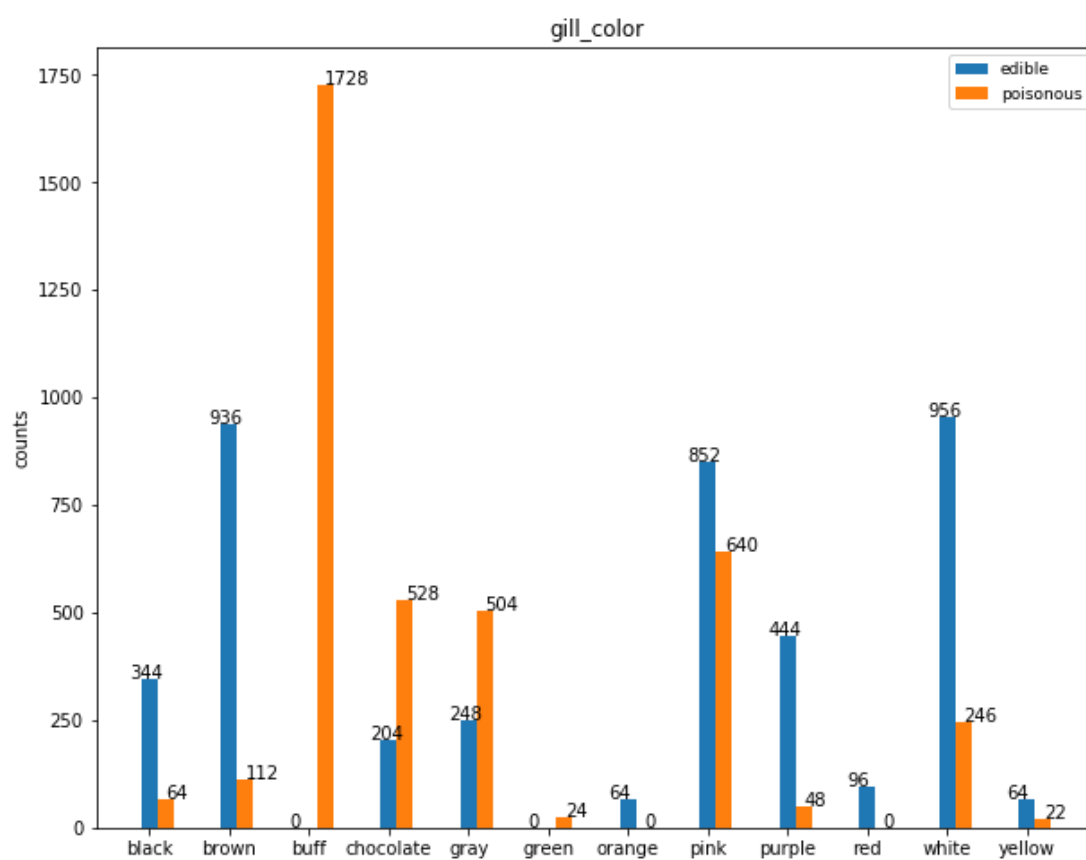
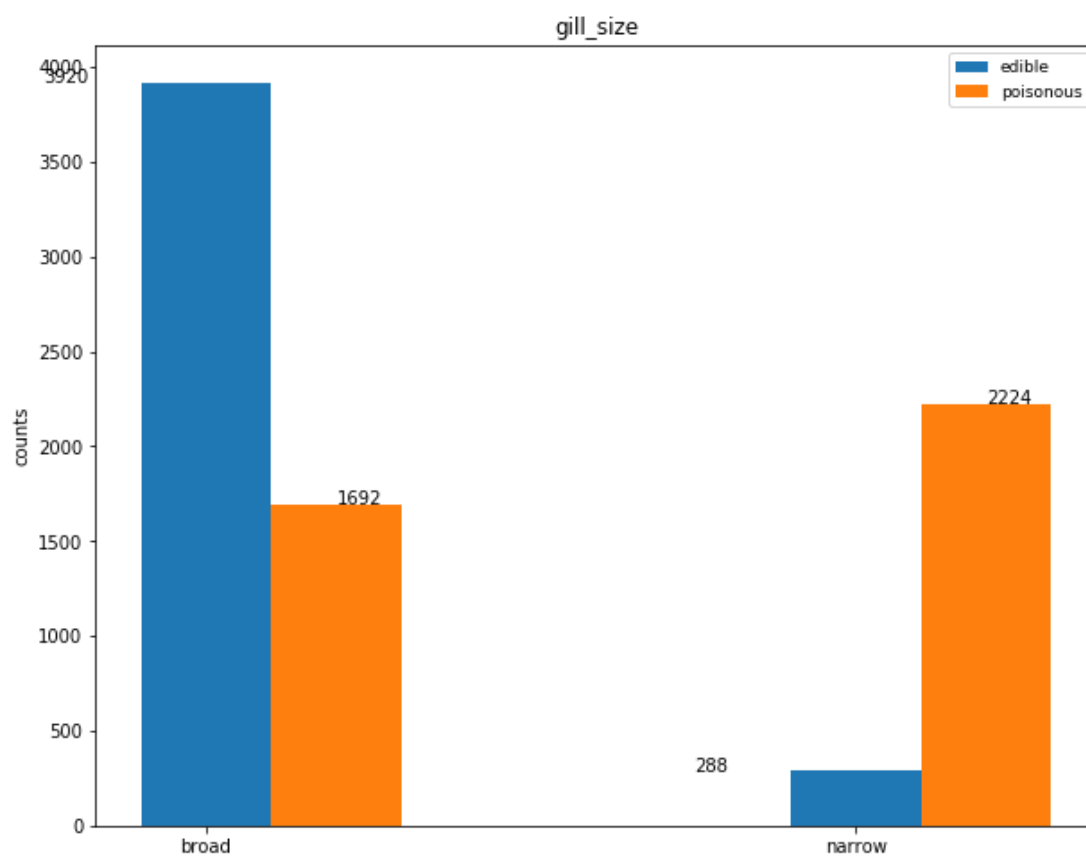
Mushroom :split data based on their labels and show the data distribution again:

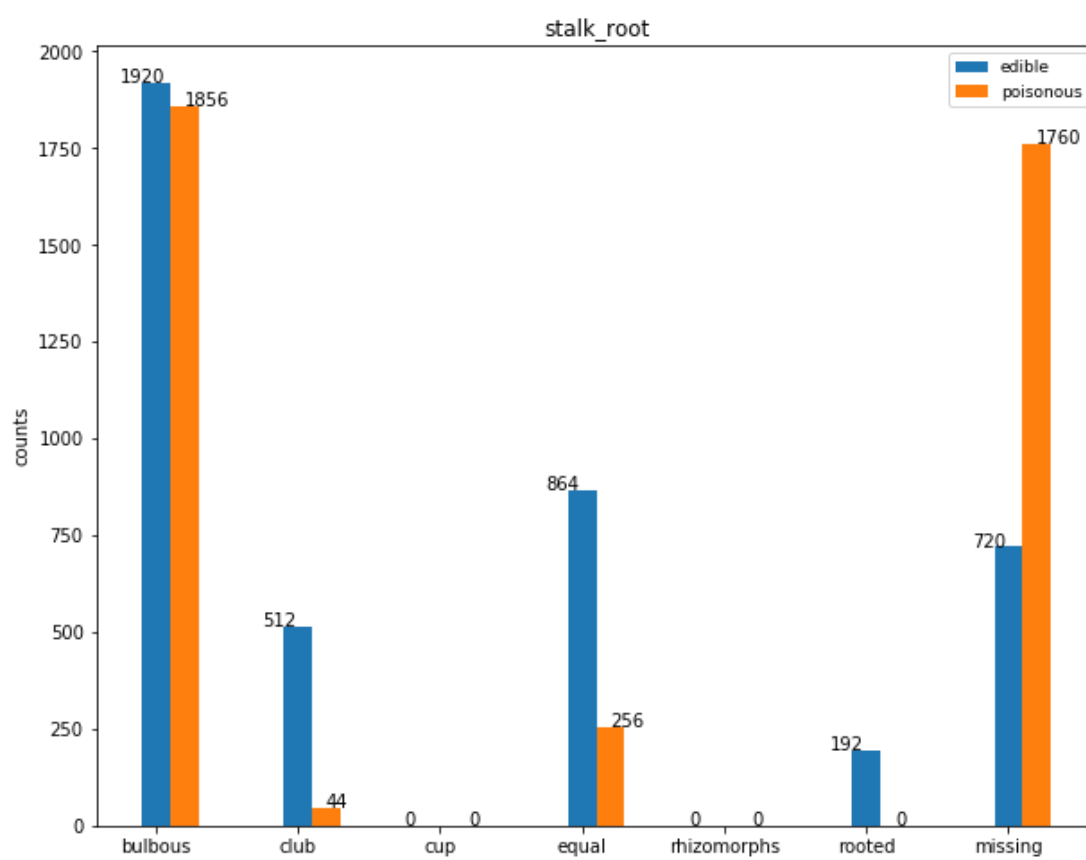
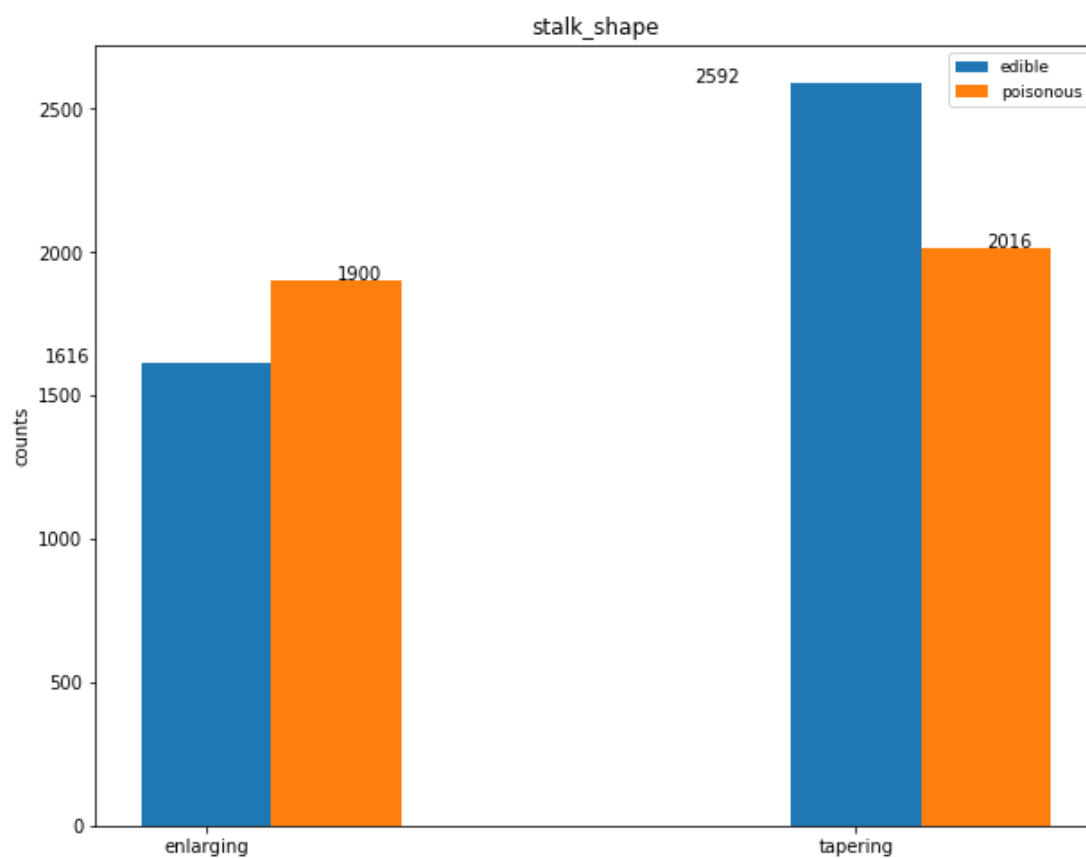


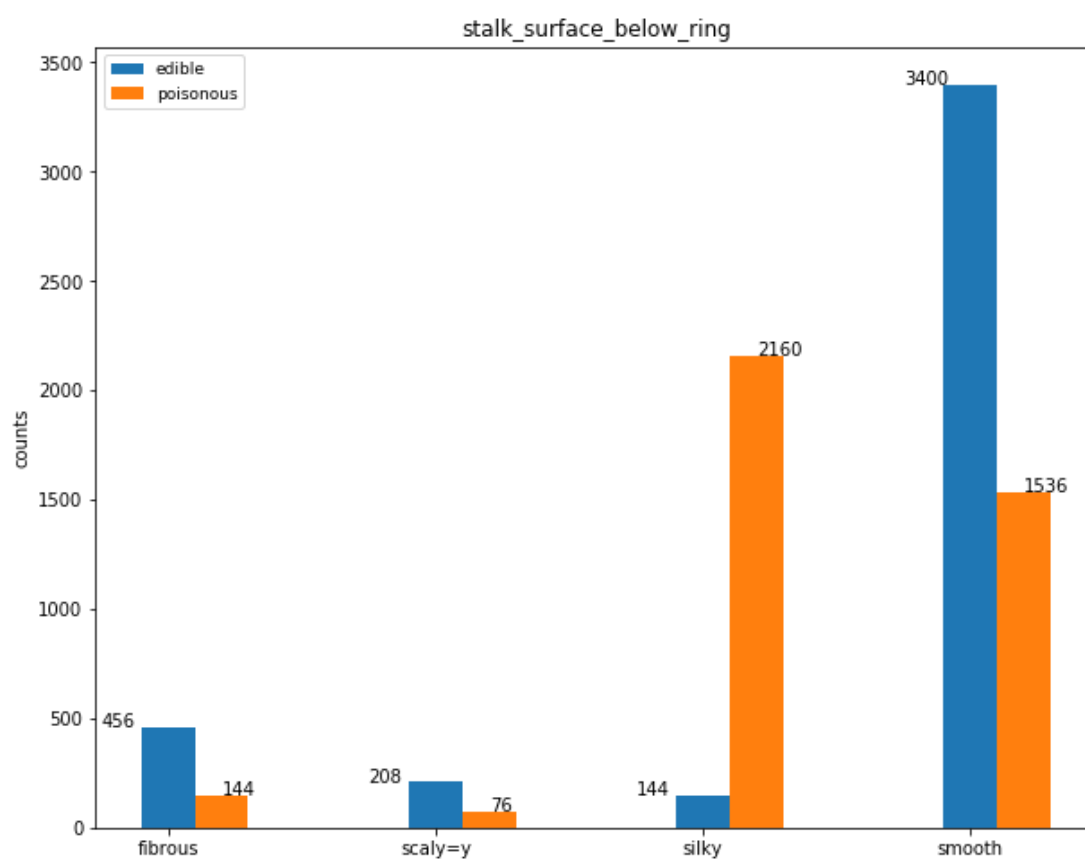
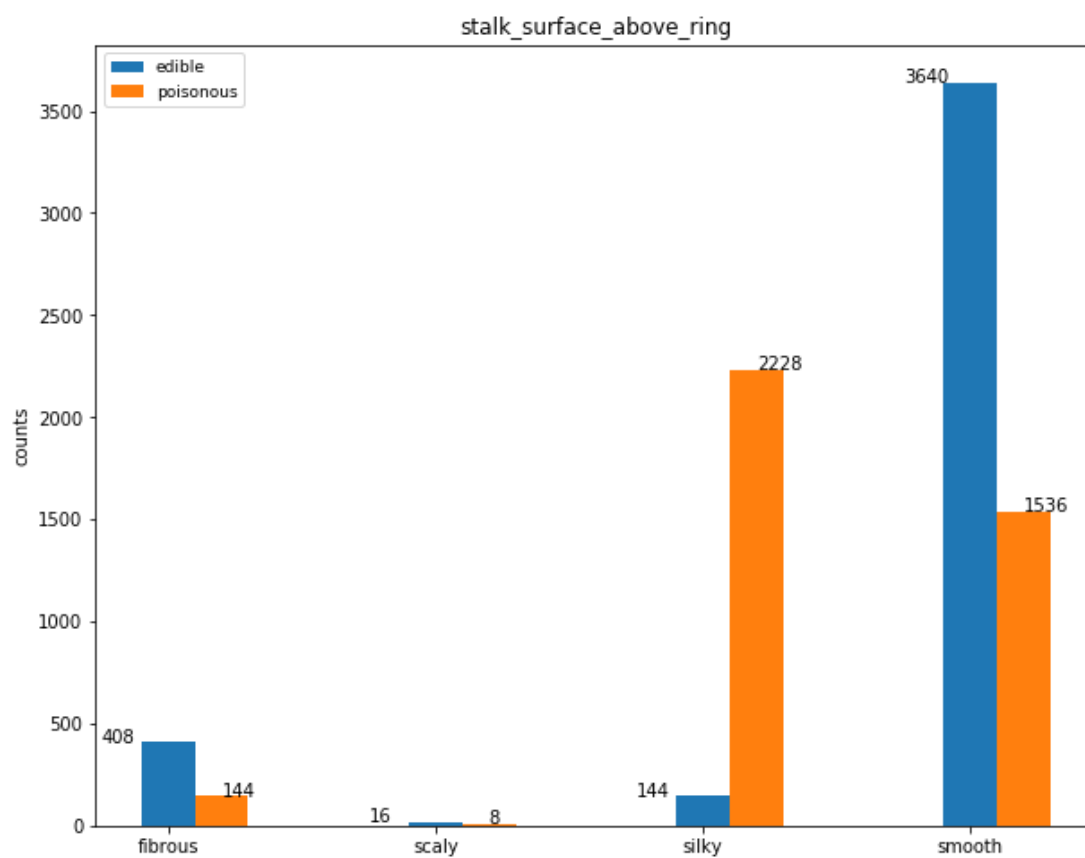


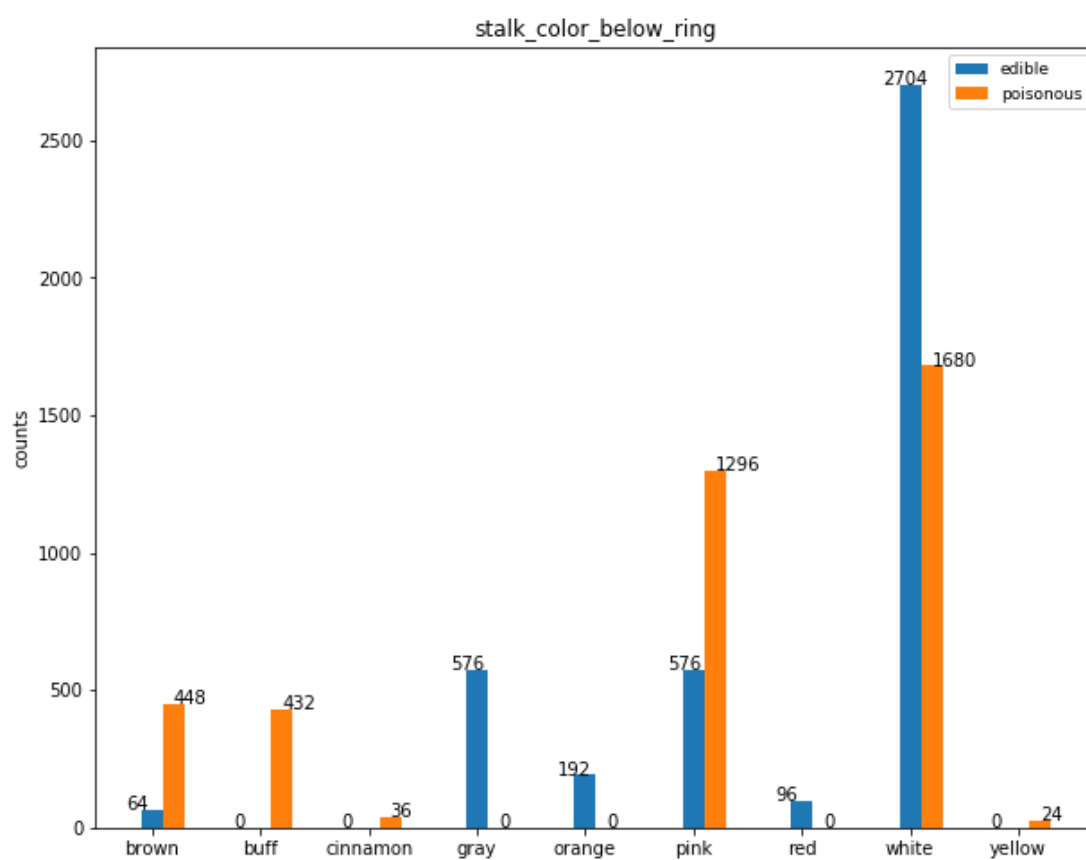
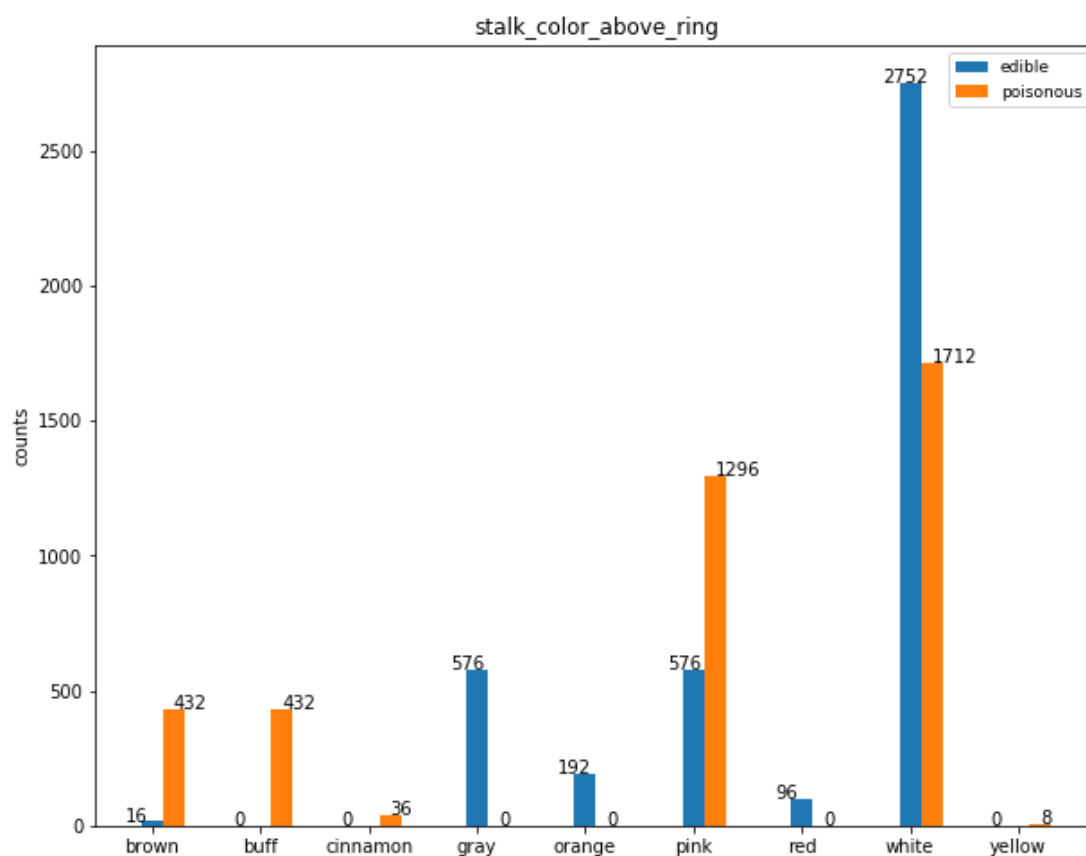


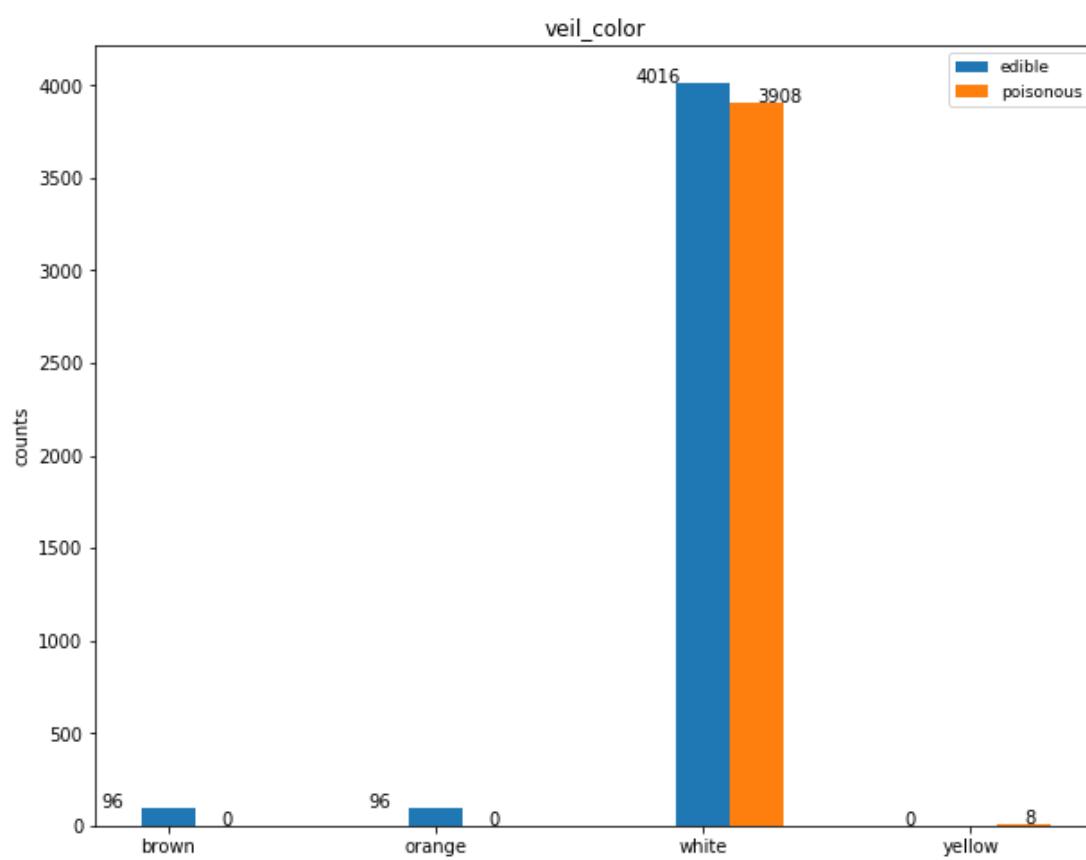
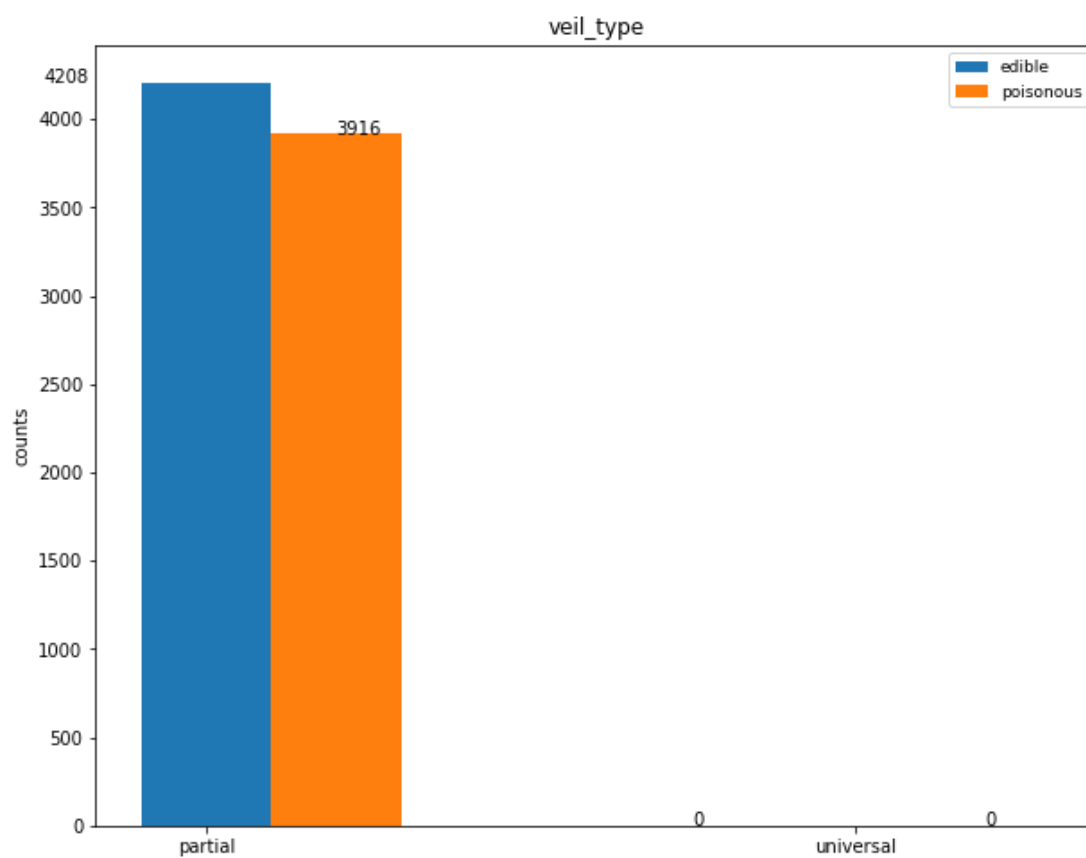


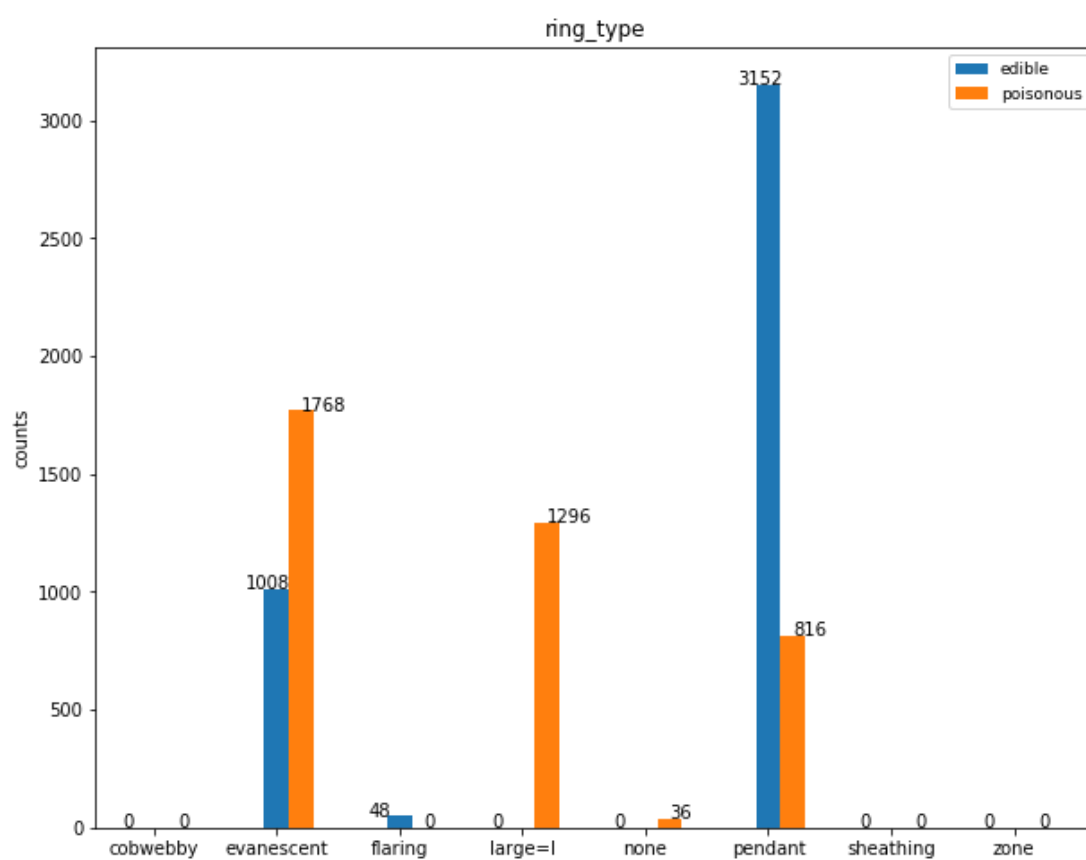
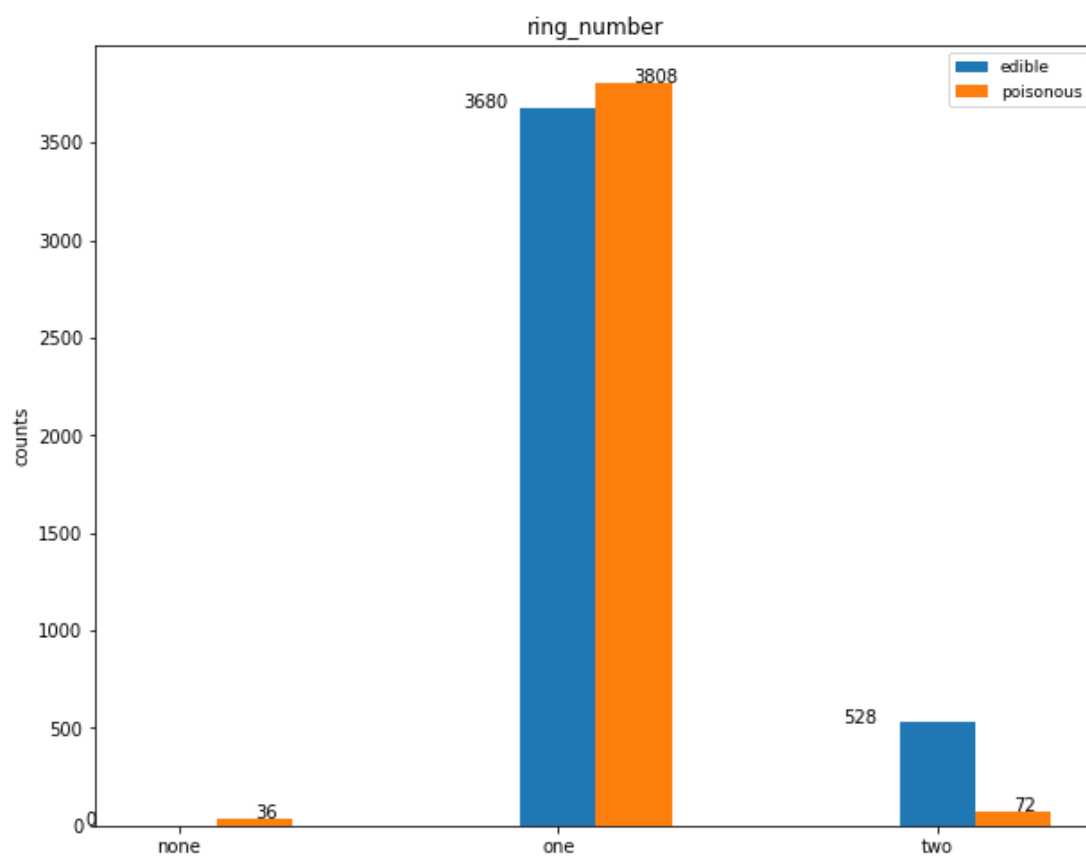


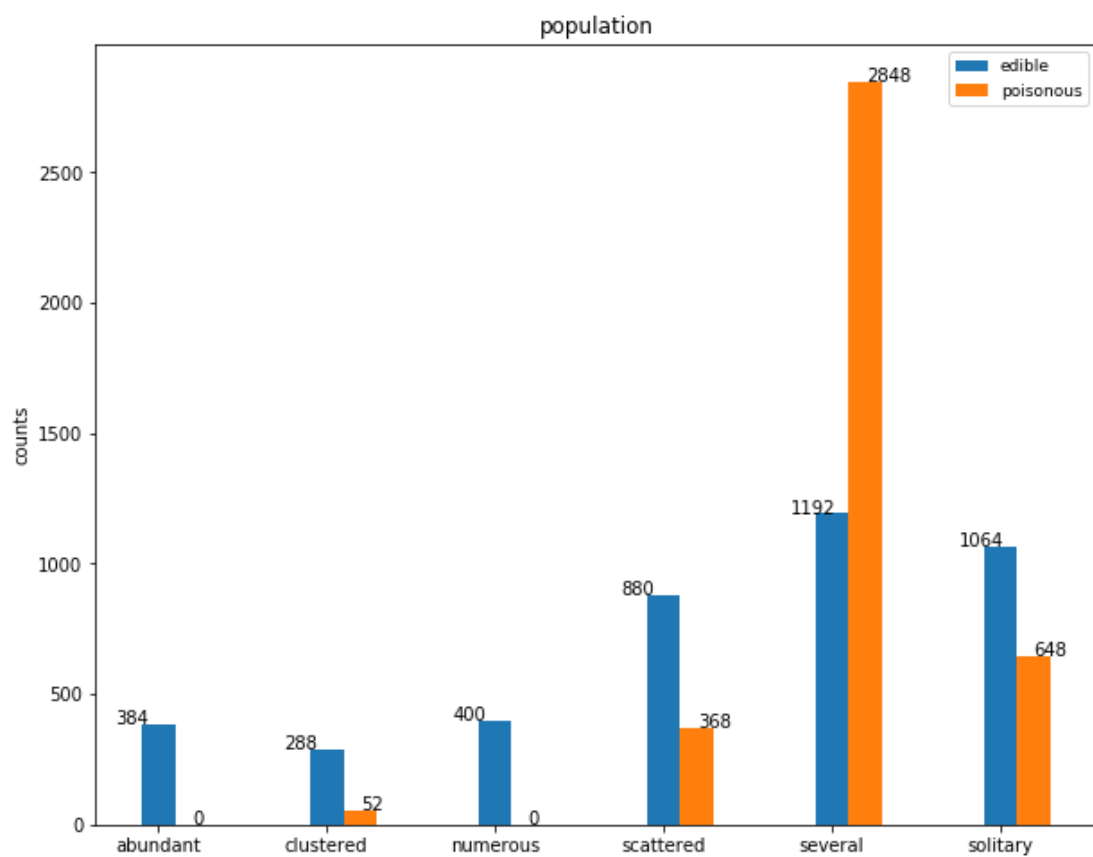
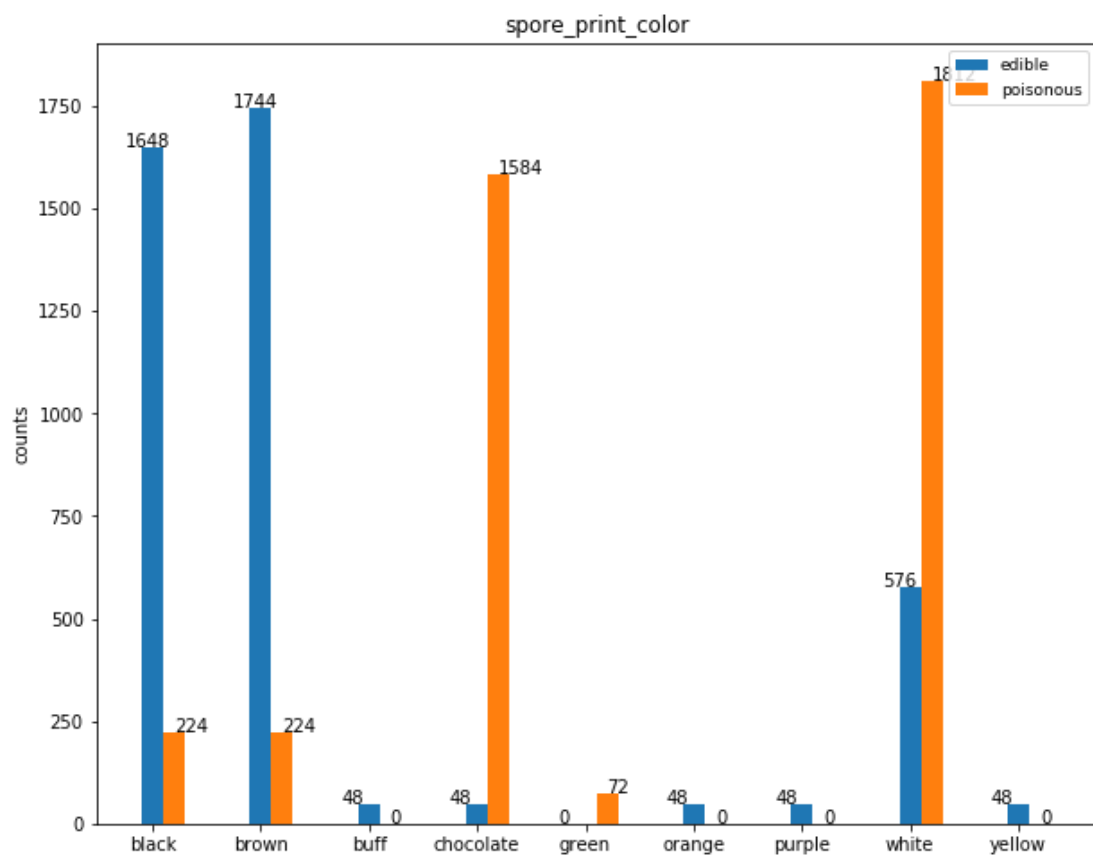


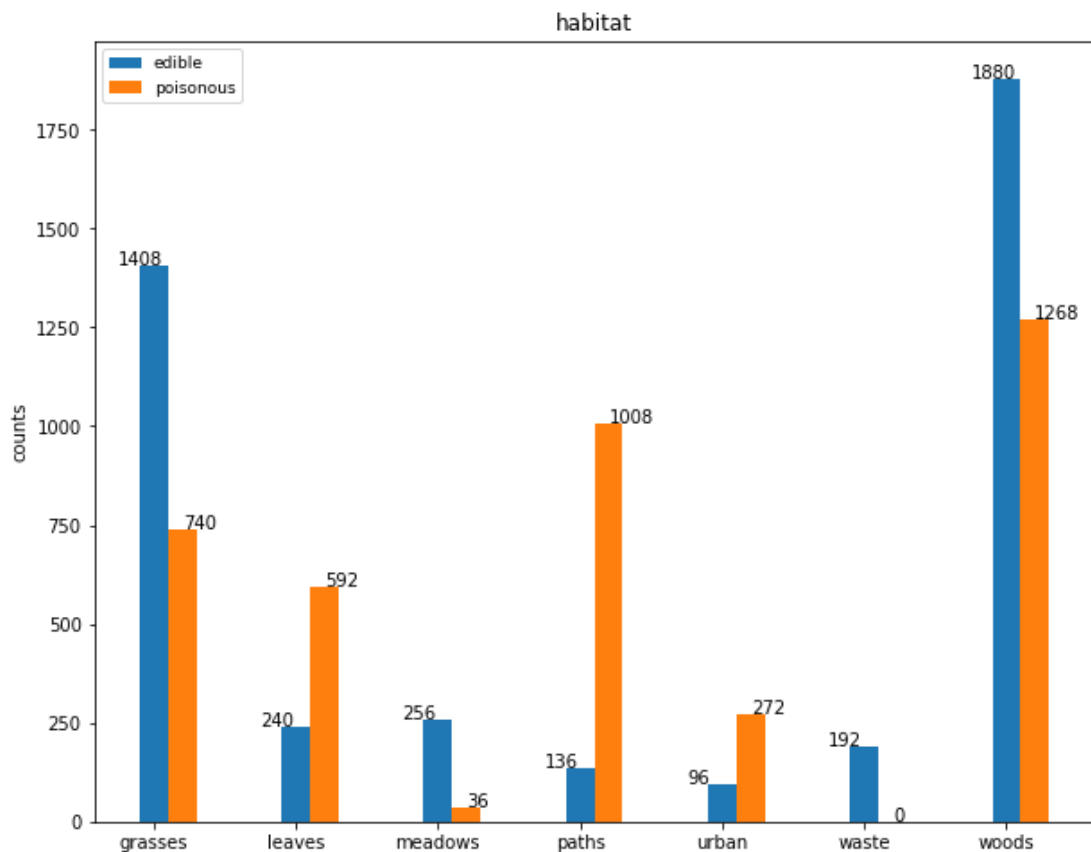






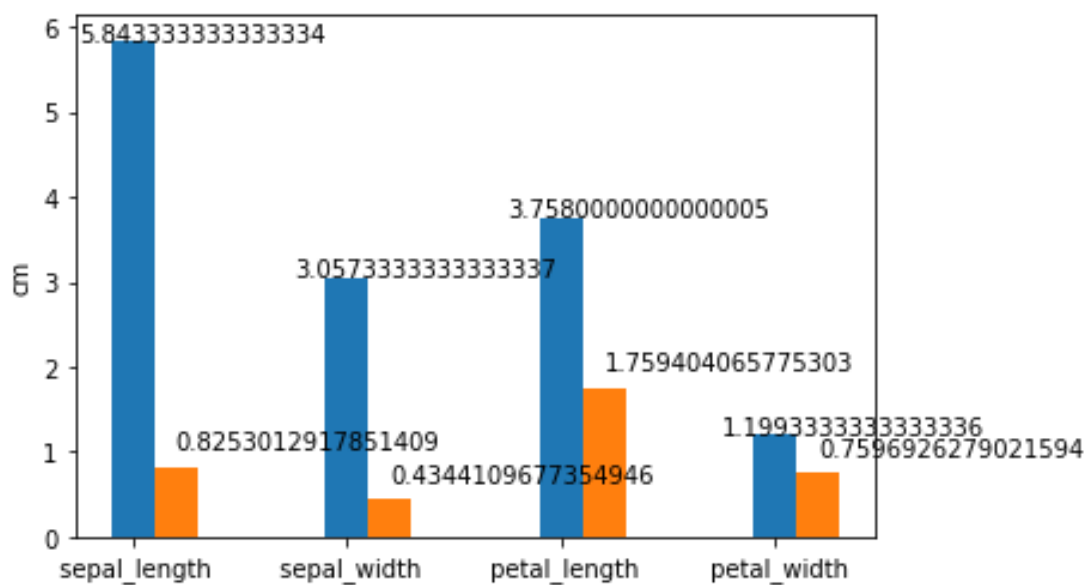




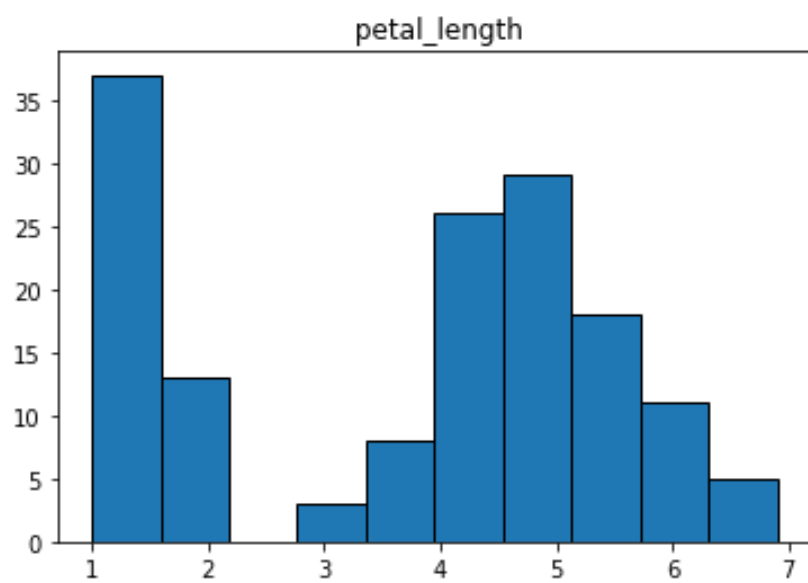
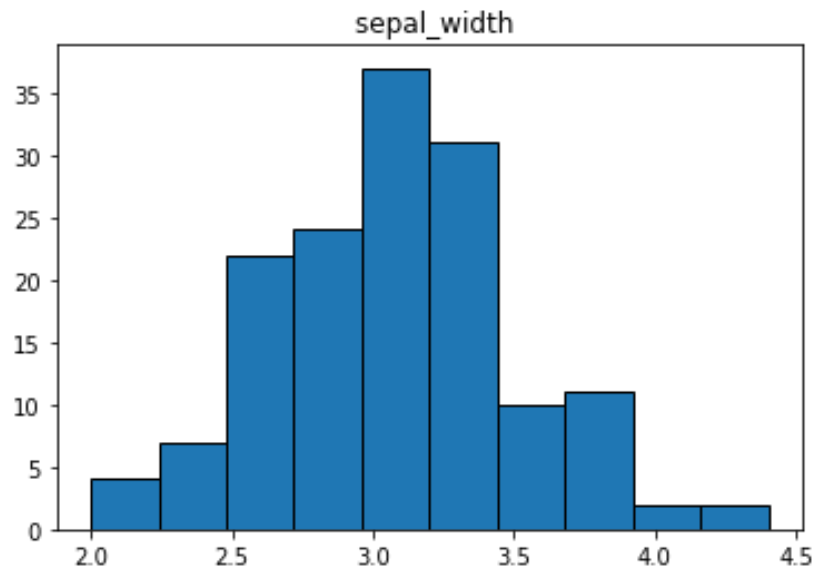
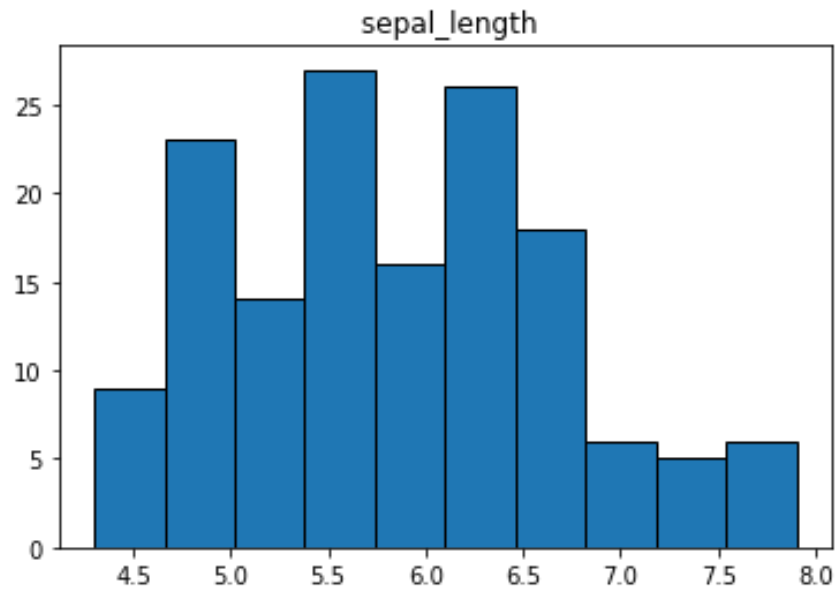


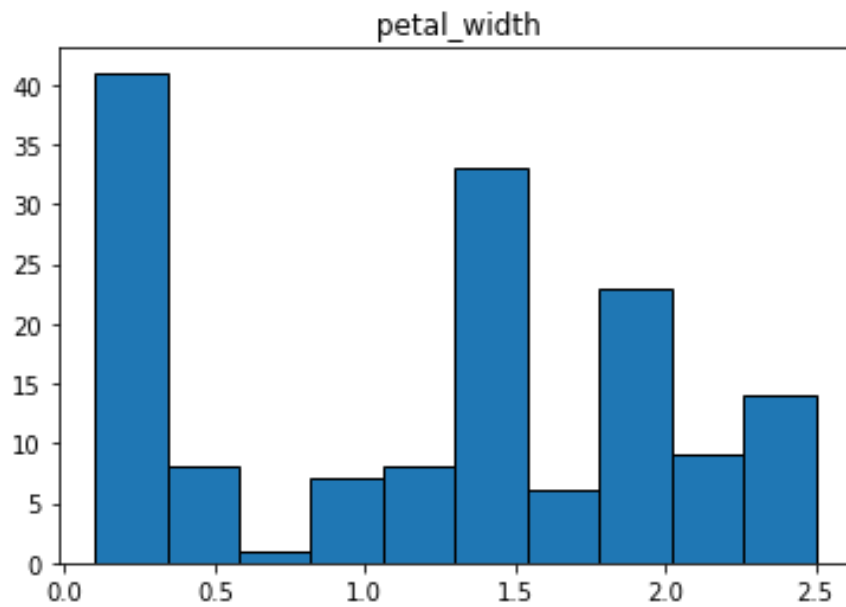
Iris dataset:

Blue : average **orange**: standard deviation



Value frequency of each feature:



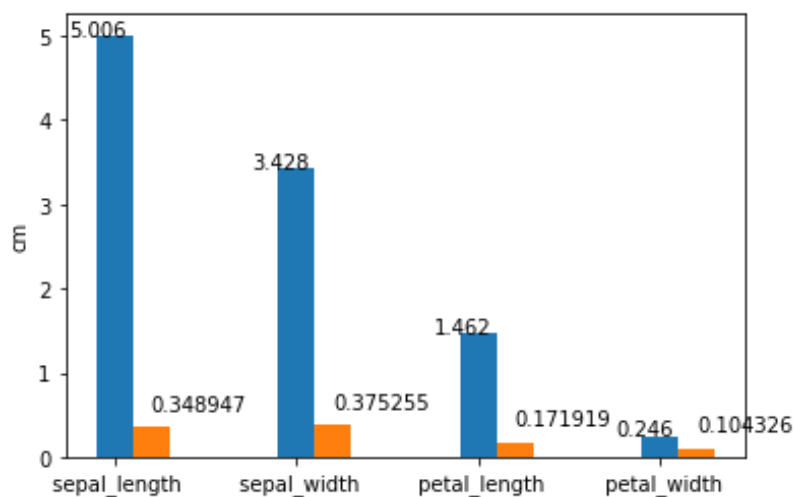


Iris dataset with split data based on their labels:

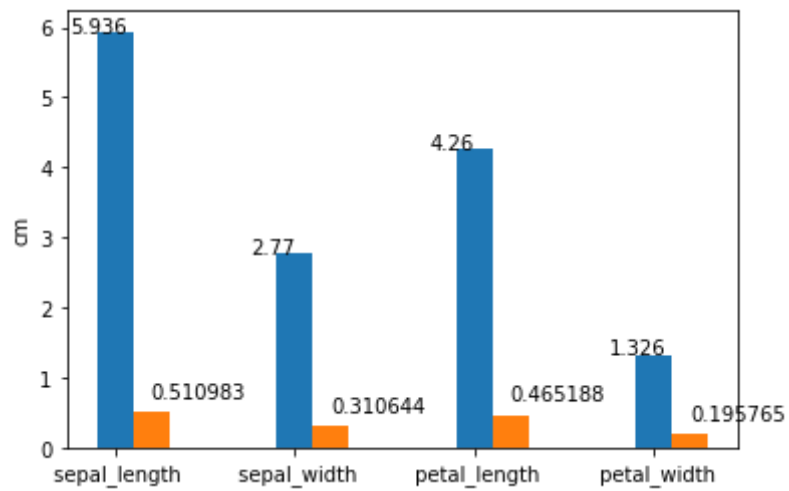
average, standard deviation:

Blue : average orange: standard deviation

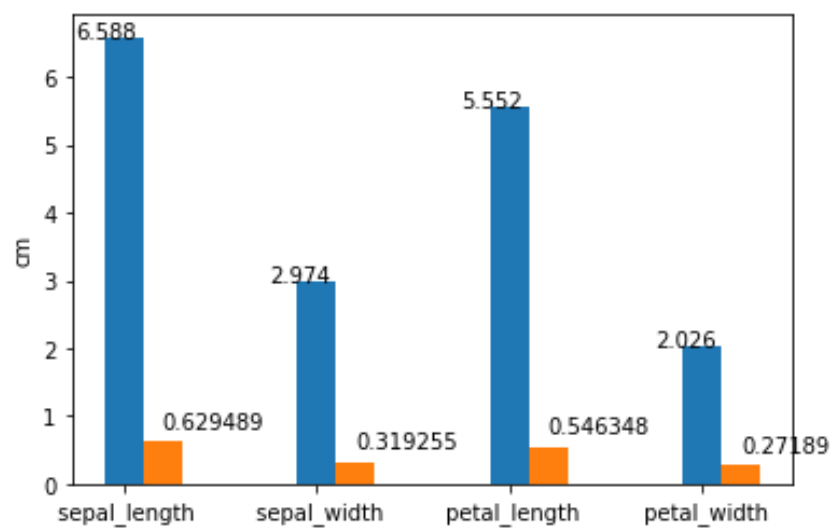
Iris-setosa:



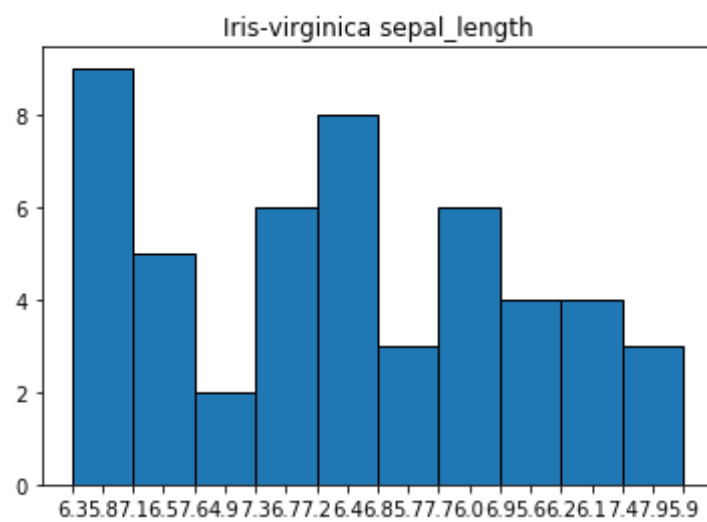
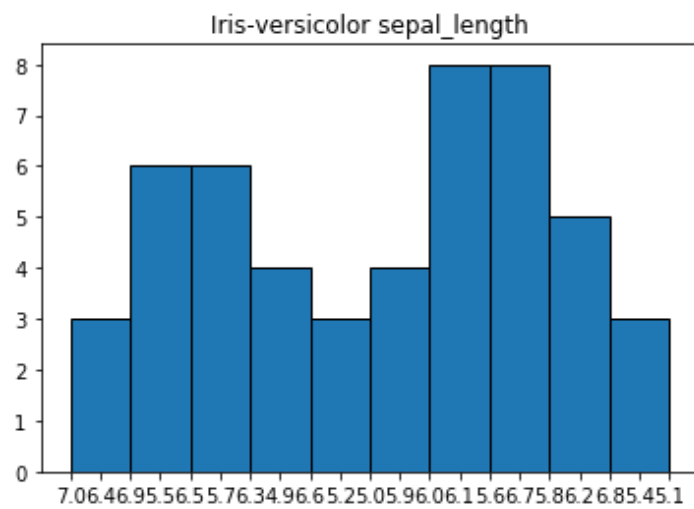
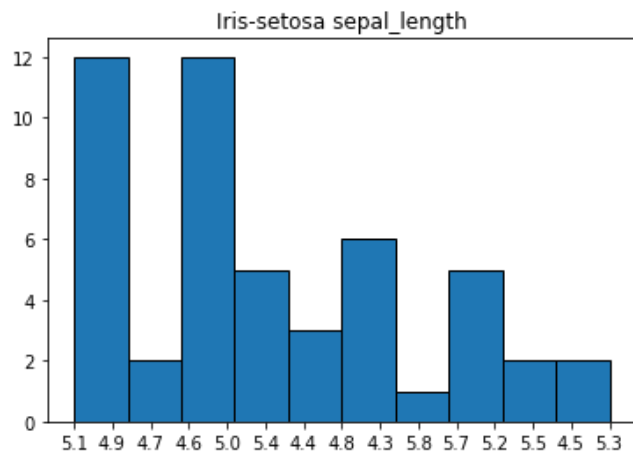
Iris-versicolor:

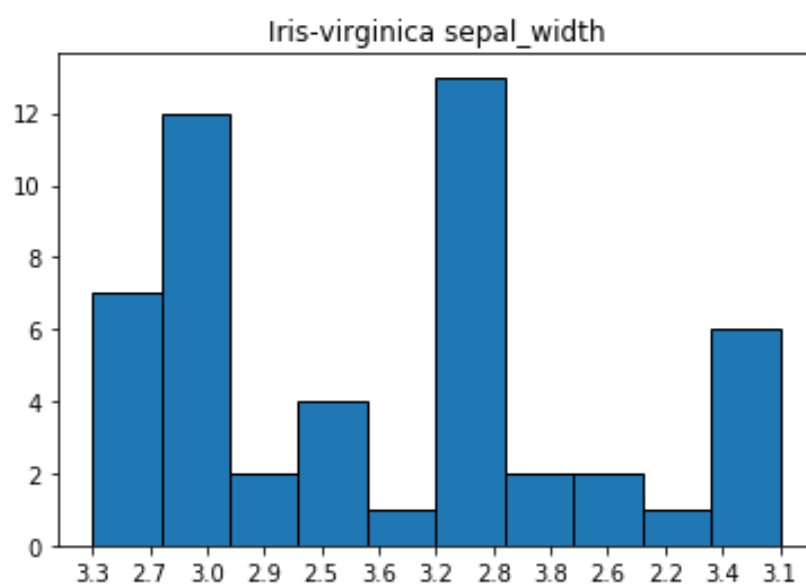
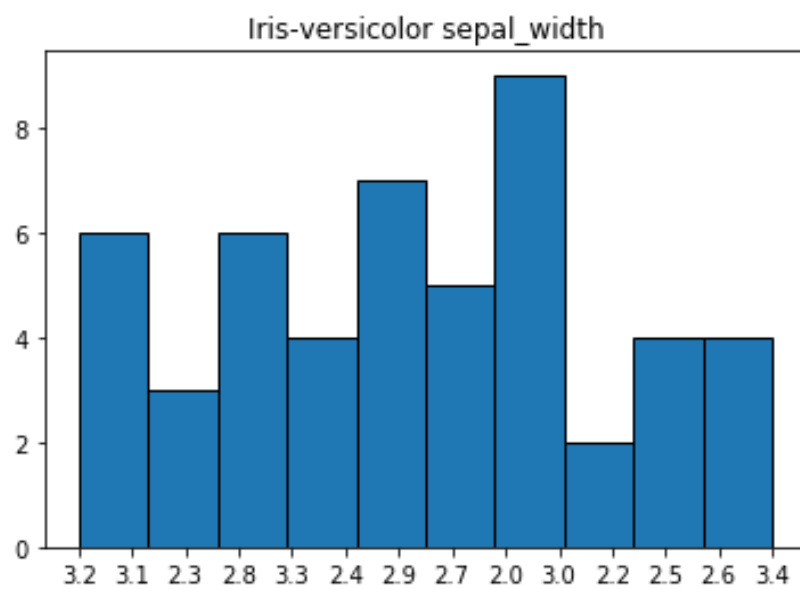
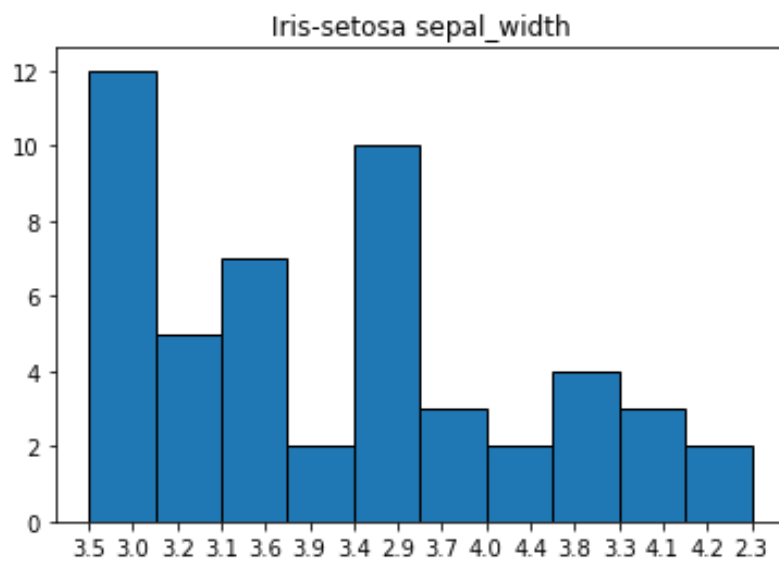


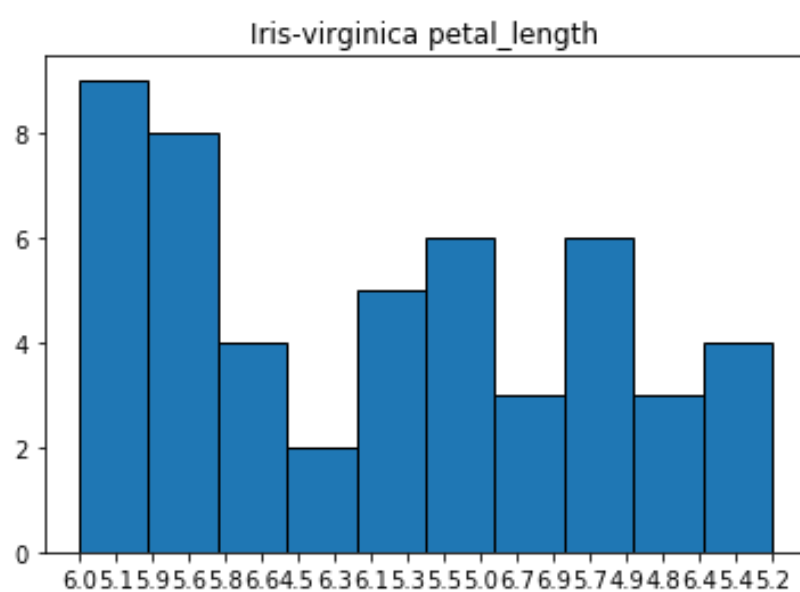
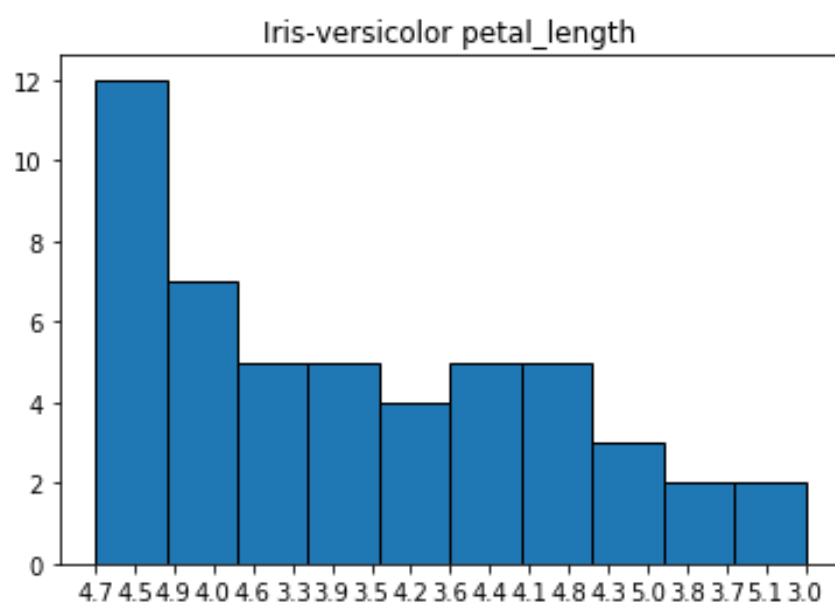
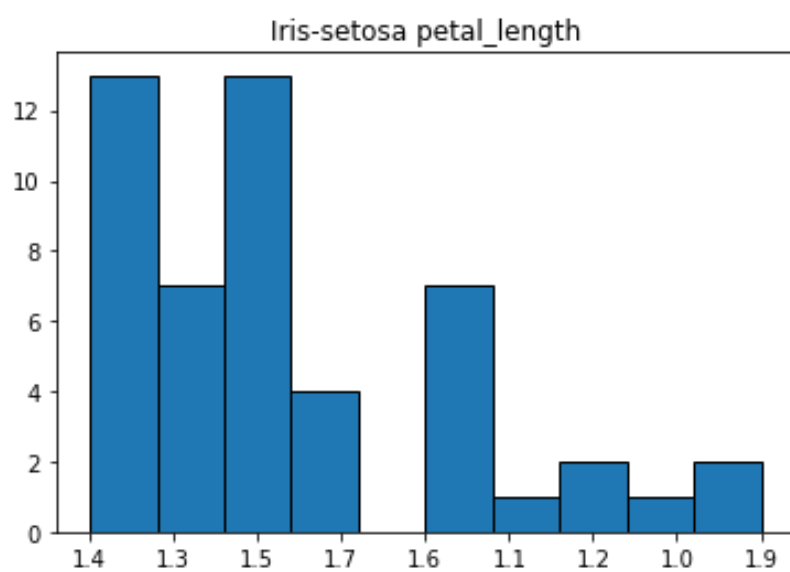
Iris-virginica

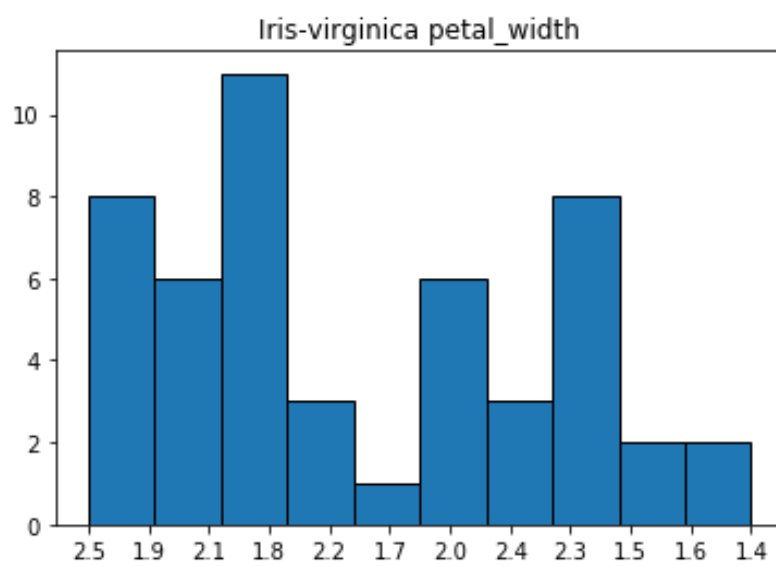
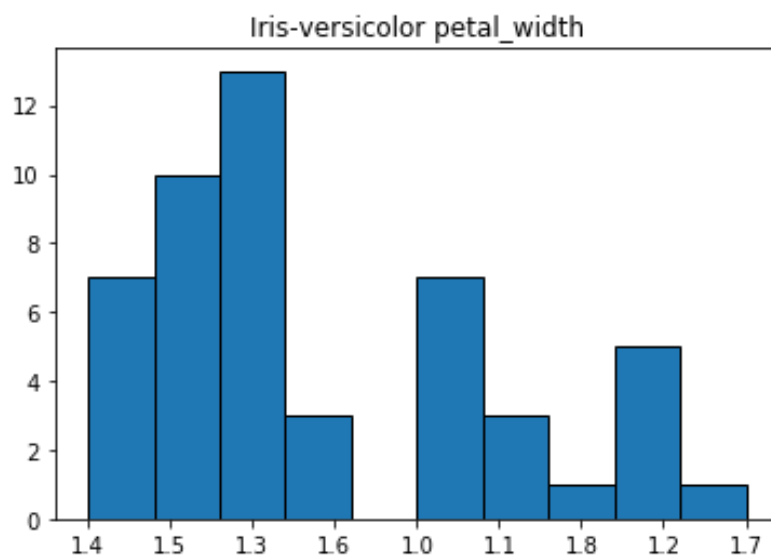
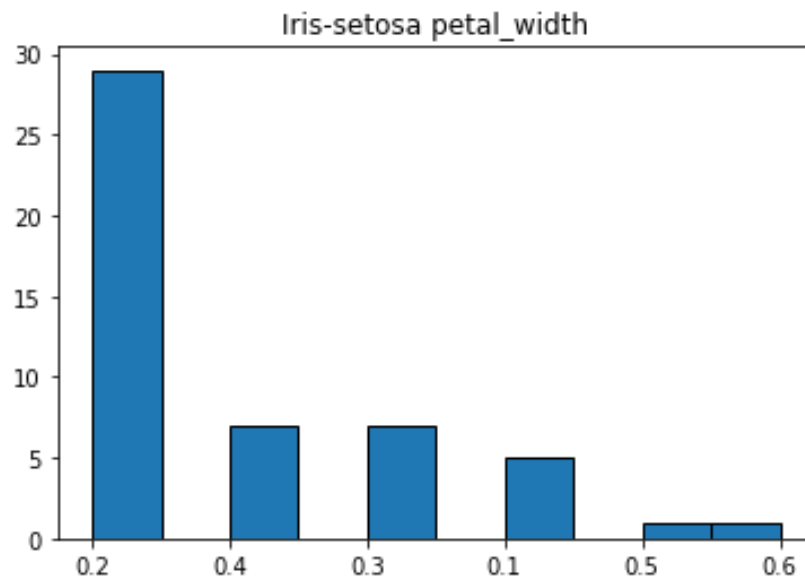


Each label's frequency:









3. Data preprocessing:

Drop feature:

don't use stalk_root to build classifier in my code because it has missing value.

Shuffle the data:

Use random.shuffle() in python to shuffle the data.

3.Data Preprocessing+4.Model Construction
+5.Train-Test-Split in code +6.**Results**
hw1_3.py,hw1_4.py.

6 result:

Mushroom:

```

Holdout validation without laplace
confusion matrix:
 1269  1
  17 1150
Classification accuracy: 0.9926138695116947
Sensitivity: 0.9992125984251968
Precision: 0.9867807153965785

Holdout validation with laplace
confusion matrix:
 1269  79
  17 1072
Classification accuracy: 0.9606073040623717
Sensitivity: 0.9413946587537092
Precision: 0.9867807153965785

K-fold cross-validation without laplace
confusion matrix:
1394.6666666666667 1.0
 8.0 1304.3333333333333
Classification accuracy: 0.9966765140324964
Sensitivity: 0.9992809104502295
Precision: 0.9943115660075713

K-fold cross-validation with laplace
confusion matrix:
1394.0 103.33333333333333
 8.666666666666666 1202.0
Classification accuracy: 0.9586410635155097
Sensitivity: 0.9310220479762409
Precision: 0.9938374076386761

]: 1

```

Iris:

Above is Holdout validation with the ratio 7:3

Below is K-fold cross-validation with $K=3$

```

confusion matrix:
 13  0  0
  0 16  1
  0  2 13

Classification accuracy: 0.9333333333333333
setosa_sensitivity: 1.0
versicolor_sensitivity: 0.9411764705882353
virginica_sensitivity: 0.8666666666666667
setosa_precession: 1.0
versicolor_precession: 0.8888888888888888
virginica_precession: 0.9285714285714286

confusion matrix:
16.666666666666668 0.0 0.0
 0.0 15.666666666666666 1.0
 0.0 1.3333333333333333 1.3333333333333333

Classification accuracy: 0.6733333333333335
setosa_sensitivity: 1.0
versicolor_sensitivity: 0.9400000000000001
virginica_sensitivity: 0.5
setosa_precession: 1.0
versicolor_precession: 0.9215686274509803
virginica_precession: 0.5714285714285715

```

[]:

7.Comparison & Conclusion:

如果有做 laplace smoothing 準確度會下降，但是可以處理 train data feature 中有項目為 0 的情況，K-fold cross-validation 跟 Holdout validation 起來多了小數點，因為它是多次平均綜合，感覺結果比較讓人信服。