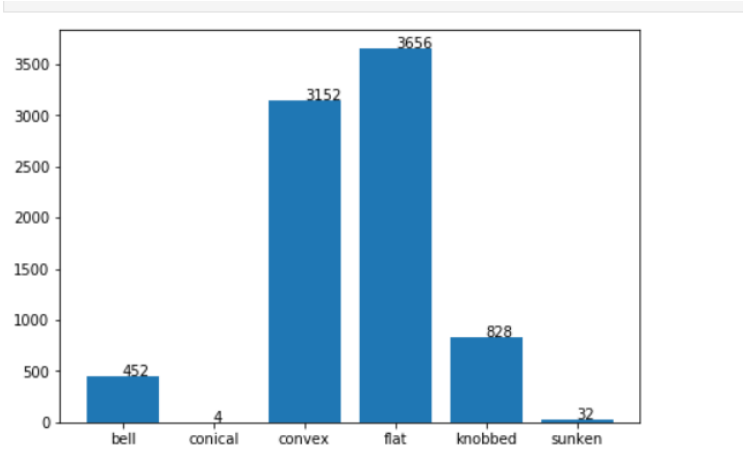
2.data visualization (hw1_1.ipynb,hw1_2.ipynb):

hw1_1.ipynb 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。

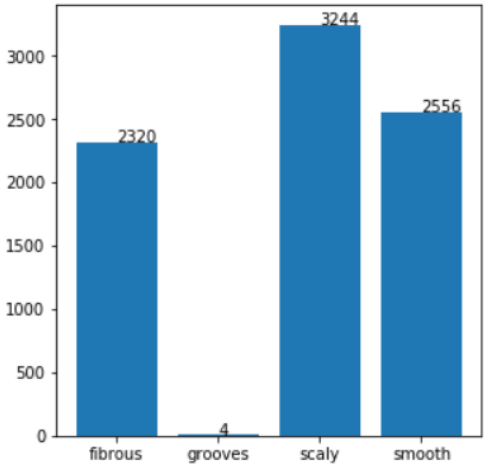
Mushroom dataset:

Value frequency of every feature:

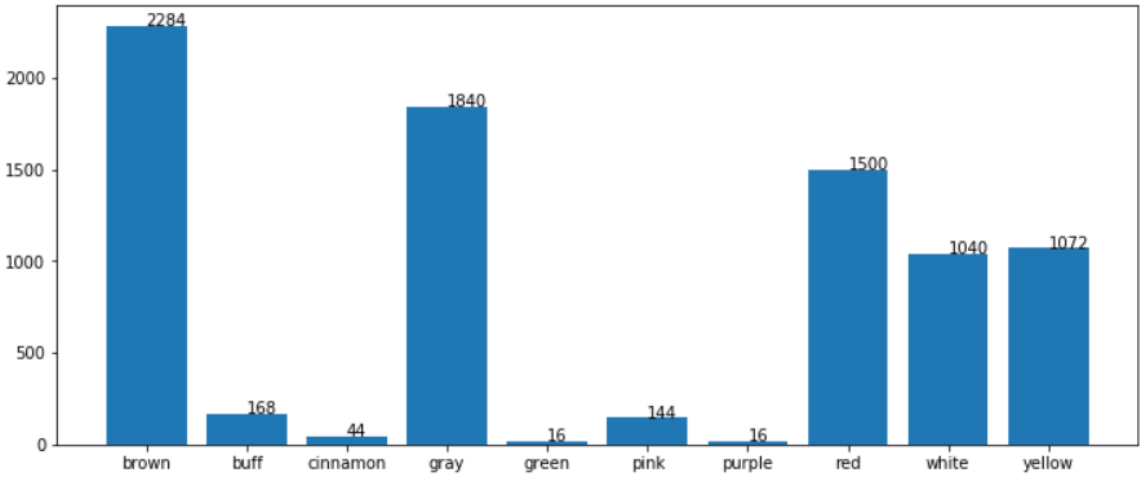
Cap_shape



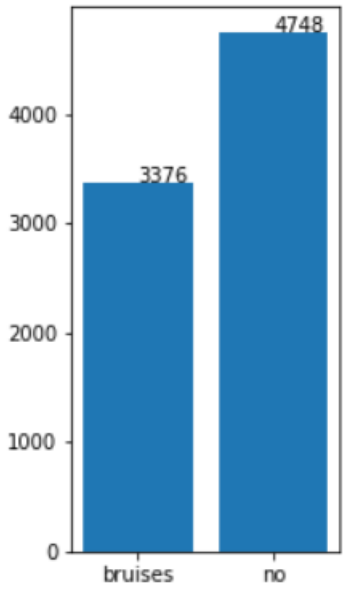
Cap_surface:



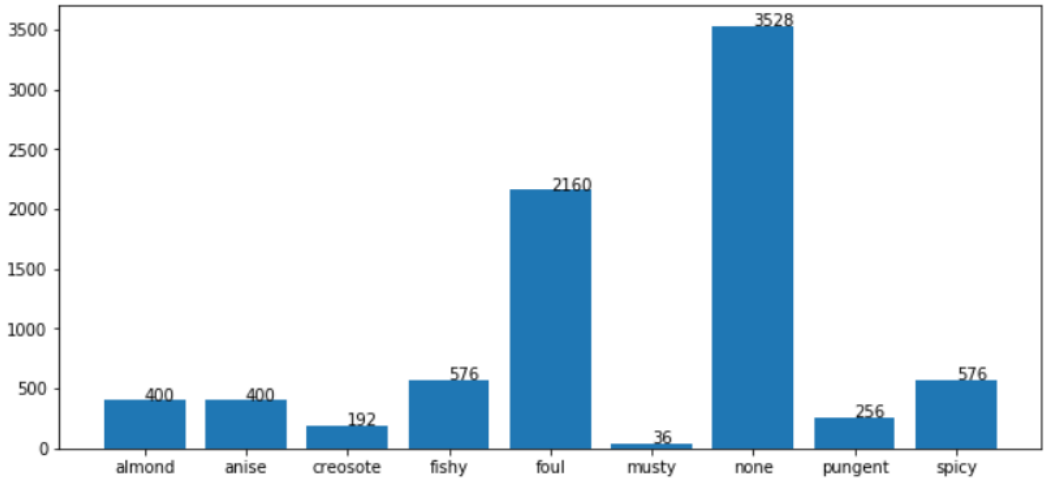
Cap_color:



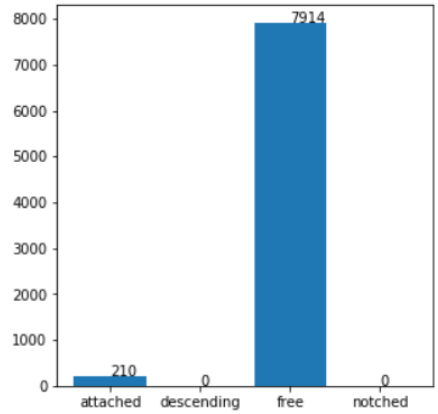
Bruises:



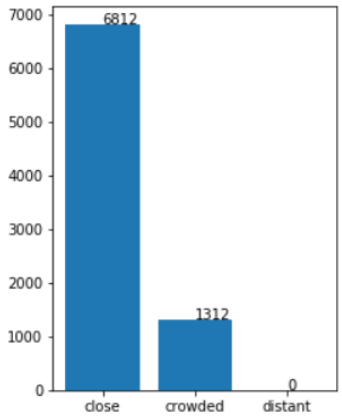
Odor:



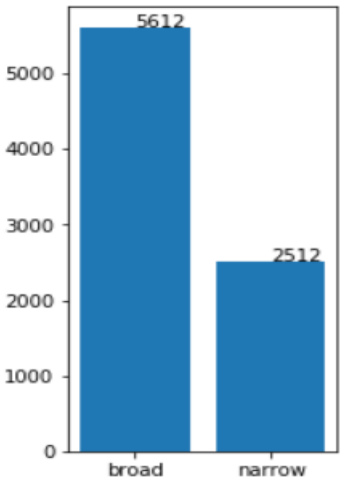
Gill_attachment:



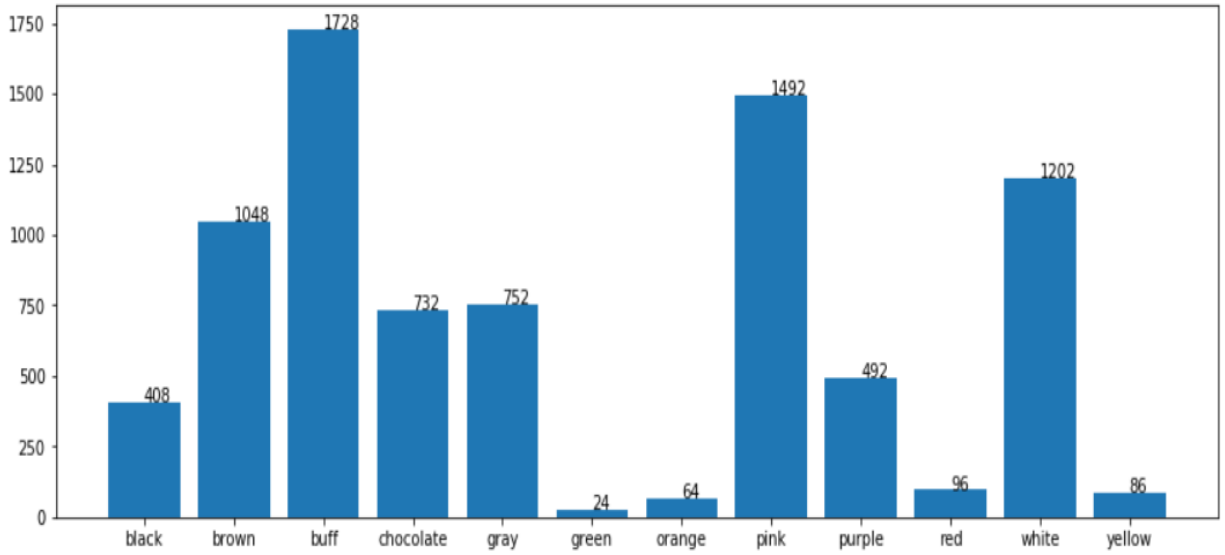
gill-spacing:



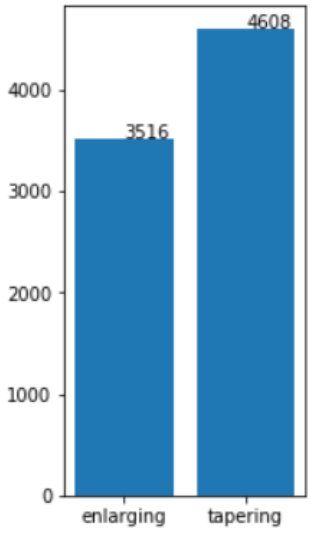
gill-size:



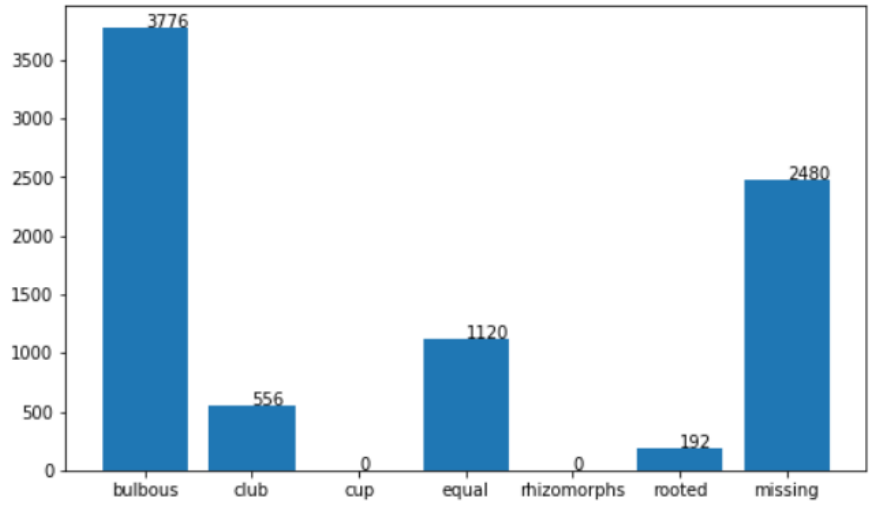
gill-color:



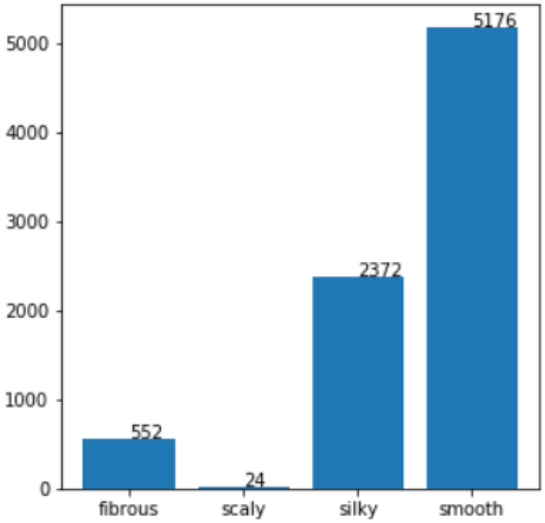
stalk-shape:



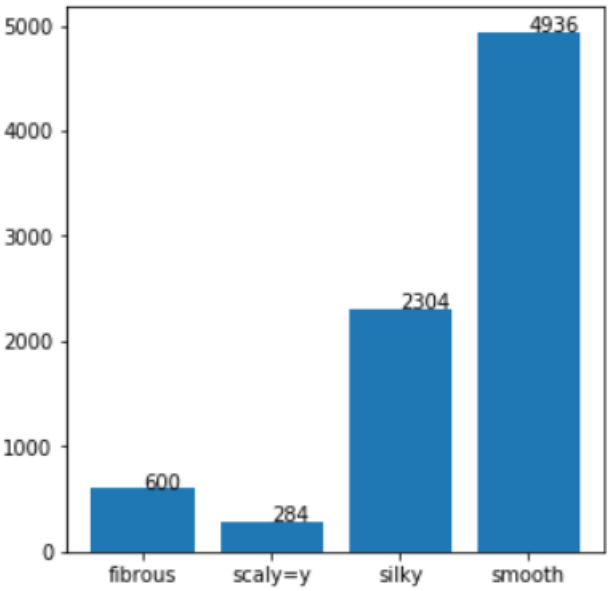
stalk-root:



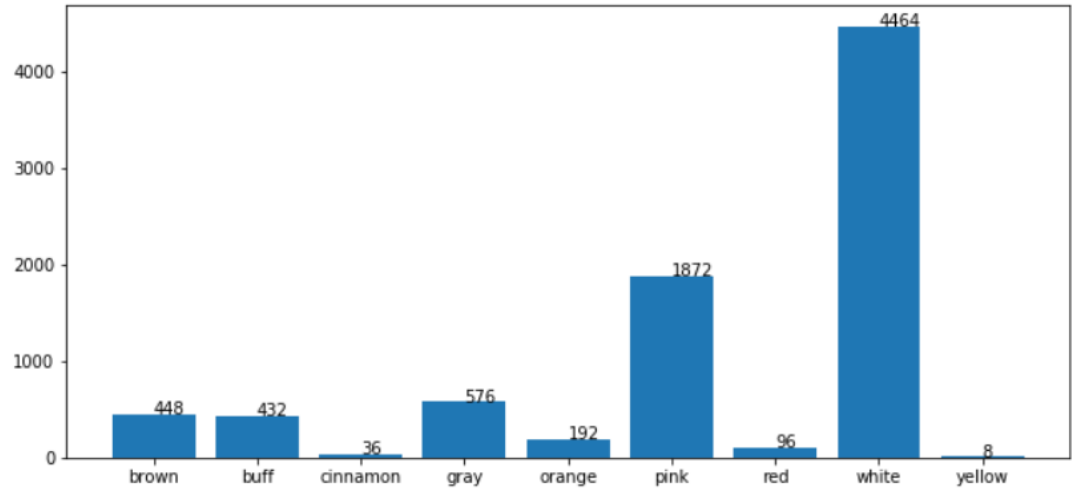
stalk-surface-above-ring:



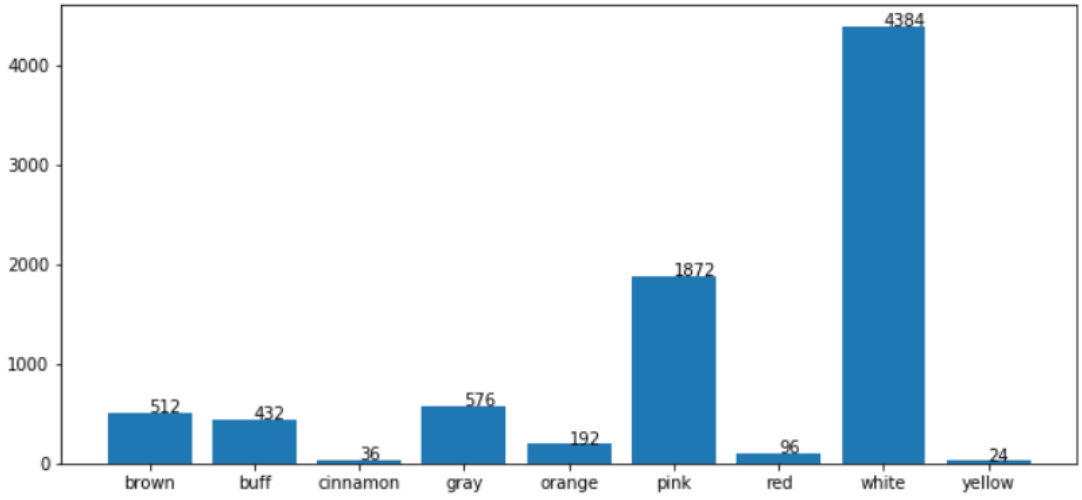
stalk-surface-below-ring:



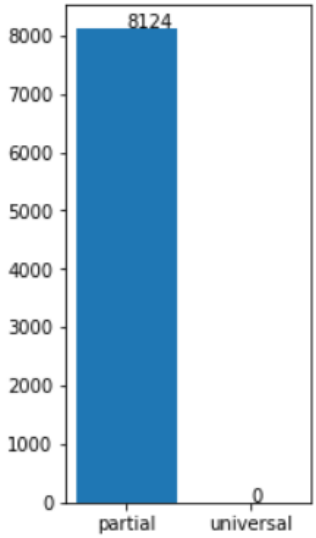
stalk-color-above-ring:



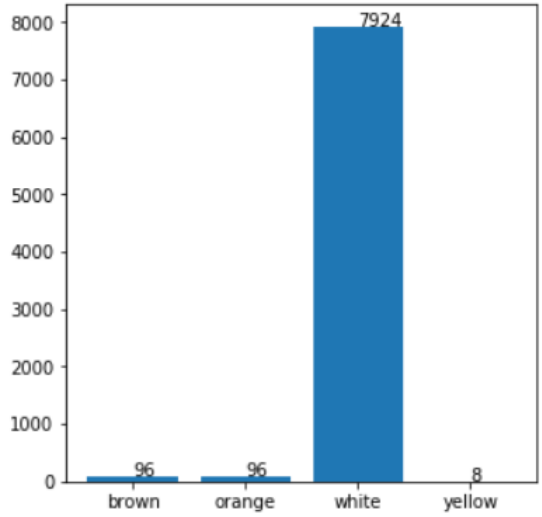
stalk-color-below-ring:



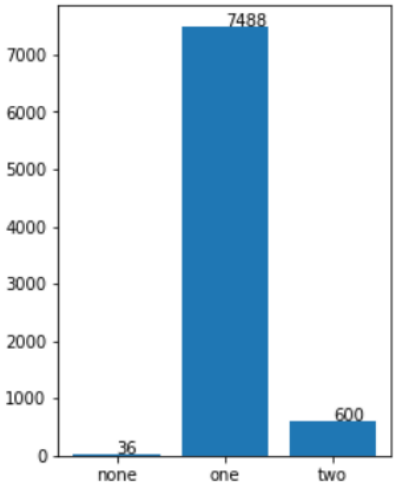
veil-type:



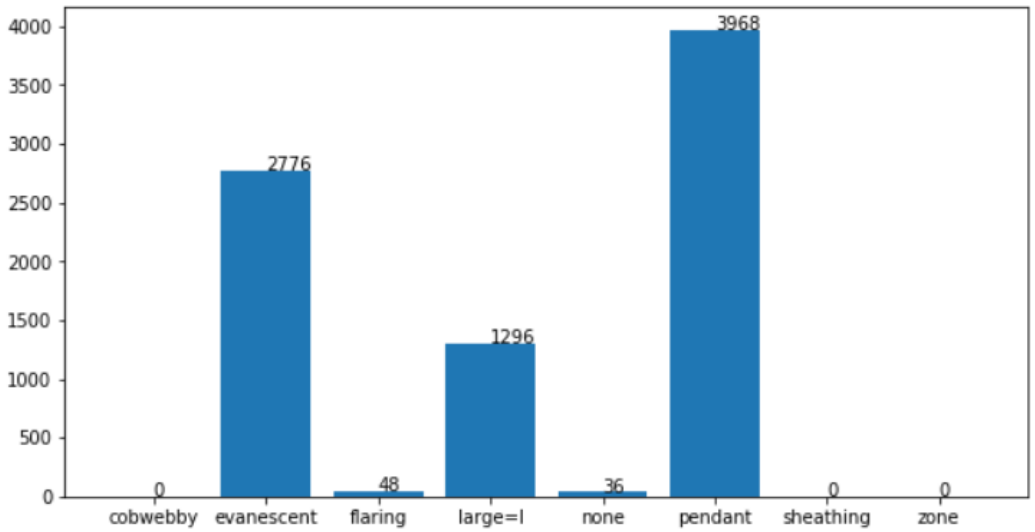
veil-color:



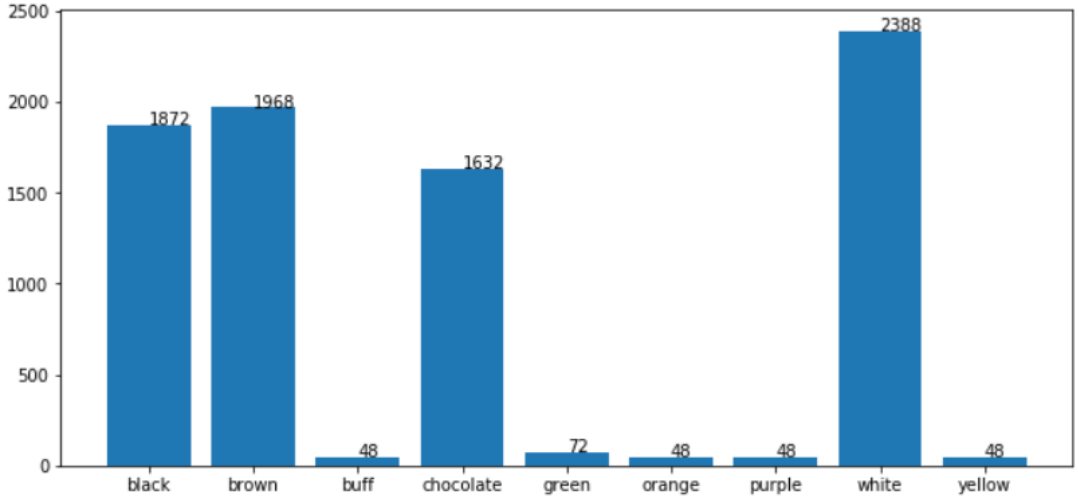
ring-number:



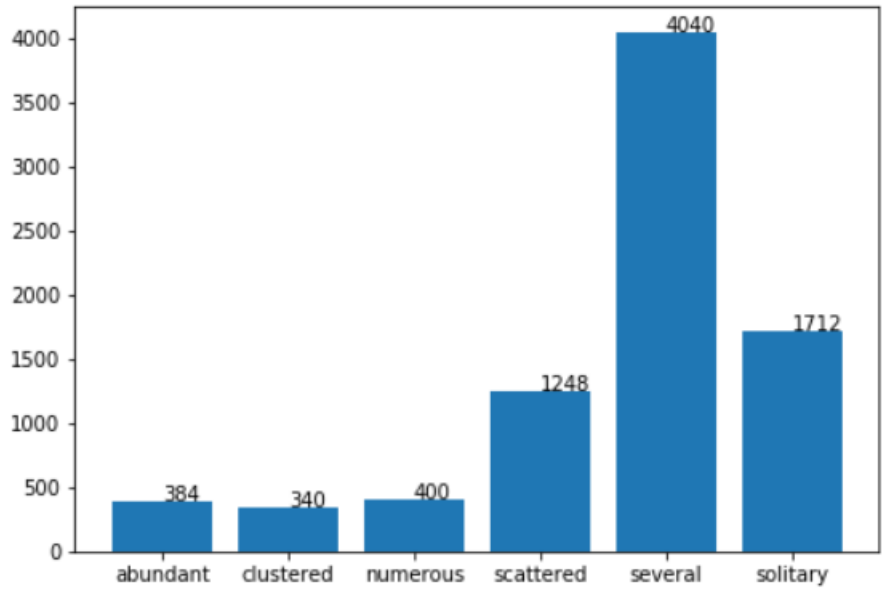
ring-type:



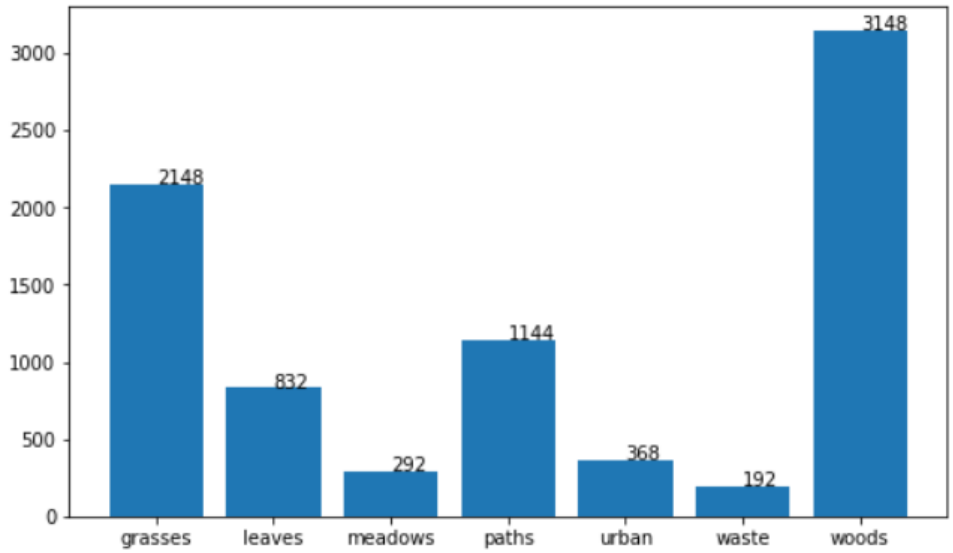
spore-print-color:



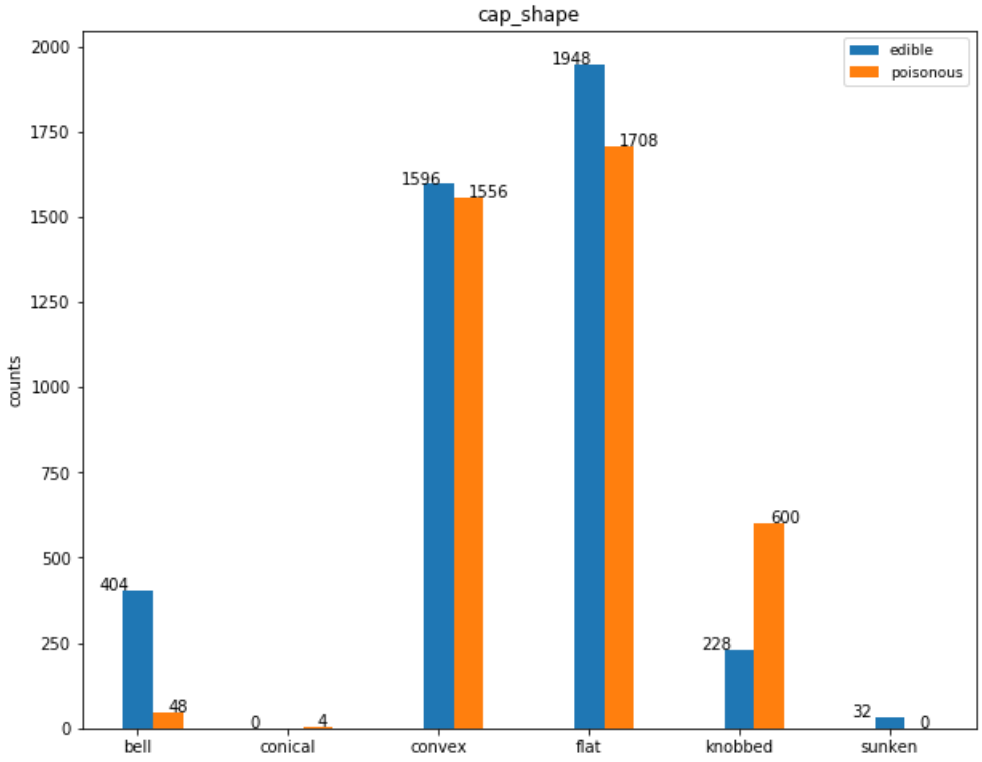
population:

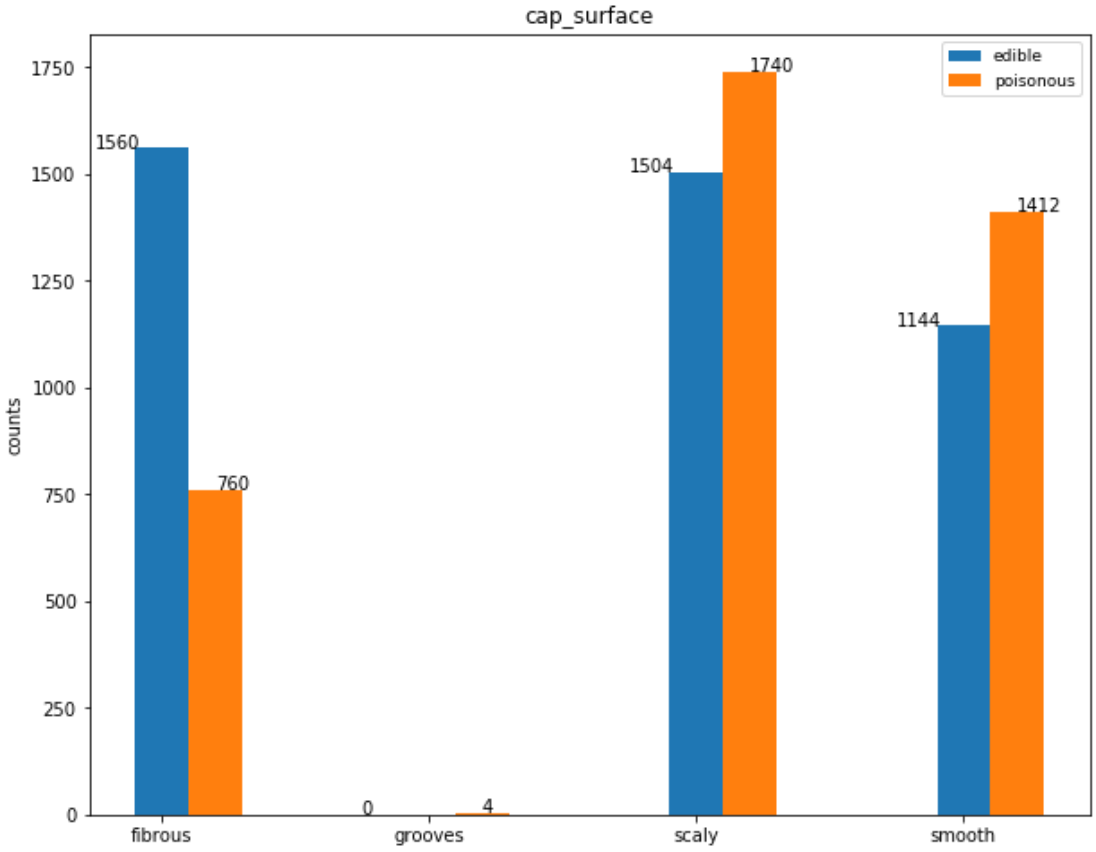


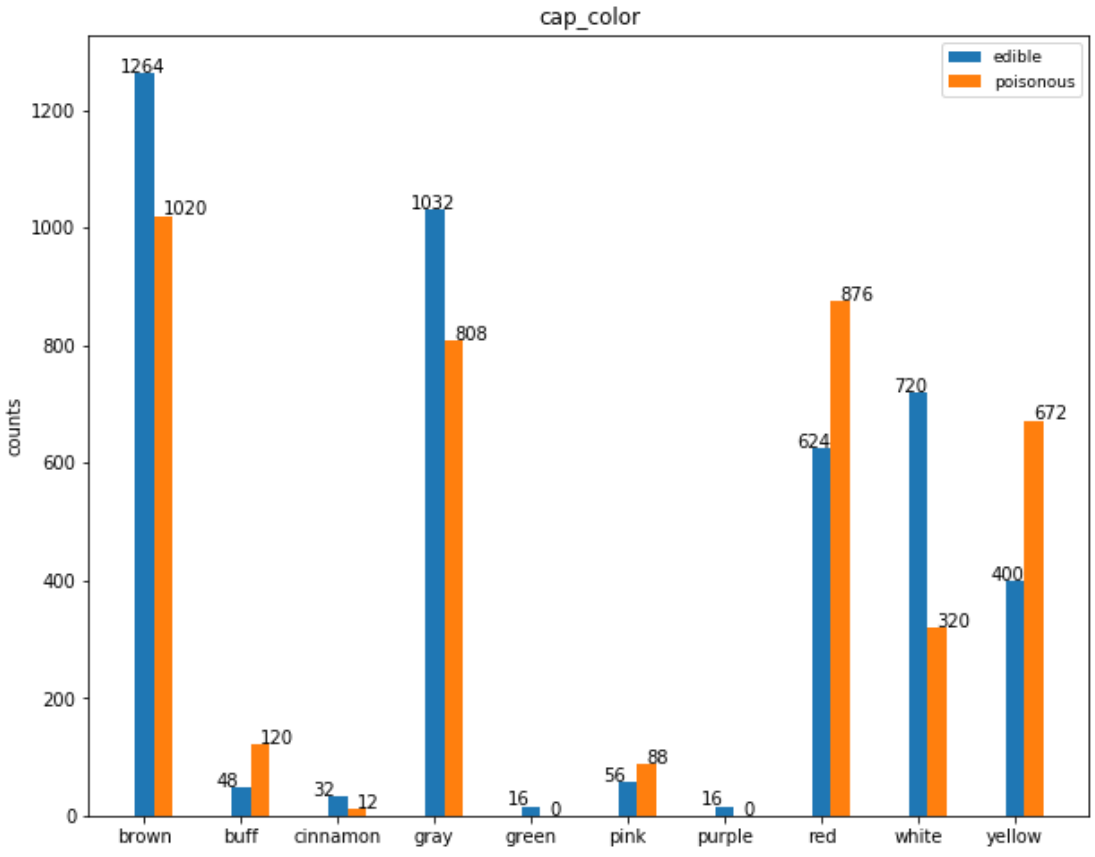
habitat:

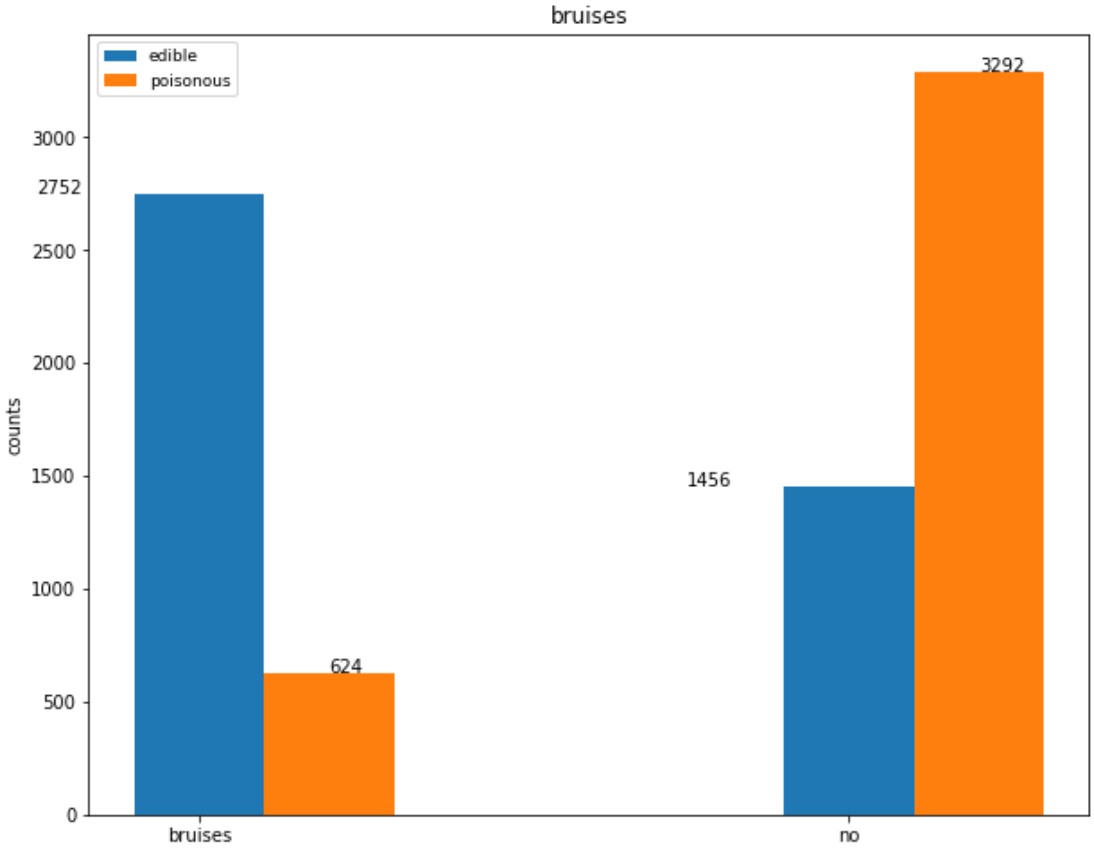


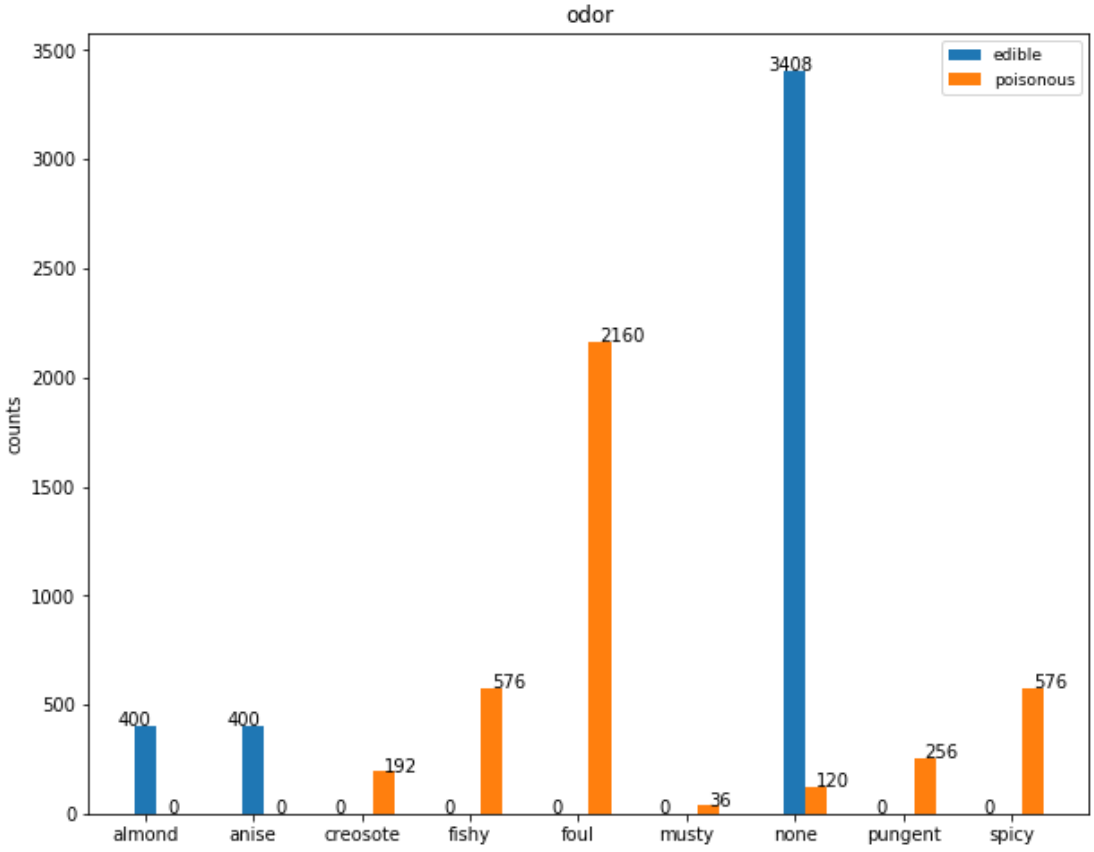
Mushroom :split data based on their labels and show the data distribution again:

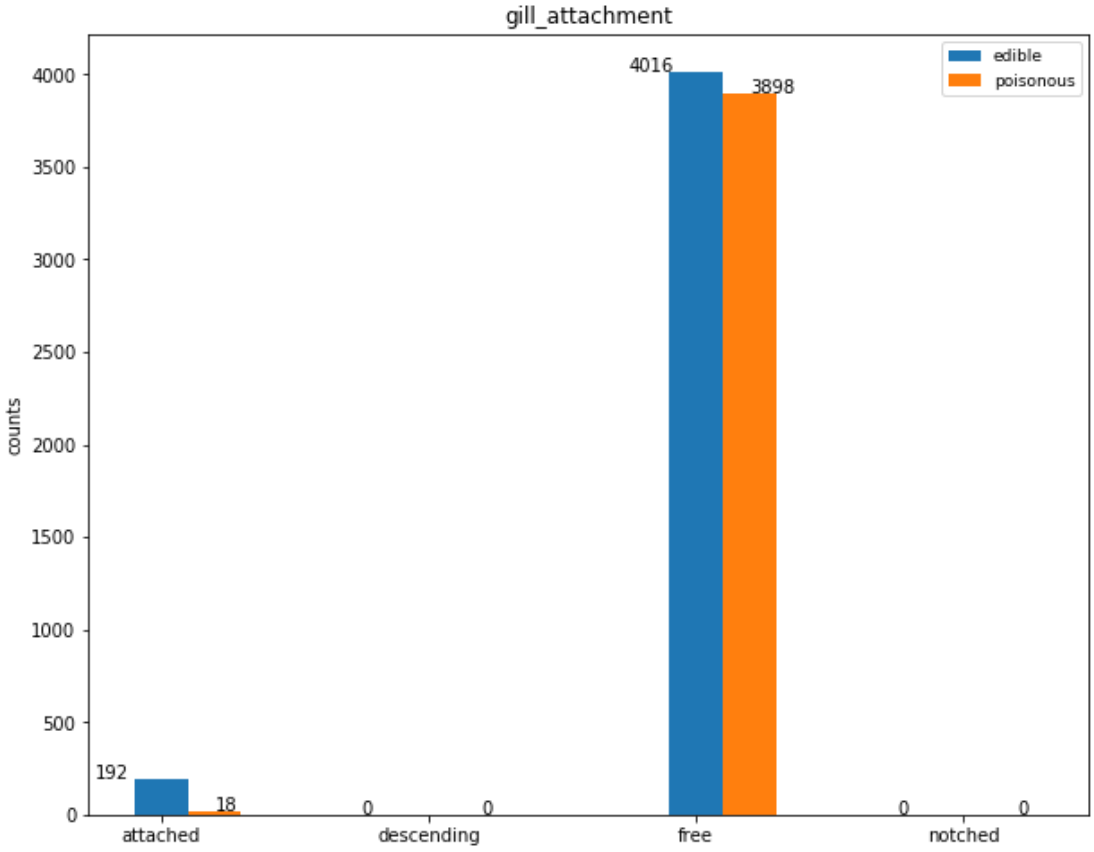


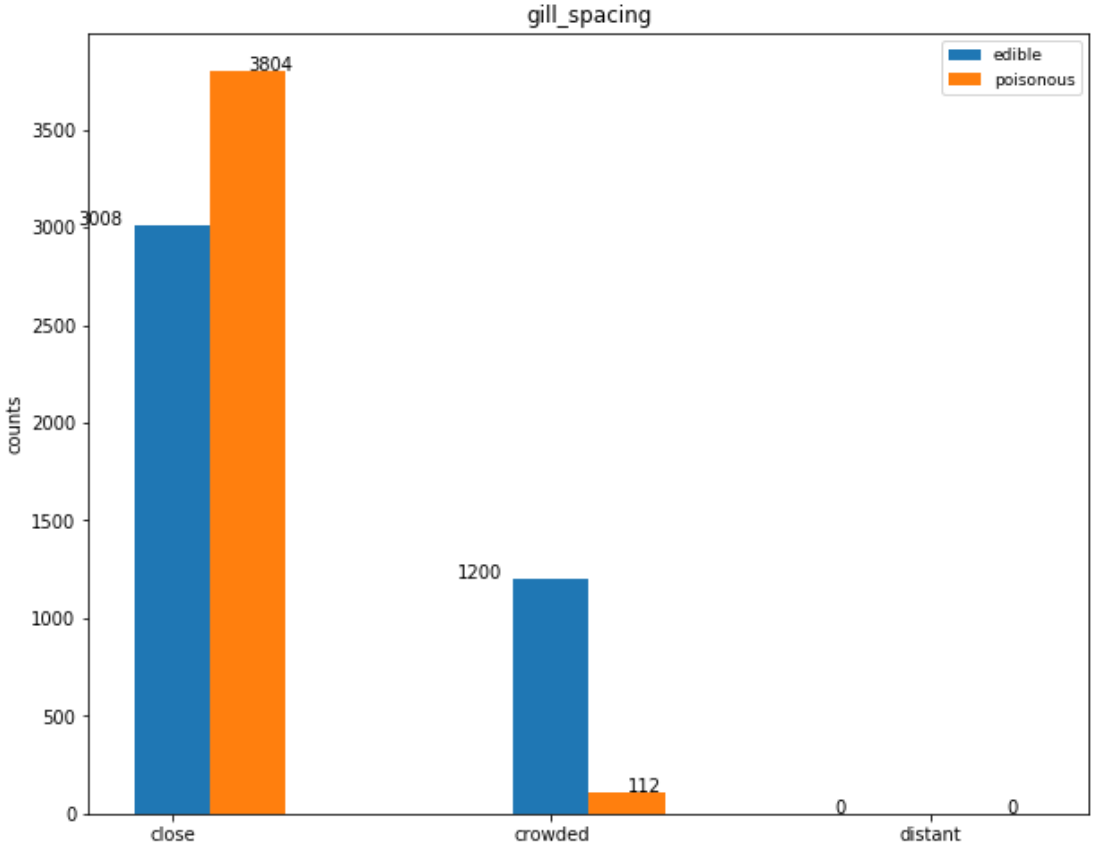

cap_surface
edible
poisonous
counts
1750
1500
1250
1000
750
500
250
0
1560
760
0
4
1504
1740
1144
1412
fibrous
grooves
scaly
smooth

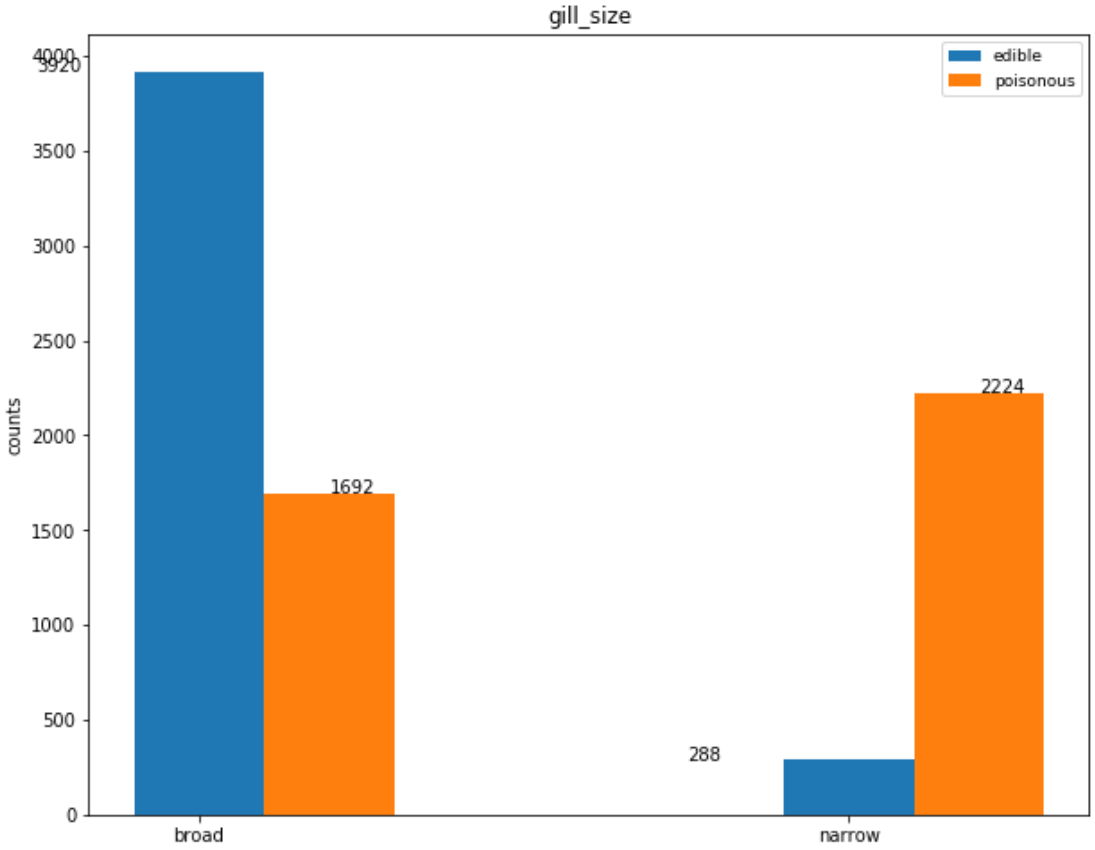

cap_color
edible
poisonous
counts
1200
1000
800
600
400
200
0
1264
1020
48
120
32
12
1032
808
16
0
56
88
16
0
624
876
720
320
400
672
brown
buff
cinnamon
gray
green
pink
purple
red
white
yellow

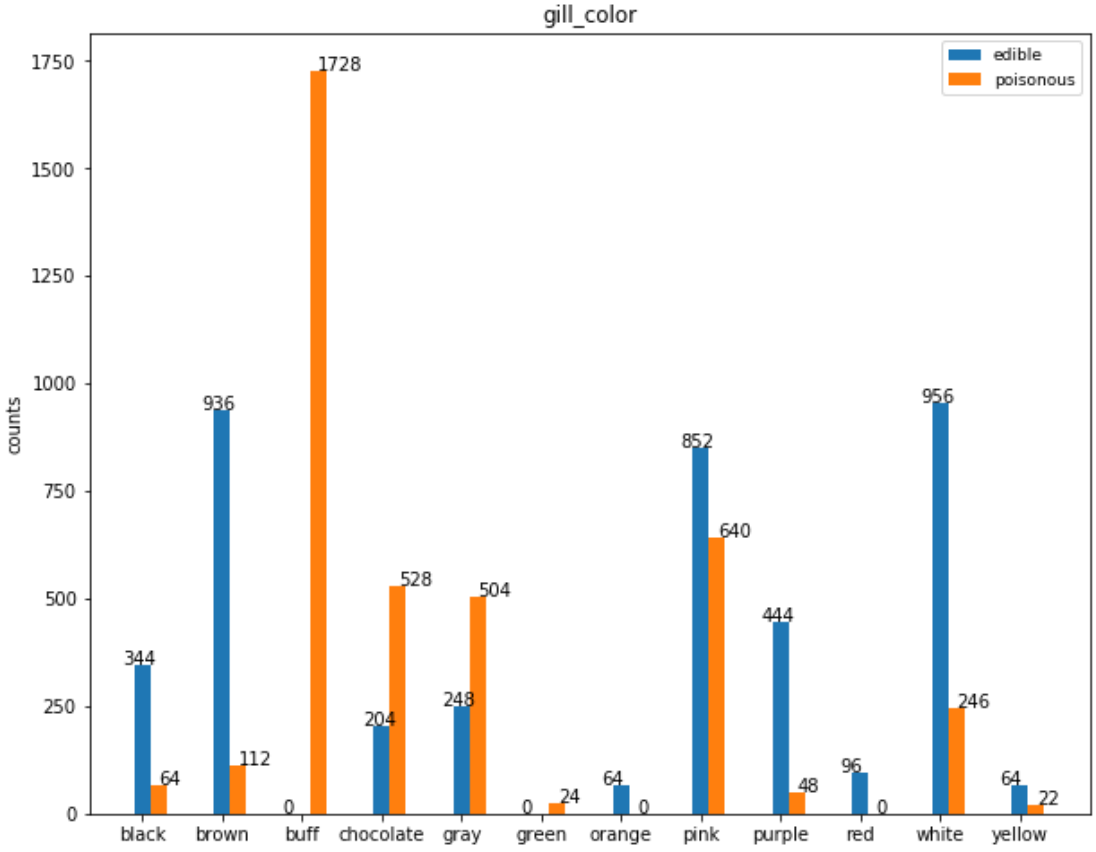

bruises
edible
poisonous
counts
3000
2752
2500
2000
1500
1000
500
0
624
1456
3292
bruises
no

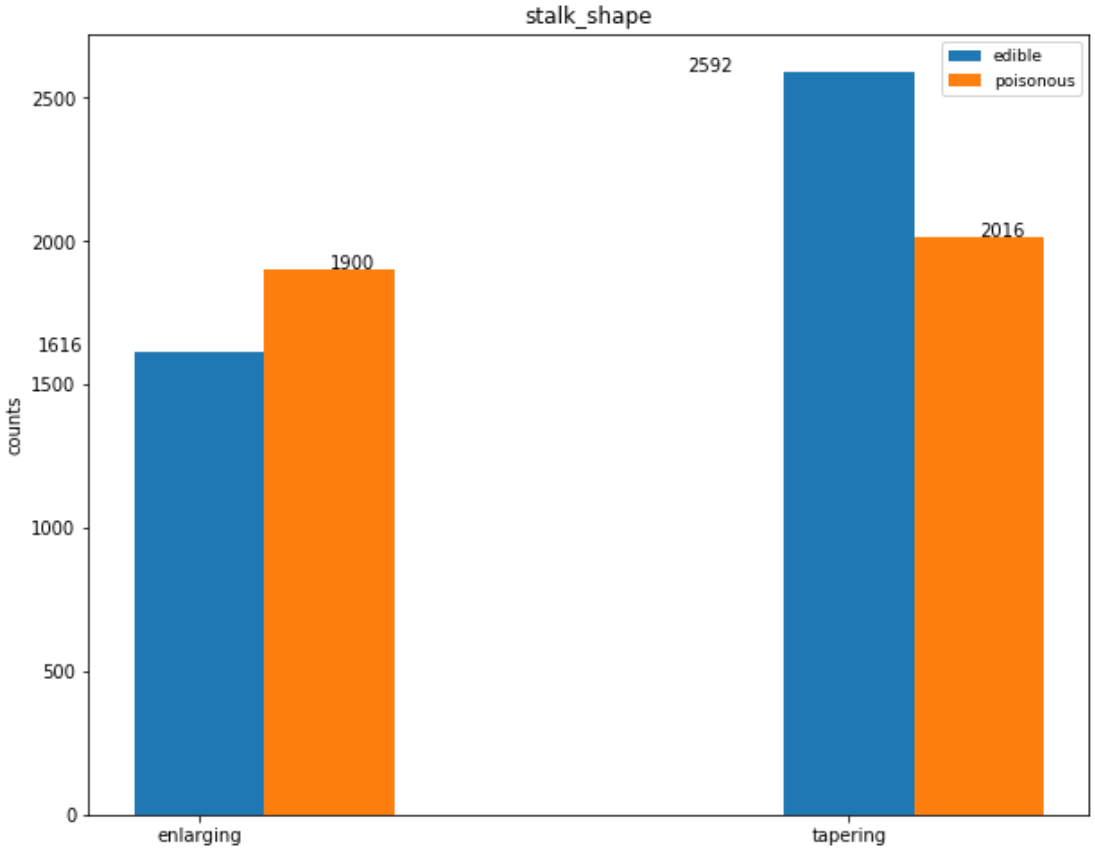

odor
edible
poisonous
counts
3500
3000
2500
2000
1500
1000
500
0
400
0
400
0
0
192
0
576
0
2160
0
36
3408
120
0
256
0
576
almond
anise
creosote
fishy
foul
musty
none
pungent
spicy

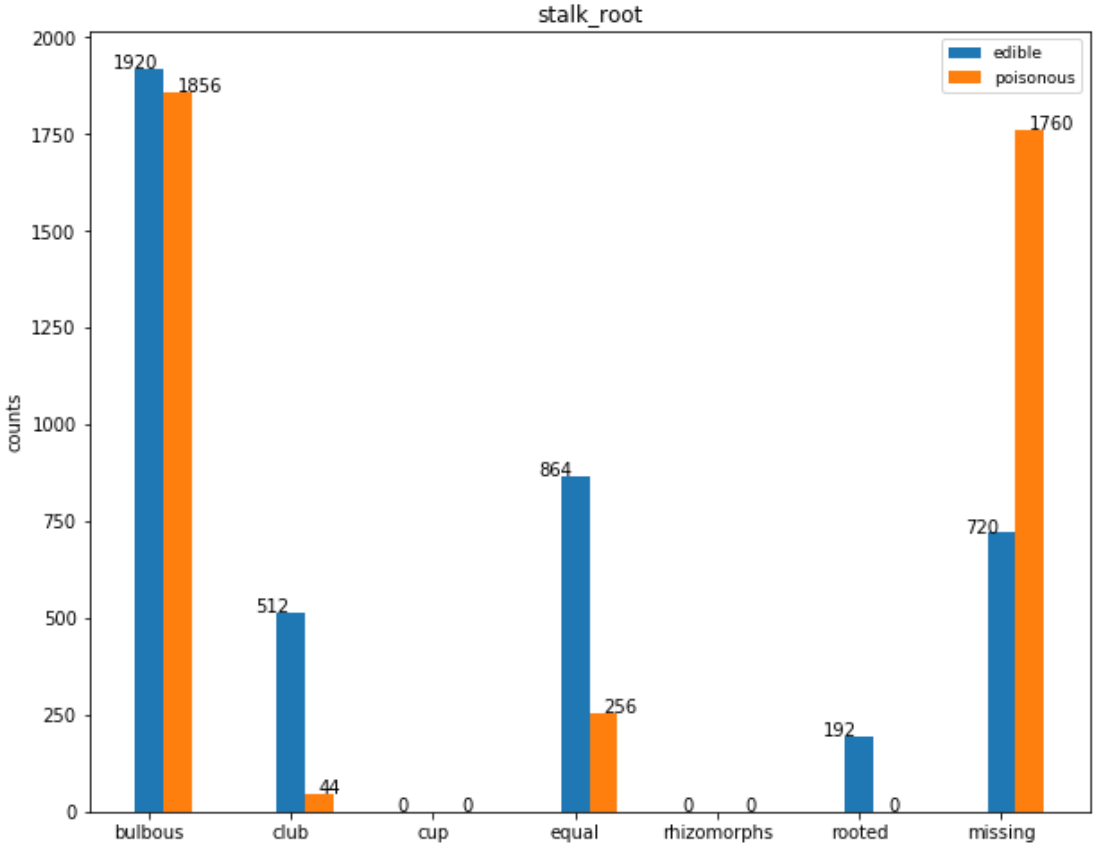

gill_attachment
edible
poisonous
counts
4000
3500
3000
2500
2000
1500
1000
500
0
192
18
0
0
4016
3898
0
0
attached
descending
free
notched

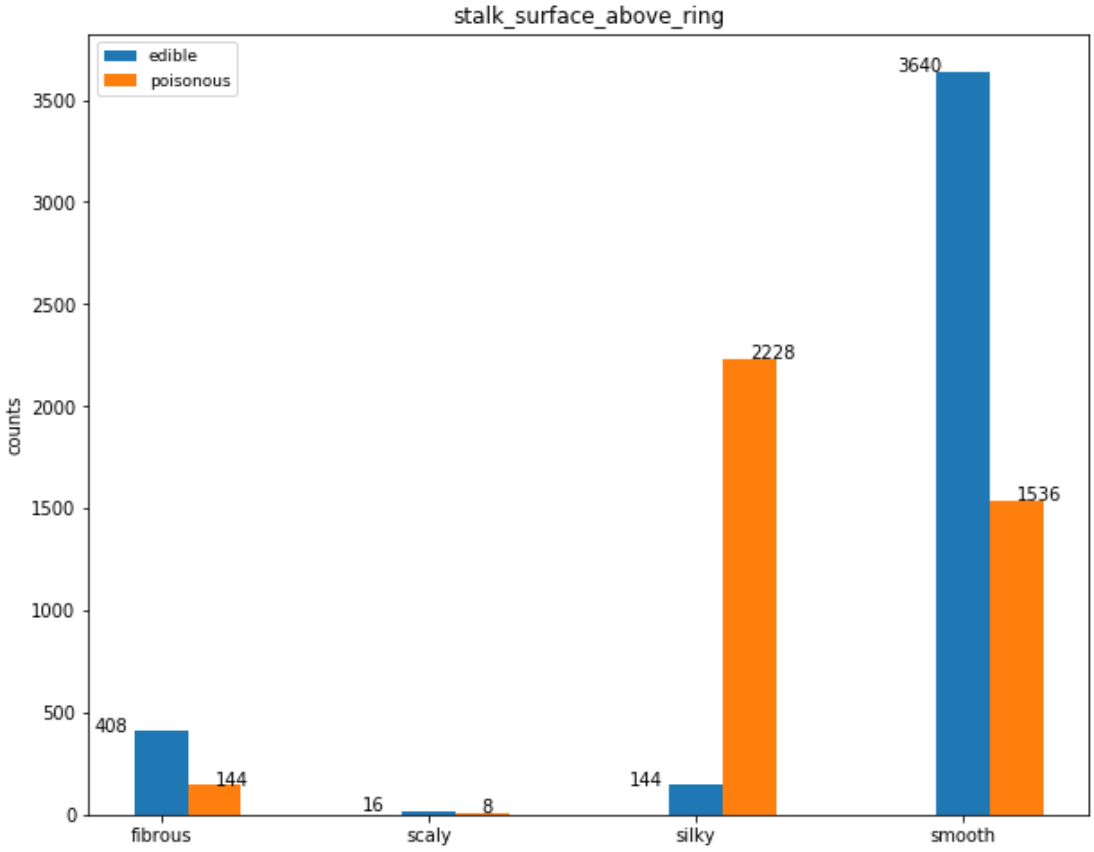

gill_spacing
edible
poisonous
counts
3500
3000
2500
2000
1500
1000
500
0
3008
3804
1200
112
0
0
close
crowded
distant

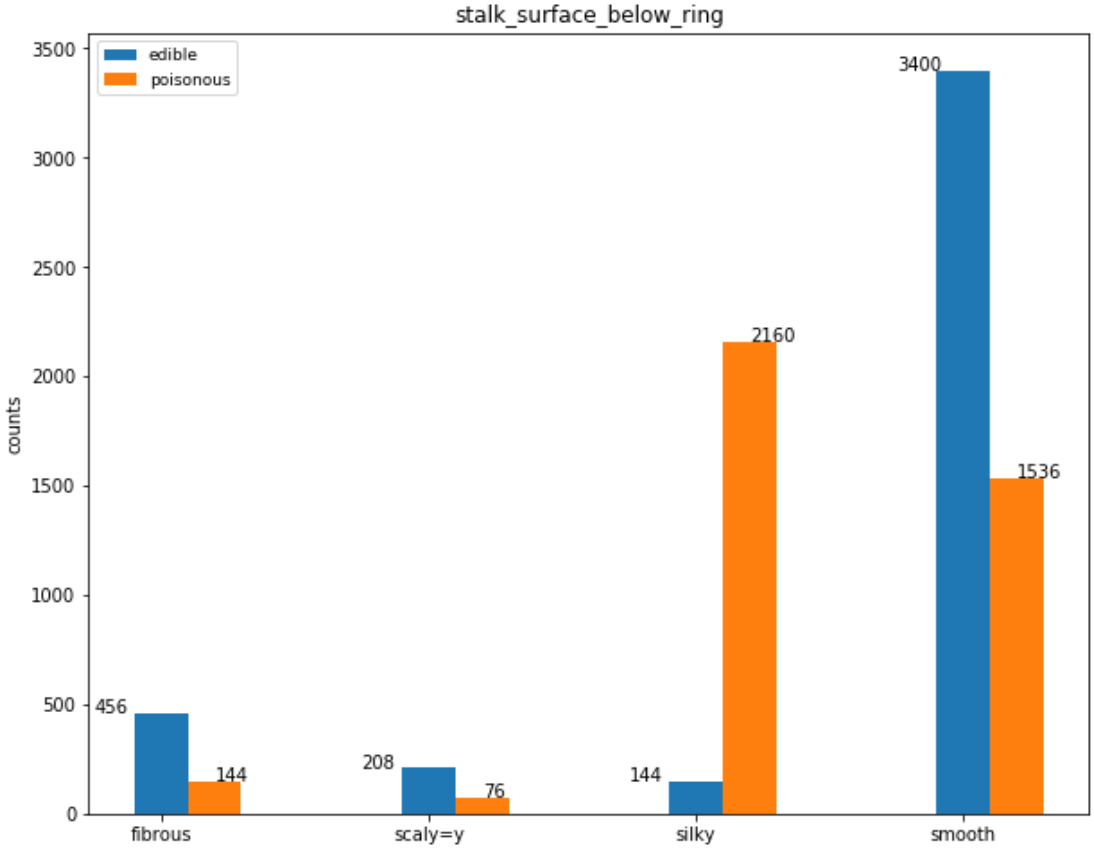

gill_size
edible
poisonous
counts
4000
3920
3500
3000
2500
2000
1500
1000
500
0
1692
288
2224
broad
narrow

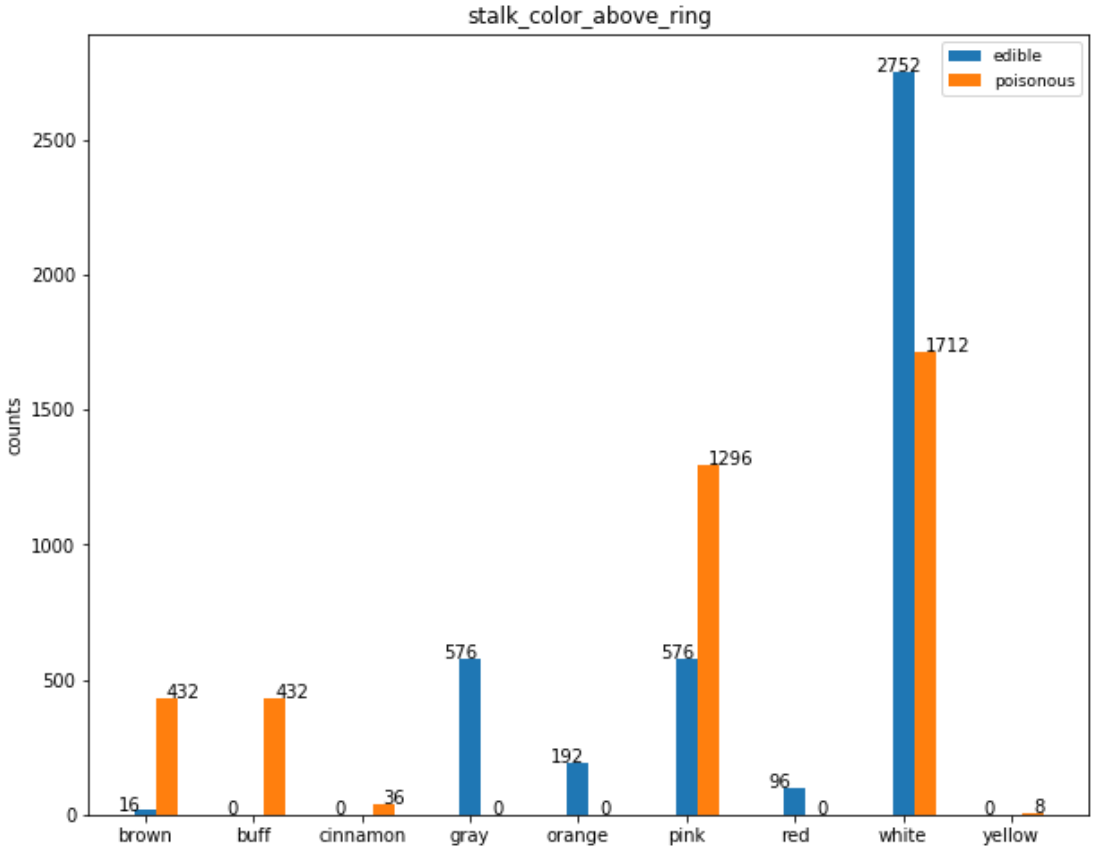

gill_color
edible
poisonous
counts
1750
1500
1250
1000
750
500
250
0
344
64
936
112
0
1728
204
528
248
504
0
24
64
0
852
640
444
48
96
0
956
246
64
22
black
brown
buff
chocolate
gray
green
orange
pink
purple
red
white
yellow

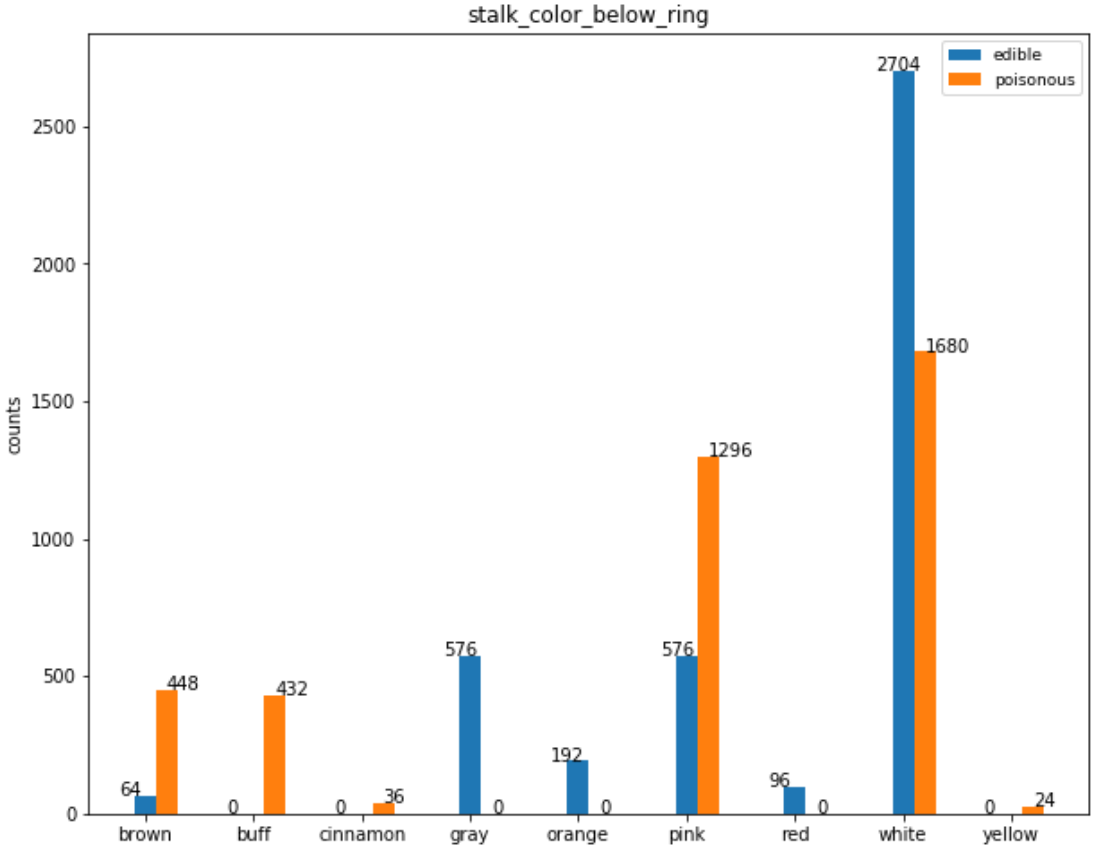

stalk_shape
edible
poisonous
counts
2500
2000
1616
1500
1000
500
0
2592
2016
1900
enlarging
tapering

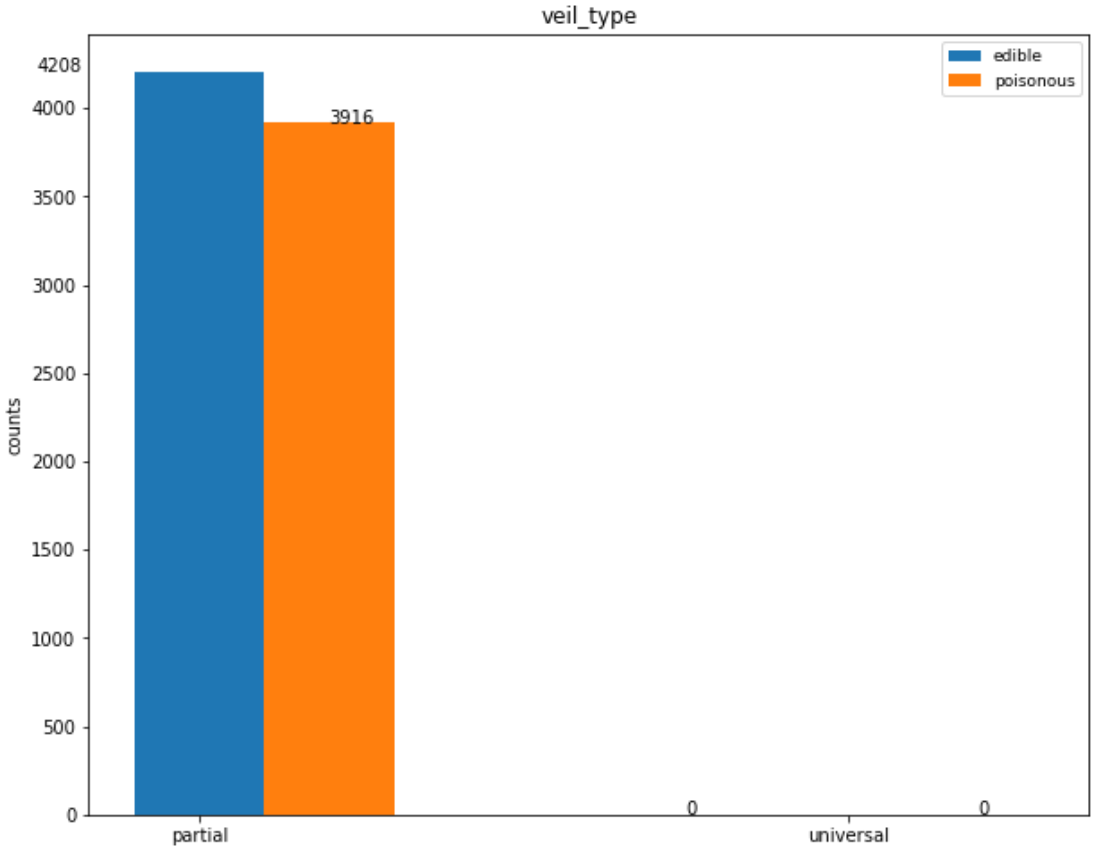

stalk_root
edible
poisonous
counts
2000
1750
1500
1250
1000
750
500
250
0
1920
1856
1760
864
720
512
256
192
44
0
0
0
0
0
0
bulbous
club
cup
equal
rhizomorphs
rooted
missing

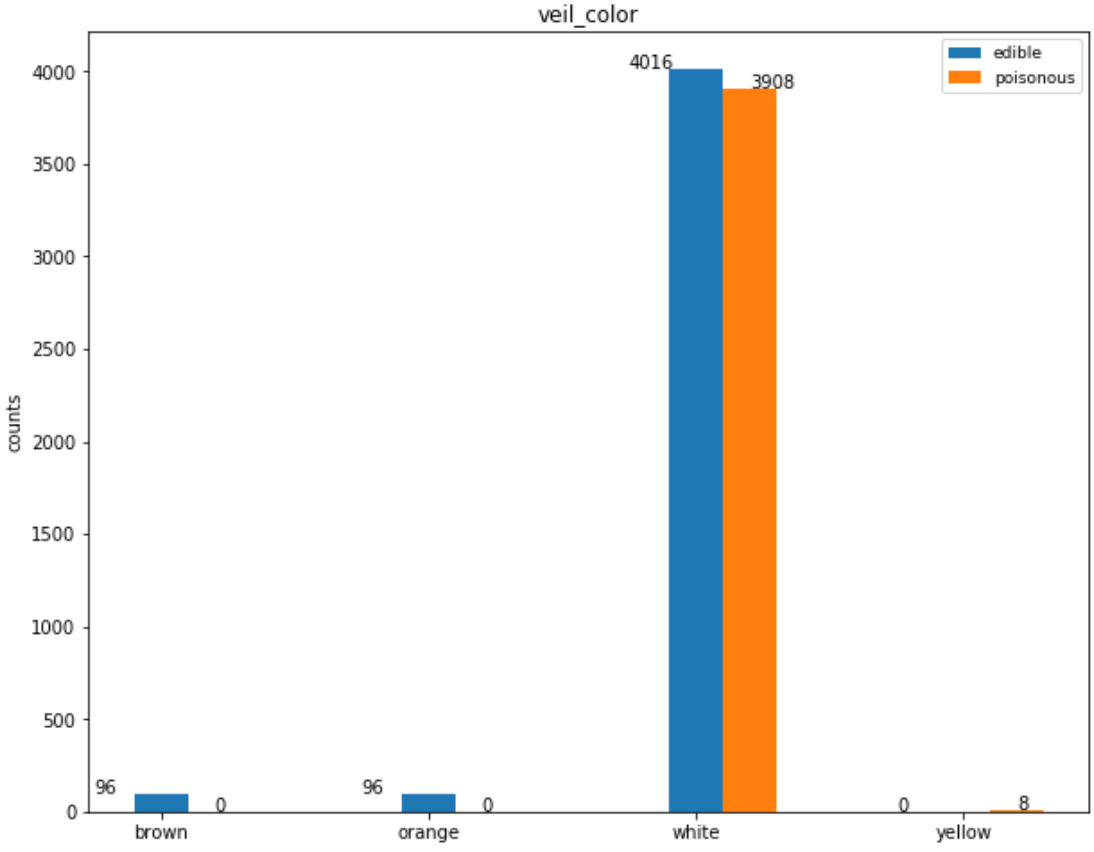

stalk_surface_above_ring
edible
poisonous
counts
0
500
1000
1500
2000
2500
3000
3500
408
144
16
8
144
2228
3640
1536
fibrous
scaly
silky
smooth

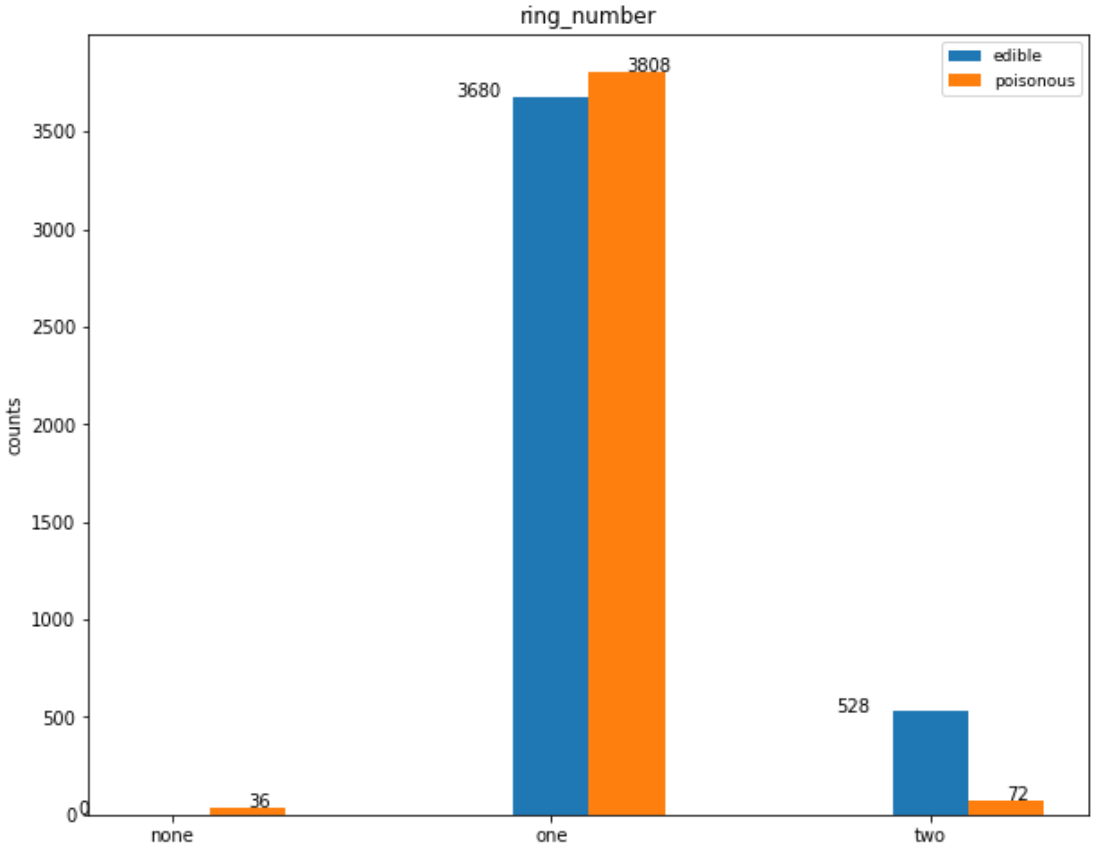

stalk_surface_below_ring
edible
poisonous
counts
0
500
1000
1500
2000
2500
3000
3500
456
144
208
76
144
2160
3400
1536
fibrous
scaly=y
silky
smooth

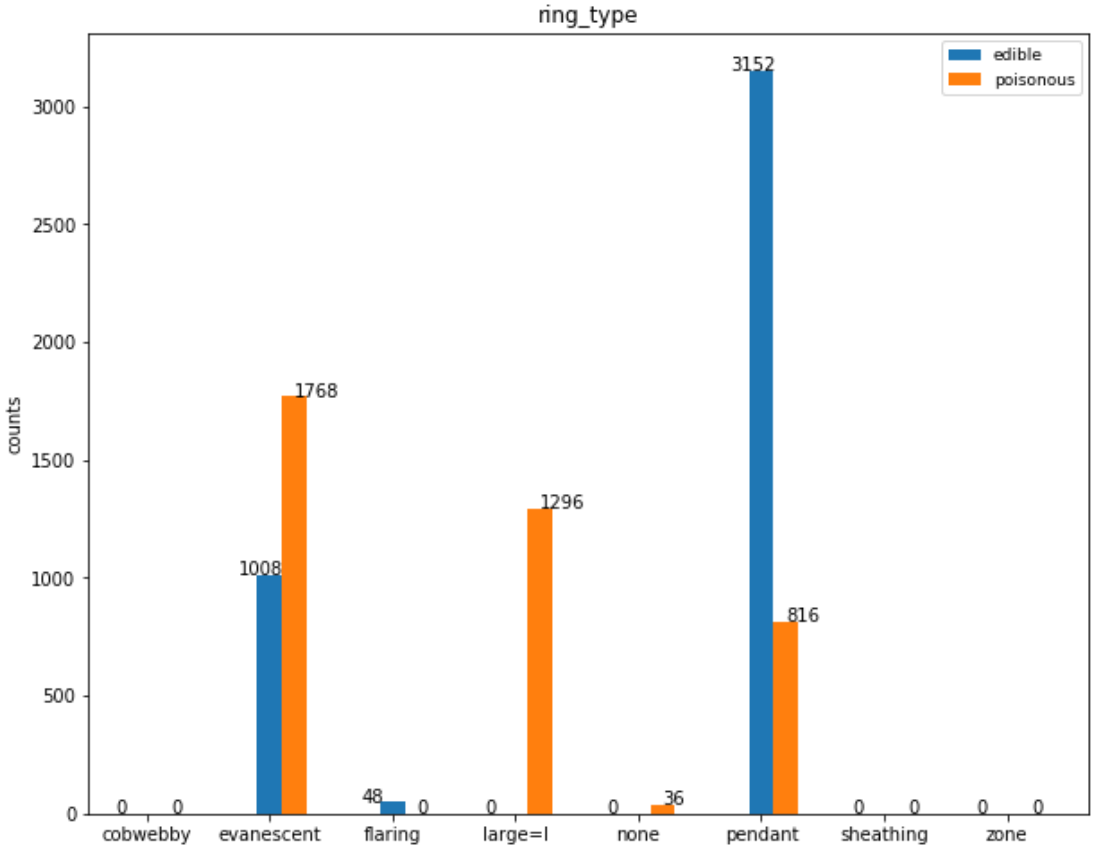

stalk_color_above_ring
edible
poisonous
counts
2500
2000
1500
1000
500
0
16
432
0
432
0
36
576
0
192
0
576
1296
96
0
2752
1712
0
8
brown
buff
cinnamon
gray
orange
pink
red
white
yellow

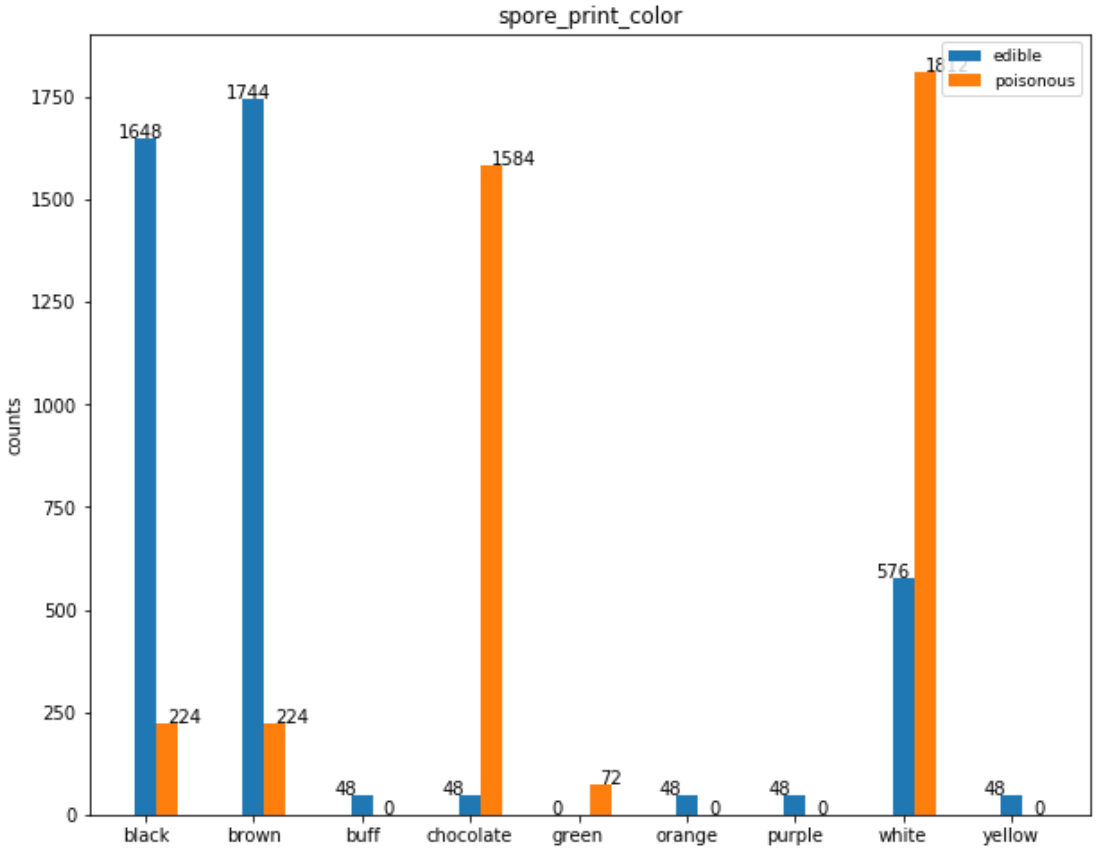

stalk_color_below_ring
edible
poisonous
counts
2500
2000
1500
1000
500
0
64
448
0
432
0
36
576
0
192
0
576
1296
96
0
2704
1680
0
24
brown
buff
cinnamon
gray
orange
pink
red
white
yellow

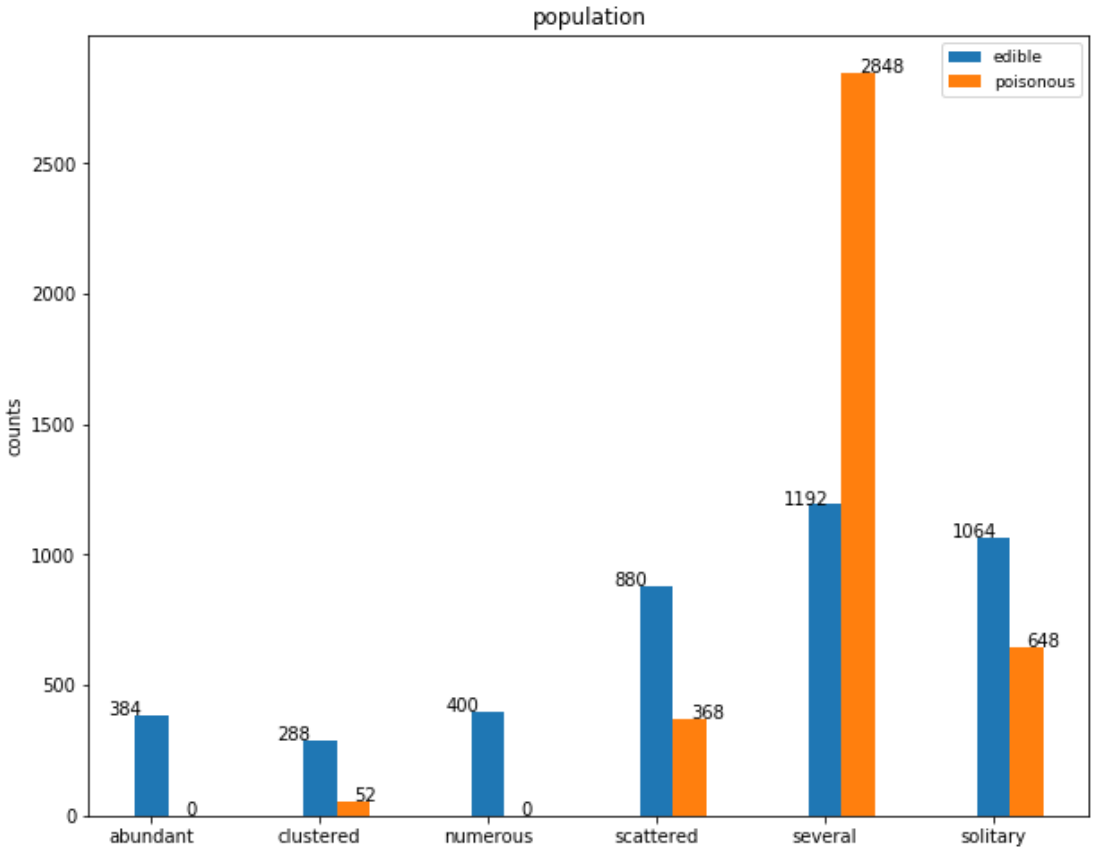

veil_type
edible
poisonous
counts
4208
4000
3500
3000
2500
2000
1500
1000
500
0
3916
0
0
partial
universal

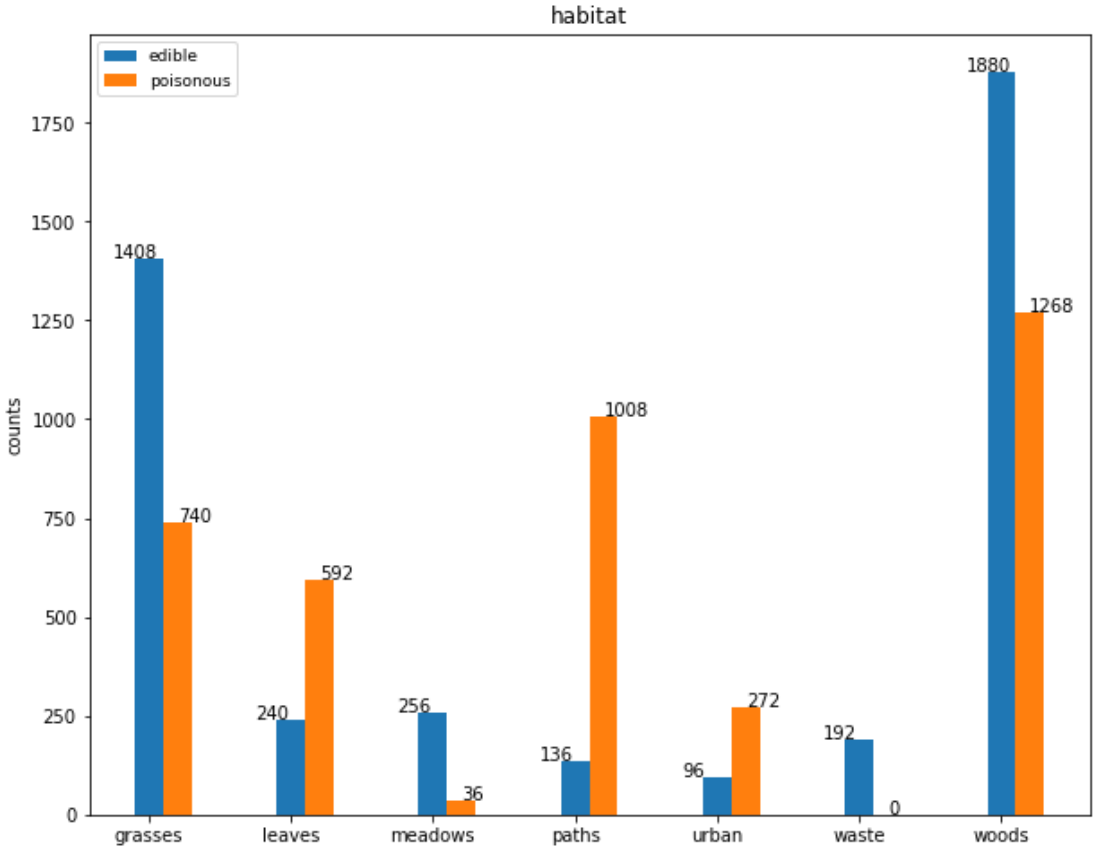

veil_color
edible
poisonous
counts
4000
3500
3000
2500
2000
1500
1000
500
0
96
0
96
0
4016
3908
0
8
brown
orange
white
yellow


ring_number
edible
poisonous
counts
3500
3000
2500
2000
1500
1000
500
0
3680
3808
528
36
0
72
none
one
two


ring_type
edible
poisonous
counts
3000
2500
2000
1500
1000
500
0
3152
1768
1296
1008
816
48
36
0
0
0
0
0
0
0
0
0
0
cobwebby
evanescent
flaring
large=l
none
pendant
sheathing
zone

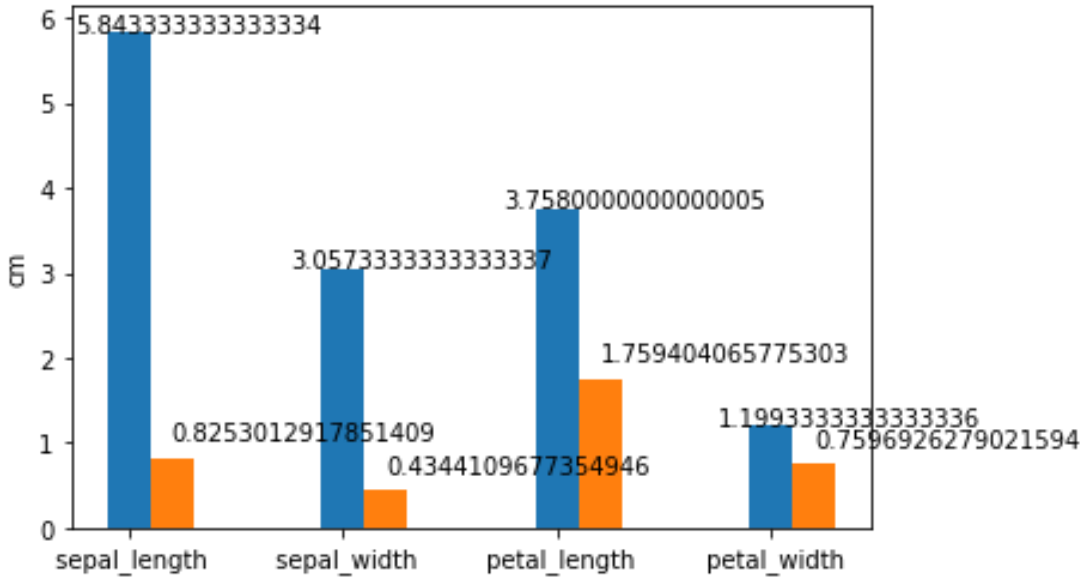

spore_print_color
edible
poisonous
counts
0
250
500
750
1000
1250
1500
1750
1648
224
1744
224
48
0
48
1584
0
72
48
0
48
0
576
1812
48
0
black
brown
buff
chocolate
green
orange
purple
white
yellow


population
edible
poisonous
counts
0
500
1000
1500
2000
2500
384
0
288
52
400
0
880
368
1192
2848
1064
648
abundant
clustered
numerous
scattered
several
solitary

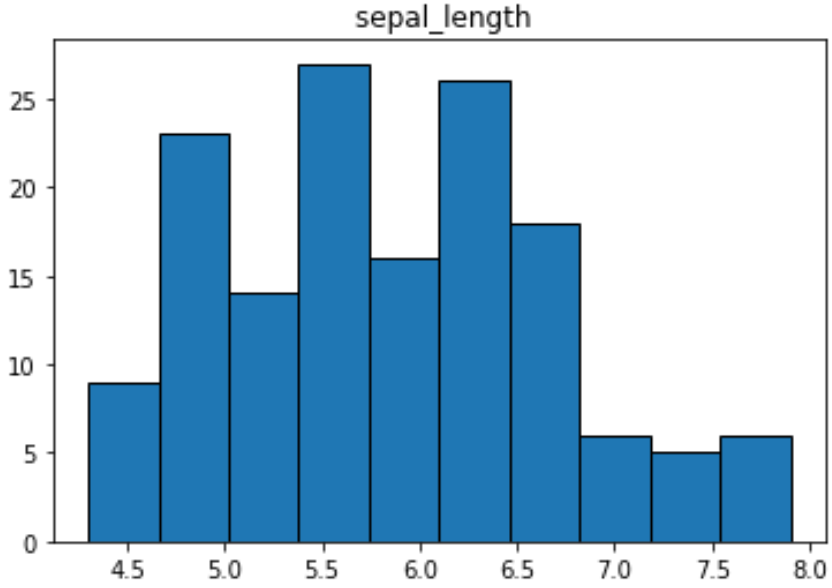


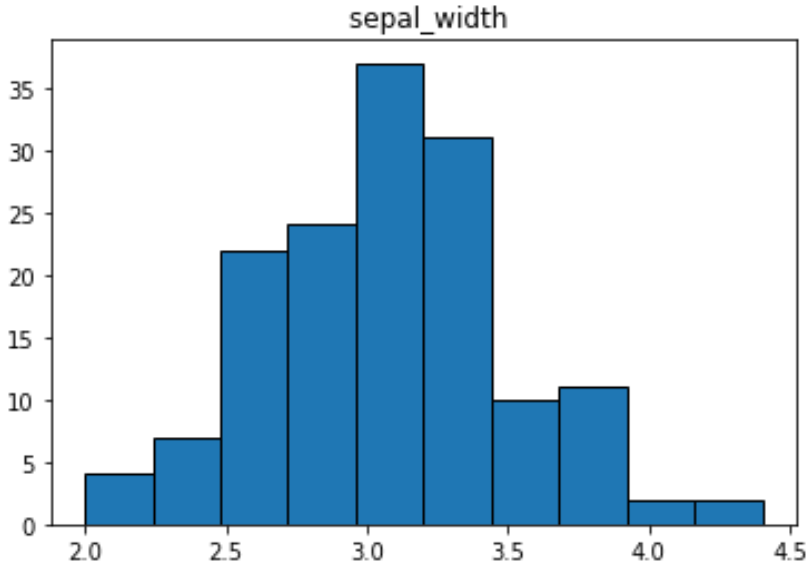
Iris dataset:

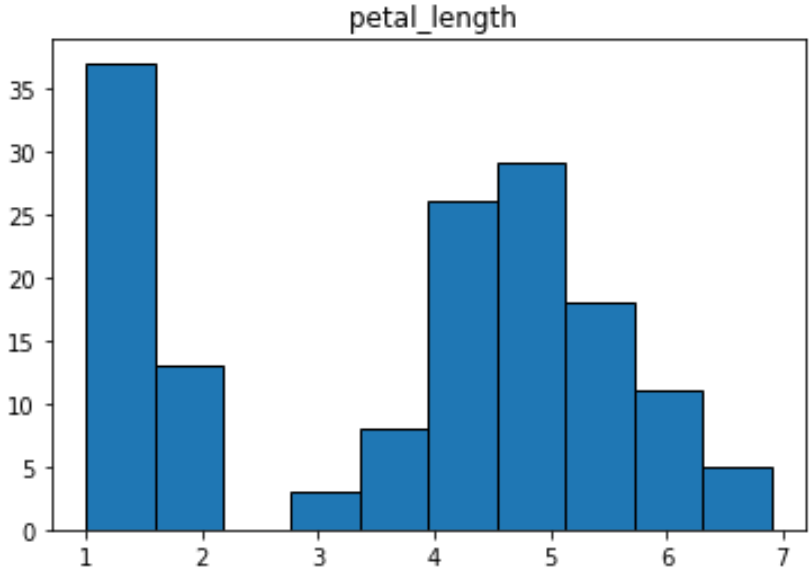
Blue : average orange: standard deviation

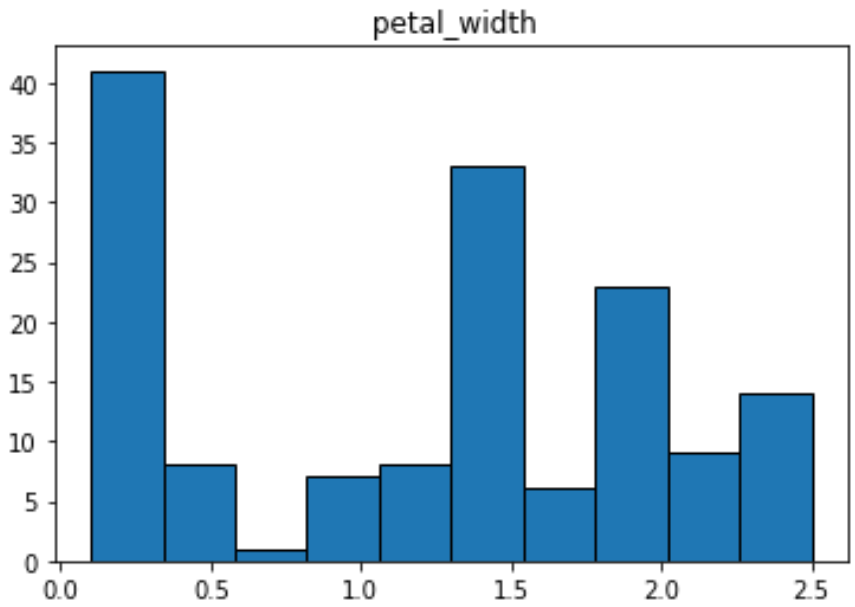


Value frequency of each feature:


sepal_length
25
20
15
10
5
0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0


sepal_width
35
30
25
20
15
10
5
0
2.0
2.5
3.0
3.5
4.0
4.5


petal_length
35
30
25
20
15
10
5
0
1
2
3
4
5
6
7

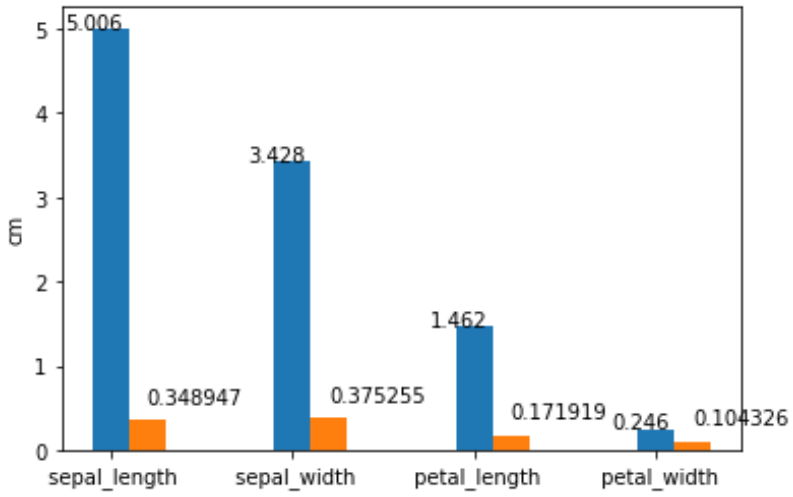


Iris dataset with split data based on their labels:

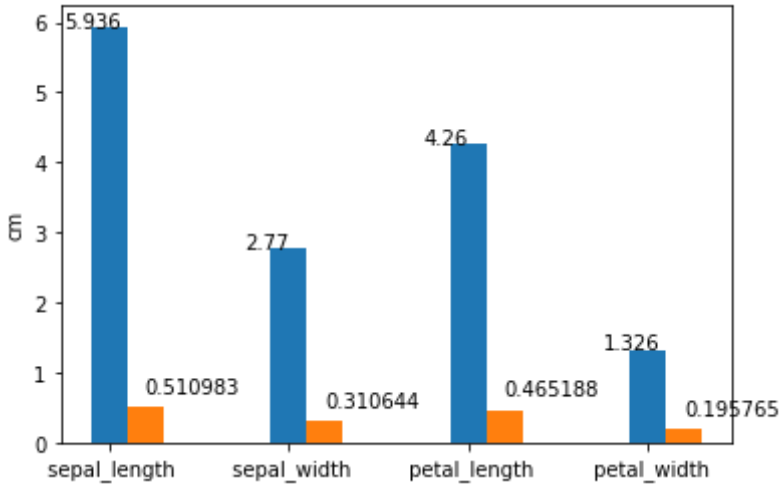
average, standard deviation:

Blue : average orange: standard deviation

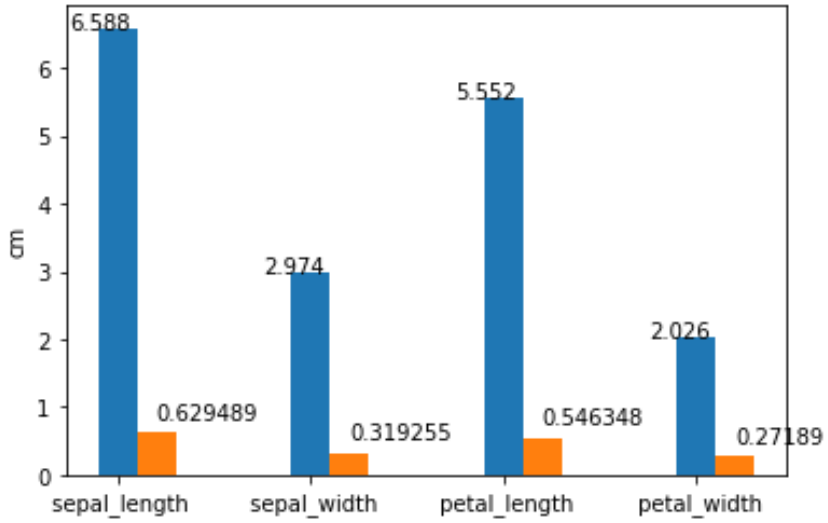
Iris-setosa:



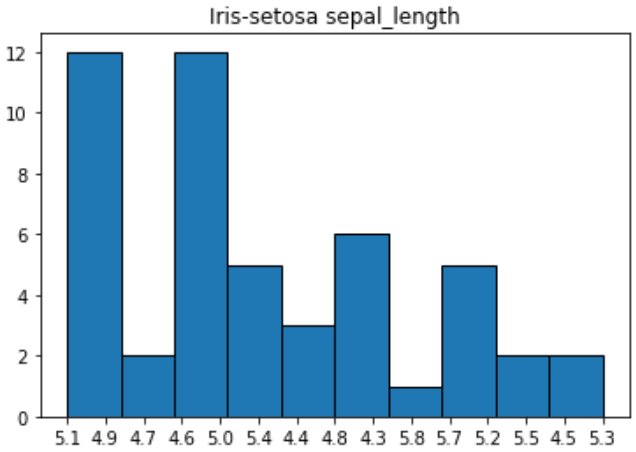
Iris-versicolor:

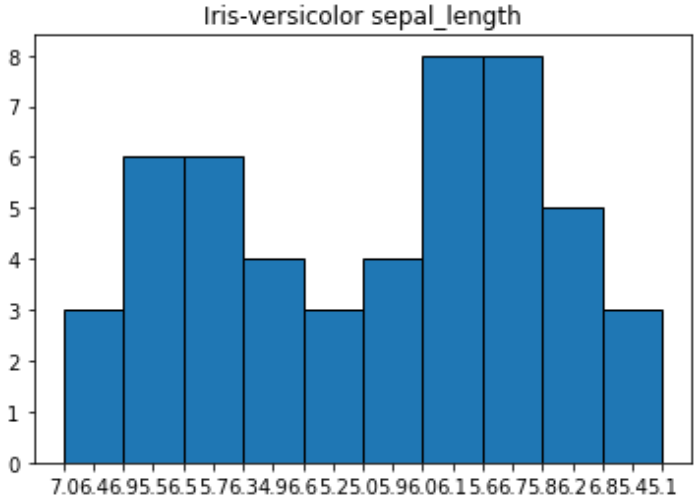


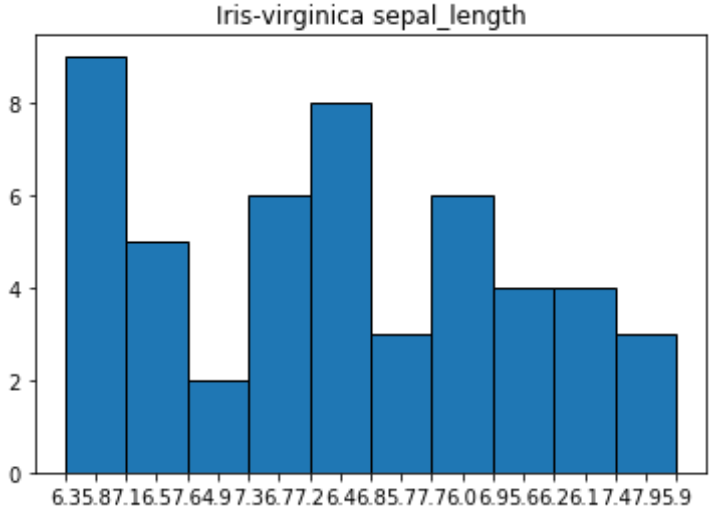
Iris-virginica

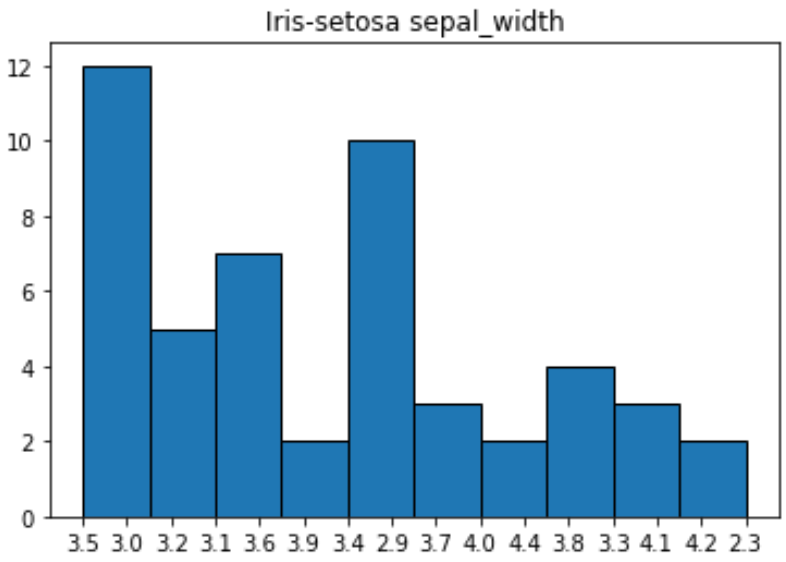


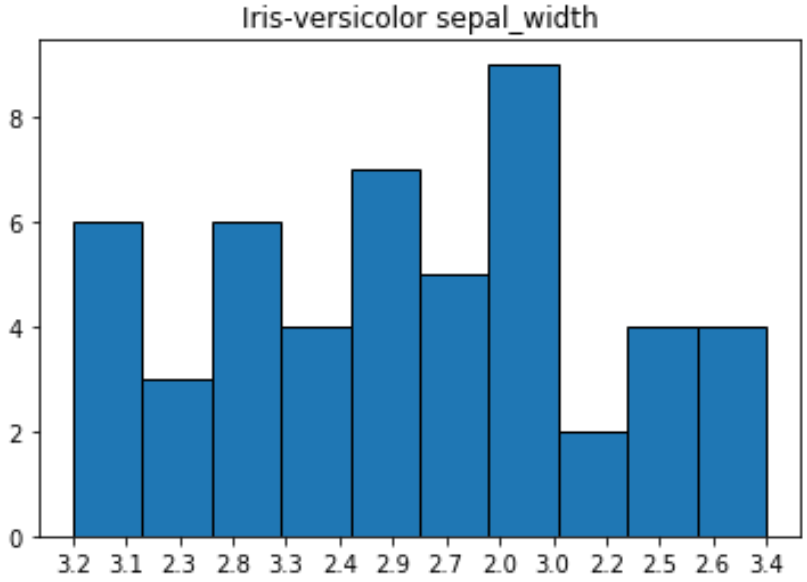
Each label’s frequency:

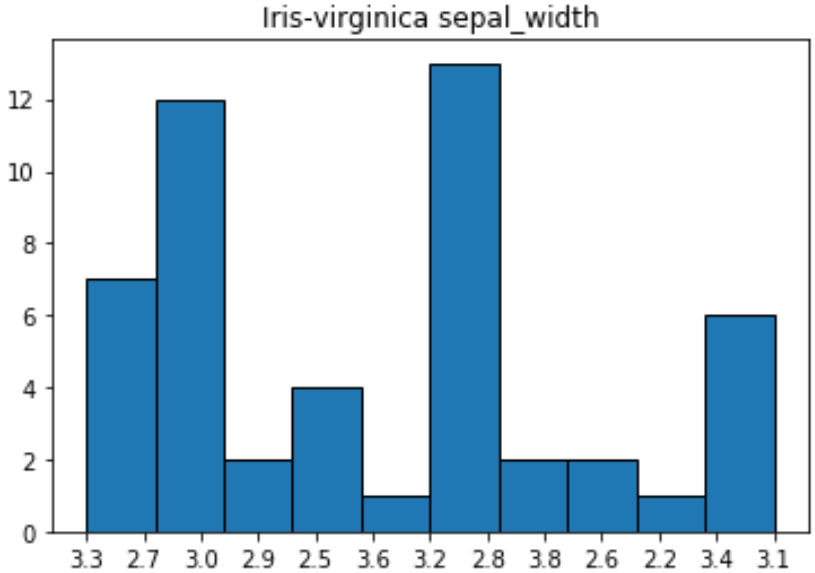

Iris-setosa sepal_length
12
10
8
6
4
2
0
5.1 4.9 4.7 4.6 5.0 5.4 4.4 4.8 4.3 5.8 5.7 5.2 5.5 4.5 5.3

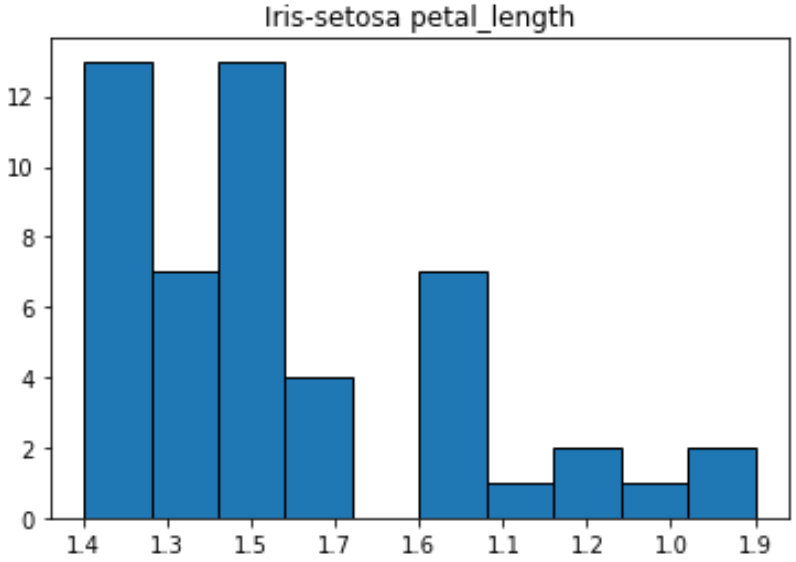

Iris-versicolor sepal_length
8
7
6
5
4
3
2
1
0
7.0 6.4 6.9 5.5 6.5 5.7 6.3 4.9 6.6 5.2 5.0 5.9 6.0 6.1 5.6 6.7 5.8 6.2 6.8 5.4 5.1

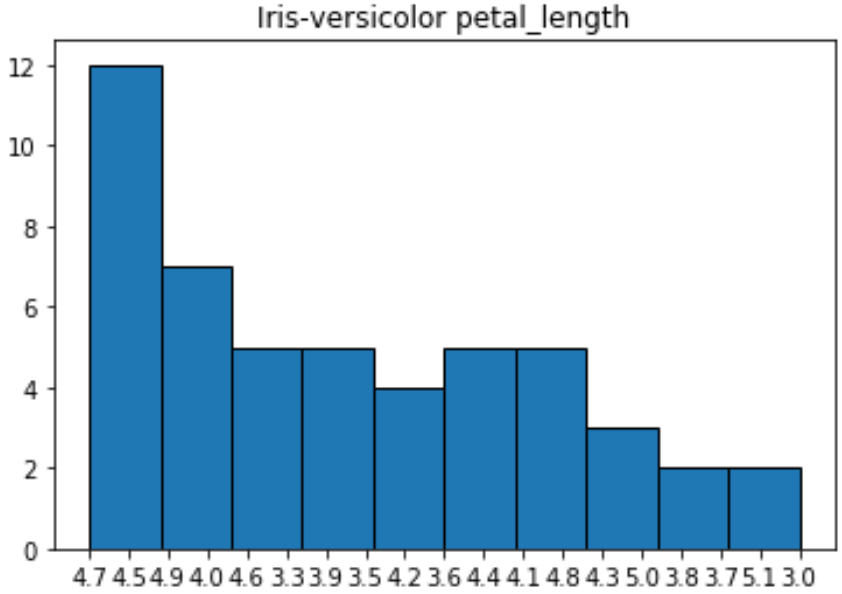

Iris-virginica sepal_length
8
6
4
2
0
6.3 5.8 7.1 6.5 7.6 4.9 7.3 6.7 7.2 6.4 6.8 5.7 7.7 6.0 6.9 5.6 6.2 6.1 7.4 7.9 5.9

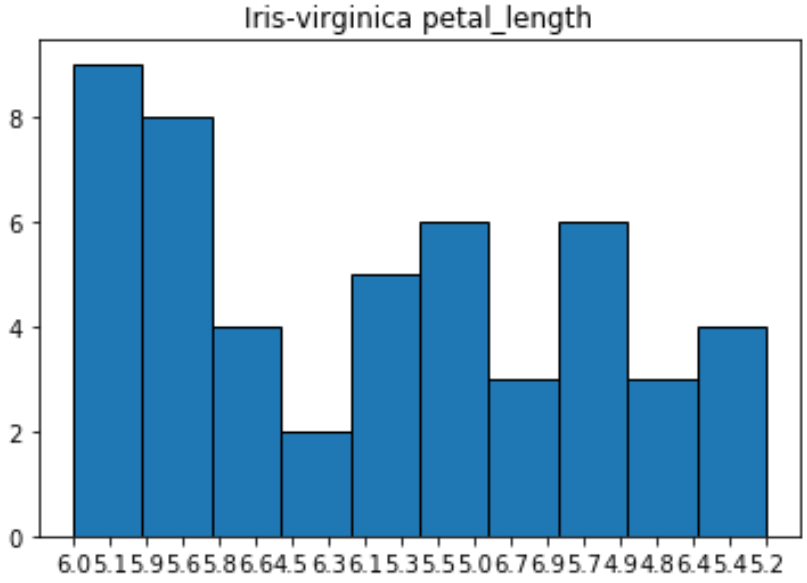

Iris-setosa sepal_width
12
10
8
6
4
2
0
3.5 3.0 3.2 3.1 3.6 3.9 3.4 2.9 3.7 4.0 4.4 3.8 3.3 4.1 4.2 2.3

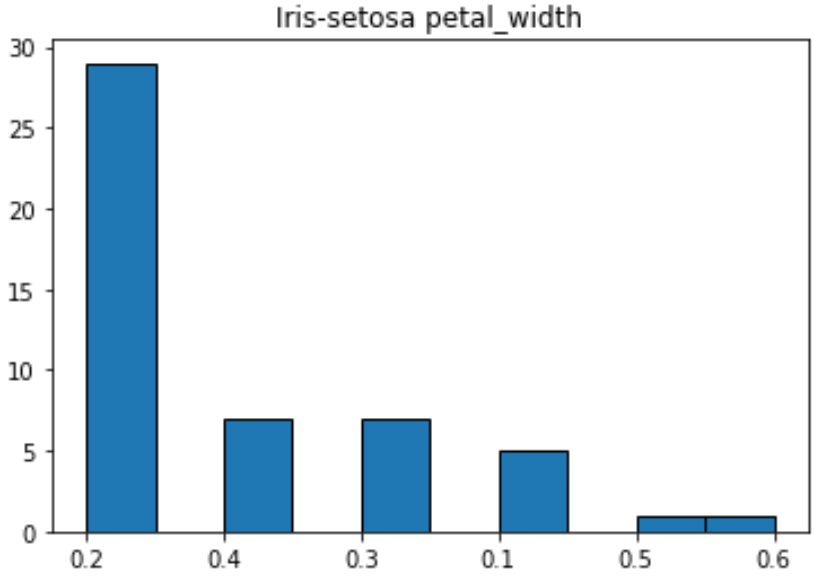

Iris-versicolor sepal_width
8
6
4
2
0
3.2 3.1 2.3 2.8 3.3 2.4 2.9 2.7 2.0 3.0 2.2 2.5 2.6 3.4

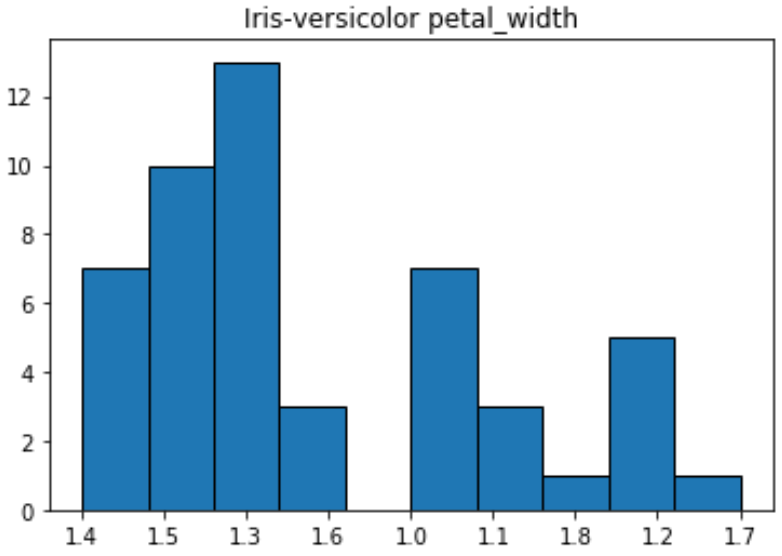

Iris-virginica sepal_width
12
10
8
6
4
2
0
3.3 2.7 3.0 2.9 2.5 3.6 3.2 2.8 3.8 2.6 2.2 3.4 3.1

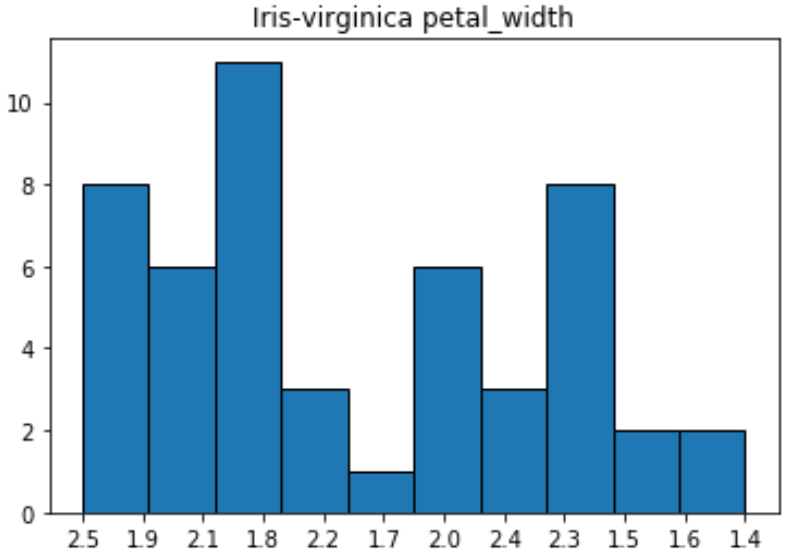

Iris-setosa petal_length
0
2
4
6
8
10
12
1.4
1.3
1.5
1.7
1.6
1.1
1.2
1.0
1.9


Iris-versicolor petal_length
0
2
4
6
8
10
12
4.7
4.5
4.9
4.0
4.6
3.3
3.9
3.5
4.2
3.6
4.4
4.1
4.8
4.3
5.0
3.8
3.7
5.1
3.0


Iris-virginica petal_length
0
2
4
6
8
6.0
5.1
5.9
5.6
5.8
6.6
4.5
6.3
6.1
5.3
5.5
5.0
6.7
6.9
5.7
4.9
4.8
6.4
5.4
5.2







3. Data preprocessing:

Drop feature:

don’t use stalk_root to bulid classfier 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.ipynb,hw1_4.ipynb.

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。