In [1]:

```
import pandas as pd
```

Attribute Information:

Classes

19 Classes:

- diaporthe-stem-canker,
- charcoal-rot,
- rhizoctonia-root-rot,
- phytophthora-rot,
- brown-stem-rot,
- powdery-mildew,
- downy-mildew,
- brown-spot,
- bacterial-blight,
- bacterial-pustule,
- purple-seed-stain,
- anthracnose,
- phyllosticta-leaf-spot,
- alternarialeaf-spot,
- frog-eye-leaf-spot,
- diaporthe-pod-&-stem-blight,
- cyst-nematode,
- 2-4-d-injury,
- herbicide-injury.

Features

1. date: april,may,june,july,august,september,october,?.
2. plant-stand: normal,lt-normal,?.
3. precip: lt-norm,norm,gt-norm,?.
4. temp: lt-norm,norm,gt-norm,?.
5. hail: yes,no,?.
6. crop-hist: diff-lst-year,same-lst-yr,same-lst-two-yrs,same-lst-sev-yrs,?.
7. area-damaged: scattered,low-areas,upper-areas,whole-field,?.
8. severity: minor,pot-severe,severe,?.
9. seed-tmt: none,fungicide,other,?.
10. germination: 90-100%,80-89%,lt-80%,?.
11. plant-growth: norm,abnorm,?.
12. leaves: norm,abnorm.
13. leafspots-halo: absent,yellow-halos,no-yellow-halos,?.
14. leafspots-marg: w-s-marg,no-w-s-marg,dna,?.
15. leafspot-size: lt-1/8,gt-1/8,dna,?.
16. leaf-shread: absent,present,?.
17. leaf-malf: absent,present,?.
18. leaf-mild: absent,upper-surf,lower-surf,?.
19. stem: norm,abnorm,?.
20. lodging: yes,no,?.
21. stem-cankers: absent,below-soil,above-soil,above-sec-nde,?.
22. canker-lesion: dna,brown,dk-brown-blk,tan,?.
23. fruiting-bodies: absent,present,?.
24. external decay: absent,firm-and-dry,watery,?.
25. mycelium: absent,present,?.
26. int-discolor: none,brown,black,?.
27. sclerotia: absent,present,?.
28. fruit-pods: norm,diseased,few-present,dna,?.
29. fruit spots: absent,colored,brown-w/blk-specks,distort,dna,?.
30. seed: norm,abnorm,?.
31. mold-growth: absent,present,?.
32. seed-discolor: absent,present,?.
33. seed-size: norm,lt-norm,?.
34. shriveling: absent,present,?.

35. roots: norm,rotted,galls-cysts,?.

In [2]:

```
columns = ['class', 'date', 'plant-stand', 'precip', 'temp', 'hail', 'crop-hist', \
            'area-damaged', 'severity', 'seed-tmt', 'germination', 'plant-growth', \
           'leaves', 'leafspots-halo', 'leafspots-marg', 'leafspot-size', 'leaf-shread', \
            'leaf-malf', 'leaf-mild', 'stem', 'lodging', 'stem-cankers', 'canker-lesion',\
           'fruiting-bodies', 'external decay', 'mycelium', 'int-discolor', 'sclerotia',\
           'fruit-pods', 'fruit spots', 'seed', 'mold-growth', 'seed-discolor', 'seed-size',\
           'shriveling', 'roots']
df = pd.read_csv('soybean-large.data', names=columns)
len(columns)
```

Out[2]:

```
36
```

In [3]:

```
df.head()
```

Out[3]:

	class	date	plant-stand	precip	temp	hail	crop-hist	area-damaged	severity	seed-tmt	...	int-discolor	sclerotia	fruit-pods	fruit spots	seed	mold-growth	seed-discolor	se s
0	diaporthe-stem-canker	6	0	2	1	0	1	1	1	0	...	0	0	0	4	0	0	0	
1	diaporthe-stem-canker	4	0	2	1	0	2	0	2	1	...	0	0	0	4	0	0	0	
2	diaporthe-stem-canker	3	0	2	1	0	1	0	2	1	...	0	0	0	4	0	0	0	
3	diaporthe-stem-canker	3	0	2	1	0	1	0	2	0	...	0	0	0	4	0	0	0	
4	diaporthe-stem-canker	6	0	2	1	0	2	0	1	0	...	0	0	0	4	0	0	0	

5 rows × 36 columns

Notice how the values are all integers or "?". The integers relate to the values gives in the attribute descriptions. So, for example, date value of 0 corresponds to April, roots value of 1 corresponds to "rotted".

If you find it useful, you can create a mapping from the integer to the string.

In [4]:

```
def create_dict(vals):
    tmp_dict = {k: v for k, v in enumerate(vals.strip().split(',')) if v != '?'}
    tmp_dict['?'] = None
    return tmp_dict

maps = {
    'date': create_dict('april,may,june,july,august,september,october,?'),
    'plant-stand': create_dict('normal,lt-normal,?'),
    'precip': create_dict('lt-norm,norm,gt-norm,?'),
    'temp': create_dict('lt-norm,norm,gt-norm,?'),
    'hail': create_dict('yes,no,?'),
    'crop-hist': create_dict('diff-lst-year,same-lst-yr,same-lst-two-yrs,same-lst-sev-yrs,?'),
    'area-damaged': create_dict('scattered,low-areas,upper-areas,whole-field,?'),
    'severity': create_dict('minor,pot-severe,severe,?'),
    'seed-tmt': create_dict('none,fungicide,other,?'),
    'germination': create_dict('90-100%,80-89%,lt-80%,?'),
    'plant-growth': create_dict('norm,abnorm,?'),
    'leaves': create_dict('norm,abnorm'),
    'leafspots-halo': create_dict('absent,yellow-halos,no-yellow-halos,?'),
    'leafspots-marg': create_dict('w-s-marg,no-w-s-marg,dna,?'),
    'leafspot-size': create_dict('lt-1/8,gt-1/8,dna,?'),
    'leaf-shread': create_dict('absent,present,?'),
    'leaf-malf': create_dict('absent,present,?'),
    'leaf-mild': create_dict('absent,upper-surf,lower-surf,?'),
    'stem': create_dict('norm,abnorm,?'),
    'lodging': create_dict('yes,no,?'),
    'stem-cankers': create_dict('absent,below-soil,above-soil,above-sec-nde,?'),
    'canker-lesion': create_dict('dna,brown,dk-brown-blk,tan,?'),
    'fruiting-bodies': create_dict('absent,present,?'),
    'external decay': create_dict('absent,firm-and-dry,watery,?'),
    'mycelium': create_dict('absent,present,?'),
    'int-discolor': create_dict('none,brown,black,?'),
    'sclerotia': create_dict('absent,present,?'),
    'fruit-pods': create_dict('norm,diseased,few-present,dna,?'),
    'fruit spots': create_dict('absent,colored,brown-w/blk-specks,distort,dna,?'),
    'seed': create_dict('norm,abnorm,?'),
    'mold-growth': create_dict('absent,present,?'),
    'seed-discolor': create_dict('absent,present,?'),
    'seed-size': create_dict('norm,lt-norm,?'),
    'shriveling': create_dict('absent,present,?'),
    'roots': create_dict('norm,rotted,galls-cysts,?')
}
```

In [5]:

```
def get_map_val(key, val):
    if val == '?':
        search_val = '?'
    else:
        search_val = int(val)

    return maps.get(key).get(search_val)

for c in maps.keys():
    df[c] = df[c].apply(lambda x: get_map_val(c, x))
```

In [6]:

```
df
```

Out[6]:

	class	date	plant-stand	precip	temp	hail	crop-hist	area-damaged	severity	seed-tmt	...	int-discolor	sclerotia	fruit-pods	fruit spots	seed	mold-growth	di
0	diaporthe-stem-canker	october	normal	gt-norm	norm	yes	same-lst-yr	low-areas	pot-severe	none	...	none	absent	norm	dna	norm	absent	
1	diaporthe-stem-canker	august	normal	gt-norm	norm	yes	same-lst-two-yrs	scattered	severe	fungicide	...	none	absent	norm	dna	norm	absent	
2	diaporthe-stem-canker	july	normal	gt-norm	norm	yes	same-lst-yr	scattered	severe	fungicide	...	none	absent	norm	dna	norm	absent	
3	diaporthe-stem-canker	july	normal	gt-norm	norm	yes	same-lst-yr	scattered	severe	none	...	none	absent	norm	dna	norm	absent	
4	diaporthe-stem-canker	october	normal	gt-norm	norm	yes	same-lst-two-yrs	scattered	pot-severe	none	...	none	absent	norm	dna	norm	absent	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
302	2-4-d-injury	None	None	None	None	None	None	None	None	None	...	None	None	None	None	None	None	
303	herbicide-injury	may	lt-normal	None	lt-norm	None	same-lst-yr	scattered	None	None	...	None	None	dna	None	None	None	
304	herbicide-injury	april	lt-normal	None	lt-norm	None	diff-lst-year	whole-field	None	None	...	None	None	dna	None	None	None	
305	herbicide-injury	may	lt-normal	None	lt-norm	None	diff-lst-year	scattered	None	None	...	None	None	dna	None	None	None	
306	herbicide-injury	may	lt-normal	None	lt-norm	None	same-lst-yr	whole-field	None	None	...	None	None	dna	None	None	None	

307 rows × 36 columns

In [7]:

```
df.isnull().sum()
```

Out[7]:

```
class               0
date                1
plant-stand         8
precip             11
temp                7
hail               41
crop-hist           1
area-damaged        1
severity           41
seed-tmt           41
germination        36
plant-growth        1
leaves              0
leafspots-halo     25
leafspots-marg     25
leafspot-size      25
leaf-shread        26
```

```
leaf-malf          25
leaf-mild          30
stem                1
lodging            41
stem-cankers       11
canker-lesion      11
fruiting-bodies    35
external decay     11
mycelium           11
int-discolor       11
sclerotia          11
fruit-pods         25
fruit spots        35
seed               29
mold-growth        29
seed-discolor      35
seed-size          29
shriveling         35
roots               7
dtype: int64
```

In [8]:

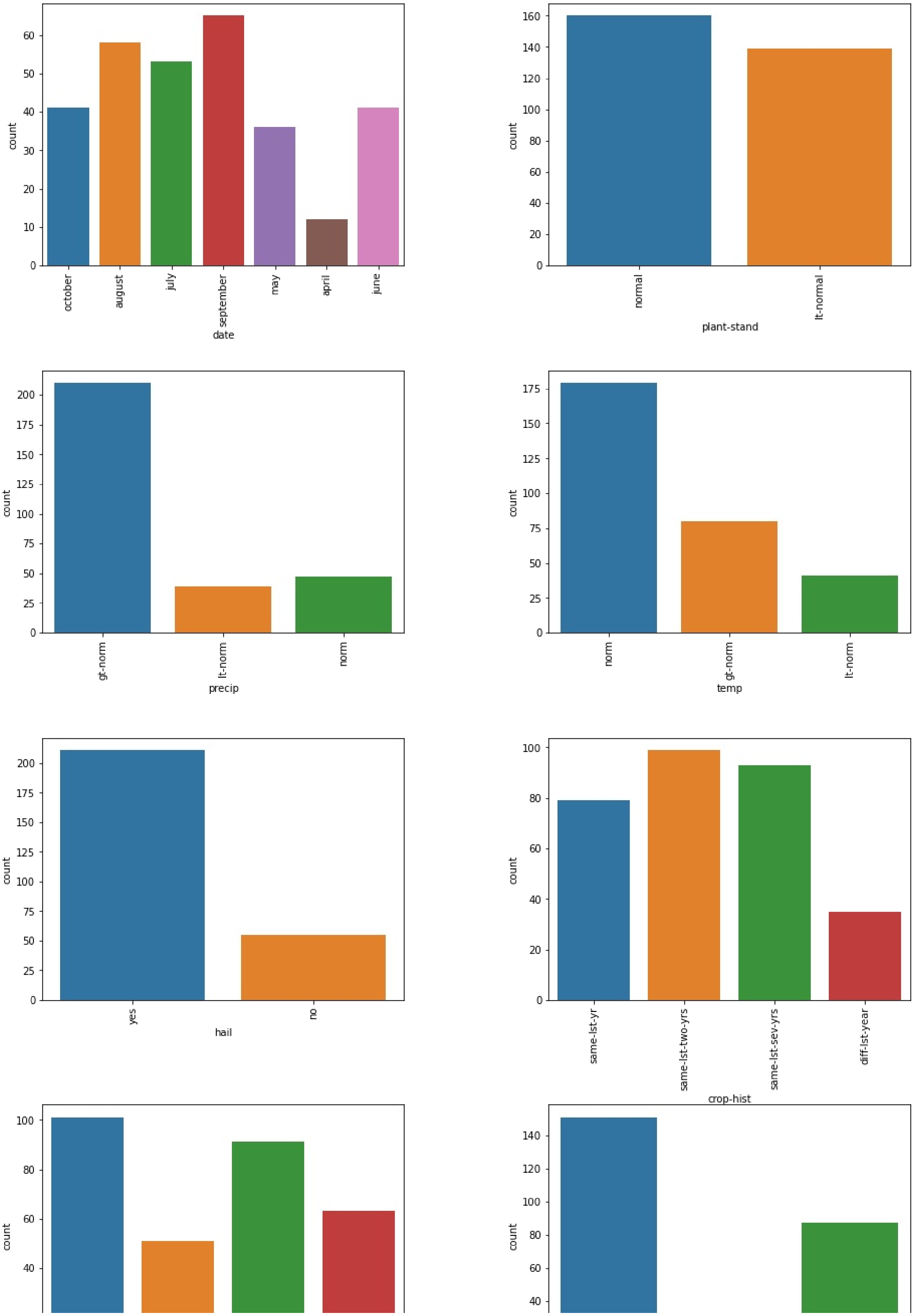
```python
df.info()
```

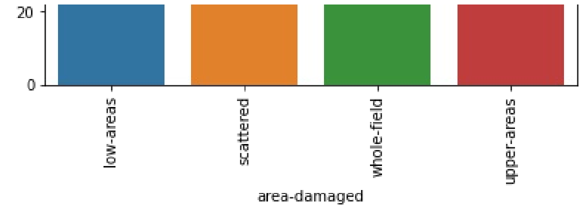
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 307 entries, 0 to 306
Data columns (total 36 columns):
 #   Column           Non-Null Count  Dtype 
---  ------           --------------  ----- 
 0   class            307 non-null    object
 1   date             306 non-null    object
 2   plant-stand      299 non-null    object
 3   precip           296 non-null    object
 4   temp             300 non-null    object
 5   hail             266 non-null    object
 6   crop-hist        306 non-null    object
 7   area-damaged     306 non-null    object
 8   severity         266 non-null    object
 9   seed-tmt         266 non-null    object
 10  germination      271 non-null    object
 11  plant-growth     306 non-null    object
 12  leaves           307 non-null    object
 13  leafspots-halo   282 non-null    object
 14  leafspots-marg   282 non-null    object
 15  leafspot-size    282 non-null    object
 16  leaf-shread      281 non-null    object
 17  leaf-malf        282 non-null    object
 18  leaf-mild        277 non-null    object
 19  stem             306 non-null    object
 20  lodging          266 non-null    object
 21  stem-cankers     296 non-null    object
 22  canker-lesion    296 non-null    object
 23  fruiting-bodies  272 non-null    object
 24  external decay   296 non-null    object
 25  mycelium         296 non-null    object
 26  int-discolor     296 non-null    object
 27  sclerotia        296 non-null    object
 28  fruit-pods       282 non-null    object
 29  fruit spots      272 non-null    object
 30  seed             278 non-null    object
 31  mold-growth      278 non-null    object
 32  seed-discolor    272 non-null    object
 33  seed-size        278 non-null    object
 34  shriveling       272 non-null    object
 35  roots            300 non-null    object
dtypes: object(36)
memory usage: 86.5+ KB
```

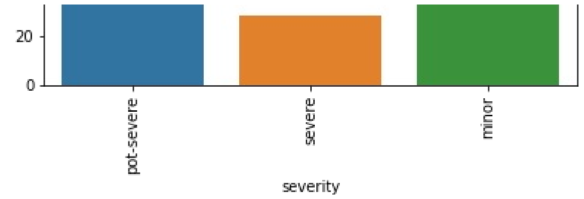
1.)Produce visualisations showing the frequency distributions for the categorical features. Are any of the distributions redundant

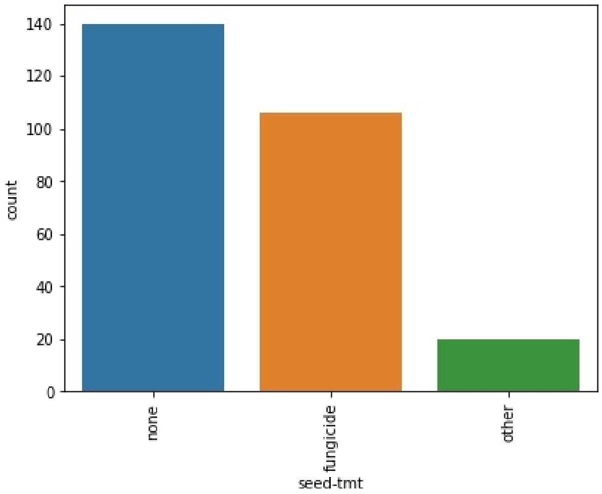
In [9]:

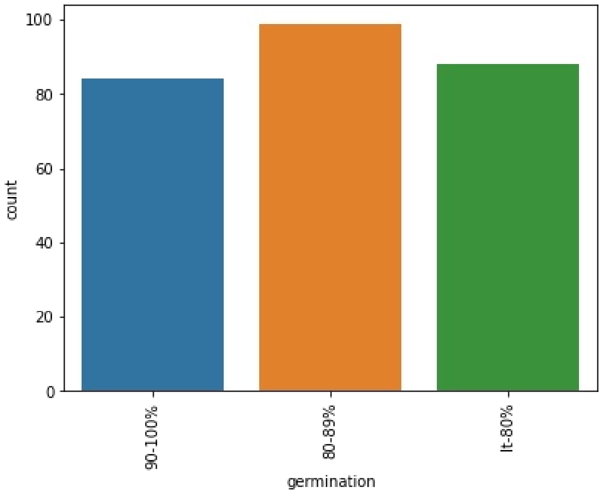
```python
import seaborn as sns
import matplotlib.pyplot as plt

cat = maps
fig, ax = plt.subplots(18, 2, figsize=(15,100))
plt.subplots_adjust(left=0.1,
                    bottom=0.1,
                    right=0.9,
                    top=1.0,
                    wspace=0.4,
                    hspace=0.4)
for variable, subplot in zip(cat, ax.flatten()):
    sns.countplot(x=df[variable], ax=subplot)
    for label in subplot.get_xticklabels():
        label.set_rotation(90)
```

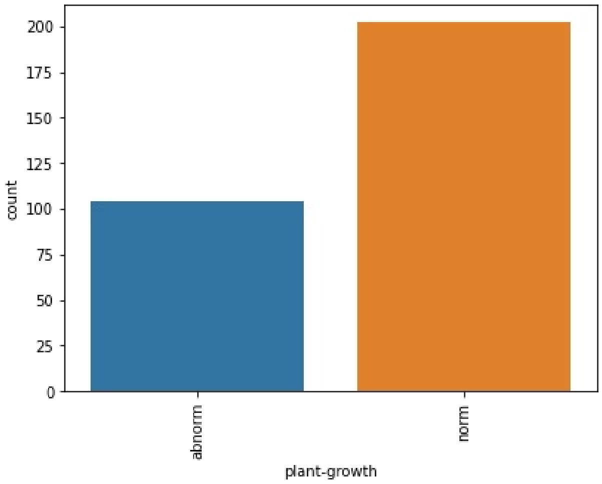


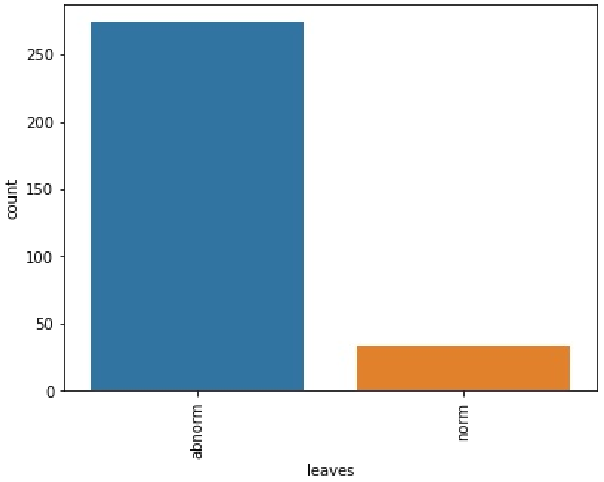

20
0
low-areas
scattered
whole-field
upper-areas
area-damaged

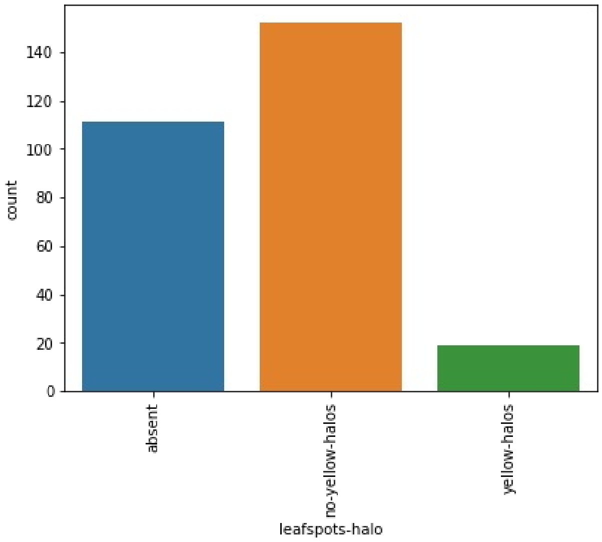

20
0
pot-severe
severe
minor
severity

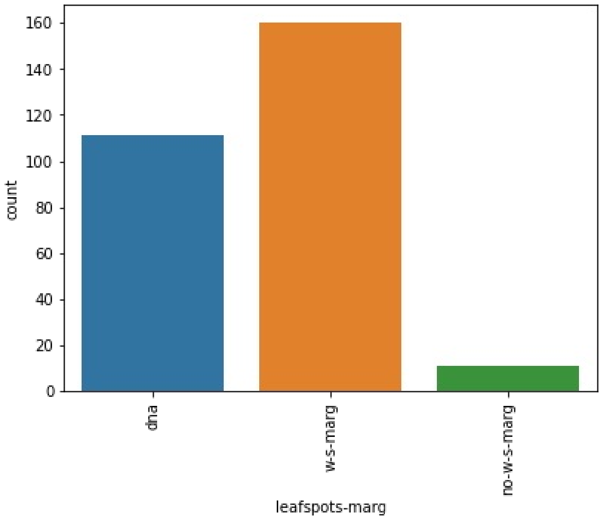

140
120
100
80
60
40
20
0
count
none
fungicide
other
seed-tmt

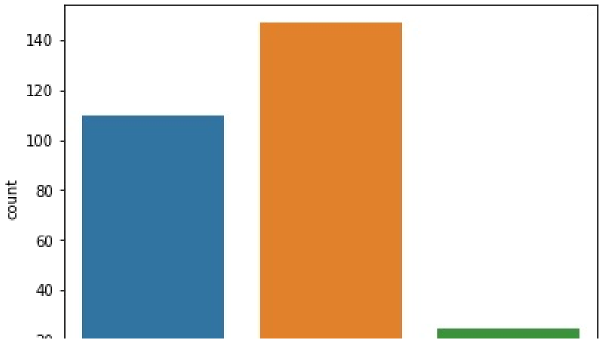

100
80
60
40
20
0
count
90-100%
80-89%
lt-80%
germination

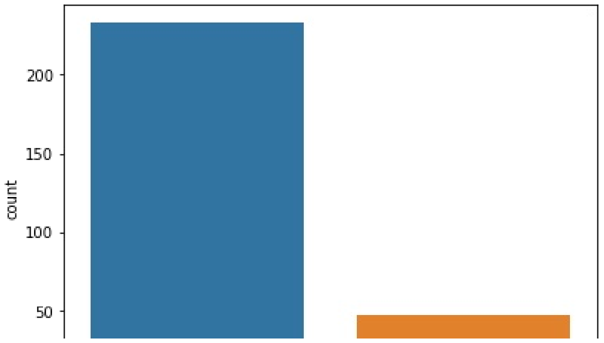

200
175
150
125
100
75
50
25
0
count
abnorm
norm
plant-growth

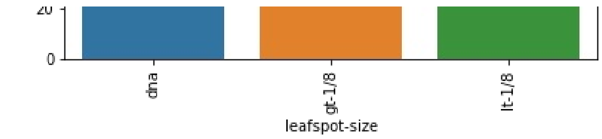

250
200
150
100
50
0
count
abnorm
norm
leaves

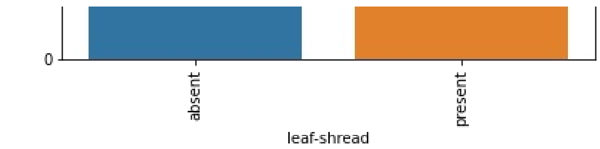

140
120
100
80
60
40
20
0
count
absent
no-yellow-halos
yellow-halos
leafspots-halo

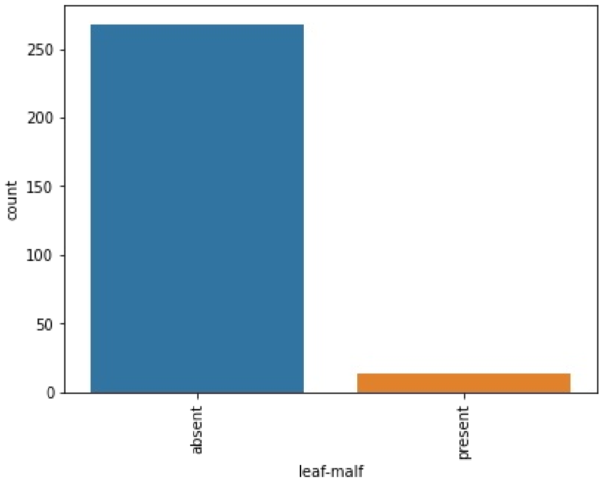

160
140
120
100
80
60
40
20
0
count
dna
w-s-marg
no-w-s-marg
leafspots-marg

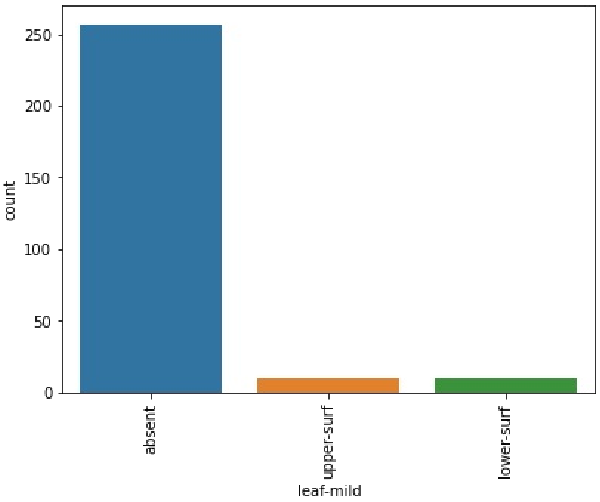

140
120
100
80
60
40
count

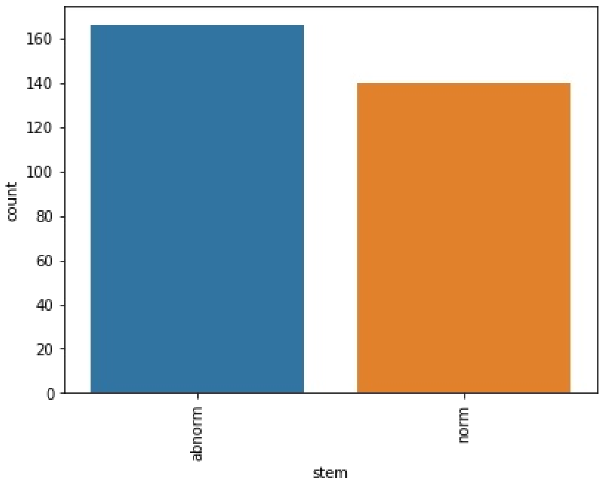

200
150
100
50
count

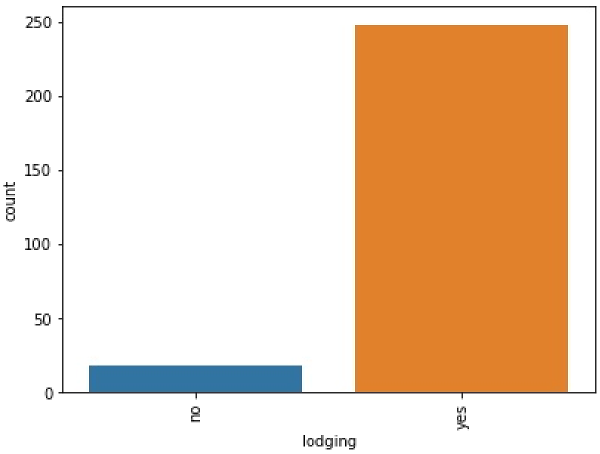

20
0
dna
gt-1/8
lt-1/8
leafspot-size

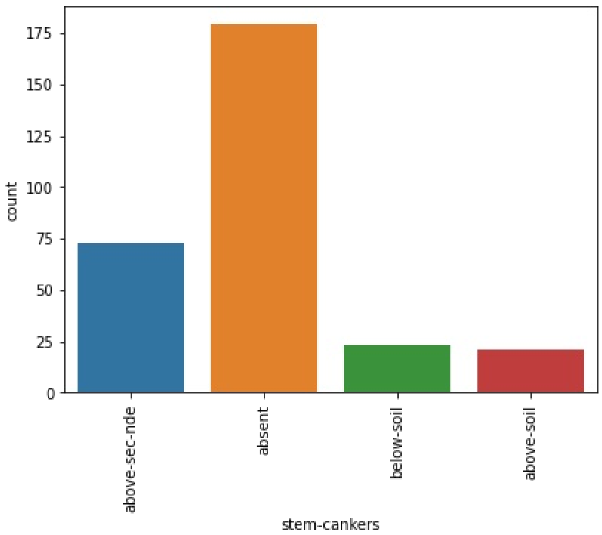

0
absent
present
leaf-shread

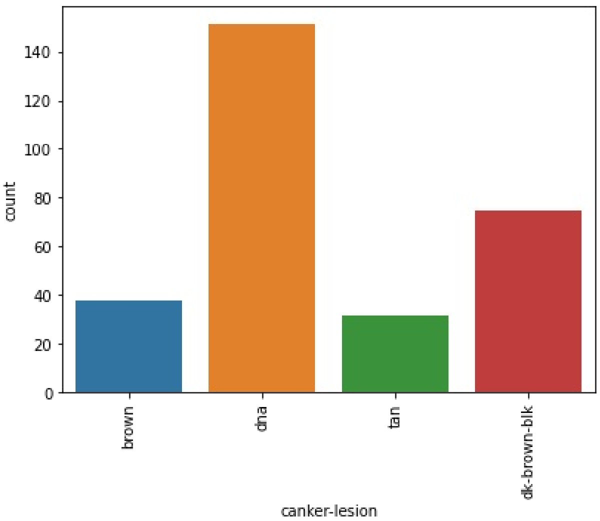

count
250
200
150
100
50
0
absent
present
leaf-malf

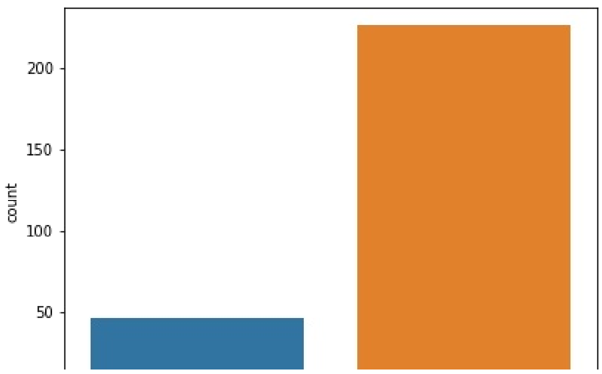

count
250
200
150
100
50
0
absent
upper-surf
lower-surf
leaf-mild

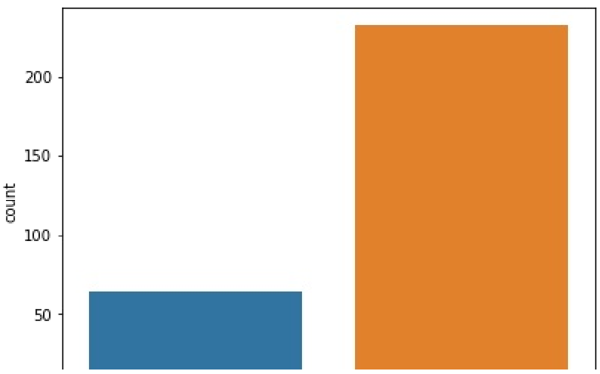

count
160
140
120
100
80
60
40
20
0
abnorm
norm
stem

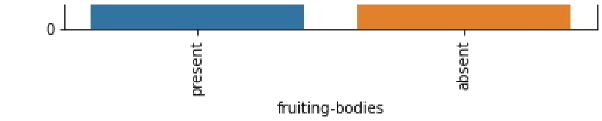

count
250
200
150
100
50
0
no
yes
lodging

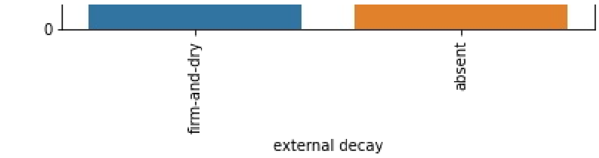

count
175
150
125
100
75
50
25
0
above-sec-nde
absent
below-soil
above-soil
stem-cankers

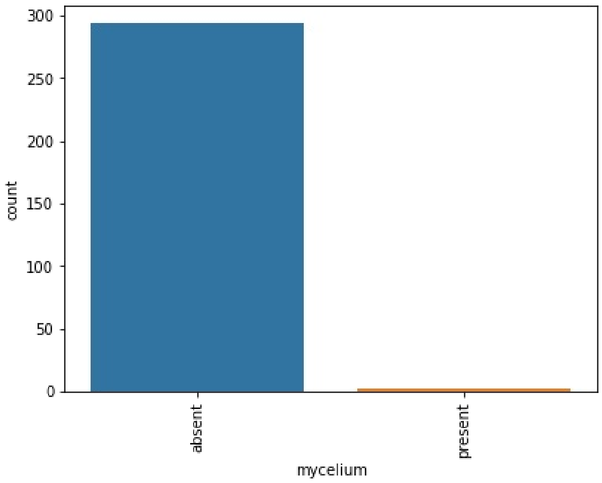

count
140
120
100
80
60
40
20
0
brown
dna
tan
dk-brown-blk
canker-lesion

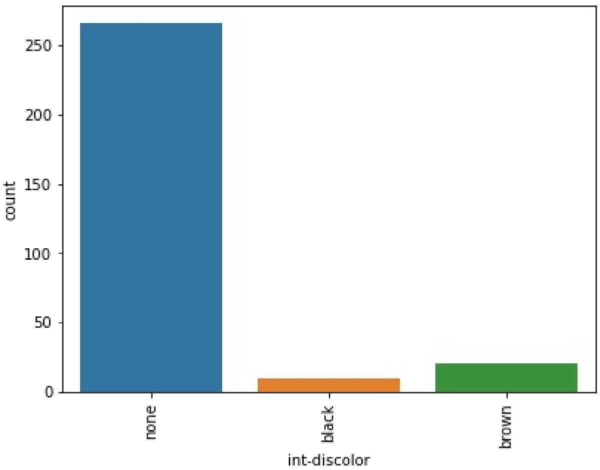

count
200
150
100
50

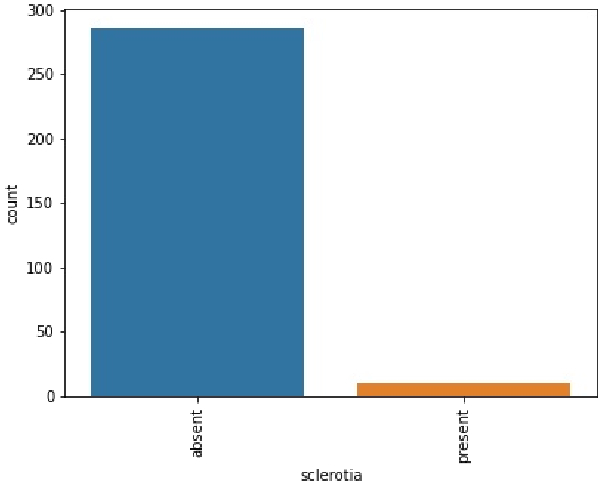

count
200
150
100
50

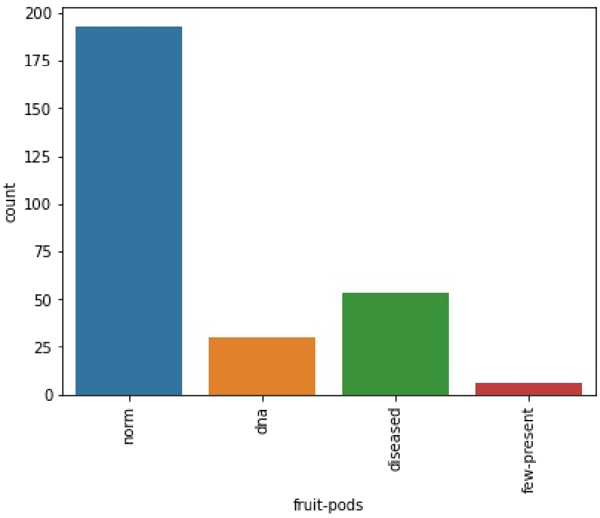

0
present
absent
fruiting-bodies

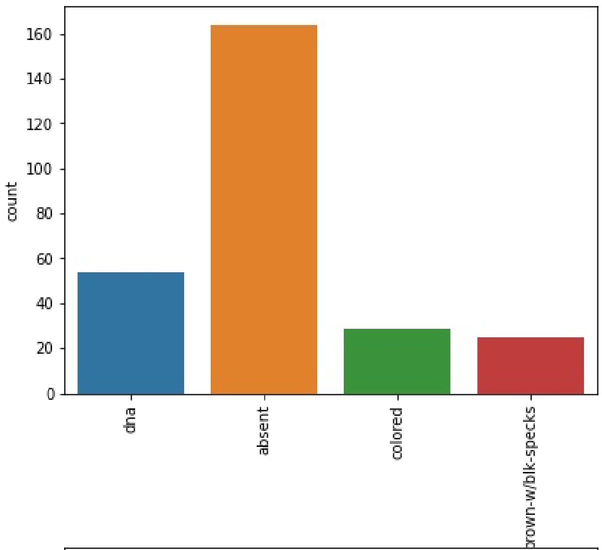

0
firm-and-dry
absent
external decay

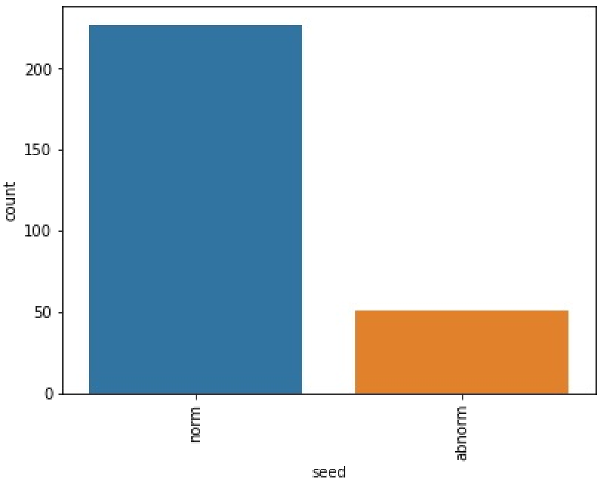

count
300
250
200
150
100
50
0
absent
present
mycelium

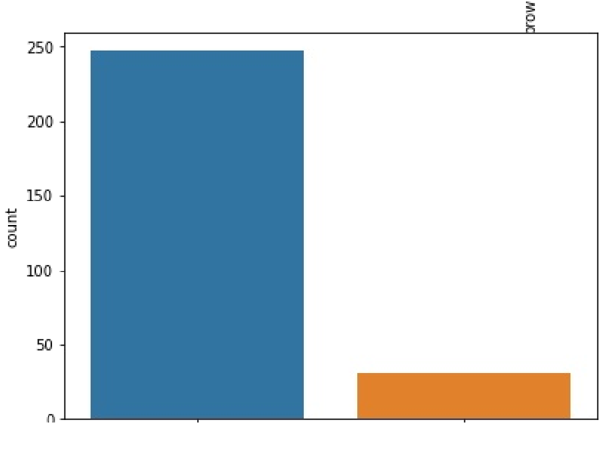

count
250
200
150
100
50
0
none
black
brown
int-discolor

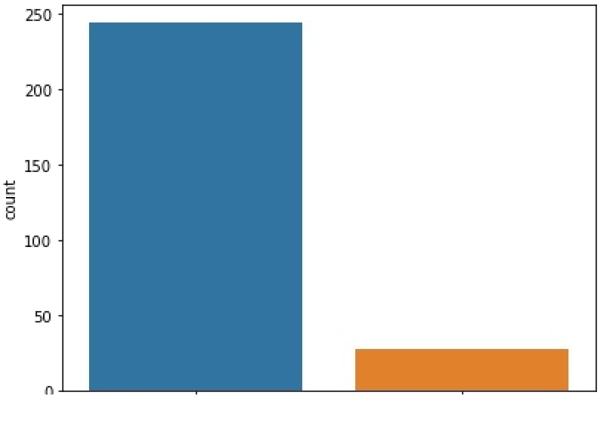

count
300
250
200
150
100
50
0
absent
present
sclerotia

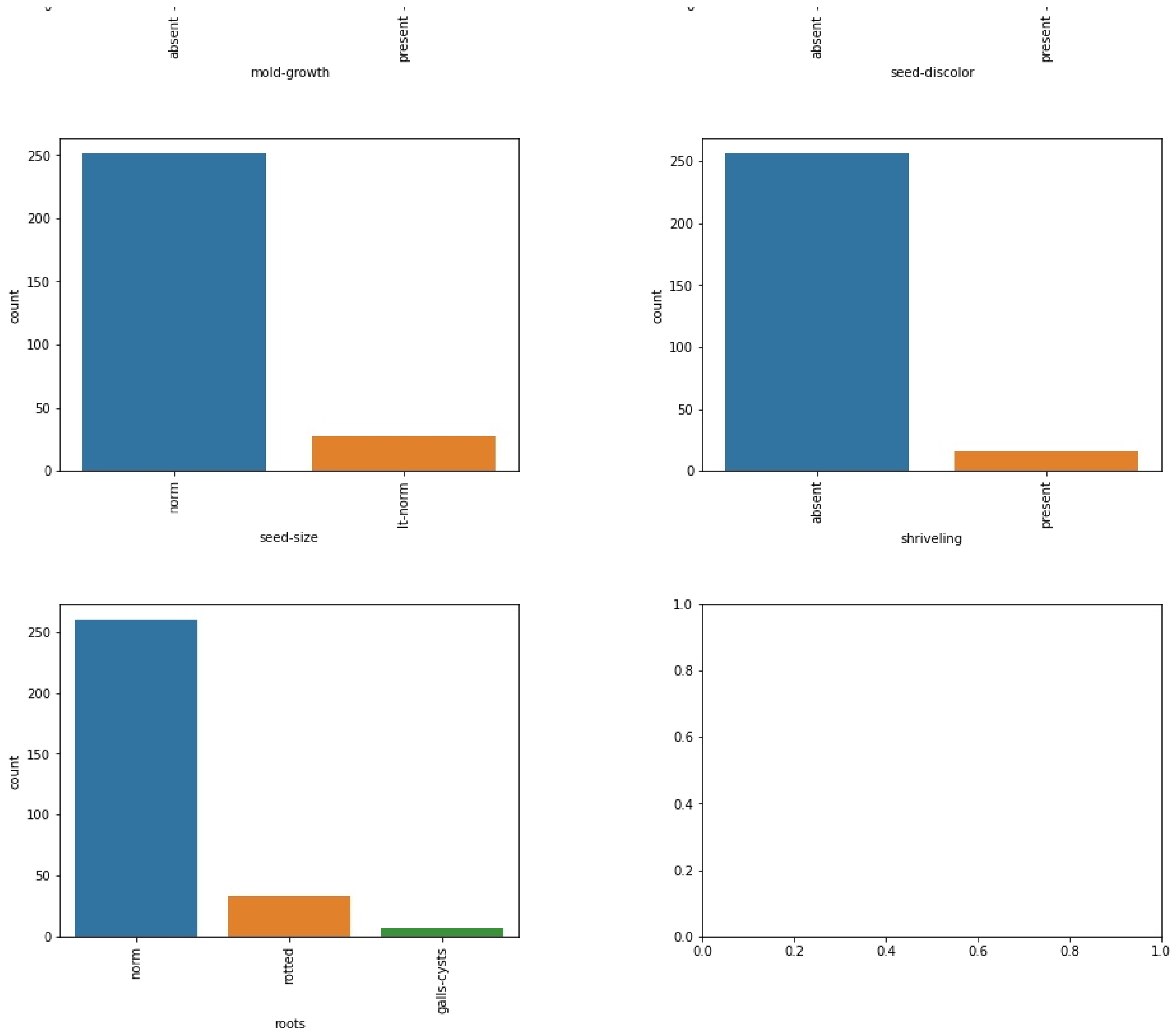

count
200
175
150
125
100
75
50
25
0
norm
dna
diseased
few-present
fruit-pods


count
160
140
120
100
80
60
40
20
0
dna
absent
colored
brown-w/blk-specks


count
200
150
100
50
0
norm
abnorm
seed


count
250
200
150
100
50
0


count
250
200
150
100
50
0



There are alot of redundant features like:

1. leafspots-halo,leafspots-marg,leafspots-size show almast identical data.
2. fruiting-bodies,external decay
3. mold-growth, seed-discolor,seed size and shivering

The features named in each row have similar,repetitive data to the rest in the row

2.)Roughly 18% of the data are missing. Are there particular features that are more likely to be missing? Does it appear to be related to the classes

In [10]:

```
df.isnull().sum()
```

Out[10]:

```
class              0
date               1
plant-stand        8
precip            11
temp               7
hail              41
crop-hist          1
area-damaged       1
severity          41
seed-tmt          41
germination       36
plant-growth       1
leaves             0
leafspots-halo    25
leafspots-marg    25
leafspot-size     25
leaf-shread       26
```

```
leaf-malf          25
leaf-mild          30
stem                1
lodging            41
stem-cankers       11
canker-lesion      11
fruiting-bodies    35
external decay     11
mycelium           11
int-discolor       11
sclerotia          11
fruit-pods         25
fruit spots        35
seed               29
mold-growth        29
seed-discolor      35
seed-size          29
shriveling         35
roots               7
dtype: int64
```

```
In [ ]:
```

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js