tu14_HW_Code

March 3, 2023

1 Breast Cancer Analysis for Each Doctors

[1]:
```
# Import Library
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

[2]:
```
# Clean Data
def bcd_load_clean():
    bcd = pd.read_csv('./data/breast_cancer_data.csv')
    bcd['patient_id'] = bcd['patient_id'].astype('string')
    bcd['doctor_name'] = bcd['doctor_name'].str.split().str[1]
    bcd['bare_nuclei'] = bcd['bare_nuclei'].replace('?', '')
    bcd['bare_nuclei'] = pd.to_numeric(bcd['bare_nuclei'])

    return bcd
```

[3]:
```
bcd = bcd_load_clean()
bcd.head()
```

[3]:

	patient_id	clump_thickness	cell_size_uniformity	cell_shape_uniformity	marginal_adhesion	single_ep_cell_size	bare_nuclei	bland_chromatin	normal_nucleoli	mitoses	class	doctor_name
0	1000025	5.0	1.0	1	1	2	1.0	3.0	1.0	1	benign	Doe
1	1002945	5.0	4.0	4	5	7	10.0	3.0	2.0	1		
2	1015425	3.0	1.0	1	1	2	2.0	3.0	1.0	1		
3	1016277	6.0	8.0	8	1	3	4.0	3.0	7.0	1		
4	1017023	4.0	1.0	1	3	2	1.0	3.0	1.0	1		

```
1              2.0          1  benign        Smith
2              1.0          1  benign          Lee
3              7.0          1  benign        Smith
4              1.0          1  benign         Wong
```

1.1 Getting the total patients + Cleaning data a little more

[4]:
```
# Cleaning a little more to get clump and bland
bcd2 = bcd.drop(labels = ['patient_id', 'cell_size_uniformity',␣
 ↪'cell_shape_uniformity',
                          'cell_shape_uniformity', 'marginal_adhesion',␣
 ↪'single_ep_cell_size',
                          'bare_nuclei', 'normal_nucleoli', 'mitoses'],
                axis = 1,              # we're selecting column - default is rows
                inplace = False)       # we could modify bcd itself with True

bcd2
```

[4]:
```
     clump_thickness  bland_chromatin      class doctor_name
0                5.0              3.0     benign         Doe
1                5.0              3.0     benign       Smith
2                3.0              3.0     benign         Lee
3                6.0              3.0     benign       Smith
4                4.0              3.0     benign        Wong
..               …                …          …          …
694              3.0              1.0     benign         Lee
695              2.0              1.0     benign       Smith
696              5.0              8.0  malignant         Lee
697              4.0             10.0  malignant         Lee
698              4.0             10.0  malignant        Wong

[699 rows x 4 columns]
```

[5]:
```
# Breif summary of the data based on class
bcd2_class = bcd2.groupby('class')
bcd2_class.describe()
```

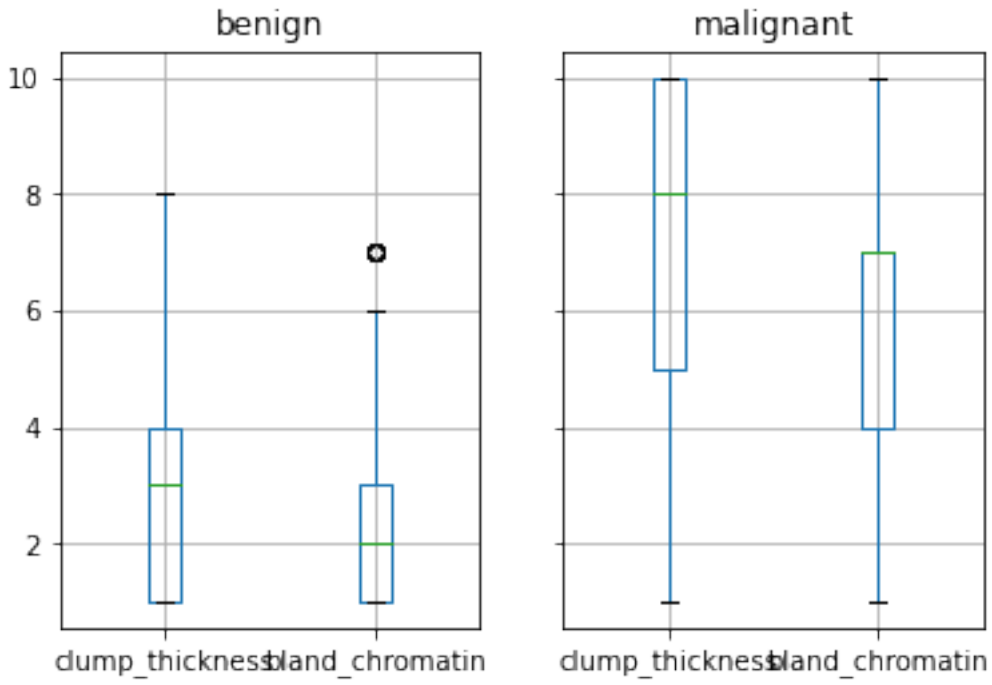
[5]:
```
           clump_thickness                                                  \
                     count      mean       std  min  25%  50%   75%   max
class
benign               458.0  2.956332  1.674318  1.0  1.0  3.0   4.0   8.0
malignant            240.0  7.204167  2.429763  1.0  5.0  8.0  10.0  10.0

           bland_chromatin
                     count      mean       std  min  25%  50%  75%   max
class
```

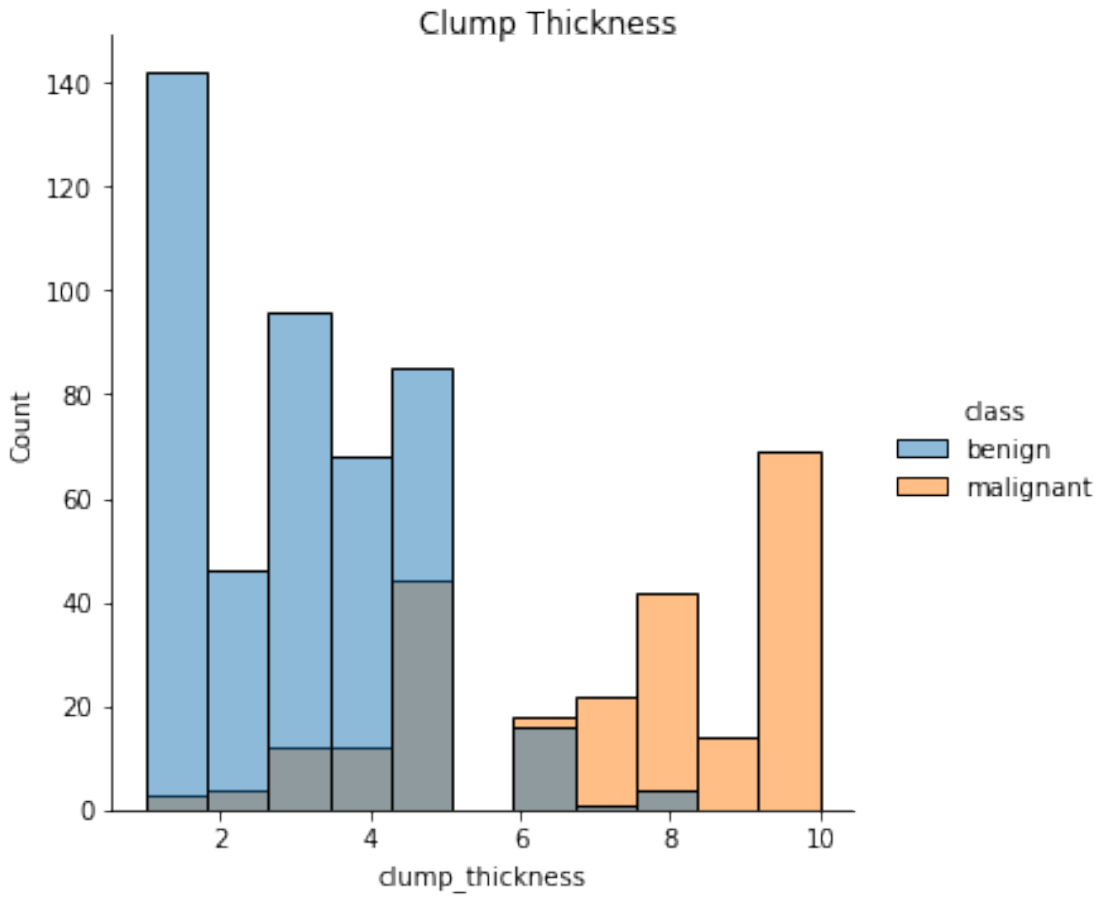
benign	455.0	2.105495	1.081417	1.0	1.0	2.0	3.0	7.0
malignant	240.0	5.991667	2.270406	1.0	4.0	7.0	7.0	10.0

- There are **699** patients in total for this data. However, there are **698** patients with clump thinkness and **695** patients with bland chromatin. The mean for clump thickness is 2.96 for benign class (std = 1.67) and 7.20 for malignant class (std = 2.43). The mean for bland chromatin is 2.11 for benign class (std = 1.08) and 5.99 for malignent class (std = 2.27). Also, benign class never reach 10 on both clump thinkness and bland chromatin.

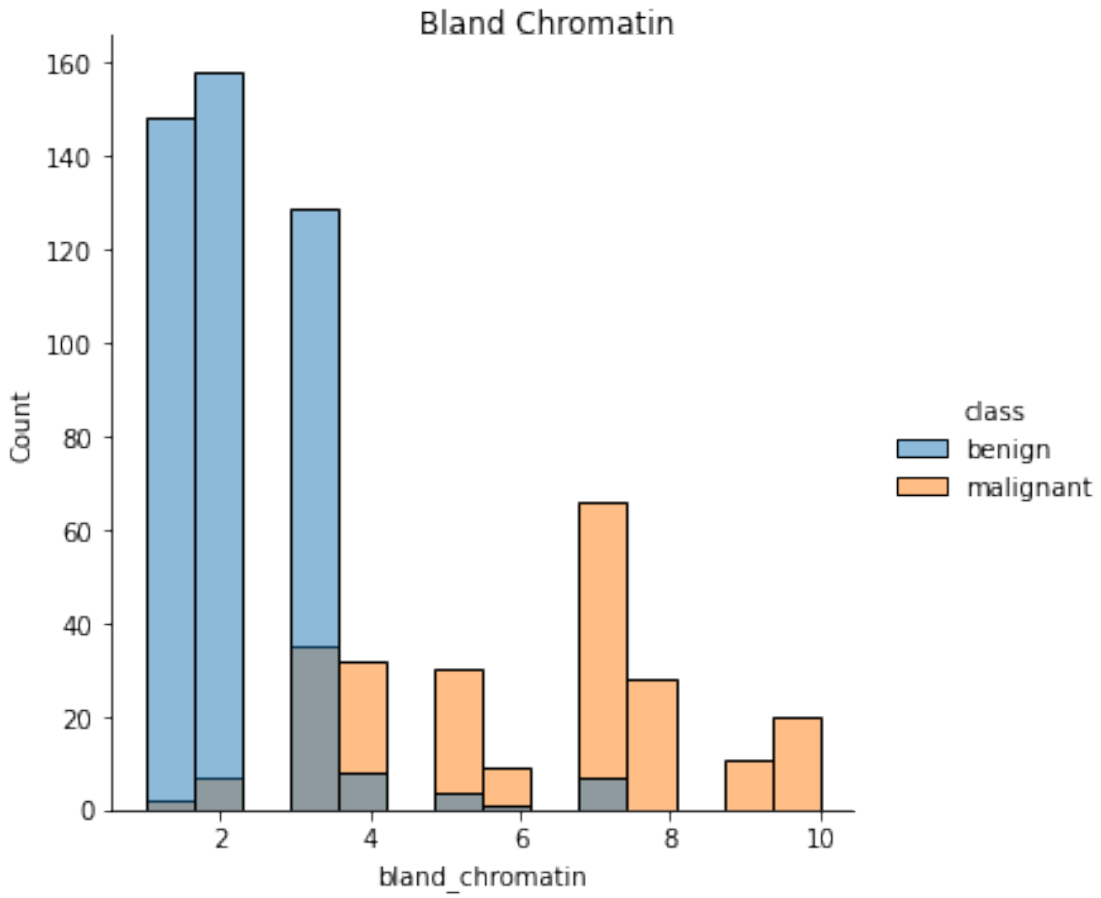
[102]:
```
bcd2_class.boxplot();
```



[58]:
```
sns.displot(data = bcd2, x = 'clump_thickness', hue = 'class').fig.
↪suptitle('Clump Thickness',y = 1, x = 0.5);
```



[59]:
```
sns.displot(data = bcd2, x = 'bland_chromatin', hue = 'class').fig.
↪suptitle('Bland Chromatin',y = 1, x = 0.5);
```



- Overall, benign class has lower value than malignant class for both clump thinkness and bland chromatin.

1.2 Data Based on Each Doctor

[9]:
```
# Get the list of doctor name
bcd2['doctor_name'].unique()
```

[9]:
```
array(['Doe', 'Smith', 'Lee', 'Wong'], dtype=object)
```

1.2.1 Overall Mean and Std

[63]:
```
bcd2.pivot_table(index = 'class',
                 columns = 'doctor_name',
                 values = 'clump_thickness',
                 aggfunc = 'mean')
```

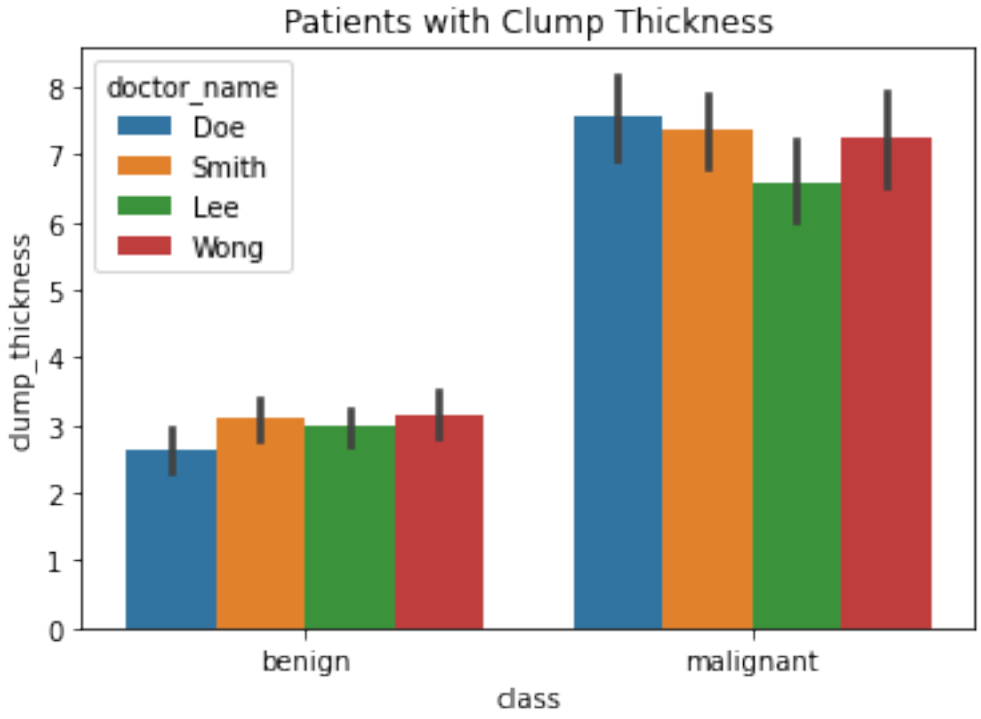
[63]:
```
doctor_name       Doe       Lee     Smith      Wong
class
```

```
benign       2.637795  2.983471  3.098039  3.166667
malignant    7.586207  6.600000  7.356164  7.265306
```

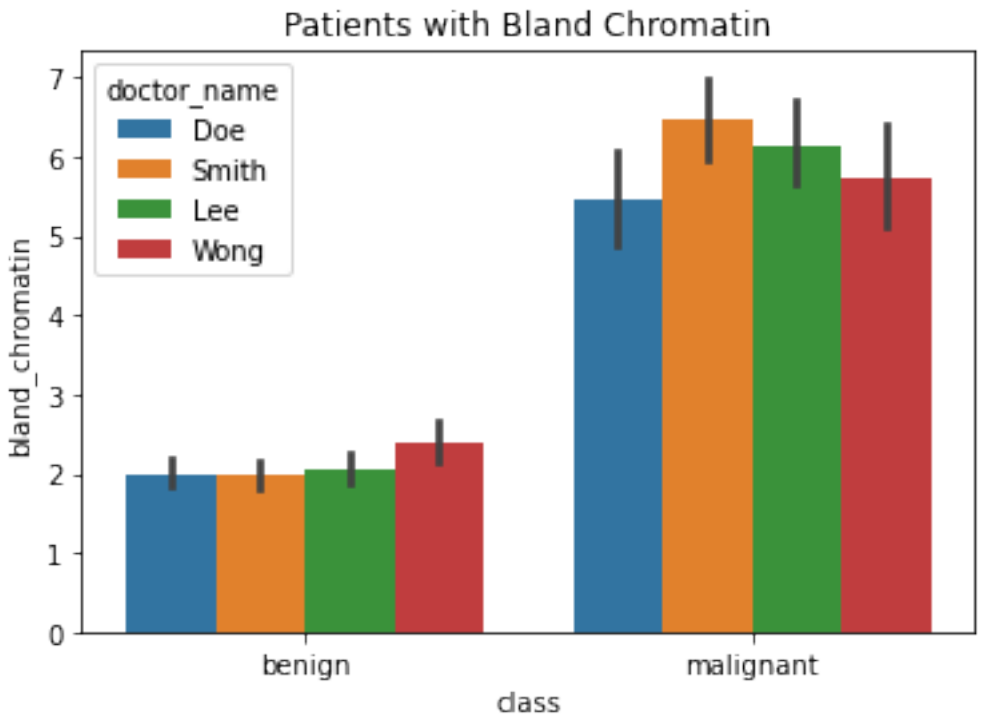
[64]:
```
bcd2.pivot_table(index = 'class',
                 columns = 'doctor_name',
                 values = 'bland_chromatin',
                 aggfunc = 'mean')
```

[64]:
```
doctor_name      Doe       Lee     Smith      Wong
class
benign       2.00000  2.067227  1.980392  2.388889
malignant    5.45614  6.150000  6.459459  5.714286
```

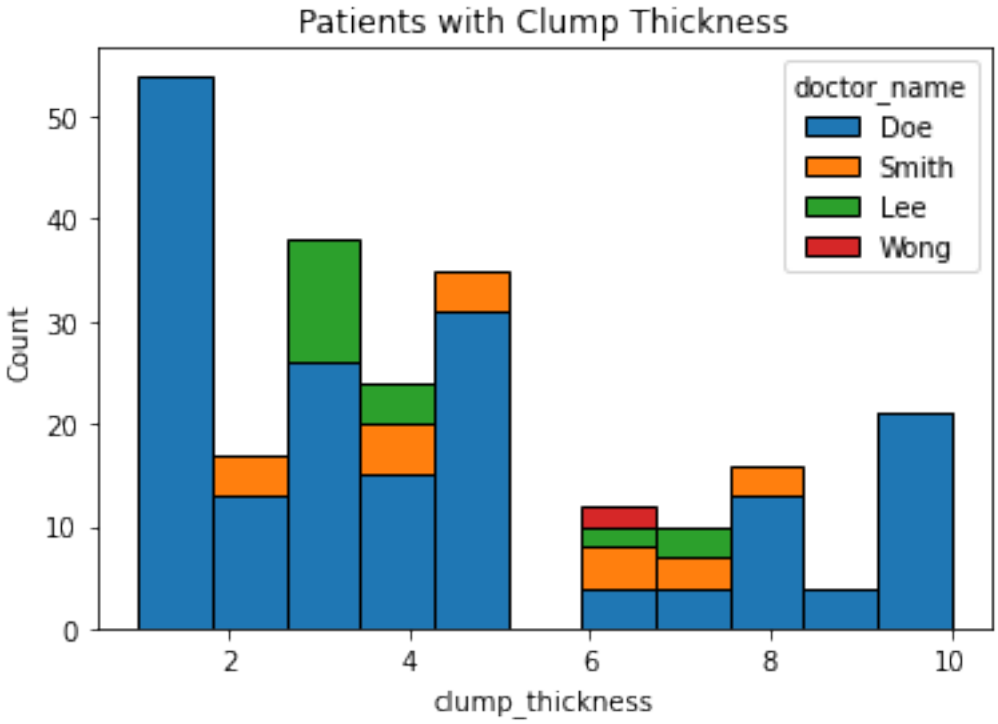
[106]:
```
sns.barplot(data = bcd2, x = 'class', y = 'clump_thickness', hue =␣
 ↪'doctor_name').set(title = 'Patients with Clump Thickness');
```



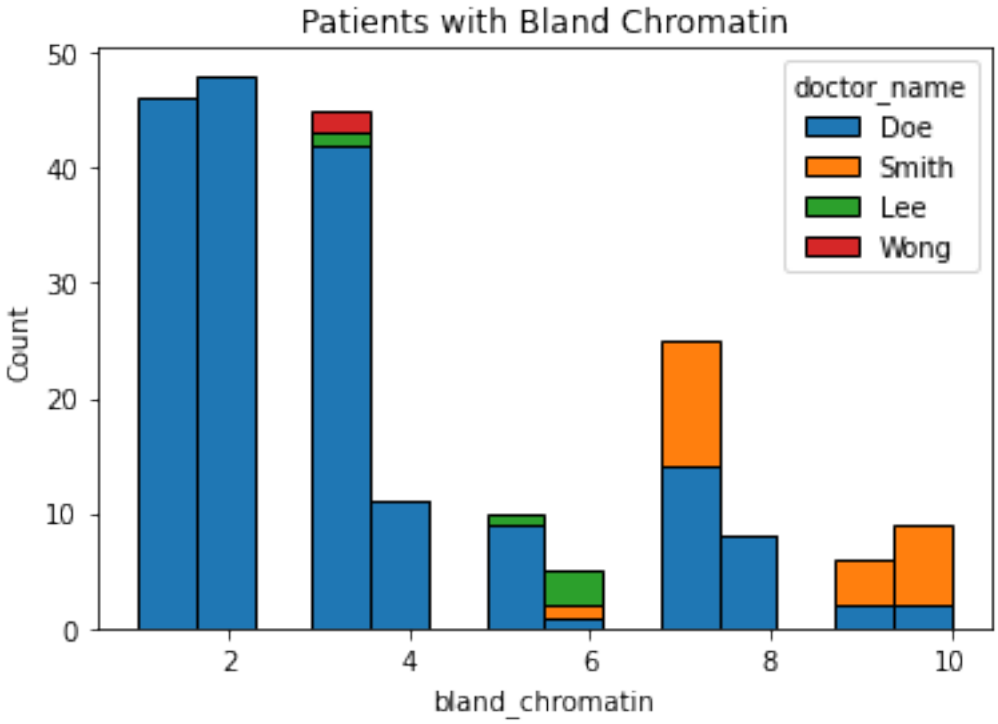
[112]:
```
sns.barplot(data = bcd2, x = 'class', y = 'bland_chromatin', hue =␣
 ↪'doctor_name').set(title = 'Patients with Bland Chromatin');
```



[111]:
```
sns.histplot(data = bcd2, x = 'clump_thickness',hue = 'doctor_name', alpha = 1).
↪set(title = 'Patients with Clump Thickness');
```

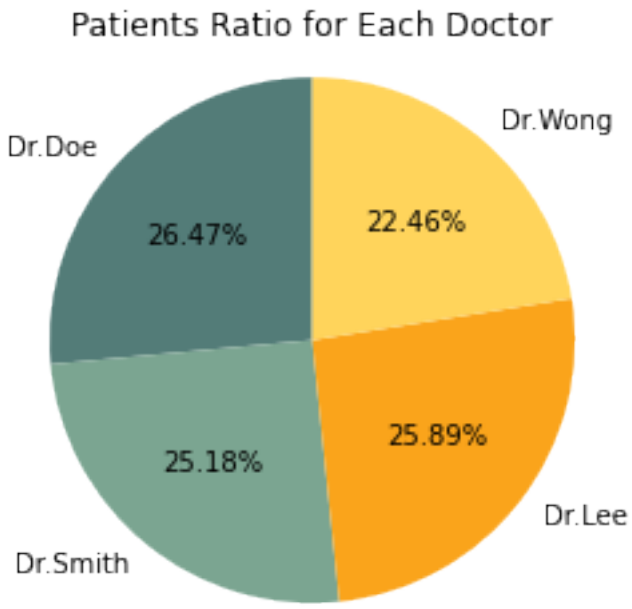


[99]:
```
sns.histplot(data = bcd2, x = 'bland_chromatin', hue = 'doctor_name', alpha =␣
↪1).set(title = 'Patients with Bland Chromatin');
```



[10]:
```
doctor_name = ['Dr.Doe', 'Dr.Smith', 'Dr.Lee', 'Dr.Wong']
doctor_patients = [185, 176, 181, 157]
doctor_color = ['#537c78','#7ba591','#faa41b','#ffd45b']

plt.pie(doctor_patients, labels = doctor_name,
        colors = doctor_color,
        autopct='%1.2f%%',
       startangle=90)
plt.title('Patients Ratio for Each Doctor')
plt.axis('equal')
plt.show()
```



- Each doctor had approximately similar number of patients.

1.2.2 Dr. Doe

[11]:
```
# Dr.Doe Data
doe = bcd2
doe = doe[doe['doctor_name'] == 'Doe']
doe.head()
```

[11]:

	clump_thickness	bland_chromatin	class	doctor_name
0	5.0	3.0	benign	Doe
6	1.0	3.0	benign	Doe
9	4.0	2.0	benign	Doe
10	1.0	3.0	benign	Doe
14	8.0	5.0	malignant	Doe

[61]:
```
doe.groupby('class').describe()
```

[61]:

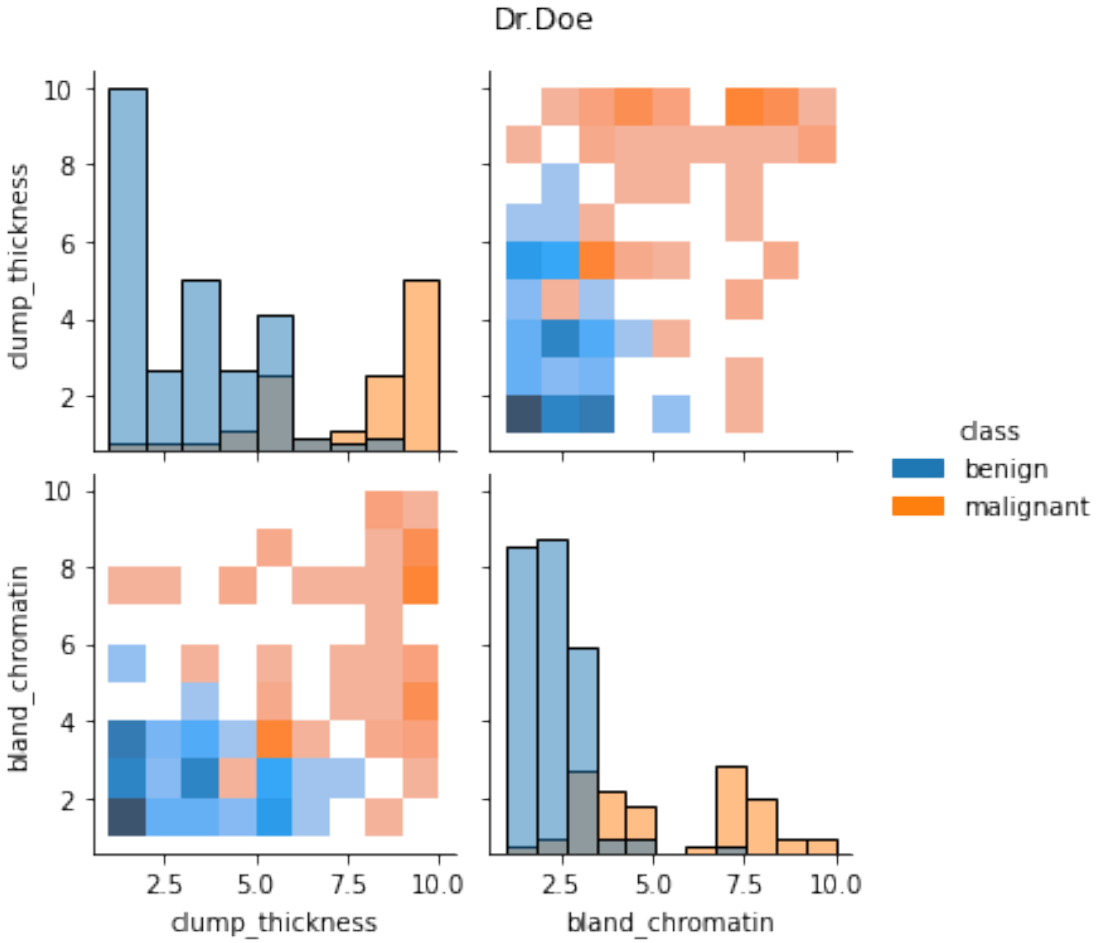
	clump_thickness								bland_chromatin							
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
class																
benign	127.0	2.637795	1.744239	1.0	1.0	2.0	4.0	8.0								
malignant	58.0	7.586207	2.464018	1.0	5.0	8.0	10.0	10.0								

```
class
benign              126.0  2.00000  1.003992  1.0  1.0  2.0  3.0   7.0
malignant            57.0  5.45614  2.260453  1.0  3.0  5.0  7.0  10.0
```

- Dr. Doe had 185 patients.

[52]:
```
doe_pair = sns.pairplot(data = doe, hue = 'class', kind = 'hist',␣
 ↪plot_kws=dict(binwidth = 1));
doe_pair.fig.suptitle('Dr.Doe',y = 1.05, x = 0.5)
```

[52]:
```
Text(0.5, 1.05, 'Dr.Doe')
```



1.2.3 Dr. Smith

[13]:
```
# Dr.Smith Data
smith = bcd2
smith = smith[smith['doctor_name'] == 'Smith']
```

```
smith.head()
```

[13]:

	clump_thickness	bland_chromatin	class	doctor_name
1	5.0	3.0	benign	Smith
3	6.0	3.0	benign	Smith
5	8.0	9.0	malignant	Smith
7	2.0	3.0	benign	Smith
8	2.0	1.0	benign	Smith

- Dr.Smith has 176 patients.

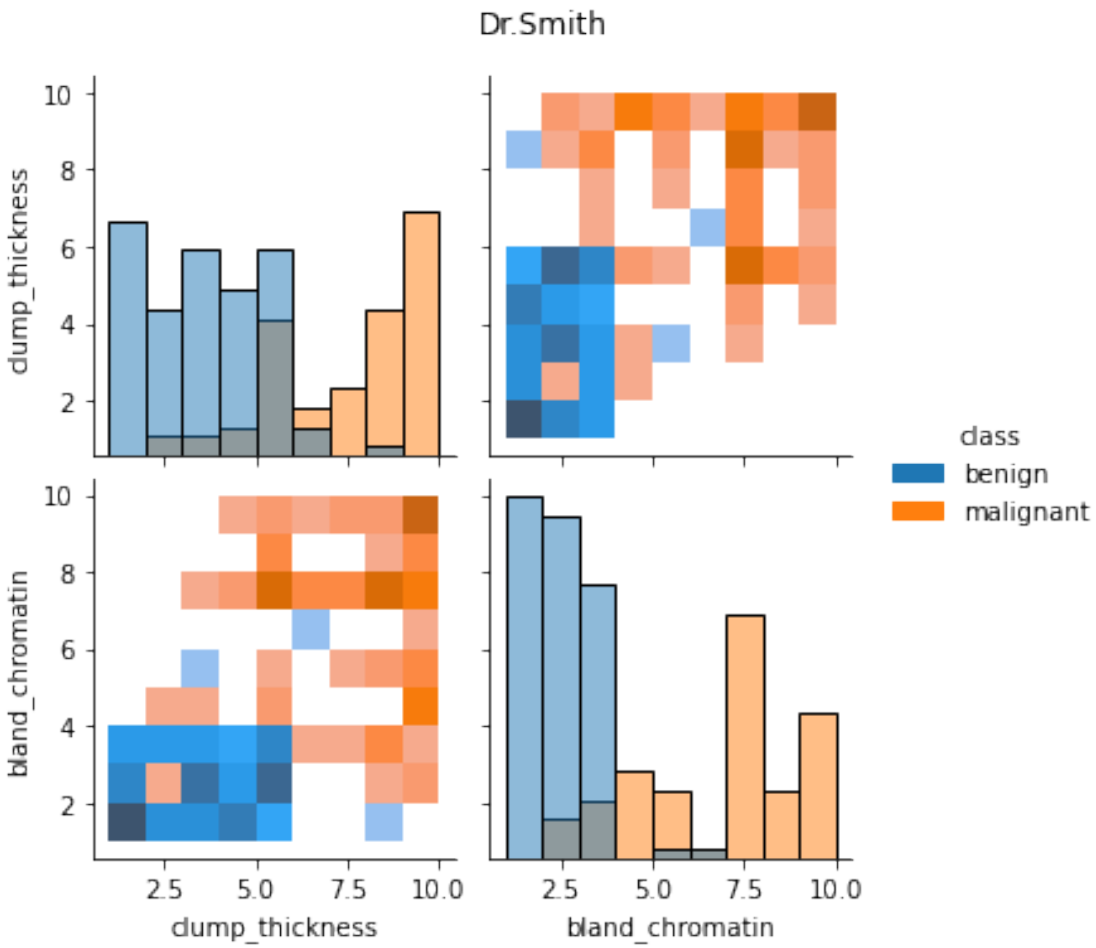
[65]:
```
smith.groupby('class').describe()
```

[65]:

	clump_thickness								bland_chromatin							
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
class																
benign	102.0	3.098039	1.613739	1.0	2.0	3.0	4.0	8.0	102.0	1.980392	0.943769	1.0	1.00	2.0	3.0	6.0
malignant	73.0	7.356164	2.299576	2.0	5.0	8.0	10.0	10.0	74.0	6.459459	2.330202	2.0	4.25	7.0	8.0	10.0

[53]:
```
smith_pair = sns.pairplot(data = smith, hue = 'class', kind = 'hist',␣
↪plot_kws=dict(binwidth = 1));
smith_pair.fig.suptitle('Dr.Smith',y = 1.05, x = 0.5)
```

[53]:
```
Text(0.5, 1.05, 'Dr.Smith')
```



1.2.4 Dr. Lee

[15]:
```
# Dr.Lee Data
lee = bcd2[bcd2['doctor_name'] == 'Lee']
lee.head()
```

[15]:

	clump_thickness	bland_chromatin	class	doctor_name
2	3.0	3.0	benign	Lee
15	7.0	4.0	malignant	Lee
16	4.0	2.0	benign	Lee
27	5.0	2.0	benign	Lee
31	2.0	3.0	benign	Lee

- Dr. Lee had 181 patients.

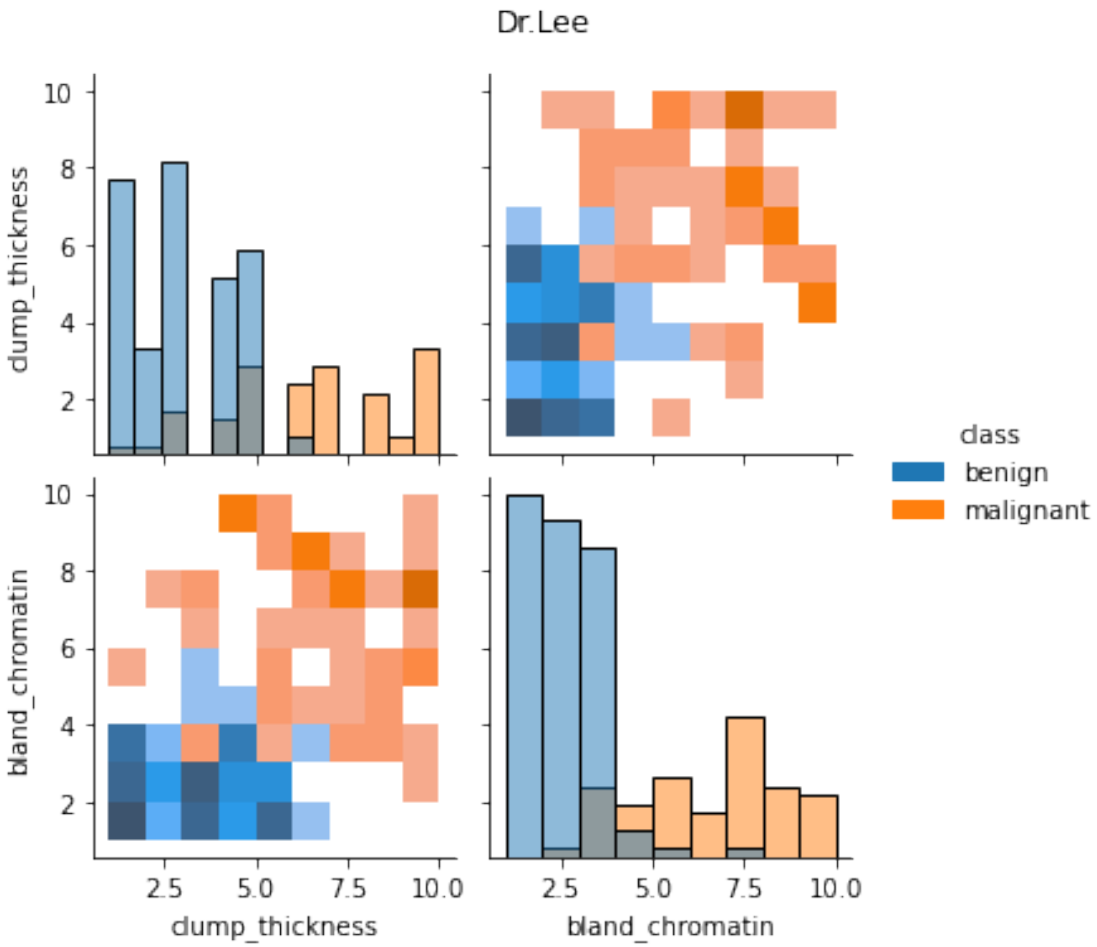
[66]:
```
lee.groupby('class').describe()
```

[66]:
```
           clump_thickness                                                \
                     count      mean       std  min  25%  50%  75%   max
class
benign               121.0  2.983471  1.488755  1.0  1.0  3.0  4.0   6.0
malignant             60.0  6.600000  2.394910  1.0  5.0  7.0  8.0  10.0

           bland_chromatin
                     count      mean       std  min   25%  50%   75%   max
class
benign               119.0  2.067227  1.014564  1.0  1.00  2.0  3.00   7.0
malignant             60.0  6.150000  2.121920  2.0  4.75  7.0  7.25  10.0
```

[54]:
```
lee_pair = sns.pairplot(data = lee, hue = 'class', kind = 'hist',␣
 ↪plot_kws=dict(binwidth = 1));
lee_pair.fig.suptitle('Dr.Lee',y = 1.05, x = 0.5)
```

[54]:
```
Text(0.5, 1.05, 'Dr.Lee')
```



1.2.5 Dr. Wong

[17]:
```
# Dr. Wong Data
wong = bcd2
wong = wong[wong['doctor_name'] == 'Wong']
wong.head()
```

[17]:

	clump_thickness	bland_chromatin	class	doctor_name
4	4.0	3.0	benign	Wong
13	1.0	3.0	benign	Wong
19	6.0	3.0	benign	Wong
37	6.0	7.0	benign	Wong
38	5.0	5.0	malignant	Wong

- Dr. Wong had 157 patients.

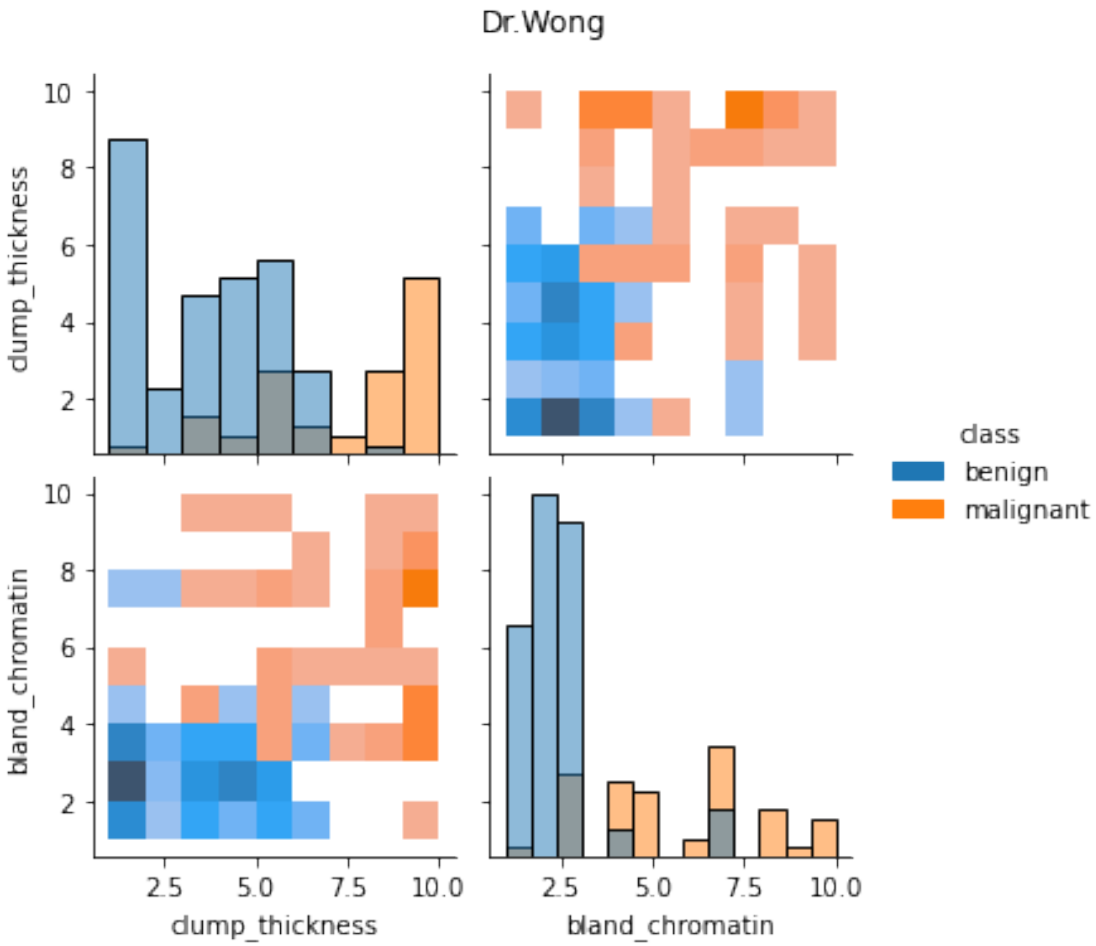
[67]:
```
wong.groupby('class').describe()
```

[67]:

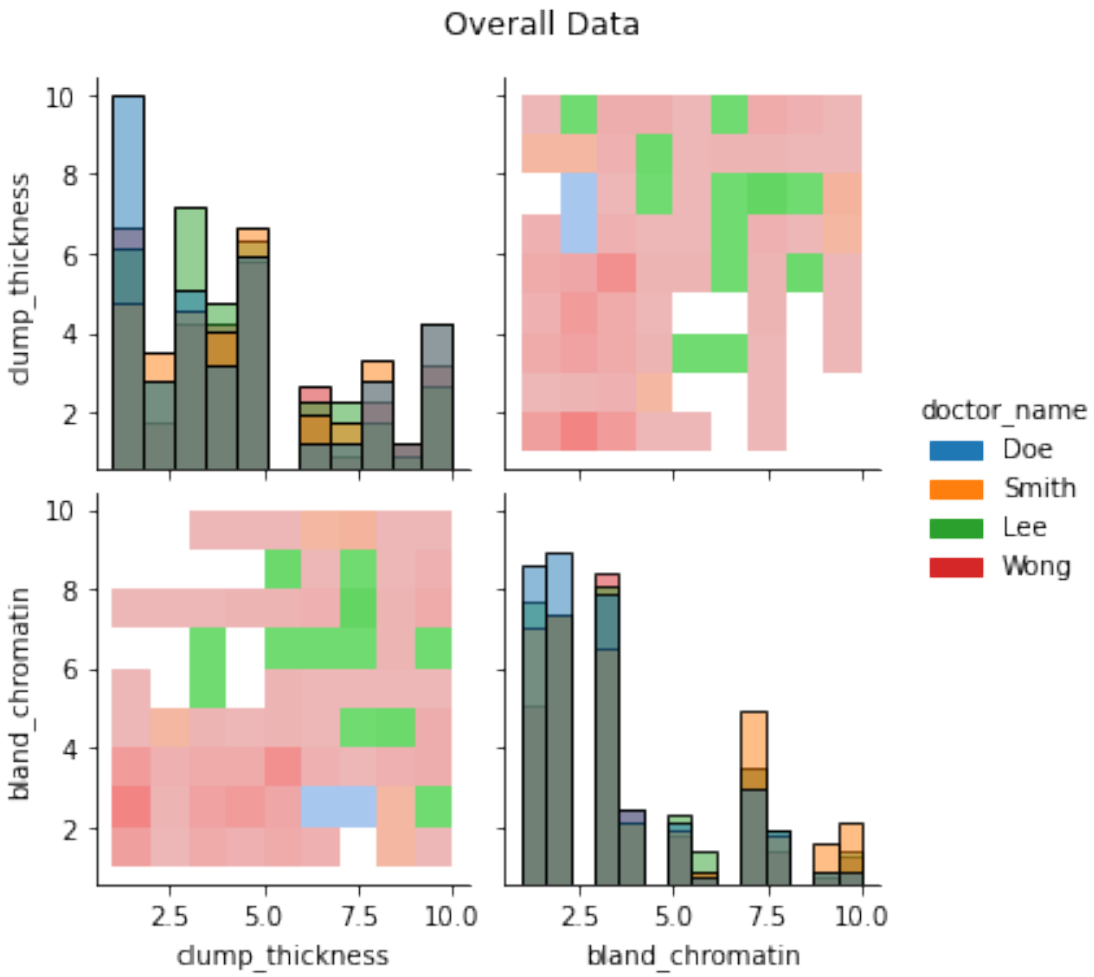
	clump_thickness								bland_chromatin							
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
class																
benign	108.0	3.166667	1.806013	1.0	1.0	3.0	5.0	8.0	108.0	2.388889	1.303004	1.0	2.0	2.0	3.0	7.0
malignant	49.0	7.265306	2.555839	1.0	5.0	8.0	10.0	10.0	49.0	5.714286	2.263846	1.0	4.0	5.0	7.0	10.0

[55]:
```
wong_pair = sns.pairplot(data = wong, hue = 'class', kind = 'hist',␣
 ↪plot_kws=dict(binwidth = 1));
wong_pair.fig.suptitle('Dr.Wong',y = 1.05, x = 0.5)
```

[55]:
```
Text(0.5, 1.05, 'Dr.Wong')
```



[100]:
```
# Maybe add another axes 2x2 comparing each of them
sns.pairplot(data = bcd2, hue = 'doctor_name', kind = 'hist',␣
 ↪plot_kws=dict(binwidth = 1)).fig.suptitle('Overall Data',y = 1.05, x = 0.5);
```


Overall Data
dump_thickness
bland_chromatin
10
8
6
4
2
2.5
5.0
7.5
10.0
doctor_name
Doe
Smith
Lee
Wong