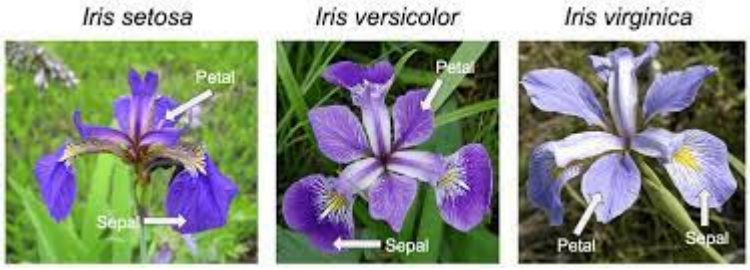
Lenara Sitshayeva

Project: IRIS dataset



Dataset load and create dataframe

Measure of centre

MEAN

MEDIAN

MODE

Measure of spread

VARIANCE

STANDARD DEVIATION

RANGE

QUARTILES AND IQR

SKEWNESS

KURTOSIS

Correlation

CORRELATION COEFFICIENTS

CORRELATION MATRIX

Dataset load and create dataframe

In [1]:
```
import pandas as pd
import matplotlib as plt
import seaborn as sns
import numpy as np
```

In [2]:
```
df = pd.read_csv("iris.csv")
```

In [3]:
```
df = df.iloc[:,1:]
df
```

Out[3]:

	sepal_length	sepal_width	petal_length	petal_width	Species
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	Iris-virginica
146	6.3	2.5	5.0	1.9	Iris-virginica
147	6.5	3.0	5.2	2.0	Iris-virginica
148	6.2	3.4	5.4	2.3	Iris-virginica
149	5.9	3.0	5.1	1.8	Iris-virginica

150 rows × 5 columns

In [4]:
```
df.iloc[:, :-1].head(4)
```

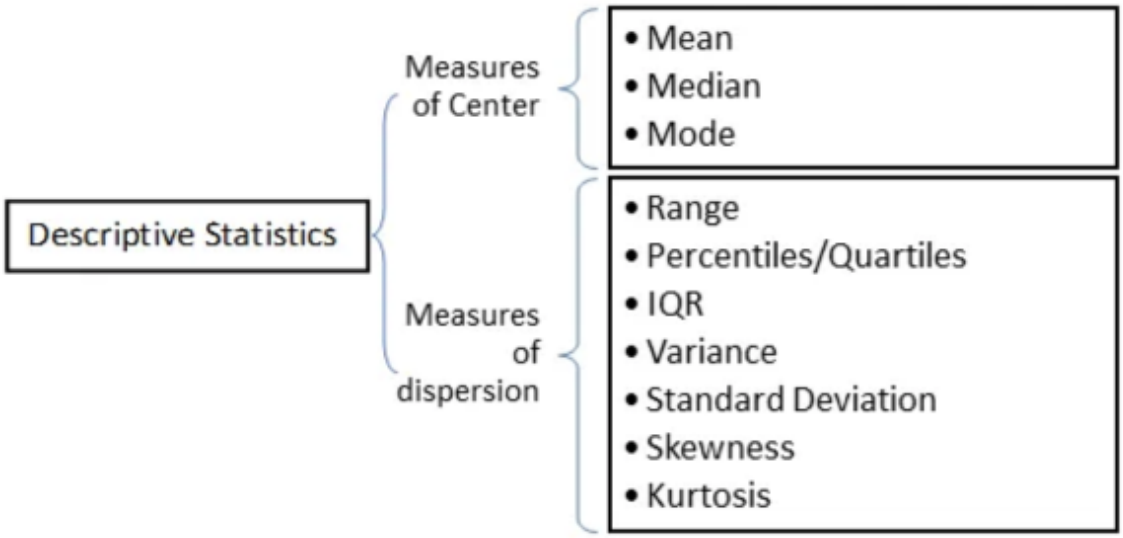
Out[4]:

	sepal_length	sepal_width	petal_length	petal_width
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2

In [5]:
```
df.describe()
```

Out[5]:

	sepal_length	sepal_width	petal_length	petal_width
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.054000	3.758667	1.198667
std	0.828066	0.433594	1.764420	0.763161
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000



Measure of centre

Measures of central tendancy of middle values of dataset: **mean, median,mode**

MEAN

In [6]:
```
df['sepal_length'].mean()
```

Out[6]:
```
5.843333333333335
```

In [7]:
```
df['sepal_width'].mean()
```

Out[7]:
```
3.0540000000000007
```

In [8]:
```
df['petal_length'].mean()
```

Out[8]:
```
3.7586666666666693
```

In [9]:
```
df['petal_width'].mean()
```

Out[9]:
```
1.1986666666666672
```

MEDIAN

In [10]:
```
df['sepal_length'].median()
```

Out[10]:
```
5.8
```

In [11]:
```
df['sepal_width'].median()
```

Out[11]:
```
3.0
```

In [12]:
```
df['petal_length'].median()
```

Out[12]:
```
4.35
```

In [13]:
```
df['petal_width'].median()
```

Out[13]:
```
1.3
```

MODE

In [14]:
```
df['sepal_length'].mode()
```

Out[14]:
```
0    5.0
Name: sepal_length, dtype: float64
```

In [15]:
```
df['sepal_width'].mode()
```

Out[15]:
```
0    3.0
Name: sepal_width, dtype: float64
```

In [16]:
```
df['petal_length'].mode()
```

Out[16]:
```
0    1.5
Name: petal_length, dtype: float64
```

In [17]:
```
df['petal_width'].mode()
```

Out[17]:
```
0    0.2
Name: petal_width, dtype: float64
```

Measure of spread

Measures of spread include the **range, quartiles and the interquartile range, variance and standard deviation, skewness.**

VARIANCE

In [18]:
```
df['sepal_length'].var()
```

Out[18]:
```
0.6856935123042505
```

In [19]:
```
df['sepal_width'].var()
```

Out[19]:
```
0.18800402684563763
```

In [20]:
```
df['petal_length'].var()
```

Out[20]:
```
3.1131794183445156
```

In [21]:
```
df['petal_width'].var()
```

Out[21]:
```
0.5824143176733784
```

STANDARD DEVIATION

In [22]:
```
df['sepal_length'].std()
```

Out[22]:
```
0.8280661279778629
```

In [23]:
```
df['sepal_width'].std()
```

Out[23]:
```
0.4335943113621737
```

In [24]:
```
df['petal_length'].std()
```

Out[24]:
```
1.7644204199522617
```

In [25]:
```
df['petal_width'].std()
```

Out[25]:
```
0.7631607417008414
```

RANGE

In [26]:
```
r_sepal_width = df['sepal_length'].max() - df['sepal_length'].min()
r_sepal_width
```

Out[26]:
```
3.6000000000000005
```

In [27]:
```
r_sepal_width = df['sepal_width'].max() - df['sepal_width'].min()
r_sepal_width
```

Out[27]:
```
2.4000000000000004
```

In [28]:
```
r_petal_length = df['petal_length'].max() - df['petal_length'].min()
r_petal_length
```

Out[28]:
```
5.9
```

In [29]:
```
r_petal_width = df['petal_width'].max() - df['petal_width'].min()
r_petal_width
```

Out[29]:
```
2.4
```

QUARTILES AND IQR

sepal_length

In [30]:
```
Q = df['sepal_length'].quantile([.25,.50,.75])
Q
```

Out[30]:
```
0.25    5.1
0.50    5.8
0.75    6.4
Name: sepal_length, dtype: float64
```

In [31]:
```
IQR = Q[.75] - Q[.25]
IQR
```

Out[31]:
```
1.3000000000000007
```

sepal_width

In [32]:
```
Q = df['sepal_width'].quantile([.25,.50,.75])
Q
```

Out[32]:
```
0.25    2.8
0.50    3.0
0.75    3.3
Name: sepal_width, dtype: float64
```

In [33]:
```
IQR = Q[.75] - Q[.25]
IQR
```

Out[33]:
```
0.5
```

petal_length

In [34]:
```
Q = df['petal_length'].quantile([.25,.50,.75])
Q
```

Out[34]:
```
0.25    1.60
0.50    4.35
0.75    5.10
Name: petal_length, dtype: float64
```

In [35]:
```
IQR = Q[.75] - Q[.25]
IQR
```

Out[35]:
```
3.4999999999999996
```

petal_width

In [36]:
```
Q = df['petal_width'].quantile([.25,.50,.75])
Q
```

Out[36]:
```
0.25    0.3
0.50    1.3
0.75    1.8
Name: petal_width, dtype: float64
```

In [37]:
```
IQR = Q[.75] - Q[.25]
IQR
```

Out[37]:
```
1.5
```

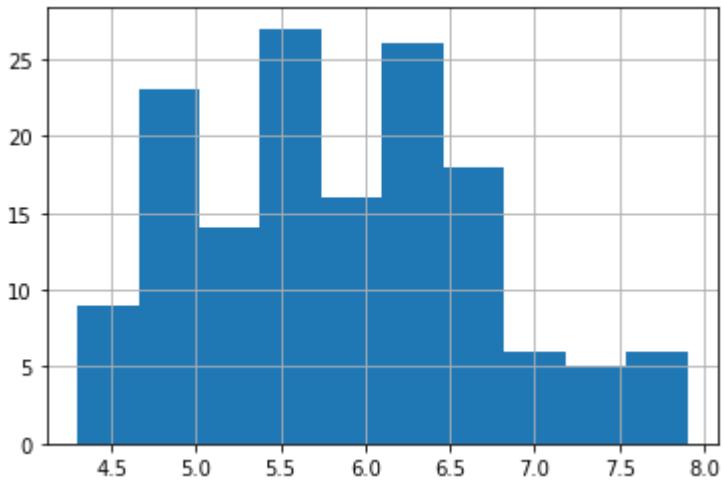
SKEWNESS

In [38]:
```
df['sepal_length'].skew()
```

Out[38]:
```
0.3149109566369728
```

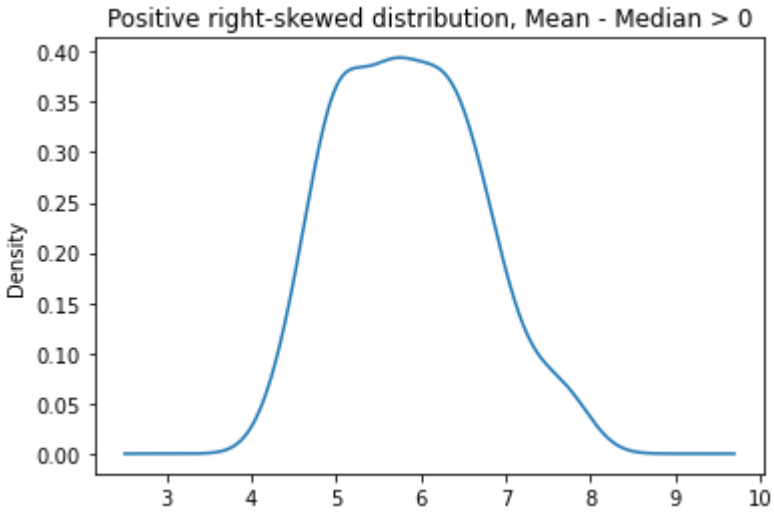
In [39]:
```
df['sepal_length'].hist(bins =10,figsize = (6,4))
```

Out[39]:
```
<AxesSubplot:>
```



In [40]:
```
df['sepal_length'].plot(kind='density', figsize = (6,4), title = 'Positive r
```

Out[40]:
```
<AxesSubplot:title={'center':'Positive right-skewed distribution, Mean - Med
ian > 0'}, ylabel='Density'>
```

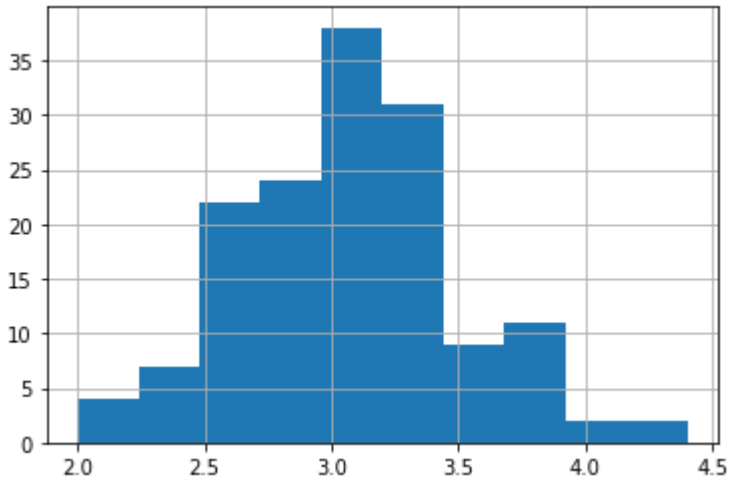


In [41]:
```
df['sepal_width'].skew()
```

Out[41]:
```
0.3340526621720866
```

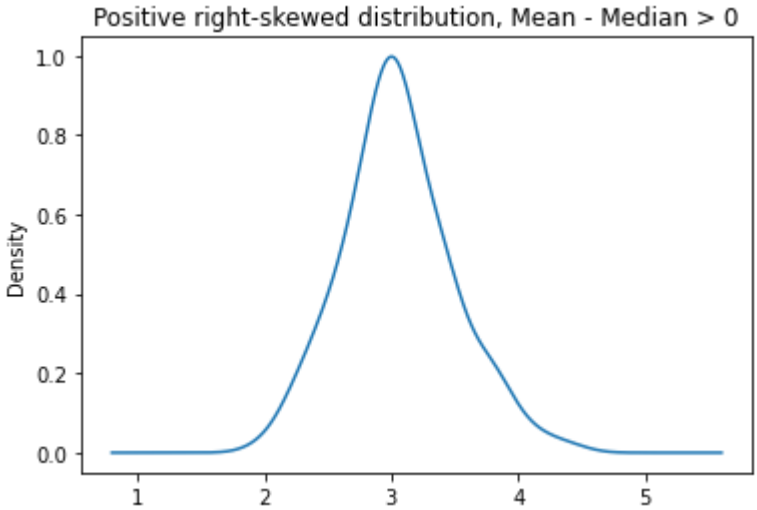
In [42]:
```
df['sepal_width'].hist(bins =10,figsize = (6,4))
```

Out[42]:
```
<AxesSubplot:>
```



In [43]:
```
df['sepal_width'].plot(kind='density', figsize = (6,4), title = 'Positive ri
```

Out[43]:
```
<AxesSubplot:title={'center':'Positive right-skewed distribution, Mean - Med
ian > 0'}, ylabel='Density'>
```

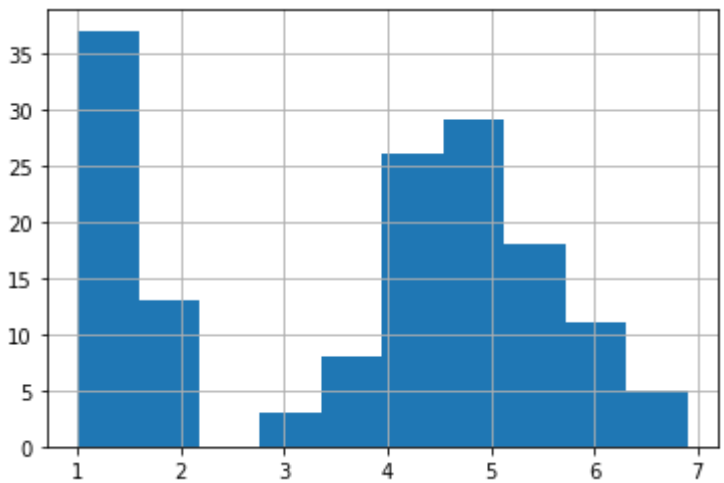


In [44]:
```
df['petal_length'].skew()
```

Out[44]:
```
-0.27446425247378287
```

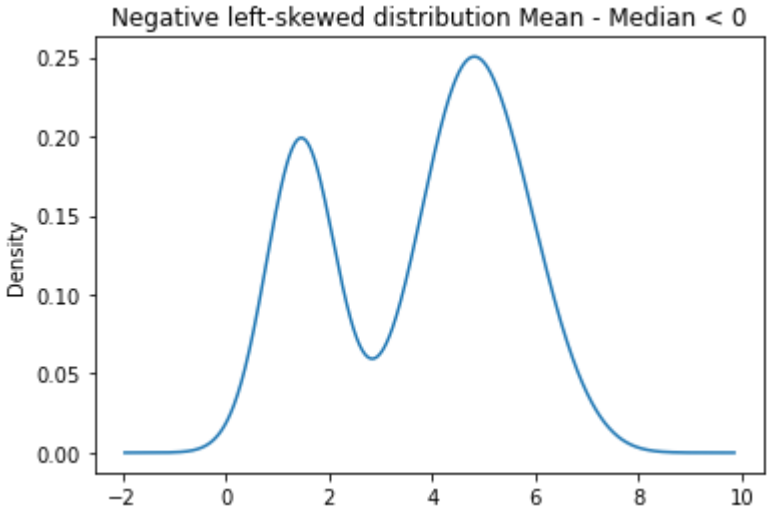
In [45]:
```
df['petal_length'].hist(bins =10,figsize = (6,4))
```

Out[45]:
```
<AxesSubplot:>
```



In [46]:
```
df['petal_length'].plot(kind='density', figsize = (6,4), title = 'Negative l
```

Out[46]: <AxesSubplot:title={'center':'Negative left-skewed distribution Mean - Median < 0'}, ylabel='Density'>

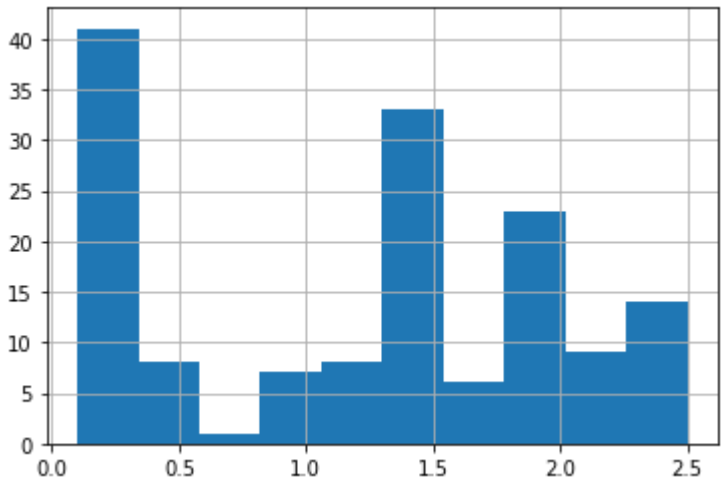


In [47]:
```
df['petal_width'].skew()
```

Out[47]: -0.10499656214412734

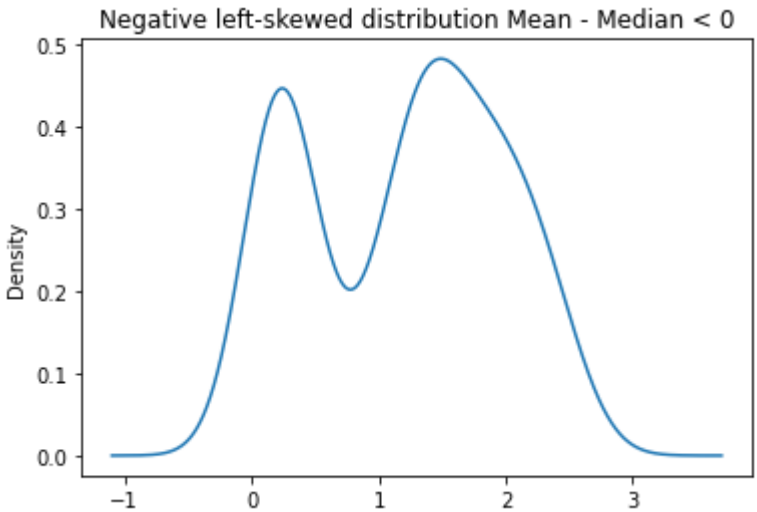
In [48]:
```
df['petal_width'].hist(bins =10,figsize = (6,4))
```

Out[48]: <AxesSubplot:>



In [49]:
```
df['petal_width'].plot(kind='density', figsize = (6,4), title = 'Negative le
```

Out[49]: <AxesSubplot:title={'center':'Negative left-skewed distribution Mean - Median < 0'}, ylabel='Density'>



KURTOSIS

In [50]:
```
df['sepal_length'].kurt()
```

Out[50]:
```
-0.5520640413156395
```

In [51]:
```
df['sepal_width'].kurt()
```

Out[51]:
```
0.2907810623654279
```

In [52]:
```
df['petal_length'].kurt()
```

Out[52]:
```
-1.4019208006454036
```

In [53]:
```
df['petal_width'].kurt()
```

Out[53]:
```
-1.3397541711393433
```

Correlation

CORRELATION COEFFICIENTS

In [54]:
```
# Pearson's r
df.corr()
```

Out[54]:

	sepal_length	sepal_width	petal_length	petal_width
sepal_length	1.000000	-0.109369	0.871754	0.817954
sepal_width	-0.109369	1.000000	-0.420516	-0.356544
petal_length	0.871754	-0.420516	1.000000	0.962757
petal_width	0.817954	-0.356544	0.962757	1.000000

In [55]:
```
# Spearman's rho
df.corr(method='spearman')
```

Out[55]:

	sepal_length	sepal_width	petal_length	petal_width
sepal_length	1.000000	-0.159457	0.881386	0.834421
sepal_width	-0.159457	1.000000	-0.303421	-0.277511
petal_length	0.881386	-0.303421	1.000000	0.936003
petal_width	0.834421	-0.277511	0.936003	1.000000

In [56]:
```
# Kendall's tau
df.corr(method='kendall')
```

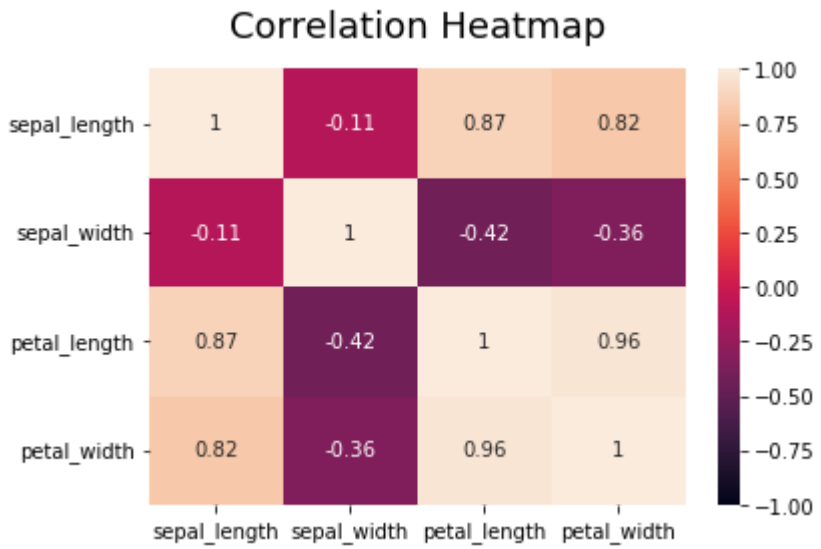
Out[56]:

	sepal_length	sepal_width	petal_length	petal_width
sepal_length	1.000000	-0.072112	0.717624	0.654960
sepal_width	-0.072112	1.000000	-0.182391	-0.146988
petal_length	0.717624	-0.182391	1.000000	0.803014
petal_width	0.654960	-0.146988	0.803014	1.000000

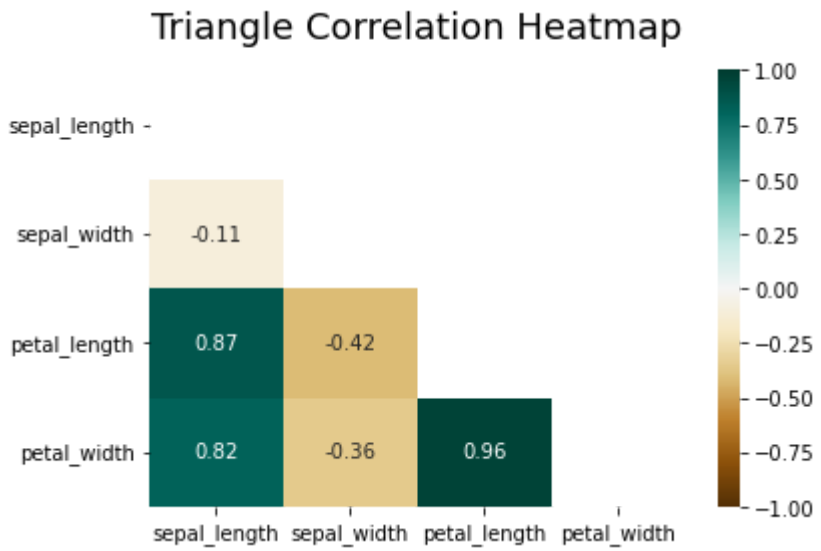
CORRELATION MATRIX

In [57]:
```
heatmap = sns.heatmap(df.corr(), annot=True, vmin=-1, vmax=1)
heatmap.set_title('Correlation Heatmap',fontdict={'fontsize':18}, pad=16)
```

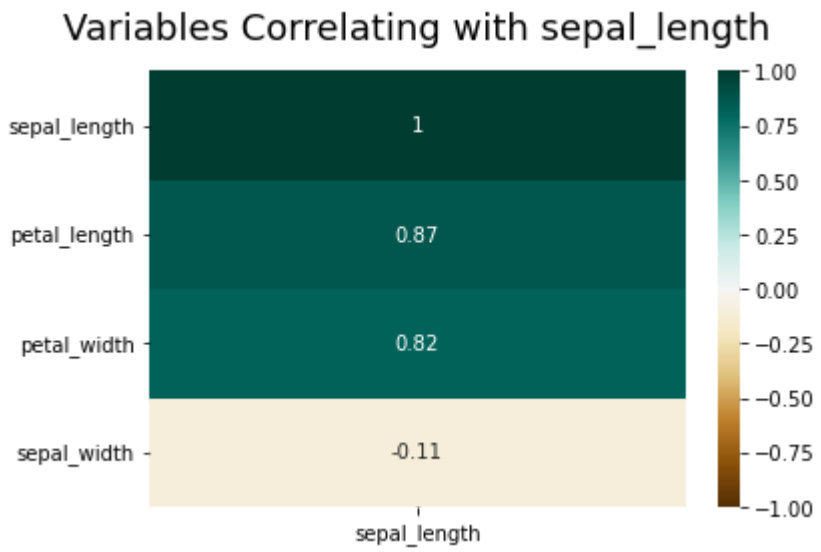
Out[57]:
```
Text(0.5, 1.0, 'Correlation Heatmap')
```



In [58]:
```
# define the mask to set the values in the upper triangle to True
mask = np.triu(np.ones_like(df.corr(), dtype=bool))
heatmap = sns.heatmap(df.corr(), mask=mask, vmin=-1, vmax=1, annot=True, cma
heatmap.set_title('Triangle Correlation Heatmap', fontdict={'fontsize':18},
```

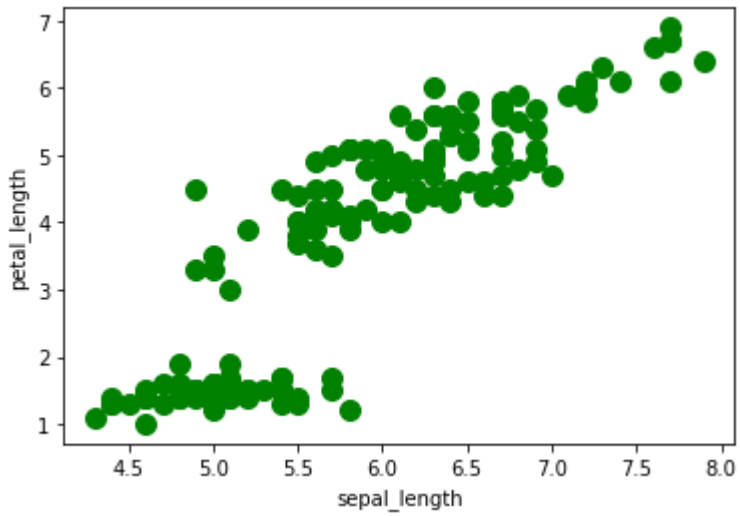


In [59]:
```
heatmap = sns.heatmap(df.corr()[['sepal_length']].sort_values(by='sepal_leng
                      annot=True, cmap='BrBG')
heatmap.set_title('Variables Correlating with sepal_length', fontdict={'font
```

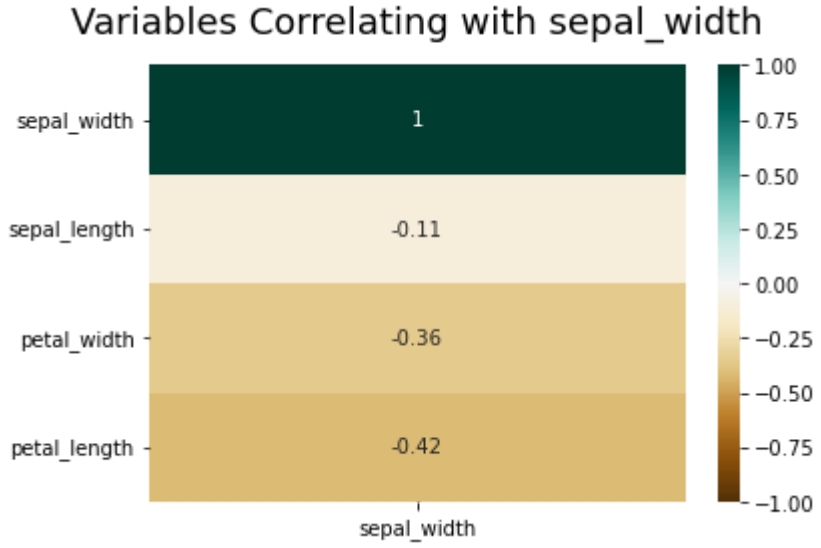


In [60]:
```
df.plot.scatter(x = 'sepal_length', y = 'petal_length', s = 100, color='gree
```

Out[60]:
```
<AxesSubplot:xlabel='sepal_length', ylabel='petal_length'>
```

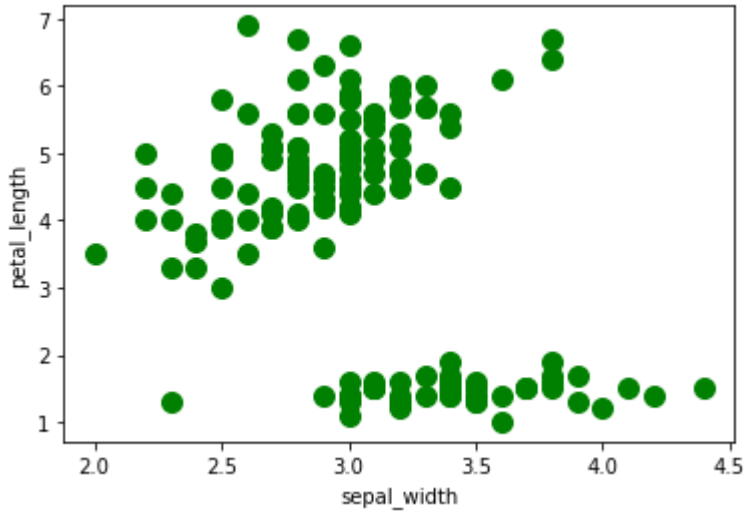


In [61]:
```
heatmap = sns.heatmap(df.corr()[['sepal_width']].sort_values(by='sepal_width
                      annot=True, cmap='BrBG')
heatmap.set_title('Variables Correlating with sepal_width', fontdict={'fonts
```

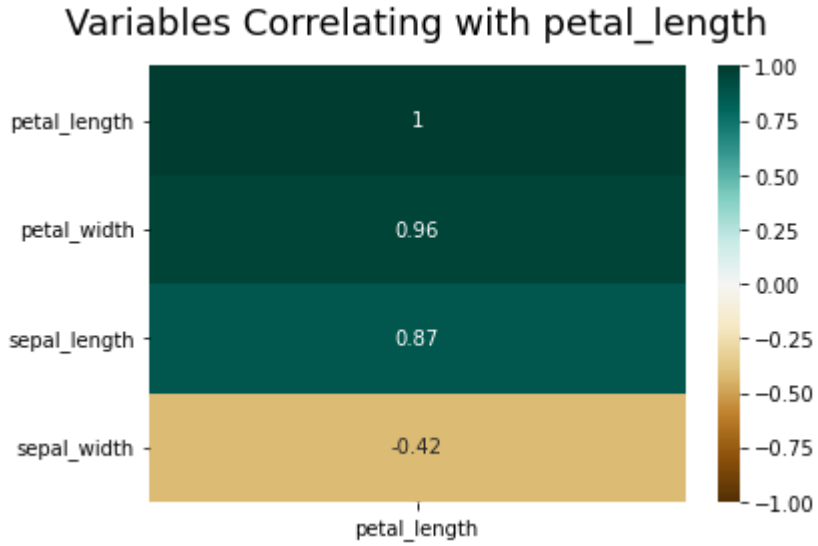


In [62]:
```
df.plot.scatter(x = 'sepal_width', y = 'petal_length', s = 100, color='green
```

Out[62]: <AxesSubplot:xlabel='sepal_width', ylabel='petal_length'>

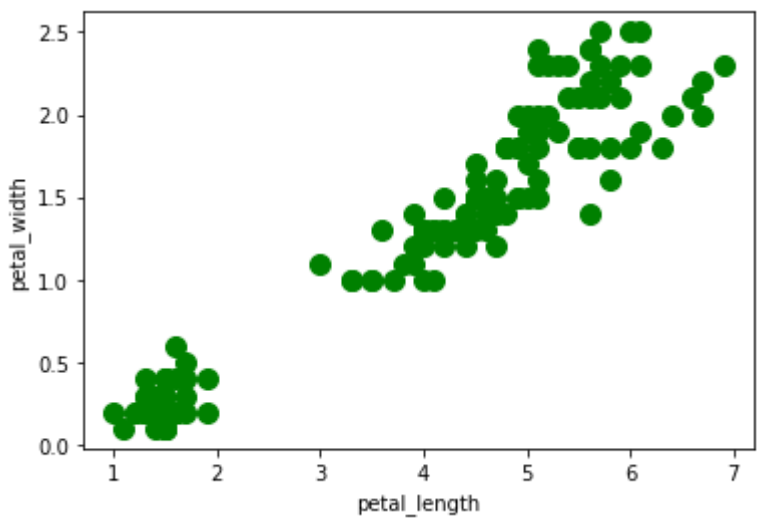


In [63]:
```
heatmap = sns.heatmap(df.corr()[['petal_length']].sort_values(by='petal_leng
                      annot=True, cmap='BrBG')
heatmap.set_title('Variables Correlating with petal_length', fontdict={'font
```



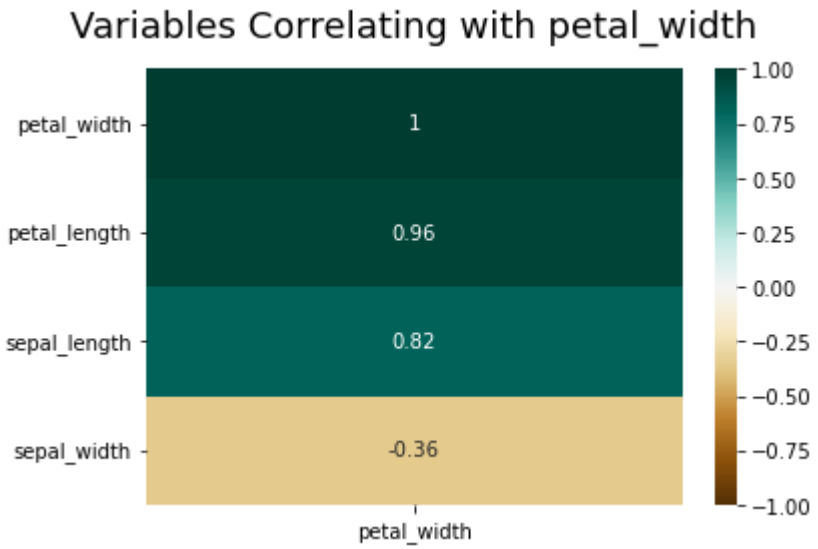
In [64]:
```
df.plot.scatter(x = 'petal_length', y = 'petal_width', s = 100, color='green
```

Out[64]: <AxesSubplot:xlabel='petal_length', ylabel='petal_width'>



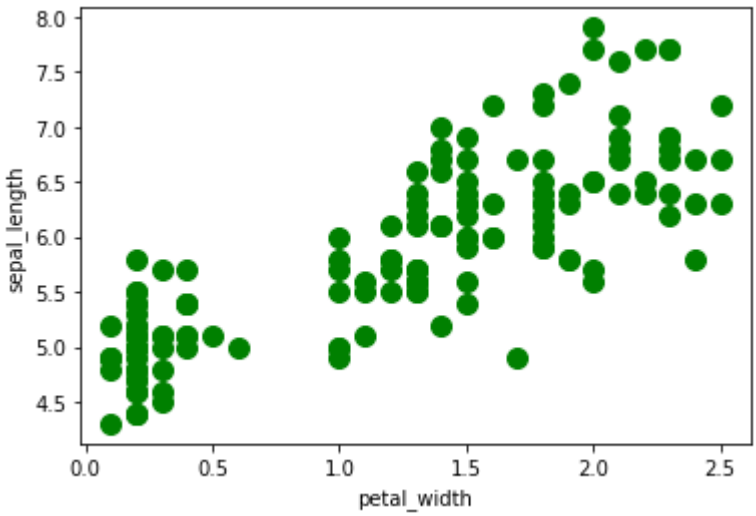
In [65]:
```
heatmap = sns.heatmap(df.corr()[['petal_width']].sort_values(by='petal_width
                      annot=True, cmap='BrBG')
```

```
heatmap.set_title('Variables Correlating with petal_width', fontdict={'fonts
```



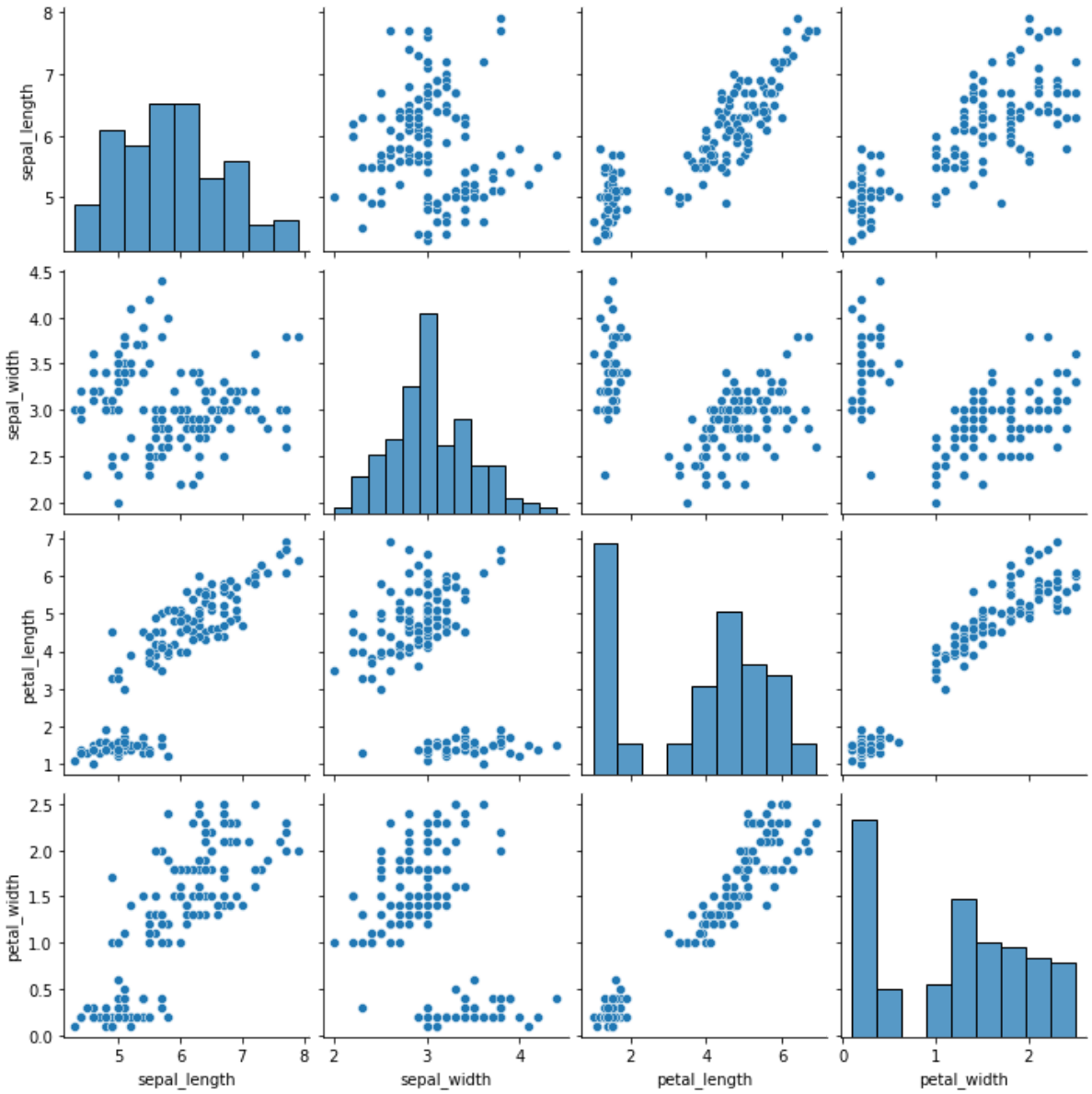
In [66]:
```
df.plot.scatter(x = 'petal_width', y = 'sepal_length', s = 100, color='green
```

Out[66]:
```
<AxesSubplot:xlabel='petal_width', ylabel='sepal_length'>
```



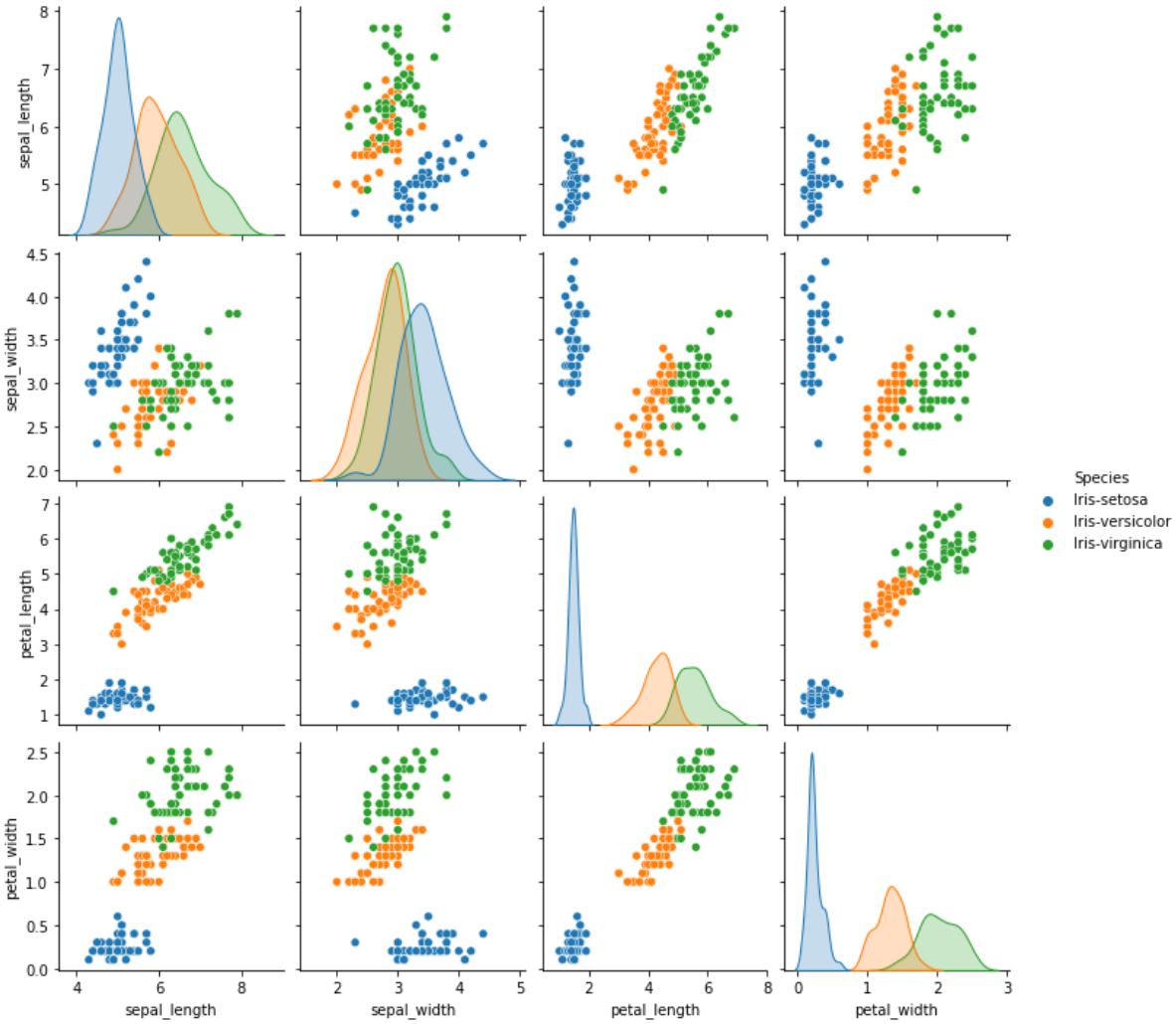
In [67]:
```
sns.pairplot(df)
```

Out[67]:
```
<seaborn.axisgrid.PairGrid at 0x7ff31bd0e7f0>
```



In [68]:
```
sns.pairplot(df, hue='Species')
```

Out[68]:
```
<seaborn.axisgrid.PairGrid at 0x7ff339651850>
```


sepal_length
sepal_width
petal_length
petal_width
Species
Iris-setosa
Iris-versicolor
Iris-virginica