

DATA EXPLORATION USING SEABORN

► In [1]:

```
1 import pandas as pd
2 import seaborn as sns
3 import matplotlib.pyplot as plt
4 import numpy as np
5
6 #Import the iris data set
7 from sklearn.datasets import load_iris
8 iris=load_iris()
9 # Assigning iris data set into data frame using features as column names
10 data=pd.DataFrame(iris.data,columns=iris.feature_names)
11 # creating label data frame with the coulms as Species types
12 label=pd.DataFrame(list(map(lambda x : iris.target_names[x],iris.target)),columns=['Species'])
13 iris=pd.concat([data,label],axis=1)
14 # Printin top 5 records of the iris data set
15 print(iris.head())
```

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	\
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	
4	5.0	3.6	1.4	0.2	

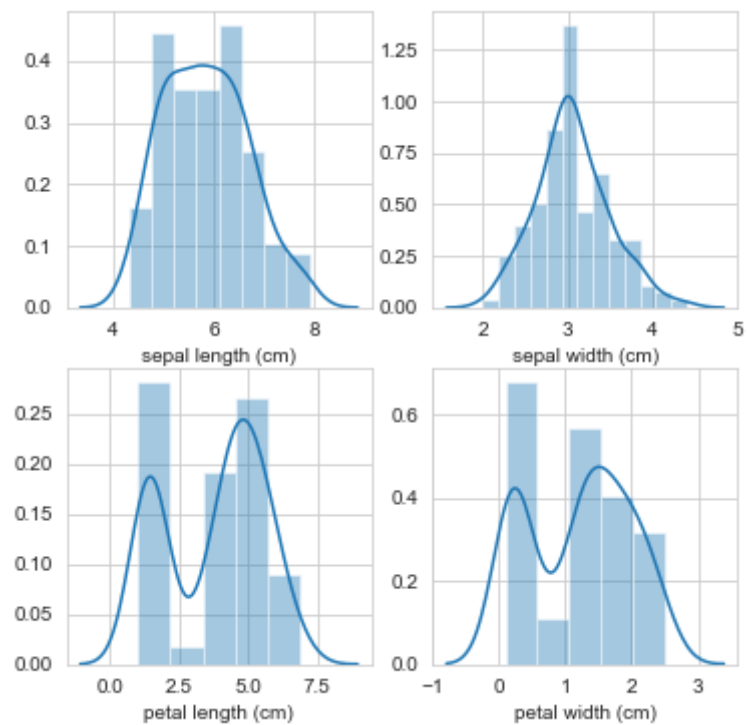
	Species
0	setosa
1	setosa
2	setosa
3	setosa
4	setosa

In [2]:

```
1  """
2  1. Use distplot() to see the diistribution of the Sepal length(cm), Sepal Width (cm), Petal length (cm) and petal width
3  features. Plot them as subplots in single image
4
5  """
6  # Settin the style as white grid as background
7  sns.set_style("whitegrid");
8  # making fig size as 6X6 inches with 2 rows and 2 columns
9  fig, axs = plt.subplots(figsize=(6,6), ncols=2, nrows=2)
10 # displaying the graphs in each quadrant from 0,0 to 1,1
11 sns.distplot(iris['sepal length (cm)'],ax=axs[0, 0])
12 sns.distplot(iris['sepal width (cm)'], ax=axs[0, 1])
13 sns.distplot(iris['petal length (cm)'], ax=axs[1, 0])
14 sns.distplot(iris['petal width (cm)'], ax=axs[1, 1])
15 # displaying the Plot
16 plt.show();
```

C:\Users\admin\Anaconda3\lib\site-packages\scipy\stats\stats.py:1713: FutureWarning: Using a non-tuple sequence for multidimensional indexing is deprecated; use `arr[tuple(seq)]` instead of `arr[seq]`. In the future this will be interpreted as an array index, `arr[np.array(seq)]`, which will result either in an error or a different result.

```
return np.add.reduce(sorted[indexer] * weights, axis=axis) / sumval
```

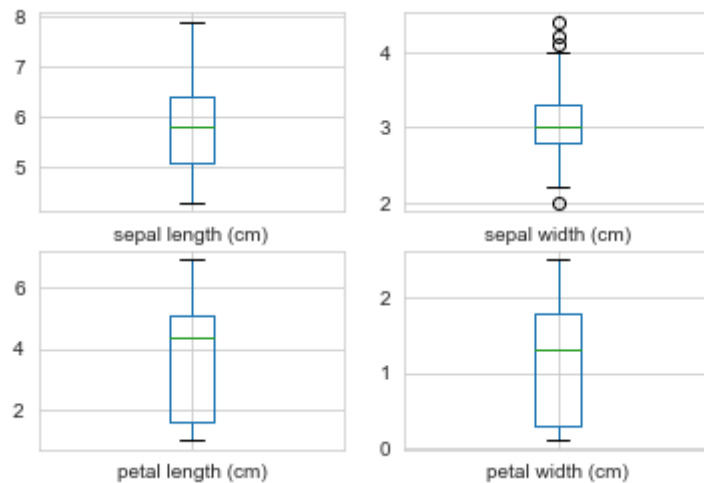


In [3]:

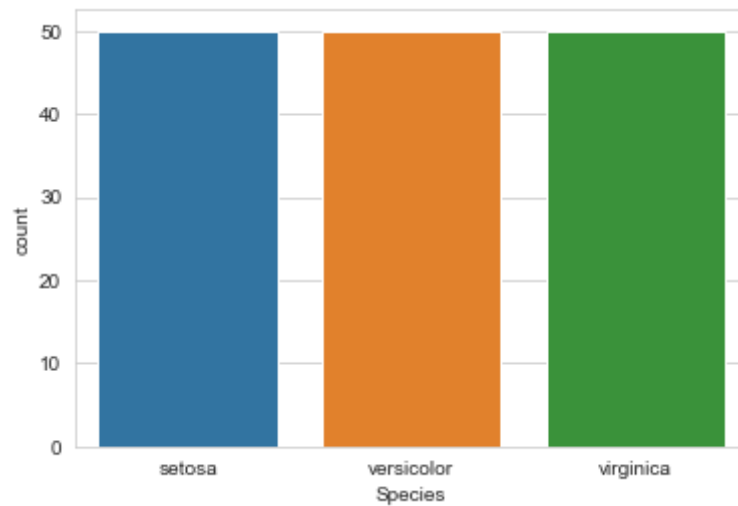
```

1  """
2  2. Do a box plot for all the features except 'Species'
3
4  """
5  # making figure and axes as 2 rows and 2 columns
6  fig, axes = plt.subplots(2,2)
7  # displaying plot for all the columns except Species which is last column(-1)
8  for i,el in enumerate(list(iris.columns.values)[: -1]):
9      a = iris.boxplot(el, ax=axes.flatten()[i])
10 # displaying the Plot
11 plt.show()

```



```
In [4]: 1 """  
2 3. Do a countplot for feature Species  
3 """  
4 # making count plot with Species column values  
5 sns.countplot(x='Species',data=iris)  
6 # displaying the Plot  
7 plt.show()
```



In [5]:

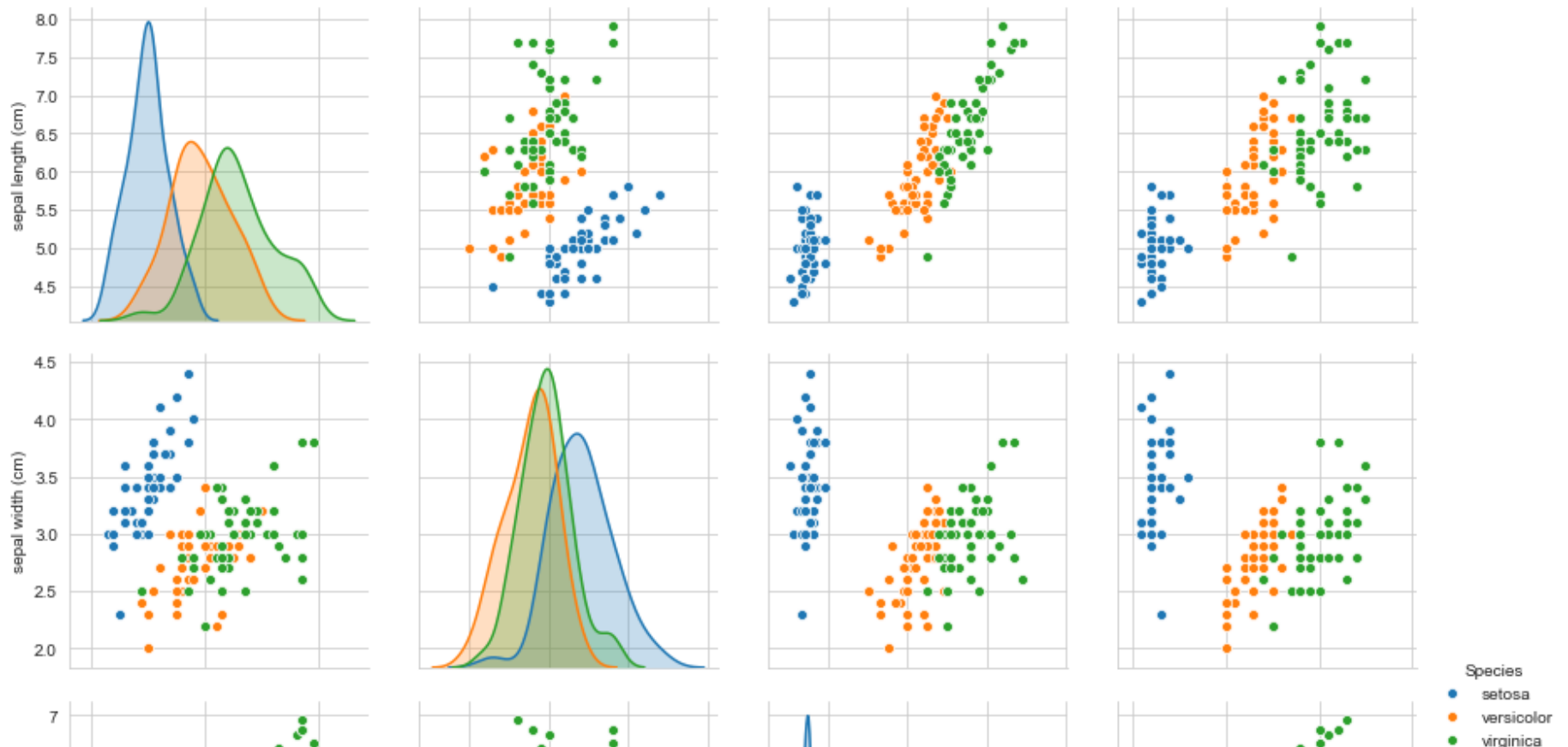
```

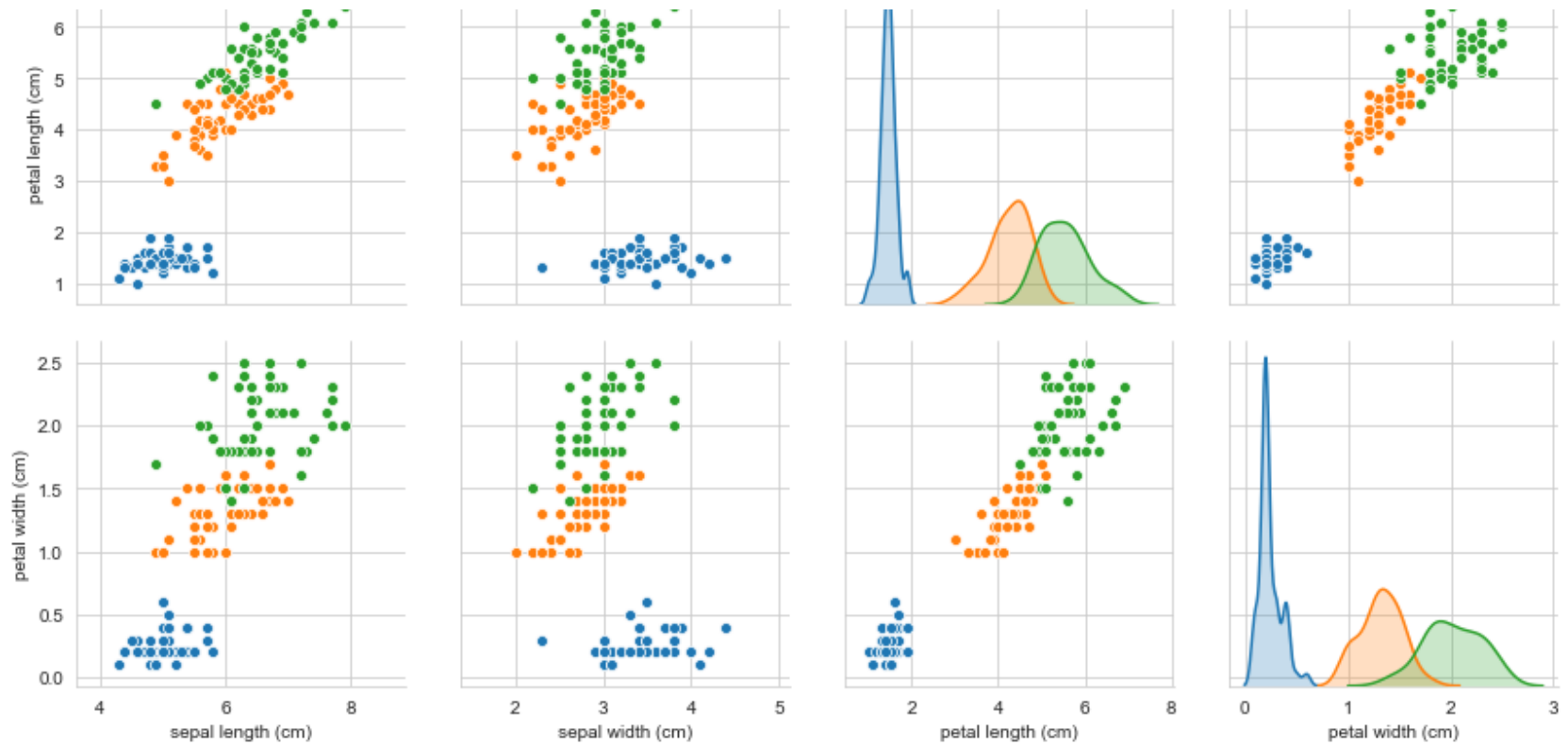
1  """
2  4. Do a pair plot on the features Sepal length(cm), Sepal Width (cm), Petal length (cm), petal width(cm) and Species
3  """
4
5  plt.close();
6  sns.set_style("whitegrid");
7  # creating pair plots using Species as hue
8  sns.pairplot(iris, hue="Species", height=3);
9  # displaying the Plot
10 plt.show()

```

C:\Users\admin\Anaconda3\lib\site-packages\scipy\stats\stats.py:1713: FutureWarning: Using a non-tuple sequence for multidimensional indexing is deprecated; use `arr[tuple(seq)]` instead of `arr[seq]`. In the future this will be interpreted as an array index, `arr[np.array(seq)]`, which will result either in an error or a different result.

return np.add.reduce(sorted[indexer] * weights, axis=axis) / sumval





In [6]:

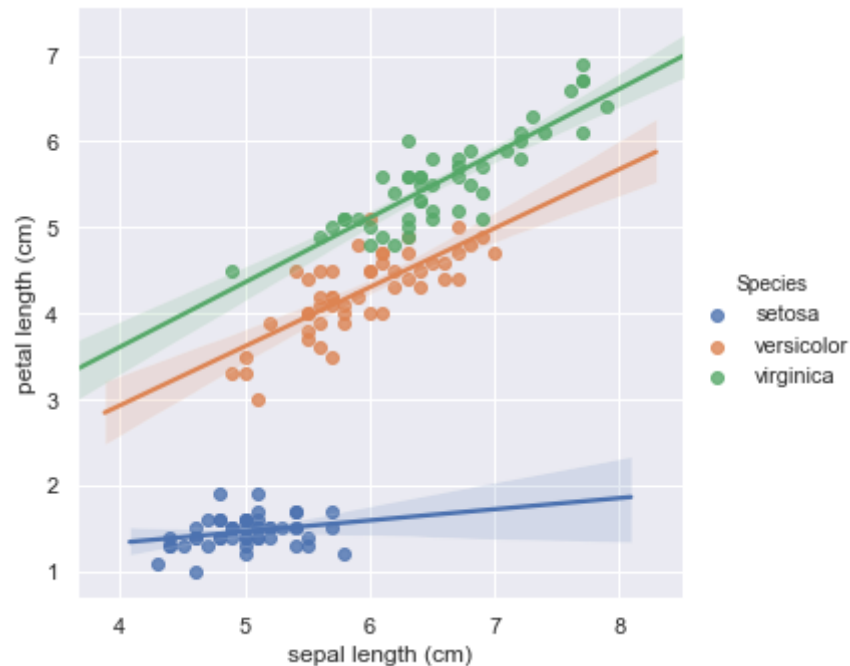
```

1  """
2  5. Do lmplo on the following Sepal length(cm), Petal length (cm). Using hue, display the different speices in
3  different color
4  """
5
6  plt.close();
7  sns.set_style("whitegrid");
8  # making color code as true to disply inbuilt color code
9  sns.set(color_codes=True)
10 # creating implot with sepal length and petal length using Species as hue
11 sns.lmplot(x='sepal length (cm)',y='petal length (cm)',hue='Species',data=iris)
12 # displaying the Plot
13 plt.show()
14

```

C:\Users\admin\Anaconda3\lib\site-packages\scipy\stats\stats.py:1713: FutureWarning: Using a non-tuple sequence for multidimensional indexing is deprecated; use `arr[tuple(seq)]` instead of `arr[seq]`. In the future this will be interpreted as an array index, `arr[np.array(seq)]`, which will result either in an error or a different result.

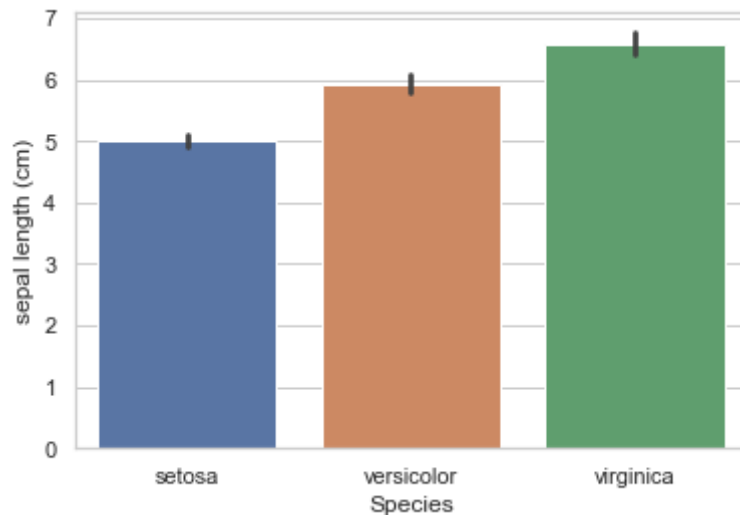
return np.add.reduce(sorted[indexer] * weights, axis=axis) / sumval



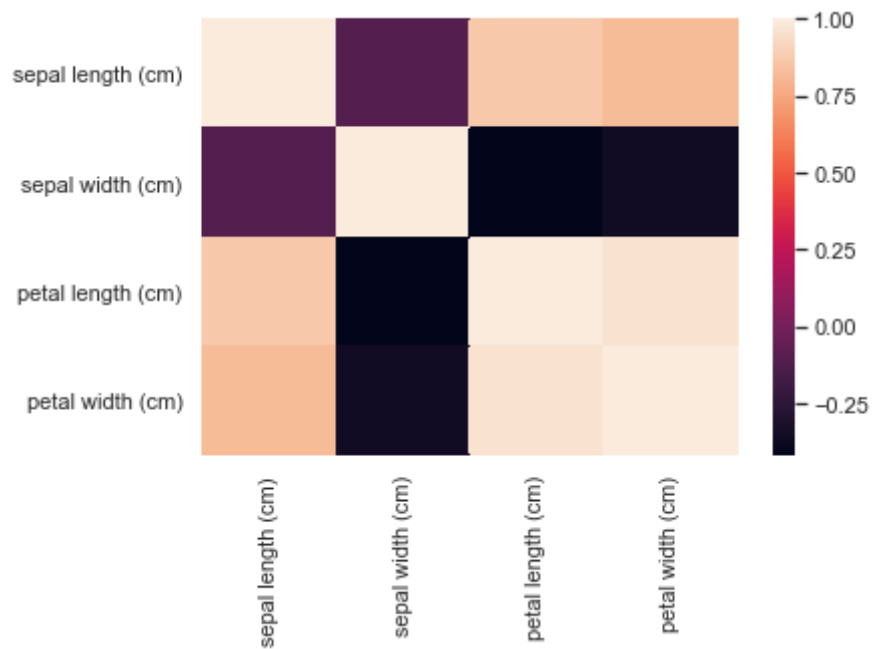

```
In [7]: 1 """
2 6. Do a bar plot of Species vs sepal length (cm)
3 """
4 # closing the existing plotting
5 plt.close();
6 # setting back ground style as white grid
7 sns.set_style("whitegrid");
8 # creating bar plot with species and sepal length
9 sns.barplot(x='Species',y='sepal length (cm)',data=iris)
10 # displaying the Plot
11 plt.show()
```

C:\Users\admin\Anaconda3\lib\site-packages\scipy\stats\stats.py:1713: FutureWarning: Using a non-tuple sequence for multidimensional indexing is deprecated; use `arr[tuple(seq)]` instead of `arr[seq]`. In the future this will be interpreted as an array index, `arr[np.array(seq)]`, which will result either in an error or a different result.

return np.add.reduce(sorted[indexer] * weights, axis=axis) / sumval



```
In [8]: 1 """
2 7. Using heatmap, plot the correlation matrix
3 """
4 # calculate the correlation matrix
5 corr = iris.corr()
6 # creating heatmap correlation column values
7 sns.heatmap(corr,xticklabels=corr.columns,yticklabels=corr.columns)
8 # displaying the Plot
9 plt.show()
```



```
In [ ]: 1
```