CALCULATING CORRELATION COEFFICIENT, MEAN, VARIANCE, SKEW AND KURTOSIS :

Importing the Libraries

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

Loading the dataset obtained by data cleaning:

In [2]:
```
data = pd.read_csv('iris_dataset_cleaned.csv')
```

Displaying properties of the cleaned dataset for further processing

In [3]:
```
data.info()
data.head()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 93 entries, 0 to 92
Data columns (total 5 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   sepal_length  93 non-null     float64
 1   sepal_width   93 non-null     float64
 2   petal_length  93 non-null     float64
 3   petal_width   93 non-null     float64
 4   species       93 non-null     object 
dtypes: float64(4), object(1)
memory usage: 3.8+ KB
```

Out[3]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.045070	2.508203	3.018024	1.164924	Iris-versicolor
1	6.325517	2.115481	4.542052	1.413651	Iris-versicolor
2	5.257497	3.814303	1.470660	0.395348	Iris-setosa
3	6.675168	3.201700	5.785461	2.362764	Iris-virginica
4	5.595237	2.678166	4.077750	1.369266	Iris-versicolor

In [4]:
```
data.tail()
```
Out[4]:

	sepal_length	sepal_width	petal_length	petal_width	species
88	4.874848	3.217348	1.592887	0.123588	Iris-setosa
89	5.564197	2.771731	3.483588	1.074754	Iris-versicolor
90	5.548047	4.249211	1.453466	0.214527	Iris-setosa
91	5.510482	2.652867	4.276817	1.298032	Iris-versicolor
92	4.538713	3.056142	1.545136	0.241424	Iris-setosa

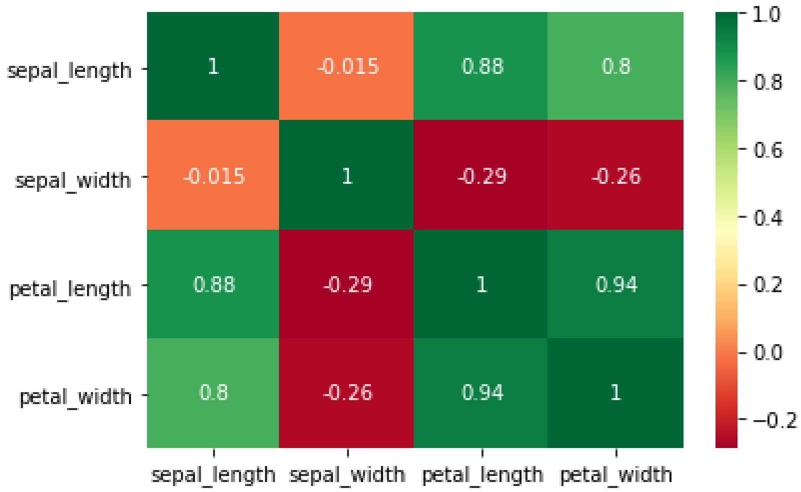
In [5]:
```
data.describe()
```
Out[5]:

	sepal_length	sepal_width	petal_length	petal_width
count	93.000000	93.000000	93.000000	93.000000
mean	5.867894	3.054935	3.808118	1.236858
std	0.892271	0.439463	1.811399	0.770872
min	4.344007	1.946010	1.033031	0.020731
25%	5.152435	2.794790	1.541564	0.343669
50%	5.636744	3.049459	4.192791	1.369266
75%	6.478961	3.239682	5.098860	1.837925
max	7.795561	4.249211	6.768611	2.603123

In [6]:
```
data.shape
```
Out[6]: (93, 5)

In [7]:
```
#correlation coefficient
correl = data.corr()
```

In [8]:
```python
sns.heatmap(correl, annot=True, vmax=1, cmap="RdYlGn")
plt.show()
```



From the above heatmap, we could observe the correlation coeficients between various features of the dataset. In our case, the colour Red shows a strong negative correlation between the features where as the colour Green indicates a strng positive correlation.

The features **"sepal_width"** and **"sepal_length"** have a correlation coefficient of **"-0.015"** which implies that they have a **weak negative** correlation between them.

The pairs of features **"petal_length"** & **"petal_width"** and **"sepal_length"** & **"petal_length"** have a correlation coefficient of **"0.96"** and **"0.88"** respectively which implies that the pair have a **strong positive** correlation between themselves.

In [9]:
```python
#mean of the dataset
mn = np.mean(data)
mn
```

Out[9]:
```
sepal_length    5.867894
sepal_width     3.054935
petal_length    3.808118
petal_width     1.236858
dtype: float64
```

In [10]:
```python
#variance(std^2) of the dataset
stddf = np.std(data)**2
stddf
```

Out[10]:
```
sepal_length    0.787586
sepal_width     0.191052
petal_length    3.245885
petal_width     0.587854
dtype: float64
```

In [11]:
```python
#skew of the dataset
from scipy.stats import skew, kurtosis
skew(data[['sepal_length','sepal_width','petal_length','petal_width']])
```

Out[11]:
```
array([ 0.41780074,  0.18375397, -0.23438962, -0.11727924])
```

Skewness is the measure of asymmetry of the normal data distrubtion. If the tail of the distrubition is long towards the left side, then the distribution is negatively skewed and if its towards the right side, then the data distribution is said to be positively skewed.

From the above values of skewness for various features, we can infer that:

"sepal_length" and **"sepal width"** has a positive skew of **0.417** and **0.183** respectively and the features **"petal_width"** and **"petal_length"** has a negative skew of **-0.117** and **-0.234** respectively.

As the skewness of all the features lies with the range **(-0.5,0.5)**, it can be said that the distribution of data points for all the above mentioned features is **approximately symmetric**.

In [12]:
```python
#kurtosis of the dataset
kr = kurtosis(data[['sepal_length','sepal_width','petal_length','petal_width']])
kr
```

Out[12]:
```
array([-0.70569072,  0.24582897, -1.40114775, -1.25662545])
```

Kurtosis is the measure of whether the data distrubtion is light tailed or heavy tailed in comparison with the normal data distribution. The kurtosis for normal data distrubtion has a value of 3.

The kurtosis is highly tailed if the value is > 3. The kurtosis is low tailed if the value is < 3.

From the kurtosis values for various features in our data:

As all the values of kurtosis are in between -3 to +3,we can deduce that the data is not having high number of outliers as well as a very low number of them.

Hence we are supposed to remove those outliers in the question CM3 for improving the efficiency of our classiication model.