Assignment 4.2: Data Visualization

Gabi Rivera || ADS501-1 || 29Sep2022

In [1]:
```
import os
os.getcwd()
```

Out[1]:
```
'/Users/gabirivera/Desktop/MSADS2/ADS-501-01/Module 4/Coding'
```

In [48]:
```
import pandas as pd
import numpy as np
```

In [5]:
```
iris = pd.read_csv('Iris.csv', sep = ',')
iris.head(5)
```

Out[5]:

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa

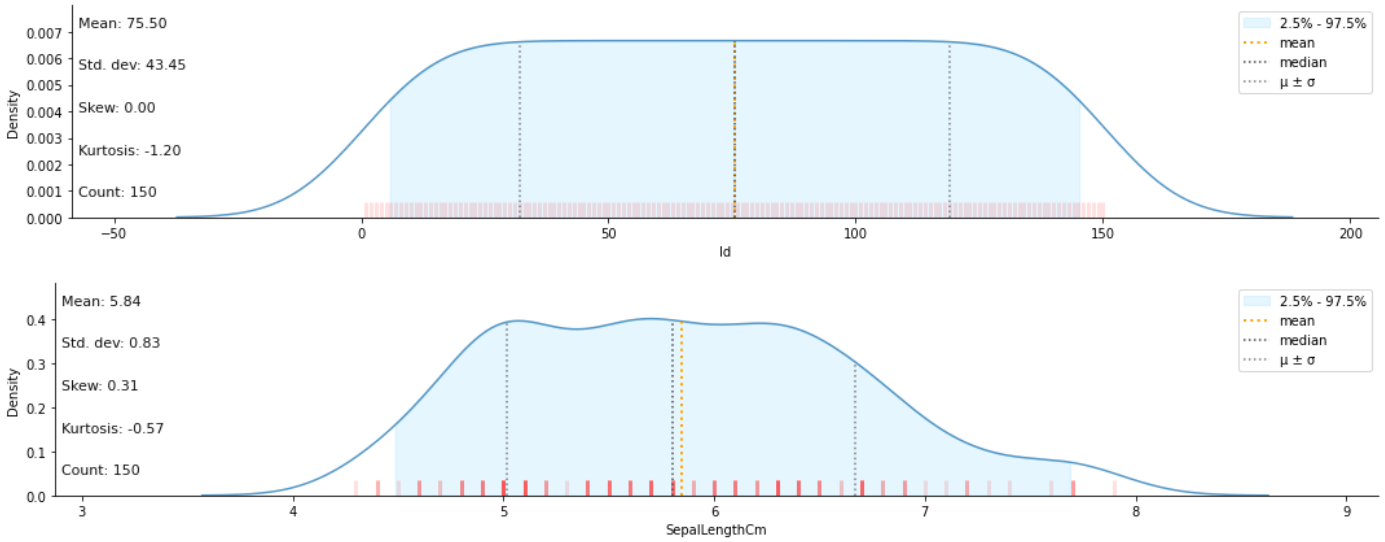
In [6]:
```
import klib
```

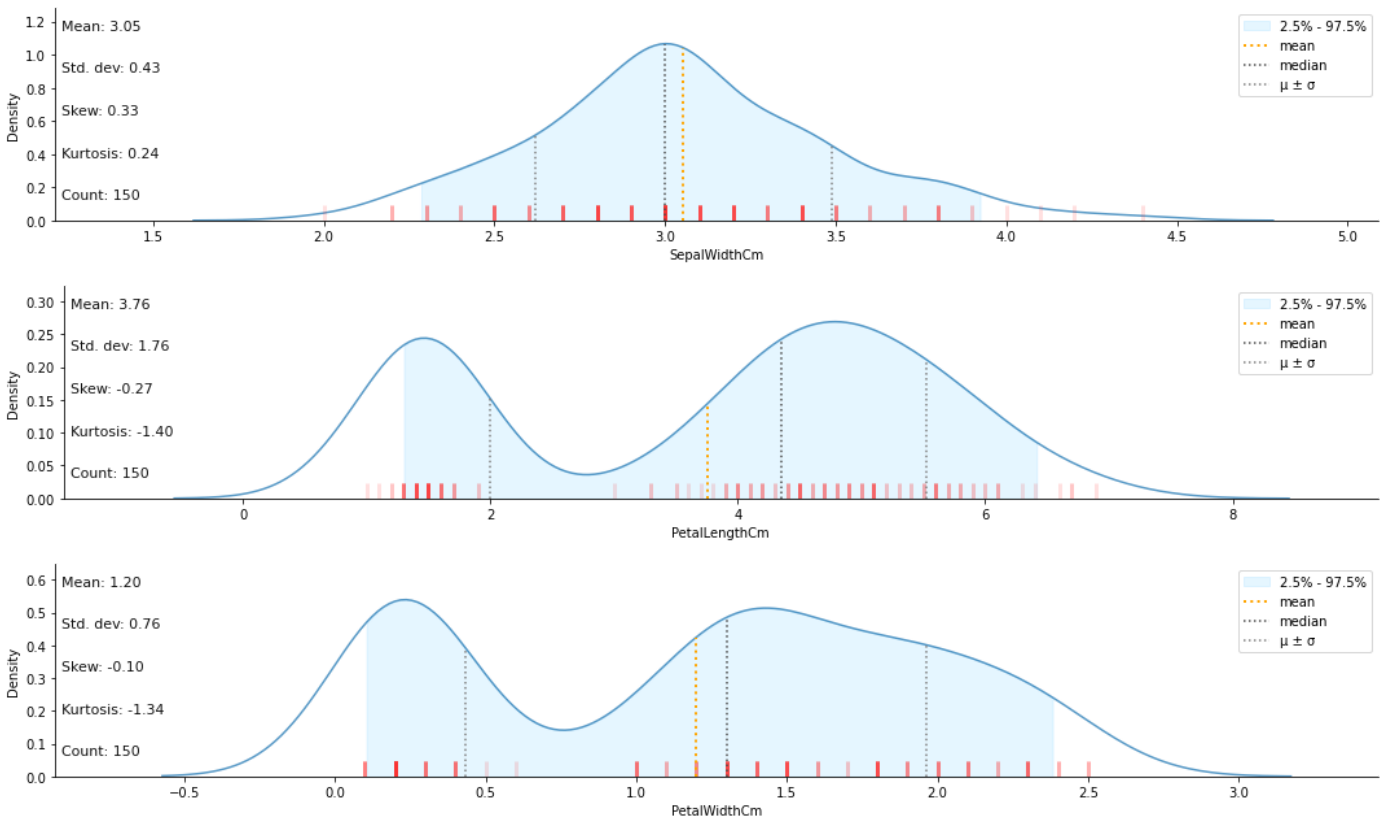
In [8]:
```
klib.missingval_plot(iris)
```

```
No missing values found in the dataset.
```

In [9]:
```
klib.dist_plot(iris)
```

Out[9]:
```
<AxesSubplot:xlabel='PetalWidthCm', ylabel='Density'>
```





In [30]:
```
iris.describe()
```

Out[30]:

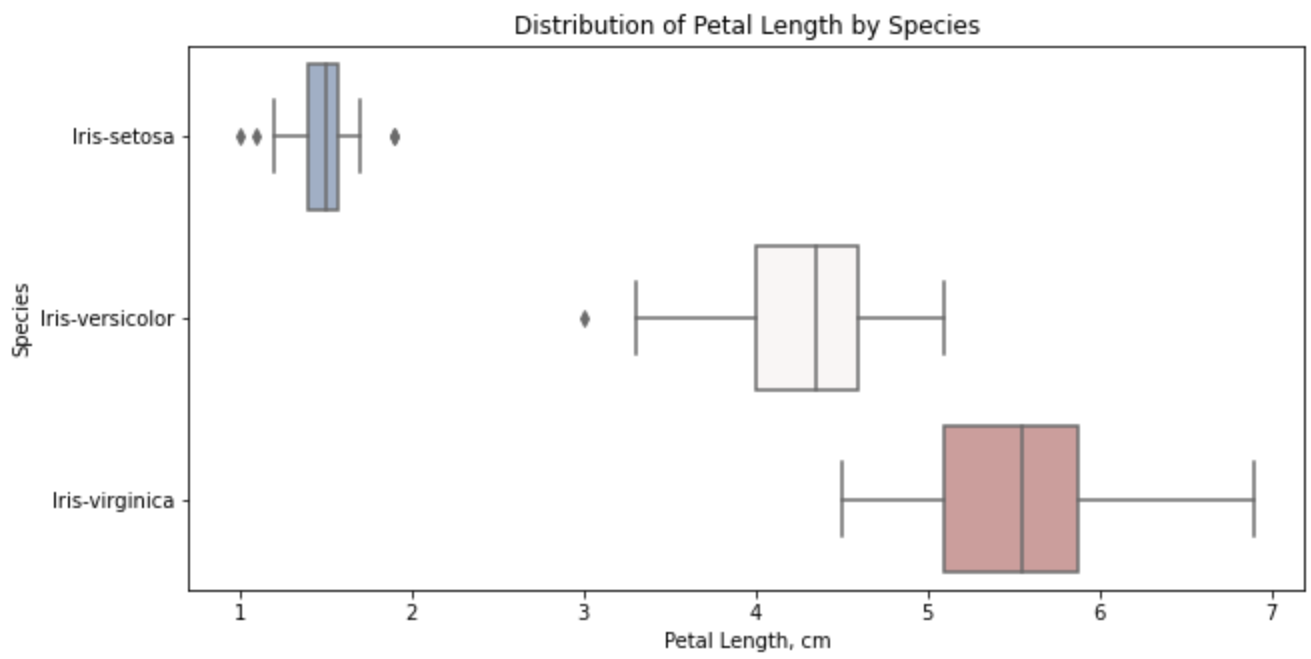
	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm
count	150.000000	150.000000	150.000000	150.000000	150.000000
mean	75.500000	5.843333	3.054000	3.758667	1.198667
std	43.445368	0.828066	0.433594	1.764420	0.763161
min	1.000000	4.300000	2.000000	1.000000	0.100000
25%	38.250000	5.100000	2.800000	1.600000	0.300000
50%	75.500000	5.800000	3.000000	4.350000	1.300000
75%	112.750000	6.400000	3.300000	5.100000	1.800000
max	150.000000	7.900000	4.400000	6.900000	2.500000

Distribution of the various petal lengths or sepal lengths of each species.

In [11]:
```
import seaborn as sns
import matplotlib.pyplot as plt
```

In [45]:
```
sns.boxplot(data=iris, x="PetalLengthCm", y="Species",
            palette = "vlag").set(title='Distribution of Petal Length by Species',
                                  xlabel='Petal Length, cm',
                                  ylabel='Species')

plt.gcf().set_size_inches(10,5)
```



In [75]:
```
iris_species = iris[['PetalLengthCm', 'Species']].groupby('Species')
iris_species.describe().T
```

Out[75]:

	Species	Iris-setosa	Iris-versicolor	Iris-virginica
PetalLengthCm	**count**	50.000000	50.000000	50.000000
	mean	1.464000	4.260000	5.552000
	std	0.173511	0.469911	0.551895
	min	1.000000	3.000000	4.500000
	25%	1.400000	4.000000	5.100000
	50%	1.500000	4.350000	5.550000
	75%	1.575000	4.600000	5.875000
	max	1.900000	5.100000	6.900000

Interpretation: For the distribution, setosa and versicolor seems to be slightly negatively skewed to the left for petal length. Setosa and versicolor's median of 1.5cm & 4.35cm are slightly higher than their mean of 1.464cm & 4.26cm, respectively. Virginica has a slight positive skew to the right with median of 5.55cm and mean of 5.552cm. Setosa has the most outliers outside of its minimum and maximum whiskers at 1cm & 1.9cm, respectively, compared to versicolor that has 1 outlier from its minimum whisker of 3cm. It's maximum whisket is at 5.1cm. Virginica species has no outliers with its minmum and maximum at 4.5cm and 6.9cm. Setosa species display the shortest petal length with IQR of 0.175cm followed by versicolor at 0.600cm and virginica at 0.775cm.

From comparing against the closely spaced distribution of sepal length boxplots below, the distribution of petal length setosa species is farther from it's cousins versicolor and virginica. The overall patterns seems to be that setosa has the shortest petal and sepal length followed by versicolor then virginica.

Extra:

In [35]:
```
sns.boxplot(data=iris, x="SepalLengthCm", y="Species",
            palette = "vlag").set(title='Distribution of Sepal Length by Species',
                                  xlabel='Sepal Length, cm',
```

```
                                ylabel='Species')

plt.gcf().set_size_inches(10,5)
```