lab-10

April 13, 2024

Data Visualization - III

Problem Statement

Download the Iris flower dataset or any other dataset into a DataFrame. (e.g.https://archive.ics.uci.edu/ml/datasets/Iris). Scan the dataset and give the inference as: 1. List down the features and their types (e.g., numeric, nominal) available in the dataset. 2. Create a histogram for each feature in the dataset to illustrate the feature distributions. 3. Create a box plot for each feature in the dataset. 4. Compare distributions and identify outliers.

[1]:
```
#imports
from sklearn.datasets import load_iris
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
iris = load_iris()
```

[2]:
```
data = pd.DataFrame(iris.data, columns = iris.feature_names)
data['label'] = iris.target
data.sample(5)
```

[2]:

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	label
147	6.5	3.0	5.2	2.0	2
39	5.1	3.4	1.5	0.2	0
70	5.9	3.2	4.8	1.8	1
13	4.3	3.0	1.1	0.1	0
84	5.4	3.0	4.5	1.5	1

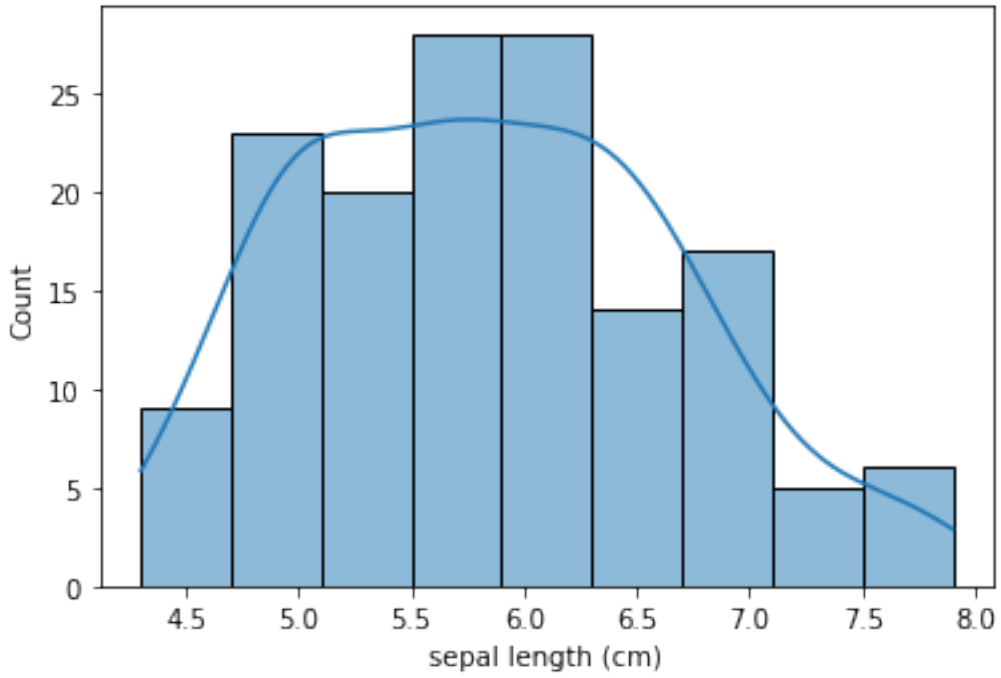
[3]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
```

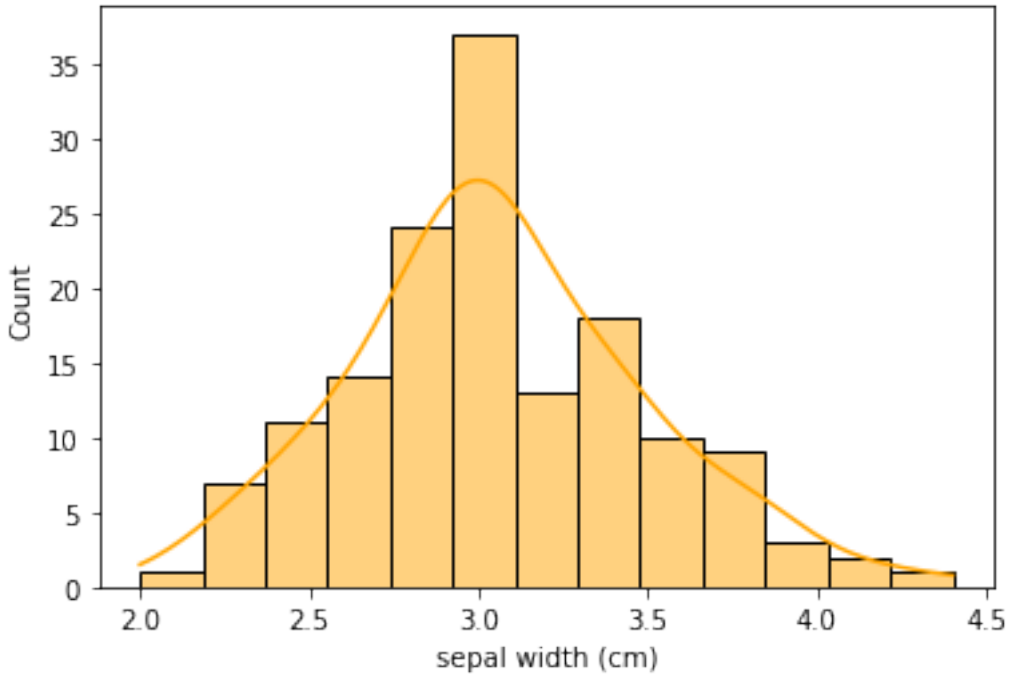
```
RangeIndex: 150 entries, 0 to 149
Data columns (total 5 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   sepal length (cm)  150 non-null    float64
 1   sepal width (cm)   150 non-null    float64
 2   petal length (cm)  150 non-null    float64
 3   petal width (cm)   150 non-null    float64
 4   label              150 non-null    int64  
dtypes: float64(4), int64(1)
memory usage: 6.0 KB
```

0.1 Histograms

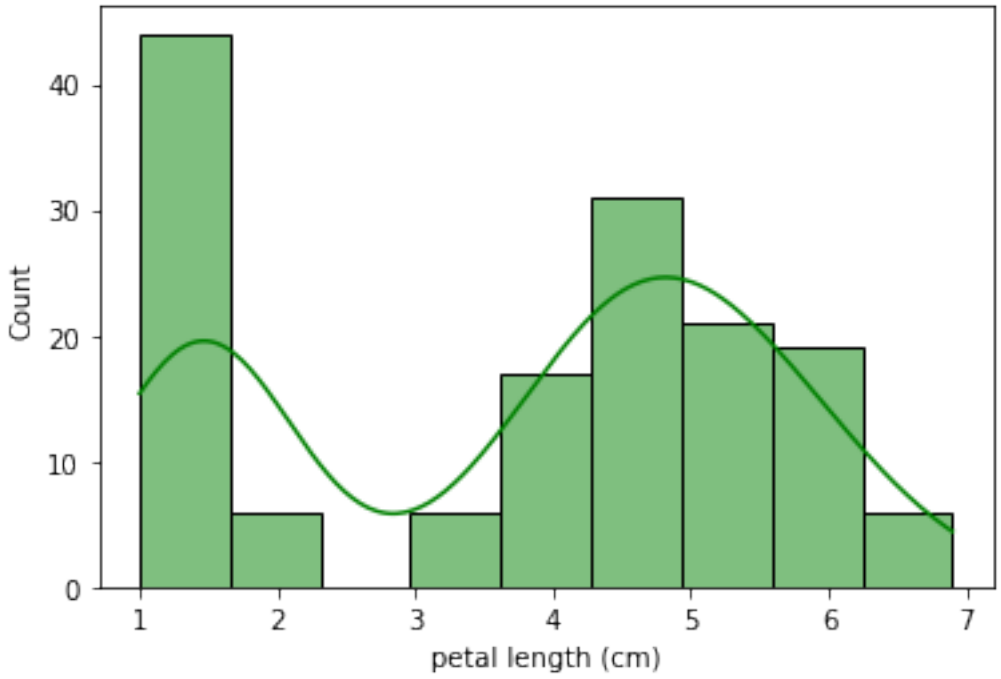
[4]:
```
sns.histplot(data = data, x = 'sepal length (cm)', kde= True);
```



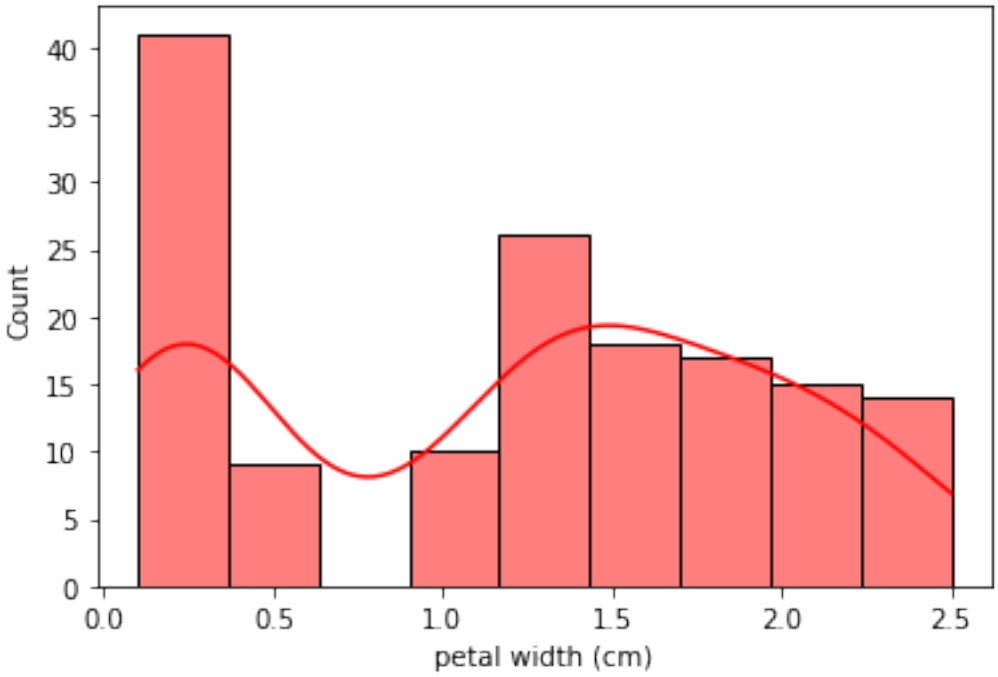
[5]:
```
sns.histplot(data = data, x = 'sepal width (cm)', kde= True, color = "orange");
```



[6]:
```
sns.histplot(data = data, x = 'petal length (cm)', kde= True, color = "green");
```

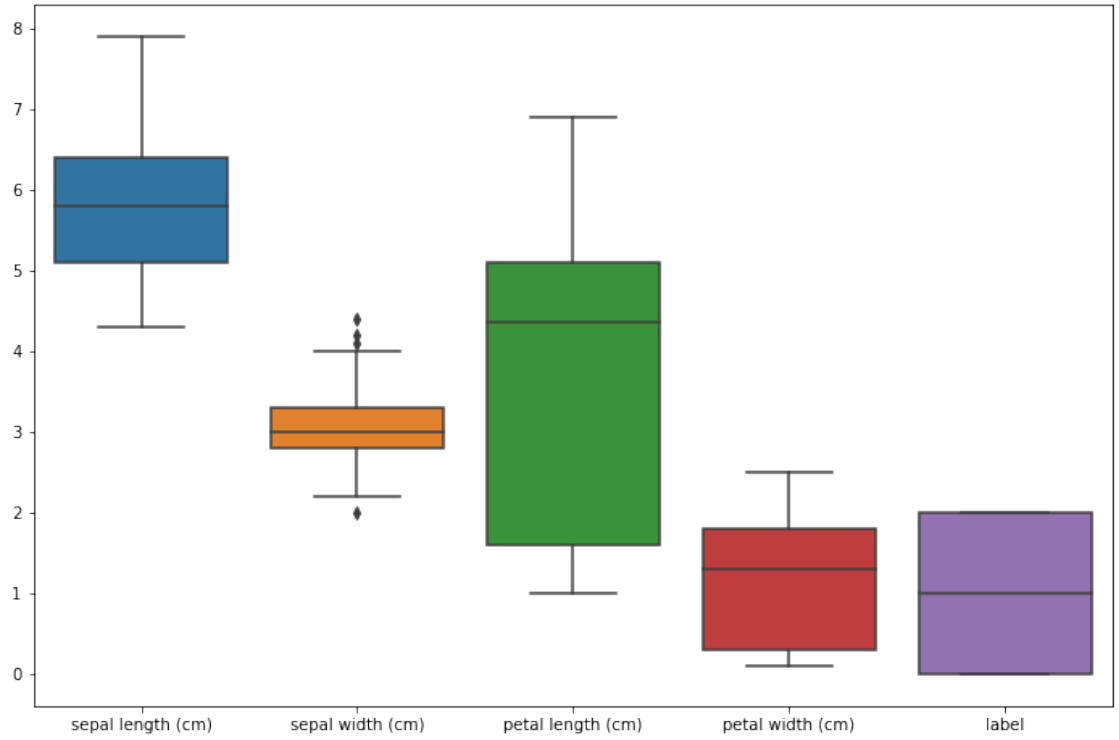


[7]:
```
sns.histplot(data = data, x = 'petal width (cm)', kde= True, color = "red");
```



0.2 Box Plots

[8]:
```
figure = plt.figure(figsize = (12,8))
sns.boxplot(data= data)
plt.show()
```



There are some outliers present in the 'sepal width (cm)' attribute. Lets identify these

[9]:
```
from matplotlib.cbook import boxplot_stats
stats = boxplot_stats(data['sepal width (cm)'])
stats
```

[9]:
```
[{'mean': 3.0573333333333337,
  'iqr': 0.5,
  'cilo': 2.9359050183971735,
  'cihi': 3.0640949816028265,
  'whishi': 4.0,
  'whislo': 2.2,
  'fliers': array([2. , 4.4, 4.1, 4.2]),
  'q1': 2.8,
  'med': 3.0,
  'q3': 3.3}]
```

[10]:
```
outliers = stats[0].get("fliers")
```

[11]:
```
outliers
```

[11]:
```
array([2. , 4.4, 4.1, 4.2])
```