**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.**

In [14]:

*#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()

In [4]:

data **=** pd**.**DataFrame(iris**.**data, columns **=** iris**.**feature\_names)

data['label'] **=** iris**.**target

data**.**sample(5)

Out[4]:

|  | **sepal length (cm)** | **sepal width (cm)** | **petal length (cm)** | **petal width (cm)** | **label** |
| --- | --- | --- | --- | --- | --- |
| **120** | 6.9 | 3.2 | 5.7 | 2.3 | 2 |
| **3** | 4.6 | 3.1 | 1.5 | 0.2 | 0 |
| **132** | 6.4 | 2.8 | 5.6 | 2.2 | 2 |
| **59** | 5.2 | 2.7 | 3.9 | 1.4 | 1 |
| **49** | 5.0 | 3.3 | 1.4 | 0.2 | 0 |

In [5]:

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 int32

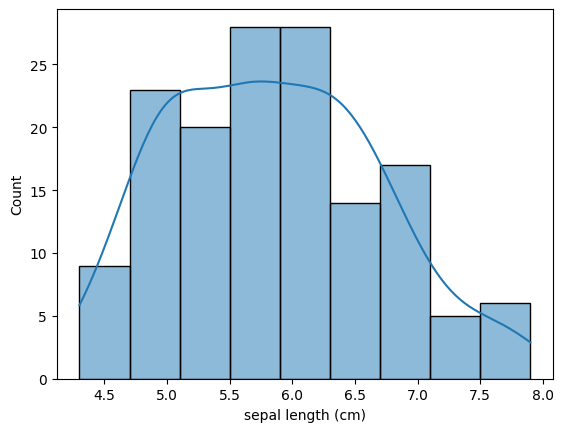
dtypes: float64(4), int32(1)

memory usage: 5.4 KB

**Histograms**

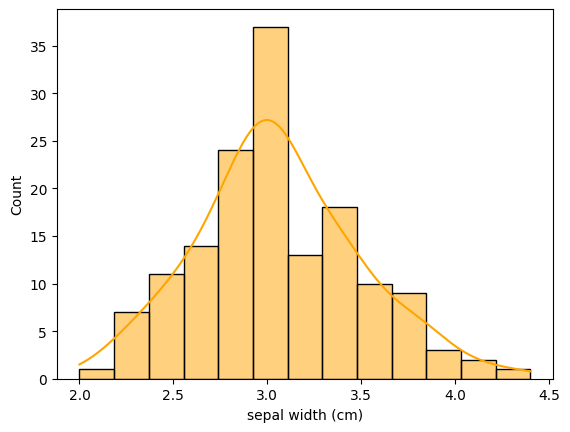
In [8]:

sns**.**histplot(data **=** data, x **=** 'sepal length (cm)', kde**=** **True**);



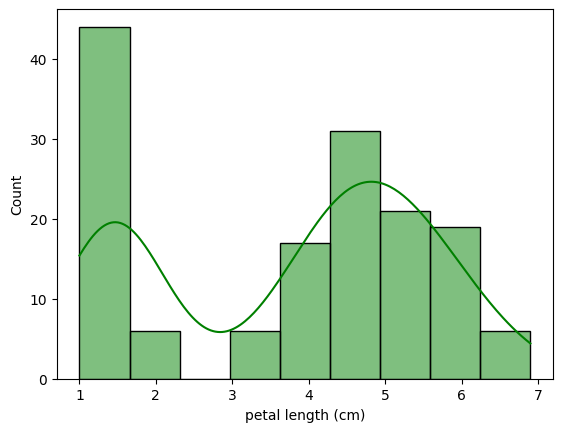
In [10]:

sns**.**histplot(data **=** data, x **=** 'sepal width (cm)', kde**=** **True**, color **=** "orange");



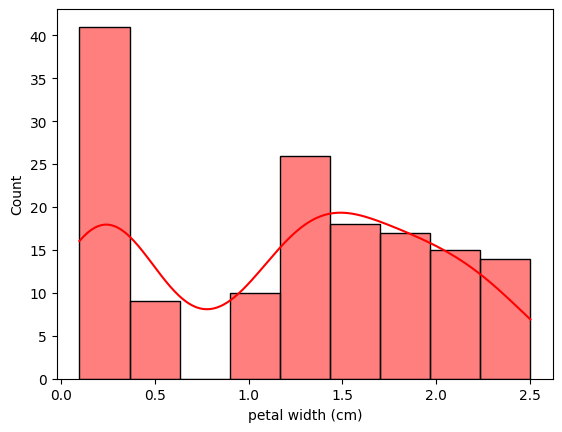
In [11]:

sns**.**histplot(data **=** data, x **=** 'petal length (cm)', kde**=** **True**, color **=** "green");



In [12]:

sns**.**histplot(data **=** data, x **=** 'petal width (cm)', kde**=** **True**, color **=** "red");



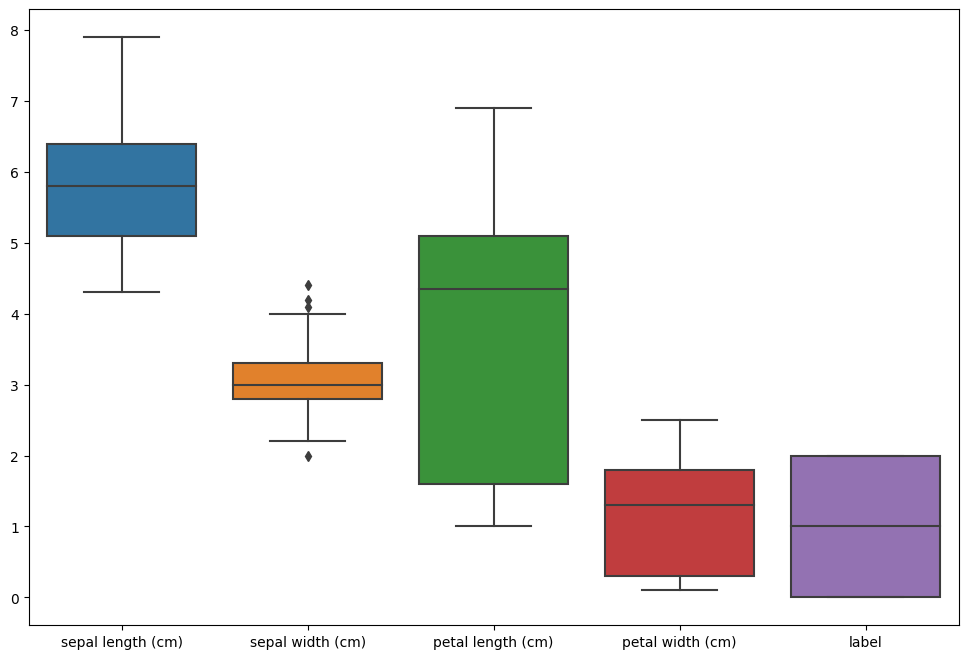
**Box Plots**

In [16]:

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

In [19]:

**from** matplotlib.cbook **import** boxplot\_stats

stats **=** boxplot\_stats(data['sepal width (cm)'])

stats

Out[19]:

[{'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}]

In [26]:

outliers **=** stats[0]**.**get("fliers")

In [28]:

outliers

Out[28]:

array([2. , 4.4, 4.1, 4.2])