omkargaikwad23 updates

History

1 contributor

```
1139 lines (1139 sloc) | 375 KB
```

Assignment 10

Omkar Gaikwad

31126

Data Visualization III 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.

Importing Libraries

In [2]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [3]:

```
sns.get_dataset_names()
```

Out[3]:

```
['anagrams',
 'anscombe',
 'attention',
 'brain_networks',
 'car_crashes',
 'diamonds',
 'dots',
 'exercise',
 'flights',
 'fmri',
 'gammas',
 'geyser',
 'iris',
 'mpg',
 'penguins',
 'planets',
 'taxis',
 'tips',
 'titanic']
```

Loading in the dataset

In [4]:

```
df = sns.load_dataset('iris')
```

In [5]:

```
df
```

Out[5]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa

3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	virginica
146	6.3	2.5	5.0	1.9	virginica
147	6.5	3.0	5.2	2.0	virginica
148	6.2	3.4	5.4	2.3	virginica
149	5.9	3.0	5.1	1.8	virginica

150 rows × 5 columns

1) Features and their data types

In [6]:

```
df.head()
```

Out[6]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

In [7]:

```
df.dtypes
```

Out[7]:

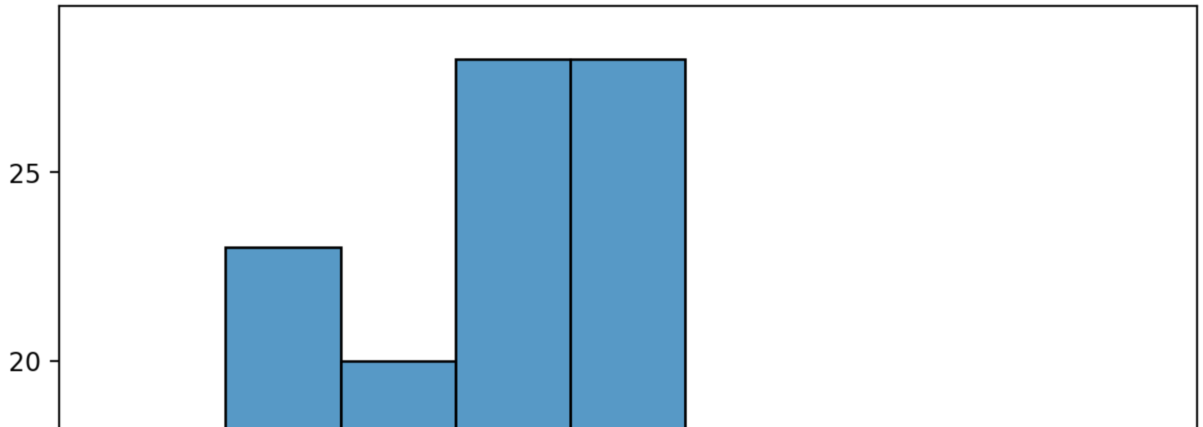
```
sepal_length    float64
sepal_width     float64
petal_length    float64
petal_width     float64
species          object
dtype: object
```

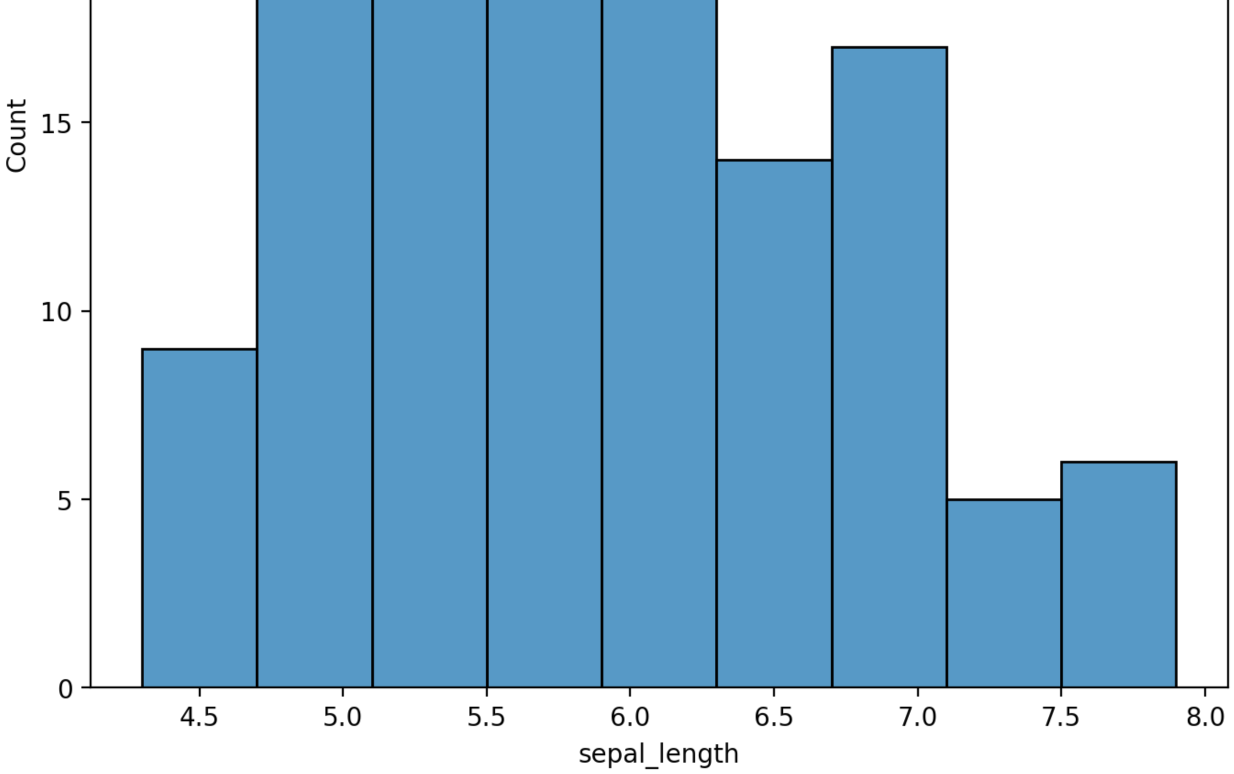
1) Nominal data type is species 2) Numeric data types are petal length, petal width, sepal length, sepal width

2) Histogram for each feature

In [8]:

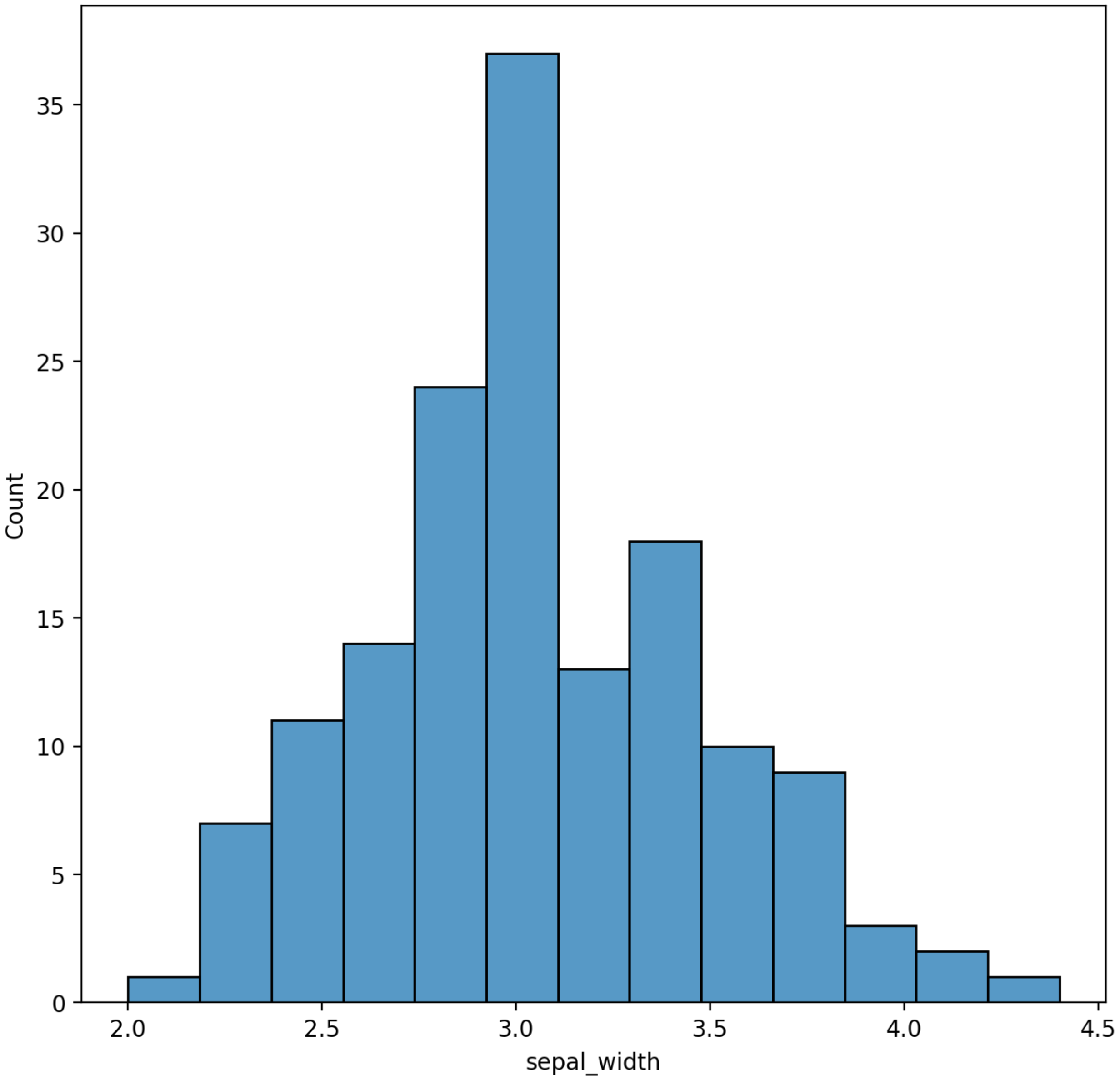
```
plt.figure(figsize=(8,8),dpi=200)
sns.histplot(x='sepal_length',data=df)
plt.show()
```





In [18]:

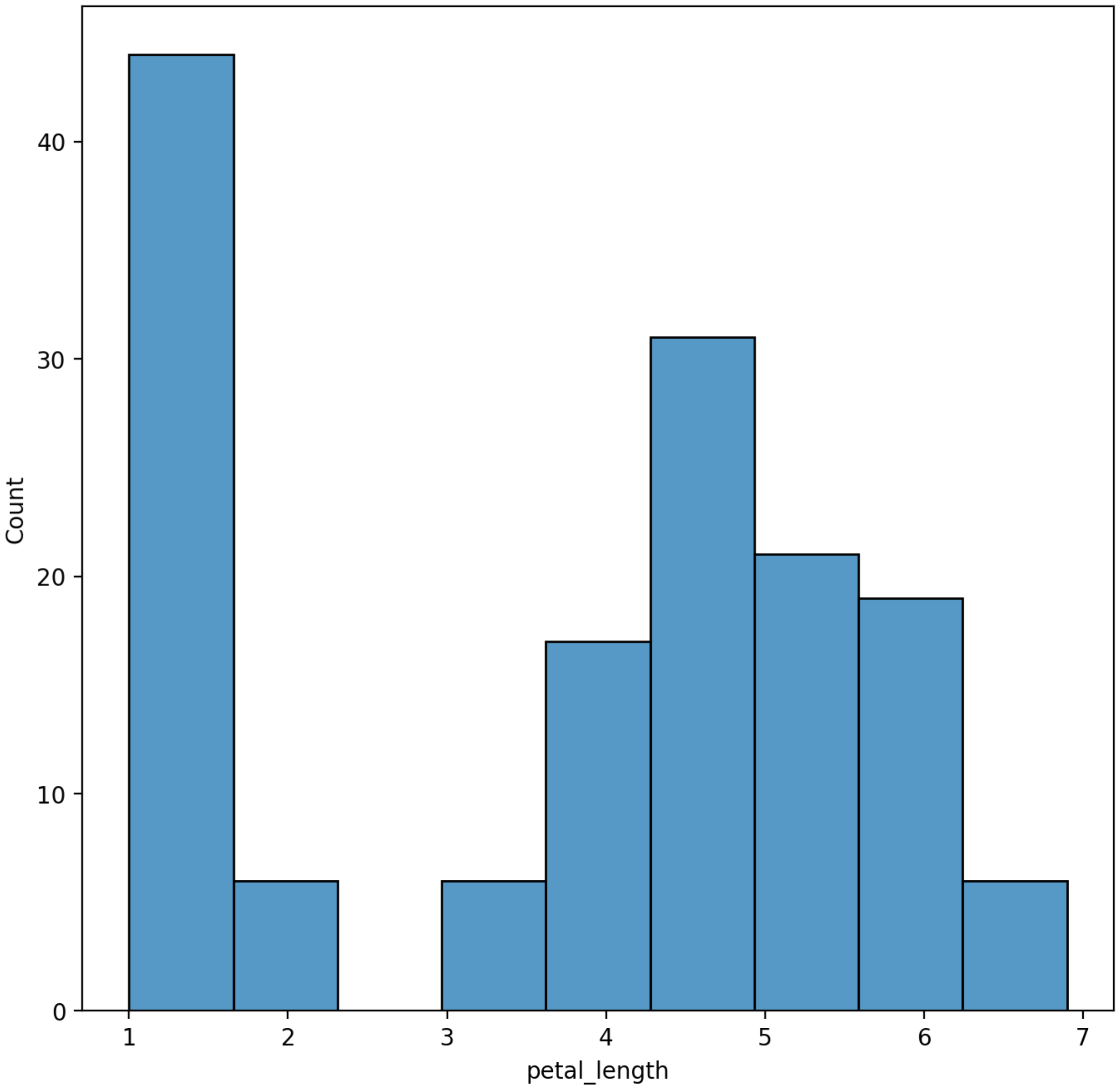
```
plt.figure(figsize=(8,8),dpi=200)
sns.histplot(x='sepal_width',data=df)
plt.show()
```



In [19]:

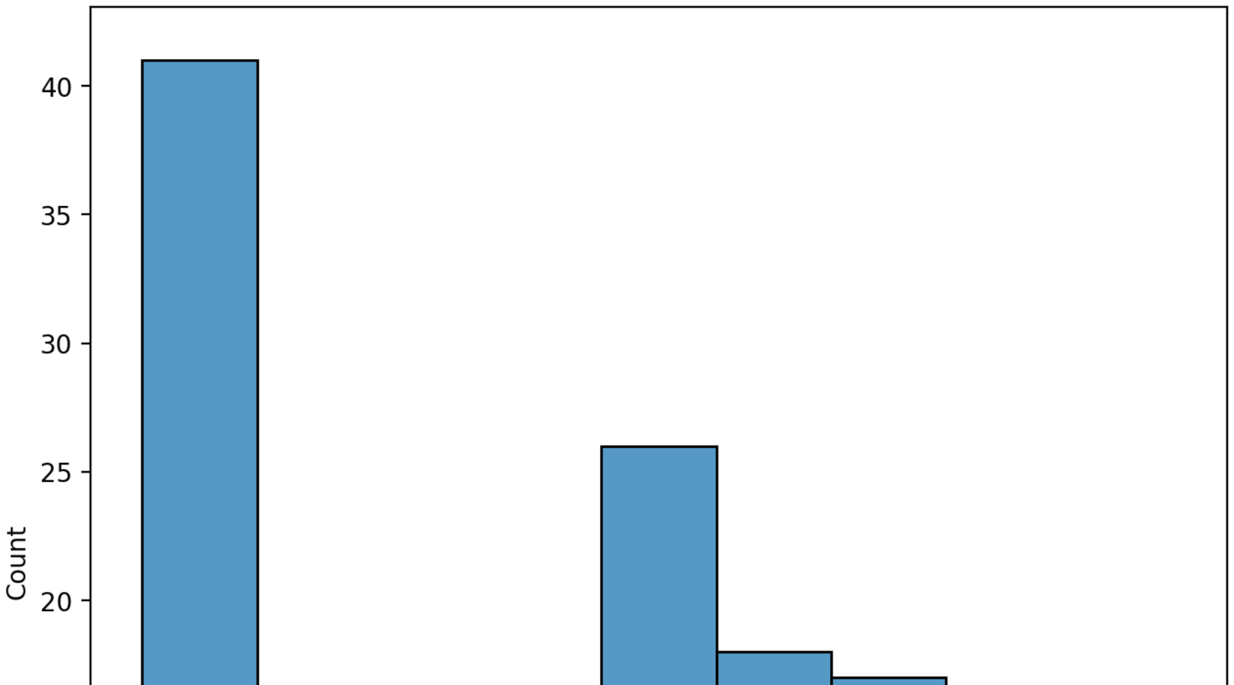
In [19]:

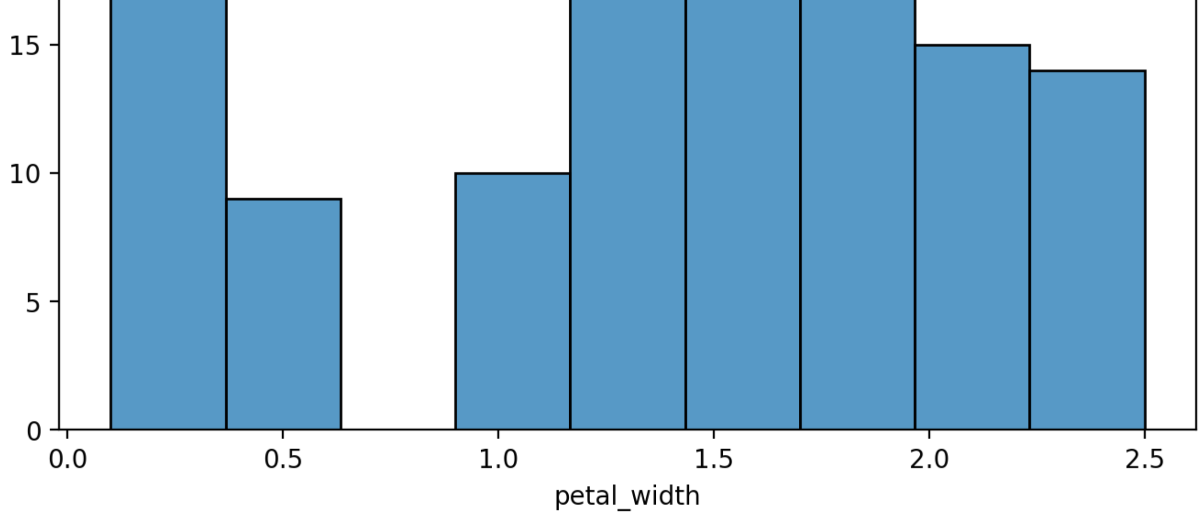
```
plt.figure(figsize=(8,8),dpi=200)
sns.histplot(x='petal_length',data=df)
plt.show()
```



In [20]:

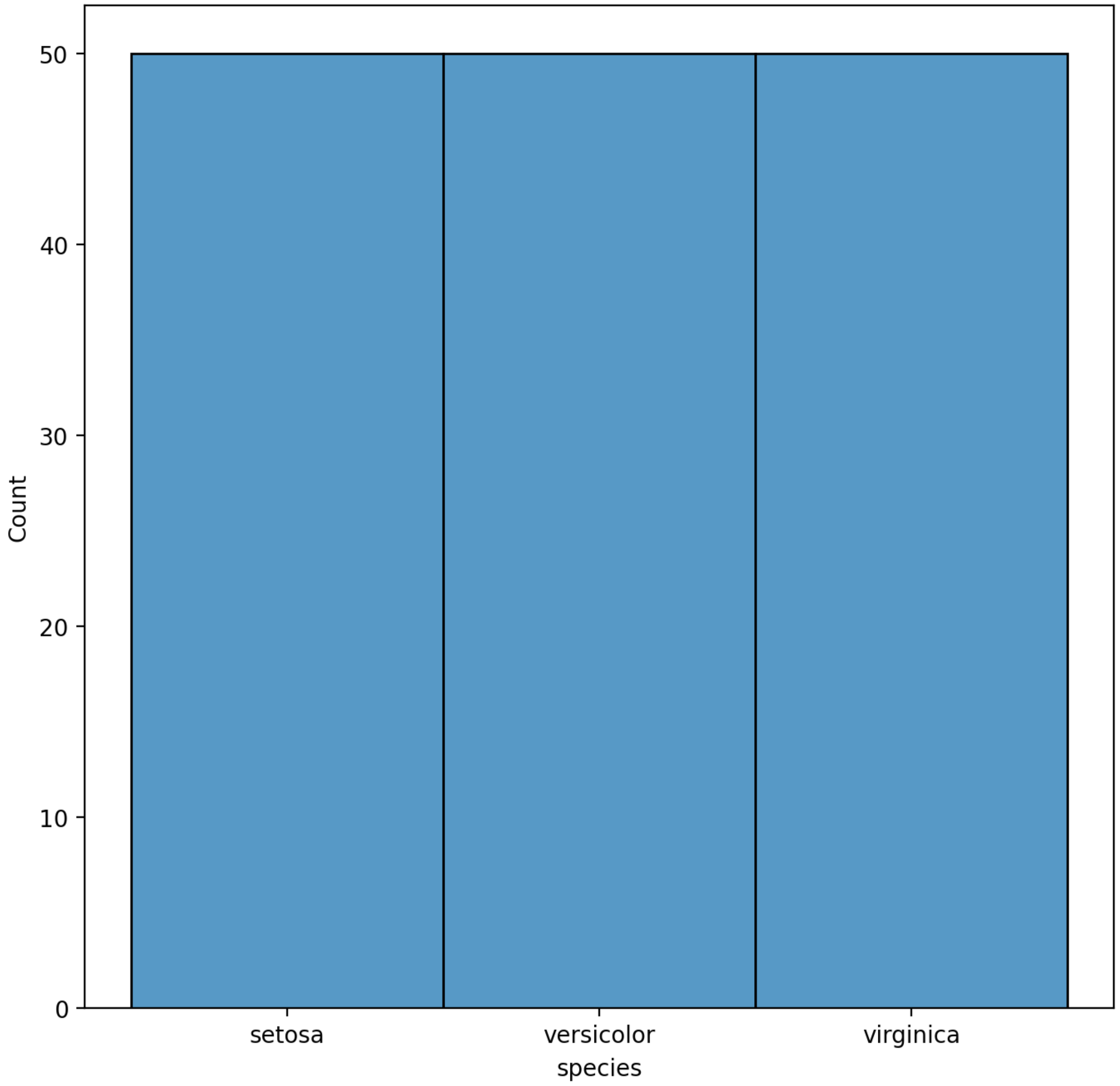
```
plt.figure(figsize=(8,8),dpi=200)
sns.histplot(x='petal_width',data=df)
plt.show()
```





In [21]:

```
plt.figure(figsize=(8,8),dpi=200)
sns.histplot(x='species',data=df)
plt.show()
```

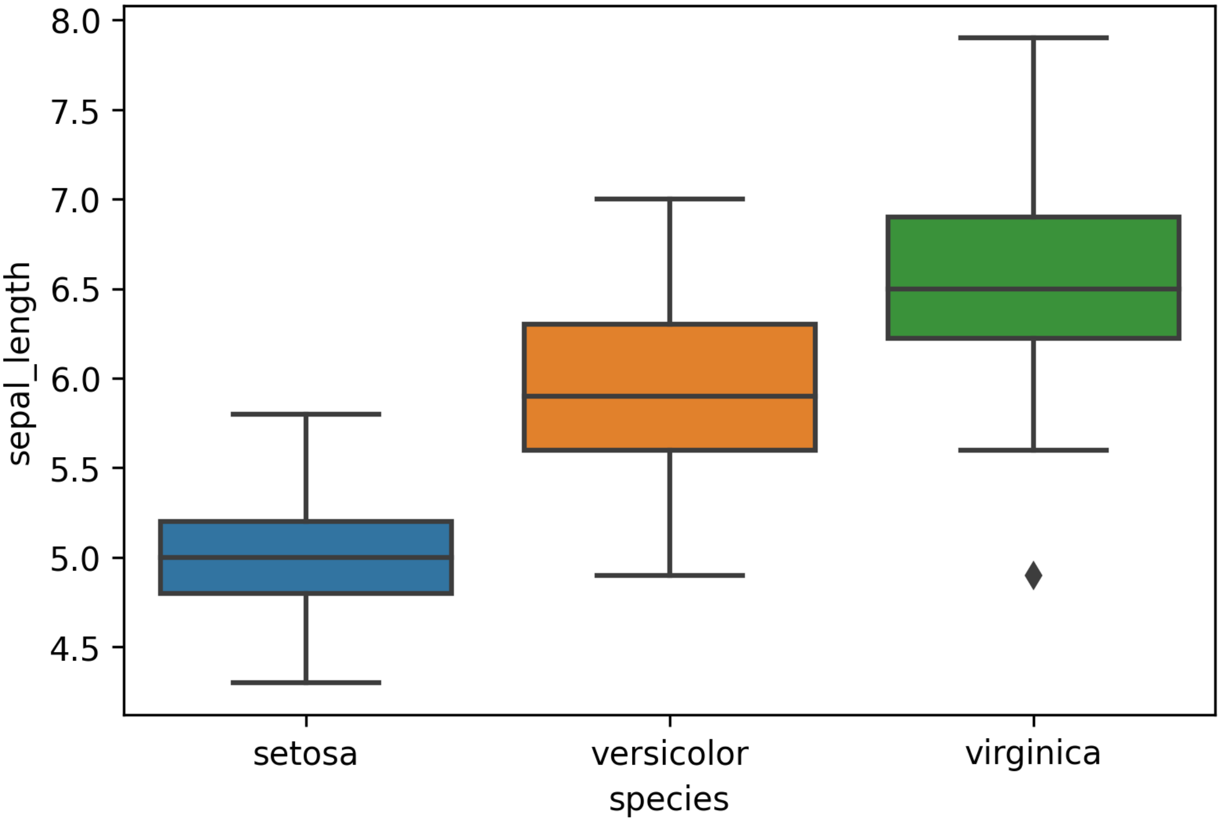


Each species have 50 instances

Boxplot

In [25]:

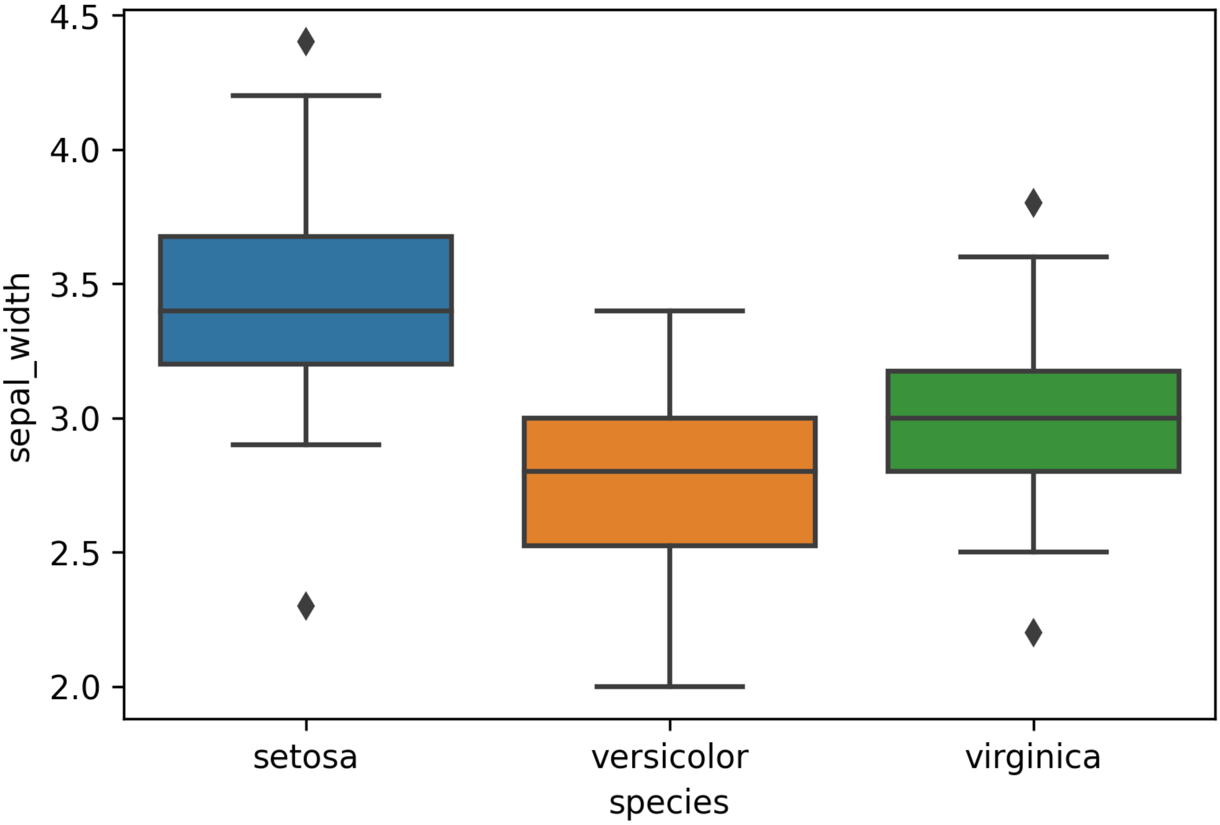
```
plt.figure(dpi=300)
sns.boxplot(x='species',y='sepal_length',data=df)
sns.boxplot(x='species',y='sepal_length',data=df)
plt.show()
```



1) Sepal length of setosa class is the lowest as compared to other two species 2) Virginica species has the highest sepal length

In [26]:

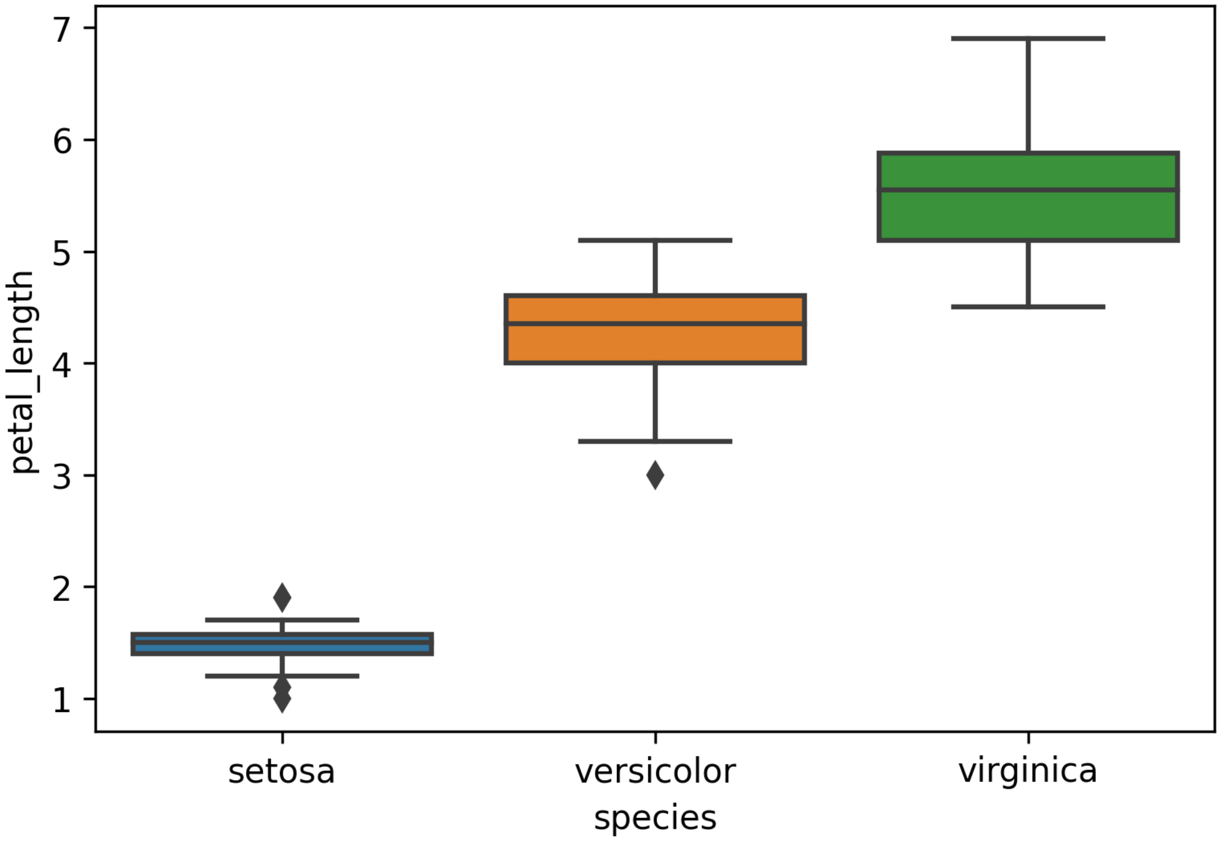
```
plt.figure(dpi=300)
sns.boxplot(x='species',y='sepal_width',data=df)
plt.show()
```



1) there are no outliers present for versicolor species 2) Setosa species has the highest sepal width compared to others

In [27]:

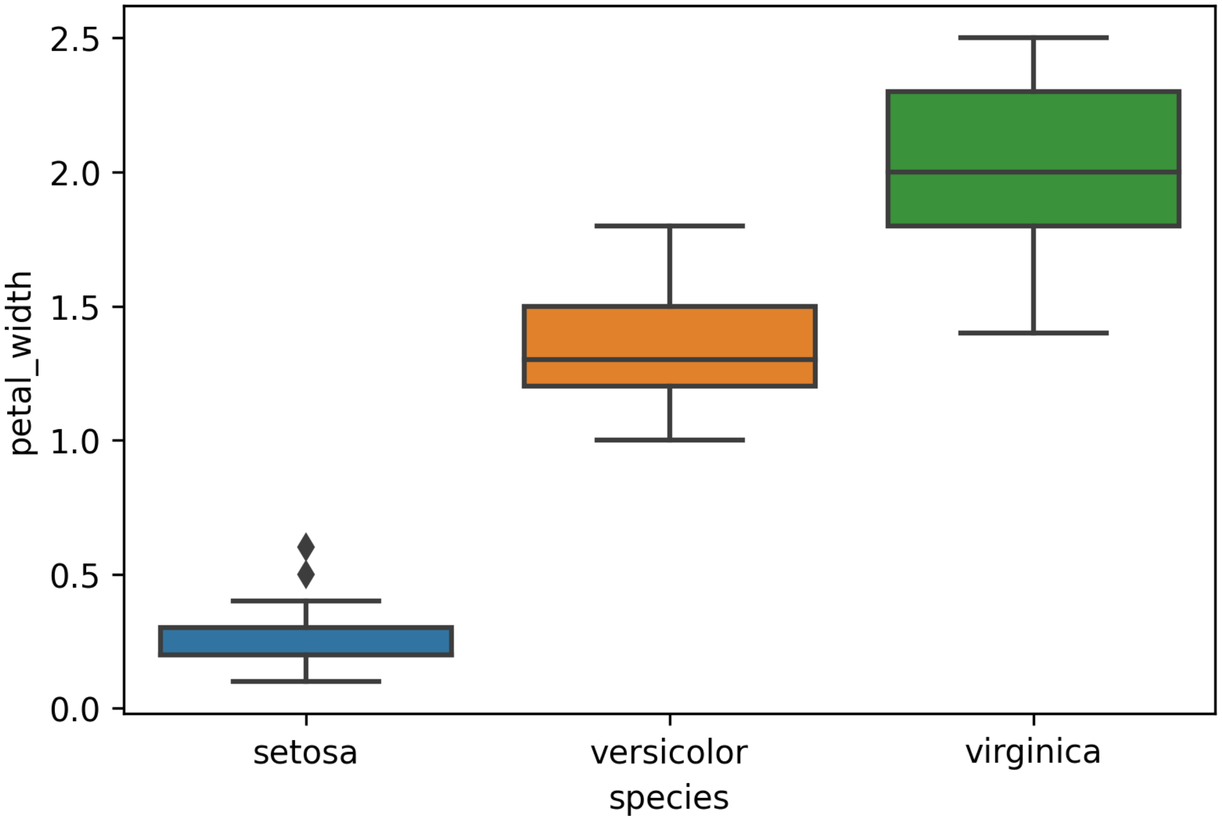
```
plt.figure(dpi=300)
sns.boxplot(x='species',y='petal_length',data=df)
plt.show()
```



1) Petal length of setosa class is lowest and for virginica it is highest

In [28]:

```
plt.figure(dpi=300)
sns.boxplot(x='species',y='petal_width',data=df)
plt.show()
```



In [2]:

```
column_list = ['sepal_length', 'sepal_width', 'petal_length', 'petal_width']

fig, axes = plt.subplots(2,2, figsize=(15,15))
axes_flat = axes.flatten()

index = 0
for axis in axes_flat:
    sns.boxplot(x='species', y=column_list[index], data=df, ax=axis)
    index += 1

plt.show()
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-2-044e6d8ac115> in <module>
      2 
      3 index = 0
----> 4 for axis in axes_flat:
      5     sns.boxplot(x='species', y=column_list[index], data=df, ax=axis)
      6     index += 1

NameError: name 'axes_flat' is not defined
```

In [23]:

```
def outlierDetection (i,df):
    Q1 = np.percentile(df[i], 25)
    Q3 = np.percentile(df[i], 75)
    IQR = Q3 - Q1
    # Upper bound
    upper = np.where(df[i] >= (Q3+1.5*IQR))
    # Lower bound
    lower = np.where(df[i] <= (Q1-1.5*IQR))

    ''' Removing the Outliers '''
    # df.drop(upper[0], axis=0, inplace = True)
    # df.drop(lower[0], axis=0, inplace = True)
    print("Species : ", i)
    print("Lower", lower[0])
    print("Upper", upper[0])
```

In [24]:

```
outlierDetection("sepal_length", df);
outlierDetection("sepal_width", df);
outlierDetection("petal_length", df);
outlierDetection("petal_width", df);
```

```
Species :  sepal_length
Lower []
Upper []
Species :  sepal_width
Lower [60]
Upper [15 32 33]
Species :  petal_length
Lower []
Upper []
Species :  petal_width
Lower []
Upper []
```

In [26]:

```
# Group the Dataset using Species
grouped_data = df.groupby('species')

# Printing the first entry in each
grouped_data.first()
```

Out[26]:

	sepal_length	sepal_width	petal_length	petal_width
species				

setosa	5.1	3.5	1.4	0.2
versicolor	7.0	3.2	4.7	1.4
virginica	6.3	3.3	6.0	2.5

In [27]:

```
grouped_data.describe()
```

Out[27]:

	sepal_length								**sepal_width**		**...**	**petal_length**		
	count	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**	**count**	**mean**	**...**	**75%**	**max**	**cou**
species														
setosa	50.0	5.006	0.352490	4.3	4.800	5.0	5.2	5.8	50.0	3.428	...	1.575	1.9	5
versicolor	50.0	5.936	0.516171	4.9	5.600	5.9	6.3	7.0	50.0	2.770	...	4.600	5.1	5
virginica	50.0	6.588	0.635880	4.9	6.225	6.5	6.9	7.9	50.0	2.974	...	5.875	6.9	5

3 rows × 32 columns

In [28]:

```
for name, group in grouped_data:
    print("\nSpecies Name: ", name, "\n")
    print(group.describe())
```

```
Species Name:  setosa

       sepal_length  sepal_width  petal_length  petal_width
count      50.00000    50.000000     50.000000    50.000000
mean        5.00600     3.428000      1.462000     0.246000
std         0.35249     0.379064      0.173664     0.105386
min         4.30000     2.300000      1.000000     0.100000
25%         4.80000     3.200000      1.400000     0.200000
50%         5.00000     3.400000      1.500000     0.200000
75%         5.20000     3.675000      1.575000     0.300000
max         5.80000     4.400000      1.900000     0.600000

Species Name:  versicolor

       sepal_length  sepal_width  petal_length  petal_width
count     50.000000    50.000000     50.000000    50.000000
mean       5.936000     2.770000      4.260000     1.326000
std        0.516171     0.313798      0.469911     0.197753
min        4.900000     2.000000      3.000000     1.000000
25%        5.600000     2.525000      4.000000     1.200000
50%        5.900000     2.800000      4.350000     1.300000
75%        6.300000     3.000000      4.600000     1.500000
max        7.000000     3.400000      5.100000     1.800000

Species Name:  virginica

       sepal_length  sepal_width  petal_length  petal_width
count      50.00000    50.000000     50.000000     50.00000
mean        6.58800     2.974000      5.552000      2.02600
std         0.63588     0.322497      0.551895      0.27465
min         4.90000     2.200000      4.500000      1.40000
25%         6.22500     2.800000      5.100000      1.80000
50%         6.50000     3.000000      5.550000      2.00000
75%         6.90000     3.175000      5.875000      2.30000
max         7.90000     3.800000      6.900000      2.50000
```

Observations

It has been observed that the attributes of Iris-versicolor and Iris-virginica are almost similar. The major difference between the 2 is Sepal length and Sepal Width. Iris-Setosa on the other hand, has a very small petal length and width as compared to the other 2.

In [ ]: