Assignment No. 10

Aim : 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 (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 boxplot for each feature in the dataset.
4. Compare distributions and identify outliers.

Code :

In [15]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [16]:
```
df1 = sns.load_dataset('iris')
```

In [17]:
```
df1
```

Out[17]:

	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

In [18]:
```
df1.dtypes
```

Out[18]:
```
sepal_length    float64
sepal_width     float64
petal_length    float64
petal_width     float64
species          object
dtype: object
```

In [19]:
```
df1.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  150 non-null    float64
 1   sepal_width   150 non-null    float64
 2   petal_length  150 non-null    float64
 3   petal_width   150 non-null    float64
 4   species       150 non-null    object
dtypes: float64(4), object(1)
memory usage: 6.0+ KB
```

In [20]:
```
df1.describe()
```

Out[20]:

	sepal_length	sepal_width	petal_length	petal_width
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.057333	3.758000	1.199333
std	0.828066	0.435866	1.765298	0.762238
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

In [21]:
```
df1.shape
```

Out[21]:
```
(150, 5)
```

In [22]:
```
df1.head()
```

Out[22]:

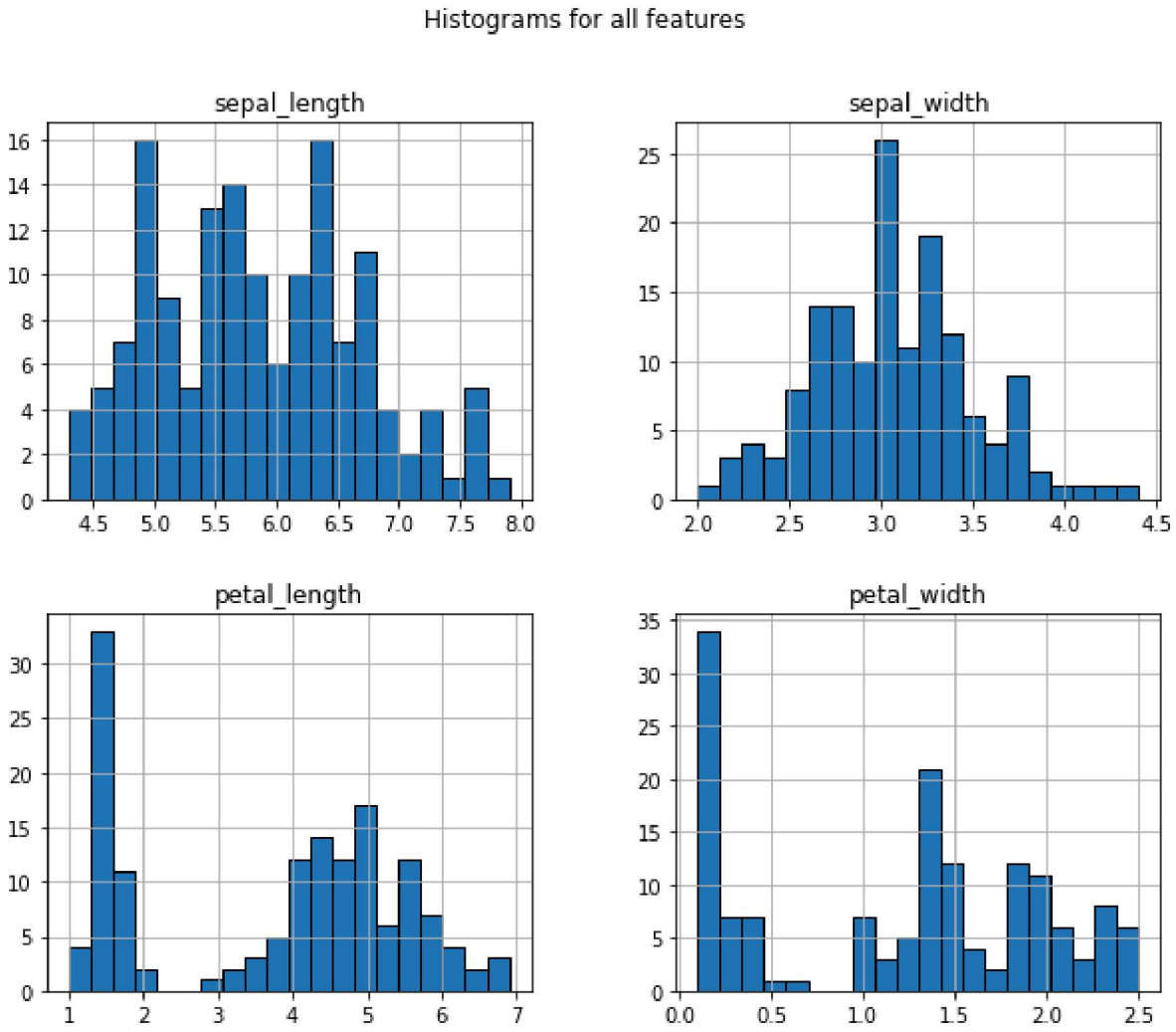
	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

In [23]:
```
df1.tail()
```

Out[23]:

	sepal_length	sepal_width	petal_length	petal_width	species
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

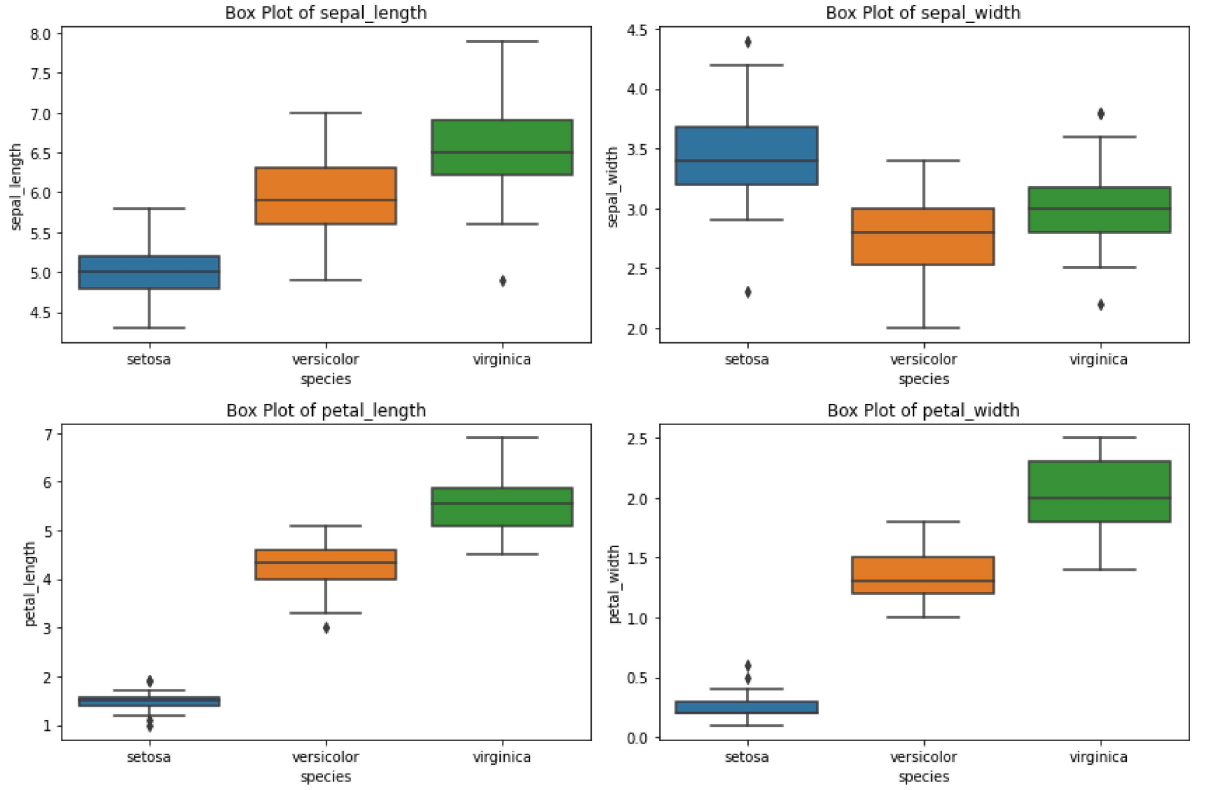
In [25]:
```
df1.drop('species', axis=1).hist(figsize=(10, 8), bins=20, edgecolor='black')
plt.suptitle('Histograms for all features')
plt.show()
```



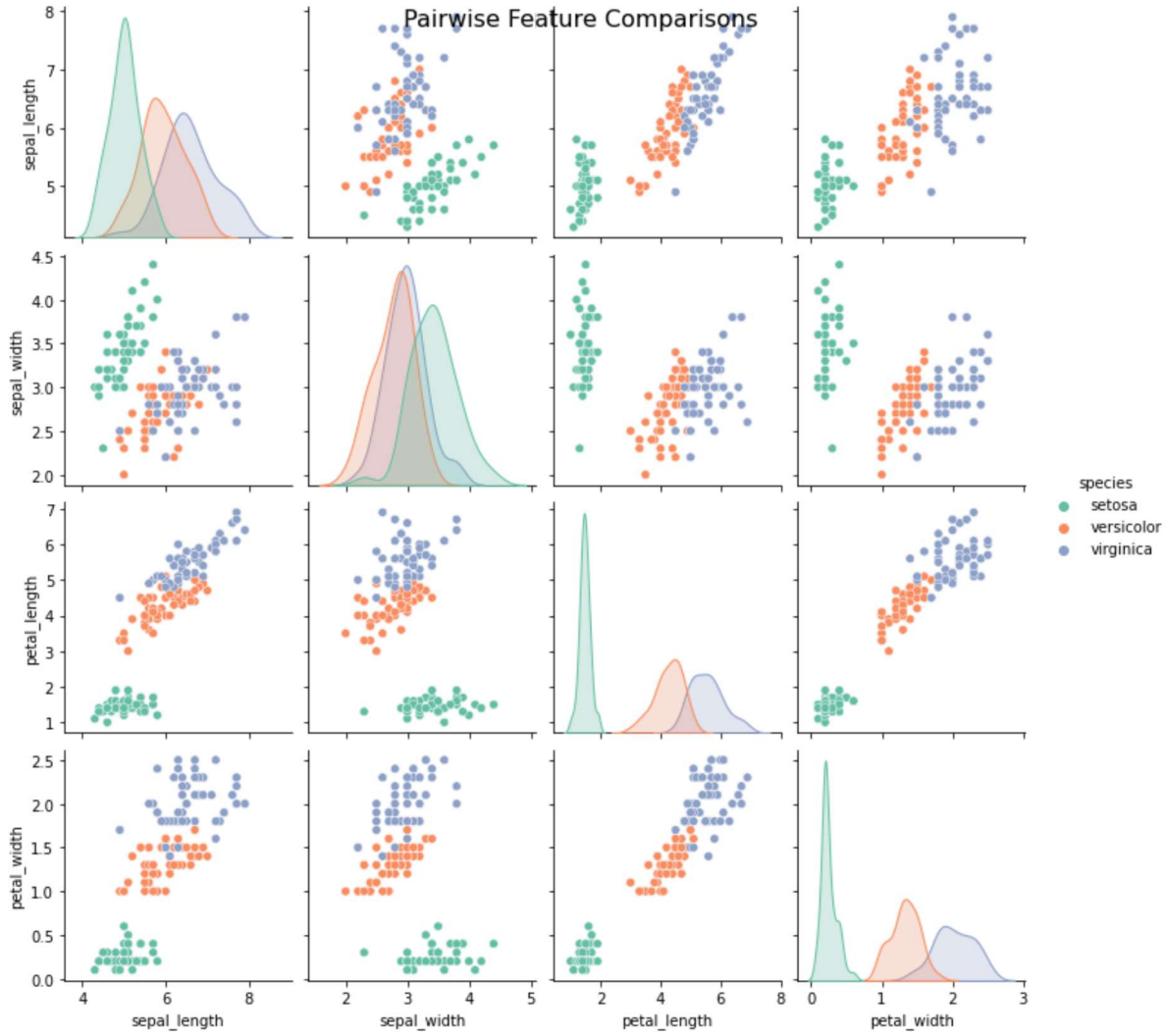
In [28]:
```
plt.figure(figsize=(12, 8))

for i, feature in enumerate(df1.drop('species', axis=1).columns):
    plt.subplot(2, 2, i + 1)
    sns.boxplot(x='species', y=feature, data=df1)
    plt.title(f'Box Plot of {feature}')

plt.tight_layout()
plt.show()
```



In [30]:
```
sns.pairplot(df1, hue='species', palette='Set2')
plt.suptitle('Pairwise Feature Comparisons', fontsize=16)
plt.show()
```



In [32]:
```
from scipy.stats import zscore

z_scores = pd.DataFrame(zscore(df1.drop('species', axis=1)), columns=df1.drop('

outliers = (z_scores.abs() > 3).sum()
print("\nOutliers (Z-scores > 3 or < -3):")
print(outliers)
```

```
Outliers (Z-scores > 3 or < -3):
sepal_length    0
sepal_width     1
petal_length    0
petal_width     0
dtype: int64
```

Name : Mohan Kadamabnde

Roll No. : 13212 (TECO-b1)