data-analysis-1

November 16, 2025

0.1 using IRIS dataset

[1]:
```
import pandas as pd
import numpy as np
```

[2]:
```
df = pd.read_csv("archive/IRIS.csv")
```

[3]:
```
df
```

[3]:
```
     sepal_length  sepal_width  petal_length  petal_width         species
0             5.1          3.5           1.4          0.2     Iris-setosa
1             4.9          3.0           1.4          0.2     Iris-setosa
2             4.7          3.2           1.3          0.2     Iris-setosa
3             4.6          3.1           1.5          0.2     Iris-setosa
4             5.0          3.6           1.4          0.2     Iris-setosa
..            …            …             …            …               …
145           6.7          3.0           5.2          2.3  Iris-virginica
146           6.3          2.5           5.0          1.9  Iris-virginica
147           6.5          3.0           5.2          2.0  Iris-virginica
148           6.2          3.4           5.4          2.3  Iris-virginica
149           5.9          3.0           5.1          1.8  Iris-virginica

[150 rows x 5 columns]
```

[4]:
```
df.head()
```

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

[5]:
```
df.tail()
```

[5]:
```
     sepal_length  sepal_width  petal_length  petal_width         species
145           6.7          3.0           5.2          2.3  Iris-virginica
146           6.3          2.5           5.0          1.9  Iris-virginica
```

```
147           6.5          3.0           5.2          2.0  Iris-virginica
148           6.2          3.4           5.4          2.3  Iris-virginica
149           5.9          3.0           5.1          1.8  Iris-virginica
```

[6]:
```
df.species
```

[6]:
```
0         Iris-setosa
1         Iris-setosa
2         Iris-setosa
3         Iris-setosa
4         Iris-setosa
             …
145    Iris-virginica
146    Iris-virginica
147    Iris-virginica
148    Iris-virginica
149    Iris-virginica
Name: species, Length: 150, dtype: object
```

[8]:
```
df.species.unique()
```

[8]:
```
array(['Iris-setosa', 'Iris-versicolor', 'Iris-virginica'], dtype=object)
```

[10]:
```
df["species"].unique()
```

[10]:
```
array(['Iris-setosa', 'Iris-versicolor', 'Iris-virginica'], dtype=object)
```

[16]:
```
df.isnull()
```

[16]:
```
     sepal_length  sepal_width  petal_length  petal_width  species
0           False        False         False        False    False
1           False        False         False        False    False
2           False        False         False        False    False
3           False        False         False        False    False
4           False        False         False        False    False
..            …          …          …          …      …
145         False        False         False        False    False
146         False        False         False        False    False
147         False        False         False        False    False
148         False        False         False        False    False
149         False        False         False        False    False

[150 rows x 5 columns]
```

[18]:
```
df.describe()
```

[18]:
```
       sepal_length  sepal_width  petal_length  petal_width
count    150.000000   150.000000    150.000000   150.000000
```

```
mean       5.843333    3.054000      3.758667     1.198667
std        0.828066    0.433594      1.764420     0.763161
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
```

[19]:
```
df = df.dropna()
```

[20]:
```
df
```

[20]:
```
     sepal_length  sepal_width  petal_length  petal_width         species
0             5.1          3.5           1.4          0.2     Iris-setosa
1             4.9          3.0           1.4          0.2     Iris-setosa
2             4.7          3.2           1.3          0.2     Iris-setosa
3             4.6          3.1           1.5          0.2     Iris-setosa
4             5.0          3.6           1.4          0.2     Iris-setosa
..              …            …             …            …               …
145           6.7          3.0           5.2          2.3  Iris-virginica
146           6.3          2.5           5.0          1.9  Iris-virginica
147           6.5          3.0           5.2          2.0  Iris-virginica
148           6.2          3.4           5.4          2.3  Iris-virginica
149           5.9          3.0           5.1          1.8  Iris-virginica

[150 rows x 5 columns]
```

[21]:
```
df.duplicated()
```

[21]:
```
0      False
1      False
2      False
3      False
4      False
       …
145    False
146    False
147    False
148    False
149    False
Length: 150, dtype: bool
```

[22]:
```
df[df.duplicated()]
```

[22]:
```
     sepal_length  sepal_width  petal_length  petal_width      species
34            4.9          3.1           1.5          0.1  Iris-setosa
37            4.9          3.1           1.5          0.1  Iris-setosa
```

```
142           5.8          2.7           5.1          1.9  Iris-virginica
```

```
[23]: df = df.drop_duplicates()
```

```
[24]: df
```

```
[24]:      sepal_length  sepal_width  petal_length  petal_width          species
     0             5.1          3.5           1.4          0.2      Iris-setosa
     1             4.9          3.0           1.4          0.2      Iris-setosa
     2             4.7          3.2           1.3          0.2      Iris-setosa
     3             4.6          3.1           1.5          0.2      Iris-setosa
     4             5.0          3.6           1.4          0.2      Iris-setosa
     ..            …            …             …            …                …
     145           6.7          3.0           5.2          2.3  Iris-virginica
     146           6.3          2.5           5.0          1.9  Iris-virginica
     147           6.5          3.0           5.2          2.0  Iris-virginica
     148           6.2          3.4           5.4          2.3  Iris-virginica
     149           5.9          3.0           5.1          1.8  Iris-virginica

     [147 rows x 5 columns]
```

```
[25]: df = df.reset_index()
```

```
[26]: df
```

```
[26]:      index  sepal_length  sepal_width  petal_length  petal_width  \
     0        0           5.1          3.5           1.4          0.2
     1        1           4.9          3.0           1.4          0.2
     2        2           4.7          3.2           1.3          0.2
     3        3           4.6          3.1           1.5          0.2
     4        4           5.0          3.6           1.4          0.2
     ..     …            …            …             …            …
     142    145           6.7          3.0           5.2          2.3
     143    146           6.3          2.5           5.0          1.9
     144    147           6.5          3.0           5.2          2.0
     145    148           6.2          3.4           5.4          2.3
     146    149           5.9          3.0           5.1          1.8

                 species
     0       Iris-setosa
     1       Iris-setosa
     2       Iris-setosa
     3       Iris-setosa
     4       Iris-setosa
     ..              …
     142  Iris-virginica
     143  Iris-virginica
```

```
144  Iris-virginica
145  Iris-virginica
146  Iris-virginica

[147 rows x 6 columns]
```

```
[27]: df["species"].unique()
```

```
[27]: array(['Iris-setosa', 'Iris-versicolor', 'Iris-virginica'], dtype=object)
```

```
[28]: df.columns
```

```
[28]: Index(['index', 'sepal_length', 'sepal_width', 'petal_length', 'petal_width',
              'species'],
             dtype='object')
```

```
[30]: features = ['sepal_length', 'sepal_width', 'petal_length', 'petal_width']

      for feature in features:
          print(df[feature].unique())
```

```
[5.1 4.9 4.7 4.6 5.  5.4 4.4 4.8 4.3 5.8 5.7 5.2 5.5 4.5 5.3 7.  6.4 6.9
 6.5 6.3 6.6 5.9 6.  6.1 5.6 6.7 6.2 6.8 7.1 7.6 7.3 7.2 7.7 7.4 7.9]
[3.5 3.  3.2 3.1 3.6 3.9 3.4 2.9 3.7 4.  4.4 3.8 3.3 4.1 4.2 2.3 2.8 2.4
 2.7 2.  2.2 2.5 2.6]
[1.4 1.3 1.5 1.7 1.6 1.1 1.2 1.  1.9 4.7 4.5 4.9 4.  4.6 3.3 3.9 3.5 4.2
 3.6 4.4 4.1 4.8 4.3 5.  3.8 3.7 5.1 3.  6.  5.9 5.6 5.8 6.6 6.3 6.1 5.3
 5.5 6.7 6.9 5.7 6.4 5.4 5.2]
[0.2 0.4 0.3 0.1 0.5 0.6 1.4 1.5 1.3 1.6 1.  1.1 1.8 1.2 1.7 2.5 1.9 2.1
 2.2 2.  2.4 2.3]
```

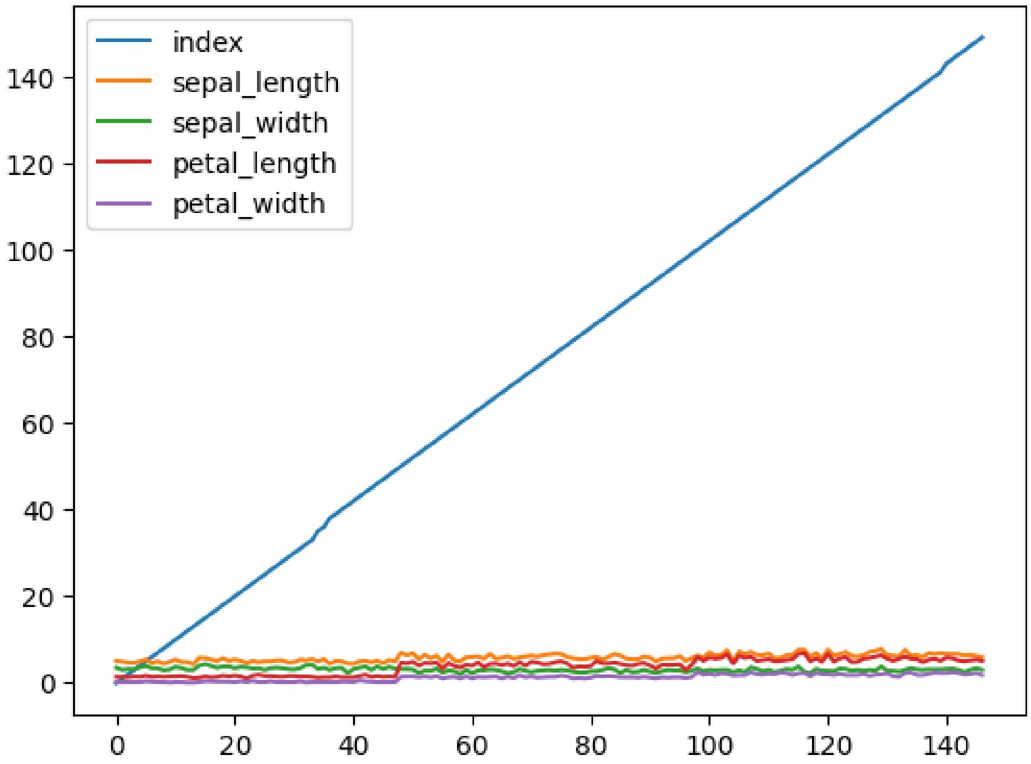
```
[31]: df["species"].value_counts()
```

```
[31]: species
      Iris-versicolor    50
      Iris-virginica     49
      Iris-setosa        48
      Name: count, dtype: int64
```

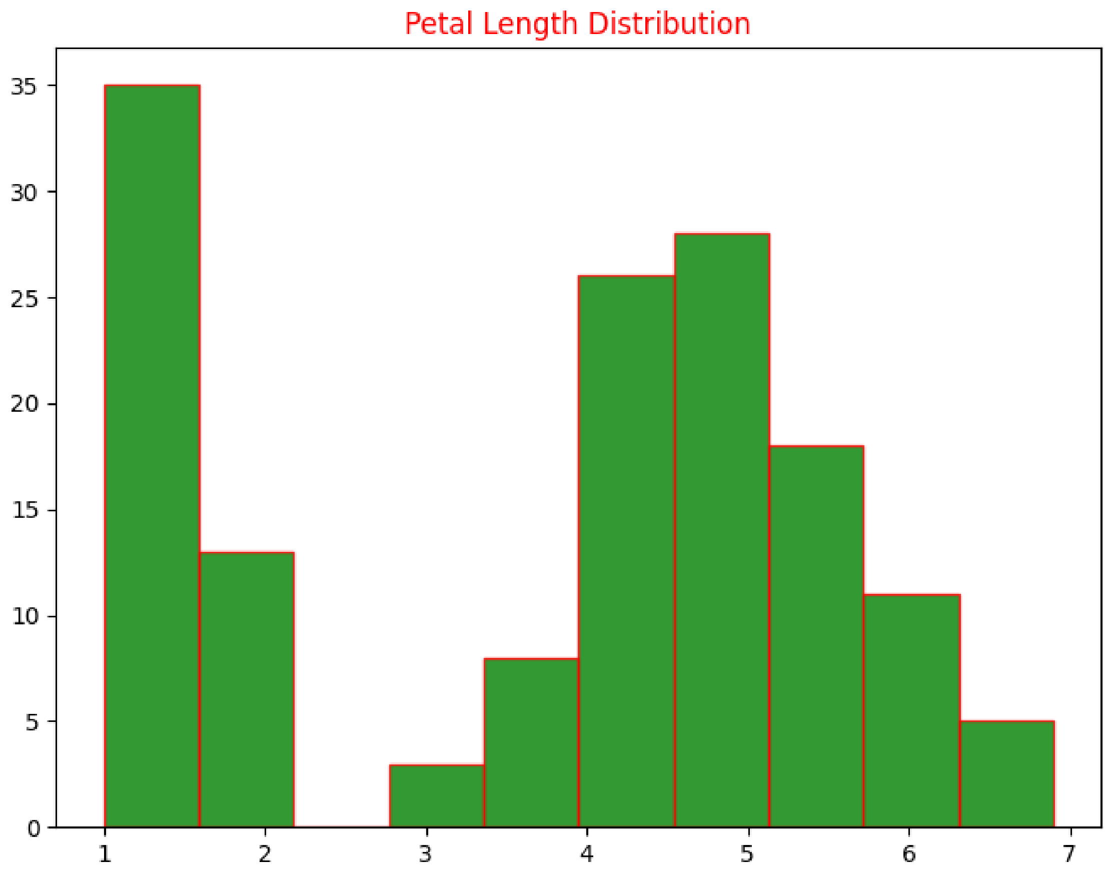
```
[32]: import matplotlib.pyplot as plt
```

```
[33]: df.plot(kind="line")
```

```
[33]: <Axes: >
```

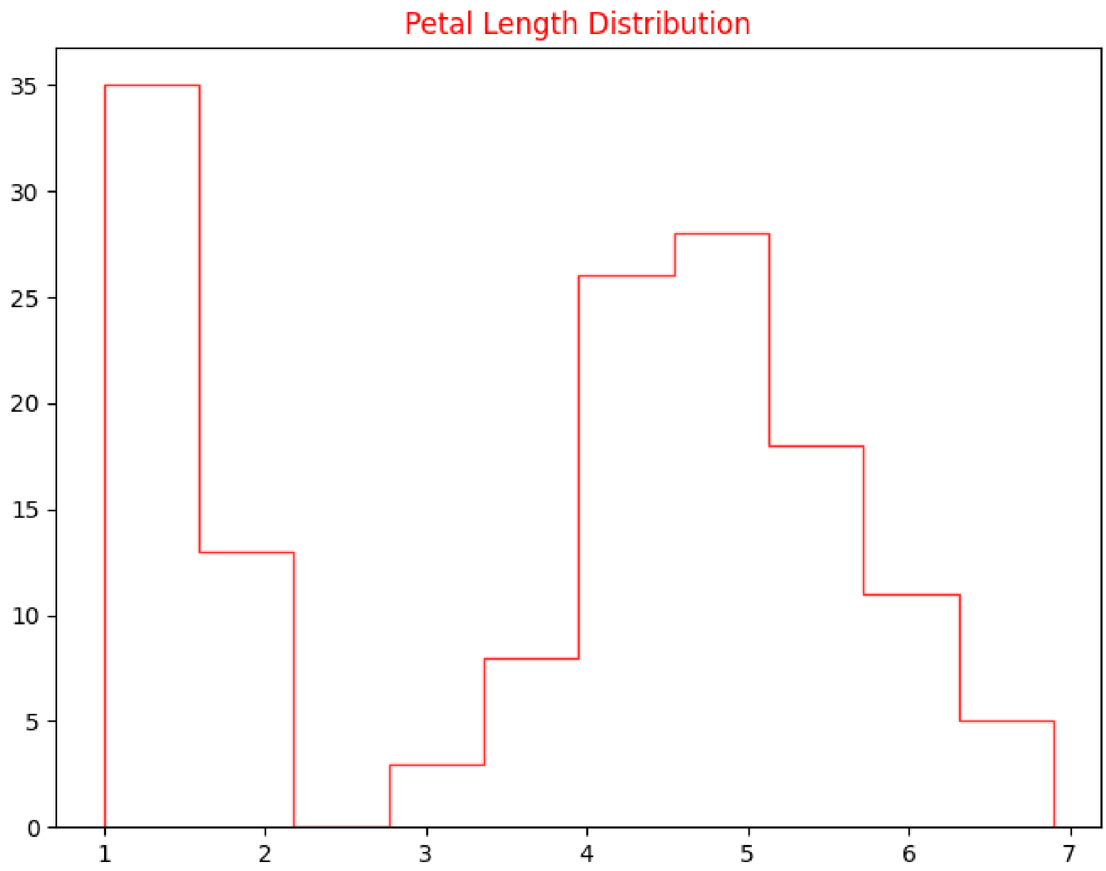


[46]:
```
plt.figure(figsize=(8, 6))
plt.hist(df["petal_length"], color="green", edgecolor="red", alpha=0.8)
plt.title("Petal Length Distribution", color="red")
plt.show()
```



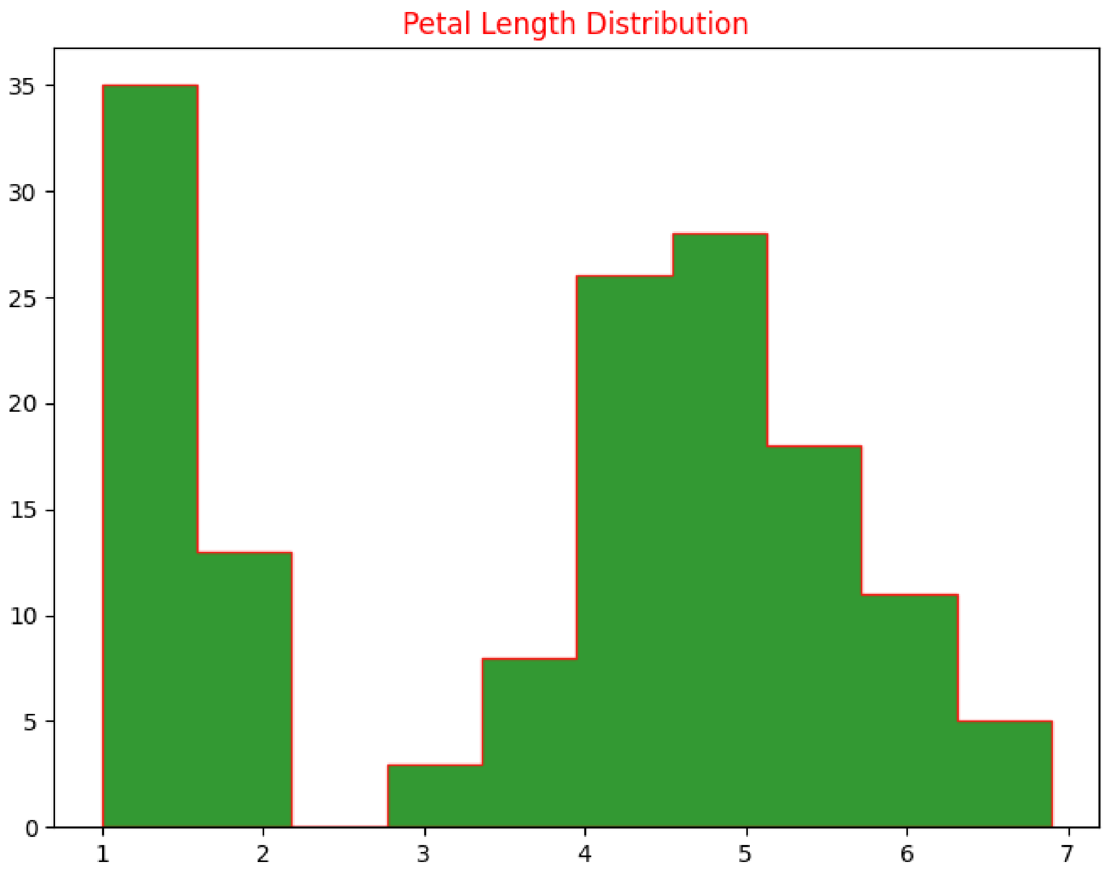
[47]:

```
plt.figure(figsize=(8, 6))
plt.hist(df["petal_length"], color="green", edgecolor="red", alpha=0.8, histtype="step")
plt.title("Petal Length Distribution", color="red")
plt.show()
```



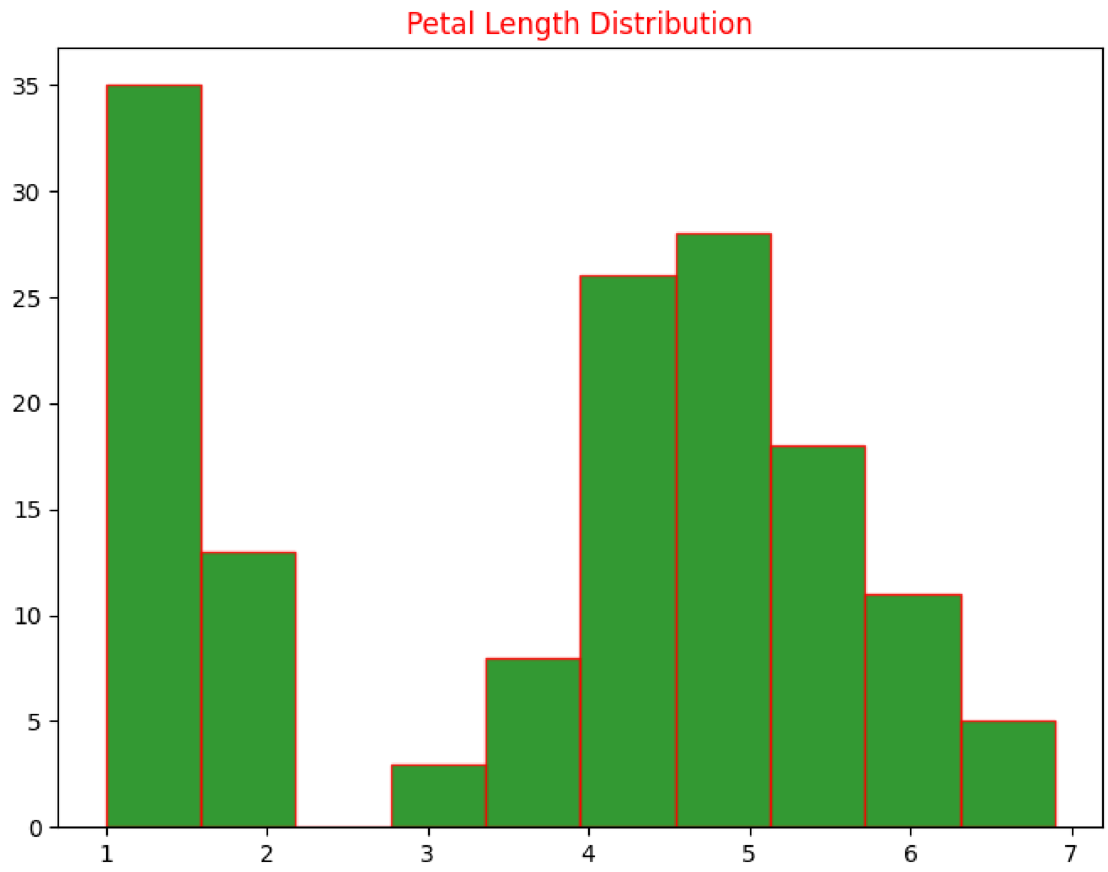
[48]:

```
plt.figure(figsize=(8, 6))
plt.hist(df["petal_length"], color="green", edgecolor="red", alpha=0.8,␣
  ↪histtype="stepfilled")
plt.title("Petal Length Distribution", color="red")
plt.show()
```



[ ]:

```python
plt.figure(figsize=(8, 6))
plt.hist(df["petal_length"], color="green", edgecolor="red", alpha=0.8, histtype="barstacked")
plt.title("Petal Length Distribution", color="red")
# plt.show()
plt.savefig("plot.png")
```



[60]:

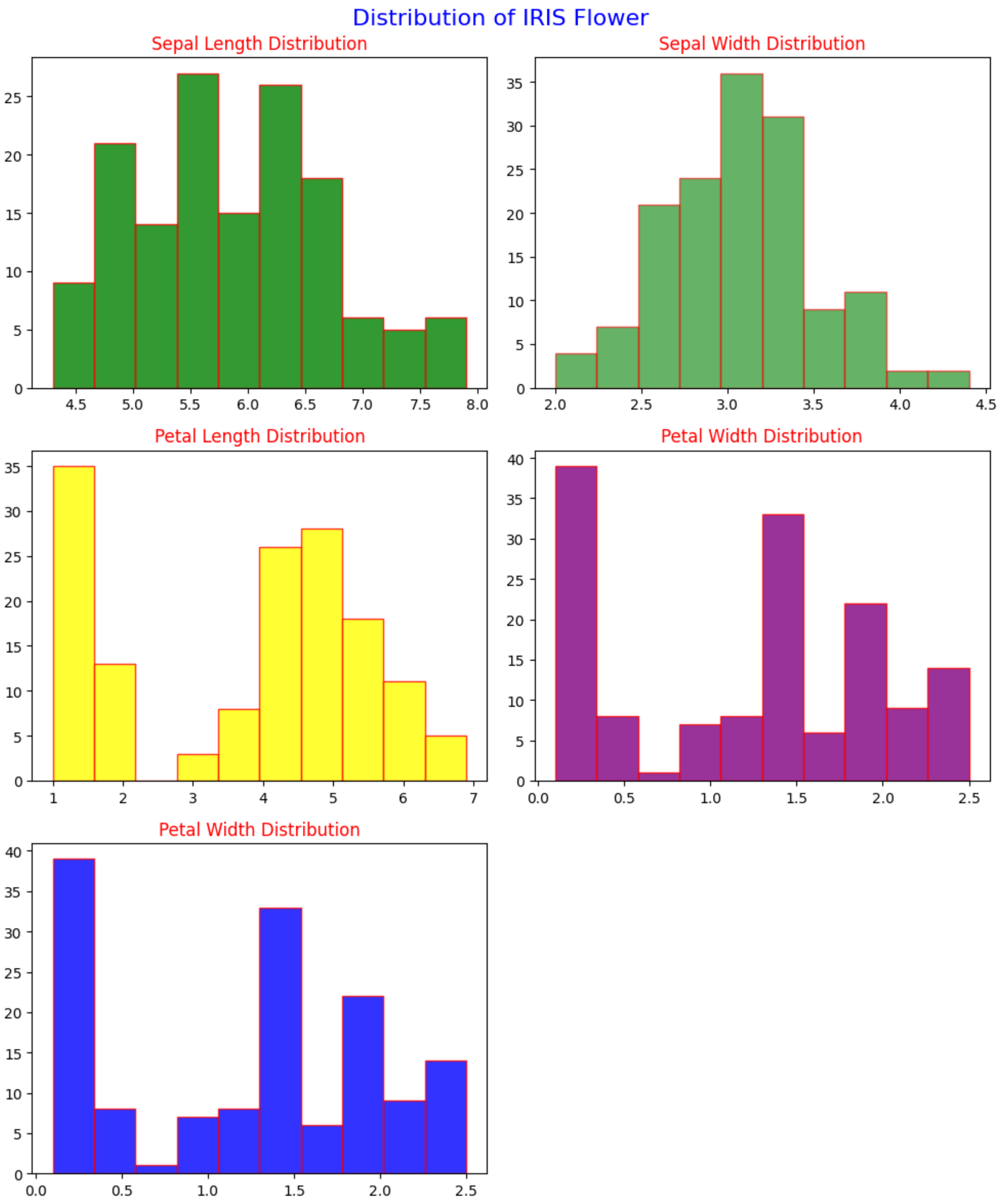
```
df.columns
```

[60]:

```
Index(['index', 'sepal_length', 'sepal_width', 'petal_length', 'petal_width',
       'species'],
      dtype='object')
```

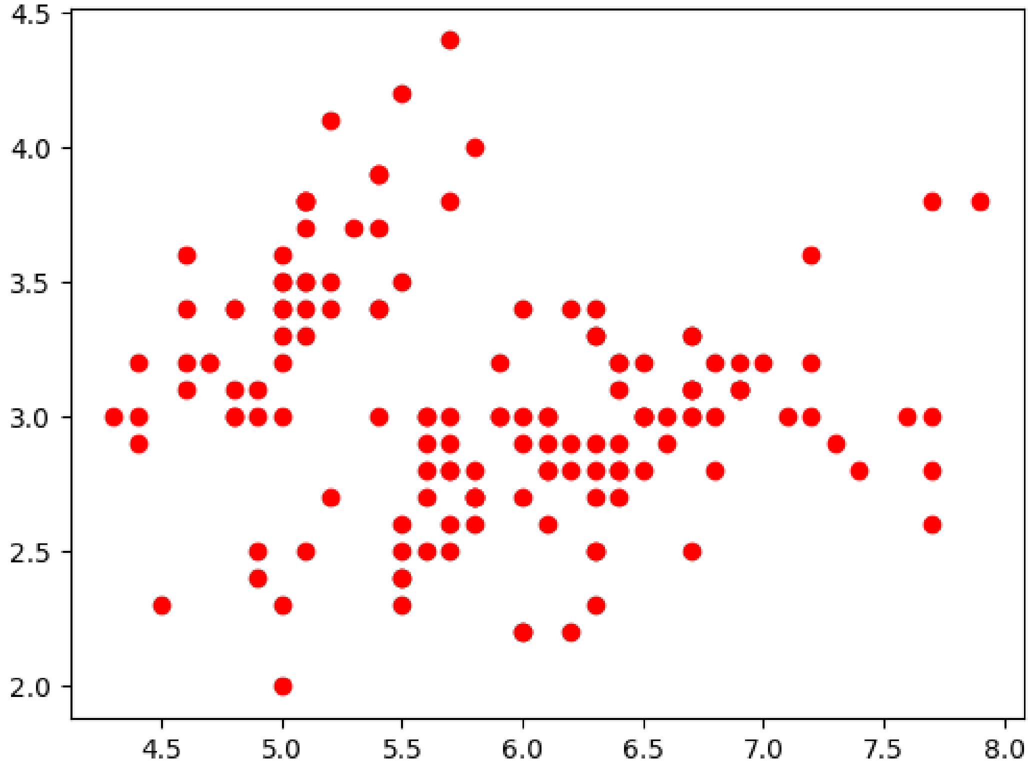
[70]:

```
# using subplot

plt.figure(figsize=(10, 12))
plt.suptitle("Distribution of IRIS Flower", fontsize=16, color="blue")

plt.subplot(3, 2, 1)
plt.hist(df["sepal_length"], color="green", edgecolor="red", alpha=0.8)
plt.title("Sepal Length Distribution", color="red")

plt.subplot(3, 2, 2)
plt.hist(df["sepal_width"], color="green", edgecolor="red", alpha=0.6)
plt.title("Sepal Width Distribution", color="red")
plt.subplot(3, 2, 3)
plt.hist(df["petal_length"], color="yellow", edgecolor="red", alpha=0.8)
plt.title("Petal Length Distribution", color="red")

plt.subplot(3, 2, 4)
plt.hist(df["petal_width"], color="purple", edgecolor="red", alpha=0.8)
plt.title("Petal Width Distribution", color="red")

plt.subplot(3, 2, 5)
plt.hist(df["petal_width"], color="blue", edgecolor="red", alpha=0.8)
plt.title("Petal Width Distribution", color="red")

plt.tight_layout()
plt.savefig("flower-distribution")
```



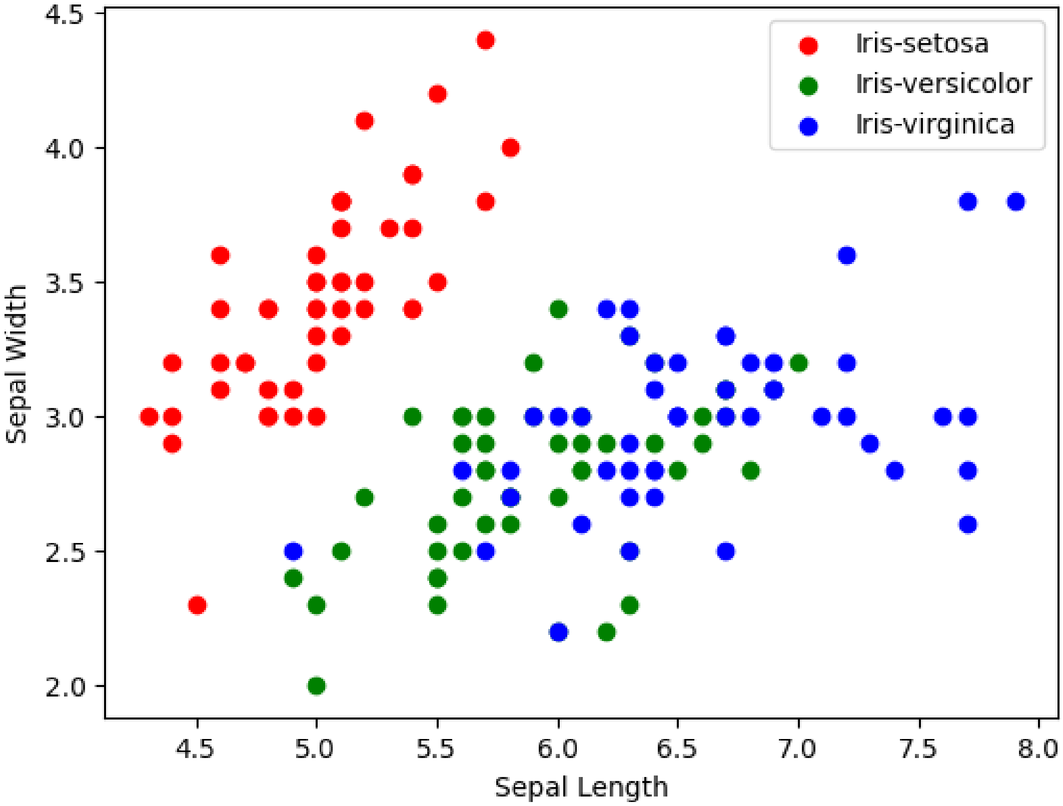
[72]:
```python
# Scatter plot of sepal length and sepal width

plt.scatter(df["sepal_length"], df["sepal_width"], color="red")
```

[72]:
```
<matplotlib.collections.PathCollection at 0x718310a0e5d0>
```



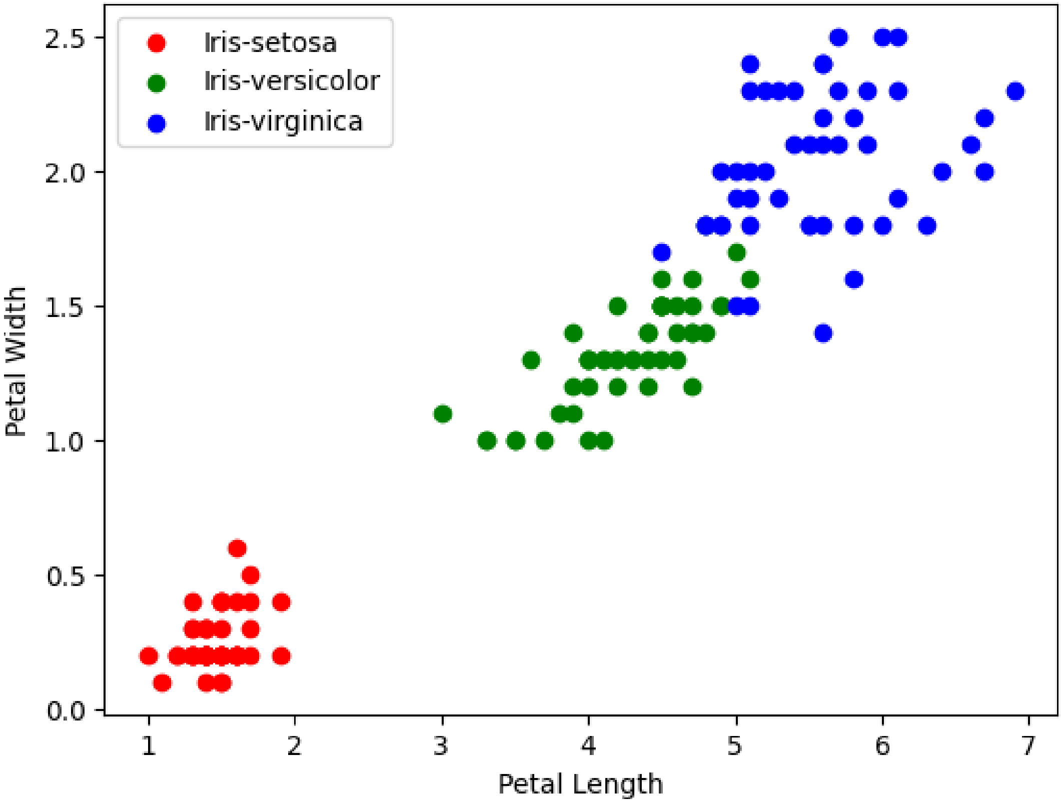
```
[79]: colors = ["red", "green", "blue"]
      species = ['Iris-setosa', 'Iris-versicolor', 'Iris-virginica']

      for i in range(3):
          data = df[df["species"] == species[i]]
          # print(data["species"].unique())
          plt.scatter(data["sepal_length"], data["sepal_width"], color=colors[i],␣
       ↪label=species[i])

      plt.xlabel("Sepal Length")
      plt.ylabel("Sepal Width")
      plt.legend()
```

```
[79]: <matplotlib.legend.Legend at 0x718310cdd950>
```



[80]:
```
colors = ["red", "green", "blue"]
species = ['Iris-setosa', 'Iris-versicolor', 'Iris-virginica']

for i in range(3):
    data = df[df["species"] == species[i]]
    # print(data["species"].unique())
    plt.scatter(data["petal_length"], data["petal_width"], color=colors[i],␣
 ↪label=species[i])

plt.xlabel("Petal Length")
plt.ylabel("Petal Width")
plt.legend()
```

[80]:
```
<matplotlib.legend.Legend at 0x718310bdaad0>
```



Iris-setosa
Iris-versicolor
Iris-virginica
Petal Width
Petal Length
0.0
0.5
1.0
1.5
2.0
2.5
1
2
3
4
5
6
7

```
[ ]:
```