05_Data Visualization

April 10, 2025

1 Data Visualization

Instructors: Jhun Brian M. Andam | Timothy Jonah E. Borromeo

Course: Introduction to Data Science

Objectives:

- Understand the necessary requirements for a data science task.
- Utilize and demonstrate the various data science tools.

[1]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import plotly.express as px
```

The Iris dataset contains 150 samples of iris flowers, each described by four features—sepal length, sepal width, petal length, and petal width—classified into three species: Setosa, Versicolor, and Virginica.

[3]:
```
iris = sns.load_dataset('iris')
iris.head()
```

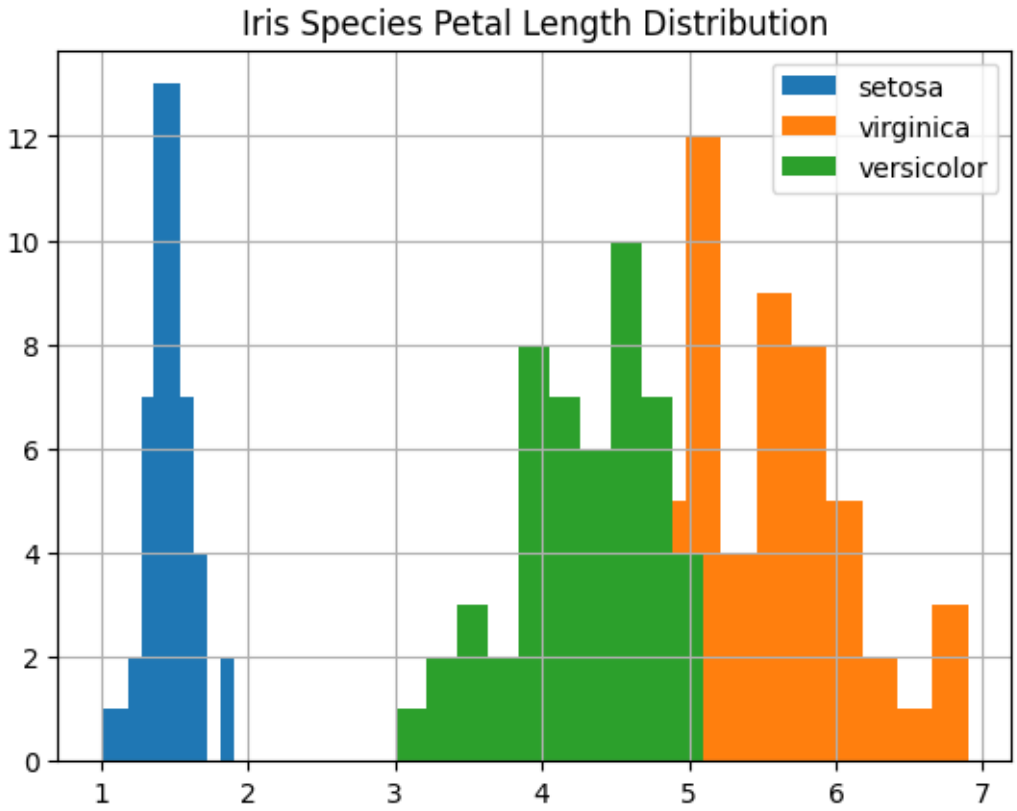
[3]:

	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

1. How do the petal lengths and widths vary across the three Iris species, and which feature shows the greatest separation between species?

[7]:
```
# how to know the unique species?

setosa = iris[iris['species'] == 'setosa']
virginica = iris[iris['species'] == 'virginica']
versicolor = iris[iris['species'] == 'versicolor']
```

[26]:
```
# how to access a specific column?

setosa['petal_length'].hist(label='setosa')
virginica['petal_length'].hist(label='virginica')
versicolor['petal_length'].hist(label='versicolor')
plt.legend()
plt.title('Iris Species Petal Length Distribution')
plt.show()
```



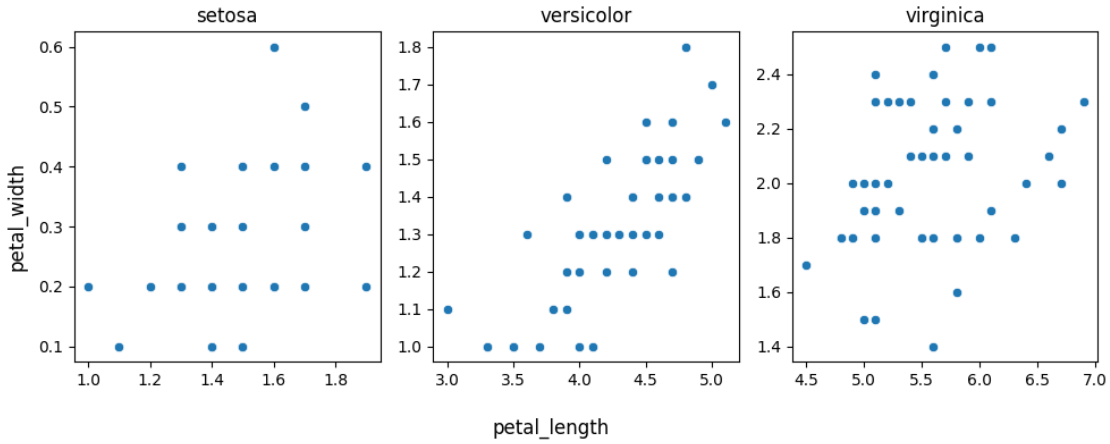
Can you think of other visualization for the petal length variation between species?

Visualize the `petal width` using the same visualization technique.

2. What is the relationship between petal length and petal width across the different Iris species, and how does this relationship differ among them?

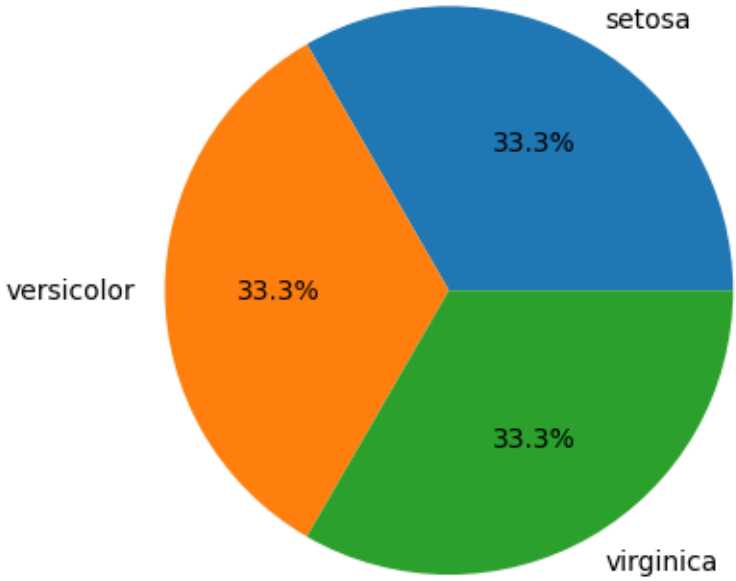
[66]:
```
fig, axs = plt.subplots(1,3, figsize=(10,4))
sns.scatterplot(data=setosa, x='petal_length', y='petal_width', ax=axs[0])
sns.scatterplot(data=versicolor, x='petal_length', y='petal_width', ax=axs[1])
sns.scatterplot(data=virginica, x='petal_length', y='petal_width', ax=axs[2])

for idx, ax in enumerate(axs):
```

```
    ax.set_xlabel('')
    ax.set_ylabel('')
    ax.set_title(label=['setosa', 'versicolor', 'virginica'][idx])

fig.supxlabel('petal_length')
fig.supylabel('petal_width')

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



What is the composition of the Iris dataset in terms of species, and how does the proportion of each species compare within the entire dataset?

[122]:
```
idx, vals = iris['species'].value_counts().index, iris['species'].
 ↪value_counts().values

plt.pie(x=vals, labels=idx, autopct='%1.1f%%');
```


setosa
33.3%
versicolor
33.3%
33.3%
virginica