Documentation on the Iris Dataset

Overview
The Iris dataset is one of the most well-known datasets in the field of machine learning and statistics. It was first introduced by the British biologist and statistician Ronald A. Fisher in 1936. The dataset consists of 150 samples of iris flowers from three different species: Iris setosa, Iris versicolor, and Iris virginica. Each sample includes four features: sepal length, sepal width, petal length, and petal width, all measured in centimeters.
Structure of the Dataset
The dataset comprises five columns:
1. Sepal Length (cm)
2. Sepal Width (cm)
3. Petal Length (cm)
4. Petal Width (cm)
5. Species: This categorical column indicates the species of the iris flower and can be one of three values: 'setosa', 'versicolor', or 'virginica'.
Source
The Iris dataset is publicly available and can be downloaded from the UCI Machine Learning Repository:
UCI Machine Learning Repository - Iris Dataset

Libraries and Data Loading
First, we import the necessary libraries and load the dataset.

[1]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
```

loading the file & Understanding Data

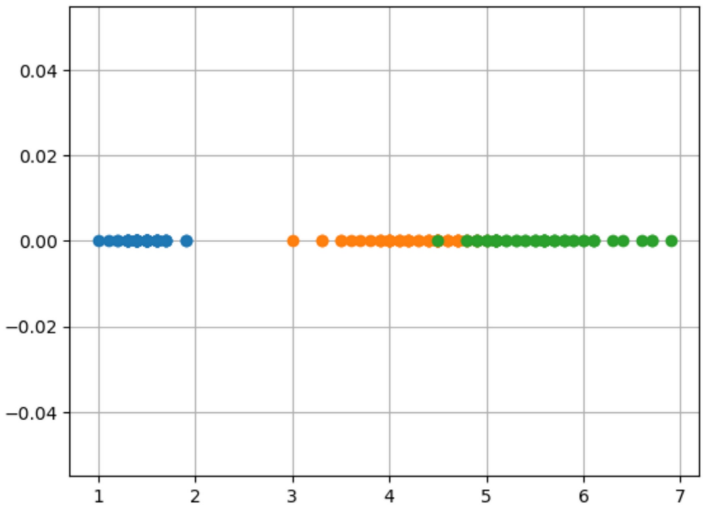
[2]:
```
iris = pd.read_csv("iris.csv")
```

[3]:
```
print(iris.shape)
```
```
(150, 5)
```

[5]:
```
print(iris.columns)
```
```
Index(['sepal_length', 'sepal_width', 'petal_length', 'petal_width',
       'species'],
      dtype='object')
```

1D Scatter plot

[28]:
```
iris_setso = iris.loc[iris["species"] == "setosa"];
iris_virginica = iris.loc[iris["species"] == "virginica"];
iris_versicolor = iris.loc[iris["species"] == "versicolor"];
plt.plot(iris_setso["petal_length"],np.zeros_like(iris_setso["petal_length"]), 'o')
plt.plot(iris_versicolor["petal_length"],np.zeros_like(iris_versicolor["petal_length"]), 'o')
plt.plot(iris_virginica["petal_length"],np.zeros_like(iris_virginica["petal_length"]), 'o')
plt.grid()
plt.show()
```



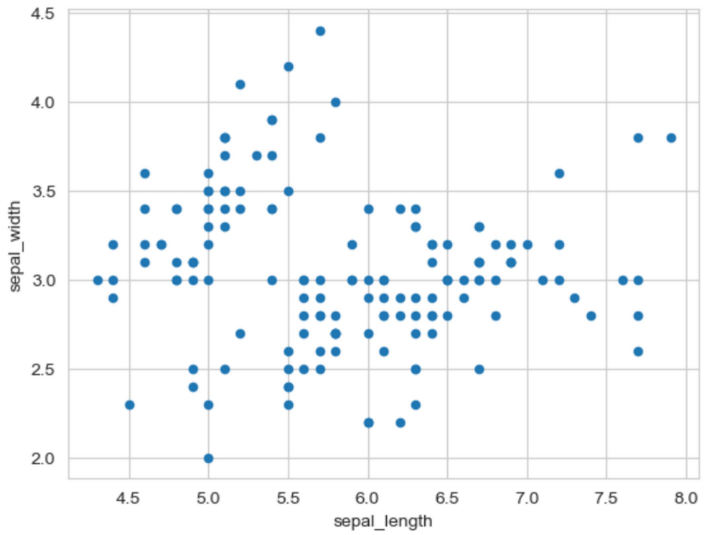
Observation() | Conclusion

Green points are Virginica, orange points are Versicolor and blue points are Setosa
Virginica and Versicolor are overlapping

1D Scatter are very hard to read and understand

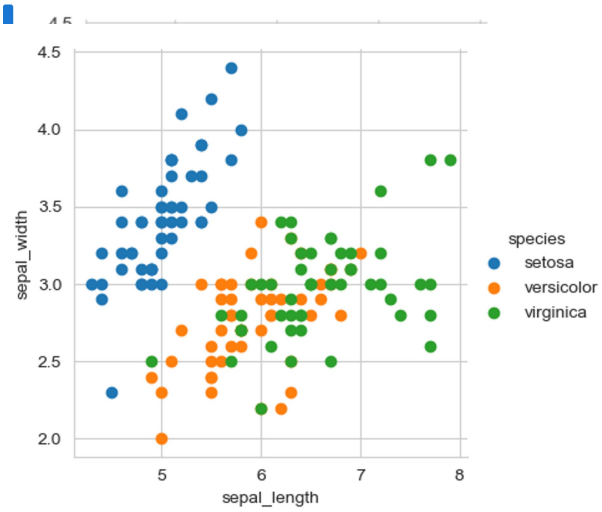
2D scatter plot

[37]:
```
iris.plot(kind="scatter",x="sepal_length",y="sepal_width")
plt.show()
```



In the above figure, we are't able to understand which is setosa or versicolor or virginica flower because all points are in same colour. It cannot make much sense out it

[9]:
```
sns.set_style("whitegrid")

# Create the FacetGrid
g = sns.FacetGrid(iris, hue="species", height=4)
g = g.map(plt.scatter, "sepal_length", "sepal_width").add_legend()

# Show the plot
plt.show()
```



Observation(s) | Conclusion

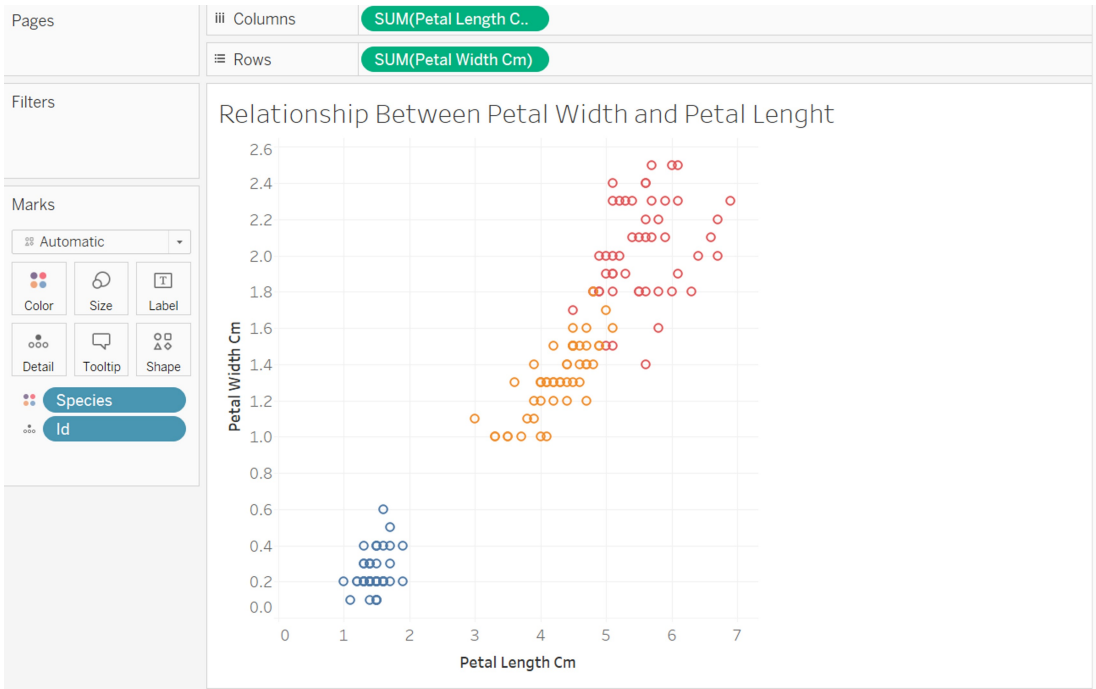
Blue points can be easily separated from red and green by drawing a line.
But red and green data points cannot be easily separated.
Using sepal_length and sepal_width features, we can distinguish Setosa flowers from others.
Separating Versicolor from Viginica is much harder as they have considerable overlap.

```
# Set the style
sns.set_style("whitegrid")

# Create the pairplot
sns.pairplot(iris, hue="species", height=3)

# Show the plot
plt.show()
```

Observation(s) | Conclusion

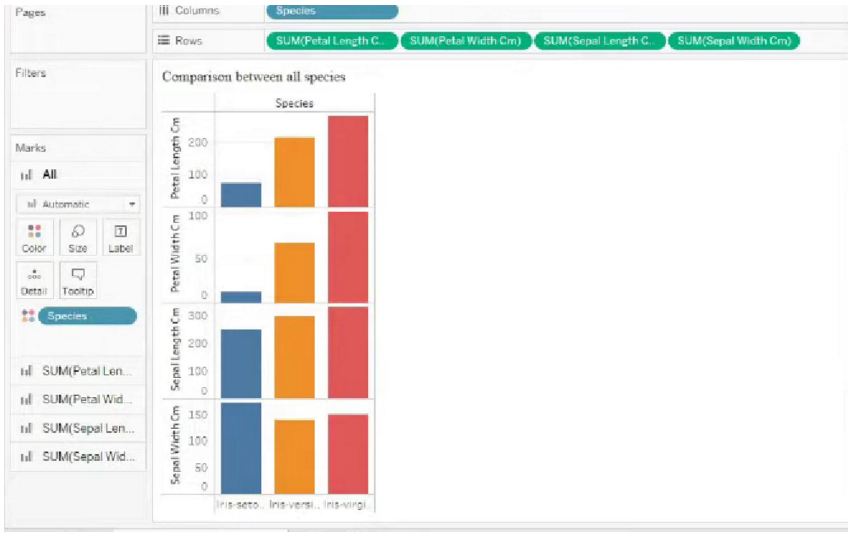
petal length and petal width are the most useful features to identify various flower types.
While Setosa can be easily identified (linearly separable), virginica and Versicolor have some overlap (almost linearly separable).
We can find “lines” and “if-else” conditions to build a simple model to classify the flower types.

Cluster Visualization: Petal Length vs. Petal Width



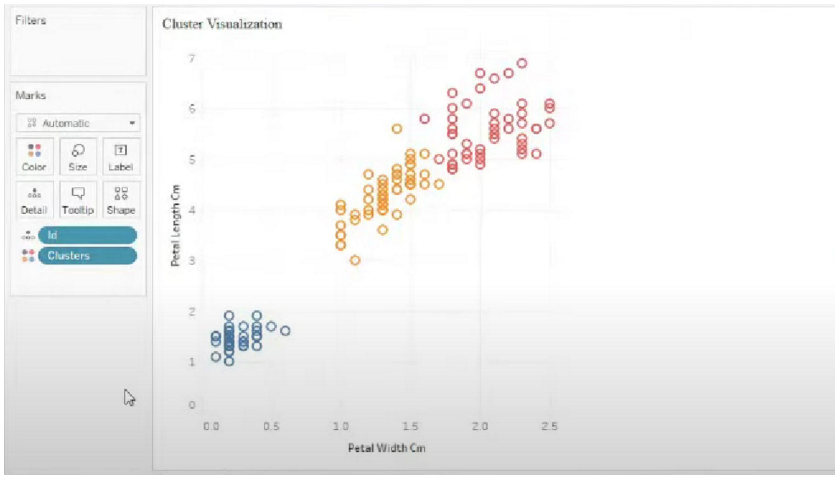
Description: This visualization applies clustering algorithms (like k-means) to group the data points based on their petal length and width.
Conclusion:
Clusters Formed: Three clusters are typically formed, each corresponding to one species.
Iris Setosa: Forms a distinct cluster, easily separable from the others due to its unique petal dimensions.
Iris Versicolor and Virginica: These species form two other clusters, which may have some overlap but are distinguishable based on their petal measurements.

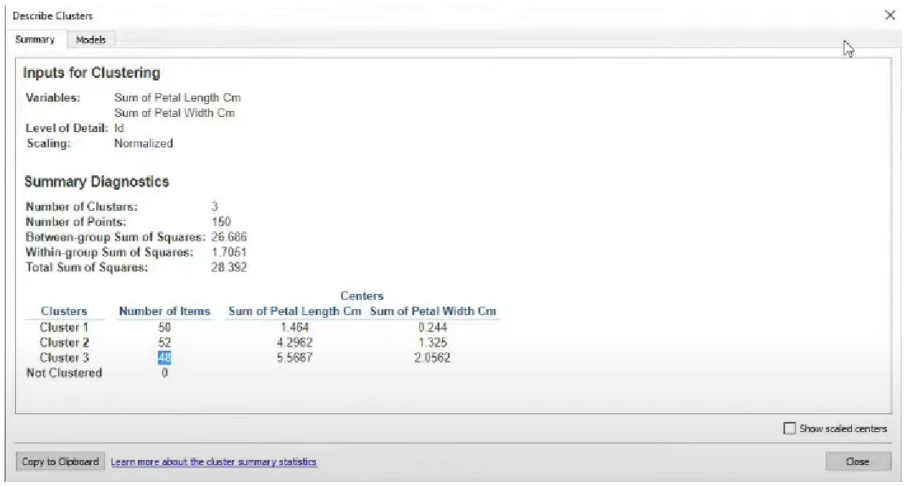
Bar Chart: Species Count



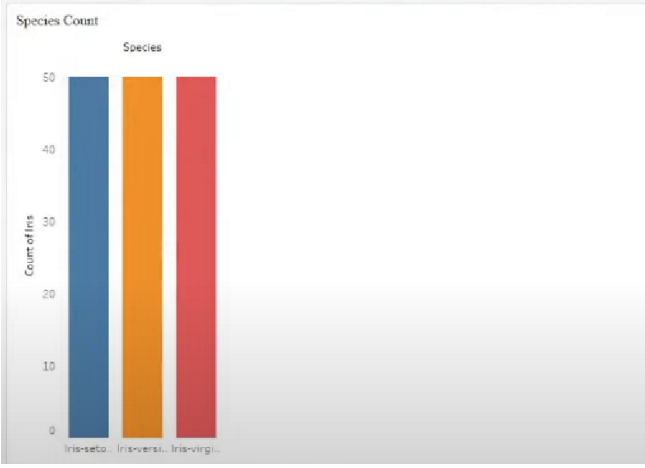
Description: This chart shows the count of each species in the dataset, with separate bars for Setosa, Versicolor, and Virginica.
Conclusion:
Distribution: Highlights the balanced distribution of the species in the dataset, showing how many samples of each species are present.
Insight: Provides a quick view of the sample size for each species, ensuring equal representation for analysis.

Scatter Plot: Sepal Length vs. Sepal Width



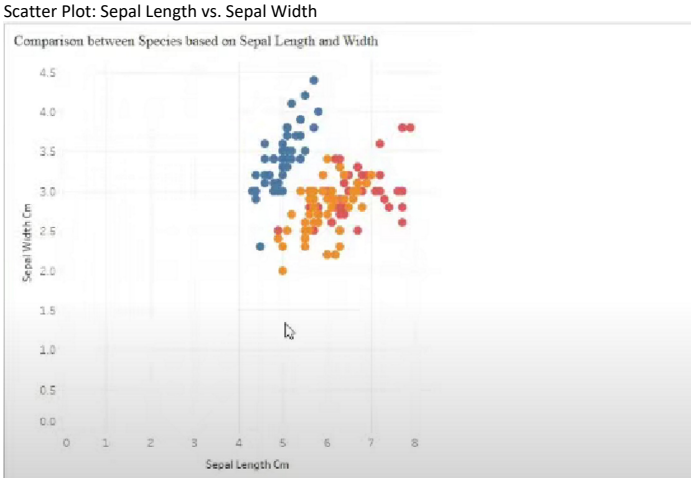


Bar Chart: Species Count

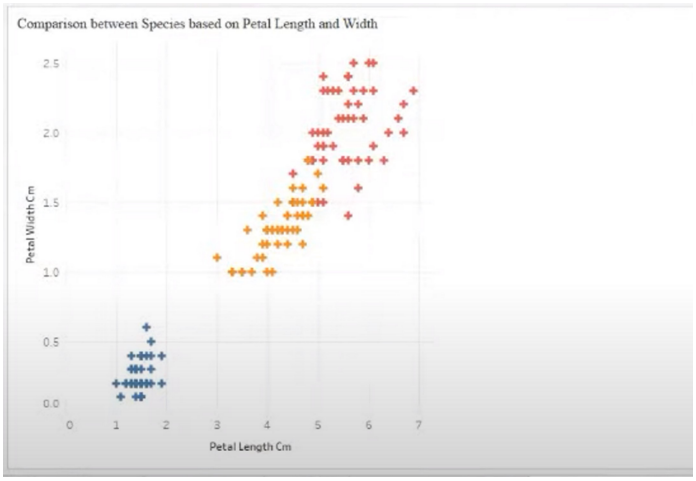


Description: This chart shows the count of each species in the dataset, with separate bars for Setosa, Versicolor, and Virginica.
Conclusion:
Distribution: Highlights the balanced distribution of the species in the dataset, showing how many samples of each species are present.
Insight: Provides a quick view of the sample size for each species, ensuring equal representation for analysis.

Scatter Plot: Sepal Length vs. Sepal Width



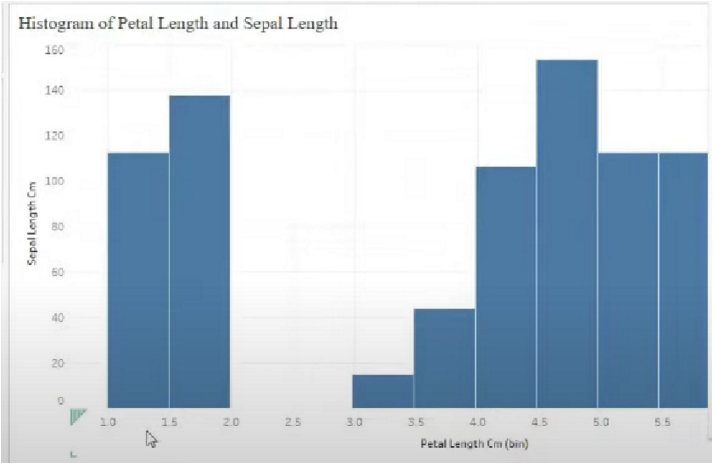
Description: This plot shows sepal length on the x-axis and sepal width on the y-axis, with each point representing a flower.
Conclusion:
Iris Setosa: Again, easily distinguishable with larger sepal width and smaller sepal length, forming a distinct cluster.
Iris Versicolor and Virginica: These species overlap significantly in this plot, making them harder to separate based on sepal dimensions alone. However, Virginica tends to have larger sepal lengths and smaller widths compared to Versicolor.
Cluster Visualization: Sepal Length vs. Sepal Width



Description: This visualization applies clustering algorithms to group the data points based on their sepal length and width.

Conclusion:
Clusters Formed: Three clusters, each generally corresponding to one species.
Iris Setosa: Forms a distinct cluster due to its unique sepal dimensions.
Iris Versicolor and Virginica: These species form clusters with significant overlap, indicating that sepal measurements alone may not be sufficient for clear separation.

Histogram: Petal Length



Description: This histogram shows the frequency distribution of petal length.
Conclusion:
Iris Setosa: Has a distinct distribution with petal lengths clustered around lower values.
Iris Versicolor and Virginica: These species show overlapping distributions, with Versicolor having intermediate petal lengths and Virginica having larger petal lengths.
Histogram: Petal Width

Description: This histogram shows the frequency distribution of petal width.
Conclusion:
Iris Setosa: Displays a distinct cluster around lower petal widths.
Iris Versicolor and Virginica: Show overlapping distributions, with Versicolor having intermediate petal widths and Virginica having larger widths.
Summary of Separation
Iris Setosa: Clearly separable from the other two species based on both petal and sepal dimensions. It forms distinct clusters and distributions with smaller petal lengths and widths and larger sepal widths.
Iris Versicolor: Overlaps with Virginica but generally has intermediate values for petal and sepal dimensions.
Iris Virginica: Overlaps with Versicolor but typically has larger petal and sepal dimensions.
By using these visualizations, we can see that Iris Setosa is the easiest to separate, while distinguishing between Iris Versicolor and Iris Virginica requires more nuanced analysis, potentially involving multiple dimensions or advanced classification techniques.