Exploratory Data Analysis (EDA) Report on Iris Dataset

Project Overview

Project ID: CC69855

Project Title: Exploratory Data Analysis (EDA) on Iris Dataset

Internship Domain: Data Science Intern

Project Level: Entry Level

Assigned By: CodeClause Internship

Name: Omar Samir Mohamed

Aim

The aim of this project is to conduct exploratory data analysis on the Iris dataset to understand its characteristics and relationships between features.

Dataset Description

The Iris dataset consists of 150 observations of iris flowers. There are five attributes in the dataset:

1. **sepal.length**: Length of the sepal in cm.
2. **sepal.width**: Width of the sepal in cm.
3. **petal.length**: Length of the petal in cm.
4. **petal.width**: Width of the petal in cm.
5. **variety**: Species of the iris flower (Setosa, Versicolor, Virginica).

Data Loading and Inspection

Code:

```
import pandas as pd
import matplotlib as plt
import seaborn as sns
iris = pd.read_csv("iris.csv")
print(iris.head())
```

Output:

	sepal.length	sepal.width	petal.length	petal.width	variety
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

Missing Values and Statistical Summary

Code:

```
missing_values = iris.isnull().sum()

print(missing_values)

statistical_summary = iris.describe()

print(statistical_summary)
```

Output:

```
sepal.length    0
sepal.width     0
petal.length    0
petal.width     0
variety         0
dtype: int64
       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
```

Observations:

- There are no missing values in the dataset.
- The mean values for each feature are as follows:
    1. Sepal Length: 5.84 cm
    2. Sepal Width: 3.06 cm
    3. Petal Length: 3.76 cm
    4. Petal Width: 1.20 cm

Data Visualization

Histograms and Box Plots

Code:

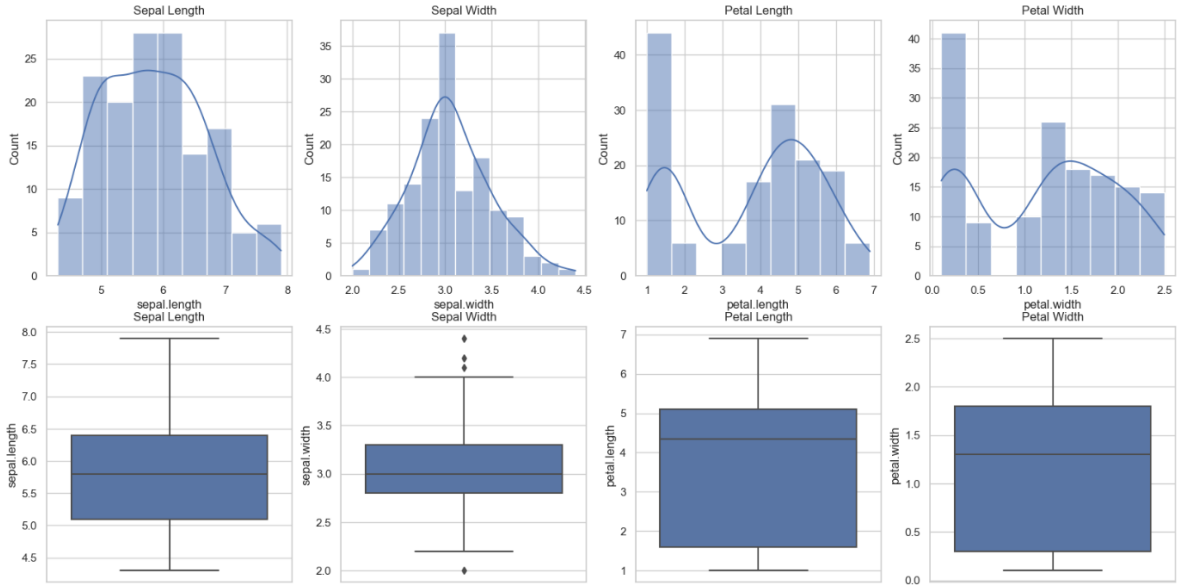
```
sns.set(style="whitegrid")


fig, axes = plt.subplots(2, 4, figsize=(20, 10))


sns.histplot(iris['sepal.length'], kde=True, ax=axes[0, 0]).set_title('Sepal Length')
sns.histplot(iris['sepal.width'], kde=True, ax=axes[0, 1]).set_title('Sepal Width')
sns.histplot(iris['petal.length'], kde=True, ax=axes[0, 2]).set_title('Petal Length')
sns.histplot(iris['petal.width'], kde=True, ax=axes[0, 3]).set_title('Petal Width')


sns.boxplot(y=iris['sepal.length'], ax=axes[1, 0]).set_title('Sepal Length')
sns.boxplot(y=iris['sepal.width'], ax=axes[1, 1]).set_title('Sepal Width')
sns.boxplot(y=iris['petal.length'], ax=axes[1, 2]).set_title('Petal Length')
sns.boxplot(y=iris['petal.width'], ax=axes[1, 3]).set_title('Petal Width')
```

Results:



Observations:

1. **Sepal Length:**
    - Distribution is slightly skewed to the right.
    - Range: 4.3 to 7.9
    - Median around 5.8
    - Few potential outliers
2. **Sepal Width:**
    - Distribution is approximately normal but with a slight left skew.
    - Range: 2.0 to 4.4
    - Median around 3.0
    - A few outliers, particularly at the lower end
3. **Petal Length:**
    - Distribution appears bimodal.
    - Range: 1.0 to 6.9
    - Median around 4.35
    - Some potential outliers
4. **Petal Width:**
    - Distribution appears bimodal.
    - Range: 0.1 to 2.5
    - Median around 1.3
    - Some potential outliers

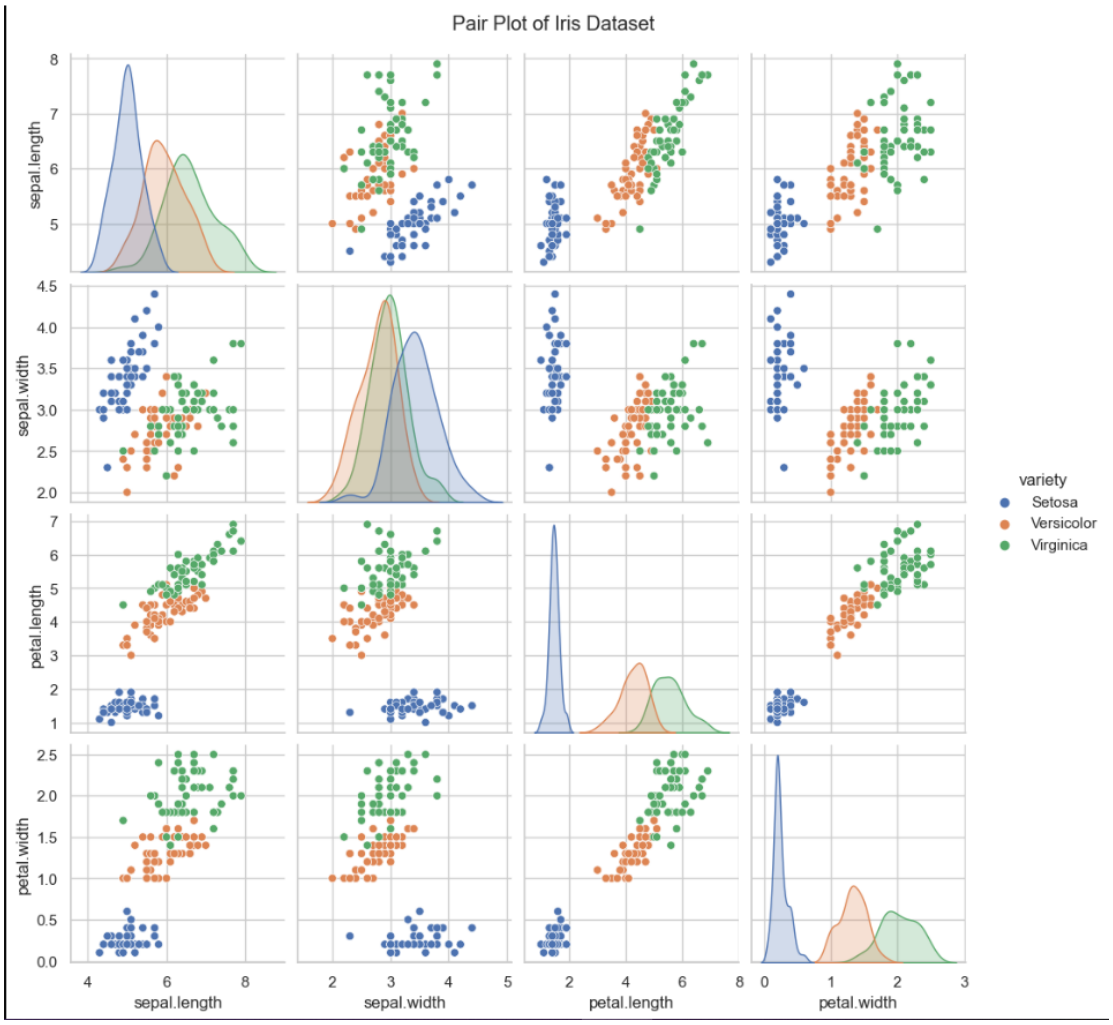
Pair Plot

Code:

```
pair_plot = sns.pairplot(iris, hue='variety', diag_kind='kde')

pair_plot.fig.suptitle('Pair Plot of Iris Dataset', y=1.02)

plt.show()
```



Visualizations:

- Pair plots to observe relationships between features and species.

Observations:

- There are clear separations between species based on the petal length and width.
- Setosa species are distinctly separated from Versicolor and Virginica in the pair plots.
- Versicolor and Virginica overlap more but can still be differentiated based on the combination of features.

Conclusion

This EDA has helped us better grasp the distributions and linkages between the attributes in the Iris dataset. Petal measurements, in particular, helped us to identify distinctive patterns that distinguish the species. This study can serve as the foundation for other modelling and classification tasks.