HexSoftware

Project 1

Exploratory Data Analysis of the Iris Dataset

By Patel Aatif

21/8/2024

Introduction

The Iris dataset is one of the most well-known datasets in the field of data science and machine learning. Collected by British biologist and statistician Ronald A. Fisher, it contains measurements of sepal length, sepal width, petal length, and petal width for 150 flower samples, distributed evenly across three species: Iris-setosa, Iris-versicolor, and Iris-virginica. This dataset is often used as a benchmark for testing and demonstrating machine learning algorithms. In this document, we will conduct a thorough exploratory data analysis (EDA) to uncover insights, patterns, and relationships within the data.

Objective

1. Understand the distribution and characteristics of the various features (sepal length, sepal width, petal length, and petal width) across the three Iris species.
2. Identify any correlations or relationships between the features that may be indicative of the differences between species.
3. Highlight any potential outliers or anomalies in the data that may require further investigation.
4. Lay the groundwork for future predictive modeling or classification tasks by understanding the underlying structure of the dataset.

Question

Basic Statistics:

What is the mean of the Sepal Length for the dataset?
What is the median of the Petal Width for the dataset?
What is the maximum Sepal Width value in the dataset?

Data Selection and Filtering:

How many entries belong to the species "Iris-setosa"?
How many records have a Sepal Length greater than 5.0 cm?
What are the records where the Petal Length is less than 1.5 cm?

Grouping and Aggregation:

What is the average Sepal Length for each species?
What is the total count of each species in the dataset?
What is the average Petal Width for the species "Iris-virginica"?

Question

Data Visualization:

Can you create a histogram of the Sepal Length for the dataset?
Can you plot a scatter plot between Sepal Length and Sepal Width to visualize the relationship?
Can you create a box plot to compare the Petal Length across the different species?

Data Cleaning:

Are there any missing values in the dataset? If yes, which columns contain them?
What would be the effect of filling missing values in Sepal Length with the mean of the column?

Advanced Selection:

Can you find the average Sepal Width for the species that have a Petal Length greater than 4.0 cm?
How many records have Sepal Length greater than 5.0 cm and belong to "Iris-versicolor"?

Data Aggregation with GroupBy:

For each species, calculate the average Petal Length and Petal Width. Then, identify which species has the highest average Petal Length-to-Width ratio.

Steps

Step 1 Load Library

```
[72]
import pandas as pd
import numpy as np
import seaborn as sns
import os
import matplotlib.pyplot as plt
```

Step 2 Load DataSet

```
[73]
df = pd.read_csv(r"C:\Users\Lord\Downloads\archive\iris.csv")
```

```
[74]
df.head()
```

	5.1	3.5	1.4	0.2	Iris-setosa
0	4.9	3.0	1.4	0.2	Iris-setosa
1	4.7	3.2	1.3	0.2	Iris-setosa
2	4.6	3.1	1.5	0.2	Iris-setosa
3	5.0	3.6	1.4	0.2	Iris-setosa
4	5.4	3.9	1.7	0.4	Iris-setosa

Step 3 Understanding DataSets

```
[6]
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 149 entries, 0 to 148
Data columns (total 5 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   5.1          149 non-null    float64
 1   3.5          149 non-null    float64
 2   1.4          149 non-null    float64
 3   0.2          149 non-null    float64
 4   Iris-setosa  149 non-null    object
dtypes: float64(4), object(1)
memory usage: 5.9+ KB
```

Steps

```
[7]  df.describe()
```

	5.1	3.5	1.4	0.2
count	149.000000	149.000000	149.000000	149.000000
mean	5.848322	3.051007	3.774497	1.205369
std	0.828594	0.433499	1.759651	0.761292
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.400000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

```
[8]  df.shape
```

```
(149, 5)
```

Step 4 Check Null Value

```
[83]  df.isnull().sum()
```

```
SepalLengthCm    0
SepalWidthCm     0
PetalLengthCm    0
PetalWidthCm     0
Species          0
dtype: int64
```

We Dont have any Null Values

```
[84]  df.nunique()
```

```
SepalLengthCm    35
SepalWidthCm     23
PetalLengthCm    43
PetalWidthCm     22
Species           3
dtype: int64
```

Questions

1. Basic Statistics

What is the mean of the Sepal Length for the dataset?

```
[98] df['SepalLengthCm'].mean()
```

```
5.8483221476510066
```

What is the median of the Petal Width for the dataset?

```
[99] df['SepalWidthCm'].mean()
```

```
3.051006711409396
```

What is the maximum Sepal Width value in the dataset?

```
[100] df['SepalLengthCm'].max()
```

```
7.9
```

2. Data Selection and Filtering:

How many entries belong to the species "Iris-setosa"?

```
[101] df['Species'].value_counts()
```

```
Species
Iris-versicolor    50
Iris-virginica     50
Iris-setosa        49
Name: count, dtype: int64
```

How many records have a Sepal Length greater than 5.0 cm?

```
[102] df[df['SepalLengthCm'] > 5.0]
```

	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
4	5.4	3.9	1.7	0.4	Iris-setosa
9	5.4	3.7	1.5	0.2	Iris-setosa
13	5.8	4.0	1.2	0.2	Iris-setosa
14	5.7	4.4	1.5	0.4	Iris-setosa
15	5.4	3.9	1.3	0.4	Iris-setosa
...	...	...	...	...	...
144	6.7	3.0	5.2	2.3	Iris-virginica
145	6.3	2.5	5.0	1.9	Iris-virginica
146	6.5	3.0	5.2	2.0	Iris-virginica
147	6.2	3.4	5.4	2.3	Iris-virginica
148	5.9	3.0	5.1	1.8	Iris-virginica

117 rows × 5 columns

Questions

What are the records where the Petal Length is less than 1.5 cm?

```
[103] df[df['PetalLengthCm'] < 1.5]
```

	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	4.9	3.0	1.4	0.2	Iris-setosa
1	4.7	3.2	1.3	0.2	Iris-setosa
3	5.0	3.6	1.4	0.2	Iris-setosa
5	4.6	3.4	1.4	0.3	Iris-setosa
7	4.4	2.9	1.4	0.2	Iris-setosa
11	4.8	3.0	1.4	0.1	Iris-setosa
12	4.3	3.0	1.1	0.1	Iris-setosa
13	5.8	4.0	1.2	0.2	Iris-setosa
15	5.4	3.9	1.3	0.4	Iris-setosa
16	5.1	3.5	1.4	0.3	Iris-setosa
21	4.6	3.6	1.0	0.2	Iris-setosa
27	5.2	3.4	1.4	0.2	Iris-setosa
32	5.5	4.2	1.4	0.2	Iris-setosa
34	5.0	3.2	1.2	0.2	Iris-setosa
35	5.5	3.5	1.3	0.2	Iris-setosa
37	4.4	3.0	1.3	0.2	Iris-setosa
39	5.0	3.5	1.3	0.3	Iris-setosa
40	4.5	2.3	1.3	0.3	Iris-setosa
41	4.4	3.2	1.3	0.2	Iris-setosa
44	4.8	3.0	1.4	0.3	Iris-setosa
46	4.6	3.2	1.4	0.2	Iris-setosa
48	5.0	3.3	1.4	0.2	Iris-setosa

3. Grouping and Aggregation:

What is the average Petal Width for the species "Iris-virginica"?

```
[12] df.groupby('Species')['PetalWidthCm'].mean().loc['Iris-virginica']
```

```
2.026
```

What is the total count of each species in the dataset?

```
[88] df['Species'].value_counts()
```

```
Species
Iris-versicolor    50
Iris-virginica     50
Iris-setosa        49
Name: count, dtype: int64
```

What is the average Sepal Length for each species?

```
] df.groupby('Species')['SepalLengthCm'].mean()
```

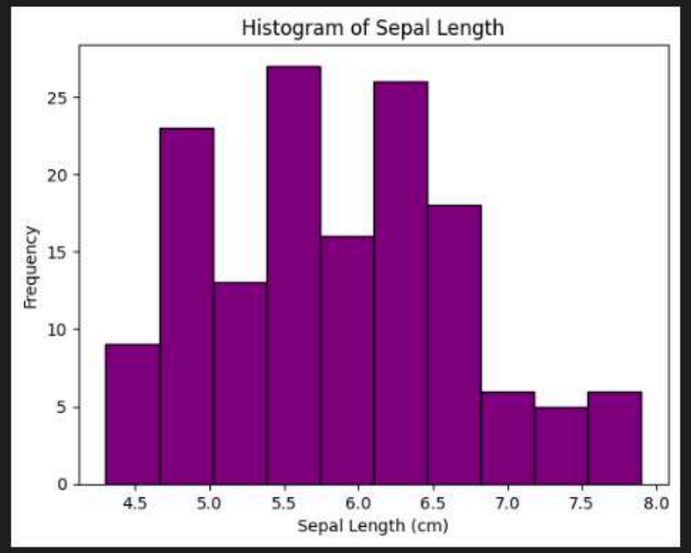
Questions

4. Data Visualization:

Can you create a histogram of the Sepal Length for the dataset?

```
plt.hist(df['SepalLengthCm'], bins=10, color='purple', edgecolor='black')
plt.xlabel('Sepal Length (cm)')
plt.ylabel('Frequency')
plt.title('Histogram of Sepal Length')
```

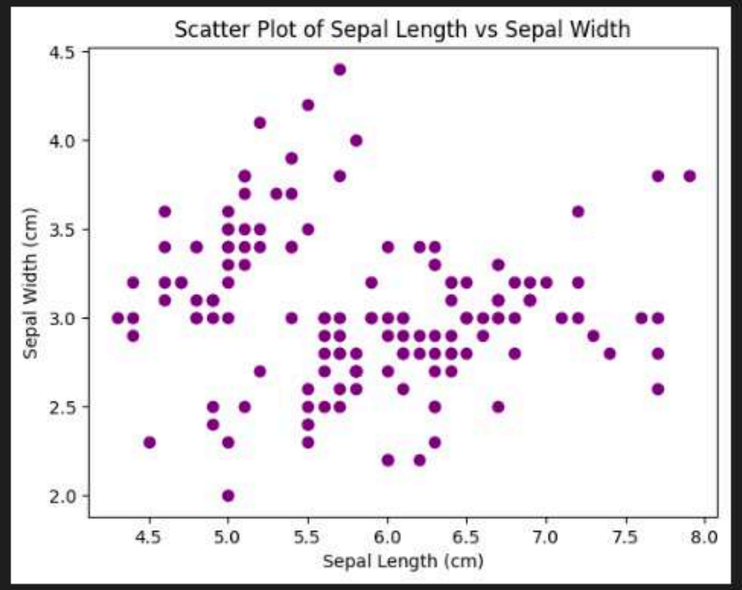
```
Text(0.5, 1.0, 'Histogram of Sepal Length')
```



Can you plot a scatter plot between Sepal Length and Sepal Width to visualize the relationship?

```
plt.scatter(df['SepalLengthCm'],df['SepalWidthCm'],color = "Purple")
plt.xlabel('Sepal Length (cm)')
plt.ylabel('Sepal Width (cm)')
plt.title('Scatter Plot of Sepal Length vs Sepal Width')
```

```
Text(0.5, 1.0, 'Scatter Plot of Sepal Length vs Sepal Width')
```

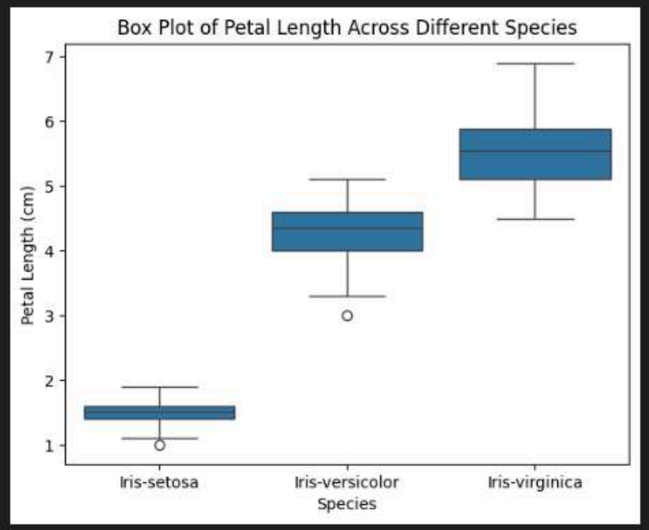


Questions

Can you create a box plot to compare the Petal Length across the different species?

```
sns.boxplot(x='Species', y='PetalLengthCm', data=df)
plt.xlabel('Species')
plt.ylabel('Petal Length (cm)')
plt.title('Box Plot of Petal Length Across Different Species')
```

```
Text(0.5, 1.0, 'Box Plot of Petal Length Across Different Species')
```



Questions

6. Advanced Selection:

Can you find the average Sepal Width for the species that have a Petal Length greater than 4.0 cm?

```
filtered_df = df[df['PetalLengthCm'] > 4.0]

average_sepal_width = filtered_df['SepalWidthCm'].mean()

average_sepal_width
```

[41] ✓ 0.1s

```
2.9452380952380954
```

How many records have Sepal Length greater than 5.0 cm and belong to "Iris-versicolor"?

```
filtered_records = df[(df['SepalLengthCm'] > 5.0) & (df['Species'] == 'Iris-versicolor')]

len(filtered_records)
```

[42] ✓ 0.1s

```
47
```

Questions

7. Data Aggregation with GroupBy:

For each species, calculate the average Petal Length and Petal Width. Then, identify which species has the highest average Petal Length-to-Width ratio.

```
average_petal_measures = df.groupby('Species').agg({'PetalLengthCm': 'mean', 'PetalWidthCm': 'mean'})
average_petal_measures
```

Species	PetalLengthCm	PetalWidthCm
Iris-setosa	1.465306	0.244898
Iris-versicolor	4.260000	1.326000
Iris-virginica	5.552000	2.026000

```
average_petal_measures['PetalLength_to_Width_Ratio'] = average_petal_measures['PetalLengthCm'] / average_petal_measures['PetalWidthCm']
average_petal_measures['PetalLength_to_Width_Ratio']
```

```
Species
Iris-setosa        5.983333
Iris-versicolor    3.212670
Iris-virginica     2.740375
Name: PetalLength_to_Width_Ratio, dtype: float64
```

```
highest_ratio_species = average_petal_measures['PetalLength_to_Width_Ratio'].idxmax()
highest_ratio_species
```

```
'Iris-setosa'
```

Thank You

patelaatif560@gmail.com

9321880477