Iris Dataset Basic Analysis

Overview

The Iris dataset is a classic multivariate dataset introduced by the British statistician and biologist Ronald Fisher in his 1936 paper "The use of multiple measurements in taxonomic problems". It contains 150 samples from three species of Iris flowers (Iris setosa, Iris virginica, and Iris versicolor). Each sample is described by 4 features: sepal length, sepal width, petal length, and petal width, all measured in centimeters.

Methodology

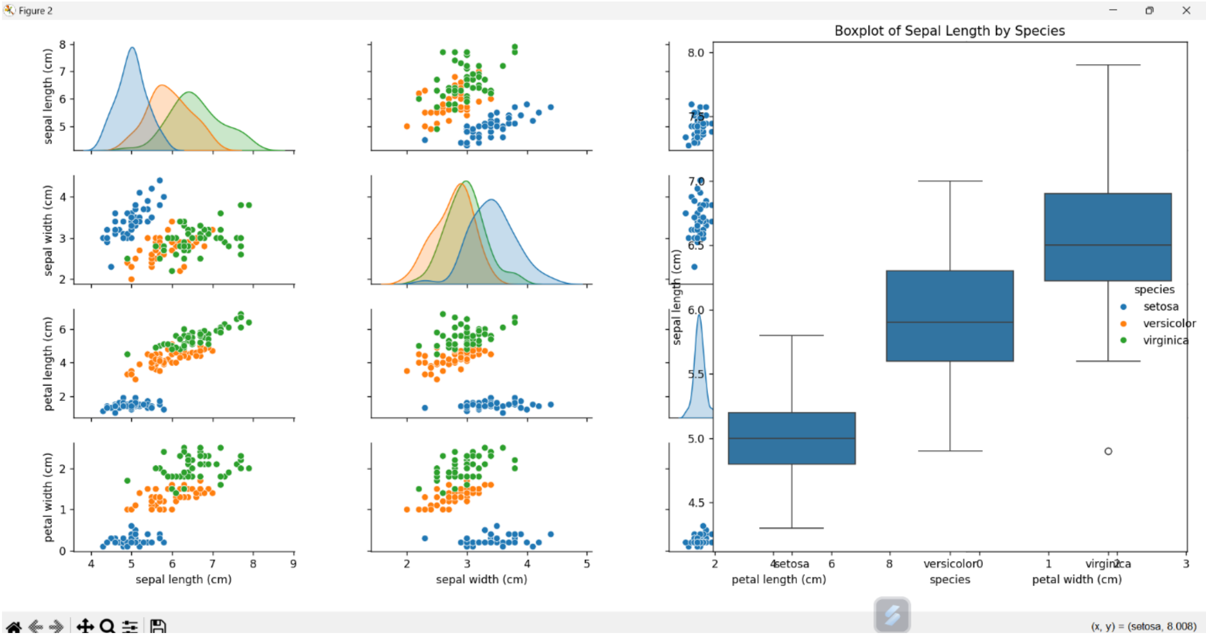
1. **Import necessary libraries**: pandas, numpy, matplotlib, and seaborn.
2. **Load the dataset**: Use pd.read_csv() to load the Iris dataset from a CSV file or use a built-in dataset from a library like scikit-learn.
3. **Basic Information**: Use df.info() to display basic information about the dataset, including the number of entries, data columns, and memory usage.
4. **Descriptive Statistics**: Use df.describe() to generate descriptive statistics for each feature, including count, mean, standard deviation, minimum, 25th percentile, median (50th percentile), 75th percentile, and maximum.
5. **Data Visualization**: Use matplotlib and seaborn to visualize the distribution of each feature and relationships between features.

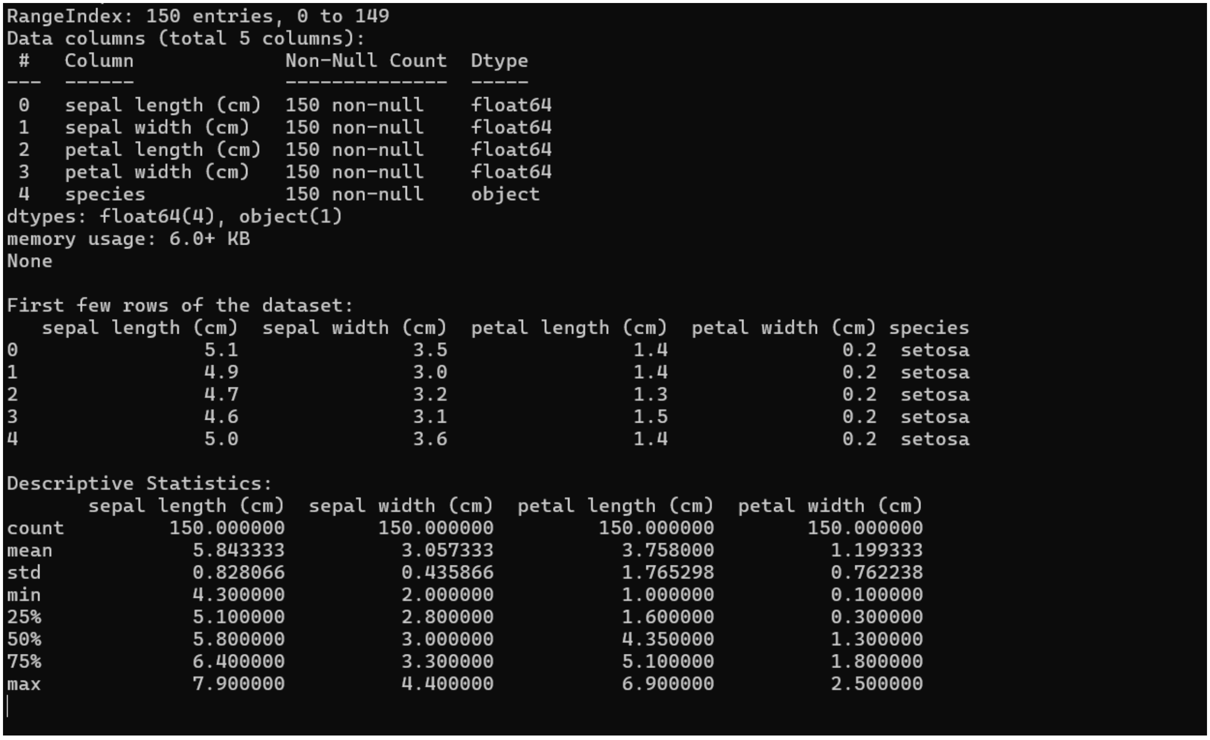
Analysis

Based on the output provided:

- The dataset has 150 entries, with 4 float columns (sepal length, sepal width, petal length, and petal width) and 1 object column (species).
- The descriptive statistics show:
  - Sepal length: mean = 5.84 cm, standard deviation = 0.83 cm, range = 4.3-7.9 cm
  - Sepal width: mean = 3.06 cm, standard deviation = 0.44 cm, range = 2-4.4 cm
  - Petal length: mean = 3.76 cm, standard deviation = 1.77 cm, range = 1-6.9 cm
  - Petal width: mean = 1.20 cm, standard deviation = 0.76 cm, range = 0.1-2.5 cm
- The dataset appears to be well-balanced, with no missing values.

Screenshots

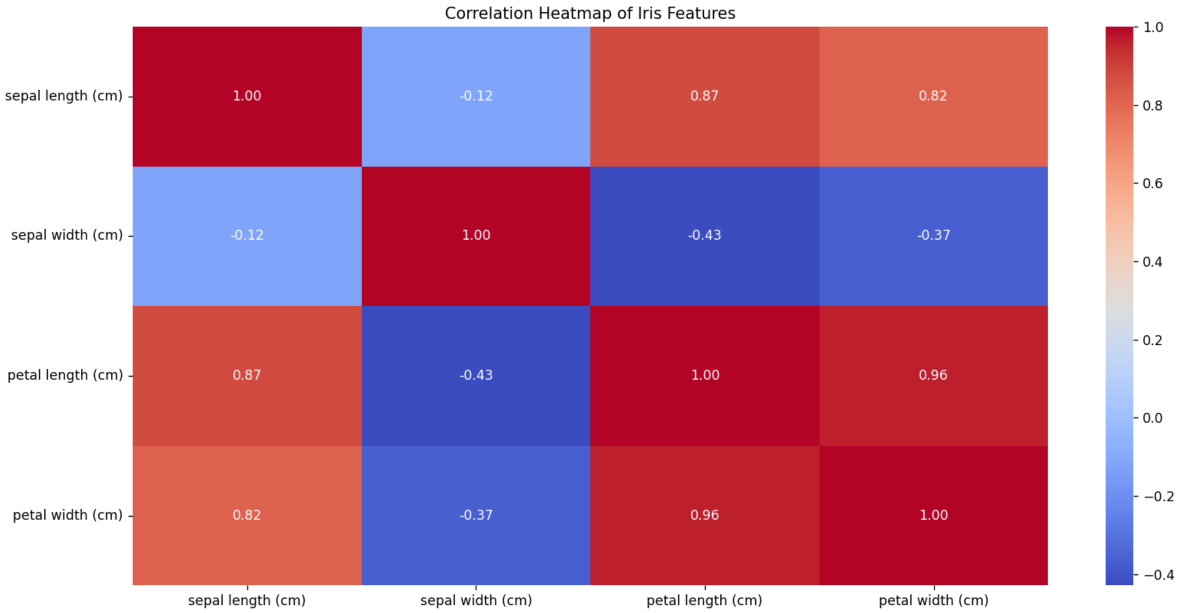




```
RangeIndex: 150 entries, 0 to 149
Data columns (total 5 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   sepal length (cm)  150 non-null    float64
 1   sepal width (cm)   150 non-null    float64
 2   petal length (cm)  150 non-null    float64
 3   petal width (cm)   150 non-null    float64
 4   species            150 non-null    object
dtypes: float64(4), object(1)
memory usage: 6.0+ KB
None

First few rows of the dataset:
   sepal length (cm)  sepal width (cm)  petal length (cm)  petal width (cm) 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

Descriptive Statistics:
       sepal length (cm)  sepal width (cm)  petal length (cm)  petal width (cm)
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:

1. High Positive Correlations:

- **Sepal Length and Petal Length:** A strong positive correlation of 0.87 suggests that as sepal length increases, petal length tends to increase as well.
- **Sepal Length and Petal Width:** A similarly strong positive correlation of 0.82 indicates a similar trend between sepal length and petal width.
- **Petal Length and Petal Width:** The highest correlation of 1.00 shows a perfect positive relationship, meaning petal length and petal width are highly interdependent.

2. Negative Correlation:

- **Sepal Width and Petal Length:** A moderate negative correlation of -0.43 suggests that as sepal width increases, petal length tends to decrease slightly.
- **Sepal Width and Petal Width:** A similar negative correlation of -0.37 exists between sepal width and petal width.

3. No Correlation:

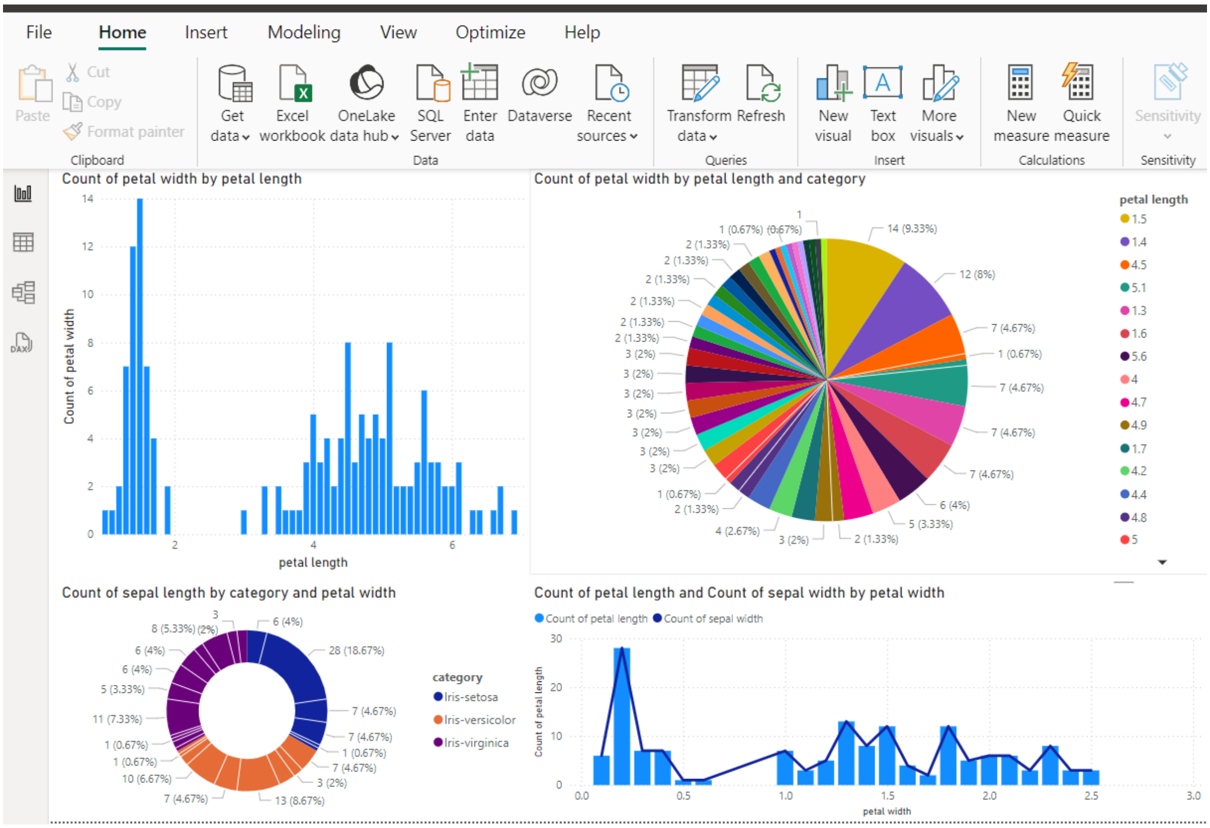
- **Sepal Length and Sepal Width:** A correlation value of 0.12 indicates a very weak relationship between these features.

Interpretation:

This heatmap reveals that petal length and petal width are strongly correlated with sepal length. Sepal width has a weaker negative relationship with both petal length and petal width. Sepal length and sepal width are almost independent of each other.

Possible Insights:

- **Redundancy:** The high correlation between petal length and petal width suggests that one of these features might be redundant in further analysis.
- **Feature Importance:** Sepal length seems to be a more important feature compared to sepal width due to its stronger correlations with other features.



Understanding the Data

The provided visualization offers insights into the Iris dataset, a classic dataset in machine learning used for classification tasks. The dataset contains measurements of sepal length, sepal width, petal length, and petal width for three species of Iris: Setosa, Versicolor, and Virginica.

1. github.com

github.com

Visualization Breakdown

- **Histogram of Petal Width by Petal Length:** This plot shows the distribution of petal widths across different petal lengths. We observe that as petal length increases, petal width tends to increase as well, suggesting a positive correlation between these two features.
- **Pie Chart of Petal Width by Petal Length and Category:** This visualization breaks down the distribution of petal widths by petal length for each Iris species. It reveals that Setosa tends to have smaller petal widths across all petal lengths compared to Versicolor and Virginica.
- **Histogram of Sepal Length by Category and Petal Width:** This plot displays the distribution of sepal lengths for each Iris species, categorized by petal width. It shows that Setosa has a distinct sepal length range compared to the other two species, which overlap to a certain extent.
- **Line Chart of Count of Petal Length and Count of Sepal Width by Petal Width:** This chart compares the count of petal lengths and sepal widths across different petal width values. It indicates that as petal width increases, the count of both petal lengths and sepal widths generally increases, but with some variations.

Key Observations

- **Distinct Features:** The Iris Setosa species clearly stands out with smaller petal lengths and widths compared to the other two species. This suggests that petal length and width can be effective features for classifying Iris Setosa.
- **Overlapping Features:** Versicolor and Virginica exhibit more overlap in terms of petal and sepal measurements, making their classification potentially more challenging based solely on these features.
- **Correlations:** The positive correlation between petal length and petal width is evident, which could be leveraged in machine learning models.
- **Potential for Classification:** The visualizations suggest that combining information about petal length, petal width, and sepal length could be effective in building a classification model to distinguish between the three Iris species.

…………………end of report……………………………