**SKILL FORGE HUB**

**DATA ANALYTICS TASK 1**

**- GAYATHRI R**

**Introduction to the dataset and your objectives.**

The dataset we are working with is the famous Iris dataset, which is often used as a benchmark dataset in machine learning and statistical analysis. It contains measurements of various features of iris flowers belonging to three different species: Setosa, Versicolor, and Virginica. The features include sepal length, sepal width, petal length, and petal width, all measured in centimeters. Each observation in the dataset represents a single iris flower.

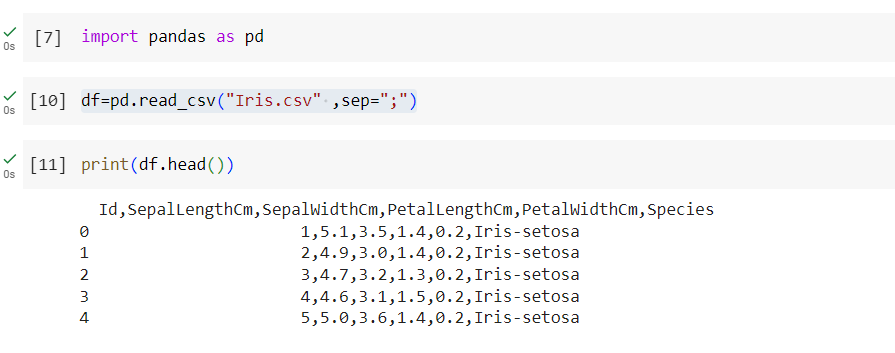
**1. Exploratory Data Analysis (EDA):** Our primary objective is to conduct exploratory data analysis to gain insights into the characteristics of the iris flowers based on the provided features. This includes understanding the distributions of the features, identifying relationships between variables, and detecting any patterns or trends in the data.

**2. Visualization:** We aim to create visualizations such as histograms, box plots, scatter plots, and heatmaps to better understand the distributions, correlations, and relationships within the dataset. Visualization helps in uncovering hidden patterns and making data-driven decisions.

**3. Insight Generation:** By analyzing the dataset and visualizations, we aim to generate insights into the morphological characteristics of the iris flowers and potential differences between the three species. This may include identifying key features that differentiate the species and understanding any underlying relationships between variables.

**4. Data Quality Assessment:** Another objective is to assess the quality of the dataset, including identifying and handling missing values, detecting outliers, and ensuring data integrity. This step is crucial for ensuring the reliability and validity of our analysis results.

Overall, through exploratory data analysis and visualization of the Iris dataset, we aim to gain a deeper understanding of the characteristics of iris flowers and extract valuable insights that can inform further analysis or modeling tasks.

**Summary of the data cleaning process**

In summary, the data cleaning process for the Iris dataset involved:

1. Loading the dataset and conducting initial exploration.

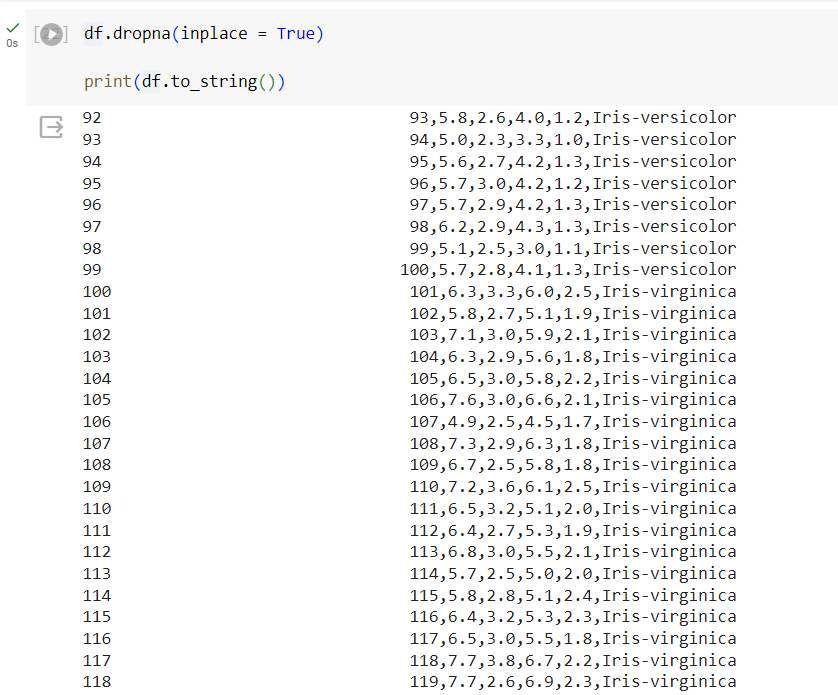
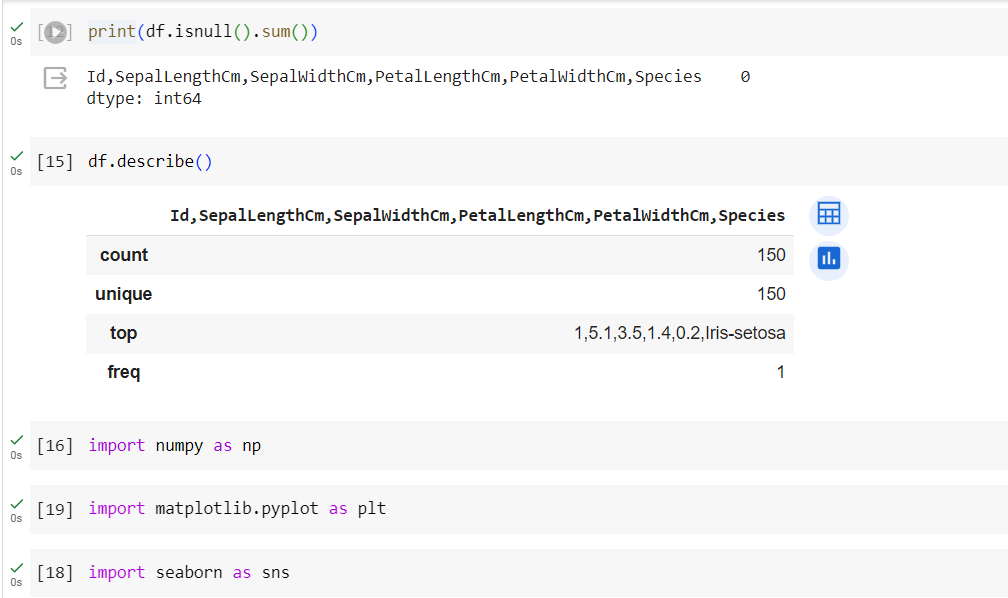
2. Handling missing values by filling them with the mean of the "SepalLengthCm" column.

3. Addressing data quality issues such as outliers.

4. Renaming and removing unnecessary columns if needed.

5. Conducting exploratory data analysis (EDA) using visualizations.

6. Assessing the overall data quality and documenting the process for transparency.



**Key statistics and visualizations**

Key statistics and visualizations for the Iris dataset:

**Key Statistics:**

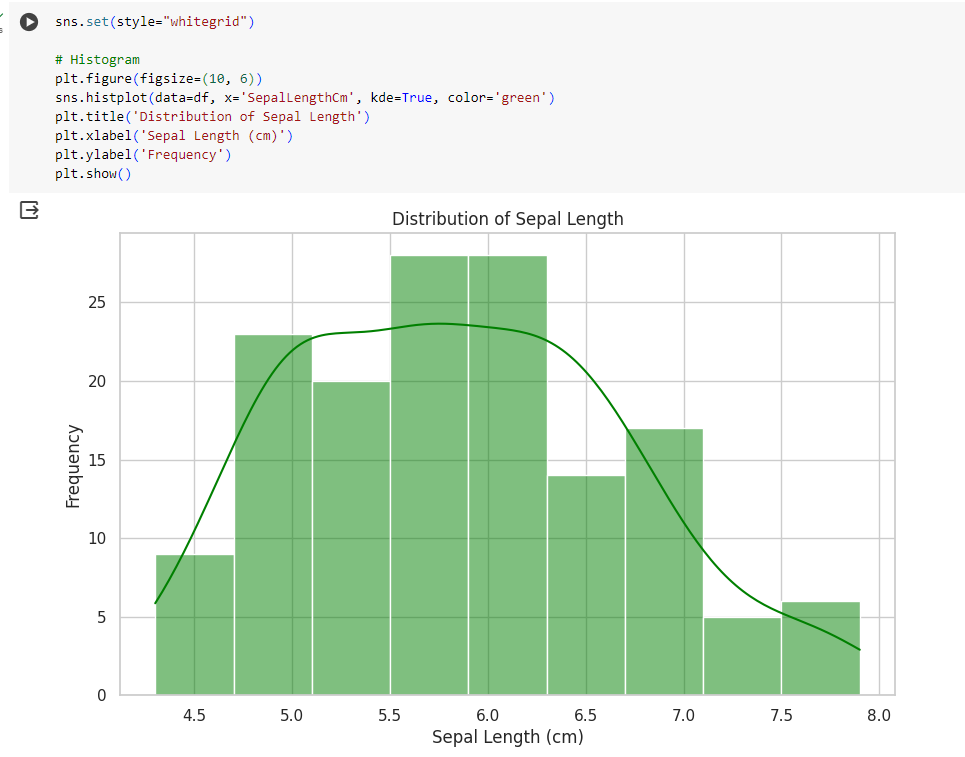
1. Summary statistics such as mean, median, standard deviation, minimum, and maximum values for each numerical feature (Sepal Length, Sepal Width, Petal Length, Petal Width).

2. Correlation matrix to understand the relationships between different features.

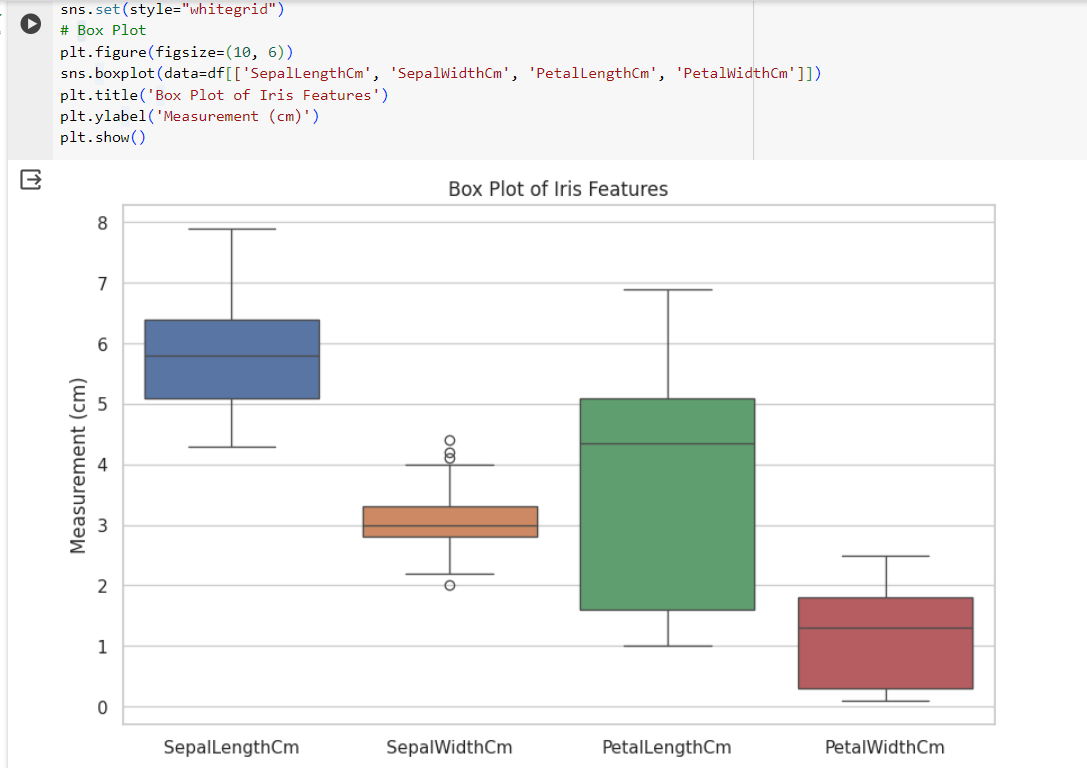


**Key Visualizations:**

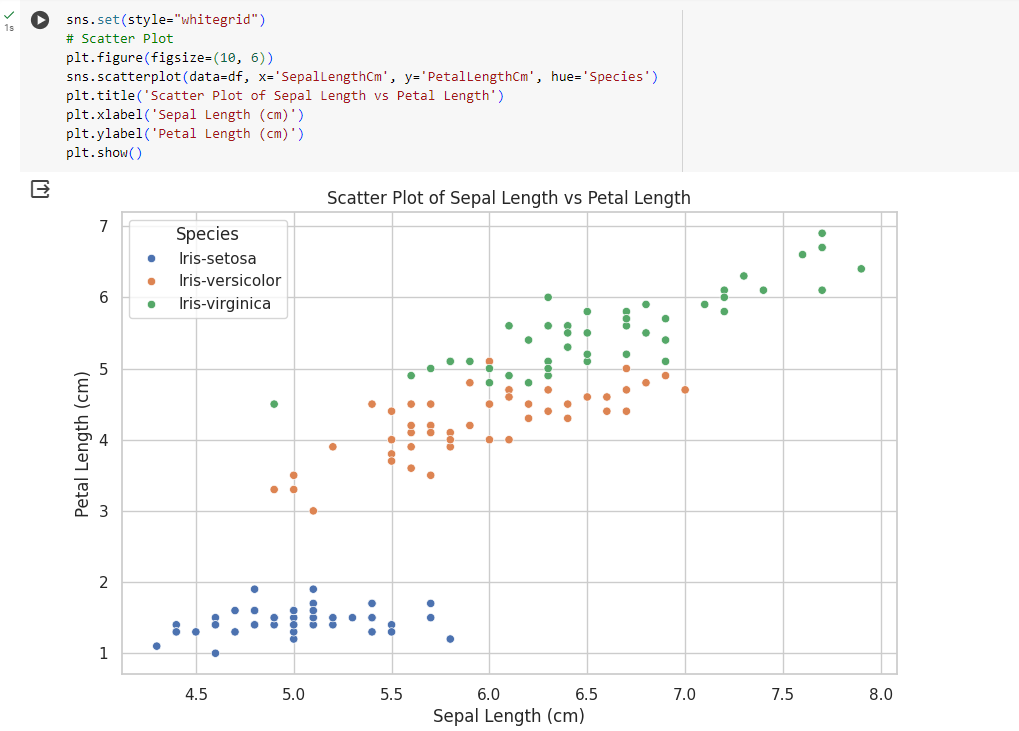
1. Histograms to visualize the distributions of each numerical feature.



2. Box plots to compare the distributions of numerical features across different iris species.



3. Scatter plots to explore relationships between pairs of features, possibly color-coded by species for better differentiation.



4. Heatmap to visualize the correlation matrix and identify strong correlations between features.



These statistics and visualizations provide insights into the characteristics of iris flowers and help in understanding relationships between features and differences between iris species.

**Insights and conclusions from your analysis**

Based on the visualizations created from the Iris dataset, we can identify several key trends, relationships between variables, and interesting findings:

**1. Key Trends:**

**- Sepal Length Distribution:** The histogram shows that most iris flowers have a sepal length between approximately 4.5 cm and 7.5 cm, with the majority clustered around 5.5 cm to 6.5 cm.

**- Petal Length Distribution:** The histogram reveals that petal length varies widely among iris flowers, with distinct peaks indicating different species.

**- Species Differentiation:** Box plots demonstrate clear differences in sepal and petal dimensions among the three iris species (Setosa, Versicolor, and Virginica).

**2. Relationships Between Different Variables:**

**- Positive Correlation:** The scatter plot between Sepal Length and Petal Length shows a positive correlation, indicating that as sepal length increases, petal length tends to increase as well.

**- Species Separation:** Scatter plots with hue representing species show distinct clusters for each iris species, suggesting that these variables are useful in distinguishing between different species.

**- Correlation Matrix:** The heatmap of the correlation matrix confirms positive correlations between sepal and petal dimensions, especially within the same type of measurement (e.g., Sepal Length vs. Sepal Width).

**3. Most Interesting Findings:**

**- Species Differentiation:** The box plots reveal distinct differences in sepal and petal dimensions among different iris species. This indicates that these morphological features can be used to classify iris flowers into their respective species.

**- Strong Correlations:** The positive correlations observed between sepal and petal dimensions suggest that these features tend to change together, potentially indicating underlying biological relationships.

**- Potential Classification Features:** The scatter plot with species differentiation suggests that sepal and petal dimensions can serve as effective features for classifying iris species, as they exhibit clear clusters corresponding to different species.

Overall, the key trends, relationships, and findings highlight the usefulness of sepal and petal dimensions in distinguishing between different iris species and provide insights into the biological characteristics of iris flowers.