# Project Report FOR SmartED: Exploratory Data Analysis (EDA) of the IRIS Dataset (BY-Siddhesh Vaishnav)

## Introduction

This project details the process of performing an Exploratory Data Analysis (EDA) on the classic Iris dataset. The primary goal is to understand the dataset’s structure, statistical properties, and relationships between its features through various visualizations. This is a crucial step in the data science workflow, as it helps in preparing the data for subsequent machine learning tasks.

## Objective

The main objectives of this project were:

* To perform an initial inspection of the Iris dataset, including its shape, data types, and a check for missing values.
* To calculate descriptive statistics for the numerical features of the dataset.
* To visualize feature distributions and relationships using scatter plots, histograms, and box plots.
* To assess the normality of key features using Q-Q plots.

## Methodology

This project was implemented using Python and several data science libraries.

**Tools and Technologies:**

* **Python:** The core programming language for the analysis.
* **Pandas:** Used for data manipulation and creating DataFrames.
* **Seaborn & Matplotlib:** These libraries were used for creating static and interactive data visualizations.
* **NumPy:** Utilized for numerical operations.
* **SciPy:** The scipy.stats module was used to generate Q-Q plots.

**Steps:**

1. **Data Loading:** The Iris dataset was loaded directly from the Seaborn library.
2. **Initial Exploration:** The head(), tail(), shape, and info() methods were used to get a basic understanding of the data. A check for null values was performed using isnull().sum().
3. **Descriptive Analysis:** The describe() method was used on the DataFrame to generate a summary of descriptive statistics for the numerical features.
4. **Data Visualization:** Plots were created to visualize the data’s characteristics.
5. **Normality Testing:** Q-Q plots were used to compare the distributions of petal\_length and sepal\_length to a normal distribution.

## Code and Implementation Details

The analysis was performed in a Jupyter Notebook. The following code snippets and their outputs show the steps taken.

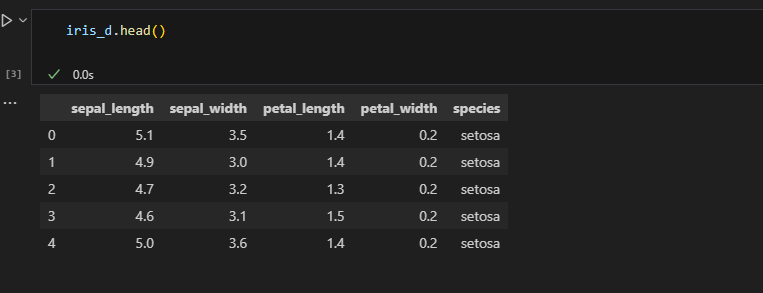
### 1. Importing Libraries and Loading Data

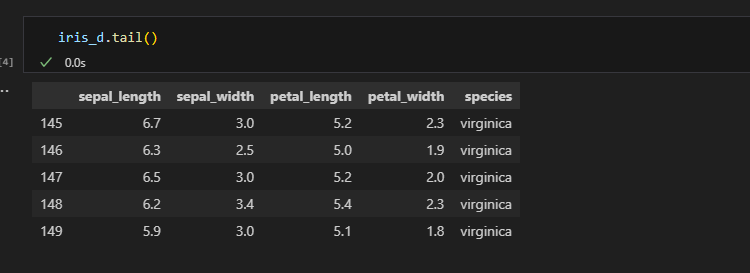
import pandas as pd  
import seaborn as sb  
import matplotlib.pyplot as plt  
import numpy as np  
  
iris\_d = sb.load\_dataset("iris")

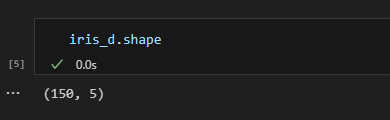
### 2. Initial Data Inspection

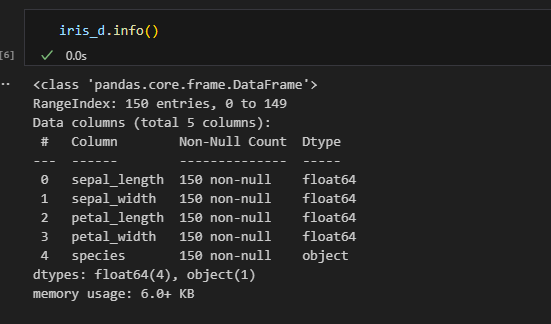
iris\_d.head()  
iris\_d.tail()  
iris\_d.shape  
iris\_d.info()  
iris\_d.isnull().sum()

**Output:**







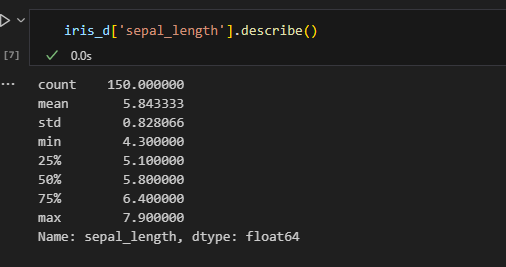


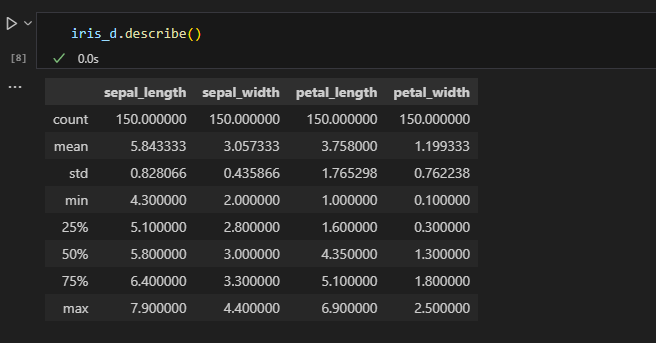
### 

### 3. Descriptive Statistics

iris\_d['sepal\_length'].describe()  
iris\_d.describe()

**Output:**

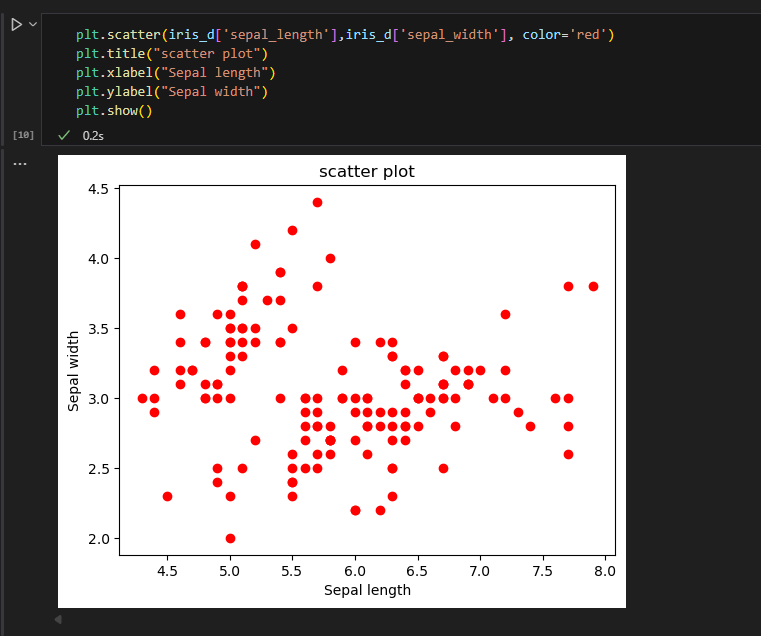




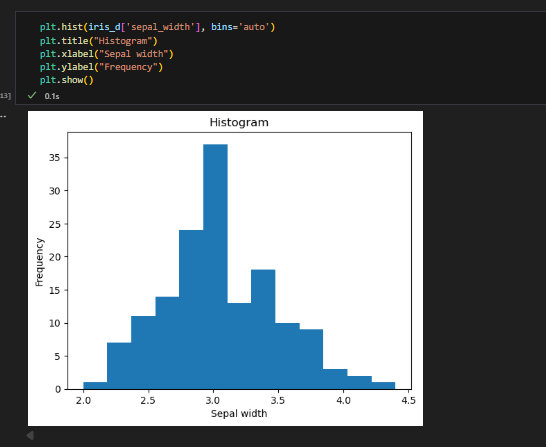
### 4. Visualizations

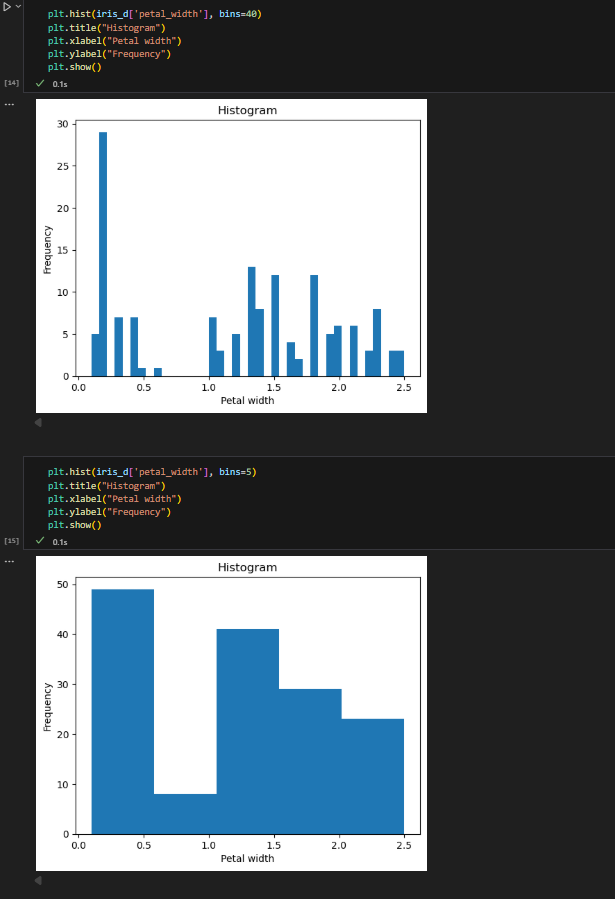
plt.scatter(iris\_d['sepal\_length'], iris\_d['sepal\_width'], color='red')  
plt.title("Scatter Plot")  
plt.xlabel("Sepal length")  
plt.ylabel("Sepal width")  
plt.show()  
  
plt.hist(iris\_d['sepal\_width'], bins=40)  
plt.title("Histogram")  
plt.xlabel("Sepal width")  
plt.ylabel("Frequency")  
plt.show()  
  
plt.hist(iris\_d['petal\_width'], bins=40)  
plt.title("Histogram")  
plt.xlabel("Petal width")  
plt.ylabel("Frequency")  
plt.show()  
  
sb.boxplot(x="sepal\_width", data=iris\_d)  
plt.title("Box Plot")  
  
sb.boxplot(x="sepal\_length", data=iris\_d)  
plt.title("Box Plot")

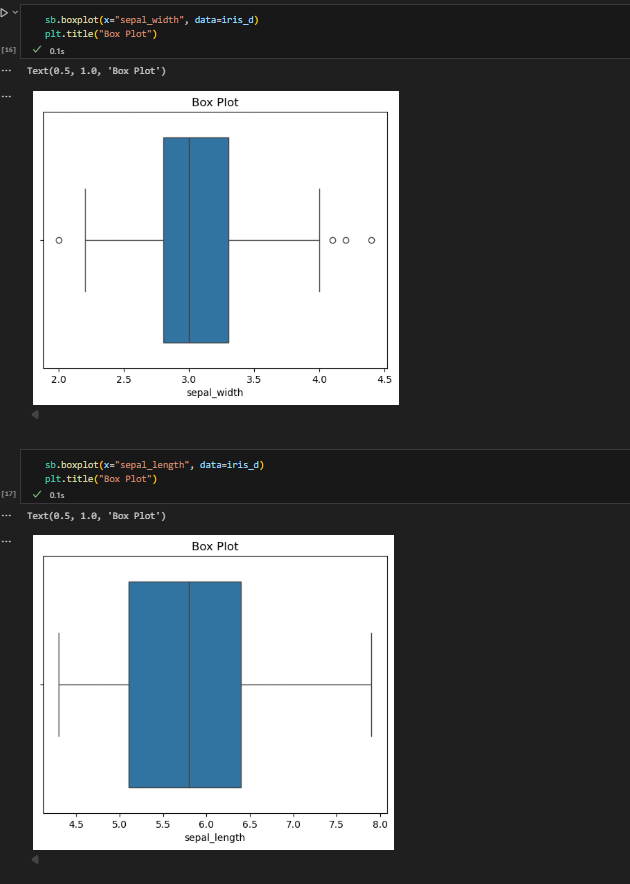
**Output:**







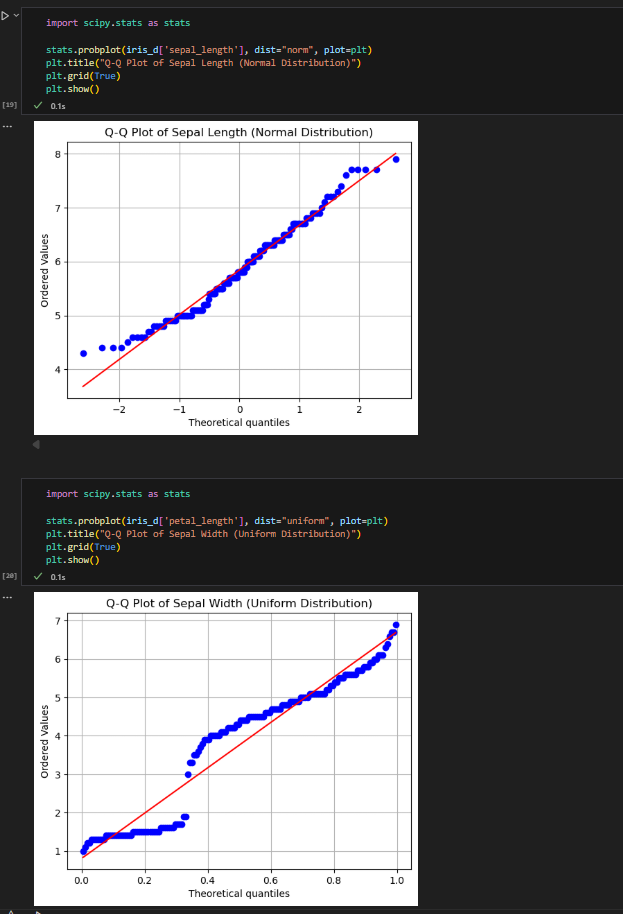
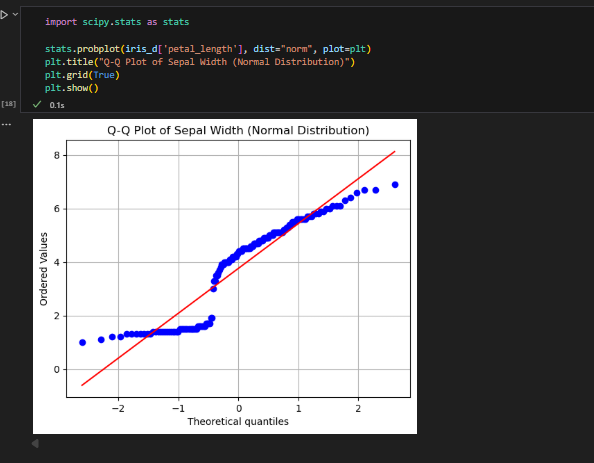




### 5. Q-Q Plots

import scipy.stats as stats  
  
stats.probplot(iris\_d['petal\_length'], dist="norm", plot=plt)  
plt.title("Q-Q Plot of Petal Length (Normal Distribution)")  
plt.grid(True)  
plt.show()  
  
stats.probplot(iris\_d['sepal\_length'], dist="norm", plot=plt)  
plt.title("Q-Q Plot of Sepal Length (Normal Distribution)")  
plt.grid(True)  
plt.show()  
  
stats.probplot(iris\_d['petal\_length'], dist="uniform", plot=plt)  
plt.title("Q-Q Plot of Petal Length (Uniform Distribution)")  
plt.grid(True)  
plt.show()

**Output:**



## Results and Observations

* The dataset contains 150 rows and 5 columns, with no missing values.
* The sepal\_length and sepal\_width features appear to be approximately normally distributed, as indicated by their histograms and Q-Q plots.
* The petal\_width and petal\_length features have non-normal distributions.
* The box plot for sepal\_width identifies some potential outliers.
* The scatter plot of sepal\_length and sepal\_width shows a negative correlation and distinct clusters, suggesting that these features could be useful for separating the different species.

## Conclusion

The EDA process provided a solid foundation for further analysis of the Iris dataset. The analysis successfully identified the data’s structure, statistical properties, and distribution characteristics. The key findings, such as the presence of outliers and non-normal distributions in certain features, are crucial for informing the next steps, including feature engineering and the selection of an appropriate machine learning model. This project successfully achieved its objective of performing a comprehensive exploratory analysis.

THANK YOU.