

|  |
| --- |
|  |
| Nexus Project |
| Data Analyst Projects:  Exploring Data Insights |
| KUMAR GAURAV | gauravt.nic@gmail.com |

|  |  |  |
| --- | --- | --- |
| Decorative |  |  |
| Introduction |  |
| **Project 1 : Iris Dataset Basic Analysis**  **Author:**  Kumar Gaurav  **Date:**  30th May, 2024  **Purpose:**  The purpose of this project is to perform an Exploratory Data Analysis (EDA) on the Iris dataset to visualize key statistics and distributions, explore correlations, patterns, and trends, and document the findings using Power BI and Python.  **Data Description**  **Dataset Information:**   * **Name**: Iris Dataset * **Source**: The dataset is publicly available from the UCI Machine Learning Repository. * **Description**: The dataset contains 150 observations of iris flowers. There are four features (sepal length, sepal width, petal length, and petal width) and one target variable (species).   **Attributes:**   1. **sepal length (cm)**: Continuous variable representing the sepal length of the iris flower. 2. **sepal width (cm)**: Continuous variable representing the sepal width of the iris flower. 3. **petal length (cm)**: Continuous variable representing the petal length of the iris flower. 4. **petal width (cm)**: Continuous variable representing the petal width of the iris flower. 5. **species**: Categorical variable representing the species of the iris flower (Setosa, Versicolor, Virginica).   **3. Data Loading and Cleaning**  **Loading Data:**  The dataset was loaded into the Python environment using the pandas library:  import pandas as pd  # Load the dataset  iris = pd.read\_csv('iris\_cleaned.csv')  **Data Cleaning:**  No missing values were found in the dataset, as shown below:  iris.isnull().sum()  The data was found to be clean with no missing values. No outliers were handled as all data points were within expected ranges.  **4. Exploratory Data Analysis (EDA)**  **Statistical Summary:**  A statistical summary of the dataset was generated using **pandas**:  iris.describe()  The summary revealed the following key statistics:   * Sepal length ranged from 4.3 to 7.9 cm. * Sepal width ranged from 2.0 to 4.4 cm. * Petal length ranged from 1.0 to 6.9 cm. * Petal width ranged from 0.1 to 2.5 cm.   **Visualizations:**   * **Histogram**:   + The histogram for sepal length showed a normal distribution with a peak around 5.5 to 6.0 cm.   import matplotlib.pyplot as plt  iris['sepal length (cm)'].hist()  plt.title('Distribution of Sepal Length')  plt.xlabel('Sepal Length (cm)')  plt.ylabel('Frequency')  plt.show()   * **Scatter Plot**:   + The scatter plot showed a positive correlation between sepal length and petal length, with different species forming distinct clusters   iris.plot(kind='scatter', x='sepal length (cm)', y='petal length (cm)', c='species', colormap='viridis')  plt.title('Sepal Length vs Petal Length')  plt.xlabel('Sepal Length (cm)')  plt.ylabel('Petal Length (cm)')  plt.show()   * **Box Plot**:   Example:  iris.boxplot(column='sepal width (cm)', by='species')  plt.title('Sepal Width by Species')  plt.xlabel('Species')  plt.ylabel('Sepal Width (cm)')  plt.show()   1. **Data Visualization with Power BI**   **Steps to Create Visualizations in Power BI:**   1. **Import Data**:    * Load **iris\_cleaned.csv** into Power BI. 2. **Create Visualizations**:    * Created a histogram for sepal length (cm).    * Created a scatter plot for sepal length (cm) vs. petal length (cm).    * Created a box plot for sepal width (cm) by species**.**   **Interactivity:**   * **Slicer**:   + Added a slicer for the **species** field to filter the data interactively.  1. **Insights and Conclusion**   **Key Findings:**   * Setosa species have significantly shorter petal lengths compared to Versicolor and Virginica. * A positive correlation exists between sepal length and petal length.   **Summary of Findings**  **Species Distribution:** The dataset contains three species of iris flowers - Setosa, Versicolor, and Virginica. Each species has 50 observations.  **Feature Distributions:**Sepal lengths and widths, as well as petal lengths and widths, show distinct patterns for different species.  Setosa species have significantly shorter petal lengths and widths compared to Versicolor and Virginica.  **Correlations:**  Positive correlation between sepal length and petal length.  Clear separation in scatter plots when colored by species, indicating that features can be used to distinguish between species.  **Conclusion**  The analysis provided valuable insights into the Iris dataset. By visualizing the distributions and relationships between features, we can effectively classify the different species of iris flowers. The use of Python for EDA and Power BI for interactive visualizations complemented each other to enhance the overall understanding of the data.  **References**  **UCI Machine Learning Repository: Iris Dataset** |  |  |
|  |  |  |
|  |  |  |