Project Documentation: Iris Dataset Analysis

1. Introduction

This project involves performing basic exploratory data analysis (EDA) on the Iris dataset using Python and Power BI. The objective is to visualize key statistics and distributions to gain insights into the dataset and identify patterns, correlations, and trends.

1.1 Project Overview

The Iris dataset contains 150 samples of iris flowers, categorized into three species: Iris-setosa, Iris-versicolor, and Iris-virginica. Each sample includes four features: sepal length, sepal width, petal length, and petal width.

2. Exploratory Data Analysis (EDA) with Python

2.1 Loading the Dataset

The Iris dataset was loaded using the pandas library.

```
import pandas as pd

# Load the Iris dataset

iris = pd.read_csv('path_to_iris_dataset.csv')
```

2.2 Basic Statistics

Basic statistics such as mean, median, and standard deviation were calculated for each feature.

```
# Summary statistics
```

```
summary_stats = iris.describe()

print(summary_stats)
```

2.3 Data Visualization

Key statistics and distributions were visualized using matplotlib and seaborn libraries.

2.3.1 Histograms

Histograms were plotted for each feature to visualize the distribution.

```
import seaborn as sns

import matplotlib.pyplot as plt

# Histogram for sepal length

sns.histplot(iris['sepal_length'], kde=True)

plt.title('Sepal Length Distribution')

plt.show()
```

Similarly, histograms for other features were plotted.

2.3.2 Scatter Plots

Scatter plots were created to visualize the relationships between features.

```
# Scatter plot for sepal length vs sepal width

sns.scatterplot(x='sepal_length', y='sepal_width', hue='species', data=iris)
```

```
plt.title('Sepal Length vs Sepal Width')

plt.show()
```

Similarly, scatter plots for other feature pairs were plotted.

3. Data Visualization with Power BI

3.1 Exporting DataFrame to CSV

The DataFrame was exported to a CSV file for import into Power BI.

```
# Export to CSV

iris.to_csv('iris_dataset.csv', index=False)
```

3.2 Importing Data into Power BI

1. Open Power BI Desktop.
2. Get Data:
   - Click on "Get Data" and select "Text/CSV".
   - Import the iris_dataset.csv file.

3.3 Creating Visualizations in Power BI

3.3.1 Scatter Plot

- Add a scatter plot visualization.
- Drag sepal_length to the X-axis and sepal_width to the Y-axis.

- Drag species to the Legend field to color code by species.

3.3.2 Histograms

- Add a clustered column chart for each feature (sepal length, sepal width, petal length, petal width).
- Drag the feature to the Values field and species to the Axis field.

3.4 Example Visualizations

Here are some of the visualizations created in Power BI:

1. Scatter Plot: Sepal Length vs. Sepal Width
2. Histograms: Distribution of Sepal Length, Sepal Width, Petal Length, and Petal Width by Species
3. Box Plots: Showing the spread and outliers of each feature by species
4. Heatmaps: Correlation matrix to identify relationships between features

4. Documentation and Insights

4.1 Summary of Findings

- Species Separation: Scatter plots show clear separation between species based on sepal and petal measurements.
- Distribution Patterns: Histograms indicate that sepal_length and petal_length have distinct distributions for each species.
- Correlation: Heatmap reveals strong correlations between petal length and petal width, as well as between sepal length and petal length.

Project Documentation: Iris Dataset Analysis

4.2 Methodology

1. Data Loading: Imported the Iris dataset using pandas.

2. Basic EDA: Calculated summary statistics and visualized distributions using Python.

3. Exporting Data: Exported the DataFrame to CSV for Power BI.

4. Power BI Visualizations: Created various visualizations to explore patterns and correlations.

4.3 Conclusion

The analysis of the Iris dataset revealed significant insights into the characteristics of different iris species. The combination of Python for initial EDA and Power BI for advanced visualizations provided a comprehensive understanding of the data.

4.4 Future Work

- Advanced Analysis: Implement machine learning models to classify species.
- Additional Datasets: Apply similar analysis techniques to other datasets for comparative studies.