PANDAS ASSIGNMENT – IRIS DATASET

Data Exploration & Analysis using Pandas

Prepared by: Riya Manoj

Introduction

- What is Pandas?
- - Python library for data manipulation and analysis.

- Purpose of this Assignment:
- - Learn Pandas operations.
- - Explore the Iris dataset, a classic dataset used in ML.

Dataset Overview

- - Iris dataset contains flower measurements.
- - Features: SepalLengthCm, SepalWidthCm, PetalLengthCm, PetalWidthCm
- - Target: Species (Setosa, Versicolor, Virginica)
- - 150 rows total
- - Checked missing values and duplicates.

Tools Used

- - Python – Programming language
- - Pandas – Data manipulation & exploration
- - Matplotlib – For visualizations
- - Google Colab – Coding environment

Data Loading & Inspection

- - Loaded dataset with pd.read_csv()
- - Previewed with head() and tail()
- - Checked shape, dtypes, unique values, and describe()

Data Quality Checks

- - Missing values check → None
- - Duplicates check → Used df.duplicated().sum()
- - Descriptive statistics with df.describe()

Data Selection & Filtering

- - **Column selection**: df['Species'], df[['SepalLengthCm','PetalLengthCm']]
- - **Row selection**: iloc[], loc[]
- - **Filtering** examples:
- df[df['SepalLengthCm']>5.0]
- df[df['Species']=='Iris-setosa']

Sorting Data

- - Sort by column:
- df.sort_values('SepalLengthCm')
- - Sort by multiple columns:
- df.sort_values(['Species','PetalLengthCm'], ascending=[True,False])

Creating New Columns

- - New column SepalRatio = SepalLengthCm / SepalWidthCm
- - New column SepalDouble = SepalLengthCm * 2 using apply()

Data Exploration

- - Unique values in Species: df['Species'].unique()
- - Count of each species: df['Species'].value_counts()
- - Correlation: df.corr()

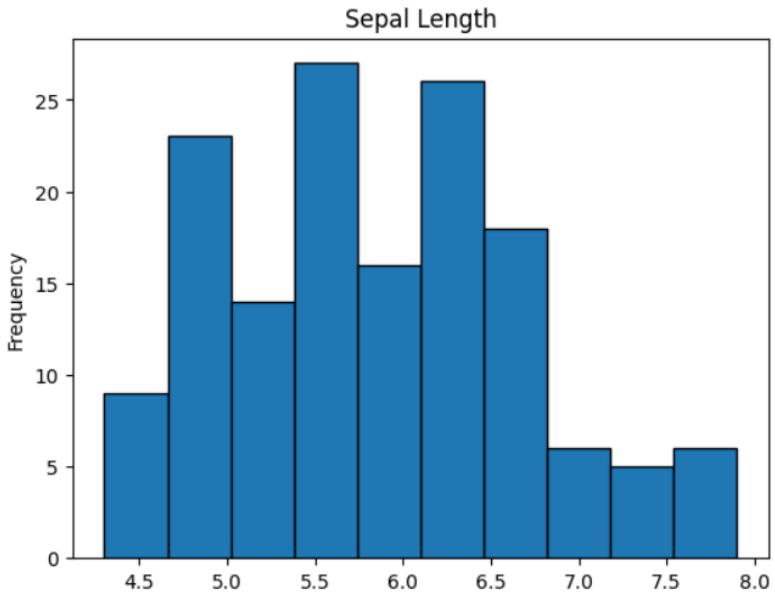
Grouping & Aggregation

- - Mean SepalLengthCm by species:
- df.groupby('Species')['SepalLengthCm'].mean()
- - Multiple aggregations:

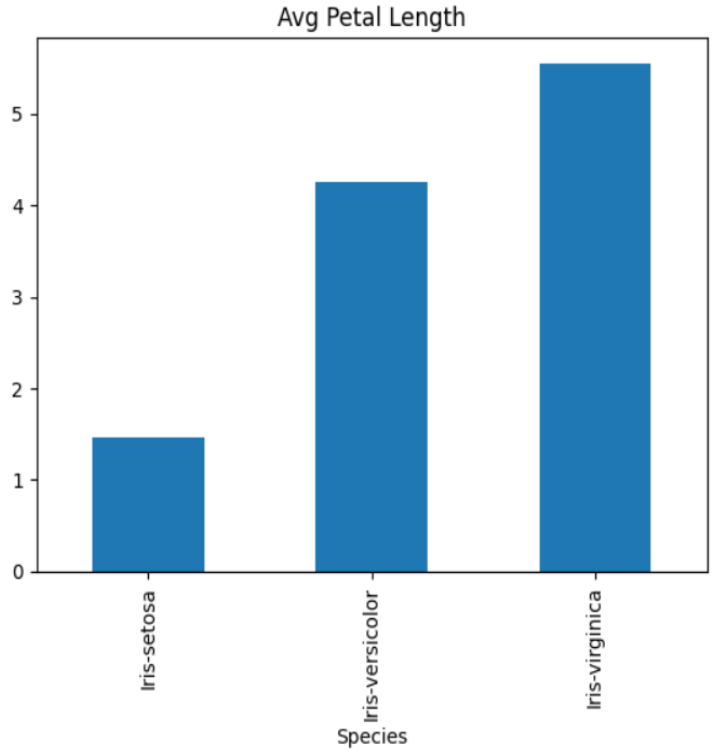
df.groupby('Species').agg({'SepalLengthCm':'mean','PetalLengthCm':'max'})

Visualization

- - Histogram of Sepal Length:



- - Bar chart of Avg Petal Length by species:



Conclusion

- - Pandas is powerful for data inspection, cleaning, and analysis.
- - Key learning outcomes:
- 1. Data loading and quality checks
- 2. Selection, filtering, sorting
- 3. Feature engineering
- 4. Grouping, aggregation, correlation
- 5. Visualizations with Pandas
- - Outcome: Hands-on experience exploring Iris dataset with Pandas