# CS 519 Applied Machine Learning I

**HW1: Basic Python Programming**

**Indronil Bhattacharjee**

**Task 1: Read data from the Iris dataset**

```
import pandas as pd
import matplotlib.pyplot as plt

# Task 1
def read_dataset():
    iris_data = pd.read_csv("iris.data", header=None) # read data
    print("Data reading completed")
    return iris_data
```

**Task 2: Counting number of rows and columns**

```
# Task 2
def row_column(dataset):
    rows, columns = dataset.shape # row and column count
    print(f"Number of rows: {rows}")
    print(f"Number of columns: {columns}")
```

**Task 3: Get the distinct values of the last column**

```
# Task 3
def distinct_values(dataset):
    distinct_values = dataset.iloc[:, -1].unique() # get distinct values with unique()
    print(f"Distinct values of the last column: {distinct_values}")
```

**Task 4: Count, Average, Minimum, Maximum**

```
# Task 4
def analyze_setosa_data(dataset):
    setosa_data = dataset[dataset.iloc[:, -1] == "Iris-setosa"]
    num_rows = setosa_data.shape[0]
    avg_first_col = setosa_data.iloc[:, 0].mean() # get average with mean()
    max_second_col = setosa_data.iloc[:, 1].max() # get maximum with max()
    min_third_col = setosa_data.iloc[:, 2].min() # get minimum with min()

    print(f"Number of rows with 'Iris-setosa': {num_rows}")
    print(f"Average value of the first column: {avg_first_col}")
    print(f"Maximum value of the second column: {max_second_col}")
    print(f"Minimum value of the third column: {min_third_col}")
```

**Task 5: Visualization of data**

```
# Task 5
def plot_scatter_plot(dataset):
    colors = {'Iris-setosa': 'red', 'Iris-versicolor': 'blue', 'Iris-virginica': 'green'}
    shapes = {'Iris-setosa': 's', 'Iris-versicolor': 'o', 'Iris-virginica': '^'}

    for species, group in dataset.groupby(dataset.iloc[:, -1]):
        plt.scatter(group.iloc[:, 0], group.iloc[:, 1], color=colors[species],
                    marker=shapes[species], label=species) # visualize with scatter plot

    plt.xlabel("First Column")
    plt.ylabel("Second Column")
    plt.legend()
    plt.show()
```

**Task 6: Readme file**

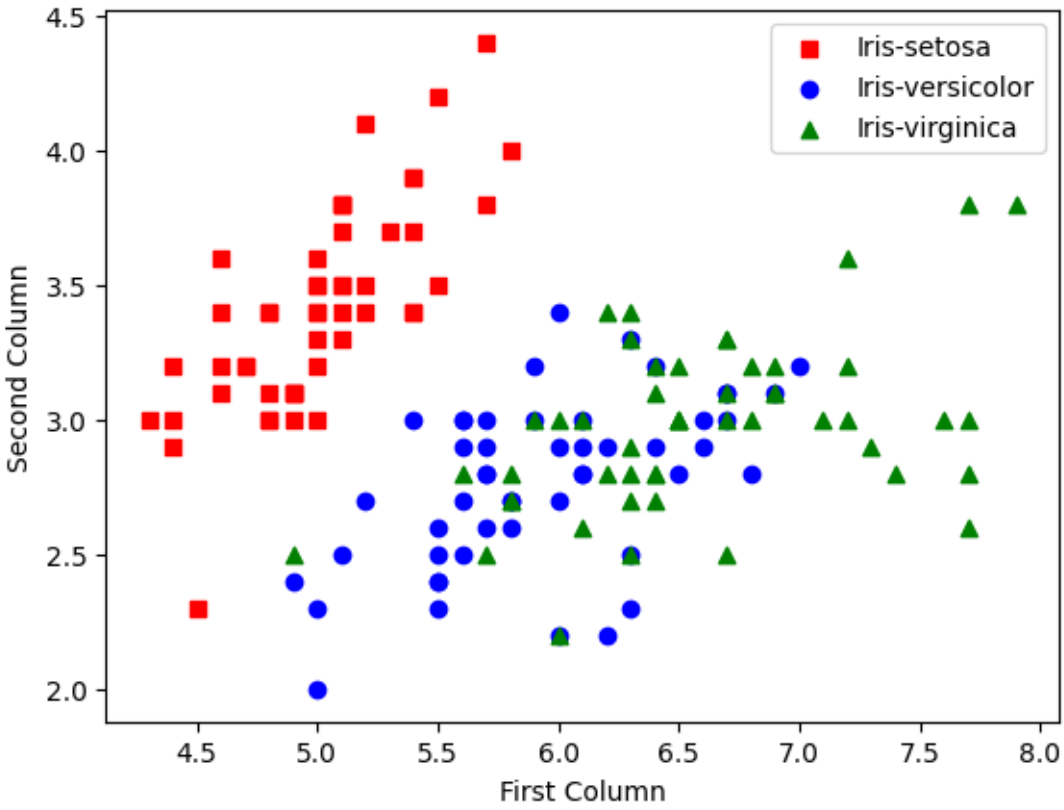
```
# Task 6
# Please read instructions in the readme.txt file
```

**Execution of the task functions**

```
# Main program
if __name__ == "__main__":
    print("Task 1")
    iris_dataset = read_dataset()
    print("_____________________________________")
    print("Task 2")
```

```
    row_column(iris_dataset)
    print("_____________________________________")
    print("Task 3")
    distinct_values(iris_dataset)
    print("_____________________________________")
    print("Task 4")
    analyze_setosa_data(iris_dataset)
    print("_____________________________________")
    print("Task 5")
    plot_scatter_plot(iris_dataset)
```

```
Task 1
Data reading completed.
--------------------------------------
Task 2
Number of rows: 150
Number of columns: 5
--------------------------------------
Task 3
Distinct values of the last column: ['Iris-setosa' 'Iris-versicolor' 'Iris-virginica']
--------------------------------------
Task 4
Number of rows with 'Iris-setosa': 50
Average value of the first column: 5.006
Maximum value of the second column: 4.4
Minimum value of the third column: 1.0
--------------------------------------
Task 5
```


Iris-setosa
Iris-versicolor
Iris-virginica
Second Column
First Column
4.5
4.0
3.5
3.0
2.5
2.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0