CS 210: Data Management for Data Science
Homework 1: Basic Tools and Manipulations

1 Theory Questions (20 points)

1. Explain the difference between a NumPy array and a Python list, and write down examples of defining a NumPy array and a Python list.

2. Explain the difference between a DataFrame and a Series in Pandas, and write down examples of defining a DataFrame and a Series in Pandas.

2 Coding Exercises (30 points)

2.1 Data Manipulation with Pandas

Download the Iris dataset and load it into a Pandas DataFrame. The dataset contains information about different iris flowers, including:

- Sepal length in cm
- Sepal width in cm
- Petal length in cm
- Petal width in cm
- Class (Iris-setosa, Iris-versicolor, Iris-virginica)

Tasks:

1. Load the dataset into a Pandas DataFrame and add appropriate column names: `['sepal_length', 'sepal_width', 'petal_length', 'petal_width', 'class']`.

2. Display the first 10 rows of the DataFrame.

3. Calculate the mean, median, and standard deviation of the sepal length for each class.

4. Filter the DataFrame to include only rows where the petal length is greater than 1.5 and display the first 5 rows of the filtered DataFrame.

Please write down the results of each task and include screenshots of your experiments in your submitted PDF document. Additionally, attach your source code files.

3 Submission Requirements

1. **Submission Format:** All files should be submitted via Canvas. Submit the report in PDF format, and attach source code files mentioned in the tasks.

2. **Note:** Make sure to test all your scripts thoroughly before submission.