## CS4431 HW2

## Team: Peter Strimbu, Megan Hoeksema, Samantha Reno, Patrick Ayers

## 2022-06-26

# Week 2 Design Document

## Introduction

**Intent of the application**

Perform exploratory data analysis on the Iris dataset to identify what data are included for which variables, visualize patterns, uncover missing or anomalous data, and find relationships between data, if any, in order to formulate appropriate questions about the data.

**Dataset to be used, including source**

Iris dataset (1936) obtained from sklearn.datasets / load\_iris

**Use case**

An alien species, tasked with performing terrestrial botanical inquiry, lands on Earth and is extremely intrigued by the morphology of a certain flower, the iris, of which three species holds particular interest. Luckily, the Iris database exists.

## Dataset Analysis

**Define labels**

The four attributes are *petal length*, *petal width*, *sepal length*, and *sepal width.* The fifth column is a class label, *species*.

**Define variables**

Species is a categorical (3) variable: *iris setosa*, *iris virginica*, *iris versicolor*.

The four measurements in centimeters, lengths and widths, are continuous qualitative values.

Each of the three species has 50 instances for a total of 150 records.

## Inputs

No user interaction is required.

## Proposed Libraries

**Libraries**

* Numpy
* Pandas
* Matplotlib.pyplot
* Sklearn.datasets
* Seaborn

**Library source**

* Numpy: <https://numpy.org/doc/stable/index.html>
* Pandas: <https://pandas.pydata.org/>
* Matplotlib.pyplot: <https://matplotlib.org/>
* Sklearn.datasets: <https://scikit-learn.org/stable/>
* Seaborn: https://seaborn.pydata.org/

## Proposed Solution

The solutions for each task are:

* Import the dataset

We will load the dataset from sklearn.datasets and review the nature of the columns, range of each feature, and relationship of the features.

* Scatter Plot – 2 dimensions

Create a scatter plot of the iris data to graphically show the relationship of the sepal length relative to the sepal width.

* Scatter Plot – add third dimension

We will add the petal length to the above scatterplot, creating a 3-dimensional plot in cartesian coordinates. We expect to use the mplot3d toolkit, ax.scatter3D function.

* Explore the dataset (python tools)

We will use python pandas to discover data types, view completeness of data, and explore data ranges.

* Calculate the dot product.

We will use 2 pandas dataframes to calculate the dot product.

* Calculate the norm of the vectors

Numpy will calculate a norm of a vector.

* Calculate the distance between vectors

We will use numpy linalg norm function to calculate the distance between vectors.

## Proposed Outputs

The assignment breaks down outputs into three categories:

Data exploration: This will be the resulting values from python pandas dataframes.

Data visualization: This will be graphs generated in matplotlib.

Data calculations: Numpy results will be displayed as values, arrays or pandas dataframes.

## Proposed Visualization

## There will be three types of visualization to explore the iris data set

1. Will be a box plot for each of the 4 attributes petal length, petal width, sepal length, and sepal width
2. A correlation matrix for each of the 4 attributes petal length, petal width, sepal length, and sepal width
3. A 3D scatter plot for the 3 attributes petal length, petal width, and sepal length and will be adding the color based on species

## Conclusions

The main goals of this assignment are exploring, visualizing, and making calculations. The outcome of this assignment should be to understand the capabilities of python in exploring and understanding a dataset as well as evaluating and displaying visualizations representative of the data.