## CS4431 HW4

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

## 2022-07-07

# Week 4 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**

**Library source**

## Proposed Solution

Elbow-plot, silhouette method, and gap-stat plots

PCA outputs and changes before and after implementation

PCA visualization in collapsed dimensions

Cluster models and appropriate visualizations

Prose analysis and conclusions from cluster models

## Proposed Outputs

## Proposed Visualization

There will be an elbow / scree plot created from the iris dataset.

There will be a scatterplot as output from the Principal Component Analysis of the iris dataset.

There will be a visualization for each of the clustering techniques employed.

## Conclusions