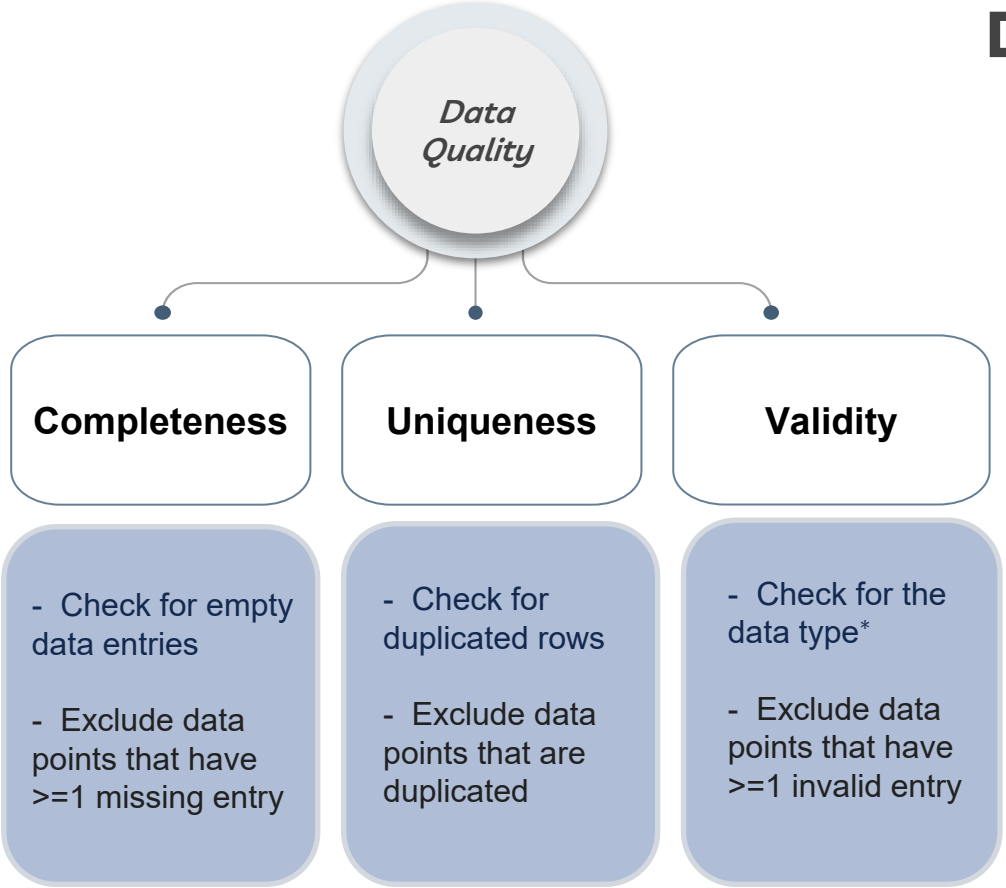
Data Science Quest 1 (Approach & Findings)

Chia Yu Ying

TABLE OF CONTENTS

Data Cleaning

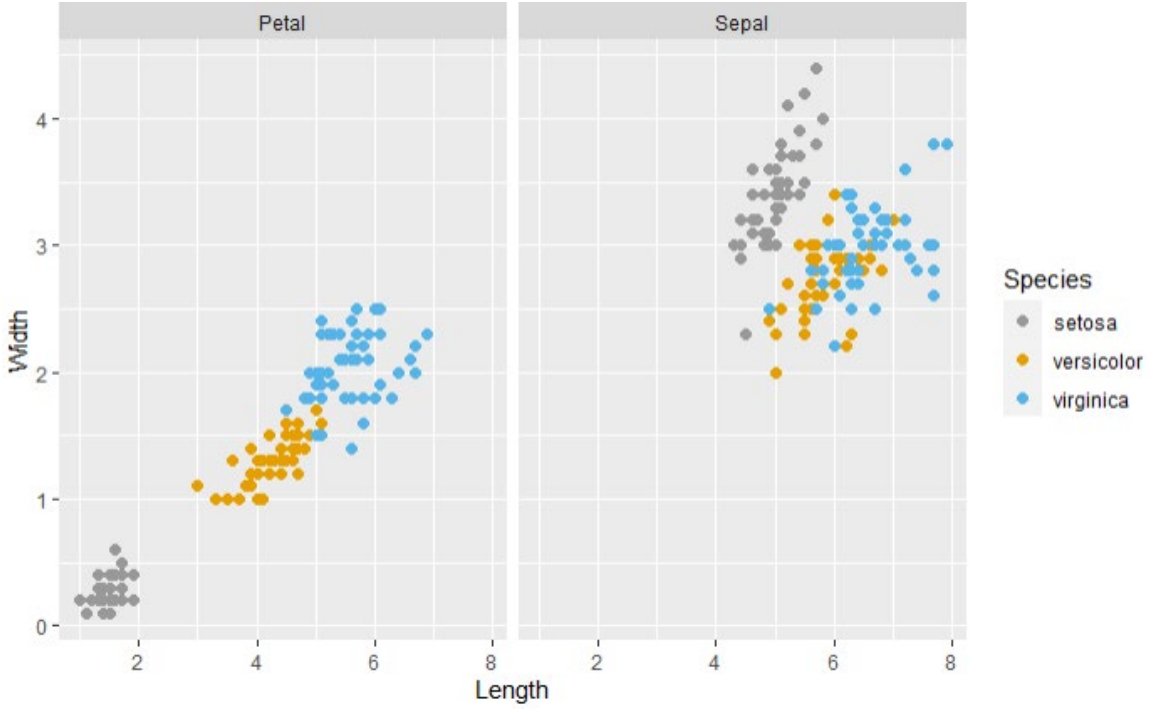
Data Cleaning



* Data Type example: We should expect numeric entries for sepal length

Data Visualization (Before Analysis)

Data Visualization (Before Analysis)

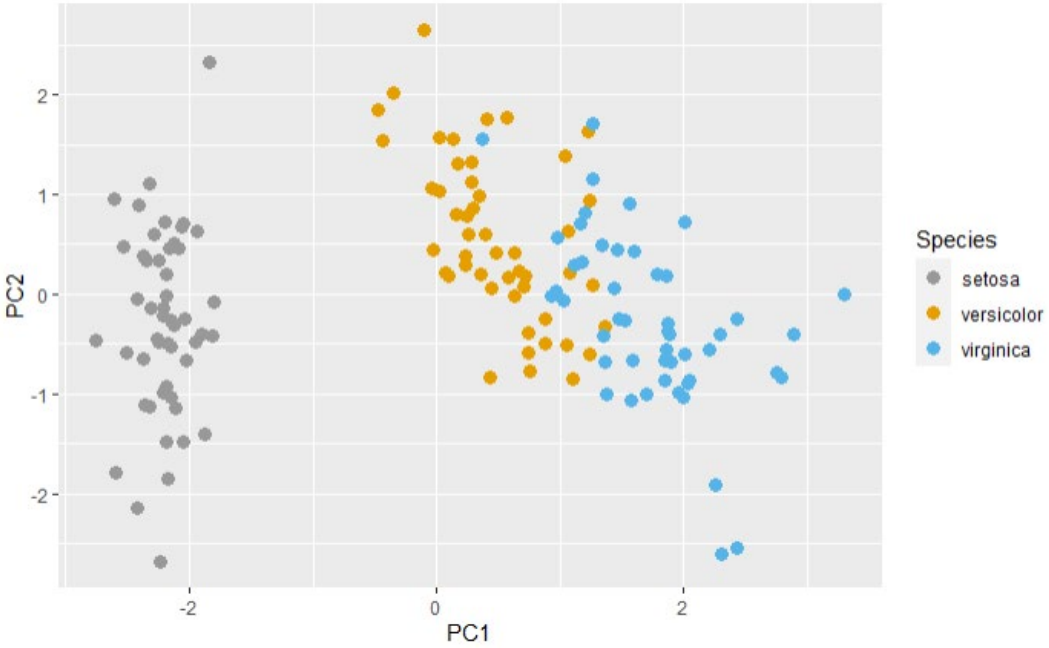


Multi-Dimensional Plot

Visualizes all 4 feature variables
(Grouped by Species Type)
Sepal Length; Sepal Width
Petal Length; Petal Width

Insight
It is generally easier identify a Setosa flower as compared to the other 2 species

Data Visualization (Before Analysis)



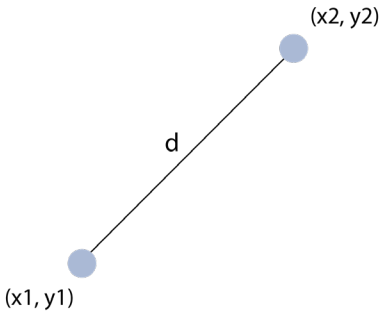
PCA Plot

Visualizes 2 Principal Components
(Grouped by Species Type)
Principal Component 1; Principal Component 2

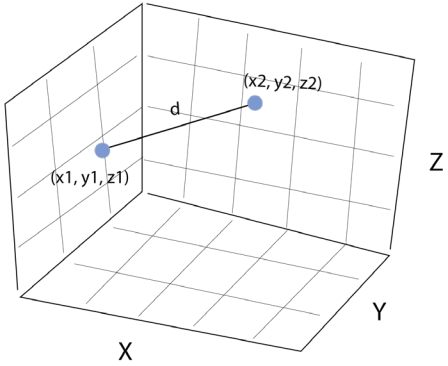
Insight
It is generally easier identify a Setosa flower as compared to the other 2 species

Approach

Similarity Measure (Euclidean Distance)



2D: Euclidean distance (d) = $\sqrt{(x1-x2)^2+(y1-y2)^2}$



3D: Euclidean distance (d) = $\sqrt{(x1-x2)^2+(y1-y2)^2+(z1-z2)^2}$

4D: Euclidean distance (d) = $\sqrt{(x1-x2)^2+(y1-y2)^2+(z1-z2)^2+(p1-p2)^2}$

Iris Data Set

Overview

Similarity measure function

Iris Data Set
(Cleaned)
+
Data measured from
new Iris plant

Input

Calculates
Euclidean Distance

+

Sort in **Ascending** order

Output
(First 10 flowers with shortest Euclidean Distance)

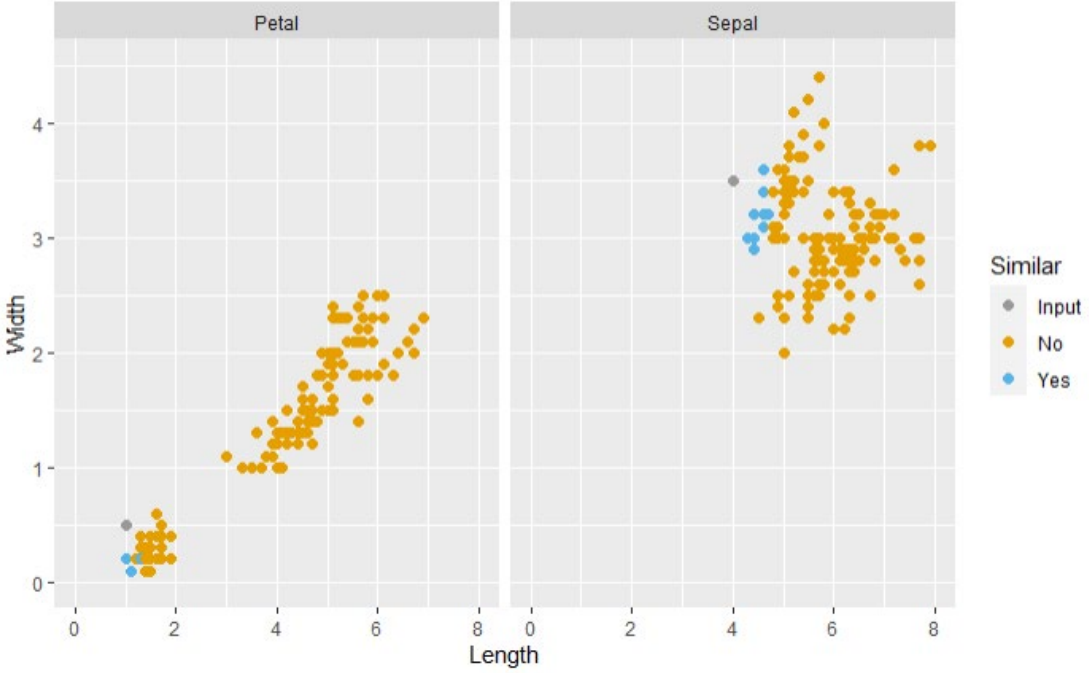
Result visualization

Result Visualization

(Scenario) Imagine the garden owner discovered a **new iris plant** with the following attributes:

Sepal Length	4.0 cm
Sepal Width	3.5 cm
Petal Length	1.0 cm
Petal Width	0.5 cm

Result Visualization

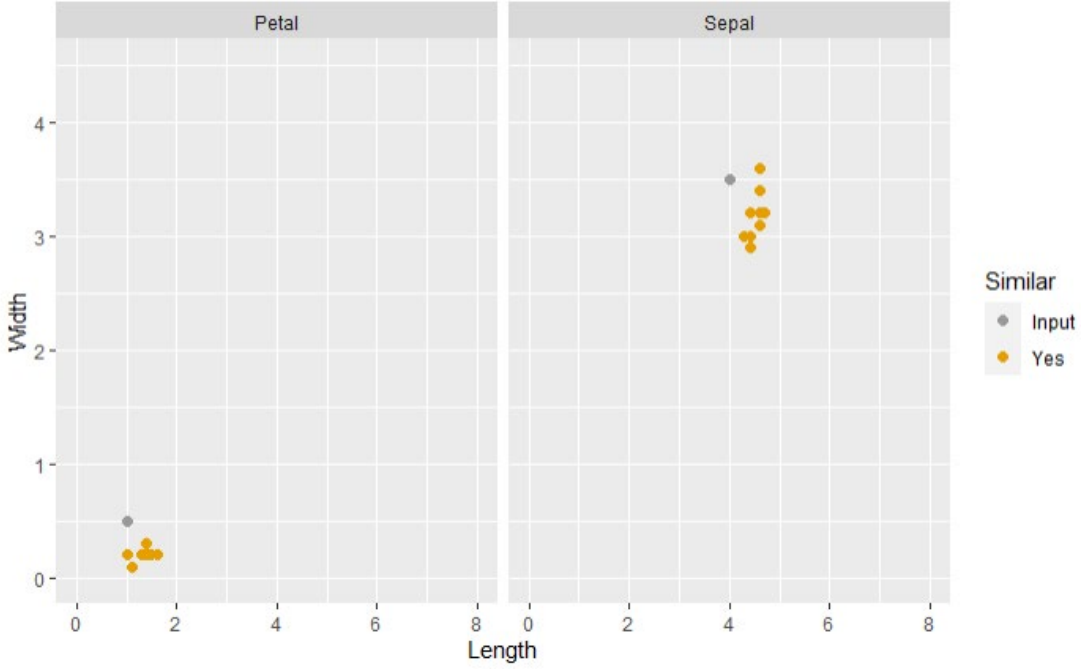


Multi-Dimensional Plot

Visualizes all 4 feature variables
Sepal Length; Sepal Width
Petal Length; Petal Width

Plot description:
Grey: Input data specified by garden owner
Blue: Top 10 flowers similar to input
Orange: Not similar to input

Result Visualization

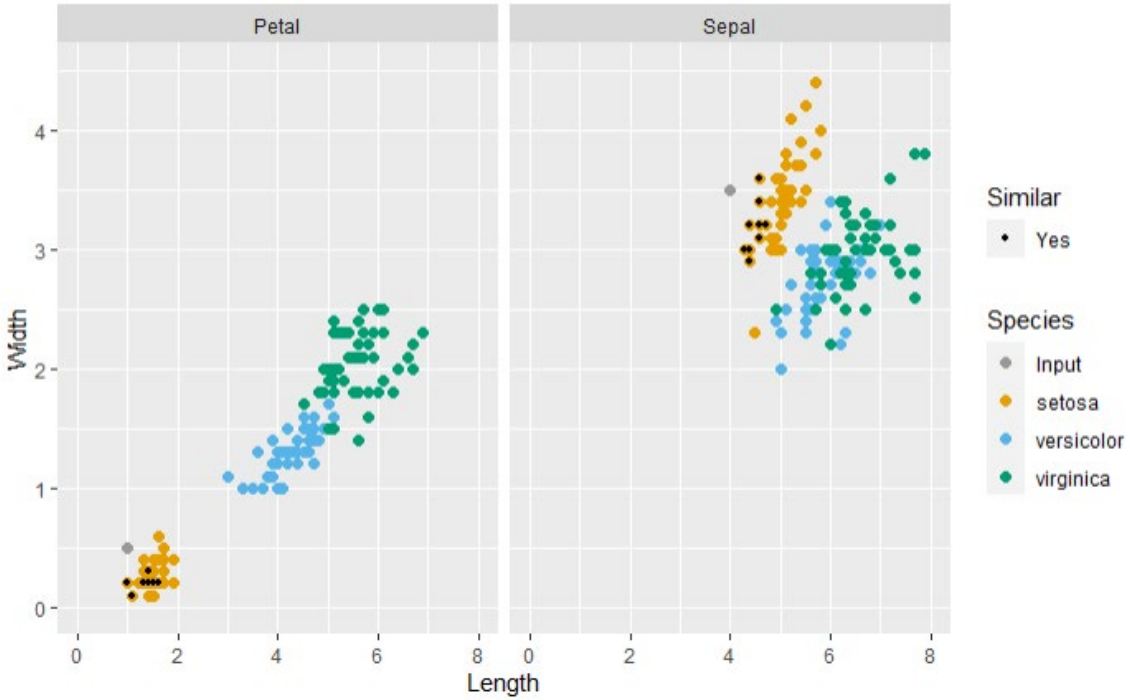


Multi-Dimensional Plot

Visualizes all 4 feature variables
Sepal Length; Sepal Width
Petal Length; Petal Width

Plot description:
Grey: Input data specified by garden owner
Orange: Top 10 flowers similar to input
(Excludes the rest of the data points)

Result Visualization

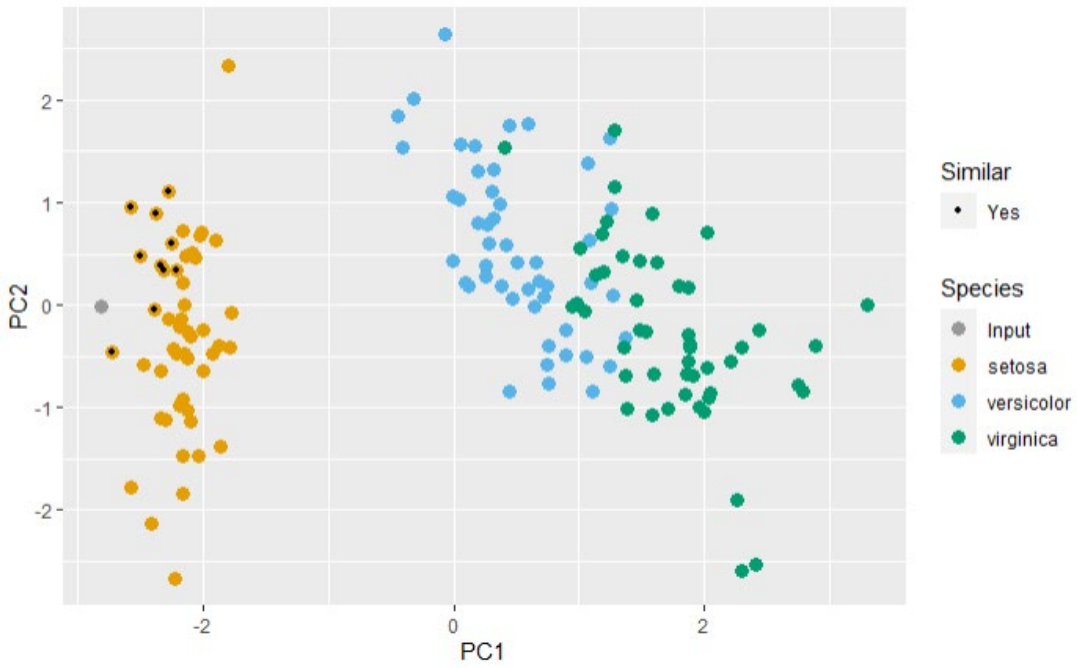


Multi-Dimensional Plot

Visualizes all 4 feature variables
(Grouped by Species Type)
Sepal Length; Sepal Width
Petal Length; Petal Width

Plot description:
Grey: Input data by garden owner
Orange: Setosa Type
Blue: Versicolor Type
Green: Virginica Type
Black inner dot: Top 10 flowers similar to input

Result Visualization



PCA Plot

Visualizes 2 Principal Components
(Grouped by Species Type)
Principal Component 1; Principal Component 2

Plot description:
Grey: Input data by garden owner
Orange: Setosa Type
Blue: Versicolor Type
Green: Virginica Type
Black inner dot: Top 10 flowers similar to input

Improvements

Improvements

Given more time . . .

Some areas for improvements:

- Explore other similarity measures such as Cosine similarity, Manhattan distance etc.
- Perform some sort of performance measures on different similarity measures and see which one is the most suitable for Iris Dataset
- Make the plots look more attractive