**Tor Thogersen**

**DA 410 -- Multivariate Analysis -- Winter 2018**

**Project 4**

**Part 1**: Complete the 5 questions in Handout 5 Code. Make sure you include the commands and outputs, as well as the interpretations of the outputs.

**# loading iris dataset in R**

> attach(iris)

> View(iris)

> head(iris)

Sepal.Length Sepal.Width Petal.Length Petal.Width Species

1 5.1 3.5 1.4 0.2 setosa

2 4.9 3.0 1.4 0.2 setosa

3 4.7 3.2 1.3 0.2 setosa

4 4.6 3.1 1.5 0.2 setosa

5 5.0 3.6 1.4 0.2 setosa

6 5.4 3.9 1.7 0.4 setosa

> names(iris) <- c("Sepal.Length", "Sepal.Width", "Petal.Length",

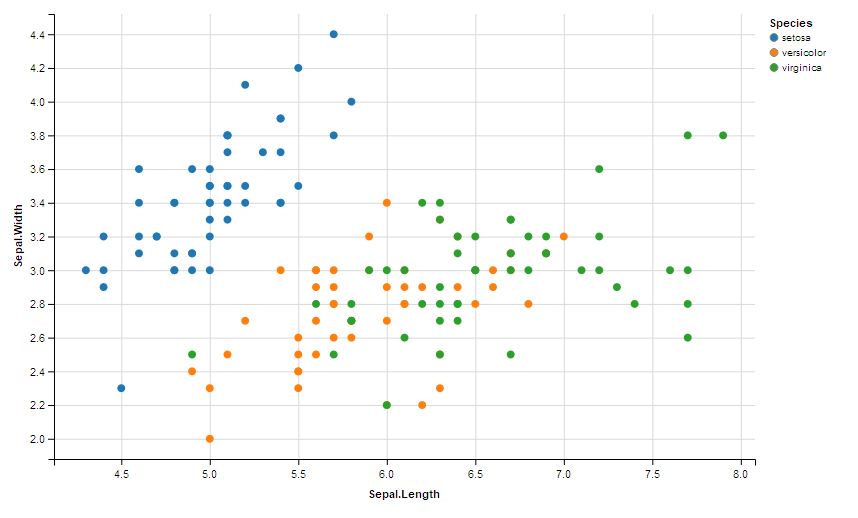
+ "Petal.Width", "Species")

> names(iris)

[1] "Sepal.Length" "Sepal.Width" "Petal.Length" "Petal.Width" "Species"

***Question 1:***

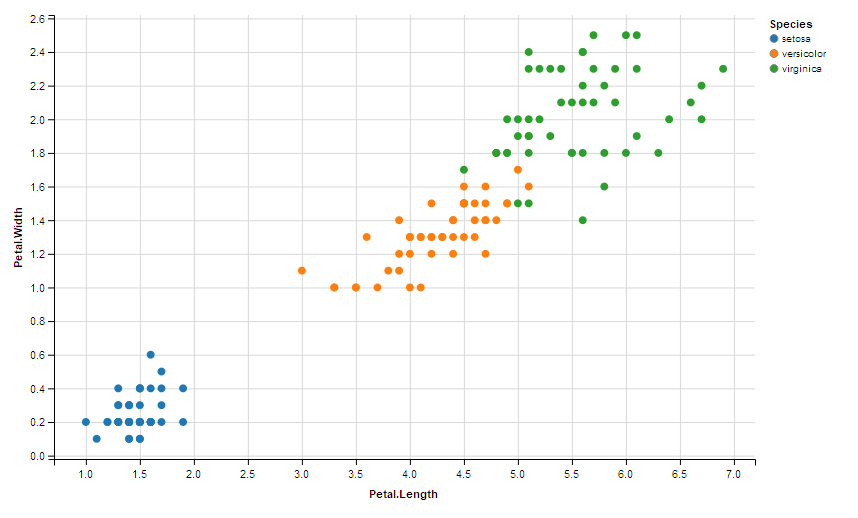
> iris %>% ggvis(~Sepal.Length, ~Sepal.Width, fill = ~Species) %>%layer\_points()



* What is the output?
  + Output shows the 3 groups of species of plants and the association between sepal.lenght verus Sepal.width
* Based on the plot, which specie(s) has the highest correlation between the sepal length and the sepal width?
  + Based on the data I believe Setosa has the highest correlation between the two X and Y variables, you see that as length increase the width exponentially increases.

***Question 2:***

> iris %>% ggvis(~Petal.Length, ~Petal.Width, fill = ~Species) %>% layer\_points()



* What is the output?
  + Similar results as the previous chart expect it shows the Petal.Width versus Petal.Length
* Based on the plot, which specie(s) has the highest correlation between the petal length and the petal width?
  + Verginica has the highest relationship between width and length of petals.
* Is this result consistent with Question 1? Why or why not?
  + No, because it shows that all three species share similar linear regression lines but the greatest relationship +/- is Virginica plant species versus Setosa species.

# setting up the training and test samples for iris dataset

> set.seed(1234)

> ind <- sample(2, nrow(iris), replace=TRUE, prob=c(0.67, 0.33))

> iris.training <- iris[ind==1, 1:4]

> iris.test <- iris[ind==2, 1:4]

> iris.trainLabels <- iris[ind==1, 5]

> iris.testLabels <- iris[ind==2, 5

# The Actual KNN Model

# Building Your Classifier

> library(class)

> iris\_pred <- knn(train = iris.training, test = iris.test, cl =iris.trainLabels, k=3)

***Question 3:***

> iris\_pred

[1] setosa setosa setosa setosa setosa setosa setosa setosa setosa setosa

[11] setosa setosa versicolor versicolor versicolor versicolor versicolor versicolor versicolor versicolor

[21] versicolor versicolor versicolor versicolor virginica virginica virginica virginica versicolor virginica

[31] virginica virginica virginica virginica virginica virginica virginica virginica virginica virginica

Levels: setosa versicolor virginica

* What is the output?
  + Output of the results from the iris\_pred stored results from the knn( ) function that takes the output from the test set and training labels.

***Question 4:***

Predicted Species Observed Species

1 setosa setosa

2 setosa setosa

3 setosa setosa

4 setosa setosa

5 setosa setosa

6 setosa setosa

7 setosa setosa

8 setosa setosa

9 setosa setosa

10 setosa setosa

11 setosa setosa

12 setosa setosa

13 versicolor versicolor

14 versicolor versicolor

15 versicolor versicolor

16 versicolor versicolor

17 versicolor versicolor

18 versicolor versicolor

19 versicolor versicolor

20 versicolor versicolor

21 versicolor versicolor

22 versicolor versicolor

23 versicolor versicolor

24 versicolor versicolor

25 virginica virginica

26 virginica virginica

27 virginica virginica

28 virginica virginica

29 versicolor virginica

30 virginica virginica

31 virginica virginica

32 virginica virginica

33 virginica virginica

34 virginica virginica

35 virginica virginica

36 virginica virginica

37 virginica virginica

38 virginica virginica

39 virginica virginica

40 virginica virginica

***What is the command that may produce above output?***

> irisTestLabels <- data.frame(iris.testLabels)

> merge <- data.frame(iris\_pred, iris.testLabels)

> names(merge) <- c("Predicted Species", "Observed Species")

> merge

**Question 5:**

> library(gmodels)

> CrossTable(x = iris.testLabels, y = iris\_pred, prop.chisq=FALSE)

Cell Contents

|-------------------------|

| N |

| N / Row Total |

| N / Col Total |

| N / Table Total |

|-------------------------|

Total Observations in Table: 40

| iris\_pred

iris.testLabels | setosa | versicolor | virginica | Row Total |

----------------|------------|------------|------------|------------|

setosa | 12 | 0 | 0 | 12 |

| 1.000 | 0.000 | 0.000 | 0.300 |

| 1.000 | 0.000 | 0.000 | |

| 0.300 | 0.000 | 0.000 | |

----------------|------------|------------|------------|------------|

versicolor | 0 | 12 | 0 | 12 |

| 0.000 | 1.000 | 0.000 | 0.300 |

| 0.000 | 0.923 | 0.000 | |

| 0.000 | 0.300 | 0.000 | |

----------------|------------|------------|------------|------------|

virginica | 0 | 1 | 15 | 16 |

| 0.000 | 0.062 | 0.938 | 0.400 |

| 0.000 | 0.077 | 1.000 | |

| 0.000 | 0.025 | 0.375 | |

----------------|------------|------------|------------|------------|

Column Total | 12 | 13 | 15 | 40 |

| 0.300 | 0.325 | 0.375 | |

----------------|------------|------------|------------|------------|

**What is the output?**

The contingency table showing the relationship between two variables of the

iris dataset, your test data against your prediction data.