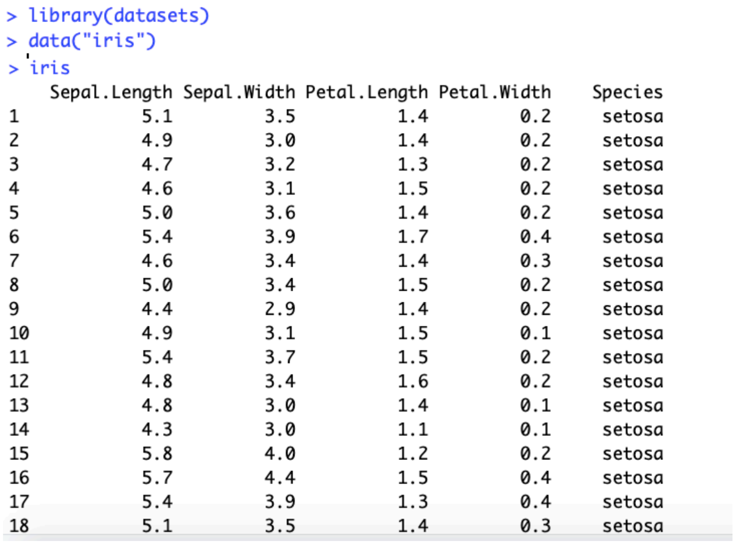
Project Description
CS 158/458 R Programming 1

Create a folder ‘CS158’ (if you are in CS 158) or ‘CS458’ (if you are in CS 458). Inside ‘CS158’ / ‘CS458’ create a subfolder named ‘FinalProject’. In RStudio, set your working directory to CS158/FinalProject or CS458/ FinalProject. Open a new script, save it as ‘FinalProject.R’ (make sure it is saved under CS158/FinalProject or CS458/FinalProject). Include your name, date and purpose of the project as the header (top) of the FinalProject.R script.

In RStudio, load the “datasets” package by invoking it through the ‘library’ command (example shown below). Next, type **iris** and you should see the following dataset



```
> library(datasets)
> data("iris")
> 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
7           4.6         3.4          1.4         0.3     setosa
8           5.0         3.4          1.5         0.2     setosa
9           4.4         2.9          1.4         0.2     setosa
10          4.9         3.1          1.5         0.1     setosa
11          5.4         3.7          1.5         0.2     setosa
12          4.8         3.4          1.6         0.2     setosa
13          4.8         3.0          1.4         0.1     setosa
14          4.3         3.0          1.1         0.1     setosa
15          5.8         4.0          1.2         0.2     setosa
16          5.7         4.4          1.5         0.4     setosa
17          5.4         3.9          1.3         0.4     setosa
18          5.1         3.5          1.4         0.3     setosa
```

Type -- ?iris to get the details of this dataset. You can load the iris dataset into the environment by invoking data(‘iris’).

Question 1) Answer the following questions about this dataset using suitable code: [**15 points]**

a) How many unique 'Species' are there in the data-frame?
b) What is the total number of records (total rows)?
c) How many records per 'Species'(number of rows per species)?
d) How many attributes in the dataset?
e) Provide a summary (min, mean, max, median) of the first four attributes.

Question 2) Pearson's correlation method measures the strength of association and direction of relationship between two continuous variables. It provides a value between -1 to +1 wherein +1 indicates the perfectly positive relationship and -1 indicates a perfectly negative relationship. So, the sign of the correlation score determines the direction whereas the magnitude determines the strength, thus a score of 0 means the two variables are independent [**20 points]**

a) Find the relationship between **Petal.Length** and **Petal.Width**.

ex: cor(iris$Petal.Length, iris$Petal.Width)

b) Find the relationship between **Sepal.Length** and **Sepal.Width**.

Describe the results obtained in few words.

Question 3) Generate scatter plots between [**20 points]**

1) **Petal.Length** and **Petal.Width (**save the file as AssoPetal.pdf**)**
2) **Sepal.Length** and **Sepal.Width (**save the file as AssoSepal.pdf **)**

Use appropriate x and y labels and suitable heading for your plot. Also see if you can verify the trend from question 2 in your plots.

Question 4) Pearson's Chi square test can also be used as a test of independence between two discrete variables (refer slides on statistical analysis for details on the usage of Chi square test)

Determine through statistical analysis if the Petal.Width and Species are dependent, that is , if there is a relationship between the two variables . You need to follow these steps: [**45 points]**

1) Utilizing the knowledge from previous lectures, use Ckmeans package to discretize the **Petal.Width**. You should get a vector of repeating integers.
   a dummy ex could be : c(2,2,2,2,1,1,1,1,2,2,2)
   Call this vector 'X'. [3 points]
2) Now create a vector Y of **Species** [3 points]
   Ex: Y <- iris$Species
3) Use the above two vectors in a method table(X, Y) to create a contingency table (or frequency table). This will be given as an input to the chisq.test(). [3 points]
4) Follow all the steps that we learned in the statistical analysis lecture and using the chisq.test() at the standard p-value cut off (p-value <= 0.05) determine if there is dependence between the **Petal.Width** and **Species**. Write all your observation and statements in a separate text. [20 points]
5) Create a boxplot names ChiPetal.pdf between Petal.Width and Species and see if you can observe a trend. Use appropriate x and y labels and title for your plot. [10 points]
6) Observing ChiPetal.pdf can you see a relationship between Petal.Width and Species such that it can be utilized by domain specialist (or you) to distinguish the three species of Iris. State your observation as well as whether or not the three Iris species can be distinguished by Petal.Width with proper explanation in either case. [6 points]

Submit FinalProject.R, AssoPetal.pdf, AssoSepal.pdf ChiPetal.pdf and separate text for your answers.