DSC Challenge #1 - Iris Data Set

The purpose:

- An introduction to the data science process.
- Set up development environment (python/R) - (http://becomingadatascientist.com/learningclub/thread-7.html)
- Carry out simple Exploratory Data Analysis.
- Share your work!

The Data:
https://archive.ics.uci.edu/ml/datasets/Iris

But it can also be accessed internally in R & Python (SciKitLearn)
Python:

```
from sklearn import datasets

iris = datasets.load_iris()
```

R:

```
head(iris)
```

Challenge:
1) Open the dataset and view the data

- How many variables?
- How many observations?
- How else can the data be grouped?
- How can you best visualise this dataset?

2) Exploratory Data Analysis (EDA) & summary statistics

- How are the observations for each variable distributed?
- What is the mean, mode, median for each dataset? Are these values meaningful?
- What relationships are present between variables?
- Can you find the variables that account for the greatest variance? (Dimensionality Reduction)
- Can you identify outliers?
- How do variables covary?

3) Further challenges

- How would you classify the data into different species?
- Is clustering a useful process for this data?
- What models/algorithms would you use?

Exercise:

- Build a 5 minute presentation of your process & findings and prepare to share with the group at the next meeting.
- Share your code and analysis to the group.
- Some ideas:
    - Use a Jupyter Notebook/Rmarkdown
    - Build an interactive tool for exploring the data.
- Complete Survey:
    - https://www.surveymonkey.co.uk/r/9BPWFLN