

# Iris Dataset Analysis

## 1 - Introduction

We use the built-in `iris` dataset within R to demonstrate a simple data visualisation. This dataset is comprised of 150 observations, each recording four characteristics of the iris flowers (petal length, petal width, sepal length, sepal width) and their species.

### *Data Loading*

Lets bring the data into our local environment and view the first few rows of the dataset.

	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

### *Data Structure*

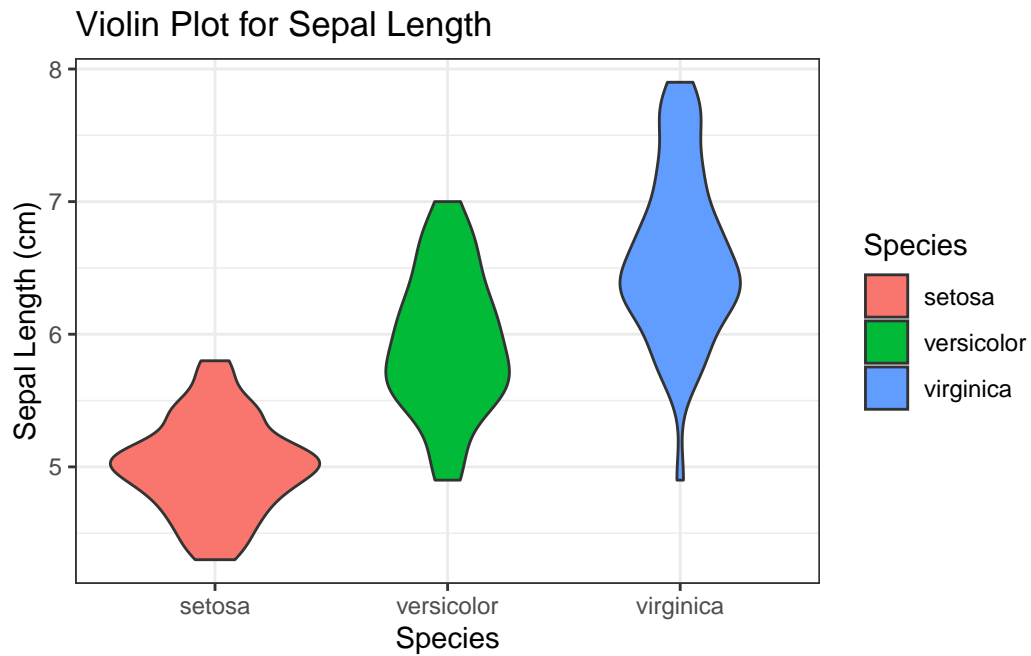
Diving deeper into the datasets composition and the various data types.

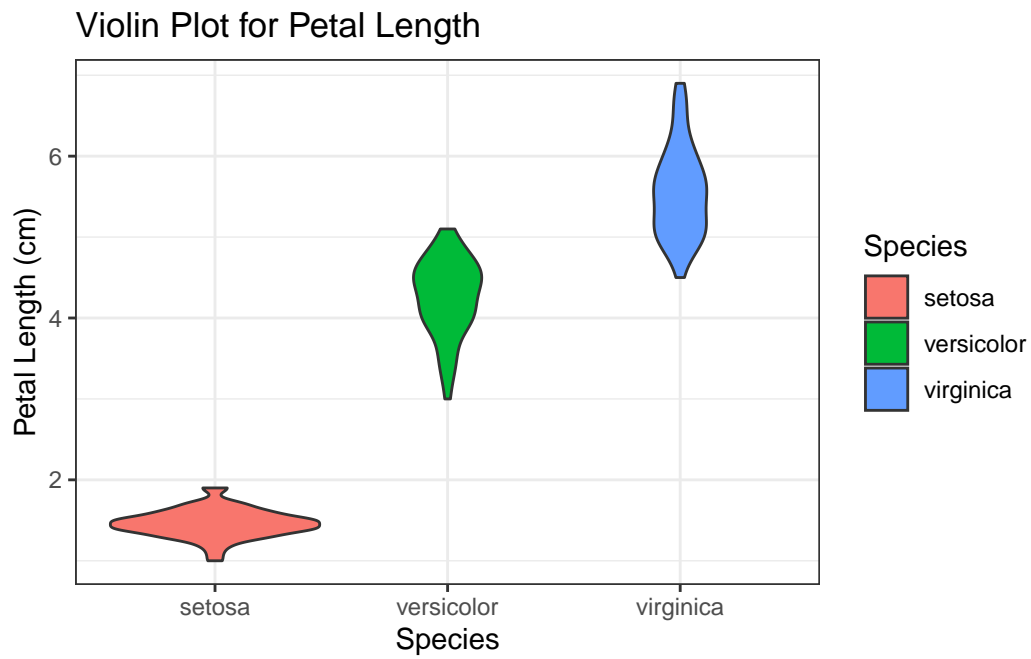
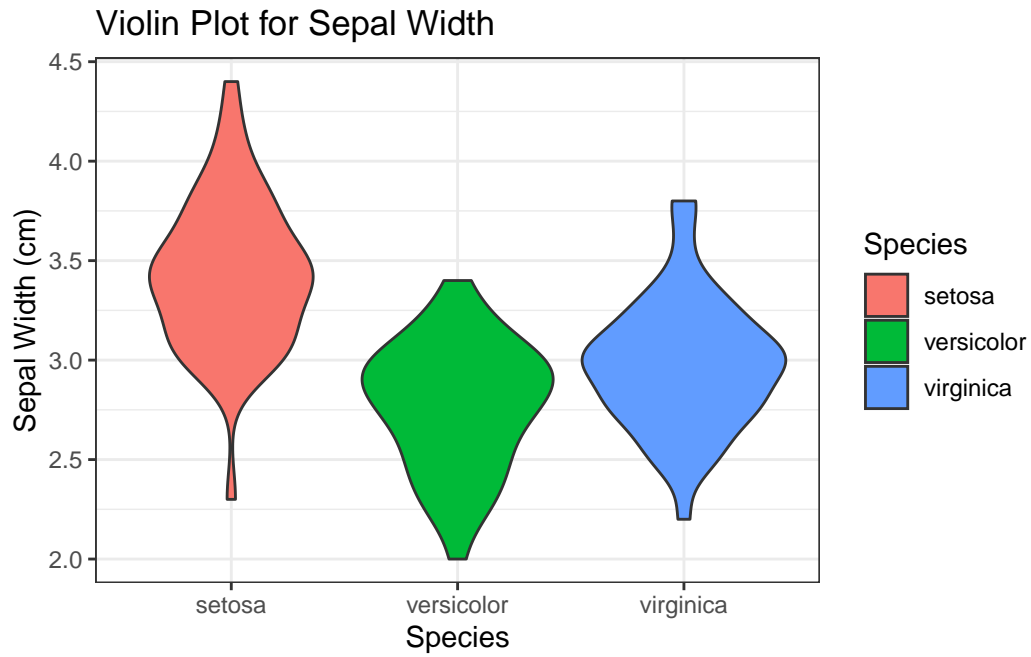
```
'data.frame': 150 obs. of 5 variables:
 $ Sepal.Length: num 5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num 3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num 0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species : Factor w/ 3 levels "setosa","versicolor",...: 1 1 1 1 1 1 1 1 1 1 ...
```

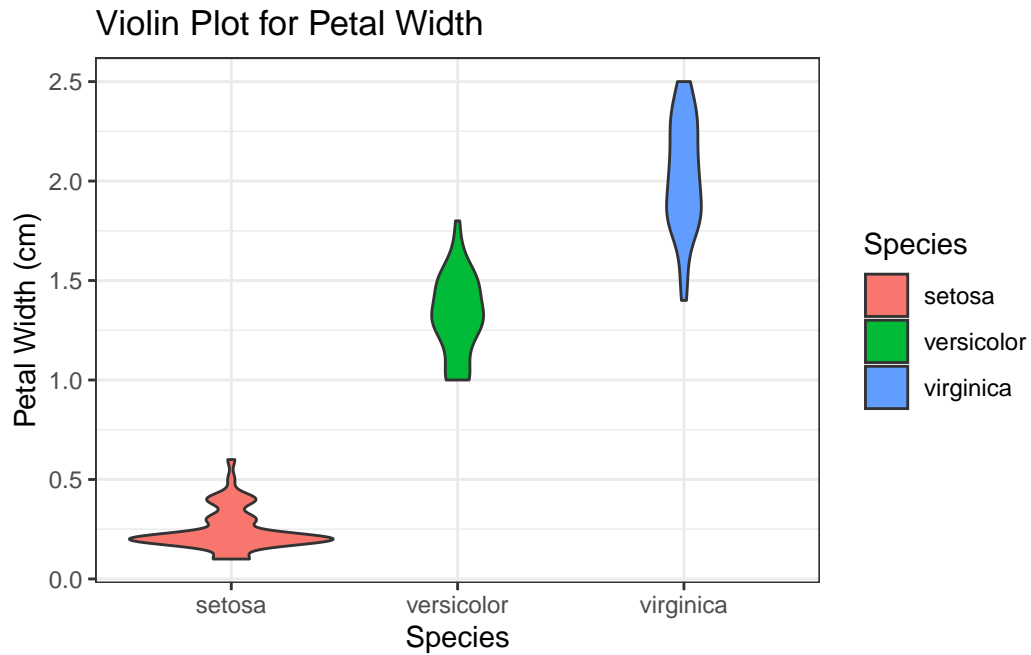
## 2 - Visualisation

### *Violin Plots for petal and sepal measurements across each species*

These provide information on both the distribution and the frequency of values which can describe the differences between the species. `ggplot2` is used to plot these.







### ***Visualisations Summary***

These plots clearly show the distinctive differences among the 3 species across both petal and sepal lengths and widths. Setosa flowers have a tight distribution in their petal length and width, around 1.5 and 0.25 respectively, whereas both versicolor and virginica have much longer lengths in this category with a wider, more spread out distribution.