1.4.2 Data Exploration Example

Stat 5100: Dr. Bean

Example: Here we will do some various data manipulations and explorations in R. We will look at the "iris" dataset, a very famous dataset that is automatically available in R.

```
# Look at the first 6 observations in the iris dataset:
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
```

```
# What variables are contained in this dataset?
names(iris)
```

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

```
# What variable types are the columns?
str(iris)
```

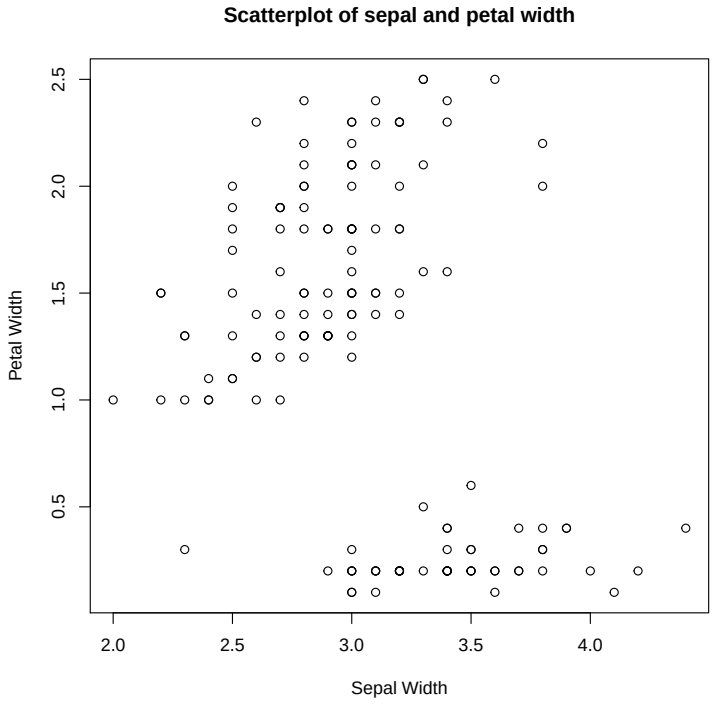
```
## 'data.frame': 150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3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

```
# How many observations (rows) does the dataset have?
nrow(iris)
```

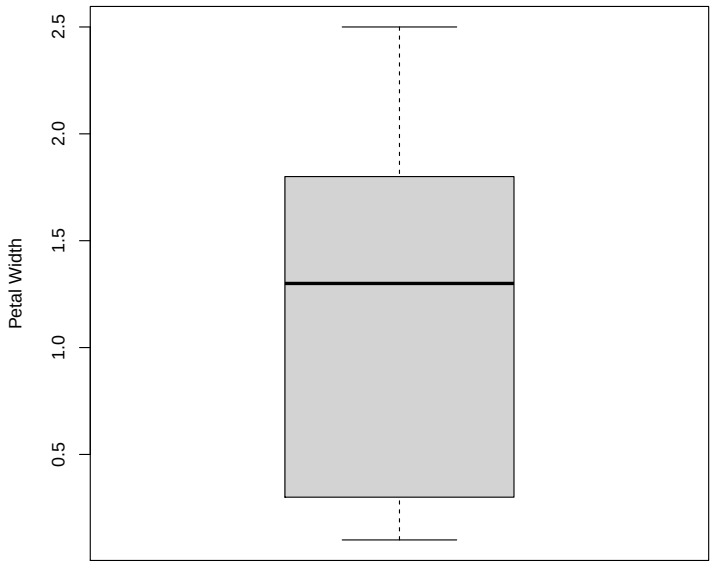
```
## [1] 150
```

Now, let's create some graphics to explore this dataset a bit more:

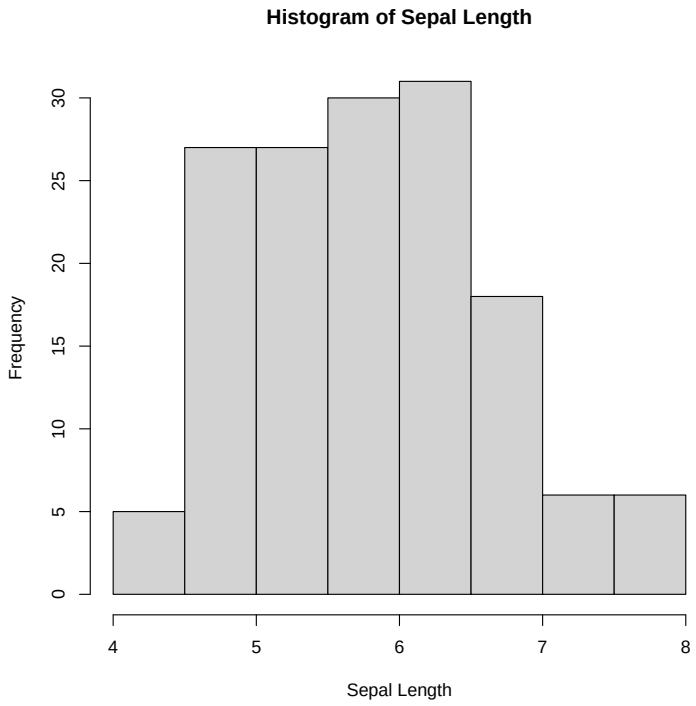
```
# Create a scatterplot of sepal width and petal width
plot(iris$Sepal.Width, iris$Petal.Width, main = "Scatterplot of sepal and petal width",
     xlab = "Sepal Width", ylab = "Petal Width")
```



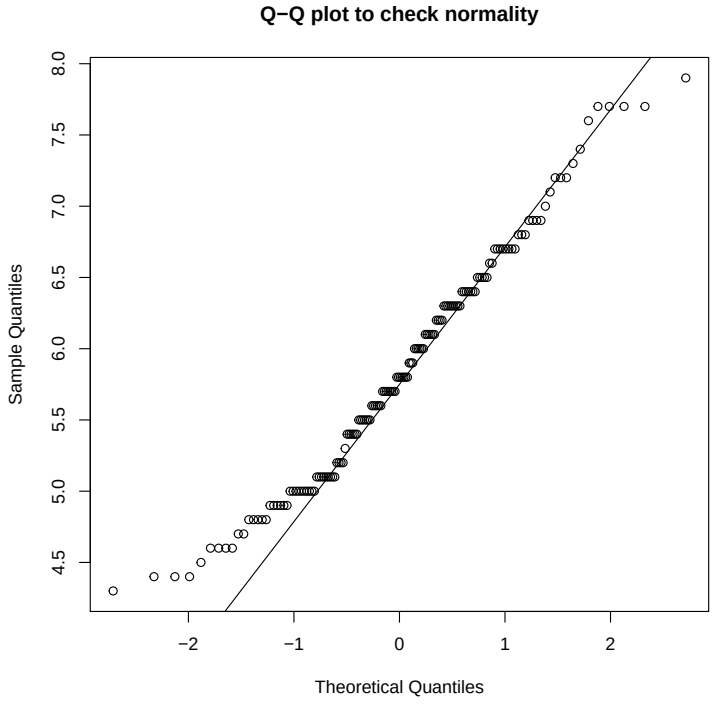

```r
# Create a box plot of petal width
boxplot(iris$Petal.Width, ylab = "Petal Width")
```




```r
# How normally distributed is the sepal length variable? Create both a
# histogram and a Q-Q plot to check.
hist(iris$Sepal.Length, main = "Histogram of Sepal Length", xlab = "Sepal Length")
```




```
qqnorm(iris$Sepal.Length, main = "Q-Q plot to check normality")
qqline(iris$Sepal.Length)
```

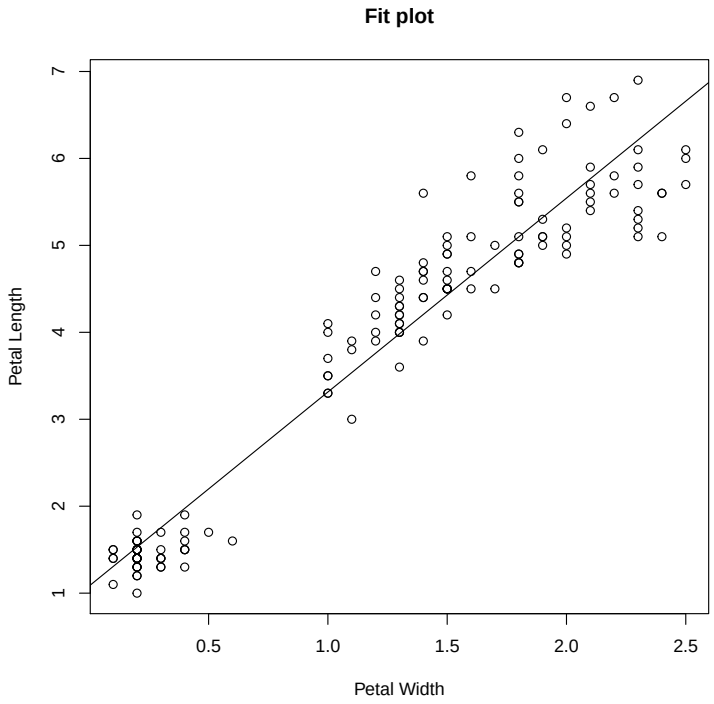


Let's create a linear model where we predict sepal length from sepal width.

```
iris_lm <- lm(Petal.Length ~ Petal.Width, data = iris)

stat5100::fit_plot(iris_lm, main = "Fit plot", xlab = "Petal Width",
```

```
                    ylab = "Petal Length")
```



What if we want to make a prediction using our linear model? Suppose that we have two flowers with petal widths of 0.3 and 2.4.

```
my_beautiful_two_flowers <- data.frame(Petal.Width = c(0.3, 2.4))

predicted_petal_length <- predict(iris_lm, my_beautiful_two_flowers)
predicted_petal_length

##        1        2
## 1.752540 6.435415
```

Based upon the above, for the flower with the petal width of 0.3 cm we would predict that the length is 1.75 cm and for the petal width of 2.4 cm we would predict that the length is 6.44 cm.