Question 3 (50 points)

a) Using R, visualize the Iris dataset with 10 different univariate techniques. Include at least 2 glyph-based approaches. For all graphs, use labels as appropriate. Submit the documented source code (40 marks)

b) submit a pdf showing screen shots and a discussion of what type of information you can extract from your graphs (10 marks)

Using R, I used 10 different univariate techniques to represent the Iris data. First, downloaded the iris dataset package. There are

3 species : Virginica, Setosa, Versicolor

4 variables : sepal.length, sepal.width, petal.length, petal.width

Iris

```
> summary(iris)
  sepal.length  sepal.width   petal.length  petal.width        species
 Min.   :4.3   Min.   :2.0   Min.   :1.0   Min.   :0.1   setosa    :50
 1st Qu.:5.1   1st Qu.:2.8   1st Qu.:1.6   1st Qu.:0.3   versicolor:50
 Median :5.8   Median :3.0   Median :4.3   Median :1.3   virginica :50
 Mean   :5.8   Mean   :3.1   Mean   :3.8   Mean   :1.2
 3rd Qu.:6.4   3rd Qu.:3.3   3rd Qu.:5.1   3rd Qu.:1.8
 Max.   :7.9   Max.   :4.4   Max.   :6.9   Max.   :2.5
> |
```

Virginica

```
> head(virginica)
  sepal.length sepal.width petal.length petal.width   species
1          6.3         3.3          6.0         2.5 virginica
2          5.8         2.7          5.1         1.9 virginica
3          7.1         3.0          5.9         2.1 virginica
4          6.3         2.9          5.6         1.8 virginica
5          6.5         3.0          5.8         2.2 virginica
6          7.6         3.0          6.6         2.1 virginica
> |
```

Setosa

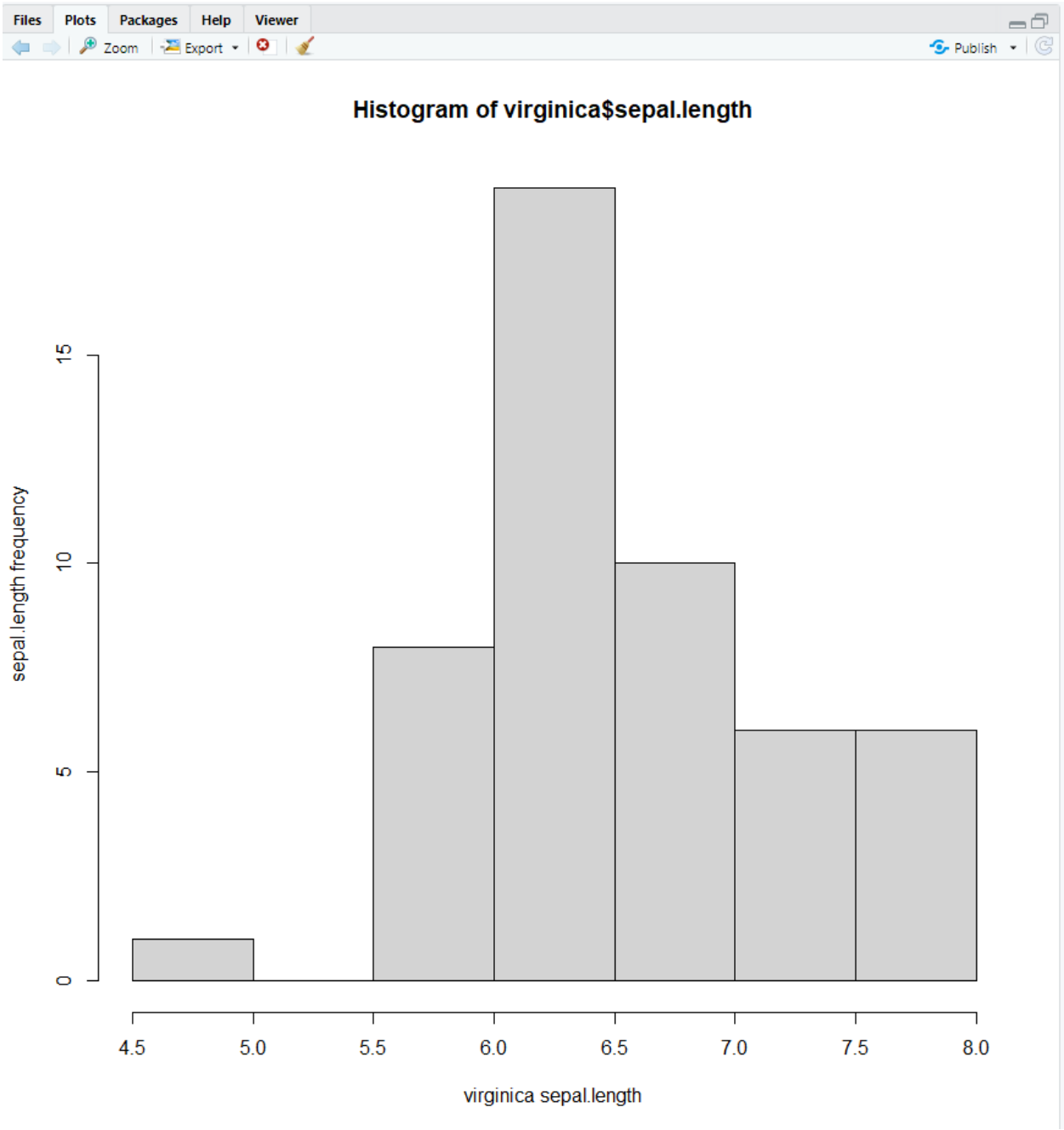
```
> head(setosa)
  sepal.length sepal.width petal.length petal.width species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
> |
```

Versicolor

```
> head(versicolor)
  sepal.length sepal.width petal.length petal.width    species
1          7.0         3.2          4.7         1.4 versicolor
2          6.4         3.2          4.5         1.5 versicolor
3          6.9         3.1          4.9         1.5 versicolor
4          5.5         2.3          4.0         1.3 versicolor
5          6.5         2.8          4.6         1.5 versicolor
6          5.7         2.8          4.5         1.3 versicolor
> |
```

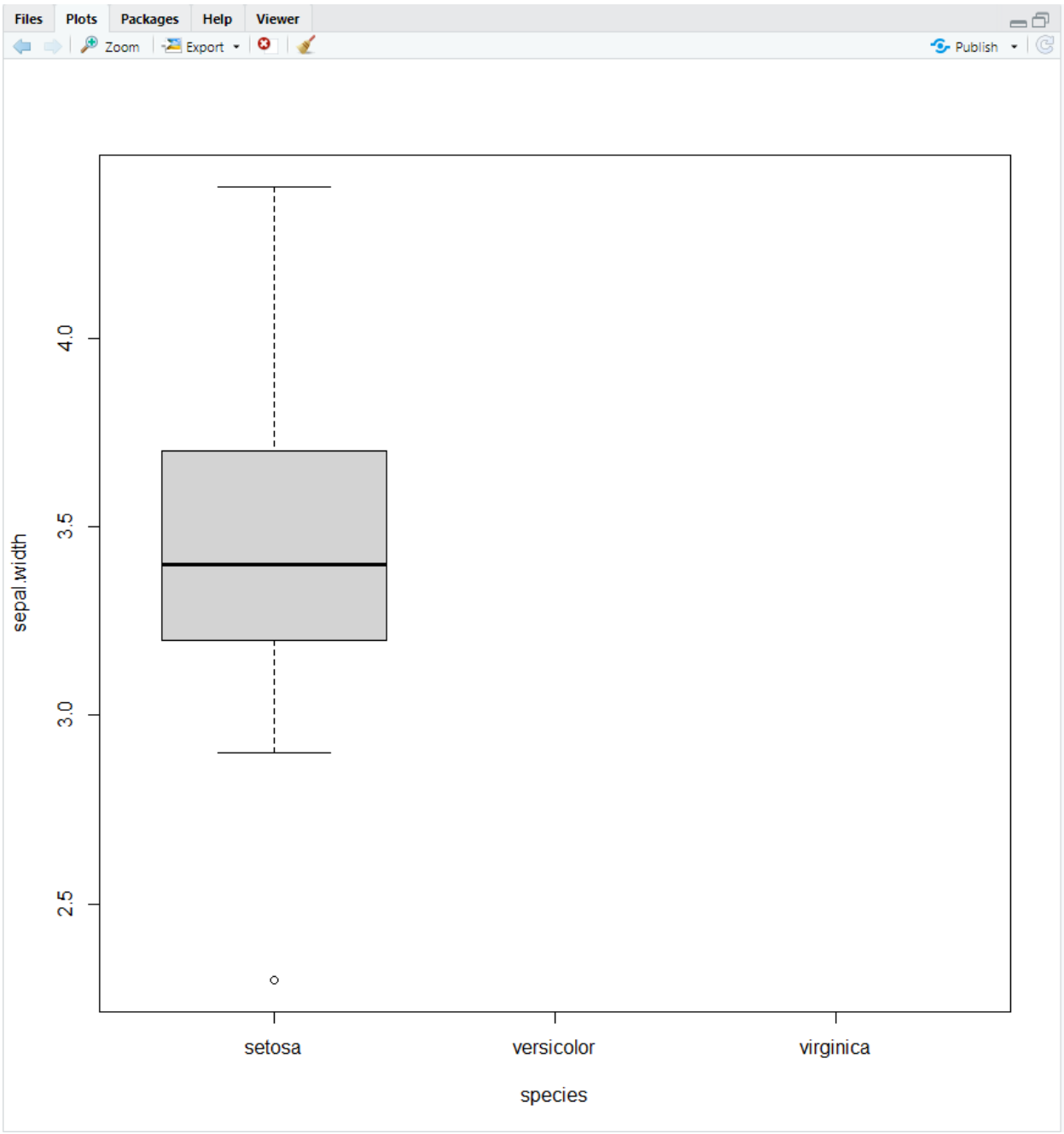
1) **Histogram**

Created a histogram for the values of sepal.length for the species virginica. The data that we can extract from this graph is first of all what data it is? So as the title says, it is a data for the sepal.length values of virginica. The y-axis says that it shows frequency of every value of sepal.length. For instance, for species Virginica, the sepal.length values between 4.5 and 5.0, the frequency/number of times this value came in the data table is



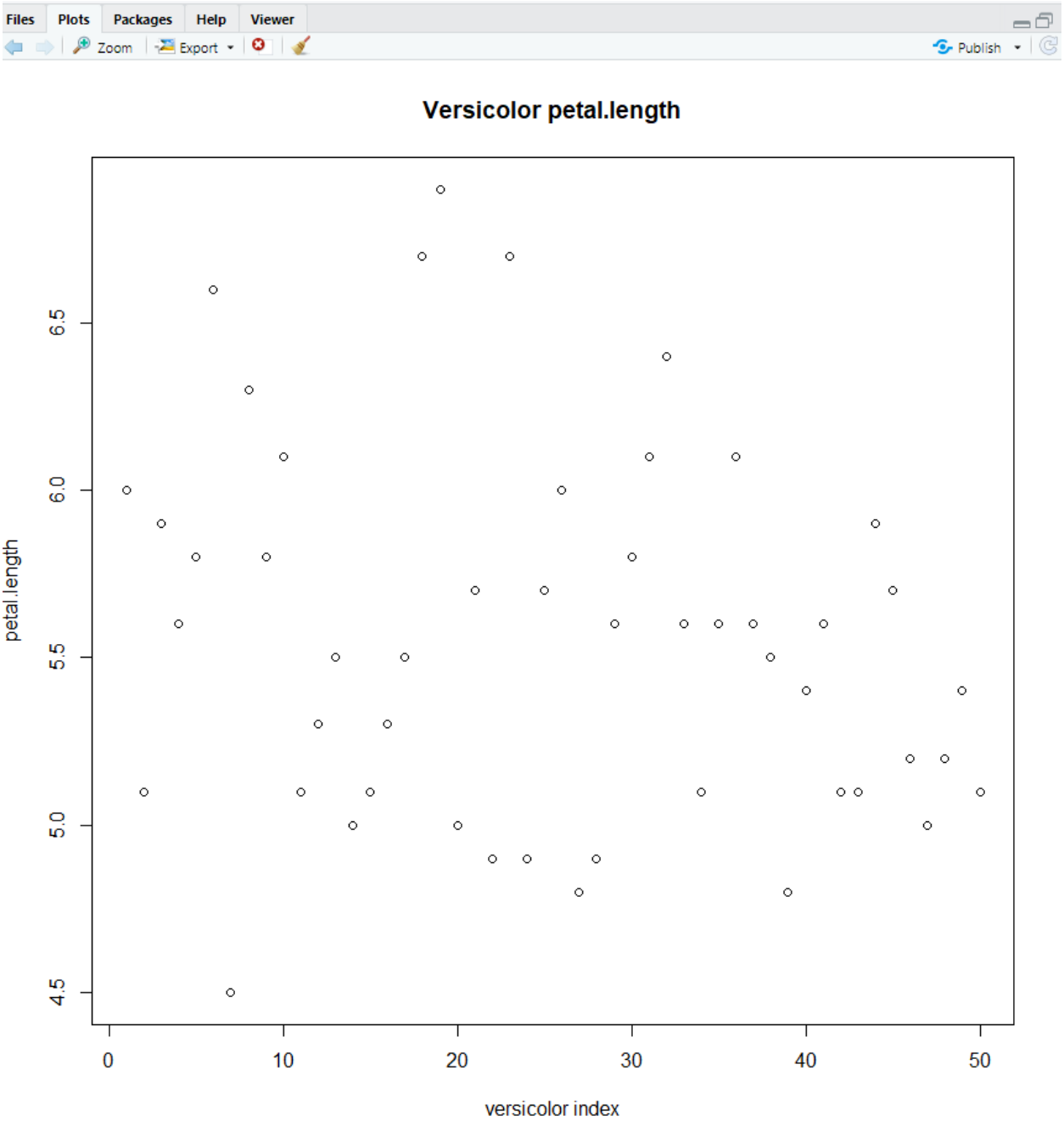
2) Boxplot

This a boxplot. The information that can be extracted from this plot is that this plot is for the sepal.width values of species Setosa. According to this, all the values lie between 3.2 to approximately 3.7



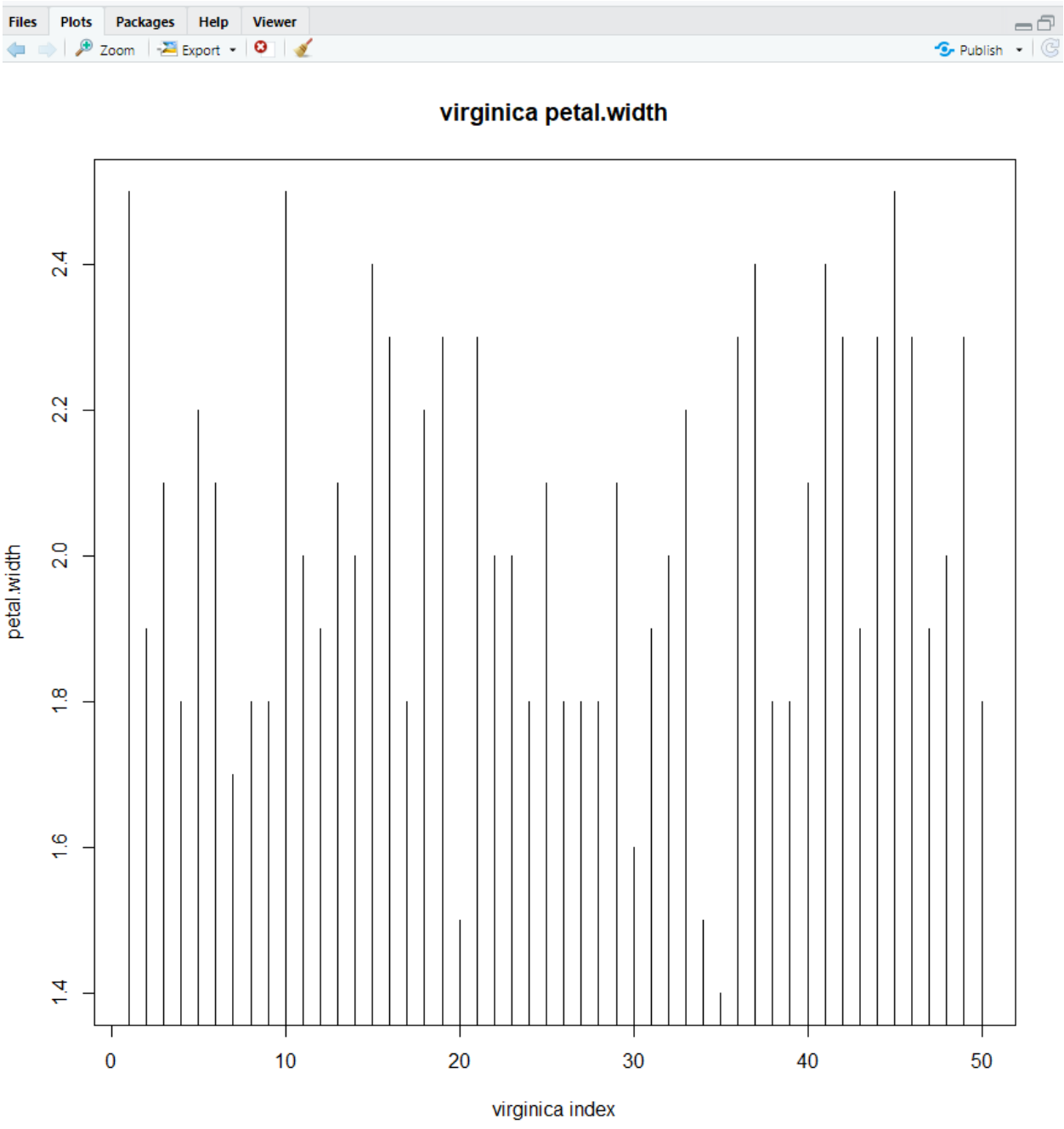
3) Univariate Scatter Plot

This univariate scatter plot shows that the data for the petal.length values for versicolor species. The petal.length values are on the y-axis and the index values(frequency) are on the x-axis. So, for example, the value 4.5 occurs almost 8 times in the petal.length data for versicolor species



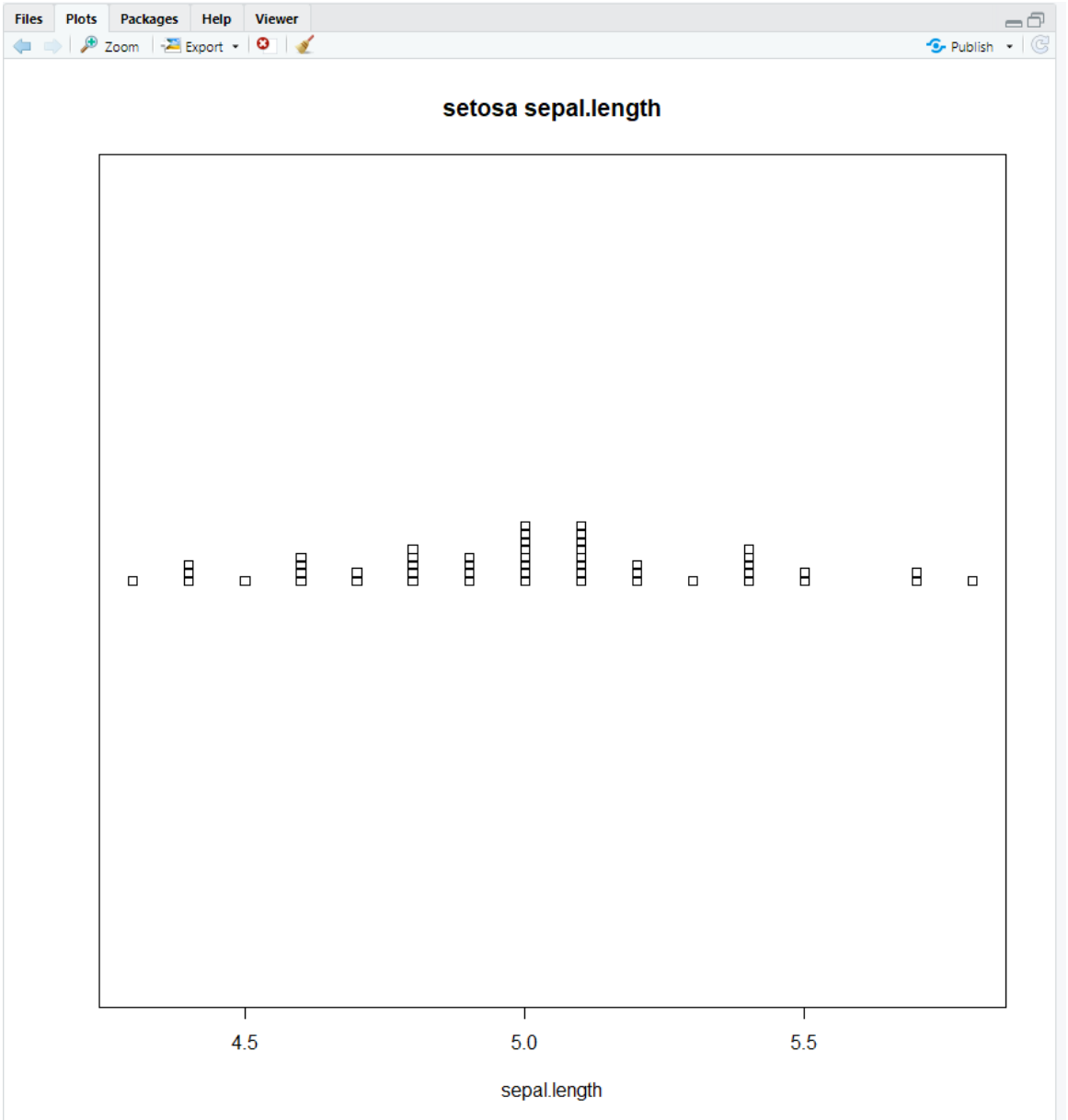
4) Y zero high density plot

The data that we can extract from this is the data values for the petal.width for virginica species. For instance, the value 2.2 petal.width has an index of 5. It is quite easy to see the range and values via this graph



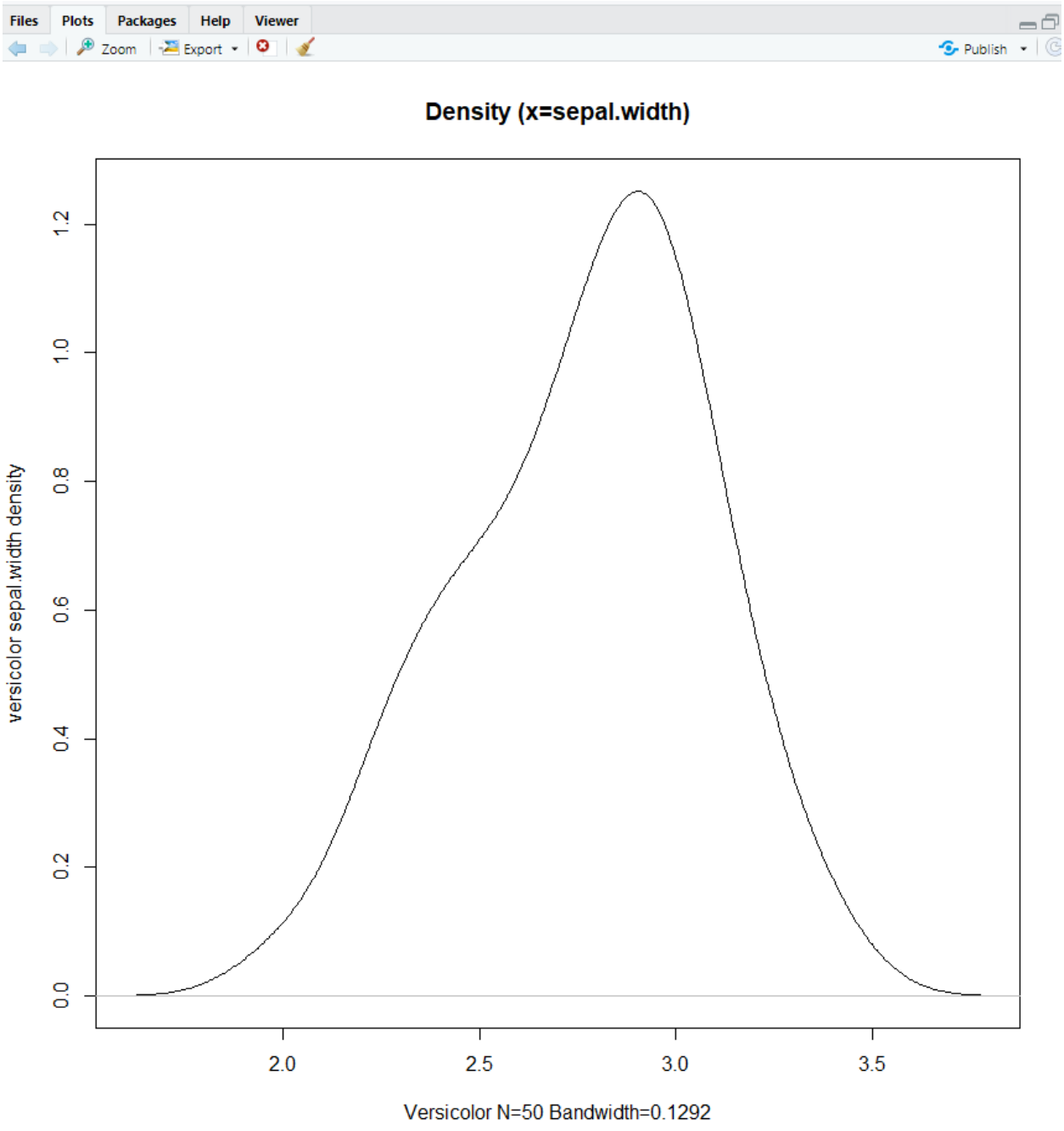
5) Stripchart

The strip chart is another univariate chart and clearly very different than the other ones seen so far. It has a very different way of depicting values. There are no values on the y-axis as we can see and the values on x-axis are the sepal.length values. The data that we can extract is actually the difference in frequency for various values, rather than the quantitative values for frequency. For instance, the frequency of 4.3 in sepal.length values of Setosa species is clearly more than 4.5 sepal.length. This type of representation is only useful if we want to depict the difference in frequency rather than numbers



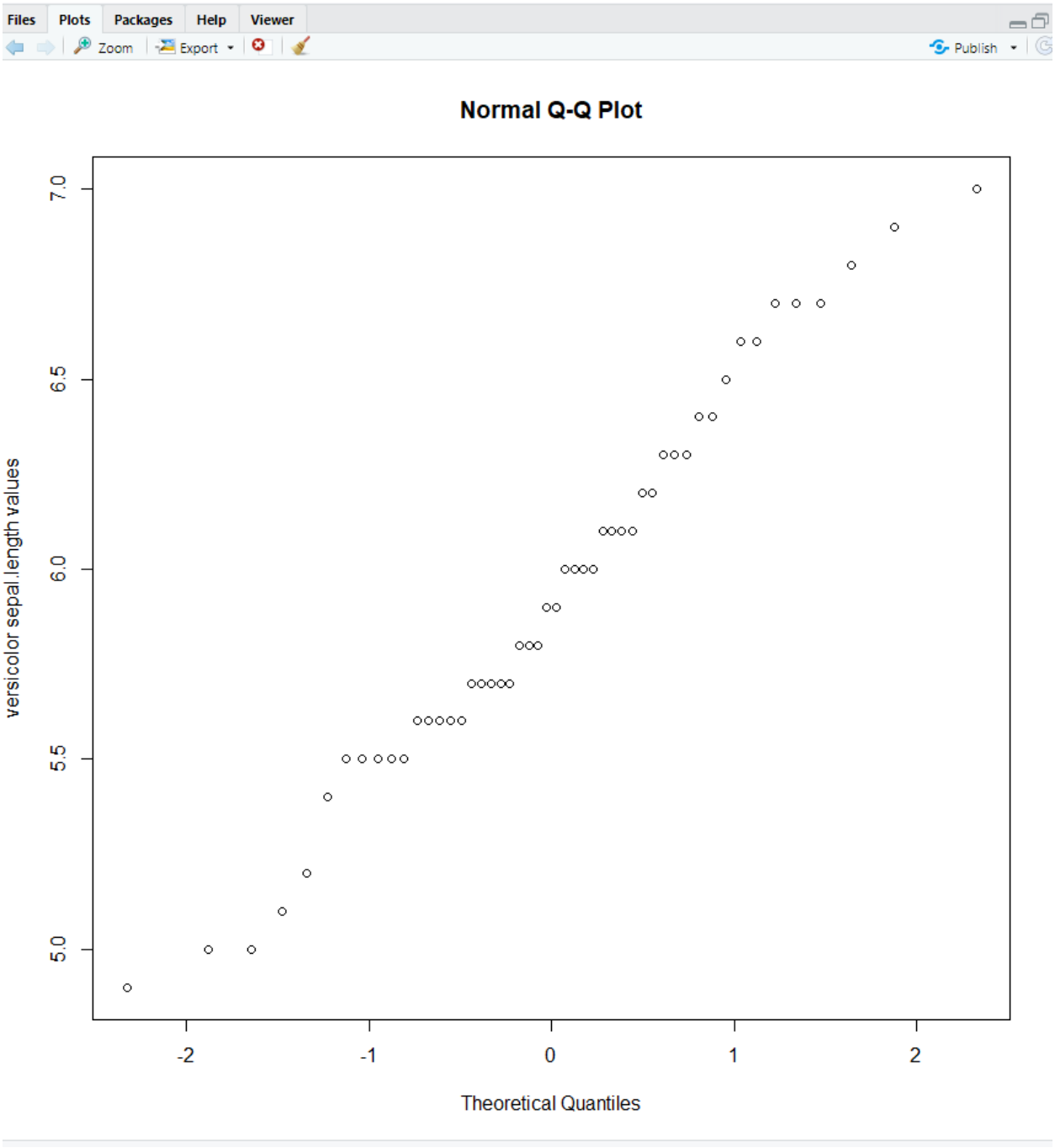
6) **Density plot**

This plot shows how the value density changes with a peak in the center. Information we can extract is that the sepal.width value near 3.0 has the highest density that goes over 1.2. Also, there is a slight change in the straight flow of ascension at sepal.width value's density as it approaches 2.5



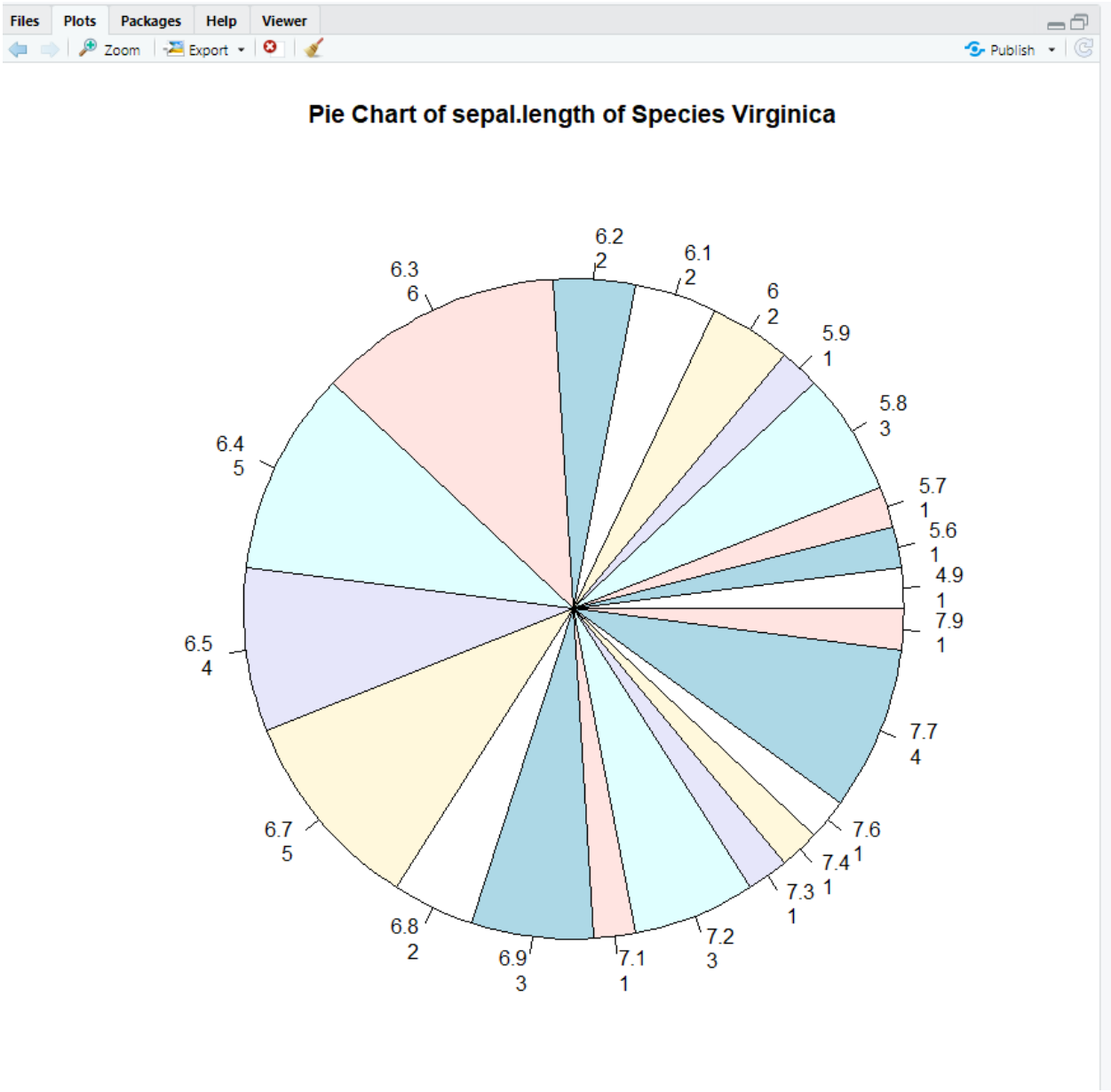
7) Q-Q plot

Normal Q-Q plot helps in determining how well a set of quantitative values distribute over a set of measurements. For example, in here we can see that the sepal.length values for 5.5 kind of overs the quantity values ranging from before -1 to after -1. Similarly we can see this variation for all the values in the versicolor sepal.length values.



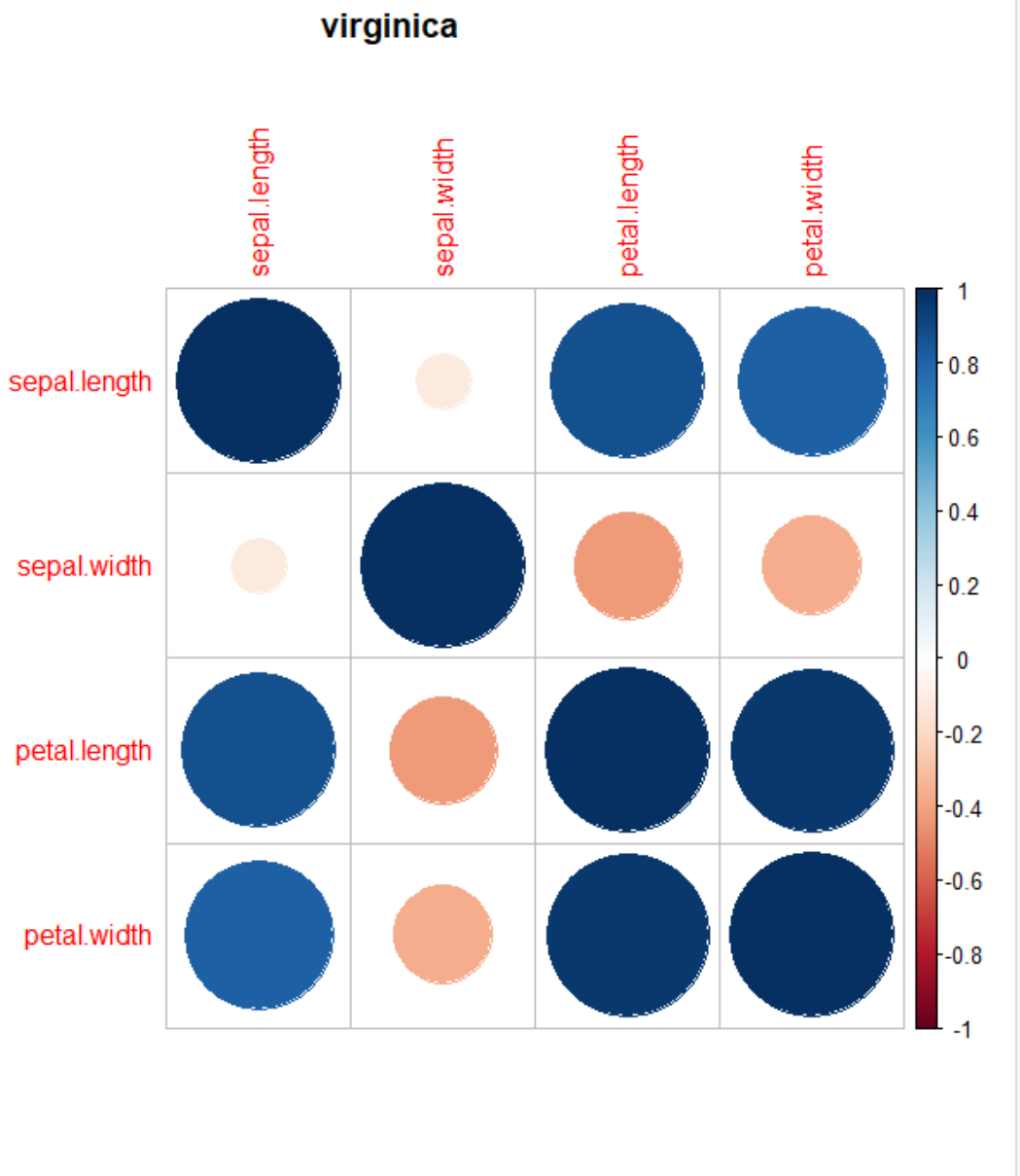
8) Pie Chart

The pie chart is another way to represent univariate values. In here we can extract that the pie is for sepal.length values of the virginica species. We can basically see the distribution based on the frequencies. For example, 6.4 sepal.length has a frequency of 4 in the data. (I did not use a 3-D representation of pie chart because I personally do not like that kind of data representation since the shadows and color and lighting kind of messes up a good representation and may lead to confusion and eye strains which are a hindrance to understanding the data we are trying to represent, especially when there are these many data, like in here).



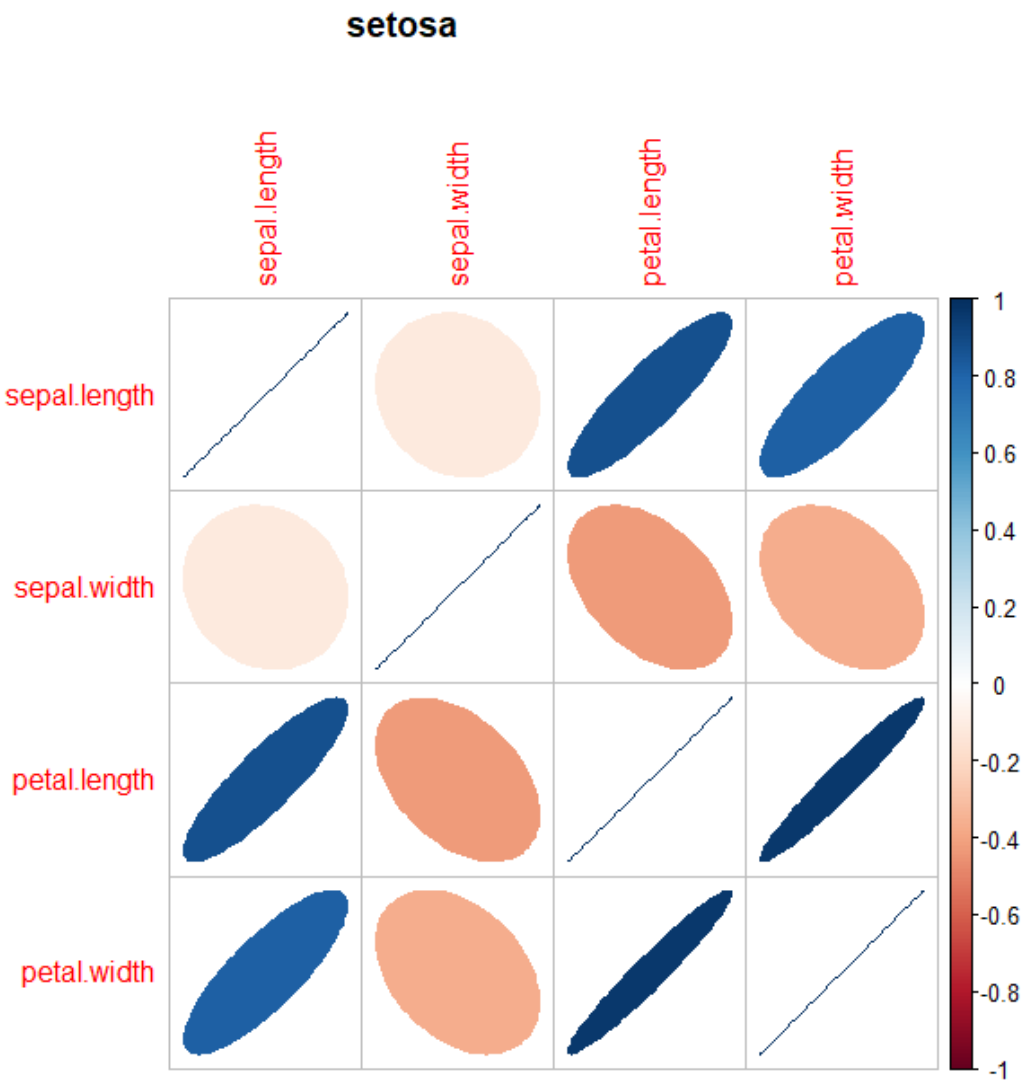
9) Glyph Chart – circles

Glyph charts are useful in representing of data using figures. They show the changing of the values based on pictures or figures sizes or color intensity. (This is technically a corrplot, because glyph library was unavailable for R, however it helps in depicting the values in the same manner and hence can be considered as a glyph approach). So the information we can extract is that the common values for example between virginica - sepal.length and petal.width is much more than between sepal.length and sepal.width.



10) Glyph Chart – ellipse

This is technically the same as above with values for Setosa to depict instead of virginica. The other difference is the representation. So the information we can extract is that the common values for example between sepal.length and petal.width is much more than between sepal.length and sepal.width.



11) Histogram (for different species)

This is a histogram for setosa values of sepal.length. We can extract from this graph the data that for instance the value 4.5 has a frequency of 4.