Project report

Applied Statistics

Dataset : Iris

Table des matières

Introduction

Data set presentation:

This dataset is a list with 5 arguments and 150 rows .

kinds of iris species are represented here with for each one of it their sepal length and width and petal length and width as well. The sample has a total of 150 data on 150 iris with 50 of each species.

It's now up to us to differentiate each species based on each criteria.

I)Data exploration

1)Choose and load the R dataset corresponding to your group subject's and Identify which variables your data set are numeric, and which are categorical (factors) if applicable

The only categorical variables contained in our data set are the species of the iris, every others, the sepal length and width and the petal length and width are numerical variables.

2)Generate summary level descriptive statistics: Show the mean, median, 25th and 75th quartiles, min, and max for each of the applicable variables in your data set

```
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa     :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

3) Determine the frequency for each of one of the categorical variables.

There are 50 specimens of each species of iris, which means that the frequency of each one of the categorical variables is of 1/3.

4)Determine the frequency for each of the one of the categorical variables, by a different categorical variable.

II)Graphic data representation

a) Univariate study

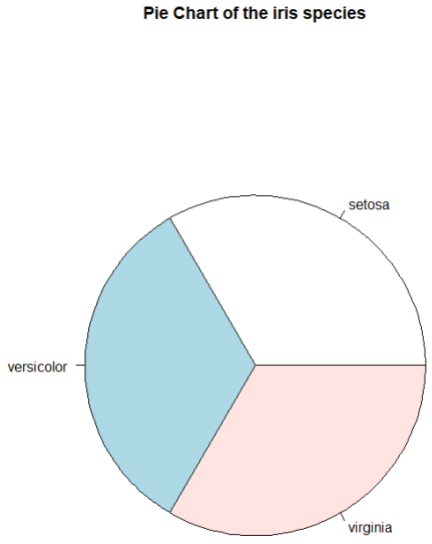
5)Use the commands pie() barplot and dotchart() to represent the categorical data. Comment

We did the barplot and dotchart for the categorical data but it was not that relevant for our study.

Here are the result for the categorical variable:

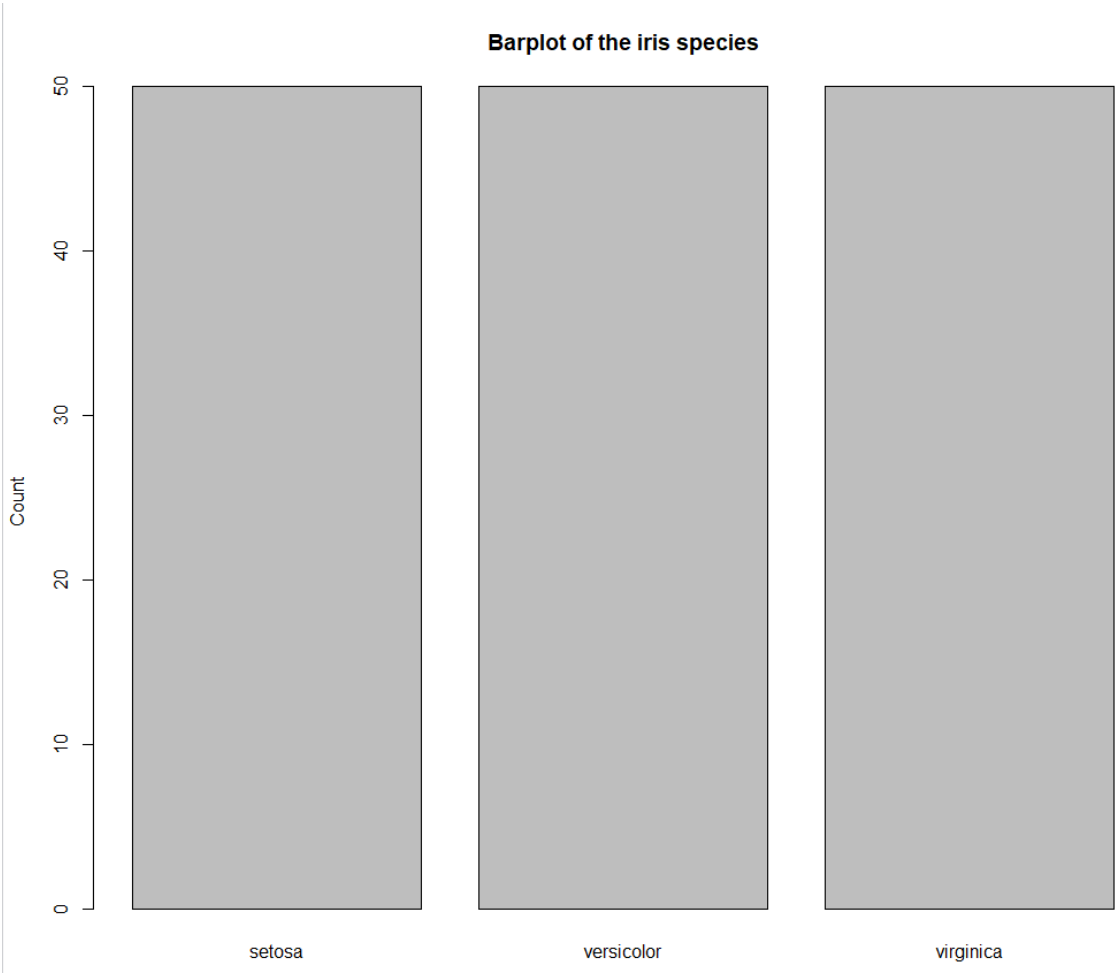
The pie chart :

```
x <- table(iris$Species)
lbls <- c('setosa','versicolor','virginia')
pie(x, labels = lbls,
   main="Pie Chart of the iris species\n (with sample sizes)")
```



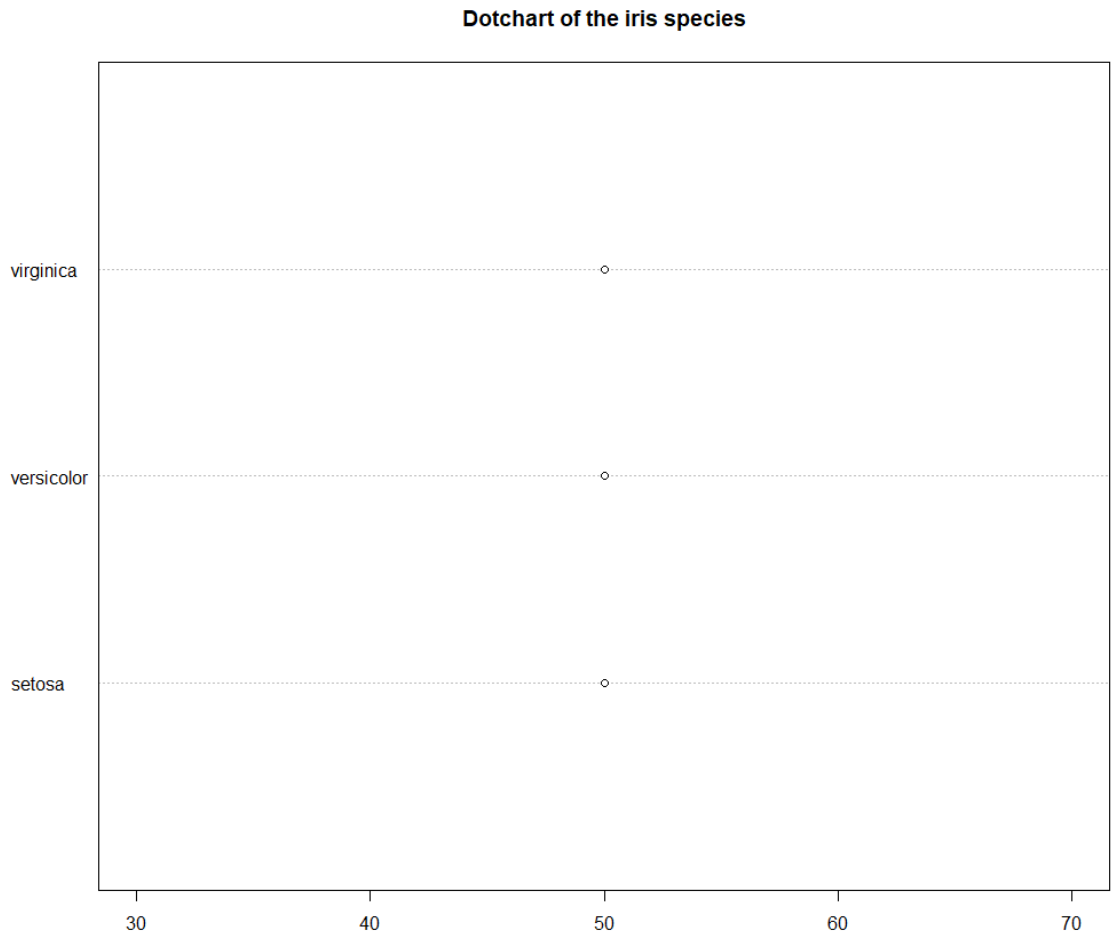
Barplot for the categorical variables:

```
x <- table(iris$Species)
lbls <- c('setosa','versicolor','virginia')
barplot(x, xlab=lbls,
        ylab="Count",
   main="Barplot of the iris species")
```



Dotchart for the categorical variables:

```
dotchart(table(iris$Species), main="Dotchart of the iris species")
```



So we also decided to apply it on the numerical data to have a better understanding of our dataset and figure out the differences between each species.

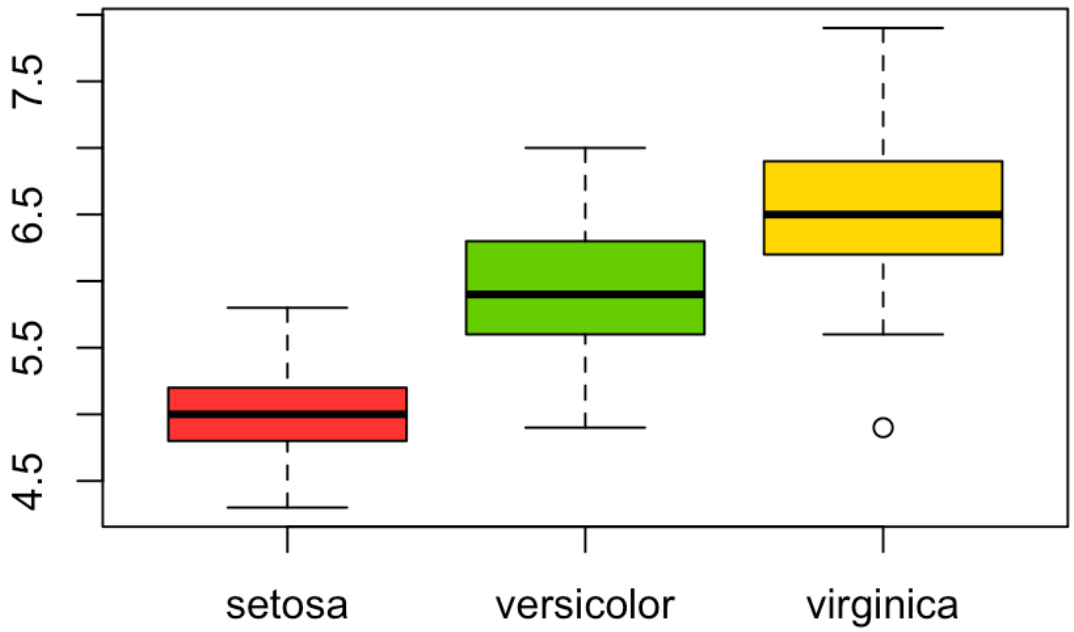
The sepal size study:

barplot():

x=iris$Species

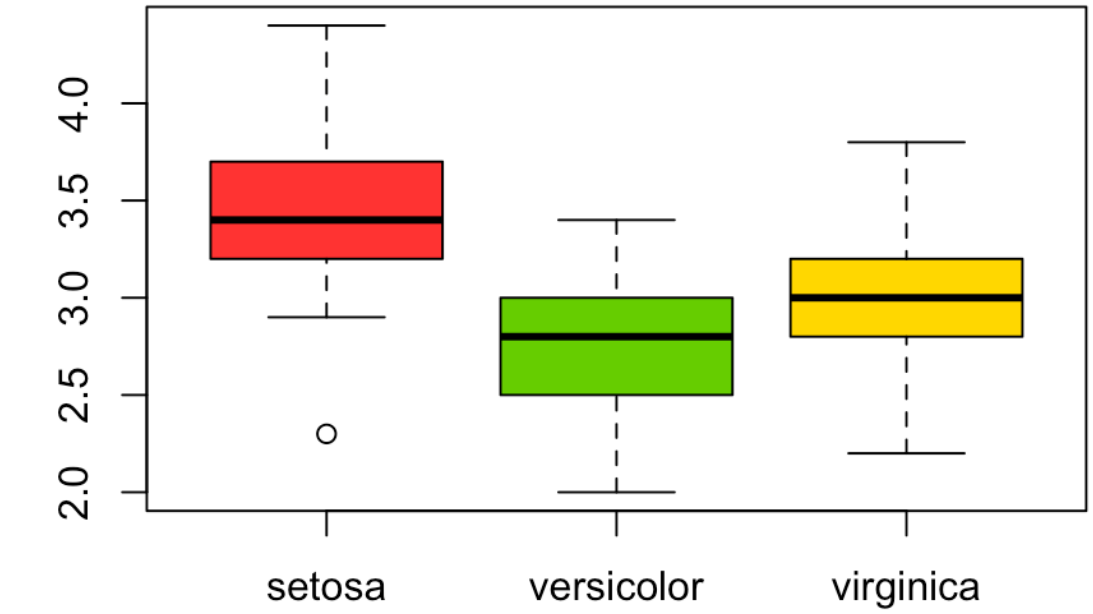
y=iris$Sepal.Length

plot(x,y,col=c("#FF3333","chartreuse3","gold"))



x=iris$Species

y=iris$Sepal.Width



Sepal study conclusion:

- Setosa is characterized by sepals fairly short (with a first and a third quartile of 4.4 and 5.9) but especially wide (with a first and a third quartile 2.9 and 4.4)

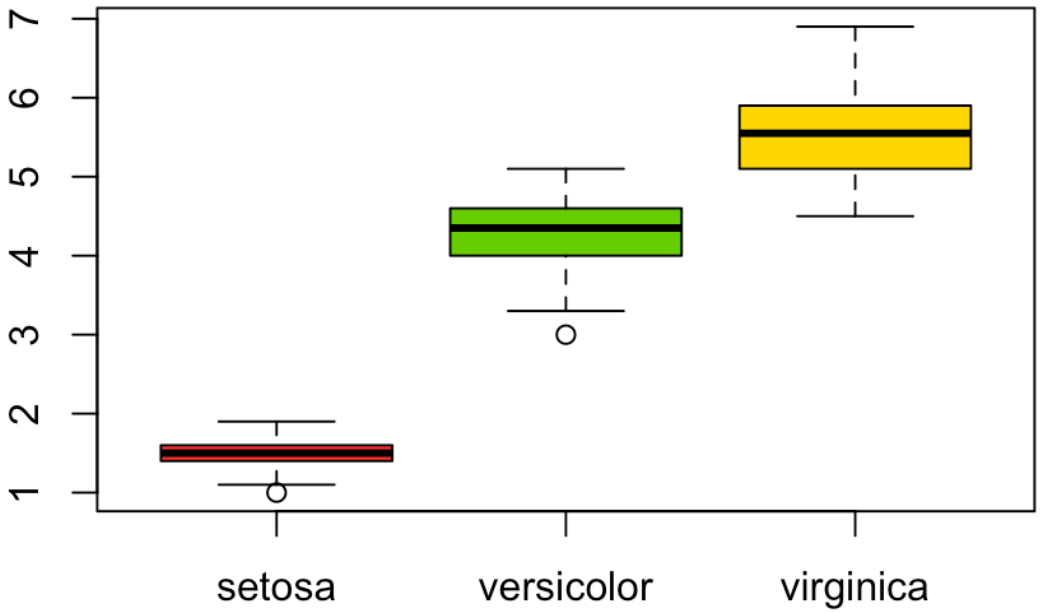
- Versicolor is characterized by a middle sized sepal (with a first and a third of quartile 4.4 and 6.5) and the less wide sepals (with a first and a third quartile of 2.1 and 3.4)

- Virginica is characterized by the longest sepals (with a first and a third quartile 5.7 and 7.9) and a low average width (with a first and a third quartile of 2.3 and 3.7)

The petal study:

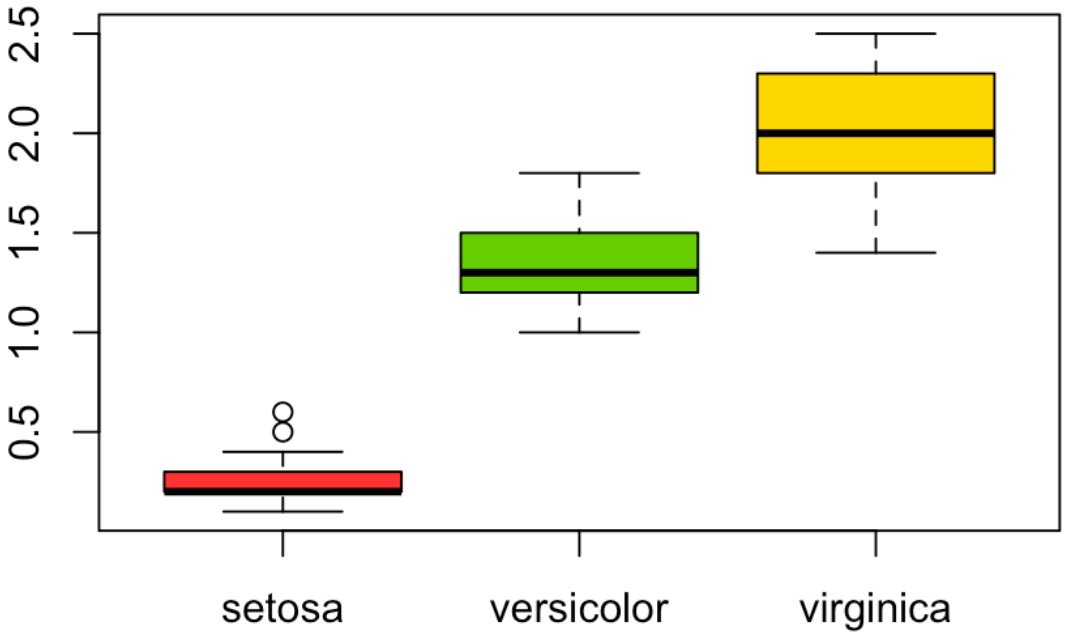
x=iris$Species

y=iris$Petal.Length



x=iris$Species

y=iris$Petal.Width

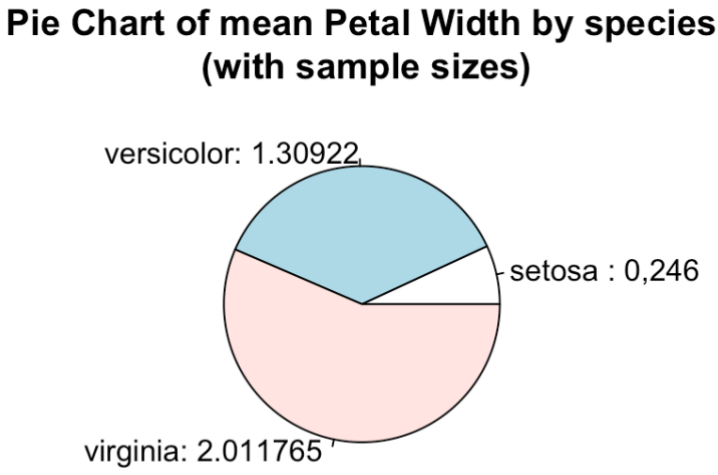


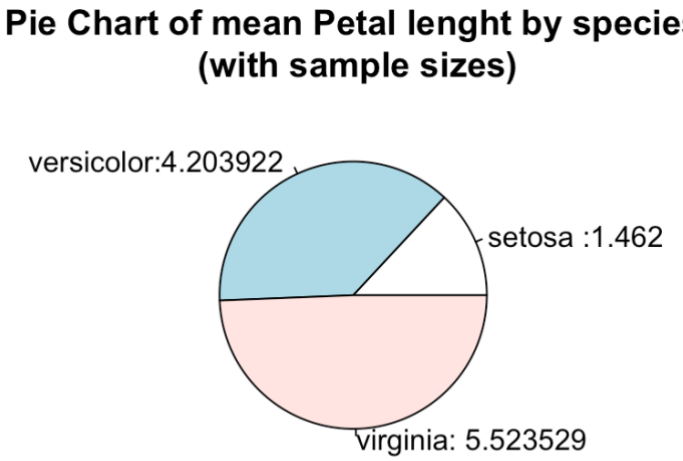
The petal study conclusion:

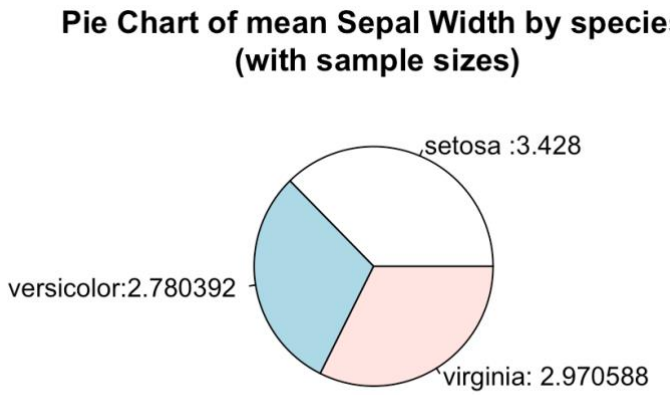
- Setosa has the shortest ad thinnest petals, the values are very located, between 1.1 and 1.9 for the length and 0.1 and 0.4 for the width (while omitting the 2 outliers)
- Versicolor as an average length and width, for the length between 3.5 and 5.1, for the width between 1.0 and 1.8.
- Virginica has especially long and wide petals, the size in length is quite consistent, between 4.6 and 6.9 however for the width the gap between the wider and the less wide is more important, between 1.4 and 2.5.

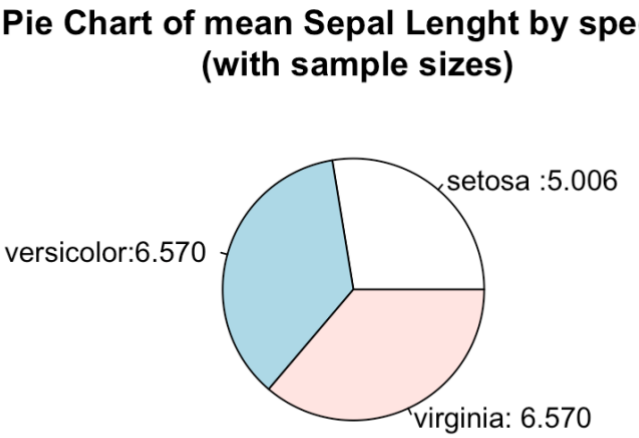
The Pie study :

As we said this representation is not really relevant for our problem because each species is equally represented so a pie representation of the species would not help the study that much. Represent each sepal or petal data while not help neither because there is too much data so I decide de represent the mean of each columns by species to see the for each species.




Pie Chart of mean Petal lenght by species
(with sample sizes)
versicolor:4.203922
setosa :1.462
virginia: 5.523529


Pie Chart of mean Sepal Width by species
(with sample sizes)
setosa :3.428
versicolor:2.780392
virginia: 2.970588

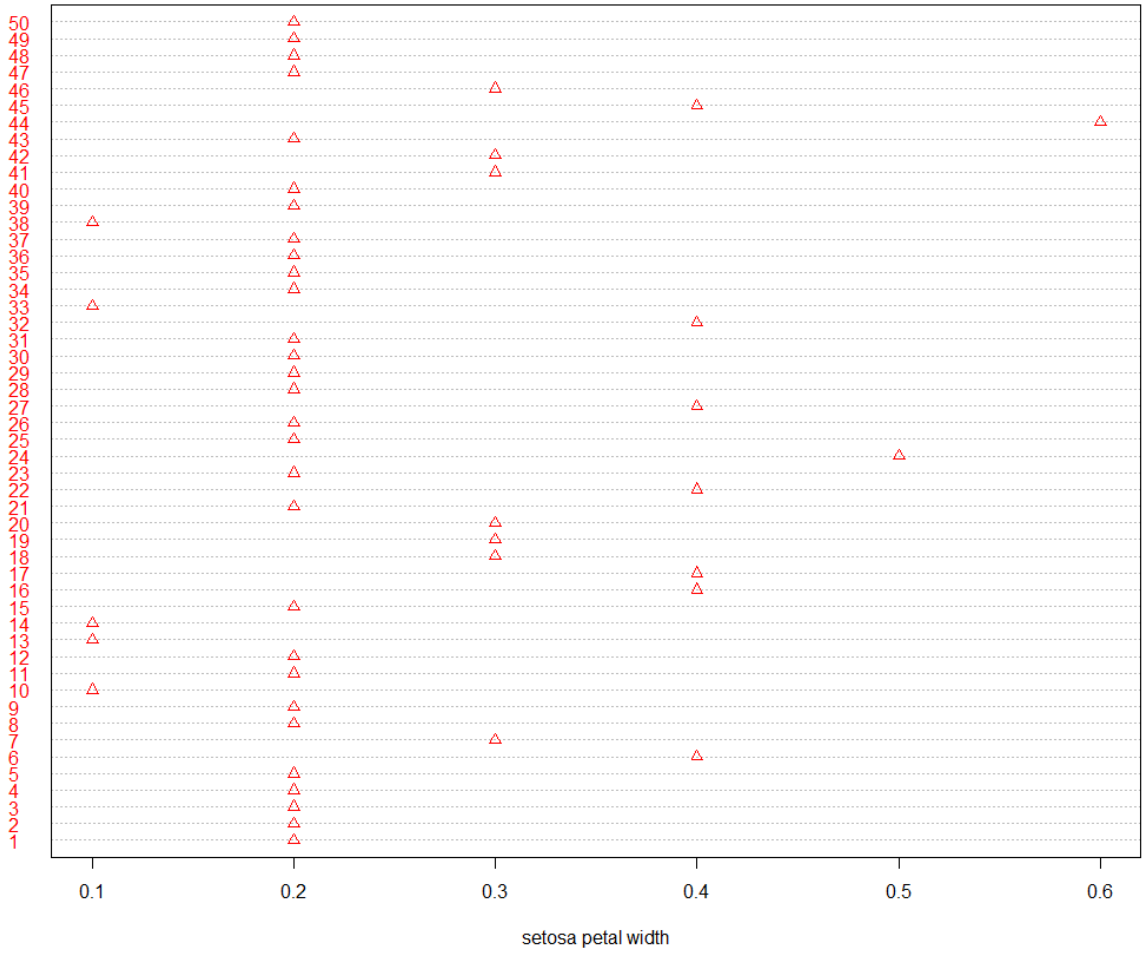


The dotchart study

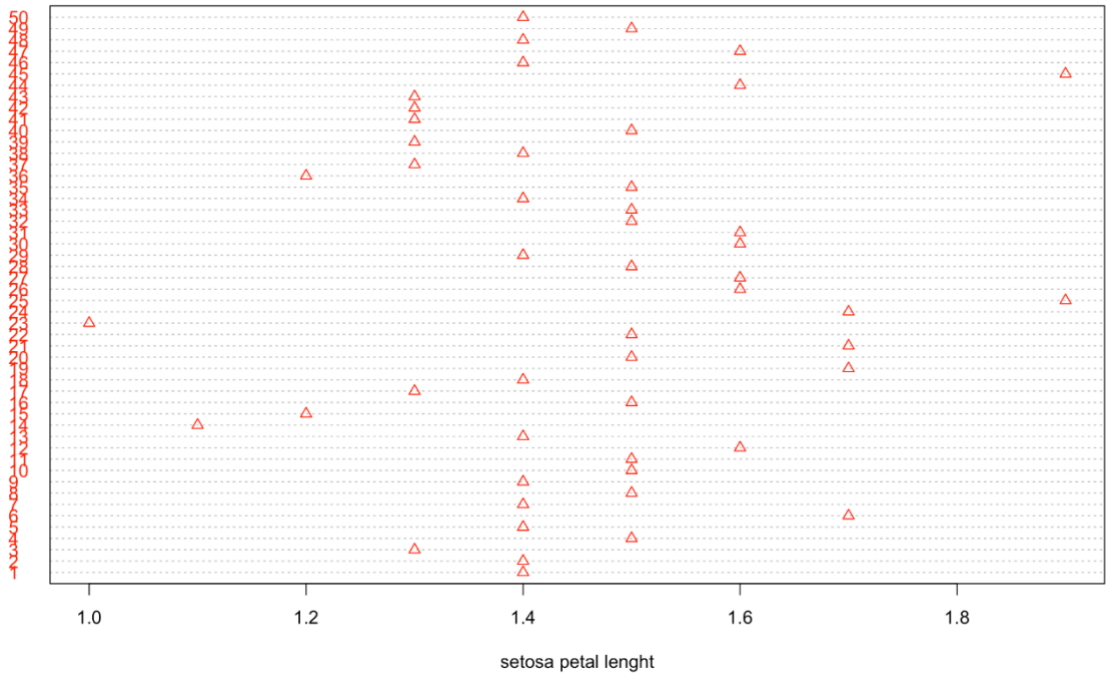
Representation of labels by species, first **setosa in red.**

Here is one example of code :

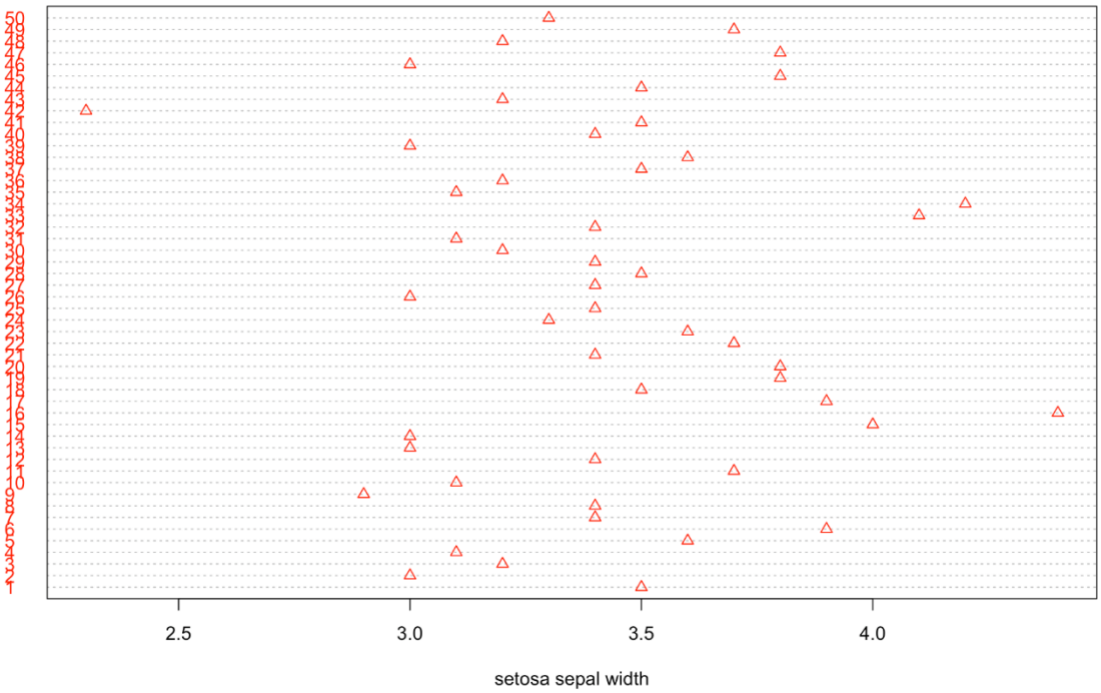
```
setosa_PW<-iris$Petal.Width[0:50]

dotchart(setosa_PW,labels=row.names(iris),col="red",pch=2,xlab="setosa petal width")

setosa_PW
```



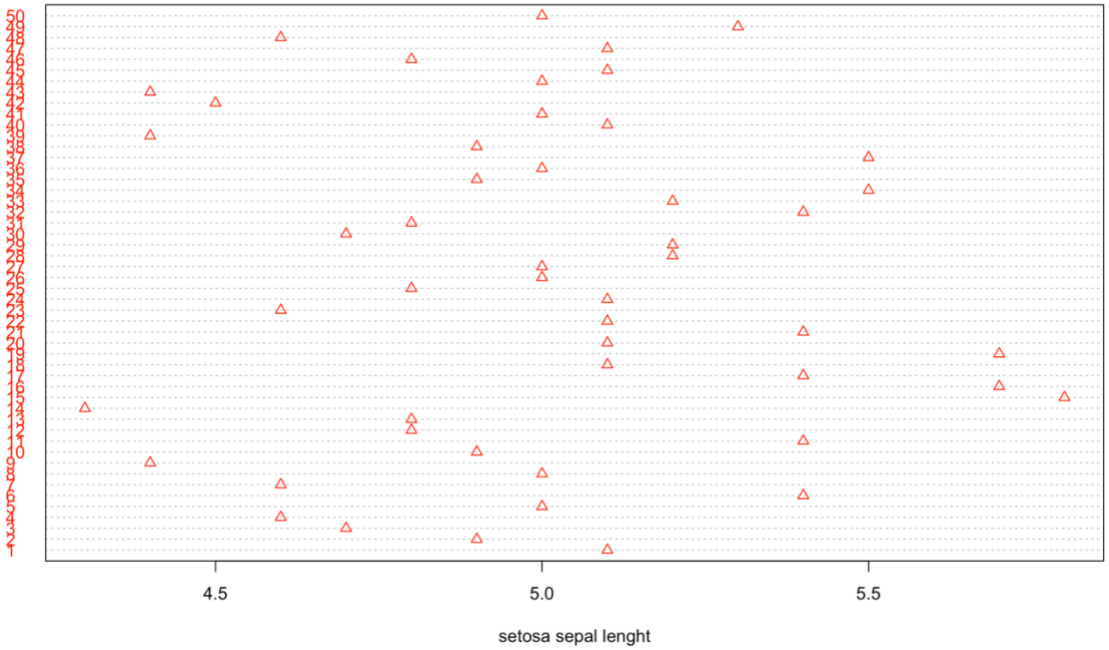
We can see the repartition of each values of setosa's petal weight in the dataset, it can be helpful to see the outlier values like here the value 0.6 which seems not to be representatives of the sample. We can also see the value the most present here is 0,2.



In the same way we can spot two outlier values, 1.0 and 1.9, the most represent values seems to be 1.4. Values seems to be mostly between 1.2 and 1.7.

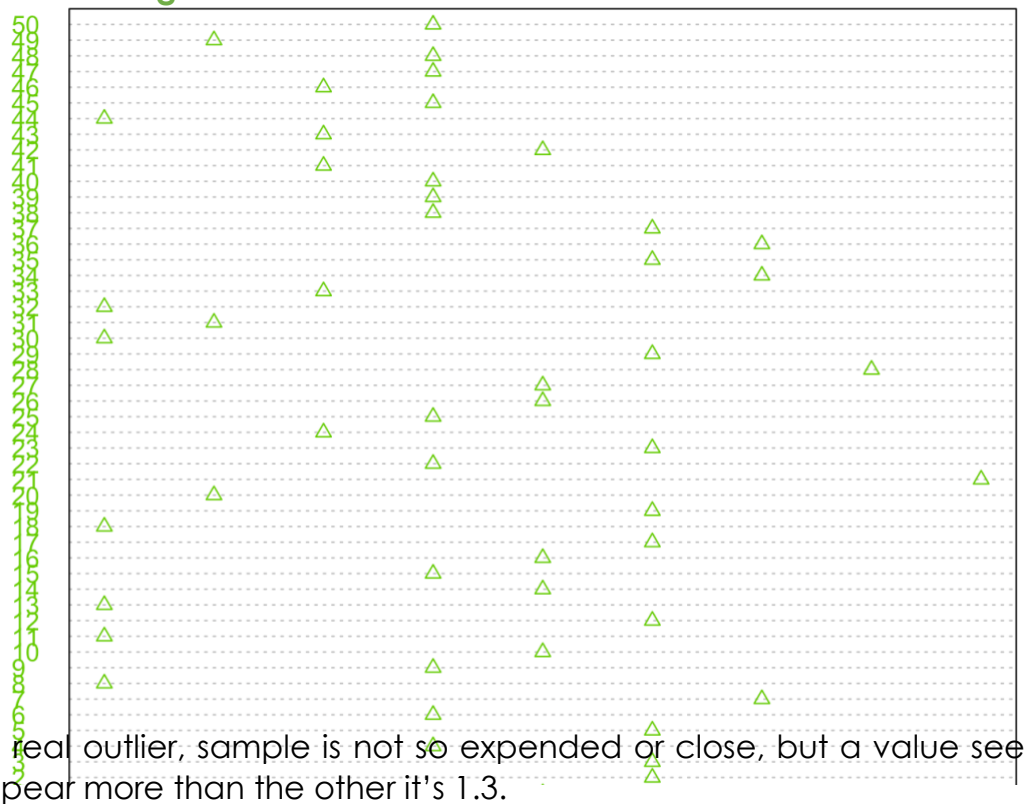


Two values seem to be outliers, 2.3 and 4.4 the rest of the values are close between 3.0 and 4.0 but very different.

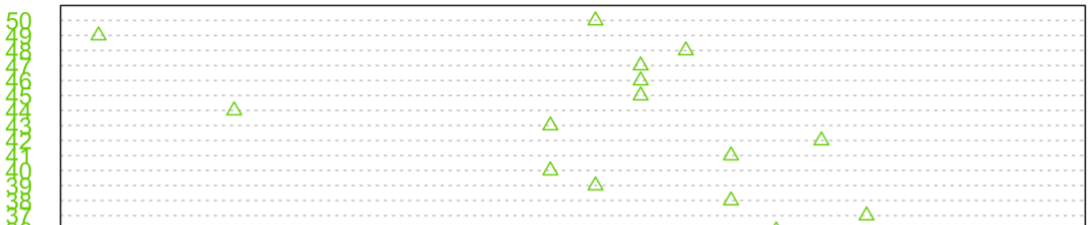


No outliers, the sample seems extended between 4.3 and 5.8. NO values seem to be particularly represented; values are very different.

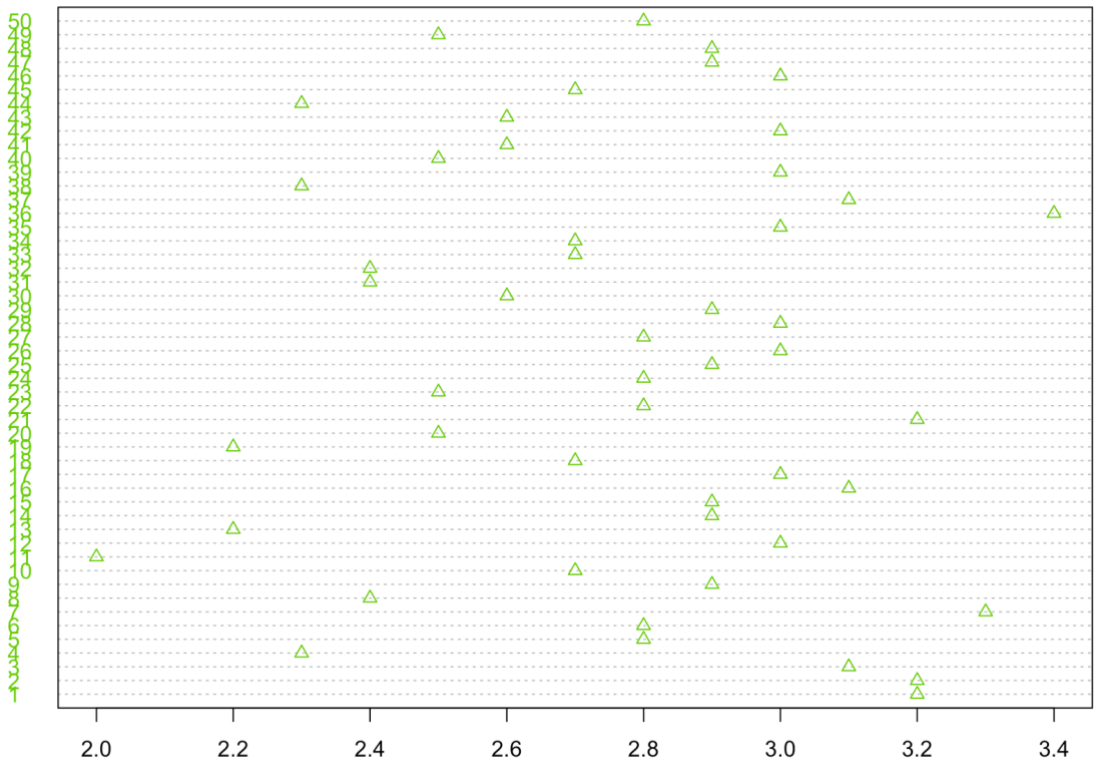
Versicolor in green



No real outlier, sample is not so expended or close, but a value seems to appear more than the other it's 1.3.

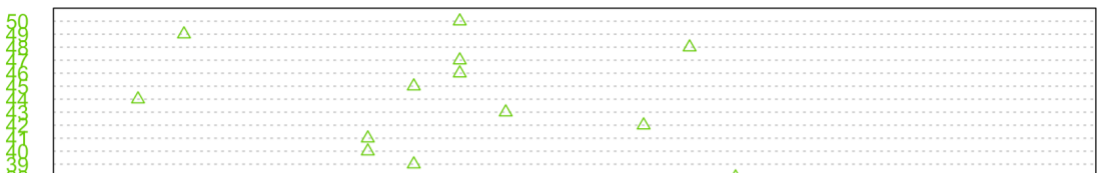


One value is an outlier, the minimum, 3.0. The sample is expended, values are diffuse on the sample.



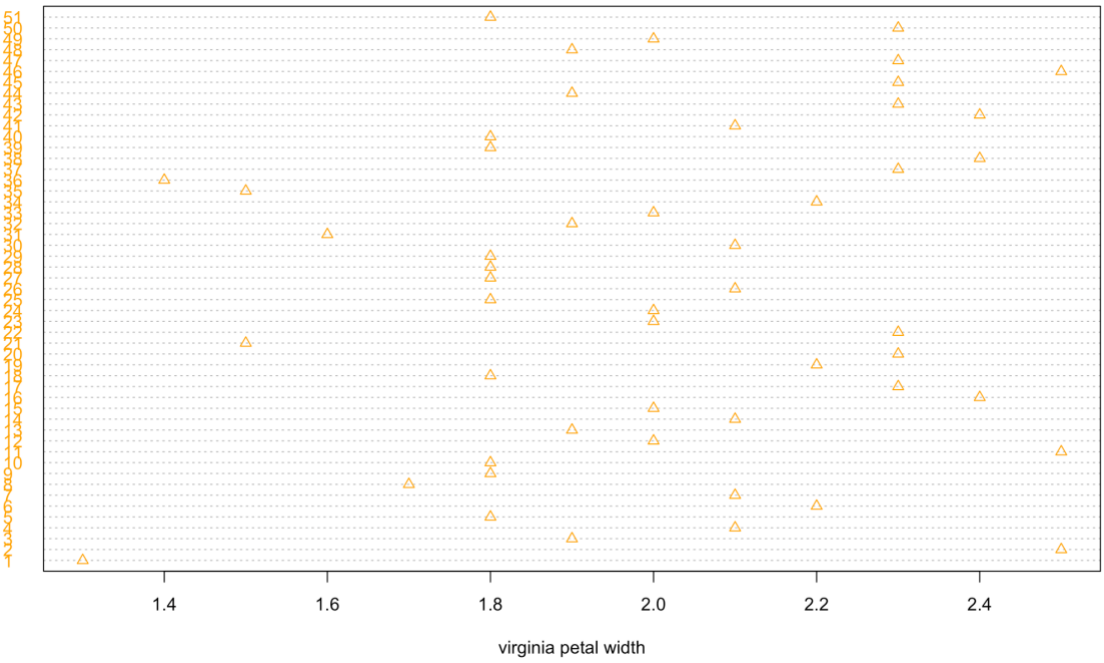
No outliers, sample spread between 2.0 and 3.4, values are diffuse.

versicolor sepal width

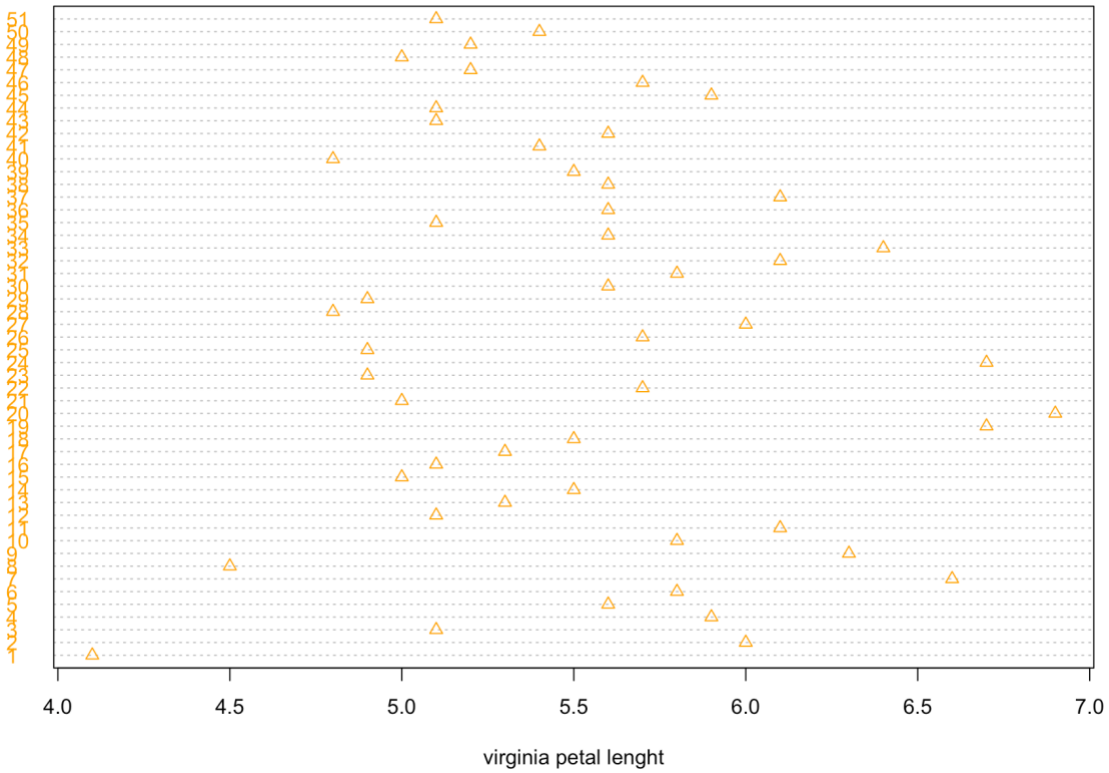


No outliers, sample spread between 5.0 and 7.0, values are diffuse, sample is large.

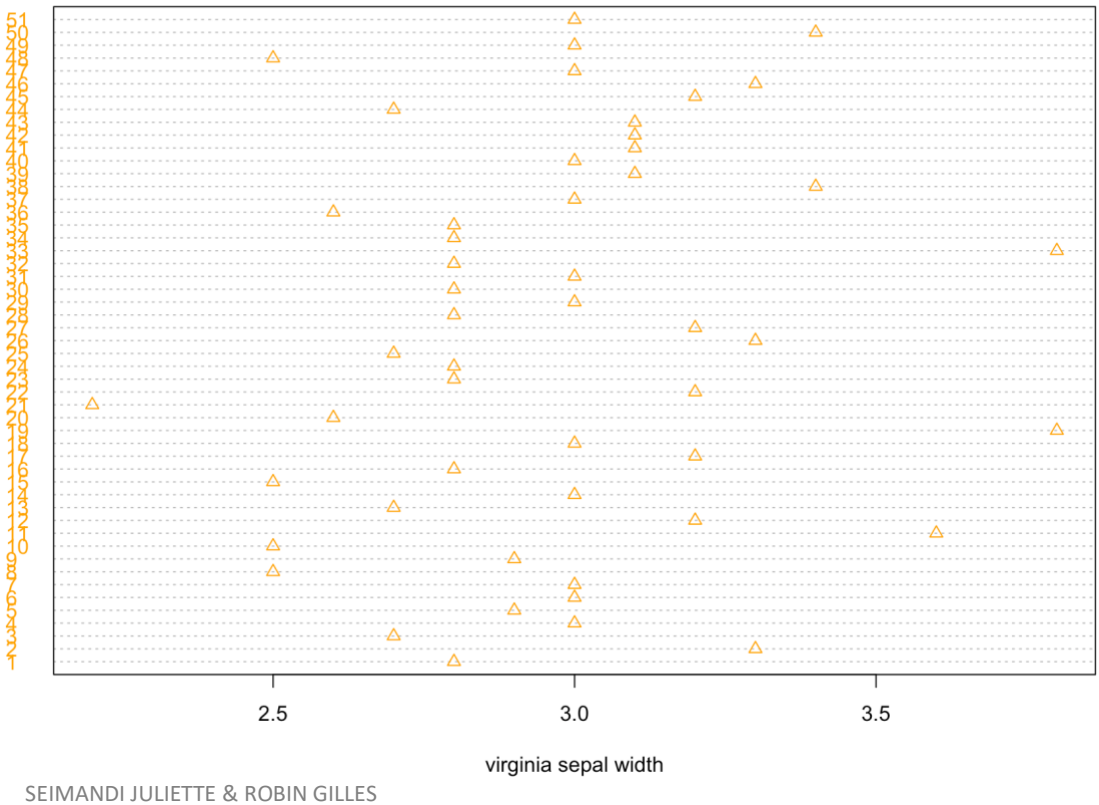
Virginica in orange:



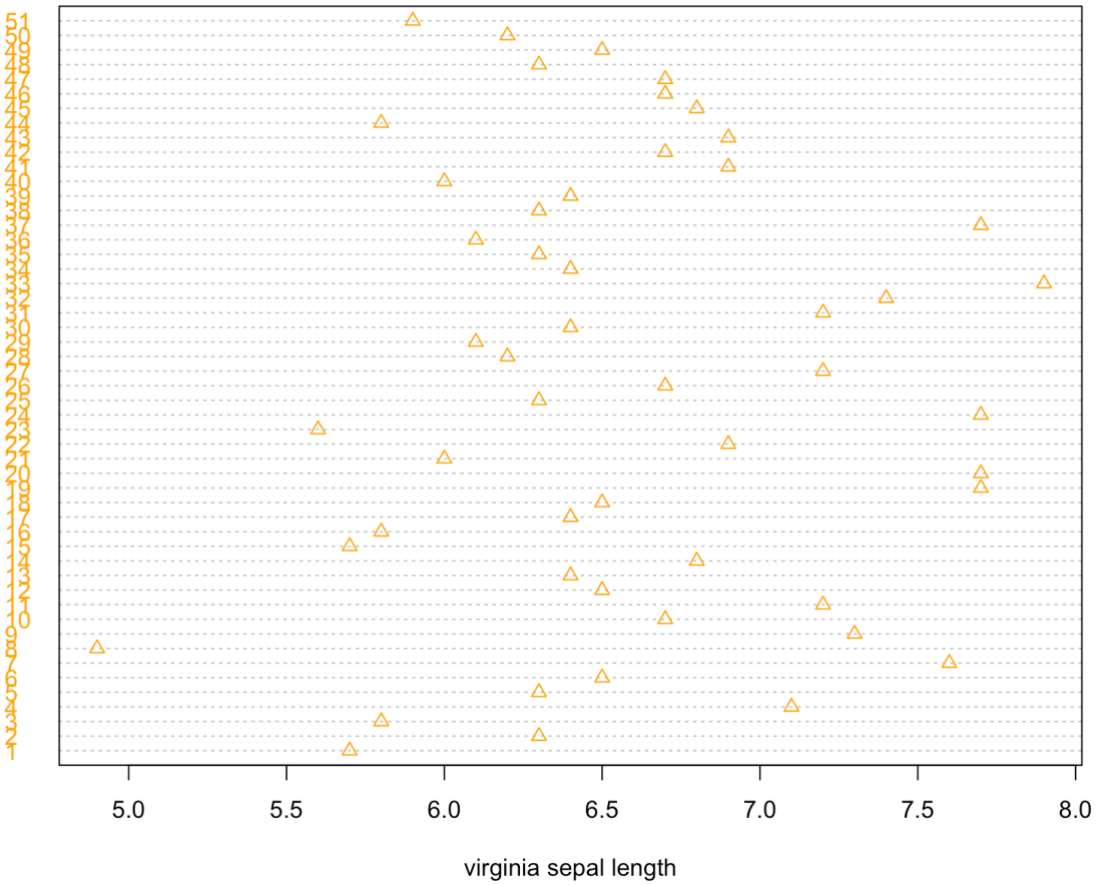
No outliers, sample spread between 1.3 and 2.5, values are diffuse. 1.8 seams to be the most present values.



4.0 can be an outlier, sample spread between 4.0 and 6.9, values are diffuse, sample is large.



No outliers, sample spread between 2.2 and 3.8, values are diffuse. 3.0 seems to be the most present values. Values seems to be focused between 2.5 and 3.3.

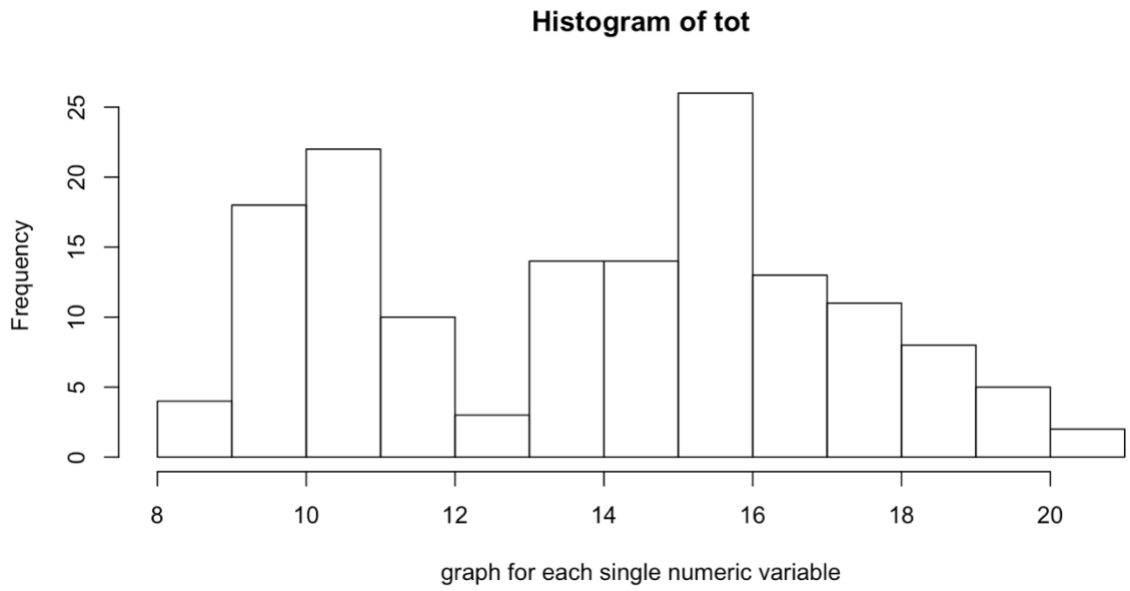


One value is an outlier, the minimum, 4.9 which is far of the rest of the sample. The sample is expended between 5.6 and 7.9, values are diffuse on the sample.

To conclude, dot chart is interesting to see the reparation of all values of a sample by labels by species, spot outliers or the minimum and the maximum and the repetition frequency. And with these data make a conclusion about the most common size or weight of sepal/petal by species.

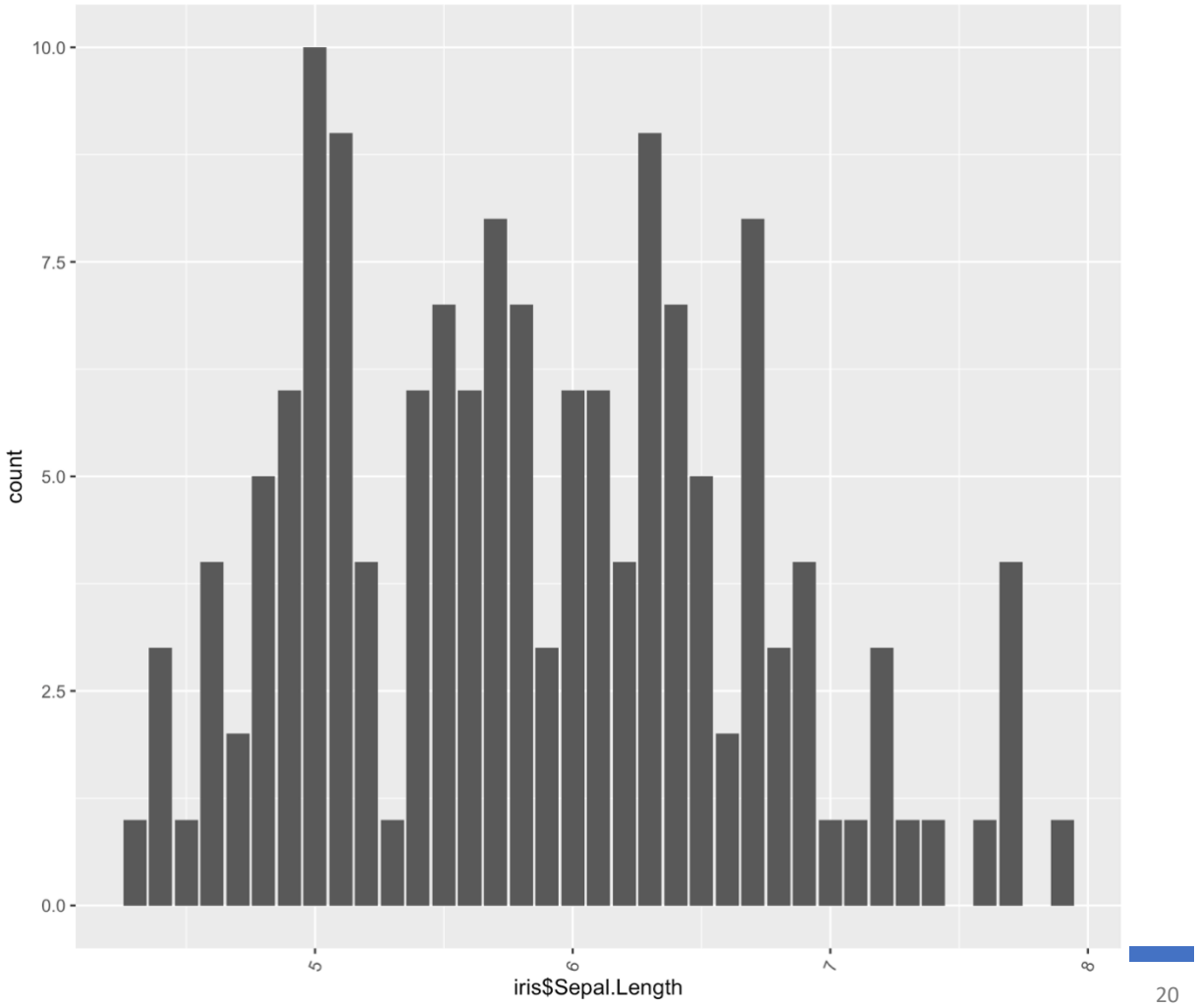
Create a graph for each single numerical variable

Histogram of all numerical values:

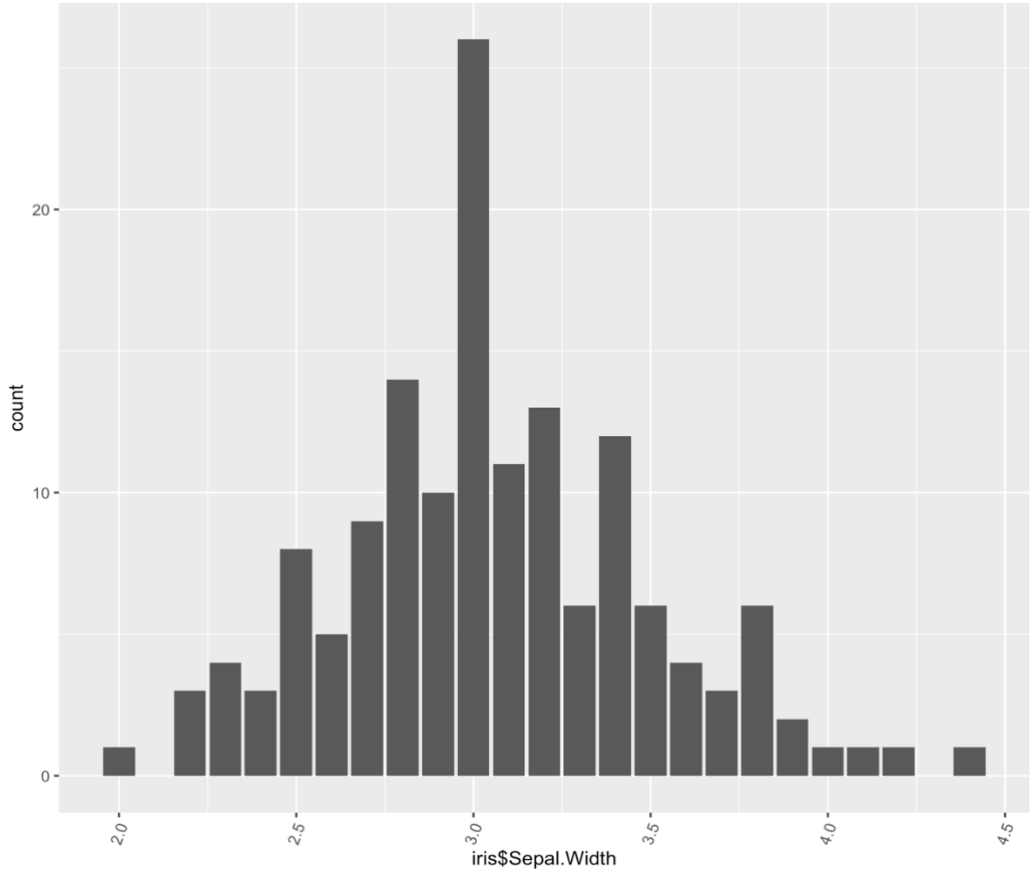


This graph shows all the values present in the dataset but is not relevant to respond to our problem.

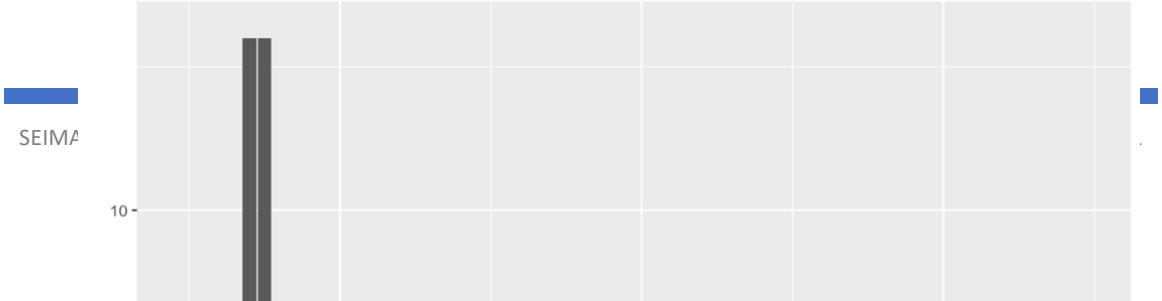
Histogram for each single numeric corresponding to different categorical variable:



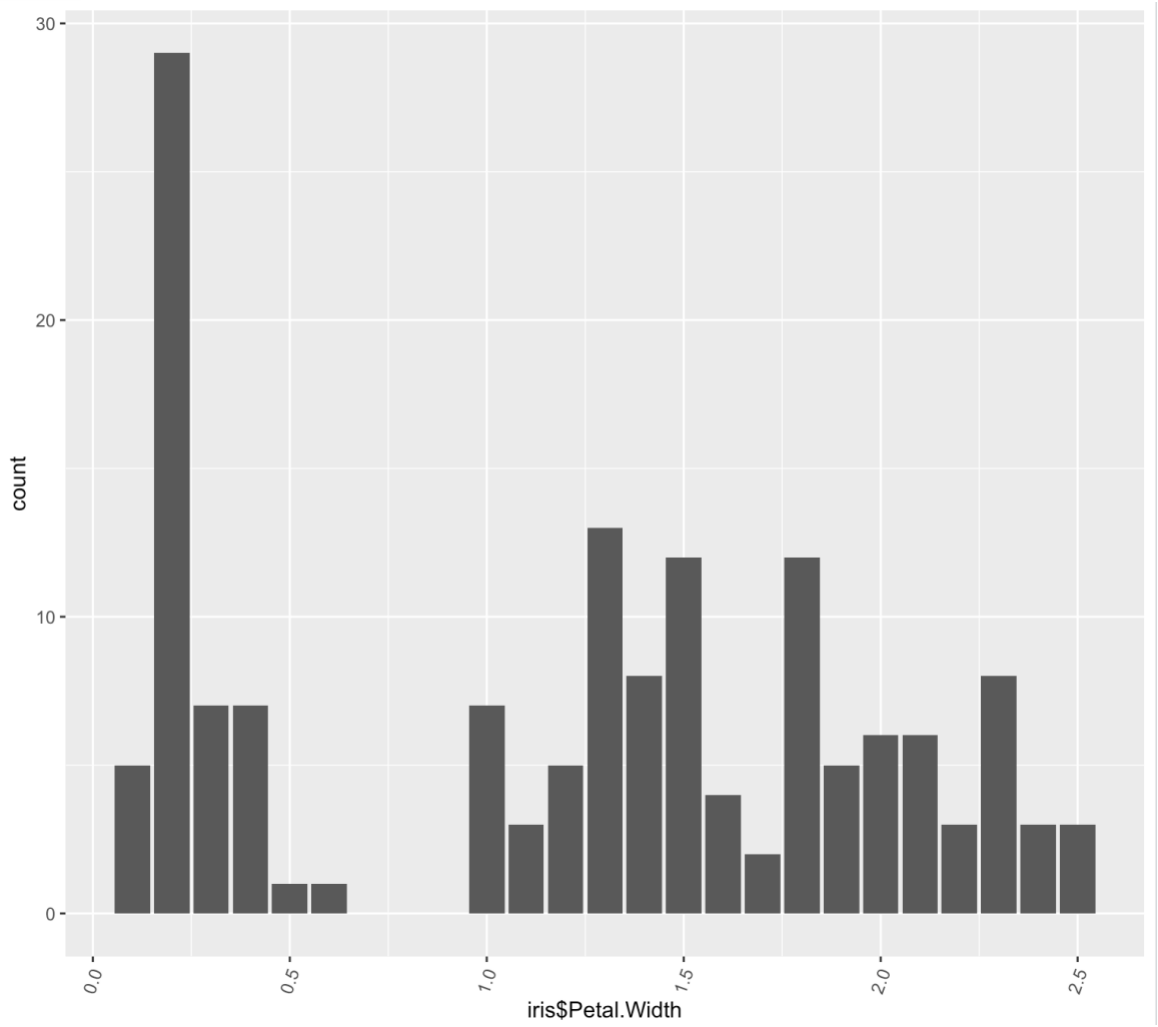
We can see that 5.0 is the most present value and some value are not represent like 7.5 or 7.7.



We can see a pic of values at 2.8/3.0. And recognize what seems to be a Gaussian repartition of the values.



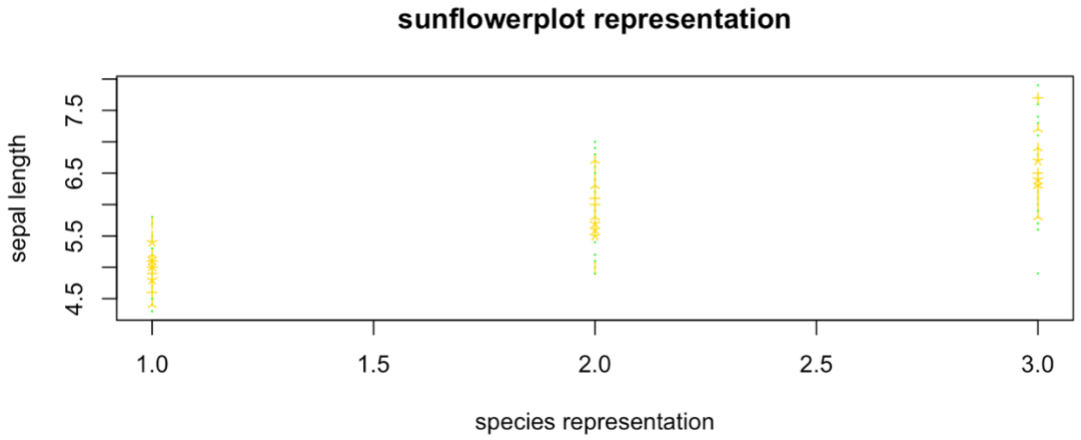
The repartition is non uniform, absence of value between 2 and 2.5.

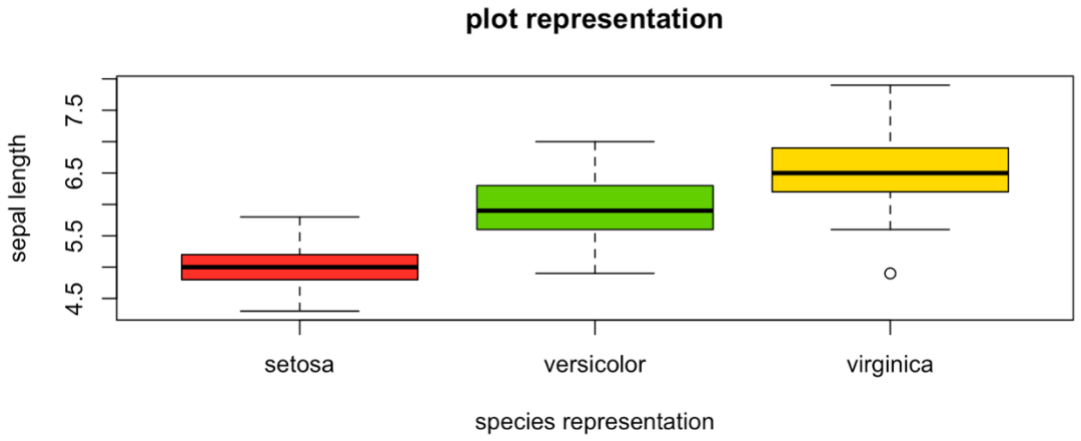


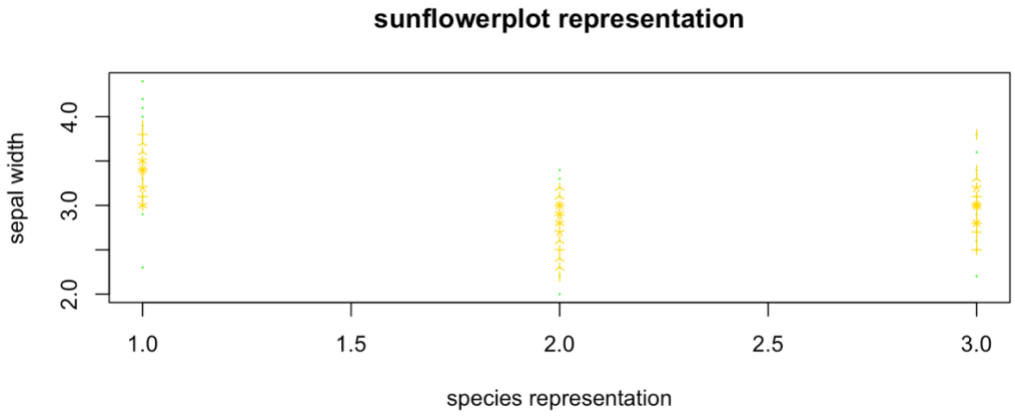
There is an important concentration of values between 0.0 and 0.5, and after a gap most of the value are between 1.0 and 2.7, but the most frequent value is 0.2

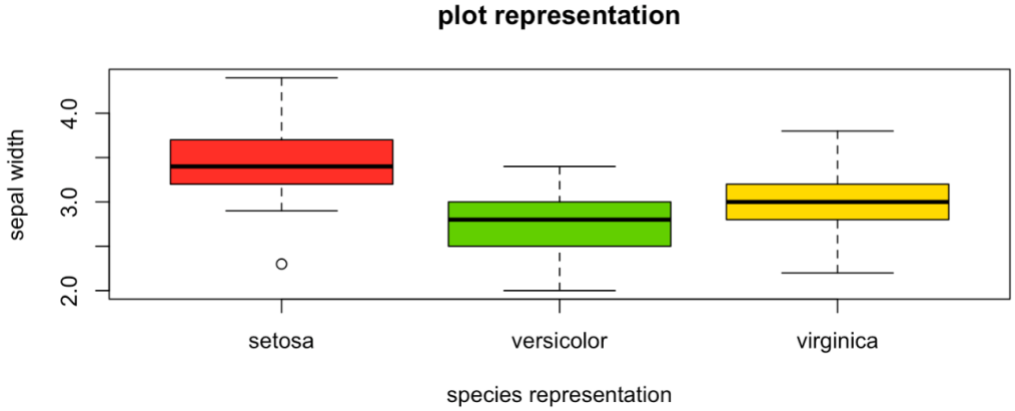
b)Bivariate study

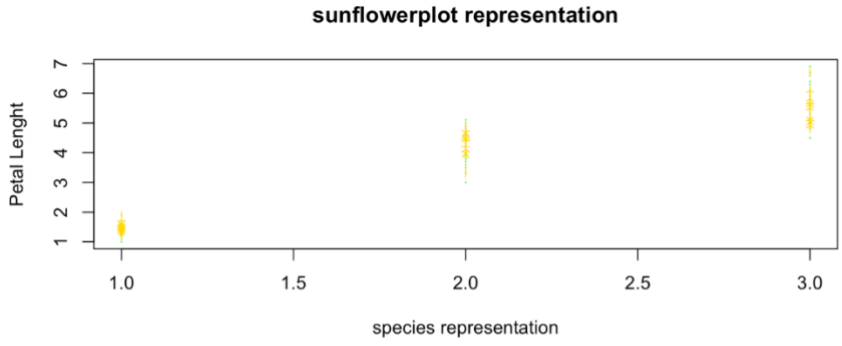
8)Use the command plot or sunflowerplot to plot the scatterplot of the dependent and independant variables. What is the diffrence between these two commands ? Comment your results.

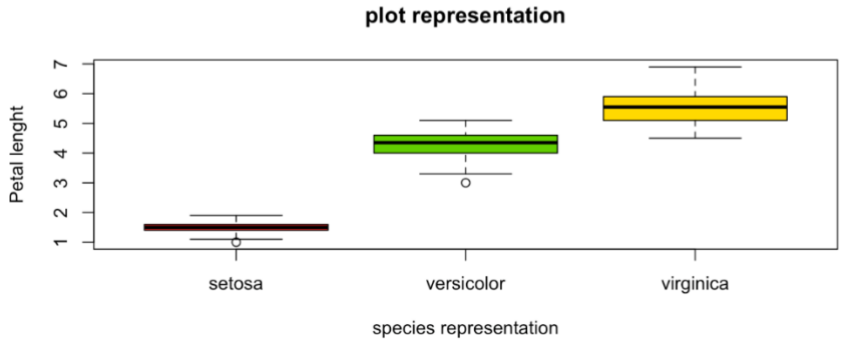


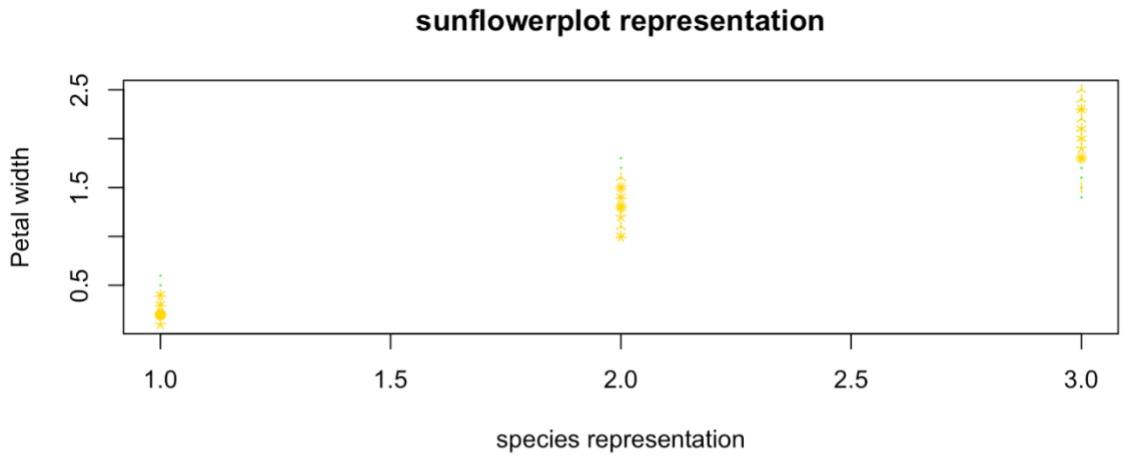


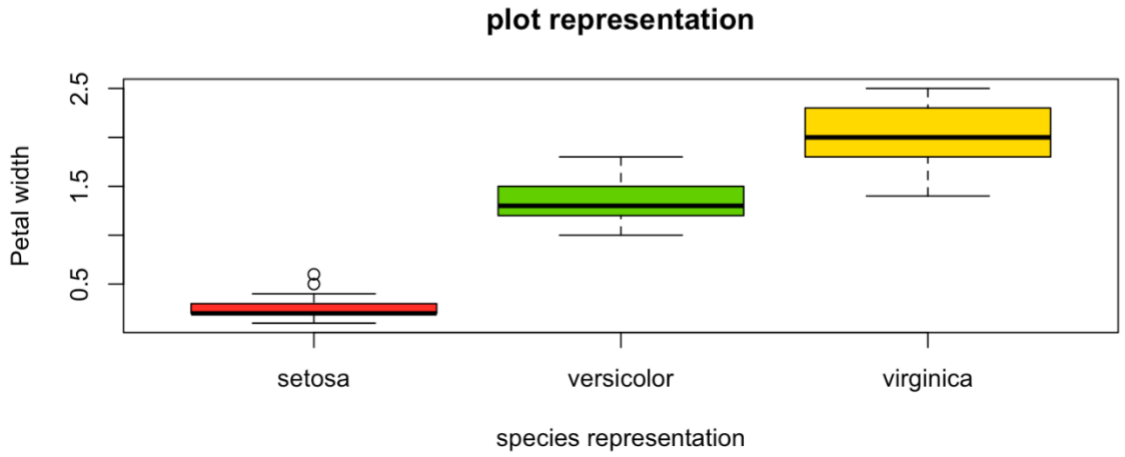

sunflowerplot representation
sepal width
2.0
3.0
4.0
1.0
1.5
2.0
2.5
3.0
species representation


plot representation
sepal width
2.0
3.0
4.0
setosa
versicolor
virginica
species representation


sunflowerplot representation
Petal Lenght
1 2 3 4 5 6 7
1.0
1.5
2.0
2.5
3.0
species representation


plot representation
Petal lenght
1 2 3 4 5 6 7
setosa
versicolor
virginica
species representation





c)Graphic representation for the different data categories

9)Represent the scatter plot for the the dependent and independant variables for each data category. Comment

III)Regression Analysis

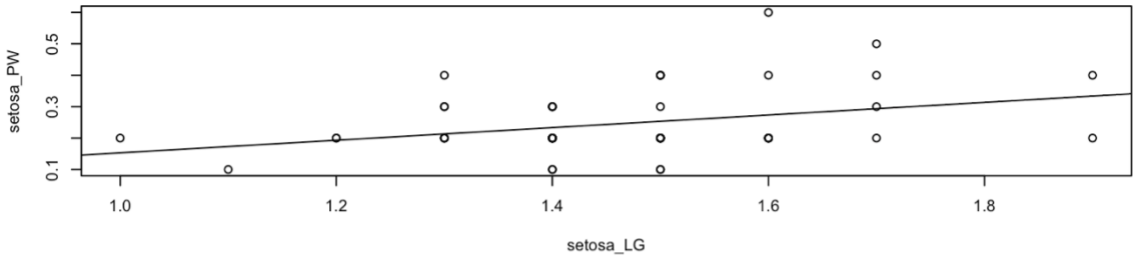
a)Testing hypothesis

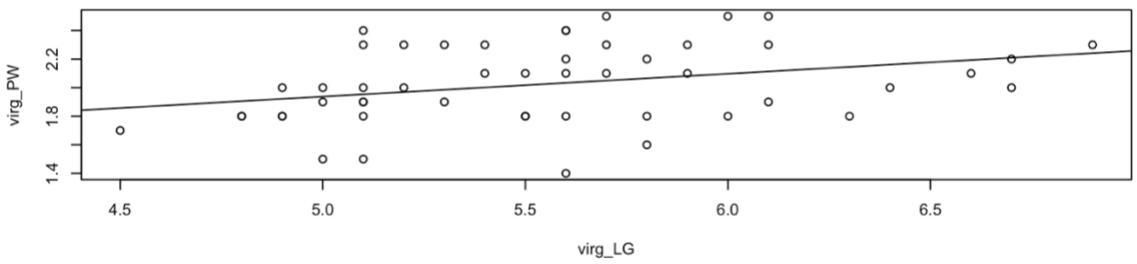
10)Use the Khi-deux test to verify the indenpendant of each data category.

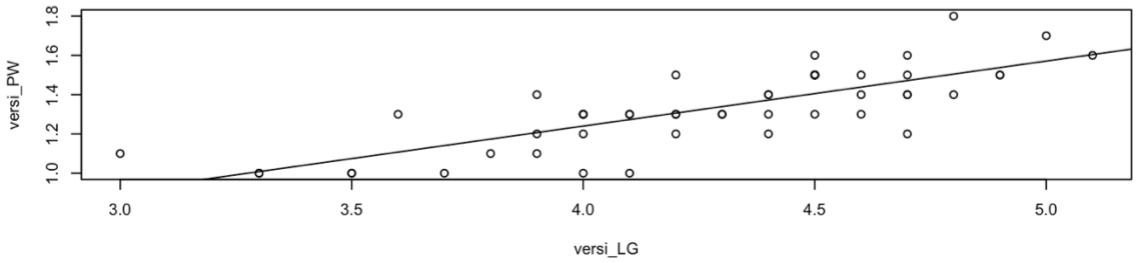
11)Testing the Standard Assumptions of linear regression

b)Build the model the regression model

c)Verify model significance (Model validation)







Coef:

graph 1: Call:

lm(formula = setosa_PW ~ setosa_LG, data = iris)

Coefficients :

(Intercept) setosa_LG

-0.04822 0.20125

Graph 2: Call:

lm(formula = virg_PW ~ virg_LG, data = iris)

Coefficients :

(Intercept) virg_LG

1.1360 0.1603

Graph 3 : Call:

lm(formula = versi_PW ~ versi_LG, data = iris)

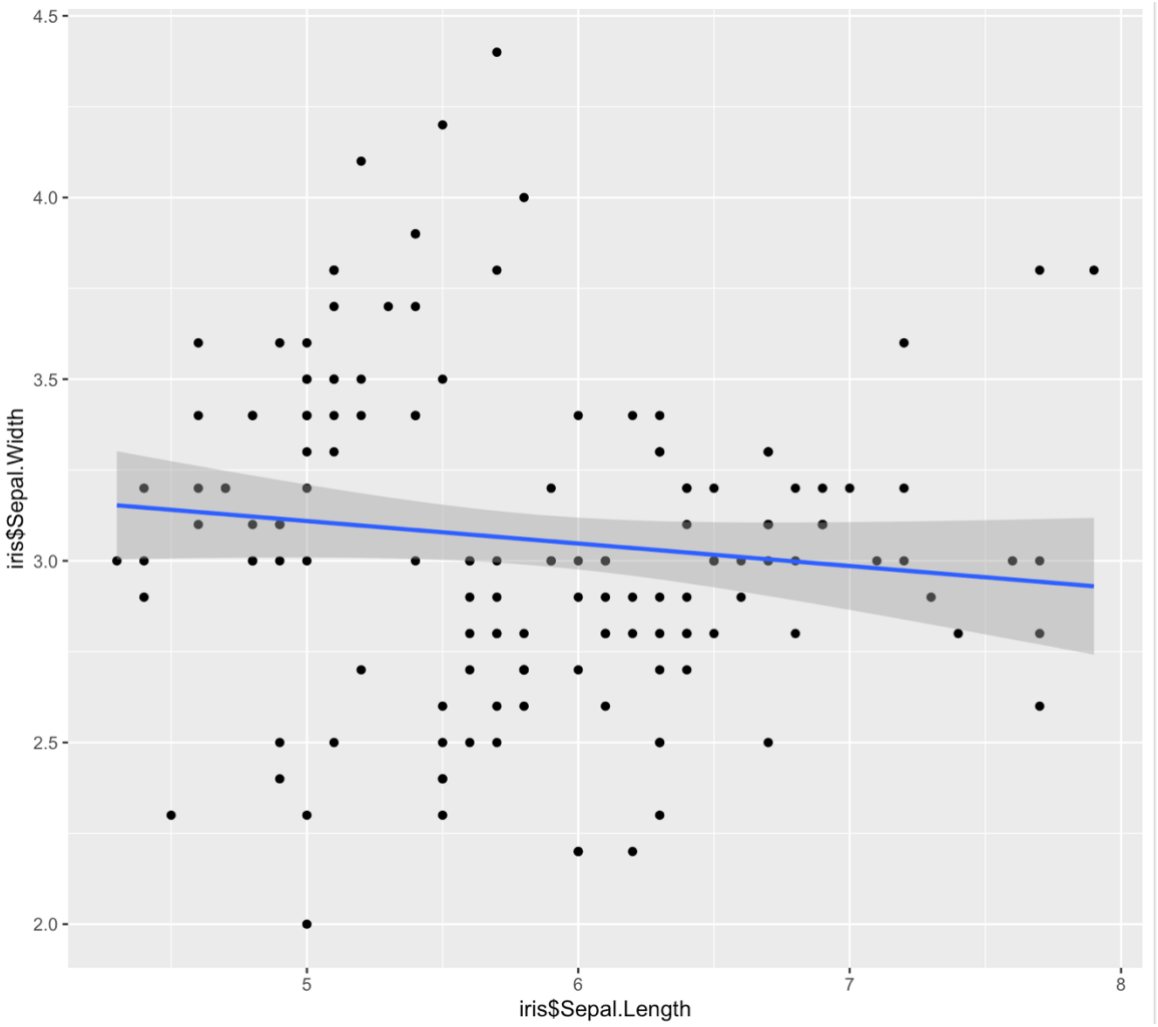
Coefficients:

(Intercept) versi_LG

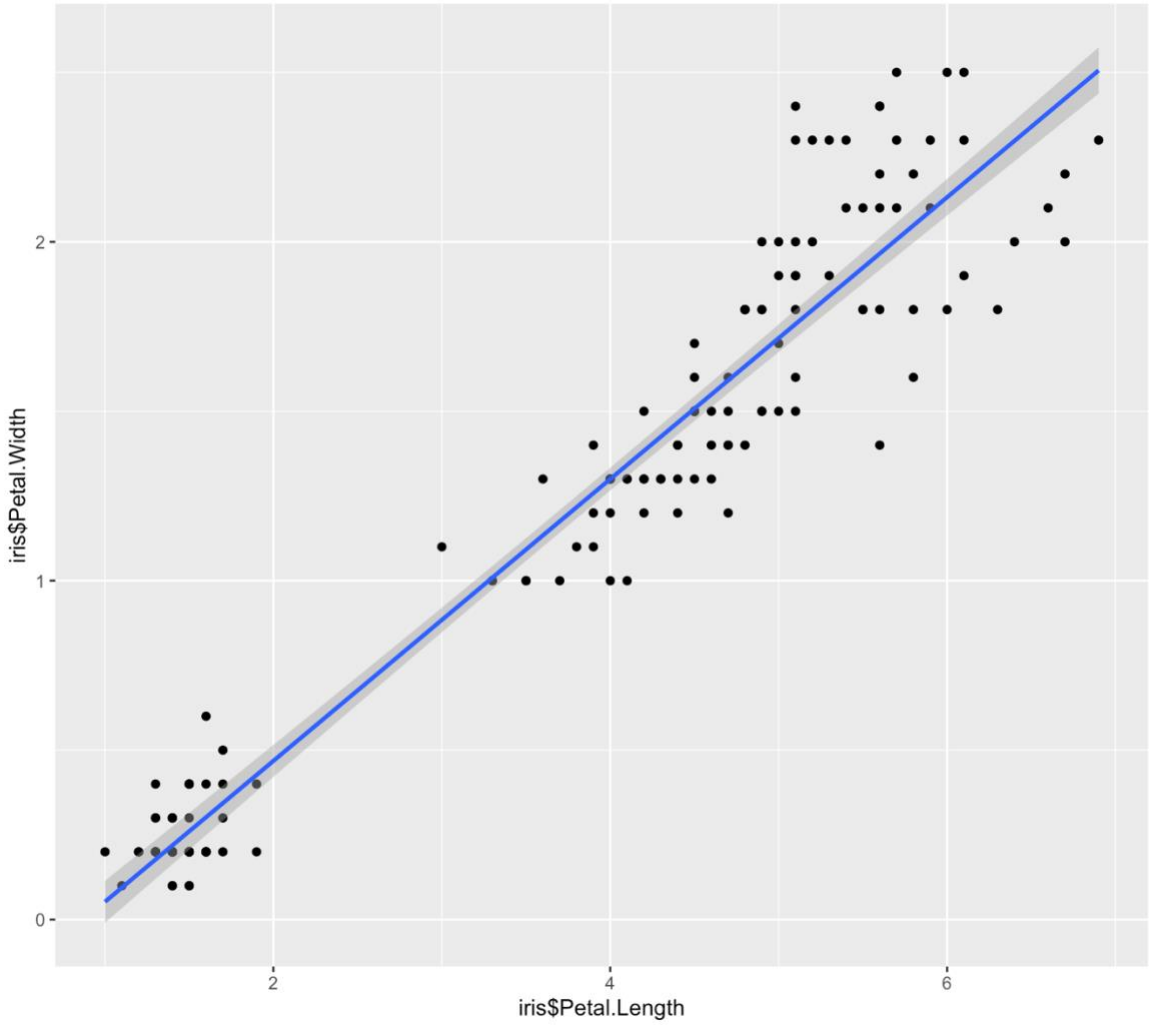
-0.08429 0.33105

we can see three different linear regression model (petal width and petal length) by species and we can see that for our dataset linear regression model is not relevant.

But if we look for correlation between length and width, we can see that for sepal it's not relevant but for petal they may have a correlation



Correlation between sepal attribute



Correlation between Petal attribute