Prédiction des fleurs d'iris avec la RL

Ousmane lom

Installation des librairies

```
library(prettyR)
library(tidyverse)
library(gt)
library(corrplot)
library(nnet)
```

Dataset des Fleurs d'iris

```
data=iris
```

Structure des données

```
'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 ...
```

Visualisez les 1er lignes

```
iris %>%
                  head(n=10) %>%
                  gt()
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa

Analyse descriptive de l'ensemble des variables

```
describe(data)
```

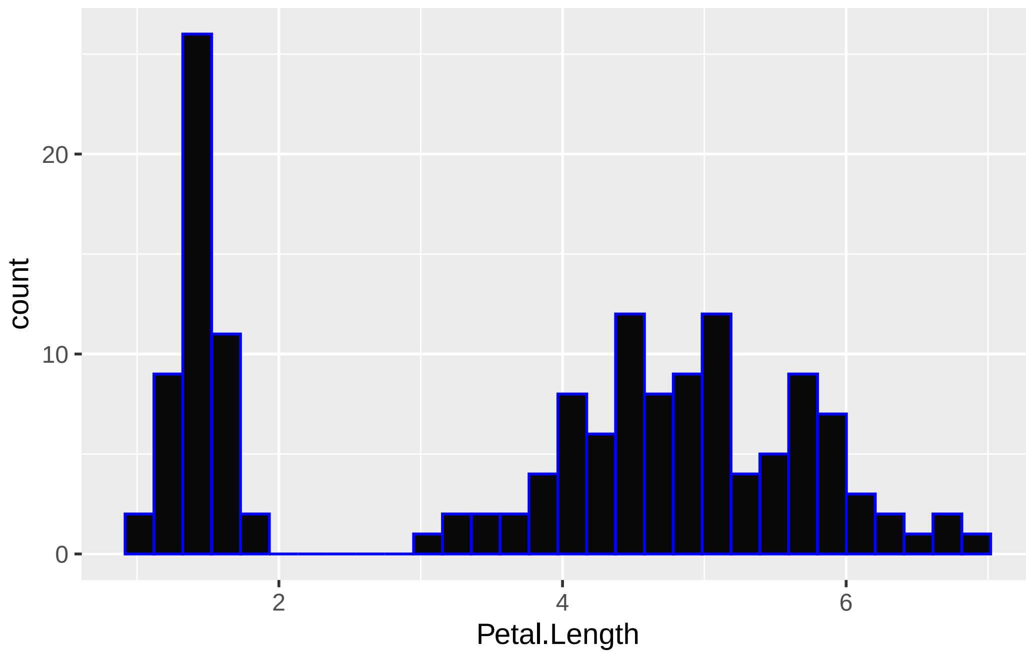
```
Description of data



 Numeric
             mean median  var   sd valid.n
Sepal.Length 5.84   5.80 0.69 0.83     150
Sepal.Width  3.06   3.00 0.19 0.44     150
Petal.Length 3.76   4.35 3.12 1.77     150
Petal.Width  1.20   1.30 0.58 0.76     150

 Factor

Species   setosa versicolor virginica
  Count    50.00      50.00     50.00
  Percent  33.33      33.33     33.33
Mode >1 mode
```

Visualisez les variables

Diagramme en histogramme de la variable Petal.length

```
ggplot(iris,aes(x=Petal.Length))+
                    geom_histogram(color="blue",fill="gray3")
```

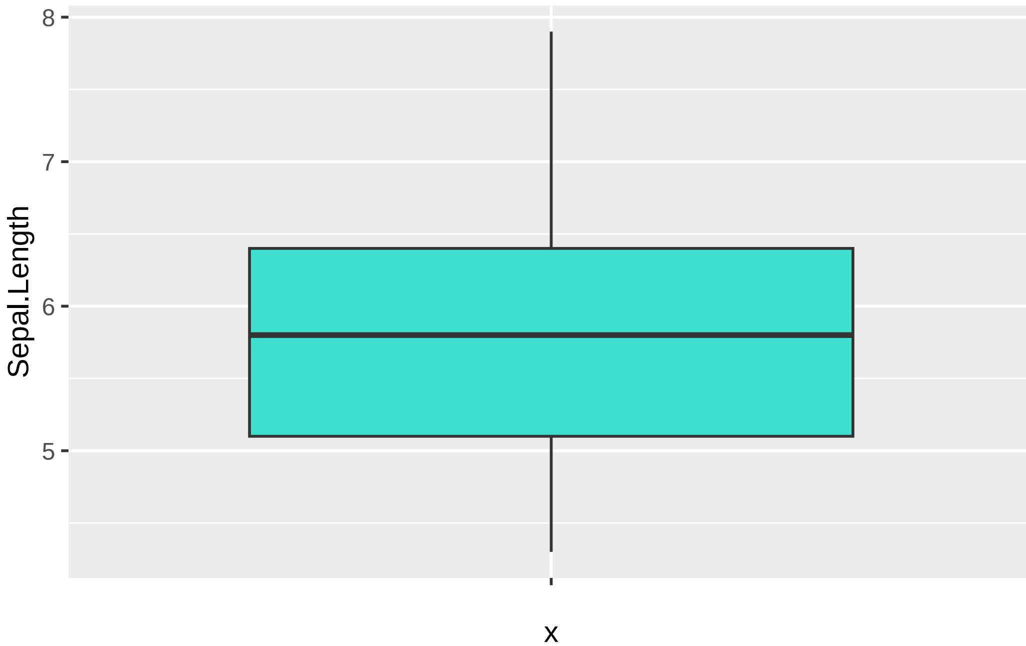
```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Distribution de la variable Petal.length

Diagramme en boxplot

```
ggplot(iris,aes(x="",y=Sepal.Length))+geom_boxplot(fill="turquoise",outliers = TRUE)
```

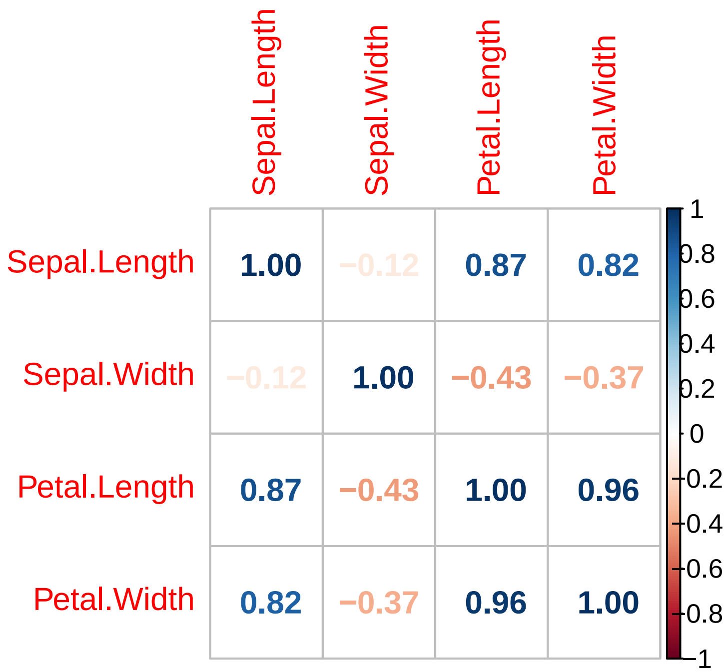


Correlation entre les variables quantitatives

```
corr=round(cor(iris[,-5]),2)
corr
```

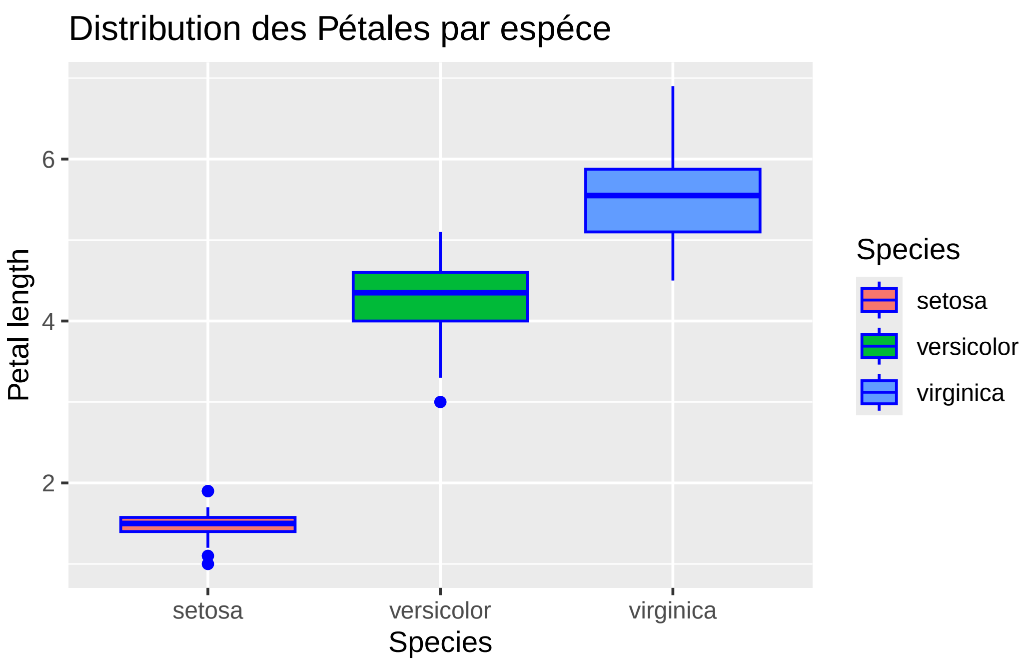
```
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length         1.00       -0.12         0.87        0.82
Sepal.Width         -0.12        1.00        -0.43       -0.37
Petal.Length         0.87       -0.43         1.00        0.96
Petal.Width          0.82       -0.37         0.96        1.00
```

```
corrplot(corr,method="number")
```



Distribution du Petal length par variété (Species)

```
ggplot(iris,aes(x=Species,y=Petal.Length,fill = Species))+
geom_boxplot(color="blue")+
                  labs(x="Species",y="Petal length",title="Distribution des Pétales par espé
```



Analyse de la variance

H0: Pas de différence significative de la longueur des pétales entre variétes

H1: Il ya une différence signicative entre la longueur des pétales entre variétes de fleurs

```
anova=aov(Petal.Length~Species,iris)
summary(anova)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)    
Species       2  437.1  218.55    1180 <2e-16 ***
Residuals   147   27.2    0.19                   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion: Donc on note un effet siginificatif entre les variétes selon la longueur des pétales

Apprentissage automatique

Prediction des fleurs par un modéle multinomial

Partitionner la base de donnée en donnée d'entrainement de notre modéle à 70% et test à 30%

```
set.seed(1234)
ind<-sample(2,nrow(iris),replace = T,prob = c(0.70,0.30))
train=iris[ind==1,]
test=iris[ind==2,]
```

Dimension train et test data

```
cat("dimension_train est:",dim(train),fill = T)
```

```
dimension_train est: 112 5
```

```
cat("dimension_test est:",dim(test),fill = T)
```

```
dimension_test est: 38 5
```

Construire le modéle

```
model=multinom(Species~., data=iris)
```

```
# weights:  18 (10 variable)
initial  value 164.791843
iter  10 value 16.177348
iter  20 value 7.111438
iter  30 value 6.182999
iter  40 value 5.984028
iter  50 value 5.961278
iter  60 value 5.954900
iter  70 value 5.951851
iter  80 value 5.950343
iter  90 value 5.949904
iter 100 value 5.949867
final  value 5.949867
stopped after 100 iterations
```

Resumé du modéle

```
summary(model)
```

```
Call:
multinom(formula = Species ~ ., data = iris)

Coefficients:
           (Intercept) Sepal.Length Sepal.Width Petal.Length Petal.Width
versicolor    18.69037    -5.458424   -8.707401     14.24477   -3.097684
virginica    -23.83628    -7.923634  -15.370769     23.65978   15.135301

Std. Errors:
           (Intercept) Sepal.Length Sepal.Width Petal.Length Petal.Width
versicolor    34.97116     89.89215    157.0415     60.19170    45.48852
virginica     35.76649     89.91153    157.1196     60.46753    45.93406

Residual Deviance: 11.89973
AIC: 31.89973
```

Prédiction du model avec les données test

```
predictions=predict(model,test)
predictions[1:10]
```

```
 [1] setosa setosa setosa setosa setosa setosa setosa setosa setosa setosa
Levels: setosa versicolor virginica
```

Calcul de l'accuracy du modéle

```
tab=table(test$Species,predictions)
tab
```

```
            predictions
             setosa versicolor virginica
  setosa         10          0         0
  versicolor      0         12         0
  virginica       0          0        16
```

Calcul de l'accuracy

$$accuray = \frac{\sum(diag(tab))}{\sum(tab)}$$

```
accuracy=sum(diag(tab))/sum(tab)
cat("accuracy est",accuracy,fill=T)
```

```
accuracy est 1
```

Le modéle prédit à 100% le type de fleur