TP: Analyse descriptive univariée et bivariée - Correction

Analyse des données

Master ISEFAR - M1

1 Durée de vie de piles

1.1 Données et caractéristiques

```
piles = read.table("piles.txt",header = TRUE,sep = "\t")
knitr::kable(head(piles))
```

MarqueA	MarqueB	MarqueC
65.1	64.4	62.8
58.4	69.1	58.6
64.9	66.9	63.3
76.0	67.5	65.3
67.8	65.8	78.8
75.1	70.4	63.1

```
str(piles)
```

```
## 'data.frame':    20 obs. of  3 variables:
##  $ MarqueA: num  65.1 58.4 64.9 76 67.8 75.1 76.7 64.2 74.9 77.6 ...
##  $ MarqueB: num  64.4 69.1 66.9 67.5 65.8 70.4 67.8 61.8 68.7 65.3 ...
##  $ MarqueC: num  62.8 58.6 63.3 65.3 78.8 63.1 76.3 64.2 61.8 73.9 ...
```

```
summary(piles)
```

```
##     MarqueA         MarqueB         MarqueC
##  Min.   :58.10   Min.   :61.80   Min.   :58.60
##  1st Qu.:65.05   1st Qu.:64.78   1st Qu.:63.25
##  Median :74.00   Median :67.20   Median :67.30
##  Mean   :70.45   Mean   :66.97   Mean   :68.83
##  3rd Qu.:76.00   3rd Qu.:68.80   3rd Qu.:74.03
##  Max.   :81.30   Max.   :72.00   Max.   :78.80
```

1.2 On réorganise le jeu de données

On veut avoir deux colonnes: une pour la marque et une pour la durée.

```r
piles.G <- piles %>% gather(key='Marque',value='duree')
head(piles.G)
```

```
##    Marque duree
## 1 MarqueA  65.1
## 2 MarqueA  58.4
## 3 MarqueA  64.9
## 4 MarqueA  76.0
## 5 MarqueA  67.8
## 6 MarqueA  75.1
```

```r
str(piles.G)
```

```
## 'data.frame':    60 obs. of  2 variables:
##  $ Marque: chr  "MarqueA" "MarqueA" "MarqueA" "MarqueA" ...
##  $ duree : num  65.1 58.4 64.9 76 67.8 75.1 76.7 64.2 74.9 77.6 ...
```

```r
piles.G$Marque <- as.factor(piles.G$Marque)
```

1.3 Moyenne et écart-type par marque

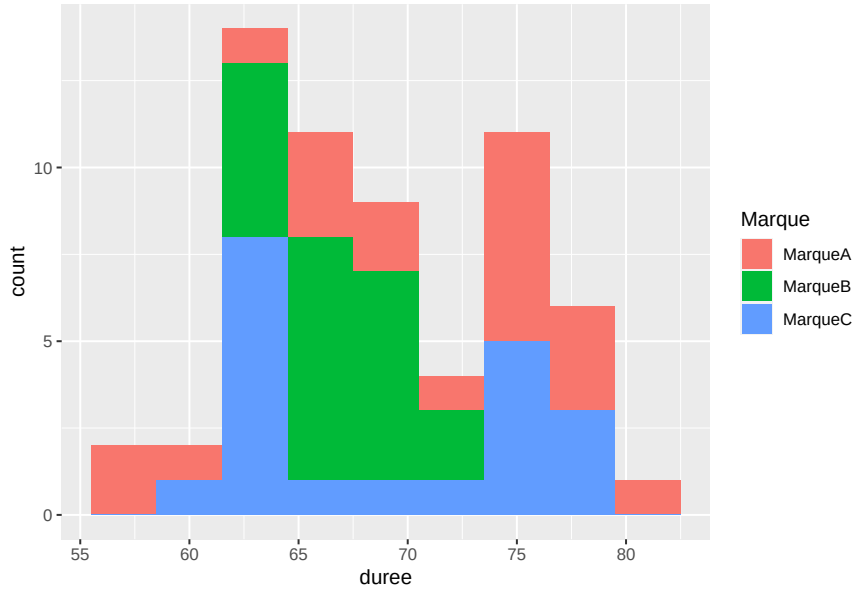
```r
piles.G %>% group_by(Marque) %>% summarise(meanD=mean(duree),stD=sd(duree)) %>% knitr::kable()
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

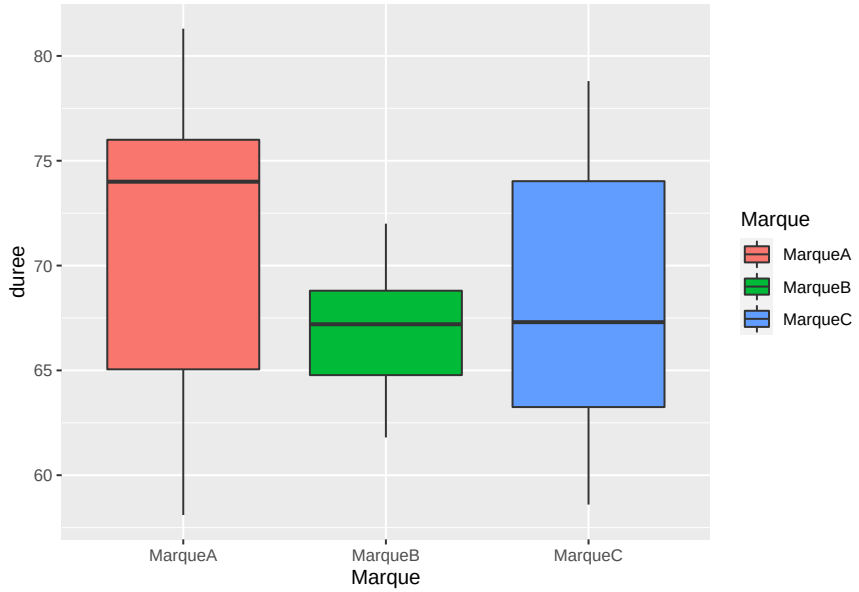
Marque	meanD	stD
MarqueA	70.45	7.092657
MarqueB	66.97	2.856221
MarqueC	68.83	6.445080

1.3.1 Représentations graphiques

```r
# Histogramme
piles.G %>% ggplot(aes(x=duree,fill=Marque))+geom_histogram(binwidth=3)
```




```
# Boxplot
piles.G %>% ggplot(aes(x=Marque,y=duree,fill=Marque))+geom_boxplot()
```



2 Les iris

2.1 Données et caractéristiques des variables

```
data(iris)
knitr::kable(head(iris))
```

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

```
dim(iris)
```

```
## [1] 150   5
```

```
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 1 ...
```

2.2 Variable Species

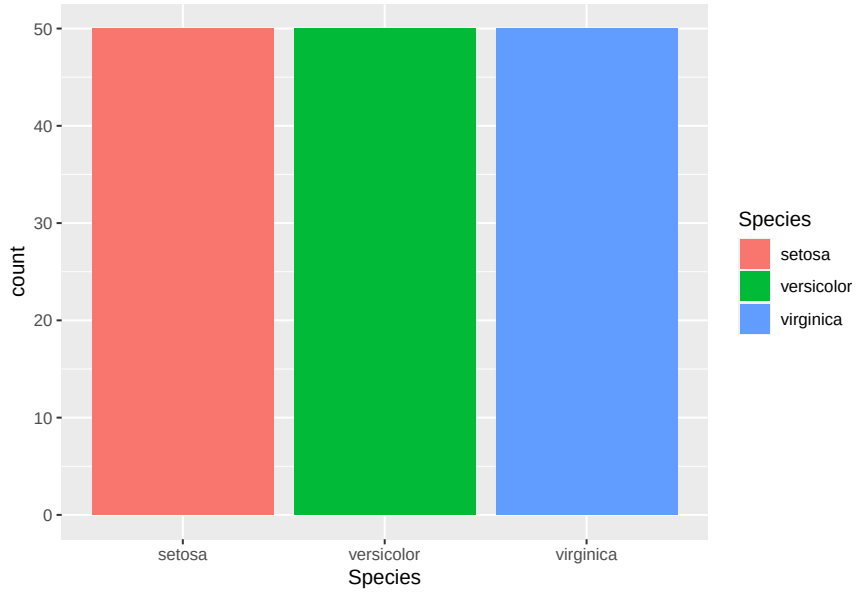
```
levels(iris$Species)
```

```
## [1] "setosa"     "versicolor" "virginica"
```

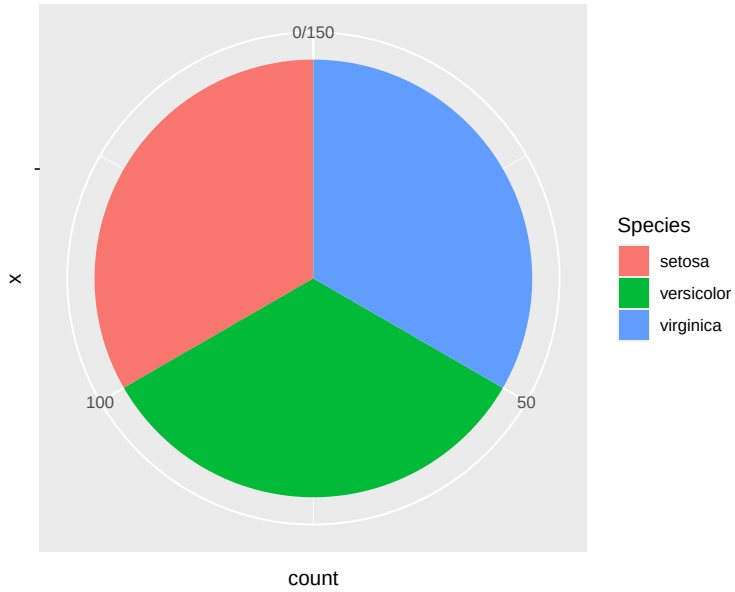
```
# Effectif par modalité
iris %>% group_by(Species) %>% count()
```

```
## # A tibble: 3 x 2
## # Groups:   Species [3]
##   Species        n
##   <fct>      <int>
## 1 setosa        50
## 2 versicolor    50
## 3 virginica     50
```

```
# diagramme en bâtons
ggplot(iris,aes(x = Species,fill=Species)) +geom_bar()
```




```
# camenbert
#pie(table(iris$Species))
ggplot(iris, aes(x="", fill=Species))+geom_bar(width = 2)+coord_polar("y")
```



2.2.1 Variables quantitatives

Quelques statistiques simples et boxplot

```
iris %>% summarise_if(is.numeric,mean)
```

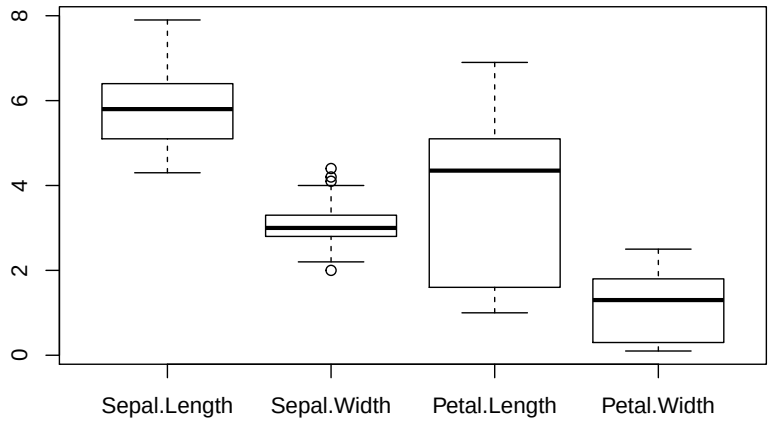
```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1     5.843333    3.057333        3.758    1.199333
```

```
iris %>% summarise_if(is.numeric,sd)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1    0.8280661   0.4358663     1.765298   0.7622377
```

```
iris %>% select(is.numeric) %>% boxplot()
```

```
## Warning: Predicate functions must be wrapped in `where()`.
##
##   # Bad
##   data %>% select(is.numeric)
##
##   # Good
##   data %>% select(where(is.numeric))
##
## i Please update your code.
## This message is displayed once per session.
```



En fonction de l'espèce

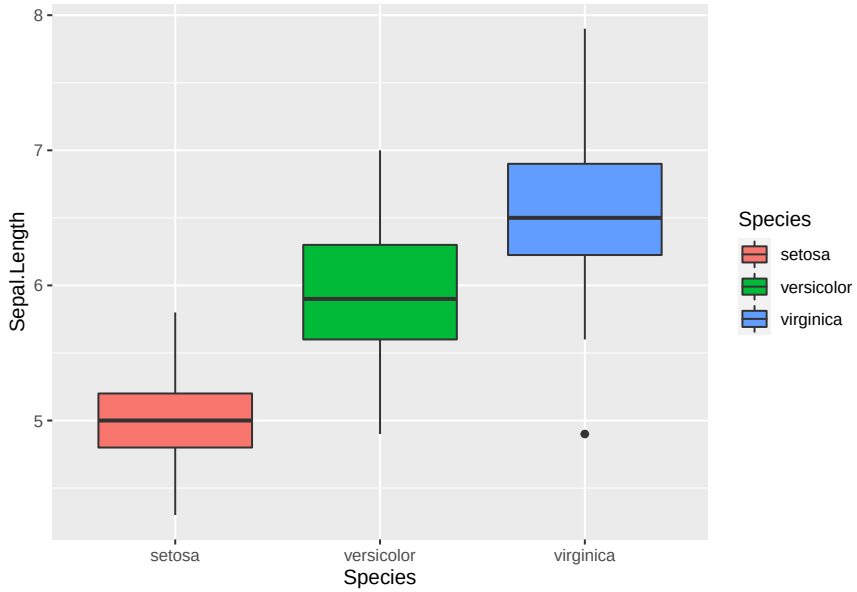
```
iris %>% group_by(Species) %>% summarise_if(is.numeric,mean)
```

```
## # A tibble: 3 x 5
##   Species    Sepal.Length Sepal.Width Petal.Length Petal.Width
##   <fct>             <dbl>       <dbl>        <dbl>       <dbl>
## 1 setosa             5.01        3.43         1.46       0.246
## 2 versicolor         5.94        2.77         4.26       1.33
## 3 virginica          6.59        2.97         5.55       2.03
```

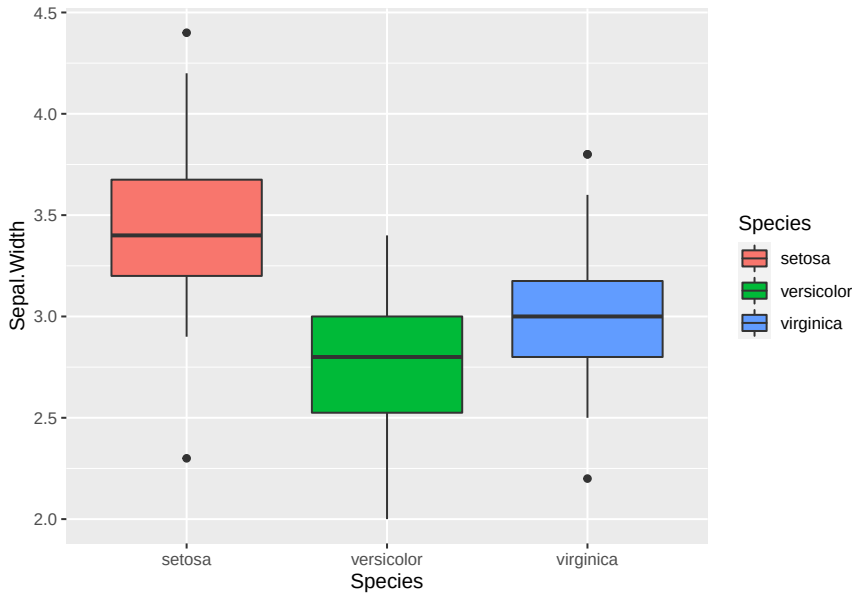
```
iris %>% group_by(Species) %>% summarise_if(is.numeric,sd)
```

```
## # A tibble: 3 x 5
##   Species    Sepal.Length Sepal.Width Petal.Length Petal.Width
##   <fct>             <dbl>       <dbl>        <dbl>       <dbl>
## 1 setosa            0.352       0.379        0.174       0.105
## 2 versicolor        0.516       0.314        0.470       0.198
## 3 virginica         0.636       0.322        0.552       0.275
```

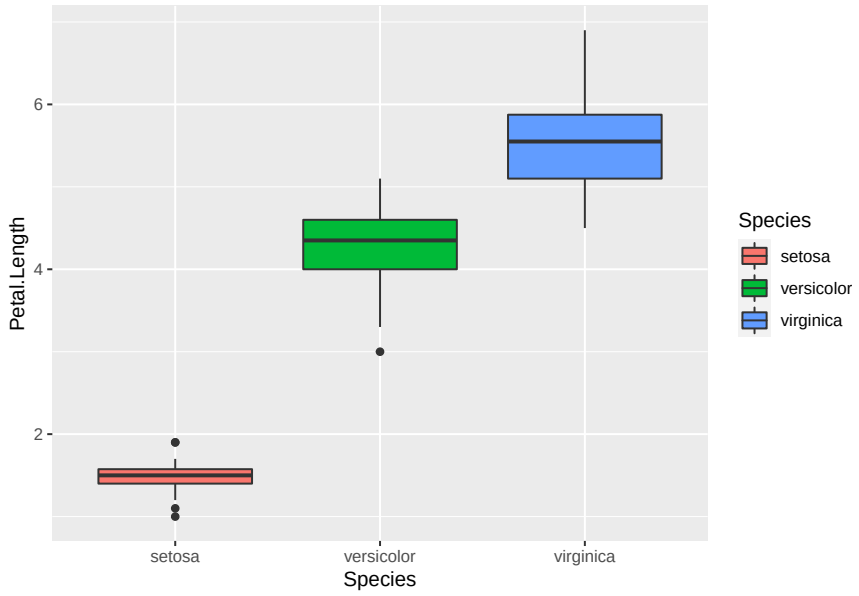
```
ggplot(iris, aes(x=Species, y=Sepal.Length, fill=Species)) +geom_boxplot()
```



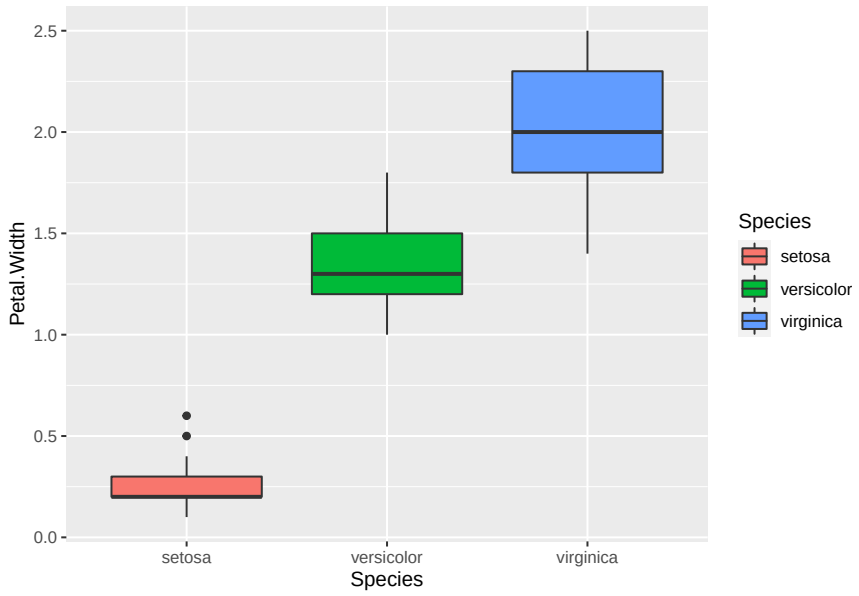

```
ggplot(iris, aes(x=Species, y=Sepal.Width, fill=Species)) +geom_boxplot()
```



```r
ggplot(iris, aes(x=Species, y=Petal.Length, fill=Species)) +geom_boxplot()
```




```r
ggplot(iris, aes(x=Species, y=Petal.Width, fill=Species)) +geom_boxplot()
```



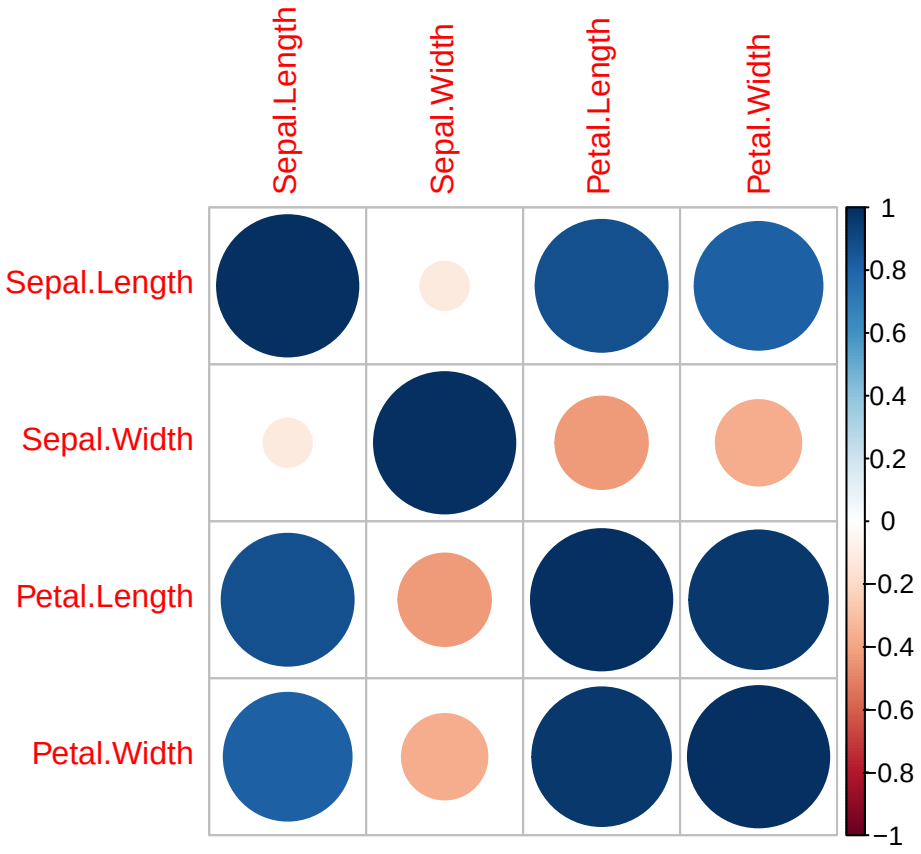
2.2.2 Lien entre variables quantitatives

Corrélations

```r
correlation <-  iris %>% select_if(is.numeric) %>% cor()
kable(correlation,digits=3)
```

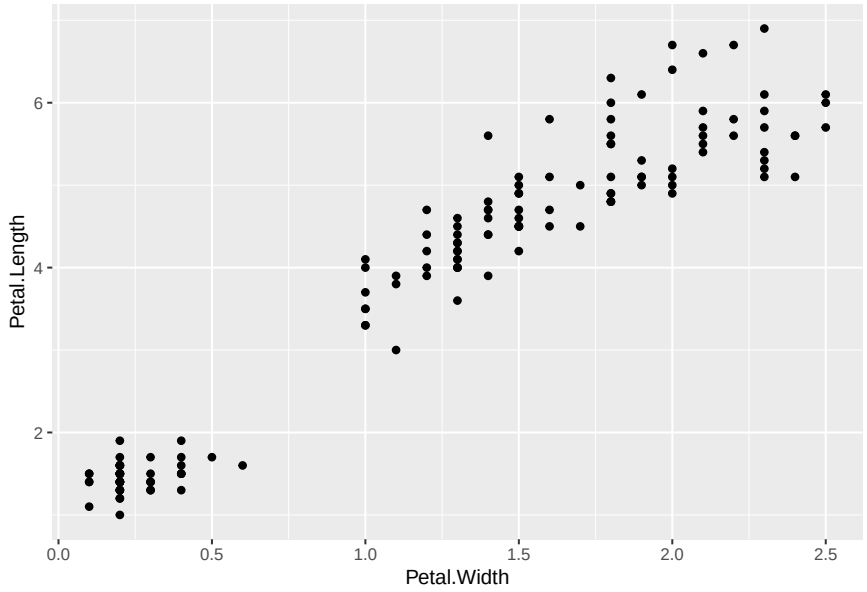
	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	1.000	-0.118	0.872	0.818
Sepal.Width	-0.118	1.000	-0.428	-0.366
Petal.Length	0.872	-0.428	1.000	0.963
Petal.Width	0.818	-0.366	0.963	1.000

```
corrplot(correlation)
```

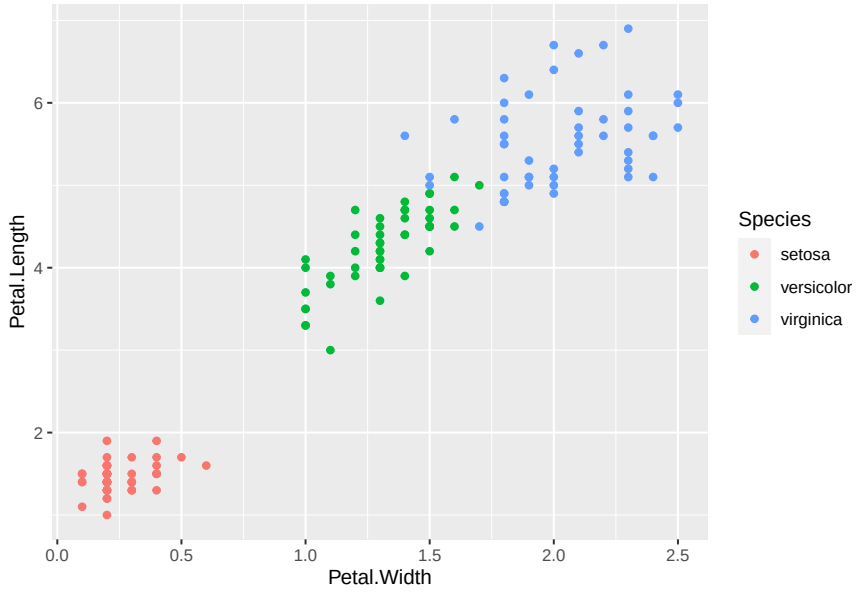


Lien entre variables 2 à 2

```
ggplot(iris,aes(x=Petal.Width,y=Petal.Length))+geom_point()
```

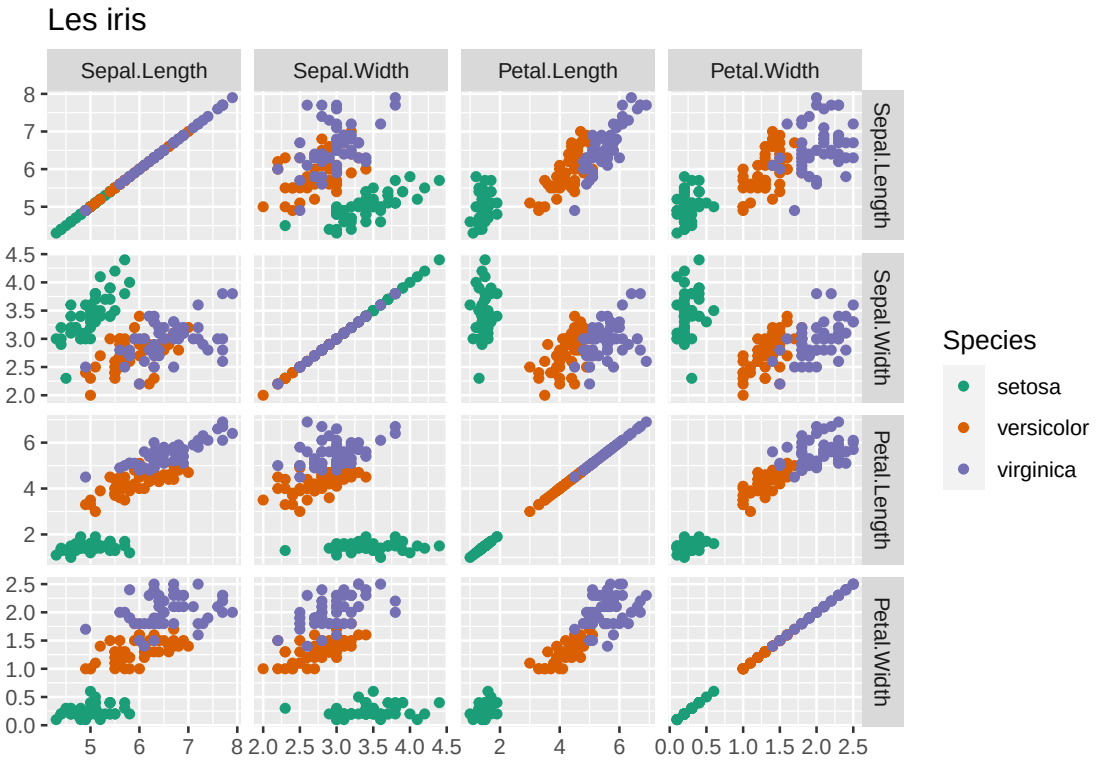



```
# Si on veut colorier selon l'espèce
ggplot(iris,aes(x=Petal.Width,y=Petal.Length,color=Species))+geom_point()
```



2.2.3 Graphe entre toutes les variables quantitatives avec coloration des points selon l'espèce

```
PairPlot(iris, colnames(iris)[1:4],"Les iris", group_var = "Species")
```



3 Couleur des cheveux et des yeux

3.0.1 Données

```
data(HairEyeColor)
dim(HairEyeColor)
```

```
## [1] 4 4 2
```

```
is.array(HairEyeColor)
```

```
## [1] TRUE
```

```
knitr::kable(HairEyeColor[1,,])
```

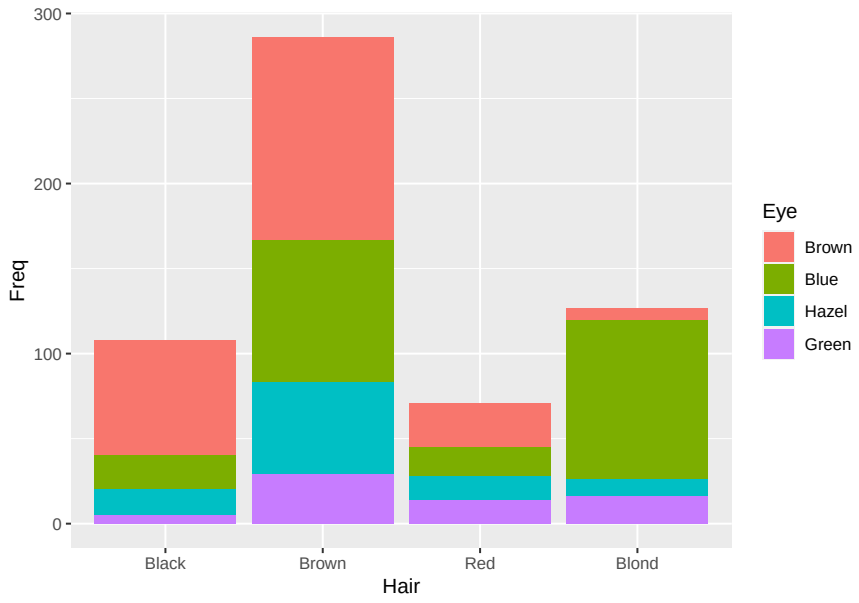
	Male	Female
Brown	32	36
Blue	11	9
Hazel	10	5
Green	3	2

3.0.2 Contruction du jeu de données d'intérêt et graphe

```
HEC = HairEyeColor[,,1] +HairEyeColor[,,2]
knitr::kable(HEC)
```

	Brown	Blue	Hazel	Green
Black	68	20	15	5
Brown	119	84	54	29
Red	26	17	14	14
Blond	7	94	10	16

```
HEC %>% as.data.frame() %>% ggplot(aes(x=Hair,fill=Eye))+geom_bar(aes(y=Freq),stat="identity")
```

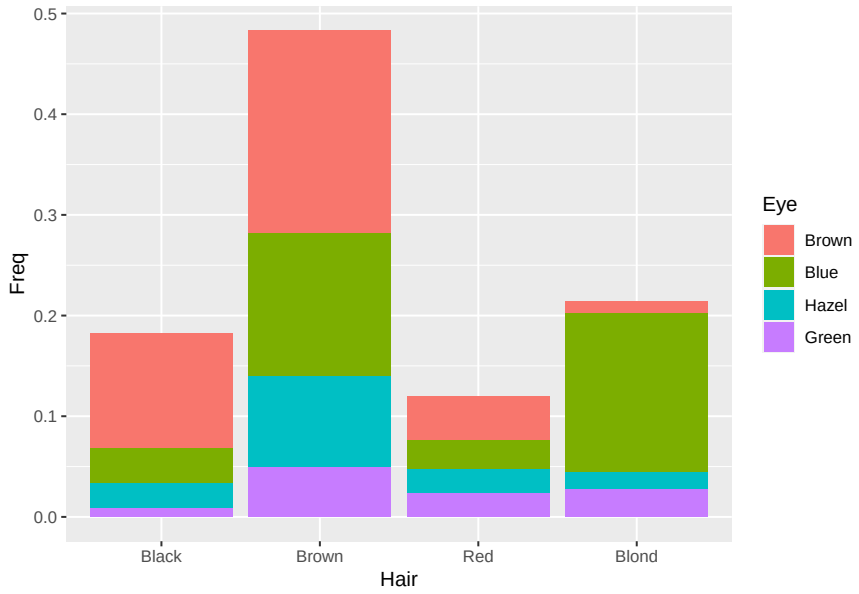


3.0.3 Fréquences coinjointes

```
F=HEC/sum(HEC)
knitr::kable(F,digits=2)
```

	Brown	Blue	Hazel	Green
Black	0.11	0.03	0.03	0.01
Brown	0.20	0.14	0.09	0.05
Red	0.04	0.03	0.02	0.02
Blond	0.01	0.16	0.02	0.03

```
F %>% as.data.frame() %>% ggplot(aes(x=Hair,fill=Eye))+geom_bar(aes(y=Freq),stat="identity")
```

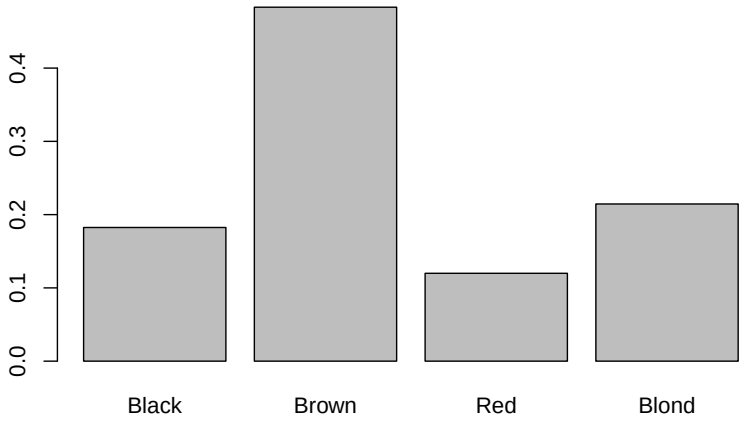


3.0.4 Fréquences marginales

```
# En ligne
FM_L = margin.table(F,1)
knitr::kable(FM_L,digits=2)
```

Hair	Freq
Black	0.18
Brown	0.48
Red	0.12
Blond	0.21

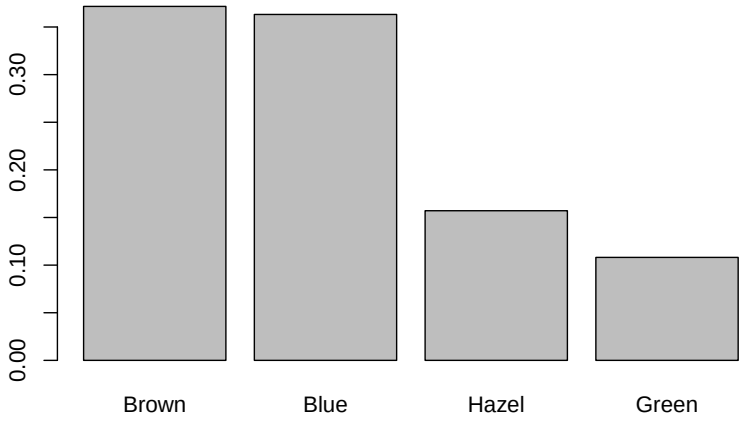
```
FM_L %>% barplot()
```




```
# En colonne
FM_C =margin.table(F,2)
knitr::kable(FM_C,digits=2)
```

Eye	Freq
Brown	0.37
Blue	0.36
Hazel	0.16
Green	0.11

```
FM_C %>% barplot()
```

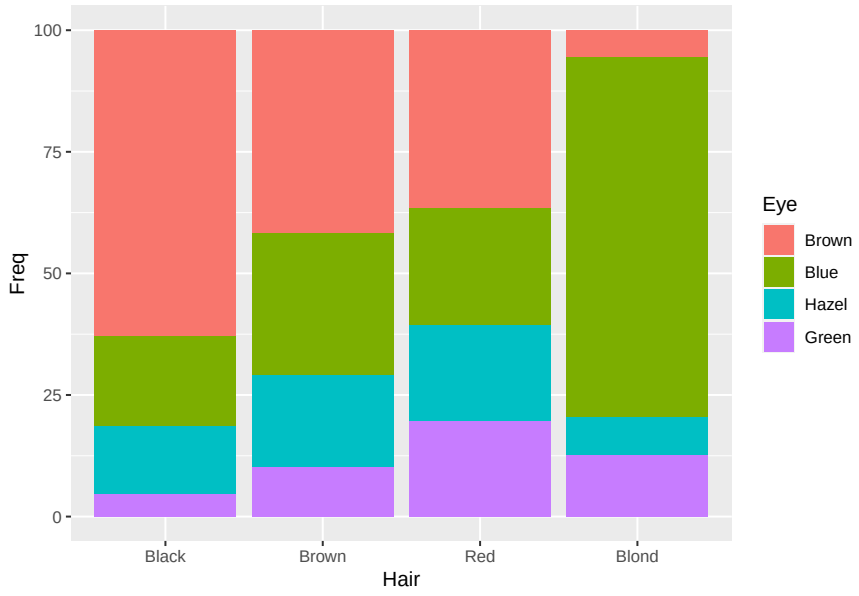


3.0.5 Fréquences conditionnelles

```
# Profil ligne
Profil.Ligne =rprop(HEC,digits = 2)
knitr::kable(Profil.Ligne)
```

	Brown	Blue	Hazel	Green	Total
Black	62.962963	18.51852	13.888889	4.62963	100
Brown	41.608392	29.37063	18.881119	10.13986	100
Red	36.619718	23.94366	19.718310	19.71831	100
Blond	5.511811	74.01575	7.874016	12.59843	100
Ensemble	37.162162	36.31757	15.709459	10.81081	100

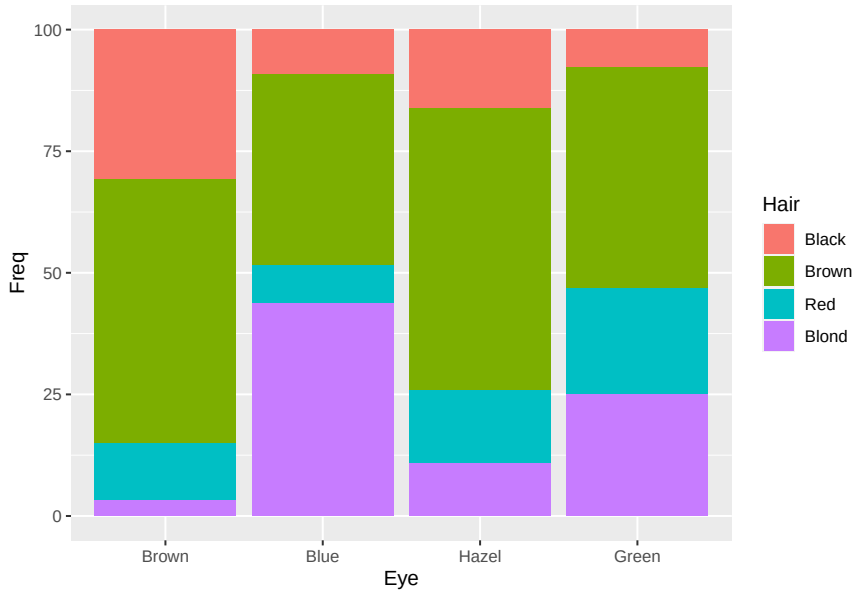
```
rprop(HEC,total=FALSE) %>% as.data.frame() %>% ggplot(aes(x=Hair,fill=Eye))+geom_bar(aes(y=Freq), stat=
```




```
# Profil colonne
Profil.Colonne =cprop(HEC,digits = 2)
knitr::kable(Profil.Colonne)
```

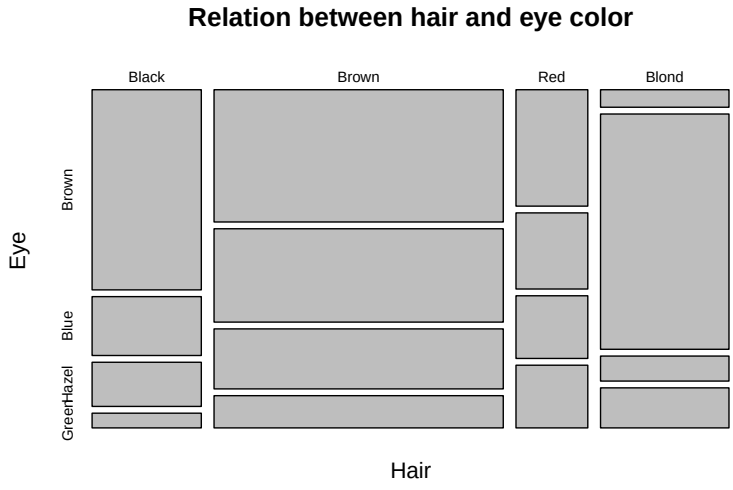
	Brown	Blue	Hazel	Green	Ensemble
Black	30.909091	9.302326	16.12903	7.8125	18.24324
Brown	54.090909	39.069767	58.06452	45.3125	48.31081
Red	11.818182	7.906977	15.05376	21.8750	11.99324
Blond	3.181818	43.720930	10.75269	25.0000	21.45270
Total	100.000000	100.000000	100.00000	100.0000	100.00000

```r
cprop(HEC,total=FALSE) %>% as.data.frame() %>% ggplot(aes(x=Eye,fill=Hair))+geom_bar(aes(y=Freq), stat=
```

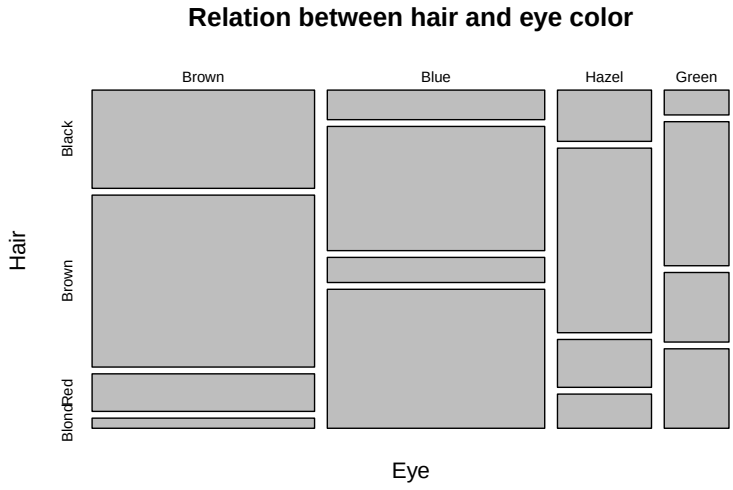


3.0.6 Mosaic

```r
mosaicplot(HEC, main = "Relation between hair and eye color")
```




```r
mosaicplot(t(HEC), main = "Relation between hair and eye color")
```



Pour le graphique 1 (HEC)

```
* Chaque rectangle représente une case du tableau

* la largeur des barres = fréquences marginales par ligne
```

```
FM_L
```

```
## Hair
##     Black     Brown       Red     Blond
## 0.1824324 0.4831081 0.1199324 0.2145270
```

```
* la longueur au sein de chaque barre = fréquences conditionnelles par ligne
```

```
Profil.Ligne
```

```
##           Eye
## Hair       Brown  Blue   Hazel  Green  Total
##   Black     62.96  18.52  13.89   4.63 100.00
##   Brown     41.61  29.37  18.88  10.14 100.00
##   Red       36.62  23.94  19.72  19.72 100.00
##   Blond      5.51  74.02   7.87  12.60 100.00
##   Ensemble  37.16  36.32  15.71  10.81 100.00
```

```
* la proportion d'individus ayant les cheveux bruns?
```

```
FM_L
```

```
## Hair
##     Black     Brown       Red     Blond
## 0.1824324 0.4831081 0.1199324 0.2145270
```

```
# donc 0.48
```

* la proportion d'individus ayant les yeux verts?

```
FM_C
```

```
## Eye
##     Brown      Blue     Hazel     Green
## 0.3716216 0.3631757 0.1570946 0.1081081
```

```
# donc 0.11
```

* Parmi les individus ayant les cheveux bruns, quelle est la proportion d'individus ayant les yeux verts

```
Profil.Ligne
```

```
##           Eye
## Hair       Brown  Blue   Hazel  Green  Total
##   Black     62.96  18.52  13.89   4.63 100.00
##   Brown     41.61  29.37  18.88  10.14 100.00
##   Red       36.62  23.94  19.72  19.72 100.00
##   Blond      5.51  74.02   7.87  12.60 100.00
##   Ensemble  37.16  36.32  15.71  10.81 100.00
```

```
# donc 0.1
```

* Parmi les individus ayant les yeux verts, quelle est la proportion d'individus ayant les cheveux bruns

```
Profil.Colonne
```

```
##        Eye
## Hair    Brown  Blue   Hazel  Green  Ensemble
##   Black  30.91   9.30  16.13   7.81  18.24
##   Brown  54.09  39.07  58.06  45.31  48.31
##   Red    11.82   7.91  15.05  21.88  11.99
##   Blond   3.18  43.72  10.75  25.00  21.45
##   Total 100.00 100.00 100.00 100.00 100.00
```

```
# donc 0.45
```

3.0.7 Test de l'indépendance entre Eye et Hair

```
h.chi = chisq.test(HEC)
h.chi
```

```
##
##  Pearson's Chi-squared test
##
## data:  HEC
## X-squared = 138.29, df = 9, p-value < 2.2e-16
```