I. Présentation du jeu de données

I-1. Importation des données

On commence par importer les différentes librairies permettant d'analyser le jeu de données :

```
library(corrplot)
library(FactoMineR)
```

On importe ensuite les données dans une table *meteo*.On obtient une table *meteo.3MIC* contenant les différentes données que l'on souhaite analyser.

```
knitr::opts_chunk$set(echo = TRUE)
meteo.3MIC<-read.table("~/meteo-3MIC.txt", quote="\"", comment.char="")
```

I-2. Nature des variables

Le jeu de données comporte 688 observations et 14 variables que l'on peut classer par nature :

qualitative ordinale	qualitative nominale	quantitative continue	quantitative discrète
month	season	td	
	wind_dir	ff	
		hu	
		t	
		precip	
		ws_arome	
		t2m_arome	
		d2m_arome	
		r_arome	
		msl_arome	
		tp_arome	

I-3. Mise en forme des données

Avant de commencer l'analyse, on change le type des variables qualitatives en *factor* :

```
meteo.3MIC$wind_dir = factor(meteo.3MIC$wind_dir,labels=c("north","west","south","east"))
meteo.3MIC$season=as.factor(meteo.3MIC$season)
meteo.3MIC$month=as.factor(meteo.3MIC$month)
```

II. Analyse uni- et bi-dimensionnelle

II-1. Analyse unidimensionelle

En ce qui concerne l'analyse unidimensionelle, on fera les choix de représentation suivants :

- Diagramme circulaire pour les variables qualitatives nominales car les différentes modalités sont représentées par des aires proportionnelles aux fréquences des modalités.

- Diagramme en bâtons des fréquences (cumulées ou non) pour les variables ordinales afin de conserver la relation d'ordre que contiennent ces différentes variables.

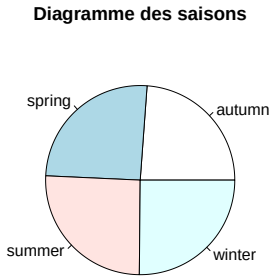
- Histogramme ou courbe des fréquences cumulées pour les variables quantitatives continues afin de conserver le caractère continu entre deux valeurs possibles pour les données.

- Pour les variables quantitatives discrètes, on utilise normalement un diagramme en bâtons ou la fonction de répartition mais l'étendue des valeurs nous oblige ici à utiliser les mêmes représentations que pour les variables quantitatives continues afin de conserver un rendu exploitable.

II-1-a. Variables qualitatives

Variable *season*

```
Season<-meteo.3MIC$season
pie(table(Season),main="Diagramme des saisons")
```



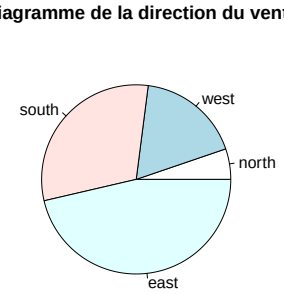

```
round(prop.table(table(Season))*100)
```

```
## Season
## autumn spring summer winter
##     24     25     26     25
```

On a 4 modalités : autumn,spring,summer,winter.On a presque le même nombre d'observations sur chaque saison(25% pour chaque saison).

Variable *wind_dir*

```
Dir<-meteo.3MIC$wind_dir
pie(table(Dir),main="Diagramme de la direction du vent")
```



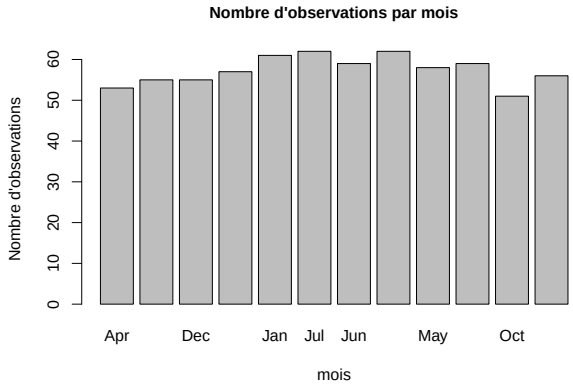

```
round(prop.table(table(Dir))*100)
```

```
## Dir
## north  west south  east
##     5    18    31    46
```

On remarque que le vent qui se dirige vers l'Est est présent dans 46% des observations.Le vent le moins représenté est celui qui se dirige vers le Nord avec seulement 4% des obsevations.Après lui on trouve le vent de l'Ouest avec 18% d'observations et enfin le vent du Sud présent à 31% des observations.

Variable *month*

```
month=meteo.3MIC$month
barplot(table(month),main="Nombre d'observations par mois",xlab="mois",
ylab="Nombre d'observations",cex.main=1,cex.lab=1)
```




```
range(table(month))
```

```
## [1] 51 62
```

```
mean(table(month))
```

```
## [1] 57.33333
```

On remarque qu'il y a plus d'observations effectuées entre Janvier et Mai que les autres mois.Mais il y n'a pas de grandes différences au total.En fait on a presque le même nombre d'obervations pour chaque mois(le max est à 62 et le min est à 51 et la moyenne est à de 57.33).

II-1-b. Variables quantitatives

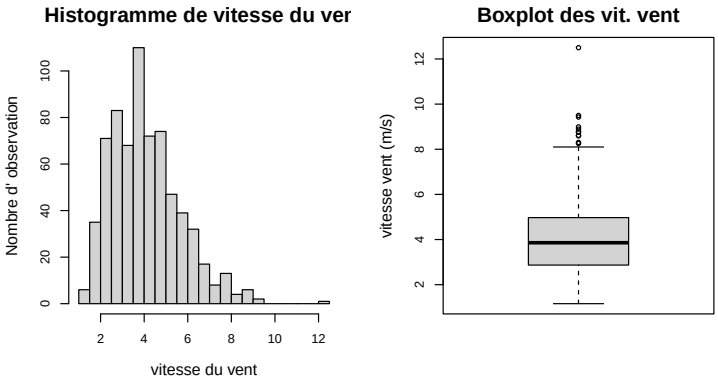
On remarque que dans ce jeu de données les 10 premières variables quantitatives peuvent être divisées en deux groupes:

observations d'aujourd'hui	prévisions de demain
t	t2m_arome
td	d2m_arome
ff	
hu	r_arome
precip	tp_arome

Dans le reste de l'analyse unidimensionelle et bi-dimensionnelle,on va juste se concentrer sur les observations d'aujourd'hui(t,td,ff,precip,hu) et aussi la variable *msl_arome*(la pression au niveau de la mer).

Variable *vitesse du vent(ff)*

```
ff<-meteo.3MIC$ff
par(cex=0.7, mai=c(0.8,0.8,0.4,0.1));par(fig=c(0,0.5,0.1,0.9))
hist(ff,main="Histogramme de vitesse du vent",xlab="vitesse du vent",
     ylab="Nombre d' observation",freq=TRUE,breaks=30,cex.main=1.5,cex.lab=1.2)
par(fig=c(0.5,1,0.1,0.9), new=TRUE)
boxplot(ff,main="Boxplot des vit. vent",ylab="vitesse vent (m/s)",cex.main=1.5,cex.lab=1.2)
```



```
summary(ff)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.158   2.870   3.859   4.082   4.970  12.500
```

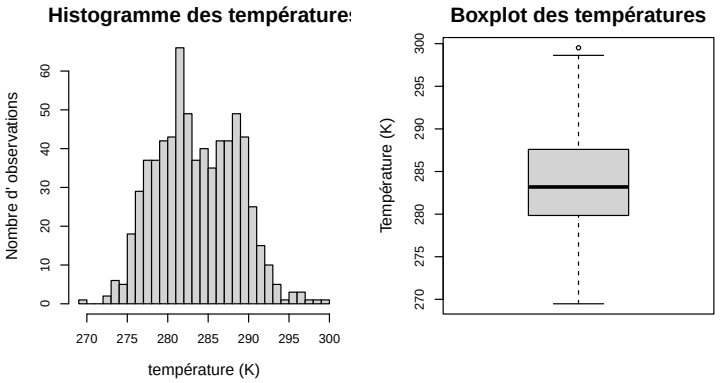
```
sd(ff)
```

```
## [1] 1.591358
```

On remarque une forte concentration des données sur la tranche de 2 m/s à 6 m/s. Cette observation est confirmée par l'écart-type : ce dernier est d'environ 2 m/s, ce qui est assez faible au vu de l'étendue des valeurs.La vitesse du vent moyenne des observations est de 4.082 m/s, et les valeurs des premier et troisième quartiles (2.870 m/s et 4.970 m/s) nous confirment la faible répartition des données. En effet, 50% des observations ont une vitessse de vent comprise entre ces deux quartiles, et cette concentration des données sur [2;6] est également visible sur le boxplot qui est très écrasé.

Variable *température(t)*

```
t<-meteo.3MIC$t
par(cex=0.7, mai=c(0.8,0.8,0.4,0.1));par(fig=c(0,0.5,0.1,0.9))
hist(t,main="Histogramme des températures",xlab="température (K)",
     ylab="Nombre d' observations",freq=TRUE,breaks=30,cex.main=1.5,cex.lab=1.2)
par(fig=c(0.5,1,0.1,0.9), new=TRUE)
boxplot(t,main="Boxplot des températures",ylab="Température (K)",cex.main=1.5,cex.lab=1.2)
```




```
summary(t);sd(t)
```

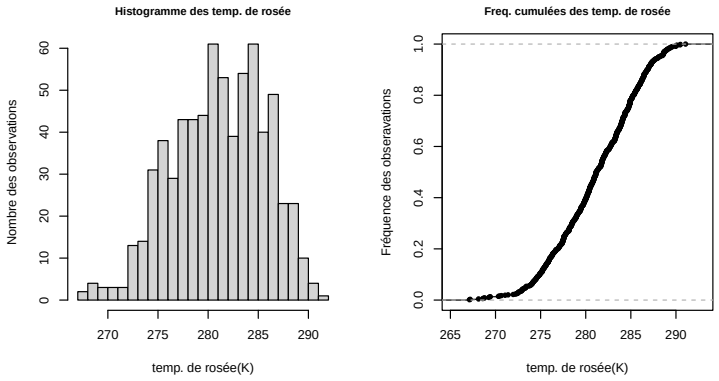
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   269.5   279.9   283.2   283.7   287.6   299.5
```

```
## [1] 4.943628
```

On constate sur l'histogramme que les températures des observations prennent des valeurs assez larges mais globalement comprises entre 275 K et 290 K. Cette observation est confirmée par la valeur du premier quartile (75% des observations ont une température supérieur à 280 K), et également par l'écart-type relativement conséquent (près de 5 K). Egalement, on peut noter sur le boxplot que l'écart inter-quartiles est relativement faible par rapport à la gamme de valeurs considérée, cela illustre la concentration des valeurs autour de la température moyenne (283.7 K). Enfin, il est important de souligner que la température moyenne est proche de la médiane (283.2 K), ce qui traduit la répartition presque symétrique des données autour de cette valeur moyenne.

Variable *température de rosée(td)*

```
td<-meteo.3MIC$td
par(cex=0.7, mai=c(0.8,0.8,0.4,0.1));par(fig=c(0,0.5,0.1,0.9))
hist(td,main="Histogramme des temp. de rosée",xlab="temp. de rosée(K)",
     ylab="Nombre des observations",freq=TRUE,breaks=30,cex.main=0.8,cex.lab=1)
par(fig=c(0.5,1,0.1,0.9), new=TRUE)
plot(ecdf(td),main="Freq. cumulées des temp. de rosée",xlab="temp. de rosée(K)",
     ylab="Fréquence des obseravations",cex.main=0.8,cex.lab=1)
```



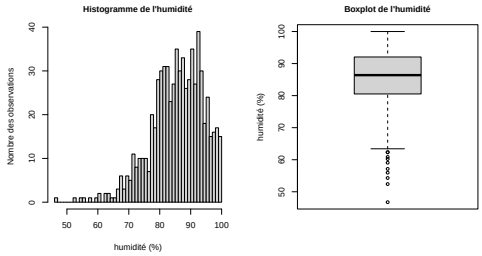

```
summary(td)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   267.1   277.6   281.2   281.1   284.7   291.1
```

On observe sur l'histogramme et sur le graphique des fréquences cumulées que les données sont essentiellement contenues dans la tranche 270 - 290 K.Cette forte concentration se retrouve dans le summary : au moins 75% des observations ont une durée inférieure à 284.7 K.Les résultats ressemblent aux résultats de l'analyse de la température.On a aussi une moyenne à 281.1K qui est trop proche de la valeur de la médiane(281.2K) ce qui traduit la répartition symétrique des données autour de cette valeur moyenne.

Variable *humidité(hu)*

```
hu<-meteo.3MIC$hu
par(cex=0.7, mai=c(0.8,0.8,0.4,0.1));par(fig=c(0,0.5,0.1,0.9))
hist(hu,main="Histogramme de l'humidité",xlab="humidité (%)",
     ylab="Nombre des observations",freq=TRUE,breaks=40,cex.main=1,cex.lab=1)
par(fig=c(0.5,1,0.1,0.9), new=TRUE)
boxplot(hu,main="Boxplot de l'humidité",ylab="humidité (%)",cex.main=1,cex.lab=1)
```




```
summary(hu);sd(hu)
```

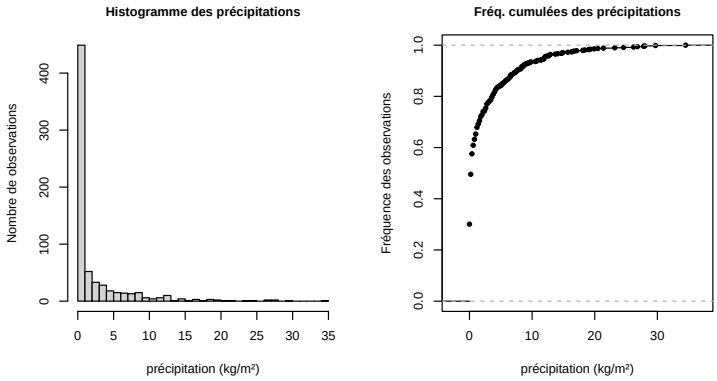
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   46.75   80.53   86.40   85.56   92.06  100.00
```

```
## [1] 8.49062
```

On constate sur l'histogramme que l'humidité dans les observations prennent des valeurs assez larges mais globalement comprises entre 80 et 95 %. Cette observation est confirmée par la valeur du premier quartile (75% des observations ont une humidité supérieur à 80.53 %), et également par l'écart-type relativement conséquent (près de 8.5 %). Egalement, on peut noter sur le boxplot que l'écart inter-quartiles est relativement faible par rapport à la gamme de valeurs considérée, cela illustre la concentration des valeurs autour de l'humidité moyenne(85.56%). Enfin, il est important de souligner que l'humidité moyenne est proche de la médiane (86.40 %), ce qui traduit la répartition symétrique des données autour de cette valeur moyenne.

Variable *précipitation*

```
pre<-meteo.3MIC$precip
par(cex=0.7, mai=c(0.8,0.8,0.4,0.1));par(fig=c(0,0.5,0.1,0.9))
hist(pre, main="Histogramme des précipitations",xlab="précipitation (kg/m²)",
     ylab="Nombre de observations",freq=TRUE,breaks=30,cex.main=1,cex.lab=1)
par(fig=c(0.5,1,0.1,0.9), new=TRUE)
plot(ecdf(pre),main="Fréq. cumulées des précipitations",xlab="précipitation (kg/m²)",
ylab="Fréquence des observations",cex.main=1,cex.lab=1)
```




```
summary(pre);sd(pre)
```

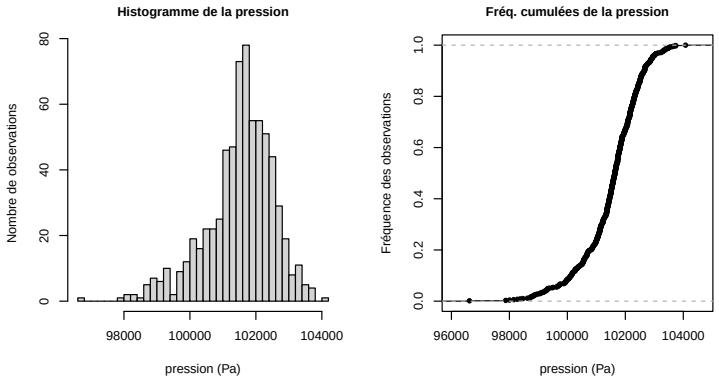
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.400   2.384   2.600  34.500
```

```
## [1] 4.624367
```

On observe sur l'histogramme et sur le graphique des fréquences cumulées que les données sont essentiellement contenues dans la tranche 0 - 5 kg/m². Cette forte concentration se retrouve dans le summary : au moins 75% des observations ont une précipitation inférieure à 2.6 kg/m². Enfin, l'écart-type est de 4.6 kg/m², donc relativement faible par rapport à la plage de valeurs (0 -30 kg/m²) ce qui traduit bien la faible dispersion des données.

Variable *la pression au niveau de la mer*

```
p<-meteo.3MIC$msl_arome
par(cex=0.7, mai=c(0.8,0.8,0.4,0.1));par(fig=c(0,0.5,0.1,0.9))
hist(p, main="Histogramme de la pression",xlab="pression (Pa)",
     ylab="Nombre de observations",freq=TRUE,breaks=30,cex.main=1,cex.lab=1)
par(fig=c(0.5,1,0.1,0.9), new=TRUE)
plot(ecdf(p),main="Fréq. cumulées de la pression",xlab="pression (Pa)",
ylab="Fréquence des observations",cex.main=1,cex.lab=1)
```




```
summary(p)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   96623  101059  101644  101515  102200  104079
```

On observe sur l'histogramme et sur le graphique des fréquences cumulées que les données sont essentiellement contenues dans la tranche 100000 - 103000 Pa.Cette forte concentration se retrouve dans le summary : au moins 75% des prévisions ont une pression à 101059 Pa.

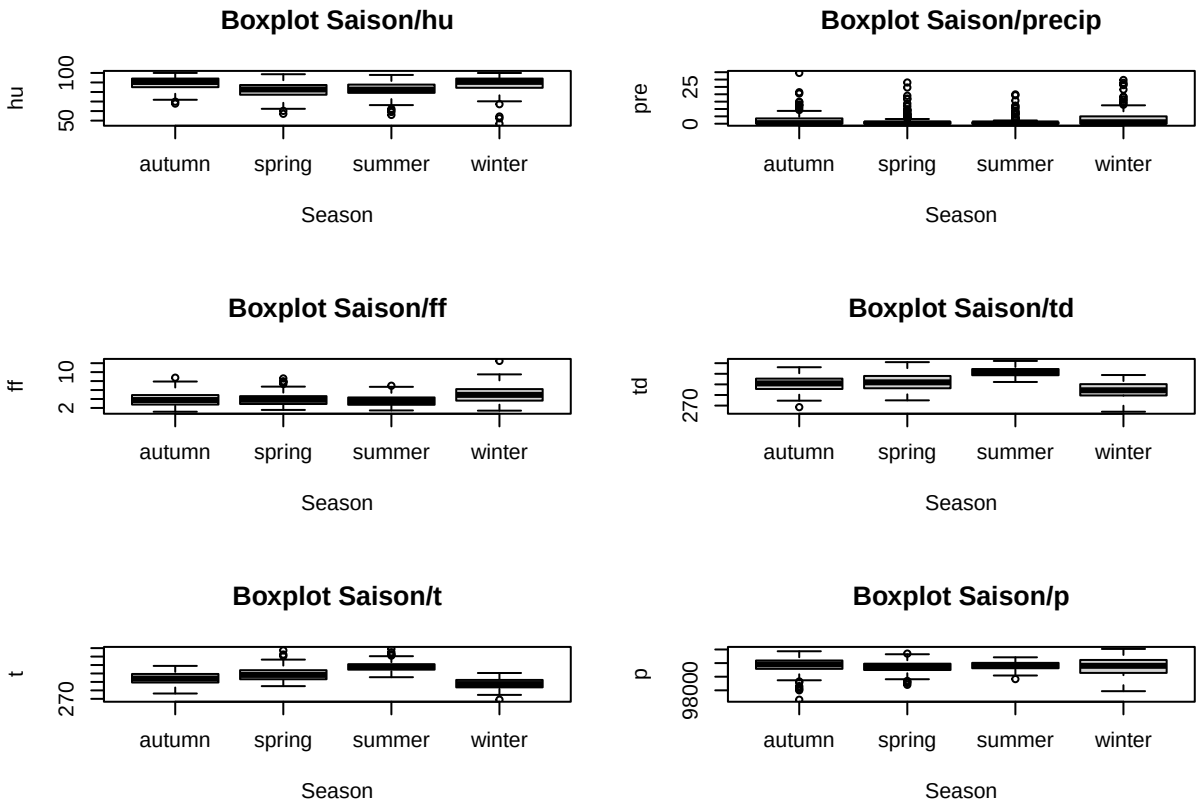
II-2. Analyse bi-dimensionnelle

II-2-a. Corrélation entre les variables quantitatives et qualitatives

Principe d'analyse de la corrélation

On va à présent s'attacher à trouver de potentielles corrélations entre les variables quantitatives et les variables qualitatives. Pour cela, on va montrer la démarche à suivre avec la variable *saison* et toutes les variables quantitatives. Une fois la démarche de recherche de corrélation autour de cette variable terminée, on fera un tableau récapitulatif des différentes analyses sur les corrélations entre les variables. Pour faire apparaître de façon claire la présence de corrélation ou non, on va afficher les boxplots entre la variable *saison* et les variables quantitatives.

```
par(mfrow=c(3,2))
boxplot(hu ~ Season,main="Boxplot Saison/hu")
boxplot(pre ~ Season,main="Boxplot Saison/precip")
boxplot(ff ~ Season,main="Boxplot Saison/ff")
boxplot(td ~ Season,main="Boxplot Saison/td")
boxplot(t ~ Season,main="Boxplot Saison/t")
boxplot(p ~ Season,main="Boxplot Saison/p")
```



On constate ici une forte corrélation entre par exemple la température et la saison de l'obseravation : la période de l'été présente une température élevée alors que celle de l'hiver présente une faible température (et donc il fait plus froid).On remarque une même corrélation avec la température de rosée que celle de la température(plus élevée pendant l'été et diminue pendant l'hiver).On voit aussi qu'il y a une corrélation relative entre la saison et l'humidité(augmente pendant l'automne et l'hiver et diminue pendant le printemps et l'été).De plus, on peut discerner une très faible corrélation avec la variable *ff(vitesse du vent)* dont les valeurs présentent une légère augmentation pendant l'hiver et la variable *p(pression)* . Enfin, la de prise des observations ne semble pas avoir de liaison notable avec la variable *précipitation*.

Tableau récapitulatif des corrélations entre les variables quantitatives et qualitatives

Le tableau suivant synthétise les différentes observations issues des différents boxplots que nous avons analysés selon la méthode présentée avec la variable *saison* précédemment. On associera à chaque binôme de variable les codes '++','+','-' et '/' qui correspondent respectivement à une très forte, notable, peu notable et absence de corrélation.

	température	**temp.de rosée**	**vitesse de vent**	**humidité**	**précip**	**pression**
saison	++	++	-	+	/	-
mois	++	++	+	+	/	-
direction du vent	-	+	++	-	/	/

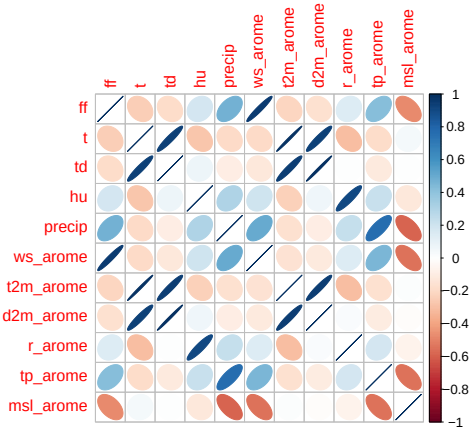
On peut donc conclure que les variables qualitatives *mois* et *saison* présentent une corrélation forte avec le reste des autres variables quantitatives, et principalement avec *température* et *température de rosée*.

II-2-b. Corrélation entre les variables quantitatives

On va désormais procéder à une analyse bidimensionnelle entre les variables quantitatives. Pour cela on peut calculer une valeur qui s'appelle la corrélation qui est comprise entre -1 et 1. Plus cette valeur est proche de -1 ou 1, plus les deux variables concernées sont correlées, et plus cette valeur est proche de 0, moins les variables sont corrélées.

Grâce au diagramme suivant, on voit les différents coefficients de corrélation entre toutes les variables quantitatives. Les ellipses rouges représentent des variables très correlées et inversement proportionnelles, les ellipses bleus représentent des variables très correlées et proportionnelles. Enfin, les ellipses presques invisibles représentent des variables peu correlées.

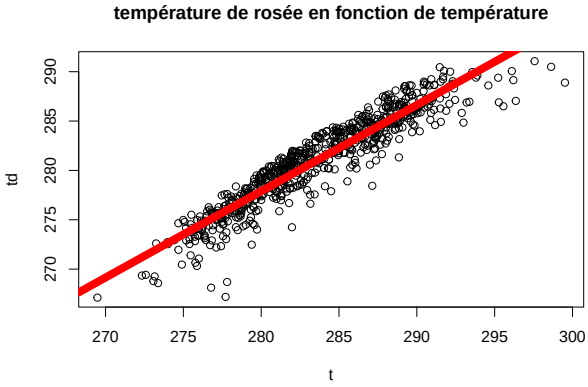
```
x = meteo.3MIC[,1:11]
corrplot::corrplot(cor(x),method='ellipse')
```



On observe ici une très forte corrélation entre les variables *température de rosée* et *température* et aussi entre les variables *msl_arome* et*précip*. Egalement, on peut noter un lien entre *la vitesse du vent* et *la température* et entre *la vitesse du vent* et *l'humidité* et aussi entre*la température* et *l'humidité*. Concrètement, cela signifie que si l'on représente le nuage de points de la variable *température de rosée* en fonction de la variable *température*, on pourra aisément identifier une tendance générale de comportement en raison de leur forte corrélation.

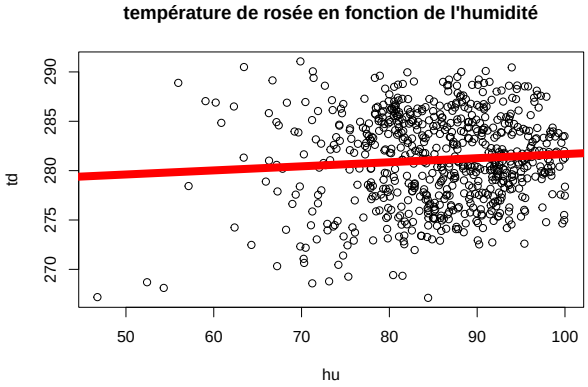
On constate bien sur le graphique ci-dessous que lorsque la température augmente, la température de rosée augmente en raison de la corrélation positive observée dans le diagramme précédent. De manière équivalente, on aurait une droite croissante sur le graphique entre les variables *la vitesse du vent* et *précipitation* en raison du signe du coefficient de corrélation.

```
plot(t,td,main="température de rosée en fonction de température")
mod=lm(td~t,x)
abline(mod,col="red",lwd=8)
```



Remarque : en ce qui concerne les variables peu corrélées, on peut interpréter cela graphiquement par une absence de tendance globale au sein des données. Ainsi, on observe sur le nuage de point ci-dessous entre les variables *température de rosée* et *humidité* (qui présente un coefficient de corrélation proche de 0) qu'il n'y a aucun lien qui semble se dégager : les données sont toutes alignées au sein de la même gamme de valeurs.

```
plot(hu,td,main="température de rosée en fonction de l'humidité")
mod=lm(td~hu,x)
abline(mod,col="red",lwd=8)
```

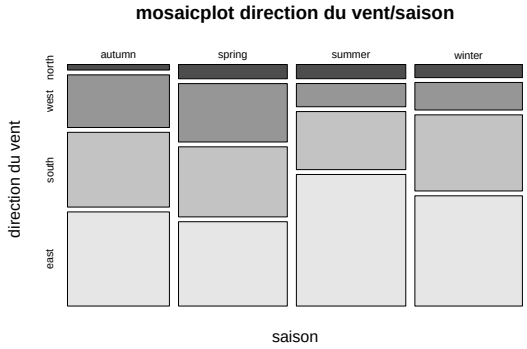


II-2-c. Corrélation entre les variables qualitatives

Principe d'analyse de la corrélation

Pour étudier la corrélation entre les différentes variables qualitatives, on va afficher les **mosaicplots** pour chaque binôme de variables afin de visualiser la répartition des différentes catégories en fonction des autres. Comme nous l'avons fait pour l'analyse de la corrélation entre les variables, nous afficherons seulement le mosaicplot entre *saison* et *direction du vent* pour expliciter notre démarche générale d'analyse.

```
mosaicplot(table(Season, Dir),main="mosaicplot direction du vent/saison",color = TRUE,xlab="saison",
           ylab="direction du vent")
```



On remarque un découpage assez marqué en terme de direction de vent observée pour différentes saisons. Tout d'abord, on constate aisément que le vent qui se dirige vers l'Est a été observé majoritairement pendant l'été (un vent chaud).On remarque que sur toutes les saisons le vent d'Est est prédominante.De plus, le vent du Nord est le même pendant les 4 saisons sauf en automne avec une petite dimunition(ce vent reste très peu observé par rapport aux autres).Enfin, les observations du vent de sud sont presque les mêmes sur les 4 saisons. On peut donc conclure que la corrélation est très forte entre les variables *saison* et *direction du vent*.

Tableau récapitulatif des corrélations entre les variables quantitatives et qualitatives

Le tableau suivant synthétise les différentes observations issues des différents mosaicplots que nous avons analysés selon la méthode présentée précédemment. Les codes utilisés sont les mêmes que dans l'analyse entre les variables qualitatives et quantitatives.

	mois	**saison**	**direction du vent**
mois		++	++
saison	++		+
direction du vent	-	-	

De manière globale, on constate une très forte corrélation entre *mois* et *saison*, ainsi qu'un lien notable entre *mois* et *direction du vent*. En revanche on ne relève aucune corrélation entre les autres binômes constitués par les variables qualitatives.

III. Analyse en Composantes Principales

La matrice de travail :

Avant de commencer, il est important de définir la **matrice de travail**. Cette dernière est la matrice des données que l'on a utilisé tout au long de l'analyse uni-dimensionnelle et bi-dimensionnelle, à l'exception que l'on ne conservera que les variables quantitatives. Etant donné le nombre de variables, on peut se permettre de toutes les inclure dans l'ACP.

Le fonctionnement global de l'ACP :

L'**Analyse en Composante Principale** est un outil en statistique qui permet de dépasser le cadre d'une analyse unidimensionnelle ou bidimensionnelle en permettant une analyse avec un nombre bien plus important de variables quantitatives. Cette analyse permet de dégager des tendances chez les individus

et donc permet de regrouper certaines variables entre elles qui seraient fortement correlées dans ce que l'on appelle des **composantes principales**. On peut ensuite représenter les individus sur des graphiques dont les axes sont ces composantes principales. On comprend dès lors qu'il est nécessaire de comprendre à quelles variables sont liées chaque composante principale pour continuer l'analyse. Ceci est possible grâce au graphe des variables.

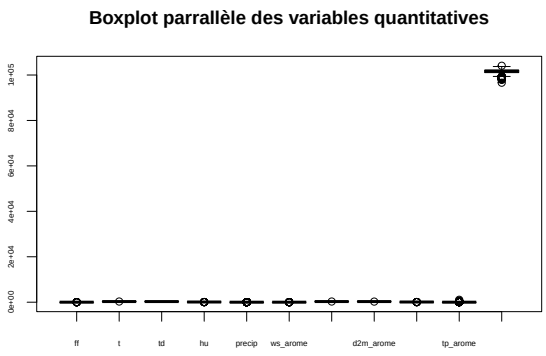
Le formalisme mathématique de l'ACP :

Nous allons à présent détailler succintement le fonctionnement de l'ACP. Tout d'abord, la matrice de travail citée dans le paragraphe précédent se nommera la matrice T. Il nous faut également introduire ce que l'on appelle une **métrique**, c'est-à-dire un moyen de calculer la "distance" qui sépare deux observations (quantifier leur degré de différence). Cette métrique se répresente sous la forme d'une matrice M qui interviendra plus tard dans les calculs (on prendra ici pour M la matrice identité). Ensuite, nous devons choisir une **matrice des pondérations** (notée W) qui va permettre de distribuer l'impact qu'aura chaque observation dans l'ACP. Or, il n'y a ici pas de raison particulière pour qu'une observation compte plus qu'une autre, ainsi on considèrera la matrice $W = \frac{1}{n} I_n$ avec n le nombre d'observation du jeu de données. Enfin, on construira la matrice $\Gamma = T^T W T$. Pour la suite, notre intérêt principal sera de vouloir quantifier la dispersion des individus. On introduit pour cela la notion d'**inertie globale** qui se calcule de la façon suivante : $I_\Omega = Tr(\Gamma M)$, avec Ω l'individu moyen et centré. On peut dès lors parler d'**inertie partielle** (ou **axiale**), qui représente la part d'inertie apportée par chacune des variables de base du jeu de données. Il y a deux façons de voir ces inerties axiales :
– On peut tout d'abord considérer la base classique et cela revient donc à diviser chaque élément diagonal de la matrice Γ.
– On peut sinon trouver une base plus intéressante : celle des vecteurs propres et c'est de cette nouvelle base que l'ACP va résulter. Les vecteurs propres a_j de Γ sont appelés les **axes principaux** (ce sont également les facteurs principaux car on utilise $M = I_n$), et les vecteurs Ta_j sont les **composantes principales**.

Choix de l'ACP :

Il est ensuite intéressant de regarder quelles sont les tendances globales du groupe soumis à l'analyse. On peut faire la distinction entre 2 types d'ACP : la centrée et la centrée-reduite. La grande différence réside dans le fait que la centrée s'utilise dans des cas où toutes les variables ont le même ordre de grandeur (comme par exemple, si toutes les variables sont des notes entre 0 et 20). On note ici que ce n'est pas le cas comme on peut le voir dans ce **boxplot** :

```
x = meteo.3MIC[,1:11]
boxplot(x, main="Boxplot parrallèle des variables quantitatives",cex.axis=0.5)
```



On constate clairement sur le graphique ci-dessus que la variable *msl_arome* écrase les autres en raison de ses valeurs trop importantes. Pour la suite de l'étude, on procèdera donc à une ACP centrée-réduite avec l'aide

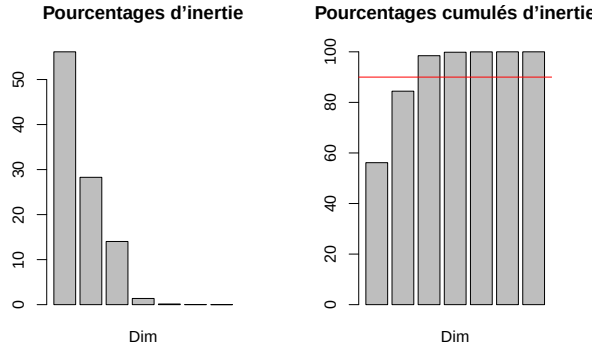
de la librairie **FactoMineR**. Il nous faut donc établir la moyenne et l'écart-type puis, pour chaque individu, soustraire la moyenne de la variable concernée (pour centrer) et diviser par l'écart-type (pour réduire) afin d'obtenir notre matrice de travail.

Afin de déterminer le nombre de composantes principales à conserver pour notre analyse, nous appueirons sur les graphiques suivants qui représentent les pourcentages d'intertie (i.e. la part d'information) de chacune des composantes :

```
res.acp <- PCA(x,scale.unit=TRUE,ncp=4,quali.sup=c(1,5,6,10),graph=FALSE)
par(mfrow=c(1,2))
summary(res.acp$eig)
```

```
##    eigenvalue       percentage of variance cumulative percentage of variance
##  Min.   :0.000218   Min.   : 0.00311       Min.   : 56.15
##  1st Qu.:0.005632   1st Qu.: 0.08046       1st Qu.: 91.45
##  Median :0.096224   Median : 1.37463       Median : 99.84
##  Mean   :1.000000   Mean   :14.28571       Mean   : 91.27
##  3rd Qu.:1.480805   3rd Qu.:21.15436       3rd Qu.: 99.99
##  Max.   :3.930684   Max.   :56.15263       Max.   :100.00
```

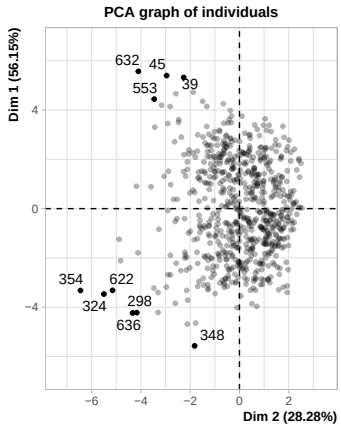
```
barplot(res.acp$eig[,"percentage of variance"],names.arg=paste("Dim",sep="."),
main="Pourcentages d'inertie")
barplot(res.acp$eig[,"cumulative percentage of variance"],names.arg=paste("Dim",sep="."),
main="Pourcentages cumulés d'inertie")
abline(90,0,col="red")
```



La droite tracée en rouge nous permet de déterminer combien de composantes principales nous devons garder en fonction d'un pourcentage d'information souhaité qui est fixé arbitrairement. Ici on va considérer qu'il est pertinent de conserver 90% d'information pour la suite de notre analyse, par conséquent il est nécessaire de garder 3 composantes principales. En effet, la droite rouge intersecte la 3ème composante ce qui signifie qu'au moins 90% de l'information est contenue dans les 3 premières composantes principales.En plus, on peut remarquer d'après le graphique des porcentages d'inertie que la totalité de l'information se trouve dans les trois premières composantes(on remarque que dans le reste des composantes l'information est presque nulle donc on peut les négliger).

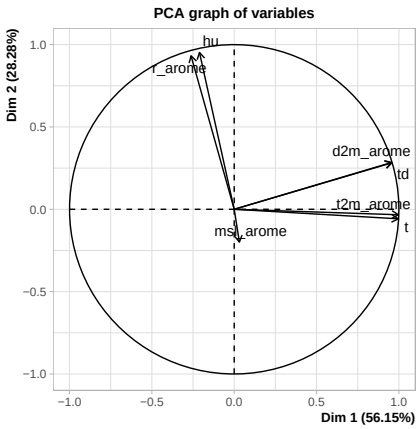
Nous pouvons maintenant afficher le nuage de points avec comme axes les deux premières composantes principales grâce aux commandes suivantes :

```
plot(res.acp,choix="ind",invisible="quali",select = "contrib 10",axes=c(2,1))
```



On remarque que le nuage de points est très concentré autour de l'origine du graphique. Toutefois, ce dernier n'est pas très intelligible sans informations sur la signification de chaque axe (qui sont ici les 2 premières composantes principales). Pour remédier à cela, on décide d'afficher le graphe des variables qui nous permettra de mieux comprendre la composition des deux premiers axes :

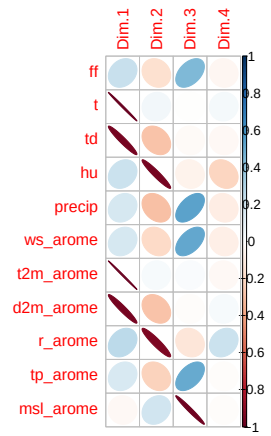
```
plot(res.acp,choix="varcor",axes=c(1,2))
```



Ce graphique nous permet de savoir comment sont correlées les variables de base et les composantes principales. Chacune des variables de base est représentée par une flèche et plus cette dernière est proche d'un axe, plus la variable concernée influe sur la composante principale associée à l'axe. De surcroît, plus la flèche se rapproche du bord du cercle, plus le poids de cette variable est important. Par exemple, on voit ici que si une observation (représenté par un point) était situé sur la droite de l'axe horizontal, elle aurait alors une température et une température de rosée élevée.En ce qui concerne l'axe vertical, une observation situé en haut aura tendance à avoir une humidité plus élevée que les autres.

Pour éviter de devoir faire cette analyse sur tous les graphiques de variable, on utilise les diagramme des corrélations suivant qui établit le lien entre nos variables de base et nos composantes principales :

```
corrplot::corrplot(cor(x,-res.acp$ind$coord),method='ellipse')
```



On retrouve donc bien ce qu'on avait analysé sur le graphique des variables des 2 premières composantes principales. En effet, pour la première composante principale(représente 56% de l'information) est très correlée avec les variables t,td,$t2m_arome$et$d2m_arome$.Ceci est illustré par l'orientation horizontale de leurs flèches dans le graphique des variables.Cette composante est peu correlée avec les variables hu et r_arome qui sont quant à elles majoritairement liées à la deuxième composante principale (qui représente 28% de l'information).On peut remarquer ce résultat dans le graphique des variables avec les flèches qui sont proche de l'axe vertical. Enfin, on remarque une corrélation relative de la première et deuxième composante avec les variables ff,$precip$,ws_arome et tp_arome et une coorélation presque nulle avec la variable msl_arome.

Pour la 3ème composante principale (16 % de l'information), on relève une forte corrélation avec la variable msl_arome de maniére négative et une corrélation de manière positive avec les variables ff,$precip$,ws_arome et tp_arome (plus importante que pour les deux premières composantes principales).

On remarque que la 4ème dimension ne possède aucune corrélation avec toutes les variables. C'est à dire que cette dernière ne représente pas une composante principale. Donc c'est un signe que les trois composantes principales ne manquent pas d'information.