Projet en Statistique Descriptive

Thanh Chung NGUYEN, Minh Duy NGUYEN, Hieu PHAM

I. Description du jeu de données

I.1. Prise en charge des données sous R

L'objectif de ce projet est d'effectuer une analyse statistique descriptive des données environnementales qui entrent en jeu dans l'étude de l'effet de serre enregistrées par Centre européen pour les prévisions météorologiques à moyen terme et de l'Agence américaine d'observation océanique et atmosphérique à l'aide du langage R.

Tout d'abord, nous démarrons par implémenter les libraries utiles pour la description et l'analyse des données:

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

Nous importons maintenant la base de données **EffetSerre**. Puis nous cherchons à voir s'il y a des valeurs manquantes (i.e les valeurs NA) dans notre jeu de données.

```
EffetSerre=read.table("/Users/corp_sysops/Desktop/Statistique/EffetSerre.txt",header=TRUE)
any(is.na(EffetSerre))
```

```
## [1] FALSE
```

Nous pouvons conclure qu'il n'y a pas de manque dans notre table.

I.2. Description de l'ensemble du jeu de données

Il s'agit d'un jeu de données environnementales qui contiennent 349 observations décrites sur la base de 12 attributs d'évaluation, qui comprennent :

- *CO2* (en molfrac ppm) : la concentration de CO2 dans l'air.
- *année*, *mois* : la date où on effectue et enregistre cette observation.
- *t2Reyjavik, t2Oslo, t2Paris, t2NewYork, t2Tunis, t2Alger, t2Beyrouth, t2Atlan, t2Dakar* (en °C) : la température enregistrée à 2 mètres du sol en 9 lieux géographiques respectivement.

En regardant la table d'observation, les données sont enregistrées sous forme de chiffres. Néanmoins, pour les données des colonnes *année* et *mois*, on se rend compte que l'objectif de ces deux colonnes de données est juste de servir de base pour observer l'évolution de la concentration en co2 et de la température ambiante. Quoi qu'elles s'agisse bien de nombres, leur nature est bien comme celle de variable catégorielle. Ainsi, on effectue le changement de type pour ces 2 colonnes :

```
EffetSerre$annee <- as.factor(EffetSerre$annee)
EffetSerre$mois <- as.factor(EffetSerre$mois)
```

En rappellant les natures des variables statistiques, nous en avons 4 qui sont qualitative nominale, qualitative ordinale, quantitative discrète et quantitative continue. Nous consultons maintenant le type de chaque variable:

```
str(EffetSerre)
```

```
## 'data.frame':    349 obs. of  12 variables:
##  $ annee      : Factor w/ 29 levels "1982","1983",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ mois       : Factor w/ 12 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ co2        : num  341 342 343 344 345 ...
##  $ t2Reykjavik: num  -5.449 0.253 -0.669 3.099 5.307 ...
##  $ t2Oslo     : num  -9.32 -2.92 3 7.71 11.73 ...
##  $ t2Paris    : num  4.38 6.38 8.91 12.22 16.39 ...
##  $ t2NewYork  : num  -5.812 -0.172 2.398 6.78 14.878 ...
##  $ t2Tunis    : num  15.7 14.7 15.4 17.5 21.4 ...
##  $ t2Alger    : num  13.7 13.1 14.9 16.7 20.1 ...
##  $ t2Beyrouth : num  15 13.1 14.9 19.1 21.3 ...
##  $ t2Atlan    : num  20.8 20.5 21 20.7 22.7 ...
##  $ t2Dakar    : num  24.6 22.3 23.2 24 24.3 ...
```

Nous voyons bien que les variables de type 'factor' sont catégorielles (ou bien qualitatives) et celles de type 'num' sont numériques (ou bien quantitatives). D'après les données, nous pouvons eventuellement les catégoriser comme suivant :

- Qualitative ordinale : *mois*, *année*
- Quantitative continue : *co2, t2Reyjavik, t2Oslo, t2Paris, t2NewYork, t2Tunis, t2Alger, t2Beyrouth, t2Atlan, t2Dakar*

On note que il y a deux observation de le mois janvier 2000. On observe que la première observation avec le nom de ligne "217", les données sont irréalistes car la concentration de co2 est de 0 et la température dans tous les endroits considérés est de 50 . Par conséquent, on utiliserait l'observation avec le nom de ligne "2171" pour les données de janvier 2000 et supprimerait la ligne nommée "217" dans l'ensemble de données. On obtient alors un table "data" que on va analyser.

```
data <- EffetSerre[-217,]
data$annee = factor(data$annee)
data$mois = factor(data$mois)
```

II. Analyse uni- et bi-dimensionnelle

Dans cette jeu de données, pour les variables qualitatives "annee" et "mois", on observe que le nombre de modalité de ces variables sont toute égales (12 individus chaque modalité de "annee" et 29 individus chaque modalité de "mois"). Par conséquent, il y aucune intérêt d'étudier statistique unidimensionnelle des variables qualitatives et d'étudier statistique bidimensionnelle entre les variables "annee" et "mois".

II-1. Analyse unidimensionelle

Dans le cadre de l'analyse unidimensionnelle, nous opterons pour différentes représentations en fonction du type de variable:

Pour les variables qualitatives nominales, nous utiliserons un diagramme circulaire pour mettre en évidence les différentes modalités proportionnelles à leurs fréquences respectives.

Pour les variables qualitatives ordinales, nous choisirons un diagramme en bâtons des fréquences (cumulées ou non) pour préserver l'ordre des données.

Pour les variables quantitatives continues, nous opterons pour un histogramme ou une courbe des fréquences cumulées pour refléter la nature continue des données.

Enfin, pour les variables quantitatives discrètes, nous utiliserons les mêmes représentations que pour les variables quantitatives continues, malgré la nature discrète de ces données, en raison de leur étendue de

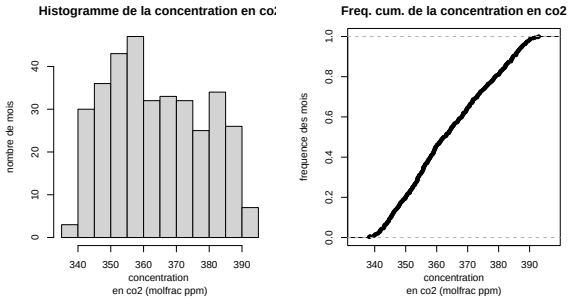
valeurs. Cela permettra de conserver une présentation claire et exploitable des données.

Dans cette jeu de données, pour les variables qualitatives “annee” et “mois”, on observe que le nombre de modalité de ces variables sont toute égales (12 individus chaque modalité de “annee” et 29 individus chaque modalité de “mois”). Par conséquent, il y aucune intérêt d’étudier statistique unidimensionnelle des variables qualitatives et d’étudier statistique bidimensionnelle entre les variables “annee” et “mois”.

Variables quantitatives

```
co2 <- data$co2
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(co2, main = "Histogramme de la concentration en co2", xlab="concentration
en co2 (molfrac ppm)", ylab = "nombre de mois")
par(fig=c(0.5,1,0.1,0.9), new=TRUE)
plot(ecdf(co2), main ="Freq. cum. de la concentration en co2", xlab="concentration
en co2 (molfrac ppm)", ylab="frequence des mois")
```

Variable “co2”




```
summary(co2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   338.2   352.3   362.5   364.0   375.9   393.0
```

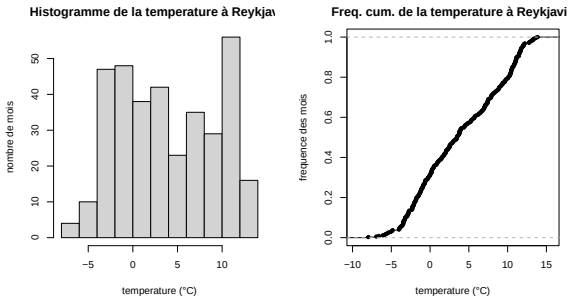
```
sd(co2)
```

```
## [1] 14.3808
```

On observe que la moyenne et la médiane sont proches (364 et 362.5). De plus, le graphique de la fréquences cumulées est presque un droite linéaire. On peut alors affirmer que cette répartition est assez homogène sur l’ensemble de la plage de valeurs bien que les fréquence à les bords soient plus faibles.

```
t2Reykjavik <- data$t2Reykjavik
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(t2Reykjavik, main = "Histogramme de la temperature à Reykjavik", xlab="temperature (°C)"
, ylab = "nombre de mois")
par(fig=c(0.5,1,0.1,0.9), new=TRUE)
plot(ecdf(t2Reykjavik), main ="Freq. cum. de la temperature à Reykjavik", xlab="temperature (°C)"
, ylab="frequence des mois")
```

Variable t2Reykjavik




```
summary(t2Reykjavik)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -7.9633 -0.9588  3.4709  3.8706  8.9377 13.8685
```

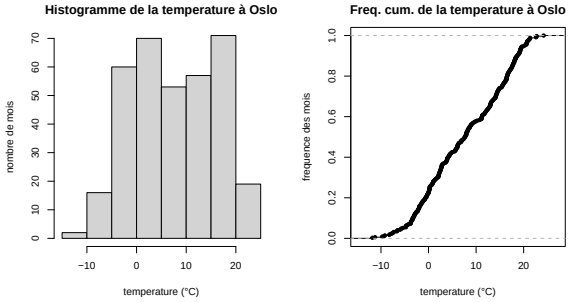
```
sd(t2Reykjavik)
```

```
## [1] 5.477197
```

On observe que la répartition des cette variable est quasi-homogène, la moyenne est près de la médianne (3,8706 et 3,4709). Néanmoins, il y a une présence plus légère sur les bords de la tranche de valeurs.Enfin, le graphique des fréquences cumulées, qui est presque une droite linéaire, nous permet d'affirmer la répartition assez homogène.

```
t2Oslo <- data$t2Oslo
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(t2Oslo, main = "Histogramme de la temperature à Oslo", xlab="temperature (°C)"
, ylab = "nombre de mois")
par(fig=c(0.5,1,0.1,0.9), new=TRUE)
plot(ecdf(t2Oslo), main ="Freq. cum. de la temperature à Oslo", xlab="temperature (°C)"
, ylab="frequence des mois")
```

Variable t2Oslo




```
summary(t2Oslo)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -11.7992   0.2801   7.7072   7.6718  15.4679  24.2041
```

```
sd(t2Oslo)
```

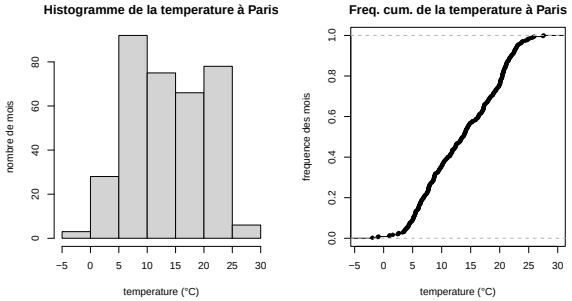
```
## [1] 8.465369
```

L'histogramme montre une distribution apparemment homogène des données sur l'ensemble de la plage de valeurs, bien qu'une présence légèrement plus faible soit notée aux extrémités. Cette observation est corroborée par le **summary** : 50 % des mois ont une température inférieure à 7,7072 (très proche de la moyenne à 7,6718) et 50 % ont une température supérieure. Enfin, le graphique des fréquences cumulées

indique clairement que les données sont presque uniformément réparties : une droite linéaire est en effet observée.

```
t2Paris <- data$t2Paris
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(t2Paris, main = "Histogramme de la temperature à Paris", xlab="temperature (°C)"
, ylab = "nombre de mois")
par(fig=c(0.5,1,0.1,0.9), new=TRUE)
plot(ecdf(t2Paris), main ="Freq. cum. de la temperature à Paris", xlab="temperature (°C)"
, ylab="frequence des mois")
```

Variable t2 Paris




```
summary(t2Paris)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -1.937   7.801  13.728  13.622  19.877  27.572
```

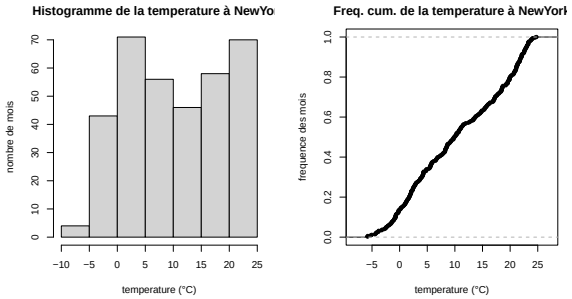
```
sd(t2Paris)
```

```
## [1] 6.567873
```

L'histogramme met en évidence que les données semblent être uniformément réparties sur toute la plage de valeurs, bien qu'une légère présence moindre soit remarquée sur les bords. Cette observation est corroborée par le "summary" : la différence entre la moyenne et la médiane est presque nulle(13,622 et 13,728) . Enfin, le graphique des fréquences cumulées confirme que les données sont presque uniformément étalées : une droite linéaire est effectivement observée.

```
t2NewYork <- data$t2NewYork
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(t2NewYork, main = "Histogramme de la temperature à NewYork", xlab="temperature (°C)"
, ylab = "nombre de mois")
par(fig=c(0.5,1,0.1,0.9), new=TRUE)
plot(ecdf(t2NewYork), main ="Freq. cum. de la temperature à NewYork", xlab="temperature (°C)"
, ylab="frequence des mois")
```

Variable t2NewYork




```
summary(t2NewYork)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -5.812   2.605   9.959  10.490  18.631  24.761
```

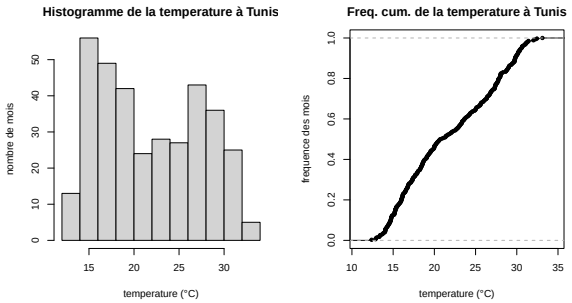
```
sd(t2NewYork)
```

```
## [1] 8.623901
```

En examinant l'histogramme, on constate que les données semblent être uniformément réparties sur l'ensemble de la plage de valeurs, bien qu'une notable diminution soit observée à minimum valeurs. Cette observation est étayée par le "summary" : la différence entre la moyenne et la médiane est presque négligeable(10,490 et 9,959). Enfin, le graphique des fréquences cumulées confirme que les données sont pratiquement uniformément réparties : une droite linéaire est en effet visible.

```
t2Tunis <- data$t2Tunis
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(t2Tunis, main = "Histogramme de la temperature à Tunis", xlab="temperature (°C)"
, ylab = "nombre de mois")
par(fig=c(0.5,1,0.1,0.9), new=TRUE)
plot(ecdf(t2Tunis), main ="Freq. cum. de la temperature à Tunis", xlab="temperature (°C)"
, ylab="frequence des mois")
```

Variable t2Tunis




```
summary(t2Tunis)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.39   16.59   20.82   21.83   27.32   33.11
```

```
sd(t2Tunis)
```

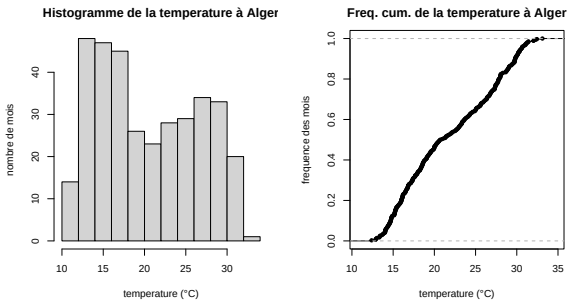
```
## [1] 5.687936
```

L'histogramme montre que les données sont uniformément réparties sur l'ensemble de la plage de valeurs, bien qu'il y ait une légère diminution aux extrémités. Cette observation est confirmé par le "summary" : la

moyenne et la médiane sont presque identiques(21,83 et 20,82). Enfin, le graphique des fréquences cumulées confirme que les données sont presque uniformément étalées : on observe une ligne droite.

```
t2Alger <- data$t2Alger
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(t2Alger, main = "Histogramme de la temperature à Alger", xlab="temperature (°C)"
, ylab = "nombre de mois")
par(fig=c(0.5,1,0.1,0.9), new=TRUE)
plot(ecdf(t2Tunis), main ="Freq. cum. de la temperature à Alger", xlab="temperature (°C)"
, ylab="frequence des mois")
```

Variable t2Alger




```
summary(t2Alger)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.31   15.21   19.14   20.43   26.04   32.10
```

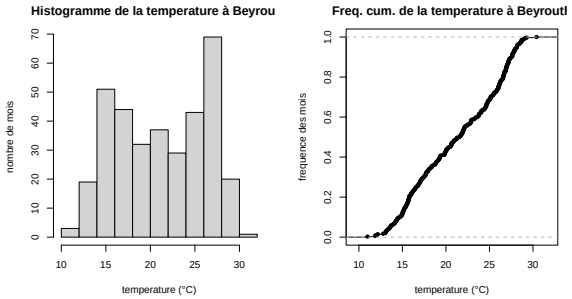
```
sd(t2Alger)
```

```
## [1] 6.07492
```

D'après l'histogramme, les données semblent être uniformément distribuées sur toute la plage de valeurs, même si l'on peut noter une notable diminution à minimum valeurs et une très forte diminution à maximum valeurs. Cette observation est soutenue par le "summary" qui montre une différence très faible entre la moyenne et la médiane(20,43 et 19,14). En outre, le graphique des fréquences cumulées confirme que les données sont pratiquement uniformément étalées, ce qui se traduit par une ligne droite visible.

```
t2Beyrouth <- data$t2Beyrouth
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(t2Beyrouth, main = "Histogramme de la temperature à Beyrouth", xlab="temperature (°C)"
, ylab = "nombre de mois" )
par(fig=c(0.5,1,0.1,0.9), new=TRUE)
plot(ecdf(t2Beyrouth), main ="Freq. cum. de la temperature à Beyrouth", xlab="temperature (°C)"
, ylab="frequence des mois")
```

Variable t2Beyrouth




```
summary(t2Beyrouth)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.99   16.62   21.56   21.24   26.03   30.40
```

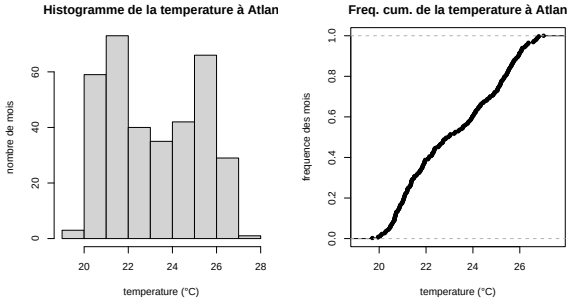
```
sd(t2Beyrouth)
```

```
## [1] 4.953816
```

On remarque sur l'histogramme que les données sont réparties uniformément sur l'ensemble de la plage de valeurs, même si l'on peut constater une notablenotable diminution sur les bords. Cette observation est soutenue par le "summary", qui indique que la différence entre la moyenne et la médiane est quasiment inexistante (21,24 et 21,56). Enfin, le graphique des fréquences cumulées confirme que les données sont réparties de manière quasi-uniforme, ce qui se reflète par une ligne droite.

```
t2Atlan <- data$t2Atlan
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(t2Atlan, main = "Histogramme de la temperature à Atlan", xlab="temperature (°C)"
, ylab = "nombre de mois")
par(fig=c(0.5,1,0.1,0.9), new=TRUE)
plot(ecdf(t2Atlan), main ="Freq. cum. de la temperature à Atlan", xlab="temperature (°C)"
, ylab="frequence des mois")
```

Variable t2Atlan




```
summary(t2Atlan)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   19.71   21.28   22.95   23.18   25.13   27.02
```

```
sd(t2Atlan)
```

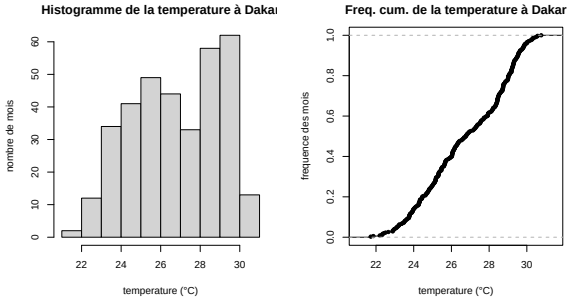
```
## [1] 2.024804
```

Les données montrent que la moyenne et la médiane sont proches (23,18 et 22,9595). De plus, le graphique de la fréquences cumulées forme presque une ligne droite. Nous pouvons donc conclure que la répartition des

données est relativement uniforme sur toute la plage de valeurs, bien que les fréquences aux extrémités soient très faibles.

```
t2Dakar <- data$t2Dakar
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(t2Dakar, main = "Histogramme de la temperature à Dakar", xlab="temperature (°C)"
, ylab = "nombre de mois")
par(fig=c(0.5,1,0.1,0.9), new=TRUE)
plot(ecdf(t2Dakar), main ="Freq. cum. de la temperature à Dakar", xlab="temperature (°C)"
, ylab="frequence des mois")
```

Variable t2Dakar




```
summary(t2Dakar)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   21.72   24.96   26.77   26.76   28.75   30.76
```

```
sd(t2Dakar)
```

```
## [1] 2.235862
```

Les valeurs de la moyenne et de la médiane sont presque identiques (respectivement 26.76 et 26,77). En outre, le graphique de la fréquence cumulée est presque une ligne droite. Cela permet de conclure que les données sont relativement uniformément réparties sur toute la plage de valeurs, bien que les fréquences aux extrémités soient plus faibles.

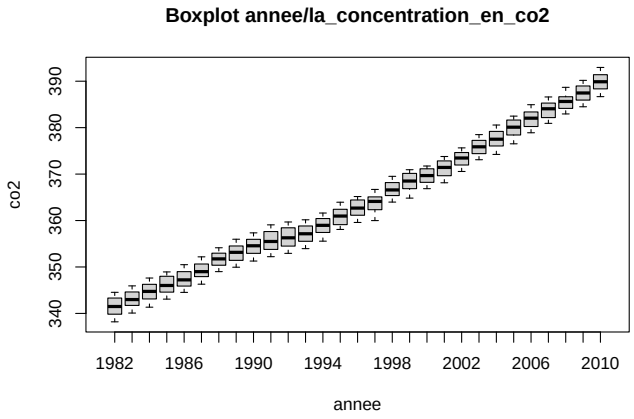
II-2. Analyse bidimensionelle

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

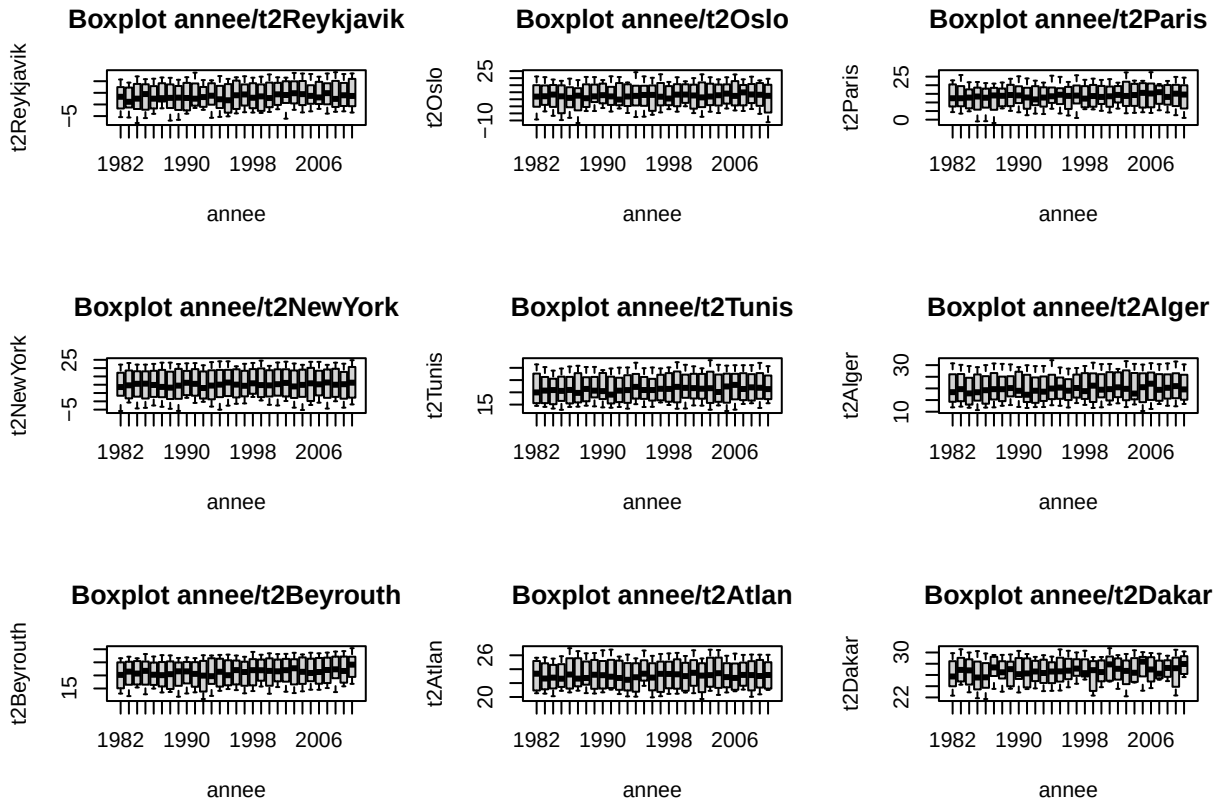
Principe d'analyse de la corrélation

Nous allons désormais nous concentrer sur l'identification de potentielles corrélations entre les variables quantitatives et qualitatives. Dans cette optique, nous présenterons la démarche à suivre en utilisant la variable "annee",la variable "mois" et toutes les variables quantitatives. Afin de mettre en évidence la présence ou l'absence de corrélations de manière claire, nous allons utiliser des diagrammes en boîte (boxplots) pour représenter la relation entre la variable "annee" et les variables quantitatives, et entre la variable "mois" et les variables quantitatives.

```
annee <- data$annee
boxplot( co2 ~ annee,main="Boxplot annee/la_concentration_en_co2")
```

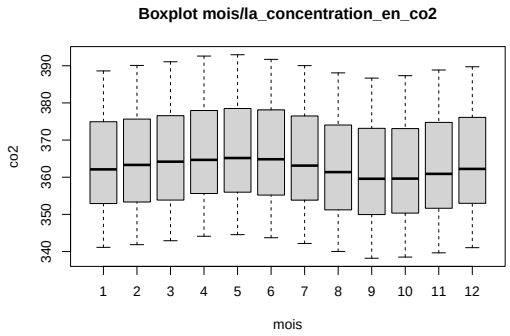



```
par(mfrow=c(3,3))
boxplot( t2Reykjavik ~ annee,main="Boxplot annee/t2Reykjavik")
boxplot(t2Oslo ~ annee,main="Boxplot annee/t2Oslo")
boxplot(t2Paris ~ annee,main="Boxplot annee/t2Paris")
boxplot(t2NewYork ~ annee,main="Boxplot annee/t2NewYork")
boxplot(t2Tunis ~ annee,main="Boxplot annee/t2Tunis")
boxplot(t2Alger ~ annee,main="Boxplot annee/t2Alger")
boxplot(t2Beyrouth ~ annee,main="Boxplot annee/t2Beyrouth")
boxplot(t2Atlan ~ annee,main="Boxplot annee/t2Atlan")
boxplot(t2Dakar ~ annee,main="Boxplot annee/t2Dakar")
```

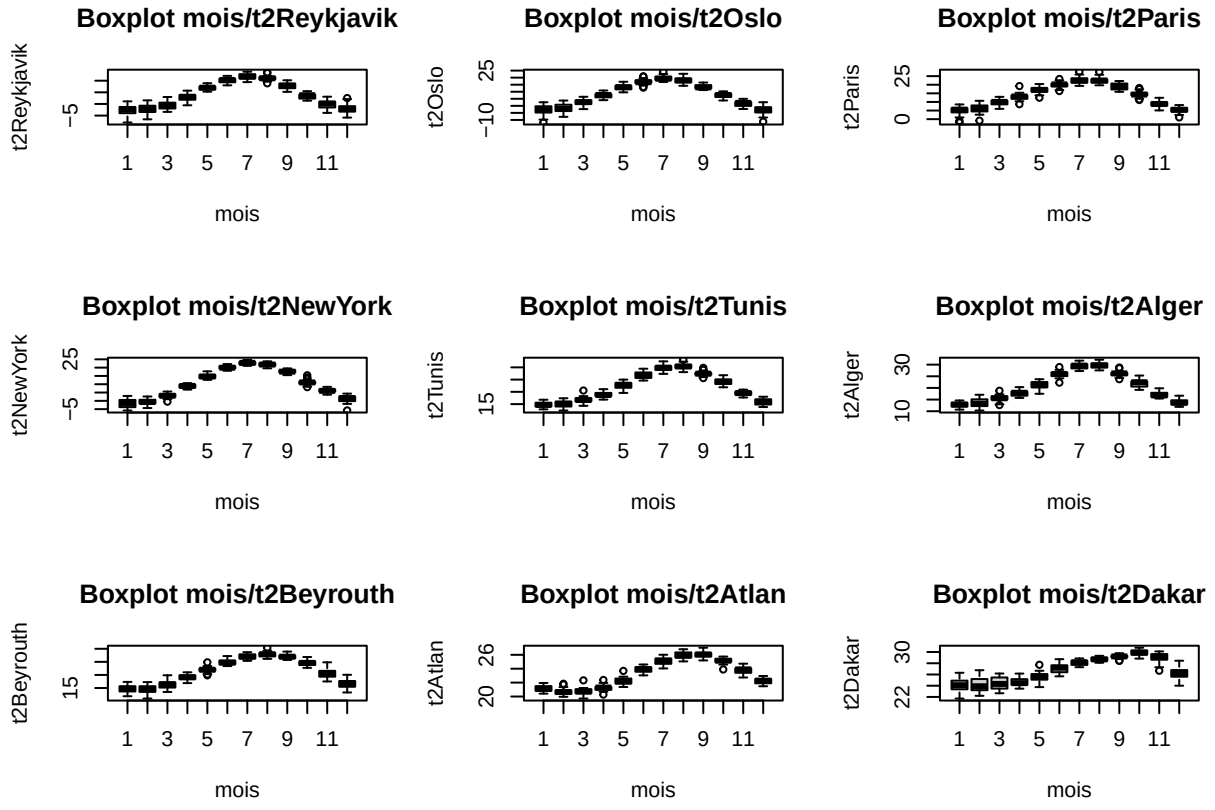


On note ici une forte corrélation entre la concentration en co2 et annee : au cours des année, la concentration en de co2 de ces derniers a connu une forte croissance. De plus,Il n'y a pas de corrélation significative entre annee et les température aux lieux considérées.

```
mois <- data$mois
boxplot( co2 ~ mois,main="Boxplot mois/la_concentration_en_co2")
```




```
par(mfrow=c(3,3))
boxplot( t2Reykjavik ~ mois,main="Boxplot mois/t2Reykjavik")
boxplot(t2Oslo ~ mois,main="Boxplot mois/t2Oslo")
boxplot(t2Paris ~ mois,main="Boxplot mois/t2Paris")
boxplot(t2NewYork ~ mois,main="Boxplot mois/t2NewYork")
boxplot(t2Tunis ~ mois,main="Boxplot mois/t2Tunis")
boxplot(t2Alger ~ mois,main="Boxplot mois/t2Alger")
boxplot(t2Beyrouth ~ mois,main="Boxplot mois/t2Beyrouth")
boxplot(t2Atlan ~ mois,main="Boxplot mois/t2Atlan")
boxplot(t2Dakar ~ mois,main="Boxplot mois/t2Dakar")
```



On remarque que il ne semble pas y avoir de corrélation significative entre les mois et la concentration en co2. Cependant, il est Il est à noter que la concentration de CO2 atteint son pic entre les mois d'avril et de juin, même si la différence est assez faible.D'un autre côté; la corrélation entre les mois et les température est visible. A Reykjavik, Oslo, Paris et New York, lla température a tendance à augmenter progressivement à

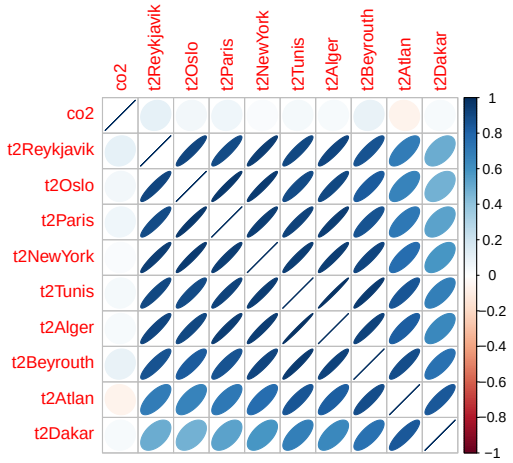
partir du mois de janvier et atteint son pic en juillet, puis elle diminue progressivement jusqu'au mois de décembre.Il y a une même tendance à Tunis, Alger et Beyrouth, sauf que la température atteint son pic en août.C'est également une tendance à Atlan, à l'exception du fait que l'augmentation démarre à partir du mois de mars, le pic est en août et la diminution est de août à mars de prochaine année.

II-2-b. Corrélation entre les variables qualitatives

Nous allons maintenant passer à l'analyse bidimensionnelle des variables quantitatives. Pour ce faire, nous pouvons calculer une mesure appelée "corrélation", qui varie entre -1 et 1. Plus cette mesure se rapproche de -1 ou 1, plus les deux variables en question sont corrélées, tandis qu'une valeur proche de 0 indique une faible corrélation.

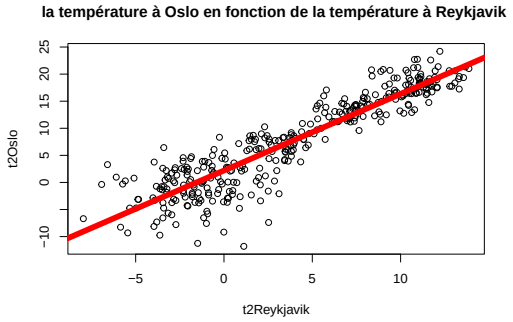
Le diagramme ci-dessous montre les différents coefficients de corrélation entre les variables quantitatives. Les ellipses rouges représentent des variables inversement proportionnelles fortement corrélées, les ellipses bleues représentent des variables proportionnelles fortement corrélées, tandis que les ellipses presque invisibles représentent des variables peu corrélées.

```
x = data[c(3:12)]
corrplot(cor(x),method='ellipse')
```



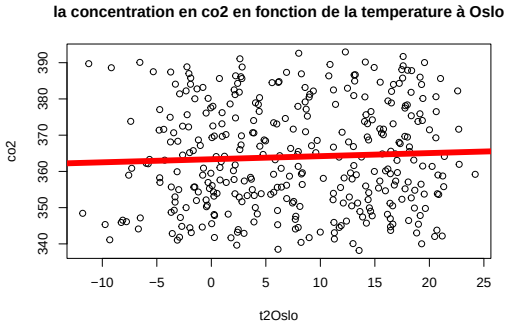
On note que il n'y a aucune corrélation entre la concentration en co2 et les températures. D'autre côté, on observe une forte corrélation entre les 7 variables: "t2Reykjavik","t2Oslo","t2Paris", "t2NewYork", "t2Tunis", "t2Alger" et "t2Beyrouth". Les corrélations entre "t2Atlan" et "t2Dakar" avec autre températures sont faibles.En d'autres termes, si l'on trace le nuage de points représentant la variable "t2Reykjavik" par rapport à la variable "t2Oslo", il sera facile d'observer une tendance générale de comportement en raison de leur forte corrélation.

```
plot(t2Reykjavik,t2Oslo,main="la température à Oslo en fonction de la température à Reykjavik ")
reg1=lm(t2Oslo~t2Reykjavik,data)
abline(reg1,col="red",lwd=7)
```



Il est possible d'observer sur le graphique ci-dessus que lorsque la température à Reykjavik augmente, la température à Oslo augmente en raison de la corrélation positive qui a été mise en évidence dans le diagramme précédent. De manière similaire, si l'on trace un graphique reliant les deux variables n'importe quel de 7 variables: "t2Reykjavik","t2Oslo","t2Paris", "t2NewYork", "t2Tunis", "t2Alger" et "t2Beyrouth", on obtiendra une droite croissante.

```
plot(t2Oslo,co2,main="la concentration en co2 en fonction de la temperature à Oslo")
reg2=lm(co2~t2Oslo,data)
abline(reg2,col="red",lwd=7)
```



Lorsque les variables sont peu corrélées, cela peut se traduire graphiquement par l'absence d'une tendance globale dans les données. Par exemple, pour deux variables : "t2Oslo" et "co2", la corrélation est presque de 0, donc la liaison est plutôt faible. En observant le diagramme ci-dessus, on constate que la pente du modèle linéaire de ces deux variables est quasiment de 0, c'est à dire que le changement de la variable "t2Oslo" n'affecte pas la variable "co2".

III. Analyse en Composantes Principales

III.1. Le principle de l'ACP

La méthode de l'Analyse en Composante Principale (ACP) est un outil statistique dont le principle est de diagonaliser la matrice de covariances ou de corrélations selon si la matrice de travail est centrée ou centrée-réduite. L'ACP nous permet de dépasser les limites de l'analyse unidimensionnelle ou bidimensionnelle en permettant une représentation d'un nombre élevé de variables quantitatives dans un espace plus petit (souvent à 2 dimensions).

Le fonctionnement de base de l'ACP est de chercher les axes (ou directions) qui expliquent "la plus grande variance" des données. Les variables initiales sont alors projetées sur ces axes pour créer de nouvelles variables, appelées **composantes principales**. Les premières composantes principales contiennent donc le maximum d'information possible sur les données initiales, tandis que les dernières contiennent le minimum.

Cette méthode permet de simplifier la structure des données et de visualiser les relations entre les observations. Elle est souvent utilisée en analyse de données pour identifier les tendances, les groupes ou les différences entre les individus ou les variables. Elle peut également être utilisée pour la réduction de dimensionnalité, la classification ou la modélisation des données.

L'interprètation des graphiques de l'ACP

Le graphique des variables, aussi appelé cercle des corrélations, permet de visualiser en deux dimensions les corrélations. Plus les variables sont proches du cercle, plus elles sont bien représentées et plus l'angle entre deux variables est faible, plus elles sont corrélées. La mesure du cosinus de l'angle formé entre deux variables est égal au coefficient de corrélation linéaire entre les 2 variables. Lorsque deux flèches, représentant deux variables, sont longues et vont dans la même direction (ou dans des directions opposées), cela indique une forte corrélation positive (ou négative) entre les variables.

Le graphique des individus permet de visualiser les similitudes entre les individus dans un espace à deux

dimensions. Les individus qui sont proches les uns des autres sur le graphique ont des profils similaires en termes de variables étudiées. Les individus qui sont éloignés du centre du graphique sont plus influencés par les composantes principales que ceux qui sont plus proches du centre.

III.2. Analyse et interprétation du résultat

Implémentation

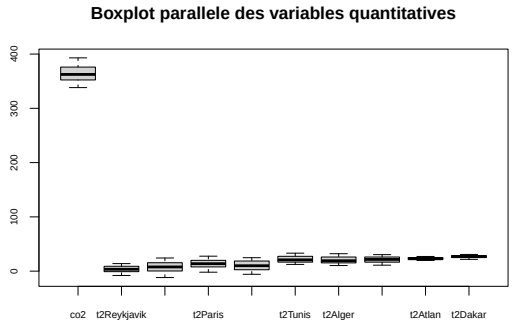
Pour faire l'ACP, nous ne retenons que les variables quantitatives (donc *co2, t2Reyjavik, t2Oslo, t2Paris, t2NewYork, t2Tunis, t2Alger, t2Beyrouth, t2Atlan, t2Dakar*). Nous les stockons ainsi dans une matrice de données initiale X en barrant les 2 premières colonnes :

```
X=data[,-(1:2)]
```

Choix de l'ACP

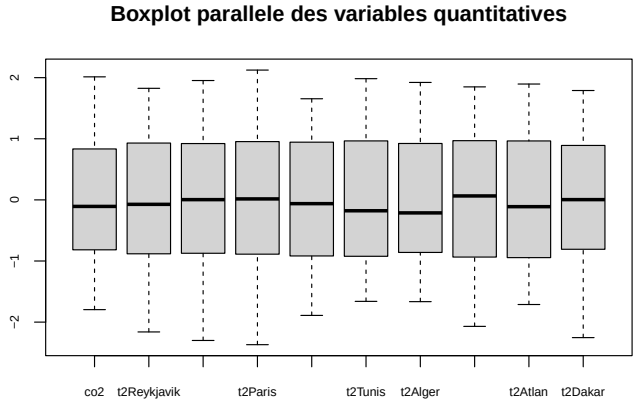
Nous avons 2 options de l'ACP à choisir : soit ACP centrée ou ACP centrée-réduite. Nous allons décider laquelle à utiliser après avoir observé le boxplot :

```
boxplot(X,main='Boxplot parallele des variables quantitatives', cex.axis=0.65)
```



Dans la figure, la variable **CO2** écrase totalement les autres ainsi que la distance entre l'élément outlier est élevée. Dans ce cas, nous optons pour l'ACP centrée-réduite. Ainsi, il nous convient maintenant centrer et normaliser la matrice des données originale X:

```
X_cr=scale(X,center = TRUE,scale = TRUE)
boxplot(X_cr,main='Boxplot parallele des variables quantitatives',cex.axis=0.65)
```



Le formalisme mathématique de l'ACP :

1) Rassemblement des données: X

Nous ayons un ensemble de données contenant p variables x1, x2, . . . , xp pour n individus, qui peuvent être représentés sous la forme d'une matrice de données X de taille n x p.

2) Création de la de travail: T

La matrice de travail est préparée en centrant et/ou réduisant les variables pour que celles-ci aient une moyenne nulle et/ou une variance unitaire. Cela permet de traiter les variables sur une même échelle et de prendre en compte les différences entre les individus dans les analyses. Puisque nous avons choisi l'**ACP centrée réduite**, la matrice **T** sera calculée comme suit:

```
T = scale(X,center = TRUE, scale = TRUE)
```

3) Définition des espaces de travail: M et W

Afin d'apporter une notion de distance entre les individus, on crée une matrice Mpp diagonale afin de munir R^p d'une norme et d'un produit scalaire. Afin d'apporter une notion de poids, d'importance entre les variables, on crée une matrice Wnn diagonale afin de munir R^n d'une norme et d'un produit scalaire.Sauf mention contraire, les matrice **M** et **W** seront calculées comme suit:

```
M = diag(length(X[1,]))  # Ip
W = diag(length(X[,1]))/length(X[,1]) # (1/n)*In
```

4) Calcul de la matrice d'inertie: Inertie

L'**Inertie** représente la distance moyenne entre chaque individu et l'individu de référence. S'il est grand, cela signifie que le nuage de points est très dispersé ; sinon, le nuage de points est assez ramassé autour de la référence individuelle.

```
gam = t(T) %*% W %*% as.matrix(T)
covX = gam %*% M
Inertie=sum(diag(covX))
```

Nous trouvons que **Inertie = 9,97** est à peu près égal au nombre de variables quantitatives **p = 10**

5) Recherche des axes principaux: a1 et a2

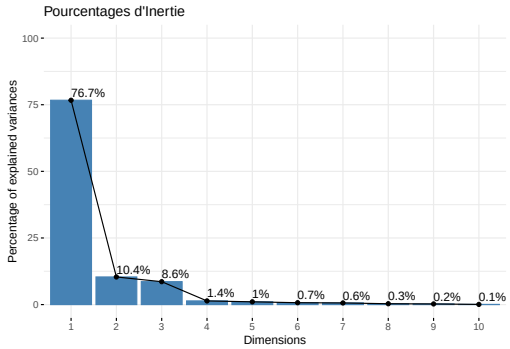
Le premier axe principal a_1 est vecteur propre associé à la plus grande valeur propre de la matrice ΓM.Le deuxième axe principal a_2 est vecteur propre associé à la deuxième plus grande valeur propre de la matrice ΓM.

```
Valeurs_Propres=eigen(covX)$values
Vecteurs_Propres=eigen(covX)$vectors
a1=Vecteurs_Propres[,1]
a2=Vecteurs_Propres[,2]
```

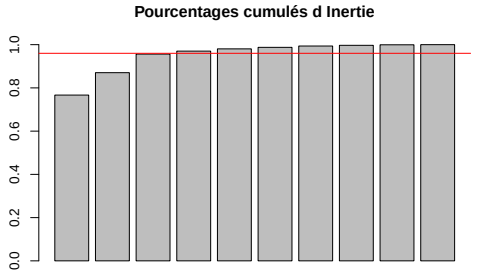
6) Calcul du pourcentage d'inertie:

Si le pourcentage d'inertie est grand pour une composante principale, cela indique que cette composante principale capture une grande partie de la variabilité des données d'origine. Dans ce cas, cette composante principale est considérée comme importante et devrait être retenue dans l'analyse. Sinon, cette composante principale peut être ignorée dans l'analyse sans perte significative d'information.

```
PC=Valeurs_Propres/sum(Valeurs_Propres)*100
res.acp <- PCA(X,scale.unit=TRUE,ncp=5,graph=FALSE)
par(mfrow=c(1,2))
fviz_eig(res.acp, addlabels = TRUE, ylim = c(0, 100),main="Pourcentages d'Inertie")
```

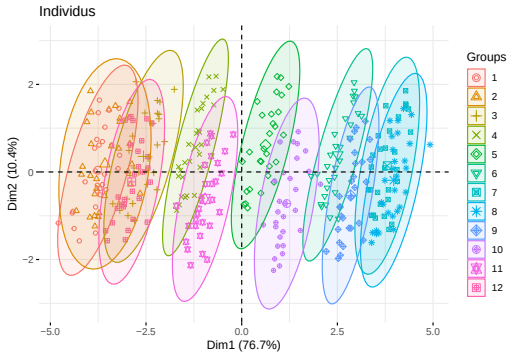



```
barplot(cumsum(PC)/sum(PC),main="Pourcentages cumulés d Inertie")
abline(h=0.96,col='red')
```



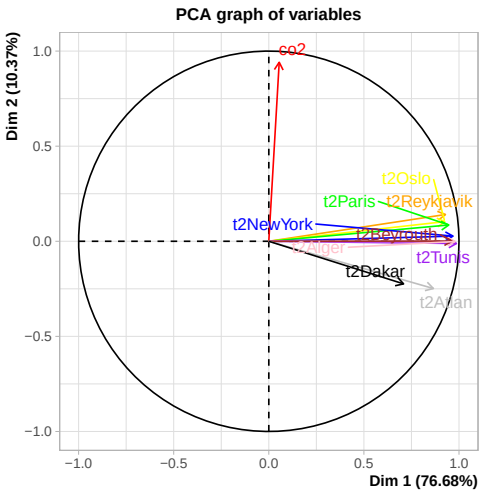
Sur la base du graphique, nous pouvons voir: les **3 composantes principales** ont représenté **96%** de l'information. **96%** est un pourcentage d'informations suffisamment important pour que nous puissions faire confiance. Nous avons donc décidé de garder les **3 composants principales** Nous pouvons maintenant afficher le nuage de points avec comme axes les deux premières composantes principales grâce aux commandes suivantes :

```
fviz_pca_ind(res.acp, geom.ind = "point",title = "Individus",col.ind = data$mois,
             addEllipses = TRUE,legend.title = "Groups")
```



Sur le nouveau système de base, les informations sont dispersées autour des deux nouveaux axes de coordonnées. Dans lequel, nous pouvons voir que **la variance sur la dimension 1 est plus grande que sur la dimension 2**.

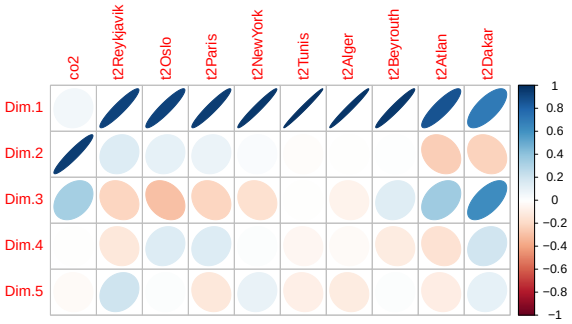
```
colors <- c("red", "orange", "yellow", "green", "blue", "purple", "pink", "brown", "gray", "black")
plot(res.acp, choix="varcor", axes=c(1,2), max.overlaps = Inf, col.var = colors)
```



Nous voyons qu'**un point est plus au-dessus de l'axe horizontal, cela signifie que la concentration de CO2 à ce moment est grande**, alors que, si le point est situé **à droite de l'axe vertical**, c'est-à-dire, **plus la température est élevée dans les villes ci-dessus**.

On utilise les diagramme des corrélations suivant qui établit le lien entre nos variables de base et nos composantes principales :

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



Nous voyons que la corrélation des variables est la plus évidente dans les **3 premières dimensions**. Dans la première composante principale (porte environ **76%** de l'information), il y a une **forte corrélation positive avec le changement de température dans les villes**, en revanche, **la corrélation avec la concentration en CO2 est très faible**. En revanche, la 2ème composante principale (portant environ **11%** de l'information) était **fortement corrélée avec la concentration de C02** alors que **la corrélation était très faible avec la température dans les villes**. Enfin, la troisième composante principale (portant environ **8%** de l'information) **n'est que fortement corrélée avec la température à Dakar dans le sens positif et la température à Oslo dans le sens négatif**. Il s'agit donc essentiellement de présenter des informations sur l'une de nos variables de base : c'est une indication que les principales composantes commencent à manquer d'informations.

IV. Conclusion

En effectuant une analyse en composantes principales (ACP), nous avons pu mettre en évidence de manière plus précise les interactions entre nos variables. La visualisation des résultats nous a permis de regrouper les paramètres de température dans plusieurs villes, de concentration en CO2 , de mois et d'année. Malgré la complexité d'interprétation des ACP, l'analyse unidimensionnelle et bidimensionnelle nous a donné une vue d'ensemble des phénomènes étudiés ainsi que des dépendances entre les variables.