Projet Analyse De Données

DIARRASSOUBA SAKARIA, DOUBA JAFUNO, OUMAR BALDE

20/12/2019

Les Données

Notre projet porte sur l'étude de la température dans plusieurs pays européens. Notre jeu de donnée décrit une comparaison intrinsèque entre les températures mensuelles et annuelles au sein de 35 pays, il contient donc 35 observations et 17 variables pour chaque ville dont:

16 quantitatives: -les *mois* de l'année, la temperature *moyenne* annuelle, l'*amplitude* thermique, la *longitude*, la *latitude*
1 qualitative: la *Région*

Valeurs manquantes

Tout d'abord étudions pour chaque variable, le nombre de valeurs manquantes s'il en existe on utilise pour cela la fonction is.na de R qui repère les variables manquantes remplacées par NA dans chacune de nos variables; Nous obtenons ce tableau suivant avec le nombre de variables manquantes (nvm):

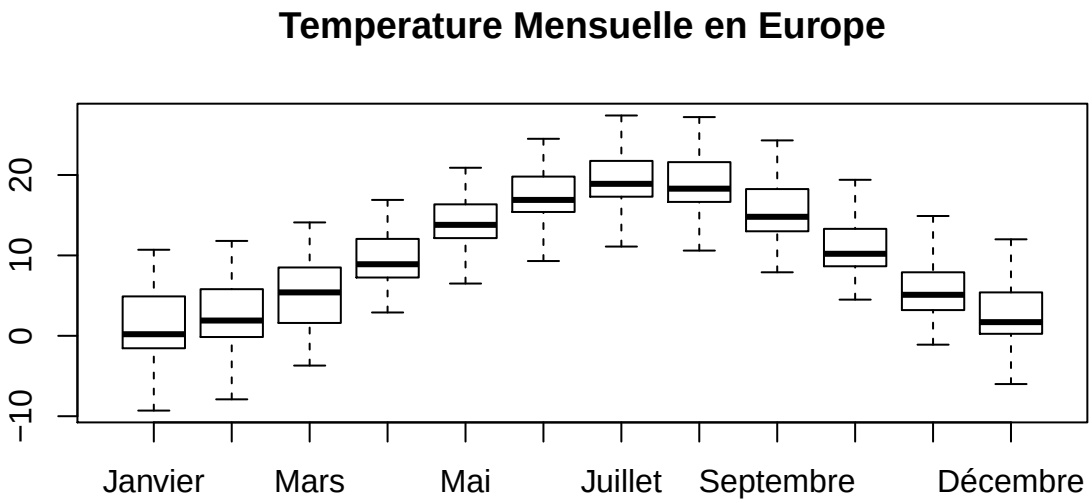
.	Janvier	Février	Mars	Avril	Mai	Juin
nvm	0	0	0	0	0	0

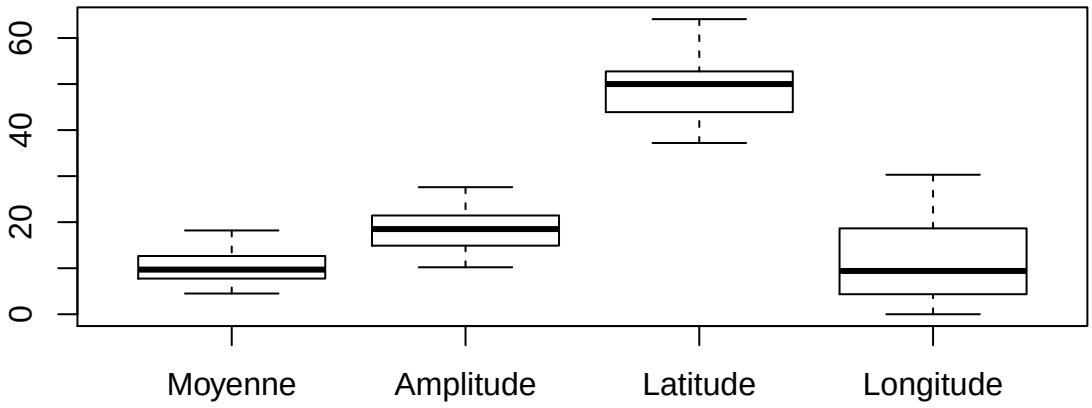
.	Juillet	Aout	Septembre	Octobre	Novembre	Décembre
nvm	0	0	0	0	0	0

Résumé et Distribution de notre jeu de donnée

.	Min	1st Qu	Median	Mean	3rd Qu	Max
Janvier	-9.300	-1.550	0.200	1.346	4.900	10.700
Février	-7,900	-0,150	1,900	2,217	5,800	11,800
Mars	-3,700	1,600	5,400	5,229	8,500	14,100
Avril	2,900	7,250	8,900	9,283	12,050	16,900
Mai	6,50	12,15	13,80	13,91	16,35	20,90
Juin	9,30	15,40	16,90	17,41	19,80	24,50
Juillet	11,10	17,30	18,90	19,62	21,75	27,40
Août	10,60	16,65	18,30	18,98	21,60	27,20
Septembre	7,90	13,00	14,80	15,63	18,25	24,30
Octobre	4,50	8,65	10,20	11,00	13,30	19,40
Novembre	-1,100	3,200	5,100	6,066	7,900	14,900
Décembre	-6,00	0,25	1,70	2,88	5,40	12,00
Longitude	0,00	4,35	9,40	11,98	18,65	30,30
Latitude	37,20	43,90	50,00	48,77	52,75	64,10
Moyenne	4,50	7,75	9,70	10,27	12,65	18,20

.	Est	Nord	Ouest	Sud
Région	8	8	9	10

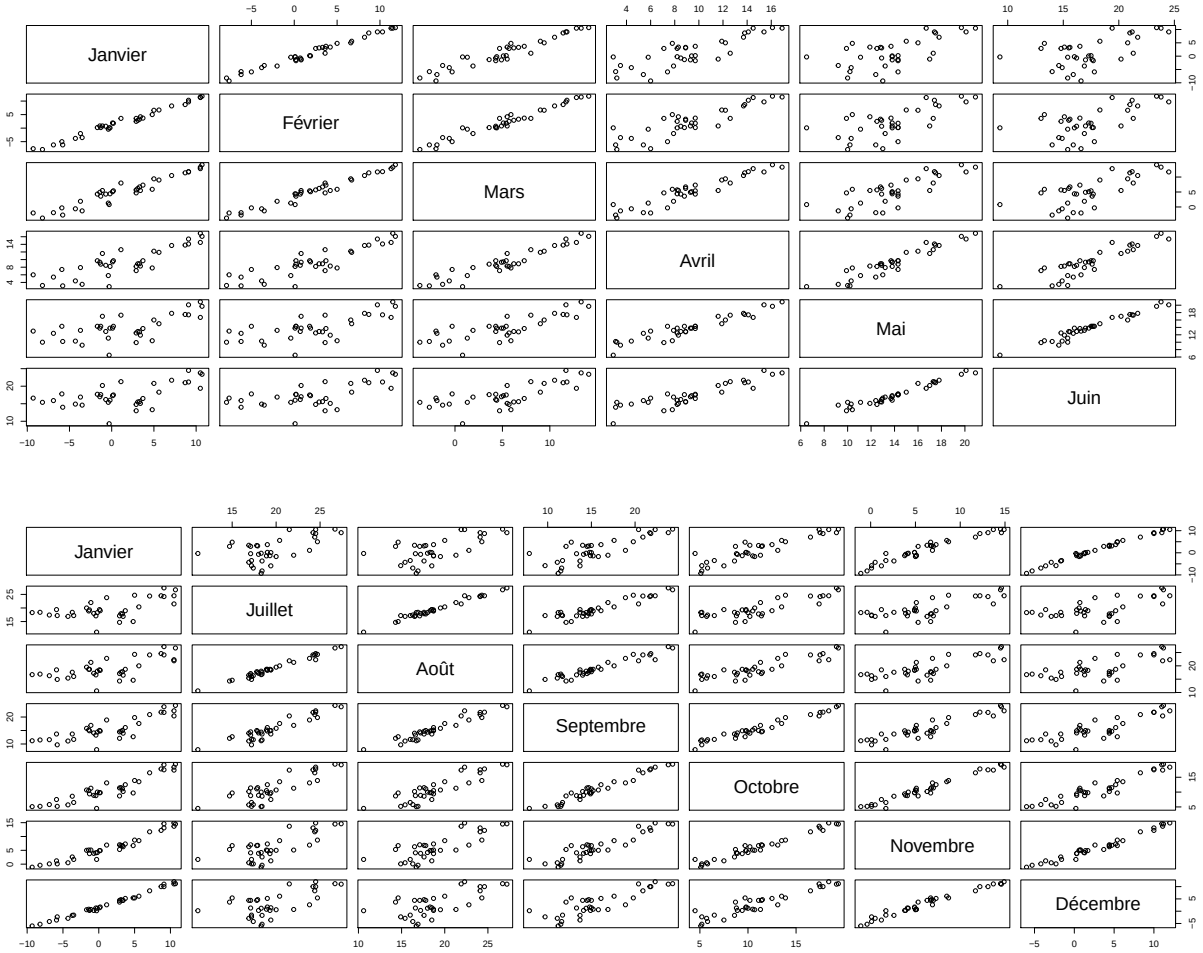




Comme on peut s'y attendre, les températures moyennes mensuelles augmentent de Janvier à Juillet en Europe avant de diminuer jusqu'à Décembre, les températures sont plus basses entre Novembre et Mars avec une distribution entre -9 dégrés en Janvier et 15 dégrés en Novembre, elles sont plus élevées en Juillet et en Aout avec une moyenne d'environ 18 dégrés en pleine saison d'été et enfin de Mars à Juin et de Septembre à Octobre les temperatures moyennes mensuelles sont environ de 12 dégrés dans ces 35 villes Européenne. Parmis nos 4 dernières variables quantitatives la Longitude a la plus grande variabilité et la latitude atteint des valeurs beaucoup plus grande et notons que l'ensemble des boxplots ne sont pas symétriques

Corrélation de nos variables mensuelles et nuages de points

Pour cela calculons et représentons la matrice .On s'intéresse aux liens qu'il pourrait avoir entre les corrélations linéaires



.	Janv	Févr	Mars	Avril	Mai	Juin	Juill	Août	Septem	Octob	Novemb	Décemb
Janv	1	0,99	0,95	0,83	0,63	0,56	0,57	0,64	0,81	0,91	0,96	0,99
Févr	0,99	1	0,98	0,88	0,69	0,62	0,62	0,69	0,85	0,93	0,97	0,98
Mars	0,95	0,98	1	0,94	0,79	0,72	0,72	0,78	0,91	0,96	0,97	0,96
Avril	0,83	0,88	0,94	1	0,94	0,89	0,86	0,89	0,97	0,96	0,92	0,85
Mai	0,63	0,69	0,79	0,94	1	0,97	0,94	0,94	0,94	0,88	0,79	0,68
Juin	0,56	0,62	0,72	0,89	0,97	1	0,98	0,96	0,93	0,83	0,74	0,61
Juill	0,57	0,62	0,72	0,86	0,94	0,98	1	0,99	0,93	0,84	0,74	0,62
Août	0,64	0,69	0,78	0,89	0,94	0,96	0,99	1	0,96	0,88	0,79	0,68
Septem	0,81	0,85	0,91	0,97	0,94	0,93	0,93	0,96	1	0,97	0,92	0,84
Octobre	0,91	0,93	0,96	0,96	0,88	0,83	0,83	0,88	0,97	1	0,98	0,93
Novemb	0,97	0,97	0,973	0,92	0,79	0,74	0,74	0,79	0,92	0,98	1	0,98
Décemb	0,99	0,98	0,96	0,85	0,68	0,60	0,62	0,68	0,84	0,93	0,98	1

On remarque que toutes nos variables mensuelles sont deux à deux significativement corrélés cependant les corrélations sont moins forte entre les mois d'hiver et d'été comme pour Janvier et Juin avec 0.56

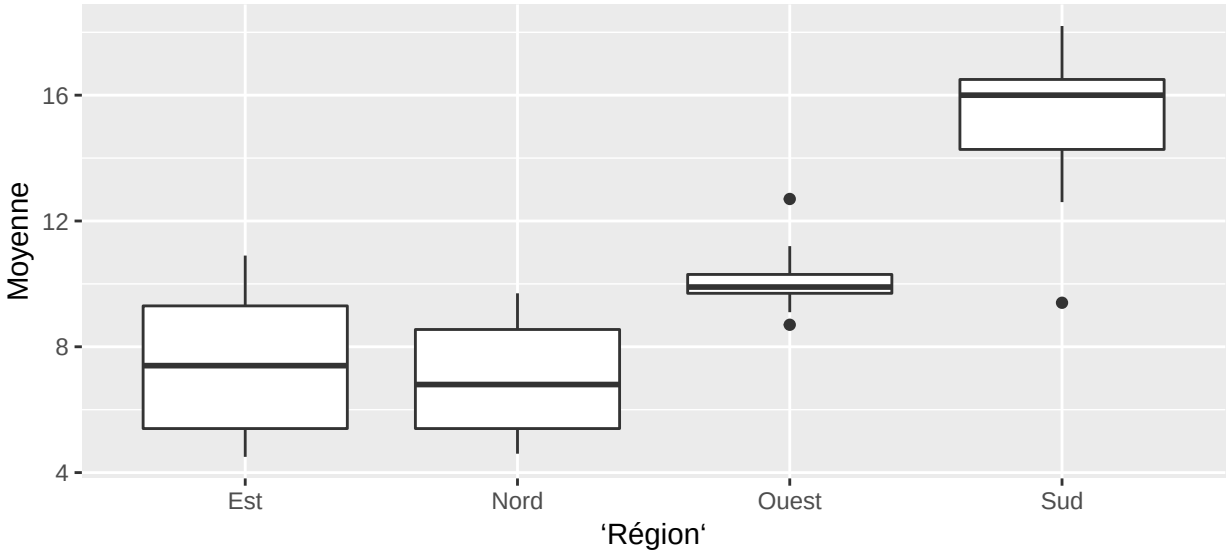
Liens avec la température moyenne annuelle

Anova à 1 facteur

Ici nous considérons le modèle à un facteur car nous avons une seule variable qualitatif qui est la Région et où la température moyenne annuelle de chaque ville est la variable d'interet et la région de la ville est le facteur

Représentation des données

On trace le boxplot pour visualiser l'impact de la région sur la température moyenne annuelle



A la fin de l'année, nous observons une différence de la répartition de la température moyenne annuelle dans chaque régions. La température des régions du sud présente plus de variabilité (avec une valeur aberrante) et est plus élevée de 9 à 18 degrés que celle des autres régions, nottament de celle du Nord qui varie de 5 à 9 degrés. Pour les régions de l'ouest la température moyenne annuelle est globalement très concentré entre 9 et 11 degrés avec des valeurs aberrantes à 13 dégrés et à 8.5 dégrés. La distribution des régions de l'est est semblable à celles des températures du Nord avec une variabilité jusqu'à 11 dégrés. Il semble donc que la Région ait une influence sur la température moyenne annuelle au vue des variabilités mais cela est-il significatif?

Région variable qualitative

On considère le modèle: $\text{MOY}_i = \beta_0 + \beta_1 \text{REGION} + \varepsilon$ ici la variable Région est interprété comme une variable quantitative. Mais en réalité la variable Région est de type qualitative, ce modèle n'a donc aucun sens d'un point de vue d'interprétation.Observons alors ce qui se passe en la définissant comme une variable qualitative, c'est à dire en la déclarant comme une variable de type factor.

Tableau des coefficients de la Fonction lm

.	Estimate	Std. Error	t value	Pr(>abs(t)
(Intercept)	7.4500	0.7516	9.913	3.95e-11
RégionNord	-0.4625	1.0629	0.435	0.6665
RégionOuest	2.7389	1.0329	2.652	0.0125
RégionSud	-0.4625	1.0629	0.435	1,11e-08

On observe qu'il s'agit d'un modèle de type analyse de la variance à 1 facteur en effet le tableau des **Coefficients** est composé de trois lignes (**Intercept**, **RégionNord**, **RégionSud** et **RégionOuest**).

Avec les observations $\text{Moyenne}_{i,j}$ on considère donc le modèle suivant:

$$\text{Moyenne}_{ij} = \mu_1 + \alpha_i + \varepsilon_{ij}, \qquad i = 1, \ldots, I \quad j = 1, \ldots, n_i.$$

ou l'indice i indique la région, μ_i les moyennes de ses groupes avec la contrainte $\alpha_1 = 0$, on a $\varepsilon_{ij} \sim N(0, \sigma?)$. Autrement dit, les α_i pour $i = 2, \ldots, I$ vérifient $\alpha_i = \mu_i - \mu_1$, on voit alors que tout les autres paramètres (pour $i > 1$) dépendent du niveau $i = 1$ la Région EST

La colonne **Estimate** contient les estimateurs de μ_1 (**Intercept**: 7.4500), de α_2 (**RégionNord**: -0.4625) de α_3 (**RégionOuest**:2.7389) et de α_4 (**RégionSud**: -0.4625). La colonne **t value** représente la valeur observée de la statistique du test d'hypothèse H_0:μ_1=α_2=α_3= 0 contre $H_1 : \mu_1 \neq 0, \alpha_2 \neq 0$, $\alpha_3 \neq 0$ et $\alpha_4 \neq 0$ Les tests de significativités des coefficients (**Pr(>|t|)**) donnent ici des p-valeurs inférieures à 0.05: 0.06665 et 0.00908 La très petite p-valeur (3.95e-11) pour la constante indique que la constante (l'intercept) doit apparaitre dans le modèle ce qui n'est pas le cas de la Région Nord . Tout cela montre bien que les coefficients **RégionOuest** et **RégionEst** sont significatives sur la température moyenne annuelle (**Moyenne**) ainsi l'hypothèse nulle H_0 de chacun des tests est rejeté au profit de l'hypothèse alternative H_1.

Pour l'analyse de la variance, la paramétrisation du modèle n'a pas d'importance, car dans les deux cas, les sous-espaces vectoriels sont les mêmes.

Tableau de l'analyse de la variance

.	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Région	3	393.91	131.304	29.056	3.879e-09 ***
Residuals 31	140.09	4.519			

La dernière colonne **Pr(>F)** est la plus importante car elle permet de conclure pour notre test. En fait, **Pr(>F)** est la p-value du test de l'absence d'effet du au facteur (REGION) dans le cas de la température moyenne annuelle.Rappelons qu'en général, si la p-value est petite (< 0.01), on rejette H_0 et on décide H_1. **Ici, on voit que la p-value du test de l'absence d'effet du à la Région est de 3.879e-09 , ce qui est clairement en faveur de l'hypothèse H_1. Il n'y a pas de doute que la Région de la ville influence significativement sur sa température moyenne annuelle**. # Liens entre Latitude et Longitude avec la température moyenne annuelle

Régression linéaire (multiple)

Dans cette partie nous allons effectuer une régression multiple pour prédire la température annuelle moyenne de chaque ville en fonction de sa latitude et de sa longitude. considérons le modèle suivant: $\text{X}_i = \alpha_i + \beta_i \times L_i + \gamma_i \times l_i + \varepsilon_i, \qquad i = 1, \ldots, I \quad j = 1, \ldots, n_i.$ avec: L=Longitude, l=latitude, $\varepsilon \sim N(0, \sigma^2)$.

.	Estimate	Std. Error	t value	Pr(>abs(t)
(Intercept)	33.65250	2.42736	13.864	4.46e-15
Latitude	-0.46502	0.05115	-9.092	2.21e-10
Longitude	-0.05903	0.04040	-1.461	0.154

On remarque que la p_value de la variable Longitude est trop grande, alors elle n'interviendra pas dans notre modèle. Par conséquent, seule la **Latitude** influe sur la température moyenne annuelle. Le modele retenu est donc

$$\text{X}_i = \alpha_i + \gamma_i \times l_i + \varepsilon_i, \qquad i = 1, \ldots, I \quad j = 1, \ldots, n_i.$$

Mise en place d'une ACP

Dans cette partie ont veut faire une comparaison entre les températures mensuelles dans les pays d'Europe en se basant sur nos données.

Choix des axes

Nous examinons les valeurs propres pour déterminer le nombre de composantes principales à prendre en considération. Les valeurs propres et la proportion de variances (i.e. information)retenues par les composantes principales peuvent être extraites à l'aide de la fonction get_eigenvalue()

.	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	9.94775042	82.8979202	82.89792
Dim.2	1.84764850	15.3970708	98.29499
Dim.3	0.12625580	1.0521317	99.34712
Dim.4	0.03829345	0.3191121	99.66623
Dim.5	0.01670941	0.1392451	99.80548
Dim.6	0.01283304	0.1069420	99.91242

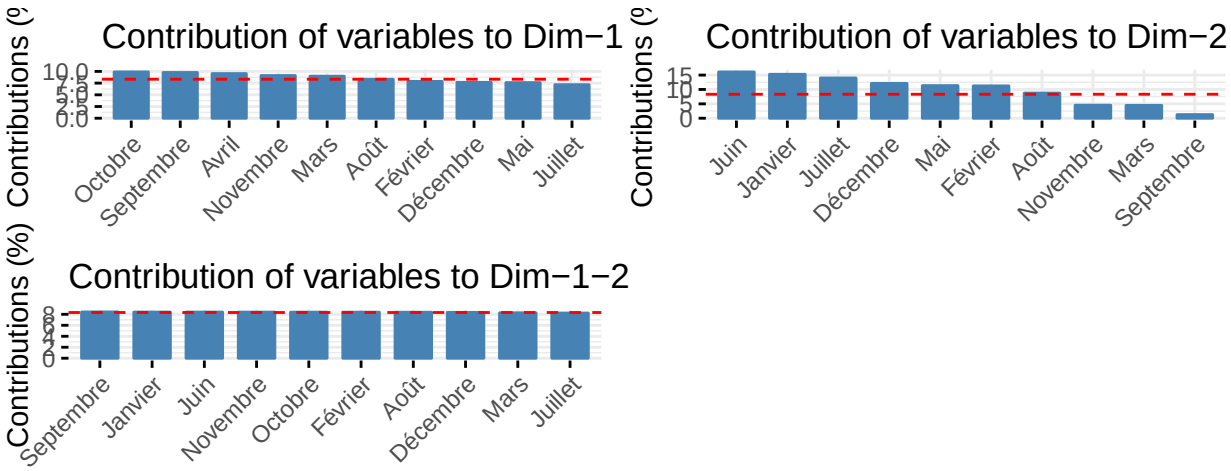
La somme de toutes les valeurs propres donne une variance totale de 12.

La proportion de variance expliquée par chaque valeur propre est donnée dans la deuxième colonne. On peut voir que environ 82.89% de la variance totale est expliquée par le premier axe. Le pourcentage cumulé expliqué est obtenu en ajoutant les proportions successives de variances expliquées.

Une valeur propre > 1 indique que la composante principale (PC) concernée représente plus de variance par rapport à une seule variable d'origine, lorsque les données sont standardisées. Ceci est généralement utilisé comme seuil à partir duquel les PC sont conservés. A noter que cela ne s'applique que lorsque les données sont normalisées.

Selon le critère de Kaiser, nous pourrions nous arrêter à la deuxième composante principale dont **98.29%** des informations (variances) contenues dans les données sont conservées par les deux premières composantes principales dans le graphique ci-dessus et qui possèdent une valeur propre supérieur à 1 par conséquent, nous retenons 2 premières axes

Contribution des variables

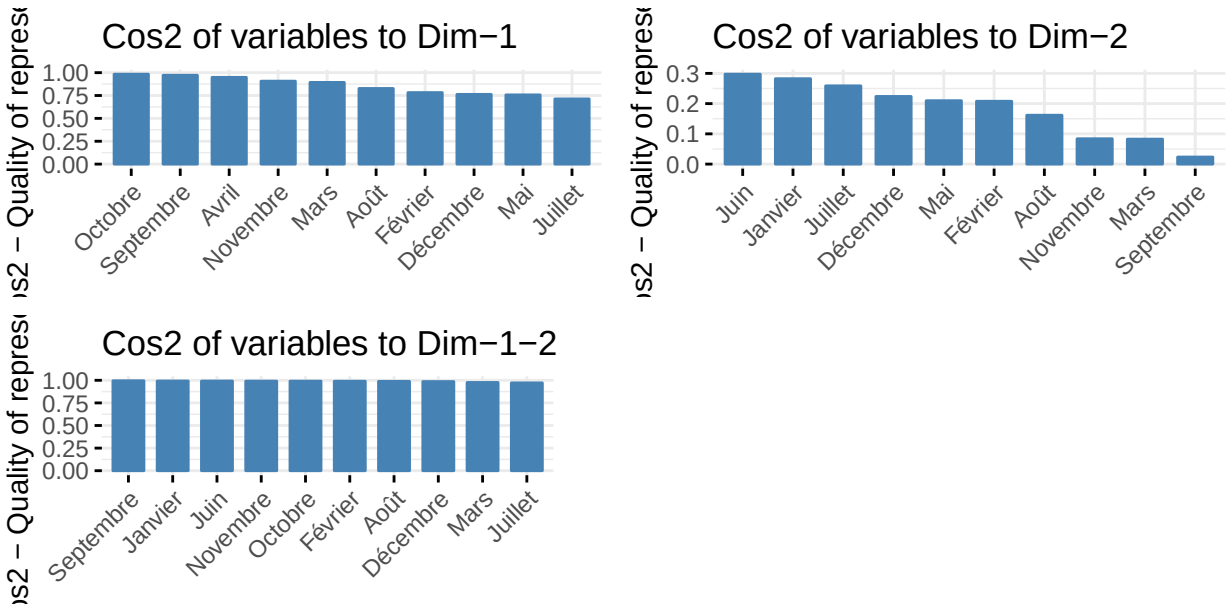


Nous constatons que 5 variables : **Octobre**, **Septembre**, Avril, Novembre et Mars ont une forte contibution supérieure à la contribution moyenne (>8%) selon Dim 1

Nous constatons que 7 variables : **Juin**, **Janvier** Juillet Décembre Mai Février Aout ont une forte contibution supérieure à la contribution moyenne (>8%) selon Dim 2

Enfin nous voyons que les variables **Septembre**, **Janvier** et **Juin** contibuent le mieux sur les 2 axes en meme temps on peut également voir cela sur le **cercle de corrélation**

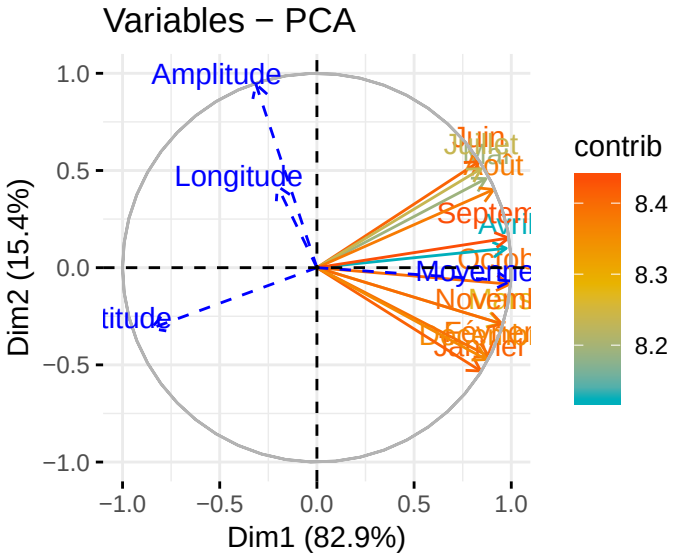
Qualité de représentation des variables



Nous constatons que ces 3 graphes sont semblables à celles de la contibution des variables on peut donc dire que les variables qui ont une bonne contribution sur 1 ou plusieurs axes ont aussi une bonne qualité de représentation sur ces meme axes

Avec le cercle de corrélation

Un cos2 élevé ou une contibution élevé = bonne représentation de la variable sur les axes principaux (3eme graphique de contribution et cos 2 sur Dim1 et Dim2) → la variable est positionnée à proximité du bord du cercle de corrélation. Un cos2 faible ou une contribution faible = la variable n'est pas bien représentée par les axes principaux → la variable est proche du centre du cercle.Voici un exemple pour la contrinution des variables:



ON peut voir sur notre **cercle des corrélations** que la 1ere composante principale est prédominante car elle resume à elle seule 82.90 % de l'inertie totale comme on l'a vu dans la première partie, la 2eme composante est relativement importante, elle résume 15,4 de l'inertie totale;

Ces 2 composantes donnent 98,3 comme inertie totale. Toutes les variables actives sont du même coté que la 2 ième composante, nous notons que les mois "Septembre" "Octobre" et "Avril" sont plus étroitement liés avec cette composante que les autres variables comme ont l'a vu précédemment.

Description des dimensions

Significativité des variables avec Dim 1

.	correlation	p.value
Moyenne	0.9975483	9.575809e-26
Octobre	0.9916246	3.734359e-20
Septembre	0.9856254	1.056964e-17
Avril	0.9738876	5.295583e-15

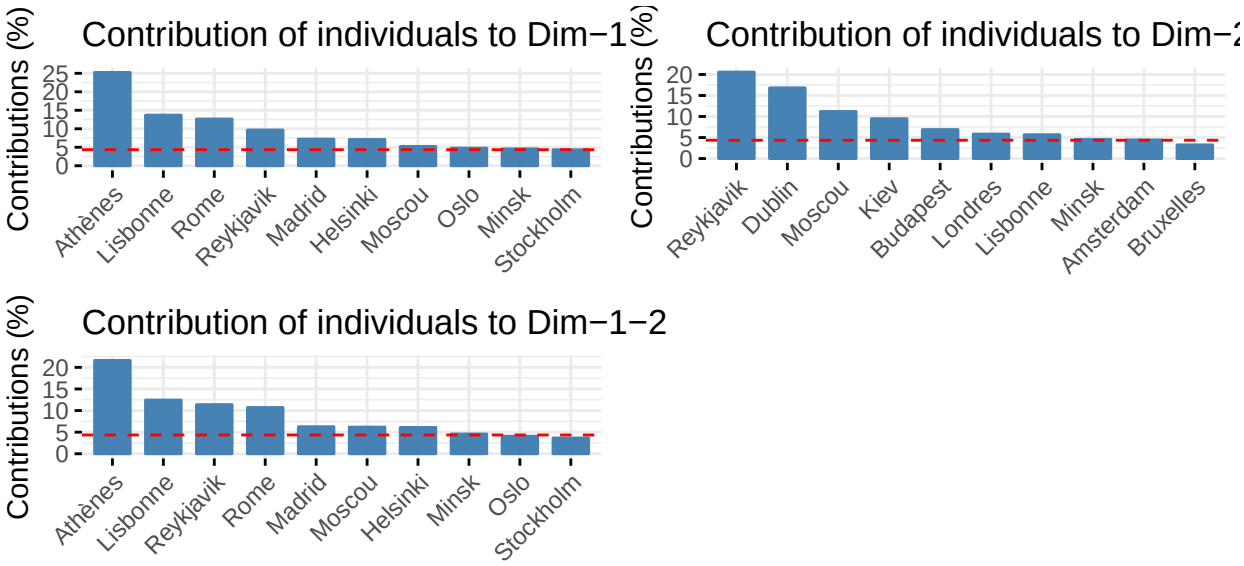
La variable "Octobre" est la plus corrélée avec le premier axe (une corrélation de 0.991)

Significativité des variables avec Dim 2

.	correlation	p.value
Amplitude	0.9444140	1.296159e-11
Juin	0.5453220	7.119942e-03
Juillet	0.5086619	1.319151e-02
Mai	0.4578116	2.804268e-02
Longitude	0.4196483	4.621251e-02

La variable "Juin" est la plus corrélée avec le 2ème axe (une corrélation de 0.545). La longitude a une corrélation de 0,41 avec cette axe.

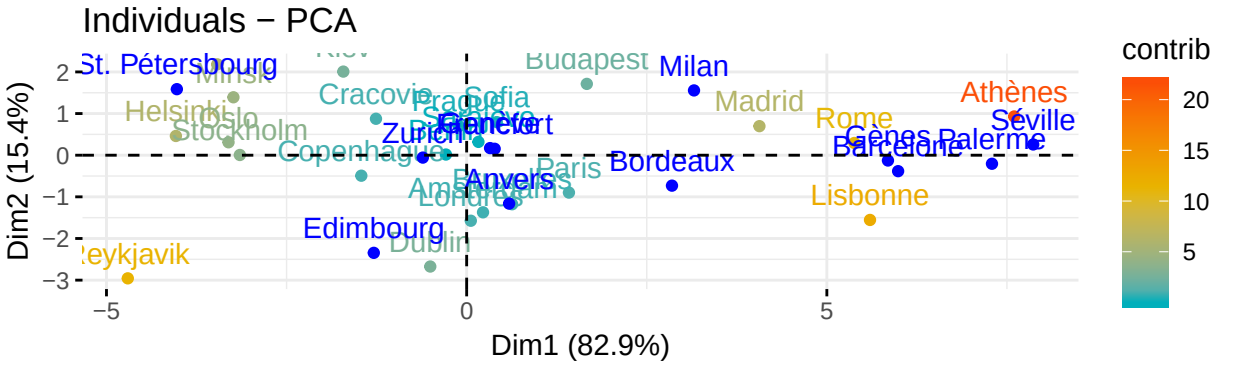
Contribution des Individus



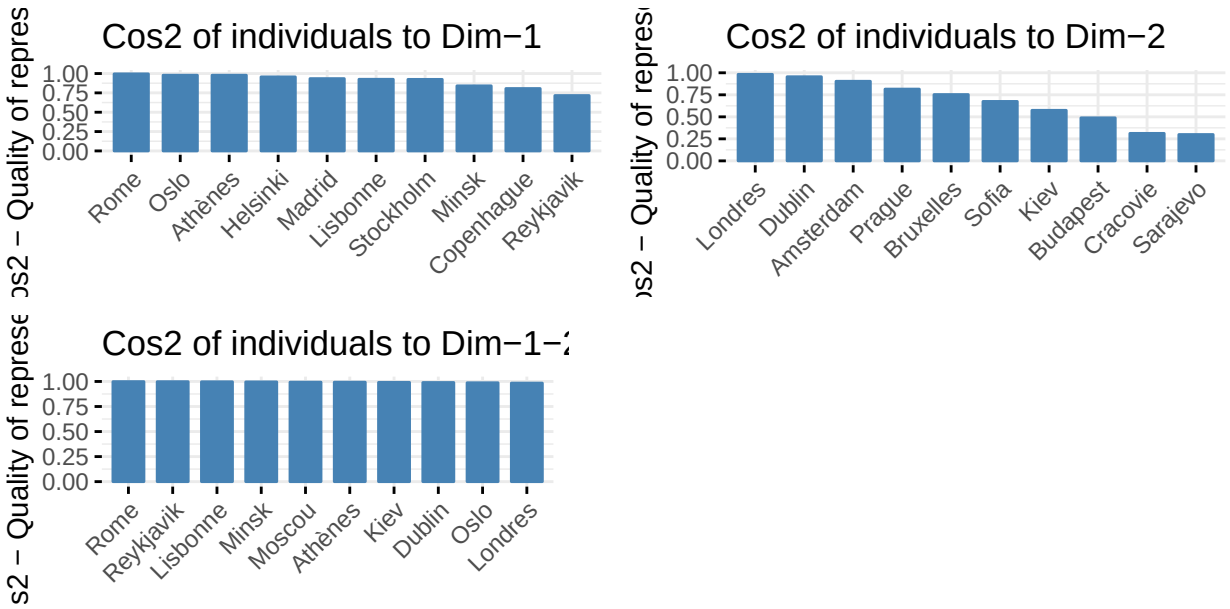
Nous constatons que 6 variables : **Athènes**, **Lisbonne**, Rome, Reyjavik, Madrid et Helsinki ont une forte contibution supérieure à la contribution moyenne (>5%) selon Dim 1

Nous constatons que 7 variables : **Reyjavik**, **Dublin**, Moscou, Kiev Budapest Londres et Lisbonnes Aout ont une forte contibution supérieure à la contribution moyenne (>5%) selon Dim 2

Enfin nous voyons que les variables **Athènes**, **Lisbonne**, **Reykjavic** ,Rome, Madrid et Helsinki contibuent le mieux sur les 2 axes en meme temps on peut également voir cela sur le **graphe des individus** qui est un peu semblable au cercle des corrélations



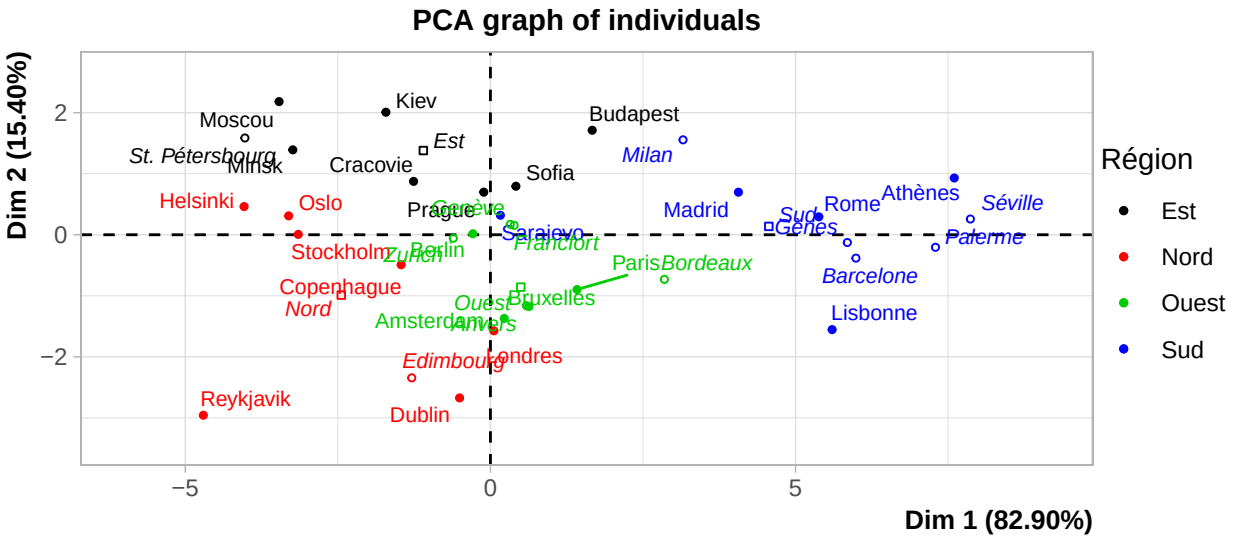
Qualité de représentation des Individus



Les villes **Rome**, **Oslo** et **Athènes** ont un Cos2 très proche 1, alors ces ville ont une bonne qualité de représentation sur DIm 1, il en est de meme pour **Londres Dublin** et **Amsterdam** sur Dim2 et **Rome Reykjavik** et **Lisbonne** sur nos deux axes

Conclusion

Voici le nuage des individus les villes



Nous avons pu synthétiser l'information relative à 35 villes en 2 dimensions à partir des différentes variables avec une précision de **98.29%**.

En se référant à ce graphique nous voyons que les villes qui sont situés au sud et qui ont des coordonnées elevés sur la 1ere composante présentent des températures annuelles elevées et sont donc des villes chaudes et donc de latitudes faibles.

Sur le cercle de corrélation on a pu voir que l'amplitude est fortement liée à la 2ème composante et les valeurs les plus elevées de cette mésure ont été observées pour des villes de l'Est et les valeurs les plus basses ont été observées pour les villes européennes de l'ouest et le nord.

La longitude est liée à cet axe mais la relation n'est pas forte avec une corrélation de 0,41 on a meme pu l'éliminer dans notre modèle de régression dans la partie précédente

On a aussi vu qu'il existe une corrélation positive entre les températures mensuelles et annuelles et plus précisément dans deux périodes importantes : la saison estivale et la saison hivernale.

Avec le graphe des individus on peut considérer deux typologies des villes :**les villes du nord froides avec des hautes latitudes et les villes du sud caractérisées par des températures mensuelles et annuelles elevées et des latitudes faibles**.