CY TECH	ING1-GIA EXAMEN DE DATA EXPLORATION 2022-2023
Durée : 2h00	Examen papier 2 feuilles R/V (manuscrites ou non) autorisées Calculatrice autorisée

Exercice 1

On considère les points suivants :

$$X_1=2,\ X_2=5,\ X_3=8,\ X_4=10,\ X_5=11.$$

a) Appliquer K-means en choisissant comme centres initiaux des 3 clusters respectivement $g_1=2$, $g_2=5$ et $g_3=10$ puis calculer le pourcentage d'inertie expliquée par la partition obtenue.

Exercice 2

On considère le jeu de données Ozone. Nous souhaitons analyser la relation entre le maximum journalier de la concentration en ozone maxO3 (en microgrammes par millilitre : µg/m3) et la température à différentes heures de la journée, la nébulosité à différentes heures de la journée, la projection du vent sur l'axe Est-ouest à différentes heures de la journée et la concentration maximale de la veille du jour considéré. Nous disposons de 112 données relevées durant l'été 2001 à Rennes (fichier ozone.txt).

A- Analyse Quantitative x Quantitative

L'objectif est d'étudier le lien entre le maximum journalier de la concentration en ozone (maxO3) et la température à 9h00 du matin (T9). Avec le logiciel R, nous avons obtenu ce résultat :

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  -25.622        11.4551  -2.235   0.0274 *
T9             6.3130     0.6151  10.263   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.24 on 110 degrees of freedom
Multiple R-squared:  0.4891,
Adjusted R-squared:  0.4845
F-statistic: 105.3 on 1 and 110 DF, p-value: < 2.2e-16
```

a) Déterminer la droite de régression de maxO3 en fonction de T9.

$y = ax + b$; $\text{maxO3} = a \cdot (T9) + b$

$y = 6{,}313\ T9 - 25.622$

b) Quel est le pourcentage de variation de maxO3 expliquée par la droite de régression ?

$R^2 = 0{,}4891$

c) Donner une estimation du coefficient de corrélation entre les deux variables maxO3 et T9.

$R^2 = (r_{xy})^2 \iff r_{xy} = \pm\sqrt{R^2} = \pm\sqrt{0.6831}$

=

d) Calculer la prévision de la valeur de maxO3 si la température à 9h00 du matin est de 15.6.

$Tg = 15,6$; maxO3 ? $\max O_3 = 6,313\, Tg - 25,622$

$\iff \hat{y} : \widehat{\max O_3} = 6,313\,(15,6) - 25,62 = 72.86$

e) On observe une concentration maximale maxO3=87 pour une température à 9h00 du matin de 15.6. Calculer le résidu dans ce cas.

$Tg = 15,6 \Rightarrow \widehat{\max O_3} = 72,86$; $\max O_3 = 87$

$e_i = \max O_3 - \widehat{\max O_3} = 87 - 72,86 = 14,32$

f) Comment peut-on vérifier les hypothèses sur les résidus. Comment peut-on valider un modèle de régression ?

Validité un modèle de rég. : il faut que résidus centrés soient entre $[-2, 2]$ ⇒ pas valeurs abergante.

On donne les indicateurs numériques suivants,

$$\bar{x} = 18.36, \bar{y} = 90.30, s_x^2 = 9.75, s_y^2 = 794.52, cov(x, y) = 61.56$$

où y=maxO3 et x=T9.

g) Retrouver les coefficients de la droite de régression à partir de ces indicateurs (aux arrondis près).

$y = ax + b$; $a = \frac{cov(x,y)}{s_x^2} = \frac{61,56}{9.75} = 6,31$

$b = \bar{y} - a\bar{x} = 90,30 - (6,31)(18,36)$

$= \ldots$

B- Analyse Qualitative x Qualitative

L'objectif est d'étudier le lien entre les deux variables vent (Nord, Est, Ouest, Sud) et la variable Pluie (Pluie, Sec).

A partir des tableaux ci-dessous, répondez aux questions suivantes.

Tableau des effectifs observés

	Est	Nord	Ouest	Sud	*Total*
Pluie	2	10	26	5	*43*
Sec	8	16	23	16	*63*
Humide	0	5	1	0	*6*
Total	*10*	*31*	*50*	*21*	*112*

Tableau des fréquences observées

	Est	Nord	Ouest	Sud	Total
Pluie	0,02	0,09	0,23	0,04	0,38
Sec	0,07	0,14		0,14	
Humide	0,00	0,04	0,01	0,00	0,05
Total	*0,09*	*0,28*	*0,45*	*0,19*	*1,00*

Tableau des profils lignes

	Est	Nord	Ouest	Sud
Pluie	0,05	0,23	0,60	0,12
Sec	0,13		0,37	0,25
Humide	0,00	0,83	0,17	0,00

Tableau des profils colonnes

	Est	Nord	Ouest	Sud
Pluie	0,20	0,32		0,24
Sec	0,80	0,52	0,46	0,76
Humide	0,00	0,16	0,02	0,00

a) Compléter les 4 cases vides dans les tableaux ci-dessus. Expliquez vos calculs.

b) Quel est le pourcentage des observations avec un vent d'Est et un temp pluvieux ?

c) Quel est le pourcentage des observations de vent de Sud ayant un temp sec ?

d) Avec quelles fréquences peut-on comparer les profils lignes ? Que pouvez-vous en conclure sur le lien entre les deux variables ? Argumentez votre réponse.

e) Quel est l'effectif théorique des relevés par temps sec et vent d'est ? Que signifie « théoriques » ?

f) On suppose que la distance de khi-deux est de 17.7, donner une conclusion sur le lien entre les variables. Justifiez votre réponse.
Voici le tableau de seuils de décision :

d.d.l	2	3	4	5	6	7	8	9	10	15
Seuil	5.99	7.82	9.49	11.075	12.59	14.07	15.51	16.92	18.31	24.99

C- ACP

Nous réalisons une analyse en composantes principales sur les variables : maxO3, T9, Ne9, Vx9 et maxO3v. (les résultats sont en Annexe).

a) Le nuage de points est constitué de combien de points ? Combien de dimensions sont nécessaires pour représenter les points ?

b) Comment est définie la quantité d'information (inertie) contenue dans le nuage de points ? A partir de quelle matrice peut-on la calculer et comment la calcule-t-on ?

c) Quel est le pourcentage manquant dans le tableau des valeurs propres ?

d) Combien d'axes faut-il retenir et pourquoi ?

e) Interpréter le graphe des variables.

Annexe

Résultats sur les valeurs propres :

	eigenvalue	percentage of variance
comp 1	3.021	60.422
comp 2	0.879	17.575
comp 3	0.610	[illegible]
comp 4	0.300	5.990
comp 5	0.191	3.822

Résultats sur les variables

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
maxO3	0.858	0.005	0.000	0.004	0.133
T9	0.631	0.141	0.088	0.130	0.010
Ne9	0.544	0.178	0.200	0.061	0.016
Vx9	0.421	0.355	0.175	0.046	0.003
maxO3v	0.567	0.200	0.147	0.058	0.029

Cos^2

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
maxO3	28.396	0.531	0.006	1.498	69.568
T9	20.894	16.033	14.369	43.571	5.134
Ne9	18.021	20.273	32.859	20.290	8.557
Vx9	13.932	40.416	28.726	15.216	1.709
maxO3v	18.757	22.747	24.039	19.425	15.031

Contribution