Projet semestriel

Consignes :

- **Résoudre** les différents exercices à l'aide du **logiciel R sous R-studio**
- Ecrire les **scripts** pour chaque traitement.
- Ne pas utiliser de fonctions R pré-définies pour les exercices 2, 3 et 4.
- Tous les résultats des fonctions à implémenter seront arrondis à 10^{-2} près
- Présenter les résultats dans un **pdf** avec tous les tracés et tableaux de calcul
- **Commenter** et interpréter les résultats lors de la soutenance

Exercice 1 : Cas d'une matrice non diagonalisable

Soit la matrice

$$A = \begin{pmatrix} 0 & 2 & 2 \\ -1 & 2 & 2 \\ -1 & 1 & 3 \end{pmatrix}$$

1. Ecrire une fonction qui demande à l'utilisateur de saisir un entier naturel $p \in \mathbb{N} - \{0, 1\}$. Retourner une matrice random, carrée d'ordre p, d'entiers compris entre 1 et 10, puis afficher toutes les valeurs propres ainsi que leurs multiplicités.
   Appliquer sur A.

2. Donner les vecteurs propres de A pour chaque valeur propre. Pourquoi A n'est pas diagonalisable ? Justifier.

3. Donner la matrice de jordan J de A ainsi qu'une matrice de passage P permettant d'effectuer le changement de bases entre A et J. Calculer P^{-1} et vérifier le changement de base

4. On pose

$$\begin{cases} a_{n+1} = 2b_n + 2c_n \\ b_{n+1} = -a_n + 2b_n + 2c_n \\ c_{n+1} = -a_n + b_n + 3c_n \end{cases}, \text{ avec } a_0 = 1, b_0 = 2, c_0 = -3$$

   Ecrire une fonction qui demande de à l'utilisatuer de saisir un entier naturel non nul n. Puis, à l'aide du changement de base de la question 3, retourner les valeurs a_n, b_n, c_n. Que constate-t-on pour $n \longrightarrow \infty$?

Séries chronologiques (éléments de cours) : **application pour les exos 2 et 3**

C'est une suite de valeurs en fonction du temps de terme X_t, qui se définie par :

1. T_t : la **tendance**, peut être obtenue par ajustement affine (équation de la droite des moindres carrés)
2. S_t : la **composante saisonnière**. Son influence peut être due au climat, coutumes. Elle est périodique de période T sur une année. Par exemple : si les données sont trimestrielles $T = 4$, si mensuelles $T = 12$
3. ε_t : la **composante accidentelle**. Son influence peut être due aux évènements occasionnels, imprévisibles (catastrophe, grèves, épidémie,...)

2 modèles possibles :

1. **addititif** : $X_t = T_t + S_t + \varepsilon_t$, amplitudes d'oscillations constantes
2. **multiplicatif** : $X_t = T_t(1 + S_t) + \varepsilon_t$ amplitudes d'oscillations proportionnelles avec la tendance

Soit $(X_t)_{t\in[\![1,\, n]\!]}$

Moyennes mobiles d'ordre $2p + 1 < n$: $M_t = \dfrac{1}{2p+1}(X_{t-p} + X_{t-p+1} + ... + X_{t+p}),\ t \in [\![p+1,\, n-p]\!]$

Moyennes mobiles d'ordre $2p < n$: $M_t = \dfrac{1}{4p}(X_{t-p} + 2X_{t-p+1} + ... + 2X_{t+p-1} + X_{t+p}),\ t \in [\![p+1,\, n-p]\!]$

Obtention de la **série corrigée des variations saisonnières** : série qui permet de suivre l'évolution du phénomène dans le temps, épuré des mouvements saisonniers de période en période.

Modèle additif :

- On néglige les variations accidentelles, puis on calcule les différences saisonnières $\widehat{S}_t = X_t - T_t$, T_t obtenu par les moyennes mobiles ou ajustement affine (moindres carrés)
- On détermine S_t en prenant la moyenne des $\widehat{S}_t$
- On obtient les **coefficients saisonniers corrigés** $S'_t = S_t - \dfrac{1}{T}\displaystyle\sum_{t=1}^{T} S_t$
- On déduit la **série corrigée des variations saisonnières** $X'_t = X_t - S'_t$

Modèle multiplicatif :

- On néglige les variations accidentelles, puis on calcule les rapports saisonniers $\widehat{S}_t = \dfrac{X_t}{T_t}$, T_t obtenu par les moyennes mobiles ou ajustement affine (moindres carrés)
- On détermine S_t en prenant la moyenne des $\widehat{S}_t$
- On obtient les **coefficients saisonniers corrigés** $S'_t = \dfrac{S_t}{\dfrac{1}{T}\displaystyle\sum_{t=1}^{T} S_t}$
- On déduit la **série corrigée des variations saisonnières** $X'_t = \dfrac{X_t}{S'_t}$

Exercice 2 : Etude de cas (1)

Les ventes trimestrielles entre 2020 et 2024 de dindes vendues dans une boucherie de village sont données dans le tableau suivant :

	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
2019	4	3	5	22
2020	3	4	6	21
2021	2	4	6	24
2022	3	4	7	26

1. Tracer le graphe de cette série chronologique. En déduire le modèle approprié. Justifier.
2. Donner la droite d'ajustement affine par la méthode des moindres carrés. La tracer.
3. Donner les moyennes mobiles d'ordre 4, la tendance et les différences saisonnières de la série chronologique
4. Déterminer les coefficients saisonniers et la série désaisonnalisée
5. Faire un graphe et donner une prévision pour le quatrième trimestre 2025 ?

Exercice 3 : Etude de cas (2)

Les ventes trimestrielles entre 2020 et 2024 d'un produit dans un magasin sont données dans le tableau suivant :

	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
2018	172	298	611	122
2019	254	414	795	198
2020	300	472	903	265
2021	466	568	1115	290
2022	352	624	1274	303

1. Tracer le graphe de cette série chronologique. En déduire le modèle approprié. Justifier.
2. Donner la droite d'ajustement affine par la méthode des moindres carrés. La tracer.
3. Donner les moyennes mobiles d'ordre 4, la tendance et les différences saisonnières de la série chronologique
4. Déterminer les coefficients saisonniers et la série désaisonnalisée
5. Faire un graphe et donner une prévision pour le troisième trimestre 2025 ?

Exercice 4 : Palindromes

1. Créer une fonction qui identifie les mots ou phrases palindromiques (mots ou phrases qui se lisent dans les deux sens, par exemple RADAR). Cette fonction prendra comme argument le mot "mot" et retournera les phrases : "mot est un palindrome" si le mot = palindrome et "mot n'est pas un palindrome", sinon.

2. Appliquer votre fonction sur les mots "radar", "bonne année", "sept", "kayak", "la mariée ira mal", "statistiques", "engage le jeu que je le gagne", "esope reste ici et se repose".

3. Créer un dictionnaire : une liste de 6000 mots constitué de 3, 4, 5, 6, 7, ou 8 lettres random (1000 mots pour chaque catégorie). Créer une fonction qui retourne tous les mots palindromiques dans ce dictionnaire.

Exercice 5 : Salles de cinéma

On mesure pour 94 départements de France métropolitaine les dix variables décrites ci-dessous. De plus, il est fourni la matrice de corrélation des variables ainsi qu'un extrait du tableau de données concernant les 20 départements les plus peuplés.

- `popu` population du département (en millions d'habitants)
- `entr` nombres d'entrées réalisées (en millions)
- `rece` recettes (en millions d'euros)
- `sean` nombre de séances (en milliers)
- `comm` nombre de communes équipées de salles de cinéma
- `etab` nombre de cinémas en activité
- `salle` nombre de salles en activité
- `faut` nombre de fauteuils disponibles
- `artes` nombre de salles d'art et essai (passant des films indépendants)
- `multi` nombre de multiplexes (au moins 8 salles)

Corrélations

	popu	entr	rece	sean	comm	etab	salle	faut	artes	multi
popu	1.00	0.71	0.68	0.77	0.61	0.77	0.85	0.87	0.72	0.80
entr	0.71	1.00	1.00	0.99	0.19	0.76	0.93	0.91	0.70	0.63
rece	0.68	1.00	1.00	0.98	0.15	0.74	0.91	0.89	0.67	0.59
sean	0.77	0.99	0.98	1.00	0.27	0.80	0.96	0.94	0.72	0.70
comm	0.61	0.19	0.15	0.27	1.00	0.75	0.49	0.53	0.64	0.52
etab	0.77	0.76	0.74	0.80	0.75	1.00	0.91	0.91	0.85	0.67
salle	0.85	0.93	0.91	0.96	0.49	0.91	1.00	0.99	0.79	0.79
faut	0.87	0.91	0.89	0.94	0.53	0.91	0.99	1.00	0.80	0.81
artes	0.72	0.70	0.67	0.72	0.64	0.85	0.79	0.80	1.00	0.55
multi	0.80	0.63	0.59	0.70	0.52	0.67	0.79	0.81	0.55	1.00

	depart	popu	entr	rece	sean	comm	etab	salle	faut	artes	multi
D59	Nord	2.555	6.868	37.459	174	35	48	151	34230	18	5
D75	Paris	2.125	30.439	192.244	698	1	92	368	72752	38	5
D13	Bouches du Rhone	1.836	6.651	39.197	193	28	49	155	27488	20	2
D69	Rhone	1.579	6.992	37.359	193	33	52	141	27023	30	3
D62	Pas de Calais	1.442	2.976	15.903	123	23	28	111	22053	10	5
D92	Hauts de Seine	1.429	3.978	21.701	107	33	39	89	20762	24	2
D93	Seine St Denis	1.383	4.803	25.543	127	26	32	97	21168	16	4
D78	Yvelines	1.354	4.625	26.700	130	29	35	95	19302	14	2
D33	Gironde	1.287	5.057	24.555	171	43	52	154	32714	22	5
D76	Seine Maritime	1.239	3.366	19.172	128	23	30	108	23436	11	3
D94	Val de Marne	1.227	4.052	23.967	122	30	36	89	23448	17	2
D77	Seine et Marne	1.194	3.298	19.964	87	30	33	89	19029	10	2
D44	Loire Atlantique	1.134	4.383	22.283	136	39	52	118	25570	26	4
D91	Essonne	1.134	2.242	11.185	84	30	33	77	14414	16	1
D95	Val d'Oise	1.105	2.160	11.581	66	21	22	56	12939	7	1
D38	Isere	1.094	3.993	22.619	142	36	49	133	25116	18	4
D31	Haute Garonne	1.046	4.674	25.114	106	25	32	82	16797	18	2
D67	Bas Rhin	1.026	3.569	20.338	124	13	18	78	16615	8	3
D57	Moselle	1.023	3.187	18.251	88	19	22	79	17866	9	3
D6	Alpes Maritimes	1.011	3.520	21.731	111	23	42	94	16764	8	1

Partie 1 : première approche

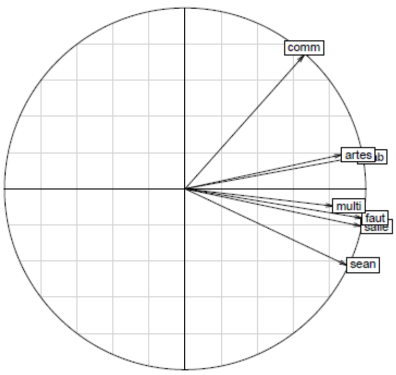
1. A l'aide du tableau des données exo5.txt, calculer la matrice de corrélation des variables et vérifier vos résultats
2. Que pouvez-vous dire à propos des corrélations entre les variables ? Commentez leurs valeurs.
3. En comparant les données brutes de Paris (D75) avec les autres fournies ici, que peut-on dire de ses particularités ?

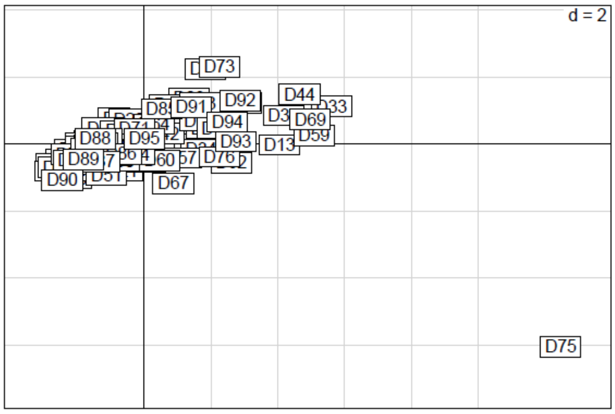
Partie 2 : ACP version1

On s'intéresse pour l'analyse aux variables concernant l'offre de cinéma : sean, comm, etab, salle, faut, artes et multi. On effectue une ACP sur les données centrées-réduites, et on obtient les valeurs propres suivantes

5.47	0.87	0.45	0.17	0.02	0.01	0.01

On donne le cercle des corrélations des variables ainsi que le plan de composantes principales avec les projections des individus :





4. Retrouver les deux tracés du dessus avec le logiciel R
5. Donner une interprétation rapide de la première composante principale à partir uniquement du cercle des corrélations. Que se passe-t-il sur la seconde composante principale ?
6. En remarquant la projection des individus sur le premier plan principal, que peut-on observer ? En s'appuyant sur la question 2, expliquer pourquoi Paris est particulier à la fois sur le premier et le second axe.
7. On se propose de diviser chaque donnée par la population du département popu (c'est-à-dire que la variable entr sera exprimée par habitant, rece en euros par habitant, sean pour 1000 habitants, etc.). Expliquer en quoi cette approche est intéressante.

Partie 3 : ACP version2

On normalise les données comme indiqué à la question précédente. On effectue l'ACP sur les données centrées-réduites sur les nouvelles variables normalisées, mais en utilisant les parts de population comme poids des individus. On obtient les données suivantes : valeurs propres, corrélations avec les quatre premiers axes et, pour une sélection (arbitraire) de 20 départements parmi les 94, les poids des individus, leurs coordonnées sur les 4 premiers axes, ainsi que la qualité de leur représentation par les 4 premiers sous espaces.

Valeurs propres

```
[1] 3.71 2.04 0.75 0.37 0.07 0.04 0.02
```

Corrélations

	Comp1	Comp2	Comp3	Comp4
sean	-0.50	-0.75	-0.41	0.04
comm	-0.68	0.64	0.30	-0.12
etab	-0.91	0.34	0.08	-0.17
salle	-0.93	-0.30	-0.09	-0.09
faut	-0.92	-0.33	-0.01	-0.08
artes	-0.64	0.54	-0.21	0.50
multi	-0.22	-0.68	0.66	0.24

	Weight
D4	0.0024
D5	0.0021
D9	0.0024
D15	0.0026
D23	0.0021
D28	0.0070
D32	0.0030
D38	0.0188
D40	0.0056
D46	0.0027
D48	0.0013
D53	0.0049
D67	0.0176
D73	0.0064
D74	0.0108
D75	0.0365
D80	0.0095
D90	0.0024
D94	0.0211
D95	0.0190

	Axis1	Axis2	Axis3	Axis4
D4	-6.19	3.93	-0.74	0.37
D5	-13.55	4.11	0.43	-3.56
D9	-2.47	3.95	-0.36	1.00
D15	-1.55	3.04	-0.44	0.30
D23	-1.83	3.24	-0.64	1.01
D28	2.56	1.00	-0.50	-0.78
D32	-5.65	5.55	-0.69	2.54
D38	-1.54	-1.16	0.87	0.25
D40	-7.00	1.56	0.90	0.91
D46	-3.44	4.45	-0.69	2.02
D48	-3.01	3.49	0.42	-1.10
D53	-1.55	0.62	1.22	1.59
D67	1.28	-1.53	0.31	0.18
D73	-10.33	1.52	1.79	-2.82
D74	-4.55	-0.75	1.93	0.12
D75	-3.39	-4.30	-2.67	0.11
D80	1.85	1.16	-0.36	-1.12
D90	-2.41	-1.29	-1.40	-2.20
D94	0.53	-0.16	-0.13	-0.05
D95	2.36	0.34	-0.09	-0.57

	Axis1	Axis1:2	Axis1:3	Axis1:4
D4	68.6	96.2	97.2	97.4
D5	85.5	93.4	93.5	99.4
D9	26.1	93.2	93.7	98.1
D15	19.9	95.9	97.5	98.3
D23	21.9	90.2	92.9	99.4
D28	77.8	89.5	92.5	99.8
D32	45.7	89.7	90.3	99.6
D38	51.7	81.0	97.3	98.6
D40	91.4	95.9	97.4	99.0
D46	32.6	87.2	88.5	99.8
D48	39.6	92.6	93.4	98.7
D53	33.7	39.1	59.9	95.4
D67	39.2	95.7	98.1	98.9
D73	88.7	90.6	93.3	99.9
D74	81.6	83.9	98.6	98.7
D75	30.9	80.7	99.9	99.9
D80	55.4	77.4	79.4	99.6
D90	25.4	32.6	41.2	62.5
D94	47.5	51.8	54.6	55.0
D95	92.1	93.9	94.1	99.5

8. Commentez la nouvelle répartition de l'inertie. Combien d'axes principaux retient-on ? La situation est-elle meilleure qu'avec la première analyse ?
9. Quelles sont les variables qui déterminent les axes que l'on retient ? Précisez les critères utilisés. Y a-t-il un effet de taille ?
10. Parmi les départements dont les données sont fournies ci-dessus, quels sont ceux qui déterminent les axes que l'on retient ? Précisez les critères utilisés. Y a-t-il des départements sur-représentés ?
11. Comment peut-on interpréter les axes à partir des deux questions précèdentes ?
12. Parmi les départements dont les données sont fournies ci-dessus, quels sont ceux dont la qualité de représentation est mauvaise sur l'espace propre retenu ? Précisez les critères utilisés.