UTBM – A2018
Examen final - AD50

abderrahim.chariete@utbm.fr

Durée : 2 heures (documents autorisés)

Exercice 1 : (5 pts) Big Data _ Hadoop

Nous considérant le code du mapper – reducer du programme WordCount suivant :

```
1.function map(LongWritable Key1, String Value1)
2. foreach word w in Value1:
3. write(w, 1)

4.function reduce(Text Key2, Iterator<intWritable> Value2)
5. set wordCount = 0
6. foreach v in Value2:
7. wordCount = wordCount + v
8. write(Key2, wordCount)
```

Soit le contenu du fichier en entrée comme suit :

1. ***Celui qui croyait au ciel***
2. ***Celui qui n'y croyait pas***
3. ***Fou qui fait le délicat***
4. ***Fou qui songe à ses querelles***

Nous considérant notre cluster Hadoop configuré pour qu'il dispose de **quatre mappers** et **deux reducers**. Chacun des mappers va travailler sur une partie du fichier en entrée, par exemple : le mapper 1 va traiter la ligne 1, le mapper 2 va traiter la ligne 2, le mapper 3 va traiter la ligne 3 et le mapper 4 va traiter la ligne 4.

Q1. **(1,5 pt)** Donner le contenu du fichier en sortie de la phase map (ensemble clé-valeur)
Q2. **(1,5 pt)** Donner le contenu intermédiaire entre la phase ***map*** et la phase ***reduce*** (après la phase ***shuffle & sorte***)
Q3. **(2 pts)** Proposer une fonction de hachage qui permet de répartir équitablement les ensembles clé-valeur sur les deux reducers. Puis, donner le contenu du fichier de sortie résultant pour chaque recucer.

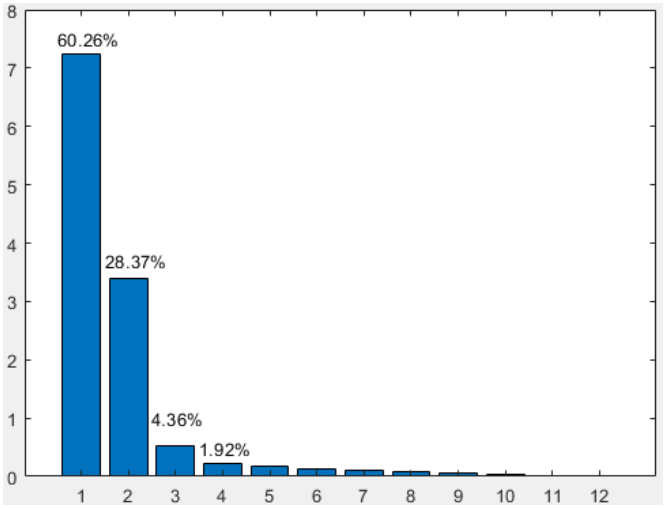
Exercice 2 : (15 pts) ACP, CAH, *k*-means et Silhouettes

Nous souhaitons analyser des données de **pluviométrie** dans les villes françaises :

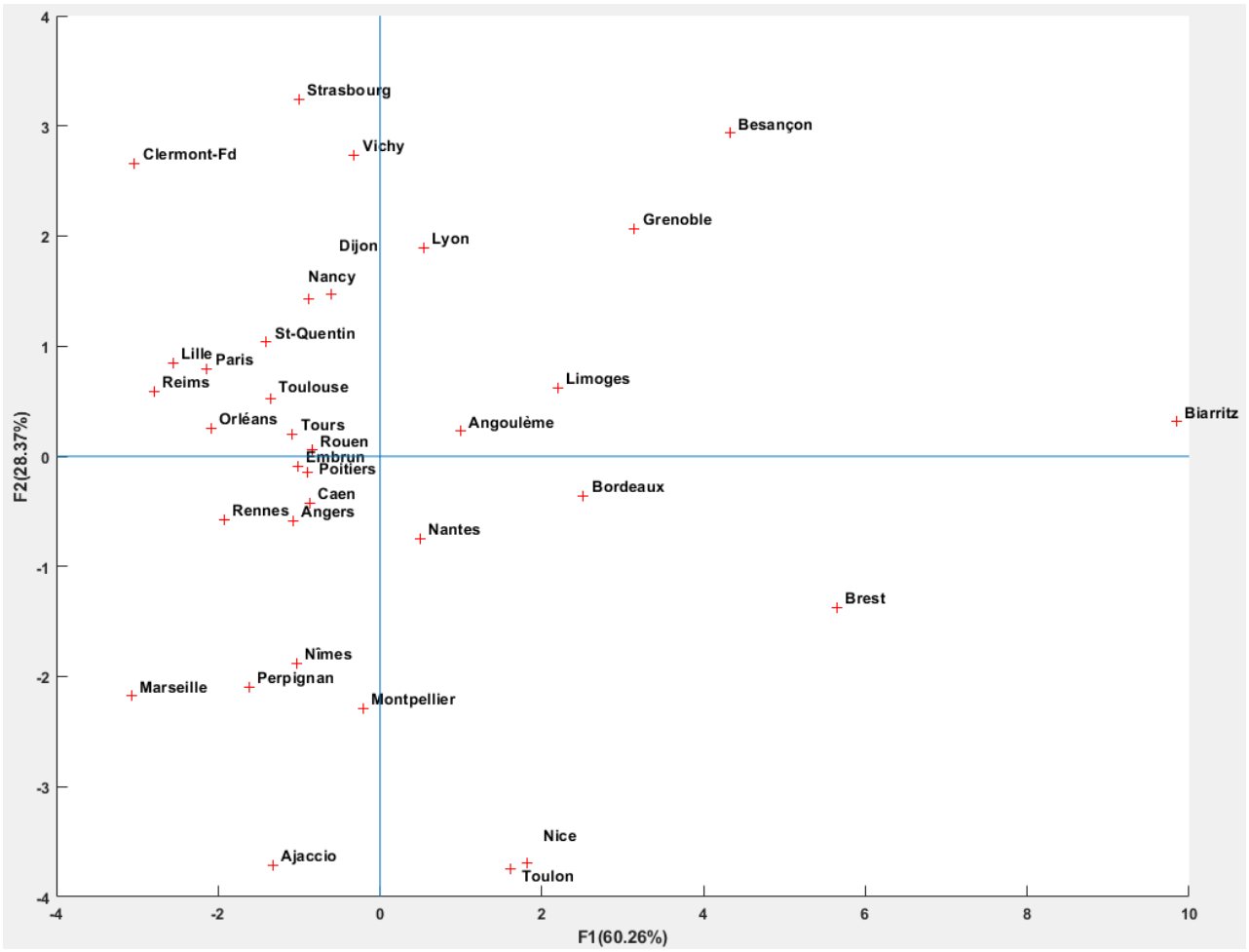
	Jan.	Fév.	Mars	Avr.	Mai	Juin	Juil.	Aout	Sept.	Oct.	Nov.	Déc.	Temp. Moy.	Ampl. Temp.	Lat.	Long.	Région
Besançon	94	87	75	74	86	107	80	116	106	78	92	93	10,04	17,6	47,15	6,02	Est
Clermont-Fd	28	27	30	41	78	79	48	70	58	43	39	30	10,94	16,8	45,47	3,05	Est
Dijon	62	48	51	48	68	79	44	79	74	53	67	61	10,5	18,3	47,19	5,01	Est
Grenoble	80	79	69	69	83	94	74	96	88	85	90	98	10,98	18,6	45,1	5,43	Est
Lyon	53	50	60	54	67	84	55	104	86	73	80	62	11,36	18,6	45,45	4,51	Est
Nancy	66	58	43	45	62	70	58	76	65	52	59	67	9,5	17,5	48,41	6,12	Est
Reims	43	44	42	37	52	53	47	58	54	43	52	50	10,06	16,4	49,15	4,02	Est
Strasbourg	51	44	42	58	71	88	73	90	61	43	51	47	9,72	18,6	48,35	7,45	Est
Vichy	50	45	51	52	84	84	63	86	75	58	58	55	10,72	16,9	46,08	3,26	Est
Caen	65	61	45	44	53	52	45	57	66	75	79	71	10,45	12,7	49,11	-0,21	Nord
Lille	45	43	38	37	45	57	62	64	53	56	56	56	9,73	14,7	50,38	3,04	Nord
Paris	53	48	40	45	53	57	54	61	54	50	58	51	11,18	15,7	48,52	2,2	Nord
Rouen	65	58	50	44	50	57	49	67	70	72	68	66	10,34	14,2	49,26	1,05	Nord
St-Quentin	52	50	46	44	52	63	61	69	67	52	63	65	9,85	16,4	49,51	3,17	Nord
Angers	65	50	60	45	50	55	35	60	55	65	80	70	11,28	14,5	47,28	-0,33	Ouest
Angoulème	79	68	64	62	70	58	53	66	69	70	79	88	12,02	14,9	45,39	0,09	Ouest
Biarritz	128	105	98	102	100	91	69	123	155	152	175	176	13,58	12,3	43,29	-1,34	Ouest
Bordeaux	100	84	66	57	64	71	52	65	88	84	99	117	13,33	15,4	44,5	-0,34	Ouest
Brest	130	98	89	77	74	60	51	80	95	108	136	159	10,77	10,2	48,24	-4,29	Ouest
Limoges	87	75	68	69	72	71	56	73	87	72	82	98	10,59	15,3	45,5	1,16	Ouest
Nantes	83	65	53	48	54	52	42	66	80	77	95	94	11,69	13,8	47,13	-1,33	Ouest
Orléans	57	48	43	46	52	54	47	54	51	54	61	54	10,53	15,7	47,55	1,54	Ouest
Poitiers	65	58	56	49	55	55	46	59	52	61	78	68	11,28	15,1	46,35	0,2	Ouest
Rennes	57	50	45	43	46	48	36	57	53	60	73	66	11,13	13,1	48,05	-1,41	Ouest
Tours	63	55	52	51	53	58	47	60	60	55	68	65	11,22	15,6	47,23	0,41	Ouest
Ajaccio	78	69	51	39	43	23	10	15	43	81	105	96	14,71	14,5	41,55	8,44	Sud
Embrun	61	55	55	48	47	63	41	65	60	60	81	62	9,49	18,4	44,34	6,3	Sud
Marseille	36	49	40	35	38	33	13	27	65	67	69	61	14,23	17,8	43,18	5,24	Sud
Montpellier	56	59	69	46	47	41	20	52	78	125	70	73	13,89	17,1	43,36	3,53	Sud
Nice	67	83	71	70	39	37	21	38	83	109	158	92	14,84	15,2	43,42	7,15	Sud
Nîmes	52	53	57	45	50	40	25	40	75	100	83	60	14,18	17,9	43,5	4,21	Sud
Perpignan	27	52	59	47	49	33	27	28	69	97	70	71	15,24	16,3	42,41	2,53	Sud
Toulon	76	86	82	60	49	35	12	31	77	105	117	107	15,28	14	43,07	5,56	Sud
Toulouse	53	50	52	55	65	65	44	43	57	49	58	65	12,68	16,2	43,36	1,26	Sud

On décide de synthétiser l'information disponible avec l'analyse en composantes principales ACP :

Q1. **(1 pt)** Combien de facteurs (composantes principales) sont nécessaires pour l'analyse. Justifier votre réponse.



Le nuage de points des individus sur le premier plan factoriel (F1-F2) est le suivant :

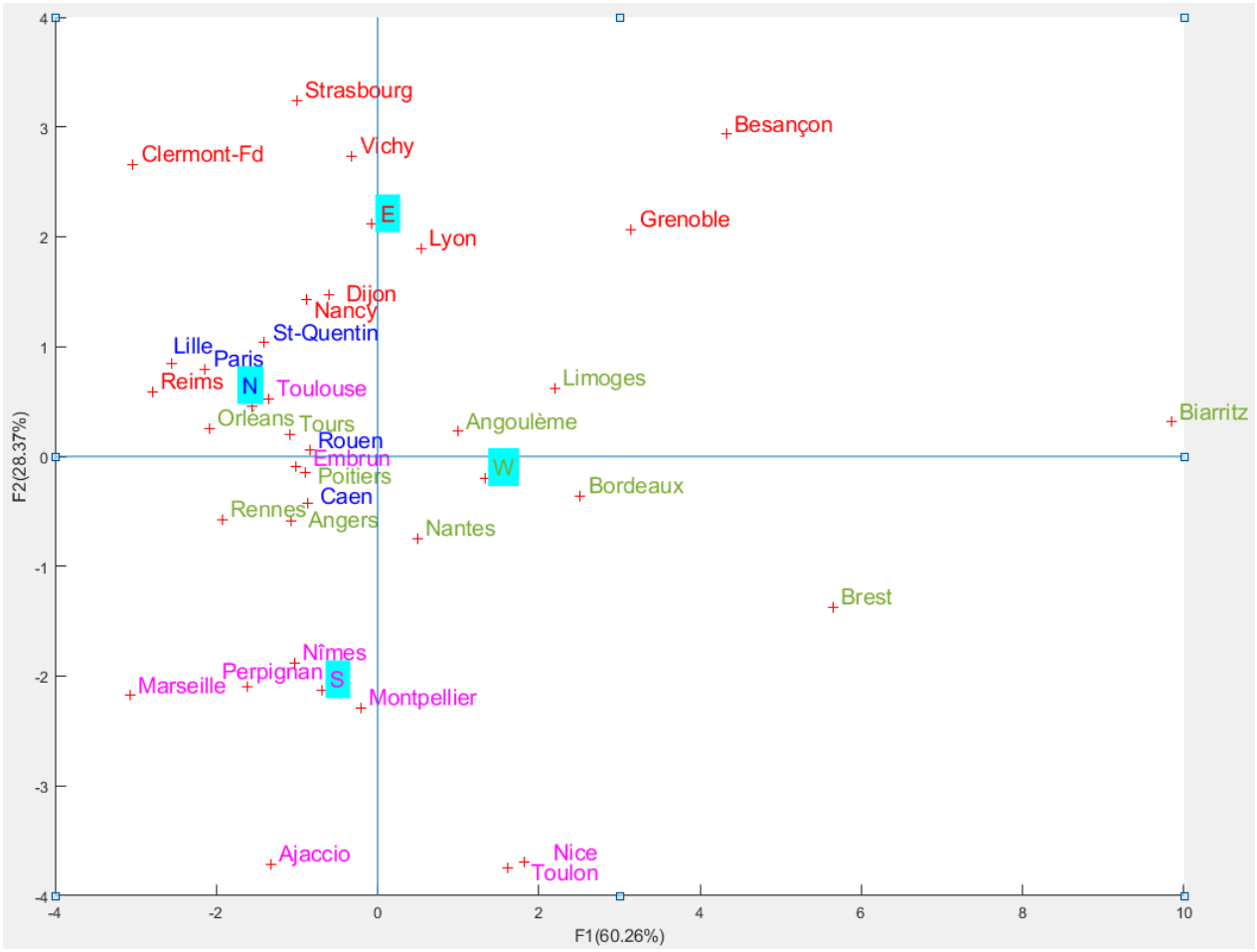


Q2. **(1 pt)** Au vu du premier plan factoriel ci-dessus, quelle interprétation pouvez-vous donner du 1er et 2ème facteur ?

Q3. **(1 pt)** Quelle interprétation pouvez-vous donner aux individus (groupes, cas isolés, corrélations, décorrélation, etc.) ?

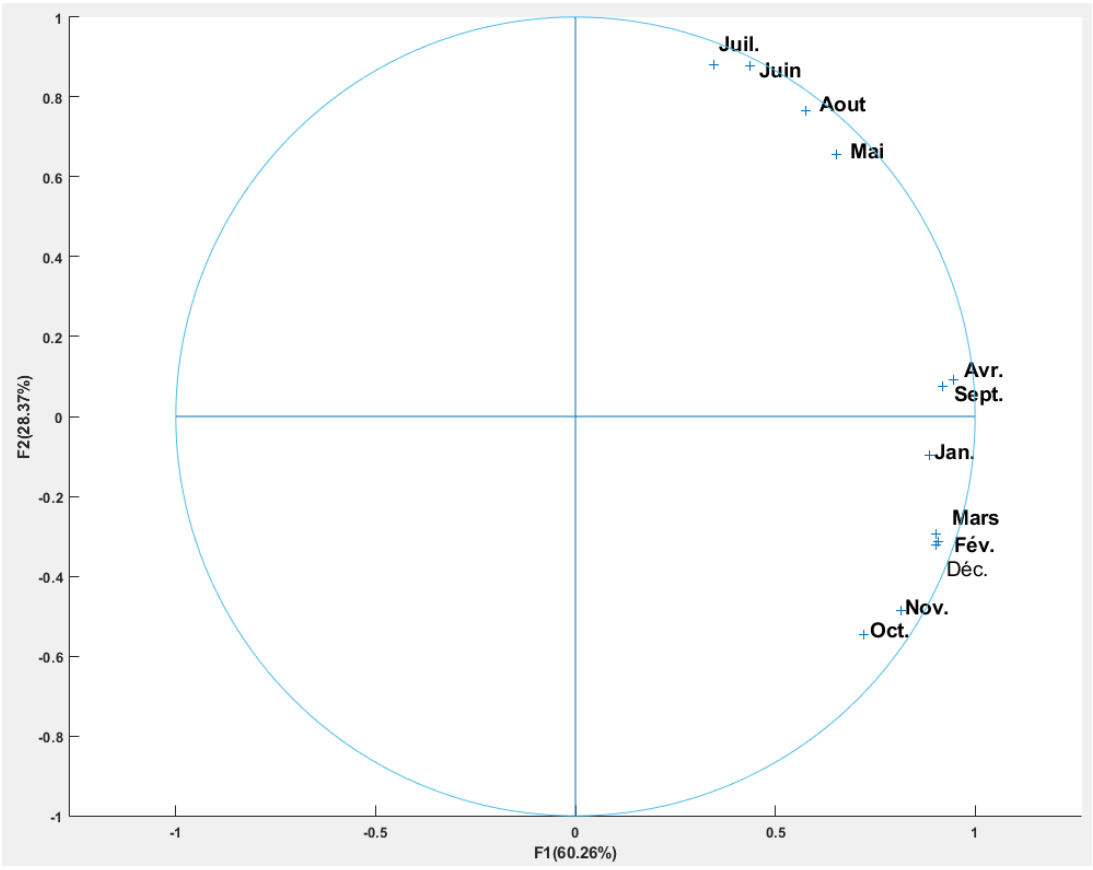
Q4. **(1 pt)** Quel est l'individu le moins bien représenté par le premier plan factoriel ? Quel est l'individu le mieux représenté ?

Nous allons utiliser une information supplémentaire concernant la région de chaque ville. Quarte modalités, ***E***, ***N***, ***W*** et ***S*** pour respectivement les régions ***Est***, ***Nord***, ***Ouest*** et ***Sud***, sont projetées sur le premier plan. Chaque modalité est projetée au centroïde des villes de la région correspondante.



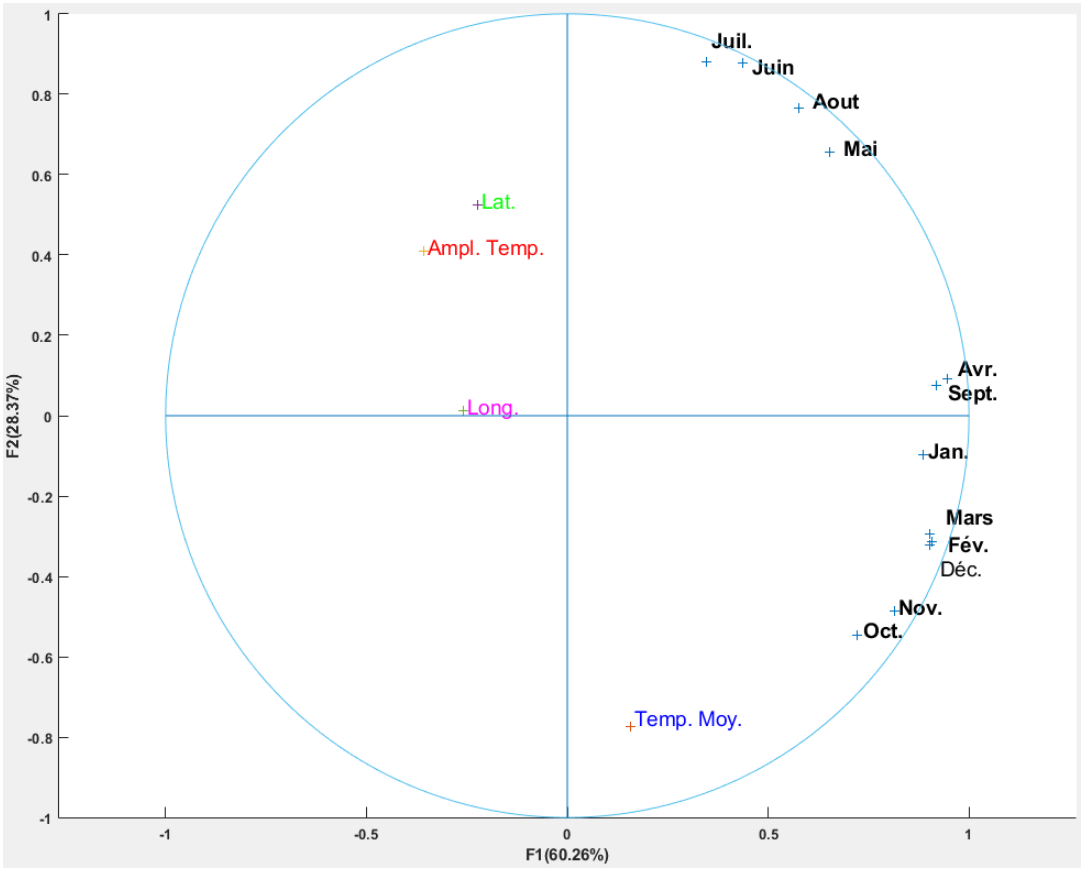
Q5. **(1 pt)** Comment interprétez-vous maintenant le premier plan factoriel ? Qu'est-ce que cette information a apporté à l'analyse ?

Le nuage de points de variables (le cercle de corrélation) de cette analyse est le suivant :



Q6. **(1 pt)** Comment interprétez-vous la participation de variables (les mois) à la construction des axes factoriels F1 et F2 ?

Nous allons utiliser des informations supplémentaires ici aussi. Il s'agit de la moyenne et de l'amplitude annuelles des températures, ainsi que la latitude et la longitude de chaque individu (les villes).



Q7. **(1 pt)** Comment interprétez-vous maintenant le cercle de corrélation des variables ? Qu'est-ce que ces informations ont apporté ?

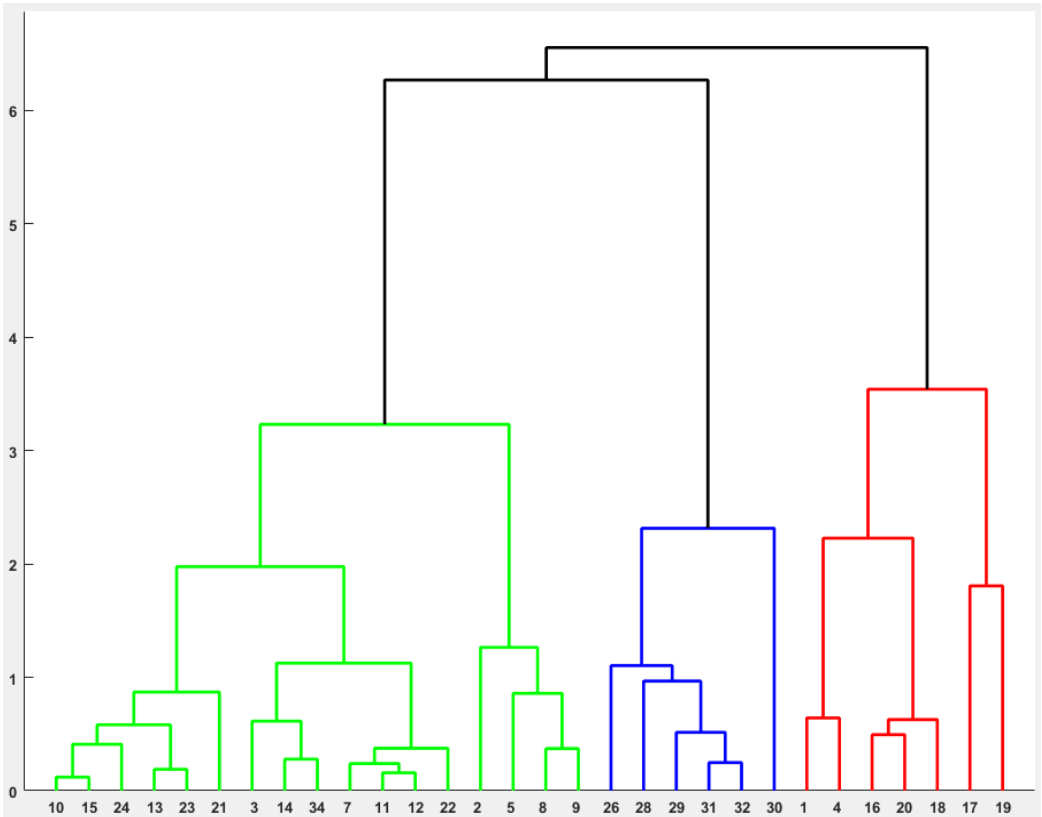
Q8. **(1 pt)** De cette interprétation, quels sont les individus dont la contribution à l'axe factoriel F1 est supérieure à la moyenne ?

Q9. **(1 pt)** Quels sont les individus dont la contribution à l'axe factoriel F2 (2ème composante principale) est de forte amplitude ?

Classification : CAH, k-means et Silhouettes.

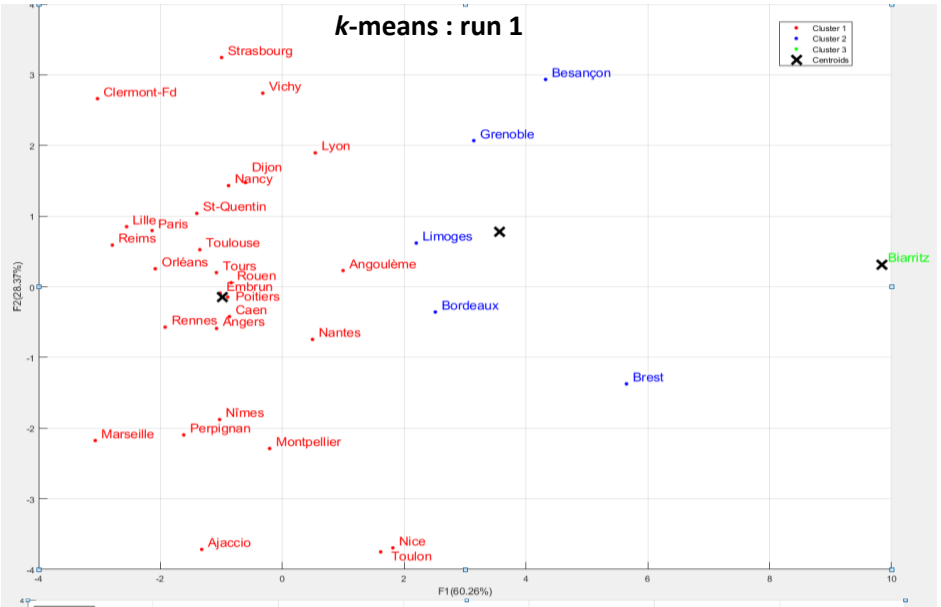
A l'issue de l'ACP, nous souhaitons réaliser une classification en utilisant les coordonnées des individus sur les deux premiers axes factoriels. L'application de la méthode de classification ascendante hiérarchique CAH a donné le dendrogramme suivant :

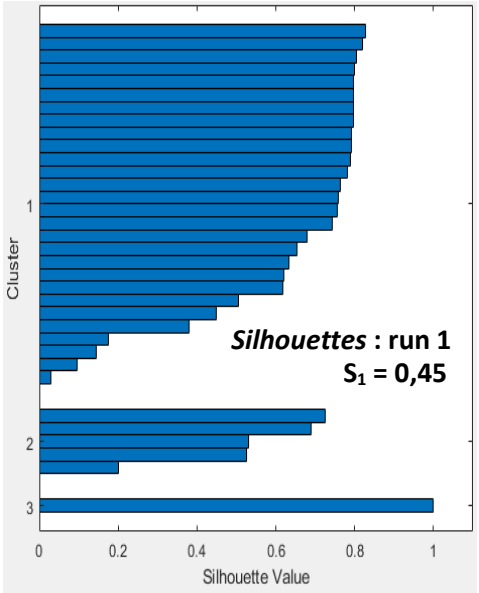
1	Besançon
2	Clermont-Fd
3	Dijon
4	Grenoble
5	Lyon
6	Nancy
7	Reims
8	Strasbourg
9	Vichy
10	Caen
11	Lille
12	Paris
13	Rouen
14	St-Quentin
15	Angers
16	Angoulême
17	Biarritz
18	Bordeaux
19	Brest
20	Limoges
21	Nantes
22	Orléans
23	Poitiers
24	Rennes
25	Tours
26	Ajaccio
27	Embrun
28	Marseille
29	Montpellier
30	Nice
31	Nîmes
32	Perpignan
33	Toulon
34	Toulouse

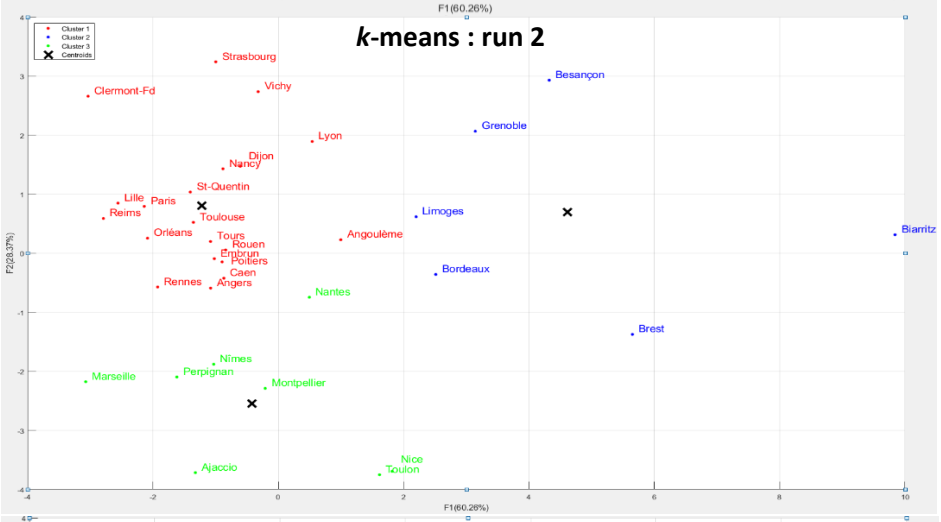


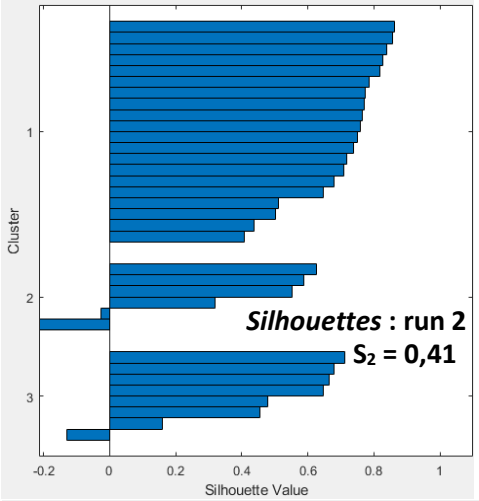
Q10. **(1 pt)** L'examen de ce graphique suggère de réaliser une partition des individus en combien de classes ? Justifier.

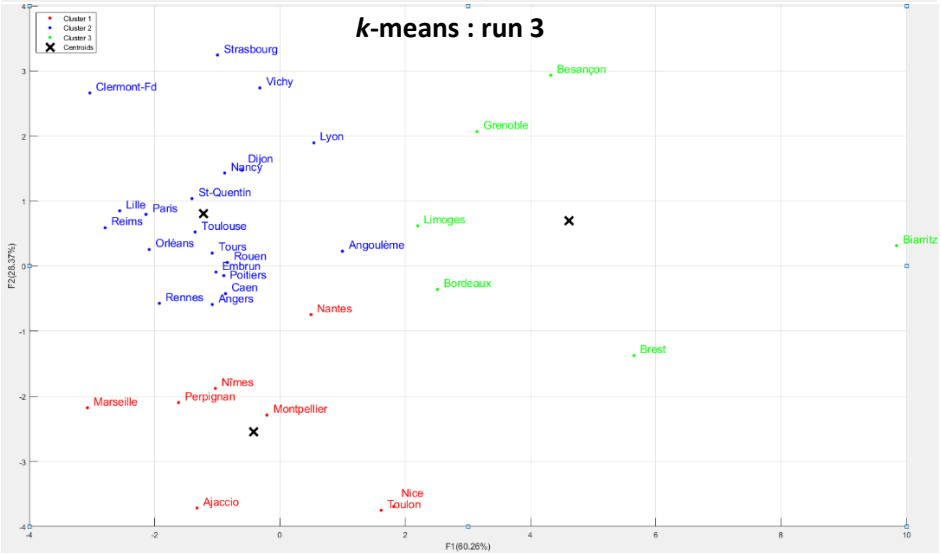
Nous allons réaliser une classification par la méthode *k*-means. Nous répétons l'opération 5 fois tout en évaluons chacune des exécutions par la méthode des silhouettes. Les figures suivantes montrent les résultats de classification de chaque exécution et son évaluation.

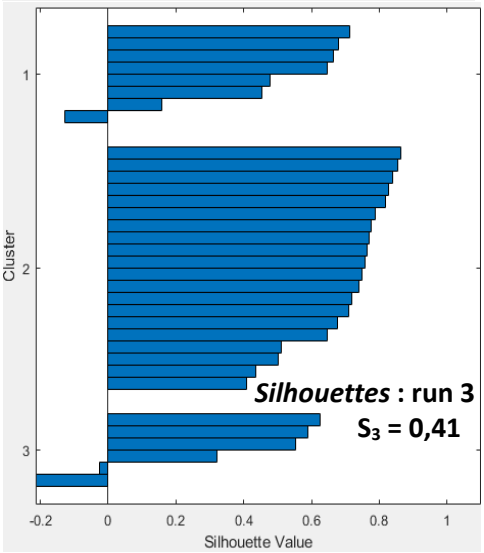


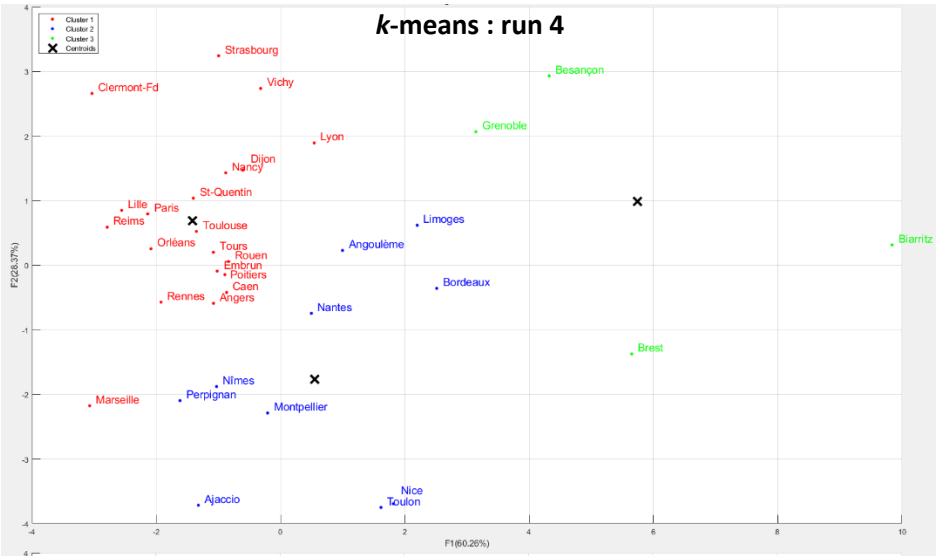


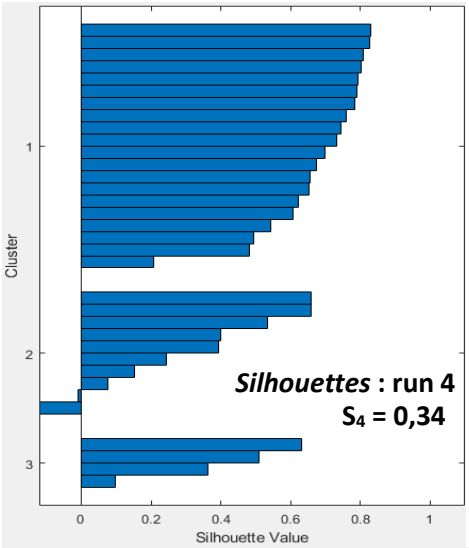


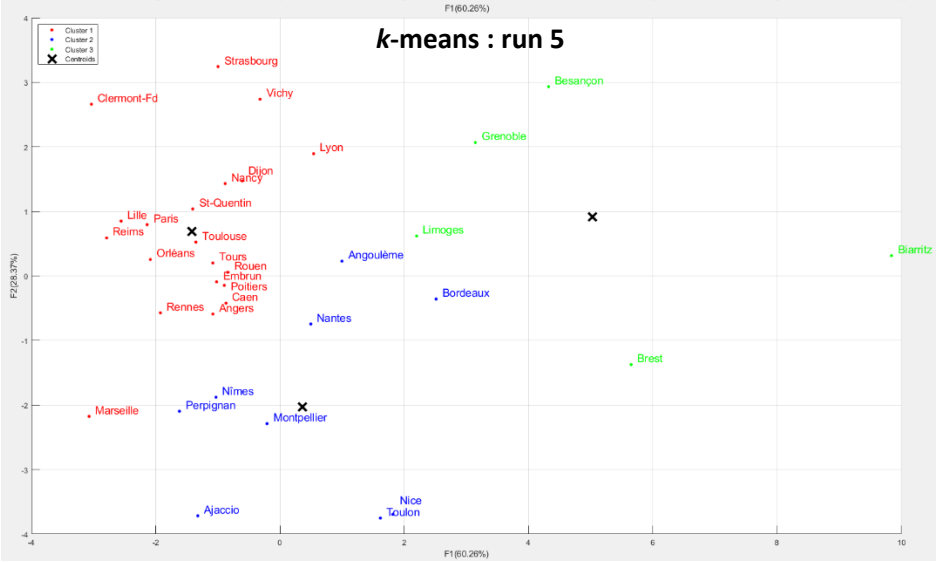


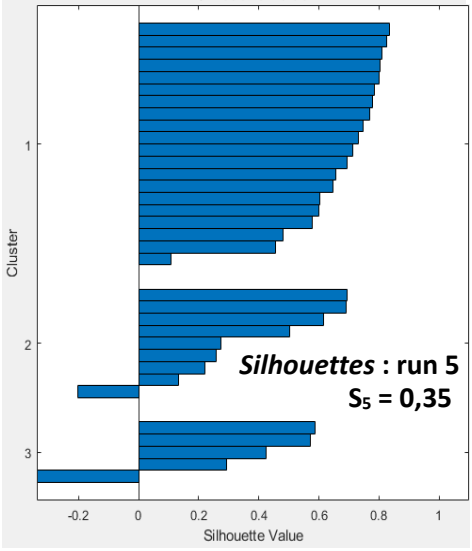












Q11. **(1 pt)** Pourquoi devrons nous réaliser plusieurs exécutions de la méthode *k*-means ?

Q12. **(1 pt)** Selon l'évaluation faite par la méthode des silhouettes, quelle serait la meilleure classification à retenir ? Justifier.

Q13. **(1 pt)** Donner le Parangon de chaque classe. Déduisez les caractéristiques des classes selon leur Parangon.

Q14. **(1 pt)** Quelle sont les points de similarité entre les individus de la même classe et les points de dissimilarité entre les individus de classes différentes ?

Q15. **(1 pt)** Faites une conclusion sur les interprétations de l'ensemble des graphiques.