EISTI	**ING1-GI : DATA EXPLORATION - EXAMEN 2019-2020**
Rédigé par : l'équipe enseignante	Durée : 1h30
A l'intention de : Elèves d'ING1-GI	Document ou matériel autorisés : Calculatrice de l'EISTI uniquement
Nom de l'élève	

Exercice 1 : Croisement quantitatif-quantitatif

Dans une population Ω de taille n, on observe deux variables :

- une variable quanitative, $x = \{x_k\}_{k=1,\dots,n}$ de moyenne $\overline{x}$ et de variance S_x^2
- une variable quantitative, $y = \{y_k\}_{k=1,\dots,n}$, de moyenne $\overline{y}$ et de variance S_y^2 .

Nous nous intéressons à la méthode de la régression linéaire.

1) Rappeler l'expression du coefficient de corrélation linéaire r_{xy}?
2) Quelle interprétation peut-on donner aux valeurs suivantes : $r_{xy} = -1$, $r_{xy} = 0$ *et* $r_{xy} = 1$

Application de la régression linéaire:

Douze personnes sont inscrites à une formation. Au début de la formation, ces stagiaires subissent une épreuve A notée sur 20. A la fin de la formation, elle subissent une épreuve B de niveau identique. Les résultats sont donnés dans le tableau suivant :

Epreuve A	3	4	6	7	9	10	9	11	12	13	15	4
Epreuve B	8	9	10	13	15	14	13	16	13	19	6	19

Avec : Moyenne(A) = 8.5833 , Moyenne(B) = 12.9166, Variance(A) = 13.5763 , Variance (B) = 15.4097 , Covariance(A,B) = 1.4652

3) On cherche à expliquer la note B à partir de la note A.
   a. Déterminer la droite de régression linéaire (en rappelant les formules utilisées pour le calcul).
   b. Rappeler et calculer le coefficient de détermination.
   c. Commenter.

4) Deux stagiaires semblent se distinguer des autres.
   a. Dire de quels points s'agit-il ?
   b. En supprimant les 2 points atypiques, on obtient : Moyenne(A) = 8.4 , Moyenne(B) = 13, Variance(A) = 10.04, Variance (B) = 10 et Covariance(A,B) = 9. Déterminer la droite de regression sur les 10 points restants.
   c. Calculer le coefficient de détermination et commenter.

Exercice 2 : Croisement qualitatif-qualitatif

Un organisme effectue une enquête d'opinion avec de la question suivante : « on en a assez se ceux qui bloquent la vie du pays par leurs revendications ». Les réponses possibles sont :

1. Pas du tout d'accord
2. Pas tellement d'accord
3. Peut-être d'accord
4. Bien d'accord
5. Entièrement d'accord

Tendance politique/ Réponse	1	2	3	4	5	Total
Extrême gauche	10	1	0	2	1	14
Gauche	134	102	94	82	60	472
Centre	22	27	58	85	62	254
Droite	5	27	49	85	148	314
Extrême droite	1	1	0	3	9	14
Indifférent	15	25	51	63	55	209
Non-réponse	17	24	52	55	45	193

Nous nous intéressons à un éventuel lien entre la réponse et la tendance politique.

1) De quel type de variables s'agit-il et quel type de test nous permettrait d'estimer le degré d'indépendance entre elles ?

2) Calculer les valeurs suivantes : $n_{\mathbf{1},\mathbf{3}}$, $n_{\mathbf{3},.}$ et $f_{.,3}$

3) Quel est le nombre de degrés de liberté d.d.l

Le tableau des profils lignes est donné par

Tendance politique	1	2	3	4	5	TOTAL
Extrême gauche	714	71	0	143	71	1 000
Gauche	284	216	199	174	127	1 000
Centre	87	106	228	335	244	1 000
Droite	16	86	156	271	471	1 000
Extrême droite	71	71	0	214	643	1 000
Indifférent	82	120	244	301	263	1 000
Non-réponse	88	124	269	285	233	1 000

4) Expliquer pour 2 lignes de votre choix comment sont obtenus ces profils ?

Après calcul de la distance du Chi2, nous obtenons la valeur suivante : $\chi^2 = 319.48$.

d.d.l	18	19	20	21	22	23	24	25
Seuil de décision	28.869	30.144	31.41	32.671	33.924	35.172	36.415	37.652

5) Rappeler la formule du χ^2 . Que peut-on conclure quant à l'indépendance ou la dépendance entre nos deux variables ?

Exercice 3 : Analyse en Composantes Principales

Dans le cadre d'une étude visant la description d'une série de véhicules (pas tout à fait récents), une ACP a été réalisée. Le but de ce qui suit est d'analyser et d'interpréter les résultats obtenus. Les données utilisées sont les suivantes :

MODELE	CYL	PUISS	LONG	LARG	POIDS	V-MAX	FINITION	PRIX	R-POIDS.PUIS
Alfasud TI	1350	79	393	161	870	165	2_B	30570	11.01
Audi 100	1588	85	468	177	1110	160	3_TB	39990	13.06
Simca 1300	1294	68	424	168	1050	152	1_M	29600	15.44
Citroen GS Club	1222	59	412	161	930	151	1_M	28250	15.76
Fiat 132	1585	98	439	164	1105	165	2_B	34900	11.28
Lancia Beta	1297	82	429	169	1080	160	3_TB	35480	13.17
Peugeot 504	1796	79	449	169	1160	154	2_B	32300	14.68
Renault 16 TL	1565	55	424	163	1010	140	2_B	32000	18.36
Renault 30	2664	128	452	173	1320	180	3_TB	47700	10.31
Toyota Corolla	1166	55	399	157	815	140	1_M	26540	14.82
Alfetta-1.66	1570	109	428	162	1060	175	3_TB	42395	9.72
Princess-1800	1798	82	445	172	1160	158	2_B	33990	14.15
Datsun-200L	1998	115	469	169	1370	160	3_TB	43980	11.91
Taunus-2000	1993	98	438	170	1080	167	2_B	35010	11.02
Rancho	1442	80	431	166	1129	144	3_TB	39450	14.11
Mazda-9295	1769	83	440	165	1095	165	1_M	27900	13.19
Opel-Rekord	1979	100	459	173	1120	173	2_B	32700	11.20
Lada-1300	1294	68	404	161	955	140	1_M	22100	14.04

1) Quels sont les individus visés par l'étude, combien y-a-t-il de variables et d'observations ?

Individu :
Nombre de variables :
Nombre d'observations

2) Précisez la nature de chacune des variables ci-dessous. Peut-on avoir plusieurs possibilités pour certaines d'entre elles ?

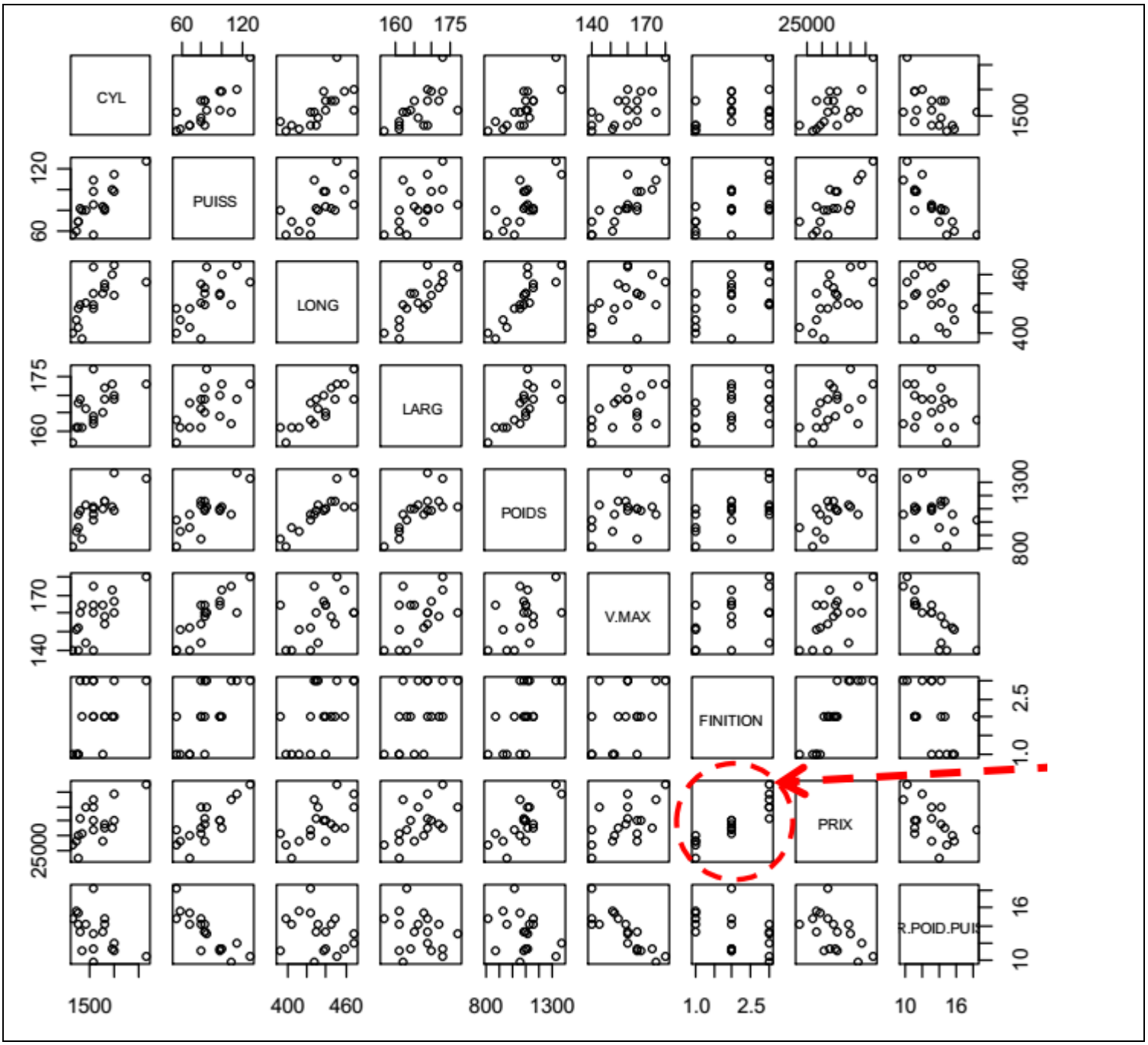
CYL :

PUISS :

LONG :

POIDS :

FINITION :



3)

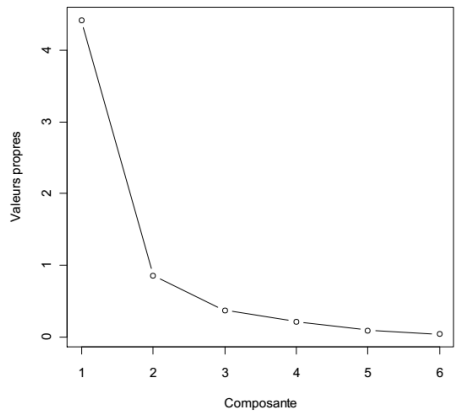
a. Que représente le tableau de graphiques ci-dessus ?
b. Est-ce qu'il y a un intérêt à tracer la variable FINITION, pointée par la flèche ?

Avant de vous présenter les résultats de l'ACP,

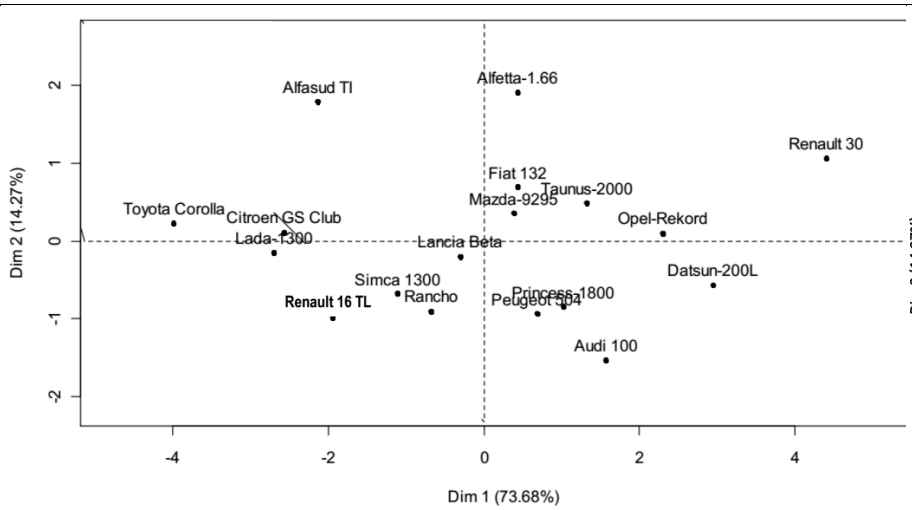
4) Rappeler en quoi consiste cette méthode d'analyse de données ?

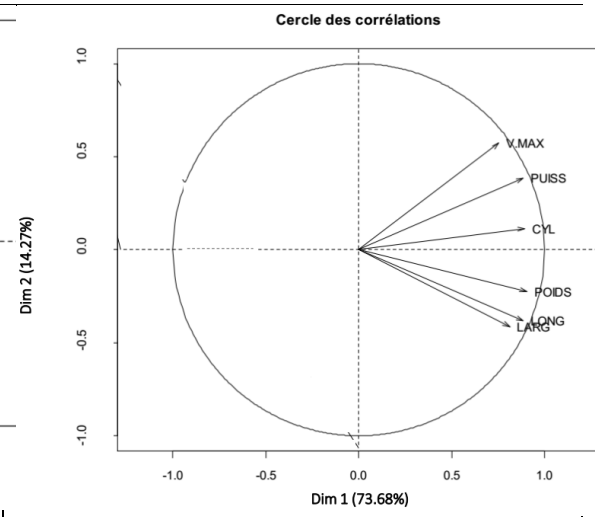
Nous opérons une ACP sur les variables : CYL, PUISS, LONG, LARG, POIDS, V-MAX et obtenons les résultats suivants

```
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4     Comp.5      Comp.6
Standard deviation     2.1025837 0.9252363 0.61079135 0.46251712 0.30463291 0.208063146
Proportion of Variance 0.7368097 0.1426770 0.06217768 0.03565368 0.01546687 0.007215045
Cumulative Proportion  0.7368097 0.8794867 0.94166440 0.97731809 0.99278495 1.000000000
```



5) D'après ces résultats,
   a. A quoi correpondent les valeurs propres représentées sur la figure ci-dessus ?
   b. Combien d'axes faut-il retenir et pourquoi ?
   c. A combien s'èlève le pourcentage de variance cumulée pour ces axes retenus ?
   d. Combien vaudrait le pourcentage de variance cumulée si on retenait tous les axes?

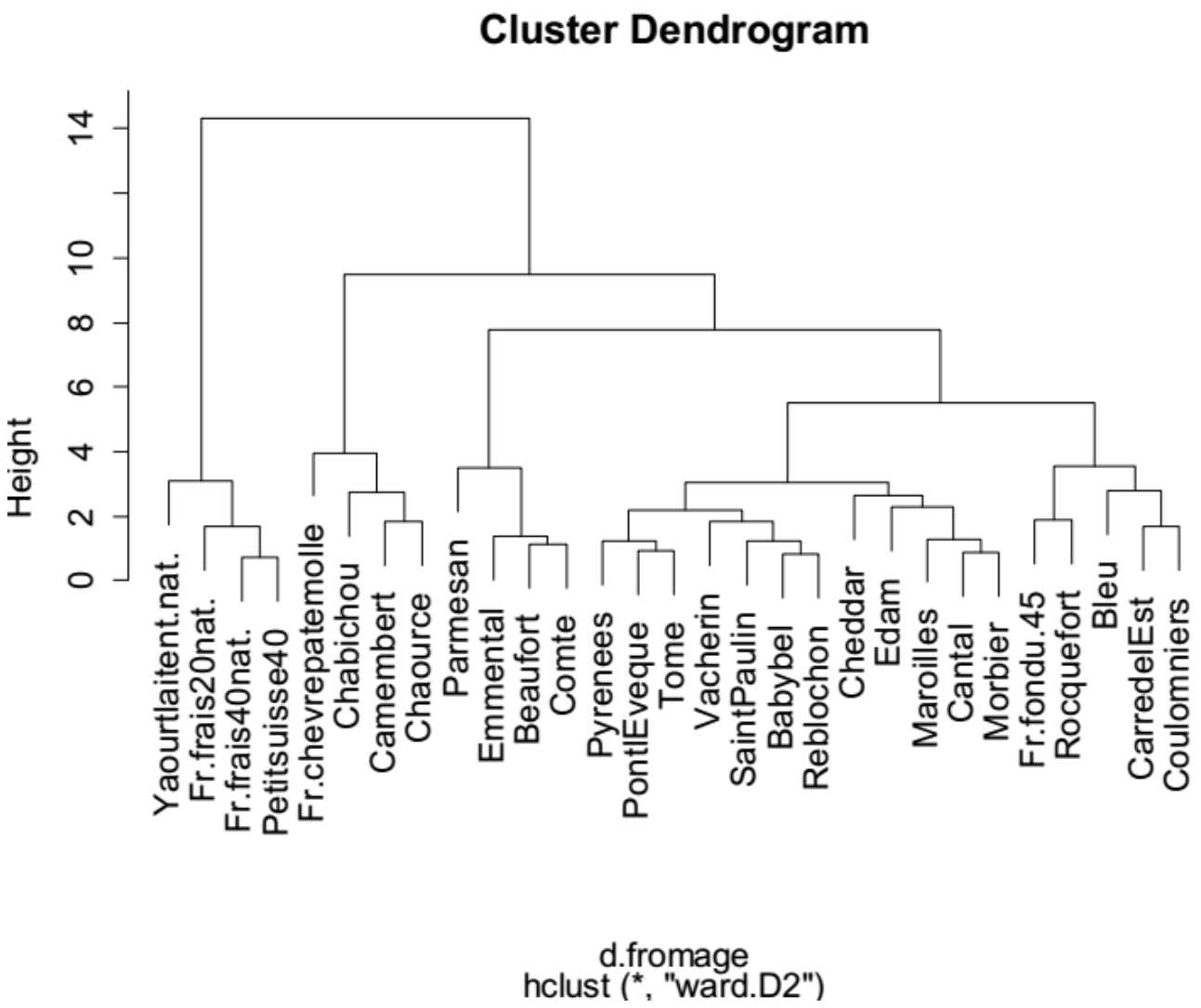




6) Que représentent les deux graphiques ci-dessus et quelles sont les interprétations, concernant les variables étudiées, que nous pouvons en faire?

Exercice 4 : Classification

Nous nous intéressons à une classification de fromages. La figure suivante est le seul résultat d'étude dont nous disposons. Nous savons également que le Camembert est un fromage à pâte molle et que le Beaufort est à pâte dure.



1) De quelle méthode s'agit-t-il et quelle est la distance utilisée?

2) Quels découpages peut-on envisager d'après ce dendrogramme ? Expliquer votre choix et décrivez les classes qui en découleraient. Penser également à tracer vos découpages sur le dendrogramme avec des couleurs différentes.

3) Comment pourrait-on exploiter ces résultats avec l'algorithme des K-means?

4) Quel est le critère de convergence du K-means ?