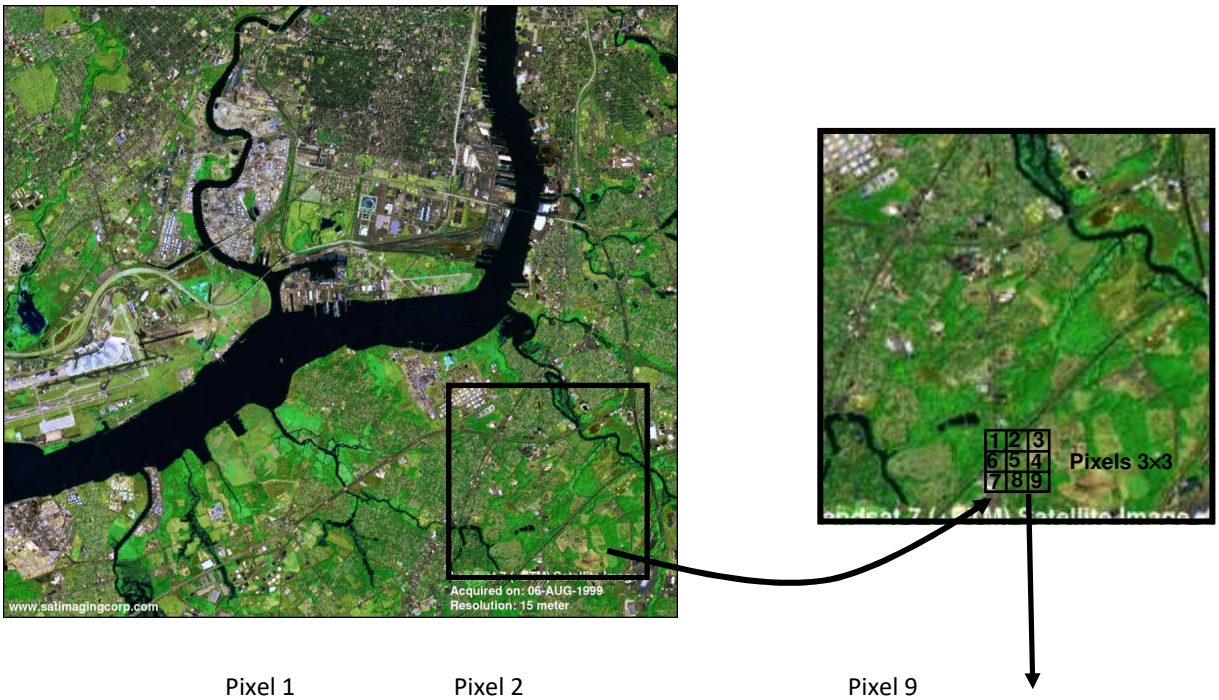
EISTI	**ING1-GI** **EXAMEN DE DATA EXPLORATION 2018-2019**
Durée : 2h30	Exammanager + Examen papier

L'examen est constitué de **deux parties** (donc deux rendus) :

- Un QCM à remplir directement sur papier. Pour répondre aux questions du QCM, vous aurez besoin des annexes en fin de sujet. Attention, il est à points négatifs.
- Un examen sur RStudio à faire sous Exammanger sur le jeu de données `Landsat.txt`. Vous trouverez un descriptif du sujet et les questions ci-dessous. Pour cette partie, vous devez rendre un document pdf contenant :
  - Les lignes de codes utilisées
  - Les graphiques obtenus
  - Les réponses aux questions
  - Et le cas échéant les screenshot demandés

Examen RStudio



	Pixel 1				Pixel 2						Pixel 9				
	x.1	x.2	x.3	x.4	x.5	x.6	x.7	x.8	x.9	…	x.32	x.33	x.34	x.35	x.36	classes
Pixels 3x3_1	92	115	120	94	84	102	106	79	84	…	100	84	107	113	87	grey soil
Pixels 3x3_2	76	94	98	76	76	98	102	72	76	…	71	79	87	93	67	damp grey soil
Pixels 3x3_3	88	107	113	85	88	103	108	85	88	…	81	86	100	104	81	grey soil
Pixels 3x3_4	46	32	124	139	46	32	124	139	42	…	139	44	31	125	139	cotton crop
…	…									…						…

Une image d'un satellite Landsat MSS consiste en quatre images numériques de la même scène dans différentes bandes spectrales. Deux d'entre elles se trouvent dans la région visible (correspondant approximativement aux régions vertes et rouges) et deux se situent dans l'infrarouge. Chaque pixel

est un mot binaire de 8 bits tel que 0 correspond au noir et 255 au blanc. La résolution spatiale d'un pixel est d'environ 80m x 80m. Chaque image contient 2340 x 3380 pixels.
La base de données est une (minuscule) sous-zone d'une scène, composée de 82 x 100 pixels.
Chaque ligne de données correspond à un voisinage 3x3 de pixels entièrement contenu dans la sous-zone.
Chaque ligne contient les valeurs des pixels dans les quatre bandes spectrales (converties en ASCII) de chacune des 9 pixels dans le voisinage 3x3 et un nombre indiquant l'étiquette de classification du pixel du centre.
Les classes sont:

red soil - cotton crop - grey soil - damp grey soil -
soil with vegetation stubble - very damp grey soil

Chaque ligne du tableau de données correspond à un ensemble de 3x3 pixels. Les colonnes correspondent aux valeurs des neufs pixels sur les quatre bandes spectrales. Les quatre premières colonnes correspondent aux quatre bandes spectrales du pixel en haut à gauche, puis les quatre suivantes pour le pixel haut-milieu, puis celles pour le pixel haut-droit, et ainsi de suite. Les pixels sont lus dans l'ordre, de gauche à droite et de haut en bas.

1) Lire le jeu de données à l'aide de l'instruction `read.table`.
   Quelle est la dimension dans laquelle évolue le nuage de points ?
   Combien y-a-t-il d'observations ?
2) Faire une représentation graphique appropriée pour la variable classes.
3) Tracer la boxplot de la variable x.1. Quelle est la médiane ? Au-dessus de quelle valeur de x.1 trouve-t-on 75% des observations ? Y-a-t-il des observations atypiques ?
4) On considère maintenant la variable x.1 pour la classe cotton crop. On obtient les quantiles suivants :

0%	25%	50%	75%	100%
39	44	46	51	89

   Tracer les boxplots de la variable x.1 par classe avec l'instruction `boxplot(x.1~classes,data=tab)`.
   La valeur moyenne de x.1 dans la classe cotton crop est

   plus grande que 46 environ 46 plus petite que 46 on ne peut pas savoir

5) Effectuer une ACP avec la fonction `PCA` du package `FactoMineR` en ajoutant la classe comme variable supplémentaire. Vous pourrez utiliser la fonction `explor` du package du même nom pour un affichage des résultats.
6) Combien d'axes doit-on retenir pour l'étude des résultats ? Justifiez votre réponse.
7) Combien observe-t-on de groupes de variables ? Pouvez-vous expliquer ces résultats par rapport à la nature des variables de ce jeu de données ?
8) Quels sont les groupes qui sont liés et les groupes qui sont non corrélés ? Expliquer ce résultat.
9) Quel groupe contribue le plus à la construction de l'axe 1 ? de l'axe 2 ?
10) Quelle est la contribution moyenne d'une observation ? Y-a-t-il des observations ayant une contribution anormalement importante sur l'axe 1 et l'axe 2 ?
11) Représenter les individus sur les axes 1 et 2 avec une couleur par classe (option de couleur dans la fenêtre graphique des individus de `explor`).
    Que pouvez-vous dire sur les individus ayant une forte contribution à la construction de l'axe 2 ?
    Que pouvez-vous dire sur les autres classes ?
12) Centrer et réduire les variables avec la fonction `scale`.

13) Construire la matrice des distances euclidiennes entre les observations centrées-réduite avec la fonction `dist`.

14) Effectuer une classification hiérarchique ascendante avec la fonction `hclust` et la distance `ward.D2`. Tracer le dendrogramme.

15) Tracer l'évolution du critère de dissimilarité (`height`) en fonction des itérations. Combien de clusters retenez-vous ?

16) Effectuer un kmeans (fonction `kmeans`) avec le nombre de clusters retenus précédemment. Quel pourcentage de variabilité des données est expliqué par ce partitionnement ? Vous ferez un screenshot du résultat pour justifier votre réponse.

17) Effectuer un tableau croisé des effectifs entre les clusters obtenus et les classes. Est-ce que certains clusters correspondent à des classes ? Vous ferez un screenshot du résultat pour justifier votre réponse.

ANNEXE 1

	Action	Animation	Aventure	Comédie	Drame
=<2	1	2	1	7	1
3-10	2	5	7	3	1
11-29	1	1	3	1	3
>=30	2	2	3	2	2
Tot	6	10	14	13	7

Tableau des effectifs observés

	Action	Animation	Aventure	Comédie	Drame	Tot
=<2	0,02	0,04	0,02	0,14	0,02	0,24
3-10	0,04	0,1	0,14	0,06	0,02	0,36
11-29	0,02	0,02	0,06	0,02	0,06	0,18
>=30	0,04	0,04	0,06	0,04	0,04	0,22
Tot	0,12	0,2	0,28	0,26	0,14	1

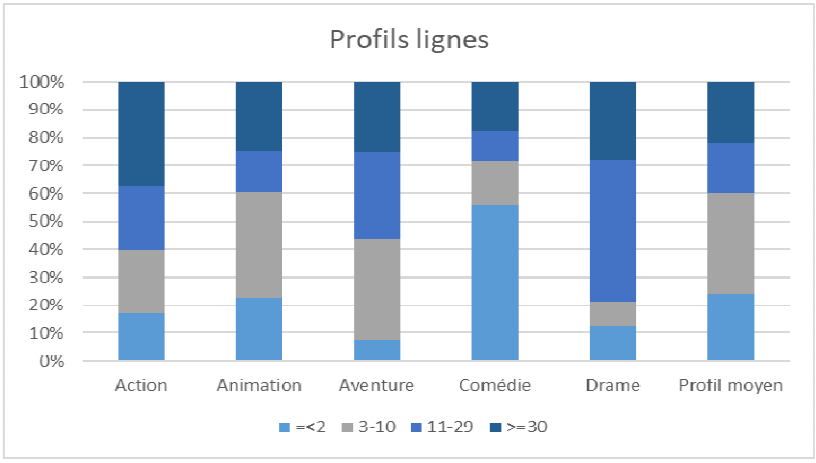
Tableau des fréquences observées

	Action	Animation	Aventure	Comédie	Drame
=<2	0,08	0,17	0,08	0,58	0,08
3-10	0,11	0,28	0,39	0,17	0,06
11-29	0,11	0,11	0,33	0,11	0,33
>=30	0,18	0,18	0,27	0,18	0,18

Profils lignes

	Action	Animation	Aventure	Comédie	Drame
=<2	0,17	0,20	0,07	0,54	0,14
3-10	0,33	0,50	0,50	0,23	0,14
11-29	0,17	0,10	0,21	0,08	0,43
>=30	0,33	0,20	0,21	0,15	0,29

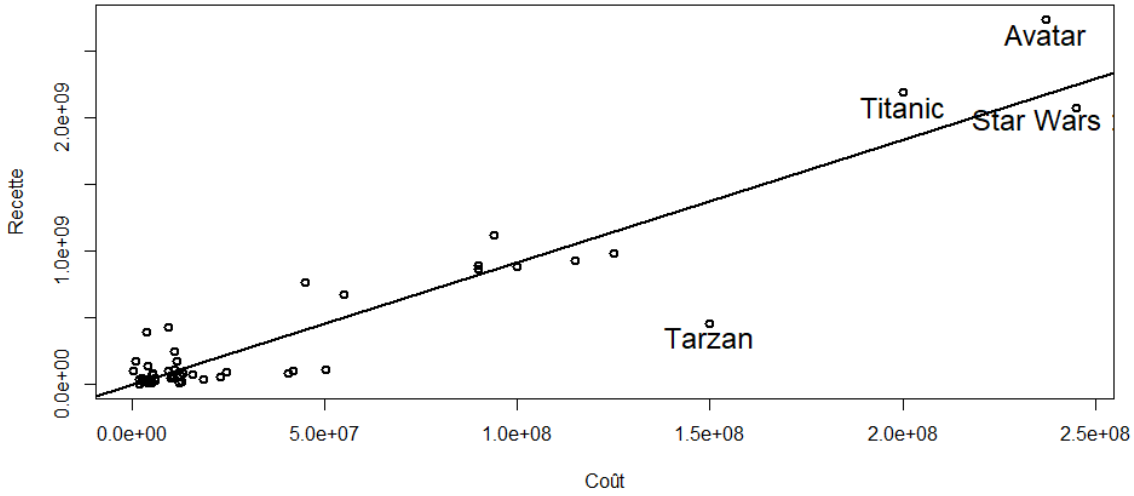
Profils colonnes



d.d.l	10	12	14	16	18	20
Seuil de décision	18,31	21,03	23,68	26,30	28,87	31,40

Tableau des seuils de décision

ANNEXE 2




```
Call:
lm(formula = Recette ~ Cout, data = tab)

Residuals:
       Min         1Q     Median         3Q        Max
-914345403  -61539077   -7564336   61577047  562410402

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.316e+06  3.689e+07  -0.171    0.865
Cout         9.177e+00  5.144e-01  17.840   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 218200000 on 48 degrees of freedom
Multiple R-squared:  0.8689,    Adjusted R-squared:  0.8662
F-statistic: 318.3 on 1 and 48 DF,  p-value: < 2.2e-16
```