UNIVERSITÉ DE SHERBROOKE

Faculté des sciences

Département d'informatique

IFT799 – Science des données

Rapport

TP1 : Visualisation des données

JEU DE DONNÉES ABR

RÉALISÉ PAR:

KAOUTAR L'HASSNAOUI 22 148 702

ZINEB EL YAMANI 22 142 129

ENCADRÉ PAR: SHENGRUI WANG

Choix du sujet:

Pour ce TP1, nous avons opté pour le sujet qui porte sur la classification des opinions à partir du jeu de données Amazon Book Review.

Notre choix fut basé sur la lucidité des explications et ainsi, notre compréhension du sujet et des données.

Choix du jeu de données:

Pour explorer et visualiser nos données, on a choisi un échantillon de données brutes.
En effet, cela nous permettra de bien comprendre nos individus et d'explorer leurs distributions, ainsi que les ressemblances et les liaisons entre variables.

Description du jeu de données:

Lignes : 50 000 lignes

Colonnes :

1. **reviewerID** - identifient de l'internaute
2. **asin** - identifiant du livre
3. **reviewerName** - Nom de l'internaute
4. **helpful** - mesure l'obligeance de l'évaluation de l'internaute
5. **reviewText** - texte exprimant l'opinion de l'internaute
6. **overall** - score du livre donné par l'internaute
7. **summary** - texte résumant l'opinion de l'internaute
8. **unixReviewTime** - temps (sous format UnixTime) fromat quand l'internaute a donné son évaluation
9. **reviewTime** - temps (sous format brute) quand l'internaute a fait son évaluation

Objectif :

Pour ce TP1, il est question de partir à l'exploration de nos données et de les visualiser. Pour cela on suivra les 3 étapes suivantes:

(a).Entre deux livres quelconques, lequel est plus apprécié ?

Dans un premier temps, on va négliger la durée d'évaluation pour construire une matrice de 5 lignes (en fonction du score : s=1,2,3,4, ou,5) et autant de colonnes qu'il y a de livres. Cette matrice va nous permettre d'évaluer les livres entre eux grâce aux calculs de mesures (somme totale, moyenne, moyenne pondérée, écart-type, médiane, quartiles, max et min).

(b).Analyse en Composantes principales sans le facteur temps

Toujours, avec la même matrice vue dans la première partie, on fera une analyse en composantes principales, puis une représentation du nuage de points projetés en 3 groupes : moins appréciés, plus-ou-moins appréciés, et les plus appréciés.

Enfin, nous ferons une analyse mensuelle. Ainsi, nous diviserons notre matrice en 12 matrices pour refaire notre ACP.

(c).Analyse de la tendance mensuelle

Il s'agit de faire intervenir le facteur temps dans notre analyse. Dans (a) et (b), nous avons fait une analyse statique en supposant que les livres ont tous été évalués en même temps. Ce qui n'est pas le cas en réalité. Ainsi, ici nous ajouterons à notre analyse une nouvelle granularité. En effet, nous allons nous ferons une exploration en divisant notre matrice en 12 matrices (12 mois). En appliquant l'ACP pour chaque matrice, nous ferons une analyse mensuelle.

22 148 702 et 22 142 129

(a)

Etapes préliminaires :

- Recherche de doublons (0 doublon)
- Types des variables

```
Data columns (total 9 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   reviewerID      50000 non-null  object
 1   asin            50000 non-null  object
 2   reviewerName    49987 non-null  object
 3   helpful         50000 non-null  object
 4   reviewText      50000 non-null  object
 5   overall         50000 non-null  int64
 6   summary         50000 non-null  object
 7   unixReviewTime  50000 non-null  int64
 8   reviewTime      50000 non-null  object
dtypes: int64(2), object(7)
memory usage: 3.4+ MB
```

- Création de la matrice **df**

	000100039X	0001055178	0001473123	0001473727	0001473905	0001712772	000171287X	0001714538	0002005395	0002006715	...	0060533226	0060533390	0060533455	0060533994	00605
1	6	0.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	...	1	0.0	0.0	0.0	
2	4	4.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	...	4	1.0	0.0	1.0	
3	8	2.0	0.0	0.0	1.0	0.0	2.0	0.0	5.0	1.0	...	1	3.0	0.0	2.0	
4	15	10.0	2.0	0.0	0.0	1.0	1.0	2.0	3.0	0.0	...	9	1.0	1.0	6.0	
5	173	2.0	13.0	7.0	5.0	13.0	9.0	3.0	6.0	4.0	...	10	12.0	4.0	10.0	

5 rows × 2641 columns

(a)

1)Pour calculer la moyenne de score de chaque livre, on regroupe par l'identifient du livre et on applique mean() à la variable overall via la fonction agg

	overall		
	mean	max	min
asin			
000100039X	4.674757	5.0	1.0
0001055178	3.555556	5.0	2.0
0001473123	4.625000	5.0	1.0
0001473727	5.000000	5.0	5.0
0001473905	4.666667	5.0	3.0

2)Pour retrouver les livres les mieux appréciés, on localise les livres qui sont dans les deux dernières lignes (overall > 3.5 en particulier)

000100039X	Un des plus apprecié
0001055178	Un des plus apprecié

Pour retrouver les livres les moins appréciés, on localise les livres qui sont dans les deux premières lignes (overall < 2.5 en particulier)

006000665X	Moins Apprecié
0060006765	Moins Apprecié

3) Pour retrouver le 1er quart des livres les plus appréciés; on trie nos livres de façon décroissante et on affiche le premier quart des livres de notre table

4) Entre deux livres, lequel est le mieux apprécié ?
On définit la fonction comparaison(i,j), qui compare la moyenne du overall du livre de la colonne i avec la moyenne du overall du livre de la colonne j elle retourne le livre dont la moyenne du overall est la plus grande.

0	000100039X	4.674757
1	0001055178	3.555556
2	0001473123	4.625000
3	0001473727	5.000000
4	0001473905	4.666667
5	0001712772	4.666667
6	000171287X	4.583333
7	0001714538	3.714286
8		

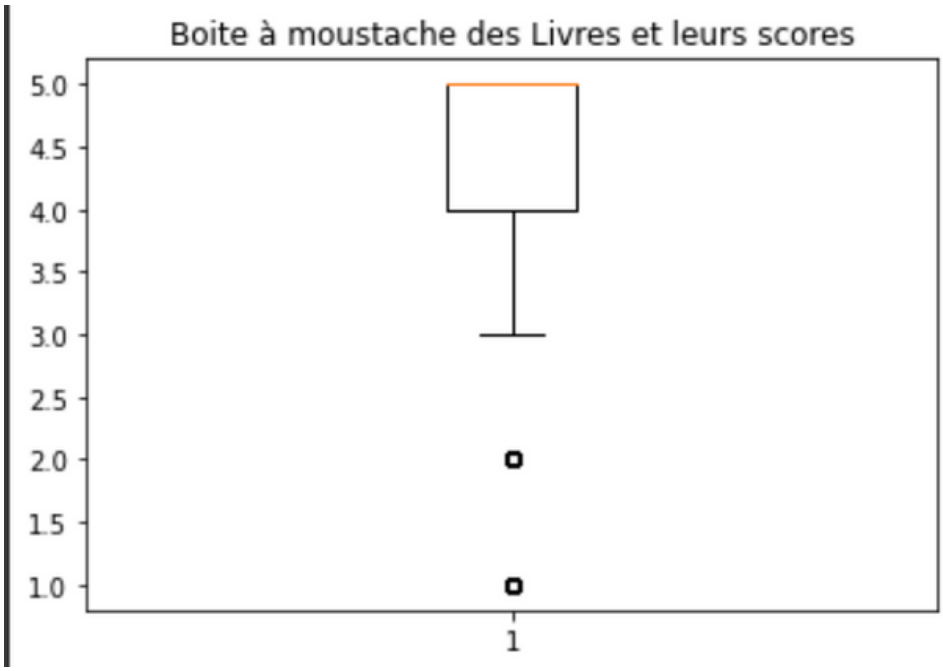
```
print(comparaison(0,1)) # tests de la fonction de comparaison
comparaison(7,2)
```

```
overall  mean    4.674757
Name: 000100039X, dtype: float64
overall  mean    4.625
Name: 0001473123, dtype: float64
```

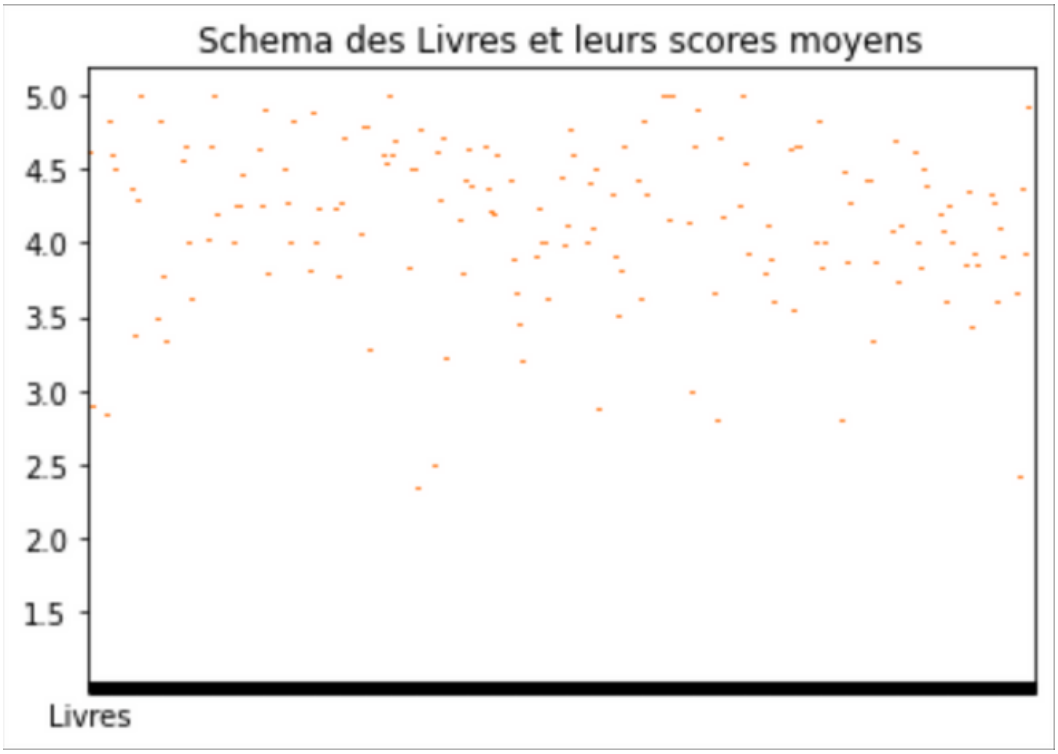
5) L'utilisation de la moyenne des scores n'est pas très représentative pour évaluer un livre par rapport à un autre.
En effet, plusieurs autres critères peuvent entrer en jeu:
-Nombre de reviews
-le facteur temporel (tendances de lecture ..)
-Le texte écrit par l'internaute
-L'ancienneté et l'expertise de l'internaute
-Préférence de lecture de l'internaute (si un internaute préfère les romans policiers, il peut ne pas aimer les romances de science-fictions)

Une alternative serait de prendre en considération ces facteurs et évaluer les corrélations et la dépendance des variables.
Il faudrait aussi faire recourt aux outils de visualisation de nos données (Boxplot, L'utilisation de L'ACP aussi s'avère très utile pour visualiser et synthétiser nos données

6)On utilise la fonction boxplot() pour afficher le diagramme en moustaches.



On remarque que la médiane se coïncide visuellement avec le premier quartile et le max. On peut mettre une hypothèse de "over rating". Pour illucider cela, on peut le visualiser dans ce qui suit, ou on représente la moyenne pondérée de chaque livre.



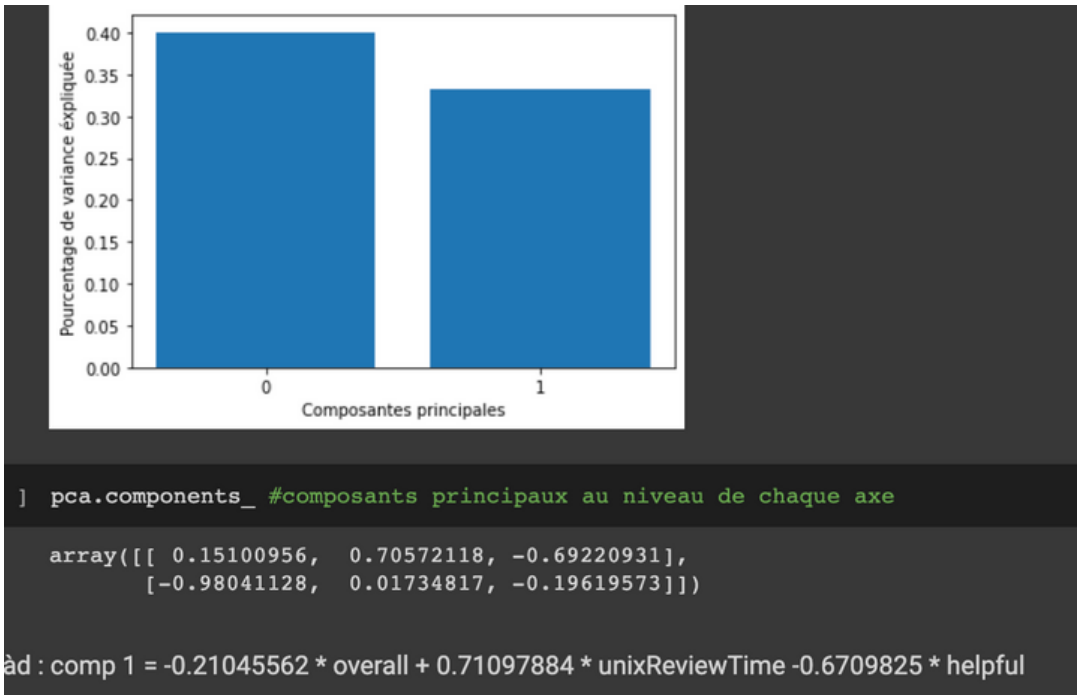
(b)

1)Intuitivement, les variables overall, helpful et unixreview seront les plus utiles pour notre Analyse en Composantes Principales.
après avoir appliquer fit_transform(), on se retrouve avec les 2 composantes principales suivantes :

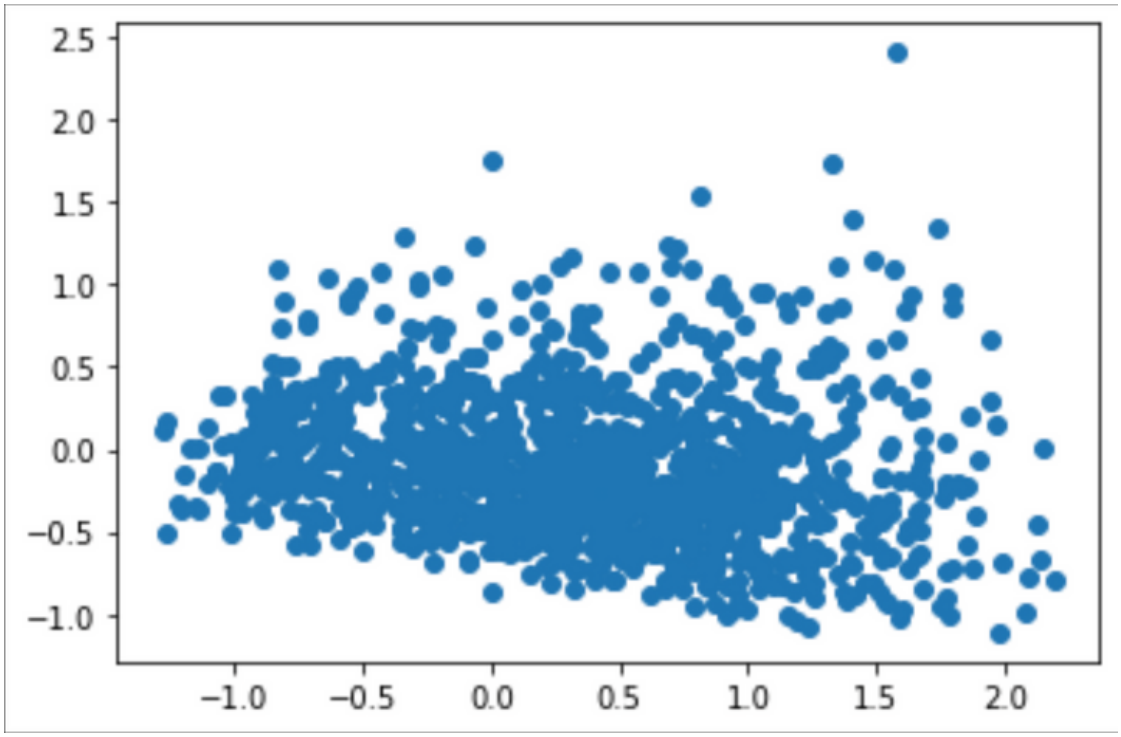
	helpful	overall	unixReviewTime
	mean	mean	mean
asin			
000100039X	0.413828	4.674757	1.236713e+09
0001055178	0.462963	3.555556	1.031966e+09
0001473123	0.207386	4.625000	1.349109e+09
0001473727	0.571429	5.000000	1.356048e+09
0001473905	0.333333	4.666667	1.332288e+09
...	...	...	...
0060006617	0.400971	4.125000	1.209481e+09

fit_transform()

	1ère composante principale	2ème composante principale
asin		
000100039X	0.213679	-0.503906
0001055178	1.552591	0.021860
0001473123	-0.710364	-0.234794
0001473727	-0.210087	-0.524683
0001473905	-0.421824	-0.300372
...	...	...
0060006617	0.423290	-0.093752



Pourcentage de variance expliquée par chaque composante principale obtenue



Le nuage de point obtenus par les deux composantes principales

On check par la suite le degrés de corrélation entre ces deux composantes pour confirmer notre choix.

```
corr, _ = pearsonr(dff.iloc[:,0], dff.iloc[:,1])
corr
```

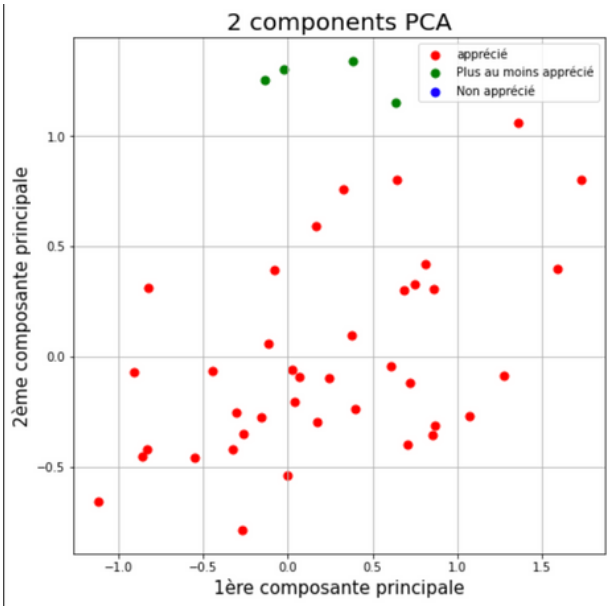
```
0.004164141416967995
```

2) On se retrouve donc avec la représentation suivante :

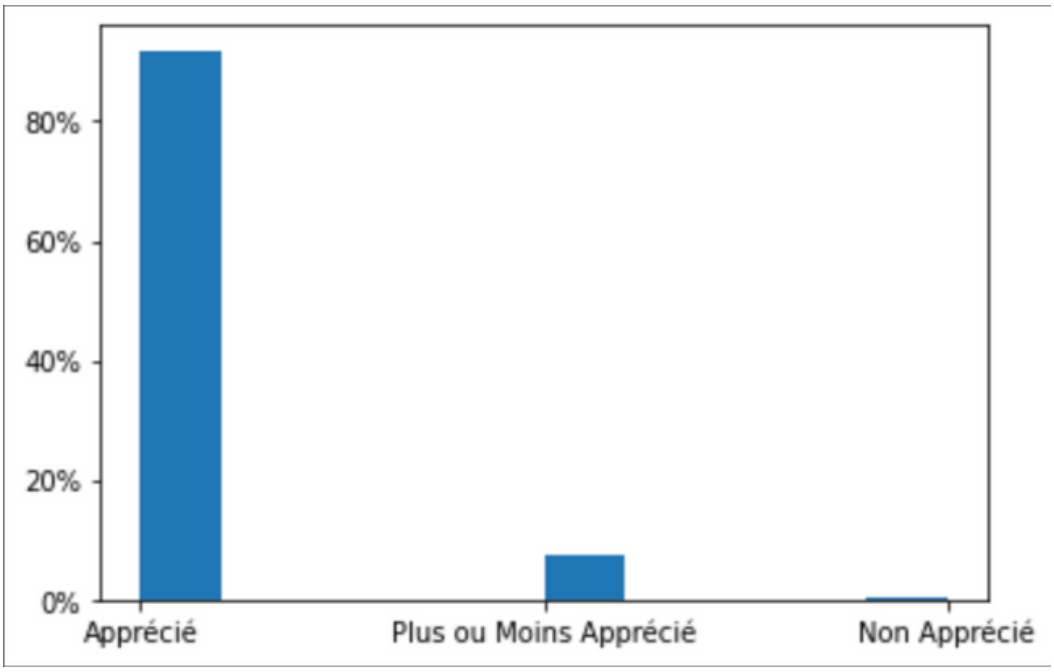
asin	1ère composante principale	2ème composante principale	target
000100039X	0.017058	-0.433517	Apprécié
0001055178	1.254119	0.340219	Apprécié
0001473123	-0.867519	-0.246437	Apprécié

Où le target est soit apprécié, plus ou moin apprécié, ou moins apprécié.

Pour illustrer cela, le nuage de points est coloré selon les target ;



La distribution de données est exprimé par cet histogramme :



On remarque encore une distribution asymétrique, qui propose l'hypothèse du "over rating" des internautes.
On remarque aussi que dans le nuage de point, il n' y a aucun livre non apprécié. Cela peut dans un premier temps porter à confusion. Cependant l'histogramme ci-dessus explique cela.

3)

(c)

df_T est le nouveau dataset inculant la colonne responsable du temps (month). On subdivise cette matrice en 12 matrices **df_mois** suivant chaque mois.

Janvier	asin	reviewTime	overall
2	000100039X	2014-01-18	5.0	1.0
5	000100039X	2014-01-27	5.0	1.0
8	000100039X	2014-01-29	5.0	1.0
10	000100039X	2009-01-15	5.0	1.0
14	000100039X	2013-01-23	5.0	1.0
...	...	...	...	...
99928	0060534389	2013-01-28	5.0	1.0
99936	0060534389	2013-01-05	4.0	1.0
99941	0060534389	2007-01-03	3.0	1.0
99946	0060534389	2013-01-21	4.0	1.0
99975	0060534397	2009-01-05	4.0	1.0

[10663 rows x 4 columns]

Juin	asin	reviewTime	overall	
15	000100039X	2012-06-27	5.0	6.0
24	000100039X	2013-06-17	5.0	6.0
34	000100039X	2000-06-24	5.0	6.0
38	000100039X	2001-06-19	5.0	6.0
40	000100039X	2014-06-18	5.0	6.0
...	...	...	...	...
100916	0060539054	2005-06-05	4.0	6.0
100950	0060539070	2009-06-25	2.0	6.0
100962	0060539097	2006-06-26	3.0	6.0
100965	0060539097	2005-06-19	4.0	6.0
100976	0060539097	2005-06-29	5.0	6.0

Février	asin	reviewTime	overall
13	000100039X	2001-02-28	5.0	2.0
18	000100039X	2012-02-15	5.0	2.0
21	000100039X	2001-02-24	5.0	2.0
42	000100039X	2012-02-08	5.0	2.0
52	000100039X	2013-02-11	5.0	2.0
...	...	...	...	...
99953	0060534389	2014-02-25	5.0	2.0
99967	0060534389	2013-02-26	4.0	2.0
99984	0060534397	2013-02-03	4.0	2.0
99990	0060534397	2013-02-19	3.0	2.0
99995	0060534397	2010-02-11	5.0	2.0

[8395 rows x 4 columns]

Octobre	asin	reviewTime	overall
4	000100039X	2002-10-07	5.0	10.0
19	000100039X	2003-10-13	5.0	10.0
26	000100039X	2004-10-03	5.0	10.0
30	000100039X	2013-10-29	5.0	10.0
41	000100039X	2013-10-03	5.0	10.0
...	...	...	...	...
100978	0060539097	2004-10-17	5.0	10.0
100981	0060539097	2005-10-12	5.0	10.0
100983	0060539097	2004-10-27	5.0	10.0
100991	0060539097	2004-10-03	4.0	10.0
100993	0060539097	2006-10-27	5.0	10.0

Mars	asin	reviewTime	overall	
6	000100039X	2008-03-28	5.0	3.0
12	000100039X	2014-03-16	5.0	3.0
23	000100039X	2008-03-21	2.0	3.0
25	000100039X	2012-03-12	5.0	3.0
39	000100039X	2008-03-29	5.0	3.0
...	...	...	...	...
100889	006053902X	2014-03-22	5.0	3.0
100905	0060539054	2006-03-16	5.0	3.0
100920	0060539054	2013-03-19	4.0	3.0
100921	0060539054	2014-03-09	5.0	3.0
100945	0060539070	2012-03-26	4.0	3.0

[9469 rows x 4 columns]

Novembre	asin	reviewTime	overall
7	000100039X	2013-11-03	5.0	11.0
11	000100039X	2013-11-20	5.0	11.0
17	000100039X	2005-11-16	5.0	11.0
53	000100039X	2009-11-29	5.0	11.0
69	000100039X	2007-11-29	5.0	11.0
...	...	...	...	...
100898	0060539046	2007-11-04	4.0	11.0
100911	0060539054	2008-11-15	4.0	11.0
100942	0060539070	2008-11-22	5.0	11.0
100961	0060539097	2007-11-30	5.0	11.0
100975	0060539097	2005-11-14	1.0	11.0

[7560 rows x 4 columns]

Avril	asin	reviewTime	overall
35	000100039X	2014-04-06	5.0	4.0
45	000100039X	2012-04-16	3.0	4.0
68	000100039X	2013-04-05	5.0	4.0
86	000100039X	2013-04-13	5.0	4.0
89	000100039X	2007-04-18	5.0	4.0
...	...	...	...	...
100923	0060539054	2008-04-03	3.0	4.0
100969	0060539097	2005-04-25	1.0	4.0
100971	0060539097	2005-04-08	1.0	4.0
100995	0060539097	2010-04-20	2.0	4.0
100998	0060539097	2013-04-04	5.0	4.0

Novembre	asin	reviewTime	overall
7	000100039X	2013-11-03	5.0	11.0
11	000100039X	2013-11-20	5.0	11.0
17	000100039X	2005-11-16	5.0	11.0
53	000100039X	2009-11-29	5.0	11.0
69	000100039X	2007-11-29	5.0	11.0
...	...	...	...	...
100898	0060539046	2007-11-04	4.0	11.0
100911	0060539054	2008-11-15	4.0	11.0
100942	0060539070	2008-11-22	5.0	11.0
100961	0060539097	2007-11-30	5.0	11.0
100975	0060539097	2005-11-14	1.0	11.0

Mai	asin	reviewTime	overall	Mor
32	000100039X	2008-05-01	5.0	5.0
37	000100039X	2000-05-03	5.0	5.0
44	000100039X	2012-05-31	5.0	5.0
49	000100039X	2009-05-15	5.0	5.0
54	000100039X	2002-05-21	5.0	5.0
...	...	...	...	...
100894	0060539046	2013-05-23	5.0	5.0
100903	0060539054	2012-05-28	4.0	5.0
100910	0060539054	2005-05-16	5.0	5.0
100959	0060539097	2008-05-15	5.0	5.0
100992	0060539097	2005-05-29	5.0	5.0

Décembre	asin	reviewTime	overall
0	000100039X	2012-12-16	5.0	12.0
1	000100039X	2003-12-11	5.0	12.0
33	000100039X	2013-12-17	3.0	12.0
36	000100039X	2004-12-04	5.0	12.0
48	000100039X	2009-12-24	5.0	12.0
...	...	...	...	...
100977	0060539097	2004-12-19	5.0	12.0
100980	0060539097	2005-12-13	5.0	12.0
100989	0060539097	2004-12-10	1.0	12.0
100990	0060539097	2005-12-29	5.0	12.0
100996	0060539097	2013-12-21	5.0	12.0

Ensuite nous devrons appliquer l'ACP pour chaque matrice en suivant les mêmes composantes principales que dans la partie (c).

FIN