Sonia DAKHLI,
Nino RAVELLA,
Maeva SOMNY

M2 IA
2023/2024

Data Mining - Projet
Sujet : Applications du Google PlayStore

I. Problématique

Ce projet s'intéresse à l'analyse d'un jeu de données contenant des informations sur les **applications du google store**.

Ces données ont été récoltées en 2018, selon la date de l'application ayant été mise à jour le plus récemment (2018-08-08).

Notre but est de réaliser une analyse permettant de déterminer **quelles sont les caractéristiques impactant la popularité d'une application**. Pour cela, nous déterminerons que plus une application est téléchargée et obtient un haut score, plus elle est populaire. Cela nous permettra alors d'établir un ou plusieurs modèles d'applications ayant un succès récurrent.

II. Répartition des tâches

- Pré-traitement des données : ensemble du groupe
- Pré-analyse : Maeva
- Clustering : Nino & Maeva
- Frequent-pattern : Sonia
- Impact des mots clés dans les titres : Maeva

III. Pré-traitement des données

Avant de débuter, nous avons fait un pré-traitement sur nos données, dans le but d'obtenir un jeu de données plus exploitable. En effet, nous nous sommes rendus compte que certaines informations sur nos applications n'étaient pas pertinentes ou nécessitaient d'être retravaillées pour être exploitables.

Nous avons donc réalisé le travail présenté en Annexe1, que vous retrouverez dans le fichier `pretraitement.ipynb`. Puis nous avons bien entendu supprimé les applications dont nous n'avions pas toutes les informations afin d'avoir un jeu de données complet et fiable, passant ainsi de 10841 applications à 7094, ce qui reste un nombre d'applications largement suffisant.

Exemple pour une application, avant le pré-traitement :

App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19M	10,000+	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up

Après le pré-traitement :

App	Category	Rating	Reviews	Size	Installs	Price	Content Rating	Genre Principal	Last Updated	code Category	code Content Rating	Genre Secondaire	code Genre Secondaire	code Genre Principal
Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19000000	10000.0	0.0	Everyone	Art & Design	07/01/2018	0	0	Aucun	0	0

IV. Pré-analyse

Pour nous familiariser avec ce projet, nous avons fait une pré-analyse de terrain sur nos données afin d'en faire ressortir certaines caractéristiques. Pour cela nous avons utilisé l'outil Tableau.

Dans un premier temps, nous avons donc cherché à établir des liens de corrélations entre plusieurs attributs de nos données et la popularité d'une application (son nombre d'installations et notation).

A. Corrélation entre genre principal, genre secondaire et rating.

Pour commencer, nous nous sommes donc demandé s'il y avait une corrélation entre le genre de l'application et sa note. En d'autre termes : **Quel genre d'application faut-il faire afin d'avoir les meilleures notes possibles ?**

Nous avons donc établi un premier visuel, disponible ici mettant en lien genre principal/genre secondaire et rating. La taille des points est déterminée par la note médiane obtenue par les applications tandis que la couleur est déterminée par la note moyenne obtenue par les applications.

Lorsque l'on se penche sur ce visuel, on remarque que ce sont les genres principaux "strategy", "adventure", "comics", "entertainment" et "health & fitness" qui obtiennent les meilleurs résultats. Ainsi, les genres les plus appréciés seraient donc liés au **jeu, au divertissement** ("strategy", "adventure", "comics", "entertainment") et **au mode de vie** ("health & fitness"). On remarque également que le fait de ne pas spécifier de genre secondaire ("Aucun") semble avoir un impact sur la notation.

Ainsi, les applications plus sérieuses ("education", "educational") ont l'air d'être moins appréciées. Cela pourrait s'expliquer par le fait que lorsque l'on souhaite utiliser une application pour quelque chose qui n'est pas du divertissement, nous sommes plus exigeants vis-à-vis des besoins auxquels celle-ci doit répondre. C'est un point que nous approfondirons dans une partie ultérieure.

Lorsque l'on filtre ces informations en ne gardant que les applications ayant été le plus installées, on remarque qu'il existe donc des applications ayant été énormément téléchargées mais n'obtenant pas toute la satisfaction des utilisateurs. Cela peut s'expliquer par le fait qu'un avis très positif à moins d'impact si beaucoup d'avis ont déjà été laissés. Mais peut-être qu'il s'agit de marchés pouvant être explorés. Si beaucoup d'utilisateurs installent et utilisent une application qu'ils n'apprécient pas fortement, peut-être est-ce l'occasion de faire un produit de meilleure qualité ?

B. Corrélation entre genre principal, genre secondaire et nombre d'installations.

Pour continuer, nous nous sommes demandés s'il y avait une corrélation entre le genre de l'application et son nombre d'installations. En d'autre termes : **Quel genre d'application faut-il faire afin d'avoir le plus d'installations possibles ?**

Une fois de plus, ce sont les **jeux** qui semblent avoir le plus de succès en termes d'installations. Nous pouvons le constater sur ce visuel, montrant que même si en moyenne, certaines applications sont très téléchargées, en médiane, ce sont les applications du genre "aventure - action et aventure", et donc des jeux qui sont téléchargés.

C. Corrélation entre catégorie et note

Nous nous sommes ensuite demandés s'il y avait une corrélation entre la catégorie (décrit le genre d'une manière plus générale) de l'application et son évaluation. En d'autres termes : **Quelle catégorie d'application est la mieux notée ?**

Ce visuel permet de remarquer que même si la catégorie jeu est bien placée, d'autres catégories ont en moyenne, mieux notées. Cela traduit donc une disparité au sein de la catégorie jeu, en comparaison à la catégorie "event" qui arrive en tête (meilleure moyenne) avec une bien moins grande disparité.

D. Corrélation entre catégorie et installations

Ainsi nous sommes donc demandés s'il y avait une corrélation entre la catégorie (décrit le genre d'une manière plus générale) de l'application et le nombre d'installations. En d'autres termes : **Quelle catégorie d'application est la plus téléchargée ?** Sans grande surprise, c'est la catégorie **jeu** qui est en tête, comme le montre ce visuel.

E. Corrélation entre prix et note/ installations

Nous avons par la suite décidé de nous intéresser à l'impact du prix sur la popularité de nos applications. **Faire une application payante aura-t-il un impact sur sa popularité ?**

Pour cela, vous trouverez le visuel ici. Ainsi, que ce soit pour optimiser l'évaluation de l'application ou son nombre de téléchargement, il est nécessaire de **faire une application gratuite**. La note est moins impactée que le nombre de téléchargements par le prix. Cependant, cela implique que l'application doit remplir certaines exigences, au risque de voir la note baisser à nouveau pour un prix trop haut. Lorsque l'on se penche sur notre catégorie la plus téléchargée, à savoir la catégorie jeu, on fait le même constat.

F. Corrélation entre classification et note/installations

Reste alors à étudier une autre caractéristique de nos applications : leur classification, afin de répondre à la question suivante : **Quelle classification est la plus populaire ?**

Ainsi, on remarque sur ce visuel que si la classification n'a pas une grande importance sur la notation d'une application, celle-ci à un impact sur le nombre de téléchargements. On aurait pu penser à première vue qu'une classification "everyone" impliquerait plus de téléchargements puisque l'application n'est pas limitée par l'âge de l'utilisateur. Mais c'est la classification "everyone 10+" qui semble avoir le plus de succès, correspondant à " Peut contenir plus de violence fictive, légère ou sous forme d'animation, des propos peu grossiers et/ou de rares thèmes suggestifs.", selon google, ce qui correspond donc majoritairement à des jeux.

G. Corrélation entre maintenance et note/installations

Enfin, nous nous penchons sur ce visuel, sur l'impact de la maintenance sur le nombre d'installations et les notes obtenues par les applications. Sans grande surprise, ces caractéristiques sont meilleures lorsque l'application est maintenue.

H. Conclusion

Si les jeux dominent les autres applications en termes de téléchargement, ils sont cependant des genres incertains, parfois très bien évalués ou au contraire mal appréciés. Des catégories comme Education et Event ont l'air d'obtenir des résultats plus stables mais moins de succès. Ajoutons à cela le fait que ce sont les applications gratuites Everyone10+ qui sont les plus téléchargées, il semblerait donc que ce soit **les jeux gratuits et adaptés à un public jeune** qui prédominent le marché.

IV. Analyse des données

A. Clustering k-means

Cette partie a été réalisée dans le fichier `clustering.ipynb`.

Clustering

Dans le cadre de cette étude, nous nous sommes ensuite tournés vers une approche de clustering afin de regrouper les applications selon les caractéristiques jugées les plus pertinentes pour observer (ou non) des similitudes entre elles. Nous avons d'abord établi **trois critères relatifs à la popularité** d'une application, à savoir **la note**, **le nombre de reviews** et **le nombre d'installations**. Puis, étant intéressés par l'impact du prix et de la catégorie, nous avons voulu les ajouter à notre cluster. Cependant, dans l'espoir de voir des résultats plus clairs, nous avons construit deux clusters, chacun avec les critères de popularité et l'un avec le prix et l'autre avec la catégorie. Ayant des données réparties sur des échelles complètement différentes, nous les avons normalisées avant d'appliquer l'algorithme de clustering. Pour le clustering en lui-même, nous avons adopté la méthode k-means avec la méthode Elbow pour trouver le nombre de clusters (4 dans les deux cas).

Une fois les clusters réalisés, nous obtenons les visualisations et coordonnées de centroïdes des figures et tableaux suivants.

Ces résultats montrent d'abord que la distinction prix/catégorie n'apporte que peu d'information puisqu'on y distingue des clusters très similaires. En s'intéressant aux coordonnées des centroïdes de nos différents clusters, nous pouvons faire les distinctions suivantes :

- Cluster 0 : Applications bien notées, mais moins installées et commentées que les plus populaires.
- Cluster 1 : Applications les moins populaires, plus particulièrement celles étant le moins installées et commentées.
- Cluster 2 : Applications ayant le plus de popularité sous tous les aspects.
- Cluster 3 : Applications les moins populaires, plus particulièrement les moins bien notées.

Analyse

Maintenant que nous avons réparti nos données en différents clusters, nous pouvons en tirer quelques conclusions. Que ce soit pour les prix ou les catégories, nous allons étudier des répartitions dans les clusters populaires (0 et 2) et ceux qui sont moins populaires (1 et 3).
Pour les prix, en regardant la répartition des application payantes et non-payantes selon nos deux différentes catégories de cluster, nous obtenons les résultats suivants :

- Pourcentage d'applications payantes dans l'ensemble des données : 7.49%
- Pourcentage d'applications payantes dans les clusters populaires : 7.55%
- Pourcentage d'applications payantes dans les clusters non-populaires : 7.25%

Malgré un pourcentage plus élevé pour les clusters populaires, une si faible différence ne nous permet pas ici de tirer de conclusion sur l'impact de la gratuité ou non d'une application.

Pour les catégories, en regardant le ratio de la présence d'une d'elles dans nos clusters sur le nombre total présent dans le dataset, nous obtenons l'histogramme suivant. Celui-ci nous permet d'observer que certaines catégories sont complètement absentes des deux clusters jugés comme les plus populaires. Notamment, la catégorie "event" pourtant bien notée, qui n'est pas présente dans les clusters populaires. Le même constat peut se faire pour la catégorie "communication", qui, même en étant l'une des catégories les plus installées est presque uniquement représentée dans les clusters moins populaires. À l'inverse, la catégorie "weather" qui était passée inaperçue jusqu'à ce point, s'avère être la plus présente dans les clusters populaires devant même les jeux, qui restent malgré tout également très populaires du point de vue de cette analyse.

B. Frequent patterns

Cette partie a été réalisée dans le fichier `frequent_patterns.ipynb`.
Nous intéressant aux applications ayant le plus grand nombre d'installations et les meilleures notes, nous avons décidé de voir quels patterns revenaient fréquemment dans les applications possédant ces caractéristiques. Nous avons donc sélectionné uniquement les données les plus pertinentes, c'est-à-dire les applications téléchargées plus de 1 million de fois et dont la note est supérieure à 4.
Pour appliquer le frequent patterns sur nos données, nous avons dû transformer ces dernières en données booléennes en s'appuyant sur la pré-analyse et en testant plusieurs combinaisons possibles :

- **Prix** : 1 si l'application est gratuite, 0 sinon. La pré-analyse nous a, en effet, montré que la gratuité d'une application était un critère important à son succès.
- **GenrePrincipal** et **Genre Secondaire**: nous avons ajouté chaque différente catégorie comme colonne de notre tableau, avec comme valeur 1 si l'application est de ce genre, 0 sinon.

Le tableau ainsi obtenu est illustré dans l'annexe 3. Cette annexe présente les premières colonnes du tableau.

Support : Installations, reviews, ratings

Nous avons, pour ce premier calcul, conservé les colonnes de notre dataframe présentant un certain nombre d'installations (**Installs** : [0, >=1M, >=5M, >=10M, >=50M, >=100M]), un certain score (**Ratings** : [0, >=4, >=4.5]) et un certain nombre d'avis (**Reviews** : [0, 100, 500, 1000, 1500, 30000]).

Les résultats des différentes combinaisons entre les trois paramètres sont présentés dans le fichier "**./output/resultats_apriori.txt**".

Ce qui ressort principalement de toutes les combinaisons testées est :

- Une application **gratuite** sera présente dans toutes les applications populaires avec un support de 1.
- Pour les applications les plus populaires (Installs>=100000000 et Ratings>=4.5) les **reviews** sont importantes et présentent également un support de 1.
- En ce qui concerne les **genres**, la plupart des applications respectant nos critères de popularité n'ont pas de genre secondaire et sont tout de même populaires.

Nous remarquons que la gratuité est déterminante, nous pouvons maintenant l'enlever puisqu'elle ne modifiera pas les couples que nous pourrons obtenir. Les reviews sont également importantes, mais ne nous indiquent pas de critère permettant de créer notre application. Intéressons-nous donc maintenant uniquement aux genres.

Support : Genres

Les résultats des différentes combinaisons entre Installs et Ratings sont présentés dans le fichier "**./output/resultats_apriori_no_price.txt.**"

- Les applications n'ayant pas de genre secondaire (***NoSecondaryGenre***) ont un support entre 93% et 100%.
- Les jeux d'***Action*** ont le meilleur second support (environ 9%) pour les applications installées entre 1 million et 50 millions de fois. Ils ont un support de 11% pour les applications installées entre 50 millions et 100 millions de fois, puis à partir de 100 millions ne font pas partie des meilleures catégories.
- Les jeux d'***Arcade*** ont un support de 16% pour les applications installées entre 50 millions et 100 millions de fois.

En ce qui concerne les applications téléchargées plus de 100 millions de fois, les genres ***Casual*** et ***Communication*** se classent en tête.

Support : Couples de genres

Les résultats des différentes combinaisons entre Installs et Ratings sont présentés dans le fichier "**./output/resultats_apriori_no_2nd_genre_category.txt**".
Les couples de genres ne nous donnent aucune précision supplémentaire. En effet, tous les couples renvoyés ne possèdent pas de genre secondaire (*NoSecondaryGenre*).

Support : Catégorie

Nous avons par la suite décidé d'ajouter la colonne **Catégorie**, pour voir si des patterns récurrents entre **Catégorie** et **GenrePrincipal** existaient dans notre dataset. Les résultats sont présentés dans le fichier "**./output/resultats_apriori_couples_categories.txt**".
Les couples (Catégorie, GenrePrincipal) revenant le plus fréquemment sont : (Game, Action), (Game, Arcade), (Communication, Communication), (Tools, Tools).

Conclusion

Bien qu'aucun pattern n'ait vraiment été trouvé, les deux points qui ressortent de cette analyse sont qu'une application doit être **gratuite** pour avoir du succès et qu'elle n'a **pas** besoin d'être étiquetée par un **genre secondaire**.
Le genre principal d'une application peut se baser sur les genres ayant les meilleures valeurs de support : Action, Arcade, Casual et Communication.

C. Analyse des mots clés dans les titres

Le code en référence à cette partie se trouve dans le fichier `keywords.ipynb`. Pour commencer cette analyse, nous avons compté l'occurrence de chaque mot présent dans les titres afin de voir quels mots étaient les plus présents.

Cela nous a permis de faire ressortir les mots “Free”, “Mobile”, “App”, “Pro” et “News”, que nous utiliserons pour notre analyse. Nous avons également choisi d’ajouter les mots “Easy” et “Official”, souvent utilisés pour attirer le consommateur. Nous avons transformé notre dataset afin d’y ajouter des colonnes, une pour chaque mot clé, qui détermine la présence ou non de celui-ci dans le titre.

Clustering

Par la suite, nous avons procédé par clustering, essayant de trouver des similitudes entre les différents groupes. La technique du coude nous à donné un premier clustering de 8, dont l’analyse n’a pas été très révélatrice. Dans ce premier clustering, 2 groupes ressortent, ne contenant chacun presque que des applications n’ayant pas de mot clé dans leurs titres, et les séparant en 2 catégories : les applications avec un haut rating mais peu d’installations et de reviews, et les applications avec un haut rating et beaucoup d’installations et de reviews. Les 6 autres clusters revenant globalement à regrouper les applications en fonction de leurs mots clés, sans donner trop d’autres informations sur leur popularité. Voir Annexe 4.

Clustering “forcé”

Afin de forcer les choses, nous avons réalisé un clustering avec 4 clusters pour reprendre le nombre de clusters précédents. Malgré cela, les deux clustering obtenus ont une allure très similaire, comme le monde l’Annexe 4. Cependant, on remarque que cette fois-ci, les applications ayant un titre contenant ou ne contenant pas de mots clés se retrouvent plus ou moins mélangées, hormis pour le cluster 4, comme le montre l’annexe 4. On remarque par ailleurs que les clusters ont à peu près tous les mêmes allures, c'est-à-dire un rating plus ou moins haut, et un nombre d’installations variable. Il semblerait donc que la présence de mot clés accrocheurs dans le titre n’ait pas d’incidence sur la popularité d’une application. On peut toutefois noter que les applications contenant les mots clés “Free”, “Easy” et “Official”, soit les mots clés les plus accrocheurs, sont toutes présentes dans le même cluster. Il s’agit du cluster 0, qui semble contenir des applications légèrement plus populaires que les 3 autres.

Matrice de similarités

Enfin, pour finaliser cette analyse, nous avons tenté de dresser une matrice permettant de mettre en avant les similarités entre les différentes applications dont les titres contenaient les mots clés accrocheurs précédemment vus. Comme le montre la matrice obtenue en Annexe 5, la seule similarité présentée concerne le nombre d'installations et le nombre de review, ce qui n’est pas surprenant et plutôt logique.

Conclusion

En conclusion, il ne semble pas que la présence de mot clés ait une incidence sur la popularité d’une application. Cependant, on peut toutefois noter que les applications classées dans les clusters contenant les applications à haut rating n'avaient pas de mot clés accrocheurs dans leurs titres, hormis pour “Free”, “Easy” et “Official”, qui appartiennent à un cluster semblant contenir des applications plutôt populaires. Cependant, au vu du manque de similarités entre les différents mots clés et de la similarités des allures des différents clusters, cette présence ne signifie pas que ces mots clés impliquent un succès. Ainsi, il semblerait préférable de ne pas utiliser ce genre de technique afin d’optimiser la popularité d’une application.

V. Conclusion

Les différentes méthodes que nous avons utilisées ont montré que les **jeux** étaient un type d'application très populaire. Ainsi, si une entreprise souhaite créer une nouvelle application en optimisant sa popularité, il serait pertinent de créer une application de jeu. Cependant, celle-ci doit correspondre à plusieurs critères.

Tout d'abord, cette application à plus de chance de fonctionner si elle appartient à la **classification *Everyone10+*** comme nous avons pu le voir dans la pré-analyse. De plus, comme cela à été montré, dans la partie frequent pattern, une application populaire est **gratuite**. Enfin, les genres ressortant le plus dans notre analyse sont **Action**, **Arcade** et **Casual**, soit des genres de jeux qu'il faudrait donc prendre en compte dans le développement d'une application. La présence de mots clés accrocheurs ne semble pas avoir d'effet positif sur le consommateur, mis à part dans certains cas très précis (avec les mots clés "Free" et "Easy") qui semblent correspondre à des applications populaires. Cependant, le lien de causalité entre ces mots clés et la popularité des applications n'étant pas établis, il serait préférable de ne pas user de cette pratique.

Ce modèle type d'application **de jeu (casual) gratuite et accessible aux plus de 10 ans** s'explique facilement par le fait que lorsqu'une personne n'a pas besoin d'une application, ses attentes sont plus basses, ce qui engendre donc une meilleure évaluation. Cela est renforcé par le fait que l'application est gratuite. Les attentes de l'utilisateur vis-à-vis d'un produit pour lequel il n'a rien dépensé sont donc encore moindres.

Ainsi, il peut être pertinent de profiter de cela en créant une application de jeu ayant un effet "bonne surprise" (avec du contenu de qualité habituellement associé à un jeu payant par exemple) qui assurera la popularité de l'application.

Dans le cas où l'entreprise ne souhaite pas s'engager dans la création d'une application de jeu, les différentes catégories pertinentes seront liées **aux outils de la vie quotidienne**, comme la **météo** ou la **communication**. Attention toutefois à ne pas proposer un produit payant au risque de passer à côté d'une partie de sa cible qui ne souhaite pas payer pour une application, avec par conséquent un moins grand nombre de téléchargements.

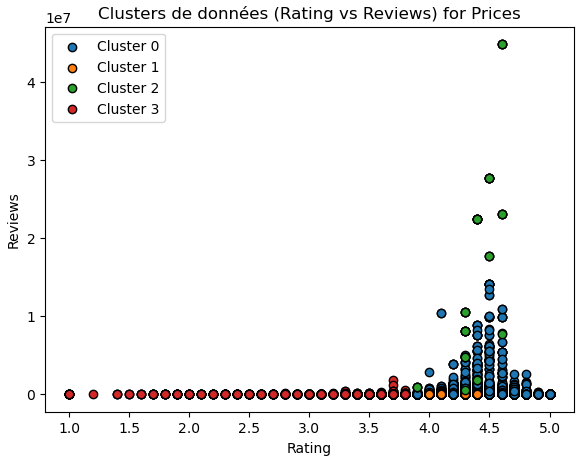
Annexes - Table des annexes

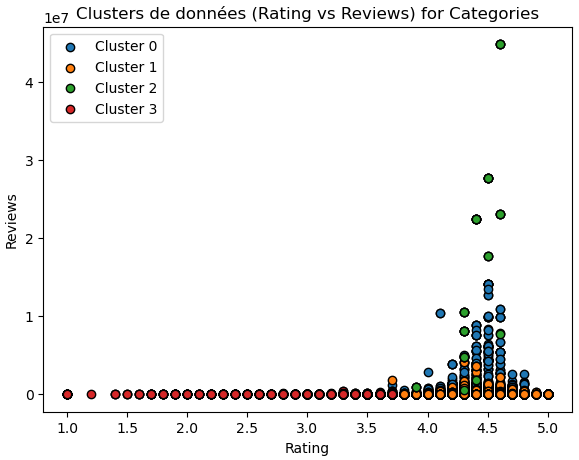
Annexe 1 - Détail du pré-traitement des données

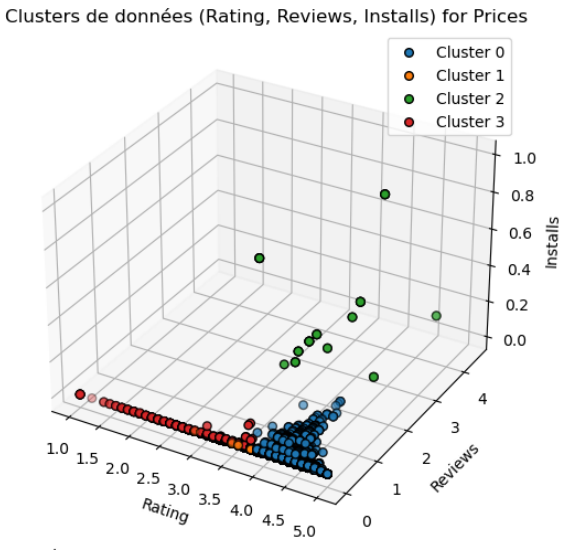
Données à l'origine	**Impact du pré-traitement**
App (nom de l'application)	Aucun changement, mis à part la conversion en colonne numérique
Category (categorie de l'application)	Ajout d'une colonne codeCategory permettant d'attribuer un code numérique à chacune des catégories
Rating (note moyenne obtenue)	Aucun changement, mis à part la conversion en colonne numérique
Reviews (nombre d'avis)	Aucun changement, mis à part la conversion en colonne numérique
Size (taille de l'application)	Retravail des valeurs afin qu'elles correspondent à des valeurs numériques (anciennement 24M → 24 000 000)
Installs (tranches de nombre d'installations)	Retravail des valeurs afin qu'elles correspondent à des valeurs numériques (anciennement 10,000+ → 10 000)
Type (gratuite ou non)	Colonne non conservée car l'information est déjà présente dans la colonne "Price"
Price (prix de l'application)	Aucun changement, mis à part la conversion en colonne numérique
Content Rating (classification en age)	Ajout d'une colonne codeContentRating permettant d'attribuer un code numérique à chacune des classifications
Genre (1 ou 2 genres définissant l'app)	Séparation de la colonne en deux : Genre principal et Genre secondaire. Ajout des colonnes codeGenrePrincipal et codeGenreSecondaire permettant d'attribuer un code numérique à chacuns des genres
Last Updated (date de dernière màj)	Retravail des valeurs afin que celles-ci soient considérées comme des dates (JJ/MM/AAAA)
Current Ver (n° version de l'app)	Colonne non conservée car la numérotation de la version est au choix du créateur, cela ne donne donc pas d'indication sur l'application supplémentaire à ce que nous fournis "Last Updated". De plus, contient trop de "Varies with device" qui ne donne pas d'information
Android Ver (Version supportée par l'app)	Colonne non conservée car beaucoup trop d'informations non fournies, présence de différents formats.

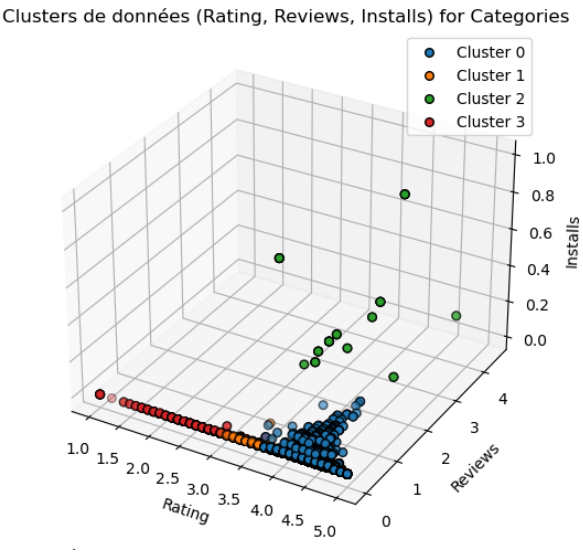
Annexe 2 - Clustering

Clusters et centroïdes





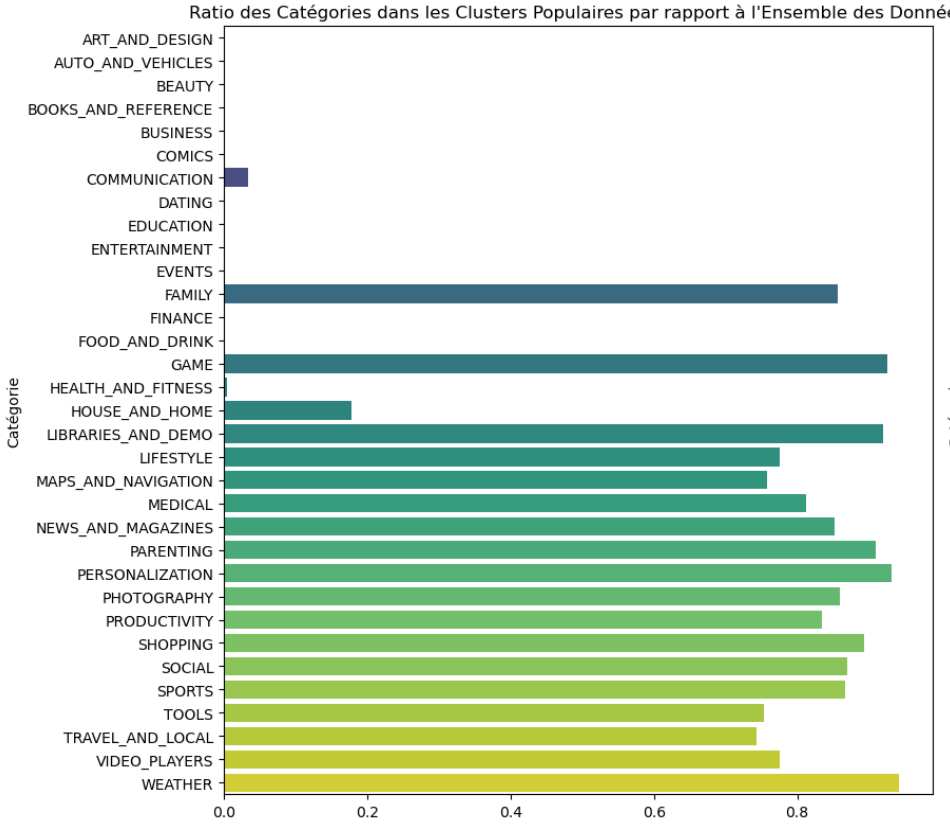


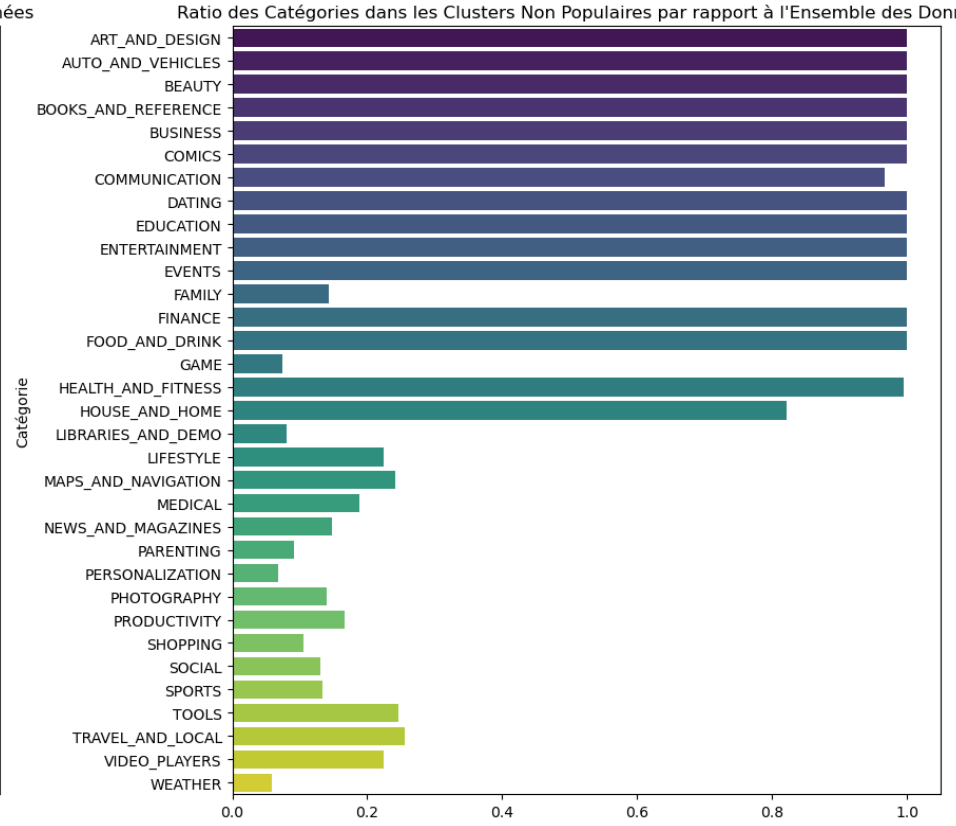


ClusterPrice	Rating	Reviews	Installs
0	4.378053	2.381361e+05	6.090760e+06
1	3.866667	6.032667e+02	1.460667e+04
2	4.387500	1.632271e+07	5.375000e+08
3	3.301986	1.106327e+04	1.000814e+06

ClusterCategory	Rating	Reviews	Installs
0	4.347316	2.772946e+05	6.847393e+06
1	4.294251	6.439931e+04	2.694921e+06
2	4.387500	1.632271e+07	5.375000e+08
3	3.139210	7.226743e+03	6.821388e+05

Répartition des catégories selon les clusters



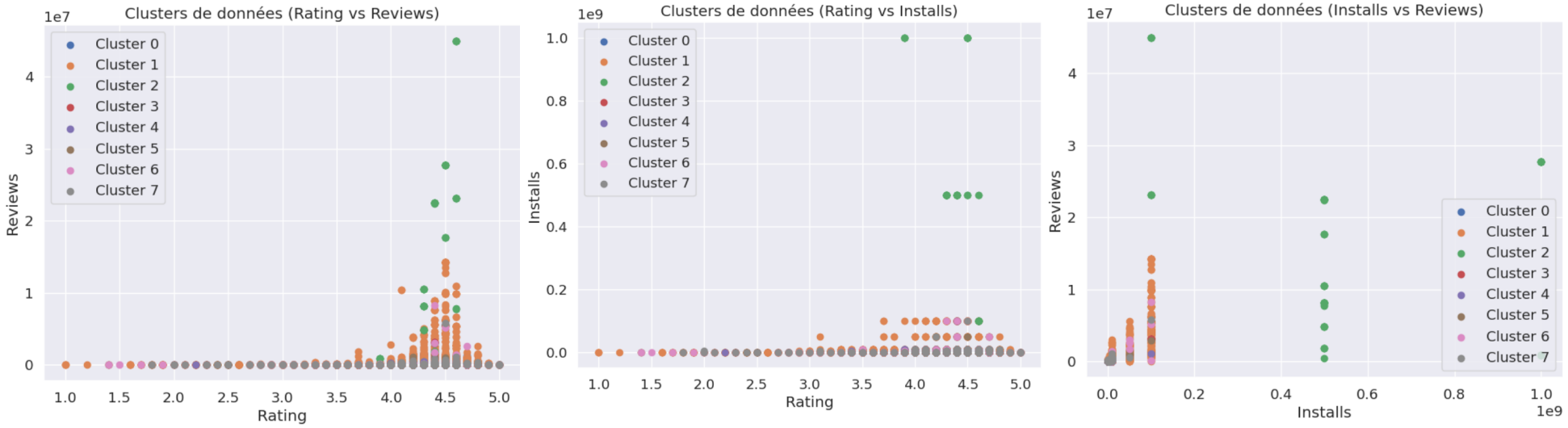


Annexe 3 - Tableau Frequent Patterns

App	Price	Action1	Adventure1	Arcade1	Art & Design1	Auto & Vehicles1	Beauty1
3D Bowling	1	0	0	0	0	0	0
3D Tennis	1	0	0	0	0	0	0
8 Ball Pool	1	0	0	0	0	0	0
APUS File Manager (Explorer)	1	0	0	0	0	0	0
APUS Launcher - Theme, Wallpaper, Hide Apps	1	0	0	0	0	0	0
Advanced Task Killer	1	0	0	0	0	0	0
Agar.io	1	1	0	0	0	0	0
Akinator	1	0	0	0	0	0	0
Alarm Clock: Stopwatch & Timer	1	0	0	0	0	0	0
Amazon Prime Video	1	0	0	0	0	0	0
Amazon Shopping	1	0	0	0	0	0	0
Anger of stick 5 : zombie	1	1	0	0	0	0	0
Angry Birds 2	1	0	0	0	0	0	0
Angry Birds Classic	1	0	0	1	0	0	0
Angry Birds Friends	1	0	0	1	0	0	0
Angry Birds Go!	1	0	0	0	0	0	0
Angry Birds Rio	1	0	0	1	0	0	0

Annexe 4 - Impact des mots clés

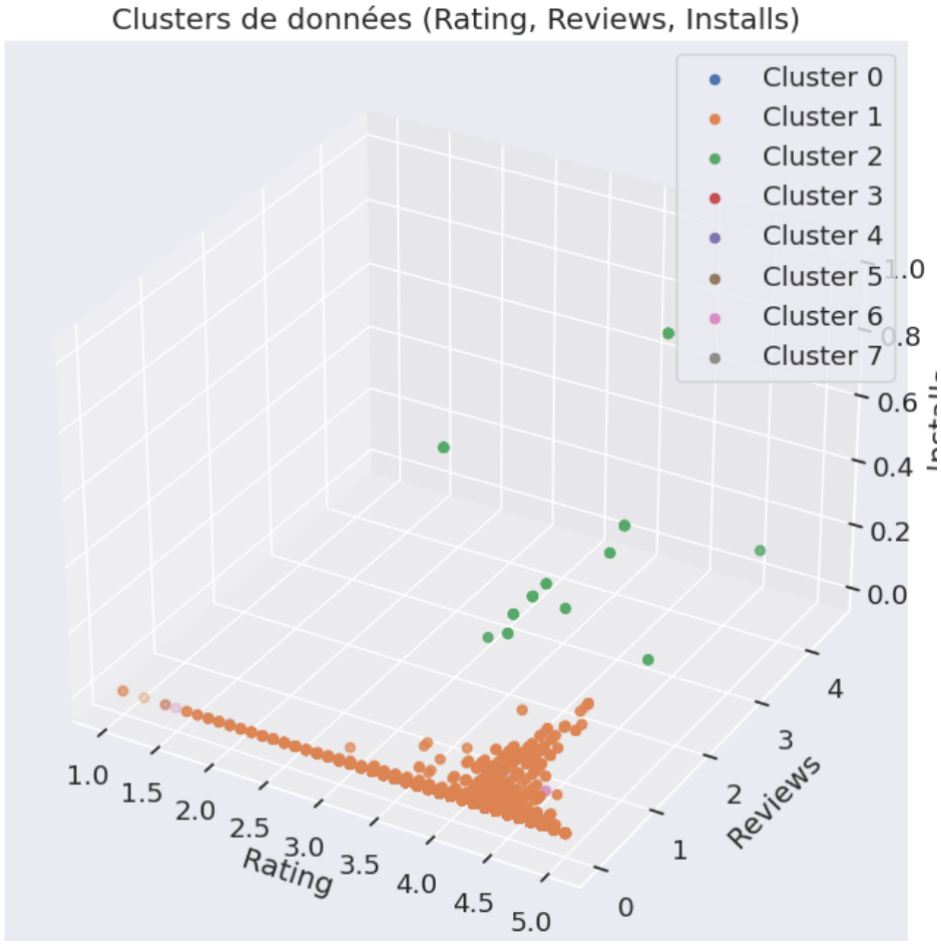
Clustering pour 8 clusters



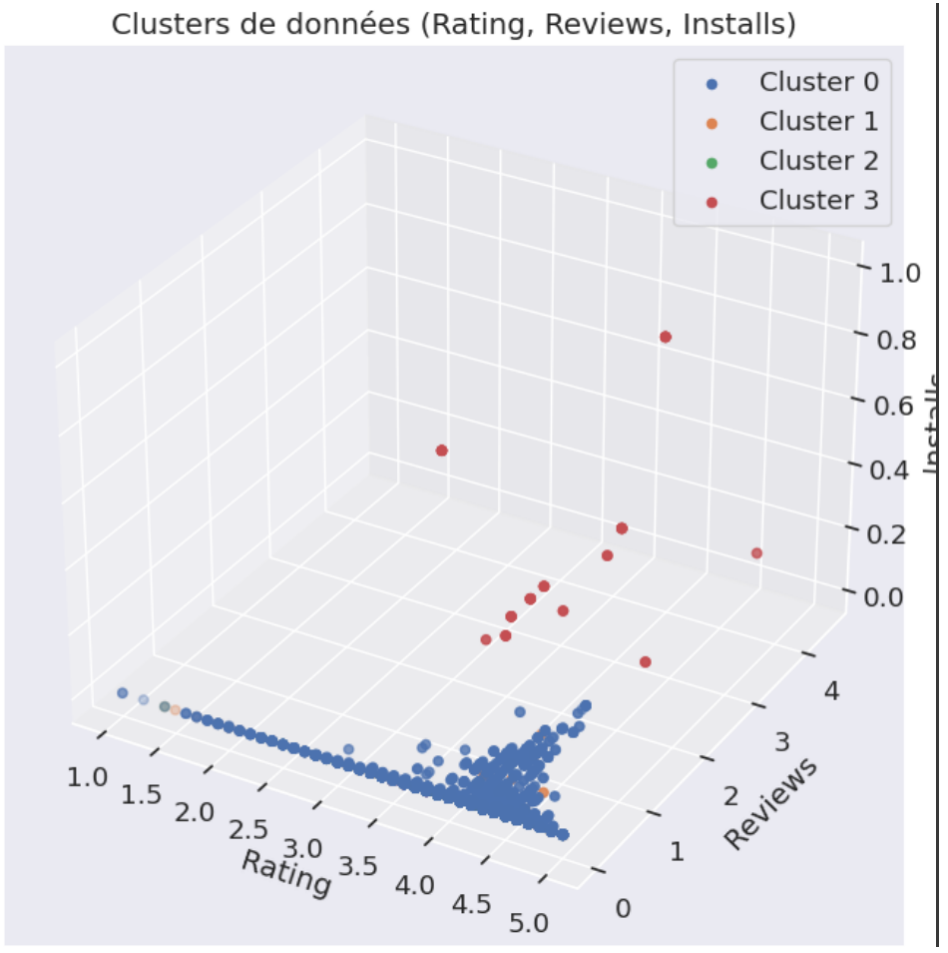
Répartition des mots clés dans les 8 cluster :

Mot clé	Cluster0	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6	Cluster7
Free	274				2	8		19
Mobile					2		171	
Pro						191	1	
News			4		101			
Easy				29				
Official								
App					3		11	189

Clustering pour 4 clusters



Pour 8 clusters

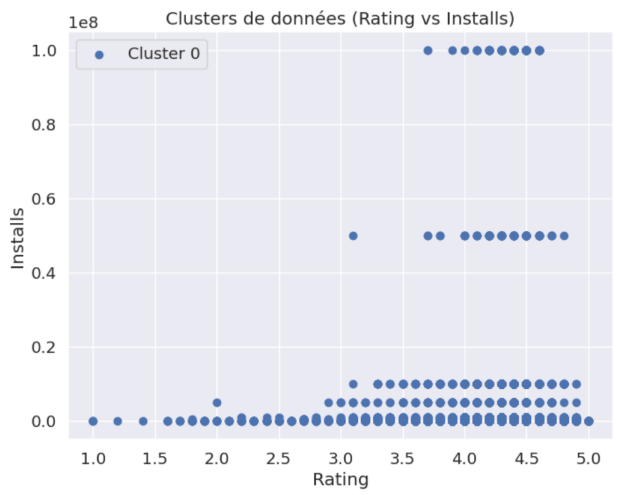


Pour 4 clusters

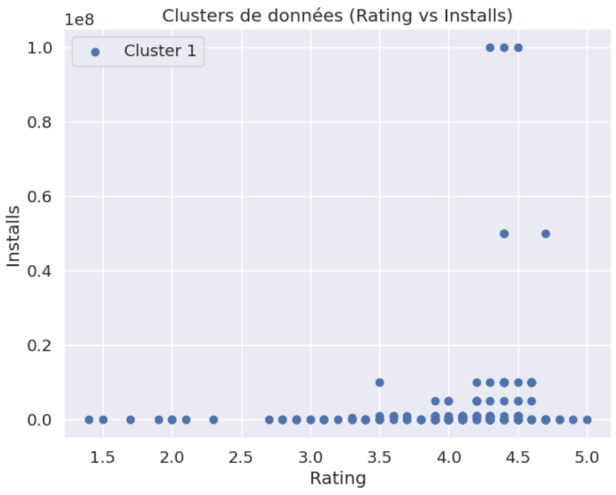
Répartitions des mots clés dans les 4 cluster :

Mot clé	Cluster0	Cluster1	Cluster2	Cluster3
Free	295		8	
Mobile		173		
Pro		1	191	
News		2		
Easy	99			4
Official	29			
App				

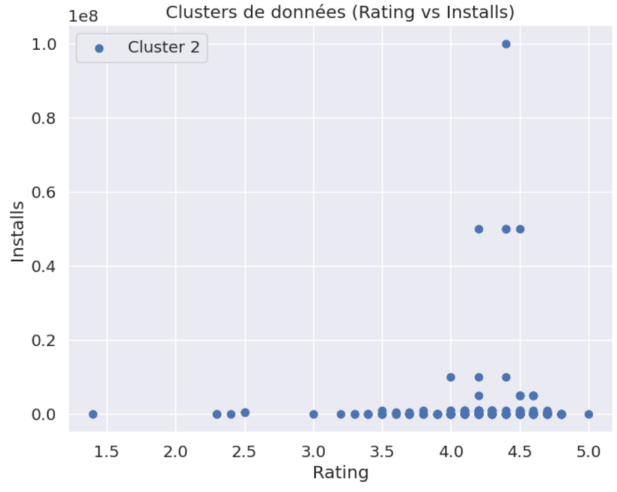
Nb installations / rating par cluster



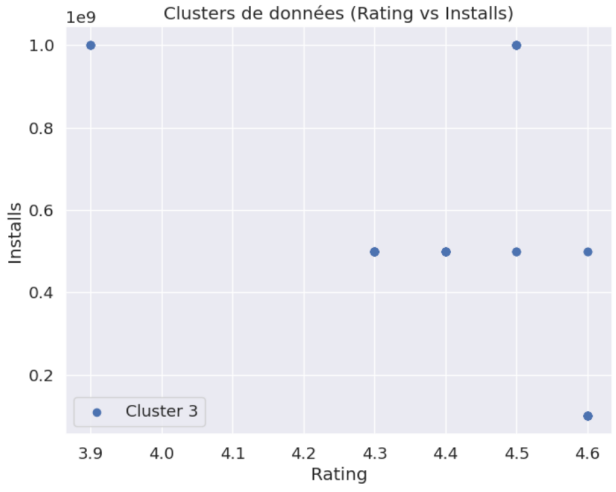
Cluster0



Cluster1

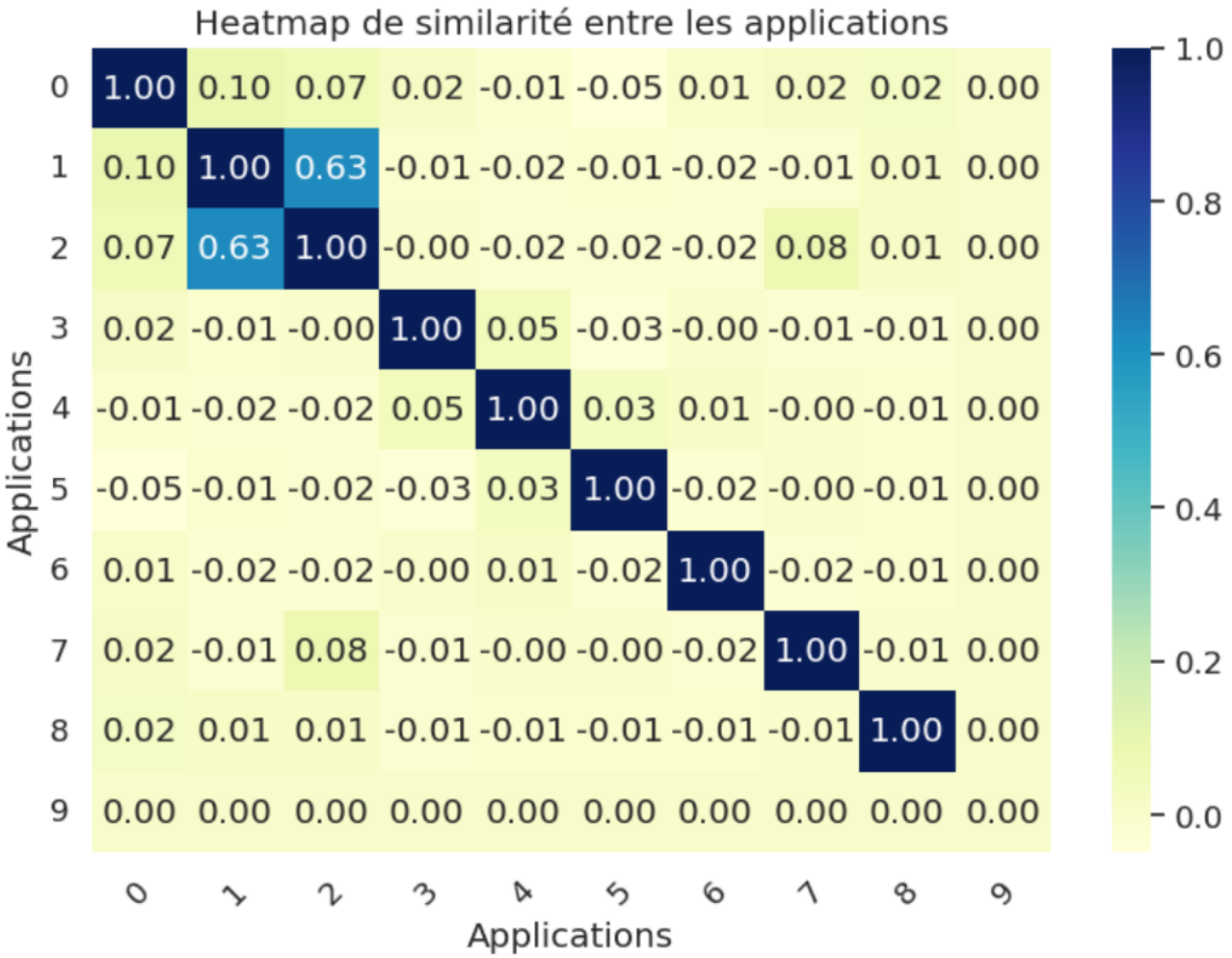


Cluster2



Cluster3

Annexe 5 - Matrice de similarités



Correspondances :
0 → 'Rating',
1 → 'Reviews',
2 → 'Installs',
3 → 'kw_Free',
4 → 'kw_App',
5 → 'kw_Mobile',
6 → 'kw_Pro',
7 → 'kw_News',
8 → 'kw_Easy',
9 → 'kw_Official'