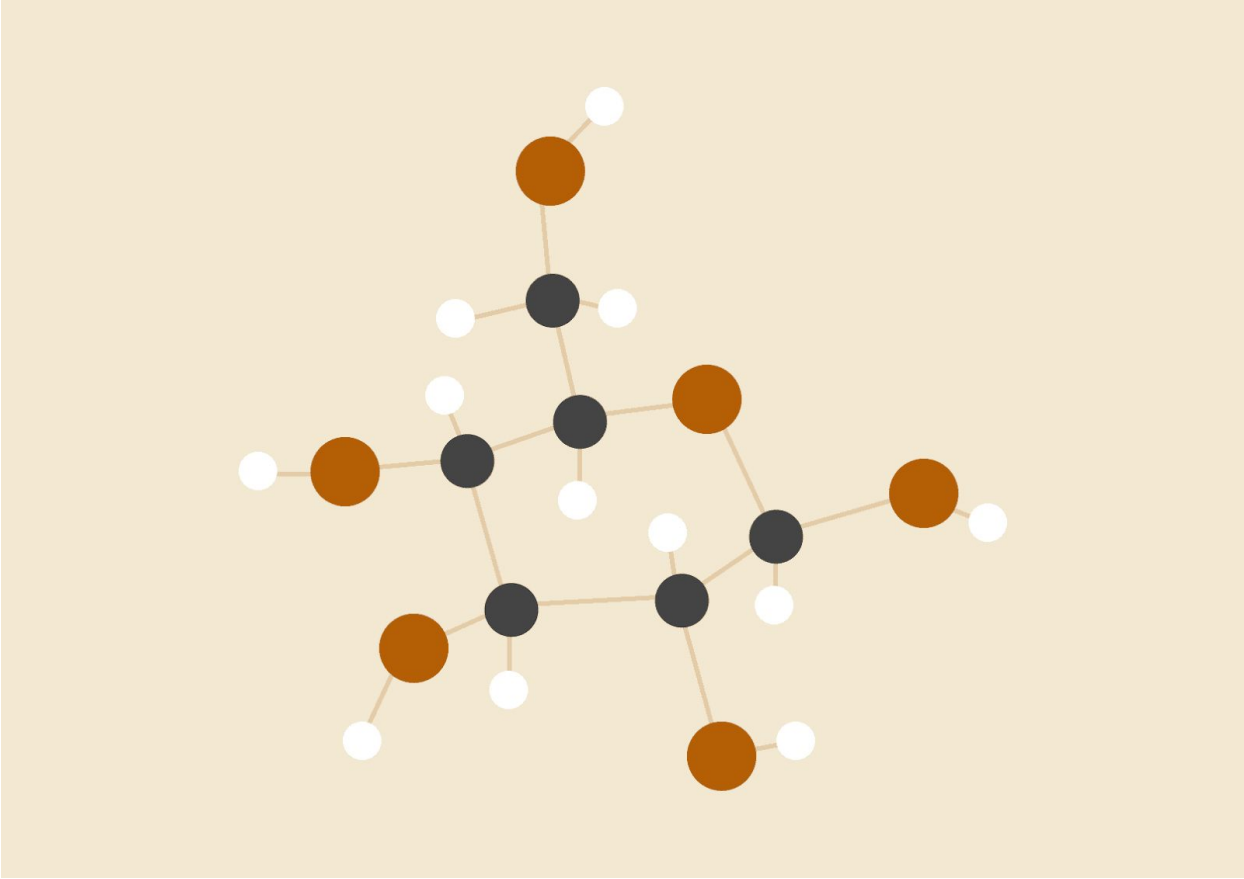
Analyse des applications Android



DE SOUSA Emilio & LAZZALI Samir

Projet MAD 2019

INTRODUCTION

Dans le cadre du projet d'analyse de données, nous avons choisi d'étudier les applications disponibles dans le googleplay store, une plateforme de téléchargement d'applications pour Android. Concernant ces applications, nous avons toutes les informations accessibles publiquement, que ce soit le nombre de téléchargement, la date de parution, ou encore la note.

Seule 2 variables étaient exploitables à l'origine. Ceci était dû aux formats des autres variables qui étaient considérées comme des chaînes de caractères. C'est pourquoi nous avons du effectuer un nettoyage de données assez lourd pour voir quelles données étaient finalement exploitables. Ainsi, par exemple, nous avons transformer le champs « Installs » qui était constitué de valeurs du type : « 50k », « 1M » en « 50 000 » et « 1 000 000 ».

A notre disposition, 10 841 observations avec 13 variables. Dans ce rapport, nous allons présenter les différentes analyses que nous avons effectuées, les observations qui en découlent. Pour cela nous avons d'abord étudié les variables unes à unes, puis deux à deux, les corrélations entre les elles pour ensuite nous concentrer sur l'analyses en composante principale.

DESCRIPTION

Importation et forme du DataSet

Apercu des données :

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	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
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50,000,000+	Free	0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2.8M	100,000+	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up

On calcule le pourcentage de valeur null en fonction des variables, Forme : (10841, 13).

	Total	Percent
Rating	1474	0.135965
Current Ver	8	0.000738
Android Ver	3	0.000277
Content Rating	1	0.000092
Type	1	0.000092
Last Updated	0	0.000000

Nettoyage des données

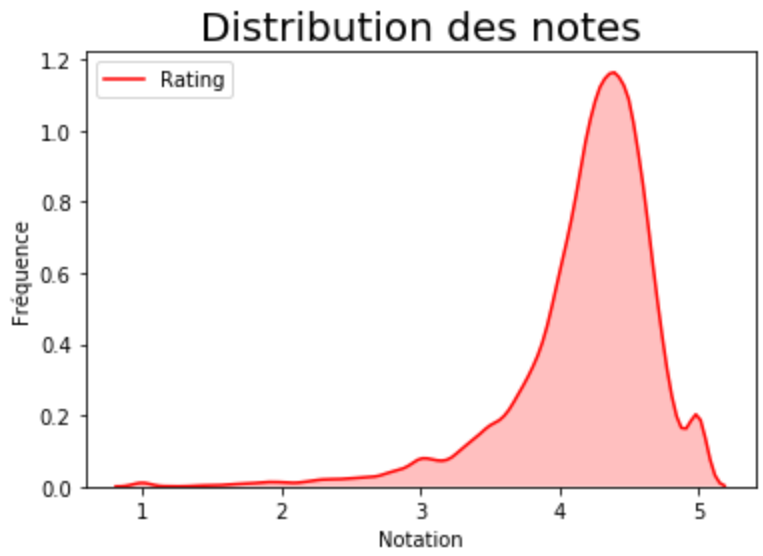
Forme : (9360, 13)

	Total	Percent
Android Ver	0	0.0
Current Ver	0	0.0
Last Updated	0	0.0
Genres	0	0.0
Content Rating	0	0.0
Price	0	0.0

STATISTIQUE DESCRIPTIVE UNIDIMENSIONNEL

Rating

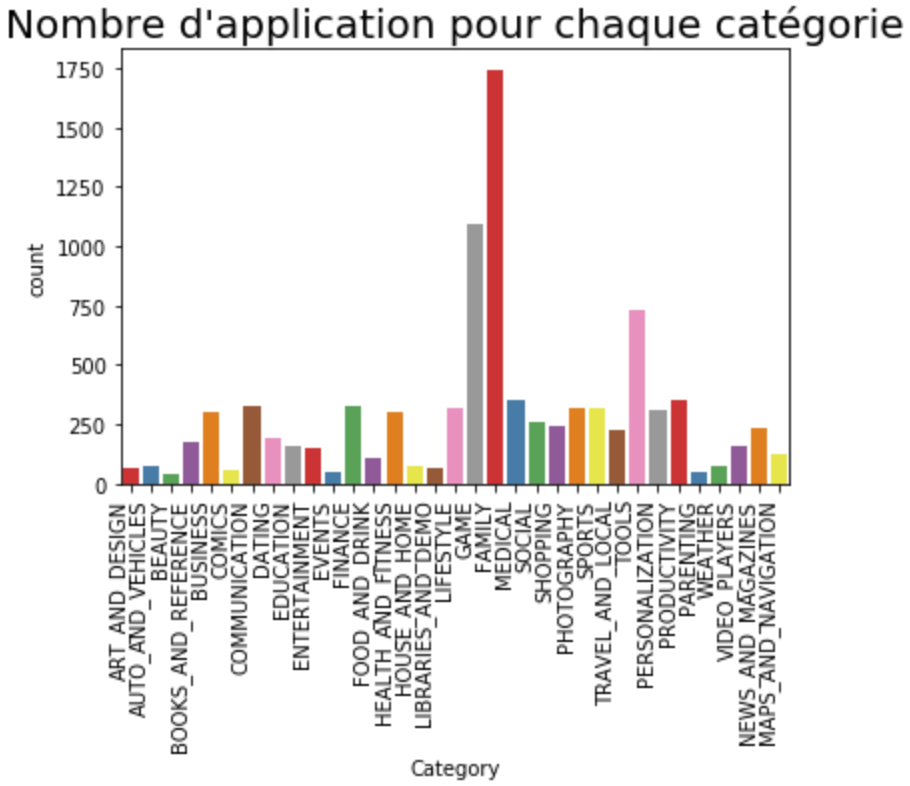
```
count    9360.000000
mean        4.191838
std         0.515263
min         1.000000
25%         4.000000
50%         4.300000
75%         4.500000
max         5.000000
Name: Rating, dtype: float64
```



Category

```
33 categories

['ART_AND_DESIGN' 'AUTO_AND_VEHICLES' 'BEAUTY' 'BOOKS_AND_REFERENCE'
'BUSINESS' 'COMICS' 'COMMUNICATION' 'DATING' 'EDUCATION' 'ENTERTAINMENT'
'EVENTS' 'FINANCE' 'FOOD_AND_DRINK' 'HEALTH_AND_FITNESS' 'HOUSE_AND_HOME'
'LIBRARIES_AND_DEMO' 'LIFESTYLE' 'GAME' 'FAMILY' 'MEDICAL' 'SOCIAL'
'SHOPPING' 'PHOTOGRAPHY' 'SPORTS' 'TRAVEL_AND_LOCAL' 'TOOLS'
'PERSONALIZATION' 'PRODUCTIVITY' 'PARENTING' 'WEATHER' 'VIDEO_PLAYERS'
'NEWS_AND_MAGAZINES' 'MAPS_AND_NAVIGATION']
```

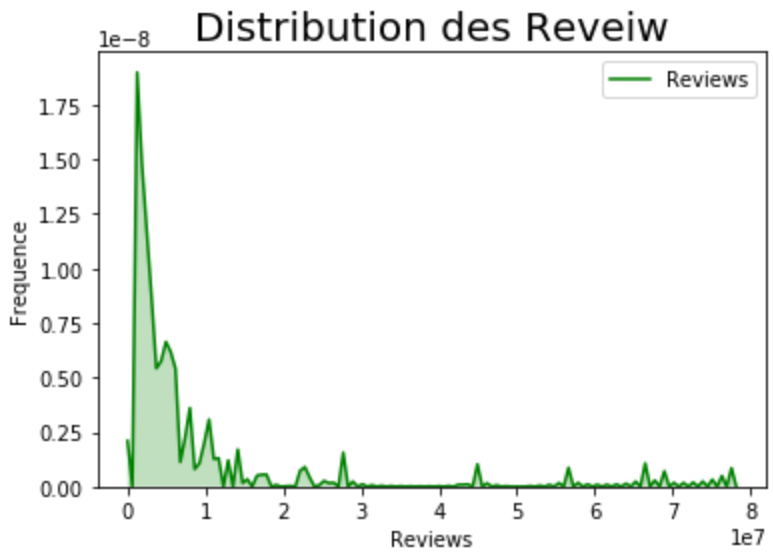


Les catégories Jeu et Famille sont les plus populaires pour les applications.

Reviews

```
0       159
1       967
2     87510
3    215644
4       967
Name: Reviews, dtype: object
```

Les données sont encore dans le type d'objet, nous avons besoin de convertir les en int.

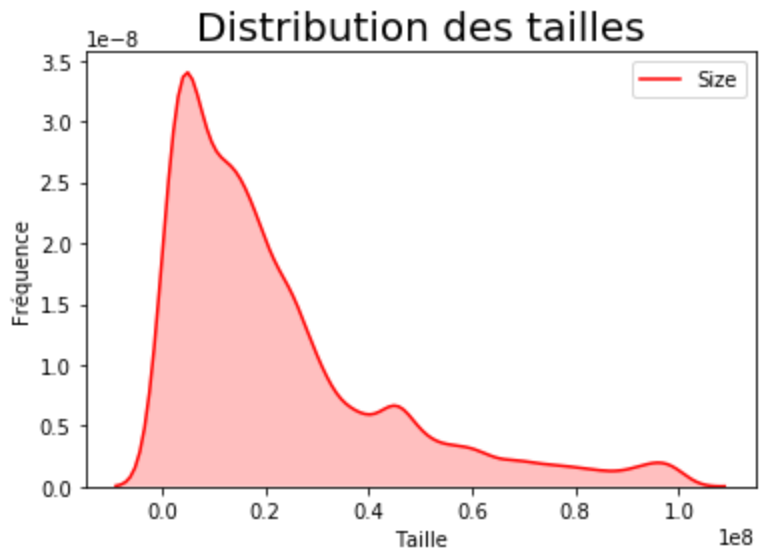


La plupart des applications ont moins d'un million d'évaluations. Évidemment, les applications bien connues ont beaucoup d'évaluations

Size

```
array(['19M', '14M', '8.7M', '25M', '2.8M', '5.6M', '29M', '33M', '3.1M',
       '28M', '12M', '20M', '21M', '37M', '5.5M', '17M', '39M', '31M',
       '4.2M', '23M', '6.0M', '6.1M', '4.6M', '9.2M', '5.2M', '11M',
       '24M', 'Varies with device', '9.4M', '15M'], dtype=object)
```

Les données sont toujours dans du type objet et contiennent l'unité, il y aussi et des **Varies with device** à supprimer.
Nettoyage des données : On les convertis d'abord en NA. Puis on supprime les unité **k** ou **M**. De plus on remplace les NA par la moyenne des tailles.



Installs

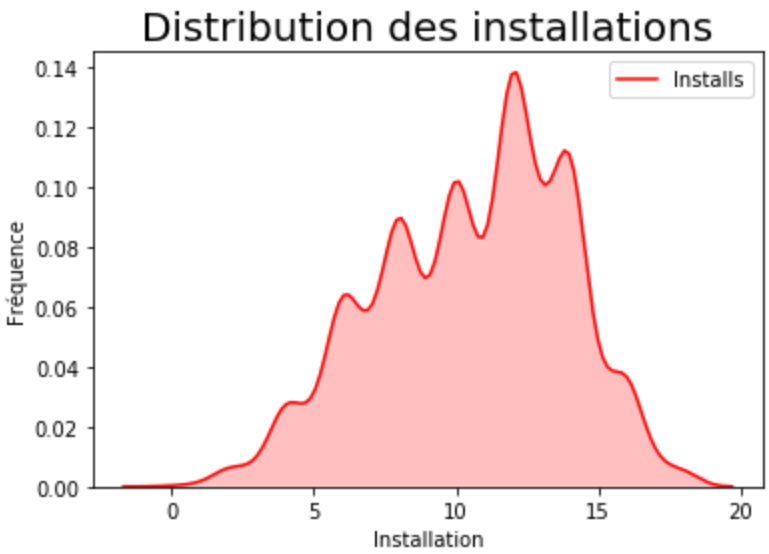
```
array(['10,000+', '500,000+', '5,000,000+', '50,000,000+', '100,000+',
       '50,000+', '1,000,000+', '10,000,000+', '5,000+', '100,000,000+'],
      dtype=object)
```

Les données sont toujours dans le type d'objet et contiennent le signe +.

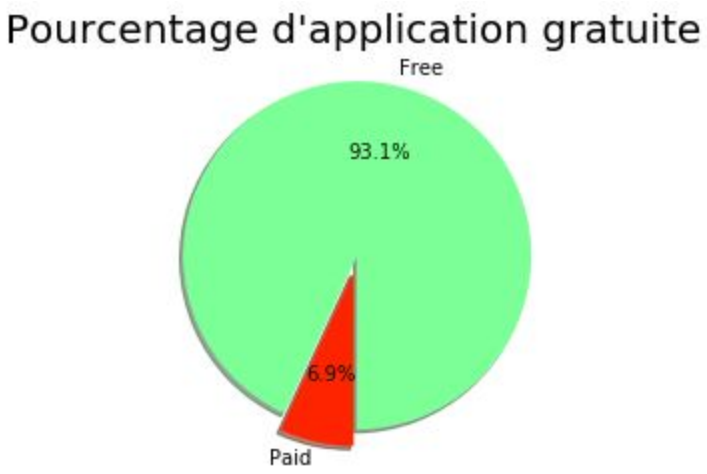
Nettoyage des donées, on les transforme

- 0 = 1+
- 1 = 5+
- 2 = 10+
- etc

```
array([     10000,     500000,    5000000,   50000000,     100000,
           50000,    1000000,   10000000,       5000,  100000000,
      1000000000,       1000,  500000000,        100,        500,
              10,          5,         50,          1], dtype=int64)
```



Type (gratuit payant)



La plupart des applications sont gratuites (93,1%).

Price

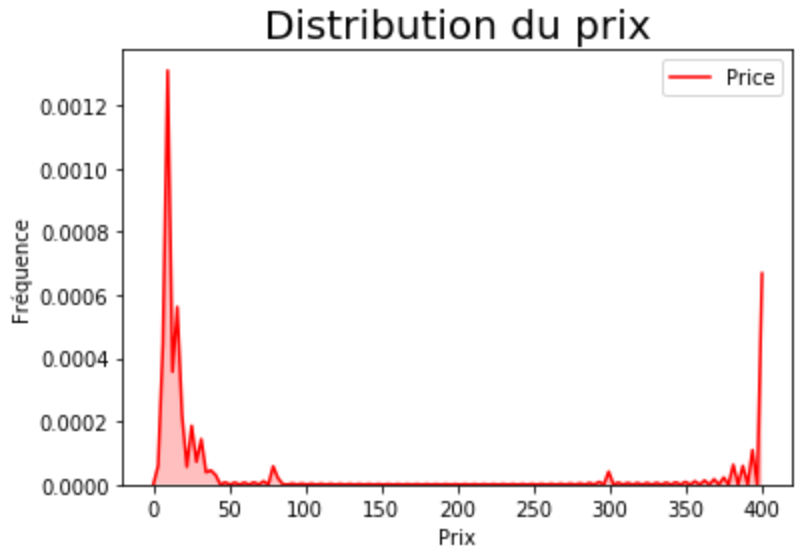
```
array(['0', '$4.99', '$3.99', '$6.99', '$7.99', '$5.99', '$2.99', '$3.49',
       '$1.99', '$9.99', '$7.49', '$0.99', '$9.00', '$5.49', '$10.00',
       '$24.99', '$11.99', '$79.99', '$16.99', '$14.99', '$29.99',
       '$12.99', '$2.49', '$10.99', '$1.50', '$19.99', '$15.99', '$33.99',
       '$39.99', '$3.95'], dtype=object)
```

Nettoyage des données, on supprime le **$** des prix avant de les convertir en float.

```
count    9360.000000
mean        0.961279
std        15.821640
min         0.000000
25%         0.000000
50%         0.000000
75%         0.000000
max       400.000000
Name: Price, dtype: float64
```

Le prix moyen est d'environ 0,96, mais la plupart sont gratuits (8715/9360).

L'application la plus chère est à **400 dollars** : **I'm Rich - Trump Edition**



Transformation du Type pour Free (0 ou 1) pour la suite de l'analyse

Content Rating

```
array(['Everyone', 'Teen', 'Everyone 10+', 'Mature 17+',
       'Adults only 18+', 'Unrated'], dtype=object)
```

Nous avons supprimé la seule ligne avec un Content Rating : *Unrated.*
Nous reviendrons sur cette variable lors de l'analyse 2D

Genres

```
115 genres
['Art & Design' 'Art & Design;Pretend Play' 'Art & Design;Creativity' 'Auto &
Vehicles' 'Beauty' 'Books & Reference' 'Business' 'Comics' 'Comics;Creativity'
'Communication' 'Dating' 'Education;Education' 'Education' 'Education;Creativity'
```

```
'Education;Music & Video' 'Education;Action & Adventure' 'Education;Pretend Play'
'Education;Brain Games' 'Entertainment' 'Entertainment;Music & Video'
'Entertainment;Brain Games' 'Entertainment;Creativity' 'Events' 'Finance' 'Food &
Drink' 'Health & Fitness' 'House & Home' 'Libraries & Demo' 'Lifestyle'
'Lifestyle;Pretend Play' 'Adventure;Action & Adventure' 'Arcade' 'Casual' 'Card'
'Casual;Pretend Play' 'Action' 'Strategy' 'Puzzle' 'Sports' 'Music' 'Word' 'Racing'
'Casual;Creativity' 'Casual;Action & Adventure' 'Simulation' 'Adventure' 'Board'
'Trivia' 'Role Playing' 'Simulation;Education' 'Action;Action & Adventure'
'Casual;Brain Games' 'Simulation;Action & Adventure' 'Educational;Creativity'
'Puzzle;Brain Games' 'Educational;Education' 'Card;Brain Games' 'Educational;Brain
Games' 'Educational;Pretend Play' 'Entertainment;Education' 'Casual;Education'
'Music;Music & Video' 'Racing;Action & Adventure' 'Arcade;Pretend Play' 'Role
Playing;Action & Adventure' 'Simulation;Pretend Play' 'Puzzle;Creativity'
'Sports;Action & Adventure' 'Educational;Action & Adventure' 'Arcade;Action &
Adventure' 'Entertainment;Action & Adventure' 'Puzzle;Action & Adventure'
'Strategy;Action & Adventure' 'Music & Audio;Music & Video' 'Health &
Fitness;Education' 'Adventure;Education' 'Board;Brain Games' 'Board;Action &
Adventure' 'Board;Pretend Play' 'Casual;Music & Video' 'Role Playing;Pretend Play'
'Entertainment;Pretend Play' 'Video Players & Editors;Creativity' 'Card;Action &
Adventure' 'Medical' 'Social' 'Shopping' 'Photography' 'Travel & Local' 'Travel &
Local;Action & Adventure' 'Tools' 'Tools;Education' 'Personalization' 'Productivity'
'Parenting' 'Parenting;Music & Video' 'Parenting;Brain Games' 'Parenting;Education'
'Weather' 'Video Players & Editors' 'Video Players & Editors;Music & Video' 'News &
Magazines' 'Maps & Navigation' 'Health & Fitness;Action & Adventure' 'Educational'
'Casino' 'Adventure;Brain Games' 'Lifestyle;Education' 'Books & Reference;Education'
'Puzzle;Education' 'Role Playing;Brain Games' 'Strategy;Education' 'Racing;Pretend
Play' 'Communication;Creativity' 'Strategy;Creativity']
```

Les genres sont variés car ils sont parfois rattaché à un sous genre : **Arcade;Action & Adventure**

Nettoyage des données :On va donc ne conserver que le premier Genres et les grouper pour connaître la répartition dans les Genres principaux. On va aussi fusionner **Group Music & Audio et Music. Il en reste 47 contre les 115 du début.**

LastUpdated

Last Update est toujours au format String, nous avons besoin de la transformer la tracer.

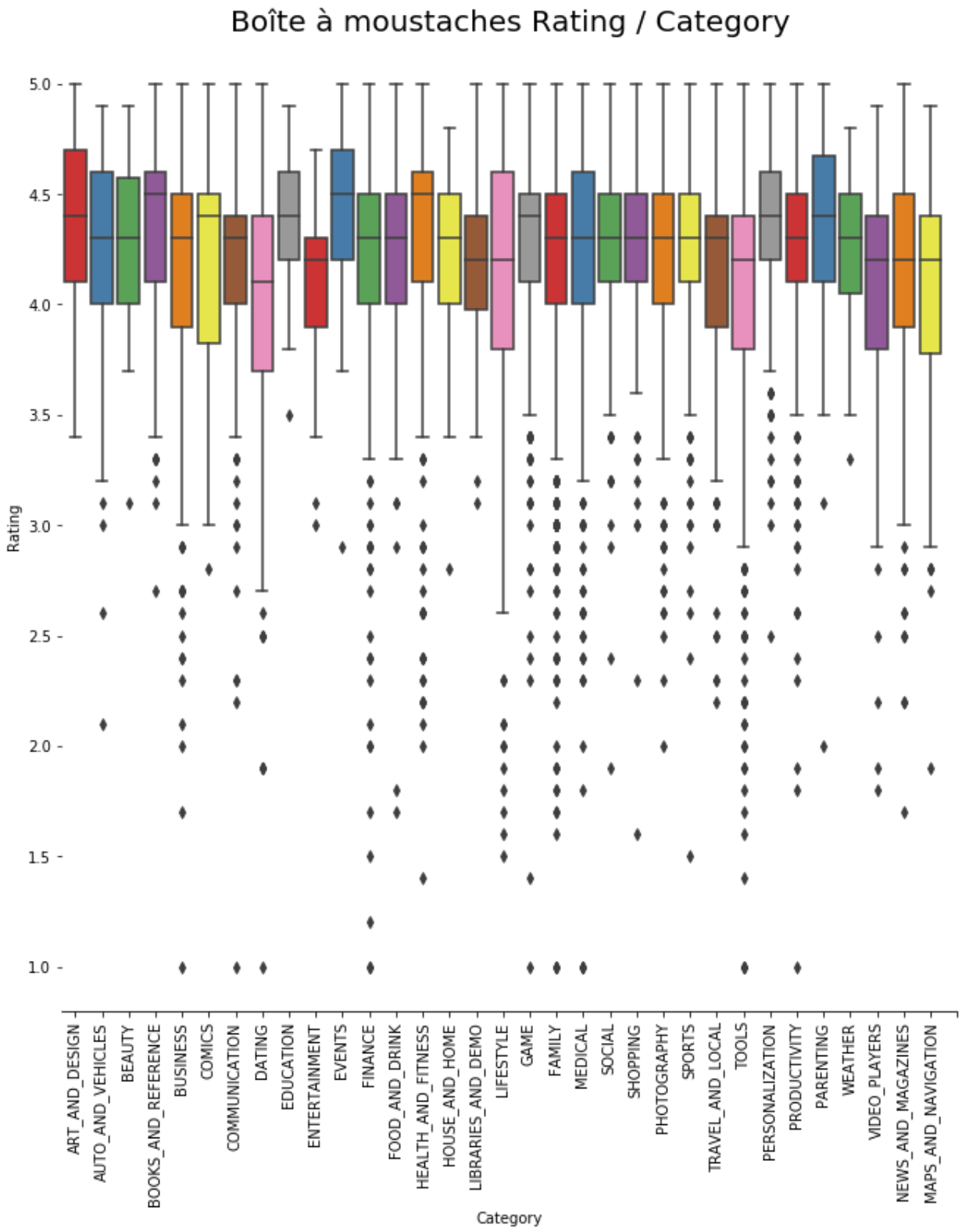
```
0     January 7, 2018
1    January 15, 2018
2      August 1, 2018
3        June 8, 2018
4       June 20, 2018
Name: Last Updated, dtype: object
```

On a décidé de le changer au format "datetime" mais il ne peut toujours pas être utilisé en sous ce format. Nous avons créé une nouvelle feature "lastupdate". Elle contiendra depuis combien de temps cette application a été mise à jour la dernière fois (... il y a quelques jours). Le tout indexé sur l'application la plus à jour : '2018-08-08 00:00:00'.
2018-08-08 00:00:00 est la date la plus à jour, lastupdate aura donc la valeur de la différence avec cette date.

```
0   -213
1   -205
2     -7
3    -61
4    -49
Name: lastupdate, dtype: int64
```

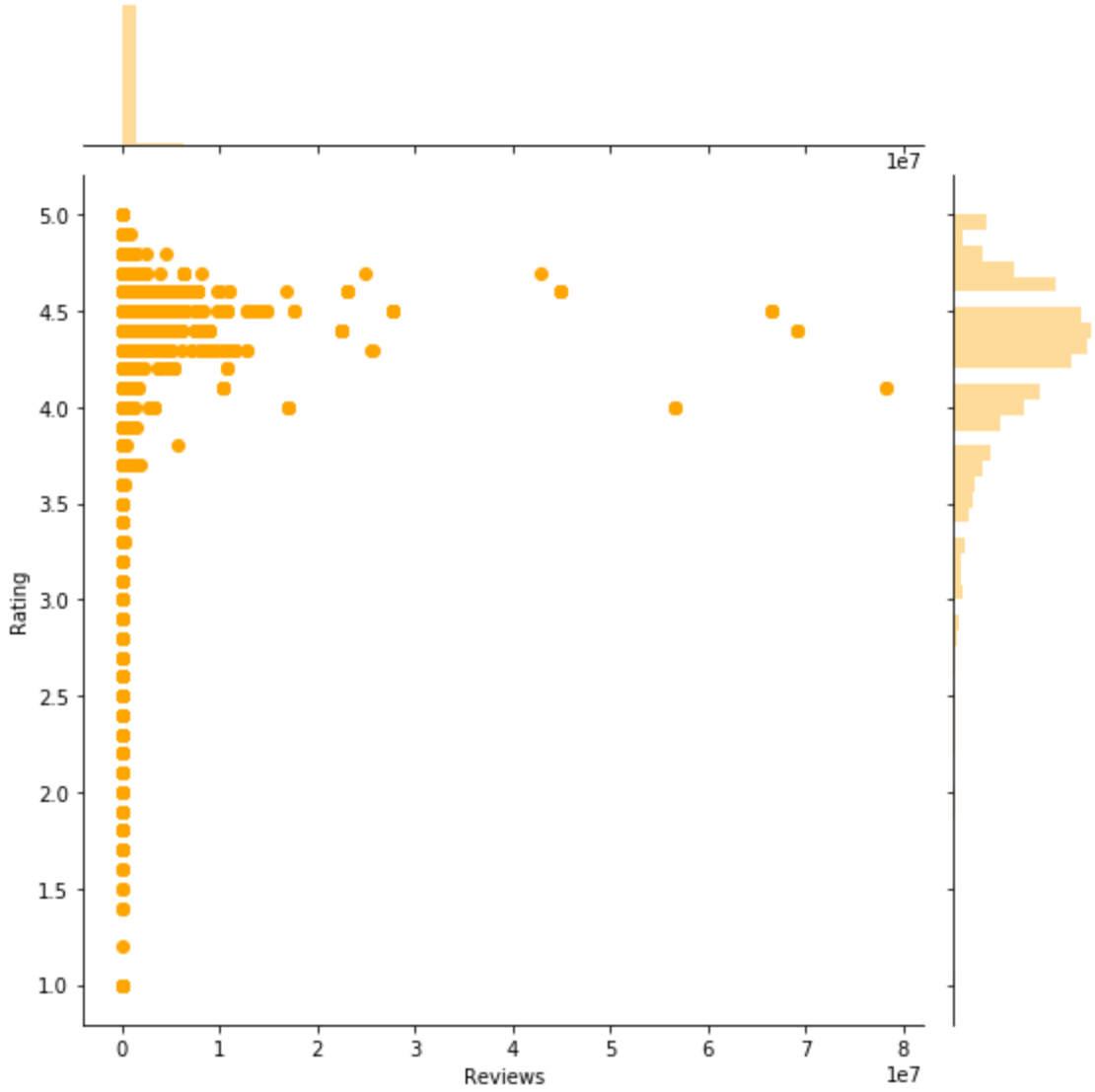
STATISTIQUE DESCRIPTIVE BIDIMENSIONNEL (par rapport au Rating)

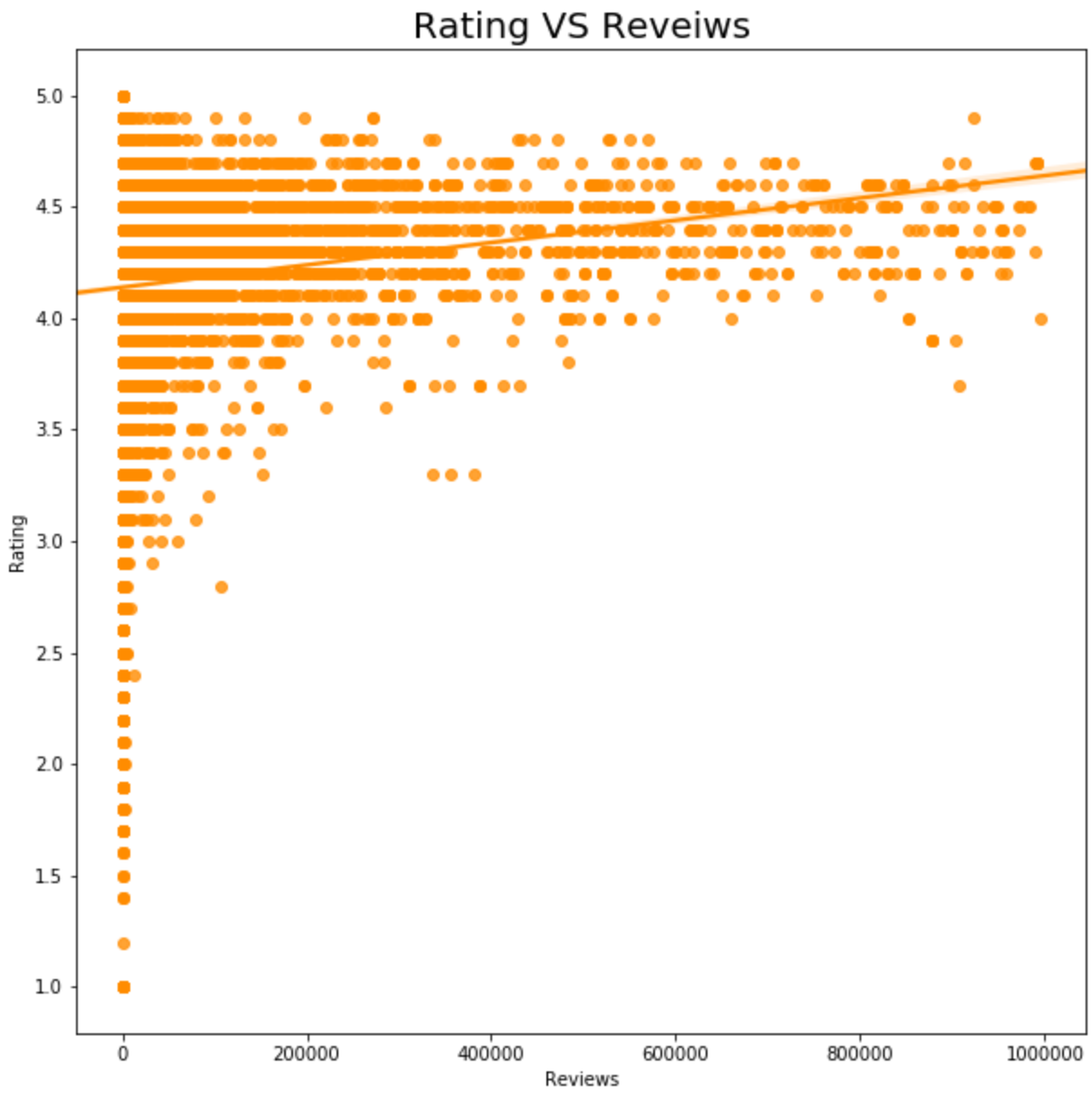
Rating / Category



Le Rating ne diffèrent pas beaucoup pour chaque Category.

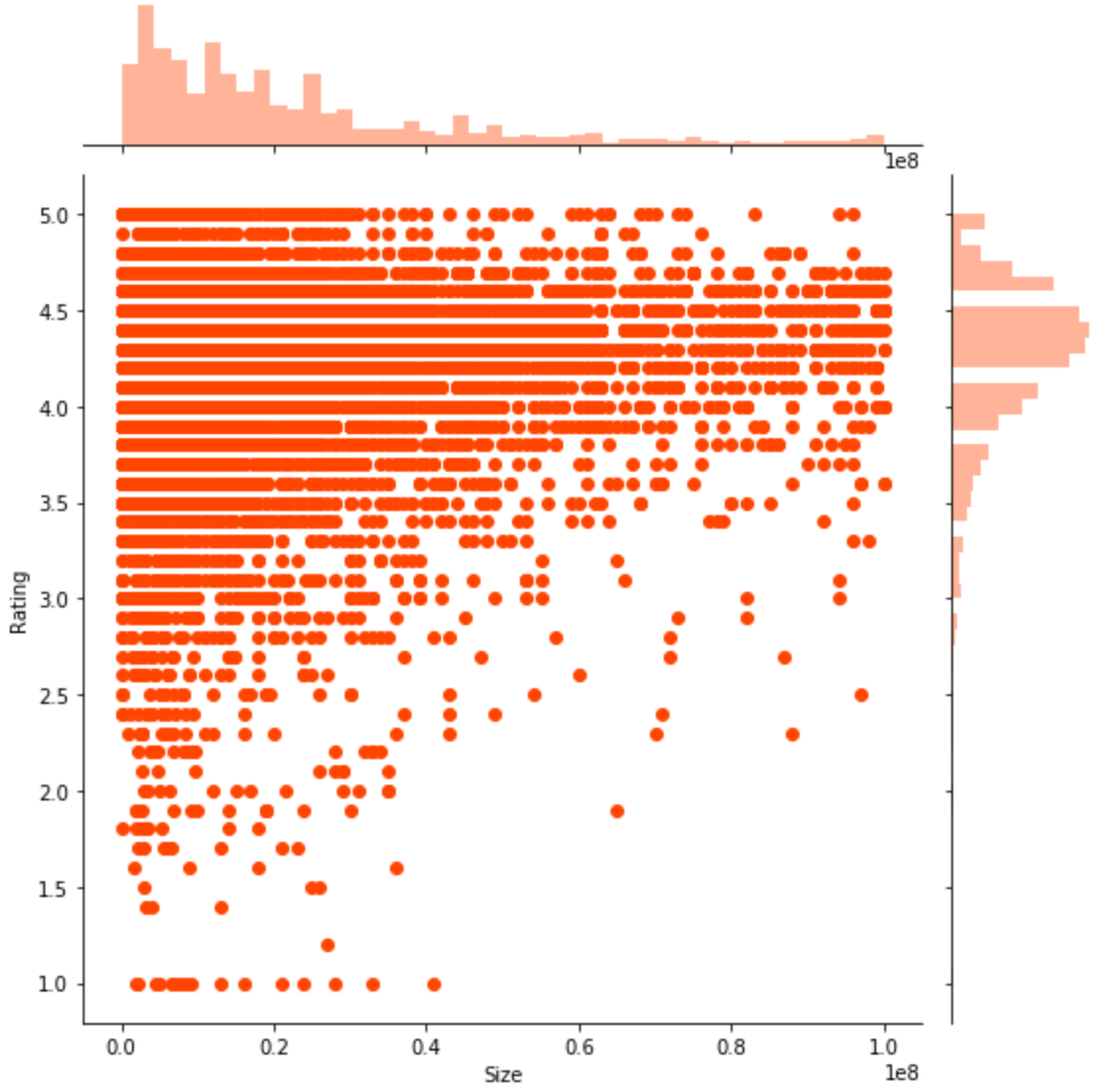
Rating / Reviews



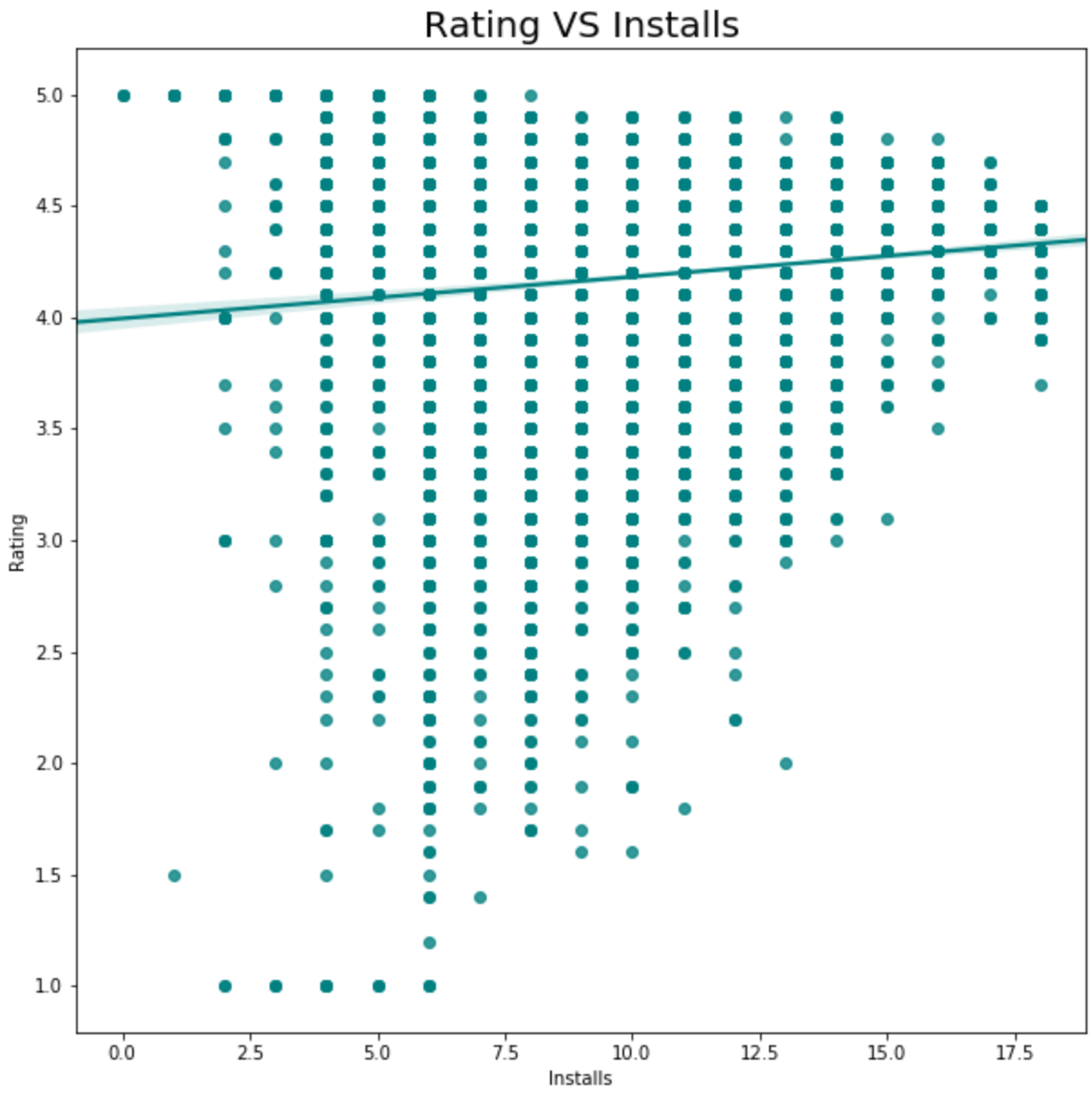


Il semble que les applications populaire obtiennent une bonne note

Rating / Size

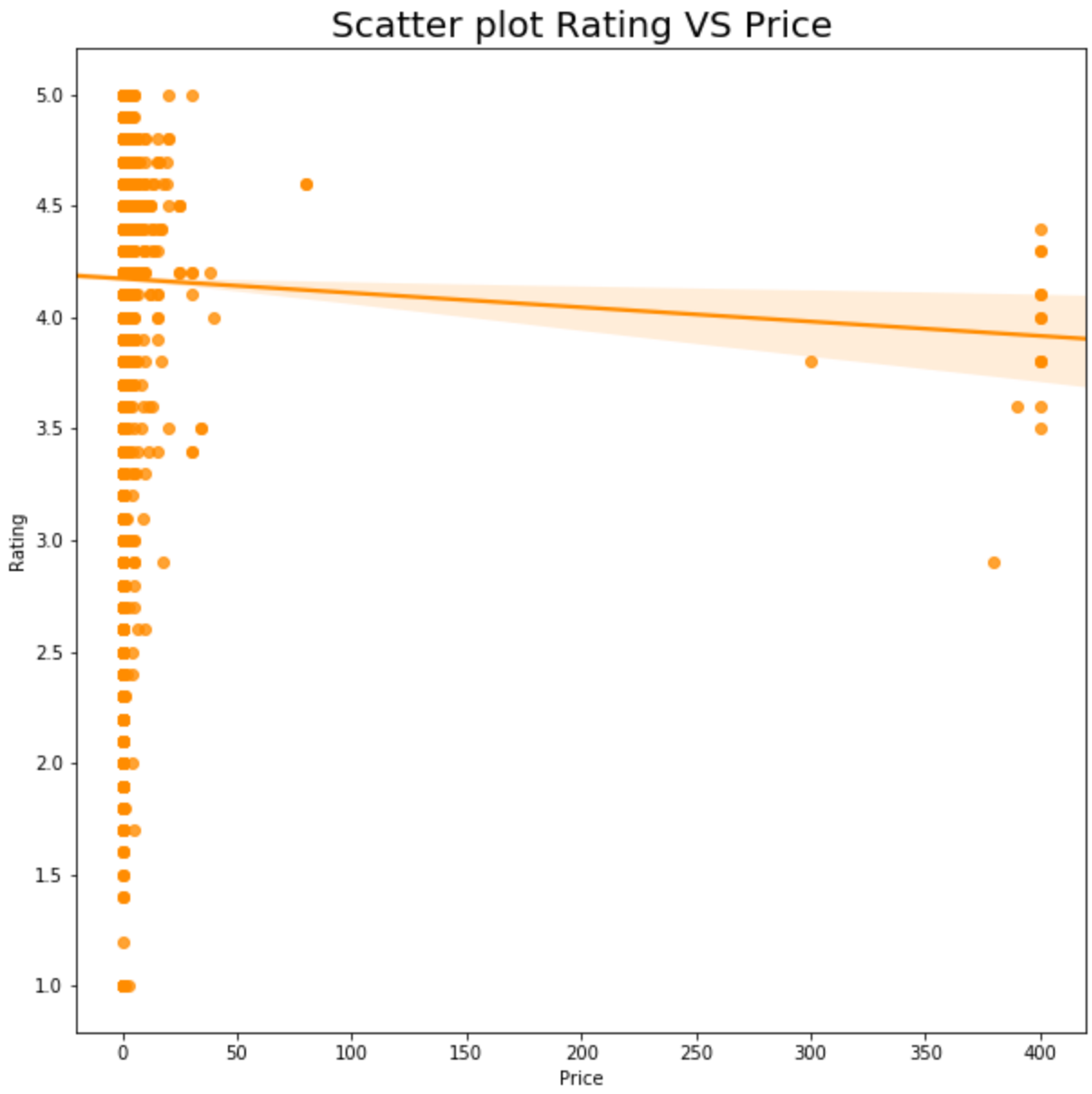


Rating / Installs



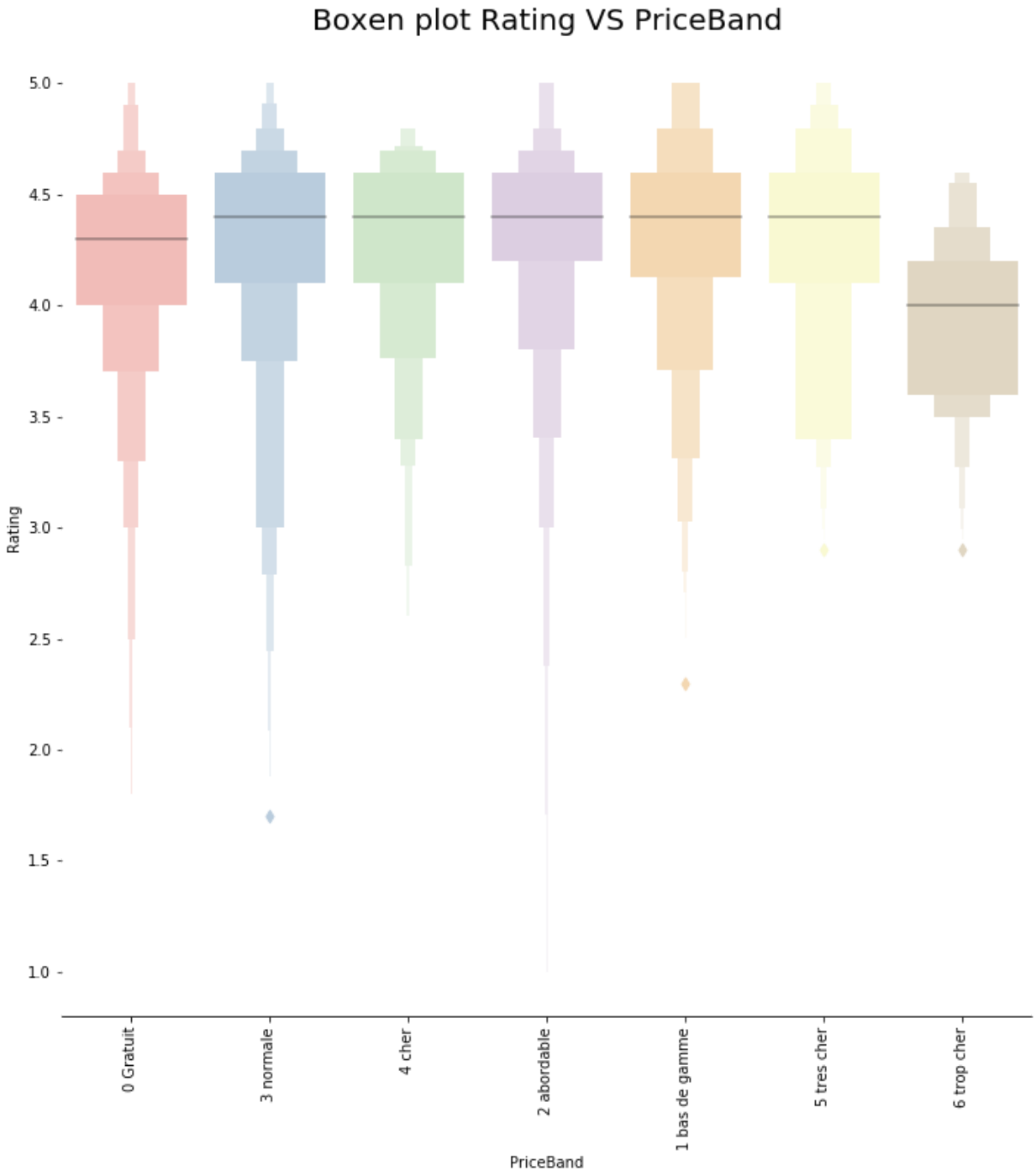
Il semblerait que le nombre d'installation affecte le Rating

Rating / Price



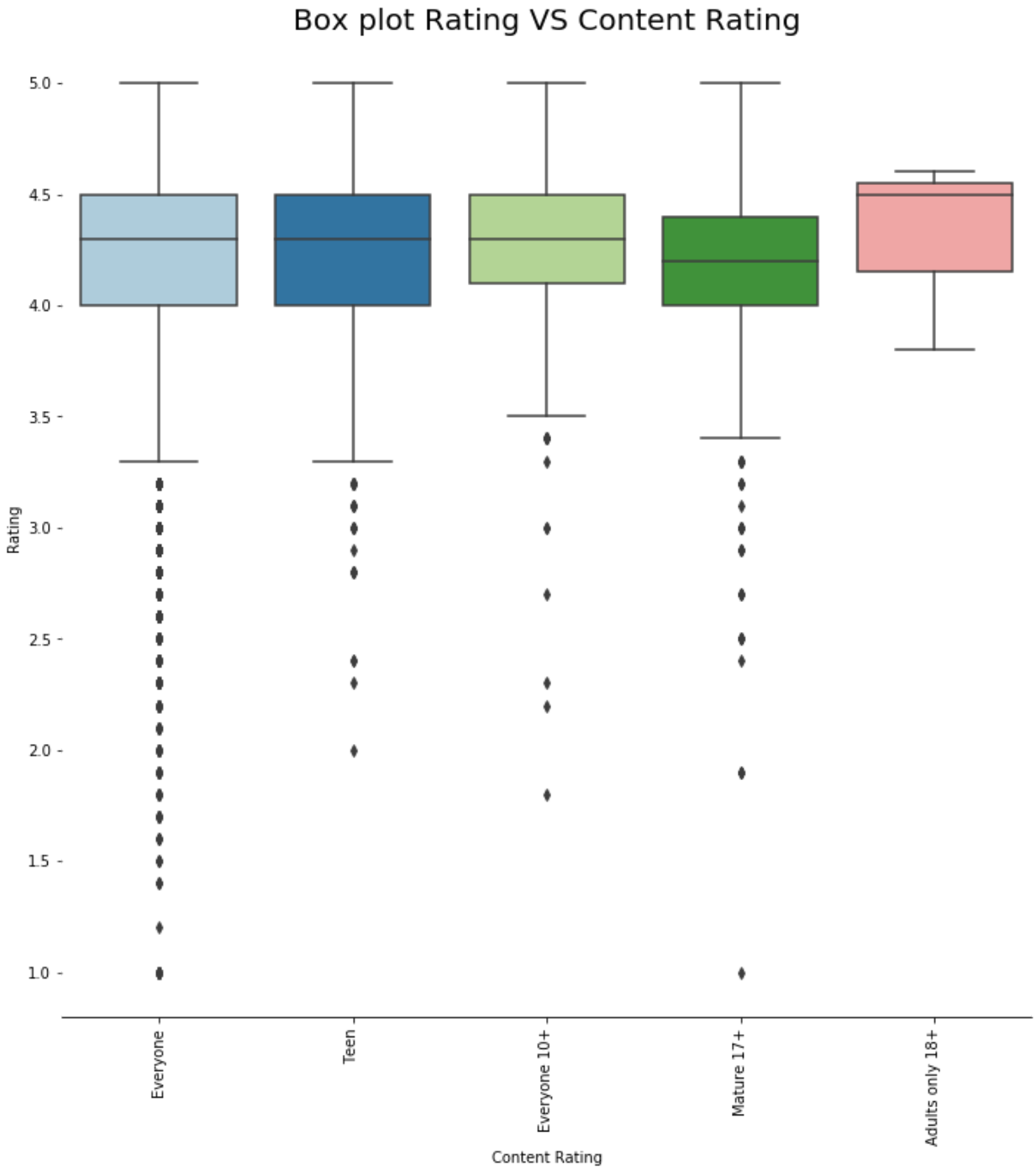
Les applications d'un prix plus élevé semble plus décevoir le client.

Pour la suite nous allons créer des fourchettes de prix :
0 : '0 Gratuit' 0.01 <= 0.99: '1 bas de gamme' 0.99 <= 2.99 : '2 abordables' 2.99 <= 4.99): '3 normale' 4.99 <= 14.99): '4 cher' 14.99 <= 29.99): '5 tres cher' supérieure à 29.99) : '6 trop cher'



Les prix n'ont pas d'effet sur le Rating , mais pour les applications trop cher le Rating peut être plus mauvais

Rating / Content Rating



Le classement du contenu n'a pas trop d'effet sur le Rating, mais dans les applications 'Matures', ils ont l'air d'être moins bien notés que les autres.

Rating / Genres

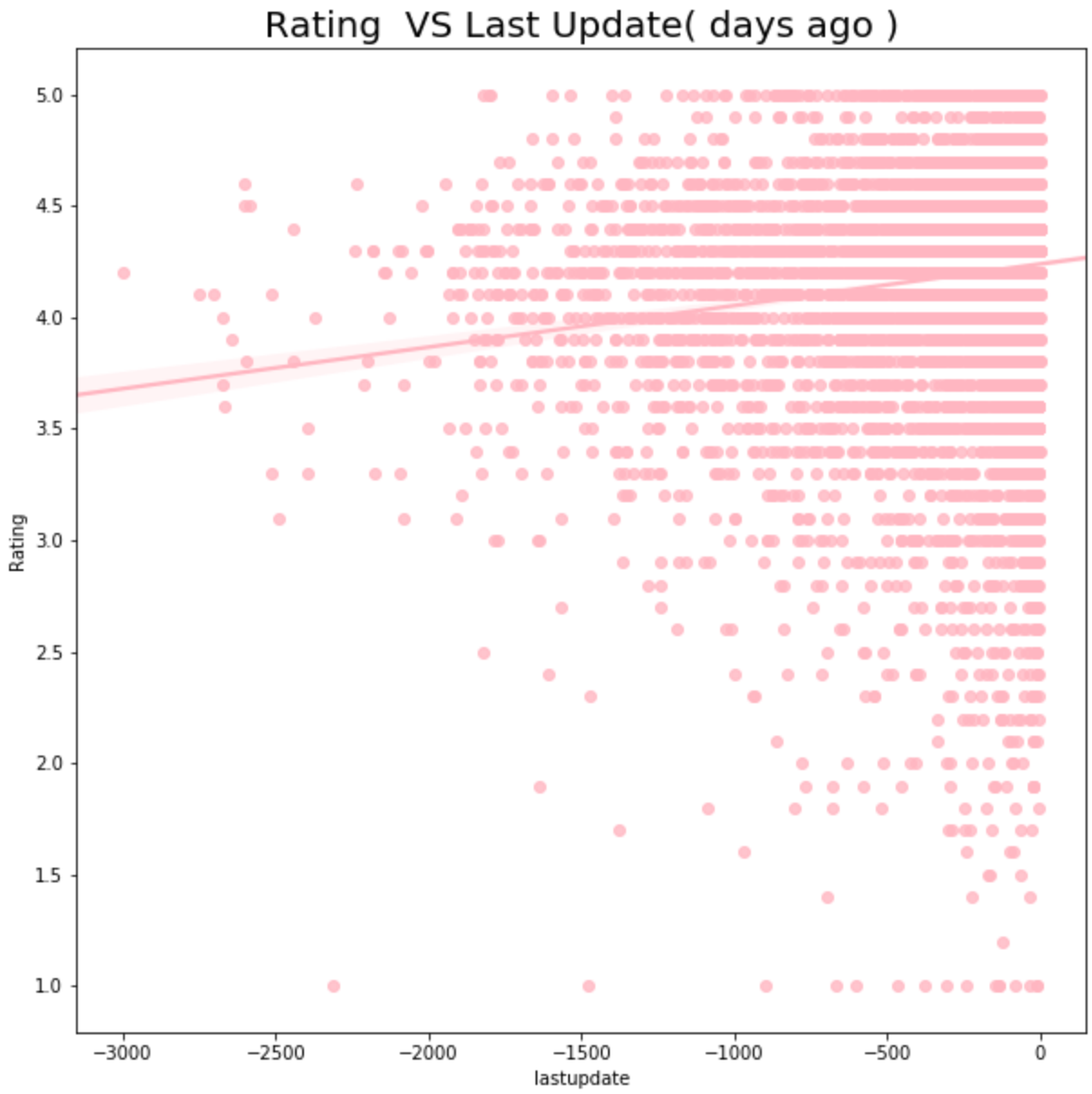
	Rating
count	47.000000
mean	4.210662
std	0.104405
min	3.970769
25%	4.132039
50%	4.198246
75%	4.282529
max	4.435556

	Genres	Rating
14	Dating	3.970769

	Genres	Rating
18	Events	4.435556

Si l'on observe à partir de l'écart-type, le genre n'a pas trop d'effet sur la notation. La plus faible d'une note moyenne sur les genres (Rencontres) est de 3,97 alors que le plus élevé (Événements) est de 4,43.

Rating / Last Update

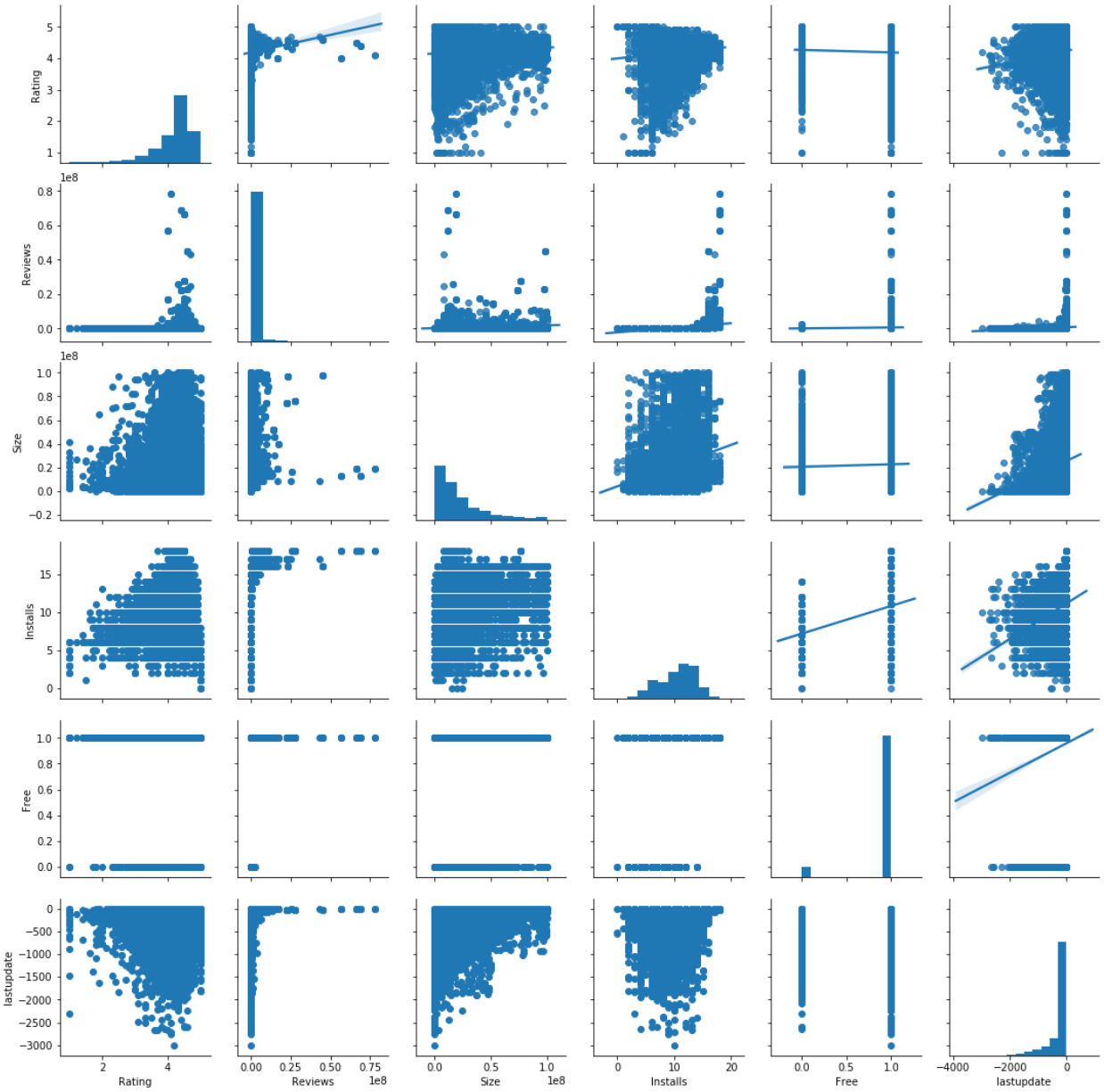


Les application les plus à jour ont les meilleurs notes.

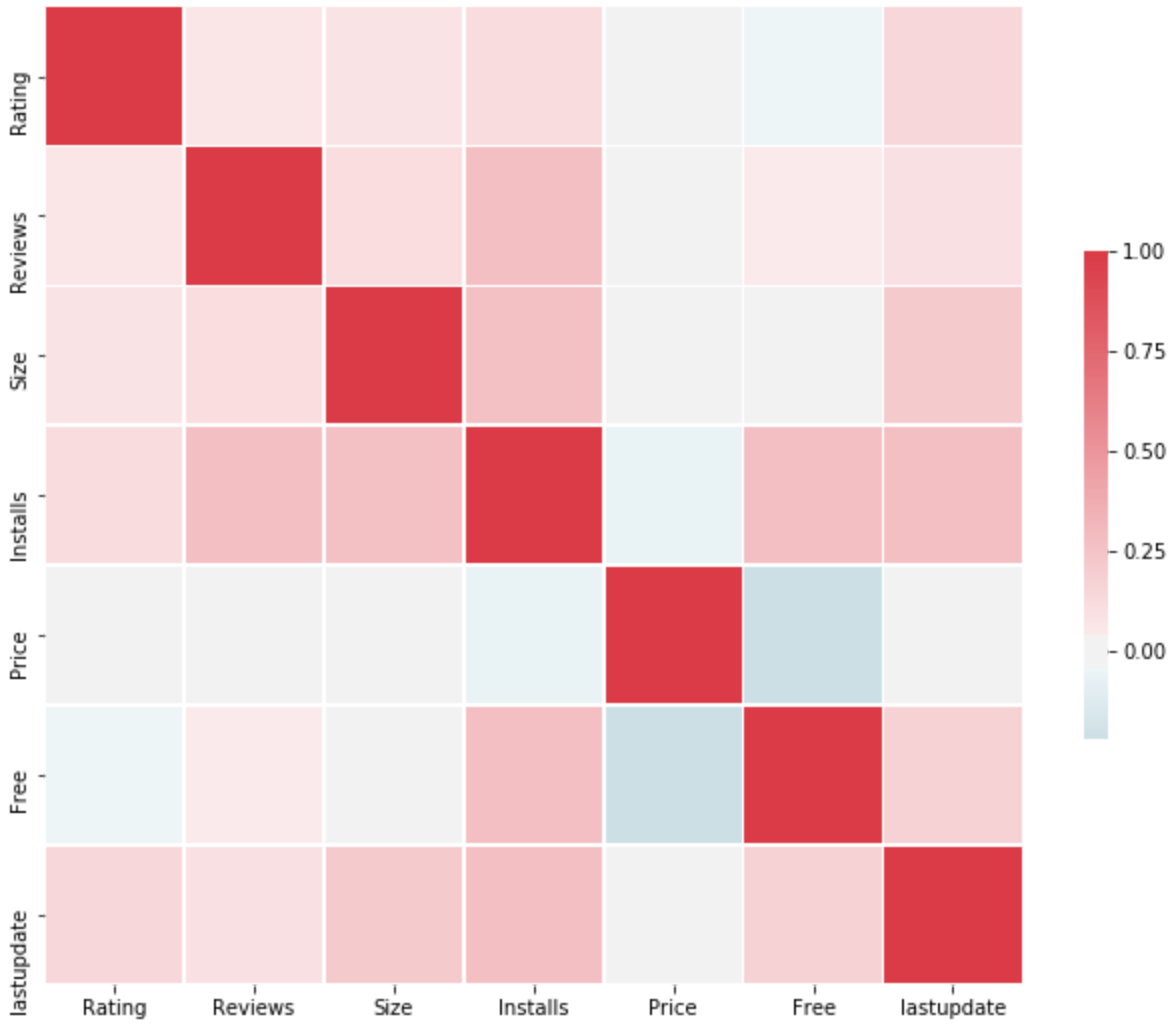
Multidimensional

Voici un extrait des données que nous allons utilisé pour la suite. C'est un récapitulatif de nettoyage résultant de notre analyse 1D.Multidimentional

	Rating	Reviews	Size	Installs	Content Rating	Free	new	lastupdate
0	4.1	159	19000000.0	8	Everyone	1	2018-01-07	-213
1	3.9	967	14000000.0	11	Everyone	1	2018-01-15	-205
2	4.7	87510	8700000.0	13	Everyone	1	2018-08-01	-7
3	4.5	215644	25000000.0	15	Teen	1	2018-06-08	-61
4	4.3	967	2800000.0	10	Everyone	1	2018-06-20	-49



MATRICE DE CORRELATION

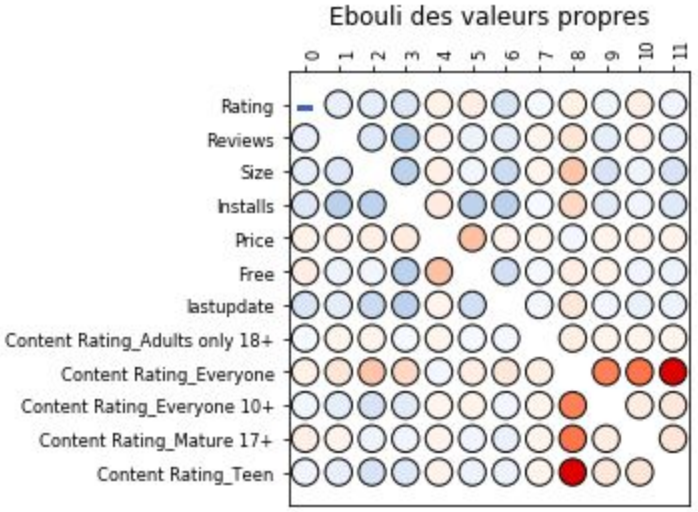


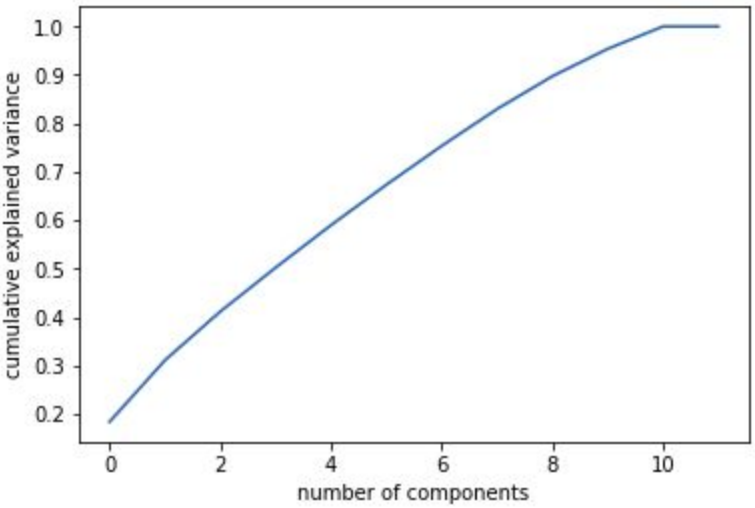
Interprétation de la matrice de corrélation

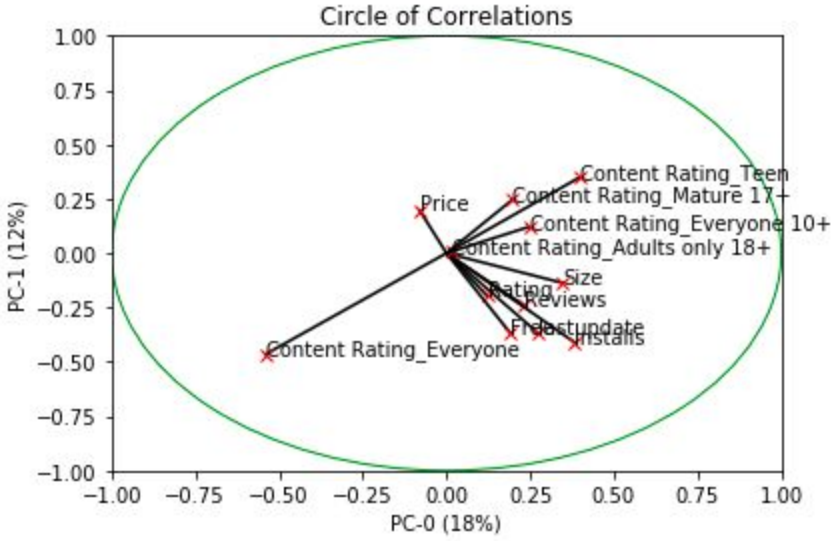
Ici la matrice de corrélation nous permet de synthétiser en quelque sorte les analyses précédentes. Ainsi nous observons que le nombre d'**installation** est corrélé positivement avec le nombre de **review**. Ce qui est plutôt intuitif si l'on considère qu'une application qui est plus téléchargée est susceptible d'avoir plus d'avis (son nombre d'utilisateur étant plus élevé)
Ensuite pour le **Prix** et le fait que l'application soit **gratuite** ou **payante**, cela va de soit. Une application dont les prix est supérieur à 0 n'est pas gratuite d'où la corrélation négative.
Un autre exemple: celui du nombre d'**installations** et le fait qu'elle soit **gratuite**: une application gratuite aura plus facilement un nombre d'installations élevé sachant qu'elle est plus simple d'accès, ce qui explique la corrélation positive.

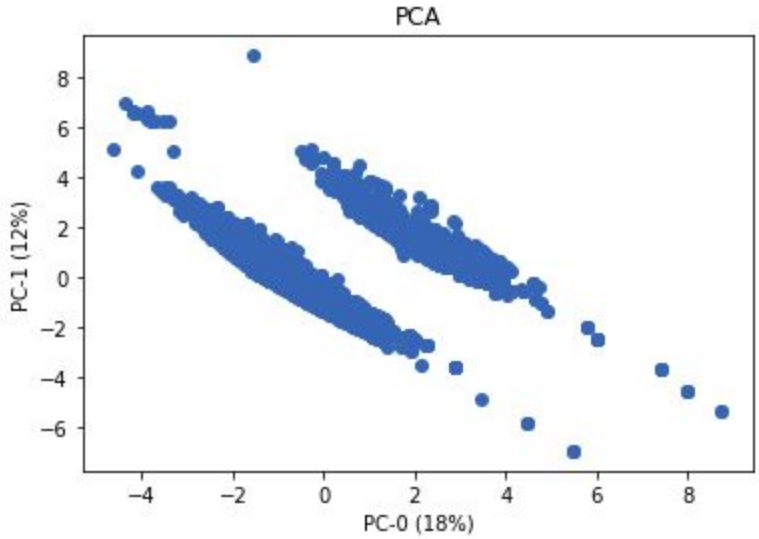
PCA

```
[1.83533033e-01 1.27182177e-01 1.00543199e-01 9.07044209e-02
 8.76357626e-02 8.34585289e-02 8.02533334e-02 7.60719639e-02
 6.77946006e-02 5.65133019e-02 4.63096787e-02 3.41343857e-33]
```






Circle of Correlations
PC-1 (12%)
PC-0 (18%)
Price
Content Rating_Teen
Content Rating_Mature 17+
Content Rating_Everyone 10+
Content Rating_Adults only 18+
Size
Rating
Reviews
Lastupdate
Installs
Content Rating_Everyone


PCA
PC-1 (12%)
PC-0 (18%)

```
(                                    PC-0      PC-1      PC-2      PC-3  \
 Rating                          0.121501 -0.196507  0.346627 -0.097728
 Reviews                         0.230823 -0.238927  0.244148 -0.016843
 Size                            0.346900 -0.137041  0.249260  0.051558
 Installs                        0.383053 -0.410459  0.030725  0.000106
 Price                          -0.080541  0.198349  0.516061  0.063047
 Free                            0.188961 -0.364162 -0.545063 -0.041475
 lastupdate                      0.271174 -0.364738  0.054982  0.096183
 Content Rating_Adults only 18+  0.015856  0.010018 -0.023041  0.015809
 Content Rating_Everyone        -0.540680 -0.465311  0.085197 -0.028739
 Content Rating_Everyone 10+     0.246570  0.121046  0.279907  0.419227
 Content Rating_Mature 17+       0.196704  0.245289 -0.315873  0.616099
 Content Rating_Teen             0.396339  0.347303 -0.069413 -0.645101

                                      PC-4      PC-5      PC-6      PC-7  \
  Rating                          0.259721  0.083297 -0.670997  0.419480
  Reviews                        -0.149587 -0.072638  0.501990  0.614111
  Size                            0.011770 -0.023265 -0.055609 -0.381734
  Installs                        0.002518 -0.031605  0.214206  0.025555
  Price                           0.361320 -0.055587  0.425275 -0.259388
  Free                           -0.111388  0.002670  0.068497 -0.137249
  lastupdate                      0.381505  0.011576 -0.094124 -0.355985
  Content Rating_Adults only 18+  0.073102  0.984545  0.141789  0.022603
  Content Rating_Everyone         0.024219 -0.018353  0.045768 -0.047306
  Content Rating_Everyone 10+    -0.643504  0.075827 -0.175194 -0.140026
  Content Rating_Mature 17+       0.443271 -0.084177  0.030363  0.254358
  Content Rating_Teen             0.070723 -0.022635  0.023842 -0.025103
```

```
                                    PC-8      PC-9     PC-10         PC-11
Rating                          0.193063 -0.236113  0.152659  5.366359e-18
Reviews                        -0.088244  0.277273  0.297146 -2.164087e-16
Size                           -0.697119 -0.241472  0.319867 -2.297180e-16
Installs                        0.065962 -0.384029 -0.696045  4.624989e-17
Price                           0.418336 -0.300350  0.192039  8.492079e-17
Free                            0.376010 -0.303482  0.513262  3.296281e-17
lastupdate                      0.192146  0.679704 -0.031222  7.065386e-17
Content Rating_Adults only 18+ -0.049765 -0.013375  0.004493 -3.005226e-02
Content Rating_Everyone        -0.111556 -0.036824  0.003614 -6.811732e-01
Content Rating_Everyone 10+     0.285330  0.066363  0.002332 -3.383523e-01
Content Rating_Mature 17+      -0.114854 -0.085494  0.010148 -3.633023e-01
Content Rating_Teen             0.042195  0.063457 -0.013164 -5.372422e-01  ,
0     1.835330e-01
1     1.271822e-01
2     1.005432e-01
3     9.070442e-02
4     8.763576e-02
5     8.345853e-02
6     8.025333e-02
7     7.607196e-02
8     6.779460e-02
9     5.651330e-02
10    4.630968e-02
11    3.413439e-33
```

Interprétation du PCA:

Tout d'abord , il a été nécessaire de centrer réduire les données pour éviter les problèmes d'ordre de grandeur et ainsi faciliter la visualisation des résultats.

En suivant le principe des variables et individus supplémentaire et en sachant que le prix renseigne déjà si un produit en payant ou non, on peut déterminer que la variable "paid" n'apporte réellement aucune information. De plus grâce au tableau de variance ([0.18353303 0.12718218 0.1005432 0.09070442 0.08763576 0.08345853 0.08025333 0.07607196 0.0677946]) et à la courbe on peut évaluer le nombre d'axes nécessaires pour avoir les principales informations. Pour dépasser 80% il est nécessaire d'avoir 7 dimensions donc et 8 pour atteindre les 90%

Grâce au cercle de corrélation, on obtient que les variables "lastupdate", "Review", "Rating", "Installs" sont les plus corrélées au PC-0. Et par exemple size est plus corrélé (certe négativement) avec PC-1. En suivant le tableau juste au dessus, on voit que la variable "Content Rating_Everyone" est négativement corrélée à PC-0 et "Content Rating_Teen" y est corrélé aussi. Ici on peut remarquer que ces variables ont un "mode" en commun c' est à dire une notion qui les unit, et c'est l'accessiblité. De plus si une application est catégorisée pour les adolescents , elle ne sera pas pour tout le monde, d'où la corrélation négative.

Passons à un autre exemple: PC-2. Ainsi, on peut observer grâce au tableau que "Price" et "Free" sont fortement corrélés à PC-2 (le second négativement). La notion partagée par ces variables est le prix et il est simple de comprendre que si le prix est différent de 0 alors l'application ne sera pas gratuite d'où la corrélation négative.

Encore un autre exemple: PC-3. "Content Rating_Mature 17+", "Content Rating_Teen" sont toutes corrélées à PC-3 (la dernière de manière négative). Intuitivement, on peut observer qu'elles ont un mode commun qui est celui des application entre 14 et 20 ans. Ainsi si une application est à destination des adolescents, elle ne le sera pas pour des personnes plus "adultes"

Conclusion

Petit à petit en continuant ces analyses des composants principaux, nous pourrons regrouper toutes les variables en variables synthétiques et ainsi éviter la **redondance d'informations!** Cependant notre Dataset est assez difficile à interprété sachant le nombre de composant nécéssaires pour obtenir toute l'information. En effet, il n'y pas 2 ou 3 variables qui contiennent l'information mais 8 voire 9 il est donc difficile de regrouper plus de 3 variables sachant que notre Dataset final avait 12 variables