INNOVATIVE SMART SYSTEMS

Big Data
Analyse d'un dataset en R

Axel BAYLE, Pierre Prié 5 ISS

I. INTRODUCTION

Pour mettre en pratique les concepts et notions que nous avons vu en cours de Big Data il nous a été proposer d'analyser un dataset d'au moins 1000 entrées. Cette analyse a été réalisé via le langage de programmation R. Nous avons utilisé un dataset qui répertorie les applications sur le Playstore. Pour chaque application nous avons accès a plusieurs informations : le nom, la catégorie, la note, le nombre de revues, la taille, le nombre d'installations, le type (payant ou gratuit), le prix, tranche d'âge visée, le(s) genre(s), date de mise à jour, version actuelle et version d'android. Nous allons essayé grâce à ce dataset de déterminer comment faire une application populaire en termes d'installations et de notes.

II. INFLUENCE DE LA CATÉGORIE

Notre première approche a été de s'intéresser aux catégories d'applications les plus populaires. C'est pourquoi nous commençons par évaluer, pour chaque catégorie, le nombre d'applications sur le Playstore. En effet, cela semble un bon indicateur car il semblerait que plus il y a d'application d'un type plus c'est une catégorie populaire. On note avec ce premier graphique 1 que la majorité des applications du Playstore sont gratuite et que les catégories qui possèdent le plus d'application sont Family, Game et Tools. Nous allons maintenant voir quelles catégories sont les plus populaires en termes de note et d'installations.

Pour ce faire nous choisissons de porter notre attention sur le nombre d'installation d'application par catégorie, ainsi que sur la médiane, le premier et le troisième quartile en termes de note. Grâce à cela on a une idée plus claire de la répartition des notes qu'avec la moyenne.

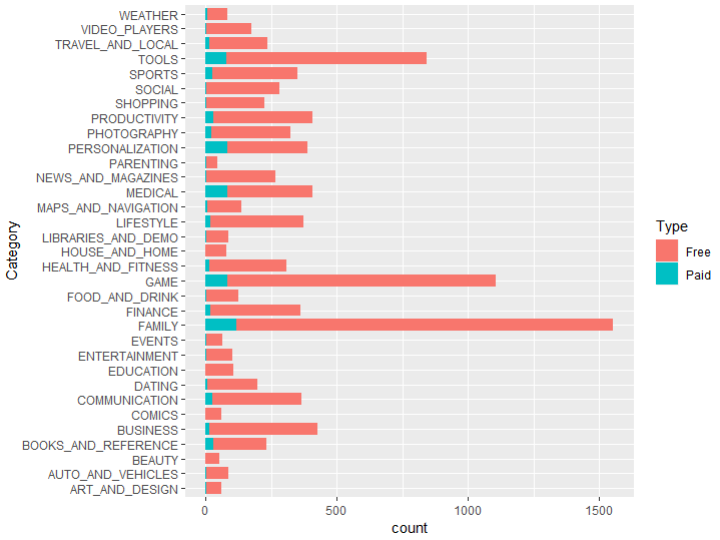

Fig. 1. Nombre d'application par catégorie

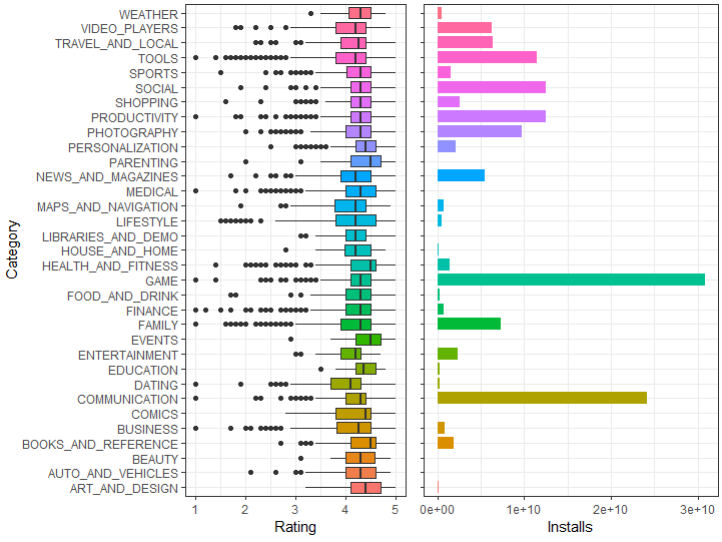

Fig. 2. Note et prix des catégories du playstore

Sur le graphique 2 nous pouvons voir que deux catégories se démarquent des autres quant à leurs nombres

d'installations : Game et Communication, ces catégorie ont donc été sélectionner d'office. Pour la suite de notre étude, nous avons également pré-sélectionner les catégories Photography, Productivity, Social et Tools. Finalement, bien que la catégorie Tools comptabilise plus d'application, la moyenne de ses notes ainsi que son troisième quartile sont plus faibles que celles des autres applications pré-sélectionner, cette catégorie n'a donc pas été retenue. On peut noter que contrairement à la catégorie Game qui comptabilise de nombreuses applications (1) mais également de nombreuse installations, les applications de la catégorie Family bien que nombreuses sont elles peut installées. Le nombre d'application d'une catégorie ne semble donc pas être un critère de succès.

En recoupant les informations que nous apportent les graphiques 1 & 2, nous avons choisi de limiter notre analyse aux catégories qui nous semblent les plus populaires. Ce sont les catégories Game, Communication, Photography, Productivity et Social. Car se sont celles qui sont le plus installées et celles qui ont une répartition des notes les plus intéressantes.

Maintenant que nous avons choisi les catégories les plus pertinentes pour nous, nous allons étudier les tranches d'âge visées par ces 5 catégories.

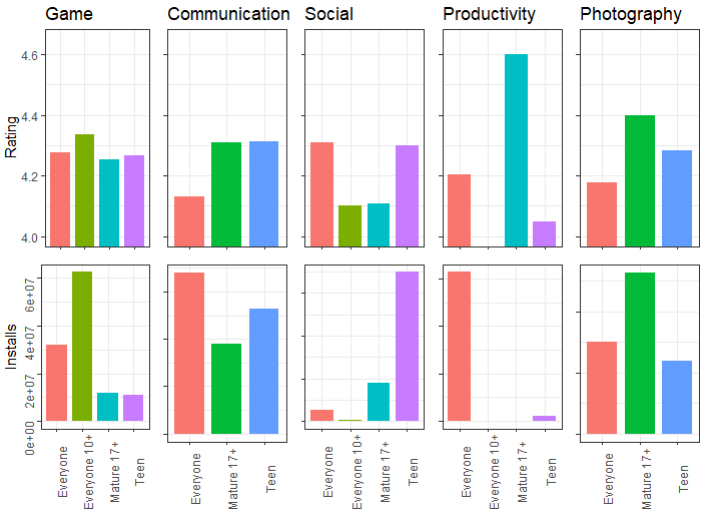

Fig. 3. Populations cibles des catégories

Sur le graphique ci dessus on peut observer plusieurs informations instructives. Tout d'abord, avec les graphiques du haut, on peut remarquer que la catégorie Game est notée de manière quasi-similaire selon que les applications soient destinées à telle ou telle population. Alors que la catégorie Productivity, quant à elle, est notée très différemment selon que l'application cible une population de plus de 17ans ou une population jeune, ou encore tout le monde.

Les catégories Communication, Social et Photographie sont entre ces deux comportement et on des notes assez similaires malgré que certaines différences existent.

Sur les graphiques du bas, on montre le nombre d'installation selon ces mêmes populations. Cette fois, on se rend compte que chaque catégorie a une population qui se démarque des autres en nombres de téléchargements. Les applications de la catégorie Game sont plus installées lorsqu'elles visent une population 10ans et plus, alors que c'est les applications visant tout le monde qui sont les plus installées dans les catégories Communication et Productivity. Dans la catégorie Social ce sont les applications visant les jeunes qui sont le plus téléchargées. Et finalement, les applications de la catégorie Photographie sont plus installées lorsqu'elles visent un public majeur.

On peut alors se demander si ces deux types de graphes sont liés. c'est à dire, si les applications d'une catégorie qui visent la population qui installe le plus vont avoir une meilleure note ou pas.

Il apparaît que la corrélation n'est pas systématique, en effet, même si c'est le cas pour la catégorie Photography et Game, on s'aperçoit que la catégorie Productivity déroge à la règle et ne suit pas du tout ce schéma. Les applications de Productivity sont les plus installées lorsqu'elles visent tout le monde mais elles ont tendance à être mieux notées lorsqu'elles visent un plublic majeur.

Ainsi alors que pour les catégories Game et Photography les critères de notes et de nombres d'installations ont le même effet sur la population à viser. Dans les autres catégories il faudra choisir si nous privilégions le critère de la note ou du nombre d'installation dans la caractérisation d'une application populaire, pour choisir ensuite la population à viser avec l'application.

III. Influence du prix

Nous avons maintenant voulu observer l'influence du prix des applications sur leurs notes moyennes et leurs nombres d'installations. Pour cela, pour chacune des 5 catégories étudiés, nous avons calculé le prix moyen des applications :

- sur l'ensemble d'entre elles pour la catégorie
- en les filtrants et en ne gardant que celles dont la note est supérieur au troisième quartile des notes de la catégorie
- en les filtrants et en ne gardant que celles dont le nombre d'installation est supérieur au troisième quartile des installations de la catégorie

Nous avions d'abord réalisé ce graphique sur l'ensemble des populations d'application, mais nous avons observé une trop forte influence des applications gratuite sur le résultat du fait de leurs surpopulation par rapport aux applications payantes, comme nous avons pu le voir avec le graphique 1. Nous avons donc réalisé un pré-filtrage des applications en ne gardant que les payantes pour réaliser notre étude

sur l'influence du prix. Nous avons obtenue le graphique suivant :

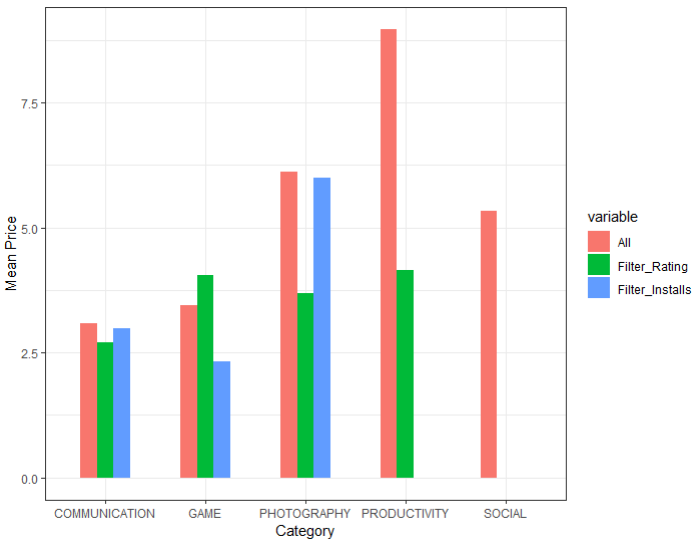

Fig. 4. Prix moyen des applications par catégorie : sans filtre, filtré par le 3ème quartile des notes, filtré par le 3ème quartile des installations

Pour la catégorie Communication, nous pouvons voir que le prix moyen des applications est de 3 dollars, ce prix reste quasi-identique chez les applications qui ont le plus de succès (en note et nombre d'installation). Pour la catégorie Game, le prix prix moyen des applications est de 3.5 dollars, ce prix est identique pour les applications les mieux notées, mais plus faible (2.1 dollars) pour les plus installées. Le résultat est identique pour la catégorie Photography mais cette fois-ci, c'est les applications les mieux notés qui ont un prix plus faible par rapport à la moyenne qui est de 6 dollars. Nous obtenons des résultats très différents pour les catégories Productivity et Social. Premièrement, pour Productivity, le prix moyen des applications est beaucoup plus important que pour les autres catégorie et est de 9 dollars. On voit que le prix moyen des applications les mieux notées est moitié plus faible (4.5 dollars), pour les plus installées le prix moyen est là de 0 euros. Ce résultat se retrouve avec la catégorie social mais cette fois-ci, toutes les applications à succés sont gratuite.

Pour résumer ce graphique, le prix n'a que peut d'influence pour les catégories Communication, Game et Photographie sur les succès des applications (variation de 1 à 2 dollars) alors que pour les catégories Productivity et Social, le prix influence le succès des applications car ce sont les applications gratuites qui fonctionnent.

On pourra également noter que de plus en plus d'applications deviennent gratuite, notamment des jeux, mais proposent alors des micros-transactions ingame. Le dataset ne prenant pas en compte ce genre d'informations, l'étude du prix des applications en fonction de leurs succés pour la catégorie Game peut être faussé.

IV. Influence de la taille

La dernière étape est de chercher si il existe un lien entre la taille d'une application et sa popularité. Pour ce faire nous avons choisi de différencier selon les deux critères de popularité.

A. sur le critère du nombre d'installations

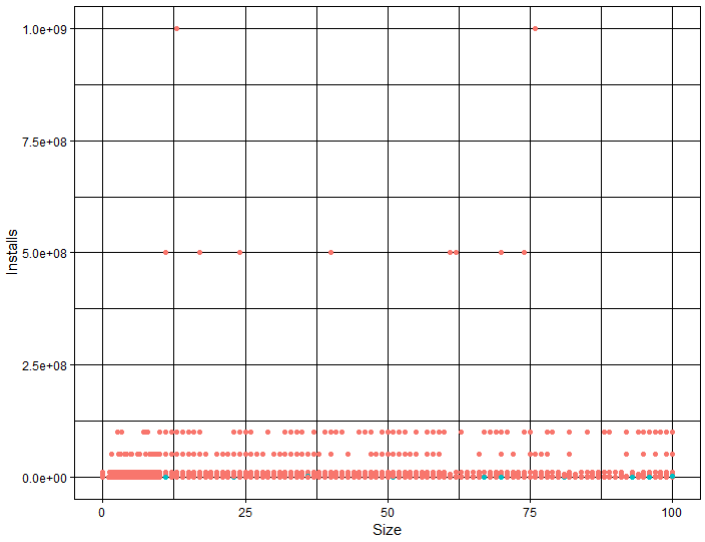

Fig. 5. Nombre d'installations en fonction de la taille

En affichant les applications sur un graphe, le constat est qu'il n'y a pas de lien entre installations et taille. En effet on trouve quasiment le même nombre d'installations quelle que soit la taille de cette application. On remarque aussi que la plupart des applications comptent peu d'installations et que même les plus installées ont des taille variables.

B. sur le critère de la note

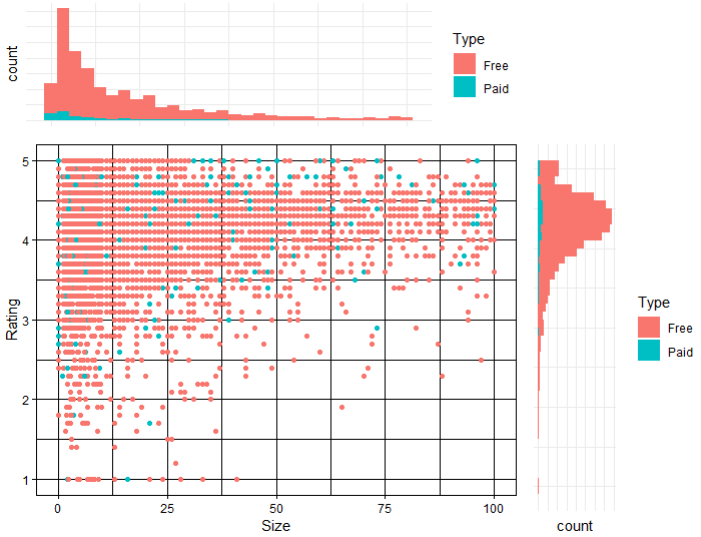

Fig. 6. Note en fonction de la taille

Contrairement au graphique 5, le graphique 6 montre un réel changement selon la taille.

D'une part, on voit que les application sont majoritairement peu volumineuse (moins de 25 Mo), d'autre part, les applications sont en majorité bien notées.

Mais en examinant plus précisément la partie centrale du graphe, on peut clairement affirmer que la grande majorité des applications les mieux notées (au dessus de 4.7) ont une petite taille.

De plus, il n'y a pas ou très peu d'applications volumineuses avec des notes en dessous de 3.5/5 ainsi qu'au dessus de 4.7.

Il apparaîtrait donc, qu'une application volumineuse serait une valeur sûre pour avoir une bonne note, mais qu'il est rare d'avoir une très bonne note. Et que si nous voulons une application très bien noté nous avons plus de chance avec une application légère mais nous prenons alors le risque d'avoir une application mal notée.

Ainsi il faut décider si on préfère le défi et le risque mais avoir plus de chance d'avoir une note remarquable, ou si la sécurité d'une note correcte est plus satisfaisante.

V. Bonus

Au cours de notre étude du data set nous avons remarqué un détail pour le moins perturbant.

En effet lors de l'affichage de différent graphes pour découvrir le data set, nous avons affiché le nombre d'application en fonction du prix de l'application (toute catégories confondu). Nous nous attendions à voir une relation évidente, beaucoup d'application gratuite ou peu chères et très peu coûtant cher.

Cela c'est avérer en parti vrai mais pas complètement. Comme le montre le graphe 7 on remarque un pic non négligeable d'applications autour des 400 $.

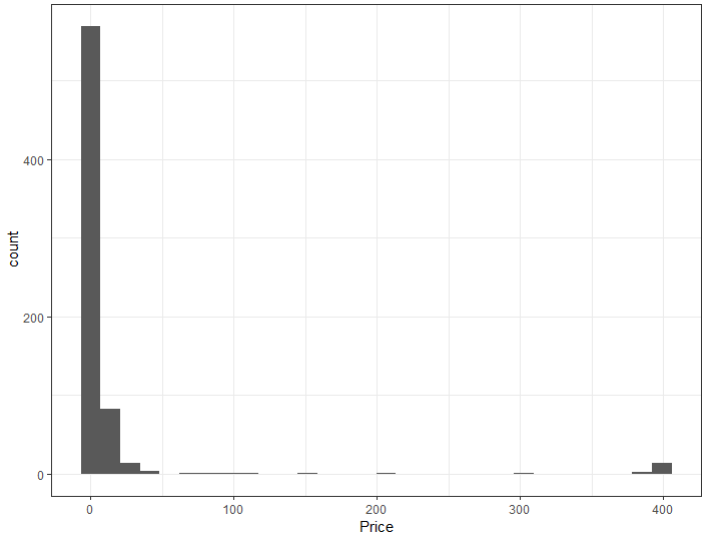

Fig. 7. Nombre d'application selon leur prix

Curieux, nous avons chercher quelles étaient ces applications. On peut voir le résultat sur la figure 8. On voit qu'il existe tout un panel d'application du type I'm Rich qui sont

	App	Price	Installs
1	most expensive app (H)	399.99	1e+02
2	ðŸ'Ž I'm rich	399.99	1e+04
3	I'm Rich - Trump Edition	400.00	1e+04
4	I am rich	399.99	1e+05
5	I am Rich Plus	399.99	1e+04
6	I Am Rich Premium	399.99	5e+04
7	I am Rich!	399.99	1e+03
8	I am rich(premium)	399.99	5e+03
9	I Am Rich Pro	399.99	5e+03
10	I am rich (Most expensive app)	399.99	1e+03
11	I am Rich	399.99	5e+03
12	I AM RICH PRO PLUS	399.99	1e+03
13	I'm Rich/Eu sou Rico/Ø£Ù†Ø§ Ø°Ù†ÙŠ/æ¯'å¾ˆæœ‰é€¢	399.99	0e+00

Fig. 8. Tableau des applications les plus chères

relativement beaucoup installées au vue de leur prix. En cherchant à part nous avons découvert que ces applications n'ont pas vraiment de contenu, et qu'elles ont juste pour but de montrer que les personnes les ayant sur les téléphones ont les moyens de se payer ce type d'application.

VI. CONCLUSION

Grâce à cette étude nous avons trouvé que pour faire une application populaires tant en termes de notes que d'installation nous avons deux possibilités.

La première est de faire une application de type Game visant un public de 10 ans et +, avec un prix de 3.5$.

La deuxième étant de développer une application de type Photography visant un public majeur avec un prix d'environ 6$.

Dans tout les cas il faudra réfléchir à la taille de l'application.

Enfin, n'oublions pas que ce qui rend une application populaire est quand même le contenu qu'elle propose et la publicité qu'elle met en place pour se vendre.