Insa de toulouse

Analyse exploratoire et visualisation de données

Traitement de masses de données

Réalisé par:

Youness LAGRINI

Ayman BENMADA

Encadré par:

M.Gilles Tredan

INSA | INSTITUT NATIONAL DES SCIENCES APPLIQUÉES TOULOUSE

Tables des matières :

Introduction :

Android continue de grossir, lors de l'événement Nexus en 2015, Pichai a annoncé que la plate-forme Google Play Store comptait désormais plus d'un milliard d'utilisateurs actifs. En 2016, la société a annoncé durant l'événement Google I/O que plus de 600 nouveaux téléphones Android ont été lancés au cours de la dernière année, plus de 65 milliards d'applications ont été installées via son magasin Google Play et que son système d'exploitation mobile avait atteint 1,4 milliard d'utilisateurs. Aujourd'hui google annonce qu'avec la possibilité de publier rapidement sur plus de 2 milliards d'appareils Android actifs, Google Play aide les développeurs à développer un public mondial pour leurs applications et leurs jeux et à générer des revenus.
Google play store propose une diversité des applications qui peuvent servir à une gamme quasi illimitée de fonctions, allant d'outils simples à des assistants numériques avancés.
Notre analyse vise à fournir une compréhension de ces applications et leurs catégories, de voir en détails les facteurs d'influences sur une application, étudier les applications les plus installées et d'essayer de savoir pourquoi et comment certaines applications réussissent et d'autres pas et finalement ce qu'il faut pour qu'une application soit considérée comme ayant du succès.
On a utilisé le jeu de données Google Play Store Apps contenant les noms des applications et leurs informations associées.

I- Exploration et préparation des données:

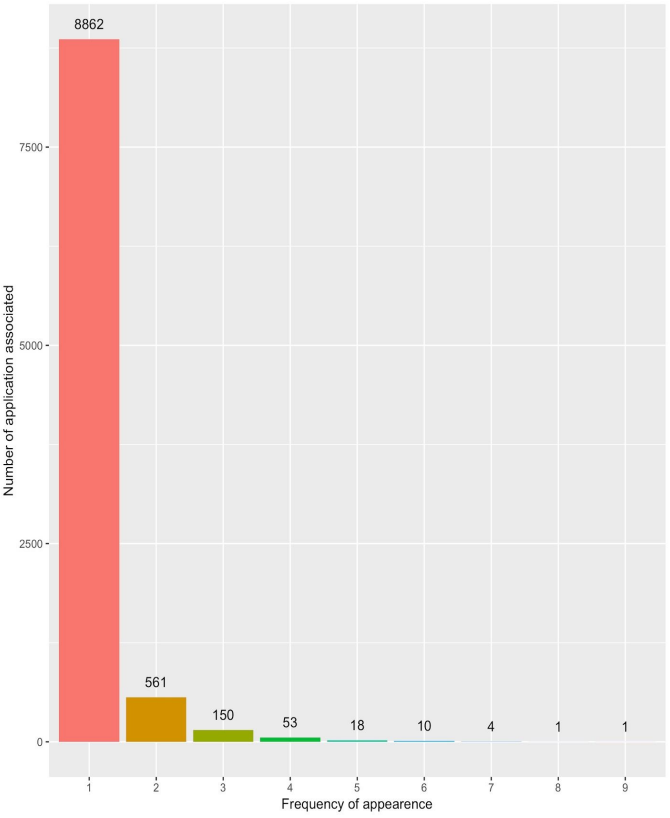
Commençons par comprendre notre jeu de données. Pour ce faire, nous commençons par une analyse exploratoire des données pour chacune des variables. Cette étape va nous permettre de répondre aux premières questions concernant l'étude des applications et des facteurs d'influences, elle est utile aussi pour mettre en évidence les valeurs manquantes et aberrantes.

1- App Column :

Le nombre d'observations dans le jeu de données égale à 10841 mais la colonne des noms des applications comporte que 9660 niveaux différents, c'est à dire qu'il y a des applications qui porte le même nom.

Le graphe suivant nous pousse à poser les questions suivantes:

+ Pourquoi a-t-on des noms des applications qui sont répétées plus d'une fois ?
+ Est ce qu'il s'agit de la même application ou de différentes applications ayant le même nom ?



Après avoir inspecté des applications portant les mêmes noms, on a constaté que ces applications avaient presque les mêmes caractéristiques, pour la majorité d'eux il y avait une différence au niveau du nombre des commentaires alors que d'autres avaient des différences au niveau de la taille, la catégorie et la date de dernière mise à jour.

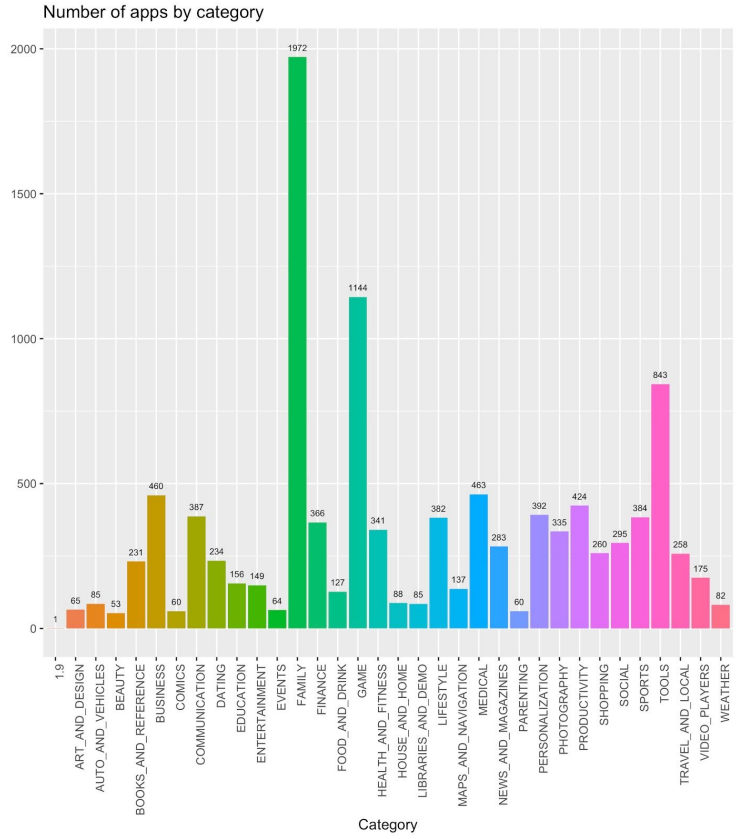
Cette constatation nous a poussé à approfondir nos recherches pour voir s'il est possible d'avoir plusieurs applications ayant le même nom d'affichage, effectivement cela est possible à condition d'avoir des noms de package différents.

On considère par la suite que ces applications dupliquées sont dues à une erreur de collection de données pour deux raisons :

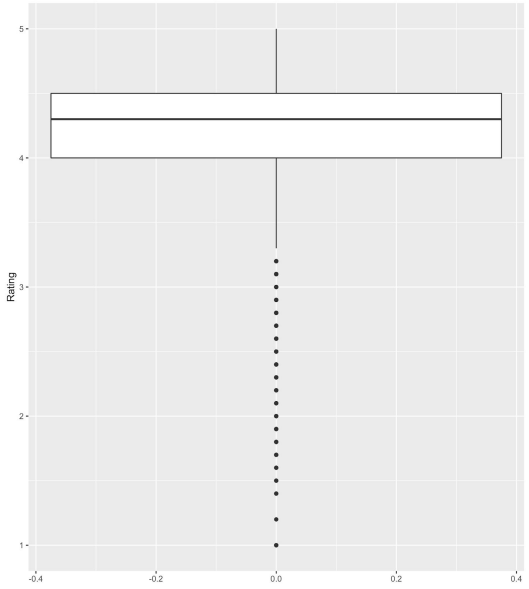
- Les applications dupliquées sont des applications reconnues ayant un grand nombre d'installation (le même nombre) et le même taux d'évaluation; comme Google Drive, Google Photos et Duolingo. Ces noms d'applications sont des marques déposées 'Trademark', donc lancer des app avec ces mêmes noms peut mener à des problèmes juridiques s'il y a eu une décision de défendre cette marque.
- Le jeu de données est publié sur Kaggle en 5 versions distinctes, avec une durée d'un mois entre la première et la dernière version. Il pourrait donc évoluer durant cette période.

2- Category column :

- Il paraît normal que la majorité des applications font parties des catégories Family et Game. Ce qui n'est pas la cas pour d'autres types d'applications.
- Ce qui ne semble pas logique ici, est le fait d'avoir une catégorie nommée 1.9 avec une seule application. En examinant cette application, on remarque qu'il s'agit d'une erreur due à un décalage à gauche.



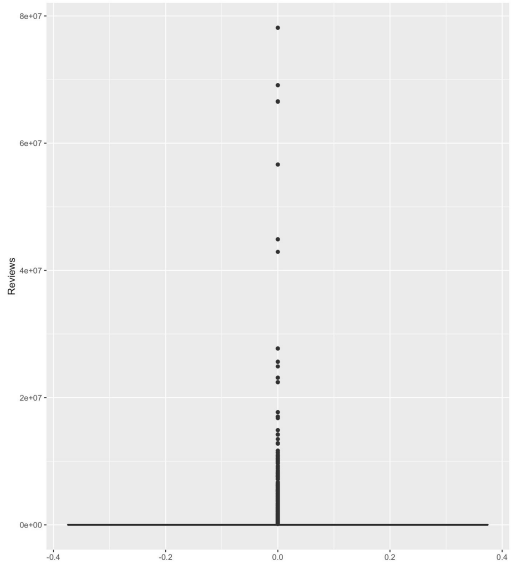
4-Rating:

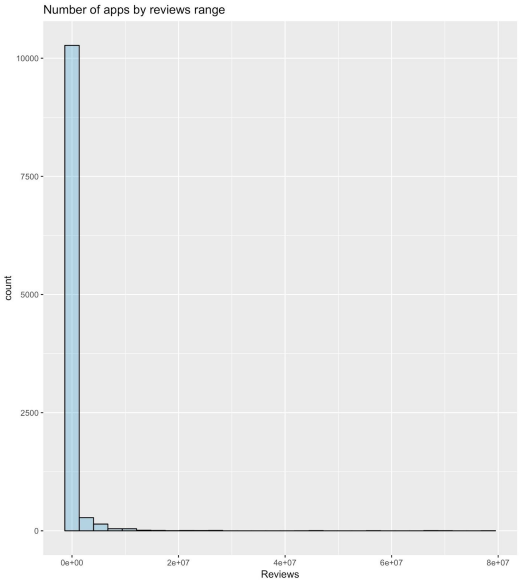


La majorité des applications ont un taux d'évaluation entre 4 et 4.5 et 1474 applications ont un taux d'évaluations manquant.
Il y'a des applications avec un taux d'évaluation égale à 5, nous étudierons ces applications plus tard, parce qu'avoir un taux égale à 5 signifie que l'application satisfait tout le monde ce qui est quasi-impossible pour les applications ayant un grand nombre d'installation.
On a pensé à créer un modèle de prédiction pour prédire les taux d'évaluation manquants à partir d'un deuxième jeux de données des commentaires des clients, mais ce dernier ne contient qu'une application, 'Blood Pressure', parmis 1474 applications ayant le taux d'évaluation non défini. Il est inutile d'entraîner un modèle pour avoir le rating d'une seule application parmis 1474.

4- Reviews Column :

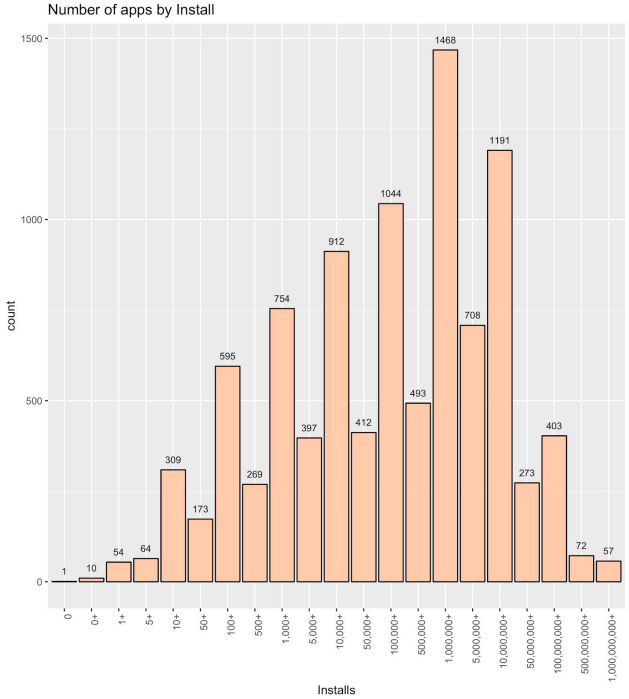
Après la conversion de ce champs en valeurs numériques, on a visualisé les deux graphiques en-dessous :.





Nous avons un histogramme dévié vers la droite (right skewed) avec 75% des applications ont un nombre de commentaires inférieur à 54776.
La boîte à moustaches et la tendance centrale nous montrent qu'il y a des valeurs aberrantes à l'ordre de 10e+07, ces valeurs peuvent être dues à des erreurs de saisie de données en ajoutant des numéros supplémentaires, ou à des valeurs aberrantes naturelles, c'est-à-dire qu'il existe vraiment des applications avec ce nombre de commentaires.
Pour vérifier ceci, on va voir par la suite s'il y a des applications avec un nombre de commentaires qui dépasse largement le nombre d'installations étant donné qu'un utilisateur peut laisser plusieurs commentaires sur la même application.

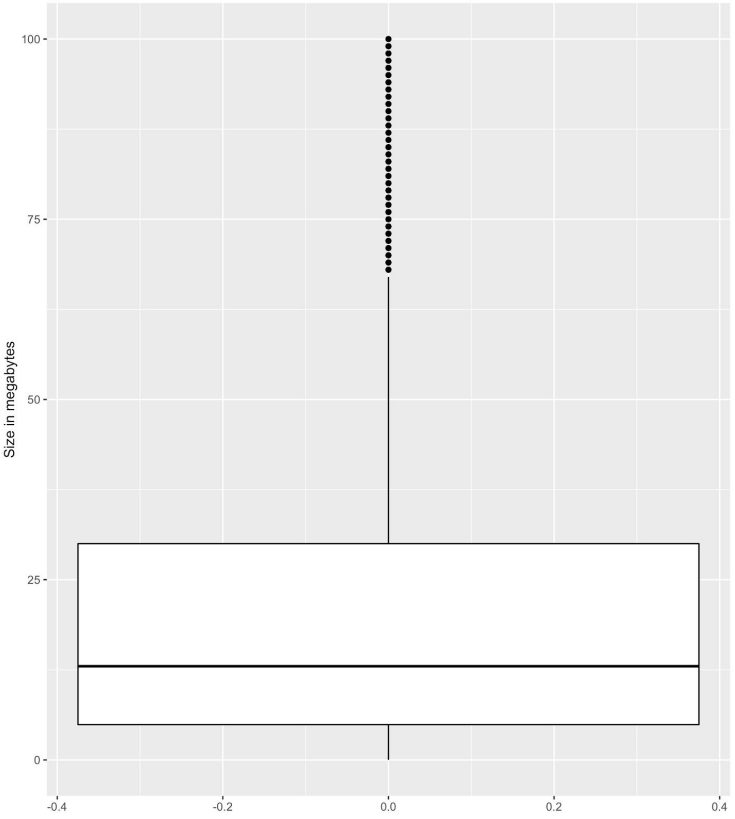
5- Installs Column :



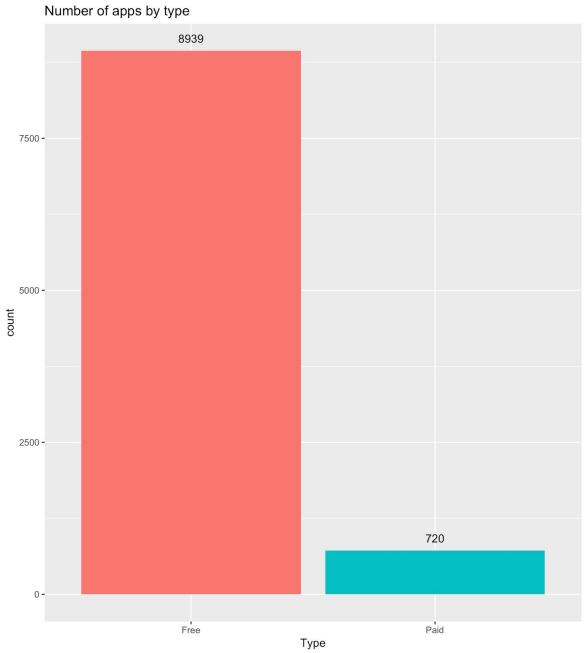
Il existe 57 applications dans notre jeu de données avec plus d'un milliard d'installations.
Après la conversion de Installs en numérique, la tendance centrale nous montre que l'histogramme de Installs est dévié à droite, vue que la valeur médiane est inférieure à la moyenne, avec presque 70% d'applications ayant un nombre d'installation entre 1000 et 5000000.
Il n'y a que 11 applications ayant le nombre d'installations inférieur au nombre de commentaires, sachant que le nombre d'installation est suivi de l'indice '+' et que le nombre de commentaires ne dépasse pas le nombre d'installations suivant, alors aucune application n'a le nombre d'installations inférieur au nombre de commentaires. Ainsi ces valeurs aberrantes sont des valeurs aberrantes naturelles.

6- Size Column :

On a 462 niveaux du facteur Size. Pour avoir une bonne visualisation, on va convertir cette colonne en valeurs numériques représentant la taille en mégaoctets.
1695 des applications ont une taille qui varie avec l'appareil utilisé, 75% des autres applications ont une taille inférieur à 30 Mo alors que 25% des applications ont une taille qui converge vers 100 Mo.
La question qui se pose est pourquoi existe-t-il des applications avec une taille qui varie selon l'appareil utilisé ? Parce que Google Play vous permet de publier différents fichiers APK pour votre application. Chacun ciblant une configuration de périphérique différente. Ainsi, chaque APK est une version indépendante de votre application, mais ils partagent la même liste d'applications sur Google Play. Ils doivent partager le même nom de package et être signés avec la même clé de publication.



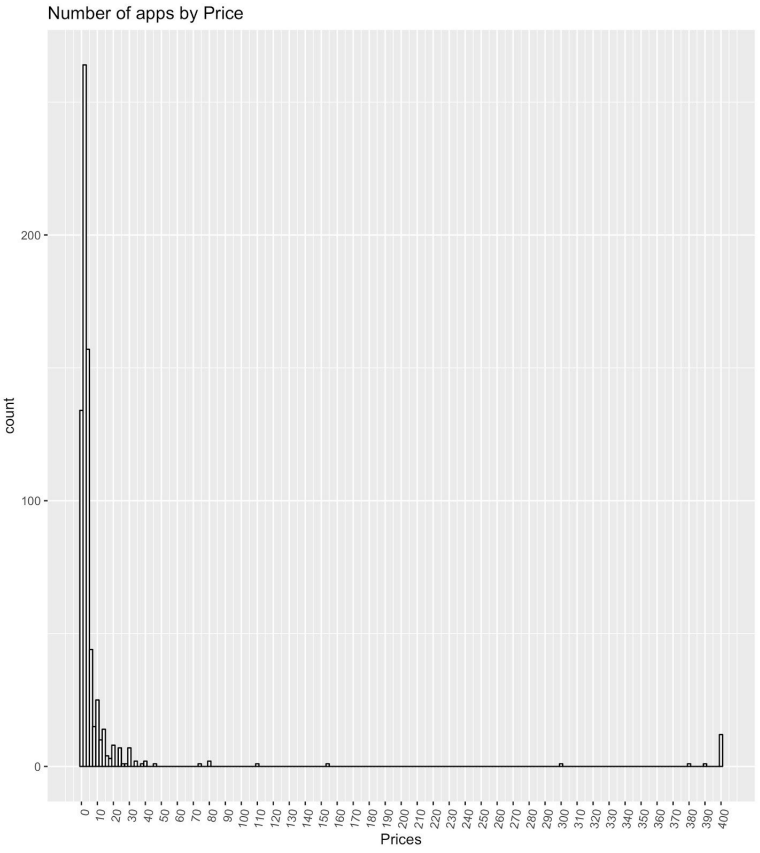
7- Type Column :



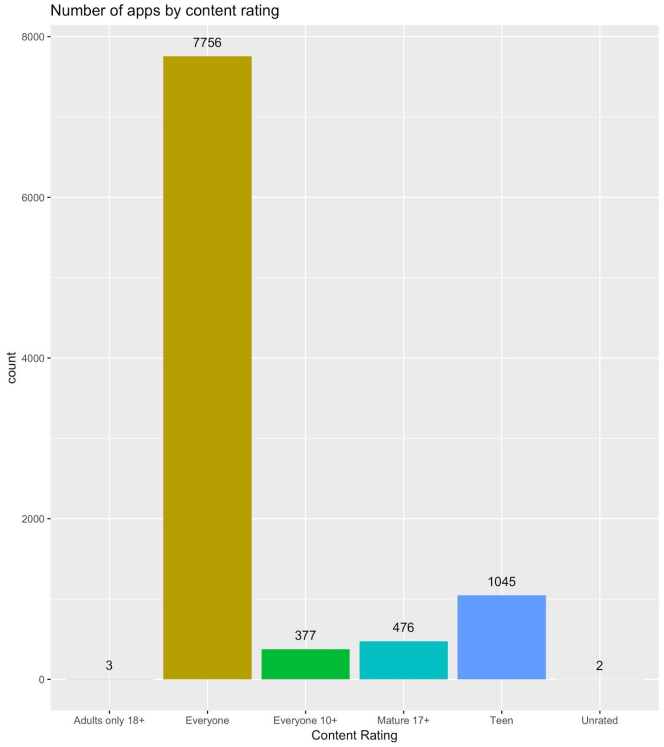
Il n'y a qu'une application avec une valeur manquante, après vérification de cette ligne on remarque que son prix égale à 0 ce qui signifie que son type est 'Free'.
Les applications gratuites sont très nombreuses par rapport aux applications payantes, elles représentent 92,54 % de notre jeu de données.

8- Price Column :

Nous savons déjà que les applications payantes ne représentent que 7,46 % de notre jeu de données. Par conséquent, on ne va s'intéresser qu'aux prix des applications payantes.
Pour ce faire, on va éliminer le symbole $ et convertir les prix en valeurs numériques.
88.05 % des applications ont des prix inférieurs ou égales à 10 alors que 1.66% ont des prix de l'ordre de 400$.
À quoi servent ces application avec un prix trop élevé et à quelle catégorie appartiennent-t-ils ?



9- Content Rating Column :



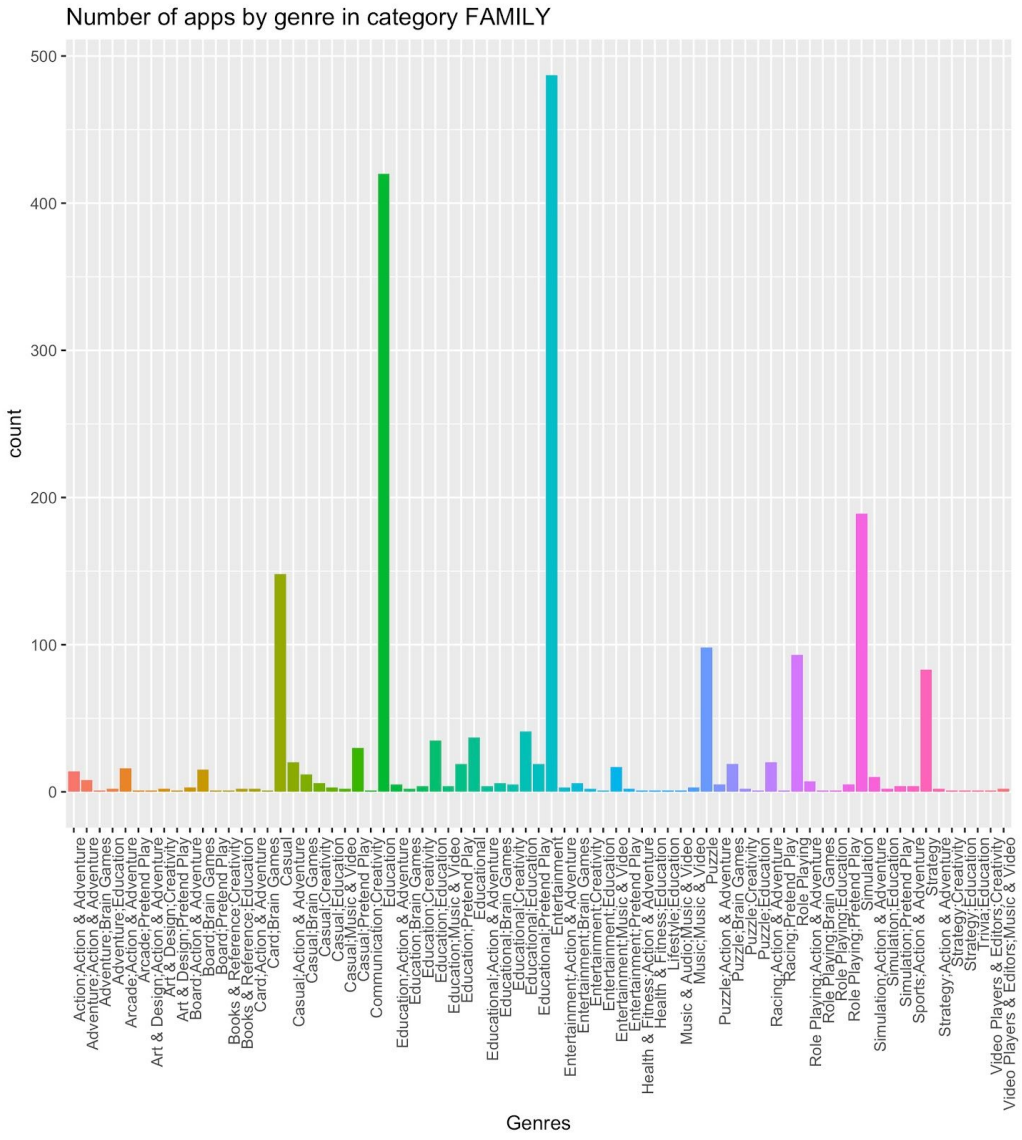
80.3 % des applications sont dédiés à tout le monde, mais pour mieux comprendre cette distribution on doit comprendre à quoi sert ce classement de contenu et que signifie-t-il.
D'après Google ce classement de contenu sert à :

1. Informer les consommateurs, en particulier les parents, du contenu potentiellement répréhensible dans une application.
2. Bloquer ou filtrer votre contenu sur certains territoires ou à des utilisateurs spécifiques.
3. Évaluer l'éligibilité de votre application aux programmes spéciaux pour développeurs.

On peut donc conclure que les développeurs ont tendance à développer des applications généralement adaptées à tous les âges, peuvent contenir un minimum de dessins animés, une violence fantasmatique ou légère et / ou un usage peu fréquent d'un langage doux pour attirer le maximum des consommateurs et que leurs applications ne soient pas bloqués.

10- Genres Column :

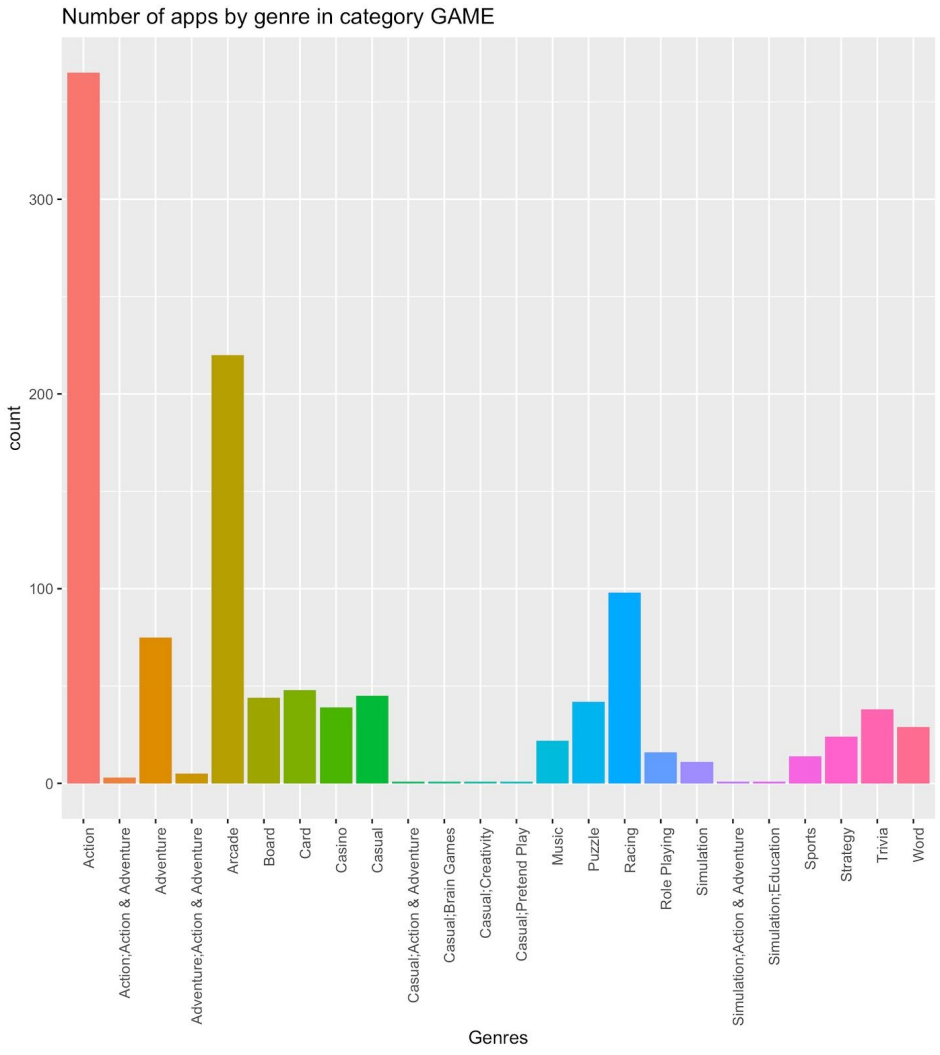
Cette colonne sert à identifier les sous-catégories des applications appartenant à Family ou Game,



Les genres les plus fréquents sont :

- Entertainment : Streaming vidéo, films, télévision, divertissement interactif.
- Education: Préparation aux examens, aide-mémoire, vocabulaire, jeux éducatifs, apprentissage des langues.

On a regardé dans Google Play Store les applications éducatives et celles de divertissement dans la catégorie famille, la majorité de ces applications étaient des applications d'éducation sous forme de jeux pour enfants ayant moins de 12 ans. Ceci est s'explique par le fait que les enfants d'aujourd'hui passent beaucoup de temps devant leurs tablettes ou un smart device avec la tolérance de leurs parents. Selon un article publié par CNN en octobre 2017 soulignant une étude faite par Common Sense Media aux États Unis, les enfants de 8 ans et moins passent en moyenne 2 heures et 19 minutes par jour avec un support d'écran et près de la moitié, 49%, regardent la télévision ou jouent à des jeux vidéo une heure avant le coucher. Selon le rapport, dans l'ensemble, 67% des parents dont les enfants utilisent les médias à l'écran déclarent que cela facilite l'apprentissage de leurs enfants.
Du coup, aujourd'hui le public ciblé par la majorité des développeurs sont les enfants moins de 12 ans. Ceci fait augmenter le nombre d'applications dont le classement du contenu est tout le monde (80.3 % dans notre jeu de données).



Les genres les plus fréquents sont :

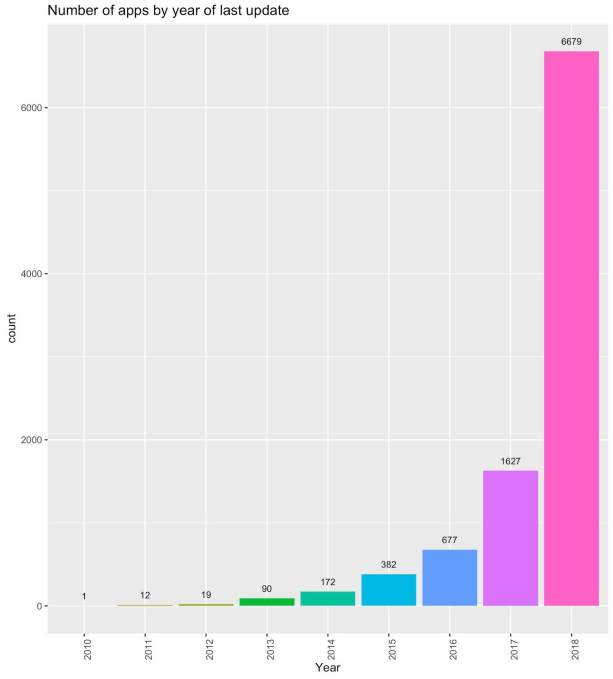
- Action : tels que les jeux de combat, les jeux de battement, les jeux de tir et les jeux de plate-forme qui sont largement considérés comme les jeux d'action les plus importants, bien que l'arène de combat en ligne multijoueur et certains jeux de stratégie en temps réel soient également considérés comme des jeux d'action.
- Arcade : tels que des flippers, des jeux électro-mécaniques, des jeux d'échange ou des marchandiseurs.

Il y'a quelques années, lorsqu'on cherche un jeu sur Google play store on trouvait une liste limité. Aujourd'hui, si on cherche un nom de jeu quelconque on trouvera une longue liste de jeux associés. On a remarqué que la majorité des applications jeux Action et Arcade sur google play store ont le même but d'application avec une différence minimale. Les jeux d'action et d'arcade étaient les plus installées, selon les gamers ces jeux sont tellement amusants et vous font sentir vivant en quelque sorte, du coup les développeurs ont pensé à développer les mêmes principes des jeux avec des retouches au niveau de l'interface et en ajoutant quelques fonctionnalités. Cette tendance a réussi au début parce que les utilisateurs cherchaient toujours du changement ainsi que la concurrence était limitée, mais aujourd'hui cette tendance ne marche plus bien, le nombre d'applications s'est élevé mais peu qui sont jugées top on verra ceci plus tard, et c'est pour cela que le règlement du programme pour les développeurs de google play store devient de plus en plus strict.

10- Last Updated Column :

Cette colonne contient 1378 niveaux, parce qu'elle est composée de la date sous forme de jour-mois-année, ce qui nous intéresse vraiment ce n'est pas précisément le jour de la dernière mise à jour, le mois et l'année sont suffisants et significatifs. Pour ce faire, on a pris que l'année de la dernière mise à jour de chaque application et les visualiser.

Le graphe situé à droite est un graphique en barre dévié à gauche, on remarque que 69.15% des application ont eu leurs dernières mise à jour en 2018. On a visualisé aussi les mois de la dernière mise à jours sont à 27.71% en Juillet 2018 et 12.3 % en Août 2018.



12- Android Ver Column :

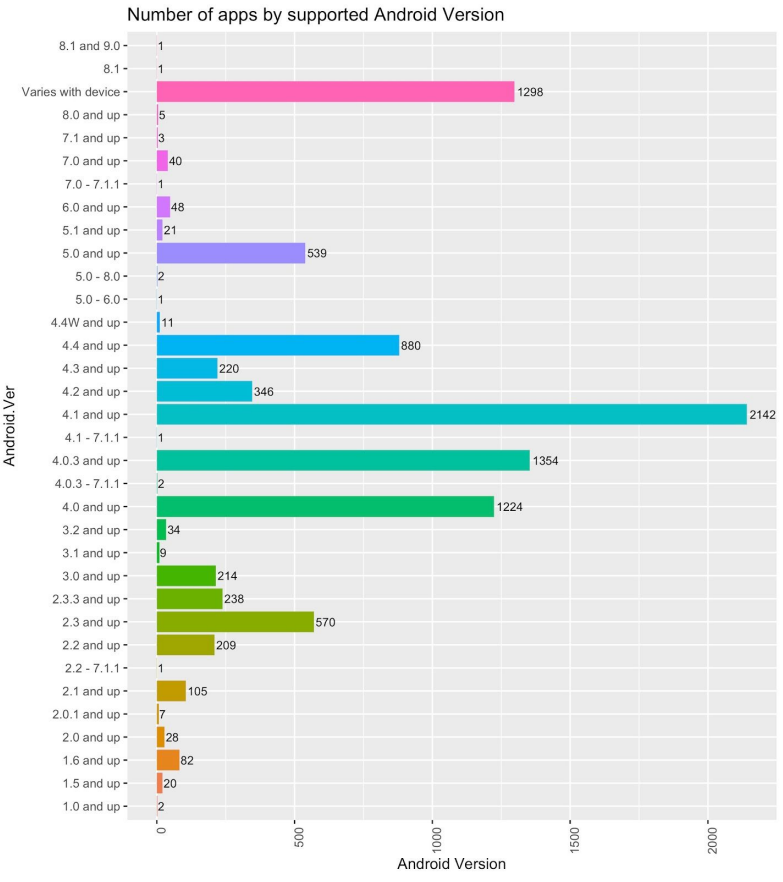
Il y'a deux applications avec ce champs manquant et 1262 applications avec la version varie selon l'appareil. On a cherché les deux applications sur google play store par leurs noms complets, et après la comparaison des caractéristiques on a mis en place leurs version android supportée.

La majorité des applications supportent les versions 4.x.x and up, et cela parce qu'il existe toujours un nombre important de périphériques exécutant ces versions de la plate-forme Android, un pourcentage de 10.9% selon le [tableau de bord de distribution de google](), du coup les développeurs ciblent des versions de la plate-forme android des périphériques à partir de 4.x.x ou plus pour gagner un public potentiel énorme pour leurs applications.

En ce qui concerne les applications ayant la version varie avec le device, on remarque que 91.67% de ces applications ont aussi version courante et la taille varie avec le périphérique utilisé.

Pourquoi existe-t-il un nombre important d'applications avec cette caractéristique de variance ? Et cela concerne quoi ?

Cet effet s'explique par la [compatibilité]() d'une application, on remarque que c'est lié principalement aux applications les plus installées, on expliquera ceci



en détail dans la section prochaine lors de l'analyse multi-variables de notre jeu de données.

II- Nettoyage des données et ingénierie des fonctionnalités:

Avant de commencer l'analyse multivariée, il faut d'abord corriger les erreurs et les valeurs aberrantes qu'on a trouvé dans la première section et faire l'extraction d'informations supplémentaires à partir des données existantes.

1- Nettoyage de données:

Pour cette étape on a effectué les changements suivants:

- Convertir Reviews en valeurs numériques.
- Éliminer les applications ayant le même nom et laisser une avec le nombre de commentaires le plus élevé.
- Elimination de la ligne qui contient NA en catégorie et en Genre, c'est la ligne où il y avait un shift à gauche.
- Suppression des virgules et des symboles '+' de Installs.
- Convertir Size en valeurs numériques (en mégaoctets).
- Correction de la ligne qui contient NaN sur Type.
- Convertir Price en valeurs numériques.
- Suppression des deux applications ayant Content.Rating égale à 'Unrated'.
- Correction des applications dont le champ Android.Ver n'est pas renseigné.

2- Ingénierie des fonctionnalités:

Pour nous faciliter l'étude de notre jeu de données et pour rendre les graphiques plus pertinents, on a ajouté les colonnes suivantes:

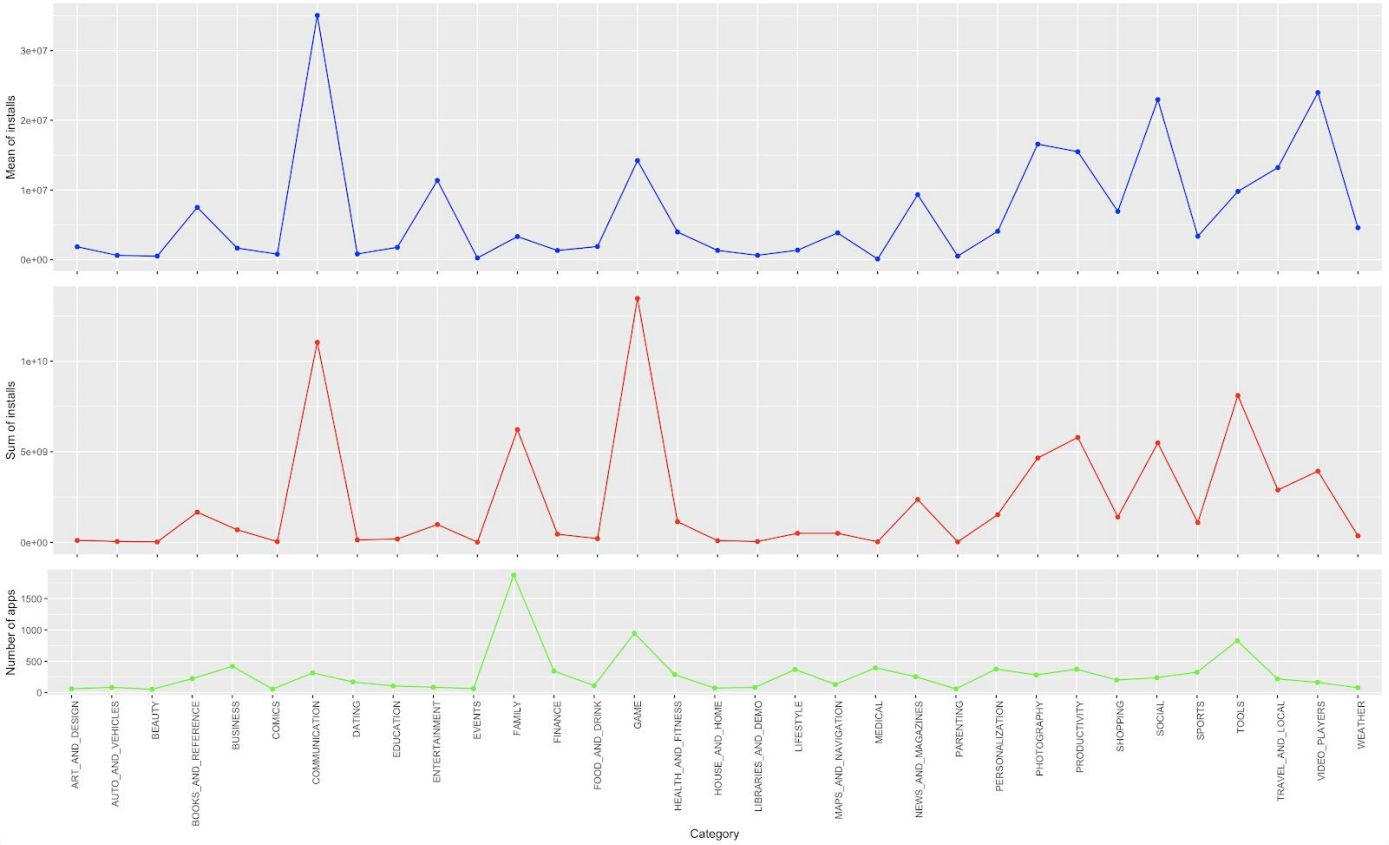
- Last.Updated.Year : factor avec les niveaux "2010", "2011", "2012", " 2013", " 2014", " 2015", " 2016", " 2017", " 2018" .
- Size.level : factor avec les niveaux suivantes :
    - NA : pour les applications ayant la taille qui varie avec le périphérique utilisé.
    - Lowest : pour les applications ayant la taille inférieur ou égale à 1 Mo.
    - Low : pour les applications ayant la taille entre 1 Mo et 10 Mo.
    - Medium : pour les application ayant la taille entre 10 Mo et 50 Mo.
    - High : pour les application ayant la taille supérieur à 50 Mo.
- Price.level : factor avec les niveaux suivants : "0","0-1","1-5","5-10","10-25","25-50","74-120","140-160","195-215","290-310","370-390","390-400".
- Le choix des niveaux et effectué selon la visualisation des prix des applications.

III- Analyse multivariée

Dans la section précédente, on a pu tirer plusieurs informations concernant notre jeu de données. Durant cette section, nous allons répondre aux questions suivantes:

- Existe-t-il une corrélation entre le nombre d'installations et le nombre d'applications dans une catégorie ?
- Quels sont les caractéristiques d'influences sur la découverte d'une application ?
- Quelles applications ont le plus grand nombre d'installations et quels sont les points communs entre elles ?
- Google respecte-t-il vraiment les conseils donnés aux développeurs pour que leurs applications et leurs listes de magasins soient découvertes par les utilisateurs sur Google Play ?

1- Facteurs d'influences sur les applications :



D'après les graphiques ci-dessus, il est clair qu'il n'y a pas une corrélation entre le nombre d'installations et le nombre d'applications dans une catégorie.

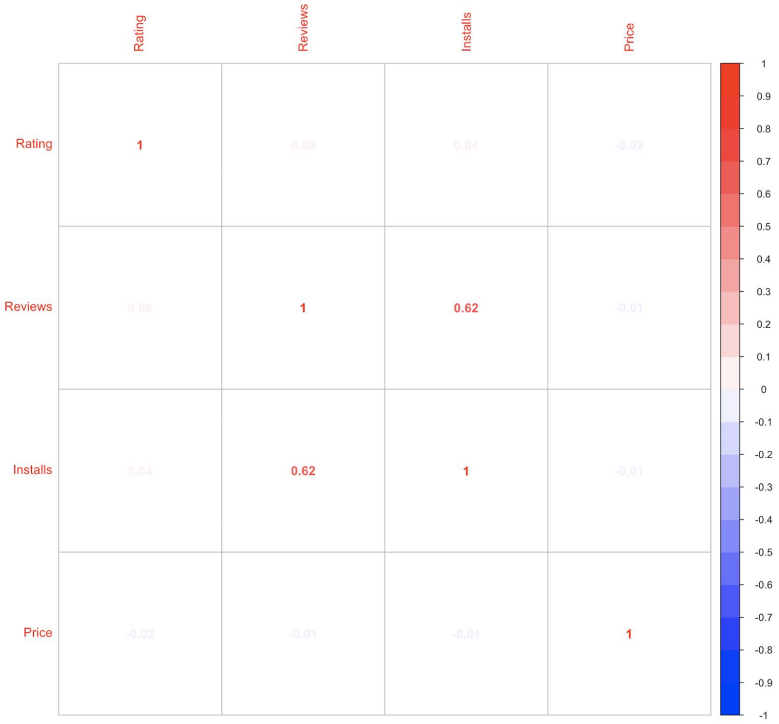
Malgré que la majorité des applications font partie de la catégorie Family, le nombre d'installations des applications appartenant à cette catégorie reste limité face aux nombres d'installations des applications appartenant aux autres catégories à savoir Communication, Game et Tools.

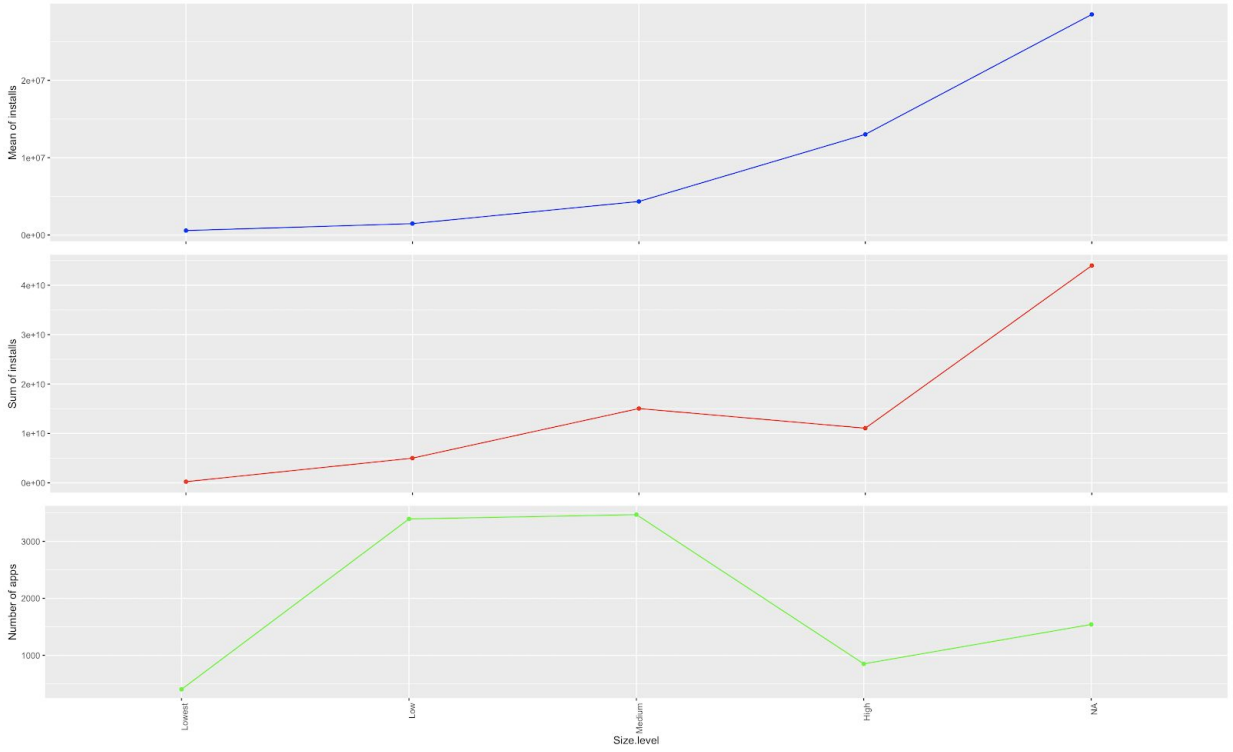
Si on analyse le graphique de la moyenne des installations par catégorie, Video_Players, Communication et Social font des pics remarquables, ce résultat n'est pas choquant parce que ces applications deviennent

aujourd'hui presque omniprésentes avec peu des fournisseurs face un grand nombre de consommateurs. La question qui se pose ici est pourquoi ces applications sont dénombrables avec un public assoiffé ?
Premièrement, le développement de ces applications de médias n'est pas une tâche facile, c'est un processus difficile à suivre de prototypage,conception et développement jusqu'à publication et marketing.
Deuxièmement, la confiance des utilisateurs dans les médias sociaux avait décliné face à la montée des préoccupations concernant le manque de réglementation des plateformes. Les utilisateurs aujourd'hui estiment que les entreprises de médias sociaux ne font pas assez pour empêcher les comportements illégaux ou contraires à l'éthique de se produire sur leurs plateformes.

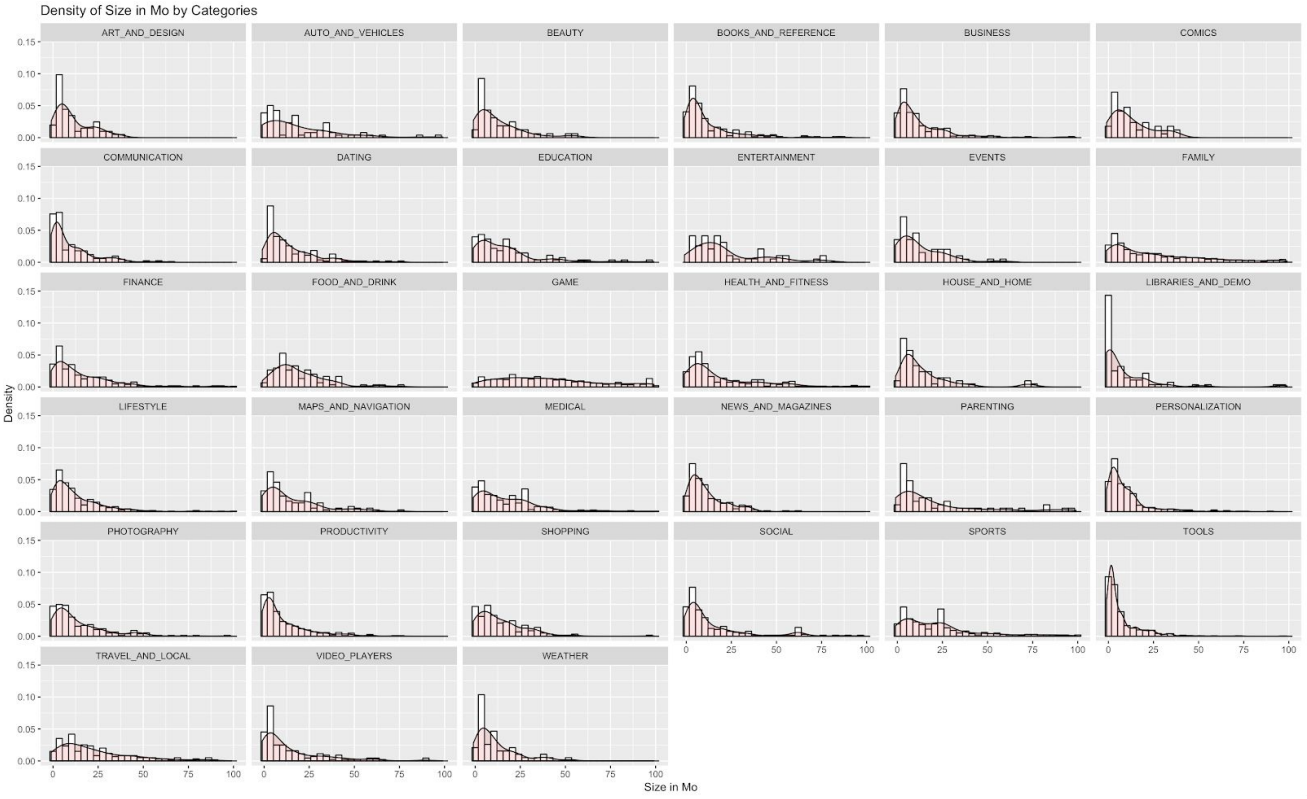
Une remarque d'après ces graphiques qui ne peuvent pas passer inaperçus, certaines catégories n'apparaissent pas sur le graphique du nombre d'applications, mais elles ont pu se démarquer dans le graphique de la moyenne d'installation, comme New and magazines, books and reference et photography. Cela prouve que les conseils de Google pour choisir avec soin la catégorie à laquelle appartient votre application vous permettent de bien cibler vos consommateurs.

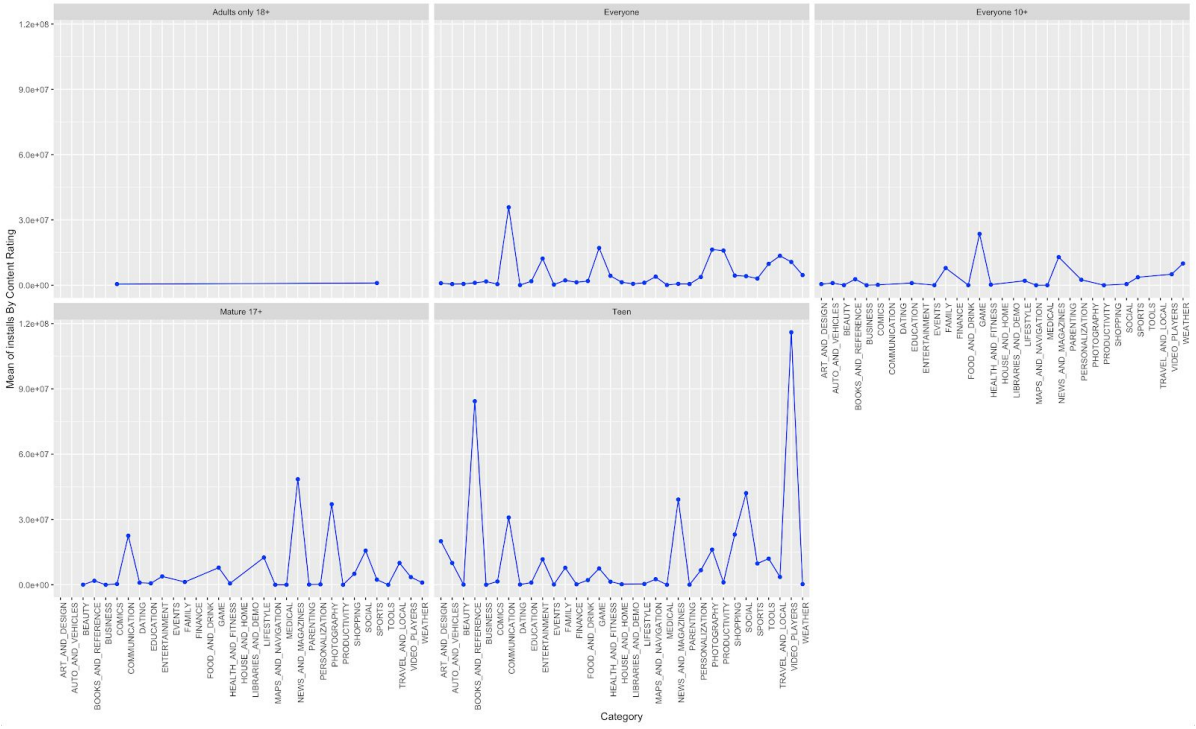
La matrice de corrélation suivante montre le coefficient de corrélation entre les différents champs de notre jeu de données après l'élimination des applications qui n'ont pas de taux d'évaluation. On voit clairement qu'il y a une corrélation entre Installs et Reviews avec un taux de 0,62.
On n'a pas une corrélation entre Rating et Installs ou Reviews, imaginez une vidéo Youtube avec des millions de vues, il n'y a aucun moyen d'avoir 100% de retour positif. Mais une vidéo avec seulement quelques visionnements (autant que les amis et les parents de youtuber), est tout à fait possible d'avoir zéro dégoûts. C'est ce qui se passe exactement avec de telles évaluations.





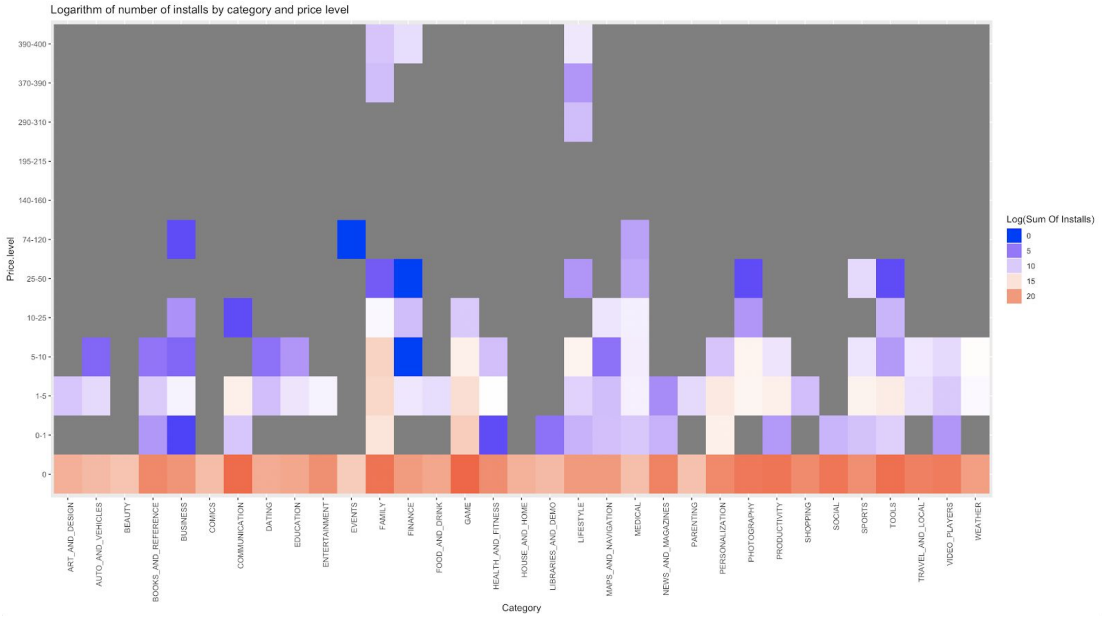
Le graphique ci-dessus montre que la courbe de la moyenne d'installation des applications est incrémentale avec la taille des applications et ceci s'explique par le fait que les applications les plus installées comme vu précédemment sont celle de gaming et le graphique ci-dessous montre que presque 50% ont une taille supérieure à 50Mo. La majorité des applications installés ont une taille qui varie avec le périphérique utilisé, cette remarque est importante, on aura son explication par la suite. La conclusion qu'on a pu générer à partir de ce graphe est que la taille d'une application n'affecte pas vraiment le nombre d'utilisateurs.





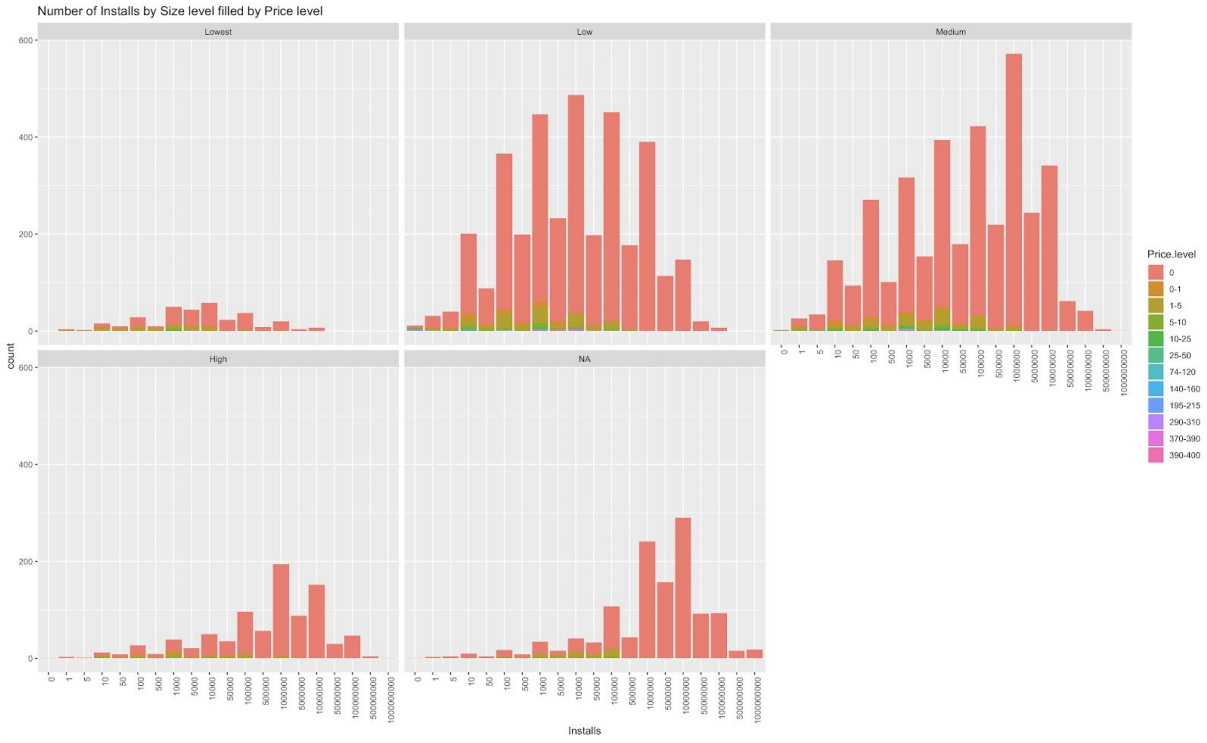
Les installations par rapport au nombre d'applications dédiées au Teen sont plus importantes que les autres classements de contenus, ces applications sont généralement adapté pour les 13 ans et plus, peuvent contenir de la violence, des thèmes suggestifs, de l'humour brut, un sang minimum, une simulation de jeu et / ou l'utilisation peu fréquente d'un langage puissant.
On peut dire que les catégories qui ont pu bien se démarquer dans ce graphique représentent un niche pour les développeurs cherchant un domaine, surtout les applications books and reference, new and magazines et photographie qui ne sont pas trop difficile à développer; vous pouvez jeter un coup d'oeil sur google play store et vous allez avoir des idées.

Il existe un nombre important des applications installées appartenant aux catégories Lifestyle, Finance et Family avec un prix trop élevé, à l'ordre de 400$.Examinons en quoi consistent ces applications.



App	Category	Rating	Reviews	Size	Installs	Type	Price
I'm Rich - Trump Edition	LIFESTYLE	3.6	275	7.3000000	1e+04	Paid	400.00
most expensive app (H)	FAMILY	4.3	6	1.5000000	1e+02	Paid	399.99
💎 I'm rich	LIFESTYLE	3.8	718	26.0000000	1e+04	Paid	399.99
I am rich	LIFESTYLE	3.8	3547	1.8000000	1e+05	Paid	399.99
I am Rich Plus	FAMILY	4.0	856	8.7000000	1e+04	Paid	399.99
I Am Rich Premium	FINANCE	4.1	1867	4.7000000	5e+04	Paid	399.99
I am Rich!	FINANCE	3.8	93	22.0000000	1e+03	Paid	399.99
I am rich(premium)	FINANCE	3.5	472	0.9423828	5e+03	Paid	399.99
I Am Rich Pro	FAMILY	4.4	201	2.7000000	5e+03	Paid	399.99
I am rich (Most expensive app)	FINANCE	4.1	129	2.7000000	1e+03	Paid	399.99
I am Rich	FINANCE	4.3	180	3.8000000	5e+03	Paid	399.99
I AM RICH PRO PLUS	FINANCE	4.0	36	41.0000000	1e+03	Paid	399.99
I Am Rich	FAMILY	3.6	217	4.9000000	1e+04	Paid	389.99
I am extremely Rich	LIFESTYLE	2.9	41	2.9000000	1e+03	Paid	379.99
I am rich VIP	LIFESTYLE	3.8	411	2.6000000	1e+04	Paid	299.99

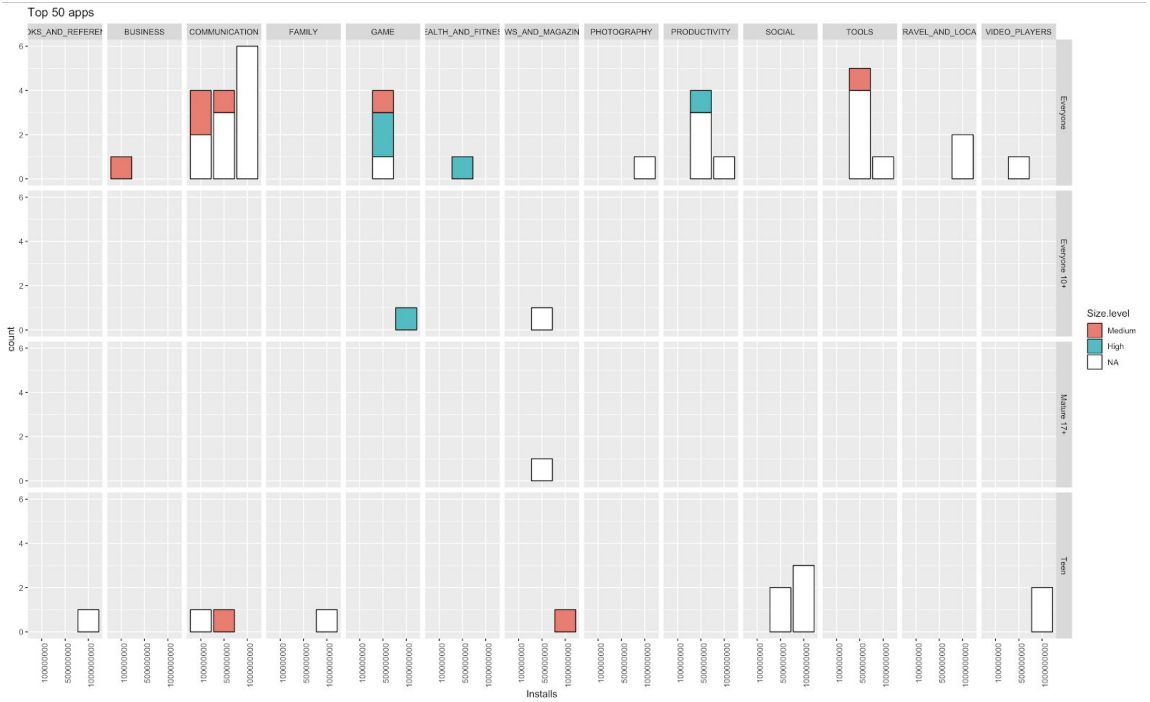
Ces applications ont des tailles petites parce qu'elles ne font rien, on s'est renseigné sur ces applications sur google play store et on a remarqué que ces application ne contiennent que des images de diamants ou dollars ayant en description "This will be the proof that you are really rich".
Mais pourquoi google autorise-t-il ce genre d'applications ?



Presque 80% des applications payées ont une taille optimale inférieur ou égale à 50 Mo. D'après le graphique précédent la majorité des applications payées avec un prix entre 0.x$ et 10$ appartiennent aux catégories Family et Game, comme on a vu précédemment plusieurs applications de ces catégories ont des tailles supérieurs ou égales à 50 Mo alors les développeurs essayent de réduire la taille de ces applications et en contrepartie les lancer sur google play store avec un prix raisonnable.

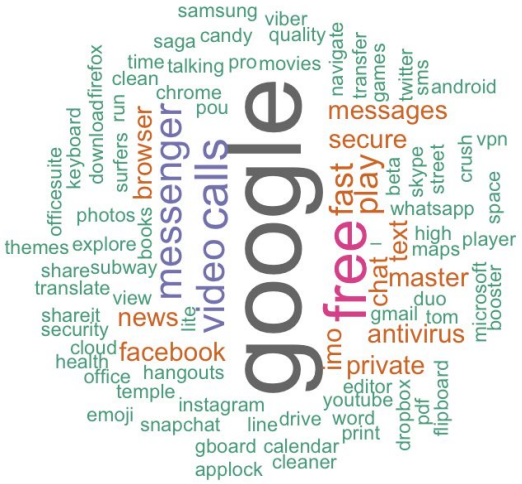
2- Top 50 des applications installées :

On a choisi de classer les applications par nombre d'installations, ainsi les applications au top sont les applications avec le plus d'installations.



Les top 50 des applications les plus fréquentes sont celles qui appartiennent à la catégorie communication, avec le plus grand nombre d'installations. Les applications des réseaux sociaux sont dédiés au public Teen parce que ces applications contiennent des thèmes suggestifs, de l'humour brut, un sang minimum, une simulation de jeu et / ou l'utilisation peu fréquente d'un langage puissant, elles sont des applications dans lesquelles toute personne connectée peut publier n'importe quel contenu, parfois on trouve des contenus relatifs à la nudité et à l'activité sexuelle ou contenus graphiques violents, c'est pour ça que les fournisseurs des applications des réseaux sociaux tentent de protéger le contenu partagé. Prenons Facebook comme exemple, il dispose de différents locaux dans le monde pour la lutte contre ces publications dans ces différentes plateformes en lignes.

Vérifions les noms de ces applications par nuage de mots-clés.



Google est le mot le plus fréquent, ainsi les applications appartenant à google qui sont les plus installées. Il est important de noter que ces applications sont toutefois des téléchargements requis pour tous les appareils Android. En d'autres termes, les utilisateurs ne choisissent pas nécessairement de télécharger ces applications, ils étaient fournis préchargés sur leurs appareils (ou du moins, ils étaient invités à les télécharger automatiquement lors de la configuration initiale de leur appareil).

Ensuite on remarque messenger et video calls, ceci confirme que les applications de communications sont au top comme vu précédemment. Finalement le mot Free, pour s'appuyer sur l'idée que cette application est gratuite d'une part, et d'autre part pour faire la différenciation parce que certains fournisseurs ou développeurs proposent des versions gratuites et des versions payantes améliorées.
Les petits mots représentent les noms des applications parce que les mots utilisés pour ce nuage de mots-clés sont dénombrables.

Qu'est ce qui est en commun entre ces applications ?

Toutes ces applications sont gratuites, c'est pour cela que le mot free est fréquent dans les noms. Dans la première section on a dit que les applications les plus installées ont la taille, la version courante et la version android supportée varient avec le périphérique utilisé, ceci s'explique par le terme "compatibilité". Ces applications sont compatibles avec chaque configuration de périphérique potentielle. Étant donné qu'Android s'exécute sur une large gamme de configurations de périphériques, certaines fonctionnalités ne sont pas disponibles sur tous les périphériques. Par exemple, certains appareils peuvent ne pas inclure de capteur de boussole. Si les fonctionnalités de base de votre application nécessitent l'utilisation d'un capteur de boussole, votre application est uniquement compatible avec les appareils comprenant un capteur de boussole. Finalement, ces applications ont fait des mise à jours durant quelques mois avant la collection des données pour ce jeu de données pour plusieurs bonnes raisons et parfois car google l'exige pour garder ces top applications en recommandation. Les mises à jour peuvent parfois avoir un effet négatif, et c'est pour ça que ces applications n'ont pas un taux d'évaluation à 100%, par exemple dernièrement vers la fin du mois Novembre, l'application Messenger à lancer une mise à jours qui a eu des retours négatifs.

Durant notre analyse, on a touché les différents facteurs qui influent sur une application dans google play store, à partir du choix d'un public et d'une catégorie jusqu'à le développement d'une application compatible avec les différents périphériques avec un nom significatif. D'après ceci, on peut dire que les conseils données par google aux développeurs sont prises en considération, d'ailleurs on a cherché sur google play store les applications ayant le nombre minimal d'installation et on a pu les retrouver par leurs noms complets, pour les applications appartenant à des catégories dont le nombre d'applications n'est pas trop grand on a pu les trouver juste avec des mots-clés.

III- Comment produire une application Android réussie?

Si vous cherchez à gagner un grand nombre d'utilisateurs pour votre applications, vous devez d'abord vous positionner dans une catégorie et vous pouvez revoir en-dessus l'analyse des catégories de google play store pour savoir bien cibler et développer votre application en suivant les conseils et consignes de google pour les développeurs. Et bien sûr relire cet article attentivement pour fixer vos idées.
Sinon, pour produire une application afin de gagner de l'argent, vous pouvez commencer à lancer des publicités sur votre application après avoir un bon nombre d'utilisateurs ou vous pouvez juste développer des applications avec quelques photos et les lancer avec le nom I'm Rich avec les différentes langues et des descriptions attirantes. Cette stratégie marche bien, vous pouvez consulter les applications les plus chers dans la section précédente.

Conclusion :

Cette analyse nous a permis d'appliquer et de mettre en pratique les méthodes de visualisation, d'exploration et d'analyse des données.

Nous avons été confronté au problème des données manquantes et des données aberrantes qu'on a pu résoudre avec l'ingénierie des fonctionnalités. Nous nous sommes également attachés à créer des graphiques permettant une bonne compréhension du jeu de données adoptés, ainsi que de comprendre les application de google play store et la logique suivie par google.

Ce travail est un bon travail d'initiation pour nos prochains projets de données massives où on va appliquer des algorithmes d'apprentissage sur une infrastructure big data, car tout commence par l'analyse, le nettoyage et la compréhension des données.

Références:

https://play.google.com/
https://www.youtube.com/watch?v=862r3XS2YB0
https://www.youtube.com/watch?v=Jc-LEG0T_4c
https://www.youtube.com/watch?v=x1AYelepG6o
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3795955/
https://www.researchgate.net/publication/220687682_Quantifying_the_trustworthiness_of_social_media_content
https://www.indiatoday.in/technology/news/story/facebook-changes-rules-to-remove-abusive-content-244810-2015-03-18
https://support.google.com/googleplay/answer/1727131?hl=fr&ref_topic=3364265
https://developer.android.com/guide/
https://developer.android.com/about/
https://developers.google.com/analytics/
https://support.google.com/googleplay/android-developer#topic=3450769
https://www.facebook.com/help/212722115425932?helpref=faq_content
https://www.facebook.com/communitystandards/safety
https://www.indiatoday.in/technology/news/story/facebook-changes-rules-to-remove-abusive-content-244810-2015-03-18
https://android-developers.googleblog.com/2017/12/improving-app-security-and-performance.html
https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/
http://www.sthda.com/english/wiki/wiki.php
http://www.greenteapress.com/thinkstats/thinkstats.pdf