Data Mining Application Play Store

Canete Valentin, Perrin Thibaud, Tisserand Jean-philippe, Bruyat Julian, Dailler Arnaud

**Introduction :**

Le but de ce rapport est d’extraire des profils d’applications assurant une réussite sur le Play Store à partir des corrélations entre les différentes caractéristiques d’applications contenu dans une base de données Play Store. Les résultats seront vérifiés par une méthode de clustering (K-mean). Deux jeux de données différents seront utiliser afin d’appuyer les résultats

Mots clés: Data mining, Clustering, Play-Store, Correlation

**Données :**

Les données sont récupérées du fichier Google-Play Store-Full.csv. Ce fichier contient 267 053 applications et leurs caractéristiques (App Name : nom de l’application, Catégories, Reviews : nombres de notes, Installs : nombre de téléchargement, Size : taille de l’application en Mo, Price : prix en dollars, Content Rating : public visé, Last Updated : date dernière Maj, Minimum version : version minimum, Latest version : dernière version). Nous avons décidé de ne pas utiliser les versions pour l’analyse.

Traitement des données : Avant de commencer toute analyse il a fallu mettre en place une phase de traitement des données : Rassembler certaines catégories, Convertir type de donnée, Supprimer les missing value (workflow 1 : Annotation “Traitement des données”)

Beaucoup d’applications portent la mention “Varies with device”. Pour gagner un nombre d’applications suffisantes, nous avons décidés de ne pas convertir cette colonne. A la fin du traitement des données il reste 267 051 Applications.

Pour la répartition du nombre d’applications par catégories on obtient

Une image contenant capture d’écran

Description générée automatiquement

Nous avons rassemblé tous les types de jeu au sein d’une même catégorie “GAME”, le détail de la répartition des applications par type de jeu est le suivant :

Une image contenant capture d’écran

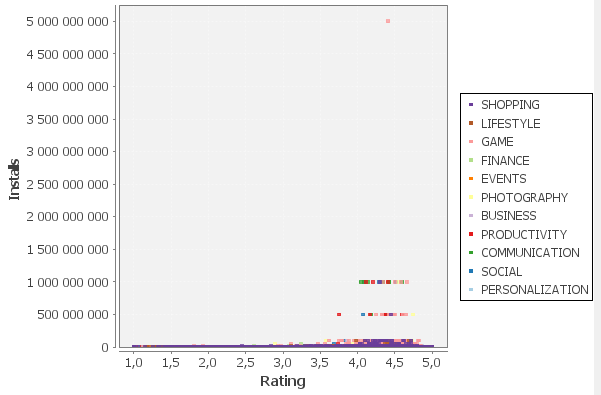
Description générée automatiquement

**Analyse des caractéristiques :**

Le nombres de téléchargements et la note d’une application sont deux paramètres importants pour indiquer la réussite sur l’App Store, dans un premier temps nous avons cherché des corrélations entre les paramètres souhaités (nombre de téléchargement) et les paramètres influent (note , taille, note, prix, public visé, catégorie) afin d’extraire les paramètres qui peuvent faire d’une application, une bonne application.

Nous présentons un scatter plot (coloré) pour montrer la répartition des applications en fonction de leurs groupes :

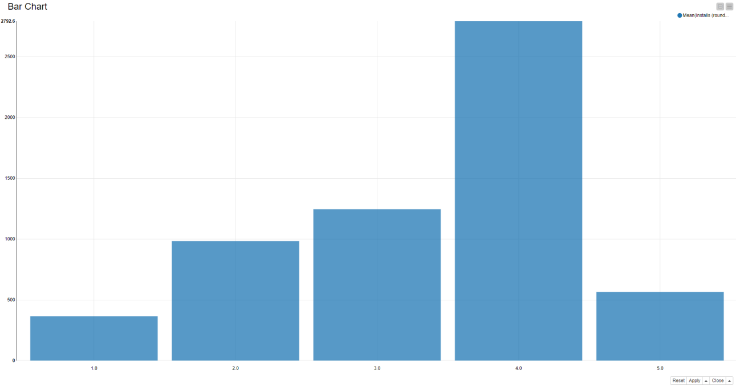
**Etude des catégories sur le nombre de téléchargements :**



Comme on peut le voir, les catégories n’ont pas réellement d’influence sur les paramètres de réussite d’une application dans le Play store, cependant on peut commencer à voir certaines tendance (jeu plus souvent téléchargé, etc..) mais on peut également voir qu’une application qui n’a pas une note inférieure à 3 ne dépassera pas les 100 000 000 de téléchargements.

**Etude de la note sur le nombre de téléchargement**

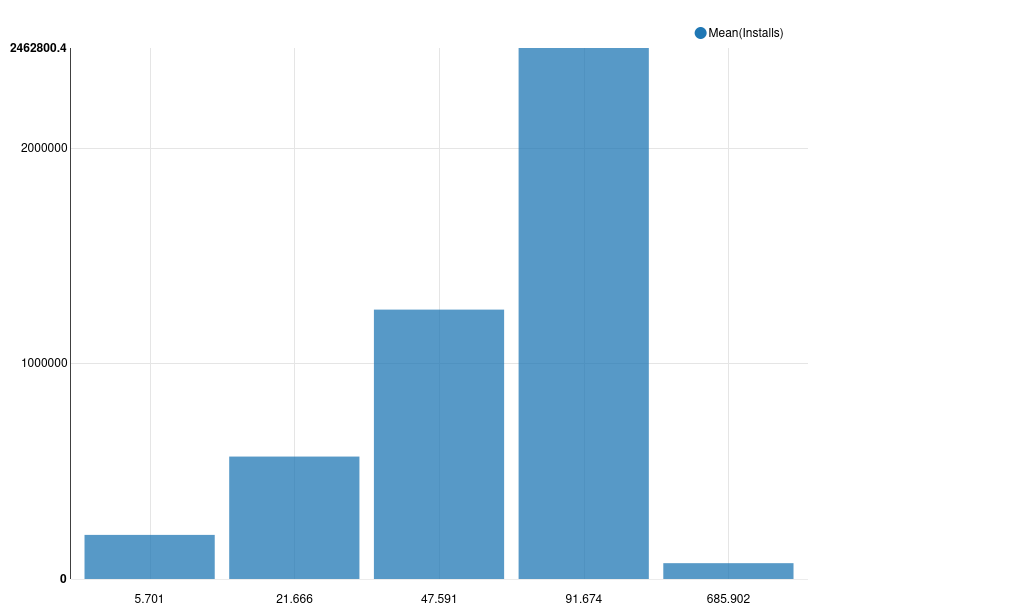
La notes et le nombre de téléchargement sont donc deux caractéristiques fortement corrélées



Il est donc conseillé de garder une note de l’application supérieure à 3 pour atteindre un nombre de téléchargement suffisamment élevé

**Etude du nombre de téléchargement selon le poids de l’application :**

Le poids est un facteur qui peut paraître important si on considère le fait qu’un utilisateur peut être réticent à l’idée d’installer une application lourde sur son téléphone. C’est donc un paramètre qui doit être étudié, pour cela on rassemble les poids par clusters et on regarde la moyenne des téléchargements par cluster



On voici ici que pour obtenir un nombre d’installation maximum, il faut veiller à créer des jeux aux alentours 90 Mo par application, une taille trop petite ou trop grande fera diminuer le nombre de téléchargement

**Etude du nombre de téléchargement selon le public visé :**

Le public visé est un facteur qui va forcément influencer le nombre de téléchargement, il existe 5 catégories de public visé, on observe la moyenne de téléchargement

Une image contenant texte

Description générée automatiquement

Selon le graphique obtenu, la catégorie 10 ans et + est la plus téléchargée. Et viser un public supérieur à 18 ans empêche un grand nombre de personne de télécharger l’application et donc baisse le nombre de téléchargement.

**Etude du nombre de téléchargement selon le prix de l’application :**

Pour le prix nous avons également regroupe par cluster pour faire des intervalles de prix et afficher le nombre de téléchargement :

Une image contenant capture d’écran

Description générée automatiquement

Comme on pouvait l’espérer, moins l’application est chère, plus le nombre de téléchargement est élevé.

Si on fait une conclusion sur les observations précédentes, une application type qui engendrerait un maximum de téléchargement serait une application de n’importe quelle catégorie, gratuite, avec un poids inférieur à 100 Mo, qui s’adresse à un public de 10 ans et plus. Il est important de conserver une note moyenne supérieur à 3.