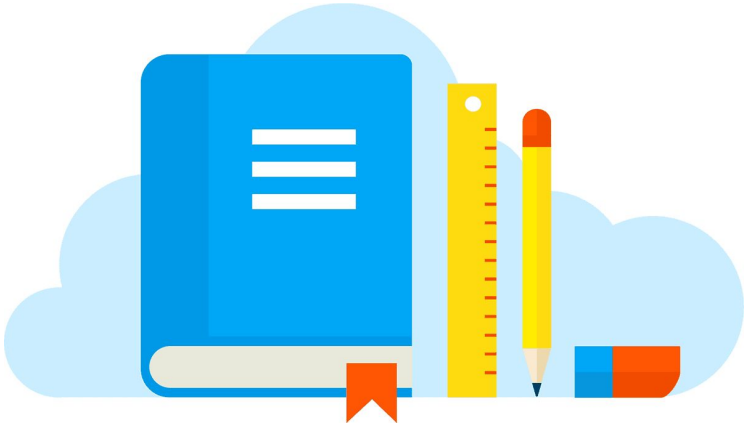


Mini Projet

Machine Learning

Lemieux Mickael IOT M1

Le mini projet de Machine Learning portera sur des statistiques d'applications de Play Store.

Lien du CSV :
https://www.kaggle.com/neomatrix369/google-play-store-apps-extended

Data

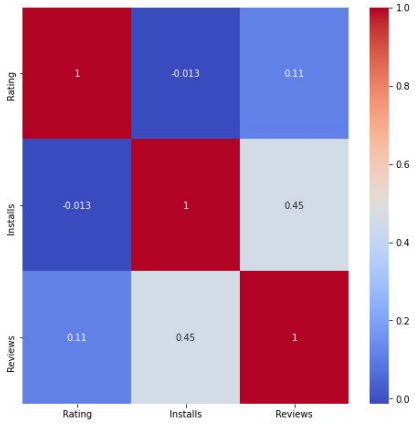
- **Partie 1** : Analyse graphique des données
- **Partie 2** : Model Building
- **Partie 3** : Features Importance

Analyse Graphique des données

Description des variables

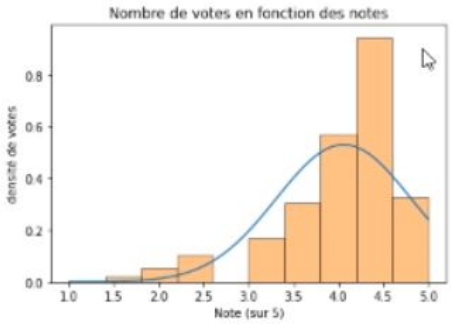
App : Nom de l'application
Rating : La note de l'application
Reviews : Nombre de revue de l'application
Size : Taille de l'application
Installs : Nombre d'installation de l'application
Content Rating : Note de maturité de l'application
Last Updated : Dernière update
Current Ver : Version actuelle de l'application
Android Ver : Version android utilisé pour l'application
Category : La catégorie
Type : Payant ou non
Genre : Le genre du jeu
Last Updated : Dernière update catégorique
No_reviews_count : Nombre de revue non compté
Reviews_present_count : Nombre de revue compté

Premièrement, il était obligatoire de devoir nettoyer les données, vérifier les types de colonnes, des NULL, et prendre les colonnes intéressantes pour le traitement futur.

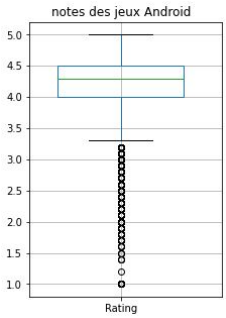


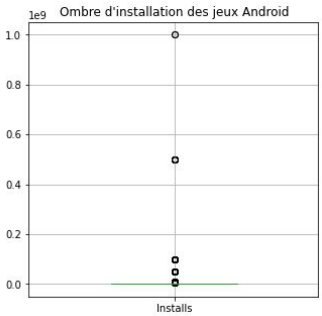
En second, il nous était donné de devoir réaliser des graphiques permettant de mettre en valeur s'il y avait une corrélation ou non avec nos choix de colonnes. Dans ce cas, la problématique étant :

-> Une application a-t-elle une corrélation entre le nombre de téléchargement et sa notation sur 5 ?



On peut alors déterminer au vu des graphiques qu'il n'y a qu'une très faible corrélation entre ces deux choix de variables, voire quasi nulle, de plus dans la boîte à moustache, il me fallait devoir supprimer les valeurs extrêmes.



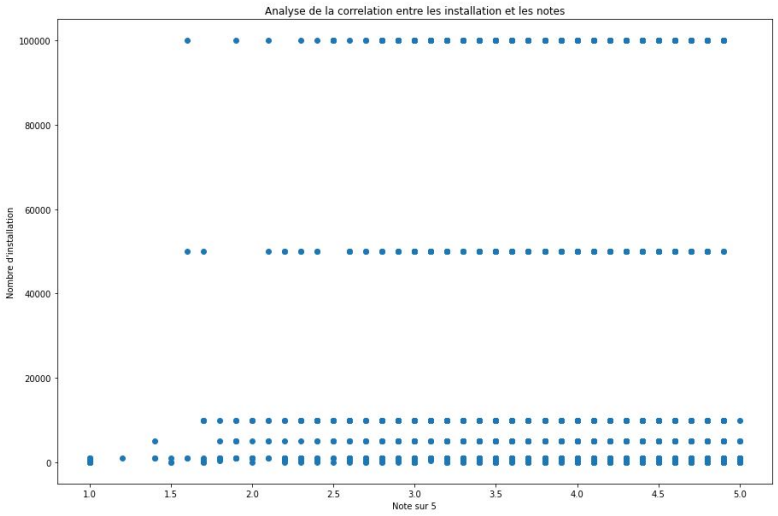


Modèle Building

Ensuite, grâce à un algorithme de régression linéaire, l'on pouvait observer s'il y avait une corrélation visuelle sur ce que l'on supposait dans la problématique.

Sur ce graphique de nuage de point, il est clairement observable qu'il n'y a aucune corrélation entre le nombre de téléchargement de l'appli et sa notation.

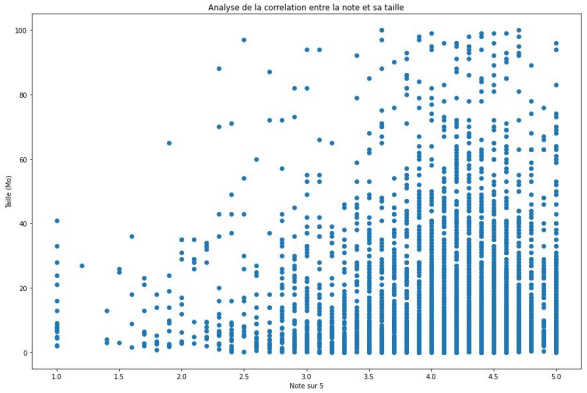
De plus son calcul de coef reste à -0.00



J'ai voulu essayer avec une autre problématique mettant en scène la corrélation entre la taille d'une application et sa notation sur 5.

Sur ce nouveau schéma, une corrélation est bien plus visible !

Elle est à hauteur de 30% mais reste toujours très faible et ne corrèle pas vraiment, puisqu'elle est en dessous de 50%.



Features Importance

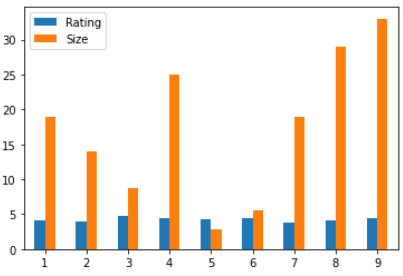
En dernier, il nous était demandé de faire un affichage graphique en BarPlot avec Pandas.

Malheureusement, mes données ne me permettaient pas de faire des graphiques visiblement et avec des données vraiment utiles.

J'ai alors pris deux colonnes de 9 datas Max pour faire un exemple lisible.

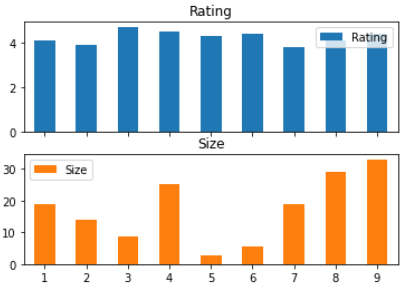
Encore une fois, l'association de la Note / Taille (Mo)

J'ai fait 3 types de diagrammes en BarPlot pour pouvoir essayer les deux colonnes sur 10 valeurs choisie.



Cela permet de voir la différence entre les colonnes et surtout de voir un écart comme on pourrait le voir avec (bleu : Taille, Rouge : Poids)

Cet exemple ne corrèle pas vraiment mais cela est là pour pouvoir mettre en valeur un exemple.



Finalement, ce type de graphique permet de montrer les comparaisons entre des catégories discrètes.