**DATA WRANGLING**

*Nettoyage du dataset*

**Introduction**

Le volume de données des datasets de nos jours ne cessent de grandir, de même que les erreurs liées à leurs données. Par conséquent, il est normal de devoir effectuer des nettoyages de données de plus en plus minutieux afin de renforcer l’intégrité et la pertinence de ceux-ci en réduisant les incohérences. Eviter les erreurs permettent de rendre les décisions prisent via les données plus précises et mieux avisées.

Le processus de nettoyage des données permet d’identifier et corriger les données altérées, inexactes ou non pertinentes. Ce processus est fondamental afin d’améliorer la cohérence, la fiabilité et la valeur de nos données.

Dans la plupart des cas, l’inexactitude des données viennent de valeurs manquantes dans le dataset, des fautes de frappes ou une donnée se trouvant au mauvais emplacement.

Plusieurs manières de nettoyage sont possible ; Correction des données, renseignement des données ou simplement suppression de celles-ci.

**Analyse**

Nous avons remarqué durant l’analyse de notre dataset certaines incohérences ou erreurs dans celui-ci.

Premièrement dans la colonne APP\_NAME nous avons remarqué que certaines valeurs était incorrect. En effet la colonne comportait énormément de nom d’application composé uniquement de point d’interrogation. Elles ne représentaient pas des noms d’applications.

Dans la colonne Category, certaines valeurs ne représentait pas du tout des catégories et était donc des valeurs incorrect.

Ensuite, la colonne reviews ne comportait pas de virgule et donc était pratiquement uniquement composé de valeurs aberrantes qui était censé être des valeurs avec un maximum de 5.

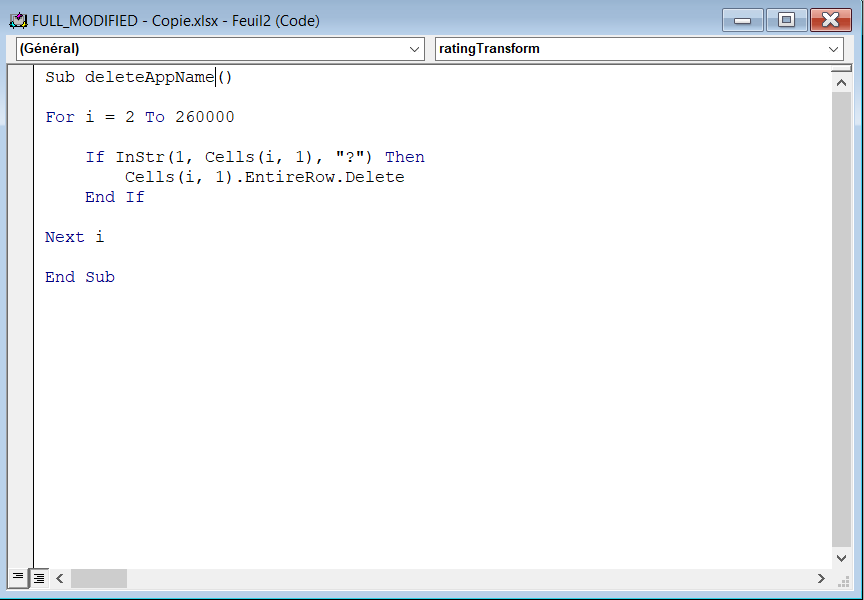
Pour continuer, la colonne Installs n’était pas une colonne en chiffre mais une colonne en « string » composé de chiffre suivi dans la plupart des cas d’un « + ». Un champ « EDUCATION » se trouvait également dans cette colonne.

Enfin, les dernières colonnes ne comportaient aucune donnée et était donc totalement inutile.

**Mise en pratique**

Pour la modification des données du dataset nous avons décidé d’utiliser EXCEL. Effectivement EXCEL fournit un accès à un espace développeur permettant de modifier les données à notre guise via un langage de programmation simple à utiliser appelé VBA.

Pour la première colonne APP\_NAME nous avons décidé de simplement supprimer toutes les lignes dont la valeur de cellule se trouvant dans APP\_NAME était composé d’au moins un point d’interrogation. Pour ce faire nous avons utiliser cette fonction que nous avons créé. Cette fonction permet de parcourir le jeu de données grâce à une boucle allant de la ligne 2 à la ligne 260000 en vérifiant si la valeur de la cellule de la première colonne (« APP\_NAME ») contient ou non un point d’interrogation. Si cela est le cas, la ligne entière est supprimée.



Ensuite la colonne CATEGORY était composé de valeurs incorrectes. Nous avons décidé de supprimer les lignes composés de celles-ci. Les valeurs étaient :

« ) »

« 6 »

« Alfabe ï¿½?ren »

« Islamic Name Boy & Girl+Meaning »

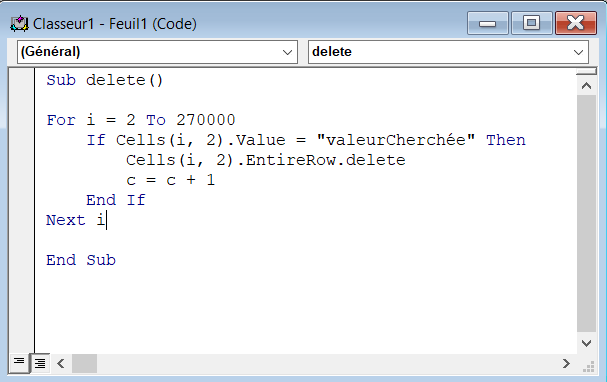
« not notified you followed – »

« Mexpost) »

« Tï¿½rk Alfabesi »

Valeurs vies

Pour ce faire nous avons utilisé une fonction que nous avons créée. Cette fonction permet de parcourir toutes les cellules de la deuxième colonne (« CATEGORY ») du dataset via une boucle. Si la valeur de la cellule correspondait à la valeur que nous voulions supprimer, la ligne entière était supprimée.

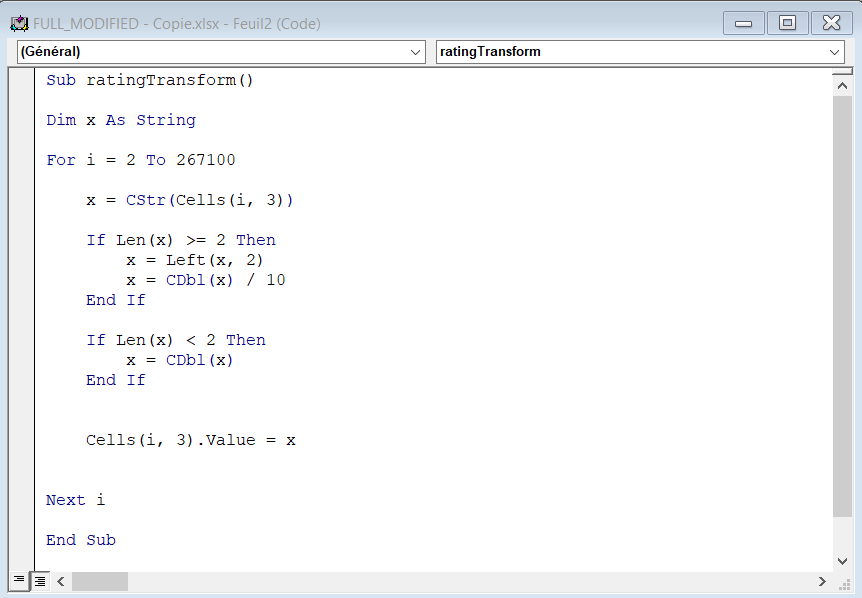


Pour continuer, la colonne reviews devait accueillir la note moyenne sur 5 des applications. Malheureusement il manquait la virgule ce qui nous donnait des nombres très important et pas du tout représentatif d’une note /5. Pour arranger ca nous avons implémenter une fonction. Cette fonction permet de transformer le chiffre en chaine de caractère.

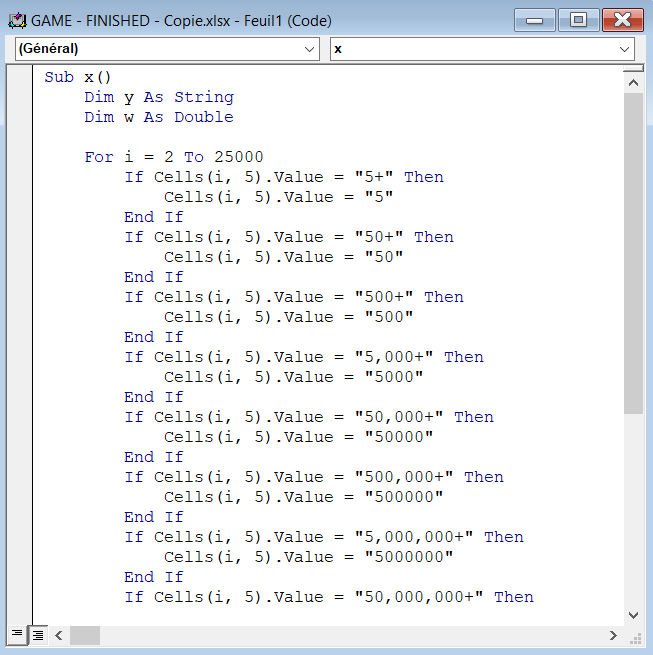
Si la longueur de la chaine est égale ou supérieur à 2, nous retenons les eux premiers caractères de celle-ci, la transformons en double et la divisons en deux.

Si dans le cas contraire, la longueur de la chaine est inférieur à 2, alors nous transformons seulement la chaine en double.

Pour finir nous remettons la nouvelle valeur dans la cellule adéquate.



Concernant la colonne « Installs », les données dans cette colonne étaient des chaines de caractères composé du nombre approximatif de téléchargement accompagné d’un « + ». Notre démarche d’étude se basant principalement sur le nombre de téléchargement, nous étions obligés de modifier cette colonne. Nous avons donc transformé chaque valeur en sa valeur correspondante via une fonction composé de conditionnelle. Nous avons également trouvé une valeur « EDUCATION » dans cette colonne qui était certainement une erreur d’encodage que nous avons simplement supprimé avec la fonction précédente recensé dans la colonne CATEGORY.



*Mise en forme du dataset*

**Introduction**

Afin de limiter son poids et optimiser son utilisation, un dataset doit être correctement construit. Les modèles de constructions les plus populaires sont des modèles de bases de données, dans lesquelles nous remplaçons les valeurs purement catégoriques via des identifiants, qui seront référencés par la suite dans des tables. Ce système permet d’alléger la table principale en remplaçant tous les champs répétés par des chiffres tandis que la vraie valeur n’est référencée qu’une fois dans sa propre table.

**Analyse**

Notre cas de dataset est simple, nous avons deux colonnes catégoriques, la colonne CATEGORY et la colonne CONTENT\_RATING, que nous pouvons changer en chiffre, pour créer des identifiants qui seront référencés sur des autres tables par après.

Il y aurait alors une table appelée « CONTENT\_RATING » composé d’une colonne identifiante et une colonne composée du label de chaque catégorie de CONTENT\_RATING.

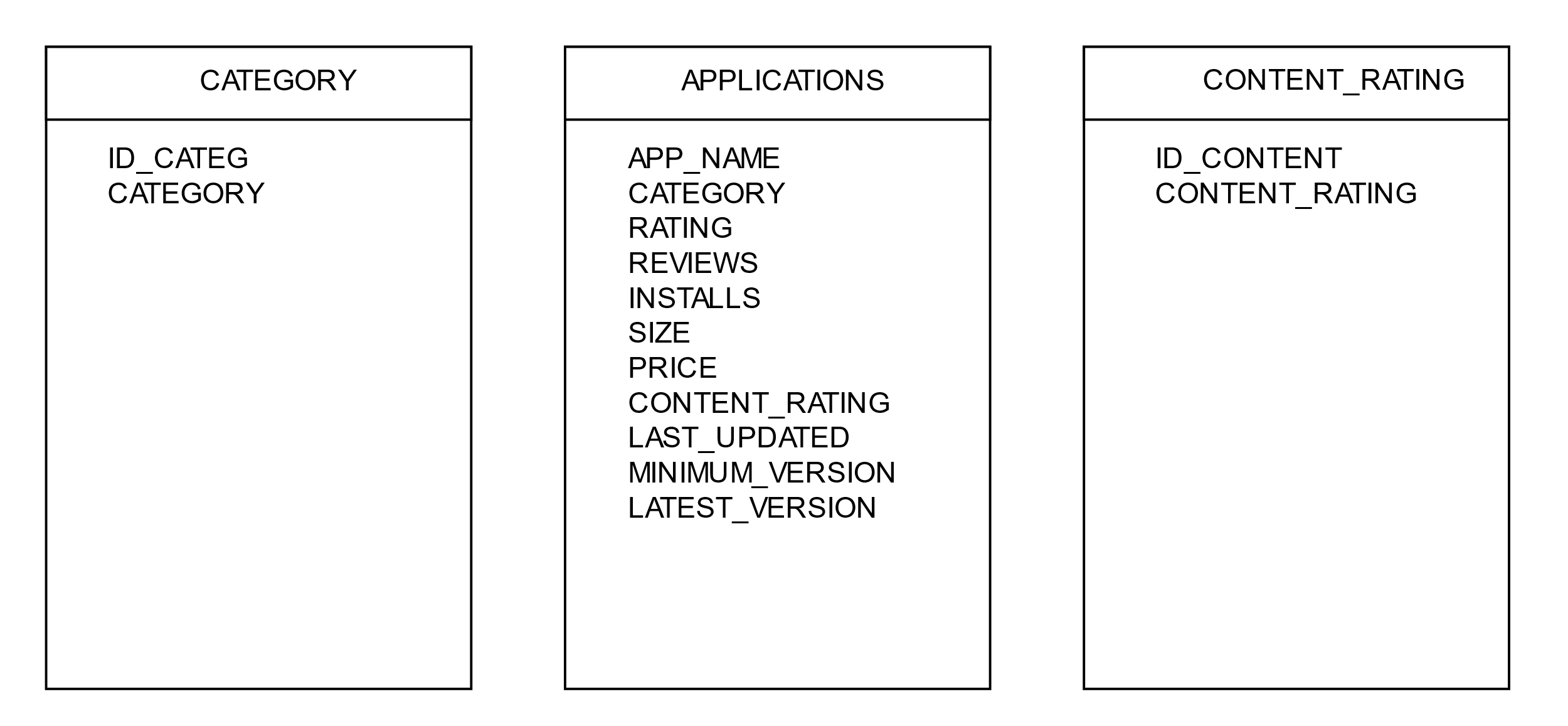
Pour ce qui concerne la table CATEGORY, le système est plus complexe. Possédant dans cette colonne beaucoup de sous-catégorie GAME (telles que « GAME\_ACTION », « GAME\_MUSIC », …), il serait intéressant de créer une étude sur un dataset où toutes les sous catégories de « GAME » seraient rassemblées dans une seule catégorie et une autre étude sur un dataset composé uniquement de toutes les sous catégories « GAME ».

**Mise en pratique**

En ce qui concerne la colonne CONTENT\_RATING, le protocole fut simple, nous avons d’abord créer une table contenant toutes les différentes valeurs de cette colonne et nous les avons accompagnés d’identifiants. Ensuite via une fonction en VBA nous avons remplacé toutes les valeurs de cette colonne par des identifiants correspondant.

Au niveau de la colonne catégorie, comme dit précédemment nous avons décidé de spliter notre dataset en deux. D’un côté un dataset avec les catégorie où les sous catégories « GAME » seraient globalisées dans une seule catégorie et d’un autre côté un dataset où il n’y aurait que les catégories de « GAME ».

La procédure fut la même que pour le CONTENT\_RATING, nous avons créer dans les deux cas une table avec les noms des catégories et un identifiants leur correspondant. Dans la table principale nous avons substituer les valeurs initiales par les identifiants correspondants.



Nous pourrions considérer cet ensemble de tables comme une base de données. Dans ce cas, dans la table CATEGORY nous pouvons considérer la colonne ID\_CATEG comme clef primaire. La colonne CATEGORY comporte les labels des catégories.

La table CONTENT\_RATING est similaire à la table CATEGORY. Nous pouvons considérer la colonne ID\_CONTENT comme clef primaire. La colonne CONTENT\_RATING comporte les labels.