Projet 2 :

Analyse et Visualisation de l'ensemble des données de l'archive de tweets « *WeRateDogs* »

RAPPORT D'ANALYSE

CONTEXTE DU PROJET

Pour le compte du deuxième projet de notre formation d'analyste de données, il nous est proposé de mener l'analyse et la visualisation d'un ensemble des données provenant de l'archive de tweets ***WeRateDogs.*** WeRateDogs est un compte Twitter qui évalue les chiens des gens avec un commentaire humoristique sur le chien. Cet ensemble de données est réparti dans trois (03) datasets différentes à savoir :

- L'archive twitter améliorée de WeRateDogs «***twitter-archive-enhanced.csv*** »;
- Les données supplémentaires regroupés dans le fichier « ***tweet-json.txt*** » ;
- Le fichier de prédiction d'image « ***image-predictions.tsv*** »

Ce travail a été guidé par les étapes de collecte, d'évaluation, de nettoyage, de sauvegarde et la visualisation de quelques variables de cet ensemble de données

LA COLLECTE DES DONNEES

L'exploration de cet ensemble de données part de la collecte des différentes dataset. Pour se faire,

- La dataset ***twitter-archive-enhanced.csv*** a été fournie ; il restait juste à l'importer dans le notebook de l'exploration.
- Les données supplémentaires devraient être collectées à partir de l'API Twitter *Tweepy* via le code twitter_api.py fourni, mais n'ayant pas pu obtenir l'autorisation d'utiliser cette API, nous avons utilisé les données résultantes stockées dans le fichier ***tweet-json.txt ;***
- Une url a été fourni pour le téléchargement des données de prédiction d'image. Nous avons utilisé cette url pour télécharger et importer le fichier ***image-predictions.tsv*** regroupant ces données.

L'EVALUATION DES DONNEES

Après la collecte, nous avons effectué une évaluation visuelle et programmatique de chaque dataset collecté.

Evaluation de « *twitter-archive-enhanced.csv »*
L'évaluation visuelle de ce dataset nous a permis de déceler des incohérences pour certaines valeurs et l'évaluation programmatique de ce dataset nous a permis de déceler des problèmes sur la structure de la table notamment la définition des types de certaines variables ainsi que les valeurs manquantes.

Evaluation de « *image-predictions.tsv »*
Ce dataset nous a paru plutôt propre car son évaluation (visuelle et programmatique) nous a déceler aucune anomalie.

Evaluation de « *tweet-json.txt »*
L'évaluation des données supplémentaires nous a principalement révélé des problèmes d'ordre structurels de la table tel quels le nom de la colonne « id » non conforme, les colonnes à valeurs multiples.

LE NETTOYAGE DES DONNEES

Nettoyage de « *twitter-archive-enhanced.csv »*
Nous avons nettoyé ce dataset en :

- Corrigeant les valeurs non conformes dont les valeurs exactes ont été retrouvé dans les textes de la colonne « text »,
- Modifiant les types de certaines variables jugés non conforme,
- Regroupant les colonnes présentant les stades de chien en une seule colonne,
- Supprimant les enregistrements relatifs aux retweets d'image,
- Supprimant les colonnes à valeurs manquantes car ces colonnes ne contenaient pratiquement pas de valeur.

Nettoyage de « *tweet-json.txt »*
Toutes les anomalies relevées sur ce dataset ont simplement été supprimé.

LA SAUVEGARDE DES DONNEES PROPRES

Après le nettoyage des ensembles de données, nous les avons fusionnés en un seul ensemble que nous avons sauvegardés en fichiers « ***twitter_archive_master.csv*** »

LA VISUALISATION DES VARIABLES

Nous avons effectué une analyse simplifiée de l'ensemble de données nettoyé et sauvegardé en produisant trois (03) observations et une (01) visualisation de l'état de certaines variables.