Titre : Analyse des publications Twitter de WeRateDogs

1. Rassembler les données

Pour ce qui est de la récolte, nous avons utilisé la méthode read_csv() de Pandas pour récupérer les données. A la seule différence que pour le fichier JSON, nous avons dû créer une fonction pour lire les données.

2. Evaluer les données

L'évaluation des données a été scindée en 3 compte tenu du nombre de DataFrame. On a pu en ressortir que :

a. Fichier des archives
   - Certaines colonnes ont des valeurs manquantes
   - Le type de données de certaines colonnes ne sont pas corrects telle que la colonne 'timestamp' qui est de type 'object' alors que celle-ci contient des dates
   - Les erreurs de frappe dans les colonnes 'rating_numerator' et 'rating_denominator' sont aussi à corriger. Certaines cotes sont déraisonnables
   - Le nom de certains chiens sont incorrects

b. Fichier des images
   Après évaluation, il se trouvait que le fichier des images ne contenait aucune valeur manquante, aberrante moins encore des valeurs dupliquées.

c. Fichiers JSON

- Les problèmes de qualité :
  - Certaines colonnes ont des valeurs manquantes, d'autres même n'en ont pas un seul
  - Le type de données de certaines colonnes ne sont pas corrects telle que la colonne 'created_at' qui est de type 'object' alors que celle-ci contient des dates
  - Les erreurs de frappe dans les colonnes 'rating_numerator' et 'rating_denominator' sont aussi à corriger. Certaines cotes sont déraisonnables
- Les problèmes d'ordre :
  - La présence inutile de certaines colonnes telle que 'in_reply_to_status_id_str' étant donné que cette dernière existe déjà avec le type adéquat

3. Nettoyage des données

Cette étape a été en deux étapes : le nettoyage lié à la qualité et le nettoyage lié à l'ordre.

Pour le nettoyage lié à l'ordre, nous avons :

- géré les valeurs manquantes : nous nous sommes convenus, premièrement, de garder uniquement les colonnes n'ayant pas plus de 75% de valeurs manquantes et puis, pour les colonnes de type numérique, les valeurs absentes seront remplacées par -1000 tandis que pour les colonnes de type 'object', nous remplacerons par 'Unknown'
- modifié le type de certaines colonnes : Celui-ci a été le plus appliqué pour les colonnes de type 'object' qui contenaient des dates. Sur ce, nous avons utilisé la méthode to_datetime() de Pandas.
- géré les valeurs aberrantes

Pour le nettoyage lié à l'ordre : Nous avons supprimé les colonnes n'avaient pas lieu d'être.

4. Stockage

Avant qu'il ne soit fait, nous avons fusionner nos 3 dataset en un seul et puis nous avons encore utilisé une methode de Pandas appelé to_csv('twitter_archive_master.csv') avec le paramètre indiqué poue ce projet.