Data Wrangle report

Nous avons travaillé sur un ensemble constitué de trois bases de données avec pour objectifs de mettre en valeur nos techniques d'évaluations de données. Le travail c'est principalement déroulé en deux phases:

1. Importation des bases de données

Au cours de cette étape, les trois bases de données ont été importées et transformées en un pandas DataFrame suivant trois différentes techniques.

- Le premier jeux de données 'twitter-archive-enhanced.csv' a été téléchargé manuellement et importée dans l'espace de travail grâce à la fonction read_cvs()
- Le second, 'image-predictions.tsv' a été téléchargé par programme grâce à la librairie requests et a l'url avant d'être importé dans l'espace de travail
- Le troisième quant à lui, 'tweet_json.txt' a été importé grâce à la librairie json.

2. Evaluation et nettoyage

Plusieurs analyses visuelles et programmatiques ont été effectuées sur les données. Les problèmes de qualité et d'ordre que nous avons repérés ont tous été fixés au cours de la phase de nettoyage.

- Les bases de données ont été fusionnées suivant la clé tweet_id
- Valeurs manquantes : elles ont ete remplacees par 'inconnu' dans la variable représentant les classes et supprimées pour les autres
- Certains noms de colonnes ne sont pas significatifs: ces noms de colonnes ont été renommés
- L'existence des valeurs dupliquées : elles ont été supprimés
- certains chiens sont multiclasses: une nouvelle variable pour les classes a ete cree et les quatres autres variables ont ete supprimees
- une colonne contient du code html: les liens de la variable source ont été remplacés par une simple chaine de caractere
- Certains chiens n'ont pas de classes: la classe 'inconnu' leur a été attribué
- Certains variables ont un mauvais type: le type de la variable timestamp a été converti en date_time
- Certains noms ne sont pas des noms de chiens: ces noms ont tous été supprimés
- Les valeurs 'None' ont ete remplacees par NaN