RAPPORT SUR LA PREPARATION DES DONNEES DU PROJET WE_RATE_DOGS

En tant que Analyste des données, bien souvent, les données que nous analysons nous sont fournies dans un sale état et en désordre. Notre devoir est alors avant toute chose, d'évaluer et de nettoyer ces données avant toute utilisation.
La préparation des données : c'est le processus qui regroupe l'évaluation et le nettoyage des données. Cette une étape très importante dans l'analyse de données car la fiabilité et l'exactitude de toutes les analyses et visualisations qu'on peut réaliser dépendent de la qualité de la préparation des données réalisée en amont.

Dans le cadre du projet **WeRateDogs**, nous avons passé la majeure partie de notre temps à évaluer et à analyser les données de manière à les rendre claires et prêtes par les analyses. Pour y arriver, nous avons fait recours au langage de programmation Python avec ses librairies(Pandas, numpy, re, tweepy,...).
A partir du fichier(twitter-archive-enhanced.csv) mis à notre disposition par WeRateDogs, nous avons procédé par :

1. **La collecte de données de données supplémentaires :**
   - De façon programmatique pour le fichier des prédictions de l'image tweet(image-predictions.tsv), grâce à l'identifiant de chaque tweet(tweet _id)
   - En utilisant l'API Twitter, pour avoir les nombre de retweets et de likes de chaque tweet contenu dans le fichier (twitter-archive-enhanced.csv)

2. **Évaluation des données**
   - De façon visuelle, nous avons constaté que certains tweet n'étaient pas évalués sur la note de 10, certains d'autres avaient une note bien au-delà de 17/10. Certains tweets, dans le texte la note attribuée contenait la partie décimale, ce qui n'a pas était pris en compte lors de la constitution de la colonne `ratting_numerator`, d'où c'était la partie décimale qui était prise en compte à la place de la partie entière.
   - De façon programmatique, nous avons constaté plusieurs anomalies, entre-autres le type de données, parfois int64 à la place de object, object à la place de datetime. Certains tweets avaient plus d'une note attribuée et qu'il fallait choisir la bonne note.

3. **nettoyage des données :** Pour chaque problème décelé, nous l'avons :
   - Défini le problème en question, en donnant une brève description de la logique à suivre pour le résoudre.
   - Résolu par le code, grâce au langage python avec ces librairies, nous avons traduit la logique qu'on a définie en amont en lignes de code pour le résoudre.
   - Vérifié si tout s'était bien passé : A la fin de chaque manipulation, nous faisions des tests pour voir si les résultats obtenus étaient conformes à la logique définie.

4. **Stockage des données**
   - A la fin, nous avons stocké les données que nous avons nettoyées dans un fichier afin de pouvoir réaliser les analyses.