Dans ce projet sur lequel nous avons travaillé , nous avons eu à travailler avec trois ensembles de données , parmi lesquels nous avons le premier ensemble de données "twitter_archive_enchanced.csv" qui a déjà été fourni et qu'on a eu à télécharger simplement , un deuxième ensemble de données "image_predictions" qu'on a télécharger à partir d'un site , et enfin le troisième dataset "tweet_json.txt" , chargé à partir de l'API twitter tweepy.
Aprés avoir télécharger tous ces datasets on les évalué et on a repérer les problèmes suivants

PROBLEME DE QUALITE :
***Dans le premier dataset twitter-archive-enhanced.csv

- timestamp est de type objet(string ou chaine) alors cela devrait etre de type datetime
- nous avons les colonnes in_reply_to_status_id, in_reply_to_user_id , retweeted_status_id , retweeted_status_user_id , retweeted_status_timestamp qui ne contiennent quasiment que des valeurs nulles respectivement 2278,2278,2175,2175, 2175 valeurs nulles ; nous avons aussi la colonne expanded_urls qui contient 59 valeurs nulles
- il y'a beaucoup de numerateurs et de dénominateurs qui contient des valeurs incohérentes comme des dénominateurs supérieures à 10 et des numérateurs avec des chiffres très grands qui sont problabement des nombres décimaux , donc des floats
- les étapes du chien doggo , fluffer , pupper , puppo doivent une seule colonne pour savoir si les chiens sont à quelle étape

***Dans le second dataset image-predictions.tsv
- nous avons un doublon au niveau des valeurs de la colonne img_url

****Dans le troixième dataset tweet-json.txt
- on supprime les retweets , on ne gardera que des tweets , ceux avec des images

PROBLEME D'ORDRE :
- le premier dataset enfreins à la premiére règle d'ordre "chaque variable doit constituer une colonne " , dans la colonne timestamp on a la date , l'heure et le mois dans la meme colonne or on devrait les séparer en ayant une colonne date , une colonne heure et une colonne mois .
-toutes les tableauxs doivent former un seul tableau

Aprés avoir évalué tous ces datasets nous avons essayé de les nettoyer et par la suite les fusionner tous en faire un seul dataset pour effectuer nos observations et et visualisations plustard .