Data Wrangling Report

Objectifs du projet:

Les objectifs du projet etaient:

- Collecter, évaluer, et nettoyer les données à partier des sources fournies.
- Stocker, mélanger et visauliser les données.
- Produire des rapports.

Step 1: Collecte les données

Ici, trois éléments de données ont été rassemblés et représentés sous forme de dataframes pandas (df) :

- L'archive Twitter de WeRateDogs (téléchargement manuel de "twitter-archiveenhanced.csv") ;
- Les prédictions d'images de tweet ('image-predictions.tsv'). Ce fichier a été téléchargé par programmation à l'aide de la bibliothèque Requests à partir d'une URL de UDACITY ;
- L'ensemble complet de données JSON de chaque tweet (avec au minimum l'identifiant de tweet, le nombre de retweets et le nombre de favoris) dans un fichier appelé « tweet_json.txt » a été stocké à l'aide de l'API Twitter et de la bibliothèque Tweepy de Python.
-

Step 2 and 3: Assessing and Cleaning Data

While working with data, a number of observations were made. In the below table there are the observations along with actions taken in the Cleaning Step.

Quality

Dataset	Observations	Solutions
Twitter archives	Valeurs manquantes dans les colonnes : in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp et extended_urls.	Nous les avons supprimés
	tweet_id est dtype int64 mais devrait être un objet	Nous avons converti le type de données
	L'horodatage doit également être un type datetime64 dtype et separer en année, mois, jour	Nous avons bien séparé en année, mois et jours
	Informations manquantes pour les certaines races canins.	Nous n'avons rien opérer comme changement ici
	Supprimer les colonnes qui ne seront pas utilisées pour l'analyse	Cela c'est bien été fait

	Corriger les numérateurs avec des décimales	Nous avons corrigé les numérateurs et dénominateurs. Toutes fois des difficultés ont été relevé face à cela
	Corriger les dénominateurs autres que 10	
	Le nom de colonne floofer doit être orthographié 'floof'	Nous avons renommé la colonne floofer en 'floof'
Tweet Count	La colonne "id_str" doit être remplacée par "tweet_id" afin de pouvoir fusionner les tables.	La colonne "id_str" et bien été remplacée par "tweet_id" afin de pouvoir fusionner les tables.
Image prédiction	Les types de chiens dans les colonnes p1, p2 et p3 avaient des lettres majuscules et minuscules.	Nous avons converti tous les noms en minuscules
	La colonne "tweet_id" doit être un objet dtype au lieu d'être int64.	Nous avons converti le type de données
	Nous devons 66 jpg_url qui sont dupliqué	Nous avons supprimé les valeurs dupliquées
	Nous devons supprimes les colonnes qui ne sont pas nécessaire pour notre analyse	Nous avons supprimé les colonnes inutiles
	Conservons la première prédiction vraie le long du niveau de confiance en tant que nouvelles colonnes	Nous avons créé une nouvelle colonne

Tidiness

Dataset	Observation	Solution
Twitter archives	La colonne source dans la table "twitter archive" semble désordonnée et encombre la table	Il on été supprimer
	Les colonnes doggo, floofer, pupper, puppo parlent toutes de la même chose, une sorte de personnalité de chien.	Nous avons crée un nouvelle colonee nommé dog_breed.