Wrangle_report

- ***Data Gathering***

The data was gathered from three primary sources:

Twitter Archive Enhanced: This CSV file contained detail of tweets such as tweet ID, timestamp, text, and dog stages (doggo, floofer, pupper, puppo).

Image Predictions: A TSV file providing image predictions for tweets, including dog breed predictons.

Tweet JSON: A json file containing additional tweet metadata such as language, user information, and retweet details.

- **Data Assessment**

After gathering the data, each data-set has it's own problems, missing values, and potential issues:

Twitter Archive Enhanced: This dataset had missing values in columns like

in_reply_to_status_id, retweeted_status_id, name, and dog stage columns.
The timestamp column was also stored as an object instead of datetime.

Tweet JSON: Missing values were observed in columns like retweeted_status, quoted_status, and place. The user column contained doubled information.

Image Predictions: This dataset had't alot of problems

- **Data Cleaning**

This is the following cleaning steps :

Handling Missing Values: useless columns with a high number of missing values (retweeted_status_id, name) were dropped.

Dog stage columns were cleaned by replacing string values with boolean (True/False) and filling missing values with False.

Type Errors: The timestamp column in "Twitter Archive Enhanced" was converted to datetime using pd.to_datetime.