This report contains documentation for data wrangling steps that is gathering the data, assessing the data and then cleaning it accordingly considering to clean to quality issues and tidiness issues.

Gathering Data:
Data is gathered in 3 different ways.
The data set 'twitter-archive-enhanced.csv' is collected manually and read using pandas.
The data set 'image_predictions.tsv' is downloaded programmatically using requests.
The data set 'tweet_json.txt' is gotten via Twitter Api.

Assessing Data
A copy of all the three above data sets is made so that even if any error is made, we can access and use the original data set.
A tidiness issue is resolved by merging all the three above data sets for ease of handling and it's copy is made as well.
Each and every column of the merged data set, df is assessed.
We come across multiple issues while doing so, like, the p1 column predicting the breeds of dogs has various words like hen, ox etc which aren't breeds of dogs. The same applies for p2 column and p3 column as well which represent the breeds of dogs.
The name column displaying names of dogs has various names like 'his' , 'a', 'the', which aren't names of dogs and should be corrected.
There are columns related to retweets like 'retweeted_status_id' , 'retweeted_status_user_id', 'retweeted_status_timestamp' which have nan values if the row is not a retweet and values otherwise. Since we need only original tweets and not the retweets, these rows need to be dropped.The issues found similarly for each column are stated in the inference section of the notebook

Cleaning data
There are various quality as well as tidiness issues which need to be rectified
Tidiness issue like merging 3 data sets into 1, there are 4 columns representing life cycle, combining them into 1 column, dropping irrelevant rows.
Quality issue like timestamp should be in date time format instead of object, corrections in the 'name column', dropping rows which are retweets, dropping rows which are replies to some tweets, dropping rows which are duplicates, correction in p1, p2 and p3 columns which have values other than breeds of dogs.