# Wrangle Report

## Data Gathering

This project contains three dataframes, first one is the Twitter-archive-enhanced.csv file which was downloaded from the course resources and read using pandas.

The second dataframe is the image_predictions dataframe which contained mainly top 3 predictions for the corresponding dog. This dataframe was obtained using the URL that was specified in the pandas read csv method. To read the TSV we needed to specify the sep parameter to be \t.

The third dataframe is the extra_archive which was obtained with the Twitter API tweepy by obtaining the authentication with the api tokens, the extra_archive dataframe contained the retweeted counts and the favorite counts. To obtain this file I gathered all the necessary information and wrote this info to a txt file that contained a string dictionary.

## Assessing Data

First of all the twitter_archive table, the table was assessed by checking the following:

Datatypes with the info() found some columns with inaccurate datatypes like timestamp.

The value count of the source column was checked.

The duplicated rows with duplicated(), no duplicated values were found.

The image_predictions table, the table was assessed by checking the following:

The first rows using the .head()

Datatypes with the info().

True predictions, some rows had other than dog animals.

The predictions with the highest confidence

The TwitterData table, the table was assessed by checking the following:

Datatypes with the info().

**Cleaning Data**

twitter_archive table:

Columns that are not needed are dropped.

Erroneous datatype in timestamp column was fixed from int to timedate format.

Erroneous datatype in tweet_id column was fixed from int to string.

Getting the values of source from the link and replacing them in the column.

retweeted_status_user_id is dropped

Predictions table:

Dropping columns that relates to 2nd and 3rd predictions.

column 'p2_conf' was renamed to 'p2_confidence'.

Erroneous datatype of tweet_id was fixed from int to string

TwitterData table:

The id column was renamed to tweet_id.

Erroneous datatype of tweet_id column was changed from int to string.

Tidy:

Four variables in four columns in twitter_archive table instead of a column (Doggo, floofer, pupper, puppo).

Merged the ArchivedTwitter table with TwiiterData table into a new table twitter_data.

Merged the twitter_data table with predictions table into twitter_master table