Wrangling report:

For this project there's five main parts, which is gathering data, Assessing data, cleaning data, then insight and visualization, but I will talk about the first four in this report and the last one in another report.

1) Gathering data:

In this part, we have to gather all the necessarily data for the next stages,

I found that there is more than one dataset that need to be gathered:

- twitter_archive_enhanced: it was the first dataset to gather, I downloaded it from udasity as csv file then included it in the project folder.
- image_predictions: For the second dataset I used a request library to download the tweet image prediction and save it as tsv format.
- The last one was a bit complicated , but I Used Tweepy library to query additional data via the Twitter API (tweet_json.txt) then saved it as data frame with three important columns that been extracted from tweet_json.txt, for the code it was taken from the udasity after I made needed changes to the code.

2) Assessing data:

For this stage, I had to explore each of the three data frames to understand the structure of each one of these data frame, then I started to record all of the mistakes that will be fixed later at cleaning data stage, and what I recorded for each data frame:

twitter_archive_df:

* The column tweet_id has type integer nested of string.

* Some dog names aren't real or valid like "a, an, mad, his, not, old, my,

such...etc"

* Some dog names has None value and it should replaced with NaN.

* The columns timestamp and retweeted_status_timestamp are type of string nested of datetime.

* The dog stages are divided into 4 columns which are "doggo, floofer, pupper or puppo" nested of combined as one column.

* The source column contain part of HTML code nested of just the source.

* The four dog stages has None nested of NaN.

* Some of the gathered tweets are retweets and it should be removed.

* There's columns that hard to read and it won't be needed.

image_predictions_df:

* There are 2356 tweets in the dataset twitter_archive_df but there are only 2075 in image_predictions_df dataset.

* The column tweet_id is an integer nested of string.

* The dog breeding type columns (p1, p2, p3) have underscores instead of white spaces.

* This dataset should be merged with the previous one.

tweets_df:

* There are 2356 tweets in the dataset twitter_archive_df but there are only 2354 in image_predictions_df dataset.

* The column tweet_id is an integer nested of string.

* This dataset should be merged with the previous two datasets.

3) Cleaning data:

For this stage I solved the issues in the previous stage and for some issues the solution was in one step, and the steps are:

- Merge the three data frames into one, and resolve the mismatch of tweets numbers.
- Clean the tweets from the retweets.
- Remove the unnecessarily columns.
- Fix the mismatch of the columns type.
- Remove some of the invalid dog names and change the None's to NaN.
- Merge the four columns of dog stage to one and replace the None to NaN.

- Remove the HTML code from the source column.
- Change the underscores at the columns p1, p2 and p3 to white spaces.

4) Storing data:

The purpose of this stage is so save the cleaned data frame to new dataset of type csv as the final one then use it for the visualization stage.