Data wrangling report.

Gathering:

1. Importing the twitter_archives csv file "downloaded manually" and turn it to dataframe.
2. Download the image predictions file programmatically from udacity server by the link provided. And extract the name of the file from the link
3. I have problems with creating twitter api account , and I do not have more time to wait. so I will proceed with the ready file. I will practise it soon¶. I used the code that is provided by udacity

Assessment:

- **I consider this step as the most important step as all steps that follow always been built on this step.**
- **The first step is the visual Assessment, at first, I will check the data on spreadsheet like excel**
- **i secondly worked on programmatically assessing.**
- **The high purpose is to collect as much as possible of issues in data even it is quality or tidiness issues**

(The issues that I managed to collect)

archive Table

Quality

1 - the column 'retweeted_status_id' has 181 non-null , these duplicated tweets should be deleted.(i will delete all retweet before deleting the column itself.

- ✓ Remove all retweets(the rows with non null values in retweeted_status_id)

2 - Missing values in columns: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, and some in expanded_urls.

3 - Columns doggo, floofer, pupper, puppo have wrong type of data as their value are none not nan.

- ✓ **create new column holds the type of the dog, and replace none with nothing, concatenate then replace empty with np.nan**

4 - Some columns of doggo, floofer, pupper, puppo have more than one value.

- ✓ **Concatenate them**

5 - timestamp has wrong type of data , should be datetime64

- ✓ **Turn its type**

6 - i think it is better to turn tweet_id to string as it is a key of the data not a num value.

- ✓ **Do this in the three dataframe before merging**

7 - column name has missing names , maybe that data got lost and got replaced by none // and there are wrong names like a and an and o

- ✓ **I will not depend on it in any valuable analysis, but won't delete it as it holds som insights as well**

8 - there is something wrong need to be checked, i think every tweet above 14 need to reviewed.

- ✓ **I checked them all and took many procedures**

9 - any number not equal 10 should be reviewed.

- ✓ **I checked them all and took many procedures**

Tidiness

1 - i think that in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp columns do not have any use and it is better to get rid of them as there hold unique values and we do not need them.(delete)

- ✓ i will delete in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp as tidiness 1

imgs Table

Quality

1 - upper & lower case missy in colums of dog type prediction p1 & p2 & p3

- ✓ I repared that issue by making all lower case

2 - i am sure that all image_predictions.p1 regestered 1 time is other thing or other animal , not a dog ,,, i need to check that

- ✓ Many of the images were something else and I excluded them through analyzing

3 - teddy & web_site has captured my eyes , need to checked

- ✓ Those are wrong of course

4 - Incorrect data type for tweet id.

- ✓ Did the transforming

api Table

Quality

- Incorrect data type for tweet id.
  - ✓ Did the transforming

Tidiness

- All three tables will eventually be merged into one.