Data Wrangling Report

1. Data Gathering

About the Data:

```
The Dataset wrangled is the Tweet archive of Twitter account
WeRateDogs--@dog_rates. Three (3) different datasets are used.
1.The archive dataset which is the main dataset and  contains 2356
tweet data.
2. The image predictions datasset which contains images based on
image predictions of the twitter archive datset and it contains the
tweetID, image URL,
and the image number with the most confident prediction
3. The tweet_json.txt data from which Retweet count,Favourite
Count,Display text range, and full text were extracted
```

Gathering the data:

```
The process used to gather each of the datasets are:
    1. Twitter archive file: This was downloaded manually using the
link provided by Udacity.
    2. Image Predictions: This was downloaded from the file hosted on
udacity server using pythons requests
       and locally saved it to image_predicitons.tsv file. The
downloaded file was then imported using Pandas as img_df
    3. Tweet_json : I downloaded the tweet_json file as provided on
the udacity servers.
    I then created a data frame 'tweet_json_df' from the the JSON
file extracting the
tweet_id,'retweet_count','favorite_count','display_text_range','full_text'
columns.
```

Assessing Data

The data quality issues were assessed based on

1. Completeness
2. Validity
3. Accuracy
4. Consistency

Quality Issues:

```
- Inappropiate datatype for tweet_id across 3 tables-- should be
string not int.
- They are a lot of missing data in some feautures in the df_archive
table - e.g "expanded_url" column.
- Missing values are also represented as "None" in
name,doggo,floofer,pupper,puppo tables of df_archive tables.
- The df_archive table contains Retweets while its only original
tweets that are needed
- Inappropiate datatype for timestamp column in df_archive table--
should be datetime instead

-There are unnecessary html tags in the 'Source" column instead of
just source name eg 'Twitter for iPhone' instead of <a
href=""http://twitter.com/download/iphone"" rel=""nofollow"">Twitter
for iPhone</a>
-The rating_numerator of df_archive table has some erroneously large
numbers  (e.g. 1176,960,420,182)
-The rating_denominator column also has other values besides
10(2,150, 170)-- some are due to the text having 2 ratings (like)
figures  in it as seen from the programmatic asessment of
denominators not 10.
-DOg names have inappropriate names
eg(very,unacceptable,this,the,such), all starting with lower cases--
(a validity issue)
-Many tweet_id(s) --186 tweet data rows from the archive table are
missing in img_df (image predictions) table.
-2 tweet_id(s) from the archive table are missing in tweet json
table.
```

Tidiness Issues:

```
  - Dogs kind 4 stages in 4 diffedifferent
columns(doggo,floofer,pupper and puppo) and they should be in one
column.
- p1, p2, and p3 in the img_df table are not properly formatted.
- All three tables need to be merged  into 1.
```

Data Cleaning

-- I made a copy of all 3 tables so as to preserve the original data in cases of mistake.

-- Also Since they all have a quality or tidiness issues one way or the other, and would be needed to be merged into one table, cleaning the 3 datasets was necessary.

-- I performed a programmatic data cleaning process using the Define, COde, Test framework on the copied version of the datasets addressing the issues alighted in my data assesment.

--I thereby merged all the 3 datasets in to 1 using a left join joining on the tweetID.

In [ ]:

Data Storage

After successfully cleaning and merging all the datasets into one, I saved the clean dataframe to file as twitter_archive_master.csv.

In [ ]: