```
#%% md
## Reporting: Wrangling Report

This report gives a step by step explanation of process carried out while cleaning the data given to us. The dataset provided by Udacity was used while working on this project. I imported the dataset using the pandas read_csv and read_json function. I also accessed the data virtually by scrolling through the data provided.
Afterwards, I moved on to access the data programmatically by using the pandas head, info, describe and duplicated function. I also made sure to check for missing values using the pandas isnull and sum function.
After access the function, I came up with the following quality and tidiness issues:

### Quality issues

#### Archived Tweets
##### 1. Missing Values

Missing values are present in the following columns:

* in_reply_to_status_id

* in_reply_to_user_id

* retweeted_status_id

* retweeted_status_user_id

* retweeted_status_timestamp

##### 2. Validity Error.
* The rating_numerator are sometimes above the rating_denominator column.

* Some dogs where without name

```

```
30 #### Predicted Images with Neural Network.
31
32 * Some Images were not Dogs.
33
34 #### Twitter API json file.
35 ##### 3. Missing Values
36 Missing values present in the following columns:
37
38 * in_reply_to_status_id
39
40 * in_reply_to_status_id_str
41
42 * in_reply_to_user_id
43
44 * in_reply_to_user_id_str
45
46 * in_reply_to_screen_time
47
48 * geo
49
50 * coordinates
51
52
53 ### Tidiness issues
54 * The columns converted to strings still has an
   integer datatype.
55 * The data included tweets that were retweeted, and
   not originally owned by the page.
56 #%% md
57 In solving the listed issues, I started with dropping
    the columns with a large number of missing values
   from the twitter_archived data and tweet_json dataset
   . Columns like the in_reply_to_status_id,
   in_reply_to_user_id, retweeted_status_id,
   retweeted_status_user_id, retweeted_status_timestamp
   and expanded_urls. I dropped off these columns since
   filling up this would only lead to inaccurate and
   inappropriate data being added to the data.
58 #%% md
59 Also for some columns present have a wrong datatype.
   I changed the datatype of this column to the right
```

datatype.

I then merged the dataset together using the pandas merge function. Then, I saved the final output of the dataset to twitter_archive_master.csv file which can be seen in the under the Github repository.