Data Wrangling

First step for our project is to gather the data required through different sources. We access the twitter archive data using pandas, read the image prediction data through the Udacity servers and finally scrape data through twitter API.

We then move on to our next task of assessing the data. Through some initital exploration of our data via Visual Assessment we found that twitter_id is the common attribute among our three tables. We would be merging them together later to correct one of the tidiness issue in our dataset and make it more viable for analysis later. We further identified 9 Data Quality issues and 2 Tidiness issues in our Dataset through Programmatic assessment using functions like info(), describe() and Visual assessment like reading through our csv files and through sample().

The issues identified were –

Quality issues

For twitter archive data(tw_arch)

1. Completeness - Missing values of dog names and stages.

2. Completeness- Missing expanded url values.

3. Validity- timestamp, retweeted_status_timestamp should be datetime instead of string.

4. Validity- tweet_id is integer instead of string

5. Accuracy- invalid dog names(all,one,the,a) in some rows

6. Consistency- Removing tweet_ids which are retweets

7. Completeness- in_reply_to_status_id, in_reply_to_user_id , retweeted_status_id retweeted_status_user_id variables not essential to our analysis

For Image prediction(image_pred)

8. Validity- tweet_id should be string instead of integer.

For additonal twitter data(add_data)

9. Accuracy - Although the retweet_count is not zero, favorite_count(number of likes for a tweet) is zero in some rows which is an unlikely scenario for a tweet.

Tidiness issues

1. Each variable forms a column- All columns dogger, floofer, pupper, puppo must be converted into a single attribute as they all represent a dog stage.

2. Each type of observational unit forms a table - We need to merge all 3 tables to get useful insights (all duplicated columns must be removed)

After identifying the issues, we started cleaning our data. We used replace() to replace certain values in the data like invalid dog names, astype() to convert data into different data types to make it more valid, drop() to delete the rows and columns which are not useful for our analysis to make our data more complete.

We further removed the retweeted user ids since we only wanted to keep the original tweet ids and only kept those tweet ids which had images and were present in the image_pred table to make data more tidy without any duplicated rows.

After this we store our master data as a csv file and proceed to Data Visualisation and providing insights through it.