Data Gathering

- Imported this file: twitter-archive-enhanced.csv into “archive_df” Data Frame
- Downloaded “image_predictions.tsv” from the link : https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv programmatically and imported it into “image_pre” Data Frame
- I had some problems to acquire twitter developer account so I used the tweet_json.txt from Udacity

Output:

- **archive_df → The archive DataFrame**
- **image_pre→ The Image Prediction DataFrame**
- **api_df→ The Twitter API DataFrame**

Data Assessment:

Visual assessment:

- **Microsoft Excel**
- **Jupyter Notebook**

programmatic assessment:

- **Jupyter Notebook**
- **pandas.DataFrame.head()**
- **pandas.DataFrame.info()**
- **pandas.DataFrame.max()**
- **pandas.DataFrame.min()**
- **pandas.DataFrame.Series.value_counts()**
- **pandas.DataFrame.duplicated().sum()**

Output:

Tidiness issues:

1. For archive_df Data Frame:
    - For column headers (doggo,floofer,pupper,puppo) are values not variables
2. tables should be merged into one

Quality issues:

For archive_df and imag_pre Data Frames:

1. Not clear column names p1,p1_conf,p1_dog
2. rating denominator have strange values like 170 and 0 where it should be only 10
3. timestamp should be in datetime type instead of string
4. Bad representation of NaN values in doggo,floofer,pupper,puppo column
5. tweets with no images in image_pre should be removed
6. expanded_urls,in_reply_to_user_id,retweeted_status_user_id,retweeted_status_timestamp,p3,p2,p3_conf,p2_conf,p2_dog,p3_dog not in use
7. dogs breed names contain underscore
8. retweets and replies should be removed from the archive table and any related tables

Data Cleaning:

- Structure Used is: Define → Code → Test

archive_df

Quality Issues	***Solution***
rating denominator have strange values like 170 and 0 where it should be only 10	**replace every strange value in rating_denominator with 10**
timestamp should be in datetime type instead of string	**convert timestamp type to datetime using pandas.to_datetime()**
Bad representation of NaN values in name, doggo, floofer, pupper, puppo column	**For None values replace each None with np.NaN in all associated columns** **Using pd.Series.replace()**
tweets with no images in image_pre should be removed	**remove tweets not in image_pre dataframe, using** **pd.DateFrame.merge()**
retweets and replies should be removed from the archive table and any related table	**drop retweets and replies from the archive table and any related tables** **using merge method , indexing and dropna method**
expanded_urls,in_reply_to_user_id,retweeted_status_user_id,retweeted_status_timestamp not in use	**Drop columns specified in issue columns using drop method**

Tidiness Issues	***Solution***
For column headers (doggo,floofer,pupper,puppo) are values not variables	**use concatenation to change values (doggo,floofer,pupper,puppo) into values of column dog_stage** **That is achieved by filling the Nan values with empty string in four columns then concatenate three into a single column called dog_stage which afterwards will be cleaned to represent empty strings by np.NAN and change unreadable strings into readable by putting ‘-‘ between words using methods fillna() , str.replace() and normal panadas addition**

image_pre

Quality Issues	***Solution***
p3,p2,p3_conf,p2_conf,p2_dog,p3_dog not in use as I only used the first level of prediction	**drop columns p3,p2,p3_conf,p2_conf,p2_dog,p3_dog in image_pre using drop method**
dogs breed names contain underscore	**replace '_' with ' ' in breed column** **using replace method**
Not clear column names p1,p1_conf,p1_dog	**Rename the columns [p1, p1_conf, p1_dog] with [breed, accuracy, is_dog]** **Using rename method**

All Tables

Tidiness Issues	***Solution***
tables should be merged into one	**Merge three tables using pd.merge()** **On tweet_id column**

Output:
twitter_archive_master.csv

Twitter_archive_master Data Frame