Wrangle Report

In this project we will go through data gathering ,data assessing ,data cleaning

Gathering

We gathered data from three resources

1. `twitter_archive_enhanced.csv`: file as given, import the data using pandas read_csv into a dataframe
2. `image_predictions.tsv`: file downloaded programmatically using the Requests library , then import the data using pandas read_csv into a dataframe
3. `tweet_json.txt`: file constructed via API(or given in my case ,import data using pandas read_json into a dataframe)

Assess

- quality issue

1. convert ids columns into object datatype , because can't applying mathematical operations on it.
2. Rename id column in tweet_json tablet tweet_id to match other tables when merging tables, and to be more descriptive
3. extract the used source from source column in tweet_json and twitter_archive table.
4. drop null columns(most null) from tweet_json('geo', ‘coordinates’,’place’,’contributors’).
5. convert timestamp ,retweeted_status_timestamp to a datetime datatype in twetter_archive table .
6. replace None in name column to null value.
7. there are 2075 rows in the image_predictions table but for twitter_archive 2356 rows and tweet_json 2354
8. After combined columns in dog_stage , there are a few cases, where a dog has more than one style: (doggofloofer, doggopupper ,doggopuppo) the stages need to be separated with(,) to be (doggo,floofer)

- tidiness issues

1. combine doggo ,floofer ,pupper ,puppo columns in one column dog_stage

2. Merge all three datasets into one dataset called twitter_archive_master

Cleaning

- tweet_json

1. covert ('id' ,'id_str' , 'in_reply_to_status_id' ,'in_reply_to_status_id_str' ,'in_reply_to_user_id' , 'in_reply_to_user_id_str' , 'quoted_status_id' ,'quoted_status_id_str') columns into object datatype,using astype(str) method
2. replace nan with null value after converting the datatype,using pandas replace metod
3. extract the used source from source column,using split() method
4. drop null columns(most null),using pandas drop() method
5. drop possibly_sensitive,possibly_sensitive_appealable not needed in analysis ,using pandas drop() method
6. rename id column to tweet_id,using pandas rename() method

- twitter_archive

1. convert timestamp ,retweeted_status_timestamp to a datetime datatype , using pandas to_datetime() method
2. convert ids columns to object,using astype(str) method
3. source column strip to extract the source, using split() method
4. combine doggo ,floofer ,pupper ,puppo columns in one column dog_stage ,using + operation
5. After combined columns in dog_stage , there are a few cases, where a dog has more than one style: (doggofloofer, doggopupper ,doggopuppo) the stages need to be separated with(,) to be (doggo,floofer),by using conditional assigning
6. replace None in (name,doggo,floofer,pupper,puppo) column to null value,using replace() method
7. Drop column. (name,doggo,floofer,pupper,puppo) columns using drop() method

- image_predictions

convert tweet_id to object ,using astype(str)

- Finally

1. Merge table clean_tweet_json , clean_twitter_archive and clean_image_predictions using

2. there are 2075 rows in the image_predictions table but for twitter_archive 2356 rows and tweet_json 2354. Because the non-pictures retweets included ,
   when merging we will use right join to exclude rows without pictures.