Data Wrangling (Gather ,Asses ,Clean)

Gathering data:

I gathered data from different sources

1. given .csv file (twitter_archive_enhanced.csv) and loaded in dataframe

2. hosted file (image_predictions.tsv) on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv and loaded in dataframe

3.query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called (tweet_json.txt) file and loaded to dataframe

Assessing data:

-I assessed data visually be loading samples in jupyter notebook and in excel sheet and I assessed data programmatically by checking datatypes ,value_counts, non-null values of data , loading samples of data , data descriptions and summaries ,listed all quality issues as changing data type ,removing or replacing null and non-valid values and tidiness issues, combining dog stage data in one column called dogtionary removing duplication ,merging all data sources in 1 dataframe using (inner) merging

Cleaning data:

- **quality issues**

Solved :

1. tweet_id to object in 3 files
2. in_reply_to_status_id to object
3. in_reply_to_user_id to object

4. timestamp to datetime
5. retweeted_status_timestamp to datetime

6. dog stage into category after melting
7. I replaced far values of rating_numerator that count 1 with mean value

8. rating_numerator and rating_domerator have data were not extracted correctly from text.

9. Tweet id must have same name in all the tables as it's the common column as tweet_id
10. using only original ratings (no retweets) that have image so I checked for any tweets with no image to be removed

11. uncompleted sentence in text columns

12. none values and wrong names in name column like a, an, the,... Didn't fix it don't seem necessary

13. uncompleted sentence in text columns Didn't fix it don't seem necessary

- **tidiness issues**

1. melted 4 columns [doggo ,floofer ,pupper ,puppo] in one column and change data-type to category
2. Solved duplication problem caused by melting , removing 2 rows of duplicates in tweet_id and dogtionary duplicates out of 3 , filtering for duplicated in tweet_id then filtering for non values in dogtionary , droping duplicated tweet_id merging data of non- duplicated values.
3. merging 3 dataframes using inner method in master dataframe

- **Storing :** I stored all data after assessing and cleaning in CSV file called twitter_archive_master.csv