Wrangle_report:

I started this project by importing data from Tweet Json, “twitter-archive-enhanced.csv” given by instructor and Image file imported from link provided to students. I did not used API from twitter as I do not want to have Twitter account for personal reasons.

After Importing all data. I started my project by assessing all DataFrames. During this process I found several Quality and Tidiness issues in the data. I am including my steps and description of resolving the issues below:

1. First, I found out that twitter-archive-enhanced has in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp columns which has null values in it. After digging deeper, I decided to go ahead by deleting rows corresponding to them as I just wanted data for tweets and not about re-tweets and reply. After deleting rows attached to them, I decided to remove these columns entirely as they have all null values in them.
2. I also decide to change Datatype of rating to float as, on analysis I found out that some ratings are in decimal in Text column. I also wanted to tackle some numerator rating that were very high. To do this I used regex function and extracted numerator and denominator rating. Next, I decided to focus on rating greater than 20. I replaced all rating greater than 20 by finding rating in text column corresponding to them, as they were extracted wrongly from text column. All other higher rating which were not mentioned correctly in Text column, are replaced by median. Similar all nan ratings are also replaced with median. I standardized the denominator to 10 for clear understanding
3. Next, I cleaned that dog name columns. dog name has many non-dog name such as such', 'a','quite','quite','not','one','incredibly', etc and are in lower case. I replaced all these names to Nan and converted None also to Nan.
4. I realized changing datatype of Timestamp is necessary for Time series, Analysis So I changed it to datetime format.
5. I wanted to see which is the highest used source, for that I extracted source name from the web link in the source Column
6. In Image Data frame I updated p1,p2,p3 to title text and replace “ “, “-“ with “_”. This makes all name in these columns of same type.
7. I also go ahead and deleted rows where prediction were not matching to dogs. This gives also only rows where predictions are dogs.
8. Later I combined stage columns (doggo, floofer, pupper, puppo) into one 'stage' column.
9. I merged all three Data Frame to make one master Data frame named “twitter_archive_master”. I wanted the common columns between twitter_arch_clean and image_clean, so I used inner join here and joined it with tweet_jason using left join so that the all tweet-ids in left data frame will be preserved
10. Finally, I exported the master Data frame to csv file named “twitter_archive_master.csv”. So, this way I go ahead to clean all three datasets as much as possible and made one single source