Wrangle Report

Gathering :

- Reading the twitter archive enhanced csv file using pandas to manipulate the data frame
- Requesting image prediction using requests library then reading it as tsv file using pandas
- Opening tweet json txt file , converting the json file to dictionaries to deal with it then using id , retweet_count , favorite_count keys of the dictionaries to make a new data frame

Assessing :

- Using describe method to know more about the columns of the data frames
- Using info method to know about data types of the attributes
- Finding any duplicates in the data frame
- Finding number of null values of each column in the data frame
- Knowing the value counts of “rating_numerator” and “rating_denominator” to find the most values that have been repeated
- Knowing the value counts of “name” attribute to find none values or entry problems
- Knowing the value counts of “favorite_counts” and “retweet_counts” attributes to find if they have outliers or not

Cleaning :

Quality issues :

- Removing the retweets (text columns starts with RT)
- Dropping in_reply_to_status_id , in_reply_to_user_id , retweeted_status_user_id , retweeted_status_id , retweeted_status_timestamp columns because all of them have null values in the most of the records in twitter archive enhanced data frame
- Dropping " source " column because all the records has the same value and it will be useless attribute
- Removing the null values of " expanded_urls " column to make twitter archive enhanced has not any null values
- Removing "+0000" from time stamp column because it has no meaning
- Converting time stamp column from string to date time data type for the operations we can do on it
- Dropping time stamp column
- Removing the records that does not have name of the dog
- Removing the records that their name is "a" *may be entry problems*
- Removing the url that in the "text" column because it is already in "expanded_urls" column
- Dropping "img_num" column from image_prediction data frame
- Removing the records that did not predict kind of the dog in the neural network that can classify breed of dogs
- Renaming the id column in the json file to " tweet_id " for the merging operation
- Dropping "Attribute" column after the melt process from data

Tideness issues :

- Making the date and time in separate columns
- Solving the stage problem by melting the 4 types of stages in the data frame in one column named “ stage ”
- Using left outer join between twitter archive enhanced and image prediction in twitter_archive data frame
- Then using left outer join between twitter_archive and tw_clean in twitter_archive_master data frame