Wrangle Report

Gathering Data:

For Gathering Data I gatherd data from the twitter account “WeRateDogs”, I downloaded the file twitter_archive_data from the project page, then I downloaded the file “image-predictions” programmatically, and last thing the twitter API which I downloaded from the project page because I didn’t have twitter account.

Assessing Data:

For assessing data I first looked at the data that I have, then typed the problems (2 tideness and 8 quality) which is:

First Tidiness Issues

1- tweet_df and image_predictions_df should be merged with twitterArchive_df

2- the coulmns puppo , pupper, floofer and doggo should be merged to be one column

Second Quality Issues:

1-tweet_id is integer instead of object.

2- converte timestamp to date type

3- both time and date are in one column (timestamp)

4- the columns in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id and retweeted_status_user_id should be string.

5- source column is not readable

6- remove the duplicated rows from jpg_url column

7- the coulmn 'name' have NAN values that assigned as None.

8- drop the columns that won't be used for analysis.

Cleaning Data:

Finally after gathering and Assessing I started Cleaning, I started with tidiness Issues first because fixing it will make the rest easier, I mostly used pandas for cleaning, It wansn’t very easy since it took me some days and I used a lot of websites and some help from a friend.