Data Wrangling Project - Wrangling report

Gathering data using 3 different ways

1- CSV file – Downloaded manually by clicking on the link
2- Image prediction file programmatically imported
3- Import data via Twitter API

Visualize and Assess data

I used two ways to visualize and test data

1- Manually by checking the three data on Jupyter NoteBook and also check Twitter Archive & Image Prediction files visually using excel.
2- programmatically using different methods like checking twitter archive info and making sure Tweets are unique and not Duplicated.

Then I defined the quality and tidiness issues As below

CSV dataset
1- Timestamp & Retweet time stamp are in object data type not DateTime format
2- 745 name are none – 55 a and we will not need the column
3- Delete unnecessary data columns (in_reply_to_status_id, in_reply_to_user_id, r etweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp)
4- Puppo, Pupper, Floofer and doggo represent the dog type or category (One Vari able) so we will combine all of them in one column

Image Prediction dataset
5- Three different predictions for the same tweet
6- Columns name isn't representative for its content

Tweet_Json DataSet
7- Id column needs to be renamed to tweet_id
8- unclean source column
9- unnecessary data columns

10 – 3 different datasets can be combined into one

Cleaning Data

1- I changed timestamp data type to Datetime type, dropped all unnecessary data columns and merged Puppo, Pupper, Floofer and doggo into one column
2- I checked the 3 predictions for the right one and assigned it new columns and dropped the rest of columns with new representative names for the columns
3- I cleaned the HTML link data from source column to leave the source only, deleted unnecessary columns and renamed Id column to tweet_id to keep a standard name while combining datasets

4- Combined all datasets into one new dataset