WRANGLE REPORT

PROJECT : WRANGLING AND ANALYZING DATA (WeRateDogs)

Project Details:

- Gathering data
- Assessing data
- Cleaning data
- Storing data

Gathering Data

I started the project by manually downloading the WeRateDogs Twitter enhanced archive data that was provided by Udacity and loaded it into a dataframe.
I then proceeded to programatically download the 'image_predictions.tsv' hosted on Udacity servers using the 'Requests' library as the second data source into a folder I programatically created. I loaded the 'image_predictions.tsv' file into a second dataframe called 'predictions'.
I then used the 'tweet_json.txt' file which is be obtained from quering the twitter API as my third data source and also loaded it into a third dataframe.

Assessing Data

I moved on to visually assess all the three datasets with Excel before printing them qith pandas in the Jupyter Notebook to get a hint of the data.
I then went on to programmatically assess all the three datasets with multiple pandas methods in order to determine all the issues and provide solutions.

Cleaning Data

Quality Issues	**Solutions**
There were retweets and multiple have plenty of empty values	Deleted retweets and dropped columns with plenty of empty values
timestamp values have a +0000 attached	Truncated the trailing +0000 from timestamp
Erroneous datatypes in tweet_id and timestamp	Corrected the datatypes
Missing column for the fraction: rating_numerator / rating_denominator	Created a new column 'fraction' for the rating fraction
Incorrect dog names listed as a, the and an	Deleted all the incorrect dog names
Source data had unwanted link tags	Removed http code from source data
Dog names listed as None instead of NaN	Replaced all 'None' enries with NaN
The dog breed names are separated by a '_' instead of space	Replaced all hyphens with spaces
Duplicated urls in jpg_url	Dropped all duplicates in jpg_url
Naming convection of id_str not consistent	Renamed 'id_str' to tweed_id

with tweet_id	

Tidiness Issues	**Solutions**
The columns doggo, puppo, pupper, floofer should be combined into a single column	Merged the columns into a single ‘dog_stage’ column
All the tables should be one dataset	Merged all the tables to form one master dataset

Storing data

All the gathered, assessed and cleaned data was merged to form a master dataset and then saved to a CSV file named "twitter_archive_master.csv".