**Project Name:**

[WeRateDogs](https://en.wikipedia.org/wiki/WeRateDogs) Wrangle and Analyse Data

**Project Motivation:**

Wrangle Provided WeRateDogs Twitter data and using techniques for gathering, assessing and cleaning data to be able to provide analysis, insights and visualizations about the cleaned data.

**Data Sources:**

* Enhanced Twitter Archive
* Additional Data via the Twitter API
* Image Predictions File

**Data wrangling consists of three phases and the whole process are iterative in any phase:**

* Gathering data.
* Assessing data.
* Cleaning data.

1. **For Gathering Data:**

* Enhanced Twitter Archive and Additional Data via the Twitter API

have been downloaded manually and imported to notebook using pandas.

* Image Predictions File

has been downloaded using requests package from

url:

<https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv>

and saved locally as'image-predictions.tsv' then imported to notebook using pandas.

1. **For Assessing Data:**

**Assess data from Enhanced Twitter Archive that imported as tweets df:**

- Started by visually seeing samples of the data in tweets data frame then getting info, stats, columns names and datatypes, checking for nulls and duplicated values.

**findings:**

* Unneeded Columns:

interesting columns are: 'tweet\_id', 'timestamp','text', 'rating\_numerator', 'rating\_denominator', 'name', 'doggo', 'floofer', 'pupper', 'puppo'.

* Data Types Issues:

tweet\_id needed to be string and timestamp needed to be datetime.

* doggo, floofer, pupper and puppo should be one column (dog stage) of type category [tidiness].
* ‘text’ column has both text and tweet url [tidiness].
* There are 745 rows with name has None value. [Missing values]
* There are 79 row with name length < 3, names have these values ['a', 'Bo', 'an', 'my', 'O', 'Mo', 'Jo', 'by', 'Al', 'Ed', 'JD' ]
* This list of bad names should be extracted from text or counted as missing name : ['a','an','Al','JD','O','my','by','the']
* There are 1976 row of tweets where all dog 'stage' has None value. [Missing values]
* There are 14 rows have multiple dog stage
* There are 23 rows with rating denominator != 10 , 20 row > 10 and 3 rows < 10 [1 row with 0 ,1 row with 2 and 1 row with 7] ,all rating < 10 should be corrected manually
* After investigating the three entries with denominator< 10, these values needed to be extracted from text column.
* Rating at tweet in index 2335 should be extracted from text
* rating\_numerator and rating\_denominator should be one column reflect (rating\_numerator/rating\_denominator) of type float [tidiness]

**Assess data from images\_predictions that imported as imgs\_predicts df:**

- Started by visually seeing samples of the data in imgs\_predicts data frame then getting info, stats, columns names and datatypes, checking for nulls and duplicated values.

**findings:**

* Data Types:

tweet\_id needed to be string

* images\_predictions df has 324 rows predicted not to be dogs in all three algorithms.
* predictions and predictions\_conf could be merged in just one column with the prediction algo with the heightest conf [tidiness]
* There are 2075 entries in images\_predictions while having 2356 entry in tweets archive [there will be missing data after join]

**Assess data from Tweeter\_api that imported api\_info df:**

- Started by visually seeing samples of the data in api\_info data frame then getting info, stats, columns names and datatypes, checking for nulls and duplicated values.

**findings:**

* Data Types:

tweet\_id needed to be string

* The interesting columns are 'id', 'full\_text', 'retweet\_count', 'favorite\_count'
* There is 1 row with 0 retweet\_count
* There are 179 rows with 0 favorite\_count
* there are 2354 entry in api\_info while having 2356 entry in tweets archive [there will be missing data after join]

1. **For Cleaning Data:**

**Quality Issues:**

* `tweets` df
* Retweet and Reply related columns needed to be dropped.
* Retweets rows needed to be dropped.
* Erroneous datatypes

tweet\_id needed to be string

timestamp needed to be datetime

* name column has entries with 'None' as value
* name column has entries with bad names from list ['a','an','Al','JD','O','my','by','the'] and should be extracted from text or counted as missing name
* 'doggo' , 'floofer' , 'pupper' , 'puppo' columns have 'None' values and there are entries with None values combined in all dog stage.
* rating\_denominator column has values less than 10 [0,2,7] should be extracted from text
* `imgs\_predicts` df
* Erroneous datatypes

tweet\_id needed to be string

* df has entries predicted not to be dogs by all applied algorithms
* `api\_info` df
* api\_info df has many columns needed to be dropped
* Erroneous datatypes

tweet\_id needed to be string

**Tidiness Issues:**

* `tweets` df
* text column should be splitted in two columns tweet\_text ,tweet\_url
* doggo, floofer, pupper and puppo should be one column (dog stage) of type category.
* rating\_numerator, rating\_denominator should be one column reflects rating\_numerator/rating\_denominator of type float.
* `imgs\_predicts` df
* prediction and prediction confifance should be one column reflects the heightest confidence
* tweets df , imgs\_predicts df and api\_info df could be merged in one df.