Wrangle Report

- **Gathering Data** :

The data we gathered were three different data frame

  - The WeRateDogs Twitter archive only contains very basic tweet information.
  - The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. We should be downloaded programmatically using the Requests library. (URL:https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv)

  - Query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file.

- **Assessing Data :**
  - Visual assessment:
    - We DISPLAY to visualize the data set and understand the structured of the data we gathered.
  - Programmatically assessments:
    - We use panda methods to understand the data type, shapes and missing data
  - quality issues:

    - Tweet_id, timestamp,img_num and dog_stages need to be changed into the right datatype.
    - Dog name is not always accurate.
    - Nulls represented as 'None' in columns 'doggo', 'floofer', 'pupper','puppo'.
    - Some errors in rating.
    - Deleting numerators and denominator columns and create new rating column
    - Some tweets don't include images
    - We only want original ratings (no retweets).So the retweets should not be there.
    - We only want ratings with images. Not all ratings have images.
    - clean source based on platform.
    -

- Tidiness Issues:
  - Merging dataframes into one using tweet_id for joining.
  - The columns 'doggo', 'floofer', 'pupper','puppo' show one variable.
  - Unnecessary columns will be deleted.
- Cleaning Data:

1. Tidiness Issues:

Untidy data has structural issues

1.1 Issues: Merging Datasets

Define

Merging the three datasets in one set.

Issue 1.2 : The columns 'doggo', 'floofer', 'pupper','puppo' merge in one column

Define:

Merge dog's stage into one column.

Issue 1.3: Delete unused columns

Define

Drop columns that are not used.

- 2. Quality:

Low quality data has content issues.

Issue 2.1:Tweet_id, timestamp, img_num and dog_stages need to be changed into the right data type

Define:

Change column type to the right types.

Issue 2.2: Dog name is not always accurate

Define

Need to replace lowercase names with np.nan

Issue 2.3: Nulls represented as 'None' in columns 'doggo', 'floofer', 'pupper','puppo'. in Dog_stage

Define: Replace missing dog stages by 'np.nan'.

Issue 2.4&2.5: Some invalid rating

Define:

- Get rid of numerator higher than 10
- Create new column 'rating'
- Delete column numerator and denominator

Issue 2.6: Some tweets don't include images

Define:

Remove all tweets that do not include image.

Issue 2.7: Delete retweets
Difine:
We don’t need retweet that start with RT

Issue2.8: Not all ratings have images.
Define:
Delete tweet with no images

Issue 2.9 : Clean Source based on platform

Define

Delete all tags and keep the platform to read easily

Save & Store New Clean Dataset:

Store the clean dataset as CSV file with name ‘master_archive_master.csv’