Data Wrangling Report

The three main steps in the data wrangling process are gathering, assessing, and cleaning (as shown in Fig.1).

WeRateDogs is a project where users tweet about their pets. Text (comments), ratings (such as 11/10, 12/10, etc.), the source of the tweet (such as iPhone), and the URLs are all wrangled. Sadly, this project requires data from a variety of sources which may come in different formats via an algorithm, site scraping, and other data documentation. Let's gather data! first, before anything else!

First: The following were the data sources used in this analysis:

1. Twitter archive enhanced: Udacity contributed this information for the purpose of data wrangling after receiving it from the WeRateDogs community. As of August 1, 2017, this data, which is in comma separated format (.csv), had more than 5000 tweets.

2. Image prediction: This information was generated by a neural network that anticipated dog photographs and classified them according to the different dog breeds. Its contents contain the tweet ID, picture URL, and image number that matched the most accurate forecast, and each tweet's JSON data is stored in a separate file called tweet json.txt.

3. Web scraping: This additional data includes a list of the tweet ids as found in the Twitter archive enhanced but with the number of favourites and retweets for each tweet. Using Python's Tweepy package, one must gather this data using the Twitter API to obtain the JSON data for each tweet, which must then be saved in a tweet json.txt file.

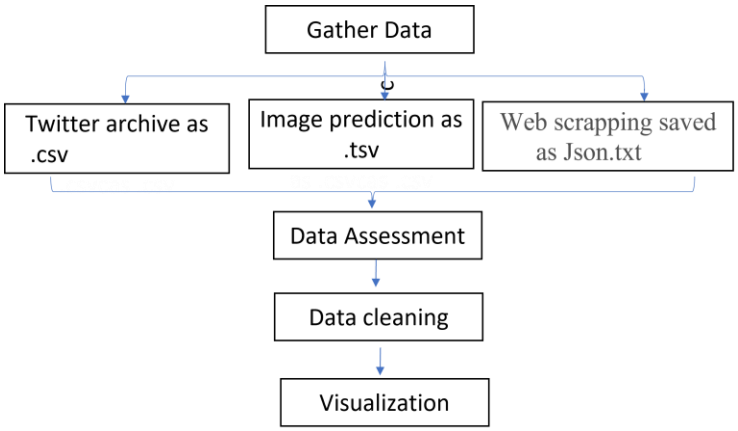

Fig 1: Process flow chart

Assessment and Cleaning

Data type	Type of assessment	Data Assessment	Data cleaning
Twitter archive enhanced as .csv	Visual	'NaN' and 'None' are the same thing +0000 appeared in every single cell The column 'source' has identical 'href' https://twitter.com/dog_rates/status is repeated in the column 'expanded URL' inconstitency issues, a mixture of capital and small letters e.g., row 42	Convert all to NaN Change datatype into datetime that removes +0000

		The column expanded URL has both text and figures	Be consistent with small letters
	Programmatic	In the text column with the index number 870 and tweet_id '761672994376806400' made no sense The names of dog such as 'none', 'a', 'an' occurred 745, 55, 7, and 8 times, respectively. The text column has URLs and hashtag	This may be converted to NaN Separate URLs from text column
Image prediction as. tsv	Visual	False images. Columns p1, p2, and p3 have other animals and other items that are not God.	Filter off the false image the columns
	Programmatic	There are about 281 missing tweet_id because TAE has 2356 while IP has 2075	
Web scrapping saved as Json.txt.csv	Visual	Column label for the tweet_id is represented with 'id' which is not consistent with TAE and IP	
	Programmatic	there are 2 missing rows when compared to the TAE with 2356 rows, RF_tweet has 2354 rows.	
All data frames		Data type of tweet_id	

Tidy issues

1. The timestamp in the twitter archive enhanced dataframe has date and time (plus +0000). The solution is to change the datatype from object into datetime.
2. Tweet_id data type is integer, and should change to string (object)

Result