Wrangle Report

This is a brief report that describes the data wrangling process involved in this project. It is worthy of note that the project leveraged tweets from the twitter account (WeRateDogs) where users give dog ratings based on humorous comment about diverse dogs. In previous times, WeRateDogs downloaded their tweets at some point in 2017 which is not up to date today, hence the need for scrapping data from diverse source for wrangling, analysis and visualization.

The dataset that was wrangled (and analyzed and visualized) is the tweet archive of Twitter user @dog_rates, also known as WeRate Dogs. This is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage

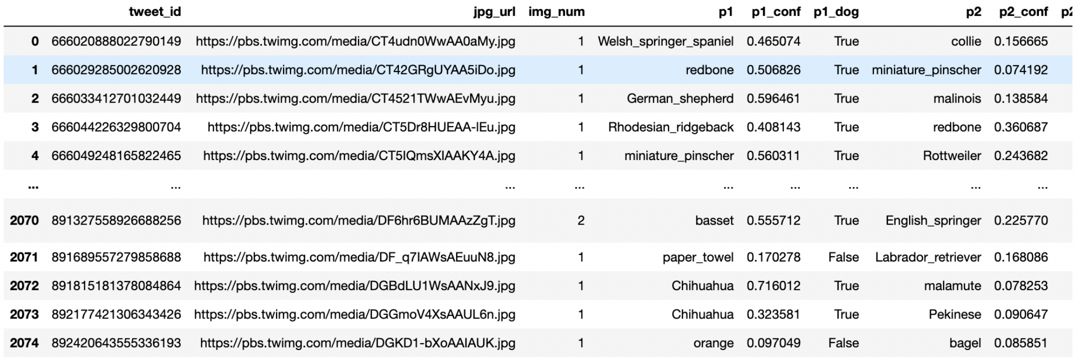
The Wrangling process is divided into three steps:

- Gathering Data
- Assessing Data
- Cleaning Data

Data Gathering

This project leveraged three (3) datasets in which the **first dataset** is the **twitter-archive-enhanced.csv** file which was downloaded through the course resources.

The second dataset is **the image prediction file** which contained mainly top 3 predictions for the corresponding dog. This dataset was obtained with the help of request library that requested to download the TSV of the URL that was specified in the function



	tweet_id	jpg_url	img_num	p1	p1_conf	p1_dog	p2	p2_conf
0	666020888022790149	https://pbs.twimg.com/media/CT4udn0WwAA0aMy.jpg	1	Welsh_springer_spaniel	0.465074	True	collie	0.156665
1	666029285002620928	https://pbs.twimg.com/media/CT42GRgUYAA5iDo.jpg	1	redbone	0.506826	True	miniature_pinscher	0.074192
2	666033412701032449	https://pbs.twimg.com/media/CT4521TWwAEvMyu.jpg	1	German_shepherd	0.596461	True	malinois	0.138584
3	666044226329800704	https://pbs.twimg.com/media/CT5Dr8HUEAA-lEu.jpg	1	Rhodesian_ridgeback	0.408143	True	redbone	0.360687
4	666049248165822465	https://pbs.twimg.com/media/CT5IQmsXIAAKY4A.jpg	1	miniature_pinscher	0.560311	True	Rottweiler	0.243682
...	...	...	...	...	...	...	...	...
2070	891327558926688256	https://pbs.twimg.com/media/DF6hr6BUMAAzZgT.jpg	2	basset	0.555712	True	English_springer	0.225770
2071	891689557279858688	https://pbs.twimg.com/media/DF_q7IAWsAEuuN8.jpg	1	paper_towel	0.170278	False	Labrador_retriever	0.168086
2072	891815181378084864	https://pbs.twimg.com/media/DGBdLU1WsAANxJ9.jpg	1	Chihuahua	0.716012	True	malamute	0.078253
2073	892177421306343426	https://pbs.twimg.com/media/DGGmoV4XsAAUL6n.jpg	1	Chihuahua	0.323581	True	Pekinese	0.090647
2074	892420643555336193	https://pbs.twimg.com/media/DGKD1-bXoAAIAUK.jpg	1	orange	0.097049	False	bagel	0.085851

The third dataset is the **additional data via the twitter API** using the Tweepy library.

Assessing

First of all, the table was assessed by checking the following:

- Datatypes with the **info()** found some columns with inaccurate datatypes like timestamp.
- The duplicated rows with **duplicated()**, no duplicated values were found.
- Whether the ids are unique or not with **nunique()**, no duplicated ids were found.
- Checking for NaN, inconsistent values and inaccurate values.

Cleaning

The following observations were cleaned after assessment:

- \`timestamp\` datatype is \`object\` instead of \`datetime\` (invalid datatype)
- Columns (\`doggo\`, \`floofer\`, \`pupper\`, \`puppo\`) has \`None\` for missing values.
- \`source\` column is html tag \`<a>\` we can extract the source of the tweet and covert it to categorical.
- \`text\` column has the link for the tweets and ratings at the end, so we can remove it.
- \`name\` column have None instead of NaN and too many unvalid values.
- \`id\` column in df_tweet_data name different than the other 2 data sets. (instead of "tweet_id")
- Invalid datatype for \`tweet_id\`(integer instead of string)
- \`expanded_urls\` column has NaN values
- The dog growth stages is divided into 4 different features (\`doggo\`, \`floofer\`, \`pupper\`, \`puppo\`) were combined into one feature “dog_growth_rate”
- Extraction of just 3 columns needed \`id\`, \`retweet_count\`, \`favorite_count\` from the twitter dataset
- All datasets should be combined into 1 dataset only